

Detecting Sporadic Recombination in DNA Alignments with Hidden Markov Models

Dirk Husmeier and Frank Wright

Biomathematics and Statistics Scotland, SCRI, Invergowrie, Dundee DD2 5DA, UK

Abstract. Conventional phylogenetic tree estimation methods assume that all sites in a DNA multiple alignment have the same evolutionary history. This assumption is violated in data sets from certain bacteria and viruses due to recombination, a process that leads to the creation of mosaic sequences from different strains and, if undetected, causes systematic errors in phylogenetic tree estimation. In the current work, a hidden Markov model (HMM) is employed to detect recombination events in multiple DNA sequence alignments. The emission probabilities in a given state are determined by the branching order (topology) and the branch lengths of the respective phylogenetic tree, while the transition probabilities depend on the global frequency of recombination. All model parameters are optimized in a maximum likelihood sense with the expectation maximization (EM) algorithm. The resulting parameter optimization scheme is applied to a synthetic benchmark problem and to real DNA sequences from the *argF* gene of four strains of the bacterium *Neisseria*. In both cases we find a significant improvement over an earlier heuristic parameter estimation approach.

1 Introduction

Conventional phylogenetic tree estimation methods assume that all sites in a DNA multiple alignment have the same evolutionary history. This is a reasonable approach when applied to DNA sequences obtained from most species. However, this assumption is violated in certain bacteria and viruses due to sporadic *recombination*, which is a process that leads to the transfer of DNA subsequences between different strains (see Figure 1). If undetected, this can lead to errors in phylogenetic tree estimation. The finding of evidence for sporadic recombination could have implications for vaccine development (e.g. in the case of HIV; see (Robertson *et al.*, 1995)).

2 A Naive Bayesian Approach

Let $\mathbf{y}_t \in \{A, G, C, T\}^m$ denote the t th column in a multiple alignment of m DNA sequences of length N , $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$, and introduce the multinomial random variable $s_t \in \{1, \dots, K\}$ to indicate the tree topology that generated the nucleotide configuration at site t . Let \mathbf{w}_{s_t} denote the vector of all branch lengths in the tree corresponding to s_t . A phylogenetic tree is a generative probabilistic model, that is, given the topology s_t and the

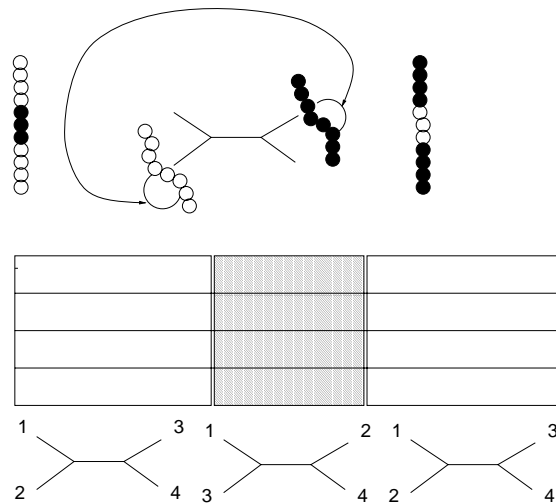


Figure 1: Schematic illustration of recombination. The exchange of DNA subsequences between different strains (top diagram, middle) results in two so-called mosaic sequences (top diagram, margins). The affected region in the multiple DNA sequence alignment (shaded area in the middle diagram) seems to originate from a different phylogenetic topology (bottom diagram).

parameters \mathbf{w}_{s_t} , we can compute the probability for an observed column vector \mathbf{y}_t in the alignment¹: $P(\mathbf{y}_t | s_t, \mathbf{w}_{s_t})$.

In the presence of recombination, the tree topology s_t becomes a site-dependent random variable, and our objective is to find the mode of

$$P(\mathbf{s} | \mathbf{Y}) = P(s_1, \dots, s_N | \mathbf{y}_1, \dots, \mathbf{y}_N) \quad (1)$$

that is, the most likely sequence of topologies, $\mathbf{s} = (s_1, \dots, s_N)$, given the data $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$. From Bayes rule we have $P(\mathbf{s} | \mathbf{Y}) \propto P(\mathbf{Y} | \mathbf{s})P(\mathbf{s})$, where the expression on the right needs to be normalized to turn the proportionality into an equality. Under the common assumption that mutations at different sites on the DNA strand are independent of each other (see, for instance, (Baldi & Brunak, 1998)), the first term in the expression on the right factorizes: $P(\mathbf{Y} | \mathbf{s}) = \prod_{t=1}^N P(\mathbf{y}_t | s_t)$. If we assume a uniform prior on the state sequences, $P(\mathbf{s}) = C$, where C denotes a constant, then (1) gives:

$$P(\mathbf{s} | \mathbf{Y}) = \prod_{t=1}^N P(s_t | \mathbf{y}_t), \quad P(s_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | s_t)}{\sum_{s_t=1}^K P(\mathbf{y}_t | s_t)} \quad (2)$$

We can then assign the t th site in the alignment to the topology s_t that maximizes (2).

3 Detecting Recombination with Hidden Markov Models

(McGuire, 1998) found that the method of the previous section led to poor prediction results. Clearly, a naive uniform prior on \mathbf{s} is sub-optimal: recombination events typically

¹More precisely, we should also have noted the dependence on the evolutionary model and its parameters (like the transition-transversion ratio), which we here assume to be known and fixed. See (Durbin *et al.*, 1998) for further details.

results in the exchange or transfer of regions of DNA consisting of many bases, which leads to strong correlations between the topologies at adjacent positions in the alignment. This is not captured by the uniform prior. In order to introduce spatial correlations at the lowest possible order, (McGuire, 1998) introduced a prior on the state sequences in form of a first-order Markov process:

$$P(\mathbf{s}) = P(s_1) \prod_{t=2}^N P(s_t | s_{t-1}) \quad (3)$$

Assuming again independent mutations, $P(\mathbf{Y}|\mathbf{s}) = \prod_{t=1}^N P(\mathbf{y}_t | s_t)$, and applying Bayes' rule, $P(\mathbf{s}|\mathbf{Y}) \propto P(\mathbf{Y}|\mathbf{s})P(\mathbf{s})$, this gives:

$$P(\mathbf{s}|\mathbf{Y}) \propto P(s_1) \prod_{t=2}^N P(s_t | s_{t-1}) \prod_{t=1}^N P(\mathbf{y}_t | s_t) \quad (4)$$

This expansion is of the form of a hidden Markov model (HMM), which is discussed at length in (Rabiner, 1989). Following (McGuire *et al.*, 2000), we define the transition probabilities $P(s_t | s_{t-1})$ as a simple function of the global recombination parameter ν (which is the probability that no recombination occurs):

$$P(s_t | s_{t-1}) = \nu \delta(s_t, s_{t-1}) + \frac{1 - \nu}{K - 1} [1 - \delta(s_t, s_{t-1})] \quad (5)$$

The emission probabilities $P(\mathbf{y}_t | s_t, \mathbf{w}_{s_t})$ are defined by the chosen evolution model (see, for instance, (Durbin *et al.*, 1998)) and depend on the topology of the phylogenetic tree, $s_t \in \{1, \dots, K\}$, and the respective vector of branch lengths, \mathbf{w}_{s_t} . To simplify our notation, we introduce the accumulated vector of all branch lengths in all possible topologies, $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_K)$, and define $P(\mathbf{y}_t | s_t, \mathbf{w}_{s_t}) = P(\mathbf{y}_t | s_t, \mathbf{w})$. This means that s_t indicates which subvector of \mathbf{w} applies.

Parameter Estimation

In (McGuire *et al.*, 2000), the branch lengths \mathbf{w} were optimized heuristically, while the recombination parameter ν was kept fixed. We here propose an improved method that aims to optimize the parameters of the model in a maximum likelihood sense, that is, maximize

$$L(\mathbf{q}) = \ln P(\mathbf{Y}|\mathbf{q}) \quad (6)$$

where \mathbf{q} is the vector of all adaptable parameters, $\mathbf{q} = (\mathbf{w}, \nu)$. In principle this could be achieved with a gradient ascent scheme, but this would involve a summation over all possible K^N combinations of hidden states: $P(\mathbf{Y}|\mathbf{q}) = \sum_{\mathbf{s}} P(\mathbf{Y}, \mathbf{s}|\mathbf{q})$. Obviously, such an approach becomes untractable for long sequences, $N \gg 1$. A viable alternative, however, is the expectation maximization (EM) algorithm (Dempster *et al.*, 1977), which is based on the following decomposition of the log likelihood (Neal & Hinton, 1999):

$$L(\mathbf{q}) = U(\mathbf{q}) + KL(Q, P) \quad (7)$$

$$U(\mathbf{q}) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln P(\mathbf{Y}, \mathbf{s}|\mathbf{q}) - \sum_{\mathbf{s}} Q(\mathbf{s}) \ln Q(\mathbf{s}) \quad (8)$$

$$KL(Q, P) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \left(\frac{Q(\mathbf{s})}{P(\mathbf{s}|\mathbf{Y}, \mathbf{q})} \right) \quad (9)$$

Here, $Q(\mathbf{s})$ is an arbitrary probability distribution over the hidden states, that is, the sequences of tree topologies, and KL represents the Kullback-Leibler divergence between the distributions Q and $P(\mathbf{s}|\mathbf{Y}, \mathbf{q})$. Note that $KL(Q, P)$ is always non-negative (and zero if and only if $Q = P$), which implies that U is a lower bound on L : $U(\mathbf{q}) \leq L(\mathbf{q})$. EM alternates between optimizing the distribution over hidden states $Q(\mathbf{s})$ (the E-step) and optimizing the parameters given $Q(\mathbf{s})$ (the M-step). The E-step holds the parameters fixed and sets Q to the posterior distribution over the hidden states given the parameters, $Q(\mathbf{s}) = P(\mathbf{s}|\mathbf{Y}, \mathbf{q})$. This sets $KL(Q, P) = 0$ and, consequently, $L(\mathbf{q}) = U(\mathbf{q})$. The M-step holds the distribution $Q(\mathbf{s})$ fixed and computes the parameters \mathbf{q} that maximize U . Since $L(\mathbf{q}) = U(\mathbf{q})$ at the beginning of the M-step, and since the E-step does not affect the model parameters, each EM cycle is guaranteed to increase the likelihood unless the system has already converged to a (local) maximum (or, less likely, a saddle point).

In the context of HMMs, the E-step is carried out with the *forward-backward* algorithm (Rabiner, 1989). Hence all that remains to be done is to derive update equations for the parameters in the M-step, that is, to maximize the function U defined in (8). Inserting (4) and (5) into (8) gives

$$\begin{aligned} U &= \sum_{\mathbf{s}} Q(\mathbf{s}) \left[\sum_{t=1}^N \ln P(\mathbf{y}_t | s_t, \mathbf{w}) + \sum_{t=2}^N \ln P(s_t | s_{t-1}, \nu) \right] + C \\ &= \sum_{\mathbf{s}} Q(\mathbf{s}) \left[\sum_{t=1}^N \ln P(\mathbf{y}_t | s_t, \mathbf{w}) + \sum_{t=2}^N \left\{ \delta(s_t, s_{t-1}) \ln \nu + [1 - \delta(s_t, s_{t-1})] \ln \left(\frac{1 - \nu}{K - 1} \right) \right\} \right] + C \end{aligned}$$

Introducing the definition

$$\Psi = \sum_{\mathbf{s}} \sum_{t=2}^N Q(\mathbf{s}) \delta(s_t, s_{t-1}) = \sum_{t=2}^N \sum_{s_t=1}^K Q(s_t, s_{t-1} = s_t) \quad (10)$$

and noting that

$$\sum_{\mathbf{s}} \sum_{t=2}^N Q(\mathbf{s}) [1 - \delta(s_t, s_{t-1})] = N - 1 - \Psi \quad (11)$$

this can be written in the following simplified form:

$$U = \sum_{\mathbf{s}} Q(\mathbf{s}) \sum_{t=1}^N \ln P(\mathbf{y}_t | s_t, \mathbf{w}) + \Psi \ln \nu + (N - 1 - \Psi) \ln \left(\frac{1 - \nu}{K - 1} \right) + C \quad (12)$$

Setting the derivative of U with respect to ν to zero, we obtain:

$$\frac{\partial U}{\partial \nu} = \frac{\Psi}{\nu} - \frac{N - 1 - \Psi}{1 - \nu} = 0 \Rightarrow \nu = \frac{\Psi}{N - 1} \quad (13)$$

This optimization is straightforward since, as seen from (10), Ψ only depends on $Q(s_{t-1}, s_t)$, which is obtained by application of the forward-backward algorithm (Rabiner, 1989). For optimizing the branch lengths, note that only the first term on the left-hand side of (12) depends on \mathbf{w} . This requires a maximization of

$$U(\mathbf{w}) = \sum_{\mathbf{s}} Q(\mathbf{s}) \sum_{t=1}^N \ln P(\mathbf{y}_t | s_t, \mathbf{w}) = \sum_{t=1}^N \sum_{s_t=1}^K Q(s_t) \ln P(\mathbf{y}_t | s_t, \mathbf{w}) \quad (14)$$

where $Q(s_t)$ is the t th marginal distribution obtained from $Q(\mathbf{s})$.

The implementation of the parameter update scheme is straightforward and can be accomplished with the following algorithm²:

1. Initialize the branch lengths \mathbf{w} and the recombination parameter ν .
2. Compute $Q(s_t)$ and $Q(s_{t-1}, s_t)$ with the forward-backward algorithm for HMMs.
3. Compute Ψ from (10) and adapt ν according to (13).
4. For $t = 1$ to N : weight the t th column in the multiple sequence alignment, \mathbf{y}_t , by $Q(s_t)$, and optimize the branch lengths \mathbf{w} so as to maximize $U(\mathbf{w})$ in (14). This can be achieved with a standard phylogeny program – the only modification required is the introduction of a weighting scheme for the sites in the alignment.
5. Test for convergence. If the algorithm has not yet converged, go back to step 2.

4 Simulation Experiments

In this section, we compare the performance of the new parameter optimization algorithm with the method of (McGuire, 1998) and (McGuire *et al.*, 2000). The latter approach is also based on an HMM, but the parameters are adapted heuristically (HEU) rather than with maximum likelihood (ML). While (McGuire, 1998) obtained significantly better results than with the naive Bayesian approach of Section 2, we show that a further considerable improvement can be achieved with the new algorithm proposed in this article.

4.1 A Synthetic Benchmark Problem

In a first study, we simulated several recombination events according to Figure 1, but with two recombinant zones rather than one. We chose the Kimura 2-Parameter model³ of evolution (see, for instance, (Durbin *et al.*, 1998)) and varied the branch lengths of the original tree, the locations and lengths of the recombinant regions, and the time at which the recombinations occurred (that is, the average number of mutations that happened *after* the recombination event). We estimated the prediction performance obtained from the Viterbi

²A software implementation of this algorithm in MATLAB can be obtained from the following web site: <http://www.bioss.sari.ac.uk/dirk/software/SERAD/INFO.html>.

³We chose a fixed transition-transversion ratio of $\tau = 2$.

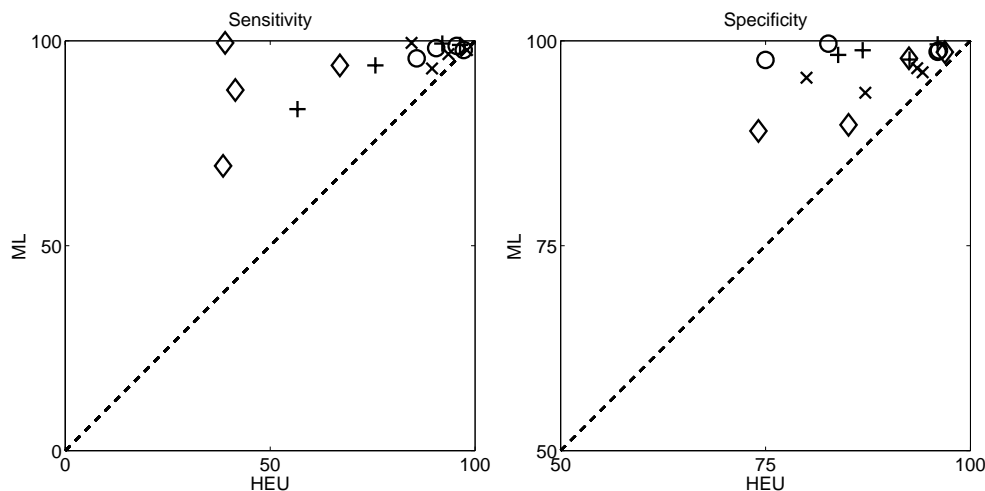


Figure 2: Scatter plots of the classification scores obtained with the heuristic method of (McGuire *et al.*, 2000) (HEU, horizontal axis) and the new algorithm described in this paper (ML, vertical axis). *Left*: Sensitivities (in percent). *Right*: Specificities (in percent). The horizontal line indicates an equal performance, for symbols above this line the new algorithm outperforms the earlier heuristic approach. The symbols represent different locations and sizes of the recombinant regions.

path, that is, the mode of $P(\mathbf{s}|\mathbf{Y}) = P(s_1, \dots, s_N|\mathbf{Y})$, with two different scores. The *sensitivity* is the probability for correctly identifying a recombinant site. The *specificity* is the probability for correctly identifying a non-recombinant site. Figure 2 compares the new method described in this paper (ML) with the earlier heuristic approach of (McGuire *et al.*, 2000) (HEU). The former is clearly seen to be superior with respect to both classification scores, which was formally confirmed with a matched *t*-test (at a 95% significant level).

4.2 Detection of Recombination in *Neisseria*

In the second simulation experiment, we applied our method to a real DNA alignment with evidence of a likely recombination event. The data used was a $N = 787$ base-pair subset of the *Neisseria argF* DNA multiple alignment, studied by (Zhou & Spratt, 1992). We selected four strains of *Neisseria*: *N. gonorrhoeae*, *N. meningitidis*, *N. cinerea*, and *N. mucosa*.⁴ Based on the Kimura 2-parameter evolution model⁵, we performed 12 parameter-estimation simulations with both the HEU and the ML methods, starting from 4 different initial branch-length vectors \mathbf{w} and 3 initial recombination parameters ν .⁶

To compare the performance of the two algorithms, Figure 3, left, shows a scatter plot of the normalized log likelihood scores obtained with the two methods, where the hori-

⁴The strains have the GenBank/EMBL accession numbers X64860, X64866, X64869, X64873, respectively. Note that (Zhou & Spratt, 1992) used a different labelling scheme, with the first nucleotide at $t = 296$, and the last one at $t = 1082$. The alignment of these sequences was carried out using CLUSTAL W with the default parameter settings.

⁵The transition-transversion ratio was set to $\tau = 2.3$, as estimated in (McGuire, 1998).

⁶ $\mathbf{w} = [0.1 \ 0.1 \ 0.1 \ 0.1 \ 0.1]$, $[0.1 \ 0.2 \ 0.1 \ 0.2 \ 0.1]$, $[0.2 \ 0.1 \ 0.2 \ 0.1 \ 0.2]$, $[0.2 \ 0.2 \ 0.2 \ 0.2 \ 0.2]$; $\nu = 0.6, 0.75, 0.9$.

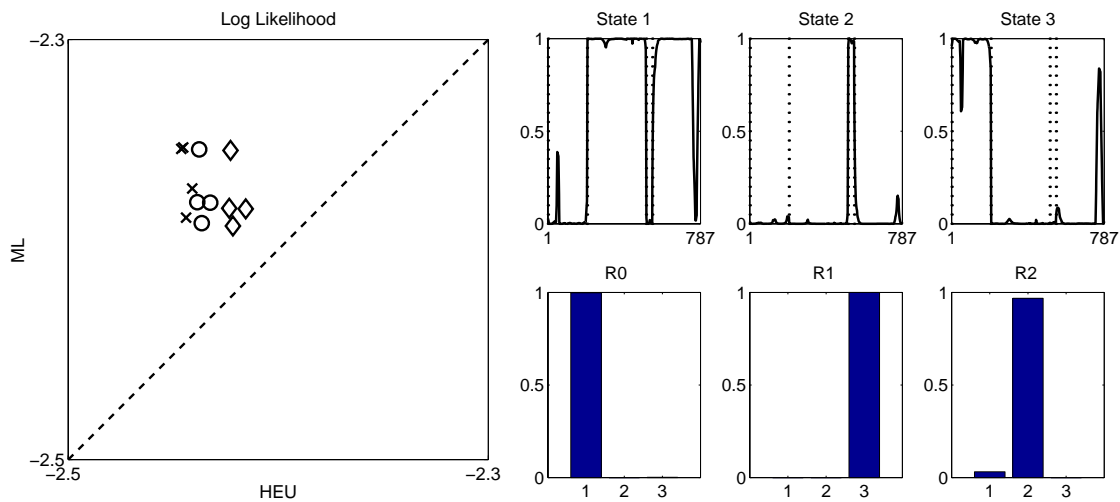


Figure 3: LEFT: Scatter plot of the normalized log likelihood $L^\circ = L/N$ obtained on the *Neisseria* data. *Horizontal axis*: heuristic algorithm, HEU. *Vertical axis*: new algorithm, ML. The diagonal line indicates an equal performance of the two methods, for entries above this line ML is superior. The symbols indicate different initial values for the recombination parameter ν ; x: $\nu_0 = 0.6$; o: $\nu_0 = 0.75$, d: $\nu_0 = 0.9$. Note the strong dependence of HEU on ν_0 , which follows from the fact that this parameter is not adapted by the training algorithm. RIGHT: Prediction of recombinant regions in the *Neisseria* data. The *top row* shows a plot of the posterior probabilities $P(s_t | \mathbf{Y})$ along the sequence alignment, where s_t represents one of the three tree topologies $s_t = 1$ (*left graph*), $s_t = 2$ (*middle graph*), and $s_t = 3$ (*right graph*). The vertical lines mark candidate regions for recombination. The histograms in the *bottom row* show the classification scores in the following regions: *Left*: R0 ($t = 203 - 506, 539 - 787$), classified as $s_t = 1$ in earlier work. *Middle*: R1 ($t = 1 - 202$), believed to result from a recombination equivalent to a transition into $s_t = 3$. *Right*: R2 ($t = 507 - 538$), an irregular region not classified before.

horizontal line represents HEU and the vertical line ML. The dashed diagonal line indicates where the two approaches are equal. All simulations lead to entries that are located far above the diagonal line, which clearly demonstrates that our new algorithm ML outperforms HEU. This was confirmed with a matched t -test, which rejected the null hypothesis of equal performance at a 95% significance level.

(Zhou & Spratt, 1992) found two anomalous regions in the *Neisseria argF* DNA sequence alignment. The two regimes occur at positions $t = 1 - 202$ (region *R1*) and $t = 507 - 538$ (region *R2*). In the rest of the sequence (region *R0*), *N.gonorrhoeae* clusters with *N.meningitidis* ($s_t = 1$) while in *R1* they found that it is grouped with *N.mucosa* ($s_t = 3$). The authors were unable to determine the cause of region *R2*.

In order to compare with this previous study, we selected the HMM with the highest likelihood score and computed the mode of $P(\mathbf{s} | \mathbf{Y}) = P(s_1, \dots, s_N | \mathbf{Y})$ (Viterbi path, see (Rabiner, 1989)). From this we determined the classification scores in the three regions *R0*, *R1*, *R2*, which are shown as histograms in the bottom row on the right of Figure 3. It is clearly seen that, in agreement with (Zhou & Spratt, 1992), most of the sites in *R0* are classified as topology 1, $s_t = 1$, while most of the sites in *R1* are classified as topology 3, $s_t = 3$. Also, region *R2* is classified as different from the topology in region *R0*, again in agreement with (Zhou & Spratt, 1992).

5 Conclusion

This article has discussed the application of hidden Markov models to the detection of sporadic recombination in DNA sequence alignments and has shown how all model parameters can be optimized in a maximum likelihood sense. We have tested the new algorithm on a set of synthetic DNA sequence alignments, where we have found a significant improvement in the prediction performance over the heuristic approach of (McGuire *et al.*, 2000). On a real DNA sequence alignment, the new algorithm achieved a significantly higher likelihood score than the heuristic method, and detected the same putative recombinant regions as an earlier, independent study (Zhou & Spratt, 1992).

References

- Baldi, P. & Brunak, P. (1998). *Bioinformatics - The Machine Learning Approach*. MIT Press.
- 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*, **B39** (1), 1–38.
- Durbin, R., Eddy, S. R., Krogh, A., & Gruber, M. (1998). *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.
- McGuire, G. (1998). *Statistical Methods for DNA Sequences: Detection of Recombination and Distance Estimation*. PhD thesis University of Edinburgh.
- McGuire, G., Wright, F., & Prentice, M. (2000). A Bayesian Method for Detecting Recombination in DNA Multiple Alignments. *Journal of Computational Biology*, **7** (1/2), 159–170.
- Neal, R. M. & Hinton, G. E. (1999). A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants. In: *Learning in Graphical Models*, (Jordan, M. I., ed) pp. 355–368, Cambridge, MA: MIT Press.
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77** (2), 257–286.
- Robertson, D. L., Sharp, P. M., McCutchan, F. E., & Hahn, B. H. (1995). Recombination in HIV-1. *Nature*, **374**, 124–126.
- Zhou, J. & Spratt, B. G. (1992). Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Molecular Microbiology*, **6**, 2135–2146.