

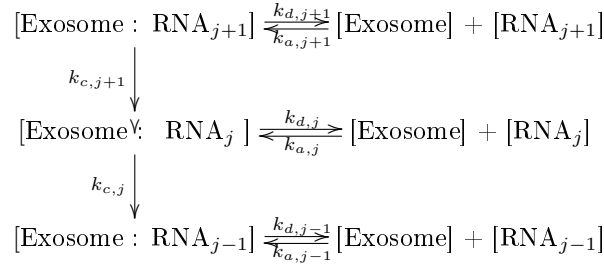
Supplements: Parameter Identification

The implementation of the method outlined in the following has been done in MATLAB [1]. The code is available from the authors on request.

1 The ODE model

We denote by $r_j = r_j(t)$ the total amount of RNA of length j , $j \in J = \{3, \dots, 30\}$. Let the corresponding amount of exosome-bound RNA be $x_j = x_j(t)$, the unbound fraction be $y_j = y_j(t)$. The (initial) amount of free exosome is denoted by $e = e(t)$. The system of ordinary differential equations (ODE) that describes the dynamics is then parametrized by the set $\Theta = \{k_{a,j}, k_{d,j}, k_{c,j} \mid j \in J\}$ (the a stands for association, d for dissociation, c for cleavage):

$$\begin{aligned} r_j(t) &= x_j(t) + y_j(t) \\ \frac{dx_j}{dt}(t) &= k_{c,j+1}x_{j+1}(t) + k_{a,j}y_j(t)e(t) - (k_{c,j} + k_{d,j})x_j(t) \\ \frac{dy_j}{dt}(t) &= k_{d,j}x_j(t) - k_{a,j}y_j(t)e(t) \end{aligned} \quad (1)$$



The initial conditions are $r_j(0) = x_j(0) = R_j$, $y_j(0) = 0$, $e(0) = E$, where the initial total amount of RNA, R_j , and the initial amount of free exosome, E , are given. The signs in (1) are chosen such that the parameter values are all positive, thus our parameter space is $\Omega = \mathbb{R}_{>0}^J \times \mathbb{R}_{>0}^J \times \mathbb{R}_{>0}^J$.

Note that we only model degradation, disregarding the reverse process of polymerization. This is because the reaction takes place in an excess of ATP, thus the reverse (synthesis) reaction can be neglected. If we want to emphasize the dependence of the results on the parameters Θ , we denote those in the superscript, e.g., we write $r_j^\Theta(t)$ instead of $r_j(t)$ etc.

The experimental data consists of a matrix $D = (R_{j,k})$, where $j = 3, \dots, 30$ runs through the lengths of the measured RNA populations, and $k = 1, \dots, K$ enumerates the measurements that were taken at times $T = t_1, \dots, t_K$ respectively. A standard ODE solver (MATLAB[1], ode15s, default parameters) is used to calculate the predictions $r_j^\Theta(t)$ of a model given by Θ . The overall goal is to determine the posterior distribution $\Pr(\Theta \mid D)$ by drawing a representative sample from it. The comparison of the measurements $R_{j,k}$ with the predictions $r_j^\Theta(t_k)$ is the basis of our sampling strategy.

2 The Sampling method

2.1 Basic approach

To determine the posterior distribution of the parameters $\Theta = \{k_{a,j}, k_{d,j}, k_{c,j} \mid j \in J\}$ of the ODE we use a Markov Chain Monte Carlo (MCMC) approach based on the Metropolis-Hastings algorithm [2]. Given Θ , we solve (1) and compare the observed values $R_{j,k}$ with the predicted value $r_j^\Theta(t_k)$. We assume that the observations $R_{j,k}$, $j \in J$, $k = 1, \dots, K$, are realizations of a Gaussian $\mathcal{N}(r_j^\Theta(t_k), \sigma_{j,k}^2)$ -distributed random variable respectively, i.e.,

$$\Pr(R_{j,k} \mid \Theta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma_{j,k}} \exp\left(-\frac{1}{2} \frac{(R_{j,k} - r_j^\Theta(t_k))^2}{\sigma_{j,k}^2}\right), \quad (2)$$

where the variances $\sigma = \{\sigma_{j,k} \mid j \in J, k = 1, \dots, K\}$ have to be given a priori, or they must be determined from the data. We will address this point later. For the sake of clarity, we will suppress the parameters σ in this section. Assuming independence of the measurement errors, the likelihood of the model becomes

$$L(\Theta) = \prod_{j \in J} \prod_{k=1}^K \Pr(R_{j,k} \mid r_j^\Theta(t_k)) \quad (3)$$

Let $\pi(\Theta)$ be any prior distribution on the parameters, and let $q(\Theta \rightarrow \Theta') = f(\Theta' \mid \Theta)$ be a proposal function, i.e., a family of distributions for $\Theta \in \Omega$, which can be calculated fast, and from which samples can be drawn easily. We briefly recall the standard Metropolis-Hastings algorithm:

1. Initialize Θ_0
2. Proposal step: Given Θ_n , draw a candidate Θ' from the proposal distribution $q(\Theta_n \rightarrow \Theta')$
- 3.1 Calculate the quantity $u = \frac{L(\Theta')}{L(\Theta_n)} \cdot \frac{\pi(\Theta')}{\pi(\Theta_n)} \cdot \frac{q(\Theta' \rightarrow \Theta_n)}{q(\Theta_n \rightarrow \Theta')}$
- 3.2 Acceptance step: With probability $\min(u, 1)$, let $\Theta_{n+1} = \Theta'$ (accept). Otherwise, let $\Theta_{n+1} = \Theta_n$ (reject)
4. Increment n by one and repeat steps 2. and 3. until convergence

2.2 Choice of proposal function

An appropriate choice of the proposal function is crucial for the convergence properties of the Markov chain. We decided to sample each individual parameter in Θ independently on a log scale from a Gaussian distribution,

$$k'_{xj} \sim \mathcal{LN}(k_{x,j}, \tau_x^2), \quad x = a, c, d; j \in J \quad (4)$$

The width of the proposal distribution is given by the variances τ_x^2 , they were determined in simulation runs (see the simulation section).

2.3 Choice of prior distribution

To avoid overfitting and to reduce the effective number of parameters, we did not choose a uniform (pseudo-) prior on Θ . It is sensible to believe that consecutive parameters $k_{x,j}$, $k_{x,j+1}$ tend to have similar values, since they reflect length (j) dependent properties of the RNA-exosome interaction. Therefore we introduced a smoothness prior for each parameter tuple $k_x = (k_{x,j})_{j \in J}$, $x = a, c, d$, by assuming that each of the consecutive differences $k_{x,j} - k_{x,j+1}$ (independently) follows a Gaussian distribution:

$$\pi(\Theta) = \pi(k_a.)\pi(k_c.)\pi(k_d.), \quad (5)$$

$$\pi(k_x.) = \prod_{j=3}^{29} \frac{1}{\sqrt{2\pi}\lambda_x} \exp\left(-\frac{1}{2} \frac{(k_{x,j} - k_{x,j+1})^2}{\lambda_x^2}\right), \quad k = a, c, d \quad (6)$$

The hyperparameters λ_a , λ_c and λ_d were determined manually in simulation runs (see the simulation section).

3 Simulation

There are compelling reasons for performing simulations before starting the analysis of the real datasets:

1. Simulations serve to assess the convergence properties of the Markov chain. Multiple, differently initialized Markov chains are run in order to verify convergence of the chain, to guarantee a sufficiently fast convergence, as well as an appropriate mixing of the chain. Importantly, the length of the Burn-in phase and the parameters of the proposal function can be determined empirically.
2. It is unclear whether all parameters Θ are identifiable in our model. Simulations help to reveal parameter dependencies. Actually, we will encounter strong dependencies in our model that require substantial changes in the sampling mechanism (see parameter identifiability).

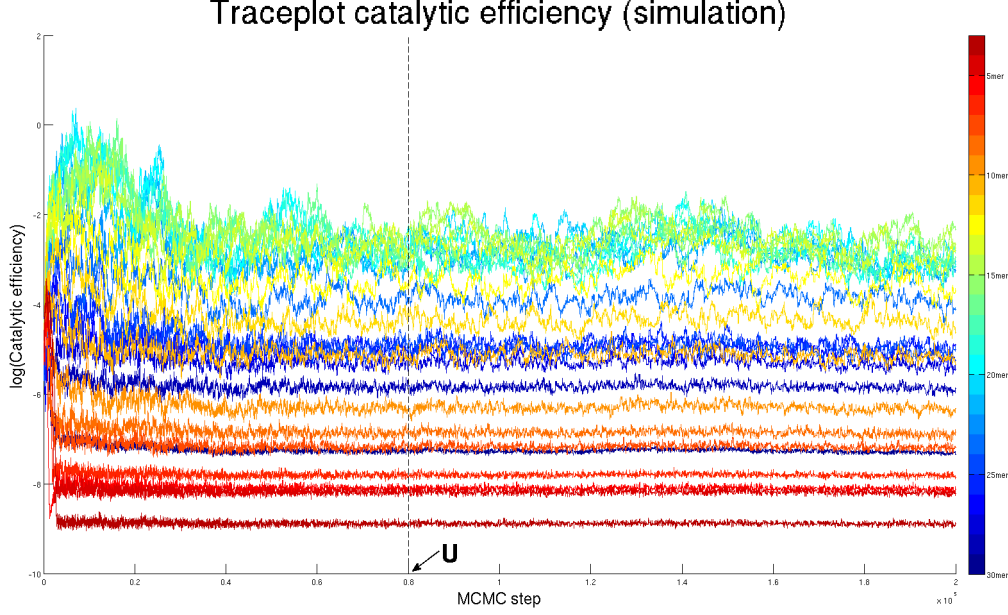


Figure 1: Choice of burnin parameter U . This figure shows the catalytic efficiency traceplot for a simulation run ($2 * 10^5$ MCMC steps, 27 parameters). The dashed line indicates the burnin parameter $U = 8 * 10^4$, the frontier between burnin phase ($s = 1, \dots, U$) and stationary phase ($s = U + 1, \dots, 2 * 10^5$). U has to be a certain trade-off between fast converging parameters (the ones with lower value in this case) and slow-converging parameters (the ones with higher values in this case).

3. The parameters λ_x of the prior distribution and the variances τ_x^2 in the proposal function need to be adjusted properly. The prior distribution introduces a bias-variance tradeoff by reducing the effective number of model parameters with decreasing λ_x , $x = a, c, d$. Our prior favors equal kinetic parameters k_{xj} , $j \in J$. This prior choice is guided by the intuition that the chemical properties of the poly-A sequences, which are the substrate of the enzymatic reaction, should not (or at least only slightly and smoothly) vary with the length of the poly-A sequence.

The simulations were carried out as follows: Starting with a realistic set of parameters Θ_{true} (which usually was obtained by applying a maximum likelihood parameter estimation method to our data), the “true” concentrations $r_j^{\Theta_{true}}(t_k)$ are computed by the above mentioned ODE solver (ode15s). We assume a common error model, namely that the measurement variance is composed of a constant “background” term β and a term which is proportional to the square of the intensity [4, 5],

$$(\sigma_{j,k}^{true})^2 = \alpha(r_j^{\Theta_{true}}(t_k))^2 + \beta, \text{ for some } \alpha, \beta \geq 0 \quad (7)$$

Then, we generate an artificial dataset $D = (R_{j,k})$ by drawing

$$R_{j,k} \sim \mathcal{N}(r_j^{\Theta_{true}}(t_k), (\sigma_{j,k}^{true})^2), \quad j \in J, \quad k = 1, \dots, K \quad (8)$$

We performed MCMC runs with error levels set to $\beta = 0$ and $\alpha = 5\%, 10\%$ and 25% .

3.1 Convergence and Mixing Properties, Model Bias and -Variance Assessment

Each model was evaluated after running $M = 5$ Markov chains for $S = 2 * 10^5$ steps, producing the parameter values $\Theta_s^m = \{k_{a,j,s}^m, k_{d,j,s}^m, k_{c,j,s}^m \mid j \in J, k = 1, \dots, K\}$, $s = 1, \dots, S$, $m = 1, \dots, M$. Convergence speed of the Markov chain is measured by the number of steps required until the chain has reached its stationary distribution. The standard visual control is offered by a convergence plot which displays the trace of each parameter along the steps s , see Fig. 1. We define by eye a value U after which the variation of the chain

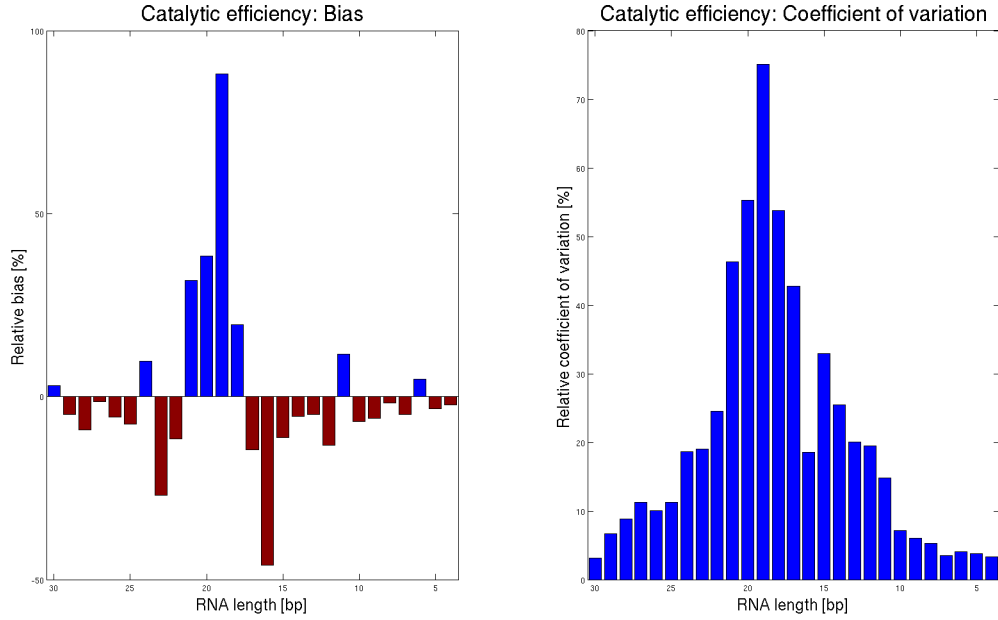


Figure 2: Relative bias (left-hand side) and coefficient of variation (right-hand side) for every catalytic efficiency parameter (see section 3.2.2). For the bias, red color indicates that the mean of the sampled parameters is shifted up with regard to the true parameter, blue color indicates that it is shifted down.

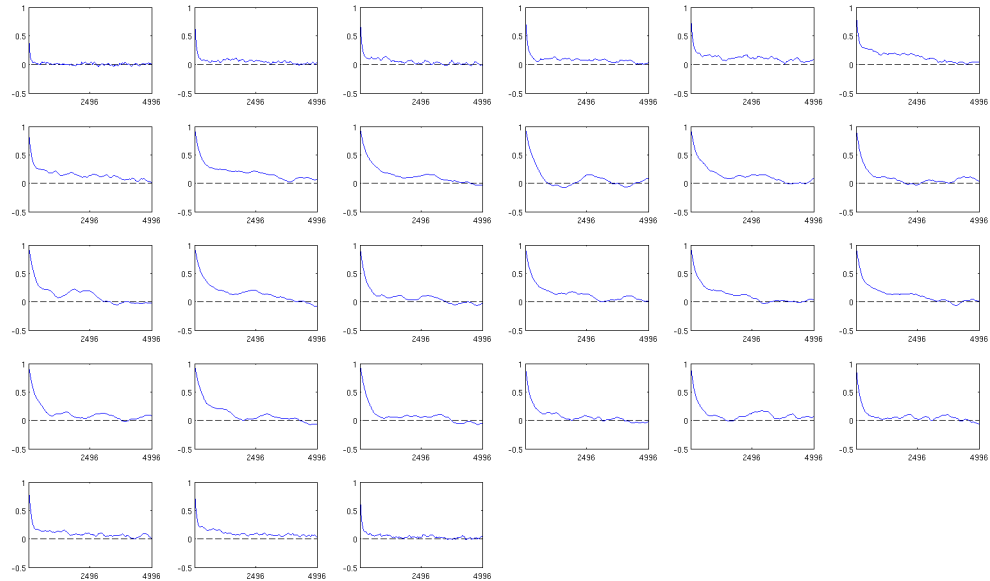


Figure 3: Autocorrelation plots of the catalytic efficiencies for an MCMC chain in stationary phase, based on the Rrp4 data. Each diagram shows the autocorrelation values of one RNA population, ranging from 30mers (top left corner) to 4mers (bottom right corner). The x-axes represent the value of k (see main text), the y-axes represent the autocorrelation values.

does not decrease further. The first steps $s = 1, \dots, U$ are declared as the burn-in phase, whereas the steps $s = U + 1, \dots, 2 * 10^5$ are called the stationary phase. The parameter values Θ_s^m of the burn-in phase are discarded for further evaluation steps. In this way we obtain an empirical sample Θ_s^m , $s > U$, $m = 1, \dots, M$, which in good approximation is drawn from the posterior distribution $P(\Theta | D)$. With $K_{x,j} = \{k_{x,j,s}^m, s > U, m = 1, \dots, M\}$, for each parameter $k_{x,j}$ of the ODE model, the relative bias and variance resp. standard deviation of its marginal posterior distribution can be assessed as

$$\text{Bias}_{x,j} = \frac{\text{mean}(K_{x,j}) - k_{x,j}}{k_{x,j}}, \quad \text{std}_{x,j} = \frac{\text{std}(K_{x,j})}{k_{x,j}}. \quad (9)$$

Bias and variance are visualized in Fig.2. Additional criteria for the quality of the sample is the acceptance rate (the proportion of proposals in the chain that have been accepted) and the “mixing”. Mixing describes the efficiency with which we sample from the whole distribution, i.e., the speed with which the empirical distribution of n consecutive individual parameter samples $\theta_1, \dots, \theta_n$ from a Markov chain converges against the true posterior distribution. The autocorrelation ([3]) function is an indicator for the mixing behavior. It can be estimated by

$$a(k) = \frac{1}{(n-k)\sigma^2} \sum_{t=1}^{n-k} (\theta_t - \bar{\theta})(\theta_{t+k} - \bar{\theta}), \quad (10)$$

with sample mean $\bar{\theta}$ and sample variance. $a(k)$ tells how a sequence of numbers correlates with a copy of itself which has been shifted by k entries. If the entries of the sequence were totally uncorrelated, $a(k)$ should be approximately zero for all k . In a sequence that has been generated by an MCMC run, consecutive entries are correlated by construction, but it is a desirable property of an MCMC chain that this correlation decreases rapidly for entries that are $k > 1$ places apart from each other. I.e., the sooner $a(k)$ decreases to zero, the better the mixing behavior of the chain. A plot of the autocorrelation functions for all parameter samples $K_{x,j}$ is shown in Fig. 3.

3.2 Parameter Identifiability and Modifications of the Model Parametrization

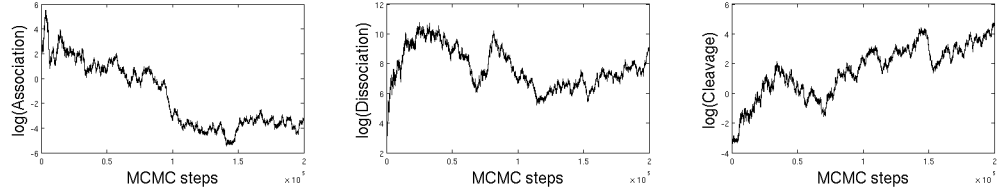
3.2.1 Problems

The complete model, as described above, consists of $27 * 4 = 108$ parameters: Association, dissociation, cleavage and polymerization parameters exist for all RNA lengths from 30 to 4. As already explained, polymerization can be neglected due to the experimental conditions, but this still leaves 81 free parameters to be estimated reliably. This is unlikely to be feasible, based on an amount of ~ 1000 individual measurements. In fact, the amount of each RNA population is only measured in total whereas the amount of free or bound RNA is not known. It can therefore not be detected whether either association or cleavage are the bottleneck for the decay process. More precisely, it is not discernible whether an average decay rate is caused by fast association and slow cleavage (most of the RNA is bound to the exosome), slow association and fast cleavage (most of the RNA is free) or by average association and average cleavage (free and bound RNA are well-balanced). Furthermore, a fast association / cleavage can be compensated by a fast dissociation, leading to a similar overall decay as a slow association / cleavage and a slow dissociation. Fig. 4 shows that even the ratios of the parameters cannot be determined precisely. In search of an identifiable quantity that describes the efficiency of decay appropriately we tested the straightforward guess $\frac{k_a * k_c}{k_d}$, and the *catalytic efficiency* $\frac{k_a * k_c}{k_d + k_c}$.

3.2.2 Catalytic efficiency

The catalytic efficiency has been shown to be meaningful when comparing the reaction rates for multiple substrates competing for one enzyme [6]. In that context, the name specificity constant was also coined for catalytic efficiency. Each MCMC chain of the original parameters leads to a derived chain for both derived quantities. We checked the convergence of that chain explicitly for the 30mer RNA. The set of parameters is then limited to only one association, one cleavage and one dissociation value, which enables the individual sampling of the parameters. Figure 4 shows how association, dissociation and cleavage relate to each other and how none of these values converges during the MCMC run. In contrast to this, Figure 5 demonstrates

A



B

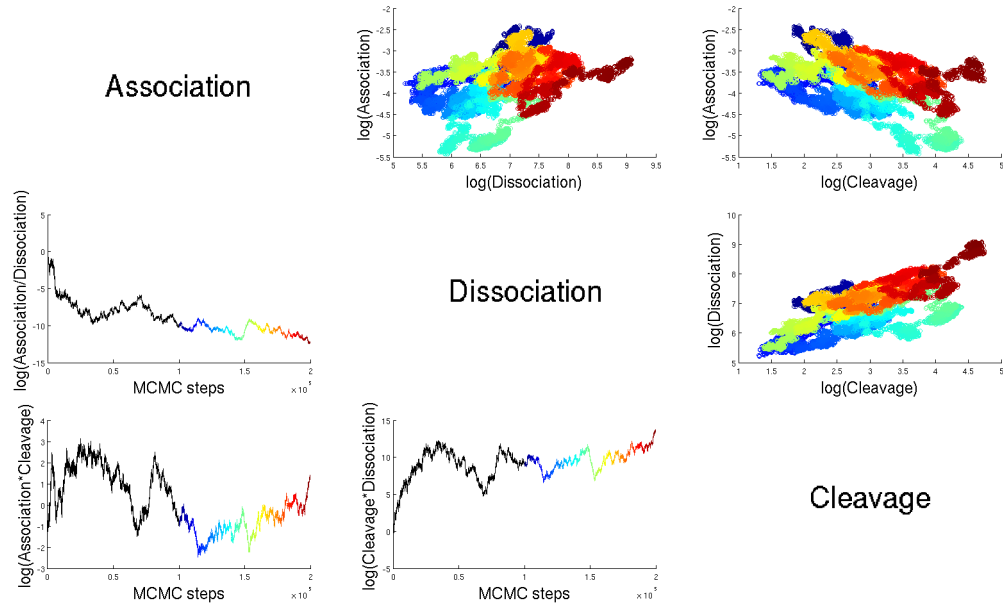


Figure 4: Parameter identifiability: These figures show how association, dissociation and cleavage can't be determined on an absolute value, but how they are related to each other, offering the possibility to define a combined value that is determinable. (A) The traceplots of the association, dissociation and cleavage parameter is shown. It can be seen that neither of them converges. (B) In the top right corner, association, cleavage and dissociation parameters are plotted against each other. Here, the colors indicate the development during the MCMC procedure, as shown in the traceplots. Importantly, note that even the derived parameters $\frac{\text{association}}{\text{dissociation}}$, $\text{association} * \text{cleavage}$ and $\text{cleavage} * \text{dissociation}$ do not converge well, they show a slight drift throughout the second half of the MCMC run.

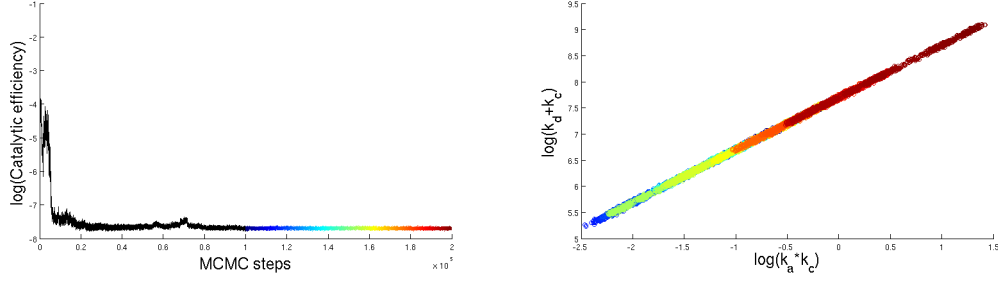


Figure 5: Parameter identifiability: The catalytic efficiency converges and has a narrow posterior distribution, as opposed to those of the individual parameters. The plot on the left-hand side shows the development of the catalytic efficiency $\frac{k_a * k_c}{k_d + k_c}$ during the MCMC procedure, while on the right-hand side nominator and denominator are plotted against each other.

that the catalytic efficiency chain converges rapidly and has a small variance, implying that the catalytic efficiency can be identified. The same does not hold for the straightforward guess $\frac{k_a * k_c}{k_d}$.

3.2.3 Summary

To determine the catalytic efficiency for each RNA length, not every parameter has to be sampled individually. Taking into account the relations between association, cleavage and dissociation we opt for the following procedure:

- **Dissociation** is fixed to one value for all RNA lengths (one manually fixed parameter). This keeps the range of possible association / cleavage parameters limited. The empirically determined value is high enough to avoid a limiting influence ($k_d = 10$).
- **Cleavage** is sampled once for all RNA lengths (one free parameter). This offers more flexibility for the association parameters.
- **Association** is sampled individually for each RNA length (27 free parameters, restricted by our prior assumptions). It accounts for the length-dependent differences of RNA decay.

We decided to use association as the flexible parameter since there are biochemical reasons to believe that it is the parameter that is primarily length-dependent: the active site, determining the cleavage rate, does not know anything about the length of the RNA. Furthermore, since the interest of this approach is to determine the catalytic efficiency of RNA decay, which is mainly driven by association and cleavage but antagonized by dissociation, we excluded the possibility of dissociation being the flexible parameter. We performed additional simulations to decide between association or cleavage as flexible parameters. Figure 6 shows that both cases yield approximately the same catalytic efficiency.

3.3 Estimation of the Measurement Error model (σ): Adaptive Likelihood MCMC

The measurement variances are a priori not known in our situation. Instead of guessing these values beforehand, we introduce an adaptive strategy for their estimation. Starting with very high initial variances, i.e. a very flat likelihood function, we estimate the variances from the comparison of predicted and measured RNA values after every 100 steps in the Markov chain, and we update the variances by moving their current values into the direction of these estimates. More specifically, let $\hat{\sigma}^{old} = \{\hat{\sigma}_{j,k}^{old} \mid j \in J, k = 1, \dots, K\}$ be the current set of variance estimates. Let $r_j^{\Theta_s}(t_k)$, $s = 1 \dots 100$, be the predictions that were produced during 100 steps in a Markov chain $(\Theta_s)_{s=1, \dots, 100}$, using $\hat{\sigma}^{old}$ as variance parameters. Assuming that the parameters Θ_s are

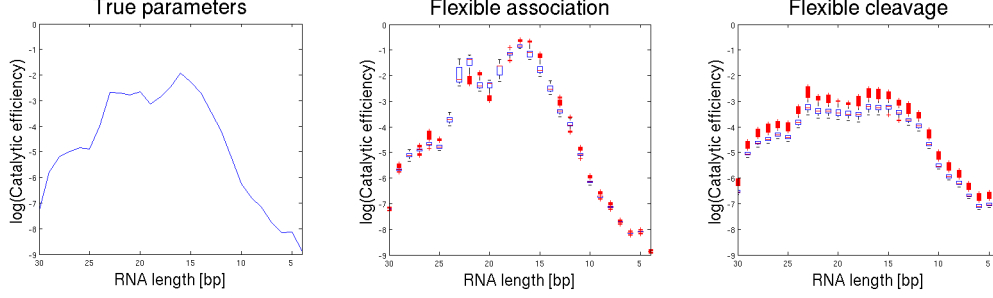


Figure 6: This figure shows the differences between association or cleavage being the flexible parameter (for which all RNA lengths are sampled). The plot on the left-hand side shows the true catalytic efficiencies, the one in the middle shows the catalytic efficiencies when all association parameter but only one cleavage parameter is sampled, the one on the right shows the catalytic efficiencies when all cleavage but only one association parameter is sampled. It can be seen that the plot in the middle fits slightly better, although the basic curve stays the same. Furthermore biological reasons as outlined in the text justify the choice of association as the flexible parameter.

close to the true parameters, the measurement error for the RNA population of length j at time point t_k is roughly $R_{j,k} - r_j^{\Theta_s}(t_k)$. A sensible guess for the variance $(\hat{\sigma}_{j,k}^{new})^2$ is thus

$$(\hat{\sigma}_{j,k}^{est})^2 = \text{mean}((R_{j,k} - r_j^{\Theta_s}(t_k))^2 \mid s = 1, \dots, 100) \quad (11)$$

We used $(\hat{\sigma}_{j,k}^{est})^2$ in place of $(\sigma_{j,k}^{true})^2$ to fit the parameters α^* , β^* of the error model in equation (7). Plugging α^* , β^* , and $r_j^{\Theta_s}(t_k)$ into (7) again produces smoothed estimates $(\hat{\sigma}_{j,k}^{est*})^2$ [7]. One could replace the old variances $(\hat{\sigma}_{j,k}^{old})^2$ by these estimates. Such a procedure has desirable properties: Since we are starting with permissive values for σ , the initial likelihood landscape is flat, which makes the chain fully explore the model space and reduces the danger of getting irreversibly caught in a local maximum. During the adaptation process, the variances tend to drop, which can be understood as a kind of annealing. Individual variances $\sigma_{j,k}$ may remain disproportionally large if the corresponding measurements are faulty/flawed, therefore providing an automatic outlier detection mechanism. On the other hand, a too fast diminishment of the variances may precipitately fix some kinetic parameters to wrong values. We guard against such effects by smoothly adjusting the old variances into the direction of the new ones:

$$(\hat{\sigma}_{j,k}^{new})^2 = \delta \cdot (\hat{\sigma}_{j,k}^{est*})^2 + (1 - \delta) \cdot (\hat{\sigma}_{j,k}^{old})^2 \quad (12)$$

Here, $\delta \in [0, 1]$ is a small constant, we set it to $\delta = 0.05$. To restrict the range of possible variances, an empirical boundary has been set to $\delta_{min} = 0.5$ and $\delta_{max} = 150$.

3.4 Choice of the Hyperparameters λ_x

We performed several MCMC runs to determine the strength of the prior. The goal is to optimize the bias-variance tradeoff by restricting the parameter fluctuation from RNA length n to length $n+1$. Since the catalytic efficiency is expected to change smoothly from step to step rather than jumping up and down this is a reasonable procedure. The results can be seen in Figure 7. Since the dissociation parameter has been set to one fixed value and the cleavage parameter is only sampled once for all RNA lengths, only a prior for the association parameters is needed. The prior is not calculated in terms of the association parameters, but it penalizes RNA length dependent variations of the catalytic efficiency parameters. Using simulation, the best trade-off was met at a hyperparameter choice $\lambda_a = 0.5$ (Fig. 7).

3.5 Final evaluation of the simulation

Figure 10 shows that our method determines parameters that describe the data very well. Figure 8 and 9 compare the relative squared error of the fits derived by our MCMC approach with those derived by a

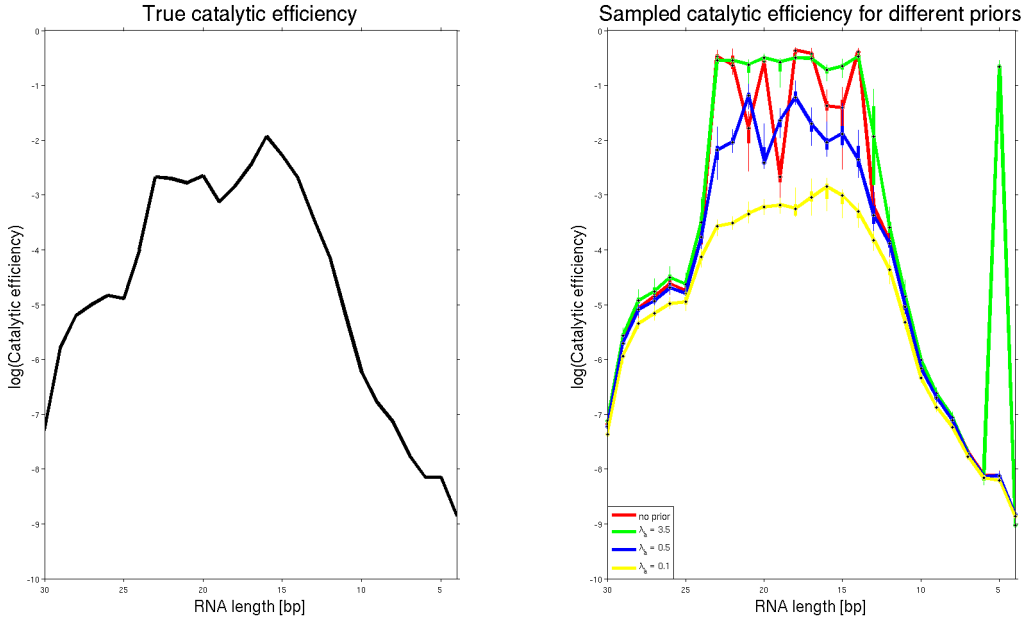


Figure 7: Choice of the hyperparameter λ : This figure shows the influence of different smoothness priors for the catalytic efficiency parameters. On the left-hand side, the true catalytic efficiencies are shown, on the right-hand one the sampled catalytic efficiencies depending on the strength of the prior. It can be seen that a prior that is not strong enough (negligible prior strength, $\lambda = 3.5$) leads to a strong variation or jumps amongst the parameters, while a prior that is too strong ($\lambda = 0.1$) does not afford the flexibility needed to adapt the parameters correctly. Based on this simulation, our prior of choice is $\lambda = 0.5$.

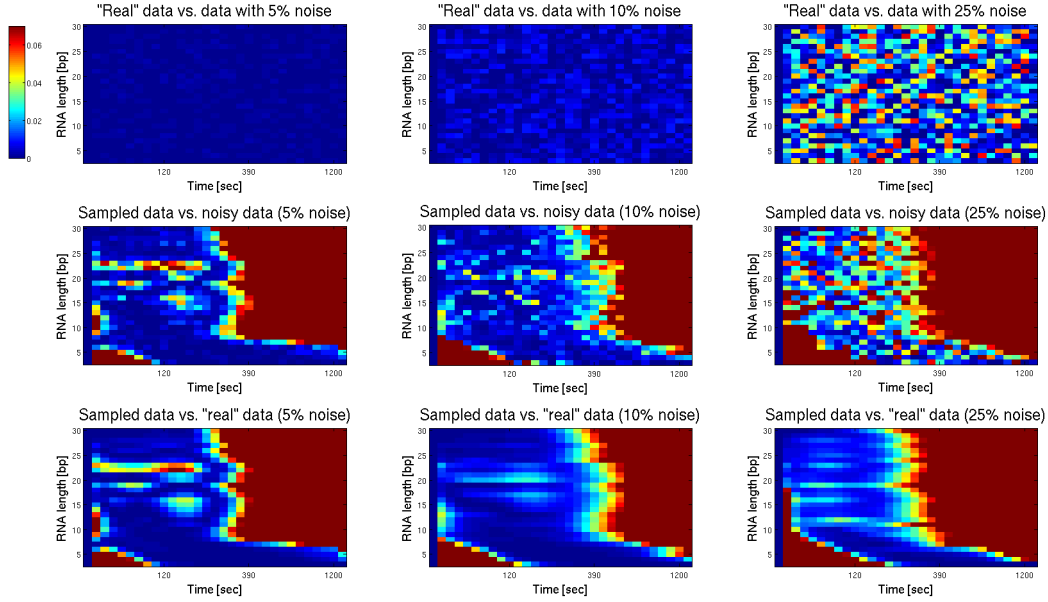


Figure 8: This figure displays the relative squared error of our fit for the simulation. To create these images, 1000 parameters have been sampled at random from the stationary phase of the Markov Chain. The averaged predictions of these 1000 models were compared to the measured data. The relative squared error has been calculated according to the following formula: $\left(\frac{|Data_{real} - Data_{sampled}|}{Data_{real}} \right)^2$. For a more detailed visualization, the color scheme has been scaled such that all values ≥ 0.07 have the same color. Every column of diagrams corresponds to a given noise ($\alpha = 5\%$, 10% and 25%) applied to the “real” (unperturbed) simulated data. The first row of diagrams displays how the perturbed data deviates from the “true” data, the second row displays the relative squared error of our fit compared to the noisy data and the third row displays the relative squared error of our fit compared to the unperturbed data. This figure allows several conclusions: First, individual RNA measurements with higher values can be fitted very well, while areas with lower amounts of RNA are fitted relatively poorly. Second, our predictions fit the real data better than the noisy data that has actually been used for fitting. Third, our procedure appears to be robust against considerable amounts of noise, since the increase in α does not visibly worsen the fit to the unperturbed data.

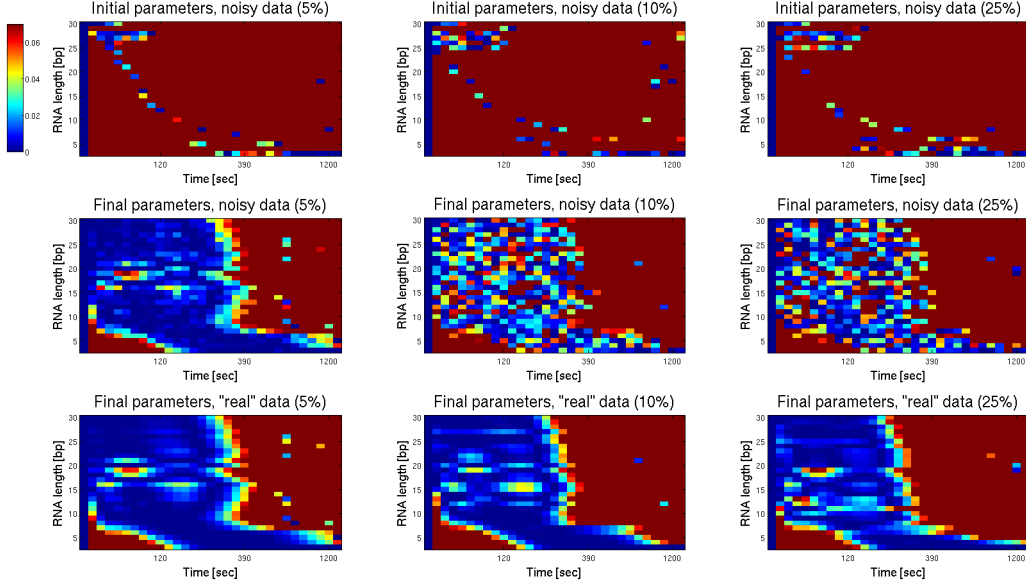


Figure 9: This figure displays the relative squared error of the straightforward optimization method's fit for the simulation (see main text). The data resulting from the optimized parameter set has been compared to the measured data using the relative squared error: $\left(\frac{|Data_{real} - Data_{sampled}|}{Data_{real}}\right)^2$. For a more detailed visualization, the color scheme has been scaled such that all values ≥ 0.07 have the same color. Every column of diagrams corresponds to a given noise ($\alpha = 5\%$, 10% and 25%) applied to the “real” (unperturbed) simulated data. The first row of diagrams displays the relative square error resulting from an optimization based on the same initial parameters we used for the MCMC approach. For the second row, the mean of a random sample from the stationary phase of the Markov Chain (the same that has been used in Figure 8) has been used as initial parameter set for the optimization and the resulting data has been compared to the noisy data (for each level of noise respectively). The third row is basically the same as the second one, except for the fact that the resulting data has been compared to the unperturbed data (again, for each level of noise respectively). It can be seen that while the straightforward optimization method yields good results (respectively, a similar one as the MCMC approach) when the initial dataset is already close to the true one, it performs very poorly when no prior information is available.

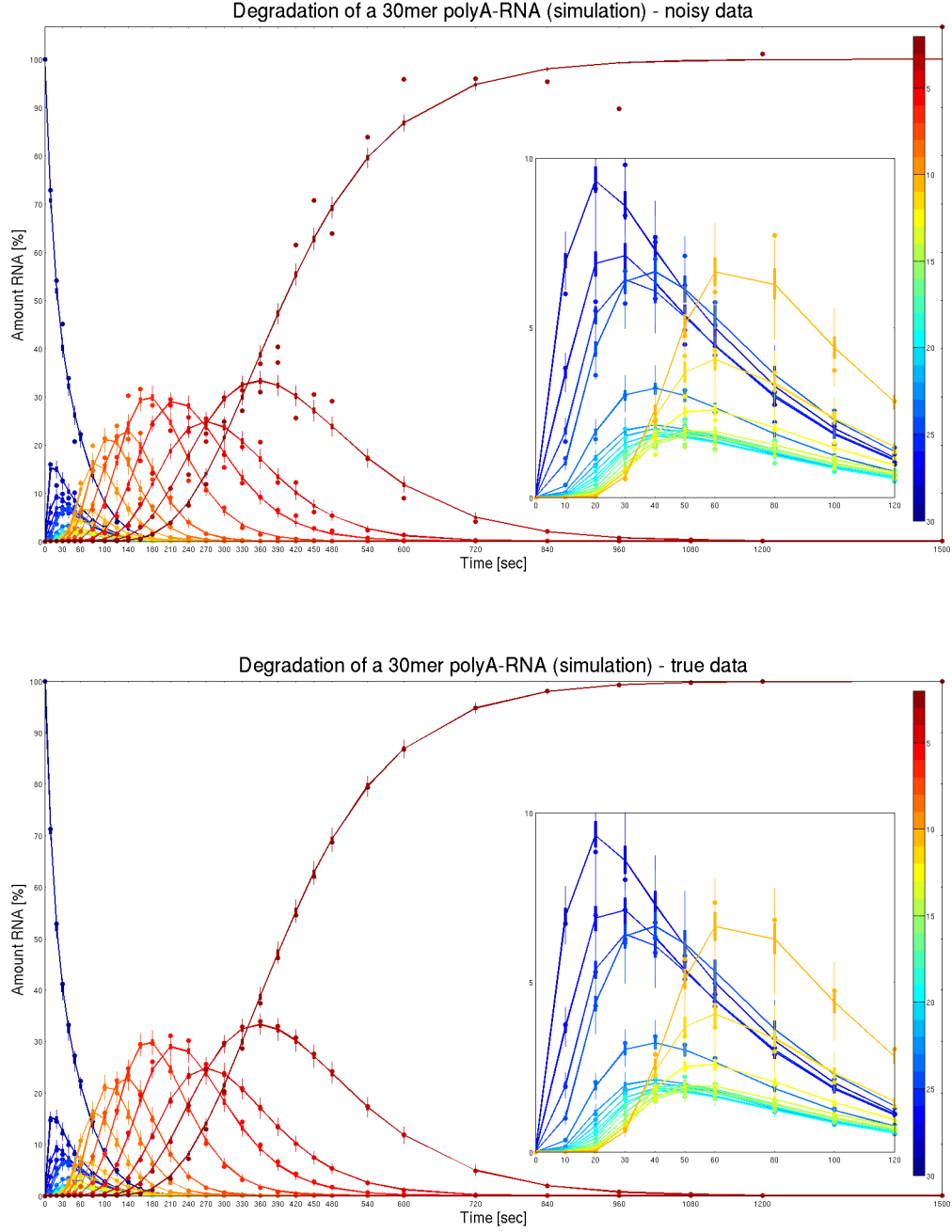


Figure 10: Goodness of fit for a simulation run. Circles in the lower plot are the unperturbed time series measurements as produced by the “true” parameter set. Circles in the upper plot represent perturbed versions of the measurements. These measurements were basis on which the MCMC chain was calculated. 1000 parameter sets have been randomly drawn from the stationary phase of the Markov chain. Thus for each RNA length and each timepoint, we obtained 1000 estimates whose distribution is displayed by boxplots. It can be seen that our model offers a better description of the “true” data (lower panel) than it does for the noisy data.

straightforward optimization method for the minimization of the quadratic loss (MATLAB[1], fminsearch, parameters see table).

| Option name | Value |
|-------------|-------------------------|
| MaxIter | $50.000 * \#parameters$ |
| MaxFunEval | $50.000 * \#parameters$ |

4 Application to real Data

For every molecule, we used an initial MCMC run to determine the set of sampled parameters that, from a chain of 100000 steps, best describes the data (least sum of squared errors). Those were used to

1. initialize two runs for each molecule that, due to the good initial parameters (derived by preceding, shorter chains), converge very fast and are used for further analysis (this is not just a continuation of the original Markov Chain, since the error model is reset to the initial one)
2. initialize three additional runs for some *other* molecule, to analyze the robustness of our approach w.r.t. initial parameter values

4.1 Single Molecule Analysis

4.1.1 Robustness w.r.t. initial parameter values

Figure 11 exemplarily displays the traceplots for the five different MCMC runs for the Rrp4 exosome. The convergence time obviously depends on the choice of the initial parameters, but for all of them it is easily reached in the given time ($\sim 50.000 - 100.000$ steps). For better visibility, one set of catalytic efficiencies is extracted in Fig. 12 to show that different initial parameters really lead to the same stable parameter after convergence.

4.1.2 Parameter identifiability

We analyzed relations between association, dissociation and cleavage (Fig. 13) as well as the identifiability of the catalytic efficiency (Fig. 14) using the initial decay of the 30mer RNA for a real dataset. Again, this has been done using the example of the Rrp4 exosome. It can be seen that the catalytic efficiency is a reliable quantity also for real datasets.

4.1.3 Results

For the final analysis of the different data sets, one of the chains initialized with convenient initial parameters has been used. The traceplots as well as the final parameter sets (burnin has been set to 150000) for all of them are shown in Figure 15. It can be seen that the convergence speed as well as the variance during the stationary phase (identifiability of the parameter) varies among the data sets. Table 1 (Csl4 exosome, Rrp4 exosome, Csl4 exosome with Y70A mutation and Csl4 exosome with R65E mutation) and 2 (capless exosome, interface mutant, crosslink mutant) specify median, 1st quartile and 3rd quartile of the catalytic efficiency value for each RNA length and each molecule. Figure 16 and 17 compare the relative squared error of the fits derived by our MCMC approach and by the straightforward optimization method (see section 3.5) for the real datasets.

The R65E mutant Although our method can describe the degradation of RNA by the exosome very well for the simulation case and for nearly all of the real molecules, this is not true for the R65E mutant. Figure 16 shows that the relative squared error for this molecule is considerably worse than for the other molecules. Furthermore, while the other molecules are stable with regard to different priors (i.e. they may be less smooth, but the catalytic efficiency remains in the same order of magnitude), the R65E mutant shows high variability (Figure 18). We thus decided to exclude this dataset from further evaluation.

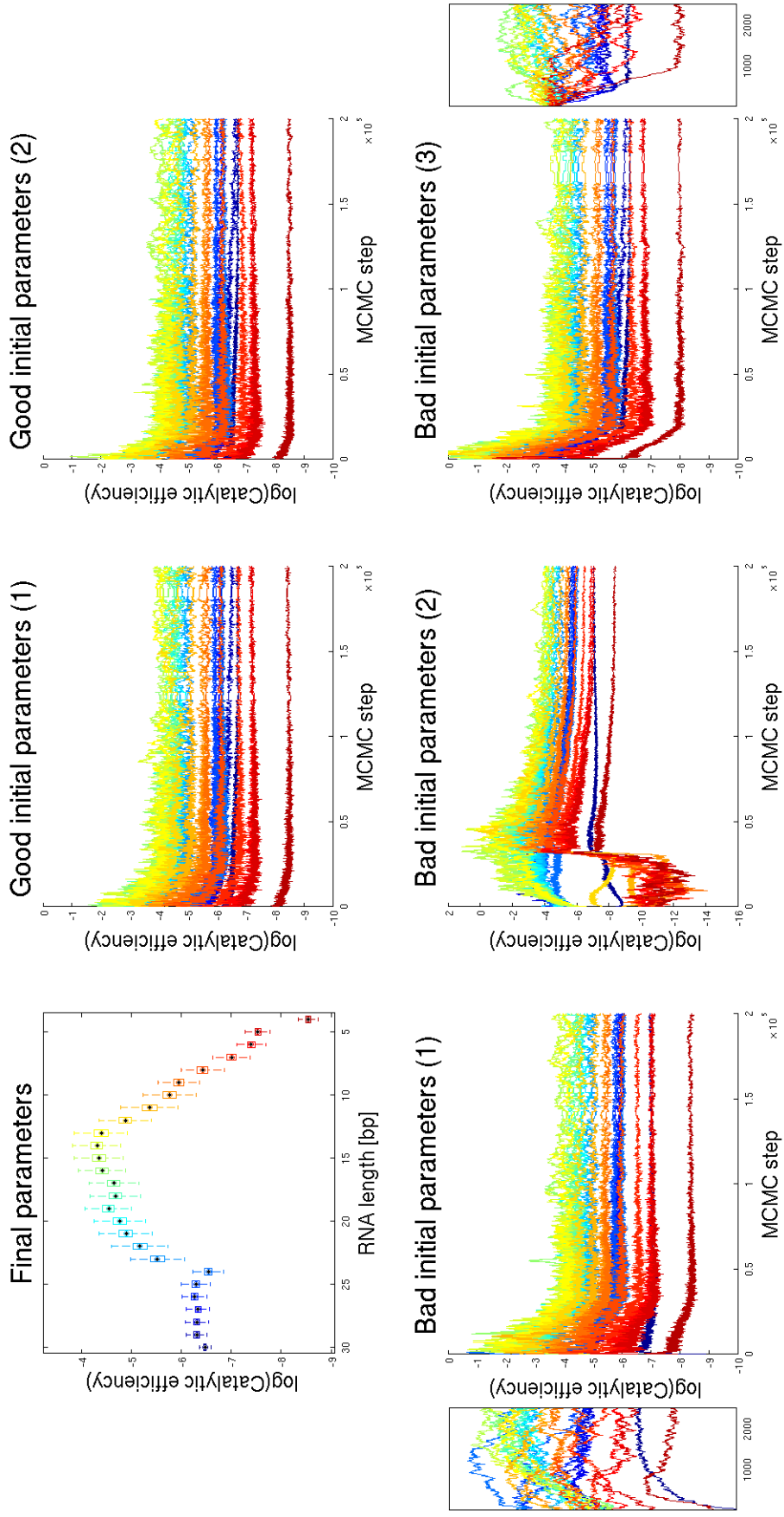


Figure 11: Influence of initial parameters: This figure shows that, independently of the choice of initial parameters, the chain converges to the same set of final parameters (see first plot) using the example of the Rrp4 exosome. While two chains have been initialized with a set of good initial parameters (plot two and three), two others have been initialized with a set of initial parameters derived by another molecule (plot four and five) and one has been initialized with all catalytic efficiencies set to the same value (sixth plot). For the fourth and sixth plot, the initial parameters and the initial 2500 steps of the chain are depicted in an enlarged extract for better visualization of the differences.

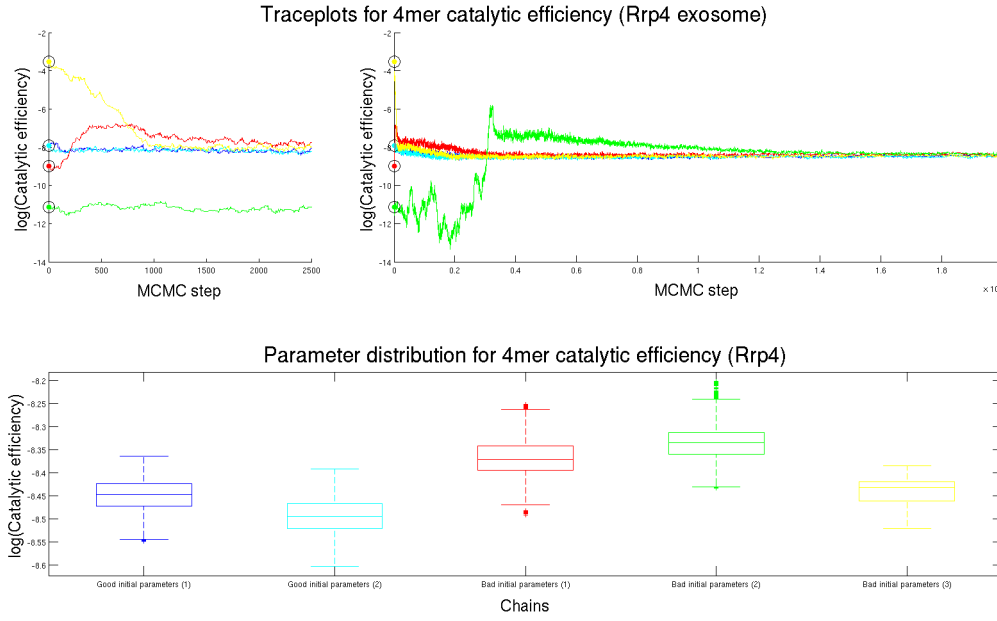


Figure 12: Influence of initial parameters: This figure shows that the choice of initial parameters has no influence on the set of final parameters, using the example of the 4mer catalytic efficiency of the Rrp4 exosome. In the top right corner, five different chains started with four different initial parameters (black circles) are depicted. The first 2500 steps are enlarged in the top left corner for a more detailed visualization. It can be seen that, although the initial parameters vary a lot, they converge to the same final parameter. The distribution of the catalytic efficiency along the chain for the individual runs is shown in the third plot.

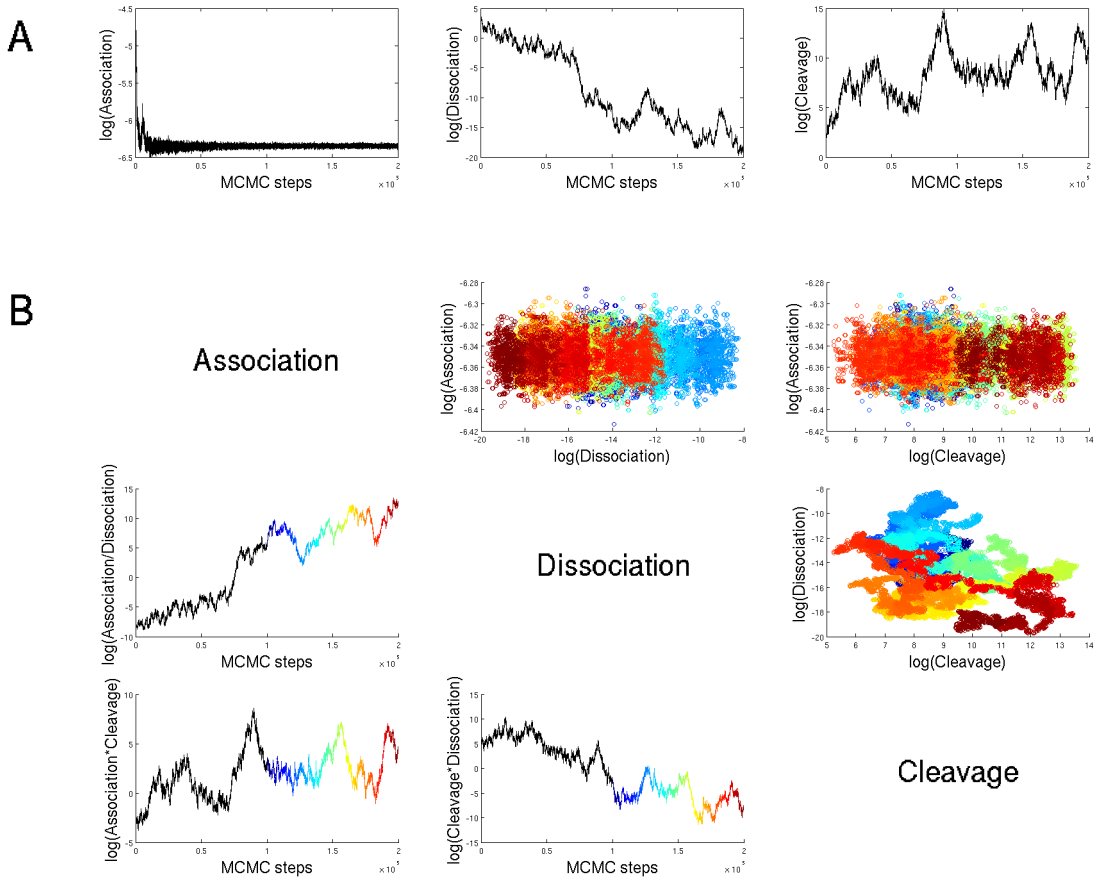


Figure 13: Parameter identifiability for real datasets (compare to fig. 4 (simulation)): These figures show that association, dissociation and cleavage on their own can not be determined. (A) The traceplots of the association, dissociation and cleavage parameter is shown. While the association parameter converges after few steps this isn't true for dissociation or cleavage. (B) In the top right corner, association, cleavage and dissociation parameters are plotted against each other. Here, the colors indicate the development during the MCMC procedure, as shown in the traceplots (bottom left corner).

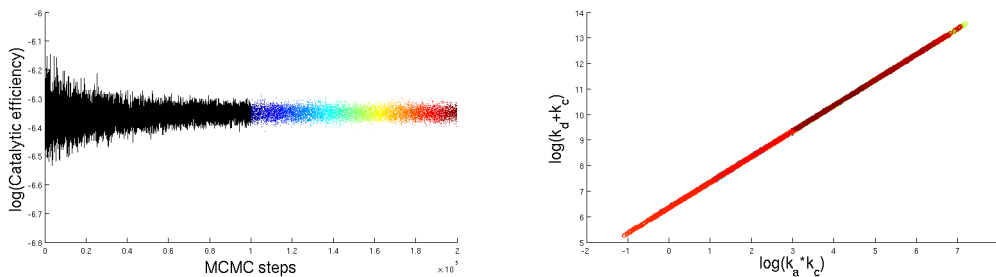


Figure 14: Parameter identifiability: This figure shows that the catalytic efficiency can be determined perfectly by our approach also for real datasets. The plot on the left-hand side shows the development of the catalytic efficiency $\frac{k_a * k_c}{k_d + k_c}$ during the MCMC procedure, while on the right-hand side nominator and denominator are plotted against each other.

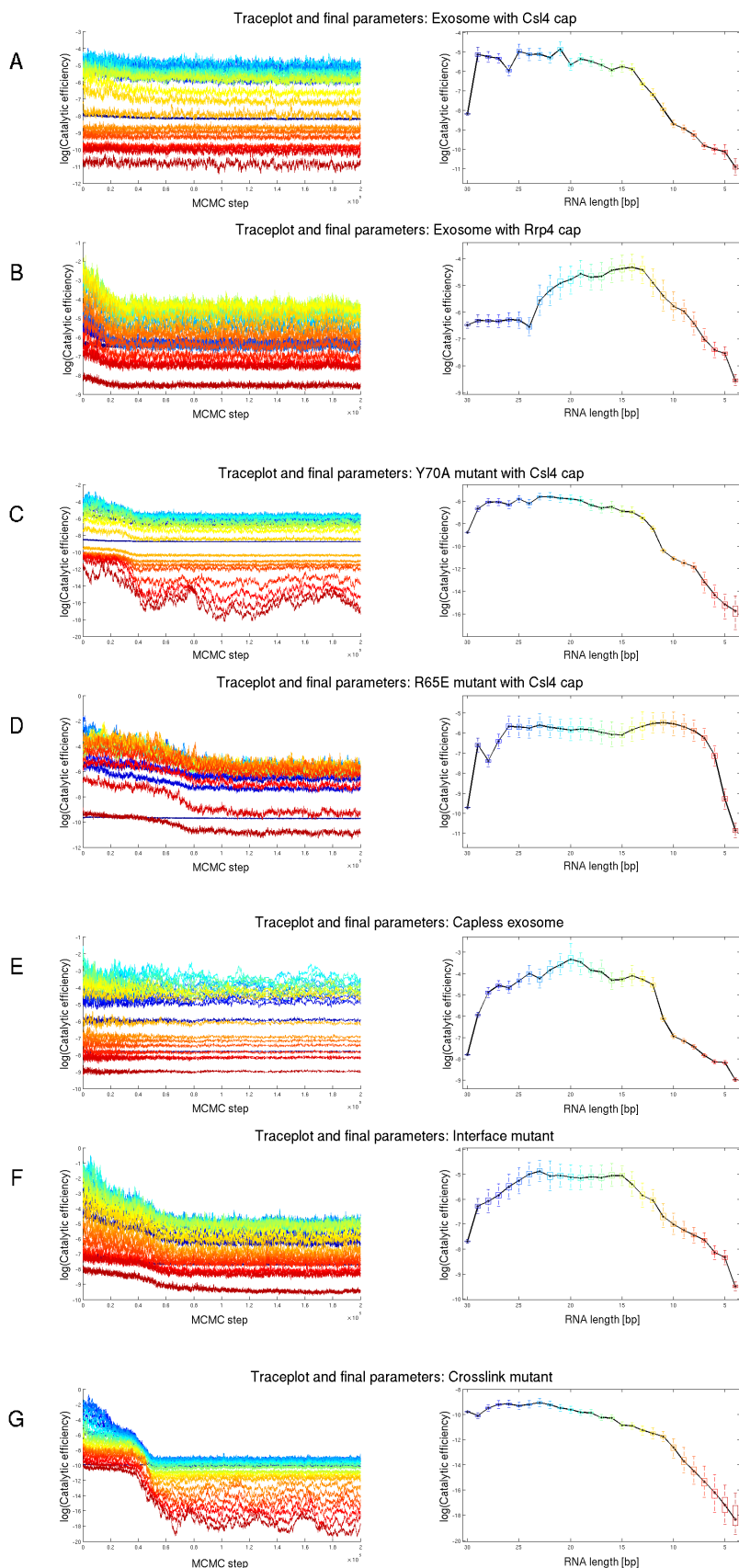


Figure 15: This figure shows the traceplot and final parameter set (burnin=150000) individually for all datasets. It can be seen that the MCMC chains vary in convergence speed as well as in variability. The boxplots on the right-hand side illustrate the main advantage of the MCMC approach: it not only offers a set of parameters that best describe the measured data, but it also yields a posterior distribution for each catalytic efficiency parameter and thus provides a more comprehensive summary of the data.

Table 1: Results - catalytic efficiency: Median [1stquartile,3rdquartile]

| | Csl4 exosome | Rrp4 exosome | Csl4-Y70A exosome | Csl4-R65E exosome |
|----|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| 30 | 2.795E-04 [2.750E-04,2.839E-04] | 1.539E-03 [1.497E-03,1.585E-03] | 1.560E-04 [1.543E-04,1.578E-04] | 6.016E-05 [5.965E-05,6.069E-05] |
| 29 | 5.875E-03 [5.373E-03,6.454E-03] | 1.814E-03 [1.721E-03,1.917E-03] | 1.258E-03 [1.178E-03,1.339E-03] | 1.362E-03 [1.254E-03,1.485E-03] |
| 28 | 5.223E-03 [4.875E-03,5.606E-03] | 1.808E-03 [1.712E-03,1.918E-03] | 2.276E-03 [2.120E-03,2.461E-03] | 6.174E-04 [5.727E-04,6.633E-04] |
| 27 | 4.842E-03 [4.558E-03,5.146E-03] | 1.768E-03 [1.673E-03,1.882E-03] | 2.297E-03 [2.138E-03,2.493E-03] | 1.643E-03 [1.511E-03,1.807E-03] |
| 26 | 2.553E-03 [2.398E-03,2.749E-03] | 1.885E-03 [1.771E-03,2.018E-03] | 1.781E-03 [1.631E-03,1.938E-03] | 3.421E-03 [2.968E-03,3.892E-03] |
| 25 | 6.895E-03 [6.350E-03,7.578E-03] | 1.849E-03 [1.719E-03,2.010E-03] | 2.974E-03 [2.762E-03,3.236E-03] | 3.364E-03 [2.945E-03,3.863E-03] |
| 24 | 5.862E-03 [5.443E-03,6.366E-03] | 1.440E-03 [1.331E-03,1.575E-03] | 1.998E-03 [1.823E-03,2.191E-03] | 3.170E-03 [2.819E-03,3.611E-03] |
| 23 | 5.954E-03 [5.534E-03,6.471E-03] | 3.824E-03 [3.333E-03,4.454E-03] | 3.654E-03 [3.413E-03,3.946E-03] | 3.655E-03 [3.237E-03,4.283E-03] |
| 22 | 4.999E-03 [4.675E-03,5.339E-03] | 5.574E-03 [4.767E-03,6.429E-03] | 3.606E-03 [3.334E-03,3.987E-03] | 3.304E-03 [2.897E-03,3.753E-03] |
| 21 | 7.714E-03 [7.037E-03,8.450E-03] | 7.352E-03 [6.372E-03,8.604E-03] | 3.174E-03 [2.935E-03,3.447E-03] | 3.090E-03 [2.745E-03,3.468E-03] |
| 20 | 3.480E-03 [3.249E-03,3.796E-03] | 8.446E-03 [7.308E-03,9.706E-03] | 2.988E-03 [2.722E-03,3.250E-03] | 2.874E-03 [2.539E-03,3.354E-03] |
| 19 | 4.663E-03 [4.354E-03,4.994E-03] | 1.041E-02 [9.314E-03,1.192E-02] | 2.621E-03 [2.364E-03,2.874E-03] | 3.022E-03 [2.641E-03,3.457E-03] |
| 18 | 4.116E-03 [3.866E-03,4.390E-03] | 9.138E-03 [8.162E-03,1.040E-02] | 1.730E-03 [1.543E-03,1.966E-03] | 2.905E-03 [2.536E-03,3.340E-03] |
| 17 | 3.408E-03 [3.181E-03,3.685E-03] | 9.433E-03 [8.392E-03,1.086E-02] | 1.350E-03 [1.207E-03,1.514E-03] | 2.574E-03 [2.237E-03,2.970E-03] |
| 16 | 2.640E-03 [2.445E-03,2.880E-03] | 1.186E-02 [1.061E-02,1.323E-02] | 1.474E-03 [1.305E-03,1.671E-03] | 2.311E-03 [1.986E-03,2.747E-03] |
| 15 | 3.172E-03 [2.942E-03,3.405E-03] | 1.267E-02 [1.121E-02,1.437E-02] | 1.005E-03 [8.912E-04,1.130E-03] | 2.254E-03 [1.987E-03,2.611E-03] |
| 14 | 2.782E-03 [2.589E-03,2.990E-03] | 1.324E-02 [1.177E-02,1.505E-02] | 9.318E-04 [8.347E-04,1.052E-03] | 2.908E-03 [2.557E-03,3.429E-03] |
| 13 | 1.295E-03 [1.209E-03,1.396E-03] | 1.203E-02 [1.061E-02,1.370E-02] | 5.594E-04 [5.096E-04,6.235E-04] | 3.481E-03 [2.917E-03,4.199E-03] |
| 12 | 7.501E-04 [6.997E-04,8.093E-04] | 7.407E-03 [6.578E-03,8.517E-03] | 2.138E-04 [2.009E-04,2.288E-04] | 4.000E-03 [3.526E-03,4.594E-03] |
| 11 | 3.516E-04 [3.239E-04,3.829E-04] | 4.556E-03 [3.939E-03,5.289E-03] | 3.070E-05 [2.949E-05,3.199E-05] | 4.172E-03 [3.695E-03,4.743E-03] |
| 10 | 1.649E-04 [1.577E-04,1.731E-04] | 3.109E-03 [2.751E-03,3.567E-03] | 1.524E-05 [1.461E-05,1.588E-05] | 3.902E-03 [3.382E-03,4.536E-03] |
| 9 | 1.304E-04 [1.245E-04,1.373E-04] | 2.572E-03 [2.331E-03,2.866E-03] | 1.023E-05 [9.582E-06,1.091E-05] | 3.403E-03 [2.975E-03,3.982E-03] |
| 8 | 9.420E-05 [8.927E-05,9.982E-05] | 1.621E-03 [1.467E-03,1.809E-03] | 7.271E-06 [6.480E-06,7.954E-06] | 2.767E-03 [2.446E-03,3.183E-03] |
| 7 | 5.395E-05 [5.168E-05,5.629E-05] | 9.026E-04 [8.251E-04,1.005E-03] | 1.921E-06 [1.474E-06,2.351E-06] | 1.933E-03 [1.715E-03,2.189E-03] |
| 6 | 4.559E-05 [4.285E-05,4.825E-05] | 6.090E-04 [5.677E-04,6.550E-04] | 6.026E-07 [4.588E-07,7.562E-07] | 8.088E-04 [7.044E-04,9.015E-04] |
| 5 | 3.937E-05 [3.588E-05,4.347E-05] | 5.328E-04 [5.014E-04,5.669E-04] | 2.553E-07 [1.795E-07,3.244E-07] | 9.304E-05 [8.226E-05,1.054E-04] |
| 4 | 1.841E-05 [1.651E-05,2.039E-05] | 1.952E-04 [1.856E-04,2.062E-04] | 1.452E-07 [9.204E-08,2.228E-07] | 1.946E-05 [1.771E-05,2.108E-05] |

Table 2: Results - catalytic efficiency: Median [1stquartile,3rdquartile]

| | Capless exosome | Interface mutant | Crosslink mutant |
|----|------------------------------------|------------------------------------|------------------------------------|
| 30 | 4.102E-04 [4.065E-04,4.139E-04] | 4.616E-04 [4.528E-04,4.701E-04] | 5.638E-05 [5.560E-05,5.717E-05] |
| 29 | 2.649E-03 [2.566E-03,2.735E-03] | 1.871E-03 [1.743E-03,2.027E-03] | 4.112E-05 [3.881E-05,4.363E-05] |
| 28 | 7.568E-03 [7.086E-03,8.075E-03] | 2.295E-03 [2.055E-03,2.594E-03] | 7.489E-05 [7.026E-05,8.029E-05] |
| 27 | 1.048E-02 [9.778E-03,1.111E-02] | 2.903E-03 [2.620E-03,3.278E-03] | 1.004E-04 [9.299E-05,1.079E-04] |
| 26 | 9.411E-03 [8.635E-03,1.024E-02] | 4.090E-03 [3.627E-03,4.662E-03] | 1.035E-04 [9.679E-05,1.123E-04] |
| 25 | 1.302E-02 [1.217E-02,1.430E-02] | 5.226E-03 [4.556E-03,5.903E-03] | 8.996E-05 [8.405E-05,9.675E-05] |
| 24 | 1.832E-02 [1.602E-02,2.018E-02] | 6.806E-03 [6.107E-03,7.643E-03] | 1.013E-04 [9.344E-05,1.088E-04] |
| 23 | 1.444E-02 [1.240E-02,1.566E-02] | 7.590E-03 [6.845E-03,8.514E-03] | 1.157E-04 [1.072E-04,1.261E-04] |
| 22 | 2.146E-02 [1.855E-02,2.557E-02] | 6.305E-03 [5.449E-03,7.340E-03] | 9.800E-05 [9.069E-05,1.056E-04] |
| 21 | 2.747E-02 [2.395E-02,3.181E-02] | 6.426E-03 [5.737E-03,7.165E-03] | 7.569E-05 [7.070E-05,8.114E-05] |
| 20 | 3.581E-02 [2.889E-02,4.279E-02] | 5.992E-03 [5.351E-03,6.855E-03] | 6.651E-05 [6.193E-05,7.145E-05] |
| 19 | 3.121E-02 [2.567E-02,3.663E-02] | 5.783E-03 [5.171E-03,6.542E-03] | 5.327E-05 [5.025E-05,5.642E-05] |
| 18 | 2.125E-02 [1.925E-02,2.306E-02] | 6.058E-03 [5.460E-03,6.797E-03] | 5.150E-05 [4.827E-05,5.524E-05] |
| 17 | 1.965E-02 [1.786E-02,2.309E-02] | 5.925E-03 [5.318E-03,6.782E-03] | 3.556E-05 [3.331E-05,3.776E-05] |
| 16 | 1.339E-02 [1.185E-02,1.462E-02] | 6.396E-03 [5.737E-03,7.296E-03] | 3.429E-05 [3.226E-05,3.669E-05] |
| 15 | 1.394E-02 [1.277E-02,1.542E-02] | 6.364E-03 [5.782E-03,7.048E-03] | 1.918E-05 [1.795E-05,2.063E-05] |
| 14 | 1.651E-02 [1.423E-02,1.825E-02] | 4.508E-03 [4.023E-03,5.168E-03] | 1.832E-05 [1.710E-05,1.955E-05] |
| 13 | 1.394E-02 [1.282E-02,1.612E-02] | 2.882E-03 [2.563E-03,3.289E-03] | 1.298E-05 [1.187E-05,1.421E-05] |
| 12 | 1.076E-02 [9.805E-03,1.170E-02] | 2.378E-03 [2.132E-03,2.659E-03] | 1.012E-05 [9.149E-06,1.128E-05] |
| 11 | 2.205E-03 [2.126E-03,2.293E-03] | 1.232E-03 [1.092E-03,1.415E-03] | 7.864E-06 [7.083E-06,8.713E-06] |
| 10 | 9.687E-04 [9.409E-04,9.954E-04] | 9.045E-04 [8.203E-04,1.023E-03] | 3.299E-06 [2.618E-06,4.042E-06] |
| 9 | 7.784E-04 [7.556E-04,8.012E-04] | 7.142E-04 [6.573E-04,7.831E-04] | 1.112E-06 [8.742E-07,1.510E-06] |
| 8 | 5.888E-04 [5.750E-04,6.086E-04] | 5.883E-04 [5.429E-04,6.366E-04] | 5.029E-07 [3.878E-07,7.019E-07] |
| 7 | 3.935E-04 [3.848E-04,4.033E-04] | 4.795E-04 [4.462E-04,5.175E-04] | 2.157E-07 [1.576E-07,2.931E-07] |
| 6 | 2.924E-04 [2.858E-04,2.989E-04] | 2.929E-04 [2.740E-04,3.150E-04] | 9.133E-08 [5.832E-08,1.239E-07] |
| 5 | 2.806E-04 [2.744E-04,2.886E-04] | 2.392E-04 [2.229E-04,2.567E-04] | 3.399E-08 [1.893E-08,6.257E-08] |
| 4 | 1.263E-04 [1.240E-04,1.290E-04] | 7.528E-05 [7.206E-05,7.893E-05] | 1.104E-08 [6.889E-09,3.346E-08] |

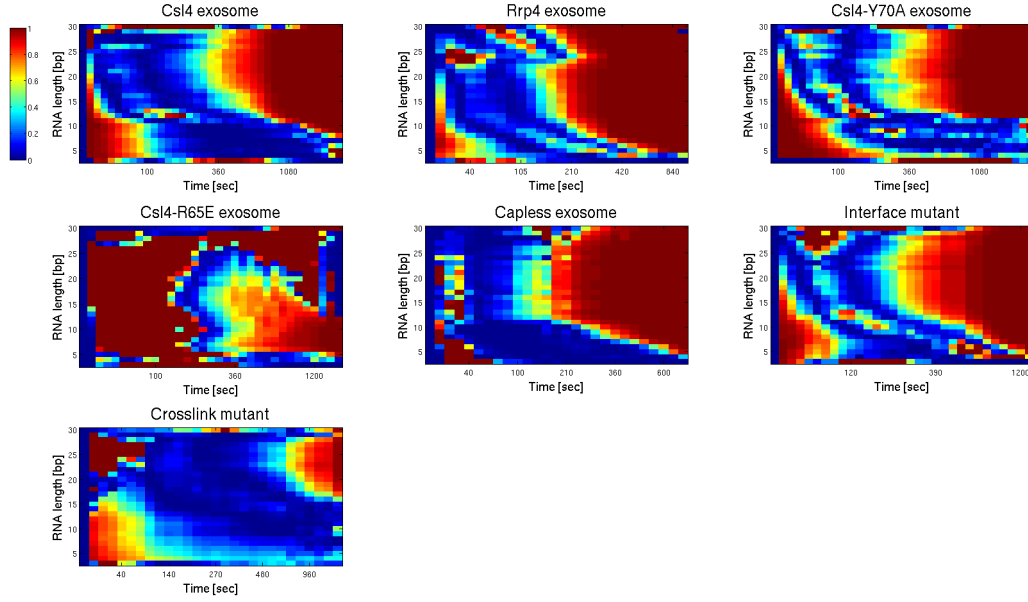


Figure 16: This figure displays the relative squared error of our fit for the different exosome variants. To create these images, 1000 parameters have been sampled at random from the stationary phase of the Markov Chain and a mean of the resulting data has been compared to the measured data. The relative squared error has been calculated according to the following formula: $\left(\frac{|Data_{real} - Data_{sampled}|}{Data_{real}}\right)^2$. For a more detailed visualization, the color scheme has been scaled such that all values ≥ 1 have the same color. Again, it can be seen that a larger amount of RNA (corresponding to the strongly stained spots in the gel) can be fitted very well by our approach, while areas with a lower amount of RNA are fitted less well.

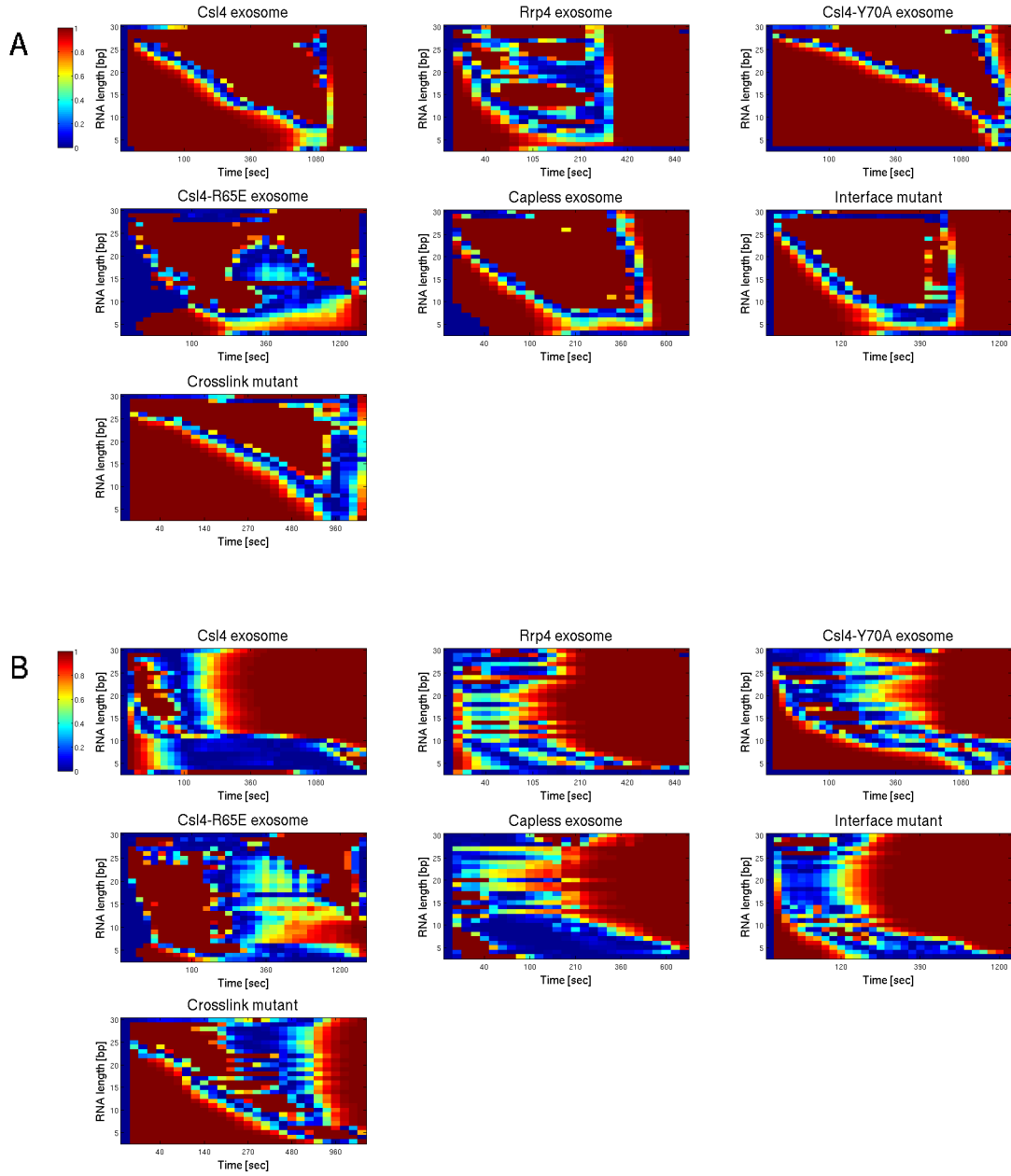


Figure 17: This figure displays the relative squared error of the straightforward optimization method's fit for the different exosome variants. The data resulting from the optimized parameter set has been compared to the measured data using the relative squared error. Color scales are as in Figure 16. In (A), the initial parameters are the same as the initial parameters used for the MCMC approach. In (B), the initial parameter set is the mean of the same random sample that has been used in Figure 16. Like for the simulated data it can be seen here that the straightforward optimization yields comparable results as the MCMC approach when a good set of initial parameters is offered, but it performs very poorly when no prior information is available.

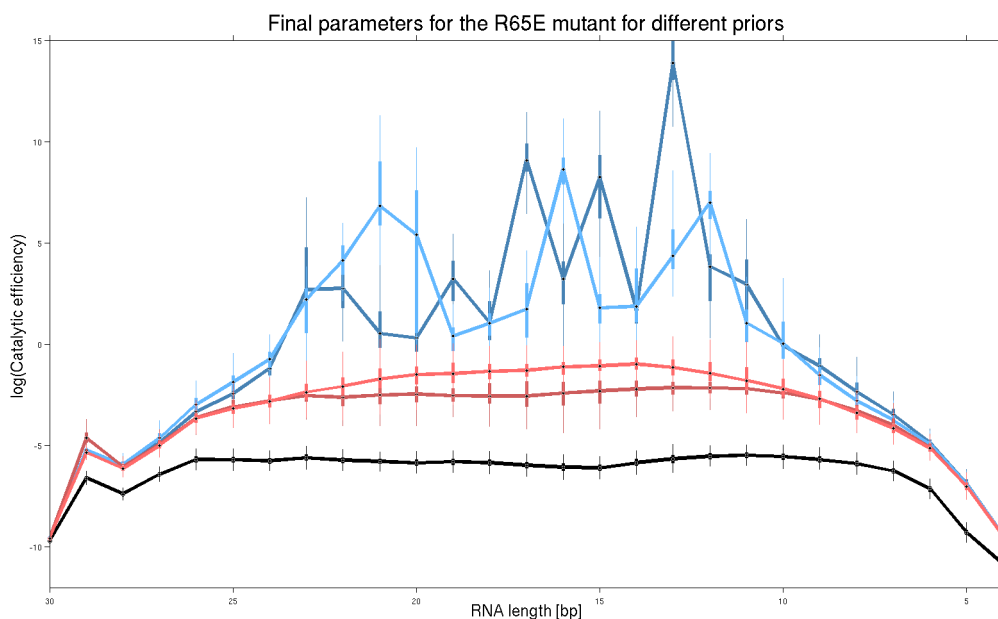


Figure 18: Influence of prior strength on the sampled parameters for the R65E mutant. The R65E mutant shows high variance with respect to different priors (blue: weak prior, red: medium prior, black: current prior). It is therefore excluded from further evaluation.

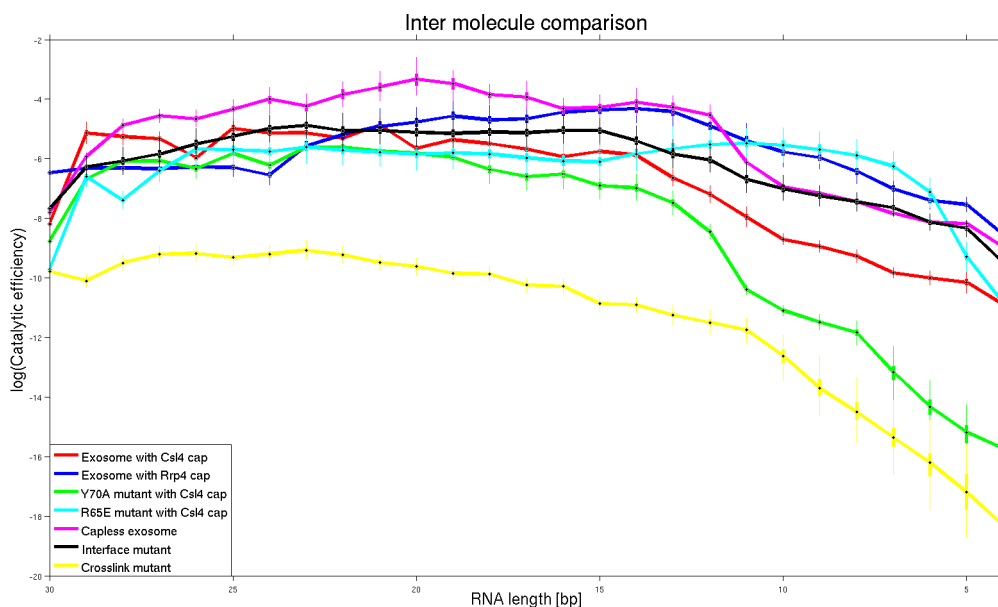


Figure 19: This figure offers a direct comparison of the final parameter sets for the different molecules as shown in Figure 15. For an interpretation of the differences see the results section of the paper.

4.2 Inter-Molecule comparison

In Figure 19, a comparison of the final parameter sets for the different molecules is offered. The differences are discussed in the results section of the paper.

References

- [1] Matlab version r2009b. Natick, Massachusetts: The MathWorks Inc., 2009.
- [2] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1):5–43, January 2003.
- [3] Patrick F. Dunn. *Measurement and Data Analysis for Engineering and Science (Engineering Series)*. McGrawHill, New York, 2005.
- [4] B.P. Durbin, J.S. Hardin, D.M. Hawkins, and D.M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *BIOINFORMATICS-OXFORD-*, 18:105–110, 2002.
- [5] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *BIOINFORMATICS-OXFORD-*, 18:96–104, 2002.
- [6] D. Koshlandjr. The application and usefulness of the ratio /. *Bioorganic Chemistry*, 30(3):211–213, June 2002.
- [7] D. M. Rocke and B. Durbin. A model for measurement error for gene expression arrays. *J. Comput. Biol.*, 8:557–569, 2001.