Published in final edited form as:

Nucleosides Nucleotides Nucleic Acids. 2012 ; 31(2): 147–156. doi:10.1080/15257770.2011.644370.

On Model Ensemble Analyses of Non-Monotonic Data

Tomas Radivoyevitch¹ and Charles A. Kunos²

¹Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106, USA, txr24@case.edu

²Department of Radiation Oncology, Case Western Reserve University and University Hospitals of Cleveland Case Medical Center, Cleveland, Ohio, USA

Abstract

Mammalian ribonucleotide reductase (RNR) activity has been reported to be non-monotonic in [ATP]. If many nonlinear models are to be fitted to such data automatically as part of a model search process, use of the same initial parameter values across all models can lead to too many poor fitting, monotonic least squares fits, i.e. false model rejections. We propose that such fits can be rescued by using as initial parameter estimates the final estimates of neighboring models that do have non-monotonic fits; here models are neighbors if complexes that they represent differ by at most one ligand. We use this approach to show that troughs in RNR activity versus [ATP] can be fitted similarly well by models that do or do not demand a third ATP binding site.

Keywords

Non-monotonic data; least squares; ribonucleotide reductase

INTRODUCTION

Deoxynucleoside triphosphate (dNTP) supply involves a *de novo* system whose substrates are ribonucleoside diphosphates (NDPs) and a salvage system whose substrates are deoxynucleosides (dNs). The main control point of dNTP supply is the *de novo* system enzyme ribonucleotide reductase (RNR:NDP→dNDP) (1, 2). RNR is the target of several anticancer drugs (3). It has two subunits, R1, and R2 or p53R2 (4, 5). RNR is regulated allosterically by ATP and dNTP binding to the R1 selectivity (*s*) and activity (*a*) sites. If the *s*-site is empty, bound by ATP, or bound by dATP, RNR reduces CDP or UDP, if it is bound by dTTP GDP is reduced, and if it is bound by dGTP ADP is reduced. Meanwhile, if the *a*-site is occupied by ATP versus dATP, RNR is activated versus inactivated. Thus, the *s*- and *a*-site have functions analogous to tuning and volume controls on a radio. The binding status of the *s*- and *a*-sites, and perhaps of a third ATP-binding *h*-site (6), interacts with an R1 monomer-dimer-(short-lived tetramer)-hexamer equilibrium (7) wherein R1 *j*-mers are bound by 1 (7) or 3 (8) R2 dimers. Many R1 complexes are thus possible ($\sim 10^2$ monomers $\Rightarrow \sim 10^{12}$ hexamers, see Appendix). Since each complex can be represented by a term in a mathematical model, or be left out of the model, the number of possible RNR models is astronomical (9). A first step in RNR data analyses then is to define model sets that are large enough to adequately represent the hypotheses of interest, yet small enough to be fitted in reasonable amounts of computing time. In this paper we define such a model space for *h*-site existence as the hypothesis and non-monotonic RNR CDP and GDP reductase activity data (6) (Figure 1) wherein troughs in activity versus [ATP] profiles are the main reason an *h*-site is suspected to exist. We also present a model ensemble fitting method whereby poor-fitting, monotonic fits are rescued by using as initial parameter estimates the final estimates of

successfully fitted neighboring models. Using this method we show that such troughs can be fitted similarly well by models that do or do not demand an *h*-site.

MATERIALS AND METHODS

Data of Kashlan *et al.* (6) was digitized with plotDigitizer (10). Models were fitted by nonlinear least squares using R (11). Parameters were exponentiated to constrain them to positive values.

RESULTS

CDP Reductase Models

Kashlan *et al.* interpreted their data in Fig. 1A as: ATP binding to the *s*-site drives inactive R1 monomers to active R1 dimers, ATP binding to the *a*-site drives active R1 dimers to inactive R1 tetramers, and ATP binding to a previously unknown *h*-site drives inactive R1 tetramers to active R1 hexamers (6), i.e. as in Fig. 2A. Later, mass distribution measurements of R1 using gas-phase electrophoretic macromolecular mobility analysis (7) and size exclusion chromatography (8) showed no signs of an R1 tetramer, i.e. R1 tetramers likely exist only as unstable intermediates whose concentrations are approximately zero and thus incapable of explaining the troughs in Fig. 1. The troughs do however imply that there is at least one ATP-loaded inactive R1 oligomer.

The model space used here is compared to Kashlan's model in Fig. 2. This space is defined by

$$0 = [R_T] - [R] - 2 \frac{[R]^2}{K_{R^2}} - 2 \frac{[R]^2[X]^{i_1}}{K_{R^2X^{i_1}}} - 6 \frac{[R]^6[X]^{i_2}}{K_{R^6X^{i_2}}} - 6 \frac{[R]^6[X]^{i_3}}{K_{R^6X^{i_3}}} \quad (1)$$

where R is R1 bound to CDP ([CDP] = 1 mM for the dataset used is high enough to saturate the catalytic site), X is ATP, mass action for an R1 *j*-mer with *i* copies of ATP bound to it,

$$[R^jX^i] = \frac{[R]^j[X]^i}{K_{R^jX^i}}, \quad (2)$$

is implicit (Eq. 2 defines the complete dissociation constants in Eq. 1), $[R_T]$ is the total R1 concentration, coefficients 2, 2, 6 and 6 in Eq. (1) are the number of copies of R1 in the two dimers and two hexamers represented by these four terms (to see this insert Eq. 2 into Eq. 1; here $[R^iX^j] = [R_iX_j]$), the integers i_1 , i_2 , and i_3 follow $i_3 > i_2$ and $i_2 > 3i_1 > 0$ (this forces R to distribute itself to the right in Eq. 1 as $[X]$ increases, since complexes with larger X/R ratios dominate equilibria with increasing $[X]$ to partition more X into a bound form (9)), and

$$k = (2k_0 [R^2] + 2k_1 [R^2X^{i_1}] + 6k_2 [R^6X^{i_2}] + 6k_3 [R^6X^{i_3}]) / R_T \quad (3)$$

is the expected specific activity where, to facilitate parameter estimate optimizations made jointly on the k 's and K 's it will be assumed that $k_2 = 0$ and that $k_0 = k_1 = k_3$. This defines a space of 225 5-parameter models where each model is uniquely identified by i_1 , i_2 , and i_3 (Figure 3A).

In Eq. (1), two dimers are represented because the first data point in Fig. 1A shows some activity at $[ATP] = 0$, which is assumed to be due to a dimer that has no ATP bound to it, and because the maximum at ~ 0.2 mM is presumably due to ATP bound active dimers. And

two hexamers are represented in Eq. (1) because an inactive ATP-loaded R1 oligomer is needed to represent the trough in activity in Fig. 2, and the global activity maximum at higher [ATP] then demands that the model also includes an active R1 hexamer. Since tetramers were not observed in (7) it follows that they must be too rare to explain the trough in activity, so the model must assume either an inactive ATP-loaded dimer or hexamer. We chose the latter because inactive R1 hexamers, albeit dATP-loaded rather than ATP-loaded, have been observed (7). Note that models with $i_1 > 4$, $i_2 > 12$, or $i_3 > 12$ demand the existence of an *h*-site (171 of the 225 models in the model space satisfy this condition, 54 do not), and that all of the models in the model space have 5 freely estimated parameters, so smaller sums of squared errors (SSE) imply better models (unless there is an additional penalty against high degrees of curvature, see Figure 1).

CDP Reductase Fits

With 225 nonlinear models (Fig. 3A) to fit to the data (Fig. 1A), it is not practical to identify initial parameter estimates for each model manually. Thus, initial estimates to be used across all models were sought. By inspection of Figure 1A, $k_0 = 0.3/\text{sec}$ ($k_2 = k_3 = k_0$) are reasonable initial estimates of the rate constants. Complete dissociation constants were then represented as base concentrations raised to $i + j - 1$ for complexes with $i + j$ reactants (i.e. R1 j -mers with i ATPs, see Eq. 2) to improve estimate convergence (9). This corresponds to binding energies of complexes being roughly proportional to the number of reactants bound in the complex, averaging R1-R1 and R1-ATP binding energies. A reasonable initialization rule is then

$$K_{R^2} = \mu\text{M}, K_{R^2X^{i_1}} = (10 \mu\text{M})^{i_1+1}, K_{R^6X^{i_2}} = (100 \mu\text{M})^{i_2+5}, \text{ and } K_{R^6X^{i_3}} = (500 \mu\text{M})^{i_3+5} \quad (4)$$

which yielded 73 non-monotonic fits and 152 monotonic fits. Those models that had monotonic fits (e.g. see Figure 4) were then placed into a queue to be refitted with initial conditions set to final estimates of a neighboring model that had a non-monotonic fit and an SSE less than 4 times the minimum SSE of all models fitted so far; neighboring models have a maximum difference in i_1 , i_2 or i_3 of one, see Figure 3A. This process was then repeated to spread non-monotonic fits into the balance of the model space. By the final iteration all 225 models yielded non-monotonic fits. The best of the final fits of models that demand an *h*-site or not could then be plotted as in Figure 1A. Median K base concentrations of all final fits then yielded as an initial condition rule

$$K_{R^2} = 3.5 \mu\text{M}, K_{R^2X^{i_1}} = (13.6 \mu\text{M})^{i_1+1}, K_{R^6X^{i_2}} = (27 \mu\text{M})^{i_2+5}, \text{ and } K_{R^6X^{i_3}} = (72 \mu\text{M})^{i_3+5}. \quad (5)$$

This initial parameter estimate rule vastly outperforms Eq. (4). Interestingly, monotonic fits formed SSE clusters (Figure 5) due to certain K values escaping to large values in the optimization, the corresponding term in Eq. (1) thus vanishing, and all models otherwise spawned by the missing term thus collapsing toward the same model, fit, and SSE.

GDP Reductase Models and Fits

To analyze the GDP reductase data in Kashlan *et al.*, consider the set of models defined by

$$0 = [D_T] - [D] - 3 \frac{[D]^3[X]^{i_2}}{K_{D^3X^{i_2}}} - 3 \frac{[D]^3[X]^{i_3}}{K_{D^3X^{i_3}}} \quad (6)$$

where D represents R1 dimers saturated with both GDP and dTTP (in the *s*-site) and $0 < i_2 < i_3 \leq 12$ since X (= ATP) can now bind only *a*- or *h*-sites. With this model Eq. (3) becomes

$$k = (2k_0 [D] + 6k_3 [D^3 X^{i_3}]) / (2 [D_T]) \quad (7)$$

where k_0 and k_3 are independently estimated. Of 66 such models, 51 demanded an h -site and 15 did not. The top two models (lowest SSEs), ($i_2 = 5$, $i_3 = 6$) and ($i_2 = 4$, $i_3 = 5$), did not demand an h -site, but the third best, ($i_2 = 6$, $i_3 = 7$), did. Differences in fit were however negligible (Fig. 1B). Applying $k_0 = k_3$ as a constraint yielded the same top 3 models and similar SSEs.

DISCUSSION

In previous work (9) the Akaike Information Criteria (AICs) (12, 13) of many models fitted to dynamic light scattering (DLS) R1 average mass versus [ATP] data in (6) suggested a lack of h -site evidence. Meanwhile, other DLS data in (6) did support h -site existence (see Kashlan's review in (9)). Unexplored, however, was the extent of h -site evidence in the *activity* versus [ATP] data in (6), such as that shown in Figure 1. This motivated the current study.

Of the 225 models fitted to Kashlan *et al.*'s CDP reductase versus ATP data (6) the best fit was (i_1 , i_2 , i_3) = (4, 13, 18), shown in black in Fig. 1A. This model supports the existence of an h -site because i_2 and i_3 are both greater than 12. That qualitative up-down-up behaviour does not imply an h -site is seen in the blue curve in Fig. 1, (i_1 , i_2 , i_3) = (2, 7, 12), the lowest SSE model that does not demand an occupied h -site. With enough of a model selection penalty against fit curvature, SSE differences of these two models (0.0026 vs. 0.0039) could perhaps be annihilated to select the blue fit, i.e. a model that does not require the existence of an h -site, but this is pure speculation. Our main result then is our method of rescuing local minima fits (Figures 4 and 5). Applying it to Kashlan *et al.*'s RNR GDP reductase activity data also yielded ambiguous results regarding h -site existence (Figure 1B).

Our model space is degenerate to the extent that it cannot disprove the existence of h -sites, as it is always possible that some may fill before all of the a -sites are filled. It can, however, prove their existence, as results in Figure 1A may suggest. However, since there was no mention of an h -site in the recent structure of human R1 (14), and no other laboratory has observed troughs as in Fig. 1 (7), conclusions regarding an h -site must remain on hold.

Assuming no h -site, with $i_2 = 7$, the blue curve in Figure 1A suggests that ATP binding one a -site nucleates R1 polymerization to an inactive hexamer, and $i_3 = 12$ suggests that remaining a -site filling causing conformational changes to an active hexamer. Meanwhile, Fig. 1B ($i_2 = 5$, $i_3 = 6$) is consistent with 5 ATPs binding to a -sites to inactivate R1 as a hexamer, and one additional ATP then binding to reactivate it. The structure of human R1 hexamers (14) suggests that either all a -sites are the same or that there are 3 of one kind and 3 of another; an average of our blue curve fit interpretations is more consistent with the latter than the former.

Kashlan's models (6) (e.g. Fig. 2A) use binary dissociation constants rather than complete dissociation constants (Eq. 2) as in our models. Binary dissociation constants allow K equality hypotheses that reduce the number of freely estimated model parameters despite the largenumber of complexes represented. Complete dissociation constant models allow hypotheses that complexes are so rare that their concentrations are approximately zero. Complete K models tend to have fewer estimated parameters than binary K models and are thus often preferable (9, 15). It was for this reason that we focused here strictly on complete K models.

CONCLUSIONS

RNR models will eventually be incorporated into dNTP supply system models (16) used to optimize therapies (17). The model space and model ensemble fitting process presented here could impact RNR models, and thus eventually, therapeutic applications of dNTP supply models.

Acknowledgments

TR was supported by NIH grant CA104791. We thank the reviewers for excellent suggestions.

APPENDIX

Number of Monomer and Hexamer States

The R1 subunit of RNR has 5 substrate site states (unoccupied or occupied with one of 4 NDPs), 5 selectivity site states (unoccupied or occupied with ATP, dATP, dTTP or dGTP), and 3 activity site states (unoccupied or occupied with ATP or dATP). This implies $5 \times 5 \times 3 = 75$ R1 states. If one now also includes 2 *h*-site states (unoccupied or occupied with ATP), the total is 150. Thus, with ~100 R1 monomer states, assuming independence (not true but models inconsistent with data can be eliminated later), R1 hexamers can then have $\sim 100^6 = \sim 10^{12}$ states.

Other Total Concentration Constraints

Equation 1 is a total concentration constraint for the R1 subunit. We assumed that it could be treated as approximately decoupled from two other constraints, one for ATP

$$0 = [X_T] - [X] - i_1 \frac{[R]^2 [X]^{i_1}}{K_{R^2 X^{i_1}}} - i_2 \frac{[R]^6 [X]^{i_2}}{K_{R^6 X^{i_2}}} - i_3 \frac{[R]^6 [X]^{i_3}}{K_{R^6 X^{i_3}}} \quad (\text{A1})$$

which we ignored because non-zero $[X_T] = 50 \mu\text{M}$ for this data (6) is large enough relative to $3[R_T] = 3.6 \mu\text{M}$ (3 ATP binding sites multiplied by $[R_T] = 1.2 \mu\text{M}$ for this data) that $[X] \approx [X_T]$ (i.e. free [ATP] $\approx$ total [ATP]) can be assumed, and one for R2 dimers, $[r_T]$, that can be ignored if the proportion of R bound to r is independent of $[X]$ and the oligomeric status of R (since non-saturating $[r]$ can then be captured by a uniform decrease in the per-site activity constants k_i).

REFERENCES

1. Thelander L, Reichard P. Reduction of ribonucleotides. *Annu Rev Biochem.* 1979; 48:133–158. [PubMed: 382982]
2. Eklund H, Uhlin U, Farnegardh M, Logan DT, Nordlund P. Structure and function of the radical enzyme ribonucleotide reductase. *Prog Biophys Mol Biol.* 2001; 77:177–268. [PubMed: 11796141]
3. Shao J, Zhou B, Chu B, Yen Y. Ribonucleotide reductase inhibitors and future drug design. *Curr Cancer Drug Targets.* 2006; 6:409–431. [PubMed: 16918309]
4. Tanaka H, Arakawa H, Yamaguchi T, Shiraishi K, Fukuda S, Matsui K, Takei Y, Nakamura Y. A ribonucleotide reductase gene involved in a p53-dependent cell-cycle checkpoint for DNA damage. *Nature.* 2000; 404:42–49. [PubMed: 10716435]
5. Nakano K, Balint E, Ashcroft M, Vousden KH. A ribonucleotide reductase gene is a transcriptional target of p53 and p73. *Oncogene.* 2000; 19:4283–4289. [PubMed: 10980602]
6. 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; 41:462–474. [PubMed: 11781084]
7. Rofougaran R, Vodnala M, Hofer A. Enzymatically active mammalian ribonucleotide reductase exists primarily as an $\alpha_6\beta_2$ octamer. *J Biol Chem*. 2006; 281:27705–27711. [PubMed: 16861739]
 8. Wang J, Lohman GJ, Stubbe J. Enhanced subunit interactions with gemcitabine-5'-diphosphate inhibit ribonucleotide reductases. *Proc Natl Acad Sci U S A*. 2007; 104:14324–14329. [PubMed: 17726094]
 9. Radivoyevitch T. Automated mass action model space generation and analysis methods for two-reactant combinatorially complex equilibria: An analysis of ATP-induced ribonucleotide reductase R1 hexamerization data. *Biol Direct*. 2009; 4:50. [PubMed: 20003203]
 10. Plot Digitizer. <http://plotdigitizer.sourceforge.net/>
 11. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and graphical statistics*. 1996; 5:299–314.
 12. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974; 19:716–723.
 13. Burnham, KP.; Anderson, DR. *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*. Springer-Verlag; 2002.
 14. Fairman JW, Wijerathna SR, Ahmad MF, Xu H, Nakano R, Jha S, Prendergast J, Welin RM, Flodin S, et al. Structural basis for allosteric regulation of human ribonucleotide reductase by nucleotide-induced oligomerization. *Nature structural & molecular biology*. 2011; 18:316–322.
 15. Radivoyevitch T. Equilibrium model selection: dTTP induced R1 dimerization. *BMC Syst Biol*. 2008; 2:15. [PubMed: 18248678]
 16. Bradshaw PC, Samuels DC. A computational model of mitochondrial deoxynucleotide metabolism and DNA replication. *Am J Physiol Cell Physiol*. 2005; 288:C989–1002. [PubMed: 15634740]
 17. Radivoyevitch T, Loparo KA, Jackson RC, Sedwick WD. On systems and control approaches to therapeutic gain. *BMC Cancer*. 2006; 6:104. [PubMed: 16638124]

Figure 1.

The RNR CDP (A) and GDP (B) reductase activity data of Kashlan et al. (6). Fits of the best fitting models of those that demand an *h*-site (black) or not (blue) are shown. **A)** Parameters (top) are defined by Eqs (1-3). Since all such models (indexed by i_1 , i_2 and i_3) have the same number of parameters (i.e. 5), smaller sums of squared errors (SSE) are better. We note however that SSE does not penalize against high curvature in the black fit at very low [ATP], inset. **B)** Using Eqs (6-7) for GDP, with dTTP saturating *s*-sites, active dimers become inactive hexamers with some [ATP] and active ones with more.

Figure 2.

A) Kashlan's CDP reductase model. In this model the trough in Fig. 1A is due to ATP (X) loaded inactive (black) R1 (R) tetramers; inactive-active tetramer equilibria in the original model are not essential here and are thus not shown. Expansions of factors $(1+X/K)^j$ for R1 j -mers (R^j) yields terms X^i/K^i that represent R1 j -mers with i ATPs. Thus, the model shown represents 1 monomer, 3 dimers, 5^2 tetramers and 7^3 hexamers, i.e. $1+3+25+343 = 372$ complexes. **B)** The CDP reductase model space of Eq. (1). Here the trough in Fig. 1A is attributed to an ATP-loaded inactive R1 hexamer. Each of these models (indexed by i_1 , i_2 , and i_3) represents only 5 complexes, i.e. very few compared to A. In A and B active enzyme complexes are green.

Figure 3. Model Spaces. Models that demand an occupied h -site are in red, others are in black. Sphere radii are proportional to $1/\text{SSE}$. **A)** The 225 models spawned by Eq. (1). **B)** The 66 models of Eq. (6).

Figure 4.

Conversion of a monotonic fit (red) to a non-monotonic fit (blue). The monotonic fit has a suboptimal SSE. Its parameter estimates were initiated by Eq. (4). Using instead of Eq. (4) the final estimate of a non-monotonic neighbouring model fit, or starting with Eq. (5), yielded the blue curve.

Figure 5. Distributions of model SSEs in the first (red) and final (blue) rounds of model fitting.