Parametric inference of recombination in HIV genomes

Niko Beerenwinkel, Colin N. Dewey,
and Kevin M. Woods

December 8, 2005

Abstract

Recombination is an important event in the evolution of HIV. It affects the global spread of the pandemic as well as evolutionary escape from host immune response and from drug therapy within single patients. Comprehensive computational methods are needed for detecting recombinant sequences in large databases, and for inferring the parental sequences.

We present a hidden Markov model to annotate a query sequence as a recombinant of a given set of aligned sequences. Parametric inference is used to determine all optimal annotations for all parameters of the model. We show that the inferred annotations recover most features of established hand-curated annotations. Thus, parametric analysis of the hidden Markov model is feasible for HIV full-length genomes, and it improves the detection and annotation of recombinant forms.

All computational results, reference alignments, and C++ source code are available at `http://bio.math.berkeley.edu/recombination/`.

1 Introduction

Retroviral recombination is a significant contributor to genetic variation in HIV-1 genomes [29]. When an individual is infected by two or more different strains of HIV-1, recombination can yield new forms of the virus that are mosaics of the original strains. Given a particular viral genome, we would like to determine whether it was formed by recombination and, if so, what *parental strains* recombined to form it. This determination is important in studying the geographic epidemiology of HIV-1 [21], and in the intra-patient evolution of immune escape and of drug resistance in response to therapy [15, 36]. In this paper, we present a method for inferring the parental strains of a recombinant genome. This method relies on a particular hidden Markov model (HMM) and involves statistical inference for *all* choices of model parameters.

Many models have been suggested for *parental inference*, i.e., for the identification and characterization of recombinant sequences. All rely, at some point, on a set of parameters. However, small differences in the parameter values may

result in substantially different predictions by the model. This is also true of the model we present, but we will use the technique of *parametric inference* to ensure that we have a complete understanding of how choices for the parameter values affect predictions.

Given a multiple alignment of a recombinant viral genome with its possible parental genomes, we use a probabilistic HMM to predict, for each position in the recombinant genome, the parental strain that gave rise to it. Our HMM is motivated by the *copy-choice model* [3], which gives one possible biological mechanism for viral recombination. This model is based on the fact that recombination in retroviruses results from two RNA molecules being packaged in one virion [10]. In multiply infected cells, two distinct strains may be packed into a single virion [13]. In the copy-choice model, a mosaic DNA molecule results from reverse transcriptase jumping between two different RNA templates during reverse transcription in the subsequently infected cell.

We analyze our HMM over all choices of parameters using the parametric inference technique, which is a general method for evaluating graphical models. For example, parametric analysis has been used successfully for pairwise sequence alignment [8, 23, 35, 37]. In this paper, we describe the methods of parametric inference as they are applied to our recombination HMM. To evaluate our parametric method, we use it to infer the parental subtypes of HIV-1 circulating recombinant forms (CRFs) obtained from the Los Alamos HIV Sequence Database.

We show that a simple HMM, combined with the ability to evaluate the model over its entire parameter space, is effective in predicting parental subtypes for recombinant HIV-1 genomes. We identify the range of parameter values that maximize the accuracy of our predictions on the test set, where we measure accuracy based on concurrence with hand-curated annotations. We demonstrate that parental subtype inference over all parameter values is feasible for HIV full-length genomes, and that it is much more informative than restricting to a particular choice of parameters.

1.1 Related work

There are many methods for the identification and characterization of recombinant sequences (for a current list, see `http://bioinf.man.ac.uk/robertson/~recombination/`). Most of these methods take as input a multiple alignment of a putative recombinant with a collection of parental sequences. These methods either output a list of parental sequences giving rise to each query position or simply state whether or not the query is a recombinant.

One method that attempts to identify the parental subtype for each position within a putative recombinant sequence was introduced in [9]. This method models each column of the input alignment by a hidden state that represents the tree topology of the column. Given fixed costs for substitutions and recombination events (changes in tree topology), the algorithm gives a most parsimonious explanation for the query sequence. This model was later formulated probabilistically as an HMM [22]. Parameter estimation and extensions for this model

have been studied in depth [11, 12]. These HMMs are similar to multiple alignment methods that have incorporated the concept of recombination [14, 18]. Indeed, our model can be regarded as a specialization of "jumping alignments" that were introduced for remote homology detection [33].

Several methods that have been applied to HIV recombinant forms use a sliding window technique [5, 20, 26, 32, 31]. In this approach, subsequences of a certain fixed length are considered along the genome and a test statistic is calculated from each segment. These methods are local in the sense that they do not solve a global optimization problem, but detect changes of locally computed characteristics along the genome.

Methods like those just mentioned and others require many parameters, which must be estimated. Although many of these methods can be effective, their performance is highly dependent on their parameter values and the rates of evolutionary events that formed the input sequences [27]. In addition, the nucleotide substitution models for these methods are usually fixed, even though different models may be better for different data sets [28]. The framework of Bayesian statistics offers a way to deal with uncertainty in model parameters, albeit at considerable computational cost due to Markov Chain Monte Carlo (MCMC) methods that are needed for estimating the posterior probabilities [34].

We remedy parameter uncertainty by determining all solutions for all parameter choices. For the presented model, a specific parametric analysis was first conducted in [19] for the detection of recombinant sequences. Employing cost minimization, the parametric solutions were analyzed in order to identify features that indicate recombination. Here, we take a probabilistic point of view and present parametric inference methods that generalize readily to more complex probabilistic graphical models. Even though the algorithms we present are quite general and return complete information about dependence on parameter choices, they can master full-length HIV genomes of about 10,000 base pairs.

2 Methods

2.1 Hidden Markov model and dynamic programming

We consider sequences over the 5-letter alphabet $\Sigma' = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{-}\}$ of nucleotides supplemented by the gap character. For a given multiple alignment $\mathcal{A}$ of n columns and $N+1$ DNA sequences $y, s^{(1)}, s^{(2)}, \ldots, s^{(N)} \in \Sigma'^n$, the task is to explain the distinguished sequence y as a recombinant of the remaining sequences $s^{(i)}$, $i \in \{1, \ldots, N\}$. An *annotation* of $y = (y_1, \ldots, y_n)$ is a sequence $x = (x_1, \ldots, x_n)$ over the alphabet $\Sigma = \{1, 2, \ldots, N\}$, where $x_j = i$ signifies that the jth character of y originates from the jth character of $s^{(i)}$. We introduce an HMM for inferring unobserved annotations from observed sequences. An *explanation* is an annotation that maximizes the *a posteriori* probability of the data, given fixed model parameters. Since the "true" values for the parameters are not known with any certainty, we present a parametric analysis that yields all maximum *a posteriori* (MAP) annotations for all parameter values.

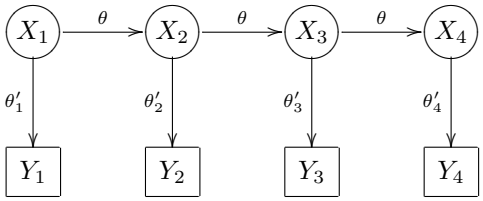

Figure 1: Graph of the recombination hidden Markov model for $n = 4$ alignment columns. For column j of a multiple alignment, the observed random variable, Y_j, is the character of the recombinant sequence for that column, and the hidden random variable, X_j, represents the parental sequence from which that character was derived.

The HMM has hidden random variables $X = (X_1, \ldots, X_n)$, encoding annotations in Σ^n, and observed random variables $Y = (Y_1, \ldots, Y_n)$, encoding sequences in Σ'^n. The underlying graph of the model is depicted in Figure 1.

The dynamics of the model are given by the $N \times N$ transition matrix θ, which is the same for all transitions, and the $N \times 5$ emission matrices θ'_j, for $1 \leq j \leq n$. The joint probability of an observed sequence $Y = y$ and an annotation $X = x$ is

$$f_{y,x} = \text{Prob}(Y = y, X = x) = \pi_{x_1} \theta'_{1,x_1,y_1} \prod_{j=2}^{n} \theta_{x_{j-1},x_j} \theta'_{j,x_j,y_j},$$

where $\pi_i = \text{Prob}(X_1 = i)$ is the initial distribution of X_1.

The likelihood of an observed sequence $Y = y$ is obtained by marginalization,

$$f_y = \text{Prob}(Y = y) = \sum_{x \in \Sigma^n} f_{y,x}.$$

The sum can be evaluated efficiently due to the factorization

$$f_y = \sum_{x_1 \in \Sigma} \pi_{x_1} \, \theta'_{1,x_1,y_1} \left(\sum_{x_2 \in \Sigma} \theta_{x_1,x_2} \, \theta'_{2,x_2,y_2} \left(\sum_{x_3 \in \Sigma} \cdots \right. \right.$$
$$\left. \left. \cdots \left(\sum_{x_n \in \Sigma} \theta_{x_{n-1},x_n} \, \theta'_{n,x_n,y_n} \right) \cdots \right) \right),$$

which yields the forward algorithm. The factorization is valid exactly because multiplication distributes over addition, a property characteristic of *semirings*. We have the following generalization.

Dynamic Programming Principle. *Let $\mathcal{S}$ be any set equipped with two binary operations $\oplus$ and $\odot$ such that $(\mathcal{S}, \oplus, \odot)$ forms a semiring. Let π, θ, and θ'_j be matrices over $\mathcal{S}$. Then the element*

$$\bigoplus_{x \in \Sigma^n} \pi_{x_1} \odot \theta'_{1,x_1,y_1} \odot \bigodot_{j=2}^{n} \left(\theta_{x_{j-1},x_j} \odot \theta'_{j,x_j,y_j} \right) \tag{1}$$

can be computed by the factorization

$$\bigoplus_{x_1 \in \Sigma} \pi_{x_1} \odot \theta'_{1,x_1,y_1} \odot \left(\bigoplus_{x_2 \in \Sigma} \theta_{x_1,x_2} \odot \theta'_{2,x_2,y_2} \odot \left(\bigoplus_{x_3 \in \Sigma} \cdots \right. \right.$$
$$\left. \left. \cdots \left(\bigoplus_{x_n \in \Sigma} \theta_{x_{n-1},x_n} \odot \theta'_{n,x_n,y_n} \right) \cdots \right) \right). \tag{2}$$

For variations and generalizations of this principle, see e.g. [1, 2, 4, 6, 7, 17, 25]. We will revisit the Dynamic Programming Principle several times, with different semirings.

The explanations of y are exactly the annotations x satisfying

$$f_{y,x} = \max_{x' \in \Sigma^n} f_{y,x'}. \tag{3}$$

This condition on x remains unchanged when passing to the logarithms of the parameters π_i, $\theta_{i,i'}$, $\theta'_{i,j,k}$. Then the maximum (3) can be computed efficiently using the factorization (2) in the $(\max, +)$ algebra. This procedure is exactly the classical Viterbi algorithm. The $(\max, +)$ algebra is also known as the tropical semiring, denoted $(\mathbb{R} \cup \{-\infty\}, \max, +)$, and solving (3) is equivalent to evaluating the polynomial (1) in tropical arithmetic [24].

2.2 Model specification

Motivated by the copy-choice model of recombination, we customize the general HMM by specializing the matrices θ and θ'_j. We assume that sequences change over time by two mechanisms: mutation and recombination. Furthermore, these evolutionary events occur uniformly over the sequence with unknown probabilities μ and ρ for mutation and recombination, respectively. For simplicity, we neglect insertions and deletions in our model and encode alignment gaps with the additional character "-". Given the multiple alignment $\mathcal{A}$ and parameters μ and ρ, we define

$$\begin{aligned}
\pi_i &= \frac{1}{N}, \\
\theta_{i,i'} &= \begin{cases} 1 - (N-1)\rho & \text{if } i = i', \\ \rho & \text{else,} \end{cases} \\
\theta'_{j,i,k} &= \begin{cases} 1 - 4\mu & \text{if } S_j^{(i)} = k, \\ \mu & \text{else,} \end{cases}
\end{aligned}$$

for $i, i' \in \Sigma$, $k \in \Sigma'$, and $j = 1, \ldots, n$.

MAP estimation in this HMM is equivalent to optimizing the following intuitive scoring scheme. Given an annotation x, we identify two salient features. The first of these is r, the number of indices j ($1 \leq j \leq n-1$) such that $x_j \neq x_{j+1}$, that is, the number of recombination events. The second is m, the number of j such that $y_j \neq s_j^{(x_j)}$, that is, the number of mutations that must have occurred given that y_j originated from the sequence $s^{(x_j)}$. We include in m instances where one of y_j or $s_i^{(x_j)}$ is a gap character and the other is a nucleotide. Given choices for two real-valued parameters R and M, corresponding to recombination and mutation, respectively, the *annotation score* of x is $R \cdot r + M \cdot m$. We seek to find the annotation x that maximizes this score. The biological meaningful solutions correspond to negative parameters R and M.

We now show that this scoring scheme is equivalent to our hidden Markov model. Given values for R and M, let

$$\rho = \frac{e^R}{1 + (N-1)e^R} \quad \text{and} \quad \mu = \frac{e^M}{1 + 4e^M}. \tag{4}$$

Note that $1-(N-1)\rho = \frac{1}{1+(N-1)e^R}$ and $1-4\mu = \frac{1}{1+4e^M}$ are the other transition and emission probabilities appearing in θ and θ'_j. Given an annotation x with r recombination events and m mutation events, one can check that

$$\begin{aligned}
&\text{Prob}(X = x, Y = y) \\
&\quad = N^{-1} \rho^r \left(1 - (N-1)\rho\right)^{n-1-r} \mu^m (1-4\mu)^{n-m} \\
&\quad = \frac{e^{R \cdot r + M \cdot m}}{N \left(1 + (N-1)e^R\right)^{n-1} \left(1 + 4e^M\right)^n}.
\end{aligned}$$

Since the denominator is constant over all annotations x, and since exponentiation is an increasing function, $\text{Prob}(X = x, Y = y)$ is maximized exactly when the score $R \cdot r + M \cdot m$ is maximized.

Conversely, given ρ and μ, let $R = \log(\rho) - \log\left(1 - (N-1)\rho\right)$ and $M = \log(\mu) - \log(1 - 4\mu)$. Since these are simply equations (4) solved for R and M, the most probable annotation given by the HMM is exactly the annotation maximizing the score $R \cdot r + M \cdot m$. Therefore our probabilistic HMM is equivalent to the scoring scheme described above, and results can easily be translated back and forth. For the following parametric analysis, the scoring scheme formulation is more natural.

2.3 Parametric inference

Every annotation can be summarized by a pair (r, m), where r and m are the number of implied recombination and mutation events, respectively. Thus, all possible *annotation summaries* can be represented as a collection of points in the two-dimensional summary space, with coordinates r and m. The *annotation polygon* is defined as the convex hull of this set of points [25, Sec. 3.2]. Each point

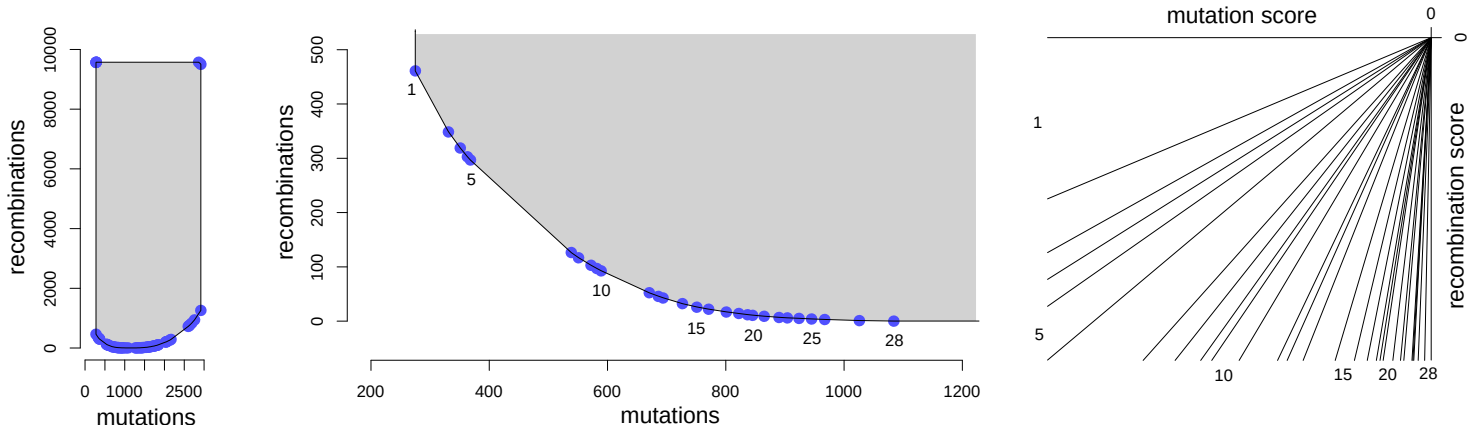

Figure 2: Annotation polygon for the sequence CY.94.CY032_AF049337. Each point of the polygon represents all annotations with the same number of mutation and recombination events. The 72 vertices of the polygon correspond to explanations for different scoring schemes. Those 28 vertices corresponding to meaningful explanations are in the lower left corner of the polygon (on the left), which is shown in detail in the middle. The right panel displays the corresponding parameter regions for which each vertex is optimal (i.e., is an explanation).

of the annotation polygon represents all annotations with the same number of recombinations and mutations, and hence with the same likelihood or annotation score. Each vertex is the summary of an explanation (optimal annotation), for some choice of parameter values. The *normal fan* of the annotation polygon is a subdivision of the parameter space (R, M) into a finite number of regions (Figure 2). In this subdivision, the same annotations are optimal for all choices of parameters that lie in the same region.

The annotation polygon can be computed by the Dynamic Programming Principle running in the *polytope algebra*. In our case, the objects of this semiring are polygons, i.e., 2-dimensional polytopes. Multiplication is defined as the Minkowski sum, and addition as the convex hull of the set union. This algorithm is known as *polytope propagation* [25, Sec. 2.3].

2.4 Concurrence of explanations

We want to compare inferred explanations x with the "true" biological annotation a, which we take to be the hand-curated Los Alamos annotation that is based on phylogenetic analyses [30]. This comparison can be done for all parameter values of the model and hence allows for determining optimal parameters.

Let y be a sequence with true annotation $a \in \Sigma^n$. When comparing inferred annotations x to the true annotation a, we consider a rating scheme that encapsulates the most important aspect of the annotation x, namely, whether it correctly states the set of sequences that have recombined to form y. We define the *parental set* of x as $I_x = \{i \mid x_j = i \text{ for some } j\}$, i.e., the set of recombining sequences that x indicates. The *concurrence* of a parental set I to the true annotation a is defined as

$$c_a(I) = \frac{|I_a \cap I|}{|I_a \cup I|},$$

and the concurrence of an annotation x to the true annotation is $c_a(x) = c_a(I_x)$. Thus, if x correctly names the parental sequences, then $I_x = I_a$ and $c_a(x) = 1$. If x and a have no recombining sequence in common, then $c_a(x) = 0$. In general, a fixed choice of parameters will yield several optimal parental sets. For a collection $\mathcal{X}$ of annotations, we average the concurrence c_a over all sets I that appear as a parental set I_x, for $x \in \mathcal{X}$. To be precise, define $\mathcal{I}_\mathcal{X} = \{I_x \mid x \in \mathcal{X}\}$, and let $c_a(\mathcal{X})$ be the average of $c_a(I)$ over all $I \in \mathcal{I}_\mathcal{X}$. Note that we use an unweighted average: the number of occurrences of a set I as a parental set I_x does not affect it.

In order to rate all optimal annotations, we need to compute their parental sets. This can be achieved by another instance of the Dynamic Programming Principle. Let $\mathcal{I}_\mathcal{X}$ denote the collection of parental sets that corresponds to a collection $\mathcal{X}$ of explanations. We consider ordered pairs $(\phi, \mathcal{I}) \in \mathcal{S} = \big(\mathbb{R} \cup \{-\infty\}\big) \times 2^{2^\Sigma}$, consisting of a number ϕ and a collection $\mathcal{I}$ of subsets of Σ. For $\phi_1, \phi_2 \in \mathbb{R} \cup \{-\infty\}$ and $\mathcal{I}_1, \mathcal{I}_2 \subset 2^{2^\Sigma}$, we define the operations

$$(\phi_1, \mathcal{I}_1) \oplus (\phi_2, \mathcal{I}_2) = \left(\max_{j=1,2} \phi_j, \bigcup \{\mathcal{I}_j \mid \phi_j = \max(\phi_1, \phi_2)\} \right)$$

and

$$(\phi_1, \mathcal{I}_1) \odot (\phi_2, \mathcal{I}_2) = (\phi_1 + \phi_2,\ \{I_1 \cup I_2 \mid I_j \in \mathcal{I}_j\}),$$

which make $(\mathcal{S}, \oplus, \odot)$ a semiring.

If we define the matrices π, θ, and θ' over this semiring by setting

$$\begin{aligned}
\pi_i &= \Big(0, \big\{\{i\}\big\}\Big) \\
\theta_{i,i'} &= \begin{cases} \Big(0, \big\{\{i'\}\big\}\Big) & \text{if } i = i', \\ \Big(R, \big\{\{i'\}\big\}\Big) & \text{else,} \end{cases} \\
\theta'_{j,i,k} &= \begin{cases} \big(0, \{\emptyset\}\big) & \text{if } s_j^{(i)} = k, \\ \big(M, \{\emptyset\}\big) & \text{else,} \end{cases}
\end{aligned}$$

then the object defined by expression (1) is exactly the pair $(\phi, \mathcal{I}_\mathcal{X})$ with ϕ the optimal score and $\mathcal{I}_\mathcal{X}$ the corresponding collection of optimal parental sets. Thus, by virtue of the factorization (2), these collections can be computed efficiently.

2.5 Data set

We considered all of the HIV-1 full-length genomes in the Los Alamos Sequence Database. A multiple DNA sequence alignment of length 12,635 was obtained from the 2003 HIV and SIV Alignments web site (`http://hiv-web.lanl.gov/`). We have omitted the CRF 01_AE, because of the lacking putative parental E strain. We further excluded all recombinants that involve a CRF as one of the recombining sequences. The resulting set of 341 genomes comprises 11 different subtypes and 11 different CRFs (Table 1).

Pure subtype sequences were used to build subtype consensus sequences. The resulting consensus alignment was trimmed such that all initial and terminal gaps were removed. The subalignment covers positions 1,626 to 11,199 relative to the original alignment. These positions correspond to 1,174 and 9,144, respectively, in the HXB2 numbering scheme [16]. By trimming, we effectively excluded the flanking LTR regions, and small portions of the *gag* (at the 3' end) and *nef* (at the 5' end) genes.

Thus, the data for each of the 341 recombination inference problems consists of a multiple alignment of length 9,574 of one of the sequences and the eleven consensus sequences.

3 Results

We have computed, for all 341 HIV-1 genomes, the annotation polygons and all concurrence ratings for all parameter values of the HMM. The Dynamic Programming Principle has been implemented in C++, and its various occurrences have been realized using templates and operator overloading, i.e., implementation of the respective semirings. The annotation polygons can be computed in

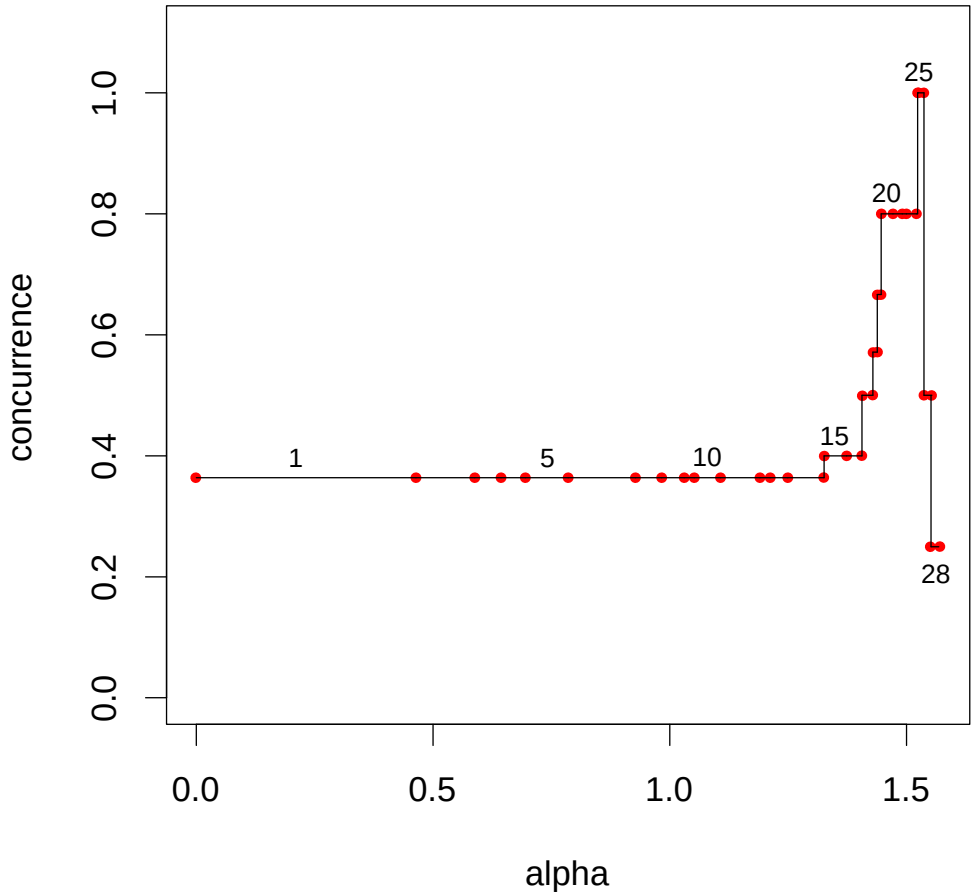

Figure 3: Concurrence of model predictions with the true annotation for sequence CY.94.CY032_AF049337, over all parameter values. For each value of the angle $\alpha = \tan^{-1} R/M$ (x-axis), ranges of which uniquely identify the biologically meaningful parameter regions in Figure 2, the concurrence (y-axis) of the optimal parental sets with the true parental set is plotted. MAP estimates remain constant over each parameter region (numbered according to Figure 2), which gives rise to a step function.

Subtype	Count	CRF	Count
A1	38	02 (A,G)	34
A2	4	03 (A,B)	3
B	57	04 (A1,G,H,K)	3
C	109	05 (D,F1)	3
D	39	06 (A1,G,J,K)	4
F1	6	07 (B,C)	4
F2	4	08 (B,C)	4
G	7	10 (C,D)	3
H	3	12 (B,F1)	5
J	2	14 (B,G)	6
K	2	16 (A2,D)	1

Table 1: Analyzed HIV subtypes and circulating recombinant forms (CRFs).

$O(N^2 n^{5/3})$ time [23], and computing a typical HIV annotation polygon takes less than one minute on a 2.6 Ghz Linux workstation. The complete computational results and the source code are available at `http://bio.math.berkeley.edu/recombination/`.

3.1 Explanations

An annotation polygon represents the set of all possible annotations of a sequence as a recombinant of the 11 consensus sequences. For example, Figure 2 shows the annotation polygon for sequence CY.94.CY032_AF049337, which is believed to be a CRF 04 comprising subtypes A1, G, H, and K in the Los Alamos database.

The annotation polygon induces a subdivision of the parameter space into regions such that the same annotations are optimal for all parameter values in a given region. Figure 2 shows the biologically meaningful part of the polygon (middle) and the corresponding subdivision of the parameter space (right) for all non-positive M and R (i.e., where mutation and recombination is penalized). Which annotations are optimal is uniquely determined by the angle $\alpha = \tan^{-1} R/M$, and each parameter region is a cone enclosed by a pair of rays. We refer to these parameter regions by (intervals of) the angle $\alpha \in [0, \pi/2]$.

For our analysis of the HIV-1 genomes, an annotation is a string over $\Sigma = \{1, \ldots, 11\}$ of length $n = 9{,}574$. In Figure 4 we have picked one explanation for each vertex in Figure 2. From bottom to top, the explanations correspond to parameter choices of increasing α, that is, of increasing ratio of the probability of a mutation to the probability of a recombination. Explanations 25 and 26 use exactly the set of recombining sequences A1, G, H, and K, although some small sequence segments of the Los Alamos hand-curated annotation are missing. However, most of these small segments appear in explanations that correspond to smaller angles α. For example, a subtype G fragment at the 3' end of the

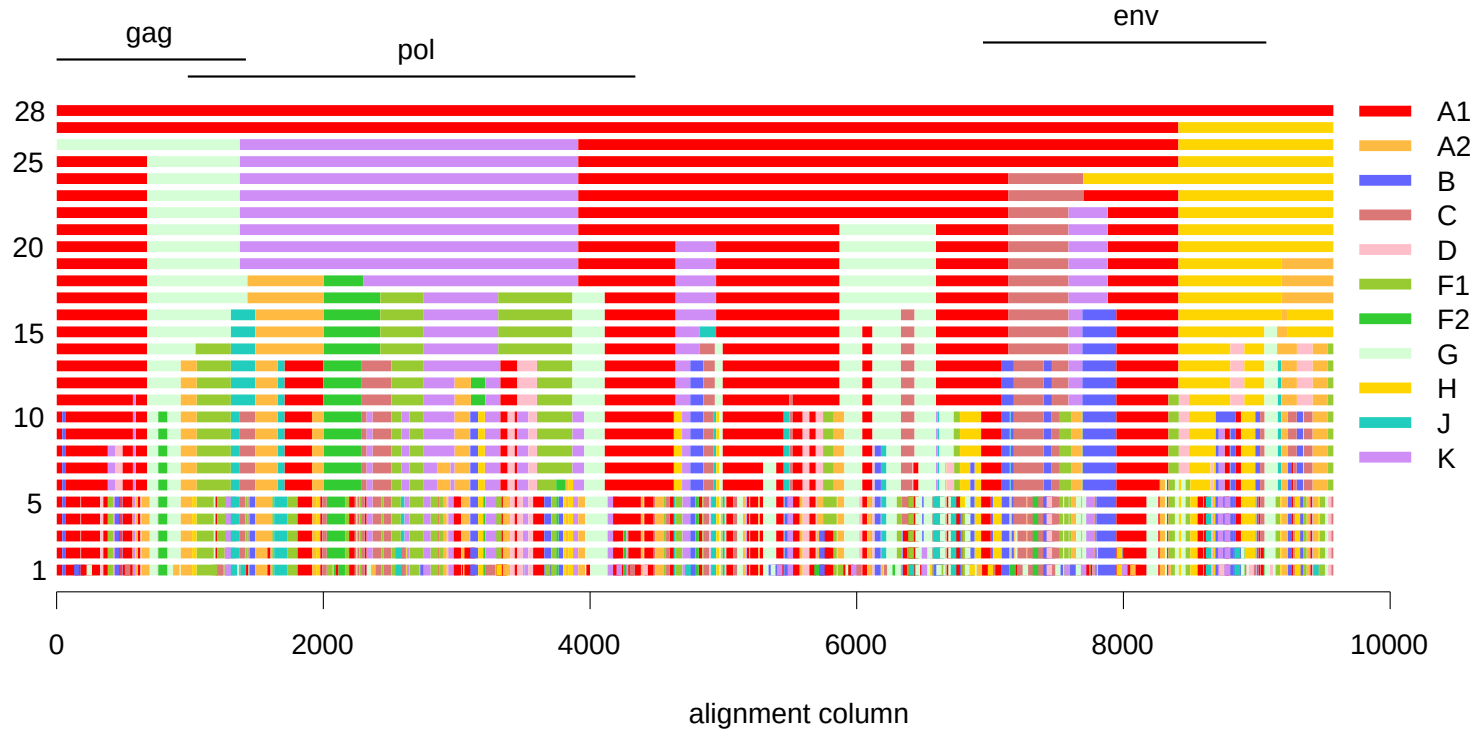

Figure 4: Explanations (optimal annotations) for sequence CY.94.CY032_AF049337 (CRF 04, subtypes A, G, H, K), over all meaningful parameter values. The numbers on the left correspond to the labeled parameter regions in Figures 2 and 3. On top the locations of the three major HIV genes are indicated.

pol gene appears in explanation 17 and below, and a subtype K segment in the *env* gene occurs in explanations 22 to 17. For explanations 24 and below, we estimate a segment in the 5' region of the *env* gene to originate from subtype C, whereas in the Los Alamos explanation this region is of unknown origin. Thus, the parametric solution of the inference problem recovers most features of the established explanation and may serve as a starting point for new sequences and for unexplained regions.

3.2 Performance

Computing the sets of recombining sequences for each vertex of the polygon yields the concurrences for all parameter values in the corresponding normal cones (Figure 3). Concurrences are plotted against the angle α. The resulting function takes values in $[0, 1]$, where 0 means that the parental sets of the predicted annotations and the true annotation have no sequence in common, and 1 means that both sets coincide. Since every parameter value in a region yields the same set of explanations, the function is a step function, constant within each region. For example, for sequence CY.94.CY032_AF049337, the maximum score of 1 is attained for the two subsequent parameter regions denoted 25 and 26, which correspond to the two explanations labeled as such in Figure 4.

We have computed the concurrences for all 341 genomes in order to analyze the performance of the HMM in recovering the established annotation from inferred explanations. In Figure 5 we first analyze the 271 pure subtypes.

For $\alpha = \pi/2 \approx 1.57$ all scores are 1, since this angle corresponds to $\rho = 0$, i.e., recombination is impossible. In fact, all $\alpha \geq 1.556$ yield the correct annotations for all subtypes and so have a score of 1. The average scores per subtype are displayed in the lower rightmost plot of Figure 5. For example, for $\alpha \geq 1.472$, the average of these averages is still ≥ 0.95. This average can be regarded as the expected score when assuming uniform distribution of the subtypes.

For the CRFs, concurrence is no longer an increasing function, because $\rho = 0$ cannot yield the true annotation (Figure 6). For many but not all CRFs, a score of 1, i.e., perfect annotation, is reached. The maximum expected score under the assumption of uniform CRF distribution is 0.874 and it is attained for $\alpha \in [1.481, 1.485]$. This unique local and global maximum is rather sharp indicating that the optimal value does not vary a lot between different CRFs. Scores ≥ 0.85 are expected for $\alpha \in [1.472, 1.494]$, a parameter region in which the expected score for the pure subtypes is ≥ 0.95.

4 Discussion

We have presented an HMM for the detection and annotation of recombinant sequences. The unifying Dynamic Programming Principle was applied for MAP estimation, parametric inference, and validation via concurrence rating. As a probabilistic model, the parameters ρ and μ of the HMM can be interpreted as the probability of observing a recombination event and a mutation event,

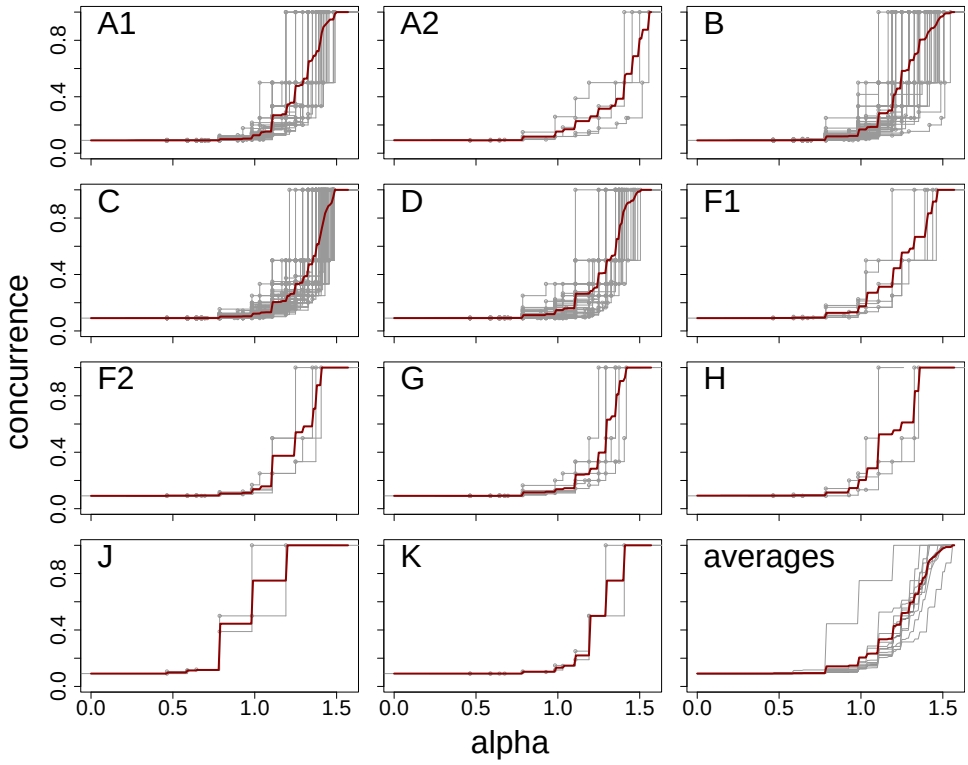

Figure 5: Concurrences of model predictions with the true annotations for established pure subtype sequences, over all values of the model parameters. Scores for individual sequences are plotted in light grey and average scores for all sequences of the same pure subtype is the thick, dark line. The average curves for all of the subtypes are plotted together (grey) in the lower right plot, along with their average (dark).

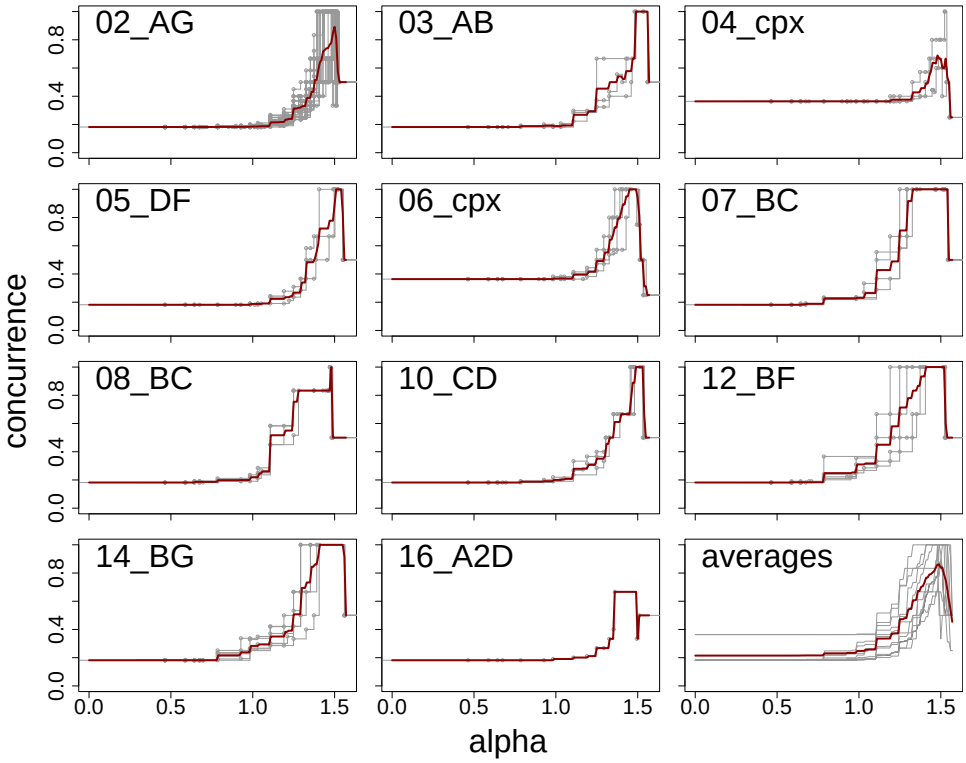

Figure 6: Concurrences of model predictions with the true annotations for established CRF sequences. Plots are analogous to those in Figure 5.

respectively, in the data set. The HMM also allows one to compute the posterior probability of each putative parent at each site of the genome.

The model can be extended in many directions, while maintaining the same conceptual and computational framework. The emission matrices θ'_j may be parametrized to implement more flexible substitution models than the Jukes-Cantor model (which we implicitly assume), e.g., by accounting for transition/transversion bias. The model might be expanded by explicit modeling of insertions and deletions, thereby simultaneously searching for recombination and aligning the query sequence to the reference alignment. This problem can also be cast in an HMM. The corresponding higher dimensional polytopes and concurrence ratings for any of these models can be computed efficiently using the Dynamic Programming Principle. In addition, other more refined rating systems could be used to determine the accuracy of the explanations returned by the model. On the other hand, the power of our model to predict recombination, as measured by concurrence ratings, is already quite high, and we need only two parameters.

Parametric inference is very efficient for the presented model. We have used it to analyze 341 HIV genomes of size approximately 10,000 characters. Furthermore, the 2-parameter model has a concise representation of its set of explanations (Figure 4), and hence the parametric inference solution can easily be inspected. Indeed, the case of HIV-1 genomes has shown that the set of all solutions for all parameters can be much more informative than a single estimate. For example, as discussed in Section 3.1, certain parameter values may be appropriate for determining the parental set of a sequence, whereas other values may pick up shorter parental subsequences. All of this information is inherent in the annotation polygon.

Solving the parametric inference problem should be regarded as an offline precomputing step. Once its solution is available, statistical inference for fixed parameters becomes very cheap. In fact, MAP estimation in the HMM corresponds to solving a linear program on the annotation polygon. In the absence of labeled training data, i.e., annotated recombinants, the model parameters may be estimated by maximum likelihood using the Expectation Maximization algorithm. In those situations the annotation polygon can be used to assess the sensitivity of MAP estimation with respect to the uncertainty inherent in parameter estimation. The relative position of the parameter estimate to the boundaries of the discrete parameter regions determines the robustness of explanations.

Parametric analysis also provided the basis for evaluating the HMM. We used the parental sets as the feature of an explanation to be compared to the true annotation. Unlike the collection of explanations for a given choice of parameters, the collection of parental sets can be efficiently computed. For example, there are 28,704 explanations for sequence CY.94.CY032_AF049337 in parameter region 25 (Figures 2,3,4), but all share the same single parental set {A1, G, H, K}. Thus, restricting to parental sets and ignoring recombination breakpoints is a compromise between computational feasibility and accuracy. The analysis of these concurrence ratings has identified optimal parameter regions. In practice,

prior knowledge about the mosaic structure of a sequence will affect the choice of the parameter region to investigate.

In summary, annotation polygons provide a powerful tool for inferring the recombinant structure of HIV genomes.

Acknowledgments

Niko Beerenwinkel was supported by the Deutsche Forschungsgemeinschaft under grant BE 3217/1-1, Colin Dewey by the NIH (HG003150), and Kevin Woods by the NSF (DMS-040214).

References

[1] SM Aji and RJ McEliece. The generalized distributive law. *IEEE Transactions on Information Theory*, 46(2):325–343, 2000.

[2] R Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NY, 1957.

[3] JM Coffin. Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. *J Gen Virol*, 42(1):1–26, Jan 1979.

[4] RG Cowell, AP Dawid, SL Lauritzen, and DJ Spiegelhalter. *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Sciences. Springer, New York, 1999.

[5] T de Oliveira, K Deforche, S Cassol, M Salminen, D Paraskevis, C Seebregts, J Snoeck, EJ van Rensburg, AMJ Wensing, DA van de Vijver, CA Boucher, R Camacho, and AM Vandamme. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, 21(19):3797–3800, Oct 2005.

[6] R Durbin, S Eddy, A Krogh, and G Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK, 1998.

[7] R Giegerich, C Meyer, and P Steffen. A discipline of dynamic programming over sequence data. *Science of Computer Programming*, 51(3):215–263, 2004.

[8] D Gusfield, K Balasubramanian, and D Naor. Parametric optimization of sequence alignment. *Algorithmica*, 12:312–326, 1994.

[9] JJ Hein. Method to reconstruct the history of sequences subject to recombination. *J Mol Evol*, 20:402–411, 1993.

[10] WS Hu and HM Temin. Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. *Proc Natl Acad Sci U S A*, 87(4):1556–1560, Feb 1990.

[11] D Husmeier and F Wright. Detection of recombination in DNA multiple alignments with hidden Markov models. *J Comput Biol*, 8(4):401–427, 2001.

[12] D Husmeier. Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics*, 21 Suppl 2:ii166–ii172, Sep 2005.

[13] A Jung, R Maier, JP Vartanian, G Bocharov, V Jung, U Fischer, E Meese, S Wain-Hobson, and A Meyerhans. Multiply infected spleen cells in HIV patients. *Nature*, 418:144, 2002.

[14] J Kececioglu and D Gusfield. Reconstructing a history of recombinations from a set of sequences. *Discrete Applied Mathematics*, 88:239–260, 1998.

[15] P Kellam and BA Larder. Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *J Virol*, 69(2):669–674, Feb 1995.

[16] BT Korber, BT Foley, CL Kuiken, SK Pillai, and JG Sodroski. Numbering positions in HIV relative to HXB2CG. In *HIV Sequence Compendium*. Theoretical Biology and Biophysics, Group T-10, Los Alamos National Laboratory, Los Alamos, NM, 1998.

[17] F Kschischang, B Frey, and H Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information*, 47(2):498–519, Feb 2001.

[18] C Lee, C Grasso, and MF Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, Mar 2002.

[19] J Maydt and T Lengauer. Recco: Recombination analysis using cost optimization. *in preparation*, 2005.

[20] JM Smith. Analyzing the mosaic structure of genes. *J Mol Evol*, 34:126–129, 1992.

[21] FE McCutchan. Understanding the genetic diversity of HIV-1. *AIDS*, 14 Suppl 3:S31–S44, 2000.

[22] G McGuire, F Wright, and MJ Prentice. A Bayesian model for detecting past recombination events in DNA multiple alignments. *J Comput Biol*, 7(1-2):159–170, 2000.

[23] L Pachter and B Sturmfels. Parametric inference for biological sequence analysis. *Proc Natl Acad Sci U S A*, 101(46):16138–16143, Nov 2004.

[24] L Pachter and B Sturmfels. Tropical geometry of statistical models. *Proc Natl Acad Sci U S A*, 101(46):16132–16137, Nov 2004.

[25] L Pachter and B Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Oxford University Press, 2005.

[26] D Paraskevis, K Deforche, P Lemey, G Magiorkinis, A Hatzakis, and AM Vandamme. SlidingBayes: exploring recombination using a sliding window approach based on Bayesian phylogenetic inference. *Bioinformatics*, 21(7):1274–1275, Apr 2005.

[27] D Posada and KA Crandall. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A*, 98(24):13757–13762, Nov 2001.

[28] D Posada and KA Crandall. Selecting models of nucleotide substitution: An application to human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.*, 18(6):897–906, 2001.

[29] DL Robertson, PM Sharp, FE McCutchan, and BH Hahn. Recombination in HIV-1. *Nature*, 374(6518):124–126, Mar 1995.

[30] DL Robertson, F Gao, BH Hahn, and PM Sharp. Intersubtype recombinant HIV-1 sequences. In *HIV Sequence Compendium*. Theoretical Biology and Biophysics, Group T-10, Los Alamos National Laboratory, Los Alamos, NM, 1997.

[31] MO Salminen, JK Carr, DS Burke, and FE McCutchan. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum Retroviruses*, 11(11):1423–1425, Nov 1995.

[32] AC Siepel, AL Halpern, C Macken, and BT Korber. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res Hum Retroviruses*, 11(11):1413–1416, Nov 1995.

[33] R Spang, M Rehmsmeier, and J Stoye. A novel approach to remote homology detection: jumping alignments. *J Comput Biol*, 9(5):747–760, 2002.

[34] MA Suchard, RE Weiss, KS Dorman, and JS Sinsheimer. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. *Syst Biol*, 51(5):715–728, Oct 2002.

[35] M Waterman, M Eggert, and E Lander. Parametric sequence comparisons. *Proc. Natl. Acad. Sci. U. S. A.*, 89:6090–6093, 1992.

[36] K Yusa, MF Kavlick, P Kosalaraksa, and H Mitsuya. HIV-1 acquires resistance to two classes of antiviral drugs through homologous recombination. *Antiviral Res*, 36(3):179–189, Dec 1997.

[37] R Zimmer and T Lengauer. Fast and numerically stable parametric alignment of biosequences. In *Proc. 1st Ann. Int. Conf. on Res. in Comput. Biol. (RECOMB 1997) , January 20–23, 1997, Santa Fe, NM, USA*, pages 344–353, 1997.

Niko Beerenwinkel, Department of Mathematics, University of California, Berkeley 94720, USA, niko@math.berkeley.edu.

Colin Dewey, Department of Electrical Engineering, University of California, Berkeley 94720, USA, cdewey@eecs.berkeley.edu.

Kevin Woods, Department of Mathematics, University of California, Berkeley 94720, USA, kwoods@math.berkeley.edu.