NSF Project III: Small: Reconstructing viral population without using a reference genome

Federal Agency and Organization Element to Which Report is Submitted: **4900**

Federal Grant or Other Identifying Number Assigned by Agency: **1724008**

Project Title: **III: Small: Reconstructing viral population without using a reference genome**

PD/PI Name: **Raj Acharya, Principal Investigator**

Recipient Organization: **Indiana University**

Project/Grant Period: **07/01/2016 - 08/31/2018**

Major goals of the project:

We propose to build a software system that will use probabilistic de Bruijn graphs for reconstruction of viral haplotypes without a reference genome. The haplotypes are assumed to arise through mutation and/or recombination processes. In the software system, when provided with a set of reads, the software developed should determine the number, sequences, and relative frequencies of the haplotypes.

Summary:

We have proposed MLEHaplo, a maximum likelihood *de novo* assembly algorithm for viral haplotypes using paired-end NGS data. The paired reads are represented in a De Bruijn graph and the pairing information of the reads is stored as pairs of k-mers. The support found in the pairs of vertices is used to score *source-sink* paths, and the scoring mechanism ensures that the paths represent possible viral haplotypes. As reconstructing a minimal cover of the graph under paired constraints is NP-hard, we have proposed a polynomial time heuristic algorithm, ViPRA, that recovers a small fraction of the total number of paths in the graph. MLEHaplo then reconstructs a maximum likelihood estimate of the viral population using the paths recovered by ViPRA based on a generative model for sampling paired reads from the viral population. MLEHaplo takes about a day and half of running time on a single core machine to process half a million reads, and was comparable in run-time to other methods.

We have tested ViPRA and MLEHaplo on simulated viral populations that consist of haplotypes arising from common ancestors based on a coalescent model. These simulations are more realistic models of viral populations rather than introducing random mutations uniformly across a known virus. MLEHaplo predicts the smallest set of viral haplotypes and has the highest recall for the true haplotypes when compared to three existing reference based reconstruction methods.

We also evaluated ViPRA and MLEHaplo on reads simulated from HCV E1/E2 genes identified in an infected patient and on a dataset of five HIV-1 strains which has been used to test haplotype reconstruction methods. MLEHaplo reconstructs haplotypes with greater than 99% sequence identity

to the true haplotypes for HCV E1/E2 genes and greater than 97% sequence identity for the five HIV-1 strains. The decrease in sequence identity in the case of real HIV-1 strains is understandable as the reads are generated from replicating viruses where one would observe additional sequence variation with respect to the reference. This has been documented as single nucleotide polymorphisms with respect to the reference in the original study and as an increase in sequence alignment score in a previous method. Nevertheless, the haplotypes reconstructed by MLEHaplo better explained the observed error corrected sequenced data than the consensus sequences of the known strains, suggesting they form a better representation of the sequenced data.

(Material from manuscript in review: IEEE TCBB and published article in Computational and Structural Biotechnology Journal)

Significant Results:

We have proposed a classifier MultiRes for detecting rare variant and erroneous *k*-mers obtained from Illumina sequencing of viral populations. Our method does not rely on a reference sequence and uses concepts from signal processing to justify using the counts of sets of *k*-mers of different sizes. We utilize the projections of sampled reads signals onto multiple frames as features for our classifier MultiRes.

We demonstrated the performance of MultiRes on simulated HIV and HCV viral populations and real HIV viral populations containing viral haplotypes at varying relative frequencies, where it outperformed the error detection algorithms used in error correction methods in terms of recall and the total number of predicted *k*-mers. Though, the error detection algorithms in the error correction methods evaluated assumed that sequenced reads originated from a single genome sequenced at uniform coverage, our method also works better than the method BayesHammer, which can tackle non-uniform sequencing coverage, and the method Seecer, which additionally incorporates methods for detecting alternative splicing and polymorphisms. (From article published in Computational and Structural Biology Journal)

We have proposed MLEHaplo, a maximum likelihood *de novo* assembly algorithm for viral haplotypes that uses paired-end NGS data to reconstruct the viral population. Reads are represented in a De Bruijn graph and their pairing information is stored as pairs of k-mers in the set PS. The support found in the set PS for pairs of vertices is used to score *source-sink* paths, and the scoring mechanism ensures that the paths represent possible viral haplotypes. As reconstructing a minimal cover of the graph under paired constraints is NP-hard, we have proposed a polynomial time heuristic algorithm, ViPRA, that recovers a small fraction of the total number of paths in the graph. MLEHaplo then reconstructs a maximum likelihood estimate of the viral population using the paths recovered by ViPRA based on a generative model for sampling paired reads from the viral population.

We have tested ViPRA and MLEHaplo on simulated viral populations that consist of haplotypes arising from common ancestors based on a coalescent model. These simulations are more realistic models of viral populations rather than introducing random mutations uniformly across a known virus.

MLEHaplo predicts the smallest set of viral haplotypes and has the highest recall for the true haplotypes when compared to three existing reference based reconstruction methods.

An advantage of MLEHaplo and ViPRA over the existing methods is that it retains the correct phylogeny of the true viral haplotypes, even when the reconstructed paths are not an exact match to the true viral haplotypes. When MLEHaplo over-estimates the number of haplotypes, they can be clustered together to further reduce their number by using phylogentic trees. Thus, the viral diversity can be correctly inferred from the reconstructed paths.

We also evaluated ViPRA and MLEHaplo on reads simulated from HCV E1/E2 genes identified in an infected patient and on a dataset of five HIV-1 strains which has been used to test haplotype reconstruction methods. MLEHaplo reconstructs haplotypes with greater than 99% sequence identity to the true haplotypes for HCV E1/E2 genes and greater than 97% sequence identity for the five HIV-1 strains. The decrease in sequence identity in the case of real HIV-1 strains is understandable as the reads are generated from replicating viruses where one would observe additional sequence variation with respect to the reference.

The usage of De Bruijn graphs for viral haplotype reconstruction has a number of advantages: The k-mers as vertices avoids the costly computations of reads overlaps. Also, because the De Bruijn graph construction is *de novo*, it contains all the variation observed in the viral population which can be lost by aligning reads to a reference genome. The current method relies on generating an acyclic De Bruijn graph for estimating the viral haploytpes. Choosing k greater than D, the size of the largest repeat in the viral genome, ensures that the De Bruijn graph obtained is acyclic. As repeats in RNA viruses are generally short, acyclic De Bruijn graphs can be readily obtained for viral populations. For HIV-1 strains and HCV strains a k-mer size of 60 allowed generation of acyclic De Bruijn graphs. For larger repeats, the algorithm can be potentially applied to reconstruct segments of the genome excluding the repeat sequences. (From article in review IEEE TCBB).

We evaluated the relative performance of MLEHaplo to those obtained from the softwares ShoRAH, QuasiRecomb and PredictHaplo on the simulated datasets, using the consensus sequence obtained by the de novo assembler Vicuna. ShoRAH and QuasiRecomb over-estimate the number of predicted haplotypes, while PredictHaplo under-estimates the number of viral haplotypes in eight out of ten datasets. None of these methods have 100% recall for any of the ten data sets and all fail to recover any true haplotypes in at least one of the datasets. In contrast, MLEHaplo predicts 6-10 haplotypes for all the ten datasets and correctly reconstructs 6-7 out of seven true haplotypes in all the ten datasets. Thus, MLEHaplo retains the correct sequences of the haplotypes while accurately predicting the number of haplotypes in each viral population.

On the HCV dataset, all methods except PredictHaplo overestimated the number of strains in the viral population by generating more than 10 paths from the read graph or the De Bruijn graph. The number of distinct paths predicted by MLEHaplo and ShoRAH were similar, while QuasiRecomb overestimated the distinct paths in the HCVU dataset. MLEHaplo generated a small number of distinct paths in both the HCV-U and HCV-P datasets. The distribution of all pairwise distances for the predicted paths indicates that while MLEHaplo, PredictHaplo, and ShoRAH have similar distributions to the true HCV strains, all of the paths generated by QuasiRecomb have small pairwise distances and the distribution of pairwise distances differs from that of the true HCV strains, which leads to small number of distinct paths in QuasiRecomb.

(Material from manuscript in review: IEEE TCBB and published article in Computational and Structural Biotechnology Journal)

Papers:

1. Malhotra, R., Wu, S., Jha, M., Rodrigo, A., Poss, M., & Acharya, R. Maximum Likelihood de novo reconstruction of viral populations using paired end sequencing data, In Review: *IEEE/ACM transactions on computational biology and bioinformatics.*

2. Malhotra, R., Mukhopadhyay, M., Poss, M., & Acharya, R. (2016). A frame-based representation of genomic sequences for removing errors and rare variant detection in NGS data. *arXiv preprint arXiv:1604.04803*.

3. Malhotra, R., Jha, M., Poss, M., & Acharya, R. (2017). A random forest classifier for detecting rare variants in NGS data from viral populations. *Computational and structural biotechnology journal*, *15*, 388-395.

Links to software:

1. MLEHaplo and ViPRA: https://github.com/raunaq-m/MLEHaplo
2. MultiRes: https://github.com/raunaq-m/MultiRes

Generating Data:

1. The simulated data is generated using the dwgsim tool available at https://github.com/nh13/DWGSIM

2. The HCV data is generated with the tool simseq: Earl D, Bradnam K, John JS, Darling A, Lin D, Fass J, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome research. 2011;21(12):2224–2241.

3. The HIV dataset used is available with the SRA number: SRR961514