BRAGOMAP – A new Perl script for high throughoutput Blast results analysis including GO and MapMan annotations.

Woycicki R.[1], Gutman W.[2], Przybecki Z[1].

[1]Dept of Plant Genetics, Breeding and Biotechnology, Faculty of Horticulture and Landscape Architecture, Warsaw University of Life Sciences SGGW, Warsaw, Poland, [2]Dept of Biochemistry, Faculty of Agronomy and Biology, Warsaw University of Life Sciences - SGGW,

Motivation

Analysis of sequences similarities is the first and most important method used to find out the function of unknown nucleotides. Searching of homologs should be done carefully not to loose any important ones. Having thousands of results from various long-read sequencing projects (ie. differentially expressed tags, genomic polymorphons or BAC ends), the by-hand ability to retrieve interesting (to our goal) similarities in hundreds of thousands of Blast results is practically not possible. Decreasing the number of retrieved sequences by giving more stringency in e-value threshold or displaying less results would lead to false deductions. Using the program it is also possible to automate gene product annotations.

Methods

To facilitate fast retrieval of interesting Blast homologies and making right deductions about the biological role of sequences, in big sequencing projects, the new Perl script BRAGOMAP was written. The program make use of some of BioPerl modules as well as the power of regex text-mining in the Perl itself (Table 1).

Results

The script gives us the possibility to retrive interesting BLAST similarities by using keywords to scann GenBank fields and giving scores for each one found. It collects all impotrant information from GenBank data and puts in separate columns of tab-delimited file for further use (Table 2). If we were interested in flower differentiation genes we could use the keywords (flower, ovule, anther, pollen, etc.) and/or filter all the homologies sequences isolated from flower tissues in a special development stage. We can also filter results by choosing similarities to interesting genes or protein products. This script retrieve also all standard information from the Blast and GenBank files as Description, ACC no., E-value, Similarity positions, Query Length, Percent of Similarity etc. Automatic annotations are done by looking for genes, protein products and /or DB references in the proper mappings files (Table 3).

Table 1 – Program parameters

```
What do You want to do [b]last or [f]iltering?: f

Filtering settings filename:? cucf2

First file number: 1
Last file number: 3883
File name: stcok_blastn_nt_030908
Most important taxon: cucurbitaceae
Less important taxon: magnoliophyta
Minor taxon: eukaryota
Important organella: mitochondrion
Minimal percent of simillarity: 60
Minimal length with 60 simillarity: 20
Number of HITs to analyze: 0

Choosen database is: nt

Keywords list filename:? cuck1

Your choosen keywords are: ACO, sex, flower, female, male, gonade, pollen,
anther, ACS, fruit, shoot apex, SRY, X chromosome, Y chromosome,
chromosome X, chromosome Y, ovary, ovule, carpel, pistil, primordium,
development, differentiation, embryo, stress,heatshock, ethylene, auxin,
hormone, steroid, seed, gibberellin, stamen, PCD, pollination, programmed
cell death, buds, drougth, ROS, monooxygenase, catalasa, apoptosis,
laddering, oxygene stress, chaperones, transport, transcription factor,
tapetum, archespor, ACC, microspore, macrospore, placenta, ethrel

Do You want to start [f]iltering or to change [s]ettings:? f

Is it an update(y/n):? n

Filtering: #1/1
ACC_NO is AM465301

Number of all unique found keywords: 0

 *** Number of all collected points *** 6

It was filtered: #1/1
***
Filtering: #2/1
```

Table 3 – Annotations chart

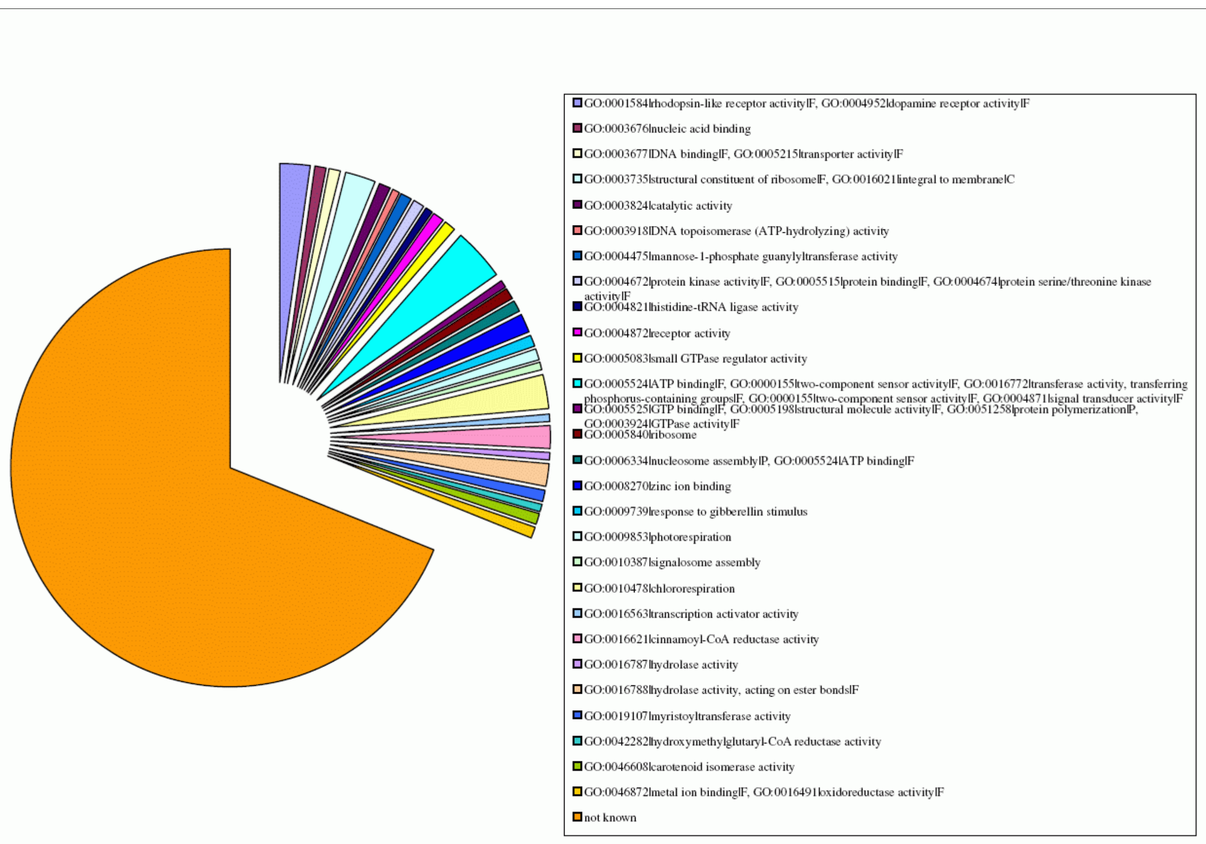

Table 2 – Collection of GenBank Data

Hit_no	Query_name	Species	Total_Score	Keys_P	Found_Keys	E-value_P	GO_AN_F	GO_AN_P	GO_AN_C	Desc_P	Org_P	Sim_P	Tax_P	Anno_P	Feat_P	Key_P	Sim_R	E-value	ACC_NO	Genes	Products	Proteins	Tissue_Type	Dev_Stage	DB_XREF
0	ZPCSPO11_C09	Cucumis sativus	12	1	transcription factor	3				0	0	2	6	0	1	0	85	0.0	AY221169		basic leucine zipper tr	AAO89566.1			taxon:3659, GI:29691164
5	ZPCSPO11_C09	Ipomoea nil	10	2	transcription factor, flower	3				0	0	2	3	1	1	1	77.976190	1,00E-151	U37437	PNIL 34	PNIL34	AAB19120.1		seedling	taxon:35883, GI:1052960
4	ZPCSPO11_C09	Populus trichocarpa	9	1	male	3				0	0	2	3	0	1	0	82.023809	1,00E-162	EF148495		unknown	ABK96463.1	Sapling trees one metre in he		taxon:3695, GI:118489317
2	ZPCSPO11_C09	Trifolium repens	8	0		3	GO:001656	GO:0009965	leaf morph	0	0	2	3	0	0	0	66.90476	1,00E-165	AM282585	zip	putative desaturase-lik	CAK54360.1	leaf		taxon:3899, GI:109450812,
10	ZPCSPO11_C09	Arabidopsis thalian	8	1	transcription factor	2	GO:000367	GO:000940	GO:000573	0	0	2	3	1	0	1	73.21428	1,00E-104	U38232	AT103	AT103	AAB18942.1		etiolated seed	taxon:3702, GI:1033195
11	ZPCSPO11_C09	Arabidopsis thalian	8	1	transport	2	GO:000367	GO:000940	GO:000573	0	0	2	3	0	1	0	73.09523	1,00E-103	NM_115553	AT103	AT103 (DICARBOXYL	NP_191253.1			taxon:3702, TAIR:AT3G569
7	ZPCSPO11_C09	Solanum lycopersic	8	1	fruit	2				0	0	2	3	0	1	0	70.47619	1,00E-121	BT013953				fruit		taxon:4081
8	ZPCSPO11_C09	Populus trichocarpa	8	1	female	2				0	0	2	3	0	1	0	8.333333	1,00E-109	EF148071		unknown	ABK96057.1	Young and mature leaves, alc		taxon:3694, GI:118488487
9	ZPCSPO11_C09	Zea mays Spermat	8	1	seed	2				0	0	2	3	0	1	0	70.59523	1,00E-105	AY108897						taxon:4577, MaizeGDB:634
1	ZPCSPO11_C09	Nicotiana tabacum	8	0		3				0	0	2	3	0	0	0	70.47619	0.0	AY221168		ZIP	AAO89565.2			taxon:4097, GI:33943103
3	ZPCSPO11_C09	Trifolium repens	8	0		3				0	0	2	3	0	0	0	66.90476	1,00E-163	AY322557		putative fatty acid des	AAP83877.1			taxon:3899, GI:32480923
6	ZPCSPO11_C09	Gossypium hirsutu	7	0		2	GO:001656	GO:0009965	leaf morph	0	0	2	3	0	0	0	78.80952	1,00E-141	AY456957	ZIP	putative leucine zipper	AAR20445.2			taxon:3635, GI:38503523
15	ZPCSPO11_C09	Arabidopsis thalian	7	0		2	GO:0003677	DNA bindi	GO:000953	0	0	2	3	0	0	0	73.09523	1,00E-101	AF236101	Crd1	putative dicarboxylate	AAF63476.1			taxon:3702, GI:7542486
17	ZPCSPO11_C09	Oryza sativa Japon	7	1	transcription factor	1				0	0	2	3	0	1	0	70.59523	5,00E-98	NM_001049	Os01g027	hypothetical protein	NP_001042745.1			taxon:39947, GeneID:43269
43	ZPCSPO11_C09	Euphorbia esula	7	1	buds	1				0	0	2	3	0	1	0	8.928571	8,00E-54	AF417577		leucine zipper-contain	AAL13304.1			taxon:3993, GI:16033631
12	ZPCSPO11_C09	Triticum aestivum S	7	0		2				0	0	2	3	0	0	0	73.09523	1,00E-103	AY322552		putative fatty acid des	AAP83872.1			taxon:4565, GI:32480913
13	ZPCSPO11_C09	Rosa davurica	7	0		2				0	0	2	3	0	0	0	75.83333	1,00E-103	AY322555		putative fatty acid des	AAP83875.1			taxon:237596, GI:32480919
14	ZPCSPO11_C09	Spinacia oleracea	7	0		2				0	0	2	3	0	0	0	75.83333	1,00E-101	AY322553		putative fatty acid des	AAP83873.1			taxon:3562, GI:32480915
18	ZPCSPO11_C09	Triticum aestivum S	6	0		1	GO:000538	GO:0010043	response t	0	0	2	3	0	0	0	70.59523	5,00E-98	AY738114	zip1	leucine zipper protein	AAW66004.1			taxon:4565, GI:58339283
26	ZPCSPO11_C09	Triticum aestivum S	6	0		1	GO:000538	GO:0010043	response t	0	0	2	3	0	0	0	70.59523	2,00E-88	AY914051	zip1	putative leucine zipper	AAX97504.1			taxon:4565, GI:62736388
64	ZPCSPO11_C09	Oryza sativa (japon	6	1	transcription factor	0				0	0	2	3	0	1	0	11.90476	1,00E-46	AP008207	Os01g0279100		BAF04659.1			taxon:39947, GI:113532276