EGENE 2 – A PLATFORM FOR AUTOMATED SEQUENCE PROCESSING AND ANNOTATION

Ferro, M.[1], Yamamoto, R.[2], Durham, A.M.[2], Gruber, A.[1]

[1]Department of Parasitology, Institute of Biomedical Sciences; [2]Institute of Mathematics and Statistics, University of São Paulo, Brazil

EGene is a flexible, modular and open-source system developed by our group (Durham *et al*., *Bioinformatics* 21: 2182-2183, 2005) to construct pipelines for sequence processing. The system is highly generic, and has been applied in several different large-scale DNA sequencing projects. We report here the development of EGene 2, an integrated platform for automated sequence annotation. We developed in total 26 new components, comprising an ORF finder and translator, modules for seven gene prediction programs (ESTScan, GENSCAN, GlimmerM, GlimmerHMM, Phat, SNAP and Twinscan), tandem repeats finders (TRF, String and mreps), tRNA gene prediction (tRNAscan-SE), cDNAs mapping onto genomic sequences (SIM4 and Exonerate), similarity searching (BLAST), protein motif finding (HMMER/Pfam, RPS-BLAST and InterProScan), transmembrane domain and signal peptide finding (SignalP, TMHMM and Phobius), and GO (Gene Ontology) term mapping. EGene2 generates feature annotation exchange files in several formats, including Feature Table (as defined by DDBJ/EMBL/NCBI) and GFF3. These output files can be used to perform manual inspection and curation, or to submit annotated sequences to public databases. EGene2 has been used and validated with a set of more than 45,000 ORESTES cDNA reads of *Eimeria* spp., a protozoan parasite. Also, we performed a full annotation of more than 14,000 cDNA clusters. The automated process is rapid, comprehensive and generates a reliable annotation. CoEd, a Java-based visual tool to facilitate pipeline construction and documentation is also provided in the package. Given the generic character of the platform, EGene2 can be used in any large-scale DNA sequencing and annotation project.

Financial support: FAPESP and CNPq
Keywords: automated annotation, pipeline construction, sequence analysis