Volume 288, number 1,2, 168–172 FEBS 010072 August 1991

ADONIS 001457939100783K

Genomic structure of human lysosomal glycosylasparaginase

Hyejeong Park, Krishna J. Fisher and Nathan N. Aronson Jr.

Department of Molecular and Cell Biology, Althouse Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

Received 29 May 1991

The gene structure of the human lysosomal enzyme glycosylasparaginase was determined. The gene spans 13 kb and consists of 9 exons. Both 5′ and 3′ untranslated regions of the gene are uninterrupted by introns. A number of transcriptional elements were identified in the 5′ upstream sequence that includes two putative CAAT boxes followed by TATA-like sequences together with two AP-2 binding sites and one for Sp1. A 100 bp CpG island and several ETF binding sites were also found. Additional AP-2 and Sp1 binding sites are present in the first intron. Two polyadenylation sites are present and appear to be functional. The major known glycosylasparaginase gene defect $G^{488} \rightarrow C$, which causes the lysosomal storage disease aspartylglycosaminuria (AGU) in Finland, is located in exon 4. Exon 5 encodes the post-translational cleavage site for the formation of the mature α/β subunits of the enzyme as well as a recently proposed active site threonine, Thr^{206}.

Glycosylasparaginase; Gene structure; Intron–exon junction; Promoter element

1. INTRODUCTION

Glycosylasparaginase (EC 3.5.1.26) is a lysosomal amidase which cleaves the GlcNAc-Asn bond that forms the linkage between peptide and oligosaccharide in Asn-linked glycoproteins [1]. The importance of this reaction in the degradation of glycoproteins is emphasized by the occurrence of aspartylglucosaminuria (AGU), a lysosomal storage disease caused by glycosylasparaginase deficiency [2]. AGU patients accumulate GlcNAc-Asn in lysosomes and typically become mentally retarded and show physical deterioration as juveniles, but can survive beyond four decades. AGU is genetically inherited as an autosomal recessive trait and is found mainly in Finland [2]. The estimated incidence of AGU in Finland is approximately 1:26000 and the heterozygote carrier frequency is as high as 1:40 in northern Finland [3].

Rat liver glycosylasparaginase has previously been purified and well characterized in this laboratory and was found to be a 49 kDa heterodimer composed of 24 and 20 kDa subunits joined by non-covalent bonding [4]. The human enzyme has been purified [5–7] and recently we [8], and later scientists at the Finnish National Public Health Institute [9], reported the cloning of a cDNA for human glycosylasparaginase. A single gene encodes the enzyme which has a deduced molecular weight of 34.6 kDa. The two subunits, 23 kDa (α) and 17–18 kDa (β), are produced by post-translational proteolytic cleavage [8,10]. The gene has been previously mapped to chromosome 4 and its size estimated to be 15 kb [9].

We have now isolated a genomic clone of human glycosylasparaginase and are reporting its structure including the 5′ untranslated sequence. This new information will allow a better characterization of the molecular nature of the disease AGU in all its forms as more patients are diagnosed.

Abbreviation: AGU, aspartylglucosaminuria

Correspondence address: N.N. Aronson Jr., Department of Molecular and Cell Biology, Althouse Laboratory, The Pennsylvania State University, University Park, PA 16802, USA.

2. MATERIALS AND METHODS

2.1. *Materials*

The sources of the major experimental materials used in this work were as follows: a λFixII human genomic library from Stratagene, restriction enzymes and random-primed labeling kit from Boehringer Mannheim, biotinyl-11-dUTP from Sigma, [α-^{35}S]dCTP from New England Nuclear-Dupont, Hybond N membranes from Amersham, Sequenase version 2.0 DNA sequencing kit from United States Biochemical Co.

2.2. *Methods*

Human glycosylasparaginase cDNA insert HPAsn.6 [8] was labeled with biotinyl-11-dUTP by random-primed labeling and used as a probe to screen the genomic library [11]. Phage DNA was prepared and analyzed by restriction enzyme digestion and Southern blotting [12]. Fragments which had exons were subcloned into pUC18, pBluescript II SK(+), or M13mp 18/19, and further analyzed or sequenced by the dideoxynucleotide chain termination method [13]. dITP was used to sequence GC rich regions. Restriction maps were constructed by multiple or partial restriction enzyme digestions.

3. RESULTS AND DISCUSSION

3.1. *Isolation and characterization of the glycosylasparaginase gene*

1×10^6 phages of the λFixII human genomic library were screened and 6 positive clones were obtained after

Published by Elsevier Science Publishers B.V.

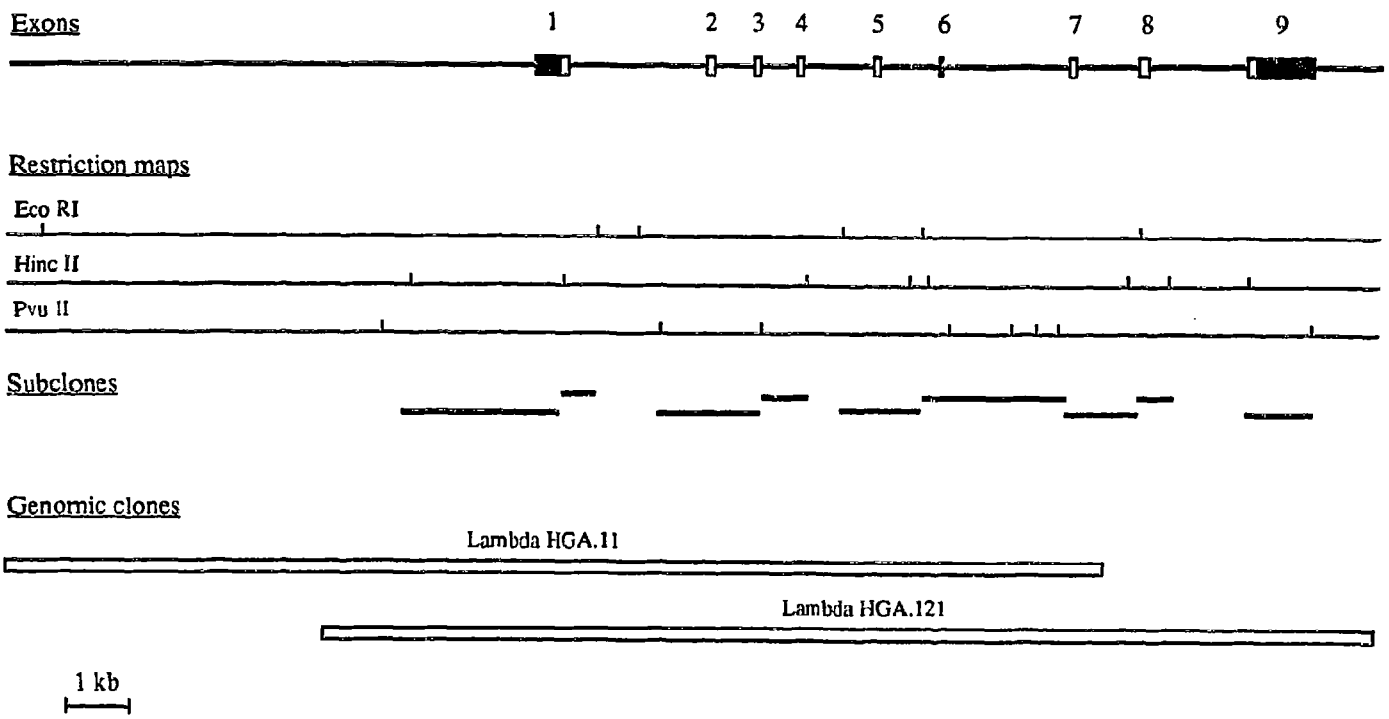

Fig. 1. Organization and restriction map of the human glycosylasparaginase gene. The schematic structure of the gene is shown at the top. Exons are represented by open boxes and black areas indicate non-coding regions. Open bars at the bottom of the figure denote inserts of two genomic clones, λHGA.11 and λHGA.121. Subclones used for sequence analysis are shown above these two primary genomic clones. Partial sequencing of these subclones included only exons and exon-intron boundaries.

four rounds of screening. By restriction mapping two of the positive clones, λHGA.11 and λHGA.121, were found to contain the longest segments of 5′ or 3′ sequence present in cDNA HPAsn.6, and they were chosen for further analysis (Fig. 1). Phage DNA was digested with *Not*I to release the genomic DNA inserts which were further fragmented with *Eco*RI. The fragments which hybridized with cDNA were eluted

Table I

Nucleotide sequence of the intron-exon boundaries in the human glycosylasparaginase gene. Exon sequences are in upper-case letters; intron sequences are in lowercase. Consensus sequences at the 5′ and 3′ borders of introns are shown at the bottom.

Exon	Exon size (bp)	Sequence of intron-exon junction								Intron size (kb)
		5′ Border				3′ Border				
1	>640	GAA	GCA	G 127	gtgcg	ggatttgattaacag	CG 128	TGG	AGG	2.1
2	154	ATG	GAT	GG 281	gtaga	tctattttcttgcag	C 282	ACT	ACT	0.7
3	113	GAG	TCA	G 394	gtatt	ttttccaatctccag	CC 395	ACC	ACA	0.9
4	113	TAT	TGG	AGG 507	gtatg	aaatcttgtttagag	AAT 508	GTT	ATA	1.2
5	115	ACT	ATT	G 622	gtaat	ttttttaacttctag	GC 623	ATG	GTT	0.9
6	76	ATA	CAT	GG 698	gttag	aataccctctcaaag	C 699	CGT	GTA	1.9
7	108	CTG	CCA	AG 806	gtatg	ctctgtttcaatcag	C 807	TAC	CAA	1.1
8	134	AGT	TAC	G 940	gtaag	cattttggccccag	GT 941	GCT	GCT	1.5
9	1040									

consensus: $\begin{matrix}A\\C\end{matrix}$AG gt (pu) ag ..... $(py)_{nn}$ (py) ag G

from the agarose gel and separately subcloned into pUC18 or pBluescript II. *Hin*cII and *Pvu*II were used for further dissection of subclones (Fig. 1), and fragments with exons were subcloned into M13mp18/19 and sequenced.

The glycosylasparaginase gene spans approximately 13 kb and consists of 9 exons. Exon 1 encodes 5′ untranslated sequence and the first 42 amino acids which include the 23 amino acid signal peptide. Exon 9 codes for the last 33 amino acids and a long 3′ untranslated sequence (939nt) which is not interrupted by an intron and includes three polyadenylation signals, two adjacent at 1152 and 1164 and one at 1950 [9]. All intron-exon junctions fit the GT/AG consensus rule (Table I) [14]. Nucleotide sequences of coding regions agreed with those of the cDNA [8,9], but a major difference was found in the 5′ non-coding region. Starting at 63 bp upsteam of the ATG translation initiation codon (−63), the sequence of genomic DNA in the 5′ direction was completely different from that of the cDNA reported by Ikonen et al. [9] (Fig. 2). We do not think this is caused by the presence of an intron which splits the first exon. First, the sequence at the required intron-exon boundary does not fit either the GT/AG rule (TT instead of AG (Fig. 2)) or other consensus structure at the 3′ border (GC-rich sequence instead of $(Py)_nN(Py)AG$) [14]. Second, the genomic sequence shows several promoter elements 5′ upstream of this divergence site (Fig. 2). These facts support the idea that the first exon contains coding region and 5′ upstream sequence which is not interrupted by an intron. No significant match to this inconsistent sequence at the 5′ end of the cDNA was found in GenBank, and the sequence may have resulted from a cloning artifact. A single discrepancy also occurred in the 3′ untranslated region at +1224 where A changed to C in the genomic DNA. This varied nucleotide might be a polymorphism.

3.2. *5′ untranslated region*

The 5′ untranslated region was sequenced to −513. Based on the 2.3 kb size of glycosylasparaginase

```
                                                    ATCAATCCGGTCCTAAGCAAACA   -491

CGCTTCACCTACACTCATAACTATTGCAGACCTCCGAGGCCTGGATCCCCAAGATATACTGAGTTTGACA   -421

AACTTTTCAACTTCAACTTTAAATTAAAAAGACAGTAAAAGAACCAAATCCATAGTACACAGCAATCGGC   -351
          AP-2
TAAAGTTCCTGGGGCCCTGCAACCCAGAGTTGAATAATTTTGTATTAAATTCCCAATATCAAGCTAAATC   -281

TATTTTAACCACAGGAAAAAGCAGCAACTTGTCTGGCTATTTTAAAAATCTGAACAGCACTTAGGAAGAA   -211
                                                          AP-2
GCACCTTAGCGCAGGGAACAGCTCAGTGCCCCGTGACACAACTCTCCCGCGGGCCCAGGGACGCCTCGTC   -141
                                                    TGTATTGAAACAATTTAATGAAAATATTAA
                                                                        Sp1
TCGCGAGAGTTGAGGGACGCCTGAGCGAACCCCCGAGAGAGCGGGCGTGGGCGCCAGGCGGGCGGGGCAC   -71
ATATTTGGTTTCAAAAGGCAGATTTATCTTCTCCCAACATTCTGTTATTTCTGATACTTTTGAAAAACTA

TGGGGATTAATTGTTCGGCGATCGCTGGCTGCCGGGACTTTTCTCGCGCTGGTCTCTTCGGTGGTCAGGG   -1
ATAAAAAC

ATGGCGCGGAAGTCGAACTTGCCTGTGCTTCTCGTGCCGTTTCTGCTCTGCCAGGCCCTAGCGCGCTGCT    70

CCAGCCCTCTGCCCCTGGTCGTCAACACTTGGCCCTTTAAGAATGCAACCGAAGCAGgtgcgggttggcg   140
  AP-2 Sp1 Sp1                                          Sp1
gcctgggcgggcggctgcccgagagcttgcgcgtggggatgactagccgccccgctgcagtcacccggc   210
                                                    Sp1
aggcggacggaagagcgccgggctgggccgggcctcgcggtccccgcccctctaggtcttcccgagtctcca   280

gcgccgctcagtctcttgcaagtcgcctttactgcatcctccccacctaccgctgttgcattttagcgtc   350

ctcccttcatgcagcccccagcgcc .tgtattttccaccctgctgctgaccatagttatggcctttgtg   420
```

Fig. 2. Nucleotide sequence of the 5′ region of the human glycosylasparaginase gene. Nucleotide A of the initiation codon ATG is desinated +1. The first intron is in lower-case letters. Putative CAAT and TATA motifs are boxed. AP-2 sites are overscored and Sp1 sites are double-underlined. The sequence between the brackets is a CpG island. The underneath sequence in capital letters depicts the sequence at the 5′ end of a full-length cDNA reported by Ikonen et al. [9] that differs from the corresponding genomic sequence.

message which was detected in human fibroblasts [10], we can estimate the approximate location of the promoter site to be near −250. There are 2 CAAT boxes both of which fall in this region (−358 and −298) and are followed by TATA-like sequences (−316, −308, −286, and −280). So far, among lysosomal enzymes it is known that β-hexosaminidase α subunit [15], α-galactosidase A [16], and cathepsin G [17] genes have both TATA and CAAT boxes in their 5′ flanking region. β-hexosaminidase β subunit [18], cathepsin L [19], α-glucosidase [20], acid phosphatase [21], and arylsulfatase A [22] genes either lack these promoter elements or have only a CAAT box and show features of housekeeping genes that lack a TATA box, e.g. the presence of a GC-rich region, GC box, and other upstream promoter elements [23,24]. The 5′ flanking region of the glycosylasparaginase gene also has some characteristics of housekeeping genes. It has two potential AP-2 sites [24] at −343 and −157 and one potential Sp1 site [24] at −81. Interestingly, more AP-2 and Sp1 sites were found in the first intron: AP-2 site at 142, Sp1 sites at 145, 149, 187, and 252. The rat cathepsin L gene also has been found to have promoter elements in its first intron [19], but it is not known whether they are functional. A 100 bp GC-rich region (78%) was located between −165 and −66, and the CpG/GpC ratio was 14/13. The length of this CpG island [25], however, is relatively short compared to others which are several hundred bp [26]. Binding motifs (CCCC or GGGG) of ETF, a transcriptional factor which stimulates transcription from TATA box-lacking promoters [27], were found in the 5′ upstream region at −340, −182, −111, −96, and −77. The 5′ upstream region of glycosylasparaginase gene has features of both TATA box-containing and TATA box-lacking promoters. Further investigations are therefore necessary to determine precisely where transcription starts and which upstream elements actually function in the expression of the glycosylasparaginase gene.

3.3. *3′ untranslated region*

Exon 9 contains the last 34 codons followed by a 939 bp 3′ untranslated region (Fig. 1, see also [8,9]. Three distinct polyadenylation signals were found in this exon, two adjacent at 1152 and 1164 [8,9] and a third at 1950 [9]. These two polyadenylation sites are separated by approximately 800 bp and interruption by an intron is unlikely, since a single cDNA spans this region [9]. From the Northern blot of human fibroblast total RNA, two glycosylasparaginase messages of 2.3 kb and 1.5 kb were found in equal amount [10]. Based on the distance separation of these polyadenylation signals in the 3′ exon, these two messages likely originate from differential use of the two signals, rather than from differential splicing of introns. β-Hexosaminidase α chain has also been reported to have two transcripts using different polyadenylation signals [15].

The gene structure of glycosylasparaginase will be helpful for characterization of mutations which cause AGU and should allow development of recombinant DNA probes to detect the disease and its genetic carriers. So far only the major Finnish AGU mutation has been determined [9,10,28], and this is in exon 4 where G^{488} has been mutated to C. This point mutation results in the amino acid change Cys^{163} to Ser and creates a new *Eco*RI restriction site. PCR primers for the detection of this major point mutation have already been designed on the basis of the intron-exon structure of the gene [10], and other non-Finnish mutations are now being studied in our laboratory [29]. Kaartinen et al. [7] recently described labelling the N-terminal threonine of the β (small) subunit of the human enzyme with an active-site inhibitor, DONV. This residue is in exon 5, and eventually natural or experimentally designed mutations coding this region of the human glycosylasparaginase gene may be discovered which can lead to a better understanding of the mechanism of this important lysosomal hydrolase.

Acknowledgement: This work was supported by United States Public Health Service Grant DK-33314 from the National Institute of Diabetes and Digestive and Kidney Diseases.

REFERENCES

[1] Makino, M., Kojima, T., and Yamashida, I. (1966) Biochem. Biophys. Res. Commun. 24, 961–966.
[2] Maury, C.P.J. (1982) J. Inher. Metab. Dis. 5, 192–196.
[3] Aula, P., Renlund, M. and Raivio, K.O. (1986) J. Ment. Defic. Res. 30, 365–368.
[4] Tollersrud, O.K. and Aronson Jr., N.N. (1989) Biochem. J. 260, 101–108.
[5] McGovern, M.M., Aula, P. and Desnick, R.J. (1983) J. Biol. Chem. 258, 10743–10747.
[6] Baumann, M., Peltonen, L., Aula, P. and Kalkkinen, N. (1989) Biochem. J. 262, 189–194.
[7] Kaartinen, V., Williams, J.C., Tomich, J., Yates, III, J.R., Hood, L.E. and Mononen, I. (1991) J. Biol. Chem. 266, 5860–5869.
[8] Fisher, K.J., Tollersrud, O.K. and Aronson Jr., N.N. (1990) FEBS Lett. 269, 440–444.
[9] Ikonen, E., Baumann, M., Gron, K., Syvanen, A.-C., Enomaa, N., Halila, R., Aula, P. and Peltonen, L. (1991) EMBO J. 10, 54–58.
[10] Fisher, K.J. and Aronson Jr., N.N. (1991) J. Biol. Chem. 266, 12105–12113.
[11] Kincaid, R.L. and Nightingale, M.S. (1988) BioTechniques 6, 42–49.
[12] Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning, a laboratory manual, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
[13] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
[14] Breathnach, R. and Chambon, P. (1981) Annu. Rev. Biochem. 50, 349–383.
[15] Proia, R.L. and Soravia, E. (1987) J. Biol. Chem. 262, 5677–5681.
[16] Quinn, M., Hantzopoulos, P., Fidanza, V. and Calhoun, D.H. (1987) Gene 58, 177–188.
[17] Hohn, P.A., Popescu, N.C., Hanson, R.D., Salvesen, G. and Ley, T.J. (1989) J. Biol. Chem. 264, 13412–13419.

[18] Neote, K., Bapat, B., Dumbrille-Ross, A., Troxel, C., Schuster, S.M., Mahuran, D.J. and Gravel, R.A. (1988) Genomics 3, 279–286.
[19] Ishidoh, K., Kominami, E., Suzuki, K. and Katunuma, N. (1989) FEBS Lett. 259, 71–74.
[20] Hoefsloot, L.H., Hoogeveen-Westerveld, M., Reuser, A.J.J. and Oostra, B.A. (1990) Biochem. J. 272, 493–497.
[21] Geier, C., Von Figura, K. and Pohlmann, R. (1989) Eur. J. Biochem. 183, 611–616.
[22] Kreysing, J., Von Figura, K. and Gieselmann, V. (1990) Eur. J. Biochem. 191, 627–631.
[23] Dynan, W.S. (1986) Trends Genet. 2, 196–197.
[24] Mitchell, P.J. and Tjian, R. (1989) Science 245, 371–378.
[25] Bird, A.P. (1986) Nature 321, 209–213.
[26] Gardiner-Garden, M. and Frommer, M. (1987) J. Mol. Biol. 196, 261–282.
[27] Kageyama, R., Merlino, G.T. and Pastan, I. (1989) J. Biol. Chem. 264, 15508–15514.
[28] Mononen, I., Heisterkamp, N., Kaartinen, V., Williams, J.C., Yates III, J.R., Griffin, P.R., Hood, L.E. and Groffen, J. (1991) Proc. Natl. Acad. Sci., USA 88, 2941–2945.
[29] Fisher, K.J. and Aronson Jr., N.N. (1991) FEBS Lett. 288, 173–178.