captions.

Figures.

Figure 1. Superimposition of fragments of the structural alphabet. The backbone is superimposed with atoms Cα, N, O and C along the 5 amino acids.

Figure 2. The learning steps of HPM (for further details see Methods section): (a) every protein is coded as Protein Blocks; (b) the structural database is created; (c) a local structure of length L=10 PBs is chosen randomly; (d) a similarity score is computed along the hybrid protein; (e) the higher score is found, (f) the hybrid protein is slightly modified, and the process starts again from (c).

Figure 3. The hybrid protein and its characteristics: (a) the final hybrid protein, with frequency of each PB: more than 35% in black, between 35% and 10% in grey, less than 10% in white; (b) the number of observations per site along the hybrid protein; (c) local entropy computed from PB distribution per site; (d) *root mean square deviation* calculated from the local structures, 10 Cα long, located at every site.

Figure 4. Three examples of local structure prototypes: (a) superimposition of local structures located in site 7 [position 3 to 12]; (b) site 73 [69:78]; (c) site 56 [52:61]; (d to e) amino acid frequencies in the previous positions, normalized into Z-scores: Z-scores more than 1.96 in black, less than -1.96 in white, in gray otherwise.

Figure 5. Amino acid specificities. (a) amino acid frequencies associated with every central position of the prototypes, normalized into Z-scores (more than 1.96 in black, less than -1.96 in white, in gray otherwise); (b) the KLd-indices computed from those amino acid frequencies, shown for a limit of χ_2 at 36 for 19 degrees of freedom and a probability of 0.05; (c) the k-means clustering on the Z-scores of the amino acid distribution into 12 clusters, numbered from left to right and up to down from 1 to 12. The Z-scores per amino acid are shown only in the range [-6;+6]. The amino acids are ranked in this order: I, V, L, M, A, F, Y, W, C, P, G, H, S, T, N, Q, D, E, R, K.

Figure 6. Positions indexes of the local structures along two cytochromes P450 in the hybrid proteins P450$_{terp}$ (a) and P450$_{BM3}$ (b) .

Figure 7. Superimpositions of the local structure pairs found to be structurally similar in 450$_{terp}$ and P450$_{BM3}$.

Tables.

Table 1. Structural alphabet. The 16 PBs obtained with their frequencies in the database, the *root mean square deviation* or *rmsd*, the *average number of repeats* or *anr*, and a crude clustering according to the standard 3-state alphabet.

Table 2. Common local structures found in 450_{terp} and $P450_{BM3}$. The 11 common local structures, with their number, the beginning and the end for $P450_{terp}$ and $P450_{BM3}$, the number of residues, the Protein Blocks identity, the amino acid identity, the *root mean square deviation* from $P450_{terp}$ and from $P450_{BM3}$ local structures, the correspondence with the Common Structural Blocks (CSBs) of Jean and coworkers and finally the secondary structure description by Haseman and co-workers.

In the last columns, / designates a local structure not found in the CSB; (1) is for the local structure **VI** which is only a part of the CSB 4; (2) local structure **VII** shows the same zone as CSB 4 for $P450_{terp}$, but is distinct for $P450_{BM3}$; (3) local structure **X** encompassed CSB 10 and a part of CSB 9; (4) the label is not found at the same place for one of the two proteins, but is in the immediate neighborhood (less than 50 residues); (5) the label does not exist in one of the two structures; and (6) the label is found for one of the two proteins a very long way from the position in the secondary structure classification (more than 50 residues).