**Motivation**

Many computational methods for RNA secondary structure prediction, and, in particular, for the prediction of a consensus structure of an alignment of RNA sequences, have been developed. Most methods however ignore biophysical factors such as the kinetics of RNA folding; no current implementation considers both evolutionary information and folding kinetics, thus losing information which, when considered, might lead to better predictions.

**Bayesian Weighting: The Distance Function**

In the basic model, distinct (pairs of) columns are assumed to be independent, contradicting biological fact. Therefore, we penalise the pairing of two columns if there exist columns between them which are likely to form a basepair incompatible with these two columns. We postulate

$$P_{\text{paired}}(s_i, s_j) = \exp\left(-\frac{d(i, j)}{K} - \beta s_i, s_j\right),$$

where $s_i, s_j$ are the usual base-pairing probabilities, and

$$d(i, j) = \begin{cases} 0 & \text{if } i = j; \\ d(i, j) & \text{if } i, j \text{ pair and } i < j; \\ \beta(i, j) + d(i + 1, j) & \text{otherwise}. \end{cases}$$

**Iterative Helix Formation**

It is known [3] that helices form quickly from a local basepair; in the sense that, once the first basepair of a helix has formed, nearby bases are more likely to pair. For this reason, we emulate the kinetics of RNA folding in a simplified way by forming helices iteratively. At each iteration, a new helix is determined conditionally on previously formed helices. By folding helices in one go, the fact that helices form quickly is taken into account. We postulate that the first helix to form is the helix (without bulges) containing the basepair

$$(i_{\text{max}}, j_{\text{max}}) = \arg \max \left\{ P_{\text{paired}}(i, j) \right\}.$$  (1)

where hats denote the posterior pairing and unpairing probabilities. In this framework, the statistic corresponding to MEA estimation is the expected difference

$$\Delta(i, j) = \hat{P}_{\text{paired}}(i, j) - \frac{1}{2} \left( \hat{P}_{\text{unpaired}}(i) + \hat{P}_{\text{unpaired}}(j) \right),$$

just as the difference in equation (2) is naturally interpreted as a measure of the energy and stability of a basepair, the corresponding basepairing probabilities can be considered as a measure of the time it takes for that basepair to form. Thus equation (1) just expresses the pairing of bases in the physical order. The algorithm is terminated when $\Delta(i_{\text{max}}, j_{\text{max}}) < \delta$ for the first time, where $\delta > 0$. This incorporates the physics that the first, local basepair needs to be stable enough for its dissociation time to be long enough for other basepairs to form. This also addresses the issue of the geometric, unphysical constraints of helix lengths in the grammar.

**Results**

Benchmarking was done on a set of 41 randomly selected sequences from the Rfam database [4]. Oxfold performed better than PPfold [5] in terms of averaged sensitivity, PPV and F-score. Additionally, it has a higher F-score and PPV than RNAalifold [6, 7]. In particular, Oxfold has a noticeably higher PPV than PPfold and RNAalifold.

**Outlook**

This work suggests the dynamical aspects of RNA folding should not be disregarded in SCFG approaches to RNA secondary structure prediction. A possible next step is therefore the inclusion of co-transcriptional effects into the model. But some general problems of SCFG approaches remain:

- **Pseudoknots.** Standard SCFG approaches cannot model pseudoknots, though iterative helix formation can be used to predict them.
- **Computational Complexity.** Standard SCFG approaches have complexity $O(t^4)$, where $t$ is sequence length, so Oxfold has complexity $O(t^3)$.
- **Non-Canonical Basepairs.** Examples suggest that the way algorithms deal with non-canonical basepairs is one of their bottlenecks.

**References**


**Acknowledgements**

This work was carried out as part of the Oxford Summer School in Computational Biology, 2012, in conjunction with the Department of Plant Sciences, and with support from the Department of Zoology. Funding was provided by EPSRC (TJW), BBSRC (PAH).