Equations for chemical probing stats paper

1 Naive analysis of chemical probing data

In chemical probing experiments, we count so-called events that can occur during reverse transcription, specifically either reverse transcription stops (termination of cDNA synthesis), or mutations in the resulting cDNA.

Let Y_{ij} represent the event counts for nucleotide i in sample j, with N total nucleotides and M total samples.

For one sample:

$$P_{event} = \frac{Y_{ij}}{C_{ij}}$$

For mutations, the coverage is simply r_{ij}, the number of reads directly covering the nucleotide of interest:

$$C_{ij} = r_{ij}$$

For stops, the coverage is the number of reads that reads the nulceotide of interest (and therefore could have stopped at the position of interest):

$$C_{ij} = \sum_{k=1}^{N} Y_{kj}$$

To combine samples, we sum all reads counts to compute event probabilities.

$$P_{event} = \frac{\sum_{j=1}^{M} Y_{ij}}{\sum_{j=1}^{M} C_{ij}}$$

The difference in probability of either an event between treated and untreated samples is:

$$\Delta P_{event} = P_{event-treated} - P_{event-untreated}$$

We rename P_{event}, depending on whether we are analyzing stop or mutation counts.

For stops:

$$\Delta P_{stop} = \Delta P_{event}$$

For mutations:

$$\Delta P_{mut} = \Delta P_{event}$$

2 Count normalization

To bring counts onto the same scale, we define pseudocounts K_{ij} that are scaled by coverage, such that every nucleotide as the same final effective coverage D_i.

Pseudocounts:

$$K_{ij} = \frac{Y_{ij}}{C_{ij}} \frac{\sum_{k=1}^{M} C_{ik}}{M}$$

Effective coverage:

$$D_i = \frac{\sum_{k=1}^{M} C_{ik}}{M}$$

3 Modeling counts using the negative-binomial distribution

To fit our observations of chemical probing data to negative binomial distributions, we start with pseudocounts, K, defined above, that are proportional to probabilities of RT stop/muation and normalized to have the same effective coverage. These pseudocounts are input into the DESeq2 pipeline, which is run with default parameters, except that count normalization is disabled, because we have already normalized counts using our experiment-specific approach. Briefly, DESeq2 estimates the parameters of the negative-binomial distribution for each nucleotide by first fitting parameters to each nucleotide by Cox-Reid adjusted profile likelihood.

The pipeline then fits the dispersion parameters α to a curve of the form $\alpha = a_0/\mu + a_1$ and re-estimates these assuming a log-normal prior.

[[Include more detail/possibly change methodology here]]

4 Chemical probing reactivity

The P_{event} values for chemical probing data are often normalized to a common scale before being used to provide constraints to secondary structure prediction. This both controls for differences in overall degree of modification, and provides a consistent value to relate to structural properties.

Here we define reactivity as:

$$R = \frac{\Delta P_{event}}{c}$$

Where c is a normalization factor, equal to the average of the top 10% of datapoints, after excluding any datapoints greater than the third quartile plus 1.5 times the interquartile range ("Boxplot method" - Deigan 2009 $PNAS$).

5 Structure prediction with soft constraints

The most common way to incorporate SHAPE data into RNA structure prediction algorithms is to convert SHAPE reactivities R for each nucleotide into pseudoenergies $E(R)$ that favor or disfavor base-pairing.

$$E(R) = a * \log(R + 1) + b$$

This method assumes a fixed reactivity. If the calculated reactivity is less than zero, it is set to zero, as there is no biochemically meaningful interpretation of negative reactivity. Since there is variability in the observed reactivity, we can instead compute pseudoenergies based on the inferred distribution of the reactivity.

$$K_t \sim NB(\hat{\mu_1}, \hat{\alpha})$$
$$K_{nt} \sim NB(\hat{\mu_2}, \hat{\alpha})$$
$$R_i(K_t, K_{nt}) = \frac{(K_t - K_{nt})}{cD_i}$$
$$\hat{E_i} = \int p(R_i)E(R_i)dR_i$$

We can easily estimate $\hat{E}$ by sampling L pseudocount values from the negative binomial distributions for the treated (K_t) and untreated (K_{nt}) that we fit to the data (see above), computing reactivity values $R(K_t, K_{nt})$, and then taking the mean of the pseudoenergies:

$$\hat{E}_i = \frac{\sum_{l=1}^{L} E(R_l(K_t, K_{nt}))}{N}$$

Finally, to avoid modification of the RNAstructure package, we compute a pseudoreactivity R^* that will cause the computed pseudoenergy $\hat{E}$ to be used for folding.

$$R^* = E^{-1}(\hat{E}_i) = \exp\left(\frac{\hat{E}_i - b}{a}\right) - 1$$