The Conditional Quadratic Ascent Procedure for Maximum-A-Posteriori Parameter Estimation of Profile Hidden Markov Models, with an Application to HIV-1 Founder Identification

Paul T. Edlefsen*

Fred Hutchinson Cancer Research Center

Abstract

Previous studies have shown that a conditional maximization variant of the Baum-Welch algorithm improves parameter estimation for the profile hidden Markov models (PHMMs) widely used in genomic sequence analysis. In this article we introduce an alternative conditional maximization approach that directly maximizes the conditional likelihood through gradient ascent, combining the Baldi-Chauvin gradient determination with a successive parabolic line search. Through simulation studies we show that this new method outperforms not only its non-conditional variant but also both Baum-Welch and Conditional Baum-Welch. We measure performance on both training data and held-out validation data, and the new method performs especially well on validation data as compared with these standard approaches, and especially well when there are few sequences available for parameter estimation. We also demonstrate an application of the new approach to estimating the parameters of profile hidden Markov models of HIV-1 data sampled in acute infection, when single founders of infection are expected to follow the star-like phylogeny implied by the PHMM. By using a simple approach to identify multiple founders of infection as violations of that model assumption, we present a first effort towards fully automated HIV-1 founder detection.

Keywords: profile hidden Markov model, genomic sequence alignment, statistical inference, HIV, CBW, CQA

* *Email address:* pedlefse@fredhutch.org (Paul T. Edlefsen)

1. Introduction

We previously introduced a new algorithm for estimating parameters of Hidden Markov Models (HMMs). The algorithm, Conditional Baum-Welch, is a variant of the Baum-Welch algorithm (Baum, 1972). Baum-Welch is a computationally efficient expectation-maximization (EM) algorithm (Dempster et al., 1977; Scott, 2002) for HMMs. Conditional Baum-Welch is as efficient as Baum-Welch for HMMs that satisfy a sparseness condition. One such HMM is the profile hidden Markov model (Profile HMM, or PHMM) (Rabiner, 1989; Durbin et al., 1998), commonly used to model families of related protein sequences. Both the Baum-Welch and the Conditional Baum-Welch algorithms converge to a local optimum. There is no guarantee that the optimum is global, and in the context of Profile HMMs, the landscape is sufficiently rugged that premature convergence to local optima is often a problem. We previously showed Profile HMM results demonstrating that the Conditional Baum-Welch algorithm converges to higher optima than the standard ("Unconditional") Baum-Welch under a range of conditions (Edlefsen and Liu, 2010). We also introduced the Dynamic Model Surgery algorithm, which further improves both Baum-Welch and Conditional Baum-Welch by detecting and escaping traps in the rugged optimization landscape.

In this article we introduce an alternative method, Quadratic Ascent, with both conditional and unconditional variants. Like the Conditional Baum-Welch algorithm, the conditional variant of this new method employs a "rotated" forward-backward procedure that takes advantage of the sparseness of the transition matrix of the underlying Markov chain to isolate and conditionally update the parameters of the model. Rather than employ the Baum-Welch Expectation Maximization procedure, which maximizes the expected value of the negative log-likelihood, the Quadratic Ascent algorithm directly ascends the likelihood by moving the parameters in the direction of steepest ascent.

The essential insight behind these conditional maximization approaches was serendipitously discovered. Although the Baum-Welch algorithm is typically described and implemented using recursions ("forward" and "backward", described below) in time, in some instances it can be implemented using recursions in space rather than time. Essentially these recursions calculate function values in the cells of a matrix, and the Baum-Welch forward-backward algorithm is typically described as calculating each column of the matrix as a (linear) function of the previous column. With the Profile HMM, the algorithm can be just as well described as calculating these values row-by-row, with each row calculated as a linear function of the previous row. Since each row of the matrix corresponds to a state of the model, this row-wise orientation inspires conditional maximization algorithms that modify state-specific parameters row-by-row, interweaved with the forward-backward calculations.

Improving parameter estimation for hidden Markov models is worthwhile wherever they are applied, but this research is particularly motivated by the need for a better approach to parameterizing Profile HMMs. While Profile HMMs are widely used to model protein sequence families, their use for modeling DNA sequence families has been slow to develop. In fact, the two major Profile HMM software packages, HMMer (Eddy and School of Medicine and Dept. of Genetics and National Human Genome Research Institute (US), 1992) and SAM (Hughey and Krogh, 1995, 1996), originally supported DNA sequences but, until recently, both retained only token support, with all development effort focused on proteins. The latest version of HMMer (version 3.1: Eddy and Wheeler, 2013; Finn et al., 2011), reintroduces support for DNA search with nhmmer (Wheeler and Eddy, 2013), but this iteration of HMMer does not support building profiles from unaligned sequences, despite the origin of the HMMer package as an implementation of the Baum-Welch algorithm (Durbin et al., 1998). It turns out that Baum-Welch is particularly prone to get stuck in local optima when the sequences are composed of only the four nucleotide residues of DNA, as opposed to the twenty amino acid residues of protein sequences. Furthermore, protein sequence alignment approaches that employ Profile HMMs for remote homology detection are prone to fail when the sequences are diverged beyond the ability of the Baum-Welch algorithm (Kececioglu et al., 2010). Improving parameter estimation for Profile HMMs in both protein and nucleic acid contexts has great potential impact in medicine and biology, and wherever improved genomic sequence alignment (and statistical uncertainty assessment therein) is valued.

The Profile HMM approach is a way to simultaneously account for all possible multiple alignments. The common reliance by bioinformatics projects on single multiple alignments is problematic, since different multiple alignments result in different "downstream" analyses. For instance it is typical for probabilistic phylogenetic studies to depend on a single multiple alignment, with no account made for the error introduced by the variation in alignments.

This problem has recently received increased attention (e.g. see Lunter et al., 2008; Wong et al., 2008). The Profile HMM can be viewed as a model of the joint probability distribution of alignments and sequences. Profile HMMs have been proposed for use as alternatives to progressive multiple alignment algorithms (e.g. Eddy, 1995), but since they represent a distribution of multiple alignments, they can als be used in more sophisticated ways, such as for drawing random alignments, or for evaluating relative probabilities of different alignments.

Here we aim to apply Profile HMMs to modeling and aligning HIV-1, a highly variable exogenous viral genome that is notoriously difficult to align. While automated alignment software exists that is specifically tailored to HIV-1 (Gaschen et al., 2001), in practice HIV-1 multiple alignment requires skilled manual labor. A critical shortcoming of this manual approach, apart from its irreproducibility, is that downstream analyses relying on HIV-1 alignments inevitably fail to account for the uncertainty. For example, Bayesian phylogenetic models misrepresent posterior credibility when they condition on a single input alignment.

Full-scale implementation of Profile HMMs for modeling viruses and other highly diverged genomic sequence families has been hindered by the relative instability of the available estimation procedures. Unfortunately we found that this is the case even when employing the Dynamic Model Surgery (DMS) procedure that we previously introduced, without which even the Conditional Baum-Welch algorithm is insufficient to escape the many optimization traps (inescapable local maxima) found in the parameter landscape. We are pleased to report that with a more careful choice of starting parameters, all of these search algorithms (Baum-Welch, Conditional Baum-Welch, and the new Quadratic Ascent algorithms) can be made to achieve optima in the absence of DMS otherwise accessible only when used with DMS. Furthermore, the Conditional Quadratic Ascent (CQA) algorithm achieves maxima comparable and in some cases superior to those achieved by Conditional Baum-Welch, and in conditions relevant to HIV-1 modeling the parameters estimated by (Conditional) Quadratic Ascent are more similar to the simulated true parameters than those estimated by (Conditional) Baum-Welch, and results on cross-validation sets are consistently superior with the new approach.

The remainder of this article is organized as follows. We first describe the Profile HMM and briefly review the Baum-Welch and Conditional Baum-Welch algorithms. We then introduce and describe the new Quadratic Ascent algorithm through its conditional variant. In the Results section, we show through simulation studies the degree of improvement conferred by the new algorithm variants in the context of both amino acid sequence families and DNA sequence families across a range of simulation conditions. We also demonstrate an application of the new approach to estimating the parameters of profile hidden Markov models of HIV-1 data sampled in acute infection, when single founders of infection are expected to follow the "star-like" phylogeny implied by the PHMM, using a simple idea to detect multiple founders of infection as violations of that model assumption. We conclude with a discussion of the implications of these findings and directions for future study.

2. Baum-Welch

2.1. The Hidden Markov Model

The hidden Markov model (HMM) is a statistical model that has been widely applied across an array of disciplines, including notably time series analysis and natural language processing, and biological sequence alignment (Rabiner, 1989; Churchill, 1989; Baldi et al., 1994; Krogh et al., 1994; Liu et al., 1999). An HMM with parameters $\mathbf{\Theta}$ describes the joint distribution $\mathbb{P}(\vec{d}, \vec{h}|\mathbf{\Theta})$ of a vector of observed data $d_1, \ldots, d_K$ and a vector of the corresponding hidden data $h_1, \ldots, h_K$. Expressed in terms of a time series, an HMMs models the distribution of an observation d_τ at time $\tau \in 1..K$ as conditionally independent from observations at other times, given the latent state h_τ of a Markov chain at time τ:

$$d_\tau \sim e_\tau(\cdot|h_\tau), \tag{1}$$

$$h_\tau \sim t_\tau(\cdot|h_{\tau-1}), \tag{2}$$

where $e_\tau(\cdot|h_\tau)$ is the "emission" distribution of the observations and $t_\tau(\cdot|h_{\tau-1})$ as the "transition" distribution of the underlying Markov process. Figure 1 depicts the dependence structure of a generic HMM.

The Profile HMM is a special case of this general approach, adapted for use in modeling a genomic sequence alignment (Durbin et al., 1998). Here, the sequence residues are the observed data $\vec{d}$, with each time τ associated with exactly one of the observed residues (d_τ). The hidden states of the Markov chain represent the positions of an ancestral reference sequence.

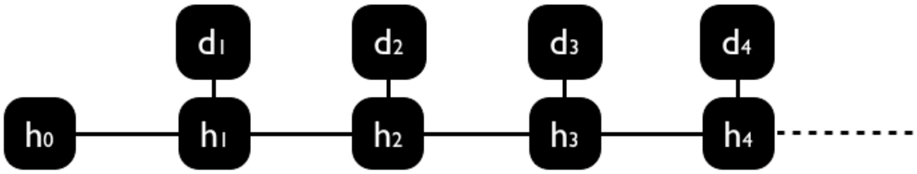


Figure 1: **The hidden Markov model.** The state of an unobserved Markov chain h_τ evolves over time τ according to transition kernel $t_\tau(\cdot|h_{\tau-1})$. At each time the observed datum d_τ is distributed according to the "emission" distribution $e_\tau(\cdot|h_\tau)$, which depends on the current state of the hidden Markov chain. [Figure 1 from Edlefsen and Liu (2010), reproduced with permission.]

Due to the processes of evolution (mutation and insertion/deletion), the correspondence between the ancestral and observed positions is obscured; its resolution constitutes a pairwise sequence alignment between the homologous pair (ancestor, decendant). Figure 2 depicts the Profile HMM.

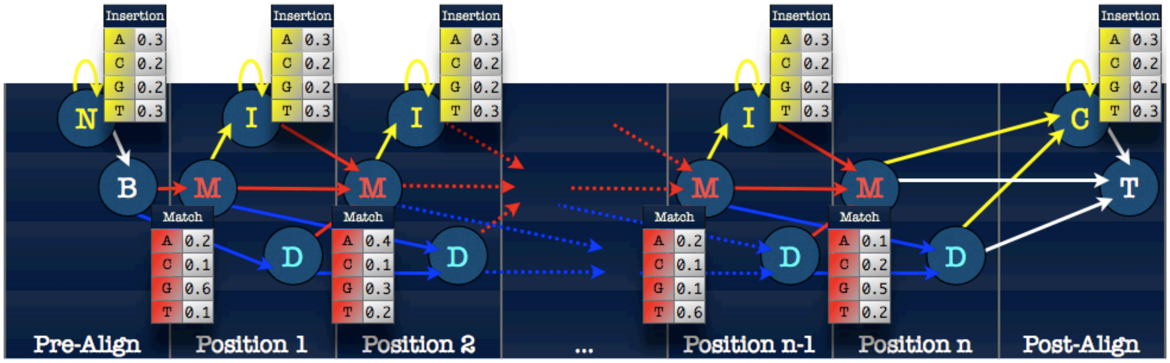


Figure 2: **The states of the Profile HMM.** There are $3n + 4$ states for a Profile HMM representing a sequence family with n ancestral positions. Each internal position has three associated states: Match, Insertion, and Deletion. Additional states represent flanking insertions. All match and insertion states have an associated emission distribution, which is multinomial over the allowed residues. The insertion emission distributions typically reflect the background residue frequencies, while the match distributions are position-specific. [Figure 2 from Edlefsen and Liu (2010), reproduced with permission.]

2.2. Baum-Welch and the forward and backward values

The transition and emission parameters of a Hidden Markov Model are generally estimable using a polynomial-time algorithm, Baum-Welch, which can be expressed as an instance of the expectation maximization (EM) algorithm for maximum-likelihood or maximum a-posteriori parameter estimation (Baum, 1972; Dempster et al., 1977; Rabiner, 1989). The algorithm uses dynamic programming to calculate "forward" values $\alpha_\tau(h_\tau) = \mathbb{P}(d^\tau, h_\tau|\mathbf{\Theta})$, where d^τ represents the initial τ residues of observed sequence $\vec{d}$.

The probability of the sequence $\vec{d}$ given the parameterized model is easily calculated from the forward value, since

$$\begin{aligned} \mathbb{P}(\vec{d}|\mathbf{\Theta}) &= \mathbb{P}(d^K|\mathbf{\Theta}) \\ &= \sum_{h_K \in \mathcal{S}} \mathbb{P}(d^K, h_K|\mathbf{\Theta}) \\ &= \sum_{h_K \in \mathcal{S}} \alpha_K(h_K), \end{aligned} \tag{3}$$

where $\mathcal{S}$ is the state space of the Markov chain, and $K = |\vec{d}|$ is the length of the sequence $\vec{d}$.

The dynamic programming recursion for computing α_τ as a function of $\alpha_{\tau-1}$ follows from the conditional independence assumptions of the Markov chain underlying the HMM:

$$\alpha_\tau(h_\tau) = \sum_{h_{\tau-1} \in \mathcal{S}} \alpha_{\tau-1}(h_{\tau-1}) t(h_\tau|h_{\tau-1}) e(d_\tau|h_\tau). \tag{4}$$

The "backward" value $\beta_\tau(h_\tau) = \mathbb{P}(d^{-\tau}|h_\tau, \mathbf{\Theta})$ is the conditional distribution of the remaining $K - \tau$ components of $\vec{d}$ (after d_τ), which we denote by $d^{-\tau}$, given the state h_τ at time τ. The recursion for calculating the backward value is similar to that of the forward value, only in reverse:

$$\beta_\tau(h_\tau) = \sum_{h_{\tau+1} \in \mathcal{S}} \beta_{\tau+1}(h_{\tau+1}) t(h_{\tau+1}|h_\tau) e(d_{\tau+1}|h_{\tau+1}). \tag{5}$$

The forward or the backward values can be used to efficiently calculate the likelihood function (Equation 3). The forward and backward values together yield the conditional distribution of the hidden state h_τ given the observed sequence $\vec{d}$, since

$$\begin{aligned} \mathbb{P}(h_\tau|\vec{d}, \mathbf{\Theta}) &= \frac{\mathbb{P}(h_\tau, \vec{d}|\mathbf{\Theta})}{\mathbb{P}(\vec{d}|\mathbf{\Theta})} \\ &= \frac{\alpha_\tau(h_\tau)\beta_\tau(h_\tau)}{\mathbb{P}(\vec{d}|\mathbf{\Theta})}. \end{aligned}$$

In a Bayesian context this can be used to compute the posterior probability that the HMM "emitted" the τ^{th} residue of sequence $\vec{d}$ from state h_τ. Assuming a degenerate prior, the joint posterior distribution of $h_{\tau-1}$ and h_τ is

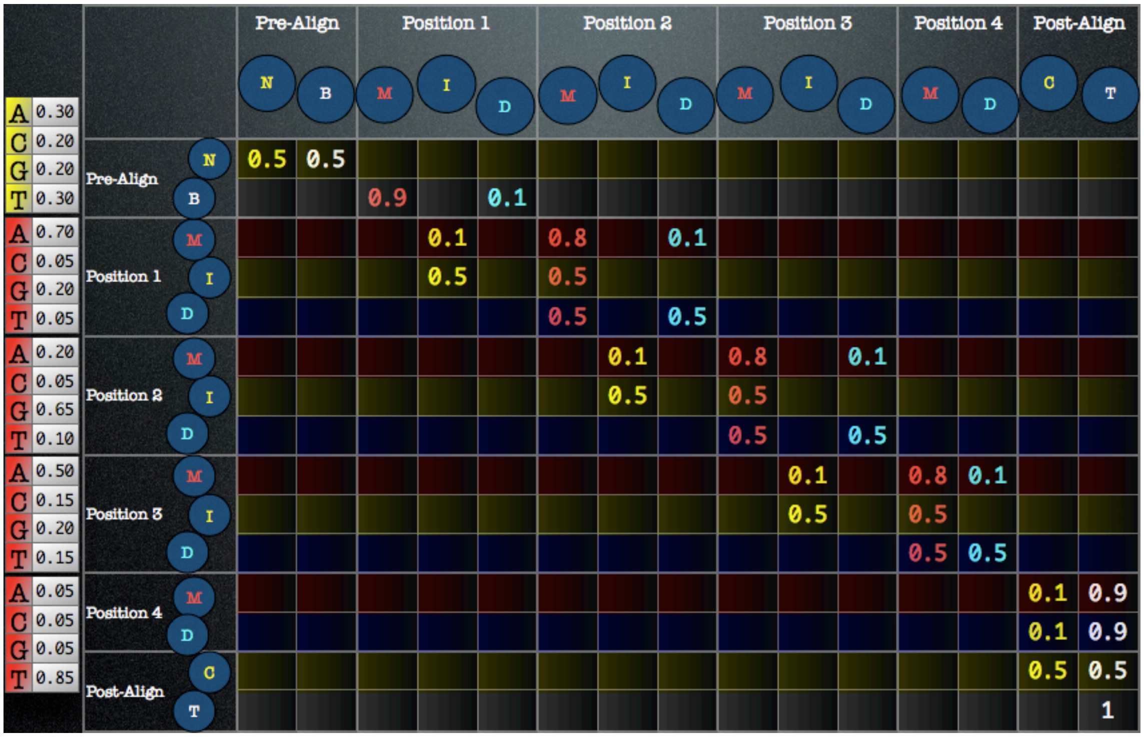
$$\begin{aligned}\mathbb{P}(h_{\tau-1}, h_\tau|\vec{d}, \mathbf{\Theta}) &= \frac{\mathbb{P}(h_{\tau-1}, h_\tau, \vec{d}|\mathbf{\Theta})}{\mathbb{P}(\vec{d}|\mathbf{\Theta})} \\ &= \frac{\alpha_{\tau-1}(h_\tau)t(h_\tau|h_{\tau-1})e(d_\tau|h_\tau)\beta_\tau(h_\tau)}{\mathbb{P}(\vec{d}|\mathbf{\Theta})}.\end{aligned}$$

The Baum-Welch procedure iteratively replaces the parameters $\mathbf{\Theta}^e$ of the "emission" distribution e and the parameters $\mathbf{\Theta}^t$ of the "transition" distribution t such that each parameter is proportional to the average (over all of the observed sequences) of the expected number of uses of the corresponding emission (or transition). We have previously shown that a conditional maximization variant of Baum-Welch (Conditional Baum-Welch) converges to better local optima than the Baum-Welch algorithm in the context of profile hidden Markov models (Edlefsen and Liu, 2010). As EM algorithms, both Baum-Welch and Conditional Baum-Welch achieve maximization indirectly by maximizing an auxiliary function that is guaranteed to lie below the objective function (an approach known generally as Minorization/Maximization, or MM (Hunter and Lange, 2004)). This auxiliary function is the so-called "Q function" of the EM algorithm. Here we introduce an alternative procedure that uses the same "rotated orientation" approach employed by Conditional Baum-Welch, but which employs a form of gradient ascent to directly maximize the posterior probability of the data.

3. Conditional Baum-Welch and the "rotated" orientation

We now briefly describe the logic of the "rotated" orientation used by both the Conditional Baum-Welch (CBW) algorithm and the CQA algorithm. We have shown through simulation studies that CBW is often able to escape the local optima that trap the Baum-Welch (UBW) algorithm. CBW depends on the same update procedure as Baum-Welch, but iteratively applies this procedure to conditional parameter distributions rather than to the complete joint likelihood/posterior. As Baum-Welch is an example of the EM algorithm, Conditional Baum-Welch is an example of a multi-cycle ECM (Expectation Conditional Maximization) algorithm (Meng and Rubin, 1993; Meng and van Dyk, 1997). Like Baum-Welch, CBW is guaranteed to increase the likelihood.

The CBW and CQA algorithms separately update the parameters associated with each model position (cf. Figure 2), holding fixed the values of the other parameters. In CBW this is iterated with standard Baum-Welch updates of parameters that apply to multiple states. See Edlefsen and Liu (2010) for further details.



A	0.30	0.70	0.20	0.50	0.05
C	0.20	0.05	0.05	0.15	0.05
G	0.20	0.20	0.65	0.20	0.05
T	0.30	0.05	0.10	0.15	0.85

		Pre-Align		Position 1			Position 2			Position 3			Position 4		Post-Align	
		N	B	M	I	D	M	I	D	M	I	D	M	D	C	T
Pre-Align	N	0.5	0.5													
	B			0.9		0.1										
Position 1	M				0.1		0.8		0.1							
	I				0.5		0.5									
	D						0.5		0.5							
Position 2	M							0.1		0.8		0.1				
	I							0.5		0.5						
	D									0.5		0.5				
Position 3	M										0.1		0.8	0.1		
	I										0.5		0.5			
	D												0.5	0.5		
Position 4	M														0.1	0.9
	D														0.1	0.9
Post-Align	C														0.5	0.5
	T															1

Figure 3: **The parameters of a four-position Profile HMM.** There are 16 states for a Profile HMM representing a sequence family with 4 ancestral positions. The transition parameters $\mathbf{\Theta}^t$ are represented as a transition matrix among the states. Blank elements of the transition matrix represent transitions disallowed by the model. Note that all transitions are relatively local, and the matrix is relatively diagonal. The emission parameters $\mathbf{\Theta}^e$ are depicted along the left edge of the transition matrix. Each position has its own multinomial Match emission distribution, and all positions share an Insertion emission distribution. The numbers given are examples; the BW, CBW, and CQA algorithms find values that are maxima in the likelihood (or posterior probability) landscape for a set of training sequences.

For CBW and CQA to be as efficient as BW, the transition probability matrix of the underlying Markov chain must be relatively sparse and relatively diagonal. In particular, there must exist an ordering of the states $\mathcal{S} = s_1, \ldots, s_S$ such that the probability $t(s_i, s_j)$ of transitioning from state i to state j is zero unless $j \geq i$ and $(j - i) \leq k$ for some fixed small constant k. As an example, consider the transition matrix depicted in Figure 3, which is an example of a transition matrix for a four-position Profile HMM. For the Profile HMM model, the condition is satisfied with $k = 5$ (from a Match state at one position to the Deletion state at the subsequent position).

When this condition is satisfied, the forward and backward dynamic programming recursions can proceed state-by-state rather than time-by-time. That is, instead of computing the forward values $\alpha_\tau(h_\tau)$ for time τ as a function of the forward values for all states at the previous time (as in Equation 4), the values can be computed for state s as a function of the k previous states at the previous time, since

$$\alpha_\tau(s_j) = \sum_{i \in (j-k), \ldots, j} \alpha_{\tau-1}(s_i) t(s_j | s_i) e(d_\tau | s_j). \tag{6}$$

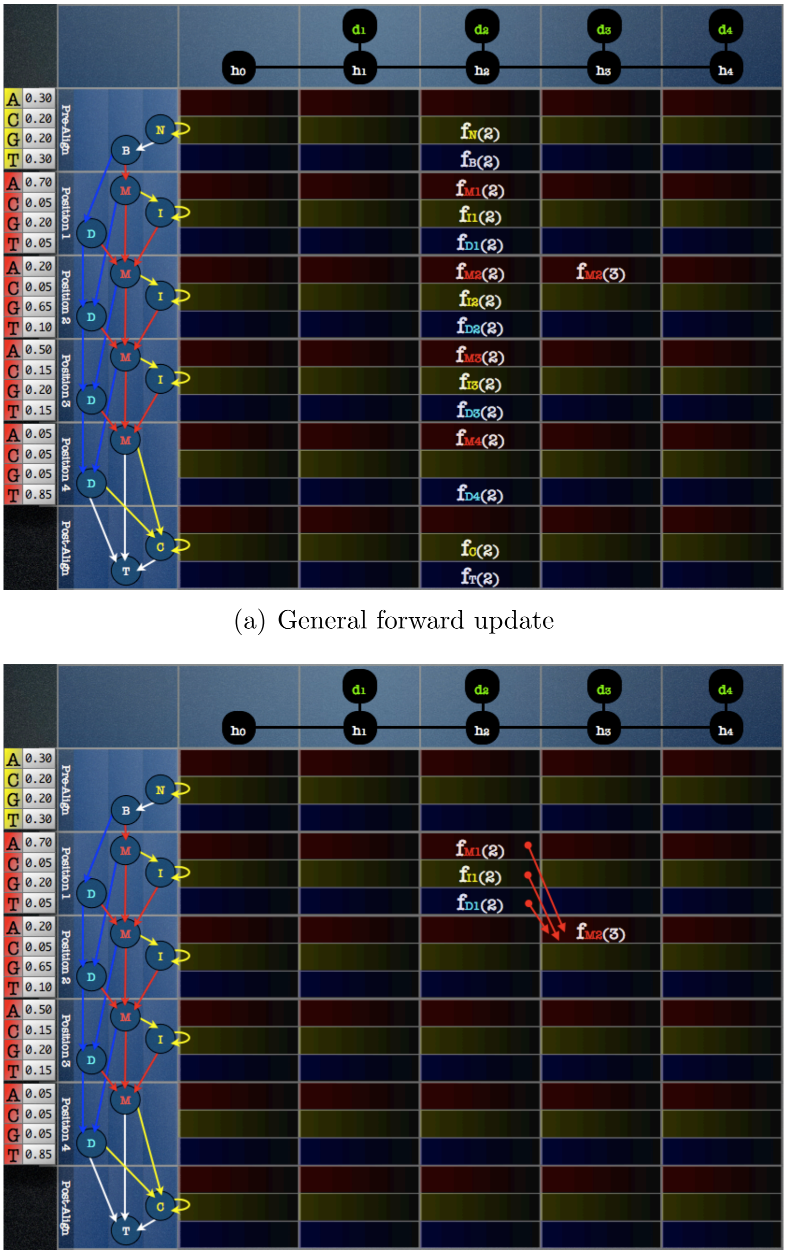
Figure 4 depicts the effect of this constraint in the context of the Profile HMM.

In some models it becomes convenient to allow for non-emitting states. In the Profile HMM, for example, we allow for Deletion states, which are non-emitting. Technically this is simply a convenient way to conceptualize and compute when certain transitions are geometrically-distributed: the Deletion states could be replaced by transitions between non-adjacent Match states. By adding non-emitting states, some HMMs that do not satisfy the constraint required for CBW and CQA can be restructured to do so.

Transitions to non-emitting states are most conveniently represented as transitions that take no time (since each time is associated with an emitted datum). When the HMM has non-emitting states, Equation 6 becomes

$$\alpha_\tau(s_j) = \sum_{i \in (j-k), \ldots, j} \big(\alpha_\tau(s_i) t(s_j | s_i) \mathbf{1}^{\not{\mathrm{e}}}(s_j) + \alpha_{\tau-1}(s_i) t(s_j | s_j) e(d_\tau | s_j) \mathbf{1}^{\mathrm{e}}(s_j) \big), \tag{7}$$

where $\mathbf{1}^{\mathrm{e}}(\cdot)$ is the function indicating whether its argument state is emitting, and $\mathbf{1}^{\not{\mathrm{e}}}(\cdot)$ is the function indicating whether its argument state is non-emitting. Figure 5 depicts the forward update for a Deletion state of the Profile HMM.



(a) General forward update

(b) Constrained forward update

Figure 4: **Forward calculation in a Profile HMM.** The general time-oriented forward calculation (of a Match state) is shown in (a). Since most transitions are not allowed, only the transitions denoted by arrows in (b) are relevant. This constraint is utilized by the Conditional Baum-Welch and Conditional Quadratic Ascent algorithms to update parameters state-by-state efficiently.

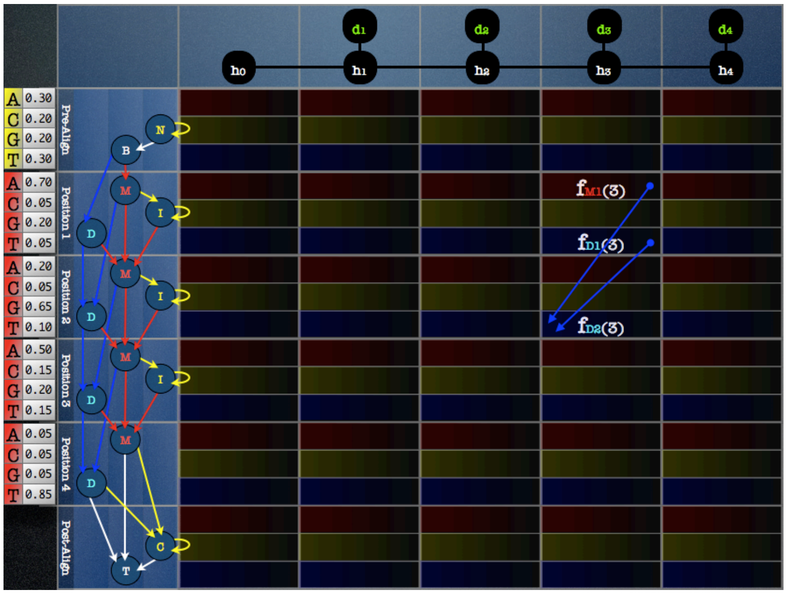
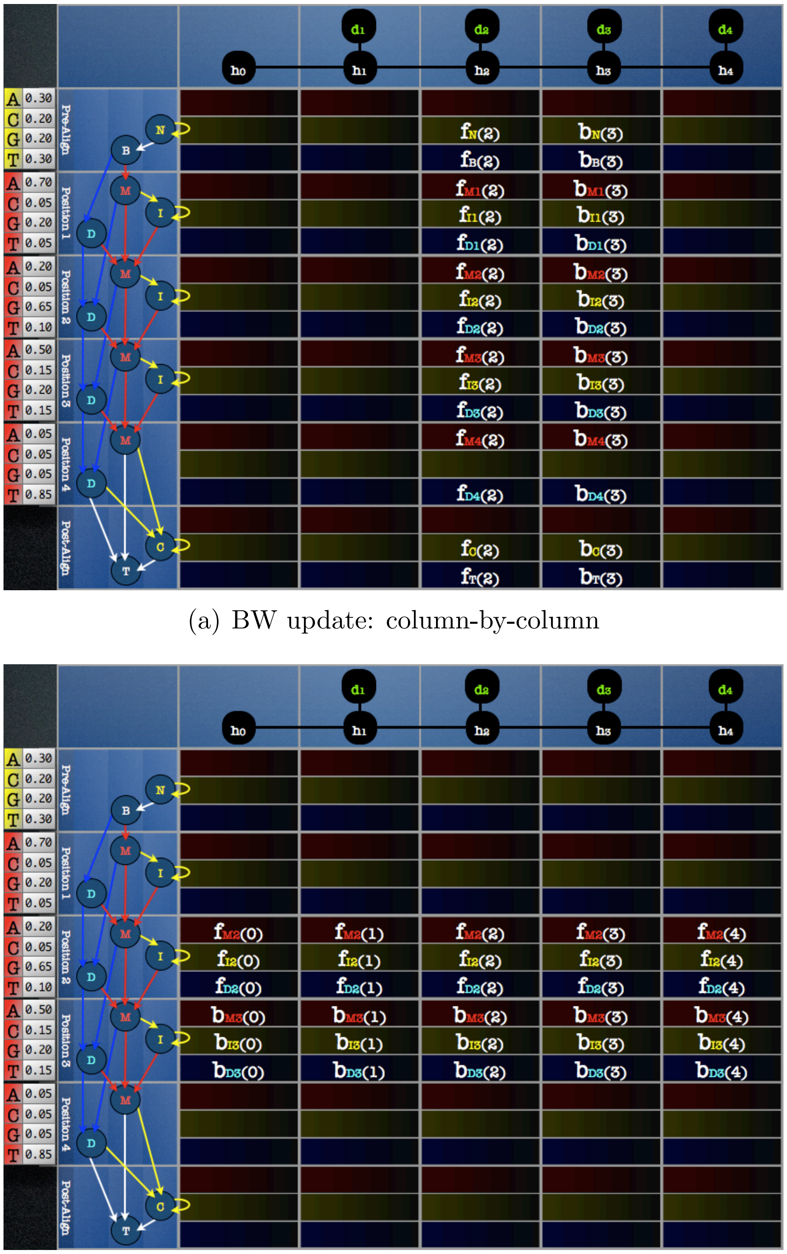


Figure 5: **Forward calculation of a non-emitting state in the Profile HMM.** Since Deletion states in Profile HMMs are non-emitting, transitions to them are within one column.

The CBW and CQA algorithms compute the forward and backward values row-by-row (state-by-state) rather than column-by-column (time-by-time) (Figure 6). Since the calculation of the forward values in a row depends only on the parameters affecting the states of that row, and on the forward values of the previous row, the CBW and CQA algorithms only need to re-compute the forward values of the affected row. The algorithms proceed row-by-row, updating the group of parameters associated with the states of that row. After the update, the algorithms calculate the forward values of the affected row, then proceed to the next row. Since each forward value needs to be calculated only once, the total computational cost of these algorithms is on the same order as that of the BW algorithm (bilinear in the number of states and the total length of all observation sequences).

We showed in Edlefsen and Liu (2010) that the CBW algorithm can outperform the BW algorithm across a range of simulation conditions, as well as in the biological context of identifying transposons in DNA sequences. In the results section below we show that the alternative approach of directly maximizing by gradient ascent (CQA) performs at least as well as the CBW procedure and better in cases relevant HIV-1 within-host evolution studies.



(a) BW update: column-by-column

(b) CBW and CQA updates: row-by-row

Figure 6: **Baum-Welch, Conditional Baum-Welch, and Conditional Quadratic Ascent in a Profile HMM.** The values used in the time-oriented Baum-Welch update calculation are shown in (a). Conditional Baum-Welch and Conditional Quadratic Ascent instead calculate updates in blocks of states. (b) shows the values used in the CBW or CQA update of the parameters affecting the three states (M, I, and D) associated with position 3. The CBW and CQA algorithms progress in blocks of states, using the updated parameters to recalculate the forward values as they proceed.

4. Quadratic Ascent

The Baum-Welch and Conditional Baum-Welch algorithms ascend the likelihood (or the posterior distribution of the parameters) by maximizing a surrogate function that is guaranteed to be below the target function. EM algorithms such as these are useful for maximizing complicated functions for which direct analytical or numerical maximization is impossible or impractical. While no analytical solution exists for directly maximizing the parameters of Profile HMMs, the gradient of the log-likelihood (after normalized exponential transformation) has a simple analytical solution, suggesting direct optimization methods such as gradient ascent. Pierre Baldi and Yves Chauvin derived this gradient and introduced gradient ascent procedures for Profile HMMs in their landmark 1994 papers (Baldi et al., 1994; Baldi and Chauvin, 1994). Mamitsuka (1996, 1998) introduced a closely related method that differs only in the determination of the step size.

To our knowledge, apart from the original articles, these gradient ascent methods have not been widely adopted for estimating parameters of Profile HMMs, despite the oft-noted tendency of the dominant Baum-Welch approach to prematurely converge. Our implementation of the Baldi-Chauvin methods reveals a possible explanation for the relative obscurity of gradient ascent for PHMMs: the parameters (temperature and learning rate) have optima that depend sensitively on the training sequences, and even with optimized parameters the methods perform poorly in comparison to the Baum-Welch approach. Since the appropriate step length is not analytically available, methods must be employed for exploring the step length. Unfortunately, such methods typically require repeatedly employing a costly recalculation of the target function (in this case, the forward algorithm to determine the likelihood).

Our original (unpublished) efforts to estimate parameters for PHMMs led us serendipitously to the approach that we introduce here as Quadratic Ascent: inspired by the Backprop algorithm (Rumelhart et al., 1986), and unaware of the applicability of existing methods for HMMs, we devised a conditional maximization approach based on isolating subsets of the parameters, transforming them to a normalized exponential representation to eliminate the boundary constraints, and then sampling perturbations to determine a gradient. This method determines locally optimal step sizes by fitting a quadratic function (a parabola) to the log-likelihood (as a function of the step size) and updates the parameters by climbing the gradient to the estimated peak. The Quadratic Ascent method is essentially identical to this original method, except that we replace the gradient estimation step with the Baldi-Chauvin analytical gradient.

As in Baldi et al. (1994), we transform the variables for each parameter distribution $\mathbf{\Phi}$ (supported on a simplex) to unconstrained values $\boldsymbol{\omega} \in \mathbf{\Omega}$ such that for each $\boldsymbol{\phi} \in \mathbf{\Phi}$, $\boldsymbol{\phi} = \frac{\exp(\lambda \boldsymbol{\omega})}{\sum_{\boldsymbol{\omega}' \in \mathbf{\Omega}} \exp(\lambda \boldsymbol{\omega}')}$. This transformation uniquely defines parameters $\boldsymbol{\omega} \in \mathbf{\Omega}$ for any fixed $\boldsymbol{\phi} \in \mathbf{\Phi}$, given temperature λ. We find that a temperature value of $\lambda = 1$ suffices and set it thus for the results presented below.

Baldi and Chauvin showed that the partial derivative of the log-likelihood with respect to transformed parameter $\boldsymbol{\omega}$ (corresponding to untransformed parameter $\boldsymbol{\theta}$) is

$$\frac{\partial \log \mathbb{P}(S \big| \mathbf{\Theta})}{\partial \boldsymbol{\omega}} = \lambda \sum_m \frac{N_m(\boldsymbol{\theta}) - \hat{\theta} Z_m \left(m(\boldsymbol{\theta}) \right)}{\mathbb{P}(\vec{d}_m \big| \mathbf{\Theta})},$$

where $\hat{\theta}$ is the Baum-Welch update for parameter $\boldsymbol{\theta}$, $N_m(\boldsymbol{\theta})$ is the expected number of uses of parameter $\boldsymbol{\theta}$ while generating the observed sequence at index m (using the current parameters, $\mathbf{\Theta}$) and $Z_m \left(m(\boldsymbol{\theta}) \right)$ is the sum of $N_m(\boldsymbol{\theta}')$ over all parameters $\boldsymbol{\theta}'$ of the multinomial distribution $m(\boldsymbol{\theta})$ of which $\boldsymbol{\theta}$ is a component.

Given a known step size η, each set of parameters is updated in the transformed space by adding $\eta \nabla \log \mathbb{P}(S \big| \mathbf{\Theta})$ to the transformed parameters $\mathbf{\Omega}$. The Baldi-Chauvin method, which also includes an "online" variant of this procedure to compute per-sequence parameter updates, does not specifically address the choice of η (Baldi and Chauvin, 1994). Our contribution is two-fold: first, we introduce a simple quadratic line search approach for dynamically computing the step size η for each update, and second, by employing the same "rotated" approach as is used in Conditional Baum-Welch, our Conditional Quadratic Ascent procedure isolates and updates subsets of parameters at a time (we do this for all sequences simultaneously, though an "online" variant would be a straightforward extension). Without employing dynamically computed step sizes, we find that the original fixed-step-size Baldi-Chauvin method behaves poorly in comparison to all other methods presented here (data not shown).

Our dynamic computation of step sizes depends on repeatedly recomputing the target function (the likelihood) at various step size proposals. The efficiency of this recomputation is important because it is repeatedly applied for each step size determination. In general, the time to compute the likelihood is on the same order as a complete Baum-Welch update (or a whole sweep of Conditional Baum-Welch updates). Without careful optimization, the time cost would be prohibitive for most real-world applications, since each Quadratic Ascent update would require many times more computation than a Baum-Welch update. This is especially true for Conditional Quadratic Ascent, which computes a separate step size for each update of each position-specific subset of the parameters. While technically the time complexity of Quadratic Ascent remains the same as that of Baum-Welch so long as the number of recalculations per step size determination is bounded (eg by setting a hard limit), the time required for Quadratic Ascent would be prohibitively worse than that of Baum-Welch.

Fortunately, we have discovered that for each parameter subset update step of the Conditional Quadratic Ascent algorithm, the recomputation of the likelihood can be performed in time proportional to the number of sequences rather than to their total length (as would be required in a naive implementation). This is accomplished by representing the probability of each individual sequence as a linear function of the parameter subset. So long as the parameter subset includes only single-use parameters (ie. any but Insertion state self-transitions), this linear function can be applied in constant time per sequence to recompute its sequence probability as a function of changed parameters.

In the Results section below we present results from simulation experiments in which the transition parameters are also constrained to be equal at all positions. With this constraint, the only position-specific parameters are the Match state emission parameters. Our present implementation employs QA only for the Match emission parameters (which are single-use), and alternates these updates with Baum-Welch updates applied to the parameters that are not position-specific. It is possible to apply the QA algorithm to parameters that are multi-use (such as Insertion self-transitions), but since the likelihood must be expensively recomputed with each step in the effort to isolate and ascend the local quadratic, it is slow compared to Baum-Welch.

The results demonstrate that the Conditional Quadratic Ascent method, even as applied only to the Match emission parameters, achieves higher and more robust maxima than Conditional Baum-Welch.

5. Results

5.1. Simulation Study

We now provide results from a simulation study in which we compared the Conditional Baum-Welch, (Unconditional) Baum-Welch, Conditional Quadratic Ascent, and Unconditional Quadratic Ascent algorithms across a series of data sets randomly generated from Profile HMMs. We split each set of generated sequences into a training set and a test set, and then assessed how well each algorithm performed using the probability of the test set under a Profile HMM with parameters estimated using the training set. The "true" profiles were designed to represent a conservation level of 0.5 and 0.75. For example a true profile with conservation level 0.5 assigns fifty percent probability to one (randomly determined) residue at each position, and assigns each of the remaining residues an equal share of the remaining probability. Transition parameters were set such that the expected number of insertions per generated sequence was 8 and the expected length of each insertion was 1.25% of the profile length, and likewise for deletions.

We evaluated two scenarios: for both DNA and AA sequence families we evaluated a scenario in which 100 sequences are available, and the profile length is 100. We found that for DNA, under this scenario all four evaluated methods performed well (and were indistinguishable) at a conservation level of 0.75, and all four methods overtrained and performed poorly on the test set at conservation level 0.5 (data not shown). However for AA sequences under this scenario we found that at both conservation levels, CQA and UQA outperformed CBW and UBW. In the second scenario (evaluated only for DNA sequences families), we more closely approximated the scenario of within-host HIV-1 evolution (see below). We evaluated profiles of length 1000, using only 20 sequences for parameter estimation, with conservation rates 0.75 and 0.90. In all cases we used 100 sequences for the test set.

For each of the true profiles we created nine "starting" profiles: one with "even" Match emission probabilities (eg. each residue having equal probability), four with Match emission probabilities chosen uniformly at random, and four with Match emission probabilities drawn from a symmetric Dirichlet with concentration parameter 100. We trained each true profile nine times per algorithm, once from each of the nine starting profiles. The algorithms were assessed by the probability they assigned to the training set and by the probability they assigned to the test set.

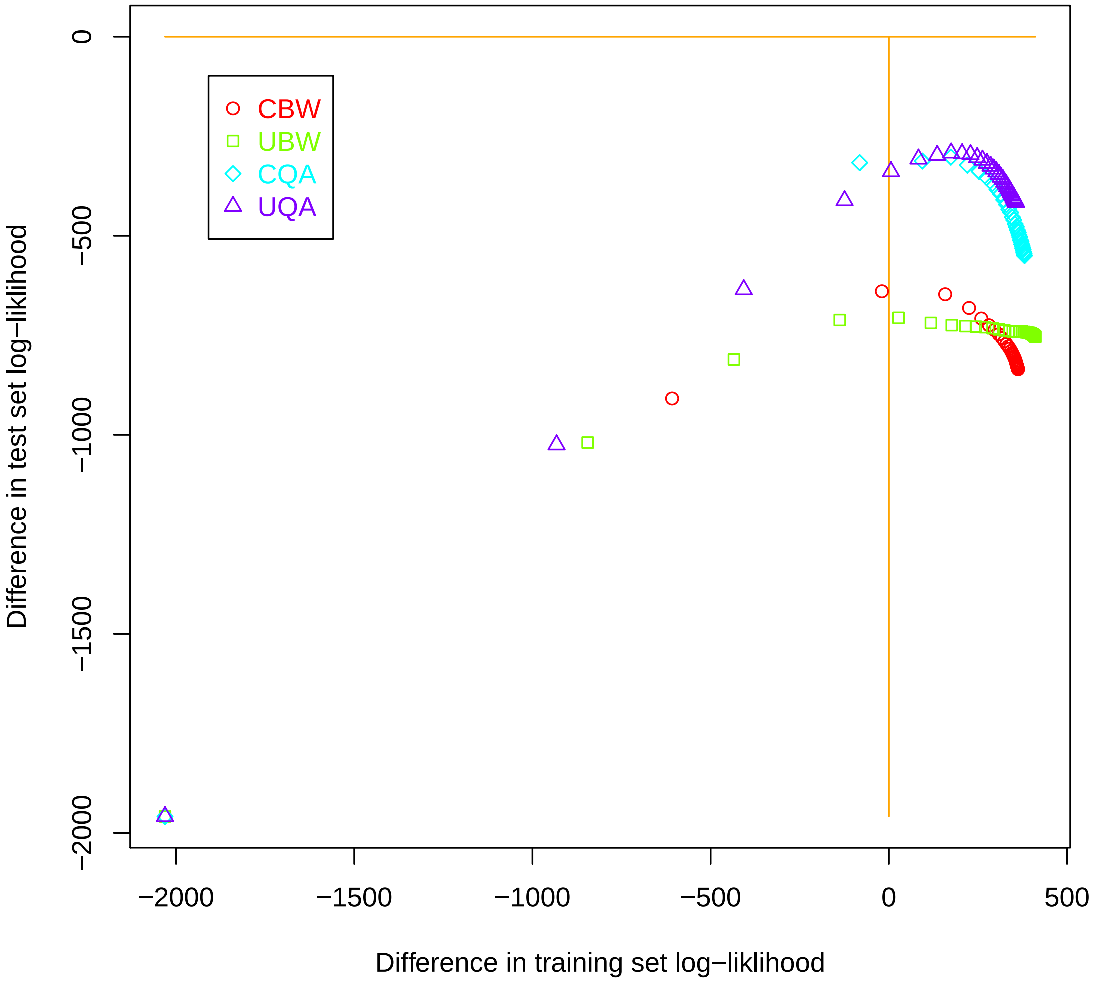


Figure 7: **AA Simulation Results for Concentration Rate 0.50.** AA results for the scenario with 100 sequences drawn from a profile of length 100 for a conservation rate of 0.50 are shown, color coded by method. The points plot the iterations of the algorithm from 0 to convergence, from left to right. The X axis shows the difference in the $\log_{10}$ likelihood of the training set data, and the Y axis shows the difference in the test set data, with differences taken to the true model from which the data were drawn. As expected there is some degree of overfitting on the training data, especially at later iterations. The Quadratic Ascent methods perform well. These plots show averages over the four starting profiles drawn from the concentrated prior; results for the "even" start are comparable, and results for the "uniform" start are consistent but somewhat poorer (data not shown).

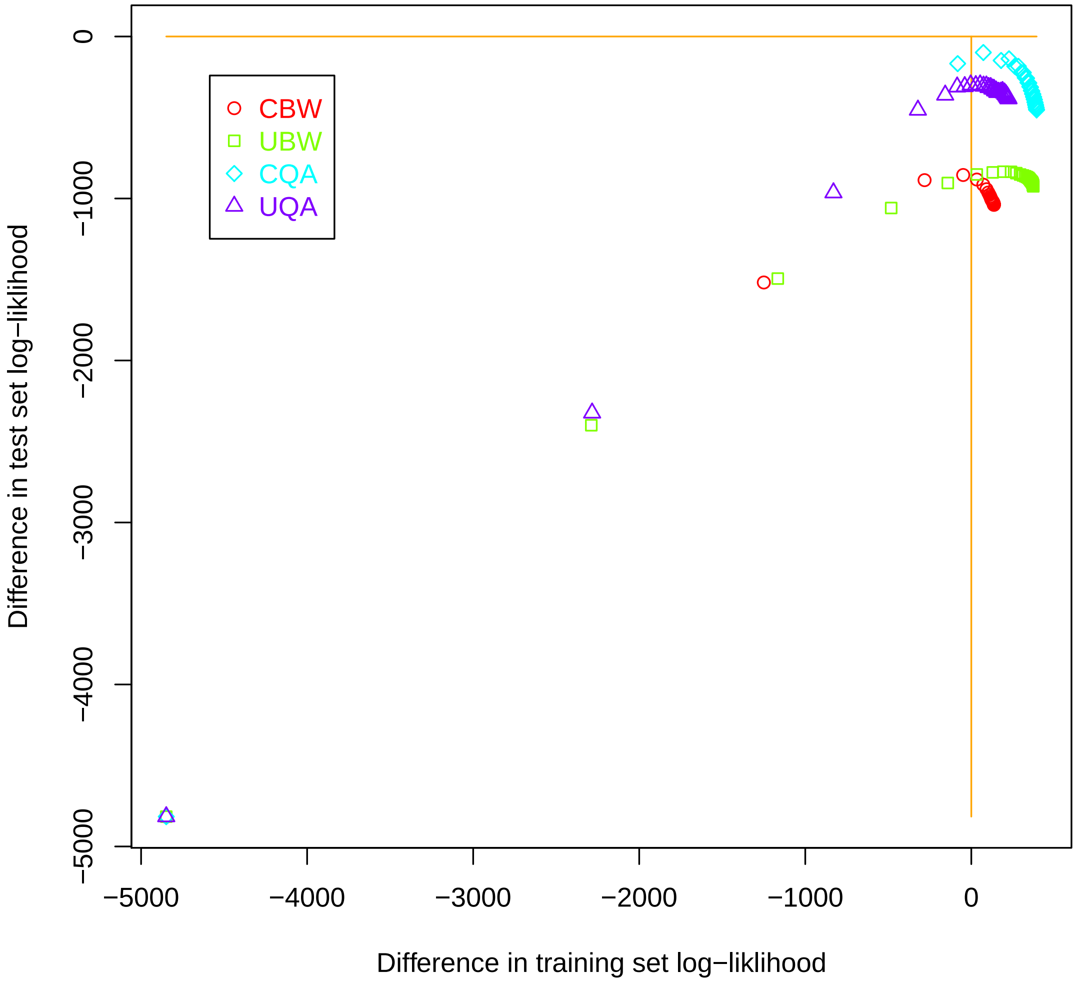


Figure 8: **AA Simulation Results for Concentration Rate 0.75.** AA results for the scenario with 100 sequences drawn from a profile of length 100 for a conservation rate of 0.75 are shown, color coded by method. The points plot the iterations of the algorithm from 0 to convergence, from left to right. The X axis shows the difference in the $\log_{10}$ likelihood of the training set data, and the Y axis shows the difference in the test set data, with differences taken to the true model from which the data were drawn. As expected there is some degree of overfitting on the training data, especially at later iterations. The Quadratic Ascent methods perform well. These plots show averages over the four starting profiles drawn from the concentrated prior; results for the "even" start are comparable, and results for the "uniform" start are consistent but somewhat poorer (data not shown).

Experience has taught us that the algorithms are not very sensitive to the starting values of the transition parameters so long as they are initially set to strongly favor Matches over gaps. Thus we drew the transition values of starting profiles from constrained uniform distributions, with the initial transition probabilities from Match states set with Match→Deletion and Match→Insertion transition probabilities each uniform over (0, 0.05). Each gap extension probability was drawn from a uniform distribution over (0, 0.5).

While it would be possible to use priors for regularization during parameter estimation (treated as pseudocounts; see Edlefsen and Liu, 2010), instead a minimum value of 1E-5 was enforced for all parameter values being trained, to prevent the algorithms from getting trapped with zero-valued parameters. The algorithms were run until convergence, with convergence defined as the average euclidean distance of all free parameters being less than 1E-5.

Code for running the simulations is available at (url: Edlefsen, 2015b) using the "Profillic" implementations of Unconditional Baum-Welch (UBW), Conditional Baum-Welch (CBW), Unconditional Quadratic Ascent (UQA), and Conditional Quadratic Ascent (CQA) (url: Edlefsen, 2015c).

Figures 7 and 8 plot the progress of the algorithms when applied to AA sequence families for the first scenario, in terms of the difference at each iteration in training set log-likelihood (on the X axis) versus the difference in the test set log-likelihood (on the Y axis), with differences taken between the estimated model and the true model that generated the data. Figures 9 and 10 show the analogous results for DNA sequence families under the second scenario. The points in the lower left coincide because all four algorithms begin with the same random starting PHMM; thereafter the points trace the algorithm's optimization path (left to right). The corresponding performance versus the true profile increases as the algorithms progress, however in every case there is evidence of some degree of overfitting, as indicated by the drop in the curves to the right of the zero line. Starting from models with position emission probabilities more concentrated around the even distribution turns out to be an important advance, as shown by contrast to Figures 11 and 12, which start with draws from a dirichlet(1,1,1,1) distribution.

5.2. HIV-1 Founder Identification

A Profile HMM represents sequences that are independent and identically distributed, which may be appropriate for endogenous and exogenous viruses that evolve with little phylogenetic structure, such as transposons and

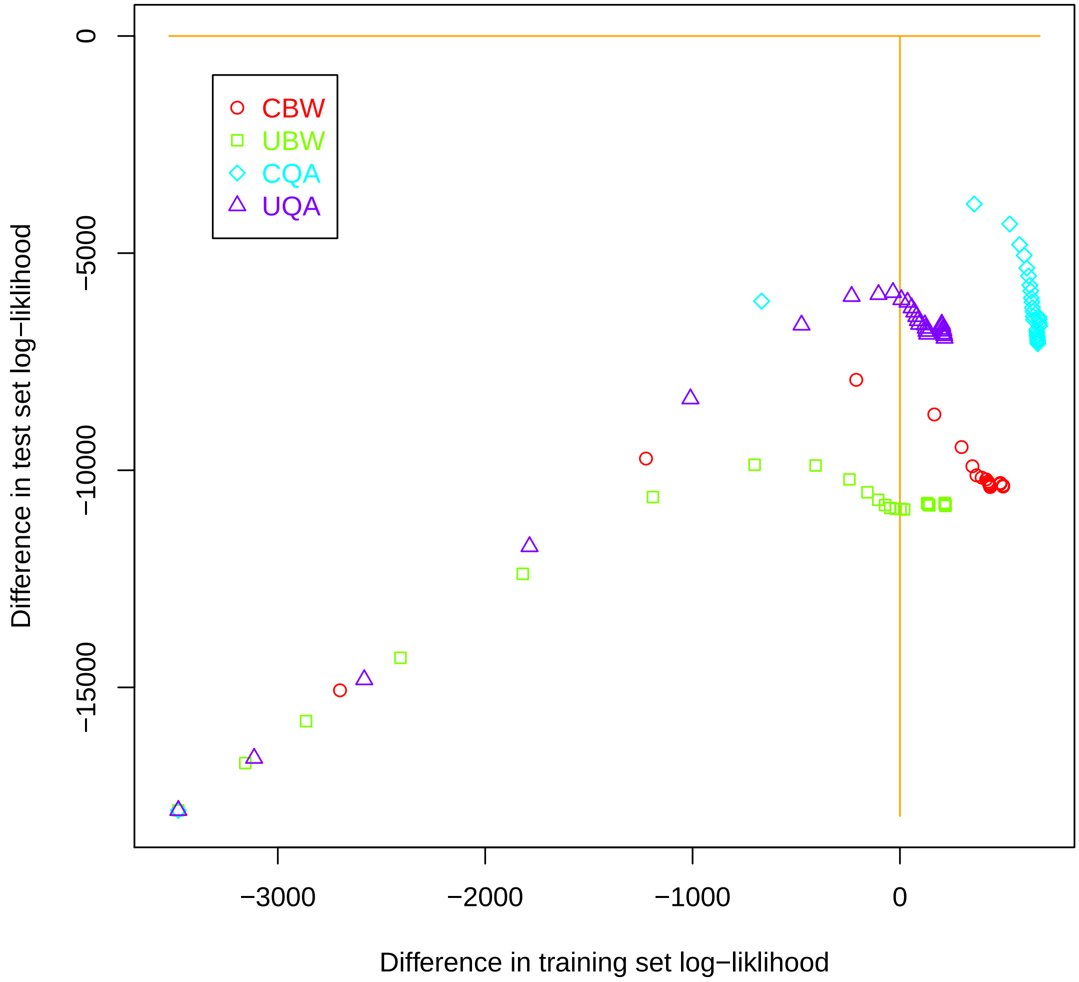


Figure 9: **DNA Simulation Results for Concentration Rate 0.75, starting from concentrated prior.** DNA results for the scenario with 20 sequences drawn from a profile of length 1000 for a conservation rate of 0.75 are shown, color coded by method. The points plot the iterations of the algorithm from 0 to convergence, from left to right. The X axis shows the difference in the $\log_{10}$ likelihood of the training set data, and the Y axis shows the difference in the test set data, with differences taken to the true model from which the data were drawn. As expected there is some degree of overfitting on the training data, especially at later iterations. The Quadratic Ascent methods perform relatively well.

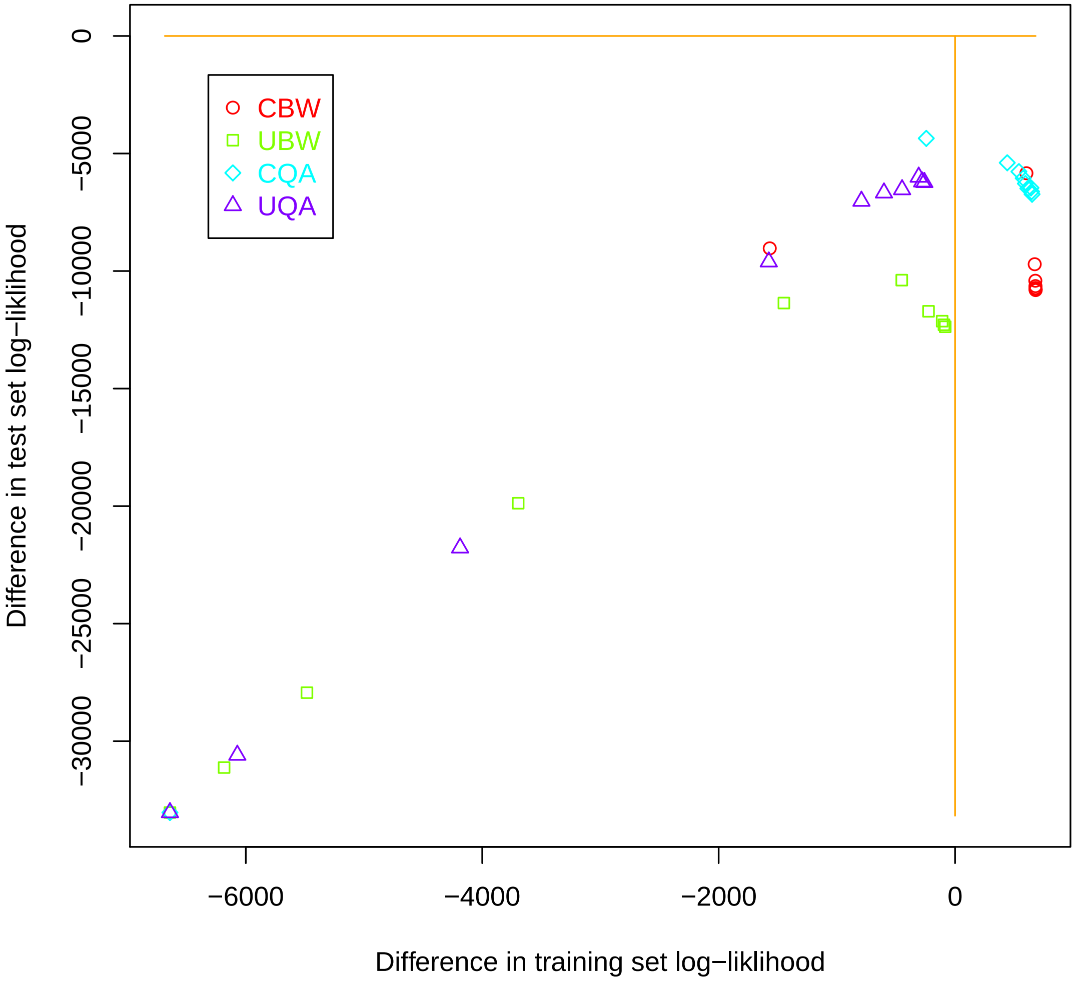


Figure 10: **DNA Simulation Results for Concentration Rate 0.90, starting from concentrated prior.** DNA results for the scenario with 20 sequences drawn from a profile of length 1000 for a conservation rate of 0.90 are shown, color coded by method. The points plot the iterations of the algorithm from 0 to convergence, from left to right. The X axis shows the difference in the $\log_{10}$ likelihood of the training set data, and the Y axis shows the difference in the test set data, with differences taken to the true model from which the data were drawn. As expected there is some degree of overfitting on the training data, especially at later iterations. The Quadratic Ascent methods perform relatively well, although the inflection point of the CBW and UBW traces are obscured between iterations.

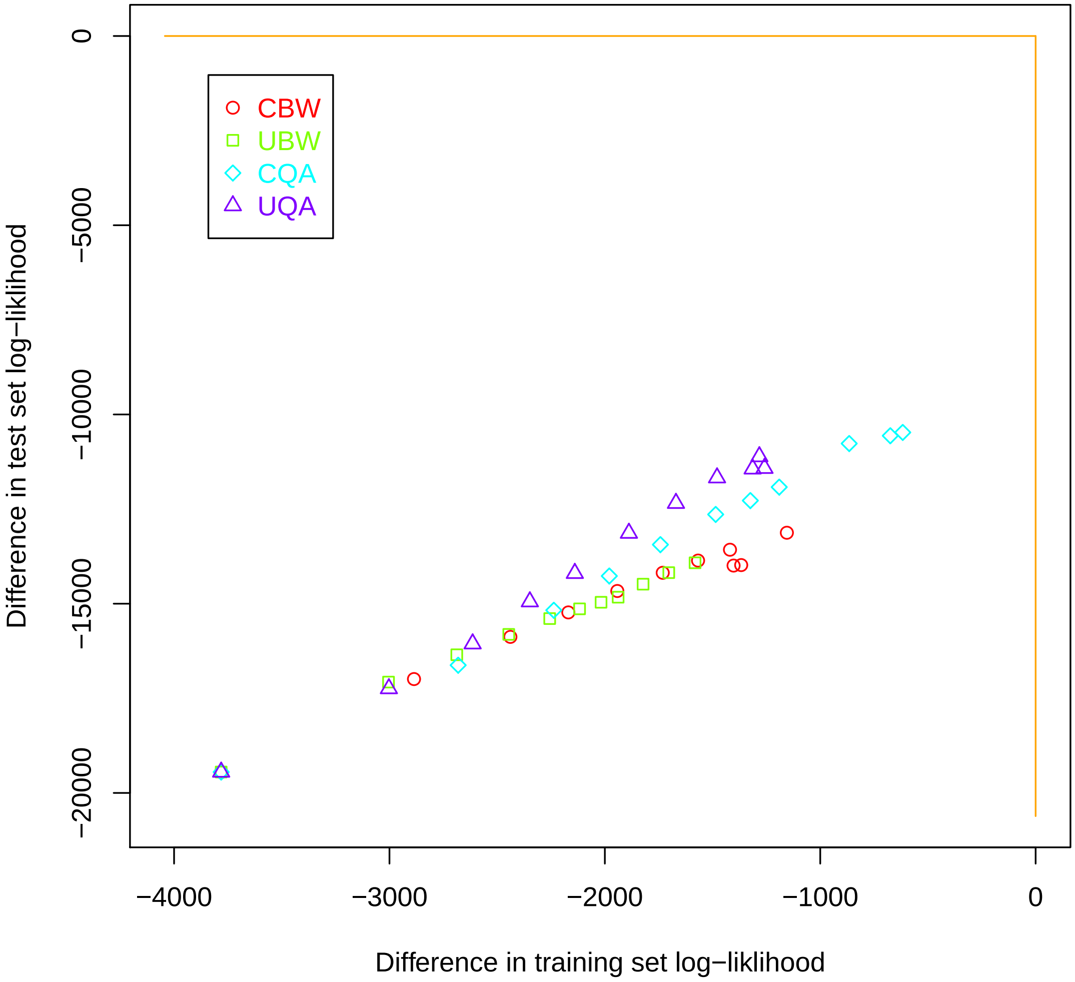


Figure 11: **DNA Simulation Results for Concentration Rate 0.75, starting from uniform positions.** DNA results for the scenario with 20 sequences drawn from a profile of length 1000 for a conservation rate of 0.75 are shown, color coded by method. The points plot the iterations of the algorithm from 0 to convergence, from left to right. The X axis shows the difference in the $\log_{10}$ likelihood of the training set data, and the Y axis shows the difference in the test set data, with differences taken to the true model from which the data were drawn. Starting from uniform positions appears to trap all of the algorithms in local optima.

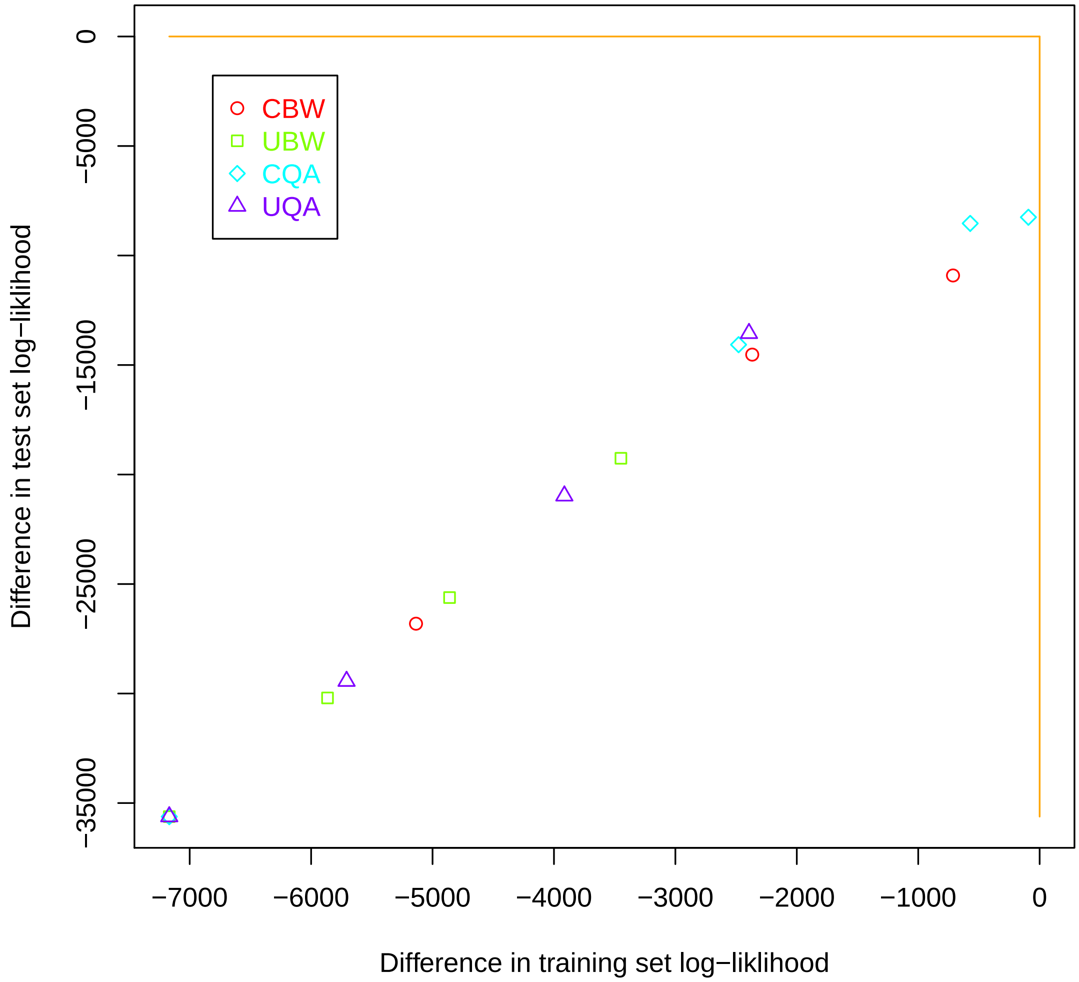


Figure 12: **DNA Simulation Results for Concentration Rate 0.90, starting from concentrated uniform positions.** DNA results for the scenario with 20 sequences drawn from a profile of length 1000 for a conservation rate of 0.90 are shown, color coded by method. The points plot the iterations of the algorithm from 0 to convergence, from left to right. The X axis shows the difference in the $\log_{10}$ likelihood of the training set data, and the Y axis shows the difference in the test set data, with differences taken to the true model from which the data were drawn. Starting from uniform positions appears to trap all of the algorithms in local optima.

.

RNA virus families (sequence weighting strategies to account for subfamily structure have been proposed; see Durbin et al., 1998). Here we discuss its application to the modeling of subpopulations of HIV-1 *in vivo*. During acute HIV-1 infection (in the first few weeks post-acquisition), HIV-1 viruses replicate at a high rate. Due to the low fidelity replication mechanism of HIV-1, new errors are introduced with each replication cycle. The sequences diversify primarily in a manner described well by a Poisson distribution (Giorgi et al., 2010), with little linkage across sites, because constant recombination among a "quasispecies cloud" of viral particles limits linkage (Zanini et al., 2015; Wu et al., 2014). This results in a sequence distribution that is well represented by a Profile HMM. However, in some fraction of infections there are multiple distinguishable founders of infection; estimates range from 20-30% of infections and vary by exposure route (Gounder et al., 2015; Sterrett et al., 2014; Li et al., 2010; Gottlieb et al., 2008). For some time these populations evolve independently, each with its own star-like phylogeny. Methods have been proposed for evaluating the impact of preventative vaccination by accounting for a possible reduction in the number of founders, rather than simply evaluating the binary endpoint of acquisition (Follmann and Huang, 2015). Furthermore, the evaluation of so-called "acquisition sieve effects" (indication of a genome-specific effect of the vaccine to prevent infection against some but not all HIV-1 viruses) requires reconstruction of the founding viruses (Edlefsen et al., 2013). Determining the number of founders is an important part of vaccine evaluation, and several authors have described methods for counting founders using a combination of automated and manual evaluation techniques (Keele et al., 2008; Herbeck et al., 2011; Rossenkhan et al., 2012), though no fully automated (and hence reproducible) method has yet been described.

Here we evaluate a simple approach to estimating HIV-1 infection founder multiplicity that leverages a unique feature of Profile HMMs that have been estimated using the techniques described here (that is, maximum likelihood or maximum a posteriori parameters): at convergence, the Baum-Welch update vectors for each participant mutually cancel (Edlefsen, 2015a). This property is guaranteed for all of the methods evaluated here, even the (C)QA methods because at convergence the gradient of the log-likelihood is zero, and Baldi et al. (1994) showed that this gradient is also the gradient of the Baum-Welch EM Q function. The per-sequence Baum-Welch update vectors are the conditional expected values of the PHMM's parameters, given that the model produced just that sequence, and so these have the same structure as the Profile HMM.

We call these per-sequence update vectors “alignment profiles” as they represent averages over all possible paths through the latent states of the model, which corresponds to all of the ways to align the sequence to the model. Under the *iid* assumptions of the PHMM, these alignment profiles should be uncorrelated vectors. Violations of the assumption are reflected in unexpected clustering among the vectors, and here we use this insight to implement a method for counting the number of founders, and ultimately for modeling each founder quasispecies population as a separate PHMM.

Our approach is straightforward, and is meant to demonstrate the utility of true statistical estimation of PHMM parameters rather than to be a final solution to the founder identification problem. For each subject’s set of (unaligned) sampled sequences, we estimate the parameters of a Profile HMM using CQA. We use this Profile HMM to estimate one alignment profile per sequence. We cluster these vectors (by performing UPGMA clustering on the matrix of Euclidean distances between them) and then cut the tree using the “Dynamic Tree Cut” method of Langfelder et al. (2008), using default options. The first cut is at a height of it 99% of the range between the 5th percentile and the maximum of the joining heights on the dendrogram, regardless of the heterogeneity of the sequences, and will thus always result in at least two clusters. To accept a call of greater than 1 founder we also require a minimum PHMM entropy (summed over the PHMM) of 20. This is a Profile HMM analogue of the requirement of a minimum sequence diversity commonly applied in the founder identification literature, but differs in that the entropy of the PHMM accounts for alignment uncertainty.

The LANL HIV sequence database (url: Los Alamos National Labs, 2015a) provides curated “special purpose” nucleotide alignments of HIV-1 sampled from acute infections, including a set of 1505 HIV-1 Clade C env nucleotide sequences collected from 69 patients (described in Abrahams et al., 2009). We follow the authors’ suggestion to pre-filter the sequences to exclude any found to have signatures of “hypermutation”, a well-described effect of a human anti-viral mechanism (we use Rose and Korber, 2000) or recombination (using RAPBeta, url: Los Alamos National Labs, 2015b). We prepare the sequences by stripping the gaps (unaligning them). We evaluate our ability to detect founders by comparing our estimated number of founders for each subject against those calculated by the study’s authors.

Tables 1, 2, and 3 show the results of this experiment, indicating that in this case the automated founder detection works quite well. The tables show,

Participant	Fiebig	nSeq	nHyper	nRcmb	Pent	Pclst	Ours	Theirs
0089	5	22	0	0	6.76	3	1	1
0114	4	27	0	4	65.50	3	3	3
0334	1-2	22	0	0	5.02	2	1	1
0393	4	22	0	0	5.57	2	1	1
0478	1-2	23	0	3	70.71	2	2	3
0595	4	28	0	0	6.93	6	1	1
0626	4	24	0	0	5.57	2	1	1
0665	4	20	0	0	6.86	2	1	1
0682	1-2	22	0	0	4.73	2	1	1
0985	1-2	23	0	0	4.61	2	1	1
1086	1-2	24	0	0	8.49	2	1	1
1172	1-2	20	0	0	4.90	2	1	1
1176	1-2	21	0	0	9.27	3	1	3
1196	1-2	23	2	0	32.78	2	2	3
1335	4	21	0	1	85.85	2	2	3
1373	1-2	22	0	0	6.36	3	1	1
1394	1-2	20	0	0	6.40	3	1	1
2010	4	23	0	0	6.78	2	1	1
2052	1-2	23	0	0	5.46	2	1	1
2060	1-2	22	0	0	5.55	2	1	1
2103	1-2	20	0	0	4.83	2	1	1

Table 1: **Profillic Founder Identification Using CQA: Table 1 of 3**

for each of the 69 participants studied, the time at which their acute infection was sampled (as Fiebig stages Fiebig et al., 2003), the number of sequences sampled, the number of these sequences excluded because they were found to have hypermutation or recombination, and the statistics computed after Profillic HMM analysis. The final two columns show the comparison of the reported number of founders from our analysis and those of the original study. The Profillic HMM statistics are the "entropy", which is actually the sum of entropy values over all of the PHMM parameters after parameter estimation using the default parameter options of the Profillic implementation of CQA for DNA sequence families, and the number of clusters obtained from the clustering of update vectors.

Participant	Fiebig	nSeq	nHyper	nRcmb	Pent	Pclst	Ours	Theirs
703010010	2	22	0	0	21.40	3	3	3
703010054	5-6	27	0	0	15.91	2	1	1
703010131	4	22	0	0	6.10	2	1	1
703010159	2	20	0	0	6.89	5	1	1
703010193	4	24	1	0	6.75	2	1	1
703010200	4	18	0	8	57.09	3	3	3
703010217	5-6	25	0	0	6.63	2	1	1
703010228	4	28	0	3	53.28	2	2	2
704010017	5-6	25	0	0	14.24	3	1	1
704010042	4	42	0	0	7.80	2	1	1
704010056	5-6	21	0	1	17.73	4	1	1
704010069	4	23	0	0	9.71	2	1	1
704010083	2	24	0	0	6.15	2	1	1
704809221	1-2	28	0	0	6.85	2	1	1
704810053	1-2	20	0	0	7.61	2	1	1
704810015	x	23	0	0	9.84	3	1	1
705010026	x	23	0	0	7.79	3	1	1
705010078	4	26	0	0	5.99	3	1	1
705010110	x	20	0	0	8.98	6	1	1
706010018	4	23	0	0	15.52	2	1	1
706010151	6	15	0	7	53.90	2	2	2
706010164	4	19	0	0	5.23	2	1	1

Table 2: **Profillic Founder Identification Using CQA: Table 2 of 3**

Participant	Fiebig	nSeq	nHyper	nRcmb	Pent	Pclst	Ours	Theirs
CAP129	4	19	0	0	5.20	2	1	1
CAP136	5	16	0	2	28.86	2	2	2
CAP174	5	21	0	0	6.37	2	1	1
CAP177	1-2	20	0	0	4.97	2	1	1
CAP188	1-2	22	0	0	5.51	2	1	1
CAP200	5	18	0	0	5.77	2	1	1
CAP206	5	21	0	0	6.30	2	1	1
CAP210	1-2	21	0	0	4.35	2	1	1
CAP217	5	20	0	0	4.88	2	1	1
CAP220	1-2	15	0	0	7.61	2	1	1
CAP221	1-2	21	0	0	7.10	2	1	1
CAP222	1-2	21	0	0	54.08	3	3	3
CAP224	5	19	0	0	21.18	2	2	2
CAP225	3	20	0	0	6.85	2	1	1
CAP237	3	22	0	0	5.87	2	1	1
CAP239	5	24	0	0	6.49	3	1	1
CAP260	5	18	0	5	47.68	2	2	2
CAP269	6	18	0	0	16.37	3	1	1
CAP37	4	20	0	2	67.29	3	3	3
CAP40	6	22	1	0	12.04	2	1	1
CAP45	1-2	16	0	0	4.81	2	1	1
CAP63	3	19	1	0	5.93	2	1	1
CAP69	1-2	20	0	9	61.61	3	3	5
CAP8	5	62	0	1	13.93	3	1	1
CAP84	4	22	1	0	5.67	1	1	1
CAP85	6	21	2	0	11.89	2	1	1

Table 3: **Profillic Founder Identification Using CQA: Table 3 of 3**

6. Discussion

Here we have presented a new algorithm for estimating the parameters of a Profile HMM. While a related method was previously proposed (Baldi and Chauvin, 1994), we describe an efficient procedure for calculation of the step size parameter, which is necessary for deployment of the approach. We show the first evidence of comparable performance as compared with the Baum-Welch algorithm and its conditional variant. We also show that the same "rotated" orientation to the algorithm allows for a conditional ascent procedure yields advantages over unconditional maximization, consistent with what was shown previously for Conditional Baum-Welch (Edlefsen and Liu, 2010). Through simulation studies we have shown that new algorithm, Conditional Quadratic Ascent, performs comparably in most scenarios to CBW, however its performance in large-sequence, low sample size settings demonstrates an improvement over that method.

The Profile Hidden Markov model represents a good candidate model for within-host HIV-1 sequence diversity within a single quasispecies cloud, because such clouds of constantly recombining variants are expected to be distributed as approximately independent representatives from a plateau in the fitness landscape (Domingo, 2002; Vignuzzi et al., 2006). Here we have demonstrated a property of true maximization of the parameters of a PHMM when the *iid* assumption holds: at convergence, the update vectors cancel and are expected to be independent and identically distributed. We introduced the notion of clustering these "alignment profiles" to detect violations from *iid*, and we showed that this can be productively applied to the important scientific aim of identifying distinct viral founder populations in acute HIV-1 infection.

The results of our simple method agreed with the founder multiplicity in most instances. Apart from participant "1176", using a threshold of PHMM entropy of 20 classified the participants into single and multiple founder variants in agreement with those of Abrahams et al. (2009). The authors noted that this participant had an "infection with 3 closely related viruses," perhaps explaining the relatively low entropy. We chose the value 20 to optimize agreement with their method; as Figure 13 shows, this represents one of multiple gaps in the distribution. However, these entropy values are not scaled for the length of the model (which here was around 2500 positions), and further work is needed to determine a general strategy for setting this threshold from the data alone, ideally one that is robust to variations in

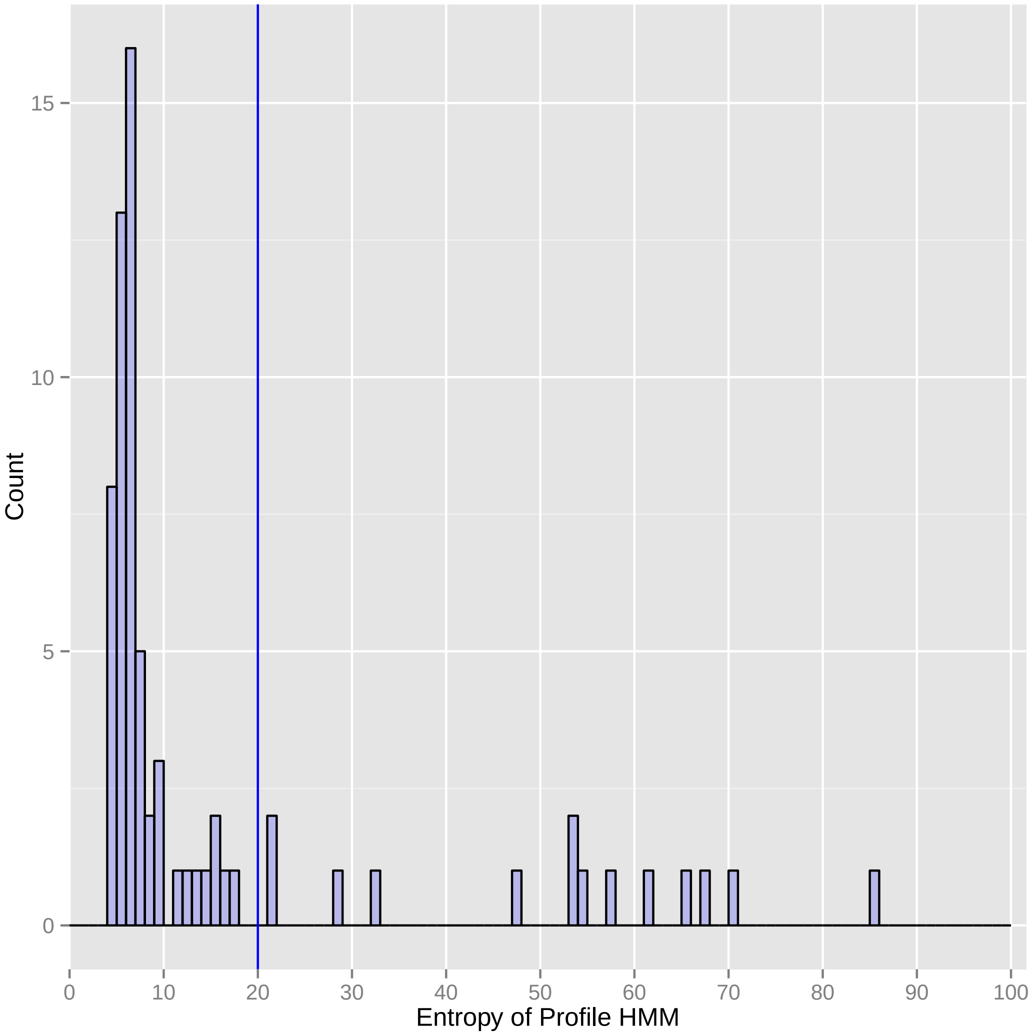


Figure 13: **Histogram of Profile HMM entropy values.** These are the sum of individual entropy values over individual multinomial distributions for the parameters of the PHMM, and are thus model-length-dependent. The value of 20 gave the best agreement with the results of Abrahams et al. (2009).

HIV-1 clade, Fiebig stage, model length and genomic region. Among the remaining 14 participants that were identified by both methods as multiple-founder variants, there was agreement on 10. In the simple analysis we conducted, the remaining four participants were assigned lower estimates by our method (by one in three cases: two instead of three; and by two in the case of CAP69: three instead of five). It turns out that for CAP69, two of the five "founders" were unique recombinants of the other three founders (C. Williamson, personal communication). Thus that particular discrepancy is due to the fact that we removed recombinants before conducting our analysis. All four of these participants' viral populations exhibited recombination or hypervariation, so the other discrepancies may have the same source, or may simply be due to the relatively simple method that we used for clustering. Further work is needed to develop the method into a full-fledged automated procedure for HIV-1 founder identification.

7. Acknowledgments

The author gratefully acknowledges the early contributions to the development of this method by Andrew F. Siegel, Eugene Kolker, and Robert Hubley.

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Numbers R37AI054165 and UM1AI068635, and by the Bill and Melinda Gates Foundation Award Number OPP1110049. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or BMGF.

Abrahams, M.-R., Anderson, J. A., Giorgi, E. E., Seoighe, C., Mlisana, K., Ping, L.-H., Athreya, G. S., Treurnicht, F. K., Keele, B. F., Wood, N., Salazar-Gonzalez, J. F., Bhattacharya, T., Chu, H., Hoffman, I., Galvin, S., Mapanje, C., Kazembe, P., Thebus, R., Fiscus, S., Hide, W., Cohen, M. S., Karim, S. A., Haynes, B. F., Shaw, G. M., Hahn, B. H., Korber, B. T., Swanstrom, R., Williamson, C., for the CAPRISA Acute Infection Study Team, the Center for HIV-AIDS Vaccine Immunology Consortium, 04 2009. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype c reveals a non-poisson distribution of transmitted variants. Journal of Virology 83 (8), 3556–3567.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2663249/

Baldi, P., Chauvin, Y., 1994. Smooth on-line learning algorithms for hidden markov models. Neural Computation 6 (2), 307–318.

Baldi, P., Chauvin, Y., Hunkapiller, T., McClure, M. A., 1994. Hidden markov models of biological primary sequence information. Proceedings of the National Academy of Sciences 91 (3), 1059–1063.

Baum, L., 1972. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3, 1–8.

Churchill, G. A., 1989. Stochastic models for heterogeneous DNA sequences. Bulletin of Mathematical Biology 51 (1), 79–94.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1), 1–38.

Domingo, E., 2002. Quasispecies theory in virology. Journal of Virology 76 (1), 463–465.
URL `http://jvi.asm.org/content/76/1/463.short`

Durbin, R., Eddy, S. R., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.

Eddy, S., 1995. Multiple alignment using hidden Markov models. In: Proc Int Conf Intell Syst Mol Biol. Vol. 3. pp. 114–120.

Eddy, S., School of Medicine and Dept. of Genetics and National Human Genome Research Institute (US), 1992. HMMER profile hidden Markov models for biological sequence analysis. Washington University School of Medicine.

Eddy, S., Wheeler, T., 2013. Hmmer user's guide, version 3.1, http://hmmer.janelia.org/.

Edlefsen, P. T., May 2015a. Poster Abstract: Identifying within-host HIV-1 subpopulations by cosegregation of Profile Hidden Markov Model update vectors. HIV Dynamics and Evolution.

Edlefsen, P. T., 2015b. Profillic Simulation Software, https://github.com/tedholzman/profillicSimulation.

Edlefsen, P. T., 2015c. Profillic Software, https://github.com/galosh/profillic.

Edlefsen, P. T., Gilbert, P. B., Rolland, M., 2013. Sieve analysis in hiv-1 vaccine efficacy trials. Current opinion in HIV and AIDS 8 (5), 432–436.

Edlefsen, P. T., Liu, J. S., 2010. Transposon identification using Profile HMMs. BMC genomics 11 (Suppl 1), S10.

Fiebig, E. W., Wright, D. J., Rawal, B. D., Garrett, P. E., Schumacher, R. T., Peddada, L., Heldebrant, C., Smith, R., Conrad, A., Kleinman, S. H., et al., 2003. Dynamics of hiv viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary hiv infection. Aids 17 (13), 1871–1879.

Finn, R. D., Clements, J., Eddy, S. R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research.

Follmann, D., Huang, C.-Y., 2015. Incorporating founder virus information in vaccine field trials. Biometrics 71 (2), 386–396.
URL `http://dx.doi.org/10.1111/biom.12277`

Gaschen, B., Kuiken, C., Korber, B., Foley, B., 2001. Retrieval and on-the-fly alignment of sequence fragments from the hiv database. Bioinformatics 17 (5), 415–418.
URL `http://bioinformatics.oxfordjournals.org/content/17/5/415.abstract`

Giorgi, E. E., Funkhouser, B., Athreya, G., Perelson, A. S., Korber, B. T., Bhattacharya, T., 2010. Estimating time since infection in early homogeneous hiv-1 samples using a poisson model. BMC Bioinformatics 11, 532–532.
URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2975664/`

Gottlieb, G. S., Heath, L., Nickle, D. C., Wong, K. G., Leach, S. E., Jacobs, B., Gezahegne, S., van 't Wout, A. B., Jacobson, L. P., Margolick, J. B., Mullins, J. I., Apr 2008. Hiv-1 variation before seroconversion in men who have sex with men: analysis of acute/early hiv infection in the multicenter aids cohort study. J Infect Dis 197 (7), 1011–1015.

Gounder, K., Padayachi, N., Mann, J. K., Radebe, M., Mokgoro, M., van der Stok, M., Mkhize, L., Mncube, Z., Jaggernath, M., Reddy, T., Walker, B. D., Ndung'u, T., 2015. High frequency of transmitted hiv-1 gag hla class i-driven immune escape variants but minimal immune selection over the first year of clade c infection. PLoS ONE 10 (3), e0119886.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4363590/

Herbeck, J. T., Rolland, M., Liu, Y., McLaughlin, S., McNevin, J., Zhao, H., Wong, K., Stoddard, J. N., Raugi, D., Sorensen, S., Genowati, I., Birditt, B., McKay, A., Diem, K., Maust, B. S., Deng, W., Collier, A. C., Stekler, J. D., McElrath, M. J., Mullins, J. I., 08 2011. Demographic processes affect hiv-1 evolution in primary infection before the onset of selective processes. Journal of Virology 85 (15), 7523–7534.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3147913/

Hoen, D., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., Kapusta, A., Lerat, E., Maumus, F., Pollock, D., Quesneville, H., Smit, A., Wheeler, T., Bureau, T., Blanchette, M., 2015. A call for benchmarking transposable element annotation methods. Mobile DNA 6 (1).
URL http://dx.doi.org/10.1186/s13100-015-0044-6

Hughey, R., Krogh, A., 1995. Sam: Sequence alignment and modeling software system, santa Cruz: University of California.

Hughey, R., Krogh, A., 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Bioinformatics 12 (2), 95–107.

Hunter, D., Lange, K., 2004. A tutorial on MM algorithms. The American Statistician 58 (1), 30–37.

Jurka, J., Kapitonov, V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110 (1-4), 462–467.

Kececioglu, J., Kim, E., Wheeler, T., 2010. Aligning protein sequences with predicted secondary structure. Journal of Computational Biology 17 (3), 561–580.

Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., Sun, C., Grayson, T., Wang, S., Li, H., Wei,

X., Jiang, C., Kirchherr, J. L., Gao, F., Anderson, J. A., Ping, L.-H., Swanstrom, R., Tomaras, G. D., Blattner, W. A., Goepfert, P. A., Kilby, J. M., Saag, M. S., Delwart, E. L., Busch, M. P., Cohen, M. S., Montefiori, D. C., Haynes, B. F., Gaschen, B., Athreya, G. S., Lee, H. Y., Wood, N., Seoighe, C., Perelson, A. S., Bhattacharya, T., Korber, B. T., Hahn, B. H., Shaw, G. M., 05 2008. Identification and characterization of transmitted and early founder virus envelopes in primary hiv-1 infection. Proceedings of the National Academy of Sciences of the United States of America 105 (21), 7552–7557.
URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2387184/`

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., Haussler, D., 1994. Hidden Markov models in computational biology : Applications to protein modeling. Journal of Molecular Biology 235 (5), 1501–1531.

Langfelder, P., Zhang, B., Horvath, S., 2008. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. Bioinformatics 24 (5), 719–720.
URL `http://bioinformatics.oxfordjournals.org/content/24/5/719.abstract`

Lerat, E., Sémon, M., 2007. Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. Gene 396 (2), 303 – 311.

Li, H., Bar, K. J., Wang, S., Decker, J. M., Chen, Y., Sun, C., Salazar-Gonzalez, J. F., Salazar, M. G., Learn, G. H., Morgan, C. J., Schumacher, J. E., Hraber, P., Giorgi, E. E., Bhattacharya, T., Korber, B. T., Perelson, A. S., Eron, J. J., Cohen, M. S., Hicks, C. B., Haynes, B. F., Markowitz, M., Keele, B. F., Hahn, B. H., Shaw, G. M., May 2010. High multiplicity infection by hiv-1 in men who have sex with men. PLoS Pathog 6 (5), e1000890.

Liu, J. S., Neuwald, A. F., Lawrence, C. E., mar 1999. Markovian structures in biological sequence alignments. Journal of the American Statistical Association 94 (445), 1–15.

Los Alamos National Labs, 2015a. LANL HIV-1 Sequence Database, http://www.hiv.lanl.gov.

Los Alamos National Labs, 2015b. RAP Beta, http://www.hiv.lanl.gov/content/sequence/RAP/RAP.html.

Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., Hein, J., 2008. Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. Genome Research 18 (2), 298–309.

Mamitsuka, H., 1996. A learning method of hidden markov models for sequence discrimination. Journal of Computational Biology 3 (3), 361–373.

Mamitsuka, H., 1998. Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. Proteins Structure Function and Genetics 33 (4), 460–474.

Medstrand, P., van de Lagemaat, L. N., Mager, D. L., 2002. Retroelement Distributions in the Human Genome: Variations Associated With Age and Proximity to Genes. Genome Research 12 (10), 1483–1495.

Meng, X., van Dyk, D., 1997. The EM algorithm–an old folk-song sung to a fast new tune. Journal of the Royal Statistical Society. Series B (Methodological), 511–567.

Meng, X.-L., Rubin, D. B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80 (2), 267–278.

Rabiner, L. R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77 (2), 257–286.

Rose, P. P., Korber, B. T., 2000. Detecting hypermutations in viral sequences with an emphasis on g a hypermutation. Bioinformatics 16 (4), 400–401.

Rossenkhan, R., Novitsky, V., Sebunya, T. K., Musonda, R., Gashe, B. A., Essex, M., 2012. Viral diversity and diversification of major non-structural genes vif, vpr, vpu, tat exon 1 and rev exon 1 during primary hiv-1 subtype c infection. PLoS ONE 7 (5), e35491.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3348911/

Rumelhart, D., Hintont, G., Williams, R., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.

Scott, S., 2002. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. Journal of the American Statistical Association 97 (457), 337–352.

Smit, A., Hubley, R., Green, P., 1996-2004. Repeatmasker Open-3.0, http://www.repeatmasker.org.

Sterrett, S., Learn, G. H., Edlefsen, P. T., Haynes, B. F., Hahn, B. H., Shaw, G. M., Bar, K. J., 09 2014. Low multiplicity of hiv-1 infection and no vaccine enhancement in vax003 injection drug users. Open Forum Infectious Diseases 1 (2), ofu056.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281816/

Venter, J., Jan 2001. Initial sequencing and analysis of the human genome. Nature 409 (6822), 860–921.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., Andino, R., 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439 (7074), 344–348.

Wheeler, T., Clements, J., Eddy, S., Hubley, R., Jones, T., Jurka, J., Smit, A., Finn, R., 2013. Dfam: a database of repetitive dna based on profile hidden markov models. Nucleic Acids Res 41, D70–D82.

Wheeler, T., Eddy, S., 2013. nhmmer: Dna homology search with profile hmms. Bioinformatics 29, 2487–2489.

Wong, K. M., Suchard, M. A., Huelsenbeck, J. P., 2008. Alignment uncertainty and genomic analysis. Science 319 (5862), 473–476.

Wu, N. C., De La Cruz, J., Al-Mawsawi, L. Q., Olson, C. A., Qi, H., Luan, H. H., Nguyen, N., Du, Y., Le, S., Wu, T.-T., Li, X., Lewis, M. J., Yang, O. O., Sun, R., 2014. Hiv-1 quasispecies delineation by tag linkage deep sequencing. PLoS ONE 9 (5), e97505.
URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4026136/

Zanini, F., Brodin, J., Thebo, L., Lanz, C., Bratt, G., Albert, J., Neher, R. A., September 2015. Population genomics of intrapatient hiv-1 evolution, arXiv:1509.02483v1 [q-bio.PE] 8 Sep 2015.