EXPLORATION OF REPRESENTATION OF POLYPEPTIDE CHAINS IN VECTORIAL MODELS FOR GENOMIC AND PROTEOMIC ANALYSIS

Ricardo Voyceik[1], José Miguel Ortega[2], Camilla Reginatto de Pierri[3], Letícia Graziela Costa Santos[3], Roberto Tadeu Raittz[3],

1 UFMG
2 Universidade Federal de Minas Gerais, Laboratório de Biodados
3 UFPR

Abstract

The volume of information in genes and proteins databases continues to increase exponentially, so vast amounts of biological information are available on public databases. Whilst some extent and requires increasing computational capacities to analysis performance, the ability to analyze this information has not been developed in the same way. Even with the today data mining techniques, useful for the treatment of large amounts of information, are limited in the exploration of biological data based on sequences, because they are unstructured and redundant data. Therefore, in this context, alignment free analysis of sequences was shown as a promising technique for analysis of proteomes and genes. With the aim of producing a mathematical model of representation of knowledge bases of genes or proteins, which can be manipulated by mathematical properties, in what it is possible to calculate the similarity between the amino acid sequences quickly, allowing the clustering by the biological characteristics given by the amino acids contained in the sequences, we propose an approach for gene and proteomic representation in a structured and reversible vector model, which has the potential to improve the capacity of analysis and data mining of biological data derived from protein sequences. Farther, we make available a graphic model to visualize this form of representation. The proposed model consists on the decomposition of sequences in sliding-windows analysis, in order to generate vectors that represent each amino acid at the sequences. As results, in the analyzes of the tests performed with the comparative among the techniques available in sequence analysis, in the aspects of clustering and similarity measures, the proposed method showed to have equivalent sensitivity, with the advantage of providing a substantially superior computational performance. We also show that the linearization of the matrices is a vector model that allows algebraic operations between the represented units, giving feasibility to operations such as geometric averages, cosine distance, centroid definitions, principal component analysis and the reduction by projection of the complete vector model for orthonormal basis. The reduction by projections to orthonormal basis down to 211 attributes, maintained the sensitivity of the model above 99%. As the vector model is storage in floating-point numbers, the data can be manipulated by GPUs with fast and precise parallel processing and possibility of clustering as well as machine learning techniques can be approached directly in the proposed model, without having to extract his characteristics or conversions, which opens new horizons in the data mining in Bioinformatics using math analysis tools.

Funding: CAPES