

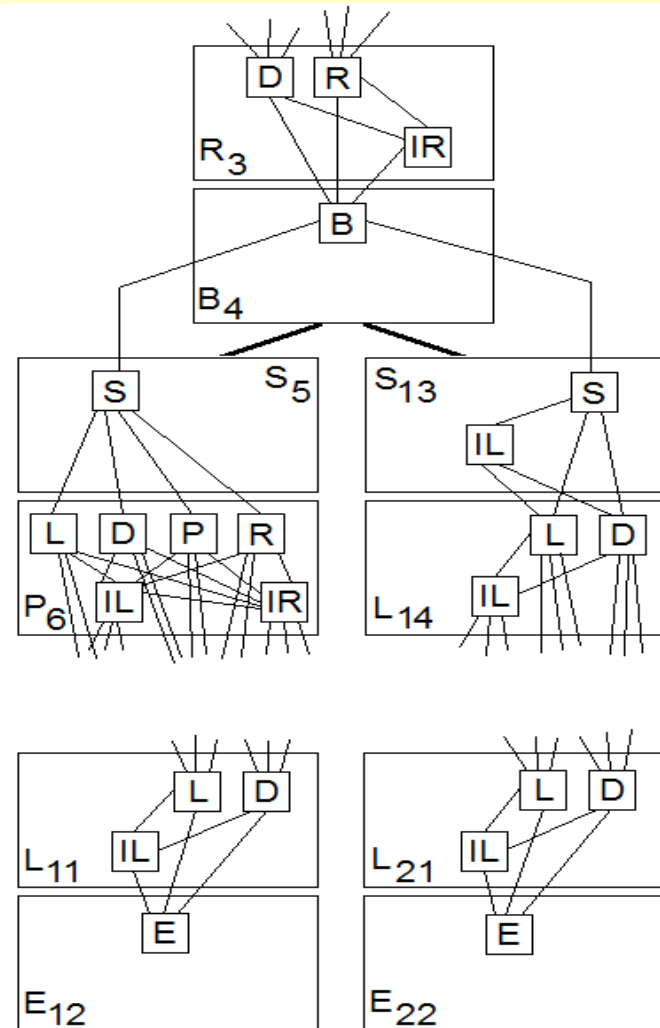
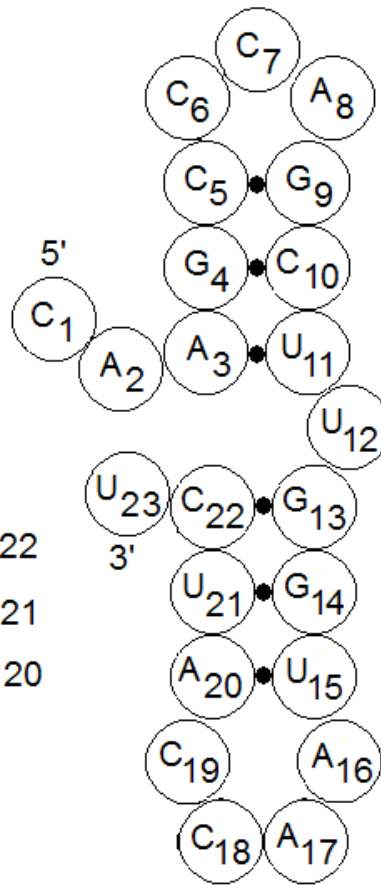
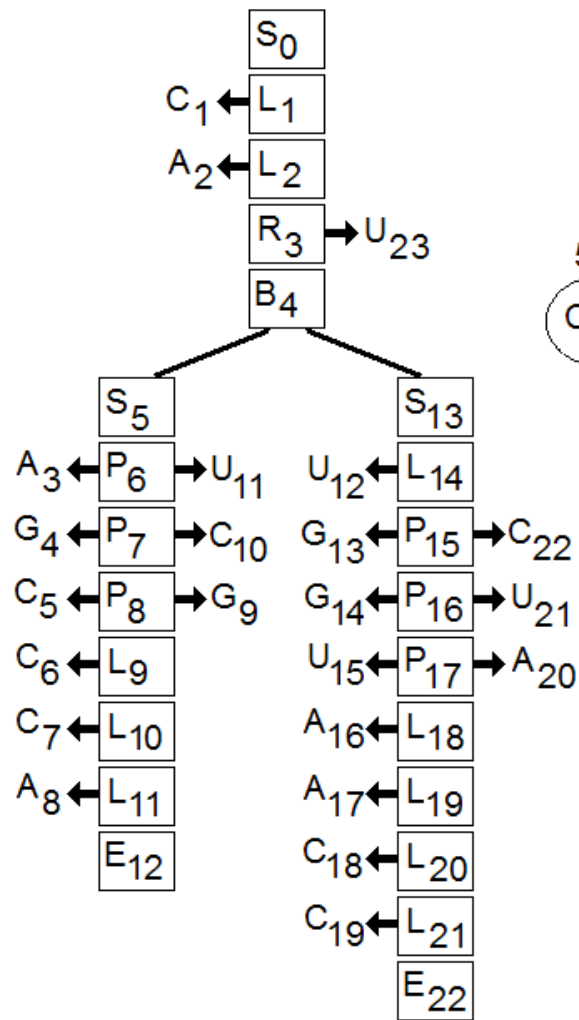
Thermodynamics-inspired ncRNA Search

Jennifer A. Smith

Electrical and Computer Engineering Dept.

Boise State University

Covariance Models



CM Parameter Estimation

Emission scores: counts of character frequencies in columns or pairs of columns

Transition scores: counts of gaps in columns or pairs of columns

Too few sequences in most training sets (families) to get reasonable parameter estimates → Use priors

```
AGCCUAAA . . . CUUAAGGG . UAAGGAAAUAUUGAUU
UGCCUAAA . . . AUUAUAAG . UAAGGAAAUAUUGAUU
UGCCUAAA . . . CUUAUGAG . UAAGGAAAUAACGAUU
UGACUUA . A . . . CUU . . . AGCUAAGGAAAUAUUGGUU
UGCCUAAA . . . AUUAUAAG . UAAGGAAAUAUUGAUU
GGUAUUAAGAGUUCUUAUGAG . UAAGGAAAUAACGAUU
UGCCUAAA . . . CUUAUGAG . UAAGGAAAUAACGAUU

--<<<<<----->>>>>-----<<<<

CM nodes:
LLPPPPPLL . . . LLLLLLLL . PPPPLLLLLLLLLLLLPPPP
```

Priors in current use:

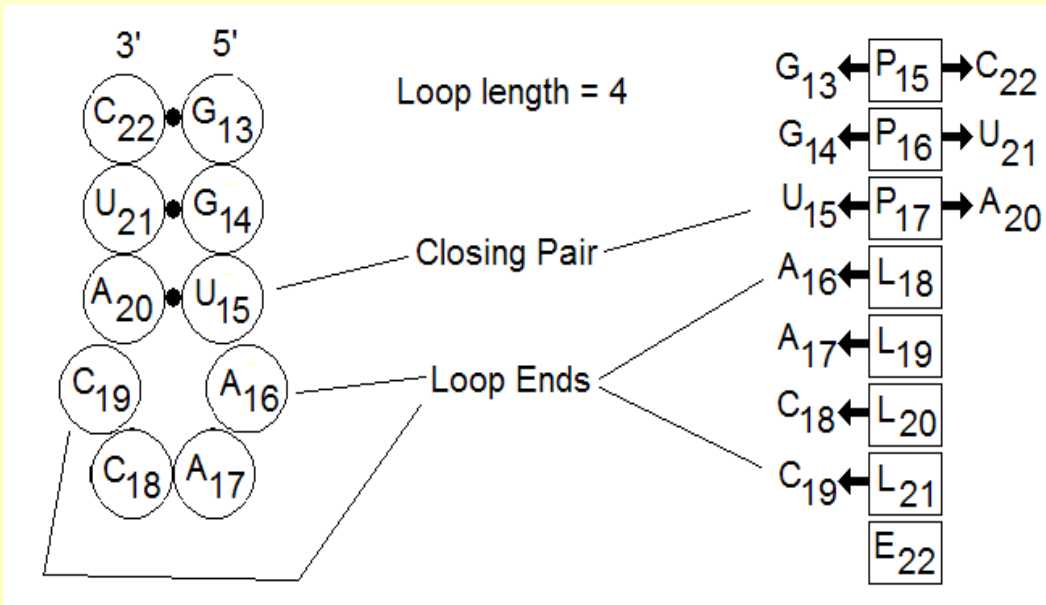
- Based on multiple-alignment column counts across many families
- Vary by type of node and type of child node
- Do not take into account:
 - a) loop lengths for L-node insertions/deletions
 - b) joint consensus nucleotides of loop ends and closing pair for mutations in these nucleotide positions

Improving Priors with Thermodynamic Data

Lab testing shows*:

1) ΔG for insertion or deletion in loops depends on loop length.

2) ΔG for mutation in closing pair or loop end nucleotides depends on identity of other closing pair and loop end nucleotides.

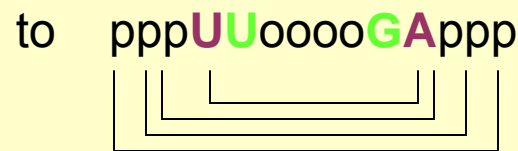
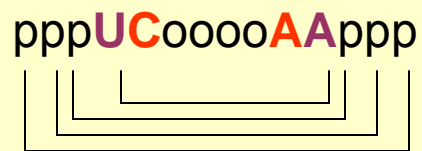


Existing methods of determining CM priors do not allow these dependences.

Mutation Probabilities at Loop/Stem Interface

RAW COUNTS AND PRIORS FOR HAIRPIN LOOP END-PAIRS

Loop End Pair	Raw Counts					Log (base 2) Likelihood Ratio Priors				
	Closing AU	Closing UA	Closing CG	Closing GC	All	Closing AU	Closing UA	Closing CG	Closing GC	All
AA	318	302	2173	1098	4054	0.16	0.03	0.98	0.48	0.65
AC	94	25	293	147	628	-0.93	-2.89	-1.24	-1.75	-1.36
AG	113	32	694	114	1013	-0.88	-2.76	-0.22	-2.33	-0.89
AU	110	66	454	208	859	-1.15	-1.94	-1.06	-1.70	-1.36
CA	671	1269	865	163	3007	1.91	2.77	0.32	-1.60	0.90
CC	301	72	128	133	692	1.43	-0.69	-1.76	-1.22	-0.55
CG	42	146	1099	86	1405	-1.64	0.11	1.12	-2.07	0.25
CU	115	104	678	175	1133	-0.41	-0.61	0.19	-1.27	-0.29
GA	175	182	1387	2270	4202	-0.25	-0.25	0.78	1.98	1.16
GC	62	43	170	92	378	-1.07	-1.66	-1.57	-1.97	-1.64
GG	94	235	285	160	844	-0.69	0.57	-1.04	-1.39	-0.70
GU	48	34	123	153	410	-1.90	-2.45	-2.49	-1.69	-1.98
UA	359	131	450	332	1318	0.55	-0.96	-1.07	-1.02	-0.75
UC	174	257	238	324	1104	0.18	0.69	-1.32	-0.38	-0.33
UG	65	23	1158	219	1495	-1.46	-3.01	0.75	-1.17	-0.11
UU	207	140	1204	2459	4102	-0.02	-0.64	0.57	2.09	1.11
All	2948	3061	11399	8133	26644					



mutation unlikely

($2^6 = 64$ times less likely than random double mutation)

Insertion/Deletion Probability Loop-length Dependence

Deletions in loops of length 3 should be heavily penalized -- result is not possible.

Insertions with loop lengths of 3 or 4 and deletions with loop lengths of 4 or 5 result in large differences in free energy of formation. These indels less likely than indels in longer loops.

Loop Length	ΔG Formation of Hairpin Loops
1	too tight
2	too tight
3	+7.4
4	+5.9
5	+4.4
6	+4.3
7	+4.1
8	+4.1
9	+4.2
10	+4.3
12	+4.9
14	+5.6
16	+6.1
18	+6.7
20	+7.1
25	+8.1
30	+8.9