Shortest Unique Representative Hidden Markov Model (SurHMM) for Detecting Protein Toxins, Virulence Factors and Antibiotic Resistance Genes

Gary Xie (xie@lanl.gov)
Los Alamos National Laboratory https://orcid.org/0000-0002-9176-924X
Jeanne M Fair
Los Alamos National Research Laboratory: Los Alamos National Laboratory

Research note

Keywords: potential toxin, VF, AR fragments, NGS, VFs, ARs

Posted Date: February 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-185430/v1

Version of Record: A version of this preprint was published at BMC Research Notes on March 30th, 2021.
See the published version at https://doi.org/10.1186/s13104-021-05531-w.

Abstract

Objective: Currently, next generation sequencing (NGS) is widely used to decode potential novel or variant pathogens both in emergent outbreaks and in routine clinical practice. However, the efficient identification of novel or diverged pathogenomic compositions remains a big challenge. It is especially true for short DNA sequence fragments from NGS, since sequence similarity searching is vulnerable to false negatives or false positives, as mismatching or matching with unrelated proteins. Therefore, this study aimed to establish a bioinformatics approach that can generate unique motif sequences for profiling searching, resulting in high specificity and sensitivity.

Results: In this study, we introduced a shortest unique representative hidden Markov model (HMM) approach to identify bacterial toxin, virulence factor (VF), and antimicrobial resistance (AR) in short sequence reads. We first construct unique representative domain sequences of toxin genes, VFs, and ARs to avoid potential false positives, and then to use HMM models to accurately identify potential toxin, VF, and AR fragments. The benchmark shows this approach can achieve relatively high specificity and sensitivity if the appropriate cutoff value is applied.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Tables

Table 1. Summary of SurHMM generated in this study

Type	Profiles
Neurotoxin	61
Shiga_toxin	10
Choliz_toxin	8
Clostridium_perfringens toxin	21
Staphylococcal_toxin	74
Total virulence factors	86136
Total antimicrobial resistance	3237

Figures

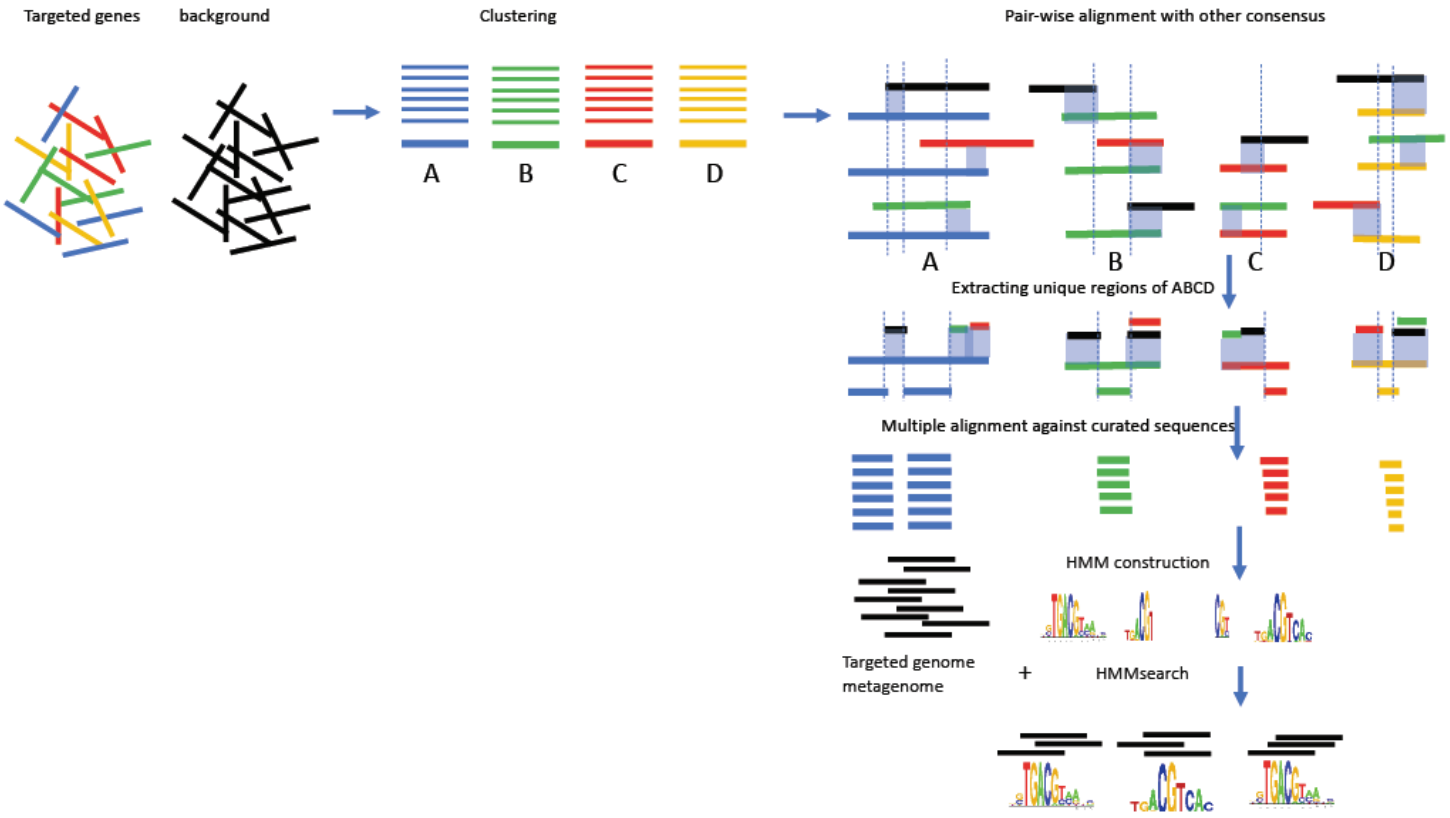

Figure 1

SurHMM approach creates Shortest Unique Representative Hidden Markov Model (SurHMM) for protein families of interest first, then identifies markers in targeted genomes and metagenomes by scanning predicted open reading frames or six-frame translation of given nucleotide reads. Drawing inspired from [7].