Modeling of overdispersion in RNA chemical probing data and application to RNA secondary structure prediction

Introduction

The base-pairing interactions of RNA molecules are essential to life, and particularly to the functions of noncoding RNAs, such as ribosomal RNAs and riboswitches. RNA secondary structure can also be important to the function and regulation of long noncoding RNAs and messenger RNAs, and determining these structures is an important step toward understanding the functional mechanisms of RNA. Chemical probing experiments can greatly improve the accuracy of RNA secondary structure prediction by providing information about which nucleotides are single- or double-stranded. In these experiments, RNA is treated with chemicals, e.g. dimethyl sulfate and selective 2' hydroxyl acylating (SHAPE) reagents, that selectively modify RNA bases based on their structural context, usually with a bias toward single-stranded nucleotides. Modified nucleotides are then read out by reverse transcriptase (RT), which stops cDNA synthesis or inserts incorrect bases at chemical adducts (we refer to RT stops and mutations more generally as RT events). Comparing results of probing experiments to controls with no chemical treatment enables calculation of nucleotide reactivities, which are then converted into probabilistic constraints for RNA secondary structure prediction \cite{19109441, reviewed in Choudhary2017QuantitativeBiol}.

An underexplored area in the analysis of chemical probing data is how best to take advantage of replicate data. Most current methods to analyze chemical probing data address the results of a single replicate\cite{19109441, 27819661, 24336214, etc}, or the pre-pooled results of multiple replicates\cite{Prober}. Ideally, if the conditions of experiments were exactly the same, such that every RNA molecule had the same probability of generating an RT event at a given nucleotide, then this would be well justified. Indeed, the Poisson distribution, which is often used to model chemical probing data\cite{ 21642536, 25332375, 26544910}, assumes that the underlying statistical process being modeled has a fixed probability. However, biological data of many types – ranging from gene expression (RNA-Seq)\cite{17728317, 19910308 } to mutation rates in cancer genomes \cite{ 26304545 } – are often overdispersed relative to Poisson statistics, due to heterogeneity in biological conditions. If chemical probing data are also overdispersed relative to Poisson statistics, statistical modeling with a replicate sensitive tool could be critical to assessing confidence in inferred nucleotide reactivities, and estimating confidence might aid downstream structure prediction. One recent method, BUM-HMM (Beta-Uniform Mixture Hidden Markov Model) used the empirical variability in control chemical

probing experiments to assess the significance of differences between treated and control experiments\cite{ 27819660}. However, to our knowledge, no existing method uses variability in both treated and control experiments to model chemical probing data.

Here, we present a new analysis tool, *structSeq*, that integrates information from replicate experiments to model chemical probing data statistically and to infer base pairing probabilities to be used for RNA secondary structure prediction. In *structSeq*, we first model chemical probing data using the negative-binomial distribution, which allows for heterogeneity in the probabilities that different RNAs within a population will produce an RT event at a given nucleotide. We use replicate data to infer distribution parameters by adapting a statistical tool from the RNA-Seq field (DESeq2) that is specially designed to infer levels of variability in experiments where many measurements are made in each replicate, but only a small number of replicate experiments are conducted. We employ this statistical model to identify nucleotides that are significantly modified in probing experiments. Finally, we create a framework to incorporate both observed nucleotide reactivities and inferred levels of variability into predictions of RNA secondary structure.

Results

RNA chemical probing data are overdispersed

[[NOTE: I should introduce to some extent how we define RT event counts, trials/coverage, pseudocounts for use in Poisson-family distributions, etc. These concepts are/will be described completely in the methods, but I have not figured out how much to address them in the results.]]

To motivate our development of a replicate-sensitive method to analyze chemical probing data, we first wanted to investigate the underlying statistical assumptions that dictate whether separation of replicate data is necessary. Specifically, if the underlying process being modeled – generation of RT stops or mutations (RT events) from a given nucleotide within an RNA molecule that undergoes chemical probing – is the same both throughout each population of RNA and between RNA populations of the same type (treated or untreated), then we would be well justified in adding results of replicates together. If this is the case, we would expect the counts of RT events we observe to follow the Poisson distribution, which models counts resulting from a statistical process with a fixed, small probability (of producing an RT event at the nucleotide of interest) and many total trials (many total RNA molecules). To test whether our data follow Poisson statistics, we used one replicate to define the \lambda parameter of the distribution, representing the mean expected counts, and computed p-values for observations from other replicates. If the Poisson distribution is a good model for chemical probing data, then the p-values should follow the uniform distribution.

Michael Rutenberg S…, 4/6/2017 7:08 AM

Comment [1]: This is redundant with the second paragraph in the intro. I would like to make this version more succinct and/or somewhat distinct in the points it makes.

However, a quantile-quantile plot comparing observed Poisson p-values to their expected uniform distribution shows that RT stop counts in Xist DMS treated samples are highly overdispersed (Fig 1a, Kolmogorov Smirnov test $p < 2.2*10^{-16}$) We visualize this finding by showing the Poisson confidence intervals estimated from a sample region of one Xist DMS replicate and the fact that many points from a second replicate lie far outside these intervals (Fig 1b). This finding is also intuitively borne out by the fact that simulated replicates using Poisson statistics (Fig 1c) display much less variability than experimental replicates (Fig 1d). Moreover, our finding that chemical probing data are overdispersed holds across many types of experiments, including counts of both RT stop and mutation counts, both *in vitro* and *in vivo*, and even untreated samples (supplemental figures).

structSeq **provides an improved model of chemical probing data using the negative binomial distribution and DESeq2**

Having established that chemical probing data are overdispersed relative to the Poisson distribution, we sought to model these data with a distribution with a flexible mean-variance relationship. This posed a challenge as like other genomics experiments, relatively few replicates are typically conducted because of cost constraints, making it hard to make accurate variance estimates for each data point (nucleotide) individually. This problem has been addressed in the RNA-Seq field by tools such as DESeq2, which take advantage of common information among the many measurements (of gene expression) made in parallel to aid inference of count distributions \cite{ 25516281 }. DESeq2 employs the negative binomial distribution, which is closely relative to the Poisson distribution but contains a dispersion parameter, \alpha, which is zero when there is no overdispersion (Poisson) but takes higher values when data are overdispersed (see Methods). DESeq2 estimates the dispersion parameter by first making estimates for each nucleotide (or gene in RNA-Seq), then observing a trend between mean counts and dispersion values, and finally adjusting dispersion values toward the trend (Fig 1e).Applying DESeq2 to RT stop counts for DMS-treated Xist yielded a substantially better fit to chemical probing data, as shown in our quantile-quantile plot of DMS-treated Xist stop counts (Fig 1a, Kolmogorov Smirnov test $p = 0.23$), as well as by the fact that negative-binomial simulated replicates (Fig 1f) are more similar to observed replicates (Fig 1d) than are Poisson replicates (Fig 1c).

[[NOTE: The negative binomial provides an excellent fit for the Xist data set (especially with moderate filtering to assure higher coverage). This isn't true for all datasets, and I'm working to figure out why/to what extent this affects downstream results.]]

Using p-values from structSeq for biochemical inference

A more accurate statistical model of chemical probing data should aid biochemical inference from these datasets. Both SHAPE and DMS are expected to modify single-stranded nucleotides preferentially over double-stranded bases, while DMS is also selective for A and C bases. We performed statistical tests comparing treated and control RT event counts to identify bases that have significantly more RT events due to treatment. We then compared the distinctive ability of p-values from our negative binomial model to that of Poisson p-values, as well as the empirical p-values produced by recently published BUM-HMM method that uses empirical differences in untreated samples as the null distribution against which to test treatment-control comparisons. Precision-recall curves, which measure precision as increasing numbers of positive identifications are made, show that negative binomial p-values have greater distinctive ability than either of the other two statistical tests for distinction of AC vs. GU nucleotides for *in* vivo, targeted DMS probing of 3 RNAs: Xist, U2 snRNA, and 7SK RNA (Fig 2a-c) The negative binomial p-values also perform comparably to the other tests for distinction of single stranded vs. double stranded nucleotides from *in vitro* SHAPE probing of the 5S rRNA. We report performance for an expanded list of RNAs, using the area under both precision-recall and receiver operator curves as metrics, in Table 1. We note that we would not expect perfect performance from a statistical metric in any of these comparisons, because the in no case is the chemical probing reagent perfectly selective for the type of base.

Michael Rutenberg …, 5/17/2017 3:05 PM
Comment [2]: I'm not sure why the other methods perform well for the in vitro dataset, but so poorly on some of the in vivo datasets. This could be due to degree of overdispersion, but that's not what I see so far.

[[Note: In figure 2A-D, the BUM-HMM p-values and the Poisson exact test perform reasonably in some cases and terribly in others. It would be good to figure out why this is (and whether it has to do with conceptual deficiencies in these methods and/or technical errors on my part.]]

***structSeq* enables replicate-aware RNA structure prediction**

[[**Note**: I had not realized until recently that the negative-binomial p-values are not obviously better than the other statistical tests for the one *in vitro* dataset that I have analyzed so far. The folding section should focus on *in vivo* data if our statistical model is not particularly important for *in vitro* data.]]

Statistical tests for differences between treated and control samples in chemical probing experiments are useful in themselves, but RNA structure prediction methods typically rely on levels of nucleotide reactivity, based on the observation that the degree of reactivity is more structurally meaningful than simply whether the reactivity is above zero \cite{Deigan2009PNAS}. Our negative binomial fits enable expansion of this concept to include confidence in estimates of reactivity. To illustrate this, we made violin plots, showing the inferred

distribution of a common measurement of reactivity for several RNAs –the increase in probability of stopping or mutation due to treatment (denoted ΔP_{stop}, ΔP_{mut}, or more generally, ΔP; see Methods)– by simulating from our treated and control count distributions \cite{ 26646615, primary citation for ΔP} (Fig 2e-h). RNA structure prediction methods typically perform one of a variety of normalization methods on ΔP or similar metrics of reactivity, and then define functions that convert these normalized reactivities into probabilities that each nucleotide is paired. These probabilities are converted to pseudoenergy terms that can be used to constrain free energy-based RNA secondary structure prediction algorithms \cite{ 19109441, 24895857].

To create a replicate-aware method to incorporate chemical probing data into RNA secondary structure prediction, we represent counts for treated and control experiments by the negative binomial distributions we fit using DESeq2. Similar to the violin plots for ΔP above, we simulated many replicates of our experimental data. We then calculated ΔP, used the so-called boxplot method (see Methods, \cite{ 19109441 }) to create normalized reactivities. These reactivities were converted to probabilities using the method proposed by Deigan et al.:

$$\Delta G = a*\log(R+1) + b$$
$$p(\text{paired})/p(\text{single-stranded}) = \exp(-\Delta G/R_{gas}T)$$
$$p(\text{single-stranded}) = 1/(1+\exp(-\Delta G/R_{gas}T))$$

Here, R represents normalized reactivity, R_{gas} is the gas constant, T is temperature, which is taken to be 37C (310.15 K). The final probability estimate was taken as the mean of the probabilities computed from sampling from treated and control count distributions. Of note is that there are two free parameters in the Deigan method that control the degree to which high reactivity indicates low pairing probability (a, above) and that no reactivity indicates higher pairing probability (b, above) (Figure 3b). The best values for a and b are typically determined based on optimization of prediction performance for RNAs of known structure.

We implemented our replicate-sensitive method for converting raw chemical probing counts to structure prediction constraints using *in vitro* SHAPE data for the 5S ribosomal RNA, an RNA of known structure \cite{ 25303992, 22976082, maybe others}. Pairing probabilities from our method were input into the RNAstructure software package for structure prediction. To evaluate our predictions, we use two metrics: sensitivity, the proportion of correct base pairs that are predicted; and positive predictive value (PPV), the proportion of predicted base pairs that are correct. Our probabilistic method enables correct prediction of the 5S rRNA structure (sensitivity = 100%, PPV = 100%, Fig 3a), in contrast to an inaccurate prediction from an unconstrained structure (sensitivity = 27.0%, PPV = 24.3%, Fig3b). We can also see that prediction accuracy is robust to parameters of the Deigan pseudoenergy function that control the degree to

Michael Rutenberg S…, 4/6/2017 7:12 AM
Comment [3]: These were previously mentioned earlier, and I think they probably should be returned to an earlier place in the results.

Michael Rutenberg …, 5/17/2017 3:06 PM
Comment [4]: Using a single replicate to predict the structure using standard methods is also accurate.

which high reactivity indicates low pairing probability (a) and that no reactivity indicates higher pairing probability (b) (Figure 3c). This demonstrates the ability of our replicate-sensitive method to aid accurate prediction of RNA structure on a model RNA.

[[**Note**: I do not yet mention the fact that folding with constraints but no statistical model is just as effective as folding with constraints and the statistical model.]]

Figure legends:

Figure 1. Overdispersion of chemical probing data and modeling using the negative-binomial distribtution and DESeq2. (Panels A-F go from left to right, then top to bottom)

A. Quantile-quantile plot for stop counts from mouse Xist treated with DMS *in vivo*. P-values of observed data against assumed distributions are compared to the uniform distribution on log10 scale. From six replicates, one was chosen to fit the Poisson model and the other five were tested against this replicate. For the negative-binomial model, five replicates were used to fit the distribution using DESeq2 and one replicate was used to test the distribution fit.
B. Comparison of two biological replicates of Xist DMS stop counts.
C. Scatterplot of Poisson simulated replicates for Xist DMS stop counts.
D. Observed biological replicates for Xist DMS stop counts.
E. Relationship between mean counts and the negative binomial dispersion parameter (which controls overdispersion) while fitting Xist DMS stop counts using DESeq2. Dispersion parameters are initially fit separately to each nucleotide (black dots), then a trend is fit to these dispersion estimates (red line), and final estimates (blue dots) are made combining information from the individual estimates and the trend.
F. Negative-binomial simulated replicates for Xist DMS stop counts. Distribution was fit using DESeq2.

Michael Rutenberg S…, 4/6/2017 8:41 PM
Comment [5]: I'm still figuring out if this is the most correct way to measure Poisson overdispersion.

Figure 2. Evaluation of biochemical inferences made with negative binomial p-values (from DESeq2) and comparison to other methods
(Panels A-H go from top to bottom, then left to right)

A. Precision recall plot for A and C bases for *in vivo* targeted DMS probing of the mouse 7SK RNA. Negative binomial p-values fit with DESeq2 are compared to those from BUM-HMM and the Poisson exact test.

B. Precision recall plot for A and C bases for *in vivo* targeted DMS probing of the mouse U2 snRNA.
C. Precision recall plot for A and C bases for *in vivo* targeted DMS probing of the mouse Xist RNA.
D. Precision recall plot for A and C bases for *in vitro* SHAPE probing of the E coli 5S rRNA.
E. Violin plots showing distributions of ΔP_{stop}, a measure of nucleotide reactivity, inferred from 10,000 simulated replicates of treated and control counts for targeted *in vivo* DMS probing mouse of the mouse 7SK RNA.
F. Violin plots showing distributions of ΔP_{stop}, a measure of nucleotide reactivity, inferred from 10,000 simulated replicates of treated and control counts for targeted *in vivo* DMS probing mouse of the mouse U2 snRNA.

(Not currently mentioned in text)

G. Violin plots showing distributions of ΔP_{stop}, a measure of nucleotide reactivity, inferred from 10,000 simulated replicates of treated and control counts for targeted *in vivo* DMS probing mouse of the mouse Xist RNA.
H. Violin plots showing distributions of ΔP_{stop}, a measure of nucleotide reactivity, inferred from 10,000 simulated replicates of treated and control counts for *in vitro* SHAPE probing of the E coli 5S rRNA.

Figure 3. Structure prediction with 5S rRNA

A. 5S rRNA structure predicted using *in vitro* SHAPE constraints. Nucleotides are colored by the probabilities of pairing produced with probing data and our negative binomial models. This predicted structure is correct (100% sensitivity, 100% positive predictive value).
B. 5S rRNA structure predicted without constraints. Nucleotides are colored by the probabilities of pairing produced with probing data and our negative binomial models. This predicted structure has sensitivity = 27.05%, positive predictive value = 24.3%,
C. Sensitivity of predicted 5S rRNA structures with SHAPE constraints, using different parameters of the Deigan pseudoenergy function.
D. Positive predictive value of predicted 5S rRNA structures with SHAPE constraints, using different parameters of the Deigan pseudoenergy function.

Non-exhaustive list of goals:

- Figure out why relative performance of statistical models in Fig 2 is different for different RNAs/experiment types.
- Fold more RNAs of known structure, evaluate ΔP only vs. sampling from count distributions.

    - Ideally train pseudoenergy function parameters on some RNAs, test on others.
    - Compare to BUM-HMM folding as well?
- Do some testing of HMM/whether it is beneficial for folding.
- Combine mutation and stop information, if possible.