BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 14 2007, pages 1866–1867
doi:10.1093/bioinformatics/btm255

Databases and ontologies

DoriC: a database of *oriC* regions in bacterial genomes

Feng Gao and Chun-Ting Zhang*
Department of Physics, Tianjin University, Tianjin 300072, China

Received on April 5, 2007; revised on April 29, 2007; accepted on May 4, 2007
Advance Access publication May 12, 2007
Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Replication origins (*oriC*s) of bacterial genomes currently available in GenBank have been predicted by using a systematic method comprising the Z-curve analysis for nucleotide distribution asymmetry, DnaA box distribution, genes adjacent to candidate *oriC*s and phylogenetic relationships. These *oriC*s are organized into a MySQL database, DoriC, which provides extensive information and graphical views of the *oriC* regions. In addition, users can Blast a query sequence or even a whole genome against DoriC to find a homologous one. DoriC will be updated timely and the latest version is DoriC 1.8, in which *oriC*s of 425 genomes (468 chromosomes) are identified.

Availability: DoriC can be accessed from http://tubic.tju.edu.cn/doric/

Contact: ctzhang@tju.edu.cn

Supplementary information: Supplementary data are available at http://tubic.tju.edu.cn/doric/supplementary.htm

1 INTRODUCTION

The initiation of replication is the central event in the bacterial cell cycle. However, *oriC* regions remain unknown in many bacterial genomes sequenced so far. Experimental methods for identifying *oriC*s *in vivo* are reliable, but time-consuming and labor-intensive. The *in silico* methods to identify *oriC*s include the GC-skew analysis (Grigoriev, 1998; Lobry, 1996) and the oligomer-skew method (Salzberg *et al*., 1998; Worning *et al*., 2006), etc. Sequence analysis revealed that an *oriC* region usually contains multiple 9mer consensus elements termed the DnaA box. Jointly using the three methods (GC-skew, location of the *dnaA* gene and distribution of DnaA boxes) resulted in better prediction of *oriC* regions (Mackiewicz *et al*., 2004).

The *Z*-curve method is an alternative technique that detects the asymmetrical nucleotide distribution around *oriC*s. Using the *Z*-curve method, three *oriC*s were predicted in the genome of the archaeon *Sulfolobus solfataricus*, e.g. see a review of (Zhang and Zhang, 2005), and the prediction is consistent with recent experimental data, e.g. see a review of (Robinson and Bell, 2005). To extensively identify *oriC*s with high accuracy and reliability, an integrated *in silico* method to predict *oriC* regions of bacterial genomes has been developed, based on the *Z*-curve method, the distributions of DnaA boxes, the indicator genes such as *dnaA* (*dnaN*, *hemE*, *gidA*... or *repC*) and phylogenetic relationships. The present work mainly consists of two parts: identifying *oriC* regions and setting up the database DoriC.

2 METHODS AND RESULTS

2.1 The procedure to identify *oriC* regions

The procedure to identify *oriC* regions (refer to Supplementary Fig. 1A) is described as follows.

(1) Extract all intergenic sequences according to the annotation files. Complete bacterial genomes and the related annotation files were downloaded from the NCBI ftp server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). In general, the number of DnaA boxes differing by no more than one position from *Escherichia coli* perfect DnaA box (TTATCCACA) was counted for each intergenic sequence. However, it should be noted that some 'species-specific' DnaA boxes were adopted for certain bacteria. In this case, the number of 'species-specific' DnaA boxes proposed by us or previously experimentally identified was counted. For example, the DnaA box motif TTTTCCACA is universal for the genomes in the phylum *Cyanobacteria*.

(2) Assign every intergenic sequence an *oriC* type (types 1–5), according to the number of DnaA boxes within it and the location related to the *dnaA* gene or the minimum of the GC disparity curve. The definition and characteristics of different *oriC* types have been summarized in Supplementary Table 1.

(3) Once the *oriC* type and related information for every intergenic region is obtained, select one or two intergenic sequences with the highest *oriC* type priority (type 1 > type 2 >···> type 5) as candidate *oriC* regions.

(4) Finally, output the location, AT content, length, DnaA box number, type and the sequence of the identified *oriC* regions.

2.2 The procedure to set up DoriC

The procedure to set up DoriC (refer to Supplementary Fig. 1B) is described briefly as follows.

(1) Extract genome information, such as organism's name, lineage, topology of chromosome, *dnaA* gene location.

(2) Calculate the genome size, GC content, coordinates of GC, AT, RY, MK disparity curves (Zhang and Zhang, 2005) and the precise coordinates of extremes of the GC disparity curve, search for DnaA boxes and *dif*-like sequences on both strands and identify *oriC* regions according to the procedure described above.

*To whom correspondence should be addressed.

(3) Output the information of genome and *oriC* region(s), and integrated plots for the original and rotated sequences to display the obtained results, such as general genome information, four disparity curves, distribution of DnaA boxes, locations of *dnaA* genes, *dif* sites and *oriC* regions. In each rotated sequence, the sequence coordinate origin begins and ends in the *dif* site or the maximum of the GC disparity curve.

(4) Organize the output information and integrated plots by using an open-source database management system, MySQL.

The identified *oriC* regions are highly reliable as reflected by the followings. Among these *oriC* regions, in 345 (72.2%) *oriC*s, all the three methods, i.e. typical base composition asymmetry, DnaA box distribution and indicator gene positions, are highly consistent with each other; in 123 (25.7%) *oriC*s, at least one method of the above three or the *dif* site position can clearly indicate the *oriC* location. Accordingly, the accuracy of predicted *oriC*s is estimated to be between 72.2 and 97.9% (72.2% + 25.7%). In addition, comparative analysis showed that the identified *oriC*s, the proposed DnaA box motif and gene arrangement around the *oriC*s are highly conserved among the bacteria within a family. Furthermore, all experimentally confirmed *oriC*s known so far are consistent with those identified in this work (refer to Supplementary Table 2). For more details, please refer to the Supplementary Material 'Analysis for the accuracy of DoriC'.

3 DATABASE CONTENT AND WEB INTERFACE

DoriC is built using a relational database (MySQL) allowing rapid retrieval of data and making resource easily maintainable. In general, one entry corresponds to one genome (chromosome). However, for some genomes (chromosomes) the *oriC* region is split into two distinct sub-regions by the *dnaA* gene, resulting in two entries for one genome (chromosome). The database access is via a web interface based on PHP script and provides various ways to search for DoriC entries, such as organism's name, accession number, lineage, *oriC* type and a keyword, etc. DoriC can be arranged in the order of organism's name, accession number, genomic GC content and *oriC* type. In addition, users can also Blast a query sequence or even a whole genome against DoriC to find a homologous one. DoriC will be updated timely and the latest version is Doric 1.8, in which *oriC*s of 425 genomes (468 chromosomes) are identified.

ACKNOWLEDGEMENTS

We would like to thank Dr Ren Zhang for invaluable assistance. Technical supports from Dr Hong-Yu Ou and Yan Lin are gratefully acknowledged. The present work was supported in part by NNSF of China (Grant No. 90408028).

Conflict of Interest: none delcared.

REFERENCES

Grigoriev,A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.*, **26**, 2286–2290.

Lobry,J.R. (1996) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, **78**, 323–326.

Mackiewicz,P. *et al.* (2004) Where does bacterial replication start? Rules for predicting the *oriC* region. *Nucleic Acids Res.*, **32**, 3781–3791.

Robinson,N.P. and Bell,S.D. (2005) Origins of DNA replication in the three domains of life. *FEBS J.*, **272**, 3757–3766.

Salzberg,S.L. *et al.* (1998) Skewed oligomers and origins of replication. *Gene*, **217**, 57–67.

Worning,P. *et al.* (2006) Origin of replication in circular prokaryotic chromosomes. *Environ. Microbiol.*, **8**, 353–361.

Zhang,R. and Zhang,C.T. (2005) Identification of replication origins in archaeal genomes based on the *Z*-curve method. *Archaea*, **1**, 335–346.