Poster: ViSpA: Viral Spectrum Assembling Method

Irina Astrovskaya*, Bassam Tork*, Serghei Mangul*, Kelly Westbrooks†, Ion Mandoiu‡, Peter Balfe§ and Alex Zelikovsky*
*Department of Computer Science, Georgia State University, Atlanta, Georgia 30303
Email: {iraa,btork,serhei,alexz}@cs.gsu.edu
†Life Technologies, Foster City, CA, Email: Kelly.Westbrooks@lifetech.com
‡Department of Computer Science &Engineering, University of Connecticut, Storrs, CT 06269
Email: mandoiu@cse.uconn.edu
§Institute of Biomedical Research, Birmingham University, Birmingham B15 2TT UK
Email:p.balfe@bham.ac.uk

***Keywords*-Next-generation sequencing, viral assembling, expectation maximization**

Like many RNA viruses, Hepatitis C virus (HCV) exists as a set of closely related sequences (*quasispecies*). The diversity of the quasispecies sequences can explain vaccines failures and virus resistance to existing therapies. Would the most virulent quasispecies are known in an infected host, the more effective treatment would be given to a patient. Since the original software of next-generation sequencing systems assumes a single genome, there is a need for a new assembler that infers viral population in a host. Thus, we focus on **Quasispecies Spectrum Reconstruction (QSR) Problem:** given a collection of 454 pyrosequencing reads taken from a sample quasispecies population, reconstruct the quasispecies *spectrum*, i.e., the set of sequences and the relative frequency of each sequence in the sample population.

The QSR problem is relatively a new topic, so few algorithmic approaches are available [1], [2]. The approaches build a *read graph*, where vertices correspond to reads, edges represent overlaps with agreement between reads, and possible sequences are paths from the leftmost to the rightmost vertices in the graph. ShoRAH [1] corrects genotyping errors via probabilistic clustering, reconstructs haplotypes via chain decomposition, and estimates haplotype frequencies by expectation maximization (EM) method. Previously, we proposed to infer haplotypes by applying network flows to transitively reduced read graph with edges, weighted by probability of the corresponding overlap [2].

In this poster, we introduce ViSpA method that significantly extends our previous approach by handling contaminated reads and overlaps with partial agreement between reads, by assembling haplotypes from per-vertex max-bandwidth paths via mutation-based clustering, and by estimating assemblies' frequencies via EM. We also suggest a procedure to fix systematic 454 errors in homopolymers if they happen in the coding region.

The ViSpA was validated on both simulated and real HCV data. In simulated studies, we created quasispecies spectrum by randomly choosing from 10 up to 40 sequences among 1739bp-long fragments of E1E2 region of 44 HCV sequences and mixing them accordingly to either uniform, or geometric, or skewed uniform distributions. Next we simulate either error-free reads or reads with systematic 454 errors [3]. In the first case, ViSpA correctly assembles all sequences out of 10 and 29 sequences out 40 quasispecies if average read length is at least 300bp. If the average read length is in 250bp-299bp range, we infer at least 8 out of 10 and 20 out of 40 sequences. In case of reads with 454 genotyping errors, ViSpA infers the most frequent quasispecies.

Then we applied ViSpA to the real 30K reads from 5.2Kbp-long region of HCV produced by 454 Life Science machine. The average read length is about 292bp. We reconstructed 10 the most frequent sequences corresponding to viable proteins. The neighbor-joining tree for these assemblies reminds a neighbor-joining tree for HCV quasispecies evolution. The most frequent sequence has been within 1% from the actual ORF obtained by cloning the quasispecies. ShoRAH was able to reconstruct only one sequence with a viable corresponding protein. This sequence has 99.94% similarity with our fourth most frequent assemblies. The rest of the ShoRAH's assemblies have defective proteins due to presence of stop codons in their amino-acid sequences. Additional experiments on 90% of read data shows that ten the most frequent assemblies are robust with respect to the assembling process in ViSpA.

Acknowledgments. This work has been partially supported by NSF awards IIS-0546457, IIS-0916401, IIS-0916948, and GSU Molecular Basis of Disease Fellowship.

References

[1] http://www.bsse.ethz.ch/cbg/software/shorah

[2] K. Wesbrooks, I. Astrovskaya, D. C. Rendon, Y. Khudyakov, P. Berman and A. Zelikovsky, *HCV Quasispecies Assembly using Network Flows*, Proc of ISBRA 2008, LNBI 4983: 159–170, 2008

[3] http://hackage.haskell.org/package/flowsim