Multi-Commodity Flow Methods for Quasispecies Spectrum Reconstruction Given Amplicon Reads

Nicholas Mancuso[1] * †, Bassam Tork[1] * †, Pavel Skums[2], Ion Măndoiu[3] *, and Alex Zelikovsky[1] *

[1] Department of Computer Science
Georgia State University
Atlanta, Georgia 30302-3994
email: {nmancuso, btork, alexz}@cs.gsu.edu
[2] Centers for Disease Control and Prevention
1600 Clifton Road NE
Atlanta, Georgia 30322
email: kki8@cdc.gov
[3] Department of Computer Science & Engineering
University of Connecticut
Storrs, CT 06269
email: ion@engr.uconn.edu

Keywords: Next-generation sequencing. Viral quasispecies. Network flows.

RNA viruses depend on error-prone reverse-transcriptase for replication within an infected host. These errors lead to a high mutation rate which creates a diverse population of closely related variants [1]. This viral population is known as a *quasispecies*. As breakthroughs in next-generation sequencing have allowed for researchers to apply sequencing to new areas, studying genomes of viral quasispecies is now realizable. By understanding the quasispecies, more effective drugs and vaccines can be manufactured as well as cost-saving metrics for infected patients implemented [2].

Given a collection of (shotgun or amplicon) next-generation sequencing reads generated from a viral sample, the *quasispecies reconstruction problem* is defined as: reconstruct the quasispecies spectrum, i.e., the set of sequences and respective frequencies of the sample population.

Reconstructing the quasispecies spectrum is difficult for several reasons. The actual amount of variants may be obfuscated by conserved regions in the genome that extend beyond the maximum read length. Additionally, the amount of possible assignments of reads to variants in overlapping segments grows quickly. Furthermore, we are required to *rank* the variants by frequency. Previous approaches have utilized min-cost flows, probabilistic methods, shortest paths, and population diversity for the quasispecies spectrum assembly problem [3–6].

* 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
† This work has been partially supported by Georgia State University Molecular Basis of Disease research fellowship.

This work extends the maximum bandwidth method of [7] by including an exact multi-commodity flow method using Integer Linear Programming. Despite **ILP** being **NP**-hard, read graphs built from viral amplicon data tend to be small enough to solve quickly.

An *amplicon* A is a multiset of reads such that, each read $r \in A$ has the same predefined starting and ending position in the genome (i.e. $start_A, end_A$). Two amplicons A_1, A_2 are said to *overlap* if and only if $start_{A_2} < end_{A_1}$. A set of amplicons $\mathcal{A} = \{A_1, \ldots, A_m\}$ is said to be *overlapping* if and only if A_i and A_{i+1} overlap for $i = 1 \ldots m-1$. Given an overlapping set $\mathcal{A} = \{A_1, \ldots, A_m\}$, we define a partial order for overlapping amplicons where $r \prec r', r \in A_i, r' \in A_{i+1}$ if and only if the suffix of r starting at $start_{A_{i+1}}$ is the same sequence as the prefix of r' ending at end_{A_i}.

Given an overlapping set $\mathcal{A} = \{A_1, \ldots, A_m\}$ an m-staged directed read-graph is defined as $G = (V = V_1 \cup \cdots \cup V_m \cup \{s, t\}, E, c)$, where $v \in V_i, 1 \leq i \leq m$ corresponds to a *distinct* read in amplicon A_i. An edge $(u, v) \in E$ if and only if $\text{read}_u \prec \text{read}_v$ for $u, v \notin \{s, t\}$, $u = s$ and $v \in V_1$, or $v = t$ and $u \in V_m$. Additionally, $c : V \to \mathbb{N}$ is the count of the read represented by $v \in V_i$ in amplicon A_i.

Lemma 1 (Each consistent overlap in amplicons A_i, A_{i+1} corresponds to a unique bipartite clique in G). *Suppose the contrary. Let $v, v' \in A_i$ and $u, u' \in A_{i+1}$, where $v \prec u, v \prec u', v' \prec u$. Since v' and u are comparable but v' and u' are not, the prefixes of u and u' must not be consistent. This implies a contradiction with $v \prec u$ and $v \prec u'$.* □

Using this simple fact, we output a new "forked" read-graph. An m-staged directed read-graph can be represented by an $(2m-1)$-staged "forked" read-graph. Given an $i \times j$ bipartite clique $K_{i,j}$ in G create an $i + j$ star graph S_{i+j} with a new "fork" vertex as the internal node. Repeating this for all bipartite cliques over V_k, V_{k+1} will produce a new "fork" stage F_k. Repeating again for all neighboring stages we see $m-1$ new fork stages. Lastly we denote $c : E \to \mathbb{N}$ to be the count function for edges. This will reduce the number of edges at the cost of additional vertices if the graph has sufficiently dense bipartite cliques. Given an edge (f, u) or (u, f) where f is a fork and u is a read vertex let $c(f, u)$ or $c(u, f)$ be $c(u)$. This will be useful for flow formulations. Figure 1 illustrates this transformation.

Given a forked read-graph, the quasispecies reconstruction problem may be restated as a network flow problem. A k-multi-commodity flow problem is defined as given k (s_i, t_i) pairs either minimize or maximize the total flow $f = \sum_{i=1}^{k} f^i$ subject to capacity and demand constraints. For the quasispecies reconstruction problem we wish to minimize the total k flows such that each read is fully covered. Additionally, we force each flow to to be unsplittable, i.e., each flow is

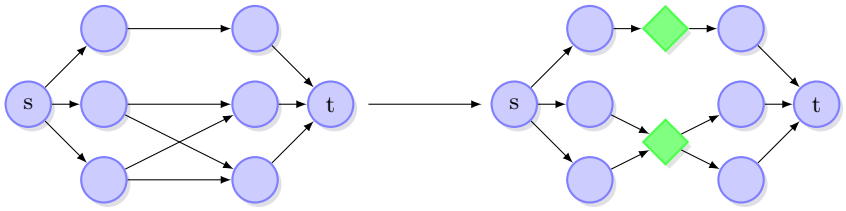

Fig. 1: Creating a "forked" read-graph from the original directed read-graph.

a simple s-t path in the read-graph, where $s_i = s, t_i = t, 1 \leq i \leq k$.

$$\textbf{Min:} \quad \sum_{\substack{0 \leq i \leq k \\ (s,u) \in E}} f^i_{s,u}$$

$$\begin{aligned}
\textbf{Subject to:} \quad & \sum_{i}^{k} g^i_{u,v} \geq c_{u,v} && \forall (u,v) \in E \\
& \sum_{u \in \text{ pred}(v)} g^i_{u,v} = \sum_{u \in \text{ succ}(v)} g^i_{v,u} && \forall v \in V, i = 1 \ldots k \\
& \sum_{u \in \text{ succ}(v)} f^i_{v,u} = 1 && \forall v \in V, i = 1 \ldots k \\
& f^i_{u,v} \geq g^i_{v,u} && \forall (u,v) \in E, i = 1 \ldots k \\
& f^i_{u,v} \in \{0,1\} && \forall (u,v) \in E, i = 1 \ldots k \\
& g^i_{u,v} \in [0,1] && \forall (u,v) \in E, i = 1 \ldots k
\end{aligned}$$

The method was run on data simulated from the E1E2 region of 44 HCV strains. Variants for each data set were produced from a uniform distribution, geometric distribution, or skewed distribution. Cross-validation was done by using Jensen-Shannon Divergence (JSD) to measure the quality of frequency assignment. JSD is defined as,

$$\text{JSD}(P||Q) = \frac{1}{2}\text{D}_{KL}(P||M) + \frac{1}{2}\text{D}_{KL}(Q||M)$$

where

$$\text{D}_{KL}(P||Q) = \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)}$$

and $M = \frac{1}{2}(P + Q)$. We also evaluate the quality of assembled quasispecies by using sensitivity and positive predicted value. Sensitivity measures the correctly assembled quasispecies out of the population, while PPV measures the correctly assembled quasispecies out of the assembled population. They are defined as,

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

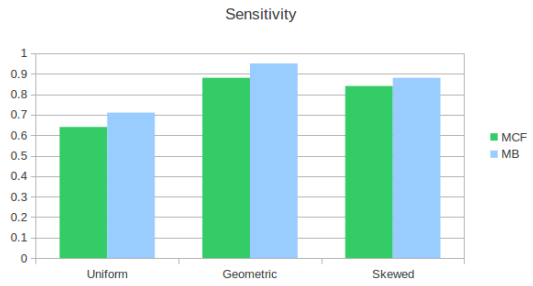

Fig. 2: Sensitivity of Multi-Commodity Flow and Maximum Bandwidth

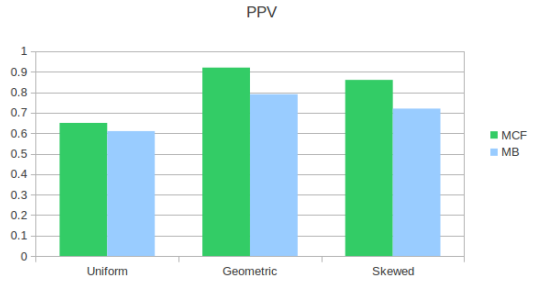

Fig. 3: PPV of Multi-Commodity Flow and Maximum Bandwidth

Under all three measures, the multi-commodity flow formulation performed competitively with Maximum Bandwidth. The flow formulation produced less variants than maximum bandwidth in all three quasispecies distributions. This lead to higher PPV (Fig. 3), but slightly less sensitivity (Fig. 2). The divergence is slightly higher than Maximum Bandwidth's due to skewing of frequencies from lower sensitivity (Fig. 4). While the current model performs quite well overall, we expect that further improvements to the flow model will lead to more accurate assemblies.

References

1. Duarte EA, Novella IS, Weaver SC, Domingo E, Wain-Hobson S, Clarke DK, Moya A, Elena SF, de la Torre JC, Holland JJ.: RNA Virus Quasispecies: Significance for Viral Disease and Epidemiology. Infectious Agents and Disease **3**(4) (1994) 201–214
2. Skums Pavel, Dimitrova Zoya, Campo David S., Vaughan Gilberto, Rossi Livia, Forbi Joseph C, Yokosawa Jonny, Zelikovsky Alex, Khudyakov Yury: Efficient error correction for next-generation sequencing of viral amplicons. In: International Symposium on Bioinformatics Research and Applications. (2011)

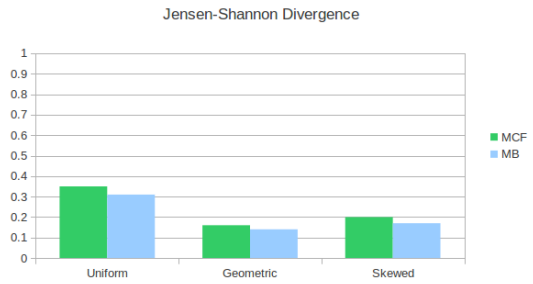

Fig. 4: Jensen-Shannon Divergence of Multi-Commodity Flow and Maximum Bandwidth

3. Westbrooks K, Astrovskaya I, Campo D, Khudyakov Y, Berman P, Zelikovsky A: HCV Quasispecies Assembly using Network Flows. In: Proc. International Symposium Bioinformatics Research and Applications. (2008) 159–170
4. Zagordi O, Klein R, Daumer M, Beerenwinkel N: Error Correction of Next-Generation Sequencing Data and Reliable Estimation of HIV Quasispecies. Nucleic Acids Research **38**(21) (2010) 7400–7409
5. Astrovskaya I., Tork, B., Mangul, S., Westbrooks, K., Măndoiu, I., Balfe, P., Zelikovsky A: Inferring Viral Quasispecies Spectra from 454 Pyrosequencing Reads. BMC Bioinformatics **12** (2011)
6. Prosperi MC, Prosperi L, Bruselles A, Abbate I, Rozera G, Vincenti D, Solmone MC, Capobianchi MR, Ulivi G: Combinatorial Analysis and Algorithms for Quasispecies Reconstruction using Next-Generation Sequencing. BMC Bioinformatics **12** (2011)
7. N. Mancuso and B. Tork and P. Skums and I. Mandoiu and A. Zelikovsky: Viral Quasispecies Reconstruction from Amplicon 454 Pyrosequencing Reads. In: Proc. 1st Workshop on Computational Advances in Molecular Epidemiology. (November 12, 2011 2011) 94–101