Journal of
Computer Science & Systems Biology

Ojha and Swati, J Comput Sci Syst Biol 2015, 8:2
DOI: 10.4172/jcsb.1000174

Research Article Open Access

ArchaeProfile: A Database of Archaea and their Origins of Replication

Krishna Kumar Ojha[1]* and Swati D[2]

[1]Department of Bioinformatics, Mahila MahaVidhyalaya Women College, Banaras Hindu University Varanasi-221005, India
[2]Department of Physics and Bioinformatics, Mahila MahaVidhyalaya Women College, Banaras Hindu University Varanasi-221005, India

Abstract

Archaea are single cell microorganisms having several unique characteristics which differentiate them from bacteria. One of the key features which make archaea distinct from bacteria is the replication process, which is very different and resembles that of the eukaryotes. *In-vivo* mapping of *the ori site* in Archae is a time consuming and tedious job due to complexity involved in the culture of archaeal colony, which puts challenges as well as opportunity to scientist to devise *in-silico* method to map the *Ori* site in archaeal genomes. Z-curve approach is a widely used *in-silico* method to predict the *Ori* site in archaea, but it is not equally successful for all archaeal genomes. Several other parameters like copy number and location of the cdc6 gene, AT rich region with the presence of origin recognition boxes (ORB) provide a better estimate of the *Ori* site in archaea. The motivation behind development of Archae Profile database is to predict the location and the number of putative *Ori* sites in archaeal genomes based on purine-pyrimidine(R-Y) and amino-keto(M-K) disparity curve along with the consensus ORB sequences, cdc6 gene copy number, their location and upstream AT richness. Quick update cycle and easy browser interface makes Archae Profile distinct from other databases. Another important feature is the integration of tools for plotting disparity plot of a given genome sequence and finding specific repeats with copy number and location in a sequence. ArchaeProfile will be updated timely and the emphasis will be to integrate other tools for genome analysis as well as new search features in the database. Presently Archaea Profile has *Ori* related data of 122 archeal genome which is likely to increase with time.

Availability: ArchaeProfile can be accessed freely from http://www.bioinfommv.in/archaeprofile. Data available on the database could be used for further analysis and tutorial purpose.

Keywords: Archaea; Disparity curve; GC skew; *Ori* site; cdc6; ORB

Abbreviation: *Ori* site: Origin of Replication Site; RY: disparity [Purine, Pyrimidine Disparity Curve]; MK: Disparity [Amino-Keto Disparity Curve]; BLOB: Binary Large Object; AJAX: Asynchronous JavaScript and XML; cdc6: Cell Division Cycle 6 Protein; ORB: Origin Recognition Box

Introduction

New trends in next generation sequencing technology have led to a proliferation of sequenced genomes and hence put new challenges before the scientific community to store and analyse the new genomic data. Complete genome sequencing initiatives are biased toward prokaryotes because of their smaller genome size, short life cycle and ease of culture. In prokaryotes itself the number of total sequenced genome projects are skewed towards the bacterial genomes rather than archaeal ones due to complexity associated with the culture and specific growth requirement of the archaea. The same bias appears in the *in-silico* analysis of data available for archaea on the web. There are very few databases available which contain basic information related to archaeal genomes in a concise way.

Archaea are a unique domain of living organism which show the biochemical, structural and morphological similarity with bacteria but have detectable similarity with eukaryotes in some other aspects like the replication process [1,2]. Archaea were classified along with bacteria without any discrimination till 1977 when Carl Woese classified them on the basis of a phylogenetic tree based on 16s-rRNA as a separate group of prokaryotes [3,4]. Most of the archaea studied till now are extremophilic in nature [5]. Nevertheless there are a large number of representatives that survive in moderate environmental conditions [6]. The first sequenced archaeal genome was a methanogen, *Methanocaldococcus jannaschii* in 1996 [7]. Archaea are hard to culture which impacts their genome sequencing, and there are only 161 archaeal genome sequence available on Gene bank as compared to 14,914 total prokaryotic species till March 2013 (ftp://ftp.ncbi.nih.gov/refseq/release/release-statistics/microbial.acc_taxid_growth.txt). The same pattern is observed in the availability of archaeal data on web. There are several dedicated databases and repositories which are available for the exploration of bacterial genomic record but only a few are available for the exploration of archaeal genomes [8,9].

Archae Profile is an effort to provide public access to information related to the *Ori* site in archaea in addition to some other information like genome length, GC%, total genes and proteins predicted to date, optimum growth temperature, pseudo genes and structural RNA for all available archaeal species. The main focus of the database is to predict the *Ori* site using an *in-silico* approach. We have used the Z-curve theory to predict *Ori* site in archaeal genomes. As a matter of fact the Z-curve method alone is not able to predict the *Ori* site in all archaeal genomes; several other *Ori* site sequence related features like presence of ORB, cdc6 gene and AT richness were also considered for prediction of *Ori* site in archaeal genomes.

In ArchaeProfile database we have given R-Y and M-K disparity curves of 122 sequenced archaeal genomes available on NCBI.

***Corresponding author:** Krishna Kumar Ojha, Department of Bioinformatics, Mahila Maha Vidhyalaya, Women College, Banaras Hindu University Varanasi -221005, India, Tel: 9795802746; E-mail: krisiids@gmail.com

Received January 15, 2015; **Accepted** January 28, 2015; **Published** February 02, 2015

Citation: Ojha KK, Swati D (2015) Archae Profile: A Database of Archaea and their Origins of Replication. J Comput Sci Syst Biol 8: 096-098. doi:10.4172/jcsb.1000174

Furthermore a tool has also been incorporated within the database which can draw the nucleotide skew plot of the nucleotide sequence supplied by users. Another tool which is helpful in finding specific repeats and their position in nucleotide sequence, is also given in the database, this can be useful in detecting short and long repeats in the *Ori* site in archaeal genomes which are very common [10,11]. Genes associated with the origin of replication in archaea like cdc6 with the full description of the location and the copy number of the above genes is also available on the database, as the association of cdc6 gene with the change in sign of the of disparity curve is a good indicator to find *Ori* site in archaeal genomes. Further the location of Origin Recognition Boxes (ORB) is necessary as a confirmation of the detection of the *Ori* site in Archae [12]. In Archae Profile database we have given all experimentally verified ORB's as well as *in-silico* predicted consensus ORB sequences and their location. The consensus ORB sequences were generated by multiple sequence alignment of predicted *Ori* region of closely related archeal genome sequences, and later on validated by matching conserved MSA region with *in-vivo* verified ORB of closely related archeal species if known.

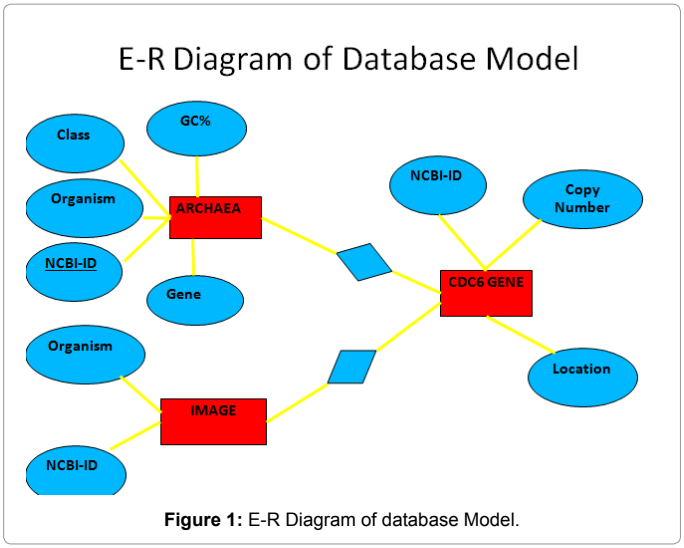

Figure 1: E-R Diagram of database Model.

Methods and Results

Most of the genomic data is collected from NCBI. (http://www.ncbi.nlm.nig.gov/genomes). The disparity curves were drawn using an Octave script (http://www.gnu.com/) based on the method given by Zhang and Zhang [13]. The cdc6 location was sought in vicinity of the change of sign of M-Y or R-K disparity curve. The upstream and downstream sequence up to 2 Kb length was scanned for AT richness and the same was investigated for the presence of consensus ORB's. As ORB sequences in archaea vary from species to species [14] and there are only few archaea for which the ORB sequences have been experimentally verified [12,15,16], we have used such sequences to find consensus ORB sequences in the closely related species using multiple sequence alignment. Presence of ORB sequence, cdc6 gene and AT rich region is a good indicator of *Ori* site in archaea.

Archaea Profile *Ori* site prediction results correspond with some of the *in-vivo* verified *Ori* site results. We have predicted three *Ori* site in Halobacterium species NRC-1 which have been confirmed by *in-vivo* studies [17]. 3 *Ori* sites is predicted by Archaea Profile in *Sulfolobus sofataricus* and *Ori in Pyrococcus abyssi* have been validated in-vivo experiment [18,19]. At Archaea Profile we have predicted single *Ori* in *Thermococcales and Methanogens,* three *Ori* site in *Sulfolobales* and Haloarchaea. Of course on some genomes we are totally unable to predict *Ori* site like *Pyrobaculum calidifontis* which has four *in-vivo* verified *Ori* site [20], nevertheless the coverage of Archae Profile *Ori* prediction is large and comprehensive.

Database implementation and web interface

Database was implemented on an open source Relational Database Management System MySQL (http://www.mysql.com) which allows rapid accession of data, hassle free update and maintenance. There are three relation tables which carries the whole record of the database (Figure 1). First table contains the data related to the archaeal species having the genome information and other numeric values. Each tuple contains the record of a single archaeal species. The second table consist of the GC graph in BLOB data type, which makes relation by NCBI-ID with first table. The third table contains the cdc6 and ORB record of all archaeal species. The outer web interface of the database for the remote user has been designed using PHP (http://www.php.net) for easy integration with database. It provides various options to search the database such as search by NCBI ID, search by Archaeal species name and search by class. Extensive module of Java (http://www.java.com) has been implemented for the real time help on the search page. AJAX library has been used to show the possible hits match with the database content submitted by user.

A disparity plotter based on BioPHP (http://www.biophp.org/minitools/skews/), and available in the public domain was also implemented with the database through which user can draw the disparity plot of any given DNA sequence up to 5 MB. The repeat finder tool has been developed on PHP and capable of parsing sequence nucleotide and protein sequence to find the exact location and copy number of a user defined pattern.

Conclusion

Several excellent databases have already been developed for prediction of *Ori* sites in prokaryotes particularly in archaea. Excellent example are DoriC [21-23] and Oriloc of Comparative Genometrics [9] databases, but their coverage is not exclusively for archaeal genomes. More than that they do not have tools integrated to draw disparity plots and find the repeats (In form of ORB'S). The Archae Profile database is exclusively for archaea, with integrated tools to draw disparity plot of user choice sequence and finding the specific repeats with their corresponding location and copy number. An important facet which makes Archae Profile distinct from other databases is update frequency, we cover almost all sequenced archaeal genomes till date and information is immediately updated on Archae Profile as soon as the complete sequence is available on Gene bank. Archae profile also has information related to some genes like Mini Chromosome Maintenance and Polynuclear Cell Nuclear Antigen, which is involved in replication of the archaeal genomes, is also given.

Tutorial aspect

In addition to the genomic information and Disparity plot provided on the databases provided on the database a separate tool has been provided to draw the disparity curve of the genomic sequence of the user choice. This will enable the users to understand the disparity hidden in a DNA sequence.

User support

Comments and feedback and questions from the scientific

community and users are highly appreciated. Please send your feedback and questions at http://www.bioinfommv.in/feedback.html

Future development

We will continue to incorporate new data related to archaeal genomes as it becomes available and update existing fields in the Archae Profile Database. With the expanded use of next-generation sequencing technologies, tracking updates to enhance accessibility and visualization of new functional data will be of growing importance. We are also implementing tools to find all possible repeats of desired length by calling octave library in PHP. The Archae Profile database will continue to focus on incorporating novel approaches to map *Ori* site in archaea.

References

1. Koonin EV, Mushegian AR, Galperin MY, Walker DR (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Molecular Microbiology 25: 619-637.
2. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proceedings of the National Academy of Sciences 86: 9355-9359.
3. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74: 5088-5090.
4. Woese CR, Achenbach L, Rouviere P, Mandelco L (1991) Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. Syst Appl Microbiol 14: 364-371.
5. van de Vossenberg JL, Driessen AJ, Konings WN (1998) The essence of being extremophilic: the role of the unique archaeal membrane lipids. Extremophiles 2: 163-170.
6. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol 6: 245-252.
7. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, et al. (1996) Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science 273: 1058-1073.
8. Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM (2006) The UCSC Archaeal Genome Browser. Nucleic Acids Res 34: D407-410.
9. Roten CA, Gamba P, Barblan JL, Karamata D (2002) Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. Nucleic Acids Res 30: 142-144.
10. Mojica F, Ferrer C, Juez G, Rodriguez-Valera F (1995) Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. Molecular microbiology 17: 85-93.
11. Capaldi SA, Berger JM (2004) Biochemical characterization of Cdc6/Orc1 binding to the replication origin of the euryarchaeon Methanothermobacter thermoautotrophicus. Nucleic Acids Res 32: 4821-4832.
12. Norais C, Hawkins M, Hartman AL, Eisen JA, Myllykallio H, et al. (2007) Genetic and physical mapping of DNA replication origins in Haloferax volcanii. PLoS Genet 3: e77.
13. Zhang R, Zhang CT (2005) Identification of replication origins in archaeal genomes based on the Z-curve method. Archaea 1: 335-346.
14. Wu Z, Liu H, Liu J, Liu X, Xiang H (2012) Diversity and evolution of multiple orc/cdc6-adjacent replication origins in haloarchaea. BMC Genomics 13: 478.
15. Dueber EL, Corn JE, Bell SD, Berger JM (2007) Replication origin recognition and deformation by a heterodimeric archaeal Orc1 complex. Science 317: 1210-1213.
16. Robinson NP, Bell SD (2005) Origins of DNA replication in the three domains of life. FEBS J 272: 3757-3766.
17. Coker JA, DasSarma P, Capes M, Wallace T, McGarrity K, et al. (2009) Multiple replication origins of Halobacterium sp. strain NRC-1: properties of the conserved orc7-dependent oriC1. J Bacteriol 191: 5253-5261.
18. Robinson NP, Dionne I, Lundgren M, Marsh VL, Bernander R, et al. (2004) Identification of two origins of replication in the single chromosome of the archaeon Sulfolobus solfataricus. Cell 116: 25-38.
19. Lopez P, Philippe H, Myllykallio H, Forterre P (1999) Identification of putative chromosomal origins of replication in Archaea. Mol Microbiol 32: 883-886.
20. Pelve EA, Lindås AC, Knöppel A, Mira A, Bernander R (2012) Four chromosome replication origins in the archaeon Pyrobaculum calidifontis. Mol Microbiol 85: 986-995.
21. Gao F, Zhang CT (2007) DoriC: a database of oriC regions in bacterial genomes. Bioinformatics 23: 1866-1867.
22. Gao F, Luo H, Zhang CT (2013) DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. Nucleic Acids Res 41: D90-93.
23. Luo H, Zhang CT, Gao F (2014) Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. Front Microbiol 5: 482.