Reconstruction of ancestral plant genomes

Chunfang Zheng[1], Lingling Jin[2], Qiaoji Xu[1], Mohammad Sindeed Islam[1], Fatemeh S. Pouryahya[1], Alma Oladi[1], and David Sankoff[1]

[1] University of Ottawa, Canada
[2] Thompson Rivers University, Canada

Reconstruction methods depending on conserved gene adjacencies tend to break down in plants, largely because the history of whole genome doubling and tripling events (WGD and WGT, respectively) in the lineages of all plants effectively scrambles gene order and disrupts most adjacencies. After these events, most of the sets of duplicate or triplicate genes are eventually reduced to a single gene, by the redundance-eliminating process known as fractionation. Duplication of a genome fragment containing genes in the order 1 2 3 4 5 6 may result in two surviving orders 1 3 5 and 2 4 6, with none of the five fragment-internal adjacencies conserved, and only one adjacency at most conserved with the chromosomal regions surrounding each copy of the fragment. The situation is compounded if there are several WGD or WGT events in the history of some of the present-day plant genomes. All this is superimposed on a background of gene family expansion or contraction through tandem duplication or other mechanisms, complete loss of single-copy genes in species for which they are no longer physiologically or ecologically essential, genome rearrangement and other processes, all of which disrupt adjacencies independently of the fractionation process.

In attempting to reconstruct as much as possible of the ancestral flowering plant genome, based on new sequences from the Amborella Genome Project (cf [1]), we attempt to mitigate this problem in a number of ways. The most important is to try to retrieve as much as possible of the adjacency structure through use of generalized adjacency [2], based on a sliding window of size w, where all the genes in a window are considered adjacent. The basic reconstruction algorithm is Maximum Weight Matching (MWM) to produce a set of linear "contigs", based only on those adjacencies of weight two, i.e. occurring in at least two genomes, with post-processing to linearize circular contigs (of which we find very few), by breaking them at the weakest point, the lowest weight adjacency.

Another key feature is the use of syntenically validated adjacencies only, restricted to genes appearing in synteny blocks identified by the comparison of some pair of the descendant genomes. This avoids generating huge gene families and astronomical numbers of adjacencies not reflective of the ancestor. At the same time it increases the computing by the square of the number of genomes in the comparison. Thus we restrain our original search for adjacencies to ten flowering plant genomes, requiring 45 comparison runs of the SynMap program of the CoGe platform [3, 4]. Though this is not very onerous in itself it does not bode well for future applications of our methods to sets of dozens of genomes. To compensate for this we adopt a procedure of local expansion of the set of adja-

cencies. This involves auxiliary studies of single branches of the tree containing one or two of the original ten genomes plus the addition of five or so additional genomes on that branch. Additional adjacencies are produced within this more closely related group, which can then be used to bolster the original set input to MWM.

Validation of the "ancestral" contigs is carried out by mapping them onto the present-day genomes. For example, the core eudicots (hundreds of thousands of them) are descendants of a WGT event called " γ " [5] and in some of them the remnants of their tripled regions can be identified. The reconstructed contigs should (and do) map in the same triplicated pattern. Another potential direction is"consolidation", in which the ancient sequence is reconstructed from two fractionated versions of original identical chromosomal fragments [6, 7].

This approach requires fine tuning at several levels. The weighting of adjacencies may require adjustment due to biases introduced by comparing descendants of the same WGT, producing up to nine copies of some adjacencies. Increasing the window size w slowly increases the number of generalized adjacencies and slowly increases the average and maximum length of contigs. At the same time, longer potential contigs compete for genes with each other and solutions become unstable, so that an optimal trade-off must be determined in the interval $5 \leq w \leq 10$. This is a greater limitation on the window-size method than the noise introduced by spurious adjacencies of genes that are coincidentally close in two genomes, but never neighbours and never even very close.

References

1. The Amborella genome and the evolution of flowering plants (Amborella genome project) Science 342, 1241089 (2013).
2. Natural parameter values for generalized gene adjacency (Z Yang, D Sankoff) Journal of Computational Biology 11, 1113–1128 (2010)
3. How to usefully compare homologous plant genes and chromosomes as DNA sequences (E Lyons, M Freeling) The Plant Journal 53, 661–673 (2008)
4. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids (E Lyons, B Pedersen, J Kane, M Alam, R Ming, H Tang *et al.*) Plant Physiology 148, 1772–1781(2008)
5. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla (O Jaillon, JM Aury, B Noel, A Policriti, C Clepet, *et al.*) Nature 449, 463–467 (2007)
6. Fractionation, rearrangement and subgenome dominance (C Zheng, D Sankoff) Bioinformatics 28, i402–i408 (2012)
7. A consolidation algorithm for genomes fractionated after higher order polyploidization (K Jahn, C Zheng, J Kováč, D Sankoff) BMC Bioinformatics 13, S19:S8 (2012).