Sequence of human adenosine deaminase cDNA including the coding region and a small intron

Dan A.Wiginton, Gwendolyn S.Adrian and John J.Hutton*

Department of Medicine, University of Texas Health Science Center and Audie L.Murphy Memorial Veterans Hospital, San Antonio, TX 78284, USA

Received 27 January 1984; Accepted 6 February 1984

ABSTRACT

The nucleotide sequence for an unusual, cloned human adenosine deaminase cDNA has been determined. Contained within a sequence of 1535 nucleotides is a coding sequence of 1089 nucleotides that encodes a protein of 40,762 daltons. The coding sequence is interrupted by a non-coding region containing 76 nucleotides. Both the 3' and 5' ends of this region have consensus sequences generally associated with splice sites. The 3' untranslated sequence contained 308 nucleotides, including a polyadenylation signal sequence 20 nucleotides from the end. The cloned cDNA appears to correspond to a nuclear mRNA precursor which contains a small intron.

INTRODUCTION

Adenosine deaminase (adenosine aminohydrolase, EC 3.5.4.4) catalyzes the irreversible deamination of adenosine and deoxyadenosine to inosine and deoxyinosine, respectively. The catalytic activity of the enzyme resides in a single polypeptide encoded by a locus on the long arm of human chromosome 20 (1). Molecular weights of 38,000 to 44,000 daltons have been reported for the polypeptide (2-4). Deficiency of adenosine deaminase activity in humans causes an autosomal recessive form of severe combined immunodeficiency disease (5-7). Numerous mutations have occurred in the ADA gene which cause human adenosine deaminase deficiency (8-12) and can be studied by molecular techniques. We have recently described the isolation of human adenosine deaminase cDNAs (13). Herein we present the nucleotide sequence of one of these cDNAs (ADA 3-3). The sequence appears to contain both the entire coding region for the ADA protein and a small intron.

METHODS

The isolation of cDNA clone ADA 3-3 from a recombinant cDNA library constructed from bacteriophage λ gt10 and poly(A)⁺ RNA from leukemic T lymphoblast cell line HPB-ALL has been described previously (13).

DNA Sequence Analysis

DNA sequencing was performed by primed synthesis with dideoxynucleotide chain termination (14) after cloning of DNA restriction fragments into phages M13

mp8 and M13 mp9 (15). Restriction endonucleases, T4 DNA ligase, large fragment of *E. coli* DNA polymerase, host *E. coli* K12JM101, RFI DNA from vectors M13 mp8 and M13 mp9, fifteen-base and seventeen-base primers, and other materials for cloning and sequencing were obtained from New England Biolabs, Beverly, MA. Selection of restriction fragments for cloning was based on the previously published restriction map of ADA 3-3 (13). Fragments were isolated after restriction endonuclease digestion by electrophoresis on low gelling temperature (LGT) agarose (16).

RESULTS AND DISCUSSION

Sequencing strategy

The strategy for sequencing ADA 3-3 is summarized in Figure 1. The entire cDNA molecule, which was originally prepared with EcoRI linker ends (13), was cloned into the EcoRI site of M13 mp8. Smaller restriction fragments were cloned into the various restriction sites of M13 mp8 and M13 mp9.

Nucleic Acid Sequence

The full sequence of the 1535 nucleotides in ADA 3-3 is shown in Figure 2. Fifty-nine bases were determined 5' to the first ATG. Beginning with this first ATG codon there is a coding sequence of 1089 bases. This reading frame is interrupted between bases 678 and 679 by a sequence of 76 bases that is shown but not numbered in Figure 2. The significance of this 76-base sequence will be discussed later.

The reading frame of 1089 bases encodes 363 amino acids which represent a protein with a molecular weight of 40,762 daltons. The predicted and actual amino acid compositions of the ADA protein are shown in Table 1. This molecular weight agrees well with the estimated molecular weight of the protein (42,000 daltons) as determined by SDS polyacrylamide gel electrophoresis (2-4). The amino acid composition predicted by the nucleotide sequence is very similar to the composition of purified human adenosine deaminase (3).

The 3' untranslated sequence contains 308 bases after the TGA terminator codon. The eukaryotic polyadenylation signal sequence AATAAA, which is typically located 11-30 nucleotides 5' to the poly(A) addition site (17,18), is 20 bases from the end of this sequence.

The assignment of the first ATG codon as the site of protein initiation is based on several lines of evidence besides the fact that the coding sequence predicts a protein of the proper molecular weight and amino acid composition. In greater than 90% of eucaryotic mRNAs the first AUG encountered from the 5' end of the molecule has been the functional initiator codon (19). Therefore, position is important but apparently not the sole criterion for a functional initiator. The sequence flanking the AUG codon is also important (19,20). Among eucaryotic mRNAs the most common

Figure 1. Strategy for sequencing human adenosine deaminase cDNA ADA 3-3. The top line is a restriction map for the entire cDNA sequence including sites that were used for cloning or as reference points. The map numbering proceeds in a 5' to 3' direction. The arrows below the map indicate the extent and direction of sequences determined. Each segment was sequenced 2-4 times.

initiation sequence is AXXAUGG, where X is any base (19). This corresponds to the sequence around the first ATG in the sense strand of our cDNA. In addition AXXAUGG has been found upstream from the functional initiator codon only once in 180 different eucaryotic mRNAs (19). Therefore on the basis of previous examples, the first ATG codon in the cDNA would be preferred as the functional initiator over any subsequent ATG codons.

If the 59 bases prior to the first ATG codon shown in Figure 2 are examined "in frame" with that ATG codon, it is found that they could code for 19 amino acids with no terminator codons. However, because this sequence is so rich in G and C residues (83%), 8 of the 19 amino acids would be Arg, including a stretch of 4 consecutive Arg residues. This would represent almost one-half the total of 17 Arg residues that are coded by the 1089 bases of open reading frame following the initial ATG codon, and if included in the sequence would cause a severe digression from the reported Arg content of purified ADA (3). It is of interest, but unknown significance, that the initial 59 base sequence does not contain a single T residue. The preponderance of circumstantial evidence would indicate that the functional initiator codon is as shown in Figure 2, although the possibility that the cDNA does not contain all of the 5' end of the coding sequence cannot be eliminated on the basis of direct evidence. The amino acid sequence of ADA has not been reported.

Sequence Variation

The sequences of two smaller adenosine deaminase cDNA molecules derived from human T-lymphoblast mRNA have recently been published (4,21). Orkin *et al* (21) published a sequence of 817 nucleotides. This sequence corresponds to nucleotide residues 421 - 1237 of the sequence we have presented (Fig. 2). The open reading frame in the shorter cDNA encodes a peptide sequence of 223 amino acids that agrees exactly with the final 223 residues of the protein sequence we have presented. The nucleotide sequences of the overlap area also agree, with the exception that residue

-59 GCCGCCGAGACCCACCACGCGCGCGGAGGGAGCAGCGCCGGGGCAGCAGGGGACCC -1

1 ATG GCC CAG ACS CCC SCC TTC BAC AAG CCC AAA GTA BAA CTB CAT GTC 48
MET Ala Gln Thr Pro Ala Phe Asp Lys Pro Lys Val Glu Leu His Val

49 CAC CTA BAC BGA TCC ATC AAG CCT BAA ACC ATC TTA TAC TAT GGC AGG 96
His Leu Asp Gly Ser Ile Lys Pro Glu Thr Ile Leu Tyr Tyr Gly Arg

97 AGG AGA BGG ATC GCC CTC CCA GCT AAC ACA GCA BAA BGG CTB CTB AAC 144
Arg Arg Gly Ile Ala Leu Pro Ala Asn Thr Ala Glu Gly Leu Leu Asn

145 GTC ATT GGC ATB BAC AAG CCG CTC ACC CTT CCA BAC TTC CTB GCC AAG 192
Val Ile Gly MET Asp Lys Pro Leu Thr Leu Pro Asp Phe Leu Ala Lys

193 TTT BAC TAC TAC ATB CCT GCT ATC GCG BGC TGC CAG BAC GCT ATC AAA 240
Phe Asp Tyr Tyr MET Pro Ala Ile Ala Gly Cys Arg Glu Ala Ile Lys

241 AGG ATC SCC TAT BAA TTT GTA BAA ATB AAG SCC AAA BAC BGC GTB GTB 288
Arg Ile Ala Tyr Glu Phe Val Glu MET Lys Ala Lys Glu Gly Val Val

289 TAT GTB BAA GTB CCG TAC AGT CCG CAC CTB CTB SCC AAC TCC AAA GTB 336
Tyr Val Glu Val Arg Tyr Ser Pro His Leu Leu Ala Asn Ser Lys Val

337 BAA CCA ATC CCC TGG AAC CAG GCT BAA GGG BAC CTC ACC CCA BAC BAA 384
Glu Pro Ile Pro Trp Asn Gln Ala Glu Gly Asp Leu Thr Pro Asp Glu

385 GTB GTB SCC CTA GTB GGC CAG GGC CTB CAG BAA GGG BAA CBA BAC TTC 432
Val Val Ala Leu Val Gly Gln Gly Leu Gln Glu Gly Glu Ala Arg Phe

433 GGG BTC AAG GCC CCG TCC ATC CTB TGC TGC ATB CBC CAC CAG CCC AAC 480
Gly Val Lys Ala Arg Ser Ile Leu Cys Cys MET Arg His Gln Pro Asn

481 TGG TCC CCC AAG GTB GTB BAA CTB TGT AAG AAG TAC CAG CAG CAG ACC 528
Trp Ser Pro Lys Val Val Glu Leu Cys Lys Tyr Gln Gln Gln Thr

529 GTB GTB SCC ATT BAC CTB GCT BGA BAT BAA ACC ATC CCA BGA AGC AGC 576
Val Val Ala Ile Asp Leu Ala Gly Asp Glu Thr Ile Pro Gly Ser Ser

577 CTC TTG CCT BGA CAT GTC CAG SCC TAC CAG BAA GCT GTB AAG AGC GGC 624
Leu Leu Pro Gly His Val Gln Ala Tyr Gln Glu Ala Val Lys Ser Gly

625 ATT CAC CGT ACT GTC CAC GCC GGG BAA GTB GGC TCG GCC BAA GTA GTA 672
Ile His Arg Thr Val His Ala Gly Glu Val Gly Ser Ala Glu Val Val

673 AAA BAA GTBAGGGCCTGGGCTGGCCATGGGGTCCCTCCTCACTGCTCTCCCACTCTG
Lys Glu

679
GCTCTATTCTGCTTCTCTACAG GCT GTB BAC ATA CTC AAG ACA BAA CCG CTB 708
Ala Val Asp Ile Leu Lys Thr Glu Arg Leu

709 BBA CAC BGC TAC CAC ACC CTB BAA BAC CAG GCC CTT TAT AAC AGG CTB 756
Gly His Gly Tyr His Thr Leu Glu Asp Gln Ala Leu Tyr Asn Arg Leu

757 CCG CAG BAA AAC ATB CAC TTC BAA ATC TGC CCC TGG TCC AGC TAC CTC 804
Arg Gln Glu Asn MET His Phe Glu Ile Cys Pro Trp Ser Ser Tyr Leu

805 ACT GGT GCC TGG AAG CCG BAC ACG BAA CAT BGA GTC ATT CCG CTC AAA 852
Thr Gly Ala Trp Lys Pro Asp Thr Glu His Ala Val Ile Arg Leu Lys

853 AAT BAC CAG GCT AAC TAC TCG CTC AAC ACA BAT BAC CCG CTC ATC TTC 900
Asn Asp Gln Ala Asn Tyr Ser Leu Asn Thr Asp Asp Pro Leu Ile Phe

901 AAG TCC ACC CTB BAC ACT BAT TAC CAG ATB ACC AAA CCG BAC ATB GGC 948
Lys Ser Thr Leu Asp Thr Asp Tyr Gln MET Thr Lys Arg Asp MET Gly

949 TTT ACT BAA BAA BAA TTT AAA BGG CTB AAC ATC AAT BCG SCC AAA TCT 996
Phe Thr Glu Glu Glu Phe Lys Arg Leu Asn Ile Asn Ala Ala Lys Ser

997 AGT TTC CTC CCA BAA BAT BAA AAG BGG BAA CTT CTC BAC CTB CTC TAT 1044
Ser Phe Leu Pro Glu Asp Glu Lys Arg Glu Leu Leu Asp Leu Leu Tyr

1045 AAA GCC TAT BGG ATB CCA CCT TCA GCC TCT BGA BGG CAG BAC CTC TBA 1092
Lys Ala Tyr Gly MET Pro Pro Ser Ala Ser Ala Gly Gln Asn Leu *

1093 AGACGCCACTCCTCCAGGCTTACCCTGTGGAGTCACCCCAACTCTGTGGGGCTGACGACCA 1155

1156 TTTTTCATTTTATTCCTTCCAGAGBACCATBATCTCAATAGTCACTTACTBATGCTCTGAA 1218

1219 CCCTATGTGTCCATTCTTGCACACACBTATACCTCGCATGG CCGCTCACTTCTCTBATTAT 1281

1282 GTGCCCTTGGCCAGGGACCGCCCTTGCACATGGGCATGGTTBATCTGAACCCCTCCTTCT 1344

1400
1345 GTGGCACTTGTACTGAAATCTGGTGTCTCAATGAAAGGCCCTGGCTGGTGGCA

TABLE 1 AMINO ACID COMPOSITION OF HUMAN ADENOSINE DEAMINASE

Amino Acid	Number of Residues	
	Predicted by nucleotide sequence	Determined by amino acid analysis*
Alanine	31	30
Arginine	17	18
Asparagine	13	34
Aspartic acid	21	
Cysteine	5	n.d.
Glutamic acid	30	45
Glutamine	15	
Glycine	24	30
Histidine	11	12
Isoleucine	17	17
Leucine	36	36
Lysine	24	24
Methionine	9	6
Phenylalanine	10	11
Proline	21	22
Serine	17	21
Threonine	17	18
Tryptophan	4	n.d.
Tyrosine	16	11
Valine	25	25

*Amino acid composition of the protein as previously published (3). n.d., not determined.

534 in our sequence is G while their sequence contains an A at that location. In one case this is the last nucleotide in a GTG codon for Val while in the other case it becomes a GTA codon for Val, therefore not resulting in a change in amino acid sequence.

Valerio *et al* (4) published a human adenosine deaminase cDNA sequence of 822 nucleotides. This sequence corresponds to nucleotide residues 581-1400 of the sequence we have presented (Fig. 2). Their sequence contains a single additional T residue on the 3' end of the molecule. There are two nucleotide differences in the region where our sequence and theirs overlap. Both differences are in the 3' non-coding region. Residue 1387 is G in our sequence while they have A at that location. They also have an additional C between residues 1251 and 1252 of our sequence. The

Figure 2: The nucleotide sequence of human adenosine deaminase cDNA (ADA 3-3). The amino acid sequence is shown beneath the nucleotide sequence from which it is deduced. The DNA sequence is numbered in both directions from the first ATG codon. The polyadenylation signal (residues 1375-1380) is underlined. The unnumbered 76 base sequence between 678 and 679 represents the intron described in Figure 3.

Figure 3. Identification of intron in human adenosine deaminase cDNA (ADA 3-3). Sequence I represents residues 665-683 of the sequence in Figure 2, without the intervening sequence of 76 bases between 678 and 679. Sequence II represents residues 671-683 of Figure 2 with the intervening sequence of 76 bases (overlined) between residues 678 and 679. Sequence III represents the consensus sequence for splice sites in eucaryotic nuclear mRNA precursors (23,24). The arrows indicate the splice points. Underlined bases are invariant. Py represents a pyrimidine and X is any base.

169 amino acids coded by their open reading frame corresponds exactly with the final 169 amino acids in the protein sequence we have presented.

Intron sequence

If the cDNA sequences of Orkin *et al* (21) and Valerio *et al* (4) are compared they overlap by 657 bases. Their sequences in this overlap area agree exactly. This overlap sequence corresponds to and agrees exactly with residues 581-1237 of our sequence (Fig. 2). The single difference is an additional sequence of 76 bases which is shown unnumbered between residues 678 and 679 in our sequence. We previously described the isolation of two different human adenosine deaminase cDNAs, ADA 2-11 and ADA 3-3 (13). Restriction maps of these two cDNA molecules indicated that they were very similar except for an additional sequence of at least 40 nucleotides in one of the Hpa II fragments of ADA 3-3 (13). This additional sequence contained an Ava II restriction site. The location of this additional sequence corresponds exactly to the location of the 76 base sequence described above, and indeed an Ava II restriction site is present in the twenty-fourth through twenty-eighth bases of this 76 base sequence. Possible explanations offered for the additional DNA in ADA 3-3 were: (a) representation of different adenosine deaminase alleles, (b) derivation of the cDNA from a poly(A)⁺ RNA not yet fully processed but containing a small intron, or (c) artifactual insertion or deletion of DNA during the cloning procedure. Another possible explanation would be different mRNA species arising from alternative splicing within the coding region such as was recently described for fibronectin (22). The explanations of difference due to different alleles or alternative mRNA splicing can be discounted because the addition of 76 bases would shift the reading frame of the sequence and totally disrupt the resulting protein sequence. Since the cDNA was prepared from

poly(A)⁺RNA isolated from total cellular RNA (13), we thought it possible that the 76 base sequence might represent a small intron which had not been spliced out of a nuclear mRNA precursor. This possibility was considered as shown in Figure 3, by examining the flanking sequences at each end of the 76 base sequence. The consensus sequence for splice sites in eucaryotic nuclear mRNA precursors is shown as sequence III in Figure 3 (23,24). The splice points are indicated by arrows. The underlined residues (GT at the 5' terminus of the intron and AG at the 3' terminus of the intron) are essentially invariant whereas the remainder of the consensus sequence is an average sequence around which there is considerable fluctuation (24). In Figure 3, sequence II shows the flanking sequences around the ends of the 76 base sequence of additional DNA in cDNA clone ADA 3-3. The overlined bases in sequence II represent the ends of that 76 base sequence and sequences II and III have been aligned so that these ends are adjacent to the splice points of the consensus sequence. Sequence II agrees with sequence III not only at the four invariant sites but also at 13 of the remaining 16 sites in which there is a base preference. This represents overwhelming evidence that the 76 base sequence corresponds to a small intron which has not been spliced out of the sequence. Sequence I represents the sequence if the 76 base intron is deleted and, as discussed earlier, is identical to sequence at this location in two other human adenosine deaminase cDNAs (4,21).

ACKNOWLEDGMENTS

This work was supported by Research Grant HD-15036 from the National Institutes of Health and by the Veterans Administration Research Service.

*To whom correspondence should be addressed

REFERENCES

1. Tischfield, J.A., Creagan, R.P., Nichols, E.A., and Ruddk, F.H. (1974) *Hum. Hered.* 24, 1-11.
2. Daddona, P.E., and Kelley, W.N. (1977) *J. Biol. Chem.* 252, 110-115.
3. Wiginton, D.A., Coleman, M.S., and Hutton, J.J. (1981) *Biochem. J.* 195, 389-397.
4. Valerio, D., Duyvesteyn, M.G.C., Meera Khan, P., Geurts van Kessel, A., de Waard, A., and van der Eb, A.J. (1983) *Gene* 25, 231-240.
5. Giblett, E.R., Anderson, J.E., Cohen, F., Pollara, B., and Meuwissen, H.J. (1972) *Lancet* ii, 1067-1069.
6. Martin, D.W., Jr., and Gelfand, E.W. (1981) *Annu. Rev. Biochem.* 50, 845-877.
7. Thompson, L.F., and Seegmiller, J.E. (1980) in *Advances in Enzymology*, Meister, A. Ed., pp. 167-210, Wiley, New York.
8. Hirschhorn, R., Roegner, V., Jenkins, T., Seaman, C., Piomelli, S., and Bor-kowsky, W. (1979) *J. Clin. Invest.* 64, 1130-1139.
9. Daddona, P.E., Frohman, M.A., and Kelley, W.N. (1980) *J. Biol. Chem.* 255, 5681-5687.
10. Wiginton, D.A., and Hutton, J.J. (1982) *J. Biol. Chem.* 257, 3211-3217.
11. Adrian, G.S., and Hutton, J.J. (1983) *J. Clin. Invest.* 71, 1649-1660.

12. Hirschhorn, R., Martiniuk, F., Roegner-Maniscalco, V., Ellenbogen, A., Perignon, J.L., and Jenkins, T. (1983) *J. Clin. Invest.* 71, 1887-1892.
13. Wiginton, D.A., Adrian, G.S., Friedman, R.L., Suttle, D.P., and Hutton, J.J. (1983) *Proc. Natl. Acad. Sci. USA* 80, 7481-7485.
14. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
15. Messing, J. (1981) in the *Third Cleveland Symposium of Macromolecules: Recombinant DNA*, Walton, A., Ed., pp. 143-153, Elsevier, Amsterdam.
16. Weislander, L. (1979) *Anal. Biochem.* 98, 305-309.
17. Proudfoot, N.J., and Brownlee, G.G. (1974) *Nature* 252, 359-362.
18. Fitzgerald, M., and Shenk, T. (1981) *Cell* 24, 251-260.
19. Kozak, M. (1983) *Microbiol. Rev.* 47, 1-45.
20. Kozak, M. (1981) *Nucleic Acids Res.* 9, 5233-5252.
21. Orkin, S.H., Daddona, P.E., Shewach, D.S., Markham, A.F., Bruns, G.A., Goff, S.C., and Kelley, W.N. (1983) *J. Biol. Chem.* 258, 12753-12756.
22. Schwarzbauer, J.E., Tamkun, J.W., Lemischka, I.R., and Hynes, R.O. (1983) *Cell* 35, 421-431.
23. Sharp, P.A. (1981) *Cell* 23, 643-646.
24. Cech, T.R. (1983) *Cell* 34, 713-716.