A Novel Technique for Detecting Putative Horizontal Gene Transfer in the Sequence Space

Sagi Snir***[1] Edward Trifonov[2]

[1] Dept. of Evolutionary Biology and the Inst. of Evolution, Haifa Univ. Israel,
ssagi@research.haifa.ac.il
[2] Genome Diversity Center, Inst. of Evolution, Haifa Univ. Israel,
trifonov@research.haifa.ac.il

Abstract. Horizontal transfer (HT) is the event of a DNA sequence being transferred between species not by inheritance. This phenomenon violates the *tree-like* evolution of the species under study turning the trees into networks. At the sequence level, HT offers basic characteristics that enable not only clear identification and distinguishing from other sequence similarity cases but also the possibility of dating the events. We developed a novel, self-contained technique to identify relatively recent horizontal transfer elements (HTE's) in the sequences. Appropriate formalism allows to obtain confidence values for the events detected. The technique does not rely on such problematic prerequisites as reliable phylogeny and/or statistically justified pairwise sequence alignment. In conjunction with the unique properties of HT it gives rise to a two-level sequence similarity algorithm that, to the best of our knowledge, has not been explored. From evolutionary perspective, the novelty of the work is in the combination of small scale and large scale mutational events. The technique is employed on both simulated and real biological data. The simulation results show high capability of discriminating between HT and conserved regions. On the biological data, the method detected documented HTE's along with their exact locations in the recipient genomes.

1 Introduction

Horizontally transferred elements (HTE's) are DNA fragments that are transferred between organisms not on a vertical descent basis. The alien fragment, thus, is inserted in the recipient genome.

The phenomenon is ubiquitous among prokaryotes, implying mainly the transfer of genes (coding sequences) and is termed horizontal gene transfer (HGT) [Doolittle, 1999a, Koonin et al., 2001, Nakamura et al., 2004, Ochman et al., 2000]. The existence of HTE's in non-coding sequences was less investigated (see however [Liu et al., 2004]). Similarity between sequences has been primarily attributed to conservation and, thus, putative functionality [Bejerano et al., 2004, Siepel et al., 2005]. However, similarity can naturally be caused by an event of horizontal transfer (HT), when not a particularly conserved sequence (say, non-coding sequence) is transferred. The very fact of the transfer in such cases may suggest some function of the transferred sequence.

HT has a fundamental evolutionary importance as it turns the traditional tree-like evolutionary history into an evolutionary network [Doolittle, 1999b,a, Wolf et al., 2002]. Genetically, HT is a primary source of new genes that are acquired by bacteria and archaea and often result in adaptations to new environments and conditions Daubin and Ochman [2004]. Therefore, identification of HT can shed light on many significant evolutionary processes that cannot be explained by the traditional tree-like approach. From medical perspective HT is a substantial factor by which bacteria develop resistance to antibiotics [Ambur et al., 2009, Chen and Novick, 2009, Koonin, 2009].

Currently, there are two prevailing methods for detecting HT. The *phylogeny based approach* takes a relatively large set of homologous (originated from a common ancestor) coding sequences, construct their corresponding phylogeny and contrast it to the phylogeny of their originating species. When conflicts are found between the two trees, they are reconciled by introducing HTE's (see e.g. [Delwiche and Palmer, 1996, Jin et al., 2007, Daubin et al., 2003, Beiko et al., 2005, Lerat et al., 2003]). While this approach has the advantage of identifying relatively old events, it is incapable of coping with events residing in non-homologous regions. Moreover, the approach is based on a very stringent assumption of where to seek the events. Finally, it also requires alignment of the sequences and inferring a reliable species tree (two major problems

* Corresponding Author
** supported by the young scientists fellowship of the Yeshaya Horowitz Association through the Center for Complexity Science.

by themselves [Linder and Warnow, 2005]).

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. [Karlin, 2001, Nakamura et al., 2004, Ochman et al., 2000, Ochman, 2003, Garcia-Vallve et al., 2000]). This approach suffers from the fact that the species involved might share similar compositional patterns. Moreover, the length of a HTE may be too short to reliably reveal these differences. As concluded in [Lawrence and Ochman, 2002, Wang, 2001], "atypical G+C content and pattern of codon usage are not reliable indicators of horizontal gene transfer events".
A recent work, DarkHorse [Podell and Gaasterland, 2007] combines between these two approaches by incorporating phylogenetic information into the sequence based search. However, it aims at protein, annotated, data, and is parameter dependent as we discuss below (Methods Section). Concluding this part, there is no gold standard for HT detection and each approach can be used either to detect unknown cases or verify rival approaches [Podell and Gaasterland, 2007].

In this work a novel technique is introduced for identifying putative horizontally transferred sequences between two organisms. It operates in the sequence space and bares no similarity to the two previously outlined approaches. It rather *complements* these approaches and bypasses their drawbacks. The inserted fragment "defines" a *HTE-environment* where similarity between the copied fragment at the donor and the inserted fragment at the recipient, is very high, whereas the corresponding flanking regions at both genomes, are base-wise unrelated (with respect to the HTE-environment, see Figure 1). Naturally, we are not interested in very old HTEs where the contrast at the HTE borders is not clearly visible. A conclusive identification of these cases is problematic. The goal is to distinguish between three types of inter genomic fragments' similarity: (1) random similarity - one that is expected to occur by chance considering the evolutionary distance between the two genomes; (2) similarity by functional conservation - within a whole family; and (3) HTE's (and rare formally similar events).

The technique proposed is not confined only to events between a donor and a direct recipient, as the donors of HTE's have not necessarily survived or their genome is not fully sequenced. Rather, it can also detect HTE's in two recipients in cases of insertion sequences (IS) or plasmids, as well as HTEs between a recipient and a close relative of the donor.

Fig. 1. (a) Two completely identical homologous sequences. (b) After a HT event of GTTG from the upper genome to the lower. Sequences are aligned so that the HTE match, forming the HTE-environment in which the segments corresponding to the HTE at both genomes are similar (or match) and flanking sequences (in the example - GTAC vs TGCT at the left and CTAA vs AATC at the right) are base-wise unrelated.

We assume a rate of point mutations that is, at the order of magnitude, approximately fixed along the time course and even across different species of the same family [Doolittle et al., 1996, Gu, 1998, Kumar and Subramanian, 2002]. Although there are many biological examples of rate acceleration during evolution [Jensen-Seaman et al., 2004], these do not violate the model as it deals with different orders of magnitude. This constant rate induces a measure of dissimilarity between different species that is proportional to the time since divergence. In the mathematical derivations of this work the simple but powerful Jukes-Cantor model of sequence evolution (JC) [Jukes and Cantor, 1969] is used. Note, that this simplification is not necessary for the detection itself, as demonstrated in the Results section, rather for the approximate dating of events.

We now give a brief, general overview of the algorithm: Given two genomes with their lengths, one can calculate the expectancy of each identical region's length. This gives a measure of confidence as to *exceptional* similarity. Next *matching seeds* for putative HT are identified such that their size is chosen to have very low probability of having occurred by chance (based on the evolutionary distance between the two genomes). This stage filters out random similarities (but not conserved regions). We comment that filtering random

similarity has also appeared in other HGT works, e.g. [Darling et al., 2004] (although with no such rigor). In order to uniquely detect HTE's, we rely on the statistical (as well as biological [Omelchenko et al., 2003]) property that a HTE normally is not inserted exactly in its homologous (orthologous) counterpart location. This implies that the flanking regions of a HTE are *non*-homologous (see Figure 1), strictly distinguishing it from conserved regions. Therefore, a special sliding window algorithm is used to detect these HTE borders, searching for sharp borders (or *walls*). Finally, as events with non-homologous flanking regions may also be a result of some rare cases of large scale mutational events (e.g. a very conserved region is duplicated or translocated), we discard any putative HTE appearing twice (or more) in a genome.

The method was applied to a set of simulated data as well as to real bacterial genomes. The simulation studies produced a challenging data with a large proportion of conserved regions and multiple HT events. We are interested in HTEs of gene size that is too short to convey some detectable statistical signal, therefore cannot be identified by traditional sequence-based techniques relying on CpG, codon usage, or alike [Karlin, 2001, Nakamura et al., 2004, Garcia-Vallve et al., 2000]. Alternatively, a simple Blast search between the genomes does detect these short regions but along with hundreds of other conserved regions. In contrast to these two later approaches, our method demonstrates high capability of detecting these events (low false negative rate) but also filters out the regions where similarity is due to simple conservation. The sensitivity of the proposed method was also checked on known documented HT cases and supported by phylogenetic data. Very accurate locations of these events were found along with other potential HTEs. Comparison to other techniques [Podell and Gaasterland, 2007, Darling et al., 2004] was also done for the task of HT detection. The code is available by request from S.S.

2 Methods

2.1 Preliminaries

We consider a (evolutionary) model in which the nucleotides along a genome 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 just a mutation for short) is the event of a nucleotide changing its value to different one. An *evolutionary model* $\mathcal{M}$ models the (stochastic) process of mutations occurring at a site as a function of *mutation rates* $\alpha_{i,j}$ 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 it has nothing to do with the type of mutation baring the same notation (see [Felsenstein, 2003] for more details). Given $\mathcal{M}$, mutation rates $[\alpha_{i,j}]$, 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 Jukes-Cantor evolutionary model is a reversible, one parameter model, postulating that at any position the rate of substitutions from one state to another, $\alpha_{i,j}$ is the same - α. Under this model, the expected number of substitutions at a site during t time units is $3\alpha t$ and is sometimes denoted as *(evolutionary) distance*[1]. The probability $\mathbb{P}_s$ of seeing the *same* nucleotide at a position in a sequence mutating at a rate α after t time units is (note that this does *not* mean the nucleotide has not been substituted):

$$\mathbb{P}_s = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \tag{1}$$

and the probability to see a *different* nucleotide at a position, $\mathbb{P}_d$, is

$$\mathbb{P}_d = 1 - \mathbb{P}_s = \frac{3}{4} - \frac{3}{4}e^{-4\alpha t}. \tag{2}$$

We denote $\mathbb{P}_s$ and $\mathbb{P}_d$ as the "no transition" and "transition" probabilities resp. and we observe that $\mathbb{P}_s \geq \frac{1}{4}$ and $\mathbb{P}_d \leq \frac{3}{4}$.

A *horizontal transfer* (HT) is the event of a subsequence of a genome, the *donor genome*, being copied and inserted at some position at another genome, the *recipient genome*. The right (left) *border* of a HT is a pair of indices indicating the right (left) endpoints of the HT segment at both genomes.

[1] We stick to the time-rate notation as it allows further flexibility in other models.

2.2 First Level: Perfect Matches

Overview We start this part with an overview. In this part we focus on modern HGT events, that is, recent events that have not yet been dissolved (heavily mutated) in the recipient genome. Consider the following (oversimplified) model: At time 0 some ancestral genome bifurcates into two identical genomes. Next, the two genomes start to accumulate mutations at an equal rate, constant over time, resulting in two contemporary genomes. Mutations are distributed randomly and uniformly along each genome. For the sake of clarity we assume the following simplification that will be removed later: A position is mutated at most once over both sequences; that is, the same nucleotide is not mutated in both sequences and no two mutations at the same nucleotide at a given sequence. It can easily be seen that a mutated nucleotide cannot maintain it's original state. Based on these assumptions, we can analyze the distribution of lengths of unmutated segments in both genomes.

Longest Identical Segment We now formalize the idea outlined in the overview. We want to analyze the probabilities of seeing identical segments among the two genomes assuming the scenario above. Note that the time between the two genomes is twice the time since divergence so the "no transition" probability under the JC model (assuming rate α and time since divergence t) is $\mathbb{P}_s = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}$. Assume G_1 and G_2 are the two genomes, both of length ℓ and assume s_x is a subsequence of G_1 of length k. Now, there are two possibilities to see s_x at G_2:

- **Convergence**: s_x appears by chance at position i in G_2. This event has probability $\frac{1}{4^k}$ and by summing over all ℓ positions, $\frac{\ell}{4^k}$ bounds this probability (Note that this bound can be greater than 1 for small values of k).
- **Retention**: There is no transition between any nucleotide of s_x and its homologue at G_1. The probability of this event is the "no transition" probability $\mathbb{P}_s$. We require this for every site in s_x. This event occurs with probability $\mathbb{P}_s^k$.

Now, since $\mathbb{P}_s \geq \frac{1}{4}$, we get that the probability $\mathbb{P}_c$ of seeing s_x by chance at G_2 is at most $(\ell+1)\mathbb{P}_s^k$. The latter bounds the probability of a single subsequence occurring by chance at both genomes. Therefore, the expected number of such s_x from G_1 being found by chance at G_2 is at most $\ell \cdot (\ell+1)\mathbb{P}_s^k$. We now use Markov's inequality stating that $\mathbb{P}[X > \beta] < \mathbb{E}[X]/\beta$ [Alon and Spencer, 1992] to bound the probability of seeing a single occurrence (i.e. $\beta = 1$). Then for X the number of occurrences of such s_x, we get

$$\mathbb{P}[X \geq 1] \leq \ell \cdot (\ell+1)\mathbb{P}_s^k. \tag{3}$$

Suppose we want to bound the probability of finding by chance one or more subsequences of length k at G_2 (i.e. the right hand side of (3)) by δ, then by substituting P_s and simple arithmetic we obtain that

$$k \geq -\log\left(\frac{\ell(\ell+1)}{\delta}\right) / \log \mathbb{P}_s. \tag{4}$$

Equation (4) shows the relationship between the probability of finding a perfect match (bounded by δ) and the length of this match (k). This probability is exponentially decreasing with the length of that match. Therefore by calibrating the probability of finding matches by chance (by setting large enough k) we can eliminate finding matches by chance and this is the subject of the next part.

We end this part with an example: For two mammalian species mutating at an average rate of 2.2×10^{-9} [Kumar and Subramanian, 2002], diverged 6 million years ago (hence $t = 12 \times 10^6$), we get "no transition" probability $\mathbb{P}_s = \frac{1}{4} + \frac{3}{4}\exp(-4 \times 2.2 \times 10^{-9} \times 12 \times 10^6) = 0.925$. Assuming each has genome of $3 * 10^9$ nucleotides, then with probability at most 0.01 we get $\log\left(\frac{\ell(\ell+1)}{\delta}\right) = 48.25$. Then $48.25/-\log 0.925$ yields that we can see by chance identical segments of $k \geq 619$. We however, work with much shorter genomes and much bigger divergence time, yielding a much smaller k.

Finding Significant HTE's The discussion above dealt with the task of how to decrease the probability of finding matches by chance. Our algorithm for detecting statistically significant HTE's starts by finding *seeds* of matches of length beyond reasonable probability (e.g. 0.01) of being found by chance. The algorithm builds the k-spectrum of one genome for large enough k. We now move sequentially (left to right) over the

other genome. Every k-mer in this genome is checked against the k-spectrum to find k-matches (identical segments of length k). Finally k-matches are expanded (obviously only forward, since backward matches were already found), until a mismatch is found and then reported.

2.3 Second Level: Coping with Mutations

The previous Section dealt with probabilities of non-mutated segments as a means for detecting HTE's between two segments. The algorithm in that part finds seeds of identities in both genomes. These seeds are identical regions in the genomes with small likelihood of being created by chance. This section extends the scope to deal with HTE's that underwent mutations. First, we argue that the probability for some seed to be detected is fairly high. Next, the *Butterfly* algorithm, that extends these seeds over regions where similarity is not complete but very high, is introduced. Therefore, one goal of the Butterfly is to bridge between neighboring perfect match segments of a horizontal transfer. More importantly, the main goal of the Butterfly is to distinguish between HTE's and similarity by conservation as will be detailed in the sequel.

Sufficient Conditions for Seeds Detection After a HT event has occurred the HTE is exposed to mutations. HTE's that have occurred sufficiently long time before present are *dissolved* in both genomes by accumulating mutations in the copied segment, so their similarity diminishes and therefore cannot be identified. In the previous section, a long enough seed was required to discard random matching. However, due to mutations over a too long time since the HTE event, there might not be long enough seeds (as explained above). Therefore, we say that a HTE *survives* if some k-match is found between the two segments corresponding to it in the two genomes. Our algorithm will fail already in the seeds stage if no seed will be found. This is equivalent to the case when every consecutive k nucleotides in the HTE have at least one mutation. The likelihood of this case is fairly low for relatively recent events.

Given a known mutation rate α, we want to determine what is the age $\hat{t}$ of HT events that can be recognized with high probability.

Let $\mathbb{P}_s$ be the "no transition" probability at a site, i.e. $\mathbb{P}_s = \frac{1}{4} + \frac{3}{4}e^{-8\alpha\hat{t}}$. The HTE H is partitioned into consecutive disjoint fragments $[H_i]$, each of length k. Note that if there is a single fragment with no transition between the genomes (i.e. a k-match), a seed will be discovered (see Figure 2). Also note that there is still the possibility that every fragment has a transition but there is a k-match between transitions of neighboring fragments (see again supplementary Figure 2 and accompanying explanation). Therefore the condition of at least one "untouched" segment, is sufficient but not necessary.
The probability of having a k-match at some H_j is $\mathbb{P}_s^k$ and $1 - \mathbb{P}_s^k$ otherwise. Denote by δ the probability that no fragment has a k-match, that is

$$\delta = \left(1 - \mathbb{P}_s^k\right)^{\ell_H/k}. \tag{5}$$

From (5) we get $\mathbb{P}_s = \left(1 - \delta^{k/\ell_H}\right)^{\frac{1}{k}}$ and by the definition of $\mathbb{P}_s$, $\hat{t}$ is obtained. Recall, by the discussion above, that $\delta \geq \mathbb{P}[H \text{ does not survive}]$. Therefore, if H is a HTE of length ℓ_H transferred $\hat{t}$ time units before present, then H will survive with probability at least $1 - \delta$ if $\hat{t}$ satisfies:

$$\hat{t} \leq -\frac{1}{8\alpha} \log\left[\frac{4}{3}\left[\left(1 - \delta^{k/\ell_H}\right)^{\frac{1}{k}} - \frac{1}{4}\right]\right]. \tag{6}$$

The Butterfly Algorithm We now describe the *Butterfly* algorithm which is the heart of the method and serves to locate the HTE borders as well as to discriminate HTEs from conserved regions. We start with an overview of the algorithm. When a segment is inserted to a genome, it creates a *HTE-environment*. In this environment, the similarity between the copied fragment at the donor and the inserted fragment at the recipient, is very high. However, beyond the borders of the HTE the corresponding flanking regions at both genomes, are base-wise unrelated. Now, suppose we color regions of high similarity by blue and regions of low similarity by red. Hence, we define the notion of a *Butterfly* to capture the intuition that when the Butterfly sits on a HTE border, one wing is red and another is blue. More formally, a (single) *wing* of the Butterfly is a pair of sequences of a specified length at two specified positions in the two genomes. The Butterfly has two adjacent, equally long, sliding windows, the end of the left wing is the starting position of the right

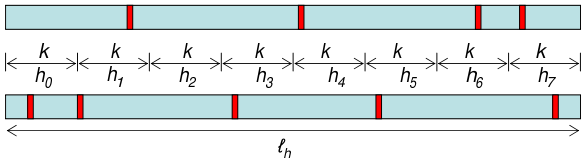

Fig. 2. The HTE H is segmented to $\frac{\ell_h}{k}$ adjacent segments of length k each ($h_0 \dots h_7$ in the figure). A transition is marked by a red (dark) bar in one of the genomes. If a certain segment (e.g. h_2) was not exposed to a transition, H will k-survive. Note also that although there was a transition in adjacent segments h_5 and h_6, there is an *untouched* fragment between two red bars of length $> k$ residing in both h_5 and h_6.

one. These are the Butterfly's wings. The Butterfly travels along the two genomes by moving sequentially the two adjacent sliding windows, seeking for high contrast between the wing colors. The *right mismatch score*, is the relative number of mismatches between the two segments of length ℓ composing the right wing. Similarly, the *left mismatch score*, is when the comparison is done between the ℓ positions composing the left wing.

Alternatively, the mismatch score is the normalized Hamming distance between the two segments in the wing/sliding window.

The Butterfly works similarly to the expansion stage of BLAST [Altschul et al., 1990] by trying to expand a seed to both sides. Its stopping conditions are however different. It starts at a (perfect match) seed found by the exact matching algorithm. Then it moves to the right one position at a time and compares the mismatch score between the two windows. The *wall* of a Butterfly is the difference between the mismatch scores of the two wings. The Butterfly stops when it finds a wall greater than some threshold parameter τ. Figure 3 illustrates all the above. After encountering a wall at the right, the process repeats from the seed to the left (now the mismatch at the right wing is subtracted from the mismatch at the left).
As two string matching algorithms, the BLAST and the Butterfly are similar. However, BLAST is not geared to detect sharp walls appearing at the HTE borders. Indeed, BLAST will detect many conserved regions and will *not* distinguish between conserved regions and HTEs (cases (2) and (3) in the introduction).

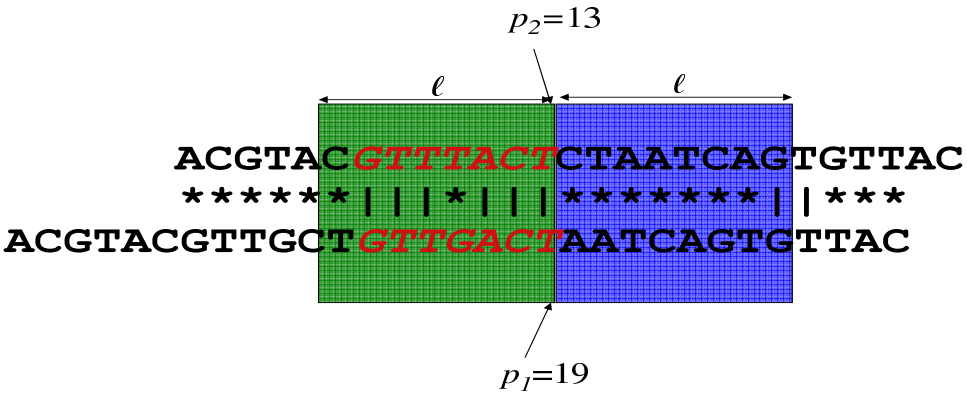

Fig. 3. A butterfly starting at positions 19, 13 in the lower and upper genomes resp. The wings are of length 8. This is a right border of seven nucleotides HTE (accumulated one substitution), hence the right mismatch score is $\frac{7}{8}$ and the left - $\frac{2}{8}$ defining a high enough wall of $\frac{7}{8} - \frac{2}{8} = \frac{5}{8}$.

Following is the formal algorithm:

Procedure *Butterfly*(p_1, p_2, ℓ, τ):

1. for $(i \leftarrow 0; ; i++)$
   (a) $m_r \leftarrow mismatch^r(p_1 + i, p_2 + i, \ell)$.
   (b) $m_l \leftarrow mismatch^\ell(p_1 + i, p_2 + i, \ell)$.
   (c) if $(m_l > \tau)$ report the region as a putative conservative region and exit.
   (d) $wall(p_1 + i, p_2 + i, \ell) \leftarrow m_r - m_l$.
   (e) if $\Big(wall(p_1 + i, p_2 + i, \ell) > \tau\Big)$ record i as a right border and goto 2.
2. for $(j \leftarrow 0; ; j++)$
   (a) $m_r \leftarrow mismatch^r(p_1 - j, p_2 - j, \ell)$.
   (b) $m_l \leftarrow mismatch^\ell(p_1 - j, p_2 - j, \ell)$.
   (c) if $(m_r > \tau)$ report the region as a putative conservative region and exit.
   (d) $wall(p_1 - j, p_2 - j, \ell) \leftarrow m_l - m_r$.
   (e) if $\Big(wall(p_1 - j, p_2 - j, \ell) > \tau\Big)$ record j as a left border and leave.
3. return $j + i + 2\ell$

The goal of the Butterfly is to locate the borders of a HTE. Assuming we are proceeding rightward, this occurs when the HTE border is exactly between the two wings, such that the right wing is completely outside the HTE and the left wing is completely inside it. Since the two processes under consideration - random similarity at the right (outer) wing and mutation rate at the left (inner) wing - are stochastic, we augment the Butterfly with two parameters to handle the noise generated by these processes. In order to have enough confidence in the decision to stop, the butterfly needs to "see" enough information. A very strong signal but with low confidence can result from too short wings (for example a 100% wall resulting from two mismatches at the right wing and two matches at the left can be pure noise). This is controlled by the wing lengths parameter ℓ. On the other hand, the wall threshold parameter τ, controls the stopping condition: A too low threshold can stop the Butterfly prematurely at a false wall inside the HTE. Figure 4 illustrates both scenarios.

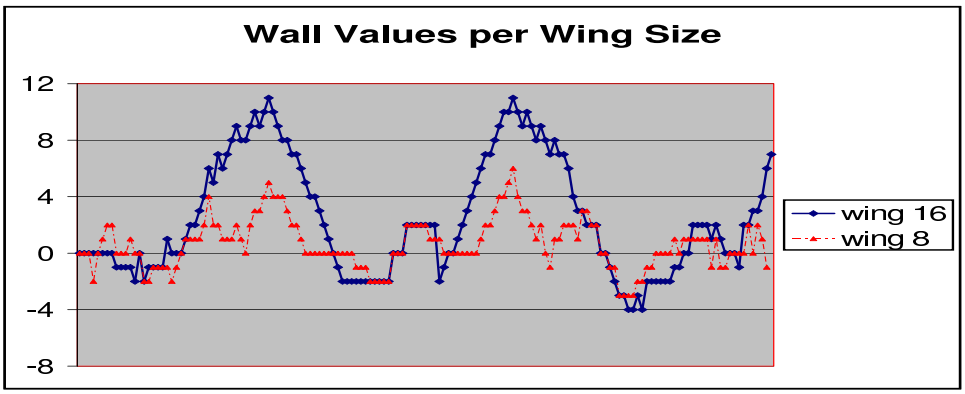

Fig. 4. Wall values of a simulated HTE-environment of length 50 (middle of figure between the two peaks). As can be seen, wing of size 8 produces a smaller signal to noise ratio compared to wing's size 16 hampering the identification of the HTE's border and consequently HTE's identification. Also note that a too small threshold value τ can lead to a false early wall identification inside the HTE as a result of noise (mutations) inside the HTE. (Remark: the wall on the right side is calculated oppositely to the left wall by moving the sliding windows leftward.)

Since the segments beyond the HTE border are base-wise uncorrelated (i.e. non-homologous), the expected relative mismatch score outside of the HTE is $\frac{3}{4}$. Therefore we chose a wall of $\frac{1}{2}$, leaving a high enough mismatch score inside the HTE, to prevent premature stopping.

2.4 Third Level: Discriminating Conserved Regions

We define a *conserved region* as two homologous segments in G_1 and G_2 with similarity above $\frac{1}{4}$ (random). Similarly to HTE's, conserved regions can have sub regions of complete similarity, producing seeds by the

exact match algorithm what can lead to false positive identification of HTE's. However, as opposed to HTE's, conserved regions have the distinguishing property that the sequences beyond a border of a conserved region are still homologous, and therefore have no clear borders. Figure 9 depicts two outputs of the Butterfly algorithm on real genomic data: In one, the borders are very clear yielding clear walls whereas in the second the walls are not clear. Figure 5 illustrates the difference of the borders at conserved regions versus real HTE's.

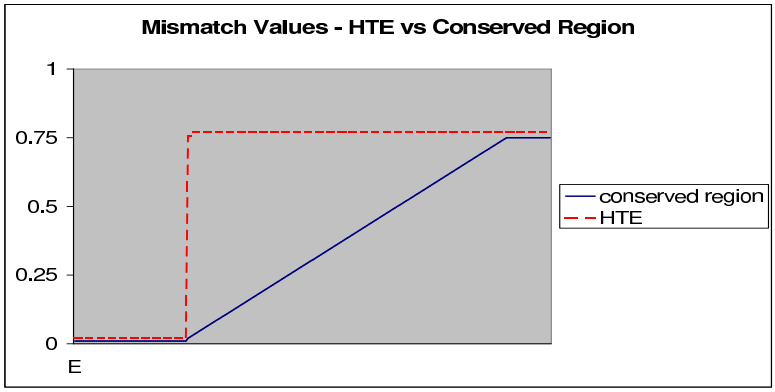

Fig. 5. While the mismatch value at HTE border jumps abruptly from 0 (or nearby values) to $\frac{3}{4}$ (red dashed line), at a conserved region the slope is much more gradual (blue solid line).

The above implies that if the Butterfly starts at a seed corresponding to a conserved region, then it might not hit a wall and might run through excessive regions with expected mismatch of $\frac{3}{4}$ at both wings. To remedy this phenomenon, the Butterfly halts the algorithm when the back wing hits a high mismatch score, and discards the putative HTE.

Accounting for Large Scale Mutational Events Large scale mutational events (also denoted as *genome rearrangement events*) are events in which large genomic fragments are either copied or moved to other location in the genome. These events are denoted as *duplications* and *translocations* respectively. Note that in such events, the flanking regions of the copied fragment and the flanking regions of some homologous fragment at another species, are not homologous between themselves, creating a spurious wall. Our method accounts for this as follows: if the translocation event is old enough, there were enough mutations in the homologous segments at both genomes, preventing a seed (first level) to be found in them. If the event is recent, then the translocated fragment is divergent enough from its homolog at the other genome, preventing a seed again. The only case where a seed is detected is when a *conserved* (in contrast to a ordinary non-conserved) region has been copied recently. To account for this case, we discard all cases where a putative HTE has a similar copy in the genome. The only case that cannot be detected is when a very conserved fragment is being translocated; this however is very rare and cannot be detected by any other method.

3 Results

The method is implemented in software, one program for the seeds (exact match) stage and another for the Butterfly, both in C. A simulation study was conducted in which artificial genomes and HT events were generated. The simulation's results were plugged into our equations in order to set the algorithms parameters. Subsequently, known biological HTEs were analyzed and located in their host genomes. Finally, the method was applied to raw genomes and our findings were compared with known annotated data.

3.1 Simulation study

The simulation study was performed to evaluate the performance of the new method. The goal was to evaluate the capability of the method to detect HTE's and to discriminate between HTE's and conserved regions. We first clarify the setting. Equation (6) gives a bound on event ages for which we are likely to

find seeds. We can not guarantee detection of the events above this age. Also, recall that distance between genomes is irrelevant as regions flanking a HTE are base-wise uncorrelated (expected similarity $\frac{1}{4}$). Indeed, simulations with one genome segment being transplanted in another genome yielded expected obvious results of 100% success. This has motivated us to formulate a more challenging input with multiple genomes and conserved regions (forming false seeds) and multiple HTE copies, intermixing with these conserved regions.

Simulated Datasets The r8s [Sanderson, 2003] software was used to produce random birth-death trees (i.e. random topology and branch lengths) over 20 taxa. In order to have enough divergence time between any two species, we normalized branch length to a minimum of $d = 0.2$ yielding a distance of $d = 0.4$ between any two species, where $d = 3\alpha t$ (see distance definition in the Methods Section).
In order to obtain bounds on random similarity we need to convert this distance to the parameters of Equation (1). Normally, distance d is defined as the expected number of substitutions and under the Jukes-Cantor model this is $3\alpha t$. Plugging this into Equation (1) yields $\mathbb{P}_s = 0.69$. So in order to set a k large enough to avoid random matching we use Equation (4). The genomes were about 3Mbp long and 90% confidence ($\delta = 0.1$) was chosen, yielding a seed length $\Big($Equation (4)$\Big)$ $k \geq 87$. Note that this seed would filter out only random similarity in non-conserved regions. However, this will not filter out the similarity among conserved regions.
Next we used Seq-Gen [Rambaut and Grassly, 1997] to evolve sequences according to the JC model, based on the trees generated in the previous step. The sequences were generated on the tree under two modes: normal rate of evolution corresponding to non-conserved regions and slow rate of evolution corresponding to conserved regions. The two sub-segments were concatenated into a segment and the whole process recurred several times to form a conserved/non-conserved alternating genome. Such a type of input generates a large number of putative seeds in the first stage and the challenge is to discriminate between them and real HT events. Note that a simple similarity based HT detection algorithm (e.g. Blast) will falsely report on a suspected HTE for every conserved segment between closely related sequences (in a relatively small subtree).

Now, a set of HT events was generated, several copies each. Every copy was inserted at a random location along a randomly chosen sequence (genome) generated at the previous stage. To simulate age of events (so that the inserted sequences acquire mutations in their host genome), we again used Seq-Gen with a star topology, uniform edge lengths tree as a model tree. Edge length (tree depth) varied randomly but in a much shorter range in comparison to the model species tree. Using Equation (6) (limiting age of HT events as a function of HTE length ℓ_h, detection confidence δ and rate of mutation α), if we set the HTE length $\ell_h = 2000$, then WHP ($\delta \leq 0.1$) we can trace HTE's of distance $\hat{d} = 3\alpha\hat{t} \leq 0.0135$ from one to another. Hence we set the edges of the star tree for the HTE generation (see Methods Section) to 0.006 yielding $\hat{d} = 0.012$. This way we guarantee low false negative rate. We comment that by the above, we limit our attention only to events recent enough, for which the algorithm has any chance to detect them. Illustration of a simulated input for a single event and three event copies is shown in supplementary Figure 6.

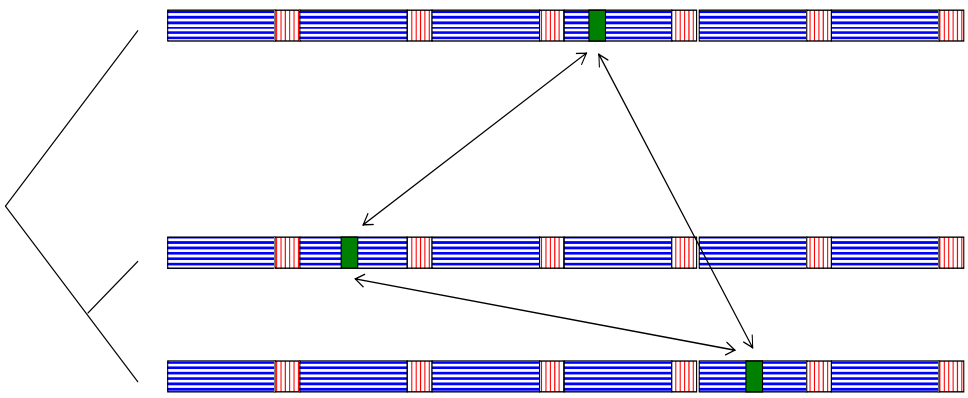

Fig. 6. A simulated input of three genomes with alternating conserved/non conserved regions, evolved along the tree in the left with three copies of a single HT event inserted at random positions.

Simulation Results As was said above, the goal was to measure the effect of false positive/negative (FP/FN) as a result of varying number of event copies. The number of HTE pairs was counted (this is simply $\sum \binom{c_i}{2}$ where c_i is the number of copies generated for HT event i). In the simulation results, we counted the events reported by our procedure that deviate in length by at most 20 (bases) from the real HTE length (this requirement was quite strict as large proportion of events were discovered but with length not in that range). Any event not in that range was considered as a false positive. Figure 7 reports results for two cases: a single HT event and four HT events (both with varying number of copies for each event). The results show that under a reasonable amount of copies for each event the FN rate is fairly low meaning that most HTE's were found and with the correct length. The rate of FP is low for any level of event copies, meaning that the method discriminates very well between HTE's and conserved regions, moreover that most of the FP events were real HTE's but with incorrect length. We comment that in a separate setting in which the HT star tree was very shallow (meaning very recent events) and longer seeds, we obtained a negligible FP rate (under 1, absolute value) and FN (under 2, percentage of events generated) for all #HT copies.

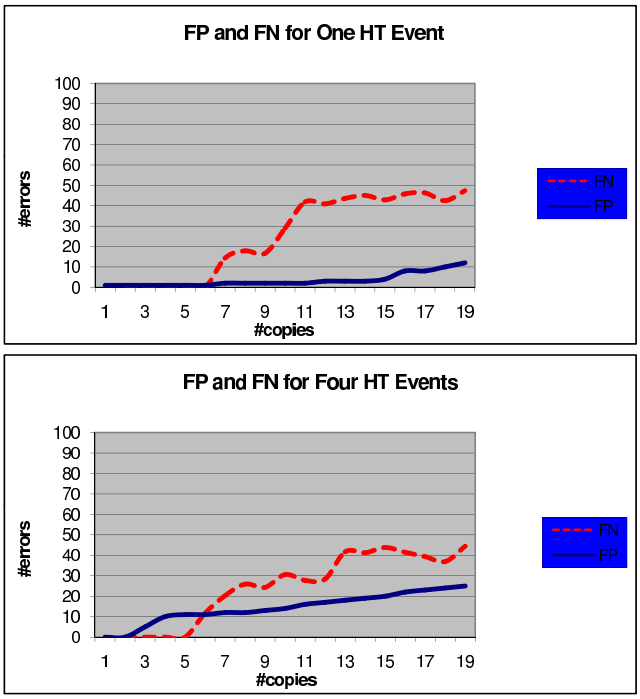

Fig. 7. False positive (lower solid line, absolute values) and false negative (upper dashed line, percentage of number of events generated) as a result of the number of HTE copies per event. Top: A single HT event. Bottom: Four HT events.

Element	stutzeri NC_009434			
	GenBank		Butterfly	
	CDS	length	CDS	length
alaT	2764257..2765468	1211	2764254..2765453	1199
pgsA	2269251..2269865	614	2269294..2270010	710
ppsA	2215516..2217801	2285	2215416..2217782	2366
purL	289386..1293282	3896	1289368..1293283	3915

Fig. 8. Comparison of location and length of HT-putative genes: NCBI GenBank vs Butterfly.

3.2 Real Data Results

The results with simulated data described above show that when the time of event is relatively recent, the method detects it with high probability. Moreover, the very low FP rate shows the power of the method in filtering out false seeds resulting from conserved regions. In its current form, our implementation does not handle effects caused by indel events as they create artificial walls[2].

Nevertheless bacterial genomes are comprised mainly from coding material which is less affected by indels. On the other hand, they exhibit high rate of HT activity and therefore are a perfect fit for demonstrating the concept behind the new method.

We started with demonstration of performance of the new method as a tool for detecting and locating the exact borders of the events. We chose an annotated HTE in the form of an insertion sequence IS53 (acc. M83932). Using Blast we found it in two species of *Pseudomonas*: *Pseudomonas stutzeri* (acc. NC_009434) and *Pseudomonas syringae* (acc. NC_004578).

In *P. stutzeri* the IS was matched in two locations, one at 673486..675897 (see supplementary material for blast output on *P. stutzeri*). Hence, we applied our software to these two genomes, seeking to find HTEs with clear borders. The technique identified this insertion and located the *exact* right (downstream) border of this insertion event (i.e. position 675897. See Butterfly output on these two genomes in supplementary material). As transposable elements are known to be major vehicles for HT, the above finding testifies on suitability of the technique for detection of HT.

Moreover, this setting of differentially located segments in two genomes fits as input for other techniques for HGT detection relying on the same characteristics. Techniques relying on protein, or protein coding, sequences [Garcia-Vallve et al., 2003, 2000, Podell and Gaasterland, 2007, Karlin, 2001, Nakamura et al., 2004, Ochman et al., 2000, Ochman, 2003] are immediately ruled out, leaving only techniques operating on raw DNA sequences. The best known is Mauve [Darling et al., 2004] which as ours, seeks for seeds in the first stage but does not employ criteria similar to the Butterfly on the second stage. Indeed Mauve could not distinguish well between conserved elements and HTEs and produced mixed output (see supplementary material). This should be compared to the well filtered output of the proposed method. Even a detailed search for the specific event of IS53 with appropriate resolution failed - it was not detected (see specific output in supplementary material).

Several highly homologous gene pairs were accurately detected as well with different (non-homologous) flanking sequences (genes *alaT*, *pgsA*, *ppsA* and *purL* of lengths 700 to 3900 bp). The accuracy of our method with respect of locating the sharp borders, as compared by the location *and* length to NCBI RefSeq, is shown in Figure 8. These gene pairs appear as differentially translocated in *P. stutzeri* and *P. syringae* genomes. The horizontal transfer from one of the genome to another, or from a third party is not excluded as well. Since this work is geared to developing a method for detection of putative HTE, the detailed analysis of the detected elements, with involvements of other genomes will be a subject of a separate study.

The next task was to locate and verify a well documented cases of HTE in bacteria. We started with a HT event in *Borrelia* [Barbour et al., 2005]. The genes *hpt*, *purA* and *purB* appear in the genomes of two species *B. hermsii* and *B. miyamotoi*, agents of human relapsing fever, but not in other species of the *Borrelia* genus (e.g. *B. burgdorferi* and *B. garinii*). In [Barbour et al., 2005] the authors concluded that these genes were acquired by a horizontal transfer from an unknown donor at some time after diversification of *Borrelia* from their common ancestor. The possibility of gene loss at the other species was discarded by maximum parsimony arguments. They also constructed the gene trees for the *purA* and *purB* genes over sets of 34 and 36 representative species respectively. By these trees, one likely candidate donor would be *Fusobacterium nucleatum*. Indeed, according to our calculation among 32 protein-coding genes, within 40 kb around the *purA* of *B. hermsii* (accession number NC_010673), the genes *purA* and *purB* show the largest sequence identity seeds (34 and 22 bases, respectively) when compared to the genome of *Fusobacterium* (accession number NC_003454). That is, application of our method of searching for the HTE to the pair of the genomes (*B. hermsii* and *Fusobacterium*) suggests that the above genes appear, indeed, to be horizontally transferred to *B. hermsii*. As to the possible donor - the likely candidate is *Fusobacterium*: the sequence similarity of the three genes between the two species *B. hermsii* and *B. miyamotoi* was found to be 88% whereas similarity between each of these and the respective genes in *Fusobacterium* is only 69%. Under the assumption that the donor is a close relative of *Fusobacterium*, this suggests that the HGT event has occurred before the speciation event of the *B. hermsii* and *B. miyamotoi*.

One more well documented case of HT event is the transfer of a 2.5 kb segment that includes gene *bioF* from *Neisseria meningitidis* and *Haemophilus influenzae RD* [Kroll et al., 1998]. At the exact match stage, we

[2] This can be overcome by joining adjacent segments found by the Butterfly.

found four seeds of length 193, 132, 100, and 100. The extension of these by the Butterfly algorithm resulted in a single HTE of length 2547 at locations 1697161-1699708 and 1621851-1624398 in the *Neisseria* and the *Haemophilus* genomes respectively (see supplementary material). We note that this is in full accordance with the original observation of [Kroll et al., 1998].

4 Discussion

4.1 Conclusions

In this study we attempted at a desirably rigorous description and handling of horizontal transfer events. Specifically, we introduced the *matching seeds* idea to sort out random matches. The next level copes with mutations inside the HTE, implemented by the sliding window of the *Butterfly* algorithm. At the top level, we distinguish HTEs from natural conserved regions by requiring "high walls" between the Butterfly wings.

This paper lays out the rigorous statistical/algorithmic groundwork for this new approach. While this seemingly restricted problem of finding HTEs between two genomes might appear simple, in fact it is key to a broader research where a group of organisms is studied. The proposed procedure and software can serve as a subroutine applied to all pairs in the group and therefore run in time quadratic in the group size. By our experiments, for groups of size of a hundred organisms, this should take not more than a day.

The examples on real data serve as a proof of concept for this new approach rather than a thorough search for HTE's among all published pairs of genomes. Moreover, as described in the Introduction and demonstrated in the Results section, an extensive benchmark between our method and *all* existing HT detection methods is problematic as each operates on different inputs, uses different assumptions and aims at other goals. Instead, in the *Borrelia* case, we showed how a combination with a complementary method, the phylogenetic approach, endows further information on the event.

Our work operates in the sequence space. Mutational events in the sequence space are divided to small scale mutational events which are point mutations, and large scale mutational events such as duplications, translocations and HT. There is a host of papers operating in the sequence space, and even beyond - in the phylogenetic space. While these works handle separately each type of events, ours combines these two types of events - small scale and large scale. We believe this novel combination is interesting and potentially important.

4.2 Future Research Directions

The research presented here gives rise to several future research directions that we intend to investigate. we list them below. The first question is whether HT is a widespread phenomenon in non-coding sequences, those with no obvious functionality, and if so, among particular regions? Previous studies on HT mainly focused at coding regions (ORF"s) and therefore this question has not been investigated sufficiently. A positive answer to this question can add another layer to the growing field of searching functionality in non-coding regions. Next, is the behavior of the mutational process behaves in HTE's? Can we infer by it as to the age of the event? In particular, *Amelioration* [Lawrence and Ochman, 1997, Ragan, 2001] is a process by which a gene that was transferred horizontally acquires features (*e.g.* GC content, the percentage of nitrogenous bases on a DNA molecule which are either guanine or cytosine) similar to its new environment. This is particularly true for recent events as this process diminishes in time and therefore relevant to our work. If the mutational process accelerates at HTEs, then calculation of events age should take this in considerations.
Finally, are there common patterns of HT among different families. Specifically, horizontal transfer was found also in plants [Bergthorsson et al., 2003], Archaea [Matte-Tailliez et al., 2002]. As the method introduced here can readily be applied to higher organisms, like we showed in our hypothetical example on two mammalians, we intend to use it to seek for HTEs among these families.

Bibliography

N. Alon and J.H. Spencer. *The Probabilistic Method*. Wiley, New York, 1992.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

OH Ambur, T. Davidsen, SA. Frye, S. Balasingham, R. Karin, L.and Torbjorn, and T. Tnjum. Genome dynamics in major bacterial pathogens. *FEMS Microbiol Rev.*, 33(3):453–70, 2009.

A. G. Barbour, A. D. Putteet-Driver, and J. Bunikis. Horizontally acquired genes for purine salvage in borrelia spp. causing relapsing fever. *Infection and Immunity*, 73(9):61656168, 2005.

R. G. Beiko, T. J. Harlow, and M. A. Ragan. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences*, 102(40):14332–14337, 2005. doi: 10.1073/pnas.0504068102. URL http://www.pnas.org/cgi/content/abstract/102/40/14332.

G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. Ultraconserved elements in the human genome. *Science*, 304(5675):1321–5, 2004.

U. Bergthorsson, K. L. Adams, B. Thomason, and J. D. Palmer. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature*, 424(6945):197–201, 2003.

John Chen and Richard P. Novick. Phage-Mediated Intergeneric Transfer of Toxin Genes. *Science*, 323 (5910):139–141, 2009.

Aaron C.E. Darling, Bob Mau, Frederick R. Blattner, and Nicole T. Perna. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research*, 14(7):1394–1403, 2004.

V. Daubin and H. Ochman. Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in E. coli. *Genome Research*, 14(6):1036–1042, 2004. doi: 10.1101/gr.2231904.

V. Daubin, N. A. Moran, and H. Ochman. Phylogenetics and the cohesion of bacterial genomes. *Science*, 301(5634):829–32, 2003.

C. F. Delwiche and J. D. Palmer. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. *Mol Biol Evol*, 13(6):873–82, 1996.

R. F. Doolittle, D. F. Feng, S. Tsang, G. Cho, and E. Little. Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science*, 271(5248):470–7, 1996.

W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–9, 1999a.

W. F. Doolittle. Lateral genomics. *Trends Cell Biol*, 9(12):M5–8, 1999b.

J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2003.

S. Garcia-Vallve, A. Romeu, and J. Palau. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res*, 10(11):1719–25, 2000.

S. Garcia-Vallve, E. Guzman, M. A. Montero, and A. Romeu. Hgt-db: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res*, 31(1):187–9, 2003.

X. Gu. Early metazoan divergence was about 830 million years ago. *J Mol Evol*, 47(3):369–71, 1998.

M. I. Jensen-Seaman, T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C.-Fu Chen, M. A. Thomas, D. Haussler, and H. J. Jacob. Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. *Genome Res.*, 14(4):528–538, 2004. doi: 10.1101/gr.1970304. URL http://www.genome.org/cgi/content/abstract/14/4/528.

G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. *Mol Biol Evol*, 24(1):324–37, 2007.

T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

S. Karlin. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends in Microbiology*, 9(7):335 – 343, 2001.

E. V. Koonin, K. S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–42, 2001.

Eugene V. Koonin. Darwinian evolution in the light of genomics. *Nucl. Acids Res.*, 37:1011–1034, 2009.

J. Simon Kroll, Kathryn E. Wilks, Jayne L. Farrant, and Paul R. Langford. Natural genetic exchange between Haemophilus and Neisseria: Intergeneric transfer of chromosomal genes between major human pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 95(21): 12381–12385, 1998.

S. Kumar and S. Subramanian. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A*, 99(2): 803–8, 2002.

J. G. Lawrence and H. Ochman. Reconciling the many faces of lateral gene transfer. *Trends Microbiol*, 10 (1):1–4, 2002.

J.G. Lawrence and H. Ochman. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol.*, 44(4):383–397, 1997.

E. Lerat, V. Daubin, and N. A. Moran. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. *PLoS Biol*, 1(1):E19, 2003.

C.R. Linder and T. Warnow. Overview of phylogeny reconstruction. In Srinivas Aluru, editor, *Handbook of Computational Molecular Biology*. CRC Press, 2005.

Y. Liu, P. Harrison, V. Kunin, and M. Gerstein. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.*, 5(9):R64, 2004.

O. Matte-Tailliez, C. Brochier, P. Forterre, and H. Philippe. Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol*, 19(5):631–9, 2002.

Y. Nakamura, T. Itoh, H. Matsuda, and T. Gojobori. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet*, 36(7):760–6, 2004.

H. Ochman. Neutral mutations and neutral substitutions in bacterial genomes. *Mol Biol Evol*, 20(12):2091–6, 2003.

H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

M. Omelchenko, K. Makarova, Y. Wolf, I. Rogozin, and E. Koonin. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biology*, 4(9):R55, 2003. ISSN 1465-6906.

S. Podell and T. Gaasterland. Darkhorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biology*, 8(2):R16, 2007.

M. A. Ragan. On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.*, 201(2): 187–191, 2001.

A. Rambaut and N. C. Grassly. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13(3):235–238, 1997. doi: 10.1093/bioinformatics/13.3.235.

M. J. Sanderson. r8s; inferring absolute rates of evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19:301–302, 2003. Available at http://loco.biosci.arizona.edu/r8s/index.html.

A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8): 1034–50, 2005.

B. Wang. Limitations of compositional approach to identifying horizontally transferred genes. *J Mol Evol.*, 53(3):244–250, 2001.

Y.I. Wolf, I.B. Rogozin, N.V. Grishin, and Koonin E.V. Genome trees and the tree of life. *Trends in Genetics*, 18(9):472 – 479, 2002.

Fig. 9. Output of the Butterfly algorithm on real data (stars represent matches). Up: clear borders with $mismatch \sim \frac{1}{4}$ outside the HTE. Down: borders are not very clear with $mismatch \sim \frac{1}{3}$ outside.