The nucleotide sequences of cDNAs coding for human salivary proline-rich proteins (PRPs) were determined. Clones cP1 and cP2 contain repetitive regions in which sites for the restriction enzyme HaeIII occur repeatedly; they code for the precursors of acidic PRPs. Clones cP3 to cP7 contain repetitive regions in which BstNI sites occur repeatedly; they code for precursors of basic and glycosylated PRPs. The clones cP3, cP4, and cP5 are identical except that cP4 and cP5 are missing 399 and 459 base pairs, respectively, from the repetitive region of cP3. The sequences at these deletion end points are homologous to the consensus sequences of RNA splicing donor and acceptor sites. This strongly suggests that all three cDNAs are derived from the transcript of a single gene via differential RNA splicing. All of the precursor proteins share a feature—the N-terminal region, following the signal peptide, is acidic, while the remainder of the molecule, made of proline-rich repeats of about 21 amino acids, is basic. Each precursor can generate multiple PRPs by various post-translational cleavages on the carboxylic side of specific arginine residues. The data show how differential RNA splicing and post-translational cleavages could generate a large number of proteins, such as those found in saliva, from a much smaller number of genes.

A wide variety of proline-rich proteins (PRPs)1 form the major protein components of human saliva. They are characterized by a predominance of the amino acids proline (25 to 42%), glycine (16 to 22%), and glutamic acid/glutamine (15 to 28%). PRPs have been classified into three groups, acidic, basic, and glycosylated, based on their electrophoretic and chemical properties (Bennick, 1982). The amino acid sequences of two acidic proline-rich proteins (Wong and Bennick, 1980) and of some small basic proline-rich proteins (Kauffman et al., 1982; Saitoh et al., 1983a, 1983b) have been determined. Each contains repeated sequences that are homologous to the repeated sequences in the others. Although all of the proline-rich proteins in saliva have not yet been characterized, Kauffman and Keller (1979) have identified at least 11 basic proline-rich proteins in the saliva of a single individual.

Protein polymorphism studies led to the description of 13 discrete but genetically linked loci coding for the synthesis of the human salivary proline-rich proteins. The loci described for genes thought to code for the acidic PRPs are: Pr, proline-rich (Azen and Oppenheim, 1973); Pa, acidic protein (Friedman and Merritt, 1975); Db, double band (Azen and Denniston, 1974); and PIF, parotid isoelectric focusing variant (Azen and Denniston, 1981). Loci for genes thought to code for the basic PRPs are: PmF, parotid middle band fast (Ikemoto et al., 1977); PmS, parotid middle band slow (Azen and Denniston, 1960); Ps, parotid size variant (Azen and Denniston, 1980); G1, major glycoprotein (Azen et al., 1979); CON1, concanavalin A1 (Azen and Yu, 1984a); CON2, concavalin A2 (Azen and Yu, 1984a); Pc (Azen and Yu, 1984b); Po (Azen and Yu, 1984b); and Pp (Karn et al., 1989). Most of the postulated 13 loci are polymorphic in all races studied, and they are unusual in that null alleles, characterized by the absence of the corresponding PRPs in homozygotes, have been described for 11 of the 13.

DNA clones corresponding to two members of the human PRP gene family, clones PRPl and PRP2, have recently been isolated and their partial nucleotide sequences have been determined (Azen et al., 1984). Both clones contain regions comprising nearly identical tandem repeats of 63 nucleotides, the decoded amino acid sequences of which are homologous to the repeated amino acid sequences found in proline-rich proteins. A probe was made from the repeated region of the PRPl clone and used to examine EcoRI digests of DNA from several individuals. The resulting Southern blots were remarkable in showing six hybridizing bands in most individuals and in failing to show any correlation between the hybridizing bands in the tested individuals and the null alleles previously assigned to some of the individuals by studying their PRPs at the protein level.

Maeda (1985) recently reviewed and extended the Southern blot data and re-examined the patterns of inheritance of the PRPs. She hypothesized that the data could be interpreted more economically in terms of six loci falling into two subfamilies rather than in terms of 13 loci with null alleles. She called the six loci PRH1, PRH2, PRB1, PRB2, PRB3, and PRB4. Her hypothesis accounted in part for the lack of correlation between the genomic DNA hybridization patterns and the postulated presence of null alleles by suggesting that the three acidic PRPs (Db, Pa, and PIF) are coded by three productive alleles at one of the six loci, rather than by a productive allele and null allele at each of three loci.

Some of the complexities seen in the PRPs are also due to post-translational processing. This has been demonstrated for the acidic proteins, protein C and protein A (Wong et al., 1983), and is suspected in the case of the basic proteins (Bennick, 1982; Maeda, 1985).

To fully understand the structures of the PRPs, their mode
of inheritance, and the evolution of the PRP multigene family, it is necessary to identify all of the genes and their products. With this purpose in mind we constructed a cDNA library from the total poly(A+) RNA of a single human parotid gland. From this library we have isolated cDNAs that code for PRPs and have sequenced seven different species. In this paper we present these sequences, which permit us to classify the PRP genes into two groups. One group codes for acidic proteins and contains a repetitive region where sites for the restriction enzyme HaeIII occur repeatedly. The other group codes for basic and glycocylated PRPs and contains a repetitive region where BstNI sites occur repeatedly. The data suggest that, at least in the transcript from one PRP gene, differential RNA splicing has occurred to give mRNAs of different lengths. The data also demonstrate that post-translational proteolytic cleavage generates multiple products from all of the PRP genes.

EXPERIMENTAL PROCEDURES

Material—A normal parotid gland, obtained as a by-product from a patient who had undergone head and neck surgery for cancer, was frozen in liquid nitrogen and stored at -80°C. Saliva of the same individual was collected for protein typing (Azen et al., 1979; Azen and Denniston, 1980, 1981) and a sample of total DNA was isolated from his white blood cells (Poncz et al., 1982).

Purification of mRNA—A urea buffer solution (10 ml of 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, plus 0.2 g of SDS, 4.2 g of urea, and 0.2 g of NaCl) and phenol (10 ml) were frozen and ground into powder under liquid nitrogen. The frozen parotid gland tissue (weight 2 g) was added to the frozen buffer powder and both were ground together under liquid nitrogen. The nitrogen was then allowed to boil off, and the frozen powder was transferred to a 50-ml plastic centrifuge tube and thawed at room temperature with continuous shaking. Chloroform (5 ml) was added to the resulting viscous solution, which was centrifuged, and the lower phase was discarded. The supernatant was extracted twice more with phenol/CHCl₃ (2:1, v/v) followed by four washes with 2 volumes of CHCl₃/isoamyl alcohol (24:1, v/v).

Solid CsCl (1.1 g/2.5 ml) was added to the aqueous solution, and the mixture was layered onto 5.8 M CsCl solution in six tubes and centrifuged at 30,000 rpm in a Beckman SW 41 rotor at 20°C for 15 h. The supernatant was removed carefully; the RNA pellet was dissolved in the urea buffer solution (above) and extracted once with phenol, and the RNA was precipitated with ethanol. Poly(A+) RNA was obtained from the total RNA after two cycles of oligo(dT)-cellulose chromatography (Aviv and Leder, 1972).

Cell-free Translation—Poly(A+) RNA was translated in a cell-free rabbit reticulocyte lysate system purchased from Bethesda Research Laboratories in the presence of [³H]proline or [³⁵S]methionine. A major band was seen at approximately M₉, 40,000, relative to globular protein size markers (data not shown). Other poly(A+)-rich bands were observed at 53,000, 45,000, 34,000, 32,000, and 30,000. Under the same conditions of SDS-polyacrylamide gel electrophoresis, Ps2 synthesized in vivo migrates to the rate expected for a globular protein of size 50,000, Ps1 at 44,000, and PsM at 24,000. Since many of the in vitro proteins are modified post-translationally by cleavage, by glycosylation, or by phosphorylation, it is not surprising that the products translated in the cell-free system could not be correlated with the in vivo products. Nonetheless, the ability to translate the poly(A+) RNA into PRPs indicates that the preparation contains translatable mRNA for PRPs. The data also suggest that the number of PRP precursor proteins may not be so large as the number of proteins found in saliva.

cDNA Clones Coding for PRPs Fall into Two Groups—Clones of cDNA were prepared from the parotid gland poly(A+) RNA as described under "Experimental Procedures" and were screened by colony hybridization using the Hinf-980 probe. Ninety-five of 750 colonies gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes. One group, about 72% of the initially positive clones, no longer gave a positive signal after a nonstringent wash. These could be classified into two main groups on the basis of their signals after more stringent washes.

RESULTS AND DISCUSSION

Protein Typing and Characterization of Poly(A+) RNA—The poly(A+)-rich proteins in the saliva of the individual whose parotid gland was used in this study had the following phenotypes: Prl-1, Db-, Ps-, PIF-, Ps1, G1-3, PmF-, PsM. The poly(A+) RNA prepared from the donor parotid gland was electrophoresed in a 1.5% agarose gel, transferred to nitrocellulose, and hybridized to a probe (Hinf-980) made from the region that contains repeats in the human PRP gene clone PRP1 (Azen et al., 1984). Two main bands of 1.2 and 0.78 kb were detected in addition to less strongly hybridizing bands of 1.4 and 0.9 kb (data not shown). This result indicates that PRP-related species of at least 1.4 kb should be represented in cDNA clones made from this RNA preparation.

Cell-free translation of the same poly(A+) RNA gave several bands showing a high proline content as judged by the ratio of proline/methionine incorporation in the presence of either [³H]proline or [³⁵S]methionine. A major band was seen at approximately M₉, 40,000, relative to globular protein size markers (data not shown). Other poly(A+)-rich bands were observed at 53,000, 45,000, 34,000, 32,000, and 30,000. Under the same conditions of SDS-polyacrylamide gel electrophoresis, Ps2 synthesized in vivo migrates to the rate expected for a globular protein of size 50,000, Ps1 at 44,000, and PsM at 24,000. Since many of the in vitro proteins are modified post-translationally by cleavage, by glycosylation, or by phosphorylation, it is not surprising that the products translated in the cell-free system could not be correlated with the in vivo products. Nonetheless, the ability to translate the poly(A+) RNA into PRPs indicates that the preparation contains translatable mRNA for PRPs. The data also suggest that the number of PRP precursor proteins may not be so large as the number of proteins found in saliva.

2 J. Brosius and O. Smithies, unpublished work.
to that of clone PRP1, which has BstNI sites (CC7/GG) spaced approximately every 63 bp. We have called these repeats BstNI repeats (Azen et al., 1984). The clones in the other group have a structure in which HaeIII sites (GGCC) occur repeatedly and we have named them HaeIII repeats. We find that both groups share a homologous 5' untranslated and secretory signal sequence (see below), and so we used a probe made from this region to isolate 19 full length cDNA clones for PRPs.

**Overall Structure of cDNAs—**The nucleotide sequences of the seven cDNAs (presented in Figs. 1–3 below) indicate that the PRP proteins are clearly related and derived from a common ancestor. They are composed of four domains (illustrated below in Fig. 4): a secretory signal sequence (S), N-terminal regions (N1 and N2), repeats BstNI (B) or HaeIII (H), and the C-terminal region. The secretory signal sequence (S) is composed of 16 predominantly hydrophobic amino acids and is similar to other secretory signal peptides (Watson, 1984). The 17-amino acid N-terminal region (N1) which follows is rich in acidic and hydrophobic residues. A second N-terminal region (N2) is unique to the HaeIII family proteins and has no counterpart in the BstNI family proteins. The repeated regions are made of about 21 amino acid residues and are rich in proline, but the number of repeats varies in the different cDNAs. A small C-terminal region (C1 or C2) follows the repeats.

**cDNA Clones with HaeIII Repeats Code for Acidic PRPs—**The nucleotide sequences of 12 full length cDNA clones with HaeIII repeats were determined. Four correspond to the sequence cP1 and five correspond to the sequence cP2 shown in Fig. 1. Two of the 12 clones had rearrangements almost certainly created during the cDNA construction. One of these two is identical to cP2 except for the addition at the 5' end of 93 bp of extra DNA the sequence of which is an exact copy of the opposite of cP2 strand from positions 381 to 473. The other clone also has an inverted sequence; in this case positions 89 to 195 of cP1 are missing and are replaced with 46 bp from the 5' end of cP1 in an inverted form. (Similar artifactual inversions have been observed in other cDNAs (Derynck et al., 1980; Fagen et al., 1980; Weaver et al., 1981; Pletnev et al., 1983).) A third cDNA was found identical to cP2 except that the 12-bp sequence GGCCATCCCCGT from positions 393 to 404 is replaced by the sequence ACCCCACCCTCCA which we have not found anywhere else. The reading frame of this cDNA is intact and would give a decoded amino acid sequence of His-Pro-Pro-Pro in place of the usual Gly-His-Arg. We suspect but cannot prove that this clone is also an artifact since six of the seven clones covering this region lacked the unusual sequence.

The nucleotide sequences presented in Fig. 1 show that, although the decoded amino acid sequences of the clones cP1 and cP2 are very similar, they code for the precursors of two different acidic PRPs. They are identical except for amino acid positions 4 and 50; cP1 has aspartic acid and asparagine at positions 4 and 50, respectively, while cP2 has asparagine and aspartic acid at these positions. The decoded amino acid sequences of cP1 and cP2 and the amino acid sequence of an acidic PRP, protein C, determined by Wong and Bennick (1980) also differ at positions 4 and 50; both positions in protein C are reported as asparagine. Since the donor of the parotid gland is producing only two acidic PRPs (P1 and PIF) cP1 and cP2 must correspond to Pn and PIF. These two proteins co-narrow the region that is composed of 7% polyacrylamide gel electrophoresis at pH 8.9, but they can be separated from each other by isoelectric focusing over the pH range 3.5 to 5.2 (Azen and Denniston, 1981) despite the fact that both have the same total amino acid composition. Our results suggest that protein C as sequenced was probably a mixture of the two proteins. (Asparagine is frequently the residue assigned when protein sequencing experiments show aspartic acid and asparagine at the same position.)

One feature shared by all of the acidic PRPs, except Pa, is their appearance as double bands after electrophoresis. Wong and Bennick (1980) have shown that the amino acid sequence of their protein A is identical to the first N-terminal 106 residues of protein C, which itself is composed of 150 residues. The salivary protease kallikrein has been shown to cleave protein C on the carboxyl side of Arg-106 (Wong et al., 1983) to give protein A plus a carboxy-terminal peptide of 44 amino acids. The identification in human saliva by Isemura et al. (1980) of a peptide P-C which is identical in sequence to the C-terminal 44 residues of protein C suggests that such proteo-
lytic cleavages occur in vivo in the salivary PRPs. Since the decoded amino acid sequences of both the cP1 and cP2 clones have the Arg-106, this mechanism is most likely responsible for the double-bandedness of PIP (PIP slow and PIP fast) and Pr1 (Pr1 and Pr3). Among all the PRP cDNA sequences that we have determined, including sequences of an additional three incomplete cDNA clones with HaeIII repeats, we have found none which is missing codons for the C-terminal 44 residues. This result strongly supports the hypothesis that the double-banded feature of acidic PRPs is due to post-translational cleavage at Arg-106 and is not due to the existence of different mRNAs coding for the two proteins making up any particular double-banded phenotype.

Possible Differential RNA Splicing in cDNA Clones with Different Numbers of BstNI Repeats—Seven full-sized cDNAs and one partial cDNA with BstNI repeats were sequenced. They form three subgroups based on their sequences. Fig. 2 shows the nucleotide sequences of the subgroup of BstNI family clones: cP3, cP4, and cP5. We obtained 2 examples of cP3, one of cP4, and two of cP5. Alignment of the nucleotide sequences shows that cP4 and cP5 are identical to cP3 (two examples) except that they are missing 399 and 459 bp, respectively, from the central part of cP3. We can exclude the possibility that they are derived from alleles at one locus because there are three different lengths from one individual. We can also exclude the possibility that recombination within the repeats during cloning might have caused the deletions, because inspection of the sequences in the appropriate alignment shows that the deletions could not have been generated by homologous crossovers between different repeating units. We therefore considered the possibility that they are differentially spliced transcripts from a single gene.

Inspection of the nucleotide sequence of cP3 shows that the 5' and 3' ends of the cP4 and cP5 deletions are homologous to the consensus sequences for RNA splicing donor sites, (C/A)AG-ATAGTA and acceptor sites, (C/T)AG-G (Mount, 1982). Consequently these nucleotide sequences support the hypothesis that the three different length mRNAs, cP3, cP4, and cP5, are all derived from a single transcript via differential RNA splicing. In cP3, there are two possible donor sites at nucleotide positions 347 and 529 and two acceptor sites at 379, 439, 499, 562, 622, 682, 745, and 805. cP4 is apparently the result of a 347 to 745 splicing and cP5 is the result of a 347 to 805 splicing. We do not know whether the other possible sites are ever used or not.

Amino Acid Sequence of the BstNI Family Clone cP3—The nucleotide sequences of the 5’ flanking, signal (S), and N terminal (Nj) portions of the BstNI family sequence cP3 are homologous to those of the HaeIII family sequences cP1 and cP2. We have already seen that protein C probably corresponds to a mixture of the translation products of cP1 and cP2. We can therefore use the sequence of protein C and the nucleotide sequence data to predict the structure of the proteins corresponding to cP3, cP4, and cP5. The N-terminal amino acids of the three secreted proteins will most likely be glutamine. They will all have an N-terminal region of 17 amino acids, followed by a differing number of the BstNI repeats, B1, B2, and B3 (B1 and B2 repeats are 21 amino acids, B3 repeats are 20 amino acids).

**FIG. 2. Nucleotide sequences of cDNA clones cP3, cP4, and cP5.** Sequences of cP4 and cP5 identical to that of cP3 are shown by *solid lines* and gaps are shown by *. Dots. Possible donor and acceptor splicing sites are shown by *—* and +, respectively. Nucleotide sequences of repeats (B1, B2, and B3) are boxed and aligned to maximize homology. Signal sequence (S), N-terminal region (Nj), and C-terminal region (Cj) are also boxed. Downward arrowheads indicate the positions of post-translational cleavages and upward arrowheads indicate the common BstNI sites in the repeats.
the first three B3 repeats are 19 and the fourth B3 repeat is 20 amino acids in length. Their consensus sequences are given in Fig. 4b.) The unit "B1-B2-B3" occurs four times in cP3. The nucleotide sequence of the C-terminal region (Cn) of cP3 is not homologous to that of the repeating units, although its decoded amino acid sequence is still rich in proline.

The overall decoded amino acid sequence of cP3 corresponds to a PRP of 315 amino acids. None of the presently sequenced salivary PRPs is as long as this. This suggests that some of the currently sequenced proteins could be post-translational cleavage products of a cP3-like precursor protein. An arginine residue occurs at the equivalent position in all but one of the B3 repeats of the cP3 precursor protein. If complete proteolytic cleavage occurred after all these arginine residues (at positions 75, 136, and 259) cP3 would give rise to four different basic proline-rich peptides. It would generate peptides of 75 residues, (from N1 and the first B1-B2-B3 unit), of 61 residues (from the second B2-B3-B1 unit), of 123 residues (two B1-B2-B3 units left together because of the replacement of arginine by glutamine at position 197), and of 56 residues (the last B2, repeat and C-terminal region). cP4 and cP5 would generate the same N- and C-terminal peptides as cP3. In addition, cP4 would produce a basic peptide of 50 residues and cP5 would produce a basic peptide of 30 residues.

Preparation of cDNA library by others on the amino acid sequences of salivary PRPs provides evidence for this interpretation of the nucleotide sequences. The amino acid sequences of the four B1-B2-B3 units of 61 amino acids we expect from cP3 are very similar although not identical to the amino acid sequences of the small basic PRPs. Residues 87 to 197 are identical to IB-9 and P-E (Kaufman et al., 1982; Isemura et al., 1982), except that the glutamine at 197 of cP3 is replaced by arginine in IB-9 and P-E. Residues 198 to 259 of cP3 are identical to P-F (Saitoh et al., 1983a) with three exceptions. Ala-249 and Gln-256 of cP3 are replaced by proline and arginine, respectively, in P-F, while Arg-259 is missing in P-F. The amino acid sequence of the basic protein P-H determined by Saitoh et al. (1983b) is identical to that from positions 260 to 315 of cP3, except that Saitoh et al. found alanine at the N-terminus of P-H rather than the serine at the equivalent position (residue 260) of cP3. Since the PRP system is genetically very polymorphic, some of these minor differences between the protein sequences and those predicted from cP3 could well be due to inherited variations between different individuals.

Arginine residues at positions 72 and 194 of cP3 could also be cleavage sites for precursors. However, the amino acid sequences of the basic proline-rich proteins IB-9 and P-F retain this Arg-Ser bond intact which suggests that this bond is resistant to cleavage. Nonetheless, in conjunction with the differential RNA splicing we have suggested, post-translational cleavage of precursor proteins could produce at least 6 different basic proline-rich proteins from the single gene corresponding to cP3.

cDNAs for Human Salivary Proline-rich Proteins

The sequence of clone cP7 (Fig. 3b) shows that clone cP7 also contains BatNI repeats but is not full length even though it is larger than 1 kb. The repeat unit of cP7 is similar to the B1-B2-B3 units of cP3. Some repeats (B0 and B3) contain a sequence for a sugar attachment site, "N-K-S." If the precursor protein of cP7 is processed post-translationally at Arg-197, Arg-198 and Gln-256 of cP3 are replaced by proline and arginine, respectively, in P-F, while Arg-259 is missing in P-F. The amino acid sequence of the basic protein P-H determined by Saitoh et al. (1983b) is identical to that from positions 260 to 315 of cP3, except that Saitoh et al. found alanine at the N-terminus of P-H rather than the serine at the equivalent position (residue 260) of cP3. Since the PRP system is genetically very polymorphic, some of these minor differences between the protein sequences and those predicted from cP3 could well be due to inherited variations between different individuals.

Arginine residues at positions 72 and 194 of cP3 could also be cleavage sites for precursors. However, the amino acid sequences of the basic proline-rich proteins IB-9 and P-F retain this Arg-Ser bond intact which suggests that this bond is resistant to cleavage. Nonetheless, in conjunction with the differential RNA splicing we have suggested, post-translational cleavage of precursor proteins could produce at least 6 different basic proline-rich proteins from the single gene corresponding to cP3.

cDNAs with BstNI Repeats Code for Two Types of Glycosylated PRPs—Fig. 3 presents the nucleotide sequences of two other types of cloned PRP cDNAs. cP6 (Fig. 3a) has a signal sequence (S) and N-terminal region (N1) very similar to those of cP3. However, the repeated region of cP6 is characterized differently from that in cP3 in that the most part it contains a fourth variety of BstNI repeat (B4) which occurs seven times. Two and one-half more repeats (like B4, B3, and part of B3) follow the seven copies of B4. The C-terminal region of cP6 is an abbreviated form (Cn) of the equivalent region (Cn) in cP3.

The B4 repeat is different because the decoded sequence includes the sugar attachment site "N-Q-S." This suggests that cP6 codes for a glycosylated PRP having six carbohydrate side chains. (The first of the seven B4 repeats has a single base pair change in the site so it would not have a side chain.) Furthermore, the amino acid sequence of the B4 repeat is very similar to the amino acid sequences of the glycopeptides, CD-II and CD-IIg, obtained from proteolytic digests of a glycosylated PRP (Shimomura et al., 1983).

If cP6 were translated and glycosylated without proteolytic cleavage, it would code for a secreted polypeptide of 231 amino acids. The sequence of the last six amino acids in the seventh B4 repeat is similar to that of the B3 repeat and contains two arginine residues. If proteolytic cleavage occurs in the protein corresponding to cP6 in the same way as we suggest it occurs in that corresponding to cP3, the secreted glycoprotein would be 161 (or possibly 158) amino acids in length (Fig. 3a). One glycosylated PRP that has been studied (Levine et al., 1979) lacks hydrophobic amino acids. This suggests that cP6 might also be cleaved after Arg-23 to give a basic glycopeptide of 138 amino acids lacking hydrophobic residues plus an N-terminal unglycosylated acidic peptide of 23 residues not rich in proline but containing hydrophobic amino acids. The C-terminal basic peptide would be 70 amino acids in length and nonglycosylated. This C-terminal sequence corresponds to the amino acid sequence of the basic proline-rich peptide P-D (Saitoh et al., 1983c) except that cP6 has alanine at position 195 whereas P-D has proline.

The sequence of clone cP7 (Fig. 3b) shows that clone cP7 also contains BatNI repeats but is not full length even though it is larger than 1 kb. The repeat unit of cP7 is similar to the B1-B2-B3 units of cP3. Some repeats (B0 and B3) contain a sequence for a sugar attachment site, "N-K-S." If the precursor protein of cP7 is processed post-translationally at Arg-9, Arg-197, and -133 in the manner we are proposing, it would produce peptides of 62 amino acids without sugar side chains and very similar to those corresponding to cP3 are also expected from the sequence of cP7. The peptide expected from residues 134 to 195 is identical to P-F except that P-F does not have an arginine at its C terminus. The C-terminal peptide (residues 196 to 252) is completely identical to P-H (Saitoh et al., 1983b, 1983c).

Assignment of cDNA Clones to Five PRP Loci—In the human genome there are probably six loci controlling the synthesis of the proline-rich proteins in saliva (Maeda, 1985). Two loci have HaeIII-type repeats and are named PRH1 and PRH2; four have BstNI-type repeats and are named PBB1, PBR2, PBR3, and PBR4. Each locus has several alleles, some of which show polymorphic length differences of the repeating region (Azen et al., 1984). We have isolated full length cDNAs corresponding to four loci and an incomplete cDNA clone corresponding to a fifth locus. Assigning these cDNA clones to the appropriate genetic locus requires the isolation and characterization of genomic clones. Genomic clones containing the genes at the PRH1 and PRH2 loci have been isolated from the DNA of the same individual whose parotid gland was used to make cDNA. By sequencing the exons of these genomic clones and by mapping genomic DNAs from other individuals with different acidic PRP phenotypes, we have been able to assign cP1 tentatively to the PRH2 locus (Pr1)
Fig. 3. Nucleotide sequences of cDNA clones cP6 (a) and cP7 (b). In addition to the symbols used as in Figs. 1 and 2, the sugar attachment sites are marked with three black squares. As the clone cP7 is not a complete cDNA, the glycine residue after the poly(G) tail is arbitrarily numbered position 1 in the amino acid sequence.

and cP2 to the PRH1 locus (PIF). The BstNI family clones, cP3, cP4, and cP5, are most likely transcripts of the PRB1 locus. Thus, a genomic clone, PRP1, which corresponds to each of the six loci is likely in the absence of selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci. We obtained 28% HueIII- and 72% selective effects to be directly proportional to the transcriptional activity of the loci.

CONCLUSION

Fig. 4a summarizes schematically the results and conclusions we draw from studying the seven types of cDNA clone described in this paper. The tentative identities of the genetic loci and the types of protein product we expect after in vivo proteolysis are also given. All full length cDNA clones have a homologous signal peptide (S) and N-terminal region (Nt). The consensus sequences of S and Nt are given in Fig. 4b. Clones cP1 and cP2, which code for acidic proline-rich proteins, differ from the others in having an additional N2 region. The two types of repeats, H and B, which give rise to the repetitive HaeIII site and BstNI sites in the DNA sequences, are homologous to each other as is clear from their consensus amino acid sequences shown in Fig. 4b. Both repeats code for basic polypeptides. The B-type repeats can be subdivided into B1, B2, B3, and B4 repeats as is also shown in Fig. 4b. Interestingly, the variations which lead to the subdivision of the B repeats are clustered within the second half of repeats while the first half is much more conserved.

H.-S. Kim and N. Maeda, unpublished results.
Two features of the PRP cDNAs that we have studied show how a relatively small number of genes (six loci) could produce a much larger number of proteins. The first feature, differential RNA splicing, was suggested by the relationships between cP3, cP4, and cP5. Fig. 4a illustrates how cP4 and cP5 could be obtained from the same transcript as cP3 by splicing out different lengths of RNA. We do not know whether this is a general feature in PRP genes and the potential splicing sites in other genes are used or not. (Clone cP6 has a potential donor site at position 475 and two potential acceptor sites 3'...
to the donor site at positions 510 and 573.) Differential RNA splicing of several genes is known to produce secreted and membrane-associated forms of immunoglobulins (Young et al., 1981), and a different number of exons in Drosophila myosin (Rozek and Davidson, 1983) and human fibronectin (Kornblitt et al., 1984). The case of a single donor site combining with several acceptor sites has been demonstrated in the rat fibronectin gene (Tamkun et al., 1984). The present examples of differential RNA splicing in a PRP gene are unique in that both alternative donor and acceptor sites occur within a single exon (see also Azen et al., 1984) without altering the reading frame or the general nature of the translated product.

The second feature leading to multiple proteins is the probable ubiquity of post-translational proteolytic cleavage in these proteins. Fig. 4a indicates the various protein products to be expected if post-translational cleavage, such as occurs at Arg-106 in the acidic PRPs, is a general process. Such processing would generate at least four basic PRPs from the cP3 precursor, all of different lengths and amino acid sequences but all having a homologous repeated region. cP6 has potential cleavage sites at Arg-23 and Arg-161 which would generate a heavily glycosylated PRP and an additional basic PRP. The cleaved products of cP7 would be basic PRPs with or without a single carbohydrate side chain. As indicated above, some of these postulated products have sequences identical or similar to known basic PRPs. In most cases we do not know to what extent the proteolytic processing occurs except that the HaellII acidic protein is known to be partially cleaved. Thus, Wong et al. (1983) found the ratio of protein C to its cleaved product protein A to be approximately 1:1 in saliva—a ratio not affected by salivary flow rate. If cleavage at the arginine residues in the precursor proteins with BstNI-type (basic and glycosylated) PRPs. If cleavages are incomplete and permuted the number will be larger.

The question remains as to function of the many different PRPs, the major protein components of saliva. Acidic PRPs bind ionic calcium (Gron and Hay, 1976), bind strongly to hydroxyapatite, and inhibit the precipitation of calcium phosphate in saliva (Hay, 1973). The functions of the glycosylated and basic PRPs are not known. PRPs also occur in the respiratory tract (Warner and Azen, 1984), which suggests that they could have some general function different from those specific to the oral cavity. It is possible that the precursor proteins rather than the processed proteins are more important functionally. All of the precursor proteins share the feature that their N-terminal region is acidic and the rest of the molecule, made up of proline-rich repeats, is basic. Because of their unusual amino acid compositions and primary structures, the molecules are also expected to have some interesting higher order structures. Conceivably the precursor molecules serve one function (perhaps in the respiratory tract), and their many prolineyct products serve a different function (perhaps in saliva). Whatever may be the case, the PRP system is unusual in the number and variety of different protein products that can be obtained from a small number of genes.

Acknowledgment—We thank L. Gottlieb, T. Warner, and G. Hafez for providing a salivary gland tissue. We are grateful to Dale Wiesman, Susan McEvoy, and Michael Koralewski for the plasmid growth and to Karen Lyons for the permission to quote her unpublished work and helpful discussion.

REFERENCES

Aviv, H., and Leder, P. (1972) Proc. Natl. Acad. Sci. U. S. A. 69, 1408-1412
Azen, E. A., and Oestreicher, P. G. (1973) Science (Wash. D. C.) 180, 1007-1009
Azen, E. A., and Denniston, C. L. (1974) Biochem. Genet. 12, 109-120
Azen, E. A., Hurley, C. K., and Denniston, C. (1979) Biochem. Genet. 17, 257-279
Azen, E. A., and Denniston, C. (1980) Biochem. Genet. 18, 483-501
Azen, E. A., and Denniston, C. (1981) Biochem. Genet. 19, 475-486
Azen, E. A., and Yu, P.-L. (1984a) Biochem. Genet. 22, 1-19
Azen, E. A., and Yu, P.-L. (1984b) Biochem. Genet. 22, 1965-1968
Azen, E., Lyons, K. M., McElmigt, T., Barrett, N. L., Clements, L. S., Maeda, N., Vannin, E. F., Carlson, D. M., and Smithies, O. (1984) Proc. Natl. Acad. Sci. U. S. A. 81, 5551-5555
Bennick, A. (1982) Mol. Cell. Biochem. 45, 53-99
Bonner, W. M., and Laskey, R. A. (1974) Eur. J. Biochem. 46, 85-98
Derynck, R., Content, J., DeClercq, E., Volckaert, G., Tavernier, J., Devos, R., and Fiers, W. (1986) Nature 285, 542-547
Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. (1980) Cell 20, 313-319
Fagen, J. B., Pastan, I., and de Crombrugghe, B. (1980) Nucleic Acids Res. 8, 3059-3064
Friedman, R. D., and Merritt, A. D. (1975) Am. J. Hum. Genet. 27, 304-314
Grin, P., and Hay, D. J. (1970) Arch. Oral Biol. 21, 201-205
Grunstein, M., and Hognes, D. S. (1975) Proc. Natl. Acad. Sci. U. S. A. 72, 2961-2965
Hay, D. I. (1973) Arch. Oral Biol. 18, 1517-1529
Ishimoto, S., Managuchi, K., Suzuki, K., and Tomita, K. (1977) Science 197, 378-379
Isemura, S., Saitoh, E., and Sanada, K. (1980) J. Biol. Chem. (Tokyo) 87, 1071-1077
Isemura, S., Saitoh, E., and Sanada, K. (1982) J. Biochem. (Tokyo) 91, 2067-2075
Karn, R. C., Goodman, P. A., and Yu, P.-L. (1985) Biochem. Genet. 23, 37-51
Kauffman, D. L., and Keller, P. J. (1979) Arch. Oral Biol. 24, 249-256
Kauffman, D. L., Wong, R. S. C., Bennick, A., and Keller, P. J. (1982) Biochemistry 21, 6556-6562
Kornblitt, A. R., Vibe-Pedersen, K., and Baralle, F. E. (1984) Nucleic Acids Res. 12, 5883-5888
Levine, M. J., Ellison, S. A., and Balb, O. P. (1973) Arch. Oral Biol. 18, 827-837
Mazia, N. (1985) Biochem. Genet. 23, 445-464
Maxam, A. M., and Gilbert, W. (1977) Proc. Natl. Acad. Sci. U. S. A. 74, 560-564
Mozg, S. (1982) Nucleic Acids Res. 10, 459-472
Okayama, H., and Berg, P. (1982) Mol. Cell. Biol. 2, 101-170
Platen, M. G., Scoble, N. A., Frolova, L. U., and Kisselev, L. L. (1983) Gene 25, 93-100
Poncz, M., Solowijczyk, K., Harpel, B., Mory, Y., Schwarz, E., and Surrey, S. (1982) Hemoglobin 6, 27-36
Rozek, E. E., and Davidson, N. (1983) Cell 32, 23-34
Saitoh, E., Isemura, S., and Sanada, K. (1983a) J. Biochem. (Tokyo) 93, 883-888
Saitoh, E., Isemura, S., and Sanada, K. (1983b) J. Biochem. (Tokyo) 94, 1991-1996
Saitoh, E., Isemura, S., and Sanada, K. (1983c) J. Biochem. (Tokyo) 95, 495-502
Shimomura, H., Kaniy, Y., and Sanada, K. (1983) J. Biochem. (Tokyo) 93, 857-863
Tamkun, J. W., Schwabauer, J. E., and Hynes, R. O. (1984) Proc. Natl. Acad. Sci. U. S. A. 81, 5140-5144
Warner, T. E., and Azen, E. A. (1984) Am. Rev. Respir. Dis. 130, 116-118
Weetman, M. E. E. (1984) Nucleic Acids Res. 12, 5145-5154
Weaver, C. A., Gordon, D. F., and Kemper, B. (1981) Proc. Natl. Acad. Sci. U. S. A. 78, 4073-4077
Wong, R. S. C., and Bennick, A. (1980) J. Biol. Chem. 255, 5945-5948
Wong, R. S. C., Madapallanmattam, G., and Bennick, A. (1983) Biochem. 211, 40-44
Young, R. A., Hagenbuchle, O., and Schiller, U. (1981) Cell 23, 451-458