BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 1 2008, pages 127–128
doi:10.1093/bioinformatics/btm449

Genome analysis

i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles

Cedric Simillion[1], Koen Janssens[2], Lieven Sterck[3,4] and Yves Van de Peer[3,4,*]

[1]Institute for Cell and Molecular Biosciences (ICaMB), Newcastle University, Newcastle-upon-Tyne, UK, [2]Departement Industriële Wetenschappen BME-CTL, Hogeschool Gent, B-9000 Ghent, [3]Department of Plant Systems Biology, VIB and [4]Bioinformatics and Evolutionary Genomics, Department of Molecular Genetics, Ghent University, B-9052 Ghent, Belgium

Received on May 28, 2007; revised on August 18, 2007; accepted on August 24, 2007

Advance Access publication October 17, 2007

Associate Editor: Alex Bateman

ABSTRACT

Summary: i-ADHoRe is a software tool that combines gene content and gene order information of homologous genomic segments into profiles to detect highly degenerated homology relations within and between genomes. The new version offers, besides a significant increase in performance, several optimizations to the algorithm, most importantly to the profile alignment routine. As a result, the annotations of multiple genomes, or parts thereof, can be fed simultaneously into the program, after which it will report all regions of homology, both within and between genomes.

Availability: The i-ADHoRe 2.0 package contains the C++ source code for the main program as well as various Perl scripts and a fully documented Perl API to facilitate post-processing. The software runs on any Linux- or -UNIX based platform. The package is freely available for academic users and can be downloaded from http://bioinformatics.psb.ugent.be/

Contact: yves.vandepeer@psb.ugent.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Identifying genomic homology (i.e. the common descent of chromosomal segments) within and between genomes is essential when studying various aspects of genome evolution such as genomic rearrangements or genome duplication. In the past years, different computational techniques have been developed to detect homology even when the actual similarity between homologous segments is low. Since point mutations rapidly reduce primary sequence similarity (Vandepoele *et al*., 2004), most of these methods detect similarity between two genomic segments by comparing their overall gene content and, optionally, gene order (Simillion *et al*., 2004b).

Depending on the strategy used, these so-called 'map-based approach' methods search for pairs of chromosomal segments between which either both gene content and gene order are conserved or gene content only. However, due to the fact that, after their divergence, homologous segments can lose a different set of genes, these methods still often fail to detect genomic homology. This can partly be overcome by inferring transitional homology relationships from multiple genome comparisons (Dietrich *et al*., 2004; Kellis *et al*., 2004; Vandepoele *et al*., 2002). However, this approach still requires that at least some pairs of segments can be found that show sufficiently significant collinearity between them. One can however imagine a homology detection problem where, within a set of mutually homologous segments (hereafter referred to as a multiplicon), one or more segments have diverged so much from the others in gene content and gene order that they no longer show any significant collinearity with any of the other segments.

Recently, an advanced algorithm, i-ADHoRe (Simillion *et al*., 2004a), has been developed that can combine gene content and gene order information of multiple genomic segments. The i-ADHoRe software first identifies pairs of homologous segments from a dataset containing one or more genomes. Next, so-called genomic profiles are built by aligning these segment pairs such that homologous genes are placed in the same column. Each profile is then used to search the entire dataset again. This is done by superimposing the collinearity-comparisons with a given genomic region of each individual segment in the profile.

If a segment is found that shows statistically significant collinearity to the profile, meaning that the number and density of pairs of collinear genes is greater than expected by chance, it is added to the multiplicon and its profile. The dataset is searched again using the updated profile and the whole process is repeated until no more additional segments can be found.

Although the first version of this tool has proved highly successful in elucidating the evolutionary past of several eukaryotic genomes (Cannon *et al*., 2006; Dujon *et al*., 2004; Palenik *et al*., 2007), it still suffered from several drawbacks. First of all, the performance speed was very slow, especially on complex multi-genome datasets. Next, the alignment algorithm used was rather inefficient causing suboptimal decisions that are sometimes taken early in the process to propagate during later steps, which results in lower sensitivity (see below). Here, we present a new version of the i-ADHoRe software that aims to overcome the aforementioned weaknesses and adds several additional improvements as well.

*To whom correspondence should be addressed.

2 IMPLEMENTATION

2.1 A new greedy graph-based alignment algorithm

In the original version of i-ADHoRe, profiles were created by aligning the first two segments with the Needlemann–Wunsch algorithm, using pairs of homologous genes as identities and non-homologous genes as mismatches. New segments were added by aligning them in the same way one by one against the entire existing profile. The drawback of this method is that any two genes that were aligned early on in the process cannot change their relative position later on in the alignment procedure. This causes erroneous decisions to propagate throughout the data, as higher-level multiplicons (multiplicons containing more segments) are detected using lower-level profiles.

We have therefore developed an alignment method that takes these considerations into account. In short, whenever a new segment is added to an existing profile, the greedy graph-based algorithm constructs a completely new alignment of all segments simultaneously rather than extending the existing alignment with the new segment. This way, erroneous decisions made in early alignment steps can still be corrected in later steps. Figure 1 illustrates the improvement in alignment quality gained by this new method. A statistical analysis shows that the fraction of misaligned genes using this new method decreases significantly (see Supplementary Material).

2.2 Improved statistical validation

When the algorithm detects a candidate multiplicon, its statistical significance is calculated as a function of the number of pairs and distance between the pairs of collinear genes (referred to as anchor points) that make up the multiplicon (Simillion *et al.*, 2004a). In brief, this calculation is performed by multiplying the probabilities p_i of finding each anchor point with each other to obtain a p-value p_m for the entire multiplicon. In the original version of i-ADHoRe, the statistical validation of candidate multiplicons that were found using a higher-level profile applied a correction of the pm-value of the entire multiplicon to compensate for the multiple segments in the profile used. This was done by multiplying this value by l-1 where l is the multiplication level (the number of segments) of the profile. The new version however applies a correction for the p-value of each individual candidate anchor point, p_i, since the use of profiles actually increases the probability of finding more anchor points.

2.3 The i-ADHoRe package and Perl API

The program was entirely rewritten in C++ resulting in a 16-fold increase in performance speed in comparison to the previous version that was written in Perl.

Because of its complex output, the i-ADHoRe package includes a set of post-processing scripts that help to visualize and interpret the results generated. Next to this, a fully documented Perl API is provided, that allows the user to quickly write scripts to perform additional analyses on the output data.

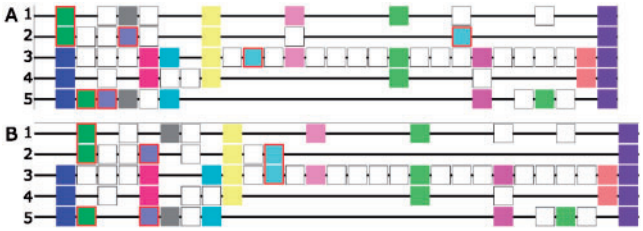

Fig. 1. Comparison of alignments of the same i-ADHoRe profile consisting of three (duplicated) *Arabidopsis thaliana* and two (duplicated) rice (*Oryza sativa*) genomic segments. (**A**) Alignment generated with the traditional, stepwise method. (**B**) Alignment with the new greedy graph-based alignment method. Each line represents a chromosome segment. The boxes on the segments represent genes. Genes of the same color are homologs. The genes marked in red are misaligned using the traditional method but are correctly aligned using the new greedy graph-based method. Note that, in order to limit the length of the entire profile, non-homologous genes are allowed in the same column. The segments 1–5 are respectively from *Arabidopsis* chromosome I, loci At1g52730.1 to At1g52830.1; *Arabidopsis* chromosome III, loci At3g15610.1 to At3g15540.1; rice chromosome I, loci Os01g08020.1 to loci Os01g08320.1; rice chromosome V, loci Os05g08430.1 to Os05g08570.1 and *Arabidopsis* chromosome I, loci At1g15460.1 to At1g15510.1.

ACKNOWLEDGEMENTS

The authors would like to thank Klaas Vandepoele for extensive testing of the software and helpful discussion. C.S. is supported by a Marie Curie Intra-European Fellowship from the European Commission.

Conflict of Interest: none declared.

REFERENCES

Cannon,S.B. *et al.* (2006) Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. *Proc. Natl Acad. Sci. USA*, **103**, 14959–14964.

Dietrich,F.S. *et al.* (2004) The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science*, **304**, 304–307.

Dujon,B. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.

Kellis,M. *et al.* (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, **428**, 617–624.

Palenik,B. *et al.* (2007) The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl Acad. Sci. USA*, **104**, 7705–7710.

Simillion,C. *et al.* (2004a) Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res.*, **14**, 1095–1106.

Simillion,C. *et al.* (2004b) Recent developments in computational approaches for uncovering genomic homology. *Bioessays*, **26**, 1225–1235.

Vandepoele,K. *et al.* (2002) Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. *Trends Genet.*, **18**, 606–608.

Vandepoele,K. *et al.* (2004) The quest for genomic homology. *Curr. Genomics*, **5**, 299–308.