The Advantage of Global Fitting of Data Involving Complex Linked Reactions

ARTICLE *in* METHODS IN MOLECULAR BIOLOGY (CLIFTON, N.J.) · JANUARY 2012

Impact Factor: 1.29 · DOI: 10.1007/978-1-61779-334-9_22 · Source: PubMed

2 AUTHORS:

Petr Herman
Charles University in Prague
85 PUBLICATIONS **1,553** CITATIONS
SEE PROFILE

James Ching Lee
University of Texas Medical Branch at Galve…
140 PUBLICATIONS **5,671** CITATIONS
SEE PROFILE

HERMAN, P.,LEE, J.C.: The advantage of global fitting of data involving complex linked reactions in A.W.W. Fenton (ed.) Allostery: Methods and Protocols. New York: Humana Press Inc., 2012, 399-421. ISBN: 978-1-61779-333-2, DOI: 10.1007/978-1-61779-334-9_22.

Chapter 22

The Advantage of Global Fitting of Data Involving Complex Linked Reactions

Petr Herman and J. Ching Lee

Abstract

In this chapter, we demonstrate the advantage of the simultaneous multicurve nonlinear least-squares analysis over that of the conventional single-curve analysis. Fitting results are subjected to thorough Monte Carlo analysis for rigorous assessment of confidence intervals and parameter correlations. The comparison is performed on a practical example of simulated steady-state reaction kinetics complemented with isothermal calorimetry (ITC) data resembling allosteric behavior of rabbit muscle pyruvate kinase (RMPK). Global analysis improves accuracy and confidence limits of model parameters. Cross-correlation between parameters is also reduced with accompanying enhancement of the model-testing power. This becomes especially important for validation of models with "difficult" highly cross-correlated parameters. We show how proper experimental design and critical evaluation of data can improve the chance of differentiating models.

Key words: Nonlinear least squares, Parameter correlation, Global fitting, Pyruvate kinase, Calorimetry, Two-state allosteric model, Monte Carlo analysis

1. Introduction

Since the discovery of allosteric regulation of proteins, the phenomena of cooperative ligand binding were recognized to play an important role in the regulation of biological functions in living cells. In considering the mechanism(s) of allostery, the investigator must be cognizant of interactions of at least two ligands with a macromolecule, which is capable of undergoing conformational change(s). Furthermore, these reactions are thermodynamically linked because each reaction is influenced by the presence or absence of the other. Decades of research have revealed highly diverse mechanisms of the allosteric regulation in different proteins. These mechanisms include the well-known concerted model proposed by Monod, Wyman, and Changeux (MWC) (1)

Aron W. Fenton (ed.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 796,
DOI 10.1007/978-1-61779-334-9_22,

or a sequential model by Koshland, Nemethy, and Filmer (KNF) (2). Eigen developed yet another model, which is a combination of these two models (3).

Owing to a high variability of possible allosteric mechanisms, their detailed understanding often requires construction of alternative molecular or thermodynamic models. Quantitative parameterization as well as validation of such models depends on validation of mathematical expressions representing the model by experimental data. It is generally accepted that models cannot be proven but can only be disproven. Once conditions are found when a particular model cannot explain the experimental observations, the model has to be rejected or modified. During the validation process, a researcher is confronted by an important task of a thorough challenge of alternative models with data by varying different solution variables (temperature, pressure, pH, ionic strength, concentration, etc.), or technical details (excitation/emission wavelength in spectroscopy, angular velocity in sedimentation analysis, etc.). Objective comparison between the experimental data and the proposed model is often done by a nonlinear least-squares (NLSQ) analysis (4). Naturally, the mathematical description of the model should contain parameters that can account for changes of external conditions and that are adjusted during the NLSQ analysis. Then the parameterized models can also predict system behavior that further helps the researcher in model validation. Thus, Fig. 1 summarizes the general philosophy of studying complex linked reactions.

Global analysis (5–7) is a powerful method for discerning between models and for accurate recovery of model parameters. The method is based on an ordinary nonlinear least-squares minimization (4) and allows for simultaneous analysis of multiple datasets. Data for the analysis can be acquired under different experimental

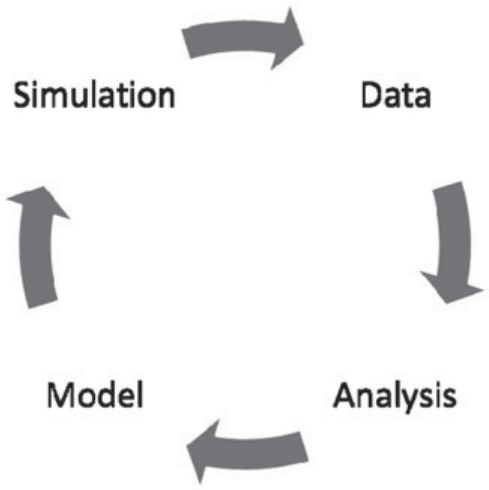

Fig. 1. Outline of general philosophy of studying complex linked reactions.

conditions and with different techniques. Consequently, some model parameters are common for multiple curves. During global fitting, the overall sum of weighed squared deviations of measured values and values calculated from a model that encompasses all experimental data is minimized. Global analysis, thus, allows determination of parameter values that are consistent with all datasets. Importantly, linkage between particular parameters and different datasets sharpens the minimum of the χ^2 surface with concomitant decrease of parameter correlation. Those parameters are recovered with higher accuracy (6). Overdetermination of parameters in the model, inherent to global analysis, helps to distinguish between alternative models by eliminating models inconsistent with data. The resolving power of global analysis is unequaled by a conventional NLSQ. Global analysis has been previously used for large variety of experimental data and techniques (5, 8–21). Recently, we have used global analysis of a dataset obtained by isothermal calorimetry (ITC), fluorescence binding experiments, and fluorescence temperature scans for evaluation of the functional energetic landscape in the allosteric regulation of rabbit muscle pyruvate kinase (RMPK) (22–24). In all cases, the global analysis approach provided performance far beyond the resolution of the conventional analysis.

In this chapter, we demonstrate the use of global analysis of multiple data curves acquired under different experimental conditions. We show that even a single properly chosen curve added to the dataset for analysis can significantly reduce correlation between parameters that often prohibits accurate recovery of parameter values and consequently lowering the chance to distinguish between alternative models.

2. Methods

2.1. Model

RMPK is an important allosteric enzyme of the glycolytic pathway catalyzing a transfer of the phosphate from phosphoenolpyruvate (PEP) to ADP:

$$\mathrm{PEP} + \mathrm{ADP} \overset{\mathrm{PK}}{\Leftrightarrow} \mathrm{Pyruvate} + \mathrm{ATP}$$

RMPK consists of four identical subunits (25), and its activity is subjected to a complex pattern of allosteric regulation. Although subject to recent challenges, numerous early experiments (13, 15, 16, 22–24, 26) suggested that all data can be described by a two-state concerted allosteric model (1) where protein assumes an active (R) and an inactive (T) form. In this model, differential affinities of ligands to the R and T-state perturb equilibrium between the states by shifting it to one or the other form. The general MWC model was customized for the tetrameric

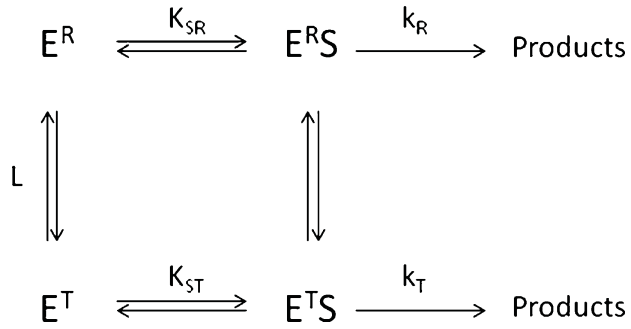

Fig. 2. Simplified model of the two-state allosteric enzyme. The enzyme equilibrates between an active R- and inactive T-state. Substrate binds to both the R- and T-states. Catalytic activities of the two states differ.

RMPK by Oberfelder et al. (13) and elaborated further by Herman and Lee (22–24).

For demonstration of the global analysis, we have chosen a model based on the reaction scheme from Fig. 2. It represents a simplified case of the general scheme published by Oberfelder (13). The substrate S binds to both the R and T-state of RMPK with dissociation constant K_{SR} and K_{ST}, respectively, which results in a catalytic activity. For the tetrameric RMPK, the steady-state reaction rate v can be described as (13):

$$v = \frac{k_R \cdot [S]/K_{SR} \cdot (1 + [S]/K_{SR})^3 + k_T \cdot L \cdot [S]/K_{ST} \cdot (1 + [S]/K_{ST})^3}{(1 + [S]/K_{SR})^4 + L \cdot (1 + [S]/K_{ST})^4} \tag{1}$$

where $[S]$ is concentration of substrate, $L = [T]/[R]$ is an equilibrium constant between the unliganded R- and T-state of the enzyme, and k_R and k_T are reaction rate constants for the R- and T-state, respectively. Assuming the T-state is fully inactive, $k_T = 0$ and binding of PEP to the T-state is negligible ($[S]/K_{ST} \to 0$ for any $[S]$) (23), Eq. 1 can be simplified to:

$$v = \frac{k_R \cdot [S]/K_{SR} \cdot (1 + [S]/K_{SR})^3}{(1 + [S]/K_{SR})^4 + L} \tag{2}$$

Temperature dependence of the equilibrium constant L can be expressed as:

$$L = \exp\left(-\frac{\Delta H_{R-T} - T \cdot \Delta S_{R-T}}{RT}\right) \tag{3}$$

where ΔH_{R-T} and ΔS_{R-T} are the enthalpy and entropy change of the $R \to T$ transition, respectively, R, the universal gas constant, and T, temperature in Kelvin (see Note 1). We use Eqs. 2 and 3 for simulation of reaction kinetics and their subsequent analysis.

2.2. Simulation of Reaction Kinetics

The task is to recover k_R, K_{SR}, and L at 40°C, i.e., close to the rabbit's physiological temperature, by fitting Eq. 2 to synthetic data supplemented with a realistic level of random noise. Then, confidence limits of the fitted values and correlation between parameters are evaluated for both the "standard" single-curve fit and the global fit of several curves of reaction kinetic data.

Employing Eq. 2 we have simulated reaction kinetics at 20, 30, 40, and 50°C using parameters values close to the ones published for RMPK (22–24) (see Note 2). Some adjustment of parameters was done to accommodate for the model simplification and to work with "pleasant round numbers". Qualitative behavior of the model, however, still closely resembles real RMPK. In particular, we used a normalized value of the rate coefficient $k_R = 1$ mol/s and the dissociation constant K_{SR} = 100 μM. Temperature dependence of the equilibrium constant between the R and T-state was calculated from Eq. 3 using ΔH_{R-T} = 44 kcal/mol and $\Delta S_{R-T} = 150$ cal/mol/deg. Then a random Gaussian noise with standard deviation (SD) of 0.01 was added to each curve. The noise mimicked noise levels obtainable in real experiments (1% relative error at saturation). For illustration, the calculated temperature dependence of L is shown together with the inactive fraction of the enzyme $f_T = L/(1 + L)$ in Fig. 3. At 40°C, the value of L is 119.

2.3. Single-Curve Fit

The simulated reaction kinetics at 40°C together with the best fit is shown in Fig. 4. The fitting was done by a standard NLSQ method that can be found in most scientific graphing packages.

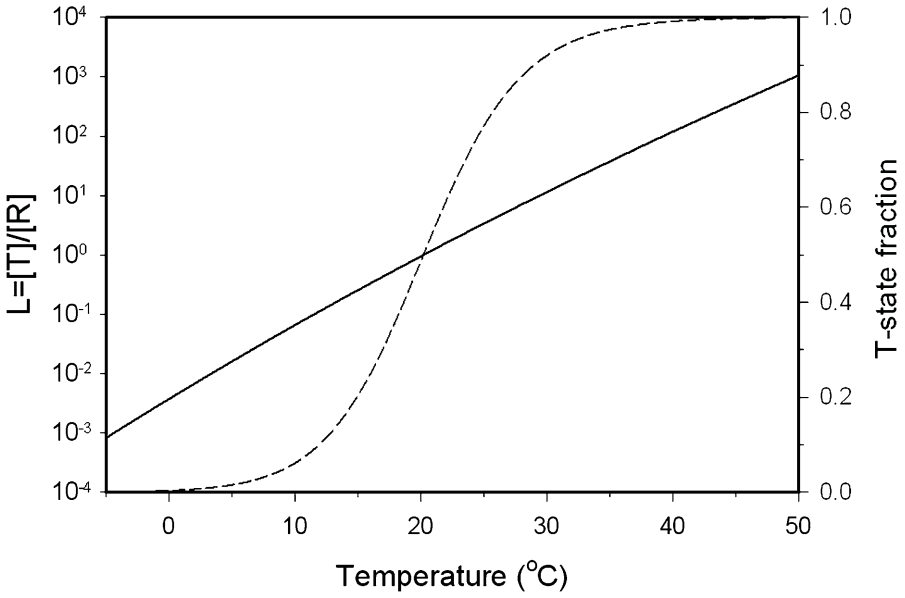

Fig. 3. Temperature dependence of the equilibrium constant between R and T state for $\Delta S_{R-T} = 150$ cal/mol/deg and $\Delta H_{R-T} = 44$ kcal/mol. The *dashed line* represents a fraction of the inactive T-state as a function of temperature.

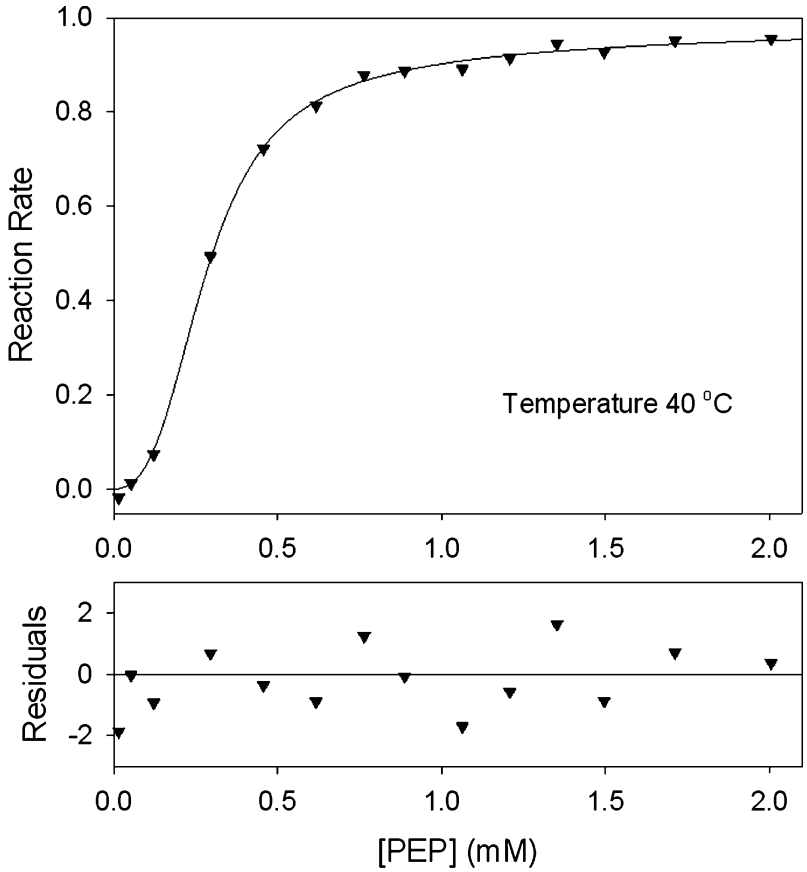

Fig. 4. Simulated steady-state reaction kinetics at 40°C. Standard deviation of the added Gaussian noise is 0.01. The *solid line* represents the best fit.

As described in numerous literatures, a function χ^2_R called reduced chi-squared is minimized during the NLSQ fit (4):

$$\chi^2_R = \frac{1}{(M-m-1)} \cdot \sum_{i=1}^{n} \left[\frac{\Upsilon_{Data}(i) - \Upsilon_{Model}(i)}{\sigma(i)} \right]^2 \qquad (4)$$

where M is the number of data points, m the number of fitted parameters, $\Upsilon_{Data}(i)$ the i-th experimental value, $\Upsilon_{Model}(i)$ the calculated value, and $\sigma(i)$ standard deviation of the corresponding data point. χ^2_R is a quantity normalized for the number of data points and model complexity. When correctly used, minimization of χ^2_R gives parameter values that have the highest probability of being correct (27) (see Notes 3 and 4). The NLSQ fit of the curve from Fig. 4 yielded parameter values $k_R = 0.998$ mol/s, $K_{SR} = 96.5$ μM, and $L = 146$. It is clear that we did not recover the exact values for parameters used in simulation. Especially, the recovered value for L is quite different from the correct value of 119. The sources of error are as follows: (a) the experimental noise that biases the true χ^2_R minimum and (b) the high correlation between parameters that causes the minimum of the χ^2_R surface to be shallow and insensitive to variations in parameters. As a consequence, a change of one parameter can be compensated by

changes of other parameters without significant change of χ^2_R. When the magnitude of the correlation increases, the estimation of parameters by NLSQ becomes difficult or even impossible. One, therefore, always needs to know the degree of correlation and the confidence intervals (CI) of the fitted parameters. Without knowing the CI of the fitted results one cannot critically evaluate their significance and make comparison with other evidence.

2.4. Parameter Correlation and Confidence Intervals

There are several method for evaluation of the CIs for NLSQ analysis, see e.g., (27) for an overview. The simplest method is an asymptotic standard error (ASE) implemented in most commercial software packages. Unfortunately, this computationally simple method, which yields symmetrical CIs, uses implicit assumptions that can hardly be fulfilled for most NLSQ problems (27–30). In particular, ASE requires a linear fitting function, a large number of data points, and the absence of any correlation between parameters. As a consequence, the ASE confidence limits are usually far from reality. The preferred approaches are bootstrap (31) or Monte Carlo (29) methods. Although computationally extensive, they use fewer assumptions and provide more realistic estimates of the CIs. Parameter cross-correlation is also inherently and rigorously evaluated during calculation of CIs.

For estimation of CIs, we chose the Monte Carlo (MC) method. Its principle is rather simple:

1. Data were analyzed to obtain a set of the most probable model parameters, see Subheading 3.
2. A noiseless data curve was generated using parameters recovered in step 1 and independent variables x_i values from the original data.
3. Random Gaussian noise with the same level as contained in the original data was added to this noiseless curve ($SD = 0.01$).
4. The resulting data were analyzed and parameters tabulated.
5. Steps 3–4 were repeated many times to obtain sufficient statistics sampling the "best parameter values".
6. Histograms representing a confidence probability distribution for each particular parameter were constructed. As seen from the description of the MC procedure, the distributions originate from the noise contained in the data only. Accuracy of the recovered distribution depends mainly on a number of MC runs and how closely noise characteristics of the simulated curves match the original data noise. For a reasonable accuracy it is necessary to perform more than several hundred runs. In our example, we used 1,500 MC cycles.
7. Cross-correlation between the p_i and p_j parameter was visualized by plotting p_i against corresponding p_j value for each MC run.

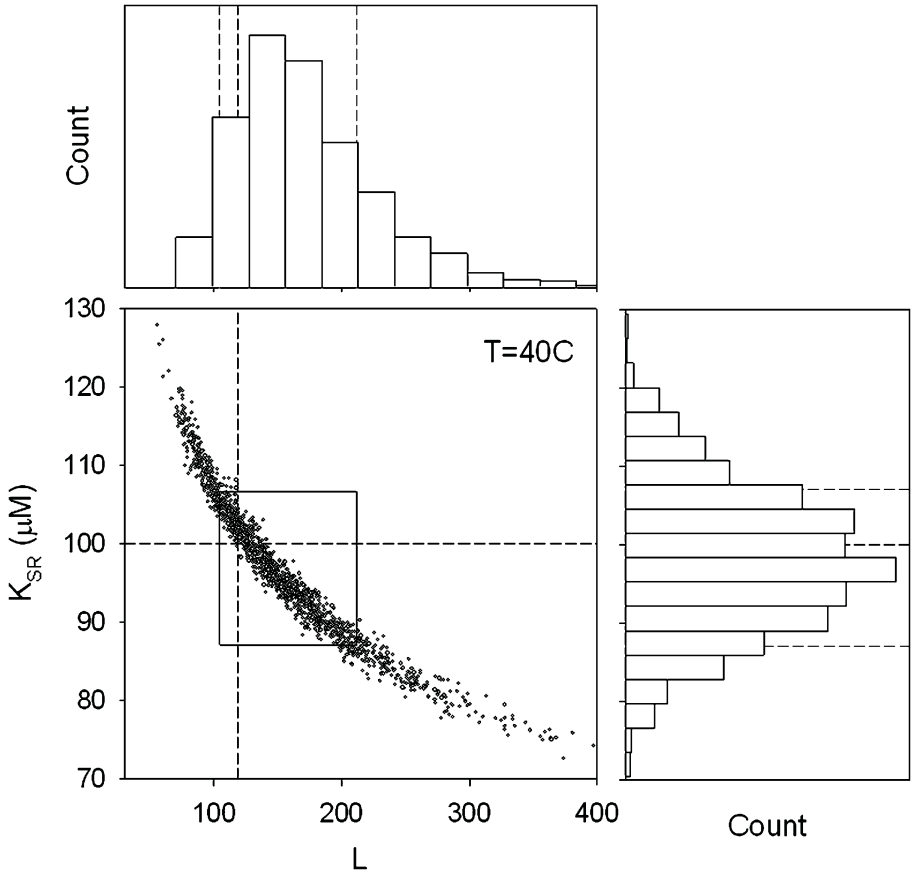

Fig. 5. Correlation between L and K_{SR} for a single curve analysis presented in Fig. 4. The correlation was obtained by 1,500 MC runs. Each point is a parameter pair from a single fit to data with unique Gaussian noise. *Black dashed lines* indicate original parameter values used for simulation. Histograms show resulting distribution of the fitted parameter values. *Gray dashed lines* border the region of ± 1 SD. This region is depicted by the *black rectangle* in the central graph.

Results of the MC analysis for the single-curve fit are visualized in Figs. 5 and 6 and quantitatively summarized in Table 1. Correlation between L and K_{SR} with corresponding confidence distributions are shown in Fig. 5. Each point in the central graph is a parameter pair from a single MC run. We can clearly see a strong negative correlation between L and K_{SR}. The parameter pairs are not randomly distributed in the parameter plane. Instead, they form a well-defined "sickle" shape indicating a dependence of K_{SR} on L. Black dashed lines mark the original parameter values used for the simulation of the "experimental" data from Fig. 4. The histograms present resulting distribution of the fitted parameter values. From the area of the histograms we evaluated confidence intervals for the 68% confidence level (1 SD) that are bordered by gray dashed lines. It is clear that the CI of the equilibrium constant L is not symmetrical and the distribution is significantly skewed to higher values. Similar pattern with slightly weaker correlation between L and k_R can be recognized in Fig. 6. Confidence limits

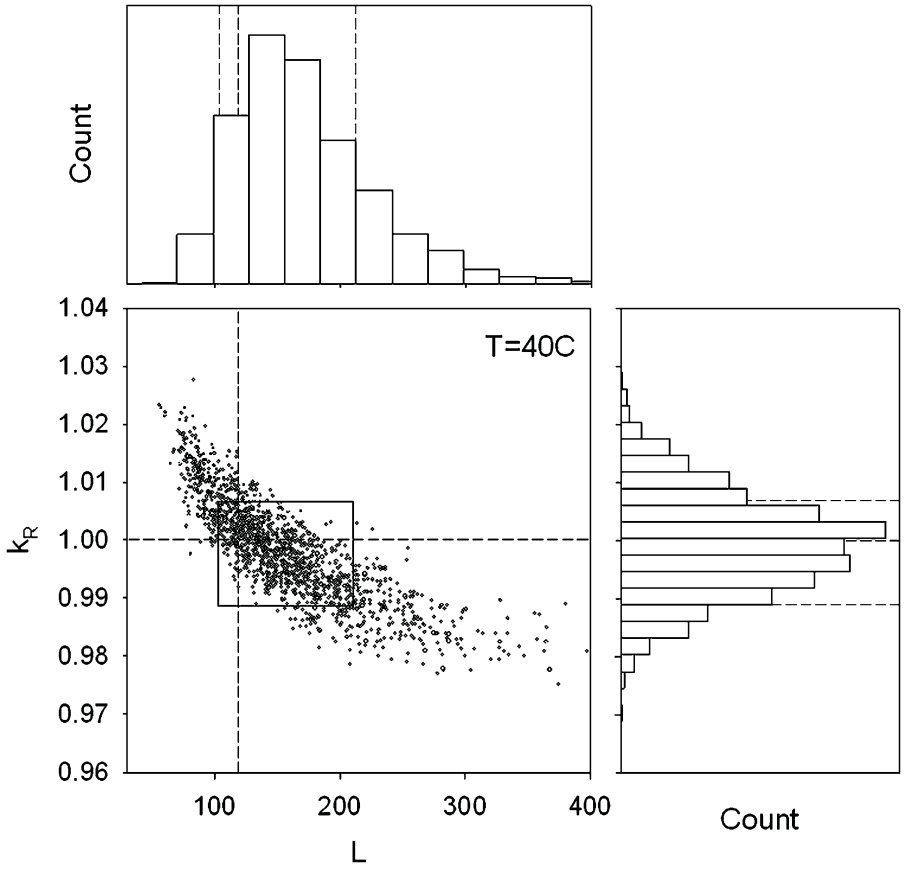

Fig. 6. Correlation between L and k_R for a single curve analysis presented in Fig. 4. The correlation was obtained by 1,500 MC runs. *Black dashed lines* indicate original parameter values used for simulation. Histograms show resulting distribution of the fitted parameter values. *Gray dashed lines* border the region of ± 1 SD. This region is depicted by the *black rectangle* in the central graph.

calculated from the histograms are summarized in Table 1. The area defined by the CIs is depicted by the black rectangle in the central correlation graph. It is seen that the area is not randomly filled with points – a sign of cross-correlation.

2.5. Global Analysis: Background

From the MC analysis of the single-reaction kinetics, we learned that parameters of our two-state model describing allosteric regulation of RMPK are strongly cross-correlated. Consequently, the correlation also influences the widths of the CIs because deviation of any parameter can be compensated by an adjustment of other parameters without significant change of χ^2_R (see Note 5). Reduction of "covariance valleys" on the χ^2_R hypersurface would certainly improve both testing capability of the model and accuracy of recovered parameters. This can be done by simultaneous analysis of multiple curves covered by a single mathematical model where groups of curves share the same parameters.

Table 1
Comparison of fitting results at 40°C

Parameter	Lower limit[a]	Recovered value	Upper limit[a]	χ^2_R
L				
Single curve analysis	104	146	212	1.12
Global analysis	109	125[b]	137	0.86
Single curve, $4N$ analysis[c]	107	137	147	0.99
Global with calorimetry	117	121[d]	124	0.98
True value	–	*119*[e]	–	–
K_{SR}				
Single curve analysis	87	96.5	107	1.12
Global analysis	97	99.9	103	0.86
Single curve, $4N$ analysis	95	96.8	104	0.99
Global with calorimetry	99	100	101	0.98
True value	–	*100*	–	–
k_R				
Single curve analysis	0.989	0.998	1.007	1.12
Global analysis	0.999	1.0026	1.006	0.86
Single curve, $4N$ analysis	0.994	0.998	1.005	0.99
Global with calorimetry	1.000	1.0025	1.005	0.98
True value	–	*1.000*	–	–

[a]Confidence limits at 68% confidence level
[b]Derived value, calculated from Eq. 3 using $\Delta H_{R-T} = 42.8$ kcal/mol and $\Delta S_{R-T} = 146$ cal/mol/deg (the best fitted values)
[c]Single curve analysis with quadruple number of experimental points
[d]Derived value, calculated from $\Delta H_{R-T} = 43.998$ kcal/mol and $\Delta S_{R-T} = 150.03$ cal/mol/deg (the best fitted values)
[e]Calculated from Eq. 3 using the primary parameters $\Delta H_{R-T} = 44$ kcal/mol and $\Delta S_{R\text{-}T} = 150$ cal/mol/deg

The global analysis is conceptually simple. Instead of minimization χ^2_k, $k = 1, 2, \ldots, K$, separately for each curve form the set of K curves, a "global" reduced $\chi^2_{R,\,GLOB}$ is minimized:

$$\chi^2_{R,\,Glob} = \frac{1}{(N-m-1)} \cdot \sum_{k=1}^{K} \chi^2_k = \frac{1}{(N-m-1)} \cdot \sum_{k=1}^{K} \sum_{i=1}^{M_k} \left[\frac{\Upsilon_{Data,k}(i) - \Upsilon_{Model,k}(i)}{\sigma_k(i)} \right]^2 \tag{4}$$

where N is the overall number of experimental points in all K analyzed curves, each of them containing $M_k$points, and m is the number of adjusted model parameters (see Note 6):

$$N = \sum_{k=1}^{K} M_k \tag{5}$$

A successful global fit requires all curves to be simultaneously well-fitted with one unique set of parameter values. Strict

parameter linkages between different curves prevent inappropriate models to fit simultaneously all curves. On the contrary, with a correct model the parameter values will be determined with better accuracy (see Note 7).

2.6. Global Analysis of Multiple Enzyme Kinetics

To demonstrate the power of global analysis, we add to our reaction kinetics at 40°C, Fig. 4, additional three RMPK kinetics "measured" at 20, 30, and 50°C. The resulting set of curves is shown in Fig. 7. It is seen that as temperature increases, shape of the kinetic curves changes from near-hyperbolic to sigmoidal. One potential interpretation is that the effect is caused by a temperature-induced shift of the equilibrium between the R and T-state toward the inactive T-state, as shown in Fig. 3 (see Note 8). Recognizing this fact, we can set the first linkage between the curves. It is the temperature dependence of the equilibrium constant L. According to Eq. 3, we can calculate L at 20, 30, 40, and 50°C:

$$L(20^{\circ}\mathrm{C}) = \exp\left(-\frac{\Delta H_{R-T} - (273.15+20)\cdot\Delta S_{R-T}}{(273.15+20)R}\right) \qquad (6)$$

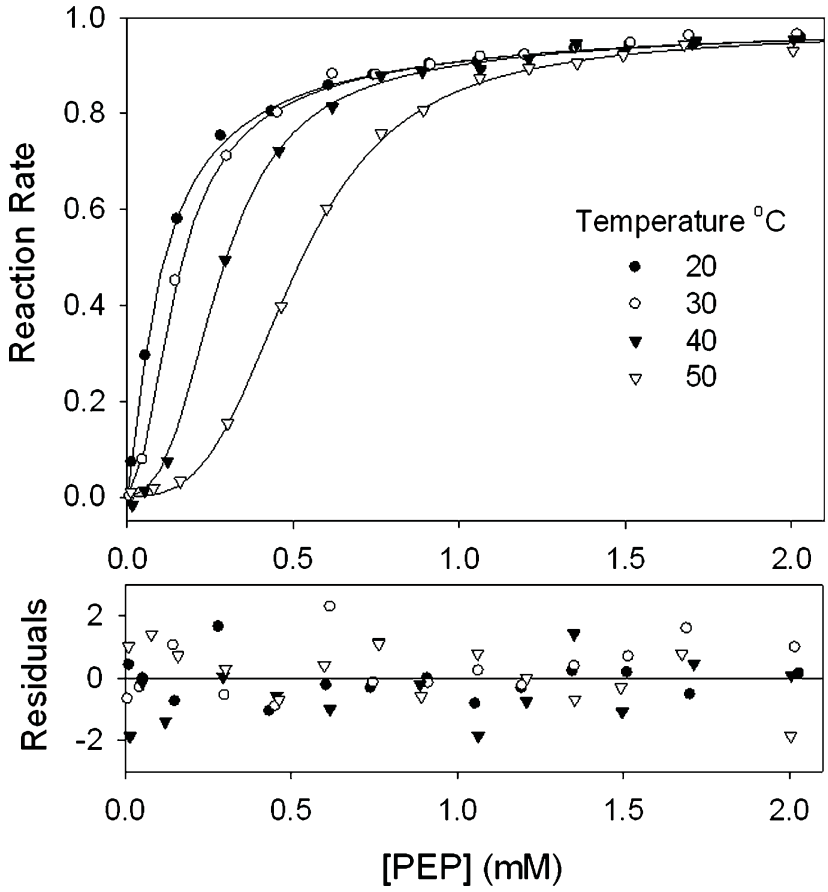

Fig. 7. Global analysis of steady-state reaction kinetics at different temperatures. Standard deviation of the added Gaussian noise is 0.01. The *solid lines* represent the best fit.

$$L(30^{\circ}\mathrm{C}) = \exp\left(-\frac{\Delta H_{R-T} - (273.15+30)\cdot\Delta S_{R-T}}{(273.15+30)R}\right) \quad (7)$$

$$L(40^{\circ}\mathrm{C}) = \exp\left(-\frac{\Delta H_{R-T} - (273.15+40)\cdot\Delta S_{R-T}}{(273.15+40)R}\right) \quad (8)$$

$$L(50^{\circ}\mathrm{C}) = \exp\left(-\frac{\Delta H_{R-T} - (273.15+50)\cdot\Delta S_{R-T}}{(273.15+50)R}\right) \quad (9)$$

At any temperature L can be expressed by two new parameters, ΔH_{R-T} and ΔS_{R-T}. Instead of four independent constants L, it is enough to fit ΔH_{R-T} and ΔS_{R-T}. As a consequence, the number of required model parameters decreases. Similar link can be made for k_{R} and K_{SR}. In real experiments, we would need to use equations similar to Eqs. 6–9 for temperature dependence of K_{SR} and the Arrhenius equation for temperature dependence of the rate constants k_{R}. In this simplified case, however, we neglected the temperature dependence as follows:

$$k_{\mathrm{R}}(20^{\circ}\mathrm{C}) = k_{\mathrm{R}}(30^{\circ}\mathrm{C}) = k_{\mathrm{R}}(40^{\circ}\mathrm{C}) = k_{\mathrm{R}}(50^{\circ}\mathrm{C}) = k_{\mathrm{R}} \quad (10)$$

and

$$K_{\mathrm{SR}}(20^{\circ}\mathrm{C}) = K_{\mathrm{SR}}(30^{\circ}\mathrm{C}) = K_{\mathrm{SR}}(40^{\circ}\mathrm{C}) = K_{\mathrm{SR}}(50^{\circ}\mathrm{C}) = K_{\mathrm{SR}} \quad (11)$$

In such a case, dimensionality of the NLSQ minimization problem is significantly reduced. Instead of 12 parameters required for independent fitting of the 4 individual curves, we need only 4 parameters, k_{R}, K_{SR}, ΔH_{R-T}, and ΔS_{R-T}, to describe all curves globally under a single model, Eq. 2. The more curves included in the global dataset, the higher the reduction of the overall number of parameters. Values of ΔH_{R-T} and ΔS_{R-T} extracted from the fit allow later reconstruction of L at any desired temperature according to Eq. 3.

The best global fit of the four enzyme kinetic curves is depicted in Fig. 7 by solid lines. Values of the corresponding parameters are summarized in Table 1. The primary adjusted parameters were ΔH_{R-T} and ΔS_{R-T} and the value of L was derived using Eq. 3. The fit yielded $\Delta H_{R-T} = 42.8$ kcal/mol and, $\Delta S_{R-T} = 146$ cal/mol/deg. Next we performed the same MC correlation and confidence analysis as described in Subheading 4. Results of 1,500 MC runs are shown in Figs. 8 and 9. In the correlation graphs, L was calculated from pairs of ΔH_{R-T} and ΔS_{R-T} as described (see Note 9). Besides the best parameter, values were closer to the expected ones, see Table 1, both the correlation graphs and the histograms indicated that addition of a few kinetic curves, i.e.,

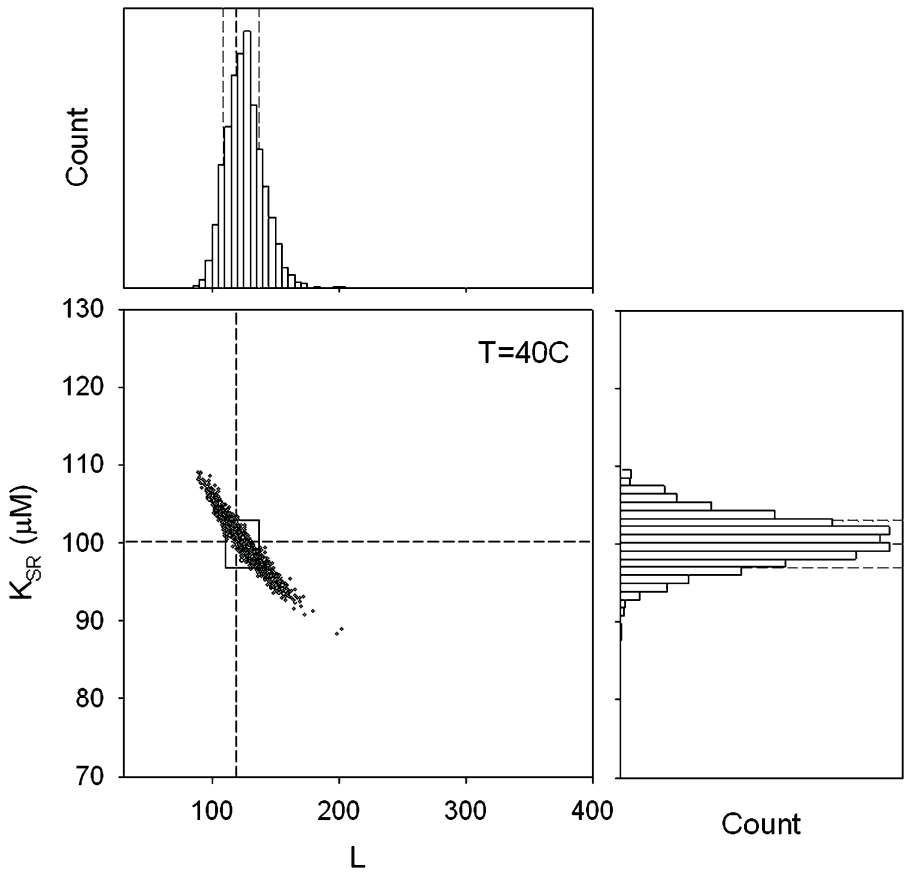

Fig. 8. Correlation between *L* and K_{SR} for global analysis of enzyme kinetics from Fig. 7. The correlation graph was obtained by 1,500 MC runs. *Black dashed lines* indicate original parameter values used for simulation. Histograms show resulting distribution of the fitted parameter values. *Gray dashed lines* border the region of ±1 SD. This region is depicted by the *black rectangle* in the central graph.

curves of the same character, to the dataset had a significant effect on accuracy and confidence intervals of all parameters. From a comparison of Figs. 5 and 8 and Figs. 6 and 9, it becomes obvious that although parameters are still correlated to some extent, the correlation is significantly reduced. In Table 1 we have also included results of a single curve fitting with quadruple number of data points that is comparable with number of data points used in the global analysis. As expected the confidence limits reduced, however, the reduction is inferior to the result of the global analysis. Addition of more curves of the same characters is but one of the first strategies to improve the accuracy of results and the limitation is due to the fact that these curves are defined by the same parameters. More significant improvement can be achieved by including data derived by alternative approaches that rely on different physical principles to monitor the changes, as shown in Subheading 7.

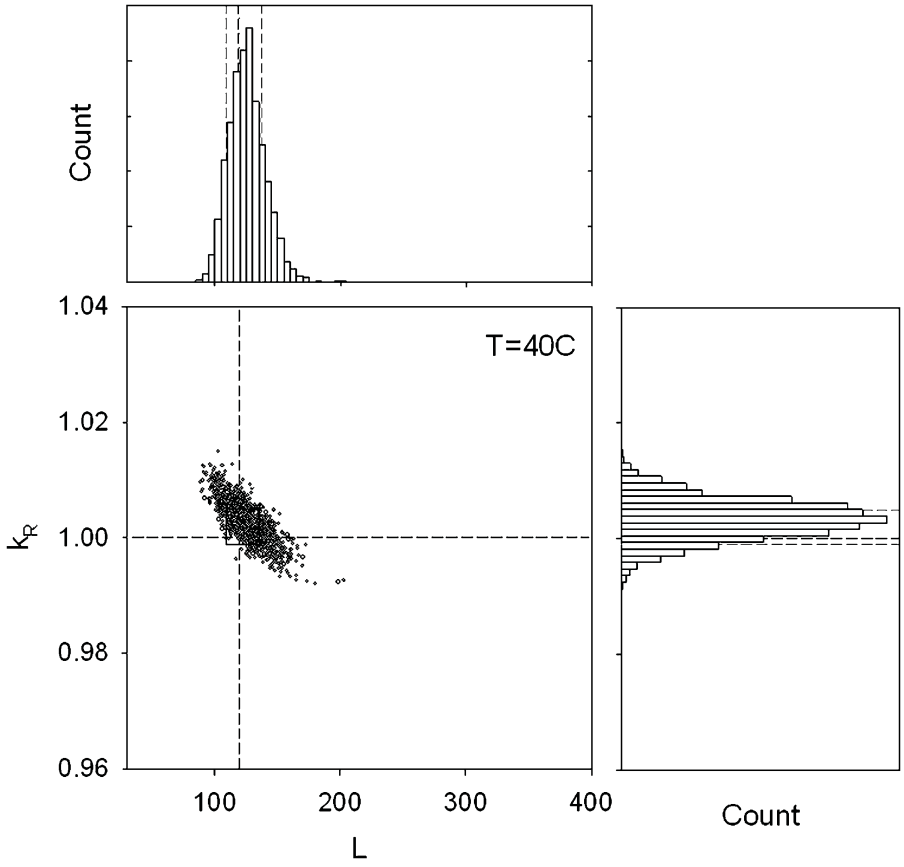

Fig. 9. Correlation between L and k_R for global analysis of enzyme kinetics from Fig. 7. The correlation was obtained by 1,500 MC runs. *Black dashed lines* indicate original parameter values used for simulation. Histograms show resulting distribution of the fitted parameter values. *Gray dashed lines* border the region of ± 1 SD. This region is depicted by the *black rectangle* in the central graph.

2.7. Global Analysis of Data with Different Character

Many different strategies can be employed to improve the accuracy of parameters. One can lower the noise of the data by collecting more precise measurements or using more sensitive approaches. More similar curves can be included in the dataset for global analysis. One could also extend the temperature range to further modulate the free energy change $\Delta G = \Delta H_{R-T} - T \cdot \Delta S_{R-T}$ and partial contributions of ΔH_{R-T} and ΔS_{R-T} to L. By rearrangement of Eq. 3 we obtain:

$$L = \exp\left(-\frac{1}{R} \cdot \left(\frac{\Delta H_{R-T}}{T} - \Delta S_{R-T}\right)\right) \qquad (12)$$

Upon examining the relative contributions of ΔH_{R-T} and ΔS_{R-T} to L in Eq. 12, it is clear that at higher temperature the contribution of ΔS_{R-T} increases. If experiments were conducted at a single temperature only, one would not be able to extract ΔH_{R-T} and ΔS_{R-T} because the ΔG, not particular values of ΔH_{R-T} and ΔS_{R-T}, is important for quantification of L. Experimental temperature range, however, cannot be arbitrarily extended. At low temperature,

we are limited by the freezing point of water and at high temperature by irreversible thermal processes degrading the enzyme structure. Using multiple curves of the same type (e.g., enzyme kinetics) has another inherent limitation. Mathematical description of all curves generated by the same approach has the same functional form and contains exactly the same parameters, Eqs. 2 and 3. For each curve, the correlation between parameters is similar. Significant improvement could be expected when a dataset includes data derived from a different approach in which only a limited subset of the globally linked parameters was needed to represent the data. When it is not necessary to include specific parameters in the mathematical description of the individual experimental curve, correlation among parameters would be reduced when multiple curves are analyzed. The extent of the effect depends on the "weight" that the curve has within the whole global dataset (see Note 10). The kind of experimental data to be included to the global fitting depends on the judicious choice of approach by the researcher.

In our next example, we include to the global fitting simulated data from isothermal titration calorimetry (ITC). Expression of these data naturally lacks the reaction rate constant $k_{R,}$ since the titration of RMPK by PEP is done in the absence of the second substrate ADP. Having in mind the rationale mentioned above, i.e., reduction of the parameter number, we do not analyze shapes of individual ITC curves because they obviously depend on the binding constant K_{SR}. Instead, we titrate RMPK with PEP to saturation and measure overall reaction heat as a function of temperature.

According to the two-state model, three major reactions contribute to the overall reaction heat Q_{tot} accompanying saturation of the enzyme by PEP. In particular, it is Q_R and Q_T that represent binding heats of PEP to the R and T state, respectively, and a heat Q_{R-T} accompanying the R → T (see Note 11):

$$Q_{tot} = f_{R,sat} \cdot Q_R + f_{T,sat} \cdot Q_T + (f_{T,sat} - f_T) \cdot Q_{R-T} \qquad (13)$$

Terms $f_{R,sat}$ and $f_{T,sat}$are fractions of RMPK in the R and T state, respectively, at the end of the titration when the enzyme is completely saturated by PEP. The term f_T is a fraction of the unliganded T-state at the beginning of the titration. Neglecting linked proton reactions, the heats can be expressed as:

$$Q_{state} = 4 M_0 V_0 \cdot \Delta H_{state} \qquad (14)$$

and

$$Q_{R \to T} = M_0 V_0 \cdot \Delta H_{R \to T} \qquad (15)$$

where M_0 stands for molar concentration of RMPK and V_0 is a sample volume. The enthalpies ΔH_{state} and ΔH_{R-T} are calculated

per mol of binding sites and per mol of tetrameric RMPK, respectively. The fraction f_T can be expressed as follows:

$$f_T = L/(1+L) \tag{16}$$

and

$$f_T + f_R = 1 \tag{17}$$

In our example we assumed that PEP does not bind to the T-state and $\Delta H_R = 0$. As a consequence, $f_{T,\mathrm{sat}} = 0$ for RMPK saturated by PEP. Then,

$$Q_{tot} = -f_T \cdot Q_{R-T} = -M_0 V_0 \cdot \frac{L}{1+L} \cdot \Delta H_{R-T} \tag{18}$$

$$\frac{Q_{tot}}{M_0 V_0} = -\frac{L}{1+L} \cdot \Delta H_{R-T}[\mathrm{kcal/mol}] \tag{19}$$

Equations 19 and 3 describe the desired temperature dependence of overall reaction heats accompanying saturation of RMPK by PEP. Data simulated according to these equations with added random noise (SD = 0.44 kcal/mol) are shown in Fig. 10.

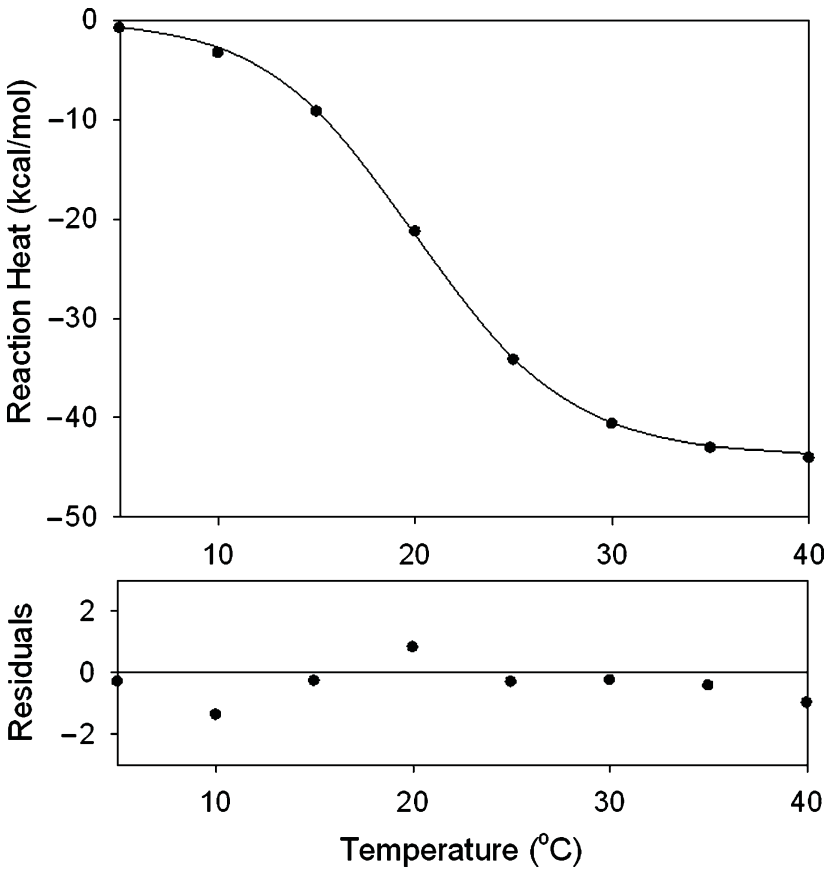

Fig. 10. Temperature dependence of overall reaction heat for PK titrated by PEP to the saturation. The *solid line* is a result of the best simultaneous fit of the calorimetric curve together with the set of 4 enzyme kinetics from Fig. 7. SD of noise is 0.44 kcal/mol.

This synthetic ITC curve depending only on ΔH_{R-T} and ΔS_{R-T} was used as the 5th dataset for the global analysis.

The result of the extended global fit is presented in Table 1, the best fit is shown in Fig. 10 by solid line. The fitted parameter values closely approached the true values used for the simulation, the ΔH_{R-T} and ΔS_{R-T} being 43.998 kcal/mol and 150.03 cal/mol/deg, respectively. Consequently, the value of equilibrium constant $L = 121$ (at 40°C) was determined with high accuracy. An outcome of the MC analysis is presented in Figs. 11 and 12, confidence limits in Table 1. Similar to the previous examples, 1,500 MC runs were performed for evaluation of correlation between parameters and confidence intervals. Figure 11 demonstrates a dramatic improvement of CI for both L and K_{SR} after addition of the ITC curve to the dataset, the highest effect being recorded for the equilibrium constant L. Although a correlation between L and K_{SR} can still be noticed from the direction of the main axis of the "ellipsoidal" scattergram, which is aligned with neither the x-axis nor the y-axis, the magnitude of the correlation

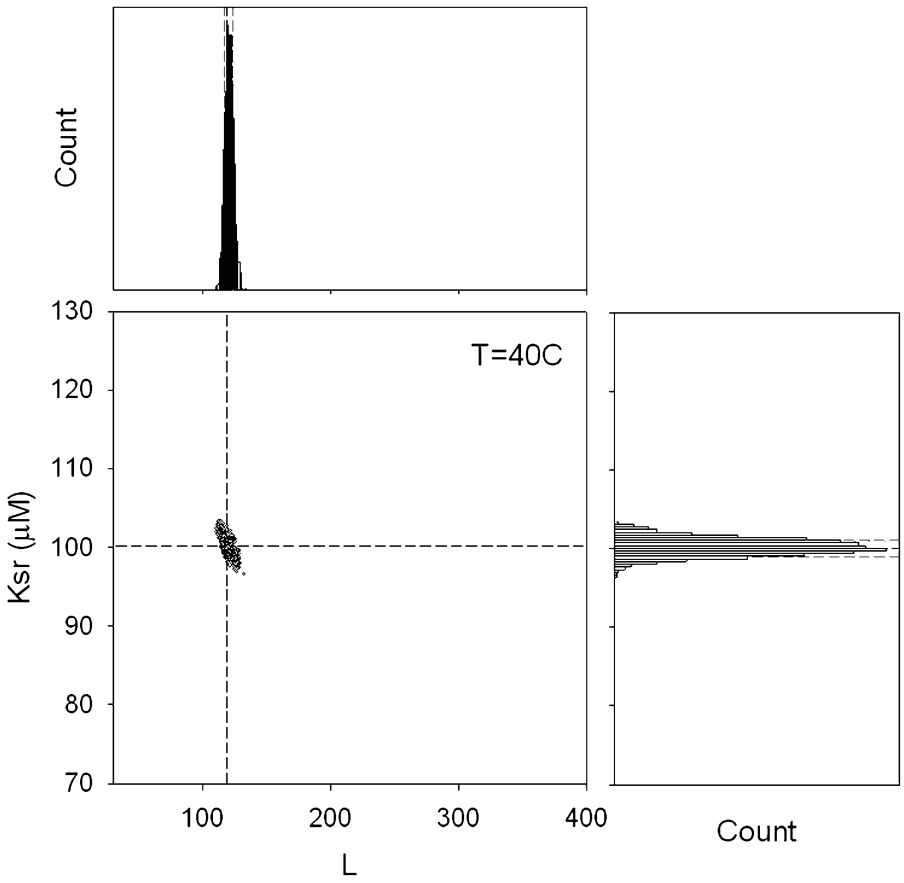

Fig. 11. Correlation between L and K_{SR} after extension of the dataset depicted in Fig. 7 by the calorimetric curve from Fig. 9. The correlation was obtained by 1,500 MC runs. *Black dashed lines* indicate original parameter values used for simulation. Histograms show resulting distribution of the fitted parameter values. *Gray dashed lines* border the region of ±1 SD. This region is depicted by the *black rectangle* in the central graph.

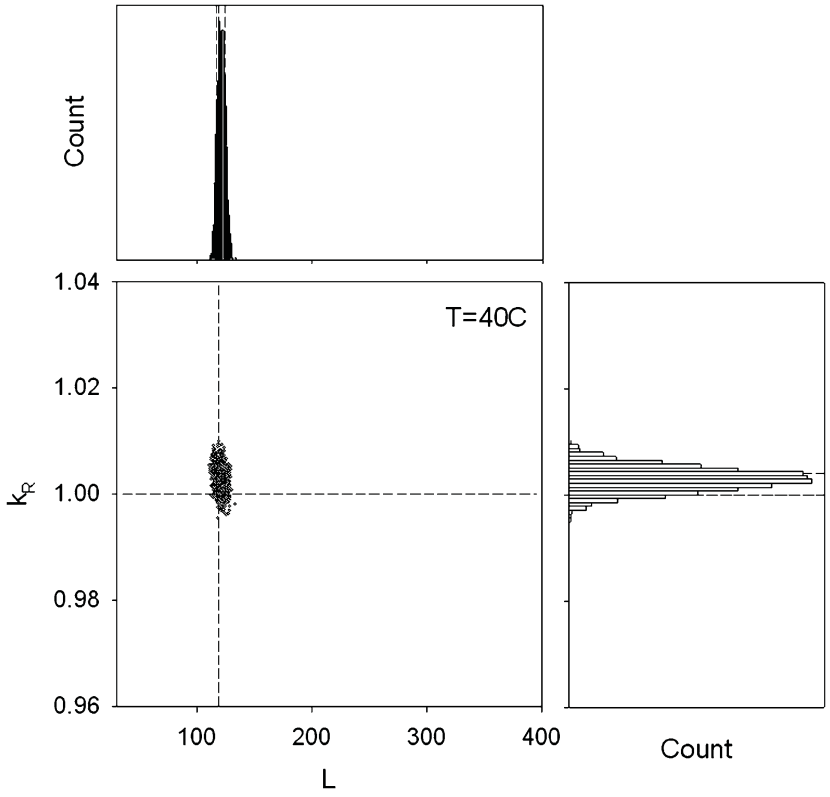

Fig. 12. Correlation between L and k_R after extension of the dataset depicted in Fig. 7 by the ITC curve from Fig. 10. The correlation was obtained by 1,500 MC runs. *Black dashed lines* indicate original parameter values used for simulation. Histograms show resulting distribution of the fitted parameter values. *Gray dashed lines* border the region of ± 1 SD. This region is depicted by the *black rectangle* in the central graph.

is much smaller. This can be visually judged from the shape and rotation of the scattergram in Fig. 11, as compared to Figs. 8 and 5. Another indication of significantly weaker correlation is the confidence area, depicted by the black rectangle in the central graph, being fully filled by the points generated during the MC analysis (compare to Figs. 8 and 5). Correlation between L and k_R has almost disappeared, as shown in Fig. 12. Last but not the least; the true values of parameter used for data generation reside within the narrow confidence intervals obtained from the MC analysis.

There are unlimited numbers of ways to achieve the same goal as expressed in Subheading 7. One may employ spectroscopic techniques, such as fluorescence, to monitor ligand binding. Spectroscopic techniques provide another set of information on the conformation changes in the protein as a consequence of ligand binding (23).

In conclusion, when properly selected, even addition of a single curve to the dataset can significantly improve accuracy and confidence limits of recovered parameters with accompanying enhancement of the model-testing power. The benefit of the analytical power of global fitting is realized particularly when data obtained under different experimental conditions or using different experimental approaches are included for the analysis.

3. Notes

1. Recent detail studies have shown that there is a considerable isothermal heat capacity change ΔCp associated with the R → T transition. The K_{SR} is slightly temperature dependent as well (22–24). For the demonstration purposes, however, we neglect the ΔCp as well as temperature dependence of K_{SR} in this chapter. Another simplification in our example rests in the assumption of the "silent" second substrate ADP that has to be present in the reaction mix to achieve enzyme activity. Specifically, we assume that ADP binds with the same affinity to the R and T-state and does not shift the R-T equilibrium. For this reason, we do not explicitly see terms associated with ADP in the equation for the normalized reaction rate, Eq. 2. In reality, the role of ADP is much more complex, see Herman and Lee (24).
2. In reality the temperature of 50°C might be too high for completing measurement due to the onset of irreversible thermal denaturation of RMPK in these experimental conditions.
3. The following conditions must be satisfied for valid application of the least-squares parameter fitting (28): (a) Errors of the measured data should be assignable to the dependent variable (the Y-axis) only. Errors caused by an inaccurate assignment of the independent variable (X-axis) should be negligible and uncorrelated with the measured values. This could be, for example, inaccurate setting or measurement of experimental temperature, pH, concentration, etc. (b) Experimental noise in measured data should be random and described by the Gaussian distribution. (c) Data should not contain any systematic error. (d) The fitted model must be correct. If specifically this condition was not satisfied, the recovered parameter values would be without physical meaning, even though the data might be fitted well.
4. The standard deviation (SD) σ_i required for the correct weighing of the i-th data point can be evaluated during the experiment. One possible way is to measure each data point many times, calculate the mean and standard error of the mean (SEM) (4). Then, the SEM can be used as σ_i. This approach is particularly useful during automatic or semiautomatic data acquisition when data points can be sampled many times. For "counting experiments" when signal is measured by counters (e.g., photon counting or radioactivity), the noise follows the Poissonian distribution and the SD is known from the theory: $\sigma_i = \sqrt{\mu}$, where μ is a mean value of the signal. In practice, however, the mean can be often substituted for the

number of counts N measured and $\sigma_i = \sqrt{N}$. This is a rigorous and preferable method for setting the weighing factors. In other cases the researcher has to rely on specifications of the instrument or perform own calibration of the data noise.

5. The parameter cross-correlation rises largely from the functional form of the fitted model. It can be modulated to some extent only by changing the span and distribution of the sampling points along the x-axis.
6. There are a number of multipurpose and specialized commercial software packages containing NLSQ minimization routines with a global analysis option. However, global analysis can be easily performed in any "open model" NLSQ program that allows user to input own fitting formulas. In this case, multiple experimental curves can be integrated to form one longer data file with appropriate input format. Then a variable is introduced to the model with the only purpose to tell the program what formula to use for fitting of each particular segment of the input data corresponding to the particular curve. The "fooled program" then automatically performs the desired global analysis. This approach is used in Sigma Plot (Systat Software, Inc.) to derive the examples used in this chapter.
7. Errors other than random noise in the data can invalidate the results derived from global fit. Therefore, the analyzed dataset must to be very *carefully inspected* for the presence of any abnormalities. Illegitimate errors, i.e., obvious mistakes and outliers, should not create problem, since they can be easily discovered, data discarded and experiments repeated. Fundamental problem can create systematic errors that are more difficult to identify and correct. Systematic errors limit accuracy and precision of experimental data by a defined and often reproducible way. This could be, for example, a wrong calibration of the instrument or experimental conditions different from the assumed ones (mismatch in temperature, pH, pressure, concentrations, sample aging, etc). Systematic errors result in inaccurate data, even though precision could be excellent. This is actually the worst situation because the correct model cannot fit the wrong data and low data noise makes the mismatch clearly visible. Researcher then can erroneously discard the correct model due to its "inconsistency with the experimental data". Therefore, careful attention should be paid to experimental design, analysis of experimental conditions and techniques used. Minimum requirement for a successful global analysis is to have an internally consistent dataset where all data curves are reproducible and measured preferably within short period of time on the same sample using the same experimental and sample handling

protocols. For identification of potentially biased data files one can analyze data subsets where individual curves or group of curves are excluded from the analysis. A consistent dataset should always yield similar parameter values and about the same quality of fit.

8. Based on the model in Fig. 2, at low temperature, almost all RMPK is in the active R-state (Fig. 3), and the steady-state kinetics follows a simple hyperbolic Michaelis–Menten equation. With increasing temperature the equilibrium between the R and T-state shifts toward the inactive T-state that is accompanied by increasing sigmoidicity of the reaction kinetics. During the PEP titration, the increasing amount of substrate, which binds preferentially to the active R-state, progressively shifts the inactive RMPK fraction to the active one with corresponding cooperative increase of the reaction rate. The effect is a consequence of homotropic allosteric interaction between four equivalent RMPK subunits causing all subunits to change conformation in a concerted manner during the T–R transition.
9. Advantage of the Monte Carlo analysis is a straightforward generation of confidence distributions and correlation graphs for quantities mathematically derived from the primary fitted parameters. Uncertainties of the constituent parameters naturally propagate to the derived quantities without need for their complex mapping.
10. In some substantiated cases, a "weighing" of the curve within the global dataset, i.e., contribution of the curve to the overall chi-squared, can be artificially increased by scaling the weighing factors $1/\sigma_k(i)$ in Eq. 4. Not to bias noise distribution within the curve, all $\sigma_k(i)$ should be scaled by the same factor. After this weighing, however, one cannot expect the reduced chi-squared of the best fit to approach unity.
11. In a real experiment, we would also need to account for buffer ionization heats due to proton absorption or release accompanying both PEP binding and R-T transition. For details, see for example ref. (22).

Acknowledgments

Supported by NIH GM 775551 and the Robert A. Welch Foundation (JCL) and grant MSM 0021620835 of the Ministry of Education Youth and Sports of the Czech Republic (PH).

References

1. Monod, J., Wyman, J., and Changeux, J. P. (1965) On the Nature of Allosteric Transitions: a Plausible Model, *J Mol Biol 12*, 88–118.
2. Koshland, D. E., Jr., Nemethy, G., and Filmer, D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits, *Biochemistry 5*, 365–385.
3. Eigen, M. (1967) Kinetics of reaction control and information transfer in enzymes and nucleic acids, in *Nobel Symposium* (Claesson, S., Ed.), pp 333–369, Almqvist & Wiksell, Stockholm.
4. Bevington, P. R., and Robinson, D. K. (2002) *Data reduction and error analysis for the physical sciences*, 3 rd ed., Mc Graw-Hill, New York.
5. Eisenfeld, J., and Ford, C. C. (1979) A systems-theory approach to the analysis of multiexponential fluorescence decay, *Biophys J 26*, 73–83.
6. Knutson, J. R., Beechem, J. M., and Brand, L. (1983) Simultaneous Analysis of Multiple Fluorescence Decay Curves - a Global Approach, *Chemical Physics Letters 102*, 501–507.
7. Beechem, J. M., Ameloot, M., and Brand, L. (1985) Global and target analysis of complex decay phenomena, *Analytical Instrumentation 14*, 379–402.
8. Ackers, G. K., Johnson, M. L., Mills, F. C., Halvorson, H. R., and Shapiro, S. (1975) The linkage between oxygenation and subunit dissociation in human hemoglobin. Consequences for the analysis of oxygenation curves, *Biochemistry 14*, 5128–5134.
9. Johnson, M. L., Correia, J. J., Yphantis, D. A., and Halvorson, H. R. (1981) Analysis of data from the analytical ultracentrifuge by nonlinear least-squares techniques, *Biophys J 36*, 575–588.
10. Barisas, B. G., and Gill, S. J. (1979) Thermodynamic analysis of carbon monoxide binding by hemoglobin trout I, *Biophys Chem 9*, 235–244.
11. Gilbert, C. W. (1980) A vector method for non-linear least squares reconvolution and fitting analysis of polarized fluorescence decay data, in *Time-resolved fluorescence spectroscopy in biochemistry and biology (Nato ASI Series)* (Dale, R. E., and Cundall, R. B., Eds.), Plenum Press, New York.
12. Beechem, J. M., Knutson, J. R., Ross, J. B. A., Turner, B. W., and Brand, L. (1983) Global Resolution of Heterogeneous Decay by Phase Modulation Fluorometry - Mixtures and Proteins, *Biochemistry 22*, 6054–6058.
13. Oberfelder, R. W., Barisas, B. G., and Lee, J. C. (1984) Thermodynamic linkages in rabbit muscle pyruvate kinase: analysis of experimental data by a two-state model, *Biochemistry 23*, 3822–3826.
14. Consler, T. G., Woodard, S. H., and Lee, J. C. (1989) Effects of primary sequence differences on the global structure and function of an enzyme: a study of pyruvate kinase isozymes, *Biochemistry 28*, 8756–8764.
15. Consler, T. G., Jennewein, M. J., Cai, G. Z., and Lee, J. C. (1992) Energetics of Allosteric Regulation in Muscle Pyruvate Kinase, *Biochemistry 31*, 7870–7878.
16. Consler, T. G., Jennewein, M. J., Cai, G. Z., and Lee, J. C. (1990) Synergistic effects of proton and phenylalanine on the regulation of muscle pyruvate kinase, *Biochemistry 29*, 10765–10771.
17. Boo, B. H., and Kang, D. (2005) Global and target analysis of time-resolved fluorescence spectra of Di-9 H-fluoren-9-yldimethylsilane: dynamics and energetics for intramolecular excimer formation, *J Phys Chem A 109*, 4280–4284.
18. Ionescu, R. M., and Eftink, M. R. (1997) Global analysis of the acid-induced and urea-induced unfolding of staphylococcal nuclease and two of its variants, *Biochemistry 36*, 1129–1140.
19. Ucci, J. W., and Cole, J. L. (2004) Global analysis of non-specific protein-nucleic interactions by sedimentation equilibrium, *Biophys Chem 108*, 127–140.
20. Verveer, P. J., Squire, A., and Bastiaens, P. I. (2000) Global analysis of fluorescence lifetime imaging microscopy data, *Biophys J 78*, 2127–2137.
21. Bednarkiewicz, A., and Whelan, M. P. (2008) Global analysis of microscopic fluorescence lifetime images using spectral segmentation and a digital micromirror spatial illuminator, *J Biomed Opt 13*, 041316.
22. Herman, P., and Lee, J. C. (2009) Functional energetic landscape in the allosteric regulation of muscle pyruvate kinase. 1. Calorimetric study, *Biochemistry 48*, 9448–9455.
23. Herman, P., and Lee, J. C. (2009) Functional energetic landscape in the allosteric regulation of muscle pyruvate kinase. 2. Fluorescence study, *Biochemistry 48*, 9456–9465.
24. Herman, P., and Lee, J. C. (2009) Functional energetic landscape in the allosteric regulation of muscle pyruvate kinase. 3. Mechanism, *Biochemistry 48*, 9466–9470.

25. Larsen, T. M., Laughlin, L. T., Holden, H. M., Rayment, I., and Reed, G. H. (1994) Structure of rabbit muscle pyruvate kinase complexed with Mn2+, K+, and pyruvate, *Biochemistry 33*, 6301–6309.
26. Oberfelder, R. W., Lee, L. L., and Lee, J. C. (1984) Thermodynamic linkages in rabbit muscle pyruvate kinase: kinetic, equilibrium, and structural studies, *Biochemistry 23*, 3813–3821.
27. Johnson, M. L., and Faunt, L. M. (1992) Parameter estimation by least-squares methods, *Methods Enzymol 210*, 1–37.
28. Johnson, M. L., and Frasier, S. G. (1985) Nonlinear Least-Squares Analysis, in *Methods in Enzymology 117* (Hirs, C. H. W., and Timasheff, S. N., Eds.), pp 301–342.
29. Straume, M., and Johnson, M. L. (1992) Monte Carlo method for determining complete confidence probability distributions of estimated model parameters, *Methods Enzymol 210*, 117–129.
30. Johnson, M. L. (1994) Use of Least-Squares Techniques in Biochemistry, in *Methods in Enzymology, vol. 240* (Abelson, J. N., Simon, M. I., Johnson, M. L., and Brand, L., Eds.), New York.
31. Davison, A. C., and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application* Cambridge University Press, Cambridge.