More on Stochastic Context Free Grammars for RNA Analysis

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Mark Craven

craven@biostat.wisc.edu

Spring 2011

Goals for Lecture

the key concepts to understand are the following

- approaches to determining the structure of an SCFG grammar
- the task of searching for sequences that “match” a given RNA structure

Where do we get a grammar?

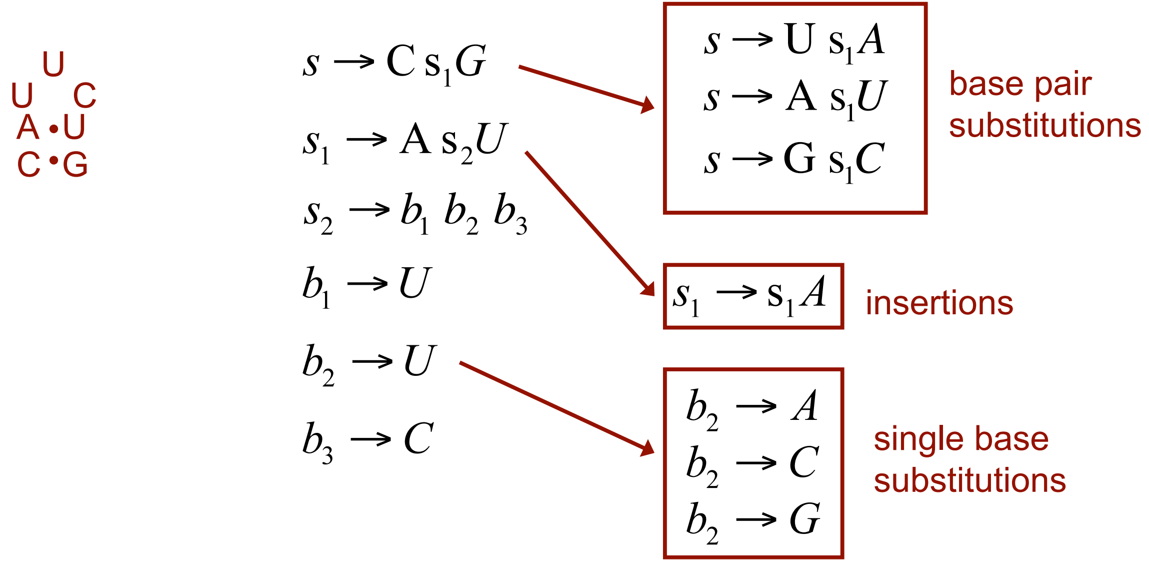
1. from a canonical secondary structure
2. through an iterative, refinement process
3. alternatively, use a simple, generic one

1. Deriving a Grammar from a Secondary Structure

given a structure

can construct a simple grammar characterizing it

can add productions to allow for variation



$s \rightarrow \mathrm{C}\ \mathrm{s}_1 G$

$s_1 \rightarrow \mathrm{A}\ \mathrm{s}_2 U$

$s_2 \rightarrow b_1\ b_2\ b_3$

$b_1 \rightarrow U$

$b_2 \rightarrow U$

$b_3 \rightarrow C$

$s \rightarrow \mathrm{U}\ \mathrm{s}_1 A$

$s \rightarrow \mathrm{A}\ \mathrm{s}_1 U$

$s \rightarrow \mathrm{G}\ \mathrm{s}_1 C$

base pair substitutions

$s_1 \rightarrow \mathrm{s}_1 A$

insertions

$b_2 \rightarrow A$

$b_2 \rightarrow C$

$b_2 \rightarrow G$

single base substitutions

2. Deriving a Grammar Through an Iterative Process

- consider the approach used by Eddy & Durbin to learn an SCFG model of tRNAs

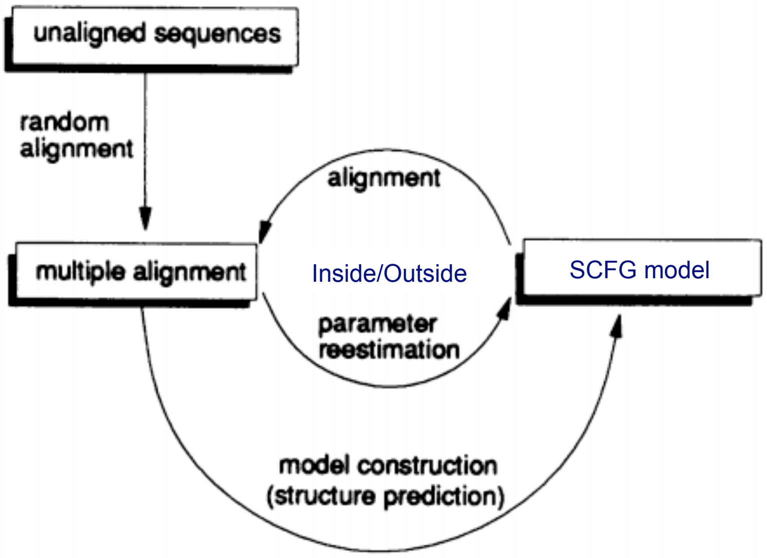
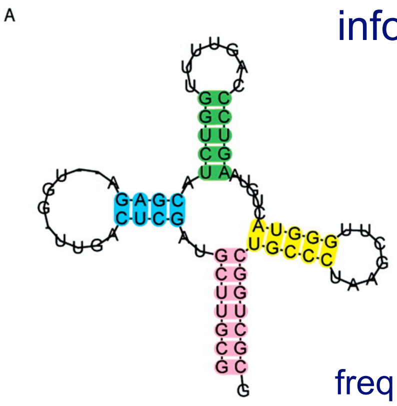


Figure from Eddy & Durbin, *Nucleic Acids Research*, 22(11):2079-2088, 1994.

Eddy & Durbin: Model Construction Step

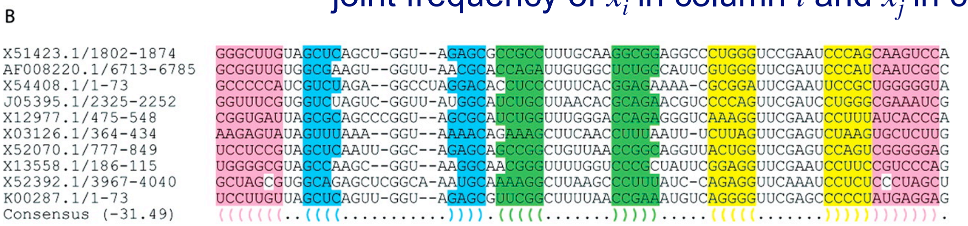
- given multiple alignment, compute mutual information between pairs of positions



$$M(i,j) = \sum_{x_i, x_j} f_{x_i x_j} \log_2 \frac{f_{x_i x_j}}{f_{x_i} f_{x_j}}$$

frequency of x_i in column i

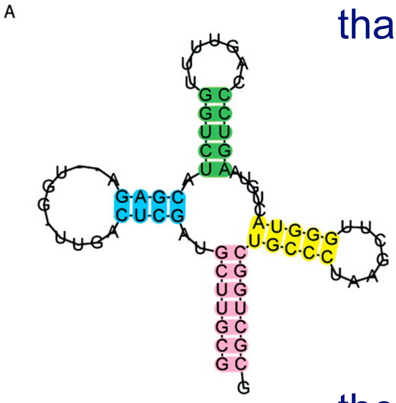
joint frequency of x_i in column i and x_j in column j



```
X51423.1/1802-1874     GGGCUUGUAGCUCAGCU-GGU--AGAGCGCCGCCUUUGCAAGGCGGAGGCCCUGGGUCCGAAUCCCAGCAAGUCCA
AF008220.1/6713-6785   GCGGUUGUGGCGAAGU--GGUU-AACGCACCAGAUUGUGGCUCUGGCAUUCGUGGGUUCGAUUCCCAUCAAUCGCC
X54408.1/1-73          GCCCCCAUCGUCUAGA--GGCCUAGGACACCUCCCUUUCACGGAGAAAA-CGCGGAUUCGAAUUCCGCUGGGGGUA
J05395.1/2325-2252     GGUUUCGUGGUCUAGUC-GGUU-AUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCG
X12977.1/475-548       CGGUGAUUAGCGCAGCCCGGU--AGCGCAUCUGGUUUGGGACCAGAGGGUCAAAGGUUCGAAUCCUUUAUCACCGA
X03126.1/364-434       AAGAGUAUAGUUUAAA--GGU--AAAACAGAAAGCUUCAACCUUUAAUU-UCUUAGUUCGAGUCUAAGUGCUCUUG
X52070.1/777-849       UCCUCCGUAGCUCAAUU-GGC--AGAGCAGCCGGCUGUUAACCGGCAGGUUACUGGUUCGAGUCCAGUCGGGGGAG
X13558.1/186-115       UGGGGCGUAGCCAAGC--GGU--AAGGCAACGGGUUUUGGUCCCGCUAUUCGGAGGUUCGAAUCCUUCCGUCCCAG
X52392.1/3967-4040     GCUAGCGUGGCAGAGCUCGGCA-AAUGCAAAAGGCUUAAGCCCUUUAUC-CAGAGGUUCAAAUCCUCUCCUAGCU
K00287.1/1-73          UCCUUGUUAGCUCAGUU-GGU--AGAGCGUUCGGCUUUUAACCGAAAUGUCAGGGGUUCGAGCCCCCUAUGAGGAG
Consensus (-31.49)     (((((((..((((...........)))).(((((.......))))).....(((((.......)))))))))))).
```

Figure from Voß, *Nucleic Acids Research*, 34:5471-5481, 2006.

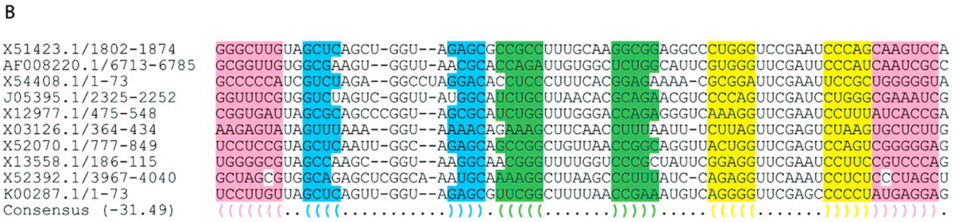
Eddy & Durbin: Model Construction Step



- use a DP like Nussinov to find folded structure that maximizes mutual information

$$\gamma(i,j) = \max \begin{cases} \gamma(i+1,j) \\ \gamma(i,j-1) \\ \gamma(i+1,j-1) + \boxed{M(i,j)} \\ \max\limits_{i<k<j}\left[\gamma(i,k) + \gamma(k+1,j)\right] \end{cases}$$

- then derive grammar from this structure



```
X51423.1/1802-1874     GGGCUUGUAGCUCAGCU-GGU--AGAGCGCCGCCUUUGCAAGGCGGAGGCCCUGGGUCCGAAUCCCAGCAAGUCCA
AF008220.1/6713-6785   GCGGUUGUGGCGAAGU--GGUU-AACGCACCAGAUUGUGGCUCUGGCAUUCGUGGGUUCGAUUCCCAUCAAUCGCC
X54408.1/1-73          GCCCCCAUCGUCUAGA--GGCCUAGGACACCUCCCUUUCACGGAGAAAA-CGCGGAUUCGAAUUCCGCUGGGGGUA
J05395.1/2325-2252     GGUUUCGUGGUCUAGUC-GGUU-AUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCG
X12977.1/475-548       CGGUGAUUAGCGCAGCCCGGU--AGCGCAUCUGGUUUGGGACCAGAGGGUCAAAGGUUCGAAUCCUUUAUCACCGA
X03126.1/364-434       AAGAGUAUAGUUUAAA--GGU--AAAACAGAAAGCUUCAACCUUUAAUU-UCUUAGUUCGAGUCUAAGUGCUCUUG
X52070.1/777-849       UCCUCCGUAGCUCAAUU-GGC--AGAGCAGCCGGCUGUUAACCGGCAGGUUACUGGUUCGAGUCCAGUCGGGGGAG
X13558.1/186-115       UGGGGCGUAGCCAAGC--GGU--AAGGCAACGGGUUUUGGUCCCGCUAUUCGGAGGUUCGAAUCCUUCCGUCCCAG
X52392.1/3967-4040     GCUAGCGUGGCAGAGCUCGGCA-AAUGCAAAAGGCUUAAGCCCUUUAUC-CAGAGGUUCAAAUCCUCUCCCUAGCU
K00287.1/1-73          UCCUUGUUAGCUCAGUU-GGU--AGAGCGUUCGGCUUUUAACCGAAAUGUCAGGGGUUCGAGCCCCCUAUGAGGAG
Consensus (-31.49)     (((((((..((((...........)))).(((((.......))))).....(((((.......)))))))))))).
```

Figure from Voß, *Nucleic Acids Research*, 34:5471-5481, 2006.

3. Using a Simple, Generic Grammar

- a grammar that could characterize almost any RNA structure

$$s \rightarrow C\, s_1\, G \;\mid\; G\, s_1\, C \;\mid\; A\, s_1\, U \;\mid\; U\, s_1\, A$$

$$s \rightarrow C\, s_1 \;\mid\; G\, s_1 \;\mid\; A\, s_1 \;\mid\; U\, s_1$$

$$s \rightarrow s_1\, G \;\mid\; s_1\, C \;\mid\; s_1\, U \;\mid\; s_1\, A$$

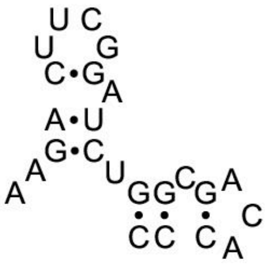
$$s \rightarrow G \;\mid\; C \;\mid\; U \;\mid\; A$$

$$s \rightarrow s\; s$$

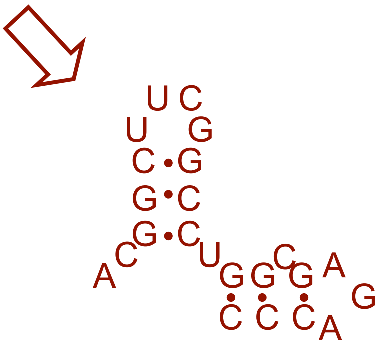
Searching Sequence for a Secondary Structure

Given

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



ACGGCUUCGGCCUGGCGAGACCC

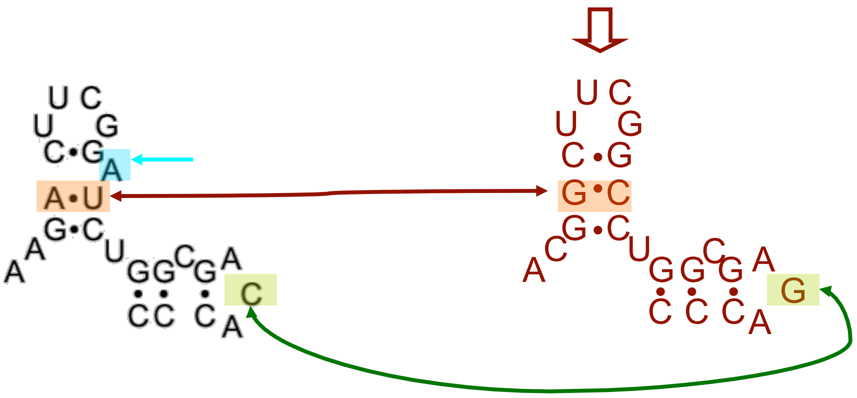


Determine if the query sequence has "same" secondary structure

Searching Sequence for a Secondary Structure

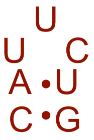
- this is analogous to pairwise alignment with primary sequences
- we take into account substitutions, insertions/ deletions, and base-pair substitutions

ACGGCUUCGGCCUGGCGAGACCC



Deriving a Grammar from Secondary Structure

given a structure



can construct a simple grammar characterizing it

$$s \rightarrow \mathrm{C}\ \mathrm{s}_1 G$$
$$s_1 \rightarrow \mathrm{A}\ \mathrm{s}_2 U$$
$$s_2 \rightarrow b_1\ l_1$$
$$l_1 \rightarrow b_2\ b_3$$
$$b_1 \rightarrow U$$
$$b_2 \rightarrow U$$
$$b_3 \rightarrow C$$

can add productions to allow for variation

$$s \rightarrow \mathrm{U}\ \mathrm{s}_1 A$$
$$s \rightarrow \mathrm{A}\ \mathrm{s}_1 U$$
$$s \rightarrow \mathrm{G}\ \mathrm{s}_1 C$$

base pair substitutions

$$s_1 \rightarrow \mathrm{s}_1 A$$

insertions

$$b_2 \rightarrow A$$
$$b_2 \rightarrow C$$
$$b_2 \rightarrow G$$

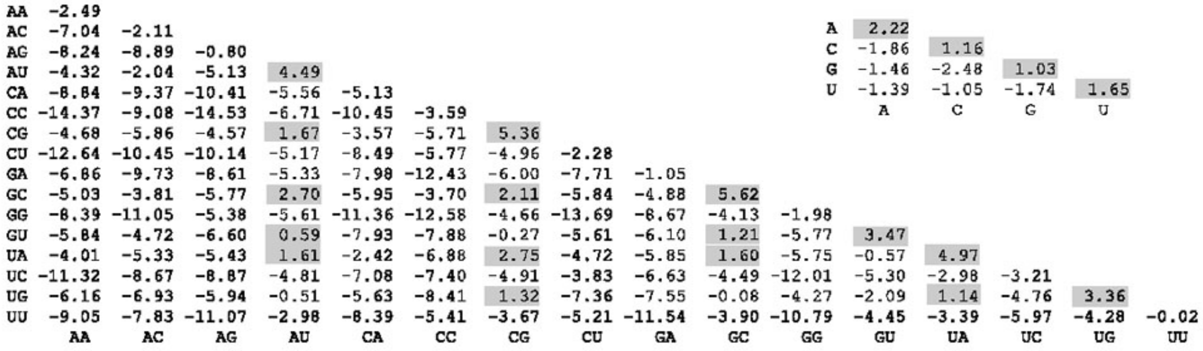
single base substitutions

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	-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
	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU

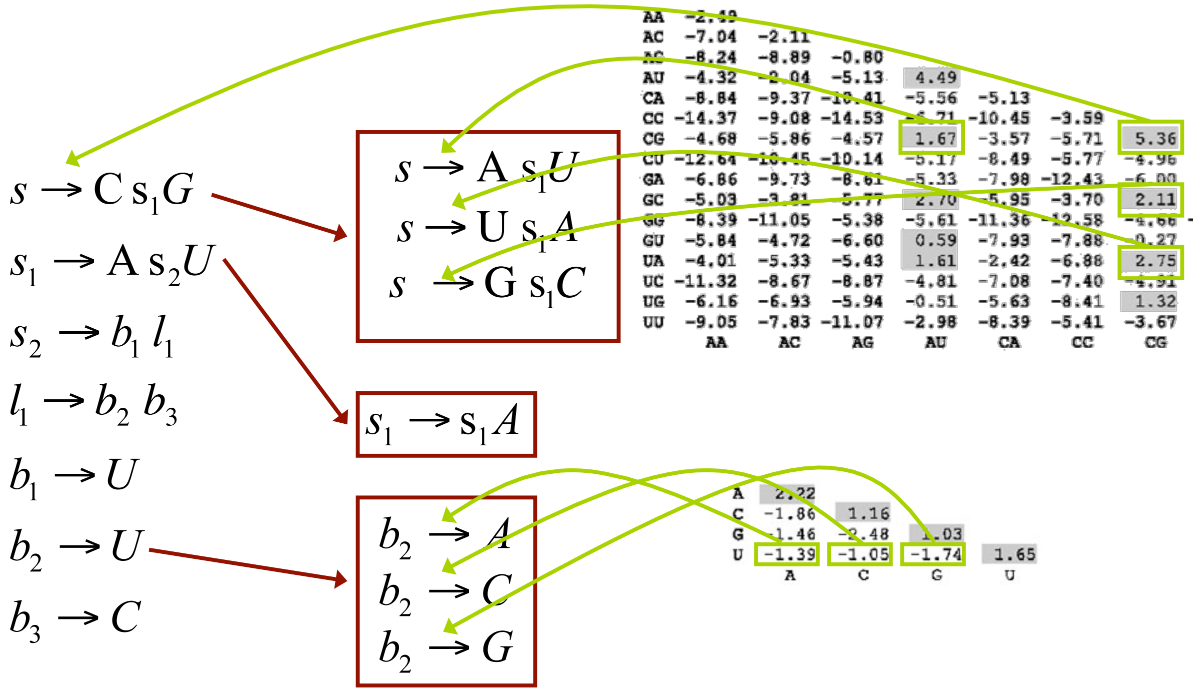
A	2.22			
C	-1.86	1.16		
G	-1.46	-2.48	1.03	
U	-1.39	-1.05	-1.74	1.65
	A	C	G	U

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

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

Setting the Parameters in the Grammar

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



$s \rightarrow \mathrm{C}\, s_1 G$

$s_1 \rightarrow \mathrm{A}\, s_2 U$

$s_2 \rightarrow b_1\, l_1$

$l_1 \rightarrow b_2\, b_3$

$b_1 \rightarrow U$

$b_2 \rightarrow U$

$b_3 \rightarrow C$

$s \rightarrow \mathrm{A}\, s_1 U$

$s \rightarrow \mathrm{U}\, s_1 A$

$s \rightarrow \mathrm{G}\, s_1 C$

$s_1 \rightarrow s_1 A$

$b_2 \rightarrow A$

$b_2 \rightarrow C$

$b_2 \rightarrow G$

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
	AA	AC	AG	AU	CA	CC	CG

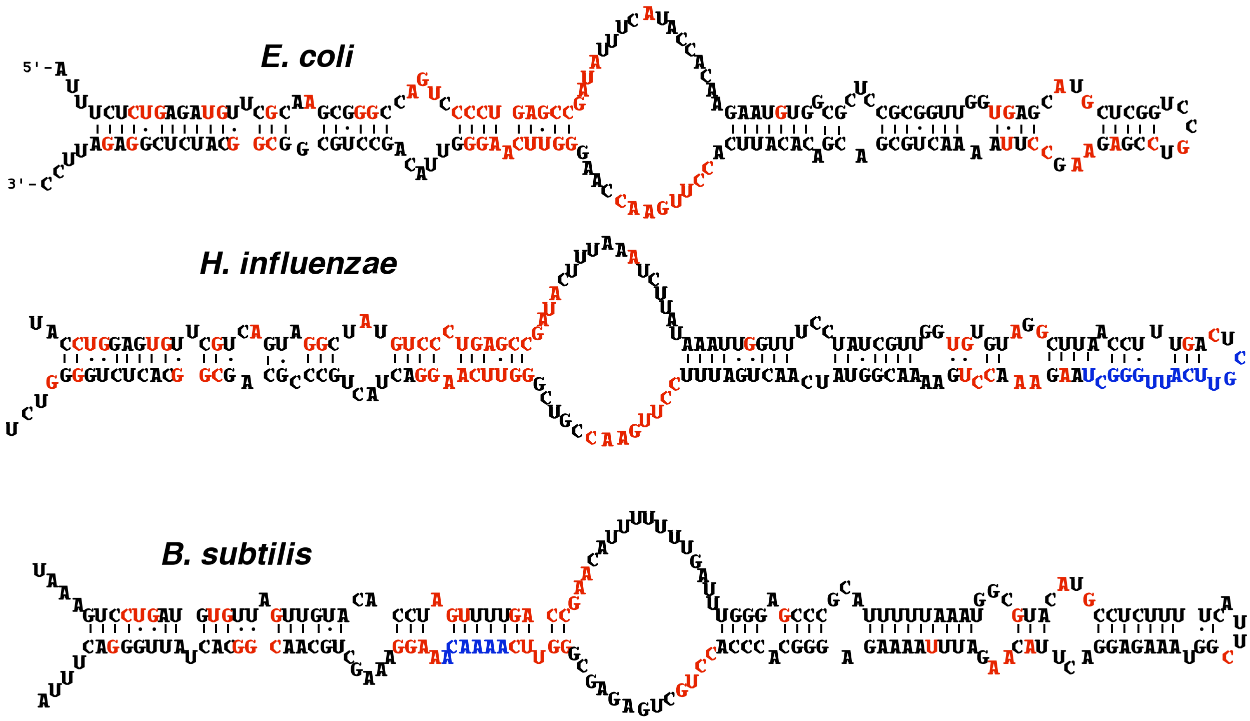
A	2.22			
C	-1.86	1.16		
G	-1.46	-2.48	1.03	
U	-1.39	-1.05	-1.74	1.65
	A	C	G	U

RSEARCH: Searching Sequence for a Secondary Structure

[Klein & Eddy, *BMC Bioinformatics* 2003]

- the RSEARCH algorithm 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

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
- others not discussed