BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 9(4)

CEPiNS: Conserved Exon Prediction in Novel Species

Shihab Hasan[1, 2*] & Christopher W Wheat[2, 3, 4]

[1]Bioinformatics, Department of Information Technology, 20014, University of Turku, Finland; [2]Department of Biological and Environmental Sciences, PL 65, Viikinkaari 1, 00014 University of Helsinki, Finland; [3]Centre for Ecology and Conservation, School of Biosciences, University of Exeter, Cornwall Campus, Penryn, Cornwall TR10 9EZ, United Kingdom; [4]Department of Zoology, Stockholm University, SE-106 91 Stockholm, Sweden; Shihab Hasan - Email: shihab.hasan@utu.fi; *Corresponding author

Received January 17, 2013; Accepted January 29, 2013; Published February 21, 2013

Abstract:
Exon structure is relatively well conserved among orthologs in several large clades of species (e.g. Mammalia, Diptera, Lepidoptera) across evolutionary distances of up to 80 million years. Thus, it should be straightforward to predict the exon structures in novel species based upon the known exon structures of species that have had their genomes sequenced and well assembled. Being able to predict the exon boundaries in the genes of novel species is important given the quickly growing numbers of transcriptome sequencing projects. CEPiNS is a new pipeline for mining exon boundaries of predicted gene sets from model species and then using this information to identify the exon boundaries in a novel species through codon based alignment. The pipeline uses the freeware SPIDEY, an exon boundary prediction tool, and BLAST (BLASTN, BLASTP, TBLASTX), both of which are part of NCBI's toolkit. CEPiNS provides an important tool to analyze the transcriptome of novel species.

Availability: http://www.cepins.org

Keywords: Exon prediction, Gene structure, Model species, Novel species, Transcriptomics, Evolutionary and Comparative genomics, Bioinformatics Software.

Background:
Genes contain exons which are regions coding for proteins and introns which are non-coding regions. In transcription process, introns are removed by RNA splicing and the exons are joined together to form the functional messenger RNA (mRNA) **[1]**. Accurate prediction of precise exon–intron boundaries in genes is an essential step in the analysis of genomic sequences **[2]**. This gene structure is conserved between closely related species for the majority of genes **[3]**. In evolution, gene structure conservation may be a record of core events **[4]**. The aim of this project is to develop a new pipeline to predict exon sequences and their boundaries for novel species comparing a model species by using the sequence similarity method.

Methodology:
CEPiNS is a bioinformatics tool for large-scale exon prediction. This application allows study of gene structure for model and novel species by predicting exon boundaries and sequences.

Given the input of a set of gene sequences and their genomic sequences for a model species, CEPiNS generates a table of exon boundaries for these genes. CEPiNS uses BLAST **[5]** to identify the orthologous genes between the genes of a reference species and the predicted genes from a novel species' assembled transcriptome. Once this orthology has been established, the exons in the genomic reference species can be transferred to the novel species. The output is therefore the predicted exons in the

novel species. The workflow for CEPiNS is illustrated in (**Figure 1).**

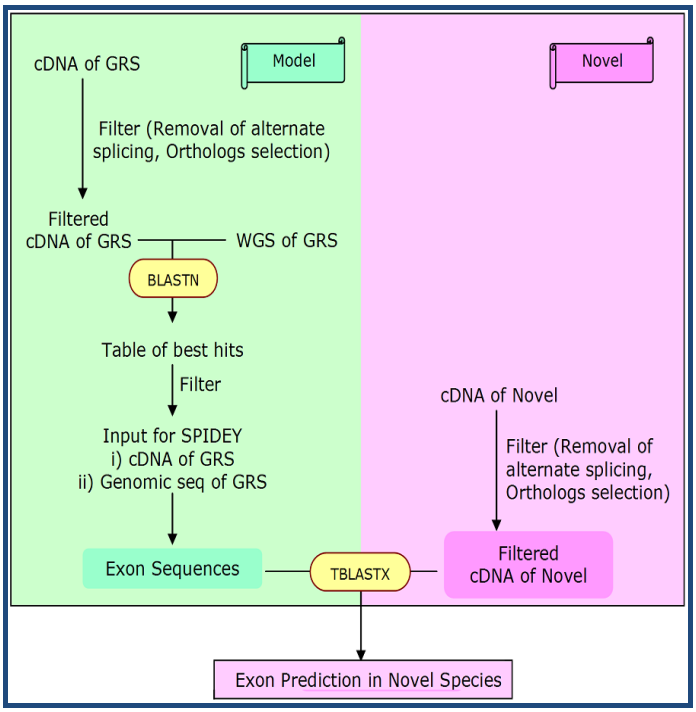

Figure 1: Workflow in CEPiNS. CEPiNS uses Exon Table obtained from SPIDEY to select the exon boundaries in cDNA of GRS. Then the Exon Sequences of GRS are used to find exon boundaries in Novel species. **GRS=Genomic Reference Species; *WGS=Whole Genome Sequence.*

Preprocessing of Dataset

CEPiNS has preprocessing tool to remove alternate splicing and to predict orthologous sequences by using BLAST. For both the model and novel species, multiple copies of the same genes within a genome are removed by identifying sequences with at least 95% similarity at the nucleotide level using BLASTN and retaining only the longest. Gene sequences in both model and novel species with at least 60% similarity at protein level CEPiNS treated as orthologous sequences using BLASTP.

Exon Prediction in Model Species

SPIDEY is an mRNA to genomic DNA alignment program **[6]**. When intron-exon boundaries are not already annotated in the reference (model) species, SPIDEY gives the exon boundaries by using a set of mRNA or cDNA sequences and their corresponding genomic sequences. CEPiNS generates a table with the gene ID, genomic sequence ID, exon boundaries in genes and genomic sequences and length of exons by using the table created by SPIDEY. It also creates a file of exon sequences in fasta format.

Exon Prediction in Novel Species

CEPiNS uses TBLASTX for the alignment at amino acid level in all six reading frames for each predicted genes from transcriptome assembly of the novel species and its corresponding exon sequences of the reference species, which has been created by SPIDEY. CEPiNS creates a table output with cDNA ID of Reference species, cDNA ID of novel species, exon Number, genomic coordinates, mRNA coordinates, length and percent identity. It also creates exon sequences fasta file for novel species.

Software Input and output:

CEPiNS requires the input of a set of transcribed genes of a model species, the genomic sequence of the same species and of a predicted gene set from transcriptome assembly of a novel species in fasta file format. The log screen keeps tracking the steps performed and results can be viewed by clicking corresponding buttons. The final outputs are the predicted exons boundaries in a text file and exons sequences in a fasta file of the novel species.

Conclusion:

CEPiNS is a package for predicting large scale exons for novel species with Graphical User Interface (GUI) so that biologists, ecologists, geneticists and people from other backgrounds can use it very easy way. The output data offer several opportunities for further work & other tool development.

Reference:

[1] Zhang MQ, *Hum Mol Genet*. 1998 **7**: 919 [PMID: 9536098]
[2] Shu JH *et al*. *Nucleic Acids Res*. 2006 **1**: 34 [PMID: 16845010]
[3] Frazer KA *et al*. *Genome Res*. 2003 **13**: 1 [PMID: 12529301]
[4] Betts MJ *et al*. *EMBO J*. 2001 **20**: 5354 [PMID: 11574467]
[5] Altschul SF *et al*. *J Mol Biol*. 1990 **215**: 403 [PMID: 2231712]
[6] Wheelan SJ *et al*. *Genome Res*. 2001 **11**: 1952 [PMID: 11691860]

Edited by P Kangueane
Citation: Hasan & Wheat, Bioinformation 9(4): 210-211 (2013)