SEQUENCE COMPARISON USING $N$-DIMENSIONAL SIMILARITY SPACE

Ramamurthy Jayalakshmi$^1$, Ramanathan Natarajan$^2$, Munusamy Vivekanandan$^3$, Ganapathy S Natarajan$^4$

$^1$ Department of Biotechnology, School of Life Sciences, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India 620 024
$^2$ Centre for Mathematical Sciences, Arunapuram, Pala, Kerala, India 686 574 & Department of Chemical Engineering, Lakehead University, Thunder Bay, Ontario, Canada P7B 5E1
$^3$ Vivekananda Arts and Science College for Women, Thiruchengode, Namakal District, Tamil Nadu 637 205
$^4$ Department of Industrial Engineering, Texas Tech University, Lubbock, TX 79409, USA

rnataraj@lakeheadu.ca

Several alignment-free sequence comparison methods are available and they use similarity based on a particular numerical descriptor of biological sequences. Any loss of information incurred in the transformation of a sequence into a numerical descriptor affects the results. A pool of descriptors that use different algorithms in their computation is expected to suffer minimum loss of information and an attempt is made in this direction to study the similarity of DNA sequences that are homogenous and heterogeneous. Several numerical descriptors for the characterization of DNA sequences based on information theoretic approach, connectivity of vertex weighted line-graphs and those derived from the matrices obtained from the graphs constructed by depicting DNA sequences as a random walk on a Euclidean plane. The information theoretic descriptors were obtained based on the $L$-tuple approach for the combination of different numbers of bases. The connectivity type descriptors were calculated by converting the DNA sequence into a line-graph in which vertices (bases) were assigned weights based on the $pK_a$ of the bases. The graphical representations were converted into numerical descriptors by constructing matrices. Computer programs were developed to calculate seventy DNA descriptors. Coding region sequences (CDS) of nine different genes were used for the analysis. 560 nucleotide sequences were retrieved for the study. After initial data analysis to eliminate almost perfectly correlated descriptors orthogonal descriptors were obtained by performing principal component analysis and they were used to construct an $N$-dimensional similarity space wherein the 560 sequences were clustered by $k$-means cluster algorithm. Five principal components (orthogonal descriptors) were extracted and found to explain 92% of data variance. The five-dimensional similarity space thus constructed grouped similar sequences in the same cluster.

Paper # 154