Viral Quasispecies Reconstruction Based on Unassembled Frequency Estimation

Serghei Mangul[a], Irina Astrovskaya[a], Bassam Tork[a], Ion Mandoiu[b], and Alex Zelikovsky[a]

[a]Department of Computer Science, Georgia State University, Atlanta, GA 30303
{serghei,iraa,btork,alexz}@cs.gsu.edu
[b] Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269
mandoiu@cse.uconn.edu

Key words: 454 pyrosequencing, expectation maximization, viral quasispecies, haplotype assembling, haplotype discovery

1 Introduction

The genomic diversity of RNA viruses (such as Hepatitis C virus (HCV), Human immunodeficiency virus (HIV), SARS and influenza) is a subject of the great interest since it is a plausible cause of vaccines failures and virus resistance to existing therapies. RNA lacks ability to detect and repair mistakes during replication, many mutations are well tolerated and passed down to descendants producing a family of co-existing related variants of the original viral genome referred to as *quasispecies* [4, 14, 11]. Knowing the sequences of the most virulent variants can help in the design of effective drugs [3, 13] and vaccines [7, 5] by targeting particular viral genome *in vivo*. This paper is devoted to the following problem.

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 has been first addressed directly in [6, 15]. Eriksson et al. [6] proposed a multi-step approach consisting of genotyping error correction via clustering, haplotype reconstruction via chain decomposition, and haplotype frequency estimation via EM method with validation on HIV data. In Westbrooks et al. [15], the focus is on haplotype reconstruction via transitive reduction, overlap probability estimation and network flows with application to simulated HCV data. Recently the results of applications of the software tool ShoRAH [16] to HIV virus have been published in [17]. A novel combinatorial method have been also applied to HIV and HBV data with similar to ShorAH results [12]. Finally, in [10] we have proposed a novel algorithm **Vi**ral **Sp**ectrum **A**ssembler (ViSpA).

Our contributions include (1) a novel **Hap**lotype **Dis**covery algorithm HapDis which adds to set of candidate strings a virtual string which emits all reads that do not fit well to candidate strings, (2) combining ViSpA with HapDis allowing ViSpA preferably assemble reads attributed by HapDis to the virtual string.

2 Haplotype Discovery

2.1 Maximum Likelihood Model

Maximum likelihood model includes a panel and an instance of sequencing machine run consisting of read spectrum i.e. the set of reads and the relative frequency of each read.

Let us define panel to be consisting of (1) a set of candidate strings (e.g. obtained from existing databases or assembled from reads) that are believed to emit the reads and (2) a weighted match between reads and strings, where weight is calculated based on the mapping of the reads to the strings.

The possible gaps in the maximum likelihood model include (a) erroneous reads (caused by genotyping errors), (b) an incorrect list of candidate strings (absence of candidates caused by gaps in current databases and presence of chimeric candidates), (c) an inaccurate read-to-string match and, finally, (d) a non-uniform emitting of reads by strings. Since the genotyping quality is improving we focus on the incompleteness of the panel, i.e. list of candidate strings.

Haplotype Discovery Problem. *Given read spectrum and a panel, i.e. set of candidate strings, weighted match between reads and strings, find strings missing from the panel.*

We measure the model quality by the deviation between expected and observed read frequencies as follows:

$$D = \frac{\sum_j |o_j - e_j|}{|R|},$$

where o_j is observed read frequencies, e_i - expected read frequencies and R is number of reads.

Expected read frequencies are calculated based on maximum likelihood frequencies estimations of strings and weighted match between reads and strings as follows:

$$e_j = \sum_i \frac{h_{i,j}}{\sum_l h_{i,l}} f_j^{ML},$$

where $h_{i,j}$ is weighted match based on mapping of read r_j to string s_i, f_j^{ML} - maximum-likelihood frequency of candidate string.

2.2 ML estimates of string frequencies

Maximum-likelihood estimates of string frequencies are calculated by the Expectation Maximization algorithm.

First, we create a bipartite graph $G = \{S \bigcup R, E\}$ such that each candidate string is represented as a vertex $s \in Q$, and each read is represented as a vertex $r \in R$. With each vertex $s \in Q$, we associate unknown frequency f_s of the candidate string. And with each vertex $r \in R$, we associate observed read frequency o_r. Then for each pair s_i, r_j, we add an edge (s_i, r_j) weighted by probability of string s_i to emit read r_j with m genotyping errors:

$$h_{s_i,r_j} = \binom{l}{m} (1-\epsilon)^{l-m} \epsilon^m,$$

where l is length of read sequence, and ϵ is the genotyping error rate.

EM algorithm starts with the set of N strings. For each string we denote by f_s its(unknown) frequency. After initializing frequencies $f_{sq \in Q}$ at random, the algorithm repeatedly performs the next two steps until convergence:

- E-step: Compute the expected number $n(j)$ of reads that come from string i under the assumption that string frequencies $f(j)$ are correct, based on weights $h_{i,j}$
- M-step: For each i, set the new value of f_s to to the portion of reads being originated by string s among all observed reads in the sample

2.3 HapDis Algorithm

The main idea of the algorithm is to add to set of candidate strings a virtual string which virtually emits reads that do not fit well to assembled sequences.

Initially all reads are connected to the virtual string with weight $h_{i,j} = 0$. The first iteration finds the ML frequency estimations of candidates strings, ML frequency estimations of virtual string will be equal to 0, since all edges between virtual string and reads $h_{vs,j} = 0$. Then these estimation are used to compute expected frequency of the reads according to formula Section 2.1. If the expected read frequency is less than the observed one (under-estimated), then the lack of the read expression is added to the weight of the read connection to the virtual string. For over-estimated reads, the excess of read expression is subtracted from the corresponding weight (but keeping it non-negative). The iterations are continued while the deviation between expected and observed read frequencies is decreasing by more than ϵ.

Algorithm 1 HapDis algorithm

$h_{i,j} = \binom{l}{m} (1-\epsilon)^{l-m} \epsilon^m$,
add virtual string vs to the set of candidate strings
initialize weights $h_{vs,j} = 0$
while D change ¿ ϵ **do**
  calculate f_j^{ML} by EM algorithm
  $e_j = \sum_i \frac{h_{i,j}}{\sum_l h_{i,l}} f_j^{ML}$
  $D = \frac{\sum_j |o_j - e_j|}{|R|}$
  $\delta = o_j - e_j$
  if $\delta > 0$ **then**
    $h_{vs,j} += \delta$
  else
    $h_{vs,j} = \max\{0, h_{vs,j} + \delta\}$
  end if
end while

Based on weight between virtual string and all reads it is possible to find set of reads that were not emitted by candidate strings. From this set of reads become possible to reconstruct set of strings missing from the panel. Based on the frequency of virtual string it is possible to decide if the panel is likely to be incomplete, i.e. if the virtual string frequency is larger then certain threshold then it is likely that some strings are missing from the panel. The total frequency of missing strings is estimated by frequency of virtual string.

3 HapDis Enhancement of VISPA

Below is the flowchart for the proposed enhancement of ViSpA. The weights on read-to-virtual-string connection obtained by HAPDIS estimate the probability of a read to be emitted by an unassembled sequence. These probabilities are fed back to ViSpA and reads with low probability (to belong to an unassembled sequence) will be assigned high weight so that s-t-paths will try to avoid using them unless s-t-connection is cut. So ViSpA will be modiofied accordingly. Newly assembled quasispecies (Qsps) are added to the original library of candidates and HapDis will estimate the frequency of unassembled sequences as well as estimate new read weights. The iterations of the big loop will be repeated until certain stopping condition is satisfied, e.g., there are no new quasispecies sequences or the virtual string has too small estimated frequency. Then final EM will estimate ML frequencies and output the resulted viral spectrum.

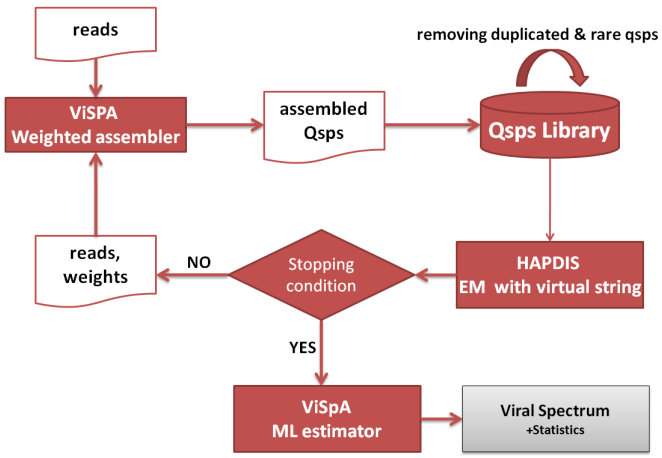

Fig. 1. Enhancement of ViSpA

References

1. National center for biotechnology information, *http://www.ncbi.nlm.nih.gov*.
2. S. Balser, K. Malde, A. Lanzen, A. Sharma, and I. Jonassen. Characteristics of 454 pyrosequencing data–enabling realistic simulation with flowsim. *Bioinformatics*, 26:i420–5, 2010.
3. N. Beerenwinkel, T. Sing, T. Lengauer, J. Rahnenfuehrer, and K. Roomp et al. Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics*, 21:3943–3950, 2005.
4. Martinez-Salas E. Sobrino F. de la Torre J.C. Portela A. Ortin J. Lopez-Galindez C. Perez-Brena P. Villanueva N. Najera R. Domingo, E. The quasispecies (extremely heterogeneous) nature of viral rna genome populations: biological relevancea review. *Gene, 40*, pages 1–8, 1985.
5. D.C. Douek, P.D. Kwong, and G.J. Nabel. The rational design of an AIDS vaccine. *Cell*, 124:677–681, 2006.
6. N. Eriksson, L. Pachter, Y. Mitsuya, S.Y. Rhee, and C. Wang et al. Viral population estimation using pyrosequencing. *PLoS Comput Biol*, 4:e1000074, 2008.
7. B. Gaschen, J. Taylor, K. Yusim, B. Foley, and F. Gao et al. Diversity considerations in HIV-1 vaccine selection. *Science*, 296:2354–2360, 2002.
8. T. Von Hahn, J.C. Yoon, H. Alter, C.M. Rice, B. Rehermann, P. Balfe, and J.A. Mckeating. Hepatitis c virus continuously escapes from neutralizing antibody and t-cell responses during chronic infection in vivo. *Gastroenterology*, 132:667–678, 2007.
9. Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M. Sharma, Philipp Khaitovich, Jörg Vogel, Peter F. Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, 09 2009.
10. Astrovskaya I., B. Tork, S. Mangul, K. Westbrooks, I. Mandoiu, P. Balfe, and Zelikovsky A. Inferring viral spectrum from 454 pyrosequencing reads. *BMC Bioinformatics (submitted)*.
11. M. Eigen M, J. McCaskill, and P. Schuster. The molecular quasi-species. *Adv Chem Phys*, 75:149–263, 1989.
12. Mattia Prosperi, Luciano Prosperi, Alessandro Bruselles, Isabella Abbate, Gabriella Rozera, Donatella Vincenti, Maria Solmone, Maria Capobianchi, and Giovanni Ulivi. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinformatics*, 12(1):5+, 2011.
13. S-Y. Rhee, T.F. Liu, S.P. Holmes, and R.W. Shafer. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS Comput Biol*, 3:e87, 2007.
14. Holland J.J. Steinhauer, D.A. Rapid evolution of rna viruses. *Annual Review of Microbiology, 41*, pages 409–433, 1987.
15. K. Westbrooks, I. Astrovskaya, D. Campo, Y. Khudyakov, P. Berman, and A. Zelikovsky. HCV quasispecies assembly using network flows. In *Proc. ISBRA*, pages 159–170, 2008.
16. Osvaldo Zagordi, Lukas Geyrhofer, Volker Roth, and Niko Beerenwinkel. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *Journal of computational biology : a journal of computational molecular cell biology*, 17(3):417–428, March 2010.
17. Osvaldo Zagordi, Rolf Klein, Martin Dumer, and Niko Beerenwinkel. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Research*, 38(21):7400–7409, 2010.