**kGEM: An EM-based Algorithm for Local Reconstruction of Viral Quasispecies**

Alexander Artymenko, Nicholas Mancuso and Alex Zelikovsky  
Department of Computer Science  
Georgia State University  
Atlanta, Georgia 30302-3994,  
email:{aartyomenko, nmancuso, alexz}@cs.gsu.edu

Pavel Skums  
Centers for Disease Control and Prevention  
Atlanta, Georgia 30333  
email: kki8@cdc.gov

Ion Mândoiu  
Department of Computer Science & Engineering  
University of Connecticut  
Storrs, CT 06269,  
email:ion@enr.uconn.edu

**Abstract**—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.

**Keywords**—Next-generation sequencing, local reconstruction, expectation maximization, error correction, viral quasispecies.

**I. INTRODUCTION**

Single amplicon NGS reads refer to reads covering a viral genome region that can be covered by a single NGS reads (e.g., 400bp for 454 technology). In this paper we are dealing with the following

Local Population Reconstruction (Error Correction) Problem. Given a set \( R \) emitted by haplotype population \( P \), find a set of haplotypes \( H \) maximizing \( \Pr(R|H) \).

**II. METHODS**

The proposed algorithm Population \( k \)-genotype EM (kGEM) initially selects candidate haplotype set \( H \) covering \( R \) with at most \( s \) mismatches (e.g., \( s = 4 \)) and then transforms haplotypes into fractional haplotypes where each nucleotide (allele) is replaced with 5 probabilities each for the allele to be \( A, C, T, G, \) or to be deleted setting probabilities of observed nucleotide to 96% and all other probabilities set to 1%. Then the following steps are repeated until convergence: (1) For each haplotype \( H_i \in H \) and read \( r \in R \) estimate \( h_{i,r} = \Pr(R = r|H = H_i) \). (2) Estimate haplotype frequencies via EM using \( h_{i,r} \)'s and observed read frequencies. (3) Compute the normalized frequency \( f_{i,m}(e) \) of each allele in the \( m \)th position of \( H_i \). (4) Set frequency of the most frequent allele to 96% and all others to 1%. Collapse duplicates and drop rare genotypes upon completion and output the resulted set of haplotypes \( H \).

**III. RESULTS**

**Simulated Data.** Using a sample of 44 HCV clones from [4], 20 simulated data sets were generated with Grinder version 0.5[1]. Each dataset consisted of 100,000 total reads from a random sample of 10 variants and was categorized by its error model and generated population distribution. All reads contained errors (substitutions and indels) uniformly distributed at a rate of 0.1 percent. In addition, 10 datasets contained reads with simulated homopolymer errors. The population distribution adhered to either a uniform or power-law model with parameter \( \alpha = 2.0 \).

kGEM was compared against QuasiRecomb [3] using sensitivity and positive predicted value (PPV) as a measure of the quality of the error-corrected data sets (Table I). Reads were aligned using the tool IndelFixer[2]. Results shown are the mean and standard error over 5 datasets. kGEM outperforms QuasiRecomb in sensitivity in all 5 datasets. Further, kGEM has comparable PPV for the datasets with homopolymer errors and higher PPV for the non-homopolymer datasets.

**IV. CONCLUSION**

In this paper we propose a new reliable and fast EM-based method for error correction of amplicon NGS reads. Our preliminary results show advantage over QuasiRecomb.

**ACKNOWLEDGMENTS**

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

**REFERENCES**


