MODELS AND FITTING

Nicholas Fitzkee, CH 4404

So far in the class, we've developed several physical models for understanding chemical phenomena. We've looked at binding models with and without cooperativity. We've examined models for describing how helices form out of random structure. Most recently, we've discussed models for describing the kinetic behavior of chemical reactions. All of these models have parameters; that is numerical constants that tell us something about the physical behavior. For binding, the parameters are the equilibrium constants (and τ) that can be used to derive free energies. Helix-coil theory used s and σ to describe complex folding behavior. And kinetic rate laws all depend on the rate constants involved. In each of these systems, the parameters can tell us something about reality (at least, that's the goal).

What I think is missing from the text is motivation for why you would want to do any of this. What is the relationship between the models that we construct and reality? Why bother? In this brief handout, I hope to convince you that there is indeed a point to all of this.

In physical chemistry, we are often concerned with distilling a complex system down to its essential parts. Thus, while real-life systems can have very sophisticated behavior, it is often possible to describe these systems with a relatively small set of parameters. As examples of parameters consider the following examples:

- The path of a thrown projectile can be described by the acceleration and drag coefficient.
- The bond distance between two atoms can be described by their covalent radii.
- The spontaneity of a chemical reaction can be described by the equilibrium constant.

In each case, a few simple parameters can describe the behavior of the system. Generally, the values of the parameters are not known and must be determined experimentally, but once the parameters have been determined they are generally applicable: The gravitational acceleration is the same whether we're in Colorado or Columbia, Mississippi or Morocco.

How do we know what models to use? That's actually a deep question, and it gets to the heart of science itself. As we start studying a system, we generally do *not* know what model applies. It is up to the scientist to examine the behavior of a system and develop a model that works. As a general rule, simple models that can completely describe the behavior of a system are preferable to complicated models that describe the same behavior. Ultimately, however, scientists need to test their models to make sure they agree with reality. The Bohr model of the atom was sufficient for describing the energy levels of the hydrogen atom, but ultimately his model was still wrong (even if it was pleasingly simple).

As we studied binding, we developed two models that could be used to describe a system of multiple binding sites. Each model has a set of primary postulates (or hypotheses):

Multiple Independent Sites

- Each ligand binds with equilibrium constant *K*
- All subsequent binding events have the same equilibrium constant

Linear Ising Model

- Each ligand initially binds with equilibrium constant *K*
- When a neighboring site is bound, equilibrium is constant τK

Once a scientist has developed a set of hypotheses for a model, he or she must try to identify the consequences of those hypotheses. In our case, we need to relate the postulates above to an observable quantity, the degree of binding (ν). A significant portion of class time this week was devoted to relating the hypotheses above to a mathematical relationship for ν. We found that the degree of binding is a function of [L], but it was also dependent on the parameters we defined: K and τ.

As mentioned in class, it is possible (using equilibrium dialysis) to determine ν as a function of the free ligand concentration [L]. The ligand concentration [L] is an independent variable, because we can control the amount of total ligand we add to the system. As you might expect, adding more total ligand will also increase the amount of unbound ligand [L]. The degree of binding ν is a dependent variable, because we only control it indirectly. Nevertheless, ν is experimentally obtainable. The key point is this: *If we can find a model that describes the observed behavior of* ν *vs.* [L] *for a given set of parameters, we have identified an instance where our model has experimental support.*

How do we do this? The first step involves choosing the right model. Good scientists can do this from experience: they'll look at the data and determine which existing model, if any, can be used to describe the experimental trends. In our case, if we were deciding between the two models above, we could look for signs of a sigmoidal binding curve. Unfortunately, even the best scientists sometimes choose the wrong model, but that's what makes science interesting.

The next step in model selection is trying to find the right parameters within a model that describe the data. In this case we "fit" the model to the data (never the other way around!). In this assignment, we will use Excel to practice finding the best values for a kinetic model. Our approach is entirely manual: you will "tweak" parameters to optimize the agreement between your model and the observed data. The Excel spreadsheet approach is one way to get parameter values assuming a particular model; another way is to use more sophisticated fitting tools like Origin or MatLab (expensive), or GNUplot or R (free). These tools systematically vary the parameter values to try to find the best fit, which is not often an easy task. Given the time constraints in our class, we will not be able to discuss different techniques for fitting parameters (or how confident we are that those parameters are correct). If you are interested, however, a good resource would be Bevington and Robinson, *Data Reduction and Error Analysis for the Physical Sciences.*

How do we assess whether our model is a good fit? For us, we will consider the *residuals* and the *sum of the squared residuals*. There are more sophisticated approaches as well

– for example, some of you may be familiar with reduced χ^2 (chi-squared). A residual is the difference between one particular observation and the same observation calculated by our given model. When we perform a binding experiment, we have a set of N pairs: (ν_1, [L_1]), (ν_2, [L_2]), … (ν_N, [L_N]), . These are our experimental data. Using our model, we can also *calculate* a predicted degree of binding for each [L_i]. Let's call this V([L_i]; K, τ). V is a function of [L_i], but it will also change as we vary K (and τ if applicable). Using this nomenclature, the residual for data point *i* would be:

$$r_i = \left(\nu_i - V([L_i]; K, \tau)\right)$$

Again, the above equation is simply a way of saying that the residual for the i^{th} data point is the difference between the observed value of ν and the value of ν that we would calculate at the same [L] using our model.

If our model captures the important elements of what's happening in the system, and if the parameters τ and K are correctly chosen, the residuals should be small. That is to say, the difference between our model's predictions and the observed data points should be small. Of course, our experiments have a certain degree of uncertainty and error in them, so it's highly unlikely that our residuals will be zero for every data point. Nevertheless, we can use the residuals to help us in finding the best values for the parameters once we have chosen our model.

By varying the parameters (K, τ), we (or Excel, or Matlab, etc.) can minimize the *sum of squared residuals (SSR)*. If we have N experimental data points, this sum is simply defined as:

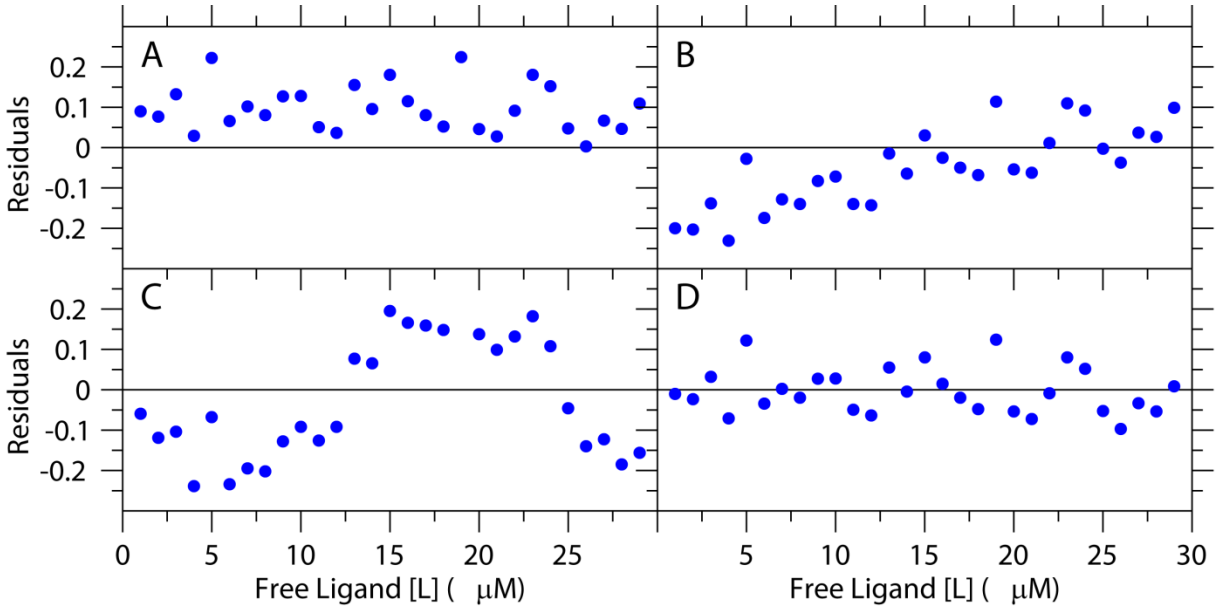
$$SSR = \sum_{i=1}^{N}\left(\nu_i - V([L_i]; K, \tau)\right)^2$$

In general, residuals can be either positive or negative, but the *squared* residual will always be positive. Thus, summing the squared residuals gives us something to minimize. If all of the residuals are small, the SSR should also should be small. If we find the minimum of the SSR by varying K and τ, we know that the values of K and τ at the minimum are the best possible values. *We also know that the agreement between our model and the data is as good as it can possibly get.*

Once we've obtained "best fit" values for our parameters, we are finally in a position to assess the model itself. Under the best possible circumstances, how well does our model describe the data? Obviously, if the best fit parameters aren't physical (e.g. a negative value for K), we know that something isn't right. There are many statistical tests which serve to describe how well a model fits a data set, but these are beyond the scope of this course. Instead, we will focus on looking at the residuals themselves.

We have already mentioned that, at the best fit values for K and τ, the residuals should be close to zero. But if the model is truly describing the data, *we should not observe any systematic trends in the residuals, either*. We are looking for a situation where the residuals are small and randomly distributed about zero. As a means of visualizing this, consider the four sets of

residuals on the next page. Remember that the residuals represent the *difference* the raw data and our best model.



Each of these examples represents a different model fit to the same data. The specific models that were used are not important here; instead, we want to assess whether the model is accurately describing our experimental data (regardless of the specific model used). In all four cases, the residuals are telling us how the model fits the data. For example, in A, we see that all the residuals are greater than zero: our model *systematically* underpredicts the data. In B, we observe a linear trend: for low Ligand concentrations the binding is overpredicted, but the model underpredicts binding at higher concentrations. The residuals in C also indicate a poor model: even at the "best fit" parameters, there is still a periodic variation in the residuals. D is the best model, because the residuals are small and there does not appear to be a trend as [L] is varied. Obviously it's good to know the uncertainty of each datapoint, and this can be helpful as you assess what is "close" to zero. However, to simplify the discussion above, error bars were not included on the data points.

Selecting a good model to describe the data is rarely an easy task, and even when the residuals look great it is possible that a key element is missing in the model. This is why it is so important to test and retest the model under different conditions. The best models are predictive and hold up under many different experimental scenarios, and these are the models that scientists tend to hold on to. However, all is not lost when a model doesn't fit well; in fact, *it could be telling you something very interesting about the system you are studying*. Observing a discrepancy between the model and the data provides an opportunity for revising the model and learning something deeper about the physical world.