Pfold: RNA secondary structure prediction using stochastic context-free grammars

Bjarne Knudsen* and Jotun Hein

BiRC (Bioinformatics Research Center), University of Aarhus, Hæegh Guldbergsgade 10, Building 090, 8000 Århus C, Denmark and 1Department of Statistics, Oxford University, The Peter Medawar Building for Pathogen Research, South Parks Road, Oxford OX1 3SY, UK

ABSTRACT

RNA secondary structures are important in many biological processes and efficient structure prediction can give vital directions for experimental investigations. Many available programs for RNA secondary structure prediction only use a single sequence at a time. This may be sufficient in some applications, but often it is possible to obtain related RNA sequences with conserved secondary structure. These should be included in structural analyses to give improved results. This work presents a practical way of predicting RNA secondary structure that is especially useful when related sequences can be obtained. The method improves a previous algorithm based on an explicit evolutionary model and a probabilistic model of structures. Predictions can be done on a web server at http://www.daimi.au.dk/~compbio/pfold.

INTRODUCTION

RNA structures are essential in many biological processes and are often conserved in evolution. Examples of such conserved structures are found in tRNA (1), rRNA (2,3), tmRNA (4), RNase P RNA (5) and SRP RNA (6). Many computational methods have been developed for predicting RNA structures. Early algorithms were made by Nussinov et al. (7) and Zuker and Stiegler (8). Zuker’s energy calculations have been further improved (9–11) and are probably the most widely used RNA secondary structure prediction method today (with the MFOLD program).

The work of Knudsen and Hein (12) (here denoted as the KH-99 algorithm) combines an explicit evolutionary model of RNA sequences with a probabilistic model for secondary structures. It assumes an alignment and gives one common structural prediction for all the sequences.

This work improves the KH-99 algorithm primarily by making it faster and more robust toward alignment errors.

A thorough evaluation of performance under different circumstances is also included. This new method called Pfold is available through a web-based server www.daimi.au.dk/~compbio/pfold.

METHODS

Pfold is based on the KH-99 algorithm, which was only useful for a limited number of sequences due to its large computation time. This work makes the algorithm practically useful for larger numbers of sequences. The main concerns are treatment of gaps, computational speed and robustness.

The KH-99 algorithm

The KH-99 algorithm uses a stochastic context-free grammar (SCFG) to produce a prior probability distribution of RNA structures. Given an alignment and a phylogenetic tree relating the sequences, posterior probabilities of the structures can be calculated using the inside–outside algorithm (13). The posterior probability is based on individual probabilities for alignment columns or pairs of columns in the case of a base-pair.

Column probabilities are calculated using the likelihood approach by Felsenstein (14). The evolution of column pairs is modelled using a rate matrix for base-pairs (i.e. a 16 by 16 matrix). The most likely structure is found using the CYK algorithm (15). The tree is estimated using a maximum likelihood approach in the SCFG model described.

Both the evolutionary and structural parameters of the KH-99 model are based on extensive tRNA and large subunit rRNA databases (12).

Gaps

Treating gaps in an appropriate way is a returning problem in biological sequence analysis. The best way to deal with gaps would probably be to make an explicit evolutionary model for insertions and deletions, and use that in the sequence analysis. Unfortunately, such calculations are often complicated, as in statistical alignment (16,17). Two simpler ideas are to treat
gaps as a fifth nucleotide or to treat them as unknown nucleotides.

When using an evolutionary model, a number of problems arise from treating a gap as a fifth nucleotide. First, the frequency of gaps will depend on how many sequences are being analysed and on their evolutionary distance. Furthermore, gaps cannot be viewed as evolving in the same way as a nucleotide, thus the rates of evolution are difficult to specify. This approach does, however, have the advantage that it includes the potentially useful information from the insertions and deletions. This method was successfully used in the RNA part of the non-coding RNA gene finding algorithm by Rivas and Eddy (18). When treating gaps as unknown nucleotides, a gapped sequence position should have probability one for any nucleotide. This has the advantage that the probability of a column with gaps is equal to the probability of the same column in an alignment without the gapped sequences, and the tree correspondingly pruned (Fig. 1).

Pfold uses the latter approach, which is often done in situations where different alignment columns are looked at individually (19). In RNA structure prediction, pairs of columns are analysed together, which can give rise to some difficulties: when treating gaps as unknowns, gaps can form pairs with nucleotides (the top of Fig. 2). This problem was handled by removing columns where less than 75% of the sequences have nucleotides (the bottom of Fig. 2).

Unknown nucleotides

In biological sequences, some nucleotides may be unknown or only partial information may be available. These situations can be treated by letting the unknown nucleotide have a probability of one for each of the possible nucleotides. This means that if a given position is known to be a pyrimidine, its probability of being a U is set to one, and its probability of being a C is also set to one. Using this method, any symbols of the extended nucleotide alphabet can be treated correctly by Pfold. This is in accordance with Felsenstein (14).

Tree estimation

In the KH-99 algorithm, the tree was estimated through a maximum likelihood method using the SCFG model. While this gave good results and was interesting with respect to phylogenetic analysis, it was slow. A much faster method is to estimate the tree first. This can be done using standard methods.

In Pfold, pairwise distances between sequences are calculated using maximum likelihood. The rate matrix used should correspond as closely as possible to what is used in the KH-99 algorithm. Since the tree is calculated before the structure has been estimated, a single rate matrix has to be used for this purpose. It was made from the KH-99 algorithm by summing the loop rate matrix and a reduction of the base-pair rate matrix to single positions. The rate matrices were weighted with the probabilities that a given position is in a loop region or a stem region, respectively. The tree is calculated from pairwise distances using the neighbour joining algorithm (20) and adjusting branch lengths to maximum likelihood estimates. This gives a large increase in speed, since the inside–outside calculations only need to be performed once, as opposed to the multiple iterations used in the KH-99 method for estimating the tree.

Robustness

Pfold assumes that all sequences have exactly the same structure. This means that a single sequence with a slightly different structure might ruin a prediction. The same situation applies for alignment errors and sequencing errors. When a single error might change a prediction significantly, the method is not robust (the top of Fig. 3).

A way to make the algorithm more robust is to let any nucleotide have a small probability of being any other nucleotide. The interpretation of this is most obvious in terms of sequencing errors, but the method works for alignment errors and structure differences, too. Figure 3 shows how
introduction of this probability changes the results. In Pf old this probability was set to 1%.

Partially known structure

Often, something is known about the structure being predicted and including this knowledge in the analysis can give improved results. Different kinds of knowledge can be used:

- That two given columns form a pair together.
- That a given column is involved in a pair.
- That a given column is unpaired.

This can all be included in the calculations by letting the structures that do not satisfy the previous knowledge have a probability of zero. This approach does not change relative prior distributions of allowed structures.

As a side note, no loops of length two are allowed in this implementation, as opposed to the KH-99 algorithm. This was implemented by disallowing pairs between positions of distance less than four.

What structure should be chosen?

A prediction program should of course report a single prediction as the best. The CYK algorithm for finding the most likely parsing from the grammar is often used (15). An alternative to this has been chosen here: Pf old reports the nested structure with the highest expected number of correct predictions. Appendix 1 (available as Supplementary Material) describes how this structure can be found. Notice that this removes some of the problems associated with using the CYK algorithm on ambiguous grammars.

Once the best nested structure has been chosen, the reliability of the prediction is evaluated for each position. This is done by finding the probability that each prediction (specific pair or unpaired) was correct, given the model and the data. The variables from the inside–outside algorithm can easily be used to give this information (Appendix 1 available as Supplementary Material). Knowing which parts of the prediction are reliable is very important when using the prediction in further work (Fig. 4).

Finding a single best structure may not always give all the information that would be useful. To give an overview of the prediction, a dot plot is produced as well. It is a square plot of pairing probabilities for all different pairs. Each probability is represented by the size of a dot in the appropriate position. Probabilities of not pairing are shown on the sides of the plot (Fig. 5). These calculations resemble the work by McCaskill (21).

Making the obvious individual structure changes

Sometimes, a structure in a single sequence will have a slightly longer stem than its homologues. This can be incorporated in the prediction by extending a stem if immediate neighbours can form base pairs. Another obvious change is to remove non-standard base pairs from individual sequences. This is done after the structure predictions given by Pf old.

INPUT AND OUTPUT

The input to the web server is an alignment of up to 40 sequences and 500 positions. The alignment should be given in the FASTA format with gaps represented by hyphens (‘-‘). Previously known structural elements can be incorporated in the prediction by adding a specific sequence with the relevant information (see web site for details). When a prediction is done, the web server returns an email to the user. This email contains a link to a web page with the results. On the web page, the following is available in multiple formats:

- A summary of the input.
- The calculated tree.
- The predicted structure given as a bracket notation.
- Reliabilities of individual predictions.
- A dot plot.

RESULTS

For all predictions in this section, two versions were made. The first version was made using the alignment from the database of the sequences being analysed. This alignment was assumed to be ‘correct’. In the second version, sequences from databases were aligned using the ClustalW alignment program by Thompson et al. (22) to imitate a realistic scenario of RNA structure prediction.

Test sets

A number of test sets were made (Table 1). The sets A, B and C are used in evaluating the prediction accuracy as a function of the number of sequences used in the analysis. Test set D is used to show how prediction accuracy varies as a function of evolutionary distance.

Prediction accuracy

An evaluation of the prediction accuracy is shown in Figure 6. The accuracy was calculated as the percentage of positions for which the secondary structure was correctly predicted. For a pairing position to be counted as correct, the position of the predicted pair had to be correct. Sequences in the test sets used here have a maximum pairwise distance of 0.50 units in the Jukes–Cantor model (23). This means that the sequences are quite diverse, but still possible to align without too many errors.

As expected, prediction accuracy rises with the number of sequences used, as more covariance information becomes
This method of illustrated some obtained RN right prediction. In were case, distances were obtainable. Another detail is that 0.5 zero makes in eukaryon within the increase here positions. Hand predicted back, length shown drops. Sequence small was cor no as back. cer zea.ma-c, is with two aspect, which are of between type RNA expected 282 estimate a computer hum.lu-c, problem. Using rat.rat, 31, discuss method two reliable these parentheses which were SRP improvement written the hom.sa-a, implement bac. ste., substantial the due alignment increases better of different Pf y within reliability hum.ja-a, also show the that The performance see show the based left RN in is a KH-99 e sal.typ., from the quality the trial, and sch.pom the of effect pairwise sequence this are these parentheses are bacterial are rather into the using cae.el-b, much model with much distances (363.8) different ara.th-b, o the the i with sequenc this are No bac.sph., a pseudoknot hum.lu-d, the algorith this length The large o a o distance (12). (exceptions to this are discussed below). An accuracy of ~75% is obtainable with six sequences and Pfold can cope with many more sequences, so even higher accuracies are possible.

Pfold was also used on the original test set of Knudsen and Hein (12). There were a few small differences in the results compared to the KH-99 algorithm and the overall performance was slightly better in this new version of the algorithm.

**Evolutionary distance effect**

There are two effects of evolutionary distance on prediction accuracy: large distances imply much covariation information, but it also means that sequences are difficult to align. Another issue with large evolutionary distances is that secondary structures may not be conserved between the sequences. Since Pfold assumes a single common structure for all sequences, this can pose a problem.

These effects are illustrated in Figure 7. When using ‘correct’ alignments, the accuracy rises with distance, as the evolutionary information increases. The accuracy levels off at ~80%, which seems to be the maximum obtainable average for two sequences of this type using Pfold. The graph made from the ClustalW alignments show how accuracy increases until an evolutionary distance of ~0.60. After this, the accuracy drops due to alignment errors. At a distance of ~0.90 the quality of alignments become so low that the structures might as well be predicted individually.

**Speed**

The algorithmical improvements made in this work reduces the computation time substantially. The largest improvement is from estimating a single evolutionary tree and fixing it, rather than estimating all the branch lengths under the RNA model. For details on the computation times, see Appendix 2 (available as Supplementary Material).

**DISCUSSION**

Some aspects of this method remain to be explored, as described by Knudsen and Hein (12). These include: base
base-pairing. The x-value of a point is the average of the 50 distances. The y-value is the accuracy of the alignment.

The accuracy can be further increased by using a grammar that takes into account the base-pairing interaction. This is achieved by using a grammar that is more closely describing real RNA structures and other models for base-pair evolution.

If base-stacking interactions and a better grammar is incorporated in Pfold, the prediction accuracies should become close to the MFOLD results for single sequences (9–11), since the methods resemble each other closely in that situation. For multiple sequences, Pfold should still be able to perform even better.

When structures are conserved in evolution, inclusion of information from multiple related sequences improves predictions. A number of methods have been developed that combine energy calculations with evolutionary information through a given alignment (24–27). These methods have proven quite reliable but lack an explicit evolutionary model.

As emphasised by this work, an important aspect of RNA structure prediction is the alignment problem. In methods that depend on a sequence alignment, the success of the method is closely linked to the quality of the alignment. Some work has been done in the field of RNA structural sequence alignment (28). Rivas and Eddy (18) developed an RNA model that takes the alignment of two sequences into consideration through a pair-SCFG. They did, however, assume the alignment to be given since computational time would be prohibitive in using their method to align sequences. The RNAGA method by Chen et al. (29) predicts consensus structures without trying to align the sequences which might be a useful approach to avoiding the alignment problem.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

B.K. would like to thank the University of Florida, the Carlsberg Foundation (grant no. ANS-0604/20) and the Danish National Science Research Council (grant no. 21-00-0283) for their support. B.K. also thanks Christian N. S. Pedersen for his great help in answering questions about Unix and programming. We acknowledge the help of Ebbe S. Andersen in making this method practically useful and in illustrating the web server. Finally, we thank Jakob S. Pedersen for helpful comments on the manuscript.
REFERENCES


