CLUSTER VISUALIZATION AND NONLINEAR PROJECTION TECHNIQUES FOR BIOLOGICAL SEQUENCES

C. Ferles[*], *A. Stafylopatis*[†]

Abstract: The present study devises two techniques for visualizing biological sequence data clusterings. The Sequence Data Density Display (SDDD) and Sequence Likelihood Projection (SLP) visualizations represent the input symbolical sequences in a lower-dimensional space in such a way that the clusters and relations of data elements are preserved as faithfully as possible. The resulting unified framework incorporates directly raw symbolical sequence data (without necessitating any preprocessing stage), and moreover, operates on a pure unsupervised basis under complete absence of prior information and domain knowledge.

Key words: *non-numerical/symbolical sequence data, DNA, RNA, protein, clustering, visualization, mapping, nonlinear projection, Self-Organizing Map (SOM), Hidden Markov Model (HMM)*

Received: May 27, 2014
Revised and accepted: May 2, 2016

DOI: 10.14311/NNW.2016.26.016

1. Introduction

Deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and chain molecules are characteristic paradigms of sequence data which are intrinsically encoded in either the four-letter alphabet of nucleotides or the twenty-letter alphabet of amino acids. As a result, machine learning approaches that are in position to incorporate and process such symbolical sequence data have proven effective for modeling, processing and analyzing biological molecules. On the contrary, machine learning techniques that are able to process numerical data can be employed only after interposing a preprocessing stage in order to transform symbolical sequences to numerical data spaces. Nevertheless, additional computational complexity, and frequently, loss of information are inevitable aftereffects of such preprocessing transformations.

One common situation in early stages of bioinformatics projects is that the only available information comes in the form of symbolical sequence data (viz. DNA, RNA and protein sequences); any other kind of prior information or domain

[*] Christos Ferles – Corresponding author, Intelligent Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Zografou, Athens, Greece, E-mail: `christos.ferles@gmail.com`

[†] Andreas Stafylopatis, Intelligent Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Zografou, Athens, Greece