Sevillya *et al. BMC Genomics* 2020, **21**(Suppl 1):106
https://doi.org/10.1186/s12864-019-6395-5

BMC Genomics

RESEARCH Open Access

Detecting horizontal gene transfer: a probabilistic approach

Gur Sevillya, Orit Adato and Sagi Snir*

From 14th International Symposium on Bionformatics Research and Applications (ISBRA'18)
Beijing, China. 8–11 June 2018

Abstract

Background: Horizontal gene transfer (HGT) is the event of a DNA sequence being transferred between species not by inheritance. HGT is a crucial factor in prokaryotic evolution and is a significant source for genomic novelty resulting in antibiotic resistance or the outbreak of virulent strains. Detection of HGT and the mechanisms responsible and enabling it, is hence of prime importance.
Existing algorithms rely on a strong phylogenetic signal distinguishing the transferred sequence from its recipient genome. Closely related species pose an even greater challenge as most genes are very similar and therefore, the phylogenetic signal is weak anyhow. Notwithstanding, the importance of detecting HGT between such organisms is extremely high for the role of HGT in the emergence of new highly virulent strains.

Results: In a recent work we devised a novel technique that relies on loss of synteny around a gene as a witness for HGT. We used a novel heuristic for synteny measurement, SI (Syntent Index), and the technique was tested on both simulated and real data and was found to provide a greater sensitivity than other HGT techniques. This synteny–based approach suffers low specificity, in particular more closely related species. Here we devise an adaptive approach to cope with this by varying the criteria according to species distance. The new approach is doubly adaptive as it also considers the lengths of the genes being transferred. In particular, we use *Chernoff* bound to decree HGT both in simulations and real bacterial genomes taken from EggNog database.

Conclusions: Here we show empirically that this approach is more conservative than the previous χ^2 based approach and provides a lower false positive rate, especially for closely related species and under wide range of genome parameters.

Keywords: Gene order, Horizontal gene transfer, Phylogenetics

Background

Genomes of bacteria and archaea are characterized by extensive gene mobility between species that is crucial not only for evolution of genome architecture but also for the functionality of prokaryotic organisms [12]. The principal mechanism accounting for gene mobility is horizontal gene transfer (HGT) [6, 14, 18] in which a gene (or a group of genes) of a donor species being acquired by a recipient organism. HGT, to a large extent, is mediated by viruses (bacteriophages), plasmids, transposons and other mobile elements. The genetic interpretation of this event is a gene being copied from the donor genome to the recipient genome (see Fig. 1).

The study of the HGT is of paramount importance for several reasons. First, from medical perspective, HGT plays a major role in the emergence of new human diseases, as well as promoting the spread of antibiotic resistance in bacteria species [21]. From the fundamental, evolutionary standpoint, HGT links distant branches in the tree of life, turning it into an evolutionary network [6, 30]. Genetically, HGT is an important, if not the primary source of new genes that are acquired by bacteria and archaea and often result in adaptations to new environments and conditions [5]. Recent advances of comparative genomics and especially metagenomics indicate that the

*Correspondence: ssagi@research.haifa.ac.il
[1]Dept. of Evolutionary and Environmental Biology, University of Haifa, 3498838 Haifa, Israel

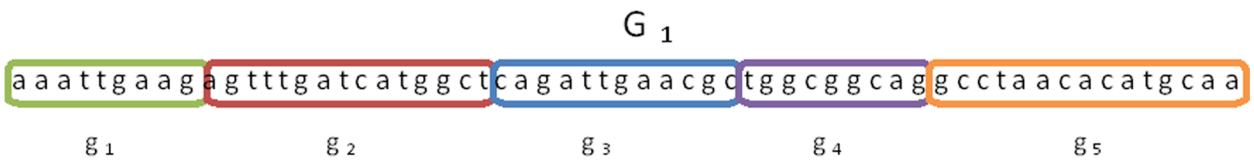

Fig. 1 A genome is viewed as a sequence of genes while a gene is a sequence of nucleotides

complexity of the genetic material that is horizontally transferred is vast and exceeds by several orders of magnitude the complexity of the set of conserved genes that are mostly vertically inherited [7]. Therefore, identification of HGT can shed light on many significant evolutionary processes that cannot be explained by the traditional tree–like approach.

Currently, there are two prevailing methods for detecting HGT. The *phylogeny based approach* takes a relatively large set of homologous (originated from a common ancestor) coding sequences, constructs their corresponding phylogeny, and contrasts it to the phylogeny of their originating species. When conflicts are found between the two trees, they are reconciled by introducing HGTs (see e.g. [13, 17]). While this approach has the advantage of identifying relatively old events, the approach is based on a very stringent assumption of where to seek the events. Finally, it also requires a *multiple* alignment of the sequences and inferring a reliable species tree (two major problems by themselves [31]). The *composition based approach* contrasts genomic sequences of different compositional structure such as G+C content, dinucleotide frequencies or codon usage biases, striving to infer different origins (e.g. [3, 11, 18, 20]). This approach suffers from the fact that the species involved might share similar compositional patterns. Moreover, the length of a transferred segment may be too short to reliably reveal these differences. As concluded in [16], "atypical G+C content and pattern of codon usage are not reliable indicators of horizontal gene transfer events".

Both the phylogenetic and the sequence based approaches rely on a strong enough signal for the HGT. Such a signal may not exists when dealing with closely related species or even strains of the same species. In two recent works [1, 24], we have defined the notion of *synteny index* (SI) between two genomes (species) and used it as a marker of evolutionary footprints. Gene synteny [8, 23] is the conservation of gene order across species along the evolutionary course. Synteny (or lack of) was already employed for defining a distance measure between genomes (species), counting the minimal number of operations to transform one genome to another [4]. Nevertheless this distance is irrelevant in the context of a particular single gene. In contrast, SI measures how much a gene, orthologous to the two species, is in its "natural place", or in other words, shares the same neighborhood in both genomes. In [24] we averaged the SI over the whole genome and used it to infer evolutionary distance. We next aimed at identifying HGTs between closely related species by means of SI [1]. The technique relies on the *constant relative mutability* (CRM) property that asserts that the ratio between the mutational rates of genes is maintained across species (even if the rates themselves change along time/species). The method was compared to several representative HGT methods, both phylogenetic methods such as *RIATA–HGT* and *PhylTR* [19, 27], and sequence based– *HGT–DB* [11]. It was also employed to real biological data, the three strains of *E. Coli* that were studied in [28] and were found to exhibit a very high rate of HGT. Understanding and detecting HGT within the strains, could be of great importance, for instance in understanding the origin of pathogenicity of certain pathogenic strains, particularly those whose ancestors were not pathogenic.

In this work we make another step forward by formulating the problem in a statistical framework. This allows us to apply probabilistic tools that are more advanced than those employed in [1]. We start by providing a tool to measure the significance of SI of a given gene over the background noise of its hosting genomes. Next we apply bounds on large deviations to assess the probability of a gene being transfered to the hosting genome or it has existed there since divergence from the donor species. This check requires the transformation between two spaces: one, in which the CRM exists and allows us to derive expected values for the gene distance, and another, in which we can assess the likelihood of the observed distance with respect to the expected one. We conducted a simulation study where we compared the current approach with that of [1] and showed it provides a greater specificity (i.e., lower false positive rate). All steps of the proposed method are very efficient as they operate between pairs of orthologous genes and therefore the complexity of the method is at most quadratic in the size of the taxa set.

Definitions and methods

We now define our working model that will serve to locate HGT between genes. A genome G is a sequence (although we sometimes treat it as an ordered set) of genes $(g_1, g_2, \ldots, g_n)$ and each gene is a sequence of DNA letters. That is, our view of a genome is at a resolution of genes,

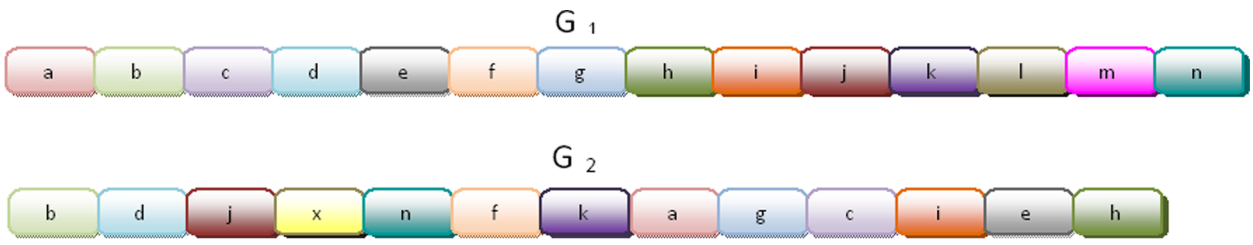

Fig. 2 Comparing G_1 with G_2 for $k = 3$: $SI(g, G_1, G_2) = 3$, $SI(x, G_1, G_2) = 0$, $SI(\ell, G_1, G_2) = 0$

and of a gene at a resolution of nucleotides (See Fig. 2). The *k–neighborhood* of a gene g_0 in genome G, $N_k(G, g_0)$ is the set of genes at distance at most k from g_0 in G (i.e. at most k genes upstream or downstream, not including gene g_0 itself). The *core set* (i.e., intersection) of genomes G_1 and G_2 is $G_1 \cap G_2$, and the symmetric difference of G_1 and G_2 is $G_1 \triangle G_2 = G_1 \cup G_2 \setminus G_1 \cap G_2$.

The conservation of the order between genes in $\cap_{G_1,G_2}$ is called *synteny*. Let $g_0 \in \cap_{G_1,G_2}$. Then the k *synteny index* (k-SI), or just SI when it is clear from the context, of g_0 in G_i, G_j is the number of common genes in the k neighborhoods of g_0 in both G_i and G_j: $SI(g_0, G_i, G_j) = |N_k(G_i, g_0) \cap N_k(G_j, g_0)|$. For the sake of completeness, for $g_0 \notin G_i \cap G_j$, $SI(g_0, G_i, G_j) = 0$. See Fig. 2 for illustration.

A genome undergoes events of gene gain and loss in which genes are added or removed respectively. As we are focused in the core set of genes that are common to two organisms, we are not interested in the latter processes. Every gene undergoes a process of sequence evolution according to some stochastic evolutionary model [9]. The evolutionary model we consider is such that the nucleotides along a gene are identically and independently distributed (IID). The value of the nucleotide is the *state* (we sometimes use just "nucleotide" to denote its state). A *single mutation* (or *point mutation* or just a mutation for short) is the event of a nucleotide changing its value to a different one (for reasons of simplicity, we use the term 'mutation' as for a point mutation that occurs and then gets fixed, i.e., 'a nucleotide substitution'). An *evolutionary model* $\mathcal{M}$ models the (stochastic) process of mutations occurring at a site as a function of *mutation rates* $\dashv_{i,j}$ modeling the rate of transitions from state i to j, and a specified time period t. We use the *transition* notation in the context of Markov chains and note that it has nothing to do with the type of mutation baring the same notation (see [9] for more details). Given $\mathcal{M}$, mutation rates $[\dashv_{i,j}]$, and a time period t, the *transition probability* $p_{i,j}$ from nucleotide i to j during t is uniquely defined by an appropriate function (determined by $\mathcal{M}$). An evolutionary model $\mathcal{M}$ is said to be *time reversible* if it is not possible to determine the direction of time given two states of a nucleotide, separated by a time period t. The *evolutionary distance* (or *mutation distance* or simply distance) is the number of mutations separating between two homologous sequences. The *hamming distance* between two homologous sequences counts the number sites with different states. These distances are usually normalized by the length of the sequences and are normally denoted by d and h respectively. In the Results section, we used the Jukes–Cantor [15] (JC) evolutionary model. See more specific details in "Results" section.

A *horizontal gene transfer* (HGT) is the event in which a gene of a genome, the *donor genome*, being copied and inserted at some position at another genome, the *recipient genome*. Since we view the genome as a circled sequence of genes, the new gene is always between two genes.

Results

Consider two genomes after speciation event. Gene order, synteny, in the two genomes is nearly the same, and hence orthologous genes have almost the same neighboring genes in the two genomes. Due to events such as HGT, this similarity decreases as the time since the divergence event grows. Hence, between closely related species (and in particular strains of a species), if a gene has exceptionally low SI, we might suspect it has undergone HGT. We denote these genes as *SI HGT suspected*.

Significant SI HGT suspected genes

We want to verify that SI suspected genes are indeed a result of a HGT and not a background noise. When the core set of genes is small, with some probability, low SI is observed even if a gene is in its original location. This is due to gene loss events around that gene. If all genes around gene g were lost, g has SI zero without being transferred. We will associate a confidence value with every SI, and set a threshold value δ_{SI} for obtaining low SI by random (i.e. not by HGT rather simply by gene gain/loss).

Lemma 1 *Consider two genomes G_1 and G_2. Let g be a gene in the core set of G_1 and G_2 ($g \in \cap_{G_1,G_2}$) with $|G_1| = |G_2| = n$ and let δ_{SI} be an arbitrary probability. Then with probability at most δ_{SI} we expect to find by chance SI of size:*

$$SI < \frac{k}{n}(2n - |G_1 \triangle G_2|) - \sqrt{-k \log_e \delta_{SI}} \quad (1)$$

Proof We denote a gene g_i as *singular* if $g_i \in G_1 \triangle G_2$.

Observation *The probability of hitting a singular gene by chance is* $\frac{|G_1 \triangle G_2|}{2n}$. □

Proof Since by assumption the length of both genomes is n, the symmetric difference $G_1 \triangle G_2$ is partitioned equally on both genomes (since the core set $\cap_{G_1,G_2}$ exists on both). Since a randomly chosen gene from a given genome is either from $\cap_{G_1,G_2}$ or $G_1 \triangle G_2$ and $n = |\cap_{G_1,G_2}| + \frac{|G_1 \triangle G_2|}{2}$, the result follows. □

Henceforth we will denote by p this probability, i.e. $p = \frac{|G_1 \triangle G_2|}{2n}$, and note that p is easily calculated.

We now focus on genes g_i for $1 \leq i \leq 2k$ in the k–neighborhood of gene g. Let $X_i = 1-p$ if gene g_i is singular, and $-p$ otherwise, and let $X = \sum X_i$. Then we observe that $\Pr[X_i = 1-p] = p$ and $\Pr[X_i = -p] = 1-p$. Hence, $E[X_i] = 0$ and X follows a distribution $B(2k,p) - 2kp$ where $B(n,p)$ is the usual binomial distribution (this is a good approximation given the reasonable assumption that $k \ll n$).

Our goal is to bound the probability of deviation from the expected value and seeing a low SI only by random. The distribution of X allows us to apply Chernoff bound [2](Thm A.4) asserting

$$\Pr[X > a] \leq e^{-2a^2/n}. \tag{2}$$

We are seeking for the minimal a such that this probability is smaller than δ_{SI}. In our case n is $2k$ and hence we set

$$e^{-2a^2/2k} = \delta_{SI} \Rightarrow a = \sqrt{-k \log_e \delta_{SI}}, \tag{3}$$

yielding:

$$\Pr\left(X > \sqrt{-k \log_e \delta_{SI}}\right) \leq \delta_{SI}. \tag{4}$$

Note that X counts the observed number of genes in only a single k–neighborhood of g (they are *not* necessarily singular) minus their expected number– $2kp$. That is $X = 2k - SI - 2kp$, where $2k - -SI$ is the number of genes in only a single k–neighborhood of g.

If we substitute in (4) $X = 2k - SI - 2kp$ we obtain:

$$\begin{aligned} &\Pr\left(2k - SI - 2kp > \sqrt{-k \log_e \delta_{SI}}\right) \\ &= \Pr\left(SI < 2k(1-p) - \sqrt{-k \log_e \delta_{SI}}\right) < \delta_{SI}. \end{aligned} \tag{5}$$

Back substituting $p = \frac{|G_1 \triangle G_2|}{2n}$, and the result follows.

Lemma 1 allows us to infer about the increase in the strength of the evidence. For that we equate a in Eq. (2) to X. As n in Eq. (2) is set constant, the only variable component is the SI in X (and in a) yielding the following:

Corollary 1 *The significance of the evidence grows exponentially in the SI.*

Lemma 1, provides an upper bound on the SI scores we expect to see by chance. However, we should be careful here. For some combinations of p and δ_{SI} our neighborhood may not be large enough. For instance, for $|G_1 \triangle G_2| = 1600$, $n = 1000$, and $\delta_{SI} = 0.05$, a neighborhood of 10 is not enough, since by our bound, under that probability we expect to see $SI < -1.47$ by chance, but $k = 30$ does suffice ($SI < 2.5$). If we increase $|G_1 \triangle G_2|$ to 1800 (i.e. $p = 0.9$), then even $k = 60$ is not enough. We therefore conclude:

Corollary 2 *Let* $p = \frac{|G_1 \triangle G_2|}{2n}$, *then for a given* δ_{SI} *we must have*

$$k \geq -\frac{\log_e \delta_{SI}}{4(1-p)^2} \tag{6}$$

Proof Since $SI \geq 0$ must hold, and according to Lemma 1, we get: $0 \geq \frac{k}{n}(2n - |G_1 \triangle G_2|) - \sqrt{-k \log_e \delta_{SI}}$. Then, $\sqrt{-k \log_e \delta_{SI}} \geq \frac{k}{n}(2n - |G_1 \triangle G_2|)$. We defined before $p = \frac{|G_1 \triangle G_2|}{2n}$, so we get $\sqrt{-k \log_e \delta_{SI}} \geq 2k(1-p)$ and isolation of k is trivial. □

Sifting between other mutational events

In the previous section we derived values under which SI is significant. However, low SI can be a result of other large scale mutational events: A *translocation* is the event where a gene moves to a different location in a genome. A *Duplication* is an identical event only that a copy of the gene remains in the original location.

The following observation follows intuitively from Fig. 2:

Observation 2 *Let* G_1 *and* G_2 *be two genomes sharing a common gene* g. *Assume* g *was either translocated or duplicated in* G_2 *(we assume* g *corresponds to the copied instance rather than the original). Assuming no other large scale mutational events occurred, then* $E[SI(g, G_1, G_2)] \approx 0$.

Proof Amuse G_1 and G_2 are two identical genomes, and now gene g is translocated in genome G_1. Then, $E[SI(g, G_1, G_2)] = 0$ except if the new position of gene g is no further than k from its original neighborhood. The probability the new position fulfill this requirement is $4k/n$. For realistic closely related genomes (genome size

of 5000 and symmetric difference < 0.8, which leads to $k < 10$), we get $E[\,SI(g, G_1, G_2)] \approx 0$. □

Indeed, based on SI only, it cannot be distinguished whether gene d in Fig. 2 has been horizontally transferred or been translocated. Therefore we cannot rely on low SI as a single evidence for HGT. To distinguish a gene undergone HGT from translocations or duplications, we rely on the fact that a translocated (duplicated) gene has been in its hosting genome since its split from another genome, in contrast to a gene recently acquired through HGT. This implies that the translocated gene was subjected to small scale substitutions (point mutation) for the time period since its split from the other genome. Hence the induced distance between orthologous genes in two genomes, is proportional to the time since their divergence.

Constant relative mutability

We now rely on a very basic evolutionary effect recently demonstrated, dubbed as *Universal Pacemaker* (UPM) of genome evolution [25, 26, 29], which serves as a useful approximation for genome evolution processes. The UPM principle states that along every lineage in the evolution of cellular life, most genes change their mutation rate in unison, as if adhering to a universal (but lineage specific) pacemaker. To harness the UPM principle to our purpose, we formulate the problem as follows: We have a gene g, *SI suspected* of having undergone HGT between two strains S_1 and S_2, by exhibiting low $SI(g, S_1, S_2) < \delta_{SI}$ for some threshold value δ_{SI}. We look for a *witness* gene, w, and two *reference* organisms R_1 and R_2 under the constraint that $w \in R_1, R_2, S_1, S_2$ and $g \in R_1, R_2$. By the UPM, genes g and w, although may mutate at different rates, maintain approximately a constant ratio between their rates. More precisely:

Definition 1 *(CRM) Let g and g′ be genes residing in a genome G mutating at (not necessarily constant) rates a and a′ . Then g and g′ have constant relative mutability (CRM) (or alternatively – conservation), if at any time, the ratio $\rho = a/a'$ is (approximately) constant.*

The result of the CRM phenomenon is that different genes, unless undergone gene specific extraordinary events, maintain the same tree topology and even tree shape. This implies that branch lengths in the corresponding gene trees differ by a multiplicative constant. Note that this property does not contradict rate heterogeneity across genes and also across organisms.

In the following, we operate in two distance spaces: The *Hamming distance* and the *mutation distance* (or *evolutionary distance* or simply, distance). The distances are always defined between two organisms and WRT a certain gene (e.g. X_1, X_2, g respectively): $h_g(X_1, X_2)$ or $d_g(X_1, X_2)$. When it is clear from the context, we ignore either the gene or the two organisms associated with the distance. In the evolutionary distance space, the distance between organisms is the number of mutations separating them, or alternatively, the number of mutations occurred at each lineage since their divergence event.

The Hamming distance is the number of different positions between the genes at the organisms and it is an underestimate for the mutation distance since multiple mutations at a site are unobserved.

To obtain the *expected* mutation distance (we don't know exactly how many mutations indeed occurred) we use some non–linear *distance correction* function, $d = corr(h)$ and an inverse correction $h = corr^{-1}(d)$.

The definition of the CRM phenomenon, operates on the substitution rates of each gene. We however, observe the Hamming distances. To use the CRM phenomenon, we need to convert the Hamming distances to evolutionary distances and then to apply the CRM rule.

Observation 3 *Assume genes g and g′, with mutation rates r_g and $r_{g'}$ respectively, satisfy the CRM hypothesis with ratio $\rho = \frac{r_g}{r_{g'}}$. Let h_g be the Hamming distance WRT gene g. Then the expected distance WRT gene g′, $\delta_{g'}$ is*

$$\delta_{g'} = \frac{corr(h_g)}{\rho} \tag{7}$$

where corr is a distance correction function to correct from the observed Hamming distance to the mutation distance.

Proof The expected number of substitutions along gene g between two sequences X_1, X_2, i.e. the real distance is defined by $d_g = corr(h_g(X_1, X_2))$. On the other hand, $d_g = tr_g$ where t is the time separating X_1, X_2, or in other words, twice the time since divergence. Since, by the CRM property $r_{g'} = \frac{r_g}{\rho}$, we obtain

$$\delta_{g'}(X_1, X_2) = tr_{g'} = \frac{tr_g}{\rho} = \frac{d_g(X_1, X_2)}{\rho} = \frac{corr(h_g(X_1, X_2))}{\rho} \tag{8}$$

□

Observation 3 derives the expected distance of a gene g' based on the CRM hypothesis and the hamming distance of another gene g. If we apply the inverse correction $corr^{-1}$ to the expected distance $d_{g'}$ we obtain the *expected Hamming distance* $h_{g'}$. This is essential since in the hamming distance space we can apply bounds on deviations from the mean that do not apply in the mutation distance space. Therefore, in order to link between the expected and the observed distance WRT gene g, we use the following Lemma:

Lemma 2 *Assume genes g and g', adhering the CRM hypothesis. Let $d_{g'}$ be the expected distance WRT to g' as derived in Observation 3. Let $h_{g'}$ be the expected Hamming distance obtained by applying the inverse correction on $d_{g'}$: $h_{g'} = corr^{-1}(d_{g'})$. Then, the difference between $h_{g'}$ and the observed Hamming distance $\hat{h}_{g'}$ satisfies:*

$$\Pr\left[\left|h_{g'} - \hat{h}_{g'}\right| > \varepsilon\right] \leq 2e^{-2n\varepsilon^2}, \tag{9}$$

where n is the length a number of nucleotides of g'.

Proof The expected Hamming distance is the probability p of observing a difference at a position between the two corresponding sequences. Let $\hat{h}_i$ be an indicator variable indicating a difference at position i at both copies of g'. By definition of $\hat{h}_{g'}$,

$$\hat{h}_{g'} = \frac{1}{n}\sum_i \hat{h}_i$$

and

$$E\left[\hat{h}_{g'}\right] = \frac{1}{n}E\left[\sum_i \hat{h}_i\right] = \frac{1}{n}\sum_i E\left[\hat{h}_i\right] = \frac{1}{n}np = h_g,$$

where the second equation is due to linearity of expectation.

Letting $X = \sum_i \hat{h}_i$, we can use again Chernoff inequality [2] to bound the deviation of a sum of IID indicator random variables from its mean. For any $\varepsilon > 0$ holds:

$$\Pr[\,|X - E[X]\,| > \varepsilon n] \leq 2e^{-2(\varepsilon n)^2/n}, \tag{10}$$

and the result follows.

□

Observation 3 and Lemma 2 relied on the CRM phenomenon and the constant ratio ρ to derive the expected distances and to bound the deviation from them. Since gene g is HGT suspected between the two strains, we cannot rely on its distance (Hamming and consequently mutation) to adhere to CRM. Therefore, we look for two *reference organisms* and a *witness* gene w, such that both g and w are present in the strains and the reference organisms. Now we can compute $\rho = r_g/r_w$ in the reference organisms, and use it to derive the expected distance between the strains WRT g. We now state our main theorem for this section that combines all this information with Observation 3 and Lemma 2 to obtain some confidence level on the observed Hamming distance $\hat{h}_g$ between the strains as a function of the distances between the reference organisms.

Theorem 1 *Let g be a gene suspected of having undergone HGT between two strain species S_1 and S_2 and let $n = |g|$ be the length of g. Let w be a witness gene while R_1 and R_2 are two different reference organisms. Finally, let $\hat{h}_g(R)$ and $\hat{h}_g(S)$ be the (observed) hamming distance WRT gene g between the reference organisms (R_1 and R_2) and between the strains (S_1 and S_2), respectively. Similarly, let $\hat{h}_w(R)$ and $\hat{h}_w(S)$ be the hamming distance between the reference organisms WRT the witness gene w. Then the probability of observing $\hat{h}_g(S)$ given $\hat{h}_g(R)$, $\hat{h}_w(R)$, $\hat{h}_w(S)$ and n assuming CRM hypothesis is:*

$$Pr\left(\left|\hat{h}_g(s) - corr^{-1}(d_g(S))\right| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}, \tag{11}$$

where, $d_g(S)$ is the expected distance between the strain species S_1 and S_2 given by

$$d_g(S) = \frac{corr\left(\hat{h}_g(R)\right)}{corr\left(\hat{h}_w(R)\right)} corr\left(\hat{h}_w(S)\right)$$

Proof We first use $\hat{h}_g(R))$ and $\hat{h}_w(R)$ to compute $\rho = \frac{corr(\hat{h}_g(R))}{corr(\hat{h}_w(R))}$ and then by Observation 3 we obtain expected distance between the strains $d_g(S)$. Finally, by using the inverse correction on $d_g(S)$ we can use Lemma 2 to bound the probability of the deviation of the observed Hamming distance between the strains, $h_g(s)$, from the expected one. □

Using Theorem 1 we can find a cutoff value for the difference between the expected and observed Hamming distances for any given confidence level δ_r:

Corollary 3 *For a given confidence value δ_r, we can refute the null hypothesis, i.e., that gene g has evolved vertically, if the difference $\left|h_g(s) - corr^{-1}(d_g)\right|$ satisfies:*

$$|h_g(s) - corr^{-1}(d_g)| > \varepsilon(\delta_r) = \sqrt{\frac{-\log_e \delta_r/2}{2n}} \tag{12}$$

Proof Equation (11) gives the probability for a deviation from the expected distance d_g. The bigger the deviation, the smaller is its probability. Therefore we can calculate the minimum deviation $\varepsilon(\delta_r)$ with probability at most the threshold value δ_r, and refute the null hypothesis for any bigger deviation. □

Figure 3 illustrates the situation. The following example applies Theorem 1 under the Jukes–Cantor [15] (JC) evolutionary model, on real data from the *E. coli* strains *CFT073* and *MG1655* [1].

Example 1 *To illustrate the use of Theorem 1 we show some real data example.*

The evolutionary model with which we work is the Jukes–Cantor [15] (JC) evolutionary model. The JC model is a

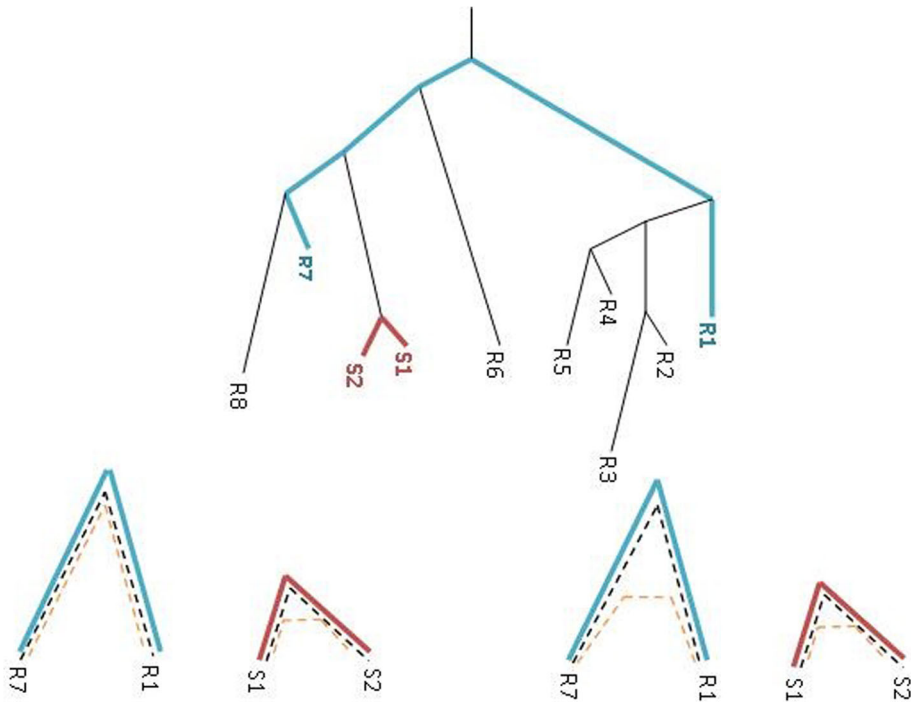

Fig. 3 The phylogeny over a group of organisms with branch lengths proportional to distances of gene g_h. g_h has undergone HGT between the two strains S_1 and S_2 and hence their distance is very short compared with two reference organisms R_1 and R_7. Bottom: The reference gene (blue, dashed line) must be a gene that accumulates mutations ever since the divergence of both the strains and reference organisms. There are two cases in which the suspicious gene evolves at the reference organism. (*A*) No HGT and then the constant relative conservation is maintained. (*B*) HGT of the SI suspicious gene at the reference organisms and the constant relative conservation is not maintained

reversible, one parameter model, postulating that at any position the rate of substitutions from one state to another, $a_{i,j}$ is the same – a. Under this model, the expected number of substitutions – that is, the evolutionary distance d_{JC} – at a site during t time units is $d_{JC} = 3at$. To obtain the distance from a given Hamming distance h,we apply the distance correction for the JC model:

$$d_{JC} = corr_{JC}(h) = -\frac{3}{4}\log_e\left(1 - \frac{4}{3}h\right), \tag{13}$$

for a given normalized Hamming distance h. Note that under this correction, $h < \frac{3}{4}$ must hold.

Inversely, the expected Hamming distance is:

$$h = corr_{JC}^{-1}(d_{JC}) = \frac{3}{4}\left(1 - e^{-\frac{4}{3}d_{JC}}\right). \tag{14}$$

Let the two strains S_1 and S_2 be the E. coli strains CFT073 and MG1655 and the reference organisms, R_1 and R_2, be Bacteroides.fragilis and Wolbachia. The HGT suspected gene is engA and the witness gene is gmk. The Hamming distances obtained are: $h_g(s) \approx 0.0237$, $h_g(r) \approx 0.583$, $h_w(r) \approx 0.541$, $h_w(s) \approx 0.008$ and the average genome size is 4743. We get:

- $h_g(s) = 0.0237$
- $h_g(r) = 0.583$
- $h_w(r) = 0.541$
- $h_w(s) = 0.008$
- $n = 1472$.

we get:

$$\begin{aligned} d_g(s) &= \frac{corr_{JC}(h_g(r))}{corr_{JC}(h_w(r))} corr_{JC}(h_w(s)) \\ &= \frac{corr_{JC}(0.583)}{corr_{JC}(0.541)} corr_{JC}(0.008) \\ &= \frac{-\frac{3}{4}\log_e\left(1 - \frac{4}{3}\cdot 0.583\right)}{-\frac{3}{4}\log_e\left(1 - \frac{4}{3}\cdot 0.541\right)} * \frac{3}{4}\log_e\left(1 - \frac{4}{3}\cdot 0.008\right) \approx 0.0095. \end{aligned}$$

Hence:

$$\begin{aligned} h_g(s) - corr_{jc}^{-1}(d_g(s)) &= 0.0237 - \left(\frac{3}{4}\left(1 - e^{-\frac{4}{3}\cdot 0.0095}\right)\right) \\ &= 0.0237 - 0.0094 \approx 0.0142 \end{aligned}$$

and the probability to see a difference greater than 0.0142 is

$$Pr(|h_g(s) - corr^{-1}(d_g)| > 0.0142) \leq 2e^{-2n\cdot 0.0142^2} \approx 0.85.$$

We can see that for such a difference, the null hypothesis is not rejected and we cannot refute by gene gmk that gene engA evolved vertically.

Alternatively, if we set $\delta_r = 0.05$ *then by Eq. (12) we get a cutoff distances* $\varepsilon(0.05)$

$$\varepsilon(0.05) = \sqrt{\frac{-\log_e 0.05/2}{2 \cdot 1472}} \approx 0.035,$$

which is greater than the value 0.0121 practically obtained.

Theorem 1 gives the tail probability for all events with $h'_g(s)$, such that $h'_g(s) \geq h_g(s)$. We refute the null hypothesis (that the gene has evolved vertically between the strains) if the probability is below some threshold value δ_r.

We conclude this part with the high–level algorithm ***SI–HGT***.

Procedure *SI–HGT*(S_1, S_2, *R*, δ_{SI}, δ_r):

1. for all $S_1, S_2 \in \mathcal{S}$
   - for every HGT suspected gene $g_s \in S_1 \cap S_2$ s.t. $P^{SI}_{value}(g_s, S_1, S_2) < \delta_{SI}$
   - let $n = |g_s|$
   - let $\varepsilon_{g_s}(\delta_r) = \sqrt{\frac{-\log_e \delta_r/2}{2n}}$
     - for $R_1, R_2 \in \mathcal{R}$ s.t. $g_s \in R_1 \cap R_2$
       * for all witness genes $g_w \in S_1 \cap S_2 \cap R_1 \cap R_2$
       * if $|\hat{h}_g(s) - corr^{-1}(d_g)| > \varepsilon_{g_s}(\delta_r)$ mark g_s as putative HGT.

It is important to note here that since we perform many tests for many witness genes and reference organisms, a correction for multiple tests bias should be done. We chose the standard *Bonferroni* correction by multiplying the bound obtained by the number of tests for a given gene.

Simulation study

We conducted a simulation study to assess the advantage of the new *chernoff* approach over the simple χ^2 approach of [1]. For that we created a simulation process as follow. At first, we simulated a random genome, i.e., a list of genes, which were named based on their order, and we created a random nucleotide sequence for each gene. We also set a mutation rate for each gene, drawn from a normal distribution with given mean and standard deviation. These three parts (gene names, gene sequences, gene mutation rates) constitute a genome object, which will be marked *GA*. Next, we created another three genomes, based on *GA*. Two of them will serve as witness genomes (marked as *GWA* and *GWB*) with identical gene order. Each gene sequence of the witness genomes was created as a copy of the corresponding gene in *GA*, then it had undergone a point mutation performed in accordance with its given mutation rate. The third genome, *GB*, was created at first as a copy of *GA*, then each gene also undergone a point mutation performed in accordance with its given mutation rate. Then *GB* had undergone a genome rearrangement process which was executed as follow. In each round, a gene was randomly chosen as well as neighborhood size. This gene, along with his neighborhood, was swapped with other randomly chosen gene with identical neighborhood size. The neighborhood size was randomly chosen from a normal distribution with a given mean and standard deviation. In addition, each gene in the neighborhoods was undergone a point mutation process accordance with its given mutation rate multiply by some *HGT mutation rate factor*, represents the fact that genes undergone HGT event tend to be more evolutionary distant. When dealing with neighborhoods swapping we refer the genome as a circle. At the end of this process we left with 4 genomes and we can execute our HGT detection algorithm to find the genes undergone rearrangement events in *GB* in relative to *GA*, while *GWA* and *GWB* serve as witness genomes. By this process we could test each approach in terms of false positive (FP, genes which the method identified incorrectly as involved in a rearrangement event).

We performed this simulation in which the two approaches of HGT detection (χ^2 and *chernoff*) were competed in terms of False positive events and results are shown in Fig. 4. As can be seen, *chernoff* approach presents much better results in terms of FP, i.e., this configuration yield only few genes which was actually not involved in any rearrangement event, especially for short genomes, while the χ^2 approach presents relatively high FP value. This is an expected outcome, in light of the inherited permissive nature of the χ^2 approach.

Real data study

In order to demonstrate the new HGT–detection method based on *Chernoff* bound we applied it on large real data set from EggNog repository [22], and compared the results to our previous approach based on χ^2. The set contains 1229 pairs of bacteria, in which all pairs are of the same taxonomy genus and species (for example, one of the pairs is *Acinetobacter baumannii AB0057* and *Acinetobacter baumannii AB307–0294*). As can be seen in Fig. 5, we found that for closely related species (SI <0.27), the χ^2 approach detects more genes than *Chernoff* –based approach. We assume that the other genes are identified by the χ^2–based approach are mostly false positive, and this finding of HGT–detection in closely related species is consistent with the simulations presented above (Fig. 4, which presents the low false positive cases of the

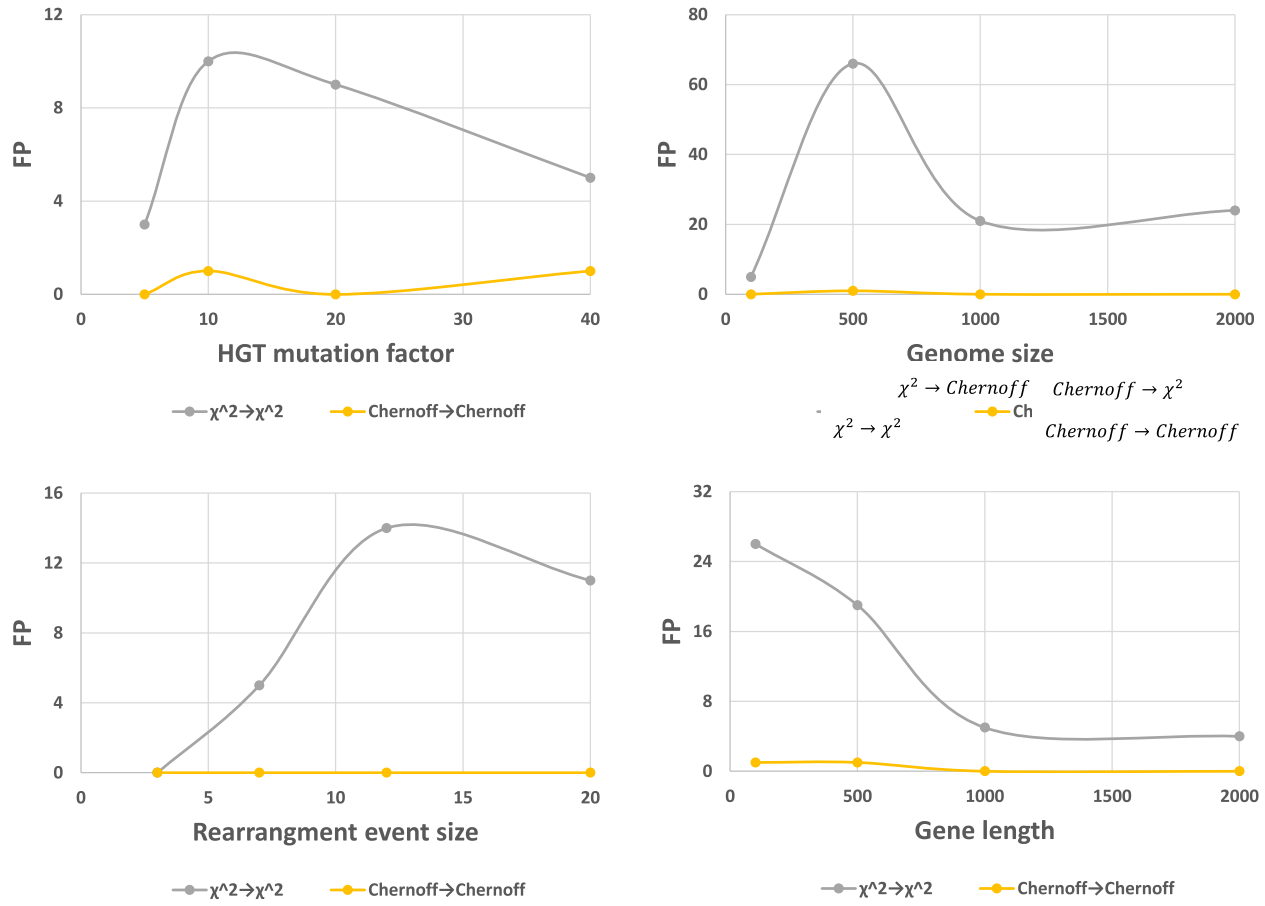

Fig. 4 Performance comparison between χ^2 and *Chernoff* approach for SI-based HGT detection under simulated data. Four artificial genomes were created, one served as original genome, one undergone rearrangement events and two served as witness genomes. The two SI-based approaches for HGT detection were executed and were compared in terms of False positive (FP). Number of HGT events- 10. Each point represents an average of 10 repeats. Default parameters: genome size=1000, gene length=1000, HGT mutation factor=10, rearrangement event size=7

Chernoff–based approach as its main advantage). For less related species, *Chernoff*–based approach detects more HGT–genes than the χ^2 approach, and we assume most of the differences are false positive cases, and this might be a results of high neighborhood size which results high calculated threshold for this low synteny similarity presented by these non–related species (see Eq. (6)). We end this section by recommendation of using this new approach for HGT–detection among closely related species.

Discussion and conclusions

In this work we have provided a probabilistic approach to detect HGTs based on the *synteny index* (SI) and the *constant relative mutability* (CRM) that were defined in [1]. The advantage of the approach portrayed here is the quantification of the statistical signal and using probabilistic bounds to decree significance. The first contribution of this work is assessment of the significance of the SI of a gene. This is essential as distantly related genomes exhibit low SI by default and therefore it is required to distinguish between background noise to signal. The next step is a rigorous probabilistic formulation of the HGT under the CRM property, such that deviations from expected values can be detected and quantified. This requires switching between two spaces– the hamming distance space where bounds on deviations are employed, and the mutation distance space where the CRM property holds. We showed by simulation that the new approach provides greater specificity (i.e., lower false positive rate) over the χ^2 criterion that was provided by [1]. We also demonstrated the specificity improvement in real biological data set. We comment that, as was demonstrated in [1], the advantage of the SI based approach over existing HGT detection techniques, is between closely related taxa where the signal is weak whatsoever. Therefore the improvement in the specificity is imperative. We comment that all steps

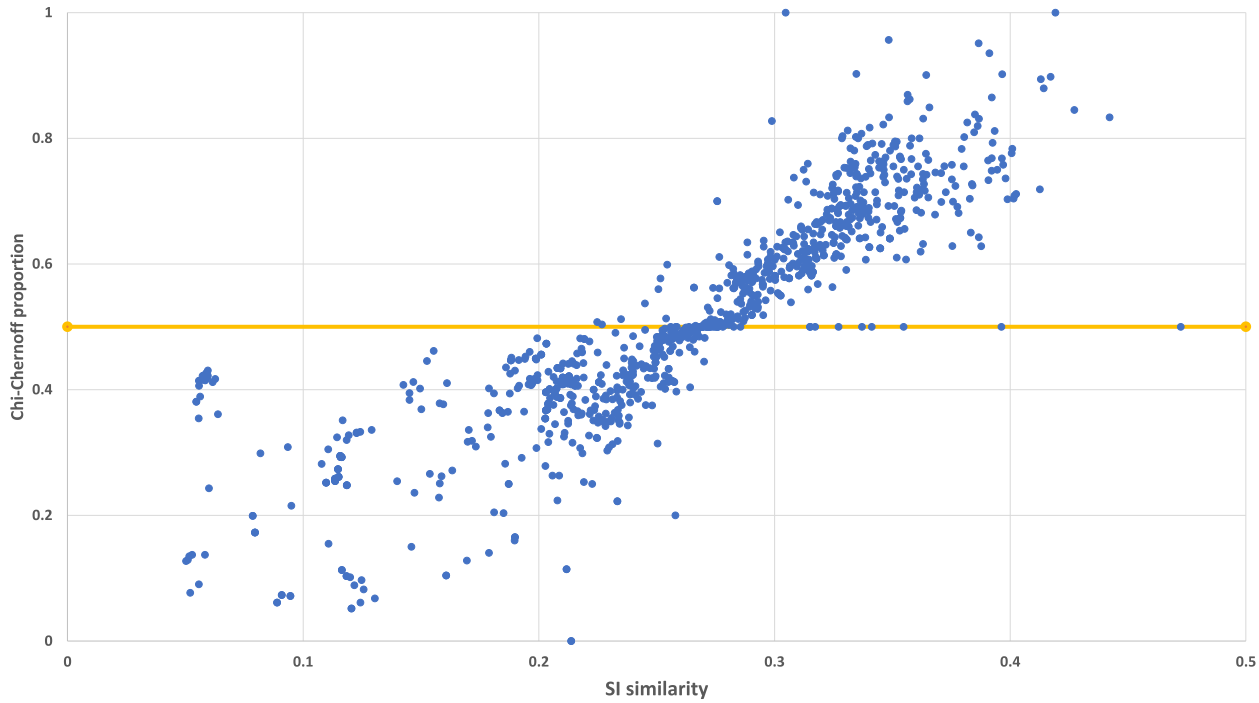

Fig. 5 Performance comparison between χ^2 and *Chernoff* approach for SI-based HGT detection under real biological data. Both the *Chernoff*-based approach and the χ^2 approach for HGT detection were applied on real biological data set contains 1229 pair of closely related species from Eggnog repository. Each point represent a pair of species. The X axis is the SI similarity among the pair. The Y axis is the proportions of the number of HGT-genes detected by the χ^2 approach divided by the amount of genes detected by both of the approaches (i.e., high value in the Y axis represents more genes detected by the χ^2 approach then by the *Chernoff* approach. The yellow line represent the boundary between the two cases)

performed in the algorithm are very fast. One bottleneck in the implementation is the identification of orthologous genes, however the same obstacle stands also in other approaches. Future directions we see in this direction include the establishment of a special repository holding the genes found as HGT putative, similarly to the HGT–DB [10]. Another challenging task is the identification of orthologous genes across many species. This problem stands at the heart of almost any task in comparative genomics. The novelty of our approach is the consideration of gene order among the genomes. While this order can serve as informative for detecting orthology, its lack thereof can allude to exceptional events such as HGT.

Abbreviations

COG: Cluster of orthologous Groups; CRM: Constant relative mutability; EggNog: Evolutionary genealogy of genes non supervised orthologous groups; HGT: Horizontal gene transfer; SI: Synteny index; UPM: Universal pacemaker

Acknowledgements

Not Applicable.

About this supplement

This article has been published as part of BMC Genomics, Volume 21 Supplement 1, 2020: Selected articles from the 14th International Symposium on Bioinformatics Research and Applications (ISBRA-18): genomics. The full contents of the supplement are available at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-21-supplement-1.

Authors' contributions

SS initiated the study. GS, OA and SS contributed to the design and implementation of the research. GS performed simulation analysis. GS and OA performed real data analysis. GS and SS wrote the manuscript. The work was done under supervision of SS. All Authors read, commented and approved the final manuscript.

Funding

We acknowledge the support of the Israeli Science Foundation (ISF) and the VolkswagenStiftung grant, project VWZN3157, for funding GS.

Availability of data and materials

All real data used in this study was taken from EggNOG repository, which is a public online sources, available at: http://eggnog.embl.de/version_3.0/.

Ethics approval and consent to participate

This research is associated with no ethical matters.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 7 October 2019 Accepted: 12 December 2019
Published: 5 March 2020

References

1. Adato O, Ninyo N, Gophna U, Snir S. Detecting horizontal gene transfer between closely related taxa. PLOS Comput Biol. 2015;10(11):e1004408. https://doi.org/10.1371/journal.pcbi.1004408.

2. Alon N, Spencer JH. The Probabilistic Method, 3rd edition. New York: Wiley; 2008.
3. Beiko RG, Harlow TJ, Ragan MA. Highways of gene sharing in prokaryotes. Proc Nat Acad Sci. 2005;102(40):14332–7.
4. Bergeron A, Stoye J. On the similarity of sets of permutations and its applications to genome comparison. J Comput Biol. 2006;13(7):1340–54.
5. Daubin V, Ochman H. Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in E. coli. Genome Res. 2004;14(6):1036–42.
6. Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999;284(5423):2124–9.
7. Edwards RA, Rohwer F. Viral metagenomics. Nat Rev Microbiol. 2005;3: 504–10.
8. Engström PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. 2007;17(12):1898–908.
9. Felsenstein J, Felenstein J. Inferring phylogenies, vol. 2. Sunderland: Sinauer associates; 2004.
10. Garcia-Vallve S, Guzman E, Montero MA, Romeu A. Hgt-db: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. 2003;31(1):187–9.
11. Garcia-Vallve S, Romeu A, Palau J. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 2000;10(11):1719–25.
12. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 2005;3(9):679–87.
13. Hein J. Reconstructing evolution of sequences subject to recombination using parsimony. Math Biosci. 1990;98(2):185–200.
14. Jin G, Nakhleh L, Snir S, Tuller T. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. Mol Biol Evol. 2007;24(1): 324–37.
15. Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian Protein Metabolism. New York: Academic Press; 1969. p. 21–132.
16. Lawrence JG, Ochman H. Reconciling the many faces of lateral gene transfer. Trends Microbiol. 2002;10(1):1–4.
17. Moret BME, Nakhleh L, Warnow T, Linder CR, Tholse A, Padolina A, Sun J, Timme R. Phylogenetic networks: modeling, reconstructibility, and accuracy. IEEE/ACM Trans Comput Biol Bioinformatics. 2004;1(1):13–23.
18. Nakamura Y, Itoh T, Matsuda H, Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet. 2004;36(7):760–6.
19. Nakhleh L, Ruths D, Wang L-S. Riata-hgt: A fast and accurate heuristic for reconstructing horizontal gene transfer. In: Wang L, editor. Computing and Combinatorics volume 3595 of Lecture Notes in Computer Science. Berlin/Heidelberg: Springer; 2005. p. 84–93.
20. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000;405(6784):299–304.
21. Pallen MJ, Wren BW. Bacterial pathogenomics. Nature. 2007;449(7164): 835–42.
22. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, Arnold R, Rattei T, Letunic I, Doerks T, Jensen LJ, Mering CV, Bork P. Eggnog v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res. 2012;40:D284–9.
23. Sankoff D, El-Mabrouk N. Genome rearrangement. Curr Top Comput Biol. 2002;135–155.
24. Shifman A, Ninyo N, Gophna U, Snir S. Phylo si: a new genome-wide approach for prokaryotic phylogeny. Nucleic Acids Res. 2014;42(4): 2391–404.
25. Snir S, Wolf YI, Koonin EV. Universal pacemaker of genome evolution in animals and fungi and variation of evolutionary rates in diverse organisms. Genome Biol Evol. 2014;6(6):1268–78. https://doi.org/10.1093/gbe/evu091.
26. Snir S, Wolf YI, Koonin EV. Universal pacemaker of genome evolution. PLoS Comput Biol. 2012;8(11):e1002785.
27. Tofigh A, Hallett M, Lagergren J. Simultaneous identification of duplications and lateral gene transfers; 2011. p 517–535.
28. Welch RA, Burland V, Plunkett III G., Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli. Proc Natl Acad Sci U S A. 2002;99(26):17020–4.
29. Wolf YI, Snir S, Koonin EV. Stability along with extreme variability in core genome evolution. Genome Biol Evol. 2013;5(7):1393–402.
30. Wolf YI, Rogozin IB, Grishin NV, Koonin EV. Genome trees and the tree of life. Trends Genet. 2002;18(9):472–9.
31. Wong KM, Suchard MA, Huelsenbeck JP. Alignment Uncertainty and Genomic Analysis. Science. 2008;319(5862):473–6.

Publisher's Note