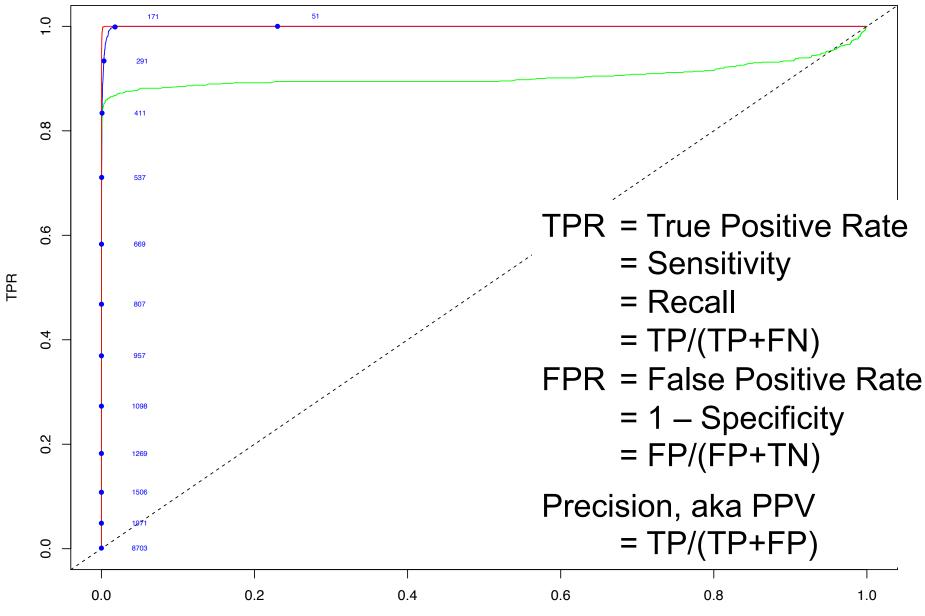
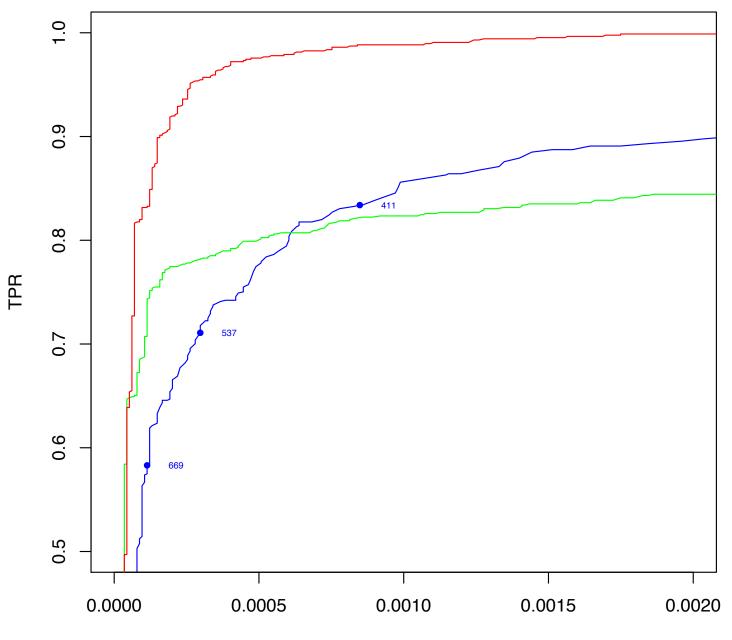


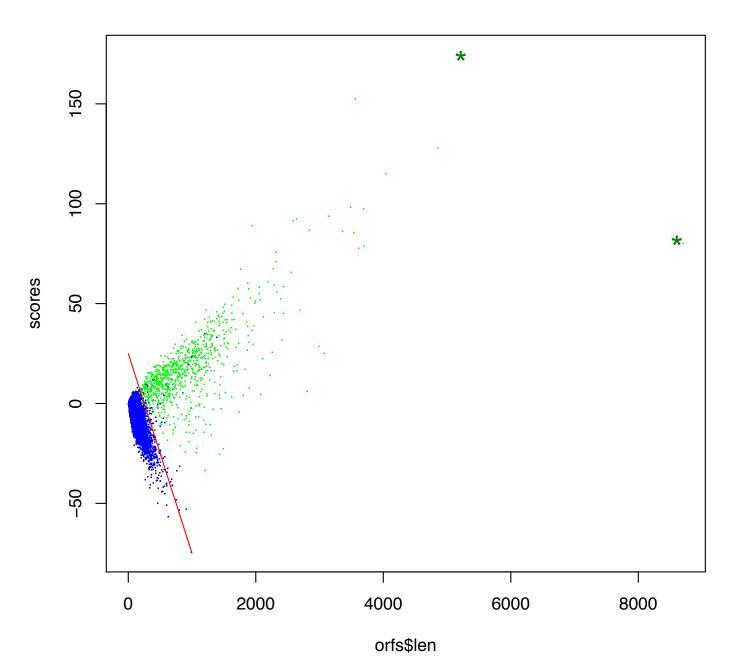
len	mun	gbk	mmtp	mmp	avg	mm	avg	mm	avg	mm	avg	
3	8381	0	0	0	0.00	0	-0.10	0	0.00	0	0.00	
6	8966	0	0	5092	-0.02	0	-0.10	5458	-0.03	5458	-0.03	5
9	9523	0	0	4946	-0.10	2690	-0.21	5064	-0.10	5064	-0.10	5
12	8622	0	0	4227	-0.15	3143	-0.25	4298	-0.14	4298	-0.14	4
15	7042	0	0	3366	-0.20	2829	-0.28	3458	-0.19	3458	-0.19	3
18	6475	0	0	2971	-0.26	2595	-0.34	3037	-0.25	3037	-0.25	3
21	6223	0	0	2754	-0.33	2500	-0.41	2883	-0.32	2883	-0.32	2
159	142	0	0	9	-7.39	10	-7.91	10	-7.93	10	-7.93	
162	105	1	1	6	-8.75	6	-9.33	6	-9.38	6	-9.38	
165	113	0	0	6	-7.83	7	-8.36	8	-8.34	8	-8.34	
168	100	0	0	6	-7.93	6	-8.41	7	-8.43	7	-8.43	
171	117	0	0	8	-8.21	9	-8.76	9	-8.80	9	-8.80	
174	80	1	1	6	-7.85	6	-8.39	6	-8.33	6	-8.33	
177	103	0	0	8	-7.91	8	-8.44	8	-8.43	8	-8.43	
5253	1	1	1	1	148.45	1	174.37	1	174.80	1	174.80	
8703	1	1	1	1	61.87	1	80.09	1	80.37	1	80.37	

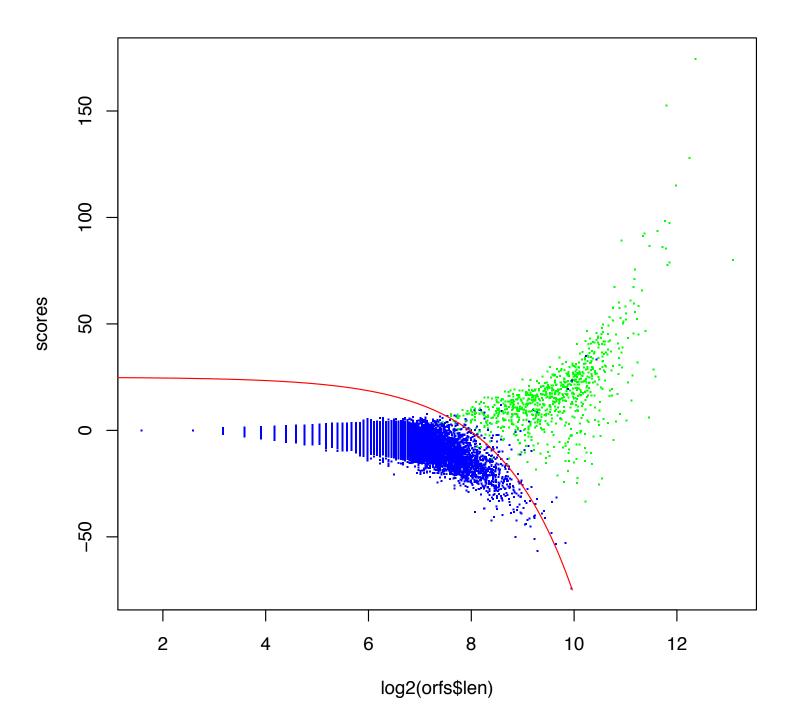
mm	avg	mm	avg	mm	avg	mm	avg	mm	avg	mm	avg
0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
458	-0.03	5458	-0.03	5458	-0.03	5189	-0.02	5458	-0.03	5150	-0.02
064	-0.10	5064	-0.10	5064	-0.10	5226	-0.09	5080	-0.10	5123	-0.10
298	-0.14	4298	-0.14	4298	-0.14	4363	-0.13	4323	-0.14	4325	-0.14
458	-0.19	3458	-0.19	3458	-0.19	3471	-0.18	3485	-0.18	3464	-0.18
037	-0.25	3037	-0.25	3037	-0.25	3061	-0.25	3058	-0.24	3051	-0.25
883	-0.32	2883	-0.32	2883	-0.32	2852	-0.31	2906	-0.31	2887	-0.31
10	-7.93	10	-7.93	11	-7.60	9	-7.90	10	-7.77	10	-7.90
6	-9.38	6	-9.38	7	-9.08	6	-9.37	7	-9.21	6	-9.38
8	-8.34	8	-8.34	8	-7.99	7	-8.35	9	-8.17	8	-8.30
7	-8.43	7	-8.43	9	-8.09	8	-8.44	7	-8.28	6	-8.39
9	-8.80	9	-8.80	10	-8.49	9	-8.78	9	-8.63	9	-8.77
6	-8.33	6	-8.33	6	-8.00	6	-8.37	6	-8.17	6	-8.26
8	-8.43	8	-8.43	10	-8.16	8	-8.45	8	-8.25	9	-8.41
1	174.80	1	174.80	1	174.80	1	172.22	1	177.83	1	180.67
1	80.37	1	80.37	1	80.37	1	71.00	1	86.31	1	84.79

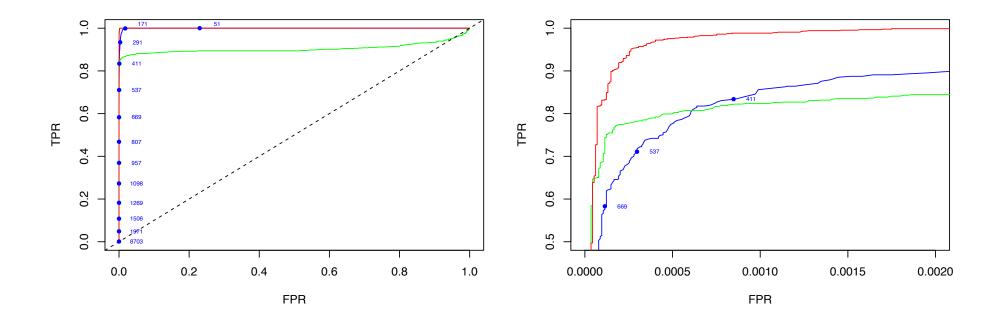
ROC

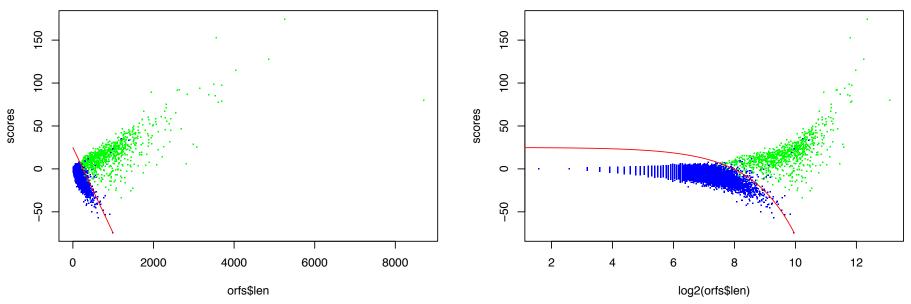






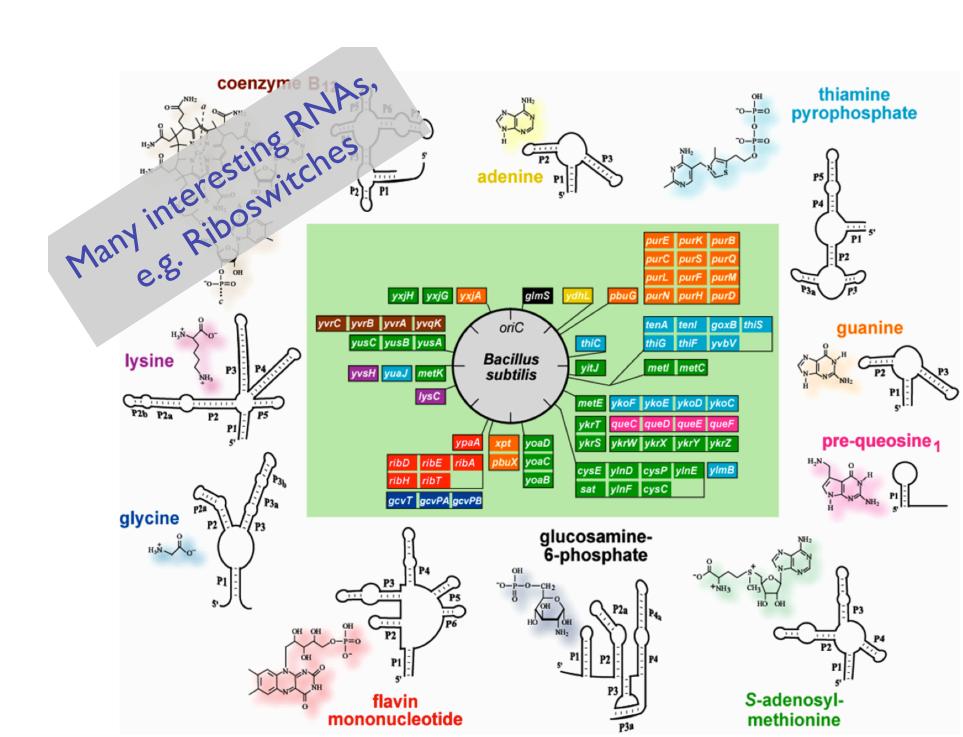






RNA Search and Motif Discovery

> CSEP 590 A Computational Biology



Structure Prediction Prediction Prediction Prediction

Maximum Pairing

- + works on single sequences
- + simple
- too inaccurate
- **Minimum Energy**
 - + works on single sequences
 - ignores pseudoknots
 - only finds "optimal" fold

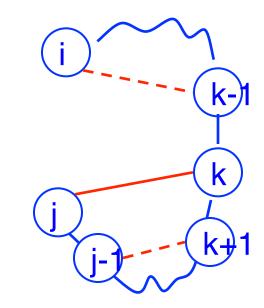
Partition Function

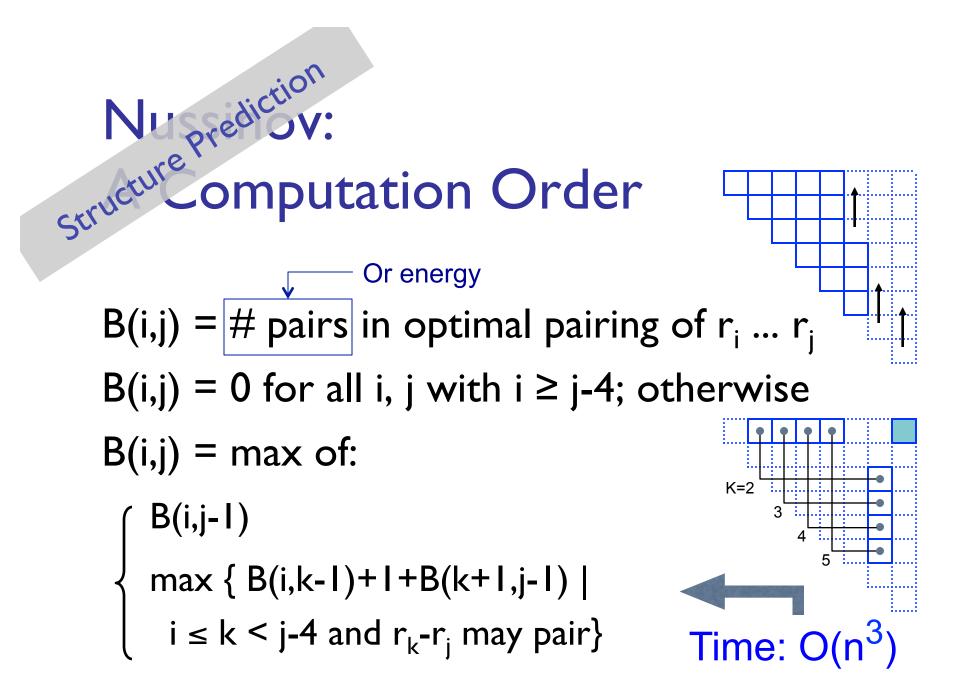
- + finds all folds
- ignores pseudoknots

"Cructure Prediction Structure Prediction pairing of r_i ... r_j" Two possibilities

- j Unpaired: Find best pairing of r_i ... r_{j-1}
- j Paired (with some k): Find best $r_i \dots r_{k-1} + best r_{k+1} \dots r_{j-1}$ plus l

Why is it slow? Why do pseudoknots matter?





Loop-based energy version is better; recurrences similar, slightly messier

Today

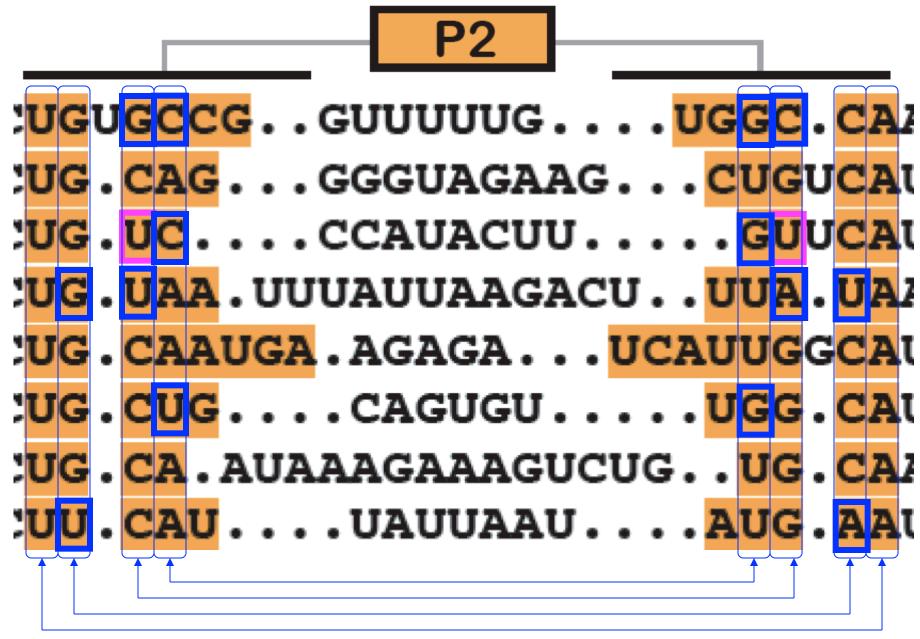
Structure prediction via comparative analysis Covariance Models (CMs) represent RNA sequence/structure motifs Fast CM search Motif Discovery

Applications in prokaryotes & vertebrates

Approaches, II

Structure Prediction Comparative sequence analysis + handles all pairings (potentially incl. pseudoknots) - requires several (many?) aligned, appropriately diverged sequences Stochastic Context-free Grammars Roughly combines min energy & comparative, but no pseudoknots

Physical experiments (x-ray crystalography, NMR)



Covariation is strong evidence for base pairing ²¹

A L19 (*rplS*) mRNA leader

35 10 TSS P2 Res		P1
B su Trester 17: Allo 11: Allo 11: Allo 12: All		TSS
Bha TYPEYEYE 17: CETYECT: 17: AUUACCAUCUUCCCCUG - CAG GEGUAGAAG EUGUCUAGAGCAUCUG.O.6. NERROF.11: NJ Brossen TYPEYEYE 17: AUUACCAUCUUCCCCUG - CAUACUU GUUCUAGAGCAUCUG.O.6. NERROF.17: NJ Brossen TYPEYEYE 11: PATCAT. 38: AUAACCAUCUUCCCCUG - CAUACUU UUA. UA CAGCAUCUU 60: NERROF.17: 0.8. NJ Brossen TYPEYEYE 17: AUCAA: 38: AUAACCAUCUUCCCCUG - CAUACUU UUA. UA CAGCAUCUU 60: NERROF.17: 0.8. NJ Brossen TYPEYEYE 17: AUCAA: 38: AUAACCAUCUUCCCCUG - CAUCAU UGE - CUGAAUCUUC 61: CERROF.08: NJ Brossen TYPEYEYE 17: AUCAA: 38: AUAACCAUCUUCCCCUG - CAUGAU UGE - CUGAAUCUUC 61: CERROF.08: NJ Brossen TYPEYEYE 17: AUAACA: 25: AUAACCAUCUUCCCCUG - CAUGAU UGE - CUGAAUCUUC 61: CERROF.08: NJ Brossen TYPEYEYE 17: AUAACA: 25: AUAACCAUUUCCCCUU - CAU AUUAUAUU UGE - CUGAAGCUUU 65: NEGROEI 08: NJ Brossen TYPEYEYE 18: AUAACCAUUUCCCCUU - CAU AUUAUAUU UGE - CUGAAGCAUCUU 65: NEGROEI 08: NJ Brossen TYPEACH: 18: AUAACCAUCUUCCCUU - CAU AUUAUAUU AUGE AUACAUACUU 65: NEGROEI 08: NJ Brossen TYPEACH: 18: AUAACCAUCUUCCCUU - CAU AUUUAUAU AUGE AUACAUACUU 65: NEGROEI 08: NJ Brossen TYPEACH: 18: AUAACCAUCUUCCCUU - CAU AUUUAUAUU AUGUUAUAAUAUUUUCCC AGCGAUCAU NEGENEE 09: NJ Brossen TYPEACH: 17: CHANAY 10: UUACCACCCCUCUUCCUU - CAU AUUUAUAUU AUGUUAUAAUAUUUCCUU AUGUUAUAAUAUUUUCCUU AUGUUAUAAUAUUUUCCUU AUGUUAUAAUAUUUU AUACUU AUGUUAUAUAUU AUGUUAUAUAUU AUGUUAUAUAUU AUGUUAUAUAUU AUGUUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAU	-35	-10 → P2 RBS Sta
Bha TYPEYEYE 17: CETYECT: 17: AUUACCAUCUUCCCCUG - CAG GEGUAGAAG EUGUCUAGAGCAUCUG.O.6. NERROF.11: NJ Brossen TYPEYEYE 17: AUUACCAUCUUCCCCUG - CAUACUU GUUCUAGAGCAUCUG.O.6. NERROF.17: NJ Brossen TYPEYEYE 11: PATCAT. 38: AUAACCAUCUUCCCCUG - CAUACUU UUA. UA CAGCAUCUU 60: NERROF.17: 0.8. NJ Brossen TYPEYEYE 17: AUCAA: 38: AUAACCAUCUUCCCCUG - CAUACUU UUA. UA CAGCAUCUU 60: NERROF.17: 0.8. NJ Brossen TYPEYEYE 17: AUCAA: 38: AUAACCAUCUUCCCCUG - CAUCAU UGE - CUGAAUCUUC 61: CERROF.08: NJ Brossen TYPEYEYE 17: AUCAA: 38: AUAACCAUCUUCCCCUG - CAUGAU UGE - CUGAAUCUUC 61: CERROF.08: NJ Brossen TYPEYEYE 17: AUAACA: 25: AUAACCAUCUUCCCCUG - CAUGAU UGE - CUGAAUCUUC 61: CERROF.08: NJ Brossen TYPEYEYE 17: AUAACA: 25: AUAACCAUUUCCCCUU - CAU AUUAUAUU UGE - CUGAAGCUUU 65: NEGROEI 08: NJ Brossen TYPEYEYE 18: AUAACCAUUUCCCCUU - CAU AUUAUAUU UGE - CUGAAGCAUCUU 65: NEGROEI 08: NJ Brossen TYPEACH: 18: AUAACCAUCUUCCCUU - CAU AUUAUAUU AUGE AUACAUACUU 65: NEGROEI 08: NJ Brossen TYPEACH: 18: AUAACCAUCUUCCCUU - CAU AUUUAUAU AUGE AUACAUACUU 65: NEGROEI 08: NJ Brossen TYPEACH: 18: AUAACCAUCUUCCCUU - CAU AUUUAUAUU AUGUUAUAAUAUUUUCCC AGCGAUCAU NEGENEE 09: NJ Brossen TYPEACH: 17: CHANAY 10: UUACCACCCCUCUUCCUU - CAU AUUUAUAUU AUGUUAUAAUAUUUCCUU AUGUUAUAAUAUUUUCCUU AUGUUAUAAUAUUUUCCUU AUGUUAUAAUAUUUU AUACUU AUGUUAUAUAUU AUGUUAUAUAUU AUGUUAUAUAUU AUGUUAUAUAUU AUGUUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAUAU		
$\begin{array}{c} cih & \mbox{rescale} 1.17 & \mbox{rescale} 1.17 & \mbox{rescale} 1.11 & res$		
Bee widewide in the interaction is characterized by the interaction of the interaction o		
$ \begin{array}{c} \text{dra} & \text{rescen} & 17. & \text{PATCAR} & 38. \text{AAAAC} CALGUUC COCUS & CANUER, ACAACAL, UCAUVAUUC UC.04. EEGENET 0.98. N) \\ \text{Bed} & \text{rescaled} & re$		
B21 TrCRCC 17. 0 AVEAU 45.4.5.4.000ACCAUGUECCEGE 000		
Bac MCANE 17, GATAGT, 35 AUBACGAUGUUC CCCUE 68 AUBAACAAAGUUC C. 16 EAR GACAUGUU G. 65 MEGARE 1.08 M Imo TYPACA 17, GATAGT, 25 AUBACGAUAUUC CCCUE 61 UAUUAAU NUG ABCACAAGUUG G. 65 MEGARE 1.08 M TYPACA 18, TAAACT 08 GUACCGUC CCUC GUUC CUCUCUCACA GAGC CAAGAACAUGUUG G. 64 MEGARE 0.89 M TYPACA 17, GATAGT, 05 UACCGACCGUC CCUC GUUC CUCUCUCACA GAGC C UGUGUUAACAACAUGUUG G. 8 MEGARE 0.81 M TYPACA 17, GATAGT, 05 UACCGACCGUC CCUC GUUC CUCUCACA AGGC UGUGUUAACAACAUGUUG G. 8 MEGARE 0.99 M Mar TYGCC 17, GATAGT, 05 UACCGACCGUC CCUCA UAU AAGCA GUU GUAACAGUC CC. 8 MEGARE 0.99 M Mar TYGCC 17, GATAGT, 05 UACCGACCGUC CCUCA UAU AAGCA GUU AGAACGUC C. 8 MEGARE 0.71 M Amar TYGCC 17, GATAGT, 10 UUACGACCGUC CCUCA UAU AAGCA GUU AUGAACGUC CAA. 07 MEGARE 0.71 M Amar TYGCC 17, GATAGT, 10 UUACGACGUC CCUCA UAU AAGCA GUU AUGAAGAUCUU 8 MEGARE 0.71 M Amar TYGCC 17, GATAGT, 10 UUACGACGUC CCUCA UAU AAGCA GUU AUGAAGAUCUU 8 MEGARE 0.71 M Amar TYGCC 17, GATAGT, 10 UUACGACUUUAUC CCUCA AAGU AGGA AGU UCCU CAAGAUUUACAAA. 04 MEGARE 0.71 M M TYTACA 17, GAACAT 26 AAACGCUUAUUC CCUCA AAGU AGGA AGU UCCU CAAGAUUUCAAA. 04 MEGARE 0.71 M M TYTACA 17, GAACAT 28 AUACACUUUAUC CCUCA AAGU CUGCCA DAAGAUUUC CCA AA 04 MEGARE 0.71 M Lp TYTACA 17, GAACAT 28 AUACACUUUAUC CCUCA AAGU CAGGUU UCCU CAAGAAUUUC CCAA. 04 MEGARE 0.71 M Lp TYTACA 17, GAACAT 28 AUACACUUC CCUCA CAAGA ACAUCUU UCCU CAAGAAUUUC CC 0. 04 MEGARE 0.71 M Lp TYTACA 17, GAACT 28 AUACACUUUC CCUCA CAAGA ACAUCUU UGCCA DAAUUUC CC 0. 04 MEGARE 0.71 M Lp TYTACA 17, TAAAAT, 26 ACAACAAUUUC CCCUC CCAC CCA AAGU UCAGUUAUAUAA. 04 MEGARE 0.71 M La TYTACA 17, TAAAAT, 26 ACAACAAUUUC CCCUC CCAC CCAC CAAGUU GUUAGAUAUUAUA AAGAAAUU AAGAAAUUUC AGCAAUUUUC CCUC CAAGAAUUUC CCUC CAAGA AUUUAUU AAGAAUUUC AGCAAUUU AAGAAUUUC AGCAAUUUUC AGCAAUUUC AGCAAUUUUC AGCAAUAUUUC AGCAAUUU AGCAAUUUUC AGCAAUUUUC AGCAAUUUUC AGCAAUUUUC CGUUAUUAUUA		
$ \begin{array}{c} Izmo \\ reverse_1, 17, \\ reverse_1, 17, \\ reverse_2, 17, \\ reverse_3, 17, \\ reverse_4, 17, \\ reverse_$		
Sau TTCAAA.1.17. CAACAT.23. AUCACTAUGCUC Cucueducaca		
Cpc UZANAC, 18: CAAACC, 08: GUACCGCCCCUCCUCUCCCCC, GAC,, CAC,, UC, CACGACCU, CAA, 17: NEGRAGE, 09: NU Chy DYCCAN, 17: DANAM, 16: AAAACUCCCCCCCUC, GAU,, AGGA,, CAGGGC,, UC, CAACCUC, CC, 03: USACACCUCU, 05: MEGAGE, 07: NU Ame DYCCCO, 17: OYAAA, 16: UNACCACCCCCUCUCU, UA, UA,, AGGA,, CAU, UA, CAACUUCA, 04: AGGAGE, 07: NU Put DYCTCCO, 17: OYAAA, 16: UNACCACCCCUCUCU, UA, UA,, AGGA,, CGU, UA, CAACUUCA, 04: AGGAGE, 07: NU Sput DYTYACC, 17: OYAAA, 16: UNACCACCCCUCUCU, UA, UA,, AGGA,, CGU, UA, CAACUUCA, 04: AGGAGE, 12: U Sput DYTYACC, 17: OYAAA, 16: UNACCACCCUUUCCCCUC, AGGA,, AGGA,, CACAUU, 04: AUAACA, 04: AGGAGE, 07: NU Sput DYTYACA, 17: OYAAA, 16: UNACCACCUUUCCCCUC, AGGA,, AGGA,, CUCU, CAACAUUCACAA, 04: AGGAGA, 05: NU Smut DYTYACA, 17: OYAAA, 22: UNACCACUUUCCCCUC, AGGA,, ACACAU,, CUCU, CAACAUUACUAA, 04: AGGAGA, 06: NC Spit DYTYACA, 17: OYAAAC, 22: UNACCACUAUUCCCCUC, AGG, C:, CAAC,, CAGUU,, CACGAUUAUUA, AA, 04: MEGAAC, 06: NU Stb DAGAAC, 29: UNACCCCUAUUCCCCUC, AGG, C: C, CAAC,, CACUUAUACUAA, 03: MEGAAC, 06: NU Spit DYTACA, 17: OYAAAC, 29: UNACCGCUAUUCCCCUC, AGG, C: C, CAAC,, CACUUAUUACUAA, 03: MEGAAC, 08: NU Spit DYTACA, 17: OYAAAC, 29: UNACCGCUAUUCCCCUC, AGG, C: CAACA, C, CAACU,, CCUCU, UAACAUUAUAU, AA, 03: MEGAAC, 06: NU Spit DYTYACA, 17: OYAAAA, 29: UNACCGCUUAUUCCCUC, C: ACG, C: C		
Chy TYCAAN 17.1 MARAM 09. UACCANGCUUC GCUG 68 CAGGGCC UG CSUGAACGUCC.03. NGGAGG 09. NU SWO TYCAAN 17. PANAM 10. UUACGCCUCCUC GCUG CAUU AAACUAA ANCUAA ANU UA CAACUU 05. NGGAGG 07. NU me TYCACC 17. PANAM 10. UUACGCCUCCUCUA UAAACUAGG UAACACCUU 05. NGGAGG 07. NU PTYCACC 17. PANAM 10. UUACGCCUCCUCUA UAAGG UAACAACUU 04. NGGAAC 12. UT PTYCCC 17. PANAM 10. UUACGCCUCCUCUA UAAGG UAACAACUU 04. NGGAAC 12. UT Smu TYTACH 17. PANAM 10. UUACGCCUCCUCUG CCUC. CUCU UGGAAGUU UAACAAUUCU.07. NGGAGG 12. UT Smu TYTACH 17. PANAM 10. UUACGCUCCUCUCUCUA UA		
$ \begin{array}{c} s_{po} & rrccAch, 17 \\ rrAAAAr, 16 \\ AAAAAr, 16 \\ AAAAAr, 16 \\ AAAAAr, 17 \\ rrAAAr, 16 \\ rracccc 17 \\ rracAch, 17 \\ rracccc 17 \\ rracAch, 17 \\ rracccc 17 \\ rracAch, 17 \\ rracccc 17 \\ rrach, 17 \\ rracccc 17 \\ rrach, 17 \\ $	_	
Ame TTCCCC 17 PAYAAT. 10. UUACGCCCCC UUA. UAC AGGA GUA LAGACACUCUA. 07. NEGROG. 07. NU Pre TTCCCC 17. PAYAAT. 10. UUACGCCCCCUC AGGA AGGA AGGA AGGA AGGA OL UAAGAACUCUA. 04. NCGARG. 12. GU pre TTACA. 17. PAAAAT. 28. AUACGCUUAUCCCCUC AGGA AGAU UCCU. CAAGAAUAU AGGAAA. 04. NCGARG. 05. AU Smu TTTACA. 17. PAAAAT. 28. AUACGCUUAUCCCCUC AGGA ACACAGCA CU UAAGAUAUCGCG. 04. NCGARG. 07. NU Pre TTTACA. 17. PAAAAT. 28. AUACGCUUAUCCCCUC AGGA ACACAGCA CU UAAGAUAUCGCG. 04. NCGARG. 07. NU Pre TTTACA. 17. PAAAAT. 28. AUUACAAUUCCCCUC AGGA ACACAGCA CU UAAGAUUAUCGC 0.03. NCGARG. 08. AU Ljo TTTACA. 17. PAAAAT. 28. AUUACAAUUCCCCUC CCCCC AAG CGGUUCAU CAAUCCCC 0.3. NCGARG. 08. AU Ljo TTTACA. 17. PAAAAT. 29. UUACGCUUAUCCCCUC CCCCC AAG CGUUCAUCAUAUUUG. CCCCU. 03. NCGARG. 08. AU Lac TTTAAA. 17. PAAAAT. 29. UUACGCUUAUCCCCUC CCCCC AAG CGUUCAUCAUA. 03. NCGARG. 08. AU Lac TTTAAA. 17. PAAAAT. 29. UUACGCUUAUCCCCUC CCCCC AAG CCUUCAUGAUCAA. 03. NCGARG. 06. AU Spy TTTACA. 17. PAAAAT. 29. UUACGCUAUUCCCCUC CCCCC AACAGAU CUUUAUCGCUAA. 03. NCGARG. 06. AU Lac TTTAAA. 17. PAAAAT. 29. UUACGCUAUUCCCCUC CCCCC AACAGAU CUUUAUAGAAUUCGCCAA. 03. NCGARG. 06. AU Lac TTTAAA. 17. PAAAAT. 29. UUACGCUAUUCCCCUC CCCUC AACAGAU CUUUAUGAUAUUCGCCCAA. 03. NCGARG. 06. AU Las TTTTAA. 17. PAAAAT. 29. UUACGCUAUUCCCCUC CCUC AACAGAU CUUUAUGAAUUUCGCCCAA. 03. NCGARG. 06. AU Ls1 TTTACA. 17. PAAAAT. 21. AUUAGAUAUUCCCCCUC CCUC AACUG CAACAUAUUCGCCCAA CAAGAUAUUCGCCCAA CAAGAUAUUCGCCCAA CAAGAUAUUCGCCCAA CAAGAUAUUCGCCCAA CAAGAUAUUCGCCCAA CAAGAUAUUCGCCCAA CAAGAUAUUCGCCCU ACUG CACAUAUUCGCCAA CAAGAUAUUCGCCCAA CAAGAUAUUCGCCCAA CAAGAUAUUCGCCCU ACU G C C G G U G A C G G G G G G G G G G G G G G G G G	-	
$ \begin{array}{c} Pre \\ rrccccc & 17 \\ spr \\ rrcccc & 17 \\ rracc & $		
$ \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c} \begin{array}{c}$		
$ \begin{array}{c} \text{Smu} & \text{TTNCR} & 17, & \text{PACAN}, 26, \text{AAACG} & \text{GCUANUC COCUE} & \text{AG} & \dots \text{ACAGAGCA} & \dots & \text{CU} & \text{UAU} & \text{GAUUAGUAA, 04} & \text{REEACH} & 07, & \text{NU} \\ \text{Lp} & \text{TTCCC}, 18, & \text{PATTCT}, 21, \text{UUAACGAUGUUC COCUE} & \text{AG} & \dots \text{CACGUU} & \text{CACCA, UAACACAUAUUC COC} & \text{AGAAA} & \text{OU} & \text{AU} \\ \text{Lf} & \text{TTTACL}, 17, & \text{PAAACT}, 22, \text{AUUACAUUCCCCUE} & \text{UGG} & \text{CA} & \text{CAAUAUUCGAUCGCC} & \text{UGA} & \text{CAUUAGUAUCCCCU} & \text{CACGAUAUUCGCC} & \text{UGA} & \text{CAUUAGUAAUCGCC} & \text{UGA} & \text{CAUUAGUAAUCGCCCU} & \text{CACGAUAUUCGAUACCCC} & \text{UGA} & \text{CAUUAGUAAUCGAUAUCCCCU} & \text{CACG} & \dots & \text{ACAGAGGU} & \text{UGCUU UAACAUUAGUAA} & \text{AO} & \text{REEACH} & 07, & \text{NU} \\ \text{Sth} & \text{TACACT}, 17, & \text{PAAACT}, 22, & \text{UUAUCGCUAUUCCCCU} & \text{CACG} & \dots & \text{ACAGAGGU} & \dots & \text{CUUUGAUCAUUAGUAA, 03} & \text{REEACH} & 07, & \text{NU} \\ \text{Sth} & \text{TACACT}, 17, & \text{PAAACT}, 22, & \text{UUAUCGCUAUUCCCCU} & \text{ACC} & \dots & \text{CUGUUAUUAGUAA, 03} & \text{REEACH} & 07, & \text{NU} \\ \text{Spy} & \text{TTTACA}, 17, & \text{PAAAT}, 22, & \text{UUACGCUAUUCCCCU} & \text{ACC} & \dots & \text{CUGUUA} & \text{CUUAUUGUAAUACAAUAUCUCGA} & \text{AO} & \text{REEACH} & 07, & \text{NU} \\ \text{Spy} & \text{TTTACA}, 17, & \text{PAAAT}, 22, & \text{UUACGCUAUUCCCCU} & \text{ACC} & \dots & \text{CUGUUA} & \text{CUUAUUCGCA} & \text{AO} & \text{REEACH} & 07, & \text{AU} \\ \text{Lsa} & \text{TTTTAA}, 17, & \text{PAAAT}, 22, & \text{UUACGCUAUUCCCCU} & \text{ACC} & \dots & \text{CUGUUA} & \text{CAUAUUUCGG} & 04, & \text{ACCA} & \text{CUUAUUCGG} & 04, & \text{ACCA} & 06 & \text{AU} & \text{AU} & \text{CUUAUUUCGG} & 04, & \text{ACCA} & 07, & \text{AU} \\ \text{Lsa} & \text{TTTTAA}, 17, & \text{PAAAT}, 12, & \text{AUUACGAUUAUCCCCU} & \text{CAAA} & \dots & \text{CUAUAUUCGG} & 04, & \text{ACCA} & 06 & \text{ACCA} & 06 & \text{ACCA} & 06 & \text{ACCA} & 00, & 07, & \text{AU} \\ \text{Fu} & \text{TTCAC}, 17, & \text{PAAAT}, 12, & \text{AUUCGAUAUUCCCCU} & \text{AACUA} & \dots & \text{UAA} & \dots & \text{UUA} & \text{AUUACUUUUU} & 04, & \text{ACCA} & 00, & 06, & 06, & 07, & 07, & 00,$		
$ \begin{array}{c} Lp1 \\ PTCGCGT, 18, PATTCT, 21, UUAAC GAUGUUC CGCUG, AC,, CAGGUU,, GU, GA CGAAUGUC GG, 04, AGGAAN, 09, AU Efa TTTACA, 17, PAAACT, 28, AUUACGAUGUC CGCUG, UGG, CA, GAGA, UGACCA, UGACAAUAUU U, 06, AGGAC, 07, AU for TTTACT, 17, PAAACT, 29, UUAUC GCUAUUC CGCUG, GCA,, AAG,, UGUGUCA, UGACUA, AUGUC, 03, AGGAGH, 08, AU Lac TTTAAA, 17, TACTT, 39, UUAUC GCUAUUC CGCUG, ACA, CACAGAGU, UGCUCU, UAAGAUUAUU, AA, 03, AGGAGH, 08, AU Lac TTTAAA, 17, TACTT, 39, UUAUC GCUAUUC CGCUG, ACA, CACAGUA,, GGUUAGUCAAUGCCGA, 03, AGGAGH, 08, AU Lac TTTAAA, 17, TACTT, 39, UUAUC GCUAUUC CGCUG, ACG,, CACAGUA,, GGUUAGUCAAUGCCGA, 03, AGGAGH, 08, AU Las TTTTAA, 17, TAATAT, 26, ACAAC GAUAUUC CGCUG, ACG,, CACAGUA,, GGUUAGUCAAUGCCG, 06, AGGAG, 06, AU Lsa TTTTAA, 17, CAAAAT, 29, UUACC GAUAUUC CGCUG, GCG,, CACAGUA,, GGUUAGUGAUAAUGUG, 06, AGGAGH, 07, AU Lsa TTTTAA, 17, CAAAAT, 29, UUACCGAUAUUC CGCUG, GCG,, CACAGUA,, GGUUAGUAAU, 03, AGGAGH, 08, AU Lsa TTTTAA, 17, CAAAAT, 29, UUACCGAUAUUC CGCUG, GCG,, CAAGA,, GGUUAAU, 04, 04, 03, AUGAA, 03, AGGAGH, 06, 10, AU Lsa TTTTAA, 17, CAAAAT, 29, UUACCGAUAUUC CGCUG, GCG,, CAAGA,, GGUUAAU, 04, 04, 04, 03, 04, 04, 05, 04, 04, 04, 04, 04, 04, 04, 04, 04, 04$	-	
$ \begin{array}{c} Ffs & TTTACA . 17 & TAAACT . 28 & AUUACAAUAUUCCCCUG . UGG . CA GAAG UGACCA . UAACAAUAUUUG$		
Ljo TTTACA 17 TAAACT 25 UUAUCGUAUUCGCUG GCAC AAG	•	
sth PAGACA 17 TAAGAT 29 UAACCGCUAAUCCGCUC AGA CAC CAGAGGU UGCUU UAA GAUUACUAA 03. AGGAGI 08. AU TTAATAA 17 TTACTT 39 UUAACGUAAUCCGCUG AGC CCUGUAACCGCAAUCCCCGA 03. AGGAGA 10 AU Spy TTTACA 17 TACAAT 29 UUACCGUAAUCCGCUA AC AC ACCAUGA ACAAGUA CUUAACAUAAUACUAA 03. AGGAGA 10 AU Spy TTTAACA 17 TAAAAT 26 ACAACGUAAUCCGCUA AC ACAAGUA CU UAACAUUACUAA 03. AGGAGA 06 AU Lsa TTTACT 17 TAATTT 24 AUACGAUAUUCCGCUG GCC AACAAGUA CU UAACAUUACUC GC .04. AGGAAT .07 AU Ls1 TTTACT 17 TAATTT 24 AUACGAUAUUCCGCUG GCC AACUC CAACAUGAAUCUUG GC .04. AGGAAT .07 AU Ls1 TTTACT 17 TAATTT 24 AUACGAUAUUCCGCUG CCC AACUC CAACAUGAAUCUUG GC .04. AGGAAT .07 AU Fnu TTGACA 17 TAAAAT 12 AAUUCGAUAUUCCGCUG .C AACUC CAACAUGAAUCUUG GC .04. AGGAAT .07 AU Fnu TTGACA 17 TAAAAT .12 AAUUCGAUAUUCCGCUU UAA UAAA UUA AAUGAAUAUCUU .04. AGGAAT .07 AU Fnu TTGACA 17 TAAAAT .12 AAUUCGAUAUUCCGCUU UAA UAAA UUA AAUGAAUCUUG .04. AGGAAT .07 AU Fnu TTGACA .17 TAAAAT .12 AAUUCGAUAUUCCGCUU UAA UAAAA UUA AAUGAAUCUU .04. AGGAAT .07 AU Fnu TTGACA .17 TAAAAT .12 AAUUCGAUAUUCCGCUU UAA UAAA UUA AAUGAUAUCUU .04. AGGAAT .07 AU Fnu TTGACA .17 TAAAAT .12 AAUUCGUUCGCUU UAA UAAA UUA AAUGAUUCUU .04. AGGAAT .07 AU G C G G G G G G G G G G G G G G G G G G		
Lac DTAAAA 17. TACATT 39. UUAUG GGUAUUC CGCUG ACC CUGGUA GGUUGAUGAAUGCC GA.03. AGAAG 10. AV Spy TTTTACA.17. TAGAAAT 29. UUACGCUAUCCGCUG ACC CAAGAAU CU.UAAGAAUAUCGAA.03. AGGAGA.06. AU Lsa TTTTACA.17. TAAAAT 26. ACAACGAUAUUCCGCUG GCC CAAGAAL CUUUAAUGAAUAUCUG.06. AGGAGA.06. AU Lsa TTTTACT.17. CATTTT 24. AUACGAUAUUCCGCUG GCC AACUG CAAGAAU GGUUAAUGAAUAUCUG.06. AGGAGA.07. AV Lsa TTTGACA.17. TAAAAT 12. AUAUCGAUAUUCCGCUG GCG AACUG CAAGAAUAUCUG.06. AGGAGA.07. AV Fau TTGACA.17. TAAAAT 12. AUUCGAUAUUCCGCUU. UAA UAAA UUA.AAUGAAUAUCUG.06. AGGAGA.07. AV Fau TTGACA.17. TAAAAT 12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA. AAUGAAUAUCUU.04. AGGAAAG.02. AV Fau TTGACA.17. TAAAAT 12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA. AAUGAAUAUCUU.04. AGGAAG.02. AV Fau TTGACA.17. TAAAAT .12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA. AAUGAAUAUCUU.04. AGGAAG.02. AV Fau TTGACA.17. TAAAAT .12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA. AAUGAAUAUCUU.04. AGGAAG.02. AV Fau TTGACA.17. TAAAAT .12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA. AAUGAAUAUCUU.04. AGGAAG.02. AV Fau TTGACA.17. TAAAAT .12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA AAUGAAUAUCUU.04. AGGAAG.02. AV Fau TTGACA.17. TAAAAT .12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA AAUGAAUUCUU.04. AGGAAG.02. AV C G G G G G G G G G G G G G G G G G G G	Ljo TTTACA.17.	TAAACT.25.UUAUGGGUAUUCCGC <mark>UG.GCAC</mark> AAG <mark>GUGU</mark> UG <mark>A</mark> UGAAUGCCGU.03. <mark>AGGAGA</mark> .07.AU
$\begin{array}{c} spy \\ represent \\ repre$	Sth TACACA 17	
LSG ITTTAAN 17. TAANAA, 26. ACAACGAUAUUCCGCUG. GCG CAAGA CGUUAAUGAUAUCUG. 06. AGGAGA. 07. AU LSI TTTACT. 17. TAATTT .24. AUAACGAUAUUCCGCUG. C AACUG CACAUGAUCUGG. 04. AGGAAA. 07. AU Fnu TTCACA. 17. TAATTT .24. AUUACGAUAUUCCGCUU. UAA UAAA UUA. AAUGAAUAUCUG. 06. AGGAAA. 07. AU Fnu TTCACA. 17. TAATTT .24. AUUACGAUAUUCCGCUU. UAA UAAA UUA. AAUGAAUAUCUG. 06. AGGAAA. 07. AU Fnu TTCACA. 17. TAATTT .24. AUUACGAUAUUCCGCUU. UAA UAAA UUA . AAUGAAUAUCUG. 06. AGGAAA. 07. AU Fnu TTCACA. 17. TAATTT .24. AUUACGAUAUUCCGCUU. UAA UAAA UUAAA UUA . AAUGAAUAUCUU. 04. AGGAAA. 07. AU Fnu TTCACA. 17. TAATTT .24. AUUACGAUAUUCCGCUU. UAA UAAA UUAAA UUA . AAUGAAUAUCUU. 04. AGGAAA. 07. AU Fnu TTCACA. 17. TAATTT .24. AUUACGAUAUUCCGCUU. UAA UAAA UUAAA UUAAA N 97% 97% N 97% 97% N 90% 90% N 75% 75% C 50% C		TAAGAT . 29. OAACGCCOAOCCGCOG . AGA . CACAGAGGO OGCOCO . OAAGAOOAGOAA. 03. AGGAGO . 08. AO
Ls 1 Fru TTCACA, 17. TALAAA, 12. AAUACGAUUUCCGCUU. UAAUAACUG		
F nu T CACA, 17. TAAAA , 12. AAUUC CAUAUUC CCCUU. UAA UAAA UUA. AAU CAUAUC UU. 04. ACGAAG , 02. AU B P P C C C C C C C C	Lac TTAAAA.17.	TTACTT.39.UUAUG <mark>GGUAUUC</mark> CGC <mark>U</mark> G. <mark>ACG</mark> CUGGUACGUUG <mark>A</mark> UGAAUGCCGA.03. <mark>AGGAGA</mark> .10.AU
B P_{a} P_{c}	Lac TTAAAA.17. Spy TTTACA.17.	TTACTT.39.UUAUG <mark>GGUAUUC</mark> CGC <mark>UG.ACG</mark> CUGGUACGUUGAUGAAUGCCGA.03.AGGAGA.10.AU TAGAAT.29.UUACC <mark>GCUAAUC</mark> CGC <mark>UA.AG</mark> ACAAGUAC <mark>U.UA</mark> AGAUUAGUAA.03. <mark>AGGAGA</mark> .06.AU
$P2 \xrightarrow{I}_{G} \xrightarrow{I}_{G$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17.	TTACTT.39.UUAUGGGUAUUCCGC <mark>UG.ACG</mark> CUGGUA <mark>CGU</mark> UG <mark>AUGAAUGCC</mark> GA.03. <mark>AGGAGA</mark> .10.AU TAGAAT.29.UUACGGCUAAUCCGC <mark>UA.AG</mark> ACAAGUA <mark>CU.UAAGAUUAGU</mark> AA.03. <mark>AGGAGA</mark> .06.AU TAAAAT.26.ACAAC <mark>GAUAUUC</mark> CGC <mark>U</mark> G. <mark>GCG</mark> CAAGA <mark>CGU</mark> UA <mark>A</mark> UGAAUAUCUG.06. <mark>AGGAGA</mark> .07.AU
$P2 \xrightarrow{I}_{G} \xrightarrow{I}_{G$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17.	TTACTT.39.UUAUGGGUAUUCCGCUG.ACGCUGGUACGUUGAUGAAUGCCCGA.03.AGGAGA.10.AU TAGAAT.29.UUACGGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AU TAAAAT.26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AU TATTTT.24.AUAACGAUAUUCCGCUG.CCAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU
$P2 \xrightarrow{I}_{G} \xrightarrow{I}_{G$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17.	TTACTT.39.UUAUGGGUAUUCCGCUG.ACGCUGGUACGUUGAUGAAUGCCCGA.03.AGGAGA.10.AU TAGAAT.29.UUACGGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AU TAAAAT.26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AU TATTTT.24.AUAACGAUAUUCCGCUG.CCAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU
$P2 \xrightarrow{G} - G = G = G = G = G = G = G = G = G = G$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17. Fnu TTGACA.17.	TTACTT.39.UUAUCGGUAUUCCGCUG.ACGCUGGUACGUUGAUGAAUGCCGA.03.AGGAGA.10.AU TAGAAT.29.UUACCGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AU TAAAAT.26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AU TATTTT.24.AUAACGAUAUUCCGCUG.CAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT.12.AAUUCGAUAUUCCGCUU.UAAUAAAUUA.AAUGAAUAUCUU.04.AGGAAG.02.AU
$P2 \xrightarrow{G} - G = G = G = G = G = G = G = G = G = G$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17. Fnu TTGACA.17.	TTACTT. 39.UUAUGGUAUUCCGCUG.ACGCUGGUACGUUGAUGAUGAAUGCCGA.03.AGGAGA.10.AU TAGAAT.29.UUACGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AU TAAAAT.26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AU TATTTT.24.AUAACGAUAUUCCGCUG.CAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT.12.AAUUCGAUAUUCCGCUG.CAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT.12.AAUUCGAUAUUCCGCUU.UAAUAAAUUA.AAUGAAUAUCUU.04.AGGAAG.02.AU
$P2 \xrightarrow{I}_{G} \xrightarrow{I}_{G$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17. Fnu TTGACA.17.	TTACTT. 39.UUAUCGGUAUUCCGCUG.ACGCUGGUACGUUGAUGAAUGCCGA.03.AGGAGA.10.AU TAGAAT.29.UUACCGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AU TAAAAT.26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUGUCGG.06.AGGAGA.07.AU TATTTT.24.AUAACGAUAUUCCGCUG.CCAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT.12.AAUUCGAUAUUCCGCUG.CAACUGQACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT.12.AAUUCGAUAUUCCGCUU.UAAUAAAUUAAAUGAAUAUCUU.04.AGGAAG.02.AU Ducleotide nucleotide
$P2 \xrightarrow{C} - G \xrightarrow{C} - $	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17. Fnu TTGACA.17.	TTACTT.39.UUAUCGGUAUUCCGCUG.ACGCUGGUACGUUGAUGAAUGCCGA.03.AGGAGA.10.AU TAGAAT.29.UUACGGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AU TAAAAT.26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AU TATTTT.24.AUAACGAUAUUCCGCUG.CCAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT.12.AAUUCGAUAUUCCGCUG.CAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT.12.AAUUCGAUAUUCCGCUU.UAAUAAAUUAA.AAUGAAUAUCUU.04.AGGAAG.02.AU Nucleotide Nucleotide Nucleotide Nucleotide
$P2 \xrightarrow{G} - G = Y = 3$ $P1 \xrightarrow{G} - R = Y = 3$ $P1 \xrightarrow{G} - R = Y = 3$ $G - C Watson-Crick \ base \ pair \\ G - C Watson-Crick \ base \ pair \\ G - C Watson-Crick \ base \ pair \\ G - C Watson-Crick \ base \ pair \\ G - C G = G = C $	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17. Fnu TTGACA.17.	TTACTT 39.UUAUCGGUAUUCCGCUG.ACGCUGGUACGUUGAUGAAUGCCGA.03.AGGAGA.10.AU TAGAAT 29.UUACGGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AU TAAAAT 26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AU TAATTTT 24.AUAACGAUAUUCCGCUG.GCGCAAGAGACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT 12.AAUUCGAUAUUCCGCUG.CAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT 12.AAUUCGAUAUUCCGCUU.UAA Inucleotide nucleotide identity present UUUU A ^C G UUUU UUUU
$P2 \xrightarrow{G} - G \xrightarrow{G} - $	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17. Fnu TTGACA.17.	TTACTT 39.UUAUCGGUAUUCCGCUG.ACGCUGGUACUUGAUGAUGAAUGCCGA.03.AGGAGA.10.AU TAGAAT 29.UUACGGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AU TAAAAT 26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AU TAATTTT 24.AUAACGAUAUUCCGCUG.GCGAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AU TAAAAT 12.AAUUCGAUAUUCCGCUU.UAAUAAAUUA.AAUGAAUAUCUU.04.AGGAAG.02.AU Inucleotide nucleotide identity present N 97% 97% 97%
$\begin{array}{c} G & - & G \\ G & - & Y \\ G & - & Y \\ G & - & Y \\ G & - & G \\$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17. Fnu TTGACA.17.	TTACTT . 39. UUAUGGUAUUCCGCUG. ACG CUGGUA CGUUGAUGAAUGCCGA. 03. AGGAGA .10. AU TAGAAT .29. UUACGGCUAAUCCGCUA. AG ACAAGUA CU . UAAGAUUAGUAA. 03. AGGAGA .06. AU TAAAAT .26. ACAACGAUAUUCCGCUG. GCG CAAGA CGUUAAUGAAUAUCUG. 06. AGGAGA .07. AU TATTTT .24. AUAACGAUAUUCCGCUG . C AACUG GACAUGAAUGUCGG. 04. AGGAAA .07. AU TAAAAT .12. AAUUCGAUAUUCCGCUU . UAA UAAA UUA . AAUGAAUAUCUU .04. AGGAAA .02. AU N 97% 97% 97% 97% G 97% G G G A U G U
$\begin{array}{c} G & - & G \\ G & - & Y \\ G & - & Y \\ G & - & Y \\ G & - & G \\$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17. Fnu TTGACA.17.	TTACTT . 39. UUAUGGUAUUCCGCUG. ACG CUGGUA CGUUGAUGAAUGCCGA. 03. AGGAGA .10. AU TAGAAT .29. UUACGGCUAAUCCGCUA. AG ACAAGUA CU . UAAGAUUAGUAA. 03. AGGAGA .06. AU TAAAAT .26. ACAACGAUAUUCCGCUG. GCG CAAGA CGUUAAUGAAUAUCUG. 06. AGGAGA .07. AU TATTTT .24. AUAACGAUAUUCCGCUG . C AACUG GACAUGAAUGUCGG. 04. AGGAAA .07. AU TAAAAT .12. AAUUCGAUAUUCCGCUU . UAA UAAA UUA . AAUGAAUAUCUU .04. AGGAAA .02. AU N 97% 97% 97% 97% G 97% G G G A U G U
$\begin{array}{c} G - Y \\ G \\ C \\ U - A \\ G \\ C \\ U - A \\ G \\ C \\ U - A \\ G \\ C \\ C \\ G \\ C \\ C \\ G \\ C \\ C \\ G \\ C \\ C$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B	TTACTT 39. UUAUGGUAUUCCGCUG. ACG CUGGUA CGUUGAUGAAUGCCGA. 03. AGGAGA. 10. AU TAGAAT 29. UUACGCUAAUCCGCUA. AG ACAAGUA CU . UAAGAUUAGUAA. 03. AGGAGA. 06. AU TAAAAT 26. ACAACGAUAUUCCGCUG. GCG AACAGUA CGUUAAUGAAUAUCUG. 06. AGGAGA. 07. AU TATTTT 24. AUAACGAUAUUCCGCUG. C AACUG GACAUGAAUGUCGG. 04. AGGAAA. 07. AU TAAAAT 12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA. AAUGAAUGUCUGG. 04. AGGAAA .07. AU TAAAAT 12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA . AAUGAAUAUCUU. 04. AGGAAG .02. AU N 97% 97% N 90% 90% N 75% 75% 75% 6 75% 6 75% 75% 75% 75% 75% 75% 75% 75% 75% 75%
$\begin{array}{c} C & J & A \\ G & C & G \\ U & A \\ U & A \\ S' & G & Y \\ S' & G & Y \\ S' & G & Y \\ \end{array} $ $\begin{array}{c} C & J & A \\ C & C & G \\ U & A \\ C & C \\ S' & G & A \\ G & C \\ G & A \\ G & G \\ G$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B	TTACTT 39. UUAUGGUAUUCCGCUG. ACG CUGGUA CUGUAUGAUGAAUGCCGA. 03. AGGAGA. 10. AU TAGAAT 29. UUACGCUAAUCCGCUA. AG ACAAGUA CU . UAAGAUUAGUAA. 03. AGGAGA .06. AU TAAAAT 26. ACAACGAUAUUCCGCUG. GCG ACAAGUA CGUUAAUGAAUAUCUG. 06. AGGAGA .07. AU TATTTT 24. AUAACGAUAUUCCGCUG. C AACUG GACAUGAAUGUCGG. 04. AGGAAA .07. AU TAAAAT 12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA . AAUGAAUGUCUGG. 04. AGGAAA .07. AU TAAAAT 12. AAUUCGAUAUUCCGCUU. UAA UAAA UUA . AAUGAAUAUCUU. 04. AGGAAG .02. AU N 97% 97% N 90% 90% N 75% 975% S 50%
$\begin{array}{c} G \\ C \\$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B	TTACTT 39. UUAUG GGUAUUC CGCUG ACG CUGGUA CUGUUGAU GAAUGCC GA.03. AGGACA.10. AU TAGAAT 29. UUACG GCUAAUC CGCUA AG ACAAGUA CU .UAAGAUUAGUAA.03. AGGACA.06. AU TAAAAT 26. ACAAC GAUAUUC CGCUG GCG CAAGA CGUUAAU GAAUAUC UG.06. AGGACA.07. AU TATTTT 24. AUAAC GAUAUUC CGCUG C AACUG GACAU GAAUGUC GG.04. AGGAAA.07. AU TAAAAT .12. AAUUC GAUAUUC CGCUU .UAA UAAA UUA . AAU GAAUAUC UU .04. AGGAAA .07. AU TAAAAT .12. AAUUC GAUAUUC CGCUU .UAA UAAA UUA . AAU GAAUAUC UU .04. AGGAAC .02. AU N 97% 97% N 90% 90% N 75% 75% C G G G A U G - G G G G G U U U U A G G G - C U U U U U A G G G - C U U U U U A A G G G - C U U U U G G - C A AGAGC
P1 $\begin{bmatrix} C & -G \\ U & -A \\ \hline C & - & C \\ \hline C & - & G \\ \hline U & - & A \\ \hline G & - & C \\ \hline U & - & A \\ \hline G & - & C \\ \hline G &$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17.	TTACTT .39. UUAUGGUAUUC CGCUG . ACG CUGGUA CGUUGAU GAAUGCC GA .03. AGGAGA .10. AU TAGAAT .29. UUACGCUAAUC CGCUA . AG ACAAGUA CU . UAAGAUUAGUAA .03. AGGAGA .06. AU TAAAAT .26. ACAACGAUAUUC CGCUG .GCG CAAGA CGUUAAUGAAUAUC UG .06. AGGAGA .07. AU TATTTT .24. AUAACGAUAUUC CGCUG .C AACUG GACAUGAAUGUC GG .04. AGGAAA .07. AU TAAAAT .12. AAUUC CAUAUUC CGCUU . UAA UAAA UUA. AAUGAAUAUC UU .04. AGGAAG .02. AU TAAAAT .12. AAUUC CAUAUUC CGCUU . UAA UAAA UUA. AAUGAAUAUC UU .04. AGGAAG .02. AU TAAAAT .12. AAUUC CAUAUUC CGCUU . UAA UAAA UUA . AAUGAAUAUC UU .04. AGGAAG .02. AU TAAAAT .12. AAUUC CAUAUUC CGCUU . UAA UAAA UUA . AAUGAAUAUC UU .04. AGGAAG .02. AU M 97% 97% 97% N 90% 90% N 75% 0 75% O 50% stem loop U G - C G G U 3' U U U G - C AGGAG C
P1 \overrightarrow{P} \overrightarrow{P} \overrightarrow{P} \overrightarrow{P} \overrightarrow{G} \overrightarrow{P} \overrightarrow{G} \overrightarrow{P} \overrightarrow{G}	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17.	TTACTT . 39. UUAUG GGUAUUC CGCUG . ACG CUGGUA CGUUGAU GAAUGCC GA. 03. AGCACA. 10. AU TAGAAT . 29. UUACG GCUAAUC CGCUA . AG ACAAGUA CU . UAA GAUUAGUAA. 03. AGCACA. 06. AU TAAAAT . 26. ACAAC GAUAUUC CGCUG . GCG CAAGA CGUUAAU GAAUAUC UG. 06. AGGACA. 07. AU TATTTT . 24. AUAAC GAUAUUC CGCUG . C AACUG GACAU GAAUGUC GG. 04. AGGAAA. 07. AU TAAAAT . 12. AAUUC GAUAUUC CGCUU . UAA UAAA
$\begin{array}{c c} G & - C & Watson-Crick base pair \\ 5' - G & - Y - 3' \\ \hline G & A & other base interaction \\ \hline G & A & other base interaction \\ \hline G & - C & U & G \\ \hline G & - C & U & G \\ \hline G & - C & U & G \\ \hline G & - C & U & G \\ \hline G & - C & - C & U & G \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C & - C & - C & - C \\ \hline G & - C & $	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 C - C - C - C - C - C - C - C - C - C -	TTACTT 39.UUAUGGUAUUC CGCUG ACG CUGUG ACG CUGGUA CGUGAUGAUGAUGCCGA.03.AGGAGA.10.AU TAGAAT 29.UUACGCUAUC CGCUG AG CAGA AG CUUAUGUAA.03.AGGAGA.06.AU TAAAAT 26.ACAACGAUAUUC CGCUG CCG CAAGA CUUAUGUAAUGUA
$\begin{array}{c c} G & - C & Watson-Crick base pair \\ 5' - G & - Y - 3' \\ \hline G & A & other base interaction \\ \hline G & A & other base interaction \\ \hline G & - C & U & G \\ \hline G & - C & U & G \\ \hline G & - C & U & G \\ \hline G & - C & U & G \\ \hline G & - C & - C & U & G \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C & - C & - C & - C \\ \hline G & - C & $	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 C - C - C - C - C - C - C - C - C - C -	TTACTT . 39. UUAUC GCUAUUC CGCUG . ACG CUGGUA CCUUGAU CAUGCC GA.03. AGGAGA .10. AU TAGAAT .29. UUACC GCUAAUC CGCUA . AG ACAAGUA CU . UAA GAUUAGU AA.03. AGGAGA .06. AU TAAAAT .26. ACAAC GAUAUUC CGCUG . GCG CAAGA CGUUAAU GAAUAUC UG .06. AGGAGA .07. AU TAATTTT .24. AUAAC GAUAUUC CGCUG . C AACUG GACAU GAAUAUC UG .06. AGGAGA .07. AU TAATTTT .24. AUAAC GAUAUUC CGCUU . UAA UAAA UUA . AAU GAAUAUC UG .06. AGGAGA .07. AU TAATAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UUA . AAU GAAUAUC UU .04. AGGAAA .07. AU TAAAAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UUA . AAU GAAUAUC UU .04. AGGAAA .07. AU TAAAAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UAAA UUA . AAU GAUAUC UU .04. AGGAAG .02. AU N 97% 97% N 90% 90% N 75% 75% Stem loop always present compensatory mutations C GC GUUGUCU
$\begin{array}{c c} G & - C & Watson-Crick base pair \\ 5' - G & - Y - 3' \\ \hline G & A & other base interaction \\ \hline G & A & other base interaction \\ \hline G & - C & U & G \\ \hline G & - C & U & G \\ \hline G & - C & U & G \\ \hline G & - C & U & G \\ \hline G & - C & - C & U & G \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C \\ \hline G & - C & - C & - C & - C & - C & - C & - C & - C \\ \hline G & - C & $	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 C - C - C - C - C - C - C - C - C - C -	TTACTT . 39. UUAUC GCUAUUC CGCUG . ACG CUGGUA CCUUGAU CAUGCC GA.03. AGGAGA .10. AU TAGAAT .29. UUACC GCUAAUC CGCUA . AG ACAAGUA CU . UAA GAUUAGU AA.03. AGGAGA .06. AU TAAAAT .26. ACAAC GAUAUUC CGCUG . GCG CAAGA CGUUAAU GAAUAUC UG .06. AGGAGA .07. AU TAATTTT .24. AUAAC GAUAUUC CGCUG . C AACUG GACAU GAAUAUC UG .06. AGGAGA .07. AU TAATTTT .24. AUAAC GAUAUUC CGCUU . UAA UAAA UUA . AAU GAAUAUC UG .06. AGGAGA .07. AU TAATAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UUA . AAU GAAUAUC UU .04. AGGAAA .07. AU TAAAAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UUA . AAU GAAUAUC UU .04. AGGAAA .07. AU TAAAAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UAAA UUA . AAU GAUAUC UU .04. AGGAAG .02. AU N 97% 97% N 90% 90% N 75% 75% Stem loop always present compensatory mutations C GC GUUGUCU
$5' - G - Y - 3'$ $G \cdot A$ other base interaction $5' - CG - CUGUU$ $G \cdot U$	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 $\begin{array}{c} & & \\ $	TTACTT . 39. UUAUG GGUAUUC CGCUG . ACC CUGGUA CGUUGAU GAAUGCC GA .03. AGGAGA .10. AU TAGAAT .29. UUACG GCUAAUC CGCUA .AC ACAAGUA CU . UAA GAUUAGUAA .03. AGGAGA .06. AU TAAAAT .26. ACAAC GAUAUUC CGCUG .GCG CAAGA CGUUAAU GAAUAUC UG .06. AGGAGA .07. AU TAATTT .24. AUAAC GAUAUUC CGCUG .C AACUG GACAU GAAUGUC GG .04. AGGAAA .07. AU TAATAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UUA .AAU GAAUAUC UU .04. AGGAAA .07. AU TAAAAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UUA .AAU GAAUAUC UU .04. AGGAAA .02. AU TAAAAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UUA .AAU GAAUAUC UU .04. AGGAAA .02. AU TAAAAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UUA .AAU GAAUAUC UU .04. AGGAAA .02. AU TAAAAT .12. AAUUC GAUAUUC CGCUU . UAA UAAA UUA . AAU GAAUAUC UU .04. AGGAAA .02. AU G G G G A U U U U A A G G G G G G G G G G G G G G G
	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 $\begin{array}{c} & & \\ $	TTACTT 39. UUAUG GGUAUUC CGCUG ACG CUGUGA CUGGUA CAAUGCC GA 03. AGGAGA 10. AU TAGAAT 29. UUACGCUAAUC CGCUG AGG CC CCAGA CA CAAGUA CUUGAAUAUC UG 06. AGGAGA 06. AU TAAAAT 26. ACAACGAUAUUC CGCUG GCC CAAGA CCAAGA CUUGGAUGAUGAUGUG GG 04. AGGAAC 7. AU TATTTT 24. AUAACGAUAUUC CGCUG C C C AACUG COUGAUGAAUGUC GG 04. AGGAAC 7. AU TAAAAT 12. AAUUC GAUAUUC CGCUU UAA CUUAAA CUUA AAUGAAUAUC UU 04. AGGAAC 02. AU TAAAAT 12. AAUUC GAUAUUC CGCUU UAA CUUAAA CUUA AAUGAAUAUC UU 04. AGGAAC 02. AU TAAAAT 12. AAUUC GAUAUUC CGCUU UAA CUUAAA CUUA AAUGAAUAUC UU 04. AGGAAC 02. AU TAAAAT 12. AAUUC GAUAUUC CGCUU UAA CUUAAA CUUAAA CUUA AAUGAAUAUC UU 04. AGGAAC 02. AU TAAAAT 12. AAUUC GAUAUUC CGCUU UAA CUUAAA CUUAAA CUUA AAUGAAUAUC UU 04. AGGAAC 02. AU C C C C G U U A A C G U U U U A A GGAAC 02. AU C G G U A A G G G G G G G G G G G G G G G
	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 $\begin{array}{c} & & \\ $	TTACTT . 39. UUAUG GGUAUUG CGC G . ACG CUGGUA CGUUGAU GAAUGCC GA.03. AGGAGA. 10. AU TAGAAT . 29. UUACG CCUAAUC CGC JA. AG ACAAGUA CU. UAA GAUUAGUAA.03. AGGAGA.06. AU TAAAAT . 26. ACAACGAUAUUC CGC G . CCG CAAGA CGUUAAUGAAUAUC G.06. AGGAGA.06. AU TAAAAT . 26. ACAACGAUAUUC CGC G . CCG CAAGA CGUUAAUGAAUAUC G.06. AGGAGA.07. AU TAAAAT . 24. AUAACGAUAUUC CGC G . C AACUG CACAU GAAUAUC GG .04. AGGAAA.07. AU TAAAAT .12. AAUUG GAUAUUC CGC U . UAA UAAA UUA . AAU GAAUAUC U 04. AGGAAG .02. AU TAAAAT .12. AAUUG GAUAUUC CGC U UAA UAAA UUA . AAU GAAUAUC U 04. AGGAAG .02. AU TAAAAT .12. AAUUG GAUAUUC CGC U UAA UAAA UUA . AAU GAAUAUC U 04. AGGAAG .02. AU V U U U G G G G A G G G G G G G G G G G G
	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 $\begin{array}{c} & & \\ $	TTACTT . 39. UUAUG GGUAUUG CGC G . ACG CUGGUA CGUUGAU GAAUGCC GA.03. AGGAGA. 10. AU TAGAAT . 29. UUACG CCUAAUC CGC JA. AG ACAAGUA CU. UAA GAUUAGUAA.03. AGGAGA.06. AU TAAAAT . 26. ACAACGAUAUUC CGC G . CCG CAAGA CGUUAAUGAAUAUC G.06. AGGAGA.06. AU TAAAAT . 26. ACAACGAUAUUC CGC G . CCG CAAGA CGUUAAUGAAUAUC G.06. AGGAGA.07. AU TAAAAT . 24. AUAACGAUAUUC CGC G . C AACUG CACAU GAAUAUC GG .04. AGGAAA.07. AU TAAAAT .12. AAUUG GAUAUUC CGC U . UAA UAAA UUA . AAU GAAUAUC U 04. AGGAAG .02. AU TAAAAT .12. AAUUG GAUAUUC CGC U UAA UAAA UUA . AAU GAAUAUC U 04. AGGAAG .02. AU TAAAAT .12. AAUUG GAUAUUC CGC U UAA UAAA UUA . AAU GAAUAUC U 04. AGGAAG .02. AU V U U U G G G G A G G G G G G G G G G G G
	Lac TTAAAA.17. Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 $\begin{array}{c} & & \\ $	TTACTT 39. UUAUGGUAUUCCGCUG ACG CUGGUA CUUGAUGAUGAUGCCGA.03. NGGAGA.10. AU TAGAAT 29. UUACGCUAAUCCGCUA. AG ACAAGUA CU. UIAGAUUAUGA.03. AGGAGA.06. AU TAGAAT 29. UUACGCUAUUCCGCUG GCG AACAGA CU. UIAGAUUAUGUG.06. NGGAGA.06. AU TAAAAT 26. ACAACGAUAUUCCGCUG GCG AACAGA CGUUGAUGAUAUGUG.06. NGGAGA.07. AU TATTT 24. AUAACGAUAUUCCGCUG G AACUG GAUGA UUA. AAUGAAUAUCUG.06. NGGAAG.07. AU TAAAAT 12. AAUUCGAUAUUCCGCUG G AACUG GAUGAUUAUCUU.04. NGGAAG.07. AU TAAAAT 12. AAUUCGAUAUUCCGCUU . UAA UAAA UUA. AAUGAAUAUCUU.04. NGGAAG.02. AU TAAAAT 12. AAUUCGAUAUUCCGCUU . UAA UAAA UUA AAUGAAUAUCUU.04. NGGAAG.02. AU TAAAAT 7. 12. AAUUCGAUAUUCCGCUU . UAA UAAA UUA AAUGAAUAUCUU.04. NGGAAG.02. AU TAAAAT 7. 12. AAUUCGAUAUUCGGCUU . UAA UAAA UUA AAUGAAUAUCUU.04. NGGAAG.02. AU TAAAAT 7. 12. AAUUCGAUAUUCGGCUU . UAA UAAA UUA AAUGAAUAUCUU.04. NGGAAG.02. AU C G G G G U A A G G G G U G G G G G U G G G G

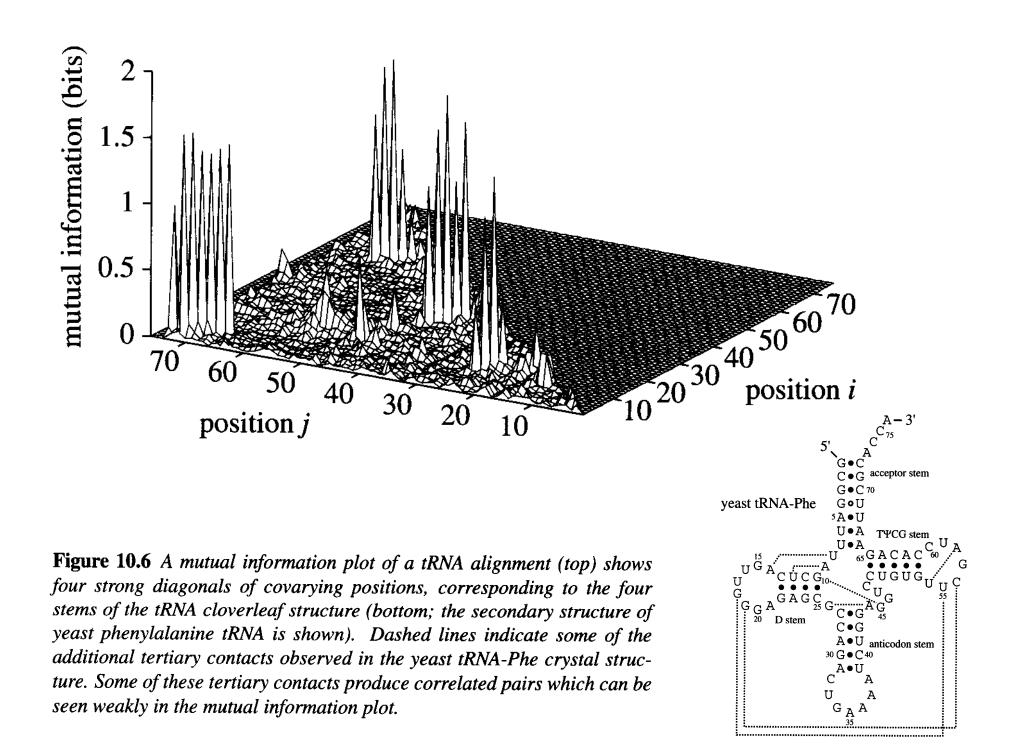
Excess L19 represses L19 (RF00556; 555-559 similar) Example: Ribosomal Autoregulation:

22

Mutual Information

$$M_{ij} = \sum_{xi,xj} f_{xi,xj} \log_2 \frac{f_{xi,xj}}{f_{xi}f_{xj}}; \quad 0 \le M_{ij} \le 2$$

Max when *no* seq conservation but perfect pairing MI = expected score gain from using a pair state (below) Finding optimal MI, (i.e. opt pairing of cols) is hard(?) Finding optimal MI *without pseudoknots* can be done by dynamic programming



Computational Problems

How to predict secondary structure
How to model an RNA "motif" (l.e., sequence/structure pattern)
Given a motif, how to search for instances
Given (unaligned) sequences, find motifs
How to score discovered motifs
How to leverage prior knowledge

Motif Description

RNA Motif Models

"Covariance Models" (Eddy & Durbin 1994) aka profile stochastic context-free grammars aka hidden Markov models on steroids
Model position-specific nucleotide preferences and base-pair preferences

Pro: accurate

Con: model building hard, search slow

Eddy & Durbin 1994: What

A probabilistic model for RNA families The "Covariance Model" ~ A Stochastic Context-Free Grammar A generalization of a profile HMM Algorithms for Training From aligned or unaligned sequences Automates "comparative analysis" Complements Nusinov/Zucker RNA folding Algorithms for searching

Main Results

Very accurate search for tRNA (Precursor to tRNAscanSE - current favorite) Given sufficient data, model construction comparable to, but not quite as good as, human experts

Some quantitative info on importance of pseudoknots and other tertiary features

Probabilistic Model Search

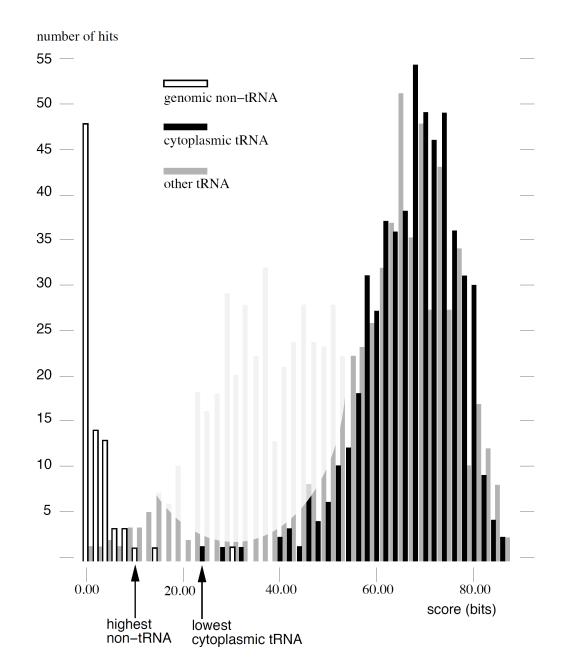
As with HMMs, given a sequence, you calculate likelihood ratio that the model could generate the sequence, vs a background model

You set a score threshold

Anything above threshold \rightarrow a "hit" Scoring:

"Forward" / "Inside" algorithm - sum over all paths Viterbi approximation - find single best path (Bonus: alignment & structure prediction)

Example: searching for tRNAs



Profile Hmm Structure

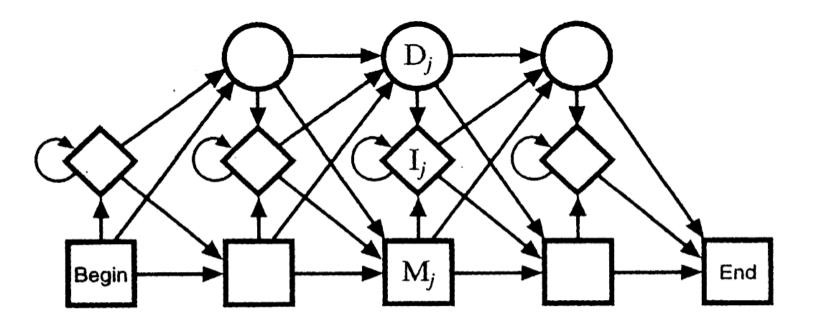


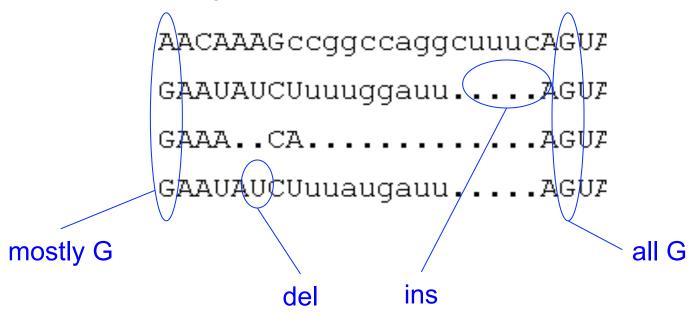
Figure 5.2 The transition structure of a profile HMM.

- M_j: Match states (20 emission probabilities)
- Ij: Insert states (Background emission probabilities)
- Dj: Delete states (silent no emission)

How to model an RNA "Motif"?

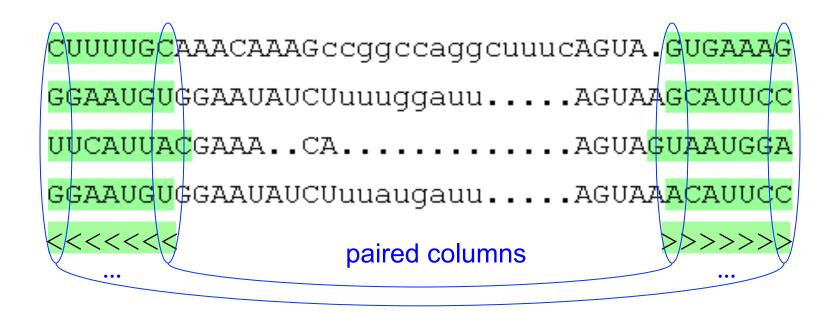
Conceptually, start with a profile HMM:

- from a multiple alignment, estimate nucleotide/ insert/delete preferences for each position
- given a new seq, estimate likelihood that it could be generated by the model, & align it to the model



How to model an RNA "Motif"?

Add "column pairs" and pair emission probabilities for base-paired regions



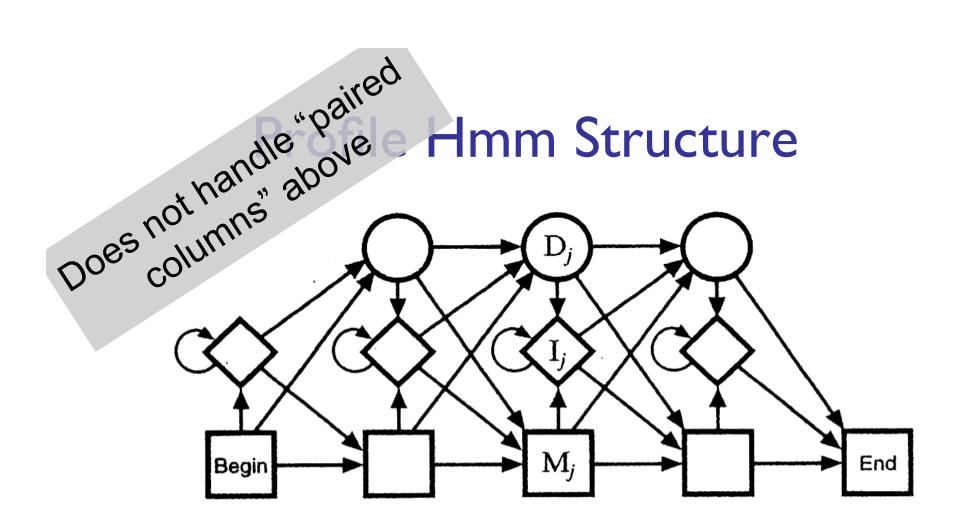


Figure 5.2 The transition structure of a profile HMM.

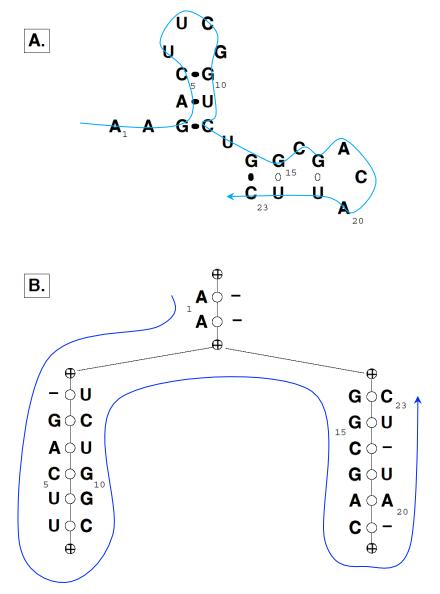
- M_j: Match states (20 emission probabilities)
- Ij: Insert states (Background emission probabilities)
- Dj: Delete states (silent no emission)

CM Structure

A: Sequence + structureB: the CM "guide tree"C: probabilities of

letters/ pairs & of indels

Think of each branch being an HMM emitting both sides of a helix (but 3' side emitted in reverse order)

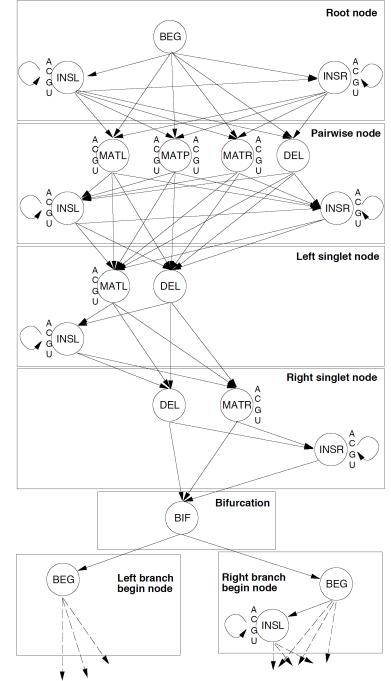


Overall CM Architecture

One box ("node") per node of guide tree

BEG/MATL/INS/DEL just like an HMM

MATP & BIF are the key additions: MATP emits *pairs* of symbols, modeling basepairs; BIF allows multiple helices



48

CM Viterbi Alignment (the "inside" algorithm)

$$x_i = i^{th}$$
 letter of input

$$x_{ij}$$
 = substring *i*,...,*j* of input

$$T_{yz} = P(\text{transition } y \rightarrow z)$$

$$E_{x_i,x_j}^{y} = P(\text{emission of } x_i, x_j \text{ from state } y)$$

 $S_{ij}^{y} = \max_{\pi} \log P(x_{ij} \text{ gen'd starting in state } y \text{ via path } \pi)$

CM Viterbi Alignment (the "inside" algorithm)

$$S_{ij}^{y} = \max_{\pi} \log P(x_{ij} \text{ generated starting in state } y \text{ via path } \pi)$$

$$\max_{z} [S_{i+1,j-1}^{z} + \log T_{yz} + \log E_{x_{i},x_{j}}^{y}] \quad \text{match pair}$$

$$\max_{z} [S_{i+1,j}^{z} + \log T_{yz} + \log E_{x_{i}}^{y}] \quad \text{match/insert left}$$

$$\max_{z} [S_{i,j-1}^{z} + \log T_{yz} + \log E_{x_{j}}^{y}] \quad \text{match/insert right}$$

$$\max_{z} [S_{i,j}^{z} + \log T_{yz}] \quad \text{delete}$$

$$\max_{i < k \le j} [S_{i,k}^{y_{left}} + S_{k+1,j}^{y_{right}}] \quad \text{bifurcation}$$

$$\prod_{compare: O(qn) \text{ for profile HMM}}$$

An Important Application: Rfam

Rfam – an RNA family DB Griffiths-Jones, et al., NAR '03, '05, '08, '11, '12

Was biggest scientific comp user in Europe - 1000 cpu cluster for a month per release

Rapidly growing:

 Rel
 1.0, 1/03:
 25 families,
 55k instances
 DB size:

 Rel
 7.0, 3/05:
 503 families,
 363k instances
 ~8GB

 Rel
 9.0, 7/08:
 603 families,
 636k instances
 ~8GB

 Rel
 9.0, 7/08:
 603 families,
 636k instances
 ~8GB

 Rel
 9.1, 1/09:
 1372 families,
 1148k instances
 ~160GB

 Rel
 10.0, 1/10:
 1446 families,
 3193k instances
 ~160GB

 Rel
 11.0, 8/12:
 2208 families,
 6125k instances
 ~320GB

RF00037: Example Rfam Family

Input (hand-curated):

MSA "seed alignment" SS_cons Score Thresh T Window Len W

Output:

CM

scan results & "full alignment" phylogeny, etc.

IRE (partial seed alignment):

DOA COC

Hom.sap.	GUUCC	JGC	UUCAA	CAGUGU	UUGGAU	GGAAC
Hom.sap.	<mark>UUUCU</mark> U	JC.	UUCAA	CAGUGU	UUGGAU	GGAAC
Hom.sap.	<mark>UUUCC</mark> U	JGU	UUCAA	CAGUGC	UUGGA.	<mark>GGAAC</mark>
Hom.sap.	UUUAU	2:	AGUGA	CAGAGU	UCACU.	AUAAA
Hom.sap.	<mark>UCUCU</mark> U	JGC	UUCAA	CAGUGU	UUGGAU	GGAAC
Hom.sap.	AUUAU	2	GGGAA	CAGUGU	UUCCC.	<mark>AUAAU</mark>
Hom.sap.	<mark>UCUUG</mark> C	2	UUCAA	CAGUGU	UUGGAC	GGAAG
Hom.sap.	UGUAU	2	GGAGA	CAGUGA	UCUCC.	<mark>AUAUG</mark>
Hom.sap.	AUUAU	2	GGAAG	CAGUGC	CUUCC.	<mark>AUAAU</mark>
Cav.por.	<mark>UCUCC</mark> U	JGC	UUCAA	CAGUGC	UUGGAC	GGAGC
Mus.mus.	UAUAU	2	GGAGA	CAGUGA	UCUCC.	<mark>AUAUG</mark>
Mus.mus.	<mark>UUUCC</mark> U	JGC	UUCAA	CAGUGC	UUGAAC	GGAAC
Mus.mus.	<mark>GUACU</mark> U	JGC	UUCAA	CAGUGU	UUGAAC	GGAAC
Rat.nor.	UAUAU	2	GGAGA	CAGUGA	ccucc.	<mark>AUAUG</mark>
Rat.nor.	UAUCU	JGC	UUCAA	CAGUGU	UUGGAC	GGAAC
SS_cons	<mark><<<<</mark> .	•••	<<<<		>>>> .	<mark>>>>></mark>

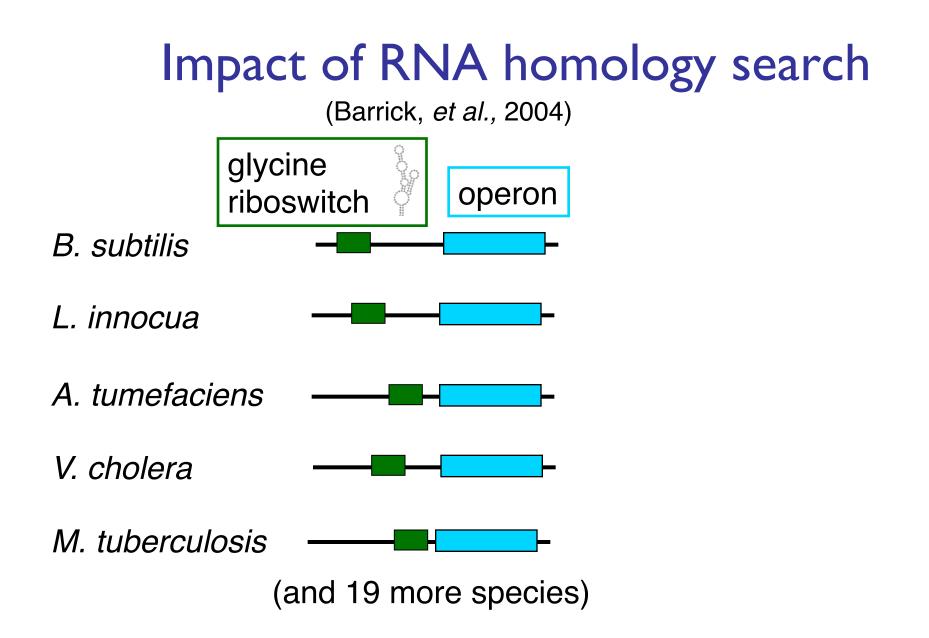
An Important Need: Faster Search

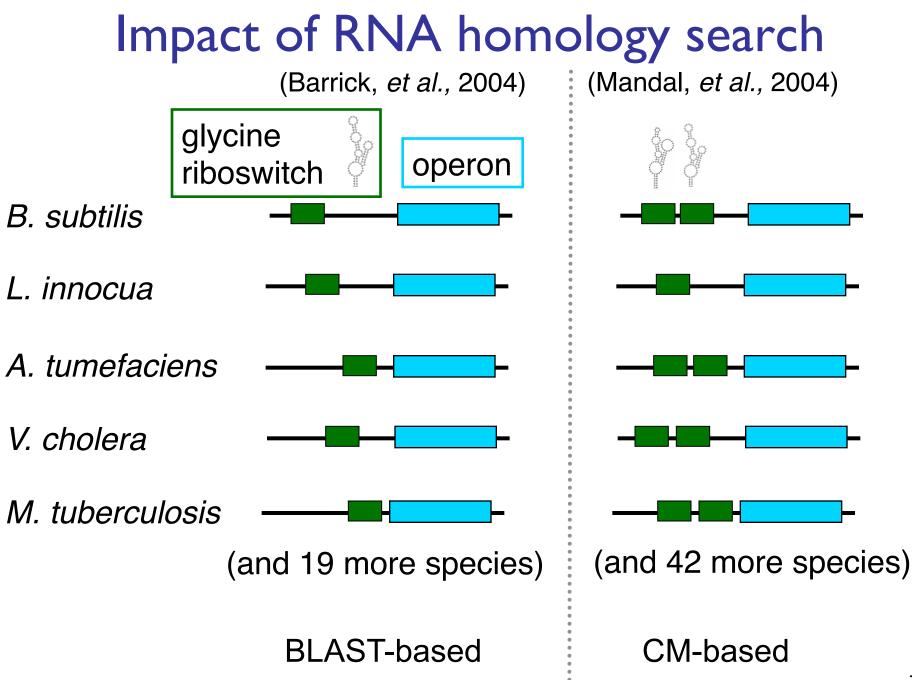
Homology search

"Homolog" – similar by descent from common ancestor Sequence-based Smith-Waterman FASTA BLAST

For RNA, sharp decline in sensitivity at ~60-70% identity

So, use structure, too



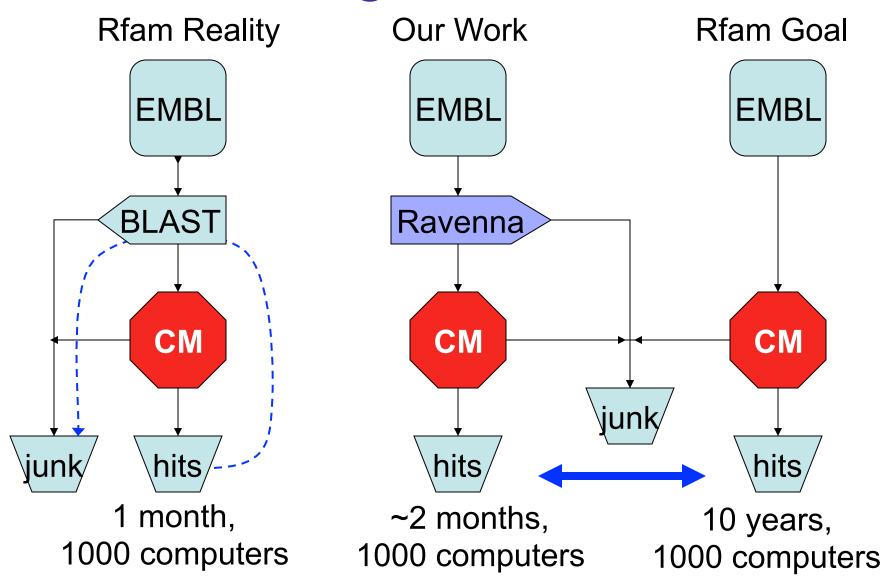


Faster Genome Annotation of Non-coding RNAs Without Loss of Accuracy Zasha Weinberg & W.L. Ruzzo Recomb '04, ISMB '04, Bioinfo '06

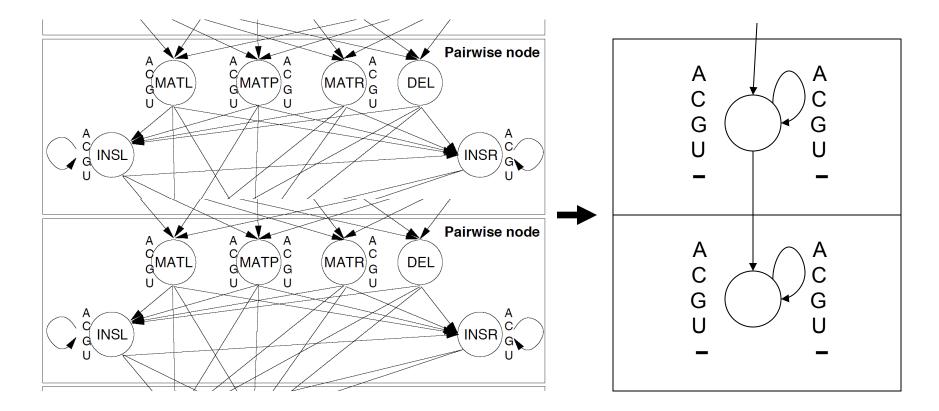
RaveNnA: Genome Scale RNA Search

Typically 100x speedup over raw CM, w/ no loss in accuracy: Drop structure from CM to create a (faster) HMM Use that to pre-filter sequence; Discard parts where, provably, CM score < threshold; Actually run CM on the rest (the promising parts) Assignment of HMM transition/emission scores is key (a large convex optimization problem)

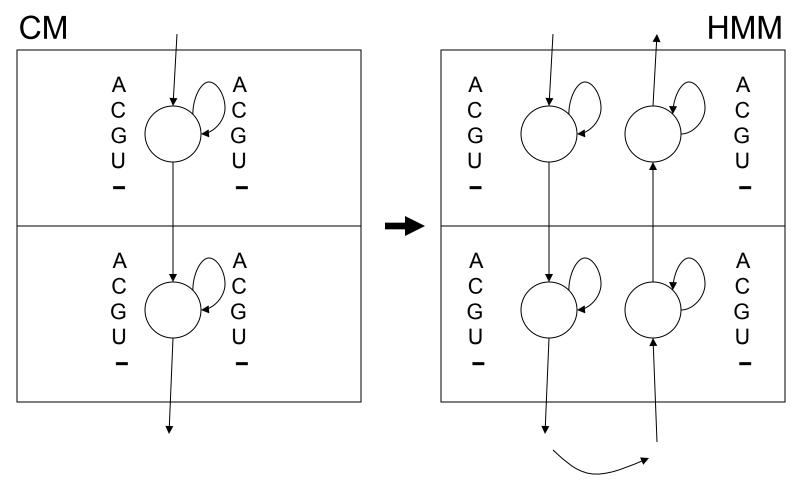
CM's are good, but slow



Oversimplified CM (for pedagogical purposes only)

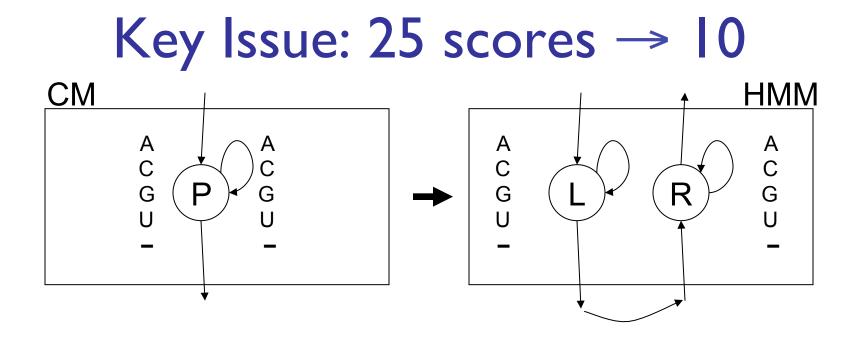


CM to HMM

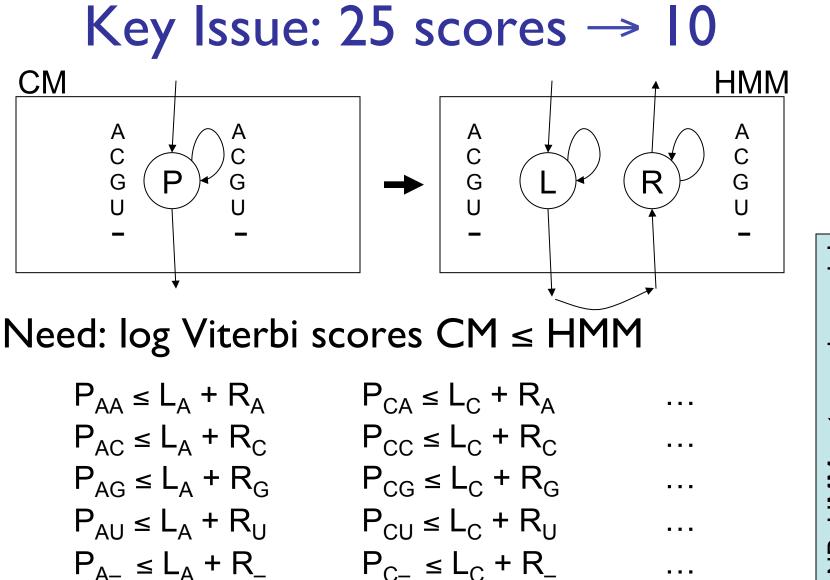


25 emisions per state

5 emissions per state, 2x states



Need: log Viterbi scores $CM \le HMM$



model NB: HMM not a prob. model

Rigorous Filtering

$$\begin{aligned} \mathsf{P}_{\mathsf{A}\mathsf{A}} &\leq \mathsf{L}_{\mathsf{A}} + \mathsf{R}_{\mathsf{A}} \\ \mathsf{P}_{\mathsf{A}\mathsf{C}} &\leq \mathsf{L}_{\mathsf{A}} + \mathsf{R}_{\mathsf{C}} \\ \mathsf{P}_{\mathsf{A}\mathsf{G}} &\leq \mathsf{L}_{\mathsf{A}} + \mathsf{R}_{\mathsf{G}} \\ \mathsf{P}_{\mathsf{A}\mathsf{U}} &\leq \mathsf{L}_{\mathsf{A}} + \mathsf{R}_{\mathsf{U}} \\ \mathsf{P}_{\mathsf{A}\mathsf{U}} &\leq \mathsf{L}_{\mathsf{A}} + \mathsf{R}_{\mathsf{U}} \end{aligned}$$

Any scores satisfying the linear inequalities give rigorous filtering

Proof:

CM Viterbi path score

- ≤ "corresponding" HMM path score
- ≤ Viterbi HMM path score

(even if it does not correspond to any CM path)

Minimizing $E(L_i, R_i)$ (subject to linear constraints)

Calculate $E(L_i, R_i)$ symbolically, in terms of emission scores, so we can do partial derivatives for numerical convex optimization algorithm

Forward:

 $f_k(i) = P(x_1 \dots x_i, \ \pi_i = k)$ $f_l(i+1) = e_l(x_{i+1}) \sum_k f_k(i) a_{k,l}$

Viterbi: $v_l(i+1) = e_l(x_{i+1}) \cdot \max_k(v_k(i) a_{k,l})$

$$\frac{\partial E(L_1, L_2, \ldots)}{\partial L_i}$$

Assignment of scores/ "probabilities"

Convex optimization problem

Constraints: enforce rigorous property

Objective function: filter as aggressively as possible

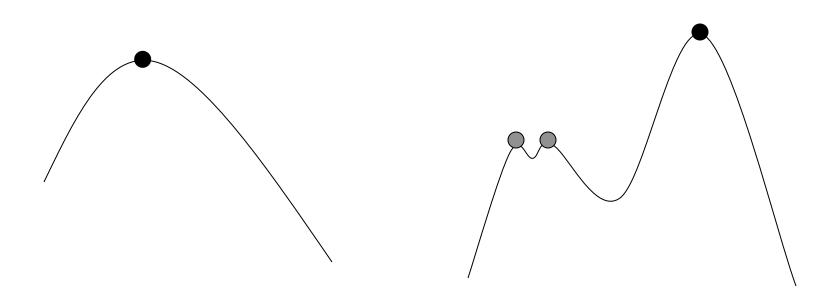
Problem sizes:

1000-10000 variables

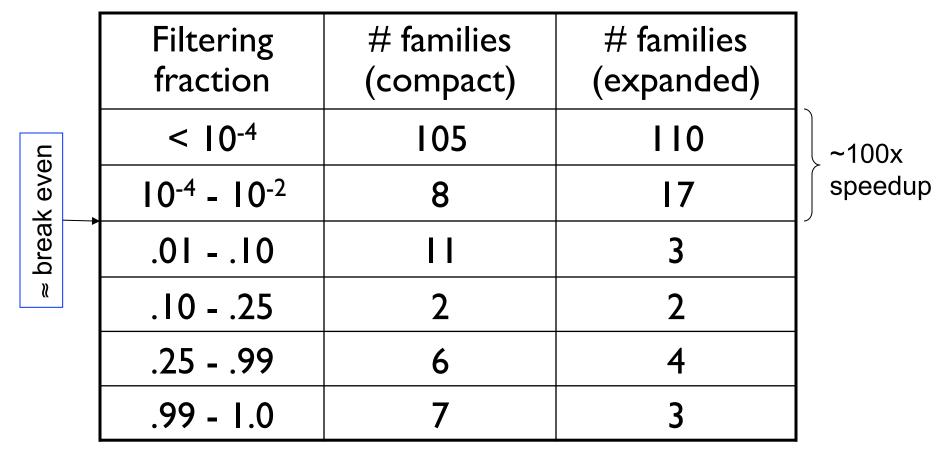
10000-100000 inequality constraints

"Convex" Optimization

Convex: local max = global max; simple "hill climbing" works Nonconvex: can be many local maxima, « global max; "hill-climbing" fails



Estimated Filtering Efficiency (139 Rfam 4.0 families)



Averages 283 times faster than CM

Motif Discovery

RNA Motif Discovery

Would be great if: given 100 complete genomes from diverse species, we could automatically find all the RNAs. State of the art: that's hopeless

Hope: can we exploit biological knowledge to narrow the search space?

RNA Motif Discovery

More promising problem: given a 10-20 unaligned sequences of a few kb, most of which contain instances of one RNA motif of 100-200bp -- find it.

Example: 5' UTRs of orthologous glycine cleavage genes from γ-proteobacteria

Example: corresponding introns of orthogolous vertebrate genes

Orthologs = counterparts in different species

Approaches

Align-First: Align sequences, then look for common structure

Fold-First: Predict structures, then try to align them

Joint: Do both together

Pitfall for sequence-alignmentfirst approach

Structural conservation ≠ Sequence conservation

Alignment without structure information is unreliable

CLUSTALW alignment of SECIS elements with flanking regions

same-colored boxes should be aligned

Approaches

Align-first: align sequences, then look for common structure

Fold-first: Predict structures, then try to align them

single-seq struct prediction only ~ 60% accurate; exacerbated by flanking seq; no biologicallyvalidated model for structural alignment

Joint: Do both together

Sankoff – good but slow

Heuristic

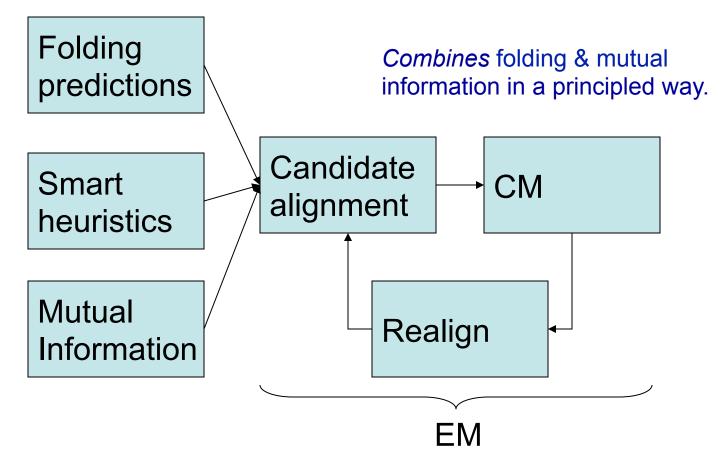
Our Approach: CMfinder

Simultaneous *local* alignment, folding and CMbased motif description using an EM-style learning procedure

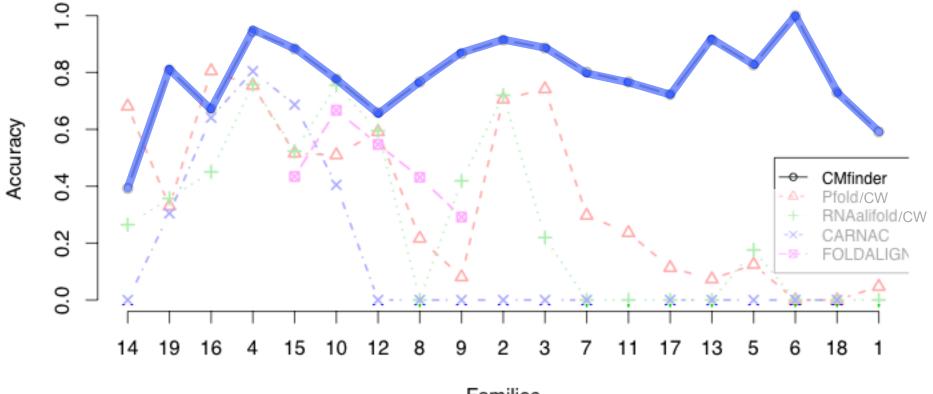
Yao, Weinberg & Ruzzo, Bioinformatics, 2006

CMFinder

Simultaneous alignment, folding & motif description Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006



CMfinder Accuracy (on Rfam families *with* flanking sequence)



Families

Discovery in Bacteria

OPEN O ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

A Computational Pipeline for High-Throughput Discovery of *cis*-Regulatory Noncoding RNA in Prokaryotes

Zizhen Yao^{1*}, Jeffrey Barrick^{2¤}, Zasha Weinberg³, Shane Neph^{1,4}, Ronald Breaker^{2,3,5}, Martin Tompa^{1,4}, Walter L. Ruzzo^{1,4}

Published online 9 July 2007

Nucleic Acids Research, 2007, Vol. 35, No. 14 4809-4819 doi:10.1093/nar/gkm487

Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline

Zasha Weinberg^{1,*}, Jeffrey E. Barrick^{2,3}, Zizhen Yao⁴, Adam Roth², Jane N. Kim¹, Jeremy Gore¹, Joy Xin Wang^{1,2}, Elaine R. Lee¹, Kirsten F. Block¹, Narasimhan Sudarsan¹, Shane Neph⁵, Martin Tompa^{4,5}, Walter L. Ruzzo^{4,5} and Ronald R. Breaker^{1,2,3} 156

Right Data: Why/How

We can recognize, say, 5-10 good examples amidst 20 extraneous ones (but not 5 in 200 or 2000) of length 1k or 10k (but not 100k)

Regulators often near regulatees (protein coding genes), which are usually recognizable cross-species So, look near similar genes ("homologs")

Many riboswitches, e.g., are present in ~5 copies per genome

(Not strategy used in vertebrates - 1000x larger genomes)

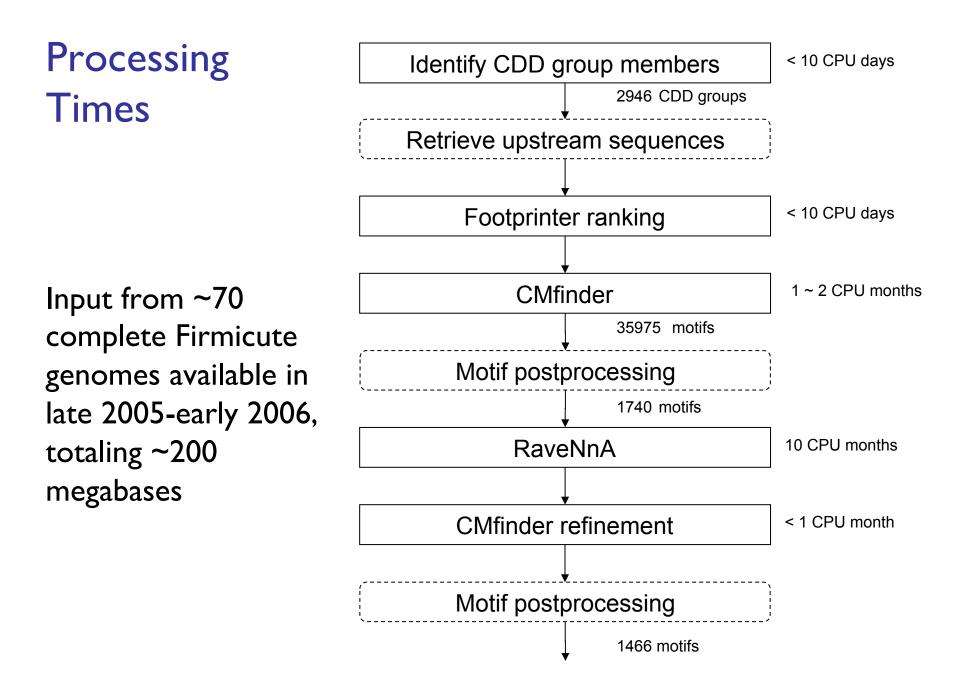


Table I: Motifs that correspond to Rfam families

	Rank S		Score	core #				Rfam	
RAV	CMF	FP		RAV	CMF	ID	Gene	Descriptio n	
0	43	107	3400	367	11	9904	llvΒ	Thiamine pyrophosphate-requiring enzymes	RF00230 T-box
1	10	344	3115	96	22	13174	COG3859	Predicted membrane protein	RF00059 THI
2	77	1284	2376	112	6	11125	MetH	Methionine synthase I specific DNA methylase	RF00162 S_box
3	0	5	2327	30	26	9991	COG0116	Predicted N6-adenine-specific DNA methylase	RF00011 RNaseP bact b
4	6	66	2228	49	18	4383	DHBP	3,4-dihydroxy-2-butanone 4-phosphate synthase	RF00050 RFN
7	145	952	1429	51	7	10390	GuaA	GMP synthase	RF00167 Purine
8	17	108	1322	29	13	10732	GcvP	Glycine cleavage system protein P	RF00504 Glycine
9	37	749	1235	28	7	24631	DUF149	Uncharacterised BCR, YbaB family COG0718	RF00169 SRP_bact
10	123	1358	1222	36	6	10986	CbiB	Cobalamin biosynthesis protein CobD/CbiB	RF00174 Cobalamin
20	137	1133	899	32	7	9895	LysA	Diaminopimelate decarboxylase	RF00168 Lysine
21	36	141	896	22	10	10727	TerC	Membrane protein TerC	RF00080 yybP-ykoY
39	202	684	664	25	5	11945	MgtE	Mg/Co/Ni transporter MgtE	RF00380 ykoK
40	26	74	645	19	18	10323	GlmS	Glucosamine 6-phosphate synthetase	RF00234 gImS
53	208	192	561	21	5	10892	OpuBB	ABC-type proline/glycine betaine transport systems	RF00005 tRNA ¹
122	99	239	413	10	7	11784	EmrE	Membrane transporters of cations and cationic drug	RF00442 ykkC-yxkD
255	392	281	268	8	6	10272	COG0398	Uncharacterized conserved protein	RF00023 tmRNA

Table 1: Motifs that correspond to Rfam families. "Rank": the three columns show ranks for refined motif clusters after genome scans ("RAV"), CMfinder motifs before genome scans ("CMF"), and FootPrinter results ("FP"). We used the same ranking scheme for RAV and CMF. "Score"

Rfam		Μ	embershi	р	Overlap			Structure		
		#	Sn	Sp	nt	Sn	Sp	bp	Sn	Sp
RF00174	Cobalamin	183	0.74 ¹	0.97	152	0.75	0.85	20	0.60	0.77
RF00504	Glycine	92	0.56 ¹	0.96	94	0.94	0.68	17	0.84	0.82
RF00234	glmS	34	0.92	1.00	100	0.54	1.00	27	0.96	0.97
RF00168	Lysine	80	0.82	0.98	111	0.61	0.68	26	0.76	0.87
RF00167	Purine	86	0.86	0.93	83	0.83	0.55	17	0.90	0.95
RF00050	RFN	133	0.98	0.99	139	0.96	1.00	12	0.66	0.65
RF00011	RNaseP_bact_b	144	0.99	0.99	194	0.53	1.00	38	0.72	0.78
RF00162	S_box	208	0.95	0.97	110	1.00	0.69	23	0.91	0.78
RF00169	SRP_bact	177	0.92	0.95	99	1.00	0.65	25	0.89	0.81
RF00230	T-box	453	0.96	0.61	187	0.77	1.00	5	0.32	0.38
RF00059	THI	326	0.89	1.00	99	0.91	0.69	13	0.56	0.74
RF00442	ykkC-yxkD	19	0.90	0.53	99	0.94	0.81	18	0.94	0.68
RF00380	ykoK	49	0.92	1.00	125	0.75	1.00	27	0.80	0.95
RF00080	yybP-ykoY	41	0.32	0.89	100	0.78	0.90	18	0.63	0.66
mean		145	0.84	0.91	121	0.81	0.82	21	0.75	0.77
median		113	0.91	0.97	105	0.81	0.83	19	0.78	0.78

Tbl 2: Prediction accuracy compared to prokaryotic subset of Rfam full alignments.

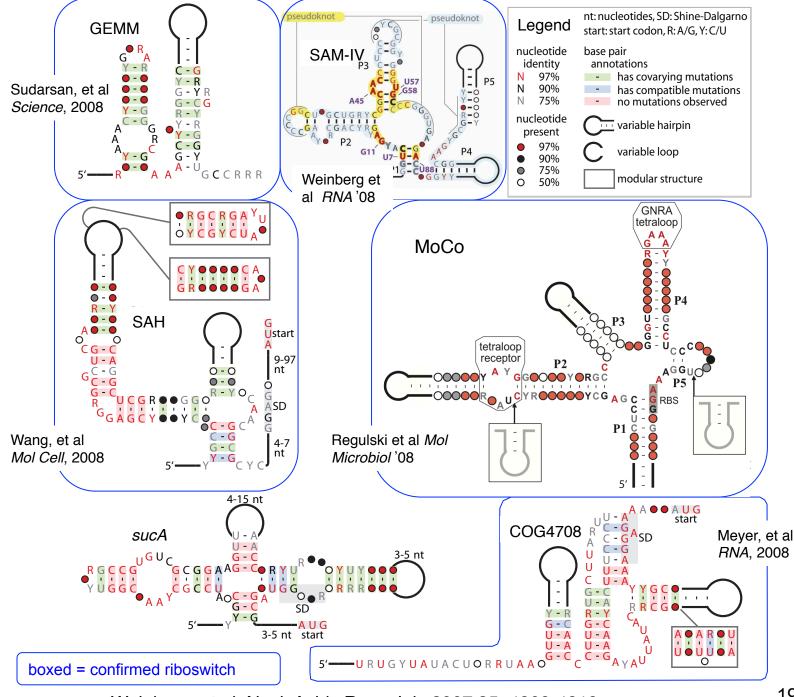
Membership: # of seqs in overlap between our predictions and Rfam's, the sensitivity (Sn) and specificity (Sp) of our membership predictions. Overlap: the avg len of overlap between our predictions and Rfam's (nt), the fractional lengths of the overlapped region in Rfam's predictions (Sn) and in ours (Sp). Structure: the avg # of correctly predicted canonical base pairs (in overlapped regions) in the secondary structure (bp), and sensitivity and specificity of our predictions. ¹After 2nd RaveNnA scan, membership Sn of Glycine, Cobalamin increased to 76% and 98% resp., Glycine Sp unchanged, but Cobalamin Sp dropped to 84%. 184

A L19 (*rplS*) mRNA leader

· •	P1
	TSS
-35	-10 → P2 RBS Start
Bsu TTGCAT.17.	TAAGAT. 40. AAAACGAUGUUCCGCUGUGCCGGUUUUUGUGGC.CAAGAGCAUCUG.05.AGGAGU.08.AUG
	TCTTCT. 17.AUUACGAUGUUCCGCUG.CAGGGGUAGAAGCUGUCAUGAGCAUCUG.06.AGGAGG.11.AUG
	TATATT. 31. UAAACGAUGUUCCGCUG. UC CCAUACUU GUUCAUGAGCAUUAG. 06. AGGAGU. 07. AUG
	TATGCT. 36. UUAACGAUGUUCCGCUG. UAA. UUUAUUAAGACU UUA. UAAGAGCAUCUG. 05. AGGAGA. 09. AUG
	TATCAT. 38. AAAACGAUGUUCCGC <mark>UG. CAAUGA</mark> . AGAGA UCAUUGGCAUGAACAUCUG. 04. AGGAGU. 08. AUG
	TATGAT. 45.AUUACGAUAUUCCGCUG.CUG.C.CAGUGU.C.CAGUGU.C.CAUGAAUGUCUG.06.AGGAGG.10.AUG
	GATAGT. 35. AUAACGAUGUUCCGCUG. CA. AUAAAGAAAGUCUG UG. CAAGAGCAUCUG. 05. AGGAGU. 08. AUG
	TAACCT. 28. AUAACGAUGUUCCGCUU. CAUUAUUAAUAUG.AAUGAUGUUUG. 05. AGGAGU. 08. AUG
	TAACAT. 23. AUCACUAUCAUCCGCUU. CAUAUAUAUUUGUCGAGGCAAGAACAUAGG.04. AGAGGA.09. AUG
	TAAACT. 08.GUACCGGCCGCUCCUCUCUCUCUCACAGAGUGUGUUAAGAACAUAGG.04.AGAGGA.09.AUG
	TATAAT. 09.UACCAAACGUUCCGC <mark>UG.GA</mark> CAGGGGCUC.CAUGAACGUCAA.17.ACGAGC.03.AGGAGG.09.AUG
	TAAAAT. 16. AAAAAGGUCGUCGCUG. CAUU AAACUAA AAUG. UAUGAACACCUU. 05. AGGAGG. 07. AUG
	TATAAT. 10. UUACGGGCCGUCCUCUA. UACAGGAGUA. UAAGAACGUCUA. 07. AGGAGG. 07. AUG
	TATAAT. 16. UUACCGACGCUCCGCUG. CCUCUGGGAAAGG.UAAGAACGUCUA.04. AGGAAG. 12. GUG
-	TAAACT. 28. AUACAGUUUAUCCGCUG. AGGAAGAUUCCU. CAAGAUUGACAA.04. AGGAGA.05. AUG
	TACAAT. 26. AAACG <mark>GCUAAUC</mark> CGC <mark>UG. AG</mark> ACAGAGCACU. UAUGAUUAGUAA.04. AGGAGA.07. AUG
_	TATTCT.21.UUAACGAUGUUCCGCUG.ACCAGGUUGU.CACGAAUGUCGG.04.AGGAAG.09.AUG
	TAAACT. 28. AUUACAAUAUUCCGCUG. UGG. CAGAAGUGACCA. UAAGAAUAUUUG. 06. AGGAGA. 08. AUG
	TAAACT. 25. UUAUGGGUAUUCCGCUG. GCACAAGGUGUUGAUGAAUGCCGU. 03. AGGAGA. 07. AUG
	TAAGAT . 29 . UAACG <mark>GCUAAUC</mark> CGC <mark>UG . AGA . CA</mark> CAGAGGU <mark>UG</mark> C <mark>UCU . UAA</mark> GAUUAGUAA . 03 . <mark>AGGAGU</mark> . 08 . AUG
	TTACTT.39.UUAUGGGUAUUCCGCUG.ACGCUGGUACGUUGAUGAAUGCCGA.03.AGGAGA.10.AUG
Spy TTTACA.17.	TAGAAT . 29 . UUACG <mark>GCUAAUC</mark> CGC <mark>UA . AG</mark> ACAAGUA <mark>. CU</mark> . <mark>UA</mark> AGAUUAGUAA . 03 . <mark>AGCAGA</mark> . 06 . <mark>AUG</mark>
Spy TTTACA.17. Lsa TTTTAA.17.	TAGAAT.29.UUACG <mark>GCUAAUC</mark> CGC <mark>UA.AG</mark> ACAAGUA <mark>CU.UA</mark> AGAUUAGUAA.03. <mark>AGGAGA</mark> .06.AUG TAAAAT.26.ACAAC <mark>GAUAUUC</mark> CGC <mark>UG.GCG</mark> CAAGA <mark>CGU</mark> UAA <mark>UGAAUAUC</mark> UG.06. <mark>AGGAGA</mark> .07.AUG
Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17.	TAGAAT.29.UUACG <mark>GCUAAUC</mark> CGC <mark>UA.AG</mark> ACAAGUA <mark>CU.UA</mark> AGAUUAGUAA.03. <mark>AGGAGA</mark> .06.AUG TAAAAT.26.ACAACGAUAUUCCGCUG.GCGCAAGA <mark>CGU</mark> UAAUGAAUAUCUG.06. <mark>AGCAGA</mark> .07.AUG TATTTT.24.AUAACGAUAUUCCGC <mark>UG.C</mark> AACUG <mark>GACA</mark> UGAAUGUCGG.04. <mark>AGGAAA</mark> .07.AUG
Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17.	TAGAAT.29.UUACG <mark>GCUAAUC</mark> CGC <mark>UA.AG</mark> ACAAGUA <mark>CU.UA</mark> AGAUUAGUAA.03. <mark>AGGAGA</mark> .06.AUG TAAAAT.26.ACAAC <mark>GAUAUUC</mark> CGC <mark>UG.GCG</mark> CAAGA <mark>CGU</mark> UAA <mark>UGAAUAUC</mark> UG.06. <mark>AGGAGA</mark> .07.AUG
Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17.	TAGAAT.29.UUACG <mark>GCUAAUC</mark> CGC <mark>UA.AG</mark> ACAAGUA <mark>CU.UA</mark> AGAUUAGUAA.03. <mark>AGGAGA</mark> .06.AUG TAAAAT.26.ACAACGAUAUUCCGCUG.GCGCAAGA <mark>CGU</mark> UAAUGAAUAUCUG.06. <mark>AGCAGA</mark> .07.AUG TATTTT.24.AUAACGAUAUUCCGC <mark>UG.C</mark> AACUG <mark>GACA</mark> UGAAUGUCGG.04. <mark>AGGAAA</mark> .07.AUG
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17.	TAGAAT.29.UUACCGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AUG TAAAAT.26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AUG TATTTT.24.AUAACGAUAUUCCGCUG.CAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AUG TAAAAT.12.AAUUCGAUAUUCCGCUU.UAAUAAAUUAA.AAUGAAUAUCUU.04.AGGAAG.02.AUG
Spy TTTACA.17. Lsa TTTTAA.17. Lsl TTTACT.17.	TAGAAT. 29. UUACCGCUAAUCCGCUA. AG ACAAGUACU. UAAGAUUAGUAA. 03. AGGAGA. 06. AUG TAAAAT. 26. ACAACGAUAUUCCGCUG. GCGCAAGACGUUAAUGAAUAUCUG. 06. AGGAGA. 07. AUG TATTTT. 24. AUAACGAUAUUCCGCUG. CAACUGGACAUGAAUGUCGG. 04. AGGAAA. 07. AUG TAAAAT. 12. AAUUCGAUAUUCCGCUU. UAAUAAAUUA. AAUGAAUAUCUU. 04. AGGAAG. 02. AUG B. subtilis L19 mRNA leader
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17.	TAGAAT 29.UUACCGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AUG TAAAAT 26.ACAACGAUAUUCCGCUG.GCGCAAGAGUUAAUGAAUAUCUG.06.AGGAGA.07.AUG TATTTT 24.AUAACGAUAUUCCGCUG.CAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AUG TAAAAT 12.AAUUCGAUAUUCCGCUG.CAACUGGACAUGAAUGUCGG.04.AGGAAA.07.AUG TAAAAT 12.AAUUCGAUAUUCCGCUU.UAAUAAAUUA.AAUGAAUAUCUU.04.AGGAAG.07.AUG TAAAAT 12.AAUUCGAUAUUCCGCUU.UAAUAAAUUA.AAUGAAUAUCUU.04.AGGAAG.02.AUG nucleotide nucleotide
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17.	TAGAAT. 29. UUACCGCUAAUCCGCUA. AG ACAAGUA
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17.	TAGAAT 29.UUACCGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGGAGA.06.AUG TAAAAT 26.ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AUG TAATTTT 24.AUAACGAUAUUCCGCUG.CAACUGGACAUGCGAUGUCGG.04.AGGAAA.07.AUG TAAAAT 12.AAUUCGAUAUUCCGCUG.CAACUGGACAUGUCGG.04.AGGAAA.07.AUG TAAAAT 12.AAUUCGAUAUUCCGCUU.UAAUAAAUUAAUUA.AAUGAAUAUCUU.04.AGGAAG.02.AUG nucleotide nucleotide identity present UUU ACG UUU UUU QUUU UUU ACG UUU
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17.	TAGAAT 29. UUACCGCUAAUCCGCUA.AGACAAGUACU.UAAGAUUAGUAA.03.AGCAGA.06.AUG TAAAAT 26. ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAAUAUCUG.06.AGGAGA.07.AUG TATTTT 24. AUAACGAUAUUCCGCUG.CAACUGGACAUGCGAUGUCGG.04.AGGAAA.07.AUG TAAAAT 12. AAUUCGAUAUUCCGCUU.UAAUAAAUUA.AAUGAAUAUCUU.04.AGGAAG.02.AUG nucleotide nucleotide identity present N 97%
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17.	TAGAAT. 29. UUACCGCUAAUCCGCUA. AG ACAAGUACU. UAAGAUUAGUAA.03. AGGAGA.06. AUGTAAAAT. 26. ACAACGAUAUUCCGCUG. GCGCAAGACGUUAAUGAAUAUCUG.06. AGGAGA.07. AUGTATTTT. 24. AUAACGAUAUUCCGCUG. CAACUGGACAUGAAUGUCGG.04. AGGAAA.07. AUGTAAAAT. 12. AAUUCGAUAUUCCGCUU. UAAUAAAUUA. AAUGAAUAUCUU.04. AGGAAG.02. AUGnucleotideidentitypresentN 97%97%N 90%00%
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B	TAGAAT 29. UUACCGCUAAUCCGCUA.AG AGACAAGUACU.UAAGAUUAGUAA.03.AGCAGA.06.AUG TAAAAT 26. ACAACGAUAUUCCGCUG.GCGCAAGACGUUAAUGAUAUCUG.06.AGGAGA.07.AUG TATTTT 24. AUAACGAUAUUCCGCUG.CAACUGGACAUGCGAUGUCGG.04.AGGAAA.07.AUG TAAAAT 12. AAUUCGAUAUUCCGCUU.UAAUAAAUUAAUUA.AAUGAAUAUCUU.04.AGGAAG.02.AUG nucleotide nucleotide identity present N 97% N 97% N 97% N 90%
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17.	TAGAAT . 29 . UUACGCUAAUC CGCUA . AG ACAAGUA CU . UAA GAUUAGUAA . 03 . AGGAGA . 06 . AUG TAAAAT . 26 . ACAACGAUAUUC CGCUG . GCG CAAGA CGUUAAU GAAUAUC UG . 06 . AGGAGA . 07 . AUG TAATTTT . 24 . AUAACGAUAUUC CGCUG . C AACUG GACAU GAAUGUC GG . 04 . AGGAAA . 07 . AUG TAAAAT . 12 . AAUUCGAUAUUC CGCUU . UAA UAAA UUA . AAU GAAUAUC UU . 04 . AGGAAG . 02 . AUG nucleotide nucleotide present N 97% 97% N 90% 90% N 75% 975% 975% C - G G U A G G G G U C - G C U L19 ?
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B	TAGAAT 29. UUACCGCUAAUC CGCUA AG. AG. ACAAGUA C. CU UAAGAUUAGUAA 03. AGGAGA 06. AUG TAAAAT 26. ACAAC GAUAUUC CGCUG GCG C. CAAGA C. CAUGAAUGAAUGUCGAUAUCUG 06. AGGAGA 07. AUG TATTTT 24. AUAAC GAUAUUC CGCUG C. AACUG C. GACAUGAAUGUCGG 04. AGGAAA 07. AUG TAAAAT 12. AAUUC GAUAUUC CGCUU UAA C. UAAA UUA AAUGAAUAUCUU 04. AGGAAG 02. AUG N 97% 97% N 90% 97% N 90% 90% N 75% 75% O 50% $U^U U A^CG U^U U$ $G^G G A^U C^G G C^G C^G C^G C^G C^G C^G C^G C^G C$
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B	TAGAAT 29. UUACG GCUAAUC CGCUA AG. AG. ACAAGUA C. CU. UAAGAUUAGUAA 03. AGGAGA 06. AUG TAAAAT 26. ACAAC GAUAUUC CGCUG CGC . CAAGA CAGAGA CUUAAUGAAUAUCUG 06. AGGAGA 07. AUG TATTTT 24. AUAAC GAUAUUC CGCUG C. AACUG CGCUG . CAAGA . CGUUAAUGAAUGUCGG 04. AGGAAG . 07. AUG TAAAAT 12. AAUUC GAUAUUC CGCUU . UAA
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B	TAGAAT 29. UUACGCUAAUC CGCUA AC ACAAGUA CU .UAAGAUUACUAA 03. AGGACA 06. AUG TAAAAT 26. ACAACGAUAUUC CGCUG GCG CAAGA CGUUAAUGAAUAUC UG 06. AGGACA 07. AUG TATTTT 24. AUAACGAUAUUC CGCUG C AACUG GACAUGAAUGUC GG 04. AGGAAA 07. AUG TAAAAT 12. AAUUCGAUAUUC CGCUU .UAA UAAA UUA .AAUGAAUGUC GG 04. AGGAAC 02. AUG nucleotide nucleotide present UUU A $AUGAAUAUCUU .04$. AGGAAC 02. AUG N 97% 97% 97% O 97% O 97% O 97% O 97% O 90% O 90
Spy TTTACA. 17. Lsa TTTTAA. 17. Ls1 TTTACT. 17. Fnu TTGACA. 17. B P2 $\begin{array}{c} & & \\ & &$	TACAAT 29. UUACGCUAAUCCGCUA AG. AG. ACAAGUA CU UAACAUUAGUAA.03. AGGAGA.06. AUG TAAAAT 26. ACAACGAUAUUCCGCUG GCG CAAGAA CGUUAAUGAAUAUCUG.06. AGGAGA.07. AUG TATTTT 24. AUAACGAUAUUCCGCUG C. AACUG. GACAUGAAUGUCGG.04. AGGAAA.07. AUG TAAAAT 12. AAUUCGAUAUUCCGCUU UAA CUAAA. UUA AAUGAAUAUCUU.04. AGGAAG.02. AUG TAAAAT 12. AAUUCGAUAUUCCGCUU UAA CUAAA. UUA AAUGAAUAUCUU.04. AGGAAG.02. AUG N 97% 97% N 90% 90% N 75% 75% O 50% C G G G G G G G G G G G G G G G G G G G
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B	TACAAT 29. UUACGCUUAAUCCGCUA AG
Spy TTTACA. 17. Lsa TTTTAA. 17. Ls1 TTTACT. 17. Fnu TTGACA. 17. B P2 $\begin{array}{c} & & \\ & &$	TACAAT 29. UUACCGCUAAUC CGCUA AG AG AG ACAAGUA CU UAAGAUUAGUAA 03. AGGAGA 06. AUG TAAAAT 26. ACAACGAUAUUC CGCUG GCG AGAGG CAAGAA CGUUAAUGAAUAUC UG 06. AGGAGA 07. AUG TAAAAT 24. AUAACGAUAUUC CGCUG C. AACUG GAAUGG GG 04. AGGAAA 07. AUG TAAAAT 12. AAUUC GAUAUUC CGCUU UAA AAUGAAA 07. AUG TAAAAT 12. AAUUC GAUAUUC CGCUU AA AAUGAAA 07. AUG TAAAAT 12. AAUUC GAUAUUC CGCUU AAA AAUGAAAA 07. AUG AAUG AAUGAAUAUC UU 04. AGGAAG 02. AUG GG GG G G G G G G G G G G G G G G G G
Spy TTTACA. 17. Lsa TTTTAA. 17. Ls1 TTTACT. 17. Fnu TTGACA. 17. B P2 $\begin{array}{c} & & \\ & &$	TACAAT 29. UUACCGCUAAUC CGCUA AG AG AG ACAAGUA CU UAAGAUUAGUAA 03. AGGAGA 06. AUG TAAAAT 26. ACAACGAUAUUC CGCUG GCG AGAGG CAAGAA CGUUAAUGAAUAUC UG 06. AGGAGA 07. AUG TAAAAT 24. AUAACGAUAUUC CGCUG C. AACUG GAAUGG GG 04. AGGAAA 07. AUG TAAAAT 12. AAUUC GAUAUUC CGCUU UAA AAUGAAA 07. AUG TAAAAT 12. AAUUC GAUAUUC CGCUU AA AAUGAAA 07. AUG TAAAAT 12. AAUUC GAUAUUC CGCUU AAA AAUGAAAA 07. AUG AAUG AAUGAAUAUC UU 04. AGGAAG 02. AUG GG GG G G G G G G G G G G G G G G G G
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 $\begin{array}{c} & & \\ & & $	TAGAAT 29. UUACG GCUAAUC CGCUA AG
Spy TTTACA. 17. Lsa TTTTAA. 17. Ls1 TTTACT. 17. Fnu TTGACA. 17. B P2 $\begin{array}{c} & & \\ & &$	TAGAAT 29. UUACGCCUAAUC CGCUA AG. AC. ACAAGUA CU .UAAGAUUAGUAA 03. AGGAGA 06. AUG TAAAAT 26. ACAACGAUAUUC CGCUG GCG CAAGA CGUUAAUGAAUAUCUG 06. AGGAGA 07. AUG TAAAAT 24. AUAACGAUAUUC CGCUG C AACUG GACAUGAAUGUCGG 04. AGGAAA 07. AUG TAAAAT 12. AAUUCGAUAUUC CGCUU .UAA UAAA UUAA UUA AAUGAUUUU 04. AGGAAA 02. AUG TAAAAT 12. AAUUCGAUAUUC CGCUU .UAA UAAA UUAA UUA AAUGAUUUU 04. AGGAAA 02. AUG TAAAAT 12. AAUUCGAUAUUC CGCUU .UAA UAAA UUAA UUA AAUGAUUUU 04. AGGAAC 02. AUG TAAAAT 12. AAUUCGAUAUUC CGCUU .UAA UAAA UUAA UUA AAUGAUUUU 04. AGGAAC 02. AUG N 97% 97% N 90% 90% N 75% 75% O 50% Stem loop always present C C G G G U G C U A C C G G G U G C G U U G C C G U U G U C U G C C G U A G C C G U G C G U U G U C U G C C G G G C U A G C C G G G G G G G G G G G G G G G G G
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B P2 $G = G = G = G = G = G = G = G = G = G $	TAGAAT . 29. UUACG GCUAAUC GCUA . AG . AG AGUA CU . UA GAUUAGU AA . 03. AGGAGA
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. P2 $\begin{array}{c} & & \\ & & $	TAGAAT 29. UUACGGCUAAUG TAAAAT 29. UUACGCUAAUG TAAAAT 26. ACAAGGAUAUUG CGCUG GCG CAAGA CUUAAGAUUAGUAA.03. AGGACA.06. AUG TAAAAT 26. ACAAGGAUAUUG CGCUG GCG CAAGA CGUUAAUGAAUGUG G.04. AGGAGA.07. AUG TAAAAT 24. AUAAGAUAUUG CGGUU .UAA UAAA UUA AAUGAAUGUG G.04. AGGAAA.07. AUG TAAAAT 12. AAUUG GAUAUUG CGGUU .UAA UAAA UUA AAUGAAUAUGUU .04. AGGAAG.02. AUG TAAAAT 12. AAUUG GAUAUUC CGGUU .UAA UAAA UUA AAUGAAUAUGUU .04. AGGAAG.02. AUG N 97% 97% N 90% 90% N 75% 75% O 50% C -
Spy TTTACA.17. Lsa TTTTAA.17. Ls1 TTTACT.17. Fnu TTGACA.17. B P2 $G = G = G = G = G = G = G = G = G = G $	TAGAAT . 29 . UUACG GCUAAUC GCUA . AG ACAAGUA CU . UAA GAUUAGU AA . 03 . AGGAGA

Example: Ribosomal Autoregulation: Excess L19 represses L19 (RF00556; 555-559 similar)

186



Weinberg, et al. Nucl. Acids Res., July 2007 35: 4809-4819.

Vertebrate ncRNAs

Some Results

Human Predictions

OUE

Evofold

RNAz S Washietl, IL Hofacker, M Luk sser, A Hutenhofer, PF S Pedersen, G Bejerano, A Siepel, K Stadler, "Mapping of conserved RNA secondary Rosenbloom, K Lindblad-Toh, ES Lander, J structures predict, thrusands of functional noncoding Kent, W Miller, D Haussler, "Identification RNAs in the hun ome." and classification of conserved RNA 2005) 1383-90. secondary structures in the human elements genome." conserved across all vertebrates. PLoS Comput. Biol., 2, #4 (2006) e33. in introns of known genes, ~1/6 in UTRs /2 located far from any known gene 48,479 candidates (~70% FDR?) FOLDALIGN CMfinder E Torarinsson, N Torarinsson, Yao, Wiklund, Bramsen, Hansen, Havgaard, M E rodkin, Kiems, Tommerup, Ruzzo and Gorodkin. "Thousands ndina Comparative genomics beyond sequence based genomic regions human and alignments: RNA structures in the ENCODE regions. primary sequence Genome Research, Feb 2008, 18(2):242-251 PMID: ommon RNA structure." 18096747 16, #7 (2006) 885-9. Res. 6500 candidates in ENCODE alone (better FDR, but candidates from 36970 (of still high) 00,000) pairs Some details below

CMfinder Search in Vertebrates

Extract ENCODE^{*} Multiz alignments

Remove exons, most conserved elements. 56017 blocks, 8.7M bps.

Apply CMfinder to both strands.

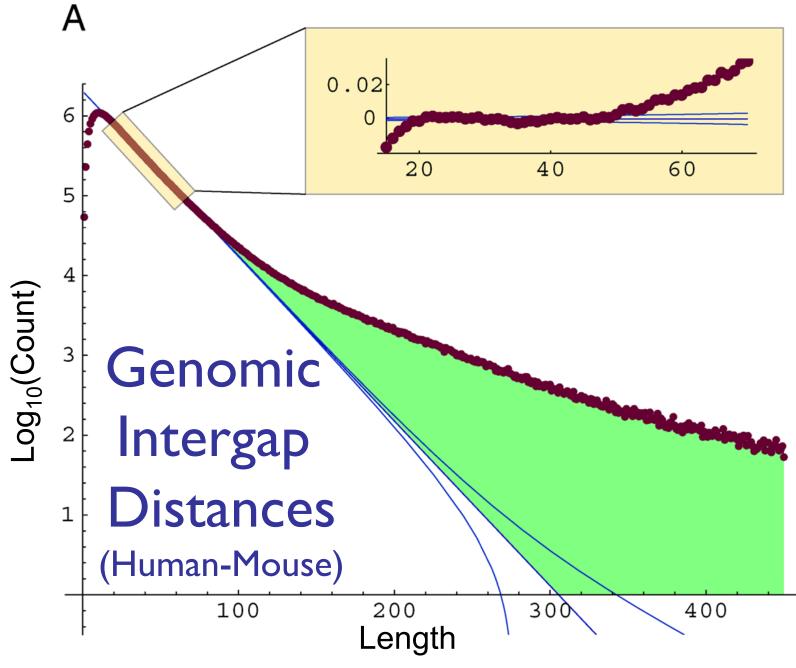
10,106 predictions, 6,587 clusters.

Trust 17-way alignment for orthology, not for detailed alignment

High false positive rate, but still suggests 1000's of RNAs.

(We've applied CMfinder to whole human genome: many 100's of CPU years. Analysis in progress.)

* ENCODE: deeply annotated 1% of human genome



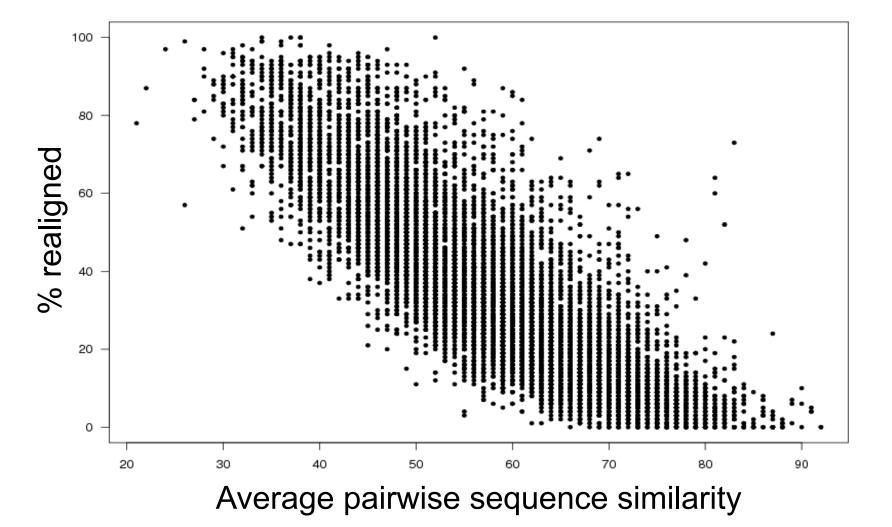
Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model Gerton Lunter, Chris P. Ponting, Jotun Hein, PLoS Comput Biol 2006, 2(1): e5.

Overlap w/ Indel Purified Segments

IPS presumed to signal purifying selection Majority (64%) of candidates have >45% G+C Strong P-value for their overlap w/ IPS

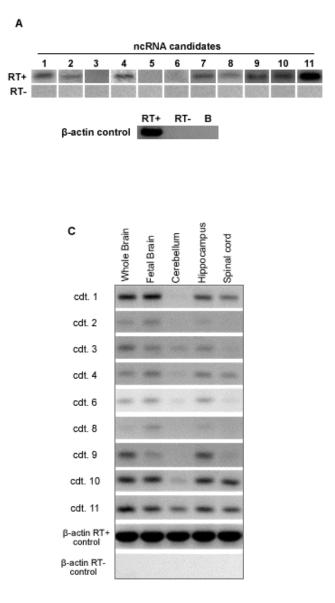
G+C	data	Р	Ν	Expected	Observed	P-value	%
0-35	igs	0.062	380	23	24.5	0.430	5.8%
35-40	igs	0.082	742	61	70.5	0.103	11.3%
40-45	igs	0.082	1216	99	129.5	0.00079	18.5%
45-50	igs	0.079	1377	109	162.5	5.16E-08	20.9%
50-100	igs	0.070	2866	200	358.5	2.70E-31	43.5%
all	igs	0.075	6581	491	747.5	1.54E-33	100.0%

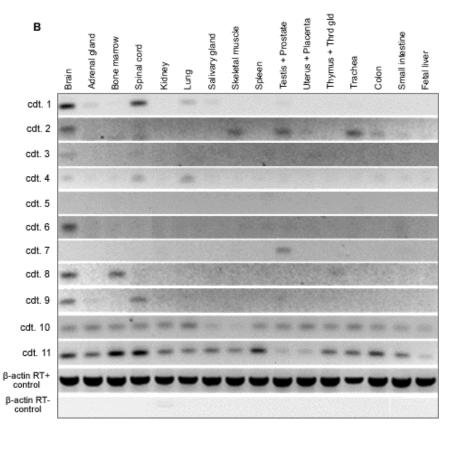
Realignment



Alignment Matters

10 of 11 top (differentially) expressed







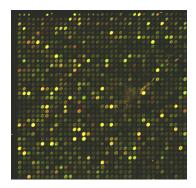
228

ncRNA Summary

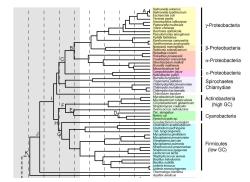
ncRNA is a "hot" topic For family homology modeling: CMs Training & search like HMM (but slower) Dramatic acceleration possible Automated model construction possible New computational methods yield new discoveries *Many open problems*

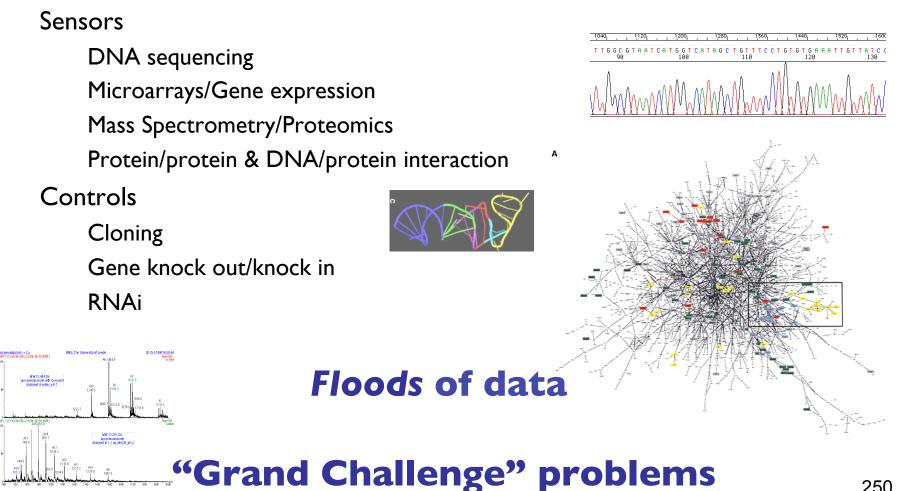
Course Wrap Up

```
What is DNA? RNA?
How many Amino Acids are there?
Did human beings, as we know them, develop
from earlier species of animals?
What are stem cells?
What did Viterbi invent?
What is dynamic programming?
What is a likelihood ratio test?
What is the EM algorithm?
How would you find the maximum of f(x) = ax^3 + b^3
bx^2 + cx + d in the interval -10<x<25?
```



"High-Throughput **BioTech**"





Exciting Times

Lots to do Highly multidisciplinary You'll be hearing a lot more about it I hope I've given you a taste of it

Thanks!