Reconstructing Ancestral Genomic Orders Using Binary Encoding and Probabilistic Models

Fei Hu[1,2], Lingxi Zhou[2], and Jijun Tang[1,2,⋆]

[1] School of Computer Science and Technology, Tianjin University, China
[2] Department of Computer Science & Engineering, Univ. of South Carolina, USA
jtang@cse.sc.edu

Abstract. Changes of gene ordering under rearrangements have been extensively used as a signal to reconstruct phylogenies and ancestral genomes. Inferring the gene order of an extinct species has the potential in revealing a more detailed evolutionary history of species descended from it. Current tools used in ancestral reconstruction may fall into parsimonious and probabilistic methods according to the criteria they follow. In this study, we propose a new probabilistic method called `PMAG` to infer the ancestral genomic orders by calculating the conditional probabilities of gene adjacencies using Bayes' theorem. The method incorporates a transition model designed particularly for genomic rearrangement scenarios, a reroot procedure to relocate the root to the target ancestor that is inferred as well as a greedy algorithm to connect adjacencies with high conditional probabilities into valid gene orders.

We conducted a series of simulation experiments to assess the performance of `PMAG` and compared it against previously existing probabilistic methods (`InferCARsPro`) and parsimonious methods (`GRAPPA`). As we learned from the results, `PMAG` can reconstruct more correct ancestral adjacencies and yet run several orders of magnitude faster than `InferCARsPro` and `GRAPPA`.

Keywords: ancestral genome, gene order, probabilistic method.

1 Introduction

1.1 Overview

Evolutionary biologists have had a long tradition in reconstructing genomes of extinct ancestral species. Mutations in a genomic sequence are made up not only at the level of base-pair changes but also by rearrangement operations on chromosomal structures such as inversions, transpositions, fissions and fusions [1]. Over the past few years, ancestral gene-order inference has brought profound predictions of protein functional shift and positive selection [2].

Methods for ancestral genome reconstruction either assume a given phylogeny that represents the evolutionary history among given species, known as the small

[⋆] Corresponding author.

Z. Cai et al. (Eds.): ISBRA 2013, LNBI 7875, pp. 17–27, 2013.

phylogeny problem (SPP); or search the most appropriate tree along with a set of ancestral genomes to fit the observed data, called the big phylogeny problem (BPP). Most of parsimony methods (such as GRAPPA [3], MGR [4,5]) typically solve the SPP exactly by searching a set of ancestral gene orders to minimize the sum of the rearrangement distance over the entire edges of the phylogeny. Ma proposed another method for the SPP in the probabilistic framework (InferCARsPro [6]) by approximating the conditional probabilities for all possible gene adjacencies in an ancestral genome.

Current methods such as GRAPPA and InferCARsPro are capable to handle modern whole-genome data due to their intrinsic high complexity. In this paper, we propose a new probabilistic method called PMAG to reconstruct ancestral genomic orders given a phylogeny. We conducted extensive experiments to evaluate the performance of PMAG with other existing methods. According to the results, our new method can outperform all the other methods under study and still run at least hundreds of times faster than GRAPPA and InferCARsPro.

1.2 Genome Rearrangement

Given a set of n genes labeled as $\{1, 2, \cdots, n\}$, a genome can be represented by an *ordering* of these genes. Each gene is assigned with an orientation that is either positive, written i, or negative, written $-i$. Two genes i and j form an *adjacency* if i is immediately followed by j, or, equivalently, $-j$ is immediately followed by $-i$. An *breakpoint* of two genomes is defined as an adjacency appears in one but not in the other.

Genome rearrangement operations can change the ordering of genes. An *inversion* operation (also called *reversal*) reverses a segment of a chromosome. A *transposition* is an operation that swaps two adjacent segments of a chromosome. In the case of multiple chromosomes, *translocation* breaks a chromosome and reattaches a portion of it to another chromosome. Later Yancopoulos et al. [8] proposed a universal double-cut-and-join (DCJ) operation that accounts for all common events which resulted in a new genomic distance that can be computed in linear time.

1.3 Parsimony Methods for Ancestral Gene-Order Reconstruction

To find a solution for SPP, parsimony algorithms typically iterate over each internal node to solve for the median genomes until the sum of all edge distances (tree score) is minimized. The median problem can be formalized as follows: give a set of m genomes with permutations $\{x_i\}_{1 \leq i \leq m}$ and a distance measurement d, find another permutation x_t such that the median score defined as $\sum_{i=1}^{m} d(x_i, x_t)$ is minimized. GRAPPA and MGR (as well as their recently enhanced versions) are two widely-referenced methods that implemented a selection of median solvers for phylogeny and ancestral gene-order inference. However solving even the simplest case of median problem when m equals to three is NP-hard for most distance measurements.

Exact solutions to the problem of finding a median of three genomes can be obtained for the inversion, breakpoint and DCJ distances. Among all the median solvers, the best one is the DCJ median solver proposed by Xu and Sankoff (`ASMedian` [9]) based on the concept of adequate subgraph. Adequate subgraphs allow decompositions of an multiple breakpoint graph into smaller and easier graphs. Though the `ASMedian` solver could remarkably scale down the computational expenses of median searching, it yet runs very slow when the genomes are distant.

1.4 Reconstructing the Ancestral Gene Order in Probabilistic Frameworks

The probabilistic approach `InferCARsPro` proposed by Ma [6] is based on Bayes' theorem such that every possible predecessor and successor of a signed gene i denoted as X_i in the ancestral genome x, given D_x representing the observed data, can be expressed as

$$P(X_i \ in \ x|D_x) = \frac{P(D_x|X_i \ in \ x)P(X_i \ in \ x)}{\sum_{j=1}^{q} P(D_x|X_j \ in \ x)P(X_j \ in \ x)} = \frac{P(D_x|X_i \ in \ x)}{\sum_{j=1}^{q} P(D_x|X_j \ in \ x)}$$

where priors are assumed equal and the likelihood $P(D_x|X_i \ in \ x)$ can be calculated recursively in a post-order traversal fashion summed over q possible configurations. Its transition matrix is defined as an extension of the Jukes-Cantor model such that probability of transition from any character to any different character is always equal.

Let $s_x(\cdot)$ denote the successor of a gene and $p_x(\cdot)$ denote the predecessor of a gene, an adjacency pair $A_x(i,j)$ can be viewed as $s_x(i) = j$ and $p_x(j) = i$ simultaneously. After finishing the calculation of conditional probabilities for every successor and predecessor relationships, the conditional probability of an adjacency $A_x(i,j)$ in genome x can be approximated as

$$P(A_x(i,j)|D_x) = P(p_x(j) = i|D_x) \times P(s_x(i) = j|D_x)$$

Finally a fast greedy algorithm is adopted to connect adjacencies into contiguous ancestral regions. Although `InferCARsPro` showed good results and speedup over parsimonious methods, it is still too slow and inaccurate when dealing with even small number of distant genomes.

We investigated the following intrinsic characteristics of `InferCARsPro` that account for its difficulties in handling complex datasets, which in turn motivated us to propose our new method.

- `InferCARsPro` uses a neutral model accounting for all changes of adjacencies, however biased model for phylogeny reconstruction has been successfully applied for genome rearrangement scenarios [11].
- The total number of states for each gene is exactly equal to $2 \times n - 2$ where n is the number of genes. Thus computing the likelihood score on such excessive number of states clearly incurs huge computational burden.

– The conditional probability of an adjacency is approximated from the predecessor and successor relations. Although such approximation is intuitive, it is more desirable to directly calculate the conditional probability of an adjacency.
– `InferCARsPro` requires branch lengths of a given phylogeny as part of its inputs, but it is not always handy to obtain in practice.

2 Algorithm Detail

Given the topology of a model tree and a collection of gene orders at the leaves, our approach first encodes the gene orders into binary sequences and estimates the parameters in the transition model for adjacency changes. Ancestral nodes in the model tree are inferred independently and in each inference, we reroot the model tree to have the target ancestor as the root of a new tree. Then we utilize a probabilistic inference tool to compute the conditional probabilities of all the adjacencies encoded in the binary sequence of the target ancestor. At last we use a greedy algorithm as used in Ma's work to connect the adjacencies into contiguous regions. We call our new approach *Probabilistic Method of Ancestral Genomics (`PMAG`).*

2.1 Encoding Gene Orders into Binary Sequences

A gene order can be expressed as a sequence of adjacency information that specifies presence or absence of all the adjacencies [10,11]. Denote the head of a gene i by i^h and its tail by i^t. We refer $+i$ as an indication of direction from head to tail ($i^h \rightarrow i^t$) and otherwise $-i$ as ($i^t \rightarrow i^h$). There are a total of four scenarios for two consecutive genes a and b in forming an *adjacency*: $\{a^t, b^t\}$, $\{a^h, b^t\}$, $\{a^t, b^h\}$, and $\{a^h, b^h\}$. If gene c is at the first or last place of a linear chromosome, then we have a corresponding singleton set, $\{c^t\}$ or $\{c^h\}$, called a *telomere*. A genome can then be expressed as a multiset of adjacencies and telomeres. For instance, a linear chromosome consists of four genes, (+1,+2,-3,-4,) can be represented by the multiset of adjacencies and telomeres $\{\{1^h\}$, $\{1^t, 2^h\}$, $\{2^t, 3^t\}$, $\{3^h, 4^t\}$, $\{4^h\}\}$. We further write 1 (0) to indicate presence (absence) of an adjacency and we consider only those adjacencies and telomeres that appear at least once in the input genomes. Table 1 shows an example of encoding two artificial genomes into binary sequences.

Given a dataset D with m species and each of n genes, let k indicate the total number of linear chromosomes in D, then there are up to $\binom{2n+2}{2}$ distinct adjacencies and telomeres. However in reality if the length of the binary sequences extracted from D is l, then l is typically far smaller. In fact, in the extreme case when genomes in D share no adjacency and telomere, l equals at most to $n \times m + k$, and since m and k are commonly much smaller than n, thus the length of the binary sequences for a dataset is usually linear rather than quadratic to the number of genes.

Table 1. Example of encoding gene orders into binary sequences

$$G_1 : (1,\ 2, -3)$$
$$G_2 : (3, -2,\ 1)$$

(a) Two signed linear genomes

	$\{1^h\}$	$\{1^t, 2^h\}$	$\{2^t, 3^t\}$	$\{3^h\}$	$\{2^h, 1^h\}$	$\{1^t\}$
G_1	1	1	1	1	0	0
G_2	0	0	1	1	1	1

(b) Binary sequences

2.2 Estimating Transition Parameters

Since we are handling binary sequences with two characters, we use a general time-reversible framework to simulate the transitions from presence (1) to absence (0) and vice versa. Thus the rate matrix is

$$Q = \{q_{ij}\} = \begin{bmatrix} \cdot & a \\ a & \cdot \end{bmatrix} \begin{bmatrix} \pi_0 & 0 \\ 0 & \pi_1 \end{bmatrix}$$

The matrix involves 3 parameters: the relative rate a, and two frequencies π_0 and π_1.

Severl models have been proposed to probabilistically characterize the changes of gene adjacencies by common types of rearrangement operations such as inversion, transposition as well as DCJ [7,11]. In this study, we use the model that has been successfully applied for phylogeny reconstruction in the context of genome rearrangement as suggested in [11]. In particular, every DCJ operation breaks two random adjacencies uniformly chosen from the gene-order string and subsequently creates two new ones. Since each genome contains $n + O(1)$ adjacencies and telomeres where n is the gene number and $O(1)$ equals to the number of linear chromosomes in the genome, thus the probability that an adjacency changes from presence (1) to absence (0) in the sequence is $\frac{2}{n+O(1)}$ under one operation. Since there are up to $\binom{2n+2}{2}$ possible adjacencies and telomeres, the probability for an adjacency changing from absence (0) to presence (1) is $\frac{2}{2n^2+O(n)}$. Therefore we come to the conclusion that the transition from 1 to 0 is roughly $2n$ times more likely than that from 0 to 1.

2.3 Inferring the Probabilities of Ancestral Adjacencies for the Root Node

In principle, our probabilistic inference is categorized as marginal reconstruction which assigns characters to a single ancestral genome at a time. Once we have the tree topology and binary sequences encoding the input gene orders, we use the extended probabilistic approach for sequence data described by Yang [12] to infer the ancestral gene orders at the root node. In the binary sequences, each site represents an adjacency with character either 0 (absence) or 1 (presence) and for each site we seek to calculate the conditional probability of observing that adjacency. As the true branch lengths are not available, we take advantage of the widely-used maximum-likelihood estimation from the binary sequences at the leaves to estimate the branch length.

Suppose x is the root of a model tree, then the conditional probability that node x has the character s_x at the site, given D_x representing the observed data at the site in all leaves of the subtree rooted at x, is

$$P(s_x|D_x) = \frac{P(s_x)P(D_x|s_x)}{P(D_x)} = \frac{\pi_{s_x} L_x(s_x)}{\sum_{s_x} \pi_{s_x} L_x(s_x)}$$

where π_{s_x} is the character frequency for s_x. The conditional probability in the form of $L_x(s_x)$ is defined as the probability of observing the leaves that belong to the subtree rooted at x, given that the character at node x is s_x. It can be calculated recursively in a post-order traversal fashion suggested by Felsenstein [13] as:

$$L_x(s_x) = \begin{cases} 1 & \text{if } x \text{ is a leaf with character} = s_x \text{ at the site} \\ 0 & \text{if } x \text{ is a leaf with character} \neq s_x \text{ at the site} \\ \left[\sum_{s_f} p_{s_x s_f}(t_f) L_f(s_f)\right] \times \left[\sum_{s_g} p_{s_x s_g}(t_g) L_g(s_g)\right] & \text{otherwise} \end{cases}$$

where f and g are the two direct descendants of x. $p_{ij}(t)$ defines the transition probability that character i changes to j after an evolutionary distance t. Following the deduction of transition probability in [13], our transition-probability matrix can be written as

$$p_{ij}(t) = \pi_j + e^{-t}(\delta_{ij} - \pi_j)$$

Here the δ_{ij} is 1 if $i = j$, otherwise δ_{ij} is 0. In order to set up the $2n$ ratio, we simply set the rate a to 1 and add a direct assignment of the two frequencies in the code. For instance, if the character frequencies are $\pi_0 = 0.1$ and $\pi_1 = 0.9$, then the rate of 0 to 1 transitions is 10 times as high as the rate of transitions in the other direction under the same evolutionary distance.

`RAxML` [14,15] is one of the most widely used program for sequence-data analysis which implements the method for ancestral sequence inference developed by Yang [12]. In this study, we modified `RAxML` to infer the conditional probabilities of gene adjacencies at all sites. Once we obtain the conditional probability of every adjacency for the target ancestor x, we can construct an adjacency graph for x in which each gene i corresponds to two nodes, i^h and i^t, and each adjacency is connected by an edge with weight equal to the conditional probability of seeing that adjacency in x. The problem of searching the longest path in such a graph by visiting each gene's head and tail exactly once is indeed NP-hard as shown in Tang and Wang's study [16]. As a trade-off for time efficiency in dealing with large-scale datasets, we adopted the same greedy algorithm used in Ma's work [18] to connect adjacencies into contiguous ancestral regions.

2.4 Rerooting the Tree Topology

To infer the genomic order of a non-root ancestral node x, if x is taken as the root of the tree such that only the leaves in the subtree of x are considered into the recursive calculation of likelihood, potentially many good adjacencies in the outgroup of the subtree will be neglected and result in a loss of information. To minimize the influence, we incorporate the technique of rerooting so that original tree is rearranged and the target node x becomes the root of a new tree. The procedure of rerooting is a standard procedure implemented in many phylogenetic tools and it also has found to be useful for ancestral genome reconstruction in [6].

3 Experimental Results

3.1 Experimental Design

Since actual ancestors are rarely known for sure, it is difficult to evaluate ancestral reconstruction methods with real datasets. In order to carry out a complete evaluation over a group of methods under a wide range of configurations, we conducted a collection of simulation experiments following the standard steps of such tests that have been extensively adopted [17,11].

In particular, a group of tree topologies were firstly generated with edge lengths representing the expected number of evolutionary operations. An initial gene order was assigned at the root so it can evolve down to the leaves following the tree topology mimicking the natural process of evolution, by carrying out a number of predefined evolutionary events. In this way, we obtained the complete evolutionary history of the model tree and the whole set of genomes it has.

We utilized the simulator proposed by Lin et al. [20] to produce birth-death tree topologies. With a model tree, we were able to produce genomes of any size and difficulty by simply adjusting three main parameters: the number of genomes m, the number of genes n, and the tree diameter d.

Predicted ancestral genomes produced from a method were evaluated in terms of the total number of correctly inferred adjacencies (i.e. those also appear in the true ancestral genomes) divided by the total number of adjacencies in both true genome and predicted genome. In particular, if D represents the set of gene adjacencies in the real genome and D' the predicted genomes. We calculate C, the rate of `correct adjacency` by:

$$C = \frac{|D \cap D'|}{|D \cup D'|} \times 100\%$$

Errors are in two parts. If a gene adjacency in D is missing in D', such a gene adjacency is called a `false negative (FN)` adjacency. The false negative rate measures the proportion of false negative adjacencies with respect to the total number of gene adjacencies in D and D'. The `false positive (FP)` rate is defined similarly, by swapping D and D'.

3.2 Comparing the Performance with Existing Probabilistic Method

Though probabilistic methods of ancestral reconstruction for rearrangement data are relatively new, they have shown great potential in both scalability and efficiency. As we have mentioned, `InferCarsPro` and `PMAG` both aim to formulate the conditional probabilities of gene adjacencies, however due to excessive number of states `InferCarsPro` has to handle, it is much more computationally demanding than `PMAG`. In this section, we compared the performance of `PMAG` to `InferCarsPro`.

Figure 1 (left) shows the assessment result of the two methods using datasets of 10 genomes and each of 1000 genes. From the figure, `PMAG` achieved better accuracies than `InferCarsPro` in all tests, with about 5 percentage points of improvements. Given datasets containing more genomes and genes, `InferCarsPro` encountered great difficulty to finish, while `PMAG` scales well to handle them within a few hours of computation (Figure 1 right).

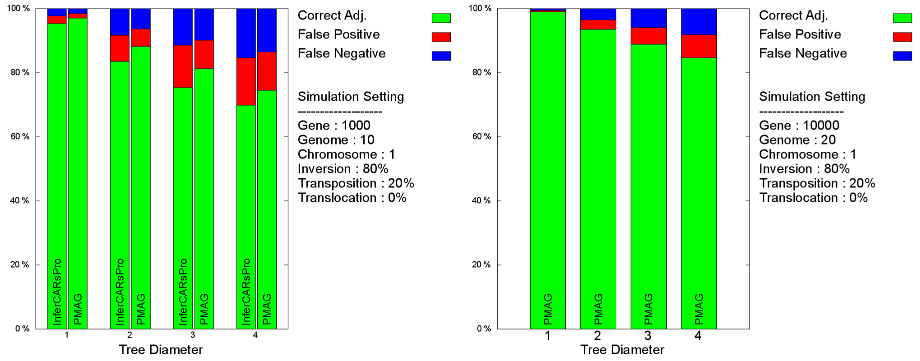

Fig. 1. Comparison between `PMAG` and `InferCARsPro`. X-axis represents the tree diameters from 1 to 4 times the number of genes.

3.3 Comparing the Performance with Parsimonious Methods

Parsimonious methods are in general time-consuming but very accurate. Their performances are sometimes referred as the upper bound of all methods [21], but such methods (`GRAPPA` for example) that directly optimize for the exact solution of the genome median problem suggested by Blanchette et al. [22] are NP-hard.

We compared the performance of `PMAG`, `InferCarsPro` and one direct optimization method `GRAPPA` with Xu's `ASMedian` solver [9] (`GRAPPA-DCJ`). Figure 2 shows the result of comparison. Because datasets are relatively easy, all methods can in average reconstruct more than 95% of true adjacencies and the differences among methods are not significant. However it is worth noting that `PMAG` receives less effect on tree diameters based on the observation that although `PMAG` performs sightly worse than `GRAPPA` methods under $0.6n$ tree diameter, it

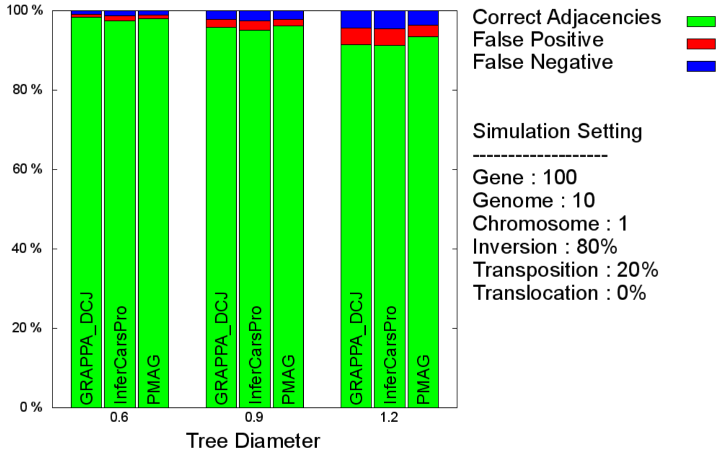

Fig. 2. Comparison among `PMAG`, `InferCARsPro` and `GRAPPA` with DCJ median solvers. X-axis represents the tree diameters that are 0.6, 0.9 and 1.2 times the number of genes.

can outperform the other methods at higher tree diameters. `InferCARsPro` is inferior to both `PMAG` and `GRAPPA` methods in the test which is consistent with the simulation results in Zhang et al.'s study [21].

3.4 Time Consumption

All tests were conducted on a workstation with 2.4Ghz CPUs and 4 GB RAM. We summarizes the running time of each method in the tests of Figure 1 and Figure 2 in Table 2 and Table 3 respectively. From table 2, we can see apparently `InferCARsPro` is computationally more demanding than `PMAG`, and hence restricted to handle small dataset. In table 3, both `InferCARsPro` and `GRAPPA` suffered significantly from high tree diameters, but tree diameter shows little impact on the running time of `PMAG`.

Table 2. Comparison of average time cost between two methods in seconds

Method	Genome#	Gene#	Tree Diameter			
			$1n$	$2n$	$3n$	$4n$
PMAG	10	1000	10	11	13	15
InferCARsPro	10	1000	5.4×10^3	1.4×10^4	2.9×10^4	7.2×10^4
PMAG	20	10000	2.4×10^3	3.6×10^3	5.7×10^3	9.5×10^3

Table 3. Comparison of average time cost between four methods in seconds

Tree Diameter	PMAG	InferCARsPro	GRAPPA-DCJ
0.6	1	300	8
0.9	1	1200	820
1.2	1	2600	7000

4 Conclusion

We introduced a new probabilistic method PMAG for ancestral gene-order inference. PMAG determines the state of each adjacency in the binary encoding to be either present or absent in an ancestral genome according to the conditional probability. Final ancestral genome is retrieved by connecting individual adjacencies into continuous regions. Experimental results show that ancestral genomes can be accurately inferred by PMAG. PMAG is also significantly faster in running time than InferCarsPro and parsimonious methods using direct optimization such as GRAPPA.

Much work remains to be done. In particular, we will try to extend our evolutionary model from rearrangements to a more general one in which other operations such as insertion (addition), duplication, or deletion (gene loss) are possible and hence introduce a new challenge to this study.

Acknowledgements. FH, LZ and JT are supported by grant NSF 0904179.

References

1. Kent, W., Baertsch, R., Hinrichs, A., Miller, W., Haussler, D.: Evolutions cauldron: duplication, deletion, andrearrangement in the mouse and human genomes. Proceedings of the National Academy of Sciences 100(20), 11484–11489 (2003)
2. Muller, K., Borsch, T., Legendre, L., Porembski, S., Theisen, I., Barthlott, W.: Evolution of carnivory in Lentibulariaceae and the Lamiales. Plant Biology 6(4), 477–490 (2008)
3. Moret, B., et al.: A New Implmentation and Detailed Study of Breakpoint Analysis. In: Pacific Symposium on Biocomputing (2001)
4. Guillaume, B., Pevzner, P.: Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Research 12(1), 26–36 (2002)
5. Max, A., Pevzner, P.: Breakpoint graphs and ancestral genome reconstructions. Genome Research 19(5), 943–957 (2009)
6. Ma, J.: A probabilistic framework for inferring ancestral genomic orders. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE (2010)
7. Sankoff, D., Blanchette, M.: Probability models for genome rearrangement and linear invariants for phylogenetic inference. In: Proceedings of the Third Annual International Conference on Computational Molecular Biology. ACM (1999)

8. Sophia, Y., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21(16), 3340–3346 (2005)
9. Xu, A.W., Sankoff, D.: Decompositions of multiple breakpoint graphs and rapid exact solutions to the median problem. In: Crandall, K.A., Lagergren, J. (eds.) WABI 2008. LNCS (LNBI), vol. 5251, pp. 25–37. Springer, Heidelberg (2008)
10. Hu, F., et al.: Maximum likelihood phylogenetic reconstruction using gene order encodings. In: 2011 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE (2011)
11. Lin, Y., Hu, F., Tang, J., Moret, B.: Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. In: Proc. 18th Pacific Symp. on Biocomputing, PSB 2013, pp. 285–296 (2013)
12. Yang, Z., Sudhir, K., Masatoshi, N.: A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141(4), 1641–1650 (1995)
13. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution 17(6), 368–376 (1981)
14. Stamatakis, A.: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21), 2688–2690 (2006)
15. Stamatakis, A.: New standard RAxML version with marginal ancestral state computationas, `https://github.com/stamatak/standard-RAxML`
16. Tang, J., Wang, L.: Improving genome rearrangement phylogeny using sequence-style parsimony. In: Fifth IEEE Symposium on Bioinformatics and Bioengineering, BIBE 2005, pp. 137–144. IEEE (2005)
17. Jahn, K., Zheng, C., Kováč, J., Sankoff, D.: A consolidation algorithm for genomes fractionated after higher order polyploidization. BMC Bioinformatics 13(suppl. 19), S8 (2012)
18. Ma, J., Zhang, L., Suh, B., Raney, B., Burhans, R., Kent, W., Blanchette, M., Haussler, D., Miller, W.: Reconstructing contiguous regions of an ancestral genome. Genome Research 16(12), 1557–1565 (2006)
19. Lin, Y., Rajan, V., Moret, B.: Bootstrapping phylogenies inferred from rearrangement data. BMC Algorithms for Molecular Biology 7, 21 (2012)
20. Lin, Y., Rajan, V., Moret, B.: Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator. J. Computational Biology 18(9), 1131–1139 (2011) (special issue on RECOMB-CG 2010)
21. Zhang, Y., Hu, F., Tang, J.: A mixture framework for inferring ancestral gene orders. BMC Genomics 13(suppl. 1), S7 (2012)
22. Blanchette, M., Bourque, G., Sankoff, D.: Breakpoint phylogenies. Genome Informatics 8, 25–34 (1997)