Workshop: A Maximum Likelihood Method For Quasispecies Spectrum Assembly

Nicholas Mancuso*, Bassam Tork*, Pavel Skums†, Lilia Ganova-Raeva†, Ion Măndoiu‡, Alex Zelikovsky*
* *Department of Computer Science*
Georgia State University
Atlanta, Georgia 30302-3994
email: {nmancuso, btork, alexz}@cs.gsu.edu
† *Centers for Disease Control and Prevention*
1600 Clifton Road NE
Atlanta, Georgia 30322
email: {kki8, lkg7}@cdc.gov
‡ *Department of Computer Science & Engineering*
University of Connecticut
Storrs, CT 06269
email: ion@engr.uconn.edu

***Keywords*-Next-generation sequencing. Viral quasispecies. Maximum likelihood.**

RNA viruses depend on error-prone RNA polymerase for replication within an infected host. These errors lead to a high mutation rate which creates a highly diverse population of 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 [2] implemented.

The problem of assembling the quasispecies spectrum is difficult for several reasons. Long conserved regions make it difficult to distinguish quasispecies in addition to the difficulty in correctly matching reads in overlapping segments. Furthermore, we are required to *rank* the quasispecies by frequency. It is not difficult to see that the quasispecies spectrum assembly problem is NP-hard via reduction from SUBSET SUM. One possible approach is to utilize a parsimonious objective. This optimization approach attempts to minimize the number of distinct quasispecies that explains the given read-data. However, a straightforward parsimony objective will not take into account frequencies over reads. Previous approaches have utilized min-cost flows, probabilistic methods, shortest paths, and population diversity for the quasispecies spectrum assembly problem [3], [6], [5], [4].

We propose a maximum-likelihood based approach for the quasispecies spectrum assembly problem inspired by minimum entropy principles. This approach is validated against simulated HCV amplicon data as well as actual HBV data.

Acknowledgements. 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, and Georgia State University Molecular Basis of Disease research fellowship.

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. (1994), *RNA virus quasispecies:significance for viral disease and epidemiology.*

[2] Skums Pavel, Dimitrova Zoya, Campo David S., Vaughan Gilberto, Rossi Livia, Forbi Joseph C, Yokosawa Jonny, Zelikovsky Alex, Khudyakov Yury. (2011). *Efficient error correction for next-generation sequencing of viral amplicons. International Symposium on Bioinformatics Research and Applications*

[3] Westbrooks K, Astrovskaya I, Campo D, Khudyakov Y, Berman P, Zelikovsky A: *HCV quasispecies assembly using network flows.* In *Proc. ISBRA* 2008:159-170.

[4] 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*(2011), *12*(1):5.

[5] Astrovskaya I., Tork, B., Mangul, S., Westbrooks, K., Măndoiu, I., Balfe, P., Zelikovsky A (2011) *Inferring Viral Quasispecies Spectra from 454 Pyrosequencing Reads. BMC Bioinformatics 12*

[6] Zagordi O, Klein R, Daumer M, Beerenwinkel N: *Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Research*(2010), *38*(21):7400-7409.

978-1-4673-1321-6/12/$31.00