RNA Search and Motif Discovery

CSE 527
Computational Biology

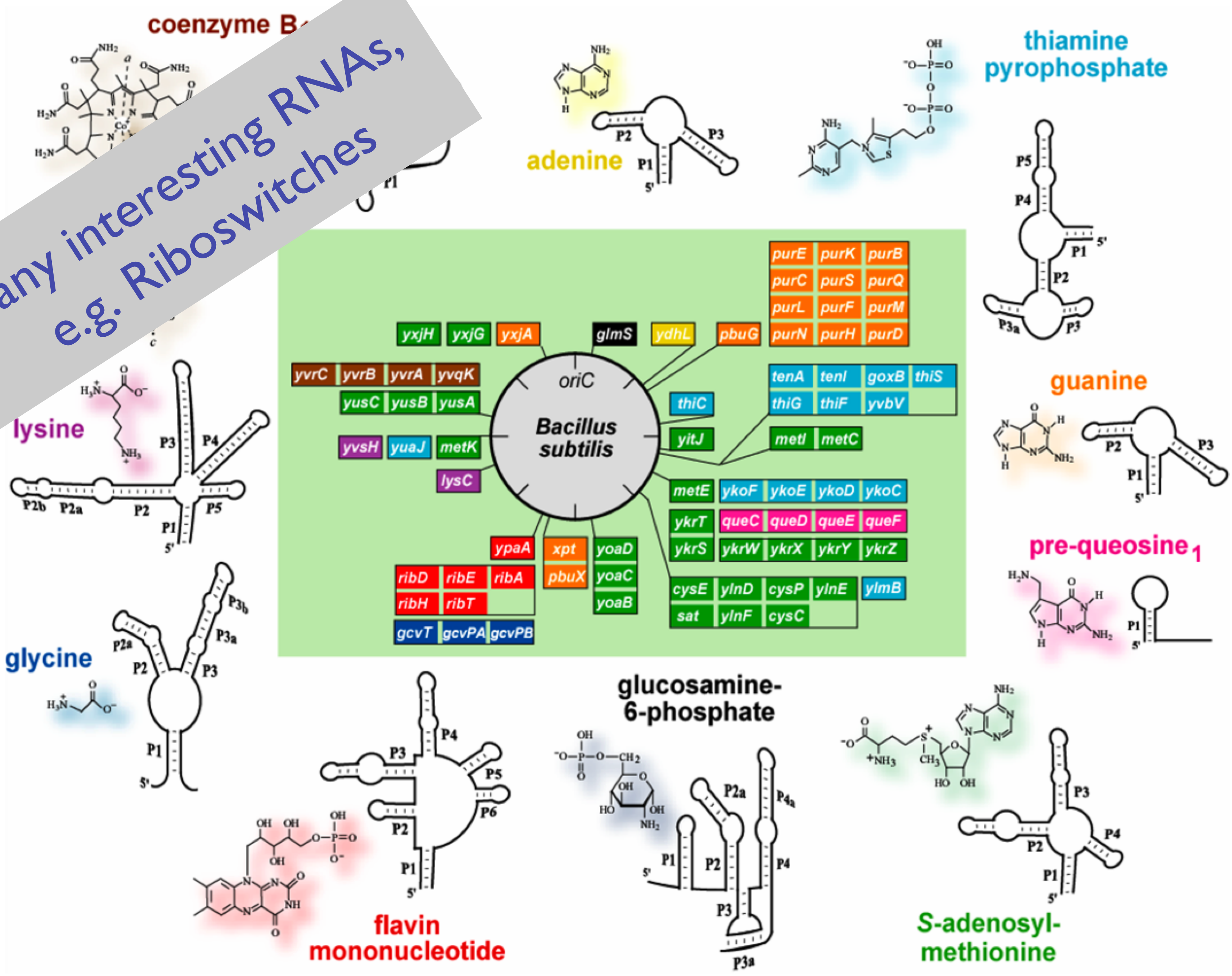
Day 1

Last lecture:
many biologically interesting roles for RNA

Today:

Covariance Models (CMs) represent conserved RNA sequence/structure motifs

Many interesting RNAs, e.g. Riboswitches



Computational Problems

How to predict secondary structure

How to model an RNA "motif"
(I.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 sloooow

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

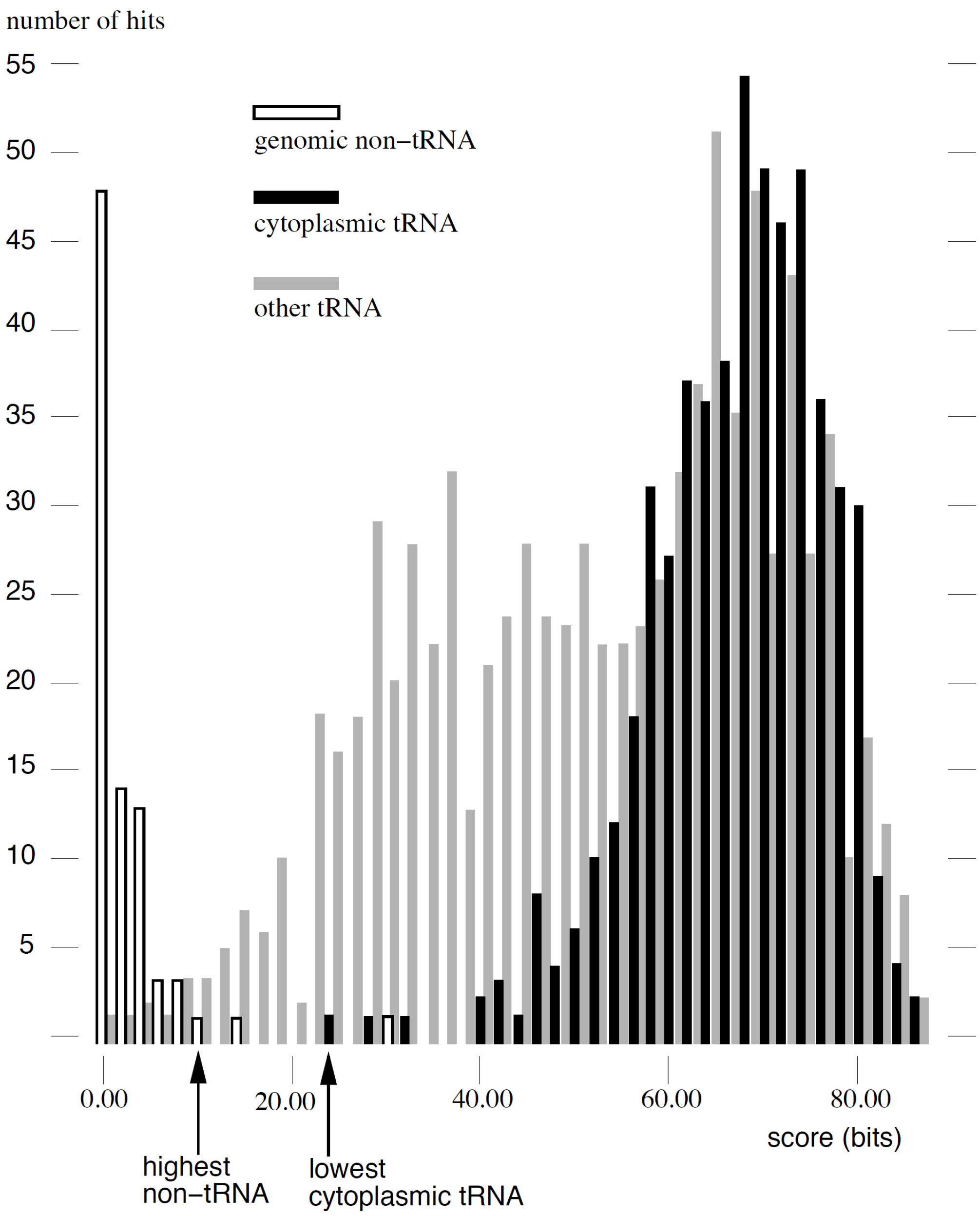
You set a score threshold

Anything above threshold → a “hit”

Scoring:

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

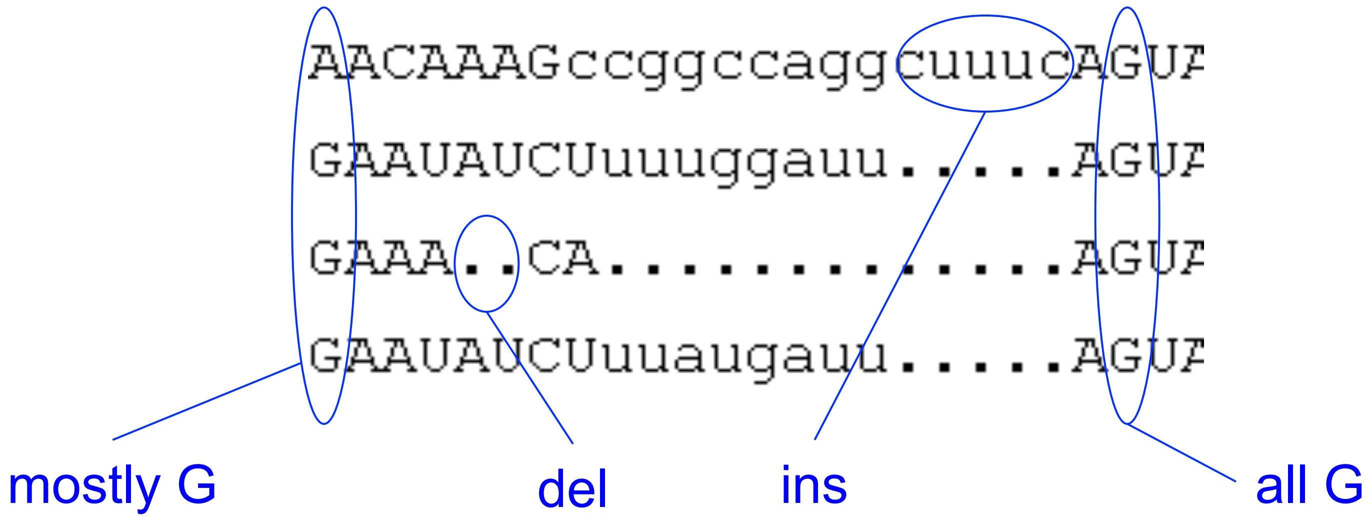
Example: searching for tRNAs



How to model an RNA “Motif”?

Conceptually, start with a profile HMM:

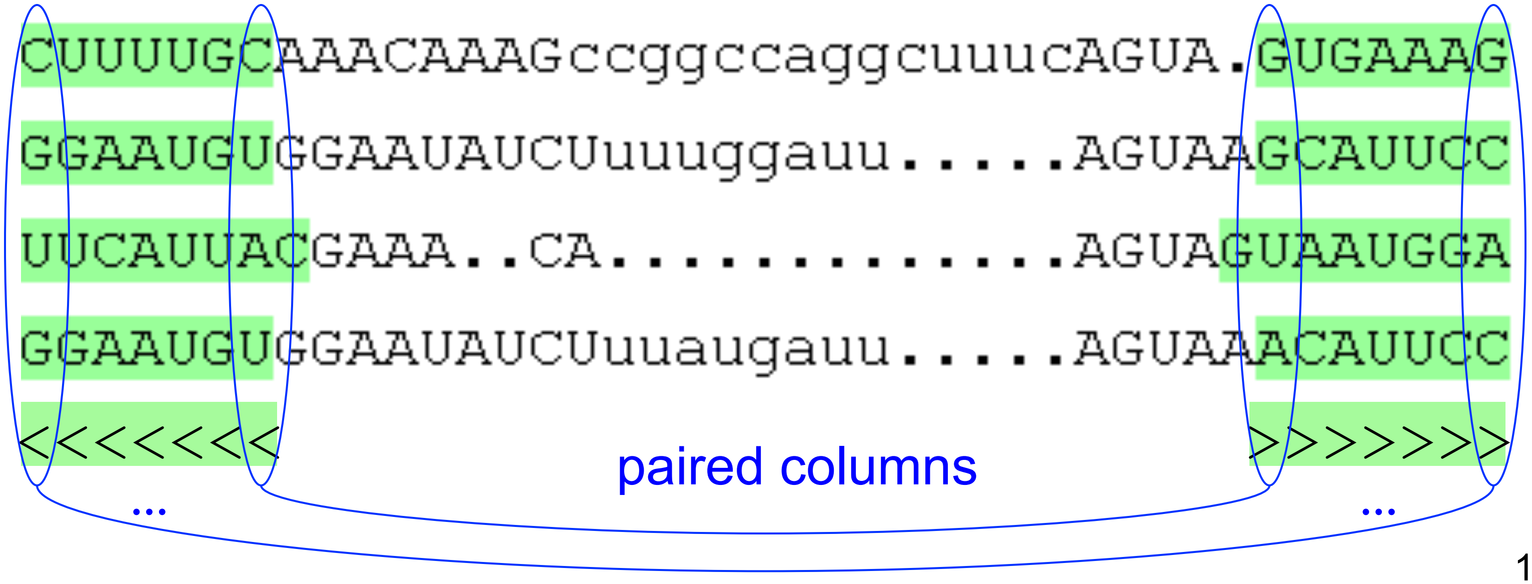
- from a multiple alignment, estimate nucleotide/ insert/delete preferences for each position
- given a new seq, estimate likelihood that it could be generated by the model, & align it to the model



```
AACAAAGccggccaggcuuucAGU
GAAUAUCUuuuggauu.....AGU
GAAA..CA.............AGU
GAAUAUCUuuaugauu.....AGU
```

How to model an RNA “Motif”?

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



```
CUUUUGCAAACAAAGccggccaggcuuucAGUA.GUGAAAG
GGAAUGUGGAAUAUCUuuuggauu.....AGUAAGCAUUCC
UUCAUUACGAAA..CA.............AGUAGUAAUGGA
GGAAUGUGGAAUAUCUuuaugauu.....AGUAAACAUUCC
<<<<<<<                            >>>>>>>
```

Profile Hmm Structure

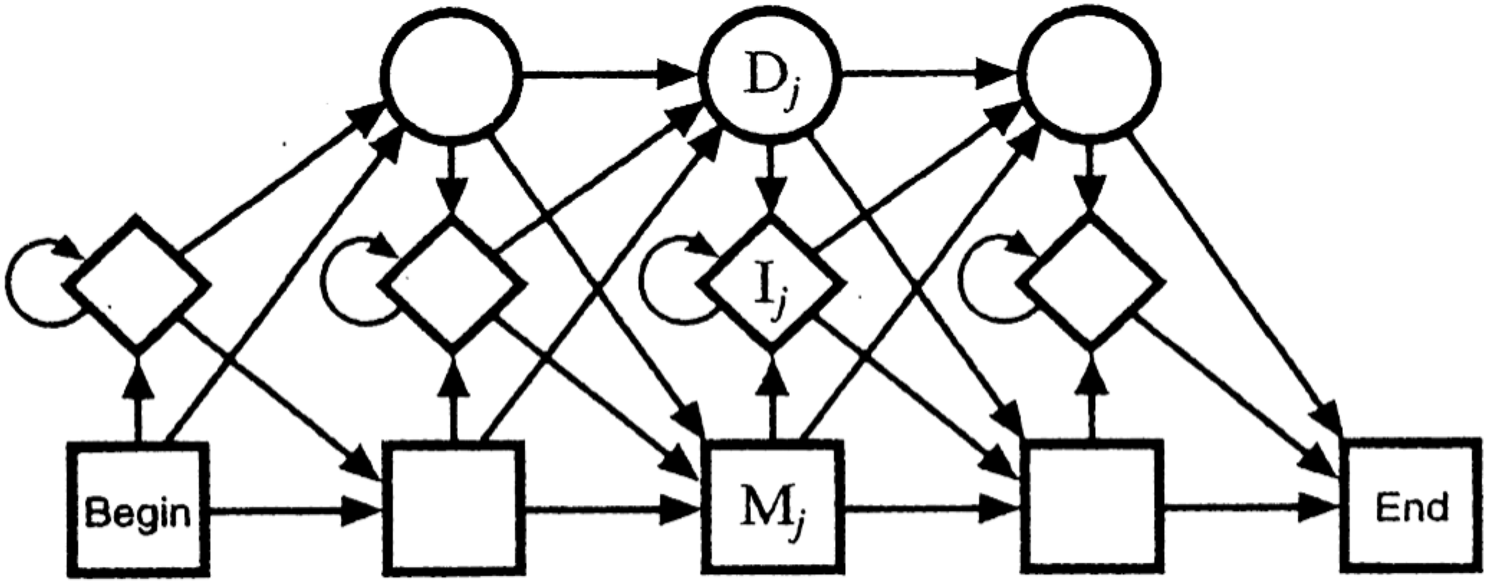


Figure 5.2 *The transition structure of a profile HMM.*

Mj: Match states (20 emission probabilities)

Ij: Insert states (Background emission probabilities)

Dj: Delete states (silent - no emission)

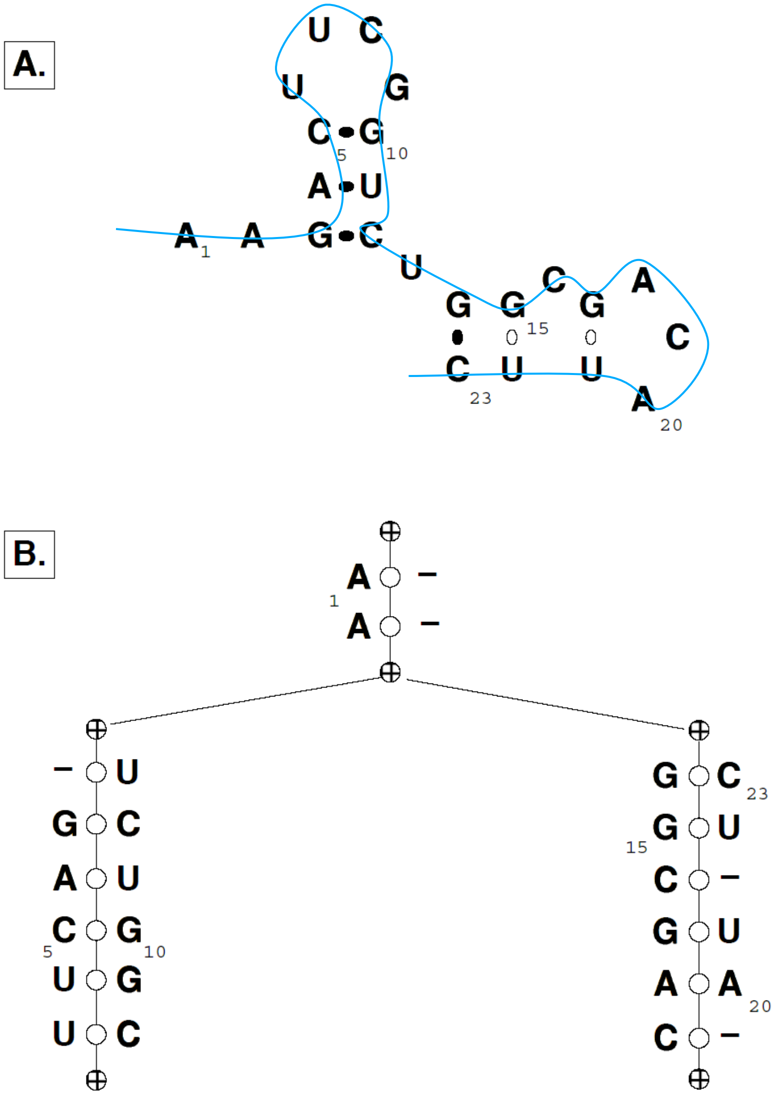
CM Structure

A: Sequence + structure

B: 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)

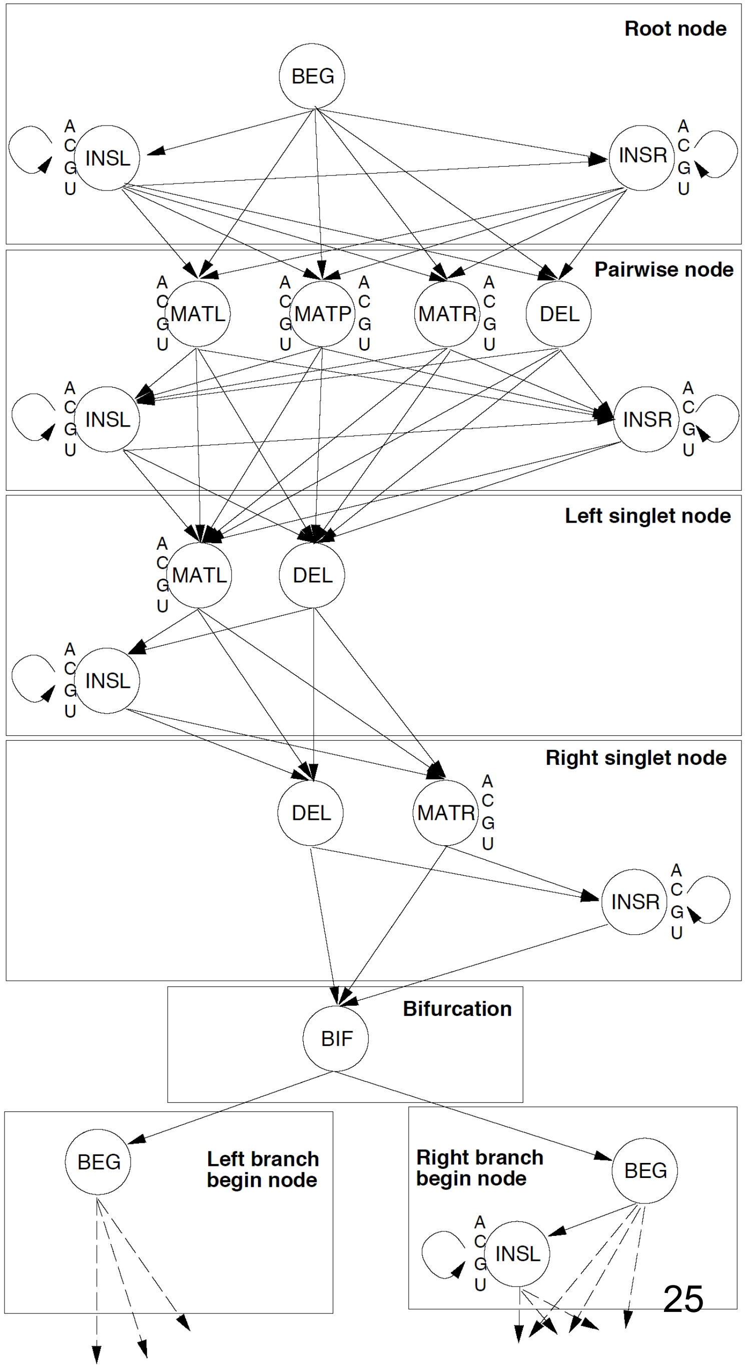


Overall CM Architecture

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 base-pairs; BIF allows multiple helices



CM Viterbi Alignment

(the “inside” algorithm)

x_i = i^{th} letter of input

x_{ij} = substring $i,\ldots,j$ of input

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

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

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

CM Viterbi Alignment

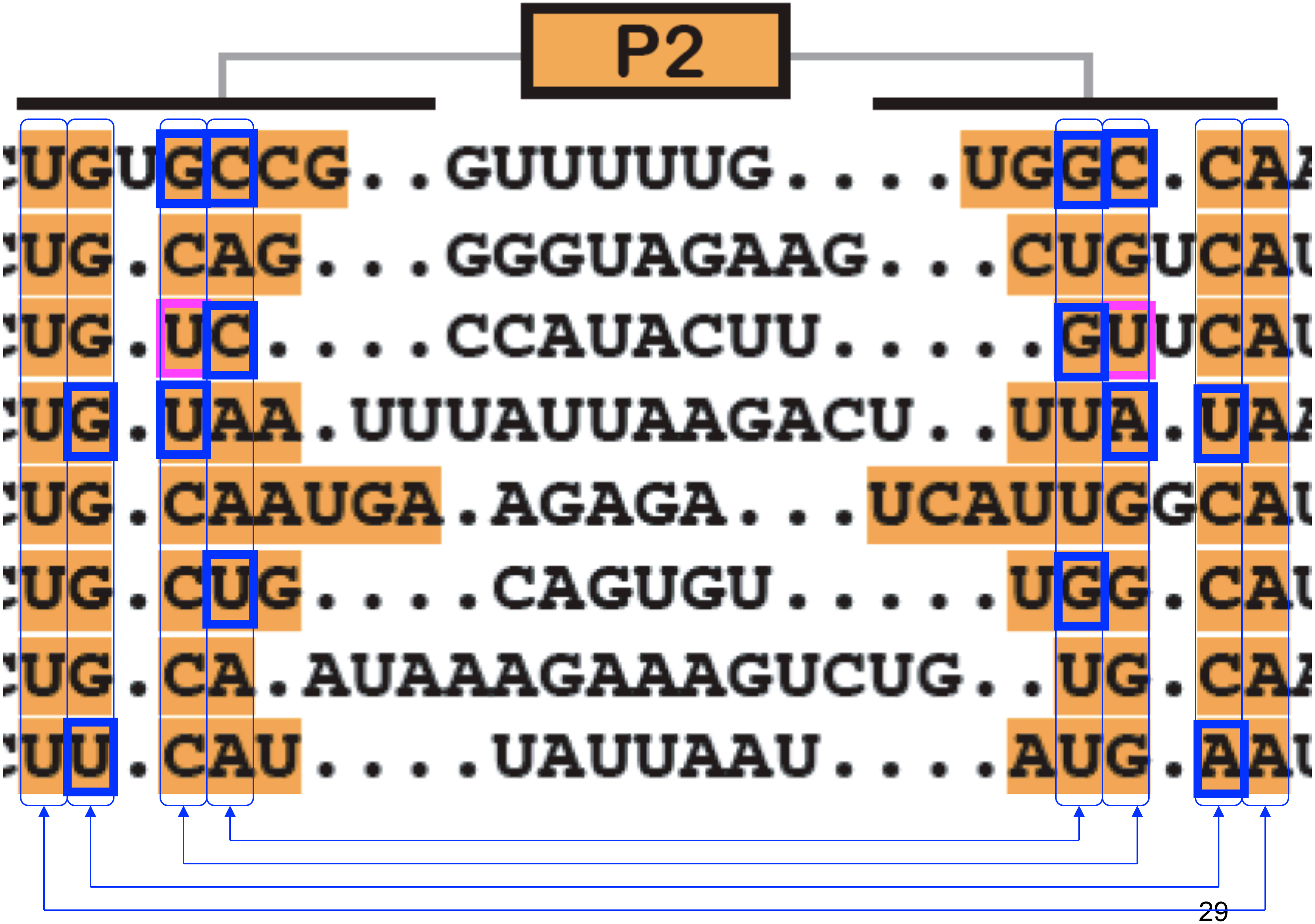
(the “inside” algorithm)

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

$$S_{ij}^{y} = \begin{cases} \max_{z}[S_{i+1,j-1}^{z} + \log T_{yz} + \log E_{x_i,x_j}^{y}] & \text{match pair} \\ \max_{z}[S_{i+1,j}^{z} + \log T_{yz} + \log E_{x_i}^{y}] & \text{match/insert left} \\ \max_{z}[S_{i,j-1}^{z} + \log T_{yz} + \log E_{x_j}^{y}] & \text{match/insert right} \\ \max_{z}[S_{i,j}^{z} + \log T_{yz}] & \text{delete} \\ \max_{i<k\leq j}[S_{i,k}^{y_{left}} + S_{k+1,j}^{y_{right}}] & \text{bifurcation} \end{cases}$$

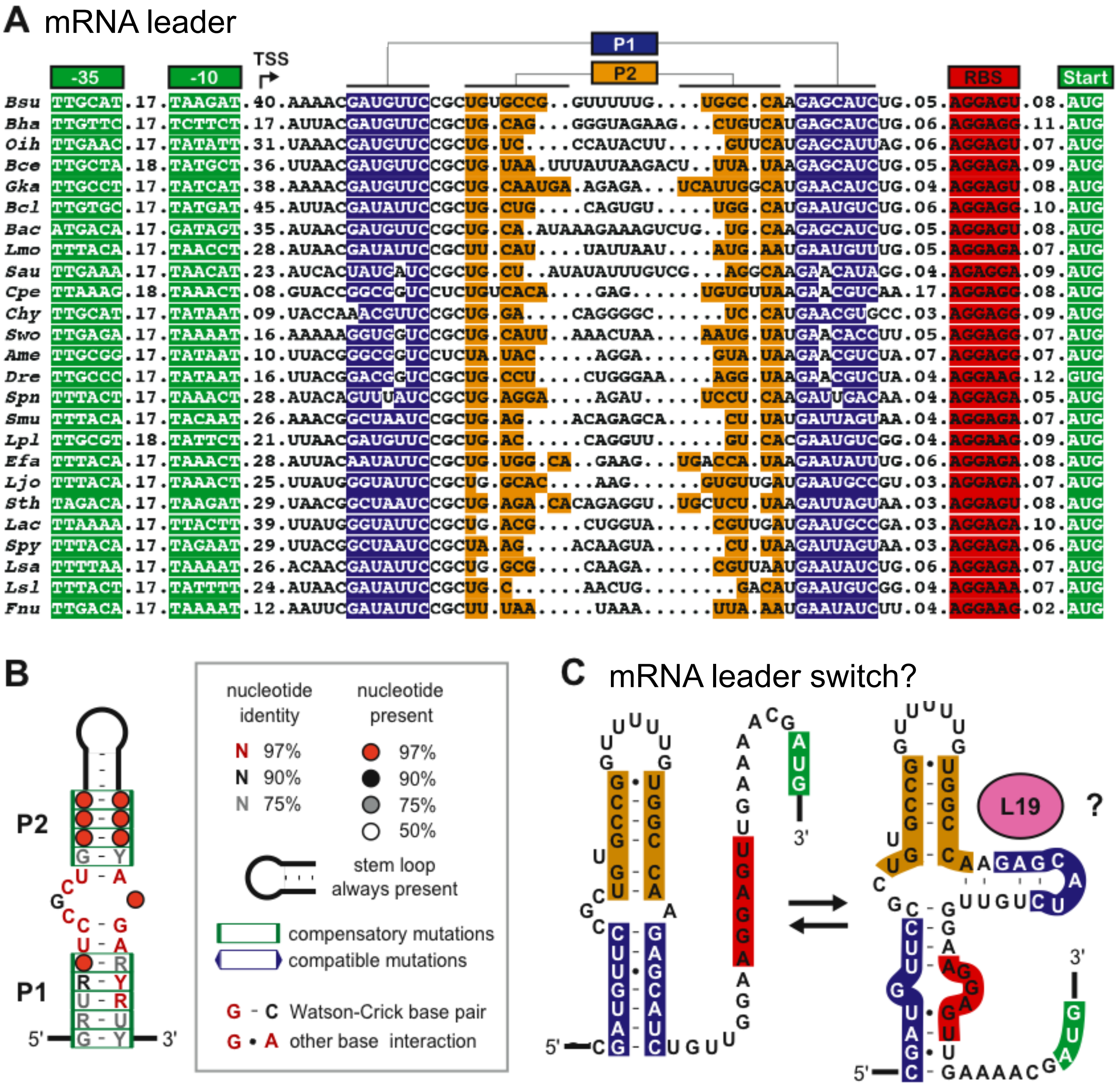
Time O(qn^3), q states, seq len n

compare: O(qn) for profile HMM



```
UGUGCCG..GUUUUUG....UGGC.CA
UG.CAG...GGGUAGAAG...CUGUCA
UG.UC....CCAUACUU.....GUUCA
UG.UAA.UUUAUUAAGACU..UUA.UA
UG.CAAUGA.AGAGA...UCAUUGGCA
UG.CUG....CAGUGU.....UGG.CA
UG.CA.AUAAAGAAAGUCUG..UG.CA
UU.CAU....UAUUAAU....AUG.AA
```

Covariation is strong evidence for base pairing



```
Bsu TTGCAT.17.TAAGAT.40.AAAACGAUGUUCCGCUGUGCCG..GUUUUUG....UGGC.CAAGAGCAUCUG.05.AGGAGU.08.AUG
Bha TTGTTC.17.TCTTCT.17.AUUACGAUGUUCCGCUG.CAG...GGGUAGAAG...CUGUCAUGAGCAUCUG.06.AGGAGG.11.AUG
Oih TTGAAC.17.TATATT.31.UAAACGAUGUUCCGCUG.UC....CCAUACUU.....GUUCAUGAGCAUUAG.06.AGGAGU.07.AUG
Bce TTGCTA.18.TATGCT.36.UUAACGAUGUUCCGCUG.UAA.UUUAUUAAGACU..UUA.UAAGAGCAUCUG.05.AGGAGA.09.AUG
Gka TTGCCT.17.TATCAT.38.AAAACGAUGUUCCGCUG.CAAUGA.AGAGA...UCAUUGGCAUGAACAUCUG.04.AGGAGU.08.AUG
Bcl TTGTGC.17.TATGAT.45.AUUACGAUAUUCCGCUG.CUG....CAGUGU.....UGG.CAUGAAUGUCUG.06.AGGAGG.10.AUG
Bac ATGACA.17.GATAGT.35.AUAACGAUGUUCCGCUG.CA.AUAAAGAAAGUCUG..UG.CAAGAGCAUCUG.05.AGGAGU.08.AUG
Lmo TTTACA.17.TAACCT.28.AUAACGAUAUUCCGCUU.CAU....UAUUAAU....AUG.AAUGAAUGUUUG.05.AGGAGA.07.AUG
Sau TTGAAA.17.TAACAT.23.AUCACUAUGAUCCGCUG.CU..AUAUAUUUGUCG...AGGCAAGAACAUAGG.04.AGAGGA.09.AUG
Cpe TTAAAG.18.TAAACT.08.GUACCGGCGGUCCUCUGUCACA....GAG......UGUGUUAAGAACGUCAA.17.AGGAGG.08.AUG
Chy TTGCAT.17.TATAAT.09.UACCAAACGUUCCGCUG.GA....CAGGGGC......UC.CAUGAACGUGCC.03.AGGAGG.09.AUG
Swo TTGAGA.17.TAAAAT.16.AAAAAGGUGGUCCGCUG.CAUU..AAACUAA....AAUG.UAUGAACACCUU.05.AGGAGG.07.AUG
Ame TTGCGG.17.TATAAT.10.UUACGGGCGGUCCUCUA.UAC.....AGGA......GUA.UAAGAACGUCUA.07.AGGAGG.07.AUG
Dre TTGCCC.17.TATAAT.16.UUACGGACGGUCCGCUG.CCU....CUGGGAA....AGG.UAAGAACGUCUA.04.AGGAAG.12.GUG
Spn TTTACT.17.TAAACT.28.AUACAGUUUAUCCGCUG.AGGA....AGAU.....UCCU.CAAGAUUGACAA.04.AGGAGA.05.AUG
Smu TTTACA.17.TACAAT.26.AAACGGCUAAUCCGCUG.AG....ACAGAGCA.....CU.UAUGAUUAGUAA.04.AGGAGA.07.AUG
Lpl TTGCGT.18.TATTCT.21.UUAACGAUGUUCCGCUG.AC.....CAGGUU......GU.CACGAAUGUCGG.04.AGGAAG.09.AUG
Efa TTTACA.17.TAAACT.28.AUUACAAUAUUCCGCUG.UGG.CA..GAAG...UGACCA.UAAGAAUAUUUG.06.AGGAGA.08.AUG
Ljo TTTACA.17.TAAACT.25.UUAUGGGUAUUCCGCUG.GCAC....AAG......GUGUUGAUGAAUGCCGU.03.AGGAGA.07.AUG
Sth TAGACA.17.TAAGAT.29.UAACGGCUAAUCCGCUG.AGA.CACAGAGGU..UGCUCU.UAAGAUUAGUAA.03.AGGAGU.08.AUG
Lac TTAAAA.17.TTACTT.39.UUAUGGGUAUUCCGCUG.ACG....CUGGUA.....CGUUGAUGAAUGCCGA.03.AGGAGA.10.AUG
Spy TTTACA.17.TAGAAT.29.UUACGGCUAAUCCGCUA.AG....ACAAGUA......CU.UAAGAUUAGUAA.03.AGGAGA.06.AUG
Lsa TTTTAA.17.TAAAAT.26.ACAACGAUAUUCCGCUG.GCG....CAAGA......CGUUAAUGAAUAUCUG.06.AGGAGA.07.AUG
Lsl TTTACT.17.TATTTT.24.AUAACGAUAUUCCGCUG.C......AACUG........GACAUGAAUGUCGG.04.AGGAAA.07.AUG
Fnu TTGACA.17.TAAAAT.12.AAUUCGAUAUUCCGCUU.UAA.....UAAA......UUA.AAUGAAUAUCUU.04.AGGAAG.02.AUG
```

Mutual Information

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

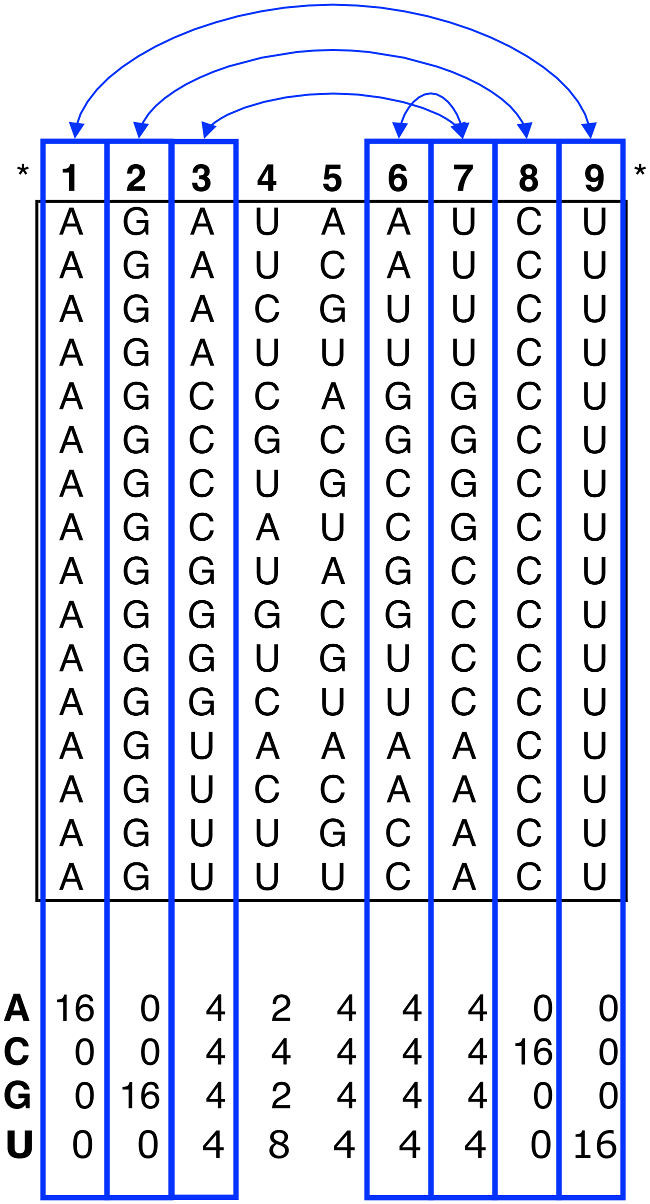
Max when *no* seq conservation but perfect pairing

MI = expected score gain from using a pair state

Finding optimal MI, (i.e. opt pairing of cols) is hard(?)

Finding optimal MI *without pseudoknots* can be done by dynamic programming

M.I. Example (Artificial)



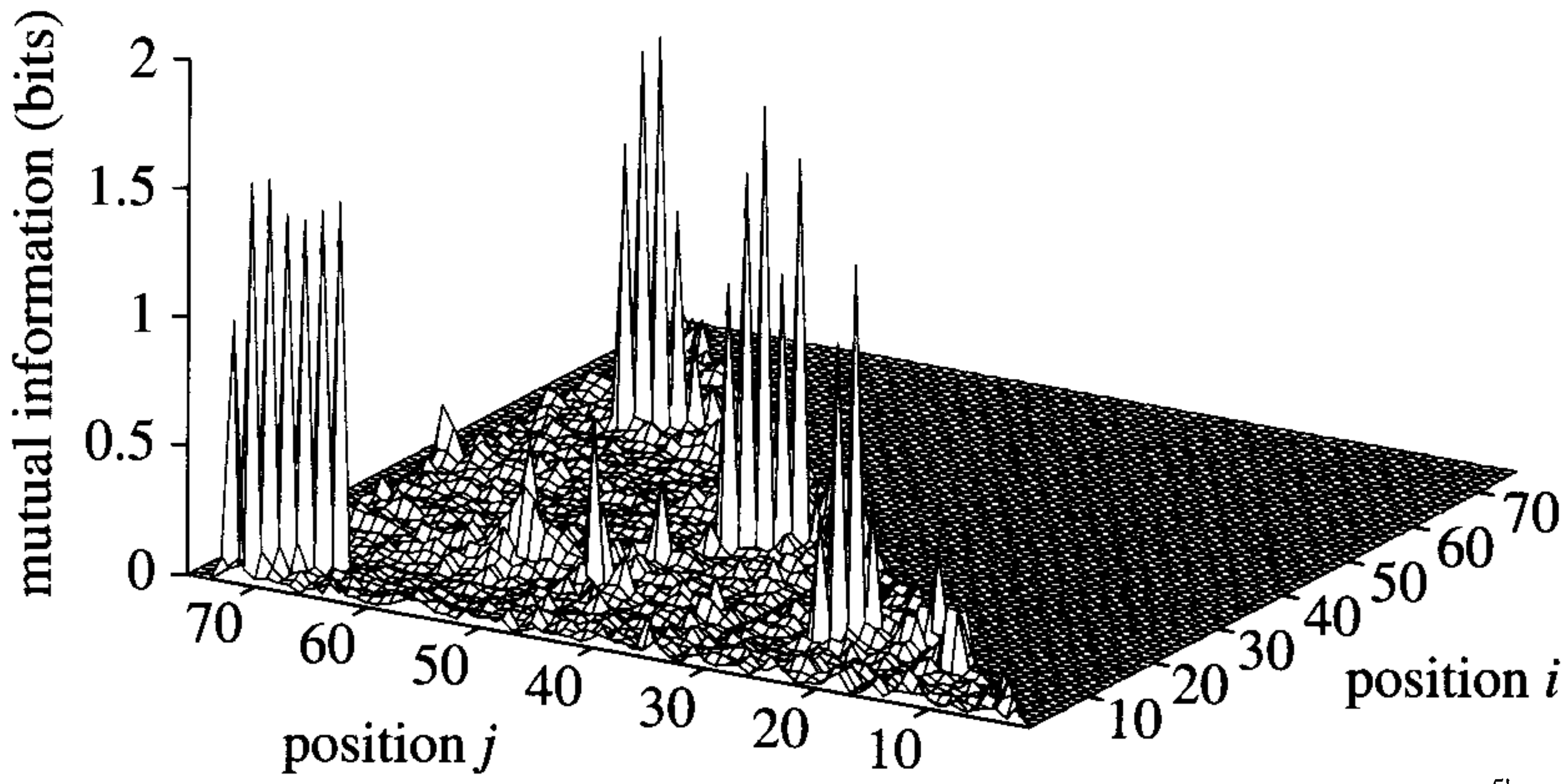
*	1	2	3	4	5	6	7	8	9	*
	A	G	A	U	A	A	U	C	U	
	A	G	A	U	C	A	U	C	U	
	A	G	A	C	G	U	U	C	U	
	A	G	A	U	U	U	U	C	U	
	A	G	C	C	A	G	G	C	U	
	A	G	C	G	C	G	G	C	U	
	A	G	C	U	G	C	G	C	U	
	A	G	C	A	U	C	G	C	U	
	A	G	G	U	A	G	C	C	U	
	A	G	G	G	C	G	C	C	U	
	A	G	G	U	G	U	C	C	U	
	A	G	G	C	U	U	C	C	U	
	A	G	U	A	A	A	A	C	U	
	A	G	U	C	C	A	A	C	U	
	A	G	U	U	G	C	A	C	U	
	A	G	U	U	U	C	A	C	U	
A	16	0	4	2	4	4	4	0	0	
C	0	0	4	4	4	4	4	16	0	
G	0	16	4	2	4	4	4	0	0	
U	0	0	4	8	4	4	4	0	16	

MI:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
9	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0		
7	0	0	**2**	**0.30**	**0**	**1**			
6	0	0	**1**	**0.55**	**1**				
5	0	0	**0**	**0.42**					
4	0	0	**0.30**						
3	0	0							
2	0								
1									

Cols 1 & 9, 2 & 8: perfect conservation & *might* be base-paired, but unclear whether they are. M.I. = 0

Cols 3 & 7: *No* conservation, but always W-C pairs, so seems likely they do base-pair. M.I. = 2 bits.

Cols 7->6: unconserved, but each letter in 7 has only 2 possible mates in 6. M.I. = 1 bit.



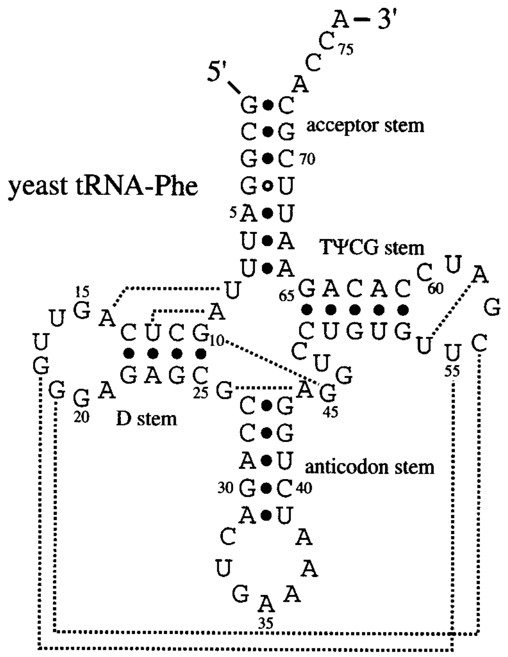


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary contacts observed in the yeast tRNA-Phe crystal structure. Some of these tertiary contacts produce correlated pairs which can be seen weakly in the mutual information plot.*

MI-Based Structure-Learning

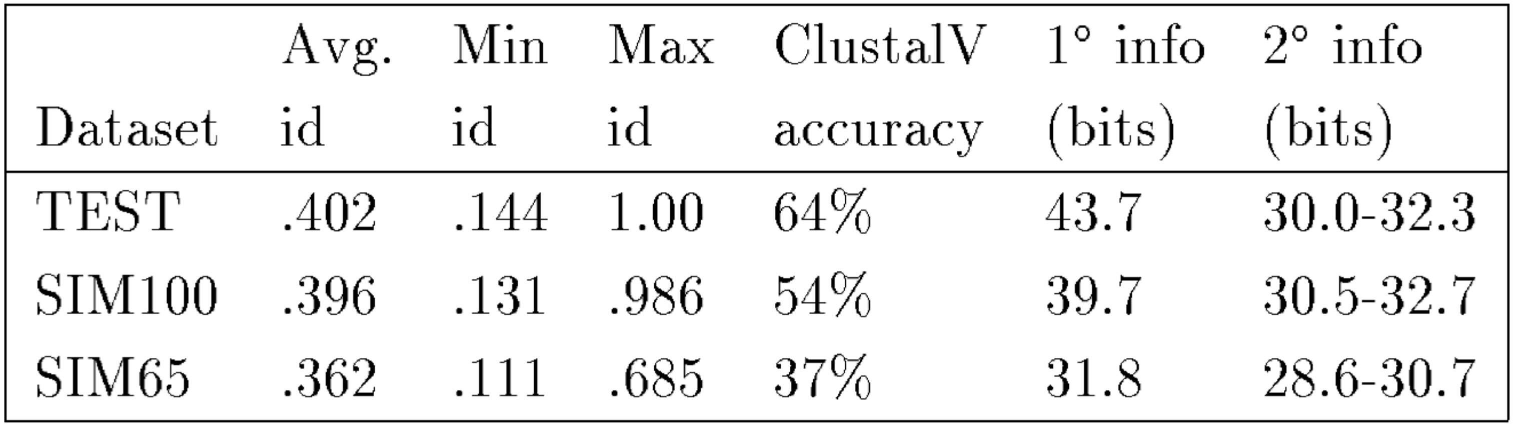
Find best (max total MI) subset of column pairs among i…j, subject to absence of pseudo-knots

$$S_{i,j} = \max \begin{cases} S_{i,j-1} & \text{j unpaired} \\ \max_{i \le k < j-4} S_{i,k-1} + M_{k,j} + S_{k+1,j-1} & \text{j paired} \end{cases}$$

“Just like Nussinov/Zucker folding”

BUT, need enough data---enough sequences at right phylogenetic distance

Primary vs Secondary Info



Dataset	Avg. id	Min id	Max id	ClustalV accuracy	1° info (bits)	2° info (bits)
TEST	.402	.144	1.00	64%	43.7	30.0-32.3
SIM100	.396	.131	.986	54%	39.7	30.5-32.7
SIM65	.362	.111	.685	37%	31.8	28.6-30.7

disallowing allowing

pseudoknots

$$\left(\sum_{i=1}^{n} \max_{j} M_{i,j}\right)/2$$

Comparison to TRNASCAN

Fichant & Burks - best heuristic then

- 97.5% true positive
- 0.37 false positives per MB

CM A1415 (trained on trusted alignment)

- > 99.98% true positives
- < 0.2 false positives per MB

Slightly different evaluation criteria

Current method-of-choice is “tRNAscanSE”, a CM-based scan with heuristic pre-filtering (including TRNASCAN?) for performance reasons.

tRNAScanSE

Uses 3 older heuristic tRNA finders as prefilter

Uses CM built as described for final scoring

Actually 3(?) different CMs

- eukaryotic nuclear
- prokaryotic
- organellar

Used in all genome annotation projects

An Important Application: Rfam

Rfam – an RNA family DB

Griffiths-Jones, et al., NAR ’03, ’05, ’08

Biggest scientific computing user in Europe - 1000 cpu cluster for a month per release

Rapidly growing:

Rel 1.0, 1/03:	25 families,	55k instances
Rel 7.0, 3/05:	503 families,	>300k instances
Rel 9.0, 7/08:	603 families,	896k instances
Rel 9.1, 1/09:	1372 families,	??? instances

Rfam database

http://www.sanger.ac.uk/Software/Rfam/

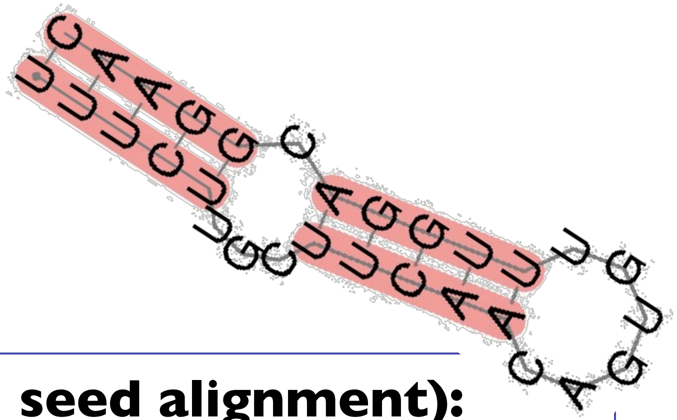
(Release 7.0, 3/2005)

503 ncRNA families

280,000 annotated ncRNAs

8 riboswitches, 235 small nucleolar RNAs, 8 spliceosomal RNAs, 10 bacterial antisense RNAs, 46 microRNAs, 9 ribozymes, 122 *cis* RNA regulatory elements, …

Example Rfam Family



Input (hand-curated):

- MSA “seed alignment”
- SS_cons
- Score Thresh T
- Window Len W

Output:

- CM
- scan results & “full alignment”

IRE (partial seed alignment):

```
Hom.sap.   GUUCCUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCUUC.UUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCCUGUUUCAACAGUGCUUGGA.GGAAC
Hom.sap.   UUUAUC..AGUGACAGAGUUCACU.AUAAA
Hom.sap.   UCUCUUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   AUUAUC..GGGAACAGUGUUUCCC.AUAAU
Hom.sap.   UCUUGC..UUCAACAGUGUUUGGACGGAAG
Hom.sap.   UGUAUC..GGAGACAGUGAUCUCC.AUAUG
Hom.sap.   AUUAUC..GGAAGCAGUGCCUUCC.AUAAU
Cav.por.   UCUCCUGCUUCAACAGUGCUUGGACGGAGC
Mus.mus.   UAUAUC..GGAGACAGUGAUCUCC.AUAUG
Mus.mus.   UUUCCUGCUUCAACAGUGCUUGAACGGAAC
Mus.mus.   GUACUUGCUUCAACAGUGUUUGAACGGAAC
Rat.nor.   UAUAUC..GGAGACAGUGACCUCC.AUAUG
Rat.nor.   UAUCUUGCUUCAACAGUGUUUGGACGGAAC
SS_cons    <<<<<...<<<<<......>>>>>.>>>>>
```

Rfam – key issues

Overly narrow families

Variant structures/unstructured RNAs

Spliced RNAs

RNA pseudogenes

- Human ALU is SRP related w/ 1.1m copies
- Mouse B2 repeat (350k copies) tRNA related

Speed & sensitivity

Motif discovery

Day 2
5 slide synopsis of last lecture

Covariance Models (CMs) represent conserved RNA sequence/structure motifs

They allow accurate search

But

a) search is slow

b) model construction is laborious

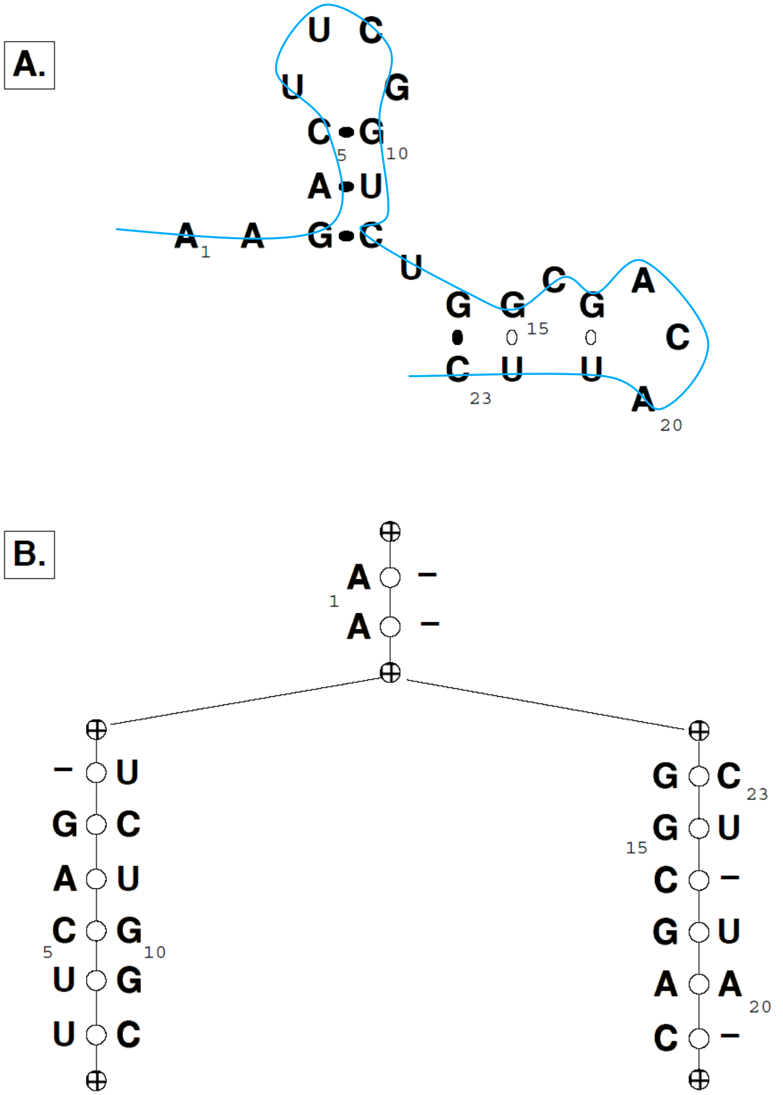
CM Structure

quence + structure

B: the CM “guide tree”

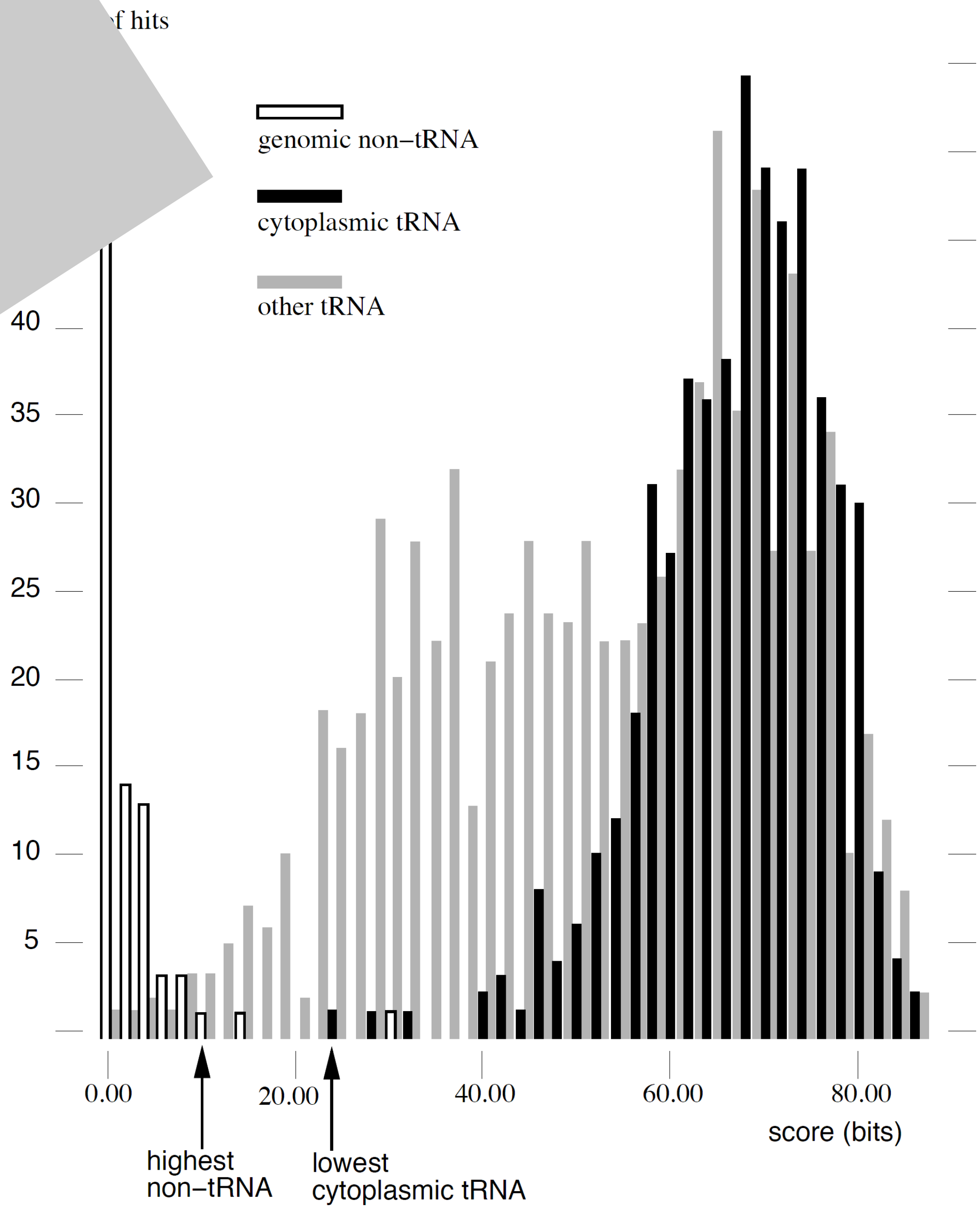
C: probabilities of letters/ pairs & of indels

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



Example:
searchin
tRN

Accurate Search



But Slow

terbi Alignment
(the "inside" algorithm)

$$\text{ax}_\pi \log P(x_{ij} \text{ generated starting in state } y \text{ via path } \pi)$$

$$S_{ij}^{y} = \begin{cases} \max_z [S_{i+1,j-1}^{z} + \log T_{yz} + \log E_{x_i, x_j}^{y}] & \text{match pair} \\ \max_z [S_{i+1,j}^{z} + \log T_{yz} + \log E_{x_i}^{y}] & \text{match/insert left} \\ \max_z [S_{i,j-1}^{z} + \log T_{yz} + \log E_{x_j}^{y}] & \text{match/insert right} \\ \max_z [S_{i,j}^{z} + \log T_{yz}] & \text{delete} \\ \max_{i<k\leq j} [S_{i,k}^{y_{left}} + S_{k+1,j}^{y_{right}}] & \text{bifurcation} \end{cases}$$

Time O(qn^3), q states, seq len n

compare: O(qn) for profile HMM

Ex… …am Family

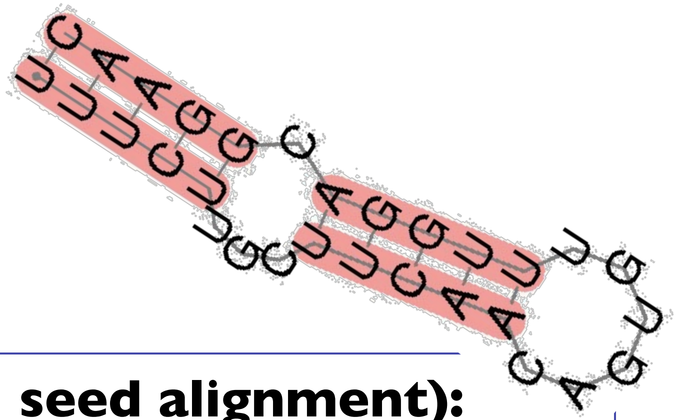
Hand-made

(hand-curated):

- MSA “seed alignment”
- SS_cons
- Score Thresh T
- Window Len W

Output:

- CM
- scan results & “full alignment”



IRE (partial seed alignment):

```
Hom.sap.   GUUCCUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCUUC.UUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCCUGUUUCAACAGUGCUUGGA.GGAAC
Hom.sap.   UUUAUC..AGUGACAGAGUUCACU.AUAAA
Hom.sap.   UCUCUUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   AUUAUC..GGGAACAGUGUUUCCC.AUAAU
Hom.sap.   UCUUGC..UUCAACAGUGUUUGGACGGAAG
Hom.sap.   UGUAUC..GGAGACAGUGAUCUCC.AUAUG
Hom.sap.   AUUAUC..GGAAGCAGUGCCUUCC.AUAAU
Cav.por.   UCUCCUGCUUCAACAGUGCUUGGACGGAGC
Mus.mus.   UAUAUC..GGAGACAGUGAUCUCC.AUAUG
Mus.mus.   UUUCCUGCUUCAACAGUGCUUGAACGGAAC
Mus.mus.   GUACUUGCUUCAACAGUGUUUGAACGGAAC
Rat.nor.   UAUAUC..GGAGACAGUGACCUCC.AUAUG
Rat.nor.   UAUCUUGCUUCAACAGUGUUUGGACGGAAC
SS_cons    <<<<<...<<<<<......>>>>>.>>>>>
```

Today's Goals

Faster Search

- Infernal & RaveNnA

Automated Model-building

- CMfinder

Faster Search

Homology search

Sequence-based

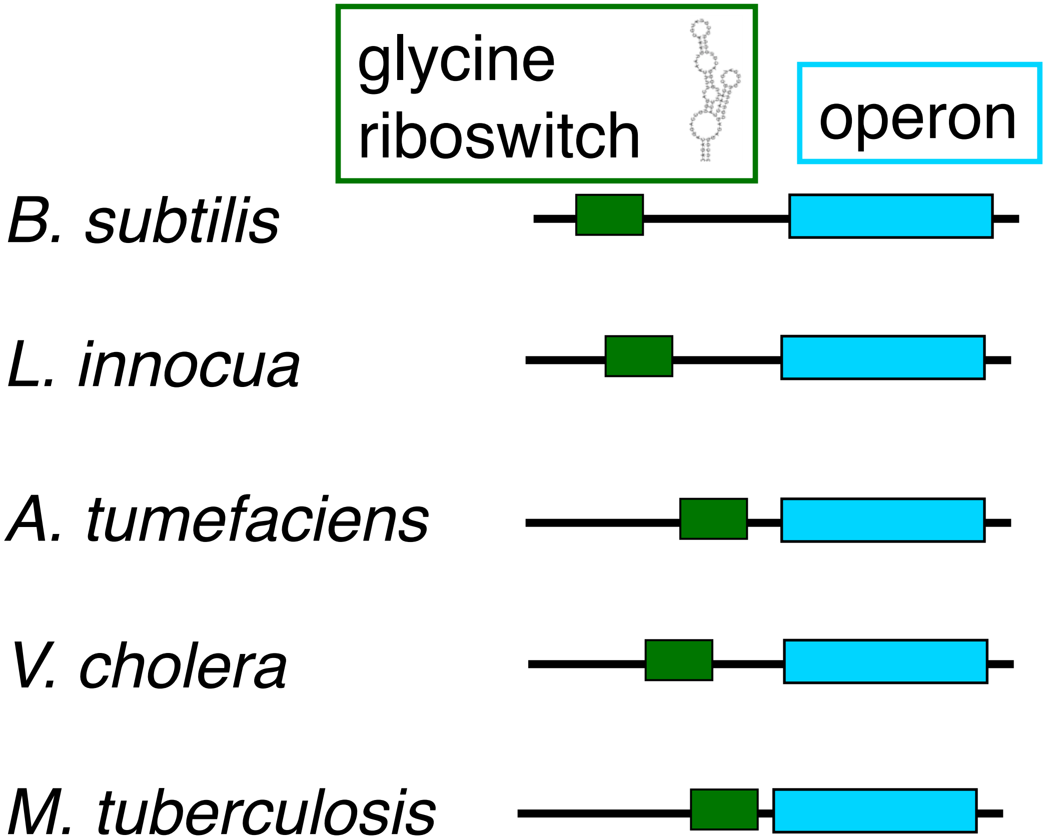
- Smith-Waterman
- FASTA
- BLAST

Sharp decline in sensitivity at ~60-70% identity

So, use structure, too

Impact of RNA homology search

(Barrick, *et al.*, 2004)



glycine riboswitch

operon

B. subtilis

L. innocua

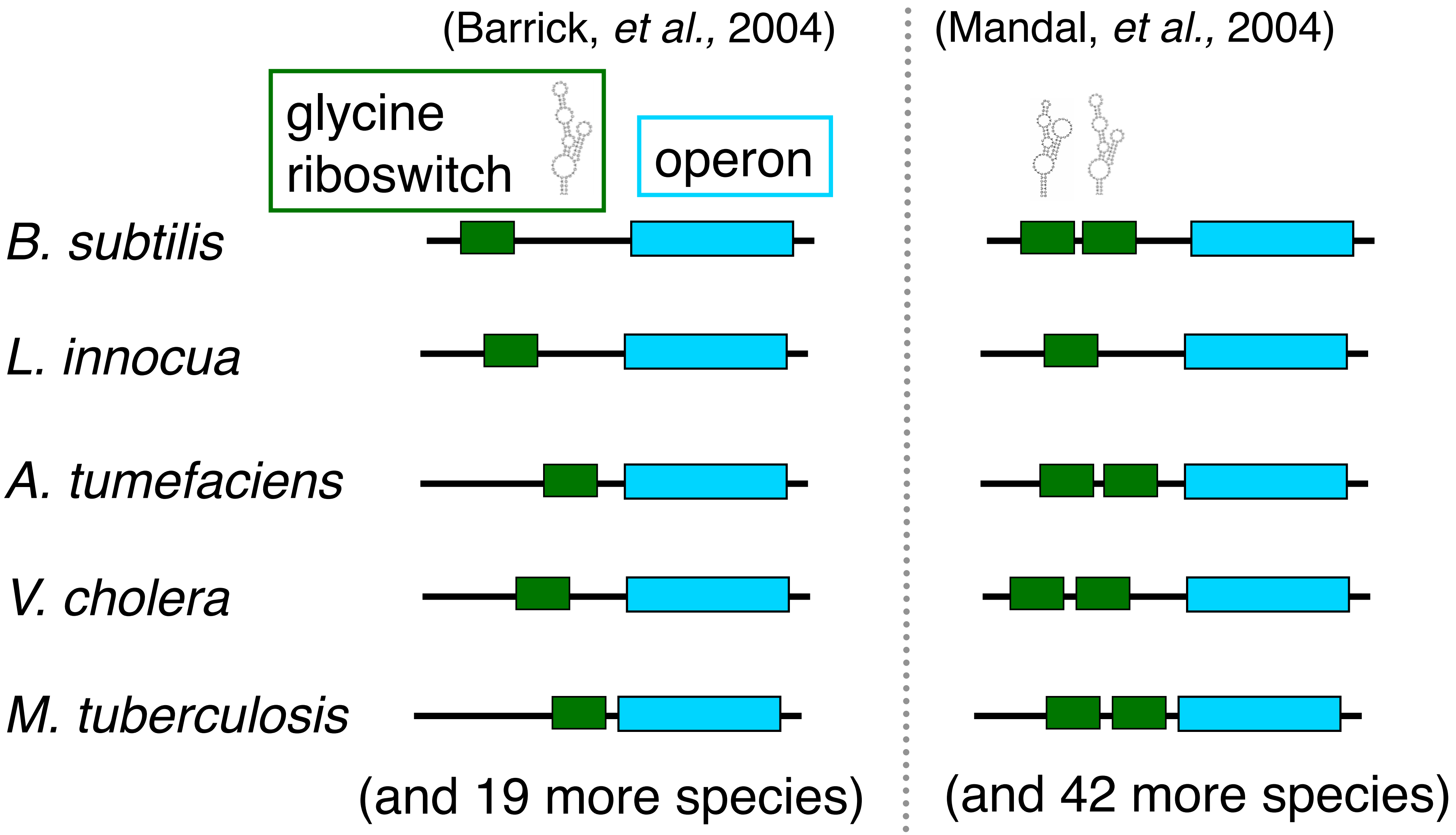
A. tumefaciens

V. cholera

M. tuberculosis

(and 19 more species)

Impact of RNA homology search



(Barrick, *et al.*, 2004)

(Mandal, *et al.*, 2004)

glycine riboswitch

operon

B. subtilis

L. innocua

A. tumefaciens

V. cholera

M. tuberculosis

(and 19 more species)

(and 42 more species)

BLAST-based

CM-based

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;

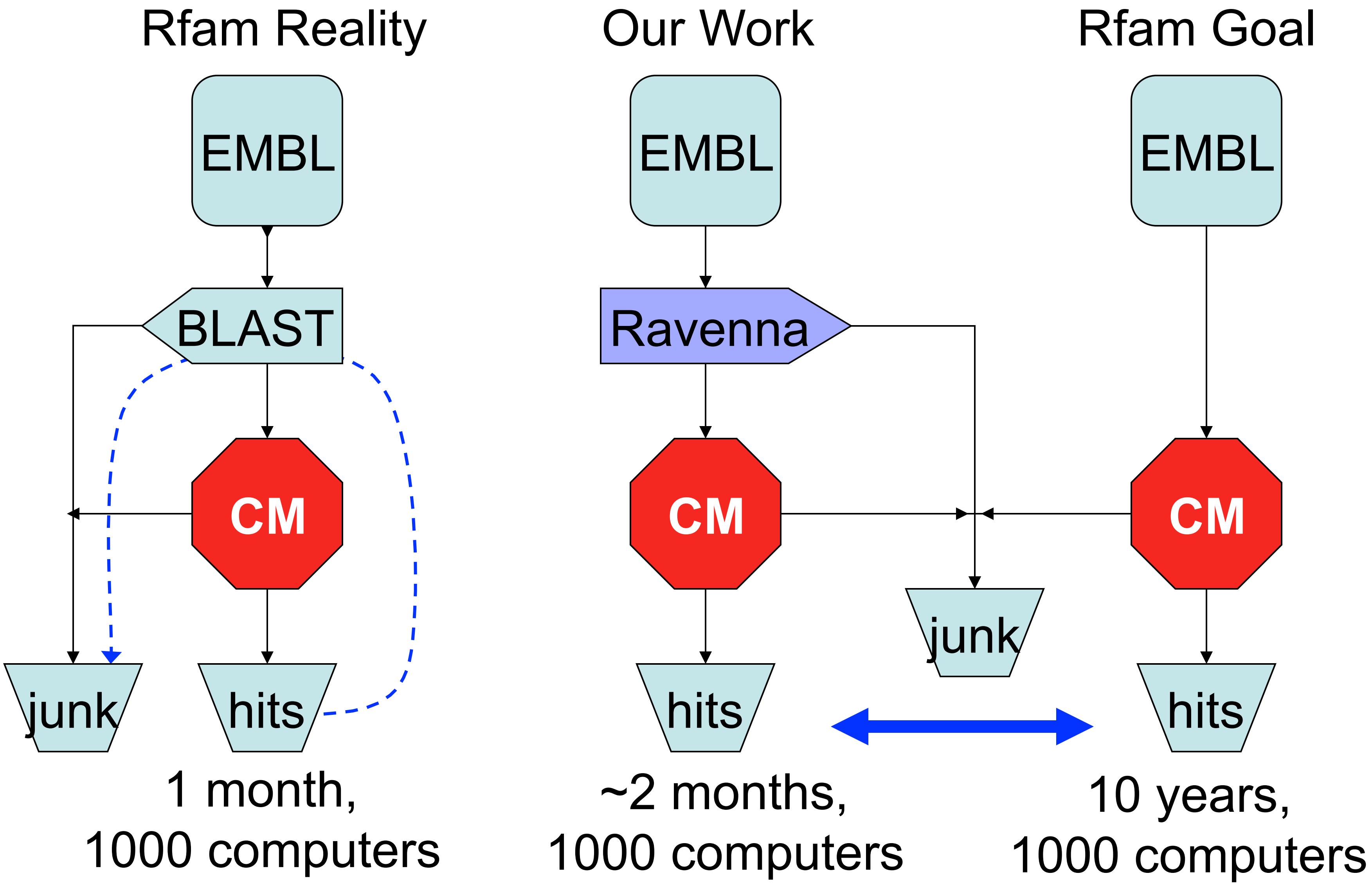
Actually run CM on the rest (the promising parts)

Assignment of HMM transition/emission scores is key

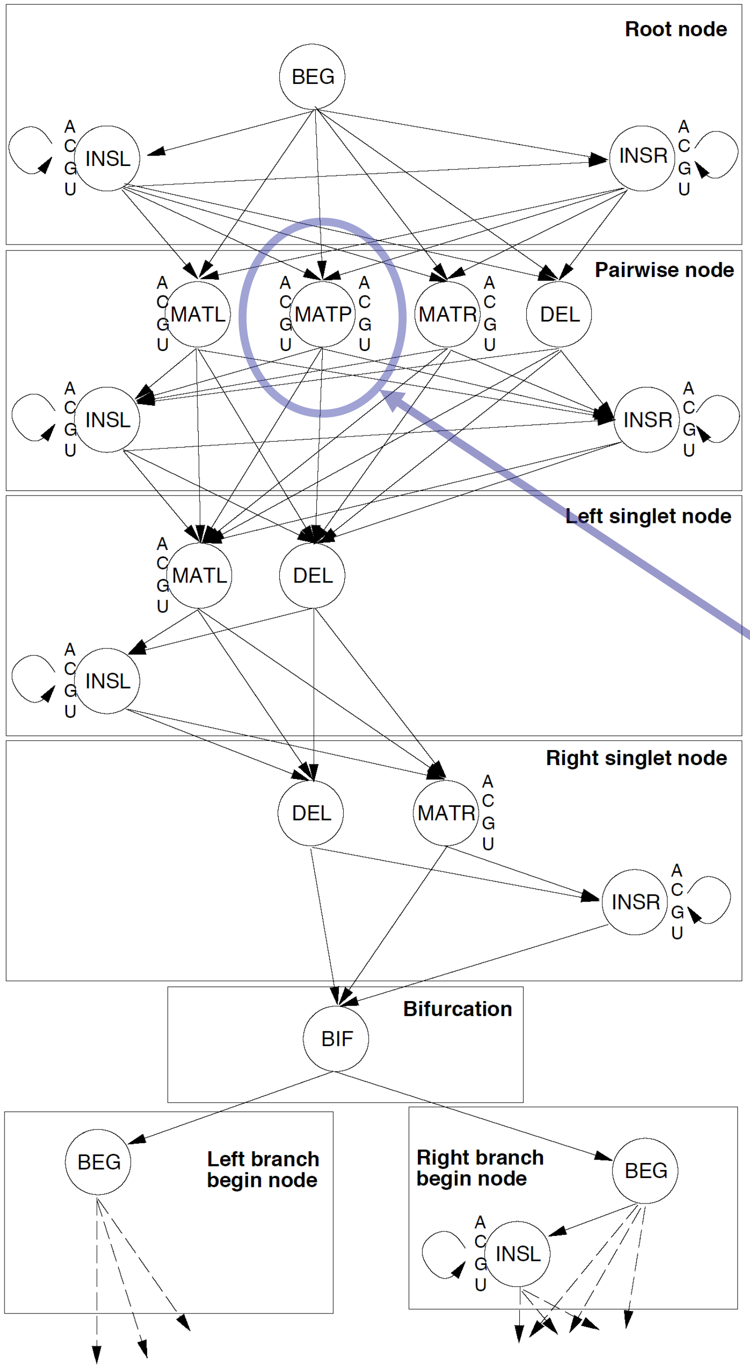
(a large convex optimization problem)

Weinberg & Ruzzo, *Bioinformatics*, 2004, 2006

CM's are good, but slow



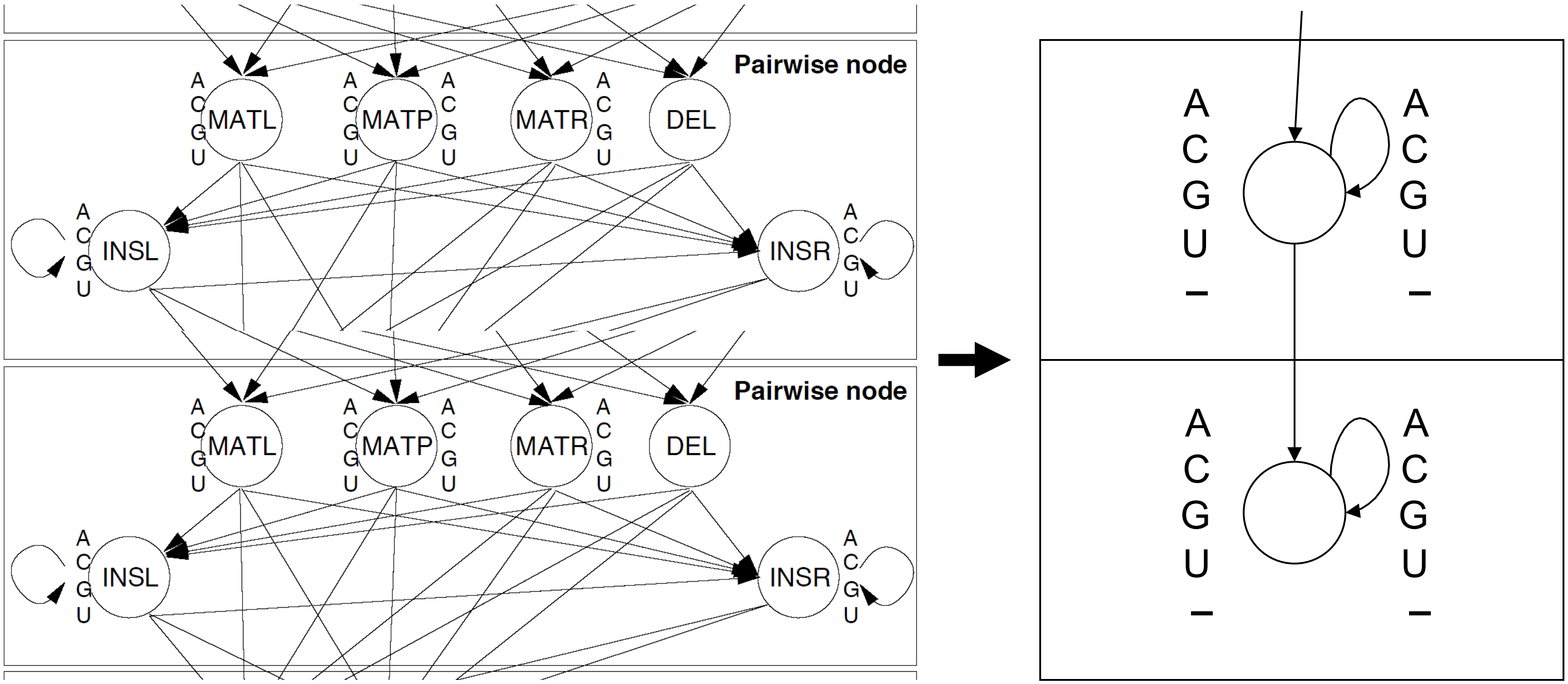
Covariance Model



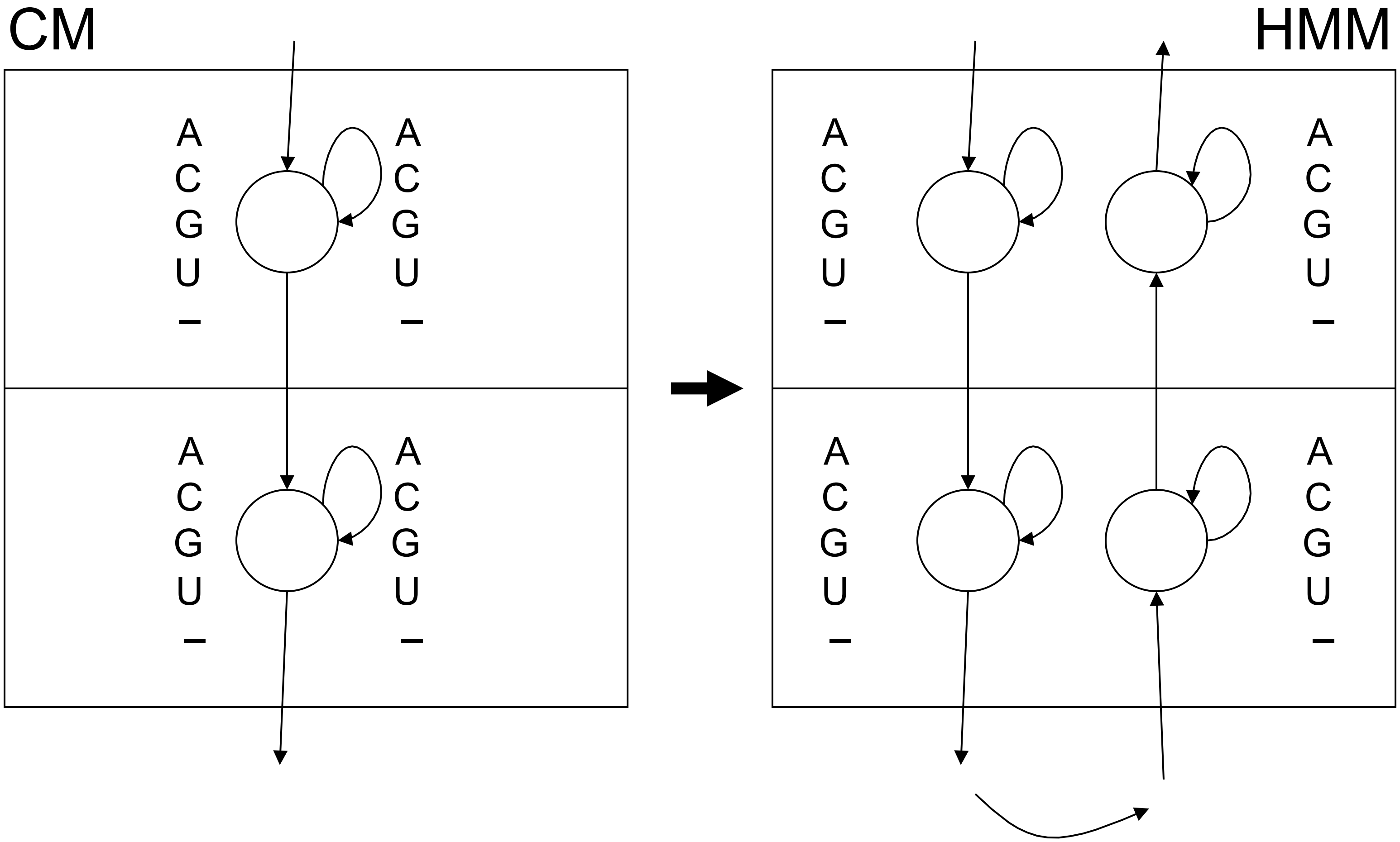
Key difference of CM vs HMM: Pair states emit paired symbols, corresponding to base-paired nucleotides; 16 emission probabilities here.

Oversimplified CM

(for pedagogical purposes only)

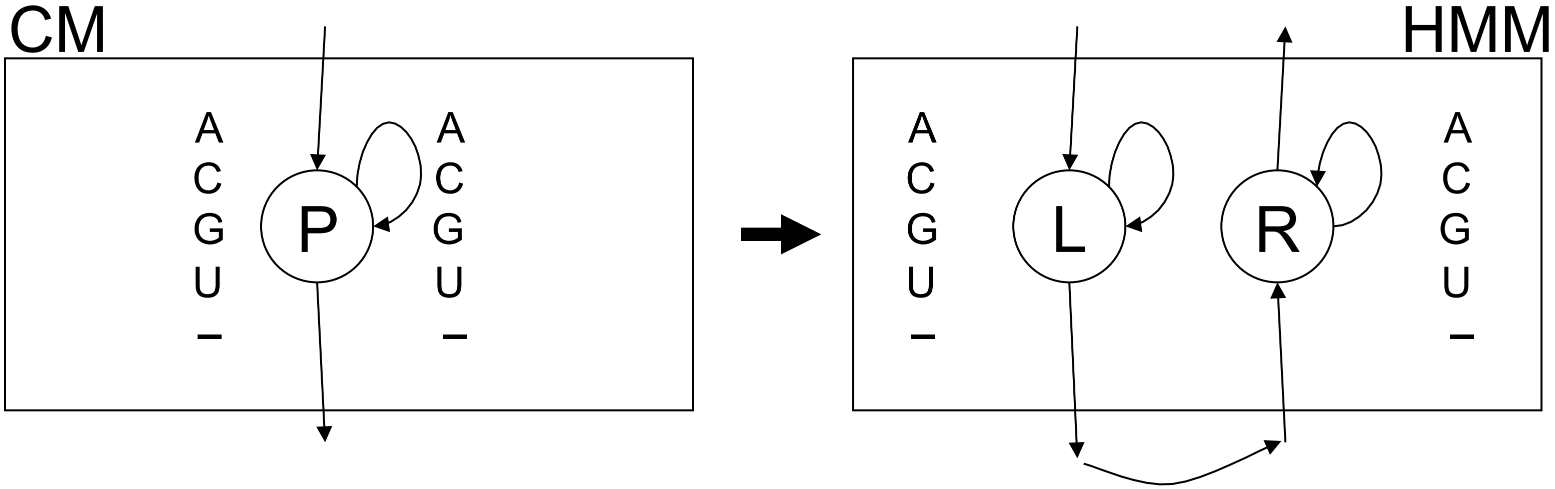


CM to HMM



25 emisions per state 5 emissions per state, 2x states

Key Issue: 25 scores → 10



Need: log Viterbi scores CM ≤ HMM

Viterbi/Forward Scoring

Path π defines transitions/emissions

Score(π) = product of "probabilities" on π

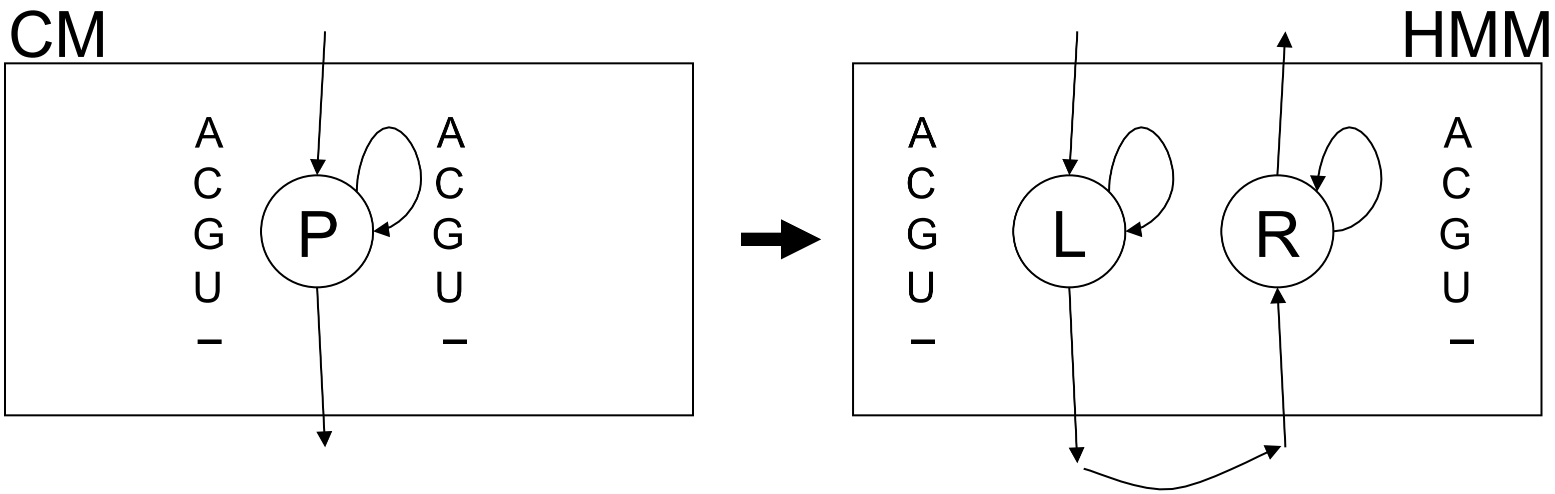
NB: ok if "probs" aren't, e.g. $\Sigma \neq 1$
(e.g. in CM, emissions are odds ratios vs
0th-order background)

For any nucleotide sequence x:

Viterbi-score(x) = max{ score(π) | π emits x}

Forward-score(x) = Σ{ score(π) | π emits x}

Key Issue: 25 scores → 10



Need: log Viterbi scores CM ≤ HMM

$P_{AA} \le L_A + R_A$	$P_{CA} \le L_C + R_A$	…
$P_{AC} \le L_A + R_C$	$P_{CC} \le L_C + R_C$	…
$P_{AG} \le L_A + R_G$	$P_{CG} \le L_C + R_G$	…
$P_{AU} \le L_A + R_U$	$P_{CU} \le L_C + R_U$	…
$P_{A-} \le L_A + R_-$	$P_{C-} \le L_C + R_-$	…

NB: HMM not a prob. model

Rigorous Filtering

$P_{AA} \le L_A + R_A$
$P_{AC} \le L_A + R_C$
$P_{AG} \le L_A + R_G$
$P_{AU} \le L_A + R_U$
$P_{A-} \le L_A + R_-$
…

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)

Some scores filter better

$P_{UA} = 1 \le L_U + R_A$

$P_{UG} = 4 \le L_U + R_G$

Option 1:

$L_U = R_A = R_G = 2$

Option 2:

$L_U = 0,\ R_A = 1,\ R_G = 4$

Assuming ACGU ≈ 25%

Opt 1:

$L_U + (R_A + R_G)/2 = 4$

Opt 2:

$L_U + (R_A + R_G)/2 = 2.5$

Optimizing filtering

For any nucleotide sequence x:

Viterbi-score(x) = max{ score(π) | π emits x }

Forward-score(x) = Σ{ score(π) | π emits x }

Expected Forward Score

$E(L_i, R_i) = \Sigma_{\text{all sequences } x}$ Forward-score(x)*Pr(x)

Under 0th-order background model

NB: E is a function of L_i, R_i only

Optimization:

Minimize $E(L_i, R_i)$ subject to score Lin.Ineq.s

This is heuristic ("forward↓ ⇒ Viterbi↓ ⇒ filter↓")

But still rigorous because "subject to score Lin.Ineq.s"

Calculating $E(L_i, R_i)$

$E(L_i, R_i) = \Sigma_x$ Forward-score(x)*Pr(x)

Forward-like: for every state, calculate expected score for all paths ending there; easily calculated from expected scores of predecessors & transition/emission probabilities/scores

Minimizing E(L$_i$, R$_i$)

Calculate E(L$_i$, R$_i$) *symbolically*, in terms of emission scores, so we can do partial derivatives for numerical convex optimization algorithm

Forward:

$$f_k(i) = P(x_1 \ldots x_i,\ \pi_i = k)$$

$$f_l(i+1) = e_l(x_{i+1}) \sum_k f_k(i) a_{k,l}$$

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 probabilities

Convex optimization problem

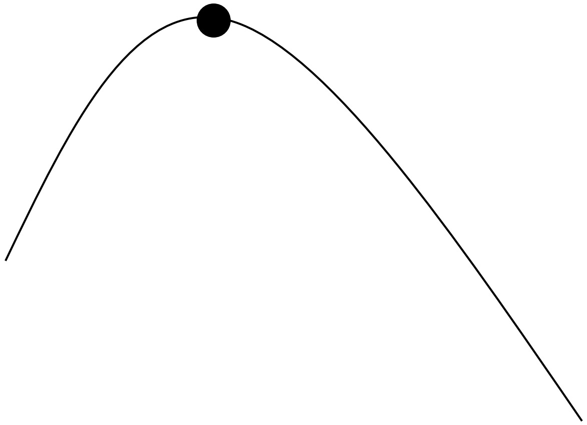
- Constraints: enforce rigorous property
- Objective function: filter as aggressively as possible

Problem sizes:

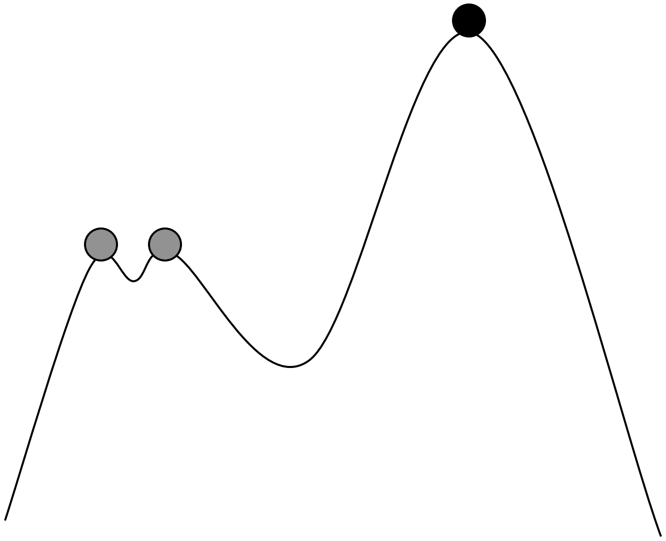
- 1000-10000 variables
- 10000-100000 inequality constraints

“Convex” Optimization

Convex:
local max = global max;
simple “hill climbing” works



Nonconvex:
can be many local maxima,
$\ll$ global max;
“hill-climbing” fails



Estimated Filtering Efficiency

(139 Rfam 4.0 families)

Filtering fraction	# families (compact)	# families (expanded)
$< 10^{-4}$	105	110
10^{-4} - 10^{-2}	8	17
.01 - .10	11	3
.10 - .25	2	2
.25 - .99	6	4
.99 - 1.0	7	3

≈ break even

~100x speedup

Averages 283 times faster than CM

Results: new ncRNAs (?)

Name	# Known (BLAST + CM)	# New (rigorous filter + CM)
Pyrococcus snoRNA	57	123
Iron response element	201	121
Histone 3' element	1004	102*
Retron msr	11	48
Hammerhead I	167	26
Hammerhead III	251	13
U6 snRNA	1462	2
U7 snRNA	312	1
cobalamin riboswitch	170	7

13 other families	5-1107	0

Results: With additional work

	# with BLAST+CM	# with rigorous filter series + CM	# new
Rfam tRNA	58609	63767	5158
Group II intron	5708	6039	331
tRNAscan-SE (human)	608	729	121
tmRNA	226	247	21
Lysine riboswitch	60	71	11
And more…			

“Additional work”

Profile HMM filters use *no* 2^{ary} structure info

- They work well because, tho structure can be critical to function, there is (usually) enough primary sequence conservation to exclude most of DB
- But not on all families (and may get worse?)

Can we exploit *some* structure (quickly)?

- Idea 1: “sub-CM”
- Idea 2: extra HMM states remember mate

} for some hairpins

- Idea 3: try lots of combinations of “some hairpins”
- Idea 4: chain together several filters (select via Dijkstra)

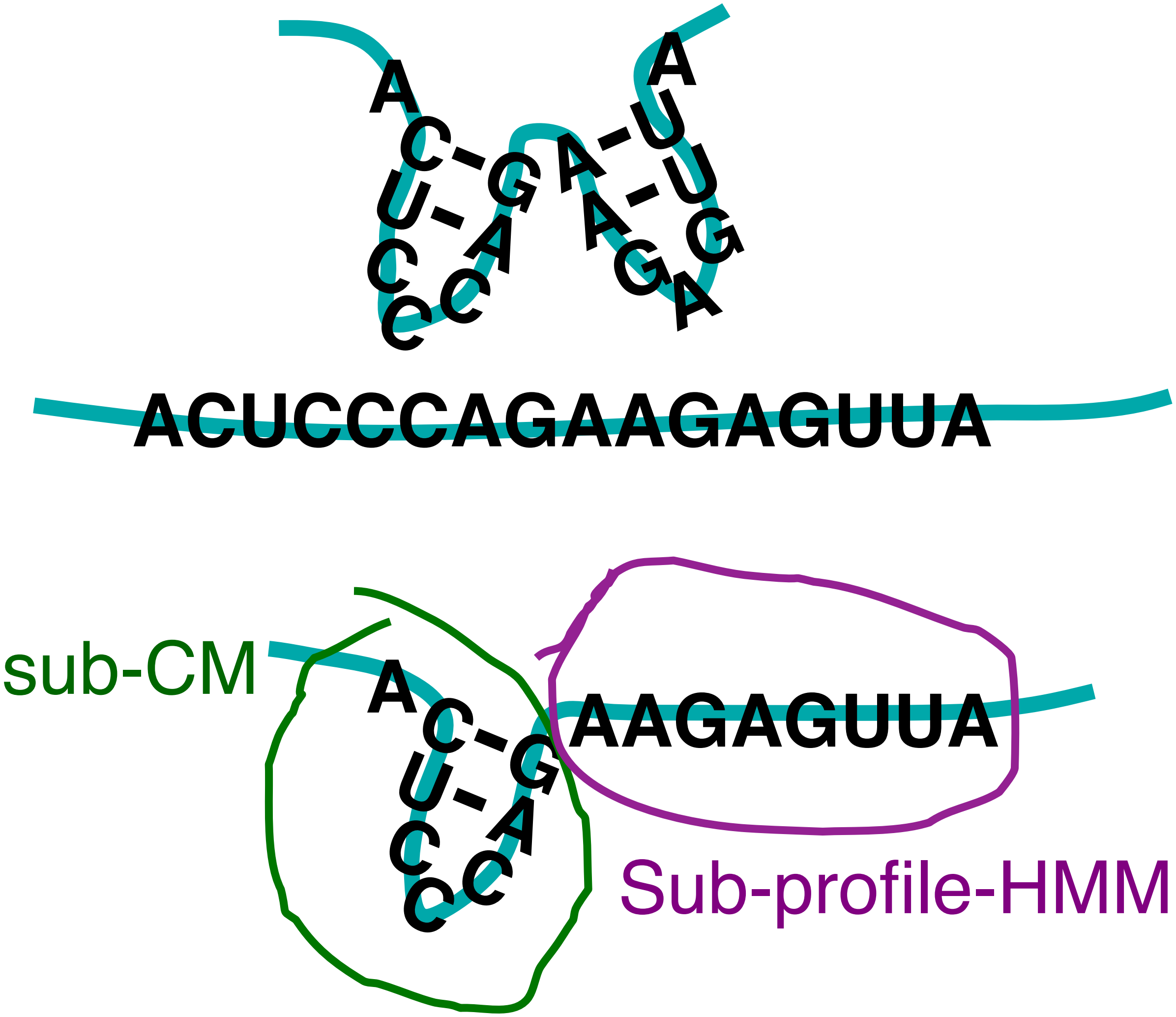
Sub-CM filters

Full CM

Profile HMM

A sub-CM

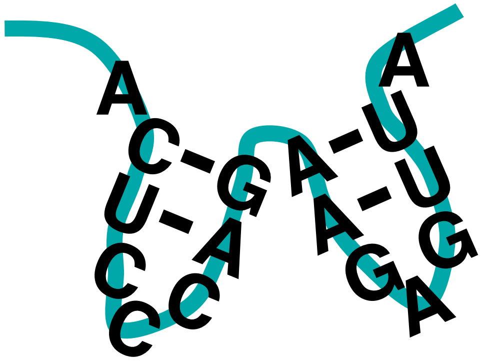
Sub-CM



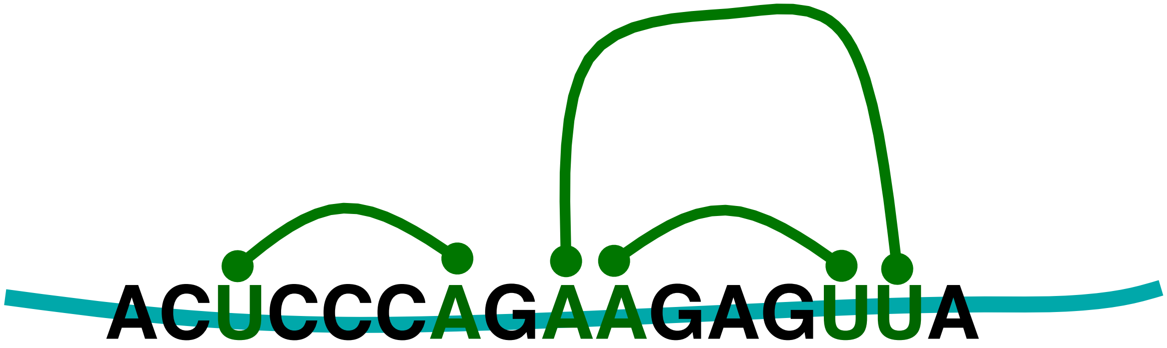
Sub-profile-HMM

Store-pair filters

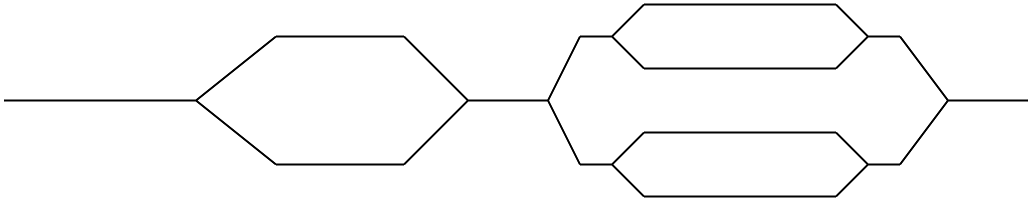
Full CM



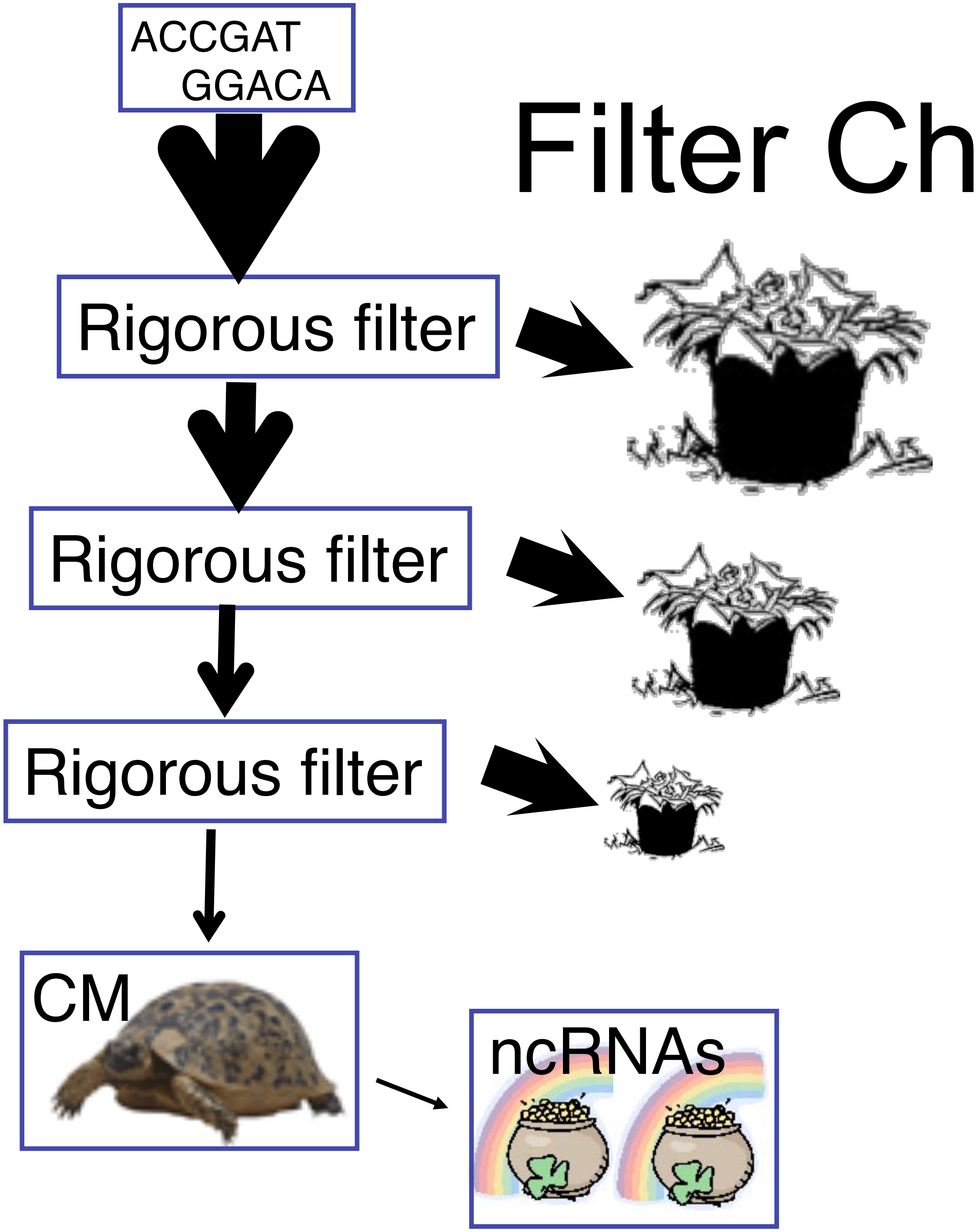
Store pair



“Profile” HMM:



Filter Chains



ACCGAT
GGACA

Rigorous filter

Rigorous filter

Rigorous filter

CM

ncRNAs

Why run filters in series?

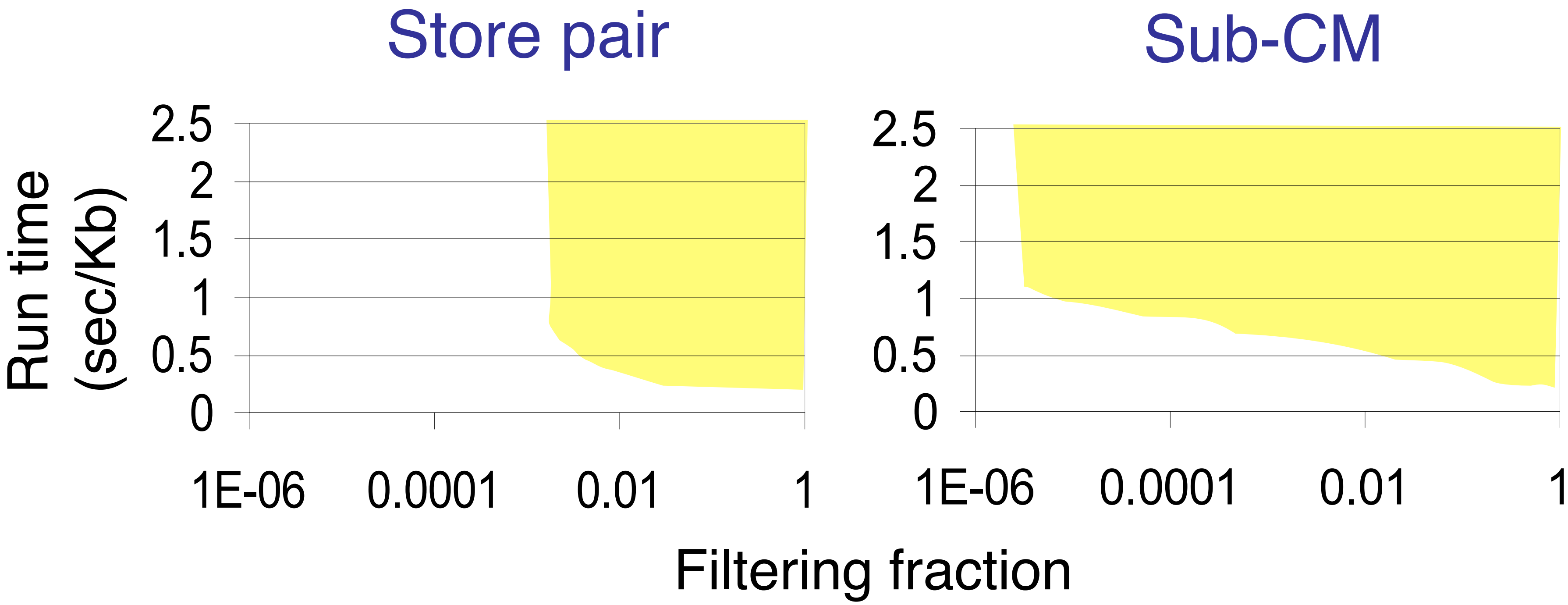
	Filtering fraction	Run time (sec/Kbase)
Filter 1	0.25	1
Filter 2	0.01	10
CM	N/A	200

CM alone: 200 s/Kb

Filter 1 → CM: 1 + 0.25*200 = 51 s/Kb

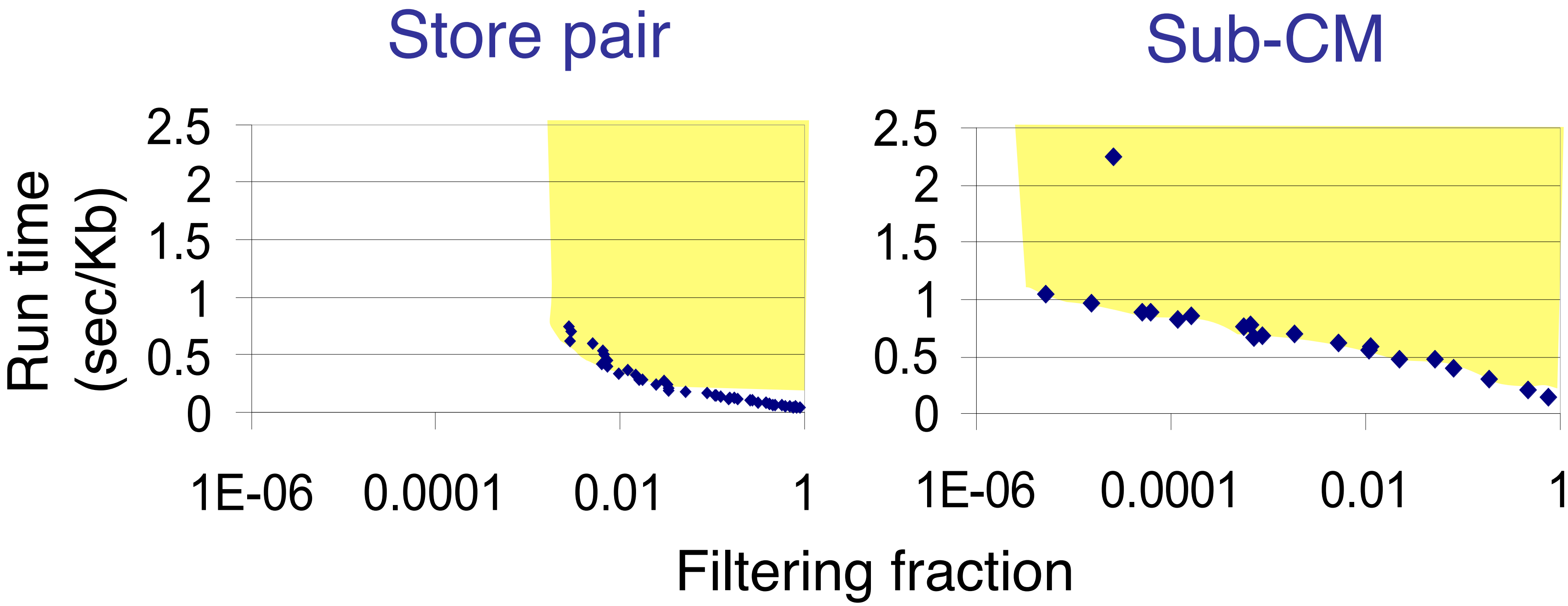
Filter 2 → CM: 10 + 0.01*200 = 12 s/Kb

Filter 1 → Filter 2 → CM:
1 + 0.25*10 + 0.01*200 = 5.5 s/Kb



Properties of a filter:

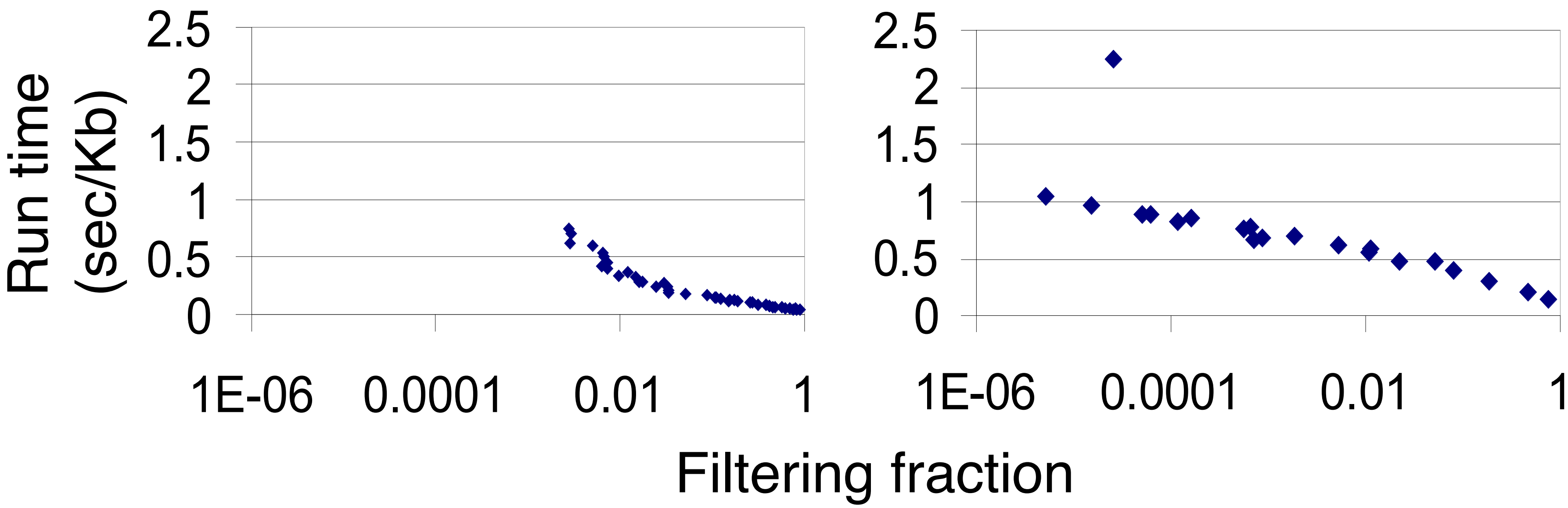
- Filtering fraction
- Run time (sec/Kb)



Simplified performance model (selectivity & speed)

Independence assumptions for base pairs

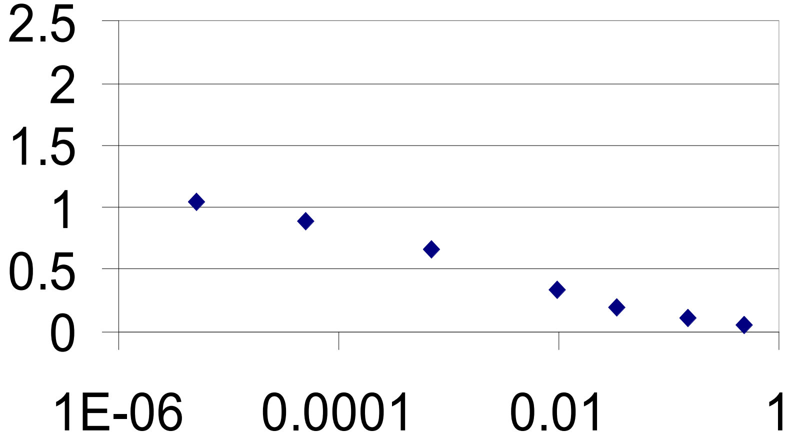
Use dynamic programming to rapidly explore base pair combinations



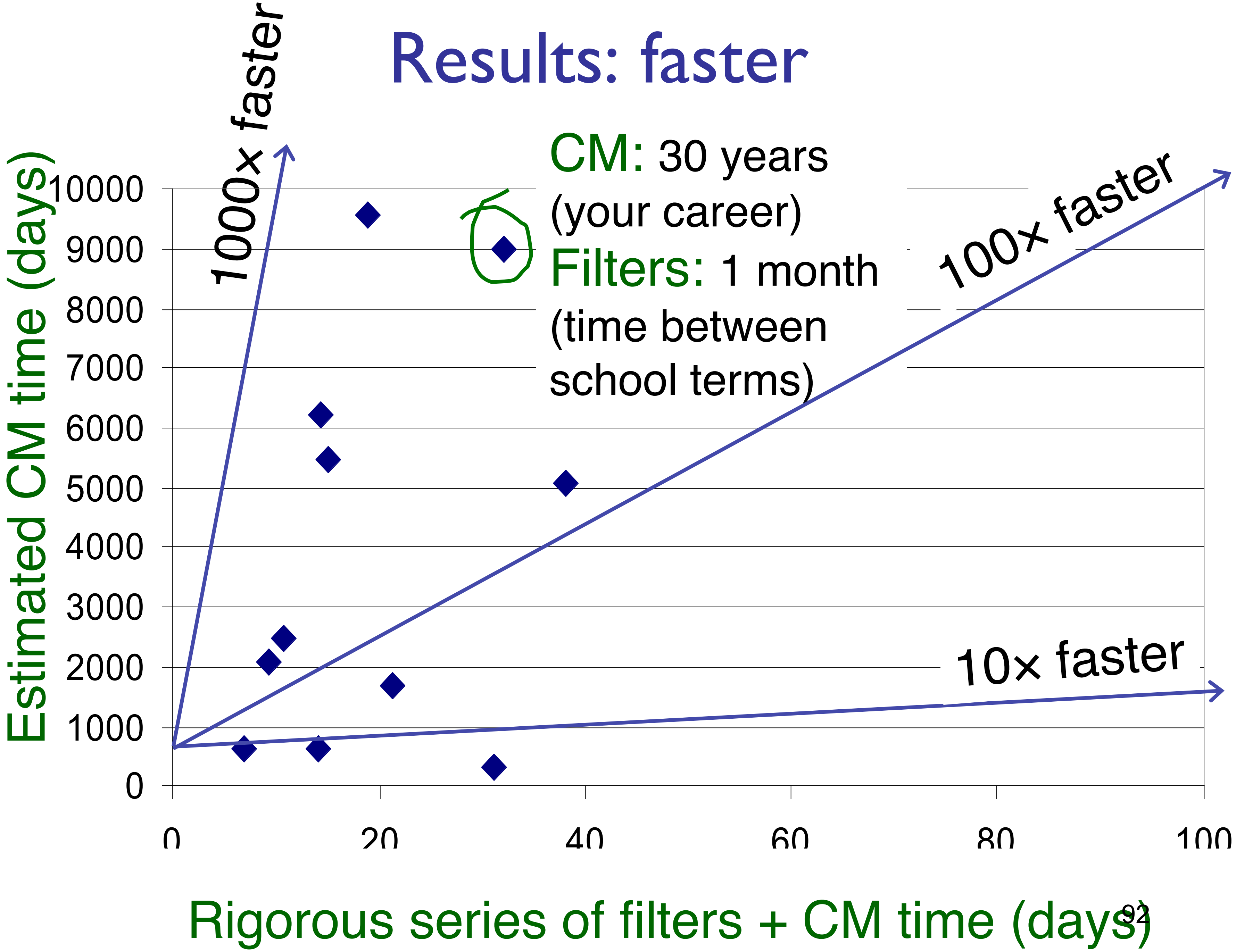
Store pair

Sub-CM

Selected rigorous filter chain



Results: faster



CM: 30 years (your career)

Filters: 1 month (time between school terms)

Results: more sensitive than BLAST

	# with BLAST+CM	# with rigorous filters + CM	# new
Rfam tRNA	58609	63767	5158
Group II intron	5708	6039	331
Iron response element	201	322	121
tmRNA	226	247	21
Lysine riboswitch	60	71	11
And more…			

Is there anything more to do?

Rigorous filters can be too cautious

- E.g., 10 times slower than heuristic filters
- Yet only 1-3% more sensitive

We want to

- Run scans faster with minimal loss of sensitivity
- Know empirically what sensitivity we're losing

Heuristic Filters

Rigorous filters optimized for worst case

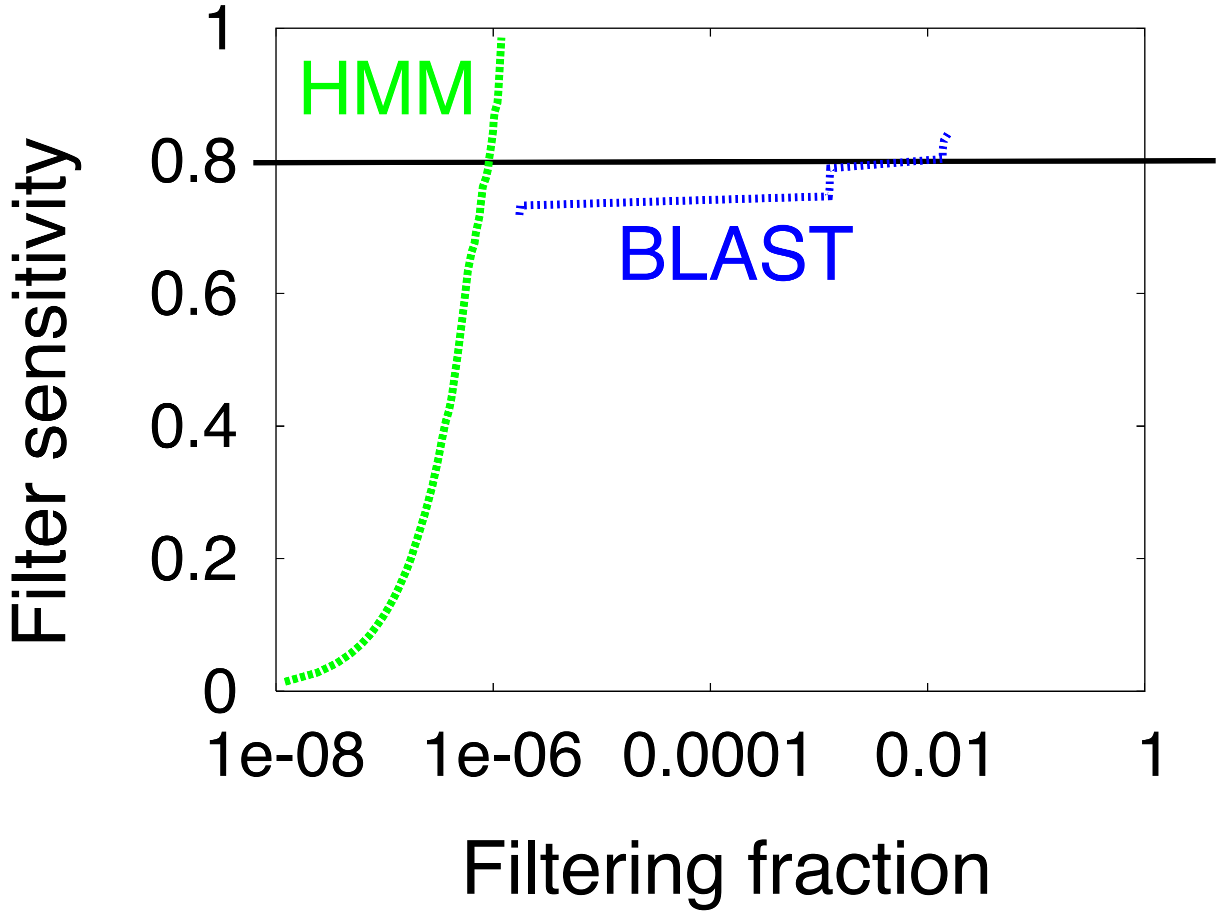
Possible to trade improved speed for small loss in sensitivity?

Yes – profile HMMs as before, but optimized for average case

“ML heuristic”: train HMM from the infinite alignment generated by the CM

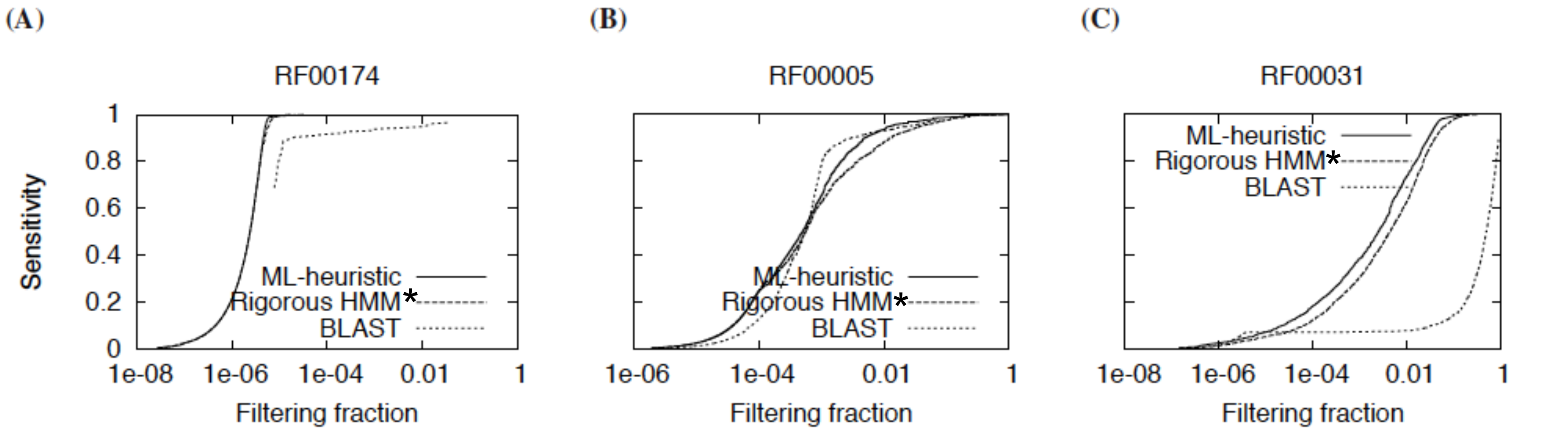
Often 10x faster, modest loss in sensitivity

Heuristic Filters
ROC-like curves
(lysine riboswitch)



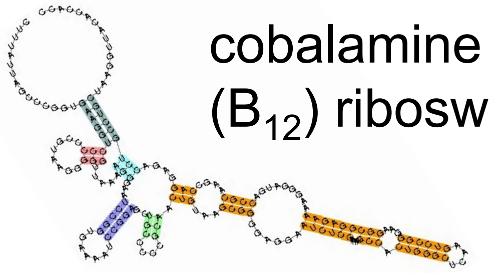
Filter sends 80% of hits to CM

Heuristic Filters

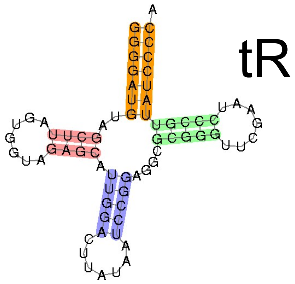


* rigorous HMM, not rigorous threshold

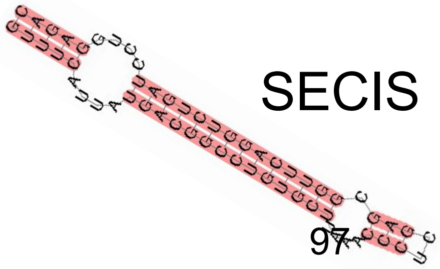
Fig. 1. Selected ROC-like curves. All plot sensitivity against filtering fraction, with filtering fraction in log scale. (A) RF00174 is typical of the other families; the ML-heuristic is slightly better than the rigorous profile HMM, and both often dramatically exceed BLAST. (B) Atypically, in RF00005, BLAST is superior, although only in one region. (C) BLAST performs especially poorly for RF00031. (Recall that rigorous scans were not possible for RF00031, so only ∼90% of hits are known; see text.) The supplement includes all ROC-like curves, and the inferior ignore-SS.



cobalamine (B_{12}) riboswitch



tRNA



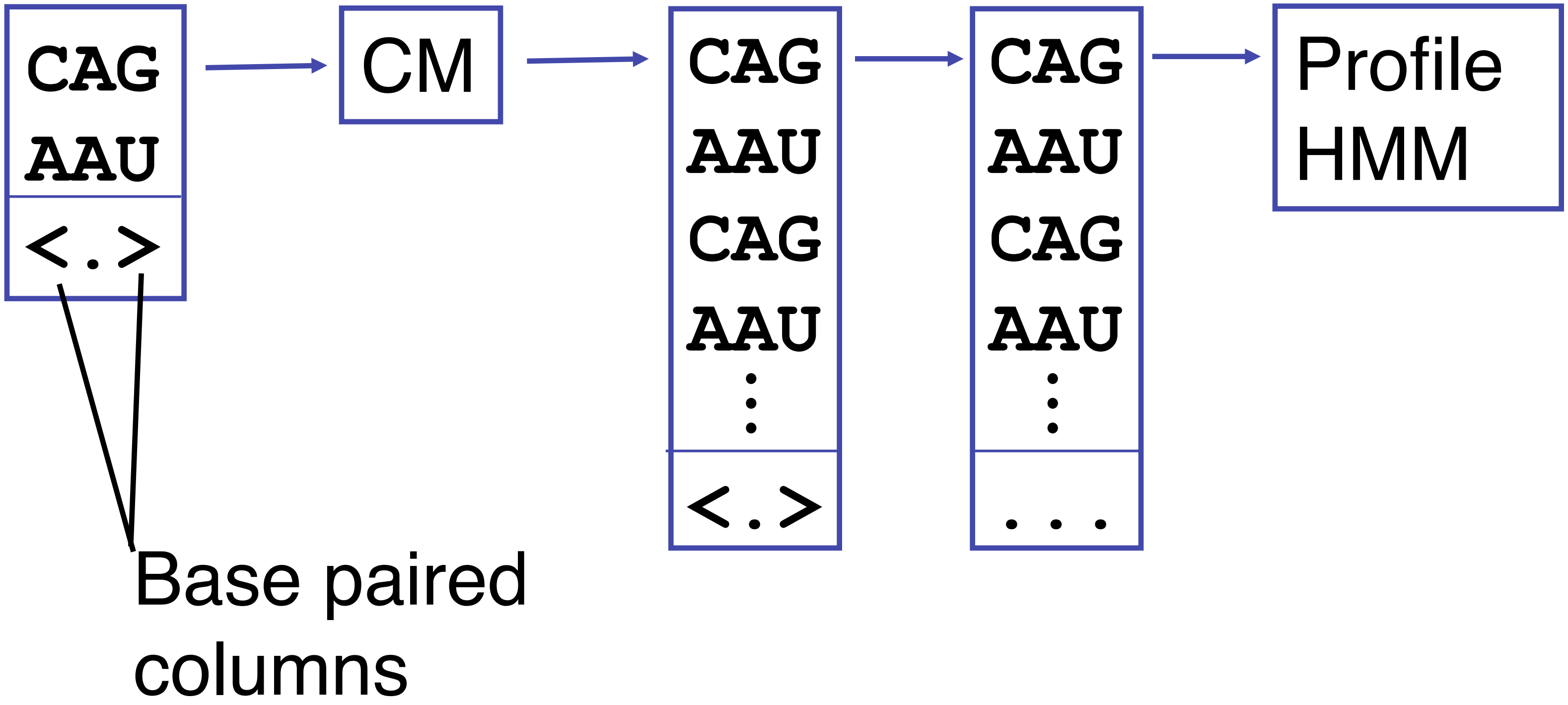
SECIS

Heuristic Profile HMMs

(**Weinberg** & Ruzzo, 2006)

Input Multiple Sequence Alignment

Infinite Multiple sequence alignments



```
CAG
AAU
<.>
```

CM

```
CAG
AAU
CAG
AAU
 ⋮
<.>
```

```
CAG
AAU
CAG
AAU
 ⋮
. . .
```

Profile HMM

Base paired columns

Software

Ravenna implements both rigorous and heuristic filters

Infernal (engine behind Rfam, for example) implements heuristic filters and some other accelerations

- E,g., dynamic “banding” of dynamic programming matrix based on the insight that large deviations from consensus length must have low scores.

CM Search Summary

Still slower than we might like, but dramatic speedup over raw CM is possible with:

- No loss in sensitivity (provably), or
- Even faster with modest (and estimable) loss in sensitivity

Day 3

Our Plot So Far:

Covariance Models (CMs) represent conserved RNA sequence/structure motifs

They allow accurate search

Basic search is slow, but substantial speedup possible

Today:

Automated model construction & ncRNA discovery in prokaryotes

Motif Discovery

RNA Motif Discovery

- CM's are great, but where do they come from?
- An approach: comparative genomics
  - 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.

RNA Motif Discovery

Typical problem: given a 10-20 unaligned sequences of 1-10kb, 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

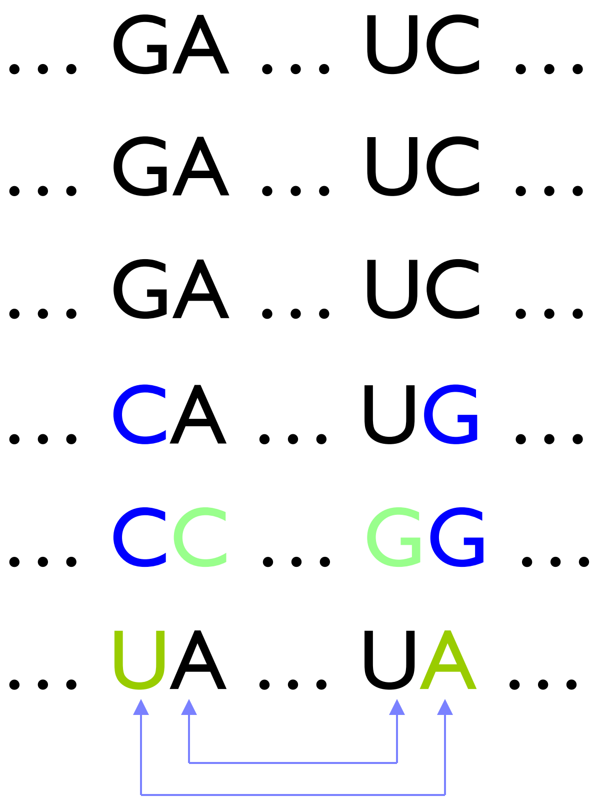
Approaches

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

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

Joint: Do both together

“Align First” Approach: Predict Struct from Multiple Alignment



… GA … UC …

… GA … UC …

… GA … UC …

… CA … UG …

… CC … GG …

… UA … UA …

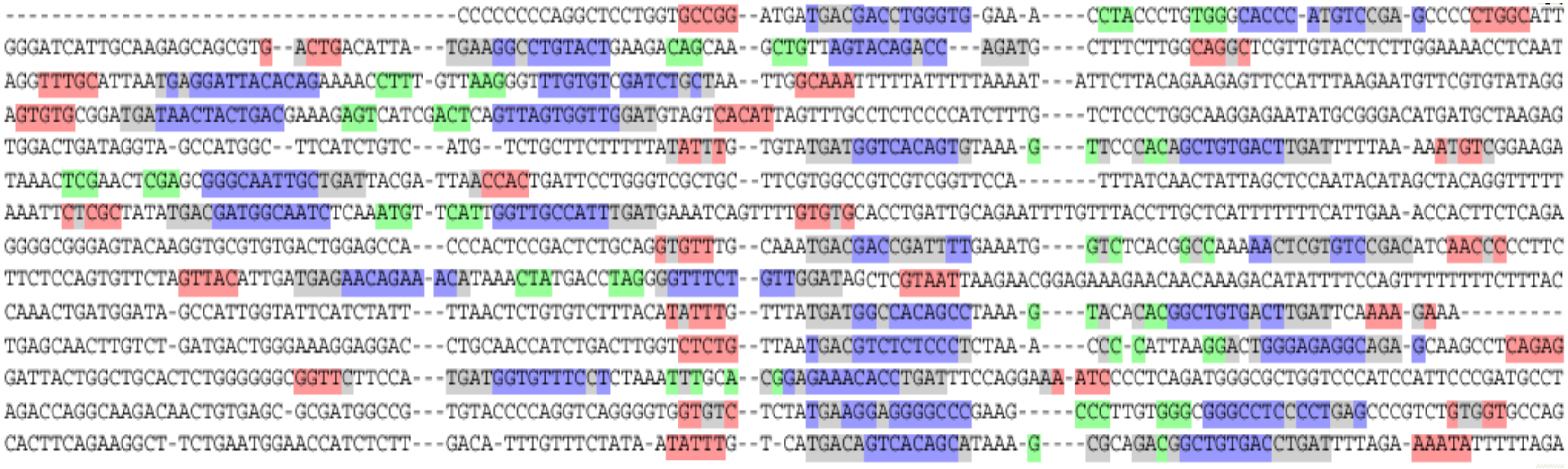
Compensatory mutations reveal structure (core of “comparative sequence analysis”) *but* usual alignment algorithms penalize them (twice)

Pitfall for sequence-alignment-first approach

Structural conservation ≠ Sequence conservation

Alignment without structure information is unreliable

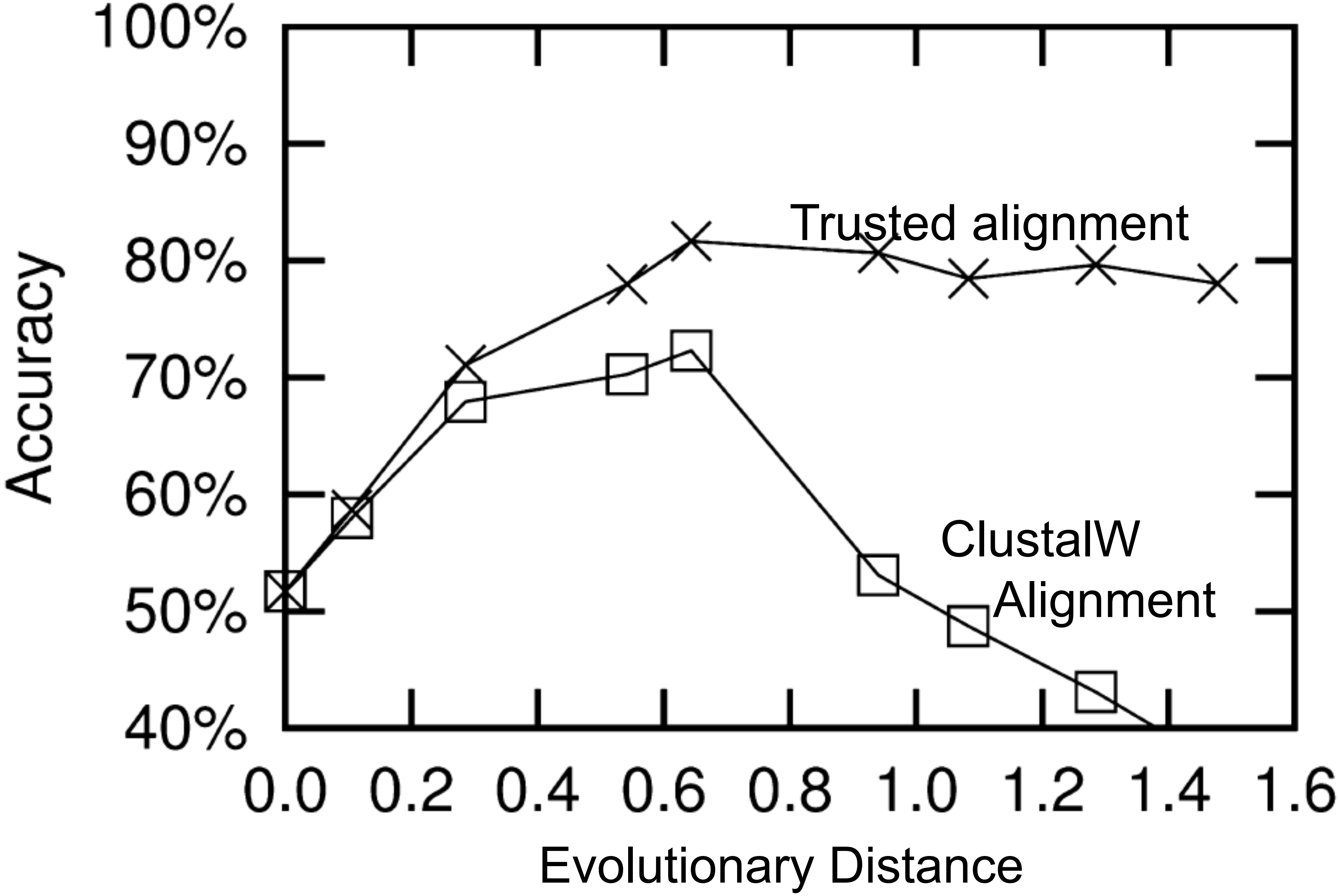
CLUSTALW alignment of SECIS elements with flanking regions



```
---------------------------------------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-ACATAAACTATGACCTAGGGTTTCT--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
```

same-colored boxes *should* be aligned

Pfold (KH03) Test Set D



Knudsen & Hein, Pfold: RNA secondary structure prediction using stochastic context-free grammars, Nucleic Acids Research, 2003, v 31,3423–3428

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 biologically-validated model for structural alignment

Joint: Do both together

- Sankoff – good but slow
- Heuristic

Our Approach: CMfinder

RNA motifs from unaligned sequences

Simultaneous *local* alignment, folding and CM-based motif description via an EM-style learning procedure

- Sequence conservation exploited, but not required
- Robust to inclusion of unrelated and/or flanking sequence
- Reasonably fast and scalable
- Produces a probabilistic model of the motif that can be directly used for homolog search

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006

Alignment → CM → Alignment

Similar to HMM, but slower

Builds on Eddy & Durbin, ‘94

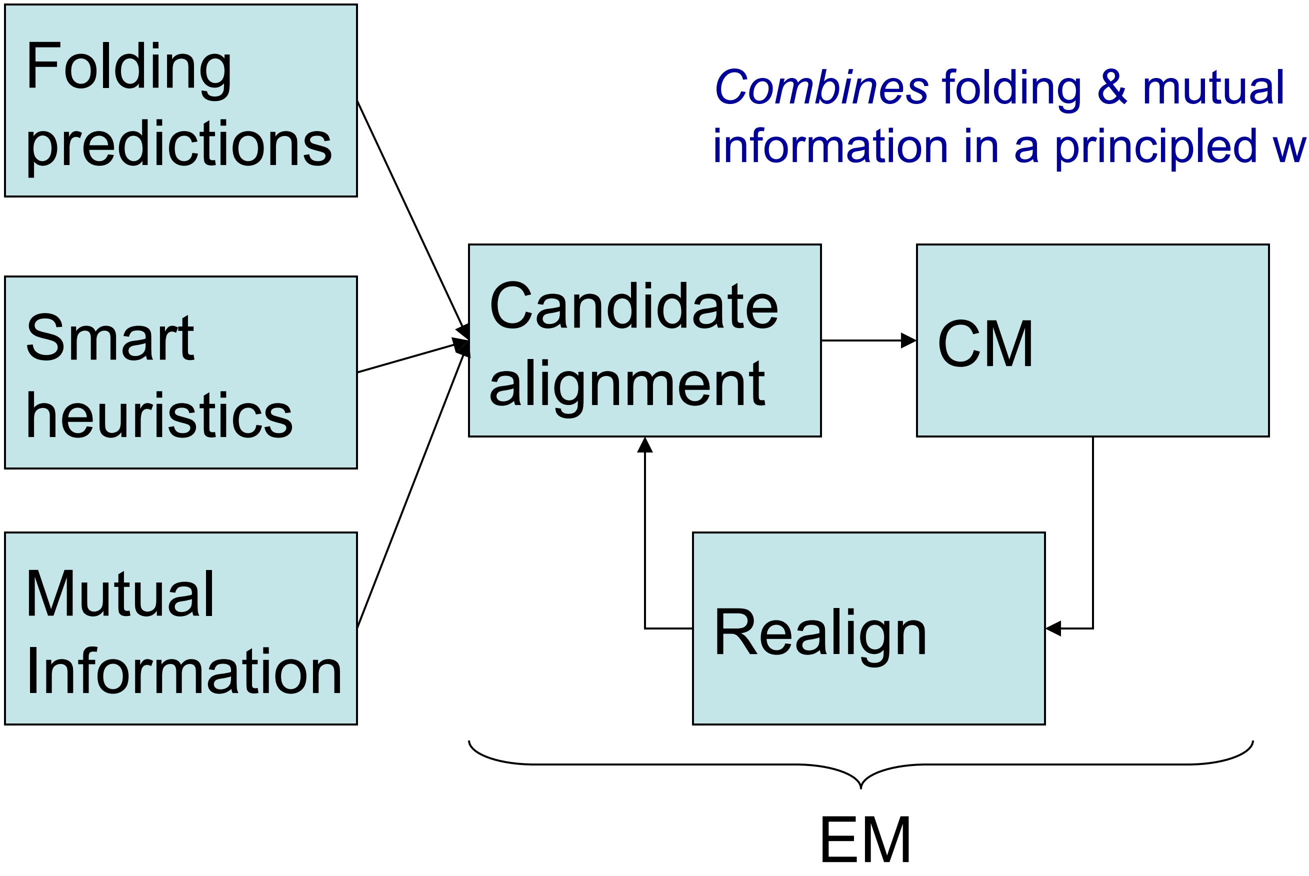
But new way to infer which columns to pair, via a principled combination of mutual information and predicted folding energy

And, it’s local, not global, alignment (harder)

CMFinder

Simultaneous alignment, folding & motif description

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006



Combines folding & mutual information in a principled way.

Initial Alignment Heuristics

fold sequences separately

candidates: regions with low folding energy

compare candidates via “tree edit” algorithm

find best “central” candidates & align to them

BLAST anchors

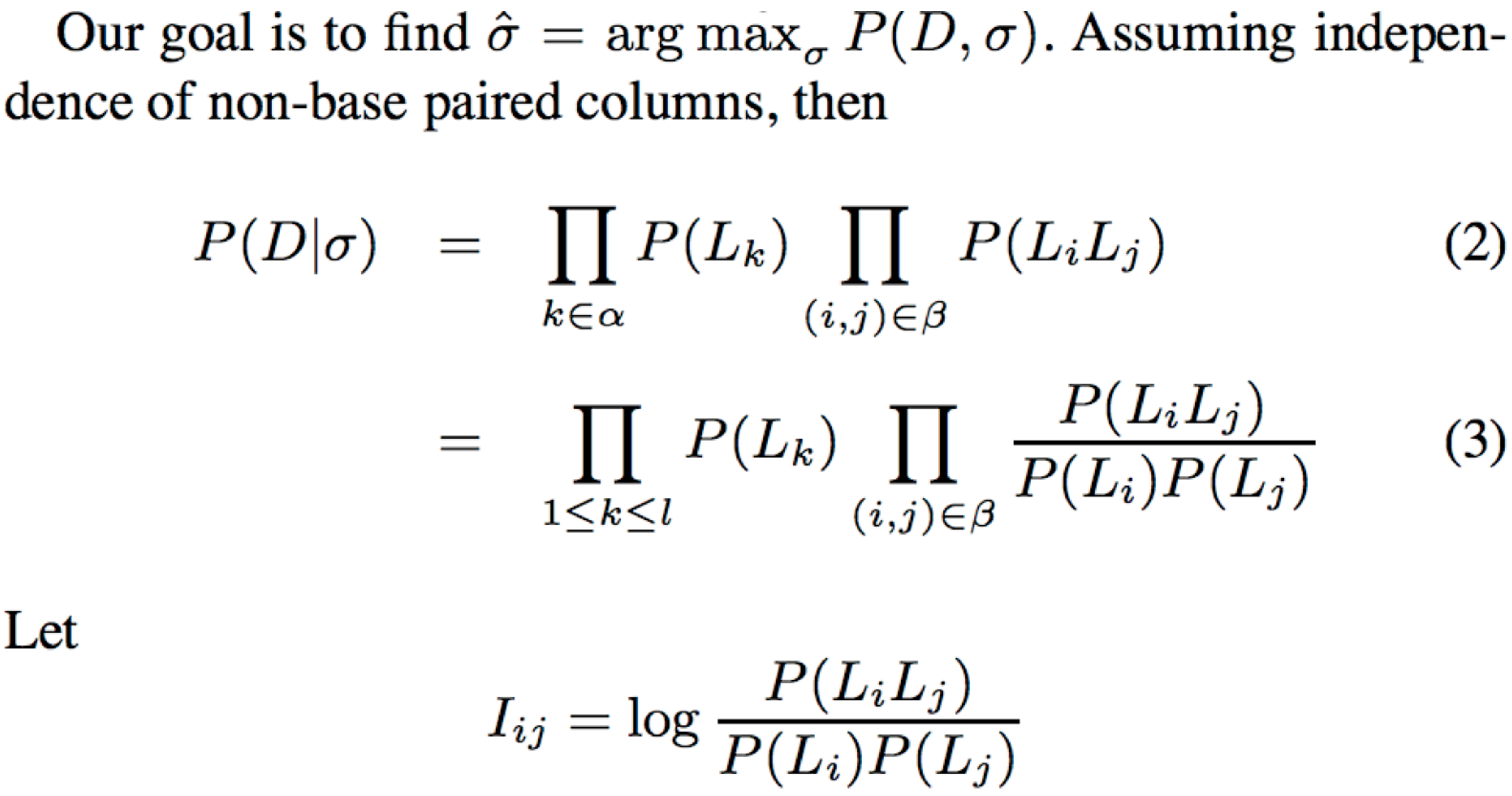
Structure Inference

Part of M-step is to pick a structure that maximizes data likelihood

We combine:

- mutual information
- position-specific priors for paired/unpaired
  - (based on single sequence thermodynamic folding predictions)
- intuition: for similar seqs, little MI; fall back on single-sequence folding predictions
- data-dependent, so not strictly Bayesian

L_i = column i; $\sigma = (\alpha, \beta)$ the 2^{ary} struct, α = unpaired, β = paired cols



Our goal is to find $\hat{\sigma} = \arg\max_\sigma P(D, \sigma)$. Assuming independence of non-base paired columns, then

$$P(D|\sigma) = \prod_{k \in \alpha} P(L_k) \prod_{(i,j) \in \beta} P(L_i L_j) \tag{2}$$

$$= \prod_{1 \leq k \leq l} P(L_k) \prod_{(i,j) \in \beta} \frac{P(L_i L_j)}{P(L_i) P(L_j)} \tag{3}$$

Let

$$I_{ij} = \log \frac{P(L_i L_j)}{P(L_i) P(L_j)}$$

With MLE params, I_{ij} is the *mutual information* between cols i and j

Let s_i be the prior for column i to be single stranded, and p_{ij} the prior for columns i,j to be base paired, then $P(\sigma) = \prod_{k\in\alpha} s_k \prod_{(i,j)\in\beta} p_{ij}$, and $P(D,\sigma)$ can be rewritten as

$$\begin{aligned} P(D,\sigma) &= P(D|\sigma)P(\sigma) \\ &= \prod_{1\leq k\leq l} P(L_k)s_k \prod_{(i,j)\in\beta} \frac{P(L_iL_j)}{P(L_i)P(L_j)} \frac{p_{ij}}{s_is_j} \end{aligned} \tag{4}$$

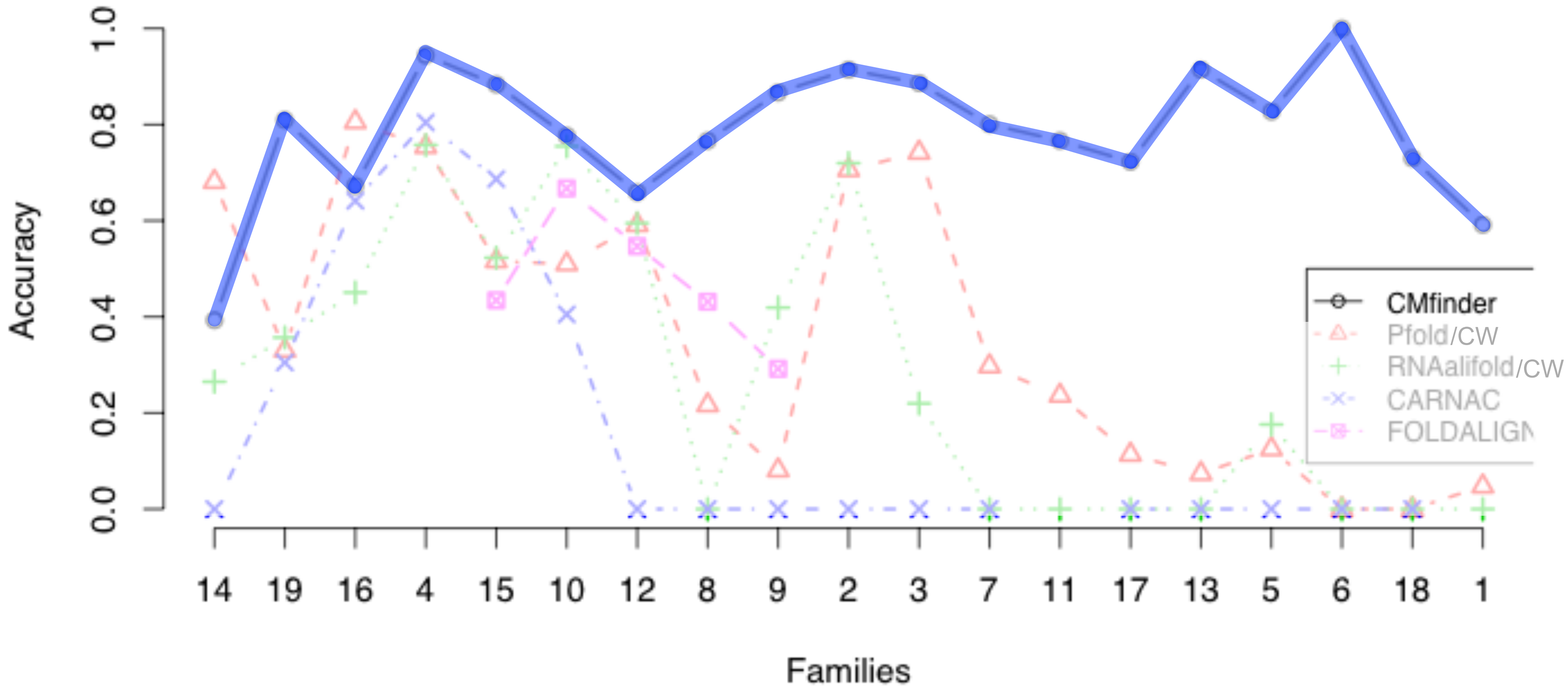
Let

$$K_{ij} = \log\left(\frac{P(L_iL_j)}{P(L_i)P(L_j)} \frac{p_{ij}}{s_is_j}\right) = I_{ij} + \log\frac{p_{ij}}{s_is_j},$$

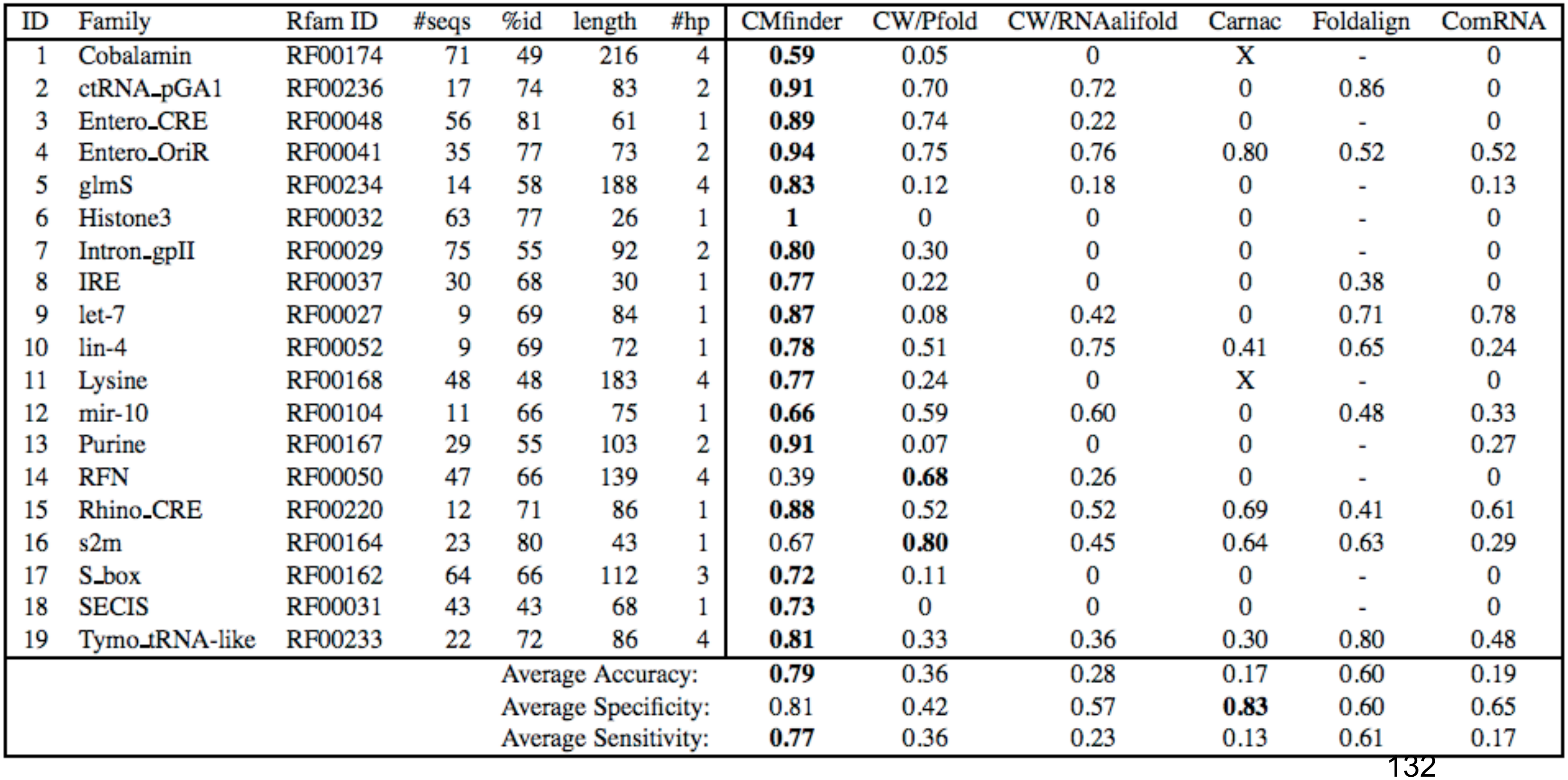
then the maximum likelihood structure σ maximizes $\sum_{(i,j)\in\beta} K_{ij}$.
Can find it via a simple dynamic programming alg.

CMfinder Accuracy

(on Rfam families *with* flanking sequence)



Summary of Rfam test families and results



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	ctRNA_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

Applications:

ncRNA discovery in prokaryotes and vertebrates

Key issue in both cases is
exploiting prior knowledge
to focus on promising data

Application I

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

Yao, Barrick, Weinberg, Neph, Breaker, Tompa and Ruzzo.
PLoS Computational Biology. 3(7): e126, July 6, 2007.

Predicting New *cis*-Regulatory RNA Elements

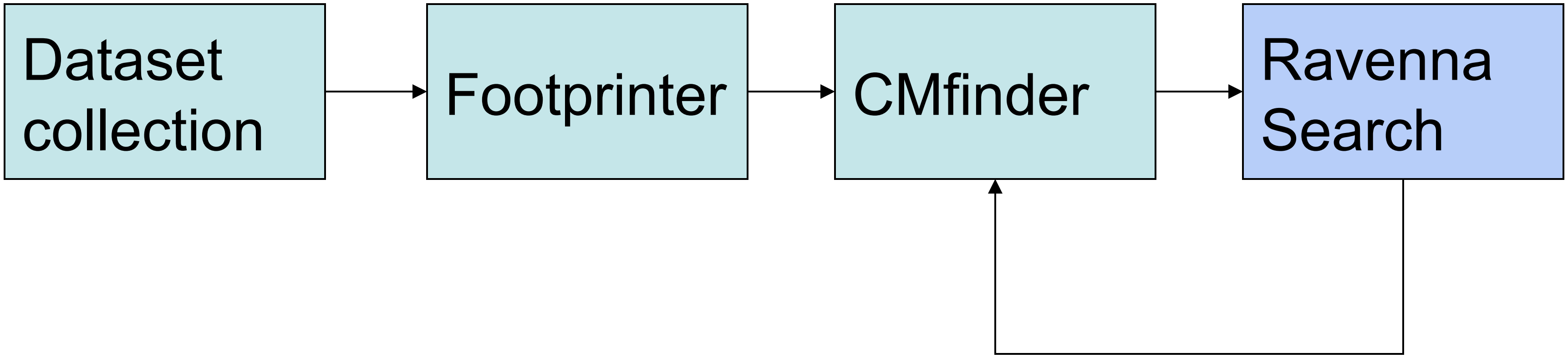
Goal:

- Given unaligned UTRs of coexpressed or orthologous genes, find common structural motifs

Difficulties:

- Low sequence similarity: alignment difficult
- Varying flanking sequence
- Motif missing from some input genes

Use the Right Data;
Do Genome Scale Search



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, find similar genes (“homologs”), look at adjacent DNA

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

Genome Scale Search: Why

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

In most close relatives

More examples give better model, hence even more examples, fewer errors

More examples give more clues to function - critical for wet lab verification

But inclusion of non-examples can degrade motif…

Approach

Get bacterial genomes

For each gene, get 10-30 close orthologs (CDD)

Find most promising genes, based on conserved sequence motifs (Footprinter)

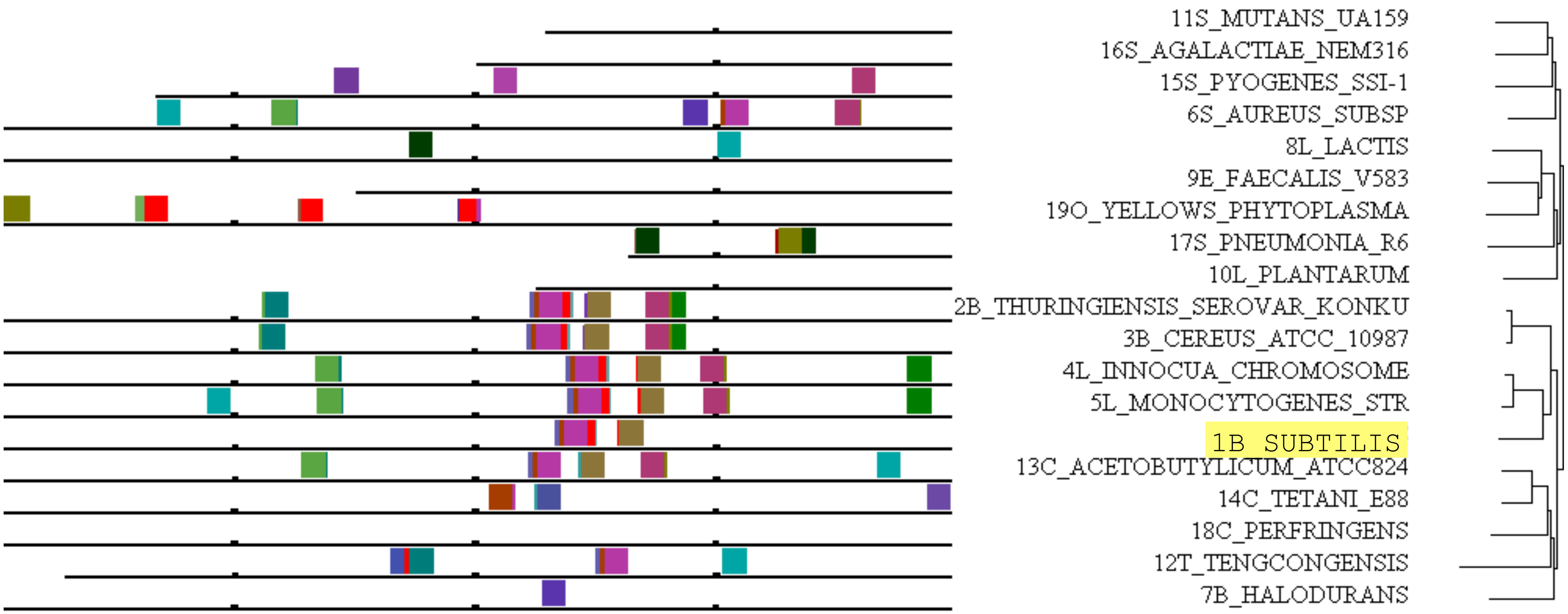
From those, find structural motifs (CMfinder)

Genome-wide search for more instances (Ravenna)

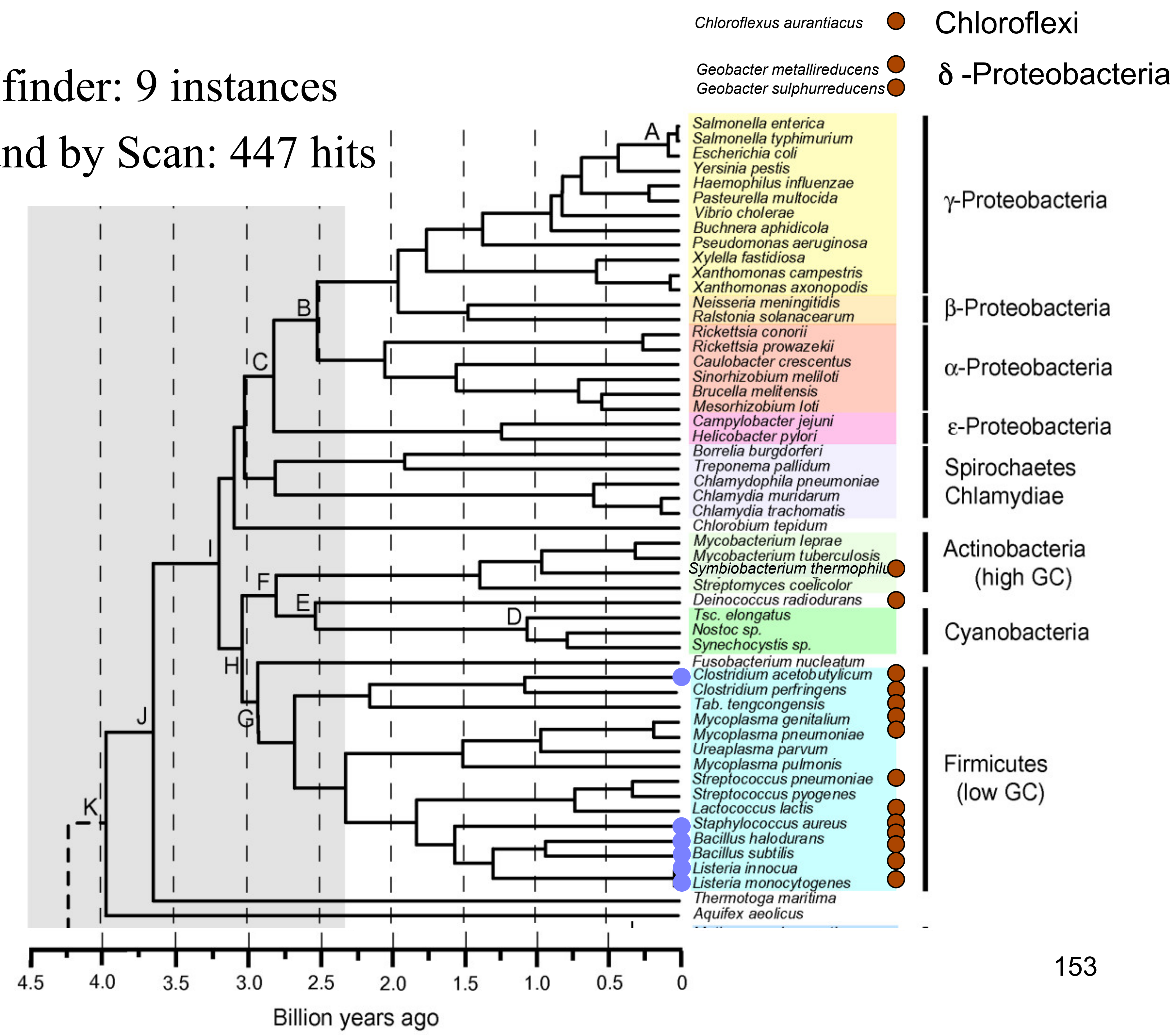
Expert analyses (Breaker Lab, Yale)

Footprinter finds patterns of conservation

Upstream of folC



- ● CMfinder: 9 instances
- ● Found by Scan: 447 hits



Processing Times

Input from ~70 complete Firmicute genomes available in late 2005-early 2006, totaling ~200 megabases

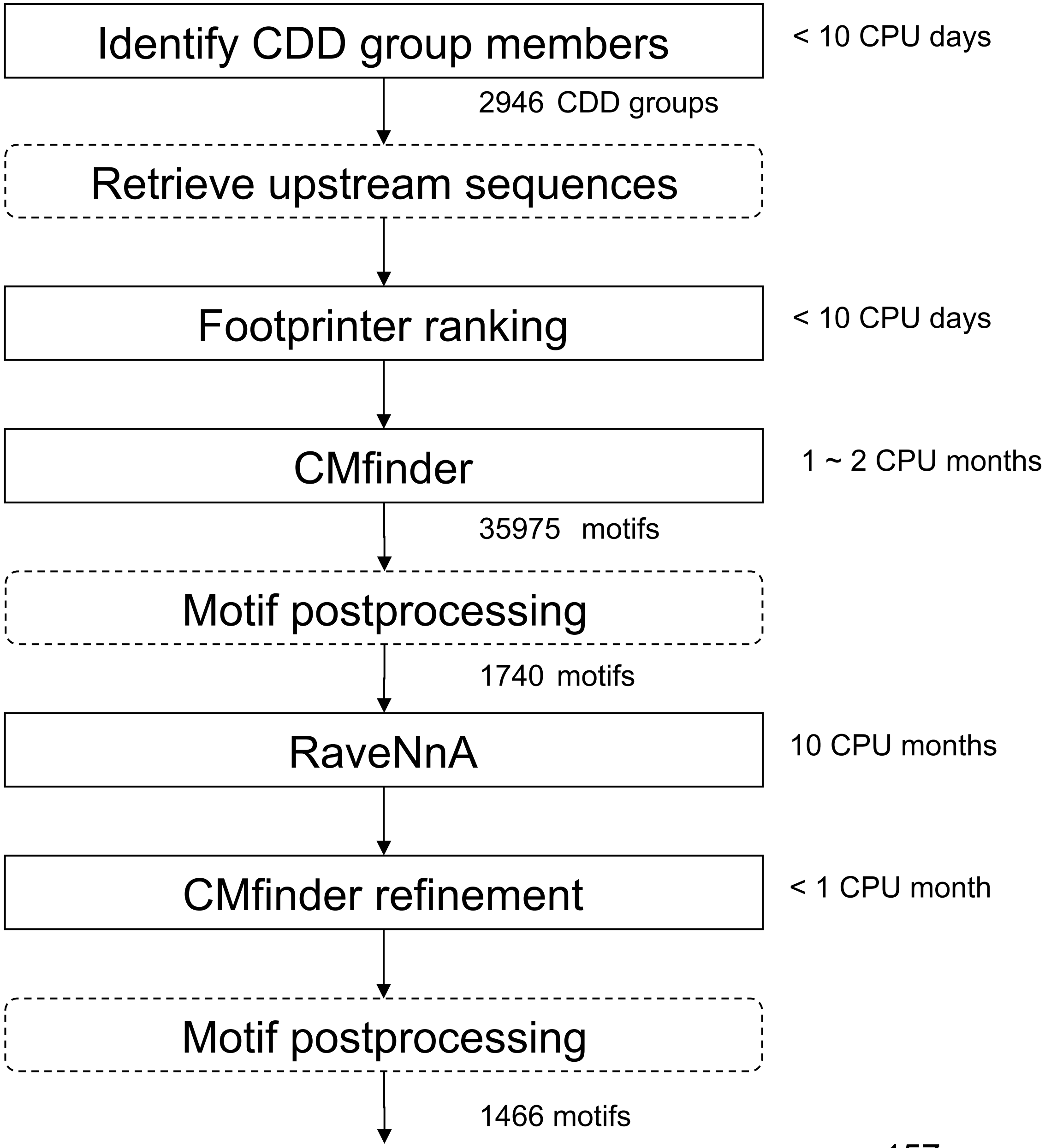


Table 1: Motifs that correspond to Rfam families

Rank			Score	#		CDD			Rfam
RAV	CMF	FP		RAV	CMF	ID	Gene	Description	
0	43	107	3400	367	11	9904	IlvB	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 glmS
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

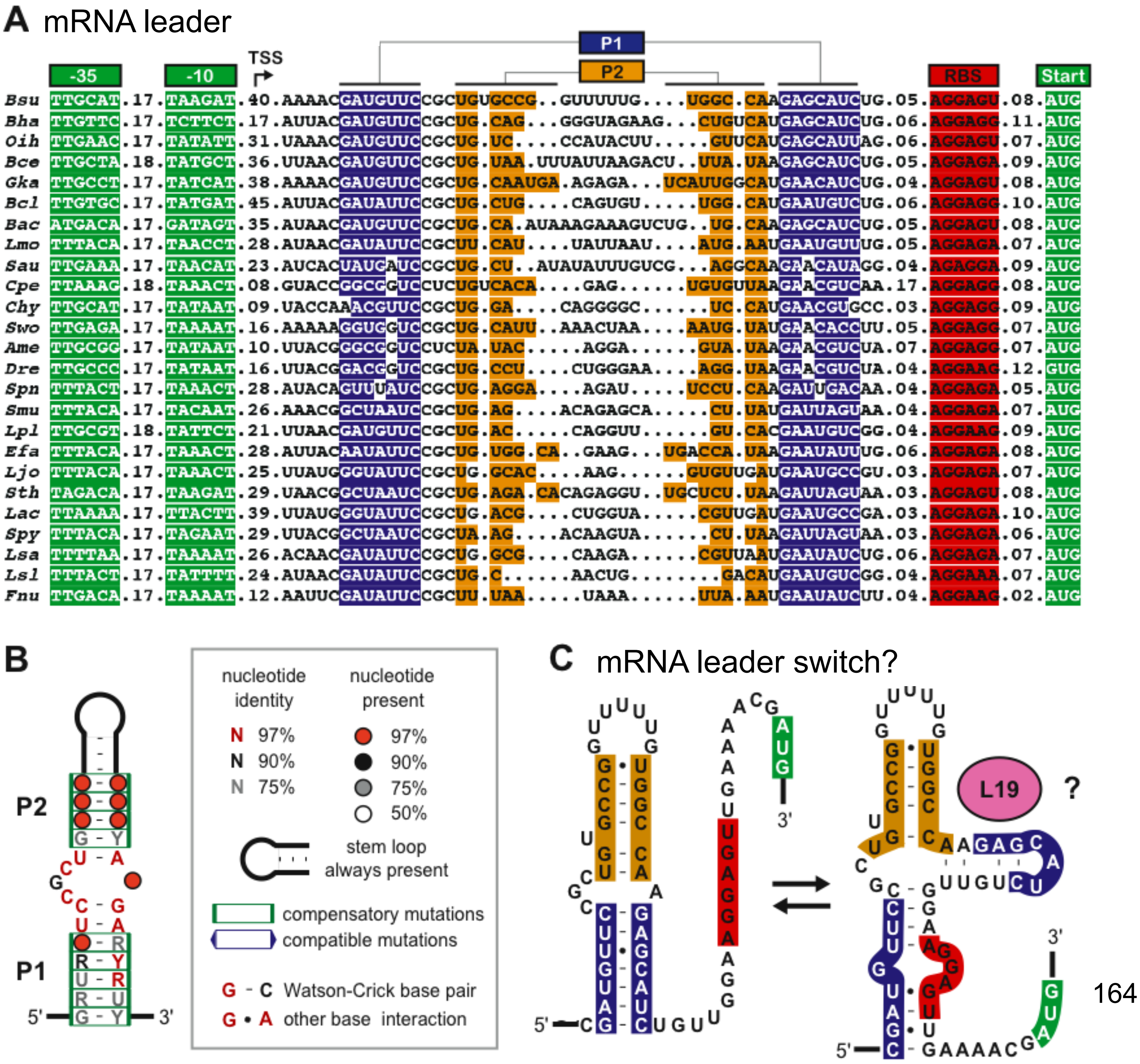
Table 1: Motifs that correspond to Rfam families. “Rank”: the three columns show ranks for refined motif clusters after genome scans (“RAV”), CMfinder motifs before genome scans (“CMF”), and FootPrinter results (“FP”). We used the same ranking scheme for RAV and CMF. “Score”

Rfam		Membership			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. [1]After 2nd RaveNnA scan, membership Sn of Glycine, Cobalamin increased to 76% and 98% resp., Glycine Sp unchanged, but Cobalamin Sp dropped to 84%.

Table 3: High ranking motifs not found in Rfam

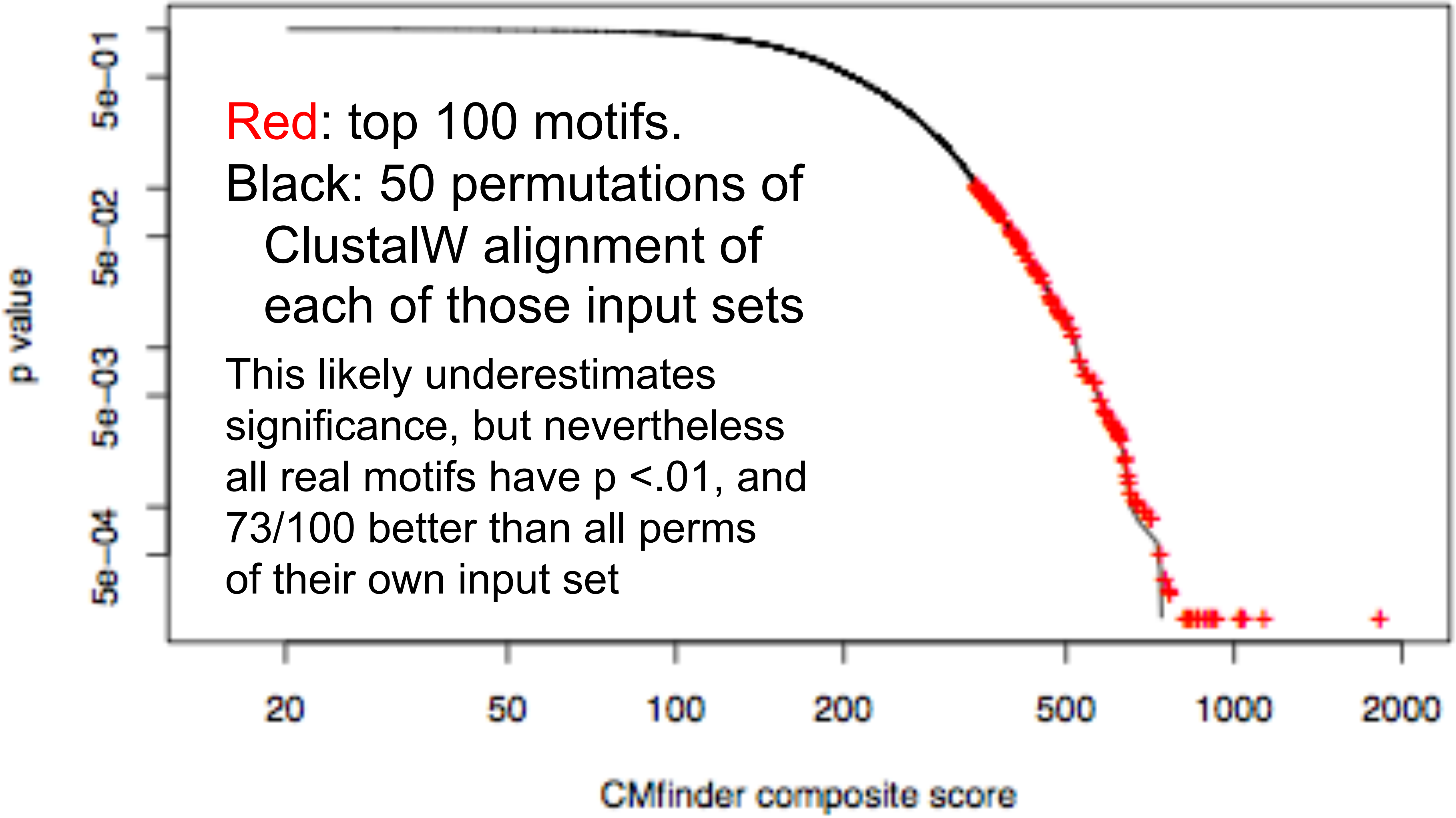
Rank	#	CDD	Gene: Description	Annotation
6	69	28178	DHOase IIa: Dihydroorotase	PyrR attenuator [22]
15	33	10097	RplL: Ribosomal protein L7/L1	L10 r-protein leader; see Supp
19	36	10234	RpsF: Ribosomal protein S6	S6 r-protein leader
22	32	10897	COG1179: Dinucleotide-utilizing enzymes	6S RNA [25]
27	27	9926	RpsJ: Ribosomal protein S10	S10 r-protein leader; see Supp
29	11	15150	Resolvase: N terminal domain	
31	31	10164	InfC: Translation initiation factor 3	IF-3 r-protein leader; see Supp
41	26	10393	RpsD: Ribosomal protein S4 and related proteins	S4 r-protein leader; see Supp [30]
44	30	10332	GroL: Chaperonin GroEL	HrcA DNA binding site [46]
46	33	25629	Ribosomal L21p: Ribosomal prokaryotic L21 protein	L21 r-protein leader; see Supp
50	11	5638	Cad: Cadmium resistance transporter	[47]
51	19	9965	RplB: Ribosomal protein L2	S10 r-protein leader
55	7	26270	RNA pol Rpb2 1: RNA polymerase beta subunit	
69	9	13148	COG3830: ACT domain-containing protein	
72	28	4174	Ribosomal S2: Ribosomal protein S2	S2 r-protein leader
74	9	9924	RpsG: Ribosomal protein S7	S12 r-protein leader
86	6	12328	COG2984: ABC-type uncharacterized transport system	
88	19	24072	CtsR: Firmicutes transcriptional repressor of class III	CtsR DNA binding site [48]
100	21	23019	Formyl trans N: Formyl transferase	
103	8	9916	PurE: Phosphoribosylcarboxyaminoimidazole	
117	5	13411	COG4129: Predicted membrane protein	
120	10	10075	RplO: Ribosomal protein L15	L15 r-protein leader
121	9	10132	RpmJ: Ribosomal protein L36	IF-1 r-protein leader
129	4	23962	Cna B: Cna protein B-type domain	
130	9	25424	Ribosomal S12: Ribosomal protein S12	S12 r-protein leader
131	9	16769	Ribosomal L4: Ribosomal protein L4/L1 family	L3 r-protein leader
136	7	10610	COG0742: N6-adenine-specific methylase	ylbH putative RNA motif [4]
140	12	8892	Pencillinase R: Penicillinase repressor	BlaI, MecI DNA binding site [49]
157	25	24415	Ribosomal S9: Ribosomal protein S9/S16	L13 r-protein leader; Fig 3
160	27	1790	Ribosomal L19: Ribosomal protein L19	L19 r-protein leader; Fig 2
164	6	9932	GapA: Glyceraldehyde-3-phosphate dehydrogenase/erythrose	
174	8	13849	COG4708: Predicted membrane protein	
176	7	10199	COG0325: Predicted enzyme with a TIM-barrel fold	
182	9	10207	RpmF: Ribosomal protein L32	L32 r-protein leader
187	11	27850	LDH: L-lactate dehydrogenases	
190	11	10094	CspR: Predicted rRNA methylase	
194	9	10353	FusA: Translation elongation factors	EF-G r-protein leader



```
Bsu TTGCAT.17.TAAGAT.40.AAAACGAUGUUCCGCUGUGCCG..GUUUUUG....UGGC.CAAGAGCAUCUG.05.AGGAGU.08.AUG
Bha TTGTTC.17.TCTTCT.17.AUUACGAUGUUCCGCUG.CAG...GGGUAGAAG...CUGUCAUGAGCAUCUG.06.AGGAGG.11.AUG
Oih TTGAAC.17.TATATT.31.UAAACGAUGUUCCGCUG.UC....CCAUACUU.....GUUCAUGAGCAUUAG.06.AGGAGU.07.AUG
Bce TTGCTA.18.TATGCT.36.UUAACGAUGUUCCGCUG.UAA.UUUAUUAAGACU..UUA.UAAGAGCAUCUG.05.AGGAGA.09.AUG
Gka TTGCCT.17.TATCAT.38.AAAACGAUGUUCCGCUG.CAAUGA.AGAGA...UCAUUGGCAUGAACAUCUG.04.AGGAGU.08.AUG
Bcl TTGTGC.17.TATGAT.45.AUUACGAUAUUCCGCUG.CUG....CAGUGU.....UGG.CAUGAAUGUCUG.06.AGGAGG.10.AUG
Bac ATGACA.17.GATAGT.35.AUAACGAUGUUCCGCUG.CA.AUAAAGAAAGUCUG..UG.CAAGAGCAUCUG.05.AGGAGU.08.AUG
Lmo TTTACA.17.TAACCT.28.AUAACGAUAUUCCGCUU.CAU....UAUUAAU....AUG.AAUGAAUGUUUG.05.AGGAGA.07.AUG
Sau TTGAAA.17.TAACAT.23.AUCACUAUGAUCCGCUG.CU..AUAUAUUUGUCG...AGGCAAGAACAUAGG.04.AGAGGA.09.AUG
Cpe TTAAAG.18.TAAACT.08.GUACCGGCGGUCCUCUGUCACA....GAG......UGUGUUAAGAACGUCAA.17.AGGAGG.08.AUG
Chy TTGCAT.17.TATAAT.09.UACCAAACGUUCCGCUG.GA....CAGGGGC......UC.CAUGAACGUGCC.03.AGGAGG.09.AUG
Swo TTGAGA.17.TAAAAT.16.AAAAAGGUGGUCCGCUG.CAUU..AAACUAA....AAUG.UAUGAACACCUU.05.AGGAGG.07.AUG
Ame TTGCGG.17.TATAAT.10.UUACGGGCGGUCCUCUA.UAC.....AGGA......GUA.UAAGAACGUCUA.07.AGGAGG.07.AUG
Dre TTGCCC.17.TATAAT.16.UUACGGACGGUCCGCUG.CCU....CUGGGAA....AGG.UAAGAACGUCUA.04.AGGAAG.12.GUG
Spn TTTACT.17.TAAACT.28.AUACAGUUUAUCCGCUG.AGGA....AGAU.....UCCU.CAAGAUUGACAA.04.AGGAGA.05.AUG
Smu TTTACA.17.TACAAT.26.AAACGGCUAAUCCGCUG.AG....ACAGAGCA.....CU.UAUGAUUAGUAA.04.AGGAGA.07.AUG
Lpl TTGCGT.18.TATTCT.21.UUAACGAUGUUCCGCUG.AC.....CAGGUU......GU.CACGAAUGUCGG.04.AGGAAG.09.AUG
Efa TTTACA.17.TAAACT.28.AUUACAAUAUUCCGCUG.UGG.CA..GAAG...UGACCA.UAAGAAUAUUUG.06.AGGAGA.08.AUG
Ljo TTTACA.17.TAAACT.25.UUAUGGGUAUUCCGCUG.GCAC....AAG......GUGUUGAUGAAUGCCGU.03.AGGAGA.07.AUG
Sth TAGACA.17.TAAGAT.29.UAACGGCUAAUCCGCUG.AGA.CACAGAGGU..UGCUCU.UAAGAUUAGUAA.03.AGGAGU.08.AUG
Lac TTAAAA.17.TTACTT.39.UUAUGGGUAUUCCGCUG.ACG....CUGGUA.....CGUUGAUGAAUGCCGA.03.AGGAGA.10.AUG
Spy TTTACA.17.TAGAAT.29.UUACGGCUAAUCCGCUA.AG....ACAAGUA......CU.UAAGAUUAGUAA.03.AGGAGA.06.AUG
Lsa TTTTAA.17.TAAAAT.26.ACAACGAUAUUCCGCUG.GCG....CAAGA......CGUUAAUGAAUAUCUG.06.AGGAGA.07.AUG
Lsl TTTACT.17.TATTTT.24.AUAACGAUAUUCCGCUG.C......AACUG........GACAUGAAUGUCGG.04.AGGAAA.07.AUG
Fnu TTGACA.17.TAAAAT.12.AAUUCGAUAUUCCGCUU.UAA....UAAA......UUA.AAUGAAUAUCUU.04.AGGAAG.02.AUG
```

164

Estimating Motif Significance



Red: top 100 motifs.

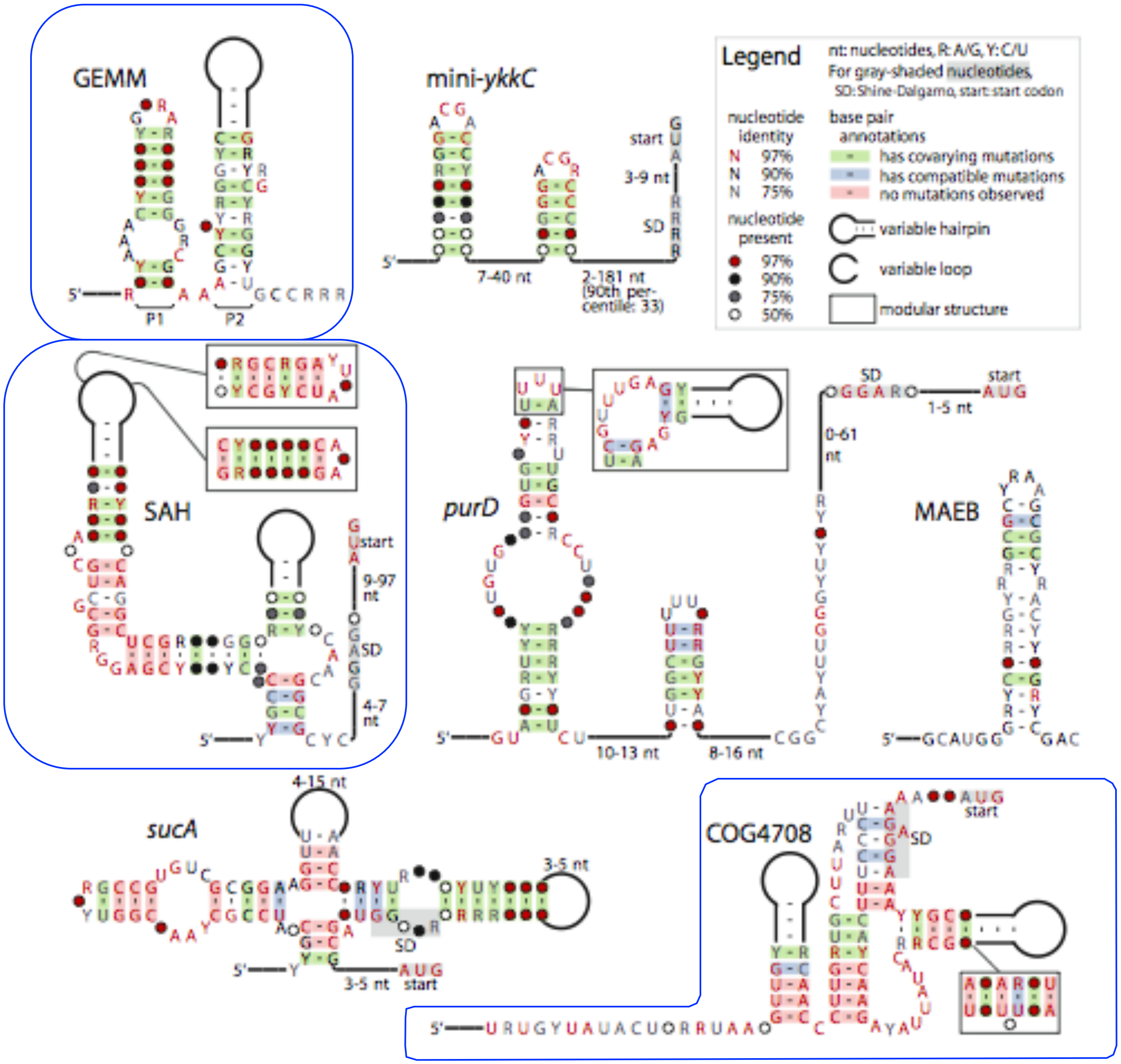
Black: 50 permutations of ClustalW alignment of each of those input sets

This likely underestimates significance, but nevertheless all real motifs have p <.01, and 73/100 better than all perms of their own input set

Application II

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

Weinberg, Barrick, Yao, Roth, Kim, Gore, Wang, Lee,
Block, Sudarsan, Neph, Tompa, Ruzzo and Breaker.
Nucl. Acids Res., July 2007 35: 4809-4819.



boxed = confirmed riboswitch (+2 more)

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

New Riboswitches

(all lab-verified)

SAM – IV	(S-adenosyl methionine)
SAH	(S-adenosyl homocystein)
MOCO	(Molybdenum Cofactor)
PreQ1 – II	(queuosine precursor)
GEMM	(cyclic di-GMP)

GEMM regulated genes

Pili and flagella

Secretion

Chemotaxis

Signal transduction

Chitin

Membrane Peptide

Other - *tfoX*, cytochrome c

GEMM senses a “second messenger” molecule (cyclic di-GMP) produced for signal transduction or for cell-cell communication.

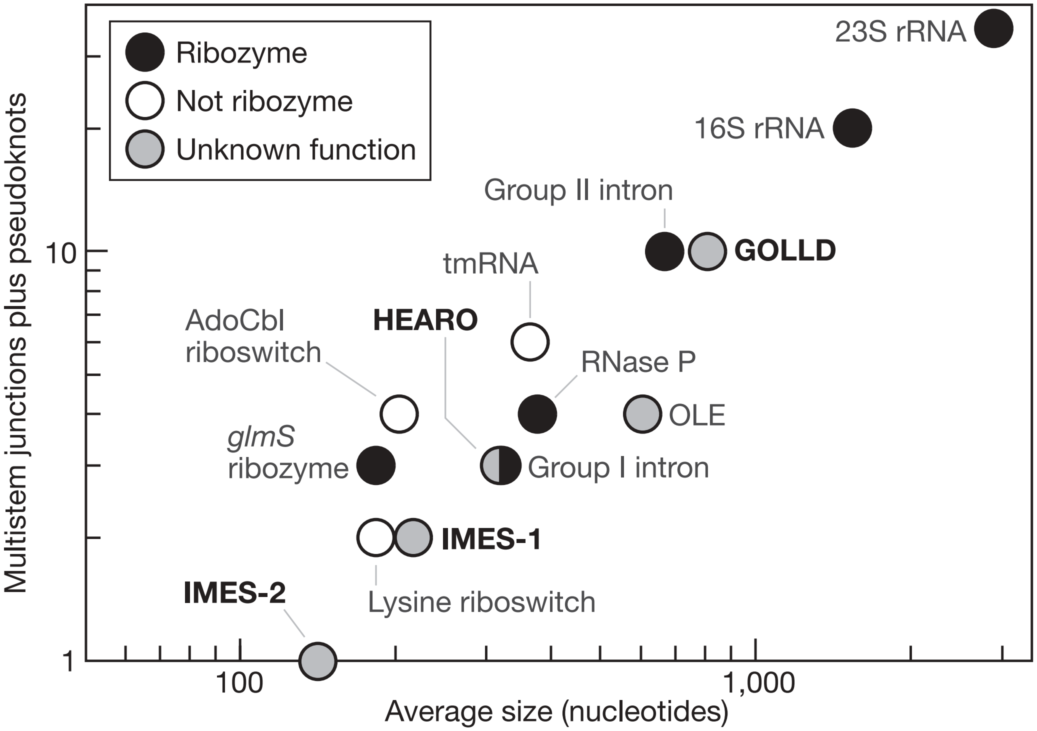
Motif	RNA?	Cis?	Switch?	Phylum/class	M,V	Cov.	#	Non cis
GEMM	Y	Y	y	Widespread	V	21	322	12/309
Moco	Y	Y	Y	Widespread	M,V	15	105	3/81
SAH	Y	Y	Y	Proteobacteria	M,V	22	42	0/41
SAM-IV	Y	Y	Y	Actinobacteria	V	28	54	2/54
COG4708	Y	Y	y	Firmicutes	M,V	8	23	0/23
sucA	Y	Y	y	β-proteobacteria		9	40	0/40
23S-methyl	Y	Y	n	Firmicutes		12	38	1/37
hemB	Y	?	?	β-proteobacteria	V	12	50	2/50
(anti-*hemB*)		(n)	(n)				(37)	(31/37)
MAEB	?	Y	n	β-proteobacteria		3	662	15/646
mini-*ykkC*	Y	Y	?	Widespread	V	17	208	1/205
purD	y	Y	?	ε-proteobacteria	M	16	21	0/20
6C	y	?	n	Actinobacteria		21	27	1/27
alpha-transposases	?	N	N	α-proteobacteria		16	102	39/99
excisionase	?	?	n	Actinobacteria		7	27	0/27
ATPC	y	?	?	Cyanobacteria		11	29	0/23
cyano-30S	Y	Y	n	Cyanobacteria		7	26	0/23
lacto-1	?	?	n	Firmicutes		10	97	18/95
lacto-2	y	N	n	Firmicutes		14	357	67/355
TD-1	y	?	n	Spirochaetes	M,V	25	29	2/29
TD-2	y	N	n	Spirochaetes	V	11	36	17/36
coccus-1	?	N	N	Firmicutes		6	246	112/189
gamma-150	?	N	N	γ-proteobacteria		9	27	6/27

nature Vol 462|3 December 2009|doi:10.1038/nature08586

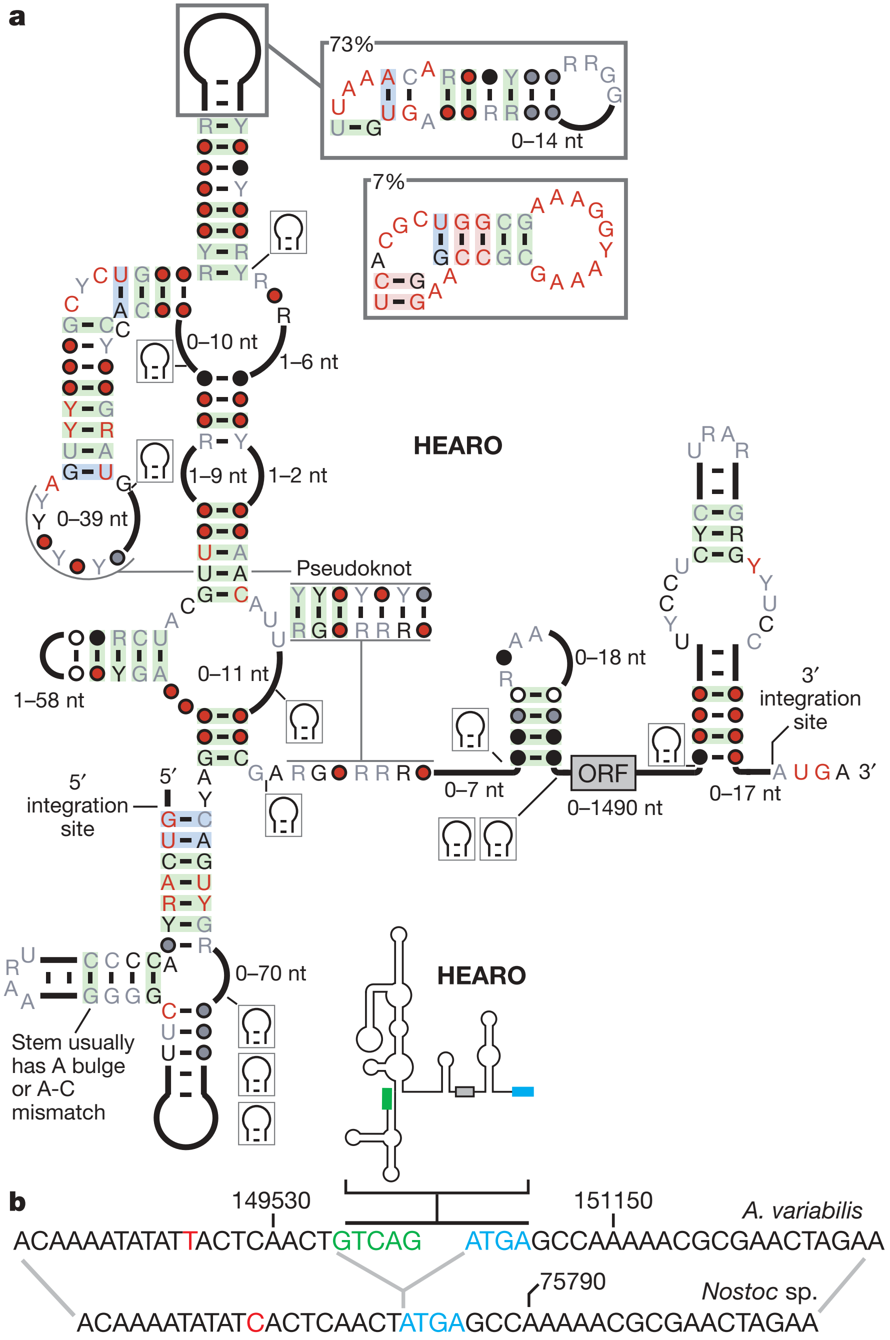
LETTERS

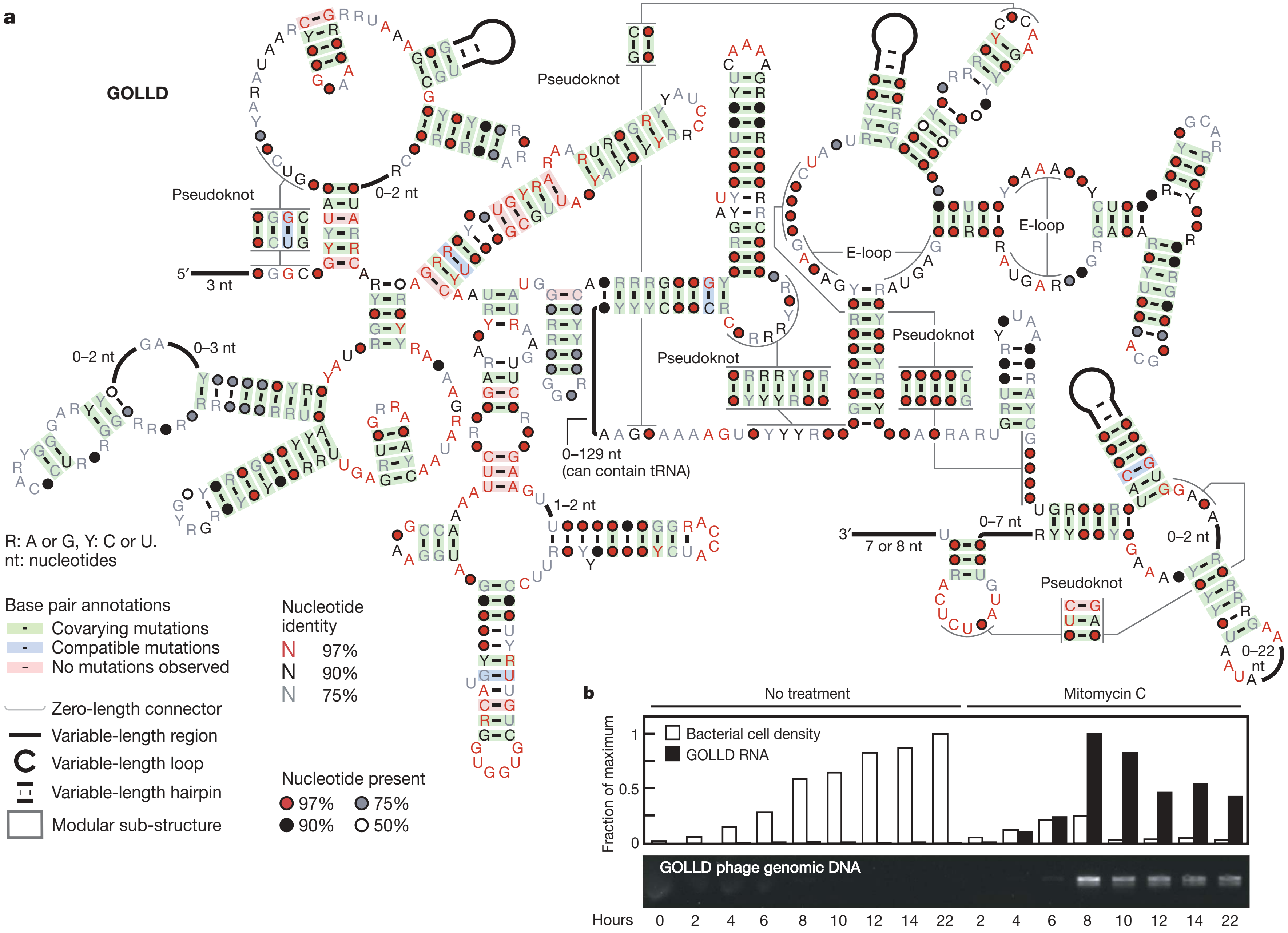
Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis

Zasha Weinberg[1,2], Jonathan Perreault[2], Michelle M. Meyer[2] & Ronald R. Breaker[1,2,3]



RNAs of unusual size and complexity

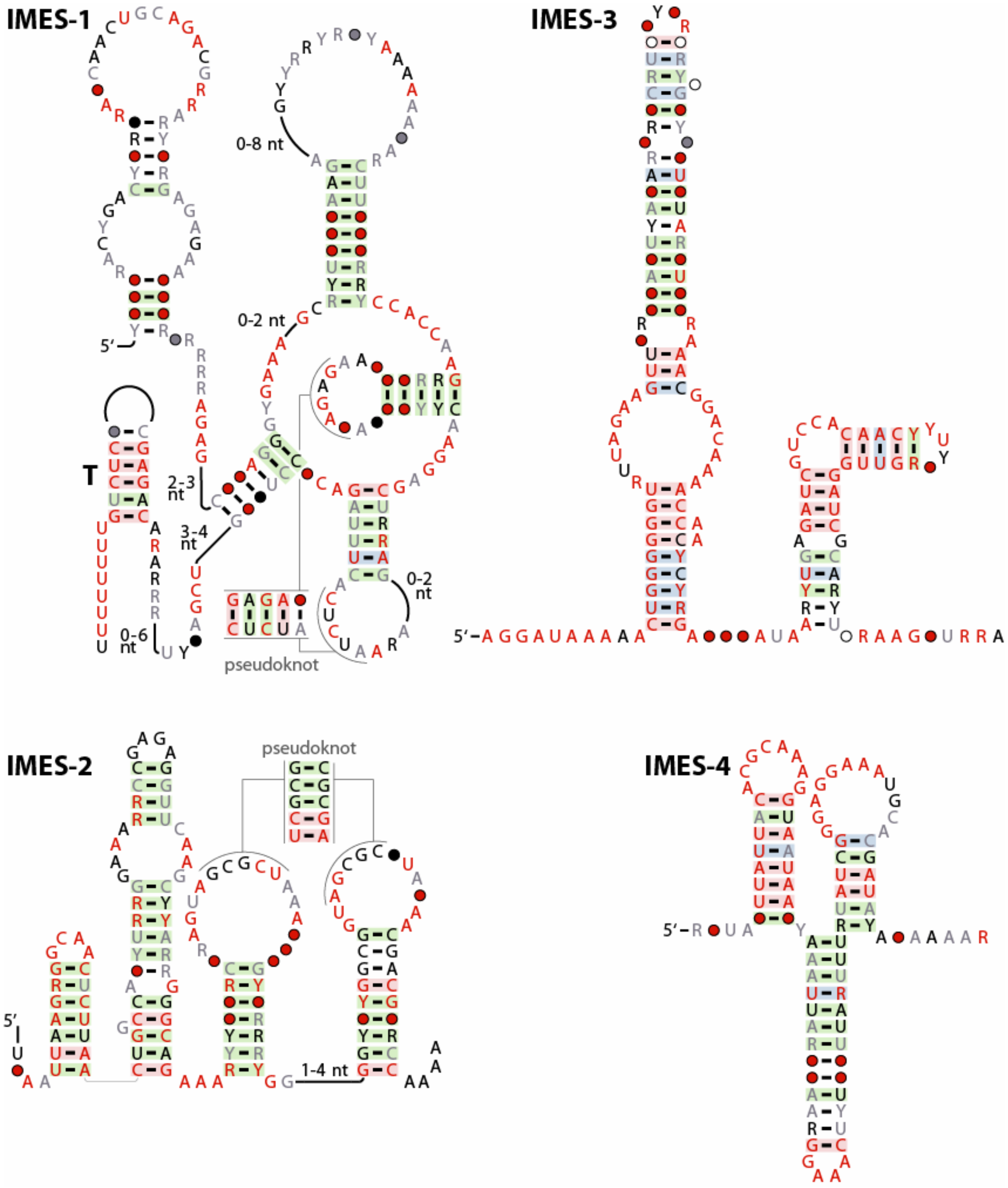



a
GOLLD
Pseudoknot
Pseudoknot
Pseudoknot
Pseudoknot
Pseudoknot
E-loop
E-loop
5′
3 nt
0–2 nt
0–2 nt
0–3 nt
1–2 nt
0–129 nt
(can contain tRNA)
0–7 nt
3′
7 or 8 nt
0–2 nt
0–22 nt
R: A or G, Y: C or U.
nt: nucleotides
Base pair annotations
🟩 Covarying mutations
🟦 Compatible mutations
🟥 No mutations observed
Zero-length connector
Variable-length region
Variable-length loop
Variable-length hairpin
Modular sub-structure
Nucleotide identity
N 97%
N 90%
N 75%
Nucleotide present
● 97%
● 75%
● 90%
○ 50%
b
No treatment
Mitomycin C
□ Bacterial cell density
■ GOLLD RNA
Fraction of maximum
1
0.5
0
GOLLD phage genomic DNA
Hours
0
2
4
6
8
10
12
14
22
2
4
6
8
10
12
14
22

RNAs of unusual abundance

More abundant than 5S rRNA

From unknown marine organisms



Day 4

Our Plot So Far:

Covariance Models (CMs) represent conserved RNA sequence/structure motifs

They allow accurate search, moderately fast (if clever)

Automated model construction / ncRNA discovery in prokaryotes, given careful choice of input data

Today:

ncRNA discovery in vertebrates

Course Project Presentations

Thursday, 12/17, Noon – 5:00, CSE 678

Aim for 20-30 minute talk, plus 5-10 minutes for questions.

Everyone's invited

Vertebrate ncRNAs

Some Results

Rfam Entries in Bacteria

Species name	#Fams	#entries	Genome bp
Roseiflexus sp. RS-1	17	848	5801598
Thermoanaerobacter tengcongensis	27	416	2689445
Clostridium difficile	23	297	4290252
Bacillus thuringiensis	30	238	5257091
Bacillus anthracis	30	232	5227293
Shewanella putrefaciens	23	221	4659220
Yersinia pestis Antiqua	46	207	4702289
Escherichia coli	73	205	5528445
Salmonella typhimurium	85	203	4857432

Rfam Entries in Eukaryotes

Species name	#fams	#	Genome bp
Homo sapiens ((549 / 7892??))	1537	8861	3603093901
Canis lupus familiaris (dog)	1425	6418	2445110183
Pan troglodytes (chimpanzee)	1293	6223	2747703341
Mus musculus (mouse)	1146	5894	2654911517
Ornithorhynchus anatinus (platypus)	169	4631	389485741
Rattus norvegicus (Norway rat)	1071	4309	2303865484
Arabidopsis thaliana (thale cress)	237	1255	93654490
Caenorhabditis elegans (worm)	144	876	100267632
Drosophila melanogaster (fruit fly)	108	493	96018145
Schizosaccharomyces pombe (yeast)	15	131	6992687
Plasmodium falciparum (malaria)	18	35	14214561

Human proteins = ~ 20-25k

# of Human hits for some Rfam families

Family	Accession	# regions in human
7SK	RF00100	1279
SNORA7	RF00409	41
Histone3	RF00032	618
U1	RF00003	682
Y_RNA	RF00019	4516
IRE	RF00037	254

Finding NOVEL vertebrate ncRNAs

Natural approach : Align, Fold, Score

UCSC Browser tracks for Evofold, RNAz

Thousands of candidates

Human Predictions

Evofold

S Pedersen, G Bejerano, A Siepel, K Rosenbloom, K Lindblad-Toh, ES Lander, J Kent, W Miller, D Haussler, "Identification and classification of conserved RNA secondary structures in the human genome." PLoS Comput. Biol., 2, #4 (2006) e33.

48,479 candidates (~70% FDR?)

RNAz

S Washietl, IL Hofacker, M Lukasser, A Hutenhofer, PF Stadler, "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome." Nat. Biotechnol., 23, #11 (2005) 1383-90.

30,000 structured RNA elements

1,000 conserved across all vertebrates.

~1/3 in introns of known genes, ~1/6 in UTRs

~1/2 located far from any known gene

Finding vertebrate ncRNAs

Previous approaches (Evofold, RNAz) have found thousands of candidates, but trusted the vertebrate genome alignments

Find even more if you don’t?

FOLDALIGN

- E Torarinsson, M Sawera, JH Havgaard, M Fredholm, J Gorodkin, "Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure." Genome Res., 16, #7 (2006) 885-9.
- 1800 candidates from 36970 (of 100,000) pairs

CMfinder

Torarinsson, Yao, Wiklund, Bramsen, Hansen, Kjems, Tommerup, Ruzzo and Gorodkin. Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions. Genome Research, Feb 2008, 18(2): 242-251 PMID: 18096747

6500 candidates in ENCODE alone (better FDR, but still high)

ncRNA discovery in Vertebrates

Natural approach : Align, Fold, Score

Previous studies focus on highly conserved regions (Washietl, Pedersen et al. 2007)

- Evofold (Pedersen *et al.* 2006)
- RNAz (Washietl *et al.* 2005)

Thousands of candidates

We explore regions with weak sequence conservation, where alignments aren't trustworthy

Thousands more

CMfinder Search in Vertebrates

Extract ENCODE Multiz alignments

Remove exons, most conserved elements.

56017 blocks, 8.7M bps.

Apply CMfinder to both strands.

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

10,106 predictions, 6,587 clusters.

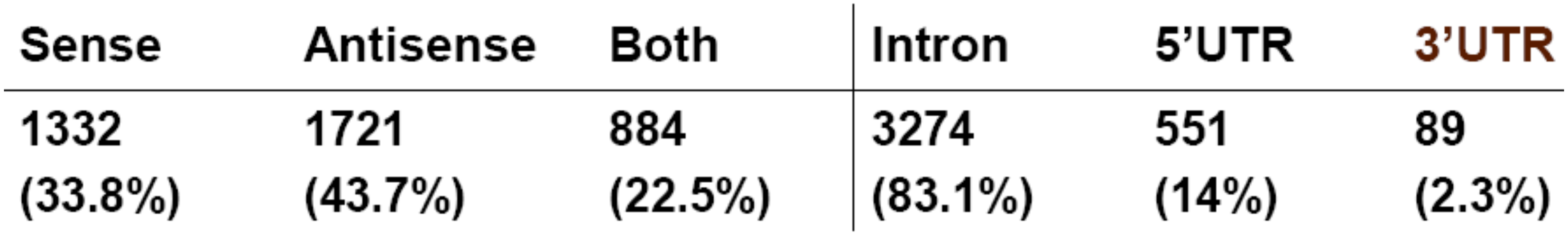
High false positive rate, but still suggests 1000's of RNAs.

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

Overlap with known transcripts

Input regions include only one known ncRNA hsa-mir-483, and we found it.

40% intergenetic, 60% overlap with protein coding gene



Sense	Antisense	Both	Intron	5’UTR	3’UTR
1332 (33.8%)	1721 (43.7%)	884 (22.5%)	3274 (83.1%)	551 (14%)	89 (2.3%)

Assoc w/ coding genes

Many known human ncRNAs lie in introns

Several of our candidates do, too, including some of the tested ones

- #6: *SYN3* (Synapsin 3)
- #10: *TIMP3*, antisense within *SYN3* intron
- #9: *GRM8* (glutamate receptor metabotropic 8)

Overlap w/ Indel Purified Segments

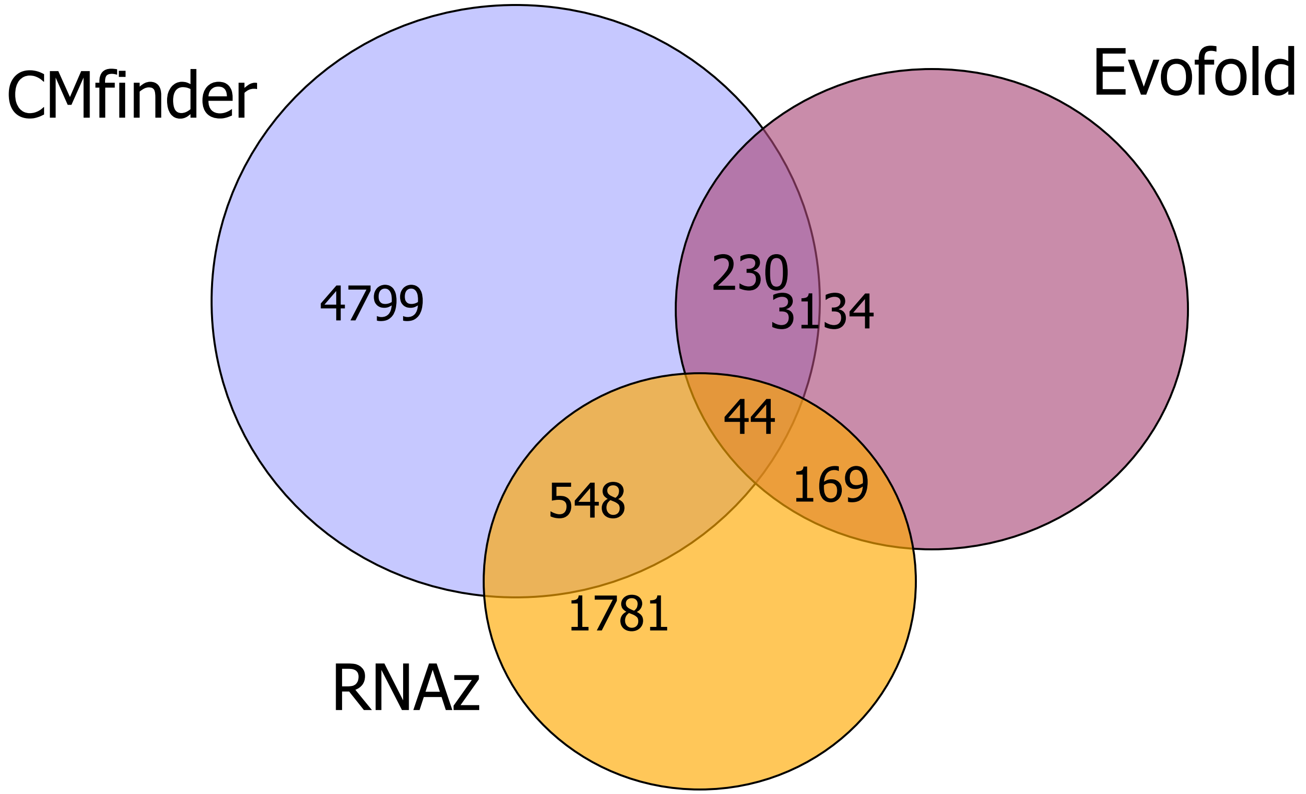
IPS presumed to signal purifying selection

Majority (64%) of candidates have >45% G+C

Strong P-value for their overlap w/ IPS

G+C	data	P	N	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%

Comparison with Evofold, RNAz



Small overlap (w/ highly significant p-values) emphasizes complementarity
Strong association with “Indel purified segments” - I.e., apparently under selection
Strong association with known genes

Alignment Matters

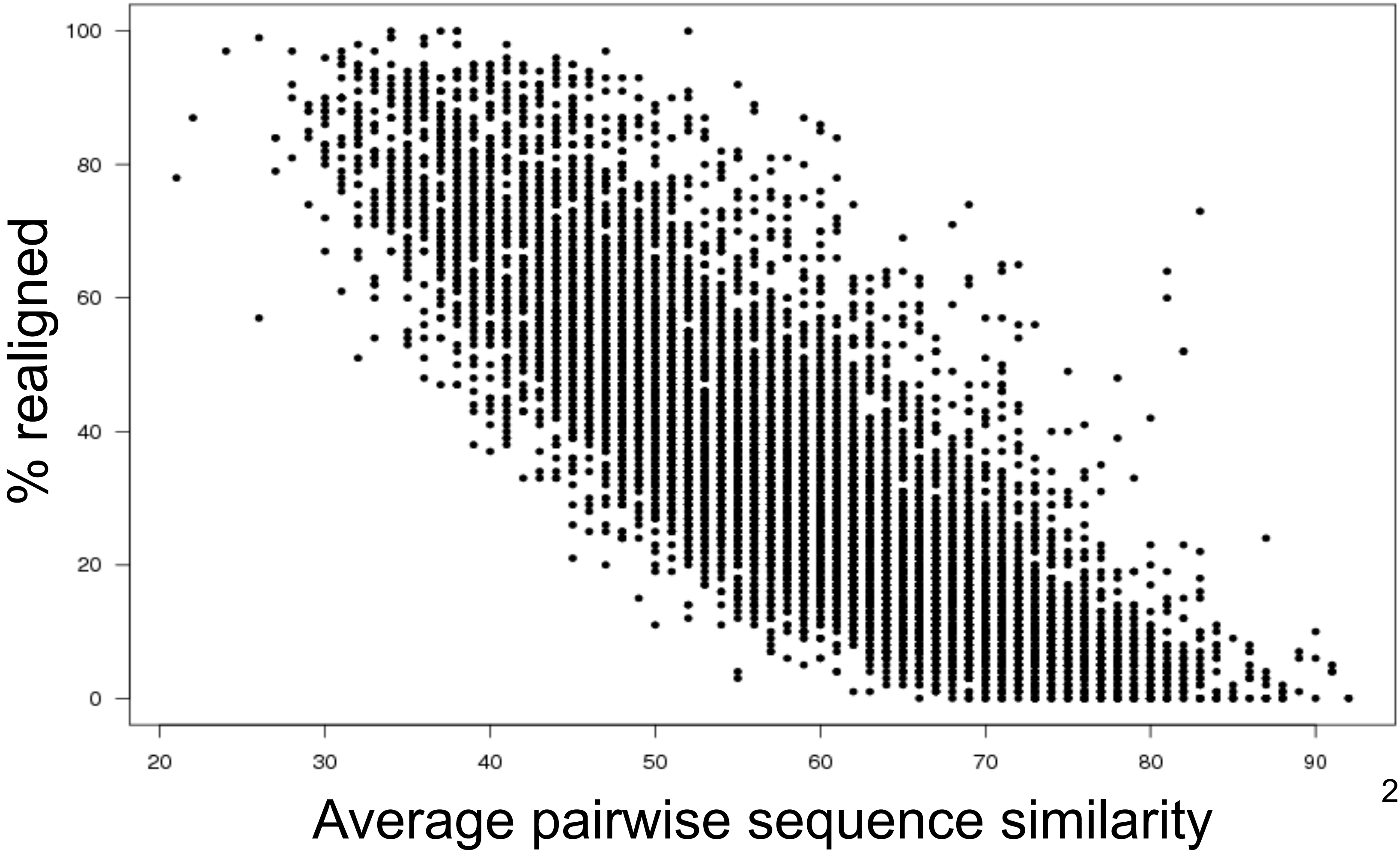
The original MULTIZ alignment without flanking regions. RNAz Score: 0.132 (no RNA)

```
Human  GGTCACTTCAAAGAGGGCTT-GTGGGCTGTGAAACCAAGAGGT----CTTAACAGTATGACCAAAAACTGAAGTT
Chimp  GGACATTTCAATGCGGGCTC-ATGGGGCTGTGAAGCCAAGAGCT----ATTAACACTATGACCAAGGACTGAAATT
Cow    GGTCATTTCAAAGAGGGCTT-ATGAGACCA--AAACCGGGAGCT----CTTAATGCTGTGACCAAAGATTGAAGTT
Dog    GGTCATTTCAAAGAGGGCTTTGTGGAACTA--AAACCAAGGGCT----CTTAACTCTGTGACCAAATATTAGAGTT
Rabbit GATCATTTCAAAGAGGGTTT-GTGGTGCTGTGAAGTCAAGAACT----CTTAACTGTATGCCCAAAGATTAAAGTT
Rhesus GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAACCAAGAGGTAGGTCTTAACAGTATAACCAAAGACTGAAGTT
Str    ((((((......(((((((...(((.........)))..))))....)))......))))))............(
```

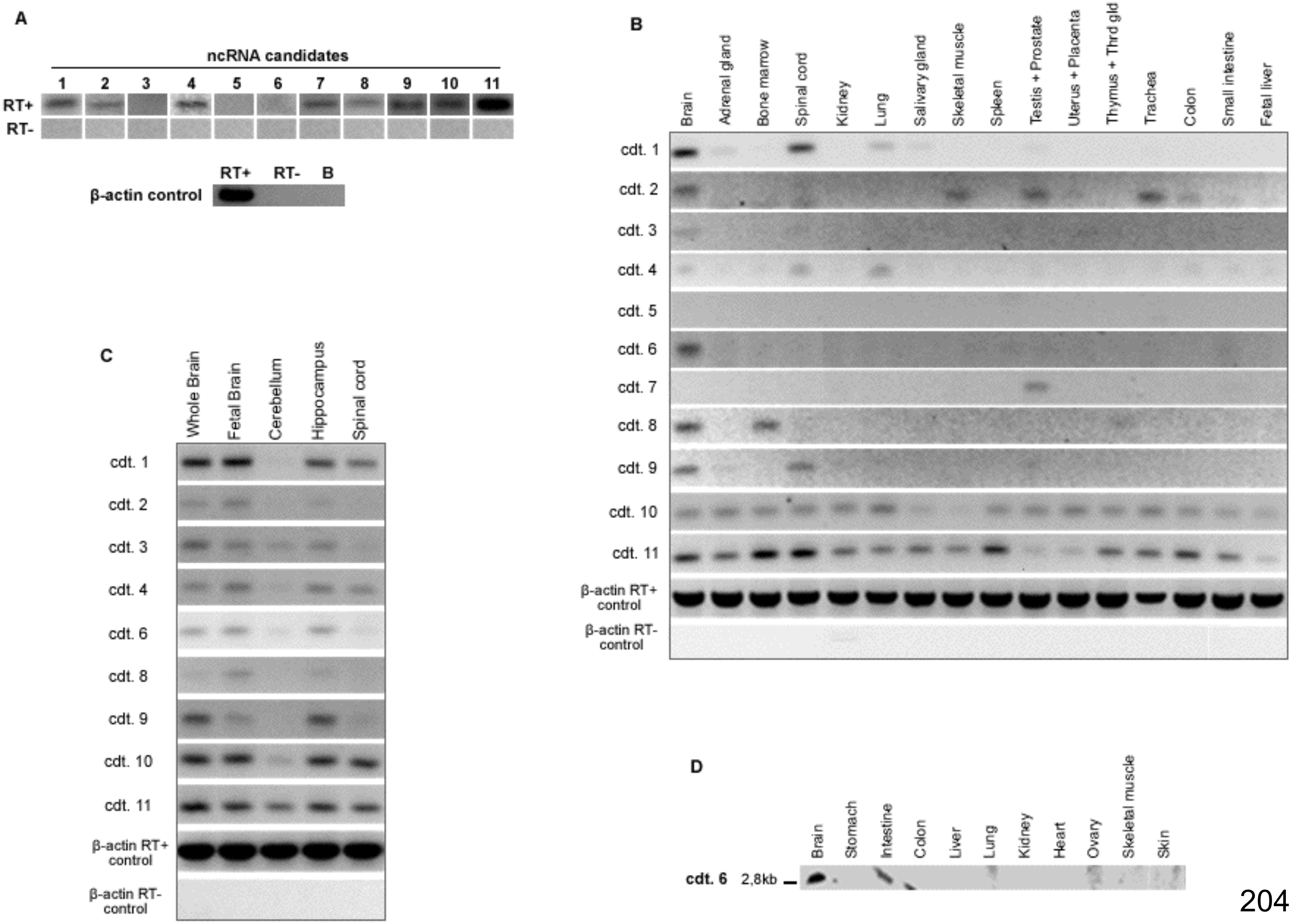
The local CMfinder re-alignment of the MULTIZ block. RNAz Score: 0.709 (RNA)

```
Human  GGTCACTTCAAAGAGGGCTT-GTGGGCTGTGAAA-CCA-----AGAGGTCTTAACAGTATGACCAAAAACTGAAG
Chimp  GGACATTTCAATGCGGGCTC-ATGGGGCTGT-GAAGCCA-----AGAGCTATTAACACTATGACCAAGGACTGAAA
Cow    GGTCATTTCAAAGAGGGCTT-ATGAGACCA--AAA-CCG-----GGAGCTCTTAATGCTGTGACCAAAGATTGAAG
Dog    GGTCATTTCAAAGAGGGCTTTGTGGAACTA--AAA-CCA-----AGGGCTCTTAACTCTGTGACCAAATATTAGAG
Rabbit GATCATTTCAAAGAGGGTTT-GTGGTGCTGT-GAAGTCA-----AGAACTCTTAACTGTATGCCCAAAGATTAAAG
Rhesus GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAA-CCAAGAGG-TAGGTCTTAACAGTATAACCAAAGACTGAAG
Str    ((((((......((((((((..(((..........)))......))))))))......))))))............
```

Realignment



10 of 11 top (differentially) expressed



Summary

Lots of *structurally* conserved ncRNA

Functional significance often unclear

But high rate of confirmed tissue-specific expression in (small) set of top candidates in humans

BIG CPU demands…

Still need for further methods development & application

Summary

ncRNA is a "hot" topic

For family homology modeling: CMs

Training & search like HMM (but slower)

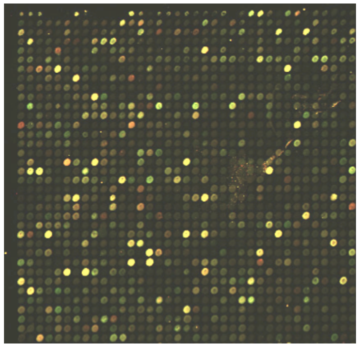
Dramatic acceleration possible

Automated model construction possible

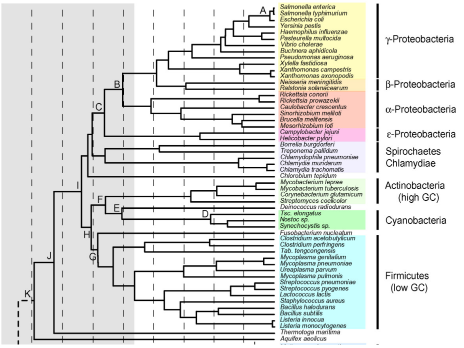
New computational methods yield new discoveries

Many open problems

Course Wrap Up

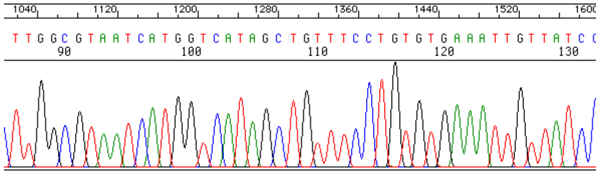


“High-Throughput BioTech”



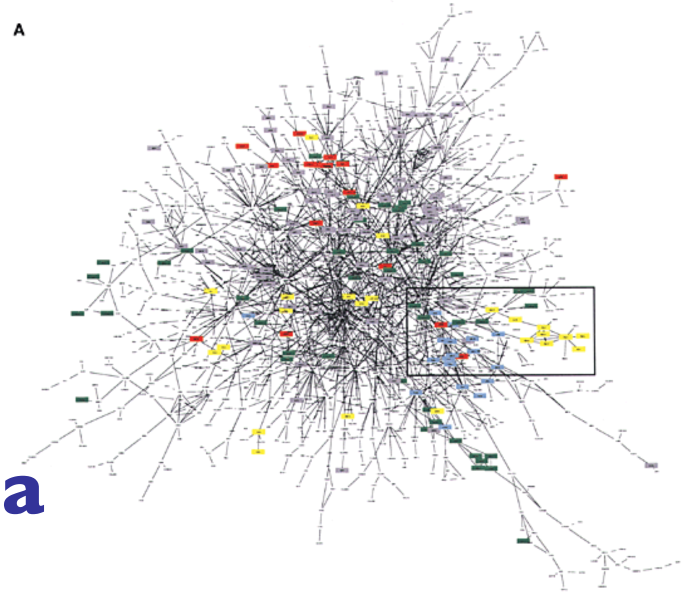
Sensors

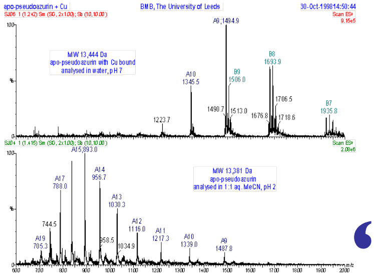
- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction



Controls

- Cloning
- Gene knock out/knock in
- RNAi





Floods of data

“Grand Challenge” problems

CS Points of Contact

- Scientific visualization
  - Gene expression patterns
- Databases
  - Integration of disparate, overlapping data sources
  - Distributed genome annotation in face of shifting underlying coordinates
- AI/NLP/Text Mining
  - Information extraction from journal texts with inconsistent nomenclature, indirect interactions, incomplete/inaccurate models,…
- Machine learning
  - System level synthesis of cell behavior from low-level heterogeneous data (DNA sequence, gene expression, protein interaction, mass spec,
- Algorithms
- …

Frontiers & Opportunities

New data:

Proteomics, SNP, arrays CGH, comparative sequence information, methylation, chromatin structure, ncRNA, interactome

New methods:

graphical models? rigorous filtering?

Data integration

many, complex, noisy sources

Systems Biology

Frontiers & Opportunities

Open Problems:

splicing, alternative splicing

multiple sequence alignment (genome scale, w/ RNA etc.)

protein & RNA structure

interaction modeling

network models

RNA trafficing

ncRNA discovery

…

Exciting Times

Lots to do

Various skills needed

I hope I’ve given you a taste of it

Thanks!