Automated Annotation of NGS Transcriptome Data using ISGA and Ergatis

Aaron Buechlein, Jeong-Hyeon Choi_, Karthik Muthuraman, Chris Hemmerich,
Center for Genomics and Bioinformatics, Indiana University (chemmeri@indiana.edu)

Project: http://isga.cgb.indiana.edu
Download: https://cgb.indiana.edu/downloads/
License: Apache Version 2.0

With next-generation sequencing technologies, a single 454 run can generate roughly 400,000,000 base pairs of transcriptome reads and a Solexa run can generate 250 times as many bases for use against a reference genome. For high-throughput research, automated tools must be used for annotation and analysis to match the pace of sequencer output. Even for biologists working with a single sequencing run, automated annotation results can provide immediate insights while labor-intensive manual curation is performed. As an evolution of our previous work on EST analysis (http://estpiper.cgb.indiana.edu) and prokaryotic genome annotation (http://isga.cgb.indiana.edu), we have built a transcriptome annotation pipeline for suitable for use with NGS data.

In our pipeline, a transcriptome comprised of contigs and singletons is compared to public databases such as non-redundant protein and EST sequence databases using BlastX. Well matching contigs and singletons to those databases are further analyzed for taxonomic distribution, GO terms, metabolic pathways, orthology and paralogy, gene duplication, and alternative splicing variants. If a close, sequenced genome is available, comparison with the genome reveal a minimal gene set in the transcriptome. Unmatched transcriptomic sequences are fed to ORF prediction programs such as ORFpredictor to find possible protein coding frames. Since contigs and singletons are parts of a gene, they are clustered using splice junction reads and an orthology database. If a transcriptome is sequenced from different organs and individuals, the transcriptome can be used to identify sequence variants such as SNPs.

The Integrative Services for Genomic Analysis (ISGA) web application is a solid platform on which to develop this pipeline. ISGA was originally developed for prokaryotic genome annotation, and provides an intuitive interface for biologists to run and customize pipelines. ISGA has a simple account system to ensure that users data is private, and allows them to easily retrieve the results from previous experiments. In addition, ISGA provides a "Tool Box" for visualizing and further analyzing pipeline results using tools such as GBrowse and Blast.

ISGA uses Ergatis (http://ergatis.sourceforge.net/) to execute and manage the provided bioinformatics pipelines. Ergatis is a workflow management system for bioinformatics utilities used at the J. Craig Venter Institute, the Institute for Genome Sciences, and other institutions, and achieves high-throughput pipeline execution through automation and leveraging distributed computing resources. Ergatis provides the necessary tools for creating a complex bioinformatics pipeline, closely monitoring its execution, and efficiently recovering from errors.