BMC Systems Biology

Methodology article

Equilibrium model selection: dTTP induced R1 dimerization

Tomas Radivoyevitch

Address: Department of Epidemiology and Biostatistics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA

Email: Tomas Radivoyevitch - txr24@case.edu

Published: 4 February 2008

BMC Systems Biology 2008, **2**:15 doi:10.1186/1752-0509-2-15

Received: 27 July 2007

Accepted: 4 February 2008

This article is available from: http://www.biomedcentral.com/1752-0509/2/15

Abstract

Background: Biochemical equilibria are usually modeled iteratively: given one or a few fitted models, if there is a lack of fit or over fitting, a new model with additional or fewer parameters is then fitted, and the process is repeated. The problem with this approach is that different analysts can propose and select different models and thus extract different binding parameter estimates from the same data. An alternative is to first generate a comprehensive standardized list of plausible models, and to then fit them exhaustively, or semi-exhaustively.

Results: A framework is presented in which equilibriums are modeled as pairs (g, h) where g = 0 maps total reactant concentrations (system inputs) into free reactant concentrations (system states) which h then maps into expected values of measurements (system outputs). By letting dissociation constants K_d be either freely estimated, infinity, zero, or equal to other K_d, and by letting undamaged protein fractions be either freely estimated or 1, many g models are formed. A standard space of g models for ligand-induced protein dimerization equilibria is given. Coupled to an h model, the resulting (g, h) were fitted to dTTP induced R1 dimerization data (R1 is the large subunit of ribonucleotide reductase). Models with the fewest parameters were fitted first. Thereafter, upon fitting a batch, the next batch of models (with one more parameter) was fitted only if the current batch yielded a model that was better (based on the Akaike Information Criterion) than the best model in the previous batch (with one less parameter). Within batches models were fitted in parallel. This semi-exhaustive approach yielded the same best models as an exhaustive model space fit, but in approximately one-fifth the time.

Conclusion: Comprehensive model space based biochemical equilibrium model selection methods are realizable. Their significance to systems biology as mappings of data into mathematical models warrants their development.

Background

Ribonucleotide reductase (RNR) has a small subunit R2 that exists almost exclusively as a dimer, and a large subunit R1 that dimerizes when dTTP, dGTP, dATP, or ATP binds to its specificity site, and hexamerizes when dATP or ATP binds to its activity site [1-6]. Thus, R1 is the backbone of a biochemical equilibrium network that contains a large number of R1 complexes. This network has more dissociation constants (K_d) than can be estimated from currently available data, so assumptions must be made to reduce the number of independent K_d. These assumptions come in two forms: those that state that for the data at hand, a K_d is too large or small to be distinguished from infinity or zero, respectively, and those that state that the

data are too weak to rule out a null hypothesis of the form $K_d = K'_d$. Model parameters such as the fraction of R1 capable of forming dimers and hexamers, and the enzymatic activities of these R1 states, also come with plausible null hypotheses. In general, different null hypotheses define different models that yield different estimates of the freely estimated parameters. Unfortunately, as modelers traverse a path of reasonable hypotheses until they arrive at a model that provides both a good fit and K_d confidence interval limits that are not too wide, they often stop at different places, and thus report different K_d values. Such K_d estimate extraction differences could be reduced, if a systematic reproducible approach to biochemical equilibria model building was established. Progress toward this goal is described in this paper.

Results

Model

Consider a dataset comprised of N steady state non-covalent binding equilibriums indexed by n in which J different complexes can potentially form from a protein R of known total concentration T_{n1} through interactions with itself and I - 1 other reactants (e.g. substrate, effectors and other proteins) of known total concentrations T_{ni} ($1 < i \leq I$). Suppose W_{ij} copies of the ith reactant exist in the jth complex and that a particular R molecule is either undamaged with probability p, and thus capable of forming each of the plausible complexes, or damaged with probability 1 - p, and thus incapable of forming any complexes. Define $T_n = (T_{n1}, T_{n2}, \ldots T_{nI})$, $F_n = (F_{n1}, F_{n2}, \ldots F_{nI})$ as the corresponding free reactant concentrations, $K = (K_1, K_2, \ldots K_J)$ as the dissociation constants (of complexes to free reactants), γ_n as the measurement(s) made at the nth steady state, and $Z_n = (Z_{n1}, Z_{n2}, \ldots Z_{nJ})$ as the concentrations of complexes predicted by W, K and F_n to be

$$Z_{nj} = \frac{\prod_{i'=1}^{I} F_{ni'}^{W_{i'j}}}{K_j}. \quad (1)$$

The relationship between the system inputs (T_n), states (F_n) and outputs (γ_n) is then modeled by I total concentration constraints

$$g(F_n, T_n, K, p) = 0$$

that must be solved for the I free reactant concentrations F_n at each n ($1 < n \leq N$) given the inputs T_n, and an output measurement model h that connects F_n to expected values of the outputs $E(\gamma_n)$

$$\gamma_n = h(F_n, K, p, L) + \varepsilon_n$$

where all of the h specific parameters (e.g. k_{cat}'s and protein masses) are contained in the vector L and, if the γ_n are vectors of measurements, the e_n are vectors of zero mean noise, potentially correlated within steady states, but uncorrelated between steady states; only scalar γ_n are considered hereafter. The model parameters K, p and L are not indexed by n because they are fitted jointly to the entire dataset, i.e. one set of estimates of these parameters describes all N steady states simultaneously as one (g, h) model of one underlying biochemical equilibrium network.

System models

The I equations of a system model $g = 0$ are

$$\begin{aligned} g_1(F_n, T_n, K, p) &= pT_{n1} - F_{n1} - \sum_{j=1}^{J} W_{1j} \frac{\prod_{i'=1}^{I} F_{ni'}^{W_{i'j}}}{K_j} = 0 \\ g_i(F_n, T_n, K) &= T_{ni} - F_{ni} - \sum_{j=1}^{J} W_{ij} \frac{\prod_{i'=1}^{I} F_{ni'}^{W_{i'j}}}{K_j} = 0 \quad (1 < i \leq I) \end{aligned} \quad (2)$$

where pT_{n1} is the total concentration of undamaged R and F_{n1} is the concentration of free R that is undamaged and thus capable of forming complexes. If all biologically plausible candidate complexes are present in these equations, the model will have as many K parameters as possible, and it will therefore be called a full model. A space of $g = 0$ models can then be generated from this full model through combinations of null hypothesis constraints on the parameters in (K, p).

Fitting a particular (g, h) to data (T, γ) to estimate parameters in (K, p, L) demands many repeated solutions of $g = 0$. These equations must be solved efficiently to fit large model spaces and models with large numbers of parameters. The approach proposed here solves $g = 0$ by letting g be the right hand side of a parent set of ordinary differential equations (ODEs) that achieves $g = 0$ at steady state. Specifically, the following ODEs were simulated to large T to solve the polynomial system in Eqs. (2):

$$\begin{aligned} \frac{dF_{n1}}{dt} &= pT_{n1} - F_{n1} - \sum_{j=1}^{J} W_{1j} \frac{\prod_{i'=1}^{I} F_{ni'}^{W_{i'j}}}{K_j} \\ \frac{dF_{ni}}{dt} &= T_{ni} - F_{ni} - \sum_{j=1}^{J} W_{ij} \frac{\prod_{i'=1}^{I} F_{ni'}^{W_{i'j}}}{K_j} \end{aligned} \quad (3)$$

where $1 < i \leq I$, $n = 1...N$ and $F_{ni}(0) = 0$. Note that the initial conditions guarantee that the system derivatives are initially positive and thus that the system always starts in an acceptable direction; model parameters are constrained to positive values, expressed internally as e^c, where c is unconstrained during optimization.

The system of polynomials in Eqs. (2) has been solved by others using other approaches. In one approach, the F_{ni} terms are pulled to the left hand side and guesses are then iteratively entered into the right hand side until the equations become self consistent [7]. This approach has more recently been shown to fail in cases of oligomerization, and modifications of the approach have been suggested [8]. The difficulties of solving systems of arbitrary nonlinear algebraic equations in general have been described [9] and a common approach (e.g. used by fsolve in Matlab) has been to minimize the sum of squares g^2 using Levenberg-Marquadt or Gauss-Newton methods. Intuitively, methods that exploit the fact that the equations are strictly polynomials should outperform these general methods. Continuation homotopy is one such method [10]. In this method, polynomials are homogenized to a larger polynomial system with known solutions, and these solutions are then traced to the desired solutions as the homogenized polynomials are continuously morphed back to the original polynomial system. On a practical level, all complex initial solutions must be tracked to find the desired final solution that is strictly real and positive, and this makes the approach slower than the R [11] implementation of Eqs. (3) provided here, which finds only the positive real root and does so rapidly because it automatically generates and compiles C code (of Eqs. 3) that is then used with the dll/so option of the ODE solver lsoda available in R [11]. To glean some insight into why Eq. (3) works, note that the g_i (i.e. right hand sides) are all initially positive, and all monotonically decreasing functions of increasing free concentrations. Free concentration differentials thus start positive and shrink toward zero as the free concentrations move out of their initial values at the origin and into the positive quadrant. When a component F_{ni} of the vector F_n crosses its steady state value, the corresponding g_i switches signs, since the g_i continue to decrease monotonically through zero, and F_{ni} is then thus driven back toward a smaller value, i.e. back toward the steady state value that it just crossed. This explains why the proposed algorithm is stable. Finally, an alternative approach to the problem is to solve $g = 0$ using full-blown kinetic equations with irrelevant time scales defined by $k_{on} = 1$ and $k_{off} = K_d$, but the number of ODEs then equals the number of complex species plus the number of reactants, rather than just the number of reactants as in Eqs. 3, and although each ODE is computationally simpler in this case, the savings per ODE do not offset the added cost of the additional ODEs. This added cost is expected to become substantial if not prohibitive in combinatorially complex scenarios wherein the number of complexes is very large relative to the number of reactants.

K *hypotheses*

In the $g = 0$ model in Eqs. (2), the elements of K are defined as

$$K_j = \frac{\prod_{i'=1}^{I} F_{ni'}^{W_{i'j}}}{Z_{nj}}. \quad (4)$$

This definition can differ by stoichiometric factors from K_d defined as k_{off}/k_{on}. For example, consider a system where R can bind a ligand t and R can also form dimers. Figure 1

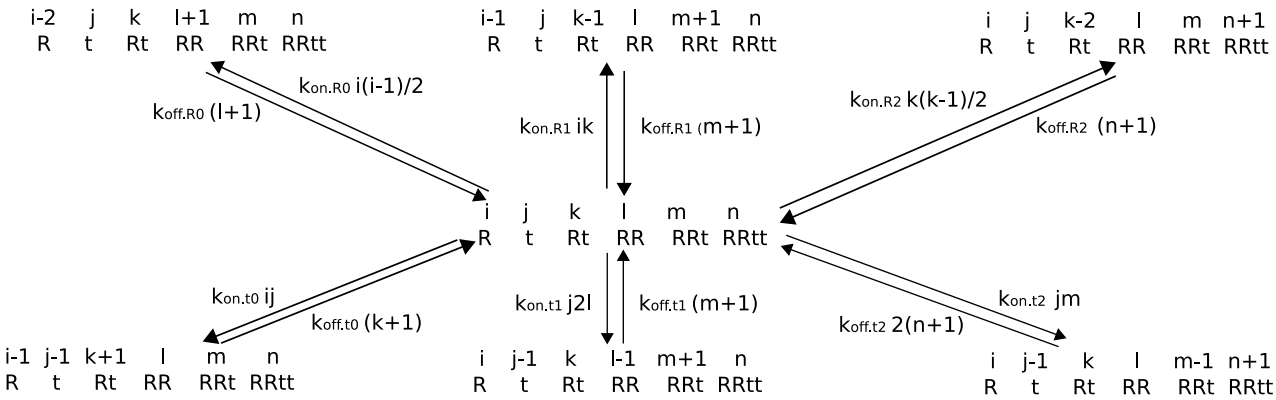

Figure 1

Rt system state transition diagram. The next states of a unit volume reaction vessel that currently has (*i*, *j*, *k*, *l*, *m*, *n*) molecules of (R, t, Rt, RR, RRt, RRtt) are shown. The k_{on}'s in this diagram are the rates at which potential interactions successfully materialize, and the k_{off}'s are the per-site rates at which ligands dissociate.

shows the state transitions of this system from a state of *i*, *j*, *k*, *l*, *m* and *n* molecules of R, t, Rt, RR, RRt and RRtt, respectively, per unit volume, where the unit volume is small enough that any reactant can react equally well with any other reactant, yet large enough that these integers are approximately equal to themselves plus or minus one or two. If net fluxes between states are zero, the system is in equilibrium and the following definitions of $K_d \equiv k_{off}/k_{on}$ arise

$$\begin{aligned} k_{on.R0} i(i-1)/2 &= k_{off.R0}(l+1) \quad \Rightarrow \\ K_{d\_R\_R} &\equiv k_{off.R0}/k_{on.R0} \\ &= \frac{i(i-1)}{2(l+1)} \approx \frac{[R][R]}{2[RR]} \end{aligned} \tag{5}$$

$$\begin{aligned} k_{on.R1} ik &= k_{off.R1}(m+1) \quad \Rightarrow \\ K_{d\_Rt\_R} &\equiv k_{off.R1}/k_{on.R1} \\ &= \frac{ik}{(m+1)} \approx \frac{[Rt][R]}{[RRt]} \end{aligned} \tag{6}$$

$$\begin{aligned} k_{on.R2} k(k-1)/2 &= k_{off.R2}(n+1) \quad \Rightarrow \\ K_{d\_Rt\_Rt} &\equiv k_{off.R2}/k_{on.R2} \\ &= \frac{k(k-1)}{2(n+1)} \approx \frac{[Rt][Rt]}{2[RRtt]} \end{aligned} \tag{7}$$

$$\begin{aligned} k_{on.t0} ij &= k_{off.t0}(k+1) \quad \Rightarrow \\ K_{d\_R\_t} &\equiv k_{off.t0}/k_{on.t0} \\ &= \frac{ij}{k+1} \approx \frac{[R][t]}{[Rt]} \end{aligned} \tag{8}$$

$$\begin{aligned} k_{on.t1} j2l &= k_{off.t1}(m+1) \quad \Rightarrow \\ K_{d\_RR\_t} &\equiv k_{off.t1}/k_{on.t1} \\ &= \frac{j2l}{m+1} \approx \frac{2[RR][t]}{[RRt]} \end{aligned} \tag{9}$$

$$\begin{aligned} k_{on.t2} jm &= k_{off.t2} 2(n+1) \quad \Rightarrow \\ K_{d\_RRt\_t} &\equiv k_{off.t2}/k_{on.t2} \\ &= \frac{jm}{2(n+1)} \approx \frac{[RRt][t]}{2[RRtt]}. \end{aligned} \tag{10}$$

In Eqs. 5 and 7, $x(x - 1)/2$ is the number of unique binary interactions of x molecules with themselves. The stoichiometric factor in Eq. (9) arises because RR has twice as many ways to gain a t as RRt has ways to lose a t, and in Eq. 10 it arises because RRtt has twice as many ways to lose a t as RRt has ways to gain a t. Eqs. 9 and 10 assume that RR and RRtt are symmetric dimers.

Regarding differences between the K_d in Eqs. (5–10) and the K_j in Eq. (4), the K_d always have units of concentration because they always correspond to two molecules binding together at one time, and the K_j have units of concentrations raised to integer powers $\sum_{i'=1}^{I} W_{i'j} - 1$ that can be greater than 1 (in such cases the K_j represent several sequential binding steps condensed into one, e.g. see Table 1). In general, the K_d are associated with grid-shaped equilibrium network graphs such as those shown in Figure 2 and the K_j are associated with spur-shaped equilibrium graphs such as those shown in Figure 3. Notationally, subscripts of the K_j will be distinguishably devoid of *d*'s and underscores, e.g. $K_{RRtt} = \frac{[R]^2[t]^2}{[RRtt]}$ is the K_j of graph M in Figure 3.

In the graphs shown in Figure 2, it is plausible to conjecture a priori that any two or all three of Kd_R_R, Kd_Rt_R and Kd_Rt_Rt are equal, i.e. that the binding of t to R has no impact on R binding to itself. Similarly, it is plausible that any two or all three of Kd_R_t, Kd_RR_t and Kd_RRt_t are equal. These two sets of hypotheses are not independent, since Kd products of two paths between the same two nodes must be equal. For example, in Figure 2A, starting with free reactants, the two paths to RRt are

$$\begin{aligned} K_{d\_R\_R\_t} &= K_{d\_R\_R} K_{d\_RR\_t} \\ &= \frac{[R][R]}{2[RR]} \frac{2[RR][t]}{[RRt]} = \frac{[R]^2[t]}{[RRt]} \\ K_{d\_R\_R\_t} &= K_{d\_R\_t} K_{d\_Rt\_R} \\ &= \frac{[R][t]}{[Rt]} \frac{[Rt][R]}{[RRt]} = \frac{[R]^2[t]}{[RRt]} \end{aligned} \tag{11}$$

and the two paths to RRtt are

$$\begin{aligned} K_{d\_R\_R\_t\_t} &= K_{d\_R\_R} K_{d\_RR\_t} K_{d\_RRt\_t} \\ &= \frac{[R][R]}{2[RR]} \frac{2[RR][t]}{[RRt]} \frac{[RRt][t]}{2[RRtt]} \\ &= \frac{[R]^2[t]^2}{2[RRt]} \\ K_{d\_R\_R\_t\_t} &= K_{d\_R\_t}^2 K_{d\_Rt\_Rt} \\ &= \frac{[R][t]}{[Rt]} \frac{[R][t]}{[Rt]} \frac{[Rt][Rt]}{2[RRtt]} \\ &= \frac{[R]^2[t]}{2[RRt]}. \end{aligned} \tag{12}$$

Table 1: K_j assignment model definitions

graph	K_{Rt}	K_{RR}	K_{RRt}	K_{RRtt}
2A	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	$K_{d\_R\_t}K_{d\_R\_R}$	$2K_{d\_R\_t}^2 K_{d\_R\_R}$
2B	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	$K_{d\_R\_t}K_{d\_R\_R}$	$2K_{d\_R\_t}^2 K_{d\_Rt\_Rt}$
2C	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	$K_{d\_R\_t}K_{d\_Rt\_Rt}$	$2K_{d\_R\_t}^2 K_{d\_Rt\_Rt}$
2D	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	$K_{d\_R\_t}K_{d\_Rt\_R}$	$2K_{d\_R\_t}^2 K_{d\_R\_R}$
2E	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	$K_{d\_R\_R}K_{d\_RR\_t}$	$2K_{d\_R\_R}K_{d\_RR\_t}^2$
3A, 2F	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	$K_{d\_R\_t}K_{d\_Rt\_R}$	$2K_{d\_R\_t}^2 K_{d\_Rt\_Rt}$
3A, 2F	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	$K_{d\_R\_R}K_{d\_RR\_t}$	$2K_{d\_R\_R}K_{d\_RR\_t}K_{d\_RRt\_t}$
2G	$K_{d\_R\_t}$	∞	$K_{d\_R\_t}K_{d\_Rt\_R}$	$2K_{d\_R\_t}^2 K_{d\_Rt\_R}$
3B, 2H	$K_{d\_R\_t}$	∞	$K_{d\_R\_t}K_{d\_Rt\_R}$	$2K_{d\_R\_t}^2 K_{d\_Rt\_Rt}$
2I	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	∞	$2K_{d\_R\_t}^2 K_{d\_R\_R}$
3C, 2J	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	∞	$2K_{d\_R\_t}^2 K_{d\_Rt\_Rt}$
2K	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	$K_{d\_R\_t}K_{d\_R\_R}$	∞
3D, 2L	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	$K_{d\_R\_t}K_{d\_Rt\_R}$	∞
2M	∞	$2K_{d\_R\_R}$	$K_{d\_R\_R}K_{d\_RR\_t}$	$2K_{d\_R\_R}K_{d\_RR\_t}^2$
3E, 2N	∞	$2K_{d\_R\_R}$	$K_{d\_R\_R}K_{d\_RR\_t}$	$2K_{d\_R\_R}K_{d\_RR\_t}K_{d\_RRt\_t}$
3F	$K_{d\_R\_t}$	∞	∞	$2K_{d\_R\_t}^2 K_{d\_Rt\_Rt}$
3G	$K_{d\_R\_t}$	∞	$K_{d\_R\_t}K_{d\_Rt\_R}$	∞
3H	$K_{d\_R\_t}$	$2K_{d\_R\_R}$	∞	∞
3I*	∞	∞	$K_{d\_R\_t}K_{d\_Rt\_R}$	$2K_{d\_R\_t}^2 K_{d\_Rt\_Rt}$
3J*	∞	$2K_{d\_R\_R}$	∞	$2K_{d\_R\_t}^2 K_{d\_Rt\_Rt}$
3K*	∞	$2K_{d\_R\_R}$	$K_{d\_R\_t}K_{d\_Rt\_R}$	∞
3L	$K_{d\_R\_t}$	∞	∞	∞
3M*	∞	∞	∞	$2K_{d\_R\_t}^2 K_{d\_Rt\_Rt}$
3N*	∞	∞	$K_{d\_R\_t}K_{d\_Rt\_R}$	∞
3O	∞	$2K_{d\_R\_R}$	∞	∞
3P	∞	∞	∞	∞
3Q	0	∞	∞	∞
3R	∞	∞	∞	0
3S	∞	∞	0	∞
3T	∞	0	∞	∞

*$K_{d\_R\_t}$ is too large to estimate in these cases, but its products with small numbers, viewed as single K_j parameters, might still be estimable.

Similarly, the two paths from the node [Rt R t] to RRtt yield

$$\begin{aligned} K_{d\_Rt\_R}K_{d\_RRt\_t} &= \frac{[Rt][R]}{[RRt]}\frac{[RRt][t]}{2[RRtt]} \\ &= \frac{[Rt][R][t]}{2[RRt]} \\ K_{d\_R\_t}K_{d\_Rt\_Rt} &= \frac{[R][t]}{[Rt]}\frac{[Rt][Rt]}{2[RRtt]} \\ &= \frac{[Rt][R][t]}{2[RRt]}, \end{aligned} \tag{13}$$

though these could have been obtained from (11) and (12). Based on Eqs. (11), either of $K_{d\_R\_t} = K_{d\_RR\_t}$ and $K_{d\_R\_R} = K_{d\_Rt\_R}$ implies the other, and based on Eqs. (13), either of $K_{d\_R\_t} = K_{d\_RRt\_t}$ and $K_{d\_Rt\_R} = K_{d\_Rt\_Rt}$ implies the other. Such constraints yield the K_d equality hypotheses shown in Fig. 2. This space of K_d equality models was generated from the fully constrained Model A by releasing pairs of R binding equality constraints and counterpart t binding constraints one at a time. When two R binding constraints are released, all three R binding constants become independent, and this leaves only one permissible t-binding constraint (Model E) or none (Model F). Models with one node less (G to N) are then considered; the two Rt nodes act as one. Models with two or more nodes removed do not allow K_d equality constraints and in these cases, K_j defined by Eq. 4 are adequate; such models are shown in Figure 3.

The Rt system full model special case of $g = 0$ in Eqs. (2), with $T_n = ([R_T], [t_T])$, $F_n = ([R], [t])$, $Z_n = ([Rt], [RR], [RRt], [RRtt])$, and thus

$$W = \begin{pmatrix} 1 & 1 & 2 & 2 \\ 1 & 0 & 1 & 2 \end{pmatrix}, \tag{14}$$

is

$$\begin{aligned} 0 &= p[R_T] - [R] - \frac{[R][t]}{K_{Rt}} - 2\frac{[R]^2}{K_{RR}} \\ &\quad -2\frac{[R]^2[t]}{K_{RRt}} - 2\frac{[R]^2[t]^2}{K_{RRtt}} \\ 0 &= [t_T] - [t] - \frac{[R][t]}{K_{Rt}} \\ &\quad -\frac{[R]^2[t]}{K_{RRt}} - 2\frac{[R]^2[t]^2}{K_{RRtt}}. \end{aligned} \tag{15}$$

These $g = 0$ equations correspond to graph A in Figure 3. As $K_j = \infty$ assumptions are applied to these equations to remove specific terms one at time, two at a time, and so on, corresponding nodes are removed from graph A to

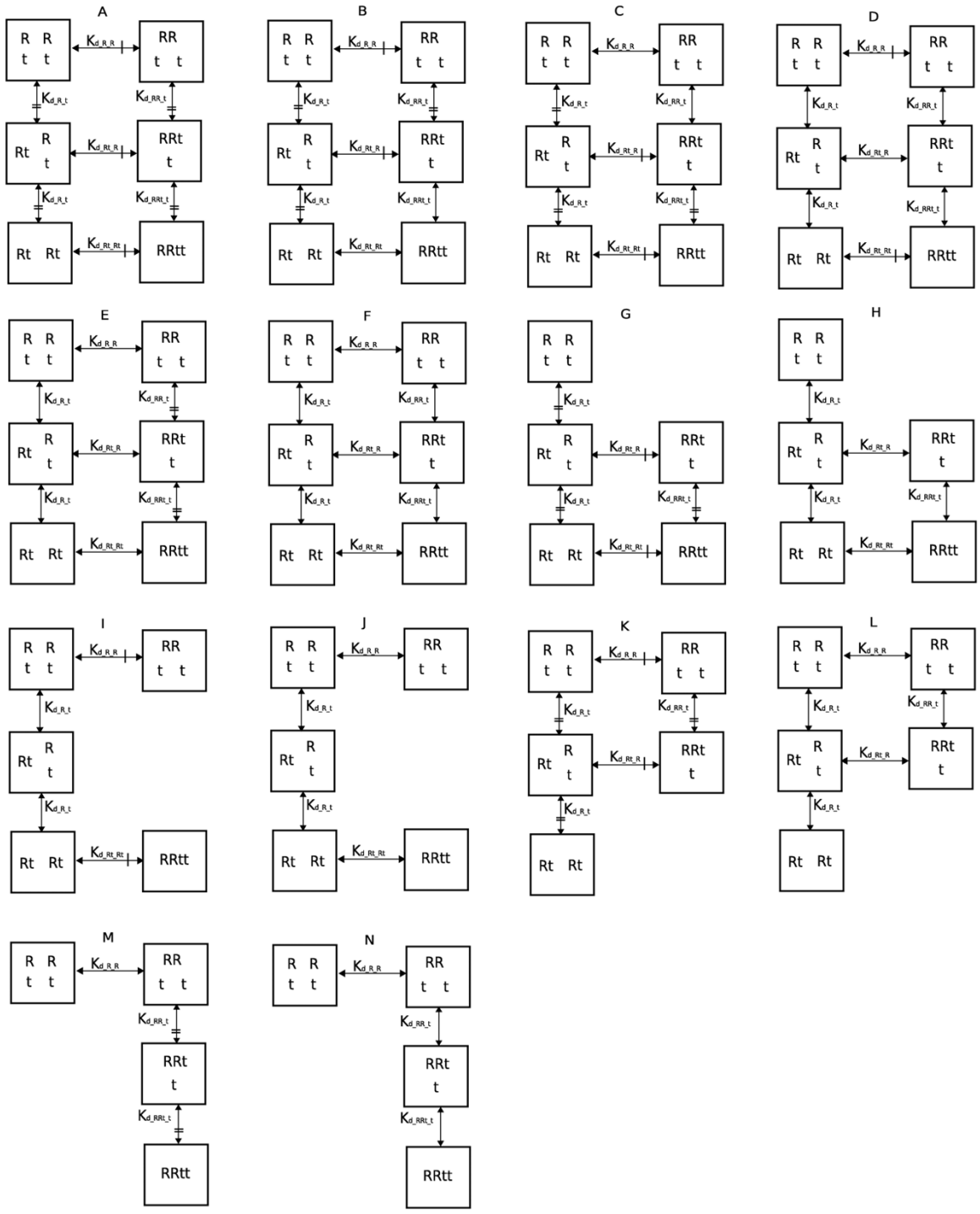

Figure 2

Space of K_d equivalence grid graph models. In these $K_d = K'_d$ grid graphs t dimension edges marked = are equal and R dimension edges marked | are equal, i.e. Model A is fully constrained. Models F, H, J, L and N have zero K_d equivalence constraints and are thus equal to Models A-E in Figure 3.

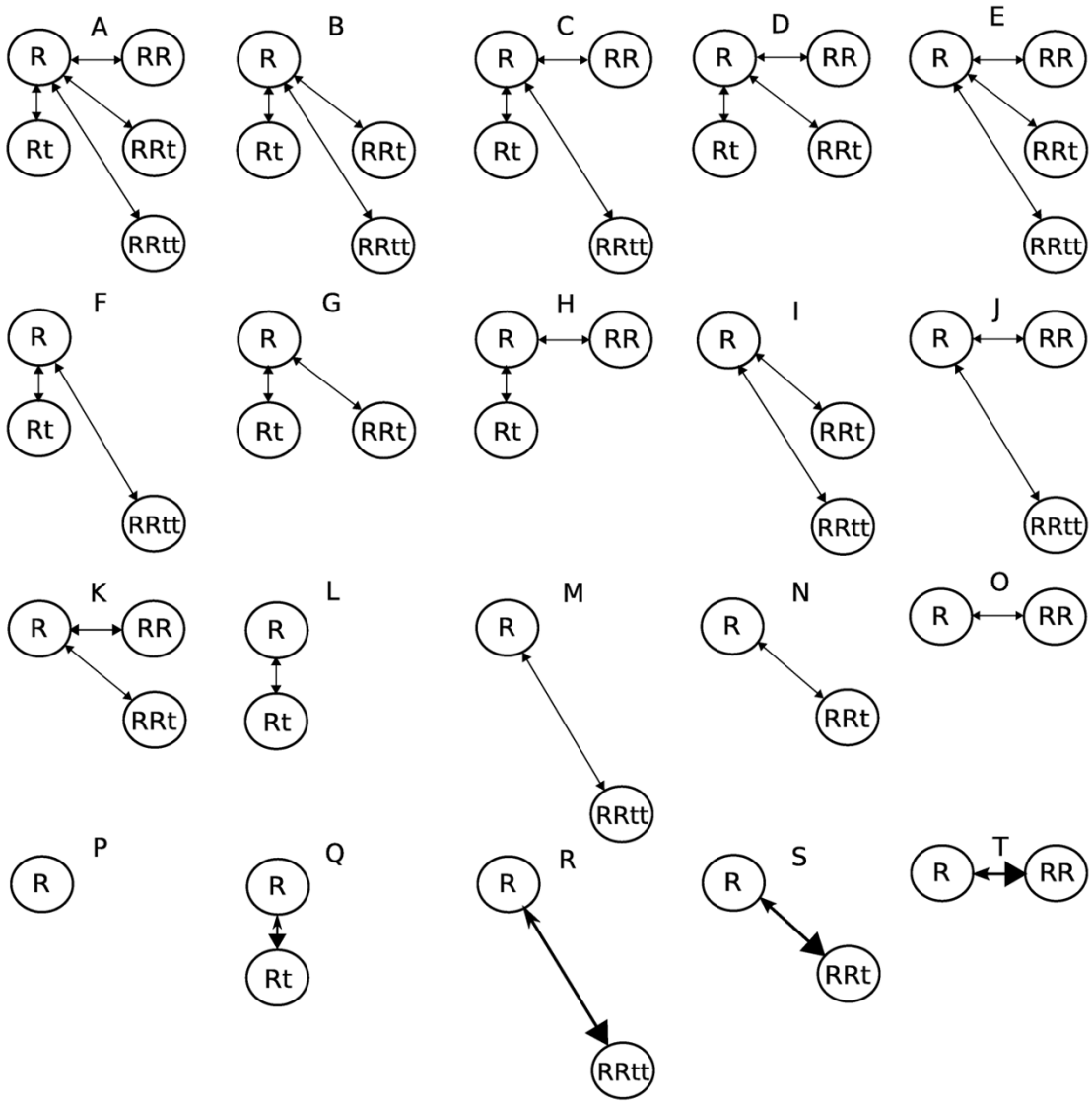

Figure 3
Space of $K_j = \infty$ or 0 spur graph models. The full spur graph in A spawns this *g* space of system models. Models Q to T correspond to infinitely tight binding. Models I, J, M, N, R and S cannot be represented by grid graphs.

create graphs B to P and thus models that conjecture that the deleted nodes/complexes are not detectable above noise. Of these models, the *J* single edge models (L to O) can have additional $K_j = 0$ assumptions applied to them to generate *J* additional *g* models (Q to T), each alleging that the free concentration of the reactant that is not in excess (i.e. ligand or R) is indistinguishable from zero (i.e. at a level too low to be detected using the data at hand). In such models, $K_j = 0$ is handled either by approximating 0 by a small number (e.g. .0001; this option is readily automated, but pushing it too far causes numerical problems) or by replacing the equations with rules (e.g. if $K_{RRtt} = 0$ as in Model 3R, the rule would be: if $[R_T] < [t_T]$, $[R] = 0$ and $[RRtt] = [R_T]/2$, else $[R_T] \geq [t_T]$ and thus $[R] = [R_T] - [t_T]$ and $[RRtt] = [t_T]/2$; this option remains to be automated). In the end, a spur graph (e.g. 3A) with *J* edges generates 2^J models via $K_j = \infty$ assumptions and an additional *J* models via $K_j = 0$ assumptions, e.g. the $2^4 + 4 = 20$ models in Fig. 3. Considering that *J* is the number of complex species, which can be large, the number of *g* models generated can be huge.

The models in Figs. 2 and 3 are characterized by their assignments to the four K_j parameters in Eq. 15 as shown in Table 1. This table defines a standard space of *K* hypothesis *g* models for ligand induced protein dimeriza-

tion equilibria. As Models F, H, J, L and N in Fig. 2 do not have any K_d equality constraints, their data fitting capabilities are equal to those of Models A through E in Fig. 3, respectively. To see this, consider the first of the two rows labeled 3A,2F in Table 1. Eqs. (5) and (8) give $K_{RR} = 2K_{d\_R\_R}$ and $K_{Rt} = K_{d\_R\_t}$, Eq. (11) gives $K_{RRt} = K_{d\_R\_t}K_{d\_Rt\_R}$, which can be adjusted independently by the factor $K_{d\_Rt\_R}$, and Eq. (12) gives $K_{RRtt} = 2K^2_{d\_R\_t}K_{d\_Rt\_Rt}K_{d\_Rt\_Rt}$, which can be adjusted independently by $K_{d\_Rt\_Rt}$. Thus, all four of the K_j parameters of 3A can be independently manipulated to arbitrary values by the four K_d parameters of 2F, and in this sense, the two models are equivalent. A major difference, however, is that 2F can be represented in more than one way. Indeed, two choices are given by the two 3A,2F rows in Table 1, and all of the graphs in Figure 2 can be parameterized as subsets of either the E-shaped or -shaped parameterization topologies given in these two full model rows.

The nine grid graphs in Fig. 2 that contain at least one $K_d = K'_d$ constraint have $|K_j| > |K_d|$ where $|K_x|$ is the number of freely estimated K_x parameters. Meanwhile, models that are equally well represented by both grid and spur graphs are characterized by $|K_j| = |K_d|$, which, in Fig. 3, is all of the graphs except I, J, M, N, R and S. These exceptions must use spur graphs to avoid non-identifiability problems, have $|K_j| < |K_d|$, include complexes without including required intermediates, and have $K_d = \infty$ in product expressions that remain finite (see Table 1 footnote). Such models are palatable only because they represent statistical null hypotheses rather than physical null hypotheses, i.e. $K_d = \infty$ is a claim that the true value of K_d is too large to estimate based on the data at hand, and not a claim that binding never occurs.

p *hypotheses*

The probability that an R molecule is undamaged can be hypothesized to be close enough to 1 that the data cannot discriminate it from being 1. If B different protein preparation batches (indexed by b) are used in the experiments, 2^B hypotheses exist. $p_b = p_{b'}$ hypotheses that two batches are equivalent can also be formulated. In the equations given above and in the data analysis given below, $B=1$ is assumed.

Measurement models h

In pairs (g, h) the system of interest g is separated from the methods used to study it in h. h maps steady states F_n of g into expected values of measurements $E(\gamma_n)$. The first step in this, common to all h models, is to convert the F_n into complex concentration predictions Z_n using Eq. (1), i.e. using W and K. The second step is to form $E(\gamma_n)$ from F_n and Z_n and any other available information (e.g. L and p; note that T_n can be reconstructed from F_n and Z_n). This second step is different for different measurement types, as illustrated below for average protein mass, fraction of protein bound to a particular ligand, and average enzymatic activity of a distribution of enzyme states.

average mass

Suppose R is the only protein in the system, that ligand masses are too small to be detected relative to protein masses, and that average protein mass measurements are mass-weighted, e.g. as in dynamic light scattering data [1-3]. The second step of h for this type of measurement is then

$$E(\gamma_n) = M_1 \frac{[R]+[R_T](1-p_b)+\sum_{j=1}^{J} Z_{nj}W_{1j}^2}{[R_T]} \quad (16)$$

where M_1 is the mass of R monomer.

fraction bound

For fraction of protein bound to ligand data, suppose the ligand of interest is the ith reactant. The fraction of R bound to ligand is then

$$E(\gamma_n) = \left(\sum_{j=1}^{J} Z_{nj}W_{ij} \right) / [R_T]. \quad (17)$$

enzyme activity

If k_{catj} is the per-active-site enzymatic activity of the jth complex, the measured average activity of an ensemble of complexes is

$$E(\gamma_n) = \left(\sum_{j=1}^{J} k_{catj}Z_{nj}W_{1j} \right) / [R_T]. \quad (18)$$

It is assumed here that R provides all of the enzymatic activity and that it has only one active site.

h *space*

Enzyme activity differs from the other two measurement types in that its parameters can have many plausible null hypotheses: the k_{catj} can be equal to zero or to each other within groups defined in various ways. Thus, Eq. (18) can generate a space of h models. When such a space is multiplied into a g space, not all h models can be paired with any g, since, for example, if a K_j is infinity in a g model, the corresponding product complex concentration is zero, so

a corresponding k_{cat} cannot be estimated. Thus, although to first order $|(g, h)| = |g||h|$ where $|x|$ is the number of x models, this is actually an upper bound.

dTTP induced R1 dimerization data analysis

Let R be the R1 subunit of ribonucleotide reductase and let t be dTTP. Using h in Eq. (16), Scott et al [1] fitted Model 2E with $p = 1$ to their dynamic light scattering data shown in Figure 4. Their final parameter estimates are shown as the initial parameter estimates in Table 2. That these estimates did not converge properly (the authors used a method similar to that of Storer and Cornish-Bowden [7] to solve their $g = 0$ equations) is evidenced by the poor fit of the solid curve in Figure 4 relative to its fully converged counterpart computed here using the $g = 0$ solver described above (Eq. 3; dotted curve). The consequences of this poor fit are seen to be substantial in Table 2, where many of the K_d estimates have initial values that differ from their final converged counterparts by an order of magnitude. The final K_d estimates are, however, very uncertain, with upper-to-lower 95% confidence interval (CI, see Methods) limit ratios of $\sim 10^6$, i.e. Model 2E is overparameterized.

Given knowledge that R has a binding site for t and that R can dimerize [12], the model space in Table 1 doubled by p free or fixed to 1 and coupled to h in Eq. 16 creates 58 (g, h) candidate models that were fitted to these data. The fitted models were ranked by the Akaike Information Criterion (AIC, see Methods) and the best model was 3Rp (p

Table 2: Parameter estimates corresponding to Figure 4

Model	Parameter	Initial Value	Optimal Value	Confidence Interval
3Rp	pRT	1.000	0.767	(0.662,0.890)
	Rt	Inf	Inf	absent
	RR	Inf	Inf	absent
	RRt	Inf	Inf	absent
	RRtt	0.000	0.000	fixed
	SSE	0.100	0.027	
	AIC	-26.948	-36.058	
	cpu	0.000	0.057	fit succeeded
3M	RRtt	1.000	17.231	(3.190,93.691)
	Rt	Inf	Inf	absent
	RR	Inf	Inf	absent
	RRt	Inf	Inf	absent
	pRT	1.000	1.000	fixed
	SSE	0.059	0.032	
	AIC	-30.657	-34.969	
	cpu	0.000	0.291	fit succeeded
3Mp	RRtt	1.000	1.838	(0.010,347.234)
	pRT	1.000	0.837	(0.656,1.067)
	Rt	Inf	Inf	absent
	RR	Inf	Inf	absent
	RRt	Inf	Inf	absent
	SSE	0.059	0.023	
	AIC	-26.457	-32.981	
	cpu	0.000	0.109	fit succeeded
3N	RRt	1.000	16.699	(6.821,40.854)
	Rt	Inf	Inf	absent
	RR	Inf	Inf	absent
	RRtt	Inf	Inf	absent
	pRT	1.000	1.000	fixed
	SSE	0.176	0.045	
	AIC	-22.987	-32.602	
	cpu	0.000	0.125	fit succeeded
2E	R_t	25.000	2.265	(0.004,1164.445)
	R_R	75.000	1451.803	(0.089,24154952.754)
	RR_t	0.550	0.024	(0.000,22.421)
	RRt_t	0.550	0.024	constrained
	pRT	1.000	1.000	fixed
	SSE	0.042	0.027	
	AIC	-21.806	-24.990	
	cpu	0.000	0.264	fit succeeded

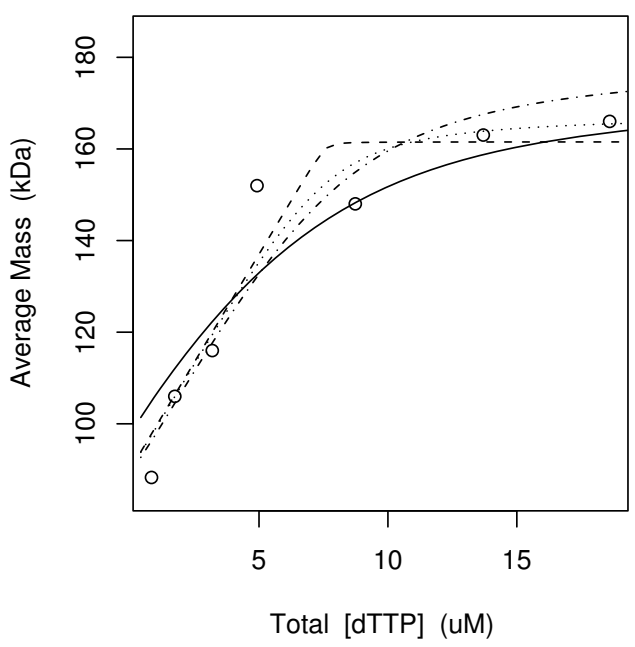

Figure 4

Scott et al. data. The parameter values of Scott et al. (Table 2, initial values of Model 2E) do not fit the data well (solid curve). The same model with fully converged parameter values does fit the data well (dotted). With *p* freely estimated, the infinitely tight binding Model 3Rp (dashed) has the lowest AIC. The second lowest AIC was achieved by Model 3M (dashed-dotted), see Table 2.

freely estimated) with K_{RRtt} = .0001 μM^3 essentially fixed to zero (dashed straight lines in Figure 4; Table 2). This model represents a tight binding titration limit wherein free molecule annihilation (the initial linear ramp in Fig. 4) continues in a one-to-one fashion with increasing [dTTP$_T$] until [dTTP$_T$] equals [R1$_T$] = 7.6 μM, the plateau point beyond which all dimerizable R has dimerized. The second best model (dashed-dotted in Figure 4) was 3M (*p* fixed to 1) with K_{RRtt} freely estimated as 17 μM^3. This second best model is the best model when recent gel filtration data [4] shown in Table 3 are also included in the analysis, see Table 4 (2E ranked 20th and 13th in Tables 2 and 4 in exhaustive model space fits and was not even fitted by the semi-exhaustive method described next).

Table 3: Rofougaran et al.'s R1 dimerization data

R_T	t_T	Dimer	Monomer	Average Mass
2.700	100	18100	910	175.692
0.135	100	693	98	168.850
2.700	0	935	19766	94.065

Semi-exhaustive model selection

The semi-exhaustive model selection algorithm is: (1) create a list of all of the candidate models; (2) sort it according to the number of freely estimated parameters in each model; (3) fit all of the models with the fewest number of parameters; (4) fit all models with one additional parameter; and (5), repeat step 4 as long as the current batch of models has an improved AIC relative to the previous batch of models. In the case of the Rt system, compared to exhaustive fits to the entire space of 58 (*g*, *h*) models, this algorithm stops before fitting the most time consuming over-parameterized models (those with three parameters or higher) though it identifies the exact same top 13 (Table 2) and top 7 (Table 4) models. CPU times to compute Tables 2 and 4, expressed as exhaustive to semi-exhaustive ratios, averaged 4.7 (4.3/.89, 5.8/1.25, in minutes/minutes) when using 4 CPUs and 5.9 (14.8/2.5, 20.3/3.5) when using 1 CPU, or, rewritten, quad processor gains averaged 3.5 (14.8/4.3, 20.3/5.8) for exhaustive fits and 2.8 (2.5/.89, 3.5/1.25) for semi-exhaustive fits, i.e. there are semi-exhaustive approach losses in parallel processing efficiency as some CPUs become idle while the last models in a batch are fitted.

Implementation

R codes are provided to insure reproducibility of the results. They are also provided because they may be useful in other ligand induced protein dimerization data analyses. The following script illustrates their use.

```
setwd("/home/radivot/case/active/rnr/Rt/
R")

load("RNR.RData") # load RNR adata

source("fRt.r") # function definitions

# the next line generates and compiles C
code

g=mkgObj("Rt", c("Rt","RR","RRt","RRtt"))

RtData=adata [c("f1a01")] # Scott et al
2001 Rt data

# these map Kd into Kj as shown in Table 1

Eshape<-function(x)

   c(x[1],        2*x[2],        x[1]*x[3],
2*x[1]^2*x[4])

nshape<-function(x)
```

Table 4: Joint Data Analysis

Model	Parameter	Initial Value	Optimal Value	Confidence Interval
3M	RRtt	1.000	18.697	(4.807,72.966)
	Rt	Inf	Inf	absent
	RR	Inf	Inf	absent
	RRt	Inf	Inf	absent
	pRT	1.000	1.000	fixed
	SSE	0.064	0.034	
	AIC	-48.066	-54.448	
	cpu	0.000	0.445	fit succeeded
3Mp	RRtt	1.000	5.558	(0.370,83.931)
	pRT	1.000	0.907	(0.787,1.044)
	Rt	Inf	Inf	absent
	RR	Inf	Inf	absent
	RRt	Inf	Inf	absent
	SSE	0.064	0.027	
	AIC	-44.852	-53.308	
	cpu	0.000	0.199	fit succeeded
3Rp	pRT	1.000	0.822	(0.736,0.918)
	Rt	Inf	Inf	absent
	RR	Inf	Inf	absent
	RRt	Inf	Inf	absent
	RRtt	0.000	0.000	fixed
	SSE	0.106	0.041	
	AIC	-42.954	-52.590	
	cpu	0.000	0.104	fit succeeded
3I	RRt	1.000	49.568	(5.755,428.375)
	RRtt	1.000	37.930	(5.003,290.035)
	Rt	Inf	Inf	absent
	RR	Inf	Inf	absent
	pRT	1.000	1.000	fixed
	SSE	0.165	0.030	
	AIC	-35.303	-52.218	
	cpu	0.000	0.223	fit succeeded
2E	R_t	25.000	143.621	(0.477,44355.855)
	R_R	75.000	956.076	(0.790,1202604.284)
	RR_t	0.550	0.106	(0.001,8.085)
	RRt_t	0.550	0.106	constrained
	pRT	1.000	1.000	fixed
	SSE	0.079	0.031	
	AIC	-38.357	-47.750	
	cpu	0.000	0.344	fit succeeded

```
   c(x[1],       2*x[2],       x[2]*x[3],
2*x[2]*x[3]*x[4])

models=list(

mkModelObj(RtData, g, "2E",

   Kdparams=c(R_t=30,  R_R=85,  RR_t=.55,
RRt_t=.55),

   Keq=c(RRt_t="RR_t"), Kd2Kj=nshape),

mkModelObj(RtData, g, "3Rp",

   Kjparams=c(Rt=Inf,   RR=Inf,   RRt=Inf,
RRtt=0),

   pparams=c(pRT=1)),

mkModelObj(RtData, g, "3M",

   Kjparams=c(Rt=Inf,   RR=Inf,   RRt=Inf,
RRtt=1))

)

fitMS(models,"MS2")
```

In this script, `load` loads the RNR data provided in Additional File 1 and `source` reads in the function definitions provided in Additional File 2. The main function, `fitMS`, fits the model space (2E, 3Rp, 3M) and writes the results to html and LaTeX files. It can be passed options to specify the number of CPUs and the choice of semi-exhaustive or exhaustive fitting. A script that fits all 58 (g, h) models is provided as Additional File 3.

Discussion

The most common approach to modeling is to manually identify several plausible models, fit them all, and accept the best in the lot, e.g. [13,14]. This approach works because human intuition carries external information that guides the choice of the initial lot. If the best model does not provide a good fit, or if it has parameters with very large confidence intervals, the lot can be augmented to include additional models with more or fewer parameters, respectively. The advantage of this approach is that only a handful of models needs to be fitted. The disadvantage is that different analysts can yield different results. In general, a model/hypothesis (e.g. that the experimental data cannot discriminate some K_j from ∞ or zero, or that some K_d equal others) is rejected if it is not among the best models selected, and supported if it is. Although inferences made from any model, including the best models, are always conditional on the truth of the model's assumptions, the likelihood of this truth increases as the model withstands elimination. This statement is valid only to the extent that alternative hypotheses are represented in the model space. For example, if a $K_d = K'_d$ model assumes symmetric oligomers (e.g. as in Eqs. 9 and 10) and the model space does not include counterpart models that assume asymmetric forms, the selection process can lend no additional support to the symmetry assumptions. On the other hand, if independent data support such symmetry assumptions, the use of a restricted model space may be acceptable. It is anticipated that large model spaces will generate many models that are roughly equally best. Overall inferences should then reflect an average of the inferences of the best models, perhaps weighted by some metric of closeness to the optimum. Methods of accomplishing this for (g, h) models is an important area of future work. Another important area is automated model space enumeration: although this can be readily achieved for K_d = ∞ or 0 spur graphs, it remains a challenge to achieve this for $K_d = K'_d$ grid graphs.

Conclusion

The process of extracting K estimates from data is inseparable from the process of (g, f) model selection. This process requires clear statements of the model space explored, the criterion used to rank models, and the method used to search the space. If standards can be developed for these entities, analyst-to-analyst variations in inferences made from identical datasets could be reduced.

Methods

Data procurement

Plot Digitizer [15] was used to digitize the data of Scot et al. shown in Fig. 4. These data were originally given with model-dependent free concentrations on the x-axis. Such x values were converted to total concentrations using the model and parameter values given by Scot et al. [1]. The data in Table 3 is from Fig. 1 of [4]. It was kindly provided by Dr. Anders Hofer.

Model selection

With P equal to the number of freely estimated model parameters, N equal to the number of steady state data points, and SSE equal to the sum of squared errors of the fitted model, the Akaike Information Criterion [16] used here has the form $AIC = 2P + N \log(SSE/N) + \frac{2P(P+1)}{N-P-1}$ [17]. This explicit metric states how much goodness of fit (SSE) one is willing to sacrifice to gain the benefit of one less parameter. For a given model, P and N are fixed, so AIC minimization reduces to SSE minimization by least squares.

Parameter estimation

Best fitting SSEs were found by nonlinear least squares using the `optim` function in R [11] with the Nelder-Mead [18] option for $P > 1$, the BFGS option for $P = 1$, and the Hessian option set to TRUE (see Additional Files). Hessians of the SSEs evaluated at the optimum were divided by 2, inverted, and multiplied by the mean squared error, $MSE = SSE/(N - P)$, to compute parameter estimate covariance matrices. From these, parameter estimate standard deviations were taken as the square roots of the main diagonal, and these were then multiplied by 1.96 to approximate 95% CIs. All parameters were estimated as e^c to constrain point estimates and CIs to positive values.

Authors' contributions

TR performed all of the work and wrote the manuscript.

Additional material

Additional File 1

RNR.RData = ***Data file***
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-15-S1.RDAT]

Additional File 2

fRt.r = ***R function definitions***
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-15-S2.R]

Additional File 3

Rt.r = ***R script used***
Click here for file
[http://www.biomedcentral.com/content/supplementary/1752-0509-2-15-S3.R]

Acknowledgements

I thank Dr. Hofer for his data (Table 3) and the referees for their suggestions. This work was supported by the National Cancer Institute under grant number K25CA104791. It does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

1. Scott CP, Kashlan OB, Lear JD, Cooperman BS: **A quantitative model for allosteric control of purine reduction by murine ribonucleotide reductase.** *Biochemistry* 2001, **40(6):**1651-61.
2. Kashlan OB, Scott CP, Lear JD, Cooperman BS: **A comprehensive model for the allosteric regulation of mammalian ribonucleotide reductase. Functional consequences of ATP- and dATP-induced oligomerization of the large subunit.** *Biochemistry* 2002, **41(2):**462-74.
3. Kashlan OB, Cooperman BS: **Comprehensive model for allosteric regulation of mammalian ribonucleotide reductase: refinements and consequences.** *Biochemistry* 2003, **42(6):**1696-706.
4. Rofougaran R, Vodnala M, Hofer A: **Enzymatically active mammalian ribonucleotide reductase exists primarily as an alpha6beta2 octamer.** *J Biol Chem* 2006, **281(38):**27705-11.
5. Ingemarson R, Thelander L: **A kinetic study on the influence of nucleoside triphosphate effectors on subunit interaction in mouse ribonucleotide reductase.** *Biochemistry* 1996, **35(26):**8603-9.
6. Wang J, Lohman GJ, Stubbe J: **Enhanced subunit interactions with gemcitabine-5'-diphosphate inhibit ribonucleotide reductases.** *Proc Natl Acad Sci USA* 2007, **104(36):**14324-9.
7. Storer AC, Cornish-Bowden A: **Concentration of MgATP2- and other ions in solution. Calculation of the true concentrations of species present in mixtures of associating ions.** *Biochem J* 1976, **159:**1-5.
8. Kuzmic P: **Fixed-point methods for computing the equilibrium composition of complex biochemical mixtures.** *Biochem J* 1998, **331(Pt 2):**571-5.
9. Press WH: *Numerical recipes in C: the art of scientific computing* Cambridge [Cambridgeshire]; New York: Cambridge University Press; 1988.
10. Sommese A, Wampler C: *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science* Singapore: World Scientific Publishing Company; 2005.
11. Ihaka R, Gentleman R: **R:a language for data analysis and graphics.** *Journal of Computational and graphical statistics* 1996, **5:**299-314.
12. Reichard P: **Interactions between deoxyribonucleotide and DNA synthesis.** *Annu Rev Biochem* 1988, **57:**349-74.
13. Schlee S, Carmillo P, Whitty A: **Quantitative analysis of the activation mechanism of the multicomponent growth-factor receptor Ret.** *Nat Chem Biol* 2006, **2(11):**636-44.
14. Kuzmic P, Cregar L, Millis SZ, Goldman M: **Mixed-type noncompetitive inhibition of anthrax lethal factor protease by aminoglycosides.** *Febs J* 2006, **273(13):**3054-62.
15. **Plot Digitizer** [http://plotdigitizer.sourceforge.net/]
16. Akaike H: **A new look at the statistical model identification.** *IEEE Transactions on Automatic Control* 1974, **19:**716-723.
17. Burnham KP, Anderson D: **Multimodel Inference: understanding AIC and BIC in Model Selection.** *Workshop on Model Selection, Amsterdam* 2004.
18. Nelder J, Mead R: **A simplex algorithm for function minimization.** *Computer Journal* 1965, **7:**308-313.