Motivation

The program MASTR (Multiple Alignment of STructural RNAs) predicts a common structure and a multiple alignment of non-coding RNAs using sampling in a simulated annealing framework.

- Structure prediction of ncRNAs is important for determining function.
- Using energy minimization, a single structure can be predicted (mfold, RNAfold).
- Including evolutionary information can improve predictions for multiple sequences.
- Simultaneous multiple alignment and structure prediction is NP-hard.
- Current programs for simultaneous multiple alignment and structure prediction: LocARNA, RNAalifold, FoldalignM, RNAsampler, RNAforester and SimulFold.

Method

The sampling approach is governed by an artificial temperature $T$ and a cost function. Initially, no structure is predicted, and a starting alignment is made by inserting gaps randomly in the sequences. The combined solution is optimized iteratively by making small changes, and the move is either accepted or rejected using the Metropolis–Hastings equation based on the cost and temperature:

$$P_{\text{accept}} = e^{-\text{COST}_{\text{NEW}}/T}$$

Cost function

The cost function considers both the quality of the sequence alignment and the predicted structure. The alignment is evaluated using the log-likelihood of the individual characters (nucleotides or gaps) and summing over all columns in the alignment:

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{L} \log[P(x_{ij}^*)]$$

The probability of a nucleotide depends on the alignment column, and the gap probabilities distinguish between opening a gap ($P_{\text{go}}$) and extending a gap ($P_{\text{go}}$) and also encourage stacking of gaps. In both cases, the preceding character also influences the probability:

$$\sum_{i;j} \log( P(x_{ij}^*) )$$

A proposed basepair is evaluated using a weighted combination of covariance and basepair probabilities. A covariance measure based on the $B$-score from RNAalifold is used (see right column). The novel measure $C_\lambda$ explicitly treats stacking of basepairs by extending to both sides of the proposed basepair:

$$C_\lambda = B_{\text{go}} + 2 \cdot B_{\text{go}} + B_{\text{go}}$$

The probability of a proposed basepair is included using McCaskills partition function as implemented in the Vienna package. The probability of a basepair between columns $i$ and $j$ is found as the average basepair probability of the sequences:

$$P(\text{BPP}_{ij}) = \frac{1}{N} \sum_{x_{ij}=A} \text{P}_{\text{go}}(x_{ij})$$

For a given structure, let $C$ be the sum of the covariance measure over all basepairs, and $P$ the sum of the logarithm of the basepair probabilities. The combined cost of the alignment and structure is a weighted sum of the log-likelihood, the covariation and the basepair probabilities:

$$\text{Cost}(MA,S) = Q + \alpha C + \beta P$$

The RNAalifold covariation measure

$$\delta \left[ \delta(a_i b_j \beta \beta) \right]: \text{Editing distance between two aligned pairs in sequences } \alpha \text{ and } \beta$$

$I_{\beta}^{ij}$: Matrix containing 1 if $(i,j)$ can form a basepair in sequence $\alpha$ and 0 otherwise

Penalty term measuring the fraction of sequences with inconsistent pairs

The covariation measure is then found as:

$$B_{ij} = \left( \frac{1}{L} \sum_{i; j} \delta(a_i b_j \beta \beta) \right) I_{\beta}^{ij} - q_{ij}$$

Changing the alignment

The optimization of the alignment is done by performing simple moves that can alter, extend and reduce the alignment:

- Move a gap block in the alignment
- Insert a gap-column in either end of the alignment
- Remove all gap-columns in the alignment

The gap block move is illustrated below:

1. Randomly locate a gap in a sequence
2. Extend gap block horizontally in sequences with some probability
3. Extend gap block horizontally in sequences with some probability
4. Move to new position in alignment

Changing the structure

Updating the structure is done similarly using simple moves under some structural constraints (e.g. allowing only non-crossing basepairs):

- Add a new basepair between columns $i$ and $j$
- Extend an already existing stem in either direction
- Remove a random basepair

Thus, the structure can be built, extended and reduced.

Results

- Tested on 5 families: tRNA, 5S rRNA, U5, TPP riboswitches and Picorna IRES.
- Covering ID from 30% to 100% in 10% intervals, and lengths from 70 to 250.
- Predictions evaluated on structure (MCC) and alignment (SPS).
- Compare with LocARNA, Clustal+Alifold, FoldalignM, RNAsampler, RNAforester
- Consistently best or comparable on structure and alignment
- Reasonably fast (few minutes) and can handle larger datasets than most programs.