Improved sequence-based orthologs identification using genomic context information and their impact on pathway analysis in plants

Tulpan, Dan; Lger, Serge; Cuperlovic-Culf, Miroslava; Pan, Youlian

Dan Tulpan[1], Serge Léger[1], Miroslava Cuperlovic-Culf[1], Youlian Pan[2]
[1] **NRC Information and Communications Technologies**, Moncton, New Brunswick, E1A 7R1, Canada
[2] **NRC Information and Communications Technologies**, Ottawa, Ontario, K1A 0R6 , Canada
dan.tulpan@nrc-cnrc.gc.ca, serge.leger@nrc-cnrc.gc.ca, miroslava.cuperlovic-culf@nrc-cnrc.gc.ca, youlian.pan@nrc-cnrc.gc.ca

Improved Sequence-Based Orthologs Identification using Genomic Context Information and Their Impact on Pathway Analysis in Plants

ABSTRACT

With the advent of sequencing techniques, a deluge of plant genome projects have emerged, all prompting for accurate and high throughput comparative genomic approaches such as orthology prediction. The current incompleteness, polyploidy and low coverage of most plant genomes prompt for further improvements of orthology prediction using evolutionary-related information such as sequence variability and gene order.

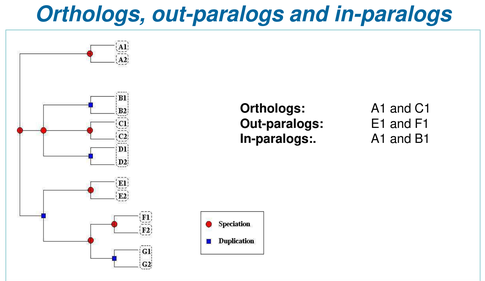
While a majority of orthology prediction approaches for large genome-scale datasets typically relies on reciprocal-best-BLAST-hits (RBBH), they suffer from insufficiencies related to incorrect prediction of paralogs as orthologs when incomplete genome sequences or gene loss are present. In addition, there is an increasing interest to identify orthologs most likely to have retained similar function.

To address these issues, we have developed a high-throughput multi-threaded computational approach that predicts orthologs using DNA and protein sequences and identifies which orthologs have similar genomic context and are likely to have similar function. First, we predict putative orthologs using commonly predicted DNA and protein based RBBHs. This dual approach is used whenever possible to reduce the number of false positives. Second, genomic context conservation is used to provide further support for orthologs assignment and to help with the identification of missing orthologs. Orthologs are predicted to have a higher likelihood of being similar in function if their relative genomic context is conserved. Third, the list of putative orthologs for pairs of plant species (e.g. *B. distachyon* and *S. bicolor*) is used to explore pathway similarities for the same biological process and discover putative enzymes omitted in some plant species.

INTRODUCTION

Orthologs are genes that have diverged after a speciation event. Orthologs most often have equivalent functions. Paralogs are genes related via a duplication event. Paralogs can be subdivided in out-paralogs and in-paralogs. Out-paralogs represent paralogous genes resulting from a duplication event preceding a speciation event. In-paralogs are paralogous genes resulting from a duplication event subsequent to a speciation event [2].

Orthologs, out-paralogs and in-paralogs



Orthologs: A1 and C1
Out-paralogs: E1 and F1
In-paralogs: A1 and B1

PROBLEM

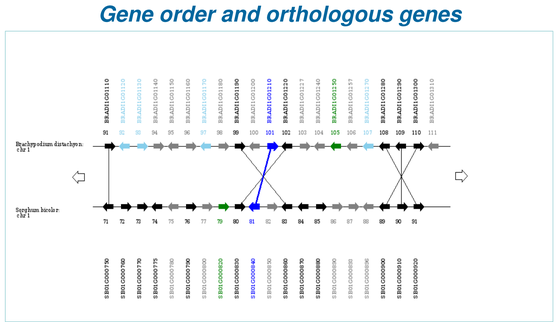
Given two plant species find a set of orthologous genes that have a high probability of having similar functions.

METHODS

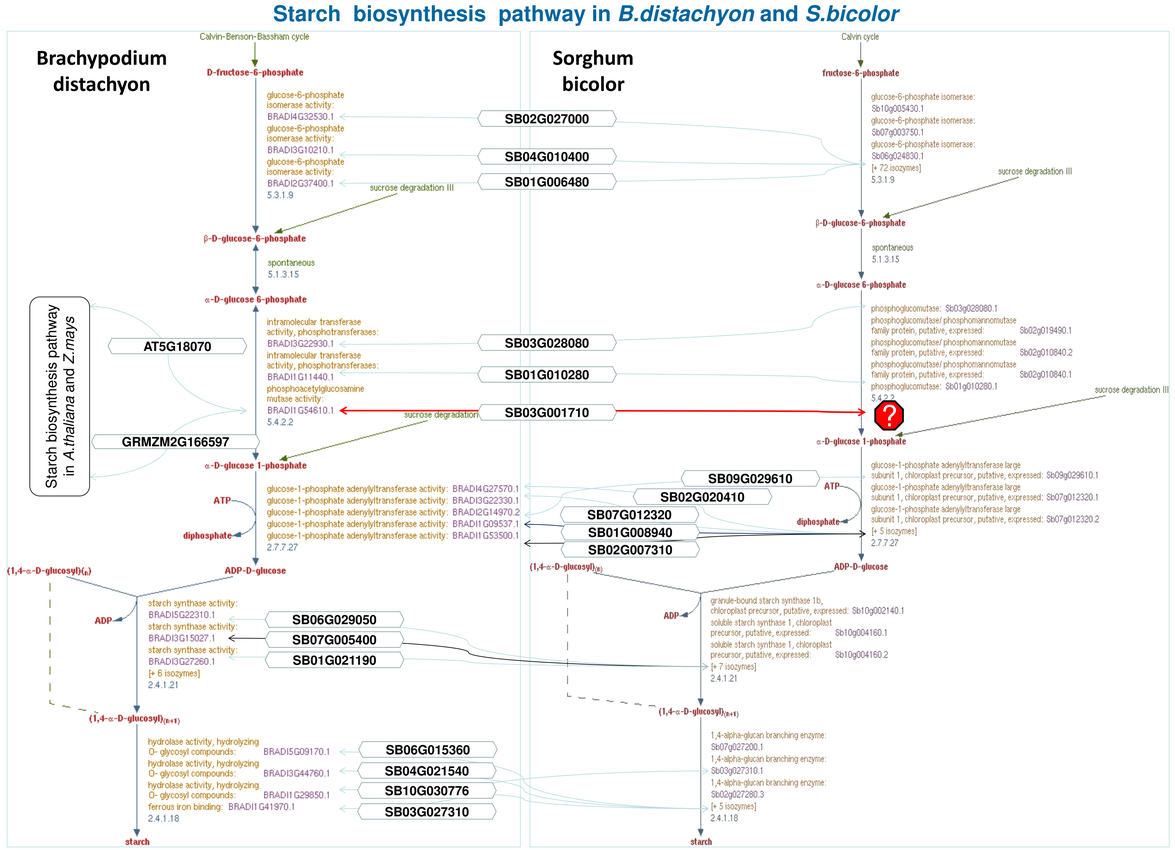
Several orthology prediction methods have been developed and they can be roughly grouped as phylogeny-based and all-versus-all sequence comparison-based. In methods relying on phylogenetic analysis via tree reconciliation, the topology of a gene tree is compared with a species tree and the two trees are reconciled on the basis of the parsimony principle (by postulating the minimal possible number of duplication and gene-loss events in the evolution of a given gene). These methods are typically limited by horizontal gene transfer events (especially in prokaryotes), and by the computationally expensive automated construction and analysis of gene trees.
All-versus-all sequence comparisons-based methods rely on two assumptions: (i) sequences of orthologous genes/proteins are more similar to each other than they are to any other genes/proteins from the compared genomes (RBH=Reciprocal Best BLAST Hits), and (ii) RBHs are most likely to be formed by orthologs. The limitations of these methods typically stem from data incompleteness such as missing genes that will lead to paralogs being predicted as orthologs.

To alleviate these insufficiencies, additional information is needed such as genomic context (e.g. gene order, annotations, etc.). Previous studies suggest that gene order is usually extensively conserved (especially in prokaryotes) between closely related species, but rapidly becomes less conserved among more distantly related organisms.

Gene order and orthologous genes



Starch biosynthesis pathway in *B.distachyon* and *S.bicolor*

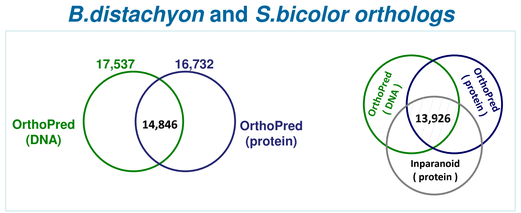


Source: Gramene [1]
B.distachyon: http://pathway.gramene.org/BRACHY/new-image?type=PATHWAY&object=PWY-622
S.bicolor: http://pathway.gramene.org/SORGHUM/new-image?type=PATHWAY&object=PWY-622&detail-level=2&EXP-ONLY=NIL

RESULTS

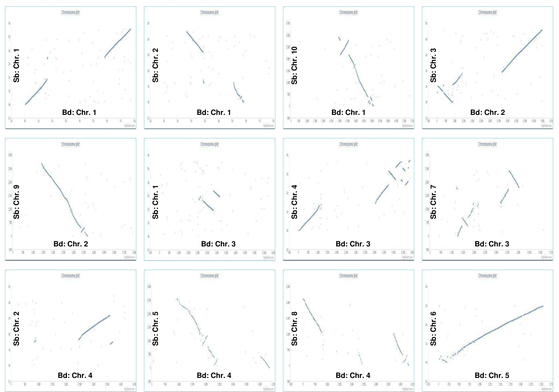
Using an in-house orthology prediction method based on all-versus-all sequence comparisons, we predicted approximately 15K *Sorghum-Brachypodium* orthologs, of which ~94% are also predicted by Inparanoid.

B.distachyon and *S.bicolor orthologs*



A chromosome-level gene order analysis reveals highly conserved areas (syntenic blocks) between the two plant species, which is in agreement with previous studies.

Syntenic dot plot for B.distachyon and S.bicolor



To further test the utility of the predicted orthologs we used the list of putative orthologs for *B.distachyon* and *S.bicolor* to explore pathway similarities for the same biological process (starch synthesis) and discover putative enzymes omitted in one of the two species (Sorghum gene; SB03G001710; located on the forward strand of chromosome 3: 1,535,581-1,538,656; putative uncharacterized protein).

CONCLUSION

Genome evolution research in different plant species via comparative genomics (e.g. orthology prediction) can shed light on the extent to which similar (orthologous) genes are involved in producing similar phenotypes and confer increased resistance to biotic and abiotic factors in phylogenetically-related species and working out how to transfer desirable genes between species to further improve cultivars. Such research corroborated with systemic studies involving metabolic pathways will lead to a better understanding of phenotypic traits in plants and their impact on breeding.

REFERENCES

[1] Youens-Clark, K., Buckler, E., Casstevens, T., Chen, C., Declerck, G., Derwent, P., Dharmawardhana, P., Jaiswal, P., Kersey, P., Karthikeyan, A.S., Lu, J., McCouch, S.R., Ren, L., Spooner, W., Stein, J.C., Thomason, J., Wei ,S., & Ware, D. (2010). Gramene database in 2010: updates and extensions. *Nucleic Acids Res. 39(Database issue)*, D1085-94.

[2] Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet. 39*, 309-38.

Poster presented at the Plant and Animal Genome Conference (PAG Asia 2013) in March 2013 in Singapore.

National Research Council Canada | Conseil national de recherches Canada

Canada