Biochimie (1996) 78, 323–326

A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria

JR Lobry

Laboratoire BGBP, CNRS UMR 5558, Université Claude-Bernard, 43, bd du 11-Novembre-1918, F-69622 Villeurbanne cedex, France

(Received 11 April 1996; accepted 2 May 1996)

Summary — A simple adaptation of Ninio's vectorial representation of DNA sequences for the detection of replication origins in bacteria is presented. The origins of replication in *Escherichia coli*, *Bacillus subtilis*, *Haemophilus influenzae* and *Mycoplasma genitalium* are well outlined with this graphical representation.

replication origin / DNA sequence / vectorial representation

Introduction

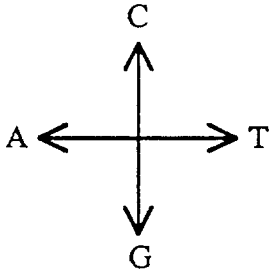
Vectorial representations of sequences were first introduced by Mizraji and Ninio [1, 2]. I present here an adaptation that allows for a detection of replication origins in bacteria.

The principle for the detection of replication origins is that under no-strand-bias conditions, the equilibrium point is such that the base frequencies in each strand always respect the [A] = [T] and [C] = [G] equalities, regardless of the initial state of the DNA sequence and of details of the substitution patterns [3, 4]. Since the mechanisms for DNA replication differ between the leading strand and the lagging strand, at least *in vitro*, mutation patterns could differ depending on which strand is being copied, yielding an asymmetry in [A] = [T] or [C] = [G] equi-frequencies. This asymmetry is expected to switch polarity at the origin of replication of the chromosome, as observed in *Escherichia coli*, *Bacillus subtilis* and *Haemophilus influenzae* [5]. The putative origin of replication suggested [6] for *Mycoplasma genitalium* has also been confirmed [7] by this method.

Materials and methods

The data set was composed of four contiguous DNA sequences from four bacterial species. *E coli*, 1 616 174 base sequences available in the EMBL/DDBJ/GenBank database as nine segments (U18997, U00039, L10328, M87049, L19201, U00006, U14003, D10483, D26562), and as a single contiguous sequence (contig) at the Universal Ressource Locator (URL) (ftp://ecoliftp.genetics.wisc.edu/pub/sequence). *B subtilis*, 193 394 base sequences available as five segments (D26185, D13303, D50303, L24376, L43593), and as a single contiguous sequence at the URL (http://ddbjs4h.genes.nig.ac.jp/cgi-bin/acnuc-search-ac?query = BS0310). *H influenzae*, 1 830 137 base sequence availables as 163 segments (L42023), and as a single contiguous sequence at the URL (ftp://ftp.tigr.org/pub/h_influenzae). *M genitalium;* 580 073 base sequences available as 56 fragments (L43967), and as a single contiguous sequence at the URL: (ftp://ftp.tigr.org/pub/m_genitalium).

The principle for the representation is that each of the four bases is represented with a vector as follows:
so that the DNA sequence is represented by a trajectory in the plane.



This is similar to the phase-plane representation for the system of two differential equations: the state variables are plotted one *versus* the other so that time is neutralized. Here, thanks to the opposite orientation of the A-vector *versus* the T-vector and of the C-vector *versus* the G-vector it are the symmetries [A] = [T] and [C] = [G] which are neutralized: when a sequence respects these equalities the corresponding trajectory is stationary because the resulting vector is the null vector. The trajectory is then controlled only by the deviation from these equalities, so that even a small bias can be emphasized. Moreover, since both deviations are represented simultaneously one can easilly appreciate simultaneous variations.

Relevant DNA sequences are usually too long to represent all elementary vectors by a graphical vector. The maximum number of graphical vectors that can be represented is hardware-dependent (the memory available in the printer) and social-dependent (the time colleagues wait for your graph to be printed before shuting down the printer). Different ways of compressing numerous elementary vectors into a sensible number of graphical vectors exist. I used the simple solution that each graphical vector represents the sum vector of N elementary vectors. The advantage of this choice is that the

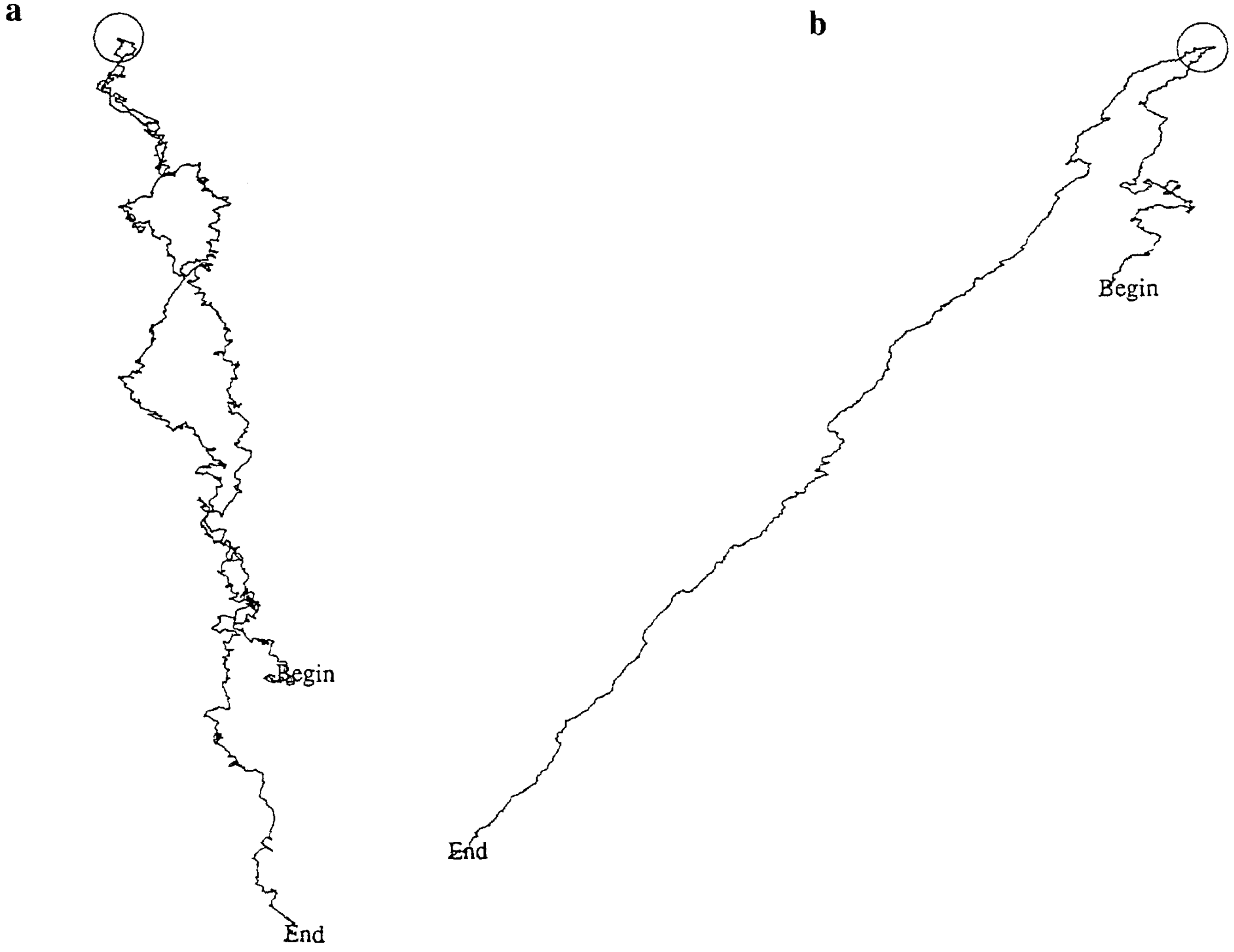


Fig 1. Vectorial representation of DNA sequences from four bacterial species. **a.** *Escherichia coli.* **b.** *Bacillus subtilis.* **c.** *Haemophilus influenzae.* **d.** *Mycoplasma genitalium.* The position of the origin of replication is outlined by a circle.

general shape of trajectories is conserved for different scalings. The values for N in bp were 300, 40, 400 and 100 for *E coli*, *B subtilis*, *H influenzae* and *M genitalium*, respectively.

Results and discussion

The trajectories for four bacterial species (fig 1) show that replication origins are easily detected with this representation: they are always close to the reverse turn of the trajectory. However, for circular genomes there are in fact two reverse turns, the second corresponding to the terminus of replication (*eg H influenzae*). What is predicted with this method is the location of the origin and of the terminus of replication, but it does not predict which one is which. Its main interest is that it yields quickly the regions of genomes we have to focus on with other methods to find the origin and terminus of replication.

Acknowledgments

Thanks are due to Jacques Ninio. This work was supported by 'Groupement de Recherches et d'Etudes sur les Génomes'.

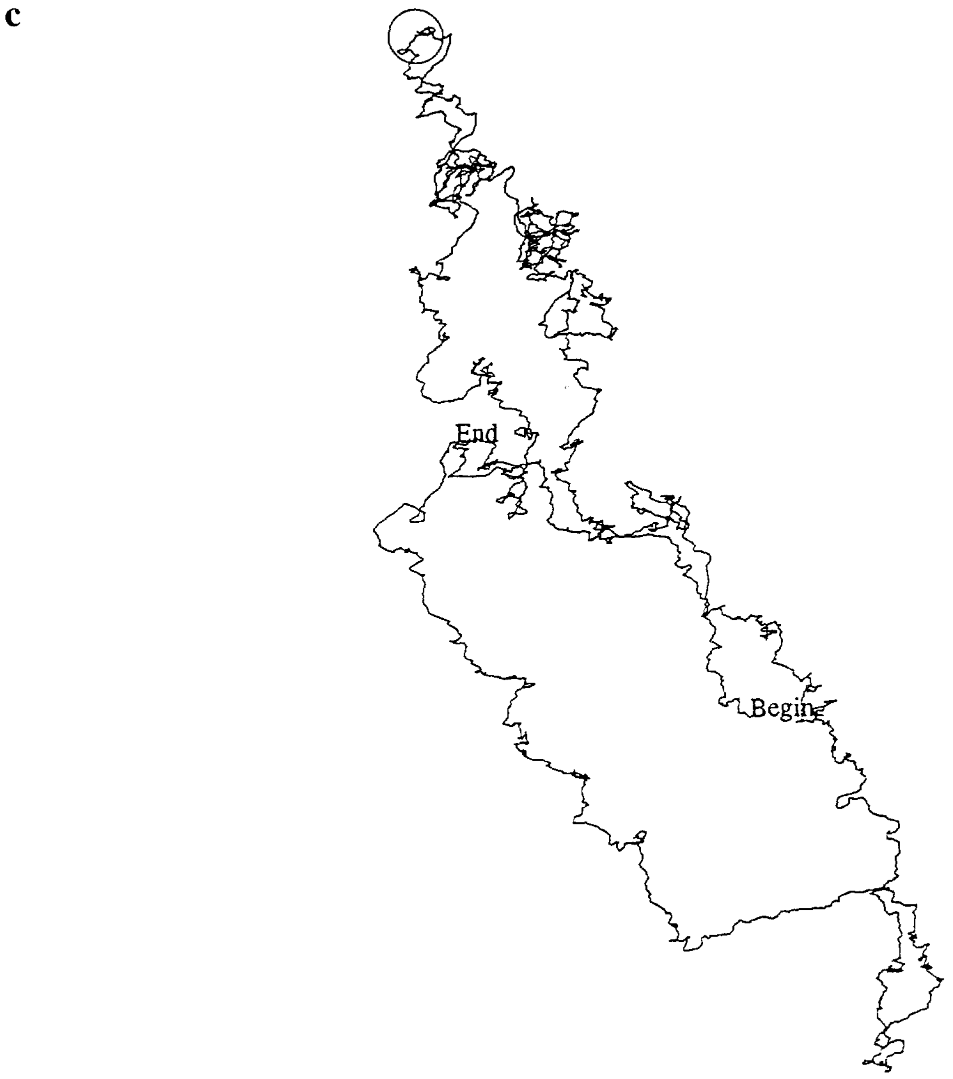


Fig 1. Continued.

References

1 Mizraji E, Ninio J (1985) Graphical coding of nucleic acid sequences. *Biochimie* 67, 445–448

2 Ninio J, Mizraji E (1995) Perceptible features in graphical representations of nucleic acid sequences. In: *Visualizing Biological Information* (Pickover CA, ed) World Scientific, Singapore, 238 p

3 Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol* 40, 318–325

4 Lobry JR (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol* 40, 326–330

5 Lobry JR (1996) Asymmetric susbtitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* 13, 660–665

6 Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman JL, Weidman JF, Small KV, Sandusky M, Fuhrmann JL, Nguyen DT, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA III, Venter JC (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397–403

7 Lobry JR (1996) Origin of replication of *Mycoplasma genitalium*. *Science* 272, 745–746

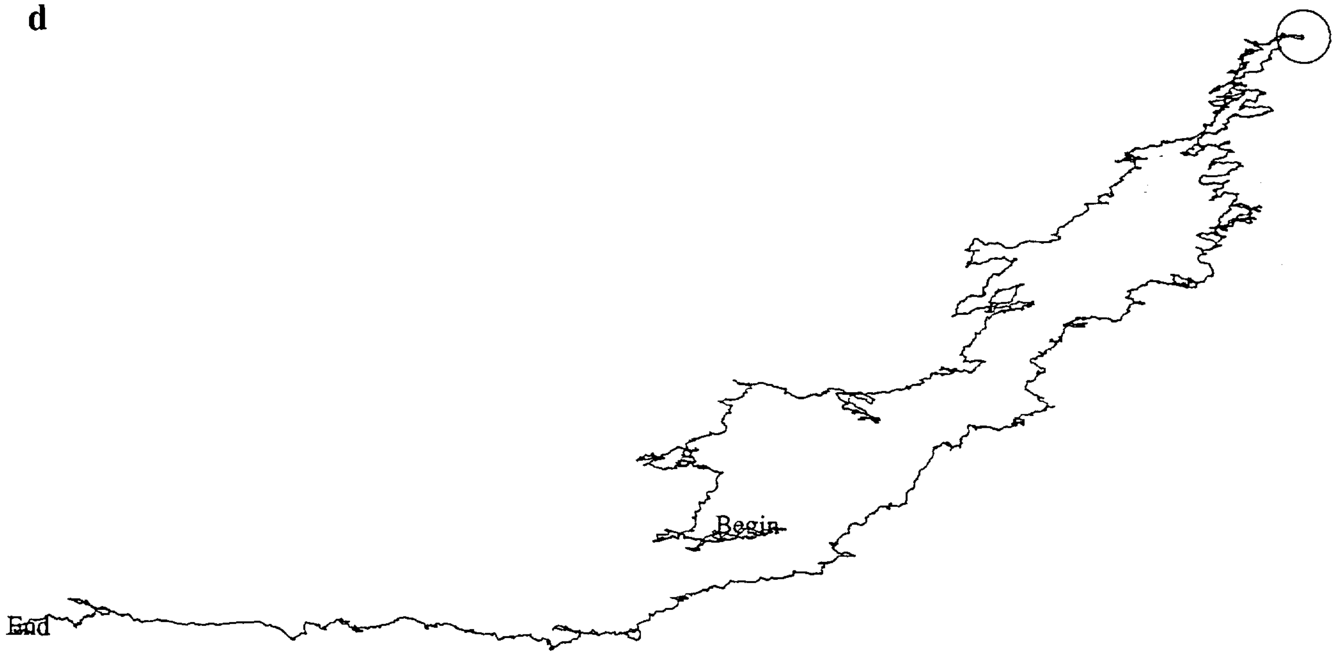


Fig 1. Continued.