Conserved RNA Structures

Ivo L. Hofacker

Institut for Theoretical Chemistry, University Vienna

http://www.tbi.univie.ac.at/~ivo/

Bled, January 2002

Energy Directed Folding

Predict structures from sequence alone, by minimizing free energy.

+ works on a single sequence
+ sophisticated energy model
+ efficient algorithms, good implementations, many variants
- unreliable, at best 70% of predicted pairs correct
- gives no information on functional vs. incidental structures.

Covariance Methods

infer structure from phylogenetic comparison, especially compensatory mutations.

+ highly accurate predictions
+ yields conserved (functional) structures
- few available programs
- requires many sequences with good alignments

Obviously a combination energy directed and covariance methods is desirable.

Score Functions for Comparative Structure Prediction

Need to derive a score for a possible pair i, j by comparing columns i and j of the alignment.

- counting compensatory mutations
- our alifold covariance score
- mutual information
- based on a phylogenetic tree
- several weird ad hoc methods

Mutual Information

Entropy of a random variable X with probability distribution $p(x)$ is

$$H(X) = -\sum_x p(x) \log_2 p(x)$$

The mutual information of two distribution is given by

$$M(X;Y) = H(X) + H(Y) - H(X,Y)$$
$$= H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Obviously we have

$$M(X;Y) = M(Y;X) \text{ and } M(X;Y) \geq 0$$
$$\text{with } M(X;Y) = 0 \iff p(x,y) = p(x)p(y)$$

Mutual Information II

Easily computed directly from frequencies in column i and j of alignment:

$$M_{i,j} = \sum_{x,y} f_{ij}(xy) \log_2 \frac{f_{ij}(xy)}{f_i(x) f_j(y)}$$

For the 4 letter alphabet $\mathcal{A} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$, $0 \leq M_{ij} \leq 2$ bits.

+ Completely parameter free
+ No model of sequence evolution or phylogenetic tree needed
± Uses no prior knowledge about secondary structures
  + can detect tertiary contacts and functional constraints
  - poor signal to noise for small data sets
  - only compensatory mutations contribute, consistent mutations (GC $\rightarrow$ GU) are neglected

Alifold Covariance Score

Let $\Pi_{ij}^{\alpha} = 1$ if sequence α can pair positions i, j;
$d_{ij}^{\alpha,\beta}$ hamming distance of α and β at positions i and j (e.g. 0,1, or 2).

$$C_{ij} = \frac{1}{N^2} \sum_{\alpha,\beta} d_{ij}^{\alpha,\beta} \Pi_{ij}^{\alpha} \Pi_{ij}^{\beta}$$

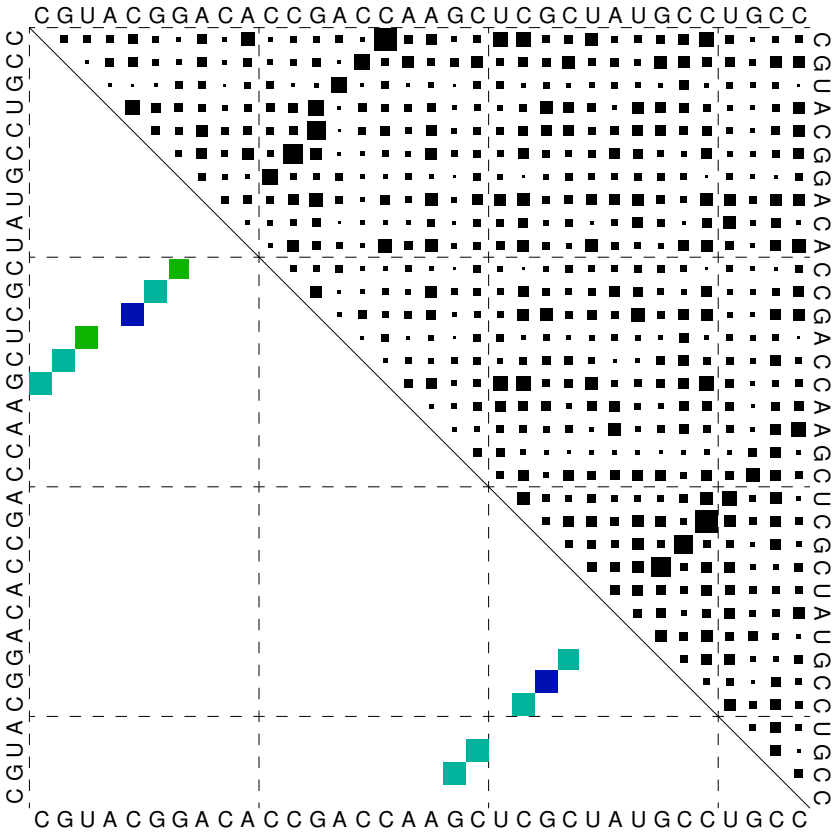
$$= \frac{1}{2N^2} \sum_{\mathrm{xy,x'y'}} f_{ij}(\mathrm{xy}) \mathbf{D}_{\mathrm{xy,x'y'}} f_{ij}(\mathrm{x'y'})$$

where $\mathbf{D}_{\mathrm{xy,x'y'}}$ contains $d_H(xy, x'y')$ if xy and $x'y'$ are allowed pairs , else 0.
Including a penalty for non-standard pairs set

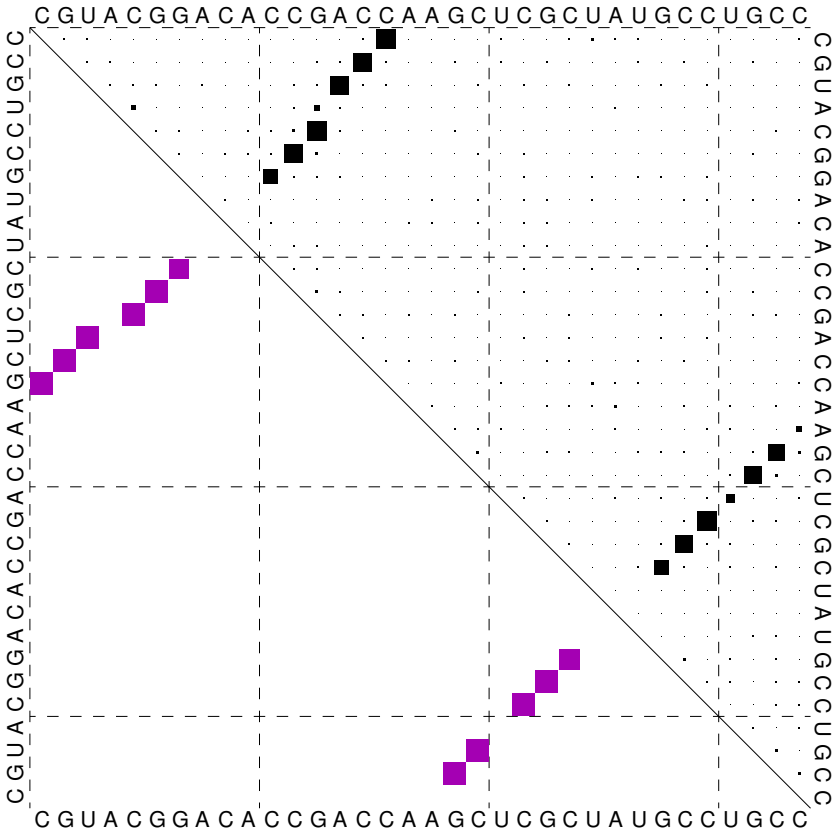
$$B_{ij} = C_{ij} - \varphi \left(1 - \frac{1}{N} \sum_{\alpha} \Pi_{ij}^{\alpha} \right)$$

Artificial Test Case

Generate sequences folding into the structure `(((.(((...)))))).((.(((...))).)).` using `RNAinverse`.



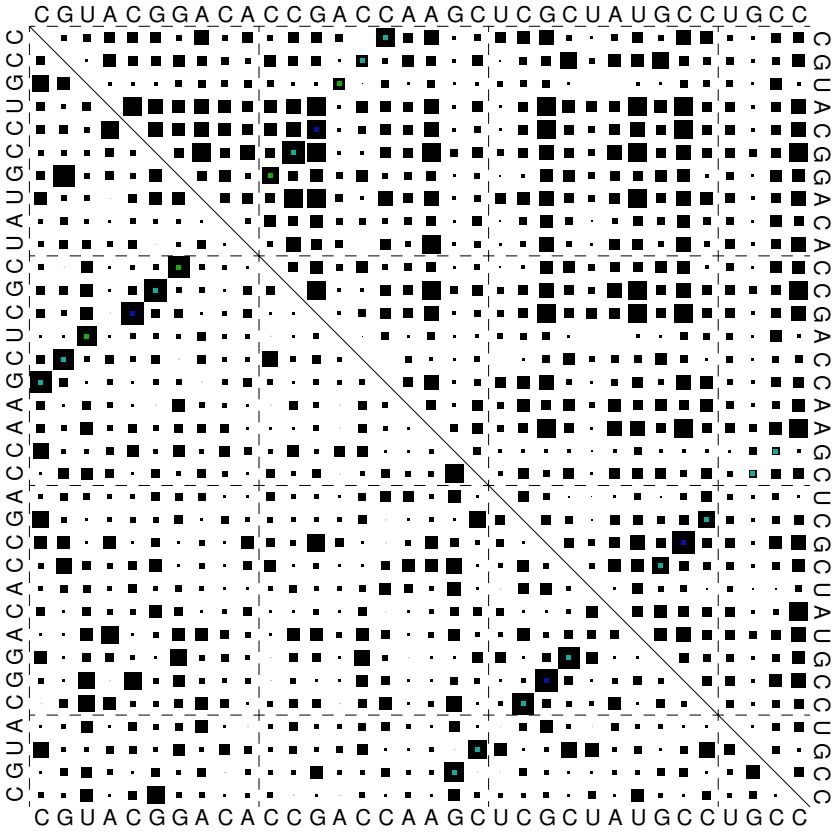
MI from 10 sequences



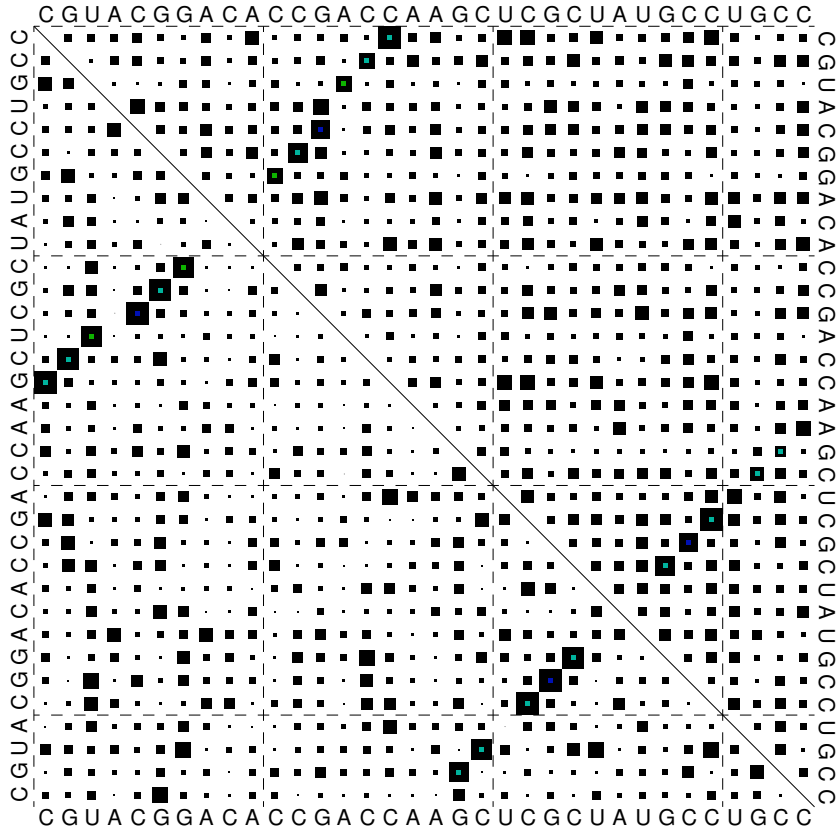
MI from 100 sequences

Artificial Test Case II

Comparing mutual information and covariance score



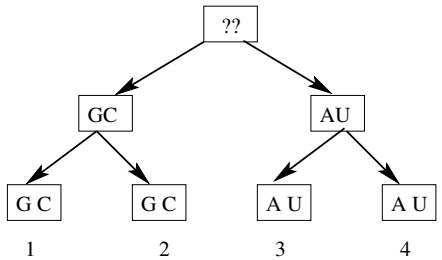
from 5 sequences



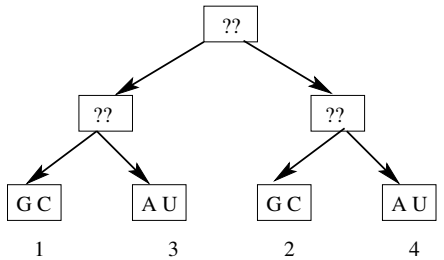
from 10 sequences

Scoring with Phylogenetic Tree

Idea: Same data pair frequencies may be produced by different histories



Single Mutation



Multiple Mutations

Right scenario gives stronger support for base pair. How to quantify this?

Scoring with Phylogenetic Tree

Haussler: Given the phylogenetic tree T data d (two columns of the alignment) compute the probability of the data given two models for a) conserved pair b) independent positions. Use the log-odds score:

$$score = \log \frac{P(d|T, \wedge pair)}{P(d|T \wedge nopair)}$$

To calculate $P(d|model)$ need to sum over all possible histories. Luckily, this can be done recursively.

Recursive calculation of probabilities

Let f be the root node of some subtree of T with children r and l; $d(f)$ denotes the data on that subtree.

$$P(d|model) = \sum_{\pi} P(d|root = \pi) * P(root = \pi)$$

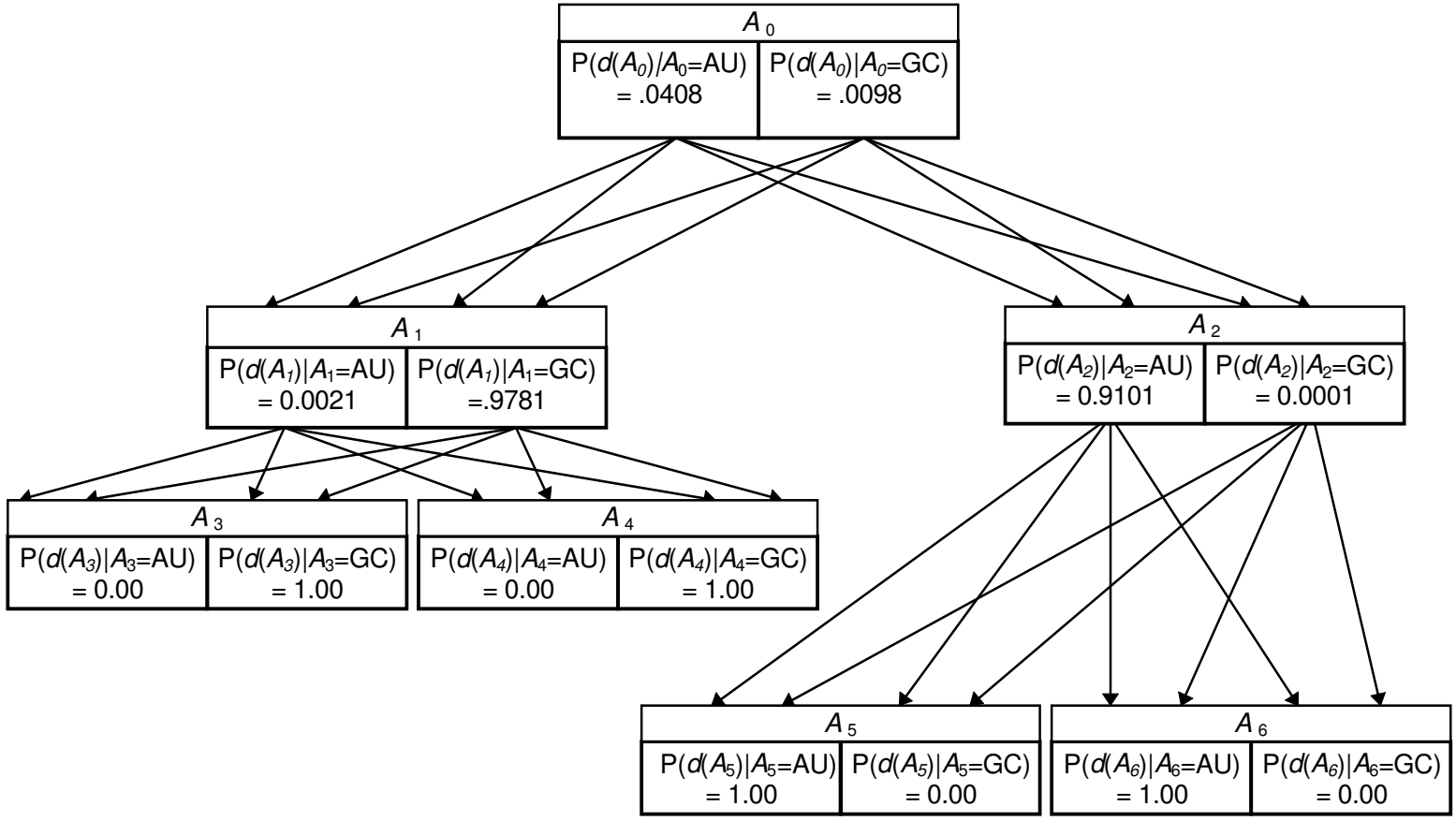
$$P(d(f)|f = \pi) = \sum_{\pi'} P(d(r)|r = \pi')P(r = \pi'|f = \pi) \cdot$$

$$\sum_{\pi'} P(d(l)|l = \pi')P(l = \pi'|f = \pi)$$

Recursion can be started at the leafs whose sequence is known.

$$\mathrm{P}(d|\mathrm{Model}) = \sum_{l=0,1} \mathrm{P}(d(A_0)|A_0 = l \wedge \mathrm{Model}) \cdot \mathrm{P}(A_0 = l|\mathrm{Model})$$

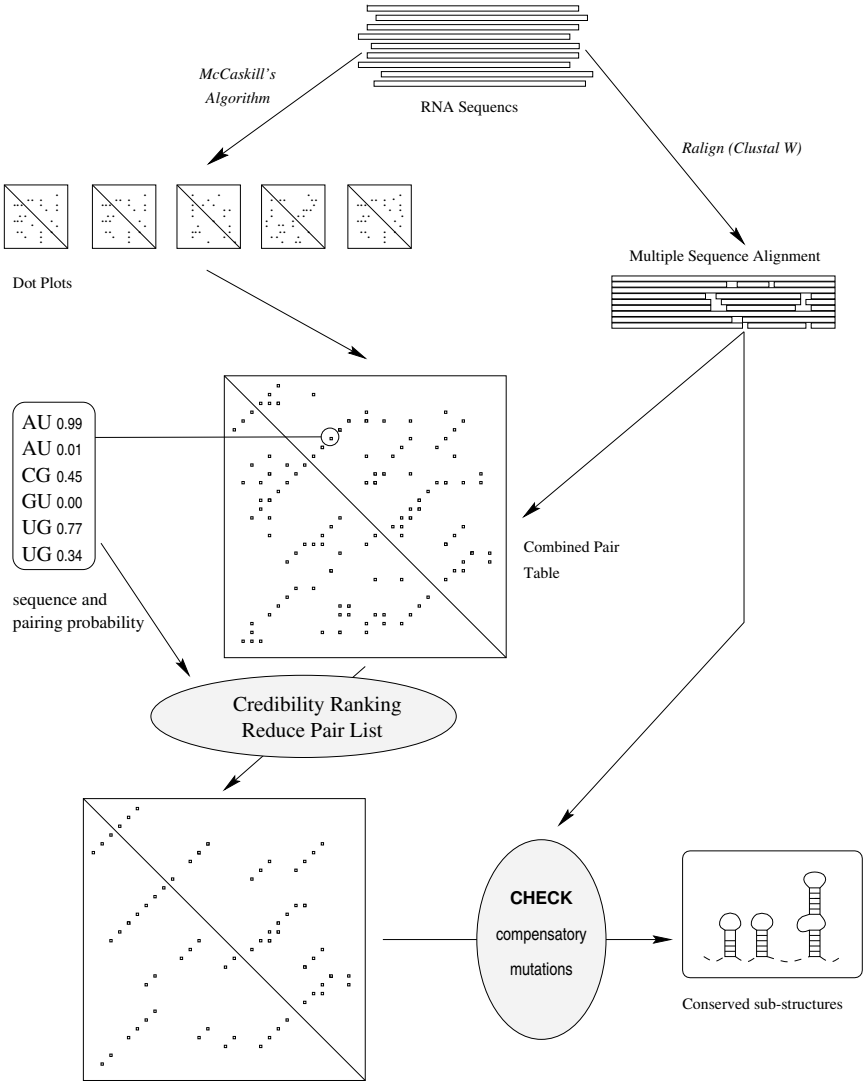
$$(.0408 \cdot .182) + (.0098 \cdot .818) = .021$$



Existing Programs I: `alidot` & `pfrali`

- Use `Vienna RNA Package` for predicting secondary structures based on thermodynamic rules.
- Combine structure prediction and with a standard (`clustalw`) sequence alignment.
- Use covariances to rank order base pairs from the prediction and extract predicted conserved structures

Best suited for large sequences with interspersed conserved structure motifs.



Flow diagram of `alidot`

AliFold

Standard dynamic programming with covariance score as bonus energy:

$$E_c(S, \Psi) = E(S, \Psi) + cv \cdot \sum_{(i,j)\in\Psi} B_{ij}$$

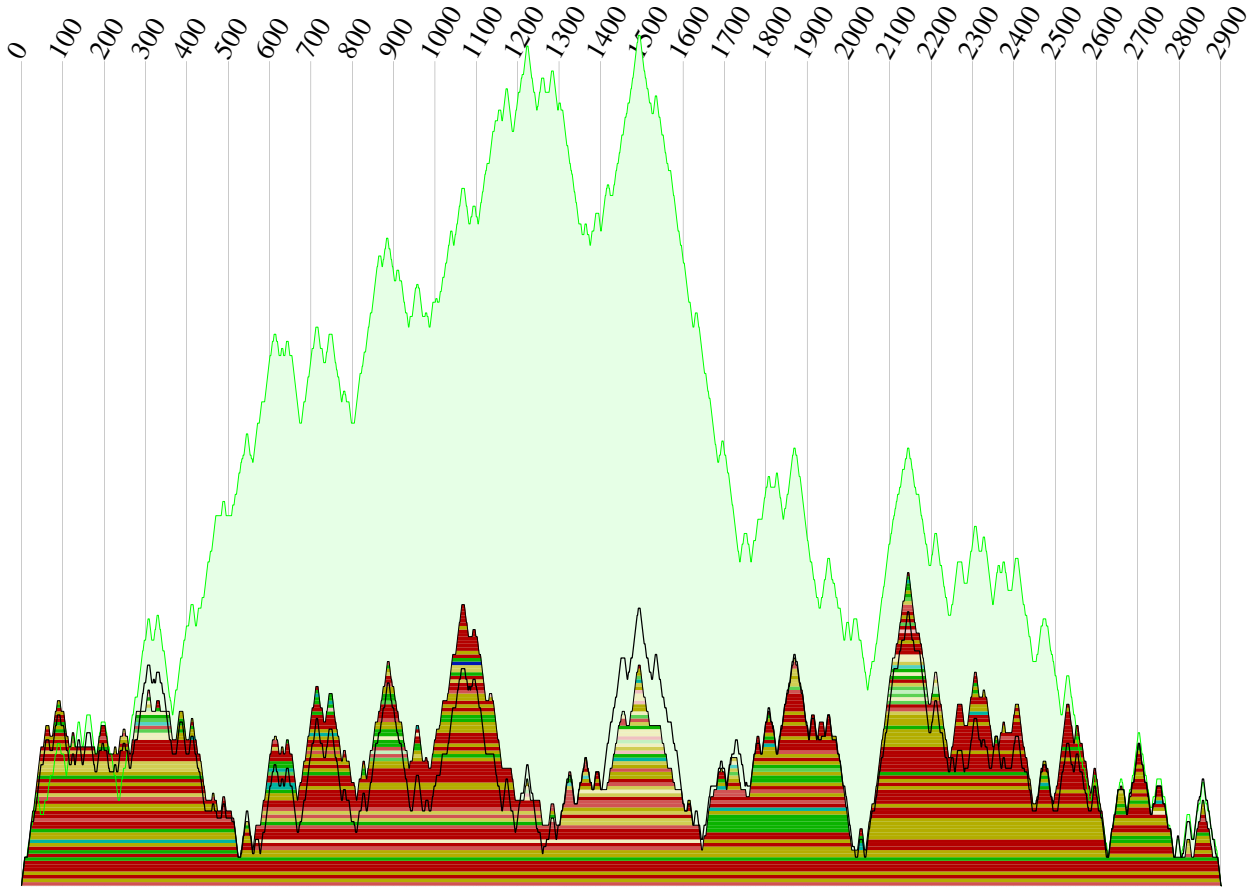
mfe and partition function algorithm implemented in Vienna RNA package.

Correctly predicted base pairs for 16S rRNA for E. Coli.

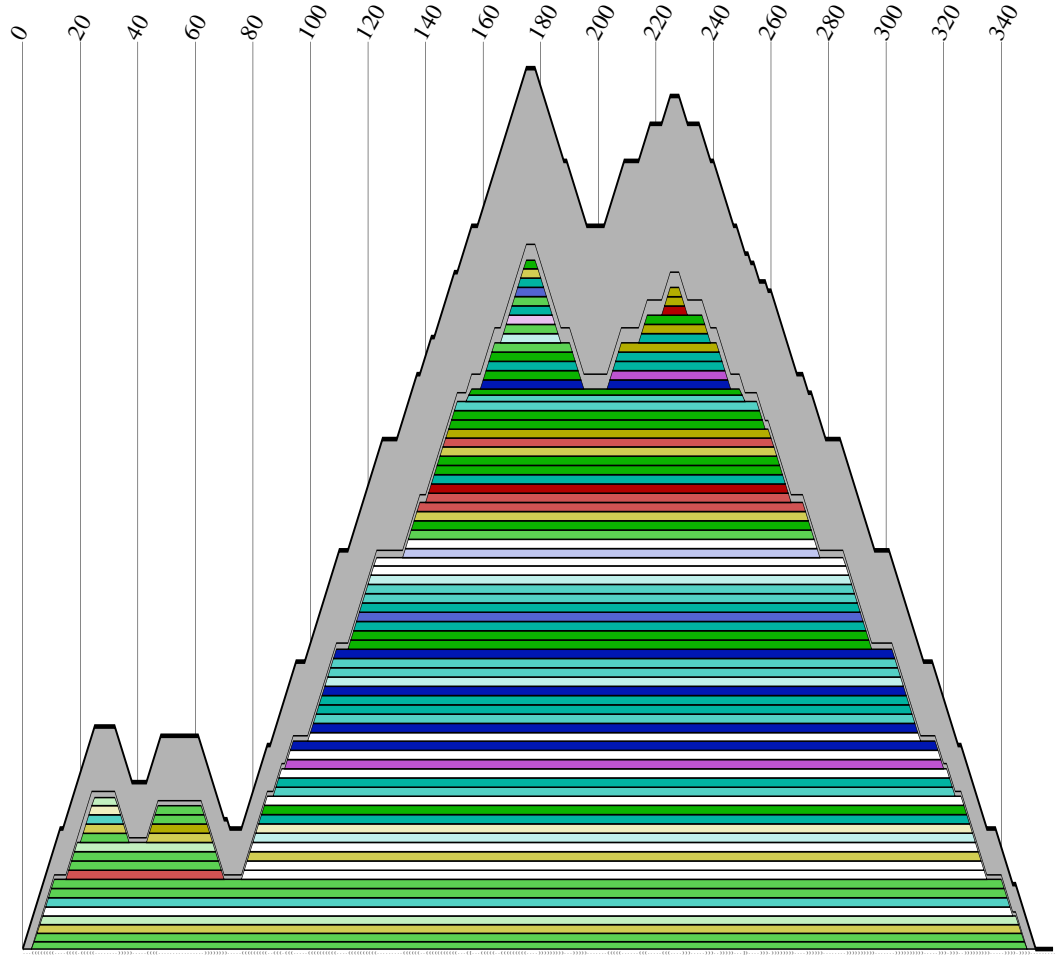
Alignment: *Ribosomal Database Project* [Maidak et al., NAR 28: 173-174 (2000)]

	Clustal W		RDB		Clustal W		RDB	
N	raw	filled	raw	filled	raw	filled	raw	filled
	E.coli 16sRNA				23sRNA			
1	47.2	N/A	47.2	N/A	52.2	N/A	52.2	N/A
2	64.7	67.1	73.8	73.4	71.0	69.4	83.7	82.6
3	74.1	77.2	78.1	79.9	71.2	73.7	85.3	84.9
5	74.5	81.2	85.2	86.6	76.2	82.4	86.6	86.8
9	74.1	82.1	85.9	88.6	74.6	82.6	86.1	86.2

Example: E.coli 23S RNA from 5 sequences

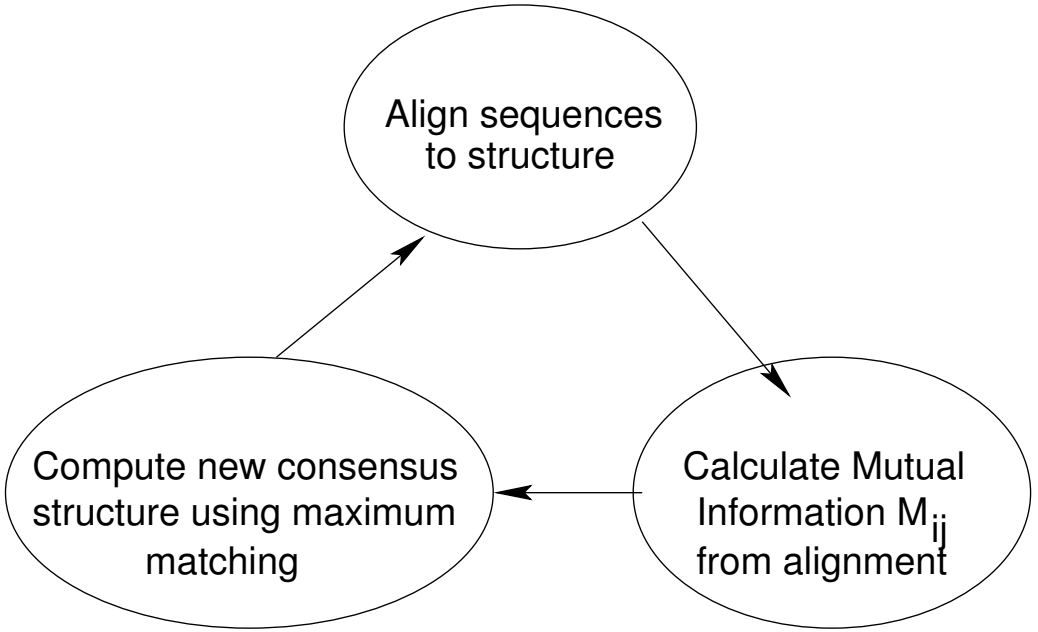


Consensus structure of 14 SRP RNA



Eddy's COVE

Implements a method to align a set of sequences to a secondary structure. Combined with simple maximum matching algorithm to iteratively improve alignment and consensus structure.



Rivas & Eddy's QRNA

Given an alignment A of two sequences, decide whether it's a coding region, structural RNA, or neither.

For the structural RNA case compute the sum over all RNA structure s

$$P(A|RNA) = \sum_s P(A|s, RNA)P(s|RNA)$$

For any of the three models Bayes rule gives

$$P(Model|A) = P(A|Model)P(Model)/P(A)$$

References

Eddy:94 S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucl. Acids. Res.*, 22:2079–2088, 1994.

B. Gulko and D. Haussler. Using multiple alignents and phylogenetic trees to detect RNA secondary structures. *Pac. Symp. Biocomput.*, pages 350–367, 1996.

I. L. Hofacker, M. Fekete, C. Flamm, M. A. Huynen, S. Rauscher, P. E. Stolorz, and P. F. Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.*, 26:3825–3836, 1998.

I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 2002. submitted, SFI Preprint 01-11-067.

I. L. Hofacker and P. F. Stadler. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. & Chem.*, 23:401–414, 1999.

E. Rivas and S. R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16:583–605, 2000.

E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(8):19 pages, 2001.

E. Rivas, R. J. Klein, T. A. Jones, and S. R. Eddy. Computational identification of noncoding RNAs in *e. coli* by comparative genomics. *Curr. Biol.*, 11:1369–1373, 2001.