frontiers

Article Title

Dennis Psaroudakis [1], Ha Vu [1], Colleen Yanarella [1], Steven Cannon [1], Darwin Campbell [1], Parnal Joshi [1], Iddo Friedberg [1,4], Kokulapalan Wimalanathan [1,2], Carolyn J. Lawrence-Dill [1,2,3*]

[1] Bioinformatics and Computational Biology, Iowa State University, Ames, IA, USA
[2] Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA, USA
[3] Department of Agronomy, Iowa State University, Ames, IA, USA
[4] Department of Veterinary Microbiology, Iowa State University, Ames, IA, USA

Correspondence*:
Carolyn J. Lawrence-Dill
triffid@iastate.edu

ABSTRACT

Abstract length and content varies depending on article type. Refer to http://www.frontiersin.org/about/AuthorGuidelines for abstract requirement and length according to article type.

Keywords: Text Text Text Text Text Text Text Text

1 INTRODUCTION

Hello, how are we doing?

2 METHODS

2.1 Clean up

All functional annotation sets were cleaned up the following way (using definitions from the Gene Ontology version 2019-07-01):

1. Any annotations where the GO accession was marked as obsolete were removed.
2. Some terms in the GO have ‘alternative ids’. When naively removing duplicates, two entries will not be recognized as duplicates if they have different accessions pointing to the same GO term. Therefore, all GO accessions were changed to their respecitve ‘main id’ and the dataset was again scanned for duplicates.

Table 1 provides information on the number of annotations that were removed this way from each dataset. All further analyses were performed on the cleaned datasets since we assume the user will only be interested in still valid and non-redundant functional annotations.

2.2 Choosing the right evaluation metric

A plethora of different metric to evaluate the quality of functional annotation predictions is available using different approaches and there seems to be no clear standard yet. TODO: THIS IS WEAK Additionally,

Table 1. Number of removed annotations during cleanup.

Genome	Dataset	Obsolete Annotations	Duplicates
Glycine max	GOMAP	203	0
Hordeum vulgarum	GOMAP	101	0
Medicago truncatula A17	GOMAP	0	0
Medicago truncatula R108	GOMAP	0	0
Oryza sativa	GOMAP	111	2
	GoldStandard	38	556
Phaseolus vulgaris	GOMAP	0	0
Triticum aestivum	GOMAP	285	0
Vigna unguiculata	GOMAP	0	0
Zea mays B73.v4	GOMAP	752	83
Zea mays Mo17	GOMAP	726	77
Zea mays PH207	GOMAP	798	76
Zea mays W22	GOMAP	754	82
Arachis hypogaea	GOMAP	0	0
Zea mays B73.v3	GOMAP	1107	70
	GoldStandard	1	0
	Gramene49	94	2
	Phytozome	54	0
Oryza sativa	Gramene61-IEA	10	14
	Gramene61-all	48	9565

each of the metrics has a different focus and lalala so choosing a metric for quality evaluation is not trivial. When we first published GOMAP (Wimalanathan et al., 2018), we used a modified version of the hierarchical evaluation metrics originally introduced in (Verspoor et al., 2006) because they were simple, clear, and part of an earlier attempt at unifying and standardizing GO annotation comparisons (Defoin-Platel et al., 2011). In the meantime, Plyusnin et al. (2018) have published an approach for evaluating different metrics showing substantial differences within the robustness of different approaches. `TODO DESCRIBE THEIR APPROACH` We have applied their method on the Gold Standards available to us to determine which evaluation metric is the most appropriate in our case. The results of this analysis can be seen in `TODO`.

We then evaluated our predictions and the other annotation sets using the best performing metrics as well as the one we previously used. `TODO`

3 RESULTS

... a quantitative comparison of the datasets in Table.

3.1 Quality Evaluation

`TODO` If it turns out that our predictions are good with hF but bad with more approriate metrics, explanation would be that score thresholds for the prediction tools used in the GOMAP pipeline have

Table 2. Quantitative metrics of the cleaned functional annotation sets. CC, BF, MP, and A refer to the aspects of the GO: Cellular Component, Biological Function, Molecular Process, and Any/All.

			Annotations[a]				Annotated Genes [%][b]				Median Ann. per G.[c]			
Genome	Genes	Dataset	CC	BF	MP	**A**	CC	BF	MP	**A**	CC	BF	MP	**A**
Arachis hypogaea	100	GOMAP	153433	132944	493799	**780176**	57667	56855	67123	**67124**	2	2	6	**10**
Glycine max		GOMAP	129215	113827	417555	**660597**	46020	47034	52871	**52872**	2	2	6	**11**
Hordeum vulgarum		GOMAP	88130	80282	272823	**441235**	35237	36470	39733	**39734**	2	2	5	**10**
Medicago truncatula A17		GOMAP	107362	99719	364065	**571146**	42325	43736	50443	**50444**	2	2	6	**10**
Medicago truncatula R108		GOMAP	112343	108031	382322	**602696**	40332	50220	55706	**55706**	1	2	5	**9**
Oryza sativa		GOMAP	72780	64685	248700	**386165**	28619	29853	35824	**35825**	2	2	6	**9**
		GoldStandard	7730	11060	19378	**38176**	5725	7383	9031	**11387**	1	1	1	**3**
		Gramene61-IEA	14633	32787	39105	**86529**	10771	15537	16705	**21446**	1	1	1	**3**
		Gramene61-all	20622	40674	54402	**115710**	13272	16962	18513	**22272**	1	1	2	**4**
Phaseolus vulgaris		GOMAP	72005	64583	229630	**366218**	25934	25539	27432	**27433**	2	2	6	**11**
Triticum aestivum		GOMAP	267741	218623	785960	**1272324**	95604	98187	107890	**107891**	2	2	6	**10**
Vigna unguiculata		GOMAP	75867	68313	243278	**387458**	27173	27124	29772	**29773**	2	2	6	**11**
Zea mays B73.v3		GOMAP	135211	87420	291251	**513882**	34866	38073	39468	**39469**	3	2	6	**11**
		GoldStandard	1565	65	299	**1929**	1548	60	151	**1634**	1	0	0	**1**
		Gramene49	20072	31056	30089	**81217**	11834	17991	15800	**21926**	1	1	1	**3**
		Phytozome	4787	19044	13100	**36931**	4524	13728	11365	**16132**	0	1	1	**2**
Zea mays B73.v4		GOMAP	88827	82251	278719	**449797**	36717	37337	39323	**39324**	2	2	6	**10**
Zea mays Mo17		GOMAP	87567	79214	277787	**444568**	33618	35105	38619	**38620**	2	2	6	**10**
Zea mays PH207		GOMAP	90617	85500	288677	**464794**	35170	36762	40556	**40557**	2	2	6	**10**
Zea mays W22		GOMAP	95390	85039	289780	**470209**	36987	37685	40689	**40690**	2	2	6	**10**

[a] How many annotations in the CC, BF, and MP aspect does this dataset contain? A = How many in total? A = CC + BF + MP

[b] How many genes in the genome have at least one GO term from the CC, BF, MP aspect annotated to them? A = How many at least one from any aspect? (A = CC ∪ BF ∪ MP)

[c] Take a typical gene that is present in the annotation set. How many annotations does it have in each aspect? A = How many in total? Ask your favorite statistician why A ≠ CC + BF + MP

Table 3. Quality evaluation of the used GO annotation sets.

Genome	Dataset	SimGIC2 score
Oryza sativa	GOMAP	0.241470
	GoldStandard	0.999971
	Gramene61-IEA	0.324831
	Gramene61-all	0.732289
Zea mays B73.v3	GOMAP	0.072971
	GoldStandard	0.998562
	Gramene49	0.126450
	Phytozome	0.032693

been chosen to maximize this hF value. It now seems reasonable to re-adjust these thresholds to maximize a different metric which will likely result in a drop in hF score but increase in other metrics. Again emphasizes the importance of choosing the right evaluation metric.

REFERENCES

Defoin-Platel, M., Hindle, M. M., Lysenko, A., Powers, S. J., Habash, D. Z., Rawlings, C. J., et al. (2011). AIGO: Towards a unified framework for the Analysis and the Inter-comparison of GO functional annotations. *BMC Bioinformatics* doi:10.1186/1471-2105-12-431

Plyusnin, I., Holm, L., and Töoröonen, P. (2018). Novel Comparison of Evaluation Metrics for Gene Ontology Classifiers Reveals Drastic Performance Differences. *bioRxiv* , 427096doi:10.1101/427096

Verspoor, K., Cohn, J., Mniszewski, S., and Joslyn, C. (2006). A categorization approach to automated ontological function annotation. *Protein Science* doi:10.1110/ps.062184006

Wimalanathan, K., Friedberg, I., Andorf, C. M., and Lawrence-Dill, C. J. (2018). Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER). *Plant Direct* 2, e00052. doi:10.1002/pld3.52