DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes

Feng Gao*, Hao Luo and Chun-Ting Zhang*

Department of Physics, Tianjin University, Tianjin 300072, China

Received August 14, 2012; Revised September 27, 2012; Accepted September 29, 2012

ABSTRACT

Replication of chromosomes is one of the central events in the cell cycle. Chromosome replication begins at specific sites, called origins of replication (oriCs), for all three domains of life. However, the origins of replication still remain unknown in a considerable large number of bacterial and archaeal genomes completely sequenced so far. The availability of increasing complete bacterial and archaeal genomes has created challenges and opportunities for identification of their oriCs in silico, as well as in vivo. Based on the Z-curve theory, we have developed a web-based system Ori-Finder to predict oriCs in bacterial genomes with high accuracy and reliability by taking advantage of comparative genomics, and the predicted oriC regions have been organized into an online database DoriC, which is publicly available at http://tubic.tju.edu.cn/doric/ since 2007. Five years after we constructed DoriC, the database has significant advances over the number of bacterial genomes, increasing about 4-fold. Additionally, oriC regions in archaeal genomes identified by in vivo experiments, as well as in silico analyses, have also been added to the database. Consequently, the latest release of DoriC contains oriCs for >1500 bacterial genomes and 81 archaeal genomes, respectively.

INTRODUCTION

The identification of replication origins will be helpful to reveal the regulatory mechanisms of the initiation step in DNA replication (1,2) and discover new broad-spectrum antibacterial drugs (3). Based on the Z-curve theory (4), we have developed a web-based system Ori-Finder for finding oriCs in bacterial genomes with high accuracy and reliability (5), and the predicted oriC regions in bacterial genomes have been organized into an online database DoriC (6). Based on the database, putative origins of replication in Sorangium cellulans, Microcystis aeruginosa (7) and Cyanothece 51142 (8), which could not be determined by using standard GC skew, have been identified by taking advantage of comparative genomics. The application of the predicted oriC selection criteria and the comparison of different cyanobacterial strains may also gain insight into the replication origins in other cyanobacteria (9). As the database was constructed in 2007, we noticed that the replication origins of Anabaena sp. PCC 7120 (10), Cytophaga hutchinsonii ATCC 33406 (11) and Synechococcus elongatus PCC 7942 (12) have been confirmed by experiments, which are all consistent with our predictions in DoriC. Because of continuous updates, our database has been widely used in the comparative genomics analysis. For example, as a source of data, DoriC has been used in the study of the relationship between the functionality of essential genes and gene strand bias in bacterial genomes (13), in the analysis of nucleotide compositional asymmetry between the leading and lagging strands of bacterial genomes (14), in the investigation of the association between growth-related traits and minimal generation times (15), in an algorithm for prediction of putative essential and core-essential genes in Mycoplasma genomes (16), in the research on coordination of spatiotemporal gene expression during the bacterial growth cycle (17) and in the study of the variation in terms of the percentage of leading strand genes across different bacteria (18), etc. It is expected that the new release of the database, DoriC 5.0, will promote the study of oriCs in both bacteria and archaea.

DATABASE UPDATES

In the current release, the database has been significantly improved compared with the initial release, and the main advances include (i) inclusion of oriCs in more bacterial genomes that increased from 435 to 1528; (ii) inclusion of oriCs in 81 archaeal genomes; (iii) inclusion of detailed information about repeats in oriCs identified by REPuter program (19); and (iv) addition of URLs that...
DATABASE DESCRIPTION

Replication origins in bacteria

To identify oriC regions of unannotated bacterial genomes, we have developed a web-based system, Ori-Finder, based on an integrated method comprising gene identification, analysis of base composition asymmetry using the Z-curve method, distribution of DnaA boxes, occurrence of genes frequently close to oriCs and phylogenetic relationships. Consequently, the predicted oriC regions have been organized into an online database, DoriC. Based on DoriC, the relationships between the conserved features associated with the oriC regions, such as adjacent genes, DnaA boxes, etc., and the taxonomic levels of the corresponding bacteria have been summarized. For example, detailed analyses have shown that the consensus sequence of the DnaA boxes in oriC regions and the distribution of genes around oriCs are strongly conserved among the bacteria in the phylum cyanobacteria (7,8). The feature that the oriC is adjacent to dnaN gene, which encodes the beta clamp processivity factor, has been found to be universal among the bacteria within the phylum cyanobacteria, and the ‘species-specific’ DnaA box motif for the phylum cyanobacteria is ‘TTTTC CACA’ instead of ‘TTATTCACA’, the DnaA box motif of Escherichia coli. These strongly conserved features indicate that the in silico identified oriCs are reliable, as they have been confirmed by comparative genomics approaches. This observation also shows that if the oriC for one of the bacteria in the phylum cyanobacteria is confirmed experimentally, the oriCs for the other bacterial genomes in this phylum may be confirmed simultaneously.

As we expected, the experimentally confirmed replication origins of Anabaena sp. PCC 7120 (10) and S. elongatus PCC 7942 (12) in the phylum cyanobacteria are all adjacent to the dnaN gene, which encodes the beta clamp processivity factor. Therefore, the proposed rules may be helpful to predict the oriC regions for some bacteria without complete genomes in the phylum cyanobacteria. In addition, the application of the proposed rules derived from DoriC would speedup the experimental confirmation and functional analysis of oriCs in bacterial genomes. Because of the rapid growth in the number of sequenced bacterial genomes, the replication origins for those unsubmitted to GenBank or not deposited in DoriC temporarily can be predicted by Ori-Finder firstly, which now has been used to analyze ∼30 newly sequenced bacterial genomes.

Replication origins in archaea

The Z-curve analysis has been used to identify one replication origin in the genomes of Methanocaldococcus jannaschii (22) and Methanosarcina mazei (23), two replication origins in the Halobacterium species NRC-1 genome (24), which have been confirmed by in vivo experiments (25,26) and three replication origins in the Sulfolobus solfataricus P2 genome (24), which have been later confirmed experimentally (27,28). Here, we collected the information of oriCs provided in the literature, such as the oriC sequences, origin recognition boxes (ORB) motifs, uncharacterized motif sequences, etc., which were identified by in vivo experiments (25–34), as well as in silico analysis (4,22–24,35). In addition, we also predicted some new replication origins by Z-curve method, with the aid of homologous sequence search against the known replication origins, analysis of ORB motifs and repeats, cde6 gene location, etc. Consequently, oriC regions in 81 archaeal genomes identified by in vivo experiments, as well as in silico analyses, have been added to our database. The number of oriCs in archaea is correlated with the phylogeny, which has been summarized in detail in the ‘Introduction’ section of the (34). Based on our results in DoriC, it shows that there is one replication origin in the genomes within the order Methanococcales (11 genomes) and within the class Thermococci (12 genomes), and three replication origins in Sulfolobus species (13 genomes). Our results and the Z-curves also show that the archaea within the Crenarchaeota phylum contain multiple origins, although some origins could not be determined at the sequence level currently. For example, Pyrobaculum calidifontis has been experimentally characterized to contain four replication origins, which is the highest number detected in a prokaryotic organism (34). However, only one origin can be determined at the sequence level (34). During the course of the prediction, we found that the location of some putative replication initiator gene besides cde6 gene can be helpful to the oriC prediction in some cases. For example, in the genome of M. jannaschii, an ORF (MJ0774), annotated as a hypothetical protein, is a distant homolog of the Cdc6 protein in fact (22). The name Mc-pPRIP for the putative replication initiator protein in Methanococcales has been used here for MJ0774 and related proteins to distinguish it from bona fide orthologous Cdc6. We also found the genes, which encode Mc-pPRIP in other 10 genomes within the order Methanococcales (Methanococcus aeolicus Nankai-3, Methanocaldococcus fervens AG86, Methanococcus maripaludis C5, M. maripaludis C6, M. maripaludis C7, M. maripaludis S2, M. maripaludis X1, Methanococcus vannielii SB, Methanococcus voltae A3 and Methanocaldococcus vulcanius M7), were annotated as ‘LysR family protein’, ‘regulatory protein ArsR’, ‘MarR family transcriptional regulator’, etc. Based on the locations of these genes, the oriCs in the aforementioned genomes were predicted reliably, which contains almost all the features of known replication origins in archaeal genomes. URLs that link to NCBI Map Viewer or UCSC Archaeal Genome Browser (if available) are also provided, which will be useful to explore and discover the conserved features around the oriC region. With the availability of an increasing number of archaeal genomes, the prediction will be more accurate and reliable, as the ORB elements or genes frequently close to oriCs can also be
analyzed by comparative genomics, and new rules for replication origins in archaeal genomes will also be extracted in the future with the continuous update of DorIC. Here, motif-based sequence analysis tools, the multiple EM for motif elicitation (MEME) Suite (36), have been used to discover motifs in the replication origins of closely related species, e.g. the archaea from the order Thermococcales. Consequently, ORB motifs and some new uncharacterized motif sequences have been found by the MEME Suite and are also included in the database.

CONCLUSION

With the increased availability of completely sequenced bacterial and archaeal genomes and experimental evidence, the database will become more useful because of including more information. The application of the rules from the database will be helpful to develop new prediction algorithms of replication origins and speedup the experimental confirmation and functional analysis of oriCs in bacterial or archaeal genomes. Systematic and functional analysis of oriC regions in bacteria and archaeal genomes will also be useful for the construction of the minimum genome and regulation of growth rate and generation time of bacteria and archaea, which play a key role in the emerging field of synthetic biology. DorIC will be updated periodically to include more entries, and to integrate more information for each entry. We also welcome any feedback or corrections to help us improve the database.

ACKNOWLEDGEMENTS

The authors thank Dr Kurtz for providing the REPuter binaries. They also thank Dr Ren Zhang for critical revision of manuscript. Helpful discussions with Yan Lin and Chong Peng are also gratefully acknowledged.

FUNDING

The National Natural Science Foundation of China [31171238, 30800642 and 10747150]. Funding for open access charge: The National Natural Science Foundation of China [31171238].

Conflict of interest statement. None declared.

REFERENCES


