STATISTICAL METHODS FOR MODEL DISCRIMINATION

Applications to Gating Kinetics and Permeation of the Acetylcholine Receptor Channel

RICHARD HORN
Department of Physiology University of California-Los Angeles School of Medicine, Los Angeles, California 90024

ABSTRACT Methods are described for discrimination of models of the gating kinetics and permeation of single ionic channels. Both maximum likelihood and regression procedures are discussed. In simple situations, where models are nested, standard hypothesis tests can be used. More commonly, however, non-nested models are of interest, and several procedures are described for model discrimination in these cases, including Monte Carlo methods, which allow the comparison of models at significance levels of choice. As an illustration, the methods are applied to single-channel data from acetycholine receptor channels.

INTRODUCTION

A biophysical examination of the properties of a type of ionic channel often results in a proposal for a kinetic model. In some cases the experimental data and subsequent analysis suggest the elimination of one type of kinetic model. Although the kinetic models are only a formalism, the details of such models may provide insight into the molecular processes underlying the phenomena being examined, and may be useful in suggestions for new experiments. A typical procedure is to examine the predictions of several models, and see which models are particularly good (or bad) at describing the experimental data. Models typically have unknown parameters that must be estimated from the data, using either casual (e.g., "by eye") fitting procedures or more rigorous statistical methods.

The problem addressed in this paper is that of choosing between models, based on the available data, by use of statistical criteria. I will use two examples for illustration of possible methods. The first involves open time data from single acetylcholine-receptor (AChR) channels, and the second involves a comparison of models for cesium permeation through the open AChR channel. The former example makes use of maximum likelihood methods. In the latter example regression is used.

In this paper I will differentiate between two general classes of model comparison. The first (and simplest) class compares any two models that are "nested" in the sense that one is a smoothly parametrized subhypothesis of the other. In this class a standard hypothesis test can be used in which the subhypothesis is considered to be the null hypothesis. The "level" of the test (i.e., the probability of rejecting the null hypothesis, when true) is selected, and a test of high power against the alternative model is desired. In general the power of the test, defined as the probability of rejecting the null hypothesis when false, increases with sample size for this class, while the level remains constant. That is, the two models are considered asymmetrically.

In the second class the models to be compared are non-nested, and are sometimes referred to as "separate." In this class any two models may be considered either symmetrically, in which case the probability of selecting the wrong model decreases for both models as the sample size increases, or else asymmetrically, in which case one model is favored as the null hypothesis. Asymmetric model choice is often called hypothesis testing, whereas the symmetric version is sometimes referred to as model discrimination. Although I will only compare two models at a time for simplicity, the generalization to several models is straightforward.

THEORY AND METHODS

Nested Models

The theory for nested models is described in many text books (e.g., Rao, 1973) and is presented in abbreviated form here. For composite hypotheses, i.e., those that require the estimation of parameters, maximum likelihood methods can be used when the probability density for each model is known. Suppose model F is a smoothly parametrized subhypothesis of model G. The probability densities for these two models may be written as f $(\mathbf{x}, \theta)$ and g$(\mathbf{x}, \beta)$, where the data are represented by the vector **x**. θ and β are unknown parameters with dimensions k_f and k_g. By

Dr. Horn's present address is Neurosciences Department, Roche Institute of Molecular Biology, Nutley, NJ 07110.

assumption $k_g > k_f$. The natural logarithm of the likelihood ratio is defined as

$$\mathrm{LLR} = \log\left\{\frac{\sup\limits_{\beta} g(\mathbf{x}, \beta)}{\sup\limits_{\theta} f(\mathbf{x}, \theta)}\right\} = \log\left\{\frac{g(\mathbf{x}, \hat{\beta})}{f(\mathbf{x}, \hat{\theta})}\right\}.$$

$\sup_\beta g(\mathbf{x}, \beta)$ denotes the supremum of $g(\mathbf{x}, \beta)$ with respect to β. Here $\hat{\beta}$ and $\hat{\theta}$, the parameter values that maximize the likelihood for each probability density, are the maximum likelihood estimates of β and θ. It is well-known (e.g., Rao, 1973) that under certain regularity conditions when model F is true, 2·LLR has a central chi-square distribution asymptotically, i.e., for large samples, with $k_g - k_f$ degrees of freedom. The LLR can thus be used to test whether model G is better than model F at an α-level of significance, where the probability of rejecting model F, when true, is asymptotically less than or equal to α (using the terminology in Lehmann, 1959). Typically, α is set to either 0.05 or 0.01.

A similar test exists for regression models (Rao, 1973). In this situation the residual sums-of-squares of errors (SSE) under the two models are defined as

$$\mathrm{SSE}_f = \min_{\theta} \sum_{i-1}^{n} [\mathbf{x}_i - E_f(\mathbf{x}_i|\theta)]^2 \text{ for model F}$$

and

$$\mathrm{SSE}_g = \min_{\beta} \sum_{i-1}^{n} [\mathbf{x}_i - E_g(\mathbf{x}_i|\beta)]^2 \text{ for model G.}$$

Here $\mathbf{x}_i$, $i = 1, \ldots, n$, are n data points, and the expected value of $\mathbf{x}_i$ predicted (for example) by model F, given the parameter θ, is $E_f(\mathbf{x}_i|\theta)$. It is assumed that the residuals about the expected values are independent, identically distributed Gaussian variables. The parameters that minimize SSE are the maximum likelihood estimates in these regression problems. If model F is a subhypothesis of model G, then the statistic T is defined as

$$\mathrm{T} = \frac{(\mathrm{SSE}_f - \mathrm{SSE}_g)}{\mathrm{SSE}_g} \cdot \frac{(n - k_g)}{k_f}.$$

T has an F distribution, asymptotically for large n, with k_f and $n - k_g$ degrees of freedom (Rao, 1973). The value of T can be compared with tabulated values of the F distribution at the desired α-level. Note that this is also a form of likelihood ratio test (Rao, 1973).

Non-nested Models

The theory for comparison of non-nested models is not as well-established as that in the previous section. I will discuss and use two general types of methods. The first type allows for convenient ranking of non-nested models, where the models are treated symmetrically. The underlying theory is applicable to both maximum likelihood and regression models, and is called a "prediction error" approach (Akaike, 1974). The convenience of this type of analysis is offset by the fact that significance levels for model discriminations are not known. Consequently one model is always ranked better or worse than another when this method is used. In other words, two models will always be discriminated, no matter how few data are available. In the second type of analysis, which uses Monte Carlo methods, it is possible to set the significance level for a hypothesis test at a chosen value, as shown below. The two types of analysis are discussed in turn.

For both types of analysis the quantities of interest are LLR (defined above) for a maximum likelihood approach and

$$\mathrm{LER} = \log \text{ (error ratio)} = \log (\mathrm{SSE}_f/\mathrm{SSE}_g)$$

for regression. Positive values of either LLR or LER tend to support model G, whereas negative values tend to support model F. In an approximate sense, values of LLR or LER near zero are suggestive that the two models are indistinguishable.

The critical values, often referred to as the "asymptotic information criteria" (AIC), of LLR or LER are simply obtained in Akaike's formulation. For maximum likelihood, model F is rejected in favor of model G if LLR > $(k_g - k_f)$ (Akaike, 1974). For regression problems, model F is rejected if LER > $2(k_g - k_f)/n$ (Akaike, 1974). Here n refers to the number of data points. There is some dispute, beyond the scope of this paper, over the merits of the AIC (Leamer, 1983), and other critical values have been proposed for discrimination of non-nested models (e.g., see Schwartz, 1978). It should be noted, however, that this approach rewards a model for parsimony, in that a model pays a cost for having extra free parameters. The critical value is zero when comparing two models with an equal number of free parameters.

Estimation of The Distribution of LLR and LER

To compare two non-nested models at a given significance level it is desirable to have some idea of the distribution of LLR or LER under an appropriate hypothesis. Cox (1961, 1962) has shown, in some cases, that LLR and LER have normal distributions asymptotically (i.e., as sample size increases) under the hypothesis that, for example, model F is true. That is

$$\lim_{n\to\infty} \mathrm{LLR}_f \rightarrow N(\mu_f, \sigma_f^2)$$

and

$$\lim_{n\to\infty} \mathrm{LER}_f \rightarrow N(\mu_f, \sigma_f^2)$$

The notation $N(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. LLR_f and LER_f denote the random variables under the assumption that model F is true. Cox (1961, 1962) showed analytically that asymptotic normality is obtained for specific examples, and White (1982) has discussed regularity conditions under which the assumption of normality holds.

Since the parameters of these Gaussian distributions are not readily calculated for the complicated kinetic models appropriate for describing ionic channels, they must be estimated by Monte Carlo simulation leading to parametric tests based on normality or non-parametric tests, which are independent of the underlying distribution of LLR_f or LER_f (Williams, 1970; Loh, 1985). Such tests will be shown below for the examples using single-channel data.

I have also estimated the distribution of LLRs using a version of resampling known as the bootstrap (Efron, 1982). In this method the data, assumed to be independent and identically distributed samples, are treated as a subpopulation that is resampled with replacement to create artificial data sets. Each data point is chosen with equal probability, leading to a data set of the same size as that of the original, but with some values omitted and some values repeated. A large number (I have used 200) of such data sets may be generated. For each bootstrap data set LLR is calculated, and the empirical distribution of LLRs is obtained. Hypothesis tests may then be formulated about this distribution. If parsimony is ignored, the null hypothesis is that the distribution of LLRs is centered at zero. This is the natural expectation if models F and G are indistinguishable. If the LLRs are consistently greater than zero for repeated realizations of the data, then model G consistently has a higher likelihood. Therefore zero may be used in obtaining a critical value for an hypothesis test. The form of the test will obviously depend on the empirical distribution of the LLRs, which turned out to be approximately Gaussian (see below). If it is desirable to reward parsimony it is also possible to use the AIC to obtain a critical value for an hypothesis test, as shown in Results. It should be noted that this procedure treats the two hypotheses symmetrically. I have limited the bootstrap method to an example where it was reasonable to assume that the data were independent samples of a univariate random variable. The bootstrap method is more problematic for regression problems, because it is inappropriate to assume that each of the X-Y pairs is equally likely to occur.

Statistical Analysis and Monte Carlo Simulations

The above procedures require a robust method for finding the global maximum of the likelihood surface, or the global minimum of the error surface in regression problems. I have used a variable metric algorithm (Powell, 1978) kindly provided by Dr. Kenneth Lange. All calculations were performed on a VAX 11/730 computer (Digital Equipment Corp., Marlboro, MA), occasionally with the assistance of an attached array processor (FPS100, Floating Point Systems, Inc., Portland, OR). Although some of the models had as many as 11 free parameters, the final values of the likelihoods or squared-errors were independent of initial guesses, except for very poor guesses.

The simulation of data for maximum likelihood analysis required the derivation of the inverse of the density function for each model (Zelen and Severo, 1964). The value of the simulated random variable was then obtained by taking the inverse function of a uniformly distributed random deviate provided by the computer's FORTRAN library.

Two methods were used for simulating data in regression models. The first method made the usual assumption that the errors around the theoretical regression line had a Gaussian distribution with a constant variance given by $\mathrm{SSE}_i/(n - k_i)$ for model i (Rao, 1973). The second method, used in a more limited study, made no assumption about the distribution of errors except that it was independent for each observation. In this method the deviates obtained from the actual data were resampled, i.e., sampled with replacement, and added to the expected values under a given regression model (see Efron, 1982) to create a simulated data set.

RESULTS

The Results are divided into two main sections. The first handles maximum likelihood analysis of kinetic models for open times; the second treats regression analysis of permeation data. For each section nested and non-nested kinetic models will be explored. It should be emphasized that the analysis and models of real data are presented only as examples, and are not meant to represent conclusive tests of the presented models.

Open Time Distributions

In the absence of acetylcholine the AChR channel opens spontaneously (Jackson, 1984). Fig. 1 shows a typical density histogram of open times for such openings, obtained by Dr. M. Jackson under the conditions described in Jackson (1984, 1986). Three models will be considered to describe these data.

Model A: One Open State, Markovian. Model A, which has one free parameter, assumes a single open state with a constant hazard function (i.e., a time-homogeneous Markov model). The probability density is

$$\mathrm{f}_A(t) = a \cdot \exp(-a \cdot t).$$

Model B: Two Open States, Markovian. Model B, which has three free parameters, assumes a time-homogenous Markov chain with two open states. The probability density (e.g., see Colquhoun and Hawkes, 1981) is

$$\mathrm{f}_B(t) = w \cdot r_1 \exp(-r_1 t) + (1 - w) r_2 \exp(-r_2 t), 0 < w < 1.$$

TABLE I
ANALYSIS OF THREE GATING MODELS FOR OPEN TIME DENSITIES

Model A: One open state, Markovian

$$\mathrm{f}_A(t) = a \cdot \exp(-a \cdot t)$$
$$\text{maximum log (likelihood)} = -21.190$$
$$\hat{\mathbf{a}} = 2.64 \pm 0.10 \text{ ms}$$

Model B: Two open states, Markovian

$$\mathrm{f}_B(t) = w \cdot r_1 \exp(-r_1 t) + (1 - w) r_2 \exp(-r_2 t), 0 < w < 1$$
$$\text{maximum log (likelihood)} = 69.648$$
$$\hat{\mathbf{r}}_1 = 6.95 \pm 0.94$$
$$\hat{\mathbf{r}}_2 = 1.24 \pm 0.16$$
$$\hat{\mathbf{w}} = 0.847 \pm 0.024$$

Model C: One open state, non-Markovian

$$\mathrm{f}_C(t) = c \cdot e^{d \cdot t} \exp[(c/d)(1 - e^{d \cdot t})]$$
$$\text{maximum log (likelihood)} = 58.098$$
$$\hat{\mathbf{c}} = 4.77 \pm 0.27$$
$$\hat{\mathbf{d}} = -1.10 \pm 0.14$$

Parameter estimates for each model are presented with their standard errors under the assumption that the model is true.

Model C: One Open State, Non-Markovian. Model C, which has two free parameters, is a time-inhomogeneous reaction with one open state. The closing rate is assumed to be an exponential function of time (Easton, 1981; Levitan, E., personal communication). The physical significance of such a model is speculative, but it could correspond to a closing rate that is proportional to a diffusible substance, whose concentration is decreasing exponentially. If the hazard function of the closing reaction is given by $c \cdot \exp(d \cdot t)$, the probability density is

$$\mathrm{f}_C(t) = c \cdot e^{d \cdot t} \exp[(c/d)(1 - e^{d \cdot t})].$$

The 742 open times over a fixed range (0.1875–5.0 ms) were fit to these three models, using a maximum likelihood procedure, with the results shown in Table I. The densities for each model were made conditional on the open times falling within the above range. The conditional density for model i, assuming a range $[t_1, t_2]$, with $t_2 > t_1$, is

$$\mathrm{f}_i(t|t_1, t_2) = \frac{\mathrm{f}_i(t)}{\int_{t_1}^{t_2} \mathrm{f}_i(s)\mathrm{d}s}.$$

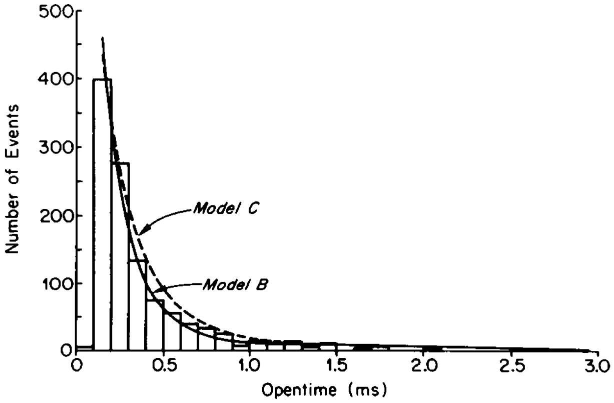

FIGURE 1 Histogram of open times for single AChR channels. Data obtained as described in Jackson (1986). Best fit curves are shown for models B and C.

Note that model A is a subhypothesis of both models B and C. Thus a likelihood ratio test, using a chi-squared table, can be used to show that it is inferior to both model B ($2 \cdot$ LLR = 181.7; $P < 10^{-4}$) and model C ($2 \cdot$ LLR = 158.6; $P < 10^{-4}$). Thus a single exponential is an inadequate density to fit these data (Jackson, 1986).

Non-exponential lifetime distributions are usually fit by sums of exponentials (Colquhoun and Hawkes, 1981; Colquhoun and Sigworth, 1983), which is the theoretical prediction of the usual Markov chain models. Model B is such a model, and its theoretical density is shown in Fig. 1. Model C is another alternative (also plotted in Fig. 1) and is a separate (i.e., non-nested) hypothesis from model B, in the sense described above.

Using the AIC criterion suggests that model B is preferable to model C. Defining LLR as the logarithm of the ratio of the maximal density of model B to that of model C leads to the following:

$$\mathrm{LLR} - (k_B - k_C) = 10.55 \, .$$

The value is greater than zero, leading to the conclusion that model B is better than model C (Akaike, 1974).

To set a significance level for a statistical comparison of models B and C, I used two different procedures (see Theory and Methods). In the first procedure 200 bootstrap data sets were generated from the data at hand, to estimate the variability of LLR. This involved sampling the original 742 open times with replacement to create resampled data sets, each having 742 values. Maximum likelihood fits were obtained for the two models for each "resampled" data set. Therefore a value of LLR was obtained for each data set. This was repeated 200 times and the distribution of LLRs is plotted as a histogram in Fig. 2. This distribution of LLRs was approximately Gaussian, as shown by the best fit theoretical curve. The LLR for the original data set (11.55) is shown by an arrow. If parsimony is ignored, it is appropriate to compare this distribution with a region bounded by zero, since a value of LLR close to zero suggests that models B and C are not significantly different. The histogram shows clearly that the distribution of LLRs is greater than zero, suggesting that model B consistently has a greater likelihood than model C. In fact all 200 LLRs from the resampled data sets were greater than zero. The probability of this occurrence under the null hypothesis is <0.02 (confidence coefficient >0.98; Lawless, 1982). The Gaussian curve in Fig. 2 has a mean of 12.97 and a standard deviation of 6.34. The probability of a value as extreme as zero being sampled at random from such a distribution is ~0.02. Therefore a standard hypothesis test here, under a Gaussian assumption, leads to a rejection of the null hypothesis in favor of model B.

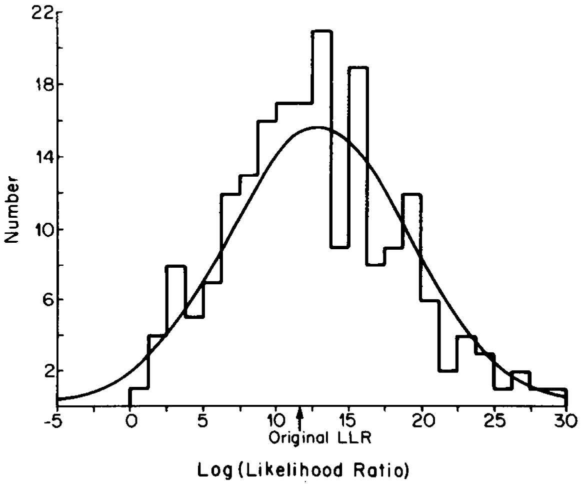

FIGURE 2 Histogram of LLRs obtained from 200 bootstrap samples of the original data. The LLR of the original data is shown by an arrow. The curve is the best fit Gaussian to the LLRs.

This hypothesis can be modified simply to reward model C for parsimony, since it has one less free parameter than model B. The rejection region will now depend on the AIC value $(k_B - k_C) = 1$, instead of zero. Under the Gaussian assumption this leads to a P-value of ~0.03. Again model B is favored.

The above bootstrap procedure treats the two models symmetrically. The second procedure I used treats the models asymmetrically, in somewhat the same style as a standard hypothesis test. In this procedure 200 data sets, each having 742 open times, were simulated assuming that model B was true. The parameters used for the simulation were the maximum likelihood estimates from the original data. For each simulated data set a value of LLR was obtained from the best fit to each model. The distribution of the LLRs obtained from these simulated data are shown on the right side of Fig. 3 in the form of a histogram. In the same fashion 200 data sets were simulated under the assumption that model C was correct. The LLRs for these data sets are shown on the left side of Fig. 3. The original LLR, obtained from the best fit to the real data, is shown below by an arrow, and is qualitatively consistent with model B, but not model C.

To make the results in Fig. 3 quantitative, two approaches were used. Cox (1961, 1962) suggested that the asymptotic distributions of LLRs for the situation shown in Fig. 3 are Gaussian. Best-fit Gaussian curves are shown in Fig. 3. Under model B, the mean and standard deviation are estimated to be 19.23 and 6.10. Thus the original LLR is a possible member of this distribution ($P \sim 0.2$). On the other hand the mean and standard deviation under model C are 0.464 and 0.988, which are not consistent with the original LLR ($P < 10^{-5}$).

It is reasonable to question the assumption of normality for the distributions in Fig. 3, since regularity assumptions made by Cox do not necessarily hold for this problem. The Anderson-Darling statistic A^2, based on the empirical distribution, was used to test normality (Stephens, 1984). Each distribution shown in Fig. 3 was significantly non-Gaussian ($P < 0.025$) by this criterion. Therefore, as above, it is possible to use a nonparametric approach to ask

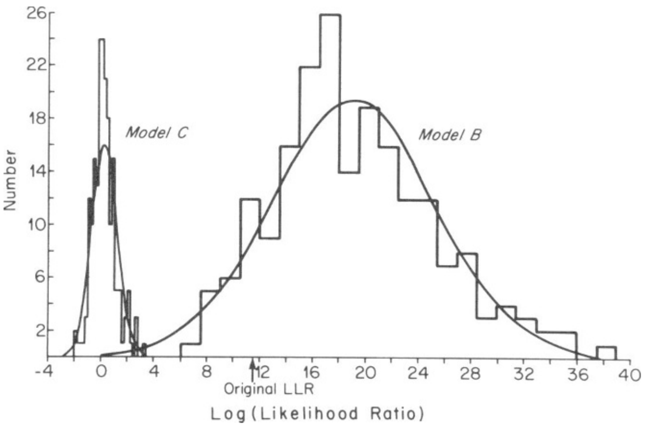

FIGURE 3 Histograms of LLRs from data simulated under either model C (*left*) or model B (*right*). Best fit Gaussian curves are displayed. The LLR from the original data set (*arrow*) is consistent with model B but not model C.

whether the original LLR is a member of either, or both, of the two empirical distributions in Fig.3. The original LLR is bigger than all values from data simulated under model C, indicating that it is an unlikely member of the distribution under model C. On the other hand if the 200 values under model B are ranked in magnitude, the original LLR falls between the 16th and 17th value. The probability of selecting the original LLR by chance then is ~16.5/200 = 0.083. Thus the nonparametric approach agrees with the above analyses based on normality, and both of these approaches agree with the simpler analysis of using the AIC value.

Note that model B is frequently better than model C, in terms of its likelihood, even when model C is used to simulate data. This fact, by itself, should not be particularly surprising, since model B has an extra free parameter. For non-nested models this result will not always hold; however this is the expected result for nested models. For example, model B, with two exponentials, will always produce a higher likelihood than model A, with one exponential, even if the latter simulated the data. The point of emphasis in Fig. 3 is that the original LLR is unlikely to be a member of the distribution of LLRs under model C.

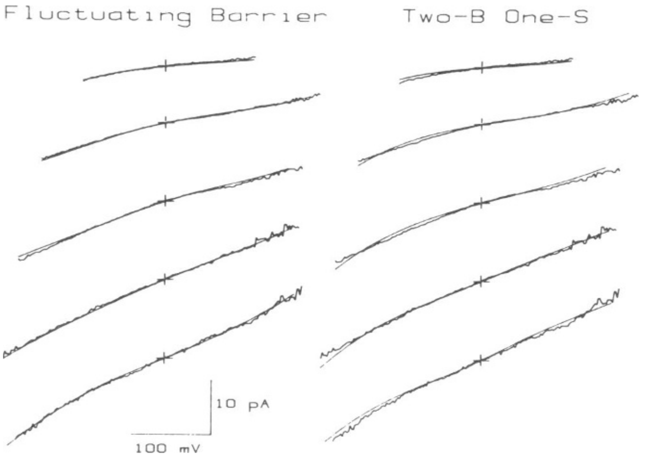

FIGURE 4 Current-voltage (I-V) curves and regression fits to models A (*left*) and C (*right*). Data obtained as described in Dani and Eisenman (1984, 1985). The symmetrical [Cs] for the five I-V curves is (from *top* to *bottom*): 7 mM, 20 mM, 45 mM, 150 mM, and 300 mM.

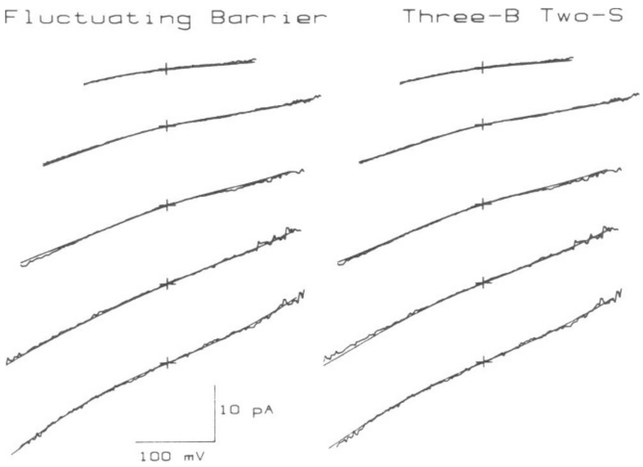

FIGURE 5 Current-voltage (I-V) curves and regression fits to models A (*left*) and B (*right*).

In conclusion the two-state Markov model B is significantly better than the one-state non-Markovian model C in describing the open time data.

Permeation

Current-voltage (I-V) relationships are shown for single open AChR channels in Figs. 4 and 5. These data were obtained by Drs. John Dani and George Eisenman using experimental methods described in Dani and Eisenman (1984). Five I-V relationships are plotted for five different concentrations of cesium ([Cs] is given in the legend of Fig. 4). The concentrations ranged from 7 mM (uppermost curve) to 300 mM (lowermost curve). The currents were recorded during application of voltage ramps. A theoretical description of these data should account for the I-V relationship over the entire range of concentrations. Fig. 6

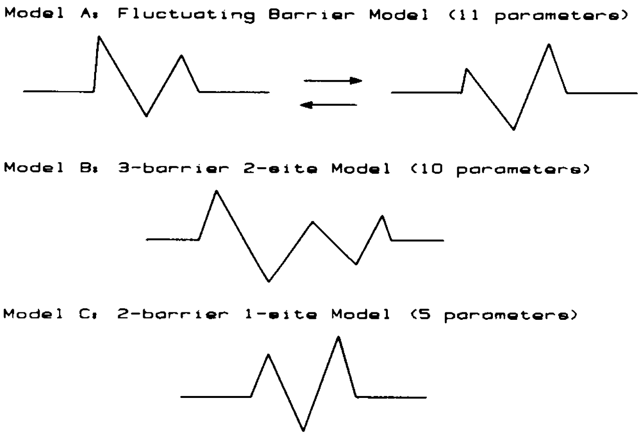
Model A: Fluctuating Barrier Model (11 parameters)
Model B: 3-barrier 2-site Model (10 parameters)
Model C: 2-barrier 1-site Model (5 parameters)

FIGURE 6 Diagrams of the three barrier models.

shows diagrams of three "Eyring-barrier" models (e.g., see Eisenman and Horn, 1983), which will be subjected to statistical analysis.

Model A: Fluctuating Two-Barrier One-Site (2B1S). Model A assumes a single saturable binding site for an ion in the middle of the electric field of the membrane. The site is flanked by two energy barriers. The location of the energy barriers in the field and their energies are free parameters (see Eisenman and Dani, 1985, for a description of this model). It is also assumed that the energy levels of the barriers and the binding site can fluctuate between two levels (Lauger et al., 1980). The rates of fluctuation are also free parameters. In total this model has 11 free parameters, two for electrical locations of the barriers, six for energy levels of barriers and wells, and three for transition rates between the two configurations (Eisenman and Dani, 1985).

Model B: Static Three-Barrier Two-Site (3B2S). In Model B an ion encounters three barriers and two sites in traversing the open channel. The energy levels and locations of the barriers and sites comprise 10 free parameters, five for energy levels and five for electrical locations. The energy profile is static (i.e., does not fluctuate). The channel may be occupied by at most one ion at a time.

Model C: Static 2B1S. This is a static version of model A and has five free parameters: three energies and two locations.

For regression analysis 129 data points were taken, at approximately regular intervals, from the raw curves shown in Figs. 4 and 5. These two figures also show the best fit regression curves for the above models. The fitting procedure is a least squares minimization of the SSEs for each model, assuming constant error variance at each point. For models A, B, and C the minimum SSEs were 5.839, 7.926, and 27.176 pA^2, respectively. Note that model C is a subhypothesis of model A, indicating that a standard hypothesis test can be used here. In this case an F statistic has a value of 86.2, with 5 and 118 degrees of freedom. This is significant ($P < 10^{-3}$), indicating that the fluctuating 2B1S model A is significantly better than the static 2B1S model C (see Fig. 4).

Model A and B are not nested, however. The question is whether model A is statistically better, in some sense, than model B, which has a larger SSE. As a first step the AIC value was calculated. If LER is defined as

$$\log\left(\mathrm{SSE_A}/\mathrm{SSE_B}\right),$$

then

$$\mathrm{LER} - 2(k_\mathrm{B} - k_\mathrm{A})/129 = -0.290.$$

The negative value here suggests that model A is better than might be expected just from its extra free parameter.

To set an α-level for hypothesis tests the empirical distribution of LERs was obtained for data simulated under the two models. The parameters for the simulations were the estimates obtained from the original data. The residuals around the theory curve for model i were sampled from a Gaussian distribution with zero mean, and variance equal to the minimum $\mathrm{SSE_i}$ divided by $(129 - k_\mathrm{i})$. For models A and B the standard deviations for these residuals were 222 and 258 fA, respectively. Fig. 7 shows the empirical distributions of LERs for data simulated under each hypothesis and also shows best-fit Gaussian curves for these distributions. In this analysis model A is significantly better than model B, since the original LER is within one standard deviation of the mean of the LERs under model A, but is nearly six standard deviations below the mean for the LERs under model B (see figure legend).

This procedure might be criticized for the method of simulation, because the distribution of residuals may not be Gaussian. Therefore I have also tried the procedure illustrated in Fig. 7 with the following modification. The empirical distribution of residuals was obtained for the original data under a given model. Then the simulations were performed by random sampling with replacement from this distribution, and a sampled residual was added to each theoretical value. Using this method 20 simulated data sets were created under each model, and these data were analysed as above. These data were, in every way, comparable to those simulated by using Gaussian residuals. The original LER was within one standard deviation of the mean of the LERs under model A, but was approximately five standard deviations below the mean for the LERs under model B.

In conclusion, at an α-level of choice, say $\alpha = 0.05$, the original LER is significantly different from LERs simulated under model B, but not those under model A. This result supports the choice of model A over model B, in agreement with the AIC value.

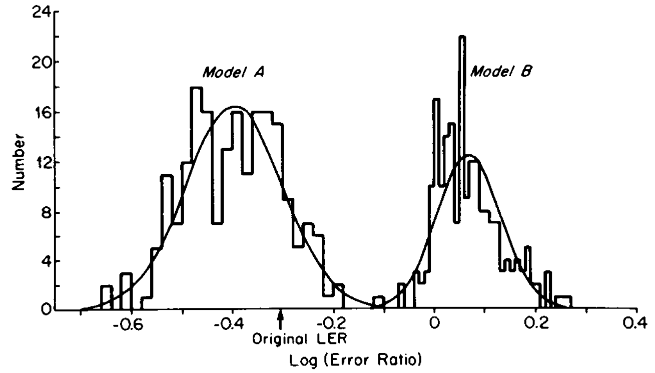

FIGURE 7 Histograms of LERs from data simulated under either model A (*left*) or model B (*right*). Best fit Gaussian curves are displayed. Under model A the mean and standard deviation of LERs were −0.397 and 0.097, respectively. Under Model B the mean and SD were 0.069 and 0.064. The LER from the original data set (*arrow*) is consistent with model A but not model B.

DISCUSSION

This paper has examined several procedures for model discrimination, using single-channel data and either likelihood or regression methods. In general, models are compared by an examination of the relative values of either likelihoods or error sum-of-squares (SSE). The procedures for comparing nested models were adapted from classical literature (Lehmann, 1959; Rao, 1973). Comparisons of non-nested, or separate, hypotheses is more difficult, both theoretically and practically. The simplest method described here, the AIC value, is based on a prediction error model. In the examples given in the text the model choice using this criterion agreed with more elaborate Monte Carlo procedures, which use resampling or simulation to examine the distribution of the logarithm of the likelihood ratio (LLR) or the logarithm of the ratio of SSEs (LER) for any two models. Although the agreement is satisfying and supports the use of the simpler method, it is necessary to realize that this is somewhat anecdotal, in that only two examples were analyzed. Furthermore, as discussed below, there are some clear differences in the goals of the two approaches, and it is possible to imagine cases in which the prediction error method will be guaranteed to disagree with the Monte Carlo methods.

The necessity or desirability of the use of any of the above methods depends heavily on the goals of a given project. In physiological experiments it is often possible to discriminate classes of models by experimental designs that obviate the need for elaborate statistics. However recent trends in biophysical studies of ion channels make this more difficult. One trend is the interest in kinetic models, which are necessarily complicated to explain the data. Often, two complicated models are both reasonably capable of describing the data, but one would like to know if one model is arguably (i.e., statistically) superior. Another trend is the use of single-channel data for studies of gating. The stochastic, at times chaotic, behavior of single channels is analyzed by probabilistic models, which are ideally suited to likelihood methods. The rationale of this approach derives from the fact that any data set can be described as the more-or-less likely result of any model. However one of the main drawbacks of statistics in such studies, as opposed to more casual comparisons, is the seductive tendency to believe that such methods can uncover "the correct model." Probably the best one can hope for are methods for eliminating incorrect or non-useful models.

All of the methods described in this paper are based on the common assumption that likelihood, and its equivalent SSE in regression models, are reasonable criteria for the goodness or badness of a model. These criteria have been evaluated elsewhere (Cox and Hinkley, 1974, Chapter 9). Another common criterion used in model selection is parsimony, a principal in which a model is penalized for having an excessive number of parameters. Parsimony is rewarded in standard hypothesis tests and in the Akaike method. It can also be introduced into the bootstrap procedure by using the AIC value in formulation of a hypothesis test. However parsimony is not rewarded in the procedure which involves simulation. This implies that a model that is unnecessarily complicated may, in principal, be chosen over a much simpler model, which can reasonably be claimed to describe the data. In this situation the use of the AIC value alone would tend to pick a simpler model than that chosen by simulation. Another obvious difference between the AIC method and both Monte Carlo methods is that the former always leads to a choice of one model over another, whereas the latter, in the style of standard hypothesis tests, may fail to reject the null hypothesis. Furthermore it is possible to set the α-level on hypothesis tests with Monte Carlo methods.

The two Monte Carlo methods used here are not equivalent. The first, which examines the distribution of LLR for resampled (i.e., bootstrap) data sets, provides insight into the properties of LLR under repeated realizations of the data. If, for example, the LLR is positive for every data set, then one model always has a higher likelihood than the other. Thus this distribution of LLRs shows whether one model is consistently higher in likelihood than another in describing the data. In the example used this distribution was approximately Gaussian, and we might say that one model is better than the other if, for example, it has a higher likelihood in 95% of the cases. Thus this bootstrap method leads to the following set-up: The null hypothesis is that two models (say, model A and B) each have the same likelihood. The two alternatives are either that model A or that model B has a greater likelihood. This has the form of a standard two-sided hypothesis test (Lehmann, 1959). The α-level is selected by choosing, for example, the proportion of LLRs that are allowed to be greater (or less) than zero, before rejecting the null hypothesis. Note that parsimony can be included in this set-up by using the AIC value, rather than zero, for the alternative hypotheses. The original problem of discriminating between model A and B has thus been reformulated into a standard hypothesis test about the distribution of LLRs under resampling. The unfortunate character of this reformulation, however, is that the interpretation of the final test is not clear (Rubin, 1981). Although the reformulated hypothesis has a standard set-up, the null hypothesis has an ill-defined relationship with the original model discrimination problem.

The second method, which uses simulation, has a stronger theoretical foundation than the above bootstrap method. This is, in part, because the null hypotheses assume that one model is true, and examine the distribution of LLRs or LERs under this assumption. Thus four decisions are possible (Williams, 1970): (*a*) model A is selected, (*b*) model B is selected, (*c*) neither model is rejected, or (*d*) both models are rejected. The decisions fall naturally out of the comparison between the LLR (or LER) from the original data and those from simulations under each model. Another difference between the two

Monte Carlo methods in this paper is that the bootstrap method treats the two models symmetrically, whereas the second method is a sequence of two asymmetric hypothesis tests, where each model is treated as the null hypothesis. The simulation method is reminiscent of Box's Bayesian model checking formalism, where model checks are based on the predictive distributions of functions of the data, here the LLR or LER (Box, 1980).

As mentioned earlier, the simulation method does not reward models for parsimony. For many practical cases this is not important, because the models often used in biophysical examples are already overparametrized, i.e., the parameters are not identifiable. This was the case in the permeation problem examined above, where the models A and B had 11 and 10 free parameters. Although it was possible to find a minimum SSE for these models, the resulting information matrix was usually not positive definite. Therefore the true number of free parameters was not known. The difficulty with overparametrized models for the Monte Carlo methods is not the ignorance about the number of free parameters. A greater problem is confidence in the robustness of the algorithm for estimating the parameters. It is expecially important that it is capable of reliably finding a global maximum (for likelihood) or minimum (for SSE). Although I did not study this property extensively, the variable metric method I used was robust in that the final likelihood or SSE was independent of initial values of parameters in several test examples.

Evaluation of the two Monte Carlo Methods

The principal purpose and advantage of the two Monte Carlo methods presented here is that they provide a mechanism for deciding the α-level of tests for the difference between two separate hypotheses. Thus it is possible to choose either of two models, or decide that they are indistinguishable, e.g., at the $\alpha = 0.05$ level. This is not possible with AIC value, which always leads to the rejection of one model for the other. It seems likely also that the method based on simulation has more power than the bootstrap method, since the former method relates the data more specifically to the underlying hypotheses. This needs further exploration.

There are two main disadvantages of these two methods. The first is that theoretical justification is more problematic, by comparison with classical hypothesis tests. This is particularly true of the bootstrap method. One of the recognized, and potentially solvable, problems with the simulation method, is that the tests are not guaranteed to be level α (Loh, 1985). The reason for this problem is that the simulations depend on the use of estimated parameters, which tend in general to lead to tests whose size is greater than level α. In other words this method leads to a greater-than-expected chance of rejecting the null hypothesis when it is true. Loh (1985) suggests procedures that yield α-level tests. However, his methods add much more computation to a method that is already computationally intensive. In practice Loh's improvement is not substantial, although he presents an extreme example where this problem is severe (Loh, 1985). The other main disadvantage with the Monte Carlo methods is that they are somewhat complicated and computer-intensive. This is an example of the brute force use of a computer when analytical methods are not feasible. This admittedly inelegant approach to theoretical problems is becoming increasingly popular as theoretical models get more complicated and computers get faster (Efron, 1979).

I thank Drs. Kenneth Lange and Ker-Chau Li for helpful suggestions and Dr. Lange for the use of his variable-metric algorithm. I also thank Drs. Meyer Jackson, Elliot Landaw, Roderick Little, and Carol Vandenberg for helpful discussions and for comments on the manuscript. Drs. Meyer Jackson, John Dani, and George Eisenman kindly provided data used for statistical analysis. Finally, I thank Chris Miller for not having the stomach to enter this field.

This work was supported by National Institutes of Health grants NS-18608, NS-00703, and National Science Foundation grant BNS-84-11033.

Received for publication 3 June 1986 and in final form 6 October 1986.

REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Automatic Control.* AC-19:716–723.

Box, G. E. P. 1980. Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Statistical Soc. Ser. A.* 143:383–430.

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

Colquhoun, D., and F. Sigworth. 1983. Fitting and statistical analysis of single-channel records. *In* Single Channel Recording. B. Sakmann and E. Neher, editors. Plenum Publishing Corp., New York. 191–264.

Cox, D. R. 1961. Tests of separate families of hypotheses. *Proc. Berkeley Symp. Math. Statistics Probability, 4th.* 1:105–123.

Cox, D. R. 1962. Further results on tests of separate families of hypotheses. *J. R. Statistical Soc. Ser. B.* 24:406–423.

Cox, D. R., and D. V. Hinkley. 1974. Theoretical Statistics. Chapman & Hall, London.

Dani, J. A., and G. Eisenman. 1984. Acetylcholine-activated channel current-voltage relations in symmetrical Na^+ solutions. *Biophys. J.* 45:10–12.

Easton, D. M. 1981. Mathematical theory of macroscopic voltage clamp currents. International Biophysics Congress Abstracts, 8th, 274.

Efron, B. 1979. Computers and the theory of statistics: thinking the unthinkable. *SIAM (Soc. Ind. Appl. Math.) Rev.* 21:460–480.

Efron, B. 1982. The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia: Society of Industrial and Applied Mathematics.

Eisenman, G., and J. A. Dani. 1985. A fluctuating 2 barrier 1 site model compared with electrical data from the acetylcholine receptor channel. *In* Water and Ions in Biological Systems. Proceedings of the Third International Conference. A. Pullman, V. Vasilescu, and L. Packer, editors. Union of Societies for Medical Sciences, Bucharest, Romania. 437–450.

Eisenman, G., and R. Horn. 1983. Ionic selectivity revisited: the role of kinetic and equilibrium processes in ion permeation through channels. *J. Membr. Biol.* 76:197–225.

Jackson, M. B. 1984. Spontaneous openings of the acetylcholine receptor channel. *Proc. Natl. Acad. Sci. USA.* 81:3901–3904.

Jackson, M. B. 1986. Kinetics of unliganded acetylcholine receptor channel gating. *Biophys. J.* 49:663–672.

Lauger, P., W. Stephan, and E. Frehland. 1980. Fluctuations of barrier structure in ionic channels. *Biochim. Biophys. Acta.* 602:167–180.

Lawless, J. F. 1982. Statistical Models and Methods for Lifetime Data. John Wiley & Sons, Inc., New York.

Leamer, E. E. 1983. Model choice and specification analysis. *In* Handbook of Econometrics. Vol. 1. Z. Grilisches and M. D. Intriligator, editors. 285–330.

Lehmann, E. L. 1959. Testing Statistical Hypotheses. John Wiley & Sons, Inc., New York.

Loh, W-Y. 1985. A new method for testing separate families of hypotheses. *J. Am. Statis. Assoc.* 80:362–368.

Powell, M. J. D. 1978. A fast algorithm for nonlinearity constrained optimization calculations. *In* Numerical Analysis. Dundee 1977. G. A. Watson, editor. Lecture Notes in Mathematics No. 630, Springer-Verlag, Berlin.

Rao, C. R. 1973. Linear Statistical Inference and Its Applications. 2nd ed. John Wiley & Sons, Inc., New York.

Rubin, D. B. 1981. The Bayesian bootstrap. *Ann. Statistics.* 9:130–134.

Schwartz G. 1978. Estimating the dimension of a model. *Ann. Statistics.* 6:461–464.

Stephens, M. A. 1974. EDF statistics for goodness of fit and some comparisons. *J. Am. Statis. Assoc.* 69:730–737.

White, H. 1982. Regularity conditions for Cox's test of non-nested hypotheses. *J. Econometrics.* 19:301–318.

Williams, D. A. 1970. Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics.* 28:23–32.

Zelen, M., and N. C. Severo. 1964. Probability functions. *In* Handbook of Mathematical Functions. M. Abramowitz and I. A. Stegun, editors. Dover Publications, Inc., New York.