kGEM: An Expectation Maximization Error Correction Algorithm for Next Generation Sequencing of Amplicon-based Data

Alexander Artyomenko*[1], Nicholas Mancuso*[1], Pavel Skums[2],
Ion Măndoiu*[3], and Alex Zelikovsky*[1]

[1] *Department of Computer Science, Georgia State University, Atlanta, Georgia 30302-3994,*

email: {aartyomenko, nmancuso, alexzz}@cs.gsu.edu

[2] *Centers for Disease Control and Prevention, 1600 Clifton Road NE, Atlanta, Georgia 30333,*

email: kki8@cdc.gov

[3] *Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269,*

email: ion@engr.uconn.edu

Introduction & Methods

RNA-based viruses exist as a heterogeneous collection of closely related variants which are referred to as a *quasispecies*. Next-generation sequencing technologies produce an unprecedented number of viral sequences, but are prone to various types of errors. The main challenge in *local* viral quasispecies reconstruction is to eliminate sequencing errors while preserving the natural heterogeneity of the viral population. This paper presents a new approach to error correction via an expectation maximization (EM) method. We briefly describe the problem and its application to quasispecies reconstruction along with preliminary results.

Given a set R of single amplicon reads emitted by haplotype population P, find a set $H^k = \{H_1, ..., H_k\}$ of k distinct haplotypes that maximizes $\Pr(R|H^k)$ which is a conditional probability of observing reads R given H^k. The estimation of the number k of distinct haplotypes can use a Bayesian approach based on a non-parametric prior known as the Dirichlet process (see [4]). The new method incorporates frequencies of nucleotides on each position and frequencies of haplotypes calculated via the EM algorithm from [2].

Results & Conclusion

Using the sample HCV clones from [3], four simulated data sets were generated with Grinder version 0.5[1]. Each dataset consisted of 20,000 reads from 10 variants and was categorized by its error model and the original population distribution. All four datasets contained mu-

*This work has been partially supported by NSF award IIS-0916401, NSF award IIS-0916948, Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture.

tation-based (i.e., substitution, insertion, and deletion) errors which were distributed uniformly at a rate of 0.1 percent, while two additionally contained homopolymer errors. The population distribution adhered to either a uniform model, or power-law model with parameter $\alpha = 2.0$. kGEM was compared against error correction algorithm KEC [3] using sensitivity and positive predicted value (PPV) as a measure of the quality of the error-corrected data sets (Figure 1). kGEM outperforms KEC in all four datasets in sensitivity while besting KEC in three out of four for PPV. The contrast is most dramatic in the case of reads generated from a power-law distributed population which contain homopolymer errors.

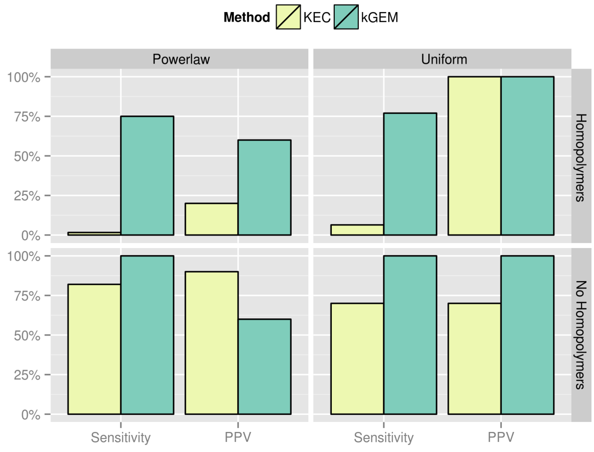

Figure 1: Sensitivity and PPV for the results on simulated data sets

References

[1] Florent E. Angly, Dana Willner, Forest Rohwer, Philip Hugenholtz, and Gene W. Tyson. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, 2012.

[2] I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I.I. Mandoiu, P. Balfe, and A. Zelikovsky. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, 12(Suppl 6):S1, 2011.

[3] P. Skums, Z. Dimitrova, D.S. Campo, G. Vaughan, L. Rossi, J.C. Forbi, J. Yokosawa, A. Zelikovsky, and Y. Khudyakov. Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics*, 13(Suppl 10):S6, 2012.

[4] E. Xing, R. Sharan, and M.I. Jordan. Bayesian haplo-type inference via the dirichlet process. In *Proceedings of the twenty-first international conference on Machine learning*, page 111. ACM, 2004.