FAT-CAT

Supplementary Materials
Benchmarking Experiments: Ortholog Identification using Subtree HMMs

Motivation: Stage 2 of the FAT-CAT web server pipeline involves classifying a sequence to an orthology group using subtree HMM scoring. This document presents experiments in which we evaluated the accuracy of this classification protocol using leave-1-out experiments. FAT-CAT subtree-HMM classification is closely related to our previous work on subfamily identification (using SCI-PHY, an information-theoretic penalized likelihood function) and subfamily HMM-based functional classification (match state parameters estimated using an information-tying protocol including Dirichlet mixture densities); both methods are described in Brown et al, "Automated Protein Subfamily Identification and Classification," PLoS Computational Biology 2007, 3(8): e160 doi:10.1371/journal.pcbi.0030160. In Brown et al, we identified subfamilies using SCI-PHY, constructed subfamily HMMs for each subfamily, and then tested subfamily classification accuracy using leave-1-out experiments. We showed that assigning a sequence to the top-scoring subfamily HMM had near-perfect precision (error rate ~1%). The basic principles are the same in FAT-CAT, except that we construct HMMs at every node in every tree (including the leaves), instead of a strict subset of those nodes corresponding to a partition of the tree into subtrees.

Summary of experiments and results: We repeated a similar set of experiments to benchmark FAT-CAT orthology prediction. We took 83 families representing multi-domain architectures (i.e., sequences in the family aligned globally to each other) from PhyloFacts, ensuring that no pair of families had the same multi-domain architecture. For each tree, we defined orthology groups using a strict consensus of OMA, OrthoMCL (subtree bracketing protocol, as described in the Supplementary Materials: Subtree Bracketing) and Kerf (70% identity cutoff). The PhyloFacts multiple sequence alignment (MSA) for that family was retrieved; highly similar sequences were removed, a new MSA was constructed and masked, and a new tree was constructed using RAxML. For each family, we then identified on the RAxML tree those subtrees whose members were restricted to the largest orthology group defined on the corresponding PhyloFacts tree. PhyloFacts trees are estimated using FastTree, so that this strict consensus of two distinct phylogenetic tree methods, in addition to OMA and OrthoMCL support ensures high reliability of orthology. We then used leave-1-out experiments, removing a sequence from the orthology group, building the subtree HMMs without that sequence, and then scoring the withheld sequence against the subtree HMMs. In over 99% of cases, the top-scoring subtree was within the original orthology group from which the test sequence was drawn.

Dataset: We used a dataset of 83 families from the PhyloFacts database. Families were selected based on the following criteria:

1 all sequences in a family share the same multi-domain architecture (MDA) (requiring pairwise bi-directional overlap of >=70%, using the FlowerPower software);

2 no pair of families share the same MDA;
3 for each family, at least one pair of sequences has <40% identity;
4 each family contains at least two groups of orthologs based on OMA or OrthoMCL;
5 each family tree contains at least one subtree supported by three orthology methods (OMA, OrthoMCL and Kerf);
6 subtrees meeting criterion 5 also had to contain at least four sequences after the alignment was made non-redundant at 99% identity (i.e., removing sequences with >=99% identity).

Justification for these criteria:
Criterion 1, requiring that all sequences in a family share the same Pfam MDA, evaluates the most common type of phylogenetic tree used by phylogenomic methods. (Apart from PhyloFacts, which includes trees for Pfam domains, most other phylogenomic databases attempt to restrict trees to sequences that share the same multi-domain architecture.)
Criterion 2 ensures that the dataset is broad and unbiased.
Criteria 3 and 4 jointly ensure some level of functional divergence across the family, mirroring the sequence and functional divergence in PhyloFacts families in general.
Criterion 5 ensures that proposed orthologs have consensus support from multiple orthology methods. (This assertion is also strengthened by our construction of trees using FastTree and RAxML and requiring orthologs to be grouped together into subtrees by both methods.)
Criterion 6 ensures that a test sequence will have at least 3 remaining sequences in the subtree of orthologs once it is removed from the subtree.

For each family taken from PhyloFacts, we removed highly similar sequences (>=99% identity) from the multiple sequence alignment (MSA) to ensure that leave-1-out experiments would not be trivial (i.e., to disallow any test sequences from having a near-exact match among those remaining in the tree). We realigned these sequences using MAFFT, and masked the alignment to remove columns with >50% gaps. These masked alignments were then used to build maximum likelihood trees using RAxML (MPI version 7.0.4) with 100 bootstrap replicates.

Leave-1-Out Experiments
For each family, we selected sequences at random from the largest subtree meeting the consensus orthology definition described above, requiring in addition a bootstrap support of 75% or higher. A total of 302 sequences were drawn. For each test sequence, we removed the sequence from the masked MSA and also from the RAxML tree. HMMs were then constructed from the modified MSA and modified tree for each subtree (i.e., for all nodes in the tree, including the leaves). The test sequence was then scored against all the HMMs in the tree, and assigned to the top-scoring subtree. The precision of that subtree classification was measured as the fraction of sequences within that subtree defined as orthologs (by the criteria shown above) (i.e., the number of actual orthologs divided by the subtree size).

Results

Of the 302 test sequences, 301 are classified to the same orthology group from which they were drawn (i.e., an error rate of 0.3%, and 99.7% precision). See Supplementary Table 2 on the following pages.

Family accession	Family Size	Alignment Length	Number of sequences in PhyloFacts Orthology group	Number of sequences in Experiment Orthology group	# of test seqs	Average precision	test seq ids	Consensus PFam MDA	OMA group	Ortho MCL group	Min PWID (%)	Taxonomic distribution of orthology group
bpg0231463	51	260	12	9	8	1	P01210 P22005 P47969 P50175 P04094 B5FXU0 E1C652 D2H957	Opioids_neuropep (100.00)	448840	OG5_143556	70	Amniota
bpg0233854	65	814	17	8	5	1	A4UMC6 Q5ZII9 D2HKD8 A1XD95 A1XD97	TIP_N, G-patch, GCFC (100.00)	27372	OG5_129333	87	Amniota
bpg0237353	69	818	16	9	6	1	Q4VSV3 Q1LV08 Q0H909 D2GTW0 Q6NX12 Q7ZX96	Nic96 (100.00)	333246	OG5_129616	79	Euteleostomi
bpg0234973	74	584	17	11	8	1	E2RNE0 Q07832 Q2TA25 Q4SMJ6 P70032 D2HV98 P62205 Q4KMI8	Pkinase, POLO_box (100.00)	34676	OG5_127396	74	Euteleostomi
bpg0226567	83	264	18	10	4	1	Q7ZV80 C1BEV6 E1BQK7 C1BW46	SMN (100.00)	369406	OG5_130055	74	Euteleostomi
bpg0171974	90	363	31	10	4	1	D1TTE9 C4UYQ5 D5AZ13 C4SUL5	DPBB_1, SPOR (100.00)	72087	OG5_133523	79	Yersinia
bpg0228948	91	773	19	11	8	1	B1WBK8 Q8BU21 E2QRQ8 D2HEW0 Q3MHH4 Q6DRJ2 B4DWJ2 Q66H61	tRNA_synt_1c_R1, tRNA_synt_1c_R2, tRNA-synt_1c, tRNA-synt_1c_C (100.00)	1512	OG5_127471	73	Euteleostomi
bpg0234489	92	194	21	15	2	1	E1C6X9 Q00709	BH4, Bcl-2 (100.00)	446513	OG5_139665	75	Amniota
bpg0172362	93	502	31	9	6	1	C4U234 C4RXD8 C4SED4 Q1CC48 C4I0I9 A1JHR5	AAA_5, DUF3763 (100.00)	46492	OG5_136473	88	Yersinia
bpg0172884	94	360	32	9	7	1	B1ERE2 C1MG47 A8ACU5 D2TV32 D4BE28	Fuc4NAc_transf (100.00)	319228	OG5_164919	88	Yersinia
bpg0253075	100	342	20	17	9	1	B4KCY3 E2BA57 B0W8P2 E2AZW2 Q9VAY7 Q16U25 E3WZQ9 D6WTJ7 Q7PYQ5	XAP5 (95.00)	347036	OG5_129889	76	Neoptera
bpg0232945	105	382	25	16	11	1	P32120 Q3TRC8 Q6DFC4 E1BDU7 Q4SPT4 P29067 P51467 A4QNQ6 P51468 Q7T2D2 A5J090	Arrestin_N, Arrestin_C (100.00)	397545	OG5_131375	76	Euteleostomi
bpg0231870	116	479	14	12	6	1	Q6P8U6 Q02157 P00591 P50903 D4P6H2 P27657	Lipase, PLAT (100.00)	429252	OG5_130874	75	Eutheria
bpg0231839	117	245	22	9	4	1	B5X8E2 Q6NTV2 B9EP79 C1BJ56	PA28_alpha, PA28_beta (100.00)	392298	OG5_131194	77	Euteleostomi
bpg0171924	118	209	115	10	2	1	A4W6A3 B1EMT2	SMP_2 (100.00)	119367	OG5_166353	86	Enterobacteriaceae
bpg0230874	121	733	19	14	7	0.90	C4YPT6 C5M9J4 C5E400 C4QYX7 Q75E97 A5DFH4 Q6CEH2	DEAD, Helicase_C, HA2, DUF1605 (100.00)	28287	OG5_126599	72	Saccharomycetales (budding yeasts)
bpg0234919	123	239	19	11	3	1	Q06600 Q5TM20 P61125	TNF (100.00)	444613	OG5_140699	72	Eutheria
bpg0230174	124	127	22	16	7	1	P81184 P11116 Q5R7M1 D2GZI6 Q862W2 E2RJL1 P09382	Gal-bind_lectin (100.00)	162570	OG5_141953	73	Eutheria
bpg0245248	128	501	11	5	1	1	C5DZC1	WD40, UTP15_C (100.00)	47533	OG5_128309	72	Saccharomycetaceae
bpg0236403	136	135	19	15	2	1	Q5FW59 P05125	ANP (100.00)	446883	OG5_143715	74	Eutheria
bpg0229237	138	224	16	13	6	1	B4ILQ8 D3TR29 B4M223 B4L8E8 B3N084 B4NE11	Cwf_Cwc_15 (100.00)	118839	OG5_128660	70	Diptera
bpg0240616	140	118	20	10	4	1	C1BXG0 B5FYA5 B5X6P0 Q6DGQ0	Spt4 (100.00)	327035	OG5_128605	84	Euteleostomi
bpg0242866	142	242	24	14	5	1	Q6BPF4 A3LNK3 Q9P8P7 B9W9D9 A5DJ56	EMG1 (95.83)	327331	OG5_127748	76	Saccharomycetales (budding yeasts)
bpg0229848	143	403	11	6	1	1	Q755W8	eIF-5_eIF-2B, W2 (100.00)	324090	OG5_127779	73	Saccharomycetaceae
bpg0230155	144	787	26	16	7	1	D7NN74 C9WCD8 D7NN73 C6F5M4 Q5MAR3 Q2VJ42 B2ZAB0	Integrin_beta, Integrin_B_tail, Integrin_b_cyt (100.00)	29749	OG5_127959	80	Eutheria

Family accession	Family Size	Alignment Length	Number of sequences in PhyloFacts Orthology group	Number of sequences in Experiment Orthology group	# of test seqs	Average precision	test seq ids	Consensus PFam MDA	OMA group	Ortho MCL group	Min PWID (%)	Taxonomic distribution of orthology group
bpg0242080	144	528	21	10	4	1	Q641X3 Q4R6G5 B1PK15 B2LSM6	Glyco_hydro_20b , Glyco_hydro_20 (100.00)	345800	OG5_127470	79	Eutheria
bpg0242486	150	193	23	17	3	1	Q6C4V5 B9WHU7 Q05498	Fcf1 (100.00)	336239	OG5_127679	73	Saccharomycetales (budding yeasts)
bpg0172504	153	482	33	9	2	1	C4SBT6 C4U2R3	Cu-oxidase_3, Cu-oxidase, Cu-oxidase_2 (100.00)	8864	OG5_171492	90	Yersinia
bpg0247846	156	337	11	7	3	1	A1CKZ2 O60032 A1D6H4	Mo25 (95.65)	336916	OG5_128076	77	Trichocomaceae
bpg0241339	166	584	36	27	9	1	D4DEG9 A4RPZ8 A1CRA5 C1GTE2 C5P956 C6H1T0 B8M5V5 A7UVV2 A6SBD7	HEAT (100.00)	341855	OG5_127382	72	leotiomyceta
bpg0235628	180	512	38	20	7	1	Q1JPZ3 Q85477 Q91952 Q80XU2 E1BIM8 D2H117 Q64993	SH3_1, SH2, Pkinase_Tyr (100.00)	43992	OG5_127750	79	Euteleostomi
bpg0241031	189	137	14	9	1	1	Q5R9M4	Cornichon (100.00)	343310	OG5_127951	81	Euteleostomi
bpg0229628	196	469	18	5	3	1	B5A9S3 P02544 P48670	Filament_head, Filament (100.00)	54423	OG5_140657	95	Eutheria
bpg0228079	196	107	64	9	1	1	B5XNJ9	Fe-S_biosyn (100.00)	165890	OG5_127222	79	Enterobacteriaceae
bpg0173448	199	665	32	9	3	1	C4SMZ0 A1JL12 B2K989	DUF1726, DUF699 (100.00)	313688	OG5_127380	72	Yersinia
bpg0172781	212	169	206	39	8	1	C4U2R3 D4I9D8 B1EGU2 Q6D1T5 A7MJ28 C4U7H9 C6DCQ4 Q2PA28	LuxS (98.04)	286151	OG5_140532	75	Enterobacteriaceae
bpg0176597	214	1126	35	15	4	1	E2A9Y4 E1G0C9 Q10578 B4N9I3	RNA_pol_Rpb2_1, RNA_pol_Rpb2_2, RNA_pol_Rpb2_3, RNA_pol_Rpb2_4, RNA_pol_Rpb2_5, RNA_pol_Rpb2_6, RNA_pol_Rpb2_7 (100.00)	1803	OG5_126694	75	Metazoa
bpg0171897	222	706	206	33	4	1	Q6DB62 C8QC48 B2VL58 D0KE75	RelA_SpoT, TGS (98.39)	278187	OG5_128389	80	Enterobacteriaceae
bpg0242992	226	374	76	26	5	1	D0E0G1 Q5USW0 D1LWV2 O35312 Q9GM97	TGFb_propeptide , TGF_beta (98.00)	425873	OG5_132822	80	Amniota
bpg0232873	233	115	42	22	3	1	D4DLM5 C4JJZ5 C5P8Y8	Ribosomal_S26e (97.50)	325775	OG5_127138	78	Pezizomycotina
bpg0231973	236	585	26	11	1	1	Q16RL8	BTB, BACK, Kelch_1 (100.00)	36140	OG5_134192	73	Coelomata
bpg0253717	251	111	13	11	1	1	Q29F91	Fer2 (100.00)	313832	OG5_126994	79	Diptera
bpg0228754	256	499	97	18	3	1	D8AQD5 D6HX25 B4TPT5	ABC_tran (95.88)	45729	OG5_164923	74	Enterobacteriaceae
bpg0171131	261	152	128	6	2	1	C1M9P0 D4BBQ3	MGS (100.00)	270786	OG5_137551	93	Enterobacteriaceae
bpg0245496	264	208	24	15	3	1	C5M6R5 A7TDU2 B9WBL5	Ribosomal_S8e (100.00)	334502	OG5_127039	77	Saccharomycetales (budding yeasts)
bpg0171752	265	533	118	9	4	1	C1M8S6 B2P1U7 A9MKF2 D4BCJ2	HATPase_c (95.83)	41622	OG5_149682	86	Enterobacteriaceae
bpg0235117	265	122	20	13	2	1	B9W924 C4YD34	NTF2 (100.00)	329382	OG5_127878	79	Saccharomycetales (budding yeasts)
bpg0171659	266	257	32	9	1	1	A9QZZ8	DUF1212 (95.37)	86357	OG5_138621	92	Yersinia
bpg0252491	267	107	37	26	5	1	D3YWY6 Q5M994 C3KHA9 Q9Y5U8 Q6DD97	UPF0041 (97.37)	324544	OG5_128371	76	Euteleostomi
bpg0172022	268	392	30	8	2	1	C4SDB0 C4RWH1	Na_H_antiport_1 (100.00)	299120	OG5_139589	90	Yersinia
bpg0171680	272	115	21	14	1	1	Q9UR17	DUF202 (100.00)	340786	OG5_130969	73	Ascomycota
bpg0219574	286	369	134	6	1	1	C9XFD5	tRNA_Me_trans (100.00)	278095	OG5_127383	94	Enterobacteriaceae

Family accession	Family Size	Alignment Length	Number of sequences in PhyloFacts Orthology group	Number of sequences in Experiment Orthology group	# of test seqs	Average precision	test seq ids	Consensus PFam MDA	OMA group	Ortho MCL group	Min PWID (%)	Taxonomic distribution of orthology group
bpg0173710	286	344	133	12	2	1	D8AGU3 D4BH51	TruD (100.00)	314525	OG5_152568	86	Enterobacteriaceae
bpg0248265	288	256	46	35	7	1	D9D841 D1LYN4 Q1HRR3 B2ZSF2 D3TQN0 Q66UB9 B6DDU1	Ribosomal_S3Ae (97.83)	337207	OG5_126852	71	Pancrustacea
bpg0171219	295	468	31	7	2	1	C4SM15 A1JT25	SelA (100.00)	318647	OG5_140169	86	Yersinia
bpg0171311	297	211	35	8	1	1	Q1C7P6	OmpW (100.00)	308760	OG5_138797	84	Yersinia
bpg0237883	304	470	38	29	3	1	Q2GQU5 C6H759 E3QLP0	SHMT (100.00)	292912	OG5_126679	75	leotiomyceta
bpg0171757	305	272	120	5	3	1	B7LS88 B3BCA3 E1WIR1	MethyltransfD12 (100.00)	317485	OG5_138839	89	Enterobacteriaceae
bpg0236558	327	313	23	17	2	1	B5X5X5 Q8BK57	IF-2B (100.00)	67477	OG5_128035	78	Euteleostomi
bpg0170285	328	644	33	15	1	1	A7G9I0	tRNA-synt_1g, tRNA_bind (100.00)	32644	OG5_127468	75	Clostridium
bpg0170510	336	346	93	8	4	1	B5Q5K7 A8ACT6 A9MJ26 A8A6N9	Glycos_transf_4 (99.06)	58693	OG5_133040	87	Enterobacteriaceae
bpg0170616	355	199	200	56	11	1	C2CC86 B6EJ44 D1RW23 A8T586 B2Q4F9 D0ZED6 A8GAV1 C4UA86 C6DB85 D0IBU8 C5BD00	RecR (96.26)	354	OG5_132580	73	Gammaproteobacteria
bpg0235236	359	110	18	9	1	1	D4A531	RNA_POL_M_15 KD, TFIIS_C (100.00)	333374	OG5_128243	87	Euteleostomi
bpg0232030	361	460	25	17	3	1	Q6BU47 A3LY71 C8ZA85	Lyase_1 (100.00)	36975	OG5_127862	71	Saccharomycetales (budding yeasts)
bpg0196734	367	252	80	16	1	1	E3G3V1	LamB_YcsF (96.25)	6098	OG5_134793	70	Enterobacteriaceae
bpg0170307	373	175	91	6	1	1	Q57JF2	Hydrolase_3 (99.13)	305835	OG5_130703	93	Enterobacteriaceae
bpg0199565	403	166	128	20	2	1	C4UBN6 A7MEH5	MarR (95.31)	9786	OG5_140512	76	Enterobacteriaceae
bpg0185688	404	189	131	16	1	1	A8AIS1	TetR_N (96.95)	128177	OG5_178460	71	Enterobacteriaceae
bpg0170731	406	703	233	71	4	1	D2TX27 D0YZW6 C9XHK0 C5BFC1	RNase_PH, RNase_PH_C, PNPase, RNase_PH, RNase_PH_C, KH_1, S1 (95.08)	263564	OG5_128725	71	Gammaproteobacteria
bpg0170559	408	210	120	10	2	1	B0H7L2 C4U8X5	Ribosomal_L3 (98.72)	1981	OG5_127133	91	Enterobacteriaceae
bpg0170450	410	211	98	8	1	1	C1MCP3	Pro_CA (99.09)	300486	OG5_127503	84	Enterobacteriaceae
bpg0225600	417	196	80	23	3	1	A9K0Q7 Q2SUA4 D8NX85	IGPD (95.65)	290561	OG5_129258	78	Burkholderiaceae
bpg0173567	429	260	112	22	1	1	B6VMT1	FAA_hydrolase (98.21)	301424	OG5_144783	75	Enterobacteriaceae
bpg0170569	442	226	131	44	4	1	D0KJV0 A9MGV9 D1S0W6 C8SYZ0	UDG (99.25)	295069	OG5_127417	70	Enterobacteriaceae
bpg0170707	467	344	145	30	4	1	C6DB25 B2VHQ4 D2T4P0 C9XX68	Queuosine_synth (96.97)	278518	OG5_132916	80	Enterobacteriaceae
bpg0202675	492	305	17	9	1	1	Q646B5	TAS2R (100.00)	92375	OG5_157453	72	Eutheria
bpg0170689	494	191	48	21	3	1	B3R701 Q8XVE9 C6BEL8	GATase (100.00)	126746	OG5_207698	95	Yersinia
bpg0170589	502	277	95	11	2	1	B7LVX0 B4A3J3	QRPTase_N, QRPTase_C (99.08)	279036	OG5_128545	85	Enterobacteriaceae
bpg0184790	503	554	62	25	4	1	B2T150 D8P050 D5W4I8 B5WTN0	ILVD_EDD (96.30)	268080	OG5_128618	84	Burkholderiaceae
bpg0171061	529	401	54	9	2	1	A6WV13 A6UE09	Arginosuc_synth (98.95)	282338	OG5_127707	92	Rhizobiales
bpg0184738	603	300	33	9	1	1	C1M9W9	TruB_N, TruB-C_2 (98.31)	263484	OG5_126700	88	Yersinia
bpg0226429	819	127	99	16	1	1	D8NP90	Glyoxalase (98.99)	295435	OG5_127627	74	Proteobacteria
bpg0184601	380	633	63	26	5	1	Q39E36 B2U8Q7 B3R5J8 Q0K8D7 D8NW45	HSP90 (87.47)	297336	OG5_126623	71	Burkholderiaceae