[Poster] Tools for genomic duplication inference under reconciliation based models.*

Jarosław Paszek$^{1[0000-0001-6442-3529]}$ and Paweł Górecki$^{1[0000-0002-2045-5892]}$

Warsaw University, Faculty of Mathematics, Informatics and Mechanics, Poland

Abstract. The search for the occurrences of gene duplications and multiple gene duplication episodes is important to understand the process of evolution. Therefore, many approaches decipher the locations of such phenomena in the Tree of Life. Such reconstruction can be done by the clustering of single gene duplications inferred by reconciling a set of gene trees with a species tree. Here, we present a software dedicated for such reconstructions.

Keywords: Gene duplication · Minimum episodes clustering · Reconciliation.

Background

The reconciliation model [4, 8] introduces a mapping for each node from a rooted gene family tree to its species tree and is either classified as a speciation event or a single gene duplication. That assignment defines a biologically consistent scenario in which evolutionary events are linked to the locations in the species tree. Such a scenario can be represented as the embedding of a gene tree into its species tree [6]. To model multiple gene duplications, the standard reconciliation approach was extended [5].

Genomic duplication clustering problem

The general problem description is as follows: given a set of gene trees, a species tree, and a cost function (to score a mapping scenario between gene trees and the species tree) find evolutionary scenarios for the collection of gene trees that yields the minimal score.

There are several variants of the above problem, in which:

- input gene trees may be rooted or unrooted,
- the model determines the episode locations in the species tree,
- clustering rules define which duplications can be clustered together,
- cost function scores the result basing on episode number and/or locations.

The first classification depends on the input which may consist of only rooted gene trees or any gene trees.

The next factor is the choice of the model of allowed evolutionary scenarios that determine the allowed locations of multiple gene duplication episodes in the species tree. The standard LCA model consists of only one scenario in which every gene tree is reconciled separately with its species tree. The model from [3] (called FHS) allows every biologically consistent scenario, while the

* The support was provided by the National Science Center grant 2017/27/B/ST6/02720.

PG model [9] allows only scenarios that preserve the minimal number of single gene duplications.

Then, the clustering rules choice determine which single gene duplications from one location in the species tree can be clustered in one multiple duplication episode. Here, we focus on episode clustering, and minimum episodes [1]. All duplications from the same location are clustered together in episode clustering. In minimum episodes clustering, a duplication and its ancestor duplication cannot be clustered together.

Finally, there can be various cost functions, e.g., the score that counts all episodes, or the score that equals the maximal number of episodes on one path [7].

Software

The poster aims to popularize the implementations of algorithms for various problems. Under the PG model and cost that optimizes the total episode number: the episode clustering [9] and minimum episodes [11] for input unrooted gene trees, and minimum episodes [10] for rooted gene trees (that tool can provide solutions for NP-hard variant [2] under FHS model). Recently, we proposed a tool for minimum episodes clustering under the PG model, where the score equals the maximal number of episodes on one path.

References

1. Bansal, M.S., Eulenstein, O.: The multiple gene duplication problem revisited. Bioinformatics **24**(13), i132–8 (2008)
2. Dondi, R., Lafond, M., Scornavacca, C.: Reconciling multiple genes trees via segmental duplications and losses. Algorithms for Molecular Biology **14**, 7 (2019)
3. Fellows, M., Hallet, M., Stege, U.: On the multiple gene duplication problem. In: 9th International Symposium on Algorithms and Computation (ISAAC'98), Lecture Notes in Computer Science 1533. pp. 347–356. Taejon, Korea (1998)
4. Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Systematic Zoology **28**(2), 132–163 (1979)
5. Guigó, R., Muchnik, I.B., Smith, T.F.: Reconstruction of ancient molecular phylogeny. Molecular Phylogenetics and Evolution **6**(2), 189–213 (1996)
6. Górecki, P., Tiuryn, J.: DLS-trees: A model of evolutionary scenarios. Theoretical Computer Science **359**(1-3), 378–399 (2006)
7. van Iersel, L., Janssen, R., Jones, M., Murakami, Y., Zeh, N.: Polynomial-time algorithms for phylogenetic inference problems. arXiv:1802.00317v2 [q-bio.PE] (9 Aug 2019)
8. Page, R.D.M.: Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Systematic Biology **43**(1), 58–77 (1994)
9. Paszek, J., Górecki, P.: Genomic duplication problems for unrooted gene trees. BMC Genomics **17**(1), 165–175 (2016)
10. Paszek, J., Górecki, P.: Efficient algorithms for genomic duplication models. IEEE/ACM Transactions on Computational Biology and Bioinformatics **15**(5), 1515–1524 (2018)
11. Paszek, J., Górecki, P.: Inferring duplication episodes from unrooted gene trees. BMC Genomics **19**(5), 288 (2018)