BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 20 2012, pages 2691–2692
doi:10.1093/bioinformatics/bts488

Structural bioinformatics

Advance Access publication August 9, 2012

PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data

Zsuzsanna Sükösd[1,2,3,*], Bjarne Knudsen[4], Jørgen Kjems[1,2] and Christian N.S. Pedersen[3,5]

[1]Interdisciplinary Nanoscience Center, Aarhus University, Ny Munkegade 120 and [2]Department of Molecular Biology and Genetics, Aarhus University, C. F. Møllers Alle 3 and [3]Bioinformatics Research Center, Aarhus University, C. F. Møllers Alle 8, DK-8000 Aarhus C, Denmark, [4]CLC bio, Finlandsgade 10-12 and [5]Department of Computer Science, Aarhus University, Aabogade 34, DK-8200 Aarhus N, Denmark

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: PPfold is a multi-threaded implementation of the *Pfold* algorithm for RNA secondary structure prediction. Here we present a new version of PPfold, which extends the evolutionary analysis with a flexible probabilistic model for incorporating auxiliary data, such as data from structure probing experiments. Our tests show that the accuracy of single-sequence secondary structure prediction using experimental data in PPfold 3.0 is comparable to *RNAstructure*. Furthermore, alignment structure prediction quality is improved even further by the addition of experimental data. PPfold 3.0 therefore has the potential of producing more accurate predictions than it was previously possible.

Availability and implementation: PPfold 3.0 is available as a platform-independent Java application and can be downloaded from http://birc.au.dk/software/ppfold.

Contact: Zsuzsanna Sükösd or zs@birc.au.dk.

Received on June 11, 2012; revised on July 6, 2012; accepted on July 27, 2012

1 INTRODUCTION

The *Pfold* package predicts (non-pseudoknotted) RNA secondary structure by combining a stochastic context-free grammar (SCFG) with an evolutionary model (Knudsen and Hein, 1999, 2003). PPfold is a recent multi-threaded re-implementation of *Pfold* (Sükösd *et al.*, 2011). The *Pfold* model has been shown to result in highly accurate predictions when the input alignment is of high quality (Gardner and Giegerich, 2004). In a different approach, data from high-throughput, quantitative RNA structure probing methods have also recently been used in thermodynamic prediction methods to increase prediction accuracy (Deigan *et al.*, 2009; Washietl *et al.*, 2012). However, phylogenetic and probing data have only been used independently in RNA secondary structure prediction so far. Here, we present PPfold 3.0, which integrates these different sources of information at the level of the model. This is expected to increase prediction accuracy beyond what is possible with either evolutionary information or experimental data alone.

*To whom correspondence should be addressed.

2 MODEL

Pfold combines two models: (1) a SCFG model M_s, which generates the prior probability distribution over secondary structures σ. The prior probabilities are denoted as $P(\sigma|M_s)$. (2) A phylogenetic model M_t, which computes the likelihood of the input alignment D, given the secondary structures. The likelihoods are denoted as $P(D|\sigma, M_t)$.

If an additional set of experimentally observed data, H, is available, the posterior probability of a secondary structure is

$$P(\sigma|D, H, M_s, M_t) = \frac{P(D, H|\sigma, M_s, M_t)P(\sigma|M_s, M_t)}{P(D, H|M_s, M_t)} . \quad (1)$$

Expanding the expression and removing dependencies,

$$P(\sigma|D, H, M_s, M_t) = \frac{P(H|D, \sigma)P(\sigma|M_s)P(D|\sigma, M_t)}{P(D, H|M_s, M_t)} , \quad (2)$$

where $P(D|\sigma, M_t)$ are the probabilities obtained from the phylogenetic part of the algorithm, and $P(\sigma|M_s)$ are the prior probabilities from the SCFG calculations. The quantity $P(H|D, \sigma)$ can be computed for any type of experimental data once a probabilistic model for its structure dependence is obtained. The posterior probability under the combined model can therefore be computed in the existing framework for optimization.

In the case of many chemical probing methods, for example Selective 2′-Hydroxyl Acylation Analyzed by Primer Extension [SHAPE; Wilkinson *et al.* (2009)], the data are assumed to be independent of nucleotide identity and sequence position, and data values for nucleotides in the same pair are not correlated. For such data, $P(H|D, \sigma) = P(H|\sigma)$, which is readily obtained by measuring known structures with the method.

Let H_i be the observed experimental value for position i, and $P_s(i)$ and $P_d(i,j)$ the likelihoods of the alignment (for single-stranded and base paired columns, respectively) calculated purely on the basis of the phylogenetic model. The combined likelihood of the alignment and data, $P'_s(i)$ (single-stranded case for alignment column i) and $P'_d(i,j)$ (base paired case for alignment columns i, j), can then be calculated as

$$P'_s(i) = P_s(i)P(H_i|j \text{ unpaired}) \quad (3)$$

$$P'_d(i,j) = P_d(i,j)P(H_i|i \text{ paired})P(H_j|j \text{ paired}). \quad (4)$$

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/28/20/2691/202359 by guest on 23 March 2024

3 IMPLEMENTATION

PPfold 3.0 has been written in Java 6 and consists of a single jar file. In addition to added support for experimental data, PPfold version 3.0 also features an intuitive graphical user interface. An arbitrary number of experimental data tracks can be added to each prediction. The following types of data tracks are currently supported:

(1) Probing data can be given in the same format as in *RNAstructure* (Mathews *et al*., 2004). The distributions $P(H|\sigma)$ must also be given as histograms; a default distribution for SHAPE data is included in the application.

(2) 'Hard' constraint data can be specified for a sequence in the alignment in the same format as in *mfold* (Markham and Zuker, 2008).

(3) Advanced data tracks are also supported, where the $P(H_i|i$ paired) and $P(H_i|i$ unpaired) values are precomputed by the user for some positions i of the sequence.

4 TESTING

We obtained SHAPE data for the *Escherichia coli* 16S and 23S rRNAs from the authors of Deigan *et al*. (2009) (personal communication), removed invalid data points and computed the $P(H|\sigma)$ distribution histograms for paired and unpaired nucleotides at a resolution of 0.01 units. SHAPE data are currently only available for few sequences with known structures, so we used the data for the *E. coli* 16S rRNA for testing.

PPfold 3.0 is designed for structure prediction based on alignments. Nevertheless, the accuracy of single-sequence structure prediction for the *E. coli* 16S rRNA sequence is greatly improved on the addition of SHAPE data and is comparable to that of *RNAstructure* (Table 1, upper block).

We also examined the effect of SHAPE data on the quality of predictions of various alignments: (a) a Clustal W2 sequence alignment of the highly divergent small ribosomal subunit (SSU) sequences from *E. coli* (accession number K00421) and *Encephalitozoon cuniculi* (accession number X98467); (b) an R-coffee RNA alignment of the same two sequences; (c) the 'SSU high similarity' alignment from BRaliBase II (Gardner and Giegerich, 2004) and (d) the 'SSU medium similarity' structural alignment from BRaliBase II.

SHAPE data improved the quality of structure predictions for all alignments (Table 1, lower block). The highest quality predictions can be obtained when a high-quality alignment is combined with experimental data. In these cases, prediction accuracies exceed what has been observed for a single sequence with *RNAstructure* with additional data, or with PPfold without additional data. However, a single-sequence structure prediction using experimental data is more accurate than a prediction using a low-quality sequence alignment alone or in combination with experimental data. The quality of the input alignment must therefore still be considered when using PPfold 3.0.

Table 1. Evaluation of the accuracy of single-sequence (upper block) and alignment (lower block) structure prediction

Program	Input	Data	PPV	Sensitivity	*F*-measure
RNAstructure	*E. coli* 16S	none	0.3246	0.3462	0.3351
RNAstructure	*E. coli* 16S	SHAPE	0.7221	0.7329	0.7275
PPfold	*E. coli* 16S	none	0.6358	0.2201	0.3270
PPfold	*E. coli* 16S	SHAPE	0.7560	0.6752	0.7133
PPfold	Clustal W2	none	0.3791	0.1709	0.2356
PPfold	Clustal W2	SHAPE	0.6590	0.5534	0.6016
PPfold	R-coffee	none	0.4519	0.2009	0.2781
PPfold	R-coffee	SHAPE	0.6743	0.5562	0.6096
PPfold	SSU high	none	0.8579	0.6838	0.7610
PPfold	SSU high	SHAPE	0.8616	0.7714	0.8140
PPfold	SSU med	none	0.8345	0.7436	0.7864
PPfold	SSU med	SHAPE	0.8721	0.8013	0.8352

PPV: positive predictive value; true positives as a fraction of predicted pairs. Sensitivity is the true positives as a fraction of reference base pairs. True positives are the correctly predicted base pairs. *F*-measure: the harmonic mean of PPV and sensitivity, and provides a measure for overall accuracy.

In conclusion, PPfold 3.0 enables phylogenetic RNA secondary structure prediction in conjunction with experimental data and has the potential of producing highly accurate predictions.

ACKNOWLEDGEMENT

We would like to thank K. Weeks, from UNC Chemistry, for providing the experimental SHAPE data used in this work.

Funding: The Danish Council for Strategic Research under project number 09-061856.

Conflict of Interest: none declared.

REFERENCES

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

Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, **30**, 140.

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.

Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol. (Clifton, NJ)*, **453**, 3–31.

Mathews,D. *et al.* (2004) Incorporating chemical modificiation constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci.*, **101**, 7287–7292.

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

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.

Wilkinson,A.C. *et al.* (2009) Influence of nucleotide identity on ribose 2′-hydroxyl reactivity in RNA. *RNA*, **15**, 1314–1321.

Downloaded from https://academic.oup.com/bioinformatics/article/28/20/2691/202359 by guest on 23 March 2024