Detailed description of the analysed test cases

1) AA_TRNA_LIGASE_II_1, AA_TRNA_LIGASE_II_2, ASP_PROTEASE, EGF_1, LIPOCALIN and RRM_RNP_1 PROSITE patterns

The AA_TRNA_LIGASE_II_1 and AA_TRNA_LIGASE_II_2 PROSITE patterns are derived from one out of the three regions, known to be conserved across the class-II aminoacyl-tRNA synthetases [1,4].

The ASP_PROTEASE PROSITE signature is built around the aspartate active site residue of the eukaryotic and viral aspartyl protease enzymes [3-5].

EGF (epidermal growth factor) [6,7] and the EGF-like domain present in very diverse proteins are characterised by six conserved cysteines and two conserved glycines. The PROSITE EGF_1 signature is built around one of such conserved glycines and includes three out of the six conserved cysteines.

The LIPOCALIN PROSITE signature is derived from a stretch of residues common to all the proteins belonging to the lipocalin protein family [8,9].

Single-stranded RNA binding proteins are characterised by one or more copies of a putative RNA-binding domain. This domain, about 90 amino acids long [10,11], displays two highly conserved regions. The first region is used to build the RRM_RNP_1 PROSITE pattern.

For each PROSITE pattern, two extended patterns, displaying better correlation than the PROSITE original one and comparable correlation (C) between them, are proposed: the former pattern with higher sensitivity and the latter with higher selectivity (see Table 4). The PROSITE region of the former is unmodified while the PROSITE region of the latter
is slightly ‘softened’ (see Table 3). Choice between the two motifs when analysing novel sequences will therefore depend on a preference for more sensitivity or selectivity.

The *extended* 1 patterns generally match a lower number of false positives and the same number of true positives and false negatives with respect to the corresponding PROSITE patterns on both the SWISS-PROT releases used as optimisation and test dataset (Table 2 and 3). There are two exceptions: the AA_TRNA_LIGASE_II_2 *extended* 1 pattern selects one more false negative on the SWISS-PROT 41.0, with respect to the corresponding PROSITE pattern, while the EGF_1 *extended* 1 pattern has three false negatives more on the SWISS-PROT 41.0, with respect to the corresponding EGF_1 PROSITE pattern. However, after scanning the SWISS-PROT 41.0 database with an EGF_1 *extended* pattern with a longer spacer between the PROSITE positions and the group of supplementary positions [x(1,50) instead of x(1,30), see regexp in Table 3], these three false negatives become true positive matches. Also the number of false positives increases, though remaining lower than that of the EGF_1 PROSITE pattern. This means that, in the case of the EGF_1 PROSITE pattern, even three *extended* patterns might be considered, for some purposes.

The *extended* 2 patterns select less false negatives than the corresponding PROSITE patterns do. This implies a greater number of true positives and, in every case but one (the AA_TRNA_LIGASE_II_1 *extended* 2 pattern), a lower number of false positives also. The AA_TRNA_LIGASE_II_1 *extended* 2 pattern matches on both the SWISS-PROT releases a slightly higher number of false positive sequences than the original PROSITE pattern does.
All the remaining true positives that have been missed (see table 2 and 3, AA_TRNA_LIGASE_IL_2, ASP_PROTEASE and LIPOCALIN) are partial sequences of the extended patterns, namely sequences that are not matched because they are fragments. The number of such partial sequences is underlined in the corresponding columns of Table 3.

2) **THIOL_PROTEASE_HIS**

The THIOL_PROTEASE_HIS PROSITE pattern is built around the histidine residue of the Cys-His-Asn catalytic triad that characterises the proteolytic enzymes belonging to the eukaryotic thiol proteases family [12]. In this case only one extended pattern was built, which displays a better performance with respect to the PROSITE original one. Such a pattern (see Table 2 and 3, last rows) matches the same number of true positives as the PROSITE one and displays selectivity and specificity values equal to one, which means that it detects NO false positives (on both the SWISS-PROT releases used).

3) **CYTOCHROME_C**

The CYTOCHROME_C pattern is built around the heme-binding site of the cytochrome c protein family [13,14] and contains two cysteine residues known to be bound to the heme group and a histidine residue, which is one of the two axial ligands of the heme iron.

In the case of the CYTOCHROME_C PROSITE pattern the computational and the visual analysis of the heme-binding site region did not highlight the presence of structurally conserved residues across the entire set of structures aligned, except for some residues
belonging to the short three-dimensional fragment corresponding to the PROSITE pattern. Therefore, the corresponding R-HET was only partially filled with Heavy Elements (see above). Indeed, the heme-binding site occupies a very small spatial region and is found in proteins with very different folds. Even if in many cases the local structure of the heme-binding site comprises similar loops, there are structures (e.g. the cytochrome c 1cpq structure) for which the 3D match occurs in a completely different local structure namely in a-helix, although still interacting with the heme. Mondal et al. (15) showed that, at least in some cases, true PDB hits of a PROSITE pattern display structural plasticity depending on the context (e.g. interaction with ligands and DNA). Therefore, when creating a 3D template for a PROSITE pattern, it might be important to take into account all the known distinct conformational states of the pattern. Similar conclusions are drawn by Lin et al. [16]. Maybe a more refined analysis of the nr-PDB true positive structures of the CYTOCHROME_C PROSITE pattern would make it possible to cluster structures with a similar 3D conformation of the pattern (e.g. bound to the heme). Structures belonging to the same group may display a greater number of conserved residues on the surface region surrounding the heme-binding site. Actually, some residues were found to be conserved in small subsets of structures and were used to build a rough extended sequence pattern. However, such CYTHOCROME_C rough extended pattern gave rise only to extended sequence patterns, which performed very poorly.
1. Cusack S, Haertlain M, Leberman R: **Sequence, structural and evolutionarily relationship between class 2 aminoacyl-rRNA synthetases.** *Nucl. Acid Res* 1991, **19**: 3489-3498.

2. Delarue M, Moras D: **The aminoacyl-tRNA synthetase family: modules at work.** *Bioessays* 1993, **15**: 675-687.

3. Foltmann B: **Gastric proteinases--structure, function, evolution and mechanism of action.** *Essays Biochem* 1981, **17**: 52-84.

4. Davies DR: **The structure and function of the aspartic proteinases.** *Annu Rev Biophys Biophys Chem* 1990, 19: 189-215.

5. Rao JK, Erickson JW, Wlodawer A: **Structural and evolutionary relationships between retroviral and eucaryotic aspartic proteinases.** *Biochemistry* 1991, **30**: 4663-4671.

6. Doolittle RF, Feng DF, Johnson, M.S: **Computer-based characterization of epidermal growth factor precursor.** *Nature* 1984, **307**: 558-60.

7. Appella E, Weber IT, Blasi F: **Structure and function of epidermal growth factor-like regions in proteins.** *FEBS Lett* 1988, **23**: 1-4.

8. Godovac-Zimmermann J: **The structural motif of beta-lactoglobulin and retinol-binding protein: a basic framework for binding and transport of small hydrophobic molecules?** *Trends Biochem Sci* 1988, **13**: 64-6.

9. Cowan SW, Newcomer ME, Jones TA: **Crystallographic refinement of human serum retinol binding protein at 2A resolution.** *Proteins* 1990, **8**: 44-6
10. Dreyfuss G, Swanson MS, Pinol-Roma S: Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem Sci* 1988, **13**: 86-91.

11. Bandziulis RJ, Swanson MS, Dreyfuss G: RNA-binding proteins as developmental regulators. *Genes Dev* 1989, **3**: 431-7.

12. Dufour E: Sequence homologies, hydrophobic profiles and secondary structures of cathepsins B, H and L: comparison with papain and actinidin. *Biochimie* 1988, **70**: 1335-1342.

13. Mathews FS: The structure, function and evolution of cytochromes. *Prog Biophys Mol Biol* 1985, **45**: 1-56.

14. Ambler RP: Sequence variability in bacterial cytochromes-c. *Biochim. Biophys. Acta* 1991, **1058**: 42-47.

15. Mondal S, Jaishankar SP, Ramakumar S: Role of context in the relationship between form and function: structural plasticity of some PROSITE patterns. *Biochem Biophys Res Commun* 2003, **305**: 1078-1084.

16. Lin KY, Wright J, Lim C: Conformational analysis of long spacers in PROSITE patterns. *J Mol Biol* 2000, **299**: 537-48.