On the theoretical properties of bipartition dissimilarity measure

Shuguang Li[1,2*], Shuying Chen[1, 2], Mengtian Cui[3]

[1] *Key Laboratory of Intelligent Information Processing in Universities of Shandong (Shandong Institute of Business and Technology), Yantai, 264005, China*

[2] *College of Computer Science and Technology, Shandong Institute of Business and Technology, Yantai, 264005, China*

[3] *School of Computer Science and Technology, Southwest University for Nationalities, Chengdu, 610041, China*

Abstract

Bipartition dissimilarity is a new measure introduced by Alix Boc et al. They proposed an algorithm for inferring horizontal gene transfer events which can rely on different optimization criteria. Simulation results suggested that the strategy based on bipartition dissimilarity provided better results than those based on other three existing tree comparison indices. However, no theoretical analysis on it has been conducted since then in the literature. The present paper reports some useful new results for this measure. The theoretical properties studied include minimum positive value, neighborhood, and local modifications.

Keywords: Tree comparison, Bipartition dissimilarity, Minimum positive value, Neighborhood, Local modifications

1 Introduction

Horizontal Gene Transfer (HGT) is a direct transfer of genetic material from one lineage to another. It plays an important role in the evolution of microorganisms. Identifying HGT accurately is one of the major challenges in molecular biology [1]. Numerous methods have appeared since the 1990s [2-10].

Maddison [2], Page [3] et al. considered evolutionary rules for modeling HGT. Hallett and Lagergren [4] described an exponential–time algorithm (called LatTrans algorithm) that maps numerous gene trees into a species tree. Mirkin et al. [5] designed an algorithm that provides in each situation a parsimonious evolutionary scenario consisting of mapping gene loss and gain events into a species tree. Moret et al. [6] presented an overview of the methods using network-based models to recover HGT. Hallett et al. [7] obtained a combinatorial model incorporating HGT and duplication events. Nakhleh et al. [8] developed the "RIATA-HGT" heuristic, the latest version of which is much faster than LatTrans while being almost equivalent in terms of HGT recovery [9].

Alix Boc et al. [10] presented a new algorithm for detecting HGT events which can rely on different criteria. They introduced the "bipartition dissimilarity" (BD) between two phylogenetic trees, and showed by simulation that the BD-based strategy outperforms least square, Robinson and Foulds distance [11], and quartet distance [12]. They also compared the BD-based algorithm with LatTrans [4] and RIATA-HGT [8, 9] and showed by simulation that the former is superior to the latters in terms of both HGT recovery and running time. The bipartition dissimilarity measure is also used in [13]

for detecting partial HGT events and provides better results than the existing algorithms.

To the best of our knowledge, there is no theoretical analysis on bipartition dissimilarity measure in the literature to date. In this paper we analyze this measure theoretically and obtain some new results. These results provide a better understanding of this novel measure and show its usefulness and powerfulness in a greater degree.

The remainder of this paper is organized as follows. In Section 2, we introduce some notations and terminology, illustrate the definition of bipartition dissimilarity measure, and compare it with the Robinson-Foulds distance. In Section 3, we analyze the minimum positive value of this measure and neighborhood. In Section 4, we study the local modifications under this measure. We conclude this paper in Section 5.

2 Preliminaries

For sets A, B, let $A \oplus B = (A \setminus B) \cup (B \setminus A)$ be their symmetric difference. Denote by $|A|$ the cardinality of set A. A *phylogenetic tree* is a tree whose leaves are labeled bijectively by a set L (species) and no vertex has degree 2. Let $|L| = n$. Denote by $\mathcal{T}_n$ the set of phylogenetic trees over L. A phylogenetic tree is *binary* if every non-leaf vertex has degree equal to 3.

A *bipartition* $A \mid B$ of L is an unordered pair (i.e., $A \mid B = B \mid A$) of its subsets, such that $A \cup B = L$ and $A \cap B = \phi$. Removing an edge from tree T disconnects the tree and induces a bipartition. Each pendant edge (one of its ends is a leaf) corresponds to a *trivial* bipartition,

---

* Corresponding authour's E-mail: sgliytu@hotmail.com

which must be present in every tree. Each internal edge corresponds to a *nontrivial* bipartition. Let $\beta(T)$ and $\beta_*(T)$ denote the set of bipartitions of T and the set of nontrivial bipartitions of T, respectively.

The Robinson-Foulds (RF) distance [11] is by far the most widely used distance measure for trees which counts the number of bipartitions present in one tree but not in the other. The RF distance between two trees $T_1, T_2 \in \mathcal{T}_n$, is defined as $RF(T_1, T_2) = |\beta_*(T_1) \oplus \beta_*(T_2)| / 2$. The main disadvantage of RF distance is that it lacks robustness in the face of small modifications. Consider a caterpillar tree as shown in Figure 1. A *caterpillar tree* is a tree in which all non-leaf vertices form a single path. When one leaf at one end of the tree is removed and reattached to the other end of the tree, we obtain a tree that is as far as possible in the RF distance i.e., $n-3$.
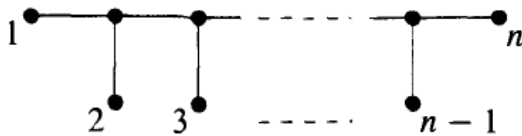

FIGURE 1 A caterpillar tree.

The *weight* of two bipartitions $A_1 | B_1$ and $A_2 | B_2$, is defined as follows [10, 14]:

$$wt(A_1 | B_1, A_2 | B_2)$$
$$= 0.5 \min\{| A_1 \oplus A_2 | + | B_1 \oplus B_2 |,$$
$$| A_1 \oplus B_2 | + | B_1 \oplus A_2 |\} \qquad (1)$$
$$= \min\{| A_1 | + | A_2 | - 2 | A_1 \cap A_2 |,$$
$$n - | A_1 | - | A_2 | + 2 | A_1 \cap A_2 |\}.$$

Given two trees T_1 and T_2. For each bipartition $A_1 | B_1$ of T_1, define its distance to T_2, $dis(A_1 | B_1, T_2)$ as $\min_{A_2 | B_2 \in \beta(T_2)} \{wt(A_1 | B_1, A_2 | B_2)\}$. Note that $dis(A_1 | B_1, T_2) = 0$ if $A_1 | B_1$ is a trivial bipartition of T_1. It is possible that for a nontrivial bipartition of T_1, $\arg\min_{A_2 | B_2 \in \beta(T_2)} \{wt(A_1 | B_1, A_2 | B_2)\}$ is a trivial bipartition of T_2. Similarly, for each bipartition $A_2 | B_2$ of T_2, $dis(A_2 | B_2, T_1)$ is defined as $\min_{A_1 | B_1 \in \beta(T_1)} \{wt(A_2 | B_2, A_1 | B_1)\}$. If $A_2 | B_2$ is a trivial bipartition of T_2, then $dis(A_2 | B_2, T_1) = 0$. It is possible that for a nontrivial bipartition of T_2, $\arg\min_{A_1 | B_1 \in \beta(T_1)} \{wt(A_2 | B_2, A_1 | B_1)\}$ is a trivial bipartition of T_1.

The *bipartition dissimilarity* between trees T_1 and T_2, $BD(T_1, T_2)$, can be calculated in the following way [10]:

$$BD(T_1, T_2)$$
$$= (\sum_{A_1 | B_1 \in \beta(T_1)} dis(A_1 | B_1, T_2) + \sum_{A_2 | B_2 \in \beta(T_2)} dis(A_2 | B_2, T_1)) / 2 \qquad (2)$$
$$= (\sum_{A_1 | B_1 \in \beta_*(T_1)} dis(A_1 | B_1, T_2) + \sum_{A_2 | B_2 \in \beta_*(T_2)} dis(A_2 | B_2, T_1)) / 2.$$

For instance, we compute the bipartition dissimilarity between trees T_1 and T_2 shown in Figure 2. We have the following nontrivial bipartitions for T_1: $ab | cde$, $abc | de$, and for T_2: $ac | bde$, $acd | be$. We get:

$$dis(ab | cde, T_2) = wt(ab | cde, a | bcde) = 1,$$
$$dis(abc | de, T_2) = wt(abc | de, ac | bde) = 1,$$
$$dis(ac | bde, T_1) = wt(ac | bde, abc | de) = 1,$$
$$dis(acd | be, T_1) = wt(acd | be, abc | de) = 2.$$

The bipartition dissimilarity between T_1 and T_2, $BD(T_1, T_2)$, is equal to $(1 + 1 + 1 + 2) / 2 = 2.5$.
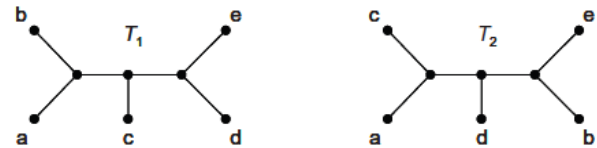

FIGURE 2 Two phylogenetic trees T_1 and T_2

The bipartition dissimilarity can be regarded as a weighted extension of the RF distance. In fact, the RF distance uses the following binary weighting scheme: the weight of two bipartitions $A_1 | B_1$ and $A_2 | B_2$, $wt_{RF}(A_1 | B_1, A_2 | B_2)$, is just 0 if they are identical and 1 otherwise. It is clear that the bipartition dissimilarity makes better use of the information in the bipartitions.

Theorem 1 Let $T_1, T_2 \in \mathcal{T}_n$. Then,

$$RF(T_1, T_2) \le BD(T_1, T_2) \le \left\lfloor \frac{n}{2} \right\rfloor RF(T_1, T_2).$$

Proof. Note that for any two non-identical bipartitions $A_1 | B_1$ and $A_2 | B_2$, $1 \le wt(A_1 | B_1, A_2 | B_2) \le \lfloor n / 2 \rfloor$. Hence we get

$$wt(A_1 | B_1, A_2 | B_2) \ge wt_{RF}(A_1 | B_1, A_2 | B_2) \text{ and}$$
$$wt(A_1 | B_1, A_2 | B_2) \le \lfloor n / 2 \rfloor \cdot wt_{RF}(A_1 | B_1, A_2 | B_2).$$

The desired inequalities follow from the above analysis. □

The bipartition dissimilarity measure is very similar to the matching split distance [14]. The main difference is that bipartition dissimilarity avoids the computation of a minimum-weight perfect matching in a complete bipartite

graph. On the other hand, the similarity between the two measures allows us to use some ideas of [14] in the next two sections.

3 Minimum positive value and neighborhood

Let $BD_{\min} = \min_{T_1,T_2 \in \mathcal{T}_n \wedge T_1 \neq T_2} BD(T_1,T_2)$. We call $BD_{\min}$ *the minimum positive value* of bipartition dissimilarity. To investigate $BD_{\min}$, we need the following definition.

Definition 1 *Nearest Leaf Interchange (NLI) means interchanging two leaves that are incident to the same internal edge.*

The generic form of a NLI operation is illustrated in Figure 3, where the operation is performed on the edge e, the circles A and B represent subtrees over sets of leaves A and B, and a and b represent two leaf labels.
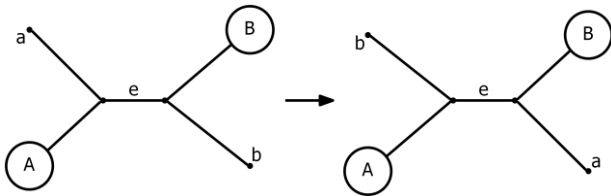

FIGURE 3 A schematic representation of the generic NLI operation

Theorem 2 $BD_{\min} = 1$, *which is achieved by performing a single NLI operation on a tree.*

Proof. Let T_1 and T_2 be two trees in $\mathcal{T}_n$. If $T_1 = T_2$, then we get $BD(T_1,T_2) = 0$. Otherwise, it must be true that there is a bipartition $A_1 \mid B_1$ which is in T_1 but not in T_2, and there is a bipartition $A_2 \mid B_2$ which is in T_2 but not in T_1. It follows that $dis(A_1 \mid B_1, T_2) \geq 1$, $dis(A_2 \mid B_2, T_1) \geq 1$. Hence we get $BD(T_1,T_2) \geq 1$.

If T_2 is obtained from T_1 by performing a single NLI operation, as shown in Figure 3, then we get

$$dis(A \cup \{a\} \mid B \cup \{b\}, T_2) = wt(A \cup \{a\} \mid B \cup \{b\}, A \cup \{a,b\} \mid B) = 1. \tag{3}$$

$$dis(A \cup \{b\} \mid B \cup \{a\}, T_1) = wt(A \cup \{b\} \mid B \cup \{a\}, A \cup \{a,b\} \mid B) = 1. \tag{4}$$

Since all the other branches of T_1 are left unchanged, we have $BD(T_1,T_2) = 1$. It follows that $BD_{\min} = 1$, which is achieved by performing a single NLI operation on a tree.

On the other hand, it is easy to see that if $T_2 \neq T_1$ and T_2 is not obtained from T_1 by performing a single NLI operation, then $BD(T_1,T_2) > 1$. $\square$

Two trees T_1 and T_2 in $\mathcal{T}_n$ are said to be *neighbors* if $BD(T_1,T_2) = 1$. The *neighborhood* of a tree $T \in \mathcal{T}_n$, denoted by $N(T)$, is the set of all trees that are neighbors of T. Theorem 2 permits us to create the neighborhood of a tree.

Theorem 3 *For a given tree T, the number of the trees in $N(T)$ is at most $n-1$, and it is possible that $\mid N(T) \mid= 0$.*

Proof. Figure 4 shows a tree T_1 with $\mid N(T_1) \mid= n-1$ and a tree T_2 with $\mid N(T_2) \mid= 0$. $\square$
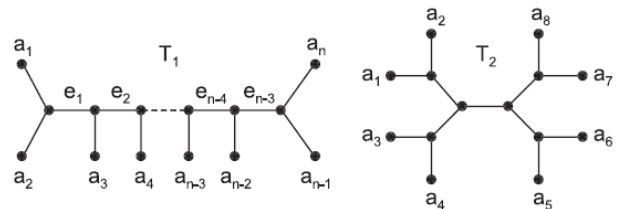

FIGURE 4 $\mid N(T_1) \mid= n-1$ and $\mid N(T_2) \mid= 0$

4 Local modifications

In order to investigate the local modifications under the bipartition dissimilarity measure, we need the following definitions.

Definition 2 [15] *Let $T \in \mathcal{T}_n$ and $X \subseteq L$. The restricted spanning tree $T(X)$ of T is the minimum subgraph of T that connects all the leaves whose labels are in X. The simplified spanning tree of T induced by X is a tree $T_{\mid X}$ obtained from $T(X)$ by replacing each maximal degree two path with an edge between the two ending vertices.*

Definition 3 [15] *Let $\mathcal{P} = \{T_1,T_2,\ldots,T_k\} \subseteq \mathcal{T}_n$ and X' be a maximum-sized subset of L for which $T_{1\mid X'} = T_{2\mid X'} = \cdots = T_{k\mid X'}$. This restricted subtree is called a maximum agreement subtree (mast) for $\mathcal{P}$.*

An example of a maximum agreement subtree is shown in Figure 5. In applications, identifying maximum agreement subtrees may help to exclude a small number of problematic species which may cause estimates of an evolutionary tree to vary greatly between different data sets.
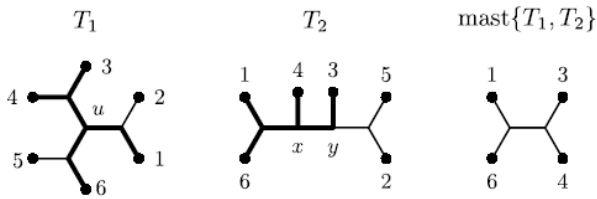
FIGURE 5 Two Trees and one of their maximum agreement subtrees

Theorem 4 *Let* $T_1, T_2 \in \mathcal{T}_n$ *and* $X = L \setminus \{x\}$.*Then*

(1) $BD(T_1, T_2) \geq BD(T_{1|X}, T_{2|X})$,

(2) $BD(T_1, T_2) \leq BD(T_{1|X}, T_{2|X}) + \left\lfloor \dfrac{3n}{2} \right\rfloor - 4.$ (5)

Proof. Each bipartition $A_1^X \mid B_1^X$ of $T_{1|X}$ corresponds to a bipartition $A_1 \mid B_1$ of T_1 such that $A_1 = A_1^X \cup \{x\}$ and $B_1 = B_1^X$, or $A_1 = A_1^X$ and $B_1 = B_1^X \cup \{x\}$. Thus, we get:

$dis(A_1 \mid B_1, T_2) \geq dis(A_1^X \mid B_1^X, T_{2|X})$, and

$dis(A_1 \mid B_1, T_2) \leq dis(A_1^X \mid B_1^X, T_{2|X}) + 1$. Since there are $2n - 4$ and $2n - 3$ bipartitions in $T_{1|X}$ and T_1 respectively, a bipartition of T_1 is left unmapped which we denote by $A_1' \mid B_1'$. We have $dis(A_1' \mid B_1', T_2) \leq \lfloor n/2 \rfloor$.

Similarly, each bipartition $A_2^X \mid B_2^X$ of $T_{2|X}$ corresponds to a bipartition $A_2 \mid B_2$ of T_2 such that $A_2 = A_2^X \cup \{x\}$ and $B_2 = B_2^X$, or $A_2 = A_2^X$ and $B_2 = B_2^X \cup \{x\}$. Thus, we get:

$dis(A_2 \mid B_2, T_1) \geq dis(A_2^X \mid B_2^X, T_{1|X})$, and

$dis(A_2 \mid B_2, T_1) \leq dis(A_2^X \mid B_2^X, T_{1|X}) + 1$. Since there are $2n - 4$ and $2n - 3$ bipartitions in $T_{2|X}$ and T_2 respectively, a bipartition of T_2 is left unmapped which we denote by $A_2' \mid B_2'$. We have $dis(A_2' \mid B_2', T_1) \leq \lfloor n/2 \rfloor$.

The desired two inequalities follow from the above analysis. □

Theorem 5 *Let* $T_1, T_2 \in \mathcal{T}_n$. *If the maximum agreement subtree of* T_1 *and* T_2 *has* $n - 1$ *leaves, then* $BD(T_1, T_2) \leq n - 2$ *and this bound is tight.*

Proof. As shown in Figure 6 (first appeared in [14]), T_1 and T_2 have different bipartitions only at the positions s_i, t_i for $i = 0, 1, \ldots, k-1$. For simplicity, we denote by s_i and t_i the bipartitions of T_1 and T_2 induced by the

edges s_i and t_i respectively, $i = 0, 1, \ldots, k-1$. Note that $wt(s_i, t_i) = 1$ for $i = 1, \ldots, k-1$. Moreover, $wt(s_0, t_0) = \min\{|C| + 1, n - |C| - 1\}$. Hence we get $BD(T_1, T_2) \leq k - 1 + n - |C| - 1 \leq n - 2$.

The bound $n - 2$ is tight. To see this, we move a single leaf labeled 1 of a caterpillar tree (shown in Figure 1) to the other end of the tree. Let T_1 and T_2 be the original tree and the obtained tree, respectively. It is easy to see that $BD(T_1, T_2) = n - 2$. □
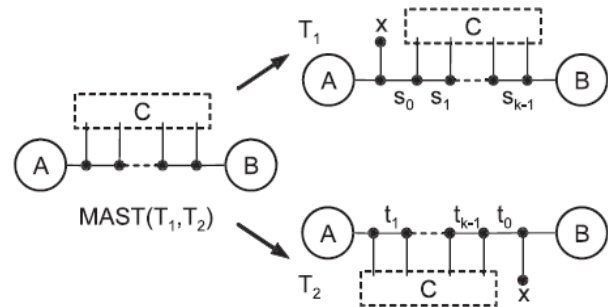

FIGURE 6 Two Trees differ only at the position of x

Although there are trees which have no neighbors as Theorem 3 shows, no isolated islands of trees distant from others exist indeed, since analysis analogous to Theorem 5.2 of [14] gives the following result.

Theorem 6 *Let* $T_a, T_b \in \mathcal{T}_n$. *There exists a sequence of trees in* $\mathcal{T}_n$, $T_a = T_1, T_2, \ldots, T_{k-1}, T_k = T_b$, *such that* $BD(T_i, T_{i+1}) \leq 2$ *where* $i = 1, \ldots, k-1$.

Proof. A *rooted caterpillar tree* is defined as a tree obtained from a unrooted caterpillar tree by inserting a degree 2 vertex as the root on one of the four outmost edges. All one-, two- and three-leaf binary rooted trees are considered to be caterpillars.

We will perform four transformations to locally modify T_a and T_b . Each transformation is reversible. Two of them are *rooting* (specifying a root for an unrooted tree) and *unrooting* (transforming a rooted tree into an unrooted one), the others consist of a series of basic operations. There are three types of basic operations which create trees at distances of 2, 0.5 and 2 respectively, as shown in Figure 7.

The first transformation transforms T_a and T_b into rooted trees RT_a and RT_b as follows: Take any leaf and let it be the roots of T_a and T_b respectively.

The second transformation transforms RT_a and RT_b into rooted caterpillar trees RCT_a and RCT_b . It starts from rooted subtrees of RT_a (RT_b) in bottom-up order by repeatedly applying the basic operations of types I and II. Suppose that we are processing the interior vertex

Li Shuguang, Chen Shuying, Cui Mengtian

v which is connected to two rooted caterpillars, as shown in Figure 7. We repeatedly transfer a leaf from the first rooted caterpillar into the second by performing Operations I and II, and finally two smaller rooted caterpillars are merged into a bigger one. Note that if RT_a and RT_b are binary rooted trees, then Operation II is unnecessary.

The third transformation transforms RCT_a and RCT_b into unrooted caterpillar trees UCT_a and UCT_b by canceling the root-specifying.
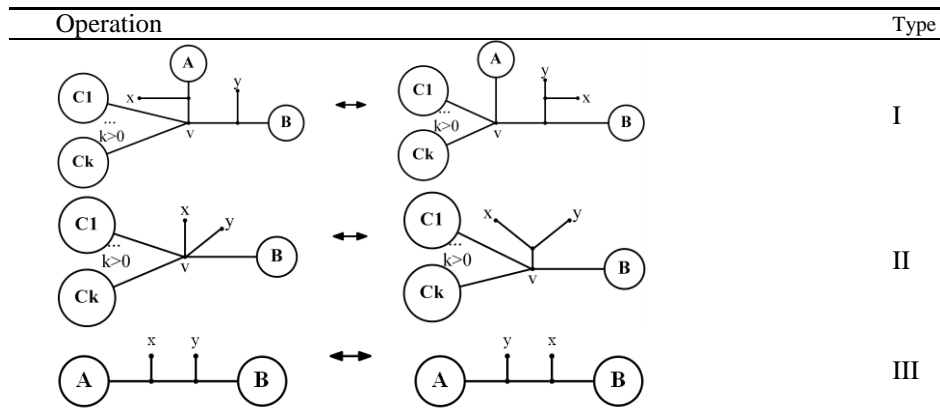

FIGURE 7 Three types of basic operations

The fourth transformation transforms UCT_a into UCT_b by repeatedly applying the basic operation of type III.

The desired sequence of trees is obtained easily from the sequence of transformations

$$T_a \rightarrow RT_a \rightarrow RCT_a \rightarrow UCT_a \rightarrow UCT_b \rightarrow RCT_b$$
$$\rightarrow RT_b \rightarrow T_b . \quad \square$$

5 Conclusions

Bipartition dissimilarity is a new measure and no theoretical analysis on it is known to date. We analyzed this measure in the paper and reported some theoretical properties of it, including minimum positive value, neighborhood, and local modifications. These results reduce the uncertainty of this measure, offer deeper insights into its behavior, and thus present a theoretical basis on which we can use this measure more efficiently and reliably. It would be interesting to investigate other properties of this measure, or use it in the other applications.

Acknowledgments

References

[1] Koonin E V 2003 Horizontal gene transfer: the path to maturity *Molecular microbiology* **50** 725-7
[2] Maddison W P 1997 Gene trees in species trees *Systematic biology* **46** 523-36
[3] Page R D, Charleston M A 1998 Trees within trees: phylogeny and historical associations *Trends in Ecology & Evolution* **13**, 356-9
[4] Hallett M T, Lagergren, J 2001 Efficient algorithms for lateral gene transfer problems *Proceedings of the fifth annual international conference on Computational biology* 149-56
[5] Mirkin B G, Fenner T I, Galperin M Y, Koonin E V 2003 Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes *BMC evolutionary biology* **3** 2
[6] Moret B. M., Nakhleh L., Warnow T., Linder C. R., Tholse A., Padolina A., et al 2004 Phylogenetic networks: modeling, reconstructibility, and accuracy *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **1** 13-23
[7] Hallett M, Lagergren J, Tofigh A 2004 Simultaneous identification of duplications and lateral transfers *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology* 347-56
[8] Nakhleh L, Ruths D, and Wang L-S 2005 RIATA-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer *Computing and Combinatorics,* Springer 84-93
[9] Than C, Nakhleh L 2008 SPR-based Tree Reconciliation: Non-binary Trees and Multiple Solutions *APBC* 251-60
[10] Boc H Ph, Makarenkov V 2010 Inferring and validating horizontal gene transfer events using bipartition dissimilarity *Systematic biology* **59** 195-211
[11] Robinson D, Foulds L R 1981 Comparison of phylogenetic trees *Mathematical Biosciences* **53** 131-47
[12] Estabrook G F, McMorris F, and Meacham C A 1985 Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units *Systematic Biology* **34** 193-200
[13] Boc H Ph, Makarenkov V 2011 Towards an accurate identification of mosaic genes and partial horizontal gene transfers *Nucleic acids research* **39** e144-e148
[14] Bogdanowicz D, Giaro K 2012 Matching split distance for unrooted binary phylogenetic trees *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **9** 150-60
[15] Semple C, Steel M. A 2003 *Phylogenetics* Oxford University Press

Authors	
	Li Shuguang, born in October, 1970, Ling County, Shandong Province, P. R. China **Current position, grades:** Associate professor in College of Computer Science and Technology, Shandong Institute of Business and Technology, China. **University studies:** B.Sc. of Mathematics from Shandong Normal University in China in 1993. M.Sc. of Operational Research and Cybernetics from Shandong University in China in 2001. Ph.D. of Operational Research and Cybernetics from Shandong University in China in 2007. **Scientific interest:** Combinatorial optimization, approximation algorithms, graph theory, and bioinformatics. **Publications:** more than 10 papers published in various journals. **Experience:** teaching experience of 18 years, 5 scientific research projects.
	Chen Shuying , born in December,1979, Dingzhou City, Hebei Province, P. R. China **Current position, grades:** lecturer in College of Computer Science and Technology, Shandong Institute of Business and Technology, China. **University studies:** B.Sc. from Shandong Institute of Business and Technology in China. M.Sc. from Liaoning Technical University in China. **Scientific interest:** Bioinformatics. **Publications:** more than 10 papers published in various journals. **Experience:** teaching experience of 7 years, 2 scientific research projects.
	Cui Mengtian, born in May, 1972, Jinjiang District, Chengdu city, P.R. China **Current position, grades:** Professor in School of Computer Science and Technology, Southwest University for Nationalities. **University studies:** Ph.D. of computer software and theory from CAS in 2010. **Scientific interest:** Trusted computing, trusted software technology and algorithms. **Publications:** more than 30 papers published in various journals. **Experience:** 16 years of experience in computer software and theory, more than 20 scientific research projects.