Bioinformatics, 2016, 1–10
doi: 10.1093/bioinformatics/btw175
Advance Access Publication Date: 21 April 2016
Original Paper

OXFORD

Sequence analysis

ProbFold: a probabilistic method for integration of probing data in RNA secondary structure prediction

Sudhakar Sahoo[1], Michał P. Świtnicki[1] and Jakob Skou Pedersen[1,2,*]

[1]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Aarhus N 8200, Denmark and [2]Bioinformatics Research Centre, Aarhus University, Aarhus C DK-8000, Denmark

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on June 22, 2015; revised on February 17, 2016; accepted on March 28, 2016

Abstract

Motivation: Recently, new RNA secondary structure probing techniques have been developed, including Next Generation Sequencing based methods capable of probing transcriptome-wide. These techniques hold great promise for improving structure prediction accuracy. However, each new data type comes with its own signal properties and biases, which may even be experiment specific. There is therefore a growing need for RNA structure prediction methods that can be automatically trained on new data types and readily extended to integrate and fully exploit multiple types of data.

Results: Here, we develop and explore a modular probabilistic approach for integrating probing data in RNA structure prediction. It can be automatically trained given a set of known structures with probing data. The approach is demonstrated on SHAPE datasets, where we evaluate and selectively model specific correlations. The approach often makes superior use of the probing data signal compared to other methods. We illustrate the use of ProbFold on multiple data types using both simulations and a small set of structures with both SHAPE, DMS and CMCT data. Technically, the approach combines stochastic context-free grammars (SCFGs) with probabilistic graphical models. This approach allows rapid adaptation and integration of new probing data types.

Availability and Implementation: ProbFold is implemented in C++. Models are specified using simple textual formats. Data reformatting is done using separate C++ programs. Source code, statically compiled binaries for x86 Linux machines, C++ programs, example datasets and a tutorial is available from http://moma.ki.au.dk/prj/probfold/.

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

Contact: jakob.skou@clin.au.dk

1 Introduction

Obtaining accurate secondary structure predictions is a crucial step toward understanding the physical properties of RNA molecules and the biological roles of structural RNA elements. However, computational predictions based only on the primary sequence are often inaccurate and therefore insufficient for in-depth interpretation. Similarly, structure probing data typically only provides partial structural information (Wan *et al.*, 2011; Weeks, 2010). Directly modeling both types of data in the folding process generally improves structure prediction accuracy (Cordero *et al.*, 2012a; Deigan *et al.*, 2009; Mathews *et al.*, 2004; Sükösd *et al.*, 2012; Swenson *et al.*, 2012; Washietl *et al.*, 2012; Zarringhalam *et al.*, 2012). While most studies have used methods based on energy-minimization, the inclusion of probing data in probabilistic folding models has not been fully explored (Sükösd *et al.*, 2012).

Here, we develop and explore a modular probabilistic approach for integrating probing data in RNA secondary structure prediction, which we call ProbFold. The focus is on how to best exploit the

structure signal of the probing data, while keeping the models general and easy to adapt to new probing data types or additional layers of probing data. We have not aimed to develop a competitive single-sequence structure prediction method. The main focus is therefore on the use of the probing data signal rather than the overall performance.

Probabilistic modeling offers a coherent framework for combining different types of evidence as they are all naturally measured on the same scale. Structure models based on energy minimization do not extend naturally to additional types of data in the same way. For instance, probing data measurements must be translated to pseudo-energy perturbations before they can be included in the models, though they do not have inherent thermodynamic interpretations.

Probabilistic approaches have previously been used to incorporate comparative evidence in structure folding (Eddy and Durbin, 1994; Knudsen and Hein, 1999, 2003; Nawrocki and Eddy, 2013; Pedersen *et al.*, 2004, 2006; Rivas and Eddy, 2001; Sakakibara *et al.*, 1994). This is another example of supplementing the primary sequence with partial structure evidence and as such closely related to the probing data modeling problem studied here. Several of these methods exploit that generative probabilistic methods can be combined and their parameters optimized in a unified approach. For instance, pfold combines previously established models of molecular evolution (phylogenetic models) with probabilistic models of RNA secondary structure (stochastic context-free grammars – SCFGs).

We aimed to develop a method, called ProbFold, that could be readily extended to disparate data types and could encompass different probabilistic models for these. In particular they should be able to capture correlations both within and between data types. This is achieved by combining SCFGs with probabilistic graphical models (PGMs). The SCFG defines a prior over secondary structures, as it does in most other probabilistic methods (Rivas *et al.*, 2012). The PGMs model the sequence and any layers of probing data given the structure. PGMs are flexible and modular models useful for capturing select dependencies in high dimensional data (Koller and Fridman, 2009). In our case, they model dependencies between the sequence and the probing data as well as dependencies along the sequence.

Standard algorithms allow efficient training of both SCFG and PGMs as well as prediction of the optimal secondary structure. Importantly, this allows ProbFold to be automatically trained without hand-setting any parameters given a sufficiently sized training set of known structures with probing data. The size of the needed training set depends on the number of free parameters in the model.

RNA structure probing has a long history and many different methods exist (Ehresmann *et al.*, 1987), including use of chemical agents (Merino *et al.*, 2005; Karaduman *et al.*, 2006; Tijerina *et al.*, 2007), RNases (Kertesz *et al.*, 2010a,b) and spontaneous cleavage (Regulski and Breaker, 2008). Generally these modify bases or the backbone preferentially at either single or paired positions, allowing positional information on base-pair status through gel electrophoresis or sequencing. In the case of the SHAPE reagent (selective 2′-hydroxyl acylation analyzed by primer extension), it is the flexibility of the backbone that determines reactivity, which is generally higher for unpaired than paired regions (McGinnis *et al.*, 2012; Merino *et al.*, 2005; Weeks, 2010).

The interpretation of structure probing data is challenged by incomplete specificity of the methods, noisy or missing data, nucleotide biases, etc., which results in incomplete labeling of the primary sequence into paired and unpaired positions. For instance, in the case of SHAPE, the distributions of reactivities for paired and unpaired bases are largely overlapping (see Section 3). There is therefore a great need for computational methods that can integrate and make optimal use of probing data, beyond interpreting the data as a definite labeling of the primary sequence.

Recently, progress have has been made on this problem with both physics-based methods (Deigan *et al.*, 2009; Mathews *et al.*, 2004; Merino *et al.*, 2005; Swenson *et al.*, 2012; Washietl *et al.*, 2012; Zarringhalam *et al.*, 2012), sampling based methods (Ouyang *et al.*, 2013; Quarrier *et al.*, 2010) and a probabilistic method (Sükösd *et al.*, 2012). These are briefly presented below. See Eddy (2014) for a detailed review and discussion of their statistical foundations.

In RNAstructure (Deigan *et al.*, 2009; Mathews *et al.*, 2004), SHAPE reactivities are converted to base-stacking pseudo-energy change terms using a linear model optimized by prediction performance on a known structure (23S rRNA from *Escherichia coli* (*E.coli*)). Zarringhalam *et al.* (2012) extends this approach in the RNAsc method, by also including pseudo-energy change terms for unpaired positions. GTfold provides a fast parallelized multi-core implementation of the energy minimization algorithm and similarly includes SHAPE data (Swenson *et al.*, 2012). In an extension of RNAfold, pseudo-energy change terms are optimized for each structure given a loss function to maximize the agreement between the energy-based prediction and the experimental observations. In particular, no change is made when the sequence prediction is in complete agreement with the SHAPE data (Washietl *et al.*, 2012). Recently, the ViennaRNA Package has been extended to include several forms of soft constraints and specifically include implementations of the Deigan *et al.* (2009), Zarringhalam *et al.* (2012) and Washietl *et al.* (2012) approaches (Lorenz *et al.* 2015). A thorough performance comparison of the three approaches is also provided (Lorenz *et al.* 2015).

Another class of approaches samples structures from the Boltzmann-weighted ensemble and selects a representative structure with minimal Manhattan distance to a probing data profile (Quarrier *et al.*, 2010). This has been extended to several layers of probing data by a method that reduces them to a single binary pairing status profile (Ouyang *et al.*, 2013).

The recently proposed probabilistic method, PPfold 3.0 (Sükösd *et al.*, 2012), extends pfold (Knudsen and Hein, 1999) by modeling both comparative sequence alignment data and experimental probing data. It uses stochastic context-free grammars (SCFGs) to model secondary structures, phylogenetic models to model alignment columns and fine-grained discrete probability distributions to model SHAPE probing data. The study provides a proof of concept for including probing data in probabilistic methods.

Compared to PPfold 3.0, ProbFold offers a more general, extendible and parameter-sparse modeling approach that is evaluated using cross-validation. We develop ProbFold using existing SHAPE data (Deigan *et al.*, 2009) combined with an extensive set of known RNA structures (Rivas *et al.*, 2012) and evaluate a hierarchy of increasingly parameter-rich models. We find that including both base-pair stacking interactions and neighbor correlations for the SHAPE data improve performance. We also show how multiple types of probing data can be included in the models and may improve prediction performance. We find that the ProbFold approach exploits the probing data well, generally yielding higher performance gains than other methods, and present automatic procedures for optimizing the models on new data types.

2 Materials and methods

2.1 SCFGs

SCFGs are the probabilistic variants of Context-Free-Grammars (CFGs). A CFG defines a formal language used for the generation of

strings and is particularly suitable for RNA modeling, as it has the ability to capture nested long-range correlations (Dowel and Eddy, 2004). This approach has been widely applied in the context of RNA modeling and structure analysis (Durbin *et al.*, 1998; Eddy and Durbin, 1994; Knudsen and Hein, 1999; Pedersen *et al.*, 2004, 2006; Rivas *et al.*, 2012; Sakakibara *et al.*, 1994). An introduction to the use of SCFGs for RNA structure modelling and how they can be extended to handle the multivariate data of our setting is given in the supplementary material and methods section (Section S1.1).

For the current study, we use (and extend) the pfold grammar (Knudsen and Hein, 1999), which models RNA secondary structures in terms of individual base pairs and unpaired nucleotides through the set of grammar rules: $S \rightarrow LS|L$; $F \rightarrow bF\hat{b}|LS$; $L \rightarrow a|cF\hat{c}$, where S, L and F are the nonterminals, a, b and c refer to the terminal symbols. However, the grammar does not explicitly model stacking interactions between consecutive base pairs, as done in nearest-neighbor energy models (Mathews *et al.*, 1999, 2004; Xia *et al.*, 1998). Apart from hydrogen bonding between paired nucleotides (base pairing), stacking interactions between adjacent base pairs are the largest contributors to helix stability in nucleic acids (Yakovchuk *et al.*, 2006). We model these interactions for consecutive base-pairs by replacing a pair-emitting rule with a stack-emitting rule in the PFold grammar ($L \rightarrow cF\hat{c}$), which takes the previous base-pair into account.

The resulting grammar has six production rules, three of which emit terminals (Fig. 1). The probability of the terminals given the transition between nonterminals is specified by emission models. When modeling only the RNA sequence, the terminals consist of nucleotides and the emission distributions can be defined simply by multinomials (Durbin *et al.*, 1998; Pedersen *et al.*, 2004). To also model probing data, the emission models instead specify a joint distribution over both RNA sequence data and probing data. To achieve the flexibility needed for specifying joint distributions over multiple, potentially heterogeneous data types, we use a probabilistic graphical model framework to define the emission models.

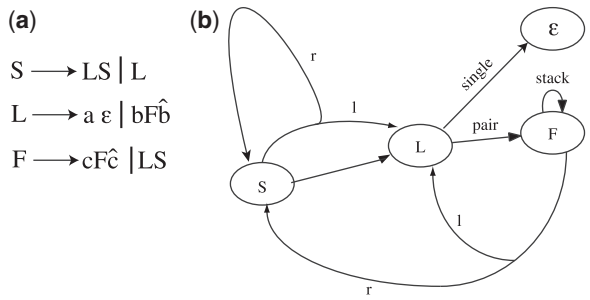


Fig. 1. (**a**) Grammar rules (see (b) for variable definitions). (**b**) Pictorial representation of the stacking grammar. The grammar has six production rules involving the four non-terminals S, L, F and ϵ. ϵ denotes the empty sequence, which cannot be derived further. Three of the rules emit terminals, named *single*, *pair* and *stack* and three are non-emitting, including two bifurcation rules. Each bifurcation rule splits into two parts, consisting of a left (l) non-terminal and right (r) non-terminal. The derivation starts in S. S can use either a bifurcation rule, which transits to L (l-part) as well as back to itself (r-part), or a non-emitting rule, which transits to L. L can use either the *single* emitting rule, which transits to ϵ and emits unpaired terminals (a), or use the *pair* rule, which transits to F and emits paired terminals ($a\hat{a}$). Finally, F can use the *stack* emitting rule, which transits back to F and emits (stacked) paired terminals ($b\hat{b}$) dependent on the previous base pair, or a bifurcation rule, which transits to L (l-part) as well as to S (r-part)

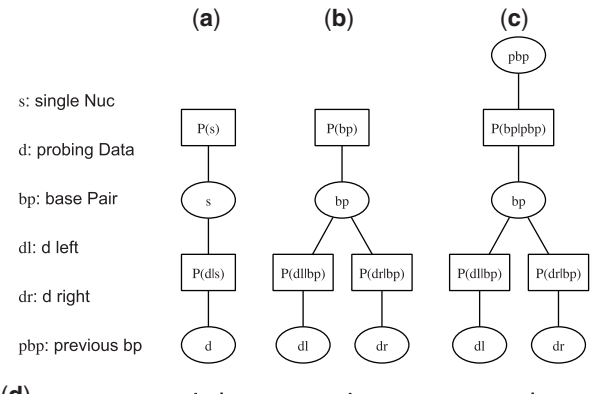
2.2 Emission distributions and probabilistic graphical models

Probabilistic graphical models (PGMs) offer a coherent and expressive framework for specifying and analyzing joint probability distributions (Koller and Fridman, 2009). PGMs are generally used to capture independence assumptions among a set of random variables and to specify their joint distribution as a factorization of local distributions each defined over subsets of variables. PGMs can be represented by mathematical graphs with nodes denoting random variables and edges denoting potential dependencies. A rich set of algorithms exist for doing inference with PGMs, which have proven a powerful tool for simplifying complex problems (Koller and Fridman, 2009).

We define the emission models as PGMs using the factor graph formalism (Fig. 2) (Bishop, 2006). In this formalism, the PGMs are specified as undirected bipartite graphs between random variable nodes (represented by circles) and factor nodes (represented by squares). The factors hold potentially unnormalized probability distributions involving neighboring random variables.

Our current factor graph implementation of PGMs only handles discrete random variables, which simplifies the implementation and speeds up likelihood calculations. Including continuous random variables generally requires potentially slow numeric integration. Probing data are therefore discretized in a preprocessing step (see below).

For a start, ProbFold was developed to take an RNA sequence with a single affiliated sequence of probing data values as input. For each sequence position we thus have observed both a nucleotide and a discretized probing data value. We need to define an emission distribution for each of the *single*, *pair* and *stack* emitting rules (Fig. 1). *Single* models only a single sequence position; *pair* models a pair of



(**d**)

	single	pair	stack
Joint prob.	P(s,d)	P(bp,dl,dr)	P(bp,pbp,dl,dr)
Factorization	P(s)P(d|s)	P(bp)P(dl|bp) P(dr|bp)	P(dl|bp)P(dr|bp) P(bp|pbp)
RNA Probing data	ACGUCACGU 001232454	ACGUCACGU 001232454	ACGUCACGU 001232454
Mapping	s←A, d←0	bp← AU dl← 0, dr← 4	bp←CG , pbp←AU dl←0 , dr ←5

Fig. 2. Probabilistic graphical models defining the (**a**) *single*, (**b**) *pair* and (**c**) *stack* emission models. The PGMs are shown as (bipartite) factor graphs, with variable nodes (circles) connected to factors (squares) defining local probability distribution. The variable abbreviations are given to the left. (**d**) For each emission model, the table gives (i) the joint probability distribution; (ii) its factorization specified by the PGM; (iii) example of short input data sequence with potential input positions highlighted. Note that the probing data has been discretized into six bins (0–5); (iv) mapping of data from highlighted sequence positions to relevant random variables of PGM

sequence positions; and *stack* models four sequence positions, consisting of two consecutive pairs (Fig. 2d).

For each of the three emission models, we specify a PGM defining a joint distribution over the relevant nucleotides and probing data values (Fig. 2). Initially, the *stack* model disregards the probing data of the previous base-pair. The PGM specification should reflect the independence structure of the modeled variables. We let the probing data at a position optionally depend on the observed nucleotides of that position. However, we let the probing data from the two sides of a base pair be independent of each other, given the observed nucleotides of the base pair (Fig. 2). These independence assumptions are evaluated separately as part of the model development below.

2.3 Datasets

The ProbFold models are potentially parameter rich. Optimally we would therefore train and test them on comprehensive sets of known RNA structures encompassing tens of thousands of base-pairs affiliated with consistently generated probing data. As such datasets do not yet exist, we complemented structure sets that include probing data with larger sequence-only sets.

Our primary structure probing set consisted of SHAPE data from *E.coli* 16S and 23S rRNAs (Deigan *et al.*, 2009; Weeks, 2012) augmented with a set of seven small RNA structures that were downloaded from the RMDB repository (Cordero *et al.*, 2012b) (RMDB set) and four taken from Rice *et al.* (2014) (see Table 1). Altogether, these include a total of 2142 unpaired positions and 1479 base pairs. The SHAPE data was preprocessed by denoting all invalid values (reactivity < 0; $n = 486$) as missing data.

For a subset of the small structures from RMDB ($n = 6$), DMS (dimethyl sulphate) and CMCT (1-cyclohexyl-(2-morpholinoethyl) carbodiimide metho-p-toluene) probing datasets were also available (Cordero *et al.*, 2012b) (RMDB set, Supplementary Table S12). These structures encompass 287 base pairs and 593 unpaired positions, for which 302 positions had missing data for DMS and 415 for CMCT. We used these to illustrate the use of ProbFold on multiple types of probing data.

In addition to the above datasets based on Sanger sequencing, we also included a probing dataset based on Next-generation sequencing (NGS). A set of eight small RNA structures with SHAPE-Seq probing data was downloaded from RMDB (Supplementary Table S11) (Lucks *et al.* 2011). In combination, these include a total of 280 base pairs and 490 unpaired positions.

The probing datasets were complemented with sets of sequence-only structures called TrainSet A (3166 structures) and TestSet A ($n = 697$), which were originally compiled by Rivas *et al.* (2012). We preprocess these sets by discarding structures with loops with less than three bases, reducing the size of TrainSet A to 2707 (130 227 base pairs) and TestSet A to 593 (25 596 base pairs). In the supplementary methods and material (Section S1.4), we describe training, testing and prediction procedures adopted in this work.

2.4 Probing data modeling and discretization

The value of probing data for secondary structure prediction depends on the difference between its distribution in single stranded and paired regions. In ProbFold, these probing data distributions (P^{single} and P^{pair}) are explicitly modeled as part of the PGMs and may be conditioned on the primary sequence. Given our use of discrete PGMs, we discretize the probing data into k bins and use normalized histogram models (i.e. multinomials), which use $k - 1$ free parameters. Initially we visualized these distributions for 16S and 23S SHAPE data using 15 equi-distant bins (Supplementary Fig. S1a). We discuss probing data modelling and discretization including optimal criterion for break points based on Kullback–Leibler (KL) divergence in the supplementary methods and materials (Section S1.3).

2.5 Cross-validation and overfitting

Unlike some of the previous approaches (Sükösd *et al.*, 2012; Washietl *et al.*, 2012), we use cross-validation in our performance evaluation (Section S1.2). We thereby avoid to train and test on the same probing data. For instance, when evaluating the performance on 16S, we train on 23S combined with RMDB and the sequence-only TrainSet A. This further limits the number of probing data related free parameters that can be learned without overfitting during development.

The number of free parameters (fp) in the probing data models is proportional to the number of bins used in the discretization. When the number of free parameters is increased, the models typically learn the properties of the training data well, but generalize poorly to other datasets. To select the optimal number of bins and to illustrate this behavior, we plot the prediction performance on both train

Table 1. Prediction performance of ProbFold and other methods on set of small RNA structures with SHAPE data (Cordero *et al.*, 2012b; Rice *et al.*, 2014)

	ProbFold		PPFold		RNAstructure		GTFold		RNAfold.zar	
Structures	*F*-Value	Δ*F*	*F*-Value	Δ*F*	*F*-Value	Δ*F*	*F*-Value	Δ*F*	*F*-Value	Δ*F*
5S RNA (Cordero *et al.*, 2012b)	0.54	0.24	0.57	0.22	0.24	0.00	0.25	0.00	**0.97**	**0.74**
Adenine riboswitch (Cordero *et al.*, 2012b)	**1.00**	**1.00**	0.96	0.51	0.96	−0.05	**1.00**	0.00	**1.00**	0.00
cidGMP riboswitch (Cordero *et al.*, 2012b)	0.63	0.14	0.55	0.21	0.73	−0.15	0.70	−0.01	**0.78**	0.08
Glycine (Cordero *et al.*, 2012b)	0.76	0.22	0.65	0.47	**0.88**	0.23	0.85	0.21	0.87	**0.48**
P4–P6 domain (Tetrahymena ribozyme) (Cordero *et al.*, 2012b)	0.87	**0.37**	0.80	0.30	**0.88**	0.01	0.84	0.07	0.73	−0.09
Ribonuclease (Cordero *et al.*, 2012b)	**0.79**	**0.11**	0.20	−0.49	0.57	−0.01	0.39	−0.39	0.64	0.01
tRNA phenylalanine (yeast) (Cordero *et al.*, 2012b)	**0.98**	**0.79**	0.44	0.12	**0.98**	0.03	0.95	0.71	**0.98**	0.00
M-Box riboswitch (Rice *et al.*, 2014)	0.71	−0.10	0.47	−0.37	0.52	**0.04**	0.71	−0.18	**0.89**	0.00
Lysine riboswitch (Rice *et al.*, 2014)	**0.28**	−0.03	0.22	−0.06	0.28	**0.06**	0.26	−0.00	0.25	0.02
Group I Intron, *T. thermpphila* (Rice *et al.*, 2014)	**0.79**	**0.23**	0.66	0.03	0.78	0.10	0.75	0.16	0.78	0.14
Group II Intron, *O. iheyensis* (Rice *et al.*, 2014)	0.51	**0.27**	0.53	0.23	**0.60**	−0.07	0.59	−0.02	0.59	−0.08
Average	0.71	**0.29**	0.55	0.11	0.67	0.02	0.66	0.05	**0.77**	0.12

Both the *F*-value and the change in *F*-value (Δ*F*) relative to the sequence-only (Seq-only) predictions are shown. The maximal performance score for each structure is denoted in bold. The Zarringhalam *et al.* (2012) SHAPE conversion approach was used for RNAfold (RNAfold.zar). See Supplementary Tables S3–S10 for the full set of performance statistics.

and test sets when using probing data from 16S and 23S (Supplementary Fig. S1d). The test performance is optimal between 3 and 15 bins, whereas train performance continues to increase and approaches perfection due to overfitting. Based on this, we use six bins for the discretization in the model evaluation below, which is the only fixed metaparameter.

3 Results

3.1 Model selection

We developed and evaluated a hierarchy of models capturing different correlations in the data with increasing number of free parameters (Supplementary Table S1). The limited amounts of training data enforces a tradeoff as models with too many free parameters will be overfitted and not be robust.

3.2 Sequence-only models

We started out with two sequence-only models: The *p*air model uses the original PFold grammar (Knudsen and Hein, 1999) and is specified by 18 free parameters (fp). The *s*tack model uses the above described grammar extension, which also includes stacking interactions (fp = 258). The *s*tack model showed a modest performance gain over the *p*air model in the ROC analysis and by *F*-measure (Fig. 3a).

3.3 Probing data models

To extend the sequence-only models to also handle probing data, we developed emission models that generate both sequence data and SHAPE reactivities. To guide the development of these, we started by analyzing correlations in the 16S and 23S SHAPE datasets.

We first evaluated if the SHAPE reactivities were correlated with the primary sequence nucleotides. To control for compositional biases and the different level of reactivities, we did this separately for single (unpaired) and paired regions (Fig. 4a,b). In both cases, there were significant differences in the distribution of the SHAPE values for the different nucleotides ($P < 4.4e{-}03$ for *single* and $P < 8.6e{-}06$ for *pair*; Kruskal-Wallis rank sum test).

We then evaluated if the SHAPE reactivities were correlated between the left and the right side of a base-pair (Fig. 4c). Surprisingly no correlation was observed (Pearson correlation coefficient, pcc=-0.042; *P*-value = 0.075). This may be explained by the overall low SHAPE reactivities of paired bases, causing experimental noise to dominate any underlying signal.

Based on these observations, we defined emission models where the SHAPE reactivities of the left and the right side of a base pair are modeled independently, but with each their own distribution.

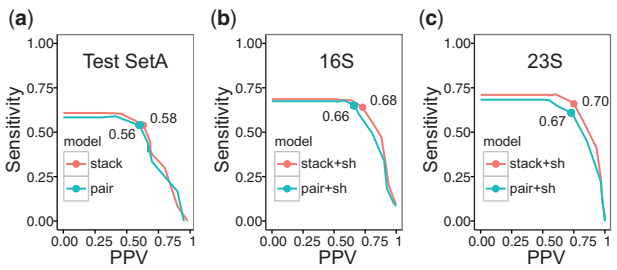


Fig. 3. ROC curves and maximal *F*-values for (**a**) sequence-only models on TestSet A, (**b**) probing data models on *E.* coli 16S rRNA and (**c**) probing data models on *E.* coli 23S. The curves are made by varying the value of γ (see text)

Altogether, the 25 free parameters are used to model the SHAPE reactivities given discretization in six bins (*single* model: fp = 5; *pair* models: fp = 2 × 5 = 10; *stack* models: fp = 2 × 5 = 10). Preferably, ProbFold should capture differences in SHAPE reactivities among nucleotides. However, capturing these dependencies requires many additional free parameters (fp = 4 × 25 – 25 = 75). Given the limited training data, we therefore model the primary sequence and the SHAPE reactivities independently (e.g. the *single* model becomes $P(s,d) = P(s) \times P(d)$, as shown in Supplementary Fig. S2).

These emission models were combined with the grammar of the sequence-only *s*tack model selected above to give the *s*tack+sh (fp = 283) model for SHAPE data. The inclusion of SHAPE data dramatically improved sensitivity and overall performance (Fig. 3b,c and Supplementary Table S2)

We finally evaluated if SHAPE reactivities correlate with neighboring positions, again analyzing single and paired regions separately (Fig. 4d,e). Significant positive correlations were observed in both cases ($P < 0.0001$), with an overall higher correlation in single regions (pcc = 0.559) than in paired regions (pcc = 0.397). The correlations may reflect overall steric constraints, which are likely to be correlated along the primary sequence. For instance, backbone flexibility of loop positions may decrease toward stems.

We extended the emission models of the *stack+sh* model to capture these sequential correlations in the SHAPE data (Supplementary Fig. S1), which requires many additional free parameters (fp = 125). The resulting model, *stack+sh+cor* (fp = 408), improves performance over simpler models when trained on the 23S dataset, but not when trained on the smaller 16S dataset. We attribute the decrease in performance on 23S (842 base pairs) to overfitting of the parameter rich correlation model when trained on the 16S dataset (468 base pairs). Overall, we recommend the *stack+sh+cor* model for SHAPE prediction applications. We make a version trained on the combined 16S and 23S datasets available for download together with the other ProbFold models (http://moma.ki.au.dk/prj/probfold/).

We also evaluated the performance of four existing methods, PPfold 3.0 (Sükösd *et al.*, 2012), RNAstructure v5.6 (Mathews *et al.*, 2004), GTFold-3.0 (Swenson *et al.*, 2012) and RNAfold (Lorenz *et al.*, 2011, 2015), with the SHAPE conversion model from Zarringhalam *et al.* (2012) (hereafter termed RNAfold.zar), on the 16S and 23S sequences with and without SHAPE data (Supplementary Table S2).

For the sequence-only predictions, the ProbFold *stack* model results in a low sensitivity (<0.30) and a high PPV (~0.80). This pattern is shared by ppfold, though it has somewhat poorer performance, which may be explained by it not modeling stack interactions. RNAstructure, GTFold and RNAfold, which all have richer structure models, have more balanced sensitivity and PPV performance. As a result they have higher *F*-values for 23S and, in some cases, also for 16S. ProbFold *stack* has the highest accuracy (ACC) on both sequences.

When including SHAPE data, ProbFold continues to have low sensitivity (47–62%) and high PPV (76–90%) compared to the other methods. However, ProbFold's tradeoff between sensitivity and PPV can be adjusted (Fig. 3). RNAstructure has the highest overall performance both by accuracy and *F*-value, closely followed by RNAfold.zar and also GTFold in the case of the 16S sequence. However, the 16S and 23S SHAPE datasets used in the development of ProbFold, have also been heavily used for developing the SHAPE models of the other methods. Specifically, PPfold was trained on both 16S and 23S and the RNAstructure parameters were chosen based on parameter-grid analysis of 23S with performance

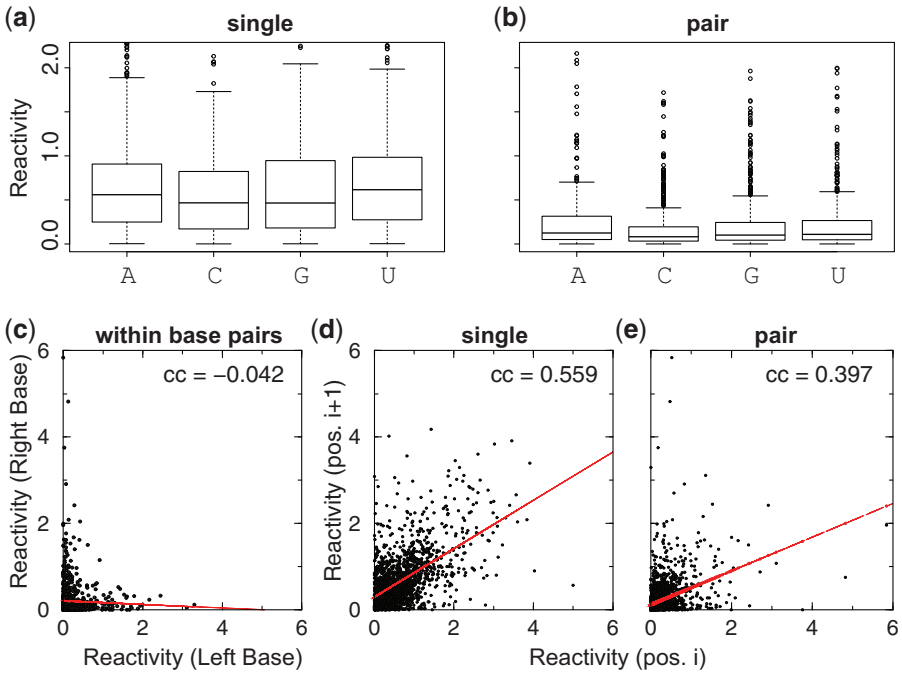


Fig. 4. Correlations in SHAPE data. Box-plots showing distribution of SHAPE reactivities for individual nucleotides for (**a**) single (unpaired) and (**b**) paired regions. Scatterplots showing (**c**) lack of correlation between left and right side of base pairs; (**d**) positive correlation along the sequence for both unpaired bases and (**e**) positive correlation along the sequence for paired bases in stems. The regression line (red) summarizes the trend in the data (Color version of this figure is available at *Bioinformatics* online.)

evaluation on 16S (information not available for GTfold and RNAfold.zar; Supplementary Table S2).

This could lead to overfitting and performance statistics that do not generalize. In contrast, the performance results for ProbFold are based on cross-evaluation, to avoid the effect of overfitting to the train dataset (Supplementary Fig. S1d). The performance results are therefore not directly comparable between methods.

As an independent test dataset, we evaluate the performance on a set of small RNA structures with SHAPE data (Table 1 and Supplementary Tables S3–S10). On this set, ProbFold achieves the highest *F*-value and accuracy on five of the eleven structures. It ranks second after RNAfold.zar in terms of overall *F*-value and accuracy across all structures (Table 1 and Supplementary Table S9).

Since our main focus is ProbFold's ability to exploit the probing data, we also evaluated the relative gain in performance when including probing data over sequence-only predictions (Table 1 and Supplementary Tables S6, S8 and S10). ProbFold showed the highest gain for five of the eleven structures, with an average *F*-value gain of 0.29. The other methods showed smaller relative gains, with PPfold at 0.11, GTFold at 0.05, RNAstructure at 0.02, and RNAfold.zar at 0.12.

We also evaluated ProbFold on an NGS-based SHAPE-Seq dataset (Lucks *et al.*, 2011), again using cross-evaluation (Supplementary Table S11). The performance is comparable to that of the capillary SHAPE data, if slightly better, with an overall accuracy of 0.76. Again, the gain from sequence-only performance is substantial (Δaccuracy of 0.29).

3.4 Robustness to noise

Experimental probing data will always include some amount of noise. When ProbFold is trained, it in effect learns the correlation between probing data levels and structure elements. If there is much noise, it should learn that the correlation is weak and that the probing data holds little information.

To evaluate robustness to noise, we simulated noisy versions of the 16S and 23S SHAPE datasets (Deigan *et al.*, 2009; Weeks, 2012). Scaled random noise was drawn from a standard normal distribution and added to the observed SHAPE value at every position. For each of a range of noise strengths, 30 datasets were generated and the structure prediction performance evaluated for both ProbFold as well as RNAfold.zar (with scoring by Zarringhalam *et al.*, 2012), RNAstructure and PPFold (Supplementary Fig. S3). ProbFold was evaluated using cross-evaluation on each simulated dataset and hence retrained on the structure not used for evaluation.

Overall, ProbFold shows greater robustness to an increase in noise levels than the other methods. In particular, the fine-grained SHAPE distributions of PPFold appear sensitive to noise, with a fast decay in performance. At high noise levels, ProbFold performance approaches the sequence-only level, as could be expected. This demonstrates the potential advantage of retraining models and learning the signal strength of individual datasets.

3.5 ProbFold with other data types

To illustrate the use of ProbFold on other types of probing data, we retrained and applied the model on publicly available DMS and CMCT datasets covering six small RNA structures (see Section 2.3). We used leave-one-out cross evaluation to train and evaluate the performance of the stack+sh model from above. Given the limited amount of training data, we use the KL approach to discretize the probing data into three bins only, which reduces the number of free parameters used to model probing data from 15 to 6. For both types of data, the overall performance improved compared to using only the primary sequence (Supplementary Tables S12, S13). Overall, using DMS resulted in better prediction performance than CMCT (*F*-values of 0.54 versus 0.48). The lower power of CMCT

compared to DMS is also apparent from the smaller separation between the probing signal intensity distributions for paired and unpaired positions (Supplementary Fig. S4).

As both DMS and CMCT probing have known strong nucleotide dependencies, we also evaluated a version of the stack+sh model where the probing signal distributions depend on the sequence nucleotides (as shown in Fig. 2). For DMS the separation of the paired and unpaired signal distributions conditional on sequence nucleotide appear to improve slightly (Supplementary Fig. S5). Whereas for CMCT no improvement is obvious (Supplementary Fig. S6). However, the prediction performance decreased for both models in cross-evaluation (*F*-values of 0.40 for DMS and 0.37 for CMCT), likely due to the limited training data and many additional free parameters ($n = 18$) introduced in this model even when discretized into three bins only.

3.6 Modeling multiple data types

When available, integration of multiple probing data types should increase prediction accuracy. We here show how ProbFold's emission models can be extended to handle multiple data types, using the *single* model as an example (Fig. 5a). Based on the emission models of Figure 2, we suggest to model multiple types of probing data (d_1, d_2, d_3, d_4 and d_5) as independent given the nucleotide of the primary sequence. For the *single* model, the joint distribution thus becomes $P(s, d_1, d_2, d_3, d_4, d_5) = P(s) \times \prod_{i=1}^{5} P(d_i|s)$. The other emission models can be simply defined following the same scheme.

To demonstrate the advantage of using multiple datasets, we performed a proof-of-principle experiment based on boot-strapped data, given the limited extent of existing real datasets for well annotated structures (see Section S1.5 for details). Performance improved when adding additional datasets, with the largest performance gains for the first two sets and gradually smaller gains for the following (Supplementary Fig. S7).

To further demonstrate the benefit of combining multiple probing datasets and to also include different data types, we applied the multiple-data-set model to the six previously described RMDB structures for which CMCT, DMS and SHAPE are all available (Cordero *et al.*, 2012b). Given the limited extent of the available datasets, each is modeled using only three bins, with KL optimized break points (Section S1.3), to retain the number of model parameters. The experiment is thus still at the proof-of-principle level. As more data become available, more bins can be used to capture the structure signal of the probing data, which is expected to improve performance. The performance was measured using leave-one-out cross evaluation and averaged across all six structures (Table 2 and Supplementary Table S14).

Both sensitivity and the overall performance as measured by accuracy and F-value increase when incrementally adding each of the three probing datasets, starting with sequence-only (Fig. 5b). A slight decrease of PPV is observed when adding DMS. Integration of SHAPE, DMS and CMCT data have previously been carried out using RNAstructure v. 5.3 and pseudo-energy terms (Cordero *et al.*, 2012a). For comparison, we evaluate the performance of the multi-data version of RNAstructure on RMDB structures (Table 2 and Supplementary Table S14). While RNAstructure performs much better on the sequence-only dataset and achieves the highest overall *F*-values, ProbFold shows the largest relative gains from including probing datasets. ProbFold also show consistent gains with each added dataset, which is not the case for the RNAstructure model (Table 2). This suggests that the ProbFold emission models are able to make good use of the available structure signal. The emission

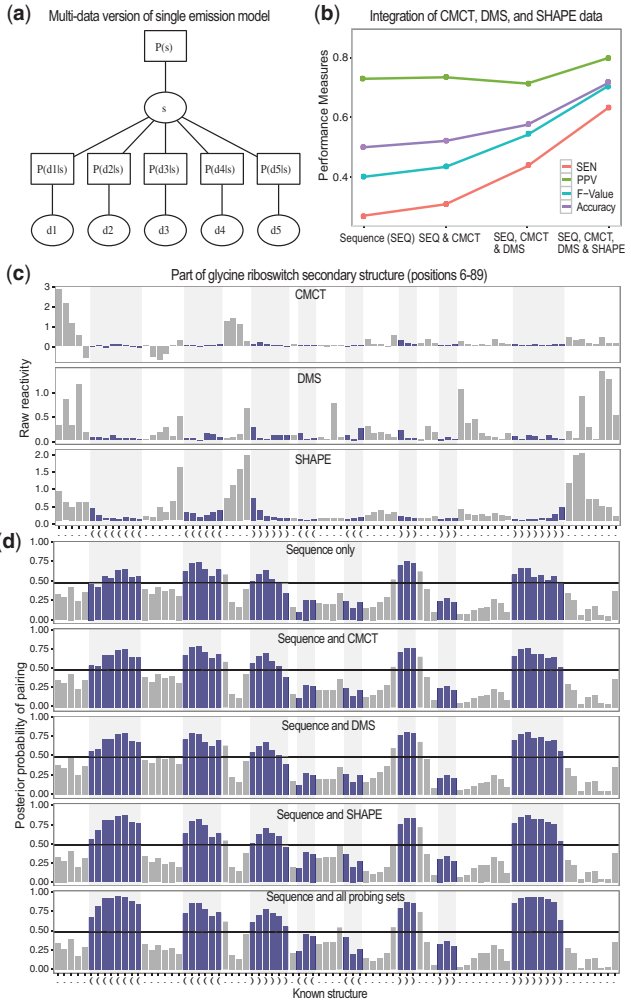


Fig. 5. Multi-data version of ProbFold. (**a**) PGM for *single* emission distribution that integrates five types of probing data, d_1, d_2, d_3, d_4 and d_5. $P(s)$ is the prior distribution of sequence data, $P(d_i|s)$, $i = 1..5$, are the conditional distributions of the different types of probing data given the sequence data. (**b**) Prediction performance of the multi-data version of ProbFold with incrementally increasing number of probing data types (CMCT, DMS and SHAPE) on six small RNA structures. Performance evaluated using cross evaluation with same measures as above. (**c**) Raw probing data reactivities for CMCT, DMS and SHAPE shown for a part of the Glycine riboswitch together with the known structure. (**d**) Positional pairing probabilities under the ProbFold model given either only the primary sequence; sequence combined with a single probing data type; or sequence and all three probing data types in combination. The pairing probability is defined as the posterior probability of pairing with any other base, which equals one minus the posterior probability of being unpaired

models can be further extended to account for dependencies both within and among multiple datasets (Section S1.6).

To further illustrate the signal contribution of the individual probing data types, we plotted the raw activities (Fig. 5c) (Cordero *et al.*, 2012b) as well as the ProbFold pairing probabilities (Fig. 5c, Supplementary Figs S9–S14) against the known RNA secondary structures. The reactivities of the different data types generally correlate well, with a few exceptions for CMCT (Fig. 5c). The SHAPE reactivities are more auto-correlated than those of CMCT and DMS, with less local variance. As shown by the above performance evaluations, all data types contribute useful structure signal for ProbFold, with an increase in pairing probabilities for paired regions (Fig. 5d). In effect,

Table 2. Average performance on six small structural RNAs of the Multi-data versions of ProbFold and RNAstructure (Mathews *et al.*, 2004) with step-wise inclusion of CMCT, DMS and SHAPE structure probing data (Cordero *et al.*, 2012a,b)

	ProbFold		RNAstructure	
Data	*F*-Value	Δ*F*	*F*-Value	Δ*F*
Seq-only	0.40	0.00	**0.73**	0.00
Seq, CMCT	0.48	0.08	**0.85**	**0.12**
Seq, CMCT, DMS	0.54	**0.14**	**0.85**	0.12
Seq, CMCT, DMS, SHAPE	0.71	**0.31**	**0.82**	0.09

Both the *F*-value and the change in *F*-value (Δ*F*) relative to the sequence-only (seq-only) predictions are shown. The maximal performance score for each combination of probing data types is denoted in bold. See Supplementary Table S14 for the full set of performance statistics.

ProbFold thus averages out the noise in the data given the structure constraints imposed by the primary sequence. Though SHAPE generally contribute the largest individual gains in pairing probabilities, the combination of all three sets further adds to the difference between the paired and unpaired reigons (Supplementary Figs S9–14).

4 Discussion

We have presented a probabilistic method for RNA secondary structure prediction that integrates experimental structure probing data. One of the virtues of our approach is its flexibility. The underlying model was initially developed on a capillary electrophoresis SHAPE dataset, but it can readily be retrained and applied on other data types given sufficient training data as well as extended to handle multiple data types. We demonstrated these extensions with proof-of-concept examples on existing datasets and on generated (bootstrapped) data. We developed and trained versions of ProbFold for SHAPE, SHAPE-seq, DMS and CMCT probing data – both individually and in combination. We also evaluated different variants of these models, for instance including nucleotide dependencies. The flexibility is achieved by the use of a highly modular probabilistic model with accompanying efficient algorithms for training and prediction (Bishop, 2006; Durbin *et al.*, 1998; Eddy, 2014; Eddy and Durbin, 1994; Knudsen and Hein, 1999, 2003; Koller and Fridman, 2009; Nawrocki and Eddy, 2013; Pedersen *et al.*, 2004, 2006; Rivas and Eddy, 2001; Rivas *et al.*, 2012; Sakakibara *et al.*, 1994).

As with most other probabilistic RNA secondary structure prediction methods (Rivas and Eddy, 2000; Metzler and Neble, 2008), we use stochastic context-free grammars to model the secondary structure. We find limited benefit of modeling stacking interactions, as have others (Rivas *et al.*, 2012). This may be because individual base-pair parameters already account for most of the stacking interaction effects (Yakovchuk *et al.*, 2006). We factorize the grammar rules into transitions and emissions. In contrast to other methods, we explicitly specify the emission distributions using probabilistic graphical models (PGMs), which may be defined over multi-variate input data. The implementation closely reflects this modularity, with separate textual specifications of the overall grammar, the transitions, the PGMs defining the emission distributions; and the mapping of observed data to PGM variables.

As part of the model development, we evaluated correlations in the probing data. We found modeling SHAPE probing data correlations along the sequence improved performance, given enough training data was present. In general, detecting and modeling the prominent dependencies in the observed data is expected to improve the fit and the discriminatory power of the emission distributions and hence overall model performance. However, including correlations makes the models more complex with many additional free parameters to learn. Given the limited size of probing datasets for training, overfitting and lack of robustness easily becomes a problem, as shown for parameter-rich versions of ProbFold (e.g. Supplementary Fig. S1d and Supplementary Table S2).

ProbFold bins the continuous probing data values to allow for the use of discrete PGMs. This simplifies the PGM implementation, for instance by avoiding computationally expensive numerical integration, and avoids use of analytical continuous distributions with a potentially poor fit. As the structural signal of the probing data depends on the differences in its distribution in different structural regions, both the number of bins and their boundaries are important parameters of the model. We show that the number of bins should be kept small given the amount of available training data to avoid overfitting. We suggest to select bin boundaries by optimizing the difference between the *single* and *pair* probing data distribution using KL divergence.

As an alternative to discretizing the data, parameter free continuous distributions, such as kernel distributions, could be used. These however easily become computationally heavy, as they in principle are specified by the full training dataset. Given knowledge of the uncertainty of the probing data observations, a more satisfying approach would be to explicitly model the uncertainty of the individual probing data values. Such knowledge would be available, e.g. with counts from NGS-based probing data, as some of transcriptome-wide approaches produce (Kertesz *et al.*, 2010a,b; Lucks *et al.*, 2011).

The RNAstructure method converts SHAPE reactivities to pseudo energy change terms, which are incorporated when predicting the minimal free energy structure (Deigan *et al.*, 2009). The conversion is done using a simple linear parametric form, which only requires two free parameters. Using a simple parametric form limits the number of free parameters, but may also introduce bias if the fit is poor in part of the probing data value range. In particular, RNAstructure has been shown to perform extremely well on 16S rRNA when introducing several preprocessing steps and filters, such as (i) selecting parameters performing well on 23S; (ii) limiting the allowed distance between base pairs; (iii) focusing on sites with useful SHAPE data and (iv) disregarding sites with clear incompatibilities with the comparative reference structure (Deigan *et al.*, 2009). Though individually helpful, manually selecting parameters and introducing many preprocessing steps risk making the approach more liable to overfitting on a concrete dataset. Direct calculation of pseudo-energy change terms based on log-likelihood ratios of being paired versus unpaired has also been suggested (Cordero *et al.*, 2012a), which is closer to the approach taken by ProbFold.

ProbFold has been designed with the aim of extendibility to multi-variate probing data measurements. This could for instance be the combination of SHAPE with other chemical or enzymatic probing agents, as in the proof-of-concept example using CMCT, DMS, SHAPE data. If the noise in the individual measurements at a site are correlated it becomes important to capture these correlations in the model to retain specificity. Such correlations could for instance be caused by tertiary structure interactions involving single stranded regions, which may affect several types of probing agents, including SHAPE. Specifically, non-canonical base pairs will often give similar signal to canonically paired bases using SHAPE but not DMS. Learning such correlations would therefore be expected to improve prediction performance. As more and more RNA tertiary structures are found, one solution could be to explicitly include tertiary

structure aspects into the model (Kopeikin and Chena, 2005; Lorenz *et al.*, 2013).

In the case of SHAPE, we did not observe any correlation between paired bases. However, such correlations may well exist for other types of probing agents, such as double stranded RNases. Even partial evidence for the presence of specific base pairs could significantly improve performance by constraining and simplifying the folding problem. Such evidence would resemble the signal exploited from compensatory base pair substitutions exploited in comparative RNA structure analysis. As suggested by Sükösd *et al.* (2012), additional power could be gained by combining experimental probing data with comparative data, though this would be limited to functional and conserved RNA structures. However, exploring these extensions is beyond the scope of the current study.

Through proof-of-principle experiments, we have illustrated the applicability of ProbFold to different types of probing data and to multiple complementary datasets (Fig. 5). The performance evaluations show that ProbFold can make efficient and competitive use of the probing data, both for SHAPE data and when combining multiple datasets (Tables 1 and 2). Retraining aids its adaptation to varying noise levels. However, model performance is limited by the small size of the available training datasets, which restricts model complexity and hence predictive power.

We hope the advent of NGS-based high-throughput structure probing techniques, as pioneered by Kertesz *et al.* (2010a,b), Underwood *et al.* (2010) and Lucks *et al.* (2011), will result in large uniform probing datasets of known structures assessed by multiple probing agents. Concretely, inclusion of comprehesive sets of RNA transcripts with known structures as spike-ins could help achive this. This would allow multi-variate versions of ProbFold or similar models to be trained, with an expected boost in performance characteristics. Ultimately such approaches could help improve RNA structure maps transcriptome-wide.

Acknowledgements

We thank Kevin Weeks for sharing SHAPE Reactivities for *E. coli* 16S and 23S rRNAs and Zsuzsanna Sükösd, Jeppe Vinther and Lukasz Jan Kielpinski for fruitful discussions and manuscript comments. This work greatly benefited from discussions at the Benaque RNA workshop in 2012.

Funding

This work was supported by a Danish Strategic Research Council grant to Center for Computational and Applied Transcriptomics (COAT) [10-092320/DSF].

Conflict of Interest: none declared.

References

Bishop,M.C. (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Cordero,P. *et al.* (2012a) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry*, **51**, 7037.

Cordero,P. *et al.* (2012b) An RNA mapping database for curating RNA structure mapping experiments. doi:10.1093/bioinformatics/bts554.

Deigan,K.E. *et al.* (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. USA*, **106**, 97–102.

Dowel,R.D. and Eddy,S.R. (2004) Evaluation of several lightweight 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, United Kingdom.

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

Eddy,S.R. (2014) Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu. Rev. Biophys.*, **43**, 433–456.

Ehresmann,C. *et al.* (1987) Probing the structure of RNAs in solution. *Nucleic Acids Res.*, **15**, 9109–9128.

Karaduman,R. *et al.* (2006) RNA structure and RNA – protein interactions in purified yeast U6 snRNPs. *J. Mol. Biol.*, **356**, 1248–1262.

Kertesz,M. *et al.* (2010a) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103–107.

Kertesz,M. *et al.* (2010b) Probing RNA structure genome-wide using high throughput sequencing. *Protoc. Exchange*, doi:10.1038/nprot.2010.152.

Knudsen,B. and Hein,J.J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, **13**, 3423–3428.

Knudsen,B. and Hein,J.J. (1999) Using stochastic context free grammars and molecular evolution to predict RNA secondary structure. *Bioinformatics*, **15**, 446–454.

Koller,D. and Friedman,N. (2009) *Probabilistic Graphical Models*. MIT Press, Cambridge, MA, USA.

Kopeikin,Z. and Chena,S.-J. (2005) Statistical thermodynamics for chain molecules with simple RNA tertiary contacts. *J. Chem. Phys.*, **122**, 094909.

Lorenz,R. *et al.* (2013) 2d meet 4g: G-quadruplexes in rna secondary structure prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **99**, 1.

Lorenz,R. *et al.* (2015) SHAPE directed RNA folding. *Bioinformatics*, doi:10.1093/bioinformatics/btv523.

Lorenz,R. *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 1.

Lucks,J.B. *et al.* (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. USA*, **108**, 11063–11068.

Mathews,D.H. *et al.* (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Mathews,D.H. *et al.* (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *PNAS*, **101**, 7287–7292.

McGinnis,J.L. *et al.* (2012) The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc.*, **134**, 12319.

Merino,E.J. *et al.* (2005) RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.*, **127**, 4223–4231.

Metzler,D and Nebel,M.E. (2008) Predicting RNA secondary structures with pseudoknots by MCMC sampling. *J. Math. Biol.*, **56**, 2008.

Nawrocki,E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **15**, 2933.

Ouyang,Z. *et al.* (2013) SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res.*, **23**, 377–387.

Pedersen,J.S. *et al.* (2004) A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.*, **32**, 4925–4936.

Pedersen,J.S. *et al.* (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.

Quarrier,S. *et al.* (2010) Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA*, **16**, 1108–1117.

Regulski,E.E. and Breaker,R.R. (2008) In-line probing analysis of riboswitches. *Methods Mol. Biol.*, **419**, 53–67.

Rice,G.M. *et al.* (2014) RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA*, **20**, 846–854.

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

Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–606.

Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.

Sakakibara,Y. *et al*. (1994) Stochastic Context-Free Grammars for tRNA Modeling. *Nucleic Acids Res*., **22**, 5112–5120.

Sükösd,Z. *et al*. (2012) PPfold 3.0: Fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics*, **28**, 2691–2692.

Swenson,M.S. *et al*. (2012) GTfold: enabling parallel RNA secondary structure prediction on multi-core desktops. *BMC Res. Notes*, **5**, 341.

Tijerina,P. *et al*. (2007) DMS footprinting of structured RNAs and RNA-protein complexes. *Nat. Protoc*., **2**, 2608–2623.

Underwood,J.G. *et al*. (2010) FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods*, **7**, 995–1001.

Washietl,S. *et al*. (2012) RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res*., **40**, 4261–4272.

Wan,Y. *et al*. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet*., **12**, 641–655.

Weeks,K.M. (2012) 16S ans 23S *E. coli* data, Personal communication.

Weeks,K.M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struct. Biol*., **20**, 295–304.

Xia,T. *et al*. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*, **37**, 14719–14735.

Yakovchuk,P. *et al*. (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res*., **34**, 564.

Zarringhalam, K. *et al*. (2012) Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS ONE*, **7**, e45160.