Supplemental Methods

Sean E. McGeary

1 Determination of miRNA–target site $\boldsymbol{K}_{\mathrm{D}}$ values from Equilibrium AGO–RNA Bind-N-Seq

We determined K_{D} values for each miRNA-target site type relative to RNA molecules lacking a site by maximum likelihood estimation (MLE). This is a statistical method by which the parameters values $\boldsymbol{\theta}$ corresponding to a particular mathematical model are determined according to their greatest agreement with associated data. More formally, the element-wise values of $\boldsymbol{\theta}$ are chosen to maximize the log-likelihood function

$$\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) = \ln p(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{\theta})), \tag{1.1}$$

where $\ln p(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{\theta}))$ is the probability of observing the sequencing count data $\boldsymbol{y}$ given the vector of model-simulated abundances $\boldsymbol{x}(\boldsymbol{\theta})$ (itself a function of $\boldsymbol{\theta}$). We first describe the derivation of $\boldsymbol{x}(\boldsymbol{\theta})$, and then $f_{cost}(\boldsymbol{x})$, a cost function scaling monotonically with $\ln p(\boldsymbol{y}|\boldsymbol{x}(\boldsymbol{\theta}))$, and therefore having a minimum value coincident with the MLE parameter estimates. We then derive the gradient for the overall cost function

$$f_{grad}(\boldsymbol{\theta}) = \nabla f_{cost}(\boldsymbol{x}(\boldsymbol{\theta})) \tag{1.2}$$

allowing for the efficient, stable solution of large sets of (>50,000) of K_{D} values via the L-BFGS-B method (using the *optim* package in R).

2 Derivation of $\boldsymbol{x}(\boldsymbol{\theta})$

The function $\boldsymbol{x}(\boldsymbol{\theta})$ calculates $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$, the concentration of each site type–containing RNA i recovered within equilibrium binding reaction sample j, where $1 \leq i \leq n$ and $1 \leq j \leq 5$. This function takes as input, the K_{D} value corresponding to each site $\boldsymbol{K} = (K_1, K_2, \ldots, K_n)$, the total concentration of each of the site types $\boldsymbol{l} = (l_1, l_2, \ldots, l_n)$, the stock concentration (i.e. the concentration at which it was stored prior to dilution into an binding reaction) of the AGO–miRNA complex a (hereafter referred to as "AGO"), the concentration of library RNA recovered nonspecifically by nitrocellulose filter binding, b, and the integer j designating which sample within the dilution series is to be simulated. The vector of total target–site type concentrations are fixed prior to the optimization routine, and are given by

$$l_i = \frac{x_i^l}{\sum_{i'} x_{i'}^l} \times 100 \text{ nM}, \tag{2.1}$$

where x_i^l is the read count corresponding to site i in the input library. In order to allow the optimization routine to proceed unconstrained (i.e., to allow the parameter values to be anywhere in the range from $-\infty$ to $+\infty$, in addition to the range of orders of magnitudes the parameter values may occupy $\boldsymbol{\theta}$ is parameterized in terms of natural logarithm of the biochemical parameters used in the model function $\boldsymbol{x}(\boldsymbol{\theta})$, such that

$$\begin{aligned}
\theta_1 &= \ln K_1, \\
\theta_2 &= \ln K_2, \\
&\vdots \\
\theta_n &= \ln K_n, \\
\theta_{n+1} &= \ln a, \\
\theta_{n+2} &= \ln b.
\end{aligned}$$

The recovered concentration of site i in sample j is given by

$$x_{ij} = c_{ij} + g_{ij}, \tag{2.2}$$

where c_{ij} and g_{ij} are the concentration of the AGO–bound and nonspecifically recovered forms of the site, respectively. By making the assumption that only the unbound sites can be nonspecifically recovered in the binding reaction, and that the total concentration (and therefore amount) of nonspecific recovered library RNA b is constant across all five samples, x_{ij} can be written as

$$\begin{aligned}
x_{ij} &= c_{ij} + b\frac{l_i - c_{ij}}{\sum\limits_{i'}(l_{i'} - c_{i'j})} \\
&= c_{ij}\left(1 - \frac{b}{L - C_j}\right) + l_i\frac{b}{L - C_j},
\end{aligned} \tag{2.3}$$

where c_{ij} and $(l_i - c_{ij})$ represent the bound and unbound form of site i in sample j, $L = \sum \boldsymbol{l} (= \sum_i l_i)$ represents the total concentration of the random library in the reaction (experimentally set to 100 nM), $C_j = \sum \boldsymbol{c}_j$ represents the total concentration of bound target sites (equivalent to the bound AGO), and $b = \exp(\theta_{n+2})$, as per the logarithmic parameterization of $\boldsymbol{\theta}$ described above. To solve for c_ij, we apply the definition of K_D:

$$K_i = \frac{a_j^f\,(l_i - c_{ij})}{c_{ij}}, \tag{2.4}$$

which contains c_{ij} and $(l_i - c_{ij})$ as in equation (2.2), as well as a_j^f representing the concentration of unbound AGO in sample j, and rearrange it to

$$c_{ij} = \frac{l_i a_j^f}{a_j^f + K_i}. \tag{2.5}$$

Since $\boldsymbol{K}$ is embedded in the parameter vector $\boldsymbol{\theta}$, and since $\boldsymbol{l}$ is a fixed quantity in the analysis, determination of $\boldsymbol{c}_j$ requires only the determination of a_j^f, which is a latent, unobserved quantity within the binding reaction. While there is no explicit equation by which to calculate this value, it is absolutely specified by the constraint:

$$\begin{aligned}
a_j &= a_j^f + \sum \boldsymbol{c}_j, \\
a_j &= a_j^f + \frac{l_i a_j^f}{a_j^f + K_i},
\end{aligned} \tag{2.6}$$

which requires that the sum of concentration of unbound AGO a_j^f and the bound form of every site c_ij (for which each is a function of a_j^f) is equal to the known, total amount of AGO in the binding reaction a_j. This value is related to the parameter value θ_{n+1} according to

$$a_j = \frac{0.4}{\sqrt{10}^{j-1}} \times \exp(\theta_{n+1}). \tag{2.7}$$

We note that the prefactor in equation (2.7) relates the sample concentration index j to the percentage dilution of AGO into that binding reaction (i.e. $j = 1$ yields 40%, $j = 2$ yields 12.65%, etc.). Substitution of (2.5) into (2.3) yields:

$$\begin{aligned} c_{ij} &= l_i \left(\frac{a_j^f}{a_j^f + K_i} \left(1 - \frac{b}{L - C_j} \right) + \frac{b}{L - C_j} \right) \\ &= l_i \left(\frac{a_j^f}{a_j^f + K_i} \left(1 - \frac{b}{L - \sum_{i'} \frac{l_{i'} a_j^f}{a_j^f + K_{i'}}} \right) + \frac{b}{L - \sum_{i'} \frac{l_{i'} a_j^f}{a_j^f + K_{i'}}} \right). \end{aligned} \tag{2.8}$$

where $K_i = \exp(\theta_i)$ for $1 \leq i \leq n$. This is the final form of the function, taking as inputs the optimization parameter vector $\boldsymbol{\theta}$, the fixed parameter vector $\boldsymbol{l}$ (and its sum L), and the free AGO concentration terms a_j^f, which are computed for each dilution sample from $\boldsymbol{\theta}$ and $\boldsymbol{l}$ using equation (2.5).

3 Derivation of $f_{cost}(\boldsymbol{x})$

The cost function $f_{cost}(\boldsymbol{x})$ is derived from the product of the negative log multinomial probability density function for each column j

$$\begin{aligned} f_{cost}(\boldsymbol{x}) &= -\ln \prod_j f_{mult}(\boldsymbol{y}_j, \boldsymbol{\pi}_j) \\ &= -\ln \prod_j \frac{Y_j! \prod_i \pi_{ij}^{y_{ij}}}{\prod_i y_{ij}!}, \end{aligned} \tag{3.1}$$

where π_{ij} is the expected frequency of each site type i in sample j according to the model parameters, and $Y_j = \sum \boldsymbol{y}_j$. Each expected frequency vector $\boldsymbol{\pi}_j$ is trivially given by $\boldsymbol{x}_j/X_j$, thereby providing the link between the model simulation and subsequent likelihood estimation. Substituting pi_{ij} and distributing the natural log yields

$$f_{cost}(\boldsymbol{x}) = \sum_j \left(Y_j \ln X_j - \sum_i y_{ij} \ln x_{ij} + \sum_i \ln y_{ij}! - \ln Y_j! \right). \tag{3.2}$$

We discard the third and fourth terms in equation (3.2) as neither contains any terms of $\boldsymbol{x}_j$, and are therefore unrelated to the MLE estimation of $\boldsymbol{\theta}$. The final cost function for our optimization routine is thus

$$f_{cost}(\boldsymbol{x}) = \sum_j \left(Y_j \ln X_j - \sum_i y_{ij} \ln x_{ij} \right). \tag{3.3}$$

4 Derivation of $f_{grad}(\boldsymbol{\theta})$

The function $f_{grad}(\boldsymbol{\theta})$ returns the derivative of the cost function with respect to each component of $\boldsymbol{\theta}$:

$$\begin{aligned} f_{grad}(\boldsymbol{\theta}) &= \nabla f_{cost}(\boldsymbol{x}(\boldsymbol{\theta})) \\ &= \left(\frac{\partial f_{cost}}{\partial \theta_1}, \frac{\partial f_{cost}}{\partial \theta_2}, \ldots, \frac{\partial f_{cost}}{\partial \theta_{n+2}} \right). \end{aligned} \tag{4.1}$$

Invoking a new subscript k, for which $1 \leq k \leq n+2$, we derive $\frac{df_{cost}}{d\theta_k}$ via chain rule:

$$\begin{aligned}
\frac{df_{cost}}{d\theta_k} &= \sum_j \sum_i \frac{\partial f_{cost}}{\partial x_{ij}} \frac{dx_{ij}}{d\theta_k} \\
&= \sum_j \sum_i \frac{\partial f_{cost}}{\partial x_{ij}} \left(\frac{\partial x_{ij}}{\partial \theta_k} + \sum_{i'} \frac{\partial x_{ij}}{\partial c_{i'j}} \frac{dc_{i'j}}{d\theta_k} \right). \qquad (4.2)
\end{aligned}$$

$\frac{\partial f_{cost}}{\partial x_{ij}}$ is obtained by differentiating equation (3.3)

$$\frac{\partial f_{cost}}{\partial x_{ij}} = \frac{Y_j}{X_j} - \frac{y_{ij}}{x_{ij}}, \qquad (4.3)$$

and both $\frac{\partial x_{ij}}{\partial \theta_k}$ and $\frac{\partial x_{ij}}{\partial c_{i'j}}$ are obtained by differentiation of equation (2.3)

$$\begin{aligned}
\frac{\partial x_{ij}}{\partial \theta_k} &= \exp(\theta_k) \frac{l_i - c_{ij}}{L - C_j} \delta_{k(n+2)} \\
&= b \frac{l_i - c_{ij}}{L - C_j} \delta_{k(n+2)}, \qquad (4.4) \\
\frac{\partial x_{ij}}{\partial c_{i'j}} &= \left(b \frac{l_i - c_{ij}}{(L - C_j)^2} + \left(1 - \frac{b}{L - C_j} \right) \delta_{i'i} \right) (1 - \delta_{k(n+2)}), \qquad (4.5)
\end{aligned}$$

where δ_{ab} (or equivalently $\delta_{a(b)}$) is the Kronecker delta function, defined as:

$$\delta_{ab} = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{otherwise.} \end{cases} \qquad (4.6)$$

Substituting (4.3) and (4.4) into (4.2) and rearranging yields

$$\begin{aligned}
\frac{df_{cost}}{d\theta_k} = \sum_j \frac{1}{L - C_j} \sum_i \Bigg(\left(\frac{Y_j}{X_j} - \frac{y_{ij}}{x_{ij}} \right) \times \\
\left(b(l_i - c_{ij}) \delta_{k(n+2)} + \left((L - C_j - b) \frac{dc_{ij}}{d\theta_k} + b \frac{l_i - c_{ij}}{L - C_j} \frac{dC_j}{d\theta_k} \right) (1 - \delta_{k(n+2)}) \right) \Bigg). \qquad (4.7)
\end{aligned}$$

Inspection of the Kronecker delta functions in equation (4.7) reveals that the derivatives associated with the K_D and AGO concentration in the reaction will only use the second term in the last factor, while by contrast the parameter describing the nonspecifically recovered RNA will only use the first term, simplifying the calculation of the derivatives. Proceeding in the ordering of the parameter vector $\boldsymbol{\theta}$, we first solve for the derivative with respect to the K_D associated with each site-type, or θ_k for $1 \leq k \leq n$. This requires solving for $\frac{dc_{ij}}{d\theta_k}$ and $\frac{dC_j}{d\theta_k}$. $\frac{dc_{ij}}{d\theta_k}$ is found by differentiating equation (2.7):

$$\begin{aligned}
\frac{dc_{ij}}{d\theta_k} &= \frac{\partial c_{ij}}{\partial \theta_k} + \frac{\partial c_{ij}}{\partial a_j^f} \frac{da_j^f}{d\theta_k} \\
&= \frac{-a_j^f l_k}{(a_j^f + K_k)^2} \exp(\theta_k) \delta_{ki} + \frac{K_i l_i}{(a_j^f + K_i)^2} \frac{da_j^f}{d\theta_k} \\
&= \frac{K_i l_i}{(a_j^f + K_i)^2} \left(\frac{da_j^f}{d\theta_k} - a_j^f \delta_{ki} \right), \qquad (4.8)
\end{aligned}$$

where δ_{ki} is a Kronecker delta function. Differentiation of equation (2.8) yields

$$\begin{aligned} \frac{0.4}{\sqrt{10}^{j-1}}\exp(\theta_{n+1})\delta_{k(n+1)} &= \frac{da_j^f}{d\theta_k} - \sum_i \frac{dc_{ij}}{d\theta_k} \\ a_j\delta_{k(n+1)} &= \frac{da_j^f}{d\theta_k} - \sum_i \frac{K_i l_i}{(a_j^f + K_i)^2}\left(\frac{da_j^f}{d\theta_k} - a_j^f\delta_{ki}\right), \end{aligned} \tag{4.9}$$

an implicit solution for $\frac{da_j^f}{d\boldsymbol{\theta}}$. We rearrange (4.9) to arrive at

$$\frac{da_j^f}{d\theta_k} = \frac{a_j\delta_{k(n+1)} - a_j^f \frac{K_i l_i}{(a_j^f+K_i)^2}\delta_{ki}}{1 - \sum_i \frac{K_i l_i \delta_{ki}}{(a_j^f+K_i)^2}}, \tag{4.10}$$

which automatically provides a solution for $\frac{dC_j}{d\theta_k}$, since $\frac{da_j^f}{d\theta_k} = -\frac{dC_j}{d\theta_k}$. Substituting (4.8) into (4.7) and restricting the derivative index to $1 \le k \le n$ yields

$$\begin{aligned} \frac{df_{cost}}{d\theta_k} = \sum_j \frac{1}{L - C_j}\Bigg(&\left(\frac{Y_j}{X_j} - \frac{y_{kj}}{x_{kj}}\right)\frac{a_j^f K_k l_k}{(a_j^f + K_k)^2}(C_j + b - L) + \\ &\frac{da_j^f}{d\theta_k}\sum_i \left(\frac{Y_j}{X_j} - \frac{y_{ij}}{x_{ij}}\right)\left(\frac{K_i l_i}{(a_j^f + K_i)^2}(L - C_j - b) - b\frac{l_i - c_{ij}}{L - C_j}\right)\Bigg) \end{aligned} \tag{4.11}$$

which, when further substituted with (4.9) yields

$$\begin{aligned} \frac{df_{cost}}{d\theta_k} = \sum_j \frac{1}{L - C_j}\frac{a_j^f K_k l_k}{(a_j^f + K_k)^2}\Bigg(&\left(\frac{Y_j}{X_j} - \frac{y_{kj}}{x_{kj}}\right)(C_j + b - L) + \\ &\frac{1}{1 - \sum_i \frac{K_i l_i \delta_{ki}}{(a_j^f+K_i)^2}}\sum_i \left(\frac{Y_j}{X_j} - \frac{y_{ij}}{x_{ij}}\right)\left(\frac{K_i l_i}{(a_j^f + K_i)^2}(L - C_j - b) - b\frac{l_i - c_{ij}}{L - C_j}\right)\Bigg). \end{aligned} \tag{4.12}$$

Proceeding similarly for $\frac{df_{cost}}{d\theta_{n+1}}$, we obtain

$$\begin{aligned} \frac{df_{cost}}{d\theta_{n+1}} &= \sum_j \frac{1}{L - C_j}\frac{da_j^f}{d\theta_{n+1}}\sum_i \left(\frac{Y_j}{X_j} - \frac{y_{ij}}{x_{ij}}\right)\left((L - C_j - b)\frac{K_i l_i}{(a_j^f + K_i)^2} - b\frac{l_i - c_{ij}}{L - C_j}\right) \\ &= \sum_j \frac{1}{L - C_j}\frac{a_j}{1 - \sum_i \frac{K_i l_i \delta_{ki}}{(a_j^f+K_i)^2}}\sum_i \left(\frac{Y_j}{X_j} - \frac{y_{ij}}{x_{ij}}\right)\left((L - C_j - b)\frac{K_i l_i}{(a_j^f + K_i)^2} - b\frac{l_i - c_{ij}}{L - C_j}\right), \end{aligned} \tag{4.13}$$

and for $\frac{df_{cost}}{d\theta_{n+2}}$

$$\frac{df_{cost}}{d\theta_{n+2}} = \sum_j \frac{b}{L - C_j}\sum_i \left(\frac{Y_j}{X_j} - \frac{y_{ij}}{x_{ij}}\right)(l_i - c_{ij}). \tag{4.14}$$