BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 6 2013, pages 704–710
doi:10.1093/bioinformatics/btt050

Structural bioinformatics

Advance Access publication February 8, 2013

Oxfold: kinetic folding of RNA using stochastic context-free grammars and evolutionary information

James W. J. Anderson[1,*], Pierre A. Haas[2], Leigh-Anne Mathieson[3], Vladimir Volynkin[4], Rune Lyngsø[1], Paula Tataru[5] and Jotun Hein[1]

[1]Department of Statistics, Oxford University, 1 South Parks Road, Oxford OX1 3TG, UK, [2]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK, [3]Department of Computer Science, University of British Columbia, 201-2366 Main Mall, Vancouver, B.C., V6T 1Z4, Canada, [4]European Bioinformatics Institute, Hinxton, Cambridgeshire CB10 1SD, UK and [5]Bioinformatics Research Centre, Aarhus University, C.F. Møllers Allé 8, DK-8000 Aarhus C, Denmark

Associate Editor: Anna Tramontano

ABSTRACT

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 that, when considered, might lead to better predictions.

Results: We present an iterative algorithm, Oxfold, in the framework of stochastic context-free grammars, that emulates the kinetics of RNA folding in a simplified way, in combination with a molecular evolution model. This method improves considerably on existing grammatical models that do not consider folding kinetics. Additionally, the model compares favourably to non-kinetic thermodynamic models.

Availability: http://www.stats.ox.ac.uk/~anderson.

Contact: anderson@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 28, 2012; revised on December 28, 2012; accepted on January 27, 2013

1 INTRODUCTION

The function of ribonucleic acid molecules (RNA) is known to depend on their 3D structure, which, in turn, depends on their secondary structure, a scaffold of base pairs formed by hydrogen bonds between nucleotides, for thermodynamic stability and molecular function. Accurate prediction of RNA secondary structures, however, falls short of being adequately solved. By contrast, the spreading of next-generation sequencing technologies and new methods in transcriptomics has increased the importance of RNA secondary structure prediction. This is exemplified by the growing amount of biological RNA data available in databases, such as Rfam (Gardner *et al*., 2011) and RNA STRAND (Andronescu *et al*., 2008).

Computational RNA secondary structure prediction methods have been used for a number of years: some of the first attempts (e.g. Pipas and McMahon, 1975) simply evaluate the free energy of all possible secondary structures, postulating that the minimum free-energy structure is the functional one. These thermodynamic models were later refined to take into account biological and thermodynamic principles and are used to great effect in algorithms such as RNAFold (Hofacker *et al*., 1994) and UNAFold (Markham and Zuker, 2008), which rely on a large number of experimentally determined parameters.

Alternative approaches use the framework of stochastic context-free grammars (SCFGs) to find the most likely structure given their training data, postulating that this is the functional structure. Among the first to describe such models were Eddy and Durbin (1994). Many different grammatical models for RNA secondary structure prediction have been implemented (Anderson *et al*., 2012; Dowell and Eddy, 2004; Knudsen and Hein, 1999, 2003).

If one seeks to build better grammatical models for RNA secondary structure prediction, one can take essentially two different routes:

- Build more complex grammars that express higher-order dependencies, such as base pair stacking, and, for instance, thereby emulate the nearest-neighbour model underlying thermodynamic approaches;
- Include additional biological and physical information about the sequences, for example, at the level of the previous pairing and unpairing probabilities of the grammar.

The former approach has been taken by Nebel and Scheid (2011) and Rivas *et al*. (2012). The latter developed a language to translate a wide variety of probabilistic and thermodynamic models for RNA secondary structure prediction into the language of SCFGs, yielding highly complex grammars with a large number of parameters.

However, we follow the latter approach, which was pioneered by Knudsen and Hein (1999, 2003), who coupled a simple grammar to an evolutionary model to obtain better estimates of the previous base pairing probabilities when folding an alignment of RNA sequences. Most current approaches to RNA secondary structure prediction are static, insofar as they assess structures based on their constituent elements like base pairs and loops but with no contribution from the path followed to form these

*To whom correspondence should be addressed.

elements, insofar as they fold sequences in one go, and thereby ignore the mechanisms of folding. The importance of folding mechanisms was noted by Tinoco *et al*. (1990) and Gultyaev *et al*. (1995), who studied the folding of intermediary stems. We note that '*the differences between real structures and the minimum energy states are believed to be determined mainly by defects in the energy rules used or by the existence of specific folding pathways capturing molecules in local minima*' (Gultyaev *et al*., 1995).

Just as comparative structure prediction is based on the observation that structure is important for function and, hence, conserved, as folding kinetics is important for either guiding or determining structure formation, we would expect evolution to exert selection on the kinetics too. Previously, evolutionary models (Knudsen and Hein, 1999, 2003) and kinetic models (e.g. Danilova *et al*., 2006; Xayaphoummine *et al*., 2005) have been implemented, but they have not been combined. It is, therefore, important to implement folding kinetics in an evolutionary framework.

In this article, we work in the framework of the fundamental problem of predicting a consensus structure for a given fixed alignment of RNA sequences. We incorporate folding kinetics, in a simplified way, into an evolutionary grammatical model in an iterative framework. Further, we introduce a distance function to incorporate information about the relationships between different pairs of columns, thus adopting the second of the aforementioned approaches. The resulting model is benchmarked against PPfold (Sükösd *et al*., 2011), a parallelized implementation of the Pfold algorithm of Knudsen and Hein (2003). Additionally, we compare it with a thermodynamic model, RNAalifold (Bernhart *et al*., 2008; Hofacker *et al*., 2002).

2 METHODS

2.1 Background: grammatical models

A context-free grammar (Chomsky, 1959) is a four-tuple $(\mathcal{N}, \mathcal{V}, \mathcal{P}, S)$ consisting of a finite set $\mathcal{N}$ of non-terminals, a finite set $\mathcal{V}$, disjoint form $\mathcal{N}$, of terminals, a finite set $\mathcal{P}$ of production rules and a distinguished starting symbol $S \in \mathcal{N}$. Each production rule replaces a non-terminal with a string of non-terminals and terminals.

For example, the context-free grammar underlying the Pfold algorithm of Knudsen and Hein (1999, 2003) is represented as

$$S \to LS|L \qquad L \to .|(F) \qquad F \to LS|(F)$$

It has non-terminals S, L, F and terminals .,(,), representing unpaired and paired nucleotides in the dot-parenthesis representation of RNA secondary structures. For example, the string ((..)) is produced by the derivation

$$\begin{array}{cccl} S & \to & L & \text{using rule } S \to L \\ & \to & (F) & \text{using rule } L \to (F) \\ & \to & ((F)) & \text{using rule } F \to (F) \\ & \to & ((LS)) & \text{using rule } F \to LS \\ & \to & ((LL)) & \text{using rule } S \to L \\ & \to & ((.L)) & \text{using rule } L \to . \\ & \to & ((..)) & \text{using rule } L \to . \end{array}$$

In the grammar of Knudsen and Hein (1999), the starting symbol S produces loops, whereas F produces stems and L determines whether a loop position should be a single base or the start of a new stem.

A SCFG is a context-free grammar with an associated probability distribution over the production rules. Thus, each string produced by the grammar (by beginning with the starting symbol and following production rules) is given a certain probability, which gives a probability distribution over RNA secondary structures. The rule probabilities are determined by inside–outside training, an expectation maximization technique (Lari and Young, 1990).

To complete the model, we require the previous probabilities of a dot representing any of the four nucleotides A, C, G and U and of a pair of parentheses representing any of the 16 corresponding base pairs. For example, the probability of producing an A from the non-terminal L is the product of the probability of the rule $L \to$. and of the previous nucleotide probability of a dot representing an A. For single sequences, these are simply the frequencies observed on training data.

For alignments of multiple sequences, rather than using the simple heuristic of just multiplying the maximum likelihood estimates of the pairing probabilities of the bases in a given pair of columns in each sequence to estimate the previous base pairing probabilities of that pair of columns, the Pfold algorithm uses an evolutionary model: each pair of columns is assumed to evolve independently according to a continuous Markov process with rates given by the branch lengths of an evolutionary tree estimated from the alignment. The base pairing probabilities are then determined by post-order traversal (Felsenstein, 1981) on the evolutionary tree. Gaps in the alignment are treated as unknown nucleotides.

One possible candidate for the consensus structure is the structure with the highest probability under grammar and evolutionary model, obtained using the Cocke-Younger-Kasami (CYK) algorithm, a dynamic programming algorithm (Durbin *et al*., 1998). However, this ignores contributions from other possible structures. Most current implementations, therefore, predict the structure with the highest expected number of correctly predicted base pairs (maximum expected accuracy, MEA, estimation). The latter is determined, using a dynamic programming algorithm, from the posterior pairing and unpairing probabilities, i.e. the pairing and unpairing probabilities given the sequence data and the model (Knudsen and Hein, 2003; Supplementary Material), which, in turn, are determined from the matrices of inside and outside probabilities associated with the SCFG (Lari and Young, 1990).

For RNA secondary structure prediction, it is most convenient to write the production rules in the double-emission form of Anderson *et al*. (2012), which only allows rules of the forms $U \to .$, $U \to VW$ and $U \to (V)$, where U, V, W denote generic non-terminals. Throughout this article, we use the grammar of Knudsen and Hein (1999) rewritten in double-emission form, viz

$$S \to LS|.|(F) \qquad L \to .|(F) \qquad F \to LS|(F)$$

This slightly reformulated version of the grammar produces the same probability distribution over strings; therefore, predictions will be the same. The generalized expressions for the inside–outside and posterior probabilities used in this article are given in this double-emission form.

2.2 Folding kinetics: iterative helix formation

The kinetics of RNA folding have been studied by Craig *et al*. (1971), who determined the speed at which helices form. They showed that helices form quickly from a local base pair, in the sense that, once the first base pair of a helix has formed, nearby bases are more likely to pair.

This motivates emulating the kinetics of RNA folding in a simplified way by forming helices iteratively. Iterative helix formation has also been used by Harmanci *et al*. (2011). Once a suitable candidate base pair has been identified, a helix containing that base pair is formed.

Local helix formation: iterative MEA estimation. We postulate that the first helix to form is the helix (without bulges) containing the base pair

$$(i_{\max}, j_{\max}) = \underset{(i,j)}{\arg\max}\{\widehat{\mathbb{P}}_{\text{paired}}(i,j)\}\ , \tag{1}$$

where we use hats to denote the posterior pairing and unpairing probabilities obtained from the grammar. From a technical point of view, taking maximal probabilities in this way can be considered as a greedy approximation of the CYK algorithm (Durbin *et al.*, 1998).

In the framework of iterative helix formation, the statistic corresponding to MEA estimation is the expected difference in the number of correctly predicted base pairs after pairing bases i and j,

$$\Delta(i,j) = \widehat{\mathbb{P}}_{\text{paired}}(i,j) - \frac{1}{2}(\widehat{\mathbb{P}}_{\text{unpaired}}(i) + \widehat{\mathbb{P}}_{\text{unpaired}}(j)) \;, \tag{2}$$

Just as the difference in Equation (2) is naturally interpreted, by analogy with thermodynamic models, as a measure of the energy and, therefore, of the stability of a base pair, the corresponding base pairing probabilities can be considered as a measure of the time it takes for that base pair to form. If we approximate base pair formation as a continuous Markov process with rate equal to the posterior pairing probability, the time until helix formation has the exponential distribution with mean equal to the inverse of the posterior pairing probability. With this interpretation in mind, Equation (1) just expresses the pairing of bases in the physical order.

At each iteration, a new helix containing $(i_{\max}, j_{\max})$ and such that, for each base pair (i,j) in the helix, $\Delta(i,j) > 0$, is determined conditional on previously formed helices. By folding helices in one go, the fact that helices form quickly is taken into account. More helices are formed until $\Delta(i_{\max}, j_{\max}) < \delta$, for some threshold $\delta > 0$.

MEA estimation hinges on the assumption that the posterior base pairing probabilities given by the grammatical model are equal to the probability that a given base pair is correct. In fact, small positive values of the difference in Equation (2) are not reliable; therefore, requiring $\delta > 0$ might be expected to increase the positive predictive value of the algorithm. From a more physical point of view, to additionally require $\Delta(i,j) > \delta > 0$ for the base pair $(i_{\max}, j_{\max})$ is to incorporate the physics that once the first base pair is formed, nearby bases are more likely to pair. This local base pair needs to be 'strong' enough for its dissociation time to be long enough for other base pairs to form. This also addresses the issue of the geometric and, therefore, unphysical distribution of helix lengths in the grammar (Knudsen and Hein, 1999).

A remark on the evolutionary model. For alignments of sequences or subsequences with high-primary sequence conservation, the evolutionary might miss 'obvious' helices, as it introduces extra uncertainty. This is especially relevant in iterative helix formation because we only ever try to pair bases that have high-posterior pairing probabilities. For this reason, rather than using the evolutionary model as in Knudsen and Hein (2003), we use a mixed model: at the start, we form base pairs with very high-posterior pairing probabilities using the simpler heuristic of just multiplying the base pairing probabilities for each sequence to obtain the previous base pairing probabilities and then switch to the full evolutionary model. The ability to mix different methods is strength of the iterative approach.

2.3 Bayesian weighting: the distance function

In the model we have built up to this point, distinct columns and pairs of columns are, insofar as previous unpairing and pairing probabilities are concerned, assumed to be independent. This does not reflect biological fact, as interlacing structures prevent each other from forming, and a column cannot pair with two different columns at the same time.

We note that this kind of crossing interaction may lead to pseudoknot formation. Standard SCFG approaches are unable to predict pseudoknots (Brown and Wilson, 1995): in standard MEA estimation, one hopes that the model predicts the more stable of the two interlacing structures. In the iterative framework, the extra information is used to address, more generally, the most obvious drawback of the iterative approach: once a helix that is incompatible with the correct structure has been formed, the final prediction is likely to be poor. From a kinetic perspective, base pairs frequently blocked by other transient base pairs would take longer to form, as the underlying continuous Markov process is only enabled during intervals where the base pair is not in conflict with other base pairs.

To include these physical dependencies between columns in the model, we look to penalize the pairing of two columns if there exist columns between that are likely to form a base pair incompatible with these two columns. As standard SCFG approaches cannot model pseudoknots; here, 'incompatible' does not only refer to base pairs that share a position with these two columns but also to base pairs that would form a pseudoknot with these two columns. Thus, we discount the previous base pairing probabilities by an exponential factor based on a distance function, so that they take the form

$$\mathbb{P}_{\text{paired}}(s_i, s_j) = \exp\left(-\frac{d(i,j)}{K|j-i|}\right)\varpi(s_i, s_j) \;, \tag{3}$$

where s_i denotes the ith column of the alignment, and where $d(i,j)$ is a distance function to be specified, K is a weighting parameter and $\varpi(s_i, s_j)$ are the usual base pairing probabilities derived from an evolutionary model or the simple heuristic mentioned previously.

We choose a distance function such that two columns are 'far away' from each other if there are columns between them that are likely to form a base pair incompatible with these two columns. Each intermediate column k, with $i < k < j$, is given a weight equal to the probability of that column forming a base pair incompatible with $(k-1, j)$. Figure 1 shows an example of the distance function on a partially folded structure. Formally, we define, for $i \le j$,

$$d(i,j) = \begin{cases} 0 & \text{if } i = j; \\ d(\bar{\imath}, j) & \text{if } i, \bar{\imath} \text{ pair and } i \le \bar{\imath} \le j; \\ \beta(i,j) + d(i+1, j) & \text{otherwise.} \end{cases} \tag{4}$$

For example, in Figure 1, we calculate the distance between i and j, following Equation (4). We have $d(i,j) = \beta(i,j) + d(i+1,j)$ and $d(i+1,j) = \beta(i+1,j) + d(i+2,j)$ and so on. Further, $d(k,j) = d(\bar{k}, j)$, as k and $\bar{k}$ are paired. Continuing, $d(\bar{k}+1, j) = \beta(\bar{k}, j) + d(\bar{k}+1, j)$, and so on; finally, $d(j,j) = 0$.

Here, $\beta(i,j)$ is the probability that column $i+1$ forms a base pair that is incompatible with (i,j), so that, the events in question being disjoint,

$$\beta(i,j) = 1 - \widehat{\mathbb{P}}_{\text{unpaired}}(i+1) - \sum_{k=i+2}^{j-1} \widehat{\mathbb{P}}_{\text{paired}}(i+1, k) \;.$$

Thus, the posterior pairing probabilities are used to guide the folding of the next iteration (and the posterior probabilities for the wholly unfolded sequence, without the exponential weighting factor, are used to calculate the distance function for the first iteration).

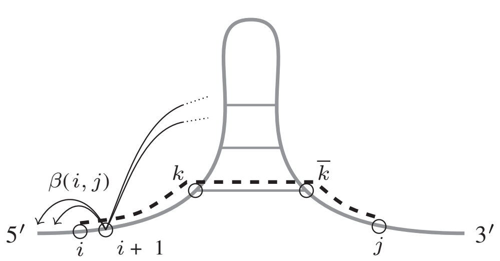
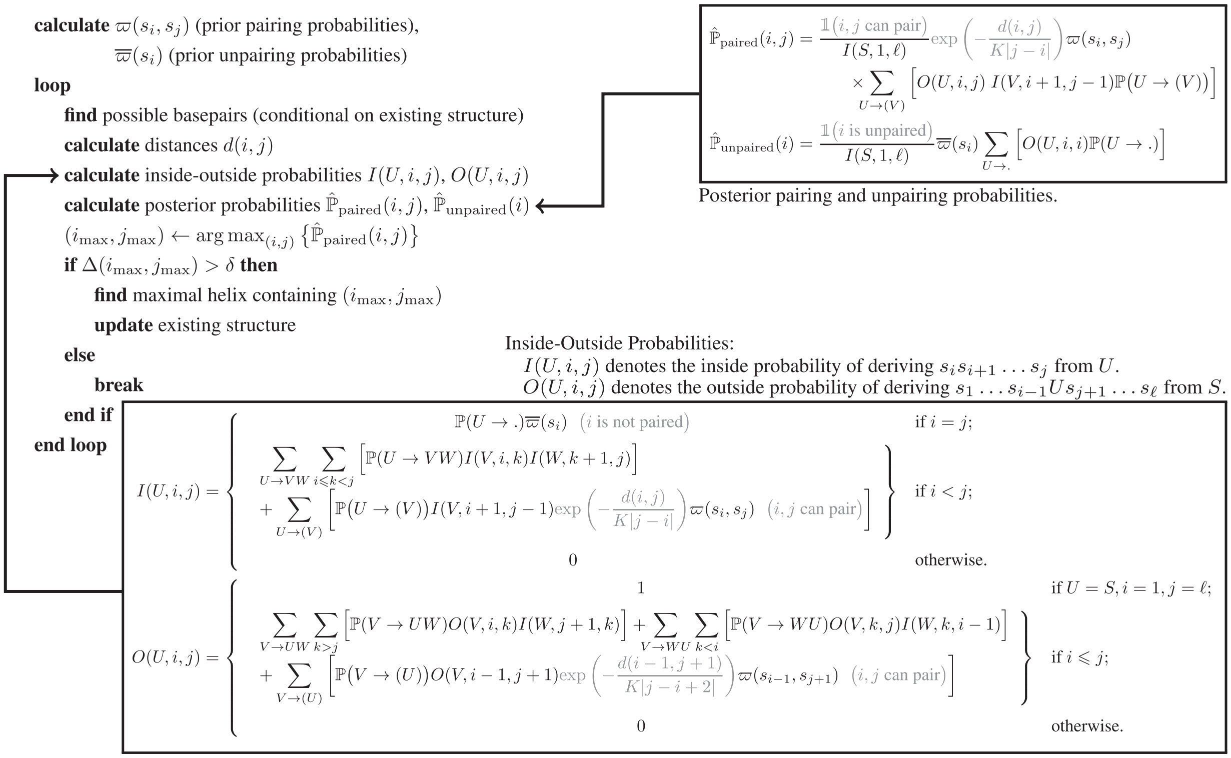


Fig. 1. Representation of the distance function. The distance between two positions i and j in the sequence is the shortest distance moving along the structure, allowing shortcuts across stems. Intermediate positions are weighted with the probabilities of certain incompatible base pairs. Informally, this can be thought of as the shortest distance between two nodes of a weighted graph with edges between paired and adjacent positions. See text for further explanation



```
calculate $\varpi(s_i, s_j)$ (prior pairing probabilities),
          $\overline{\varpi}(s_i)$ (prior unpairing probabilities)
loop
    find possible basepairs (conditional on existing structure)
    calculate distances $d(i, j)$
    calculate inside-outside probabilities $I(U, i, j)$, $O(U, i, j)$
    calculate posterior probabilities $\hat{\mathbb{P}}_{\text{paired}}(i, j)$, $\hat{\mathbb{P}}_{\text{unpaired}}(i)$
    $(i_{\max}, j_{\max}) \leftarrow \arg\max_{(i,j)} \{\hat{\mathbb{P}}_{\text{paired}}(i, j)\}$
    if $\Delta(i_{\max}, j_{\max}) > \delta$ then
        find maximal helix containing $(i_{\max}, j_{\max})$
        update existing structure
    else
        break
    end if
end loop
```

$$\hat{\mathbb{P}}_{\text{paired}}(i,j) = \frac{\mathbb{1}(i,j \text{ can pair})}{I(S,1,\ell)} \exp\left(-\frac{d(i,j)}{K|j-i|}\right) \varpi(s_i, s_j) \times \sum_{U \to (V)} \Big[O(U,i,j)\; I(V,i+1,j-1)\mathbb{P}\big(U \to (V)\big)\Big]$$

$$\hat{\mathbb{P}}_{\text{unpaired}}(i) = \frac{\mathbb{1}(i \text{ is unpaired})}{I(S,1,\ell)} \overline{\varpi}(s_i) \sum_{U \to .} \Big[O(U,i,i)\mathbb{P}(U \to .)\Big]$$

Posterior pairing and unpairing probabilities.

Inside-Outside Probabilities:
$I(U,i,j)$ denotes the inside probability of deriving $s_i s_{i+1} \ldots s_j$ from U.
$O(U,i,j)$ denotes the outside probability of deriving $s_1 \ldots s_{i-1} U s_{j+1} \ldots s_\ell$ from S.

$$I(U,i,j) = \begin{cases} \mathbb{P}(U \to .)\overline{\varpi}(s_i) \quad (i \text{ is not paired}) & \text{if } i = j; \\ \displaystyle\sum_{U \to VW} \sum_{i \leqslant k < j} \Big[\mathbb{P}(U \to VW) I(V,i,k) I(W,k+1,j)\Big] + \sum_{U \to (V)} \Big[\mathbb{P}\big(U \to (V)\big) I(V,i+1,j-1) \exp\left(-\frac{d(i,j)}{K|j-i|}\right) \varpi(s_i,s_j) \quad (i,j \text{ can pair})\Big] & \text{if } i < j; \\ 0 & \text{otherwise.} \end{cases}$$

$$O(U,i,j) = \begin{cases} 1 & \text{if } U = S, i = 1, j = \ell; \\ \displaystyle\sum_{V \to UW} \sum_{k > j} \Big[\mathbb{P}(V \to UW) O(V,i,k) I(W,j+1,k)\Big] + \sum_{V \to WU} \sum_{k < i} \Big[\mathbb{P}(V \to WU) O(V,k,j) I(W,k,i-1)\Big] + \sum_{V \to (U)} \Big[\mathbb{P}\big(V \to (U)\big) O(V,i-1,j+1) \exp\left(-\frac{d(i-1,j+1)}{K|j-i+2|}\right) \varpi(s_{i-1},s_{j+1}) \quad (i,j \text{ can pair})\Big] & \text{if } i \leqslant j; \\ 0 & \text{otherwise.} \end{cases}$$

Fig. 2. Simplified pseudocode summarizing the full kinetic folding algorithm. The inside–outside and posterior probabilities are written in the double emission form of Anderson *et al.*, 2012, and include the distance function and structural constraints. The modifications needed to take account of structural constraints and to introduce the distance function are shown in grey. Capital letters denote generic non-terminals, whereas lower-case letters denote column indices and s_i denotes the ith column of the alignment; ℓ denotes the total sequence length. See text for detailed explanation

This completes the set-up of the kinetic folding algorithm. A summary of the algorithm is given as pseudocode in Figure 2, which also shows the expressions for the inside–outside and posterior probabilities conditional on existing base pairs.

As the distance function, as shown in Figure 1, allows shortcuts across stems, this distance function implements the physics that, once a base pair has been formed, nearby bases are more likely to pair (Craig *et al.*, 1971). We note that the distance function imposes a certain hierarchy on substructures, making it more attractive to pair interior stems before pairing exterior ones.

It is important to note the effect of the denominator $K|j-i|$ in the exponential factor in Equation (3): we do not penalize absolute distance between base pairing partners, but rather, the penalization is relative, comparing the probabilistic weights β to the parameter K.

From a more fundamental point of view, this is a Bayesian approach: the observed 'horizontal' relationships between the columns in the alignment are used to update the previous information from the evolutionary model, which evaluates the 'vertical' relationships between the sequences in the alignment. Thus, these two parts of the model complement each other.

3 DISCUSSIONS

The full kinetic model for RNA secondary prediction, Oxfold, was benchmarked against PPfold (Sükösd *et al.*, 2011), a parallelized implementation of the Pfold model of Knudsen and Hein (1999, 2003). We also evaluate the performance of RNAalifold (Bernhart *et al.*, 2008; Hofacker *et al.*, 2002), a thermodynamic model without evolutionary information.

Benchmarking data and parameters. For benchmarking purposes, we have created a curated RNA dataset based on the Rfam database (Gardner *et al.*, 2011). Alignments of homologous RNA sequences with their consensus secondary structure were extracted from among those Rfam seed alignments that bear the 'published' tag. From these, 50 alignments with at least 50 sequences were randomly selected. We note that random data selection ensures reliability of results. In a pre-filtering step, we discarded outlier sequences with many/long insertions and deletions from each family, using a similar approach to that of PPfold (Sükösd *et al.*, 2011), which does not consider columns with >75% of gaps for pairing. Indels were first determined relative to a family consensus sequence; then a total mismatch score was calculated based on indel lengths; and sequences that had significantly larger mismatch score than the family mean were deleted. Further random selection was performed to reduce these to alignments of five sequences each; the results of Knudsen and Hein (1999, 2003) suggest that this suffices to take into account the evolutionary information. Because of the computational complexity of the model (which we discuss further in the text), we restricted to 41 alignments of length up to 214, with an average length of 105.

The consensus secondary structures given in Rfam may be slightly different from the secondary structures that the individual sequences fold into. In this sense, we cannot say that our secondary structures are experimentally verified, but the approach of comparing predictions to these secondary structures is commonly used in analysis of comparative prediction methods. As with all benchmarks of this nature (Bernhart *et al*., 2008; Knudsen and Hein, 1999), this should be taken into account.

It is known that grammar performance depends on datasets (Rivas *et al*., 2012). Consequently, it is important to monitor dataset dependence, in particular to avoid overfitting. For these reasons, the grammar parameters and evolutionary trees used for benchmarking purposes were those of PPfold. In particular, the grammar underlying our present approach is essentially the simple grammar for which Rivas *et al*. (2012) did not find evidence of overfitting. The other parameters were chosen heuristically: the parameter δ was set to 0.5 to make the first, local base pair stable enough for its dissociation time to be long enough for other base pairs to form. This is a trade-off between losing sensitivity at high values of δ and losing positive predictive value (PPV) at low values of δ. Similarly, K determines the amount of penalization by the distance function; setting $K = 0.5$ leads to a maximum penalization of about one order of magnitude.

Benchmarking statistics. We assess the performance of these models on a single alignment by calculating the sensitivity, PPV, defined by

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where TP, FP and FN denote the number of true positives (number of correctly predicted base pairs), false positives (wrong base pairs predicted) and false negatives (true base pairs not predicted), respectively. We also determine the F-score, which is the harmonic mean of sensitivity and PPV:

$$\text{F-score} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}}.$$

The averaged values of the F-scores, sensitivities and PPVs of the structures predicted by the kinetic models are compared, in Table 1, with those of PPfold and RNAalifold. The predictions

Table 1. Comparison of the performance of different algorithms on the test dataset of 41 alignments

Algorithm	F-Score	Sensitivity	PPV
RNAalifold	0.704	**0.748**	0.689
PPfold	0.673	0.650	0.728
Iterative without evolutionary model, without distance function	0.684	0.698	0.694
Iterative without evolutionary model, with distance function	0.688	0.696	0.703
Iterative with evolutionary model, without distance function	0.698	0.666	0.780
Oxfold (full kinetic model)	**0.723**	0.688	**0.800**

The algorithms presented in this article are compared with PPfold (Sükösd *et al*., 2011) and RNAalifold (Bernhart *et al*., 2008; Hofacker *et al*., 2002). See Section 2 for details. The values shown are the averages of the F-score, sensitivity and PPV of the alignments in the test dataset. The values shown in bold type are the maximum values in the respective column.

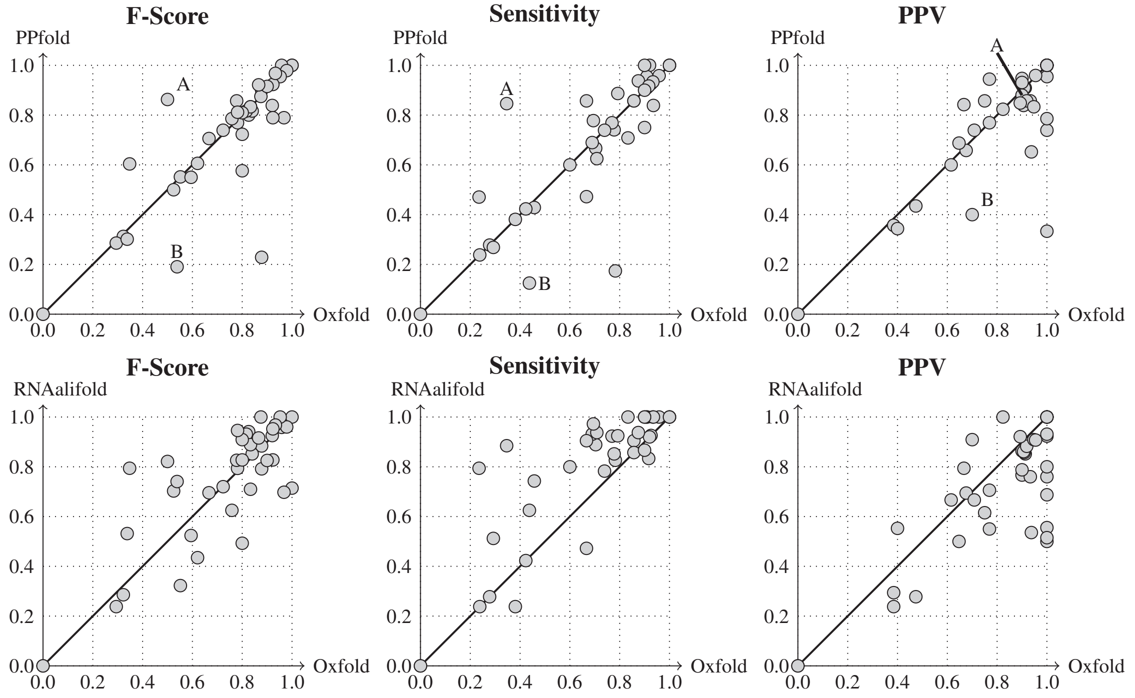


Fig. 3. Comparison of the F-score, sensitivity and PPV of the consensus structures predicted by Oxfold (the full kinetic model presented in this article) and those of PPfold (Sükösd *et al*., 2011) and RNAalifold (Bernhart *et al*., 2008; Hofacker *et al*., 2002) respectively, for the sequences in the test dataset of 41 alignments. The outliers marked A and B are discussed in the text

of the full kinetic model presented in this article are compared with those of PPfold and RNAalifold in Figure 3.

Discussion of results. We note that Oxfold performs better, on average, than PPfold in terms of averaged sensitivity, PPV and F-score. Additionally, it has a higher F-score and PPV than RNAalifold, although the thermodynamic model has a higher sensitivity. In particular, we observe that Oxfold has a noticeably higher PPV than PPfold and RNAalifold.

We also note that including an evolutionary model decreases the sensitivity of the algorithm, but greatly increases the PPV. Moreover, the distance function does not seem to be of much use without an evolutionary model. Both of these observations are compatible with the conclusion that the posterior probabilities without the evolutionary model are less reliable (in the sense that they somehow correlate with correctness) than those obtained with an evolutionary model. The fact that the iterative model without distance function, both with the mixed and the unmixed evolutionary model, has essentially the same PPV, gives further weight to this conclusion.

Also, the model with evolutionary information, but without distance function, performs mildly better than PPfold. This is notable because this method is the iterative method applied to the standard PPfold model. By conditioning on known structure, the grammar model no longer has probability mass contributed from incompatible structures, which one might hope would lead to a better structure prediction.

As PPfold and Oxfold share the same grammatical framework, outliers in Figure 3 can be attributed to the iterative method developed in this article. We discuss two such outliers, marked A and B in Figure 3 (see Supplementary Data for the corresponding predictions). In outlier A, the prediction cut-off δ is not met by a base pair predicted by PPfold, on which Oxfold terminates prediction (consequently, the PPVs are similar, but the sensitivity of PPfold is higher). By contrast, outlier B is an example of an alignment where Oxfold performs noticeably better than PPfold. Without the distance function, Oxfold has zero F-score on this alignment; with this distance function, the F-score rises to 0.5, illustrating how the distance function facilitates the prediction of correct base pairs.

With the interpretation of the difference in Equation (2) as a measure of the stability of a base pair, and of the posterior pairing probabilities as a measure of the inverse time it takes to form a base pair (discussed in Section 2), the fact that this model works indicates that RNA folds a stable scaffold before less stable substructures with short dissociation times start to appear and disappear (rather than folding its functional secondary structure while such substructures appear and disappear).

4 CONCLUSIONS

In this article, we have incorporated kinetic effects into a grammatical model for RNA secondary structure prediction by iterative formation of helices and by taking into account of some relationships between columns of an alignment by means of a distance function. Conceptually, introducing a distance function is the analogue, at the level of the emission probabilities of the grammar, of including (albeit possibly different) information about the relationships between columns in the alignment by making the production rules of the grammar more complex. The performance of the kinetic model suggests that the dynamical aspects of RNA folding should not be disregarded in SCFG approaches to RNA secondary structure prediction.

Incorporating co-transcriptional effects into the model might, therefore, be a possible next step: Kramer and Mills (1981) have shown that RNA folds as it is being transcribed, usually in the 5′–3′ direction. Thus, the 5′-end of the RNA molecule is allowed to fold before the sequence has been entirely transcribed, resulting in intermediate structures that do not necessarily exist in the final functional structure, as the speed of stem formation greatly exceeds the speed of transcription (Gultyaev *et al*., 1995). Moreover, Meyer and Miklós (2004) have demonstrated '*with statistical significance that co-transcriptional folding strongly influences RNA sequences in two ways*: (i) *alternative helices that would compete with the formation of the functional structure during co-transcriptional folding are suppressed* and (ii) *the formation of transient structures that may serve as guidelines for the co-transcriptional folding pathway is encouraged*'. Gultyaev *et al*. (1995), for instance, have incorporated co-transcriptional effects into a thermodynamic model by using a genetic algorithm.

Nevertheless, the fundamental problems affecting SCFG algorithms listed by Knudsen and Hein (1999) still remain much topical. Here, we discuss three issues of particular relevance to our algorithms:

Pseudoknots. As mentioned previously, standard SCFG approaches cannot predict pseudoknots (Brown and Wilson, 1995). Leaving the framework of SCFGs, but still using a formal grammar, it is possible to predict pseudoknotted structures (Rivas and Eddy, 2000). The iterative method presented in this article could be adapted to predict pseudoknots either by adapting the definition of permissible base pairs or by using the methods described by Rivas and Eddy (2000). Similarly, we would expect biophysical folding mechanisms to be conserved in pseudoknotted structures as in non-pseudoknotted structures.

Computational complexity. Standard SCFG algorithms for RNA secondary structure prediction have computational complexity $\mathcal{O}(\ell^3)$, where ℓ is the sequence length (Knudsen and Hein, 1999). Hence, the kinetic model presented in this article has complexity $\mathcal{O}(\ell^4)$, and a co-transcriptional algorithm along the lines of the algorithm of Gultyaev *et al*. (1995) (i.e. folding longer and longer subsequences, starting at the 5′-end) would have a complexity $\mathcal{O}(\ell^5)$, which makes the algorithms even more expensive than standard prediction approaches. Although it is straightforward, even intrinsic, to reuse computations for shorter subsequences in current methods for thermodynamic models, this may seem much more complex in a kinetic model, as the optimum pathway may completely change on the elongation of a subsequence. However, when the aim is to include co-transcriptional effects, it is not unreasonable to assume that relevant models can be formulated allowing algorithms with complexity lower than $\mathcal{O}(\ell^5)$. This matters especially for co-transcriptional folding, as one expects transcriptional effects to be stronger for longer sequences.

Non-canonical base pairs. The existence of non-canonical base pairs is a possible complication in RNA secondary prediction, for long, correct helices with non-canonical base pairs may seem less attractive than short, yet spurious helices. This observation ties the non-canonical base pairs issue somewhat to the geometric, and, therefore, unphysical, distribution of the helix lengths in the grammar (Knudsen and Hein, 1999). Here, we have addressed that issue by making the pairing of the first base pair of a helix more expensive than that of later pairs (as discussed in Section 2). The effect of more complex distributions of helix lengths has previously been studied by Dowell and Eddy (2004) and Rivas *et al*. (2012), who considered more complex grammars, allowing for stacking non-terminals in the grammar.

Nonetheless, RNAalifold (Bernhart *et al*., 2008; Hofacker *et al*., 2002) does not allow non-canonical base pairs in its default settings, whereas PPfold (Sükösd *et al*., 2011) associates very low-pairing probabilities with non-canonical base pairs. Some gain in sensitivity might, therefore, be possible by allowing some non-canonical base pairs in pairs of alignment columns at low-probability cost, but more insight into the role of non-canonical base pairs (and a corresponding model) may well be required.

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 the Department of Zoology. The authors thank the Oxford Supercomputing Centre for computational resources and support, and the EU grant COGANGS.

Funding: Engineering and Physical Sciences Research Council (J.W.J.A.); Biotechnology and Biological Sciences Research Council (P.A.H.).

Conflict of Interest: none declared.

REFERENCES

Anderson,J.W.J. *et al.* (2012) Evolving stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **13**, 78.

Andronescu,M. *et al.* (2008) RNA strand: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.

Bernhart,S.H. *et al.* (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Brown,M. and Wilson,C. (1995) RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search. *Pac. Symp. Biocomput*., 109–125.

Chomsky,N. (1959) On certain formal properties of grammars. *J. Math Control Inf*., **2**, 137–167.

Craig,M.E. *et al.* (1971) Relaxation kinetics of dimer formation by self complementary oligonucleotides. *J. Mol. Biol*., **62**, 383–401.

Danilova,L.V. *et al.* (2006) RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA. *J. Bioinform. Comput. Biol*., **4**, 589–596.

Dowell,R.D. and Eddy,S.R. (2004) Evaluation of several leightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 71.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*., **22**, 2079–2088.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*., **17**, 368–376.

Gardner,P.P. *et al.* (2011) Rfam: wikipedia, clans and the "decimal" release. *Nucleic Acids Res*., **1**, 5.

Gultyaev,A.P. *et al.* (1995) The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol*., **250**, 37–51.

Harmanci,A.O. *et al.* (2011) TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC Bioinformatics*, **12**, 108.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Chem. Mon*., **125**, 167–188.

Hofacker,I.L. *et al.* (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol*., **319**, 1059–1066.

Knudsen,B. and Hein,J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.

Knudsen,B. and Hein,J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*., **31**, 3423–3428.

Kramer,F.R. and Mills,D.R. (1981) Secondary structure formation during RNA synthesis. *Nucleic Acids Res*., **9**, 5109–5124.

Lari,K. and Young,S.J. (1990) The estimation of stochastic context-free grammars using the inside-outside algorithm. *Comput. Speech Lang*., **4**, 35–56.

Markham,N. and Zuker,M. (2008) UNAFold: software for nucleic acide folding and hybridization. In: Keith,J.M. (ed.) *Bioinformatics, Volume II. Structure, Function and Applications*. Humana Press, Totowa, pp. 3–31.

Meyer,I. and Miklós,I. (2004) Co-transcriptional folding is encoded within RNA genes. *BMC Mol. Biol*., **5**, 10.

Nebel,M.E. and Scheid,A. (2011) Evaluation of a sophisticated SCFG design for RNA secondary structure prediction. *Theory Biosci*., **130**, 313–336.

Pipas,J.M. and McMahon,J.E. (1975) Method for predicting RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **72**, 2017–2021.

Rivas,E. and Eddy,S.R. (2000) The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics*, **16**, 326–333.

Rivas,E. *et al.* (2012) A range of complex probabilistic models for RNA secondary structure prediction that include the nearest neighbor model and more. *RNA*, **18**, 193–212.

Sükösd,Z. *et al.* (2011) Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars. *BMC Bioinformatics*, **12**, 103.

Tinoco,I. *et al.* (1990) RNA folding. In: Eckstein,F. and Lilley,D.M.J. (eds) *Nucleic Acids and Molecular Biology*. Springer, New York, pp. 205–226.

Xayaphoummine,A. *et al.* (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res*., **33**, W605–W610.