BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 1 2004, pages 122–124
DOI: 10.1093/bioinformatics/btg378

GRIL: genome rearrangement and inversion locator

Aaron E. Darling[1,*], Bob Mau[2], Frederick R. Blattner[3] and Nicole T. Perna[2]

[1]Department of Computer Science and [2]Department of Animal Health and Biomedical Science, University of Wisconsin-Madison, 1656 Linden Dr, Madison, WI 53706, USA and [3]Department of Genetics, University of Wisconsin-Madison, 445 Henry Mall, Madison, WI 53706, USA

Received on February 26, 2003; revised on May 30, 2003; accepted on July 22, 2003

ABSTRACT

Summary: GRIL is a tool to automatically identify collinear regions in a set of bacterial-size genome sequences. GRIL uses three basic steps. First, regions of high sequence identity are located. Second, some of these regions are filtered based on user-specified criteria. Finally, the remaining regions of sequence identity are used to define significant collinear regions among the sequences. By locating collinear regions of sequence, GRIL provides a basis for multiple genome alignment using current alignment systems. GRIL also provides a basis for using current inversion distance tools to infer phylogeny.

Availability: GRIL is implemented in C++ and runs on any x86-based Linux or Windows platform. It is available from http://asap.ahabs.wisc.edu/gril

Contact: darling@cs.wisc.edu

Supplementary information: The GRIL web site contains a detailed description of the GRIL's algorithms, an example of applying GRIL to five genomes, and verification of the correctness of the results.

INTRODUCTION

As genomes evolve, inversions and other rearrangements lead to a loss of collinearity between chromosomes of related organisms. Analyses of these events can be used to reconstruct phylogenies or to gain insight into the rates and patterns of the events themselves. Many previous efforts have focused on inferring a series of reversal events that transform one genome organization to another, given a set of ordered landmarks from each genome (Moret *et al.*, 2001; Ajana *et al.*, 2002; Bourque and Pevzner, 2002; Larget *et al.*, 2002). Orthologous genes are a common choice for landmarks for analyses of rearrangements among mitochondrial or bacterial genomes.

However, genome rearrangements need not correspond directly to gene boundaries. GRIL is a software tool that automates identification of genome rearrangements independently of gene boundaries. The rearrangements identified by GRIL delineate collinear sequence regions shared by all genomes under study called locally collinear blocks (LCBs). Each LCB contains regions of homologous sequence and may contain regions of genome-specific sequence, but does not contain any rearrangements of homologous sequences that meet user defined criteria (e.g. length of rearrangement). The boundaries of each LCB found by GRIL correspond to the breakpoints of significant chromosomal rearrangements. As such, the LCBs are suitable input for phylogenetic reconstructions based on inversion distances.

The LCBs found by GRIL are also useful in the context of multiple genome alignment. All current multiple genome alignment systems assume collinearity among the sequences being aligned (Hohl *et al.*, 2002; Brudno *et al.*, 2003). Inversions and rearrangements in otherwise alignable genomes prevent a successful alignment. By locating the collinear regions within each genome, GRIL can provide the information necessary to complete a multiple genome alignment using existing tools.

Previous work related to identifying collinear regions of sequence includes a software tool called GRIMM-Synteny (Pevzner and Tesler, 2003). GRIMM-Synteny takes a set of potentially homologous regions between two sequences as input and outputs the location of synteny blocks. The synteny blocks reported by GRIMM can then be used as input to their phylogenetic reconstruction algorithm. GRIL differs from GRIMM-Synteny in two fundamental ways. First, it can locate LCBs shared by more than two genomes. Second, it has a built-in algorithm to search for regions of sequence identity, so that it can take sequence data as input and produce the location of LCBs in multiple genomes as output.

METHODS

The algorithm used by GRIL to compute LCBs is composed of three basic steps. First, a string-matching algorithm locates

[*]To whom correspondence should be addressed.

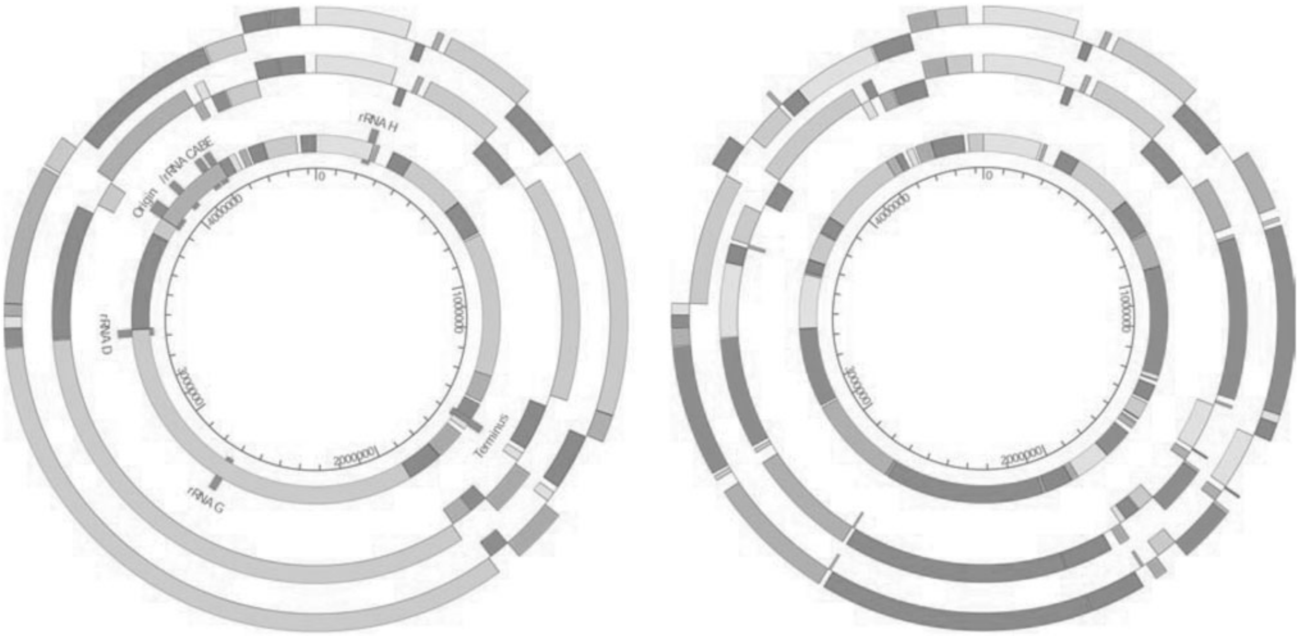

Fig. 1. The left set of concentric circles depicts collinear regions found by GRIL using a 10 kb minimum-length threshold for each LCB. The right set of circles shows the same genomes using a 5 kb minimum LCB length threshold. The inner circle is *E.coli* K-12 MG1655, the middle is *S.flexneri* 2A str 301 and the outer is the *S.flexneri* 2A str 2457T. The K-12 genome provides a reference orientation for inversions in the *Shigella* genomes. Forward oriented LCBs are placed outside each genome's baseline circle and inverted LCBs are placed inside the baseline circle. Image generated by GenVision (DNAStar). This figure can be viewed in colour as supplementary data at *Bioinformatics* online.

regions of sequence identity. Second, some of the matching regions found in the first step are filtered based on user-specified criteria. Finally, the remaining regions of sequence identity are organized into LCBs.

Each region of sequence identity located by GRIL is a maximal unique match (MUM) (Delcher *et al.*, 1999) existing in all sequences under consideration. GRIL locates MUMs using a simple seed-and-extend style hashing algorithm. First, GRIL creates a sorted mer list by sorting a list of the positions of all mers by the alphabetical ordering of the mer sequences. The sorted mer lists are used to identify exactly matching seeds of a user-specified minimum size. Matching seeds unique in each sequence are extended until a mismatch is encountered. The resulting set of MUMs may contain small MUMs arising from random sequence matches. The second step of GRIL's algorithm is intended to discard such matches.

GRIL filters matches based on four user-specified parameters: minimum match size, maximum offset difference, length of a collinear region and density of matches in a collinear region. A more detailed description of the effects of each parameter is given on our web site at http://asap.ahabs.wisc.edu/gril/. In the third and final step, GRIL groups the remaining matches into LCBs and reports the boundaries of the LCBs in each sequence.

The number of MUMs found in the first step depends on the particular combination of genomes being compared. For example, GRIL reports 8434 MUMs found in three *Escherichia coli* and two *Salmonella enterica* genomes, over 40 000 MUMs for two *E.coli* and *Shigella flexneri*, but only 2277 in a single *E.coli* versus *Yersinia pestis* pair-wise comparison. MUMs found during step one are written to a file that can be read in so that filtering and LCB determination steps of GRIL can be rerun using different filtering parameters without repeating the string-matching step.

We illustrate using two comparisons of *E.coli* K-12 with two separately sequenced isolates of *S.flexneri* (Fig. 1). The number of LCBs almost doubles (from 21 to 36) when the minimum LCB length parameter is changed from 10 to 5 kb. The additional rearrangements revealed by the lower threshold are a testament to the genome plasticity of *S.flexneri*. Rearrangements between *S.flexneri* and *E.coli* are likely to have been mediated by numerous copies of certain insertion sequences that punctuate *S.flexneri* but are absent in *E.coli* . A detailed evaluation of rearrangements identified by GRIL in multiple genomes and a comparison to previously published pair-wise results is on-line at http://asap.ahabs.wisc.edu/gril/example.html

Note that GRIL can be applied to any type of orthologous landmarks. One must create a file for the landmarks in GRIL format: landmark length, starting coordinates in the reference genome and then starting coordinates in the remaining genomes. The reference genome is chosen to assign a forward orientation to each landmark: landmarks in the reverse complement orientation in the other genomes are denoted by a minus sign, affixed to the corresponding start position. GRIL can then be used to identify all LCBs, and hence rearrangements, defined by these landmarks.

The size of sequences that GRIL can be applied to is dependent on the amount of available memory. GRIL requires approximately 16 bytes of memory per base of the longest input sequence. Furthermore, GRIL places only a modest demand on CPU resources. For example, detecting rearrangements in three bacterial genomes, the largest of which is 5.5 MB, requires approximately 88 MB of memory and about 2 min of time on a 600 MHz Pentium III system.

GRIL is a tool that identifies significant collinear regions among a set of bacterial-size genomes. GRIL solves a problem that has inhibited application of multiple genome aligners to genomes with significant rearrangements. Results from GRIL also provide a useful basis for performing phylogenetic inference using inversion distance.

ACKNOWLEDGEMENTS

We thank Mark Craven and Jeremy Glasner for insightful comments on this manuscript. This work is supported by NIH Grant GM62994-02 and NLM Grant 1T15LM007359-01 Graduate Training Fellowship to A.E.D.

REFERENCES

Ajana,Y., Lefebvre,J.F., Tillier, E. and El-Mabrouk, N. (2002) Exploring the set of all minimal sequences of reversals—an application to test the replication-directed reversal hypothesis. *Second International Workshop, Algorithms in Bioinformatics*.

Bourque,G. and Pevzner,P.A. (2002) ‘Genome-scale evolution: reconstructing gene orders in the ancestral species.’ *Genome Res.*, **12**, 26–36.

Brudno,M., Do,C., Cooper,G., Kim,M.F., Davydov,E., NISC Sequencing Consortium, E.D.G., Arend Sidow and and Batzoglou,S. (2003) ‘LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.’ *Genome Res.*, **13**, 721–731.

Delcher,A.L., Kasif,S., Fleischmann, R.D., Peterson, J., White, O. and Salzberg,S.L. (1999) ‘Alignment of whole genomes.’ *Nucleic Acids Res.*, **27**, 2369–2376.

Hohl,M., Kurtz,S. and Ohlebusch,E. (2002) ‘Efficient multiple genome alignment.’ *Bioinformatics*, **18** (Suppl. 1), S312–S320.

Larget,B., Simon,D. and Kadane,J. (2002) ‘Bayesian phylogenetic inference from animal mitrochrondrial genomes.’ *J. R. Stat. Soc. Ser. B*, **64**, 681–693.

Moret,B.M., Wang,L.S., Warnow, T. and Wyman,S.K. (2001) ‘New approaches for reconstructing phylogenies from gene order data.’ *Bioinformatics*, **17** (Suppl. 1), S165–S173.

Pevzner,P. and Tesler,G. (2003) ‘Genome rearrangements in Mammalian evolution: lessons from human and mouse genomes.’ *Genome Res.*, **13**, 37–45.