A mercury-thiol affinity system for rapid generation of overlapping labeled DNA fragments for DNA sequencing

---

James L.Hartley^{*}, Kenneth K.Chen and John E.Donelson

Department of Biochemistry and Diabetes and Endocrinology Research Core Center, University of Iowa, Iowa City, IA 52242, USA

Received 8 February 1982; Revised 17 May 1982; Accepted 25 May 1982

ABSTRACT

We describe an *in vitro* protocol for quickly generating overlapping terminal-labeled restriction fragments for DNA sequence analysis via the Maxam-Gilbert technique. The protocol involves introducing mercurated nucleotides into one end of a region to be sequenced, partial digestion with several restriction enzymes and terminal-labeling, separation of the mercurated restriction fragments from non-mercurated ones on a thiol column and resolution of the different mercurated fragments on one preparative agarose gel. The protocol was used to determine the nucleotide sequence of a 980 base pair cDNA that contains the coding region for a variable surface glycoprotein of *Trypanosoma brucei*. It could just as quickly and easily be used to obtain many terminal-labeled overlapping restriction fragments covering a region of several kilobases.

INTRODUCTION

Rapid DNA sequencing has played a pivotal role in recent molecular biology research. The chemical modification method of Maxam and Gilbert (1) and the chain termination procedures of Sanger (2,3) are quick and reliable, to the point that the generation of suitable DNA fragments, rather than the sequencing steps themselves, has become the principle factor limiting sequencing speed.

"Shotgun" sequencing strategies (1,2,4-6) yield large amounts of data, but typically require increasing effort as the number of bases needed to complete the remaining gaps decreases. Thus the time needed to obtain the final 10% of a sequence may be equal to that needed for the initial 90%. Subcloning strategies have been described (7,8) which alleviate some of this difficulty. However such methods entail cloning steps, plasmid purifications, and fragment isolations that are themselves time consuming.

We report here an *in vitro* protocol which quickly generates overlapping labeled DNAs for Maxam-Gilbert sequencing. The method includes mercuration of plasmid DNA, thiol agarose affinity purification, and end labeling by DNA polymerase repair. One preparative gel electrophoresis step yields many fragments which are immediately ready for the sequencing reactions, without strand separa-

tion or secondary cutting. Furthermore, the map positions of all the labeled ends are determined by inspection of the relative migrations of the labeled fragments on the preparative gel. We have used the procedure to determine a 980 base pair cDNA sequence cloned at the Pst I site of plasmid pBR322. In practice the procedure could just as easily be used to determine sequences of several kilobases in length cloned at either this Pst I site or at other plasmid restriction sites.

MATERIALS AND METHODS

Materials

Plasmid pcB-1.1 is plasmid pBR322 which has at its Pst I site a cDNA insert that contains the coding sequence for the variable surface glycoprotein of *Trypanosoma brucei* clone 1ltat 1.1. It was maintained in *E. coli* HB 101 (9) and plasmid DNA was prepared by cesium chloride/ethidium bromide gradient centrifugation (10). The construction, identification and preliminary characterization of pcB-1.1 is described in reference 11. Restriction enzymes were from Bethesda Research Laboratories (BRL) or New England BioLabs. T4 DNA polymerase was from BRL, and *E. coli* DNA polymerase I was from New England BioLabs. The 5-mercuri-2'-deoxycytidine 5'-triphosphate (Hg-dCTP), 5-mercuri-2'-deoxyuridine 5'-triphosphate (Hg-dUTP), and thiol agarose ("AgThiol") were obtained from P-L Biochemicals. [α - ^{32} P]dCTP and [α - ^{32} P]dATP, 2-3000 Ci/mmol, were from Amersham. Other deoxynucleotide triphosphates were from Sigma.

Methods

Trichloroacetic acid (TCA) precipitations. To aliquots of DNA containing ^{32} P incorporated in vitro was added TCA stop mix (0.1 M sodium pyrophosphate, 50 mM EDTA, 100 μ g/ml salmon sperm DNA) to a final volume of 505 μ l. Five μ l were added to scintillation cocktail for direct counting, and 2 ml cold 10% TCA were added to the remaining 500 μ l. After mixing and incubating on ice 5 min, the mixture was filtered (Whatman GF/A) and counted in scintillation cocktail. The proportion of incorporated ^{32} P was calculated from the direct and TCA-precipitable counts.

Mercuriation of DNAs. Linear pcB-1.1 DNA molecules were made by digesting 300 μ g (86 pmoles) of intact plasmid with either EcoR I or Ava I. After ethanol precipitation, DNAs were redissolved in 67 mM potassium acetate, 33 mM Tris acetate pH 7.8, 10 mM magnesium acetate, 10 mM β -mercaptoethanol, 100 μ g/ml bovine serum albumin to a final concentration of 500 μ g DNA/ml. The reducing agent is required for incorporation of mercurated nucleotides (12). T4 DNA polymerase (0.2 units/ μ g DNA) was added and incubated (without added triphos-

phates) for 60 min at 37°, then transferred to an ice bath. Stock solutions (10 mM) of dGTP, Hg-dCTP, Hg-dUTP, and [α -³²P]dATP (0.1 Ci/mmol) were added to final concentrations of 100 μ M each, and the tube was incubated another 10 min at 37°. Tubes were placed in ice and aliquots were precipitated with trichloroacetic acid to determine the extent of mercuration. The reaction mixtures were extracted with phenol and applied to small (bed volumes of 10-20 times the reaction volumes) columns of BioGel A15M (BioRad) equilibrated and eluted in 10 mM Tris-HCl pH 7.4, 150 mM NaCl, 10 mM MgCl₂. Void volume fractions were located by incorporated ³²P and pooled. Flushness of the mercurated ends was tested by adding 1 μ l (10 μ Ci) each of aqueous [α -³²P]dATP and [α -³²P]dCTP (2-3000 Ci/mmol) to a 25 μ l aliquot of the pooled void volume fractions, adding 5 units of *E. coli* DNA polymerase I, and incubating on ice for 5 min. The mercurated DNA was judged acceptable if less than 5% of the ends were labeled with one nucleotide under the assay conditions.

Secondary cut and partial restriction enzyme digestions. All the steps described here prior to the affinity chromatography were conducted in the high salt buffer used for the BioGel A15M gel filtration, i.e., 10 mM Tris HCl pH 7.4, 150 mM NaCl, 10 mM MgCl₂. The mercurated DNAs were cut to completion with BamH I at 37°. Aliquots of each DNA were removed and partially digested with Msp I, Taq I, Hinf I, or Hind III, the amount of DNA and the degree of cutting desired being determined by the number of sites of interest for each enzyme (see Results). The extent of digestion was judged by analytical 0.8% agarose gels, or by a quantitative assay as follows. A small aliquot of DNA was cut to completion with each enzyme to be used. To 50 μ l of freshly prepared partial digest assay mix (2 μ M each of dGTP and dTTP, containing 2.5 μ Ci (100 pmol) each of [α -³²P]dATP and [α -³²P]dCTP, and 5 units *E. coli* DNA polymerase I in 70 mM NaCl, 70 mM Tris HCl pH 7.4, 10 mM MgCl₂) were added 5 μ l of either a partial digest or a complete digest containing about 1 μ g DNA. After incubating at 4° for 10 min, the TCA precipitable counts were compared for the partial and complete digests. Taq I cut unexpectedly well at 65° under the high salt conditions. Digestions were kept on ice during the assay or gel, and re-incubated as necessary.

Labeling and affinity chromatography. The [α -³²P]dATP or [α -³²P]dCTP, 2-3000 Ci/mmol, was added to appropriate partial digests at a concentration of 0.5-2 pmol label per pmol of "labelable" termini, and DNA polymerase I was added to 100 units/ml. After 10 min incubation on ice, a mixture of dGTP, dCTP, dATP, and dTTP was added to a final concentration of 25 μ M each and incubated 5 min on ice to eliminate possible heterogeneity of label incorporation. One-tenth

volume of 0.5 M EDTA, pH 8.5, was added to stop the reactions.

The thiol agarose was prepared for affinity chromatography by stirring 1 ml aliquots twice for 10 min in 10 ml DTT buffer (0.1 M Tris HCl pH 7.4, 0.3 M NaCl, 1 mM EDTA, 20 mM dithiothreitol, freshly prepared), collected by centrifugation, and washed twice in T100N buffer (10 mM Tris HCl pH 7.4, 100 mM NaCl). The centrifuged pellets were suspended in 0.5 ml T100N and transferred to plugged 1 ml (blue) disposable plastic pipet tips to make columns of about 0.7 ml bed volume. Gel beds were washed with at least 5 bed volumes of T100N. Labeling mixtures were applied directly to the thiol agarose columns, and the non-adsorbed material was reapplied to the columns two more times. After washing with 5 bed volumes T100N, adsorbed DNAs were eluted with T100N containing 0.1 M β -mercaptoethanol. Fractions (100 μ l) were counted without addition of scintillation fluid by Cerenkov radiation, and peak fractions were pooled and applied directly to a preparative 1.2% agarose gel. Because earlier fractions contained the larger adsorbed DNAs, these were selected over later fractions. The preparative gel was run overnight at 4°. Bands of interest were electroeluted (13), and the labeled DNA was concentrated by adsorbing to small (200 μ l bed vol) columns of DEAE cellulose (Whatman DE52) equilibrated in 10 mM Tris HCl pH 7.4, 1 mM EDTA, then eluting with 0.5 M Tris base, 1.5 M sodium acetate.

Sequencing reactions. Sequence determinations of the isolated mercurated, ^{32} P-labeled fragments were carried out by the Maxam and Gilbert procedure (1) using the A>G, G, C and C+T reactions. The cleavage products were fractionated on thin 6% polyacrylamide gels (0.4 mm x 25 cm x 87 cm) containing 7 M urea. Three applications of each set of cleavage products, spaced about 10 hours apart were made. The first 100 nucleotides could be read from the last application, the next 100 from the second and another 100 from the first for a total of about 300 nucleotides.

RESULTS

The general strategy employed is outlined in Fig. 1. A plasmid is opened with a restriction enzyme near the region to be sequenced (a cloned cDNA insert in this case) and about two hundred nucleotides are removed from the 3' ends by T4 DNA polymerase. When mercurated and non-mercurated deoxynucleoside triphosphates are added, the 5'→3' polymerase activity of the enzyme overwhelms the 3'→5' exonuclease function (14), resulting in repaired, flush ends containing mercury atoms in the major grooves (15). The extent of mercuration is thus a function of the extent of the 3'→5' exonuclease digestion.

After inactivation of the polymerase by phenol extraction and removal of

Figure 1. Schematic representation of the protocol for generating terminal labeled overlapping DNA fragments for Maxam-Gilbert sequence analysis. In general the region to be sequenced is in a plasmid (shown as a double circle). First the plasmid is linearized with restriction enzyme A whose recognition site occurs only once in the plasmid. Then about 200 nucleotides are removed from each 3'-end using the 3' exonuclease activity of T4 DNA polymerase. The four deoxynucleoside triphosphates (dNTPs), two of which are mercurated, are added to the reaction and the polymerase activity of T4 DNA polymerase repairs the 3' ends. One mercurated end is removed with a second restriction enzyme and partial digestions with restriction enzyme C (and additional restriction enzymes if necessary) are conducted. After terminal labeling with $[\alpha\text{-}^{32}\text{P}]$ dNTPs using *E. coli* DNA polymerase I, the fragments are passed through a thiol agarose column. The bound mercurated fragments are eluted with β -mercaptoethanol and applied immediately to a preparative agarose gel for separation on the basis of size. Fragments of interest are eluted and subjected to the cleavage reactions for sequence determination.

unincorporated triphosphates by gel filtration, the end of the DNA furthest from the insert is cut off with a second restriction enzyme. This prevents a ladder

of partials from arising from this end, and simplifies the preparative agarose gel (below). The mixture of the two fragments is then divided into aliquots and subjected to partial digestions by restriction enzymes which produce protruding 5' ends. These ends are labeled with DNA polymerase I and [α - ^{32} P] deoxynucleoside triphosphates, under conditions in which the mercurated, flush ends are not affected. The second restriction enzyme cut, the partial digestions, and the labeling are all done in the same buffer if possible, and products of these reactions are then applied directly to small columns of thiol agarose (one for each restriction enzyme used in the partial digests). The unmercurated partial digestion products are washed off the column and the bound fragments are eluted with β -mercaptoethanol and applied immediately to a preparative gel. Fragments of interest are eluted for sequencing.

The method was applied to a plasmid pcB-1.1 which contains a Trypanosoma brucei variable surface glycoprotein cDNA of about 980 base pairs cloned at the Pst I site of pBR322 (Fig. 2). Aliquots of the plasmid were opened with either EcoR I or Ava I and treated with T4 DNA polymerase (0.2 units/ μ g DNA for 60 min at 37°). Upon addition of [α - ^{32} P]dATP (0.1 Ci/mmol), dGTP, Hg-dCTP and Hg-dUTP, about 110 mercury atoms were incorporated into the EcoR I ends, and about 128 into the Ava I ends as determined by monitoring the amount of incorporated [^{32} P]dAMP and assuming an equal distribution of the four nucleotides in the sequence. The rate of incorporation of mercurated deoxynucleotides was very similar to that of non-mercurated deoxynucleotides (data not shown), in the presence of reducing agent (12). After phenol extraction and desalting on a small column of BioGel A15M, each DNA preparation was checked to verify that the T4 DNA polymerase had extended the 3'-ends all the way back to generate linear double-stranded DNA molecules with flush double stranded ends. This was tested by incubating an aliquot of the DNA with [α - ^{32} P]dATP, [α - ^{32} P]dCTP (2-3000 Ci/mmol each) and E. coli DNA polymerase I and measuring the incorporation of these high specific activity radionucleotides. If less than 5% of the ends were labeled with a radioactive nucleotide, then the procedure was continued. In practice these small scale assays did not reveal difficulties at this step. Both DNA preparations were then cut to completion with Bam HI.

Taq I, Hinf I, Msp I, and Hind III were chosen for the partial restriction cuts on the basis of preliminary mapping of the cDNA insert. Five to 10 pmol of plasmid were cut for each restriction site between the mercurated end and the furthest site of interest. For example, four Msp I fragments were sought from the EcoR I-opened plasmid, extending counter-clockwise from the mercurated EcoR I end to the Msp I sites at map positions 4880, 4638 and 3548 in the vector por-

Figure 2. The strategy used to sequence the cDNA insert of plasmid pcB-1.1 which was cloned at the Pst I site of plasmid pBR322 using the oligo dG: oligo dC tailing procedure (11). The plasmid was linearized at either the EcoR I or Ava I site in pBR322 (the site for enzyme A in Figure 1), mercurated nucleotides were introduced at these ends as shown by the symbol Hg, and Bam HI was used for the secondary cut (enzyme B of Figure 1). The other restriction enzymes (enzyme C of Figure 1) shown in the figure were used to obtain partial restriction fragments for sequence analysis. Numbers indicate the locations of the restriction sites in base pairs using the single EcoR I site as the reference position 1. Arrows indicate the locations and lengths of the sequenced regions on both strands. G₃₀ and C₂₅ show the locations and sizes of the two oligo dG:oligo dC boundaries. A₁₆ shows the location of a 16 nucleotide oligo dA region that corresponds to the poly rA at the 3' end of the mRNA from which this cDNA was synthesized.

tion (equivalent to Msp I sites at positions 3900, 3658 and 3548 in pBR322), and at position 4229 in the cDNA insert (see Fig. 2). As a consequence, 40 pmol of the EcoR I-cut, mercurated plasmid were cut to 25% completion with Msp I, as judged by gel electrophoresis, or by a DNA polymerase I labeling assay (see Materials and Methods). Other digestions of the EcoR I-cut, mercurated plasmid were (i) a 50% Taq I digestion of 10 pmol DNA, (ii) a total Hind III digestion of 5 pmol DNA and (iii) a total Hinf I digestion of 5 pmol DNA. For the Ava I-cut, mercurated plasmid the digestions were (i) a 50% Taq I cut of 20 pmol, (ii) a 14% Hinf I cut of 70 pmol, and (iii) a 100% Hind III cut of 5 pmol. Ends were labeled with high specific activity [α -³²P]dCTP or [α -³²P]dATP (2-3000 Ci/mmol) and *E. coli* DNA polymerase I at low temperature (ice bath) and high salt (150 mM NaCl). After addition of EDTA to about 50 mM, labeling mixtures were passed

three times through small (0.7 ml bed volume) columns of thiol agarose. The columns were washed with five bed volumes of buffer, then eluted with buffer containing 0.1 M β -mercaptoethanol. About 90-95% of the radioactivity remaining after the washing was eluted with β -mercaptoethanol, and 85-95% of the eluted counts were collected in three separate 100 μ l fractions.

The analytical gel shown in Fig. 3 shows the partial digestion products applied to the column (lane 1), the fragments that bound (lane 2) and the fragments that did not bind (lane 3) to the thiol column for three of the mercurated partially-digested DNAs. The non-mercurated fragments produced by the partial restriction enzyme digestion did not adsorb to the thiol agarose at all, while the adsorbed mercurated fragments bound to varying extents and were eluted with β -mercaptoethanol. The bands in Fig. 3 vary in intensity due to the different sensitivities of the restriction sites during the partial digestion. The adsorbed bands were precisely those predicted from the maps of the cDNA insert and the vector. The mercury substituents of the DNAs did not alter the appearance of the bands. Nearly 100% of the mercurated DNAs smaller than 1 kb adsorbed to the thiol column, while 30-50% of the DNAs of 3-4 kb were bound. The use of an affinity medium with more thiol groups, such as a recently reported sulhydryl-cellulose (16) may improve the binding of the large fragments but in practice, the binding of only 30-50% of the larger fragments was not a problem. The peak fractions of bound fragments were collected from the thiol columns and immediately applied to horizontal preparative 1.2% agarose gels run at 4°. Bands of interest were cut out, electroeluted, and concentrated by adsorbing and eluting from DEAE cellulose (200 μ l bed volume) columns. Eluted DNAs were precipitated with ethanol and used directly for Maxam/Gilbert sequencing reactions.

Figure 4 shows the preparative gel of the same three mercurated partially-digested DNAs discussed in Fig. 3. The bands indicated by arrows are the same fragments indicated by arrows in Fig. 3 and were eluted for sequence determination. Autoradiographs for two of the eight fragments sequenced (see Fig. 2) are shown in Fig. 5. The presence of the mercury in the DNA did not have a noticeable effect on any of the sequencing manipulations nor on the sequencing gel itself. The coding strand sequence of the entire cDNA insert is shown in Fig. 6. Fragments derived from the EcoRI-opened pcB-1.1 plasmid yielded sequences from the coding strand of the cDNA insert, while the non-coding strand portions were obtained from plasmid mercurated at the Ava I ends (the inside and outside strands respectively as shown in Figure 2).

Figure 3. An analytical 0.85% agarose gel stained with ethidium bromide showing aliquots of the mercurated, partially restricted, terminal-labeled plasmid DNA that was applied to the thiol agarose column (lanes 1), the partial restriction fragments that bound to the column (lanes 2) and the partial restriction fragments that passed through the column (lanes 3). The first three lanes show the fragments obtained from the plasmid that was mercurated at the *EcoR* I site and partially digested with *Msp* I; the middle three lanes and last three lanes show the fragments obtained from plasmid mercurated at the *Ava* I site and partially digested with *Hinf* I or *Taq* I respectively. Arrows show fragments that were eluted from a preparative agarose gel (Figure 4) for DNA sequence determination. Bands on the far right are standard length markers.

DISCUSSION

The objective of this work was to develop a method for rapidly generating a

Figure 4. Ethidium bromide stain of a preparative 0.85% agarose gel of three sets of mercurated, labeled, partial restriction fragments that bound to the thiol agarose column (see the three lanes 2 in Figure 3). Lanes 1-4 in this figure are multiple applications of the same three samples described in Figure 3. Arrows show the fragments that were eluted for sequencing and point to the same fragments as do the arrows in Figure 3. Bands on the far right are standard length DNA fragments.

series of overlapping labeled DNA fragments suitable for Maxam-Gilbert sequencing. Since we can reliably determine sequences of about 300 nucleotides on our Maxam and Gilbert sequencing gels, we tried to introduce the radioactive label at sites that were located 100 to 300 base pairs apart as shown by the arrows in

Figure 5. Autoradiograms of DNA sequencing gels of two fragments used to obtain the sequence shown in Figure 6. Three applications of the four cleavage reactions were made, spaced about 10 hours apart. From left to right on the gel, the cleavage reactions are specific for A>G, G, C+T, and C. (A) Sequencing gel of a fragment mercurated at the *Ava* I site and labeled at the *Taq* I site (position 3912 in Figure 2 or position 679 in Figure 6). The sequence that can be deduced from the gel is of the outside (non-coding) strand shown in Figure 2. Note in the right-hand lanes the 16 T residues (complementary to the poly rA of the mRNA) followed by the 25 residue dG boundary of the cDNA. These sequences occur about 300 nucleotides from the labeled *Taq* I site. (B) Sequencing gel of a fragment mercurated at the *Eco*RI site and labeled at the same *Taq* I site. This time the sequence that can be read from the gel goes the other direction from the *Taq* I site and is of the inside (coding) strand shown in Figure 2.

```

    val ala cys lys asn gly gly gly ala cys ser ala ala ser ser ser asp lys ile his ile thr ile
30(G)A GTT GCC TGC AAA AAC GGC GGC GGC GCA TGC AGT GCA GCC AGC AGC AGC GAC AAA ATA CAT ATA ACG ATC
    100

    glu thr asp ser lys asn lys gly thr ala ala ser thr met asp ser gln asn thr ala val ala phe ala
    GAA ACG GAC AGC AAA AAC AAA GGG ACA GCG GCC AGC AGC ATG GAC TCC CAA AAC ACA GCG GTC GCA TTC GCA

    thr glu leu gln ile gly ser trp lys ser gln his ile asp ser asn ile thr ala leu ala asn ala leu
    ACG GAG CCG CAG ATT GGA AGC TGG AAA AGC CAG CAC ATA GAC AGC AAC ATA ACG GCT TTG GCG AAT GCC TTA
    244

    thr ala leu asp ser ile pro asp leu thr asp pro thr ala tyr thr ala asp ala ala phe lys gln leu
    ACA GCA CTA GAC AGC ATT CCT GAC CTC ACA GAC CCA ACG GCT TAC ACA GCA GAC GCG GCG TTC AAA CAA CTA

    val ala thr val ala leu asn lys pro pro thr thr glu leu thr gly glu val leu asp ala val asn arg
    GTA GCC ACA GTA GCG CTC AAC AAG CCG CCA ACC ACG GAG CTA ACC GGA GAA GTT CTA GAC GCA GTA AAC AGA
    388

    ala cys ala asp asn tyr gly thr ser ala ser glu leu thr thr lys ile trp glu pro leu asn glu gln
    GCA TGT GCC GAC AAC TAC GGA ACC TCG GCA AGC GAG CTG ACG AC AAA ATT TGG GAA CCG TTA AAT GAA CAG
    ↑
    thr ala ser tyr tyr ser glu lys thr ile lys thr asp gln leu lys leu leu thr ser asn gln gln leu
    ACA GCC AGC TAC TAC AGC GAA AAA ACT ATC AAA ACC GAC CAA CTA AAG CTA CTA ACA AGC AAC CAA CAG CTG
    532

    thr thr ala leu gly val ala leu ala lys ala ile asn val lys glu ala ser lys lys glu cys asn leu
    ACA ACT GCA CTA GGA GTA GCA CTG GCG AAA GCA ATA AAC GTA AAA GAA GCT TCC AAA AAG GAA TGC AAC CTA

    his gly his glu thr asp ala thr cys glu ala lys gly val gly asp asn cys lys pro pro cys lys glu
    CAC GGC CAT GAA ACA GAC GCT ACG TGC GAG GCA AAA GGA GTA GGA GAT AAC TGT AAA CCC CCA TGC AAA GAA
    676

    val glu glu gly gly lys lys lys cys lys leu asp lys glu glu ala lys arg val ala glu gln ala ala
    GTC GAA GAA GGT GGA AAA AAG AAG TGC AAA TTG GAT AAA GAG GAG GCC AAA AGG GTA GCA GAA CAA GCA GCA

    thr asn gln glu thr glu gly lys asp gly lys thr thr asn thr thr gly ser asn ser phe ala ile his
    ACA AAT CAA GAA ACA GAA GGC AAA GAT GGA AAA ACC ACA AAC ACC ACA GGA AGG AAT TCT TTT GTC ATT CAT
    820

    lys ala gln leu phe leu ala phe leu leu phe ***
    AAG GCA CCA CTT TTT CTT GCG TTT TTG CTT TTT TAA TTTTCCCTTTAAAAATTTCTCCTTTAAAAACAAATTTTTTCTACTT
    904

    GAAAACTCTGATATTGATATATTTTAAACACCTTTAAAAA(A)25

```

Figure 6. Sequence of the coding strand of the cDNA insert of plasmid pcB-1.1 and the deduced amino acid sequence. This cDNA was shown previously by hybrid-arrested translation *in vitro* (11) to contain part of the coding sequence for the variable surface glycoprotein of *T. brucei* clone ILTAT 1.1. The arrow points to a single base pair deletion in the cDNA thought to have occurred either during the cDNA construction *in vitro* or during its subsequent cloning in *E. coli* as discussed in the text. The symbol *** indicates the termination codon. This cDNA sequence is compared with other surface glycoprotein cDNA sequences in reference 17.

Figure 2. The nearly complete sequence of the coding (inside) strand of the 980 base pair cDNA insert was very quickly obtained from four overlapping sequences and a fifth sequence which was about forty nucleotides short of overlapping one

of the others. These five sequences were obtained from the mercurated EcoR I-cut DNA partially digested with the enzymes shown in Figure 2. The sequence of about three-fourths of the other (non-coding) strand was obtained using three fragments generated from mercurated Ava I-cut DNA. This non-coding (outside) sequence information covered the forty nucleotide gap in the coding (inside) strand and also clarified a region of about five nucleotides in the coding strand that was ambiguous due to a compression in the sequencing gel (nucleotides 342-346 in Figure 6).

The sequence of one strand of pcB-1.1 is shown in Fig. 6. Initially it was clear that this was the coding strand because it contains 16 deoxyadenosine residues at one end which almost surely corresponds to the poly A at the 3' end of the mRNA from which the cDNA was originally synthesized (11). In addition the 84 non-translated nucleotides between the (DNA) termination codon, TAA, and the poly A resembles the 3' non-translated region of mRNAs for other variable surface glycoproteins (17). However, there was no open translation reading frame preceding this termination codon nor was one present in the other 2 reading frames of this strand. This confusing situation was clarified when portions of two other cloned cDNA molecules containing the same coding sequence were determined and the complete sequence of the corresponding chromosomal gene was obtained (20). The other two cDNAs and the genomic DNA sequence all contain five adjacent deoxyadenosines at the location indicated by the arrow in Fig. 6 instead of the four deoxyadenosines (and four thymidines in the complementary strand) that are present in pcB-1.1. In addition a Hind III-Xba I fragment of pcB-1.1 containing this region was isolated without mercuration and sequenced to verify that this difference was not generated in vitro by the mercury (not shown). Thus it appears that pcB-1.1 contains a single base pair deletion that destroys the translation reading frame. Errors have been reported previously in other cloned cDNA molecules (21,22) so it is likely that this deletion is the result of a mistake by enzymes used either (i) during the construction of the cDNA in vitro or (ii) during the subsequent transformation and replication of the recombinant plasmid in E. coli. Another less likely possibility is that the deletion occurred in an mRNA molecule during its transcription in vivo. If nothing else, this observation demonstrates the hazard of relying on the sequence of a single cloned cDNA molecule to predict an mRNA and/or gene sequence. A comparison of the sequence shown in Fig. 6 with four other T. brucei surface glycoproteins cDNAs has already been published (17) so will not be discussed here.

Incorporation of mercurated ribo- and deoxyribonucleotides into RNA and DNA

has been studied previously and can be accomplished in vitro by a number of enzymes (12). We found during the present work that T4 DNA polymerase would efficiently add Hg-dCTP and Hg-dUTP to DNA by virtue of its 5'→3' polymerase activity. Initially a major concern was that the T4 DNA polymerase would not efficiently incorporate the mercuroated nucleotides all the way back to the 3'-end but instead leave recessed 3'-ends that would label efficiently with the high specific activity ³²P-nucleotides so that both ends of the partial restriction fragments would be labeled, rendering them unsuitable for sequence analysis. This possibility did not happen. Neither small scale assays to test for recessed 3'-ends after mercuration nor subsequent sequence determinations of the partial restriction fragments indicated that the mercuroated 3'-end was labeled during the conventional DNA polymerase I repair reaction (Materials and Methods) used to label the other end for sequencing. In addition, the presence of about 100 mercury atoms in the major groove (15) near the ends of the DNA did not have a noticeable effect on the activity of restriction enzymes or *E. coli* DNA polymerase I, and did not influence the sequencing reactions. Mercuroated and non-mercuroated DNA fragments were indistinguishable on gels (Fig. 3).

In establishing the optimal experimental conditions for the overall protocol outlined in Figure 1, a number of preliminary experiments were performed. The conditions given in Materials and Methods were found to give the best results. Several characteristics of the protocol described here influence its application to a variety of sequencing projects.

- 1) All steps are in vitro. The starting material is the intact plasmid. No strand separation or secondary restriction enzyme digests are required if labeling is done by DNA polymerase I repair. These features allow the rapid generation of an entire set of end-labeled fragments. Because the largest portion of time is devoted to confirming the various steps (completeness of the restriction cuts, extent of mercuration, etc.), experience permits all the steps through the affinity chromatography to be conducted in one or two days, followed by overnight preparative gel electrophoresis and recovery of individual fragments. Therefore many terminal-labeled overlapping fragments ready for sequencing can be generated quickly.

- 2) A number of different partial restriction digests can be performed on the same mercuroated DNA substrate, and run on the same preparative gel after the labeling and affinity steps. Fragments can then be selected from the gel on the basis of their spacing. Thus, of the Taq I, Hinf I, etc., fragments produced, only those 200-300 bp different in length, and not necessarily from the same digest, need be sequenced to get overlapping information. Other fragments can

be (i) saved in case the expected overlaps are not achieved or (ii) discarded if they appear to be mixtures of two or more fragments because of closely spaced sites for the same restriction enzyme. The result is much faster and less labor-intensive than "shotgun" sequencing since each fragment is chosen on the basis of its migration relative to others on the preparative gel.

3) All the sequence obtained from a particular mercurated DNA will be from the same strand. To obtain the sequence of the opposite strand, the plasmid must either be opened on the other side of the insert (e.g. the *Ava* I site of pcB-1.1, Fig. 2), or the insert must be cloned in both orientations. Alternatively, it should also be possible to label the 3' ends of the partial digest with DNA polymerase, and the 5' ends with T4 polynucleotide kinase. In the latter instance, the label on the mercurated end would have to be cut off at some point, perhaps with a restriction enzyme that cleaves near the end. For example, in the case of pcB-1.1, it would be possible to open the plasmid at its single *Cla* I site (position 25), mercurate, secondary cut with *Bam* H I, partially-restrict, 5'-end label and then remove the mercurated label by cleavage with *Eco* R I at position 1 without removing so many mercurated nucleotides that the fragments would not bind to the thiol column. The sequence of both strands could then be determined from different aliquots of the same set of fragments.

4) Although over 30 commercially available restriction enzymes produce recessed 3'-ends that can be labeled by DNA polymerase repair, those which have short recognition sequences are the most likely candidates for use in generating the partial digestion products because they cut frequently. Among these are restriction enzymes with 4 or 5 long recognition sites such as *Taq* I, *Msp* I (*Hpa* II), *Sau* 3A I, *Hinf* I, *Dde* I, *Eco* R II or *Bst* N I, *Sau* 96 I, and *Ava* II or *Sin* I. Since we have sequenced DNA molecules of over 6000 nucleotides using DNA polymerase repair labeling exclusively (19) combinations of these enzymes should suffice for most requirements.

5) Preliminary mapping of the insert by the method of Smith and Birnstiel (18) is rapid and useful. The amount of DNA and the degree of completion of the partials needed can be judged, and the enzymes for the partial digestions can be judiciously chosen.

6) The DNA region to be sequenced must not contain sites for either of the enzymes used to open the plasmid or to cut off one end after mercuration. Since pBR322 contains at least sixteen unique restriction sites (see Appendix of New England BioLabs catalog), this requirement should not generally pose a problem. The varying sensitivities of restriction sites to the partial digestions, and the fact that each plasmid molecule yields only one (mercurated) "sequenceable"

fragment, dictate that 5-10 pmol of plasmid be used for each fragment to be prepared. As a consequence the plasmid is ideally opened quite close to the insert, so that relatively few sites occur between the mercurated end and the DNA to be sequenced.

Finally, the nucleotide sequence reported here is 980 base pairs long. However, the preparative gel shown in Fig. 4 demonstrates that "sequenceable" fragments of 3000 bp and more can be obtained by the approach described above. Because the labor involved is independent of the number of fragments until the final step of the protocol (elution of the labeled DNAs from the preparative gel), the time required is virtually the same whether the number of fragments sought is 4 or 40. Given suitable restriction sites in the vector and sufficient sites in the insert, the method presented here should enable rapid sequencing of cloned DNAs of several thousand base pairs or more with little data handling and a minimum of sequencing runs. These characteristics suggest the approach has wide applicability to projects requiring determination of DNA sequences. In addition the procedure can be used to generate DNA fragments of variable length for other projects - for example, different DNA hybridization probes of variable length but containing a common sequence.

ACKNOWLEDGMENTS

We thank David Dorfman for doing some of the DNA sequence determinations and John Young, Phelix Majiwa and Richard Williams of ILRAD, Nairobi, Kenya, for providing plasmid pcB-1.1. The research was supported by USPHS grants AM25295 and AI 16950 and NSF grant PCM 76-1341.

*Present address: Bethesda Research Labs., Inc., P.O. Box 6009, Gaithersburg, MD 20877, USA

REFERENCES

1. Maxam, A.M. and Gilbert, W. (1979) In Methods of Enzymology Vol. 65 (L. Grossman and K. Moldave, eds.) Academic Press, Inc., pp. 499-560.
2. Sanger, F. and Coulson, A.R. (1975) J. Mol. Biol. **94**, 441-448.
3. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. USA **74**, 5463-5468.
4. Rüther, V., Koenen, M., Otto, K. and Müller-Hill, B. (1981) Nucleic Acids Research **9**, 4087-4097.
5. Messing, J., Crea, R. and Seeburg, P.H. (1981) Nucleic Acid Research **9**, 309-321.
6. Sutcliffe, J.G. (1979) Cold Spring Harbor Sym. Quant. Biol. **43**, 77-90.
7. Frischauf, A.M., Garoff, H. and Lehrach, H. (1980) Nucleic Acids Research **23**, 5541-5549.
8. Barnes, W.M. (1980) In Genetic Engineering Vol. 2 (J.K. Setlow and A. Hollaender, eds.) Plenum Press, N.Y., pp. 185-200.

9. Bolivar, F. and Backman, K. (1979) In Methods in Enzymology Vol. 68, (R. Wu, ed.) Academic Press, Inc., pp. 245-267.
10. Clewell, D.B. and Helinski, D.R. (1972) J. Bact. 110, 1135-1146.
11. Majiwa, P.A.O., Young, J., Englund, P.T., Shapiro, S.Z. and Williams, R.O. (1982) Nature (in press).
12. Dale, R.M.K., Livingston, D.C. and Ward, D.C. (1973) Proc. Nat. Acad. Sci. USA 70, 2238-2242.
13. Smith, H.O. (1979) In Methods in Enzymology Vol. 65 (L. Grossman and K. Moldave, eds.) Academic Press, Inc., pp. 371-379.
14. O'Farrell, P.H., Kutter, E. and Nakanishi, M. (1980) Molec. Gen. Genetics 179, 421-435.
15. Dale, R.M.K. and Ward, D.C. (1975) Biochemistry 14, 2458-2469.
16. Feist, P.C. and Danna, K.J. (1981) Biochemistry 20, 4243-4246.
17. Rice-Ficht, A., Chen, K.K. and Donelson, J.E. (1981) Nature 294, 53-57.
18. Smith, H.O. and Birnstiel, M. (1976) Nucleic Acids Research 3, 2387-2398.
19. Hartley, J.L. and Donelson J.E. (1980) Nature 266, 860-865.
20. Rice-Ficht, A., Chen, K.K. and Donelson, J.E. (1982) Nature (in press).
21. Fagen, J.B., Pasten, I. and deCrombrughe, B. (1980) Nucleic Acids Research 8, 3055-3064.
22. Gopinathan, K.P., Weymouth, L.A., Kunkel, T.A. and Loeb, L.A. (1979) Nature 278, 857-859.