cSAGE and the Serial Analysis of Gene Expression (SAGE) in Arabidopsis thaliana

Chris Lewis Steve Robinson Tony Kusalik Isobel Parkin

April 30, 2003

The Serial Analysis of Gene Expression (SAGE) is based on the ability of a short sequence to uniquely identify a single gene in an organism. The protocol employs a restriction enzyme (i.e. NlaIII) to cleave mRNA at the 3'-end and produce short fragments (SAGE tags) that uniquely identify the genes from which they were produced. Pairs of tags are ligated to form ditags, which are then concatenated to form ditag chains. A Polymerase Chain Reaction (PCR) phase is applied to amplify the ditags prior to sequencing.

A sequence read contains 400-600 bases comprised of 16-25 ditags seperated by the restiction enzyme's recognition sequence (i.e. 'CATG' for NlaIII). The SAGE software extracts valid ditags from the sequence read—valid ditags have a defined length (24-26 bases for NlaIII) and are non-duplicates—and extracts tags from the ditags. The SAGE protocol calls for the rejection of duplicate ditags—ditags formed from the same pair of tags—to reduce the likelihood of bias introduced by preferential amplification during PCR. Additional error may be introduced by sequencing error or mistaken PCR replication, the cloning of tags containing artifacts such as linker sequence, and the formation of ditags containing vector sequence.

cSAGE is an application written in C to extract SAGE tags and to assign matches to genomic sequence such as a collection of ESTs or the annotated genes of a model organism. cSAGE is intended for use as a small component in a larger analysis pipeline; for instance a PERL script is used to compare two cSAGE reports and display tags with significance changes in expression. A modular design has been applied to facilitate extention of the software for new applications.

Tags are extracted from the sequence reads and genomic sequences using a near-linear state machine. Sequence tags are stored in an 5-ary tree with nodes representing the bases A,C,G,T,N, which enables the rapid detection of duplicate ditags and allows efficient tag-to-gene matching. Known vector and contaminant tags can be excluded from analysis by placing them in an exclude file. Sequence reads may be provided in either fasta format of PHD format (output from phred), and genomic sequence is expected in fasta format.

cSAGE has been applied to a cold tolerance experiment in Arabidopsis thaliana. Highlights from this experiment include: 184,580 sequence tags, of which 146,178 had an average quality greater than 20. After removing polyA and linker tags a final set of 145,170 tags was analysed. The final pool contained 29,663 (20.6%) unique tags, of which 16,664 (11.6%) were singletons. When assigning matches to the tags, 46% of the unique tags were found to match the canonical (3' most) recognition site of the genes, and 43% of the unique tags matched a non-canonical recognition site; 89% of tags matched a gene. The non-canonical matches indicated potential alternate splicing for the genes, or possibly anti-sense transcription; a small number of tags have been confirmed as alternate transcription tags.

More information on cSAGE can be found at http://homepage.usask.ca/ ctl271/csage.