Correction. In the article "Molecular cloning and nucleotide sequence of human glucocerebrosidase cDNA" by Joseph Sorge, Carol West, Beryl Westwood, and Ernest Beutler, which appeared in number 21, November 1985, of Proc. Natl. Acad. Sci. USA (82, 7289–7293), the authors request that the following be noted. Because of typographical errors in the laboratory, one of two consecutive GAT sequences was omitted from the glucocerebrosidase sequence shown in Fig. 2. The relevant portion of the sequence should read

\[
\text{CTGATGATTT.}
\]

Nucleotide 1616 in the original sequences should have been T instead of C. The correct length of the coding region is increased to 1548 nucleotides, representing 516 amino acids with a molecular weight of 55,498 for the mature protein.
Molecular cloning and nucleotide sequence of human glucocerebrosidase cDNA

(.expression vector/genetic deficiency/A library/cDNA clone/chromosome localization)

JOSEPH SORGE, CAROL WEST, BERYL WESTWOOD, AND ERNEST BEUTLER

Department of Basic and Clinical Research, Scripps Clinic and Research Foundation, 10666 North Torrey Pines Road, La Jolla, CA 92037

Contributed by Ernest Beutler, July 15, 1985

ABSTRACT Mutations in the human glucocerebrosidase gene cause Gaucher disease. A cDNA clone containing the entire human glucocerebrosidase coding region from normal cells has been isolated using λgt11 expression libraries. The complete nucleotide sequence, a restriction map, and a hydrophathy profile are presented. Hybridization to chromosome-specific DNA localizes the human glucocerebrosidase gene to chromosome 1. The likely precursor protein is 515 amino acids long. The NH2-terminal 19 amino acids constitute a leader sequence that is cleaved from the mature protein. The predicted molecular weight of the mature protein is 55,384, without glycosylation or carboxyl-terminal processing.

Gaucher disease is caused by a hereditary deficiency in the enzyme glucocerebrosidase. It is an autosomal recessive disorder that is inherited in a Mendelian fashion. Affected individuals cannot adequately catabolize glucocerebrosidase, which then accumulates in macrophages. Cloning of the glucocerebrosidase gene has been reported (1); however, the restriction endonuclease map and nucleotide sequence were not disclosed. In an earlier publication, we presented the nucleotide sequence of a partial glucocerebrosidase cDNA clone that enabled us to study the genetics of Gaucher disease (2). We now report the cloning of the full-length cDNA for human glucocerebrosidase by using a randomly primed λgt11 cDNA expression library, show the complete nucleotide sequence of the cDNA and a restriction map, and demonstrate the localization of the cloned sequence to chromosome 1.

MATERIALS AND METHODS

Cells and RNA. WI-38 human fibroblasts were obtained from the American Type Culture Collection and grown in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum. Cellular RNA was prepared using the guanidinium thiocyanate/cesium chloride method (3). The poly(A)^+ fraction of RNA was prepared by two purifications with oligo(d-T)-Sepharose (4).

Synthesis of Randomly Primed cDNA. Five micrograms of poly(A)^+ RNA was dissolved in 1 mM methylmercuric hydroxide and incubated at room temperature for 5 min. 2-Mercaptoethanol was then added to 0.5%. The reaction was carried out in 0.4 ml of 50 mM Tris-HCl, pH 8.3/20 mM KCl/10 mM MgCl2/5 mM dithiothreitol/1 mM dATP/dCTP/dGTP/dTTP containing 15- to 25-nucleotide-long random calf thymus oligonucleotides at 1 mg/ml, placental ribonuclease inhibitor (Amersham) at 500 units/ml, [32P]dATP at 125 μCi/ml (600 μCi/mmol; 1 Ci = 37 GBq), and avian myeloblastosis virus reverse transcriptase at 800 units/ml.

After incubation for 2 hr at 37°C the randomly primed first strand was separated from the unincorporated nucleotides on a Sephadex G-50 column. The cDNA was made 0.3 M with NaOH and incubated at 68°C for 1 hr. The mixture was then neutralized with acetic acid and the cDNA was precipitated with 2.5 volumes of ethanol. The second strand of the cDNA was synthesized using reverse transcriptase followed by the large fragment of DNA polymerase as described (5). The hairpins were nicked using S1 nuclease (Sigma) at 100 units/ml and the ends were made blunt with the large fragment of DNA polymerase (5). The double-stranded cDNA was treated with EcoRI methylase (New England Biolabs) according to the recommended procedure. Prephosphorylated EcoRI oligonucleotide linkers were then ligated to the cDNA and digested with EcoRI (5). The cDNA was size fractionated on Sepharose 4B (Pharmacia).

λgt11 Library. One hundred nanograms of cDNA, >1000 base pairs long, was ligated to 1 μg of alkaline phosphatase-treated λgt11 EcoRI arms (Vector Cloning Systems, San Diego, CA). The DNA was packaged in vitro using a high-efficiency packaging extract (Vector Cloning Systems) and the recombinant phage were plated on strain Y1088 of Escherichia coli (6). Approximately 3 × 10^6 plaques were obtained with >90% containing cDNA inserts. An amplified phage stock was made from these plates and was used for subsequent screening.

Antibody Purification. Rabbit antiserum raised against purified human glucocerebrosidase was purified using two methods. In the first method, an ammonium sulfate fraction was allowed to react with an affinity column to which purified glucocerebrosidase had been bound using cyanogen bromide. The bound antibodies were then removed by elution with 0.1 M glycine, pH 2.6. In the second method, the affinity column was prepared by first attaching a very-high-affinity anti-glucocerebrosidase monoclonal antibody to Sepharose with cyanogen bromide. Then, purified glucocerebrosidase was bound to the monoclonal antibodies attached to the column resin, and the column was washed extensively with 0.1 M glycine, pH 2.6. Most of the glucocerebrosidase protein remained bound to the column, as shown by measuring enzyme activity in the resin. A rabbit anti-glucocerebrosidase serum was then allowed to react with the column, the resin was washed extensively at neutral pH, and then the bound rabbit antibodies were removed by elution with 0.1 M glycine, pH 2.6. This purified serum was found to be highly specific for human glucocerebrosidase on immunoblots of human fibroblast proteins.

Library Screening. Approximately 1 × 10^6 recombinant phage from the amplified library were plated on E. coli strain Y1090 (6) at a density of 5 × 10^4 phage per 150-mm plate (5). After 3 hr of incubation at 37°C, the bacterial lawns were covered with a 137-mm nitrocellulose filter (HATF 137, Millipore) that had been soaked in 5 mM isopropyl

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: bp, base pair(s).
thiogalactoside (IPTG) and dried. Incubation at 37°C was continued for 5 hr, and then the filter was removed, a second IPTG-treated filter was placed on the bacterial lawn, and the plate was incubated overnight at 37°C. The filters were dried at room temperature and then soaked first in phosphate-buffered saline (P/NaCl) for 10 min and then in P/NaCl containing 3% bovine serum albumin at 37°C. The filters were washed at room temperature for 2 hr with P/NaCl containing 3% bovine serum albumin and affinity-purified anti-glucocerebrosidase antibody at a dilution of 1:1000 at room temperature for 2 hr. The filters were washed at room temperature once with P/NaCl for 10 min and once with P/NaCl containing 0.1% Triton X-100 for 10 min and then with P/NaCl for 10 min. They were then incubated for 1 hr with 125I-labeled protein A (specific activity, 20 mCi/mg) in P/NaCl (5 × 105 cpm per filter), washed with P/NaCl and with P/NaCl containing Triton as above, and air dried; and XAR-2 x-ray film was exposed to the filters for 1–2 days at −80°C using an intensifying screen. Positive phage were picked and repeatedly rescreened at lower density until a pure population was obtained.

**DNA Subcloning and Sequencing.** The clone G5A-1Y was shown to contain an 1100-base-pair (bp) EcoRI insert that was subcloned into the EcoRI site of plasmid pBR322 (5). This plasmid, G5A-1Y, was then nick-translated and used as a probe to screen an oligo(dT)-primed human WI-38 λgt11 cDNA library prepared as above, except that the first-strand cDNA was primed with oligo(dT) instead of with random calf thymus primers. Positive clones were identified at a frequency of 1 in 600,000. Several plaques were purified and one clone, ID9-bb, was found to have a 1500-bp insert. This insert was subcloned into plasmid pBR322, yielding plasmid ID9-bb. Fig. 1 illustrates the portions of the glucocerebrosidase cDNA contained within plasmids G5A-1Y and ID9-bb. A full-length cDNA clone was constructed by taking advantage of the unique Sca I site in the region where G5A-1Y and ID9-bb overlap. The plasmid containing the full-length clone is called GCS-2kb.

The full-length clone was sequenced by the method of Maxam and Gilbert (7). The complete DNA sequence was analyzed with computer programs from the University of Wisconsin. A hydrophobicity–hydrophilicity profile was generated using the method of Kyte and Doolittle (8).

**Chromosome Localization.** The nick-translated 1100-bp G5A-1Y probe was hybridized to chromosome-specific DNA attached to nitrocellulose filters according to published procedures (9, 10).

**Protein Sequencing Methods.** Glucocerebrosidase was purified from human placenta by the method of Dale et al. (11) and further purified by immunoaffinity purification over an anti-glucocerebrosidase monoclonal antibody column. Specifically, 4 mg of an anti-glucocerebrosidase mouse monoclonal antibody (16-B-3) was coupled to CNBr-activated agarose. The partially purified glucocerebrosidase was applied to the column in 50 mM NaCl/50 mM citrate, pH 6, and the column was washed extensively with this starting buffer. Glucocerebrosidase was removed by elution with 0.2 M glycine, pH 2.6. The eluted protein was analyzed by NaDodSO4/PAGE and showed a single band.

The NH2-terminal sequence was determined by solid-phase Edman degradation with HPLC analysis of the phenylthiohydantoin-derivatized amino acid cleavage products.

**RESULTS**

**Library Screening.** Approximately 1 × 106 recombinant phage were screened with an affinity-purified anti-glucocerebrosidase antibody and a single, strongly positive phage was obtained. The same phage was positive with a second, highly purified serum. The phage DNA contained an 1100-bp EcoRI insert. When the same randomly primed library was screened using this 1100-bp DNA insert as a probe, 6 positives were found in approximately 1 × 106 phage. This suggested an mRNA frequency of 0.0006% or roughly 1 or 2 mRNA molecules per cell. When the same library was screened with a human β-actin DNA probe (12), ≈0.7% of the phage were positive.

**Restriction Map and Sequence Analysis.** The composite restriction map of human glucocerebrosidase cDNA is shown in Fig. 1 and the complete nucleotide sequence of the coding region is shown in Fig. 2. Since the insert of the 5′ clone, G5A-1Y, had been fused in frame to bacterial β-galactosidase in λgt11, it was necessary for an open reading frame to extend from the extreme left end of G5A-1Y. As shown in Fig. 2, an open reading frame extends from nucleotide 1 to nucleotide 1700.

A hydrophobicity profile for the protein sequence predicted from the full-length cDNA sequence is shown in Fig. 3.

**Chromosome Mapping.** The 1100-bp G5A-1Y DNA insert was nick-translated and hybridized with chromosome-specific human DNA (Fig. 4). Hybridization was specific for chromosome 1, which has been shown to contain the gene for human glucocerebrosidase (14, 15).

![Fig. 1. Restriction map of the GCS-2kb insert. ATG represents the probable initiator methionine codon and UGA represents the translational terminator codon. The boundaries of the partial cDNA clones G5A-1Y and ID9-bb are indicated with horizontal bars. We have previously used clone G5A-1Y to perform restriction polymorphism analysis (2). Enzymes that do not digest the GCS-2kb insert include Xho I, Xba I, Cia I, EcoRI, EcoRV, Sma I, and Bgl II.](image)
Confirmation with the Protein Sequence. The sequence of the amino terminus of the glucocerebrosidase protein is as follows:

Xaa-Xaa-Pro-Xaa-Ile-Pro-Xaa-Xaa-Phe.

Although there are several unknown residues, all confirmed residues match the amino acids predicted from the nucleotide sequence. The probability of this occurring by chance is 1:160,000.

DISCUSSION

We have cloned and sequenced 2227 nucleotides of cDNA containing the entire coding region of the human glucocerebroside.
The glucocerebrosidase gene. We initially attempted to clone the gene by using an affinity-purified antisera and an oligo(dT)-primed λgt11 cDNA library, but we were unsuccessful, presumably because the cDNA was not long enough to contain sufficient protein coding region to bind antibodies. We circumvented this problem by constructing a λgt11 cDNA library with random primers to initiate first-strand cDNA synthesis. Since the random primers initiate cDNA synthesis at any point along the mRNA molecule, the random library does not have the "3' bias" that conventional oligo(dT)-primed libraries have. With the randomly primed library, we were able to identify a positive clone on the first screening of 1 × 10^6 plaques. Our first clone, G5A-1Y, extended from midcoding region to a point near the 5' end of the mRNA molecule.

Although the left end of the open reading frame from nucleotides 1–1700 could be part of the coding sequence of the glucocerebrosidase molecule, the sequence data (see below) suggest that the ATG at nucleotides 154–156 is the initiator ATG for glucocerebrosidase. It is interesting, however, that the 5' untranslated region of the glucocerebrosidase mRNA contains such a long open reading frame that is in-frame with the coding sequence for the protein. We have not yet mapped the 5' end of the mRNA molecule. If another ATG lies upstream of what we refer to as nucleotide 1, it is possible that a larger protein precursor could be synthesized as well (at least 6000 daltons larger). The ATG at nucleotides 94–96 is also in the same open reading frame. However, the sequences surrounding this ATG do not match Kozak's consensus sequence for initiator ATGs (16). The sequences surrounding the ATG at nucleotides 154–156 match Kozak's consensus sequence quite well.

If the ATG at positions 154–156 is the initiator codon, the progglucocerebrosidase protein is 515 amino acids long. The protein sequence data suggest that the first 19 amino acids are removed from the mature protein. This conclusion has also been reached by Tsuji et al. (17). Such cleavage of a leader sequence is common for secreted proteins; glucocerebrosidase is secreted into lysosomes. Thus the mature glucocerebrosidase protein would contain 496 amino acids. The leader sequence (nucleotides 154–210) is the most hydrophobic portion of the protein, followed closely in hydrophobicity by the carboxyl terminus, where the active site has been shown to be located (G. A. Grabowski, personal communication). The protein would have an unglycosylated molecular weight of 55,384. We do not yet know whether the carboxyl terminus identified in the nucleotide sequence is present on the mature protein or whether it also is cleaved.

The fact that the gene we have cloned is glucocerebrosidase is established by (i) amino-terminal protein sequence data, (ii) chromosome localization, (iii) antibody reactivity, and (iv) the fact that amino acids 362–376 perfectly match a previously published glucocerebrosidase oligopeptide sequence (1).

The isolation and characterization of a full-length cDNA clone makes it possible to attempt expression experiments in human cells with the hope that these studies may eventually lead us to an effective treatment of Gaucher disease through gene therapy.

**Note Added in Proof.** We have now found that our cDNA clone directs the synthesis of functional glucocerebrosidase when inserted into mammalian cells.

We thank Y. W. Kan for his efforts in performing the chromosome-specific hybridizations and L. Kedes and P. Gunning for their human actin gene clone. This work was supported in part by Grant CA36448 from the National Cancer Institute and Grant 1-955 from the March of Dimes Birth Defects Foundation. This is publication no. 3874BCR from the Research Institute of Scripps Clinic.