Biophysical Journal, Volume 99

Supporting Material

Proline-rich salivary proteins have extended conformations

Helene Boze, Therese Marlin, Dominique Durand, Javier Perez, Aude Vernhet, Francis Canon, Pascale Sarni-Manchado, Veronique Cheynier, and Bernard Cabane

Proline-rich salivary proteins have extended conformations

Hélène Boze[1], Thérèse Marlin[1], Dominique Durand[2], Javier Pérez[3], Aude Vernhet[1], Francis Canon[1], Pascale Sarni-Manchado[1], Véronique Cheynier[1], Bernard Cabane[4*]

1 INRA, Montpellier SupAgro, UMR 1083 Sciences pour l'Œnologie, 2 place Viala, F-34060 Montpellier, France

2 IBBMC, CNRS UMR 8619, Université Paris-Sud, 91405 Orsay, France

3 Beamline SWING, Synchrotron SOLEIL, BP 48, 91192 Gif-sur-Yvette, France

4 PMMH, CNRS UMR 7636, ESPCI, 10 rue Vauquelin, 75231 Paris cedex 05, France

* Author for correspondence

Supplementary information

Bioinformatic analysis

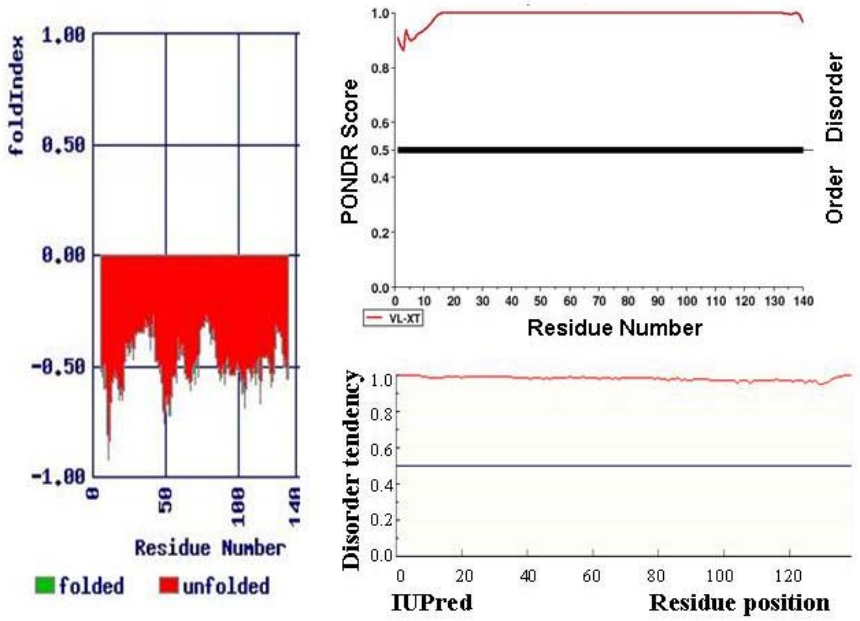

Figure SI-1. Predictions of disorder for the structures of the salivary protein IB5 in solution, based on the sequence given in figure 2. Left panel: Foldindex predicts no folding at all (http://bip.weizmann.ac.il/fldbin/findex). Upper right panel: PONDR (Predictor of Naturally Disordered Regions) predicts maximum disorder (www.pondr.com). Lower right panel: Iupred (Prediction of Intrinsically Unstructured Proteins) also predicts maximum disorder (http://iupred.enzim.hu/).

Reconstruction of data-compatible conformations

The single conformation approach: reconstruction of a single conformation that reproduces the experimental scattering.

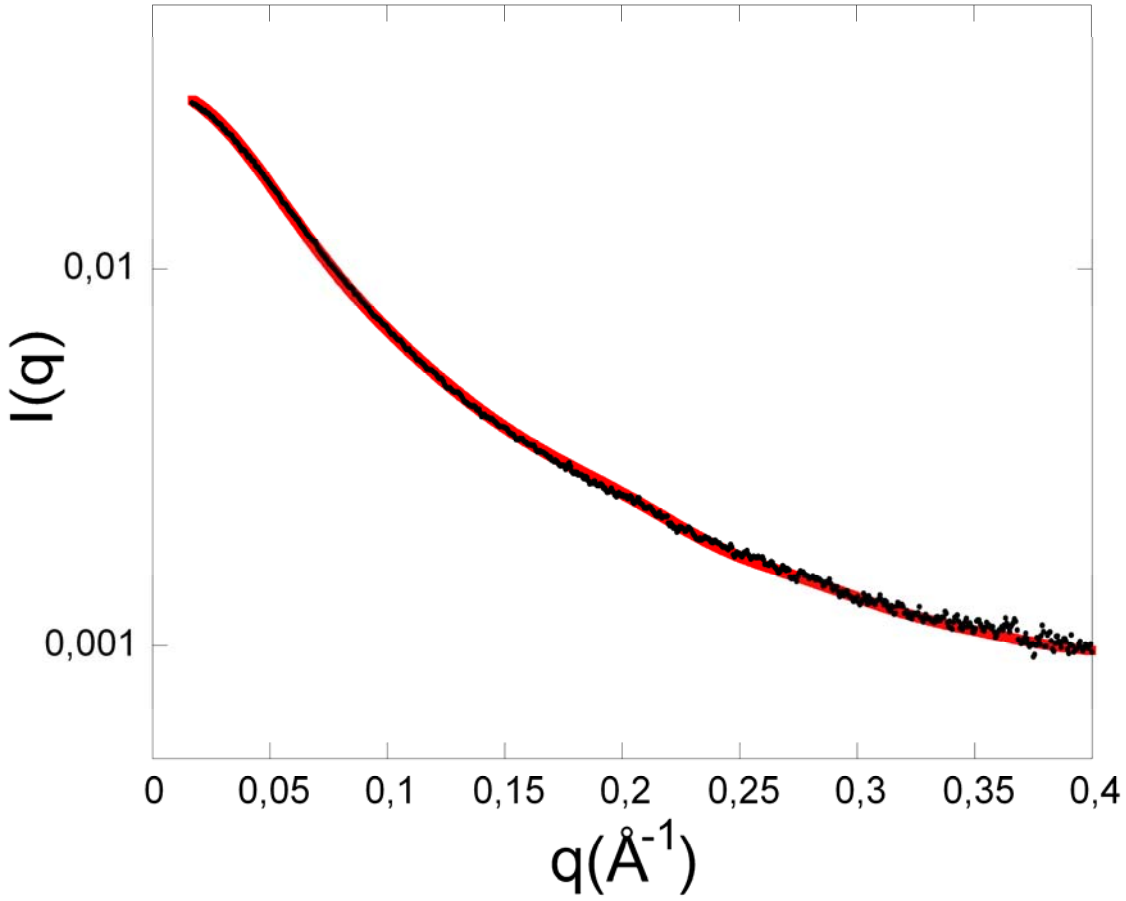

Figure SI-2. Calculated scattering curve from a single "reconstructed" conformation for protein IB5 in solution (shown in Figure 7 of the paper). SAXS spectrum of IB5 in solution (dots), compared with the prediction from BUNCH (red line)

The other approach: choosing a subset from a large pool of conformations

The program EOM (Ensemble Optimized Method) describes the protein by using an ensemble of typically 50 conformations extracted from a very large (typically 10,000) and assumed exhaustive pool of conformations [59]. A genetic algorithm progressively refines the composition of the ensemble so that the average scattering pattern of the molecular conformations within the ensemble fits the experimental data within error bars. The process is repeated typically 100 times and the distribution of structural parameters such as the radius of gyration and the maximum diameter are calculated and compared with those derived from the entire starting pool.

We applied EOM using the same rigid domains as those used for BUNCH (i.e. 3 polyproline elements of 3, 4 and 5 residues respectively) but also a fully random chain without polyproline structure elements. The results are exactly the same. We chose 20, 50, 100 or 200 conformations per ensemble. As shown on Figure S3a, the distribution of radii of gyration values of the selected, optimized ensembles depends on the number of curves per ensemble and more surprisingly on the noise of the data. So the noisy Nanostar curve, shown on Figure S3b, provides distributions clearly different of those obtained using the Swing curve. Nevertheless it appears that all distributions are centred around a root mean square value higher than the value corresponding to the pool. This confirms that IB5 adopts conformations more extended than the average of all accessible conformations.

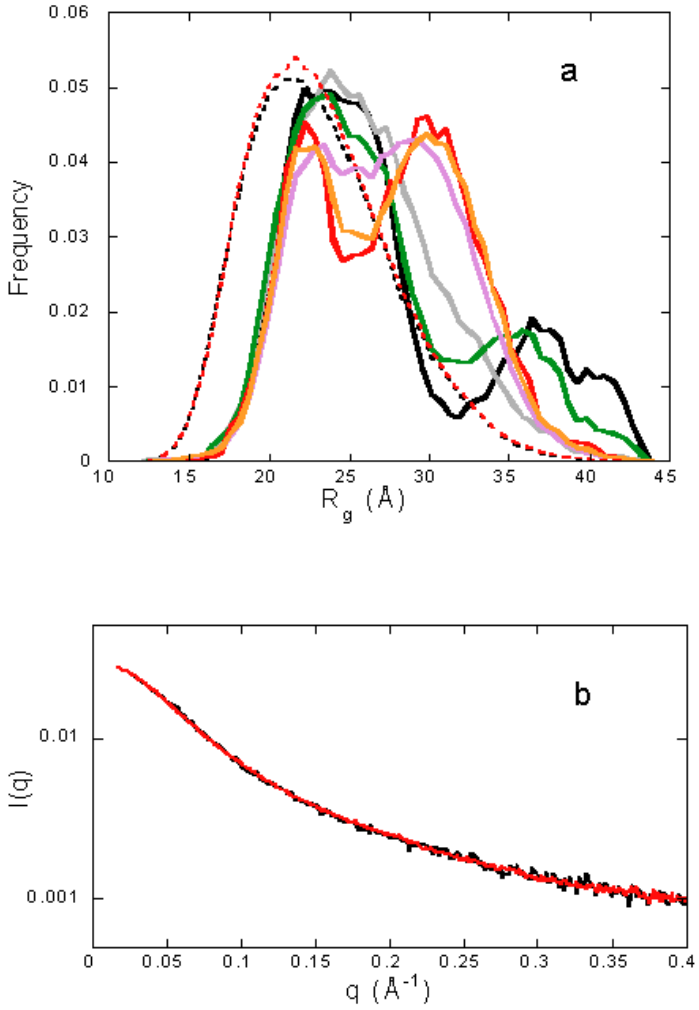

Figure SI-3

(a) Distribution of radius of gyration values of conformations from ensembles selected by the program EOM so that the average scattering curve fits the experimental data for the protein IB5. Red, orange and pink continuous lines: obtained from the Swing data using 20, 100 and 200 conformations per ensemble respectively. Black, green and grey continuous lines: obtained from the Nanostar curve using 20, 100 and 200 conformations per ensemble respectively. Dashed lines: corresponding distributions for the complete pool of 10000 randomly generated conformations (red: Swing data; black: Nanostar data)

(b) Experimental spectra used to obtain the previous distributions. Red line: Swing data; black line: Nanostar data. The fitted curves are not shown because they are exactly superimposed to the experimental ones.