BMI3_Mini_project_proposal_Group4.pdf

by 江睿颖 (JIANG Ruiying) ZJU-UOE Institute

Submission date: 02-Nov-2025 03:48PM (UTC+0800)
Submission ID: 2800300373
File name: BMI3_Mini_project_proposal_Group4.pdf (94.16K)
Word count: 1475
Character count: 9030

A Hidden Markov Model Framework for Detecting Horizontally Transferred Genes in *Legionella pneumophila*

Group 4

***Abstract*—Horizontal Gene Transfer (HGT) is a pivotal mechanism driving microbial evolution and pathogenesis, enabling pathogens like *Legionella pneumophila* to rapidly acquire virulence traits. However, accurately identifying horizontally acquired genes, especially those partially assimilated into the host genome, remains a significant computational challenge. This project proposes a three-state Hidden Markov Model (HMM) to detect foreign genes in *Legionella* by integrating both sequence composition and lightweight contextual biological features. Our model distinguishes between 'Host-like,' 'Foreign-like,' and an intermediate 'Ameliorated' state, explicitly modeling gene contiguity to identify genomic islands. The model will be rigorously validated on simulated datasets and then applied to the *Legionella* genome to produce a high-confidence catalog of candidate genes. The key deliverables are an original, open-source HGT detection tool and a curated list of foreign genes in *Legionella*, providing a valuable resource for understanding pathogen evolution.**

I. Survey of Background Literature

A. Overview

HORIZONTAL gene transfer (HGT) is a major evolutionary force enabling bacteria to rapidly acquire new traits such as antibiotic resistance and virulence. *Legionella pneumophila*, the causative agent of Legionnaires' disease, has extensively integrated genes from environmental amoebae, evolving into a proficient intracellular pathogen. Detecting these transferred genes is challenging because foreign sequences gradually assimilate, eroding their distinctive signatures. This project aims to develop an HMM-based bioinformatics tool to identify both recent and partially assimilated foreign genes in *L. pneumophila* by modeling genomic composition and contextual signals. Expected outcomes include an open-source algorithm and a curated list of HGT-derived genes, contributing to a deeper understanding of pathogen evolution.

B. Background

Unlike vertical inheritance, HGT involves the direct exchange of genetic material between organisms, serving as an engine for rapid microbial adaptation. Given the close associations with multiple microorganisms and eukaryotic hosts, *legionella* serves as a perfect model for studying both inter and intra HGT. Moreover, the identification of which genes have been acquired horizontally is of great importance considering *legionella*'s formidable pathogenicity. Standard methods like compositional and phylogenetic analysis can be imprecise or computationally intensive. We introduce a novel framework using Hidden Markov Models (HMMs), which can statistically distinguish 'native' from 'foreign' DNA and efficiently pinpoint integrated regions, providing a unified method for mapping HGT in *Legionella*.

C. Relevance/Impact

This research aims to identify HGT-derived genes in *Legionella* that contribute to virulence and adaptation, informing the validation of new drug targets and antivirulence therapies. Beyond a single pathogen, this HMM framework provides a scalable tool for identifying acquired virulence arsenals in diverse multidrug-resistant pathogens, representing a critical first step towards designing next-generation therapeutics.

II. Proposed Methodology

A. Algorithmic overview

We will detect horizontally transferred genes using a three-state HMM. The latent states represent Host-like, Ameliorated (partially assimilated), and Foreign-like regimes. The HMM models gene contiguity to recover foreign "islands" and combines intrinsic sequence composition with lightweight contextual cues.

B. Feature representation

Each coding gne is embedded by two complementary channels. The composition channel comprises GC/GC3, codon usage (RSCU), and amino acid usage, standardized and, if needed, modestly reduced in dimension for numerical stability. The context channel encodes sparse indicators such as proximity to mobility related annotations (e.g., integrase/transposase), local GC shift around gene boundaries, replicon type where available, and simple motif/domain flags associated with eukaryotic like effectors (e.g., ankyrin/F box/U box/SEL1). This yields orthogonal evidence without relying on exhaustive annotations.

C. Model architecture

We adopt a three state HMM with a dual channel, evidence weighted emission. The composition channel is modeled by a

multivariate Gaussian; the context channel modulates likelihood via a low dimensional Bernoulli component, combined as a product of experts with a scalar weight to balance their influence. Transition probabilities favor self transitions to encode island contiguity and prefer Host ↔ Ameliorated ↔ Foreign moves over direct Host ↔ Foreign jumps, reflecting gradual assimilation. An expected island length provides a weak prior on contiguity strength.

D. Training and inference

Parameters will be estimated by Baum–Welch with simple clustering based initialization; covariances will be lightly regularized to improve robustness under limited sample sizes. Viterbi decoding will produce a genome segmentation into the three states and delineate candidate island boundaries, while posterior probabilities will provide per gene scores for Foreign like and Ameliorated assignments. Core HMM routines (forward–backward, EM, Viterbi) will be implemented from scratch to satisfy the ICA requirement for original algorithmic development; standard utilities may be used for preprocessing.

E. Evaluation plan

We will tune parameters on simulated datasets with inserted foreign islands at varying evolutionary distances. The final model will be applied to *Legionella*, with predictions validated via GO/KEGG enrichment and domain screens.

F. Rationale and alternatives

An HMM is preferred over independent classifiers and Gaussian mixtures because horizontal transfers in *Legionella* frequently occur in contiguous blocks; transitions explicitly capture this structure and improve boundary localization. A three state design is favored over binary partitioning because partial assimilation is common and otherwise confounds native–foreign separation; an explicit intermediate state yields more stable and interpretable assignments. Incorporating a light context channel addresses known failure modes of composition only methods by adding mechanistic evidence linked to mobility and eukaryotic like effectors, while remaining annotation light and broadly applicable. Compared with GC thresholds, GMM or purely supervised classifiers, the proposed formulation bett aligns with our goals: principled detection of foreign islands, robustness across donor distances and assimilation levels, and clear biological interpretability.

III. Research Plan

Stage 1. Data Preparation Feature Engineering (10.30 - 11.3): Acquire genomes and annotations. Calculate compositional and contextual features.

Stage 2. Model Implementation (11.3 - 11.17): Build the 3-state HMM. Code dual-channel emission probability. Implement Baum-Welch and Viterbi algorithms.

Stage 3. Validation Tuning (11.18 - 11.24): Create simulated genomes. Benchmark performance (sensitivity, precision). Tune hyperparameters.

Stage 4. Biological Discovery Analysis (11.28 - 12.1): Apply the model to the real *Legionella* genome. Perform functional enrichment analysis (GO/KEGG).

Stage 5. Project Finalization (12.2 - 12.8): Prepare the final report and presentation. Finalize code and documentation for submission.

IV. Resources

A. Data Resources

The host genome is *Legionella pneumophila* strain Philadelphia 1 (NC_002942.5), with complete genome sequences and annotations downloaded from NCBI. Conserved single-copy orthologs, such as ribosomal proteins and RNA polymerase subunits, will be screened using eggNOG/COG to form a host core dataset, reducing interference from potential HGT genes during training.

Foreign genes are divided into three evolutionary distance groups relative to the host. Near (γ-Proteobacteria) includes *Escherichia coli* K-12 MG1655 (NC_000913.3), *Pseudomonas aeruginosa* PAO1 (NC_002516.2), and *Vibrio cholerae* N16961 (NC_002505.1). Moderate distance includes *Agrobacterium tumefaciens* C58 (NC_003062.2), *Bacillus subtilis* 168 (NC_000964.3), and *Bacteroides thetaiotaomicron* VPI-5482 (NC_004663.1). Distant taxa include archaea, higher eukaryotes, and bacteria with extreme GC content, such as *Methanobacterium formicicum* DSM1535 (NC_014409.1), *Arabidopsis thaliana* Col-0 (NC_003070.9), *Streptomyces coelicolor* A3(2) (NC_003888.3), and *Clostridium perfringens* 13 (NC_003366.1).

Each foreign dataset is split into two parts: one for model training and one for insertion into the host genome to construct simulated datasets for model tuning, with a training-to-simulation ratio of approximately 3:1.

The final model will be applied to the unannotated host genome and the related strain *Legionella longbeachae* NSW150 (NC_013861.1) for real-world application and generalization testing.

B. Computational Resources

Model training and feature extraction are implemented in Python using Biopython, NumPy, pandas, and other basic libraries. Core HMM algorithms, including Baum–Welch and Viterbi, will be independently implemented to ensure algorithmic originality.

C. Reference Databases

NCBI RefSeq, eggNOG/COG for gene selection and annotation; GO/KEGG for functional enrichment analysis and validation of results.

V. References

Gogarten, J.P. and Townsend, J.P. (2005) "Horizontal gene transfer, genome innovation and evolution," Nature Reviews Microbiology, 3(9), pp. 679–687. Available at: https://doi.org/10.1038/nrmicro1204.

Gomez-Valero, L. and Buchrieser, C. (2019) "Intracellular parasitism, the driving force of evolution of Legionella pneumophila and the genus Legionella," Genes and Immunity, 20(5), pp. 394–402. Available at: https://doi.org/10.1038/s41435-019-0074-z.

Juhas, M. et al. (2009) "Genomic islands: tools of bacterial horizontal gene transfer and evolution," FEMS microbiology reviews, 33(2), pp. 376–393. Available at: https://doi.org/10.1111/j.1574-6976.2008.00136.x.

Murray, C.J.L. et al. (2022) "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis," The Lancet, 399(10325), pp. 629–655. Available at: https://doi.org/10.1016/S0140-6736(21)02724-0.

[NOTE: be brief, strictly no more than 2 pages.]

12

BMI3_Mini_project_proposal_Group4.pdf

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

This is a very good proposal. It follows the template (except for the number of pages).

You understand the general question and are considering the expected elements of the ICA: a new algorithm, benchmark and some kind of optimisation but more can be considered.

The preferred moves from Host to Ameliorated to Foreign is a very good addition over a simplistic model of any state to any other state transition. This may be considered as an optimisation if finally implemented. If not, consider any other option for true time or space optimisation (dimensionality reduction?).

You have added very important and useful elements for the detection of HGT events: proximity to mobility.

Nevertheless, consider that is it better to have a very simple algorithm that is functional (even if not accurate at predicting host/foreign/intermediate) than one that did not run or nothing is concluded because it was too difficult to implement. I suggest you make sure an initial version works that considers something as simple as GC content, and from that, you add more and more features.

Part of your research plan is how the work will be split into the 5 group members.

Regarding the dataset again, you do not need to make it too complicated. The core of the ICA is the algorithm and the benchmarking. Not the dataset. Have a very well classified dataset that can give you values of accuracy but you probably do not need entire genomes and plenty of them.

For the Biological Discovery analysis consider: Gomez-Valero, 2011, Table 2.

Your text lacks references (cited within the text).

Pay attention to the instructions, including the technical ones like the maximum number of pages.

PAGE 1

Comment 1

it could also detect contamination during sample preparation and sequencing.

Comment 2

sequences? loci?

Comment 3

unclear what do you mean

Comment 4

your abstract does not clarify that you will look for Eukaryote-to-procaryote transfers. So do not mix things.

Keep it either "foreign-not foreign" or Eukaryotic vs prokaryotic. Either is fine.

Comment 5

correct, if detected by annotation they would be orthogonal to the HMM algorithm. So use this orthogonality for benchmarking rather than as an *if* in the detection.

Comment 6

unclear: do you mean exchange of genes between Legionella bacteria? Even if this happens, it is probably out of the scope of your tool.

Comment 7

Reference missing

PAGE 2

Comment 8

this is unclear. Check the reference for already-annotated HGT events.

Comment 9

This sounds very good if you manage to implement it. Again, consider a simplified dataset: just some genes (from these different evolutionary distances) randomly inputed into the host. I do not see why you would need to use entire genomes

Comment 10

you can use the proteins in table 2 of this
reference: https://link.springer.com/article/10.1186/1471-2164-12-536

Comment 11

abbreviation not defined

PAGE 3

Comment 12

Which you did not do