Heliyon 10 (2024) e32103

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Detecting genetic gain and loss events in terms of protein domain: Method and implementation

Boqian Wang, Yuan Jin, Mingda Hu, Yunxiang Zhao, Xin Wang, Junjie Yue, Hongguang Ren *

Beijing Institute of Biotechnology, State Key Laboratory of Pathogen and Biosecurity, Beijing, China

ARTICLE INFO

Keywords:
Bacterial evolution
Protein domain
Phylogenetic tree
Shigella

ABSTRACT

Continuous gain and loss of genes are the primary driving forces of bacterial evolution and environmental adaptation. Studying bacterial evolution in terms of protein domain, which is the fundamental function and evolutionary unit of proteins, can provide a more comprehensive understanding of bacterial differentiation and phenotypic adaptation processes. Therefore, we proposed a phylogenetic tree-based method for detecting genetic gain and loss events in terms of protein domains. Specifically, the method focuses on a single domain to trace its evolution process or on multiple domains to investigate their co-evolution principles. This novel method was validated using 122 *Shigella* isolates. We found that the loss of a significant number of domains was likely the main driving force behind the evolution of *Shigella*, which could reduce energy expenditure and preserve only the most essential functions. Additionally, we observed that simultaneously gained and lost domains were often functionally related, which can facilitate and accelerate phenotypic evolutionary adaptation to the environment. All results obtained using our method agree with those of previous studies, which validates our proposed method.

1. Introduction

The gain and loss of genes are the major driving forces of bacterial evolution and environmental adaptation [1,2]. The purpose of detecting gene evolution is to understand how genes have changed over time, identify genetic variations, and gain insights into the evolutionary relationships among species, which could have a significant impact on both economic and biological development [2,3].

Currently, genetic evolutionary processes are primarily detected at the nucleotide sequence level. Some methods, such as RDP4, focus on gene fragments to identify and quantify recombination events between different regions of the genome by analyzing sequence alignment data [4]. Some methods detect genetic gain and loss events over the entire gene length range [1,2,5–8]. Other methods rely on statistical models, such as k-mers, to detect genetic gain and loss events in genes of fixed lengths [9–12]. However, nucleotide sequence analysis may not capture all aspects of gene evolution such as structural changes. Furthermore, analyzing gene evolution based solely on nucleotide sequences can be computationally intensive, particularly when dealing with large datasets or complex evolutionary histories that require significant computational resources and time [9,10].

The use of domains for the phylogenetic evolutionary analysis of bacteria offers unique advantages over nucleotide sequences. Protein domains are regions within a protein molecule that can independently maintain specific functional structures, typically

* Corresponding author.
E-mail address: bioren@163.com (H. Ren).

https://doi.org/10.1016/j.heliyon.2024.e32103
Received 15 March 2024; Received in revised form 8 May 2024; Accepted 28 May 2024
Available online 31 May 2024
2405-8440/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

composed of a series of amino acids [13,14]. These domains play crucial roles in protein functionality and can perform various biological functions, such as binding specific molecules, catalyzing chemical reactions, and enabling proteins to execute diverse and complex biological functions within cells, thus forming the biological basis for specific phenotypes in bacteria [15]. The rationale for utilizing structural domains in phylogenetic analysis is their high conservation and functionality compared with nucleotide sequences, which may undergo significant changes due to mutations. Therefore, utilizing domains can provide more stable and reliable analytical results and better reveal evolutionary relationships in biology [16,17].

Our approach was based on a phylogenetic tree to detect gene gain and loss events in bacteria during evolution from a domain perspective. To explore and trace credible bacterial evolutionary processes, we first used bac120, a more commonly used standard in bacteria, to construct a phylogenetic tree [18]. Subsequently, based on the principle of maximum parsimony, which follows the idea that biological systems tend to favor the simplest solutions during evolution, we annotated the domain scenarios possessed by each evolutionary node, ensuring that the labeling results align better with the evolutionary history [19,20]. Finally, relying on the systematically constructed phylogenetic tree and node annotation outcomes, we detected gene gain and loss events in bacterial evolution by examining the differences between parent and child nodes in terms of the domain. To further validate the reliability of this method, we applied it to 122 *Shigella* isolates to investigate its evolutionary process and verify its feasibility.

2. Methods

The overall architecture of our method includes three main parts marked in different colors (Fig. 1): phylogenetic tree interface, tag decision and node clustering, and gain and loss detection for single/multiple domains.

2.1. Phylogenetic tree interface

The proposed method is based on a phylogenetic tree to detect gain and loss events in terms of the domain. Given the sequences of evolution-related isolates, we first established a phylogenetic tree and rooted it according to a preset outgroup. The leaf nodes in the phylogenetic tree represent the target isolates and the root/internal nodes represent the ancestors during the evolutionary process of the target isolates. It utilizes a Finite State Machine (FSM), which is an automated computer-based analysis process, to represent the phylogenetic tree in 'Newick' format (Supplementary File 1).

A given phylogenetic tree records the position of each node (root, internal, or leaf) and the distance from each child node to its father node, which is utilized by the following processes.

2.2. Tag decision & nodes clustering

A bacterial isolate usually contains thousands of domains that can be searched using PfamScan [14,21] either online or offline. All the domains in each isolate formed a domain set. For each domain, the leaf nodes of the phylogenetic tree were tagged according to the

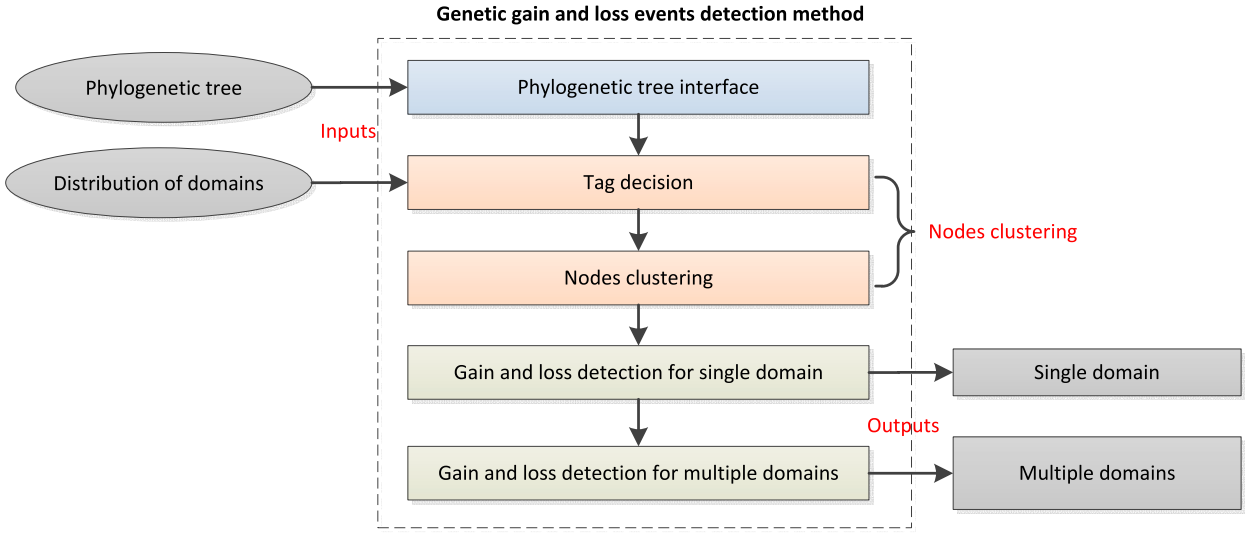

Fig. 1. Structure of the proposed method. The method consists of three parts (phylogenetic tree interface, nodes clustering, and gain and loss detection) and takes the phylogenetic tree and domain distribution file as inputs to detect genetic gain and loss events in terms of the domain. The first part utilizes an FSM-based phylogenetic tree interface to resolve the phylogenetic tree structure to facilitate the automated process in the next two parts. The second part tags each phylogenetic node to determine whether it contains the target domain and clusters them according to their tags and neighboring relationships in the phylogenetic tree. The third part detects genetic gain and loss events according to the clustering results and generates the analysis results from two aspects.

following principle (Fig. 2a).

- The leaf node that included the domain was marked as '1'.
- The leaf node that excluded the domain was marked as '0'.

Based on a customized maximum parsimony principle (Fig. 2a and b), the method determines the tag of each root and internal node to infer whether they contain the domain (Algorithm: Supplementary File 2). The customized maximum parsimony principle can minimize the number of gain and loss events during evolution, and has been widely used in various biological fields [19,22,23]. Nodes connected to the same tag are clustered together (Algorithm: Supplementary File 3), and each cluster is represented by its top node, which is utilized in the next step (Fig. 2c and d).

2.3. Gain and loss events detection

2.3.1. Single domain

Based on the cluster above, the method detected domain gain and loss events during the evolutionary process of the target bacteria, following the principles listed below. It is evident that gain and loss events occurred only in the top node of each cluster. In this example, nodes '000′ and '0100′ gain domains (Fig. 2c and d).

- Gain event: The child node includes the domain, whereas the father node excludes it.
- Loss event: The child node excludes the domain, whereas its father node includes it.

2.3.2. Multiple domains

In the above process, we focused on each domain to detect gain or loss events at each phylogenetic node. In this subsection, we focus on each phylogenetic node to collect all domains gained or lost compared to its father node, which is based on the statistical analysis of the results of the single domain. For example, we first detected the evolutionary processes of domains 1 and 2 separately (Fig. 3a and b). Then, focusing on each node ('000' & '0100'), we combined the detection results for statistical analysis (Fig. 3c). The domains in each phylogenetic node can be analyzed from two perspectives.

- Domain combination: We defined a domain combination as a combination of domains that were always simultaneously gained or lost by ancestral nodes in the phylogenetic tree.
- Number statistics: Number of domains involved in gain and loss events for ancestral nodes in the phylogenetic tree during evolution.

Compared to a single domain, the results of multiple domains can provide a more comprehensive understanding of their functional relationships and influence on evolution.

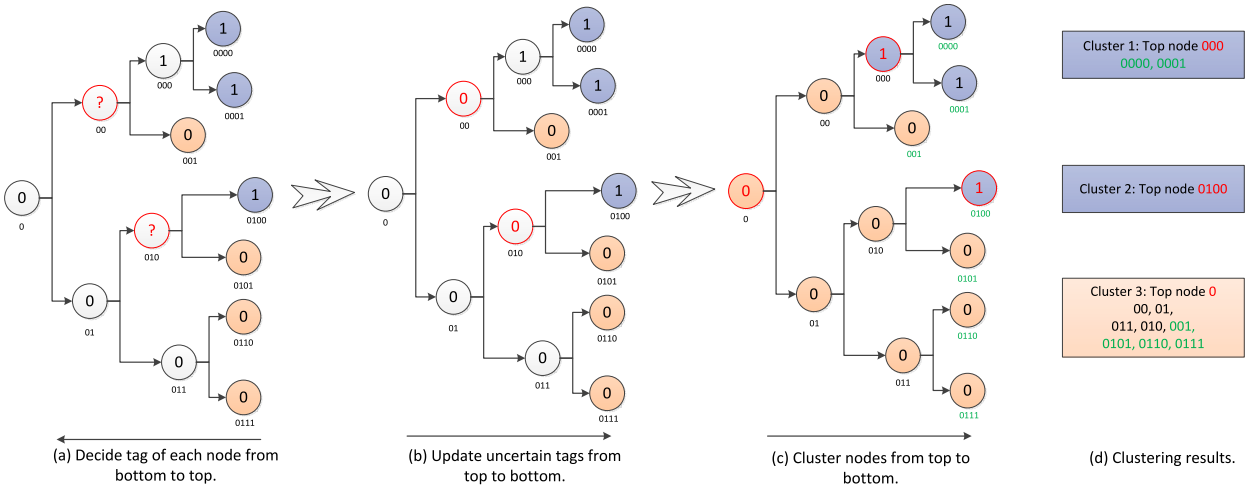

Fig. 2. (a and b) Tag-decision process. First, the tags of the middle and root nodes are decided from bottom to top according to the customized maximum parsimony principle. Then, the node with an uncertain tag is decided again from top to bottom, according to the tag of its father node. c and d) Clustering process. Based on the tags determined above, nodes with the same tag and neighboring nodes are clustered together. Each cluster is represented by its top node, which will later be used to detect gain and loss events.

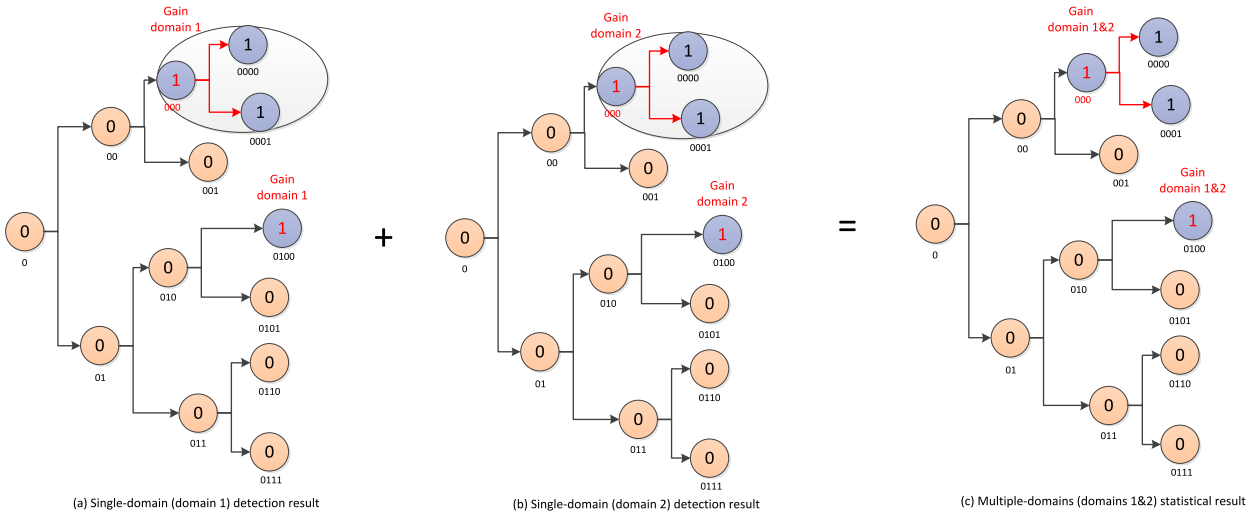

Fig. 3. (a and b) Gain and loss events of domains 1 and 2. For simplification, we adopted the results from Fig. 2, assuming that the detection results for domains 1 and 2 are identical. c) Statistical results for domains 1 and 2. We combined the gain and loss events of domains 1 and 2 across various nodes in the phylogenetic tree. Extrapolating this scenario to all domains, statistical analysis of the gain and loss events of all domains involved at each node in the phylogenetic tree can be conducted.

3. Implementation

3.1. Data selection

We selected complete and annotated *Shigella* genomic sequences from the National Center for Biotechnology Information (NCBI) database and downloaded 122 proteomes (Supplementary File 4). Additionally, two reference sequences, *Salmonella bongori* and *Salmonella enterica*, were included as outgroups to determine the root of the phylogenetic tree.

3.2. Phylogenetic tree

The 16s rRNA standard has the limitations of low phylogenetic resolution at the highest and lowest taxonomic ranks and missing diversity due to primer mismatches [24–27]. To address these issues, we used bac120, a more commonly used standard in bacteria, to construct our phylogenetic tree [18]. The analysis results greatly depend on the phylogenetic tree, which requires good markers, such as bac120 for bacteria, to make the results more consistent with the true evolutionary process.

The 120 ubiquitous single-copy proteins (bac120) were extracted and aligned according to the Genome Taxonomy Database (GTDB) by GTDB-Tk [16,18,28]. Based on the aligned sequences, a phylogenetic tree was constructed using an IQ-tree [29] under the JTT + F + R2 model (determined by the model finder), with a preferred bootstrapping value of 1000. The IQ-tree employs the Maximum Likelihood method as the default method to construct trees, which attempts to find the most probable evolutionary tree based on a probability model given observed data, such as sequence alignments. In theory, other softwares using the Maximum Likelihood for tree construction, such as RAxML [30] and MEGA [23], would similarly apply our method. We use an IQ tree as an example. The results are illustrated in Figtree [31] according to the outgroup species.

3.3. Distribution of domains

The domains included in the 122 isolates were searched using PfamScan [14,21] online or offline, and further polished to select the best match among the domains with overlaps. For each domain, we recorded the distribution of all the isolates.

3.4. Extension implementation

To demonstrate the universal applicability of our method, we randomly selected 100 strains of *Escherichia coli* from the NCBI database (Supplementary File 5) and applied them to evolutionary analysis of *E. coli*. The application process follows the procedure described above.

4. Results

4.1. Single domain

We selected PF00161.20 (Fig. 4), which is an important domain found in subunit A of the Shiga toxin 2 protein, as an example to detect gain and loss events during the evolutionary process of *Shigella* and visualize the results using iTOL [32]. In the phylogenetic tree, our method marked the main branch nodes where the domain was gained or lost during evolution. After each leaf node, which represents the isolate to be analyzed, our method will show all the gain and lost events its ancestor nodes experienced during the evolution process from the very beginning of the root node. The results showed that the domain was first acquired during the evolutionary process of *S. dysenteriae*, and then transferred to two isolates of *S. sonnei*, forming the main distribution in *S. dysenteriae* and a sporadic distribution in *S. sonnei*.

4.2. Multiple domains

We further showed the number of combinations of domains gained or lost simultaneously during the evolutionary process (Fig. 5). A total of 49 combinations were involved in gain events (green bar) and 57 combinations were involved in loss events (red bar). This combination consisted of nine domains.

In particular, we examined the functional relationships among the domains in combination. We collected 24 domain combinations that always appeared simultaneously in both the gain and loss events. Their names and functional annotations from the Pfam website (https://pfam.xfam.org/) are provided in Table 1. Interestingly, we observed that the domains in this combination were functionally related. The first record consisted of five domains associated with the bacterial secretion system. These domains work together to facilitate the secretion of virulence proteins during invasion.

Additionally, in the major differentiation nodes, we collected the number of domains gained or lost compared to its father node (Fig. 6). The results showed that *S. dysenteriae* lost 403 domains to differentiate it from the other three species, while *S. boydii* lost 189 domains to distinguish it from *S. flexneri* and *S. sonnei*. In addition, in *S. sonnei*, continuous loss of 121 and 84 domains was observed prior to its differentiation from *S. flexneri*.

4.3. Results for E. coli

PF00161.20 is present in the majority of *E. coli* strains, which is not the case for *Shigella* strains (Supplementary File 6). This indicates that Shiga toxin 2 is only found in a few *Shigella* species but is widespread in *E. coli* species [33]. This domain-level observation highlights *E. coli* as the main host of Shiga toxin 2. Additionally, even among closely related *E. coli* strains, there were instances of domain loss, indicating that both loss and gain events in this domain occur relatively frequently within this genus [34].

Furthermore, unlike the abundance of loss events observed during *Shigella* evolution, gain and loss events at the differentiation nodes of *E. coli* were relatively balanced (Supplementary File 6). Significant genomic differences exist among individual *E. coli* strains, with frequent domain loss and acquisition events. This also corresponds to the fact that while core genes among various species within the *Escherichia* genus exhibit minor differences, there are substantial differences in their entire genomes [35].

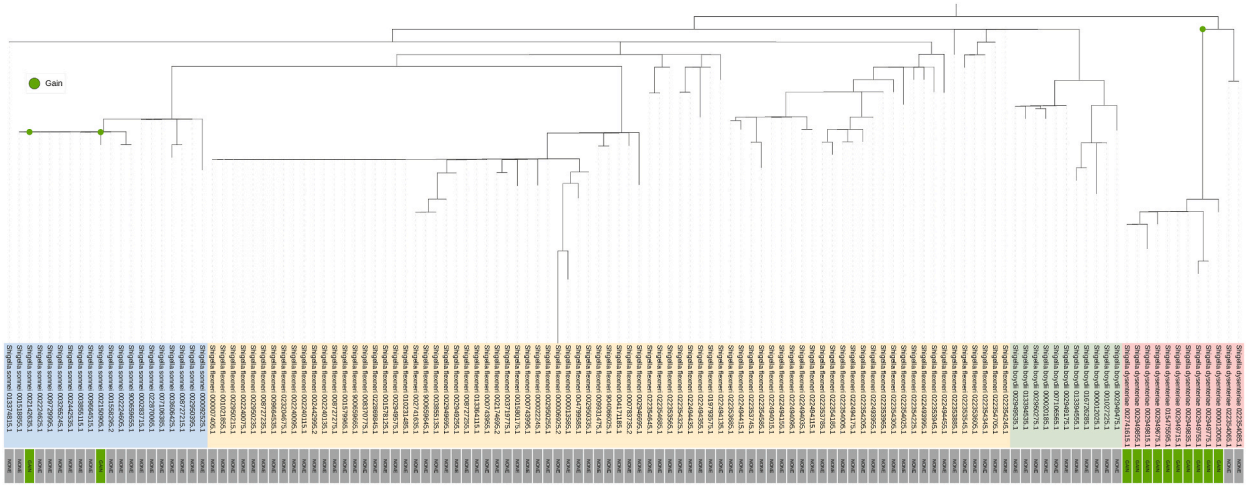

Fig. 4. Gain and loss events of PF00161.20 in the phylogenetic tree of the 122 *Shigella* isolates. PF00161.20 is a domain in subunit A of the Shiga toxin 2 protein. We marked the important evolutionary nodes where the gain event of PF00161.20 happened. The events that may have occurred during the entire evolution of a species are listed next to the species name. In this case, PF00161.20 was firstly gained by the ancestor of *S. dysenteriae* and then transferred to *S. sonnei*.

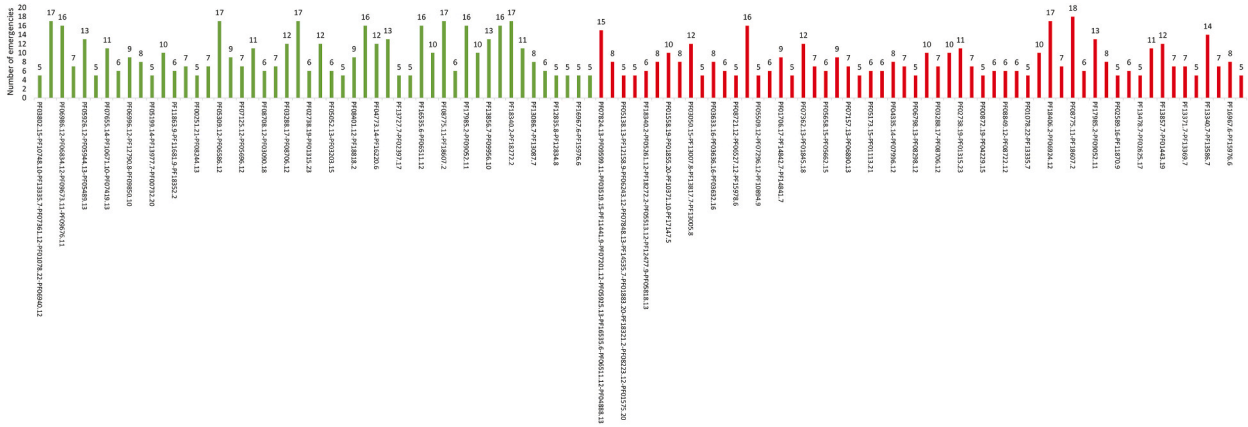

Fig. 5. Statistical analysis of the different domain combinations. The histograms depict the combination of domains involved in at least five gain or loss events. Specifically, the green bars represent 49 domain combinations associated with gain events and the red bars represent 57 domain combinations associated with loss events. Among the 106 results, a combination can consist of a maximum of nine domains.

Table 1
All domain combinations that always appear simultaneously in both gain and loss events during the evolutionary process of our selected 122 *Shigella* species. The names and functional annotations were obtained from Pfam (https://pfam.xfam.org/), and the domains in a combination are usually functionally related.

Combinations of domains	Frequency	Annotation of domains
PF07824.13-PF05925.13-PF03519.15-PF09599.11-PF11441.9	17	Type III secretion system - virulence protein - invasion protein - secretin
PF09676.11-PF09673.11-PF06986.12-PF06834.12	16	Conjugative transfer system (type IV, type-F, TraN, TraU)
PF10671.10-PF07419.13-PF07655.14	11	Pilus biosynthesis -pilin transport - secretary
PF05261.12-PF05513.12	17	Conjugative transfer and resistance (TraM, TraA)
PF05135.14-PF05521.12	7	Phage (head-tail connector protein)
PF08244.13-PF00251.21	7	Glycoside hydrolase
PF08706.12-PF03288.17	8	D5 protein
PF02738.19-PF01315.23	11	Xanthine dehydrogenase (molybdenum)
PF05662.15-PF05658.15	6	Trimeric autotransporter adhesin
PF06890.13-PF07157.13	7	Phage (Spike protein, DNA circularization)
PF16535.6-PF06511.12	17	T3SSs (IpaD, SipB)
PF11100.9-PF09679.11	16	Conjugative transfer system
PF17482.3-PF04984.15	10	Phage tail sheath protein
PF13728.7-PF07916.12	16	F pilus assembly (TraG, TraF)
PF14000.7-PF05354.12	9	Phage (DNA-packing, head-tail attachment)
PF09052.11-PF17985.2	13	Salmonella invasion protein A
PF18607.2-PF08775.11	18	Par system (ParA, ParB)
PF18340.2-PF18272.2	17	DNA relaxase TraI
PF13087.7-PF13086.7	7	AAA proteins
PF12293.9-PF11393.9	11	Type IV secretion system
PF15976.6-PF16967.6	8	Fimbrial proteins
PF13637.7-PF12796.8	5	Ankyrin repeat
PF11650.9-PF11134.9	6	Phage (P22, stabilization)
PF13335.7-PF01078.22	5	Magnesium chelatase

5. Discussion

In this study, we present a phylogenetic tree-based method to detect genetic gain and loss events during evolution in terms of domains. For ease of use, the results are formatted for easy reading, analysis, and visualization using well-developed programs or platforms such as iTOL and Flourish [36].

In the implementation, we first focused on PF00161.20 to detect its evolution process (Fig. 4). This is an important domain of Shiga toxin 2 protein, which is a virulence factor produced by *S. dysenteriae* [37]. *S. dysenteriae* primarily acquires virulence factors through vertical gene transfer, whereas a few *S. sonnei* strains occasionally acquire virulence factors through horizontal transfer [34–36]. Although no related cases have been reported for the other two species, Shiga toxin 1 has been found in *Shigella flexneri*, indicating the potential for the other two species to carry Shiga toxin 2 [37,38]. Therefore, in clinical treatment, special attention should be paid to *S. dysenteriae* with respect to this virulence factor, whereas other species, particularly *S. sonnei*, should not be overlooked because of their potential harm.

By combining the results of the single domains, we found that a fixed combination of multiple domains usually involves many gain

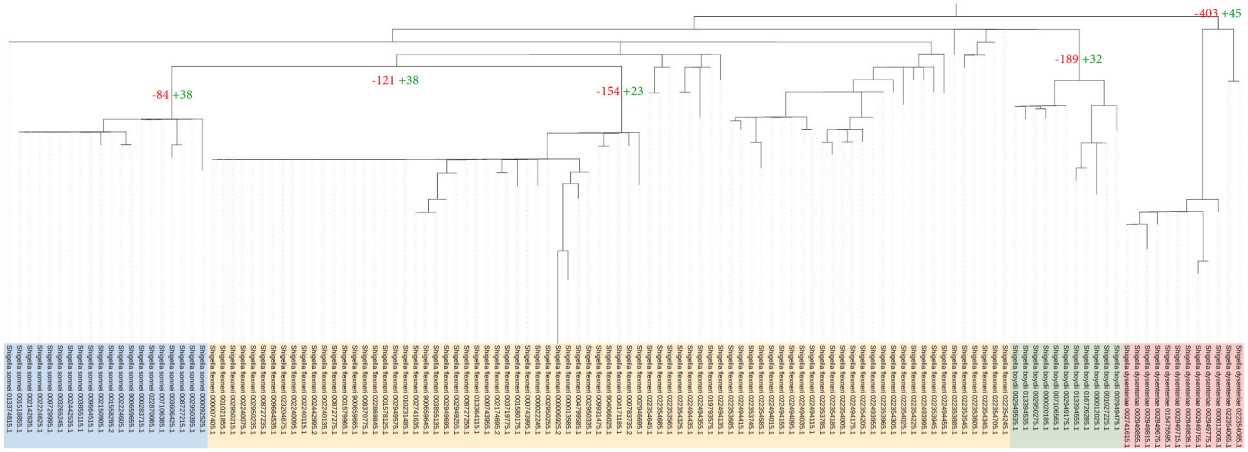

Fig. 6. Statistical results of gain and loss events in key differentiation nodes. In the phylogenetic tree of the 122 *Shigella* isolates, *S. dysenteriae* lost 403 domains to differentiate it from other three species, whereas *S. boydii* lost 189 domains to differentiate it from *S. flexneri* and *S. sonnei*. Additionally, in *S. sonnei*, a continuous loss of 121 and 84 domains was observed prior to its differentiation from *S. flexneri*.

and/or loss events (Fig. 5). Interestingly, these domains are functionally related and can potentially form biological phenotypes (Table 1) because bacterial gene evolution relies heavily on structures, such as transposons and plasmids, which can acquire and lose multiple genes simultaneously [38,39]. Several of these genes are functionally related. For example, in the classic mobile colistin resistance (mcr-1) transposon structure "*ISApl1-mcr-1-pap2-ISApl1*", the PAP2 gene promotes the expression of the mcr-1 resistance gene [40]. Similarly, many genes in the Tn125 transposon (*ISAba125-bla*$_{NDM}$*-ble-trpF-tat-dct-groES-groEL-ISAba125*) are associated with expression of the New Delhi metallo-beta-lactamase resistance gene [41]. Therefore, the evolution of these genes is often interrelated, which explains why we observed functionally related genes that were involved in the same gain or loss events. This phenomenon helps an isolate rapidly acquire or lose a certain phenotype to adapt to environmental changes, which is a beneficial mechanism for accelerating the evolutionary process [42,43].

Regarding the specific case of *Shigella* evolution, we identified a large number of loss events compared to gain events in terms of the domain during the evolution of the 122 *Shigella* isolates (Fig. 6). In particular, many loss events were detected and observed for divergent nodes in the phylogenetic tree. These results indicated that the loss of domains could be the primary driving force for *Shigella* evolution [34,44]. By examining the rate of gene loss in two groups of facultative pathogenic bacteria, pathogenic *E. coli* and *Shigella*, the results showed that *Shigella* strains lost genes at an accelerated rate relative to pathogenic *E. coli*. This demonstrates that a genome-wide reduction in the effectiveness of selection contributes to the observed increase in the rate of gene loss in *Shigella*, which is more like an intelligent upgrade process characterized by continuously discarding numerous non-essential phenotypes, reducing energy expenditure, and acquiring crucial beneficial phenotypes [44]. This phenomenon primarily arises from the convergent evolution of *Shigella* niche adaptation, mostly owing to loss of function and negative selection pressure [34]. This result indirectly validates the effectiveness of our approach in producing accurate outcomes.

In conclusion, our method can detect genetic evolution in both single and multiple domains. From the perspective of a single domain, this provided a detailed evolutionary process for a certain domain. From the perspective of multiple domains, this offers the opportunity to delve deeply into the rapid evolutionary process of bacterial phenotypes, thereby providing a more comprehensive understanding of the evolutionary relationships between bacterial genotypes and phenotypes. By implementing this method on 122 *Shigella* isolates, we successfully traced their evolutionary processes over a considerable historical period in terms of domains and explored the evolutionary relationship between their genotypes and phenotypes from the multiple-domain dimension, which could provide precise data and theoretical basis for relevant research. Furthermore, related studies corroborated the application of our method to *E. coli*, indicating the general applicability of our approach.

Code and data availability

Our proposed method (domain_gain_loss_detection.py) and programs for searching the distribution of domains (pfam_scan.py and pfam_domain_postprocess.py) were uploaded to GitHub (https://github.com/wr-sky/Domain-Gain-Loss-Detection/tree/main/program/) to facilitate the implementation of the method on their own dataset. The relevant materials, including the inputs, intermediate results, and outputs, were also uploaded to GitHub (https://github.com/wr-sky/Domain-Gain-Loss-Detection/tree/main/input_output/).

Ethics statement

Review and/or approval by an ethics committee was not required for this study because it involved only a retrospective analysis of bacterial genome data from public databases, which did not involve interactions with human subjects or interventions on their behalf.

Informed consent was not required for this study because it only involved the analysis of bacterial genome data from public databases, and did not involve human samples or information.

Data availability statement

All data associated with this study have been deposited in a publicly available repository to help other researchers evaluate our findings and build on our work. The codes used in this study have been uploaded to GitHub (code and data availability), and all genomes utilized in this study were downloaded from NCBI with the accession numbers recorded in Supplementary Files 4 and 5.

CRediT authorship contribution statement

Boqian Wang: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Yuan Jin:** Supervision, Formal analysis. **Mingda Hu:** Resources, Investigation. **Yunxiang Zhao:** Validation. **Xin Wang:** Funding acquisition, Formal analysis. **Junjie Yue:** Supervision. **Hongguang Ren:** Supervision, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China [grant numbers 32070025, 31800136, and 62102439] and the Research Project of the State Key Laboratory of Pathogen and Biosecurity [grant numbers SKLPBS1807 and SKLPBS2214].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e32103.

References

[1] S.J. Sibbald, L. Eme, J.M. Archibald, A.J. Roger, Lateral gene transfer mechanisms and pan-genomes in eukaryotes, Trends Parasitol. 36 (2020) 927–941.
[2] J. Van Etten, D. Bhattacharya, Horizontal gene transfer in eukaryotes: not if, but how much? Trends Genet. 36 (2020) 915–925.
[3] C.X. Chan, R.G. Beiko, A.E. Darling, M.A. Ragan, Lateral transfer of genes and gene fragments in prokaryotes, Genome biology and evolution 1 (2009) 429–438.
[4] D.P. Martin, B. Murrell, A. Khoosal, B. Muhire, Detecting and analyzing genetic recombination using RDP4, Bioinformatics: Volume I: data, sequence analysis, and evolution (2017) 433–460.
[5] C.X. Chan, R.G. Beiko, M.A. Ragan, Lateral transfer of genes and gene fragments in Staphylococcus extends beyond mobile elements, J. Bacteriol. 193 (2011) 3964–3977.
[6] S. Lytras, J. Hughes, D. Martin, A. de Klerk, R. Lourens, S. Pond, W. Xia, X. Jiang, D.L. Robertson, Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination, 2021. COVID-19 Research.
[7] Y. Wang, J. Zeng, C. Zhang, C. Chen, Z. Qiu, J. Pang, Y. Xu, Z. Dong, Y. Song, W. Liu, New framework for recombination and adaptive evolution analysis with application to the novel coronavirus SARS-CoV-2, Briefings Bioinf. 5 (2021).
[8] D.P. Martin, A. Varsani, P. Roumagnac, G. Botha, S. Maslamoney, T. Schwab, Z. Kelz, V. Kumar, B. Murrell, RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets, Virus Evolution 7 (2021) veaa087.
[9] J. Ren, X. Bai, Y.Y. Lu, K. Tang, Y. Wang, G. Reinert, F. Sun, Alignment-free sequence analysis and applications, Annual Review of Biomedical Data Science 1 (2018) 93–114.
[10] A. Zielezinski, S. Vinga, J. Almeida, W.M. Karlowski, Alignment-free sequence comparison: benefits, applications, and tools, Genome Biol. 18 (2017) 1–17.
[11] S. Sarmashghi, K. Bohmann, M.T.P. Gilbert, V. Bafna, S. Mirarab, Skmer: assembly-free and alignment-free sample identification using genome skims, Genome Biol. 20 (2019) 1–20.
[12] A. Zielezinski, H.Z. Girgis, G. Bernard, C.A. Leimeister, W.M. Karlowski, Benchmarking of alignment-free sequence comparison methods, BioMed Central 20 (2019) 1–8.
[13] M.F. Aziz, G. Caetano-Anollés, Evolution of networks of protein domain organization, Sci. Rep. 11 (2021) 12075.
[14] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L. Sonnhammer, S.C. Tosatto, L. Paladin, S. Raj, L.J. Richardson, Pfam: the protein families database in 2021, Nucleic Acids Res. 49 (2021) D412–D419.
[15] G. Apic, R.B. Russell, Domain recombination: a workhorse for evolutionary innovation, Sci. Signal. 3 (2010) pe30, pe30.
[16] D.H. Parks, M. Chuvochina, P.-A. Chaumeil, C. Rinke, A.J. Mussig, P. Hugenholtz, A complete domain-to-species taxonomy for Bacteria and Archaea, Nat. Biotechnol. 38 (2020) 1079–1086.
[17] I. Sarkar, M. Gtari, L.S. Tisa, A. Sen, A novel phylogenetic tree based on the presence of protein domains in selected actinobacteria, Antonie Leeuwenhoek 112 (2019) 101–107.
[18] D.H. Parks, M. Chuvochina, C. Rinke, A.J. Mussig, P.-A. Chaumeil, P. Hugenholtz, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy, Nucleic Acids Res. 50 (2022) D785–D794.
[19] Z. Yan, Z. Cao, Y. Liu, H.A. Ogilvie, L. Nakhleh, Maximum parsimony inference of phylogenetic networks in the presence of polyploid complexes, Syst. Biol. 71 (2022) 706–720.

[20] J.H. Fong, L.Y. Geer, A.R. Panchenko, S.H. Bryant, Modeling the evolution of protein domain architectures using maximum parsimony, J. Mol. Biol. 366 (2007) 307–315.
[21] R.D. Finn, B. Alex, C. Jody, C. Penelope, R.Y. Eberhardt, S.R. Eddy, H. Andreas, H. Kirstie, H. Liisa, M. Jaina, Pfam: the protein families database, Nucleic Acids Res. 42 (2014) D222–D230.
[22] D.T. Hoang, L.S. Vinh, T. Flouri, A. Stamatakis, A. von Haeseler, B.Q. Minh, MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation, BMC Evol. Biol. 18 (2018) 1–11.
[23] K. Tamura, G. Stecher, S. Kumar, MEGA11: molecular evolutionary genetics analysis version 11, Mol. Biol. Evol. 38 (2021) 3022–3027.
[24] M.T. Caudill, K.A. Brayton, The use and limitations of the 16S rRNA sequence for species classification of Anaplasma samples, Microorganisms 10 (2022) 605.
[25] N.S. Muhamad Rizal, H-m Neoh, R. Ramli, P.R. A/LK Periyasamy, A. Hanafiah, M.N. Abdul Samat, T.L. Tan, K.K. Wong, S. Nathan, S. Chieng, Advantages and limitations of 16S rRNA next-generation sequencing for pathogen identification in the diagnostic microbiology laboratory: perspectives from a middle-income country, Diagnostics 10 (2020) 816.
[26] F. Schulz, E.A. Eloe-Fadrosh, R.M. Bowers, J. Jarett, T. Nielsen, N.N. Ivanova, N.C. Kyrpides, T. Woyke, Towards a balanced view of the bacterial tree of life, Microbiome 5 (2017) 1–6.
[27] D.H. Parks, M. Chuvochina, D.W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life, Nat. Biotechnol. 36 (2018) 996–1004.
[28] P.-A. Chaumeil, A.J. Mussig, P. Hugenholtz, D.H. Parks, GTDB-tk: a Toolkit to Classify Genomes with the Genome Taxonomy Database, Oxford University Press, 2020.
[29] B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. Von Haeseler, R. Lanfear, IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era, Mol. Biol. Evol. 37 (2020) 1530–1534.
[30] A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (2014) 1312–1313.
[31] A. Rambaut, FigTree, Tree Figure Drawing Tool (2009).
[32] I. Letunic, P. Bork, Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation, Nucleic Acids Res. 49 (2021) W293–W296.
[33] O. Nyholm, T. Lienemann, J. Halkilahti, S. Mero, R. Rimhanen-Finne, V. Lehtinen, S. Salmenlinna, A. Siitonen, Characterization of Shigella sonnei isolate carrying Shiga toxin 2–producing gene, Emerg. Infect. Dis. 21 (2015) 891–892.
[34] R. Lan, P.R. Reeves, Escherichia coli in disguise: molecular origins of Shigella, Microb. Infect. 4 (2002) 1125–1132.
[35] M. Riley, B. Labedan, Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module, J. Mol. Biol. 268 (1997) 857–868.
[36] L. Ivica, B. Peer, Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation, Nucleic Acids Res. 49 (2021) 293–296.
[37] J. Bergan, A.B.D. Lingelem, R. Simm, T. Skotland, K. Sandvig, Shiga toxins, Toxicon 60 (2012) 1085–1107.
[38] J. Rodríguez-Beltrán, J. DelaFuente, R. Leon-Sampedro, R.C. MacLean, A. San Millan, Beyond horizontal gene transfer: the role of plasmids in bacterial evolution, Nat. Rev. Microbiol. 19 (2021) 347–359.
[39] J. Botelho, H. Schulenburg, The role of integrative and conjugative elements in antibiotic resistance evolution, Trends Microbiol. 29 (2021) 8–18.
[40] Y. Wang, C. Xu, R. Zhang, Y. Chen, Y. Shen, F. Hu, D. Liu, J. Lu, Y. Guo, X. Xia, Changes in colistin resistance and mcr-1 abundance in Escherichia coli of animal and human origins following the ban of colistin-positive additives in China: an epidemiological comparative study, Lancet Infect. Dis. 20 (2020) 1161–1171.
[41] M. Acman, R. Wang, L. van Dorp, L.P. Shaw, Q. Wang, N. Luhmann, Y. Yin, S. Sun, H. Chen, H. Wang, Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene bla NDM, Nat. Commun. 13 (2022) 1131.
[42] B. van Dijk, F. Bertels, L. Stolk, N. Takeuchi, P.B. Rainey, Transposable elements promote the evolution of genome streamlining, Philosophical Transactions of the Royal Society B 377 (2022) 20200477.
[43] M.A. Brockhurst, E. Harrison, Ecological and evolutionary solutions to the plasmid paradox, Trends Microbiol. 30 (2022) 534–543.
[44] R. Hershberg, H. Tang, D.A. Petrov, Reduced selection leads to accelerated gene loss in Shigella, Genome Biol. 8 (2007) 1–11.