Applications of Lightweight Stochastic Context Free Grammars for RNA Analysis

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Mark Craven

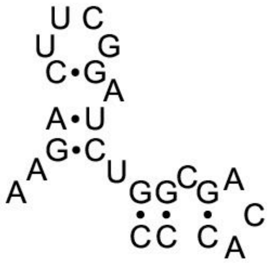
craven@biostat.wisc.edu

Spring 2009

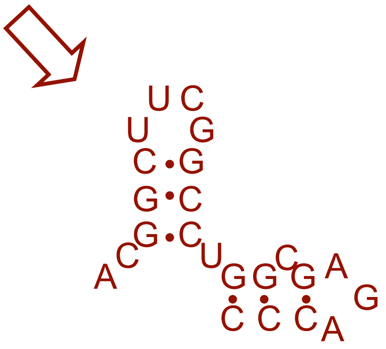
Searching Sequence for a Secondary Structure

Given

- a single RNA sequence with its secondary structure
- another RNA query sequence



ACGGCUUCGGCCUGGCGAGACCC

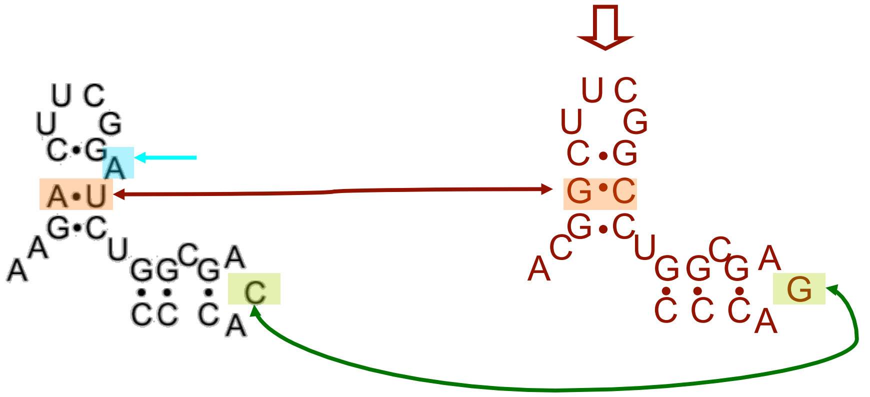


Determine if the query sequence has "same" secondary structure

Searching Sequence for a Secondary Structure

- this is analogous to pairwise alignment with primary sequences
- we take into account substitutions, insertions/deletions, and base-pair substitutions

ACGGCUUCGGCCUGGCGAGACCC

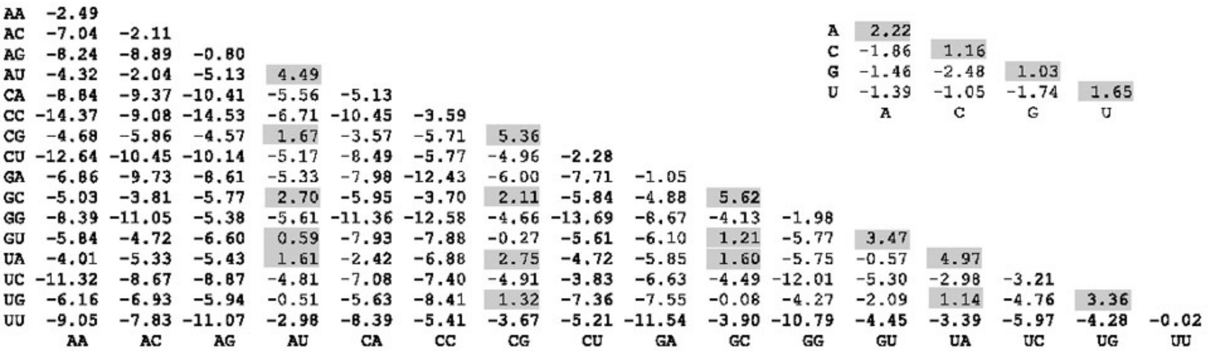


The RIBOSUM Matrices [Klein & Eddy]

observed frequency of *i* aligned to *j* in homologous RNAs

$$s_{ij} = \log_2 \frac{f_{ij}}{g_i g_j}$$

background frequency of *i*



	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	-2.49															
AC	-7.04	-2.11														
AG	-8.24	-8.89	-0.80													
AU	-4.32	-2.04	-5.13	4.49												
CA	-8.84	-9.37	-10.41	-5.56	-5.13											
CC	-14.37	-9.08	-14.53	-6.71	-10.45	-3.59										
CG	-4.68	-5.86	-4.57	1.67	-3.57	-5.71	5.36									
CU	-12.64	-10.45	-10.14	-5.17	-8.49	-5.77	-4.96	-2.28								
GA	-6.86	-9.73	-8.61	-5.33	-7.98	-12.43	-6.00	-7.71	-1.05							
GC	-5.03	-3.81	-5.77	2.70	-5.95	-3.70	2.11	-5.84	-4.88	5.62						
GG	-8.39	-11.05	-5.38	-5.61	-11.36	-12.58	-4.66	-13.69	-8.67	-4.13	-1.98					
GU	-5.84	-4.72	-6.60	0.59	-7.93	-7.88	-0.27	-5.61	-6.10	1.21	-5.77	3.47				
UA	-4.01	-5.33	-5.43	1.61	-2.42	-6.88	2.75	-4.72	-5.85	1.60	-5.75	-0.57	4.97			
UC	-11.32	-8.67	-8.87	-4.81	-7.08	-7.40	-4.91	-3.83	-6.63	-4.49	-12.01	-5.30	-2.98	-3.21		
UG	-6.16	-6.93	-5.94	-0.51	-5.63	-8.41	1.32	-7.36	-7.55	-0.08	-4.27	-2.09	1.14	-4.76	3.36	
UU	-9.05	-7.83	-11.07	-2.98	-8.39	-5.41	-3.67	-5.21	-11.54	-3.90	-10.79	-4.45	-3.39	-5.97	-4.28	-0.02

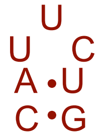
	A	C	G	U
A	2.22			
C	-1.86	1.16		
G	-1.46	-2.48	1.03	
U	-1.39	-1.05	-1.74	1.65

$$s'_{ij\,kl} = \log_2 \frac{f'_{ij\,kl}}{g_i g_j g_k g_l}$$

observed frequency of two base pairs *i-j* and *k-l* aligned to each other in homologous RNAs

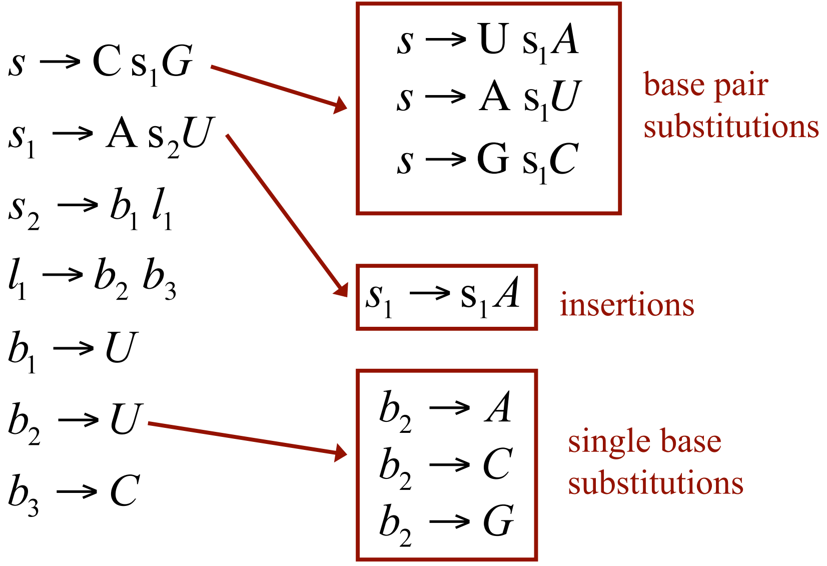
Using a Lightweight SCFG to Search for Secondary Structure

given a structure



can construct a simple grammar characterizing it

can add productions to allow for variation



$s \rightarrow C\ s_1 G$

$s_1 \rightarrow A\ s_2 U$

$s_2 \rightarrow b_1\ l_1$

$l_1 \rightarrow b_2\ b_3$

$b_1 \rightarrow U$

$b_2 \rightarrow U$

$b_3 \rightarrow C$

$s \rightarrow U\ s_1 A$

$s \rightarrow A\ s_1 U$

$s \rightarrow G\ s_1 C$

base pair substitutions

$s_1 \rightarrow s_1 A$

insertions

$b_2 \rightarrow A$

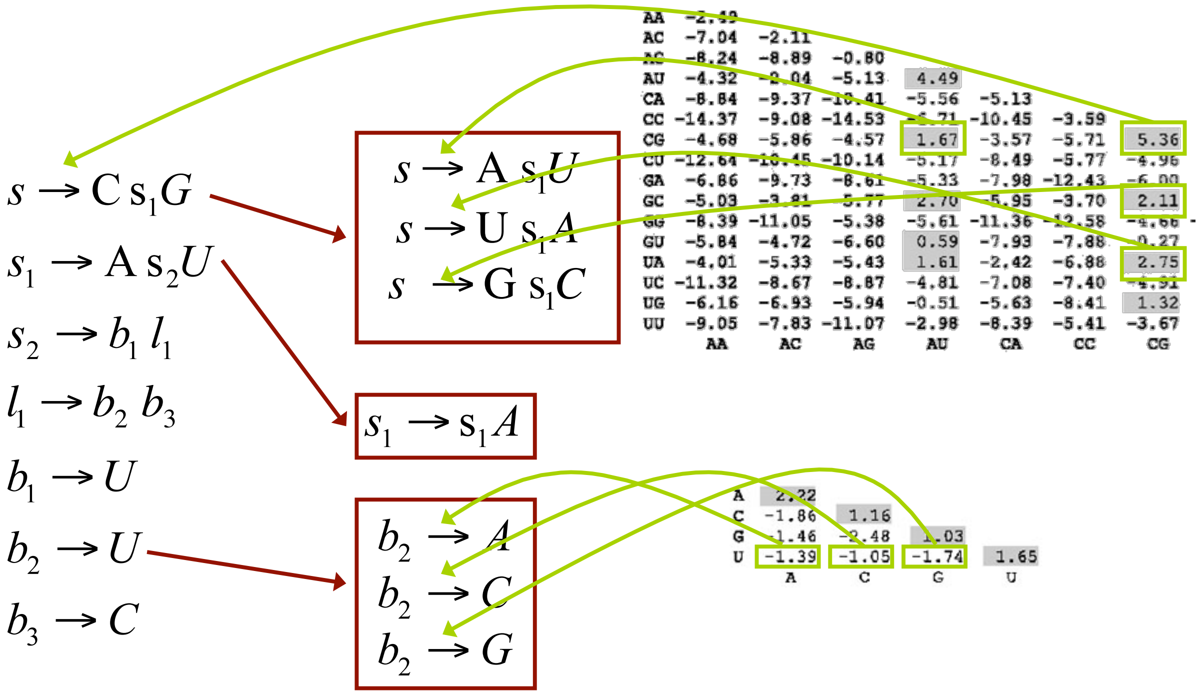
$b_2 \rightarrow C$

$b_2 \rightarrow G$

single base substitutions

Setting the Parameters in the Grammar

- Infer them from the parameters from the RIBOSUM matrices (taking into account the latter are log-odds scores)



$s \rightarrow C\ s_1 G$

$s_1 \rightarrow A\ s_2 U$

$s_2 \rightarrow b_1\ l_1$

$l_1 \rightarrow b_2\ b_3$

$b_1 \rightarrow U$

$b_2 \rightarrow U$

$b_3 \rightarrow C$

$s \rightarrow A\ s_1 U$

$s \rightarrow U\ s_1 A$

$s \rightarrow G\ s_1 C$

$s_1 \rightarrow s_1 A$

$b_2 \rightarrow A$

$b_2 \rightarrow C$

$b_2 \rightarrow G$

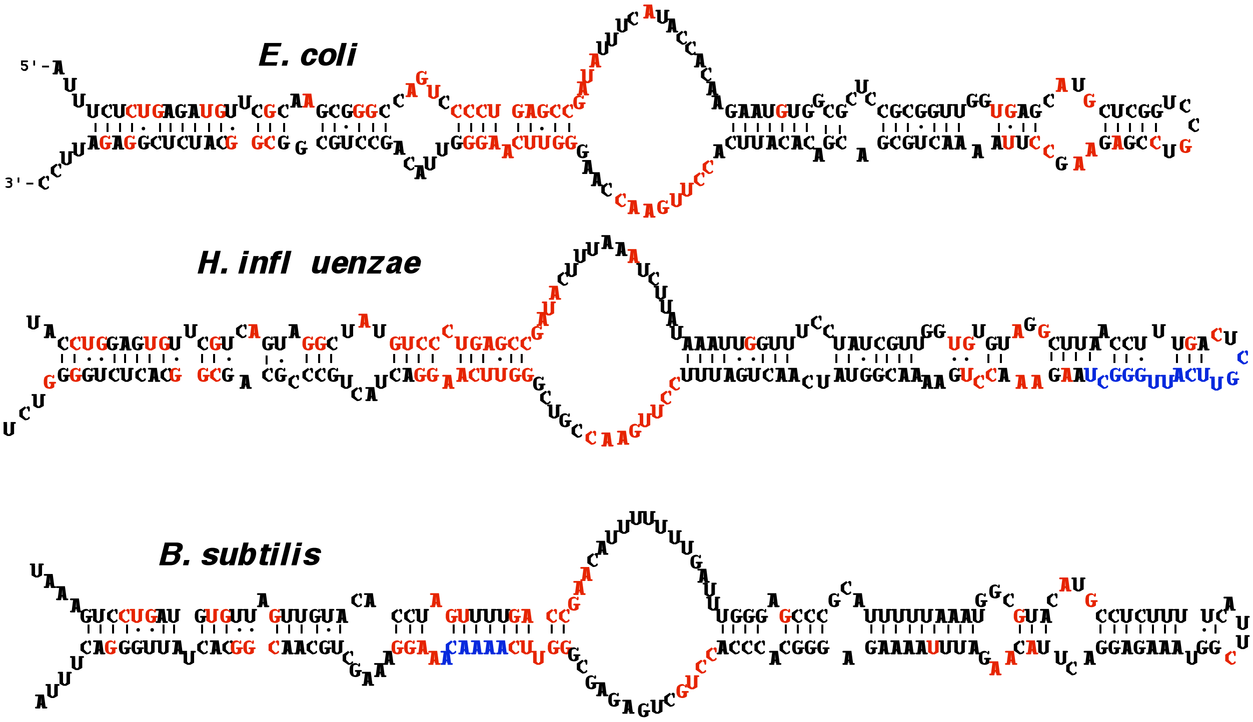
	AA	AC	AG	AU	CA	CC	CG
AA	-2.49						
AC	-7.04	-2.11					
AG	-8.24	-8.89	-0.80				
AU	-4.32	-2.04	-5.13	4.49			
CA	-8.84	-9.37	-10.41	-5.56	-5.13		
CC	-14.37	-9.08	-14.53	-6.71	-10.45	-3.59	
CG	-4.68	-5.86	-4.57	1.67	-3.57	-5.71	5.36
CU	-12.64	-10.45	-10.14	-5.17	-8.49	-5.77	-4.96
GA	-6.86	-9.73	-8.61	-5.33	-7.98	-12.43	-6.00
GC	-5.03	-3.81	-5.77	2.70	-5.95	-3.70	2.11
GG	-8.39	-11.05	-5.38	-5.61	-11.36	-12.58	-4.66
GU	-5.84	-4.72	-6.60	0.59	-7.93	-7.88	-0.27
UA	-4.01	-5.33	-5.43	1.61	-2.42	-6.88	2.75
UC	-11.32	-8.67	-8.87	-4.81	-7.08	-7.40	-4.91
UG	-6.16	-6.93	-5.94	-0.51	-5.63	-8.41	1.32
UU	-9.05	-7.83	-11.07	-2.98	-8.39	-5.41	-3.67

	A	C	G	U
A	2.22			
C	-1.86	1.16		
G	-1.46	-2.48	1.03	
U	-1.39	-1.05	-1.74	1.65

RSEARCH: Searching Sequence for a Secondary Structure

- the RSEARCH algorithm [Klein & Eddy, *BMC Bioinformatics* 2003] implements this idea
- but uses a somewhat different SCFG formulation – covariance models (see section 10.3 in Durbin et al.)

6S RNA Secondary Structure



An RSEARCH Case Study

- finding 6S genes in bacterial genomes
  - we used E. coli 6S as the query structure
  - searched 14 other genomes with known 6S genes
    - ~ 5,000 intergenic sequences on average
  - the top-scoring RSEARCH hit in all 14 genomes was the known 6S gene

RNA Gene Detection

[Rivas & Eddy, BMC Bioinformatics 2001]

Given: a pair of putatively homologous sequences

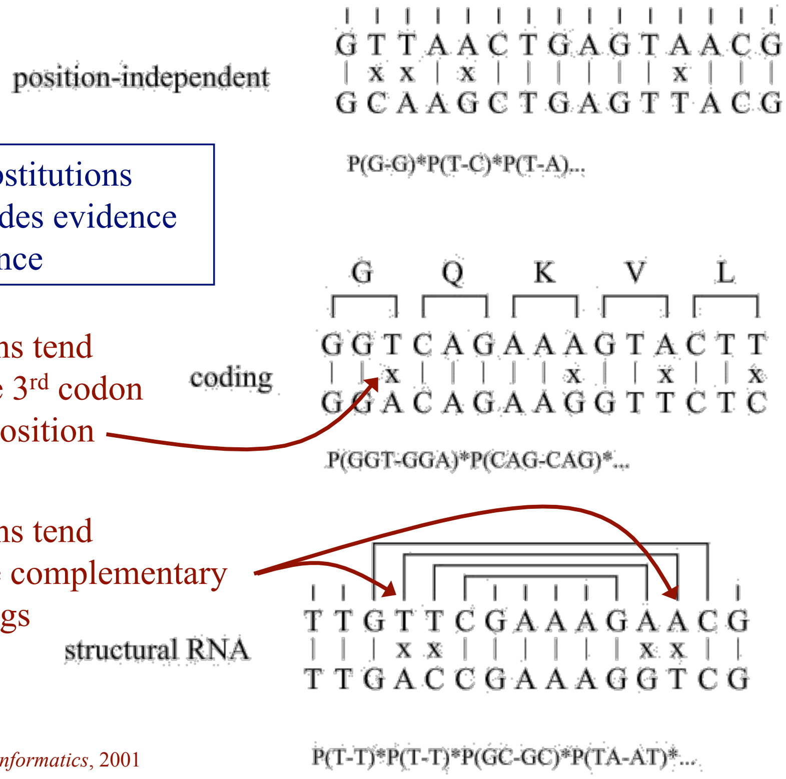
Identify novel RNA genes in the sequences

TAGTCATGCAGTCAGCTATCATCAGCATCGATCGATCGACTAGCTACGTACGACTAGGACTAGCTACGTACGACTAGGACTAGCTACGTACGAACTGACTGACTAGGGGGGGATATTCTCTGGGCCCTCATCTACTGAGCTATCATCATCGTACTA

TCAAACTGACGTACTAGCTAGTCATGCAGTCAGCTATCATCAGCATCGATAACTGACGTACTAGCTAGTCATGCAGTCAGCTATCATCAGCATCGATGCTATCATCAGCATCGATCGATCGACTAGCTACGTACGACTAGGACTAGCTACGTACGAC

RNA Gene Detection

key idea: the pattern of substitutions in the two sequences provides evidence about the role of the sequence


position-independent
G T T A A C T G A G T A A C G
| x x | x | | | | | | | x | | |
G C A A G C T G A G T T A C G
P(G-G)*P(T-C)*P(T-A)...
G Q K V L
coding
G G T C A G A A A G T A C T T
| | x | | | | | | x | | x | | x
G G A C A G A A A G G T T C T C
P(GGT-GGA)*P(CAG-CAG)*...
structural RNA
T T G T T C G A A A G A A C G
| | | x x | | | | | | x x | |
T T G A C C G A A A G G T C G
P(T-T)*P(T-T)*P(GC-GC)*P(TA-AT)*...

substitutions tend to be in the 3rd codon (wobble) position

substitutions tend to preserve complementary base pairings

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001

RNA Gene Detection

- illustrative examples of emission scores for three models
  (numbers before parens are log-odds with respect to a model of no alignment)

OTH	$P^{\text{OTH}}\begin{pmatrix} G \\ \vert \\ G \end{pmatrix}$ +0.76(-3.20)	$P^{\text{OTH}}\begin{pmatrix} C \\ \vert \\ C \end{pmatrix}$ +0.72(-3.52)	$P^{\text{OTH}}\begin{pmatrix} U \\ \vert \\ C \end{pmatrix}$ -0.19(-4.41)	$P^{\text{OTH}}\begin{pmatrix} A \\ \vert \\ U \end{pmatrix}$ -0.53(-4.45)
COD	$P^{\text{COD}}\begin{pmatrix} A\,A\,C \\ \vert\;\vert\;\vert \\ A\,A\,C \end{pmatrix}$ +3.31(-8.19)	$P^{\text{COD}}\begin{pmatrix} A\,A\,C \\ \vert\;\vert\;\vert \\ A\,A\,U \end{pmatrix}$ +3.31(-8.19)	$P^{\text{COD}}\begin{pmatrix} A\,A\,C \\ \vert\;\vert\;\vert \\ A\,U\,C \end{pmatrix}$ -0.52(-12.31)	$P^{\text{COD}}\begin{pmatrix} U\,C\,U \\ \vert\;\vert\;\vert \\ A\,G\,C \end{pmatrix}$ +1.29(-10.95)
RNA	$P^{\text{RNA}}\begin{pmatrix} G \cdot\cdot C \\ \vert \quad \vert \\ G \cdot\cdot C \end{pmatrix}$ +3.81(-4.37)	$P^{\text{RNA}}\begin{pmatrix} G \cdot\cdot U \\ \vert \quad \vert \\ G \cdot\cdot C \end{pmatrix}$ +1.36(-6.82)	$P^{\text{RNA}}\begin{pmatrix} G \cdot\cdot A \\ \vert \quad \vert \\ G \cdot\cdot A \end{pmatrix}$ -8.82(-16.42)	$P^{\text{RNA}}\begin{pmatrix} G \cdot\cdot C \\ \vert \quad \vert \\ C \cdot\cdot G \end{pmatrix}$ +2.43(-5.76)

both code for Asn

both code for Ser

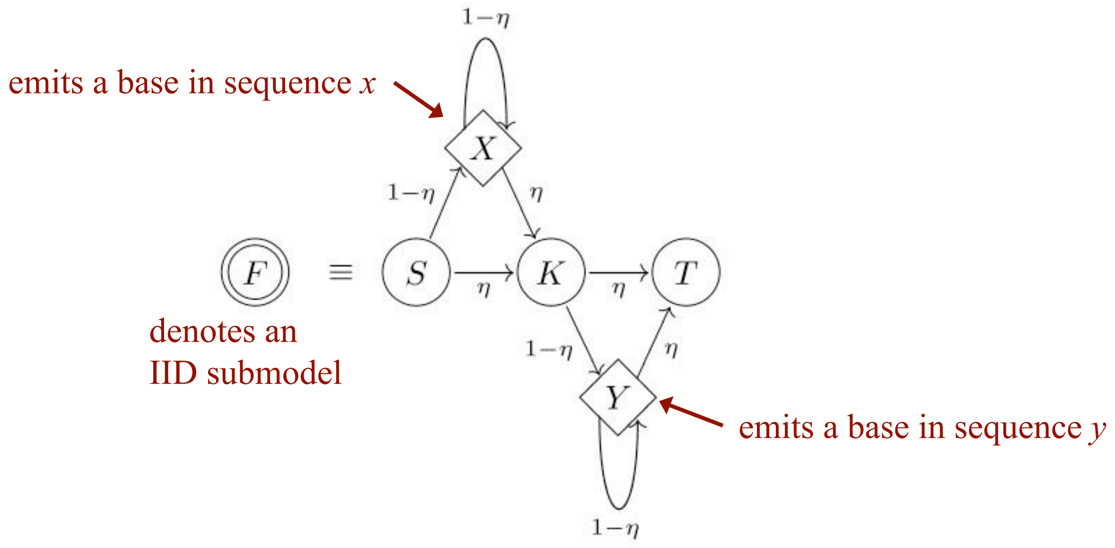
Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001

RNA Gene Detection via Comparative Sequence Analysis

- given sequences *x* and *y*, want a model that can distinguish
  - homologous RNA subsequences
  - homologous coding subsequences
  - "other" homologous subsequences
  - non-homologous subsequences
- allow these to be interleaved, have gaps

RNA Gene Detection: The IID Model

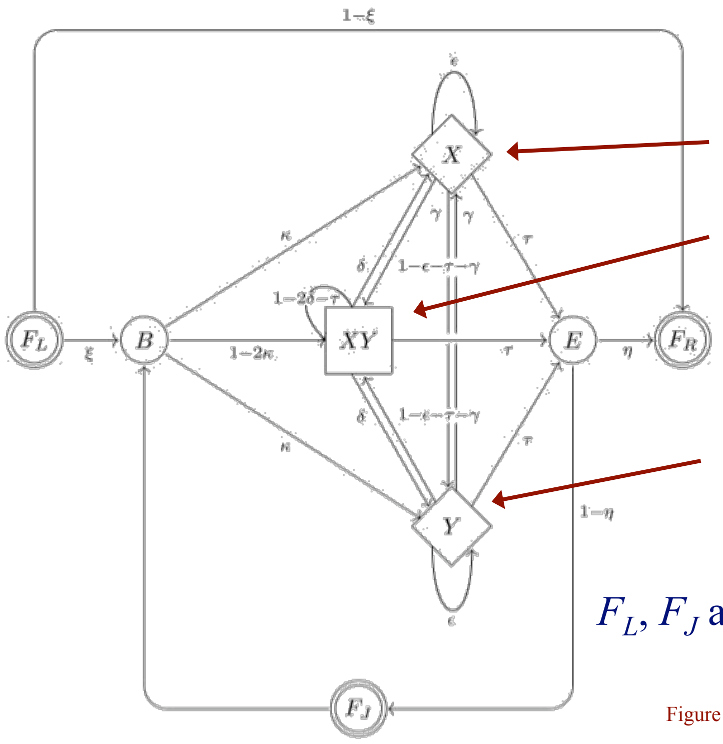
- models non-homologous sequences, *x* and *y*



- *S*, *K* and *T* are silent states

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001

RNA Gene Detection: The "Other" Homologous Sequence Model



emits a base in sequence *x*

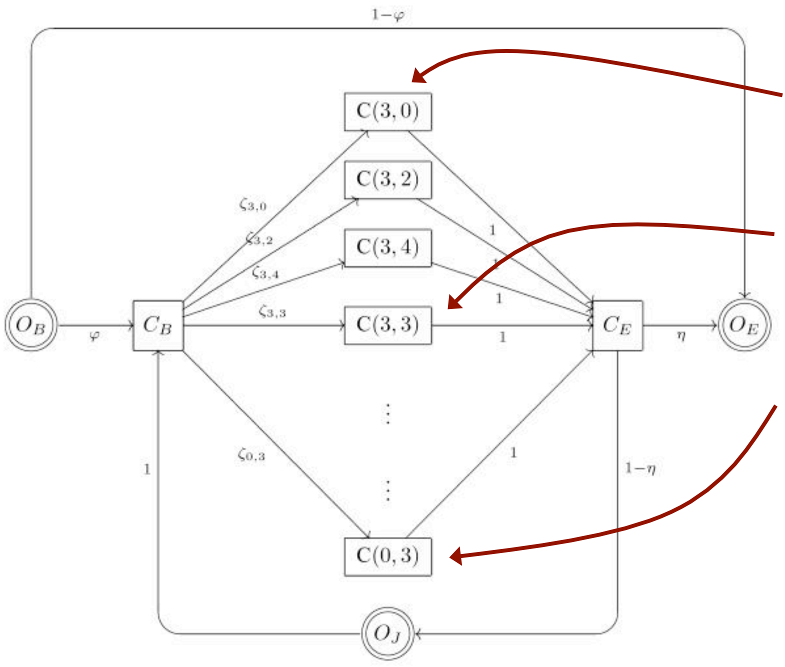
emits bases in *x* and *y*

emits a base in sequence *y*

F_L, F_J and F_R are IID submodels

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001

RNA Gene Detection: The Coding Sequence Model



emits a codon in *x* only

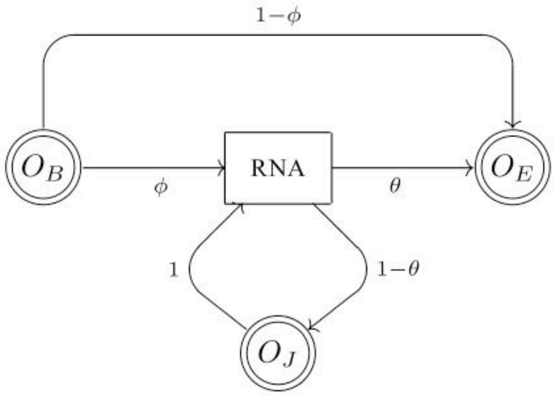
emits codons in *x* and *y*

emits a codon in *y* only

O_B, O_J and O_E are "other" submodels

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001

RNA Gene Detection: The RNA Model



O_B, O_J and O_E are "other" submodels

Figure from Rivas & Eddy, *BMC Bioinformatics*, 2001

- here, the RNA box is a "lightweight" pairwise SCFG

Summary of RNA Analysis Tasks

- given a sequence, predict its secondary structure
- given a set of related RNA sequences, construct a model of the set
  - parameter learning (Inside-Outside)
  - structure refinement
- given a model of an RNA class, find sequences that belong to the class (Inside or CYK)
- given a sequence/structure, find other sequences with similar structure
- given a pair of related genomic sequences, find subsequences that seem have similar secondary structure (RNA gene finding)