Structures of Two HaeIII-type Genes in the Human Salivary Proline-rich Protein Multigene Family*

Hyung-Suk Kim and Nobuyo Maeda

From the Laboratory of Genetics, University of Wisconsin-Madison, Madison, Wisconsin 53706

Two members of the human salivary proline-rich protein (PRP) multigene family have been isolated and completely sequenced. These PRP genes, PRHI and PRH2, are of the HaeIII-type subfamily and code for acidic PRP proteins. Both genes are approximately 3.5 kilobase pairs (kb) in length and contain four exons. Exon 3 encodes the proline-rich part of the protein and includes five 63-base pair (bp) repeats. CAT and ATA boxes and several possible enhancer sequences occur in a 1-kb region 5' to exon 1. Two sets of repeats occur in the sequenced region in addition to the 63-bp repeats: one pair of about 140 bp flanks 500 bp of DNA in the first intervening sequence, and the other pair of 72 bp is tandemly repeated 1.4 kb 5' to the PRHI gene. The 4-kb region of sequenced DNA from PRHI differs by an average of 8.7% from the same region in PRH2, but the nucleotide sequences of the exon 3 of the two genes differ by only 0.2%. This result suggests the occurrence of a recent gene conversion event. The regions containing the 5-fold repeated sequences of 63 bp are identical in the two genes, PRHI and PRH2. A comparison of the human HaeIII and BstNI subfamily repeats and a comparison of the human ham, mouse, and rat repeats suggest that the individual repeats have evolved in a concerted fashion within each gene and within the PRP gene family as a whole.

Proline-rich proteins (PRPs) constitute about 70% of human salivary proteins. They are characterized by a predominance of the amino acids proline (25–42%), glycine (16–22%), and glutamic acid/glutamine (15–28%), which together make up 70–85% of the proteins. Protein studies have shown that the human salivary PRP system is highly complex and polymorphic in all races studied (Bennick, 1982; Azen and Denniston, 1981; Azen and Yu, 1984). A set of genes localized on chromosome 12 (Azen et al., 1986) controls the synthesis of the PRPs. We have recently cloned a number of cDNAs coding for PRPs (Maeda et al., 1985) and have shown that the complexity of the PRP protein system is in part due to post-translational processing of precursor proteins that can generate multiple PRP products from one gene. In addition, our cDNA study suggested that, in at least one gene, differential splicing may cause the generation of three transcripts with different lengths, possibly contributing an additional source of complexity.

The PRP proteins in humans are classified into three groups: acidic, basic, and glycosylated. Protein studies led to the proposal of four loci coding for the acidic PRPs. They are Pr, proline-rich (Azen and Oppenheim, 1973); Db, double band (Azen and Denniston, 1974); Pa, acidic protein (Friedman and Merritt, 1975); and PIF, parotid isoelectric focusing variant (Azen and Denniston, 1981). The locus Pr has two common alleles, Pr1 and Pr2, and an infrequent variant Pr1'. One productive allele and one null allele have been described at each of the Db, Pa, and PIF loci on the basis of protein data. Our DNA studies, in contrast, indicate that only two loci encode the acidic PRPs and that these two loci are a subfamily of the larger (six loci) PRP gene family. Thus, Maeda (1985) re-examined the pattern of inheritance of the PRPs in conjunction with DNA blotting data and hypothesized that the three acidic PRPs (Db, Pa, and PIF) are coded by three alleles (proposed nomenclature PRHI', PRH2', and PRH4') at the single locus, PRHI. PRH2', PRH4', and PRH2* were suggested to be alleles at a second locus, PRH2, that also codes for acidic PRPs, Pr1, Pr2, and Pr1', respectively.

The PRHI and PRH2 genes that code for the acidic PRPs form a PRP subfamily in which sites for the restriction enzyme HaeIII occur repeatedly. The other subfamily codes for the basic and glycosylated PRPs and consists of four BstNI-type genes, PRB1, PRB2, PRB3, and PRB4. These BstNI-type genes have a region where BstNI sites occur repeatedly. Two clones corresponding to the BstNI group, PRP1 and PRP2, have been isolated and partially sequenced (Azen et al., 1984). Tandem repetitive sequences of 63 nucleotides encoding for proline-rich repeated amino acid sequences were observed within the BstNI-type genes.

Multigene families are of particular interest because of the possibility that recombinational events between family members play a role in their evolution. The PRP multigene system shows signs of being involved in such events because polymorphic length differences are frequently observed in the PRP genes of different individuals (Azen et al., 1984). This suggests that unequal crossing-over between the repeated units within the genes may have been occurring during the evolution of the family.

To gain a better understanding of the complexity of the PRP genes and of the evolutionary inter-relationships of the genes in this multigene family, we have isolated both human salivary PRP genes of the HaeIII-type, PRHI and PRH2, and in this paper we describe their complete DNA sequences. Comparison of these nucleotide sequences suggests that a recent gene conversion between the two genes rendered the exons containing the repeated regions more alike than other parts of the genes.

---

*This work was supported by National Institutes of Health Grants AM20126 and GM20069 (to Oliver Smithies) and DE03658 (to Edwin A. Azen). This is Paper 2850 from the Laboratory of Genetics, University of Wisconsin-Madison. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

1 The abbreviations used are: PRPs, proline-rich proteins; kb, kilobase pairs; bp, base pairs.
RESULTS AND DISCUSSION

The PRH1 and PRH2 Loci—The 18-kb region which contains the PRH1 gene (allele PRH1*) was obtained from DNA of R. D. as a series of overlapping phage clones 1-4 (Fig. 1). R. D. is homozygous for the PRH1* allele (old nomenclature, PRP*). Three libraries containing BglII and BamHI digests of genomic DNA of R. D. were constructed from complete BglII and BamHI digests of genomic DNA of R. D. The digests were ligated into λ phage Charon 35 (Loenen and Blattner, 1981) and packaged in vitro (Hohn, 1979). An EcoRI library was made from the same individual using the vector Charon 35 (Loenen and Blattner, 1983). A library was made of 12-18-kb size-selected fragments from a partial MboI digest of DNA from O. S. ligated into the BamHI sites of Charon 35. All libraries were screened without amplification. A 500-bp fragment (HaeIII 500) of a human PRP cDNA clone which encodes all of the repetitive region (plus a little of both the 5' and 3' regions) was cloned into the Smal site of the plasmid pL10 (Rothenstein et al., 1979). This fragment was used as a probe during most of the cloning work. The HaeIII 500 probe hybridizes strongly to both HaeIII- and BstNI-type genes under non-stringent washing conditions at 68°C. After hybridization under less stringent conditions (0.1× SSC at 68°C), we could distinguish the strongly labeled HaeIII-type clones from the now weakly labeled BstNI-type clones.

DNA Sequencing and Analysis—DNA sequencing was carried out by the method of Maxam and Gilbert (1977) with slight modifications (Slightom et al., 1980). All regions were sequenced in both directions. Sequences were analyzed using software provided by the University of Wisconsin Genetic Computer Group (Devereux et al., 1984).

EXPERIMENTAL PROCEDURES

Materials—Human genomic DNA was prepared from white blood cells of the same individual (R. D.) whose parotid gland was used in our earlier cDNA study (Maeda et al., 1985). R. D. has the acidic PRP* protein phenotypes Db*, Pa*, and PIP* (genotypes PRH1*/PRH2* and Prl-1 (genotype PRH2*/PRH22)). Other DNA was from cultured fibroblast cells of an individual (O. S.) who has the acidic PRP* protein phenotypes Db*, Pa*, and PIP* and Prl-2 (genotypes PRH11/PRH1 and PRH21/PRH2').

Plasmid pLL was made from the same individual using the vector Charon 32 (Loenen and Blattner, 1981). All regions were sequenced in both directions. Sequences were analyzed using software provided by the University of Wisconsin Genetic Computer Group (Devereux et al., 1984).
Human Salivary Proline-rich Protein Multigene Family

Fig. 2.
Human Salivary Proline-rich Protein Multigene Family

Exon 1 is 64 bp in length and codes for the secretory signal sequence and for the first five NH₂-terminal residues of acidic PRPs.

Exon 2, located approximately 1 kb downstream from exon 1, contains only 36 bp and codes for the next 12 residues of the NH₂-terminal region of the proteins.

Exon 3, located approximately 360 bp downstream from exon 2, encodes for the main repeated region of the two proteins where the HaeIII-type repeats occur tandemly five times. These repeated regions are not interrupted by any introns. The termination codon TAA also occurs in exon 3.

Exon 4, located after a long intervening sequence of about 4}
1200 bp, contains only the 3'-untranslated region and the poly(A) addition signal sequence AATAAA (Fitzgerald and Shenk, 1981).

The nucleotide positions 1615 in PRH1 and 850 in PRH2 are indicated by bent arrows in Fig. 2. They are the 5' ends of the nucleotide sequences of the cDNAs cP2 and cP1, respectively. The glutamine residue in exon 1, indicated by an asterisk, is the NH2-terminal amino acid of the secreted protein as judged by the amino acid sequence of the acidic PRP, protein C, determined by Wong and Bennick (1980). The poly(A) attachment sites in PRH1 and PRH2 were deduced from the nucleotide sequence of the cDNAs cP1 and cP2. They occur after the sequence TTGC, as indicated by the bent arrows in exon 4 in Fig. 2, at nucleotide positions 4835 and 3926.

Several sequences associated with transcriptional initiation are found in the 5' regions of the two genes, including an ATA box (boxed in Fig. 2; Goldberg, 1979) and a possible CAT box (Efstratiadis et al., 1980) located 27 bp upstream of the ATA box. The sequence TGGAAAG, the core sequence of some viral enhancers (Khoury and Gruss, 1983), occurs twice in the 5' region, once at 483 bp and once at 345 bp upstream from the ATG codon. Similar sequences, TGAAAAA, TGAAAAG, and TGAAAAC, occur at 534, 376, 364, and 331 bp upstream from codon 1. We do not know their significance. The transcription units of the genes from the start of transcription to the poly(A) attachment site extend 3714 bp in PRH1 and 3578 bp in PRH2.

A proline-rich protein gene, MPz, from a mouse has recently been sequenced (Ann and Carlson, 1985). This mouse PRP, protein C, determined by Wong and Bennick (1980), is the NH2-terminal amino acid of the secreted protein as judged by the amino acid sequence of the acidic PRP genes is a transposon of length 497 or 496 bp with long terminal repeats of 142 or, less likely, 126 bp. However, genomic Southern blot hybridization of human DNA to a probe made from most of this region from PRH1 or to a probe containing the whole 5'-terminal repeat failed to detect any other copies of the sequence in the genome (data not shown), nor were we able to find any homologous sequences in the GenBank library (version of February 22, 1985). Thus, there is no evidence from the presence of other copies to support the idea that the region is a transposon.

A comparable set of sequences occurs in the PRH2 gene except that there appears to have been a deletion of 119 nucleotides in the 3' end of the second repeat. Direct repeats of this type are frequently associated with transposable elements (Calos and Miller, 1980), and it is possible that this region of the PRP genes is a transposon of length 497 or 496 bp with long terminal repeats of 142 or, less likely, 126 bp. However, genomic Southern blot hybridization of human DNA to a probe made from most of this region from PRH1 or to a probe containing the whole 5' -terminal repeat failed to detect any other copies of the sequence in the genome (data not shown), nor were we able to find any homologous sequences in the GenBank library (version of February 22, 1985). Thus, there is no evidence from the presence of other copies to support the idea that the region is a transposon.

A pair of 72-bp direct repeats having 93% sequence identity with each other are present in the PRH1 gene 1429 and 1357 bp upstream from exon 1. The equivalent region of PRH2 has not been sequenced. This 72-bp tandem repeated sequence in the PRH1 gene has only limited similarity (35% identity) to the enhancer element in SV40 gene (Benoiist and Chambon, 1981; Gruss et al., 1981), but the sequences TGGAAC and CAAACCA, which occur within the PRH1 repeats, are identical to virus enhancer core sequences. This suggests that the 72-bp tandem repeats in the PRH1 gene may be important for its transcription.

---

**Fig. 3. Diagrammatic comparison of PRH1 and PRH2 genes.** Upper part, the nucleotide sequences of the two aligned genes are represented by horizontal lines with their exons shown by black bars. Nucleotide differences between the genes are shown by long vertical lines between the horizontal lines. Length differences are shown by short vertical lines between the two horizontal lines, as well as by a second small vertical line close to the gene with fewer nucleotides. Lower part, the percentage differences between different parts of the PRH1 and PRH2 genes are illustrated by a bar diagram with the relevant percentages indicated. In each exon (E1, E2, E3, and E4), synonymous differences are indicated by stippled bars, and nonsynonymous substitutions are shown by black bars.
a) Hae III Type Repeats

1
CCTCCTCAgGAAAgGCCACCAAGGACCACCCCAACACcAGGAGGGGCACTCCTCCCCCCT

2
CCTCCTCAAGGAAgGCCACCAAGGACCACCCCAACACcAGGAGGGGCACTCCTCCCCCCT

3
CCTCCTCAAGGAAgGCCACCAAGGACCACCCCAACACcAGGAGGGGCACTCCTCCCCCCT

4
CCTCCTCaAGGAAgGCCACCAAGGACCACCCCAACACcAGGAGGGGCACTCCTCCCCCCT

5
CCTCCTCaAGGAAgGCCACCAAGGACCACCCCAACACcAGGAGGGGCACTCCTCCCCCCT

b) Consensus Sequences

Hae III
CCTCCTCAgGAAAgGCCACCAAGGACCACCCCAACACcAGGAGGGGCACTCCTCCCCCCT

Bst NI
CCTCCTCAAGGAAgGCCACCAAGGACCACCCCAACACcAGGAGGGGCACTCCTCCCCCCT

c) Other Species

Rat pRP33
CCcCCaAgGGGaggCCcCAGcagagCCCCCTCACGGccGGaAaAcCaCAAGGc........CCA

Mouse MP2
CcCaCCaCCAgGGGgCCACAGcgcAGCGgGACCCCCCT.................CAAGGc......CCA

Fig. 4. Repeated sequences in coding exon 3. a, The five tandemly repeated sequences that occur in the PRH1 and PRH2 genes (both genes are identical in this region). Bases which differ from the consensus sequence of HaeIII-type repeats are indicated by small letters. Asterisks denote nucleotides different from the consensus sequence of the HaeIII-type repeat but identical to that of the BstNI-type repeat. b, The HaeIII- and BstNI-type consensus sequences. Bases which differ between the two consensus sequences are shown by asterisks. The HaeIII and BstNI recognition sites are underlined. c, consensus sequences of repeats from a rat cDNA clone, pRP33, and a mouse gene, MP2. Bases which differ from the human consensus sequences are shown by small letters.

Uneven Distribution of Sequence Differences between Two Genes—The differences between the nucleotide sequences of the two HaeIII-type genes are diagrammatically summarized in Fig. 3 in which the two sequences are homologously aligned with their exons indicated by solid black bars. The overall difference between the two genes is 8.7% when each base mismatch and each gap is counted as one difference (4031 bp at its 5' end, whereas the second and third repeated sequences are 12 bp shorter near their 3' ends). 63% of all positions are identical in all five repeated sequences. A consensus sequence for the HaeIII-type PRP repeat has been constructed from these five repeated sequences and is compared in Fig. 4c to the consensus sequence of a BstNI-type PRP repeat (Azen et al., 1984). The total difference between the consensus sequences of these two gene types is 19%, which suggests that the duplication leading from a single human salivary PRP ancestral gene to the present HaeIII- and BstNI-type genes is not recent. When the individual HaeIII-type repeats in the two genes are compared to the consensus sequence, more differences are found in the first and the fifth repeats. Conceivably, the middle three repeats have been more homogenized by gene conversions and unequal crossing-over events between the repeats than have the outer repeats.

HaeIII-type Repeats—The repeated region in exon 3, consisting of five tandem repeats of 63 bp, is completely identical in PRH1 and PRH2, as discussed in the preceding paragraph. The five repeated sequences are aligned to show their relatedness in Fig. 4a. Each repeat has the common HaeIII site, GGCC, which translates to Gly-X. The first repeat lacks 15 bp at its 5' end, whereas the second and third repeated sequences are 12 bp shorter near their 3' ends. 63% of all positions are identical in all five repeated sequences. A consensus sequence for the HaeIII-type PRP repeat has been constructed from these five repeated sequences and is compared in Fig. 4b to the consensus sequence of a BstNI-type PRP repeat (Azen et al., 1984). The total difference between the consensus sequences of these two gene types is 19%, which suggests that the duplication leading from a single human salivary PRP ancestral gene to the present HaeIII- and BstNI-type genes is not recent. When the individual HaeIII-type repeats in the two genes are compared to the consensus sequence, more differences are found in the first and the fifth repeats. Conceivably, the middle three repeats have been more homogenized by gene conversions and unequal crossing-over events between the repeats than have the outer repeats.

Proline-rich proteins have been found in the saliva of other mammals. A cDNA clone, pRP33, of a rat contains six repeats of 19 amino acids (Ziemer et al., 1984). A mouse PRP gene, MP2, contains 13 repeats of 14 amino acids (Ann and Carlson, 1985). The consensus nucleotide sequences of the repeats in these genes are shown in Fig. 4c for comparison. Clearly, the repeats are all related and derived from a common ancestor.
The two HaeIII- and BstNI-type repeats of human are, however, more related to each other than either is to rat or mouse repeats. This suggests that the divergence of HaeIII- and BstNI-type repeats may be more recent than the divergence of human and rodents. Alternatively, the human genes may have evolved in a concerted fashion. We are in the process of sequencing other human PRP genes, as well as investigating the organization of the six human PRP loci, in order to understand better the evolution of this gene family.

In conclusion, our studies on the PRH1 and PRH2 genes show that the evolution of these genes is more complicated than simple gene duplication and divergence and suggest the occurrence of recombinational events between genes in the family, as well as between the repeats within each gene.

Acknowledgments—We gratefully thank Dr. Oliver Smithes for his support and invaluable advice. We also thank Dr. Edwin A. Azen, Dr. Patricia A. Goodman, and Karen M. Lyons for helpful discussions and Natalie Borenstein, Susan McEvoy, and Tom Gettelman for their technical assistance.

REFERENCES

Azen, E. A., Goodman, P. A., and Lalley, P. A. (1985) Am. J. Hum. Genet. 37, 418–424
Bennick, A. (1982) Mol. Cell. Biochem. 45, 83–99
Benoist, C., and Chambon, P. (1981) Nature 290, 304–310
Calos, M. P., and Miller, J. H. (1980) Cell 20, 579–585
Devereux, J., Haeberli, P., and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395
Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forges, B. C., Weissman, S. M., Slightom, J. L., Blech, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., and Proudfoot, N. J. (1980) Cell 21, 653–668
Fitzgerald, M., and Shenk, T. (1981) Cell 24, 251–260
Friedman, R. D., and Merrit, A. D. (1975) Am. J. Hum. Genet. 27, 304–314
Goldberg, M. (1979) Ph.D. thesis, Stanford University
Gruss, P., Dhar, R., and Khoury, G. (1981) Proc. Natl. Acad. Sci. U. S. A. 78, 943–947
Hohn, B. (1979) Methods Enzymol. 68, 299–309
Khoury, G., and Gruss, P. (1983) Cell 33, 313–314
Loenen, W. A. M., and Blattner, F. R. (1983) Gene (Amst.) 26, 171–179
Maeda, N. (1985) Biochem. Genet. 23, 455–464
Maeda, N., Kim, H.-S., Azen, E. A., and Smithies, O. (1985) J. Biol. Chem. 260, 1123–1130
Maxam, A. M., and Gilbert, W. (1977) Proc. Natl. Acad. Sci. U. S. A. 74, 560–564
Mount, S. (1982) Nucleic Acids Res. 10, 459–472
Rothstein, R. J., Lau, L. F., Bahl, C. P., Nayang, S. A., and Wu, R. (1979) Methods Enzymol. 68, 98–109
Slightom, J. L., Blech, A. E., and Smithies, O. (1980) Cell 21, 627–638
Southern, E. M. (1975) J. Mol. Biol. 98, 503–517
Wong, R. S. C., and Bennick, A. (1990) J. Biol. Chem. 255, 5943–5948
Ziemer, M. A., Swain, W. F., Rutter, W. J., Clements, S., Ann, D. K., and Carlson, D. M. (1984) J. Biol. Chem. 259, 10478–10489