Use of Voronoi tessellation in confidence-assessed prediction of protein structural class

C. J. RICHARDSON AND D. J. BARLOW

Department of Pharmacy, King's College London, Manresa Road, London SW3 6LX

The exponential growth of protein primary structure data furnished through the various genome sequencing programmes has emphatically marked the need for reliable means to predict protein 3D structure from amino acid sequence (Taubes 1996). Without such facilities the hard-earned genomic data has negligible practical utility, contributing nothing of immediate significance to facilitate the task of rational (structure-based) drug design.

Of considerable importance is knowledge of the principal type of secondary structural elements composing a novel protein, that is its structural class (all-helical, all-β-sheet, etc.). This basic knowledge assists significantly in predicting the location of secondary structural elements and may even help in deciding the probable function of the protein. Previous work has shown how a protein's structural class can be predicted simply from a knowledge of the molecule's sequence, using methods such as cluster analysis (Nishikawa *et al.* 1983). The level of success achieved in these predictions is typically around 75%. Encouraging though this seems, it must be noted that all structural class predictions made for novel proteins will therefore have an uncertainty of 25% (regardless of the nature of the protein's sequence) and in many ways it would be preferable to have a prediction method that gave a case-dependent assessment of prediction confidence.

Here we report on a novel program QhullProp that we have developed for this purpose. It takes protein amino acid sequence data as input and maps this into 2 dimensions with coordinates provided by the mean α- and β-forming propensities (Levitt 1978) per residue. The 2D maps of protein data points are then partitioned according to a Voronoi tessellation (Aurenhammer 1991) with each fully-bounded cell within the tessellation labelled with the structural class of the protein data point it contains (Fig. 1a).

In predicting the structural class of a novel protein an assignment is made by noting the structural class labels for the cells which surround the novel protein cell in the two-dimensional transformation. If a novel protein's cell is completely surrounded by cells with the same structural class label (Fig. 1b), the novel protein is assigned this structural class. Novel proteins whose cells are surrounded by others with a mixture of class labels (Fig 1c) are left with no structural class prediction. Although this reduces the number of predictions made from 920 to 82, the 9% of predictions made are 100% accurate and are made with 100% confidence. The results from this preliminary investigation of a novel prediction methodology are very encouraging. It is anticipated that improved methods for mapping from a protein sequence to a Voronoi tessellated plane will increase the proportion of novel sequences for which a trustworthy prediction can be made.

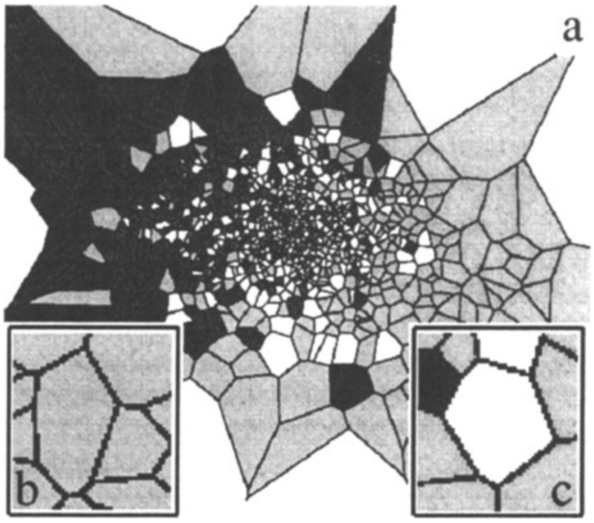

Figure 1. a) Voronoi tessellated 2D map of sequence data for 920 proteins of classes all-α (light grey), all-β (dark grey) and α/β (white) from the Structural Class of Proteins Database (Murzin *et al.* 1995). Details of Voronoi cells where a prediction b) is and c) is not made.

Aurenhammer, F. 1991. ACM Comp. Surv. **23**:345-465
Levitt, M. 1978. Biochem. **17**:4277-4285.
Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. J. Mol. Biol. **247**:536-540.
Nishikawa, K., Kubota, Y. and Ooi, T. J. 1983. Biochem **94**:997-1007.
Taubes, G. 1996. Science **273**:588-590.