

EVALUATING THE CLUSTERING EFFICIENCIES OF THE THREE METHODS ON SIMULATED METAGENOMIC DATA SETS USING GENEALOGICAL SORTING INDEXES (GSIs)

The clustering efficiencies of the three methods (i.e. HabiSign, average tetranucleotide frequency based approach, dinucleotide relative abundance approach) were evaluated on two simulated metagenomic data sets. The first set contained metagenomic groups having very distinct taxonomic composition. All three methods are expected to cluster these groups with equal efficiency. In contrast, the second simulated set contained metagenomic groups having similarity with respect to the constituent species. However, there were subtle differences in the relative abundance of the constituent species. A method with a high clustering efficiency is expected to capture these subtle differences and appropriately cluster the metagenomic samples belonging to different groups.

The details of the simulated metagenomic data sets, the methodology adopted for evaluation and the results obtained are summarized in this document.

SIMULATED DATA SET 1

This data set consisted of nine simulated metagenomic samples (M11, M12, M13, M21, M22, M23, M31, M32, M33). The taxonomic composition of these samples were simulated in such a manner that they could be classified into three different groups. The details of these groups are provided below. Note that the species constituting these three groups are entirely different.

(A) GROUP1: Consists of metagenomic samples M11, M12 and M13. The relative abundance of the constituent species (in terms of the number of sequences) are also indicated below.

SPECIES	No. of sequences in metagenome		
	M11	M12	M13
<i>Aeromonas hydrophila</i>	1436	6351	2101
<i>Akkermansia muciniphila</i>	5912	5182	2477
<i>Bifidobacterium longum</i>	2670	7323	5259
<i>Buchnera aphidicola</i>	6056	4457	2405
<i>Desulfovibrio desulfuricans</i>	7199	6053	1792
<i>Dichelobacter nodosus</i>	7555	2469	1546
<i>Dictyoglomus thermophilum</i>	1374	4548	4392
<i>Gloeobacter violaceus</i>	3098	224	4293
<i>Gramella forsetii</i>	5659	2492	7176
<i>Halothermothrix orenii</i>	1457	892	203
<i>Herpetosiphon aurantiacus</i>	7268	1973	1180
<i>Legionella pneumophila</i>	6219	7916	411
<i>Marinomonas</i> sp. MWYL1	1421	6708	3167
<i>Methanospirillum hungatei</i>	3206	5048	918
<i>Natronaerobius thermophilus</i>	4933	791	4675
<i>Natronomonas pharaonis</i>	2144	4449	1840
<i>Nautilia profundicola</i>	4163	4392	6049
<i>Nitrosopumilus maritimus</i>	4888	4076	3433
<i>Prochlorococcus marinus</i>	7052	7385	2732
<i>Rhodospirillum centenum</i>	1107	3037	3405
<i>Roseiflexus</i> sp. RS-1	6685	3169	4779
<i>Rubrobacter xylanophilus</i>	3437	2491	6572
<i>Solibacter usitatus</i>	5597	4103	3656
<i>Sphingopyxis alaskensis</i>	7524	7520	954
<i>Thermoanaerobacter</i> sp. X514	1117	116	7981
<i>Thermodesulfovibrio yellowstonii</i>	6283	529	4262
<i>Thermoplasma volcanium</i>	2390	3244	7991
<i>Thiobacillus denitrificans</i>	2352	5345	3473

B) GROUP 2: Consists of metagenomic samples M21, M22 and M23. The relative abundance of the constituent species (in terms of the number of sequences) are also indicated below.

SPECIES	No. of sequences in metagenome		
	M21	M22	M23
<i>Acholeplasma laidlawii</i>	6115	2567	3980
<i>Acidithiobacillus ferrooxidans</i>	7735	1758	480
<i>Acidobacterium capsulatum</i>	1627	3036	6384
<i>Archaeoglobus fulgidus</i>	5517	4818	7229
Aster yellows witches-broom phytoplasma	7558	7	736
<i>Bacteroides fragilis</i>	2659	4771	5854
<i>Desulfotalea psychrophila</i>	6397	4152	4129
<i>Fusobacterium nucleatum</i>	2222	3129	2847
<i>Gemmatimonas aurantiaca</i>	4564	7122	7439
<i>Lawsonia intracellularis</i>	6408	3868	4190
<i>Mesoplasma florum</i>	5267	5295	5149
<i>Methanococcus maripaludis</i>	5606	2361	3546
<i>Methanopyrus kandleri</i>	3542	3638	6308
<i>Methanosarcina barkeri</i>	7164	1934	6516
<i>Methanosphaera stadtmanae</i>	1437	7907	3140
<i>Methylobacillus flagellatus</i>	5184	21	1026
<i>Methylococcus capsulatus</i>	4090	5492	648
<i>Nitrosococcus oceani</i>	3193	4671	5277
<i>Nitrospira multiformis</i>	3070	4043	2748
<i>Pelobacter propionicus</i>	5250	106	7675
<i>Phenylobacterius zucineum</i>	6513	3429	4690
<i>Pyrococcus abyssi</i>	4860	6136	6838
<i>Salinibacter ruber</i>	2284	6574	7787
<i>Silicibacter</i> sp. TM1040	4689	6056	7654
<i>Sorangium cellulosum</i>	1932	1433	4073
<i>Staphylothermus marinus</i>	5634	3503	3787
<i>Sulfolobus tokodaii</i>	7462	3942	1326
<i>Syntrophus aciditrophicus</i>	3658	6826	6158
<i>Thermofilum pendens</i>	7550	5577	6554
<i>Thermomicrobium roseum</i>	391	3387	2354
<i>Thermus thermophilus</i>	7367	1984	5157
<i>Vibrio cholerae</i>	4647	7589	5372
<i>Yersinia pseudotuberculosis</i>	2738	6934	7595

(C) GROUP 3: Consists of metagenomic samples M31, M32 and M33. The relative abundance of the constituent species (in terms of the number of sequences) are also indicated below.

SPECIES	No. of sequences in metagenome		
	M31	M32	M33
<i>Acidovorax avenae</i>	4191	6497	5482
<i>Agrobacterium tumefaciens</i>	2433	7035	4643
<i>Aquifex aeolicus</i>	5217	4209	6380
<i>Bacillus halodurans</i>	1	1600	3018
<i>Bdellovibrio bacteriovorus</i>	4622	6298	6021
<i>Bordetella pertussis</i>	7381	5296	751
<i>Burkholderia cenocephalia</i>	3985	2338	2179
<i>Campylobacter jejuni</i>	5322	2700	5715
<i>Chlamydomonas pneumoniae</i>	1760	7212	4724
<i>Chlorobium chlorochromatii</i>	3952	1653	3657
<i>Clostridium tetani</i>	4772	6032	7837
<i>Corynebacterium diphtheriae</i>	6837	1084	2626
<i>Escherichia coli</i>	716	7475	5299
<i>Francisella philomiragia</i>	6171	7739	561
<i>Haemophilus influenzae</i>	2932	3484	2325
<i>Helicobacter pylori</i>	5473	1413	7003
<i>Magnetococcus</i> sp. MC-1	2013	2456	4254
<i>Mycobacterium tuberculosis</i>	3199	5722	7795
<i>Mycoplasma pneumoniae</i>	4953	2008	3294
<i>Neisseria meningitidis</i>	4884	2997	3028
<i>Nostoc</i> spp. 7120	4133	956	6205
<i>Pseudomonas aeruginosa</i>	3765	6623	4209
<i>Rhodospseudomonas palustris</i>	5738	4250	4312
<i>Rickettsia bellii</i>	3920	7592	385
<i>Salmonella enterica</i>	5154	6456	1790
<i>Staphylococcus aureus</i>	66	4501	949
<i>Streptococcus pyogenes</i>	7688	3077	4681
<i>Thermotoga maritima</i>	2460	7956	5354
<i>Wigglesworthia brevipalpis</i>	2054	7916	3290
<i>Xanthomonas campestris</i>	1485	6806	4444

COMPUTATION AND COMPARISON OF GSI VALUES FOR SIMULATED DATA SET 1

The nine simulated metagenomic samples (belong to the above three groups) were clustered using HabiSign, the average tetranucleotide frequency approach, and the dinucleotide relative abundance based approach. Hierarchical trees were obtained by providing the individual distance matrices (generated using the three methods) as inputs to the neighbor joining algorithm (with default parameters) of the Phylip package. The GSIs for the three different groups (obtained using the three methods) were computed by providing the hierarchical trees generated for each of the individual methods as inputs to the online GSI computation web-server available at <http://www.genealogicalsorting.org/>. A comparison of GSI values obtained for the three different

groups using the three methods is given below

GROUPs	GSI Values obtained using		
	HabiSign	Average Tetranucleotide frequencies	Dinucleotide relative abundances
GROUP1	1	1	1
GROUP2	1	1	1
GROUP3	1	1	1

ANALYSIS

As expected, it is observed that all the three methods obtained a maximum GSI value for the three different groups. These results indicate that all the three methods perform equally when the different groups have entirely different taxonomic composition.

SIMULATED DATA SET 2:

This simulated set contains metagenomic groups having similarity with respect to the constituent species. However, across groups, there are subtle differences with respect to the relative abundance of these species. A method with a high clustering efficiency is expected to capture these subtle differences and appropriately cluster the metagenomic samples belonging to different groups.

Group 1 (consisting of the simulated samples G11, G12 and G13) had all the constituent species present in roughly equal proportions. On the other hand, the Groups 2 and 3 displayed the classic 'long tail' effect wherein the first 9 species were represented in much higher amounts (number of reads > 4000) as compared to the other species (number of reads in the range 200 – 400). This pattern of taxonomic composition is generally observed in real metagenomic data sets.

Furthermore, the metagenomes constituting groups 2 and 3 displayed subtle variations in their taxonomic composition. Metagenomes belonging to Group 2 had a relatively lower abundance of species *Francisella philomiragia*, *Wigglesworthia brevipalpis* and *Staphylococcus aureus* (number of reads ~ 200) as compared to Group 3. On the other hand, metagenomes belonging to Group 3 had a relatively lower abundance of species *Rickettsia bellii*, *Campylobacter jejuni*, *Haemophilus influenzae* and *Clostridium tetani* (number of reads ~ 200) as compared to Group 2.

The details of the three groups constituting this simulated metagenomic data set is given below.

(A) GROUP 1: Consists of metagenomic samples G11, G12 and G13. The relative abundance of the constituent species (in terms of the number of sequences) are also indicated below.

SPECIES	No. of sequences in metagenome		
	G11	G12	G13
<i>Acidovorax avenae</i>	1135	1183	1050
<i>Bordetella pertussis</i>	1104	1138	1066
<i>Pseudomonas aeruginosa</i>	1067	1036	1179
<i>Mycobacterium tuberculosis</i>	1165	1073	1065
<i>Xanthomonas campestris</i>	1051	1022	1055
<i>Rhodopseudomonas palustris</i>	1189	1084	1093
<i>Magnetococcus</i> sp. MC-1	1091	1159	1087
<i>Burkholderia cenocephalia</i>	1126	1111	1190
<i>Agrobacterium tumefaciens</i>	1053	1092	1123
<i>Corynebacterium diphtheriae</i>	1091	1177	1065
<i>Francisella philomiragia</i>	1015	1103	1093
<i>Staphylococcus aureus</i>	1078	1151	1064
<i>Bdellovibrio bacteriovorus</i>	1198	1066	1011
<i>Chlamydomonas pneumoniae</i>	1190	1056	1163
<i>Aquifex aeolicus</i>	1024	1024	1144
<i>Escherichia coli</i>	1028	1124	1104
<i>Thermotoga maritima</i>	1112	1108	1175
<i>Nostoc</i> spp. 7120	1094	1093	1022
<i>Bacillus halodurans</i>	1021	1163	1012
<i>Streptococcus pyogenes</i>	1164	1190	1097
<i>Helicobacter pylori</i>	1126	1143	1023
<i>Rickettsia bellii</i>	1185	1188	1092
<i>Chlorobium chlorochromatii</i>	1184	1167	1178
<i>Wigglesworthia brevipalpis</i>	1035	1124	1006
<i>Salmonella enterica</i>	1000	1179	1137
<i>Mycoplasma pneumoniae</i>	1094	1118	1071
<i>Campylobacter jejuni</i>	1093	1170	1074
<i>Haemophilus influenzae</i>	1044	1112	1078
<i>Clostridium tetani</i>	1044	1179	1058
<i>Neisseria meningitidis</i>	1114	1101	1077

(B) GROUP 2: Consists of metagenomic samples G21, G22 and G23. The relative abundance of the constituent species (in terms of the number of sequences) are also indicated below.

SPECIES	No. of sequences in metagenome		
	G21	G22	G23
Acidovorax avenae	4084	4152	4112
Bordetella pertussis	4008	4031	4144
Pseudomonas aeruginosa	4073	4002	4133
Mycobacterium tuberculosis	4109	4086	4083
Xanthomonas campestris	4164	4096	4021
Rhodopseudomonas palustris	4156	4009	4142
Magnetococcus sp. MC-1	4107	4198	4030
Burkholderia cenocephalia	4037	4065	4080
Agrobacterium tumefaciens	4187	4060	4135
Corynebacterium diphtheriae	403	408	407
<i>Francisella philomiragia</i>	273	227	232
<i>Staphylococcus aureus</i>	220	232	275
Bdellovibrio bacteriovorus	459	418	491
Chlamydophila pneumoniae	449	435	465
Aquifex aeolicus	456	490	447
Escherichia coli	474	466	401
Thermotoga maritima	410	464	405
Nostoc spp. 7120	466	499	412
Bacillus halodurans	463	442	400
Streptococcus pyogenes	422	400	498
Helicobacter pylori	454	475	457
Rickettsia bellii	494	471	474
Chlorobium chlorochromatii	444	434	423
<i>Wigglesworthia brevipalpis</i>	222	234	283
Salmonella enterica	422	401	414
Mycoplasma pneumoniae	463	444	422
Campylobacter jejuni	418	440	405
Haemophilus influenzae	421	407	450
Clostridium tetani	464	496	485
Neisseria meningitidis	402	405	464

(C) GROUP 3: Consists of metagenomic samples G31, G32 and G33. The relative abundance of the constituent species (in terms of the number of sequences) are also indicated below.

SPECIES	No. of sequences in metagenome		
	G21	G22	G23
<i>Acidovorax avenae</i>	4059	4153	4193
<i>Bordetella pertussis</i>	4177	4122	4003
<i>Pseudomonas aeruginosa</i>	4051	4162	4194
<i>Mycobacterium tuberculosis</i>	4138	4138	4049
<i>Xanthomonas campestris</i>	4172	4140	4015
<i>Rhodopseudomonas palustris</i>	4041	4036	4088
<i>Magnetococcus sp. MC-1</i>	4033	4026	4159
<i>Burkholderia cenocephalia</i>	4045	4048	4059
<i>Agrobacterium tumefaciens</i>	4190	4066	4055
<i>Corynebacterium diphtheriae</i>	482	412	401
<i>Francisella philomiragia</i>	496	434	406
<i>Staphylococcus aureus</i>	440	452	419
<i>Bdellovibrio bacteriovorus</i>	495	493	480
<i>Chlamydomonas pneumoniae</i>	422	458	426
<i>Aquifex aeolicus</i>	498	488	414
<i>Escherichia coli</i>	487	456	466
<i>Thermotoga maritima</i>	437	415	420
<i>Nostoc spp. 7120</i>	419	498	401
<i>Bacillus halodurans</i>	439	412	494
<i>Streptococcus pyogenes</i>	451	449	456
<i>Helicobacter pylori</i>	420	422	468
<i>Rickettsia bellii</i>	230	270	202
<i>Chlorobium chlorochromatii</i>	458	498	493
<i>Wigglesworthia brevipalpis</i>	473	423	493
<i>Salmonella enterica</i>	466	485	456
<i>Mycoplasma pneumoniae</i>	457	415	424
<i>Campylobacter jejuni</i>	281	269	256
<i>Haemophilus influenzae</i>	235	294	272
<i>Clostridium tetani</i>	206	241	223
<i>Neisseria meningitidis</i>	415	477	414

COMPUTATION AND COMPARISON OF GSI VALUES FOR SIMULATED DATA SET 2

The nine simulated metagenomic samples (belong to the above three groups) were clustered using HabiSign, the average tetranucleotide frequency approach, and the dinucleotide relative abundance based approach. Hierarchical trees were obtained by providing the individual distance matrices (generated using the three methods) as inputs to the neighbor joining algorithm (with default parameters) of the Phylip package. The GSIs for the three different groups (obtained using the three methods) were computed by providing the hierarchical trees generated for each of the individual methods as inputs to the online GSI computation web-server available at <http://www.genealogicalsorting.org/>. A comparison of GSI values obtained for the three different groups using the three methods is given below

GROUPs	GSI Values obtained using		
	HabiSign	Average Tetranucleotide frequencies	Dinucleotide relative abundances
GROUP1	1	1	1
GROUP2	1	0.56	1
GROUP3	1	0.33	0.4

It was observed that the three different methods attained a GSI value of 1 for Group 1. Moreover, HabiSign is observed to have GSI values of 1 for group 2 and 3 as well. These results indicate that the HabiSign approach is able to efficiently identify the subtle variations in the composition of metagenomes.

In contrast, using the average tetranucleotide frequencies based approach is observed to result in lower GSI values (0.56 and 0.33 respectively) for groups 2 and 3. Similarly, although the dinucleotide relative abundance based approach was able to attain a GSI value of 1 for Group 2, the GSI value obtained for Group 3 was observed to be 0.4.