BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 6 2000
Pages 560–561

Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes

A. C. Frank and J. R. Lobry*

Laboratoire BBE-CNRS UMR 5558, Université Claude Bernard, 43 Bd. du 11 Novembre 1918, F-69622 Villeurbanne cedex, France

Received on November 26, 1999; revised and accepted on January 21, 2000

Abstract

Summary: *A program called Oriloc has been developed for the prediction of bacterial replication origins. The method builds on the fact that there are compositional asymmetries between the leading and the lagging strand for replication. The program works with unannotated sequences in fasta format and therefore uses glimmer 2.0 outputs to discriminate between codon positions so as to increase the signal/noise ratio.*

Availability: *The ANSI C source code is freely available for academic use at ftp://pbil.univ-lyon1.fr/pub/logiciel/oriloc/oriloc.c.*

Contact: *lobry@biomserv.univ-lyon1.fr*

Supplementary information: *ftp://pbil.univ-lyon1.fr/pub/logiciel/oriloc/oriloc.ps*

Advances in understanding the control of DNA replication in bacterial chromosomes require that we are able to locate initiation sites at the sequence level in replicating genomic DNA. Asymmetry in base composition between the leading and the lagging strands provides us with a method for the prediction of an *a priori* origin of replication, which can then be confirmed experimentally. In the absence of bias between the two DNA strands for mutation and selection, the base composition within each strand should be such that A = T and C = G (Lobry and Lobry, 1999), but in bacterial genomes there are local and systematic deviations, changing sign at the origin and terminus of replication (Lobry, 1996; Frank and Lobry, 1999).

The program analyses simultaneously the deviations from A = T and C = G by performing a DNA walk (e.g. Cebrat and Dudek, 1998) with the following axis assignment:

$$x - \text{axis}\{A\} = (-1, 0) \quad \text{and} \quad \{T\} = (1, 0)$$
$$y - \text{axis}\{C\} = (0, 1) \quad \text{and} \quad \{G\} = (0, -1).$$

The x-axis and y-axis coordinates, corresponding to the cumulated AT and GC skew, respectively, are an output of the program. The 2D DNA walk is then reduced to a single skew index by projection onto the orthogonal regression line through the origin so as to obtain a simple index versus map position representation (Figure 1). This index is similar to the 'keto excess' and 'purine excess' indices (Freeman *et al.*, 1998) except that here the synthetic skew index is not predefined but optimized for each genome to maximize its total variance.

In bacteria the leading strands for replication are enriched in keto (G or T) bases while the lagging strand is enriched amino bases (A or C) (Perrière *et al.*, 1996; Rocha *et al.*, 1998) so that the sign of the skew index was

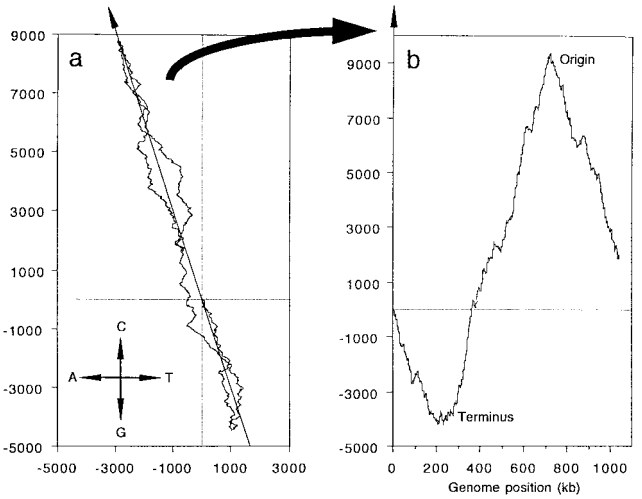

Fig. 1. The complete genome sequence of *Chlamydia trachomatis* (AE001273) was used to illustrate the method used by oriloc. a, A DNA walk is performed by reading the sequence in the third codon positions predicted by glimmer and walking into the plane according to the four directions defined by the four bases as indicated on the bottom left of the figure. The resulting DNA walk is then summarized by projection onto the orthogonal regression line pointing out at about 11 o'clock in the figure. b, The projected values are used as a composite skew index plotted versus map position on the chromosome. The origin is predicted at the maximum skew value while the terminus is predicted at the minimum.

*To whom correspondence should be addressed.

set so that its maximum value correspond to the origin of replication and its minimum to the terminus of replication.

The DNA walk is performed on a coding sequence by coding sequence basis for three reasons: (i) in all known cases the origin of replication is an intergenic region so that origin of replication should be between two steps in the walk, (ii) to make the method more sensible, only third codon positions are regarded, where deviations are more pronounced because of the reduced selective pressure, and (iii) this removes the choice of a window size parameter.

In order to make the program easy to use on newly sequenced genomes that are unannotated, Oriloc uses the genome sequence in fasta format and glimmer 2.0 (Delcher *et al.*, 1999) outputs (the file g2.coord produced by the script run-glimmer 2). Glimmer 2.0 produces without human intervention a list of putative coding sequences with a high sensitivity (98.6–99.8%) allowing to discriminate between codon positions. No appreciable differences were found when using genome annotations instead of glimmer outputs, or third codon positions that are completely synonymous (e.g. CCN for Pro) instead of all third codon positions.

Once the origin and terminus position have been assigned by Oriloc, the program computes the percentage of coding sequences on the leading strand. This is a control: if this percentage is less than 50% the prediction of the location of the origin is probably incorrect because in bacteria there is a selective pressure increasing the percentage of coding sequences on the leading strand (Brewer, 1988).

The program was tested on genomes for which the origin has been experimentally determined. The program succeeded to locate the origin of *Borrelia burgdorferi* and *Bacillus subtilis* and located the *Escherichia coli* origin 7 kb from the true origin with a secondary peak at *oriC*, a point that was also reported by Salzberg *et al.* (1998). This outlines the importance of a human inspection of skew diagrams for the prediction of putative replication origins. More illustrations of Oriloc outputs can be found at the supplementary information site.

Oriloc was primarily developed for the detection of replication boundaries in bacteria and assumes that the chromosome replicates bi-directionaly from a single origin and evolves with an asymmetric substitution pattern between the two DNA strands. Oriloc is therefore not suited for the prediction of replication boundaries when at least one of these conditions are not fulfilled, or when the bias between the two strands is too weak with respect to the amount of available data to generate a clear signal.

References

Brewer,B.J. (1988) When polymerases collide: replication and the transcriptional organization of the *E.coli* chromosome. *Cell*, **53**, 679–686.

Cebrat,S. and Dudek,M.R. (1998) The effect of DNA phase structure on DNA walks. *Eur. Phys. J. B.*, **3**, 271–276.

Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Freeman,J.M., Plasterer,T.N., Smith,T.F. and Mohr,S.C. (1998) Patterns of genome organization in bacteria. *Science*, **279**, 1827a.

Frank,A.C. and Lobry,J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, **238**, 65–77.

Lobry,J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.

Lobry,J.R. and Lobry,C. (1999) Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol. Biol. Evol.*, **16**, 719–723.

Perrière,G., Lobry,J.R. and Thioulouse,J. (1996) Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. *Comput. Applic. Biosci.*, **12**, 519–524.

Rocha,E.P.C., Viari,A. and Danchin,A. (1998) Universal replication biases in bacteria. *Mol. Microbiol.*, **32**, 11–16.

Salzberg,S.L., Salzberg,A.J., Kerlavage,A.R. and Tomb,J.-F. (1998) Skewed oligomers and origins of replication. *Gene*, **217**, 57.