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 Generating Predictions

Used GOMAP on condo lalala. Input files are (usually) published along results.

2.2 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.
3. Any annotations with modifiers (NOT, contributes_to...) were removed since no tool used in the further analysis can handle them.

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.

2.3 Quantitative Evaluation

lalala lololo table xyz

2.4 Quality Evaluation

Quality evaluation of gene function predictions is not trivial and usually done by comparing the set of predicted functions of a gene against a *gold standard* consisting of annotations that are assumed to be correct. We used annotations that were created or in some way curated with human participation for gold standards. There are a plethora of different metrics to perform the comparison of predictions against this gold standard. 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 (Table `TODO`).

2.5 Phylogenetic Tree Construction

To demonstrate that a more top-level and holistic use of whole-genome functional predictions can still be useful we devised some simple ways of applying phylogenetic methods to our predictions. ### Distance Based ### Character Based

2.6 Ensuring Reproducibility

containerization, github...

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 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. Also shows how comparison between different pipelines/predictions can be difficult if chose different metric or optimized for different metric. Also: if an annotation is not present in the gold standard, there is no way of knowing whether that gene truly doesn't have that function or whether it has just never been characterized/examined. So we cannot distinguish between a biologically true negative and an actually false negative in the gold standard. This

Table 1. Number of removed annotations during cleanup.

Genome	Dataset	Obsolete Annotations	Duplicates	Annotations with Modifiers
Arachis hypogaea	GOMAP	0	0	912
Brachypodium distachyon	GOMAP	696	43	789
Glycine max	GOMAP	203	0	930
Gossypium raimondii	GOMAP	184	0	822
Hordeum vulgarum	GOMAP	101	0	815
	GoldStandard	0	4	0
Medicago truncatula A17	GOMAP	0	0	798
Medicago truncatula R108	GOMAP	0	0	803
Oryza sativa	GOMAP	111	2	869
	GoldStandard	38	556	0
	Gramene61-IEA	10	14	0
Phaseolus vulgaris	GOMAP	0	0	783
Sorghum bicolor	GOMAP	690	59	783
Triticum aestivum	GOMAP	285	0	1132
	GoldStandard	0	10	0
	Gramene61-IEA	47	48	0
Vigna unguiculata	GOMAP	0	0	811
Zea mays B73.v3	GOMAP	1106	70	709
	GoldStandard-OFF	1	0	0
	Gramene49	94	2	0
	Phytozome	54	0	0
Zea mays B73.v4	GOMAP	752	83	848
	GoldStandard	55	174	0
	Gramene61-IEA	99	157	0
Zea mays Mo17	GOMAP	726	77	823
Zea mays PH207	GOMAP	798	76	830
Zea mays W22	GOMAP	754	82	840

Download this table (CSV)

poses a problem when annotations are predicted that are not found in the gold standard: Is this truly a wrong prediction or is the gold standard incomplete? Especially in our case where the predictions not only contain more annotations than the gold standard, but are also more diverse. In effect this means that a quality score as calculated above may not only describe the quality of the prediction, but to some extent also the completeness of the gold standard itself. At least we can see here that gold standards with a median of 3 annotations per gene resulted in higher quality scores than gold standards with less annotations per gene, even though predictions were generated the same way in all cases. `TODO maybe put a figure with regression quality score/median annotions per gene or something` In conclusion this means that truly making a statement about the quality of a prediction set would require the ideal and complete gold standard. The scores we can generate so far are by far not as meaningful.

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

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.

			Genes Annotated[%][a]				Annotations[b]				Median Ann. per G.[c]			
Genome	Genes	Dataset	CC	BF	MP	**A**	CC	BF	MP	**A**	CC	BF	MP	**A**
Arachis hypogaea	67,124	GOMAP	85.90	84.69	100.00	**100.00**	153,052	132,624	493,588	**779,264**	2	2	6	**10**
Brachypodium distachyon	34,310	GOMAP	81.37	85.36	100.00	**100.00**	75,536	69,452	255,616	**400,604**	2	2	6	**10**
Glycine max	52,872	GOMAP	87.02	88.93	100.00	**100.00**	128,836	113,491	417,340	**659,667**	2	2	6	**11**
Gossypium raimondii	37,505	GOMAP	93.06	92.37	100.00	**100.00**	96,442	85,234	307,727	**489,403**	2	2	6	**11**
Hordeum vulgarum	39,734	GOMAP	88.67	91.77	100.00	**100.00**	87,780	80,015	272,625	**440,420**	2	2	5	**10**
		GoldStandard	0.02	0.05	0.05	**0.07**	7	23	45	**75**	0	1	1	**2**
Medicago truncatula A17	50,444	GOMAP	83.90	86.69	100.00	**100.00**	107,019	99,452	363,877	**570,348**	2	2	6	**10**
Medicago truncatula R108	55,706	GOMAP	72.39	90.14	100.00	**100.00**	111,991	107,769	382,133	**601,893**	1	2	5	**9**
Oryza sativa	35,825	GOMAP	79.87	83.31	100.00	**100.00**	72,415	64,386	248,495	**385,296**	2	2	6	**9**
		GoldStandard	15.98	20.61	25.21	**31.79**	7,730	11,060	19,378	**38,176**	1	1	1	**3**
		Gramene61-IEA	30.07	43.37	46.63	**59.86**	14,633	32,787	39,105	**86,529**	1	1	1	**3**
Phaseolus vulgaris	27,433	GOMAP	94.52	93.07	100.00	**100.00**	71,658	64,328	229,449	**365,435**	2	2	6	**11**
Sorghum bicolor	34,129	GOMAP	82.48	85.98	100.00	**100.00**	76,343	69,937	259,229	**405,509**	2	2	6	**10**
Triticum aestivum	107,891	GOMAP	88.60	90.98	100.00	**100.00**	267,317	218,186	785,689	**1,271,192**	2	2	6	**10**
		GoldStandard	0.89	0.57	1.54	**1.73**	1,590	923	4,807	**7,323**	1	0	2	**3**
		Gramene61-IEA	26.74	55.24	48.72	**70.24**	38,975	109,319	109,518	**257,832**	0	1	1	**2**
Vigna unguiculata	29,773	GOMAP	91.26	91.08	100.00	**100.00**	75,513	68,040	243,094	**386,647**	2	2	6	**11**
Zea mays B73.v3	39,469	GOMAP	88.33	96.42	99.99	**100.00**	134,917	87,166	291,091	**513,174**	3	2	6	**11**
		GoldStandard-OFF	3.92	0.15	0.38	**4.14**	1,565	65	299	**1,929**	1	0	0	**1**
		Gramene49	29.98	45.58	40.03	**55.55**	20,072	31,056	30,089	**81,217**	1	1	1	**3**
		Phytozome	11.46	34.78	28.79	**40.87**	4,787	19,044	13,100	**36,931**	0	1	1	**2**
Zea mays B73.v4	39,324	GOMAP	93.36	94.93	100.00	**100.00**	88,468	81,963	278,518	**448,949**	2	2	6	**10**
		GoldStandard	21.23	25.60	30.82	**38.07**	11,510	15,019	25,737	**52,428**	1	1	1	**3**
		Gramene61-IEA	37.57	56.11	60.94	**74.13**	20,265	47,657	58,110	**126,525**	1	1	2	**3**
Zea mays Mo17	38,620	GOMAP	87.04	90.88	100.00	**100.00**	87,221	78,938	277,586	**443,745**	2	2	6	**10**
Zea mays PH207	40,557	GOMAP	86.71	90.62	100.00	**100.00**	90,267	85,223	288,474	**463,964**	2	2	6	**10**
Zea mays W22	40,690	GOMAP	90.89	92.59	100.00	**100.00**	95,043	84,750	289,576	**469,369**	2	2	6	**10**

Download this table (CSV)

[a] 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)

[b] How many annotations in the CC, BF, and MP aspect does this dataset contain? A = How many in total? 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? Please note that A ≠ CC + BF + MP

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

		SimGIC2			TC-AUCPCR		
Genome	Dataset	CC	BF	MP	CC	BF	MP
Hordeum vulgarum	GOMAP	0.309334	0.424286	0.180110	NaN	0.000493	0.000388
Oryza sativa	GOMAP	0.499374	0.448762	0.213827	0.274482	0.261767	0.142014
	Gramene61-IEA	0.416283	0.415342	0.324981	0.170662	0.257685	0.126104
Triticum aestivum	GOMAP	0.474972	0.418110	0.202528	0.014568	0.005485	0.009287
	Gramene61-IEA	0.384973	0.346840	0.191962	0.004446	0.006115	0.004811
Zea mays B73.v4	GOMAP	0.498781	0.429594	0.212130	0.276072	0.245183	0.133683
	Gramene61-IEA	0.368491	0.411399	0.323139	0.159213	0.229186	0.130063

Download this table (CSV)

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