Information

Phil Nelson
University of Pennsylvania

For these slides see:
www.physics.upenn.edu/~pcn

Sunday, December 16, 2012
Focus attention

Of what use is theory, anyway?

Well... Theory is often needed to extract information from data:

✴ Sometimes theory suggests a new kind of measurement that tests a model more stringently, or distinguishes two different models more completely, than previous measurements.

✴ Sometimes our model is not obviously connected with what we can actually measure experimentally, but theory makes a connection.

✴ Sometimes the model that interests us involves the behavior of actors that we can only see indirectly in our data; theory may be needed to separate them out from each other, and from noise.

There’s more, of course, but that’s enough to get started.
Part I: Changepoint analysis in single-molecule TIRF

JF Beausang (now at Stanford), Yale Goldman, PN

Many thanks to Haw Yang. See also Lucas P. Watkins and Haw Yang *J. Phys. Chem. B* 2005
Myosin V Processivity

We’d like to know things like: How does it walk? What are the steps in the kinetic pathway? What is the geometry of each state?
One classic approach is to monitor the position in space of a marker (e.g. a bead) attached to the motor. But this does not address the geometry of each state.

The approach I’ll discuss involves attaching a bifunctional fluorescent label to one lever arm. The label has a dipole moment whose orientation in space reflects that of the arm.
Fluorescence illumination by the evanescent wave eliminates a lot of noise, and importantly, maintains the polarization of the incident light.

To tickle the fluorophore with every possible polarization, we need the incoming light to have at least two different beam directions.
Previous state of the art

It’s a bit more meaningful to convert from lab-frame angles $\theta, \phi$ to actin-frame angles $\alpha, \beta$. Unfortunately, existing analyses gave pretty noisy determinations, with pretty poor time resolution.

You could easily miss a short-lived state -- e.g. the elusive diffusive-search step (if it exists).

Can’t we do better?

JN Forkey et al. Nature 2003

Sunday, December 16, 2012
Unfortunately, the total photon counts from a fluorescent probe may not be very informative. Here we divided a time period of interest into 20 bins. There is some Poisson noise in the photon counts, of course.

If we classify the photons by polarization and bin them separately, that reveals a definite changepoint. But when exactly did it occur? Probably not at the dashed line shown, but how can we be more precise?

If we choose wider bins, we’ll get worse time resolution; if we choose narrower bins, we’ll get worse shot-noise errors. **Can we evade the cruel logic of photon statistics?**

Key point: *binning the data destroyed some information.* Something magical happens if instead of binning, we just plot photon arrival time versus photon sequence number. Despite some ripples from Poisson statistics, it’s obvious that each trace has a sharp changepoint, and moreover that the two changepoints found independently in this way are simultaneous. (A similar approach in the context of FRET was pioneered by Haw Yang.)

*JF Beausang, YE Goldman, and PCN, Meth. Enzymol. 487:431 (2011).*

Sunday, December 16, 2012
Now that I have your attention

• Why did that trick work? How did we get such great time resolution from such cruddy data?
• How well does it work? If we have even fewer photons, for example because a state is short-lived, how can we quantify our confidence that any changepoint occurred at all?
• Could we generalize and automate this trick? Ultimately we’ll want to handle data with multiple polarizations, and find lots of changepoints.

The appropriate tool is **maximum-likelihood analysis** or its big brother, **Bayesian inference**: Focus on just one “flavor” of photons (e.g. one polarization).

Suppose that in total time \( T \) we catch \( N \) photons at times \( t_1, \ldots, t_N \).

We wish to explore a family of hypotheses, that photons are arriving in a Poisson process with rate \( R \) from time 0 to time \( t_* \), and thereafter arrive in another Poisson process with rate \( R' \).

We want to find our best estimates of the three parameters \( t_* \), \( R \), and \( R' \), find confidence intervals for them, and compare the null hypothesis that there was no changepoint.

To do this, we ask for the “Likelihood,” the probability that the data we actually observed would have been observed in a world described by our model with particular values of the unknown fit parameters:

\[
\log P(t_1, \ldots, t_N | R, R', t_*) = \sum_{k=1}^{t_*/\Delta t} \log \begin{cases} R \Delta t & \text{if a photon in this slice} \\ (1 - R \Delta t) & \text{otherwise} \end{cases} + \sum_{k'=t_*/\Delta t+1}^{T/\Delta t} \log \begin{cases} R' \Delta t & \text{if a photon in this slice} \\ (1 - R' \Delta t) & \text{otherwise} \end{cases}
\]
Left: Some data, shown in traditional binned form and in the improved version.

Below: Likelihood function for placement of the changepoint. Dashed line, maximum-likelihood point. Black triangle: Actual changepoint used to generate the simulated data. The analysis found a robust changepoint, even though there were a total of just 200 photons in the entire dataset.

Payoff

Oh, yes—it also works on real experimental data, multiple-channel data, data with many different changepoints...

Now we can get back to the original motivation. Previously, people would take data from multiple polarizations, bin it, and pipe the inferred intensities into a maximum-likelihood estimator of the orientation of the fluorophore. That procedure leads to the rather noisy dots shown here. One problem is that if a transition happens in the middle of a time bin, then the inferred orientation in that time bin can be crazy.

Here the solid lines are the inferred orientations of the probe molecule during successive states defined by changepoint analysis. We see a nice alternating stride in $\phi$.

We got a 50-fold improvement in time resolution for finding changepoints, compared to the binning method, without changing the apparatus.

*JF Beausang, YE Goldman, and PCN, Meth. Enzymol. 487:431 (2011); and Biophys J in press.*

Sunday, December 16, 2012
When you only get a million photons, you’d better make every photon count.

A simple maximum-likelihood analysis accomplishes this.

In the context of TIRF it can dramatically improve the tradeoff between time resolution and accuracy.

“Requiring physical sciences”? Well, that’s where the model to be fit came from.
Part II: Go long

Big Data means Big Headaches. Here’s a tiny fragment of what comes out of a multielectrode array every few milliseconds.

Unfortunately many events are complex, with multiple overlapping spikes in many locations. And of course these may be the most interesting ones!

Our algorithm assessed “which neuron fired when” by using a Bayesian inference approach.

Many authors say bursts are a big problem, but here is a nice fit that we obtained with no special effort.

We even handle overlapping spikes, which some algorithms do not attempt.

*JS Prentice, J Homann, KD Simmons, G Tkacik, V Balasubramanian, PCN, PLoS ONE 6(7): e19884 (2011).*
Rob says we should provoke

Why Most Published Research Findings Are False

John P. A. Ioannidis

It is Time to Stop Teaching Frequentism to Non-statisticians

William M. Briggs
300 E. 71st Apt. 3R, New York, NY 10021
matt@wmbriggs.com
January 13, 2012

Maybe we need to rethink our “biostatistics” course. Maybe Bayesian inference belongs in every course.

“Overfitting: The biggest scientific problem you’ve never heard of.”
Data can’t always speak for themselves

P. Nelson, *Physical models of living systems* (WH Freeman and Co TBA)

Sunday, December 16, 2012
Power laws are ubiquitous

OK, Rob, we arrived at 10000 meters. But it would be the same figure if I were talking about enzyme turnovers.

How much weight should I give to data in the tail? Depends on how reliable they are, but also on what the model says there. Different points convey different amounts of information.

Likelihood takes care of all that.

Again… these two models have vastly different likelihood scores.
Often, when we want to justify physical modeling, we scratch our heads and say, “Well Hodgkin and Huxley was a big deal.” Indeed. But that sort of cherry-picking approach can leave the impression that this is something that happens every 50 years or so. It’s also too reverent.

Another context in which theory enters laboratory discussions is, “We need some theory to get this thing published. Go do some theory, run some ANOVA, whatever.” I’d just like to suggest that this attitude, though common, misses out on some of what theory can do for you.

Getting back to the opening point, I’d say theory is needed every day. It’s our microscope; our Geiger counter; it helps us to see the invisible. It squeezes out the information from the data.
Oh, and cells are themselves Bayesian inference machines, as they try to interpret their environment, even predict it, and respond appropriately. (Think chemotaxis.)

Not to mention brains, which are more sophisticated Bayesian inference machines.

So it’s no surprise that scientists -- who are living organisms too -- can get some mileage out of these ideas.

At first data analysis sounds like a bag of unrelated tricks -- I found that Bayesian logic ties a lot of things together.
(Things I wish I’d read earlier)

M. Denny and S. Gaines
*Chance in biology*
Princeton Univ. Press 2000

D. S. Sivia and J. Skilling
*Data analysis: A Bayesian tutorial*
Oxford Univ. Press 2006

H. J. C. Berendsen
*A student’s guide to data and error analysis*
Cambridge Univ. Press 2011

P. Nelson
*Physical models of living systems*
WH Freeman and Co TBA
Thanks

University of Pennsylvania

NSF DMR, BIO, IBN

NSF NSEC

National Eye Institute Training grant

Computational Neuroscience Training grant

For these slides see: www.physics.upenn.edu/~pcn

Sunday, December 16, 2012