Determination of association constants from titration data and any system of equilibria

René Becker

April 21, 2020

1 Numerical calculation of equilibrium concentrations

Description of a system of chemical equilibria. To start, we have to describe our chemistry in a concise manner. Suppose we are titrating a *host* species H with a *guest* species G, and that H can bind two G molecules to form the assemblies HG and HGG in two consecutive binding steps, each with their own equilibria:

$$H + G \overset{S_1 K_a}{\rightleftharpoons} HG \tag{1}$$

$$HG + G \overset{S_2 K_a \alpha_1}{\rightleftharpoons} HGG \tag{2}$$

We can then formulate a set of non-linear equations that describe our system exhaustively. First, for each equilibrium in the system we add an equation that relates the equilibrium constant to the species concentrations:

$$S_1 K_a \cdot [H][G] - [HG] = 0 \tag{3}$$

$$S_2 K_a \alpha_1 \cdot [HG][G] - [HGG] = 0 \tag{4}$$

Next, for each *initial* species in the system (here H and G) we add a mass-balance equation:

$$[H] + [HG] + [HGG] - [H]_{initial} = 0 \tag{5}$$

$$[G] + [HG] + 2 \cdot [HGG] - [G]_{initial} = 0 \tag{6}$$

Solving these equations for the species concentrations, one obtains the species concentrations at chemical equilibrium.

Numeric *vs.* algebraic Solving for using algebraic expressions can be done for a limited number of systems (*cf.* Hunter's method). Solving for using numerical methods can be done for any system of equilibria (*cf.* the next paragraphs and sections) and is sufficiently fast to be used inside an optimization routine.

Matrix description. A general way of finding equations like (3) through (6) - *for any system that can be described in terms of elementary equilibria!* - can be set up by writing the system description in matrix form.

Let's define N_E as the number of equilibria (here: 2), N_S as the number of species (here: 4) and $N_{S,i}$ as the number of initial species (here: 2, H and G). First of all, we define the $(N_S \times 1)$ species concentration matrix $\boldsymbol{C}$ as:

$$\boldsymbol{C} = \begin{bmatrix}[H] & [G] & [HG] & [HGG]\end{bmatrix} \tag{7}$$

We store the initial concentrations (the amounts you add to the mixture at each titration point) in the $1 \times N_{S,i}$ matrix $\boldsymbol{C}_{initial}$:

$$\boldsymbol{C}_{initial} = \begin{bmatrix}[H]_{initial} \\ [G]_{initial}\end{bmatrix} \tag{8}$$

Then we define two $(N_S \times N_E)$ matrices $\boldsymbol{E}_{LHS}$ and $\boldsymbol{E}_{RHS}$, where we store the occurrence of a species in the left-hand side (LHS) and right-hand side (RHS) of the equilibria. So if species H occurs *once* in the LHS of (1), we put a 1 in $\boldsymbol{E}_{LHS}$ in row 1 at the column that belongs to species H, etc. In case the species occurs *twice* (e.g. dimerization equilibrium) we simply add a 2.

$$\boldsymbol{E}_{LHS} = \begin{bmatrix}1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0\end{bmatrix} \tag{9}$$

$$\boldsymbol{E}_{RHS} = \begin{bmatrix}0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1\end{bmatrix} \tag{10}$$

The association constants are put in the $(1 \times N_E)$ matrix $\boldsymbol{K}$:

$$\boldsymbol{K} = \begin{bmatrix}K_1 \\ K_2\end{bmatrix} = \begin{bmatrix}S_1 K_a \\ S_2 K_a \alpha_1\end{bmatrix} \tag{11}$$

Before we go into the next section, we need to define one more matrix containing information about the mass balance. Suppose we have an equilibrium somewhere in our system, where some assembly is split into two parts. From only the information we gathered so far, we can not know what is inside these parts, and as such we cannot assemble our mass balance equations without this knowledge. We define a $N_S \times N_{S,i}$ mass-balance matrix $\boldsymbol{M}$ that contains in each column the number of initial species that is contained in each species. For our system, that looks like this:

$$\boldsymbol{M} = \begin{bmatrix}1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2\end{bmatrix} \tag{12}$$

We can see that the left part forms a $\boldsymbol{I}_{N_{S,i}}$, because H contains one H and zero G, and G contains zero H and one G. Then, HG contains one H and one G (the third column) and HGG contains one H and two Gs (the fourth column).

Generating the equations from the matrices. To generate the equations from the matrices, we use the (Matlab-style) element-wise power operation $\boldsymbol{A}\hat{}\boldsymbol{B}$ which raises every element $\boldsymbol{A}_{ij}$ to the power $\boldsymbol{B}_{ij}$.

The element-wise power of the concentration matrix (expanded to N_E rows by multiplying it with the N_E identity matrix $\boldsymbol{I}_{N_E}$) with the $\boldsymbol{E}_{LHS}$ matrix yields

the left-hand side terms of the equilibrium equations. Taking the product over the rows yields a $(1 \times N_E)$ matrix:

$$\prod_{rows} (\boldsymbol{I}_{N_E} \boldsymbol{C}) \hat{\cdot} \boldsymbol{E}_{LHS} = \prod_{rows} \begin{bmatrix} [H] & [G] & 1 & 1 \\ 1 & [G] & [HG] & 1 \end{bmatrix} = \begin{bmatrix} [H][G] \\ [G][HG] \end{bmatrix} \tag{13}$$

The full set of equilibrium equations can thus be derived by element-wise multiplication ($\odot$) of this expression with $\boldsymbol{K}$ and subtracting from it the similar expression for the RHS:

$$\begin{aligned} 0 &= \prod_{rows} (\boldsymbol{I}_{N_E} \boldsymbol{C}) \hat{\cdot} \boldsymbol{E}_{LHS} \odot \boldsymbol{K} - \prod_{rows} (\boldsymbol{I}_{N_E} \boldsymbol{C}) \hat{\cdot} \boldsymbol{E}_{RHS} \\ &= \begin{bmatrix} [H][G] \\ [G][HG] \end{bmatrix} \odot \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} - \begin{bmatrix} [HG] \\ [HGG] \end{bmatrix} \\ &= \begin{bmatrix} [H][G] \cdot S_1 K_a - [HG] \\ [G][HG] \cdot S_2 K_a \alpha_1 - [HGG] \end{bmatrix} \end{aligned} \tag{14}$$

The mass-balance equations can then be generated by multiplication of the mass-balance matrix with the transposed concentration matrix, and subtraction of the initial concentrations of the initial species:

$$\begin{aligned} 0 &= \boldsymbol{M} \boldsymbol{C}^T - \boldsymbol{C}_{initial} \\ &= \begin{bmatrix} [H] + [HG] + [HGG] - [H]_{initial} \\ [G] + [HG] + 2 \cdot [HGG] - [G]_{initial} \end{bmatrix} \end{aligned} \tag{15}$$

We have now found a general method to find equilibrium equations (14) and mass-balance equations (15). This general method is implemented in a straightforward manner in computer code. The equations are the only information necessary to calculate the equilibrium concentrations, e.g. by using them inside a non-linear solver.

Example Python code. A Python implementation is given below. It assumes that the matrices $\boldsymbol{E}_{LHS}$, $\boldsymbol{E}_{RHS}$ and $\boldsymbol{M}$ are available in the target function, and that the matrices $\boldsymbol{C}_{initial}$ and $\boldsymbol{K}$ are available to be used in the `SciPy` `fsolve` function call.

```
def equations_to_solve(C, C_initial, K):
    return [
        *(np.prod( np.power(C, E_LHS), axis=1) * K \
          - np.prod( np.power(C, E_RHS), axis=1)),
        *(np.matmul(M, C) - C_initial)
    ]

C_equilibrated = fsolve( equations_to_solve,
                         x0=C_initial, args=(C_initial, K) )
```

The `NumPy` routines `matmul` and `power` internally take care of the expansion (broadcasting) and transposing of $\boldsymbol{C}$.

2 Fitting of experimental binding curves

Titration methods and observables. When titrating a solution (e.g. containing H) with another solution (e.g. containing both H and G), we observe a changing property which is usually measured by some analytical instrument. For example, a UV-vis spectrometer will give us a spectrum/absorption at each titration point, and a pH meter will give us a pH value at each titration point. The observables O) (spectrum/absorption, pH) are a function of the species in the equilibrium under study. Again for our example system:

$$O = f(H, G, HG, HGG) \tag{16}$$

The form of the function f is radically different for different measurements:

O	$f(\cdots)$
pH	$-\log[H]$
Light absorption @ λ	$\epsilon_{H,\lambda}[H] + \epsilon_{G,\lambda}[G] + \epsilon_{HG,\lambda}[HG] + \epsilon_{HGG,\lambda}[HGG]$
NMR @ nucleus h in H	$(\delta_{h,H}[H] + \delta_{h,HG}[HG] + \delta_{h,HGG}[HGG])/[H]_{initial}$
NMR @ nucleus g in G	$(\delta_{g,G}[G] + \delta_{g,HG}[HG] + 2\delta_{g,HGG}[HGG])/[G]_{initial}$
ITC	$\Delta_{HG}[HG] + \Delta_{HGG}[HGG]$

Only titrations where one of the species concentrations is directly observable (e.g. pH titrations) can be analyzed directly. Titration data where the observable is a function of two or more species concentrations can be analyzed by the method described in this document. A prerequisite is that this function is known, which is often not the case with e.g. potentiometric titrations.

If the function is known, all it takes to calculate the binding curve are the values of the unknown parameters (ϵ, δ, Δ, etc.) and the equilibrium concentrations of the species. Let's call the set of unknown parameters $\boldsymbol{\Lambda}$. From the previous section we know that we can calculate the equilibrium concentrations from the association constants $\boldsymbol{K}$ and the initial concentrations $\boldsymbol{C}_0$ ($\boldsymbol{C}_{initial}$, but expanded to each titration point). We can then define the binding curve as:

$$\boldsymbol{O}_{calc} = f(\boldsymbol{K}, \boldsymbol{C}_0, \boldsymbol{\Lambda}) \tag{17}$$

In a titration experiment:

- $\boldsymbol{C}_0$ is **known** but is **different** at each titration point
- $\boldsymbol{K}$ and $\boldsymbol{\Lambda}$ are **unknown** but are the **same** at each titration point

We introduce N_P as the number of titration points and N_O as the number of observables (e.g. the number of wavelengths in a UV-vis titration or the number of peaks in an NMR titration). We also introduce N_F as the number of unknown parameters in $\boldsymbol{\Lambda}$. The titration problem can then be described with the following matrices and their dimensions:

$$\begin{aligned} \boldsymbol{C}_0 &\quad N_{S_i} \times N_P \\ \boldsymbol{K} &\quad 1 \times N_E \\ \boldsymbol{\Lambda} &\quad N_O \times N_F \end{aligned}$$

Determination of unknown parameters by least-squares minimization. Given a set of experimental binding curves $\boldsymbol{O}_{exp}$, we can determine the optimal values for $\boldsymbol{K}$ and $\boldsymbol{\Lambda}$ by minimizing the residual sum of squares between calculated and experimental binding curves:

$$\begin{aligned} minimize RSS(\boldsymbol{K}, \boldsymbol{\Lambda}) &= \sum \left[\boldsymbol{O}_{exp} - \boldsymbol{O}_{calc}\right]^2 \\ &= \sum \left[\boldsymbol{O}_{exp} - f(\boldsymbol{K}, \boldsymbol{C}_0, \boldsymbol{\Lambda})\right]^2 \end{aligned} \tag{18}$$

The minimization routine itself and the interpretation of the quality of the fit falls outside the scope of this document and can be found in Thordarson's paper *Death of the Job Plot*.