Evolution by Whole Genome Duplication

Yves Gagnon
DIRO, University of Montreal
y.gagnon@umontreal.ca

Mathieu Blanchette
McGill Centre for Bioinformatics
blanchem@mcb.mcgill.ca

Nadia El-Mabrouk
DIRO, University of Montreal
mabrouk@iro.umontreal.ca

Whole genome duplication (WGD) is a rare evolutionary event that has played a dramatic role in the diversification of most eucaryotic lineages. Given a set of species known to have evolved from a common ancestor through one or many rounds of WGDs, together with a set of genome rearrangements and gene losses, and given a phylogenetic tree for these species, the goal is to infer the ancestral genome just preceding the first WGD event. In [1] we have developed a two step approach: (1) Compute a score for each possible ancestral adjacency; (2) Combine adjacencies to form ancestral chromosomes. The main contribution of our methodology compared to other similar two-steps (local) approaches for inferring ancestral genomes, is the computation of a rigorous score for each potential ancestral adjacency (a, b), reflecting the maximum number of times a and b can be adjacent in the whole phylogeny, for any setting of ancestral genomes. This first step is done using a dynamic programming algorithm. The second step however is more heuristic and consists in chaining adjacencies in a way maximizing an objective function. In [1], we achieved this by solving a Traveling Salesman Problem (TSP) on a complete undirected graph where vertices correspond to genes, edges to potential adjacencies, and edge-weight to the adjacency score computed in the first step. The problem with this approach is that all genes are included in the final "best solution" of the TSP, even those for which no clear phylogenetic signal about their ancestral adjacencies is available. This is a serious drawback, as predicting less information but with high support is usually more useful than predicting everything but with no measure of reliability. Following this observation, we have developed a variant of the two-step previous approach, allowing to extend the notion of adjacencies to δ-adjacencies, for δ being an integer, where a δ-adjacency is a pair of genes in a genome separated by at most δ genes. More precisely, we generalize the dynamic programming algorithm used in the first step to compute δ-adjacency scores. We then iterate the two-step approach described above for increasing values of δ ($\delta = 0,\ 1,\ 2\ \cdots$). At the end of each iteration, we remove from the solution of the TSP, all edges with a score bellow a sufficiently selective threshold, leading to a large set of stretches or CARs (Contiguous Ancestral Regions). The following iteration is then performed on CARs' endpoints and free genes. The procedure terminates as soon as no more CARs can be chained together. We will test our new method on simulated datasets, and then use it to infer the pre-duplicated ancestor of the four completely sequenced grass species: Rice, Brachypodium, Maize and Shorgum.

References

[1] Bertrand, D., Gagnon, Y., Blanchette, M., El-Mabrouk, N.: Reconstruction of Ancestral Genome Subject to Whole Genome Duplication, Speciation, Rearrangement and Loss. WABI 2010, LNBI 6293, pp. 78- 89, (2010).