Table 2. Median protein lengths in eukaryotic, bacterial and archaeal organisms

Species[a]	All species Number[b]	Median[c]	Classified in COG Number[b]	Median[c]	Classified in Pfam-A Number[b]	Median[c]
Eukarya	**104 394**	**361**	**5177**	**471**	**71 584**	**419**
HUMAN	33 869	375	–	–	21 686	416
DROME	14 226	373	3092	492	13 091	475
CAEEL	21 124	344	–	–	13 316	391
YEAST	6315	379	2085	438	3953	448
ARATH	28 860	356	–	–	19 538	407
Bacteria	**191 541**	**267**	**83 513**	**304**	**131 915**	**306**
ECOLI	4289	*278*	3289	*309*	3483	*303*
SALTY	4553	*271*	3408	*303*	3527	*300*
SALTI	4767	*253*	3258	*300*	3118	*300*
YERPE	4083	*268*	2991	*299*	3003	*304*
SHIFL	4180	*261*	–	–	2613	*304*
WIGBR	654	*268*	–	–	571	*291*
BUCAI	574	*282*	558	*284*	544	285
BUCAP	545	*279*	–	–	536	285
HAEIN	1709	*262*	1470	*286*	750	*314*
PASMU	2014	*286*	1740	*302*	780	*289*
XANCP	4181	*291*	–	–	2976	*325*
XANAC	4312	*286*	–	–	3056	*326*
XYLFA	2832	201	1549	*305*	1544	*305*
PSEAE	5565	*291*	4355	*310*	4309	*309*
VIBCH	3828	*259*	2794	*315*	2731	*312*
Chromosome 1	2736	273	2133	314		
Chromosome 2	1092	225	661	316		
SHEON	4778	245	–	–	2913	*308*
NEIMB	2025	239	1448	*291*	1310	*305*
RALSO	5116	*276*	–	–	3518	*310*
Chromosome 1	3440	271				
Chromosome 2	1676	296				
AGRT5	5402	*280*	3984	*316*	4062	*307*
Circular chr.	2785	258	2098	305		
Linear chr.	1876	302	1424	329		
Plasmids	741	273	462	316		
RHILO	7275	*269*	5184	*305*	5107	*303*
Chromosome	6746	270	4888	304		
Plasmids	529	243	296	327		
RHIME	6205	*281*	4614	*312*	4669	*308*
Chromosome	3341	276	2602	302		
Plasmid A	1294	265	890	310		
Plasmid B	1570	303	1122	330		
BRUME	3198	*263*	–	–	2322	*300*
Chromosome 1	2059	252				
Chromosome 2	1139	279				
BRUSU	3264	*254*	–	–	2351	*301*
Chromosome 1	2116	239				
Chromosome 2	1148	278				
CAUCR	3737	*275*	2551	*317*	2686	*312*
RICPR	834	*283*	687	*295*	672	*299*
RICCN	1374	173	861	247	769	264
HELPY	1566	*266*	1083	*303*	1052	*315*
CAMJE	1634	*268*	1309	*294*	1197	*298*
MYCTU	3918	*287*	2554	*322*	2213	*326*
MYCLE	1605	*282*	1145	*326*	1138	*324*
STRCO	7897	*278*	–	–	5330	*317*
CORGL	2993	*275*	1954	*307*	1985	*314*
COREF	2950	*287*	–	–	1961	*323*
BIFLO	1729	*321*	–	–	1286	*341*
OCEIH	3496	*261*	–	–	2583	*295*
BACSU	4100	*256*	2818	*298*	2974	*297*
BACHD	4066	*261*	2838	*303*	2916	*300*
STAAN	2625	*257*	1801	*300*	1542	*293*
LISIN	3043	*255*	2176	*289*	2234	*291*
LISMO	2846	*267*	2206	*289*	2211	*292*
LACLA	2266	*251*	1602	*288*	1690	281
STRA5	2124	*254*	–	–	1480	*290*
STRMU	1960	*250*	–	–	1445	*282*
STRPN	2043	243	1465	*287*	1409	*291*
STRPY	1696	*263*	1178	*294*	1159	*299*
CLOAB	3848	*262*	2487	*298*	2634	*299*

Table 2. *Continued*

Species[a]	All species Number[b]	Median[c]	Classified in COG Number[b]	Median[c]	Classified in Pfam-A Number[b]	Median[c]
CLOPE	2723	*268*	–	–	1997	*303*
THETN	2588	*269*	–	–	1903	*306*
UREPA	614	*286*	409	*298*	395	*303*
MYCGE	484	*292*	384	*292*	375	*304*
MYCPN	677	*286*	407	*299*	507	*299*
MYCPU	782	*297*	489	*302*	498	*320*
MYCPE	1037	*304*	–	–	664	*315*
FUSNN	2067	*261*	–	–	1432	*303*
SYNY3	3169	*274*	2141	*318*	2344	*306*
ANASP	6129	*256*	–	–	3600	*320*
SYNEL	2475	*272*	–	–	1759	*315*
CHLTR	894	*289*	615	*316*	639	*327*
CHLMU	916	*290*	644	*321*	641	*330*
CHLPN	1052	*289*	646	*324*	716	*333*
BORBU	1637	220	635	*318*	981	265
Chromosome	850	286	–	–		
Plasmids	787	179	–	–		
TREPA	1031	*293*	708	*331*	691	*337*
LEPIN	4727	207	–	–	2243	*309*
Chromosome 1	4360	206				
Chromosome 2	367	223				
DEIRA	3182	*264*	2249	*303*	2050	*307*
Chromosome 1	2629	257	1873	294		
Chromosome 2	368	304	265	347		
Plasmids	185	303	111	336		
CHLTE	2252	239	–	–	1431	*311*
THEMA	1846	*284*	1509	*303*	1459	*304*
AQUAE	1560	*272*	1321	*291*	1231	*297*
Archaea	**37 141**	**247**	**18 219**	**283**	**24 067**	**288**
PYRAB	1765	265	1450	281	1407	282
PYRHO	1801	257	1398	283	1312	285
PYRFU	2065	253	1627	273	1477	281
ARCFU	2420	243	1887	270	1720	276
THEAC	1482	269	1233	293	1083	301
THEVO	1499	259	1247	287	1074	304
METTH	1869	242	1382	273	1325	277
METJA	1770	241	1298	266	1260	272
METAC	4540	256	–	–	2677	306
METMA	3371	255	–	–	2141	294
METKA	1691	257	–	–	1067	272
HALN1	2622	242	1746	297	1471	303
AERPE	1840	239	1191	293	1067	301
PYRAE	2603	208	–	–	1411	267
SULSO	2977	251	1983	294	1917	293
SULTO	2826	226	1777	284	1658	279

[a]See Table 1 for abbreviations.
[b]Number of proteins in each set.
[c]Median length.

Table 1. Proteomic collections

	Species	Abbreviation
Eukaryota	*Homo sapiens*[a]	HUMAN
	Drosophila melanogaster	DROME
	Caenorhabditis elegans[a]	CAEEL
	Saccharomyces cerevisiae	YEAST
	Arabidopsis thaliana[a]	ARATH
Euryarchaeota	*Pyrococcus abyssi*	PYRAB
	Pyrococcus horikoshii	PYRHO
	Pyrococcus furiosus	PYRFU
	Archaeoglobus fulgidus	ARCFU
	Thermoplasma acidophilum	THEAC
	Thermoplasma volcanium	THEVO
	Methanothermobacter thermoautotrophicus	METTH
	Methanococcus jannaschii	METJA
	Methanosarcina acetivorans[a]	METAC
	Methanosarcina mazei[a]	METMA
	Methanopyrus kandleri[a]	METKA
	Halobacterium sp. NRC-1	HALN1
Crenarchaeota	*Aeropyrum pernix*	AERPE
	Pyrobaculum aerophilum[a]	PYRAE
	Sulfolobus solfataricus	SULSO
	Sulfolobus tokodaii	SULTO
γ-Proteobacteria	*Escherichia coli* K12	ECOLI
	Salmonella typhimurium	SALTY
	Salmonella enterica	SALTI
	Yersinia pestis CO92	YERPE
	Shigella flexneri[a]	SHIFL
	Wigglesworthia brevipalpis[a]	WIGBR
	Buchnera sp. APS	BUCAI
	Buchnera aphidicola[a]	BUCAP
	Haemophilus influenzae	HAEIN
	Pasteurella multocida	PASMU
	Xanthomonas campestris[a]	XANCP
	Xanthomonas axonopodis[a]	XANAC
	Xylella fastidios[a]	XYLFA
	Pseudomonas aeruginosa	PSEAE
	Vibrio cholerae	VIBCH
	Shewanella oneidensis[a]	SHEON
β-Proteobacteria	*Neisseria meningitides* MC58	NEIMB
	Ralstonia solanacearum[a]	RALSO
α-Proteobacteria	*Agrobacterium tumefaciens* C58 Cereon	AGRT5
	Mesorhizobium loti	RHILO
	Sinorhizobium meliloti	RHIME
	Brucella melitensis[a]	BRUME
	Brucella suis[a]	BRUSU
	Caulobacter crescentus	CAUCR
	Rickettsia prowazekii	RICPR
	Rickettsia conorii	RICCN
ε-Proteobacteria	*Helicobacter pylori* 26 695	HELPY
	Campylobacter jejuni	CAMJE
Actinobacteria	*Mycobacterium tuberculosis* H37 Rv	MYCTU
	Mycobacterium leprae	MYCLE
	Streptomyces coelicolor[a]	STRCO
	Corynebacterium glutamicum	CORGL
	Corynebacterium efficiens[a]	COREF
	Bifidobacterium longum[a]	BIFLO
Firmicutes	*Oceanobacillus iheyensis*[a]	OCEIH
	Bacillus subtilis	BACSU
	Bacillus halodurans	BACHD
	Staphylococcus aureus N315	STAAN
	Listeria innocua	LISIN
	Listeria monocytogenes	LISMO
	Lactococcus lactis	LACLA
	Streptococcus agalactiae 2603 V/R[a]	STRA5
	Streptococcus mutans[a]	STRMU
	Streptococcus pneumoniae TIGR4	STRPN
	Streptococcus pyogenes M1	STRPY
	Clostridium acetobutylicum	CLOAB
	Clostridium perfringens[a]	CLOPE
	Thermoanaerobacter tengcongensis[a]	THETN
	Ureaplasma urealiticum	UREPA
	Mycoplasma genitalium	MYCGE
	Mycoplasma pneumoniae	MYCPN
	Mycoplasma pulmonis	MYCPU
	Mycoplasma penetrans[a]	MYCPE
	Fusobacterium nucleatum[a]	FUSNN
Cyanobacteria	*Synechocystis* sp. PCC 6803	SYNY3
	Nostoc sp. PCC 7120[b]	ANASP
	Thermosynechococcus elongatus[a]	SYNEL
Chlamydiae	*Chlamydia trachomatis*	CHLTR
	Chlamydia muridarum	CHLMU
	Chlamydophila pneumoniae CWL029	CHLPN
Spyrochaetes	*Borrelia burgdorferi*	BORBU
	Treponema pallidum	TREPA
	Leptospira interrogans[a]	LEPIN
Others	*Deinococcus radiodurans*	DEIRA
	Chlorobium tepidum[a]	CHLTE
	Thermotoga maritima	THEMA
	Aquifex aeolicus	AQUAE

[a]Proteins from these species are not classified in the COG database and are excluded from the functional group analyses.
[b]The COG classification of proteins from this species does not follow the standard coding and has been excluded from the COG analyses.

lengths were compared with respect to taxonomic, functional and ecological classes.

Datasets of selected proteins

We evaluated results using the set of proteins classified in the COG (Clusters of Orthologous Groups of proteins) database (13,14) and the set of genomic proteins included in the Pfam (15–17) database of functional/structural domain alignments verified by human intervention (Pfam-A). The COG database classifies orthologous proteins in functional groups. At the onset of this study, classification of proteins into COGs was available for 2 eukaryotic proteomes (yeast and *Drosophila melanogaster*), 12 archaeal proteomes and 44 bacterial proteomes (Tables 1 and 2). COG data were obtained for prokaryotic organisms and yeast from the corresponding tables (*.ptt) available from the NCBI genomes ftp site (ftp://ftp.ncbi.nih.gov/genomes). Data for *D.melanogaster* were obtained from the classification table available at the COG website (http://www.ncbi.nlm.nih.gov/COG). All proteomes included in our analysis are also represented in the Pfam-A database.

Statistical significance evaluations

We compared median protein lengths (the midpoint of all lengths arranged in order of magnitude) rather than average lengths between two sets of proteins to reduce the effect of outliers. The statistical significance of the difference in median length between the proteins of two sets was evaluated estimating the distribution of median length differences between samples created by randomly redistributing all sequences from the two sets into two new sets of the original sizes. For each determination, from 200 to 1000 independent data shufflings were implemented. We highlight the differences in median length observed in <1% of all shuffled samples ($P < 0.01$, boldfaced in the tables) and observed in the range 1–10% of all shuffled samples ($0.01 \leqslant P < 0.1$, underlined in the tables).