BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 5 2013, pages 654–655
doi:10.1093/bioinformatics/btt025

Structural bioinformatics

Advance Access publication January 17, 2013

StatAlign 2.0: combining statistical alignment with RNA secondary structure prediction

Preeti Arunapuram[1], Ingolfur Edvardsson[2], Michael Golden[3], James W. J. Anderson[4], Ádám Novák[4,*], Zsuzsanna Sükösd[5] and Jotun Hein[4]

[1]Department of Computer Science, University of North Carolina at Chapel Hill, NC 27599, USA, [2]Department of Mathematics, Reykjavik University, Reykjavik, Iceland, [3]Department of Clinical Laboratory Sciences, University of Cape Town, Rondebosch, Cape Town, 7701, South Africa, [4]Department of Statistics, 1 South Parks Road, University of Oxford, OX1 3TG, UK and [5]Bioinformatics Research Centre, Aarhus University, C.F. Møllers Allé 8, DK-8000 Aarhus C, Denmark

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Comparative modeling of RNA is known to be important for making accurate secondary structure predictions. RNA structure prediction tools such as PPfold or RNAalifold use an aligned set of sequences in predictions. Obtaining a multiple alignment from a set of sequences is quite a challenging problem itself, and the quality of the alignment can affect the quality of a prediction. By implementing RNA secondary structure prediction in a statistical alignment framework, and predicting structures from multiple alignment samples instead of a single fixed alignment, it may be possible to improve predictions.

Results: We have extended the program StatAlign to make use of RNA-specific features, which include RNA secondary structure prediction from multiple alignments using either a thermodynamic approach (RNAalifold) or a Stochastic Context-Free Grammars (SCFGs) approach (PPfold). We also provide the user with scores relating to the quality of a secondary structure prediction, such as information entropy values for the combined space of secondary structures and sampled alignments, and a reliability score that predicts the expected number of correctly predicted base pairs. Finally, we have created RNA secondary structure visualization plugins and automated the process of setting up Markov Chain Monte Carlo runs for RNA alignments in StatAlign.

Availability and implementation: The software is available from http://statalign.github.com/statalign/.

Contact: novak@stats.ox.ac.uk

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

Received on September 28, 2012; revised on November 30, 2012; accepted on January 12, 2013

1 INTRODUCTION

Our recently created Statistical Alignment package, StatAlign (Novak *et al.*, 2008), performs simultaneous multiple sequence alignment and phylogenetic tree reconstruction using a statistical alignment strategy (Hein *et al.*, 2000). This method operates in a Bayesian framework and uses the Markov chain Monte Carlo (MCMC) method to sample from the joint posterior space of alignments, trees and model parameters. It works with a realistic insertion–deletion model on the gene level, TKF92 (Thorne *et al.*, 1992) and offers a selection of substitution models. The clear advantage of this approach is that the complete evolutionary history is modelled and estimated in a single sound statistical framework.

RNA secondary structure prediction is an important problem in molecular biology—the function of the RNA molecule depends on the way it folds. Stochastic Context-Free Grammars (SCFGs) have been used to great effect in programs such as Pfold (Knudsen and Hein, 2003), which has recently been reimplemented [PPfold; Sükösd *et al.*, 2011]. They use the homology of RNA sequences to help make accurate predictions, as the base pair complementarity must be kept for structural conservation.

Multiple sequence alignment in itself is a challenging problem, as most multiple alignment programs rely on score-based heuristics rather than attempting to model the underlying biological process of substitution, insertion and deletion. It is therefore desirable to implement RNA folding in a statistical alignment framework to gain the advantages from statistical alignment, including being able to predict from multiple alignment samples, and statistical measures associated with the probabilistic model.

2 RNA PLUGIN

We have extended the program StatAlign to include several RNA-specific features. A modified secondary structure-prediction approach that makes use of sampled multiple alignments has been implemented. Each alignment sample is folded using PPfold or RNAalifold (Bernhart *et al.*, 2008), and a base pairing probability matrix is produced. These matrices are averaged to obtain a single base pairing probability matrix, and maximum posterior-decoding (as described in Knudsen and Hein, 2003) is used to obtain a secondary structure. We refer to this method as the *base pair averaging* method. Output is produced on the GUI through the VARNA package [(Darty *et al.*, 2009), see Supplementary Fig. S1] and additionally written to file.

In a second approach, which also makes use of multiple alignment samples, we calculate phylogenetic probabilities as in Knudsen and Hein (2003) by averaging over the phylogenetic probabilities of each of our alignment samples. PPfold then takes these probabilities and produces a secondary structure prediction using the standard SCFG method. Given that this

*To whom correspondence should be addressed.

alignment requires an evolutionary model in order to produce phylogenetic probabilities, this second method of consensus structure prediction is only available for the SCFG method. We refer to this method as the *phylogenetic averaging* method. To benchmark the methods, we took reference alignments from Rfam (Griffiths-Jones *et al*., 2005). In a pre-filtering step, we discarded outlier sequences with many/long insertions and deletions from each family. 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. From 44 families that remained, we took a random sample of 5 sequences, as Knudsen and Hein (1999) suggest this is adequate to capture the evolutionary information. We then re-aligned the sequences using MUSCLE (Edgar, 2009), MAFFT (Katoh *et al*., 2005), Clustal Omega (Sievers *et al*., 2008) and StatAlign, and predicted structures on the resulting alignments. Predictive accuracy was assessed by F-score, the harmonic mean of sensitivity and positive predictive value (Knudsen and Hein, 2003).

Table 1 gives the results of the benchmark. By considering multiple alignments, the average F-score is at least as good or better as many current alignment programs. In particular, the averaging methods improve upon the structure prediction on the StatAlign consensus alignment. When sequences are easy to align, we might expect methods to perform similarly, but when alignment is difficult, a method which can consider multiple alignments will produce fewer poor predictions.

Naturally, additional computational time is required for the RNA secondary structure predictions, for which the algorithms are cubic in the length of the alignment. For the base pair averaging method, we require PPfold or RNAalifold to fold each of the sample alignments. The phylogenetic averaging method is in practice slightly faster (but with the same complexity), as the inside-outside algorithms only runs once.

The user is also provided with additional scores relating to the quality of secondary structure predictions. Firstly, the information entropy (Sükösd *et al*., 2012) is calculated to give the user an indication of how clear the signal of secondary structure prediction is. Paired and unpaired column probabilities are calculated using the Pfold model over alignment samples and averaged. PPfold then uses these probabilities along with SCFG probabilities to calculate the consensus entropy. The second score is a reliability score, designed to give the user an indication of how many base pairings are expected to be correctly predicted in a particular secondary structure. The reliability score we calculate is a modified version of the PPfold reliability score, which averages the posterior SCFG probabilities of base pairs and unpaired bases. Our score considers only base pairs, and correlates more strongly with the F-score measure of predictive accuracy. Evaluating the new reliability score on the benchmarking dataset showed that the R^2 value between this reliability score and the PPfold F-score increased from 0.2685 for the existing PPfold reliability score to an R^2 value of 0.4021 for our modified reliability score, when performing linear regression. These calculations are only available for the SCFG method.

Lastly, an option to automate StatAlign's MCMC parameters for RNA alignments has been included—previously, the user was expected to specify the parameters of the MCMC run prior to execution. The burn-in process is terminated when we get a significant decline in the log likelihood graph. The sample rate is calculated by taking many samples and measuring the distances between alignments: the rate is chosen to be the number of steps such that the distance between alignments approaches a constant. The MCMC run is terminated when adding more alignment samples no longer results in significant changes to the consensus alignment. Automating the process in this way allows the user to obtain desirable results which otherwise might not be the case for a poor choice of parameters.

Table 1. Average F-scores for RNA secondary structure prediction benchmarks on 44 re-aligned Rfam alignments

Alignments	PPfold	RNAalifold
Reference alignments	0.710	0.746
MUSCLE alignments	0.539	0.572
MAFFT alignments	0.567	0.599
Clustal Omega alignments	0.554	0.598
StatAlign alignments	0.561	0.584
StatAlign: base pair averaging	0.562	0.605
StatAlign: phylogenetic averaging	0.614	NA

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. We would like to thank the Oxford Supercomputing Centre for computational resources and support.

Funding: EU COGANGS Grant, BBSRC (Á.N.) and EPSRC (J.W.J.A.).

Conflict of Interest: none declared.

REFERENCES

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

Darty,K. *et al*. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974.

Edgar,R.C. (2009) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.

Griffiths-Jones,S. *et al*. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*., **33** (**Suppl. 1**), D121–D124.

Hein,J. *et al*. (2000) Statistical alignment: computational properties, homology testing and goodness-of-fit. *J. Mol. Biol*., **302**, 265–279.

Katoh,K. *et al*. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*., **33**, 511–518.

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.

Novak,A. *et al*. (2008) StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics*, **24**, 2403–2404.

Sievers,F. *et al*. (2008) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol*., **7**, 539.

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

Sükösd,Z. *et al*. (2012) Characterising RNA secondary structure space using information entropy. *BMC Bioinformatics*, **14** (**Suppl. 2**), S22.

Thorne,J.L. *et al*. (1992) Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol*., **34**, 3–16.