Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 8370–8374, December 1987
Biochemistry

cDNA sequence, protein structure, and chromosomal location of the human gene for poly(ADP-ribose) polymerase

(DNA binding protein/DNA-strand-break repair/chromosomes 1, 13, and 14/restriction-fragment-length polymorphism)

BARRY W. CHERNEY*, O. WESLEY MCBRIDE†, DEFENG CHEN*, HUSSEIN ALKHATIB*, KISHOR BHATIA*, PRESTON HENSLEY*, AND MARK E. SMULSON*‡

*Department of Biochemistry, Georgetown University Schools of Medicine and Dentistry, Washington, DC 20007; and †Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892

Communicated by Roscoe Brady, August 18, 1987 (received for review July 1, 1987)

ABSTRACT **Recently we described a full-length cDNA for the human nuclear enzyme poly(ADP-ribose) polymerase. Here, we report the chromosomal localization and partial map of the human gene for this enzyme as well as the complete coding sequence for this protein. The nucleotide sequence reveals a single 3042-base open reading frame encoding a protein with a predicted M_r of 113,135. A comparison of this deduced amino acid sequence with the amino acid sequence of three peptides derived from human poly(ADP-ribose) polymerase revealed a match of 27 amino acid residues. A computer-derived structural analysis of the enzyme and a search for similarities with other proteins confirmed that the polymerase belongs to a subfamily of DNA/NAD-binding proteins and DNA-repair proteins. Possible Zn^{2+}-binding "fingers," a nucleotide-binding fold, and a nuclear transport signal were noted. Additionally, chromosomal mapping has identified polymerase-hybridizing sequences on human chromosomes 1 (the active gene), 13, and 14 (processed pseudogenes). Using the polymerase cDNA as a probe, we also have detected several DNA restriction fragment length polymorphisms in normal humans.**

Considerable work in the past few years has centered on the possible role of the nuclear enzyme, poly(ADP-ribose) polymerase, in modulating the DNA replication/repair processes of mammalian cells (1). This enzyme uses NAD as a substrate in the formation of poly(ADP-ribose) chains at sites on many nuclear proteins. The enzyme binds tightly to DNA and requires DNA strand breaks for enzymatic activity (2). Since cells carrying a mutated polymerase are not yet available, a reasonable approach to answering questions about the biochemical and biological functions of this protein is through gene cloning. We have described (3) the cloning of poly(ADP-ribose) polymerase cDNA from a phage λgt11 library. This 1.3-kilobase (kb) insert was used in both hybrid-selection and arrest studies to confirm the identity of the partial cDNA. Subsequently, a 3.7-kb full-length cDNA contained in the Okayama–Berg expression vector was isolated, and transfection with this recombinant plasmid caused significant levels of transient polymerase expression in COS cells (3). The hyperexpression of this cloned cDNA in COS cells was also shown to increase the ability of these cells to repair DNA lesions caused by γ radiation (unpublished data). These results suggest that this gene encodes a protein that is functionally involved with DNA repair in eukaryotic cells. A partial polymerase cDNA clone has recently been reported by Suzuki *et al.* (5).

The enzyme is composed of three functionally distinct regions, an NH_2-terminal DNA-binding domain, a central automodification domain, and a COOH-terminal NAD-binding region (6). We now report the DNA§ and amino acid sequences of this enzyme and a structural analysis of its domains.

We also used various regions of the poly(ADP-ribose) polymerase cDNA as probes to chromosomally map the human gene for this nuclear enzyme by Southern analysis of DNAs isolated from human–rodent somatic cell hybrids. The results provide a molecular basis for studying the structure and expression of poly(ADP-ribose) polymerase at both the nucleic acid and protein levels.

MATERIALS AND METHODS

Purification of Peptides for Solid-Phase Sequencing. Purified human polymerase (40–50 μg) was pelleted, dissolved in 2 M urea, and digested with endoproteinase Lys C, 1:100 (wt/wt) at 37°C for 8 hr as described (7). The peptides were separated on a Vydac C_{18} reverse-phase HPLC column, and selected peptides were sequenced by the solid-phase procedure (8).

DNA Sequence Analysis. The full-length cDNA was subcloned as overlapping DNA fragments into phage M13 mp18 and mp19. Sequencing reactions were carried out with ^{35}S-substituted adenosine 5′-[γ-thio]triphosphate (ATP-[γ^{35}S]) by the dideoxy chain-termination method (9). Alternatively, restriction fragments were end-labeled and sequenced by the method of Maxam and Gilbert (10). Sequence similarity analysis was performed with the National Biomedical Research Foundation programs. Statistical significance (i.e., *Z* values) of potentially related sequences was determined by the method of Lipman and Pearson (11). Secondary structure and hydropathic analysis was based on established methods (12, 13).

Chromosomal Location Studies. The isolation and characterization of the human–rodent hybrids have been described (14). In general, hybrid cell lines were characterized by standard isoenzyme analyses, which distinguished all chromosomes except Y. Hybrid cell lines were also characterized by Southern analysis of the DNAs with probes that had been localized to specific chromosomes and by karyotyping in some cases. DNA was isolated from each hybrid cell line digested with *Eco*RI, size-fractionated by 0.7% agarose gel electrophoresis, transferred to nylon membranes, hybridized, and washed as described (14). Restriction-fragment-length polymorphism (RFLP) was determined from DNAs

Abbreviations: RFLP, restriction-fragment-length polymorphism; ORF, open reading frame.

‡To whom reprint requests should be addressed.

§The sequence reported in this paper is being deposited in the EMBL/GenBank data base (Bolt, Beranek, and Newman Laboratories, Cambridge, MA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J03030).

isolated from peripheral lymphocytes of normal unrelated individuals and from fibroblast cell lines.

RESULTS AND DISCUSSION

DNA Sequence and Derived Amino Acid Sequence of Human Poly(ADP-Ribose) Polymerase cDNA. The cDNA insert [3681 base pairs (bp)] encoding poly(ADP-ribose) polymerase was digested with restriction enzymes and subcloned in M13 for sequencing (Fig. 1). The nucleotide sequence of the polymerase cDNA is shown in Fig. 2. A single long opening reading frame (ORF) extends 3042 bases from the first ATG initiation codon to a TAA termination codon at positions 3202–3204. This ATG is flanked by sequences that fulfill the Kozak criteria (15) for initiation codons (Fig. 2). From the putative TAA termination codon to the start of the poly(A) sequence, there are 11 in-phase termination codons. Analysis of the other two reading frames showed 73 and 52 termination codons dispersed throughout the insert. The 113,135-Da protein deduced from this sequence contains 1014 amino acids. Its molecular weight is close to estimates for the purified protein (112–116 kDa) and is flanked by untranslated sequences of 159 bases 5′ and 459 bases 3′. The 5′ untranslated sequence has a 74% G+C content. A potential polyadenylylation signal (AATAAA) is present 16 bases upstream from the poly(A) sequence (Fig. 2). The deduced protein composition is in excellent agreement with previously published amino acid compositions of the polymerase from calf (16) and pig thymus (17) as well as Ehrlich ascites cells (17). There is a relatively high percentage of lysine and glutamic acid residues that are nonrandomly distributed within the protein (Fig. 3). The NH_2-terminal region is rich in lysine (14%) and glutamic acid (8%) and has a net basic charge. The NAD-binding COOH-terminal region (6) is composed of 9% lysine and an average amount of glutamic acid (6%). The central 19-kDa automodification domain (6) is similar in amino acid composition to the DNA-binding region.

To confirm the identity of the single long ORF as the correct frame, a comparison was made between the deduced amino acid sequence and the sequence of three polymerase peptide fragments obtained as described. Perfect matches for oligopeptides 2 and 3 (see Fig. 2, solid lines) were observed, while five of seven residues matched oligopeptide 1.

Computer-Derived Structure Predictions for Human Poly(ADP-Ribose) Polymerase. In agreement with an analysis of the CD spectrum of the calf thymus enzyme (7, 16), the Chou–Fasman secondary structure analysis (not shown) predicted a protein with a low β-sheet content (10%), a high random coil content (38%), a large number of β-turns (34%), and 26% helical content. The structure shows several helix-turn-helix units (centered at residues 82, 97, 230, and 257), characteristic of some DNA-binding proteins (18). Hydropathic analysis (Fig. 3) for the polymerase predicts a protein with low hydrophobic character. One feature is the prediction of two hydrophilic domains between residues 433 and 448 and between 484 and 521 with a central hydrophobic element (residues 449–483). This region encompasses the automodification domain of the polymerase. Interestingly, the hydrophilic sections contain an unusually high amount of lysine and glutamic acid residues (38% lysine/31% glutamic acid and 20% lysine/17% glutamic acid, respectively). Since these hydrophilic elements are probably accessible to modification and contain a high percentage of glutamic acid residues that participate in protein ADP–ribose linkages, it is possible that automodification is largely restricted to these elements. This speculation is consistent with data indicating that automodification occurs at ≈15 sites (19).

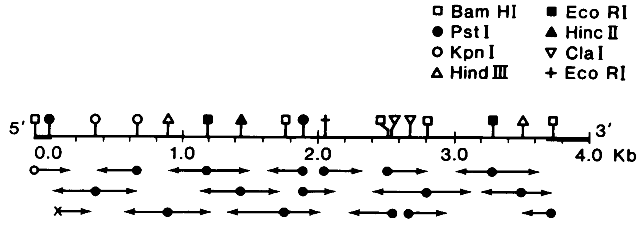

FIG. 1. Restriction endonuclease map and sequence determination strategy for pcD-poly(ADP-ribose) polymerase cDNA clone. The arrow beginning with an open circle denotes sequences determined by chemical modification. The arrows beginning with closed circles denote a sequence subcloned into phage M13 and determined by dideoxy chain termination. The arrow beginning with an "x" denotes a sequence derived by dideoxy chain termination using a pcD-poly(ADP-ribose) polymerase-derived synthetic oligomer as primer. The "+" represents a synthetic *Eco*RI linker used in the construction of a partial cDNA. There was complete sequence overlap of all regions.

Homology to Other Proteins. Sequence similarity comparison of the polymerase with the National Biomedical Research Foundation's protein data bases revealed no extensive identities (i.e., optimization score > 60). This analysis confirms that the enzyme belongs to a unique subfamily of DNA/NAD-binding proteins with some short but interesting identities. One interesting feature is the presence of a putative nucleotide-binding fold similar to many DNA-repair enzymes. A comparison of ATP-requiring enzymes has lead to the suggestion (20) of a consensus nucleotide-binding fold consisting of an "A" site (Gly-Xaa_4-Gly-Lys-Gly-Xaa_6-Val) and a "B" site [Arg-Xaa_{2-3}-Gly-Xaa_3(hydrophobic)$_{4-6}$-Asp]. Similar sites are observed in the NAD-binding domain of the polymerase (A site, residues 888–901; B site, residues 940–957). Another statistically significant identity (40% identity over residues 873–897; *Z* value = 10) is with the putative catalytic site of the ricin A chain (21). This cytotoxic plant protein inactivates the 60S ribosome by cleaving the N-glycosidic bond of 28S rRNA residue adenosine-4324 in a hydrolytic fashion (22). In part this similarity may reflect the presence of a consensus nucleotide-binding fold; however, this explanation does not completely account for the observed identity. We also detected a similarity to several nuclear transport signals (most significantly simian virus 40 large tumor antigen). This identity occurs in polymerase residues (219–226) that are located in a hydrophilic portion of a helix-turn-helix unit. Polymerase residues 361–416 also showed a slight homology (*Z* value = 5) with the product of *NMYC*. Since this region is just outside the DNA-binding domain of the polymerase, this identity may reflect some other feature of protein structure. Indeed these sequences share a resemblance to proline/glutamic acid/serine/threonine-rich sequences reported by Rogers *et al.* (4) to be involved in protein degradation. The polymerase also exhibits some slight homologies (*Z* values, 2–4) with a number of DNA-binding proteins such as DNA and RNA polymerases. The identities occur in two regions spanning polymerase residues 47–120 and 160–260. Consistent with a possible DNA-binding function, secondary structure predictions indicate these homologous regions in the enzyme exist as helix-turn-helix units. In addition, the -turn regions are rich in positively charged amino acids, suggesting DNA binding. The polymerase contains regions within the DNA-binding domain that are highly suggestive of a Zn^{2+} binding "finger" (18). These sequences are of the form Cys-Xaa_2-Cys-Xaa_{27}-His-Xaa_2-Cys (residues 21–56), His-Xaa_2-Cys-Xaa_5-His-Xaa_3-His (residues 53–66), and Cys-Xaa_2-Cys-Xaa_{30}-His-Xaa_2-Cys (residues 125–162). Since the NH_2-terminal portion of the polymerase specifically binds Zn^{2+}, Zn^{2+} binding fingers may be involved in polymerase–DNA recognition events. Therefore, this protein may use multiple DNA recognition strategies for DNA binding.

```
                                                                                           -150      -140      -130
                                                                                  AATCTATCAGGGAACGGCGGTGGCCGGTGCGGCGTGTTC
-120      -110      -100      -90       -80       -70        -60       -50       -40       -30       -20       -10
GGTGCGCTCTGGCCGCTCAGGCCGTGCGGCTGGGTGAGCGCACGCGAGGCGGCGAGGCGG CAAGCGTGTTTCTAGGTCGTGGCGTCGGGCTTCCGGAGCTTTGGCGGCAGCTAGGGGAGG
                            10                            20                             30                            40
MetAlaGluSerSerAspLysLeuTyrArgValGluTyrAlaLysSerGlyArgAlaSer CysLysLysCysSerGluSerIleProLysAspSerLeuArgMetAlaIleMetValGln
ATGGCGGAGTCTTCGGATAAGCTCTATCGAGTCGAGTACGCCAAGAGCGGGCGCGCCTCT TGCAAGAAATGCAGCGAGAGCATCCCCAAGGACTCGCTCCGGATGGCCATCATGGTGCAG
                    -----GLU---THR------ #1
                            50                            60                             70                            80
SerProMetPheAspGlyLysValProHisTrpTyrHisPheSerCysPheTrpLysVal GlyHisSerIleArgHisProAspValGluValAspGlyPheSerGluLeuArgTrpAsp
TCGCCCATGTTTGATGGAAAAGTCCCACACTGGTACCACTTCTCCTGCTTCTGGAAGGTG GGCCACTCCATCCGGCACCCTGACGTTGAGGTGGATGGGTTCTCTGAGCTTCGGTGGGAT
                            90                           100                            110                           120
AspGlnGlnLysValLysLysThrAlaGluAlaGlyGlyValThrGlyLysGlyGlnAsp GlyIleGlySerLysAlaGluLysThrLeuGlyAspPheAlaAlaGluTyrAlaLysSer
GACCAGCAGAAAGTCAAGAAGACAGCGGAAGCTGGAGGAGTGACAGGCAAAGGCCAGGAT GGAATTGGTAGCAAGGCAGAGAAGACTCTGGGTGACTTTGCAGCAGAGTATGCCAAGTCC
                                                                                --------------------------- #2
                           130                           140                            150                           160
AsnArgSerThrCysLysGlyCysMetGluLysIleGluLysGlyGlnValArgLeuSer LysLysMetValAspProGluLysProGlnLeuGlyMetIleAspArgTrpTyrHisPro
AACAGAAGTACGTGCAAGGGGTGTATGGAGAAGATAGAAAAGGGCCAGGTGCGCCTGTCC AAGAAGATGGTGGACCCGGAGAAGCCACAGCTAGGCATGATTGACCGCTGGTACCATCCA
                           170                           180                            190                           200
GlyCysPheValLysAsnArgGluGluLeuGlyPheArgProGluTyrSerAlaSerGln LeuLysGlyPheSerLeuLeuAlaThrGluAspLysGluAlaLeuLysLysGlnLeuPro
GGCTGCTTTGTCAAGAACAGGGAGGAGCTGGGTTTCCGGCCCGAGTACAGTGCGAGTCAG CTCAAGGGCTTCAGCCTCCTTGCTACAGAGGATAAAGAAGCCCTGAAGAAGCAGCTCCCA
              -------------------------------- #3
                           210                           220                            230                           240
GlyValLysSerGluGlyLysArgLysGlyAspGluValAspGlyValAspGluValAla LysLysLysSerLysLysGluLysAspLysAspSerLysLeuGluLysAlaLeuLysAla
GGAGTCAAGAGTGAAGGAAAGAGAAAAGGCGATGAGGTGGATGGAGTGGATGAAGTGGCG AAGAAGAAATCTAAAAAAGAAAAAGACAAGGATAGTAAGCTTGAAAAAGCCCTAAAGGCT
                           250                           260                            270                           280
GlnAsnAspLeuIleTrpAsnIleLysAspGluLeuLysLysValCysSerThrAsnAsp LeuLysGluLeuLeuIlePheAsnLysGlnGlnValProSerGlyGluSerAlaIleLeu
CAGAACGACCTGATCTGGAACATCAAGGACGAGCTAAAGAAAGTGTGTTCAACTAATGAC CTGAAGGAGCTACTCATCTTCAACAAGCAGCAAGTGCCTTCTGGGGAGTCGGCGATCTTG
                           290                           300                            310                           320
AspArgValAlaAspGlyMetValPheGlyAlaLeuLeuProCysGluGluCysSerGly GlnLeuValPheLysSerAspAlaTyrTyrCysThrGlyAspValThrAlaTrpThrLys
GACCGAGTAGCTGATGGCATGGTGTTCGGTGCCCTCCTTCCCTGCGAGGAATGCTCGGGT CAGCTGGTCTTCAAGAGCGATGCCTATTACTGCACTGGGGACGTCACTGCCTGGACCAAG
                           330                           340                            350                           360
CysMetValLysThrGlnThrProAsnArgLysGluTrpValThrProLysGluPheArg GluIleSerTyrLeuLysLysLeuLysValLysLysGlnAspArgIlePheProProGlu
TGTATGGTCAAGACACAGACACCCAACCGGAAGGAGTGGGTAACCCCAAAGGAATTCCGA GAAATCTCTTACCTCAAGAAATTGAAGGTTAAAAAGCAGGACCGTATATTCCCCCCAGAA
                           370                           380                            390                           400
ThrSerAlaSerValAlaAlaThrProProProSerThrAlaSerAlaProAlaAlaVal AsnSerSerAlaSerAlaAspLysProLeuSerAsnMetLysIleLeuThrLeuGlyLys
ACCAGCGCCTCCGTGGCGGCCACGCCTCCGCCCTCCACAGCCTCGGCTCCTGCTGCTGTG AACTCCTCTGCTTCAGCAGATAAGCCATTATCCAACATGAAGATCCTGACTCTCGGGAAG
                           410                           420                            430                           440
LeuSerArgAsnLysAspGluValLysAlaMetIleGluLysLeuGlyGlyLysLeuThr GlyThrAlaAsnLysAlaSerLeuCysIleSerThrLysLysGluValGluLysMetAsn
CTGTCCCGGAACAAGGATGAAGTGAAGGCCATGATTGAGAAACTCGGGGGGAAGTTGACG GGGACGGCCAACAAGGCTTCCCTGTGCATCAGCACCAAAAAGGAGGTGGAAAAGATGAAT
                           450                           460                            470                           480
LysLysMetGluGluValLysGluAlaAsnIleArgValValSerGluAspPheLeuGln AspValSerAlaSerThrLysSerLeuGlnGluLeuPheLeuAlaHisIleLeuSerPro
AAGAAGATGGAGGAAGTAAAGGAAGCCAACATCCGAGTTGTGTCTGAGGACTTCCTCCAG GACGTCTCCGCCTCCACCAAGAGCCTTCAGGAGTTGTTCTTAGCGCACATCTTGTCCCCT
                           490                           500                            510                           520
TrpGlyAlaGluValLysAlaGluProValGluValValAlaProArgGlyLysSerGly AlaAlaLeuSerLysLysSerLysGlyGlnValLysGluGluGlyIleAsnLysSerGlu
TGGGGGGCAGAGGTGAAGGCAGAGCCTGTTGAAGTTGTGGCCCCAAGAGGGAAGTCAGGG GCTGCGCTCTCCAAAAAAAGCAAGGGCCAGGTCAAGGAGGAAGGTATCAACAAATCTGAA
                           530                           540                            550                           560
LysArgMetLysLeuThrLeuLysGlyGlyAlaAlaValAspProAspSerGlyLeuGlu HisSerAlaHisValLeuGluLysGlyGlyLysValPheSerAlaThrLeuGlyLeuVal
AAGAGAATGAAATTAACTCTTAAAGGAGGAGCAGCTGTGGATCCTGATTCTGGACTGGAA CACTCTGCGCATGTCCTGGAGAAAGGTGGGAAGGTCTTCAGTGCCACCCTTGGCCTGGTG
                           570                           580                            590                           600
AspIleValLysGlyThrAsnSerTyrTyrLysLeuGlnLeuLeuGluAspAspLysGlu AsnArgTyrTrpIlePheArgSerTrpGlyArgValGlyThrValIleGlySerAsnLys
GACATCGTTAAAGGAACCAACTCCTACTACAAGCTGCAGCTTCTGGAGGACGACAAGGAA AACAGGTATTGGATATTCAGGTCCTGGGGCCGTGTGGGTACGGTGATCGGTAGCAACAAA
                           610                           620                            630                           640
LeuGluGlnMetProSerLysGluAspAlaIleGluGlnPheMetLysLeuTyrGluGlu LysThrGlyAsnAlaTrpHisSerLysAsnPheThrLysTyrProLysLysPheTyrPro
CTGGAACAGATGCCGTCCAAGGAGGATGCCATTGAGCAGTTCATGAAATTATATGAAGAA AAAACCGGGAACGCTTGGCACTCCAAAAATTTCACGAAGTATCCCAAAAAGTTTTACCCC
                           650                           660                            670                           680
LeuGluIleAspTyrGlyGlnAspGluGluAlaValLysLysLeuThrValAsnProGly ThrLysSerLysLeuProLysProValGlnAspLeuIleLysMetIlePheAspValGlu
CTGGAGATTGACTATGGCCAGGATGAAGAGGCAGTGAAGAAGCTCACAGTAAATCCTGGC ACCAAGTCCAAGCTCCCCAAGCCAGTTCAGGACCTCATCAAGATGATCTTTGATGTGGAA
                           690                           700                            710                           720
SerMetLysLysAlaMetValGluTyrGluIleAspLeuGlnLysMetProLeuGlyLys LeuSerLysArgGlnIleGlnAlaAlaTyrSerIleLeuSerGluValGlnGlnAlaVal
AGTATGAAGAAAGCCATGGTGGAGTATGAGATCGACCTTCAGAAGATGCCCTTGGGGAAG CTGAGCAAAAGGCAGATCCAGGCCGCATACTCCATCCTCAGTGAGGTCCAGCAGGCGGTG
                           730                           740                            750                           760
SerGlnGlySerSerAspSerGlnIleLeuAspLeuSerAsnArgPheTyrThrLeuIle ProHisAspPheGlyMetLysLysProProLeuLeuAsnAsnAlaAspSerValGlnAla
TCTCAGGGCAGCAGCGACTCTCAGATCCTGGATCTCTCAAATCGCTTTTACACCCTGATC CCCCACGACTTTGGGATGAAGAAGCCTCCGCTCCTGAACAATGCAGACAGTGTGCAGGCC
                           770                           780                            790                           800
LysValGluMetLeuAspAsnLeuLeuAspIleGluValAlaTyrSerLeuLeuArgGly GlySerAspAspSerSerLysAspProIleAspValAsnTyrGluLysLeuLysThrAsp
AAGGTGGAAATGCTTGACAACCTGCTGGACATCGAGGTGGCCTACAGTCTGCTCAGGGGA GGGTCTGATGATAGCAGCAAGGATCCCATCGATGTCAACTATGAGAAGCTCAAAACTGAC
                           810                           820                            830                           840
IleLysValValAspArgAspSerGluGluAlaGluIleIleArgLysTyrValLysAsn ThrHisAlaThrThrHisSerAlaTyrAspLeuGluValIleAspIlePheLysIleGlu
ATTAAGGTGGTTGACAGAGATTCTGAAGAAGCCGAGATCATCAGGAAGTATGTTAAGAAC ACTCATGCAACCACACACAGTGCGTATGACTTGGAAGTCATCGATATCTTTAAGATAGAG
                           850                           860                            870                           880
ArgGluGlyGluCysGlnArgTyrLysProPheLysGlnLeuHisAsnArgArgLeuLeu TrpHisGlySerArgThrThrAsnPheAlaGlyIleLeuSerGlnGlyLeuArgIleAla
CGTGAAGGCGAATGCCAGCGTTACAAGCCCTTTAAGCAGCTTCATAACCGAAGATTGCTG TGGCACGGGTCCAGGACCACCAACTTTGCTGGGATCCTGTCCCAGGGTCTTCGGATAGCC
                           890                           900                            910                           920
ProProGluAlaProValThrGlyTyrMetPheGlyLysGlyIleTyrPheAlaAspMet ValSerLysSerAlaAsnTyrTyrHisThrSerGlnGlyAspProIleGlyLeuIleLeu
CCGCCTGAAGCGCCCGTGACAGGCTACATGTTTGGTAAAGGGATCTATTTCGCTGACATG GTCTCCAAGAGTGCCAACTACTACCATACGTCTCAGGGAGACCCAATAGGCTTAATCCTG
                           930                           940                            950                           960
LeuGlyGluValAlaLeuGlyAsnMetTyrGluLeuLysHisAlaSerHisIleSerArg LeuProLysGlyLysHisSerValLysGlyLeuGlyLysThrThrProAspProSerAla
TTGGGAGAAGTTGCCCTTGGAAACATGTATGAACTGAAGCACGCTTCACATATCAGCAGG TTACCCAAGGGCAAGCACAGTGTCAAAGGTTTGGGCAAAACTACCCCTGATCCTTCAGCT
                           970                           980                            990                          1000
AsnIleSerLeuAspGlyValAspValProLeuGlyThrGlyIleSerSerGlyValAsn AspThrSerLeuLeuTyrAsnGluTyrIleValTyrAspIleAlaGlnValAsnLeuLys
AACATTAGTCTGGATGGTGTAGACGTTCCTCTTGGGACCGGGATTTCATCTGGTGTGATA GACACCTCTCTACTATATAACGAGTACATTGTCTATGATATTGCTCAGGTAAATCTGAAG
                          1010
TyrLeuLeuLysLeuLysPheAsnPheLysThrSerLeuTrpEND
TATCTGCTGAAACTGAAATTCAATTTTAAGACCTCCCTGTGGTAATTGGGAGAGGTAGCC GAGTCACACCCGGTGGCTGTGGTATGAATTCACCCGAAGCGCTTCTGCACCAACTCACCT
GGCCGCTAAGTTGCTGATGGGTAGTACCTGTACTAAACCACCTCAGAAAGGATTTACAG AAACGTGTTAAAGGTTTTCTCTAACTTCTCAAGTCCCTTGTTTGTGTTGTGTCTGTGGG
GAGGGGTTGTTTTGGGGTTGTTTTGTTTTTCTTGCCAGGTAGATAAAACTGACATAGA GAAAAGGCTGGAGAGAGATTCTGTTGCATAGACTAGTCCTATGGAAAAAACCAAAGCTTC
GTTAGAATGTCTGCCTTACTGGTTTCCCCAGGGAAGGAAAAATACACTTCCACCCTTTTT TCTAAGTGTTCGTCTTTAGTTTTGATTTTGGAAAGATGTTAAGCATTTATTTTTAGTTAA
AATAAAAACTAATTTCATACT(A)21
```

FIG. 2. Nucleotide sequence of the Okayama–Berg pCD-poly(ADP-ribose) polymerase cDNA insert and the deduced amino acid sequence of the 113,135-Da protein. The protein contains sequences coding for three poly(ADP-ribose) polymerase peptides (underlined and sequentially numbered), in the 3′ untranslated region, a putative mRNA processing signal (AATAAA) is underlined. In the 5′ region two nucleotides that correspond to the Kozak criteria for initiation are underlined (15).

Chromosomal Location of the Human Gene. A 3.7-kilobase (kb) full-length insert of the polymerase cDNA was used as a probe for Southern analysis (Fig. 4) of *Eco*RI-digested DNAs isolated from somatic cell hybrids described in detail previously (14). A complex pattern of 2.3-, 5.3-, 6.4-, 6.8-, 7.0-, 8.2-, and 25-kb hybridizing bands were detected in

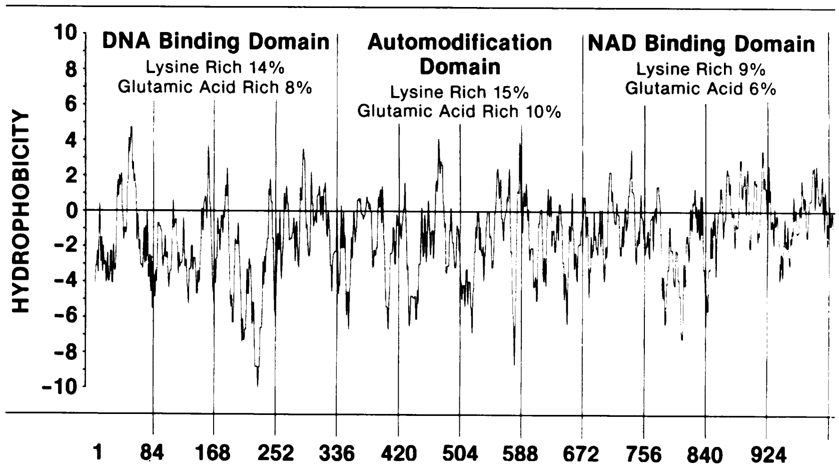

FIG. 3. A computer-derived hydropathy plot for poly(ADP-ribose) polymerase. Seven-residue averages have been plotted by using a program written by Rose *et al.* (12) based on the procedure of Hopp and Woods. All values above the horizontal line in the hydropathy plot indicate hydrophobic character, values below the line indicate hydrophilic character of the peptide. The three domains of the polymerase are illustrated along with the atypical amino acid compositions.

human DNA (Fig. 4, lane H) while 2.0-, 22-, and 24-kb or 1.2-, 2.3-, 2.9-, 10- and 11-kb cross-hybridizing fragments were found in CHO cell (Fig. 4, lane C) or mouse (not shown) DNAs, respectively. Based upon the distribution of specific hybridizing human bands found in different somatic cell hybrids, it was apparent that these sequences were located on at least three different human chromosomes (Fig. 4).

Analysis of 95 human–mouse and human–hamster *Eco*RI digested DNAs with the 3.7-kb cDNA probe demonstrated that the 25-, 7.0-, 8.2-, 6.4-, and 2.3-kb fragments all segregate concordantly with human chromosome 1 and discordantly (≥13%) with all other chromosomes (Table 1). Similarly, the 5.3-kb and 6.8-kb sequences could be assigned unambiguously to human chromosomes 13 and 14, respectively. These results indicate that polymerase-related sequences are located on human chromosomes 1, 13, and 14, and this localization was confirmed by Southern analysis of *Hin*dIII digests of the somatic cell hybrid DNAs (not shown).

Examination of three hybrids containing spontaneous breaks involving chromosome 1 allowed regional localization of the polymerase sequences on this chromosome. One hamster–human hybrid retained most of the short arm (1p) including the gene for phosphoglucomutase-1 and the *MYCL* protooncogene but not the *NRAS* protooncogene or any long arm (1q) markers. Two mouse–human hybrids retained all chromosome 1p markers including *NRAS* but contained breaks proximal to the *abl*-related protooncogene at 1q24-25. The polymerase sequences were absent from all three hybrids, indicating that they are located in chromosome 1q or the short arm proximal to *NRAS* (1p11–1p22). The 6.8-kb *Eco*RI sequence was regionally assigned to chromosome 14q13–q32 by similar methods. The human parental fibroblasts in one series of human–hamster hybrids contained a well-characterized 14; X(q32;q13) translocation. Seven hybrids retaining the 14q32 translocation chromosome in the absence of normal chromosome 14 all retained the 6.8-kb sequence. In contrast, two hybrids retaining the reciprocal translocation chromosome did not contain this sequence. Three additional mouse–human hybrids containing a spontaneous break distal to the nucleoside phosphorylase locus *NP* and proximal to 14q32 also contained the 6.8-kb sequence. The three discordancies (Table 1, column 4) with chromosome 14 were also informative for regional localization of the sequence. One was a hamster–human hybrid retaining part of the 14pterq32; Xq13-qter translocation chromosome with a spontaneous break distal to *NP* and absence of chromosome 14 sequences above this break point. This hybrid retained the 6.8-kb human sequence. Two mouse–human hybrids with breaks distal to *NP* also lost the 6.8-kb sequence. These combined results all strongly suggest that the polymerase sequence on chromosome 14 is located distal to *NP* (14q13) and proximal to 14q32. Both discordancies of chromosome 13 with the 5.3-kb sequence represented the presence of the hybridizing sequence and absence of esterase D activity. These hybrids were positive for three other sequences on chromosome 13, including two anonymous DNA segments from a chromosome-specific library and a calmodulin gene or pseudogene (O.W.M., unpublished data). These discordancies could be explained by differing sensitivities for detection of esterase D and the hybridizing DNA sequences or by spontaneous breaks involving this chromosome.

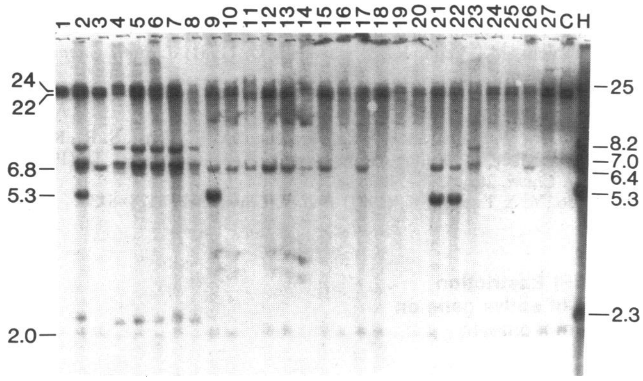

FIG. 4. Southern hybridization of representative *Eco*RI-digested human–hamster somatic cell hybrid DNAs with a full-length 3.7-kb human polymerase cDNA probe. A different hybrid cell DNA is present in lanes 1–27. Parental Chinese hamster (lane C) and human placental (lane H) DNAs are also shown. The sizes of hybridizing human (2.3, 5.3, 6.4, 6.8, 7.0, 8.2, and 25 kb) and hamster (2.0, 22, and 24 kb) fragments are depicted. The 5.3-kb human band was detected in lanes 2, 9, 21, and 22, whereas the 6.8-kb band was observed in lanes 2–15, 17, 21–23, and 26; the remaining human bands were found in lanes 2, 4–8, and 23. The independent segregation of hybridizing human sequences indicates that they are present at three different loci.

To determine whether the five hybridizing *Eco*RI fragments on chromosome 1 represented one or multiple genes, the probe was removed from the Southern blots, and they were hybridized sequentially with specific portions of the polymerase cDNA insert. Probe A was a 0.9-kb 5′ fragment (*Pst* I–*Hin*dIII); B, a 1.9-kb 5′ *Pst* I–*Pst* I fragment; C, an internal 1.2-kb predominantly 3′ fragment; and D, a 1.6-kb 3′ *Pst* I–*Hin*dIII fragment (Fig. 5). The 5.3-kb sequence on chromosome 13 was detected with all of these probes, whereas the 6.8-kb *Eco*RI band on chromosome 14 did not hybridize with either of the 5′ probes (A or B). In contrast, the different probes identified different *Eco*RI bands and series of bands on chromosome 1q (Fig. 5). We conclude that the polymerase probes identify a single gene on chromosome 1q, and this analysis allows the cluster of *Eco*RI fragments to

Table 1. Segregation of poly(ADP-ribose) polymerase gene in human–rodent hybrids

	Discordant segregation with chromosomes, %																						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X
Gene	0	13	29	32	20	24	40	36	23	27	23	27	32	35	38	36	40	43	19	23	51	24	52
P1 fragment	29	33	46	54	45	43	21	47	34	38	40	34	2	35	27	38	34	28	35	43	42	44	56
P2 fragment	41	43	61	51	39	20	44	39	43	60	51	33	47	3	31	32	43	42	46	42	46	51	45

The poly(ADP-ribose) polymerase gene was detected as 2.3-, 7.0-, 8.2-, and 25-kb hybridizing bands (27 positive hybrids) in *Eco*RI digests of human–rodent somatic cell hybrid DNAs or as a 5.3-kb sequence (P1) or 6.8-kb band (P2). Detection of each sequence or group of sequences is correlated with the presence or absence of each human chromosome in the somatic cell hybrids. Discordancy indicates the presence of hybridizing sequences in the absence of the chromosome or absence of the hybridizing bands despite the presence of the chromosome; the sum of these numbers divided by total hybrids examined (×100) represents percent discordancy. The human–hamster hybrids consisted of 26 primary clones and 15 subclones, and the human–mouse hybrids contained 13 primary hybrids and 42 subclones. The 5.3-kb and 6.8-kb human poly(ADP-ribose) polymerase sequences were detected in 35 and 54 hybrid cell DNAs, respectively.

be ordered as follows: 5′ flank–25 kb–7.0 kb–8.2 kb–2.3 kb–6.4 kb–3′ flank (Fig. 5). The 6.4-kb band (not shown in Fig. 5) was identified with the 3.7-kb full-length probe (Fig. 4) but was not detected with the other probes described. This suggests that the 6.4-kb band shares identity with cDNA sequence 3′ of the second *Hin*dIII site. The cDNA contains no *Eco*RI site at the junction of the 7.0- and 8.2-kb fragments, and this site must occur within an intron. Any fragments that are exclusively intronic would not be detected with the cDNA probes. We interpret these results to suggest that a large (≥18–42 kb) functional gene is located on chromosome 1q, and the sequences on chromosomes 13 and 14 most likely represent processed pseudogenes.

Peripheral DNAs from 10 unrelated individuals were digested with restriction endonucleases, and Southern blots were hybridized with the cDNA probes. Several simple two allele polymorphisms (RFLPs) were detected including 2.6-kb and 2.9-kb *Hin*dIII alleles (0.12:0.88 frequencies) on chromosome 13. RFLPs were also detected with *Pst* I, *Sst* I, *Taq* I, and *Msp* I (not shown) but chromosomal localization of each of these RFLPs is incomplete.

The poly(ADP-ribosylation) of nuclear proteins represents an important modulation of chromatin structure in response to DNA strand breaks during various biological processes such as DNA repair or replication. The identification of the locus for the functional gene and construction of an *Eco*RI restriction map of this locus provides a strategy for cloning the active gene. Additionally, identification of the specific RFLP will permit an evaluation of the potential role of this gene in various DNA-repair diseases.

Note Added in Proof. The following corrections in the sequence in Fig. 2 should be noted: amino acid 50 should read Asp (GAC); amino acid 827 should read Asn (AAT).

We gratefully thank Dr. Ken Williams (Yale University) for his excellent suggestions on proteolytic cleavage and his help in peptide sequences. We also wish to thank Drs. Caroline Tahourdin, Shiuan Chen, Ellis Merlin, and Vicente Notario for their contributions. This research was supported by Grants CA 13195 and CA 25344 from the National Cancer Institute and Grant OSR-86-0024 from the U.S. Air Force, Office of Scientific Research.

1. Smulson, M. E. & Sugimura, T., eds. (1980) *Novel ADP-Ribosylations of Regulatory Enzymes and Proteins* (Elsevier, New York).
2. Benjamin, R. C. & Gill, D. M. (1980) *J. Biol. Chem.* **255**, 10502–10508.
3. Alkhatib, H. M., Cheng, D., Cherney, B., Bhatia, K., Notario, V., Giri, C., Stein, G., Slattery, E., Roeder, R. G. & Smulson, M. E. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 1224–1228.
4. Rogers, S., Wells, R. & Rechsteiner, M. (1986) *Science* **234**, 364–369.
5. Suzuki, H., Uchida, K., Shima, H., Sato, T., Okamoto, T., Teruyuki, K. & Miwa, M. (1987) *Biochem. Biophys. Res. Commun.* **146**, 403–409.
6. Kameshita, I., Matsuda, M., Nishikimi, M., Ushiro, H. & Shizuta, Y. (1986) *J. Biol. Chem.* **261**, 3863–3868.
7. Jugle, P. A., Wicher, J. W. & Beintema, J. J. (1983) *Anal. Biochem.* **134**, 347–354.
8. Hemmings, H. C., Williams, K. R., Konigsberg, W. H. & Greengard, P. (1984) *J. Biol. Chem.* **259**, 14486–14490.
9. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
10. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
11. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227**, 1435–1440.
12. Rose, G. D., Gierasch, L. M. & Smith, J. A. (1985) *Adv. Protein Chem.* **37**, 1–109.
13. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol.* **47**, 45–148.
14. McBride, O. W., Zmudzka, B. Z. & Wilson, S. H. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 503–507.
15. Kozak, M. (1983) *Microbiol. Rev.* **47**, 1–40.
16. Ito, S., Shitzuta, Y. & Hayaishi, O. (1979) *J. Biol. Chem.* **254**, 3647–3651.
17. Holtlund, J., Kristensen, T., Ostuold, A. & Laland, S. G. (1981) *Eur. J. Biochem.* **119**, 23–29.
18. Berg, J. M. (1986) *Nature (London)* **319**, 264–265.
19. Kawaichi, M., Veda, K. & Hayaishi, O. (1981) *J. Biol. Chem.* **256**, 9483–9489.
20. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. (1982) *EMBO J.* **1**, 945–951.
21. Montfort, W., Villafranca, J. E., Monzingo, A. F., Ernst, S. R., Katzin, B., Rutenber, E., Xuong, N. J., Hamlin, R. & Robertus, J. D. (1987) *J. Biol. Chem.* **262**, 5398–5403.
22. Endo, Y. & Tsurugi, K. (1987) *J. Biol. Chem.* **262**, 8128–8130.

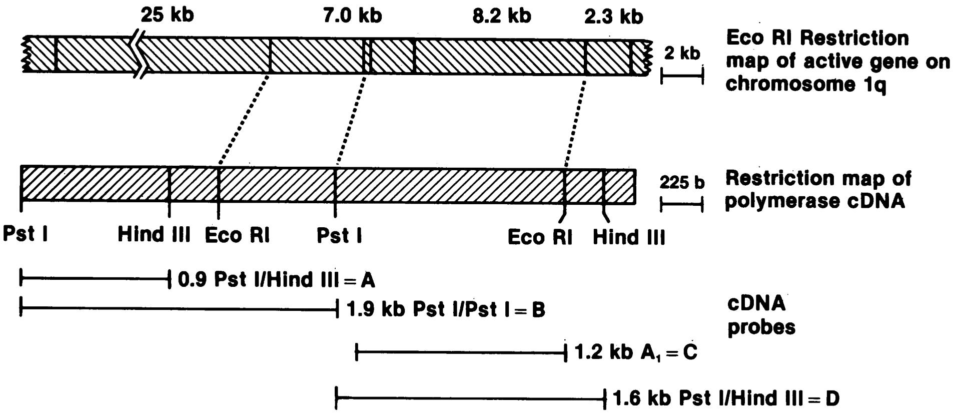

FIG. 5. Genomic organization of poly(ADP-ribose) polymerase. The alignment of similar *Eco*RI sites present in the genomic and cDNA map is illustrated together with cDNA probes used to order the *Eco*RI fragments. The approximate location in the genomic map for the *Pst* I cDNA site is also shown. (Separate scales have been used to construct the genomic and cDNA maps.) b, Base.