NS.SCI.38

A COMPUTER BASED STATISTICAL TOOL TO ANALYZE THE CORRELATION AMONG DNA SEQUENCES

P. G. S. S. Jayarathna[1], R. D. Yapa[1], S. D. S. S. Sooriyapathirana[2]

[1]Department of Statistics and Computer Science,
[2]Department of Molecular Biology and Biotechnology,
Faculty of Science, University of Peradeniya

DNA is the molecule of life. DNA sequence analysis is the key for understanding many biological questions. In bioinformatics, statistical techniques such as frequency distribution techniques, alignment algorithms, hypothesis testing, and clustering techniques are used to analyze the correlation among DNA sequences. Furthermore, comparing lengths, GC-content, AT/GC ratio, repetition of small sub-sequences and the analysis about restriction sites are the most basic analysis on the DNA sequences. Pie charts and the frequency tables can be used to analyze nucleotide distribution among DNA sequences. In DNA sequence analysis, sequence alignment is one of the most important steps to identify the similarity regions between DNA sequences, because it reflects functional, structural, or evolutionary relationships among them. Since the process of alignment algorithms like Smith-Waterman's are very time consuming, the BLAST algorithm can be used as a time efficient procedure because it addresses the fundamental problems and the algorithm emphasizes speed over sensitivity. Cluster Analysis is also associated widely in DNA sequence analysis. The DNA analysis by using different statistical techniques requires several statistical tools and demands considerable expertise in statistics.

Therefore, an attempt was made to design a user friendly computer based statistical tool to analyze one or more DNA sequences in different paths of statistics and make a sequence alignment efficiently.The DNA Sequence Analysis Tool (DSAT) was developed and implemented. by using vb.net programming language in Microsoft Visual Studio 8. *MSChart* and *MSChartVisualStudioAddOn* tools were used to display graphic outputs of the tool. An analysis can be conducted under five options named as Nucleotide Distribution Analysis, Basic Analysis (GC content, AT/GC ratio and repetitions), Multiple Analysis, Pair wise Analysis and Cluster Analysis.

The DSAT contains a collection of several statistical techniques in one application and quick in aligning DNA sequences. This statistical tool can be used by biologists and students with limited statistical knowledge in quick time to get more detailed information about the correlation among DNA sequences.