Many Protein Products From a Few Loci: Assignment of Human Salivary Proline-Rich Proteins to Specific Loci

Karen M. Lyons,* Edwin A. Azen,* Patricia A. Goodman† and Oliver Smithies*

**Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706, and †Department of Neurology, University of California, San Francisco, California 94143*

Manuscript received January 14, 1988
Revised copy accepted May 12, 1988

ABSTRACT

Earlier studies of protein polymorphisms led to the description of 13 linked loci thought to encode the human salivary proline-rich proteins (PRPs). However, more recent studies at the DNA level have shown that there are only six genes which encode PRPs. The present study was undertaken in order to reconcile these observations. Nucleotide and decoded amino acid sequences from each of the six genes were compared with the available protein sequence data for PRPs. This analysis allowed assignment of the PmF, PmS and Pe proteins to the *PRB1* locus, the G1 protein to the *PRB3* locus, the Po protein to the *PRB4* locus, the Ps protein to the *PRB2* locus, and the CON1 and CON2 proteins to the *PRB4* locus. Correlations between insertion/deletion RFLPs and PRP protein phenotypes were observed for the PmF, PmS, Gl and CON2 proteins. Our overall analysis indicates that in many instances several proteins previously considered to be the products of separate loci are actually proteolytic cleavage products of a large precursor specified by one or other of the six genes identified at the DNA level. Our analysis also demonstrates that some of the "null" alleles proposed to occur at 11 of the 13 loci in the earlier genetic studies, are actually productive alleles having alterations at proteolytic cleavage sites within the relevant precursor protein. The absence of cleavage leads to the persistence of longer precursor peptides not resolved electrophoretically, concurrently with an absence of the smaller PRPs seen when cleavage occurs.

THE human proline-rich proteins (PRPs) are a heterogeneous group of more than 20 proteins that constitute approximately 70% of the protein content of human saliva. They are characterized by an abundance of the amino acids proline (25–42%), glycine (16–22%) and glutamic acid/glutamine (15–28%), which together make up 70–88% of their total amino acid content. PRPs have been classified as either acidic, basic, or glycosylated. They all contain a series of proline-rich tandem repeats 16–21 amino acids in length (reviewed by BENNICK 1987).

Electrophoretic studies of polymorphic PRPs led to the description of 13 linked loci, each of which encoded one of these polymorphic PRPs (reviewed by AZEN and MAEDA 1988). Linkage studies indicated that the loci covered a genetic distance of 15 cM, suggesting that either the PRP gene cluster spanned a very large physical distance, or that recombination frequently occurs within the multigene family (GOODMAN *et al.* 1985). The loci proposed to encode acidic PRPs were *Pr* (AZEN and OPPENHEIM 1973; AZEN and DENNISTON 1974), *Pa* (FREIDMAN, MERRITT and RIVAS 1975; AZEN 1977), *Db* (AZEN and DENNISTON 1974) and *PIF* (AZEN and DENNISTON 1981). The loci proposed to encode basic PRPs were *PmF* (IKEMOTO *et al.* 1977), *PmS* (AZEN and DENNISTON 1980; ANDERSON, KAUFFMAN and KELLER 1982), *Ps* (AZEN and DENNISTON 1980), *Pc* (KARN, GOODMAN and YU 1985), *Po* (AZEN and YU 1984a), and *Pe* (AZEN and YU 1984a). The glycosylated PRPs were proposed to be encoded by three loci: *CON1* (AZEN and YU 1984b), *CON2* (AZEN and YU 1984b), and *Gl* (AZEN, HURLEY and DENNISTON 1979). Null alleles, characterized by the absence of the corresponding PRP on electrophoresis gels, were postulated for eleven of the thirteen loci. Null alleles were not described for the *Pr* and *Pc* loci.

DNA clones corresponding to members of the human PRP multigene family were isolated by AZEN *et al.* (1984). Subsequent DNA studies (MAEDA 1985; Maeda *et al.* 1985) suggested that the human PRP multigene family contains only six genes rather than the thirteen proposed in the earlier genetic studies. A determination of the physical organization of the PRP gene complex (H.-S. KIM, unpublished observation) has confirmed that there are only six genes encoding PRPs.

The DNA studies also revealed that the PRP multigene family can be divided into two subfamilies based on the DNA sequences of the tandem repeats that comprise the third exons of these genes: *PRH1* and *PRH2* form one subfamily (the *PRH* genes) and encode acidic PRPs (KIM and MAEDA 1986; AZEN *et al.* 1987); *PRB1*, *PRB2*, *PRB3* and *PRB4* (the *PRB* genes)

form the second subfamily (MAEDA 1985). Length variants, caused by insertions or deletions of DNA, have been identified at each of the *PRB* loci (O'CONNELL *et al.* 1987; LYONS, STEIN and SMITHIES 1988).

MAEDA *et al.* (1985) demonstrated by cDNA studies that differential mRNA splicing and proteolytic processing of products of the PRP loci can potentially generate multiple PRPs from a single transcription unit, and BENNICK (1987) has compared the available protein sequences with the decoded amino acid sequences for these cDNAs. While both of these studies show that a single locus can encode multiple PRPs, they only allow a partial reconciliation of the number of loci proposed in the genetic studies and the number demonstrated in DNA studies. For example, the discrepancies between the proposed mode of inheritance of the acidic PRPs and the DNA studies were resolved by MAEDA (1985), and the extension of these ideas to the basic and glycosylated PRPs was hypothesized; she suggested that the electrophoretic data, originally used to support the existence of the four loci *Pr*, *Db*, *Pa* and *PIF*, could be interpreted more economically in terms of two loci with no null alleles. This reinterpretation has since been confirmed by DNA studies (KIM and MAEDA 1986; AZEN *et al.*, 1987) which demonstrate that *PRH1* encodes the Pa, Db and PIF proteins and *PRH2* encodes the Pr protein.

In our present study, we have reexamined the inheritance of the basic and glycosylated PRPs in order to further resolve the discrepancy between the results of the original genetic studies and the DNA data. By comparing the amino acid sequences decoded from the DNA sequences of the four *PRB* genes with the available protein sequences, and by comparing the segregation of allelic DNA length variants with the segregation of PRP proteins in family members and in unrelated individuals, we have been able to assign almost every basic and glycosylated polymorphic PRP to a particular locus. An essentially complete outline of the genetic and phenotypic complexities of the PRP system is generated by this analysis.

MATERIALS AND METHODS

Electrophoretic typing of PRPs in saliva: Parotid saliva samples were collected as described by AZEN and DENNISTON (1974). Ps, PmF and PmS protein polymorphisms were typed by electrophoresis in acid-lactate polyacrylamide gels stained with Coomassie brilliant blue R-250 (AZEN and DENNISTON 1980). The Po protein polymorphism was typed on immunoblots treated with anti-Ps or anti-Pr serum as described by AZEN and YU (1984a). The Gl protein polymorphism was typed by electrophoresis in acid-lactate polyacrylamide gels stained with periodic acid-Schiff as described by AZEN, HURLEY and DENNISTON (1979). CON1 and CON2 protein polymorphisms were typed by electrophoresis in SDS gels with a concanavalin A stain as described by AZEN and YU (1984b).

Southern blot hybridizations: High molecular weight DNA was prepared from peripheral blood leukocytes by the method of PONCZ *et al.* (1982). Genomic DNA (5 μg per lane) was digested with *Eco*RI, electrophoresed in 0.8% agarose gels, and transferred to nitrocellulose according to the method of SOUTHERN (1975) with modifications as described by WAHL, STERN and STARK (1979). Filters were hybridized to a nick-translated 980 bp *Hin*fI fragment from *PRB1* (AZEN *et al.* 1984). Hybridization conditions were as described by VANIN *et al.* (1983).

RESULTS AND DISCUSSION

Overall assignments: Our general strategy for assigning the basic and glycosylated PRPs to the *PRB* loci has been to compare the decoded amino acid sequences derived from the nucleotide sequences of the third exons of various alleles of these genes (LYONS, STEIN and SMITHIES 1988; MAEDA *et al.* 1985) with protein sequences determined by other investigators (KAUFMAN *et al.* 1982, 1986; KAUFFMAN and KELLER 1979; SAITOH, ISEMURA and SANADA 1983a). This strategy is practical because the third exon contains nearly all of the protein coding portion of PRP genes (AZEN *et al.* 1984). Where protein sequence data are incomplete, we have used small peptide sequences (SHIMOMURA, KANAI and SANADA 1983) or data on amino acid composition (GOODMAN *et al.* 1985). We have also used electrophoretic comparisons of purified PRPs with the polymorphic PRPs (AZEN 1988). Finally, we compared the segregation in families of DNA length variants detected by Southern blot analysis with the segregation of PRP variants detected by electrophoresis.

In order to facilitate the presentation in the following sections of the detailed considerations on which our assignments are based, we first present in Figure 1 a summary of our overall conclusions. For the sake of completeness we have included in the figure the assignments of the acidic proteins Pa, Pr, PIF and Db as proposed by MAEDA (1985) and confirmed by AZEN *et al.* (1987). All six PRP loci are therefore illustrated in the figure along with specific alleles for each locus, and the proteins that are the products of these specific alleles.

Reexamination of the inheritance of the basic proline-rich proteins: Polymorphisms of the basic PRPs, PmS, PmF, Pe and Po had been interpreted as being due to the autosomal inheritance of one expressed and one unexpressed allele at each of four loci. The basic PRP Pc was ascribed to the autosomal inheritance of two expressed alleles at a fifth locus (KARN, GOODMAN and YU 1985). A sixth locus with two expressed alleles and one unexpressed allele was proposed to control the expression of one of the basic PRPs, Ps (AZEN and DENNISTON 1980).

The *PRB1* locus encodes the PmF and PmS proteins: The amino acid sequences of the PmF and PmS proteins were determined by KAUFFMAN *et al.* (1982 and 1986, respectively). The decoded amino acid

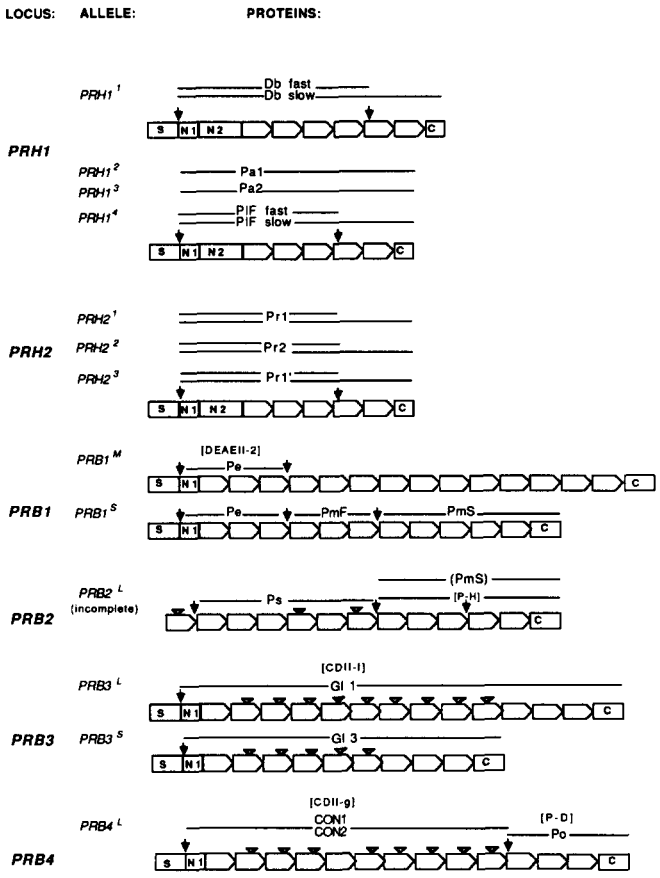

FIGURE 1.—Summary of assignments of specific proline-rich proteins to the six PRP loci. The names of specific loci are shown in boldface capitals, specific alleles in lightface capitals, encoded proteins in light type. The names of encoded proteins or small peptides with sequence identity to the proteins indicated (see text) are enclosed in small square brackets. The coding sequences of specific alleles are illustrated schematically as arrays of open boxes and open arrows. Small vertical arrows indicate potential proteolytic cleavage sites. The boxes labeled "S" represent a signal peptide. The boxes labeled "N1" and "N2" represent amino-terminal sequences. Each open arrow represents one proline-rich tandem repeat. The box labeled "C" represents the carboxyl-terminal region. Stippled triangles indicate *N*-linked glycosylation sites. The allele of *PRB2* presented in the figure is an incomplete cDNA copy of a transcript derived from $PRB2^L$ (MAEDA *et al.* 1985). The localization of the coding regions for each protein are shown by lines. Where partial proteolytic cleavage is known to occur, the lengths of both protein products are shown, *e.g.*, Db fast is coded by the $PRH1^1$ allele of the *PRH1* locus by N1, N2, and 4¼ repeats, Db slow by N1, N2, and 6 repeats plus C. The tentative assignment of the PmS protein (in parentheses) in $PRB2^L$ requires that the rightmost proteolytic cleavage site indicated in $PRB2^L$ is not functional.

sequences for the third exons of two alleles at the *PRB1* locus, $PRB1^M$ and $PRB1^S$, were determined by LYONS, STEIN and SMITHIES (1988). These alleles were cloned from a female whose saliva was typed electrophoretically as PmS^+PmF^+. The segregation of the PmS and PmF proteins in her family members indicates that she must carry one chromosome encoding a PmF^+PmS^+ phenotype and one chromosome encoding a PmF^-PmS^- phenotype. The association studies described in the next section indicate that the $PRB1^S$ allele is associated with the PmF^+PmS^+ phenotype and the $PRB1^M$ allele is associated with the PmF^-PmS^- phenotype.

Figure 2A shows that a portion of the decoded amino acid sequence of the allele $PRB1^S$, amino acid residues 59 to 119, is identical to the protein sequence determined for PmF (KAUFMAN *et al.* 1982). We therefore conclude that, as illustrated in Figure 1, part of the allele $PRB1^S$ codes for PmF. Figure 2A also shows that amino acid residues 120 to 237 coded by $PRB1^S$ comprise a second decoded amino acid sequence which is identical to the amino acid sequence of PmS (KAUFMAN *et al.* 1986) except for a single substitution. The PmS protein sequence has a glycine at amino acid position 123 whereas the $PRB1^S$ allele encodes an arginine at this position; this difference is unlikely to result in a PmS^- phenotype. We therefore conclude that a different part of the allele $PRB1^S$ encodes PmS, as illustrated in Figure 1.

For these conclusions to be valid, it is necessary that the junction between the decoded amino acid sequences encoding the PmF and PmS peptides (between amino acid residues 119 and 120) in $PRB1^S$ include a proteolytic cleavage site for a salivary protease. Although most characterized sites of post-translational proteolytic cleavage are a sequence of two basic amino acids (BOND and BUTLER 1987), SCHWARTZ (1986) has shown that a single arginine residue, often preceded by a proline can serve as a site of proteolytic processing.

SCHWARTZ (1986) pointed out that the general structure basic-x-x-arginine is common to a variety of monobasic proteolytic cleavage sites. He also noted that charged and polar amino acids are often found near the site of proteolytic cleavage. Our data are well explained by the assumption that the related structure, arginine-serine-x-arginine-serine serves to generate potential proteolytic cleavage sites in the human PRPs, with cleavage occurring after the last arginine residue. This structure contains a basic amino acid (arginine) three residues upstream of the site of proteolytic cleavage, an arginine immediately upstream of the site of cleavage, and polar amino acids (serine residues) near the potential proteolytic cleavage site, features discussed by SCHWARTZ (1986) as common elements of monobasic proteolytic cleavage sites. We cannot determine whether all amino acid sequences that contain the above structure have functional proteolytic cleavage sites in PRPs *in vivo*. However, comparison of the amino acid sequences decoded from the DNA sequences with the actual amino acid sequences of PRP proteins determined by other investigators, as discussed in the following sections, shows that all of the proteolytic products of the PRPs characterized to date conform to this rule.

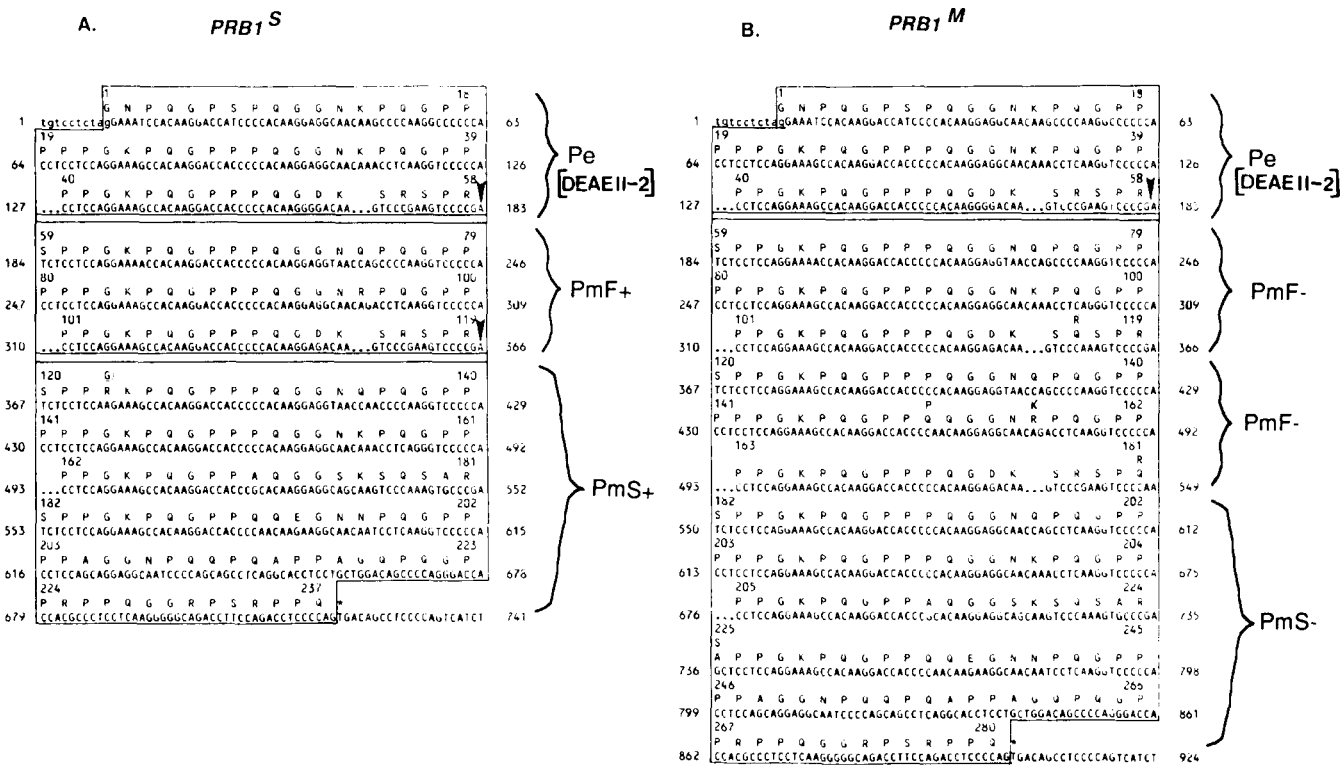

FIGURE 2.—Nucleotide and decoded amino acid sequences from the third exons of (A) *PRB1*^{S} and (B) *PRB1*^{M} and comparison to the PmF, PmS and DEAEII-2 protein sequences determined by KAUFFMAN *et al.* (1982 and 1986, respectively). Lower case letters represent nucleotides in the intron preceding exon 3. Upper case letters represent nucleotides in exon 3. The translated portion of exon 3 is boxed, and the decoded amino acid sequence of this region is presented above the corresponding nucleotide sequence. Gaps have been introduced into the nucleotide sequences in order to permit alignment of tandem repeats. The stop codon is indicated by an asterisk. Potential proteolytic cleavage sites are indicated by vertical arrows. Each peptide potentially produced by proteolytic cleavage is included in a separate box. The portions of the decoded amino acid sequence which are most similar to the PmF, PmS and Pe protein sequences are indicated by brackets to the right of the sequence. The Pe protein was assigned based on its electrophoretic identity to the DEAEII-2 peptide. In the positions where the decoded amino acid sequences and the protein sequences determined by other investigators differ, the amino acid present in the protein sequence is presented circled above the corresponding decoded amino acid.

Figure 2A illustrates the relevance of this assumption. Only the two proteolytic cleavage sites indicated in *PRB1*^{S} in Figure 2A contain the structure arginine-serine-x-arginine-serine. Thus the PmF and PmS proteins are readily accounted for as being the products of a single locus specifying a precursor protein which is then cleaved by a protease that recognizes such monobasic sites.

The decoded amino acid sequence of the allele *PRB1*^{M} is presented in Figure 2B. Although this allele is associated with the PmF^{-}PmS^{-} phenotype, it also contains two regions closely related to the protein sequence for PmF (KAUFFMAN *et al.* 1982). The first region occurs, as it does in the allele *PRB1*^{S}, from amino acid positions 59 to 119. However, amino acid residue 116 in *PRB1*^{M} is glutamine, whereas it is arginine in *PRB1*^{S} and in the protein sequence determined by KAUFFMAN *et al.* (1982). The loss of this arginine residue destroys the arginine-serine-x-arginine-serine motif and, we suggest, thereby greatly decreases or abolishes the proteolytic cleavage necessary to generate the PmF^{+} phenotype.

The second region in *PRB1*^{M} encoding a PmF-like peptide extends from amino acid residues 120–181, but it contains three amino acids differing from the PmF protein sequence. The first, at residue 151 in *PRB1*^{M}, is a glutamic acid in the decoded sequence in place of a proline present in the sequence determined by KAUFFMAN *et al.* (1982). The second, at residue 156, is an arginine in place of a lysine. The third difference, at amino acid position 181, is a glutamine in place of an arginine. Arginine 181 is in the position associated with proteolytic cleavage, and its absence in *PRB1*^{M} is, we suggest, again consistent with a PmF^{-} phenotype, which this allele produces.

The *PRB1*^{M} allele also contains a region with a decoded amino acid sequence identical to that of the PmS protein sequence determined by KAUFFMAN *et al.* (1986) except for a single amino acid difference; the decoded amino acid sequence contains an alanine at amino acid position 225, whereas a serine occurs at the corresponding position in the PmS protein sequence. This amino acid substitution would not be expected to result in a PmS^{-} phenotype. However,

the other substitutions in $PRB1^M$, at amino acid positions 116 and 181, alter two potential proteolytic processing sites so that the final product of the $PRB1^M$ allele is a peptide so much larger than PmS that it would be classified as a PmS^- phenotype.

The Pe protein is encoded by the *PRB1* locus: The Pe protein polymorphism has been described in terms of one expressed (Pe^+) and one unexpressed (Pe^-) autosomal allele (AZEN and YU 1984a). AZEN *et al.* (1987) showed that the protein DEAEII-2 (KAUFFMAN and KELLER 1979) has identical electrophoretic properties to the Pe protein and proposed that DEAEII-2 and Pe are identical. BENNICK (1987) has reported that a decoded cDNA sequence derived from *PRB1* (MAEDA *et al.* 1985), contains a region identical to the partial sequence of DEAEII-2. The region of identity includes an amino terminal portion encoded by exons 1 and 2 plus residues 1–45 from exon 3 (Figure 2). A potential proteolytic cleavage site occurs at amino acid position 58, suggesting that the complete sequence of DEAEII-2 will extend to this site as illustrated in Figure 1. Thus, the *PRB1* locus encodes the Pe protein as well as the PmS and PmF proteins. The individual from whom the $PRB1^M$ and $PRB1^S$ alleles were cloned has a Pe^+ phenotype, which is consistent with this assignment.

Associations between DNA length variants of the *PRB1* locus and PmF plus PmS protein phenotypes: Studies of the inheritance of the basic PRPs had revealed a strong correlation between the presence or absence of the PmS and PmF proteins, suggesting that these two proteins are inherited as a unit, with PmF^+S^+ and PmF^-S^- being the most common phenotypes. Occasionally the phenotype PmF^+S^- is seen, but the PmF^-S^+ phenotype has not been observed (AZEN and DENNISTON 1980). The above comparisons of the decoded amino acid sequences with the determined protein sequences for PmS and PmF suggest that some alleles at the *PRB1* locus encode the PmF and PmS proteins, but that others do not. We have therefore compared the PmF and PmS protein phenotypes with the presence of specific alleles of the *PRB1* locus.

The segregation of three allelic length variants of *PRB1* ($PRB1^S$, $PRB1^M$ and $PRB1^L$) were compared with the segregation of the PmS and PmF protein phenotypes in three families (20 individuals). The comparison reveals a complete correlation between the presence of the $PRB1^S$ allele and the PmF^+PmS^+ phenotype. An example is presented in Figure 3. In this family, and in two others examined, we found a complete correlation between the presence of the $PRB1^S$ allele and the PmF^+PmS^+ phenotype.

Table 1 presents the results from a similar comparison of 24 unrelated individuals. All six individuals with the PmF^+PmS^+ phenotype carry the $PRB1^S$ allele, and no individual who carries $PRB1^S$ was typed as either PmF^- or PmS^-. No other PRP locus displays this absolute correlation between PmS and PmF protein phenotype and segregation of one of its length variants.

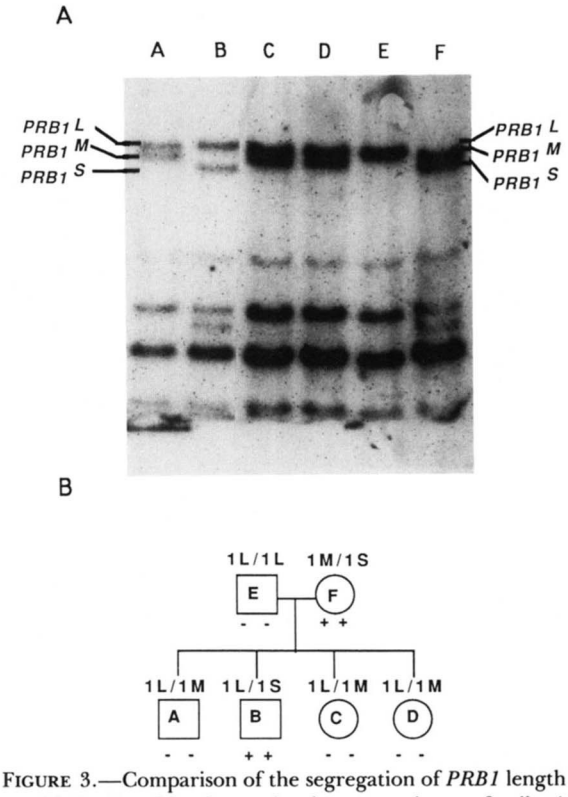

FIGURE 3.—Comparison of the segregation of *PRB1* length variants with PmF and PmS protein phenotypes in one family. (A) A Southern blot of *Eco*RI-digested DNA samples hybridized to the 980 bp *Hin*fI fragment from *PRB1*. Three alleles of *PRB1*, giving rise to bands labeled as 1L, 1M or 1S, are segregating in this family. The unmarked bands correspond to other PRP loci. (B) Pedigree of the family in Figure 3A. The *PRB1* alleles present in each individual are indicated above the circle or square. The PmF and PmS protein phenotypes are indicated below the circle or square.

TABLE 1

Comparison of PmF and PmS protein phenotypes with PRB1 length variants in 24 unrelated individuals

PRB1 allele(s)[a]	Protein phenotype		No. of individuals
	PmF	PmS	
1L/1L	−	−	3
1L/1M	−	−	6
1L/1M	+	−	1
1L/1S	+	+	1
1M/1M	−	−	7
1M/1M	+	−	1
1M/1S	+	+	5

[a] 1L = $PRB1^L$, 1M = $PRB1^M$, 1S = $PRB1^S$.

Two individuals in Table 1 were typed electrophoretically as PmF^+PmS^-. Both of these individuals carry the $PRB1^M$ allele. However, 13 other individuals carrying at least one copy of the $PRB1^M$ allele have a PmF^-PmS^- protein phenotype. This result suggests

that the *PRB1*^{M} length variant represents at least two alleles, one that encodes the PmF^{+}PmS^{-} phenotype, and one that encodes a PmF^{-}PmS^{-} phenotype. A comparison between protein phenotypes and the segregation of *PRB1* alleles in a family of seven segregating the PmF^{+}PmS^{-} phenotype (data not shown) shows a complete correlation between the PmF^{+}PmS^{-} phenotype and the presence of the *PRB1*^{M} allele. Segregation analysis of PmF plus PmS protein phenotypes and *PRB1* length variants indicates that the *PRB1*^{M} allele presented in Figure 2B encodes a PmF^{-}PmS^{-} phenotype in agreement with the argument already presented. A *PRB1*^{M} allele encoding a PmF^{+}PmS^{-} phenotype could potentially be generated by a single nucleotide change leading to an arginine at position 116 in place of the glutamine encoded by the allele in Figure 2B. This single substitution would be expected to generate a functional proteolytic cleavage site, resulting in a PmF^{+}PmS^{-} phenotype. The *PRB1*^{L} allele is associated with a PmF^{-}PmS^{-} phenotype in these comparisons. The DNA sequence of this allele has not been determined, but we suspect that the PmF^{-}PmS^{-} phenotype encoded by this allele is a result of alterations at proteolytic cleavage sites that reduce or abolish post-translational cleavage. There appears to be no *a priori* reason for the absence of any alleles encoding a PmF^{-}PmS^{+} phenotype; a more extensive search might reveal their existence. For example, a single change at position 181 in *PRB1*^{M} that results in the formation of a functional proteolytic cleavage site would generate a PmF^{-}PmS^{+} phenotype.

The PmS protein may also be encoded by the *PRB2* locus: We have examined the decoded amino acid sequences from other PRP loci for homology to the PmS and PmF protein sequences. The decoded amino acid sequence of an incomplete cDNA copy of the *PRB2* locus (MAEDA *et al.* 1985) is presented in Figure 4. The three regions in *PRB2*^{L} that contain the sequence arginine-serine-x-arginine-serine and thus serve as potential proteolytic cleavage sites are indicated in the figure. Residues 134–251 in *PRB2* have only two differences from the sequence for PmS determined by KAUFFMAN *et al.* (1986). The first difference, at residue 185, results in the substitution of a proline for the alanine present in the PmS protein sequence. The second substitution, at residue 192, results in replacement of a glutamine present in the PmS protein sequence with an arginine residue. This substitution is likely to result in production of a functional proteolytic cleavage site within the region of identity with the PmS protein because it places a basic residue three positions in front of a potential proteolytic cleavage site. We conclude that the allele of *PRB2* presented in Figure 4 probably does not encode a PmS protein because of this alteration, but we suggest that other unsequenced alleles of *PRB2* might have

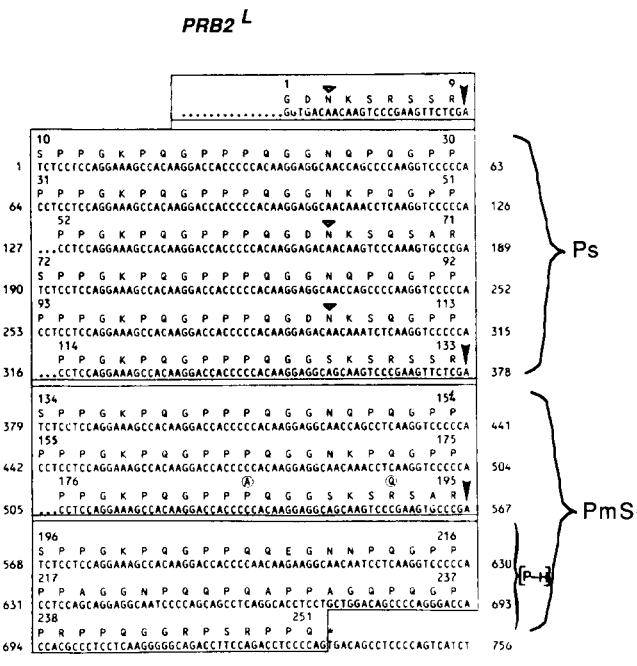

FIGURE 4.—Nucleotide and decoded amino acid sequences from the third exon of a truncated cDNA clone derived from *PRB2*^{L} (MAEDA *et al.* 1985). Coding sequences and potential proteolytic cleavage sites are identified as described in the legends to Figures 1 and 3. Potential *N*-linked glycosylation sites are identified by triangles. The region spanned by amino acids 196 through 251 is identical to the P-H peptide (SAITOH, ISEMURA and SANADA, 1983a) (see text).

amino acid differences which would abolish this proteolytic cleavage site, and so lead to the PmS^{+} phenotype.

The Ps protein is encoded by the *PRB2* locus: Amino acid sequence data are not available for the Ps proteins, the proposed products of the *Ps* locus (AZEN and DENNISTON 1980), although the amino acid compositions of the Ps1 and Ps2 allelic proteins, determined by GOODMAN *et al.* (1985), are very similar to those of the PmF and PmS proteins and are distinct from those of the acidic and glycosylated PRPs. In addition, AZEN and DENNISTON (1980) have demonstrated that the Ps protein can be stained faintly for carbohydrate with the periodic acid-Schiff reagent. Comparison of the decoded amino acid sequences from the *PRB1* and *PRB2* loci (Figures 2 and 4) demonstrates that they encode peptides with amino acid sequences very similar to each other. However, the decoded amino acid sequence of *PRB2* contains several asparagine-lysine-serine tripeptides, which conform to the amino acid sequence asparagine-x-serine recognized as a site for *N*-linked sugar addition, whereas the decoded amino acid sequence of *PRB1* does not. The decoded amino acid sequence of the truncated cDNA clone from *PRB2* contains three potential proteolytic cleavage sites, which suggests that up to four peptides are encoded by *PRB2*. Two of these peptides contain *N*-linked glycosylation sites.

Amino acids 10 through 133 of the decoded sequence of *PRB2*^{L} comprise a glycosylated peptide with an amino acid composition nearly identical to that determined for the Ps proteins (GOODMAN *et al.* 1985). The individual from whom this *PRB2* clone was isolated produces the Ps1 protein. Consequently, we suggest that this region, corresponding to amino acid residues 10 through 133, encodes the glycosylated Ps1 protein. Amino acid residues 196 through 251 comprise a peptide that is identical to the P-H peptide sequenced by SAITOH, ISEMURA and SANADA (1983a). The P-H peptide has not been correlated with any of the polymorphic PRPs described in genetic studies (AZEN 1988).

The segregation of *PRB2* length variants and Ps proteins could not be compared because there were no *PRB2* length variants in our families. Furthermore, only one individual out of our random sample of 25 carried an allelic length variant for *PRB2*. Thus, although our data do not allow detection of potential correlations between the Ps protein phenotype and the *PRB2* gene, the nearly identical amino acid compositions of Ps and the *PRB2*-encoded peptide, plus the presence of potential glycosylation sites in the *PRB2*-encoded peptide provide strong support for our assignment.

The Po protein is encoded by the *PRB4* locus: Previous genetic studies described one expressed (*Po*$^{+}$) and one unexpressed (*Po*$^{-}$) allele at the *Po* locus (AZEN and YU 1984b). A basic proline-rich peptide, P-D, sequenced by SAITOH, ISEMURA and SANADA (1983b), has been suggested to be a product of the *Po* locus based on the identical mobilities of the P-D and Po proteins in several electrophoretic gel systems (AZEN 1988). Figure 5 presents the decoded amino acid sequence for the third exon of the allele *PRB4*^{L} and compares it to the sequence of the P-D protein. They are identical from amino acid residues 208–277. Proteolytic cleavage must occur at amino acid position 207 in order to result in production of the Po protein. This site is reasonable since it contains the structure arginine-serine-x arginine-serine. The individual from whom this allele of *PRB4* was cloned does produce the Po protein, which further supports this assignment. We therefore conclude that the Po protein is encoded by *PRB4* and results from the proteolytic cleavage of a larger precursor.

DNA sequences have been determined for two other alleles of *PRB4* (LYONS, STEIN and SMITHIES 1988). One of these alleles encodes a peptide identical to the Po protein. The other encodes a peptide with a single difference: an alanine occurs at amino acid position 239 rather than the proline found in the P-D protein. This substitution is unlikely to result in a Po^{-} phenotype, and we are therefore not surprised that all three individuals from whom these *PRB4*

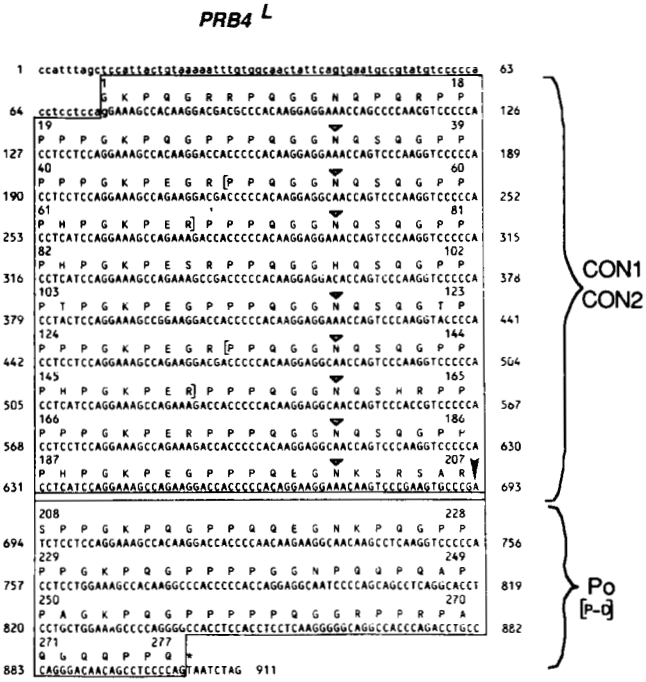

FIGURE 5.—Nucleotide and decoded amino acid sequences from the third exons of *PRB4*^{L} compared to the P-D protein sequence determined by SAITOH, ISEMURA and SANADA (1983b). The P-D peptide is electrophoretically identical to the Po protein (see text). Intronic sequences, coding sequences, potential proteolytic cleavage sites, and *N*-linked glycosylation sites are identified as described in the legends to Figures 1 and 4. Regions of identity with the CD-IIg peptide sequence determined by SHIMOMURA, KANAI and SANADA (1983) are enclosed in small square brackets.

alleles are cloned have a Po^{+} phenotype. We were unable to correlate Po protein phenotypes with the presence of length variants at *PRB4* because of the small number of individuals typed for the Po protein. However, no other PRP locus contains a region of complete or even near identity to the P-D peptide.

The Pc protein is probably a proteolytic cleavage product of a *PRB* locus: The assignments of the basic polymorphic PRPs to the loci indicated above demonstrate that all of these proteins, previously considered to be products of separate loci, represent proteolytic cleavage products of larger precursor proteins. The decoded amino acid sequences presented in Figures 2–5 demonstrate that there are likely to be additional potential proteolytic cleavage products encoded by these alleles. We anticipate that the unassigned basic protein, Pc, will be encoded by one of these fragments.

Reexamination of the inheritance of the heavily glycosylated PRPs: Genetic studies led to the description of three loci, *Gl*, *CON1* and *CON2*, which encode heavily glycosylated PRPs (AZEN, HURLEY and DENNISTON 1979; AZEN and YU 1984b).

The Gl protein is encoded by the *PRB3* locus: Polymorphisms for the glycosylated protein Gl were described in terms of four common productive alleles

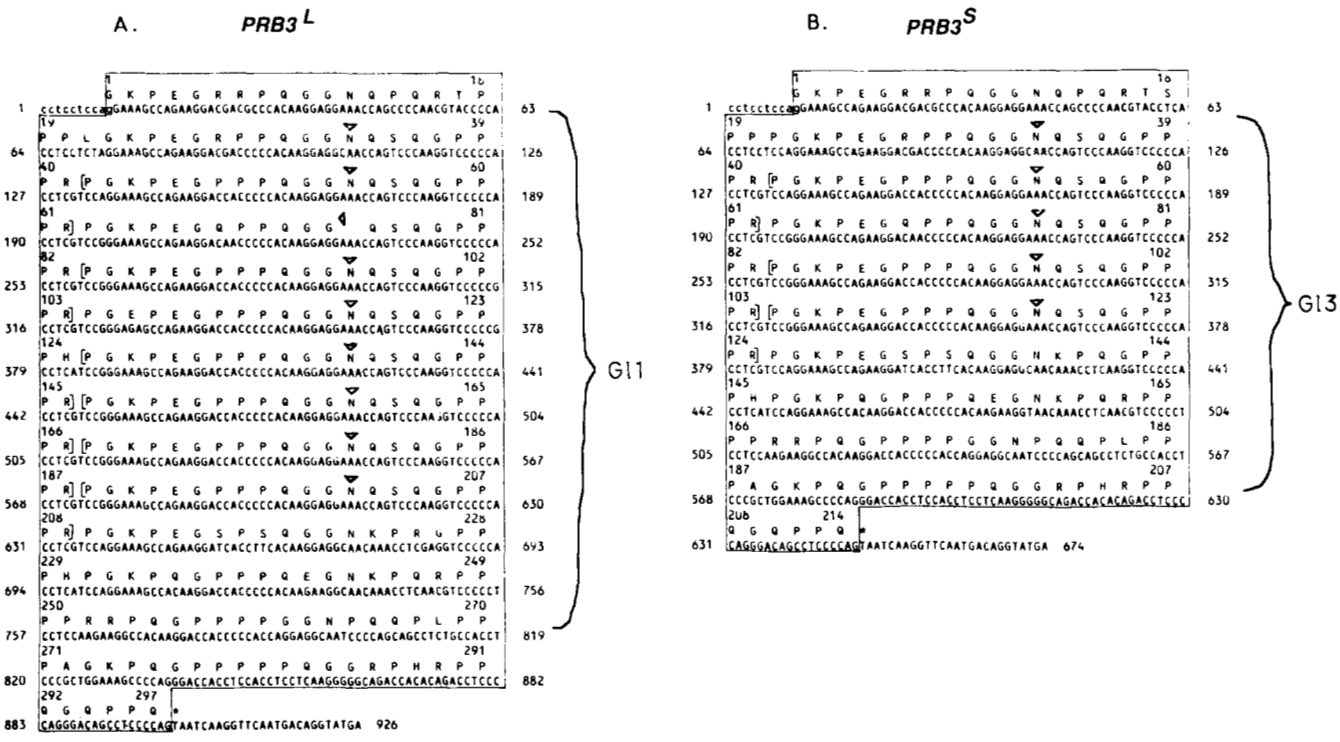

FIGURE 6.—Nucleotide and decoded amino acid sequences from the third exon of $PRB3^L$ (A) and $PRB3^S$ (B). Intronic sequences, coding sequences, and *N*-linked glycosylation sites are identified as described in the legends to Figures 1 and 3. Regions of identity with the CD-IIf peptide sequence determined by SHIMOMURA, KANAI and SANADA (1983) are enclosed in small square brackets.

and one null allele. The products of the four productive alleles differ in their molecular weights due to differences in the length of their protein backbones (AZEN, HURLEY and DENNISTON 1979). Although a complete protein sequence has not been determined for the Gl protein, a partial amino acid sequence of the peptide, CD-IIf, isolated from the major glycopeptide found in saliva and thought to be the Gl protein, has been determined by SHIMOMURA, KANAI and SANADA (1983). The decoded amino acid sequences for the two most common alleles at the *PRB3* locus, $PRB3^L$ and $PRB3^S$, are presented in Figure 6 and the regions identical to CDII-f are enclosed by square brackets. The CD-IIf sequence is encoded six times in $PRB3^L$ and three times in $PRB3^S$. No other PRP locus encodes this peptide sequence. The decoded amino acid sequences of $PRB3^L$ and $PRB3^S$ differ from those of the other PRP loci in that there are no sites within exon 3 which contain the general features described above for monobasic proteolytic cleavage sites. We therefore conclude that *PRB3* encodes the Gl protein and no other PRP.

Associations between *PRB3* length variants and Gl protein phenotype: Analysis of the segregation in families and in unrelated individuals of alleles at the *PRB3* locus provides strong support for the conclusion that Gl is encoded by *PRB3*. Figure 7 compares *PRB3* length variants and Gl protein phenotypes in eight unrelated individuals. The assignments of the $PRB3^L$, $PRB3^M$, and $PRB3^S$ alleles were confirmed by Southern blot analysis using other restriction enzymes (data not shown). The results of the comparison in all 24 unrelated individuals examined are presented in Table 2. A complete correlation between Gl protein phenotype and *PRB3* length variants was found. No other locus displays a correlation with Gl protein phenotype. Furthermore, the relative sizes of the *PRB3* DNA length variants ($PRB3^{VL} > L > M > S$) correspond to the relative sizes of the Gl proteins (Gl4 > 1 > 2 > 3) determined by AZEN, HURLEY and DENNISTON (1979). The segregation of G1 protein phenotypes and *PRB3* alleles was also examined in three families segregating *PRB3* length variants: a complete correlation was found. This analysis thus indicates that the decoded amino acid sequences presented in Figures 6A and 6B for $PRB3^L$ and $PRB3^S$ represent the Gl1 and Gl3 proteins, respectively.

AZEN, HURLEY and DENNISTON (1979) described a null allele, Gl^O, associated with the Gl protein polymorphism. We have examined the DNA from a Gl^OGl^O individual. No alterations in the *PRB3* or any other PRP locus were detected in a Southern blot of DNA from this individual (data not shown). Thus, the molecular basis for this null phenotype does not appear to involve a large DNA rearrangement such as a deletion. Its basis remains obscure.

The CON1 and CON2 proteins are products of the *PRB4* locus: The inheritance of the remaining

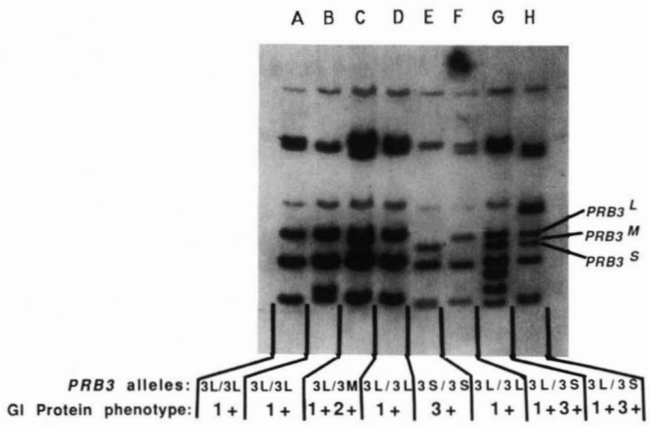

FIGURE 7.—Comparison of the segregation of *PRB3* DNA length variants with Gl protein phenotypes in eight unrelated individuals. Southern blot of *Eco*RI-digested DNA samples hybridized to the 980 bp *Hin*fI fragment from the third exon of *PRB1*. The positions of the bands derived from the *PRB3* alleles are marked. The types of *PRB3* alleles and Gl protein phenotypes for each individual are indicated below each lane.

TABLE 2

Comparison of Gl protein phenotype with *PRB3* length variants in 23 unrelated individuals

PRB3 allele(s)[a]	Gl protein phenotype	No. of individuals
3L/3L	1^+	12
3L/3M	1^+2^+	1
3L/3S	1^+3^+	8
3L/3VL	1^+4^+	1
3S/3S	3^+	1

[a] 3L = $PRB3^L$, 3M = $PRB3^M$, 3S = $PRB3^S$, 3VL = $PRB3^{VL}$.

glycosylated polymorphic PRPs, CON1 and CON2, have each been described as being determined by one productive and one nonproductive allele (AZEN and YU 1984b). The CON1 and CON2 proteins stain intensely with concanavalin A, suggesting that they are heavily glycosylated. Although no data on amino acid composition or sequence are available for these proteins, SHIMOMURA, KANAI and SANADA (1983) determined the amino acid sequence of a glycopeptide, CD-IIg, which was isolated from a heavily glycosylated PRP. The decoded amino acid sequence of *PRB4* (Figure 5) contains this peptide sequence (enclosed by square brackets). This sequence is found only in alleles at the *PRB4* locus, indicating that CD-IIg is derived from a heavily glycosylated PRP distinct from the other heavily glycosylated protein, Gl. As discussed above, the Gl protein appears to be the only product of the *PRB3* locus, suggesting that the CON1 and CON2 proteins must be encoded by a different locus. The only other PRP locus capable of encoding a heavily glycosylated PRP is *PRB4*. These observations suggest that *PRB4* encodes CON1 and CON2.

Our present proposal that the CON1 and CON2 proteins are in fact the products of a single locus, *PRB4*, requires reconciliation with the earlier assignments of the CON1 and CON2 proteins to two separate loci, each encoding one productive and one nonproductive allele (AZEN and YU 1984b). The data can be reconciled if there are actually two productive alleles, with gene frequencies in whites of 0.396 and 0.034, corresponding to the $CON1^+$ and $CON2^+$ protein phenotypes, respectively (AZEN and YU 1984b), and one "null" allele with a gene frequency in whites of 0.570. A reexamination of the inheritance of the CON1 and CON2 proteins in 24 of the families (including 132 individuals originally studied by AZEN and YU 1984b) show no inconsistencies with this proposal. The "null" allele, which still must be proposed, may represent a functional allele(s) whose protein products have not yet been identified electrophoretically as a consequence of loss or gain of proteolytic cleavage sites.

We have compared CON1 and CON2 protein phenotypes with length variants at the *PRB4* locus in families in order to further investigate this proposal. Only one family (seven members) in our sample was polymorphic for *PRB4* length variants, and in this family, a complete correlation was found between the presence of the $PRB4^M$ allele and the presence of the CON2 protein (data not shown). We were unable to investigate the correlation of $PRB4^M$ and the CON2 protein in our collection of unrelated individuals because none of them carried the $PRB4^M$ allele. Five of the 18 unrelated individuals examined did, however, have a $CON2^+$ phenotype, and all of them carried at least one copy of the $PRB4^S$ allele. This suggests that the CON2 protein may also be encoded by at least some alleles of the $PRB4^S$ type. CON1 protein phenotypes could not be compared with the presence of length variants because all individuals in our sample produced CON1. The available evidence thus indicates that the CON1 and CON2 proteins are encoded by *PRB4*.

Conclusions: The proposed assignments of the basic and glycosylated PRPs to the four PRP loci that we have considered in detail are summarized in Figure 1. The assignments proposed by MAEDA (1985) and confirmed by DNA studies (AZEN *et al.* 1987) for the acidic PRPs are also included in the figure. Together these results demonstrate that 12 of the 13 polymorphic PRPs so far described, which were assigned to 12 distinct loci in the earlier genetic studies, can now be assigned to the six PRP genes characterized in the DNA studies. The inheritance of the PRP polymorphisms described at the protein level in the earlier studies have been correlated with the presence of specific DNA length variants in most cases.

As originally suggested by MAEDA *et al.* (1985) following their characterization of cDNAs encoding PRPs, many of these proteins are actually proteolytic cleavage products of larger precursors rather than the

products of separate loci. An example of this from our present work is found in the case of the PmS, PmF and Pe proteins, which we suggest are derived by proteolytic cleavage from a single larger precursor encoded by *PRB1*. Our assignment of PmS, PmF, and Pe to a single transcript from *PRB1* predicts that strong associations should be evident among all three of these proteins. A strong association between the PmF and PmS proteins has already been noted by AZEN and DENNISTON (1980), but no strong associations were noted between the Pe protein and either PmF or PmS individually (AZEN and YU 1984a). However, a reexamination of the data derived from 108 individuals whose Pe, PmF and PmS protein phenotypes are known confirms our expectation—all individuals with a Pe^- phenotype are also phenotypically PmF^-PmS^- even though most chromosomes that encode a PmF^-PmS^- phenotype also encode a Pe^+ phenotype. This association of the Pe^- phenotype with the PmF^-PmS^- phenotype can be explained by the occurrence of a single amino acid substitution that inhibits proteolytic cleavage after residue 58 in an allele otherwise identical to $PRB1^M$ (Figure 2B).

We have concluded that the Po and CON proteins are proteolytic cleavage products of a precursor encoded by *PRB4*. However, no strong association was noted between the Po^+ and CON^+ protein phenotypes in previous genetic studies (AZEN and YU 1984a). A reexamination of the data derived from 108 individuals whose Po and CON protein phenotypes are known still revealed no clear associations between these proteins. We suspect that the low level of polymorphism (only six individuals have a Po^- phenotype) in the sample available to us precludes the identification of a strong association, and that an association would be evident in a larger sample.

We have concluded that the Ps proteins are proteolytic cleavage products of precursors encoded by *PRB2*. The Pc protein could not be assigned to a specific locus, but we expect that it will also prove to be a proteolytic cleavage product of a large precursor.

The existence of null alleles at 11 of the 13 loci was an unusual feature of the original genetic description of the PRP system. The sequence of the $PRB1^M$ allele (Figure 2), which encodes a PmS^-PmF^- phenotype, appears to encode a protein that is much larger than either PmS or PmF because of amino acid substitutions that remove proteolytic cleavage sites. Our data suggest that many of the null alleles proposed for other PRP loci are also due to losses of proteolytic cleavage sites accompanied by the production of fewer, but much larger PRPs, from a single transcript. These larger products would not be easily resolved in the electrophoretic systems used to type PRP polymorphisms, thus preventing their recognition as allelic to the much smaller PRPs. Similarly, the gain of a proteolytic cleavage site within a sequence corresponding to a known longer product, such as occurs in the PmS-like region in $PRB2^L$ (Figure 4), can also lead to a null phenotype. We suspect that the large differences in size, and in some cases in amino acid composition, among the peptides generated by proteolytic cleavage from a single transcript prevented the detection of associations between some of the cleavage products in earlier genetic studies. Incomplete proteolytic processing at some or all of the potential proteolytic cleavage sites discussed above would also be likely to hinder the recognition that several proteins of very different lengths can be derived from the same allele.

In summary, we have been able to construct an essentially complete outline of the genetic and phenotypic complexities of the PRP system. This system is remarkable in the way in which it produces a large number of proteins from only a few loci.

We thank LINDA M. SABATINI for providing the Southern blot in Figure 7. This work was supported by National Institutes of Health Grants GM20069 to O.S. and DEO 3658-23 to E.A.

LITERATURE CITED

ANDERSON, L. C., D. L. KAUFFMAN and P. J. KELLER, 1982 Identification of the Pm and PmS human parotid salivary proteins as basic proline-rich proteins. Biochem. Genet. **20:** 1131–1137.

AZEN, E. A., 1977 Salivary peroxidase (SAPX): genetic modification and relationship to the proline-rich (Pr) and acidic (Pa) proteins. Biochem. Genet. **15:** 9–29.

AZEN, E. A., 1988 Genetic protein polymorphisms of human saliva. In *Clinical Chemistry of Human Saliva*, Edited by J. TENOVUO, CRC Press, Inc., Cleveland, Ohio (in press).

AZEN, E. A., and C. L. DENNISTON, 1974 Genetic polymorphism of human salivary proline-rich proteins: further genetic analysis. Biochem. Genet. **12:** 109–120.

AZEN, E. A., and C. L. DENNISTON, 1980 Polymorphism of Ps (parotid size variant) and detection of a protein (PmS) related to the Pm (parotid middle band) system with genetic linkage of *Ps* and *Pm* to *Gl*, *Db* and *Pr* genetic determinants. Biochem. Genet. **18:** 483–501.

AZEN, E. A., and C. L. DENNISTON, 1981 Genetic polymorphism of PIF (parotid isoelectric focusing variant) proteins with linkage to the PPP (parotid proline-rich protein) gene complex. Biochem. Genet. **19:** 475–785.

AZEN, E. A., and N. MAEDA, 1988 Molecular genetics of human salivary proteins and their polymorphisms. In: *Advances in Human Genetics*, Plenum Press, New York, New York (in press).

AZEN, E. A., and F. G. OPPENHEIM, 1973 Genetic polymorphism of proline-rich human salivary proteins. Science **180:** 1067–1069.

AZEN, E. A., and P. L. YU, 1984a Genetic polymorphism of Pe and Po salivary proteins with probable linkage of their genes to the salivary protein gene complex (SPC). Biochem. Genet. **22:** 1065–1080.

AZEN, E. A., and P. L. YU, 1984b Genetic polymorphism of CON1 and CON2 salivary proteins detected by immunologic and concanavalin A reactions on nitrocellulose with linkage of Con 1 and Con 2 genes to the SPC (salivary protein complex). Biochem. Genet. **22:** 1–19.

AZEN, E. A., C. K. HURLEY and C. L. DENNISTON, 1979 Genetic

polymorphism of the major parotid salivary glycoprotein (Gl) with linkage to the genes for Pr, Db and Pa. Biochem. Genet. **17:** 257–279.

AZEN, E. A., K. M. LYONS, T. MCGONIGAL, N. L. BARRETT, L. S. CLEMENTS, N. MAEDA, E. F. VANIN, D. M. CARLSON and O. SMITHIES, 1984 Clones from the human gene complex coding for salivary proline-rich proteins. Proc. Natl. Acad. Sci. USA. **81:** 5561–5565.

AZEN, E. A., H.-S. KIM, P. GOODMAN, S. FLYNN and N. MAEDA, 1987 Alleles at the *PRH1* locus coding for the human salivary proline-rich proteins (PRPs) Pa, Db, and PIF. Am. J. Hum. Genet. **41:** 1035–1047.

BENNICK, A., 1987 Structural and genetic aspects of proline-rich proteins. J. Dent. Res. **66:** 457–461.

BOND, J. S., and P. E. BUTLER, 1987 Intracellular Proteases. Annu. Rev. Biochem. **56:** 333–364.

FRIEDMAN, R. D., A. D. MERRITT and M. L. RIVAS, 1975 Genetic studies of human acidic salivary protein (Pa). Am. J. Hum. Genet. **27:** 292–303.

GOODMAN, P. A., P. L. YU, E. A. AZEN and R. C. KARN, 1985 The human salivary complex (SPC): a large block of related genes. Am. J. Hum. Genet. **37:** 785–797.

IKEMOTO, S., K. MINAGUCHI, K. SUZUKI and K. TOMITA, 1977 New genetic marker in human parotid saliva (Pm). Science **197:** 378–379.

KARN, R. C., P. A. GOODMAN and P. L. YU, 1985 Description of a genetic polymorphism of a human proline-rich salivary protein, Pc, and its relationship to other proteins in the salivary protein complex (SPC). Biochem. Genet. **23:** 37–51.

KAUFFMAN, D. L., and P. J. KELLER, 1979 The basic proline-rich protein in human parotid saliva from a single subject. Arch. Oral Biol. **24:** 249–256.

KAUFFMAN, D., R. WONG, A. BENNICK and P. KELLER, 1982 Basic proline-rich proteins from human parotid saliva: complete covalent structure of protein IB-9 and partial structure of protein IB-6, members of a polymorphic pair. Biochemistry **21:** 6558–6562.

KAUFFMAN, D., T. HOFMANN, A. BENNICK and P. KELLER, 1986 Basic proline-rich proteins from parotid saliva: complete covalent structures of proteins IB-1 and IB-6. Biochemistry **25:** 2387–2392.

KIM, H.-S., AND N. MAEDA, 1986 Structure of two *Hae*III-type genes in the human salivary proline-rich protein multigene family. J. Biol. Chem. **261:** 6712–6718.

LYONS, K. M., J. H. STEIN and O. SMITHIES, 1988 Length polymorphisms in human proline-rich protein genes generated by intragenic unequal crossing over. Genetics **120:** 267–278.

MAEDA, N., 1985 Inheritance of human salivary proline-rich proteins: a reinterpretation in terms of six loci forming two subfamilies. Biochem. Genet. **23:** 455–464.

MAEDA, N., H.-S. KIM, E. A. AZEN and O. SMITHIES, 1985 Differential RNA processing and post-transitional cleavages in the proline-rich protein gene system. J. Biol. Chem. **260:** 11123–11130.

O'CONNELL, P., G. M. LATHROP, M. LAW, M. LEPPERT, Y. NAKAMURA, M. HOFF, E. KUMLIN, W. THOMAS, T. ELSNER, L. BALLARD, P. GOODMAN, E. AZEN, J. E. SADLER, G. Y. LAI, J.-M. LALOUEL and R. WHITE, 1987 A primary genetic linkage map for human chromosome 12. Genomics **1:** 93–102.

PONCZ, M., D. SOLOWIEJCZYK, B. HARPEL, Y. MORI, E. SCHWARTZ and S. SURREY, 1982 Construction of human gene libraries from small amounts of peripheral blood: analysis of β-like globin genes. Hemoglobin **6:** 27–36.

SAITOH, E., S. ISEMURA and K. SANADA, 1983a Further fractionation of basic proline-rich peptides from human parotid saliva and complete amino acid sequences of basic proline-rich peptide P-H. J. Biochem. **94:** 1991–1995.

SAITOH, E., S. ISEMURA and K. SANADA, 1983b Complete amino acid sequence of a basic proline-rich peptide, P-D, from human parotid saliva. J. Biochem. **93:** 495–502.

SCHWARTZ, T. W., 1986 The processing of peptide precursors. FEBS Lett. **200:** 1–10.

SHIMOMURA, H., Y. KANAI and K. SANADA, 1983 Amino acid sequences of glycopeptides obtained from basic proline-rich glycoprotein of human parotid saliva. J. Biochem. **93:** 857–863.

SOUTHERN, E. M., 1975 Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:** 503–517.

VANIN, E. F., P. S. HENTHORN, D. KIOUSSIS, F. GROSSVELD and O. SMITHIES, 1983 Unexpected relationships between four large deletions in the human β-globin gene cluster. Cell **35:** 701–709.

WAHL, G. M., M. STERN and G. R. STARK, 1979 Efficient transfer of large DNA fragments from agarose gels to diazobenzyloxymethyl-paper and rapid hybridization by using dextran sulfate. Proc. Natl. Acad. Sci. USA **76:** 3683–3687.

Communicating editor: R. E. GANSCHOW