Gene 582 (2016) 143–147

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Overlapping genes: A significant genomic correlate of prokaryotic growth rates

Deeya Saha, Soumita Podder, Arup Panda, Tapash Chandra Ghosh *

Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, India

ARTICLE INFO

Article history:
Received 16 November 2015
Accepted 3 February 2016
Available online 4 February 2016

Keywords:
Generation time
Archaea
rRNA multiplicity
tRNA multiplicity
Codon usage bias
Eubacteria
Overlapping genes

ABSTRACT

Elucidating the genomic features influencing prokaryotic growth rates has always been a study of interest. Previously, it was observed that overlapping genes (OGs) play a crucial role in the prokaryotic genome size reduction. This study is focused to explore whether OGs act as a potential correlate of prokaryotic growth rates. For this purpose, we compiled a dataset of 25 archaeal and 117 eubacterial genomes and analyzed the inter-correlation between the proportion of overlapping regions in these genomes with their growth rates. Here, we observed that the proportion of overlapping region holds a significant negative correlation with generation time in archaeal domain, whereas no correlation was observed in the eubacterial domain. However, after masking the effect of tRNA, rRNA multiplicity and environmental diversity, OGs show an independent effect over growth rates in the eubacterial domain as well as in the archaeal domain. Moreover, the influence of OGs on prokaryotic growth rates provides different delineations in archaeal and eubacterial domains. In archaea, both long overlap frequency (LOF) and short overlap frequency (SOF) influence the growth rates by increasing the degree of operonization. On the contrary, in the case of bacteria, neither SOF nor LOF plays any significant role in achieving faster growth rates.

1. Introduction

Prokaryotic growth rates vary from few minutes to several hours (Vieira-Silva and Rocha, 2010) Among host-associated bacteria, competition often results in increased virulence through selection for higher growth rates as these have an outstanding role in the trade-off between rapid horizontal dissemination and slow clearance from the host (Read, 1994; vanBaalen and Sabelis, 1995). In order to attain fast growth rates, prokaryotes acquire different genomic signatures such as enhanced tRNA (Dong et al., 1996; Ikemura, 1985), rRNA copy numbers (Klappenbach et al., 2000), elevated codon usage bias (Vieira-Silva and Rocha, 2010), replication-associated gene dosage (Couturier and Rocha, 2006) and pronounced strand bias (Mao et al., 2012). The growth rates of prokaryotes are also largely dependent on extrinsic factors like favorable temperature, nutrient supply and most importantly on bacterial lifestyle (Freilich et al., 2009).

Genome streamlining is an attractive concept, but many of the previous research could hardly explain the role of genome streamlining in the attainment of fast growth in prokaryotes as there was no correlation between genome size, coding density and generation time in the prokaryotic world (Kuo et al., 2009). Firstly, to minimize replication time one needs to shorten genome length. However, till date, all the previous studies failed to demonstrate any correlation between genome size and growth rates (Vieira-Silva and Rocha, 2010) and it has been argued that replication in prokaryotes could be initiated prior to previous round's completion (Vieira-Silva and Rocha, 2010). On the other hand, several other reports hypothesized that genome size reduction and simplification of prokaryotic genomes are elementary processes to achieve fast growth rates (Lynch, 2006). Previously, genome streamlining was observed in thermophilic prokaryotes where gradual elimination of intergenic regions facilitated the minimization of genome size (Sabath et al., 2013). However, to what extent the process of genome streamlining facilitates bacterial speed of multiplication is largely unknown. The incidence of genome streamlining is also expected to result in the formation of overlapping genes. Thus, it was proposed that OGs are involved in genome compaction (Sakharkar et al., 2005).

Archaea and eubacteria differ in various genomic and molecular adaptations (Mizuguchi et al., 2007). In our study, we intend to explore whether OGs play any significant role in the attainment of fast growth rates in archaeal and eubacterial domains and if yes what are the underlying mechanisms behind the correlation of OG with growth rates.

Abbreviations: OG, Overlapping gene; OP%, Percentage of genes in operons; LOF, Long overlap frequency; SOF, Short overlap frequency; CAI_{avg}, Codon Adaptation Index; TF, Transcription factor.

* Corresponding author.
E-mail addresses: deeya@jcbose.ac.in (D. Saha), soumita@jcbose.ac.in (S. Podder), arup@jcbose.ac.in (A. Panda), tapash@jcbose.ac.in (T.C. Ghosh).

http://dx.doi.org/10.1016/j.gene.2016.02.002
0378-1119/

2. Materials and methods

2.1. Retrieval of genomic information

Overlapping gene data were retrieved from PairWise Neighbors Database (Palleja et al., 2009). Generation time data used in this study were obtained from the supplementary files of (Vieira-Silva and Rocha, 2010). Out of 214 prokaryotic genomes for which generation time data were available, we found overlapping gene data for 206 prokaryotic genomes (25 archaea and 181 eubacteria). From these 206 genomes with both generation time and gene overlapping gene data, two prokaryotes (generation time greater than 72 h) were eliminated from the analysis as their generation time falls in the outlier category (Supplementary Table S1). Previously it was ascertained that host-associated bacteria do not follow streamlining process for their genome size compaction (Moran and Wernegreen, 2000). Therefore, we have also excluded 62 host-associated bacteria from our analysis. Finally, we performed our analysis with a dataset of 25 archeal genomes and 117 eubacterial genomes, a total of 142 prokaryotic genomes.

Environmental diversity data was retrieved from the supplementary table of Freilich et al. (Freilich et al., 2009). The number of transcription factors (TFs) in a given prokaryotic genome was retrieved from Transcription Factor Prediction database (DBD) (http://www.transcriptionfactor.org/). Habitat classifications were retrieved from NCBI genome projects. Information regarding tRNA and rRNA copy number in a given bacterial genome was derived from rRNdb (http://rrndb.umms.med.umich.edu/) (Klappenbach et al., 2001). Codon usage bias of genomes (measured as average codon adaptation index (CAI)) were taken from the supplementary table of Botzman and Margalit (2011) (Botzman and Margalit, 2011).

2.2. Calculation of overlapping gene frequency

The proportion of overlap of a given genome is calculated as the total length of overlapping regions of each overlap in a given genome divided by genome length. We eliminated overlap with a length greater than 59 bp from our dataset since those overlaps were suggested to be due to the consequence of misannotation (Palleja et al., 2008).

Overlapping gene frequency was calculated as the total number of adjacent gene pairs that overlap divided by a total number of adjacent gene pairs in that genome (Sabath et al., 2008). Long overlaps were defined as overlaps spanning a region of 7–50 nucleotides, whereas, short overlaps are defined as overlaps spanning a region of 1–4 nucleotides (Fonseca et al., 2014).

2.3. Statistical analyses

All statistical tests including bivariate correlations were done using SPSSv13. Spearman's partial correlation were conducted using TANAGRA 1.4. All the correlations of generation time with other factors mentioned in this manuscript were corrected for phylogenetic inertia using coevol1.4b (Lartillot and Poujol, 2011). Coevol is a Bayesian Monte Carlo Markov Chain (MCMC) program that is widely used for comparative analysis combining molecular data and quantitative traits. The program takes three types of input files, (*i*) multiple alignment file, (*ii*) phylogenetic tree, and (*iii*) a matrix of quantitative traits (e.g. life history traits). The posterior probability determines the significance level of correlation between life history traits. Lower the posterior probability higher the level of significance. The 16s rRNA sequences for archaea were downloaded from (https://rdp.cme.msu.edu/) and aligned using the program aligner and tree builder provided in https://rdp.cme.msu.edu/taking *E. coli* as an outgroup. The alignment and phylogenetic tree are available as Supplementary Fig. 1 and Fig. 2 respectively.

3. Results

3.1. Overlapping regions share significant correlation with generation time in archaeal domain

To study the effect of gene overlapping on prokaryotic growth rates here, we have compiled a dataset of 25 archeal genomes and 117 eubacterial genomes whose generation time as well as overlapping gene data were available. Here, we observed that in archaeal domain proportion of overlapping region share a strong negative correlation with generation time (Spearman's $\rho = -0.623$; $P = 1 \times 10^{-3}$) (Fig. 1A). On the contrary, we obtained no significant correlation between the proportion of overlapping region with generation time in the eubacterial domain (Spearman's $\rho = 0.110$; $P = 0.143$) (Fig. 1B). Phylogenetic dependencies could be a potential confounding factor in our analysis. Therefore, to test whether the correlations observed in the archeal domain are due to phylogenetic artifacts we tested the above associations using coevol1.4b (Lartillot and Poujol, 2011). The correlations after correction for phylogenetic inertia are as follows $r_{\text{OG-generation time}} = -0.595$; with posterior probability of 0.004; $r_{\text{genome size-generation time}} = 0.520$ with posterior probability 0.009). This in turn suggests that these correlations are independent of any phylogenetic artifacts.

3.2. Independent effect of OGs on generation time in archaea and eubacterial domain

Previous studies have shown that translational efficiency measured by tRNA multiplicity, rRNA multiplicity, codon usage bias (CAI_{avg}) are the key determinants of cell division rate in prokaryotes (Vieira-Silva and Rocha, 2010). However, the pronounced effect of OGs over generation time in archaeal domain prompted us to investigate the relative strength of transcriptional and translational efficiency in archaeal and eubacterial domains separately. We calculated four factors for each genome in our dataset that are related to the transcription and translation efficiency of prokaryotic genomes i) number of transcription factors (TFs) in a given genome ii) tRNA multiplicity iii) rRNA multiplicity iv) codon usage bias measured as CAI_{avg}. Moreover, it was proposed that habitat variability might influence prokaryotic growth rates (Freilich et al., 2009). In order to determine habitat variability, we used environmental complexity index (ENV_DI) as computed by (Freilich et al., 2009). The ENV_DI is an estimate of environmental diversity in which a given bacterium grows. For instance, *Desulfotalea psychrophila* inhabits specialized type of habitat and thus possesses a low ENV_DI, whereas *Pseudomonas aeruginosa* that has larger metabolic capabilities with more diverse environmental conditions possesses high ENV_DI. Then to exemplify the relative contribution of each genomic correlates with generation time, we performed partial correlation analysis between generation time and proportion of overlapping region after controlling all the above-mentioned factors. The result delineated in Table 1 represents that in both the prokaryotic domains (i.e. archaea and eubacteria) OGs play a significant role in attaining fast growth rates. Although, in the eubacterial domain, there is no correlation between the proportion of overlap and generation time, in partial correlation analyses we found a strong association of OGs with growth rates. This observation depicts that the influence of OGs on growth rates may be masked by the effect of translation and other habitat-related factors and becomes evident only when these effects are controlled for. Another major difference between archaeal and eubacterial domain is the impact of translation related factors (tRNA, rRNA and CAI_{AVG}). In the archaeal domain, there is no significant effect of translation related factors on growth rates, whereas, in eubacteria, rRNA copy number is one of the major determinants of growth rates accompanied by other translation related factors like CAI_{avg}. Our result thus suggests that archaea and eubacteria adopt two different mechanisms for attaining fast growth rates.