Pure & Appl. Chem., Vol. 65, No. 7, pp. 1527–1532, 1993.
Printed in Great Britain.

Data Treatment: Considerations when applying binding reaction data to a model

Dan Hallén

Division of Thermochemistry, Chemical Center, University of Lund, P.O. Box 124, S-221 00 Lund, Sweden.

Abstract

A discussion is held to emphazise the importance of performing simulation experiments in order to optimise the experimental situation for data treatment of binding curves. The optimatisation is done to decrease the influence of covariance contribution to the estimated errors of the fitted parameters. This is especially important when the model contains parameters that are linearly correlated.

INTRODUCTION

Thermodynamic data on macrocyclic compounds play an essential role for our understanding of their properties in solution and as ligands. Together with structural and dynamic properties the thermodynamics give us information relevant to the systemisation of the properties of the ligands which should then allow the synthesis of tailor-made ligands for specific applications. The aim of the thermodynamic study is to obtain precise stability constants and other thermodynamic properties such as enthalpies, entropies, and, especially for aqueous solutions, heat capacities for the complex formation. Calorimetry is in this respect a useful and accurate method to obtain all these thermodynamic properties. Irrespective of the method used for the binding experiment the raw data have to be treated in a regression procedure to obtain the stability constant(s). As for all physical properties, the values of the parameters obtained from the regression are without any value if no statistical uncertainties are reported together with the parameters. The literature contains more or less correct methods of estimating the uncertainties of the parameters, but there are statistical methods to estimate these values. It is of crucial importance that the experiments are performed to minimise the uncertainties of the parameters obtained from the regression. I will in this paper discuss the importance of optimising the experimental set-up for treatment of experimental data from binding studies. The examples used often refer to calorimetric titration techniques, but the general tendencies and conclusions made in this paper are valid for any experimental method.

REGRESSION PROCEDURES

The equilibrium constant(s) from binding experiment are obtained from regression of raw data to a function describing the model we want to use to rationalise our data. In all regressions the aim is to find the minimum of the χ^2 -function, which is defined as,

$$\chi^2 \equiv \Sigma \frac{(y_i - f_i)}{\sigma_i^2} \qquad (1)$$

Where y_i is the experimental value of point *i*, f_i is the fitted value of point *i* calculated from the function representing our model, and σ_i^2 is the variance of point *i* or the weighting factor of point *i*. There are two different forms of regression functions: functions that are linear with respect to the fitting parameters, and functions that are non-linear with respect to the parameters. The general form of a linear function can be written as,

$$f \equiv A_1X_1 + A_2X_2 + ... + A_nX_n \qquad (2)$$

In eq.2 A_k are the fitting parameters, and X_k are the independent variables. The function is linear because the differential of the function with respect to any fitting parameter is equal to the analogous independent variable.

$$\frac{\delta f}{\delta A_k} = X_k \tag{3}$$

An analogous description of a non-linear function is,

$$f = f(A_1, A_2, \ldots, A_n; X_1, X_2, \ldots, X_n) \tag{4}$$

The differential of the function with respect to any of the parameters result in new functions containing some or all parameters from the original function,

$$\frac{\delta f}{\delta A_k} = g(A_1, A_2, \ldots, A_n; X_1, X_2, \ldots, X_n) \tag{5}$$

There is an analytical solution for the linear regression function regarding both the values of the parameters as well as the error assignment. For the non-linear regression function the solution is obtained by numerical iterative methods. The functions used to obtain equilibrium constants are always non-linear with respect to the the fitting parameters.
The errors of the parameters are calcualted from the diagonal of the variance-covariance matrix, **E**, which is defined as the inverse of the curvature matrix, **C**,

$$\mathbf{E}=\mathbf{C}^{-1} \tag{6}$$

The elements of the curvature matrix is defined as,

$$\mathbf{C}_{kl} = \frac{\delta^2 \chi^2}{\delta A_k \delta A_l} \cong \sum \frac{\delta f_i}{\delta A_k} \frac{\delta f_i}{\delta A_l} \tag{7}$$

The summation on the right-hand side of eq.7 is correct for the linear regression function, while it is an approximation for the non-linear regression function that is valid near the minimum of the χ^2 -function. From the elements of the variance-covariance matrix a correlation matrix, **G**, is defined,

$$g_{kl} = \frac{e_{kl}}{\sqrt{e_{kk} e_{ll}}} \tag{8}$$

The correlation matrix contains important *qualitative* information about the correlation between the parameters. The value of the elements ranges from -1 to 1, where $|g_{kl}|=1$ describes maximum correlation. The diagonal elements are by definition unity, and the off-diagonal elements give us information about covariance contribution to the errors on the parameters. For the binding experiment it is therefor important to select concentrations and volumes for minimising the covariance contribution.
At titration experiments the signal detected at each step is proportional to the number of mol of complex(es) formed at the titration step. The data can in principle be treated in two ways: either treated as differential data or accumulated data of the differential signals. We can in most cases assume that the absolute uncertainty of the detected differential signal is independent on the magnitude of the signal. For differential treatment of data this will simplify eq. 1,

$$\chi^2 = \frac{1}{\sigma^2} \sum (y_i - f_i)^2 \tag{9}$$

This means that all points are equally weighted. In contrary, if the data are accumulated the variance term in eq. 1 will not be the same for all points due to propagation of error. The variance of point *i* is then,

$$\sigma_i^2 = i \cdot \sigma^2 \tag{10}$$

Using accumulated as input data in regressions will result in degrading of the information originally obtained from the experiment. The data points of most important interest for the resolution of the parameter(s) are not as much weighted as if the same raw data are treated as differential signals. The effects of treating the data in as the differential signals or accumulated is valid for any technique used. In the discussion I will only consider differential measurements.

ONE PARAMETER FIT

A one parameter fit of a binding curve is done using a technique where only the equlibrium constant can be calculated (spectroscopy, potentiometry, etc.). The model is then based upon the assumption that there is only one complex formed,

$$M + L = ML \tag{11}$$

where

$$K = \frac{[ML]}{[M][L]} = \frac{\alpha_M}{(1-\alpha_M)(C_L - \alpha_M)} \quad ; \quad \alpha_M = \frac{[ML]}{C_M} \tag{12}$$

The total concentration of L is , C_L, and the total concentration of M is C_M. We can define a new constant, D

$$D = KC_M = \frac{\alpha_M}{(1-\alpha_M)(r-\alpha_M)} \tag{13}$$

The variable r in eq. 13 is the ratio of total number of mol of L and M ($r = n_{L,tot} / n_{M,tot}$). At step-wise titration the increment of r , Δr, at each step is normally constant throughout the titration serie. This means that the parameters that we can change to optimise the experiment are D and Δr. It has been shown by number of authors that D should be in the range of $1 \leq D \leq 1000$ (ref. 1-2). In Fig. 1 the relative amount of the titrand, L, that binds to the host molecule. M, at each step is plotted against the number of injections for different choices of Δr, and D=1and 100.

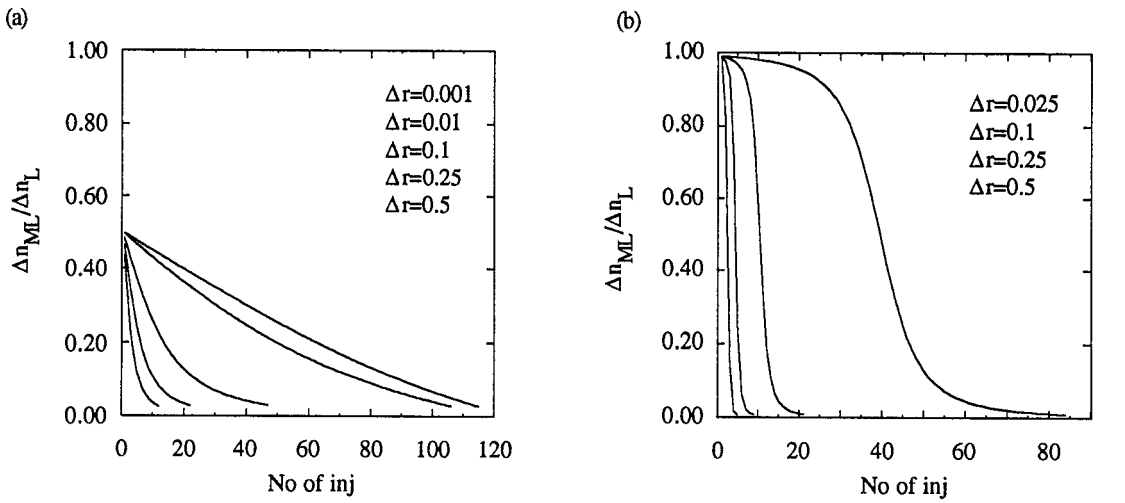

Fig. 1. The relative amount of the ligand added to a solution that is bound to a host molecule at each titration step, $\Delta n_{ML}/\Delta n_L$, for different Δr when (a) D=1 and (b) D=100.

The shape of the binding curve will directly effect the quality of the regression. We know from experience that we will obtain good results if the there is an inflection point in the binding curve. With good results means normally that the fitting parameter is well defined. The estimated uncertainty of K depends on the choice of Δr for a given value of D. This is illustrated in Fig. 2 where χ^2 is plotted against D for different Δr. The curves was genereated by simulating one serie of a binding curve where and D=1 for each Δr. All series used the same error assignment on the signal which randomly varied within 0.1 %. As can be seen from the plot the minimum is not well defined if we use Δr=0.01. The χ^2 -function is very flat which means that there is a great risk of obtaining systematic errors on the parameter. Furthermore, the shape of the curve effects directly the assigned error on K, since the error on K is calculated from the inverse of the second derivative of the χ^2-function with respect to K,

$$\sigma_K^2 = \frac{1}{(\delta^2\chi^2/\delta^2 K)} \tag{14}$$

The choice of Δr-parameter depends on the magnitude of D, and the range of Δr is smaller the larger D-value. This is illustrated on Fig. 3, where the Δr-range is plotted against log D. The plot shows the Δr-values that correspond to 12-25 titration steps in one series. For systems where the solubilty of one or both of the compounds are limited the quality of the regression will not improve significantly even if the technique used is improved with orders of magnitude. It also shows that it is not always benifitial to have many data points when fitting data to a function.

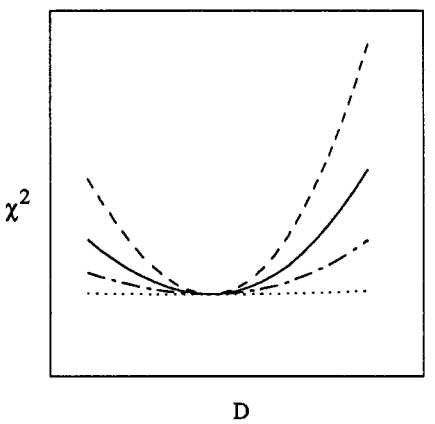

Fig. 2. The graph illustrates the shape of the χ2-function for different Δr when making a regression on one parameter. Series of binding curves have been generated with assigned errors on each point. All series have the same error assignment on the data points. The values of Δr are from top to bottom 0,2, 0.1, 0.05, and 0.01.

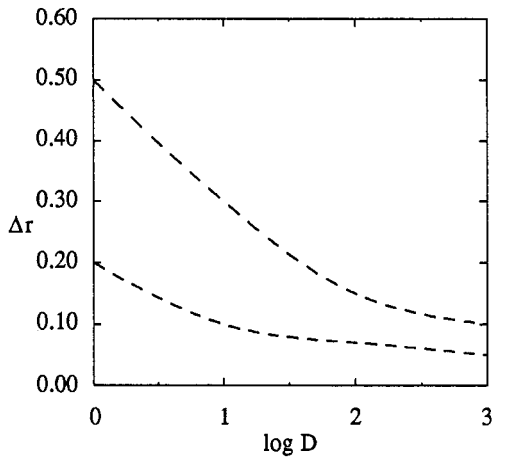

Fig. 3. The plot illustrates the range of choice for Δr for a given D. The ranges are analogous to 12-25 titration steps of the ligand to reach 95% saturation of the host molecule.

TWO PARAMETER FIT

If the model contains more than one parameter there is a risk that the covariance contribution is significant to the errors of the individual parameters. It is in principle impossible to have a situation where the covariance contribution is diminished, because the variance-covariance matrix will always be non-orthogonal. However, by optimising the experimental variables we can minimise the covariance contribution. From the correlation matrix we obtain qualitative information about the correlation between the fitting parameters. Is there limit of the off-diagonal element at which the covariance contribution is minimised? There is probably not a general answer to that, because that will depend on the fitting function. Here I will only show some examples on what effect the off-diagonal element of the correlation matrix from a two parameter fit has on the error assignment of the parameters. Calorimetric step-wise titration experiments are used to illustrate the effects.
In a calorimetric experiment the amount of heat measured at each step, q_i, is directly proportional to the amount of complex form at each step, Δn_{ML}. The proportionality constant is the enthalpy of complex formation, ΔH°,

$$q_i = \Delta H^\circ \Delta n_{ML} \tag{15}$$

where

$$n_{ML}=\alpha_M C_M \tag{16}$$

The equilibrium constant is calculated from the change in Δn_{ML} , that is curvature of the titration curve. The two parameters, ΔH° and K, are linear correlated, which means that in the function, q_i, there is a product of ΔH° and K, thus there will be covariance contribution to the assigned errors. In Fig. 4 and Fig. 5 the graphs show double-plots of the off-diagonal element in the correlation matrix, g_{12}, and the error on K obtained from the fit, σ_K, against Δr. Fig. 4 is based upon simulations of calorimetric experiments where K=1000, ΔH° =15kJ mol^{-1} , C_M =0.01M, vessel volume=1ml, and the mean standard deviation of one point, s_i , is 20μJ. The same series of assigned errors for each point were used for all the series for the different Δr. As shown the error on K flattens out when $|g_{12}|$ is close to 0.8. It should also be noted that even if g_{12} and e_{12} < 0 the result is an increase in σ_K. The same sort of simulations have been done with the exception that the

relative error of the first signal is the same for all the series. The results of the simulations are shown in Fig. 5. The reason for this approach is to show that if we improve our calorimetric technique with orders of magnitude there will not be any improvements of the results on the error estimates of K when using Δr in the experiment that is less favourable. The reason for this is that the general shape of the χ^2-function will not change even if we improve the precision on the experiments (cf. Fig. 2) and the parameters, K and ΔH°, are linear correlated.

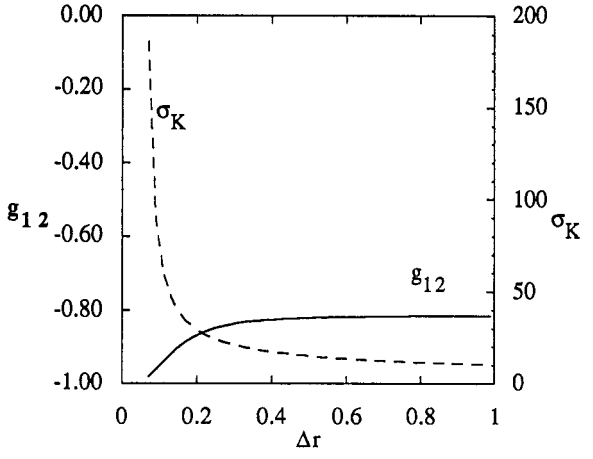

Fig. 4. Double plot of the off-diagonal element in the correlation matrix, g_{12}, and the estimated error on K, σ_K, as function of added increment of the ligand, Δr. Data has been generated assuming K=1000 M^{-1}, ΔH°=15 kJ mol^{-1}, C_M=0.1 M, V=1 ml. The data points in each series was $q_i+\varepsilon_i$, where q_i was calculated according to eq. 15 and ε_i was randomly varied within the range ±20μJ. The same assigned error on the data point was used for all the series.

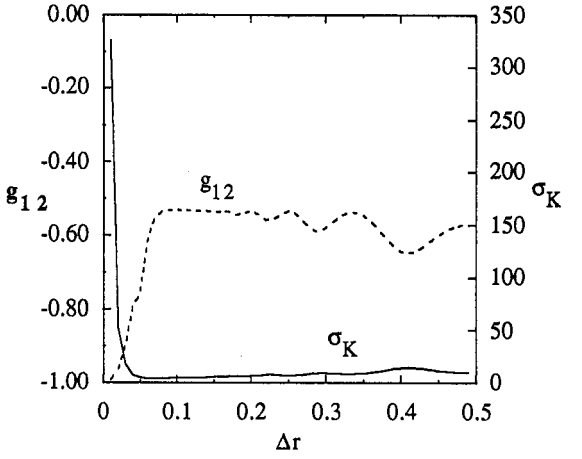

Fig. 5. Double plot of the off-diagonal element in the correlation matrix, g_{12}, and the estimated error on K, σ_K, as function of added increment of the ligand, Δr. Data has been generated assuming K=1000 M^{-1}, ΔH°=15 kJ mol^{-1}, C_M=0.1 M, V=1 ml. The data point in each series was $q_i+\varepsilon_i$, where q_i was calculated according to eq. 15 and ε_i was randomly varied within the range ±20μJ for Δr=0.5. For the other series the ranges for the error assignment were match so that the relative error were the same for all series.

Figure 4-5 shows the correlation of the parameters which is partly due to the experimental parameters, but also due to the function used in the regression (eq. 15). The function can be expressed in another way to avoid linear correlation and minimising the correlation between the parameters.

$$q_i = \Delta H^\circ(C_L + C_M) + R - \left[(\Delta H^\circ)^2(C_M - C_L)^2/4 + S(C_M + C_L) + R^2\right] \quad (17)$$

where $R=\Delta H^\circ/(2K)$ and $S=(\Delta H^\circ)^2/(2K)$.

If the model contains two complexes so that the total concentration of the host molecule can be expressed according to the Adair equation,

$$C_M = [M] + \beta_1[M][L] + \beta_2[M][L]^2 \quad (18)$$

the general features of error analysis are analogous with the case where the second parameter is linear such as the enthalpy. However, here the choice of C_M and Δr are governed by the magnitudes of both β_1 and β_2. It can be of advantage to perform the experiments at different C_M to have good resolution of the parameters in the regression. The resolutions of the parameters are determined by the difference in the relative amount of the species that contains the host molecule after each titration step (α_0=[ML], α_1=[ML]/CM, and α_2=[ML_2]/C_M). Systems that are limited by solubility the resolution of the parameters can be difficult to obtain due to the same effect that occur on a one parameter fit as discussed earlier. It is possible that the minimum χ^2-function is well defined for the first parameter, while it is less pronounced for the second parameter. The situation of large covariance contribution will in that case be severe. There will always be some contribution from the covariance in a non-linear regression since the data space is not orthogonal. Applying a model that contains more than one complex to calorimetric data will result in addition of enthalpy terms to each complex equilibrium, which will result in large covariance contributions on the errors of the fitted parameters. The reason is that there is no analytical way to avoid linear combinations of the enthalpies and the analogous equilibrium constants if the model contains more than one complex.

CONCLUSIONS

I have in this paper tried to illustrate the importance of performing simulation experiments for systems where we assume complex formation before doing the actual measurements where the aim is to obtain stability constant(s) of the complex(es) formed and other thermodynamic properties such as enthalpies. In the analysis of the data treatment the correlation matrix is of high interest to study, because it gives qualitative information about the correlation between the fitting parameters. Addition of linear parameters to the model, such as enthalpy, will induce strong correlation unless the function is expressed in a new way to avoid linear correlation. At step-wise titration the data should always be treated as differential data so all data points are equally weighted.

REFERENCES

1. D.J. Eatough, E.A. Lewis and L.D. Hansen, Analytical Solution Calorimetry, Ed. J.K. Grime, p.137-161, John Wiley & Sons, Inc., New York, (1985)
2. T. Wiseman, S. Willliston, J. Brandts and L. Lin, Anal. Biochem. 179, 131-137 (1989).