MEGI: a comprehensive annotation dataset of mobile elements for genomic island detection

1. Introduction

Genomic islands are formed through horizontal gene transfer and typically contain functional genes, such as those related to pathogenicity, metabolic regulation, and antibiotic resistance, as well as mobile elements like insertion sequences (IS), integrative and conjugative elements (ICE), mobile genetic elements (MGE), and bacteriophages. In turn, utilizing these mobile elements can aid in the identification of genomic islands. We proposed a comprehensive dataset (MEGI) of mobile elements for genomic island detection. There are 1,103 genomic islands from 153 genomes, which were annotated with mobile genetic elements including the insertion sequences (IS), integrative and conjugative elements (ICE), transposons (MGE) and bacteriophages (phage).

1. Genomic island datasets

We compiled four datasets: GIs/non-GIs dataset, *S. enterica Typhi* CT18 dataset, L-data dataset, and IslandViewer.

1. GIs/non-GIs

GIs/non-GIs dataset is one of the standard datasets used for predicting genomic islands (GIs) and is commonly used to validate the efficiency of genomic island identification algorithms. This dataset includes 771 GIs and 3770 non-GIs segments ranging from 8 to 30 kb, compiled from 118 prokaryotic genomes [1].

1. *S. enterica Typhi* CT18 Dataset:

This dataset is derived from the complete genome of Salmonella Typhi CT18. The literature [2-5] has identified 21 large genomic islands in this genome, which have been experimentally confirmed to exist, making this dataset suitable for subsequent validation experiments.

1. L-data Dataset:

Through whole-genome comparison, Wei et al. [6] collected and organized information on 11 complete genomes from published literature and annotated genomic islands within these genomes. The 14 sequence IDs are NC\_002505.1, NC\_002506.1, NC\_002935.2, NC\_003198.1, NC\_004431.1, NC\_009480.1, NC\_009778.1, NC\_010161.1, NC\_010170.1, NC\_010554.1, NC\_011000.1, NC\_011001.1, NC\_011002.1, and NC\_012471.1. There are total 114 GIs from the above genomes.

1. IslandViewer

IslandViewer provides visualization tools for genomic islands and includes the comparative genomics-based IslandPick, as well as the Hidden Markov Model-based methods SIGI-HMM [7] and IslandPath-DIMOB [8]. From 3000 genomes, we selected sequences that met the following criteria: 1) the length of GIs identified by each prediction method must account for at least 40% of the total sequence length; 2) at least two of the three prediction methods must predict the genomic islands, and the predicted segments must be larger than 20 kb. The 20 sequences that met these criteria comprised the dataset used to validate the algorithm's efficiency. Since the true positions of genomic islands for 20 genomes from the IslandViewer database were not provided, we used the IslandPath method to identify 195 genomic islands.

1. Mobile element datasets

We downloaded the FASTA sequences for each genome from the GenBank database and manually organized the corresponding mobile elements from databases such as ISFinder, ICEfinder, INTEGRALL, GyDB [9], and PHAST [10]. We identified the presence of insertion sequences (IS), integrative and conjugative elements (ICE), integrons, transposons (MGE), and bacteriophages (phage) within the genomes. From the four datasets mentioned above, we identified a total of 165,986 ISs, 108 ICEs, 942 MGEs, and 908 phages.

Reference

[1] Xu Z, Hao BL. CVTree updata: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res, 2009, 37: W174-W178.

[2] Aaron JA, Rajeev K, Azad AR, et al. Detection of genomic islands via segmental genome heterogeneity. Nucleic Acids Res, 2009, 37: 5255-5266.

[3] Vernikos GS, Parkhill J. Interpolated variable order motifs for identiﬁcation of horizontally acquired DNA: revisiting the salmonella pathogenicity islands. Bioinformatics, 2006, 22: 2196–203.

[4] Kingsley RA, van AK, Kramer N, et al. The shdA gene is restricted to serotypes of Salmonella enterica subspecies I and contributes to efficient and prolonged fecal shedding. Infect Immun, 2000, 68: 2720-2727.

[5] Kingsley RA, Humphries AD, Weening EH, et al. Molecular and phenotypic analysis of the CS54 island of Salmonella enterica serotype Typhimurium: identification of intestinal colonization and persistence determinants. Infect Immun 2003, 71: 629-640.

[6] Garcia-Vallve,S., Janssen,P.J. and Ouzounis,C.A. Genetic variation between Helicobacter pylori strains: gene acquisition or loss？Trends Microbiol, 2002, 10: 445-447.

[7] Merkl R. SIGI: score-based identification of genomic islands. BMC Bioinformatics, 2014, 5: 22.

[8] Hsiao W, Wan I, Jones S J, et al. IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics, 2003, 19(3): 418-420.

[9] Carlos L, Ricardo F, Laura C, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research, 2011, 39: D70-D74.

[10] You Z, Liang YJ, *et al.* PHAST: a fast phage search tool. Nucleic Acids Research, 2011, 39: W347-W352.