Basic Exploratory Proteins Analysis with Statistical Methods Applied on Structural Features

Eugenio Del Prete, Serena Dotolo, Anna Marabotti, and Angelo Facchiano

Abstract Exploratory Data Analysis (EDA) is an approach for summarizing and visualizing the important characteristics of a data set, in order to make a prearranged data screening and display multivariate data in a graphical way, to render them more comprehensible. Moreover, it reveals hidden aspects within the simple evaluations. In particular, EDA is suitable for datasets with comparable variables, as structural-geometrical protein features. In this work, we analyzed some proteins belonging to ten different architectural families. After retrieval, feature selection and normalization stages, the dataset has been processed by means of simple correlation, partial correlation and principal component analysis (PCA), highlighting family-independent or family-specific relationships, and possible outliers for the dataset itself. The results can be useful to connect these features to functional protein properties.

Keywords Correlation • Exploratory data analysis • Global features • Principal component analysis • Protein structure

1 Background

Exploratory Data Analysis (EDA) is the process of looking through data to get a basic idea of their structures and attributes, often with visualizations. EDA is a graphical-statistical approach, almost a philosophy of research, applied to data in order to make some aspects clearer and answer some questions about them. It is like a magnifying glass that helps in:

E. Del Prete (✉) • S. Dotolo • A. Facchiano
Institute of Food Science, National Research Council, Via Roma 64, 83100 Avellino, Italy
e-mail: eugenio.delprete@isa.cnr.it; serenadotolo@hotmail.it; angelo.facchiano@isa.cnr.it

A. Marabotti
Institute of Food Science, National Research Council, Via Roma 64, 83100 Avellino, Italy

Department of Chemistry and Biology, University of Salerno, Via Giovanni Paolo II 132, 84084 Fisciano, SA, Italy
e-mail: amarabotti@unisa.it

V. Zazzu et al. (eds.), *Mathematical Models in Biology*,
DOI 10.1007/978-3-319-23497-7_13

- leading towards a right interpretation of data;
- showing and summarizing data in a clear way;
- finding underlying relationships among observations and, main thing, among variables.

It can be univariate or multivariate and can use graphical or not graphical methods. A historical explanation can be found in [1]. Main point is that EDA is essential in understanding data, because it can reinforce or undermine *a priori* knowledge about observations and prepare data for the following inference step.

In the analysis of large data sets, an inevitable phase that must anticipate the statistical analysis concerns getting and cleaning data. Data can be obtained from a variety of sources: downloaded from online repositories, streamed on-demand from online sources, automatically generated by physical apparatus interfaced to a computer, generated by a computer software, manually entered in a spreadsheet or text file. Data origin, management and storage are other issues related to the getting part of the data analysis. Raw data retrieved are probably not in a convenient format, because of semantic errors, missing entries, inconsistent formatting. Thus, it is recommended to make a control on all variables and, if necessary, integrate new ones from different sources that are coherent with the previous ones, in order to create a tidy final dataset [2].

Investigations on protein structure and function represent a field of research in which experimental techniques as well as computational methods are widely applied [3–6]. Nevertheless, many aspects are still unsolved, in particular concerning the relationships between structure and function of proteins. While successful methods have been developed to "predict" the complex three-dimensional structure of a protein from a simple structural information as the amino acid sequence, and are largely applied in literature and by our research group [7–9], it is less investigated the deep nature of the structural features and their relationships with protein function. In other words, evolution may modify the amino acid sequence of an ancestral protein at a large extent among living species, thus affecting the lower level of structural organization of a protein family member. This has low impact on the three-dimensional structure, i.e. the higher level of structural organization, so that the protein family maintains its specific biochemical function over the species. On the other hand, a single amino acid substitution within a protein can strongly affect structure and function, as in human pathologies due to genetic diseases [10–12]. However, it is still unclear in detail how the modification of amino acid sequence is softened or emphasized when it is reflected at the functional level. In this context, we are interested to exploit graphical and mathematical methods, poorly used in protein science up to now, in order to explore protein structure and function relationships from a new point of view.

In this study, multivariate graphical methods have been used, because of tabular (observations—variables) data type. Data are composed of protein families chosen for their functional similarity; it is interesting to examine protein structure and analyze conformational features within a family and among the families, in order to find relationships that could be related to functional properties. Ten protein families have been chosen, depending on CATH different architectural classification [13].

2 Materials and Methods

2.1 *Analysis Workflow*

The workflow (Fig. 1) consists of four steps, of which the first three concern getting and cleaning data, whereas the fourth step is the real EDA. More in details:

- ***Step 1***. 153 crystallographic structures (Table 1) have been retrieved from RCSB PDB [14]. The structures have been selected to represent ten structural protein families, and different architectural classes in CATH. Rules applied to select structures to be analyzed are: families for which a similar number of structures is available (i.e., in the range 13–19); within each family, only one chain per protein (in homo-multimeric proteins, A chain), and structures which differ for less than 50 residues in length;
- ***Step 2***. Different online and local tools have been used to extract protein structural properties from PDB files: *Vadar* [15], for secondary structure (also confirmed with DSSP [16]), hydrogens bonds, accessible surface areas, torsion angles, packing defects, charged residues numbers, free energy of folding; *McVol* [17] for volumes, with a mean difference of 4 % from that extracted from Vadar, but using a more robust algorithm; an in-house-developed *R-script* for automatic search of salt bridges conditions [18]. Parsing of Vadar results has been performed by means of regular expression commands in *R*;

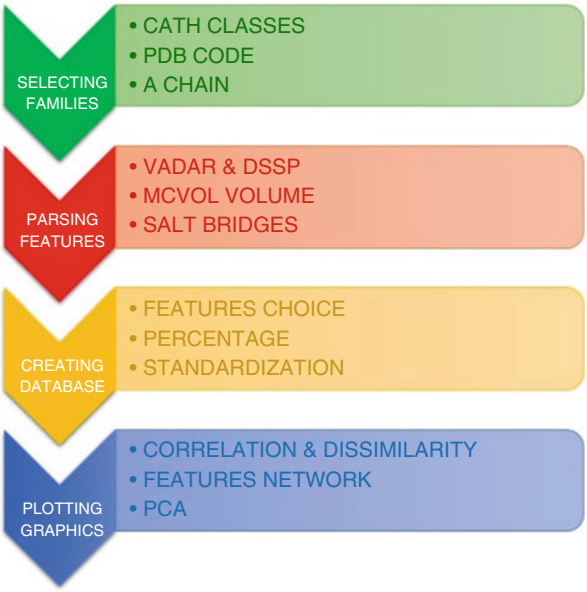

Fig. 1 Analysis graphical workflow. There are four simply identifiable steps: steps 1–3 composed the getting and cleaning data part, step 4 is the Exploratory Data Analysis

Table 1 Protein families and PDB structures

CATH code	PDB files
Beta-Lactamase (BLA) 3.30.450, α/β 2-layer sandwich	1D1J, 1EW0, 1F2K, 1N9L, 1P0Z, 2V9A, 2VK3, 2VV6, 2ZOH, 3BW6, 3BY8, 3CI6, 3CWF, 3EEH
Cathepsin B (CTS) 3.90.70, α/β complex	1AEC, 1B5F, 1S4V, 2B1M, 2BDZ, 2DC6, 2P7U, 2WBF, 3AI8, 2BCN, 3CH2, 3LXS, 3P5U
Ferritin (FTL) 1.20.1260, α up-down bundle	1JI4, 1QGH, 1R03, 1S2Z, 1TJO, 2FKZ, 2XJM, 2YW6, 3AK8, 3E1J, 3KA3, 3MPS, 3R2H, 3RAV
Glycosyltransferase (GTF) 1.50.10, α-α barrel	1GAH, 1HVX, 1KRF, 1KS8, 1NXC, 1R76, 1X9D, 2NVP, 2P0V, 2XFG, 2ZZR, 3P2C, 3QRY
Hemoglobin (HGB) 1.10.490, α orthogonal bundle	1CG5, 1FLP, 1GCW, 1HLM, 1RQA, 1UVX, 2C0K, 2QSP, 2VYW, 3BJ1, 3NG6, 3QQR, 3WCT, 4IRO, 4NK1
Lipocalin 2 (LCN) 2.40.128, β-β barrel	1AQB, 1BEB, 1CBI, 1CBS, 1GGL, 1GM6, 1IIU, 1JYD, 1KQW, 1KT6, 1LPJ, 1OPB, 1QWD, 2CBR, 2NND, 2RCQ, 2XST, 3S26, 4TLJ
Lysozyme (LYS) 1.10.530, α orthogonal bundle	1BB6, 1FKV, 1GD6, 1GHL, 1HHL, 1IIZ, 1JUG, 1QQY, 1REX, 1TEW, 2EQL, 2GV0, 2IHL, 2Z2F, 3QY4
Proliferating Cell Nuclear Antigen (PCNA) 3.70.10, α/β box	1AXC, 1B77, 1CZD, 1DML, 1T6L, 1UD9, 1UL1, 1HII, 1IJX, 2OD8, 3HI8, 3LX1, 3P83, 3P91, 4CS5
Purine Nucleoside Phosphorylase (PNP) 3.40.50, α/β 3-layer sandwich	1A9O, 1JP7, 1M73, 1ODK, 1PK9, 1QE5, 1TCU, 1V4N, 1VMK, 1XE3, 1Z33, 2P4S, 3KHS, 3OZE, 3SCZ, 3TL6, 3UAV, 4D98
Superoxide Dismutase (SOD) 1.10.287, α orthogonal bundle 2.60.40, β sandwich	1BSM, 1IDS, 1JCV, 1MA1, 1MMM, 1MY6, 1Q0E, 1WB8, 2ADP, 2JLP, 2W7W, 3BFR, 3ECU, 3EVK, 3LIO, 3QVN, 3SDP

Left column reports the name of the protein, short name, CATH code, and architecture; right column reports the PDB codes

- ***Step 3***. Among all the variables extracted by means of Vadar, percent features have been preferred for their intrinsic homogeneity. More in details, the features related to residues have been transformed in a percent form by means of protein sequence length; on the other hand, the ones related to surfaces have been transformed by means of total accessible surface area. Furthermore, they have been normalized in a standard score form for a better stability relative to the EDA. That is, mean value has been subtracted from the data and the result has been divided by the standard deviation: details are described in the Sect. 2.3. Redundant features, as expected values, have been ignored;
- ***Step 4***. Variables have been transformed into correlation and dissimilarity matrices, through the procedures explained under Sect. 2.3.1. Then, they have been used as features for an overall PCA, in order to verify the existence of common information. A comparison with a features network has been showed.

All the work has been executed with *R* [19] inside *R Studio IDE*, using some specific *R packages* to perform getting and cleaning phase and EDA. In particular: *stringr*, to rearrange file names [20]; *RCurl*, to manage connection for downloading [21]; *bio3d*, to compute DSSP inside R and read PDB files [22]; *corrplot*, to plot graphical correlation matrix [23]; *Hmisc*, to calculate correlation matrix with p-value [24]; *ppcor*, to calculate partial and semi-partial correlations with p-value [25]; *dendroextras*, to readjust and color dendrogram [26]; *ggplot2*, to plot PCA clustering [27]; *GeneNet*: to plot features network [28].

2.2 Statistical Methods

As part of EDA, two proven statistical procedures have been chosen for our work: correlation and principal component analysis [29], with different developments and additional interpretations.

Correlation has been performed as Pearson's correlation coefficients between pairwise features. Its practice must be carefully implemented, because of a batch of well-known traps (causality, multi-collinearity, outliers and so on). Statistical validation, performed here, procures only a quantitative robustness: an incisive analysis, together with a knowledge of data, allows to reach non-misleading conclusions. Partial correlation can helps with collinearity problem, taking away the effects of another variable, or several other variables, on a relationship. Moreover, it can be used to detect possible redundant features.

Principal component analysis (PCA) is a very common multivariate statistical method, simple and powerful: it is an unsupervised approach and it is considered an EDA method. It allows summarizing initial variables in new ones, so-called components, which represent data in a more compact way and their tendency. Furthermore, given the intrinsic orthogonality of the components, PCA can be applied to obtain a kind of clustering [30], depending on inner information derived

from explained variance. This grouping helps in seeking possible outliers when executed on a dataset (it is a good habit searching for outliers, because they could polarize inferred results).

2.3 Mathematical Overview

2.3.1 Correlation, Partial Correlation and Dissimilarity

Given two variables with continuous values $X = (x_1, \ldots, x_r)$ and $Y = (y_1, \ldots, y_r)$., where r is rows-observations number and c column-variables number, the density $f(x_i, y_j)$ is represented by a single element in the normalized data table, and it is just a sort of bivariate distribution in a numerical form. A measure of strength and direction of association between the variables is provided by the covariance:

$$\sigma_{xy} = E\left[(X - \mu_x)\left(Y - \mu_y\right)\right] = E[XY] - \mu_x \mu_y \tag{1}$$

where

$$E[XY] = \sum_{i=1}^{r} \sum_{j=1}^{c} x_i y_j f\left(x_i, y_j\right) \tag{2}$$

where μ_x, μ_y are the expected values for a single variable. An index of covariation between X and Y is provided by the correlation coefficient:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3}$$

where σ_x, σ_y are the standard deviations for a single variable. Given a third variable Z, the partial correlation coefficient between X and Y after removing the effect of Z is:

$$\rho_{yx \cdot z} = \frac{\rho_{yz} - \rho_{yx}\rho_{zx}}{\sqrt{1 - \rho_{yx}^2}\sqrt{1 - \rho_{zx}^2}} \tag{4}$$

and it is possible to extend the formula in case of removing the effect of all the variables but one [31, 32]. Furthermore, a transformation from correlation to dissimilarity, by means of the formula:

$$d_{xy} = 1 - \left|\rho_{xy}\right| \tag{5}$$

allows to obtain a distance matrix, consistent with a cluster dendrogram on the variables themselves. d_{xy} is also known as Pearson's distance [33, 34]. Finally, every correlation coefficient has been validated with a t-test for significance, with the statistic:

$$t = \rho \sqrt{\frac{n-2}{1-\rho^2}} \tag{6}$$

where ρ is a generic correlation and $n-2$ are the degrees of freedom [32].

2.3.2 Principal Component Analysis

Give a data table in a matrix form, it is possible to create new variables as linear combination of the old ones:

$$\begin{cases} PC_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1c}X_c \\ PC_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2c}X_c \\ \qquad\qquad \cdots \\ PC_l = a_{l1}X_1 + a_{l2}X_2 + \cdots + a_{lc}X_c \end{cases} \tag{7}$$

that have the largest variance. For a single principal component loading vector $a_m = (a_{11}, \ldots, a_{1c})^T$, with $m = 1, \ldots, l$, it is required to resolve an optimization problem:

$$\max_{a_m} \left\{ \frac{1}{r} \sum_{i=1}^{r} \left(\sum_{j=1}^{c} a_{1c} x_{ij} \right)^2 \right\} \tag{8}$$

subject to $\sum_{j=1}^{c} a_{1j}^2 = 1$. This is an eigenvalues-eigenvectors problem, numerical and computationally resolvable with Single Value Decomposition factorization, with a_m determined by:

$$(\Sigma - \lambda_m I)\, a_m = 0 \tag{9}$$

where Σ is the covariance/correlation matrix of the original data, λ_m are eigenvalues in descending order associated with a_m eigenvectors and I is the identity matrix. After calculating the contribution of every eigenvalue $\lambda_m / \sum_{k=1}^{l} \lambda_n$, it is possible to choose the first several λ_m that cover a preset quantity of explained variability. In other words, the new data table composed by scores PC_k, always in matrix form,

represents the old one with a reduced dimensionality. Scores and loading vectors are plotted in a single biplot display [35, 36]. The challenge with this method is the new variables interpretation in the reality: that is, they are not so intuitive and their understanding is often delegated to investigator's experience.

2.3.3 Standardized Variables

Also known as z-score or standard score, a standardized variable has a mean equal to zero and a variance (standard deviation) equal to one, and it is possible to obtain it by means of the linear transformation:

$$z_x = \frac{\mathrm{x} - \mu_x}{\sigma_x} \tag{10}$$

useful for comparing same variables from different distributions or variables with different units of measurement. This kind of normalization is recommended when correlations have to be used [32].

3 Results

Dendrogram in Fig. 2, obtained following formula (5), highlights relationships between the features chosen for the entire proteins dataset: it is the landmark about structural and geometrical features, but only in reference to the proteins chosen for assembling the dataset. There are four evident clusters: from the left, the first and the third concern torsion angles, the second concerns volume, free energy of folding and residues buried for the most part, and the fourth concerns secondary structures and residues convolved in hydrogen bonds.

Features network in Fig. 3 has been plotted by means of partial correlations and graphical Gaussian model (GGM) [37]: it helps in seeking spurious correlations and pruning excessive features. For this dataset, torsion angles information results peripheral in the network, therefore they can be considered as unnecessary for the purpose of the work.

PCA, performed on the whole dataset, allows to extract the real important features in term of variability, producing a sort of clustering. In Fig. 4, the first principal component is composed by structural features (%A, %RHB) and second principal component by energy-geometrical ones (VOL, FEF, %RB95). This statistical technique is useful for outliers detection: for example, in the same plot, an isolated protein results so distant that it must be consider an outlier not only for its family (SOD), but also for the entire dataset. PCA performed only on SOD family has confirmed the result.

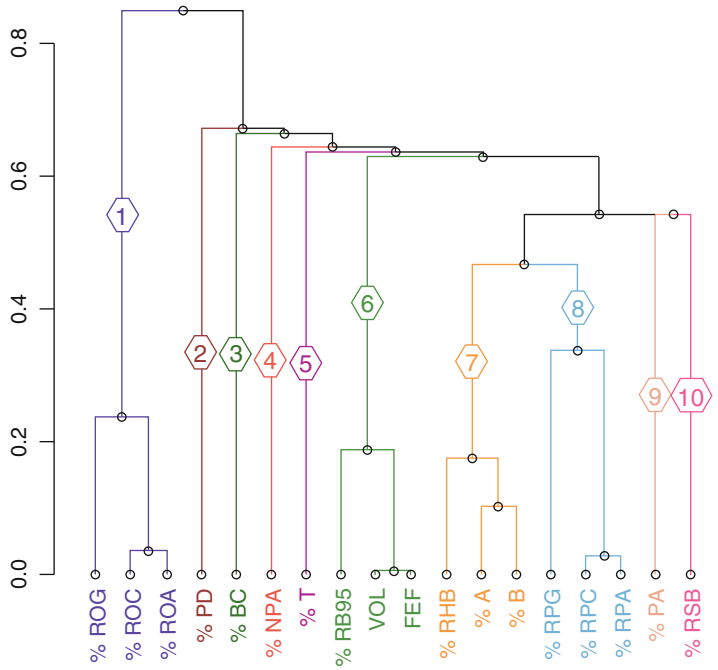

Fig. 2 Dissimilarity dendrogram for proteins dataset. Every number (and color) indicates a cluster for the features. Cut-off has been put at 0.4, as deduced from the grafico. *Legend*: *ROx* omega angle core/allowed/generous, *PD* packing defect, *BC* buried charge, *NPA* non polar accessible surface area, *T* turn, *RB95* buried 95 %, *VOL* volume, *FEF* free energy folding, *RHB* hydrogen bond, *A* alpha helix, *B* beta sheet, *RPx* phi-psi angles core/allowed/generous, *PA* polar accessible surface area, *RSB* salt bridge

Moreover, previous plot questions if some relationships between the features are family-independent. A graphical correlation matrix for a single family protein may answer to this query. For example, choosing SOD family in Fig. 5 as test, it is possible to notice a strong family-specific "four-relationship" in the bottom left corner, between buried charged residues, secondary structure and free energy of folding. In this work, strong correlation threshold is 0.65, deduced from data. By contrast, some relationships are family-independent: for example, because of intrinsic physical-conformational connection (secondary structure and residues involved in hydrogen bonds) or prediction formula (volume and free energy of folding [15]).

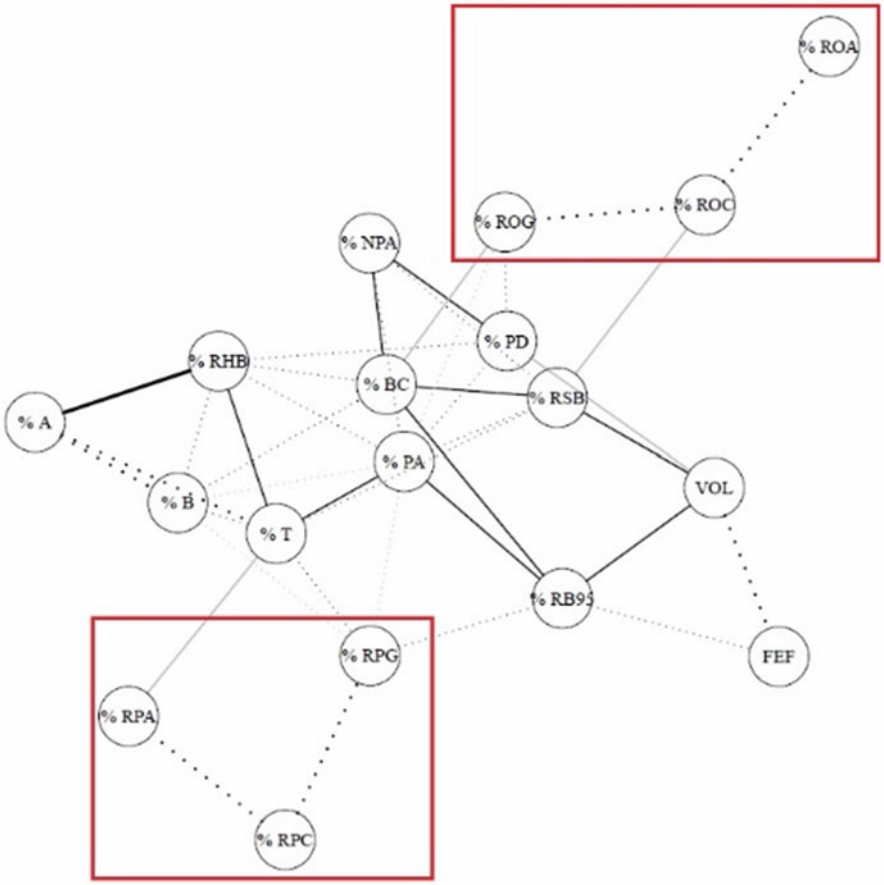

Fig. 3 Features network for proteins dataset. *Continuous line* represent partial correlation, *dotted line* represent partial anticorrelation (with the support of GGM). Peripheral subnetworks have been showed in the *squares*, which contain phi, psi and omega angle features. Meaning of acronyms as in Fig. 2

4 Conclusions and Perspectives

All the procedures that are part of EDA are well-suited for this kind of multivariate data: (a) distance dendrogram shows an overview about features interactions; (b) partial correlation indicates some possible redundant feature, if integrated in a network algorithm; (c) simple correlation helps in seeking family-specific features relationships; (d) principal component analysis is useful in finding family-specific connections to features and possible outliers. Therefore, these graphical multivariate procedures may be good tools in order to create a sort of fingerprint for the protein families themselves.

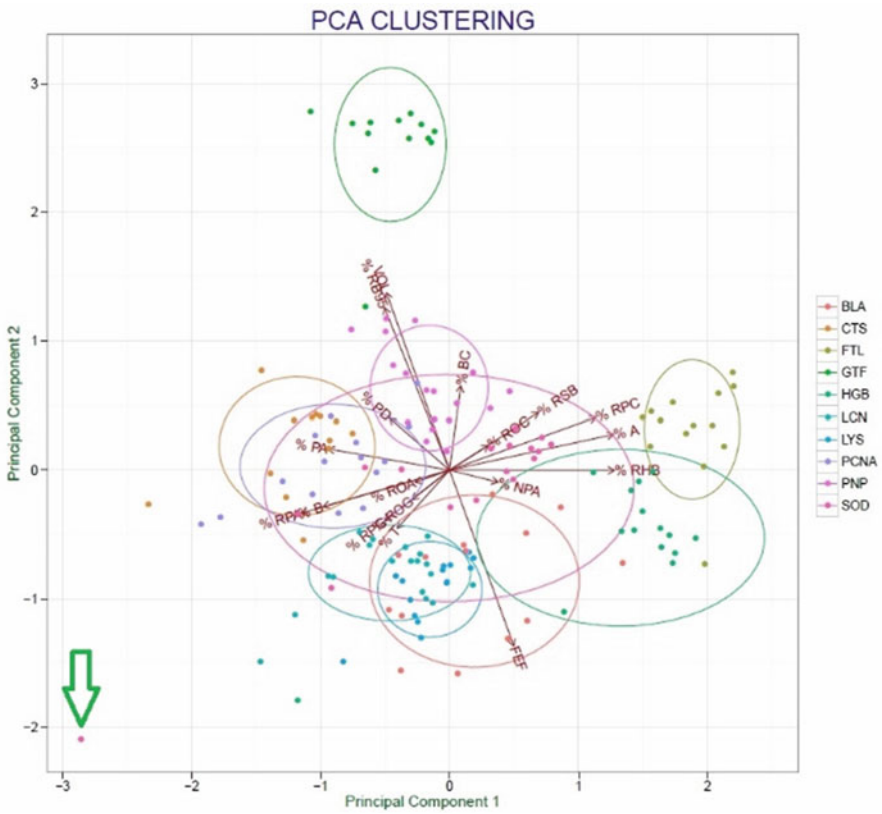

Fig. 4 PCA for protein dataset. Centralizing ellipses enclosed each protein family. GTF family is polarized near positive PC2, FTL family near positive PC1 and SOD family is wide open. *Bottom left arrow* points to an outlier: *Pseudomonas putida* SOD A chain (PDB code: 3SDP). Protein families short names refer to legend in Table 1

As future perspective, there are two directions of work enhancement: using advanced regression analysis to make a more robust features selection—partial correlation aids to do this, on the other hand PCA is not a real feature selection technique: it is rather a sort of "compression features" method—and integrating functional information (for example, by the analysis of protein interaction networks, as shown in Fig. 6) to highlight connections with the structural-geometrical ones.

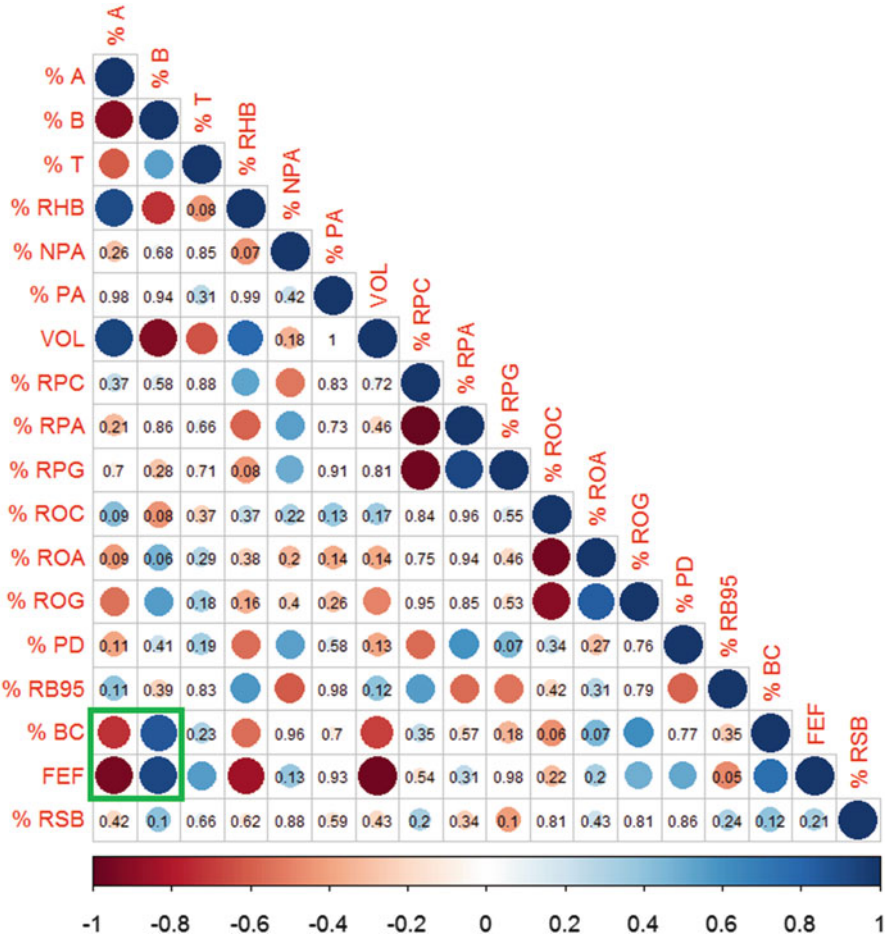

Fig. 5 Circular (lower) triangular correlation matrix for SODs. *Circle dimension* represents correlation strength, while *circle color* represent correlation direction (*blue*: correlation, *red*: anticorrelation). In the *green square*, there is a closed family-specific "four-relationship". Numbers in the matrix show correlation statistically non-significant (p-value > 0.05)

Acknowledgments This work is partially supported by the *Flagship InterOmics Project* (PB.P05, funded and supported by the Italian Ministry of Education, University and Research and Italian National Research Council organizations).

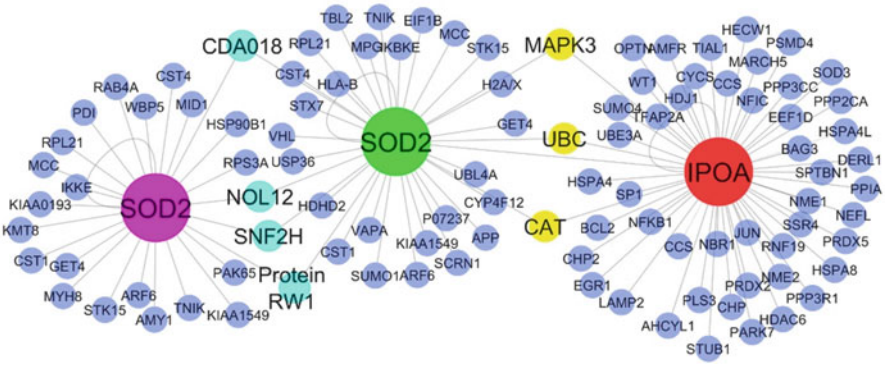

Fig. 6 SODs partial interaction network. IPOA is an alternative name of SOD1 soluble; every SOD2 mitochondrial has its gene connections taken from different online database. Network has been drawn with *Cytoscape* [38]

References

1. Tukey, J.W.: Exploratory Data Analysis. Behavioral Science: Quantitative Methods. Addison-Wesley, Reading (1977)
2. De Jong, E., Van der Loo, M.: An Introduction to Data Cleaning with R. Statistics Netherlands, The Hauge (2013)
3. Branden, C., Tooze, J.: Introduction to Protein Structure, 2nd edn. Garland Publishing Inc, New York (1999)
4. Facchiano, A.M., Colonna, G., Ragone, R.: Helix stabilizing factors and stabilization of thermophilic proteins: an X-ray based study. Protein Eng. **11**(9), 753–760 (1998)
5. Marabotti, A., Spyrakis, F., Facchiano, A., Cozzini, P., Alberti, S., Kellogg, G.E., Mozzarelli, A.: Energy-based prediction of amino acid-nucleotide base recognition. J. Comput. Chem. **29**, 1955–1969 (2008)
6. Russo, K., Ragone, R., Facchiano, A.M., Capogrossi, M.C., Facchiano, A.: Platelet-derived growth factor-BB and basic fibroblast growth factor directly interact in vitro with high affinity. J. Biol. Chem. **277**, 1284–1291 (2002)
7. Buonocore, F., Randelli, E., Bird, S., Secombes, C.J., Facchiano, A., Costantini, S., Scapigliati, G.: Interleukin-10 expression by real-time PCR and homology modelling analysis in the European sea bass (Dicentrarchu Labrax L.). Aquaculture **270**, 512–522 (2007)
8. Casani, D., Randelli, E., Costantini, S., Facchiano, A.M., Zou, J., Martin, S., Secombes, C.J., Scapigliati, G., Buonocore, F.: Molecular characterisation and structural analysis of an interferon homologue in sea bass (Dicentrarchus labrax L.). Mol. Immunol. **46**, 943–952 (2009)
9. Marabotti, A., D'Auria, S., Rossi, M., Facchiano, A.M.: Theoretical model of the three-dimensional structure of a sugar binding protein from Pyrococcus horikoshii: structural analysis and sugar binding simulations. Biochem. J. **380**, 677–684 (2004)
10. Marabotti, A., Facchiano, A.M.: Homology modelling studies on human galactose-1-phosphate uridylyltransferase and on its galactosemia-related mutant Q188R provide an explanation of molecular effects of the mutation on homo- and heterodimers. J. Med. Chem. **48**, 773–779 (2005)
11. Facchiano, A., Marabotti, A.: Analysis of galactosemia-linked mutations of GALT enzyme using a computational biology approach. Proteins Eng. Des. Sel. **23**, 103–113 (2010)

12. d'Acierno, A., Facchiano, A., Marabotti, A.: GALT protein database: querying structural and functional features of GALT enzyme. Hum. Mutat. **35**, 1060–1067 (2014)
13. Sillitoe, I., Cuff, A.L., Dessailly, B.H., Dawson, N.L., Furnham, N., Lee, D., Lees, J.G., Lewis, T.E., Studer, R.A., Rentzsch, R., Yeats, C., Thornton, J.M., Orengo, C.A.: New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acids Res. **41** (Database issue):D490–8 (2013). URL http://www.cathdb.info/
14. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. **28**, 235–242 (2000)
15. Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R.F., Sykes, B.D., Wishart, D.S.: VADAR: a web server for quantitative evaluation of protein structure quality. Nucleic Acids Res. **31**(13), 3316–3319 (2003)
16. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**(12), 2577–2637 (1983)
17. Till, M.S., Ullmann, G.M.: McVol - a program for calculating protein volumes and identifying cavities by a Monte Carlo algorithm. J. Mol. Model. **16**, 419–429 (2010)
18. Costantini, S., Colonna, G., Facchiano, A.M.: ESBRI: a web server for evaluating salt bridges in proteins. Bioinformation **3**(3), 137–138 (2008)
19. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2014). http://www.R-project.org/
20. Wickham, H.: stringr: Make it easier to work with strings. R package version 0.6.2 (2012). URL http://CRAN.R-project.org/package=stringr
21. Temple Lang, D.: RCurl: General network (HTTP/FTP/...) client interface for R. R package version 1.95-4.3 (2014). URL http://CRAN.R-project.org/package=RCurl
22. Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A., Caves, L.S.: Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics **22**(21), 2695–2696 (2006)
23. Wei, T.: corrplot: Visualization of a correlation matrix. R package version 0.73 (2014) URL http://CRAN.R-project.org/package=corrplot
24. Harrell Jr, F.E., Dupontm, C. and al.: Hmisc: Harrell Miscellaneous. R package version 3.14- 5 (2014) URL http://CRAN.R-project.org/package=Hmisc
25. Kim, S.: ppcor: Partial and Semi-partial (Part) correlation. R package version 1.0 (2012). URL http://CRAN.R-project.org/package=ppcor
26. Jefferis, G.: dendroextras: Extra functions to cut, label and colour dendrogram clusters. R package version 0.2.1 (2014). URL http://CRAN.R-project.org/package=dendroextras
27. Wickham, H.: A layered grammar of graphics. J. Comput. Graph. Stat. **19**(1), 3–28 (2010)
28. Schaefer, J., Opgen-Rhein, R., Strimmer, K.: GeneNet: Modeling and Inferring Gene Networks. R package version 1.2.10 (2014). URL http://CRAN.R-project.org/package=GeneNet
29. Ding, Y., Cai, Y., Han, Y., Zhao, B., Zhu, L.: Application of principal component analysis to determine the key structural features contributing to iron superoxide dismutase thermostability. Biopolymers **97**(11), 864–872 (2012)
30. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the 21st International Conference on Machine Learning, Banff, 2004
31. Jobson, J.D.: Applied Multivariate Data Analysis. Volume I: Regression and Experimental Design. Springer Texts in Statistics, 4th edn. Springer, New York (1999)
32. Edwards, A.L.: Multiple Regression and the Analysis of Variance and Covariance, 2nd edn. W.H. Freeman and Company, New York (1985)
33. Quinn, G.P., Keough, M.J.: Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge (2002)
34. Fulekar, M.H.: Bioinformatics: Applications in Life and Environmental Sciences. Springer, Heidelberg (2009)
35. Jolliffe, I.T.: Principal Component Analysis. Springer Series in Statistics, 2nd edn. Springer, New York (2002)

36. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: With Application in R. Springer Texts in Statistics. Springer Science + Business Media, New York (2013)
37. Schaefer, J., Strimmer, K.: An empirical Bayes approach to inferring large scale gene association networks. Bioinformatics **6**(21), 754–764 (2005)
38. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Ideker, T.: Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics **27**, 431–432 (2011)