Fast Evaluation of the Likelihood of an HMM: Ion Channel Currents with Filtering and Colored Noise

Donald R. Fredkin and John A. Rice

***Abstract*—Hidden Markov models (HMMs) have been used in the study of single-channel recordings of ion channel currents for restoration of idealized signals from noisy recordings and for estimation of kinetic parameters. A key to their effectiveness from a computational point of view is that the number of operations to evaluate the likelihood, posterior probabilities and the most likely state sequence is proportional to the product of the square of the dimension of the state space and the length of the series. However, when the state space is quite large, computations can become infeasible. This can happen when the record has been lowpass filtered and when the noise is colored. In this paper, we present an approximate method that can provide very substantial reductions in computational cost at the expense of only a very small error. We describe the method and illustrate through examples the gains that can be made in evaluating the likelihood.**

I. Introduction

HIDDEN Markov models (HMMs) have recently found application to the analysis of single-channel recordings, both for the construction of an idealized quantal signal from a noisy recording [4], [9] and for estimation of kinetic parameters directly from the recording rather than from an idealized reconstruction [2], [10], [19], [17]. HMMs have also been used in a variety of other areas, for example, in speech recognition [18] and gene finding [14]. A key to their computational effectiveness is that the number of operations required to evaluate the likelihood or its gradient or to evaluate posterior probabilities is proportional to the product of the square of the dimension (D) of the state space and the length of the record (T) [3].

Filtering and colored noise complicate the application of hidden Markov methodology to ion channel recordings. In principle, the state space can be enlarged to include "metastates" [9], [19], and the standard algorithms can be used. In practice, however, the dimensionality of the new state space can easily become so large that computations are intractible. For example, if the underlying state space has cardinality six and a filter of length five is used, the number of operations required to evaluate the likelihood is of order 6^6T rather than 6^2T—a factor of more than 1000. The problem of large state-space dimension also occurs in other extensions of HMMs, for example, [11].

Manuscript received August 9, 1999; revised October 24, 2000. The associate editor coordinating the review of this paper and approving it for publication was Prof. Scott C. Douglas.

D. R. Fredkin is with the Department of Physics, University of California, San Diego, La Jolla, CA 92093 USA.

J. A. Rice is with the Department of Statistics, University of California, Berkeley, Berkeley, CA 94720 USA (e-mail: rice@stat.berkeley.edu).

Publisher Item Identifier S 1053-587X(01)01408-8.

In this paper, we propose and illustrate an approximation strategy that can radically decrease the number of operations required to evaluate the likelihood while entailing little loss in accuracy. The basic idea is to ignore metastates that are either *a priori* or *a posteriori* highly unlikely. In an example to be presented in detail below, the number of operations is reduced by a factor of about 400.

The remainder of this paper is organized as follows. In Section II, we describe the HMM that relates a kinetic model to an observed noisy digital recording and show how it can be extended to account for filtering and colored noise. We then show how the basic recursions of [3] can be accomplished for the extended model and introduce approximations that produce lower bounds on the likelihood. Finally, in Section II-E, we describe the way we have implemented evaluation of the likelihood and our approximations. A collection of examples motivated by models that have been proposed for ion channel kinetics are presented in Section III. Here, we examine in some detail the savings that can be accomplished via our approximations and the size of the errors consequently incurred. Section IV contains a summary, conclusions, and discussion of further directions.

II. Theory

A. Model

We assume that an N_s–state Markov process underlies the kinetics. We consider a discrete time process since in practice, the data are samples at times $k\Delta t$. The one-step transition probabilities P_{ij} for the transition $i \to j$ are related to the generator Q_{ij} of a continuous time Markov process by matrix exponentiation: $P = \exp(Q\Delta t)$.

Current levels $\mathcal{I}_i$ are associated with the states, with the values being, in general, not all distinct. For example, a system with two closed states and one open state would have $\mathcal{I}_1 = \mathcal{I}_2 = 0, \mathcal{I}_3 \neq 0$. Denote the temporal sequence of states by $s(t)$. In the absence of filtering and noise, the observed current would be $x(t) = \mathcal{I}_{s(t)}$. In practice, because of filtering and noise, the observed current is $I(t) = (a * x)(t) + W(t)$, where $a * x$ denotes the convolution $\sum_k a(k)x(t-k)$ and $a(0), a(1), \cdots, a(N_f)$ are filter coefficients; $W(t)$ is additive noise.

In this paper, we assume that the noise $W(t)$ is independent of the state $s(t)$. We will usually assume the noise to be independent identically distributed (IID) Gaussian random variables with mean zero and variance σ^2. However, because we are already prepared to consider the effect of a filter, we can easily consider noise that is an autoregressive (AR) random process

driven by IID Gaussian noise: $b * W = w$, where $b(0) = 1$, and $w(t)$ is IID Gaussian noise with mean zero and variance σ^2. The FIR filter with coefficients b can be considered a prewhitening filter [19]. Applying this filter to the observations $I(t)$, we arrive at

$$y = b * a * x + w = f * x + w \tag{1}$$

where $y = b * I$, and $f = b * a$. The coefficients $b(k)$ can be determined by some variant of the Levinson–Durbin algorithm from the autocorrelation sequence of the noise [19]. If the maximum lag in the sequence $b(k)$ is N_n, the effective filter $f(k)$ has maximum lag $N_e = N_f + N_n$. From now on, we will work with (1), referring to $w(t)$ as the noise and $y(t)$ as the observation at time t. There are T observations at $t = 1 \cdots T$. For most purposes, we do not need the detailed structure of (1); it is sufficient that conditional on the state sequence $s = s(-N_e+1) \cdots s(T)$, the observations $y(t)$ are independent, and the probability density $P(y(t)|s)$ depends only on $s(t) \cdots s(t - N_e)$: $P(y(t)|s) = g(y(t)|s(t) \cdots s(t - N_e))$.

B. Recursive Calculation of the Likelihood

We can include the filter in (1) by extension of the state space [9] and working with a Markov chain whose states are the $N_s^{N_e+1}$ "metastates" $(s_0 \cdots s_{N_e})$. However, the transition matrix among the metastates is sparse, and we find it slightly simpler to work with the original state space and extend the usual recursive procedure [3].

Define

$$\alpha_t(s(t) \cdots s(t - N_e)) = P[y_1 \cdots y_t,\, s(t) \cdots s(t - N_e)] \tag{2}$$

which can be computed recursively. With equilibrium probabilities $\pi(s)$ and transition probabilities $P(s'|s)$, we have

$$\alpha_0(s(0) \cdots s(-N_e)) = \prod_{k=0}^{-N_e+1} P(s(k)|s(k-1))\pi(s(-N_e))$$

and, for $t = 1 \cdots T$

$$\begin{aligned}
&\alpha_t(s(t) \cdots s(t - N_e)) \\
&\quad = \sum_{s(t-N_e-1)} P[y_1 \cdots y_t \,\&\, s(t) \cdots s(t - N_e - 1)] \\
&\quad = \sum_{s(t-N_e-1)} P[y_1 \cdots y_{t-1} \,\&\, s(t-1) \cdots s(t - N_e - 1))] \\
&\qquad \times P[s(t)|y_1 \cdots y_{t-1} \,\&\, s(t-1) \cdots s(t - N_e - 1))] \\
&\qquad \times P[y_t|y_1 \cdots y_{t-1} \,\&\, s(t) \cdots s(t - N_e - 1))] \\
&\quad = \sum_{s(t-N_e-1)} \alpha_{t-1}(s(t-1) \cdots s(t - N_e - 1))) \\
&\qquad \times P(s(t)|s(t-1))g(y_t|s(t) \cdots s(t - N_e)) \\
&\quad = \tilde{\alpha}_{t-1}(s(t-1) \cdots s(t - N_e))P(s(t)|s(t-1)) \\
&\qquad \times g(y_t|s(t) \cdots s(t - N_e))
\end{aligned} \tag{3}$$

where, in the last line, we defined

$$\tilde{\alpha}_t(s(t) \cdots s(t - N_e + 1)) = \sum_{s(t-N_e)} \alpha_t(s(t) \cdots s(t - N_e)).$$

The likelihood is

$$L = P[y] = \sum_{s(T) \cdots s(T-N_e)} \alpha_T(s(T) \cdots s(T - N_e)).$$

In practice, we must renormalize the α_t to avoid underflow. (This procedure was used in [15].) We define

$$N_t = \sum_{s(t) \cdots s(t-N_e)} \alpha_t(s(t) \cdots s(t - N_e)),$$

$$\hat{\alpha}_t(s(t) \cdots s(t - N_e)) = \alpha_t(s(t) \cdots s(t - N_e))/N_t$$

and $\hat{N}_t = N_t/N_{t-1}$. Note that $N_0 = 1$, using the definition of α_0, and N_T is the likelihood. We have

$$\begin{aligned}
&\hat{N}_t\hat{\alpha}_t(s(t) \cdots s(t - N_e)) \\
&\quad = \sum_{s(t-N_e-1)} \hat{\alpha}_{t-1}(s(t-1) \cdots s(t - N_e - 1)) \\
&\qquad \times P(s(t)|s(t-1))g(y_t|s(t) \cdots s(t - N_e))
\end{aligned} \tag{4}$$

and

$$L = \prod_{t=1}^{T} \hat{N}_t. \tag{5}$$

The $\hat{N}_t$ are determined by the requirement that

$$\sum_{s(t) \cdots s(t-N_e)} \hat{\alpha}_t(s(t) \cdots s(t - N_e)) = 1. \tag{6}$$

C. Related Recursive Algorithms

Our focus is on calculation of the likelihood, but we digress briefly to give the form of the EM [3] and Viterbi [20] algorithms using the formalism of Section II-B. We do not necessarily advocate use of the EM algorithm. Some form of quasi-Newton method [8] may be more effective. However, the recursions needed for the EM algorithm can also be regarded as calculations of the posterior probabilities of states given the data, and, as such, can be useful for reconstruction of the ideal signal based on a fictitious HMM. Similarly, the Viterbi algorithm consists of recursions needed to find the most probable state sequence. All of these recursions are complicated by the large numbers of metastates, and our approximations can be applied to all of them.

1) EM Algorithm: Define

$$\begin{aligned}
&\beta_t(s(t) \cdots s(t - N_e + 1)) \\
&\quad = P[y(t+1) \cdots y(T)|s(t) \cdots s(t - N_e + 1)]
\end{aligned} \tag{7}$$

which, like α, can be computed recursively as

$$\beta_T(s(T) \cdots s(T - N_e + 1)) = 1$$

and, for $t < T$

$$\begin{aligned}
&\beta_t(s(t) \cdots s(t - N_e + 1)) \\
&\quad = \sum_{s(t+1)} P(s(t+1)|s(t)) \\
&\qquad \times g(y(t+1)|s(t+1) \cdots s(t - N_e + 1)) \\
&\qquad \times \beta_{t+1}(s(t+1),\, s(t) \cdots s(t - N_e + 2)).
\end{aligned}$$

It is straightforward to show, using a lemma from [3], that the EM algorithm leads to the iteration scheme $p^0 \mapsto p$ for the transition probabilities, where

$$P(b|a) = \frac{\mathcal{N}_{ab}}{\mathcal{D}_a}$$

with

$$\begin{aligned}\mathcal{N}_{ab} &= \sum_s \sum_{t=0}^{T} \tilde{\alpha}_t^0(a,\, s_1 \cdots s_{N_e-1}) p^0(b|a) \\ &\quad \times g(y(t+1)|b,\, a,\, s_1 \cdots s_{N_e-1}) \\ &\quad \times \beta_{t+1}^0(b,\, a,\, s_1 \cdots s_{N_e-2}) \\ \mathcal{D}_a &= \sum_s \sum_{t=0}^{T-1} \tilde{\alpha}_t^0(a,\, s_1 \cdots s_{N_e-1}) \beta_t^0(a,\, s_1 \cdots s_{N_e-1})\end{aligned}$$

and α^0 and β^0 are computed with p^0. (We use the notation $s_1 \cdots s_{Ne}$ to emphasize that these dummy variables are not associated with specific times.)

2) Viterbi Algorithm: The Viterbi algorithm [20] is a dynamic programming method for finding the sequence of states $\{\hat{s}(t)\}$ that is most likely given the observed data. It has been used by [17] for finding a idealized record from which the kinetics parameters are estimated by maximizing the likelihood of the resulting sequence of dwell times. It has also been used in the context of speech recognition by by Juang and Rabiner [13], who maximized the joint likelihood of the kinetic parameters and the sequence of unobserved states rather than the marginal likelihood of the kinetic parameters as in standard maximum likelihood estimation. To formulate the Viterbi algorithm in the case of filtering and colored noise, we follow the notation of [9]. Let

$$H_t(s(t-N_e),\, \cdots,\, s(t)) = \{\hat{s}(-N_e),\, \cdots,\, \hat{s}(t-N_e-1)\} \quad (8)$$

be the most likely state sequence up to and including time $t - N_e - 1$. It maximizes

$$\begin{aligned}&P(s(-N_e),\, \cdots,\, s(t),\, y_0,\, \cdots,\, y_r) \\ &\quad = P(s(-N_e),\, \cdots,\, s(0)) g(y_0|s(-N_e,\, \cdots,\, s(0)) \\ &\qquad \times \prod_{k=0}^{t-1} P(s(k+1)|s(k)) \\ &\qquad \times g(y_{k+1}|s(k-n_e+1),\, \cdots,\, s(t),\, y_0,\, \cdots,\, y_t).\end{aligned}$$

Let

$$\begin{aligned}L_t(s(t-N_e), \cdots, s(t)) = P(&\hat{s}(-N_e), \cdots, \hat{s}(t-N_e-1) \\ &s(t-N_e), \cdots, s(t), y_0, \cdots, y_t).\end{aligned}$$

Then, L_t satisfies the recursion

$$\begin{aligned}&L_{t+1}(s(t-N_e+1),\, \cdots,\, s(t+1)) \\ &\quad = \max_{s(t-N_e)} L_t(s(t-N_e),\, \cdots,\, s(t))\, P(s(t+1)|s(t)) \\ &\qquad \times g(y_{t+1}|s(t-N_e+1,\, \cdots,\, s(t+1)). \quad (9)\end{aligned}$$

Denote the maximizer by $\hat{s}(t-N_e)$. H_t then also satisfies the recursion relation

$$\begin{aligned}&H_{t+1}(s(t-N_e+1),\, \cdots,\, s(t+1)) \\ &\quad = H_t(s(t-N_e)\, \cdots,\, s(t)) \circ \hat{s}(t-N_e)\end{aligned}$$

where $\circ$ denotes concatenation. Note that L_T is the likelihood of the state sequence s, which is to be maximized.

D. Approximations

Consider the computational cost of using (4)–(6) to compute the likelihood. For each time t, we compute $N_s^{N_e+1}$ values of α (one for each of the metastates $s_0 \cdots s_{N_e}$), and each such computation requires order N_s operations. Calculation of $\hat{N}_t$ requires $N_m - 1$ additions. Calculation of the likelihood thus takes $O(N_s^{N_e+2}T)$ floating-point operations. If we compare this with the computational cost when there is neither a filter nor autoregressive noise coloration, we see that the work is multiplied by a factor $N_s^{N_e}$. For a simple scheme involving three states ($N_s = 3$) and maximum lag due to filtering and noise coloration $N_e = 10$, we have a cost amplification of $3^{10} = 59\,049$. If, to be optimistic, we could compute the likelihood for $N_e = 0$ in 1 μs, we now require a full second to compute the likelihood once, and we will need to compute the likelihood many times to maximize it.

The key to speeding up the calculation of the likelihood is the observation that the exact scheme, whether in the efficient form (4)–(6) or in the raw form

$$L = P[y] = \sum_s P[y \,\&\, s] \quad (10)$$

where y and s are histories ($y_1 \cdots y_T$ and $s_{-N_e+1} \cdots s_T$), involves a large number of improbable and numerically unimportant sequences of states. For example, in a two-state model ("closed" and "open"), the transition probabilities (P_{12}, P_{21}) are likely to be extremely small. If they were not, we would say that the sampling interval Δt was too large. If $N_e = 4$, say, we expect to encounter metastates containing multiple transitions (like $C \to O \to C \to O \to C$) rarely, and their contribution to the sum in (10), or the role of any α for such a metastate, might be negligible.

Our primary approximation is to choose a small tolerance ϵ_1 and neglect any metastate $s_0 \cdots s_{N_e}$ for which the conditional probability

$$P[s_1 \cdots s_{N_e}|s_0] < \epsilon_1. \quad (11)$$

We discuss quantitatively the effective reduction in the number of metastates and in the computation time in Section III for a variety of realistic examples and choices of ϵ_1.

The selection of metastates to be neglected based on (11) is made once at the beginning of the calculation of the likelihood. The selection depends, of course, on the transition probabilties; therefore, the selection must be made repeatedly in the course of maximization of the likelihood once each time the likelihood is evaluated.

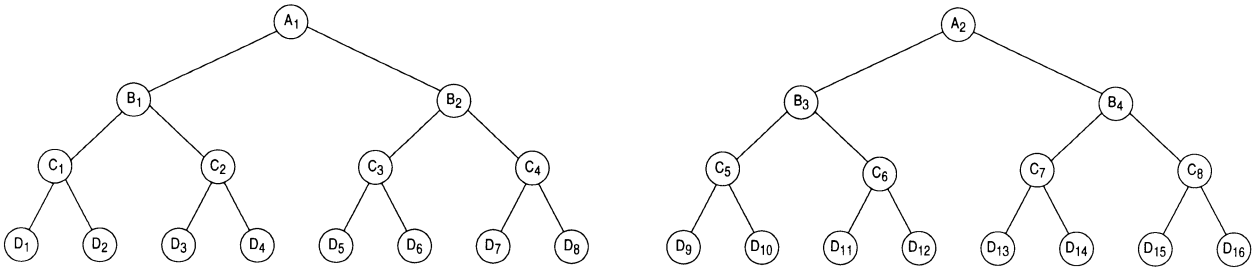


Fig. 1. Full forest, before any approximations, for $N_s = 2$ and $N_e = 3$. The labels have no particular significance. See Table I for information associated with the various nodes.

We can make a second approximation of a more dynamical character. Whenever, in evaluating (4), we encounter a value

$$\sum_{s(t-N_e-1)} \hat{\alpha}_{t-1}(s(t-1)\cdots s(t-N_e-1)) < \epsilon_2 \qquad (12)$$

we replace the sum by zero. Note that this sum is the renormalized version of $\tilde{\alpha}_{t-1}(s(t-1)\cdots s(t-N_e))$. The elimination of terms using (12) depends on the data y, whereas the simplification using (11) depends only on the model and not at all on the data. The utility of this approximation is also discussed in Section III.

As will be shown in Section III, quite substantial reduction in computational cost can be achieved with small values of ϵ_1 and ϵ_2 (e.g., 10^{-6}).

Similar approximations can be applied to the EM and Viterbi algorithms. For example , in the Viterbi algorithm, note that one has to update L_t as in (9) for each of its $N_s^{N_e+1}$ arguments (metastates). An approximation that discards those metastates that have small *a priori* probability can drastically reduce the total number of calculations. In addition, if $g(y_{t+1}|s(t-N_e+1),\cdots,s(t+1))$ is small, an approximation can be made in which $L_{t+1}(s(t-N_e+1),\cdots,s(t+1))$ is set equal to zero and then ignored in the step $t+1 \rightarrow t+2$. A similar approximation of the latter kind is used in speech processing [16].

E. Implementation

We use (4)–(6) to compute the likelihood. In this section, we discuss some design decisions we made when implementing the calculation on a computer.

We must store values of $\hat{\alpha}_t(s_0 \cdots s_{N_e})$ and update them as t ranges from 0 to T. There are many indices, each with a modest range, and the number of indices depends on the model. This suggests that a multidimensional array, with many nested loops to manipulate the values as t progresses from 0 to T, might not be the best scheme. We prefer to keep track of the various values in a forest of N_s ordered trees. (We use the terminology of [1] throughout this section.) Let us use a simple example for ease of exposition. The model structure is defined by $N_s = 2$ and $N_e = 3$, and the transition matrix is, for illustrative purposes

$$P = \begin{pmatrix} 0.99 & 0.01 \\ 0.005 & 0.995 \end{pmatrix}.$$

The general case does not involve anything new, and the discussion would become excessively abstract. The general case is documented in our source code, using the C programming language.

TABLE I
INFORMATION STORED IN THE NODES OF FIG. 1. "NODE" IS THE LABEL IN FIG. 1. "HISTORY" IS THE SEQUENCE OF STATES REPRESENTED BY THE NODE. "P" IS THE CONDITIONAL PROBABILITY OF THE PARTIAL HISTORY. THE LAST COLUMN INDICATES WHETHER OR NOT THE NODE IS ELIMINATED ("PRUNED") WHEN $\epsilon_1 = 0.001$. NOTE THAT D_5 AND D_6 ARE AUTOMATICALLY PRUNED BECAUSE C_3 IS, AND, SIMILARLY, D_{11} AND D_{12} ARE ELIMINATED WHEN C_6 IS PRUNED

Node	History	P	Prune?
A_1	0	1.	no
A_2	1	1.	no
B_1	00	0.99	no
B_2	01	0.01	no
B_3	10	0.005	no
B_4	11	0.995	no
C_1	000	0.9801	no
C_2	001	0.0099	no
C_3	010	0.0005	yes
C_4	011	0.00995	no
C_5	100	0.00495	no
C_6	101	0.00005	yes
C_7	110	0.004975	no
C_8	111	0.990025	no
D_1	0000	0.970299	no
D_2	0001	0.009801	no
D_3	0010	0.0000495	yes
D_4	0011	0.0098505	no
D_5	0100	0.0000495	yes
D_6	0101	0.0000005	yes
D_7	0110	0.00004975	yes
D_8	0111	0.00990025	no
D_9	1000	0.0049005	no
D_{10}	1001	0.0000495	yes
D_{11}	1010	0.00000025	yes
D_{12}	1011	0.00004975	yes
D_{13}	1100	0.00492525	no
D_{14}	1101	0.00004975	yes
D_{15}	1110	0.004950125	no
D_{16}	1111	0.985074875	no

We start by constructing N_s trees (Fig. 1). Each node represents a partial state history, starting at the roots, corresponding to individual states, and descending to the leaves, which represent metastates, so that the history corresponding to a node of depth d has length $d+1$ (see the second column of Table I). We store the probability of the partial state history, conditional on the initial state, in each node; these values are built up recursively as the tree is built (see the third column of Table I). In general, all operations that one might think of performing by means of multiple nested loops are, in fact, done by recursive tree traversals.

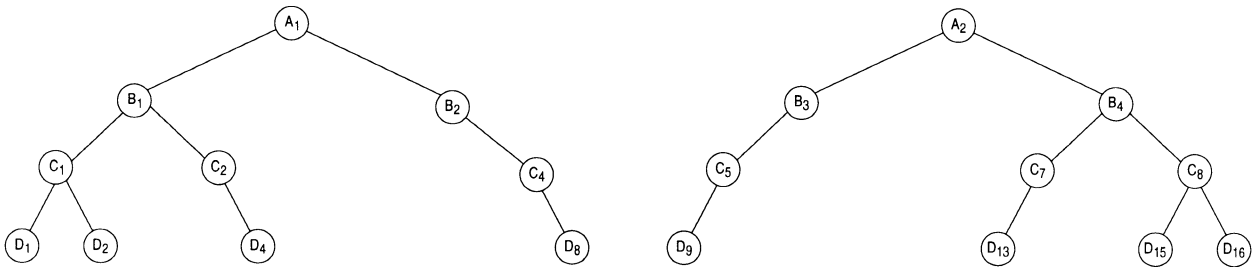


Fig. 2. Forest of Fig. 1 after pruning with $\epsilon = 0.001$.

In practice, we need not build the full tree because we invoke (11) to "prune" the tree as we build it, eliminating any node for which $P < \epsilon_1$ and all of its children. For our example, suppose we choose $\epsilon = 0.001$. Then, we actually build the forest in Fig. 2. It can happen that (11) eliminates all the children of a node without eliminating the node itself; in this case, the node is pruned. At the end of the pruning process, there are no leaves at levels greater than zero.

After pruning, we multiply the stored probabilities in the leaves by the equilibrium probabilities associated with the roots of the trees to obtain values of $\hat{\alpha}_0$. During the same tree traversal, the means of $y(t)$ conditional on the metastate are constructed and stored in the leaves.

We still need to discuss the updating process in which, starting from a forest with $\hat{\alpha}_t$ stored in the leaves, we arrive at a new forest, with the same topology, with $\hat{\alpha}_{t+1}$ in the leaves. Mathematically, we must sum over the oldest state, which is at the roots, to obtain the normalized version of $\tilde{\alpha}_t$, and then, we use the last form of (3). All of the index manipulation in (3) will be done automatically by recursive tree traversals. Consider the subtrees rooted at B_1 and B_3. The "sum" of these will become the part of the new tree rooted at A_1 of level greater than zero, and $\tilde{\alpha}_t$ will be stored in its leaves, which are the nodes of level one in the final tree. In general, when "adding" two trees, we add the αs stored in the leaves, except when some leaves are missing because of pruning. Similarly, the part of the new tree rooted at A_2 of level greater than zero is obtained as the sum of the subtrees rooted at B_2 and B_4. It is then straightforward to compute and store the values of α_{t+1} and carry out the normalization process described by (4)–(6).

III. Examples

We illustrate the computational savings of our method by simulations from three models that have appeared in the ion channel literature. Model I was proposed in [5] for an acetylcholine receptor. When sampled at 10 kHz, the transition matrix of the five-state scheme is

$$P_{\mathrm{I}} = \begin{pmatrix} 0.7373 & 0.0044 & 0.0004 & 0.2325 & 0.0253 \\ 0.0001 & 0.9723 & 0.0219 & 0.0053 & 0.0004 \\ 0.0002 & 0.6579 & 0.1614 & 0.1588 & 0.0217 \\ 0.0012 & 0.0020 & 0.0020 & 0.8142 & 0.1807 \\ 0.0000 & 0.0000 & 0.0000 & 0.0009 & 0.9991 \end{pmatrix}. \tag{13}$$

The channel is open in the first two states ($\mathcal{I} = 1$) and closed in the last three ($\mathcal{I} = 0$). We note that the fifth is a long-lived closed state.

Model II was proposed in [6] for a batrachotoxin-modified sodium channel. It too is a five-state scheme, which when sampled at 10 kHz yields the transition matrix

$$P_{\mathrm{II}} = \begin{pmatrix} 0.9903 & 0.0096 & 0.0000 & 0.0000 & 0.0000 \\ 0.0577 & 0.9330 & 0.0092 & 0.0001 & 0.0000 \\ 0.0017 & 0.0554 & 0.9141 & 0.0274 & 0.0014 \\ 0.0001 & 0.0025 & 0.0822 & 0.9152 & 0.0001 \\ 0.0000 & 0.0003 & 0.0086 & 0.0001 & 0.9910 \end{pmatrix}. \tag{14}$$

The channel is closed in the first three states and open in the last two. The first closed state and the last open state are particularly long lived, with mean durations of about 100 sampling units.

Model III was used in [10] and is derived from another model for a batra chotoxin-modified sodium channel [12]. This model has three states, the first two of which are closed, and when sampled at 10 kHz produces a transition matrix

$$P_{\mathrm{III}} = \begin{pmatrix} 0.9996 & 0.0004 & 0.0000 \\ 0.0090 & 0.9860 & 0.0049 \\ 0.0000 & 0.0093 & 0.9907 \end{pmatrix}. \tag{15}$$

These models share a feature that makes our approximation schemes effective. Many of the entries of the transition matrices are quite small, and the diagonal entries are relatively large, implying that a substantial fraction of metastates have very small probability. Particularly improbable are those with many transitions between different states.

In our simulations, we used a digital approximation to an eight-pole Bessel filter with a cutoff at 2 kHz (a moving average with coefficients $[0.0348,\ 0.4515,\ 0.4556,\ 0.0621,\ -0.0064]$). Our two noise models were white noise and an autoregressive scheme from [19] with coefficients $[1.0,\ 0.7152,\ 0.4900,\ 0.3056,\ 0.1427]$. The convolution of these two sequences, truncated after eight terms and normalized to sum to one, gave a net composite filter with coefficients $[0.0131,\ 0.1799,\ 0.3004,\ 0.2341,\ 0.1526,\ 0.0867,\ 0.0305,\ 0.0026]$. Three different SNRs were used, the innovation standard deviations being 0.05, 0.25, and 0.75. For each of the three kinetic models, for each of the two noise models, and for each of the three signal to noise levels, we simulated 100 000 points, or 10 s of

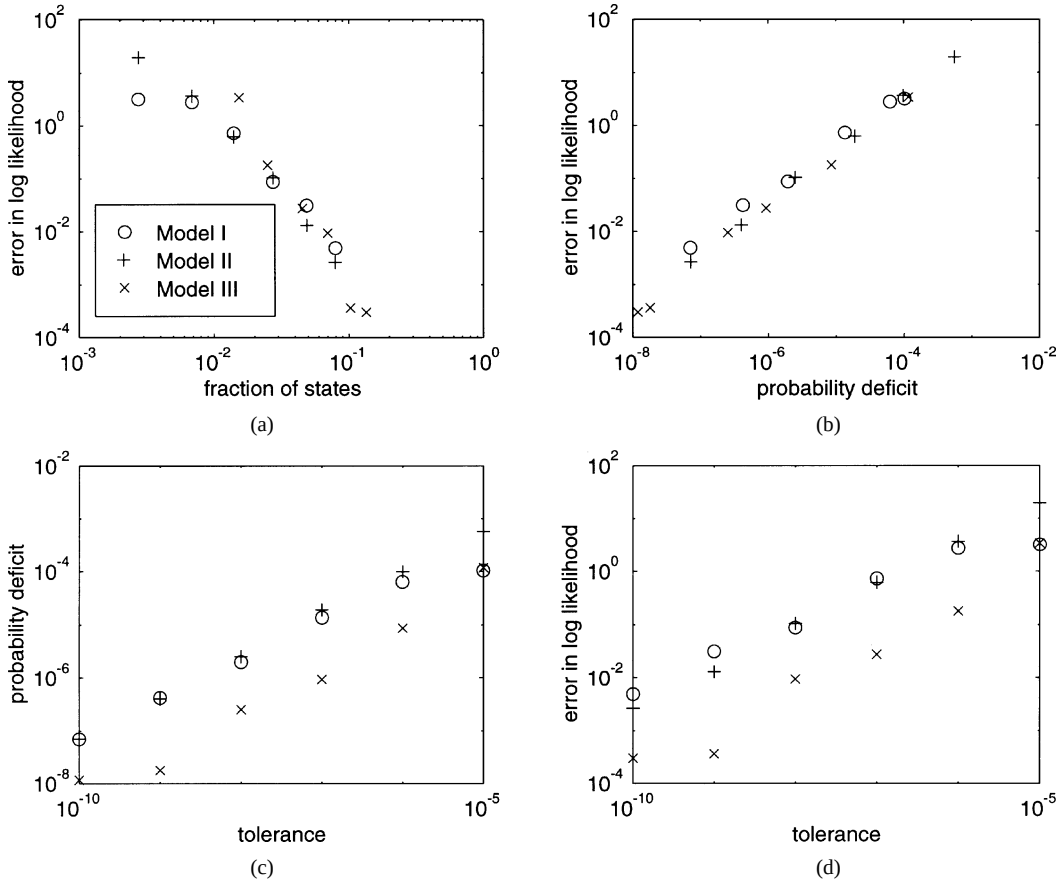


Fig. 3. (a) Error in the approximation to the log likelihood as a function of the fraction of meta-states retained. (b) Error in the approximation to the log likelihood as a function of the total equilibrium probability of the metastates pruned from the tree (the probability deficit). (c) Probability deficit as a function of the tolerance ϵ_1. (d) Error in the approximation to the log likelihood as a function of the tolerance ϵ_1.

data. The computations we report were performed on a Sun UltraSparc 2. Our programs were written in C and linked to Matlab[1] .

We first discuss the results for the autoregressive noise model with innovation standard deviation $\sigma = 0.05$. For a composite filter of length eight, the total number of metastates are $5^8 = 390\,625$ for models I and II and $3^8 = 6561$ for model III. As discussed in the last section, the computational prices to be paid over a model with no filtering and white noise are factors of $5^7 = 78\,125$ and $3^7 = 2187$. For example, if the likelihood took 1 s to evaluate with no filtering and white noise (this figure is roughly accurate), it would take approximately 22 h to evaluate in models I and II, allowing for colored noise and filtering.

As explained in the previous section, we can decrease the effective number of metastates and, proportionally, the time to evaluate the likelihood, by increasing the parameter ϵ_1. Fig. 3(a) shows the resulting error in approximating the likelihood as a function of the fraction of the number of metastates remaining after pruning. (The actual log likelihoods were of order 10^5 for each of the three models.) For example, if the number of metastates of model I is reduced by a factor of 362, the resulting error in the log likelihood is 3.15 out of 1.57×10^5. For model III, reduction of the number of metastates by a factor of 65 resulted in an error of 3.37 out of 1.57×10^5. Although these reductions are large, even with them, computational times are quite substantial. For example, after the number of metastates of model I is reduced by factor of 362, 1077 effective metastates still remain. In fact, evaluation of the likelihood allowing for filtering and colored noise, pruning the number of effective metastates to 1077, took 1687 s, as compared with 1.3 s for evaluation of the likelihood of a model with no filtering and white noise. For model III, the computation of the likelihood took 158 s after reduction of the number of metastates by a factor of 65.

Without specification of the use of the approximate log likelihood, it is difficult to determine an acceptable level of error, but we suggest the following heuristic as a guide. Suppose that $\hat{\theta}$ is the maximum likelihood estimate of an m-dimensional vector of rate constants. A standard large sample theory result [7] is that an approximate $100(1-\alpha)\%$ confidence region for θ is $\{\theta | 2(\ell(\hat{\theta}) - \ell(\theta)) \leq \chi^2_m(\alpha)\}$, where $\ell(\theta)$ is the log likelihood, and $\chi^2_m(\alpha)$ is the upper α percentage point of the chi-square distribution with m degrees of freedom. For example, the underlying kinetic model for model II has six free rate constants that determine the rate matrix Q from which $P_{\rm II} = \exp(Q\Delta t)$ was found. The upper 5% point of the chi-square distribution with six degrees of freedom is 12.59. Thus, the effect of an approximation error of order one in the log likelihood is comparable with the variation in the likelihood due to parameter uncertainty. The effect of the approximation error on optimization is discussed in the concluding section.

[1]The MathWorks, Inc., Natick, MA.

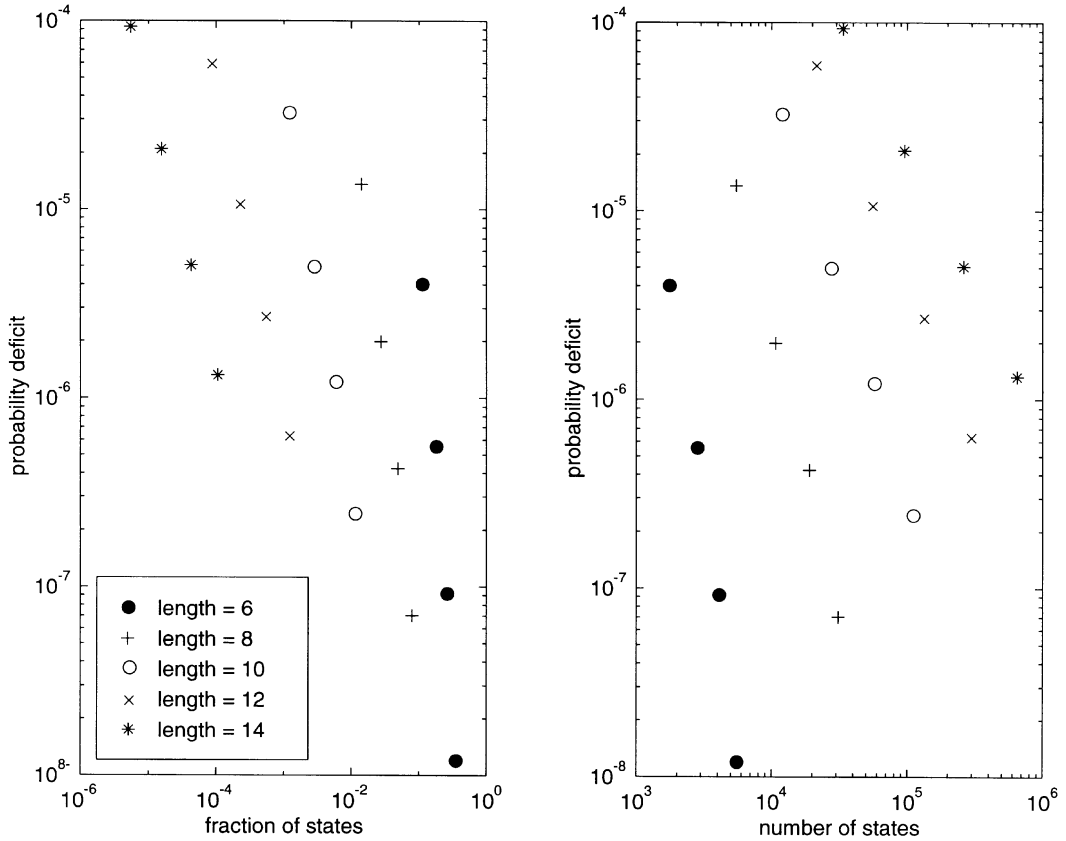


Fig. 4. (a) Probability deficit for model I as a function of the fraction of metastates retained for various filter lengths. (b) Probability deficit as a function of the number of metastates retained for filter lengths as in (a).

As described in the previous section, we prune the number of effective metastates by setting the tolerance parameter ϵ_1. Let the sum of the equilibrium probabilities of the metastates that have been discarded be termed the "probability deficit." Fig. 3(b) shows that the error in the log likelihood is proportional to the probability deficit with a constant of proportionality of order 10^4. The probability deficit induced by pruning of the degree discussed in the examples above is roughly of order 10^{-5}, which we believe is negligble when viewed from a broad perspective in which the model itself is a crude approximation to physical reality. Fig. 3(c) shows how the probability deficit is determined by the tolerance. To complete the picture, Fig. 3(d) shows how the error in the log likelihood is determined by the tolerance ϵ_1. From these figures, we see that the tolerance, the probability deficit, and the fraction of metastates remaining are all equivalent ways of specifying the amount of pruning. We have found it algorithmically most natural to control the amount of pruning by setting ϵ_1 since the pruning can be accomplished as the forest of metastates is traversed.

Very similar results were found at the lower SNRs in that the errors induced in estimating the log likelihood by using a small fraction of the total number of metastates were comparable in order of magnitude with those described above for the three models. For example, for model I, with $\sigma = 0.75$, the total log likelihood was -1.13×10^5, and the error when 1077 metastates were used was 1.00.

We next briefly contrast the results discussed above to those obtained when a lowpass filter is used but noise is white rather than colored. The length of the filter is thus five rather than eight, and the relative gains are smaller. On an absolute scale, the computations are less forbidding. For models I and II and a filter of length five, there are $5^5 = 3125$ metastates as compared with $5^8 = 390\,625$ for a filter of length 8, and gains by factors of about 10 are possible while incurring an error of order one.

Generally, as the length of a filter is increased, the fraction of metastates needed to maintain a given probability deficit decreases rapidly. Fig. 4(a) shows this phenomena for model I and various filter lengths. However, the total number of remaining metastates, and, hence, the time to evaluate the likelihood, continues to increase, as shown in Fig. 4(b). It thus appears that additional computational strategies, such as distributing the computations over a network of workstations, are still needed for very long filters.

Finally, we discuss the savings that can be accomplished by imposing the second tolerance $\epsilon_2 > 0$. In our simulations, we found that with $\epsilon_1 > 0$, decreases in computation time of factors of two to three, with little additional inaccuracy in the approximated log likelihood, could be accomplished by setting ϵ_2 to small values, such as 10^{-9}, when the SNR was high. Further increasing ϵ_2 did not result in substantial consequent savings as most metastates that were *a posteriori* unlikely had already been eliminated. At lower SNRs the effectiveness of ϵ_2 decreased and became insubstantial at $\sigma = 0.75$. This is to be expected since using the second tolerance eliminates, at each time point, metastates that are *a posteriori* unlikely given the observed data, and with a high noise level, the data are relatively uninformative.

As an example, for model II with $\sigma = 0.05$, setting $\epsilon_1 = 10^{-6}$ reduced the number of metastates by a factor of 200—from 390 625 to 1946. With $\epsilon_2 = 0$ the error in the log likelihood of 3.63, setting $\epsilon_2 = 10^{-10}$ reduced the computation time by a further factor of 2.1, giving a net reduction by a factor of about 400, whereas the additional error in the log likelihood was less than 10^{-4}. With this setting of ϵ_2, the average number of metastates discarded per time point was 765 (out of 1946). Examination of the results revealed that when the channel was closed, about 750 metastates were typically discarded, and when it was open (which was less frequent), about 1150 were discarded.

IV. Discussion

We have explained and demonstrated methods that provide dramatic computational gains in the evaluation of the likelihood of an HMM for single-channel recordings contaminated by filtering and colored noise. These gains are achieved by discarding the contributions to the likelihood from metastates that are either *a priori* or *a posteriori* unlikely. We have found it convenient and effective to organize the computations in a tree structure, but other approaches are possible. With our implementation, the greatest gains are made by discarding metastates that are *a priori* unlikely since the pruned branches of the tree are subsequently never traversed during the iterated passes through it. Our methods can be applied to approximate not only the likelihood but its gradient and posterior probabilities as well. The effectiveness of the approach depends on the kinetics of the model. If the kinetics are very fast, relatively few metastates may be ignorable, and the method will be less effective.

It is difficult to give a simple, concrete recipe for choosing the critical parameter ϵ_1, but we can offer some considerations that may help guide its choice.

Although it is algorithmically most natural to use ϵ_1, it is more intuitive to work with the equivalent probability deficit. On *a priori* grounds, one might feel that the suitability or implications of the model should not depend on the inclusion or exclusion of a set of metastates having total probability of order 10^{-5} or 10^{-6}. For example, a comparison of two models should not hinge on such fine structure.

There are rough theoretical grounds for believing that the error in log likelihood is proportional to the probability deficit, as we found empirically in the previous section. Consider first the effect of deleting from the summation in (10) all sequences containing a metastate $(s_t, \cdots, s_{t-N_e})$ occurring at time t. The difference between the exact and approximate likelihood is

$$L_{\text{exact}} - L_{\text{approx}} \approx P(s_t, \cdots, s_{t-N_e})P(y_t|s_t) \cdots P(y_{t-N_e}|s_{t-N_e})L_{\text{exact}} \tag{16}$$

$$= \Delta C_t L_{\text{exact}} \tag{17}$$

where Δ is the probability of that metastate. Now, in the summation (10), there are T sequences in which that metastate occurs exactly once, and the consequent reduction when they are all deleted is $T\Delta C L_{\text{exact}}$, where C is the average of the C_t. There are $T(T-1)/2$ sequences in which the metastate occurs twice, but if Δ is sufficiently small so that $T\Delta C < 1$, the contribution of these and higher numbers of multiple occurrences is relatively negligible. Extending this argument to the omission of M metastates, we then have

$$L_{\text{exact}} - L_{\text{approx}} \approx L_{\text{exact}} T \sum_{m=1}^{M} \Delta_m C_m \tag{18}$$

$$\approx L_{\text{exact}} TC\Delta \tag{19}$$

if the C_m do not vary much or if their covariance with the Δ_m is small relative to the product of their means, and where Δ is the total probability deficit, and C is the average of the C_m. We thus have $\log(L_{\text{approx}}) \approx \log(L_{\text{exact}}) - TC\Delta$. This argument, coupled with the empirical results of the previous section, suggests that if the log likelihood is plotted against the probability deficit, the intercept, and thus the error of the approximation, can be roughly gauged to an order of magnitude.

The desirable accuracy of the approximation depends on how the results are going to be used. In the previous section, we discussed the effects on construction of confidence intervals. If the goal is to construct the Viterbi approximation to the underlying sample path, a sensible way to proceed would be to start with a relatively large tolerance and then relax it, stopping when the changes in the reconstruction became practically negligible.

In this paper, we have concentrated on efficient approximate evaluation of the likelihood but not directly on its maximization. We have not systematically investigated the impact of approximation to the likelihood on the maximum likelihood estimates themselves. However, we did find for Model III and a filter of length five that choosing $\epsilon_1 = 10^{-5}$ and $\epsilon_2 = 0$ led to estimates within a percent of the maximum likelihood estimate ($\epsilon_1 = 0$ and $\epsilon_2 = 0$), with a relative time savings per evaluation of the likelihood of 2.5. More substantial gains were made when the filter was longer. Many additional issues come into play in maximizing the likelihood, but in any case, evaluation of the likelihood function is a key component. Other important components include the choice of starting values and the search strategy. For choice of starting values, it may be effective to maximize the likelihood or an approximation to it on a relatively small segment of data. When working with the full data set, one could initially use these maximizers as starting values and relatively large tolerances to find a new maximum. The tolerances could then be decreased, and the process could be continued until there was little change in the maximizers. Since our approximations work by discarding metastates, they produce lower bounds to the likelihood; the success achieved in maximizing such lower bounds rather than the likelihood itself depends in part on how uniform the bounds are over the relevant parameter space. We have not yet investigated this question, but the observed proportionality of the error in the log likelihood to the probability deficit provides some reason for optimism that maintaining a fairly constant probability deficit as the parameters change would produce nearly uniform lower bounds. Given the time that it takes to evaluate the likelihood function, it is clearly important to use a search strategy that entails a minimum number of function evaluations.

Although we have developed and illustrated the methods in the context of single-channel recordings, we believe that they

may have relevance to other phenomena modeled by HMMs in which the dimensionality of the state space makes exact computation of the likelihood prohibitive or impractical. Within the context of the statistical analysis of patch clamp recordings, we believe that our methods will be especially effective in evaluating the likelihood of superpositions of independent channels. Such superpositions produce a very high dimensional state space that has hindered the successful application of otherwise promising HMM techniques [2].

Our code is written in C to be driven by Matlab, and we will be pleased to share it with anyone who is interested.

REFERENCES

[1] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*. Reading, MA: Addison-Wesley, 1974.

[2] A. Albertson and U.-P. Hansen, "Estimation of kinetic rate constants from multi-channel recordings by a direct fit of the time series," *Biophys. J.*, vol. 67, pp. 1393–1403, 1994.

[3] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, pp. 164–171, 1970.

[4] S. H. Chung, J. B. Moore, L. Xia, L. S. Premkumar, and P. W. Gage, "Characterization of single channel currents using digital signal processing techniques based on hidden Markov models," *Philos. Trans. R. Soc. London*, vol. 329, pp. 265–285, 1990.

[5] D. Colquhoun and A. G. Hawkes, "The principles of the stochastic interpretation of ion-channel mechanisms," in *Single-Channel Recording*, B. Sakmann and E. Neher, Eds. New York: Plenum, 1995, ch. 18.

[6] A. M. Correa, F. Benzanilla, and R. Latorre, "Gating kinetics of batrachotoxin-modified Na^+ channels in the squid giant axon," *Biophys. J.*, vol. 61, pp. 1332–1352, 1992.

[7] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. London, U.K.: Chapman & Hall, 1974.

[8] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. Nedw York: Wiley, 1987.

[9] D. R. Fredkin and J. A. Rice, "Bayesian restoration of single channel patch clamp recordings," *Biometrika*, vol. 48, pp. 427–448, 1992a.

[10] ——, "Maximum likelihood estimation and identification directly from single-channel recordings," *Proc. R. Soc. London*, vol. 249, pp. 125–132, 1992b.

[11] Z. Ghahramani and M. Jordan, "Factorial hidden Markov models," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, vol. 8.

[12] L. Huang, N. Moran, and G. Ehrenstein, "Gating kinetics of batra\-chotoxin-modified sodium channels in neuroblastoma cells determined from single-channel measurements," *Biophys. J.*, vol. 45, pp. 313–324, 1984.

[13] B. Juang and L. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1639–1641, 1990.

[14] A. Krogh, S. Mian, and D. Haussler, "A hidden Markov model that finds genes in e. coli dna," *Nucl. Acids Res.*, vol. 22, pp. 4769–4778, 1994.

[15] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of probabilistic functions of Markov processes to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, pp. 1035–1074, 1983.

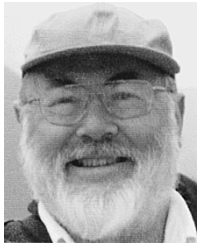
[16] H. Ney and X. Aubert, "Dynamic programming search strategies: From digit strings to large vocabulary word graphs," in *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Boston, MA: Kluwer, 1996, ch. 16.

[17] F. Qin, A. Chen, A. Auerbach, and F. Sachs, "Extracting channel kinetic parameters using hidden Markov techniques," *Biophys. J.*, vol. 66, p. 392, 1994.

[18] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speed processing," *Proc. IEEE*, vol. 77, pp. 257–285, 1989.

[19] L. Venkataramanan, J. L. Walsh, R. Kuc, and F. J. Sigworth, "Identification of hidden Markov models for ion channel currents—Part I: Colored background noise," *IEEE Trans. Signal Processing*, vol. 46, pp. 1901–1915, July 1998.

[20] J. Viterbi, "Error bounds for convolution codes an asymptotically optimal decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 260–269, 1967.



Donald R. Fredkin was born in New York, NY, on September 28, 1935. He received the A.B. degree in mathematics in 1956 from New York University, New York, and the Ph.D. degree in mathematical physics in 1961 from Princeton University, Princeton, NJ.

Since 1961, he has been with the University of California, San Diego, La Jolla, where he is Professor of physics. His research deals with theoretical condensed matter physics, biophysics, and statistical problems arising in neurophysiology. At various times, he has also been associated with Bell Laboratories, the Aerospace Corporation, C.E.N. Saclay, A.E.R.E. Harwell, the West Los Angeles Medical Center of the Veterans' Administration, the Nanogen Corporation, and Seashell Technology.

Dr. Fredkin is a member of the American Physical Society and the Society for Industrial and Applied Mathematics.

John A. Rice was born in New York, NY, on June 14, 1944. He received the B.A. degree in mathematics from the University of North Carolina, Chapel Hill, in 1966 and the Ph.D. degree in statistics in 1972 from the University of California, Berkeley.

He was with the Department of Mathematics, University of California, San Diego, La Jolla, from 1973 to 1991, and since 1991, he has been with the University of California, Berkeley, where is a Professor of statistics. His research interests include applied and theoretical statistics.

Dr. Rice is a member of the Institute of Mathematical Statistics, the American Statistical Association, and the International Statistical Institute.