dissociates from the membrane after proteolysis. Furthermore, it seems likely that there is a narrow exposed region between the two domains and close to the surface of the membrane, which is a preferred site of proteolysis by trypsin. This site probably is found between amino acid residues 335 and 351 since this stretch contains three lysine residues (335, 338, and 340) and two arginine residues (350 and 351) (Lee & Saier, 1983), and cleavage in this region would produce fragments approximately equal in size to those shown in this report. The model also depicts the mannitol permease as a dimer in the membrane, held together primarily by interactions between the intramembrane domains, although direct evidence for an intramembrane dimer remains a subject for further work (Stephan & Jacobson, 1986). This model, with a large cytoplasmic domain, fulfills the requirement for multiple protein binding sites on the inside of the cell needed to perform the multiple functions of the mannitol permease.

Finally, we believe that the techniques described in this report could be of general use for studying the overall disposition of *E. coli* membrane proteins provided that (1) the gene for the protein of interest can be cloned and expressed in a minicell system and (2) the protein is sensitive to proteolysis at either or both faces of the cytoplasmic membrane.

Registry No. Mannitol permease, 91386-44-6.

References

Burnette, W. N. (1981) *Anal. Biochem. 112*, 195-203.

Chamberlain, J. P. (1979) *Anal. Biochem. 98*, 132-135.

Foster, D. L., Boublik, M., & Kaback, H. R. (1983) *J. Biol. Chem. 258*, 31-34.

Jacobson, G. R., Kelly, D. M., & Finlay, D. R. (1983a) *J. Biol. Chem. 258*, 2955-2959.

Jacobson, G. R., Lee, C. A., Leonard, J. E., & Saier, M. H., Jr. (1983b) *J. Biol. Chem. 258*, 10748-10756.

Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105-132.

Laemmli, U. K. (1970) *Nature (London) 227*, 680-685.

Lee, C. A., & Saier, M. H., Jr. (1983) *J. Biol. Chem. 258*, 10761-10767.

Lee, C. A., Jacobson, G. R., & Saier, M. H., Jr. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 7336-7340.

Lengeler, J., Auberger, A.-M., Mayer, R., & Pecher, A. (1981) *Mol. Gen. Genet. 183*, 163-170.

Lowry, O. H., Rosebrough, N. J., Farr, A. L., & Randall, R. J. (1951) *J. Biol. Chem. 193*, 265-275.

Postma, P. W., & Lengeler, J. W. (1985) *Microbiol. Rev. 49*, 232-269.

Roossien, F. F., & Robillard, G. T. (1984) *Biochemistry 23*, 5682-5685.

Saier, M. H., Jr. (1985) *Mechanisms and Regulation of Carbohydrate Transport in Bacteria*, Academic Press, New York.

Segall, E. J., Ishihara, A., & Berg, H. C. (1985) *J. Bacteriol. 161*, 51-59.

Stephan, M. M., & Jacobson, G. R. (1986) *Biochemistry 25*, 4046-4051.

Towbin, H., Staehelin, T., & Gordon, J. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 4350-4354.

Waygood, E. B., Mattoo, R. L., & Peri, K. G. (1984) *J. Cell. Biochem. 25*, 139-159.

Weber, K., & Osborn, M. (1969) *J. Biol. Chem. 244*, 4406-4412.

Complete Sequence and Structure of the Gene for Human Adenosine Deaminase†

Dan A. Wiginton, David J. Kaplan, J. Christopher States, Ann L. Akeson, Charles M. Perme, Ihor J. Bilyk, Arthur J. Vaughn, David L. Lattier, and John J. Hutton*

Children's Hospital Research Foundation and Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio 45229

Received June 9, 1986; Revised Manuscript Received August 19, 1986

ABSTRACT: The nucleotide sequence of the human adenosine deaminase gene was determined. The gene was isolated in a series of overlapping λ phage clones containing human germ line DNA. A total of 36 741 base pairs were sequenced, including 32 040 base pairs from the transcription initiation site to the polyadenylation site, 3935 base pairs of 5′-flanking DNA, and 766 base pairs of 3′-flanking DNA. The gene contains 12 exons separated by 11 introns. The exons range in size from 62 to 325 base pairs while the introns are 76–15 166 base pairs in size. The area sequenced contains 23 copies of Alu repetitive DNA and a single copy of an "O" family repeat. All but one of these repeat sequences are located in the first three introns or the 5′-flanking region. The apparent promoter region of the gene lacks the "TATA" and "CAAT" sequences often found in eucaryotic promoters and is extremely G/C rich. Contained within this region are areas homologous to other G/C-rich promoters, including six decanucleotide sequences that are highly homologous to sequences identified as functional binding sites for transcription factor Sp1.

Adenosine deaminase (adenosine aminohydrolase, EC 3.5.4.4), a part of the purine catabolic pathway, catalyzes the irreversible deamination of adenosine and deoxyadenosine. Deficiency of adenosine deaminase activity in humans is associated with an autosomal recessive form of severe combined immunodeficiency disease (Giblett et al., 1972; Martin & Gelfand, 1981; Thompson & Seegmiller, 1980).

Adenosine deaminase purified from erythrocytes (Schrader et al., 1976; Daddona & Kelley, 1977) and granulocytes (Wiginton et al., 1981) is a soluble monomeric protein with a molecular weight of approximately 41 000. Adenosine de-

† This work was supported in part by National Institutes of Health Research Grant HD 19919, Training Grant T32 Hd 07200 (I.J.B.), and BIONET Grant 1 U41 RR-01685-03.

* Author to whom correspondence should be addressed.

0006-2960/86/0425-8234$01.50/0

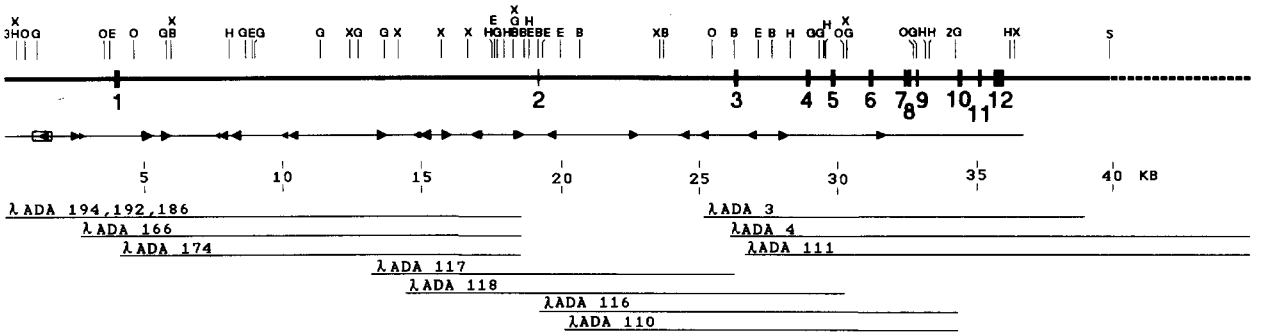

FIGURE 1: Structure of the human adenosine deaminase gene. A partial restriction endonuclease map of genomic DNA inserts in 12 λ phage clones encompassing the gene is shown. The locations and sizes of the 12 exons are indicated by bars on the heavy upper line. The locations, orientations, and relative sizes of the 23 Alu sequences are shown as arrowheads on the lower line. The arrowheads point toward the A-rich tails of the repeats. The open box around the leftmost Alu sequence shows the location of the O family repeat. The dashed section of the heavy line indicates an area in the 3′ flank not completely mapped. Restriction endonuclease sites indicated are B, *Bam*HI; E, *Eco*RI; G, *Bgl*II; H, *Hin*dIII; O, *Xho*I; S, *Sal*I; and X, *Xba*I. Direction of transcription (5′ to 3′) is from left to right.

aminase cDNA[1] has been cloned and sequenced from human T-cell cDNA libraries (Wiginton et al., 1984; Daddona et al., 1984; Valerio et al., 1984; Adrian et al., 1984). The human adenosine deaminase gene has been isolated in clones from a cosmid library constructed from human placental DNA, and the exon–intron junctions have been identified and sequenced (Valerio et al., 1985).

In the present study, we report the isolation of the human adenosine deaminase gene in a series of overlapping clones from a bacteriophage λ library constructed from germ line DNA. The characterization of the structure and sequence of the entire gene and significant flanking regions is presented.

Experimental Procedures

Materials. Avian myeloblastosis virus reverse transcriptase was purchased from Life Sciences, Inc. Sodium 2′,3′-dideoxynucleoside 5′-triphosphates were obtained from Pharmacia P-L Biochemicals, Inc. The 17-base M13 sequencing primer (1212) was obtained from New England BioLabs, Inc. Radiolabeled nucleotides were obtained from New England Nuclear Corp. or Amersham Corp. Bacteriophage λEMBL3 DNA was obtained from Promega Biotec. pUC-12 plasmid DNA was obtained from Pharmacia P-L Biochemicals, Inc. M13 mp10, mp11, mp18, and mp19 phage RF DNA was obtained from New England BioLabs, Inc., or Pharmacia P-L Biochemicals, Inc. Calf intestinal alkaline phosphatase was obtained from Boehringer Mannheim Biochemicals. T4 DNA ligase was from New England BioLabs, Inc. Restricion endonuclease enzymes were purchased from New England BioLabs, Inc., Pharmacia P-L Biochemicals, Inc., Boehringer Mannheim Biochemicals, International Biotechnologies, Inc., and Bethesda Research Laboratories. Human adenosine deaminase cDNA clones ADA211 and ADA33 were isolated and characterized previously (Wiginton et al., 1983, 1984; Adrian et al., 1984).

Genomic DNA Library in Bacteriophage λEMBL3. The genomic library from which clones containing human adenosine deaminase gene sequences were isolated was prepared by Jeremy Nathans, Department of Biochemistry, Stanford University School of Medicine, Stanford, CA. The library was generated by partial Sau3A digestion of human germ line DNA and insertion of the resulting fragments into the *Bam*HI arms of bacteriophage λEMBL3 (Frischauf et al., 1983) generated by removal of the phage stuffer fragment. Recombinant phage in the library were selected after packaging by plating on *Escherichia coli* NM539, a P2 lysogen, thus utilizing the Spi selection system (Maniatis et al., 1982). The library contained approximately 5×10^5 independent recombinants.

Screening of Genomic Library. Genomic clones in the library containing human adenosine deaminase sequences were identified by plaque hybridization screening techniques (Benton & Davis, 1977; Maniatis et al., 1982). Initially the library was screened with previously isolated human adenosine deaminase cDNA as probe (Wiginton et al., 1983). Later some genomic clones were isolated by screening with intronic fragments from prior genomic isolates. Positive genomic clones isolated were first plaque purified by plating at low density and were then characterized and correlated by standard methods of DNA isolation, restriction mapping, and blot hybridization (Maniatis et al., 1982).

M13 Cloning and Sequencing. Specific restriction fragments of the inserts in the genomic clones were either subcloned directly into the M13 cloning vectors mp10, mp11, mp18, or mp19 or subcloned first in the plasmid pUC-12 and then into M13. M13 clones that contained inserts too large to be fully sequenced were used to make RF DNA, which was used as a source for smaller restriction fragments from which were generated new M13 clones. All M13 recombinant constructs were transformed into and grown on *E. coli* JM109 (recA−). Single-stranded M13 templates were sequenced by the quasi end labeling adaptation of the dideoxy chain-termination method (Duncan, 1985). Over 220 clones were sequenced to generate the composite sequence. A total of over 90 000 bases were sequenced. Approximately 80% of the sequence was determined on both strands. DNA sequences were recorded, stored, and analyzed by the computer programs of Queen and Korn (1984).

Results and Discussion

Isolation and Mapping of the Human Adenosine Deaminase Gene. A human genomic DNA library (5×10^5 phage) in λEMBL3 was screened with radiolabeled cDNA probe for human adenosine deaminase. The initial screening produced a large number of clones spanning the region from the 5′ end of λADA117 to the 3′ end of λADA111 (Figure 1). Restriction mapping and blot hybridization indicated that no clones extended beyond the 5′ end of λADA117. Blot hybridization to a 130-bp *Bam*HI fragment from the 5′ end of the ADA211 human adenosine deaminase cDNA clone indicated that these genomic clones did not contain the 5′ end of the gene. The genomic library was then screened with a

[1] Abbreviations: cDNA, complementary DNA; bp, base pair(s); RF, replicative form.

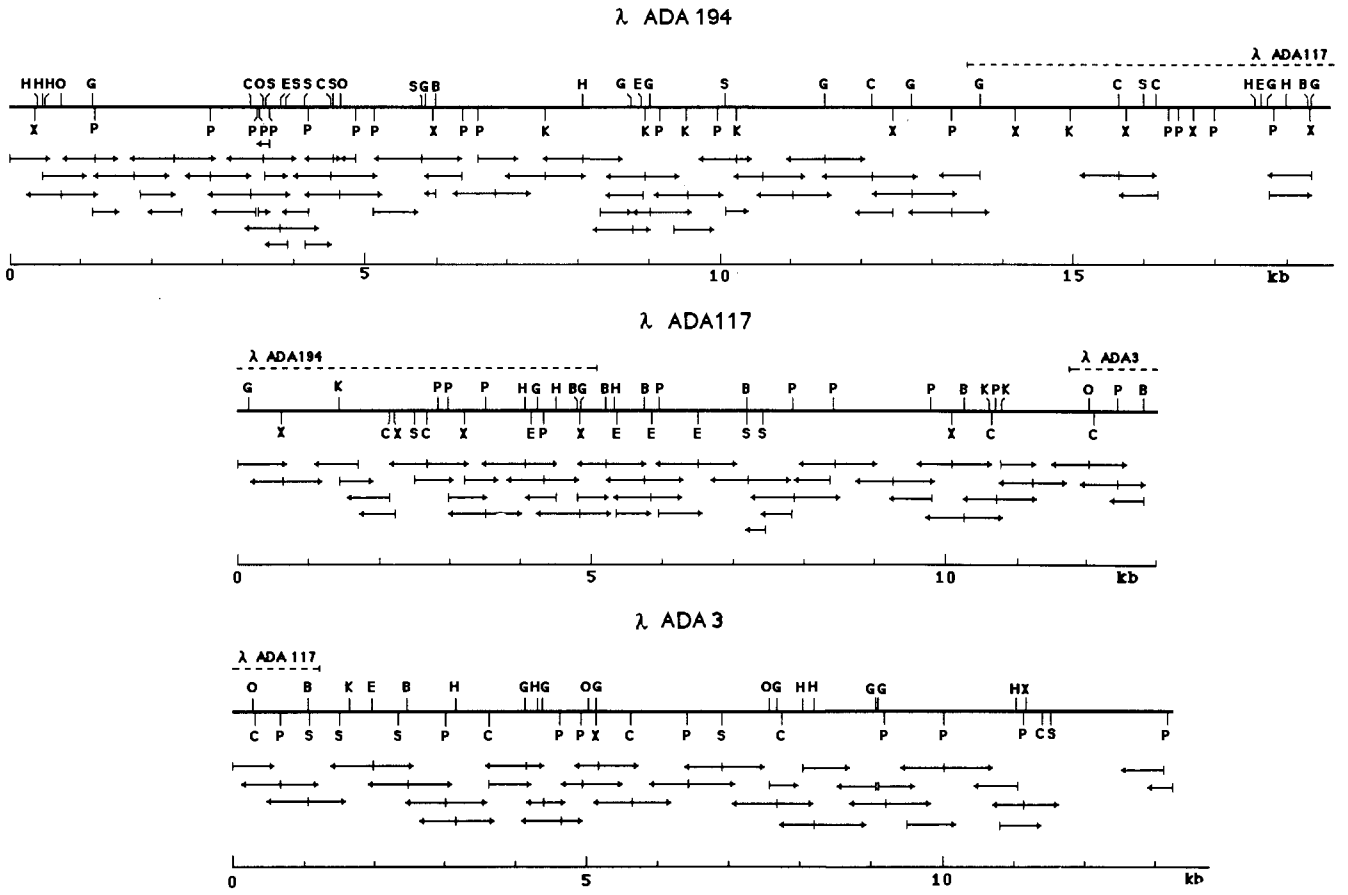

FIGURE 2: Detailed partial restriction maps and sequencing strategies for overlapping genomic DNA inserts contained in the three λ phage clones used for sequencing the human adenosine deaminase gene and flanks. The origin, length, and direction of sequencing by the dideoxy chain-termination method are shown by the arrows and crosshatch marks. The clonal overlaps are indicated by dashed lines. Restriction endonuclease sites indicated are B, *Bam*HI; C, *Sac*I; E, *Eco*RI; G, *Bgl*II; H, *Hin*dIII; K, *Kpn*I; O, *Xho*I; P, *Pst*I; S, *Sma*I; and X, *Xba*I.

radiolabeled 70-bp *Nco*I fragment from the 5′ end of the cDNA. This fragment is extremely rich in G and C residues (80%), and a number of genomic clones containing G/C-rich areas were isolated, but none were related to adenosine deaminase. Next a 192-bp fragment (*Dra*I to *Xba*I) located approximately 500 bp from the 5′ end of λADA117 was used as a probe, and the remainder of the clones encompassing the 5′ end of the gene were isolated.

In Figure 1, a composite map of the human adenosine deaminase gene and flanking regions is shown, including a partial restriction map and a representative sample of the overlapping genomic clones from which the composite gene is derived.

Nucleotide Sequence of the Human Adenosine Deaminase Gene. The overlapping genomic clones λADA194, λADA117, and λADA3 (Figure 1) were chosen for primary sequencing. These clones were sequenced by the methods described under Experimental Procedures. The sequencing strategy employed for each of the three clones is shown in Figure 2. In the areas of overlap among the three clones the sequence was determined completely for one clone and partially for the other. In the overlap between λADA194 and λADA117, the 1760 nucleotides sequenced in both agreed totally. In the overlap between λADA117 and λADA3, the 1090 nucleotides sequenced in both agreed totally except for the Alu repeat sequence, which will be described below.

The entire nucleotide sequence for the gene is shown in Figure 3. The nucleotide numbering was designated by assigning the proposed transcription initiation site (Valerio et al., 1985) as nucleotide number 1. We confirmed this site as at least the major transcription initiation start site by S1 nuclease and primer extension analyses (unpublished results). The genomic sequence contains 32 040 nucleotides from the first base of the first exon to the last base of the last exon. In addition, 3935 nucleotides of 5′-flanking sequence and 766 nucleotides of 3′-flanking sequence were determined.

Gene Organization. The human adenosine deaminase gene contains 12 exons and 11 introns (Figures 1 and 3). The exon–intron junctions (Figure 4) were deduced by comparison of the genomic sequence to the human cDNA sequence (Wiginton et al., 1984; Adrian et al., 1984) and to a consensus splice-junction sequence in eucaryotic nuclear mRNA precursors (Sharp, 1981; Cech, 1983). The splice junctions agree with those described previously for the human adenosine deaminase gene (Valerio et al., 1985). Analysis of the junctions indicates that in general the intronic sequence near the junction is more conserved than is the exonic sequence relative to the consensus sequence.

The locations and sizes of the exons and introns are summarized in Table I. The exons range in size from 62 to 325 bp. The average size of 125 bp is similar to but slightly less than the average size of 150 bp reported for exons in higher eucaryotes (Naora & Deacon, 1982; Blake, 1983). Exon 1 contains the entire 5′-noncoding region plus the first 11 amino acid codons. Exon 12 contains all of the final three amino acid codons as well as the entire 3′-noncoding region. The introns vary enormously in size, ranging from 15 166 to 76 bp. Intron 1 comprises more than 47% of the total exon–intron sequence. Intron 7, the smallest intron, corresponds to the unspliced residual intron found previously within the ADA33 cDNA sequence (Wiginton et al., 1984). S1 nuclease protection studies indicate the presence within human cells of low but detectable amounts of mRNA containing this unspliced intron

Table I: Location and Size of Exons and Introns in the Gene for Human Adenosine Deaminase

exon	nucleotide positions	length (bp)	amino acids	intron	nucleotide positions	length (bp)
1	1–128	128	1–11[a]	1	129–15 294	15166
2	15 295–15 356	62	12–32	2	15 357–22 408	7052
3	22 409–22 531	123	32–73	3	22 532–24 972	2441
4	24 973–25 116	144	73–121	4	25 117–25 887	771
5	25 888–26 003	116	121–160	5	26 004–27 240	1237
6	27 241–27 368	128	160–202	6	27 369–28 489	1121
7	28 490–28 561	72	203–226	7	28 562–28 637	76
8	28 638–28 739	102	227–260	8	28 740–28 915	176
9	28 916–28 980	65	261–282	9	28 981–30 418	1438
10	30 419–30 548	130	282–325	10	30 549–31 164	616
11	31 165–31 267	103	326–360	11	31 268–31 715	448
12	31 716–32 040	325	360–363[a]			

[a] Exon 1 also contains 95 nucleotides of 5′-noncoding sequence. Exon 12 also contains 311 nucleotides of 3′-noncoding sequence following the TGA termination codon.

(Adrian et al., 1984). The significance of this is uncertain.

The body of the gene (nucleotides 1–32 040) is slightly G/C rich (A, 22.1%; C, 26.6%; G, 26.6%; and T, 24.8%, on the mRNA-like strand). Analysis of the dinucleotide frequencies compared to those expected from random distribution shows that only CG (1.7% found, 7.1% predicted) is significantly different from the predicted frequency. Underrepresentation of CG dinucleotide is a well-established characteristic of eucaryotic DNA (Nussinov, 1981).

Repetitive Sequences. Several types of repetitive sequences are present in the human adenosine deaminase gene and flanking region. The most common type is the ubiquitous Alu family of dispersed repeats (Jelinek & Schmid, 1982). Twenty-three copies of this repeat are present within the sequenced area (Figures 1 and 3), accounting for 18% of the total sequence. The Alu repeats are found in both orientations (Figure 1). Four of the Alu repeat sequences are truncated relative to the normal repeat unit of approximately 300 nucleotides, and each of these truncated repeats is located near the tail end of another complete Alu unit (Figure 1). All but one of the Alu repeat areas are flanked by unique short direct repeats of 5–20 nucleotides. Seventeen of the Alu repeats are clustered in the first two introns.

The Alu repeat sequence indicated as nucleotides 20 546–20 838 in Figure 3 is located near the 3′ end of λADA117. This area overlaps λADA3, and the first 236 bp on the 5′ end of λADA3 correspond to part of this Alu repeat. However, this Alu sequence is only 84% homologous in the two clones. The remaining 850 nucleotides in the overlap between the two genomic clones are identical. Included in the sequence differences between the repeats is a *Taq*I restriction site present in λADA3 but not in λADA117. Twenty-five other independent genomic clones spanning this region were screened for this *Taq*I site, and all were found to agree with λADA117. The pertinent Alu repeat regions from two of these independent clones, λADA118 (Figure 1) and λADA175, were sequenced and found to agree totally with λADA117. The Alu repeat difference between λADA3 and λADA117 might be an allelic difference, but the number of genomic clones agreeing with λADA117 would argue against that. The difference may represent a rearrangement or cloning artifact, although a mechanism involving only the Alu repeat is not immediately evident. Genomic DNA from the original donor was not available to clarify this point by restriction analyses. Three Alu repeat sequences in the overlap between λADA117 and λADA194 are identical in the two genomic clones.

All of the repetitive sequences underlined in Figure 3 are of the Alu family type except for nucleotides −2911 to −2579 and −2256 to −2219, which taken together represent a copy of the "O" family of repetitive sequences (Sun et al., 1984). These repeats of approximately 350 bp have recently been proposed as the long terminal repeats for a transposon-like element (Paulson et al., 1985). Solitary copies of the repeat are also found that are not associated with complete transposon-like elements. The repeat in the adenosine deaminase gene appears to be of this type. The copy of the O family repeat flanking the adenosine deaminase gene is interrupted by an Alu family repeat (nucleotides −2574 to −2264) that appears to have inserted into it (Figures 1 and 3), thus generating an unusual "hybrid" repeat sequence. Interestingly, another human hybrid repeat of this type has recently been identified in which an Alu repeat is inserted at exactly the same location in another O family repeat element (Lloyd et al., 1986).

Two unusual direct-repeat sequences found within the human adenosine deaminase gene are shown in Figure 5. Panel I shows a direct repeat (solid rectangle) of 53 nucleotides, the two copies of which are 98% homologous. The repeats are 2600 nucleotides are apart and are associated with Alu repeat sequences, although they are not homologous to Alu repetitive sequence themselves. One copy of the direct repeat has a truncated Alu repeat just downstream (3′) that is highly homologous (97%) to half of the Alu repeat just downstream of the other copy of the direct repeat (Figure 5), although the tails are divergent. The downstream copy of the direct repeat also has an unusual truncated Alu repeat sequence just upstream from it. This direct repeat sequence probably represents a duplication resulting from some type of interaction of the neighboring Alu repeats, although the mechanism is not clear.

Panel II is Figure 5 shows a repeated sequence (53 nucleotides, 93% homologous) located approximately 400 nucleotides distant on either side of exon 5. The significance of this repeat unit is unknown. The repeat sequence was used to search against the GenBank genetic sequence data bank, and no significant homologies with known sequence were discovered. Because of the location of these repeat units relative to exon 5 and because of their high degree of homology, it is possible that they might be involved in gene rearrangements that would be destructive to gene function. At least one mutant adenosine deaminase species with an apparent deletion in the region of exon 5 has been described previously (Adrian et al., 1984).

5′ End of the Human Adenosine Deaminase Gene. The sequence of the promoter region of the gene is shown in Figure 6. As described previously (Valerio et al., 1985), this region is extremely G/C rich (82%) and appears to lack the classical TATA and CAAT sequences present in most eucaryotic promoters. The sequence TAAGAA indicated in Figure 6 at −26 to −21 is at the proper location for a TATA box but is

```
                                                                                                                    GATCT GGGTAAAGGG TTTTCCAGGT GTCAGGATGG  -3901
AAGTGACTAA GGTGCAGAGG CTGGAGGGCT GGGGCAGGTA GAAGCAAGCA TTCCTGTTAC CTACTGCTGT GTGACAATCT CCCCCTAAAA CACAATGGCT  -3801
TAAAATAACA TCCATTTCAT TACATATCTC AATACTATAG GTCAGGAATT TGGGCTGGGC TTACTTGGGT AATTCTTCTG TCCCACATGG CATTGACCAA  -3701
AGCCTGGTTT TCAGTGGGCA GCTGGGCTGG ATGGCCCAAC ACAGCTTCGC TAACATGATT GCTGTCTTCG TAGGGATGGT GGAAGCCTGG GCTCAGTGGG  -3601
ACTGTCAACT GGAATGGCCA TATGTGGACT CTCTTAGCAT GATGGTCTCT TCTAGAAGCT TGGGTTCCCA GAGAGAATGT TCAAGAGGCC CCAAAGGACA  -3501
CCACAAAGCT TCTTTATGAC CAAGGCTCGG AAATCCAGGA AGCTTGCTCC CATCACGCTC TATTACTCCA ACAAGTCACT CAGGCCAGCC CAGGTCCAAG  -3401
AGGAGGAAAC CTAGACTCCA TCTTGCAATG TGAAGAATTG CAAATAATTT GTGTCACCCT TAAGCAACCA GCAACTCATC TAGGTTGATT GGCATTTCAG  -3301
CAATGTGGTG GGAAGTGGTG GGACTGATGT TGAAGAGGGA CTTGAATGTC ATGAGAGGCT GGGGAGGCAA TAAGGTGGGG AGTGAAGTTT CTCGAGTCAG  -3201
ATTCAAATTT AAACCCCAGT TTTGCCACTT ACAACCCATG AGCCAAGCAG GCTGTCTCTC TATCTGAACC TCAGTGTCCT CATCTGTAAA ATGAGGAGAA  -3101
CACCTCCTAC ATCTGAGGAT GACTGTAAAG ATGAAATGGG ATGGGTGCTT ATAAAGTGCT TCCCAGTGTA CCTGGCTCCA AACCTGTCTC AGTAAATGGC  -3001
AGCCCCTATT ATTGAACCCG AGTAACACAG AGAGCCAAGA AAGGATCTTA CAAAAAACTC CCCTGGCTTT GACAATGTAT GAGACCCACT GATAGGGTTT  -2901
GGCTTTGTGT CCTCACCCAA ATCTCATCTA GTAGCTCCCA TAATTCCTAC ATGTTGTGGG AGAGACTCGG CGGGAGATAA TTGAATCATG GGGGATGGTC  -2801
TTTCCCATGC TGTTCTTGTG ATAGTAAATA AGTCTCACAA GATCTGATGG TTTTAAAAAT GGGAGTTTCC CTGCAGGCGC TCTCTCTTTG TCTACTGCCA  -2701
TCCATGTAAG ACGTGACTTG CTCCTCCTTT GCCTTCTGCC ATGATTGCAA GGCCTCCCCA CCATTGTGGA ACTGTAAGTC TATTAAAGCC TCTTTCTTTT  -2601
GTAAATTACC CAGTCTCAGG TATGTCTTTT TTTTTTTTTT CATGAGATGG AGTTTCGCTC TTGTTGCCCA GGCTGGAATG CAATGGTGTA ATCTTGGCTC  -2501
ACCACAACCT CCACCTCCCA GGTTCAAGCG ATTCTCCTGC CTCAGCCTCC CGAGTAGCTG GGATTACAGT CATACACCAC CACGCCTGGC TAATTTTGTA  -2401
TTTTTTTTTT TTTTTTAGT AGAGACGGGG TTTCACCATG TTGGTCAGGC TGGTCTCAAA CTCCCGACCT CAGGTGATCC TCCTGCCTTG GCCTCCCAAA  -2301
GTCCTGGGAT TACAGGCATG AACCACTGCG CCCAGGCTCG GGTATGTCTT CATCAGTAGC ATGAAAATAA TGGACTAATA CAGCCACCCT CTCCCTCACT  -2201
CCCACATACA ACCAAACCCC AAATCCAGCT GATTTTACAC CCTAAATGCA GCTTGAATAT GAGTTTCTCC ACTTCCCCCA CTGACATCAC TATGCCCTAC  -2101
CCAGACCATG GCAGTTGCCT CCTTCCTGGT ATCCTGTCCT CCCTCACCCC CGCTGGCCCC CTGTAATGCC CTCCCCTCAC AGCAGGGAGC CCAGGCTTCT  -2001
CAAAGTGCCC TGTGGGTGCG AACCACCTGG GGGTCCTGTT TGTATAAAAT ACAGATTCTA CTTCAGTAGG TCTGGGATGG GGTCTGAAAG TCTGCATTTG  -1901
TAGTCAGCTC CCAGGTGATG TGGGTGCTGA TGATCCCTGG ATCACACTTT CAGTAGCTGG AGAATATTTT TTCCAAATAA AAGGGTGATT TTGTCTGCCC  -1801
TCCACTTAAA ACACTCCACT GACTTCCTAG GAATCCCACA CCATCGCTGG GTCCCACATC CCTGGCAGGA TTCAGCTCCC ATCAGACCTT CTAGCCCCTT  -1701
GCTCTCCACT CTCCCACTCT CTCTTTCCCC CTTGTTTATG GGTTTGTTAA TTTATTTATG ATGAAATGAA ATGAAGCTAC CATCCACCCC AGTACTGGAA  -1601
CATTATCAAT AACCTGTGTG TGGCCAGGCG TGGTGGCTCA TGCCTGTAAT CACGCCTTGG GAAGCCGAGG TGGGTGGATC ATGTGAGGTC AGGTGTTCGA  -1501
GACCAGCCTG GCCAACATGG TGAAACCCCG TCTCTACTAC AAATCCAAAA CTTAGCAGGG CACGGTGCCA CGCGCCTGTA ATCCCAGCTA CTCGGGAGGC  -1401
TGAGGCCGAG AACTGCTTAA AATCCAGGAG GTGGAGGTTG CAGTGAGCCG AGATTTCGCC ACTGCACTCC AGCCTGGGCG ACAGAGCAAG AGTCCATCTC  -1301
AAAAAAACAA AAACAAAAAC AAAAAAACAA AAAACAAAAA TTAGCCAGGC GTGGTTGTGG GCGCCTATAA TCCCAGCTAC TCGGGAGGCT GAGACAGGAA  -1201
AATCGCTTGA AACGCTGGGG GTGCGGGGGG GCGGTGGGGA GGAGGCGGGC CAGAGGGGCA GAGGTTGCAG TGAGCCCAGA TCGCGCCACT TCACTGCAGC  -1101
CTCCGCGAAA GAGCGAAACT CCGTCTCAGT AAATAAATAA ATAAATAAAT AAATAAATAA ATAAATAACC TGTACCCGCG TGTTATTTCC CTCCGTCCTT  -1001
ACCTCCTCCC GGCTCCTTCC CTTTCACCTG AGATAACCAC TCTTCTCGTA TCTATGCTCA TCTTTCCCTT GCTTTACATT TTTTCCACCG ATGCATGTGT   -901
CTAAACATAC ATACTTTTGG TTTTGCTTTT ACACATTCTA AAAGTTGCAC CATTGTATGC AGTTTTCCGC AACTTAGTTT TTTTCACTCA ACATTGTTTC   -801
TGAGACATTG TTTCTGTTGT TGTCTGGCTG AAGTTCATTC CGTTTCACTG CTGTCTAACG TTTCATGGTG TGAATATTCC GGTTTATTTG CCCACTCGCC   -701
CGTGGAGGGG CATTTGAGGG TGTTTCCAAT GTTCCTGTTA TTCGGAATAG CGCTGGTGTG AACATTCTGC ACAGGTCTCT GGCTGCGCCT GGGCGGGTTT   -601
CTTAAAGGTG AATGCCCAGG AGGGGACTGT CTGTGTTCTC CCTCCCTCCG AGCTCCAGCC TTCCTCGCCT CCTTTCACTC CCAGCTCCCT GGAGTCTCTC   -501
ACGTAGAATG TCCTCTCCAC CCCCACCCAC CCCTGATGAA CTCCTGCAGG TTCTGCAGGC CACGGCTGGC CCCCCTCGAA AGTTCCTTAA CTATACAATT   -401
ATGGTGTGTG TTTCTGCGAC GAGCGTCCGT CTATCCGGTG GAAGGCACGC CGCTCGAGGC TTGCGATGCT CCCGGGGTCC CCGCTTCTAG CTTGGGCCTG   -301
GCGCACAGCA GCGCCCAGAC TGCAGGGGGA CGCTTGAAAG TTGCTGGAGG AGCCGGGGGG AAGGCAGCGC CCAGCGAGGC GGCTGGAGCG CGCGCCCACA   -201
GGTGGGTCCG GTCGGGCGCC GCGGGGCCGT AGTTTTCGGG TCGGCGGGCG AGGACGCCGG GTCCAGAATT CCAGGAAATG CGCGATCCAG GCCGGCGGGC   -101
GGGGCGGGGG CTCCGGCGAG AGGGCGGGCC CCGGGAACGG CGGCGGGCGG GGCGGGAGGC GGGGCCCGGC CCGTTAAGAA GAGCGTGGCC GGCCGCGGCC     -1

                                                                                                                                                    MetAl
.
ACCGCTGGCC CCAGGGAAAG CCGAGCGGCC ACCGAGCCGG CAGAGACCCA CCGAGCGGCG GCGGAGGGAG CAGCGCCGGG GCGCACGAGG GCACCATGGC     100

aGlnThrPro AlaPheAspL ysProLys
CCAGACGCCC GCCTTCGACA AGCCCAAA
                                  GT GAGCGCGCGC GGGGGCTCCG GGGACGGGGG TCCGGCGCCT GGGCGGCCCG AGGGGCTTAG CGGGGCCCAG     200
CCCGGGGCGT CCAAACCCTG GGAACGAACG GGGGCTCCTG CAGGCGAGTT CTTCCTTCGG CTTAGGCCGT GGCTTGCTTG CGGGCTAATC AGGGACAATG     300
GGGCAGAGAA GGTCCAGAAC CCGGAGGCCT CCAGAGTCTG CTTCTGCCCC TGACTTGACC CCTCTGGGTC TCAGTTTCGC TGTCTGTCAA GTGGGCATCC     400
TAGCACCGCT GAGCGCTGTG TGGGCCTGGG CAGGGACTTG AGGTCTCTGA AGCTCAGCTG TATGATCAGG CCCGATGTCT ACGCCGGATA GCGACCTAGT     500
GCTGTGCCCC GCGCCTACTG AGTGCTCAGT GAATGGAAGC AGCTTTGTAC GCCAGCGTTA TGGTGGTGAG CGCCAAGGAG CTCAGGTTTG TGGATGCGCC     600
CCGGGGAAGA ACCGTGAGCC CTGCCAGAAA GGGGAGGGAG GGGAGCAGAG CACCCCCCTT CCCCCGCGCG GGAAGAACAG GAGCTAGGTA GGCCCTGGGT     700
TTGGGGCCCT AGCAGGGTTC ACTCGAGGCC AAGCCATGGC CCACTGGCCC CAGGGGAGAA TCCCCTTGTT TCTCCGCCCA CCAGCTGTGG CGTCTTGGGA     800
CTGTTGGGGT CAGGGAGGGT CTGGACCCCC TTGGCCTGTC TCAGAGTCCG AGAGGAGGGG CCCAGGAGTC TGCCAAGCAG GGTGAGTCAG CCAGTAGGGT     900
GTGAGAGTGG TTGGGGAAGG AGTCAGCTGC AGTCAGCCTC AACTTACCCT TCTAAGAAAT AGGTGTGAGT GGCCCAGGAG GTTGGCTCAC GCCTGTAATC    1000
CCAGCACTTT GTGAGGCTGA GGCGGGAGGA TCATTTGAGT CCAGGAGTTT GAGACTAGCC TGGACAACAA AACTAGACCC CGTCTCTCCA AAAAATAAAA    1100
AAAGTTAGGG GAAGTGTGTG TGGTGGTGCA CTCCCGTAGT CCCAGCTACT CAGGAGGCTG AGGCGGGAGG ATCGCTTGAG CCCAGGAGGT TGAGGCTGCA    1200
GTGAGGTGTG ATGGTGCCAC TGACCTTCAG CCTGGGAGAC AGAGCGAGAC CCTGTCTCAA AAAAAAAGAG AAGAAAAAGA AAAGAAAAGA AATAGGTGTG    1300
AATGATGATG ACAGCTATCA CAAAAGTGCC GGTGAGAATC CAGTGAGTGT GCATGTGTCA GTGAGGGAGA CAGGCTGTGG AGAGCCCACC TACCTTCTGA    1400
GGAGGGTGAG GCCTGGCCCC CACTACTGAT GCCCCCAGCC CAGGGAAAAT GCTCAGCTAC TCCCCGTCAG AAGCTGGAAC GACTGAGGTG CTGTACAAGC    1500
CCTCCTACCC CCACCCCTGC CTCCTTCACG TCTTACTGGA GCTGGGGCCC ATGATTGGCG CCTCCCCTTT GCAGTCTTTT TATTAAATGC TCTGGGCTCC    1600
CTCTGCCCTT GGGCTGGGGA CCCACTGTAC CCTGATGTGA ATCCTATGGC AGTAGCAAAG CTCTTTGATT GGCGGGGTGC AGTGGCTCAC GCCTGTAATC    1700
CCAGCACTTT GGGAGGCAAA GGTGGGTGGA TCATGAGGCC AGGAGTTCGA GACCAGCCTG GCCAACATGG CAAAACCCCA TTTCTACTAA AAATACAAAA    1800
AATTAGCTGG GCATGGTGCG GGCGCCTGTA GTCCCACGTA CGCAGAAGGC TGAGGCAGGA GAATGGCATA AACCCGGGAG GTGGAGCTTG CAGTGAGCCG    1900
AGATCTCGCC ATTGCACTCC AGCCTGGGTG ACAGAGTGAG ACTCTGTCTC AAAAAAAAAA AAAAAAAAAA AAAGGCTCCT TGATTGCGAA CATGTTGGGA    2000
GTTATGGAGA GAACAGCAGG GCCCACTTCT AGAGCACTTG TTGCAGACAC CCATTGGATC CTTGCAGTTC TTCTGTAACA GCCCATCAAG GGAGGGGCTC    2100
ATATTATTAT CCCCATTTTT TGGCCTTGCT CAGTCCTCCC ATCTGATTCA AGCTGGCAGA TCATTTTCCC TATTGGGACC TCAGTGTCCA CACCTGGAGG    2200
ATGGAACATC AGCTGCTTAT GTGGGTGTCC CGTGTCCTGA GTCCCAAGGC CACAAGGTGA TGCTTGAGAG TGAAGGTAGA ATGTTACCTG CCATGTGTTT    2300
GAGGCGTGAC AAATCTTGTA TGATTGTGAG GAGGAACTTG TGTGAGCTGG CAGGAGAAGT GGGAAGGAGT GTGAATCTCA GAGCCACTGT GACCAGAGCC    2400
AGCTCCCTGC CCTCTTGTGG GAGGGACAGA TGACAGTTAT AATTATTAGC ATTACTAGCT GCAGCTAATG GAGTGTTGAT GTTTCTGCCA GGCACCGTTC    2500
TAAACACATT ATCTGCATTT TTTATTTAAT CCAGGCACAG AGAGGTTAAC TAGGCCCAAG ATCACACAGC TAGGAAATGT CCAACTCTGG GGTTTGAGTC    2600
CAAGGGAGGC TGGCTTCGAA ATCCCATGCC TCTAACCATC TTTCCTAAAC TACCTCTGCA GAAGCCTTTG GGGATAGAGG TGCCAGTGCC CCAGGTGCAA    2700
ACCTCCTGAG ACAGGAGCCT TTGCTGTGTC CTTCAGCTTC TCATACCTGC CACCAGCTGA GGCCTGGGAC CTGGTCAGCT AGAAGAAAGC AGAGCAGGGC    2800
AGCGCTTTTC AAACTGCACT CAAGTGGCCT GACTTTTAAT GTTCACACTG TGATTCTGTG TGGGTCGGGT TGGGGCCTGC GATGCTGCAC TGCTGACCAG    2900
CTCCCAGGAA ATGCTAATGT CAACGATCCA GGAACACACT TTGCTTAGCA AGGCCCTAGG CAGCTGCCTT CTGTTGTGCG GGACCCCTAT TGACTCCAAT    3000
GGATATAGCA CCAGGTTCAA GAGGCTACCT TCTTTGGAAG AGGTAGCAAA CAAGATACGG GGTTTTACTG GGGGCTTAGA CACAGGGAAG AGAGTCCAGT    3100
GGCGGCAGAC TGAGCAGAAG AACCGCAACC ACTTGCAAAT CATGCAGTTT ATGTAGCATT TTCATTTAAC ACCTTCTCCC AACCATCTCC ACCTAGTAAC    3200
CTTCATTTAA CCCAAAACAA AGGGCCTCGG TCCCTATACC CCTGTATGGT CAGTGTCCCG TGGGAATGGG GTGGGGCTCA GATGTTCCTC ATAGATAACG    3300
ACTGGATCTC CAGGTTGGCC ACTCTTGGAT TCCTTCGCTC AGAACTCTGA ACACCCATTC AAGTGTGCCT GCCATGCAGG GTCATCGTCA GGGGATGCCC    3400
AAGTCAAGTT TGCCTGTCGG GTGTGCCTCC CATACCCCCA CCTGGTTTGA CTTAGCACCT GCTGGGCACT GGAAGAAGTG CAAAGGGGGG TTGCAGGGGT    3500
GGCCCTTATC AGCCTATGTT CACAGGTGGC ACCAGGCACT CAGGCATTCT GCATCCTGGA GGCCAGTGCT GATCACATGC CTGTTACAAT AATCATAACA    3600
ATAGCTGTCC TTGAAGTAGT CCTGGGTACC AGGTGCCTTC AGTGACTTTT TCTTCTTTGC CAGAATCTCA CTCTGTCGCC CAAGCTGGAG TGCAGTGGCA    3700
AGATTTGGG TCGCTGCAAC CTCTGCCTCC TGGGTTCATG CGATCCTCCT GCCTCAGCCT CCCAAGTAGC TGGGACTACA GGCGTGTGCC GCAGTCTCAC    3800
TCTGTTGCCC AGGCTGGAGT GCAGTGGTGT GATCCTGGCT CACTACAACC TCCACCTCCC GAGTTCAAGC CATTCTTCTG CCTCAGCCTC CGGAGTAGCT    3900
GGGATTACAG GCGTCCACCA CCACGCCCGG CTAATTTTTG TATTTTTAGT AGAGACAGGG TTTCACCACG TTAGCCAGCT GGTCTCGAAC TCCTGATCTC    4000
AGGTGATCCT CCCACCTTGG CTTCCCAAAG CGCTGGGATT ACAGGTGTGA GCCACTGTGC CCGGCTAGTA ACTTTTATCT CACGGAATCC TCTGGACGAC    4100
TTGACAAGGC ATGGGTCTTC ATCCCCATTT ACAGATGAAG AAACTGAAGC TTAGGGAGTG GAGGGACTTG CCAGGGCTAC ACAAAATCTG AGAGCCTTGA    4200
```

```
AGCTGTAGAC TGGCAAGTGA ACAGGTACAG GCTGGGACAG CAGTTTCTTT CTTTTTTTCT TTTTTAGAC AGAGTTTCGC TCTTGTTGCC CAGGCTGGAG  4300
TGCAATGGCA CGACCTCGGC TCACTGCAAC CTTCGCCTCC CAGGTTCAAG TGATTCTTCT GCCTCAGCCT CCCAAGTAGC TGGAATTACA GGCATGCACC  4400
ACCATGCCCG GCTAATTTTT TGTATTTTTA GTAGAGACGG GGTTTCTCCT TGTTGGCCAG GCTGGTCTCG AACTCCCGAC TTCAGGTGAT CCGCCCACCT  4500
CAGCCTCCCA AAGTGCCGGG ATTACAGGCA TGAGCCACCG CACCCGGCCA AGGGACAGCA GTTTCTAAAC TGTCCCTCTC TGATGCAGAG GGGAATTGGG  4600
GCTAAATCAG CAATGTGCCT TTTCTGTCTC ATATTTGAAT GTCTACTCTG CACGAGGCGC TGTCCTGCTT TGCATACAGT GACTCATTTA ATGTTTATGT  4700
CAGCCCTCTG AGGAAGGTCC TGTCCTATTA TTAACTTCAC TTATTATGAG GAAACTGAGA CTCAGAGAGG GGAGGGAACT TGCCAAAGTC ACACAGCTGG  4800
CAAGCAGCAG AGCTAGACTT GAACCCAGAT CTGCCTGCAC TCAAGTAGAA GCTGTTCATT GCTTTGCTCA TTTGCCAATT CCACTTTATG CAAAAAAGAG  4900
GGGGCAGTGT GGGGGGAAGA GTTAGAATCA GGGTGGCAGG GTGGGCCAGT GCATTAGCCC TGGGCTTCAG ATGTACTGGG GTTGAATTCC TGCCTGCCGC  5000
TTAGCAGCTA GGGTACCTCA GGTAGACAAC TCCTGAAACT CAGCTTCCCC CTCTGTAAAA TGGGGTGACA AAACCAAGAT CTTGGGGTTC TTGGGGAAAC  5100
TGACATGCTG ATTGGTTTTT GTACAGTGCC TGGCTGGTAA CAGCAGGCCC TCAGGGGTGC GTTTCCTTCC TGGGGACTGG AGTGGGGGTT GCAGTAGACT  5200
CTGGGAGGCC TCTCCAGCTG CAGAATCTCC CTCCTCCCTC CTCCTTTTTG TCTTCCTGAC ACAAAACCCA CCAGCTGCAC TTCTTTGGGC TTGCAGTGGC  5300
TTTCAGTTAC CAGAGCCACC TGTTAAAACA AAAATGTGCC TAGGAAGAGC CTGCCTTACC CATTTTGACT CACATGGCAG TTGGTGGTGG AGGGGAACAA  5400
AGGAGACTGA GTTTCATCGA AGCCTTTTGC TTCGGAGGAG GAAGGGAGGA TCAGAGAGAG GAAGTGGTCT GTGTTCACAC AGGGAGGCAG GGGAGGCCAG  5500
GCAGCTTCCC AATCCTGCAT TCAACCTCAG GGTGGGCTTG ACCTGGGTGG CTGGGGGCCC TGTGATCCAG GAGAGACTTG TCCACCTGCT CAGGTGTCTT  5600
GAAGGGGTCC CTGTGGTACC CCCTGGGCGG GGCAAGGTAG TAGGACCATG GTCTGGCTGG GGAGGTGGAG AGGAGCAGGC TGTGGGCGCA GAGTGAGGTT  5700
GGAATCTGTA TTTACCCAAG GTGTTGGGGG TAGGCTTGCC CTCAGCCCTT AATGTTCTCA GGCCCCTGAG CAGTTGTGGG GGATAACCTC TGCACTCCTA  5800
GTGACCAGGG AGCTAGAACA GCAAGGAATT TGAACTTGGA CACCAGCTGG GGTCAGGCTC TCTGGGTCTG AGTCCTGATT TCCCACTTTC CAGCTAGAGG  5900
AGCTTGAATG AGTCATTTAA CTTCACGGTG CCTCAGTTTC CCCTCTCTAA AATGAGAATT ATACCCATAC CCACCTCTCA AACACCAAGT GCAGGCCTGG  6000
CTCAGAGCAG GTGCTGCAGC AATAGCTGCC ATTGGTCAGC ATCATCATCA TGGTTGGTAA TGGTCCTACT TTGACTTTTG AGACAGAGTC TCACTCTGTC  6100
GCCCAGGCTG GAGTGCAGTG GTGCAATCTC GGCTCACTAC AACCTCTGCT CCCGGGTTCA AGTGATTCTT CTGCCTCAGT CTCCCAAGTA GTTGGGATTA  6200
CAGGTGTGCG CCACCATGCC TGGCTAATTT TTGTGTTTTT AGTAGAGACA GGGTTTCACC ATGTTGGCCA TAACAATGGC TGTCCTTGAA GTAGTCCTGG  6300
GTACCAGGTG CCTTCAGTGA CTTTTTTTTT TTTTTTTTTT TTGAGCTGGA GTCTTCCTCT GTCACCCAAG CTGGAGTGCA GTGGCACGAT TTTGGCTCAC  6400
TGCAACCTCT GCCTCCTGGG TTCATGCGAT CCTCCTGCCT CAGCCTCCCA AGTAGCTGGG ACTTGGGATA CACTTGCCCC CGCTGGTCCT CCCTTCCACC  6500
TCTGTGAAGA GGAGGTCTCA AACTCCTGGC CTCAAGTGAT CCACCCACCT CAGCCTCCCA AAGTGCTGGG ATTTCAAGAG TGAGCCACCG CACCTGGCCC  6600
CTGTTTAGAT GTTAGCATCA GTGACCCAGC ACCTTGCTAT GTGGCATGCA GGGAGCGTGC TGCTAGACCT CCGGGTTTAG AGTCAAATAG CTTCCTGGCT  6700
GTGGTGTGCA TTAGACTTTC TAACTCAAGG TCCTCCCACT CTCTGAGCCT CAGTCTTGTT GCCTTTAAAA CGAGTTTAAG TGTGCTGAGT CCCTATGCTG  6800
TGGCTCCACA GGAATTTCCC CAGGTGGAAG ACACATCTTG CCTTCTGTGA AACCTCTCAG CAGCAGAGCT GTCAGGCCCC GTCAGCAGGA GACACTGTGG  6900
GGACTGCTCA GTCCCTTCCA CTGTGTACCT CGGAGCTGGC GGAGCCTAGA TGAGGCTGAG CATAGAGGGC TTCCTGGAGG AAGTGGAGCT GAAACAGTTT  7000
CTCAGCCCAG GGCTGCTCTG TCTCCTGGCC TCACACTAAA AGTCAGTTGA GAGGCCATAG TGGCATAAGT CACTGACCCT GGCACTGCCC AGCTCATCAC  7100
CAAAAGCAGG GCTAGGGAGG GAGGGGACAT TCGATTGGCA GTGGGCACCT GTGGCTCATC TGGGTTCTGG CCACGGTGCT CAGGTTCTGT GAGCTGACCA  7200
GGCAGCCCTG GCTCCTCTGC CCCCGTGTGG GTTCTGCCAG GTCCCATGGG GCAGGTCAGC CCCTTCCTTG TTGCAGGGAG AGCACCCAGC ATTGCTGACA  7300
TGGGACAGGG AAACGAGGAA ATAACGGTGT GGTCATTGAA CACAGAGAGC ACTAGGTGCT GTGCGAGGTG CTGAGGACAC GACATGATGA CACAGACAAG  7400
GTCCCCCCTC TCAGCAAACG GCTCATGAGG GAGACAGACA TGTTACATAC ATGAACCCAA AAAGTCAGAC GAAAACAAAA CAGAGCGATG TGTTTGGGAG  7500
GCAAACCCAA CTGCCGGAGG GCGAGCAGTT GGGAACGTGG AAACATGAGT CAGATCTGGG AGTATCTGTC CCAGGAGTCC AAGACCTGGG TCCTCATGGT  7600
AGCTCTGCCA CCGACACACT GAGTGACCTT GGGTAAGTGA ACCCACCGCC CTGGACCTCT CTGGCACGCA TCTCTTGAGA GCAGGGACTT AGTGCATTTC  7700
CCGAGGGCCT CCACGGTGCC TGGCACATAG TGGGGCTTAG TAAATATTTG TTGGTAACTG AGGATGCTTC CTGTTCACAT CAGCGCTGGG AGGATTTCCT  7800
GCTGTTCAGA CAAATGCTGG GCTGGCTGTG AGTCAGCCTT GCAGAGAGCA AAGGCAGTGG GAAGGGGCGT GAGATTCCCC TCTGGAGAGG TCAGGAGGCC  7900
AGGCACTGTC TCGACATGAG TGCCAGGGAG GGGGTGTGGC CTGTGGGCAG GGCTTGGGCT GAGGCAGAGG GACTTGAGTT CCACCCTAGC TCTACCACCA  8000
TCAATTTTGT GTAACTCTGG ACAGGCCACT GAACTTCTCC GGGCTTAGCC TGGCAAGTCC ATTTCCCCAT CTGTAACATG GGCCGATATG TACATTGCCT  8100
AGGGATTAAA TGAGATAAAG GGTCTGAAAA CAGTAGGTAG CTGCTTTATC ATTATTATTA TTTCTGTATT ATTGATGTCT GAGGCTAGGC CCACAGAGGC  8200
AGTACAGTAG AGTGGTTAGG AGCTCAAGAA TCAGACTAGG GTTCAAATTC TGACTCCATC ACTGACTGTT TTGGGGTACT TCTTTGAACC TCAGTTTCTT  8300
CATCAGTAAA ATGGGAGTGA AGTCTCTACC TTGCTGGTTG TAAGGATGAA ATAAGATAAT GCATATAGAT GGTCTAGCAC ATAGTAGATA CTCAAAAGTT  8400
TGAGGCCACT GCTGACCCTT TTCCCTGAAA GGAGACAGGA GAGCGGGGTC GCCACCCCAT TGTCATTGTC ATCTGGAATA GGCTGACAGA CTTCCCATGG  8500
TGTGTTGCAG TTTTCTAGAA AATTCAGTAG GAGGCCTGCC TGAGCTTGAG CCACCTGTGG AGGTGCTTCC TGCCTCTGCT CCACACCTGA AACGCGTCTG  8600
GGCCTCTTCT CAGGCAGCCG TGAGAAGGGA TGAGTGCTAC TGGTCATGGT GGGCAGCTGG CTCTGCTTTC CCCCTTCCCA GAGGCGCTCC TGCCTCCTGC  8700
CCAGCTCCCT GAACCCCTAG CTTCTGCACC CCGGCACTGT CTGGCTTCTG CCCCGCTGAG CACCCACTGT CTCTGACGCT GCCTTGAGTA CTTCCCGCAT  8800
GTTATTCAAA TCCCAATCAG ATCTTCCCTC CCCCAGTAGC TGGTCTTCTG TTCTGGCTTC CTGCCATCCT GTCCTCCACA CAGCAGCCGG GAAAGGTTTT  8900
TTTAAAGGGG ACTCTCCGAT TTAACACACT TGGGTGGAAA ACCCTTTGCT TCGGCCTCTG CAATCTCCCT GCCCCCTCTC CACTTTGCCC TGGCCTCATT  9000
TCTCACCACT AACCTCACTC TGCACTCTGG CCAACTCCCC GCCTGCTTCC TGATTCAGAC ACTAAGCACA CGCAGCTCCC CTGCCTGGAG CCATTCTCCC  9100
TCTCCTTCTT TCTTCTCCCT GGAGAACTCC CCCTTTAAGT GATCTTTTCC CAACACACTT TCTAAATTGC CCCCACCCCA GTGTGATTTT TCTTTATCTC  9200
ATAGCACTTG GTCTGCTTCT TATCACAGTT TGCAAGGCTG AGTTCAGAAA GGTGTGTTTG CTCATTCTGA GGCAGGAGAG GCTACCTTGT GCTGCTGTGG  9300
TAACAAACAG CCCCCAGGTC TGAGGGGTCT GCAGAGACCC AGGTTGACCT CATACTGCTT GTCCCTCCAG GGCCTCCAGT GAGGTTTCGG CTCCTTGGAT  9400
CACTCAGGGC CCCAGGCAGA TGGGAAGATT CCACTCTGAA CATTGCCAAT TGTTGTGCCA GAGTAAAGCA GAGCTGGGAG GTGGGCTCTT GAATTGGCAT  9500
TTAAATACTT TTGCCAGGCA GGGTAAGGCA GCTCACGCCT GTAATCATAA CACTTTGGGA GGCCTAGGTG GGTGGATCAC CTGAGGTCAG GAGTTCAAAA  9600
CCAGCCTGGC CAACATGGTG AAACCCTGTC TCTACTAAAA GTACAAAAAT TAGCCGGGCA TGGTGGTGGG CGCCTGTAAT CCCAGCTACT TGGGAGGCTG  9700
AGGCACGAGA ATCCCTTGAA CCTGGGAGGC AGAGGCTGCA ATGAGCTGAG ATCTTGCCAC TGCACTCCAG CCTGGGCAAC AGAGCCAGAC TCCATCTCAA  9800
AAAAAAAAAA ACAACAACAA CAAATAAATA AATGAATAAA TACTTTAGCC AGAAGTAGCC ATGCAGACCT CCCCCCACCA GTCCCACCCA CAAGCGGACG  9900
TGACTACCGC CCCCATTCAC TGCCTGATCC TCCTGTTCTC AGGGGCTCCA AGGCCAGGCC TGGTTTGACC TTCTGACTTT CTGACTTCCT CCTACCTTCC 10000
CAGTAACCTC ATGCAACTCC TTTCACTCAG CCTCAATCAT CCCCATGGGT GTTTAAACTT GCCCAAGACA TGCCCCTTTG AAAAAGCCTG CCATTCTCTT 10100
GACCCACATG CACGTCCTGC CCCCTCCAAG GCTGCTAGTT CCTTTAGGGG CAAAATTGTG AAAGAGTAGT CTAAACCTTC TTCCTCTTCT TACCTCCACT 10200
TCTTTCTTAC CTTATTCCCA TGTGGATTCT ACCCTCACTC AGGCCTCTAG AACGGTTCCT CTACGGCAGT GGTTCCCAAT CTTGACTACG TGTTTTTTTA 10300
AAAAAAGTCC TCCACCTGGG CCTGCCACCA AGGATTTTTC TTTAATTGAC CTCAGATGGG GTTGAGGCCT TGGGAACTGG CCAGAACTTC CCGTGCTCCT 10400
AACTTGCAGC CGGGGTTAAG AACTACTCCT CTGAAGCCCC CAGTGCCTGC GCTTTTAGCC CGACGGACAA GTTTCTGCCC TTCCATCCTG TGACCTCCAG 10500
CAGGGCCTGA CCATGTGAGT TTTCTGTGGC TGCCGTGACA AGTTGCCACA CCCTGCATGG CTTCAACCAA CAGAAACGTG TGCCCTGGCA GTTCTGGGGG 10600
CCACAAGTCC AACATCAAGA TATCATCAGA GCCACATGCC CACTGAAGGC TCTCGGGGGA ATCCATTCCT TGCCTCTTCT GGTTGCTGGT GGCTCTAGGC 10700
ATTCCTTGGC TTGTGGCTGC ATCATTCCAG TCTCTGCCTC TGAGGTCACG TTGCTGCTTC CTCTTGTGTG TGTTTCTCTT AAAACTCTCT GCTTCTGTCT 10800
TATAAGGATA CATGTGATTG CATCTAGGGC CCAACCAGAT AATCCAGGAT AAACTCTTCC TGTCAAGACA TTTAATAATC ACACTTTGCC ATATAAGGTA 10900
ATTTTTTTTT TTTTTTGAGG TGGAGTTTTG CACTTTCACC CAGGCTGGAG TAAAGTGATT TAATCTCGGC TCACTGGAAT CTCTGCCCCC AGGTTCAAGC 11000
AATTCTCCTG CCTCAGCCTC CTGAGTAGCT GGGATTATAG GTACCTGCCA CCATGCCCAG CTAACTTTTG TATTTTTAGT AGACATGGGG TTTCACCATG 11100
TTGGCCAGGC TGGTCTCGAA CTCCTGACCT CAGGTGATCC ACCCGCCATA AGTTAATATT TTTTTTTTGA GAGGGAGTAT TGCTCTGTTG CCCAGGCTGG 11200
AGTGCTAGTG GCTCAATCTC GGCTCACTGC AACCTCCGCC TCCCAGGTTC AAATGATTCT CCTACCTCAG TCTCCTGAGT AGCTGGGACT ACAGATGCAT 11300
GCCACCATGC CTGGCTGATT TTTGTATTTT TAATAGAGAG GGGATTTCAC CATGTTGGCC AGGCTGGTGT TGAACTCCTA ACCTCAAGTG ATCCACCCAC 11400
CTCAGCCTCC CAAAGTGTTG GGATTACAGG CATGAACCAC CACGCCCGAC CCATATAAGG TAATATTTAC AGGTTCTGGG GATTAGGATT AGCATGTAGA 11500
CAGCTTTGTG GGGGCCACCA TTCAGCCCAC TATGCTAACC CTGTGAACCG TTGCTCGCTT CTCCTTGACA TCTGACGGCC TGGCCTTCTG CATACCACAC 11600
ACCCTCCCAC CTCTCTGGCC ACAGTTCTGT AGGCTCAGCC TCCTCCGTAA GGCCATTAAG TGCTTGTGCT GGTCAAAGTT TCATCCTAGG CCTTTTCCTT 11700
ACCTCCCTTG ATATTTTCTC CCTAGGTGAG CTCCTTCAAG CCCACAGCTT CTGTGCTTAC CCACACTCCT ACCTACATTC CCAGCTTGGG CTTCTCAGGC 11800
CAGCTCTAGA CTCTTGTATC CCACTGGGTT CTTCCACTTA CCTTTGGATA TCTCAAAGGC ATCTCCAGTT GGGTGGGCAC GATGGTTCAC ACCTGTAACC 11900
CCAGCACTTT GGGAGGCCGA GGTGGGCAGA TCACTTGAGG TCAGGAGTTC AAGACCAGCC TGGCCAATAT GGTGAAACCC CATCTCTACT AAAAATACAA 12000
AAATTAGCTG GGCATGGTGG TGGGTGCCTG TAGTCCCAAC TACTCGGGAG GCTGAGGCAG GAGAATCGCT TGAACCCGGG AGGTGGAGGT TTCCGTGAGC 12100
TGAGCTGGAG CCACTGCACT CCAGCCTGGG CAACAGAGTG AAACTCCGTC TTAAAAAAAC AAAAAACAAA AGGTGTCTCT AGTGTAACAT AACTAAAACC 12200
AAACCAATCA TGCCTCCCTC CCCCGCATCC TCCCTCCTGG AGGGAGCTCC AGGACTTGGT CTTCTCTTCC AGAGTTCTCT GTCTCAAACT GCGGGAATTG 12300
CTCCCCACCC AGGCCTAACC TGAAGTGTGA GCCTTGGCAT CTCTTTCTAT CCACCTGTTT TTCCTCTATG CACCTCACAA CCCTGGTCCA AGCCACCGTC 12400
ATCTTTCAAA TGGCTGCAGT AGCCTCTAAC TGGCCTTGGA GGAGCCATCC TCTTTCTCTA ACCAGCTGCC AACCCTGCAA TGGCCTCTGT GTGCTTTCCA 12500
GATAAAGCCT GACTCCTCGT GGCCCGCACA GCCCTGCCTG GGTGGTCCTA TCCTGCAGCC TCTCCAGTAC CATGAACCCT CCCTTCTCTG AACCTCTATT 12600
TAATCCATTT CATATACCCC GTTTTCTCCT GCCATAGGGC CTTGCACATG CTGTTCCTTC TGCCTGGAAT TTTCTTCCTG CCTCCCTCCG CACCCCTGCC 12700
TTGTGTTGTG GGTTCCTCGC TATCCTCTAG CTTTTCGCTC AGGCTCATTG TTGGCCCTCT AGATGTATTC ACTTCTCTTG TTTGTTACCC TCTGTCATAG 12800
GACTGTGTTC GTACTTCCCA AGGAGTCGTC TTGGTTTGTG ACTGTACATT TTCCCATGTG ACATTTGCTT AATGCCTCTC CCACTCTGGG GCCTGTACAA 12900
```

```
GCCCCAGGAA CAGGACTTGG ACCCTCCTGT TTAACTCTAC AATCTAGCAT CCAGCAGGCG CGCAGGCCTT CGTTGACTTT TATTTTATTC TTATTTTTTA  13000
TTTTTGAGAT GCAGTTTCGC TCTTGTCGCC CAGGCTGGAG TGCAGTGGCG TAATCTCGGC TCACTGCAGC CTCTGCCTCC CAGGTTCAGG TGATTCTCCT  13100
GTCTCAGCCT CCGAAGTAGC TGGGATTACA CGTGTGCCGC CACACGCCTG GCTAATTTTT TGCATTTTTA GTAGAGATGG GGTTTCACCA TGTTGGCCAG  13200
GCTGGTCTCA AACTCCTGGC CTCAGGTGAT CCACCCACCT CGGCCTCCCA AAGTGGCTGG ATTACAGGGG TGAGCCCCCA TGCCCAGCCT TCATTGACTT  13300
TTAGTTGACA ACTATTTAGC ATTTGCTATG TGCCAAGAAC TCCCTGCCTA CTAATGCAGT TAACCCTCAT GAAGCCTAGA AGGAAGGACT GCCATTCTCC  13400
CCACTTAACA GATGAGGATG CCGAGGCACA GGAAGTGAAG TGACTTTCTC AGGGTCAAGC AGGGAGTGAG TGGAGGAGCC GAGATTCCAG CTCTAACCGC  13500
ATGATGCTCT ATACAGTGTG ACTCCGGCTC TCTGGCTGGG CCCTCTCCAT AGCCCTGTGA GGGTTAAGGA TAGAAAACAG AGGCTCAGAG AGTTGAGGTC  13600
CCTTGCCTGA GGTCACACAG CTGGTTGGCC GTTCCCTGGG CTATAAGCTT CAGTATTCCC AATGCTGAGC ATATTTTGAG AACCCGAGAA ACAGACGTTT  13700
GGCTGGGTGG GAACTGAACT CATTTTGTCA GGGAATTCAA CAACTAAGTT GGCCCTGAGA CTGGGTGTGA AGACCGCTCT GTCCCCTGCC AGCTGGATGA  13800
CCTCAGGAGA GATCTGATGA CTCTGAGGTC CTGCTGATAG GACCTCTGGT GTCTCTGTTC CCTGCTGGCC TCCCCTGGGC CTGGGTTGGG TTTCCTCTGC  13900
AGGAGGCAGC TCATGTATGT GCTCCTAGAC GCCCTTGGGC CAGCAGCTCC TTGGCTGTTC CTCCCTGAGC CAGGGCAGCC AACTTTCTTA TCCAGCTCTC  14000
CATGCTCCCC ACCCCAGCAT GAGATGTCAG CTGAGAGTTT TCTGGATCTC CCCTAGCTAG GGGGAAAGCT TCCATCATTT GGAACAGGAA CAGCAGGAAC  14100
AGCAAAGTCC CTTTCCCCAC CATCTCCCAC TGCCTGCTGT GCTTCTCCTA ACAGCTCATG GTAAACACCC TGACTGAGCG GCAGGGGCTG TTTCCTTTGG  14200
GCTATCCATG TCCACCTACA CTGCCCTTTT TAATCCTTAC AATTTTTCTT GGACACGGGG GCATAATATT CCATTGTTTT TCAGTTGAGG AAACTGAGGC  14300
TCAGAGAGGT CAAGTGTCTT GTCTGAGGTC ACACAGCAGA ACTGGGAGTC AAGCCAGATG GGCTGCCTCC AAGGATCCTA CTCTTAAACT CTAGAGTACT  14400
AGAAAGATCT TCCGTTGCCT AATATTGATT CCTGATAGGC TATGCTTGAG TAGCATCTGC TTTTGAAAAT GGAGCCTGGG TCGGTTGCGG TGGCACATAC  14500
CTGTAATCCC AGCACTTTGG GAGGCTGAGG TGGGTGGACA CCTGAGGTCA GGAGTTCGAG ACTAGCCTGA GCAACATGGT GAAACCCTGT CTCTACTAAA  14600
AATACAAAAA TTAACTGGGT GTGGTGGCAC CTGCCTATAG TCCCAGCTAC TCCGGAGGCT GAGGCACAAG AATTGCTTGA ACCCAGGAGG TGGAGGTTGC  14700
AGTGAGAGGA GATCACGTCA CTGCACTCCA GCCTGGGAGA CAGAGCGAGA CTCCATCCGT CTCAAAAAAA AGAAAACGAA AATGGATCCT GAATTTTGAA  14800
ATATGCTGTG ACTCTTCCCT AGTTTGGGAC ATCTGGGTCA ATCCCTTTTG TTAAAGTAGT TTATTTAGTT GGCTGAGAGC GGGAGCTGCC TACGTGACCT  14900
GGAGCACAAG CTTTGGAATT GGGCTTGGGT TAGAATTCCG CCTCTGCCAC TCACCAGCTG CGATTAAGAA CAAAGATACT GGGTTGGGCT CCTGCCTCTA  15000
TTACTTGCAA TCTGTGTGGC CTTGGATGAG ATATTTAACA CCTCCGAACC TCAGTGTCCT CAATTGTGAA AGAGATCGAG ATAACAGCTG AACCCACATC  15100
CCAGGAGCGG ATTAAATGAG ATAGTGCAGT ACAGAGTTTA CCGAAGTATA TGGGGTCAGC AGCCAGCCAG TAAAATGGTG GCTAATGGTT ATCATGATTA  15200
ATGTTAACAT TAAGCTCTGA AAGGTCCTTC GTGAACTCAT AGGTATTTGT TCTCTCTCTC CCTTTCTCTC TCTCTTCCCC CTGCCCCCTT GCAG
                                                                                                    ValGlu
                                                                                                    GTAGAA  15300

LeuHisValH isLeuAspGl ySerIleLys ProGluThrI leLeuTyrTy rGlyAr
CTGCATGTCC ACCTAGACGG ATCCATCAAG CCTGAAACCA TCTTATACTA TGGCAG
                                                        GTAA GTCCATACAG AAGAGCCCTC TCTCCCTGGG ATTTGAGTGG  15400
GGTCCCCAGC TCCACCCAGA GGCCCCTGGG GAATTCCAGG GTCACTGTTC CTTCCTGTCT CCCTGTGGGA ATCAAGCCAG CTCCAGGCCA GAAGTGGGAC  15500
TGTGAGGACA TGGAGGCCTC GGCACTGAGC TGCAGACCCG CAGACCAACT CCTGAGCTTT CTGGGCCTCT GAGTCTTGTC CTCCTGGTGT CAGGTGAGCC  15600
AGGCCTGAGC CTGCTCTCCC CACCCACCCA CATACGTGCA TGAAGGTAGT TCCCAGGGCT GAATCCGTCT TTTTTTTTTT CTTTTGAGAT AGAGTCTTGC  15700
TCTGTCGCCC AGGCTGGAGT GCAGTGGCAT GATCTCGGCT CACTGCAACC TCCACCTCCT GGGTTCAAGT GATTCTCCTG CCTCAGCCTC CTGAGTAGCT  15800
GGGATTACAA GCACATGCCA CCACATCCAG CTAATTTTTG TATTTTTAGC GGAGATGGGG TTTCACATGT TGGCCAGGCT GGTCTCGAAC TCCTGACCTC  15900
AAGTGATCCA CCCAGCTTGG CCTCCCACAG TGCTGGGATT ACAGGCATGA GCCACTGTGC CTGGCTCCTG TCTTTTGACT TAACTGAGAG CCTATATATA  16000
GCAGGTGATG TGCTCACATG AGATGCCAGT ACAATTTCTT GAGCATCTCC TAGAGCTGGG CTGGGCTTTA TCAGCTCATT GAATTCCTCC ACGCTTGGAA  16100
GAGGAGGATA CGCTCTCTGC ATTTTACTGA GGAGGGAATG GGCTCAGCCA AGACAGTTGT CCACGGTCAC ACAAATTAAT AGCAGATCAA GAGTTGAACC  16200
CAAGGCTGTC TGACCCCTAA GGCTTTACTA CATCATCAGG GTCATAACCT GCTAGGAGTC ACGGAAAAGT GGCTCCCCAA CTCTGGGCCT AAATCTCTGC  16300
ATCTTCCAAG TGAGAACACA CTTCCTGCCT CAGCTCTCAG AGATGCTAGG GGGCCAGAGG GTCCCCCTGT TCCCCAGCGA GGAAGGTTCT TCCCTTCCTA  16400
CCCAGACCTC AAGGGCTCAC AGCAGCTCCT CTCTTAGGAC CAGCTTTTAA GGGCAGGGAC TTTAAAGGCC AGTGGATCTG GATTCAAATT TGGACATATT  16500
ATCTCCTGTC TGCGAACTTG GTCTCTATCA ACTGAGGCTA AGAACAGGCC CTCCCTAGAG AGATGACCTA GGAGCTAGGG GCTCCTTGTC CACCCAGCCC  16600
TGCCCCCGCA GACCTGTGTT CCTCGGATGT TTGCACAACA CTCATTTTGT TTGGAGCTGA AAGAACTCAG CCTCTCTGTC ACAGTCTTGA AATTCAGCTC  16700
GGGACCCAAA TTTGAACATT TCTGCTCCAT AAGCCAGAAT CCTGTTATTC AGAGGCCTGC CCTCATGGAG AGAATGAGGG ATCCCGGGGG TTGCCCCCAA  16800
CTCTCGGGAG CATCTCCACC AACTCCCTGA GAGATTTCTG GTAAGTCCAC TATTCTCCAT CTTTTCACAC TTCCAGGGAC CTTCTTCTGC CCCAGGAAGC  16900
TGCCATTGAT TTAATTCCTA TTTAACTGCA AGGCATAAGC ACAGTAGCAC CTCCTGTGTG CCAAACACTC CTTTAAGTGC GTTACCCGGG TTAAGTTATT  17000
GAAGCCTCAC AACAATTTGT AAGATAGGAA CTCTATTGCC GTCATTTACA GATGAGGAGA CTGAGCCGTG GTAGGTGGAG TAAGGTGCCC AGTAAGCACA  17100
GGGCGGAGGT TTGAACCCAG ATAGTCTGCC CCCGAGTCCA TGGCCCTGGC CATTACCCCC TGTCAGTTAG AGGTTTTGGT AAGTGATGCC CGTAAAATGC  17200
TTAGTTCAGG GCCTAGCACA CATTAATGTG CTCCATAAAT GTCACTTAAT GATAATATTC TTATTAATTG GAGCTTATAT CTCTAAGTGG GGTGAAACCT  17300
CTTGGCTTAT CTCTGCCTGG CCTTTGCCCA TGTCAAGCCG CCAACTTGCC ACAAGGCCCC TAATGAGGTC GTTCAGTGGG GCACCAAGAT GAGATCGAAC  17400
CCAGGCACTC ATTAAGGGGT CACGGAGGGC TCATCAGCTG CAGCCAGGGG CTGGGAGCGC CGGGTGGGGC TAAGAGAAAG GGGAAAGGAG CCGCCGGGAG  17500
GGGCACTGGT CTGATCGTCC ATTCCTCACA CCACCTCTGG GCCTTGGAGA TGGCGTGCGG CAGGTGCCAG CTGGAGCTTG GCCTGAAGTC AGCAGGCAGG  17600
GGACTGGGGA GTTTGTCACA CTCAGATATG GGTGTCTGTA AATGCACACA AATATGGGCT AAGAATGGAA GGAGGAGGGG AGCCCCTGGC CTGAGCCCTG  17700
CTAGGCCCAA TTCAGTGGCC CTTTTTCCAG CTCTGGGACT CAGGCCTGCC TCATTAACTG TCCTCACCCA TTTCTCCTTC CTCCAGTTCC CAGGATTCTG  17800
GCCTTTTCAG GGGCCTCTCC AACCTCTTTC TCAGTCTTGT TTATAACCCT GTCAACTATT TCTACAGAGA TTCTGAAACT GGCTGCTCTT TCCTCCGATC  17900
ACTGCCCTGG TCTGGGCCAC CACTGCCCCT CCCTGGTGCT GTGGCCTCCT GATTGGTCTC AGCCATCTAC TCTGGCCTTC CTCTCTACGG GCCCTGCAGT  18000
GCTGTAGTTG GAGCAAGAGC CTTAACCCAT GGTCTTCCCA GCTCATTCCC CAGCTTCCCC ATCTCACTCA GAGTCAAAGC CAAAGTCCAC ACATGGGCCT  18100
TAAAGTTCTG CAAAGCCTGC ATTGCCTCTC TGACCTCTCT AAGGCTCCTT GCTTAGTCCA CACTGGATGT TTTTCAAACA TGCCAGACCT AGGAAACAGA  18200
GAGTCTGGGT TACTTGCCCA AGGTCACACA GCCTTTAAGT CACAGAGCTG GGATTCAAAC CCAGACCACT GGGCTTCAGA GTCTGCTCTT TCTCATGACA  18300
CACAAAGTTT CATTTCTTCC TCTGTGCACC CCTACATGGA AAATATTATG TTTTACTGAC AAGGGCACCA AGGGCCTTAG AGGGGAGCGC TCCTGCCTGG  18400
GATGATGTGG TAAATAGGGG TGGGAGATGG ACTTGACCTG CAACCCCTGC GCTCATCCTC CCTCCCTCCC TGGGCTCCTG ATGGTGGGCT TCTTGTGACT  18500
GTGTTGCCCA CCAAGGCCGG AAGAGGACCA GACAGTGCCC CAGCACAGCA GCTGTGGCTG ACCAGGGAGT AGGGATCATC TAAGAACAGA GCGTGCATGG  18600
TGCTCACGCC TGTAATCCCA GCACTTTGGG AGGCCAAGGC GGGTGGATCA CCTGAGGTCA GGAGTTCAAG ACCAGCGTGG CCAACATGGG AAACCCCGTG  18700
TCTACTAAAC ATACAAAAAA TTAGCCAGGC ATGGTGGTGG GCATCTATAA TCCCAGCTAC TTGAGAGGCT GAGGCAGGAG AATCACTTGA ACCAGGGAGG  18800
TGAAGGTTGC AGTGAGTCGA GGTCGTGCCA TTGCACTCCA GCCTGGGCAA CAAGAGCAAG ACTCCGTCTC AAAAAAACAA AACAAAAGAA AAAACAGAGG  18900
GTGGCCCTAT GAGGAGCCTT CGCTTGTGTG GGTGGCCAGG GACAGCAAGA GGTGCCAGGG CCCTAGGAAC AGCTCTTTCC TGCTTCAACT TTGGGCTCCA  19000
GATGGGCGCT TTCCAGCTCA GTCTGAGCAG CTTCGGGAAG CTGTGTCCCA TGGGAGACAC TGGGAGTCCC CTGTGCTCTT TGTCTCCTGT CGGGCCCCCA  19100
CATTAGCTCT CTGGCCTCAG CTCTGGCTTC CCTCCAATTT GTTTCCCACG CAGCAGCCAG AGGAGCTTTC AAAAAGGTAA ATTATTTCAT GCTAGTCCCC  19200
TGCTTGAAAT CCTACAGTGC CTTCCCAGTG CTTTCAGCCA AAGCCCCAGT CCCTTCCTAA GCCCAGCCTG GCCCTGCCTC CCTGGTGCAT CATCTGCACA  19300
AATGCCTGCT CTCTGACCTC CAGCCACCCT GCACTTCCAA TGCCCGCGGC TTCCTGCCTG CAGCTTTAGT ACAGACCCCT CCCCTGCCCA GAACTGCCCC  19400
CACCCCAAGG CTTCTGCTGA AATGTCACCT CCTCAGAGAG GCCTTCCCTG GCTGCTCTGT CTAAACTCTG TGTTGAGAAG TTCCTTCTTG ATGGTTGTTG  19500
AGGAGGGAGG CTGGAGAAGA AGAATCAAAG AGGAGAAATA GAAAGCAAAA TAATTTGTTC TTGGGGACGG GCTGGTGCTG GGCACGGGGA GGCGCCCGTC  19600
TCTGGTGTGG GCAGCTGGGT AGATGGAGGA GCCGTATTTG GAAATGTGGA ACCCAGGAAG GGAGTGATCT AGAGGGAGGG GAAAGGTGGC GCGAGATGCC  19700
TGCCTCTCAA CAGGTAGCCA GACACATGGG TCTGTCTTGG TCACTGCTAT CTGCCCAGTG CCCAGCACAT CACAGGCCCT CAGTGGTGGT GTGTGGGCAT  19800
AGAGAATTAG AAGCTGTGGA CCTCTGGATC CGGAGCTGAA AACCACCAAA GGAGATGAGT TGGCCTGGCC AGGTGTGTAA AAGGCAGAGT CTGAGAGAGA  19900
ACGACCAGAG GGCAGAGCCC CGCAGGTGGA GTCCTGGGGG CTGGAGGGAG ACCATTAGGA GAATCGCACA TGGCTGGCGC AGCAGGTCCC AGGCAAATGT  20000
GGCCACTGGG TTTGGCAATA TGGGAGCCAG AGCCCTAGTG TCATCTCCCT GCCTTCTACC CAGCAGTTCC CAGAGTGATA TCCCCAACAG TGTTTGACAA  20100
CTGGTACAGG CTCTTCAGCG GCCACAGTTA CTGGGCAAGG CCTTGTGAGG GTGACTTTGG GGCAGCTGGC CAGCAGTGGG AGGGGAAGCA GTCTCAGGGG  20200
TACCTGAGGC ACTGAGCTCC GACCTCCAGG TGCCAATGCC GCACCAGGGC ACCGTTCCCC TGCAGGCTCT TACAGGGATT AGGGGCTGGT AAGGAGCAGT  20300
GATTAGGGGC TGACTAGCAG GCTGGTGGGC ACCAGCATGA CCCCTTGGTG GTACCCTCTG GGCACTCATG GGGACTTGGG CTAACAGATG GGGAAGGGAG  20400
CACATTCAGG GGGCTTAGGA AACATATTTA TGTAGGGAAG CATTTTAATA TTTTAGTAAC AGAAGCTATT AAAGGACTTA CAAACTTACT TACATACACT  20500
AAAACACTAT TTGGTCAAAC TTCTGTTTCT TTGGCACTTT CCTCCTTTAT TCTTTTTTAT TTTTTTGAGA CAGGGTCTTG CTCTGTCACC CAAGCTGGAG  20600
TGCAGTGGTG CAATCTTGGC CCGCAGTAGC CTTGACTTCC AGGCTCAGGT GGTCCTCCCA CCTTAGCCTC CCAAGTAGCT GGGACTACAG GTGCACGCCA  20700
CCACGCCTGG TGAATTTTTG TTTTGAAGGG GTTTCACTGT GTTGCCCAGG CTGGTTTCAA ACTCCTGGGC TTAAGTGATC CGCCAGCCTT GGCTTCCCAA  20800
AGTACTGTGA TTACAGGTAT GAGCCACTGC ACCCGGCCTC CTATTTTTCT GCTTCTGCTT TGTGGATAAT TGGATGCTTG GACCTCCTGA TTTAATCTTC  20900
TAATTTCCTT AACTGTTTAC TCCTATTTTT CATCATCTTG TCTTTTTGTT CTACTTTGTG GAGGATTTCT TCACTTTTAG CTTCCAGTTC TTTTCTTACA  21000
TCGTGACAGT TGCTGCCGCA TTCTCTTGTA AATTTCCGAG GGCTCGTTCT TGGGTTCTGA ATGTTCCCTC CTTTCAAGGA TCTTCTCATC TCTTTGAGGA  21100
```

```
TATTCATGTC TTTTTTGTTT TGGTTCTTAG GTTTTCATCT GTTCTCTGTG CTGTTTCCTC GGAGTGCTTT TGTCTATTCT GTTGTTTTGT CCCTCATGTT  21200
AGAAGCATTT CTTTTTTTTT TGTTTTTTTT TTTGTGATAC AGAGTCTTGC TCTGTCACCA GGCTGGAGTG CAGTAGCATG ATCTCGGCTC ACCACAGCCT  21300
CTGACTCCCT GGTTCAAGTG ATTCTCCTGC CTCAGCCTCC TGAGTAGCTG GGATTACAGG CACACACCAC CACACCCAAC TAATTTTTGT ATTTTTGGTA  21400
GAGACGGGGT TTCACCATGT TGGCCAGGAT AGTCTCAATC TCCTGACCTC ATGATCCTCC GACCTTGCCT GGGAGGCCAA AGTGCTGGGA TTACAGGCGT  21500
GAGCCACCAT GCCCAGCCTA GAAGCATTTC TTAATGTCTG GTGTTCTCTG GCTGTTGTAT CTTAAAAAAA AAAGGGGGGG GAAACTGAGG CTCGAGGTGA  21600
CCTTGTGAGC TGGAGCAGAG CCGGGATGGG ATGAGGAGGC AGGAGCGTGT GCAGAAGAGA GGGAGCCCCC CTGAGCTCGC ACCCTGCTTC CCGTGGCTGG  21700
GAGGGGAGGC CGAGATGCTT GGGGAGAAAT GGAGGCTCCA AGCCAGAGGG GCTGTTTCCA GCACGCTCTT ACTGAGCGCT GCTGTAGTCC AGCTTGGTGT  21800
GGCGGCTGTG GGCAGGGAGG GGAGAGAGGT CTGAGCTGGC TGGCGGCCCA CTGGGCCCCT CCCCTGAGCC TCCACCGGCC CTCTCCCAGT GCGCTGGGCT  21900
GGGCAAGCCT CTGATGTGCC AGCCAGATGG AGGGTGAAGT CCTGATGCCT GCCCCTACCC TGGGAATTGT GATGCTGCAG TTACTGCCCC TGATAACCCC  22000
TGACTGGGCA TAGGACCAGC TGGCTGAGCC AGCTCCTGGG GCTGAGGAGG AAGCCATGAA CTTGACCTGG CACTTTCCTT GTCTCCAAGC ATCAGTCAAC  22100
CAAGGATATG GAGGGGGTGT GTGCATGTGT GCACACATAC ACACACACAC ACACACACAC TTCAACCTGT TTATCCCCCT TGAGATTTGC TGACTTGTGC  22200
ATTGGGGGTA GAAGGTGCTG GAAAAATTCC GGTCCTGGTT CTCAGTTTCC CCATCTGTCC AGTGGGAGCA GCTGGACTGA GAGACGCCCA TGTCTCCTGC  22300
TGTGGTCCTG CAAGGAGGCT GGCGCTCCTG AGTCTGCTCC ATCCTGGCCT GTCAGGCCTG CCTGGATCCT GCCCCGGGTT GGTCCACCAC TCACTGTTTT  22400
GTTTCCAG
        gA rgArgGlyIl eAlaLeuPro AlaAsnThrA laGluGlyLe uLeuAsnVal IleGlyMetA spLysProLe uThrLeuPro AspPheLeuA
        GA GGAGAGGGAT CGCCCTCCCA GCTAACACAG CAGAGGGGCT GCTGAACGTC ATTGGCATGG ACAAGCCGCT CACCCTTCCA GACTTCCTGG  22500

laLysPheAs pTyrTyrMet ProAlaIleA l
CCAAGTTTGA CTACTACATG CCTGCTATCG C
                                  GTGAGTTGC CCCCAACCCA CAGGTCCTAG GGCAGCATTG ATCCCTATGA CTAGGACCAG GCCTGTCCCT  22600
CAGCCTGTGG GGGCCAGAGA AGTTGCTCTG AAACCACAGC TGTCTTTCTC ACCATTGTGT ACACTTAGTG AGTCTCTCCA GTGCCTTTAG GCCTCAGTTT  22700
TCCCTTCTGA GATGTGGGTG TGATGGACTG AAATTGCTTC AAGTTCTACA GAGAAATGGC AGAATATGGG AGCTAAGAAC ACAGGGTCAG AGGCAGTGCA  22800
GGGCTTGAAC CCGGGCCATC TATCTCCTAG TTCAGGGCTT CGTGTTGTGA GGGGAGGAGA GGCCTGAATA TAGGGTGGGG GCGGGGAGAT GTGGGGAAGA  22900
TTCTCCAAAA GGCTTTTTCT TTTTCTTGTC TTGAGTCGCC AGGGAACAGC ACTAGGTACC GAAAAGGCCA GAAGGGGTAT GGGCGAGTAC TAGAGAGAAA  23000
TTTCCATGAC TGCTTTATTT ATTTATTTAT TTATTTATTT ATTTATTTAT TGAGACAGAG TCTCACTCTG TTGCCCAGGC TGAAGTGCAG TGGTGCGATC  23100
TCAGCTCACT GCAACCTCCA CCTCCCAGTT TAAGGGATTC TCCTGCTTTA GCCTCCCAAG TAGCTGGGAT CACAGGCACC CACCATCACA CCCAACTAAT  23200
GGTTTTGTAT TTTTAGTAGA GATGGGGTTT TACTATGTTT GCCAGGCTGG TCTCGAATTC CTGACCTCAG GTGATCTGCC CGCCTCGGCC TCCCAAAATG  23300
CTGGGATTAC AGGCGTGAGC CACTGCGCCT GGCCTCCATC CTCATCCTGA AGATGCAAGA ACTTCTGGTG ACCCCTTCTC CTGAGAGTGG CCTGATCTCC  23400
CCTGGGCAGG GCACTTTCTT CCCACGCTGG GCTCTCCCAC GACTTGTGTG CCTTCCCTCA CACATTCTAG TAACCACTTC ATTTTCACTC TTCATGGTGG  23500
GAACTTCCAG CTAAGCACAG TCCACCGTTA CGTGATCAAC ACAGTGGCCC TGGCAGGCCA ATTTGTGCCT TGCTTCTGGA ACAAACATGC AGTAATAACA  23600
ACGAAAATGT TTTGAGCATT TGTCCGCTCT GCTCCAAGCA CTGACCCGGG TGGGGTTTAT GAAGTTTGAC TCATTTGTCC CCGCAATAAC TCCTTGACCT  23700
AGGTGTCAGA GGGTGACTAA CCAGGGGTCA CACAGCAGAT AAGTGTGGGC ACAAGGATCC AAGTCCATGA CTGTATCCCA CGTGTCTCCC ACATCCAGGC  23800
ATCCCTCTGG ACTTGTCCAG CTGTGTCCTT TTCTCTCATT TCTCTTCCCT GCCAGCCTTA ACTCCATCAC CAACAAATAT TGGGCTACTC TGTCCTAGGC  23900
ATGGTCCTCA GCTGAGAGGT CGCAGCCATC CCAAGACAGA GGGGTCCTTG CCACATGGAG ACTGCATTCT AGTAGGGAAT ACAGCAAACT GGCTGATAAG  24000
CCATATGACA CACAATGTTG AGTAGTGATA AGGACCTGGG AGAAAAAGAA AGCCCAGGAG AATGGTGGAG GGGCCGTTTT AAGATAAGGC GGTCTGGGCC  24100
AGGTACAGTG GCTCACGCCT GTATCCCCAG CACTTTGGGA GGCTGAGGTG GGCGGATCAT GAGGTCAGGA GATCGAGACC ATCCTGGCTA ACACAGCGAA  24200
ACGCTGTCTC TACTAAAAAT ACAAAAAATT AGCCGGGCGT GGTGGCATGC GCCTGTAATC CCAGCTACTT GGGAGGCTGA GGCAGAGAAT TCACTTGAAC  24300
CCAGGAGGCA GAGGCTGCAG TGAGCTGAGA TGGCGCCACT GCACTCCAGC CTGGGCGACA GAGCAAGATT CTGTCTCAAA AAAAAAAAAA AAAGATAAGG  24400
TGGTCAGGGA AGGCCTCTCT GAGGAGGTGA AGCTTCAGCT GGCTCTAAAC CAGGGGAGCG GGAGAGACGC AGTGTAGGAC AGTATCGGGG AAGAGCAGGC  24500
CTGTGTCTTC TCCGGTGGCC TCAGGGAATG AGGGAGAAGG AAGGTGCTGG GGAGGCTGGC AAGGCTGGAG GATGCAGGCT TGTGGGCAGG ACCTGGGAGT  24600
TGCGATGTCA CTCTCCGTGG CAGGAAGCTA CTGGGGCTTC GAGGGGAGAA GTGATATGCT TTGATTTACC TTCTTAAAAG ATTGCCCCAA CTGCTGGGTG  24700
GAGAACAGGA TGACAGGGGC AAGCATGGAG ACAGGGAGGC CAGTTAGAGA TGGCGTGATT CAGGCCAGGA TGAGGGGTGA GAACTGGTAT GCAGTTCCAA  24800
AGTAGAGCTG ATAGGACTTG CCCAGTGTCT GGATCTTATC CAGTGGATGC CCAGAGCTTG GGTCTGGGGA TGAAGTGGGT TTAATCTGCC AAGGGTTGGG  24900
GATGTCATTT GCTCCTGGAG CTCCCAAGGG ACTTGGGGAA GGTTGTTCCC AACCCCTTTC TTCCCTTCCC AG
                                                                              aGlyCysA rgGluAlaIl eLysArgIle
                                                                              GGGCTGCC GGGAGGCTAT CAAAAGGATC  25000

AlaTyrGluP neValGluMe tLysAlaLys GluGlyValV alTyrValGl uValArgTyr SerProHisL euLeuAlaAs nSerLysVal GluProIleP
GCCTATGAGT TTGTAGAGAT GAAGGCCAAA GAGGGCGTGG TGTATGTGGA GGTGCGGTAC AGTCCGCACC TGCTGGCCAA CTCCAAAGTG GAGCCAATCC  25100

roTrpAsnGl nAlaGl
CCTGGAACCA GGCTGA
               GTGA GTGATGGGCC TGGAAGGGGC CATGCTGAGG GTGTGGCTGG GAGGCTCAGC TCTGAGACTG GAAGGGCGAA CTGCTGGGAA  25200
TCCCTGACCC AAGCAAGACC TTGTTCTTGC CCCCAGTCTG GTCCATGGCC TCAGAAAGAT GGGTTTAACT CTGTCACAAG AGACGTGGTT CCCATCCTCC  25300
CTTTGCCGTT ATGTTCTTAC CTTGGGCACA AGTGTTTGGC TGTGTCTTGC TCTGGCCACA GGCCTGCTGT CCAGGAATGT TAACCTGCTT AGCCACCCAG  25400
GATTTCTGAG GGGTCTCCCT TGTCACTGAT GCTGATCAGA TCTCTAAAGG CCCTAAAGGT CCTGCTCTAA CTTCATAACT GAAGTGAGTC TGGCCCATTT  25500
CTAGCCCCCT GCCTGGGCCC CCATGGATCT CTAAGTGGTA TCACAAAACC ACCCTGCCCC ATTTTCTGAG CCATGATTCT GATACATATA GAATGTGAAC  25600
ATCATGGCAG GCCCAAGCTT AGCAATGCTG TCCATCTGGG GGTGGGGAGG GCCATGTTGA CACCCCACAC CTCCCACTAA GATCTAGGAG CACCCAGCTG  25700
CTTTAAGAGC TAGAGGGACA TGCTAGGGCC TGGGGGCATC TCTGCCAGTC TTTCCTCTGA GGCAGTGGGT CAGTGGGGGA GGAGGGTCCT CCCCAAAGCC  25800
TCCTCTTCCT CCTCTGTCCC AGTCCCAGAG CTGCCCTTTA GGCCTTCCTT TTGCCTCAGG CCCATCCCTA CTCCTCTCCT CACACAG
                                                                                        uGl yAspLeuThr
                                                                                        AGG GGACCTCACC  25900

ProAspGluV alValAlaLe uValGlyGln GlyLeuGlnG luGlyGluAr gAspPheGly ValLysAlaA rgSerIleLe uCysCysMet ArgHisGlnP
CCAGACGAGG TGGTGGCCCT AGTGGGCCAG GGCCTGCAGG AGGGGGAGCG AGACTTCGGG GTCAAGGCCC GGTCCATCCT GTGCTGCATG CGCCACCAGC  26000

roA
CCA
   GTGAGTA GGATCACCGC CCTGCCCAGG GCCGCCCGTC TCACCCTGGC CCTGACCTCC TGGCCTAGCA GTGGGGCTGT ACCTGATCTC CCCTGTGCCC  26100
CACAGCCCCA TGGTGTCCCC TTGAGCCCAC TGGCATGAAC TTCGGGGCTTC ATGAAACAAC TGGAGACCTC CTAGGCAGGC TCAGAACTTC TGGAGATGTT  26200
CTCCCCAGGG ACACCATGCC TTTATAGCCA CCCTGCAGGA AGCTCAACAC CAAATAGGAA CGTAACTATT GAAAAAAAAA TCTAGGCTAG ATTCTGATCA  26300
GCCCATAGTC CTCCCTCGAG ACCCAGTGGA CCAGGCCCCA TCCTGTCTGG GCCTGAATAG GTCTGATTTC CAAGATTTCT GAGGGGTCTC CCTTGTCACT  26400
GACGCAGATC AGATCTCTAG AGTTTGTGCC TCATGGTGCA CAGCCTCACT GTGTGATATT GGGCAGGTCA CACTGCTGCT CTGGTTATGC ACCAAGACAC  26500
CTCAGTTGTG CACTGTCACA AGGAGATGAT CACACTTACT TCATTCCTCT ACCCTCAGGA TTAGTAAGAA CCAAAGAGCT ACCTGCACGC ATTTCCTCTA  26600
ATCCTCGCAG CAGCCTGCAA AGCAGAACTA CCATTGCTTA GTCCCATTTG ACAGATGAGG AAACTGAGGT GGAGTGAGGT GCAGCCTCTT GCAAGGCACA  26700
AACCCTGGAT TTGTATCCGG GGACATCTAG TTCCAAAGCC TGTGTTCATT CATTCTTTCT TAAACACTTC AGAATAACTT TATTGGTTAA GAGTACCTAA  26800
TACATTAGCG AGATACTTCC CAATACTAGT GTGAGTTCTA TTTTAGATGA CGTGTTAAAC GGTCCTCCGT TTCCTCATCT GCGCATGGGA ATAAGCCTAC  26900
CATGAGTGTT GTTGGAAACA CCAGGTGAGA GAAGGGTCCG TGTCATTTAC TGAGCTCAGG CCCCGTCCTT GGTGCTTTAC ACACATGGCC TCGGCAAAGC  27000
CTGGCCGTGA CCCTGTGCAA TAGCTGGCAG GGTTCTTTCT GAAAAGGGCG GAAACTGAGG CCATAAGCAG AGCAGTTTTC CGCAGCCATG TGGTTAGGAC  27100
ATAGCAGTTA GGATTTGAAG ACACTGAGCC CTGTTTTGTG CTGGCCTCCC ATGGGGGGTT TGGGTGGGAC AGCAGGCAGG TAGGCTGGGA GGTCTCTCCA  27200
TGGTGCTGGT GACAGAGCCT GGGTGGGCAT CTGCCCACAG
                                        snTrpSerPr oLysValVal GluLeuCysL ysLysTyrGl nGlnGlnThr ValValAlaI
                                        ACTGGTCCCC CAAGGTGGTG GAGCTGTGTA AGAAGTACCA GCAGCAGACC GTGGTAGCCA  27300

leAspLeuAl aGlyAspGlu ThrIleProG lySerSerLe uLeuProGly HisValGlnA laTyrGln
TTGACCTGGC TGGAGATGAG ACCATCCCAG GAAGCAGCCT CTTGCCTGGA CATGTCCAGG CCTACCAG
                                                                    GT GGGTCCTGTG AGAAGGAATG GAGAGGCTGG  27400
CCCTGGGTGA GCTTGTCTCC CACCCATAGT TGGGAGAAAT CACAAGAACC AGGGACCATG GTGTCTCCTG AGTTCTGAAG TGTGTCTTTG TTGGGTCTTA  27500
```

```
AGGCTTGGAA CTGGAATCCC CCTGGGCCAG GCGTGGTGGT TCATGCCTGT GATCCCAGCA CTTTGGGAGG CGAGGCAGGA GGATTGCTTG AGCCTAGGAG  27600
TTTGAGACCA GCCAGGGCAA CATAGTGAGA TCCATCTCTG CAAATACAAA AAAAAGTAGT CAGGCATGGT GGTGCATGCC TGTAGTCCCA GCTACTTGGG  27700
AGGCTGAGGT GGGAGAATTG CTTGAGTCCA GGAAGTCAAA GCTGCAGTGA GCTGTGATAA TGCGACTGCA CTCCAGCCTG GGTGACAGAG GGAGACCCTG  27800
TCTCAAAAAA AAAAAAAAGG AAGAAAGAAG AAAGAGAAAA GAAAGAGAAA GAAAGAGAGG AAGGAAGGAA AAAGAGGAAG GGAGGGAGGG AGGAAGGAAG  27900
GAAAGAAGGA AGGAAGGGAG AGAGAAAGAA AAGCCTCCAC TTGGTGTTGG GAGTCCTGTG CTGAGCCTGC TTCTGGCTGT GATTTGCTGT GTGAACCTGG  28000
GCAACACTGT GTCTTCTCTG GGCCTCTGTT TCTTCTATTG GGATGACTGA GTTGGAGCCG ACATCTCAAA AGTCGCTTCC AGCGTGATGA TGAATGGGCC  28100
TCCTGTGGAG GGTGCAGCAT GGTGGAGAAG TCAGGGCTCT GGAGTCCCAC TGCCCGGGCT CAGAGCTTGG TTCCACACTT CCTGTCTGAC CTTGGTCACA  28200
TTACTTGAAT CTCCTGAGCT TCAGTCCTTC ATCATAAAAT GGGTGGGATA ATAGTTGTGA ATATTAGATA ATGTATACAA GTCACTTCAT ATACTACCTG  28300
ACACATGGTA ACTGGCTAAT GAGTGACAGC TACCACTTAG ATAAGGACTT GGAGGGTAAA AGACCAGGTT TCCCCATGCT GTTGAAGCAG GCAGCATGAC  28400
TAGGATGGTT CAATCTCCAC AGCATGGTCA AGGCAGGCTG CCGGGGCCCT CCGGCTAGGG CAGCCATGAC CTGGCTCTCC CCCTTCCAG
                                                                                                    G luAlaValLy
                                                                                                    G AGGCTGTGAA  28500

sSerGlyIle HisArgThrV alHisAlaGl yGluValGly SerAlaGluV alValLysGl u
GAGCGGCATT CACCGTACTG TCCACGCCGG GGAGGTGGGC TCGGCCGAAG TAGTAAAAGA G
                                                                GTGAGGGCC TGGGCTGGCC ATGGGGTCCC TCCTCACTGC  28600
CTCCTCCCAT ACTTGGCTCT ATTCTGCTTC TCTACAG
                                    Ala ValAspIleL euLysThrGl uArgLeuGly HisGlyTyrH isThrLeuGl uAspGlnAla
                                    GCT GTGGACATAC TCAAGACAGA GCGGCTGGGA CACGGCTACC ACACCCTGGA AGACCAGGCC  28700

LeuTyrAsnA rgLeuArgGl nGluAsnMet HisPheGlu
CTTTATAACA GGCTGCGGCA GGAAAACATG CACTTCGAG
                                         G TAAGCGGGCC AGGGAGTGGG GAGGAACCAT CCCCGGCTGT CCCAACTTCC TGTATAGAGA  28800
GGCAGAAAGC AGGGCGGGTC CCAGGAACTC GAGGGGTGGC CCCAGGCCCA GACATGGGGG GAGGAATCAG CATGGCCTGG GGCCATCCCT GCCAGCCACA  28900
CACCTGCTCT TCCAG
               IleCy sProTrpSer SerTyrLeuT hrGlyAlaTr pLysProAsp ThrGluHisA laValIleAr
               ATCTG CCCCTGGTCC AGCTACCTCA CTGGTGCCTG GAAGCCGGAC ACGGAGCATG CAGTCATTCG
                                                                                GTGAGCTCTG TTCCCCTGGG  29000
CCTGTTCAAT TTTGTTCCAG GAAGGCCAAA GAGGGAAGAA ACTTTAGGGA TTGGGCATCA GCCCATGCCG CGTCTTTTAG ATATGAAATC TCTTCGACAC  29100
CCTGGGAAGC AGGCATTGCC GTCCTCATCT TACAAATGAG GAATCCGAGG CCCAGATGTG CTGTGGCTTG ACTGGGATTA CCCAGCTGCT AACCAGCAGA  29200
GCTGGGGCCC TACAGCTCAT CAGCTGGAGC AGAACGCTCC ATTACTCTGA GGGAAGCTTC CACACTTCCA ATTCTCCCAA CTCTGCCCCC TGGGCATCGC  29300
ATAGGAAGCA GGAGTCCCTC TGGCCAGCAT GTTCTCTCTT CCTGACACCT GGCCCTTGGG ACCCCTGGGC ATTCCCCTGA GCGCCATCTT GAAGCTTTCC  29400
ACCGGAGGTC TGTTCCACCC TGCCTGGCTC CCATCCTGGA GTCTAACCAG GGTCAAGGCC CTCCTTCCGT CCTGTCGCCA AGCCACAGGA GCAGTATCAG  29500
GCCTTAGGAA AAAGCCGCCT TCCCCAAGAC AAGGACAGCA AGAACTCAGG GTGACCATGG TCAGGCCAGC ACTTATCCAT CTGCCAGGCA TATGAGAAGG  29600
GGAGGGGCTT CGGCTCTGAT GTTCTGATGA CAAGGGGGTC TTGGGGCTTG CTTAGGGACA CGTGGCACCT GTGGAGGTTC TTGGAGGCAT GTGGGTATAC  29700
CATGGGCTGG AAAAAGATCC AGGAGTCATC TGCACAGATA TGGTGGCTGA AGGAGAAGCA GTGGCCCCAG GAGGTGGTGG AGCAAGAAGG GCCTAGGATA  29800
GAACCCAGAA GGACAATGGT ATTTAAGGGA CCAGCAAAAG AGACAAGTAG GAGGAAAGTC AAAAGTGTGG TGTCACAGAA ATCCAGGGAA AAGGTTTCAA  29900
GAAACAGTCA ACAGTGTGAA ATTCTGCTAT GCAAGTCGAT TATGGTCAGA GCTAGGAAAG ATCCATTAGA TACAACAAGA TGGTGGTCAG GGATCGTGCC  30000
AAGAACAGCT TCCATGGTAT GTTGGAGTAG CCAGCTCCCA GTGGGACTGA GGAACAAGCA GGGTAGGGTG CAGAGGGGAA GGCTGGAGAG GGTGGCAGCC  30100
GGAGGGGGAT GTTGCTTTCT TGGCTCCCAC CCCCACGCCC CCACCGGCTG CCATTCTGCC TGGTTCCCAT GTCTGGCCCC TCTGCTGCCT TTGCCCAGCT  30200
CTGGTCTTCA GGATGGGCTG GATTCTGGAC TTTCTGGTTA CATAGACTTG AACAAGTCAC CTAAGTTCTG AATTTATTTC CCCCTCTGCA CAAGGATCAG  30300
ATCTTTCAGA TCTGTTTGAG GCTGCTGTGA GGATCAAAGG CGGGTGAACG TCAATGTGTT CTGACTATTT ATGTAAGAGT AAAAGGAGGC TGATTCTCTC  30400
CTCCTCCCTC TTCTGCAG
                  gL euLysAsnAs pGlnAlaAsn TyrSerLeuA snThrAspAs pProLeuIle PheLysSerT hrLeuAspTh rAspTyrGln
                  GC TCAAAAATGA CCAGGCTAAC TACTCGCTCA ACACAGATGA CCCGCTCATC TTCAAGTCCA CCCTGGACAC TGATTACCAG  30500

MetThrLysA rgAspMetGl yPheThrGlu GluGluPheL ysArgLeu
ATGACCAAAC GGGACATGGG CTTTACTGAA GAGGAGTTTA AAAGGCTG
                                            GT GAGTGGGTGT GAGCCATACT GGCCTTGACT CGGGTTTGGG AGTATGGTAT  30600
CTACAGGTCC AGTCCGGGGC CTGGAATCTT TGGAGAGAGG GAGTGAGTCT GCCTCAACAG TCCAAGACAA GCCCAACCTA GACACTTTCC ACAGAGAAGA  30700
CATCTTTGTG TTGACGTCCT GACCTAGGAC CAGGTTTTTG ATCCTTTGCT TGGGTTGAGT GCCTTTAAAG AATCCAGTGA AAGCTGTCAA CCCTCTCCCC  30800
AGAAAGGTGT GTGCAGCAGC TATGAAGTCT TGCACACTCT CTTCAGGTTG TTCTTAAATC CCAGGCTGAA TAAGTCCATT CCTGCACGTG TCTGCGAGGT  30900
GTCTCTGGCC CCCTACATGC CACCCTGTCT CTCAAAGGTT TCTCCAACTT CCTTCTCACA GCCCTTTTTC ATGTAATGAC AAATTAAGAA CACGACCTCA  31000
TGGTCTCTAC TCTGGCACTT GCTGCCGTGT GACAGTGGAC AAATCCTTCC CCCTCTAAGC GTATCTGCCC ATGTTGAGTG AAGAGGATGG ACTATCACTA  31100
CATTGCTAAG AGCTGCCTTC TTTGTTCTCT GGTTCCATGT TGTCTGCCAT TCTGGCCTTT CCAG
                                                                  AsnIle AsnAlaAlaL ysSerSerPh eLeuProGlu
                                                                  AACATC AATGCGGCCA AATCTAGTTT CCTCCCAGAA  31200

AspGluLysA rgGluLeuLe uAspLeuLeu TyrLysAlaT yrGlyMetPr oProSerAla SerAlaG
GATGAAAAGA GGGAGCTTCT CGACCTGCTC TATAAAGCCT ATGGGATGCC ACCTTCAGCC TCTGCAG
                                                                     GTA GGTTCCTGTC TGGGCTTCTG GGCAGTTGCC  31300
TGTCCTGGCC CCAGTGTGGC TTTCTGTGGG ACTTCTAGCA AGATGCCCTT CCATTCTTGG GCAGCGCATG AATGTGTGAT GACTCCCTGG TTTCTGGGCC  31400
CTGGCTGGGA GCAGCGTCTC ATTAGATCGG TTTGTTTTCT ATAAAAGTTC TTGAGAGGCT GTTCTAAGGG GAGACTTTCT GAAGCCCAGT CCCAAAGGTC  31500
TGGGCAGTTG GGGACACCTC CATGGCTGCC CAAAGCCAAG GGCAGGGAGA GGGGCCCAGG CTGTTCTGCT CCTTTCTTCC TATGTGGTCT TGGCAAGGCA  31600
TCTTCTTGCC ATCATAGGAA GGAGTTCCTT TCTGGTTCTG GTGTTCTATG ATTTTTACAA CATCCTGGGT ACTACAAGTT GCCTGATCTT TTTGCTTCTC  31700
TGAACCAACG AGCAG
               lyGln AsnLeuEnd
               GGCAG AACCTCTGAA GACGCCACTC CTCCAAGCCT TCACCCTGTG GAGTCACCCC AACTCTGTGG GGCTGAGCAA CATTTTTACA  31800

TTTATTCCTT CCAAGAAGAC CATGATCTCA ATAGTCAGTT ACTGATGCTC CTGAACCCTA TGTGTCCATT TCTGCACACA CGTATACCTC GGCATGGCCG  31900
CGTCACTTCT CTGATTATGT GCCCTGGCCA GGGACCAGCG CCCTTGCACA TGGGCATGGT TGAATCTGAA ACCCTCCTTC TGTGGCAACT TGTACTGAAA  32000
                                                •
ATCTGGTGCT CAATAAAGAA GCCCATGGCT GGTGGCATGC
                                            AGCAGGTGGC ATGTAATTTG GTGGTCTTGG GCGGGCCGAT GTGGGCAGGA TGAGCATGGA  32100
GGGAGCTGGG TCAGCCTGCT CAGCAGCAGG GCCTGAGCCT AAGGGTGGCT GTGAATGCCA GGCCAGAGAT CCCAATGCTG TGGGCCAAGA GGGGTCCAGA  32200
GGCTGTCCTC CTTCCAGAAG AAATAAGGCT TCTCTGGTTG TTGCTCAAAC ATTCCCTGAA CTCTCAGCCC CTCCTAACTC TAGGTTTTAA GGAGTAAAGC  32300
TTCCTTTTGG GTTCCTGAAG CTGGCAGTTG GGGTGAGAGC AGATGAGATG GAAGAGGGCT CATCAGACAC TGGCCTTGGA GGGTGCTGGC CTCTGCAGAA  32400
CGCCAGCATC TTCTCAGAAT CGTATGTTCT AGAAGCCTGG GCGAAGTCCG GCTAATTGTG GACTTGGGGA AAATAAGGCC CAACCCCTGT TTTTGCAAGG  32500
TTAAGGAGAA ATAATCTTAA ACCAGTCACA CAAATCATCG GCATTTATTT CCTGGGTCCT AGGTGTCACT TATCCTGGTG GACAGGGCAG AGGTGGTCAG  32600
ATCGTTTTGA GCCAAAATCC CTTCCCTAAA AATGGATCTG TGGAGCTCCA TGAGGGAACC TCAGAGATGC ACAATGACAG TTTAGCTAAA ATGGCTTAAA  32700
AAATGTGAAT TGATTGTCAG CTCTCTCCAT ATCTGCTGAA AAAAGGTTTA AAATTTTTAA AAAGTTTAAA AGTGTTTTCT AAAAAAGGGA CAAGCAGGTC  32800
TGGACC
```

FIGURE 3: Complete nucleotide sequence of the human adenosine deaminase gene. The putative transcription initiation site (nucleotide 1) and the poly(A) addition site (nucleotide 32 040) are identified by solid dots. The 3935 nucleotides of 5′-flanking sequence are numbered in a negative fashion. The sequence is broken and continued one or two lines below to indicate the exon–intron splice junctions. The deduced amino acid sequence is shown by the three-letter code above the coding sections of the exons. The polyadenylation signal is indicated by dashed underlining. Repetitive sequences of the Alu or O family type are all underlined and are distinguished in the text.

Consensus	$\begin{matrix}A\\C\end{matrix}AG\downarrow GT\begin{matrix}A\\G\end{matrix}AGT$	$\begin{matrix}CCCCCC\\TTTTTT\end{matrix}XCAG\downarrow G\begin{matrix}G\\T\end{matrix}$
	intron	
Intron #		
1	AAa↓GTGAGc	CCCCTTGCAG↓GT
2	CAG GTAAGT	TTgTTTCCAG Ga
3	Cgc GTGAGT	CCCTTCCCAG GG
4	tga GTGAGT	CCTCaCACAG aG
5	Cca GTGAGT	CTgCCCACAG ac
6	CAG GTGgGT	CCCCTTCCAG Ga
7	gAG GTGAGg	TTCTCTACAG Gc
8	gAG GTAAGc	gCTCTTCCAG aT
9	tcG GTGAGc	TCTTCTGCAG GC
10	CtG GTGAGT	gCCTTTCCAG aa
11	CAG↑GTAgGT	CaaCgaGCAG↑GG

FIGURE 4: Exon–intron splice-junction sequences in the human adenosine deaminase gene. These sequences are compared to the consensus sequence at the top. Matches are indicated by capital letters; mismatches, by lower case letters. Arrows indicate the locations of the splice sites. The invariant GT and AG residues at the respective 5′ and 3′ ends of the introns are underlined.

only superficially similar to the consensus sequence of TATARAR (R = A or T) (Breathnach & Chambon, 1981). Functional assays will be required to evaluate the significance of this sequence in transcriptional initiation, if any.

A number of genes have been isolated recently that have similar promoters that are G/C rich and lack the classical TATA and CAAT box sequences, including mouse and human hypoxanthine phosphoribosyltransferase genes (Melton et al., 1984; Patel et al., 1986; Kim et al., 1986), mouse and human dihydrofolate reductase genes (Crouse et al., 1982; Chen et al., 1984), and hamster 3-hydroxy-3-methylglutaryl-coenzyme A reductase (Reynolds et al., 1984). When the human adenosine deaminase promoter region was compared to the promoter regions of these genes, a large number of significant homologies were noted. Examples of these are shown boxed in Figure 6. Sequence I is 80% homologous to a similar area in the human dihydrofolate reductase promoter (Chen et al., 1984). Sequences II and III are 81% and 89% homologous, respectively, to sequences in an area of the human hypoxanthine phosphoribosyltransferase promoter that have recently been described as critical to expression of that gene (Patel et al., 1986). The equivalent area has also been shown to be important for expression of the mouse hypoxanthine phosphoribosyltransferase gene (Melton et al., 1986). Contained within the areas of sequence homology shown in Figure 6 are six underlined decanucleotide sequences that are highly homologous (five, 90%; one, 80%) to the consensus decanucleotide sequence for transcription factor Sp1 binding sites (Dynan et al., 1986). Similar sites in the mouse dihydrofolate reductase promoter were shown to bind Sp1, and Sp1-containing preparations were shown to stimulate transcription from this promoter in vitro (Dynan et al., 1986). Contained within five of the indicated decanucleotide sequences of the adenosine deaminase promoter is the GGGCGGG motif that was noted previously (Valerio et al., 1985) and is homologous to the GC motifs of the SV40 promoter region which contribute to the activity of this promoter in vivo and in vitro (Barrera-Saldano et al., 1985).

Since the sequence of the entire human adenosine deaminase gene and large flanking areas is now available, as well as the large number of M13 subclones spanning the entire region, contribution to regulation of gene expression not only of the elements in the immediate 5′-flanking promoter region but also of any elements in the far upstream region, in the downstream 3′ flank, or within the exon–intron superstructure itself can be assessed. Hereditary adenosine deaminase deficiency with severe combined immunodeficiency has been proposed as one of the most promising candidates for initial attempts at human gene therapy (Anderson, 1984). Knowledge of the sequence, structure, and regulation of the normal human adenosine deaminase gene may be critical to the success of such an endeavor.

Comparison to cDNA Sequences. The exonic sequences of the gene were compared to the previously published ADA211 cDNA sequence (Adrian et al., 1984). Three sequence differences were noted. The G at position 25915 in exon 5 of the gene instead of A (at 390) in ADA211 and the A at position 27296 in exon 6 of the gene instead of G (at 534) in ADA211 represent wobble in the third base of a Val codon. These first two differences probably represent allelic variation since both of the alternative nucleotides at these locations have been observed previously in normal human adenosine deaminase cDNAs (Wiginton et al., 1984; Daddona et al., 1984; Valerio et al., 1984; Bonthron et al., 1985). The third difference is a C at position 84 in the first exon of the gene which is missing in the published ADA211 sequence. When this cDNA was sequenced again, this difference was revealed as an artifactual error in the original sequencing gels, and a C should be present at this location in the ADA211 cDNA sequence as well. Sequence differences among adenosine deaminase cDNAs have been tabulated previously (Bonthron et al., 1985).

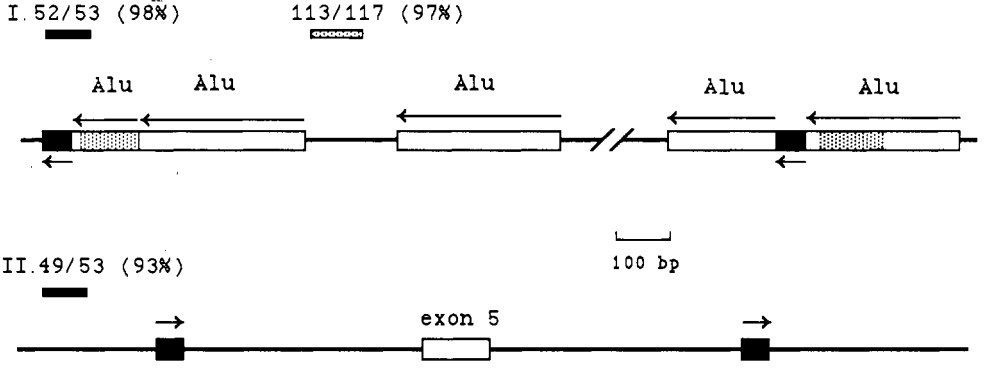

FIGURE 5: Direct-repeat sequences within the human adenosine deaminase gene. (I) Arrows indicate the location, extent, and orientation of Alu repeat sequences (arrows above the line) and of non-Alu repeat sequences (arrows below the line) in a 3100-bp region of the first intron (nucleotides 3550–6650 in Figure 3). The solid rectangles are a highly homologous (98%) non-Alu 53-nucleotide direct repeat. The open rectangles are Alu repeat sequences. The stippled areas within two Alu repeat sequences represent a 117-bp segment that is extremely homologous (97%). (II) The arrows and solid rectangles indicate a highly homologous (93%) direct repeat of 53 bases flanking exon 5.

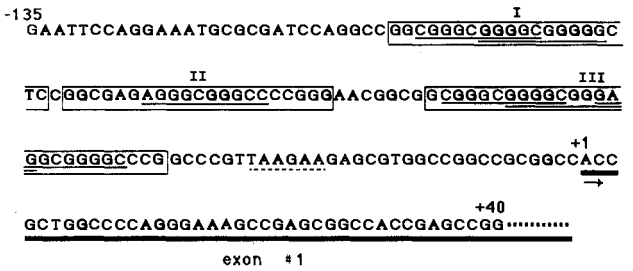

FIGURE 6: Sequence of the promoter region of the human adenosine deaminase gene. The first 40 nucleotides of the first exon (heavy underlining) are shown, beginning with the putative transcription start site (arrow) at nucleotide +1. Also shown are the first 135 nucleotides of the 5'-flanking region (negatively numbered). Boxed G/C-rich sequences I, II, and III in this region are highly homologous to sequences in similarly G/C-rich promoters of other human genes (see text). Within the boxes, six decanucleotide sequences are underlined that are highly homologous to the consensus Sp1 transcription factor binding sequence. Also indicated by dashed underlining is the only sequence resembling a classical TATA box.

The availability of the entire gene sequence should be beneficial in determining the basis of mutations resulting in adenosine deaminase deficiency, not only those involving point mutations and amino acid substitutions, as has been done in two cases (Bonthron et al., 1985; Valerio et al., 1986), but also those involving larger segments of DNA (Adrian et al., 1984), possibly including intronic sequences.

ACKNOWLEDGMENTS

We thank Richard Friedman for preliminary screening of the genomic library.

REFERENCES

Adrian, G. S., Wiginton, D. A., & Hutton, J. J. (1984) *Mol. Cell. Biol. 4*, 1712–1717.

Anderson, W. F. (1984) *Science (Washington, D.C.) 226*, 401–409.

Barrera-Saldana, H., Takahashi, K., Vigneron, M., Wildeman, A., Davidson, I., & Chambon, P. (1985) *EMBO J. 4*, 3839–3849.

Benton, W. D., & Davis, R. W. (1977) *Science (Washington, D.C.) 196*, 180–182.

Blake, C. (1983) *Nature (London) 306*, 535–537.

Bonthron, D. T., Markham, A. F., Ginsburg, D., & Orkin, S. H. (1985) *J. Clin. Invest. 76*, 894–897.

Breathnach, R., & Chambon, P. (1981) *Annu. Rev. Biochem. 50*, 349–383.

Cech, T. R. (1983) *Cell (Cambridge, Mass.) 34*, 713–716.

Chen, M. J., Shimada, T., Moulton, A. D., Cline, A., Humphries, R. K., Maizel, J., & Nienhuis, A. W. (1984) *J. Biol. Chem. 259*, 3933–3943.

Crouse, G. F., Simonsen, C. C., McEwan, R. N., & Schimke, R. T. (1982) *J. Biol. Chem. 257*, 7887–7897.

Daddona, P. E., & Kelley, W. N. (1977) *J. Biol. Chem. 252*, 110–115.

Daddona, P. E., Shewach, D. S., Kelley, W. N., Argos, P., Markham, A. F., & Orkin, S. H. (1984) *J. Biol. Chem. 259*, 12101–12106.

Duncan, C. (1985) *NEN Prod. News 4*, 6–7.

Dynan, W. S., Sazer, S., Tjian, R., & Schimke, R. T. (1986) *Nature (London) 319*, 246–248.

Frischauf, A., Lehrach, H., Poustka, A., & Murray, N. (1983) *J. Mol. Biol. 170*, 827–842.

Giblett, E. R., Anderson, J. E., Cohen, F., Pollara, B., & Meuwissen, H. J. (1972) *Lancet ii*, 1067–1069.

Jelinek, W. R., & Schmid, C. W. (1982) *Annu. Rev. Biochem. 51*, 813–844.

Kim, S. H., Moores, J. C., David, D., Respess, J. G., Jolly, D. J., & Friedmann, T. (1986) *Nucleic Acids Res. 14*, 3103–3118.

Lloyd, J., Lamb, A., & Potter, S. (1986) *Mol. Biol. Evol.* (in press).

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Martin, D. W., Jr., & Gelfand, E. W. (1981) *Annu. Rev. Biochem. 50*, 845–877.

Melton, D. W., Konecki, D. S., Brennand, J., & Caskey, C. T. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 2147–2151.

Melton, D. W., McEwan, C., McKie, A. B., & Reid, A. M. (1986) *Cell (Cambridge, Mass.) 44*, 319–328.

Naora, H., & Deacon, N. J. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6196–6200.

Nussinov, R. (1981) *J. Mol. Biol. 149*, 125–131.

Patel, P. I., Framson, P. E., Caskey, C. T., & Chinault, A. C. (1986) *Mol. Cell. Biol. 6*, 393–403.

Paulson, K. E., Deka, N., Schmid, C. W., Misra, R., Schildler, C. W., Rush, M. G., Kadyk, L., & Leinwand, L. (1985) *Nature (London) 316*, 359–361.

Queen, C., & Korn, L. J. (1984) *Nucleic Acids Res. 12*, 581–599.

Reynolds, G. A., Basu, S. K., Osborne, T. F., Chin, D. J., Gill, G., Brown, M. S., Goldstein, J. L., & Luskey, K. L. (1984) *Cell (Cambridge, Mass.) 38*, 275–285.

Schrader, W. P., Stacy, A. R., & Pollara, B. (1976) *J. Biol. Chem. 251*, 4026–4032.

Sharp, P. A. (1981) *Cell (Cambridge, Mass.) 23*, 643–646.

Sun, L., Paulson, K. E., Schmid, C. W., Kadyk, L., & Leinwand, L. (1984) *Nucleic Acids Res. 12*, 2669–2690.

Thompson, L. F., & Seegmiller, J. E. (1980) *Adv. Enzymol. Relat. Areas Mol. Biol. 51*, 167–210.

Valerio, D., McIvor, R. S., Williams, S. R., Duyvesteyn, M. G. C., Van Ormondt, H., Van der Eb, A. J., & Martin, D. W., Jr. (1984) *Gene 31*, 147–153.

Valerio, D., Duyvesteyn, M. G. C., Dekker, B. M. M., Weeda, G., Berkvens, Th. M., Van der Voorn, L., Van Ormondt, H., & Van der Eb, A. J. (1985) *EMBO J. 4*, 437–443.

Valerio, D., Dekker, B. M. M., Duyvesteyn, M. G. C., Van der Voorn, L., Berkvens, Th. M., Van Ormondt, H., & Van der Eb, A. J. (1986) *EMBO J. 5*, 113–119.

Wiginton, D. A., Coleman, M. S., & Hutton, J. J. (1981) *Biochem. J. 195*, 389–397.

Wiginton, D. A., Adrian, G. S., Friedman, R. L., Suttle, D. P., & Hutton, J. J. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 7481–7485.

Wiginton, D. A., Adrian, G. S., & Hutton, J. J. (1984) *Nucleic Acids Res. 12*, 2439–2446.