Title: Tools and Analysis of PANTHER Genomic Data Annotations

Authors: M. Fong[1], I. Holmes[1]

Affiliations:[1]Department of Bioengineering, University of California, Berkeley

*Correspondence to: mfong92@berkeley.edu

The PANTHER (Protein ANalysis THrough Evolutionary Relationships) database contains gene families, trees, and functional annotations and is located at http://pantherdb.org. This following discussion is a description of a tool created for retrieving these annotations and an analysis of the number and quality of the annotations that are provided through the PAINT tool. The results indicate a possible deficiency in the annotations that are provided through the Gene Ontology (GO).

Two scripts were created for this analysis: the **Branch Generation Script**, which parses Newick tree files and returns a list of branches (as tuples containing two AN numbers and a branch length), and the **Annotation Mapping Script**, which translates each AN number to the corresponding GO annotations that are curated on each node. This was done using the resources from PANTHER and GO online repositories.

The summary of the analysis can be seen in Figure 1. This gives a general idea about the prevalence of GO annotations within the PANTHER data set. Of the 17761 total nodes over the 53 different families that were surveyed, only about 25% of them were annotated (including nodes that had annotations solely because they were propagated from their ancestors).

Figure 2 brings into question the actual usefulness of the annotations of the PANTHER data set. The left columns show the unique number of annotation sets across entire gene families, where a set of annotations is counted as the same as another set if all of the annotations of one compose of all of the annotations of the other. Only 5% of the total nodes have unique sets of annotations, which suggest that the majority of annotations come from the propagation down towards a node's descendants. The right columns demonstrate a potential problem with the above claim. Logically, it would be assumed that nodes that share an annotation would be closer together, on average, because of the effect that distance has on evolution. However, the opposite is true of the gene families shown in Figure 3, as the nodes that share an annotation are further apart than nodes that don't share an annotation. Here, distance is calculated as distance that is necessary to have been traveled through the tree. Although this is not a perfect calculation of genetic distance, this finding does point again towards the possible lack of quality of these annotations.

Figure 3 extends this argument by plotting the average distance between nodes for pairs with the corresponding number of shared annotations. We would expect a steady negative slope for this relationship, yet it is mostly flat from between 0-3 annotations shared. This seems to show that the strength of the closeness of two nodes (ie. the number of annotations that they share) has no relationship with the distance between them, which could indicate a failure to properly annotate a random sample of nodes.

As it stands, the annotations that are curated to the PANTHER trees appear not to be on a random selection of nodes, but on groups of nodes where the annotations are the same. More analysis needs to be done to see whether this currently is a useful data set to conduct evolutionary comparison analyses.

PTHR ID	Gene Family Description	Total Nodes	Annotated Nodes	Total Annotations	CC Annotations	MF Annotations	BP Annotations	NOT Annotations
PTHR10000	PHOSPHOSERINE PHOSPHATASE	208	45	124	28	69	27	0
PTHR10003	SUPEROXIDE DISMUTASE [CU-ZN]-RELATED	319	188	1117	395	532	190	0
PTHR10005	SKI ONCOGENE-RELATED	152	24	109	29	37	43	0
PTHR10006	MUCIN-1-RELATED	20	2	4	4	0	0	0
PTHR10009	PROTEIN YELLOW-RELATED	61	35	114	26	31	57	0
PTHR10012	SERINE/THREONINE-PROTEIN PHOSPHATASE 2A REGULATORY SUBUNIT B	142	76	525	214	215	96	0
PTHR10019	SNF5	97	52	329	52	1	276	0
PTHR10025	TETRAHYDROFOLATE DEHYDROGENASE/CYCLOHYDROLASE FAMILY MEMBER	326	91	451	119	249	83	89
PTHR10027	CALCIUM-ACTIVATED POTASSIUM CHANNEL ALPHA CHAIN	206	32	102	30	47	25	1
PTHR10032	ZINC FINGER PROTEIN WITH KRAB AND SCAN DOMAINS	697	365	782	334	372	76	2
PTHR10046	ATP DEPENDENT LON PROTEASE FAMILY MEMBER	283	70	577	101	268	208	0
PTHR10048	PHOSPHATIDYLINOSITOL KINASE	669	354	1090	504	582	4	1
PTHR10063	TUBERIN	203	26	113	38	18	57	0
PTHR10073	DNA MISMATCH REPAIR PROTEIN (MLH, PMS, MUTL)	410	77	465	186	171	108	0
PTHR10125	P2X PURINOCEPTOR	225	37	90	20	49	21	0
PTHR10130	PEROXISOMAL TARGETING SIGNAL 1 RECEPTOR (PEX5)	134	34	142	64	34	44	2
PTHR10138	TRYPTOPHAN 2,3-DIOXYGENASE	70	21	64	1	42	21	0
PTHR10146	PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN	136	29	29	0	29	0	0
PTHR10150	DNA REPAIR ENDONUCLEASE XPF	109	20	114	16	32	66	0
PTHR10159	DUAL SPECIFICITY PROTEIN PHOSPHATASE	1460	316	973	178	287	508	36
PTHR10169	DNA TOPOISOMERASE/GYRASE	334	77	457	102	69	286	27
PTHR10170	HUNTINGTON DISEASE PROTEIN	69	8	16	8	4	4	0
PTHR10194	RAS GTPASE-ACTIVATING PROTEINS	489	65	281	114	55	112	0
PTHR10202	PRESENILIN	124	26	383	262	27	94	1
PTHR10218	GTP-BINDING PROTEIN ALPHA SUBUNIT	860	185	1251	342	631	278	4
PTHR10224	ES1 PROTEIN HOMOLOG, MITOCHONDRIAL	40	1	1	0	0	1	0
PTHR10233	TRANSLATION INITIATION FACTOR EIF-2B	480	116	458	136	142	180	122
PTHR10250	MICROSOMAL GLUTATHIONE S-TRANSFERASE	186	67	316	110	147	59	13
PTHR10290	DNA TOPOISOMERASE I	139	15	25	6	4	15	4
PTHR10322	DNA POLYMERASE CATALYTIC SUBUNIT	429	70	470	115	106	249	68
PTHR10371	NADH DEHYDROGENASE [UBIQUINONE] FLAVOPROTEIN 2, MITOCHONDRIAL	143	42	116	39	40	37	1
PTHR10373	T-CELL-SPECIFIC TRANSCRIPTION FACTOR	155	29	429	25	116	288	0
PTHR10394	40S RIBOSOMAL PROTEIN S8	146	29	103	24	26	53	0
PTHR10442	40S RIBOSOMAL PROTEIN S21	119	25	129	21	21	87	0
PTHR10459	DNA LIGASE	345	75	406	136	63	207	16
PTHR10496	40S RIBOSOMAL PROTEIN S24	158	30	94	45	2	47	0
PTHR10516	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE	1136	214	848	240	383	225	2
PTHR10529	LOW-DENSITY LIPOPROTEIN RECEPTOR-RELATED	782	39	222	25	45	152	5
PTHR10559	TRANSCOBALAMIN-1/GASTRIC INTRINSIC FACTOR	78	10	18	0	8	10	0
PTHR10569	GLYCOGEN DEBRANCHING ENZYME	108	11	40	15	16	9	0
PTHR10663	GUANYL-NUCLEOTIDE EXCHANGE FACTOR	923	140	488	191	121	176	96
PTHR10668	PHYTOENE DEHYDROGENASE	376	40	74	29	19	26	0
PTHR10682	POLY(A) POLYMERASE	175	95	311	107	91	113	1
PTHR10732	40S RIBOSOMAL PROTEIN S17	159	32	85	29	1	55	0
PTHR10769	40S RIBOSOMAL PROTEIN S28	128	26	73	22	26	25	0
PTHR10795	PROPROTEIN CONVERTASE SUBTILISIN/KEXIN	1227	281	449	135	236	78	1
PTHR10845	REGULATOR OF G PROTEIN SIGNALING	871	150	551	290	148	113	0
PTHR10856	CORONIN	324	52	179	73	37	69	13
PTHR10871	30S RIBOSOMAL PROTEIN S13/40S RIBOSOMAL PROTEIN S18	212	65	256	104	51	101	9
PTHR10877	POLYCYSTIN-RELATED	604	398	1185	379	417	389	70
PTHR10878	SEGMENT POLARITY PROTEIN DISHEVELLED	168	34	209	53	22	134	3
PTHR10890	CYSTEINYL-TRNA SYNTHETASE	226	46	140	58	41	41	0
PTHR10927	RIBOSOME MATURATION PROTEIN SBDS	121	25	111	42	24	45	0
	TOTAL	17761	4412	17488	5616	6204	5668	587

Figure 1. Summary table of the types of annotation present in each gene family of the analysis.

PTHR ID	Gene Family Description	Total Nodes	Total Unique Annotation Sets	% of Unique Annotation Sets	At Least 1 Shared Annotation		No Shared Annotations		
					Pairs of Nodes	Average Distance	Pairs of Nodes	Average Distance	Difference
PTHR10000	PHOSPHOSERINE PHOSPHATASE	208	7	3.365384615	956	4.571638075	20572	4.219736875	0.3519012
PTHR10003	SUPEROXIDE DISMUTASE [CU-ZN]-RELATED	319	35	10.97178683	17327	3.370855659	33394	2.742057495	0.62879816
PTHR10005	SKI ONCOGENE-RELATED	152	16	10.52631579	268	3.003406716	11208	2.875069593	0.12833712
PTHR10006	MUCIN-1-RELATED	20	3	15	1	1.124	189	1.24921164	-0.12521164
PTHR10009	PROTEIN YELLOW-RELATED	61	11	18.03278689	537	3.629810056	1293	3.158932715	0.47087734
PTHR10012	SERINE/THREONINE-PROTEIN PHOSPHATASE 2A REGULATORY SUBUNIT B	142	11	7.746478873	2846	2.889373858	7165	2.23755492	0.65181894
PTHR10019	SNF5	97	16	16.49484536	1288	2.959091615	3368	2.434377078	0.52471454
PTHR10025	TETRAHYDROFOLATE DEHYDROGENASE/CYCLOHYDROLASE FAMILY MEMBER	326	19	5.828220859	3974	2.937868898	49001	2.942764637	-0.00489574
PTHR10027	CALCIUM-ACTIVATED POTASSIUM CHANNEL ALPHA CHAIN	206	11	5.339805825	447	3.561548098	20668	3.554066189	0.00748191
PTHR10032	ZINC FINGER PROTEIN WITH KRAB AND SCAN DOMAINS	697	31	4.447632712	61896	8.642831411	180660	8.131445206	0.5113862
PTHR10046	ATP DEPENDENT LON PROTEASE FAMILY MEMBER	283	21	7.4204947	2226	3.759056155	37677	3.35030711	0.40874905
PTHR10048	PHOSPHATIDYLINOSITOL KINASE	669	28	4.185351271	62436	5.216610385	161010	4.406688963	0.80992142
PTHR10063	TUBERIN	203	17	8.374384236	319	5.308705329	20184	4.92280871	0.38589662
PTHR10073	DNA MISMATCH REPAIR PROTEIN (MLH, PMS, MUTL)	410	25	6.097560976	2854	5.188041345	80991	4.80274654	0.38529481
PTHR10125	P2X PURINOCEPTOR	225	13	5.777777778	649	3.062964561	24551	2.632734634	0.43022993
PTHR10130	PEROXISOMAL TARGETING SIGNAL 1 RECEPTOR (PEX5)	134	5	3.731343284	528	2.753632576	8383	2.682073721	0.07155886
PTHR10138	TRYPTOPHAN 2,3-DIOXYGENASE	70	3	4.285714286	210	1.461028571	2205	1.410672562	0.05035601
PTHR10146	PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN	136	3	2.205882353	406	2.730453202	8774	2.246188056	0.48426515
PTHR10150	DNA REPAIR ENDONUCLEASE XPF	109	10	9.174311927	184	3.647673913	5702	3.144271133	0.50340278
PTHR10159	DUAL SPECIFICITY PROTEIN PHOSPHATASE	1460	87	5.95890411	38059	5.192499383	1027011	4.993835216	0.19866417
PTHR10169	DNA TOPOISOMERASE/GYRASE	334	21	6.28742515	2744	3.746473761	52867	3.768275181	-0.02180142
PTHR10170	HUNTINGTON DISEASE PROTEIN	69	9	13.04347826	19	1.596947368	2327	3.03136141	-1.43441404
PTHR10194	RAS GTPASE-ACTIVATING PROTEINS	489	16	3.27198364	2039	5.627784698	117277	5.45167865	0.17610605
PTHR10202	PRESENILIN	124	14	11.29032258	300	2.160966667	7326	1.904738329	0.25622834
PTHR10218	GTP-BINDING PROTEIN ALPHA SUBUNIT	860	82	9.534883721	16096	3.117376305	353274	2.952318263	0.16505804
PTHR10224	ES1 PROTEIN HOMOLOG, MITOCHONDRIAL	40	2	5	0	N/A	780	0.900638462	N/A
PTHR10233	TRANSLATION INITIATION FACTOR EIF-2B	480	15	3.125	4883	4.471405079	110077	4.266658584	0.2047465
PTHR10250	MICROSOMAL GLUTATHIONE S-TRANSFERASE	186	23	12.3655914	2064	3.819453973	15141	3.447479757	0.37197422
PTHR10290	DNA TOPOISOMERASE I	139	2	1.438848921	0	N/A	9591	1.910268168	N/A
PTHR10322	DNA POLYMERASE CATALYTIC SUBUNIT	429	26	6.060606061	2293	4.512476232	89513	4.414259672	0.09821656
PTHR10371	NADH DEHYDROGENASE [UBIQUINONE] FLAVOPROTEIN 2, MITOCHONDRIAL	143	5	3.496503497	816	2.833448529	9337	1.875329763	0.95811877
PTHR10373	T-CELL-SPECIFIC TRANSCRIPTION FACTOR	155	19	12.25806452	381	1.636732283	11554	2.01719344	-0.38046116
PTHR10394	40S RIBOSOMAL PROTEIN S8	146	7	4.794520548	401	1.348331671	10184	1.298155931	0.05017574
PTHR10442	40S RIBOSOMAL PROTEIN S21	119	9	7.56302521	294	1.442520408	6727	1.376968783	0.06555163
PTHR10459	DNA LIGASE	345	21	6.086956522	2606	5.309938219	56734	4.322866235	0.98707198
PTHR10496	40S RIBOSOMAL PROTEIN S24	158	12	7.594936709	373	1.475241287	12030	1.266154613	0.20908667
PTHR10516	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE	1136	33	2.904929577	22258	7.268916524	622422	7.374198324	-0.1052818
PTHR10529	LOW-DENSITY LIPOPROTEIN RECEPTOR-RELATED	782	24	3.069053708	362	6.749651934	305009	9.74411538	-2.99446345
PTHR10559	TRANSCOBALAMIN-1/GASTRIC INTRINSIC FACTOR	78	3	3.846153846	45	2.689911111	2958	2.714791751	-0.02488064
PTHR10569	GLYCOGEN DEBRANCHING ENZYME	108	6	5.555555556	49	1.603734694	5729	2.304497469	-0.70076278
PTHR10663	GUANYL-NUCLEOTIDE EXCHANGE FACTOR	923	35	3.791982665	9477	5.896578242	416026	5.566500512	0.33007773
PTHR10668	PHYTOENE DEHYDROGENASE	376	15	3.989361702	585	5.588805128	69915	6.055902868	-0.46709774
PTHR10682	POLY(A) POLYMERASE	175	11	6.285714286	4458	2.771760655	10767	2.163001486	0.60875917
PTHR10732	40S RIBOSOMAL PROTEIN S17	159	8	5.031446541	431	1.248786543	12130	1.052334955	0.19645159
PTHR10769	40S RIBOSOMAL PROTEIN S28	128	4	3.125	325	0.758495385	7803	0.666034602	0.09246078
PTHR10795	PROPROTEIN CONVERTASE SUBTILISIN/KEXIN	1227	41	3.341483293	31785	8.115946988	720366	8.574990354	-0.45904337
PTHR10845	REGULATOR OF G PROTEIN SIGNALING	871	22	2.525832377	10597	5.425885439	368288	5.129542809	0.29634263
PTHR10856	CORONIN	324	21	6.481481481	1177	3.589627867	51149	3.561149309	0.02847856
PTHR10871	30S RIBOSOMAL PROTEIN S13/40S RIBOSOMAL PROTEIN S18	212	17	8.018867925	1914	2.213934169	20452	2.163962449	0.04997172
PTHR10877	POLYCYSTIN-RELATED	604	21	3.476821192	77841	11.03156934	104265	9.90374608	1.12782326
PTHR10878	SEGMENT POLARITY PROTEIN DISHEVELLED	168	33	19.64285714	517	2.855524178	13511	3.112988084	-0.25746391
PTHR10890	CYSTEINYL-TRNA SYNTHETASE	226	7	3.097345133	1030	2.831790291	24395	2.524547571	0.30724272
PTHR10927	RIBOSOME MATURATION PROTEIN SBDS	121	8	6.611570248	296	2.286290541	6964	2.1582803	0.12801024
	TOTAL	17761	964	5.427622319					

Figure 2. Table indicating the possible lack of annotation quality in the gene families involved in the analysis. An average of 5% of the nodes have unique sets of annotations, and this can cause the phenomenon seen in the right hand columns – nodes with at least one shared annotation are further apart than those with none.

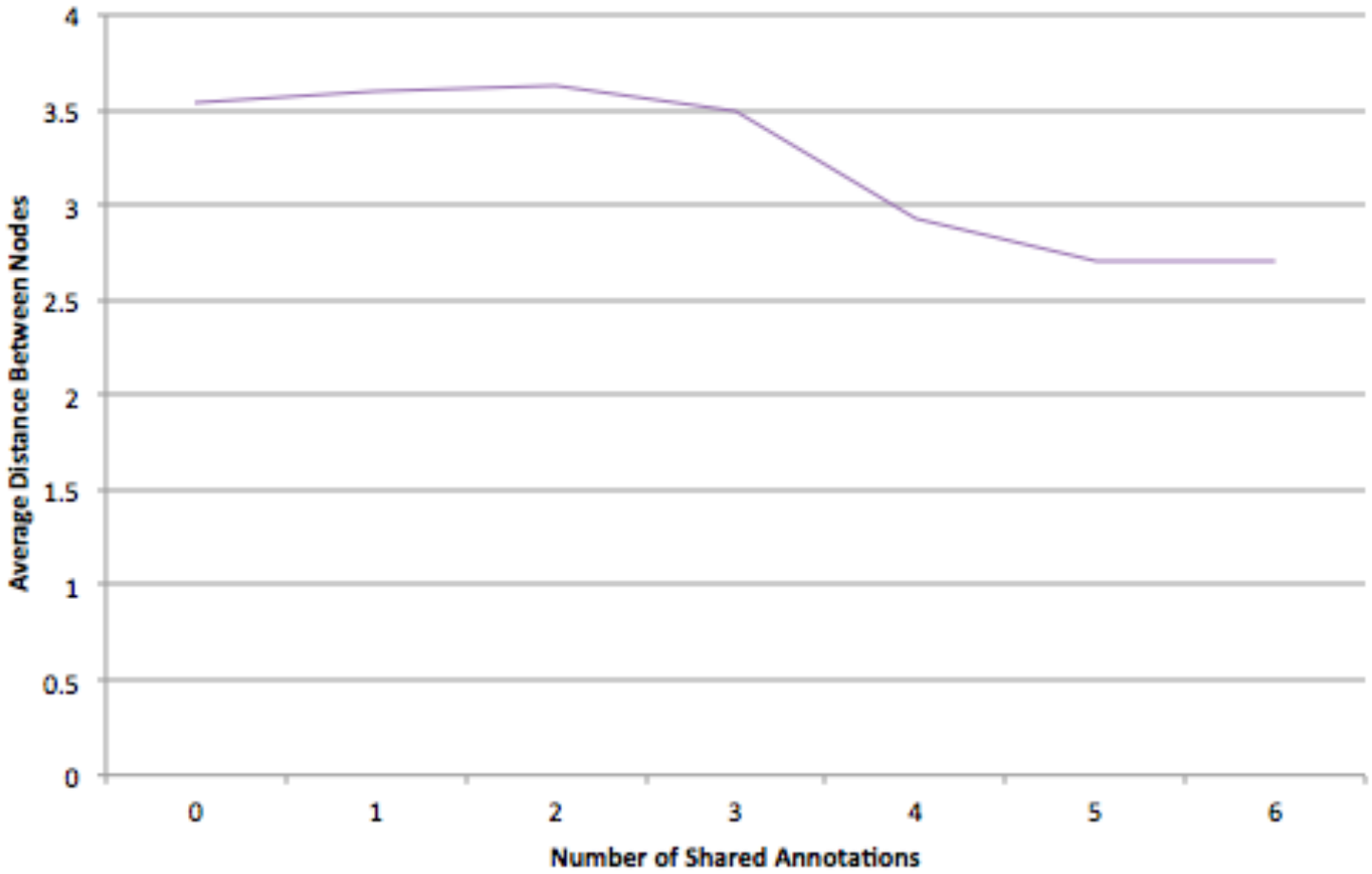

Figure 3. An extension of the problem seen in Figure 2, plotting the average distance between nodes as a function of the number of annotations that they share.