Metrics on RNA secondary structure ensembles

Michael Golden, Alain Laederach, Jotun Hein

March 3, 2015

RNA secondary structure is defined as the set of base-pairing interactions between the constituent bases of a RNA sequence. The function of a RNA is largely determined by its secondary structure. Consequently, if two or more RNAs share a similar structure it is likely that they share a similar function. There exist various metrics that allow one to measure similarity between two RNA secondary structures [4].

A typical assumption made in RNA structure studies is that a particular RNA sequence forms a single, well-defined structure, such RNAs are referred to as *stable* or *highly structured*. However, it is more accurate to think of RNA sequences as able to adopt a range of possible conformations with varying degrees of probability [2, 3, 1, 5]. RNAs that have a tendency to adopt a large number of distinct conformations are termed *flexible*. The untranslated regions (UTRs) of many mRNAs are generally considered to be so [2]. The probability distribution over all possible structural conformations a RNA sequence can form under a given set of conditions is referred to as its *ensemble*.

The goal of this project is define a set of metrics that allow one to quantify the spread of structural conformations within an ensemble (a measure of flexibility) and measure distances between structural ensembles (as opposed to just the distance between the most probable conformation in each of the two ensembles, as is currently the standard). As a final outcome these metrics are expected to be applied to real-world data, to test whether associations exist between the measured properties of the RNA ensembles and different functional annotations. For example, we expect RNA sequences that have structural ensembles with large variability (high structural flexibility) to be associated with certain types of annotations, e.g. annotated as UTRs more frequently than can be accounted for by chance.

Short-term objectives:

1. Let $\mathcal{E}_1$ and $\mathcal{E}_2$ denote structural ensembles corresponding to two sequences, $\mathcal{S}_1$ and $\mathcal{S}_2$, folded under a given set of conditions. Design two computationally efficient metrics for structural ensembles. The first metric, $\mathbb{V}(\mathcal{E}_1)$, should approximate the within-ensemble variance (partially a proxy of structural flexibility). The second metric, $d(\mathcal{E}_1, \mathcal{E}_2)$ should approximate distances between ensembles and must have the property that $d(\mathcal{E}_1, \mathcal{E}_2) \approx 0$, when $\mathcal{E}_1 = \mathcal{E}_2$.

2. Investigate properties of the within-ensemble variance metric $\mathbb{V}(\mathcal{E}_1)$ with respect to RNA sequence length, GC content and folding temperature.

3. Benchmark $\mathbb{V}(\mathcal{E}_1)$ against known stable RNA structures and known flexible RNA structures. Flexible structures are expected to have higher $\mathbb{V}(\mathcal{E}_1)$.

4. Investigate ensemble distance, $d(\mathcal{E}_1, \mathcal{E}_2)$, for ensembles corresponding to wild-type RNAs and disease-associated RNAs that differ from wild-type by a single nucleotide and test whether ensemble distance from wildtype structures correlate with disease (see [2]).

Long-term objectives:

1. Perform a large scale analysis of human RNAs, and test whether flexible structures (as measured using $\mathbb{V}(\mathcal{E}_1)$) are associated with certain functional annotations.

2. Develop a robust and computationally efficient method for pairwise alignment of RNA structural ensembles. This would be particularly valuable, as alignment becomes necessary when comparing two structural ensembles corresponding to two RNA sequences of differing length. This would eventually be expected to be applied on a genome-wide scale.

Tasks:

1. Basic introduction to computer representations of RNA secondary structure (dot-bracket notation).

2. Understand the difference between thermodynamic and SCFG approaches to RNA secondary structure prediction.

3. Read [4]. Implement the mountain metric for measuring distances between two RNA secondary structures.

4. Read [2]. Motivates why considering structural ensembles is important and provides good demonstrations of structural ensembles.

5. Before computing the desired statistics, i.e. $\mathbb{V}(\mathcal{E}_1)$ and $d(\mathcal{E}_1, \mathcal{E}_2)$, you need to be able to generate statistically representative samples (RNA secondary structures) from RNA ensembles. There are multiple approaches to sampling:

   (a) Recommend approach. Use ViennaRNA's RNAsubopt to stochastically draw samples with probability proportional to their Boltzmann weights. Requires programmatically shelling to the RNAsubopt executable. Thermodynamic only. For example, the following command: `RNAsubopt.exe --stochBT=1000 --temp=37.0 < input_sequence.fas > output_sample.txt`, will take as input a sequence, fold it at 37.0C and generate a sample of 1000 secondary structures from the folded ensemble.

   (b) Implement a MCMC algorithm to sample from either a thermodynamic or SCFG ensemble. Requires designing valid proposal between RNA secondary structures. Will likely require implementing a parallel tempering algorithm due to the multi-modal nature of RNA ensembles. Has the disadvantage that it will be slow and samples are dependent (auto-correlated).

   (c) Sample using inside-outside probabilities. For SCFGs only, programmatically difficult, but will generate independent samples. Similar to RNAsubopt's stochastic traceback.

6. [Possible task] Do PCA analysis, as in [2].

   (a) Generate samples from the wild-type $\mathcal{E}_w$ and mutant ensembles: $\mathcal{E}_1, \ldots, \mathcal{E}_m$.

   (b) Compute mountain vectors[4] for each of the ensemble samples.

   (c) Perform PCA analysis seperately on each of the ensemble samples.

7. [Possible task] A comparison of ensembles generated via the two different approaches (thermodynamic vs. SCFGs).

8. Design a metric, $\mathbb{V}(\mathcal{E}_1)$, for approximating within-ensemble variance. As a test case apply to a single sequence and plot $\mathbb{V}(\mathcal{E}_i)$ versus temperature T_i. We expect within-ensemble variance to be positively correlated with temperature.

9. Design a metric, $d(\mathcal{E}_1, \mathcal{E}_2)$, for measuring a distance between two ensembles. As a test case compute distances between a wild-type RNA ensemble and disease-associated RNA ensembles that differ from wild-type by a single nucleotide. Test whether ensemble distance from the wildtype ensemble correlates with disease-association.

References

[1] Stephan H Bernhart, Hakim Tafer, Ulrike Mückstein, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for Molecular Biology*, 1(1):3, 2006.

[2] Matthew Halvorsen, Joshua S Martin, Sam Broadaway, and Alain Laederach. Disease-associated mutations that alter the RNA structural ensemble. *PLoS genetics*, 6(8):e1001074, 2010.

[3] David H Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190, 2004.

[4] Vincent Moulton, Michael Zuker, Michael Steel, Robin Pointon, and David Penny. Metrics on RNA secondary structures. *Journal of Computational Biology*, 7(1-2):277–292, 2000.

[5] Jérôme Waldispühl and Peter Clote. Computing the partition function and sampling for saturated secondary structures of RNA, with respect to the turner energy model. *Journal of Computational Biology*, 14(2):190–215, 2007.