BBCC2015
Dec.4th,2015
CNR-ISA, Avellino, Italy

FEATURE SELECTION ON A DATASET OF PROTEIN FAMILIES: FROM EXPLORATORY DATA ANALYSIS TO STATISTICAL VARIABLE IMPORTANCE

E. Del Prete[1,2], S. Dotolo[1,3], A. Marabotti[1,4], A. Facchiano[1]

1 Istituto di Scienze dell'Alimentazione, CNR, Via Roma 64, 83100 Avellino
2 Dip. di Scienze, Università della Basilicata, Viale dell'Ateneo Lucano 10, 85100, Potenza
3 Dip. di Biochimica, Biofisica e Patologia Generale, SUN, via De Crecchio 7, 80138, Napoli
4 Dip. di Chimica e Biologia "Adolfo Zambelli", Università degli Studi di Salerno, Via Giovanni Paolo II 132, 84084 Fisciano (SA)

Email contact address: angelo.facchiano@isa.cnr.it

Proteins are characterized by several typologies of features (structural, geometrical, energy). Most of these features are expected to be similar within a protein family. We are interested to detect which features can identify proteins that belong to a family, as well as to define the boundaries among families. Some features are redundant: they could generate noise in identifying which variables are essential as a fingerprint and, consequently, if they are related or not to a function of a protein family. We defined an original approach to analyze protein features for defining their relationships and peculiarities within protein families.

A multistep approach has been mainly performed in R environment: getting-cleaning data, exploratory data analysis and predictive modeling for classification. Ten protein families have been chosen by their CATH classification (different architectures), with rules over the number of structures, the length of the sequence and the choice of the chain. Properties investigated are secondary structures, hydrogen bonds, accessible surface areas, torsion angles, packing defects, number of charged residues, free energy of folding, volume and salt bridges. Kernel density estimation helps in discovering unusual multimodal profiles. Pearson's correlation highlights statistical links between pairwise variables and Pearson's distance provides a dendrogram with a clusterization of the features. PCA clusterizes the protein families and it detects outliers, sparse PCA performs a feature selection. Many classification algorithms have been used: decision trees (classical, boosting and bagging), SVMs (flexible discriminant analysis), centroid (nearest shrunken). The interest is on variable importance estimation. A 10-fold x 10 cross validation has been applied over the training set. Accuracy, K coefficient, sensitivity and specificity have been calculated for each methods.

From the density plots, the percentage of mostly buried residues is significantly different for each family. Dissimilarity dendrogram shows separated clusters for secondary structures, torsion angles, defects and geometrical features. From the features network, torsion angles and surface variables result as peripheral (i.e. redundant) from the core of the graph. PCA biplot gives a good clustering for the protein families and sparse PCA confirm dendrogram results. Unifying all the results, these features are typical for our dataset: helix, strand, coil, turn, hydrogen bond, polar and charged accessible surface area, volume and residue buried for the most part. Random forest algorithm has the best performance values.

Graphical multivariate procedures are good tools for the characterization of possible fingerprints about the protein families. Predictive models for classification and variable importance estimation help in performing feature selection. The work can be improved by the use of multivariate regression models and the increase of the protein families number.

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).

REFERENCES

Del Prete E, Dotolo S, Marabotti A, Facchiano A: "Statistical analysis of protein structural features: relationships and PCA grouping". Lecture Notes in Computer Sciences, 8623, 33-43, 2015

Grömping U, "Variable Importance Assessment in Regression: Linear Regression versus Random Forest", The American Statistician, 63:4, 308-319, 2009

Kuhn M, "Building Predictive Models in R Using the caret Package", Journal of Statistical Software, Vol. 28 (5), 2008, URL http://topepo.github.io/caret/index.html

Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA, "New functional families (Fun-Fams) in CATH to improve the mapping of conserved functional sites to 3D structures", Nucleic Acids Research, 2013

Zou H, Hastie T, Tibshirani R, "Sparse Principal Component Analysis", Journal of Computational and Graphical Statistics, 15 (2): 265-286, 2006