Computational Identification of Orthologs in Soybean and Arabidopsis Using Sequence and Expression Data

Wengang Zhou, Eve Wurtele, Ling Li, Julie Dickerson

We propose that the use of confirmatory expression data can strengthen a hypothesis that two probes are orthologous and help eliminate false positive BLAST hits based solely on sequence similarity. This is critical for soybean, an economically important species for which most genes have not yet been characterized. First, reciprocal BLAST hits are used to find up to five possible orthologous genes in Arabidopsis for each soybean microarray probe set. This analysis led to a total number of 58,702 ortholog pairs. Second, comparable Arabidopsis and soybean microarray data sets characterizing developmental stages at five time points during seed development with three repetitions were clustered using K-means clustering (K=200). The co-occurrence of ortholog pairs within clusters across species was computed. This analysis was repeated using microarray data from 5 different soybean lines. The results were similar across soybean lines and gave a set of 1200 ortholog pairs. In the final step, the soybean data was compared to a superset of Arabidopsis microarray data. The data from each species was clustered separately, and the soybean clusters were matched to the Arabidopsis clusters by selecting the cluster pairs with the highest number of orthologs. For the Arabidopsis genes in the same cluster, their putative orthologs were often in the same cluster in soybean. Also, most of these ortholog pairs overlapped the pairs from step 2. Thus, two independent lines of evidence suggest that these putative orthologs may be functionally-related: sequence similarity and commonality of expression pattern.