EXTRACTION OF LARGE QUANTITY OF GENE START REGIONS AND START CODON REGIONS FROM HUMAN GENOME - FIE TOOL

Alen Chong and Guanglan Zhang

BioDiscovery Group, Laboratories for Information technology, Singapore

Promoters are stretches of DNA sequences, upstream of or overlapping the transcription start site (TSS) of genes, that control the expression of genes. FIE (5'-end Information Extraction: freely available at http://sdmc.krdl.org.sg/FIE) is a web-based program designed primarily to extract the sequence of the promoter region and the region around the translation initiation site (TIS) for a particular gene. First, FIE analyzes the alignment of currently available mRNA sequences (including Reference Sequence [RefSeq]; http://www.ncbi.nlm.nih.gov/LocusLink/RSfaq.html) for a gene against the relevant genomic contig. The alignment of these sequences is obtained from NCBI's LocusLink (http://www.ncbi.nlm.gov/LocusLink). Second, FIE extracts a user-specified length of DNA sequence around the promoter or TIS region for that gene, based on the information gained from the alignment analysis and the working draft sequence segments of the human genome. The importance of these sequence extractions performed by FIE lie in its usefulness for follow-up experiments in the lab and in silico in current research efforts to understand the regulatory mechanism of gene expression. The sequences extracted by FIE were also recently used to compile a dataset for training and testing our gene finding prediction system, Dragon Promoter Finder, Dragon ATG Finder and Dragon Gene Start Finder (http://sdmc.krdl.org.sg:8080/promoter). To date, we are not aware of any other promoter or TIS sequence extraction program that utilizes the sequence information of the human genome working draft sequence segments in this fashion. FIE was tested with a list of 210 genes from human chromosome 22 and found to be capable of identifying and extracting the relevant information for these genes where information about them were available.