Figures legends :

Figure 1 : The learning step.

Protein structures are described by their dihedral angles. Window ***V****(i)* of 8 consecutives angles are attributed to the closest Protein Block ***W****(k)* according to the *rmsda* criterion. For the first C learning cycles (C is an user-fixed parameter) only *rmsda* criterion is used in the PB selection. For the following cycles, the PB choice is based on the search of maximal transition frequency relation to the previous selected PBs (see text).

Figure 2 : Superimposition of fragments of the structural alphabet.

The backbone in superimposed with the atoms Cα, N, O and C along the 5 amino acids of fragments for the 16 PBs. PBa to PBp from left to right and from top to bottom .

Figure 3 : Coding of the ubiquitin conjugating enzyme (2aak).

(a) 3D structure coding of 2aak in terms of Protein Blocks. (b) Variation of the *rmsd* (Å) along the sequence.

Figure 4 : Description of Protein Blocks PBp, PBb, PBd and PBm.

(a - d) 3D representation of superimposed Cα coordinates fragments associated with the PBs. (e - h) Z-score matrices of amino acid distributions in the window [-7:+7]. Amino acids are ordered according to a decreasing hydrophobicity scales. The images are displayed in 5 gray levels according to the Z-scores (thresholds : -4.4, -2.0, 2.0, 4.4). These matrices characterize the amino acid over and under-representations in the protein sub-sequence encompassing the protein block. (i - l) KLd profiles defined in the window [-7:+7]. They characterize the positions where certain amino acids contribute to the specificity of these fold patterns (see text).

Figure 5 : KLd profile of PBb's sequence family.

The Kld profile of PBb is in straigth line. The KLd profiles of the families PBb_1 and PBb_2 are indicated by crosses and black dots respectively.

Figure 6 : Improvement of prediction rate after introducing sequence families.

The figure shows the plot between the initial prediction rate Q(1) at the first rank (in abcissae), and the difference between prediction rates (Q(1)* - Q(1)) (in ordinate); Q(1)* is the prediction rate at the first rank after introducing sequence families (see text). The proteins are notified all-α proteins (*), all-β (O) and non-all-α or all-β (.).

Figure 7 : Protein block prediction of the the ubiquitin conjugating enzyme (*2aak*).

(a) *Neq* variation along the protein sequence. This index *Neq* ("equivalent number of PBs") based on the Shannon entropy quantifies the prediction uncertainty. (b) Rank of the true PB in the prediction. (c) Number of selected blocks for an average prediction rate of Q_g of 75% with the "global stategy" (see text); the dots show the positions where the true PB is found again at the first rank, the smaller ones when it is found among the selected blocks. (d) Number of sites selected for a prediction rate of $Q_l = 75\%$ and 3 PBs per site, with the "local stategy" (see text); the dots indicate the positions where the true PB is found again among the selected solutions. 49 among 62 selected sites give a correct prediction. (d) Number of sites selected for a prediction rate of $Q_l = 70\%$ and 3 PBs per site, the dots indicates the positions where the true PB is found again among the selected solutions. 95 among 122 sites give correct prediction.

Figure 8 : Relationship between the *Neq* index and the prediction rate for different ranks.

The prediction rate Q*(r) is assessed according to the *Neq* value (range = [1;8]) or the associated fraction of selected sites in the database. The rank r varies between 1 to 6.

Table legends :

Table I : Description of the Protein Blocks.

For each protein block (labelled from PBa to PBp), the occurrence frequency, the average root mean square deviation (*rmsd*), the average number of repeats (*anr*), the 3 main PB transitions, the repartition in secondary structures (helix-α, β-sheet and coil) of the central residue and a coarse characterization are given.

Table II : Significant amino acids in protein blocks.

For each position (indexed from -7 to +7) of the 16 PBs, the highest amino acid over-representations (Z-score $>$ +4.4) and under-representations (Z-score $<$ -4.4) labelled by the symbols + and - respectively are given. The zone corresponds to the positions detected the most informative by thresholding the KLd profile.

Table III : Example of prediction of the *2aak*'s N-terminal.

For each window of 15 amino acids, the true PB, the *Neq* index and the three most probable PBs with their own scores R_k are given. The underlined PB corresponds to the true one.