Construction of a Phylogenetic Tree Based on the Average Amino Acid Composition of Exomes

Esumi, Genshiro

Department of Pediatric Surgery, Hospital of the University of Occupational and Environmental Health, Kitakyushu, Japan

Abstract

All extant cellular organisms are believed to have descended from a common ancestor, and numerous phylogenetic trees have been constructed using various genetic and molecular data. In this study, I hypothesized that the average amino acid composition of an exome—computed across all exons—could serve as an index reflecting an organism's characteristic amino acid usage and could be used to construct a phylogenetic tree. To test this hypothesis, I analyzed publicly available exome data from 81 species. For each species, I counted the fractional composition of each amino acid in each exon, and then calculated the average amino acid composition. I measured the pairwise distances between species using the angular distance based on these average compositions. Hierarchical clustering with Ward's method was applied to construct a phylogenetic tree. The resulting tree showed a reasonable degree of similarity to previously established phylogenies, suggesting that this exome-based amino acid composition approach may offer some utility for inferring evolutionary relationships. To my knowledge, this is the first demonstration of constructing a phylogenetic tree solely from average exome-wide amino acid composition.

Keywords: Phylogenetic tree, Amino acid composition, Exome, Distance function, Evolution

Email: esumi@clnc.uoeh-u.ac.jp

Background

Throughout the history of evolutionary biology, numerous phylogenetic trees have been constructed under the assumption that all organisms are derived from a common ancestor, and that evolutionary change proceeds at a pace slow enough to be traced over time. Early efforts to infer evolutionary relationships focused largely on morphological phenotypes. However, with advancements in molecular biology, current approaches increasingly rely on evaluating changes in protein amino acid sequences, genes, and entire genome sequences [1].

Thanks to the completion of whole-genome sequencing projects for many species, the exomes of these organisms have become well-characterized. In this study, I hypothesize that the average amino acid composition of each exon within an exome can serve as a set of indices reflecting the organism's amino acid usage. I further assume that these indices have gradually diverged from a common ancestor. Under this assumption, I investigate whether a phylogenetic tree with at least some degree of validity can be reconstructed solely by calculating pairwise distance matrices of these average amino acid compositions across species.

Materials and Methods: Phylogenetic Analysis Based on Average Amino Acid Compositions of Exomes

Exome sequence data from 81 species listed in a publicly available dataset named "reference proteomes" on the EMBL-EBI website were used to calculate the average amino acid composition for each exome [2]. Briefly, all exon-derived amino acid sequences within the exome of each species were retrieved from this dataset. For each exon, the number of each amino acid was counted, and these counts were then summed exon by exon. Subsequently, to obtain the average amino acid composition of each exome, the total counts of each amino acid were normalized so that the sum of all fractions equals 1, by dividing by the total number of residues in each exon, and then averaged across all exons.

Next, an 81 × 81 distance matrix was constructed by computing the angular distance between each pair of species, based on their respective average amino acid compositions. The angular distance is defined as:

$$d(x, y) = \arccos\left(\frac{x \cdot y}{\|x\|\|y\|}\right).$$

This distance matrix was subsequently used to perform hierarchical clustering with Ward's method using JMP Pro 18 software (SAS Institute Inc., Cary, NC, USA). The resulting dendrogram was interpreted as a phylogenetic tree, illustrating the relationships among the 81 species based on their exome-wide amino acid composition patterns.

Results

Table 1 provides the scientific names of the 81 species examined in this study, the short labels used for the phylogenetic trees, and the number of annotated proteins registered for each species. Table 2 presents the average exome-wide composition of the 20 standard amino acids in each species' proteins. I generated a horizontal dendrogram based on these amino acid compositions and a radial dendrogram from the same dataset, shown in Figure 1 and Figure 2, respectively. In the figures, archaea are depicted in blue, bacteria in red, and eukaryotes in green. In addition, within the radial dendrogram (Figure 2), unicellular organisms are represented by small dots, whereas multicellular organisms are highlighted with larger dots and bold labels.

As a result of these procedures, a phylogenetic tree that appears reasonably valid at first glance was obtained. The validity of this tree will be evaluated in the subsequent Discussion section.

Discussion

In this study, I investigated whether it is possible to infer—or more precisely, restore—a phylogenetic tree based solely on the average amino acid composition of organismal exomes. Although my analysis was limited to 81 species, the resulting tree showed that groups of closely related organisms, as previously understood from existing phylogenies [3], largely clustered together. For example, mammals formed a distinct cluster, with primates (human, monkey, chimpanzee, and gorilla) and rodents (mouse and rat) each grouping on their own respective branches. These observations suggest that exome amino acid composition—at least in terms of its average values—shifts slowly enough to preserve certain phylogenetic relationships, allowing for the construction of such a tree (Figure 1, 2).

On the other hand, I also observed instances where organisms from entirely different lineages appeared on adjacent branches. One likely explanation is that while traditional phylogenetic reconstructions rely on highly multidimensional data (e.g., morphological or genetic diversity), the amino acid composition of exomes is comparatively low-dimensional. This may increase the chance

of coincidental similarities or convergence under evolutionary pressure. As a result, the overall validity or information content of the exome-based tree is not sufficient to replace or override existing phylogenies.

A further question arises as to the significance of an exome-based phylogenetic tree. Since the average amino acid composition does not account for protein abundance, it has often been presumed not to directly reflect the amino acid composition that an organism actually utilizes. However, my earlier work found that exome-wide amino acid composition tends to follow an approximately binomial (bell-shaped) distribution [4]. If the exome indeed exhibits such a distribution, then the proteome—composed of proteins synthesized from these exons—may be constrained by the exome's amino acid composition. In the same report, I also proposed that this bell-shaped pattern could itself be influenced by constraints imposed by the proteome [4]. Hence, the exome and proteome likely exist in a mutually constraining relationship that gives rise to the exome's bell-shaped distribution. If so, the phylogenetic tree constructed here may reflect how each organism's cytoplasmic or cellular amino acid composition has evolved into its present form—that is, the evolutionary trajectory as seen from the perspective of amino acid composition.

An intriguing observation is that maize and a pathogenic fungus infecting maize lie on adjacent branches. This could indicate that the fungus converged toward maize's amino acid composition, its nutritional source. In other words, it could be said that the exome of the pathogenic fungus may have adapted to the proteome of maize—a form of convergent evolution—although further study is required to clarify this possibility.

It has already been pointed out that an organism's genome and GC content can influence its exome's amino acid composition [5]. Even so, my results demonstrate that exome-level amino acid composition not only exhibits such variation, but also—albeit partially—contains the information needed to restore certain aspects of a phylogenetic tree. Moreover, since I found that each organism's exome tends toward a bell-shaped distribution—implying convergence—exomes will continue to shift while maintaining a reciprocal constraint with their proteomes under evolutionary pressures [4]. Even if this concept appears straightforward once the underlying premises are considered, to my knowledge it has not been explicitly emphasized before, and thus may represent a novel insight. In this light, I hope this study provides a small yet meaningful contribution to understanding life's complex evolutionary tapestry.

Reference

1. Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics,* 21(7), 428–444. https://doi.org/10.1038/s41576-020-0233-0

2. EMBL-EBI. (2024). *Reference Proteomes* (Release 2024_02). Retrieved January 7, 2025, from https://www.ebi.ac.uk/reference_proteomes/

3. Siepel, A. (2009). Phylogenomics of primates and their ancestral populations. *Genome Research*, 19(11), 1929–1941. https://doi.org/10.1101/gr.084228.108

4. Esumi, G. (2023). The Distributions of Amino Acid Compositions of Proteins in an Organism's Proteome Uniformly Approximate Binomial Distributions. *Jxiv.* https://doi.org/10.51094/jxiv.408

5. Du, M.-Z., Liu, S., Zeng, Z., Alemayehu, L. A., Wei, W., & Guo, F.-B. (2018). Amino acid compositions contribute to the proteins' evolution under the influence of their abundances and genomic GC content. *Scientific Reports*, 8(1). https://doi.org/10.1038/s41598-018-25364-1

No.	Scientific Name	Short Label	Domain	Cell Organization	Protein Count
1	*Halobacterium salinarum* (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium)	Halobacterium	archaea	unicellular	2427
2	*Thermococcus kodakarensis* (strain ATCC BAA-918 / JCM 12380 / KOD1) (Pyrococcus kodakaraensis (strain KOD1))	Thermococcus	archaea	unicellular	2301
3	*Methanosarcina acetivorans* (strain ATCC 35395 / DSM 2834 / JCM 12185 / C2A)	Methanosarcina	archaea	unicellular	4468
4	*Methanocaldococcus jannaschii* (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440) (Methanococcus jannaschii)	Methanocaldococcus	archaea	unicellular	1787
5	*Saccharolobus solfataricus* (strain ATCC 35092 / DSM 1617 / JCM 11322 / P2) (Sulfolobus solfataricus)	Saccharolobus	archaea	unicellular	2937
6	*Korarchaeum cryptofilum* (strain OPF8)	Korarchaeum	archaea	unicellular	1602
7	*Nitrosopumilus maritimus* (strain SCM1)	Nitrosopumilus	archaea	unicellular	1795
8	*Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv)	M. tuberculosis	bacteria	unicellular	3999
9	*Escherichia coli* (strain K12)	E. coli	bacteria	unicellular	4416
10	*Helicobacter pylori* (strain ATCC 700392 / 26695) (Campylobacter pylori)	Helicobacter	bacteria	unicellular	1554
11	*Streptomyces coelicolor* (strain ATCC BAA-471 / A3(2) / M145)	Streptomyces	bacteria	unicellular	8039
12	*Neisseria meningitidis* serogroup B (strain MC58)	Meningococcus	bacteria	unicellular	2001
13	*Leptospira interrogans* serogroup Icterohaemorrhagiae serovar Lai (strain 56601)	Leptospira	bacteria	unicellular	3676
14	*Fusobacterium nucleatum* subsp. nucleatum (strain ATCC 25586 / DSM 15643 / BCRC 10681 / CIP 101130 / JCM 8532 / KCTC 2640 / LMG 13131 / VPI 4355)	Fusobacterium	bacteria	unicellular	2046
15	*Pseudomonas aeruginosa* (strain ATCC 15692 / DSM 22644 / CIP 104116 / JCM 14847 / LMG 12228 / 1C / PRS 101 / PAO1)	Pseudomonas	bacteria	unicellular	5564
16	*Bacillus subtilis* (strain 168)	B. subtilis	bacteria	unicellular	4267
17	*Aquifex aeolicus* (strain VF5)	Aquifex	bacteria	unicellular	1553
18	*Bradyrhizobium diazoefficiens* (strain JCM 10833 / BCRC 13528 / IAM 13628 / NBRC 14792 / USDA 110)	Bradyrhizobium	bacteria	unicellular	8253
19	*Bacteroides thetaiotaomicron* (strain ATCC 29148 / DSM 2079 / JCM 5827 / CCUG 10774 / NCTC 10582 / VPI-5482 / E50)	Bacteroides	bacteria	unicellular	4782
20	*Rhodopirellula baltica* (strain DSM 10527 / NCIMB 13988 / SH1)	Rhodopirellula	bacteria	unicellular	7271
21	*Deinococcus radiodurans* (strain ATCC 13939 / DSM 20539 / JCM 16871 / CCUG 27074 / LMG 4051 / NBRC 15346 / NCIMB 9279 / VKM B-1422 / R1)	Deinococcus	bacteria	unicellular	3084
22	*Geobacter sulfurreducens* (strain ATCC 51573 / DSM 12127 / PCA)	Geobacter	bacteria	unicellular	3402
23	*Mycoplasma genitalium* (strain ATCC 33530 / DSM 19775 / NCTC 10195 / G37) (Mycoplasmoides genitalium)	Mycoplasma	bacteria	unicellular	483
24	*Thermotoga maritima* (strain ATCC 43589 / DSM 3109 / JCM 10099 / NBRC 100826 / MSB8)	Thermotoga	bacteria	unicellular	1852
25	*Gloeobacter violaceus* (strain ATCC 29082 / PCC 7421)	Gloeobacter	bacteria	unicellular	4406
26	*Chlamydia trachomatis* (strain D/UW-3/Cx)	Chlamydia	bacteria	unicellular	895
27	*Thermodesulfovibrio yellowstonii* (strain ATCC 51303 / DSM 11347 / YP87)	Thermodesulfovibrio	bacteria	unicellular	1982
28	*Chloroflexus aurantiacus* (strain ATCC 29366 / DSM 635 / J-10-fl)	Chloroflexus	bacteria	unicellular	3850
29	*Dictyoglomus turgidum* (strain DSM 6724 / Z-1310)	Dictyoglomus	bacteria	unicellular	1743
30	*Synechocystis sp.* (strain PCC 6803 / Kazusa)	Synechocystis	bacteria	unicellular	3508
31	*Chlamydomonas reinhardtii* (Chlamydomonas smithii)	Chlamydomonas	eukaryota	unicellular	18832
32	*Physcomitrium patens* (Spreading-leaved earth moss) (Physcomitrella patens)	Physcomitrella	eukaryota	**multicellular**	47782
33	*Arabidopsis thaliana* (Mouse-ear cress)	Arabidopsis	eukaryota	**multicellular**	41596
34	*Zea mays* (Maize)	Maize	eukaryota	**multicellular**	63281
35	*Leishmania major*	Leishmania	eukaryota	unicellular	8038
36	*Paramecium tetraurelia*	Paramecium	eukaryota	unicellular	39461
37	*Caenorhabditis elegans*	C. elegans	eukaryota	**multicellular**	28553
38	*Helobdella robusta* (Californian leech)	Leech	eukaryota	**multicellular**	23328
39	*Ixodes scapularis* (Black-legged tick) (Deer tick)	Tick	eukaryota	**multicellular**	20496
40	*Tribolium castaneum* (Red flour beetle)	Tribolium	eukaryota	**multicellular**	18505
41	*Anopheles gambiae* (African malaria mosquito)	Mosquito	eukaryota	**multicellular**	14411
42	*Drosophila melanogaster* (Fruit fly)	Fruit fly	eukaryota	**multicellular**	23539
43	*Ciona intestinalis* (Transparent sea squirt) (Ascidia intestinalis)	Sea squirt	eukaryota	**multicellular**	17311
44	*Branchiostoma floridae* (Florida lancelet) (Amphioxus)	Amphioxus	eukaryota	**multicellular**	38648
45	*Lepisosteus oculatus* (Spotted gar)	Spotted gar	eukaryota	**multicellular**	22463
46	*Danio rerio* (Zebrafish) (Brachydanio rerio)	Zebrafish	eukaryota	**multicellular**	46840
47	*Oryzias latipes* (Japanese rice fish) (Japanese killifish)	Medaka	eukaryota	**multicellular**	36138
48	*Xenopus laevis* (African clawed frog)	X. laevis	eukaryota	**multicellular**	61769
49	*Xenopus tropicalis* (Western clawed frog) (Silurana tropicalis)	X. tropicalis	eukaryota	**multicellular**	37693
50	*Gallus gallus* (Chicken)	Chicken	eukaryota	**multicellular**	43968
51	*Macaca mulatta* (Rhesus macaque)	Rhesus	eukaryota	**multicellular**	44416
52	*Gorilla gorilla gorilla* (Western lowland gorilla)	Gorilla	eukaryota	**multicellular**	44726
53	*Pan troglodytes* (Chimpanzee)	Chimpanzee	eukaryota	**multicellular**	48794
54	*Homo sapiens* (Human)	Human	eukaryota	**multicellular**	104573
55	*Canis lupus familiaris* (Dog) (Canis familiaris)	Dog	eukaryota	**multicellular**	43672
56	*Bos taurus* (Bovine)	Cow	eukaryota	**multicellular**	37871
57	*Mus musculus* (Mouse)	Mouse	eukaryota	**multicellular**	63289
58	*Rattus norvegicus* (Rat)	Rat	eukaryota	**multicellular**	49582
59	*Monodelphis domestica* (Gray short-tailed opossum)	Opossum	eukaryota	**multicellular**	36221
60	*Thalassiosira pseudonana* (Marine diatom) (Cyclotella nana)	Thalassiosira	eukaryota	unicellular	11612
61	*Daphnia magna*	Daphnia	eukaryota	**multicellular**	26600
62	*Plasmodium falciparum* (isolate 3D7)	Plasmodium	eukaryota	unicellular	5369
63	*Oryza sativa* subsp. japonica (Rice)	Rice	eukaryota	**multicellular**	49224
64	*Dictyostelium discoideum* (Social amoeba)	Dictyostelium	eukaryota	unicellular	12746
65	*Nematostella vectensis* (Starlet sea anemone)	Sea anemone	eukaryota	**multicellular**	24445
66	*Monosiga brevicollis* (Choanoflagellate)	Monosiga	eukaryota	unicellular	9156
67	*Phytophthora ramorum* (Sudden oak death agent)	Phytophthora	eukaryota	unicellular	15349
68	*Giardia intestinalis* (strain ATCC 50803 / WB clone C6) (Giardia lamblia)	Giardia	eukaryota	unicellular	4900
69	*Cryptococcus neoformans* var. neoformans serotype D (strain JEC21 / ATCC MYA-565) (Filobasidiella neoformans)	Cryptococcus	eukaryota	unicellular	6746
70	*Candida albicans* (strain SC5314 / ATCC MYA-2876) (Yeast)	Candida	eukaryota	unicellular	6037
71	*Ustilago maydis* (strain 521 / FGSC 9021) (Corn smut fungus)	Corn smut	eukaryota	unicellular	6805
72	*Yarrowia lipolytica* (strain CLIB 122 / E 150) (Yeast) (Candida lipolytica)	Yarrowia	eukaryota	unicellular	6454
73	*Schizosaccharomyces pombe* (strain 972 / ATCC 24843) (Fission yeast)	S. pombe	eukaryota	unicellular	5132
74	*Phaeosphaeria nodorum* (strain SN15 / ATCC MYA-4574 / FGSC 10173) (Glume blotch fungus) (Parastagonospora nodorum)	Phaeosphaeria	eukaryota	unicellular	15998
75	*Aspergillus fumigatus* (strain ATCC MYA-4609 / CBS 101355 / FGSC A1100 / Af293) (Neosartorya fumigata)	Aspergillus	eukaryota	unicellular	9648
76	*Neurospora crassa* (strain ATCC 24698 / 74-OR23-1A / CBS 708.71 / DSM 1257 / FGSC 987)	Neurospora	eukaryota	unicellular	10266
77	*Trichomonas vaginalis* (strain ATCC PRA-98 / G3)	Trichomonas	eukaryota	unicellular	50190
78	*Puccinia graminis* f. sp. *tritici* (strain CRL 75-36-700-3 / race SCCL) (Black stem rust fungus)	Stem rust	eukaryota	unicellular	15808
79	*Saccharomyces cerevisiae* (strain ATCC 204508 / S288c) (Baker's yeast)	S. cerevisiae	eukaryota	unicellular	6091
80	*Sclerotinia sclerotiorum* (strain ATCC 18683 / 1980 / Ss-1) (White mold) (Whetzelinia sclerotiorum)	Sclerotinia	eukaryota	unicellular	14445
81	*Batrachochytrium dendrobatidis* (strain JAM81 / FGSC 10211) (Frog chytrid fungus)	Batrachochytrium	eukaryota	unicellular	8610
				Total	1547370

Table 1. Species Information, Short Labels, and Protein Counts

This table lists the scientific names of the 81 species analyzed in this study, along with their corresponding short labels (used for phylogenetic tree visualization), domain classification (archaea, bacteria, eukaryotes), cell organization (unicellular or multicellular), and the total number of annotated proteins ("Protein Count") obtained from publicly available datasets. The color codes in the "Domain" column correspond to blue for archaea, red for bacteria, and green for eukaryotes.

No.	Short Label	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
1	Halobacterium	0.12407	0.00889	0.0888	0.06979	0.03058	0.08201	0.02265	0.03687	0.01821	0.08469	0.0191	0.02153	0.0464	0.02691	0.06561	0.05464	0.06794	0.0949	0.01126	0.02514
2	Thermococcus	0.07304	0.00605	0.04609	0.09242	0.04258	0.07421	0.01543	0.07038	0.0718	0.10734	0.02497	0.02987	0.04216	0.01729	0.06113	0.04855	0.04367	0.08396	0.01264	0.03645
3	Methanosarcina	0.06629	0.01453	0.04976	0.07941	0.04731	0.06994	0.01726	0.07562	0.07015	0.09589	0.02711	0.04246	0.03945	0.02558	0.04721	0.06736	0.05082	0.068	0.01016	0.03571
4	Methanocaldococcus	0.05447	0.01407	0.05329	0.08678	0.04249	0.06184	0.01392	0.1071	0.10671	0.09512	0.02504	0.05102	0.0327	0.01459	0.03919	0.04432	0.03922	0.06867	0.00701	0.04246
5	Saccharolobus	0.05343	0.00756	0.04685	0.07053	0.0438	0.06169	0.01298	0.09485	0.08221	0.10399	0.02327	0.04882	0.03589	0.02106	0.04918	0.06635	0.04584	0.07477	0.00995	0.04701
6	Korarchaeum	0.06846	0.00938	0.04855	0.08472	0.03745	0.07669	0.01348	0.08305	0.05775	0.10988	0.02782	0.02675	0.04342	0.01481	0.07305	0.07177	0.03527	0.07344	0.01108	0.03318
7	Nitrosopumilus	0.05963	0.01131	0.05955	0.07541	0.04374	0.06223	0.01819	0.08494	0.09008	0.08403	0.02842	0.04776	0.03691	0.03234	0.03552	0.06945	0.05444	0.06666	0.00898	0.03041
8	M. tuberculosis	0.13255	0.01007	0.05865	0.04819	0.02836	0.09087	0.02317	0.04236	0.0218	0.09669	0.02099	0.02323	0.05782	0.03129	0.07804	0.05515	0.05916	0.08598	0.01504	0.02057
9	E. coli	0.09305	0.01318	0.05002	0.0576	0.03944	0.07025	0.02335	0.0617	0.04742	0.10621	0.03054	0.03902	0.04268	0.04376	0.05607	0.05831	0.0534	0.07084	0.01514	0.028
10	Helicobacter	0.06687	0.01225	0.04646	0.06903	0.05481	0.05555	0.02189	0.07188	0.09247	0.11448	0.02528	0.05466	0.03218	0.03653	0.0363	0.06709	0.0417	0.05719	0.00745	0.03593
11	Streptomyces	0.13599	0.00884	0.06061	0.05768	0.02622	0.09479	0.0239	0.02903	0.02075	0.10077	0.0185	0.01645	0.06188	0.02648	0.08583	0.05005	0.06105	0.08621	0.01524	0.01972
12	Meningococcus	0.09655	0.01324	0.05037	0.06095	0.04411	0.07361	0.02219	0.05997	0.05986	0.09822	0.02674	0.04048	0.04248	0.03915	0.05665	0.05563	0.05087	0.06652	0.01217	0.03026
13	Leptospira	0.05031	0.01006	0.04632	0.07204	0.0582	0.06183	0.01719	0.08154	0.08268	0.10395	0.02057	0.05026	0.03772	0.03269	0.04483	0.07716	0.0494	0.05664	0.01145	0.03517
14	Fusobacterium	0.05234	0.00858	0.0527	0.0799	0.05173	0.05921	0.01188	0.10135	0.1049	0.09525	0.02538	0.06217	0.02491	0.02151	0.03207	0.05801	0.04602	0.06088	0.00634	0.04486
15	Pseudomonas	0.11635	0.0114	0.05248	0.06156	0.03557	0.08272	0.02222	0.04153	0.02988	0.12462	0.02175	0.02523	0.05084	0.04232	0.07702	0.05463	0.04113	0.06885	0.01504	0.02486
16	B. subtilis	0.07337	0.00897	0.05087	0.07343	0.0458	0.06647	0.02285	0.07454	0.07458	0.09641	0.03031	0.04034	0.03454	0.03849	0.04144	0.06179	0.05295	0.0672	0.01024	0.03541
17	Aquifex	0.05813	0.00893	0.04224	0.09674	0.05124	0.06659	0.01555	0.07351	0.09662	0.10561	0.0205	0.0356	0.04013	0.02066	0.0495	0.04767	0.04139	0.07947	0.00937	0.04055
18	Bradyrhizobium	0.12369	0.01047	0.05365	0.05283	0.03696	0.08185	0.02124	0.05213	0.0374	0.09763	0.02633	0.02647	0.05292	0.0316	0.07496	0.05847	0.05299	0.07343	0.01345	0.02154
19	Bacteroides	0.06771	0.0139	0.05323	0.06664	0.04671	0.06457	0.01886	0.07206	0.06924	0.0925	0.02863	0.04971	0.03669	0.03432	0.04655	0.06205	0.05549	0.06357	0.01271	0.04485
20	Rhodopirellula	0.08859	0.01789	0.05437	0.05502	0.03876	0.07142	0.02544	0.04779	0.03579	0.09128	0.02706	0.03306	0.05542	0.03938	0.07983	0.0791	0.05724	0.06811	0.01593	0.01851
21	Deinococcus	0.12104	0.00765	0.0497	0.05762	0.03138	0.09042	0.02157	0.03185	0.02746	0.11759	0.02005	0.02303	0.06085	0.04118	0.07576	0.05194	0.05744	0.07643	0.01467	0.02237
22	Geobacter	0.10093	0.01443	0.05256	0.06639	0.03912	0.08291	0.02089	0.05658	0.04475	0.10136	0.02614	0.02788	0.0467	0.0273	0.07124	0.05314	0.05342	0.0785	0.0102	0.02556
23	Mycoplasma	0.05619	0.00984	0.04591	0.05393	0.06123	0.04645	0.01697	0.08423	0.10079	0.10697	0.01714	0.07179	0.02895	0.04575	0.03396	0.06381	0.05272	0.06213	0.00941	0.03183
24	Thermotoga	0.05742	0.00847	0.04814	0.08921	0.05199	0.06882	0.01575	0.07238	0.07894	0.10056	0.02553	0.03494	0.03891	0.01979	0.05662	0.05611	0.04436	0.0866	0.01079	0.03467
25	Gloeobacter	0.10994	0.01139	0.04863	0.06323	0.03664	0.08016	0.01917	0.04493	0.0308	0.11385	0.0197	0.02606	0.05579	0.04251	0.0758	0.0543	0.05052	0.07519	0.01542	0.02599
26	Chlamydia	0.07537	0.01766	0.04345	0.0657	0.04805	0.06224	0.02262	0.06605	0.06105	0.11303	0.02206	0.03348	0.04284	0.04167	0.05089	0.07965	0.04938	0.06615	0.00928	0.0294
27	Thermodesulfovibrio	0.05953	0.01214	0.04616	0.08053	0.0499	0.06168	0.01614	0.10021	0.09635	0.0966	0.02422	0.04223	0.0373	0.02707	0.04214	0.05656	0.04469	0.06233	0.00845	0.03577
28	Chloroflexus	0.11165	0.00913	0.04999	0.05489	0.03395	0.07508	0.02203	0.06084	0.01942	0.11375	0.0219	0.0257	0.05778	0.04189	0.07782	0.04814	0.05758	0.07602	0.01582	0.02663
29	Dictyoglomus	0.05175	0.00781	0.04687	0.08251	0.05036	0.06613	0.01364	0.09922	0.08789	0.10536	0.02077	0.04564	0.03887	0.0212	0.04543	0.05539	0.04123	0.06691	0.01081	0.04221
30	Synechocystis	0.08285	0.01119	0.0493	0.06173	0.04029	0.07021	0.01904	0.0628	0.04545	0.1143	0.02225	0.03915	0.05043	0.05466	0.05253	0.05863	0.05362	0.06619	0.01606	0.02931
31	Chlamydomonas	0.15477	0.01565	0.04349	0.0509	0.02383	0.10551	0.0216	0.02349	0.03227	0.08804	0.02077	0.02148	0.07044	0.04449	0.065	0.07292	0.04813	0.0649	0.01302	0.01929
32	Physcomitrella	0.07484	0.02155	0.04843	0.06009	0.04002	0.06676	0.02582	0.04853	0.05496	0.09671	0.02654	0.03897	0.0497	0.03905	0.06068	0.08663	0.05225	0.06816	0.01383	0.02647
33	Arabidopsis	0.06273	0.01994	0.05248	0.0656	0.04369	0.0642	0.02281	0.05348	0.06493	0.09354	0.02598	0.0433	0.04848	0.03445	0.05462	0.09068	0.05136	0.06676	0.01238	0.02861
34	Maize	0.09753	0.01959	0.05225	0.05851	0.03508	0.07529	0.02553	0.04017	0.04821	0.0923	0.02409	0.03175	0.0599	0.03554	0.0682	0.08228	0.04772	0.06834	0.01287	0.02485
35	Leishmania	0.11223	0.01983	0.04876	0.0607	0.0325	0.06396	0.02691	0.03322	0.03921	0.09125	0.02512	0.02826	0.05521	0.04032	0.07125	0.08224	0.05839	0.07266	0.01129	0.0267
36	Paramecium	0.03532	0.01676	0.04883	0.06691	0.04863	0.03576	0.01871	0.08293	0.09275	0.09577	0.02274	0.06816	0.03047	0.09121	0.03745	0.069	0.04489	0.04457	0.00769	0.04145
37	C. elegans	0.064	0.0217	0.05155	0.06312	0.04829	0.05487	0.02308	0.06134	0.06279	0.08519	0.02846	0.04835	0.0496	0.04117	0.05239	0.08043	0.05844	0.06179	0.01113	0.03232
38	Leech	0.05147	0.02567	0.05646	0.05683	0.04471	0.04953	0.02556	0.06147	0.07335	0.08879	0.02498	0.07006	0.03854	0.04047	0.04827	0.08343	0.05604	0.06032	0.01101	0.03304
39	Tick	0.08029	0.02489	0.04887	0.05652	0.04033	0.06947	0.02683	0.03855	0.04984	0.09581	0.02322	0.03298	0.05936	0.03782	0.07051	0.07833	0.05572	0.07014	0.01288	0.02766
40	Tribolium	0.06016	0.0219	0.05065	0.0638	0.04406	0.05509	0.02475	0.05852	0.06969	0.09195	0.02324	0.0507	0.05071	0.03975	0.0555	0.07415	0.05709	0.06388	0.01131	0.0331
41	Mosquito	0.07664	0.02044	0.05163	0.06165	0.03768	0.06547	0.02669	0.04995	0.05456	0.08998	0.02429	0.04352	0.05282	0.04528	0.0567	0.07537	0.05846	0.06528	0.01046	0.03311
42	Fruit fly	0.07466	0.01998	0.05093	0.0619	0.03684	0.0634	0.02629	0.05018	0.05675	0.09008	0.02489	0.04678	0.05417	0.05009	0.05564	0.08075	0.05563	0.05963	0.01036	0.03105
43	Sea squirt	0.05675	0.02469	0.05016	0.05876	0.04303	0.05446	0.02561	0.05727	0.06568	0.08966	0.02772	0.05207	0.04579	0.04095	0.05217	0.0817	0.06162	0.06652	0.01237	0.03304
44	Amphioxus	0.0698	0.02301	0.05797	0.06595	0.03545	0.06886	0.02527	0.04221	0.05654	0.084	0.02618	0.03857	0.05582	0.04528	0.05799	0.07488	0.06323	0.06654	0.01258	0.02989
45	Spotted gar	0.0674	0.02433	0.04895	0.06803	0.03908	0.06396	0.02496	0.04619	0.05868	0.09678	0.023	0.0374	0.05473	0.04547	0.05754	0.08232	0.05429	0.06445	0.01257	0.02987
46	Zebrafish	0.06315	0.02343	0.05175	0.06919	0.03743	0.06012	0.02709	0.04646	0.06054	0.09262	0.02514	0.03983	0.05454	0.04697	0.05588	0.08788	0.05704	0.06193	0.01124	0.02777
47	Medaka	0.06564	0.02437	0.04994	0.06538	0.0399	0.0628	0.02691	0.04478	0.05759	0.09657	0.02436	0.03807	0.05582	0.04544	0.05741	0.08627	0.05484	0.06393	0.01244	0.02755
48	X. laevis	0.0607	0.02291	0.05061	0.0704	0.03798	0.05895	0.02587	0.05151	0.06417	0.09248	0.02458	0.04323	0.05475	0.04659	0.05256	0.08566	0.05624	0.05961	0.0114	0.0298
49	X. tropicalis	0.06238	0.02337	0.04966	0.06975	0.03797	0.06065	0.02556	0.05127	0.06322	0.09356	0.02402	0.04213	0.0554	0.04635	0.05229	0.08451	0.05626	0.05986	0.01158	0.03019
50	Chicken	0.07187	0.02402	0.04757	0.06957	0.03727	0.06569	0.0254	0.04561	0.05998	0.09608	0.02312	0.03725	0.05805	0.0453	0.05749	0.08078	0.05228	0.06145	0.01284	0.02839
51	Rhesus	0.07029	0.02363	0.04586	0.06801	0.03871	0.06623	0.02679	0.04407	0.0577	0.10017	0.02191	0.03524	0.06247	0.04662	0.05799	0.08248	0.05252	0.05923	0.01343	0.02665
52	Gorilla	0.07005	0.02395	0.04662	0.06869	0.03786	0.06607	0.02597	0.0443	0.05895	0.10004	0.02366	0.03552	0.06075	0.04644	0.05765	0.08017	0.05231	0.06023	0.01341	0.02736
53	Chimpanzee	0.07068	0.02395	0.04636	0.06835	0.03773	0.06657	0.02598	0.04408	0.05832	0.09981	0.02347	0.03538	0.06142	0.04635	0.05804	0.08043	0.0522	0.0601	0.01343	0.02734
54	Human	0.07089	0.02315	0.04663	0.06878	0.03692	0.0669	0.02553	0.04258	0.05684	0.10069	0.0249	0.03473	0.06186	0.04699	0.05886	0.08145	0.05215	0.05991	0.0138	0.02644
55	Dog	0.07254	0.02311	0.04653	0.06768	0.03749	0.06831	0.02539	0.04417	0.05761	0.09954	0.02268	0.0351	0.06315	0.04556	0.0594	0.07978	0.0516	0.06017	0.0129	0.0273
56	Cow	0.0713	0.02382	0.04636	0.06718	0.03858	0.0669	0.02538	0.0446	0.05717	0.10151	0.02267	0.03483	0.06163	0.04533	0.05833	0.08009	0.05221	0.06139	0.01314	0.0276
57	Mouse	0.06952	0.024	0.04673	0.06759	0.03784	0.06468	0.02568	0.04322	0.05658	0.10134	0.02516	0.03479	0.06035	0.04689	0.05773	0.08368	0.05318	0.06072	0.0131	0.02724
58	Rat	0.06832	0.02372	0.04726	0.06758	0.03902	0.06446	0.02572	0.0453	0.05891	0.10025	0.02371	0.03573	0.059	0.04595	0.05676	0.08176	0.05387	0.06188	0.01267	0.02814
59	Opossum	0.06337	0.02324	0.04617	0.06856	0.04076	0.06326	0.02571	0.05124	0.06268	0.09972	0.02359	0.03937	0.05748	0.04605	0.05437	0.08115	0.05338	0.05803	0.01277	0.02911
60	Thalassiosira	0.07733	0.01676	0.06188	0.06541	0.03605	0.07067	0.02319	0.04822	0.05527	0.08523	0.02611	0.04426	0.04645	0.03681	0.05342	0.08862	0.05928	0.06554	0.01169	0.02779
61	Daphnia	0.06545	0.02211	0.04769	0.05763	0.04517	0.05589	0.02655	0.05446	0.05946	0.09247	0.02491	0.04584	0.05232	0.04182	0.05925	0.08426	0.05903	0.06343	0.01281	0.02944
62	Plasmodium	0.02533	0.01834	0.05968	0.06987	0.04696	0.03194	0.02273	0.0924	0.11662	0.08264	0.02265	0.12125	0.02226	0.02818	0.02996	0.06374	0.04238	0.04154	0.00589	0.05479
63	Rice	0.10101	0.02119	0.04967	0.05574	0.03353	0.07968	0.0267	0.03922	0.0423	0.09099	0.02443	0.02969	0.06261	0.03226	0.07658	0.08252	0.04755	0.0678	0.01409	0.02246
64	Dictyostelium	0.03294	0.01611	0.05093	0.0582	0.04944	0.04721	0.01802	0.08516	0.08054	0.08467	0.01823	0.10532	0.03813	0.04601	0.02889	0.09229	0.05858	0.04532	0.00769	0.03632
65	Sea anemone	0.06266	0.0278	0.0502	0.05619	0.04028	0.06259	0.02953	0.05458	0.0609	0.09165	0.0242	0.04234	0.05043	0.03775	0.05596	0.07574	0.06042	0.06676	0.01377	0.03624
66	Monosiga	0.1039	0.01909	0.05702	0.05765	0.03427	0.0609	0.02881	0.03788	0.039	0.10162	0.0233	0.0333	0.05456	0.0458	0.06438	0.07597	0.05964	0.06657	0.01199	0.02435
67	Phytophthora	0.09438	0.01622	0.05824	0.06615	0.03737	0.06446	0.0231	0.03814	0.05226	0.09103	0.02554	0.03314	0.04708	0.04151	0.06074	0.07889	0.05956	0.07304	0.01248	0.02666
68	Giardia	0.08114	0.02487	0.05416	0.06029	0.03465	0.05496	0.02393	0.0567	0.05261	0.102	0.02345	0.03905	0.04728	0.03946	0.05605	0.08587	0.06189	0.06132	0.00718	0.03311
69	Cryptococcus	0.08439	0.01117	0.05258	0.06512	0.03483	0.07231	0.02338	0.04871	0.05388	0.08853	0.02211	0.03394	0.0654	0.03767	0.05888	0.08756	0.05785	0.06111	0.01411	0.02647
70	Candida	0.05197	0.01137	0.05762	0.06384	0.04413	0.05221	0.02096	0.07141	0.07562	0.09152	0.01838	0.06506	0.04436	0.04456	0.03835	0.08693	0.06096	0.0548	0.00984	0.0361
71	Corn smut	0.10327	0.01213	0.05634	0.05243	0.03353	0.06561	0.0263	0.04315	0.04756	0.08919	0.02011	0.0343	0.05665	0.04384	0.06263	0.09638	0.06023	0.05991	0.01242	0.02402
72	Yarrowia	0.07856	0.01291	0.05863	0.06502	0.03924	0.06433	0.02381	0.04992	0.06197	0.08679	0.02347	0.04042	0.05265	0.04181	0.05065	0.07893	0.06037	0.06737	0.01231	0.03084
73	S. pombe	0.06338	0.0154	0.05183	0.06532	0.04602	0.05104	0.02269	0.06145	0.06787	0.09658	0.02207	0.05059	0.04659	0.03825	0.05073	0.08978	0.05405	0.06082	0.01123	0.03431
74	Phaeosphaeria	0.08965	0.01512	0.0546	0.05947	0.03661	0.06767	0.02579	0.04783	0.05028	0.08506	0.02565	0.03634	0.05971	0.0398	0.06176	0.07982	0.0613	0.06069	0.01536	0.0275
75	Aspergillus	0.08656	0.01378	0.0541	0.06038	0.03735	0.06902	0.02433	0.0496	0.04711	0.0905	0.0221	0.03559	0.05949	0.03955	0.06283	0.08212	0.05931	0.06276	0.01493	0.02861
76	Neurospora	0.08676	0.01307	0.05429	0.06272	0.03435	0.07338	0.02471	0.04442	0.05181	0.0828	0.02244	0.03684	0.06379	0.04134	0.06171	0.08237	0.06223	0.06061	0.01412	0.02625
77	Trichomonas	0.05199	0.01552	0.05795	0.07639	0.04988	0.04147	0.01959	0.07515	0.08908	0.0833	0.02467	0.06123	0.04428	0.04052	0.04005	0.07025	0.05467	0.05306	0.0088	0.04214
78	Stem rust	0.07212	0.01573	0.0524	0.05697	0.0361	0.06061	0.0281	0.05123	0.05661	0.09193	0.02135	0.04427	0.06631	0.04628	0.05736	0.09364	0.05969	0.05225	0.01279	0.02424
79	S. cerevisiae	0.05717	0.01413	0.0555	0.06335	0.04529	0.0515	0.02196	0.06456	0.07492	0.09471	0.0224	0.05807	0.04343	0.03951	0.04671	0.0861	0.05799	0.05745	0.01078	0.03444
80	Sclerotinia	0.07255	0.01488	0.05248	0.06552	0.03796	0.06927	0.02306	0.05681	0.05839	0.0845	0.02715	0.04194	0.05239	0.0367	0.05929	0.08422	0.05947	0.0593	0.01489	0.02923
81	Batrachochytrium	0.06999	0.01564	0.0559	0.05387	0.03911	0.05439	0.02688	0.06248	0.0596	0.09164	0.02551	0.04485	0.04863	0.04543	0.04799	0.09136	0.06376	0.06177	0.01104	0.03017

Table 2. Average Exome-Wide Amino Acid Composition for 81 Species

This table shows the average fractional composition of each of the 20 standard amino acids (Ala, Cys, Asp, Glu, Phe, Gly, His, Ile, Lys, Leu, Met, Asn, Pro, Gln, Arg, Ser, Thr, Val, Trp, and Tyr) in the exomes of 81 species. Each row corresponds to a species, identified by the “No.” and “Short Label” columns. For each species, all exon-derived amino acid sequences were aggregated, normalized to ensure the total of all fractions equals 1, and then averaged across all exons. These data were used to calculate the angular distances and perform hierarchical clustering as described in the text.

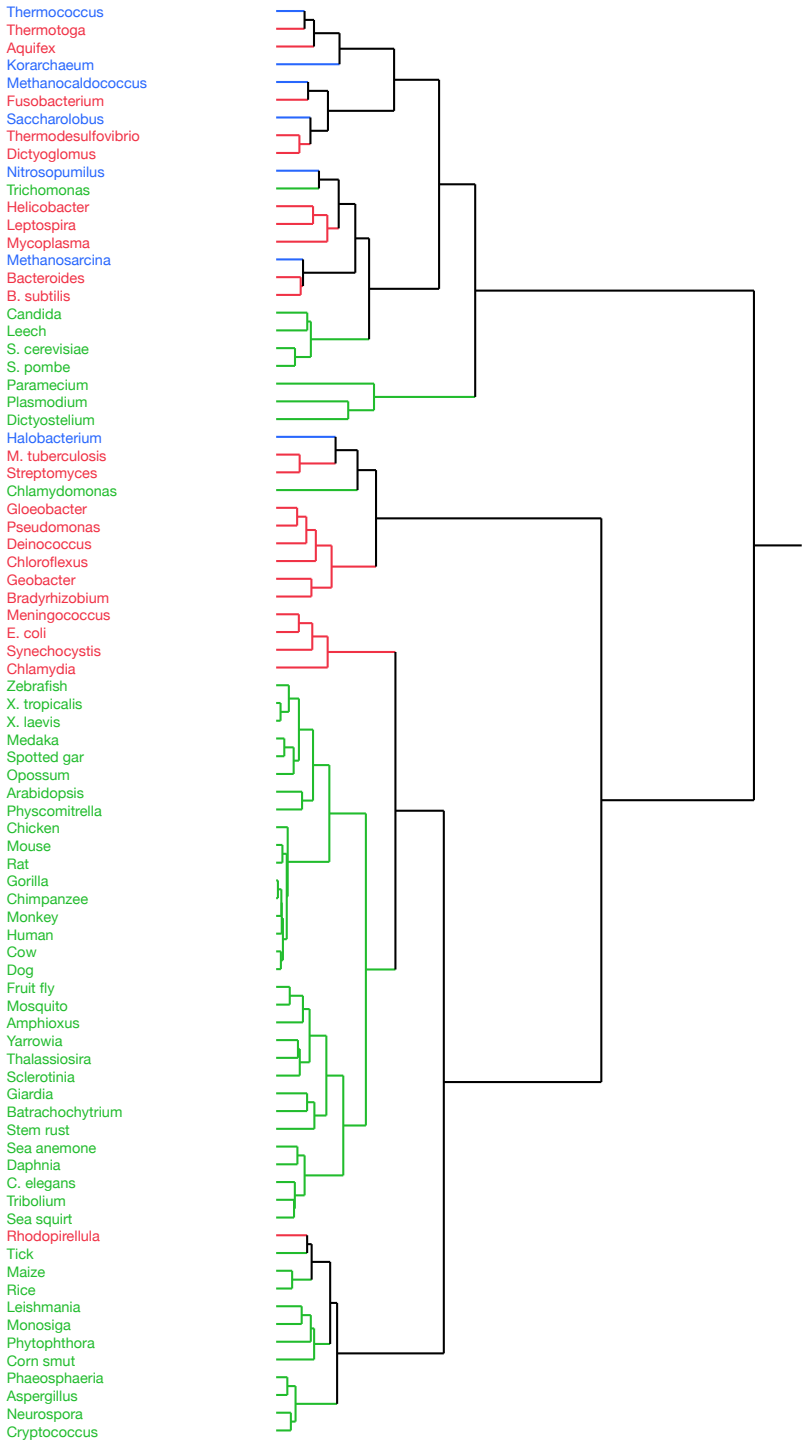

Figure 1. Horizontal Phylogenetic Tree Based on Average Exome-Wide Amino Acid Composition

This dendrogram was generated using hierarchical clustering with Ward's method, applied to the pairwise angular distance matrix calculated from the average amino acid compositions of 81 species' exomes. Taxonomic groups are color-coded: blue for archaea, red for bacteria, and green for eukaryotes. Short labels for each species are displayed on the left; see Table 1 for full scientific names. The branch lengths reflect the relative distances in amino acid composition between species.

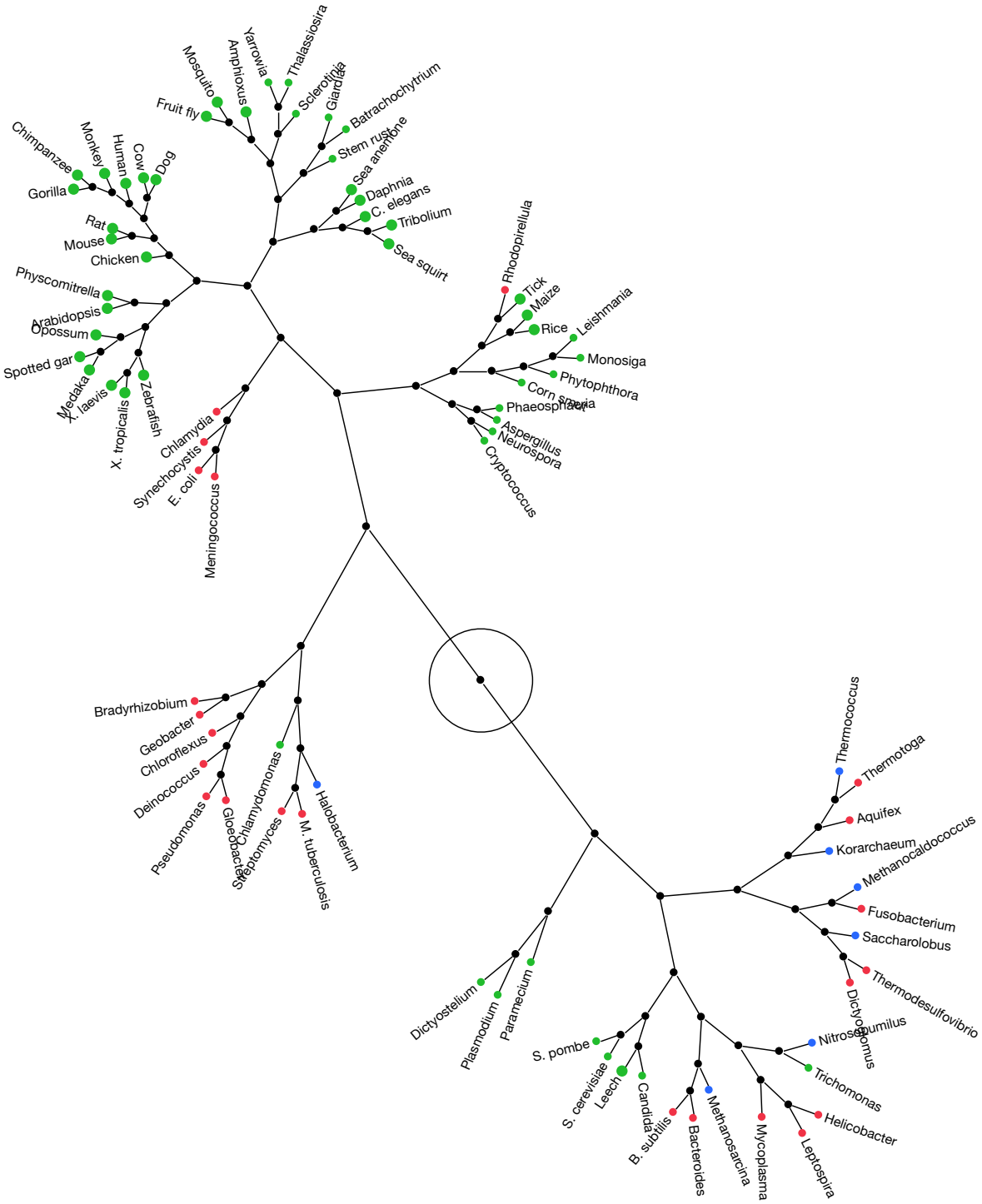

Figure 2. Radial Phylogenetic Tree Based on Average Exome-Wide Amino Acid Composition

This radial (divergent) dendrogram was generated using the same dataset and clustering approach described in Figure 1, but displayed in a circular layout to illustrate the evolutionary divergence among the 81 species. As in Figure 1, archaea are shown in blue, bacteria in red, and eukaryotes in green. Unicellular organisms are indicated with small nodes, whereas multicellular organisms are represented by larger nodes and bold labels. Short labels correspond to those in Table 1.