The EMBO Journal vol.10 no.1 pp.51 – 58, 1991

# Aspartylglucosaminuria: cDNA encoding human aspartylglucosaminidase and the missense mutation causing the disease

**Elina Ikonen[1], Marc Baumann[2], Kristiina Grön[3], Ann-Christine Syvänen[1], Nina Enomaa[1], Ritva Halila[1], Pertti Aula[3] and Leena Peltonen[1]**

[1]National Public Health Institute, Laboratory of Molecular Genetics, Mannerheimintie 166, SF-00300 Helsinki, [2]University of Helsinki, Department of Medical Chemistry, Siltavuorenpenger 10 A, SF-00170 Helsinki and [3]University of Turku, Department of Medical Genetics, Kiinamyllynkatu 10, SF-20520 Turku, Finland

Communicated by K.I.Kivirikko

**We have isolated a 2.1 kb cDNA which encodes human aspartylglucosaminidase (AGA, E.C. 3.5.1.26). The activity of this lysosomal enzyme is deficient in aspartylglucosaminuria (AGU), a recessively inherited lysosomal accumulation disease resulting in severe mental retardation. The polypeptide chain deduced from the AGA cDNA consists of 346 amino acids, has two potential *N*-glycosylation sites and 11 cysteine residues. Transient expression of this cDNA in COS-1 cells resulted in increased expression of immunoprecipitable AGA protein. Direct sequencing of amplified AGA cDNA from an AGU patient revealed a G → C transition resulting in the substitution of cysteine 163 with serine. This mutation was subsequently found in all the 20 analyzed Finnish AGU patients, in the heterozygous form in all 53 carriers and in none of 67 control individuals, suggesting that it represents the major AGU causing mutation enriched in this isolated population. Since the mutation produces a change in the predicted flexibility of the AGA polypeptide chain and removes an intramolecular S–S bridge, it most probably explains the deficient enzyme activity found in cells and tissues of AGU patients.**

*Key words:* aspartylglucosaminidase/aspartylglucosaminuria/cDNA/lysosomal/mutation

## Introduction

Aspartylglucosaminidase (AGA, E.C. 3.5.1.26) is a key enzyme in the catabolism of *N*-linked oligosaccharides of glycoproteins (Aronson and Kuranda, 1989). It cleaves the asparagine from the residual *N*-acetylglucosamines as one of the final steps in the lysosomal breakdown of glycoproteins. Deficient activity of AGA in man results in aspartylglucosaminuria (AGU, McKusick 208400), the only known lysosomal storage disorder caused by an amidase deficiency. At the cellular level this is reflected by the accumulation of the abnormal breakdown products, mainly aspartylglucosamine, in the lysosomes (Maury, 1980). The clinical manifestations of this recessive disorder demonstrate the consequences of the enzyme deficiency in various organs: progressive mental retardation starting from the age of two to five years and minor connective tissue changes (Autio, 1972). The mutation is strongly enriched in the Finnish population with a gene frequency of 1:100, but occasional cases have been reported from several other countries and various ethnic groups (Aula *et al.*, 1986; Borud and Torp, 1976; Gehler *et al.*, 1981; Hreidarsson *et al.*, 1983).

Only limited amounts of AGA have been purified from human sources (McGovern *et al.*, 1983; Baumann *et al.*, 1989). Largely due to this fact, the detailed structure of the human enzyme is still unknown. The gene locus of human AGA has been assigned to the long arm of chromosome 4 using interspecies cell hybrids and taking advantage of the different thermal stabilities of human and hamster AGA (Aula *et al.*, 1984). Our recent linkage analyses in AGU families suggest the same location for the AGU mutation (Grön *et al.*, 1990). These findings support the idea of a mutation in the gene coding for the AGA protein as the cause of the AGU disease.

Here we report the cloning, nucleotide sequence and expression of the full-length cDNA for human AGA. Sequence analysis of the *in vitro* amplified AGA cDNA from an AGU patient revealed a missense mutation resulting in the conversion of a cysteine residue to a serine residue at the amino acid position 163. The same mutation was subsequently identified in 20 AGU patients by Southern blot analysis. This mutation is very likely to be the predominant cause of AGU in the Finnish population.

## Results

### *AGA peptide sequences*

In order to synthesize oligonucleotides for library screening, the AGA protein was purified to homogeneity essentially as described (Baumann *et al.*, 1989). Failure to sequence the amino-terminal end of the purified AGA protein suggested the blockage of the amino terminus. Two structurally different fragments, of sizes 17 and 24 kd, are constantly seen in SDS–PAGE analysis of the homogeneous enzyme and represent the post-translational cleavage products of the enzyme with a mol. wt of ~60 kd in gel filtration (Baumann *et al.*, 1989; Halila *et al.*, 1991). These fragments were isolated and sequenced. The amino terminus of the 24 kd fragment was again blocked but the 17 kd fragment yielded the amino-terminal sequence: TIGMVVIHKTGHIAAGTSTNGIKFKIHGRV. The sequence analysis of peptides produced by digesting these fragments with endoproteinases Asp-N or Lys-C provided us with the sequence of a total of 187 amino acids. No high frequency homologies were found when the protein sequence data bank (SWISS-PROT 13) was screened for homologies with these peptide sequences.

### *Characterization of the AGA cDNA*

The amino acid sequences used to synthesize oligonucleotide mixtures based on the most probable mammalian codon usage (Lathe, 1985) are marked in Figure 1. A human fetal liver λgt10 cDNA library was screened with these oligo-

```
        -169
          TGTATTGAAACAATTTAATGAAAATATTAAATATTTGGTTTCAAAAGGCAGATTTATCTTCTCCCAACATTCTGTTATTT   -91

CTGATACTTTTGAAAAACTAATAAAAACAATTGTTCGGCGATCGCTGGCTGCCGGGACTTTTCTCGCGCTGGTCTCTTCGGTGGTCAGGG   -1

ATGGCGCGGAAGTCGAACTTGCCTGTGCTTCTCGTGCCGTTTCTGCTCTGCCAGGCCCTAGTGCGCTGCTCCAGCCCTCTGCCCCTGGTC    90
 M  A  R  K  S  N  L  P  V  L  L  V  P  F  L  L  C  Q  A  L  V  R  C  S  S  P  L  P  L  V     30
                                                                  ↑
GTCAACACTTGGCCCTTTAAGAATGCAACCGAAGCAGCGTGGAGGGCATTAGCATCTGGAGGCTCTGCCCTGGATGCAGTGGAGAGCGGC   180
 V  N  T  W  P  F  K  N  A  T  E  A  A  W  R  A  L  A  S  G  G  S  A  L  D  A  V  E  S  G     60
                      *
TGTGCCATGTGTGAGAGAGAGCAGTGTGACGGCTCTGTAGGCTTTGGAGGAAGTCCTGATGAACTTGGAGAAACCACACTAGATGCCATG   270
 C  A  M  C  E  R  E  Q  C  D  G  S  V  G  F  G  G  S  P  D  E  L  G  E  T  T  L  D  A  M     90

ATCATGGATGGCACTACTATGGATGTAGGAGCAGTAGGAGATCTCAGACGAATTAAAAATGCTATTGGTGTGGCACGGAAAGTACTGGAA   360
 I  M  D  G  T  T  M  D  V  G  A  V  G  D  L  R  R  I  K  N  A  I  G  V  A  R  K  V  L  E    120

CATACAACACACACTTTTAGTAGGAGAGTCAGCCACCACATTTGCTCAAAGTATGGGGTTTATCAATGAAGACTTATCTACCAGTGCT   450
 H  T  T  H  T  L  L  V  G  E  S  A  T  T  F  A  Q  S  M  G  F  I  N  E  D  L  S  T  S  A    150

TCTCAAGCTCTTCATTCAGATTGGCTTGCTCGGAATTGCCAGCCAAATTATTGGAGGAATGTTATACCAGATCCCTCAAAATACTGCGGA   540
 S  Q  A  L  H  S  D  W  L  A  R  N  C  Q  P  N  Y  W  R  N  V  I  P  D  P  S  K  Y  C  G    180
                   •  ▲
CCCTACAAACCACCTGGTATCTTAAAGCAGGATATTCCTATCCATAAAGAAACAGAAGATGATCGTGGTCATGACACTATTGGCATGGTT   630
 P  Y  K  P  P  G  I  L  K  Q  D  I  P  I  H  K  E  T  E  D  D  R  G  H  D  T  I  G  M  V    210
                                                                  o→
GTAATCCATAAGACAGGACATATTGCTGCTGGTACATCTACAAATGGTATAAAATTCAAAATACATGGCCGTGTAGGAGACTCACCAATA   720
 V  I  H  K  T  G  H  I  A  A  G  T  S  T  N  G  I  K  F  K  I  H  G  R  V  G  D  S  P  I    240

CCTGGAGCTGGAGCCTATGCTGACGATACTGCAGGGGCAGCCGCAGCCACTGGGAATGGTGATATATTGATGCGCTTCCTGCCAAGCTAC   810
 P  G  A  G  A  Y  A  D  D  T  A  G  A  A  A  A  T  G  N  G  D  I  L  M  R  F  L  P  S  Y    270

CAAGCTGTAGAATACATGAGAAGAGGAGAAGATCCAACCATAGCTTGCCAAAAAGTGATTTCAAGAATCCAGAAGCATTTTCCAGAATTC   900
 Q  A  V  E  Y  M  R  R  G  E  D  P  T  I  A  C  Q  K  V  I  S  R  I  Q  K  H  F  P  E  F    300

TTTGGGGCTGTTATATGTGCCAATGTGACTGGAAGTTACGGTGCTGCTTGCAATAAACTTTCAACATTTACTCAGTTTAGTTTCATGGTT   990
 F  G  A  V  I  C  A  N  V  T  G  S  Y  G  A  A  C  N  K  L  S  T  F  T  Q  F  S  F  M  V    330
                      *
TATAATTCCGAAAAAAATCAGCCAACTGAGGAAAAAGTGGACTGCATCTAATCCATCTTTACTGTCAACATCTGTATTTAAAGAAGAAAG  1080
 Y  N  S  E  K  N  Q  P  T  E  E  K  V  D  C  I  -                                          346

AAACAAAGGCTGAAAAGGCTGCTCACTCTCATCATCTAGTGTTCCTCATGTGTTCTAAAGTCTTTTTGTAAAATAAACACAAAAATAAAT  1170

TTATGTTGAATTGGCACTTTCTATATCTGAAATTGTATTATCTATTTTTATTTAACCTTTCTTATATTATCTGAAGATGAATATGTATGT  1260

GGCTGTATTTTGTATATATTTAAGTTTTAAGTGACACTTGGATTTCTGATCAGTGTTGCAAAAAATTGCTACTTTGAGATTTATTTCCAC  1350

AGTGATATATTAAACCCAAAAGGAATCATCATATTTGGATGCTTAAAATAGAGTATACAAGTTTTTTAGACCCAAAGAAAAAAACATGCA  1440

GAGTCTCAGCATGATCAGTCCTGATTGAATTTCATTATGTCACAGTGATGCATGTCTTAGCACTGATTATTGGTTTTACTTAATATCCCC  1530

TAAAAACTTGAGGAAGGAAGATGATAGAACATGAGATACTCTTTGTTTTGTACATACTAGAAAATACAGCTATCTCCTGATGAGCAATTG  1620

TGTTGATTTTCTTTGCTTTTTTCCTCTTATGGAGCTCTCAGTGTGTGAAAGTTCTAAAGCCACAAGAATCAATGATGTTTTAGCTGCTAA  1710

TATTTAGTATTGATCGTTTCTTTTAAAAATGATTTGGGGAACTTGATCTGTACTGTAATTTTACTAATTCCTTGTGCAGAGGTGGGTATG  1800

GCAATTAGAAGTCAAGAAATGGGTTGTTTAGCTTGGTTTGTTTCCCTAACTTCTGATTAACTCTCTGTATGACAACTCTACAGAAGTTGT  1890

GCGCGTGCTTTCTCAGCAGCATTTTTCCTTCAAAATCATCTTTTTAATCAACAGTCATTAATAAATGTACATGTACGATATTCAAACTTT  1980

AAAAAAAAAA                                                                                   1990
```

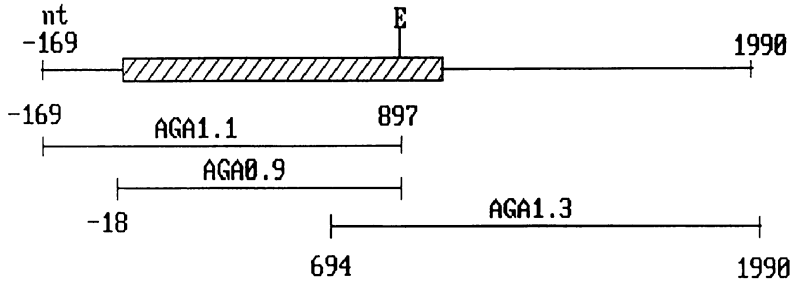


**Fig. 1.** (**a**) The cDNA and deduced amino acid sequence of human AGA. Nucleotides are numbered in the 5′ → 3′ direction, and given negative numbers upstream of the predicted translation initiation ATG codon. The putative signal peptidase cleavage site is marked by ↑, and the two potential *N*-glycosylation sites by *. Amino acid sequences obtained from peptide sequencing are underlined. The peptide sequences used to synthesize oligonucleotides for library screening are underlined with double dashed lines, those peptides used to synthesize oligonucleotides for the confirmation of the enriched clones are underlined with single dashed lines (see text). The amino-terminal end of the 17 kd fragment is indicated by o→. The polyadenylation signal is boxed. The changes found in the nucleotide sequence of the AGU cDNA are marked with ▲ and °. (**b**) Schematic presentation of the overlapping AGA cDNA clones isolated from a liver λgt10 library. The nucleotides are numbered as in (a) and the positions of the clones in the AGA cDNA are indicated by the first and last nucleotide of the clone. The internal *Eco*RI site (E) is indicated.

nucleotide mixtures as hybridization probes. The enriched clones were further confirmed in Southern hybridizations of the cDNA inserts using additional oligonucleotides as probes. Those cDNAs identified with at least one additional oligonucleotide were selected for further analyses. A clone containing a 1.1 kb insert cross-hybridized with all the oligonucleotides and 24 enriched smaller clones. This insert, AGA1.1, was subcloned into the *Eco*RI site of

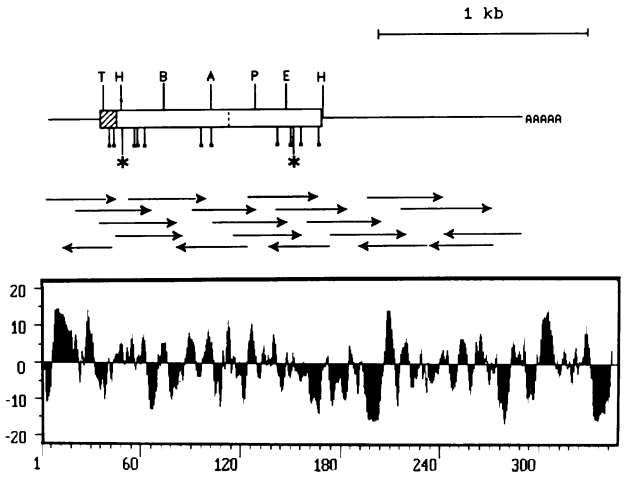


**Fig. 2. Upper part**. Schematic presentation of the AGA cDNA and the sequencing strategy used. The coding region is marked as a box with the localized restriction sites for *Taq*I (T), *Hin*dII (H), *Bgl*II (B), *Ava*II (A), *Pst*I (P) and *Eco*RI (E). The hatched part of the box represents the signal peptide and the line non-translated areas of the AGA cDNA. The arrows indicate the extent and direction of the sequencing reactions obtained with each 15mer sequencing primer. The asterisks (*) indicate possible *N*-glycosylation sites and squares (↓) the cysteine residues of the deduced polypeptide chain. **Lower part**. The hydropathy profile of the AGA polypeptide chain predicted from the AGA cDNA. Positive and negative numbers correspond to hydrophobic and hydrophilic regions, respectively.

pGEM7Z-vector and sequenced. The sequencing strategy is summarized in Figure 2.

The sequence of AGA1.1 contained an open reading frame from position −66 to 897 but was not followed by a termination codon. Instead this clone and several other clones obtained had identical 3′ sequences, corresponding to a cleaved *Eco*RI site. Obviously in the construction of the cDNA library, the *Eco*RI site in these clones was not protected by *Eco*RI methylase. To find additional clones containing the missing 3′ end of the cDNA, PCR amplification was applied. Selected clones were amplified using as the 5′ primer 22 nucleotides of the cDNA sequence and a 27 or 31 bp sequence of both λgt10 arms as the other primer. We obtained amplification products from two clones and their inserts were subsequently subcloned and sequenced. In one clone, AGA1.3, we found an open reading frame overlapping with the nucleotide sequence of the AGA1.1 clone. In addition this clone contained a 10 nucleotide long poly(A)$^+$ tail following the termination codon after an untranslated region of 949 nucleotides. The polyadenylation signal (AATAAA) was detected 25 nucleotides prior to the poly(A)$^+$ tail. The total length of the AGA cDNA obtained was 2159 nucleotides, which contained a coding region of 1041 nucleotides (Figure 1).

The open reading frame of the AGA cDNA begins at nucleotide position −66, and the first methionine following this is given the amino acid position 1 (Figure 1). The initiation sequence with a purine in nucleotide position −3 and a G in position +4 agrees with the consensus sequence for an initiator methionine in eukaryotic mRNAs (Kozak, 1987). The 30 amino acids following the initiator methionine contained a stretch of 10 hydrophobic residues (amino acids 7−16). Potential signal peptidase cleavage sites were identified in two positions by the weight matrix method of von Heijne: i.e. between amino acids 19 and 20 (score 7.62) and between amino acids 23 and 24 (score 7.31). Neither of these sites is preceded by an α-helix breaker (such as proline or glycine), but otherwise the region demonstrates characteristics typical of a signal peptide (von Heijne, 1986). The cleavage site between amino acids 23 and 24 is located at the end of the predicted membrane spanning region, which suggests that this location is the actual cleavage site.

The predicted amino acid sequence encoded by the AGA cDNA contained only two potential *N*-glycosylation sites, both located in the hydrophilic β-turn areas (Figures 1 and 2) which is consistent with their probable surface location in the folded protein. The hydropathy profile of AGA polypeptide determined according to Kyte and Doolittle (1982) shows that hydrophilic regions are dispersed along the entire sequence, the longest hydrophilic stretch being the 11 amino acid region from position 195 to 205 (Figure 2). The only hydrophobic region of at least 10 amino acids was the predicted signal peptide. The number of cysteine residues was as high as 11, nine of these located in two clusters, five in the beginning and four at the end of the polypeptide chain. The predicted amino acid sequence showed 94% identity (176 amino acids out of 187) with the sequences obtained from the proteolytic peptides (see Figure 1). The nucleotide structure of the open reading frame predicted a polypeptide chain of 346 amino acids, which is in good agreement with the size of the deglycosylated human AGA having a mol. wt of 40 kd (Halila *et al.*, 1991). The predicted site of post-translational cleavage to the 24 and 17 kd polypeptides is the protease sensitive hydrophilic stretch of 11 amino acids at positions 195−205. Neither the nucleotide sequence of the AGA cDNA nor the predicted amino acid sequence revealed significant homologies in the EMBL nucleic acid and protein databases.

### *Northern and Southern analyses, chromosomal location*

When poly(A)$^+$ RNA from cultured skin fibroblasts was hybridized with the AGA cDNA, a major RNA species with a length of 2.1 kb was detected, but two additional minor signals corresponding to a mRNA species of ~3.4 kb were also consistently present (Figure 3). The size of the major hybridization signal was not significantly different in the Northern analyses of mRNA isolated from cultured fibroblasts of AGU patients. However, in this preliminary study, the level of transcription was slightly decreased in the skin fibroblasts of Finnish AGU patients when compared with that in control fibroblast lines (see Figure 3).

In Southern hybridization analyses of human genomic DNA digested with different restriction enzymes, the AGA cDNA visualized 1−5 fragments which summed up to ~15 kb, independently of the restriction enzyme used. This suggested that the human genome contains a single copy gene of ~15 kb coding for the AGA protein. To establish the chromosomal location of the genomic area corresponding to AGA cDNA, a panel of human−hamster somatic cell hybrids was hybridized with AGA cDNA. The signals obtained were in concordance only with localization to chromosome 4 (Figure 4).

### *Transient expression of AGA cDNA in COS-1 cells*

Preliminary experiments demonstrated that both human fibroblasts and COS-1 cells have considerable endogenous AGA activity, which is also immunoprecipitated with specific antibodies against human AGA. However, during short

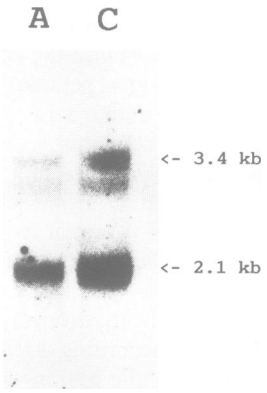


**Fig. 3.** Northern hybridization of human fibroblast poly(A)$^+$ RNA with the AGA cDNA. Lane A contains 3 μg of poly(A)$^+$ RNA isolated from cultured fibroblasts of an AGU patient. Lane C contains 3 μg of fibroblast poly(A)$^+$ RNA from an age-matched control individual (non-carrier of AGU). The indicated sizes of the mRNA species detected by AGA cDNA are based on the relative migration of the ribosomal subunits used as mol. wt markers and visualized with ethidium bromide staining.

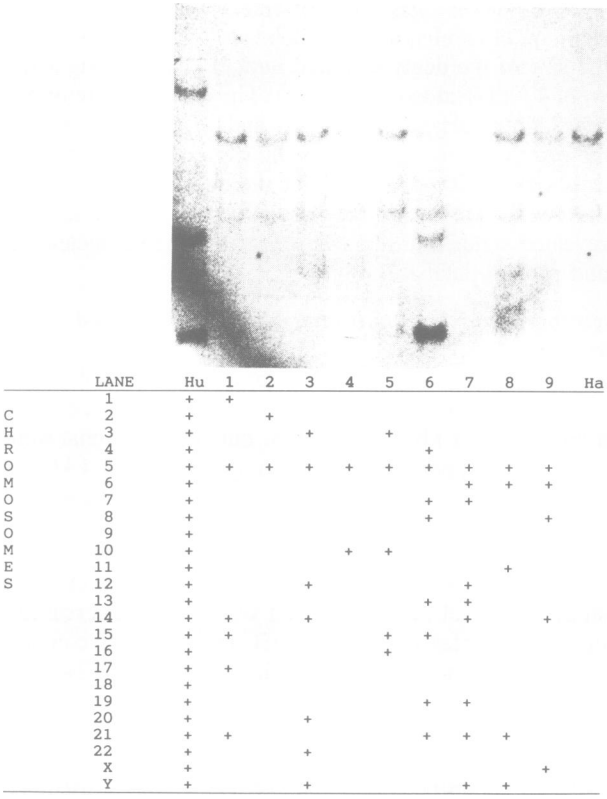


**Fig. 4.** Southern hybridizations of representative *Taq*I digested human–hamster hybrid cell DNAs with the $^{32}$P-labeled AGA cDNA clone. Each lane contains a different hybrid cell DNA and the presence of the respective human chromosome is indicated with (+) in the panel. The human AGA-specific DNA fragments were detected only in lane 6 correlating with the presence of chromosome 4.

pulses of radioactive labeling, the high relative expression of the cDNA construct compared with the endogenous expression could be expected.

The longest AGA cDNA clone having the translation initiation codon as the first ATG of the sequence (Figure 1) was chosen for expression. This clone (AGA0.9, Figure

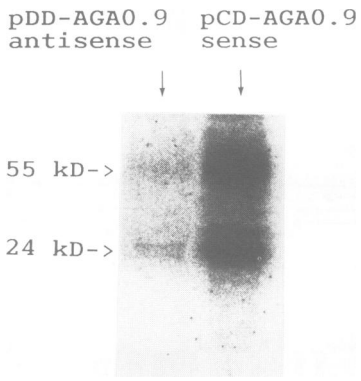


**Fig. 5.** Transient expression of the 0.9 kb AGA cDNA constructs in COS-1 cells. The cDNA constructs in the sense and antisense orientations were transfected into COS-1 cells. After 48 h the cells were labeled with [$^3$H]Leu for 5 h and the labeled proteins were immunoprecipitated from the cells with AGA antibodies, analyzed on SDS–polyacrylamide gel and visualized by fluorography. Molecular sizes were calculated by comparison with protein markers. Exposure time was 3 days.

1), which lacked the most 3′ end of the cDNA, was cloned in both sense and antisense orientations into a derivative of the mammalian expression vector pCD-X to transfect COS-1 cells, and newly synthesized radiolabeled proteins were immunoprecipitated with AGA antibodies. The typical intracellular AGA polypeptides, the uncleaved 55–60 kd fragment and the post-translational cleavage product of 24 kd, were detected from the COS-1 cells with the sense AGA cDNA, and significantly lower amounts of endogenous AGA were immunoprecipitated in COS-1 cells with the antisense construct (Figure 5). As expected, the pCD-AGA0.9-sense construct resulted in the expression of a slightly smaller protein than the endogenous AGA of the COS-1 cells, and no increase beyond endogenous enzyme activity could be detected in transfected cells.

### *The mutation resulting in AGU*

To search for the mutation the cDNA from an AGU patient was amplified using two sets of PCR primers designed on the basis of the AGA cDNA sequence and directly sequenced. A missense change, G → C, which results in the substitution of a cysteine with a serine at amino acid position 163 and creates a new *Eco*RI site, was detected (Figure 6). The same mutation was subsequently identified in the cDNAs amplified from several AGU patients by cleavage of the PCR product by *Eco*RI, while the cDNAs from healthy individuals remained uncleaved (Figure 7). The relevance of this mutation for the AGU disease was further confirmed by Southern hybridizations in 13 non-related families with one or more affected siblings. The new *Eco*RI site was constantly detected in the homozygous form in AGU patients and in the heterozygous form in their parents and other individuals known to be carriers based on the activity of AGA in their lymphocytes (Figure 8). In the screening of 68 non-related phenotypically healthy individuals, we found one heterozygote for this mutation. When measuring the enzyme activity in her leukocytes, she turned out to be a carrier of AGU as well. So far we have not detected a single AGU patient without this mutation among 20 Finnish patients studied.

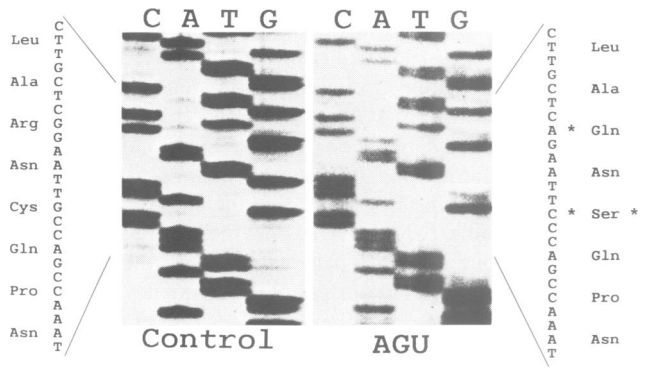


**Fig. 6.** The AGU mutation. DNA sequencing gels of the AGA cDNA of an AGU patient and a control. The autoradiography of sequences of the coding strand in the 5′ → 3′ direction from the top to the bottom of the gel is shown. The missence mutation, G → C, found in all AGU patients so far, and the following substitution Cys163 → Ser is marked with two stars. The other change, G → A, resulting in the substitution Arg161 → Gln is marked with one star.

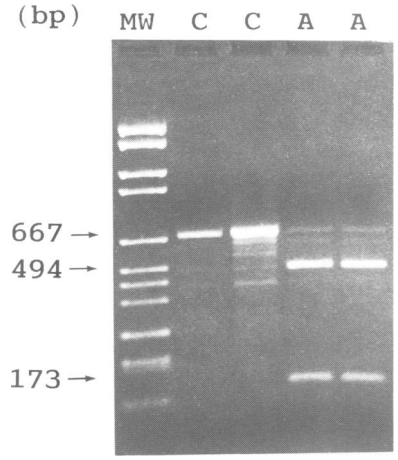


**Fig. 7.** Demonstration of the AGU mutation by *Eco*RI digestions of the PCR products from cDNAs of two non-related AGU patients (A) and two controls (C). The mutation creates an *Eco*RI site resulting in the cleavage of the 667 bp PCR product into two fragments of 494 bp and 173 bp, whereas no cleavage occurs in the controls.

We found one additional missense change in the sequenced AGU cDNA, G → A, which results in the conversion of arginine to glutamine at the amino acid position 161, only 5 bp upstream from the AGU mutation (Figure 6). This change creates a new restriction site for *Dde*I. We have identified this change in the amplified cDNAs of three unrelated AGU patients in both alleles but not in four controls.

The significance of both amino acid changes identified was further analyzed by predicting the flexibility at each point of the AGA polypeptide chain using the method of Karplus and Schulz (1985). The Cys → Ser mutation results in a dramatic increase in the flexibility in the middle of the AGA polypeptide chain (Figure 9) changing this region to the second most flexible region of the protein. The Arg → Gln mutation does not produce any changes in the predicted flexibility of the polypeptide chain. Further, despite the charge difference between arginine and glutamine, the conversion Arg → Gln has no effect on the predicted pI value of the encoded polypeptide chain when occurring with the Cys → Ser mutation.

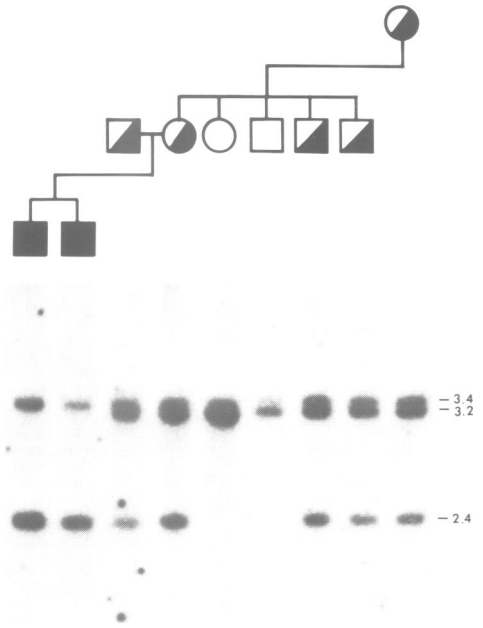


**Fig. 8.** Southern hybridization of an AGU family with the AGA cDNA demonstrating the segregation of the AGU mutation. In AGU patients (■) both alleles are cleaved by *Eco*RI, resulting in 2.4 kb fragments, in the carriers (◪, ◑) both the uncleaved 3.2 kb and the cleaved 2.4 kb allele are present. The non-carriers (□, ○) demonstrate only the uncleaved allele. An additional constant 3.4 kb fragment is visualized in all the samples.

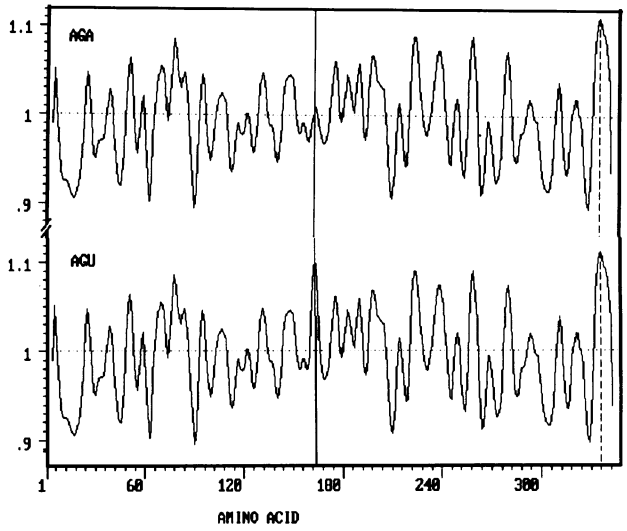


**Fig. 9.** The predicted chain flexibility of the AGA polypeptide plotted at each point of the protein sequence. The vertical axis shows the units of flexibility, with an average flexibility of 1. The horizontal axis shows the amino acid number along the protein sequence. The solid vertical line indicates the mutation region, and the dashed line the polypeptide region with the highest flexibility.

## Discussion

cDNA encoding human AGA was isolated from a fetal liver cDNA library. This tissue is known to contain significant amounts of active AGA (McGovern *et al.*, 1983; Baumann *et al.*, 1989). Authenticity of this cDNA for AGA was confirmed by several methods. Firstly, the amino acid sequence deduced from the nucleotide sequence of this

cDNA is colinear with that determined by direct amino acid sequencing of AGA peptides. Secondly, expression of the partial cDNA in monkey COS-1 cells showed that the cDNA contains the necessary nucleotide sequence for increased expression of immunodetectable AGA in heterologous cells. Thirdly, in a panel of hamster–human somatic cell hybrids, the sequences homologous to human genomic AGA segregated with human chromosome 4, which is in agreement with the location established earlier by the enzyme assay in somatic cell hybrids (Aula *et al.*, 1984). Finally, a missense mutation was detected in the coding region of this cDNA in several AGU patients.

The structural characteristics of the AGA polypeptide based on the nucleotide sequence of the cDNA resemble the structural features of several other lysosomal enzymes. The sequence contains no predicted membrane spanning segments other than the putative signal peptide; thus the cleavage of this domain will result in a soluble mature protein. Two potential *N*-glycosylation sites are located on the predicted surface areas of the folded protein. Based on earlier glycosidase digestions of the purified AGA protein of rat and man, both of these sites are most probably utilized (Tollersrud and Aronson, 1989; Halila *et al.*, 1991). All the peptides sequenced from the 24 kd fragment were found in the amino-terminal part, and all the 17 kd peptides in the carboxy-terminal end of the nucleotide sequence encoded by the AGA cDNA. The amino-terminal sequences of both fragments were highly homologous with the reported 24 and 20 kd fragments of rat AGA (Tollersrud and Aronson, 1989).

The mRNA identified in human fibroblasts by the AGA cDNA is 2.1 kb long, thus the 2159 bp cDNA reported here must represent the full-length cDNA. The explanation for other minor mRNA species could be alternative splicing or the use of another polyadenylation signal.

Due to the limited amount of this housekeeping-type enzyme in tissues and cell lines, the detailed structural analyses of AGA would require the production of this enzyme by *in vitro* expression. The AGA0.9-sense construct used here for transient expression is far from ideal. The missing poly(A)$^+$ tail along with the 3′ end of the cDNA decreases the half-life of the mRNA resulting in low efficiency of protein production. However, this construct demonstrates that immunoreactive AGA can be expressed from our cDNA clone. We are currently pursuing expression experiments with complete AGA cDNA constructs in different cell lines to produce large quantities of the human AGA enzyme in an active form.

AGU is the only human lysosomal storage disorder demonstrating the biological consequences of deficient amidase activity (Aronson and Kuranda, 1989). Interestingly, all the AGU patients so far reported have some residual enzyme activity in their tissues and cells (Aula *et al.*, 1982, 1986; Gehler *et al.*, 1981). As total absence of AGA activity has never been reported in a living individual, it is most probably incompatible with life.

Until now, the defective gene and protein in AGU have not been identified, although the disease itself was reported for the first time in 1968 (Pollitt *et al.*, 1968). The association between AGU and deficient AGA activity has been known ever since, but this does not necessarily imply that mutation of the AGA gene would explain the disturbed function of this amidase. Examples of defects in different regulatory proteins have been reported in other lysosomal accumulation diseases (Conzelman and Sandhoff, 1979; d'Azzo *et al.*, 1982; Inui *et al.*, 1983; Christomanou *et al.*, 1986). Our previous data demonstrating the consistent chromosomal location for both the AGA activity and the AGU disease (Aula *et al.*, 1984; Grön *et al.*, 1990) are also only suggestive due to the limitations of the methods used. Here we demonstrate, for the first time, that the structural mutation in a gene coding for the AGA enzyme results in AGU in the majority of the Finnish patients, arising from the population with the highest reported AGU frequency.

The mutation described here changes one of the cysteine residues to a serine. The number of cysteine residues in the polypeptide chain coded by the AGA cDNA is surprisingly high (11 residues) for a 40 kd polypeptide chain. Based on our protein data we know that intramolecular S–S bridges do not hold the post-translational cleavage products, i.e. the 17 and 24 kd fragments, together (Baumann *et al.*, 1989; Halila *et al.*, 1991). Consequently, the S–S bridges form only within these fragments and are most probably essential for the secondary structure of the active enzyme. This assumption is supported by the missense mutation reported here. This mutation most probably has a dramatic effect on the enzymatic activity since it affects the predicted flexibility of the AGA polypeptide chain by disturbing the formation of intra-chain S–S bridges. Our recent studies on the purified AGU protein demonstrate that these predicted changes actually take place *in vivo* (T.Heiskanen, in preparation).

The second amino acid substitution found in the AGA cDNA (Arg → Gln) does not result in changes in the predicted flexibility or pI of the AGA polypeptide chain and could thus represent a protein polymorphism. However, this will be difficult to prove, especially if the polymorphism is rare. Firstly, this change is tightly coupled to the AGU mutation due to the extremely short distance between these two nucleotide changes. Secondly, although the change creates a new restriction site, Southern hybridizations are inapplicable for family and population analyses due to the small sizes of *Dde*I fragments. We are currently developing a PCR strategy to amplify this region directly from genomic DNA. However, most probably only expression of the active enzyme and the use of focused *in vitro* mutagenesis of the AGA cDNA will finally clarify the significance of the Arg substitution for the pathogenesis of AGU.

## Materials and methods

### *Library screening*

A total of $2 \times 10^6$ recombinant plaques from several unamplified cDNA libraries were screened by the high density method of Benton and Davis (1977) using mixtures of oligonucleotides as probes. The oligonucleotides were end-labeled with [$\gamma$-$^{32}$P]ATP using T4 polynucleotide kinase. Duplicate nylon filters were hybridized at 42°C for 20 h in 6 × SSC (1 × SSC: 150 mM sodium chloride, 15 mM sodium citrate), 0.5% sodium sarcosine, 1 × Denhardt's solution (Denhardt, 1966), 1 mM EDTA and 20 μg/ml yeast tRNA. The filters were washed with 6 × SSC, 0.1% SDS several times at room temperature, then at 42°C for 30 min and at 50°C for 10 min, and exposed to Kodak X-AR film. Plaques positive on both filters were picked into SM buffer (0.1 M Tris–HCl, pH 8.0, 0.15 M NaCl, 10 mM $MgCl_2$) for further screening, and enriched single clones were isolated after tertiary screening. Of the enriched clones, 5–50 μl of phage stock in SM buffer was boiled for 15 min and used for PCR amplification in conditions specified below. Each clone was amplified using a primer

from the AGA cDNA (5′-TGATTTCAAGAATCCAGAAGCA, nucleotides 868–887) together with either the λgt10 forward or reverse primer (Promega).

### *Synthesis of oligonucleotides*

The method for purification of AGA from human liver has been published elsewhere (Baumann *et al.*, 1989). To obtain large quantities of the enzyme, human leukocytes originating from 140 l of blood from voluntary blood donors, and originally used to produce α-interferon, were used as enzyme source (Halila *et al.*, 1991). The purification procedure consisted of affinity chromatography, gel filtration, chromatofocusing and reverse phase HPLC and resulted in a 13 000-fold purified homogeneous enzyme preparation confirmed by SDS–PAGE; 150–400 μg homogeneous enzyme was produced in each purification. The 17 and 24 kd fragments of the AGA polypeptide chain were separated by reverse phase HPLC and digested with endoproteinase Lys-C or Asp-N. The peptides produced were separated again by reverse phase HPLC. Amino acid sequencing from these peptides was performed by automated Edman degradation using a modified Applied Biosystems 477A/120A on-line pulsed liquid phase/gas phase sequencer in the gas phase mode. The oligonucleotides were synthesized on the basis of the amino acid sequence data on a DNA synthesizer (Applied Biosystems 381A) using cyanoethyl phosphoramidite chemistry.

### *Subcloning and sequencing*

The two partially overlapping AGA cDNA clones of 1.1 and 1.3 kb from the λgt10 liver library were digested with *Eco*RI and subcloned into pGEM-7Z (Promega). The clones were subjected to restriction endonuclease mapping. The nucleotide sequence of the clones was determined by the dideoxynucleotide chain termination method (Sanger *et al.*, 1977) using the Sp6 and T7 promoter primers and 15mer sequencing primers. Figure 2 summarizes the sequencing strategy and shows a partial restriction map of the AGA cDNA.

### *Transient expression of AGA cDNA in COS-1 cells*

The AGA cDNA clone containing the region from nucleotide −18 to 897 (AGA0.9) was excised from the λgt10 vector by digestion with *Eco*RI and subcloned into the *Eco*RI site of the mammalian expression vector pCD-X (Okayama and Berg, 1983) in both orientations. In the pCD-AGA0.9-sense construct the 5′ end of the cDNA was oriented towards the SV40 promoter, whereas in the antisense construct the orientation was reversed. The constructs were transfected separately into COS-1 cells by employing the liposome transfection method (Felgner *et al.*, 1987) (Lipofectin reagent, BRL). Control cells were treated in the same manner as the transfected cells except that no DNA was added. After 48 h the transfected and control cells were incubated with [$^3$H]leucine (Amersham) for an additional 5 h in leucine-free medium. The cells were harvested by trypsinization and homogenized into 20 mM sodium phosphate buffer containing 0.5% Triton X-100. Radiolabeled proteins were immunoprecipitated with the IgG fraction of antiserum raised against homogeneous AGA protein (Halila *et al.*, 1991). After washing the pellet three times wth 0.01 M Tris–HCl, pH 8.6, 0.6 M NaCl, 0.1% SDS and 0.05% NP-40, the immunoprecipitated proteins were analyzed on a 12% SDS–PAGE (Laemmli, 1970) and visualized by fluorography. The fixed (15% methanol, 30% acetic acid) gel was impregnated with En$^3$Hance solution (New England Nuclear), dried and exposed for 3–7 days.

### *Southern analyses with AGA cDNA, chromosome blots*

Southern analyses (Southern, 1975) were performed in 13 unrelated AGU families with a total of 20 patients and 90 healthy family members (18 of these married into the AGU families) and 50 additional unrelated control individuals. The carriership of the family members was demonstrated by reduced AGA activity in isolated lymphocytes (Aula *et al.*, 1976). Fifty-two of the family members turned out to be carriers. Genomic DNA was extracted from blood samples (Kunkel *et al.*, 1977) and 4 μg of DNA was digested to completion with *Eco*RI. The DNA fragments were fractionated on 0.8% agarose gels and transferred to Hybond N nylon filters (Amersham). The AGA cDNA was labeled with $^{32}$P by the random primer extension reaction and hybridized to the filters in 6 × SSC, 5 × Denhardt's solution (Denhardt, 1966), 100 μg/ml sonicated herring sperm DNA and 0.5% SDS at 65°C overnight. The filters were washed once in 2 × SSC, and twice in 2 × SSC, 0.1% SDS at 65°C. The restriction patterns were observed by autoradiography after exposure at −80°C for 16 h or longer. The chromosome panel blots were hybridized with the labeled AGA cDNA according to the instructions of the manufacturer (Bios Corporation).

### *RNA extractions and Northern analysis*

Poly(A)$^+$ RNA was prepared from cultured fibroblasts using the guanidine isothiocyanate method (Chirgwin *et al.*, 1979) and oligo(dT)–cellulose chromatography. For Northern analyses (Alwine *et al.*, 1977), 1–4 μg of poly(A)$^+$ RNA was electrophoresed in an 0.8% agarose gel containing formaldehyde, transferred to a Biodyne nylon membrane (Pall) and then hybridized to a $^{32}$P-labeled AGA cDNA under conditions recommended by the manufacturer. After washing in 2 × SSC, 0.1% SDS for 30 min at 25–45°C and 0.1 × SSC, 0.1% SDS for 15 min at 65°C, the filters were exposed to X-ray film at −80°C for 1–4 days.

### *PCR amplification, restriction enzyme digestion and sequencing of cDNA from AGU patients*

First strand cDNA was synthesized by reverse transcription from 1 μg of total RNA using an AGA-specific primer (primer 1b, see below). The cDNA was extracted with chloroform and precipitated with ethanol before amplification by PCR (Mullis and Faloona, 1987). The cDNA was amplified as two overlapping fragments using two primer pairs:

Primer 1a: 5′-TGGTCAGGGATGGCGCGG (nt −9 to +9)
Primer 1b: 5′-AGCAATATGTCCTGTCTTATGG (nt 657–636)
Primer 2a: 5′-CCTACAAACCACCTGGTATC (nt 542–561)
Primer 2b: 5′-GATGGATTAGATGCAGTCCAC (nt 1047–1027).

PCR was carried out with 100 pmol of primers in 100 μl of a solution of 0.2 mM each of dATP, dCTP, dGTP, dTTP, 20 mM Tris–HCl, pH 8.8, 15 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 0.1% Tween 20, 0.01% gelatin and 2.5 U of *Thermus aquaticus* DNA polymerase (Amplitaq, Perkin-Elmer/Cetus) in a programmable heating block (Techne PHC1) for 30 cycles for 1 min at 96°C, 1 min at 60°C and 3 min at 72°C.

An aliquot of ~200 ng of the 5′ PCR product amplified with the primers 1a and 1b was digested with *Eco*RI or *Dde*I and analyzed in a 2% agarose gel. For sequencing the amplified DNA was purified from a 1% agarose gel. The gel slice containing the DNA was excised from the gel, crushed into TE buffer, extracted with phenol, quickly frozen in liquid $N_2$ and centrifuged. After extraction with chloroform the DNA was precipitated with ethanol and sequenced by the chain termination method.

## Acknowledgements

We thank Ms Ritva Timonen and Ms Tuula Manninen for excellent technical assistance and Ms Helena Rantanen for help in secretarial work. We acknowledge Drs Fisher, Tollersrud and Aronson, Jr for sharing with us their unpublished sequence data of partial AGA cDNA, which helped us to confirm our data. This work was supported by grants from the Academy of Finland, Duodecim Foundation and the Research and Science Foundation of Farmos.

## References

Alwine,J.C., Kemp,D.J. and Stark,G.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5350–5354.
Aronson,N.N., Jr and Kuranda,M.J. (1989) *FASEB J.*, **3**, 2615–2622.
Aula,P., Raivio,K.O. and Autio,S. (1976) *Pediatr. Res.*, **10**, 625–629.
Aula,P., Autio,S., Raivio,K.O. and Rapola,J. (1982) *Genetic Errors of Glycoprotein Metabolism*. Springer-Verlag, Berlin.
Aula,P., Astrin,K.H., Francke,U. and Desnick,R.J. (1984) *Am. J. Hum. Genet.*, **36**, 1215–1224.
Aula,P., Renlund,M., Raivio,O. and Koskela,S.-L. (1986) *J. Ment. Defic. Res.*, **30**, 365–368.
Autio,S. (1972) *J. Ment. Defic. Res. Monogr.*, **Ser I**, 1–39.
Baumann,M., Peltonen,L., Aula,P. and Kalkkinen,N. (1989) *Biochem. J.*, **262**, 189–194.
Benton,W.D. and Davis,R.W. (1977) *Science*, **196**, 180–182.
Borud,O. and Torp,K.H. (1976) *Lancet*, **i**, 1082–1083.
Chirgwin,J.M., Przybyla,A.E., MacDonald,R.J. and Rutter,W.J. (1979) *Biochemistry*, **18**, 5294–5299.
Christomanou,H., Aignesberger,A. and Linke,R. (1986) *Hoppe-Seyler's Z. Physiol. Chem.*, **367**, 879–890.
Conzelman,E. and Sandhoff,K. (1979) *Hoppe-Seyler's Z. Physiol. Chem.*, **360**, 1837–1849.
d'Azzo,A., Hoogeveen,A.T., Reuser,A.J.J., Robinson,D. and Galjaard,H. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 4535–4539.
Denhardt,D.T. (1966) *Biochem. Biophys. Res. Commun.*, **23**, 641–646.
Felgner,P.L., Gadek,T.R., Holm,M., Roman,R., Chan,H.W., Wenz,M., Northrop,J.P., Ringold,G.M. and Danielsen,M. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 7413–7417.
Gehler,J., Sewell,A.C., Becker,C., Hartmann,J. and Spranger,J. (1981) *Helv. Paediatr. Acta*, **36**, 179–189.

Grön,K., Aula,P. and Peltonen,L. (1990) *Hum. Genet.*, **85**, 233–236.
Halila,R., Baumann,M., Ikonen,E., Enomaa,N. and Peltonen,L. (1991) *Biochem. J.*, in press.
Hreidarsson,S., Thomas,G.H., Valle,D.L., Stevenson,R.E., Taylor,H., McCarty,J., Coker,S.B. and Green,W.R. (1983) *Clin. Genet.*, **23**, 427–435.
Inui,K., Emmett,M. and Wenger,D.A. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3074–3077.
Karplus,P.A. and Schulz,G.E. (1985) *Naturwissenschafter*, **72**, 212–213.
Kozak,M. (1987) *Nucleic Acids Res.*, **15**, 8125–8148.
Kunkel,L.M., Smith,K.D., Boyer,S.H., Borganonkar,D.S., Wachtel,S.S., Miller,O.J., Breg,W.R. and Jones,H.W. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 1245–1249.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Laemmli,U.K. (1970) *Nature*, **227**, 680–685.
Lathe,R. (1985) *J. Mol. Biol.*, **183**, 1–12.
Maury,P. (1980) *J. Lab. Clin. Med.*, **96**, 838–840.
McGovern,M.M., Aula,P. and Desnick,R.J. (1983) *J. Biol. Chem.*, **258**, 10743–10747.
Mullis,K.B. and Faloona,F.A. (1987) *Methods Enzymol.*, **155**, 335–350.
Okayama,H. and Berg,P. (1983) *Mol. Cell. Biol.*, **31**, 280–289.
Pollitt,R.J., Jenner,F.A. and Merskey,H. (1968) *Lancet*, **ii**, 253–255.
Sanger,F., Micklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Southern,E.M. (1975) *J. Mol. Biol.*, **98**, 503–517.
Tollersrud,O.K. and Aronson,N.N., Jr (1989) *Biochem. J.*, **260**, 101–108.
von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.

*Received on October 9, 1990; revised on October 29, 1990*