Secondary structure prediction of human salivary proline-rich proteins

Hilda Cid, Veronica Vargas, Marta Bunster and Sergio Bustos*

Department of Molecular Biology, Faculty of Biological Sciences and Natural Resources, University of Concepcion, Concepcion, Chile and **Department of Oral Biology-Biochemistry, Medical College of Georgia, Augusta, GA 30912, USA*

Received 18 December 1985; revised version received 20 January 1986

Conformations associated with secondary structure in human salivary proline-rich proteins A (PRPA), C (PRPC), P-D and P-E were predicted by analysis of their respective hydrophobicity profiles by computer programming. Structurally, PRPA and PRPC would present a globular head and a tail that consists of type 3_{10} polyproline helices. P-D and P-E would be fibrilar molecules with helical zones of the polyproline 3_{10} type. Alternatively for PRPA and PRPC, the head and tail would form one globular domain with the tail folding upon itself at places where random coils occur.

Proline-rich protein *Conformation* *Computer prediction (Human saliva)* *Sequence homology* *Structural model*

1. INTRODUCTION

The salivary secretion of human parotid and submandibular salivary gland is a complex mixture of macromolecules that include a group of acidic and basic proteins collectively known as proline-rich proteins [1–3]. Two of the predominant acidic proline-rich proteins have been named A and C [1] and their amino acid sequences determined.

More recently, the complete amino acid sequences of two of the basic proline-rich proteins, P-D [4] and P-E [5] (IB-9) from human parotid saliva have been reported. P-E is identical to IB-9 [6] with the exception of amino acid residue 22.

All 4 of these proteins have been well characterized as to their molecular masses and primary structures. The calcium-binding properties as well as the inhibitory action of protein A and C on calcium phosphate crystal growth have also been reported.

There have been few and conflicting reports [7–11] on the conformational aspects of proteins A, C, P-D and P-E. This fact prompted us to examine the available data and to attempt prediction of their secondary structures.

2. METHODS AND RESULTS

We used two prediction methods which have been reported to have a reliability of about 80% when applied to globular proteins: (i) the method of Corrigan and Huang [12] which uses the Chou and Fasman [13] rules of prediction to obtain a computer graph of the secondary structure of the protein based only on information derived from the amino acid sequence; and (ii) the method of Cid et al. [14] that predicts the secondary structure based on the hydrophobicity profile of the protein.

Both methods imply that the amino acid sequences contain all the necessary information for the folding of the polypeptide chain. Since sequence studies [7] have shown that protein A is 100% identical in primary structure with the first 106 amino acid residues of protein C, the prediction of the secondary structure was made only for protein C, P-D and P-E.

In the method of Corrigan and Huang, the evaluation program analyzes the α and β confor-

* To whom correspondence should be addressed

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/86/$3.50 © 1986 Federation of European Biochemical Societies

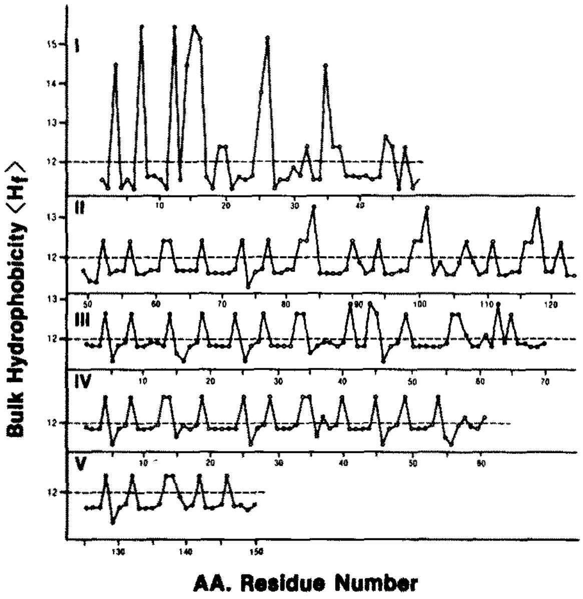


Fig.1. Hydrophobicity profiles of protein C, P-D and P-E aligned to display their similarities. The sequence of protein C has been divided into 3 parts: (I) residues 1–48; (II) residues 49–124 and (V) residues 125–150. P-D and P-E correspond to III and IV, respectively.

mational parameters, P_α and P_β defined by Chou and Fasman [13]. These conformational parameters represent the normalized frequency of occurrence of each amino acid in that particular type of secondary structure, as obtained from a data base of 29 globular proteins whose tertiary structures are fully known from X-ray diffraction methods. The resulting text file containing the data from these calculations is stored and the transfer program activated. This program examines and checks the values of the conformational parameters P_α, P_β, and P_t and proceeds to plot them. If the conditions for α-helix, β-pleated sheet and β-turn are not met, a random coil is plotted.

The second method used was that of Cid et al. [14] which considers that the folding of the polypeptide chain occurs in a way that allows the hydrophilic amino acids to locate themselves in the protein surface, whereas the hydrophobic ones can be buried in the interior of the molecule. A 'bulk hydrophobic character', as defined in [15] for each of the 20 natural amino acid residues, is used to draw the hydrophobicity profile of the protein. A computer program systematically reads the sequence of amino acids, assigning to each one the bulk hydrophobic character $\langle H_f \rangle$ (calculated as an average in a data set of 21 globular proteins of known tertiary structure), location with respect to the protein surface [16], and polarity and charge [17]. Four typical profiles are defined: an exposed helical structure, an exposed and a buried β-strand and a β-turn. The prediction of the secondary structure by this method consists simply of the

Table 1

Prediction of the secondary structure of proteins C and A

Segment	Structure (method of Cid et al.)	Segment	Structure (method of Chou and Fasman)	Segment	Structure (joint prediction)
1–7	random coil	1–5	random coil	1–7	random coil
8–11	β-turn	6–11	α-helix	8–11	β-turn
12	random coil			12	random coil
13–17	β-strand	12–17	β-strand	13–17	β-strand
18–22	random coil	18–20	random coil	18–22	random coil
23–28	α-helix	21–27	α-helix	23–28	α-helix
29–33	random coil	28–30	random coil	29–33	random coil
		31–34	β-turn		
34–38	β-strand	35–39	β-strand	34–38	β-strand
39–42	random coil	40–45	random coil	39–42	random coil
43–48	β-strand	46–49	β-turn	43–48	β-strand
49–106	random coil	50–106	random coil	49–106	random coil
107–150[a]	random coil	107–150[a]	random coil	107–150[a]	random coil

[a] Only for protein C

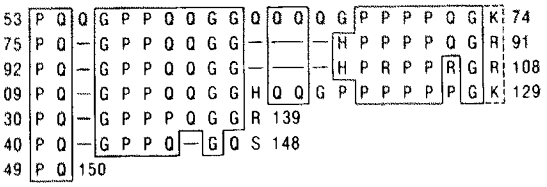
identification of these basic patterns in the hydrophobicity profile of the protein [14].

The hydrophobicity profiles of protein C, P-D and P-E, aligned to show their similarities, are shown in fig.1. It is clear that the profile of the first 50 amino acid residues of protein C is completely different from the rest. The secondary structure prediction by both methods indicates a random-coiled structure for P-D and P-E and for the 100 amino acid 'tail' of protein C. Table 1 compares the predictions made independently by both methods.

3. DISCUSSION

As shown in table 1, both methods yield a large percentage of random-coiled structure. This is not surprising, since both predictive methods are bound to data bases of globular proteins whose tertiary structures are fully determined. None of the proteins included in the data base has an amino acid composition similar to those of the proline-rich proteins.

Analyzing the sequences of proteins A, C, P-D and P-E, a very striking repetition of groups of amino acids in sequence is observed. In protein C, for example, the sequence Pro-Gln-Gly is found 9 times and the sequence Pro-Gln-Gln 5 times. In general, the collagen-like sequences Gly-X-Pro appear several times in the 4 proteins analyzed. Fig.2 shows the repetition of groups of amino acids in the tail of protein C. Fig.3 shows the conservation of the primary structure between protein C, P-D and P-E. If the first 48 and the last 37 residues of protein C are not considered, there is 43% homology between protein C, P-D and P-E and 72% between P-D and P-E. Notice that long strands of amino acid residues such as 17–31 (P-D) with 38–52 (P-E), or 20–31 (P-D) with 21–32 (P-E) are identical. It is also important to note that the sequence 126–138 of protein C is repeated as 2–14 in P-E. This points towards the generation of protein C, A, P-D and P-E by a single precursor, of the 'head-tail' type, and that P-D and P-E are parts of a tail.

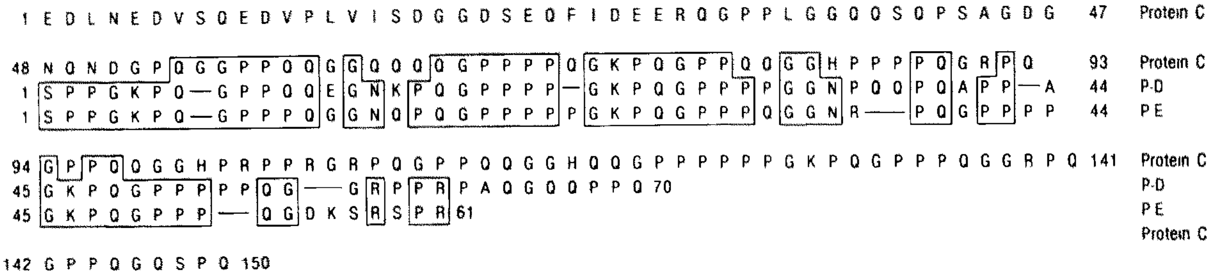


```
 53 P Q Q G P P Q Q G G Q Q Q Q G P P P P Q G K 74
 75 P Q — G P P Q Q G G — ——— — H P P P P Q G R 91
 92 P Q — G P P Q Q G G — ——— — H P R P P R G R 108
109 P Q — G P P Q Q G G H Q Q G P P P P P P G K 129
130 P Q — G P P P Q G G R 139
140 P Q — G P P Q — G Q S 148
149 P Q 150
```

Fig.2. The repetition of groups of amino acids in the tail of protein C. Note that these sequences include tripeptides that form 3_{10} helices (cf. fig.4).

X-ray diffraction studies of synthetic polypeptides of sequences similar to those found repeated in the saliva proteins have been published by Traub et al. [18–22] and others [23,24]. They have shown that the sequences (Gly-Pro-Pro)$_n$, (Pro-X-Gly)$_n$, (Pro-Pro-X)$_n$, (Pro-Pro-Pro)$_n$ and (Gly-X-Pro)$_n$ present a helical structure of type 3_{10}. Such sequences are found in protein C, P-D and P-E, and therefore we postulate that the 'random-coiled' zones predicted by both methods could in fact be helical zones of the polyproline type (fig.4).

In summary, the following structural models can account for the data discussed above: P-D and P-E are fibrillar molecules with several helical zones of type 3_{10}, and would need a stabilization similar to that of collagen, i.e., 3 molecules would form a superhelix. Proteins C and A would present a globular head (residues 1–48) and a tail with several polyproline-like helical regions that stretches from residue 49 to 150 or 106, respective-



```
  1 E D L N E D V S Q E D V P L V I S D G G D S E Q F I D E E R Q G P P L G G Q Q S Q P S A G D G   47  Protein C

 48 N Q N D G P Q G G P P Q Q G G Q Q Q Q G P P P P Q G K P Q G P P Q Q G G H P P P P Q G R P Q   93  Protein C
  1 S P P G K P Q — G P P Q Q E G N K P Q G P P P P — G K P Q G P P P P G G N P Q Q P Q A P P — A  44  P-D
  1 S P P G K P Q — G P P P Q G G N Q P Q G P P P P P G K P Q G P P P Q G G N R ——— P Q G P P P P  44  P E

 94 G P P Q Q G G H P R P P R G R P Q G P P Q Q G G H Q Q G P P P P P P G K P Q G P P P Q G G R P Q 141  Protein C
 45 G K P Q G P P P P P Q G ——— G R P P R P A Q G Q Q P P Q 70                                      P-D
 45 G K P Q G P P P ——— Q G D K S R S P R 61                                                        P E
                                                                                                     Protein C
142 G P P Q G Q S P Q 150
```

Fig.3. Comparison of the primary structure of protein C, P-D and P-E. Homologous residues are enclosed in boxes.

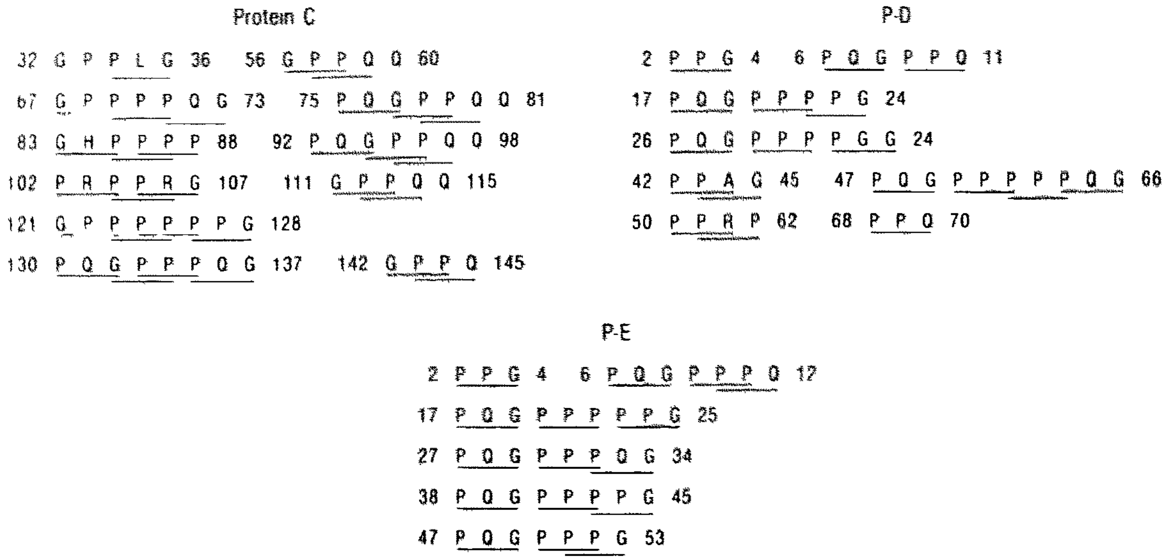


Fig.4. 3_{10} helices in protein C, P-D and P-E. The tripeptides responsible for the formation of the helices are underlined.

ly. Like P-D and P-E the tails would need 3 neighboring molecules to form a collagen-type triple helix, with the heads protruding outwards. Another possibility is that both, head and tail, would form one globular domain, with the tail folding upon itself at places where there are random-coiled zones: these are sequences 49–55, 61–66 and 88–92 for protein A, and 115–120 and 137–142 for protein C. However, for this type of structure, the very special amino acid composition of these proteins becomes unnecessary (fig.5).

The circular dichroism (CD) data obtained by Bennick et al. [7–9] do not support the occurrence of the poly(L-proline) form II conformation in protein A, C and IB-9.

The conformation of the two basic proline-rich polypeptides P-D and P-E from parotid saliva has been studied by Shibata et al. [10] using CD and NMR. Their CD data suggest the existence of the poly(L-proline) form II conformation in P-D. The CD spectrum obtained by Bhatnagar et al. [11] for protein A shows that its conformation is related to poly(L-proline) II. Their data support the structure we postulate for P-D and P-E, and for the tails of proteins C and A. The structure for the heads of the latter two proteins would be that proposed in the joint prediction in table 1, with the only addition of a possible 3_{10} helix in the region 32–36, as shown in fig.5.

We are aware that the validity of the models proposed must be tested experimentally, however its importance resides in the fact that they can direct further studies in these proteins.

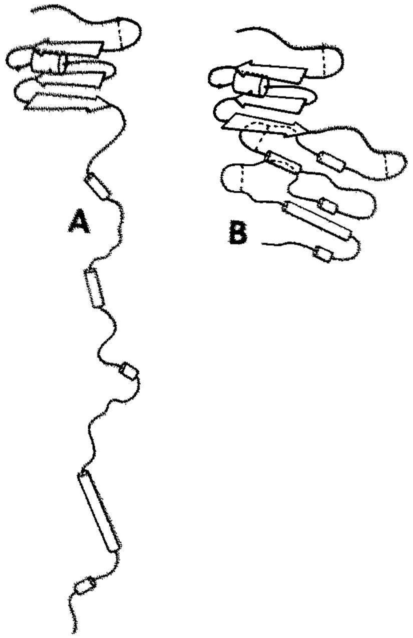


Fig.5. Postulated models for the structure of protein C and A. Fibrillar model (A), globular model (B).

ACKNOWLEDGEMENTS

This work was partially supported by Research Project 20.33.16 from the Research Office, University of Concepcion, Concepcion, Chile and by research funds from the Oral Biology-Biochemistry Department, Medical College of Georgia, Augusta, GA. The authors acknowledge the help of Ms Beth Robinson from the Health Systems and Computer Science Department in the computer programming for plotting the secondary structure of the proteins reported in this study. The authors are also indebted to Mr Gonzalo Castillo who wrote the computer programs to detect the repeated strands in all the proteins studied.

REFERENCES

[1] Bennick, A. and Connell, G.E. (1971) Biochem. J. 123, 455–464.

[2] Oppenheim, F.G., Hay, D.I. and Franzblau, C. (1971) Biochemistry 10, 4233–4238.

[3] Kauffman, D.L. and Keller, P.J. (1979) Arch. Oral Biol. 24, 249–256.

[4] Saitoh, E., Isemura, S. and Sanada, K. (1983) J. Biochem. 93, 459–502.

[5] Isemura, S., Saitoh, E. and Sanada, K. (1982) J. Biochem. 91, 2067–2075.

[6] Kauffman, D.L., Wong, R.S.C., Bennick, A. and Keller, P.J. (1982) J. Dent. Res. 61, 294.

[7] Wong, R.S.C. and Bennick, A. (1980) J. Biol. Chem. 255, 5943–5948.

[8] Kauffman, D., Wong, R., Bennick, A. and Keller, P. (1982) Biochemistry 21, 6558–6562.

[9] Bennick, A. (1977) Biochem. J. 163, 229–239.

[10] Shibata, S., Asakura, J., Isemura, T., Isemura, S., Saitoh, E. and Sanada, K. (1984) Int. J. Peptide Protein Res. 23, 158–165.

[11] Nordbo, A.H., Darwish, S., Sorensen, K.R. and Bhatnagar, R.S. (1984) J. Dent. Res. 63, 227.

[12] Corrigan, A.H. and Huang, P.C. (1982) Comput. Prog. Biomed. 15, 163–168.

[13] Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 211–244.

[14] Cid, H., Bunster, M., Arriagada, E. and Campos, M. (1982) FEBS Lett. 150, 247–254.

[15] Ponnuswamy, P.K., Prabhakaran, M. and Manavalan, P. (1980) Biochim. Biophys. Acta 623, 301–316.

[16] Chothia, C. (1976) J. Mol. Biol. 105, 1–14.

[17] Olsen, K.W. (1980) Biochim. Biophys. Acta 622, 259–267.

[18] Traub, W. and Yonath, A. (1966) J. Mol. Biol. 16, 404–414.

[19] Yonath, A. and Traub, W. (1969) J. Mol. Biol. 43, 461–477.

[20] Segal, D.M. and Traub, W. (1969) J. Mol. Biol. 43, 487–496.

[21] Segal, D.M., Traub, W. and Yonath, A. (1969) J. Mol. Biol. 43, 519–527.

[22] Traub, W., Shmueli, U., Suwalsky, M. and Yonath, A. (1967) in: Conformation of Biopolymers (Ramachandran, G.N. ed.) vol.2, pp.449–467, Academic Press, New York.

[23] Olsen, B.R., Berg, R.A., Sakakibara, S., Kishida, Y. and Prockop, D.J. (1971) J. Mol. Biol. 57, 589–595.

[24] Brown, F.R., Di Corato, A., Lorenzi, G.P. and Blout, E.R. (1972) J. Mol. Biol. 63, 85–99.