Journal of Advanced Zoology
ISSN: 0253-7214
Volume 44 Issue S-5 Year 2023 Page 240:245

Phylogenetic Comparative Methods Demonstrate The Non-Directional Evolution Of Genomic Features In The *Streptococcus* Genus

Akash Ajay[1]

[1]*School of Environmental Sciences, Jawaharlal Nehru University, New Delhi,110067, India*

Article History

Received: 06 June 2023
Revised: 11 Aug 2023
Accepted: 05 Sept 2023

Abstract

Streptococcus, a genus of Gram-positive bacteria, encompasses many species with diverse ecological roles, from commensals in the human microbiota to pathogens causing a spectrum of infections. Understanding the biology and evolution of *Streptococcus* is crucial for unraveling its significance in health and disease. In this paper, I shall study the trait evolution of 4 genomic features - Genome size, Genomic GC content, Genomic repeat fraction, and Number of coding genes in the Streptococcus genus. Using phylogenetic generalized least squares, I find a strong positive correlation between genome size and the number of coding genes, while other features are unrelated.

Keywords – Phylogenetic comparative modeling, trait evolution, streptococcus, PGLS

Introduction

The *Streptococcus* genus is a prominent and diverse group of bacteria characterized by their spherical or cocci-shaped cells (B.Spellberg., 2015)[1]. These Gram-positive bacteria are widely distributed across various environments, including soil, water, and the human body. One of their distinctive features is their tendency to form chains or pairs during cell division, which can be observed under a microscope (B. Lara et al., 2005)[2]. The genus encompasses various species, each with unique characteristics and adaptations.

Streptococci exhibit significant variability in terms of their ecological roles and interactions (JO Mundt, 1982)[3]. While some species are commensals (Kreth et al., 2017 [4]; Salvadori et al., 2019[5]), coexisting harmlessly with their host organisms, others can be opportunistic pathogens, causing various infections in humans and other animals (Cunningham, 2000)[6]. The adaptability and versatility of Streptococcus make it an intriguing subject of study in microbiology as researchers explore the factors determining whether these bacteria become beneficial residents or disease-causing agents within a host.

Moreover, the Streptococcus genus has practical importance in various industries and fields. Some species are involved in cheese and yogurt fermentation processes, contributing to their flavor and texture. Additionally, they have been studied extensively for their genetics and physiology, offering insights into fundamental

biological processes. Understanding the diverse roles and behaviors of Streptococcus bacteria contributes to our knowledge of microbiology, evolutionary biology, and the complex relationships between microorganisms and their hosts.

An organism's genome provides a detailed record of its evolutionary history and can give insights into its phenotype, metabolomics, and proteome (Ellegren , 2014)[7]. There are many genomic features that are studied in global level genome size, genomic GC, number of genes, non-coding regions, 3D genome organization, and local genomic contexts like - local mechanical and shape properties. I restrict myself to the study of whole genomic features in streptococcus genera. To study these features between closely spaced species, we need to account for the phylogenetic non-independency of data points. I shall use trait evolution models and phylogenetically corrected regressions to understand the pattern of genomic features in *Streptococci*.

Trait evolution models are essential tools in evolutionary biology and provide a framework for understanding how traits change over time in populations and species (Munkemuller et al., 2012)[8]. These models correct for phylogeny similarity of the species, which may play a role in trait similarity, and then explore questions related to adaptation and diversification of life forms. There are several trait evolution models in the literature; some of the popular ones are - Brownian motion (Felsenstein , 1985)[9], OU model (Butler and King, 2005)[10], EB model (Clavel et al., 2019)[11], lambda model (Ho et al., 2014)[12], and white noise model (Cooper et al., 2010)[13]. Brownian motion model, the most basic model of trait evolution, assumes that trait evolution occurs as a random walk, with trait values changing continuously along branches of a phylogenetic tree. Under this model, traits evolve without any specific directionality, and the variance of trait change is proportional to the time elapsed. Ornstein-Uhlenbeck's (OU) model, which incorporates stabilizing selection, posits that traits are subject to a restoring force that pulls them toward an optimal value, providing a mechanism for trait convergence. The OU model is especially relevant when examining traits adapted to specific ecological niches. The EB (Early Burst) model suggests that during the early stages of a lineage's evolutionary history, a rapid burst of diversification leads to the emergence of a wide variety of traits or species. Subsequently, this rate of diversification slows down over time. The Lambda model assumes that trait evolution follows a Brownian motion process, where trait values change continuously along phylogenetic tree branches. However, it allows for the rate of trait change to vary across branches, with λ serving as a scaling factor. A λ value of 1 indicates that the trait evolves constantly across the tree, while values greater than 1 suggest accelerated trait evolution along some branches, and values less than 1 indicate decelerated evolution. The white model shows how traits change over time and across different species. Unlike models that assume specific evolutionary processes like Brownian motion or Ornstein-Uhlenbeck, the White Noise Model assumes that trait evolution is entirely random, with no correlation between traits of closely related species.

This study investigates the evolution of 4 genomic features in the streptococcus genus - genome size, genomic GC, number of coding regions, and genomic repeat fraction. We explore the relationship between them with the help of phylogenetic regression, study their trait evolution with the help of the above-mentioned trait evolution models, and examine which one is most appropriate for each trait.

2. Methods

2.1 Plotting Phylogenetic tree to visualize the phylogenetic relationships

The genome for 48 *Streptococci* bacteria was downloaded from the NCBI database (Federhen et al 2012)[14], and their accessions are given in Table 1. The phylogenetic tree for the streptococcus genera was plotted using the TYGS server (Kolthoff et al., 2019)[15], and a 16S phylogenetic tree was taken. The phylogenetic tree is shown in Figure 1. The phylogenetic tree was also taken as a Newick format for further analysis.

2.2 Computation of genomic features

The genomic features - genome size, GC, and coding genes were obtained from NCBI sites. The SSR repeats in the genome were detected using the repeat finder plugin of Geneios Prime 2023. The repeat finder uses a k-mer approach to detect repeats and is database-independent, making it suitable for species for which repeat databases are unavailable (Benson, 1999)[16]. The repeat lengths were summated and divided by the total genome size to obtain the genomic fraction. The table showing the genomic features for each species is given in Table 1.

2.3 Phylogenetic modeling of trait evolution

For statistical analysis, we used the programming language R v.4.01 (RC Team., 2000)[17]. We used the Pearson correlation test to check for the correlations using the cor.test() function of the base package. The phylogenetic least squares regression was performed using ape (Paradis et al., 2019)[18] and caper packages (Orme et al., 2013)[19] in R. The phylogenetic comparative modeling was performed using the Geiger package of R and Brownian motion, OU model, EB model, lambda model, and white noise model was fitted, and their AICs were checked using aic function of the base package.

3. Results

Table 1 shows the species taken for analyses, their NCBI accession numbers, and their genomic features. We can see that 95% confidence intervals for genome size are (1.959 - 2.083) MB, for genomic GC (38.636 - 39.946) %, for coding genes (1831 - 1941), and for genomic repeat fraction (16.053 - 20.933)%. We can see a greater spread in the distribution of the values for genomic repeat fraction than genomic GC, which has a narrower range. The Pearson correlation between the genomic features was computed and plotted as a correlation heatmap.

Table 1

Species	Accession number	Genome size (MB)	Genome GC	coding genes	genomic repeat fraction(%)
Streptococcus agalactiae	NZ_CP012480.1	2.082	35.4	1993	26.171
Streptococcus anginosus	NZ_CP012805.1	1.966	38.8	1846	6.888
Streptococcus australis	NZ_LR134285.1	2.013	42.1	1858	15.188
Streptococcus caballi	GCF_000379985.1	2.122	40.4	2000	10.95
Streptococcus canis	GCF_010993845.2	2.073	39.69	1896	22.607
Streptococcus castoreus	GCF_000425025.1	1.883	37.8	1723	11.751
Streptococcus constellatus	GCF_900459125.1	1.903	38	1782	23.403
Streptococcus criceti	GCF_900459215.1	2.425	42.2	2072	27.342
Streptococcus cristatus	GCF_900475445.1	2.047	42.4	1936	46.558
Streptococcus didelphis	GCF_000380005.1	1.905	36	1344	7.405
Streptococcus downei	52087_B01	2.235	43.55	1918	33.322
Streptococcus dysgalactiae	NZ_CP066069.1	2.117	39.4	1961	21.645
Streptococcus equi	NZ_CP065054.1	2.141	41.2	1955	22.573
Streptococcus equinus	GCF_900102715.1	1.875	37.3	1773	7.361
Streptococcus ferus	GCF_900475025.1	1.872	42.7	1782	10.999
Streptococcus gallolyticus	NZ_CP054015.1	2.246	37.5	2153	16.28
Streptococcus henryi	GCF_900104235.1	2.425	38.55	2268	13.509
Streptococcus hyovaginalis	GCF_000420785.1	2.019	39.9	1877	8.75
Streptococcus ictaluri	GCF_000188015.2	2.23	38.2	1998	27.763
Streptococcus infantis	ASM18746v1	1.792	39.5	1729	11.818
Streptococcus iniae	ASM30091v1	1.995	36.6	1858	25.726
Streptococcus intermedius	NZ_LS483436.1	1.942	37.6	1807	12.217
Streptococcus lutetiensis	GCF_900475675.1	1.843	37.5	1745	21.512
Streptococcus macacae	GCF_900459485.1	1.923	37.8	1774	17.999
Streptococcus marimammalium	GCF_000380045.1	1.505	33.2	1402	10.95
Streptococcus massiliensis	GCF_900459365.1	1.864	41.6	1843	32.547
Streptococcus merionis	GCF_900187085.1	2.357	41.8	2094	21.701
Streptococcus minor	GCF_000377005.1	1.927	41.1	1885	7.148
Streptococcus mitis	ASM96000v1	1.984	40	1831	16.041
Streptococcus mutans	NZ_CP044221.1	1.984	36.8	1829	16.531
Streptococcus oralis	NZ_LR134336.1	1.973	41.1	1866	32.77
Streptococcus orisratti	GCF_000380105.1	2.097	38.5	2178	17.574
Streptococcus ovis	GCF_000380125.1	2.358	40.1	2291	12.817

Streptococcus parauberis	ASM290038v1	2.084	35.5	1977	19.433
Streptococcus pasteurianus	GCF_004843545.1	2.14	37.3	2080	15.142
Streptococcus peroris	GCF_000187585.1	1.286	39.4	1586	9.978
Streptococcus phocae	GCF_001302265.1	1.679	39.55	1555	9.16
Streptococcus plurextorum	GCF_000423745.1	2.103	41.1	2022	11.514
Streptococcus pneumoniae	NZ_CP020549.1	2.085	39.6	1954	13.485
Streptococcus pseudopneumoniae	NC_015875.1	2.172	39.8	2019	21.052
Streptococcus pseudoporcinus	48128_D02	2.124	37.35	1925	19.701
StreptococcuspyogenesNCTC12064	NZ_LS483338.1	1.791	38.4	1665	15.548
Streptococcus ratti	:GCF_008803015.1	2.079	40.8	1915	18.366
Streptococcus salivarius	Ssal_L25	2.185	39.5	1920	22.809
Streptococcus sobrinus	NZ_CP029491.1	2.112	43.5	1977	22.871
Streptococcus suis	NC_012926.1	2.106	41.1	1963	26.035
Streptococcus thermophilus	GCF_001657915.1	1.931	39.5	1830	26.288

We see a significant correlation between all the genomic parameters- genome size is positively correlated with genomic GC (R = 0.341, P = 0.009, N= 48), genomic repeat fraction (R = 0.318, P = 0.029, N = 48), number of coding genes (R = 0.865, P = 2.2E-16, N= 48). Genomic GC is positively correlated with genomic repeat fraction (R = 0.317, P = 0.030, N= 48) and number of coding genes (R = 0.386, P = 0.003, N= 48). Genomic repeat fraction is nearly correlated with the number of coding genes (R = 0.272, P = 0.065, N= 48). Some of these correlations can be understood in the context of pre-existing literature, which talks about the trends of genomic features. We also know that organismal complexity is correlated with genome size in organisms with smaller genomes (Gregory, 2000). Since protein-coding genes can be considered as a proxy of organismal complexity, we find them to be positively correlated. These trends are, however not seen in larger multicellular organisms, which tend to have much larger genomes and don't show a similar increase in geomic complexity. Since streptococci are prokaryotes, the positive scaling between genomic size and the number of coding genes is the expected behavior.

Figure 1

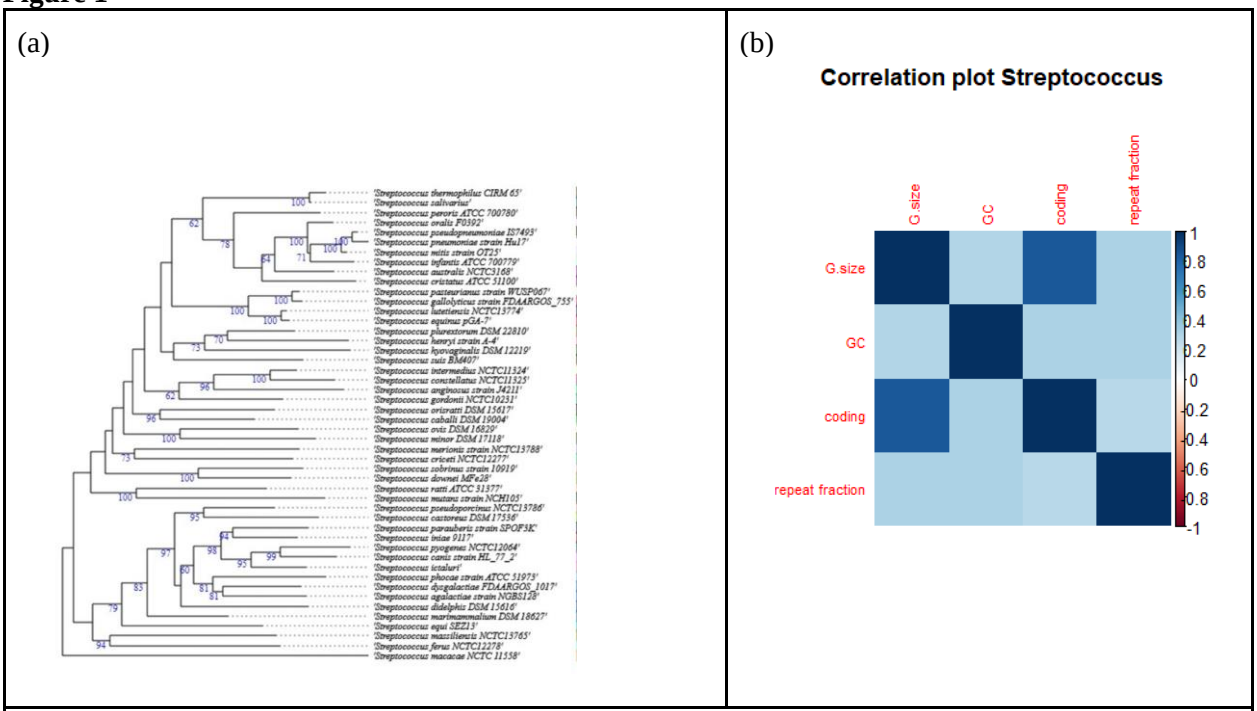

Figure 1 - (a) 16S phylogenetic tree of the streptococcal species considered under this study. (b) The correlation plot shows the correlation between the genomic features of the species.

We find that nearly all these traits are strongly correlated with each other, but this correlation can be misleading since all the members belong to the same genus and are thus closely related. To correct for phylogenetic influence, we applied for phylogenetic least squares (Symonds and Bloomberg, 2014)[21]. The results of phylogenetic least squares are shown in Table 2. Upon correcting for phylogenetic closeness, we see that most of these correlations disappear, indicating they are merely scientific artifacts. Only the positive correlation between genome size and the number of coding genes remains, indicating that the relationship between genome size and genomic complexity is also seen in streptococci, as in other prokaryotic phyla.
We also see a significant positive correlation between genomic GC and repeat fraction, but after phylogenetic correction, it has a low R^2, indicating a negligible interaction.

Table 2

Independent variable	Dependent variable	R^2	P	kappa	lambda	delta
Genome size	Coding	0.7057	9.44E-14	1	1	1
Genome size	GC	-0.0002874	0.3258	1	1	1
Genome size	genomic repeat fraction	0.05344	0.0643	1	1	1
GC	Coding	0.01411	0.2044	1	1	1
GC	genomic repeat fraction	0.06761	0.04305	1	1	1
genomic repeat fraction	Coding	0.07757	0.0325	1	1	1

To check for the trait evolution of genome size, genomic GC , genomic repeat fraction and coding genes in Streptococcus genus we fit the trait values to Brownian motion, Ornstein -Uhlenbeck model, early burst model, lambda model and white noise model. The AIC values of these models are tabulated in Table 3

.

Table 3

AIC Values	Brownian Motion	OU model	EB model	lambda model	white noise
Genome size	**-7.508**	2.561	-5.508	-5.508	**-7.508**
Genomic GC	189.385	189.378	**185.747**	191.385	213.808
Coding genes	**629.081**	629.261	631.918	631.117	630.293
Genomic repeat fraction	**337.423**	362.793	339.423	338.242	**337.423**

In the above table we have marked the best performing model for each of the traits in bold. We can see that Brownian motion best explains the evolution of genomic size, genomic repeat fraction, and number of coding genes. This may indicate that during evolution , the descendant nodes of the common ancestor might have undergone genome expansion and contraction, giving rise to a wider range of values. It seems to indicate a lack of directionality for genome size evolution and complexity in streptococci. Unlike other traits, genomic GC is the only trait for which the best-fit model is not Brownian motion , and fits better to an early burst model. The white noise and Brownian motion models perform well for both genome size and repeat fraction.

4. Discussion

Our study of the evolution of 4 genomic features- genome size, genomic GC , genomic repeat fraction, and number of coding genes in the *streptococcus* genus showed that the random walk or Brownian motion model was the best performing model for 3 out of 4 features. The white noise model also explains the trait evolution well for two of 4 genomic features. Both Brownian motion and white noise models agree on the lack of any directional trend in the evolution of the traits concerned. This implies that genome size, complexity, and genomic repeat fraction tend to be non-directional in their evolution in streptococci and may even be independent of phylogeny. Genomic GC shows some phylogenetic influence in diversification across the genus. It shows rapid diversification in the early stages, followed by a slow pace.
Our results also show that these genomic features are not related to each other if corrected for phylogeny. The only exceptions are genome size and number of coding genes, which show a strong correlation post-phylogenetic correction as well, which was expected in a prokaryotic taxon. Genomic GC and repeat fraction are significantly correlated post phylogenetic corrections, though the correlation is weak given its low R^2. From

our study, we conclude that there is no directional trend for the evolution of the 4 genomic features in streptococci, and these 4 features, except for genome size and complexity, are independent of each other.

References -

1. Spellerberg, B., & Brandt, C. (2015). Streptococcus. *Manual of clinical microbiology*, 383-402.
2. Lara, B., Rico, A. I., Petruzzelli, S., Santona, A., Dumas, J., Biton, J., ... & Massidda, O. (2005). Cell division in cocci: localization and properties of the Streptococcus pneumoniae FtsA protein. *Molecular microbiology*, *55*(3), 699-711.
3. Mundt, J. O. (1982). The ecology of the streptococci. *Microbial ecology*, *8*, 355-369.
4. Kreth, J., Giacaman, R. A., Raghavan, R., & Merritt, J. (2017). The road less traveled–defining molecular commensalism with Streptococcus sanguinis. *Molecular Oral Microbiology*, *32*(3), 181-196.
5. Salvadori, G., Junges, R., Morrison, D. A., & Petersen, F. C. (2019). Competence in Streptococcus pneumoniae and close commensal relatives: mechanisms and implications. *Frontiers in cellular and infection microbiology*, *9*, 94.
6. Cunningham, M. W. (2000). Pathogenesis of group A streptococcal infections. *Clinical microbiology reviews*, *13*(3), 470-511.
7. Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in ecology & evolution*, *29*(1), 51-63.
8. Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., & Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, *3*(4), 743-756.
9. Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1-15.
10. Butler, M. A., & King, A. A. (2004). Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The american naturalist*, *164*(6), 683-695.
11. Clavel, J., Aristide, L., & Morlon, H. (2019). A penalized likelihood framework for high-dimensional phylogenetic comparative methods and an application to new-world monkeys brain evolution. *Systematic Biology*, *68*(1), 93-116.
12. Tung Ho, L. S., & Ané, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic biology*, *63*(3), 397-408.
13. Cooper, N., Jetz, W., & Freckleton, R. P. (2010). Phylogenetic comparative approaches for studying niche conservatism. *Journal of evolutionary biology*, *23*(12), 2529-2539.
14. Federhen, S. (2012). The NCBI taxonomy database. *Nucleic acids research*, *40*(D1), D136-D143.
15. Meier-Kolthoff, J. P., & Göker, M. (2019). TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nature communications*, *10*(1), 2182.
16. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, *27*(2), 573-580.
17. Team, R. C. (2000). R language definition. *Vienna, Austria: R foundation for statistical computing*, *3*(1).
18. Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claude, J., Cuong, H. S., & Desper, R. (2019). Package 'ape'. *Analyses of phylogenetics and evolution, version*, *2*(4), 47.
19. Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., & Pearse, W. (2013). The caper package: comparative analysis of phylogenetics and evolution in R. *R package version*, *5*(2), 1-36.
20. Ryan Gregory, T. (2002). Genome size and developmental complexity. *Genetica*, *115*, 131-146.
21. Symonds, M. R., & Blomberg, S. P. (2014). A primer on phylogenetic generalised least squares. *Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice*, 105-130.