A Direct Optimization Approach to Hidden Markov Modeling for Single Channel Kinetics

Feng Qin, Anthony Auerbach, and Frederick Sachs
Department of Physiology and Biophysics, State University of New York at Buffalo, Buffalo, New York 14214 USA

ABSTRACT Hidden Markov modeling (HMM) provides an effective approach for modeling single channel kinetics. Standard HMM is based on Baum's reestimation. As applied to single channel currents, the algorithm has the inability to optimize the rate constants directly. We present here an alternative approach by considering the problem as a general optimization problem. The quasi-Newton method is used for searching the likelihood surface. The analytical derivatives of the likelihood function are derived, thereby maximizing the efficiency of the optimization. Because the rate constants are optimized directly, the approach has advantages such as the allowance for model constraints and the ability to simultaneously fit multiple data sets obtained at different experimental conditions. Numerical examples are presented to illustrate the performance of the algorithm. Comparisons with Baum's reestimation suggest that the approach has a superior convergence speed when the likelihood surface is poorly defined due to, for example, a low signal-to-noise ratio or the aggregation of multiple states having identical conductances.

INTRODUCTION

Patch-clamp recording is a primary tool for studying the function of ion channels. The technique allows the ionic currents flowing through an individual channel protein molecule to be measured directly. The data provide information on many aspects of channel behaviors. The amplitude of the current describes the permeability of ions through the conducting pathway, while the time course of the currents provides a real-time view of the conformational changes of the channel molecule. Quantitative analysis of the currents can thus guide modeling of the gating mechanisms.

In practice, analysis of single channel currents can be complicated. Gating is typically modeled by a Markov process, in which each state of the model represents a physical conformation of the channel. However, biological channels often have multiple conformations with the same conductance, concealing transitions between states. Transitions among these conformations can only be deduced statistically from their lifetime distributions. In the literature of ion channel modeling, this is called an aggregated Markov process.

Traditionally, single channel currents are analyzed in two steps: noisy records are first idealized, and the resulting dwell-time distributions are fitted by a model. Idealization is typically done through half-amplitude threshold detection. The idealized dwell-times can be analyzed in different ways. The simplest approach is to fit the histograms of the dwell-times at each conductance level (Colquhoun and Sigworth, 1995; McManus et al., 1987; Blunck et al., 1998). For equilibrium processes these are predicted to be sums of exponentials with as many components as states of a particular conductance (Colquhoun and Hawkes, 1981, 1983). However, the one-dimensional histogram doesn't include the correlation information between adjacent events; thus it only applies to simple models. The problem can be improved by using two-dimensional histograms that contain all the information available in the equilibrium data (Fredkin et al., 1985; Magleby and Weiss, 1990a). Although enlarging the dimensions can lead to more complete use of available information, approaches based on histogram fitting are not applicable to nonstationary data, such as those from voltage-gated channels where the channel activity is not constant in time. To resolve the limitations of histogram fitting, Horn and Lange (1983) introduced the maximum likelihood approach, in which the model is fitted directly to the observed dwell-time sequence such that its probability is maximized. The approach is computationally more demanding, but more efficient in the use of available information and minimizing the variance of the optimal estimates. Ball and Sansom (1988) extended the approach to the models with substates and considered simplifications of the likelihood evaluation. A more complete treatment of the approach was recently provided by Qin et al. (1996, 1997). The new algorithm provides an explicit correction for missed events and uses a variable metric optimizer with analytically calculated derivatives for efficiently searching the likelihood space.

The idealization-based approach works reasonably well with many different kinds of channels. Although brief events may be missed during idealization, the errors can largely be corrected. The approach fails, however, when idealization is limited by poor signal-to-noise ratio. When the channel spends little time in stable states, a condition commonly referred to as "buzzed mode", there may be other difficulties. In our earlier paper (Qin et al. 1996) we used a first-order missed events correction given by Roux and Sauve (1985), which applies to all kinetics except buzz

Received for publication 18 January 2000 and in final form 14 June 2000.
Address reprint requests to Dr. Feng Qin, Dept. of Biophysical Sciences, SUNY at Buffalo, 124 Sherman Hall, Buffalo, NY 14214. Tel.: 716-829-3289; Fax: 716-829-2028; E-mail: qin@acsu.buffalo.edu.

0006-3495/00/10/1915/13 $2.00

mode. While incomplete, the solution allows us to use the analytical derivatives of the likelihood function to improve speed and stability. Hawkes et al. (1990) have published an exact missed events solution for short times that is not subject to the kinetic limitations of buzz mode. However, for this solution it is difficult to calculate the analytical derivatives for use by the optimizer. In practice, the buzz mode dwell-time analysis suffers from a more fundamental problem than missed events. The presence of many event durations comparable to the filter's time response results in coupling amplitude and duration, so that idealization requires the additional assumption of binary conductances.

To address these extreme cases it is necessary to analyze the noisy data directly. One such approach is the simulation method proposed by Magleby and Weiss (1990b). Starting with an initial model, it simulates a set of single channel currents and then analyzes them in the same way as the experimental data. The histograms of the resulting dwell-times are constructed and compared to those from the experimental data. The parameters of the model are adjusted so that the fit between the two is optimal. Although the approach eliminates the necessity for a missed events correction, it still relies on a proper idealization of the data. It is also computationally intensive and inefficient.

A more efficient approach is hidden Markov modeling (HMM) (Rabiner, 1989). The approach is also based on the use of the maximum likelihood criterion, but it models both the signal and noise simultaneously. The underlying signal is assumed to be a discrete Markov chain and the noise is assumed to be white and Gaussian. It uses the joint probability of the sequence of the discrete observation samples as the likelihood function. The parameters of the model, including those for both the signal and noise, are adjusted to maximize the likelihood values. The general theory of HMM was established by Baum et al. in the 1960s (Baum et al., 1970). Since then it has been successfully applied to a variety of fields ranging from speech recognition to gene location. The effectiveness of the technique for ion channel current analysis was demonstrated by Chung et al. (1990, 1991). They showed that reliable estimates of model parameters could be achieved at a signal-to-noise ratio too low to permit idealization. Recently the technique has been extended to the more general case with correlated noise (Venkataramanan et al., 1998a, 1998b).

Standard HMM relies on Baum's reestimation procedure to optimize the likelihood function. When applied to single channel analysis, the algorithm has some major limitations. Channel kinetics are usually described by a continuous Markov process with rate constants as the parameters of interest, but Baum's algorithm estimates the discrete transition probabilities. As a consequence, the model topology generally cannot be constrained to utilize a priori knowledge because the estimation process assumes that the transitions between all pairs of states are allowed. Due to the lack of the explicit control of rate constants, it is also difficult with the algorithm to impose constraints such as detailed balance and to fit data sets at different experimental conditions simultaneously. In light of these problems, Fredkin and Rice (1992) considered the use of a general mathematical programming procedure to optimize the likelihood, but the algorithm is very computationally intensive.

In this paper we present a new approach for hidden Markov modeling of single channel currents. Instead of optimizing transition probabilities, the algorithm works directly on rate constants. As a result, it has the capability to account for specific model topology. It also has the advantage of allowing for constraints on model parameters and global fitting of multiple data sets across different experimental conditions. The traditional HMM procedures optimize transition probabilities and do not have such flexibilities. To improve the efficiency of optimization, we derive the analytical derivatives of the likelihood function and use a gradient-based variable metric method for searching the likelihood surface. Finally, we show by examples the efficiency and accuracy of the algorithm and compare it with standard Baum reestimation procedure and traditional dwell-time analysis.

THE MODEL

Channel kinetics are modeled as a time-homogeneous Markov process with a discrete number of states. Each state represents an energetically stable conformation of the channel. Consider a channel with N states. Transitions among the states are governed by first-order rate constants, which are collectively designated by a matrix $\mathbf{Q} = [k_{ij}]_{N\times N}$ whose (i, j)th off-diagonal element is the rate from state i to state j and whose diagonal elements have values such that the sum of each row equals zero. At any time t the state transition probabilities are determined by the Kolmogorov equation

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{Q} \tag{1}$$

where $\mathbf{P}(t) = [p_{ij}(t)]$ is an N-by-N matrix whose (i, j)th element defines the probability being in state j at time t given that it was in state i at time 0 (Cox and Miller, 1965).

Changes in the channel conformations are not observable directly; instead, only the current conducted by each state is measured. There may be multiple distinct states with the same conductance, a phenomenon known as "aggregation." Assume a channel has M conductances whose current amplitudes are represented by $I_1 \ldots I_M$, respectively. The state space of the channel can be accordingly partitioned into M classes. Transitions between the states within the same class are not visible even in the absence of noise; instead, they can only be inferred statistically. This can give rise to inherent difficulty for the general identifiability of such models (Kienker, 1989).

While the kinetics are modeled as a continuous Markov process, the observations consist of discrete samples. A sampled Markov process is known as a Markov chain. Instead of using the infinitesimal rate constant matrix **Q,** the transitions of a Markov chain are often characterized by a discrete transition probability matrix $\mathbf{A} = [a_{ij}]$, where a_{ij} represents the probability of the channel being in state j at the next sampling clock given that it was in state i at the previous sampling clock. Given a rate constant matrix **Q** and a sampling duration Δt, the transition probability matrix **A** can be obtained from the Kolmogorov equation as

$$\mathbf{A} = \exp(\mathbf{Q}\Delta t) \tag{2}$$

where the matrix exponential is defined in the usual manner. It is interesting to note that although a given rate constant matrix **Q** always results in a meaningful transition probability matrix **A,** the reverse may not be true. For example, a 2-by-2 transition probability matrix with $a_{11} = a_{22} = 0.1$ and $a_{12} = a_{21} = 0.9$ doesn't even satisfy the positive definite condition that a matrix exponential should satisfy. In other words, certain Markov chains do not have a physically meaningful interpretation as a sample of a continuous Markov process.

The HMM approach allows the channel properties, including both kinetics and current amplitudes, to be extracted directly from the noisy recordings. The data are modeled using a hidden Markov model, i.e., it is considered as a Markov process embedded in noise. In this paper we assume the noise is white and follows the Gaussian distribution (non-white noise is considered in a following paper). The noise is allowed to have a different standard deviation at each different conductance. Denote by σ_i, $i = 1 \ldots M$, the standard deviation of the noise associated to the ith conductance level. The estimation criterion used by HMM is maximum likelihood. That is, we wish to choose the model parameters including the rate constants (k_{ij}), the current amplitudes (I_i), and the noise standard deviations (σ_i) to maximize the probability of the observed data samples $Y = Y_1 \ldots Y_T$, given these parameters. Such a criterion is appealing because it is guaranteed to be asymptotically unbiased and have a minimum variance as compared to other estimators (Cox and Hinkley, 1965). The likelihood itself also provides a natural measure for the goodness-of-fit, which can be used to identify appropriate model topology. In the subsequent sections we discuss how to evaluate and optimize the likelihood function.

THE LIKELIHOOD FUNCTION

Maximum likelihood estimation is appealing, but often technically demanding; for many problems its formulation can be mathematically intractable. Fortunately, for the HMM problems there exist efficient algorithms for evaluating the likelihood. In the following, we give a brief description of these algorithms. For a detailed description see, for example, Rabiner (1989).

The basic idea for the evaluation of the likelihood involves three steps: enumerating all possible state sequences, calculating the probability of the observed data given each state sequence, and averaging the results by weighting them according to the probability of the state sequences. This, however, only serves as a theoretical guide. Direct implementation of this approach is usually computationally infeasible in practice because there are an astronomical number of state sequences even with a small data set, and the computational complexity increases exponentially with the length of the observations.

A more efficient approach is the so-called forward-backward procedure. Let $\alpha_t(i)$, $1 \le i \le N$, $1 \le t \le T$, and $\beta_t(j)$, $1 \le j \le N$, $1 \le t \le T$ denote the forward and backward variables, respectively, where $\alpha_t(i)$ is the joint probability of the partial observation sequence up to time t assuming the channel is in state i at time t, and $\beta_t(j)$ is the joint probability of the complement observations from the last sample back to time t assuming the channel is in state j at time t. Mathematically,

$$\alpha_t(i) = \Pr[Y_1 \ldots Y_t, s_t = i]$$

$$\beta_t(j) = \Pr[Y_{t+1} \ldots Y_T, s_t = j]$$

where s_t represents the underlying state sequence. By making use of Bayes law, the forward and backward variables can be formulated recursively as

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_{ij}(t+1), \qquad j = 1 \ldots N, \quad t = 1 \ldots t \tag{3}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} \beta_{t+1}(j) b_{ij}(t+1), \qquad i = 1 \ldots N, \quad t = T \ldots 1 \tag{4}$$

where a_{ij} is the transition probability from state i to j, and $b_{ij}(t)$ is the probability distribution of the observation samples at time t given the knowledge that the channel is in state j at time t and state i at time $t - 1$.

The fact that the observation distribution $b_{ij}(t)$ can depend on the previous state of the channel in addition to the current state makes it possible to exploit a correlated noise model in the framework of standard HMM without introducing additional complications (Qin et al., 1994). The capability of this approach, however, is limited, and only a first-order regression model can be accommodated. In this paper we are only concerned with white Gaussian noise, in which case

the observation distribution is simply a Gaussian function:

$$b_j(t) = \frac{1}{\sqrt{2\pi}\sigma_{\kappa(j)}} \exp\left[-\frac{(Y_t - I_{\kappa(j)})^2}{2\sigma^2_{\kappa(j)}}\right]$$

where $\kappa(j)$ represents the conductance mapping of a given state j. The subscript i in the previous notation $b_{ij}(t)$ is omitted because the distribution is dependent only on the current state.

Equations 3 and 4 imply that the forward and backward variables can be calculated in a recursive manner by making use of previous results. The initial conditions for the recursions are given by $\alpha_0(i) = \pi_i$ for the forward variables and $\beta_T(j) = 1$ for the backward variables, where π_i values are the starting probabilities of the channel. At every time t, it takes $2N^2$ operations to evaluate all N variables assuming that the Gaussian distributions are calculated ahead of time. For a data set with T samples, the total complexity of the forward-backward procedure is of the order $4N^2T$.

In practice, the implementation of the forward-backward procedure requires appropriate scaling to avoid numeric underflow. This can be done by normalizing the forward variables at each time t and then multiplying the backward variables with the same scaling factor. For the details of scaling, see Rabiner (1989).

From the forward and backward variables, the likelihood function can be formulated as

$$\Pr(\mathbf{Y}) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i). \tag{5}$$

The equality holds for every $t = 1 \ldots T$. In particular, taking $t = T$, we obtain the likelihood as the sum of the last forward variables over all states. In other words, the forward variables alone suffice for evaluating the likelihood. But as we will see later, the backward variables are needed for evaluating the derivatives of the likelihood function.

We have assumed that the transition probabilities a_{ij} and the initial probabilities π_i are known in the calculation of the forward and backward variables. These probabilities can be determined from the rate constants. The transition probability matrix $\mathbf{A}$ is related to $\mathbf{Q}$ by the matrix exponential in Eq. 2, which can be evaluated using spectral decomposition, i.e.,

$$\mathbf{A} = \exp(\lambda_1 \Delta t)\mathbf{A}_1 + \cdots + \exp(\lambda_N \Delta t)\mathbf{A}_N \tag{6}$$

where λ_i is the ith eigenvalue of $\mathbf{Q}$ and $\mathbf{A}_i$ consists of the product of the corresponding left and right eigenvectors, respectively (Qin et al., 1997). It should be noted that the spectral decomposition technique requires the matrix to be diagonable (Ben-Israel and Greville, 1974). Most $\mathbf{Q}$ matrices can satisfy the condition, given the assumption of microscopic reversibility (Fredkin et al., 1985). For the matrices that do not satisfy the condition, other matrix exponential computation techniques, such as series expansion, have to be applied to compute **A.**

The initial probabilities, if unknown, are determined as the equilibrium probabilities of the channel at the holding condition. Let $\mathbf{Q}_h$ be the rate constant matrix corresponding to the holding condition. Then the initial probabilities are obtained by solving

$$\pi^\tau \mathbf{Q}_h = 0 \tag{7}$$

where $\pi = [\pi_1 \ldots \pi_N]^\tau$. Equation 7 is homogeneous and can be solved using the singular value decomposition technique. Normally, a rate constant matrix has a rank of $N - 1$, which is equal to the number of independent variables in π because the probability variables are subject to the probability totality constraint. However, at certain holding conditions, $\mathbf{Q}_h$ may have a reduced rank because some rates in the model may vanish. In those cases, some of the states will not be reachable and we can introduce further constraints by forcing the initial probabilities of those states to be zero. Therefore, the equation continues to have a unique solution.

GRADIENTS OF THE LIKELIHOOD FUNCTION

For efficient optimization it is necessary to have the gradient information of the objective functions. Although there are optimization algorithms that require only function evaluations, such as the downhill simplex method (Nelder and Mead, 1965) and Powell's direction set method (Fletcher, 1980), these derivative-free approaches are usually not efficient and limited to applications only with few parameters. In this section we describe how to analytically calculate the derivatives of the likelihood function. A related discussion for general HMM can also be found in the literature (Levinson et al., 1983).

The calculation of the derivatives can be divided into two major steps. In the first step, we calculate the derivatives of the likelihood function with respect to the standard HMM variables, i.e., the initial probabilities, the transition probabilities, the current amplitudes, and the noise standard deviations. For this purpose, it is helpful to formulate the likelihood into a more compact form using matrix notation. Let α_t and β_t be the column vectors of the forward and backward variables at time t, respectively. Then the forward-backward recursion can be written in matrix form as

$$\alpha^\tau_{t+1} = \alpha^\tau_t \mathbf{A}\mathbf{B}_{t+1}$$

$$\beta_t = \mathbf{A}\mathbf{B}_{t+1}\beta_{t+1}$$

where $\mathbf{B}_t$ is a diagonal matrix of the observation distributions at time t,

$$\mathbf{B}_t = \begin{bmatrix} b_1(t) & & \\ & \ddots & \\ & & b_N(t) \end{bmatrix}.$$

The likelihood function, which is equal to $L = \alpha_T^\tau 1$, can be expressed explicitly in terms of the model parameters as

$$L = \pi^\tau \mathbf{B}_1 \cdot \mathbf{AB_2} \cdot \cdot \cdot \mathbf{AB}_T \cdot 1.$$

It is interesting to notice that the expression is in a similar form to that of the likelihood in the dwell-time maximum likelihood approach (Qin et al., 1997). They are both products of transition probabilities. Intuitively, the equation says that the likelihood is equal to the initial probability of entering the states, multiplied by the probability of observing the first sample, and then multiplied by the probability to make a transition to the next sample, followed by the probability of observing the second sample, and so on. The unit vector at the end serves to sum the probabilities over all possible states since there is no explicit knowledge about which state the channel goes to at the end of the observations.

By considering the likelihood as the product of matrices, we can derive its derivatives using the chain rule. Specifically, we take the derivatives of each term $\mathbf{AB}_t$ in the product while holding others as constants, and then sum the results together. For a variable x, this leads to

$$\frac{\partial L}{\partial x} = \frac{\partial \pi^\tau \mathbf{B_1}}{\partial x} \cdot \beta_1 + \sum_t a_t^\tau \frac{\partial \mathbf{AB}_{t+1}}{\partial x} \beta_{t+1}$$

where the following equalities were used in the derivation:

$$\alpha_t^\tau = \pi^\tau \mathbf{B}_1 \cdot \mathbf{AB}_2 \ldots \mathbf{AB}_t$$

$$\beta_t = \mathbf{AB}_{t+1} \ldots \mathbf{AB}_T \cdot 1.$$

Letting x be the initial probabilities, the transition probabilities, the current amplitudes, and the noise standard deviations, respectively, we can obtain the corresponding derivatives as

$$\frac{\partial L}{\partial \pi_i} = \beta_1(i) b_i(1) \tag{8}$$

$$\frac{\partial L}{\partial a_{ij}} = \sum_t \alpha_t(i) \beta_{t+1}(j) b_j(t+1) \tag{9}$$

$$\frac{\partial L}{\partial I_k} = \sum_t \sum_{\kappa(i)=k} \alpha_t(i) \beta_t(i) \sigma_k^{-2} (Y_t - I_k) \tag{10}$$

$$\frac{\partial L}{\partial \sigma_k^2} = \sum_t \sum_{\kappa(i)=k} \alpha_t(i) \beta_t(i) \left[-\frac{1}{2\sigma_k^2} + \frac{(Y_t - I_k)^2}{2\sigma_k^4} \right] \tag{11}$$

where $\kappa(i)$ has the same definition as before.

In the above derivation, we have worked on the likelihood itself. In practice, the log likelihood is computed. With the scaled forward and backward variables, the derivatives of the log likelihood can be formulated as

$$\frac{\partial \ln L}{\partial \pi_i} = \hat{\beta}_1(i) b_i(1) \tag{12}$$

$$\frac{\partial \ln L}{\partial a_{ij}} = \sum_t \hat{\alpha}_t(i) \hat{\beta}_{t+1}(j) b_j(t+1) \tag{13}$$

$$\frac{\partial \ln L}{\partial I_k} = \sum_t \sum_{\kappa(i)=k} \gamma_t(i) \sigma_k^{-2} (Y_t - I_k) \tag{14}$$

$$\frac{\partial \ln L}{\partial \sigma_k^2} = \sum_t \sum_{\kappa(i)=k} \gamma_t(i) \left[-\frac{1}{2\sigma_k^2} + \frac{(Y_t - I_k)^2}{2\sigma_k^4} \right] \tag{15}$$

where $\hat{\alpha}_t(i)$ and $\hat{\beta}_t(i)$ are the scaled forward and backward variables.

The second step in the calculation of the derivatives of the likelihood function is to calculate the derivatives of the transition probabilities and initial probabilities with respect to rate constants. This is essential for the capability to optimize the rate constants directly. The derivatives of the transition probabilities can be derived by expanding $\mathbf{A}$ into its Taylor series and then calculating the derivatives of each individual term in the series, i.e.,

$$\frac{\partial \mathbf{A}}{\partial x} = \sum_{n=1}^{\infty} \frac{\Delta t^n}{n!} \sum_{k=0}^{n-1} \mathbf{Q}^k \frac{\partial \mathbf{Q}}{\partial x} \mathbf{Q}^{n-k-1} \tag{16}$$

where x is any variable of interest. Although this formula itself can be used for the derivative evaluation, it can be further simplified if we have the spectral expansion of $\mathbf{Q}$. Recalling the spectral expansion of $\mathbf{Q}$ and the orthogonal conditions of $\mathbf{A}_i$, we have

$$\mathbf{Q}^k = \sum_{i=1}^{N} \lambda_i^k \mathbf{A}_i .$$

Substituting it into Eq. 16 we obtain

$$\frac{\partial \mathbf{A}}{\partial x} = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{A}_i \frac{\partial \mathbf{Q}}{\partial x} \mathbf{A}_j f(\lambda_i, \lambda_j, \Delta t) \tag{17}$$

where $f(\lambda_i, \lambda_j, \Delta t)$ is a scalar function defined by

$$f(\lambda_i, \lambda_j, \Delta t) = \sum_{n=1}^{\infty} \frac{\Delta t^n}{n!} \sum_{k=0}^{n-1} \lambda_i^k \lambda_j^{n-k-1}$$

and can be simplified into

$$f(\lambda_i, \lambda_j, \Delta t) = \begin{cases} \dfrac{g(\lambda_j) - g(\lambda_i)}{\lambda_j - \lambda_i} & \text{if } \lambda_j \neq \lambda_i \\ g'(\lambda_j) & \text{otherwise} \end{cases}$$

where $g(\lambda) = \exp(\lambda\Delta t)$. Equation 17 suggests that the infinite power series defining the derivatives of the transition probabilities in Eq. 16 can be evaluated explicitly. Therefore, no series expansion is needed, making not only the computation more efficient, but also the result more accurate.

To derive the derivatives of the initial probabilities, we differentiate Eq. 7, leading to

$$\frac{\partial \boldsymbol{\pi}^{\tau}}{\partial x}\mathbf{Q}_{\mathrm{h}} = \boldsymbol{\pi}^{\tau}\frac{\partial \mathbf{Q}_{\mathrm{h}}}{\partial x}. \tag{18}$$

This is a linear system equation similar to Eq. 7. They have the same coefficient matrix and therefore can be solved using the same technique. One difference to note is that the unknowns here are normalized to have a sum of zero instead of one.

Finally, the derivatives of the likelihood function with respect to the rate constants are obtained by combining the derivatives of the likelihood with respect to the transition probability, and initial probability with the derivatives of these probabilities to the rate constants. That is,

$$\frac{\partial \ln L}{\partial k_{\mathrm{ij}}} = \frac{\partial \ln L}{\partial \boldsymbol{\pi}}\cdot\frac{\partial \boldsymbol{\pi}}{\partial k_{\mathrm{ij}}} + \frac{\partial \ln L}{\partial \mathbf{A}}\cdot\frac{\partial \mathbf{A}}{\partial k_{\mathrm{ij}}} \tag{19}$$

where $\partial\boldsymbol{\pi}/\partial k_{\mathrm{ij}}$ and $\partial\mathbf{A}/\partial k_{\mathrm{ij}}$ are obtained from Eqs. 18 and 17, respectively. This equation, along with Eqs. 12–15, give the overall derivatives of the likelihood function with respect to the unknowns.

The overall computation of the derivatives of the likelihood function is dominated by the computation of the derivatives with respect to the transition probabilities, which takes on the order of $2N^2T$, as seen from Eq. 13. The computation of the derivatives with respect to the current amplitudes and the noise variances takes on the order of $2NT$, which is about one order of magnitude lower than those for the transition probabilities. The computation of the derivatives of the transition probability and initial probability with respect to the rate constants is usually negligible, because it only depends on the number of states, which is much less than the data length. Therefore, the overall computation of the derivatives is on the order of $2N^2T$. If we also include the calculations for the forward and backward variables, the total complexity will be on the order of $6N^2T$, which is about the same as for the standard Baum's reestimation.

MAXIMIZATION OF THE LIKELIHOOD FUNCTION

Before describing the optimization procedure, we need to decide what parameters to optimize. So far we have assumed that the rate constants, the current amplitudes, and the noise variances are the independent unknowns. In practice, however, we don't optimize the rate constants directly. Instead, we have followed the same approach as we did with the dwell-time maximum likelihood estimation (Qin et al., 1996). Specifically, we represent the rate constants by

$$q_{\mathrm{ij}} = C_{\mathrm{ij}}\exp(\mu_{\mathrm{ij}} + \nu_{\mathrm{ij}}V) \tag{20}$$

where μ_{ij} and ν_{ij} are the new variables, C_{ij} is the concentration of the ligand to which the rate q_{ij} is sensitive, and V is the voltage or force. For the rates that are not ligand or voltage-sensitive, we set $C_{\mathrm{ij}} = 1$ or enforce $\nu_{\mathrm{ij}} = 0$, respectively. The new parameters μ_{ij} and ν_{ij} are intrinsic to a channel and do not vary with the experimental conditions.

Parametrizing the transition rates with μ_{ij} and ν_{ij} rather than the rate constants offers several advantages. First, it allows us to combine the datasets obtained at different experimental conditions to be fit simultaneously. Second, it reduces the detailed balance constraints, which are nonlinear in the rate constants, to be linear in μ_{ij} and ν_{ij}. This is important from the computational point of view because linear constraints can be handled analytically, as shown below. Other constraints, such as holding rates at fixed values or scaling between two rates, remain linear. Therefore, all these constraints can be cast into a set of linear equations

$$\Gamma\mathbf{x} = \mathbf{g} \tag{21}$$

where Γ is the coefficient matrix, $\mathbf{g}$ is the constant vector, and $\mathbf{x}$ is the collection of the new variables μ_{ij} and ν_{ij}.

Optimization subject to linear constraints is equivalent to searching within the nullspace of the coefficient matrix of the constraint. By making use of matrix decompositions, we can express the constrained variables $\mathbf{x}$ explicitly in terms of a set of unconstrained independent ones, say

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{b} \tag{22}$$

where $\mathbf{z}$ can be considered as a nullspace vector. The matrix $\mathbf{A}$ and vector $\mathbf{b}$ can be readily determined from Γ and $\mathbf{g}$ in Eq. 21 using the standard matrix decomposition techniques such as SVD or QR (Press et al., 1992). By choosing to optimize $\mathbf{z}$, the originally constrained problem becomes an equivalent unconstrained one, which is easier to solve.

The derivatives of the likelihood with respect to the unconstrained variables $\mathbf{z}$ can be obtained by combining the derivatives of the log likelihood with respect to the rate constants $\partial\ln L/\partial q_{\mathrm{kl}}$ and the derivatives of the rate constants with respect to the unconstrained variables $\partial q_{\mathrm{kl}}/\partial z_{\mathrm{j}}$ using the chain rule. The derivatives $\partial q_{\mathrm{kl}}/\partial z_{\mathrm{j}}$ can be readily derived from Eqs. 20 and 22 as

$$\frac{\partial q_{\mathrm{kl}}}{\partial z_{\mathrm{j}}} = \frac{\partial q_{\mathrm{kl}}}{\partial x_{2\mathrm{i}-1}}\cdot\frac{\partial x_{2\mathrm{i}-1}}{\partial z_{\mathrm{j}}} + \frac{\partial q_{\mathrm{kl}}}{\partial x_{2\mathrm{i}}}\cdot\frac{\partial x_{2\mathrm{i}}}{\partial z_{\mathrm{j}}} = q_{\mathrm{kl}}(a_{2\mathrm{i}-1\mathrm{j}} + a_{2\mathrm{ij}}V) \tag{23}$$

where $\mu_{\mathrm{kl}} = x_{2\mathrm{i}-1}$ and $\nu_{\mathrm{kl}} = x_{2\mathrm{i}}$.

Maximization of the likelihood can be achieved with a variety of methods. We have used a quasi-Newton method based on the Davidon-Fletcher-Powell algorithm with an approximate line search (Press et al., 1992). The method is

appealing because it has a quadratic convergence near the optimum and gives an estimate of the curvature of the likelihood surface. A small modification was made to the method in our implementation. We found that the method was sometimes too aggressive in the sense that it often took a step size too large to cause an excessive number of function evaluations for backtracking. Because this can occur at each iteration and the likelihood function evaluation is usually costly, it significantly degraded the performance of the algorithm. We have taken a more conservative strategy by imposing a hard limit on the maximal step size. A limit that is approximately equal to the initial rate constants was found to work well in practice.

The quasi-Newton method provides an approximate inverse Hessian matrix of the likelihood surface at the maximum. It contains the curvature information, from which one can obtain the standard errors of the estimates. It is known that

$$\mathrm{Cov}(\mathbf{z}^*) = -\mathbf{H}^{-1}(\mathbf{z}^*)$$

where $\mathbf{z}^*$ is the optimum variable vector, Cov is the covariance matrix, and $\mathbf{H}$ is the Hessian matrix of second-order partial derivatives of the likelihood function. The diagonal elements of the covariance matrix define the variance of the transformed variable $\mathbf{z}$. The estimates for the variances of rate constants, denoted by $\mathrm{var}(q_{kl})$, can be calculated from $\mathrm{var}(z_i)$ using the following relation:

$$\mathrm{var}(q_{kl}) \approx \sum_i \left(\frac{\partial q_{kl}}{\partial z_i}\right)^2 \mathrm{var}(z_i) \qquad (24)$$

where the derivatives $\partial q_{kl}/\partial z_j$ are given by Eq. 23. Although there is no rigorous proof in theory, experiments have shown that the error estimates obtained in this way provide a fairly good approximation to those calculated from the exact Hessian matrix (Salamone et al., 1999). The exact Hessian matrix can be calculated numerically because the first-order derivatives of the likelihood function are known analytically.

EXPERIMENTAL RESULTS

The algorithm described in the previous sections has been implemented and runs on both PC and Unix workstations. The spectral expansion for the calculation of matrix exponentials was performed using the EISPACK routines, which was downloaded from the netlib repository (http://www.netlib.org). The linear homogeneous equations for the starting probabilities and their derivatives were solved using the SVD routine in Numerical Recipes (Press et al., 1992). The same routine was also used for the decomposition of constraints on model parameters. The optimization of the likelihood was done using the variable metric method in Numerical Recipes. The procedure was modified slightly by using a restricted maximal step size as described above. The simulation of the single channel currents was done by first generating exponentially distributed dwell-times followed by discrete sampling at given intervals.

The program has a variety of features such as permitting constraints on rate constants, allowing global fitting of multiple data sets across different experimental conditions, optimizing the initial probability implicitly based on preconditioning pulses, and so on. We will not examine these features individually here since their uses are straightforward. Instead, we focus mainly on the basic performance of the algorithm.

Our initial testing concerns the limits of the HMM approach. As mentioned previously, the HMM approach is mainly intended for some extreme cases where the more efficient dwell-time approach doesn't work reliably. One limit is when the kinetics of the channel are too rapid and where the missed events cannot be accurately corrected. The other is when the signal-to-noise ratio is too small for currents to be idealized. In this example, we attempt to draw some insights about the limits that the HMM approach can reach in these two cases. For simplicity, a two-state model was used. The currents were chosen to be 0 pA for the closed state and 1 pA for the open state. The two rates in the model were set to be equal to create a "buzz mode" activity, and their values were fixed at 100,000 s^{-1}. The noise standard deviation and the sampling duration were subject to testing.

We consider the kinetics first. To minimize the effect of noise, we used only a small amount of noise with a standard deviation of 0.1 pA. The length of the data was chosen such that there were ~500 dwell-times in the discrete samples that were at least twice as long as the sampling duration. A set of six sampling intervals was tested, corresponding to $k\Delta t = 0.3$, 0.5, 2, 3, and 4, respectively. This led to approximately 1600, 2700, 3700, 14,000, 70,000, 360,000 discrete samples containing a total number of 503, 1327, 3684, 27,945, 210,870, and 1,440,095 dwell-times in each case. As $k\Delta t \geq 2$, the data contained more samples than dwell-times. This happened because the kinetics was too fast compared to the sampling rate, so that one sampling interval may contain multiple transitions. With a starting value of 1000 s^{-1} for the rate constants, we found that the algorithm could work up to $k\Delta t = 4$, i.e., when the rates were four times as fast as the sampling rate. However, when the starting value was changed to be larger than the true values (200,000 s^{-1}), we obtained reliable estimates only when $k\Delta t \leq 2$.

Fig. 1 shows the cross sections of the likelihood surface cut parallel to the direction where the two rates are equal. It is seen that the function has a relatively well-defined curvature for $k\Delta t < 1$. When $k\Delta t$ increases, the upper half of the likelihood function becomes increasingly flat. As $k\Delta t$ increases up to 2, the surface becomes almost flat visually, yet the curvature is still large enough for the algorithm to work. But when $k\Delta t$ increases further, even the algorithm

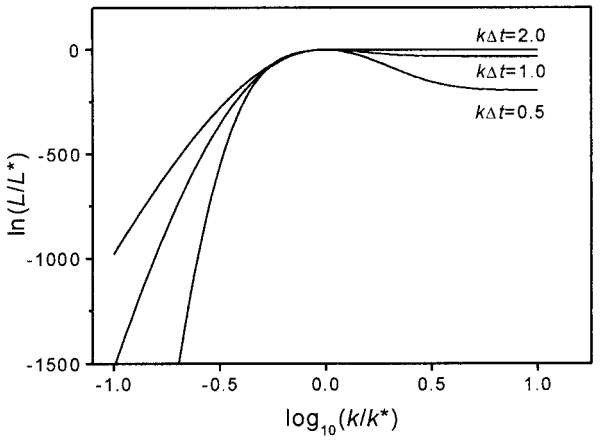

FIGURE 1 Cross sections of the log likelihood surface at different sampling intervals. k^* denotes the true value of the rate constants and L^* is the optimal likelihood value at $k = k^*$. The sections are cut along the line where the two rates are equal. Three sections are shown corresponding to $k\Delta t = 0.5$, 1.0, and 2.0. As the sampling interval increases, the likelihood surface becomes increasingly asymmetric, where the lower half ($k \leq k^*$) still maintains a good curvature, but the upper half ($k \geq k^*$) becomes flat. The maxima of the likelihood can be reliably identified until $k\Delta t = 2$.

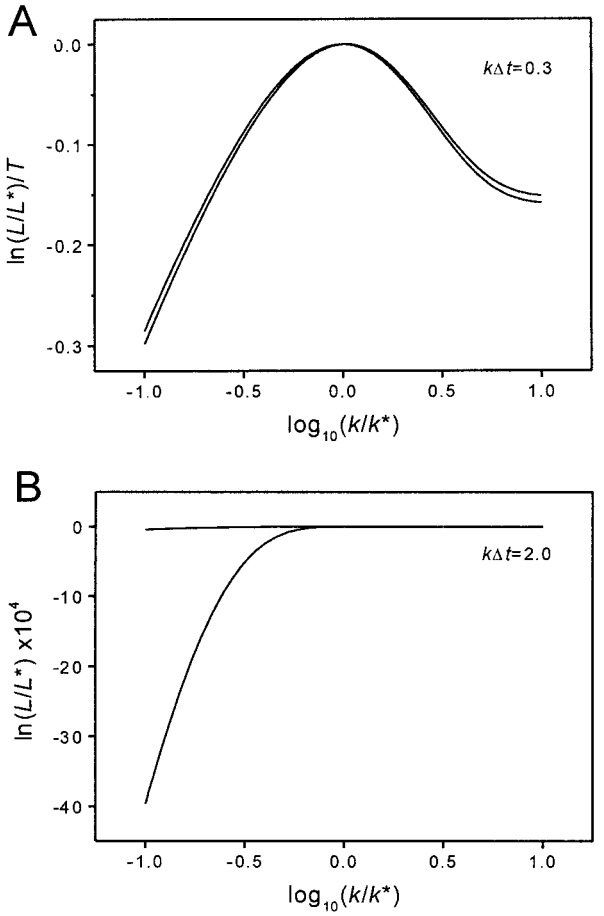

FIGURE 2 Cross sections of the log likelihood surface cut along the direction where the two rates are equal. (*A*) Three log likelihood functions at $k\Delta t = 0.3$ with different data lengths are shown. The log likelihood functions were normalized by the data length, which increased successively by a factor of 10. The three normalized log likelihood functions almost exactly overlap, suggesting that increasing the data length can improve the curvature of the likelihood surface proportionally. (*B*) Two log likelihood functions at $k\Delta t = 2$ with the data length different by a factor of 100 are displayed. Increasing the data length in this case improves the curvature of the lower half ($k \leq k^*$) of log likelihood function, but has little effect on the flatness of the upper half ($k \geq k^*$).

cannot see any curvature. The derivative is nearly zero at every point as long as the rates are greater than their true values. It is also seen from the figure that the likelihood function exhibits a skewed asymmetry, which explains why the results become unreliable if the starting value was larger than the true one. The algorithm can tell whether a rate is too slow to fit the data, but cannot discriminate very fast rates because they are all equally likely. At $k\Delta t = 2$, the algorithm resulted in $k_{12} = 124348 \pm 31428$ and $k_{21} = 127715 \pm 32305$, staring from 200,000 for both rates. The results remained about the same when a low starting value (1000) was used.

We also checked whether increasing the data length can resolve the ambiguity. Normally, the curvature of the likelihood function increases with the number of independent observations. Fig. 2 *A* shows the cross sections of the log likelihood functions for three different datasets. The data length increases successively by a factor of 10, and the sampling duration corresponds to $k\Delta t = 0.3$. The log likelihood functions were normalized by the data lengths. We see that the three log likelihood functions, after normalization, almost exactly overlap, suggesting that the curvature of the log likelihood at the maxima is proportional to the data length. This is in a good agreement with the theory that the second-order partial derivatives of the log likelihood function give the asymptotic standard errors for the parameter estimates, which are inversely proportional to the square root of the number of samples (Cox and Hinkley, 1965). As $k\Delta t$ increases, however, the effect of the data length on the flatness of the likelihood function becomes diminished. Fig. 2 *B* shows the cross sections of the log likelihood functions at $k\Delta t = 2$ for two different data lengths, with one 100 times the other. It is seen that the curvature of the upper half has little change, though the lower half increases dramatically.

The fact that increasing the data length has little effect on the flatness of the likelihood function with large $k\Delta t$ suggests that the algorithm has an intrinsic limit with respect to the kinetics. Theoretically, there should be no limit because even when the transition rates are faster than the sampling rate, there are still a number of events in the data that are longer than the sampling interval. Given a sufficiently large data set, one should be able to sample enough events to extrapolate the distribution of the dwell-times, since it is only a single exponential. Therefore, the limit is likely due to numerical errors. One possible explanation is that there are too few long events in the data. For a single exponential

distribution, the probability that a dwell-time is at least twice the sampling duration is equal to $\exp(-2k\Delta t)$. With the kinetics at $k\Delta t = 2$, this is ~0.018. That is, only 1.8% of the dwell-times are captured as true events. In other words, we only see a very small tail of the exponential distribution, which can be extrapolated in many different ways given the numerical errors. Increasing the data length simply repeats this same pattern. The absolute number of events in the data only affects the confidence of the results, but not the identifiability of the model. To have a unique fit, a larger portion of the exponential distribution must be observed, i.e., a smaller sampling duration has to be used.

To explore the noise limit, we used a fixed sampling duration at $\Delta t = 5$ μs, corresponding to $k\Delta t = 0.5$. The rates are fixed at their true values, so that the noise effect can be examined independently. Fig. 3 *A* illustrates the cross sections of the log likelihood surface along the direction in which the two rates are equal. For a fixed data length, the likelihood surface becomes flatter as the data get noisier. With 5000 samples, the algorithm reaches the limit at about $\sigma = 2$ pA, which is twice as large as the channel current. Such a signal-to-noise ratio would be beyond the limit that the dwell-time approach can work. If the half-amplitude threshold detection is used to idealize the data, the noise needs to be no more than a quarter of the current amplitude, which requires filtering at least at $f_c = 1/64$ kHz if a Gaussian filter is used. The rise time of such a filter is $\sim T_r = 0.3396/f_c = 22$ ms, which gives a dead time of ~11 ms. Given the channel kinetics at 100,000 s^{-1}, the chance to have an event to be detected is e^{-1000}. In other words, all events will be wiped out in the output. Therefore, examples like this have to be analyzed directly, without idealization.

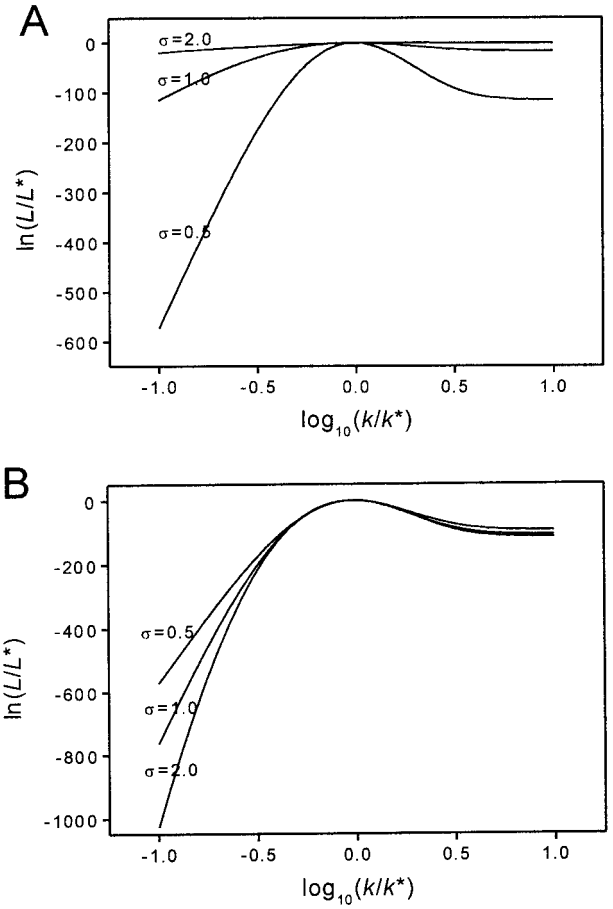

FIGURE 3 Effects of noise on the log likelihood surface around the maxima. (*A*) For a fixed data length, the likelihood surface becomes flatter as the data get noisier. The algorithm reaches the limit at a standard deviation about 2 pA, corresponding to a signal-to-noise ratio 0.5. (*B*) When the noise increases, the data length is also increased so that the log likelihood function maintains about the same curvature at the maxima as in the low noise case. The number of samples is 5000, 32,000, and 300,000 for $\sigma = 0.5$, 1, and 2 pA, respectively.

Although the kinetics have a hard limit that cannot be improved by increasing the data length, the noise doesn't seem to have such a cutoff. It appears that an increase in the noise standard deviation can always be compensated by enlarging the data set. In the above example, when we increased the data length from 5000 samples to 32,000 samples for $\sigma = 1$ pA and 300,000 samples for $\sigma = 2$ pA, the resultant log likelihood functions have approximately equal curvature at the maxima, as illustrated in Fig. 3 *B*. Therefore, the three cases, though with very different amounts of noise, have basically the same level of technical difficulty to solve given an unlimited computational resource. Intuitively, this could be understood in a way similar to ensemble averaging, where the same effective noise level can be achieved either with a short data set at low noise or a large data set at high noise. The difference here is that the HMM approach is based on the calculation of the expected means of the samples rather than the simple arithmetic averaging.

Next, we compare the optimization-based approach with the standard Baum's reestimation for HMM. A two-state model with $k_{co} = 10{,}000$ s^{-1} and $k_{oc} = 100{,}000$ s^{-1} was used. The data were sampled at $\Delta t = 10$ μs, and the current amplitudes were fixed at their true values, which were 0 pA for the closed channel and 1 pA for the open channel, respectively. Baum's algorithm does not estimate the rate constants directly. Neither has it the ability to constrain the model. For the sake of comparison, we simply obtain the rate constants by taking the logarithm of the transition probability matrix at the end of the estimation and then retain the rates for the transitions that are specified in the model.

The comparison was done at two different noise levels, one at $\sigma = 0.5$ pA and the other at $\sigma = 1.5$ pA. For $\sigma = 0.5$ pA, Baum's reestimation took about 20 iterations to converge, while the gradient approach took 17 iterations and 34 likelihood function evaluations, starting from $k_{co} = 100$ s^{-1} and $k_{oc} = 1000$ s^{-1} for both algorithms. When changing the starting values to $k_{co} = 100$ s^{-1} and $k_{oc} = 1{,}000{,}000$ s^{-1}, Baum's algorithm converged in 21 iterations, while the

gradient approach took 14 iterations and 25 function evaluations. In both cases, Baum's algorithm slightly outperforms the optimization method, which is found to be typically true with low noise data. At high noise levels, however, the optimization approach becomes much better. For example, with $\sigma = 1.5$ pA, the optimization approach took about 16 iterations and 23 function evaluations to converge starting from $k_{co} = 100\ s^{-1}$ and $k_{oc} = 1000\ s^{-1}$. The resulting estimates were $k_{co} = 10452 \pm 1359\ s^{-1}$ and $k_{oc} = 111091 \pm 12249\ s^{-1}$, with a maximal log likelihood value -184343.19. Baum's algorithm, however, took as many as 200 iterations to get comparable results: $k_{co} = 9711\ s^{-1}$ and $k_{oc} = 104536\ s^{-1}$ with the maximal likelihood value -184343.34. Fig. 4 shows the convergence trajectories of the two algorithms at both noise levels. The two algorithms proceeded basically in the same direction toward the minimum. The steps taken by Baum's algorithm were relatively smooth, while the gradient approach often made large turns. The major difference between the two algorithms is around the maximum, where the gradient approach converged rapidly, while Baum's algorithm moved very slowly by taking small steps at high noise level, presumably because the likelihood surface was flat.

The influence of noise and lack of true model constraints are not the only factors that limit the performance of Baum's reestimation algorithm. Another common problem is the aggregation of states of identical conductance, which often makes the convergence of Baum's algorithm very slow. As the last example, we check the influence of state aggregation on the optimization method and compare the convergence performance of the two algorithms. We consider the following nicotinic acetylcholine receptor (nAChR) channel model:

$$C_1 \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} C_2 \underset{k_{32}}{\overset{k_{23}}{\rightleftharpoons}} C_3 \underset{k_{43}}{\overset{k_{34}}{\rightleftharpoons}} O$$

The channel has three aggregated closed states. The rate constants were chosen to be $k_{12} = 200$, $k_{21} = 500$, $k_{23} = 400$, $k_{32} = 25{,}000$, $k_{34} = 60{,}000$, and $k_{43} = 240\ s^{-1}$. The two rates from C_3 are relatively fast, making its lifetime short. Thus the data appear with relatively long closures and openings mixed with brief closures occurring as bursts of activity. Two different noise levels were tested, one with $\sigma = 0.25$ pA and the other with $\sigma = 0.5$ pA. The channel currents were fixed at 0 pA for the closed states and 1 pA for the open state. The data were simulated with a sampling duration of 10 μs, and a total of 500,000 samples were generated for $\sigma = 0.25$ pA and 1,000,000 samples for $\sigma = 0.5$ pA. The initial values were set to $100\ s^{-1}$ for the slow rates k_{12}, k_{21}, k_{23}, and k_{43}, and $1000\ s^{-1}$ for the other two fast rates k_{32} and k_{34}.

Figure 5 depicts the time course of the convergence of the two algorithms in terms of the likelihood values and the root-mean-square errors of the rate constants. Even at the

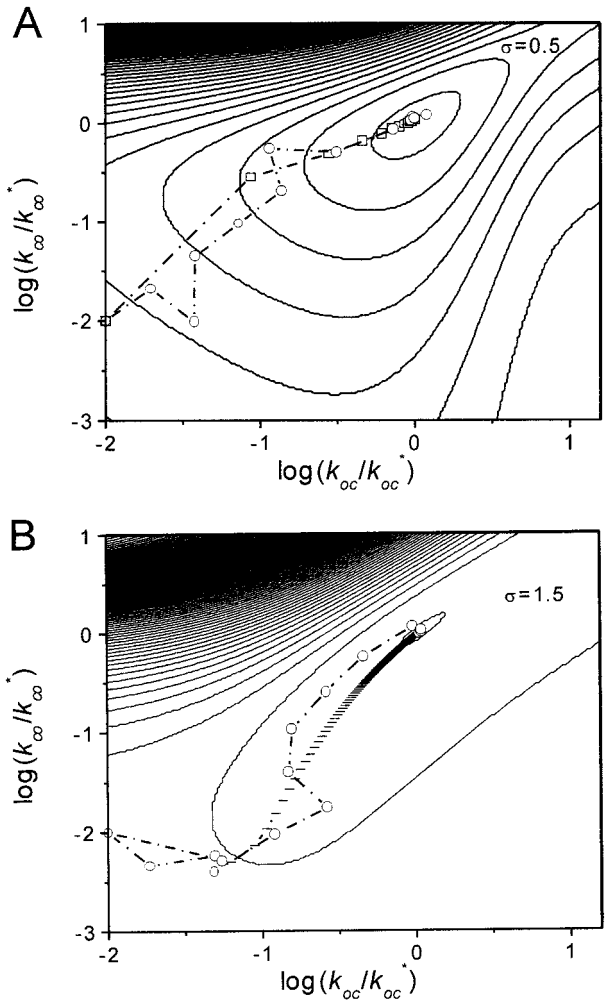

FIGURE 4 Comparison of the convergence behavior between Baum's reestimation and the optimization method at two different noise levels with (*A*) $\sigma = 0.5$ pA and (*B*) $\sigma = 1.5$ pA. The two approaches have a comparable convergence performance at the low noise level, in which case the log likelihood surface is relatively well-defined. When the noise is high, Baum's algorithm moves very slowly near the maxima, while the optimization approach maintains a quadratic convergence independent of the poor curvature. The circles represent the optimization approach, and the squares in (*A*) and the bars in (*B*) correspond to Baum's reestimation. Each symbol represents one iteration. The contours are ~330 natural log units apart in (*A*) and 200 natural log units in (*B*).

low noise level $\sigma = 0.25$ pA, Baum's algorithm exhibited a relatively slow convergence. It converged rapidly within the first 20 iterations, during which the likelihood increased to about five natural log units away from the ultimate maximum, but it then converged slowly and took about another 80 iterations to eventually reach the maximum. Such a biphasic convergence behavior was also observed in other examples and seems to be characteristic to Baum's reestimation. Because the noise in this example was low, the slow convergence is likely due to the aggregation of the multiple closed states. When the noise was increased to $\sigma = 0.5$ pA, the convergence became further degraded, as shown in Fig.

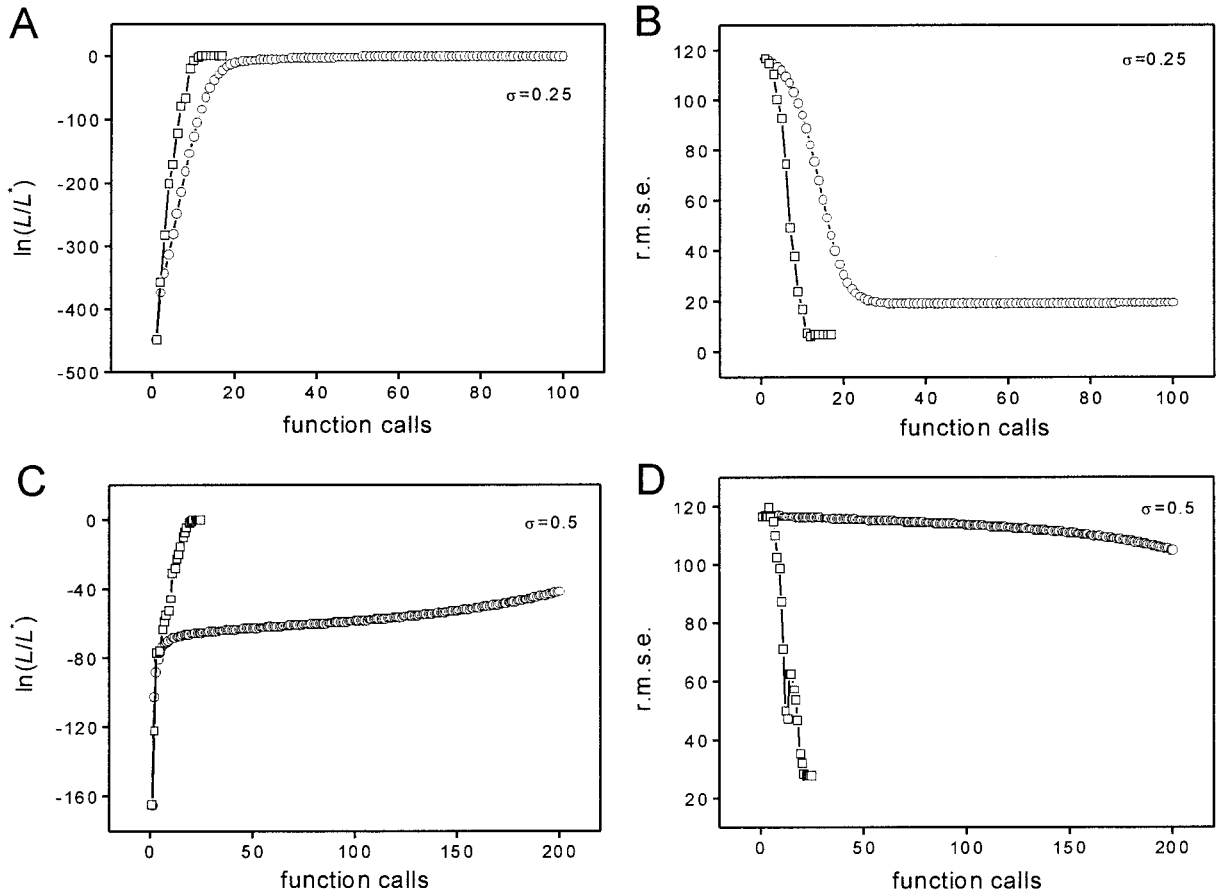

FIGURE 5 Evolutions of the log likelihood values and the root-mean-square errors of the rate constant estimates. The squares correspond to Baum's reestimation and the circles correspond to the optimization method. Two noise levels are shown, one at $\sigma = 0.25$ pA and the other at $\sigma = 0.5$ pA. The starting rates were 100 s^{-1} for all rates except k_{32} and k_{34}, which were set to 1000 s^{-1}.

5 *C*. The algorithm entered the slow convergence stage when the likelihood was only about halfway from the maximum, and it didn't reach the maximum within 200 iterations. The final rates also remained far from their true values, as seen in Fig. 5 *D*. The optimization approach, however, worked well in both cases and showed little dependence on either the noise level or the aggregation of states. The convergence near the maximum was all approximately quadratic. The final estimates of the rate constants obtained by the two algorithms are listed in Table 1 for the low-noise case and Table 2 for the high-noise case.

DISCUSSION

We have presented an optimization approach for hidden Markov modeling of single channel kinetics. The algorithm optimizes the rate constants directly and uses analytically calculated derivatives to search the likelihood surface. It allows the imposition of constraints on the model and can simultaneously fit multiple datasets obtained with different experimental conditions. It also has the feature to allow the initial probabilities to be specified by holding conditions, which is often necessary for the studies of rapidly inactivating voltage channels. The traditional Baum's reestimation algorithm lacks these features.

We compared the performance of the algorithm with that of Baum's reestimation. In addition to the advantage of optimizing rate constants rather than transition probabilities directly, the optimization approach also shows a better convergence performance when the likelihood surface is

TABLE 1 Parameter estimates for the ACh receptor channel with $\sigma = 0.25$ pA

	True Value	Initial Value	Optimization	Baum
k_{12}	200	100	155 ± 33	152
k_{21}	500	100	569 ± 204	531
k_{23}	400	100	500 ± 111	768
k_{32}	25,000	1000	23,184 ± 2165	29,336
k_{34}	60,000	1000	57569 ± 5157	51,192
k_{43}	240	100	222 ± 14	249
LL		−19,741.57	−19,292.81	−19,292.83

relatively flat because of low signal-to-noise ratio or state aggregation. In these cases, Baum's algorithm exhibits a biphasic convergence behavior. The likelihood increases rapidly at the beginning, but convergence becomes very slow as it approaches the maximum. It is interesting to notice that similar behaviors are also observed with other maximum likelihood problems and it appears to be a common symptom of the expectation-maximization (EM) approach (Titterington et al., 1985). The gradient-based optimization approach, however, shows little dependence on the conditioning of the likelihood surface.

Baum's reestimation algorithm, however, outperforms the optimization approach when the likelihood surface is well-defined. Its convergence is more stable and smooth, making it more suitable for extracting the amplitude information. For those parameters, the likelihood function usually has a large curvature, because the information is contained in each sample and the total information content is abundant. In the future, we plan to integrate the two methods, as they appear to be complementary; Baum's reestimation has a good convergence at the initial stage, while the gradient approach is more efficient near the maximum.

The HMM algorithm is likely to be more useful in the extreme limiting cases where the more efficient dwell-time approach cannot work reliably, when the channel currents are too small to be idealized, or the kinetics is too rapid for the first-order correction for missed events. With simulations, we showed that the algorithm could work up to the point where the kinetics is about as fast as the sampling rate. Beyond that, the likelihood surface became too flat, making the estimates unreliable. Unlike the kinetics, the effect of noise doesn't appear to have a hard limit, and a high noise could be compensated by an increase in the number of samples.

We have assumed that the background noise is white and the data are unfiltered. This assumption is rarely satisfied in practice. The recording system always has a finite bandwidth that introduces memory to the signal. The noise in patch-clamp recording arises from a variety of sources. The instrumental noise has a power spectrum increasing quadratically with frequency due to input capacitance, and the noise from other sources, such as channel openings, may have unknown characteristics. It has been shown that the noise correlation can give rise to significant biases on parameter estimates. In the following paper we describe how to extend the current approach to allow for filtering and correlated noise.

TABLE 2 Parameter estimates for the ACh receptor channel with σ = 0.5 pA

	True Value	Initial Value	Optimization	Baum
k_{12}	200	100	240 ± 47	175
k_{21}	500	100	841 ± 323	141
k_{23}	400	100	539 ± 124	416
k_{32}	25,000	1000	24,286 ± 2197	7663
k_{34}	60,000	1000	44,014 ± 9501	3290
k_{43}	240	100	203 ± 28	90
LL		−731,387.72	−731,222.78	−731,264.41

The program is publicly available and can be downloaded from our website at http://www.qub.buffalo.edu.

This work was supported by National Institutes of Health Grant RR-11114 and W. M. Keck Foundation.

REFERENCES

Ball, F. G., and M. S. P. Sansom. 1988. Aggregated Markov processes incorporating time interval omission. *Adv. Appl. Prob.* 20:546–572.

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

Ben-Israel, A., and T. Greville. 1974. Generalized Inverses: Theory and Applications. John Wiley & Sons, New York.

Blunck, R., U. Kirst, T. Riessner, and U. Hansen. 1998. How powerful is the dwell-time analysis of multichannel records? *J. Membr. Biol.* 165: 19–35.

Chung, S. H., V. Krishnamurthy, and J. B. Moore. 1991. Adaptive processing techniques based on hidden Markov models for characterizing very small channel currents buried in noise and deterministic interferences. *Proc. R. Soc. Lond. B.* 334:357–384.

Chung, S. H., J. B. Moore, L. G. Xia, L. S. Premkumar, and P. W. Gage. 1990. Characterization of single channel currents using digital signal processing techniques based on hidden Markov models. *Proc. R. Soc. Lond. B.* 329:265–285.

Colquhoun, D., and A. G. Hawkes. 1981. On the stochastic properties of single ion channels. *Proc. R. Soc. Lond. B.* 211:205–235.

Colquhoun, D., and A. G. Hawkes. 1983. The principles of the stochastic interpretation of ion-channel mechanisms. *In* Single-Channel Recording. B. Sakmann and E. Neher, editors. Plenum Publishing Corp., New York. 135–175.

Colquhoun, D., and F. J. Sigworth. 1995. Fitting and statistical analysis of single channel records. *In* Single-Channel Recording. B. Sakmann and E. Neher, editors. Plenum Publishing Corp., New York. 483–585.

Cox, D. R., and D. V. Hinkley. 1965. Theoretical Statistics. Chapman and Hall, London.

Cox, D. R., and H. D. Miller. 1965. The Theory of Stochastic Processes. Methuen, London.

Fletcher, R. 1980. Practical Methods of Optimization. John Wiley & Sons, Chichester.

Fredkin, D. R., M. Montal, and J. A. Rice. 1985. Identification of aggregated Markovian models: application to the nicotinic acetylcholine receptor. Proceedings of the Berkeley Conference in Honor of Jerzy Neymann and Jack Kiefer. Belmont, CA.

Fredkin, D. R., and J. A. Rice. 1992. Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. R. Soc. Lond. B.* 239:125–132.

Hawkes, A. G., A. Jalali, and D. Colquhoun. 1990. The distribution of the apparent open times and shut times in a single channel record when brief events cannot be detected. *Phil. Trans. R. Soc. Lond. A.* 332:511–538.

Horn, R., and K. Lange. 1983. Estimating kinetic constants from single channel data. *Biophys. J.* 43:207–223.

Kienker, P. 1989. Equivalence of aggregated Markov models of ion-channel gating. *Proc. R. Soc. Lond. B.* 236:269–309.

Levinson, S. E., L. R. Rabiner, and M. M. Sondhi. 1983. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Syst. Tech. J.* 62: 1035–1074.

Magleby, K. L., and D. S. Weiss. 1990a. Identifying kinetic gating mechanisms for ion channels by using two-dimensional distributions of simulated dwell times. *Proc. R. Soc. Lond. B. Biol. Sci.* 241:220–228.

Magleby, K. L., and D. S. Weiss. 1990b. Estimating kinetic parameters for single channels with simulation: a general method that resolves the missed event problem and accounts for noise. *Biophys. J.* 58:1411–1426.

McManus, O. B., A. L. Blatz, and K. L. Magleby. 1987. Sampling, log binning, fitting, and plotting durations of open and shut intervals from single channels and the effects of noise. *Pflugers Arch.* 410:530–553.

Nelder, J. A., and R. Mead. 1965. A simplex method for function minimization. *Comput. J.* 7:308–313.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. Numerical recipes in C. Cambridge University Press, Cambridge.

Qin, F., A. Auerbach, and F. Sachs. 1996. Estimating single channel kinetic parameters from idealized patch-clamp data containing missed events. *Biophys. J.* 70:264–280.

Qin, F., A. Auerbach, and F. Sachs. 1997. Maximum likelihood estimation of aggregated Markov processes. *Proc. R. Soc. Lond. B.* 264:375–383.

Qin, F., J. L. Chen, A. Auerbach, and F. Sachs. 1994. Extracting kinetic parameters using hidden Markov techniques. *Biophys. J.* 66:392*a* (Abstr).

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE.* 77:257–286.

Roux, B., and R. Sauve. 1985. A general solution to the time interval omission problem applied to single channel analysis. *Biophys. J.* 48:149–158.

Salamone, F. N., M. Zhou, and A. Auerbach. 1999. A re-examination of adult mouse nicotinic acetylcholine receptor channel activation kinetics. *J. Physiol. (Lond.).* 516(Pt. 2):315–330.

Titterington, D., A. Smith, and V. Markov. 1985. Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons, New York.

Venkataramanan, L., R. Kuc, and F. J. Sigworth. 1998b. Identification of hidden Markov models for ion channel currents. II. State-dependent excess noise. *IEEE Trans. Signal Processing.* 46:1916–1929.

Venkataramanan, L., J. L. Walsh, R. Kuc, and F. J. Sigworth. 1998a. Identification of hidden Markov models for ion channel currents. I. Colored background noise. *IEEE Trans. Signal Processing.* 46: 1901–1915.