DOI 10.18699/MM-HPC-BBB-2018-11

Principal Component Analysis for any type Sequences (PCA-Seq)

V. Efimov[1, 2, 3, 4]*, K. Efimov[5], V. Kovaleva[2]
[1] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2] *Institute of Systematics and Ecology of Animals SB RAS, Novosibirsk, Russia*
[3] *Novosibirsk State University, Novosibirsk, Russia*
[4] *Tomsk State University, Tomsk, Russia*
[5] *Moscow Institute of Physics and Technology* (*State University*)*, Moscow, Russia*
** e-mail: efimov@bionet.nsc.ru*

Key words: time series, PCA, PCo, SSA, molecular sequences

Motivation and Aim: In the 40s of the last century, Karhunen and Loève proposed a method for processing a one-dimensional numerical time series by a multidimensional method of principal components. In the 1980s, Takens showed in fact that this method makes it possible to obtain an attractor and, accordingly, phase portraits of the dynamic system from observing only one variable of this system [1]. The method was independently arised and applied in practice, including by us for the analysis of the animals abundance dynamics [2, 3], and other [4]. The method can be extended for a sequence of any type elements, including numbers, symbols, figures, etc. and, as a special case, for molecular sequences. It is the point of this abstract.

Methods and Algorithms: Let there be a sequence $X = \{x_1, x_2, \dots, x_N\}$ of any type elements. Choose a lag L, $N > L > 1$. Denote by X_i the fragment X of length L terminated by the element x_i, $X_i = (x_{i-L+1}, x_{i-L+2}, \dots x_{i-1}, x_i)$, $N \geq i \geq L$. Compute the matrix of Euclidean distances $D = (d_{ij} = d(X_i, X_j))$ between all fragments (this is always possible, for example, using the number of unmatched elements, but not only). Apply the method of principal coordinates to the D and obtain the principal components of it [5]. Call this method PCA-Seq.

Results: The amino acid sequence of the *Homo sapiens Cytb* gene (AFJ22730.1, GenBank) was processed by PCA-Seq with parameters $N = 380$, $L = 8$. The root of the *p*-distance is used as the Euclidean distance. The first component (18.2 % of the common variance) clearly reflects the content of Leucine in each fragment and manifest the evident cyclicity, which is most likely determined by the secondary structure of the *Cytb* protein. Jacobi 4 package was used for calculations [6].

Conclusion: PCA-Seq is promising for processing molecular sequences, but not only.

Acknowledgements: Supported by budget project (No. 0324-2018-0017).

References

1. Takens F. (1981). Detecting strange attractors in turbulence. In Dynamical systems and turbulence, Warwick 1980 (pp. 366-381). Springer, Berlin, Heidelberg.
2. Efimov V.M., Galaktionov Y.K. (1983) On the possibility of predicting cyclic changes in the abundance of mammals. Zh. Obshch. Biol. (3):343-352. (in Russian)
3. Efimov V.M., Galaktionov Y.K., Shushpanova N.F. (1988). Analysis and prediction of time series by the principal component method. Novosibirsk: Nauka. 70p. (in Russian)
4. Golyandina N., Nekrutkin V., Zhigljavsky A.A. (2001) Analysis of time series structure: SSA and related techniques. Chapman and Hall/CRC.
5. Gower J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53(3/4):325-338.
6. Polunin D.A., Shtaiger I.A., Efimov V.M. (2014) Development of software system JACOBI 4 for multivariate analysis of microarray data, Vestnik NSU. Information Technology. 12(2):90-98. (in Russian)