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. (2019) 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	3437	13	912
Brachypodium distachyon	GOMAP	2512	49	789
Cannabis sativa	GOMAP	1714	6	757
Glycine max	GOMAP	3333	10	930
Gossypium raimondii	GOMAP	1781	7	822
Hordeum vulgare	GOMAP	1877	8	815
	GoldStandard	0	4	0
Medicago truncatula A17	GOMAP	2673	10	798
	GoldStandard	0	7	0
	Gramene62-IEA	429	251	0
Medicago truncatula R108	GOMAP	4168	7	803
Oryza sativa	GOMAP	1642	7	869
	GoldStandard	44	581	0
	Gramene61-IEA	242	28	0
Phaseolus vulgaris	GOMAP	1190	6	783
Pinus lambertiana	GOMAP	1839	4	587
Sorghum bicolor	GOMAP	2384	66	783
Triticum aestivum	GOMAP	9624	17	1132
	GoldStandard	0	10	0
	Gramene61-IEA	706	88	0
Vigna unguiculata	GOMAP	1269	6	811
Zea mays B73.v3	GOMAP	1805	92	709
	GoldStandard	1	11	0
	Gramene49	221	6	0
	Phytozome	132	0	0
Zea mays B73.v4	GOMAP	2077	89	848
	GoldStandard	65	207	0
	Gramene61-IEA	600	178	0
Zea mays Mo17	GOMAP	2346	83	823
	GoldStandard	1	60	0
Zea mays PH207	GOMAP	2676	82	830
	GoldStandard	1	70	0
Zea mays W22	GOMAP	2681	88	840
	GoldStandard	1	52	0

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

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	85.85	84.68	100.00	**100.00**	150,525	132,144	493,145	**775,814**	2	2	6	**10.0**
Brachypodium distachyon	34,310	GOMAP	81.33	85.35	100.00	**100.00**	74,172	69,213	255,397	**398,782**	2	2	6	**10.0**
Cannabis sativa	33,677	GOMAP	94.22	95.48	100.00	**100.00**	85,755	73,614	262,741	**422,110**	2	2	6	**11.0**
Glycine max	52,872	GOMAP	86.95	88.92	100.00	**100.00**	126,470	113,068	416,989	**656,527**	2	2	6	**11.0**
Gossypium raimondii	37,505	GOMAP	93.00	92.37	100.00	**100.00**	95,419	84,910	307,470	**487,799**	2	2	6	**11.0**
Hordeum vulgare	39,734	GOMAP	88.57	91.76	100.00	**100.00**	86,489	79,727	272,420	**438,636**	2	2	5	**10.0**
		GoldStandard	0.02	0.05	0.05	**0.07**	7	23	45	**75**	0	1	1	**2.0**
Medicago truncatula A17	50,444	GOMAP	83.79	86.69	100.00	**100.00**	104,902	99,155	363,608	**567,665**	2	2	6	**10.0**
		GoldStandard	0.03	0.06	0.05	**0.07**	18	35	48	**101**	0	1	1	**2.5**
		Gramene62-IEA	34.15	50.66	38.98	**65.71**	33,350	62,800	39,230	**135,713**	1	1	1	**3.0**
Medicago truncatula R108	55,706	GOMAP	72.10	90.14	100.00	**100.00**	108,388	107,499	381,831	**597,718**	1	2	5	**9.0**
Oryza sativa	35,825	GOMAP	79.78	83.31	100.00	**100.00**	71,306	64,150	248,304	**383,760**	2	2	6	**9.0**
		GoldStandard	15.98	20.61	25.21	**31.79**	7,729	11,033	19,375	**38,145**	1	1	1	**3.0**
		Gramene61-IEA	29.76	43.21	46.63	**59.79**	14,475	32,703	39,101	**86,283**	0	1	1	**3.0**
Phaseolus vulgaris	27,433	GOMAP	94.48	93.06	100.00	**100.00**	70,987	64,022	229,230	**364,239**	2	2	6	**11.0**
Pinus lambertiana	31,007	GOMAP	92.67	95.91	100.00	**100.00**	71,247	68,315	212,248	**351,810**	2	2	5	**10.0**
Sorghum bicolor	34,129	GOMAP	82.44	85.98	100.00	**100.00**	75,145	69,659	259,004	**403,808**	2	2	6	**10.0**
Triticum aestivum	107,891	GOMAP	88.53	90.98	100.00	**100.00**	259,318	217,467	785,051	**1,261,836**	2	2	6	**10.0**
		GoldStandard	0.89	0.57	1.54	**1.73**	1,590	923	4,807	**7,323**	1	0	2	**3.0**
		Gramene61-IEA	26.47	55.03	48.72	**70.23**	38,593	109,013	109,507	**257,133**	0	1	1	**2.0**
Vigna unguiculata	29,773	GOMAP	91.21	91.08	100.00	**100.00**	74,791	67,734	242,847	**385,372**	2	2	6	**11.0**
Zea mays B73.v3	39,469	GOMAP	88.33	96.41	99.99	**100.00**	134,622	87,007	290,824	**512,453**	3	2	6	**11.0**
		GoldStandard	3.89	0.15	0.38	**4.10**	1,554	65	299	**1,918**	1	0	0	**1.0**
		Gramene49	29.98	45.58	40.03	**55.55**	20,066	30,936	30,084	**81,086**	1	1	1	**3.0**
		Phytozome	11.46	34.77	28.79	**40.87**	4,787	18,966	13,100	**36,853**	0	1	1	**2.0**
Zea mays B73.v4	39,324	GOMAP	93.16	94.92	100.00	**100.00**	87,648	81,665	278,305	**447,618**	2	2	6	**10.0**
		GoldStandard	21.23	25.60	30.82	**38.07**	11,505	14,986	25,732	**52,385**	1	1	1	**3.0**
		Gramene61-IEA	37.12	55.97	60.94	**74.11**	19,870	47,547	58,093	**126,003**	1	1	2	**3.0**
Zea mays Mo17	38,620	GOMAP	86.98	90.87	100.00	**100.00**	86,074	78,650	277,395	**442,119**	2	2	6	**10.0**
		GoldStandard	3.22	0.14	0.35	**3.42**	1,266	64	277	**1,607**	1	0	0	**1.0**
Zea mays PH207	40,557	GOMAP	86.55	90.61	100.00	**100.00**	88,962	84,910	288,208	**462,080**	2	2	6	**10.0**
		GoldStandard	3.15	0.14	0.33	**3.34**	1,302	63	266	**1,631**	1	0	0	**1.0**
Zea mays W22	40,690	GOMAP	90.77	92.58	100.00	**100.00**	93,622	84,450	289,364	**467,436**	2	2	6	**10.0**
		GoldStandard	2.92	0.13	0.29	**3.08**	1,205	59	241	**1,505**	1	0	0	**1.0**

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

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

Plyusnin, I., Holm, L., and Törönen, P. (2019). Novel comparison of evaluation metrics for gene ontology classifiers reveals drastic performance differences. *PLOS Computational Biology* 15, e1007419. doi:10.1371/journal.pcbi.1007419

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

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

		SimGIC2			TC-AUCPCR		
Genome	Dataset	CC	BF	MP	CC	BF	MP
Medicago truncatula A17	GOMAP	0.260687	0.192466	0.091357	0.000332	0.000573	0.000789
	Gramene62-IEA	0.755529	0.241984	0.265019	0.000702	0.000736	0.011422
Oryza sativa	GOMAP	0.489140	0.460077	0.214754	0.282023	0.265899	0.143394
	Gramene61-IEA	0.399104	0.422182	0.326020	0.170803	0.255999	0.127307
Triticum aestivum	GOMAP	0.455167	0.421998	0.202177	0.014502	0.005619	0.009458
	Gramene61-IEA	0.367460	0.352624	0.194847	0.004466	0.006177	0.004972
Zea mays B73.v3	GOMAP	0.198179	0.359422	0.094640	0.034380	0.001644	0.001021
	Gramene49	0.252350	0.365922	0.163637	0.049989	0.003238	0.001972
	Phytozome	0.152410	0.355794	0.100792	0.013784	0.003064	0.000539
Zea mays B73.v4	GOMAP	0.486568	0.440989	0.213537	0.283686	0.247096	0.135719
	Gramene61-IEA	0.349744	0.416108	0.324970	0.157917	0.230439	0.131515
Zea mays Mo17	GOMAP	0.232016	0.292290	0.104776	0.040932	0.001932	0.000895
Zea mays PH207	GOMAP	0.234422	0.272876	0.096926	0.037857	0.001743	0.000813
Zea mays W22	GOMAP	0.235595	0.276437	0.103866	0.038546	0.001730	0.000868

Download this table (CSV)

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