Comparative approach to discovering ncRNA

Project Question

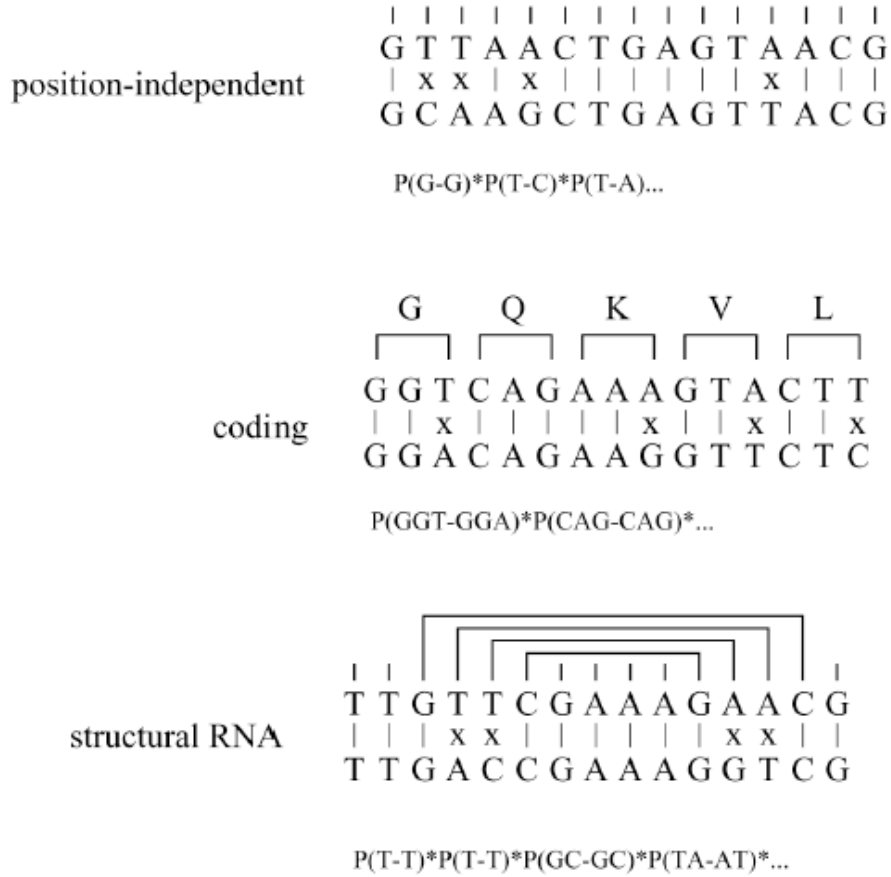
- Akutsu's algorithm gives a recursive formulation for pseudoknots.
- Sequence-structure alignments enable structure prediction through comparison to a homologous RNA (non-pseudoknotted) structure.
- Can sequence-structure alignments be extended to handle pseudoknots?

Comparative prediction of ncRNA

- In genome-genome comparisons, many sequences are found to be conserved.
- Can you use the pattern of conservation to detect if these are ncRNA sequences?
- *QRNA* is a software to do that.

QRNA: Approach

- Compute the 3 probabilities
  - Pr($\overline{XY}$|COD)
  - Pr($\overline{XY}$|RNA)
  - Pr($\overline{XY}$|OTH)



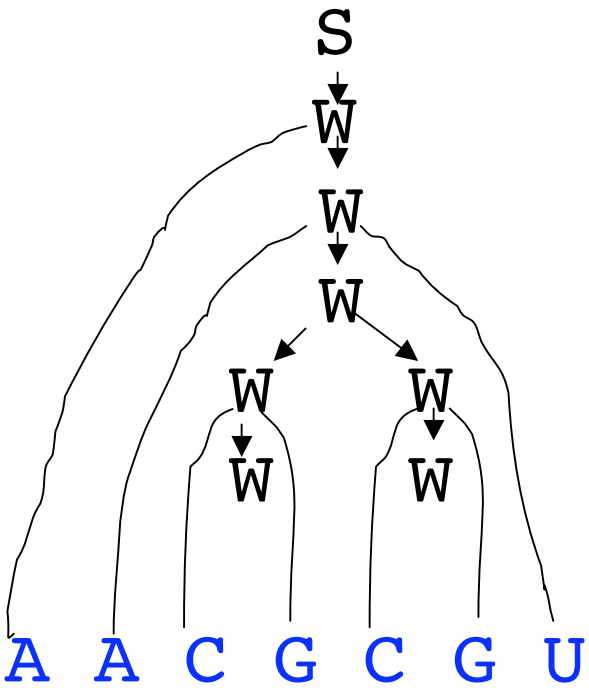
$\Pr(\overline{XY}|RNA)$

- $\Pr(\overline{XY}|RNA) = \sum_s \Pr(\overline{XY}|s,RNA)\Pr(s|RNA)$
- While there are many structures, this expression can be computed efficiently.
- We start by describing a different formalism for computing RNA structure.
- We will show that the probabilistic, and energy frameworks are essentially equivalent

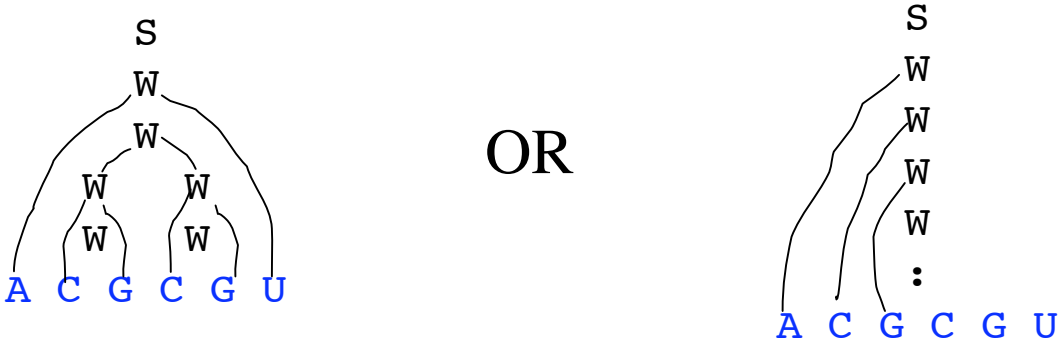
Stochastic Context Free Grammars

- Consider the following CFG:
  - S -> W (* Start *)
  - W -> x W y x,y ∈ {A,C,G,U} (* base-pairing *)
  - W -> x W (* unpaired bases *)
  - W -> WW (* branching *)
  - W -> ε (* termination *)
- The CFG generates RNA sequences, with an associated structure.

Example



Computing RNA structures



- Consider the inverse problem: Given an RNA string, find the best parse (a sequence of Context Free rules that generate the sequence).
- This is equivalent to computing structure.

Computing the optimum parse

Stochastic Context Free Grammars

- Associate a probability with each rule. Hence, SCFG.
  - Pr(S -> W) (* Start *)
  - Pr(W -> x W y) x,y ∈ {A,C,G,U} (* base-pairing *)
  - Pr(W -> x W) (* unpaired bases *)
  - Pr(W -> WW) (* branching *)
  - Pr(W -> ε) (* termination *)
- Let γ_{ij} be the probability that the RNA subsequence s[i..j] was generated by the SCFG.
  - γ_{ij} = Pr(s[i..j] | SCFG) = $\sum_{\pi}$ Pr(s[i..j] | π,SCFG) Pr(π |SCFG)
- It is sufficient to compute γ_{ij} for all i,j.

Computing $\beta_{i,j}$

- $\beta_{i,j}$ = Pr (W -> s[i] W s[j]) $\beta_{i+1,j-1}$
  + Pr (W -> s[i] W) $\beta_{i+1,j}$
  + Pr (W -> W s[j]) $\beta_{i,j-1}$
  + $\sum_k$ Pr (W -> WW) $\beta_{i,k-1}$ $\beta_{k,j}$
- Computing the most likely parse:
- $v_{i,j}$ = max { Pr (W -> s[i] W s[j]) $v_{i+1,j-1}$,
  Pr (W -> s[i] W) $v_{i+1,j}$
  Pr (W -> W s[j]) $v_{i,j-1}$
  $\max_k$ Pr (W -> WW) $v_{i,k-1}$ $v_{k,j}$
  }

SCFGs versus Energy minimization

- The two approaches (most likely parse and energy minimization) give equivalent answers.
- The full-likelihood function (Π) might sometimes be more meaningful from the max-likelihood parse. It helps answer the question: is the string s an RNA sequence?
- The probabilistic approach makes it easier to train parameters (using Bayesian methods).

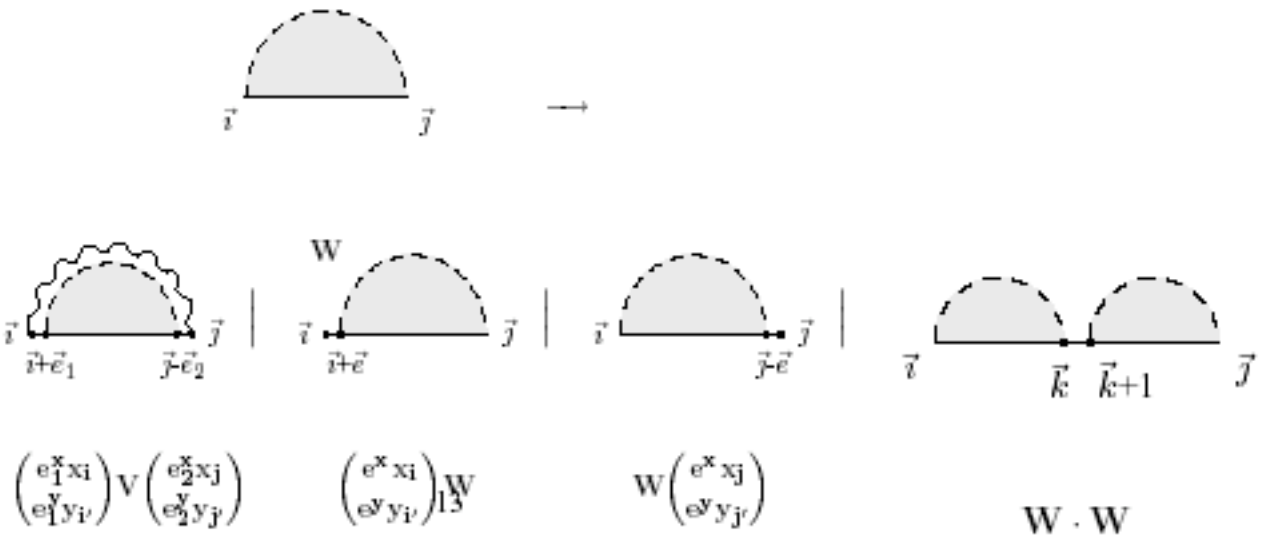
Probability of RNA alignments

- How can we compute

$\Pr(\overline{XY} \mid \text{RNA})$

$= \sum_{\pi} \Pr(\overline{XY} \mid \pi, \text{RNA}) \Pr(\pi \mid \text{RNA})$?

Probability of an RNA alignment



- $\gamma_{i,j}$ = Pr (W -> A[i] W A[j]) $\gamma_{i+1,j-1}$
  + Pr (W -> A[i] W) $\gamma_{i+1,j}$
  + Pr (W -> W A[j]) $\gamma_{i,j-1}$
  + $\sum_k$ Pr (W -> WW) $\gamma_{i,k-1}$ $\gamma_{k,j}$

Computing RNA emission probabilities

$$
\begin{aligned}
P^{\mathrm{RNA}}(a_L a_R b_L b_R \,|\, t) &= P(b_R \,|\, a_L a_R b_L t) P(a_L a_R b_L \,|\, t), \\
&= P(b_R \,|\, a_L a_R b_L t) P(a_R \,|\, a_L b_L t) P^{\mathrm{OTH}}(a_L b_L \,|\, t), \\
&\simeq P^{\mathtt{pair}}(b_R \,|\, b_L t) P^{\mathtt{pair}}(a_R \,|\, a_L t) P^{\mathrm{OTH}}(a_L b_L \,|\, t), \\
&= \frac{P^{\mathtt{pair}}(b_L b_R \,|\, t) P^{\mathtt{pair}}(a_L a_R \,|\, t) P^{\mathrm{OTH}}(a_L b_L \,|\, t)}{P(b_L \,|\, t) P(a_L \,|\, t)}.
\end{aligned}
$$

Other models in QRNA

$$P(\overline{XY}|\mathrm{COD}) = \sum_f P(\overline{XY}|f,\mathrm{COD})P(f|\mathrm{COD}),$$

$$P^{\mathrm{COD}}(a_1a_2a_3, b_1b_2b_3 \,|\, t) \simeq \sum_{A,B} P(a_1a_2a_3 \,|\, A)P(b_1b_2b_3 \,|\, B)P(A,B \,|\, t),$$

Is the sequence RNA, coding or OTH?

- Pr(XY | Model) can be computed for the 3 models (RNA, COD, OTH)
- $Pr(Model_i \mid XY) = P(XY \mid model_i)\, P(model_i)/P(XY)$
- $P(XY) = \sum_j P(XY \mid model_j)\, P(model_j)$

QRNA results

- Multiple alignment of 63 Eukaryotic SRP-RNAs, and 52 RNaseP RNA
  - Use pair-wise alignments from the structural alignment
    - Alignments are classified according to sequence diversity
  - Use each sequence as query to Blast against other family members
- Sensitivity: fraction pairs predicted to be RNA
- Specificity:1-(fraction predicted to be RNA after shuffling)

Sensitivity and Specificity

	# align	% sensitivity	% specificity
% ID			
0 < 10	140	42.8 (60)	100.0 (0)
10 < 20	827	59.6 (493)	100.0 (0)
20 < 30	503	71.4 (359)	100.0 (0)
30 < 40	764	75.1 (574)	100.0 (0)
40 < 50	283	58.6 (166)	100.0 (0)
50 < 60	434	81.3 (353)	100.0 (0)
60 < 70	88	80.7 (71)	100.0 (0)
70 < 80	70	91.4 (64)	97.1 (2)
80 < 90	73	97.3 (71)	79.4 (15)
90 < 100	61	93.4 (57)	27.9 (44)
100	99	93.9 (93)	29.3 (70)
% CC			

QRNA results: experiment 2

- Each of the sequences was chosen in turn, and compared against members of its own family (WU-Blastn2).
  - Poor quality of alignments
  - Bias towards conserved sequences
  - 1003 (out of 3342 pairs) alignments were selected

Table 3: Similar analysis to the one presented in Table 2 for 586 BLASTN alignments of SRP RNAs and 417 BLASTN alignments of RNaseP RNAs.

	# alignments	% sensitivity	% specificity
% ID			
60 < 70	419	15.3 (64)	99.5 (2)
70 < 80	269	26.8 (72)	98.5 (4)
80 < 90	131	61.1 (80)	89.5 (19)
90 < 100	78	97.4 (76)	67.9 (53)
100	106	92.4 (98)	24.5 (80)
% GC			

QRNA: Results

- Comparison of E. coli and S. typhii
- E. coli was partitioned into 115 RNA, 4290 ORFs, and 2367 intergenic
- Each region blasted against S. typhii, and QRNA was used on "quality" alignments
  - 354 alignments to RNA
  - 4946 to ORFs
  - 11509 alignments to intergenic regions (Repeats?)

Genomic comparison Results

ncRNA	148 regions annotated	33 RNA 115 OTH	
ORF	7422	88 RNA 3397 COD 3937 OTH	
intergenic	1974	351 RNA 61 COD 1562 OTH	

Conclusions

- Blastn does not produce good alignments from a structural viewpoint.
- Can we use paired SCFGs to redo the alignment and the structure? In principle, yes, but it is expensive.
- Rivas and Eddy did not use a true comparison of orthologs. Would that help?

Computing Structural alignments

For all intervals (i,j) in $s_{1..n}$, and (k,l) in $t_{1..m}$

$$S(i,j,k,l) = \max \begin{cases} S(i+1, j-1, k+1, l-1) + \sigma(i,j,k,l) \\ S(i+1, j, k, l) + \sigma(s[i], -) \\ S(i+1, j, k+1, l) + \sigma(s[i], t[j]) \\ \vdots \\ \max_{j',l'} \{ S(i, j'-1, k, l'-1) + S(j', k, k', k) \} \end{cases}$$

Project Question

- Can you improve upon QRNA with the following:
  - Structural alignments to obtain better results
  - Filtering to make search efficient. Most pairs should be discarded without computing a structural alignment.

RNA multiple alignments

- Why should we compute multiple (structural) alignments for RNA?

Structural Alignment

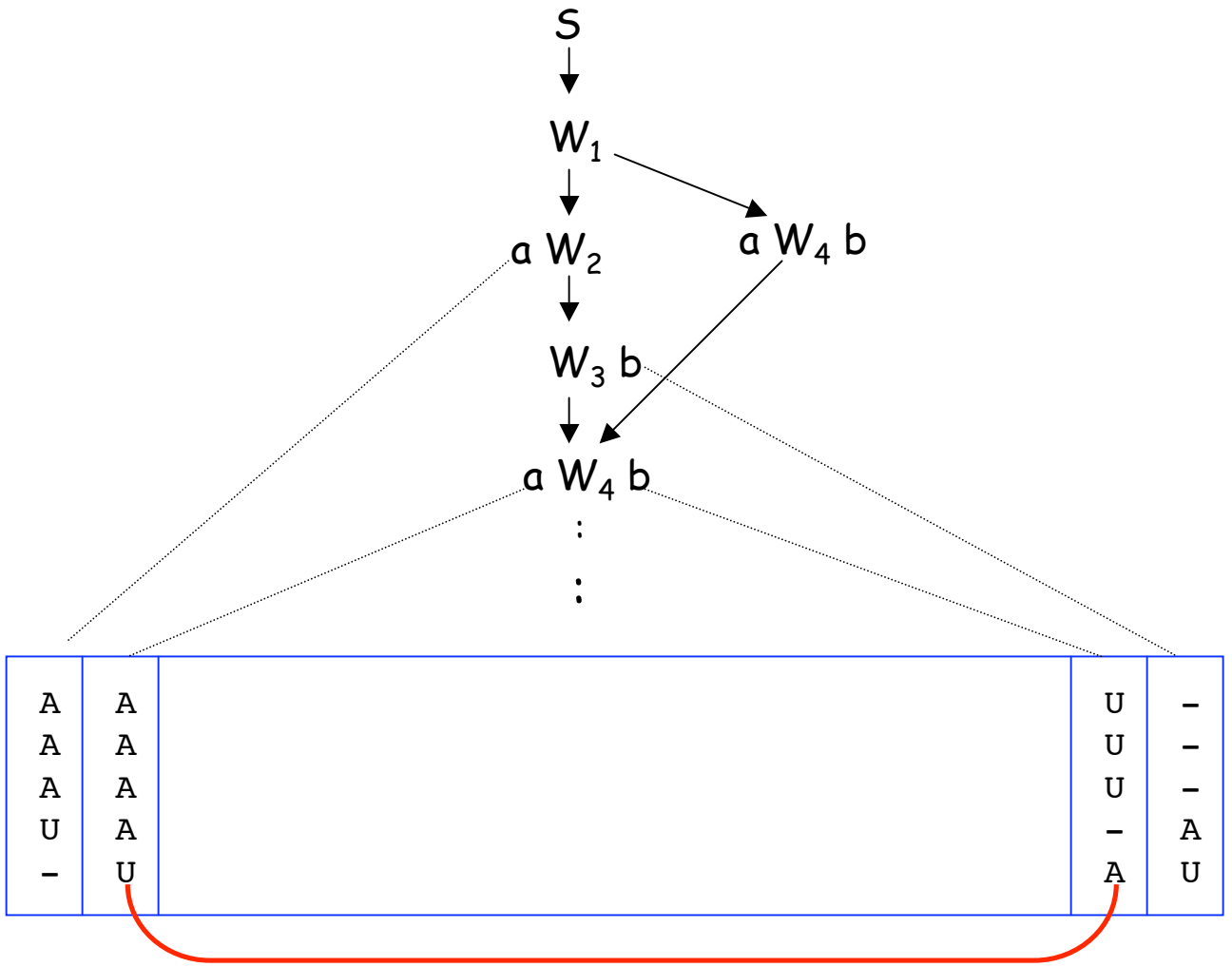
```
X07545      ..ACCCGGC.CAUA...GUGGCCG.GGCAA.CAC.CCGG.U.C..UCGUU
M21086      ..ACCCGGC.CAUA...GCGGCCG.GGCAA.CAC.CCGG.A.C..UCAUG
X05870      ..ACCCGGC.CACA...GUGAGCG.GGCAA.CAC.CCGG.A.C..UCAUU
U05019      ..ACCCGGU.CAUA...GUGAGCG.GGUAA.CAC.CCGG.A.C..UCGUU
M16530      ..ACCCGGC.AAUA...GGCGCCGGUGCUA.CGC.CCGG.U.C..UCUUC
X01588      ..ACCCGGU.CACA...GUGAGCG.GGCAA.CAC.CCGG.A.C..UCAUU
AF034619    ...GGCGGC.CACA...GCGGUGG.GGUUGCCUC.CCGU.A.C..CCAUC
L27170     AGUGGUGGC.CAUA...UCGGCGG.GGUUC.CUCCCCGU.A.C..CCAUC
X05532     AGGAACGGC.CAUA...CCACGUC.GAUCG.CAC.CACA.U.C..CCGUC
#=GC       <<<<<<<<<<.......<<.<<<<.<...<.<...<<<<.<.<.......
```

Conserved sequences, and conserved structure are more apparent in multiple alignments.

Computing Structural Alignments

- Analogy: In sequence alignment, the score for aligning a column is position independent.
- In profiles, or HMMs, position specific scoring is used to distinguish conserved positions from non-conserved positions
- Similar ideas can be used for RNA.

Covariance models=RNA profiles



Aligning a sequence to a covariance model

- We align each node of the covariance model (it is tree like, but may be a graph).
- The alignment score follows the same recurrence as in Lecture 7, but with position specific probabilities.
- Example:
  - $A[W_i,(i,j)] = -\log (Pr[W_i \rightarrow s[i]\ W_j\ s[j]\ ) + A[W_j,(i+1,j-1)]$
- If we wish to compute the probability that a sequence belongs to a family, we compute the total likelihood (sum over all probabilities)
- If we wish to compute the structure of an unknown sequence by comparison to a covariance model, we compute the max likelihood parse in this graph.

Covariance models and ncRNA discovery

- Given a family of ncRNA sequences, scan a genomic sequence with a covariance model and retrieve all high scoring sub-sequences.
- This is the most common method, but it is expensive.
- Assume covariance model has m states, and the substring has at most n symbols, and the database has L symbols.
- Alignment cost = $O(n^2m_1+n^3m_2)$
- Total time =?

Computing covariance models

- If we are given a CM, a multiple structural alignment is 'easy'.
  - In turn, align each sequence to the CM.
- If we are given a multiple alignment, computing the covariance model is easy
- For simultaneous prediction, a Bayesian iterative approach is used
  - Compute a seed alignment
  - Use the alignment to compute a CM
  - Use the CM to compute a new alignment
  - Iterate

Project

- Compute an RNA multiple alignment.
- Existing methods do not work well without good seed alignment, and require excessive hand curation.