Novel Gene Identification through Family-based Classification of EST Sequences

X. Guan, Jurgen Lehmann[2], Ray Grimaila, Eric Lai[1], and Lei Du

Departments of Bioinformatics and [1]Molecular Genetics, Genetics Directorate,
Department of [2]Endocrinology, Glaxo Wellcome Research and Development.
5 Moore Drive, Research Triangle Park, NC 27709, USA

Abstract

EST databases are valuable sources for novel gene discovery. Due to the high redundancy and relatively low quality in most collections of EST sequence databases, searching for novel genes by simple homology comparison becomes very laborious and inefficient. We have developed a new approach in which we automatically scan through public EST databases and map homologous ESTs to a protein family such that known genes and novel genes of the protein family can be separated. Multiple evidences from DNA and protein homology comparisons are presented to the end user such that novel genes can be recognized with high confidence. The system consists of an automated database mining script, an interactive Java interface, and an Oracle database allowing users to periodically examine EST databases and identify novel genes while filtering out all ESTs that have been processed before.