#### Leman Nur Erkan

A General Review on the *Probabilistic Inference of Viral Quasispecies Subject to Recombination[[1]](#footnote-0)*

## 

## Background

**The Next Generation Sequencing** is a technique that provides the future genetic data of the organisms (Qin et al., 2019). Since Next Generation Sequencing (NGS) is very costly and it has big data that is noisy and incomplete, and it provides sampling and sequencing errors; to characterize and sort the data in a meaningful way is very hard.

**Haplotypes** are alignments of the variants on the genetic data in a single chromosome (Dervan, 2017). **Quasispecies** are the viral groups in a viral population and quasispecies have the same genetic composition as the structure of the haplotypes (Domingo, 1999). **Biological signal generators[[2]](#footnote-1)** are the generators that make signals which are well known and therefore the network of a biological system can be detected both *in silico* and *in vivo* (Taylor, 2019).

**In this study,** they presented a probabilistic model based on HMM and this model infers the haplotype distribution from viral quasispecies by using NGS data.

There are two main causes of the haplotype variation in a viral population: **recombinations** and **mutations**. Even mutations and recombination events are high in the non-conserved regions, in the conserved regions there are rare and in viral quasispecies, there is a dominant haplotype that shows very low recombination events in their evolutionary history, and many mutant types are around these major dominant haplotypes.

## Aim of the Paper

In this study, they aim to infer viral quasispecies that is obtained from NGS data which has very high error rates and size, via modeling the mutation and recombination events. In specific, they tended to identify haplotype probabilities - which is the architecture of the viral population in a host- that are shaping host interactions.

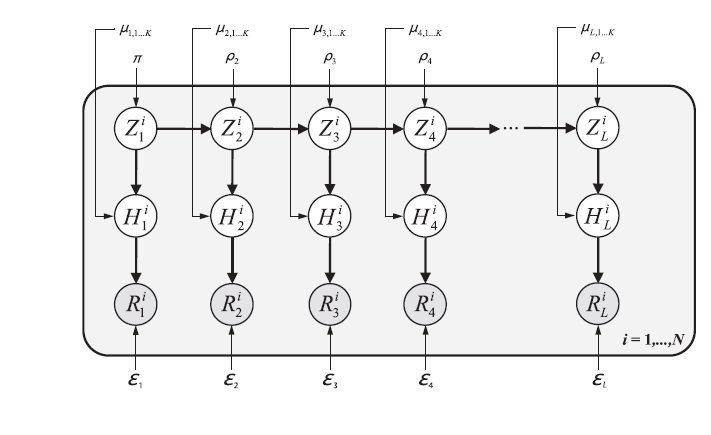
## The Application and Inference Problem

They used HMM to generate viral populations. They defined viral populations as haplotype distributions and created haplotypes from a small number of generations as a switch state of the HMM which is the *jumping HMM.* Even jumping HMM is used in biological applications before, this study differs from previous studies because of the erroneous sequence read data, high mutation and recombination rates, and the large number of haplotypes, which are not found in the previous studies. They used the Expectation-Maximization (EM) algorithm and Maximum a posteriori (MAP) estimation.

In this study, they used Hidden Markov Models as the *probabilistic reasoning system* in their study and used a forward-backward algorithm that gets the evidence and query and gives the answer. But for the limitations of large data, etc. they did not use the forward algorithm, rather, they used sampling to estimate the desired probabilities. They constructed the model, modified the algorithms in the model to achieve their intention and applied all to a Java program, *QuasiRecomb*.[[3]](#footnote-2)

## The PPL and the Model

They made a descriptive probabilistic model. The model parameters are the π-the probability of the generator, ρ-the transition probability, ε-rates, and μ-sequence profiles. For parameter estimation, they used the maximum likelihood (ML) approach and developed an Expectation-Maximization (EM) algorithm for ML estimation (MLE). They define the prior parameter distributions to enforce sparse maximum a posteriori (MAP) solutions and they modified the EM to estimate the MAP for the parameters. In Figure 1, the model is summarized.



**Figure 1.** The general representation of the model. Z is for the hidden random variable with state space, H is for the parental sequence that is generating the recombinant event, R is for the observed read of the sequence that is obtained from a haplotype that includes sequencing errors. For other denotations, see *The PPL and the Model*.

They made a descriptive probabilistic model. The model parameters are the π-the probability of the generator, ρ-the transition probability, ε-rates, and μ-sequence profiles. For parameter estimation, they used the maximum likelihood (ML) approach and developed an Expectation-Maximization (EM) algorithm for ML estimation (MLE). They define the prior parameter distributions to enforce sparse maximum a posteriori (MAP) solutions and they modified the EM to estimate the MAP for the parameters.

## The Inference Technique

Since RNA viruses show low recombination levels, and they have no real diversity in the conserved regions of the genome, the haplotype distributions are very low and mutatiıns are accumulated around the one or two main haplotypes. Therefore, to solve the MAP estimation problem, they used a Variational Bayes approach.

To get the haplotype distribution, which is the structure of the quasispecies, and since there are many haplotype possibilities, sampling was chosen rather than enumeration.

## Results and Discussion

***Results:*** In this study, there did one simulation study and one study that uses a real data set (for HIV). For simulation study, they construct seven data sets that have different recombination breakpoints and detect the haplotype distributions of each dataset. They realized that the minimum reads on the dataset should be 500 and more to correctly select the proper K (K=2) generating sequences. They also tested the sensitivity and for this, they implemented some errors to the error-free reads from the original haplotypes. In each case, QuasiRecomb identifies the recombinants even if the data has high errors. For a real HIV dataset study, they used NGS reads which belong to the real data of an HIV-infected patient. After the estimation, they found the quasispecies of the HIV-infected patient is dominated by a single haplotype (with a frequency of 31%). Clusters of the mutants surround this main haplotype and in this haplotype, the sequence similarities are between 93% and 99%. This means that the estimation of the real data is correctly achieved since the haplotype distribution is like that: there is one haplotype in the middle of the viral population and there should be some mutants around the dominant haplotype. To compare the results of the program, they did the same analysis in the software ShoRAH (Zagordi, 2011), and in this software, the results include nearly 15 haplotypes for the same data.

***Discussion:*** They offer that besides estimating the haplotype distribution by sampling, an approach that computes suboptimal haplotypes from the recombinant sequence generators (K) can be used. For this, an extension of HMM can be considered.

In addition to QuasiRecomb properties which include local read alignments that they used in their study, they mention that QuasiRecomb can estimate both local and global read alignments. This means the population data can be inferred from the genomic regions that are known.

Lastly, they point that the quasispecies inference approach in this study is based on NGS-long reads and while the reads are getting shorter, the accuracy of the method will be lower. However, they say that this is not a big problem for the future since the NGS technologies are developing and the long read numbers and techniques are increasing compared to the short reading of the genome.

**REFERENCES**

**Main Article**:

Töpfer A, Zagordi O, Prabhakaran S, Roth V, Halperin E, Beerenwinkel N. Probabilistic inference of viral quasispecies subject to recombination. J Comput Biol. 2013 Feb;20(2):113-23. doi: 10.1089/cmb.2012.0232. PMID: 23383997; PMCID: PMC3576916.

**Other References:**

Andrew Dervan, Jay Shendure, Chapter 3 - The State of Whole-Genome Sequencing, Editor(s): Geoffrey S. Ginsburg, Huntington F. Willard, Genomic and Precision Medicine (Third Edition), Academic Press, 2017, Pages 45-62, ISBN 9780128006818, https://doi.org/10.1016/B978-0-12-800681-8.00003-7.

Esteban Domingo, QUASISPECIES, Editor(s): Allan Granoff, Robert G. Webster, Encyclopedia of Virology (Second Edition), Elsevier, 1999, Pages 1431-1436, ISBN 9780122270307, <https://doi.org/10.1006/rwvi.1999.0240>.

Schwarz, Gideon. Estimating the Dimension of a Model. Ann. Statist. 6 (1978), no. 2, 461--464. doi:10.1214/aos/1176344136. https://projecteuclid.org/euclid.aos/1176344136

Taylor D. Scott, Kieran Sweeney, Megan N. McClean, Biological signal generators: integrating synthetic biology tools and in silico control, Current Opinion in Systems Biology, Volume 14, 2019, Pages 58-65, ISSN 2452-3100, https://doi.org/10.1016/j.coisb.2019.02.007.

Töpfer A and Beerenwinkel N. QuasiRecomb: prediction of recombinant viral quasispecies. F1000Posters 2012, **3**:1381 (poster)

Qin D. (2019). Next-generation sequencing and its clinical application. *Cancer biology & medicine*, *16*(1), 4–10. https://doi.org/10.20892/j.issn.2095-3941.2018.0055

1. By TÖPFER et al., 2013. [↑](#footnote-ref-0)
2. In this study, they used K for the notation of the generator. [↑](#footnote-ref-1)
3. Webpage: <https://bio.tools/quasirecomb> , for more information about QuasiRecomb, see *Töpfer, 2013 (poster).* [↑](#footnote-ref-2)