CMfinder - A covariance model based algorithm

To appear, *Bioinformatics*

Zizhen Yao

Zasha Weinberg

Walter L. Ruzzo

University of Washington, Seattle

12/8/05

Searching for noncoding RNAs

- CM's are great, but where do they come from?
- A comparative genomic approach
  - Search for motifs with common secondary structure in a set of functionally related sequences.
- Challenges
  - Three related tasks
    - Locate the motif regions.
    - Align the motif instances.
    - Predict the consensus secondary structure.
  - Motif Search space is huge!
    - Motif location space, alignment space, structure space.

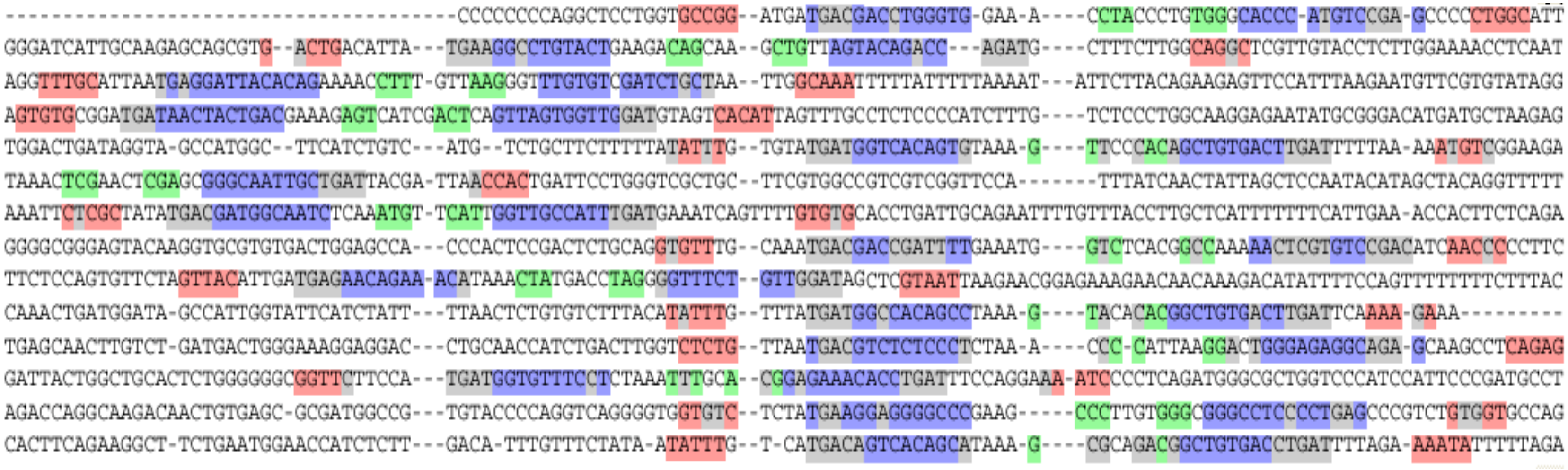
Approaches

- Align sequences, then look for common structure
- Predict structures, then try to align them
- Do both together

Pitfall for sequence alignment approach

- Structural conservation ≠ Sequence conservation
  - Alignment without structure information is unreliable

CLUSTALW alignment of SECIS elements with flanking regions



```
---------------------------------------CCCCCCCCAGGCTCCTGGTGCCGG--ATGATGACGACCTGGGTG-GAA-A----CCTACCCTGTGGGCACCC-ATGTCCGA-GCCCCCTGGCATT
GGGATCATTGCAAGAGCAGCGTG--ACTGACATTA---TGAAGGCCTGTACTGAAGACAGCAA--GCTGTTAGTACAGACC---AGATG----CTTTCTTGGCAGGCTCGTTGTACCTCTTGGAAAACCTCAAT
AGGTTTGCATTAATGAGGATTACACAGAAAACCTTT-GTTAAGGGTTTGTGTCGATCTGCTAA--TTGGCAAATTTTTATTTTTAAAAT---ATTCTTACAGAAGAGTTCCATTTAAGAATGTTCGTGTATAGG
AGTGTGCGGATGATAACTACTGACGAAAGAGTCATCGACTCAGTTAGTGGTTGGATGTAGTCACATTAGTTTGCCTCTCCCCATCTTTG----TCTCCCTGGCAAGGAGAATATGCGGGACATGATGCTAAGAG
TGGACTGATAGGTA-GCCATGGC--TTCATCTGTC---ATG--TCTGCTTCTTTTTATATTTG--TGTATGATGGTCACAGTGTAAA-G----TTCCCACAGCTGTGACTTGATTTTTAA-AAATGTCGGAAGA
TAAACTCGAACTCGAGCGGGCAATTGCTGATTACGA-TTAACCACTGATTCCTGGGTCGCTGC--TTCGTGGCCGTCGTCGGTTCCA-------TTTATCAACTATTAGCTCCAATACATAGCTACAGGTTTTT
AAATTCTCGCTATATGACGATGGCAATCTCAAATGT-TCATTGGTTGCCATTTGATGAAATCAGTTTTGTGTGCACCTGATTGCAGAATTTTGTTTACCTTGCTCATTTTTTTCATTGAA-ACCACTTCTCAGA
GGGGCGGGAGTACAAGGTGCGTGTGACTGGAGCCA---CCCACTCCGACTCTGCAGGTGTTTG--CAAATGACGACCGATTTTGAAATG----GTCTCACGGCCAAAAACTCGTGTCCGACATCAACCCCCTTC
TTCTCCAGTGTTCTAGTTACATTGATGAGAACAGAA-ACATAAACTATGACCTAGGGGTTTCT--GTTGGATAGCTCGTAATTAAGAACGGAGAAAGAACAACAAAGACATATTTTCCAGTTTTTTTCTTTAC
CAAACTGATGGATA-GCCATTGGTATTCATCTATT---TTAACTCTGTGTCTTTACATATTTG--TTTATGATGGCCACAGCCTAAA-G----TACACACGGCTGTGACTTGATTCAAAA-GAAA---------
TGAGCAACTTGTCT-GATGACTGGGAAAGGAGGAC---CTGCAACCATCTGACTTGGTCTCTG--TTAATGACGTCTCTCCCTCTAA-A----CCC-CATTAAGGACTGGGAGAGGCAGA-GCAAGCCTCAGAG
GATTACTGGCTGCACTCTGGGGGGCGGTTCTTCCA---TGATGGTGTTTCCTCTAAATTTGCA--CGGAGAAACACCTGATTTCCAGGAAA-ATCCCCTCAGATGGGCGCTGGTCCCATCCATTCCCGATGCCT
AGACCAGGCAAGACAACTGTGAGC-GCGATGGCCG---TGTACCCCAGGTCAGGGGTGGTGTC--TCTATGAAGGAGGGGCCCGAAG----CCCTTGTGGGCGGGCCTCCCCTGAGCCCGTCTGTGGTGCCAG
CACTTCAGAAGGCT-TCTGAATGGAACCATCTCTT---GACA-TTTGTTTCTATA-ATATTTG--T-CATGACAGTCACAGCATAAA-G----CGCAGACGGCTGTGACCTGATTTTAGA-AAATATTTTTAGA
```

Approaches

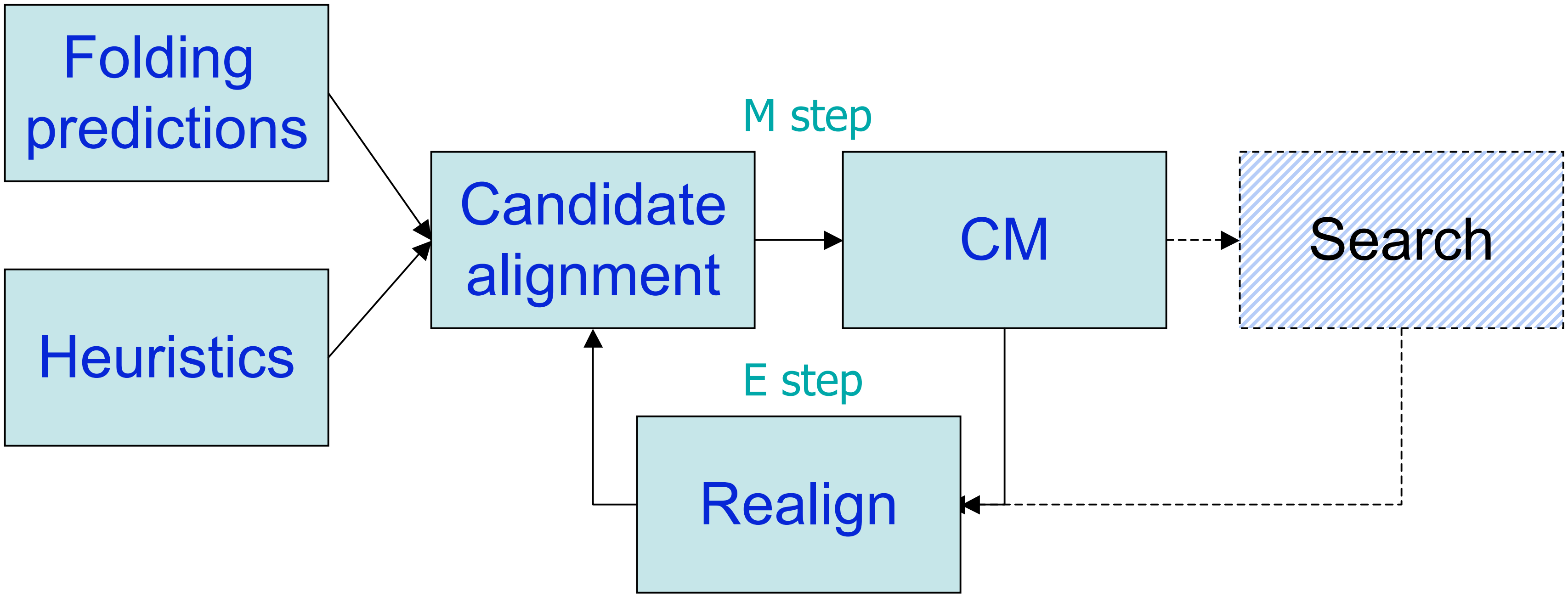
- Align sequences, then look for common structure
- Predict structures, then try to align them
  - single-seq struct prediction only ~ 60% accurate; exacerbated by flanking seq; no biologically-validated model for structural alignment
- Do both together
  - Sankoff – good but slow
  - Heuristic

Design goal

Search for RNA motifs in unaligned Sequences.

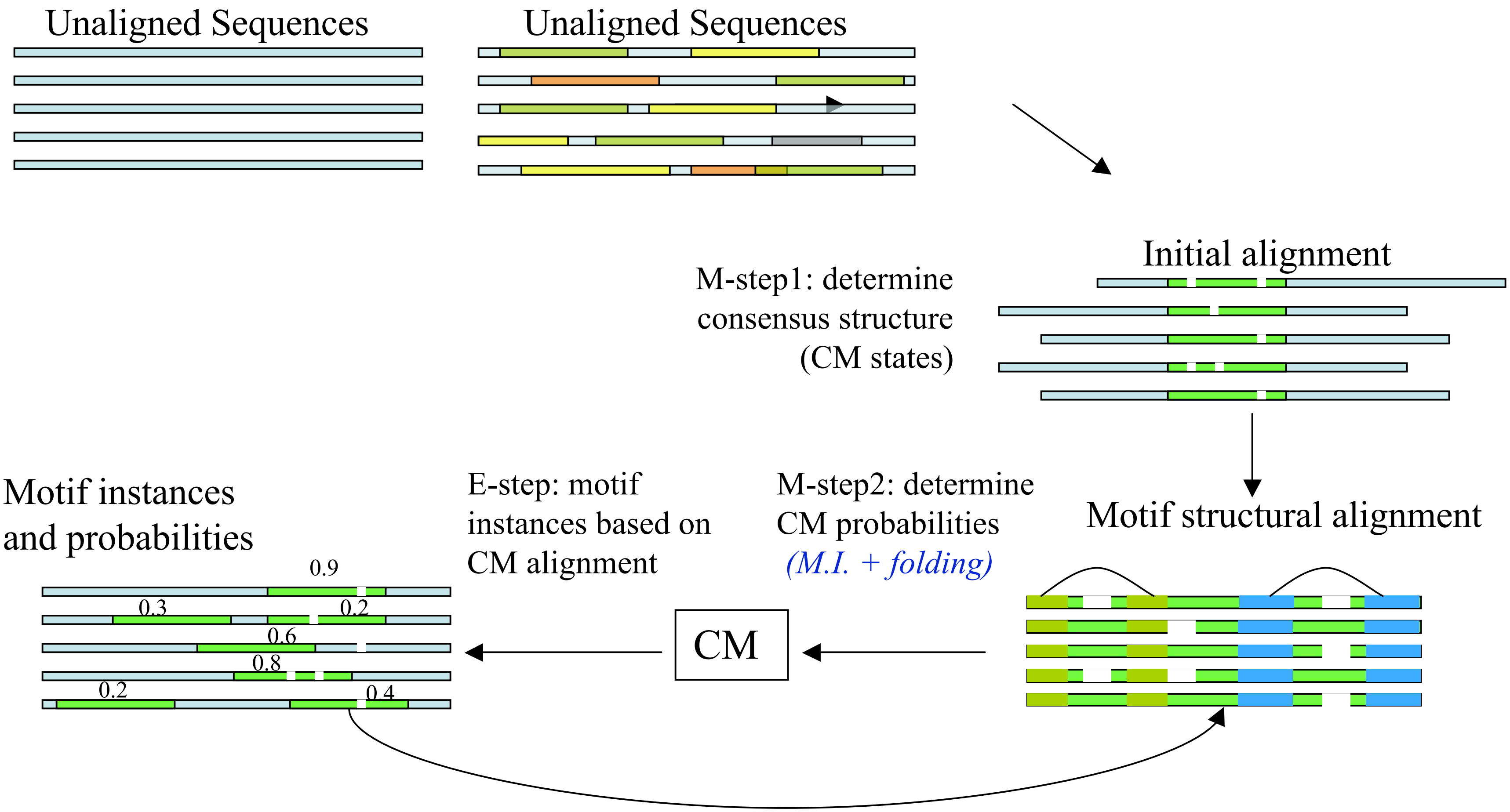
- Perform Local alignment
- Exploit but do not require sequence conservation
- Robust to inclusion of unrelated sequences.
- Reasonably fast and scalable.
- Produce a probabilistic model of the motif that can be directly used for homolog search.

CMfinder Outline



12/8/05

CMfinder at work



Unaligned Sequences

Unaligned Sequences

Initial alignment

M-step1: determine consensus structure (CM states)

Motif structural alignment

M-step2: determine CM probabilities *(M.I. + folding)*

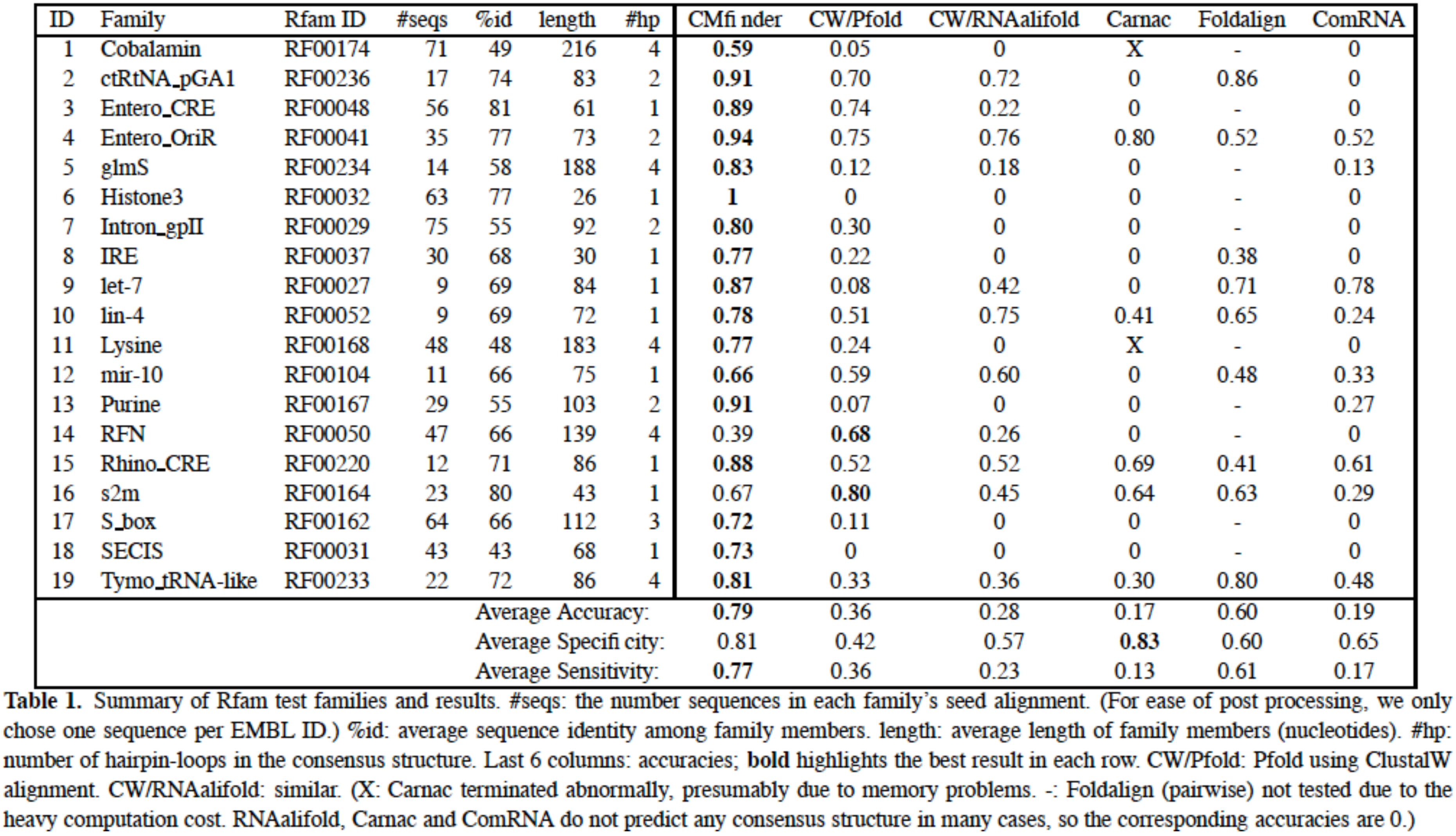
CM

E-step: motif instances based on CM alignment

Motif instances and probabilities

Performance on unaligned sequences

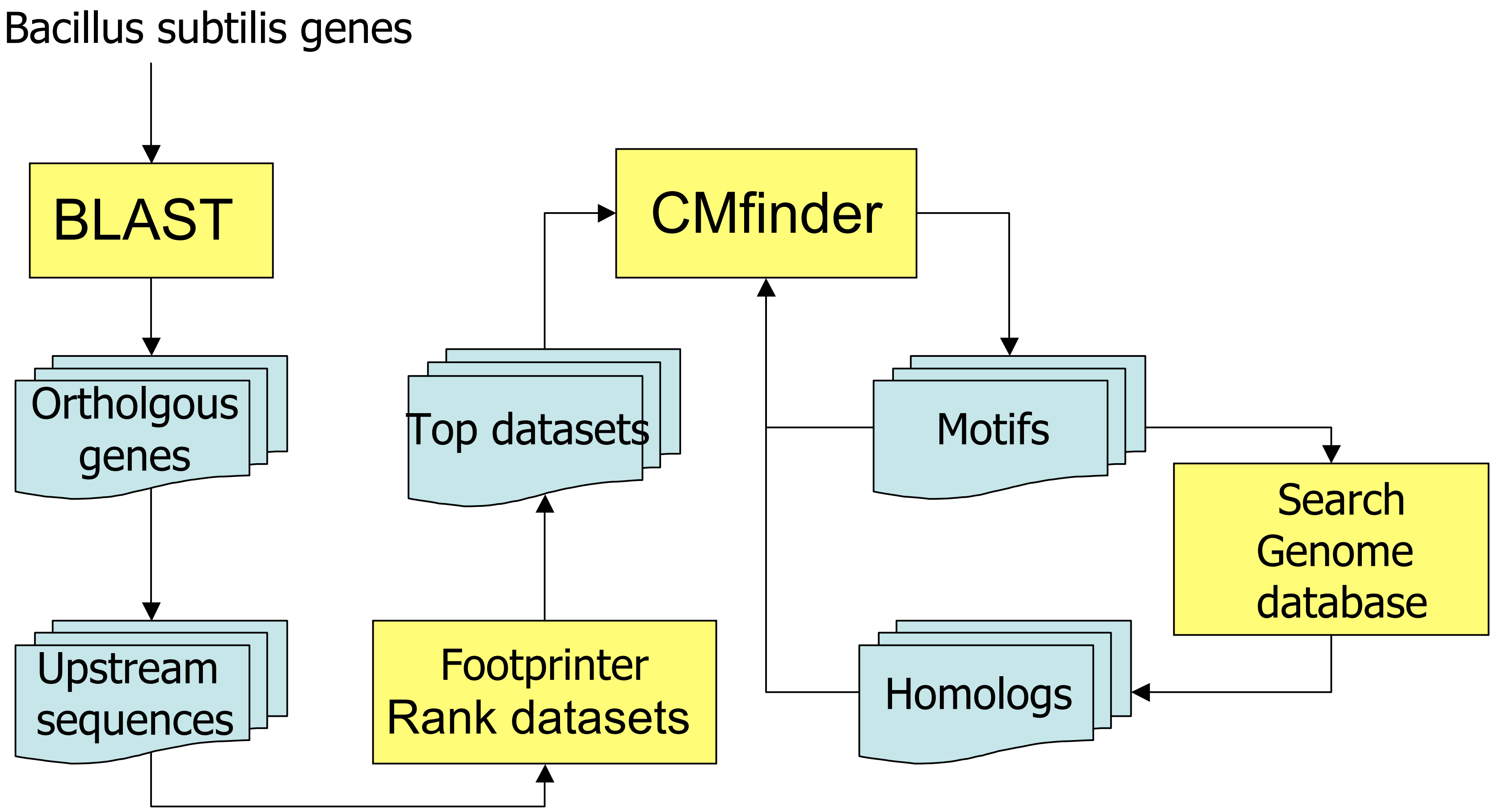
Including 200 base flanking region, distributed randomly between 3’ and 5’ side



ID	Family	Rfam ID	#seqs	%id	length	#hp	CMfinder	CW/Pfold	CW/RNAalifold	Carnac	Foldalign	ComRNA
1	Cobalamin	RF00174	71	49	216	4	**0.59**	0.05	0	X	-	0
2	ctRtNA_pGA1	RF00236	17	74	83	2	**0.91**	0.70	0.72	0	0.86	0
3	Entero_CRE	RF00048	56	81	61	1	**0.89**	0.74	0.22	0	-	0
4	Entero_OriR	RF00041	35	77	73	2	**0.94**	0.75	0.76	0.80	0.52	0.52
5	glmS	RF00234	14	58	188	4	**0.83**	0.12	0.18	0	-	0.13
6	Histone3	RF00032	63	77	26	1	**1**	0	0	0	-	0
7	Intron_gpII	RF00029	75	55	92	2	**0.80**	0.30	0	0	-	0
8	IRE	RF00037	30	68	30	1	**0.77**	0.22	0	0	0.38	0
9	let-7	RF00027	9	69	84	1	**0.87**	0.08	0.42	0	0.71	0.78
10	lin-4	RF00052	9	69	72	1	**0.78**	0.51	0.75	0.41	0.65	0.24
11	Lysine	RF00168	48	48	183	4	**0.77**	0.24	0	X	-	0
12	mir-10	RF00104	11	66	75	1	**0.66**	0.59	0.60	0	0.48	0.33
13	Purine	RF00167	29	55	103	2	**0.91**	0.07	0	0	-	0.27
14	RFN	RF00050	47	66	139	4	0.39	**0.68**	0.26	0	-	0
15	Rhino_CRE	RF00220	12	71	86	1	**0.88**	0.52	0.52	0.69	0.41	0.61
16	s2m	RF00164	23	80	43	1	0.67	**0.80**	0.45	0.64	0.63	0.29
17	S_box	RF00162	64	66	112	3	**0.72**	0.11	0	0	-	0
18	SECIS	RF00031	43	43	68	1	**0.73**	0	0	0	-	0
19	Tymo_tRNA-like	RF00233	22	72	86	4	**0.81**	0.33	0.36	0.30	0.80	0.48
				Average Accuracy:			**0.79**	0.36	0.28	0.17	0.60	0.19
				Average Specificity:			0.81	0.42	0.57	**0.83**	0.60	0.65
				Average Sensitivity:			**0.77**	0.36	0.23	0.13	0.61	0.17

Table 1. Summary of Rfam test families and results. #seqs: the number sequences in each family’s seed alignment. (For ease of post processing, we only chose one sequence per EMBL ID.) %id: average sequence identity among family members. length: average length of family members (nucleotides). #hp: number of hairpin-loops in the consensus structure. Last 6 columns: accuracies; **bold** highlights the best result in each row. CW/Pfold: Pfold using ClustalW alignment. CW/RNAalifold: similar. (X: Carnac terminated abnormally, presumably due to memory problems. -: Foldalign (pairwise) not tested due to the heavy computation cost. RNAalifold, Carnac and ComRNA do not predict any consensus structure in many cases, so the corresponding accuracies are 0.)

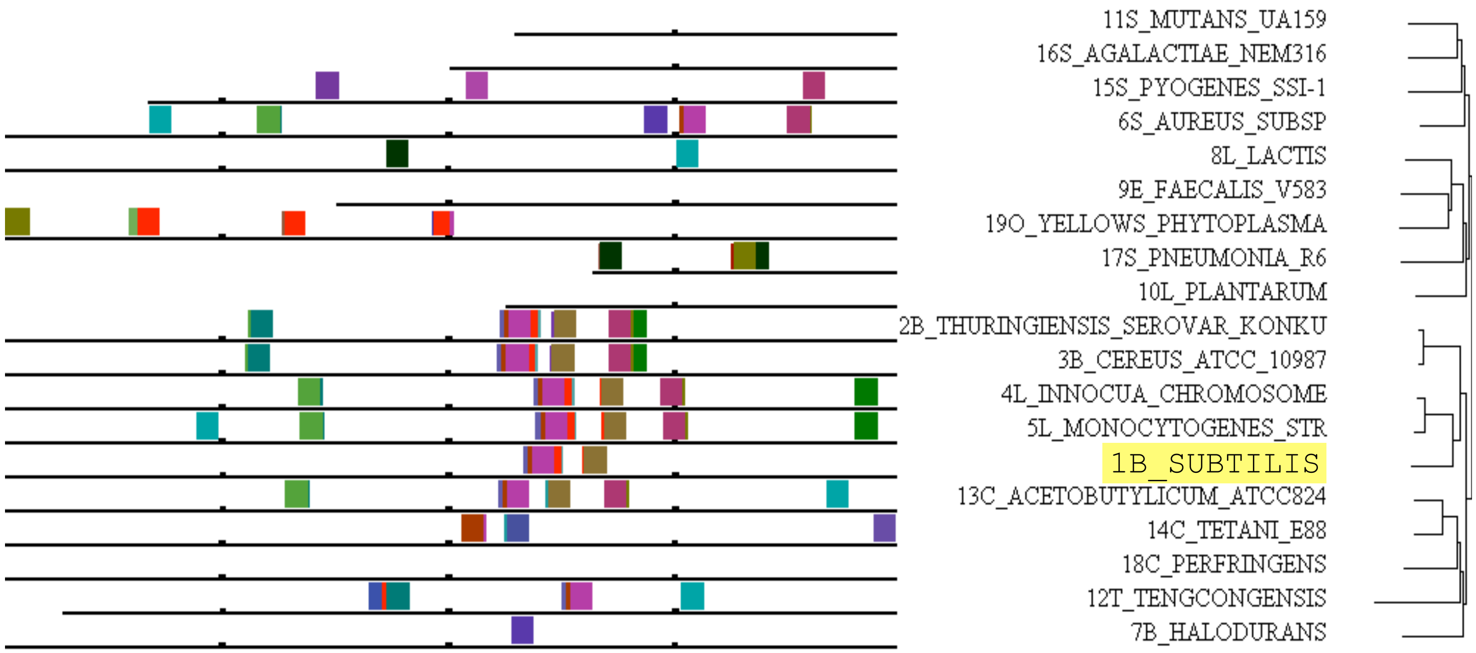
A pipeline for RNA motif genome scans



12/8/05

Footprinter find patterns of conservation

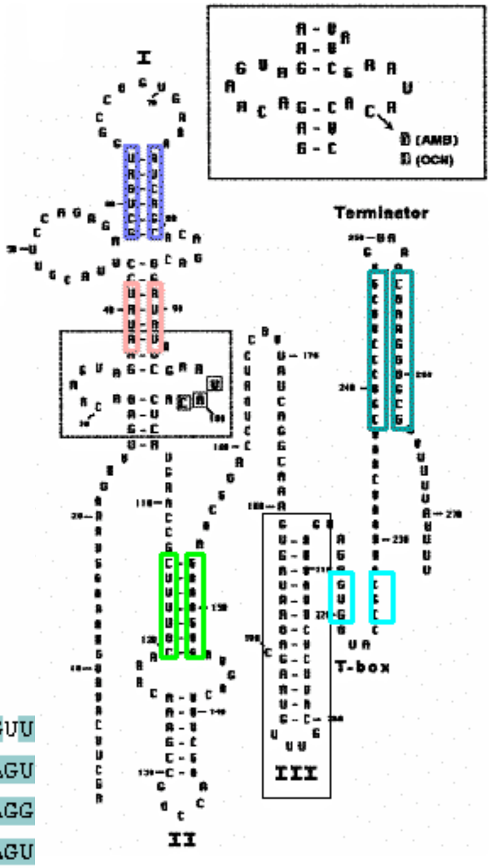
Upstream of folC



A blind test

1ST genome scan:	234 sequences
2ND genome scan:	447 sequences
The motif turned out to be T box	
Match to RFAM T box family:	299 OF 342
False Positives:	89/148 are probable (upstream of annotated tRNA-synthetase genes)

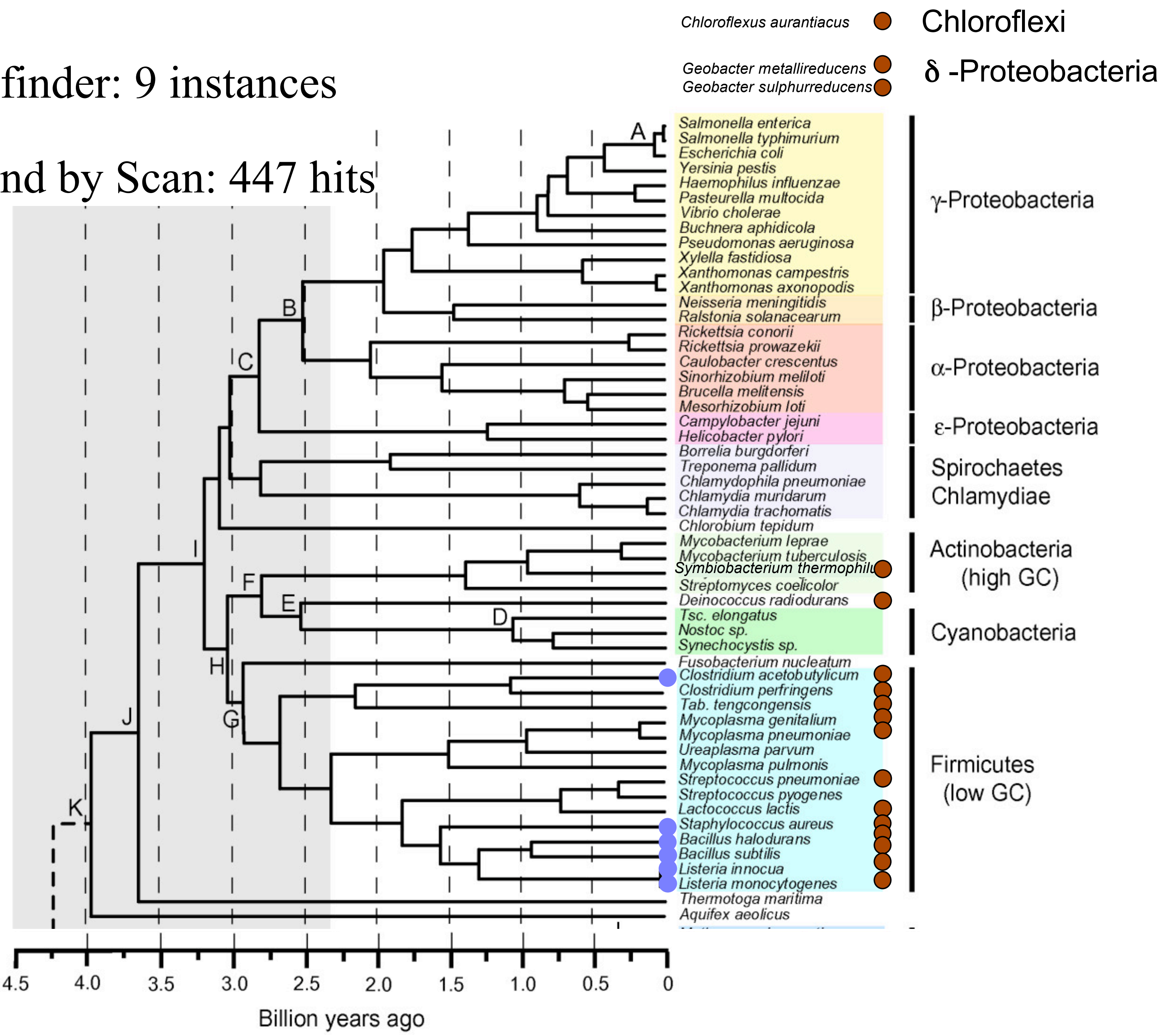
```
AUAUC.CUUACGU..UCCAGAGAGCUGAUGGCCGGUGAAA.AUCAGCACAGACGGAUAUAU
CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUUG
AAAAGUAGAACCG.AUCUAGCGAAUUGAGGAU.GGUGUGAGCUCAGUGC.GGAAAGCUUUU
CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUUG

CGAA..UACACUCAUGAACCGCUUUUGCAAACAAAGccggccaggcuuucAGUA.GUGAAAG
UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuuggauu.....AGUAAGCAUUCC
AGAAAAUC.ACUCUUGAGUU.UUCAUUACGAAA..CA.............AGUAGUAAUGGA
UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuaugauu.....AGUAAACAUUCC

aCGGAC.CUGAUCCGUUAUCAGGCAAAGUGGUACCGCGAUAAUCAAUCGUCCCUUCGUGUAAaCGAAGGGGCGUU
.CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAuAGGGACGGGAGU
.CGGUUcAUC.UCCGUUAUCGAUCUUAGUGGUACCGCGA.....GUCUUCUCGUCCCUUUU..GGGAUUAGAAGG
.CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAuAGGGACGGGAGU
```



tyrS T box structure

● CMfinder: 9 instances

● Found by Scan: 447 hits



Preliminary results of genome scan

Top 115 datasets (some are redundant)
13 T box, 22 riboswitches, 30 ribosomal genes
RNase P, tRNA, CIRCE elements and other DNA binding sites

Gene	#motif	#hits	RFAM_fam	#Rfam_seed	#Rfam_full	#TP	specificity	sensitivity
metK	13	150	S_box	71	151	145	0.967	0.960
ribB	9	106	RFN	48	114	97	0.915	0.851
folC	9	447	T_box	67	342	299	0.669	0.874
xpt	14	106	Purine	37	100	97	0.915	0.970
glmS	16	33	glmS	14	37	33	1.000	0.892
thiA	16	305	THI	237	366	305	1.000	0.833
ykoY	10	34	yybP-ykoY	74	127	33	0.971	0.260

Genome Scans in Progress

- Firmicutes
  - e.g. anthrax
- Actinobacteria
  - source of penicillin & most other antibiotics
- Cyanobacteria
  - Primary producer of oxygen
- Gamma-proteobacteria
  - e.g. E. coli.