The quality of a clustering solution

1. Classification quality indices. In order to estimate the quality of each clustering solution, we introduced three empirical indices: the "group homogeneity"(GrH), "functional homogeneity,"(FunH), and "uncertainty"(Unc). The first and the second indices indicate the percentage of COGs from the same group/functional category in the cluster (we used definitions of groups and functional categories from the COGs database). The Unc is computed as the percent of poorly characterized COGs in the cluster. Three more indices reflect the statistical properties of the cluster, namely "consistency"(Cons), "average distance between cluster members"(AveD) and "in-cluster variance" (Var) (see comments in additional data file 1 for computational details). The consistency is computed for every cluster member x^j_i (j=1,…,N, i=1,..n_j; N is the total number of clusters and n_j is the number of members in j^{th} cluster) separately, as the probability of x^j_i belonging to the cluster ($p_{x^j_i}$), and then is averaged over all members. Consider j^{th} cluster and let k_i be the number of COGs in clusters other than j, which are closer (have smaller distance) to x^j_i than are on the average COGs in j^{th} cluster. Then $p_{x^j_i}$ is computed as $p_{x^j_i} = 1 - k_i / \sum_{i=1}^{N} n_i$, and the consistency of the j^{th} cluster is estimated as $\hat{c}_j = \sum_{i=1}^{n_j} p_{x_i^j} / n_j$. AveD is the averaged distance among all members of given cluster and Var is the average of all distance variances for all cluster members.

Table 1. The values of classification quality indices for UPGMA/NJ algorithms with different distance measures

Members' number thresholds (MNT)	Numb. of clus-ters	Aver. clus-ter size	GrH	FunH	Unc	Cons	AveD	In clus-ter Var.	To-tal co-verage: (lost COGs)	Co-verage in %
Distance measure d_r										
A)UPGMA										
Without MNT	67	21.194	0.712	0.598	0.371	0.988	0.174	0.085	3169	69.1
50	157	21.038	0.632	0.458	0.431	0.944	0.358	0.133	1286	28.0
100	115	31.870	0.619	0.439	0.430	0.914	0.394	0.145	924	20.1
150	96	38.812	0.618	0.436	0.431	0.901	0.418	0.153	863	18.8
200	82	46.451	0.615	0.431	0.432	0.887	0.439	0.155	780	17.0
250	70	54.886	0.626	0.438	0.449	0.875	0.455	0.161	747	16.3
B)NJ										
Without MNT	72	17.806	0.700	0.580	0.344	0.985	0.177	0.095	3307	72.1
50	167	21.521	0.616	0.432	0.451	0.929	0.393	0.154	995	21.7
100	111	34.351	0.596	0.416	0.445	0.887	0.436	0.158	776	16.9

150	90	42.867	0.605	0.422	0.458	0.867	0.460	0.165	731	15.9
200	74	53.243	0.614	0.428	0.468	0.841	0.479	0.170	649	14.1
250	56	73.196	0.621	0.428	0.455	0.792	0.496	0.175	490	10.7

Distance measure d_{r2}

A)UPGMA

Without MNT	60	16.617	0.733	0.611	0.386	0.984	0.379	0.163	3592	78.3
50	197	20.071	0.609	0.413	0.424	0.928	0.596	0.158	635	13.8
100	135	31.474	0.595	0.392	0.404	0.887	0.633	0.157	340	7.4
150	109	39.578	0.587	0.391	0.425	0.862	0.666	0.158	275	6.0
200	94	47.032	0.581	0.384	0.421	0.840	0.688	0.157	168	3.7
250	83	53.386	0.588	0.382	0.415	0.822	0.711	0.157	158	3.4

B)NJ

WithoutMNT	80	14.025	0.701	0.576	0.412	0.975	0.378	0.172	3467	75.6
50	170	24.694	0.607	0.405	0.444	0.903	0.631	0.170	391	8.5
100	97	45.289	0.582	0.373	0.417	0.817	0.675	0.167	196	4.3
150	64	69.031	0.574	0.368	0.405	0.740	0.721	0.173	171	3.7
200	52	85.327	0.566	0.367	0.399	0.688	0.736	0.172	152	3.3
250	42	105.952	0.555	0.350	0.404	0.629	0.755	0.169	139	3.0

Distance measure $d_{|r|}$

A)UPGMA

WithoutMNT	60	16.250	0.720	0.618	0.357	0.982	0.186	0.099	3614	78.8
50	173	19.075	0.634	0.457	0.426	0.952	0.348	0.128	1289	28.1
100	120	30.708	0.637	0.454	0.424	0.921	0.370	0.133	904	19.7
150	100	38.000	0.634	0.451	0.420	0.909	0.389	0.138	789	17.2
200	91	42.275	0.634	0.451	0.421	0.902	0.403	0.140	742	16.2
250	82	47.232	0.642	0.452	0.408	0.890	0.421	0.145	716	15.6

B)NJ

WithoutMNT	91	15.868	0.702	0.584	0.413	0.984	0.173	0.092	3145	68.5
50	167	21.623	0.618	0.427	0.434	0.928	0.378	0.144	978	21.3
100	110	35.355	0.627	0.422	0.421	0.889	0.423	0.152	700	15.3
150	88	45.193	0.621	0.418	0.423	0.862	0.444	0.161	612	13.3
200	70	57.986	0.599	0.403	0.447	0.834	0.465	0.165	530	11.5
250	62	65.887	0.599	0.407	0.459	0.832	0.464	0.163	504	11.0

2. Predictive power. Using the descriptions of 52 metabolic pathways and functional systems (http://www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?sys=all.), we compared the predictive power (PPs) of UPGMA and NJ with different distance measures (Table 2).

Table 2. The performance (predictive power) of UPGMA/NJ algorithms with different distance measures

Pathways and functional systems (http://www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?sys=all)	NJAC250*	NJAC250†	UPAC250‡	UP250§
AMINOACYL-tRNA_SYN	11.5	11.5	53.8	11.5
ARCHAEAL-VACUOLAR-TYPE H+-ATPASE SUBUNITS	77.8	88.9	66.7	88.9
ARGININE_BIOSYNTHESIS	63.6	27.3	81.8	90.9
BASAL_REPL_MACHINERY	38.5	30.8	34.6	38.5
BASAL_TF	81.8	45.5	72.7	54.5
BIOTIN_BIOSYNTHESIS	66.7	50	50	50
COBALAMIN_BIOSYNTHESIS	77.8	72.2	77.8	77.8
COENZYME_A_BIOSYNTHESIS	33.3	22.2	33.3	33.3
DEOXYXYLULOSE_PATHWAY_OF_TERPENOID BIOSYNTHESIS	100	100	100	100
DNA_POLIII_SUB	37.5	50	37.5	50
DNA-DEPENDENT_RNA_POL	40	40	46.7	40
ENTNER-DOUDOROFF_PATHWAY	50	50	25	75
F0F1-TYPE_ATP_SYN	77.8	77.8	77.8	77.8
FAD_BIOSYNTHESIS	44.4	66.7	33.3	44.4
FATTY_ACID_BIOSYNTHESIS	38.5	38.5	38.5	46.2
FLAGELLUM	66.7	66.7	75.8	75.8
GLUCONEOGENESIS	35.7	14.3	28.6	7.1
GLYCOLYSIS	35.7	14.3	35.7	7.1
GLYOXYLATE_BYPASS	100	100	100	100
HEME_BIOSYNTHESIS	21.4	21.4	21.4	21.4
HISTIDINE_BIOSYNTHESIS	83.3	66.7	75	75
ISOLEUCINE_BIOSYNTHESIS	83.3	83.3	66.7	100
LEUCINE_BIOSYNTHESIS	50	80	80	90
LIPID_BIOS	100	100	100	100
MENAQUINONE_BIOSYNTHESIS	25	25	31.2	31.2
METHIONINE_BIOSYNTHESIS	40	30	50	50
MULTISUBUNIT_NA+-H+_ANTIPORTER	75	62.5	75	62.5
NA+-TRANSPORTING_NADHUBIQUINONE OXIDOREDUCTASE_SUBUNITS	85.7	85.7	85.7	85.7
NAD_BIOSYNTHESIS	42.9	28.6	42.9	42.9
NADHUBIQUINONE_OXIDOREDUCTASE SUBUNITS	86.7	86.7	86.7	86.7

PENTOSE_PHOSPHATE_PATHWAY	33.3	55.6	22.2	55.6
PHENYLALANINE-TYROSINE_BIOSYNTHESIS	42.9	35.7	57.1	57.1
PREPROTEIN_TRANSLOCASE_SUBUNITS	44.4	44.4	44.4	55.6
PROLINE_BIOSYNTHESIS	60	60	40	80
PURINE_BIOSYNTHESIS	33.3	33.3	66.7	83.3
PURINE_SALVAGE	20	40	40	20
PYRIDOXAL_PHOSPHATE_BIOSYNTHESIS	37.5	37.5	50	25
PYRIMIDINE_BIOSYNTHESIS	42.9	50	57.1	57.1
PYRIMIDINE_SALVAGE	40	30	40	30
PYRUVATE_DECARBOXYLATION	42.9	42.9	42.9	42.9
RIBOFLAVIN_BIOSYNTHESIS	42.9	85.7	42.9	42.9
RIBOSOMAL_PROTEINS_LS	45.1	27.5	56.9	35.3
RIBOSOMAL_PROTEINS_SS	34.4	31.2	43.8	31.2
TCA_CYCLE	18.8	12.5	18.8	18.8
TF_AND_INVOLVED_ENZYMES	40.9	40.9	63.6	40.9
THIAMINE_BIOSYNTHESIS	20	20	30	60
THREONINE_BIOSYNTHESIS	80	80	80	80
THYMIDYLATE_BIOSYNTHESIS	10	30	20	10
TRANSCR_REG	10	11.4	20	11.4
TRYPTOPHAN_BIOSYNTHESIS	52.9	52.9	64.7	64.7
UBIQUINONE_BIOSYNTHESIS	40	20	26.7	26.7
VALINE_BIOSYNTHESIS	66.7	83.3	83.3	100
Average:	50.6	49.3	53.8	54.7
Gene Displacements	83	47	53	43

*NJ with distance measure d_{r2};
†NJ with distance measure d_r;
‡UPGMA with distance measure d_{r2};
§UPGMA with distance measure d_r

3. Weighting schemes. We were interested whether weighting of phyletic patterns may improve discovery of functional links. For example, patterns shared by many COGs may have different functional significance than patterns including only a few COGs. Second, the probability that two patterns belong to the same pathway or functional system could be higher when they have similar rate of evolutionary gain and loss, so one can introduce a term accounting for the rate of pattern changes in evolution. Third, one can take into account the phylogenetic breadth at which co-inheritance is observed – if two genes are co-gained or co-lost once in bacteria and another time in archaea, this may have different significance than two such events within, say, alphaproteobacteria. We tried to weight the distance measure by accounting for each of these effects individually, measured an

increase in recovery (predictive power) of a pathway or functional system. None of these weighting approaches improved the PP of the pairs “distance measure + algorithm”.