Probabilistic Inference of Viral Quasispecies Subject to Recombination

Osvaldo Zagordi[1,2,⋆,⋆⋆], Armin Töpfer[1,2,⋆], Sandhya Prabhakaran[3], Volker Roth[3], Eran Halperin[4,5], and Niko Beerenwinkel[1,2,⋆⋆⋆]

[1] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2] SIB Swiss Institute of Bioinformatics, Switzerland
[3] Computer Science Department, University of Basel, Switzerland
[4] Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Israel
[5] International Computer Science Institute, Berkeley, California, USA
niko.beerenwinkel@bsse.ethz.ch

Abstract. RNA viruses are present in a single host as a population of different but related strains. This population, shaped by the combination of genetic change and selection, is called quasispecies. Genetic change is due to both point mutations and recombination events. We present a jumping hidden Markov model that describes the generation of the viral quasispecies and a method to infer its parameters by analysing next generation sequencing data. The model introduces position-specific probability tables over the sequence alphabet to explain the diversity that can be found in the population at each site. Recombination events are indicated by a change of state, allowing a single observed read to originate from multiple sequences. We present an implementation of the EM algorithm to find maximum likelihood estimates of the model parameters and a method to estimate the distribution of viral strains in the quasispecies. The model is validated on simulated data, showing the advantage of explicitly taking the recombination process into account, and applied to reads obtained from two experimental HIV samples.

Keywords: Molecular sequence analysis, Sequencing and genotyping technologies, Next-generation sequencing, Viral quasispecies, Hidden Markov model.

1 Introduction

Next-generation sequencing (NGS) technologies have transformed experiments previously considered too labour intensive into routine tasks [11]. One application of NGS is the sequencing of genetically heterogeneous populations to quantify their genetic diversity. The genetic diversity is of primary clinical relevance, for example, in infection by RNA viruses, such as HIV and HCV. In these

⋆ These authors contributed equally.
⋆⋆ Current affiliation: Institute of Medical Virology, University of Zurich, Switzerland.
⋆⋆⋆ Corresponding author.

B. Chor (Ed.): RECOMB 2012, LNBI 7262, pp. 342–354, 2012.

systems, the high mutation rate of the pathogen, together with recombination, which can occur when different viral particles infect a single cell, give rise to a population of different but related individuals, referred to as viral quasispecies. We denote the different viral strains in this population as haplotypes. Studying the features of the viral quasispecies can shed light on the mechanisms of pathogen evolution in the host and it is of direct clinical relevance. In fact, the diversity of the quasispecies is known to affect virulence [19], immune escape [12], and drug resistance [10].

The quasispecies equation is a mathematical model for RNA virus populations evolving according to a mutation-selection process [6]. The dynamics of the model are described by a mutation term accounting for transformation of one viral haplotype (or strain) into another at the time of replication and a selection term that accounts for varying replication rates of different strains. The mutation process is generally considered as the result of point mutations only, although recombination is known to be frequent in many clinically relevant viruses, including HIV and HCV. For example, the recombination rate of HIV is estimated to be about ten fold higher than its point mutation rate. Therefore, the quasispecies model has been extended to account for both mutation and recombination [5]. At equilibrium, the model predicts the viral population to be dominated by one or a few haplotypes, which are surrounded by a cloud of mutants they constantly generate.

Recent NGS technologies allow for observing viral quasispecies at an unprecedented level of detail by producing millions of DNA reads in a single experiment. However, this high yield comes at a cost, because reads are usually short (up to 700bp with the latest technology and much shorter than the smallest viral genomes) and error-prone [8]. As a result, since the data obtained are incomplete and noisy, a meaningful characterization of viral populations by means of NGS requires careful analysis of the sequencing data [4].

In this manuscript, we aim at making inference of the viral quasispecies based on NGS data by explicitly modeling mutation and recombination. To this end, we use a hidden Markov model (HMM) to generate viral populations, i.e., haplotype distributions, and their probing by means of NGS. In our model, the haplotypes are originating from a small number of generating sequences via recombination (described as change of state in the HMM that selects from which sequence the haplotype derives) and mutation (described by position-specific probability tables for the generating sequences). The sequencing reads are obtained from the haplotypes subject to observation error. HMMs allowing for a switch between generating sequences, termed jumping HMMs, have been applied, for example, to sequence alignment of protein domains [18] and to detecting inter-host HIV circulating recombinant forms [17].

In order to reliably identify the haplotypes shaping intra-host quasispecies, including variants of low frequencies, sequencing errors must be corrected, as they will confound the true variation present in the sample. This has been approached, for example, by clustering reads or flowgrams and removing the within-cluster variation [15,21,7,22]. Rather than addressing error correction, in

the present manuscript, we present a novel generative probabilistic model for making inference of viral quasispecies, i.e., for estimating the intra-patient viral haplotype distribution. Specifically, we assume that the true genetic diversity is generated by a few sequences, called generators, through mutation and recombination, and that the observed diversity results from additional sequencing errors. Throughout, the sequencing error rate is assumed to be known (either from control experiments or from complementary analyses) and fixed.

We present the model for local haplotype inference, meaning that we aim at inferring the population structure in a genomic region of a size that can be covered by individual reads. Extending this model to global haplotype inference, i.e., to longer genomic regions, is straightforward and will be discussed briefly. Nevertheless, local inference will generally be more reliable and sufficient for many applications. For example, the HIV protease gene, an important target of antiretroviral therapy, is 297 bp long, and it is now standard to obtain reads of 400 bp and longer with the Roche/454 GS Junior platform, a common pyrosequencing platform for clinical diagnostics. Local haplotype reconstructions can also be used as a starting point for global reconstruction [20,7,14,2,13].

We show that our model is able to estimate the true distribution of the haplotypes with high reliability, by applying it to simulated data, where we have access to the ground truth. We also present an application to experimental data obtained from sequencing mixtures of HIV clones.

2 Methods

2.1 Hidden Markov Model

During infection a viral strain can change either by point mutation, when a single base is copied with error, or by recombination, when a cell is infected by more than one viral particle, and viruses in subsequent generations produce a sequence that is a mosaic of those of the progenitors. The model we present here does not aim at representing these evolutionary processes mechanistically. Rather, it is a descriptive probabilistic model, in which the quasispecies is generated by switching among K different generating sequences, each of length L. These generators are defined as sequence profiles (μ_{jkv}) indicating the probability over the alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{-}\}$ of base $v \in \mathcal{A}$ at position j of the k-th generating sequence. The set of sequences generates viral haplotypes $H \in \mathcal{A}^L$ by mutation (modeled by the probability tables (μ_{jkv})) and by recombination (modeled by switching to a different profile) as follows.

Let Z_j be the hidden random variable with state space $[K] = \{1, \ldots, K\}$ indicating the parental sequence generating H_j, the haplotype character at position j. We denote by ρ_j the probability of recombination, i.e., of switching the generating sequence right before position j. Each observed read R with bases R_j is obtained from a haplotype subject to noise (sequencing errors) at rate ε. The model is depicted in Figure 1 and defined as

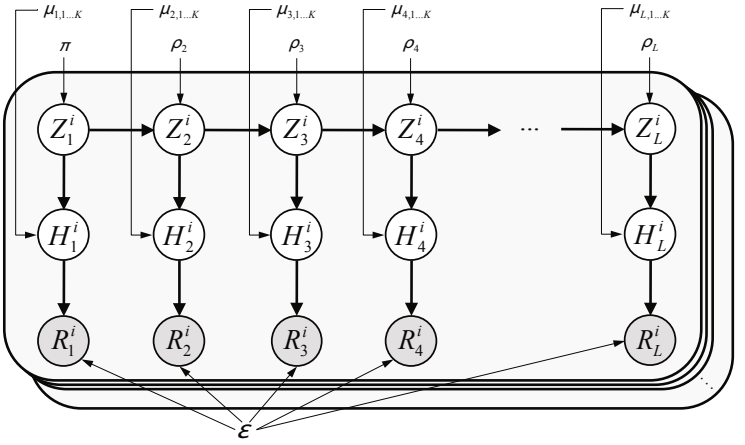


Fig. 1. Graphical representation. of the model. Only one observation i is depicted; for the full model, the graph is replicated for $i = 1, \ldots, N$.

$$\Pr(Z_1 = k) = \pi_k \tag{1a}$$

$$\Pr(Z_j = k \mid Z_{j-1} = l) = \begin{cases} \rho_j, & \text{if } k \neq l \\ 1 - (K-1)\rho_j, & \text{otherwise} \end{cases} \tag{1b}$$

$$\Pr(H_j = v \mid Z_j = k) = \mu_{jkv} \tag{1c}$$

$$\Pr(R_j = b \mid H_j = v) = \begin{cases} \varepsilon, & \text{if } b \neq v \\ 1 - (n-1)\varepsilon, & \text{otherwise} \end{cases} \tag{1d}$$

where $n = |\mathcal{A}|$ is the size of the alphabet.

The full model consists, for each observation $i = 1, \ldots, N$, of the hidden random variables Z_j^i (indicating generator sequences) and H_j^i (the haplotypes of the quasispecies), and the observed reads R_j^i, for all sequence positions $j = 1, \ldots, L$. The model parameters are summarized as $\theta = (\pi, \rho, \mu)$, and ε is a fixed constant.

2.2 EM Algorithm (Baum-Welch)

In the following we compute the likelihood $\Pr(R \mid \theta)$ and develop an expectation-maximization (EM) algorithm for finding the maximum likelihood estimates (MLE) of the parameters θ. The likelihood factorizes into the product over independent reads and, for each read, it can be computed efficiently using the Markov property. We have

$$\begin{aligned}\Pr(R) &= \prod_i \sum_{Z^i,H^i} \Pr(Z^i, H^i, R^i) \\ &= \prod_i \sum_{Z^i,H^i} \prod_j \Pr(R^i_j \mid H^i_j) \Pr(H^i_j \mid Z^i_j) \Pr(Z^i_j \mid Z^i_{j-1}),\end{aligned}$$

where we use the definition $\Pr(Z^i_1 \mid Z^i_0) = \Pr(Z^i_1)$. Using the distributive law, each sum in this expression can be factored along the Markov chain, which gives rise to the forward algorithm [16]. In this manner, the likelihood can be computed in $O(NLK^2)$ time.

The EM algorithm is an iterative procedure to find local maxima of the likelihood as a function of θ by maximizing the auxiliary Baum's function $Q(\theta, \theta')$, the expected hidden log-likelihood of the data with respect to the posterior distribution of (Z, H) given θ'. Here, θ' is the previous estimate of the parameters (π, ρ, μ). Baum's function is defined as

$$Q(\theta, \theta') = \mathrm{E}_{Z,H|\theta'}[\log \Pr(R, Z, H \mid \theta)].$$

It bounds the log-likelihood from below, and repeated iterations of the maximization step with respect to θ (M-step) alternated with estimations of the distributions $\Pr(Z, H \mid R, \theta)$ (E-step) are guaranteed to find a local maximum of the likelihood function.

For the E-step, we compute

$$\begin{aligned}Q(\theta, \theta') = \sum_{j,k,v} N_j(k, v) \log \mu_{jkv} + \sum_{j=2}^{L} \left[N_j^{\neq} \log \rho_j + N_j^{=} \log(1 - (K-1)\rho_j) \right] + \\ + N^{\neq} \log \varepsilon + N^{=} \log(1 - (n-1)\varepsilon) + \sum_k N_1(k) \log \pi_k,\end{aligned}$$

where $N_1(k)$ is the expected number of times a Markov chain starts in state k at position 1, $N_j^{\neq}$ is the expected number of times that a Markov chain switches from a state k to a state $l \neq k$ at position j, $N_j^{=}$ is the expected number of times it does not switch, and $N_j(k, v)$ is the expected number of times the Markov chain is in state k and emits haplotype character v at position j. These expected counts are estimated for all reads by computing posterior probabilities of the hidden variables H, Z given the data and the current estimate of θ, using the forward and backward algorithm [16].

In the M-step, the parameters θ are updated by solving

$$\theta^{\mathrm{new}} = \arg\max_{\theta} Q(\theta, \theta') \ .$$

This is achieved by setting

$$\pi_k = \frac{N_1(k)}{\sum_k N_1(k)}, \qquad \rho_j = \frac{1}{K-1} \frac{N_j^{\neq}}{N}, \qquad \mu_{jkv} = \frac{N_j(k, v)}{\sum_v N_j(k, v)}.$$

The two steps are iterated until convergence (defined here as a relative change of the log-likelihood smaller than 10^{-6}). Since the EM algorithm is only guaranteed to find a local maximum, we performed 50 random restarts and chose the solution with the largest likelihood. The initial parameter values were drawn at random from the following distributions:

$$\begin{aligned} \pi &\sim \mathrm{Dir}(2,\ldots,2), \\ \mu_{jk} &\sim \mathrm{Dir}(2,2,2,2), \\ \rho_j &\sim \frac{1}{K-1+M}\mathrm{Beta}(2,2), \end{aligned}$$

where Dir and Beta are, respectively, the Dirichlet and Beta distributions, and M the number of positions where ρ_j is non-zero (see below). Reads are hashed at the beginning in order to identify identical ones and to avoid unnecessary computations. The sequencing error rate ε was fixed to 0.01%

2.3 Sparse Recombination Rates

The model (1a–d) allows for recombination at each site $j = 1,\ldots,L$ with rate ρ_j, but it is not possible to estimate, for each sequence position of a read separately, from which generating sequence it has originated. However, despite high recombination rates, real RNA virus populations always display genomic regions that are conserved or nearly conserved (and that define the virus). In these regions, the different generating sequences cannot be distinguished because there is no or little diversity. Therefore, recombination among different sequences is expected to occur only at a small fraction of genomic sites, i.e., most recombination rates ρ_j are expected to be zero.

Sparse estimators of ρ can be found by considering the regularized maximum likelihood problem

$$\max_{\theta} \ \log \Pr(R \mid \theta) - f(\theta).$$

A natural choice for the regularization term is $f(\theta) = \lambda||\theta||_p$, where λ controls the degree of regularization. Values of p between 0 and 1 result in the desired sparsity of the parameters. The choice $p = 1$ is the lasso regularization, and approximation algorithms for solving it have been proposed [9].

Here, we consider the combinatorial model selection ($p = 0$) and select an optimal small subset of non-zero transition parameters ρ_j. We only allow ρ_j to be different from zero at M out of the $L-1$ positions. In practice, this is achieved by choosing $\rho_j = 0$ as initial value for $L-1-M$ positions, as they will remain zero throughout the EM procedure. In each re-start, the M positions for which we allow $\rho_j > 0$ are sampled uniformly at random. We choose $M = 3$ because more than three recombination sites are unlikely to be detectable for the size and the diversity of the genomic regions considered here.

2.4 Estimating the Haplotype Distribution

The quantity of greatest interest we derive from the model is the haplotype distribution $P(H)$, i.e., the structure of the viral quasispecies. For given model

parameters θ, we can compute the probability of each haplotype efficiently using the forward algorithm. The distribution $P(H)$ might be estimated by computing the probability of each haplotype. However, since there are 4^L possible haplotypes, enumeration is infeasible. We estimate $P(H)$ by sampling from the model using equations 1a–1c. We sampled 10,000 haplotypes at the MLE of θ obtained in the previous step. This procedure is efficient because almost all model parameters μ_{jkv} and ρ_j are very close to either zero or one and as a result the probability mass will be centered on a few haplotypes.

3 Results

3.1 Simulated Datasets

We assessed the performance of our model on four different datasets, corresponding to different distributions of haplotypes of 300 bp length. In the first dataset, the haplotypes have frequencies 80%, 10%, 5% and 5% respectively. The two most frequent ones have a mutual sequence identity of 94%. They recombine to produce the latter two at recombination breakpoint at position 198. The second population serves as a negative control. It consists of three haplotypes with pairwise similarities 94%, 91%, and 91%, and we sampled reads at frequencies 80%, 10% and 10% without recombination. The third dataset comprises the same three haplotypes at frequencies 28%, 28% and 26%, and six others at frequency 3% each obtained by recombining the dominating ones at position 198. The fourth dataset consists again of the same three dominating haplotypes at frequencies 22%, 21% and 21%, and 24 recombinants at frequency 1.5% each obtained by recombining the three at two breakpoints at positions 198 and 280.

We performed model selection by evaluating the BIC score, defined as

$$\log \Pr(R \mid \theta) - \frac{\nu \log N}{2},$$

where ν is the number of free parameters of the model. For each distribution of haplotypes, we sampled 50 datasets of 2000 reads with point mutations at an error rate of 0.03% per base and evaluated the BIC score. Figure 2 reports the BIC scores for the three datasets. In all cases, the BIC score is maximum at the correct number of generators (two for the first dataset, three for the others). For comparison, we additionally learned a model in which recombination is not possible, i.e., where $\rho_j = 0$ for all j. For datasets 1 and 3, the BIC score is maximized at $K = 2$ and $K = 3$, respectively, but without recombination the correct number of generators (and of haplotypes) is $4 = 2 + 2$ and $9 = 3 + 6$, respectively. Thus, the recombination-free model fails to reconstruct the quasispecies structure in these cases.

In order to study the impact of the sample size, we sampled instances of the first dataset of different sizes and repeated the model selection procedure. We analyzed datasets of 500, 750, 1000 and 2000 reads. The results are reported in Figure 3. For the smallest sample, the BIC score erroneously selects $K = 1$,

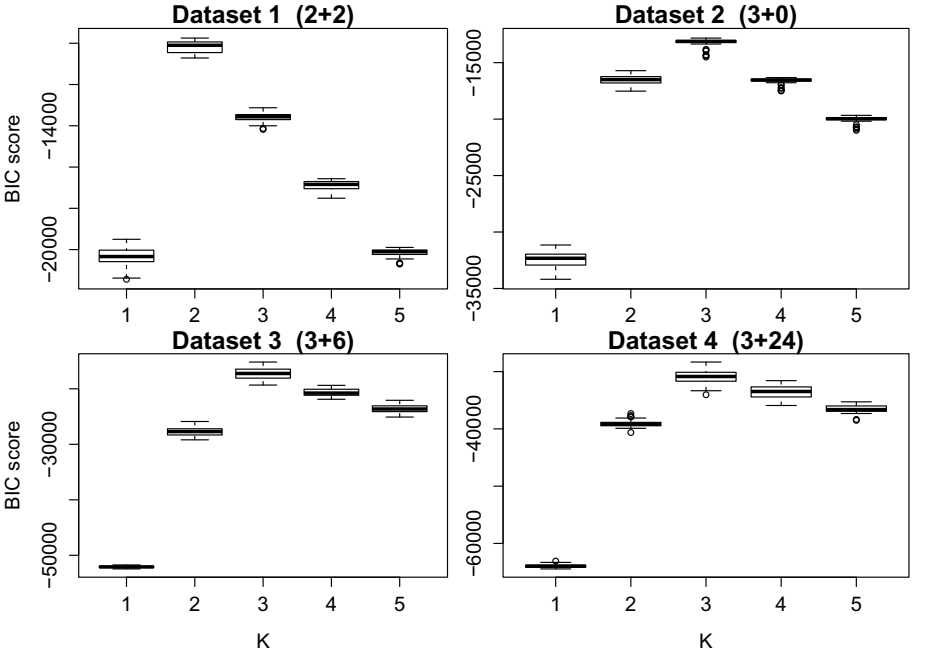


Fig. 2. BIC score for the four simulated datasets. The model correctly chooses $K = 2$ for the first dataset and $K = 3$ for the others. The boxplots summarize results of 50 independent datasets. The numbers in parentheses report the number of original generators plus the number of recombinants.

whereas for 750 and more reads, the procedure correctly selects $K = 2$ generating sequences indicating that sufficient coverage is an important prerequisite.

For parameter estimation, we sampled an additional set of reads from the haplotype distribution, and ran the EM algorithm with the value of K inferred before. Then, we inspected the MLE of the parameters μ, π and ρ. For datasets 1 and 2, where $K = 2$ and $K = 3$, respectively, were chosen, the estimates of μ_{jkv} are very close to either zero or one. The recombination parameter ρ is different from zero before the recombination hotspot and close to it.

In the second dataset, where no recombinants are present, the three generating sequence profiles corresponded exactly to the original haplotypes (i.e., all μ_{jkv} were either close to zero or to one) and no recombination was found (i.e., $\rho_j = 0$ for all j). In this case, π represents the frequency of the original haplotypes. Its estimate was very close to the original distribution and the remaining discrepancy can be explained by the sampling variance of the reads alone.

We also inspected the estimated parameters when all ρ_j are constrained to zero (no recombination), for dataset 1 with $K = 2$ and dataset 3 with $K = 3$. The estimates of μ_{jkv} reflected the distribution of bases at each individual position, but, as expected, the generating sequences cannot predict the original haplotypes. This is a consequence of the poor performance of the model selection in this case.

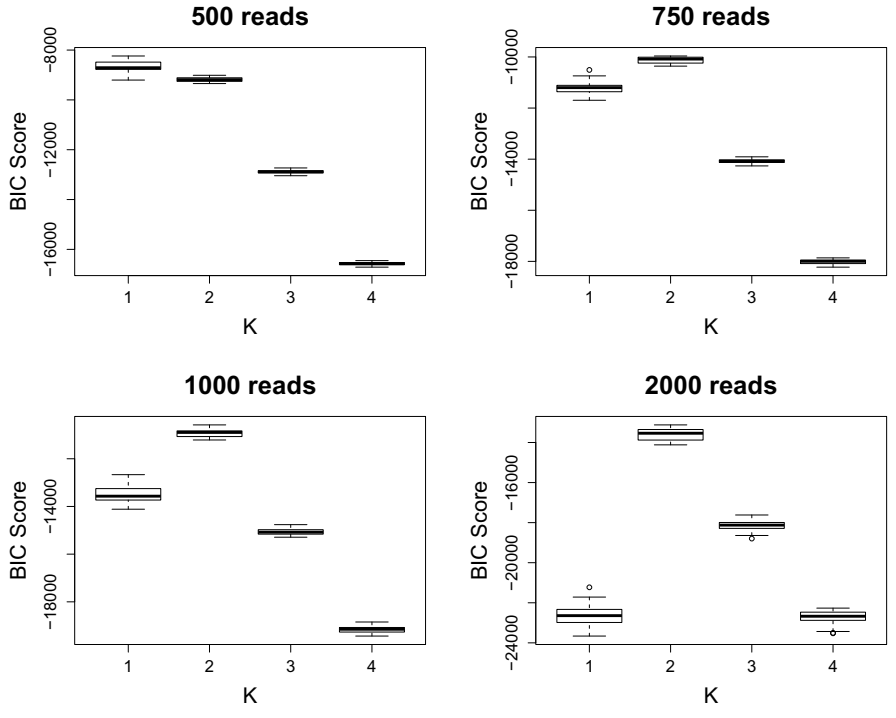
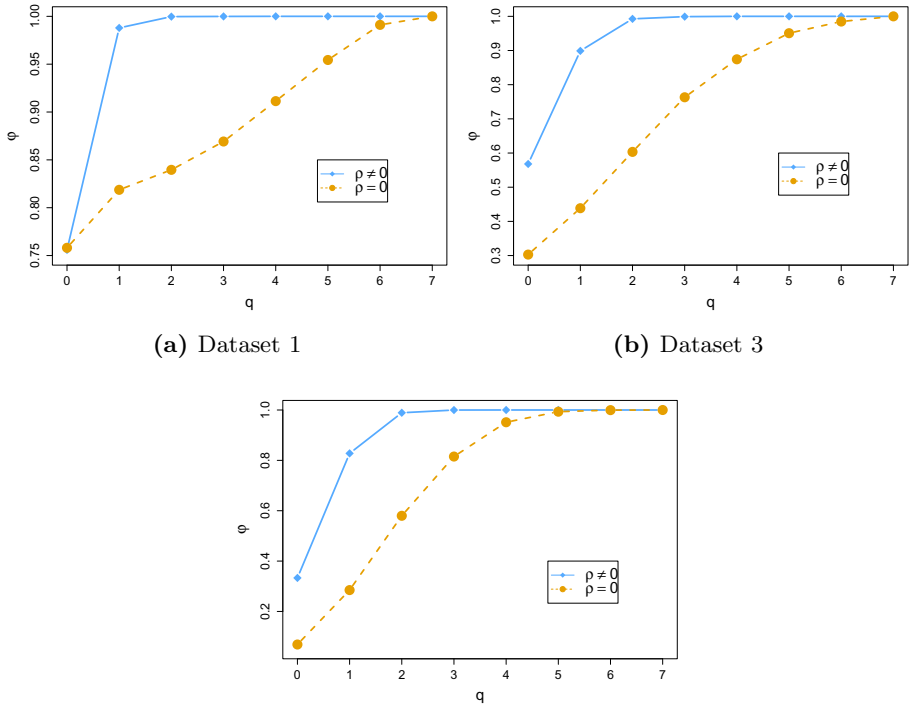


Fig. 3. BIC score for simulated dataset 1 at different sample sizes between 500 and 2000 reads. The model selection correctly selects $K = 2$ already with 750 reads. The boxplots summarize results on 50 independent datasets.

We assessed the reconstruction of the population by comparing it to the original set of sequences using the proportion close measure, φ_q, defined as the fraction of inferred haplotypes that match an original one with at most q mismatches [7]. Figure 4 reports the proportion close, as a function of the number of allowed mismatches, for the models with and without recombination estimated from datasets 1 and 3. The advantage of modeling recombinants is evident as the fraction of the population reconstructed is always higher than in the recombinant-free case. The proportion close is at least 99% for $q \geq 3$.

3.2 Real HIV Dataset

Using our model, we analyzed two sets of experimental NGS reads downloaded from the Sequence Read Archive (SRA). The first one (SRA run SRR069887) was obtained by sequencing a clinical sample from an HIV-infected patient in the context of a study of viral tropism [1]. We selected 1517 reads overlapping a 179 bp long region of the *env* gene (positions 6321–6499 in the HXB2 reference strain). We ran the EM algorithm on 50 datasets, generated by bootstrapping 1517 reads each, and selected the model with $K = 2$ generators (Figure 5, left).



(a) Dataset 1

(b) Dataset 3

(c) Dataset 4

Fig. 4. Proportion close, φ_q, as a function of q. The fraction of the population reconstructed with q mismatches is higher than 99% already for $q = 3$ if one allows recombination, and, depending on the dataset, for $q = 6$ or 7 if one does not allow recombination.

By visual inspection of the estimated parameters, we found two positions where ρ is significantly different from zero. This translates into 8 possible different reconstructed recombination patterns and additional haplotypes of low frequencies generated by mutation.

The second sample (SRA run SRR002680) consists of 3954 reads overlapping a region of length 155 bp of the *env* gene (HXB2 positions 7085–7239). Under the same bootstrapping procedure used for the other dataset, model selection resulted in $K = 3$ generators (Figure 5, right).

In order to appreciate the compactness of the model inferred with the jumping HMM, we compared its solutions with those of another tool to reconstruct haplotypes, implemented in the software ShoRAH [20]. This method does not take recombination into account and it identified 15 haplotypes in the dataset SRR069887, which can be further reduced to 10 if one excludes those with frequencies lower than 1% and those which harbor a frameshift due to a deletion. Similarly, more than 50 haplotypes were found in the dataset SRR002680, which can be reduced to 18 by the same analysis.

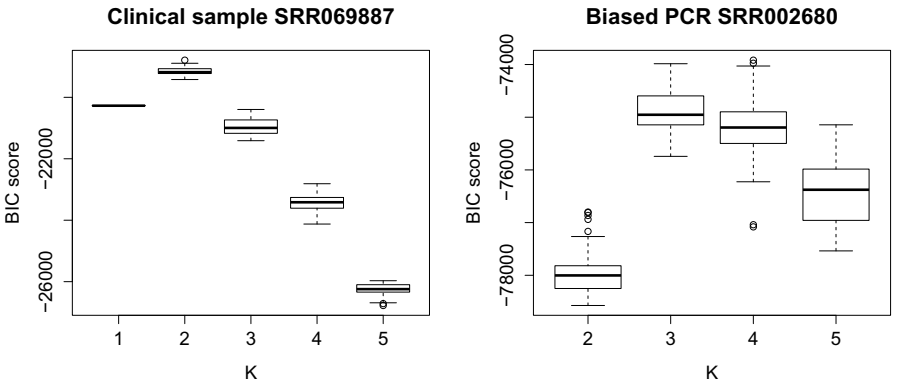


Fig. 5. BIC scores for experimental datasets based on 50 bootstrap samples

4 Discussion

We have presented a probabilistic model based on a HMM that infers the distribution of haplotypes in a viral quasispecies from NGS data. The model describes these different viral strains present in the population as originating from different generating sequences by means of two processes: point mutation and recombination. Point mutation is captured by the fact that the sequences are modeled as probability tables over the sequence alphabet. Recombination is modeled via a change of the sequence from which the haplotype is drawn as indicated by a change of state, or a jump, in the hidden Markov chain. Due to the possibility to switch between sequences, the number of tables necessary to describe the population structure remains small, whilst offering an excellent fit to the data. This results in a more compact and structured description of the viral population.

Using the EM algorithm, we find the MLE of the model parameters, namely the emission and transition probabilities. We have also introduced sparse recombination rates accounting for the fact that only few sites exist where generators recombine. Our results on simulated data demonstrate the usefulness of incorporating such sparsity while inferring haplotypes from recombinant reads.

There are several ways to extend the methodology presented here. For example, a different strategy for the regularization of the transition parameters ρ_j could be explored. We have observed in some cases a slow optimization of the likelihood, a behavior that might be related to the unidentifiability of the model, which implies that there are regions in the parameter space where the likelihood is flat. A Bayesian approach in which one could run a modified EM algorithm to maximize the posterior distribution of the parameters might solve this issue. In this framework, an appropriate prior might also provide efficient regularization.

Additionally, the transition matrix might be generalized allowing a different parameter for each pair of different states. This would account better for recombinant sequences present in the population at different frequencies. In this case, an efficient regularization of the ρ_j would be even more necessary.

Previous work on the analysis of NGS data to estimate genetic diversity have approached model selection in a non-parametric way by using the Dirichlet process mixture [21]. Extension of the HMM in this direction have been proposed and might be explored in this context as well [3].

Currently, we have presented our results in a local reconstruction setting, but our method generalizes to global inference. In this scenario, the population structure inferred locally is extended to longer genomic regions (longer than the typical read length). This can be achieved, for example, by allowing for longer generating sequences, along with two additional silent states to describe the unobserved regions before and after each read in the same fashion as the pair-HMM can be used for semi-global sequence alignment.

Although no current sequencing technology produces reads longer than 1000 bp, new platforms are foreseen to push this limit further. With such long reads, the probability to observe recombinations on a single read will be higher, and the necessity to keep the number of generators small will be even more compelling.

References

1. Archer, J., Rambaut, A., Taillon, B.E., Harrigan, P.R., Lewis, M., Robertson, D.L.: The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through timean ultra-deep approach. PLoS Comput. Biol. 6(12), e1001022 (2010), http://dx.doi.org/10.1371/journal.pcbi.1001022
2. Astrovskaya, I., Tork, B., Mangul, S., Westbrooks, K., Mandoiu, I., Balfe, P., Zelikovsky, A.: Inferring viral quasispecies spectra from 454 pyrosequencing reads. BMC Bioinformatics 12(suppl. 6) (2011), doi:10.1186/1471-2105-12-S6-S1
3. Beal, M., Ghahramani, Z., Rasmussen, C.: The infinite hidden Markov model. Advances in Neural Information 14, 577–584 (2002), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.8840&rep=rep1&type=pdf
4. Beerenwinkel, N., Zagordi, O.: Ultra-deep sequencing for the analysis of viral populations. Current Opinion in Virology (January 2011) (in press), http://dx.doi.org/10.1016/j.coviro.2011.07.008
5. Boerlijst, M., Bonhoeffer, S., Nowak, M.: Viral quasi-species and recombination. Proceedings: Biological Sciences 263(1376), 1577–1584 (1996), http://www.jstor.org/stable/50405
6. Eigen, M.: Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften (January 1971), http://www.springerlink.com/index/Q47866457218X543.pdf
7. Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R.W., Beerenwinkel, N.: Viral population estimation using pyrosequencing. PLoS Computational Biology 4(4), e1000074 (2008), http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000074
8. Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T., Martin, J.F.: Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12, 245 (2011), http://www.biomedcentral.com/1471-2164/12/245
9. Graça, J., Ganchev, K., Taskar, B., Pereira, F.: Posterior vs. parameter sparsity in latent variable models. In: NIPS 2009 (2009)

10. Johnson, J.A., Li, J.F., Wei, X., Lipscomb, J., Irlbeck, D., Craig, C., Smith, A., Bennett, D.E., Monsour, M., Sandstrom, P., Lanier, E.R., Heneine, W.: Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naïve populations and associate with reduced treatment efficacy. Plos Med. 5(7), 158 (2008)
11. Metzker, M.L.: Sequencing technologies - the next generation. Nat. Rev. Genet. 11(1), 31–46 (2010)
12. Nowak, M.A., Anderson, R.M., McLean, A.R., Wolfs, T.F., Goudsmit, J., May, R.M.: Antigenic diversity thresholds and the development of AIDS. Science 254(5034), 963–969 (1991), http://www.sciencemag.org/cgi/reprint/254/5034/963
13. Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N., Roth, V.: HIV-haplotype inference using a constraint-based dirichlet process mixture model. In: Machine Learning in Computational Biology (MLCB) NIPS Workshop 2010, pp. 1–4 (October 2010)
14. Prosperi, M.C., Prosperi, L., Bruselles, A., Abbate, I., Rozera, G., Vincenti, D., Solmone, M.C., Capobianchi, M.R., Ulivi, G.: Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. BMC Bioinformatics 12(1), 5 (2011), http://www.biomedcentral.com/1471-2105/12/5
15. Quince, C., Lanzen, A., Davenport, R.J., Turnbaugh, P.J.: Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12, 38 (2011)
16. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition (with erratum). Proceedings of the IEEE 77(2), 257–286 (1989), doi:10.1109/5.18626
17. Schultz, A.K., Zhang, M., Leitner, T., Kuiken, C., Korber, B., Morgenstern, B., Stanke, M.: A jumping profile hidden Markov model and applications to recombination sites in HIV and HCV genomes. BMC Bioinformatics 7, 265 (2006)
18. Spang, R., Rehmsmeier, M., Stoye, J.: A novel approach to remote homology detection: jumping alignments. J. Comput. Biol. 9(5), 747–760 (2002)
19. Vignuzzi, M., Stone, J., Arnold, J., Cameron, C., Andino, R.: Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439(7074), 344–348 (2006), http://www.nature.com/doifinder/10.1038/nature04388
20. Zagordi, O., Bhattacharya, A., Eriksson, N., Beerenwinkel, N.: Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics 12, 119 (2011), http://www.biomedcentral.com/1471-2105/12/119
21. Zagordi, O., Geyrhofer, L., Roth, V., Beerenwinkel, N.: Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. J. Comput. Biol. 17(3), 417–428 (2010), http://www.liebertonline.com/doi/abs/10.1089/cmb.2009.0164
22. Zagordi, O., Klein, R., Däumer, M., Beerenwinkel, N.: Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res. 38(21), 7400–7409 (2010), http://nar.oxfordjournals.org/lookup/pmid?view=long&pmid=