Model selection: reliability and bias

Dear Sir:

We have two levels of response to the letter of Liebovitch (1989). With the growing complexity of Markov chain models, we and others would welcome a model that accurately describes the gating of single channels with just two free parameters. Unfortunately, the basic conclusion of both papers (Korn and Horn, 1989; McManus et al., 1989) is that, when tested statistically, dwell time histograms of single channel data are described better by specific Markov chain models than by Liebovitch's fractal model. Liebovitch does not argue with this conclusion.

The real issue here, however, is how to choose a particular model. Should one choose a model that best describes the details of the data and provides insight (however naive) that leads to testable hypotheses, or should one choose a model that ignores the experimentally observed details and utilizes a physically uninterpretable parameter to "emphasize the overall memory" that may exist in the channel gating process? Although our choice is probably clear, we will address this question in section one below. Liebovitch also raises a number of criticisms of our methods and of Markov chain models in general. We will respond to some of these comments in section two.

1. How can one choose among "incorrect" models?

Liebovitch (1989) has shown that a "physically meaningless" polynomial model fits our data better than a three-state Markov model. The detailed aspects of this comparison may be criticized.[1] His point, however, is that statistical criteria alone may lead to nonsensical decisions among models. This fact is undeniable, especially since it is extremely rare that one of the models under consideration is exactly true (Leamer, 1983). The (usually implicit) goal is to choose the best among the class of useful models.

Nonstatistical bias

Given the option, scientists almost always discriminate models by nonstatistical criteria, a fact that is hardly surprising. An "obvious or intuitive truth" is always more satisfying than a "statistical truth." Among the common nonstatistical criteria are:

(*a*) *Graphical fits.* Plots of the data may be consistent with one, but not another, model. For example, we found that a plot of the open time density of K channels, with the abscissa logarithmically transformed, had large multiple humps (Horn and Korn, 1989). This is inconsistent with Liebovitch's fractal model, which only has a single peak, but consistent with Markov models, which generate a peak with each additional state.

(*b*) *Simplicity.* Everything else being equal (which it almost never is), simple models tend to be selected over complicated models. Simple models are generally more appealing, and easier to communicate and remember. Parsimony is one variant of simplicity that can be handled objectively.

(*c*) *Predictive power.* A model that can predict the results of new experiments is obviously preferred over a model that, like the polynomial model discussed above, leads nowhere. In this regard, Markov chain models have a long history of predictive power, in terms of such factors as voltage- and agonist concentration-dependence of rate constants. Markov models also support the concept of memory, just not at the level of an individual state.

(*d*) *Consistency with other types of data.* Good models should be consistent with data acquired under a wide variety of conditions and experimental designs. For example, a good model of calcium channel gating should be consistent with macroscopic currents, single-channel currents, gating currents, flux measurements of calcium transport, and structural studies of the protein itself.

(*e*) *Physical intuition.* A good model of channel gating should provide insight into the physical processes underlying the kinetic behavior.

(*f*) *Sex appeal.* Models, like styles of clothing, are subject to fads. One class of model may be passé, while another is très à la mode. Markovian models have been around since their formulation by Andrei Andreevich Markov, about a century ago. Fractal models, by contrast, have been around for about a decade, and are currently very popular for computer graphics. Suppose Liebovitch's model had a name other than "fractal"? Would it have the same popularity? For example, the Liebovitch model, where the kinetic rate $k = At^{1-D}$, shares many similarities with the expo-exponential model (Easton, 1978), where the kinetic rate $k = Ae^{Dt}$. Both models have time-dependent transition rates that smoothly expand the time range of a single exponential process (e.g., see Horn, 1987, p. 257). Yet the expo-exponential model has received virtually no attention.

Handling bias in model selection

We used a Bayesian statistical framework (Jeffreys, 1961) for discrimination between two nonnested models, say model A and model B. The power of this approach to model selection is that data-independent, nonstatistical criteria may be incorporated explicitly into the statistical test. Given the data Y, the posterior odds in favor of hypothesis A (H_A) over hypothesis B (H_B) is

$$\{P(H_A|Y)/P(H_B|Y)\} = \{P(H_A)/P(H_B)\} \cdot \{P(Y|H_A)/P(Y|H_B)\},$$

where $\{P(H_A|Y)/P(H_B|Y)\}$ is the posterior odds ratio and $\{P(H_A)/P(H_B)\}$ is the prior odds ratio for the two models. The posterior odds are determined after looking at the data. The prior odds ratio, by contrast, accounts for all of the bias that is independent of the data Y. $P(Y|H_A)$ is the maximum likelihood of the data, given model A, and $\{P(Y|H_A)/P(Y|H_B)\}$ is the

1. For example, the use of an F test for comparing nonnested models produces a meaningless probability. Also, the parameters in Table 1 of Korn and Horn (1989) lead to a much better fit of the data (see our Fig. 2 *B*) than shown in Fig. 1 of Liebovitch (1989).

0006-3495/89/02/379/03 $2.00

likelihood ratio for the two models. The likelihood contains all of the information that the data can provide, with respect to a given model (Rao, 1973). If the posterior odds ratio is significantly greater than 1.0, then model A is selected. Otherwise, model B is selected.

All of the analyses in Korn and Horn (1989) and in McManus et al. (1989) are directed toward the calculation of the likelihood ratio of the Markov and fractal models. We have implicitly assumed that the prior odds ratio equals 1.0; in other words, we have not biased our analysis for one model versus the other by preconceived notions or preferences. This is by choice, but not a necessity. We could, for example, have decided that fractal models should be favored for a variety of nonstatistical reasons, and introduced this bias into the prior odds ratio. For our potassium-channel data, the prior odds must favor the fractal model by a factor of $>e^{804}$ before this method would lead to the selection of the fractal model over the Markov model. In our opinion, this is a hefty bias.

2. Reliability of our methods

Fitting sums of exponentials

The data generated by Markov models of channel gating are fit by weighted sums of exponentials. How reliable are such fits? Is it possible to estimate the number of exponential components? Such questions fall into the realm of standard statistics (Rao, 1973), because the class of Markov models are "nested" in the sense that models with fewer components are smoothly nested subhypotheses of models with more components. This problem is exactly analogous to that of estimating the number of terms in a power series. The standard method is a sequential likelihood ratio test, where the number of terms is increased until the fit does not improve at a significance level of choice (usually 0.05). Does this method work? If the time constants are sufficiently different, the weight of each component is large enough, and there are enough data, then the answer is always affirmative. The data must be sufficient to make reliable estimates, no matter how simple the model. This is a fundamental limit of estimation. The sequential likelihood ratio test chooses the simplest model that is both consistent with and supported by the data. More complicated models, with more terms, may provide marginally better fits of the data. However, the extra terms do not improve the likelihood significantly over the (simpler) model of choice. When the model becomes overdetermined (i.e., too many parameters), the likelihood no longer increases. More complicated models could be chosen, but not because they were supported by the data, and not because they had a higher likelihood.

We cannot address the reliability of parameter estimates in general, because that depends on the data and on the model that generated them. However the two data sets we examined in our paper (Korn and Horn, 1989) were reliably fit by three exponential terms. How do we know this? First, the estimates always converged to the same values, regardless of the initial guesses. Second, the standard errors of estimates were small, even using the limited data set from corneal endothelium channels (see Table 1 in Korn & Horn, 1989). This "identifiability" means that no other parameter values could give a likelihood as high as the one we calculated. In other words, the solutions were unique. Third, the addition of a fourth component did not improve the fits. These results are in contrast with Liebovitch's assertion that such estimation problems are "extremely ill conditioned." Such a malady typically manifests itself as a need for increasing numbers of exponential components as the size of data sets increase. However, there are examples in both papers (Korn and Horn, 1989; McManus et al., 1989) where large data sets are well represented by few components.

How many parameters are too many?

In a standard statistical test, two models are nested, and the simpler model is the null hypothesis, which is favored unless overwhelming evidence supports a more complicated alternative. The number of free parameters is handled automatically in this context. The comparison of two nonnested models (e.g., fractal versus Markov) is typically very different. Neither model is favored as the null. The models are treated symmetrically, and one has a choice of penalizing a model for extra parameters. Had we underestimated the number of exponential components, our choice between fractal and Markov models would not have been altered. The only effect of an error of this type is in knowing the number of free parameters for the Markov model. This type of error does not affect the likelihood ratio, the criterion we used for model selection.

The Asymptotic Information Criterion (AIC) may be used to reward nonnested models for parsimony. If two models have the same number of free parameters, the AIC chooses the model with the higher likelihood. We did not, in fact, rely on the AIC for our definitive choice of models. If the data were insufficient to make a choice, as we found for Liebovitch's data set (Korn and Horn, 1988), neither model was selected. In the problem proposed by Liebovitch for choosing the number of sides of a polygon, where Nature has tricked us by supplying a circle, the question of number of parameters is irrelevant. If the polygon is regular (i.e., equilateral and equi-angular), then each model (triangle, square, pentagon, etc.) has one free parameter to estimate, the radius of the circumscribed circle. Since the fit will improve with each additional side, the AIC will choose a polygon with an infinite number of sides, i.e., a circle. The flaw in Liebovitch's argument is the assumption that each additional side increases the degrees of freedom.

Use of log k_{eff} versus log t_{eff} plots

Liebovitch has argued for the use of plots of "effective rate constants," k_{eff}, in the evaluation of kinetic models. This "model-independent tool" is not precise enough for anything except the most casual inspection of the data. The method (Liebovitch et al., 1987) involves an arbitrary estimation procedure; straight lines are fit to the 2nd to 4th bins of semilogarithmic histograms of dwell times (why not the 3rd to 7th bins?). This method ignores much data in these histograms and produces numbers that reflect merely a vague silhouette of the observations. If models are to be evaluated by nonstatistical criteria such as these, one should carefully evaluate the bias of the methods.

Received for publication 6 September 1988 and in final form 20 October 1988.

REFERENCES

Easton, D. M. 1978. Exponentiated exponential model (Gompertz kinetics) of Na^+ and K^+ conductance changes in squid giant axon. *Biophys. J.* 22:15–28.

Horn, R. 1987. Statistical methods for model discrimination. Applications for gating kinetics and permeation of the acetylcholine receptor channel. *Biophys. J.* 51:255–263.

Horn, R., and S. J. Korn. 1989. Graphical discrimination of Markov and fractal models of single channel gating. *Comments Theor. Biol.* In press.

Jeffreys, H. 1961. Theory of Probability. 3rd ed. Oxford University Press, London.

Korn, S. J., and R. Horn. 1989. Statistical discrimination of fractal and Markov models of single channel gating. *Biophys. J.* 54:871–877.

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

Liebovitch, L. S. 1989. Testing fractal and Markov models of ion channel kinetics. *Biophys. J.* 55:00–00.

Liebovitch, L. S., J. Fischbarg, and J. P. Koniarek. 1987. Ion channel kinetics: a model based on fractal scaling rather than multistate Markov processes. *Math. Biosci.* 84:37–68.

McManus, O. B., D. S. Weiss, C. E. Spivak, A. L. Blatz, and K. L. Magleby. 1989. Fractal models are inadequate for the kinetics of four different ion channels. *Biophys. J.* 54:859–870.

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

Richard Horn and Stephen J. Korn

Neurosciences Department

Roche Institute of Molecular Biology

Nutley, New Jersey 07110