# SCOP protein families fingerprints

Fabris F[1]

## Motivation

The tree structure of the SCOP database proteins classification is based on four levels: families, superfamilies, folds and classes. Proteins of the same family are structurally and biologically related, with a classification guided by human expertise. Nevertheless SCOP suffers an important draw-back, which is the lack of numerical parameters characterizing each family, so as to check the possible membership to a family by using some numerical quantifications. Based on the BLOSUM Spectrum [1], we characterize each family by means of a cloud of points, the fingerprint of the family, that lies on a Cartesian diagram. This can be used for: 1) discriminating stronger from weaker related sequences inside the same family; 2) disclosing a fine-grained classification inside each family, based on sub-clusters sharing a very close structure; 3) classifying a novel protein on the basis of the parameters defined above.

## Methods

Since the Mutual Information I(X,Y) and the Informational Divergence D(Fxy//Pab) characterize each couple of aligned sequences in terms of correlation (sequence convergence) and philogenetic distance (target frequency divergence), we inspect their behavior by performing all-versus-all alignments between members of the same SCOP family. We call Protein Family Fingerprint (PFF) the cloud of points we obtain in the D(Fxy //Pab) vs I(X,Y) diagram. Different families, even the ones belonging to the same fold or superfamily, are characterized by clouds of different shape. The alignments can be made with or without gaps. To generate each PFF, we downloaded the file containing all sequences from the SCOP database (release 1.71, 3004 families) and computed the all-vs-all alignments between the possible say N sequences, generating N(N-1)/2 points constituting the fingerprint. The PFF can lay on different areas of the same diagram: high correlation and typical matching, high correlation and untypical matching, and the grey area of weakly related sequences. Clusters of points inside a PFF derive from clusters of more related proteins inside the corresponding family. Also, the PFF could be used to tentatively classify a novel protein, not already inserted inside the SCOP database, by comparing the query with all the sequences of a certain family, and by repeating this procedure for all families conjectured to be good candidates for accepting the query as their own member. We call relative fingerprint the cloud of points derived from the query vs family sequences alignments. If the query effectively belong to

[1] Dipartimento di Matematica e Informatica, Universita' di Trieste

the inspected family, then the relative fingerprint should belong to the fingerprint area of the PFF, with points scattered inside the possible clusters.

**Results**

We have built a fingerprint for each of the 3004 families. Figure 1 shows some examples (colored points). As for the classification method, we have qualitatively tested it by checking if a certain probe sequence, surely belonging to a family, could have been correctly classified by the method. We have performed three tests. In the first two we have chosen the first sequence of the a.1.1.2 Globin and of f.13.1.1 Bacteriorhodopsin-like family to build the relative fingerprints associated to a certain set of families, one for each SCOP class. Then we have decided to test the procedure for two families belonging to the same superfamily, fold and class, that are usually difficult to be separated; they are b.60.1.1 Retinol binding protein-like and b.60.1.2 Fatty acid binding protein-like. The relative fingerprints for b.60.1.2 are reported in Figure 1 (black points), and as expected the relative fingerprint associated to the correct family is constituted by scattered points that overlap the correct PFF, while in all other cases the points are outside the fingerprint, and/or are not scattered. [1] Fabris F, Sgarro A. and Tossi A. Splitting the BLOSUM score into numbers of biological significance J. on Bioinformatics and Systems Biology, 2007

**Availability**

http://www.dmi.units.it/~fabris/BLOSpectrum

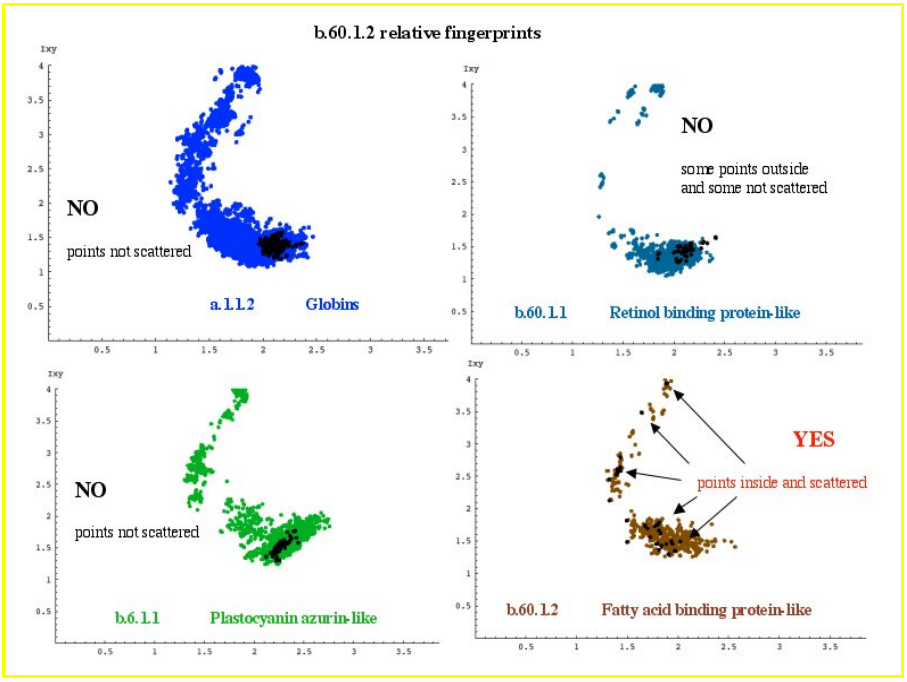
**Contact e-mail**

frnzfbrs@dmi.units.it

**Supplementary information**

This work was supported by FIRB 2003 (LIBI)

## Image


b.60.1.2 relative fingerprints
Ixy
NO
points not scattered
a.1.1.2 Globins
NO
some points outside
and some not scattered
b.60.1.1 Retinol binding protein-like
NO
points not scattered
b.6.1.1 Plastocyanin azurin-like
YES
points inside and scattered
b.60.1.2 Fatty acid binding protein-like