

ПРИЛОЖЕНИЯ
к статье Е.В. Малюгина, Д.А. Афонникова
«OrthoML2GO: предсказание функций белков по гомологии
с использованием ортогрупп и алгоритмов машинного обучения»

Таблица S1. Информация о последовательностях и аннотациях

Название вида	Источник аннотации	Число последовательностей	Источник последовательностей	Ссылка на аннотацию
<i>Arabidopsis thaliana</i>	TAIR	27 655	Araport11	https://www.arabidopsis.org/download/list?dir=GO_and_PO_Annotations%2FGene_Ontology_Annotations
<i>Homo sapiens</i>	EBI Gene Ontology Annotation Database	19 763	UniProtKB	https://current.geneontology.org/products/pages/downloads.html
<i>Drosophila melanogaster</i>	FlyBase	28 543 (Включая изоформы)	FlyBase	https://current.geneontology.org/products/pages/downloads.html
<i>Solanum tuberosum</i>	SpuDB	40 722 (Включая изоформы)	SpudDB	https://spuddb.uga.edu/dm_v6_1_download.shtml
<i>Danio rerio</i>	ZFIN	33 428 (Включая изоформы)	UniProtKB	https://current.geneontology.org/products/pages/downloads.html
<i>Chlamydomonas reinhardtii</i>	PhycoCosm	16 090	PhycoCosm	https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=ChlreiCC4532_1
<i>Oryza sativa</i>	RGAP	34 226 (Включая изоформы)	RGAP	https://rice.uga.edu/download_osa1r7.shtml

Дата доступа ко всем базам данных: 01.2025.

Таблица S2. Параметры для машинного обучения

Краткое название	Описание
1. Identity	Доля идентичных аминокислотных остатков в выровненных участках query и target (%)
2. Query coverage	Доля последовательности query, которая была выровнена с target (%)
3. Target coverage	Доля последовательности target, которая была выровнена с query (%)
4. Coverage ratio	Отношение покрытия query к target (%)
5. Query length	Общее количество аминокислотных остатков в последовательности query
6. Target length	Общее количество аминокислотных остатков в последовательности target
7. Length difference	Разница между длинами query и target
8. Bitscore	Нормализованная оценка качества выравнивания, вычисленная алгоритмом USEARCH
9. Number of differences	Суммарное количество неидентичных позиций в выравнивании. Включает в себя как <i>mismatches</i> (несовпадения аминокислот), так и <i>gaps</i> (зэпы)
10. GOk%	Частота встречаемости термина GO для query среди всех его гомологов
11. RMSD	Среднеквадратичное отклонение частот аминокислот между query и target

Таблица S3. Распределение последовательностей на обучающую и тестовую для комбинированной выборки

Название организма	Доля (%)	Для обучения	Для тестирования
<i>Arabidopsis thaliana</i>	16.2	8117	3247
<i>Homo sapiens</i>	11.6	5802	2321
<i>Drosophila melanogaster</i>	8.7	4339	1735
<i>Solanum tuberosum</i>	25.1	12 565	5026
<i>Danio rerio</i>	14.2	7082	2833
<i>Chlamydomonas reinhardtii</i>	4.1	2044	818
<i>Oryza sativa</i>	20.1	10 051	4020
Итого	100	50 000	20 000

Таблица S4. Зависимость точности предсказания на белках *Arabidopsis thaliana* от параметра *k* для методов KNN, OG и KNN+OG

Жирным шрифтом выделены наибольшие значения F1 для каждого метода предсказания.

Arabidopsis thaliana	KNN		<i>k</i> = 5	<i>k</i> = 10	<i>k</i> = 15	<i>k</i> = 20	<i>k</i> = 25	<i>k</i> = 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		<i>k</i> = 5	<i>k</i> = 10	<i>k</i> = 15	<i>k</i> = 20	<i>k</i> = 25	<i>k</i> = 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		<i>k</i> = 5	<i>k</i> = 10	<i>k</i> = 15	<i>k</i> = 20	<i>k</i> = 25	<i>k</i> = 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

Таблица S5. Зависимость точности предсказания на белках *Homo sapiens* от параметра *k* для методов KNN, OG и KNN+OG
Жирным шрифтом выделены наибольшие значения F1 для каждого метода предсказания.

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

Таблица S6. Зависимость точности предсказания на белках *Drosophila melanogaster* от параметра *k* для методов KNN, OG и KNN+OG

Жирным шрифтом выделены наибольшие значения F1 для каждого метода предсказания.

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

Таблица S7. Зависимость точности предсказания на белках *Solanum tuberosum* от параметра *k* для методов KNN, OG и KNN+OG

Жирным шрифтом выделены наибольшие значения F1 для каждого метода предсказания.

<i>Solanum tuberosum</i>	KNN		<i>k</i> = 5	<i>k</i> = 10	<i>k</i> = 15	<i>k</i> = 20	<i>k</i> = 25	<i>k</i> = 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		<i>k</i> = 5	<i>k</i> = 10	<i>k</i> = 15	<i>k</i> = 20	<i>k</i> = 25	<i>k</i> = 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		<i>k</i> = 5	<i>k</i> = 10	<i>k</i> = 15	<i>k</i> = 20	<i>k</i> = 25	<i>k</i> = 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

Таблица S8. Зависимость точности предсказания на белках *Danio rerio* от параметра *k* для методов KNN, OG и KNN+OG

Жирным шрифтом выделены наибольшие значения F1 для каждого метода предсказания.

Danio rerio	KNN		<i>k</i> = 5	<i>k</i> = 10	<i>k</i> = 15	<i>k</i> = 20	<i>k</i> = 25	<i>k</i> = 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		<i>k</i> = 5	<i>k</i> = 10	<i>k</i> = 15	<i>k</i> = 20	<i>k</i> = 25	<i>k</i> = 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		<i>k</i> = 5	<i>k</i> = 10	<i>k</i> = 15	<i>k</i> = 20	<i>k</i> = 25	<i>k</i> = 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

Таблица S9. Зависимость точности предсказания на белках *Chlamydomonas reinhardtii* от параметра *k* для методов KNN, OG и KNN+OG

Жирным шрифтом выделены наибольшие значения F1 для каждого метода предсказания.

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

Таблица S10. Аннотация белковых последовательностей *Ostreococcus lucimarinus* с помощью метода OrthoML2GO

Число последовательностей для анализа	7603
Число последовательностей с найденными гомологами	7187
Число последовательностей с предсказанной аннотацией:	5273
1. Биологические процессы (BP)	4144
2. Молекулярные функции (MF)	4422
3. Клеточные компоненты (CC)	4314
Время выполнения (248 ядер)	1 час 27 минут 49 секунд

Таблица S11. Список 5 наиболее частых терминов GO предсказанных для *Ostreococcus lucimarinus* по словарям

Словарь	Термин GO	Количество	Определение
BP	GO:0016310	275	Фосфорилирование (phosphorylation)
BP	GO:0006412	267	Трансляция (translation)
BP	GO:0055085	206	Трансмембранный транспорт (transmembrane transport)
BP	GO:0032259	192	Метилирование (methylation)
BP	GO:0006508	165	Протеолиз (proteolysis)
CC	GO:0016020	1596	Мембрана (membrane)
CC	GO:0005634	1113	Ядро (nucleus)
CC	GO:0005737	1030	Цитоплазма (cytoplasm)
CC	GO:0005829	583	Цитозоль (cytosol)
CC	GO:0005739	460	Митохондрия (mitochondrion)
MF	GO:0005524	767	Связывание АТФ (ATP binding)
MF	GO:0000166	760	Связывание нуклеотидов (nucleotide binding)
MF	GO:0016740	752	Трансферазная активность (transferase activity)
MF	GO:0016787	730	Гидролазная активность (hydrolase activity)
MF	GO:0046872	667	Связывание ионов металлов (metal ion binding)

Таблица S12. Параметры алгоритмов

Параметры алгоритма градиентного бустинга (XGBoost):	Параметры алгоритма случайного леса (Random Forest):
<pre>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)</pre>	<pre>rf_model <- randomForest(x = x_train, y = y_train, ntree = 250, mtry = 3, importance = TRUE, do.trace = 50, maxnodes = 20)</pre>