Supplementary material

Global or local? Predicting secondary structure and accessibility in mRNAs
Sita J. Lange1,*, Daniel Maticzka1,*, Mathias Möhl1, Joshua N. Gagnon2, Chris M. Brown2, Rolf Backofen1,**

1 Department of Computer Science, Albert-Ludwigs-University, Freiburg, Germany
2 Department of Biochemistry and Genetics Otago, University of Otago, P.O. Box 56, 710
Cumberland St, Dunedin 9054, New Zealand
* Authors have contributed equally to this work
** E-mail: backofen@informatik.uni-freiburg.de

Algorithms for secondary structure prediction

We give a general overview of previously published methods for secondary structure prediction to enable a better comprehension of the results presented in this work. Furthermore, we give the execution calls used in the performance comparisons.

Global folding

To fold an RNA sequence globally means that structures are predicted for the entire input sequence for which all possible pairwise base-pairs are considered. Global base-pair probabilities are predicted via a partition function algorithm that considers the entire ensemble of all possible structures weighted by their free energies [1]. These free energies are calculated with a nearest neighbor energy model using thermodynamic parameters determined by the Turner group [2,3]. The algorithm is implemented in RNAfold [4], RNAstructure [5], and UNAfold (formerly known as mfold) [6] and has a complexity $O(n^3)$ time and $O(n^2)$ space. Each implementation has different additional features that are not relevant to this work, however, they are all based on the same algorithm for computing base-pair probabilities. We used RNAfold from the Vienna Package Version 1.8.4 as a representative of the global folding approach. The options used in this study are RNAfold -d2 -p -noLP. For folding under the constraint of the consensus structure, we used the additional option -C.

RNAfold does not compute accessibilities, but position-wise accessibilities (as measured in the YeastUnpaired dataset and used in our evaluation) can be computed from the base-pair probabilities as follows:

$$pu(i) = 1 - \sum_{j=1}^{n} p(i,j),$$  \hspace{1cm} (1)

where $pu(i)$ is the probability for base $i$ to be unpaired (i.e. its accessibility); $n$ is the length of
the RNA sequence; \((i, j)\) is a base-pair between base \(i\) and base \(j\); and \(p(i, j)\) is the probability for the base-pair \((i, j)\) according the the McCaskill algorithm [1].

**Local folding**

In local folding there is a concept of locality, which means only locally stable structures are predicted. Locality is determined by the base-pair span, i.e. the number of nucleotides between the two bases of a base-pair \((i, j)\):

\[
bp-span(i, j) = j - i + 1, \quad i < j.
\]  

The first known approach for the prediction of stable local secondary structures was presented in [7]. Compared to the global approach, it introduces a maximal base pair span \(L\) such that the predicted structure contains no base-pairs \((i, j)\) with \(bp\text{-}span(i, j) > L\). As a result, the predicted structures are local in the sense that they do not contain any long-range base-pairs that connect distant parts of the sequence. However, the predicted structures are still global in the sense that the energy of all local structures that cover the entire sequence is considered in a single calculation. In particular, although there are no long-range base-pairs, there might be long-range dependencies among the base-pairs. As a consequence of the severe search-space reduction, the prediction for a sequence of length \(n\) and a maximal base-pair span of \(L\) can be done in time \(O(nL^2)\) and space \(O(n + L^2)\); thus allowing for genome-wide surveys.

We use Rfold [8] in our analysis to represent this approach for base-pair probabilities and Raccess [9] for accessibilities. The commands for Rfold and Raccess were `run_rfold -max_pair_dist=L -print_prob=true` and `run_raccess -max_span=L -access_len=1`.

RNAplfold [10, 11] is currently the most cited method for computing base-pair probabilities of local secondary structures. It also includes the maximum base-pair restriction parameter \(L\). The partition function, however, is not computed for the entire sequence simultaneously (as for Rfold or Raccess), but for all subsequences (windows) of size \(W\) independently. The probability of a base-pair is then derived by averaging over all windows that contain this base-pair. The algorithm has the same time and space complexity as Rfold, but depends on the size of \(W\) instead of \(L\). The execution call for RNAplfold is: `RNAplfold -noLP -W W -L L -u u 1`.

**Curated benchmark set of cis-regulatory elements**

From all 222 families labelled as “Cis-reg” in the Rfam database version 10.0 [12,13], we have selected 98 with experimental evidence, which are likely to have well defined structures. These families comprise sequences from eukaryotic, bacterial and viral genomes with diverse cis-regulatory functions. More information about each one is available through the CisReg website, [http://lancelot.otago.ac.nz/CisRegRNA/](http://lancelot.otago.ac.nz/CisRegRNA/), with links to Rfam.
We extracted the seed alignments for each family from the database and used BLAST [14] to locate the positions of each element. Subsequently, we extracted the element sequences from the original sequences within contexts of 100, 200, 500 nt to either side of the functional element where possible. If there was not enough context, the sequences were extended to the beginning and/or end of the mRNA. We further extracted full-length mRNAs up to maximum context of 3000 nt. Some of the original sequences are genomic from bacterial or viral genomes, so that possibly non-mRNA sequences are within the dataset. This fact, however, should not significantly influence the comparison of the prediction methods. We divided the dataset into originating from mRNA or genomic context to separately test the trends observed. To gain the exact base-pairs for each structural element in the family, we mapped the given consensus structure to the individual sequences. Any base-pair not consisting of GC, AU, or GU were omitted. Furthermore, any base-pairs that did not allow for at least three unpaired bases within a hairpin loop were also omitted.

The consensus structure only includes base-pairs that are common to all elements within a given family, although more base-pairs are likely to form in the individual element to improve its stability. To find the most stable structure, each element was folded with RNAfold using the consensus base-pairs as a constraint and the resulting minimum free energy structure was used. In the process of mapping the consensus structure to the single elements, non-conforming base-pairs were deleted. Therefore, we filtered out any elements that (i) did not retain at least 80% of the original base-pairs in the consensus structure and (ii) did not retain at least 80% of the mapped base-pairs in the final constraint structure as folded by RNAfold (constraint folding with RNAfold sometimes results in a constraint base-pair being left unpaired – see RNAfold manual). This means, only sequences that were very similar to the consensus structure in Rfam were used.

The final dataset, referred to as CisReg, consists of 2500 individual structural elements from 95 cis-regulatory elements located on the mRNA across many different species. RF00632, RF00227, and RF00524 were not used because they did not pass the filtering steps described above. Possible reasons for this are as follows: RF00632 (sxy 5’ UTR element) includes only two sequences in the seed alignment and the 16 sequences in the full alignment from H. influenzae have 97% identity, i.e. poor evidence for the consensus structure. RF00227 (ftz instability element 3’) is mainly unstructured with a small nine base-pair stem in the centre. RF00524 (R2 RNA element) is a large computationally predicted structure that has a functional ribozyme within it. Subsequent updates to CisReg and Rfam entries should address these deficiencies. With this dataset, we evaluated over 85,000 base-pairs.

**Dataset redundancy evaluation**

During the manual curation the families were chosen to exclude similar families.
**Similarity within families**

To assess the sequence redundancy of our dataset, we took the sequences with 100 nt context to either side of the element. We selected this context, because the direct context is the most influential region for the structure prediction of the regulatory element. We subsequently clustered these sequences using BlastClust [14]. This program groups sequences according to sequence similarity. To avoid the problem of overlapping sequences as described above, we set the coverage of both sequences to 100%. To assess the amount of sequence similarity, we varied the percent identity of the pairwise alignments from 10% to 100% in steps of ten. We received the following number of clusters: 2460 (100% identity), 1759 (≥ 90% identity), 1671 (≥ 10− ≥ 80% identity). Therefore, at least 1671 sets are different problems with respect to structure prediction. Even a single mutation can alter the element structure at specific locations. Modest sequence differences (e.g. > 20%) usually result in different foldings and thus form different problem sets. As most of the redundancy is due to similar sequences within a family, we separated the analysis into families in Supplementary Figure 1. Here we observe the same trends as we presented in the main paper for most of the families. In addition to the support of the YeastUnpaired results, this analysis exhibits the reproducibility of our results.

The program call for BlastClust (pI = percent identity) was:

```
blastclust -i sequences_context100.fasta -o blastclust.out -p F -L 1 -b T -S pI.
```

We have also reported the overall pairwise similarities of the seed alignments in the online database.

**Similarity between families**

The clustering analysis on primary sequence indicate that there are distinct sets, however to access the redundancy in folds we have used cmcompare [15] to do pairwise comparisons of each of the covariation models to one another and to the whole Rfam 10 set of models. For cmcompare scores of over 20 are considered ‘worthy of note’ and 7.4% of the entire Rfam database had such scores. However, no pairs within the CisReg set used here had scores over 20. Although there were some notable matches to other Rfam families not in the set analysed here. We also compared the primary sequence of the first member of each family to all the sequences in the other families using blastn. Only 9 of 98 had matches with E < 1.0 in the other sequences. These were all short regions of identity < 13 bases long, too short to contribute substantial common secondary structures.

**References**

1. McCaskill, J. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7), 1105–19.
2. Mathews, D., Sabina, J., Zuker, M., and Turner, D. (1999) Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J. Mol. Biol.*, 288(5), 911–40.

3. Turner, D. and Mathews, D. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, 38(Database issue), D280–2.

4. Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie - Chemical Monthly*, 125(2), 167–188.

5. Reuter, J. and Mathews, D. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11, 129.

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

7. Hofacker, I., Priwitzer, B., and Stadler, P. (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20(2), 186–190.

8. Kiryu, H., Kin, T., and Asai, K. (2008) Rfold: an exact algorithm for computing local base pairing probabilities. *Bioinformatics*, 24(3), 367–73.

9. Kiryu, H., Terai, G., Imamura, O., Yoneyama, H., Suzuki, K., and Asai, K. (2011) A detailed investigation of accessibilities around target sites of siRNAs and miRNAs. *Bioinformatics*, pp. 1–11.

10. Bernhart, S., Hofacker, I., and Stadler, P. (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5), 614–5.

11. Bernhart, S., Mückstein, U., and Hofacker, I. (2011) RNA Accessibility in cubic time. *Algorithms Mol. Biol.*, 6(1), 3.

12. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S., and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, 33(Database Issue), D121–4.

13. Gardner, P., Daub, J., Tate, J., Moore, B., Osuch, I., Griffiths-Jones, S., Finn, R., Nawrocki, E., Kolbe, D., Eddy, S., and Bateman, A. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, 39(Database issue), D141–5.

14. Altschul, S., Gish, W., Miller, W., Myers, E. W., and Lipman, D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215(3), 403–10.

15. Höner zu Siederdissen, C. and Hofacker, I. L. (2010) Discriminatory power of RNA family models. *Bioinformatics*, 26(18), i453–9.
Figure 1. The median *bp-accuracy* is shown separately for each of the 95 *Rfam* families within the CisReg data using sequence contexts of 500 nt. The families are sorted by the maximum base-pair span of their elements, ranging from 15 to 551. This information is more relevant than the actual element length, because this corresponds to the parameter $L$ used. *RNAfold* only performs better than the other methods when the base-pair spans of the structure greatly exceed the maximum base-pair span parameter $L = 150$. In general, we see similar trends across most families and no bias due to data redundancy is evident.
Figure 2. Average accessibilities per window position for the 400 mRNAs used for Figure 3, split by GC-content of the windows. While average accessibilities decrease with increasing GC-content, border nucleotides are distinctly more accessible for all instances.
Figure 3. Average accessibilities per window position for ten 15,000 nt random sequences ranging in GC-content from 10 – 100%. Sequences were folded with $L = 100$ and $W = 150$ (lower) and $W = 100$ (upper). Folding of each sequence resulted in $15,000 - W + 1$ independent folding windows.
Figure 4. Base-pair accuracy box-plots for the CisReg data for several window sizes $W$ and span $L = 150$, using RNAplfold. For the comparison on the CisReg data shown in Figure 6, the span $L$ was optimized using Rfold and thus independently from RNAplfold or LocalFold. Given this $L$, we selected the optimal $W$ for RNAplfold. Results for $W = 200$ show the highest median accuracy and the smallest variance.
Figure 5. AUCs for separating high-scored and low-scored nucleotides from the YeastUnpaired data for several window sizes $W$ and span $L = 100$, using RNAplfold. The comparison of YeastUnpaired in Figure 8 was done for several $L$ (fixing $W$ at $L+50$). The best result for RNAplfold was reached using parameters $L=100$ and $W=150$. This is the optimal $W$ for this span for RNAplfold.