Estimating RNA structure chemical probing reactivities from reverse transcriptase stops and mutations.

Angela M Yu[1,2], **Molly E. Evans**[1], **and Julius B. Lucks**[1,*]

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL

[2]Tri-Institutional Training Program in Computational Biology and Medicine, Cornell University, Ithaca, New York, Weill Cornell Medical College, New York, New York, Memorial Sloan-Kettering Cancer Center, New York, New York, USA

[*]jblucks@northwestern.edu

ABSTRACT

Chemical probing experiments interrogate RNA structures by creating covalent adducts on RNA molecules in structure-dependent patterns. Adduct positions are then detected through conversion of the modified RNAs into complementary DNA (cDNA) by reverse transcription (RT) as either stops (RT-stops) or mutations (RT-mutations). Statistical analysis of the frequencies of RT-stops and RT-mutations can then be used to estimate a measure of chemical probing reactivity at each nucleotide of an RNA, which reveals properties of the underlying RNA structure. Inspired by recent work that showed that different reverse transcriptase enzymes show distinct biases for detecting adducts as either RT-stops or RT-mutations, here we use a statistical modeling framework to derive an equation for chemical probing reactivity using experimental signatures from both RT-stops and RT-mutations within a single experiment. The resulting formula intuitively matches the expected result from considering reactivity to be defined as the fraction of adduct observed at each position in an RNA at the end of a chemical probing experiment. We discuss assumptions and implementation of the model, as well as ways in which the model may be experimentally validated.

Introduction

Chemical probing has developed into a powerful experimental approach to interrogate RNA structures *in vitro* and *in vivo*[1–46]. In these experiments, chemical reactions between an RNA and a probe creates covalent adducts at positions in the RNA in a pattern that is determined in part by the underlying structure of the RNA[47]. Uncovering the distribution of adduct positions across a population of RNAs is then a means by which to measure structural properties of those RNAs.

Recently, a collection of experimental techniques have been developed that use sequencing technologies to recover the adduct distribution of chemically modified RNAs as accurately as possible in order to infer RNA structures[9,10,18,19,29–32,34–37,40,43,44,46,48,49].

These techniques all use indirect methods to detect adduct positions, since direct detection of chemical probing adducts on an RNA molecule has not been shown to be possible with current sequencing technologies. The most convenient indirect method is to first convert the modified RNA into a DNA molecule using an enzymatic process called reverse transcription (RT). In this process, a reverse transcriptase enzyme (also referred to as RT) catalyzes the synthesis of a complementary DNA (cDNA) molecule in a $3' \rightarrow 5'$ direction (Fig. 1). When RT encounters an adduct, one of two scenarios is possible: either the RT stops 1 nt before the adduct[50] (at site $k-1$) which we call an RT-stop, or the RT proceeds through the adduct and introduces a mutation at site k, which we call an RT-mutation.

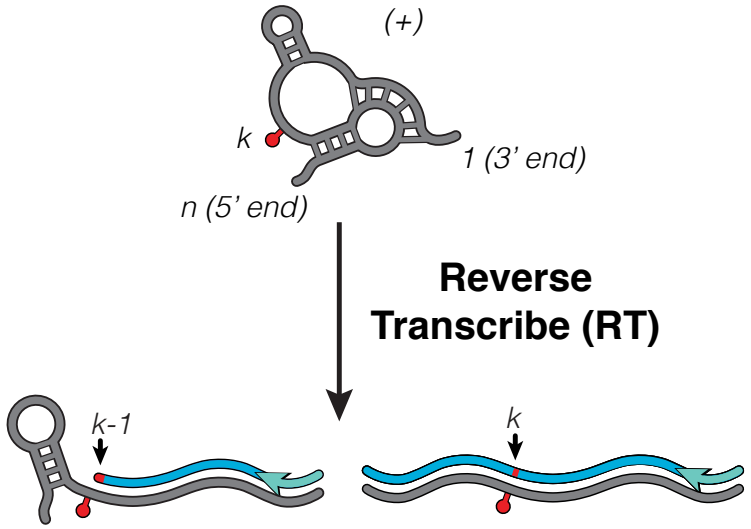

Figure 1. Reverse transcription (RT) encodes the positions of RNA adducts as either stops or mutations. RT converts an RNA (grey) into a complementary DNA (cDNA, blue) in a $3' \rightarrow 5'$ direction. (A defined RT priming site is shown in teal). If RT encounters an adduct at site k (red pin), one of two scenarios is possible: either the RT falls off 1 nt before the adduct (at site $k-1$, indicated as a red segment) to generate a k-fragment (left), or the RT proceeds through the adduct and intruduces a mutation at site k (indicated as a red segment, right). (+) refers to these RNAs being present in a population that has already been modified by the chemical probe to form covalent adducts. There are a range of RT priming strategies that can be used with this approach including those that can recover adduct positions at the 3' end of the molecule[14,18,19,31,32,36,43,44,46,51–57].

Subsequent steps in the experimental protocols then analyze cDNAs for signatures of RT-stops and RT-mutations. Early versions of the experiments analyzed cDNAs by separation techniques such as capillary electrophoresis that enabled the identification of RT-stops as distribution of different length cDNAs[22,24,26,27,58–64]. Later innovations developed high-throughput sequencing (HTS) methods to analyze cDNAs[10,18,19,40]. These experiments generally involve a series of ligation, size selection, and PCR steps in order to format cDNAs into a sequencing library with vendor-specific adapter sequences prior to sequencing[65].

The earliest versions of HTS experiments mapped RT-stops by mapping cDNA ends, and importantly enabled chemical probing to be performed on complex mixtures of RNAs since cDNA signatures from different RNAs could be distinguished through bioinformatic sequence alignment[18,66,67]. Later innovations showed that HTS approaches could also be used to map RT-mutation sites through bioinformatic analysis of sequence mutations[32,68]. These sequencing-based approaches represent an important enhancement in the amount and speed in which adduct detection information could be gleaned from these

experimental approaches.

Once adduct signatures are detected and collected as distributions of RT-stops or RT-mutations across each nucleotide of the corresponding RNA sequence, they can then be used to estimate a value for the 'reactivity' of each nucleotide to the chemical probe. Reactivities contain information about the underlying RNA structure, and specific reactivity values are a result of differences in the propensity of the chemical probe to react with bases that differ in structural context[69,70]. The main goal of data analysis procedures then is to estimate these underlying reactivity values as accurately as possible given the observed distributions of RT-stops and RT-mutations.

Historically, chemical probing experimental and data analysis approaches used the signatures from RT-stops to estimate reactivities at each position[66,67]. More recently, groups have begun push to use RT-mutations in order to define reactivity at a given position. These groups argue that inherent enzymatic biases in library prep cause distortions in RT-stop data that lead to calculated reactivities that do not directly correspond to the intrinsic reactivity of a given position. However, this assumes that the information given by RT-stops is correlated with information given by RT-mutations. However, recent papers[52,71] have shown not only that these metrics are poorly correlated but also, in some contexts, completely orthogonal. Specifically, recent analysis of DMS probing RT-stop and RT-mutation signatures on the same pool of modified RNAs show that different reverse transcriptase enzymes and reaction conditions show distinct biases for detecting adducts as either RT-stops or RT-mutations[52,71]. The orthogonality of RT-stops and RT-mutations and the variability between RT-stop and RT-mutations between different enzymes strongly suggest that approaches that only incorporate either RT-stops or RT-mutations miss information about adduct distributions and therefore the resulting reactivities may be incomplete and lacking in accuracy.

Conversely, these observations suggest that improvements in chemical probing accuracy can be achieved by incorporating both RT-stops and RT-mutations in the estimation of reactivities. To address this, here we developed a formalism for estimating chemical probing reactivities using both RT-stops and RT-mutations in a single experiment. Following the work of Aviran *et al.*[66,67], we extend a maximum-likelihood derivation of reactivities and present a reactivity formula that uses this combined information. Interestingly, this formula matches an intuitive interpretation of chemical probing reactivities as the fraction of adduct formed at each nucleotide at the end of the probing reaction. We discuss assumptions of this model, and end with a discussion on experimental approaches to validate this model.

Results

Model Setup

For an RNA of length n, we define the very 3' nucleotide of the RNA as position 1, and the very 5' nucleotide of the RNA as position n, and only consider cDNAs that start at position 1 (Fig. 1). We define a k-fragment as a cDNA whose 5' end begins at and complements position 1 of the RNA, and whose 3' end complements position $k-1$ of the RNA. We emphasize that we **assume that RT-stops and RT-mutations at a given adduct position in a single RNA molecule are mutually exclusive events: since an RT that stops due to an adduct at position k stops at position $k-1$, it cannot introduce a mutation at**

position k. This will be an important feature of the formula for recovering reactivities from observations of RT-stop and RT-mutation events, and we discuss implications for this assumption below.

The outcome of a typical high throughput sequencing (HTS) experiment to map RNA chemical probing adducts is a series of cDNA sequences that encode the locations of RT-stops and RT-mutations. These raw cDNA sequencing reads are then processed through read alignment software to generate a list of RT-stops and RT-mutations at each position within the target RNA. To keep track of these observed patterns, we define several variables that indicate stops (S) and mutations (M).

Since an RT that stops due to an adduct at position k stops and transcribes through position $k-1$, we define $S_k^{(+)}$ as the number of observed cDNA fragments whose 3' ends map to position $k-1$ and 5' ends map to position 1, where (+) indicates these fragments were observed from samples that had been treated with the chemical probe, called (+) channel samples. We additionally call $S_k^{(+)}$ the number of k-fragments because information on adduct formation would come from position k even though the sequenced length is of $k-1$. Note that if an RT does not stop internally, then RT will transcribe through the end of the cDNA and reach position $k=n$. Thus, $S_{n+1}^{(+)}$ represents the number of full-length reads observed in the (+) channel. Similarly, we define $M_{k,l}^{(+)}$ as the number of k-fragments observed that have at least one mutation and includes a mutation at position l in the (+) channel, where $l<k$.

It is also possible that RT stops or mutates at positions due to natural processes and/or there are cDNA sequencing errors that are not caused by chemical probe adducts. These events confound the measurement of adduct distributions and must be accounted for in data analysis to extract their confounding influence. To do this, control experiments are run that process the RNA in the same manner, but do not include the addition of the chemical probe. Such (-) channel experiments then generate a set of observed RT-stops and RT-mutations which we denote as $S_k^{(-)}$ and $M_{k,l}^{(-)}$, respectively.

Figure 2 shows several examples of possible RT-stop and RT-mutation scenarios in the (+) and (-) channel and how they would each contribute to $S_k^{(\pm)}$ and $M_{k,l}^{(\pm)}$.

Once the raw data is processed, our goal is to use the observed patterns of stops and mutations, $S_k^{(\pm)}$ and $M_{k,l}^{(\pm)}$, in a formula that more completely and accurately estimates the reactivity of each nucleotide of the interrogated RNAs to the chemical probe. Below we state the results of our derivation of this formula and discuss its implementation and limitations.

Estimating chemical probing reactivity from RT-stops and RT-mutations

We define the 'reactivity' of site k in an RNA molecule, r_k, as the probability of an adduct forming at that site during a chemical probing experiment. r_k contains structural information about the molecular fold of the RNA sample and is the primary data we want to extract from the probing experiment. Since RT can both stop and mutate at adducts, we also define β_k to be the probability that RT stops due to an adduct at site k following syntax in[66], and μ_k to be the probability that RT mutates due to an adduct at site k. Recent evidence suggests that there are strong context preferences for RT to favor either stops or mutations at a given position of an RNA, and so both must be accounted for in our estimate of r_k[52,72]. Since RT-stops and RT-mutations are mutually exclusive events when detecting adducts, we define the reactivity at site k to be the sum of these two probabilities: $r_k = \beta_k + \mu_k$.

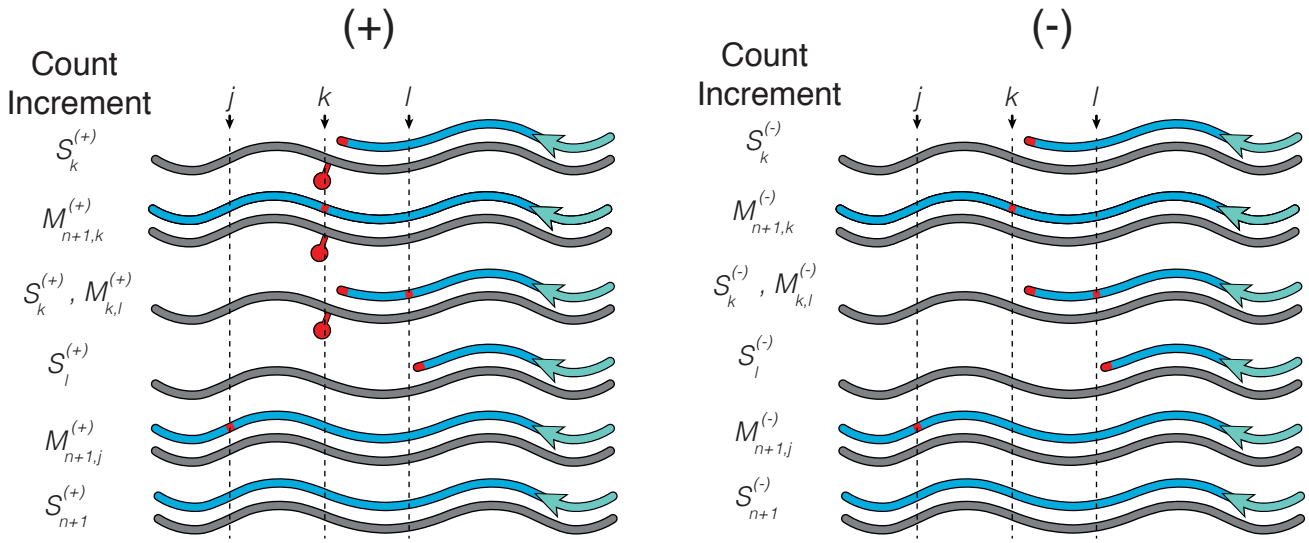

Figure 2. Examples of RT-stops and RT-mutations in the (+) and (-) channels and how they contribute to $S_k^{(\pm)}$ and $M_{k,l}^{(\pm)}$. Note that a single read may contribute to both stop and mutation counts. Adduct positions in the (+) channel are denoted by red pins.

Below we present a maximum-likelihood framework derivation following[66] for generating our best estimate of these probabilities $\{r_k^*, \beta_k^*, \mu_k^*\}$ given the observations of $S_k^{(\pm)}$ and $M_{k,l}^{(\pm)}$. Using this framework we obtain

$$r_k^* = \beta_k^* + \mu_k^*, 1 \leq k \leq n \tag{1}$$

$$\beta_k^* = \max\left\{\frac{\frac{stop_k^{(+)}}{depth_k^{(+)}} - \frac{stop_k^{(-)}}{depth_k^{(-)}}}{1 - \frac{stop_k^{(-)}}{depth_k^{(-)}}}, 0\right\} \tag{2}$$

$$\mu_k^* = \max\left\{\frac{\frac{mut_k^{(+)}}{depth_k^{(+)}} - \frac{mut_k^{(-)}}{depth_k^{(-)}}}{1 - \frac{mut_k^{(-)}}{depth_k^{(-)}}}, 0\right\} \tag{3}$$

where $depth_k^{\pm}$ is the number of sequencing reads that cover position k in each channel, which can be calculated from

$$depth_k^{(\pm)} = \sum_{l=k}^{n+1} S_l^{(\pm)} \tag{4}$$

and the mutations observed at position k, mut_k is defined as

$$mut_k^{(\pm)} = \sum_{l=k+1}^{n+1} M_{l,k}^{(\pm)} \tag{5}$$

We rewrite $S_k^{(\pm)}$ as $stop_k^{(\pm)}$ in Eq. 2 to more clearly delineate RT-stop and RT-mutation information.

$$stop_k^{(\pm)} = S_k^{(\pm)} \tag{6}$$

The quantities β_k^* and μ_k^* are the best feasible estimates for the reactivity components due to RT-stops and RT-mutations, respectively. As explained below, they come from $\widehat{\beta}_k, \widehat{\mu}_k$ which are the estimated parameters from observed data. Since $\widehat{\beta}_k$ and $\widehat{\mu}_k$ are each calculated as the difference between terms calculated from the (+) and (-) channel data, there is no guarantee that they will be strictly nonnegative. Therefore, in the case that the calculation results in $\widehat{\beta}_k < 0$ or $\widehat{\mu}_k < 0$, the feasible solution is to set them to zero[66], hence the use of the maximum function in defining β_k^* and μ_k^*. Note this arises in cases when there is severe dropoff or mutations in the (-) channel indicating a large amount of noise in the measurement at this position. It is therefore adviseable to keep track of scenarios where $\widehat{\beta}_k < 0$ or $\widehat{\mu}_k < 0$ since they are generally indications of poor data quality, especially for larger negative values, where experimental conditions may potentially be optimized.

Discussion

An intuitive link between reactivity estimates and fraction of adduct formed.

Interestingly Eq.1 has an intuitive interpretation that is related to the concept of the fraction of adduct formed at the end of a chemical probing reaction. Since $depth_k$ is an indication of how many adduct detection events are possible at each position, both RT-stops and RT-mutation data are incorporated in terms that have the form

$$\frac{\text{Number of RT Events Observed at } k}{\text{Number of Possible RT Events Observable at } k} = \text{Fraction of RT Events Observed at } k = f_k(\text{RT Event}),$$

where RT Event refers to either RT-stops or RT-mutations. Therefore we can write

$$\widehat{r_k} = \widehat{\beta}_k + \widehat{\mu}_k, 1 \leq k \leq n$$
$$\widehat{\beta}_k = \frac{f_k^{(+)}(stop) - f_k^{(-)}(stop)}{1 - f_k^{(-)}(stop)}$$
$$\widehat{\mu}_k = \frac{f_k^{(+)}(mut) - f_k^{(-)}(mut)}{1 - f_k^{(-)}(mut)}$$

Thus both $\widehat{\beta}_k$ and $\widehat{\mu}_k$ represent the fraction of RT events observed in the (+) channel corrected for the fraction of RT events observed in the (-) channel for stops and mutations, respectively. The denominators in each term represent the fraction of signal due to adduct that is possible to observe in the (+) channel. These denominators arise because an RT event observation can be due to an adduct *or* a background process, but not both[66] – i.e. if a fraction of RT events is observed in the (-) channel, the fraction of events that then *can* be observed in the (+) channel is reduced by that amount in order to estimate the fraction of events due to true signal. The denominators effectively correct for scenarios in which there is high background that obfuscates signal due to adducts. In cases where there is little or no background, these denominators can be approximated to be ~ 1 and we have

$$r_k^* \approx (f_k^{(+)}(stop) + f_k^{(+)}(mut)) - (f_k^{(-)}(stop) + f_k^{(-)}(mut))$$

Since the (+) channel has signal due to adducts and background processes, while the (-) channel only has signal due to background processes, the subtraction amounts to

$$r_k^* \approx f_k(stop\ due\ to\ adduct) + f_k(mut\ due\ to\ adduct),$$

where $f_k(event\ due\ to\ adduct)$ denotes RT-stops or RT-mutations at site k due to adduct and not due to natural fall off and mutations.

Since RT-stops and RT-mutations are the two ways to detect adducts, then

$$r_k^* \approx \text{fraction of adduct observed at position } k$$

Importantly, the fraction of adduct formed at any given position is a quantity that is determined by the chemical kinetics of the probing reaction and the structure-dependent fluctuations of each nucleotide of the RNA[47,69]. By estimating reactivities that correspond to the fraction of adduct observed, reactivity values should most closely align with the kinetics of the chemical probing reaction, which should allow a deeper understanding of data from high throughput RNA structure probing experiments.

Model Assumptions and Implementation

Two main assumptions were made in the above model:

1. Mutations at different positions are independent of each other.

2. Observing an RT-mutation at a given position is exclusive to observing an RT-stop at that position.

The first assumption is reasonable for low modification rates. However, some studies[47,73] indicate that the chemical adducts caused by certain probe reagents may alter the structural dynamics of the RNA once formed, which could in turn influence the formation of additional adducts if high modification rates are used - i.e. that certain chemical probes could potentially destabilize individual RNA molecules such that additional adduct formation to the same molecule would not reflect the native structures of interest. More work is needed to understand how multiple probe modifications can impact the ability to estimate reactivities which may depend on the nature of the probe, where it chemically modifies the RNA base, and any sequence/structural contexts of these scenarios.

The second assumption is more nuanced and impacts the implementation of the read mapping and application of the above reactivity formulas. In particular, the case of the very 3' end of a cDNA presents a challenging mapping case if it is mutated. According to the assumption, a mutated cDNA 3' end at position $k-1$ would contribute to two counts: S_k *and* $M_{k,k-1}$. However, it could be possible that in the process of stopping, RT introduces the mutation at the same time and thus this event should only be counted once. More work on biochemically defined adducts would be needed to validate or modify this assumption.

Experimental Validation

The major innovation in the above formula for chemical reactivities is to formally incorporate the observed signatures of *both* RT-stops and RT-mutations when estimating reactivities. The major motivation for this is the data and results of[52,72] which clearly show that RT drop and mutation signatures on the same pool of modified RNA differs depending on the specific RT enzyme and associated buffer conditions used to perform the conversion into cDNA. The goal of the RT process to convert every RNA adduct into a detectable signature in cDNAs therefore justifies our assertion that both RT-stops and RT-mutations should be included simultaneously in the reactivity estimation.

While the above formula for chemical probing reactivities makes intuitive sense as the fraction of adduct detected at each position, it still requires experimental validation. Accordingly, we expect two important improvements from applying the combined RT-stop+map model: improvement in chemical probing reactivity accuracy, and an invariance of reactivities to RT conditions.

Improvements in chemical probing reactivity accuracy are naturally expected since current approaches that focus solely on RT-stops[3,11,18,30,34,36] or RT-mutations[32,46] will inherently miss information[52,72]. Many approaches to assess chemical probing accuracy rely on an indirect method to first utilize reactivities in RNA secondary structure prediction algorithms, and then assess accuracy of reactivity data based on the improvements in structural prediction[74–77]. While we expect that RT-stop+map reactivities may improve the accuracy according to this benchmark, we anticipate that improvements may only be modest as it appears that the current RNA structure prediction algorithms are starting to reach the inherent limits of their accuracy given the assumptions used in their calculations[78]. Therefore methods that assess accuracy of reactivities through more direct analysis and/or new RNA structure prediction algorithms may be needed to uncover improvements when using both RT-stops and RT-mutations.

The other improvement suggested by the RT-stop+map model is an invariance of reactivities to RT conditions. In other words, estimated reactivity values should be the same no matter what RT enzyme or RT conditions are used. This is because the RT process is a means to detect RNA adducts. If the same pool of modified RNA is used, then this adduct distribution will not change, making it a goal of adduct detection methods to uncover the same distribution independent of the method conditions. Interestingly, this invariance is also suggested when observing the individual RT-stop and RT-mutate reactivities from Sexton *et al.*[52] and Novoa *et al.*[71], which strongly suggest that adding the two together would create reactivities that are highly similar between RT enzymes and conditions.

While invariance to RT conditions is strongly suggested as an outcome when using both RT-stops and RT-mutations, it is not guaranteed, mainly because it should only be true when reactivity estimates converges to the true fraction of adduct formed value. This can breakdown for simple reasons such as inadequate sequencing depth needed to overcome high background stops and mutations, or more complex reasons related to biases in specific library preparation steps. In particular, biases introduced by ligation or PCR steps that prevent adducts in specific sequence contexts to be uniformly sampled would interfere with more accurately estimating reactivities. More work is needed to test different experimental library preparation protocols in the context

of the RT-stop+map reactivity estimation in order to examine these effects. Interestingly, searching for library preparation strategies that are invariant to RT conditions may be a means to identify the most accurate experimental strategy. Thus, future work in this area may produce useful insights in both experimental protocols as well as more accurately estimating reactivities.

Conclusion

In this work, we derive an equation for estimating chemical probing reactivities that uses information from both RT-stops and RT-mutations. This is based off of recent work[52,72] that gives strong evidence that RT-stop and RT-mutation detection methods give complementary information when used with DMS probing - i.e. each method has context dependence such that they tend to map adducts in unique scenarios rather than mapping the same adduct positions. Therefore, we propose that reactivity estimation that considers both RT-stop and RT-mutation will be more accurate than methods that consider only one source of adduct detection. Future work will require the above formulas to be tested in a range of experimental contexts to demonstrate that the conclusions drawn from it are robust. We hope that these efforts will lead to improvements in RNA structure interrogation methods that are becoming increasingly important in answering questions about how RNA structure impacts a broad range of processes in biology.

Full Derivation

Model Setup

Chemical probing reactivities represent probabilities that adduct formation will occur at each nucleotide in an RNA during the probing reaction. Once the reaction proceeds to completion, these probabilities will naturally manifest themselves as a fraction of adduct formed at each position, which is defined as the proportion of those nucleotides that have the adduct out of the total population.

A given pattern of reactivities will naturally generate a distribution of RT-stops and RT-mutations across a population of cDNAs when the RNAs are reversed transcribed. Thus given a set of known reactivities, an observed pattern of RT-stops and RT-mutations could be calculated. However, in chemical probing experiments the information available is the converse - we know the pattern of RT-stops and RT-mutations, but do not know the underlying reactivities that gave rise to those patterns. The process of deriving a formula to estimate reactivities is thus to search over all possible reactivity values that lead to RT-stop and RT-mutate distributions that most accurately match the observed data. Fortunately this can be done exactly to yield a closed-form expression for the most accurate reactivity values possible from observed patterns of RT-stops and RT-mutations.

The derivation of equation (1) follows the maximum likelihood approach originally developed in[66] for the case of just detecting RT-stops. The maximum likelihood approach describes adduct detection by RT as a probabalistic process, where as RT processes through cDNA synthesis there are certain probabilities for it to fall off or mutate due to either encountering an

adduct or due to natural processes. We define the probabilities

$$\beta_k = \text{probability RT falls off due to adduct at site k,} \qquad 0 \leq \beta_k \leq 1$$

$$\mu_k = \text{probability RT mutates due to adduct at site k,} \qquad 0 \leq \mu_k \leq 1$$

where the ranges for β_k and μ_k are set since they are probabilities. When describing a complete model, we also need to account for the probabilities for RT to fall off or mutate due to natural processes, which are described by two additional probabilities

$$\gamma_k = \text{probability RT falls off due to natural processes at site k,} \qquad 0 \leq \gamma_k \leq 1$$

$$\delta_k = \text{probability RT mutates due to natural processes at site k,} \qquad 0 \leq \delta_k \leq 1$$

where again ranges for γ_k and δ_k are set since they are probabilities. If known, these probabilities define the reactivity information we desire from the chemical probing experiment from equation (1). However, for a given RNA these probabilities are unknown, and it is the goal of the maximum likelihood framework to estimate these probabilities given the information obtained from the sequencing reads. Estimated parameters from the maximum likelihood estimation are then $\{\widehat{\beta}_k, \widehat{\mu}_k, \widehat{\gamma}_k, \widehat{\delta}_k\}$ which we then enforce nonnegativity to obtain the estimated probabilities $\{\beta_k^*, \mu_k^*\}$ and thus the estimated reactivity $\{r_k^*\}$. For shorthand we define $B = \{\beta_k\}, \Gamma = \{\gamma_k\}, M = \{\mu_k\}, \Delta = \{\delta_k\}$.

Construction of the Likelihood Function

Even though the probabilities $\{B, M, \Gamma, \Delta\}$ are initially unknown, we can still use them to construct the overall probability of observing a specific type of sequencing read in the experiment. For example, if we only consider RT-stops, the probability that we would observe a k-fragment in the (-) channel would be

$$\begin{aligned} Prob(k\text{-}fragment\ in\ (-)) &= Prob(\text{RT-stops at } k | \text{RT does not stop before } k) \cdot Prob(\text{RT does not stop before } k) \\ &= \gamma_k \prod_{i=1}^{k-1} (1 - \gamma_i) \end{aligned}$$

The first term in the equation is the probability RT stops at position k, which is γ_k. The second term is the probability that RT does *not* stop before reaching position k. Since the probability of *not* stopping at a position i is $(1 - \gamma_i)$, then the probability of not stopping before k is the product of all the $(1 - \gamma_i)$ terms where $i < k$. Note that since we were describing events in the (-) channel, we use $\Gamma = (\gamma_1, ..., \gamma_n)$ since they describe RT stop events due to natural processes which are the only causes of RT stop events in the (-) channel. Following this logic, we can write the probability of observing a full length fragment in the (-)

channel as the product of RT not stopping the whole length of the molecule, or

$$\begin{aligned} Prob(full\ length\ fragment\ in\ (-)) &= Prob(\text{RT does not stop}), \forall i, 1 \leq i \leq n \\ &= Prob(\text{RT does not stop naturally}), \forall i, 1 \leq i \leq n \\ &= \prod_{i=1}^{n} (1-\gamma_i) \end{aligned}$$

When considering RT-stops in the (+) channel, things are slightly more complex since RT can stop both due to adducts present *and* natural processes as well. Thus when we write down probabilities for observing fragments in the (+) channel, these probabilities will involve both $B = (\beta_1, ..., \beta_n)$ and $\Gamma = (\gamma_1, ..., \gamma_n)$. When considering the probability for observing a full length fragment in the (+) channel, we simply need to include $(1-\beta_i)$ with $(1-\gamma_i)$ in the probability for RT *not* stopping at position i leading to

$$\begin{aligned} Prob(full\ length\ fragment\ in\ (+)) &= Prob(\text{RT does not stop}), \forall i, 1 \leq i \leq n \\ &= Prob(\text{RT does not stop naturally}) \cdot Prob(\text{RT does not stop due to adduct}), \forall i, 1 \leq i \leq n \\ &= \prod_{i=1}^{n} (1-\gamma_i)(1-\beta_i) \end{aligned}$$

For the probability of observing k-fragments in the (+) channel there are two independent causes for dropoff at k: either the k-fragment was due to an adduct at k or due to a natural dropoff at k:

$$\begin{aligned} Prob(k\text{-}fragment\ in\ (+)\ due\ to\ adduct) &= Prob(\text{RT-stop at } k \text{ due to adduct} | \text{RT does not stop before } k) \\ &\quad \cdot Prob(\text{RT does not stop before } k) \\ &= Prob(\text{RT-stop at } k \text{ due to adduct} | \text{RT does not stop before } k) \\ &\quad \cdot Prob(\text{RT does not stop naturally before } k) \\ &\quad \cdot Prob(\text{RT does not stop due to adduct before } k) \\ &= \beta_k \prod_{i=1}^{k-1} (1-\beta_i)(1-\gamma_i) \end{aligned}$$

Here, the first term in the equation is the probability RT stops at position k due to adduct, which is β_k. The second term is the probability that RT does *not* stop before reaching position k, and since the probability of *not* stopping at a position i in the (+) channel is $(1-\beta_i)(1-\gamma_i)$ (which is the probability of not stopping due to adduct *and* not stopping due to natural processes), then the probability of not stopping before k is the product of all the $(1-\beta_i)(1-\gamma_i)$ terms where $i < k$.

RT can also fall off in the (+) channel due to natural processes, so we must account for this probability as well:

$$
\begin{aligned}
Prob(k\text{-fragment in } (+) \text{ due to natural processes}) &= Prob(\text{RT-stop at } k \text{ due to natural processes}|\text{RT does not stop before } k) \\
&\quad \cdot Prob(\text{RT does not stop before } k) \\
&= Prob(\text{RT-stop at } k \text{ due to natural processes}|\text{RT does not stop before } k) \\
&\quad \cdot Prob(\text{RT does not stop naturally before } k) \\
&\quad \cdot Prob(\text{RT does not stop due to adduct before } k) \\
&= \gamma_k \prod_{i=1}^{k-1} (1-\beta_i)(1-\gamma_i)
\end{aligned}
$$

The first term, γ_k represents the probability of falling off due to natural processes. The other term in this equation describing the probability of no RT-stop before k remains the same.

However, we must disentangle RT fall off in the (+) channel due to natural processes versus adduct formation. This is particularly important for positions that have both a tendency for natural process RT-stops as well as adduct formation. In other words, a position k with the probability of adduct formation $\beta_k < 1$ could also create a k-fragment from natural processes as governed by the probability γ_k. We account for this in the following equation:

$$
\begin{aligned}
&Prob(k\text{-fragment in } (+) \text{ due to only natural processes}) \\
&\quad = Prob(\text{RT-stop at } k \text{ due to only natural processes}|\text{RT does not stop before } k) \cdot Prob(\text{RT does not stop before } k) \\
&\quad = Prob(\text{RT-stop at } k \text{ due to natural processes, No RT-stop at } k \text{ due to adduct}|\text{RT does not stop before } k) \\
&\qquad \cdot Prob(\text{RT does not stop before } k) \\
&\quad = Prob(\text{RT-stop at } k \text{ due to natural processes}|\text{RT does not stop before } k) \\
&\qquad \cdot Prob(\text{No RT-stop at } k \text{ due to adduct}|\text{RT does not stop before } k) \\
&\qquad \cdot Prob(\text{RT does not stop before } k) \\
&\quad = \gamma_k(1-\beta_k) \prod_{i=1}^{k-1} (1-\beta_i)(1-\gamma_i)
\end{aligned}
$$

Note the slightly different form of this probability. The first two terms represent the probability of falling off due to natural processes *and not* due to adduct, which is $\gamma_k(1-\beta_k)$. The last term in this equation describing the probability of no RT-stop before k remains the same. Since we have taken into account the possible scenarios, to find the probability of a k-fragment in the (+) channel due to either type of process, we simply sum them:

$$
\begin{aligned}
Prob(k\text{-}fragment\ in\ (+)) &= Prob(k\text{-}fragment \text{ in } (+) \text{ due to adduct or due to natural processes}) \\
&= Prob(k\text{-}fragment \text{ in } (+) \text{ due to adduct}) \\
&\quad + Prob(k\text{-}fragment \text{ in } (+) \text{ due to natural processes}) \\
&\quad - Prob(k\text{-}fragment \text{ in } (+) \text{ due to natural processes and adduct}) \\
&= Prob(k\text{-}fragment \text{ in } (+) \text{ due to adduct}) \\
&\quad + Prob(k\text{-}fragment \text{ in } (+) \text{ due to only natural processes}) \\
&= \beta_k \left[\prod_{i=1}^{k-1} (1-\beta_i)(1-\gamma_i) \right] + \gamma_k (1-\beta_k) \left[\prod_{i=1}^{k-1} (1-\beta_i)(1-\gamma_i) \right] \\
&= (\beta_k + \gamma_k (1-\beta_k)) \prod_{i=1}^{k-1} (1-\beta_i)(1-\gamma_i) \\
&= [1-(1-\gamma_k)(1-\beta_k)] \prod_{i=1}^{k-1} (1-\beta_i)(1-\gamma_i)
\end{aligned}
$$

The first term in square brackets in this equation represents the nature of RT falling off due to an adduct or a natural processes. Here $(1-\gamma_k)(1-\beta_k)$ is the probability of not falling off due to a natural process *and* not falling off due to adduct. Therefore $1-(1-\gamma_k)(1-\beta_k)$ is the probability of falling off due to one or the other. The last term in this equation describing the probability of no RT-stop before k remains the same.

The RT-stop-only terms are the same as used in[66] to derive a reactivity formula in terms of RT-stop events. Here we extend this to include the observation of mutations in cDNA products as well, which are governed by the $\{M, \Delta\}$ probabilities. Therefore we must extend the probability terms above to account for the different scenarios of observing specific patterns of mutations across the cDNA molecules. There are two important aspects of RT-mutations that we need to incorporate when constructing these probabilities. The first is the *assumption* that an RT-mutation at position k is *mutually exclusive* to an RT-stop at the same position. This is reasonable because RT stops one nucleotide *before* the roadblock it encounters (either adduct or some interfering element that contributes to natural drop off). Therefore an RT-stop due to a road block at position k results in a cDNA that ends at position $k-1$, which could therefore not have a mutation at position k. This mutually exclusive nature modifies the probabilities for observing a k-fragment in the (-) channels in the following way:

$$
Prob(k\text{-}fragment\ in\ (-)) = \gamma_k (1-\delta_k) \prod_{i=1}^{k-1} (1-\gamma_i) \tag{7}
$$

where the $\gamma_k(1-\delta_k)$ term reflects the mutually exclusive nature that if an RT-stops due to a natural process at position k (with probability γ_k), then it cannot introduce a mutation at position k (with the probability of not mutating $(1-\delta_k)$). Similarly,

the probability for observing a k-fragment in the (+) channel is modified to:

$$Prob(k\text{-}fragment\ in\ (+)) = [1-(1-\gamma_k)(1-\beta_k)](1-\delta_k)(1-\mu_k)\prod_{i=1}^{k-1}(1-\beta_i)(1-\gamma_i)$$

where we have incorporated the probability that there was no mutation at position k due to natural processes $(1-\delta_k)$ and no mututation at position k due to adduct $(1-\mu_k)$ with the $(1-\delta_k)(1-\mu_k)$ term.

The second aspect of mutations that we need to incorporate is the specific pattern of mutations that may be present across the rest of the cDNA molecule. We *assume* that mutations at different positions are independent of each other – i.e. if a mutation can occur at position i due to natural processes with probability δ_i, then the probability of observing mutations at positions i and j is just the product $\delta_i\delta_j$. If a full length cDNA in the (-) channel only had mutations at i and j and nowhere else, then the probability of observing this fragment would be $\delta_i\delta_j\prod^{n}_{\substack{l=1\\ l\notin\{i,j\}}}(1-\delta_l)$ since every other position other than i and j would not be mutated. In the same way, the probability of observing any pattern of mutations across a cDNA of length $k-1$ in the (-) channel can be written as:

$$Prob(mutations\ at\ \{l\}\ in\ (-)|fragment\ length\ of\ k-1) = \prod_{l}\delta_l\prod_{\substack{j=1\\ j\notin\{l\}}}^{k-1}(1-\delta_j) \tag{8}$$

where the notation $j\notin\{l\}$ indicates that the second product covers the positions j that are not mutated. For (+) channel cDNAs, we have a similar scenario as with RT-stops – RT-mutate events due to an adduct or natural processes are independent. An RT-mutate event due to natural processes at position k occurs with probability δ_k, while that due to an adduct alone occurs with probability $\mu_k(1-\delta_k)$. Summing these two then gives the probability that an RT-mutate event occurs at postion k in the (+) channel is $(1-(1-\delta_k)(1-\mu_k))$.[1] With this we can write

$$Prob(mutations\ at\ \{l\}\ in\ (+)|fragment\ length\ of\ k-1) = \prod_{l}[1-(1-\delta_k)(1-\mu_k)]\prod_{\substack{j=1\\ j\notin\{l\}}}^{k-1}(1-\delta_j)(1-\mu_j)$$

While we have written down all of the different probabilities for observing different types of cDNA fragments, we still do not know the true underlying $\{B,M,\Gamma,\Delta\}$ probabilities. However, we can estimate these numbers given the observed cDNA reads using maximum likelihood estimation. The overall concept of maximum likelihood estimation is that if we can find the set of underlying probabilities $\{\hat{B},\hat{M},\hat{\Gamma},\hat{\Delta}\}$ that is most consistent with our observed data, then this will be our best estimate of these parameters. To do so we first construct a likelihood function, $\mathscr{L}(\{B,M,\Gamma,\Delta\})$, which represents the likelihood that we would observe a given set of cDNA reads given the set of probabilities $\{B,M,\Gamma,\Delta\}$. To then find the set of $\{\hat{B},\hat{M},\hat{\Gamma},\hat{\Delta}\}$ most consistent with our data, we then maximize $\mathscr{L}(\{B,M,\Gamma,\Delta\})$ with respect to these parameters given our observed data.[2]

[1] Note an easy interpretation of this term is that the probability of not mutating due to a natural process and not mutating due to an adduct at k is $(1-\delta_k)(1-\mu_k)$, therefore the probability of mutating at this position is just 1 minus this.

[2] The maximum likelihood approach can be explained with the example of a coin flip experiment. Suppose we have a coin that has the probability of h for observing a heads and $1-h$ for observing a tails. Given h, the likelihood of us observing m heads and n tails in a series of $m+n$ flips is: $\mathscr{L}(h) = h^m(1-h)^n$. Suppose we do not know h, but we have observed m and n. The question is, what is our best estimate of h? We can achieve this by maximizing $\mathscr{L}(h)$, by

To construct $\mathscr{L}(\{B,M,\Gamma,\Delta\})$, we raise the probability of a particular observed event (p) to the N^{th} power, or p^N, where N is the number of times the event was observed. Since we have already constructed the probabilities for observing the different types of cDNA events, we can simply raise these probabilities to powers equal to the number of those types of reads observed. For example from equation (7), we can write

$$\mathscr{L}(observing\ S_k^{(-)}\ fragments) = \left[\gamma_k(1-\delta_k)\prod_{i=1}^{k-1}(1-\gamma_i)\right]^{S_k^{(-)}} \tag{9}$$

Note however that this does not consider the pattern of mutations that may be present in specific (-) channel k-fragments. Since mutations away from the stop site are assumed to occur independently, we can simply multiply the term above to the likelihood of observing a specific pattern of mutations. To construct the likelihood functions for mutations in a k-fragment, we must take into account two different observations: the number of k-fragments observed with a mutation at a specific position l, $M_{k,l}$, and the number of fragments observed that were *unmutated* at position l, $U_{k,l}$. (This is similar to accounting for the number of heads in a coin flip and the number of not heads (tails) in the footnote example.) The latter must be taken into account to account for all observed fragments. Since mutation events are independent, the likelihood of observing $M_{k,l}, U_{k,l}$ is then according to the equation (8)

$$\mathscr{L}(observing\ M_{k,l}^{(-)}, U_{k,l}^{(-)} | fragment\ length\ of\ k-1) = \prod_{l=1}^{k-1}\delta_l^{M_{k,l}^{(-)}}(1-\delta_l)^{U_{k,l}^{(-)}} \tag{10}$$

Note the above equation is essentially obtained by multiplying versions of equation (8) together for every cDNA observed. Rearranging all of the different products together will naturally collapse the terms into the form above which raises the probabilities of observing a mutation or not observing a mutation to the number of times those events were observed. Combining equations (9) and (10) we have

$$\mathscr{L}(observing\ S_k^{(-)}\ fragments\ with\ M_{k,l}^{(-)}, U_{k,l}^{(-)}) = \left[\gamma_k(1-\delta_k)\prod_{i=1}^{k-1}(1-\gamma_i)\right]^{S_k^{(-)}}\prod_{l=1}^{k-1}\delta_l^{M_{k,l}^{(-)}}(1-\delta_l)^{U_{k,l}^{(-)}}$$

Similarly, for full length fragments in the (-) channel, we have:

$$\mathscr{L}(observing\ S_{n+1}^{(-)}\ fragments\ with\ M_{n+1,l}^{(-)}, U_{n+1,l}^{(-)}) = \left[\prod_{i=1}^{n}(1-\gamma_i)\right]^{S_{n+1}^{(-)}}\prod_{l=1}^{n}\delta_l^{M_{n+1,l}^{(-)}}(1-\delta_l)^{U_{n+1,l}^{(-)}}$$

maximizing $\log(\mathscr{L}(h))$. Since $\log(\mathscr{L}(h)) = m\log(h) + n\log(h-1)$, $\frac{\partial\log(\mathscr{L}(h))}{\partial h} = \frac{m}{h} - \frac{n}{1-h} = 0$, which has the solution $\hat{h} = \frac{m}{m+n}$. It is easy to show that $\frac{\partial^2\log(\mathscr{L}(h)}{\partial h^2} < 0$ so this is a maximum. Thus our best estimate of h is just the fraction of heads observed. The maximum likelihood approach used here to estimate reactivities is just a more elaborate example of this coin flip experiment.

Note the $(1-\delta_n)$ term is incorporated into the right hand side of this equation. The full likelihood function takes into account all of the observed $S_k^{(+)}$ and $S_k^{(-)}$ fragments from $k=1,\ldots,n$, as well as the full length fragments $S_{n+1}^{(\pm)}$. We can therefore write

$$\begin{aligned}\mathscr{L}(\{B,M,\Gamma,\Delta\}) = \prod_{k=1}^{n}\Bigg\{ &\left[\gamma_k(1-\delta_k)\prod_{i=1}^{k-1}(1-\gamma_i)\right]^{S_k^{(-)}}\prod_{l=1}^{k-1}\delta_l^{M_{k,l}^{(-)}}(1-\delta_l)^{U_{k,l}^{(-)}} \\ &\times\left[(1-(1-\gamma_k)(1-\beta_k))(1-\delta_k)(1-\mu_k)\prod_{i=1}^{k-1}(1-\beta_i)(1-\gamma_i)\right]^{S_k^{(+)}} \\ &\times\prod_{l=1}^{k-1}(1-(1-\delta_l)(1-\mu_l))^{M_{k,l}^{(+)}}((1-\delta_l)(1-\mu_l))^{U_{k,l}^{(+)}}\Bigg\} \\ &\times\left[\prod_{i=1}^{n}(1-\gamma_i)\right]^{S_{n+1}^{(-)}}\prod_{l=1}^{n}\delta_l^{M_{n+1,l}^{(-)}}(1-\delta_l)^{U_{n+1,l}^{(-)}} \\ &\times\left[\prod_{i=1}^{n}(1-\gamma_i)(1-\beta_i)\right]^{S_{n+1}^{(+)}}\prod_{l=1}^{n}(1-(1-\delta_l)(1-\mu_l))^{M_{n+1,l}^{(+)}}((1-\delta_l)(1-\mu_l))^{U_{n+1,l}^{(+)}}\end{aligned} \tag{11}$$

where we have combined terms for k-fragments observed in the (-) and (+) channels, and complete fragments observed in the (-) and (+) channels in order.

Maximizing the Likelihood Function

Our next goal is to maximize $\mathscr{L}$ with respect to $\{B,M,\Gamma,\Delta\}$ (Eq. 11), given our observations of $S_k^{(\pm)}$, $M_{k,l}^{(\pm)}$, and $U_{k,l}^{(\pm)}$. To do this, we first take the logarithm of $\mathscr{L}$ to separate variables in order to more easily find the ML estimates and obtain:

$$\begin{aligned}\log(\mathscr{L}(\{B,M,\Gamma,\Delta\})) = \log\Bigg(\prod_{k=1}^{n}\Bigg\{ &\left[\gamma_k(1-\delta_k)\prod_{i=1}^{k-1}(1-\gamma_i)\right]^{S_k^{(-)}}\prod_{l=1}^{k-1}\delta_l^{M_{k,l}^{(-)}}(1-\delta_l)^{U_{k,l}^{(-)}} \\ &\times\left[(1-(1-\gamma_k)(1-\beta_k))(1-\delta_k)(1-\mu_k)\prod_{i=1}^{k-1}(1-\beta_i)(1-\gamma_i)\right]^{S_k^{(+)}} \\ &\times\prod_{l=1}^{k-1}(1-(1-\delta_l)(1-\mu_l))^{M_{k,l}^{(+)}}((1-\delta_l)(1-\mu_l))^{U_{k,l}^{(+)}}\Bigg\} \\ &\times\left[\prod_{i=1}^{n}(1-\gamma_i)\right]^{S_{n+1}^{(-)}}\prod_{l=1}^{n}\delta_l^{M_{n+1,l}^{(-)}}(1-\delta_l)^{U_{n+1,l}^{(-)}} \\ &\times\left[\prod_{i=1}^{n}(1-\gamma_i)(1-\beta_i)\right]^{S_{n+1}^{(+)}}\prod_{l=1}^{n}(1-(1-\delta_l)(1-\mu_l))^{M_{n+1,l}^{(+)}}((1-\delta_l)(1-\mu_l))^{U_{n+1,l}^{(+)}}\Bigg)\end{aligned} \tag{12}$$

We reduce Eq. 12 further:

$$
\begin{aligned}
\log(\mathscr{L}(\{B,M,\Gamma,\Delta\})) &= \sum_{k=1}^{n}\Bigg\{S_k^{(-)}\left[\log(\gamma_k)+\log(1-\delta_k)+\sum_{i=1}^{k-1}\log(1-\gamma_i)\right]+\sum_{l=1}^{k-1}\left[M_{k,l}^{(-)}\log(\delta_l)+U_{k,l}^{(-)}\log(1-\delta_l)\right]\\
&+S_k^{(+)}\left[\log(1-(1-\gamma_k)(1-\beta_k))+\log(1-\delta_k)+\log(1-\mu_k)+\sum_{i=1}^{k-1}\log[(1-\beta_i)(1-\gamma_i)]\right]\\
&+\sum_{l=1}^{k-1}\left[M_{k,l}^{(+)}\log(1-(1-\delta_l)(1-\mu_l))+U_{k,l}^{(+)}\log((1-\delta_l)(1-\mu_l))\right]\Bigg\}\\
&+S_{n+1}^{(-)}\sum_{i=1}^{n}\log(1-\gamma_i)+\sum_{l=1}^{n}\left[M_{n+1,l}^{(-)}\log(\delta_l)+U_{n+1,l}^{(-)}\log(1-\delta_l)\right]\\
&+S_{n+1}^{(+)}\sum_{i=1}^{n}\log[(1-\gamma_i)(1-\beta_i)]+\sum_{l=1}^{n}\left[M_{n+1,l}^{(+)}\log(1-(1-\delta_l)(1-\mu_l))+U_{n+1,l}^{(+)}\log[(1-\delta_l)(1-\mu_l)]\right]\\
&=\sum_{k=1}^{n}\Bigg\{S_k^{(-)}\log(\gamma_k)+(S_k^{(-)}+S_k^{(+)})\left[\log(1-\delta_k)+\sum_{i=1}^{k-1}\log(1-\gamma_i)\right]\\
&+\sum_{l=1}^{k-1}\left[M_{k,l}^{(-)}\log(\delta_l)+(U_{k,l}^{(-)}+U_{k,l}^{(+)})\log(1-\delta_l)\right]\\
&+S_k^{(+)}\left[\log(1-(1-\gamma_k)(1-\beta_k))+\log(1-\mu_k)+\sum_{i=1}^{k-1}\log(1-\beta_i)\right]\\
&+\sum_{l=1}^{k-1}\left[M_{k,l}^{(+)}\log(1-(1-\delta_l)(1-\mu_l))+U_{k,l}^{(+)}\log(1-\mu_l)\right]\Bigg\}\\
&+(S_{n+1}^{(-)}+S_{n+1}^{(+)})\sum_{i=1}^{n}\log(1-\gamma_i)+\sum_{l=1}^{n}\left[M_{n+1,l}^{(-)}\log(\delta_l)+(U_{n+1,l}^{(-)}+U_{n+1,l}^{(+)})\log(1-\delta_l)\right]\\
&+S_{n+1}^{(+)}\sum_{i=1}^{n}\log(1-\beta_i)+\sum_{l=1}^{n}\left[M_{n+1,l}^{(+)}\log(1-(1-\delta_l)(1-\mu_l))+U_{n+1,l}^{(+)}\log(1-\mu_l)\right]\\
&=\sum_{k=1}^{n}\left\{S_k^{(-)}\log(\gamma_k)+(S_k^{(-)}+S_k^{(+)})\log(1-\delta_k)+S_k^{(+)}\left[\log(1-(1-\gamma_k)(1-\beta_k))+\log(1-\mu_k)\right]\right\}\\
&+\sum_{k=1}^{n+1}\Bigg\{\sum_{i=1}^{k-1}\Big[(S_k^{(-)}+S_k^{(+)})\log(1-\gamma_i)+M_{k,i}^{(-)}\log(\delta_i)+(U_{k,i}^{(-)}+U_{k,i}^{(+)})\log(1-\delta_i)+S_k^{(+)}\log(1-\beta_i)\\
&+M_{k,i}^{(+)}\log(1-(1-\delta_i)(1-\mu_i))+U_{k,i}^{(+)}\log(1-\mu_i)\Big]\Bigg\}
\end{aligned}
\tag{13}
$$

Using the identities $\sum_{k=1}^{n+1}\sum_{i=1}^{k-1}a_{k,i}b_i=\sum_{k=1}^{n}\sum_{i=k+1}^{n+1}a_{i,k}b_k$ and $\sum_{k=1}^{n+1}\sum_{i=1}^{k-1}a_kb_i=\sum_{k=1}^{n}\sum_{i=k+1}^{n+1}a_ib_k$ we can rearrange (13) to

more easily find the maximum of the likelihood function.

$$\begin{aligned}
\log(\mathscr{L}(\{B,M,\Gamma,\Delta\})) &= \sum_{k=1}^{n}\left\{S_k^{(-)}\log(\gamma_k)+(S_k^{(-)}+S_k^{(+)})\log(1-\delta_k)+S_k^{(+)}\left[\log(1-(1-\gamma_k)(1-\beta_k))+\log(1-\mu_k)\right]\right\} \\
&+\sum_{k=1}^{n}\left\{\sum_{i=k+1}^{n+1}\left[(S_i^{(-)}+S_i^{(+)})\log(1-\gamma_k)+M_{i,k}^{(-)}\log(\delta_k)+(U_{i,k}^{(-)}+U_{i,k}^{(+)})\log(1-\delta_k)+S_i^{(+)}\log(1-\beta_k)\right.\right. \\
&\left.\left.+M_{i,k}^{(+)}\log(1-(1-\delta_k)(1-\mu_k))+U_{i,k}^{(+)}\log(1-\mu_k)\right]\right\} \\
&= \sum_{k=1}^{n}\left\{S_k^{(-)}\log(\gamma_k)+(S_k^{(-)}+S_k^{(+)})\log(1-\delta_k)+S_k^{(+)}\left[\log(1-(1-\gamma_k)(1-\beta_k))+\log(1-\mu_k)\right]\right. \\
&+\sum_{i=k+1}^{n+1}\left[(S_i^{(-)}+S_i^{(+)})\log(1-\gamma_k)+M_{i,k}^{(-)}\log(\delta_k)+(U_{i,k}^{(-)}+U_{i,k}^{(+)})\log(1-\delta_k)+S_i^{(+)}\log(1-\beta_k)\right. \\
&\left.\left.+M_{i,k}^{(+)}\log(1-(1-\delta_k)(1-\mu_k))+U_{i,k}^{(+)}\log(1-\mu_k)\right]\right\} \\
&\equiv \sum_{k=1}^{n} l_k
\end{aligned} \tag{14}$$

Let $A=\sum_{i=k+1}^{n+1}(S_i^{(-)}+S_i^{(+)})$, $B=\sum_{i=k+1}^{n+1}M_{i,k}^{(-)}$, $C=\sum_{i=k+1}^{n+1}(U_{i,k}^{(-)}+U_{i,k}^{(+)})$, $D=\sum_{i=k+1}^{n+1}S_i^{(+)}$, $E=\sum_{i=k+1}^{n+1}M_{i,k}^{(+)}$, and $F=\sum_{i=k+1}^{n+1}U_{i,k}^{(+)}$. Then from (14):

$$\begin{aligned}
l_k &= S_k^{(-)}\log(\gamma_k)+(S_k^{(-)}+S_k^{(+)})\log(1-\delta_k)+A\log(1-\gamma_k)+B\log(\delta_k)+C\log(1-\delta_k) \\
&+S_k^{(+)}\left[\log(1-(1-\gamma_k)(1-\beta_k))+\log(1-\mu_k)\right]+D\log(1-\beta_k)+E\log(1-(1-\delta_k)(1-\mu_k))+F\log(1-\mu_k)
\end{aligned} \tag{15}$$

From (15) we can calculate partial derivatives with respect to each parameter in the likelihood function. By setting these paritial derivatives to 0, we can find a critical point for each parameter, which we later show these critical points are maxima and thus are maximum likelihood estimators. Since (14) is a sum of independent terms, we can maximize these independently. Starting with $\frac{\partial l_k}{\partial \beta_k}$:

$$\frac{\partial l_k}{\partial \beta_k} = \frac{S_k^{(+)}(1-\gamma_k)}{1-(1-\gamma_k)(1-\beta_k)} - \frac{D}{1-\beta_k} = 0$$

$$S_k^{(+)}(1-\gamma_k)(1-\beta_k) - D + D(1-\gamma_k)(1-\beta_k) = 0$$

$$\therefore (1-\gamma_k)(1-\beta_k) = \frac{D}{S_k^{(+)}+D} \tag{16}$$

$$\therefore 1-(1-\gamma_k)(1-\beta_k) = \frac{S_k^{(+)}}{S_k^{(+)}+D} \tag{17}$$

We then use (16) and (17) to calculate $\widehat{\gamma_k}$ from $\frac{\partial l_k}{\partial \gamma_k}$.

$$\frac{\partial l_k}{\partial \gamma_k} = \frac{S_k^{(-)}}{\gamma_k} - \frac{A}{1-\gamma_k} + \frac{S_k^{(+)}(1-\beta_k)}{1-(1-\gamma_k)(1-\beta_k)} = 0$$

$$S_k^{(-)}(1-\gamma_k)(1-(1-\gamma_k)(1-\beta_k)) - A(\gamma_k)(1-(1-\gamma_k)(1-\beta_k)) + S_k^{(+)}(\gamma_k)(1-\gamma_k)(1-\beta_k) = 0$$

$$S_k^{(-)}(1-(1-\gamma_k)(1-\beta_k)) = \gamma_k\left[(S_k^{(-)}+A)(1-(1-\gamma_k)(1-\beta_k)) - S_k^{(+)}(1-\gamma_k)(1-\beta_k)\right]$$

The solution $\widehat{\gamma_k}$ then is

$$\begin{aligned}\widehat{\gamma_k} &= \frac{S_k^{(-)}(1-(1-\gamma_k)(1-\beta_k))}{(S_k^{(-)}+A)(1-(1-\gamma_k)(1-\beta_k)) - S_k^{(+)}(1-\gamma_k)(1-\beta_k)} = \frac{\frac{S_k^{(-)}S_k^{(+)}}{S_k^{(+)}+D}}{\frac{(S_k^{(-)}+A)S_k^{(+)}}{S_k^{(+)}+D} - \frac{S_k^{(+)}D}{S_k^{(+)}+D}} = \frac{S_k^{(-)}}{S_k^{(-)}+A-D} \\ &= \frac{S_k^{(-)}}{S_k^{(-)}+\sum_{i=k+1}^{n+1} S_i^{(-)}} = \frac{S_k^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}\end{aligned} \tag{18}$$

Let $G = \frac{D}{S_k^{(+)}+D}$. Using G, (16), (17), and (18) we can solve for $\widehat{\beta_k}$:

$$1-\beta_k = \frac{G}{1-\gamma_k}$$

$$\widehat{\beta_k} = 1 - \frac{G}{1-\gamma_k} = \frac{1-G-\gamma_k}{1-\gamma_k} = \frac{\frac{S_k^{(+)}}{S_k^{(+)}+D} - \gamma_k}{1-\gamma_k} = \frac{\frac{S_k^{(+)}}{\sum_{i=k}^{n+1} S_i^{(+)}} - \frac{S_k^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}}{1-\frac{S_k^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}} \tag{19}$$

We can solve for $\widehat{\mu_k}$ and $\widehat{\delta_k}$ similarly:

$$\frac{\partial l_k}{\partial \mu_k} = -\frac{S_k^{(+)}}{1-\mu_k} + \frac{E(1-\delta_k)}{1-(1-\delta_k)(1-\mu_k)} - \frac{F}{1-\mu_k} = 0$$

$$E(1-\delta_k)(1-\mu_k) - (S_k^{(+)}+F)(1-(1-\delta_k)(1-\mu_k)) = 0$$

$$(1-\delta_k)(1-\mu_k)[E+S_k^{(+)}+F] = S_k^{(+)}+F$$

$$\therefore (1-\delta_k)(1-\mu_k) = \frac{S_k^{(+)}+F}{E+S_k^{(+)}+F} \tag{20}$$

$$\therefore 1-(1-\delta_k)(1-\mu_k) = \frac{E}{E+S_k^{(+)}+F} \tag{21}$$

We then use (20) and (21) to calculate $\widehat{\delta_k}$ from $\frac{\partial l_k}{\partial \delta_k}$.

$$\frac{\partial l_k}{\partial \delta_k} = -\frac{S_k^{(-)}+S_k^{(+)}}{1-\delta_k} + \frac{B}{\delta_k} - \frac{C}{1-\delta_k} + \frac{E(1-\mu_k)}{1-(1-\delta_k)(1-\mu_k)} = 0$$

$$B(1-(1-\delta_k)(1-\mu_k)) - (\delta_k)[S_k^{(-)}+S_k^{(+)}+C+B)(1-(1-\delta_k)(1-\mu_k)) - E(1-\delta_k)(1-\mu_k)] = 0$$

$$\delta_k[(S_k^{(-)}+S_k^{(+)}+C+B)(1-(1-\delta_k)(1-\mu_k)) - E(1-\delta_k)(1-\mu_k)] = B(1-(1-\delta_k)(1-\mu_k))$$

$$\begin{aligned}\widehat{\delta_k} &= \frac{B(1-(1-\delta_k)(1-\mu_k))}{(S_k^{(-)}+S_k^{(+)}+C+B)(1-(1-\delta_k)(1-\mu_k)) - E(1-\delta_k)(1-\mu_k)} = \frac{BE}{(S_k^{(-)}+S_k^{(+)}+B+C)E - E(S_k^{(+)}+F)} \\ &= \frac{B}{S_k^{(-)}+B+C-F} = \frac{\sum_{i=k+1}^{n+1} M_{i,k}^{(-)}}{S_k^{(-)}+\sum_{i=k+1}^{n+1}(M_{i,k}^{(-)}+U_{i,k}^{(-)})} = \frac{\sum_{i=k+1}^{n+1} M_{i,k}^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}\end{aligned} \tag{22}$$

Where we have used $M_{i,k}^{(\pm)} + U_{i,k}^{(\pm)} = S_i^{(\pm)}$ since the former is all i-fragments that are either mutated at k or not, which is just $S_i^{(\pm)}$. Let $H = \frac{S_k^{(+)}+F}{E+S_k^{(+)}+F}$. Using H, (20), (21), and (22) we can solve for $\widehat{\mu_k}$:

$$1-\mu_k = \frac{H}{1-\delta_k}$$

$$\widehat{\mu_k} = 1 - \frac{H}{1-\delta_k} = \frac{1-H-\delta_k}{1-\delta_k} = \frac{\frac{E}{E+S_k^{(+)}+F} - \delta_k}{1-\delta_k} = \frac{\frac{\sum_{i=k+1}^{n+1} M_{i,k}^{(+)}}{\sum_{i=k}^{n+1} S_i^{(+)}} - \frac{\sum_{i=k+1}^{n+1} M_{i,k}^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}}{1 - \frac{\sum_{i=k+1}^{n+1} M_{i,k}^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}} \tag{23}$$

Thus, we have calculated the maximum likelihood estimators $\{\widehat{\beta_k}, \widehat{\mu_k}, \widehat{\gamma_k}, \widehat{\delta_k}\}$ as defined in equations 19, 23, 18, 22.

We additionally use Eq. 5, 4, 6 and propose that the estimated reactivity $\widehat{r_k}$ is then:

$$\begin{aligned}
\widehat{r_k} &= \widehat{\beta_k} + \widehat{\mu_k} \\
&= \frac{\frac{S_k^{(+)}}{\sum_{i=k}^{n+1} S_i^{(+)}} - \frac{S_k^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}}{1 - \frac{S_k^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}} + \frac{\frac{\sum_{i=k+1}^{n+1} M_{i,k}^{(+)}}{\sum_{i=k}^{n+1} S_i^{(+)}} - \frac{\sum_{i=k+1}^{n+1} M_{i,k}^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}}{1 - \frac{\sum_{i=k+1}^{n+1} M_{i,k}^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}} \\
&= \frac{\frac{S_k^{(+)}}{\sum_{i=k}^{n+1} S_i^{(+)}} - \frac{S_k^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}}{1 - \frac{S_k^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}} + \frac{\frac{mut_k^{(+)}}{\sum_{i=k}^{n+1} S_i^{(+)}} - \frac{mut_k^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}}{1 - \frac{mut_k^{(-)}}{\sum_{i=k}^{n+1} S_i^{(-)}}} \\
&= \frac{\frac{S_k^{(+)}}{depth_k^{(+)}} - \frac{S_k^{(-)}}{depth_k^{(-)}}}{1 - \frac{S_k^{(-)}}{depth_k^{(-)}}} + \frac{\frac{mut_k^{(+)}}{depth_k^{(+)}} - \frac{mut_k^{(-)}}{depth_k^{(-)}}}{1 - \frac{mut_k^{(-)}}{depth_k^{(-)}}} \\
&= \frac{\frac{stop_k^{(+)}}{depth_k^{(+)}} - \frac{stop_k^{(-)}}{depth_k^{(-)}}}{1 - \frac{stop_k^{(-)}}{depth_k^{(-)}}} + \frac{\frac{mut_k^{(+)}}{depth_k^{(+)}} - \frac{mut_k^{(-)}}{depth_k^{(-)}}}{1 - \frac{mut_k^{(-)}}{depth_k^{(-)}}}
\end{aligned}$$

However in practice with real data, $\widehat{\beta_k}, \widehat{\mu_k}, \widehat{\gamma_k}, \widehat{\delta_k}$ are not guaranteed to be between 0 and 1 inclusive for all k. Thus we enforce nonnegativity and our final reactivity calculation r_k^* is as outlined above in Eq. 1, 2, 3 and below:

$$\begin{aligned}
r_k^* &= \beta_k^* + \mu_k^*, 1 \le k \le n \\
\beta_k^* &= \max\left\{ \frac{\frac{stop_k^{(+)}}{depth_k^{(+)}} - \frac{stop_k^{(-)}}{depth_k^{(-)}}}{1 - \frac{stop_k^{(-)}}{depth_k^{(-)}}}, 0 \right\} \\
\mu_k^* &= \max\left\{ \frac{\frac{mut_k^{(+)}}{depth_k^{(+)}} - \frac{mut_k^{(-)}}{depth_k^{(-)}}}{1 - \frac{mut_k^{(-)}}{depth_k^{(-)}}}, 0 \right\}
\end{aligned}$$

We note the formula for $\{\beta_k^*\}$ is identical to that of[66].

Maximum Likelihood Estimators are Maxima

Next, we show that $\{\widehat{\beta_k}, \widehat{\mu_k}, \widehat{\gamma_k}, \widehat{\delta_k}\}$ are maxima by taking second partial derivatives of the likelihood equation l_k.

For $\widehat{\gamma_k}$:

$$\frac{\partial^2 l_k}{\partial \gamma_k \partial \gamma_k} = \frac{-S_k^{(-)}}{\gamma_k^2} - \frac{A}{(1-\gamma_k)^2} - \frac{S_k^{(+)}(1-\beta_k)^2}{(1-(1-\gamma_k)(1-\beta_k))^2} \leq 0$$
$$\frac{\partial^2 l_k}{\partial \gamma_k \partial \beta_k} = \frac{-S_k^{(+)}(1-\beta_k)(1-\gamma_k)}{(1-(1-\gamma_k)(1-\beta_k))^2} \leq 0$$
$$\frac{\partial^2 l_k}{\partial \gamma_k \partial \mu_k} = 0$$
$$\frac{\partial^2 l_k}{\partial \gamma_k \partial \delta_k} = 0$$

For $\widehat{\beta_k}$:

$$\frac{\partial^2 l_k}{\partial \beta_k \partial \gamma_k} = \frac{-S_k^{(+)}(1-\beta_k)(1-\gamma_k)}{(1-(1-\gamma_k)(1-\beta_k))^2} \leq 0$$
$$\frac{\partial^2 l_k}{\partial \beta_k \partial \beta_k} = \frac{-S_k^{(+)}(1-\gamma_k)^2}{(1-(1-\gamma_k)(1-\beta_k))^2} \leq 0$$
$$\frac{\partial^2 l_k}{\partial \beta_k \partial \mu_k} = 0$$
$$\frac{\partial^2 l_k}{\partial \beta_k \partial \delta_k} = 0$$

For $\widehat{\delta_k}$:

$$\frac{\partial^2 l_k}{\partial \delta_k \partial \gamma_k} = 0$$
$$\frac{\partial^2 l_k}{\partial \delta_k \partial \beta_k} = 0$$
$$\frac{\partial^2 l_k}{\partial \delta_k \partial \mu_k} = \frac{-E(1-\delta_k)(1-\mu_k)}{(1-(1-\delta_k)(1-\mu_k))^2} \leq 0$$
$$\frac{\partial^2 l_k}{\partial \delta_k \partial \delta_k} = \frac{-(S_k^{(-)} + S_k^{(+)})}{(1-\delta_k)^2} - \frac{B}{\delta_k^2} - \frac{C}{(1-\delta_k)^2} - \frac{E(1-\mu_k)^2}{(1-(1-\mu_k)(1-\delta_k))^2} \leq 0$$

For $\widehat{\mu_k}$:

$$\frac{\partial^2 l_k}{\partial \mu_k \partial \gamma_k} = 0$$
$$\frac{\partial^2 l_k}{\partial \mu_k \partial \beta_k} = 0$$
$$\frac{\partial^2 l_k}{\partial \mu_k \partial \mu_k} = \frac{-S_k^{(+)}}{(1-\mu_k)^2} - \frac{E(1-\delta_k)^2}{(1-(1-\mu_k)(1-\delta_k))^2} - \frac{F}{(1-\mu_k)^2} \leq 0$$
$$\frac{\partial^2 l_k}{\partial \mu_k \partial \delta_k} = \frac{-E(1-\delta_k)(1-\mu_k)}{(1-(1-\delta_k)(1-\mu_k))^2} \leq 0$$

When $0 < \beta_k, \mu_k, \gamma_k, \delta_k < 1$ and $S_k^{(-)}, S_k^{(+)} > 0$, $\left\{ \frac{\partial^2 l_k}{\partial \gamma_k \partial \gamma_k}, \frac{\partial^2 l_k}{\partial \gamma_k \partial \beta_k}, \frac{\partial^2 l_k}{\partial \beta_k \partial \gamma_k}, \frac{\partial^2 l_k}{\partial \beta_k \partial \beta_k}, \frac{\partial^2 l_k}{\partial \delta_k \partial \mu_k}, \frac{\partial^2 l_k}{\partial \delta_k \partial \delta_k}, \frac{\partial^2 l_k}{\partial \mu_k \partial \mu_k}, \frac{\partial^2 l_k}{\partial \mu_k \partial \delta_k} \right\} < 0$ and thus $\mathscr{L}(\{\widehat{\beta_k}, \widehat{\mu_k}, \widehat{\gamma_k}, \widehat{\delta_k}\})$ maxima in respect to the variables of these partial derivatives. Otherwise the second partial derivatives are all ≤ 0.

Estimating the Rate of SHAPE Adduction Formation

Most experimental designs when using RT-stops aim for a single SHAPE adduct formation per RNA present[79]. However, these experimental considerations do not guarantee single hit kinetics so it is useful to estimate the rate of SHAPE adduct formation from the sequencing reads. This estimation then informs the effectiveness of the probing step and its effects on picking up signal. Following[66] if we assume the probability of SHAPE adduct formation follows a Poisson distribution with rate c, then:

$$P(n\ modifications) = \frac{c^n e^{-c}}{n!}$$

The probability of no modification would then be the following, considering our model both considers RT-stops and RT-mutations:

$$P(n=0) = e^{-c} = \prod_{k=1}^{n} (1-\beta_k)(1-\mu_k)$$

Therefore:

$$\hat{c} = -\log\left(\prod_{k=1}^{n}(1-\widehat{\beta}_k)(1-\widehat{\mu}_k)\right) = -\sum_{k=1}^{n}\left[\log(1-\widehat{\beta}_k)+\log(1-\widehat{\mu}_k)\right]$$

$$= -\sum_{k=1}^{n}\left[\log\left(1-\frac{\frac{S_k^{(+)}}{\sum_{i=k}^{n+1}S_i^{(+)}}-\frac{S_k^{(-)}}{\sum_{i=k}^{n+1}S_i^{(-)}}}{1-\frac{S_k^{(-)}}{\sum_{i=k}^{n+1}S_i^{(-)}}}\right)+\log\left(1-\frac{\frac{mut_k^{(+)}}{\sum_{i=k}^{n+1}S_i^{(+)}}-\frac{mut_k^{(-)}}{\sum_{i=k}^{n+1}S_i^{(-)}}}{1-\frac{mut_k^{(-)}}{\sum_{i=k}^{n+1}S_i^{(-)}}}\right)\right]$$

$$= -\sum_{k=1}^{n}\left[\log\left(\frac{1-\frac{S_k^{(+)}}{\sum_{i=k}^{n+1}S_i^{(+)}}}{1-\frac{S_k^{(-)}}{\sum_{i=k}^{n+1}S_i^{(-)}}}\right)+\log\left(\frac{1-\frac{mut_k^{(+)}}{\sum_{i=k}^{n+1}S_i^{(+)}}}{1-\frac{mut_k^{(-)}}{\sum_{i=k}^{n+1}S_i^{(-)}}}\right)\right]$$

$$= -\sum_{k=1}^{n}\left[\log\left(\frac{1-\frac{stop_k^{(+)}}{depth_k^{(+)}}}{1-\frac{stop_k^{(-)}}{depth_k^{(-)}}}\right)+\log\left(\frac{1-\frac{mut_k^{(+)}}{depth_k^{(+)}}}{1-\frac{mut_k^{(-)}}{depth_k^{(-)}}}\right)\right]$$

1 Acknowledgements

We would like to thank Aaron Coraor for informative discussions about the chemical kinetic view of reactivities, as well as Adam Silverman and Eric Strobel for similar discussions and comments on the detailed derivation. We also thank Chaitan Khosla for inspiring the connections between the chemical and statistical perspectives of reactivities.

References

1. Adilakshmi, T., Lease, R. A. & Woodson, S. A. Hydroxyl radical footprinting in vivo: mapping macromolecular structures with synchrotron radiation. *Nucleic Acids Res* **34**, e64 (2006). DOI 10.1093/nar/gkl291.

2. Climie, S. C. & Friesen, J. D. In vivo and in vitro structural analysis of the rplj mrna leader of escherichia coli. protection by bound l10-l7/l12. *J Biol Chem* **263**, 15166–75 (1988).

3. Ding, Y. *et al.* In vivo genome-wide profiling of rna secondary structure reveals novel regulatory features. *Nat.* **505**, 696–700 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24270811. DOI 10.1038/nature12756.

4. Ehresmann, C. *et al.* Probing the structure of rnas in solution. *Nucleic Acids Res* **15**, 9109–28 (1987). URL https://www.ncbi.nlm.nih.gov/pubmed/2446263.

5. Favorova, O. O., Fasiolo, F., Keith, G., Vassilenko, S. K. & Ebel, J. P. Partial digestion of trna–aminoacyl-trna synthetase complexes with cobra venom ribonuclease. *Biochem.* **20**, 1006–11 (1981). URL https://www.ncbi.nlm.nih.gov/pubmed/7011369.

6. Feng, C. *et al.* Light-activated chemical probing of nucleobase solvent accessibility inside cells. *Nat Chem Biol* **14**, 276–283 (2018). DOI 10.1038/nchembio.2548.

7. Holley, R. W. *et al.* Structure of a ribonucleic acid. *Sci.* **147**, 1462–1465 (1965).

8. Inoue, T. & Cech, T. R. Secondary structure of the circular form of the tetrahymena rrna intervening sequence: a technique for rna structure analysis using chemical probes and reverse transcriptase. *Proc Natl Acad Sci U S A* **82**, 648–52 (1985). URL https://www.ncbi.nlm.nih.gov/pubmed/2579378.

9. Incarnato, D., Neri, F., Anselmi, F. & Oliviero, S. Genome-wide profiling of mouse rna secondary structures reveals key features of the mammalian transcriptome. *Genome Biol* **15**, 491 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/25323333. DOI 10.1186/s13059-014-0491-2.

10. Kertesz, M. *et al.* Genome-wide measurement of rna secondary structure in yeast. *Nat.* **467**, 103–7 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20811459. DOI 10.1038/nature09322.

11. Kladwang, W., Cordero, P. & Das, R. A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model rna. *RNA* **17**, 522–34 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21239468. DOI 10.1261/rna.2516311.

12. Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. A two-dimensional mutate-and-map strategy for non-coding rna structure. *Nat Chem* **3**, 954–62 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/22109276. DOI 10.1038/nchem.1176.

13. Knapp, G. Enzymatic approaches to probing of rna secondary and tertiary structure. *Methods Enzym.* **180**, 192–212 (1989). URL https://www.ncbi.nlm.nih.gov/pubmed/2482414.

14. Kwok, C. K., Ding, Y., Tang, Y., Assmann, S. M. & Bevilacqua, P. C. Determination of in vivo rna structure in low-abundance transcripts. *Nat Commun* **4**, 2971 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/24336128. DOI 10.1038/ncomms3971.

15. Lee, B. *et al.* Comparison of shape reagents for mapping rna structures inside living cells. *RNA* **23**, 169–174 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/27879433. DOI 10.1261/rna.058784.116.

16. Li, F. *et al.* Global analysis of rna secondary structure in two metazoans. *Cell Rep* **1**, 69–82 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22832108. DOI 10.1016/j.celrep.2011.10.002.

17. Lockard, R. E. & Kumar, A. Mapping trna structure in solution using double-strand-specific ribonuclease v1 from cobra venom. *Nucleic Acids Res* **9**, 5125–40 (1981). URL https://www.ncbi.nlm.nih.gov/pubmed/7031604.

18. Lucks, J. B. *et al.* Multiplexed rna structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (shape-seq). *Proc Natl Acad Sci U S A* **108**, 11063–8 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21642531. DOI 10.1073/pnas.1106501108.

19. Loughrey, D., Watters, K. E., Settle, A. H. & Lucks, J. B. Shape-seq 2.0: systematic optimization and extension of high-throughput chemical probing of rna secondary structure with next generation sequencing. *Nucleic Acids Res* **42** (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/25303992. DOI 10.1093/nar/gku909.

20. McGinnis, J. L., Duncan, C. D. & Weeks, K. M. High-throughput shape and hydroxyl radical analysis of rna structure and ribonucleoprotein assembly. *Methods Enzym.* **468**, 67–89 (2009). URL https://www.ncbi.nlm.nih.gov/pubmed/20946765. DOI 10.1016/S0076-6879(09)68004-6.

21. McGinnis, J. L. *et al.* In-cell shape reveals that free 30s ribosome subunits are in the inactive state. *Proc Natl Acad Sci U S A* **112**, 2425–30 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25675474. DOI 10.1073/pnas.1411514112.

22. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. Rna structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (shape). *J Am Chem Soc* **127**, 4223–31 (2005). URL https://www.ncbi.nlm.nih.gov/pubmed/15783204. DOI 10.1021/ja043822v.

23. Mitchell, r., D. *et al.* Glyoxals as in vivo rna structural probes of guanine base pairing. *RNA* **24**, 114–124 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/29030489. DOI 10.1261/rna.064014.117.

24. Moazed, D., Stern, S. & Noller, H. F. Rapid chemical probing of conformation in 16 s ribosomal rna and 30 s ribosomal subunits using primer extension. *J Mol Biol* **187**, 399–416 (1986). URL https://www.ncbi.nlm.nih.gov/pubmed/2422386.

25. Mortimer, S. A. & Weeks, K. M. A fast-acting reagent for accurate analysis of rna secondary and tertiary structure by shape chemistry. *J Am Chem Soc* **129**, 4144–5 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17367143. DOI 10.1021/ja0704028.

26. Noller, H. F. & Chaires, J. B. Functional modification of 16s ribosomal rna by kethoxal. *Proc Natl Acad Sci U S A* **69**, 3115–8 (1972). URL https://www.ncbi.nlm.nih.gov/pubmed/4564202.

27. Qu, H. L., Michot, B. & Bachellerie, J. P. Improved methods for structure probing in large rnas: a rapid 'heterologous' sequencing approach is coupled to the direct mapping of nuclease accessible sites. application to the 5' terminal domain of eukaryotic 28s rrna. *Nucleic Acids Res* **11**, 5903–20 (1983). URL https://www.ncbi.nlm.nih.gov/pubmed/6193488.

28. Rice, G. M., Leonard, C. W. & Weeks, K. M. Rna secondary structure modeling at consistent high accuracy using differential shape. *RNA* **20**, 846–54 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24742934. DOI 10.1261/rna.043323.113.

29. Ritchey, L. E. *et al.* Structure-seq2: sensitive and accurate genome-wide profiling of rna structure in vivo. *Nucleic Acids Res* **45**, e135 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28637286. DOI 10.1093/nar/gkx533.

30. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of rna structure reveals active unfolding of mrna structures in vivo. *Nat.* **505**, 701–5 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24336214. DOI 10.1038/nature12894.

31. Seetin, M. G., Kladwang, W., Bida, J. P. & Das, R. Massively parallel rna chemical mapping with a reduced bias map-seq protocol. *Methods Mol Biol* **1086**, 95–117.

32. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. & Weeks, K. M. Rna motif discovery by shape and mutational profiling (shape-map). *Nat Methods* **11**, 959–65 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/25028896. DOI 10.1038/nmeth.3029.

33. Spitale, R. C. *et al.* Rna shape analysis in living cells. *Nat Chem Biol* **9**, 18–20 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23178934. DOI 10.1038/nchembio.1131.

34. Spitale, R. C. *et al.* Structural imprints in vivo decode rna regulatory mechanisms. *Nat.* **519**, 486–90 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25799993. DOI 10.1038/nature14263.

35. Strobel, E. J., Watters, K. E., Nedialkov, Y., Artsimovitch, I. & Lucks, J. B. Distributed biotin-streptavidin transcription roadblocks for mapping cotranscriptional rna folding. *Nucleic Acids Res* **45**, e109 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28398514. DOI 10.1093/nar/gkx233.

36. Talkish, J., May, G., Lin, Y., Woolford, J., J. L. & McManus, C. J. Mod-seq: high-throughput sequencing for chemical probing of rna structure. *RNA* **20**, 713–20 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24664469. DOI 10.1261/rna.042218.113.

37. Tang, Y. *et al.* Structurefold: genome-wide rna secondary structure mapping and reconstruction in vivo. *Bioinforma.* **31**, 2668–75 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25886980. DOI 10.1093/bioinformatics/btv213.

38. Tian, S., Cordero, P., Kladwang, W. & Das, R. High-throughput mutate-map-rescue evaluates shape-directed rna structure and uncovers excited states. *RNA* **20**, 1815–26 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/25183835. DOI 10.1261/rna.044321.114.

39. Tijerina, P., Mohr, S. & Russell, R. Dms footprinting of structured rnas and rna-protein complexes. *Nat Protoc* **2**, 2608–23 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17948004. DOI 10.1038/nprot.2007.380.

40. Underwood, J. G. *et al.* Fragseq: transcriptome-wide rna structure probing using high-throughput sequencing. *Nat Methods* **7**, 995–1001 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/21057495. DOI 10.1038/nmeth.1529.

41. Van Stolk, B. J. & Noller, H. F. Chemical probing of conformation in large rna molecules. analysis of 16 s ribosomal rna using diethylpyrocarbonate. *J Mol Biol* **180**, 151–77 (1984). URL https://www.ncbi.nlm.nih.gov/pubmed/6210372.

42. Vary, C. P. & Vournakis, J. N. Rna structure analysis using methidiumpropyl-edta.fe(ii): a base-pair-specific rna structure probe. *Proc Natl Acad Sci U S A* **81**, 6978–82 (1984). URL https://www.ncbi.nlm.nih.gov/pubmed/6209709.

43. Watters, K. E., Abbott, T. R. & Lucks, J. B. Simultaneous characterization of cellular rna structure and function with in-cell shape-seq. *Nucleic Acids Res* **44**, e12 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26350218. DOI 10.1093/nar/gkv879.

44. Watters, K. E., Strobel, E. J., Yu, A. M., Lis, J. T. & Lucks, J. B. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nat Struct Mol Biol* **23**, 1124–1131 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27798597. DOI 10.1038/nsmb.3316.

45. Zheng, Q. *et al.* Genome-wide double-stranded rna sequencing reveals the functional significance of base-paired rnas in arabidopsis. *PLoS Genet.* **6**, e1001141 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20941385. DOI 10.1371/journal.pgen.1001141.

46. Zubradt, M. *et al.* Dms-mapseq for genome-wide or targeted rna structure probing in vivo. *Nat Methods* **14**, 75–82 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/27819661. DOI 10.1038/nmeth.4057.

47. Mlynsky, V. & Bussi, G. Molecular dynamics simulations reveal an interplay between shape reagent binding and rna flexibility. *The J. Phys. Chem. Lett.* **9**, 313–318 (2018). URL https://doi.org/10.1021/acs.jpclett.7b02921. DOI 10.1021/acs.jpclett.7b02921. PMID: 29265824.

48. Homan, P. J. *et al.* Single-molecule correlated chemical probing of rna. *PNAS* **111**, 13858–13863 (2014).

49. Cheng, C. Y., Kladwang, W., Yesselman, J. D. & Das, R. Rna structure inference through chemical mapping after accidental or intentional mutations. *Proc Natl Acad Sci U S A* **114**, 9876–9881 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28851837. DOI 10.1073/pnas.1619897114.

50. Brunel, C. & Romby, P. Probing rna structure and rna-ligand complexes with chemical probes. *Methods Enzymol.* **318**, 3–21 (2000).

51. Watters, K. E., Yu, A. M., Strobel, E. J., Settle, A. H. & Lucks, J. B. Characterizing rna structures in vitro and in vivo with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (shape-seq). *Methods* **103**, 34–48 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27064082. DOI 10.1016/j.ymeth.2016.04.002.

52. Sexton, A. N., Wang, P. Y., Rutenberg-Shoenberg, M. & Simon, M. D. Interpreting reverse transcriptase termination and mutation events for greater insight into the chemical probing of rna. *Biochem.* **56**, 4713–4721 (2017).

53. Smola, M. J., Calabrese, J. M. & Weeks, K. M. Detection of rna-protein interactions in living cells with shape. *Biochem.* **54**, 6867–75 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26544910. DOI 10.1021/acs.biochem.5b00977.

54. Smola, M. J. *et al.* Shape reveals transcript-wide interactions, complex structural domains, and protein interactions across the xist lncrna in living cells. *Proc Natl Acad Sci U S A* **113**, 10322–7 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27578869. DOI 10.1073/pnas.1600008113.

55. Kladwang, W. *et al.* Standardization of rna chemical mapping experiments. *Biochem.* **53**, 3063–5 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24766159. DOI 10.1021/bi5003426.

56. Cheng, C. Y. *et al.* Consistent global structures of complex rna states through multidimensional chemical mapping. *Elife* **4**, e07600 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26035425. DOI 10.7554/eLife.07600.

57. Fang, R., Moss, W. N., Rutenberg-Schoenberg, M. & Simon, M. D. Probing xist rna structure in cells using targeted structure-seq. *PLoS Genet.* **11**, e1005668 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26646615. DOI 10.1371/journal.pgen.1005668.

58. Vicens, Q., Gooding, A. R., Laederach, A. & Cech, T. R. Local rna structural changes induced by crystallization are revealed by shape. *RNA* **13**, 536–48 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17299128. DOI 10.1261/rna.400207.

59. Karabiber, F., McGinnis, J. L., Favorov, O. V. & Weeks, K. M. Qushape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* **19**, 63–73 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23188808. DOI 10.1261/rna.036327.112.

60. McGinnis, J. L., Dunkle, J. A., Cate, J. H. & Weeks, K. M. The mechanisms of rna shape chemistry. *J Am Chem Soc* **134**, 6617–24 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22475022. DOI 10.1021/ja2104075.

61. Chamberlin, S. I. & Weeks, K. M. Mapping local nucleotide flexibility by selective acylation of 2 '-amine substituted rna. *J. Am. Chem. Soc.* **122**, 216–224 (2000). URL <GotoISI>://WOS:000084999000004. DOI DOI 10.1021/ja9914137.

62. Soukup, G. A. & Breaker, R. R. Relationship between internucleotide linkage geometry and the stability of rna. *RNA* **5**, 1308–25 (1999). URL https://www.ncbi.nlm.nih.gov/pubmed/10573122.

63. Lavery, R. & Pullman, A. A new theoretical index of biochemical reactivity combining steric and electrostatic factors. an application to yeast trnaphe. *Biophys Chem* **19**, 171–81 (1984). URL https://www.ncbi.nlm.nih.gov/pubmed/6372881.

64. Peattie, D. A. & Gilbert, W. Chemical probes for higher-order structure in rna. *Proc Natl Acad Sci U S A* **77**, 4679–82 (1980). URL https://www.ncbi.nlm.nih.gov/pubmed/6159633.

65. Metzker, M. L. Sequencing technologies - the next generation. *Nat Rev Genet.* **11**, 31–46 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/19997069. DOI 10.1038/nrg2626.

66. Aviran, S., Lucks, J. B. & Pachter, L. Rna structure characterization from chemical mapping experiments. *Proc. 49th Annu. Allerton Conf. on Commun. Control. Comput.* 1743–1750 (2011).

67. Aviran, S. *et al.* Modeling and automation of sequencing-based characterization of rna structure. *Proc Natl Acad Sci U S A* **108**, 11069–74 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21642536. DOI 10.1073/pnas.1106541108.

68. Busan, S. & Weeks, K. M. Accurate detection of chemical modifications in rna by mutational profiling (map) with shapemapper 2. *RNA* (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/29114018. DOI 10.1261/rna.061945.117.

69. Gherghe, C. M., Mortimer, S. A., Krahn, J. M., Thompson, N. L. & Weeks, K. M. Slow conformational dynamics at c2'-endo nucleotides in rna. *JACS* **130**, 8884–8885 (2008).

70. Bindewald, E. *et al.* Correlating shape signatures with three-dimensional rna structures. *RNA* **17**, 1688–96 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21752927. DOI 10.1261/rna.2640111.

71. Novoa, E. M., Beaudoin, J.-D., Giraldez, A. J., Mattick, J. S. & Kellis, M. Best practices for genome-wide rna structure analysis: combination of mutational profiles and drop-off information. *bioRxiv* (2017). URL https://www.biorxiv.org/content/early/2017/08/21/176883. DOI 10.1101/176883.

72. Novoa, E. M., Beaudoin, J.-D., Giraldez, A., Mattick, J. S. & Kellis, M. Best practices for genome-wide rna structure analysis: combination of mutational profiles and drop-off information. *BioRxiv* 1–24 (2017). DOI 10.1101/176883.

73. Krokhotin, A., Mustoe, A., Weeks, K. M. & Dokholyan, N. V. Direct identification of base-paired rna nucleotides by correlated chemical probing. *RNA* **23**, 6–13 (2016).

74. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate shape-directed rna structure determination. *Proc Natl Acad Sci U S A* **106**, 97–102 (2009). URL https://www.ncbi.nlm.nih.gov/pubmed/19109441. DOI 10.1073/pnas.0806929106.

75. Cordero, P., Kladwang, W., VanLang, C. C. & Das, R. Quantitative dimethyl sulfate mapping for automated rna secondary structure inference. *Biochem.* **51**, 7037–9 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22913637. DOI 10.1021/bi3008802.

76. Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. Understanding the errors of shape-directed rna structure modeling. *Biochem.* **50**, 8049–56 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21842868. DOI 10.1021/bi200524n.

77. Quarrier, S., Martin, J. S., Davis-Neulander, L., Beauregard, A. & Laederach, A. Evaluation of the information content of rna structure mapping data for secondary structure prediction. *RNA* **16**, 1108–17 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20413617. DOI 10.1261/rna.1988510.

78. Sukosd, Z., Swenson, M. S., Kjems, J. & Heitsch, C. E. Evaluating the accuracy of shape-directed rna secondary structure predictions. *Nucleic Acids Res* **41**, 2807–16 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23325843. DOI 10.1093/nar/gks1283.

79. Aviran, S. & Pachter, L. Rational experiment design for sequencing-based rna structure mapping. *RNA* **20**, 1864–77 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/25332375. DOI 10.1261/rna.043844.113.