Supplementary text for manuscript

A. Exponential decay of $SI'(p)$.

The following table demonstrates how we found the exponential decay constant, $\frac{dN}{dt}$.

Table 1: calculating $E(SI)'$ and $E(SI)''$ in different periods of time:

Time periods in units of n/(6k) HGT events	$E(SI)$	Next event SI contribution			$E(SI)'$ (N in the exponential decay terms)	$E(SI)''$ ($\frac{dN}{dt}$ in the exponential decay terms)
		Old neighborhood	new neighborhood	The gene itself		
0	0	$2k$	$2k$	$2k$	$\frac{3}{n}$	--
$\frac{n}{6k}$	$\frac{1}{2k}$	$2k-1$	$2k-1$	$2k$-1	$\frac{3}{n}-\frac{3}{2kn}$	$-\frac{3}{2kn}$
$\frac{2n}{6k}$	$\frac{2}{2k}$	$2k-2$	$2k-2$	$2k$-2	$\frac{3}{n}-\frac{6}{2kn}$	$-\frac{3}{2kn}$
$\frac{mn}{6k}$	$\frac{m}{2k}$	$2k-$m	$2k-$m	$2k$-m	$\frac{3}{n}-\frac{3m}{2kn}$	$-\frac{3}{2kn}$

B. From SI to number of HGT events.

In the main text we present an expression for SI after given number of HGT events ($d = \lambda t$). Here we will show the transposed expression, which gives the expected number of HGT events for a given SI between two species, that we used in our real data evaluations.

We start with the equation

$$\bar{SI}_k(G_1, G_2) = (1 - e^{-\frac{3-\frac{5k}{n-1}}{n}d})\,(1 - \frac{2k}{n-1})$$

Or, for simplicity:

$$\bar{SI}_k = (1 - e^{-\frac{3-\frac{5k}{n-1}}{n}d})\,(1 - \frac{2k}{n-1})$$

After dividing by $1 - \frac{2k}{n-1}$ we get:

$$\frac{\bar{SI}_k}{(1 - \frac{2k}{n-1})} = (1 - e^{-\frac{3-\frac{5k}{n-1}}{n}d})$$

Then,

$$e^{-\frac{3-\frac{5k}{n-1}}{n}d} = 1 - \frac{\bar{SI}_k}{(1 - \frac{2k}{n-1})}$$

$$-\frac{3 - \frac{5k}{n-1}}{n}d = \ln(1 - \frac{\bar{SI}_k}{(1 - \frac{2k}{n-1})})$$

And finally:

$$d = -\frac{n}{3 - \frac{5k}{n-1}}\ln(1 - \frac{\bar{SI}_k}{(1 - \frac{2k}{n-1})})$$

C. Real data analysis-

Using with the equation for inferring the number of HGT events between two genomes based on their SI value, we analyzed a large set of real biological data, the EGGnog database. This database contains protein sequences of 1133 species, most of them bacteria. In addition, this database clusters all proteins into COGs (Clusters of Orthologous Groups). Hence we can represent each organism as a list of its genes which can serve as the input for SI method. This pre-processing stage is widely described in (1), and there we also showed that SI induced 39 native clusters of closely related species, which are much correlated with the conventional genus term. In the following table we present the average SI among each clique, and the average number (as % of the genome size) of HGT events separating between each pair of species in each clique. We found that this parameter is normally distributed (Shapiro-Wilks test: $p = 0.238$) with mean of 52.7%, median of 54.1% and SD of 23.78%. In other words, we found that the number of HGT events separating between each pair of species inside the genus groups is about 50% (±20). This is an interesting finding since SI values themselves (before the transformation to number of HTG events) are not normally distributed (Shapiro-Wilks test: $p = 0.024$).

Table 2: Distribution of SI and the estimated number of HGT events, among closely related species (sharing the same clique). In green- values within the range of 1SD from the mean. In blue values higher more than 1SD from the mean. In yellow, values below 1SD from the mean.

Clique number	Genus list	Avg SI	Clique size	Estimated number of HGT events (average)	Genome size (average)	Number of HGT events as % of genome size
1	{'Borrelia'}	0.554	8	320.0	1158.8	27.6
2	{'Burkholderia', 'Ralstonia', 'Cupriavidus'}	0.819	25	3657.1	6367.9	57.4
3	{'Pelodictyon', 'Chlorobaculum', 'Chlorobium', 'Prosthecochloris'}	0.851	10	1478.8	2271.3	65.1
4	{'Shewanella'}	0.764	19	2072.1	4262.0	48.6
5	{'Streptococcus'}	0.856	10	1331.7	2000.9	66.6
6	{'Rickettsia'}	0.654	13	434.3	1192.7	36.4
7	{'Methanococcus'}	0.632	6	584.6	1721.0	34.0
8	{'Exiguobacterium', 'Oceanobacillus', 'Macrococcus', 'Bacillus', 'Geobacillus', 'Anoxybacillus', 'Staphylococcus', 'Listeria'}	0.877	25	2283.6	3202.6	71.3
9	{'Streptococcus', 'unknow'}	0.713	14	863.8	2037.3	42.4
10	{'Corynebacterium', 'Mycobacterium'}	0.86	10	1568.4	2331.7	67.3
11	{'Thermotoga'}	0.501	6	439.3	1867.8	23.5
12	{'Bartonella', 'Brucella'}	0.669	15	961.1	2572.6	37.4
13	{'Rhodopseudomonas', 'Nitrobacter', 'Bradyrhizobium', 'Oligotropha'}	0.855	11	3222.2	4947.1	65.1
14	{'Mycobacterium'}	0.817	19	2761.3	4827.3	57.2
15	{'Francisella'}	0.608	9	518.0	1627.3	31.8
16	{'Staphylococcus'}	0.226	12	228.0	2647.8	8.6
17	{'Shigella', 'Cronobacter', 'Serratia', 'Photorhabdus', 'Pectobacterium', 'Citrobacter', 'Klebsiella', 'Salmonella', 'Dickeya', 'Erwinia', 'Sodalis', 'Yersinia', 'Edwardsiella', 'Escherichia', 'Proteus', 'Enterobacter'}	0.796	85	2405.4	4492.3	53.5
18	{'Candidatus', 'Buchnera'}	0.938	8	605.9	486.0	124.7
19	{'Azotobacter', 'Pseudomonas'}	0.826	18	3167.8	5382.3	58.9
20	{'Clostridium'}	0.749	13	1630.0	3496.0	46.6
21	{'Photobacterium', 'Vibrio', 'Aliivibrio'}	0.797	14	2369.4	4410.4	53.7
22	{'Chlamydia', 'Chlamydophila', 'unknow'}	0.444	14	194.2	966.0	20.1
23	{'Lactobacillus', 'Pediococcus'}	0.884	12	1574.2	2122.0	74.2
24	{'Xanthomonas', 'Stenotrophomonas'}	0.717	10	1864.1	4391.6	42.4
25	{'Desulfotomaculum', 'Candidatus', 'Carboxydothermus', 'Pelotomaculum', 'Moorella', 'Ammonifex'}	0.923	6	2172.2	2447.6	88.7
26	{'Neisseria'}	0.524	7	518.1	2065.3	25.1
27	{'Agrobacterium', 'Rhizobium', 'Sinorhizobium', 'Ochrobactrum'}	0.866	12	4149.9	6132.2	67.7
28	{'Acidovorax', 'Variovorax', 'Delftia', 'Comamonas'}	0.88	6	3517.6	4910.2	71.6
29	{'Prochlorococcus', 'Synechococcus'}	0.706	16	904.7	2179.1	41.5
30	{'Bifidobacterium'}	0.787	8	981.5	1857.1	52.8
31	{'Bacillus', 'Geobacillus'}	0.612	15	1687.6	5315.8	31.7
32	{'Streptococcus'}	0.61	19	604.3	1893.1	31.9
33	{'Helicobacter'}	0.528	8	390.5	1531.4	25.5
34	{'Acinetobacter'}	0.616	7	1121.1	3481.7	32.2
35	{'Ehrlichia', 'Anaplasma'}	0.679	8	390.0	992.9	39.3
36	{'Geobacter'}	0.892	6	2926.0	3871.7	75.6
37	{'Campylobacter'}	0.76	8	839.4	1721.1	48.8
38	{'Methylobacterium'}	0.676	6	2147.7	5681.5	37.8
39	{'Sulfolobus'}	0.453	6	559.6	2756.5	20.3

D. Phylogenetic analysis.

In (1) we presented phylogenetic trees of closely related species based on solely the SI measure. Using the corrected measure developed in this work - the expected number of HGT events between pair of species – we can compare between the two measures. While this concerns only real data, and hence cannot be accurately validated, we believe the new corrected measure is more accurate than the crude SI. In table 3 we depict all the cliques from (1) and for each such clique the Robinson-Foulds (RF) symmetric difference (2) between the two types of trees for this cluster. Although some differences are observed, the two trees are mostly similar (average normalized RF 0.145, median 0.125, SD 0.159) in both approaches.

Table 3: Comparison between trees which generated based on SI to trees generated based on the distance measure developed here, of all the cliques of (1). Here we calculated the RF (Robinson–Foulds) measure between trees in order to evaluate the difference between them.

Clique Number	RF	Tree size	Normalized RF
1	0	8	0.000
2	4	25	0.091
3	2	10	0.143
4	4	17	0.143
5	0	10	0.000
6	4	13	0.200
7	2	6	0.333
8	2	25	0.045
9	4	14	0.182
10	2	10	0.143
11	4	6	0.667
12	10	14	0.455
13	2	11	0.125
14	12	19	0.375
15	0	8	0.000
16	0	12	0.000
17	26	82	0.165
18	0	7	0.000
19	2	17	0.071
20	4	13	0.200
21	0	14	0.000
22	2	14	0.091
23	0	12	0.000
24	2	10	0.143
25	0	5	0.000
26	2	7	0.250
27	2	12	0.111
28	0	6	0.000
29	10	16	0.385
30	0	7	0.000
31	0	14	0.000
32	12	18	0.400
33	2	8	0.200
34	0	7	0.000
35	2	8	0.200
36	0	6	0.000
37	2	8	0.200
38	2	6	0.333
39	0	6	0.000

Bibliography

1. **Sevillya, Gur and Snir, Sagi.** Synteny Footprints Provide Clearer Phylogenetic Signal than Sequence Data for Prokayotic Classi. *Molecular Phylogenetics and Evolution.* 2019.

2. **Robinson, D. R. and Foulds, L. R.** Comparison of phylogenetic trees. *Mathematical Biosciences.* 1981.