

ConiferEST: an integrated bioinformatics system of data processing, integration and mining for Expressed Sequence Tags (ESTs) in conifers

Chun Liang, Gang Wang and Lin Liu

The Bioinfo Lab, Department of Botany, Miami University, Oxford 45056, Ohio

Although relatively small in terms of species numbers, conifers are dominant plants over a huge area of land. As an ancient group of plant with a fossil record that extends back at least 300 million years, conifers are clearly of immense ecological, economical and scientific importance. Among them, loblolly pine (*Pinus taeda*) alone provides ca 16% of the world's annual timber supply. Conifer genomes are uniformly large with highly repetitive sequences, presenting a serious impediment for complete genome sequencing. As an alternative, however, Expressed Sequence Tags (ESTs) remain a dominant approach for accessing the protein-coding portion of the genome and are widely used for gene discovery, gene expression profiling, marker development and genome annotation.

ConiferEST is aimed to be a freely available, web-based and integrated bioinformatics resource for data processing, integration and mining for ever-expanding conifer EST projects. (1) **WebTraceMiner**. Through easy-to-use web interfaces, users can upload trace files into our server, and sequence cleaning and quality trimming can be done automatically because we adopt NCBI UniVec and other databases for vector and contamination screening. Upon the reception of clone linkage information for 3' and 5' EST reads from users, putative full-length cDNAs will be detected. All EST features, including length and position of vector fragments, length and location of good quality regions, length, position and identity of poly(A)/(T) segments, and type and fidelity of poly(A) signals can be easily mined. All processed data can be saved locally in FASTA, XML or tab-delimited formats, and are ready for GenBank submission. (2) **BlastMiner**. Different from most of other online blast interfaces, BlastMiner provides users the capability to mine their blast data due to data warehousing schema in database design. For example, when the user selects two different parameters (i.e., PAM 30 and PAM250 for close or loose evolutionary relations) for the same query sequences against the same target database, BlastMiner is capable of presenting the common and different blast outputs between the two parameter settings. All blast results can be filtered using a user-adjustable E-value threshold as well as other filters, and they are searchable using different criteria. Each meaningful blast hit can be automatically linked back to original sources such as NCBI *Entrez* and *UniProt* to facilitate data mining for biologists. We are now in a stage of developing a web tool for EST clustering, which allows users to choose from different algorithms (i.e., d2-cluster, TGICL, NCBI Unigene and our own algorithm) for clustering. Using this tool, biologists can visualize and compare the different EST clusters from different algorithms and explore all important features, such as SSR, SNP, polyadenylation, splice variants and so on. For registered users, data with more advanced in-depth analyses will be available using our server through special requests. ConiferEST can also be a freely available and generic bioinformatics system for processing, integrating and mining ESTs for a wide variety of species other than conifers.