Reconstruction of Influenza A Virus Variants from PacBio Reads

Alexander Artyomenko*, Nicholas C. Wu‡, Serghei Mangul†, Eleazar Eskin†, Ren Sun‡, and Alex Zelikovsky*
*Department of Computer Science, Georgia State University
Atlanta, GA 30302-3994,
email:{aartyomenko, alexz}@cs.gsu.edu
†Computer Science Department, ‡Molecular and Medical Pharmacology, University of California, Los Angeles
Los Angeles, CA 90095
email: {serghei, eeskin}@cs.ucla.edu, wchnicholas@ucla.edu, rsun@mednet.ucla.edu

***Abstract*—Pacific Biosciences (PacBio) sequencing is providing thousands of reads with the length up to 10,000 bases. In most cases this length is enough to cover entire region of interest however this technology has high ($\approx 15\%$) error rate. We propose a method for viral haplotype reconstruction generalizes k-means clustering with Hamming distance and capable of handling up to 25% random errors. When applied to PacBio reads from an Influenza A Virus (IAV) sample with ten variants, our method was able to reconstruct the four most frequent.**

***Keywords*—PacBio, viral quasispecies, clustering.**

I. Introduction

PacBio reads cover a viral genome region of length up to 10,000bp [2]. In this paper we are dealing with the following problem:

Maximum Liklehood Haplotyping of PacBio Reads. Given a set of PacBio reads R emitted by haplotype population, find a set of haplotypes H maximizing $\Pr(R|H)$.

This problem has been succesfully solved by kGEM [1] for 454 HCV amplicon reads. Unfortunately, the original kGEM doesnt work for the PacBio reads due to long insertions and gaps. In this paper we propose a modified version of kGEM applicable to PacBio reads.

II. Clustering Method

PacBio reads were aligned to the reference using the tool InDelFixer [3]. Than mutliple sequence alignment is applied to the aligned reads. Haplotypes found by kGEM represents initial cluster centers [1] run on all reads. The set of clusters is repeatedly expanded with the following procedure:

(1) Pick a read r maximizing Hamming distance to the closest cluster center.
(2) Find the set of reads S which are closer to r than to any cluster center.
(3) Get new cluster centers by running kGEM.

This expansion is repeated until the set S becomes sufficiently small.

III. Results

Sequencing Experiments. Error-prone PCR was performed on the influenza A virus (A/WSN/33) PB2 segment using GeneMorph II Random Mutagenesis Kits (Agilent Technologies, Westlake Village, CA) according to manufacturer's instruction. For the first experiment, a single clone was amplified. For the second experiment, 10 independent clones, ranging from 1 to 13 mutations, were selected. These 10 clones were mixed at a geometric ratio with two-fold difference in occurrence frequency for consecutive clones.

The 2kb region was generated from the viral population and subjected to PacBio RS II sequencing using 2 SMRT cells with P4-C2. The average read length was 1973b and ranges from 200 to 5k. In the first experiment there were 11907 reads and in the second experiment there were 33558 reads.

The single clone experiment. The average Hamming distance between the recovered haplotype and reads is 14.4%. The modified kGEM has been applied for reads. The result of this run perfectly matches the original clone.

The multiple clones experiment. We ran modified kGEM on reads obtained from 10 IAV clones. Our method reported 6 haplotypes: the 4 most frequent haplotypes exactly match 4 most frequent clones and 2 least frequent haplotypes do not exactly match any clone. The correlation between the estemated and true frequencies of the 4 correctly reconstructed haplotypes is 99.4% .

Acknowledgments

We would like to thank H. Hao for performing the PacBio sequencing at Johns Hopkins Deep Sequencing & Microarray Core Facility. S.M. and E.E is supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448 and 1320589, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-MH101782 and R01-ES022282. S.M. was supported in part by Institute for Quantitative & Computational Biosciences Fellowship, UCLA. N.C.W was supported by UCLA Molecular Biology Whitcome Pre-Doctoral Fellowship.

References

[1] Alexander Artyomenko, Nicholas Mancuso, Alex Zelikovsky, Pavel Skums, and Ion Mandoiu. kgem: An em-based algorithm for local reconstruction of viral quasispecies. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on*, pages 1–1. IEEE, 2013.

[2] Alice McCarthy. Third generation dna sequencing: pacific biosciences' single molecule real time technology. *Chemistry & biology*, 17(7):675–676, 2010.

[3] Armin Töpfer. Indelfixer. http://www.bsse.ethz.ch/cbg/software/InDelFixer.

978-1-4673-1321-6/12/$31.00