QIAGEN

Application Note

Whole genome functional annotation of Solanum lycopersicum

Sample to Insight

The Blast2GO plugin for CLC Genomics Workbench allows for functional annotation of novel sequence data. Functional annotation is relevant to complete the functional characterization of de novo sequenced genomes and transcriptomes. This application note describes the complete functional annotation of the tomato genome, Solanum lycopersicum, using the Blast2GO plugin for CLC Genomics Workbench. We present the basic analysis workflow, describe several issues regarding quality and quantity of the obtained results, and comment on the genome wide analysis of the functional compositions of this dataset.

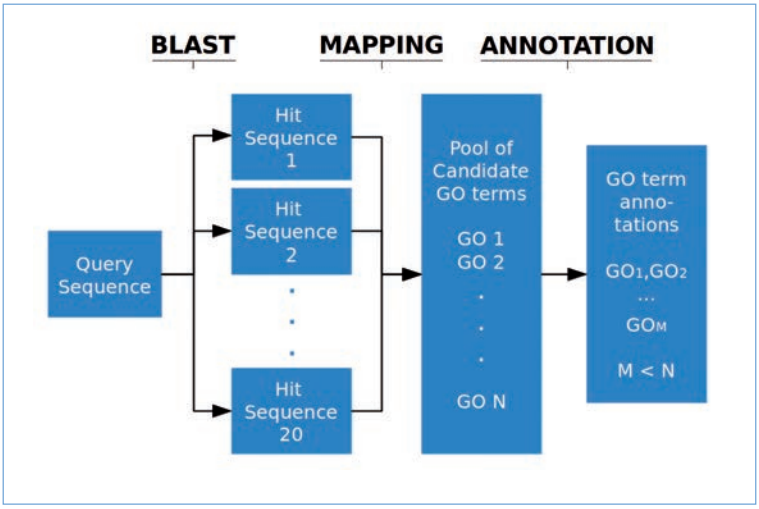

Figure 1: The Blast2GO annotation process performed for each query sequence.

Data

In the following analysis we use a whole genome dataset of Solanum lycopersicum (Taxa: 4081). The official CDS annotation for the genome are obtained from the SL2.40 genome build released by the International Tomato Annotation Group (Release 2.3, 2011-04-26). (ftp://ftp.solgenomics.net/tomato_genome/annotation/ITAG2.3_release/ITAG2.3_cds.fasta)

Software: Blast2GO PRO Plugin 1.1.0 for CLC Genomics Workbench 6.5.

Analysis Workflow

The input data is a text file in fasta format containing 34,727 CDS of the tomato genome.

Sequence alignment via BLAST

First, the tomato CDS sequences are used as queries in a blastx search launched from CLC Genomics Workbench. The BLAST search is run against the non-redundant (NR) database at NCBI with an e-value of 1xE-6 and we keep the top 20 alignments for each sequence. The resulting MultiBlast data collection is then converted into a Blast2GO project: Toolbox → Blast2GO → Manage Projects → Convert Data to Blast2GO Project.

Domain search via InterProScan

An InterPro domain search is performed directly on the FASTA input file. InterPro combines different protein signature recognition methods and the identified domains can be directly translated into Gene Ontology terms. Once the process is finished, we export the results in XML format and import them to the Blast2GO project generated. In this way, InterProScan and BLAST searches can be performed in parallel and then combined: Menu → Import FASTA → Manage Projects → Convert Data to Blast2GO Project → Blast2GO → InterProScan.

Gene Ontology mapping

Mapping is the process by which Blast2GO retrieves functional information for all BLAST Hits from the Gene Ontology (GO) database. The GO database contains several millions of functionally annotated gene products

for hundreds of different species. Blast2GO uses different public resources provided by the NCBI, PIR and GO to link the different protein IDs (names, symbols, GIs, UniProts, etc.) to these GO annotations. Moreover, annotations in the GO database contain an Evidence Code qualifier that provides information about the quality of this functional assignment, which are also retrieved by Blast2GO. The result of this step is a set of GO candidate annotation terms for each tomato query sequence: Toolbox → Blast2GO → Mapping → Mapping.

Functional Annotation

The annotation algorithm selects GO terms from the pool of candidate GOs obtained by the Mapping step and assign these to the query sequences. GO annotation is carried out by applying the Blast2GO annotation rule, which computes an annotation score for each candidate GO term. This score considers the similarity between hit and query sequences, the Evidence Code of each GO term and the existence of neighboring GO term candidates. We use an annotation threshold of 55 to select the GO term candidates for final assignment to the query sequence, leaving other parameters by default to achieve an optimal balance between quality (minimum of 55% of sequence similarity for hits with experimental evidence codes, higher for other types of evidence) and quantity (number of annotated sequences). The Blast2GO annotation rule is described in S. Götz et al., 2008: Toolbox → Blast2GO → Annotation → Annotation. Figure 1 represents schematically the Blast2GO annotation process.

GO-slim summary

GO-Slims are reduced versions of the Gene Ontology that contain a selected number of relevant functions. Different GO-Slims are available, via Blast2GO, for different types of organisms. To summarize the functional information of this tomato genome dataset we perform a plant specific GO reduction*: Toolbox → Blast2GO → GO-Slim → GO-Slim.

*Plant GO-Slim (http://www.geneontology.org/GO.slims.shtml)

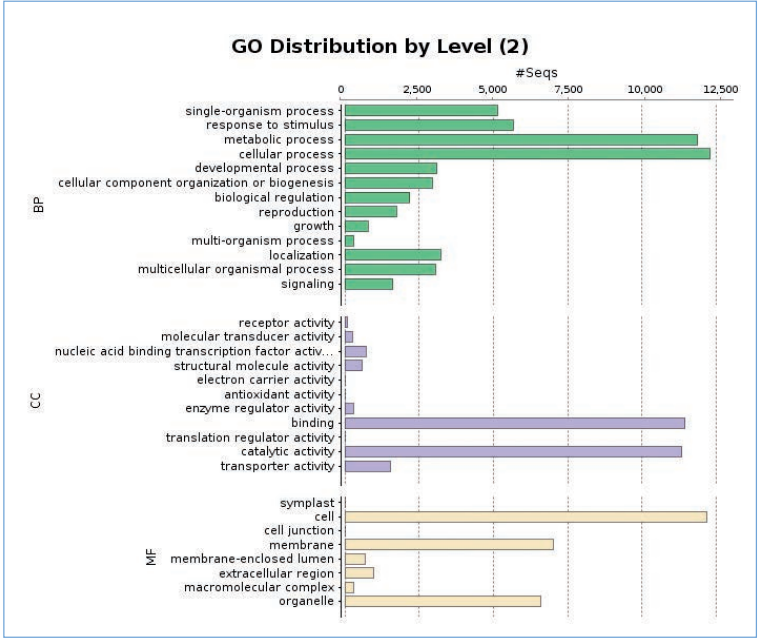

Figure 2. Number of sequences annotated to the different GO terms at hierarchy level 2 for all three GO categories.

Combined graph and pie chart

Once we reduce the functional diversity to an appropriate size we can obtain a Combined Graph. The Combined Graph provides a birds eye view of the GO terms annotations of the entire dataset within the structure of the Gene Ontology. Since the Gene Ontology is divided into three categories, we can generate separate graphs for Molecular Functions (MF), Cellular Component (CC), and Biological Process (BP), respectively. As Combined Graphs can still be large and difficult to navigate, a more accessible representation of the functional data is obtained by pie charts. Pie charts create a transversal cut of the GO graph at the selected

Ontology level. In this example, we choose to generate a pie chart at level 2 to get a broad overview of the functional distribution in this dataset: Toolbox → Blast2GO → Create Combined Graph; Toolbox → Blast2GO → Create Pie Chart (from the Combined Graph).

Summary statistics charts

The Blast2GO plugin contains many additional visualization tools to graphically summarize annotation results. Charts are directly generated from within the plugin: Toolbox → Blast2GO → Analysis → Statistics.

Some of the most useful ones are:

- Data Distribution, summarizes the annotation process.
- Species Distribution, indicates the species present in BLAST hits.
- Annotation bar charts, an alternative representation to pie charts (Figure 2).

Results

From a total of 34,727 CDS sequences, 11% do not obtain a BLAST hit, 7% of the sequences cannot be linked to Gene Ontology entries, and 8% of the sequences with GO mapping do not reach the quality for an annotation assignment. Additionally, we retrieve protein domain information for about 75% of the sequences, 70% of which have BLAST-based GO terms. Overall, we can assign functional labels to 74% of the input sequences, and around 7% of these assignments are merely domain based. Blast2GO can also assign enzyme codes to more than 20% of the sequences.

The sequence alignments (BLAST and InterProScan) are the most time consuming steps which requires normally several days using public resources (10 days in this case). This step can be accelerated by running these steps in parallel and by using local installations in a cluster or grid environment.

References:

- A. Conesa, S. Götz, J. M. Garcia-Gomez, J. Terol, M. Talon and M. Robles. "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research", Bioinformatics, Vol. 21, September, 2005, pp. 3674-3676.
- A. Conesa and S. Götz. "Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics", International Journal of Plant Genomics, Vol. 2008. 2008, pp. 1-13.
- S. Götz et al. "High-throughput functional annotation and data mining with the Blast2GO suite", Nucleic Acids Research, Vol. 36, June, 2008, pp. 3420-3435.

biobam BIOINFORMATICS SOLUTIONS b2g blast2go

The Blast2GO PRO Plugin is developed and maintained by BioBam Bioinformatics, a knowledge-based company dedicated to creating user-friendly software for the scientific community. BioBam is internationally recognized for its expertise in functional annotation and genome analysis. For more information about BioBam and Blast2GO please visit them online at www.blast2go.com and www.biobam.com.

For futher information, please refer to clcagro.com/science