SUPPLEMENTARY MATERIALS
to the article E.V. Malyugin, D.A. Afonnikov
"OrthoML2GO: homology-based protein function prediction using orthogroups and machine learning"

Table S1. Information about sequences and annotations

Species name	Annotation source	Number of sequences	Sequence source	Annotation link
Arabidopsis thaliana	TAIR	27,655	Araport11	https://www.arabidopsis.org/download/list?dir=GO_and_PO_Annotations%2FGene_Ontology_Annotations
Homo sapiens	EBI Gene Ontology Annotation Database	19,763	UniProtKB	https://current.geneontology.org/products/pages/downloads.html
Drosophila melanogaster	FlyBase	28,543 (includes isoforms)	FlyBase	https://current.geneontology.org/products/pages/downloads.html
Solanum tuberosum	SpuDB	40,722 (includes isoforms)	SpudDB	https://spuddb.uga.edu/dm_v6_1_download.shtml
Danio rerio	ZFIN	33,428 (includes isoforms)	UniProtKB	https://current.geneontology.org/products/pages/downloads.html
Chlamydomonas reinhardtii	PhycoCosm	16,090	PhycoCosm	https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=ChlreiCC4532_1
Oryza sativa	RGAP	34,226 (includes isoforms)	RGAP	https://rice.uga.edu/download_osa1r7.shtml

Access date for all databases: January 2025.

Table S2. Machine learning parameters

Short name	Description
1. Identity	Percentage of identical residues in the aligned regions of query and target (%)
2. Query coverage	Percentage of the query sequence aligned to the target (%)
3. Target coverage	Percentage of the target sequence aligned to the query (%)
4. Coverage ratio	Ratio of query coverage to target coverage (%)
5. Query length	Total number of amino acid residues in the query sequence
6. Target length	Total number of amino acid residues in the target sequence
7. Length difference	Difference between query and target lengths
8. Bitscore	Normalized alignment quality score computed by the USEARCH algorithm
9. Number of differences	Total number of non-identical positions in the alignment. *Includes both mismatches and gaps*
10. GOk%	Frequency of a GO term for the query among all its homologs
11. RMSD	Root mean square deviation of amino acid frequencies between query and target

Table S3. Distribution of sequences in the training and test sets for the combined dataset

Species name	Proportion (%)	For training	For testing
Arabidopsis thaliana	16.2	8117	3247
Homo sapiens	11.6	5802	2321
Drosophila melanogaster	8.7	4339	1735
Solanum tuberosum	25.1	12 565	5026
Danio rerio	14.2	7082	2833
Chlamydomonas reinhardtii	4.1	2044	818
Oryza sativa	20.1	10 051	4020
Total	**100**	**50 000**	**20 000**

Table S4. Dependence of the prediction performance of *Arabidopsis thaliana* proteins on the parameter *k* for the KNN, OG, and KNN+OG

The highest F1-score values for each prediction method are highlighted in bold.

Arabidopsis thaliana	KNN	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$
	TP	50964	78288	81848	83370	84427	85075
	FP	55076	75830	83714	88792	93149	97234
	FN	97282	69958	66398	64876	63819	63171
	Precision (%)	48.06	50.80	49.44	48.43	47.54	46.67
	Recall (%)	34.38	52.81	55.21	56.24	56.95	57.39
	Accuracy (%)	41.22	51.80	52.32	52.33	52.25	52.03
	F1-score (%)	40.08	51.78	**52.16**	52.04	51.82	51.47
	OG	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$
	TP	51904	52646	52714	52682	52628	52591
	FP	57056	56813	57144	57603	57966	58399
	FN	96342	95600	95532	95564	95618	95655
	Precision (%)	47.64	48.10	47.98	47.77	47.59	47.38
	Recall (%)	35.01	35.51	35.56	35.54	35.50	35.48
	Accuracy (%)	41.32	41.80	41.77	41.65	41.54	41.43
	F1-score (%)	40.36	**40.86**	40.85	40.75	40.66	40.57
	KNN+OG	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$
	TP	59523	80183	83425	84866	85897	86566
	FP	66608	79463	86882	92156	96729	101193
	FN	88723	68063	64821	63380	62349	61680
	Precision (%)	47.19	50.23	48.99	47.94	47.03	46.10
	Recall (%)	40.15	54.09	56.27	57.25	57.94	58.39
	Accuracy (%)	43.67	52.16	52.63	52.59	52.49	52.25
	F1-score (%)	43.39	52.09	**52.38**	52.18	51.92	51.53

Table S5. Dependence of the prediction performance of *Homo sapiens* proteins on the parameter *k* for the KNN, OG, and KNN+OG

The highest F1-score values for each prediction method are highlighted in bold.

Homo sapiens	KNN	*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25	*k* = 30
	TP	131087	170362	188604	198406	204878	209641
	FP	40373	53339	59523	63140	65925	68633
	FN	172018	132743	114501	104699	98227	93464
	Precision (%)	76.45	76.16	76.01	75.86	75.66	75.34
	Recall (%)	43.25	56.21	62.22	65.46	67.59	69.16
	Accuracy (%)	59.85	66.18	69.12	70.66	71.62	72.25
	F1-score (%)	55.25	64.68	68.43	70.28	71.40	**72.12**
	OG	*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25	*k* = 30
	TP	159475	166703	163245	145178	142387	141883
	FP	48116	49864	49336	44618	43978	44116
	FN	143630	136402	139860	157927	160718	161222
	Precision (%)	76.82	76.98	76.79	76.49	76.40	76.28
	Recall (%)	52.61	55.00	53.86	47.90	46.98	46.81
	Accuracy (%)	64.72	65.99	65.32	62.19	61.69	61.55
	F1-score (%)	62.45	**64.16**	63.31	58.91	58.18	58.02
	KNN+OG	*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25	*k* = 30
	TP	182129	197391	203072	203664	207994	211708
	FP	56022	61754	64624	65376	67618	70047
	FN	120976	105714	100033	99441	95111	91397
	Precision (%)	76.48	76.17	75.86	75.70	75.47	75.14
	Recall (%)	60.09	65.12	67.00	67.19	68.62	69.85
	Accuracy (%)	68.28	70.65	71.43	71.45	72.04	72.49
	F1-score (%)	67.30	70.21	71.15	71.19	71.88	**72.40**

Table S6. Dependence of the prediction performance of *Drosophila melanogaster* proteins on the parameter *k* for the KNN, OG, and KNN+OG

The highest F1-score values for each prediction method are highlighted in bold.

Drosophila melanogaster	KNN	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$
	TP	58928	66842	69357	70072	70551	70865
	FP	32540	38720	41629	42565	43314	43943
	FN	40287	32373	29858	29143	28664	28350
	Precision (%)	64.42	63.32	62.49	62.21	61.96	61.72
	Recall (%)	59.39	67.37	69.91	70.63	71.11	71.43
	Accuracy (%)	61.91	65.35	66.20	66.42	66.53	66.58
	F1-score (%)	61.81	65.28	65.99	66.15	**66.22**	**66.22**
	OG	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$
	TP	50104	51961	52658	52954	53131	53282
	FP	25948	26727	27009	27122	27217	27279
	FN	49111	47254	46557	46261	46084	45933
	Precision (%)	65.88	66.03	66.10	66.13	66.13	66.14
	Recall (%)	50.50	52.37	53.07	53.37	53.55	53.70
	Accuracy (%)	58.19	59.20	59.59	59.75	59.84	59.92
	F1-score (%)	57.17	58.41	58.87	59.07	59.18	**59.28**
	KNN+OG	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$
	TP	67145	71077	72753	73298	73725	73977
	FP	36543	40152	42425	43214	43924	44511
	FN	32070	28138	26462	25917	25490	25238
	Precision (%)	64.76	63.90	63.17	62.91	62.67	62.43
	Recall (%)	67.68	71.64	73.33	73.88	74.31	74.56
	Accuracy (%)	66.22	67.77	68.25	68.39	68.49	68.50
	F1-score (%)	66.18	67.55	67.87	67.95	**67.99**	67.96

Table S7. Dependence of the prediction performance of *Solanum tuberosum* proteins on the parameter *k* for the KNN, OG, and KNN+OG

The highest F1-score values for each prediction method are highlighted in bold.

Solanum tuberosum	KNN	*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25	*k* = 30
	TP	71253	100451	108212	112471	113883	114766
	FP	73685	107688	122962	131408	135105	137629
	FN	216080	186882	179121	174862	173450	172567
	Precision (%)	49.16	48.26	46.81	46.12	45.74	45.47
	Recall (%)	24.80	34.96	37.66	39.14	39.63	39.94
	Accuracy (%)	36.98	41.61	42.24	42.63	42.69	42.71
	F1-score (%)	32.97	40.55	41.74	42.35	42.47	**42.53**
	OG	*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25	*k* = 30
	TP	87068	88704	88922	88998	89148	89125
	FP	88238	89133	89296	89542	90253	90715
	FN	200265	198629	198411	198335	198185	198208
	Precision (%)	49.67	49.88	49.90	49.85	49.69	49.56
	Recall (%)	30.30	30.87	30.95	30.97	31.03	31.02
	Accuracy (%)	39.98	40.38	40.42	40.41	40.36	40.29
	F1-score (%)	37.64	38.14	38.20	**38.21**	38.20	38.16
	KNN+OG	*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25	*k* = 30
	TP	93603	106071	111584	115589	117025	117899
	FP	100298	117031	129606	137810	141956	144823
	FN	193730	181262	175749	171744	170308	169434
	Precision (%)	48.27	47.54	46.26	45.62	45.19	44.88
	Recall (%)	32.58	36.92	38.83	40.23	40.73	41.03
	Accuracy (%)	40.43	42.23	42.55	42.92	42.96	42.95
	F1-score (%)	38.90	41.56	42.22	42.75	42.84	**42.87**

Table S8. Dependence of the prediction performance of *Danio rerio* proteins on the parameter *k* for the KNN, OG, and KNN+OG

The highest F1-score values for each prediction method are highlighted in bold.

Danio rerio	KNN	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$
	TP	88585	113577	119431	120557	121046	121447
	FP	14537	20267	23113	24919	26502	27980
	FN	96997	72005	66151	65025	64536	64135
	Precision (%)	85.90	84.86	83.79	82.87	82.04	81.28
	Recall (%)	47.73	61.20	64.35	64.96	65.23	65.44
	Accuracy (%)	66.82	73.03	74.07	73.92	73.63	73.36
	F1-score (%)	61.37	71.11	72.80	72.83	72.67	72.50
	OG	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$
	TP	109259	109618	109852	110044	110154	110319
	FP	17350	17290	17334	17324	17355	17373
	FN	76323	75964	75730	75538	75428	75263
	Precision (%)	86.30	86.38	86.37	86.40	86.39	86.39
	Recall (%)	58.87	59.07	59.19	59.30	59.36	59.44
	Accuracy (%)	72.59	72.72	72.78	72.85	72.87	72.92
	F1-score (%)	69.99	70.16	70.25	70.33	70.37	70.43
	KNN+OG	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$	$k = 30$
	TP	123510	124789	125567	126028	126355	126725
	FP	21765	23970	25952	27545	29069	30522
	FN	62072	60793	60015	59554	59227	58857
	Precision (%)	85.02	83.89	82.87	82.06	81.30	80.59
	Recall (%)	66.55	67.24	67.66	67.91	68.09	68.29
	Accuracy (%)	75.79	75.56	75.27	74.99	74.69	74.44
	F1-score (%)	**74.66**	74.65	74.50	74.32	74.11	73.93

Table S9. Dependence of the prediction performance of *Chlamydomonas reinhardtii* proteins on the parameter *k* for the KNN, OG, and KNN+OG

The highest F1-score values for each prediction method are highlighted in bold.

Chlamydomonas reinhardtii	KNN		*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25	*k* = 30
		TP	10141	10252	10299	10306	10340	10351
		FP	27821	29027	29345	29555	29719	29898
		FN	15362	15251	15204	15197	15163	15152
		Precision (%)	26.71	26.10	25.98	25.85	25.81	25.72
		Recall (%)	39.76	40.20	40.38	40.41	40.54	40.59
		Accuracy (%)	33.24	33.15	33.18	33.13	33.18	33.15
		F1-score (%)	**31.96**	31.65	31.62	31.53	31.54	31.48
	OG		*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25	*k* = 30
		TP	10115	10173	10199	10197	10218	10219
		FP	30947	31022	31003	30954	30956	30942
		FN	15388	15330	15304	15306	15285	15284
		Precision (%)	24.63	24.69	24.75	24.78	24.82	24.83
		Recall (%)	39.66	39.89	39.99	39.98	40.07	40.07
		Accuracy (%)	32.15	32.29	32.37	32.38	32.44	32.45
		F1-score (%)	30.39	30.50	30.58	30.60	30.65	**30.66**
	KNN+OG		*k* = 5	*k* = 10	*k* = 15	*k* = 20	*k* = 25	*k* = 30
		TP	10655	10745	10784	10789	10819	10830
		FP	34104	35215	35494	35685	35834	35997
		FN	14848	14758	14719	14714	14684	14673
		Precision (%)	23.81	23.38	23.30	23.22	23.19	23.13
		Recall (%)	41.78	42.13	42.29	42.30	42.42	42.47
		Accuracy (%)	32.79	32.76	32.79	32.76	32.81	32.80
		F1-score (%)	**30.33**	30.07	30.05	29.98	29.99	29.95

Table S10. Annotation of *Ostreococcus lucimarinus* protein sequences using the OrthoML2GO method

Number of analyzed sequences	7603
Number of sequences with homologs found	7187
Number of sequences with predicted annotation:	5273
1. Biological Process (BP)	4144
2. Molecular Function (MF)	4422
3. Cellular Component (CC)	4314
Execution time (248 cores)	1 hour 27 minutes 49 seconds

Table S11. List of the 5 most frequent predicted GO terms for *Ostreococcus lucimarinus* by ontology

GO aspect	GO term	Count	Definition
BP	GO:0016310	275	Phosphorylation
BP	GO:0006412	267	Translation
BP	GO:0055085	206	Transmembrane transport
BP	GO:0032259	192	Methylation
BP	GO:0006508	165	Proteolysis
CC	GO:0016020	1596	Membrane
CC	GO:0005634	1113	Nucleus
CC	GO:0005737	1030	Cytoplasm
CC	GO:0005829	583	Cytosol
CC	GO:0005739	460	Mitochondrion
MF	GO:0005524	767	ATP binding
MF	GO:0000166	760	Nucleotide binding
MF	GO:0016740	752	Transferase activity
MF	GO:0016787	730	Hydrolase activity
MF	GO:0046872	667	Metal ion binding

Table S12. Algorithm parameters

The parameters of the gradient boosting (XGBoost) algorithm	The parameters of the Random Forest algorithm
params <- list(objective = "binary:logistic", eval_metric = "auc", eta = 0.01, max_depth = 4, gamma = 0.5, subsample = 0.7, colsample_bytree = 0.7, min_child_weight = 1, lambda = 3, alpha = 0.2, tree_method = "hist") final_model <- xgb.train(params = params, data = dtrain, nrounds = 1000, early_stopping_rounds = 20, print_every_n = 10, maximize = TRUE)	rf_model <- randomForest(x = x_train, y = y_train, ntree = 250, mtry = 3, importance = TRUE, do.trace = 50, maxnodes = 20)