Data Mining Strategy for "Gene Prediction" with Special Reference to Cotton Genome

KSHIRSAGAR Manali[1], BALASUBRAMANI G[2], SINGH Col Gurmit[3]

(1. *Yeshwantrao Chavan College of Engineering, Hingna Road, Wanadongri, Nagpur, Maharashtra* 441110, *India*; 2. *Central Institute of Cotton Research, Wardha Road, Nagpur, Maharashtra, India*; 3. *Prof. and Head with the Department of IT and Electronics, AAIDU, Allahabad, U. P., India*)

This paper presents an integrated approach towards solving the problem of "Gene Prediction". The "Gene Prediction" problem solving undergoes well defined stages starting with a DNA sequence as input and lab treatment and computational analysis go hands in hands throughout the process. Many bioinformatics tools are available for analysis at different stages of "Gene Prediction", but a simplified and integrated approach is needed to support and speed up the task of a life scientist. A data mining strategy has been proposed in this paper to explore the comparatively less expressed cotton genome. The work is being carried out in CICR, Nagpur, and comparative genomics is being used to predict cotton genes. This strategy reveals the fundamentals and mathematics of the entire process based on which a complete software can be developed which can help in automating the process of "Gene Prediction". This strategy involves 9 steps towards exploring cotton genome and also presents a well-defined model for the same. Using this approach even the research assistants, research fellows and others working at a lower level in the research labs can easily complete the computational tasks involved in the process of "Gene Prediction" and this can be of great help to the principal investigators.

Key words: data mining; codon; DNA; genome; mRNA; protein; splicing; transcription; translation