Searching for Noncoding RNA

Larry Ruzzo

Computer Science & Engineering
Genome Sciences
University of Washington

http://www.cs.washington.edu/homes/ruzzo

BioC 2006, Seattle, 8/4/2006

Outline

Noncoding RNA

Why are they hard to discover?

Key computational problems:

- Motif discovery
- Motif search

Sketch new methods

Application:

- cis-regulatory motifs in actinobacteria

Non-coding RNA

Messenger RNA - codes for proteins

Non-coding RNA - all the rest

- Before, say, mid 1990's, 1-2 dozen known (critically important, but narrow roles: e.g. tRNA)

Since mid 90's dramatic discoveries

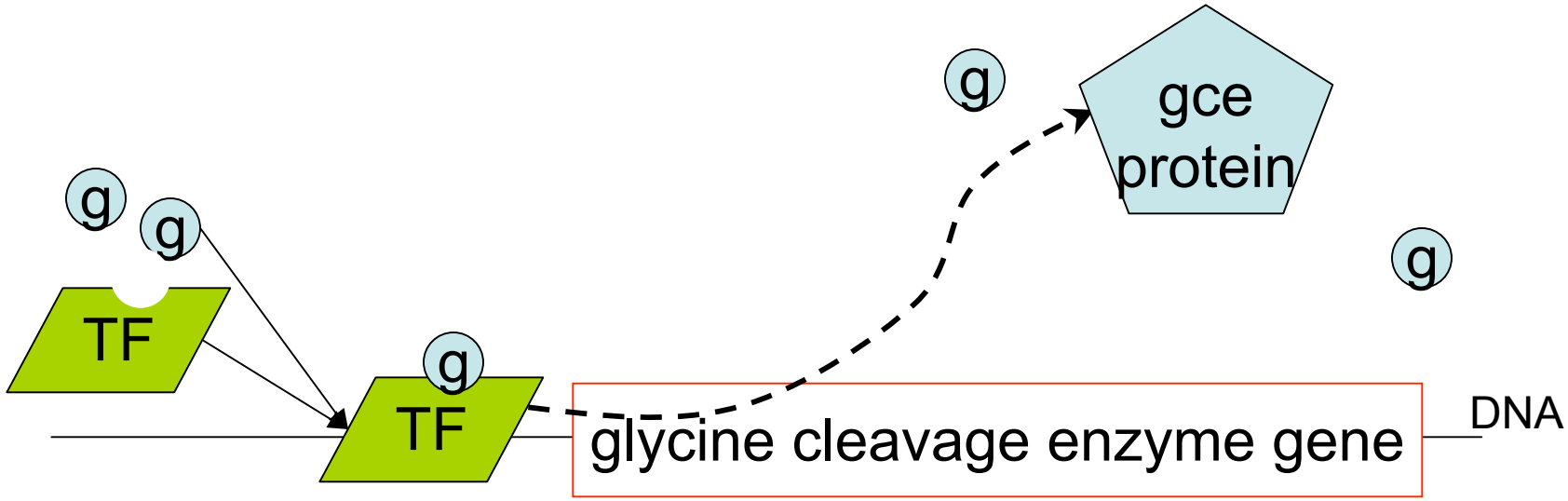
- Hundreds of new families
- Regulation, transport, stability/degradation
- E.g. "microRNA": ≈ 100's in humans

By some estimates, ncRNA >> mRNA

Example: Glycine Regulation

How is glycine level regulated?

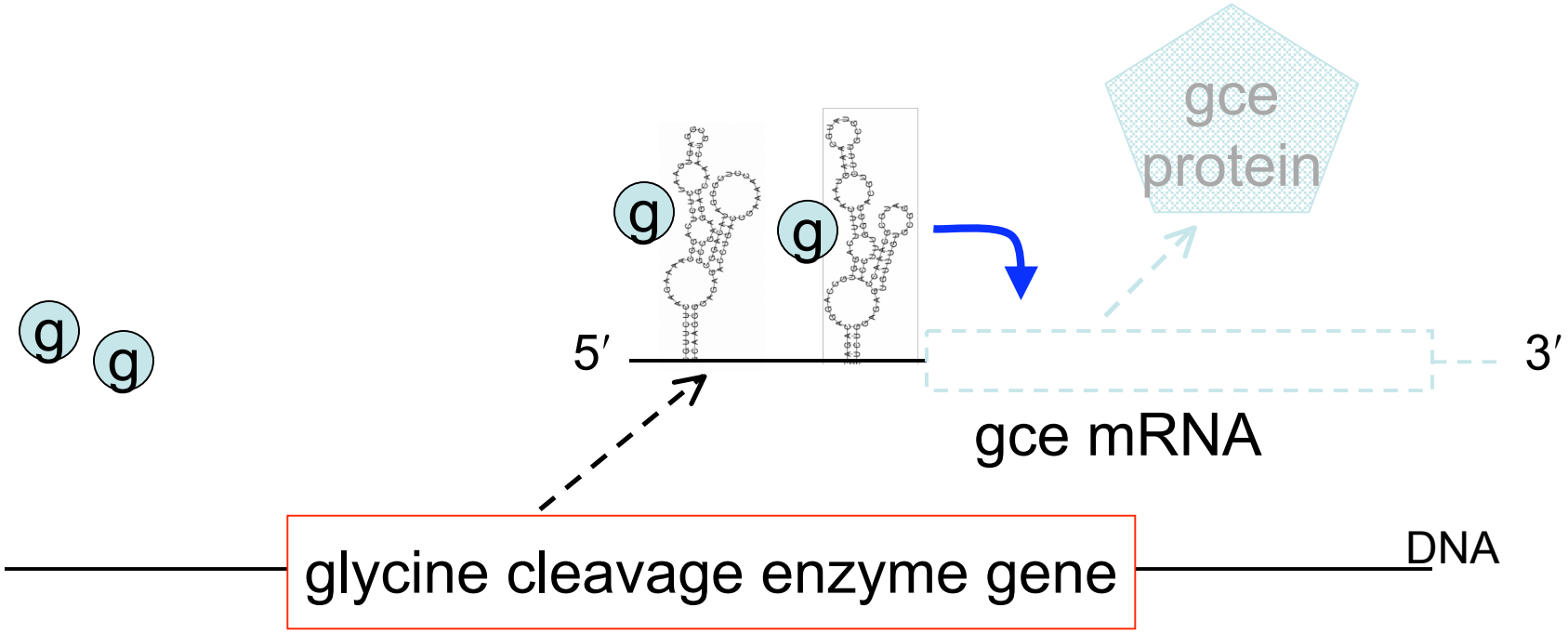
Plausible answer:



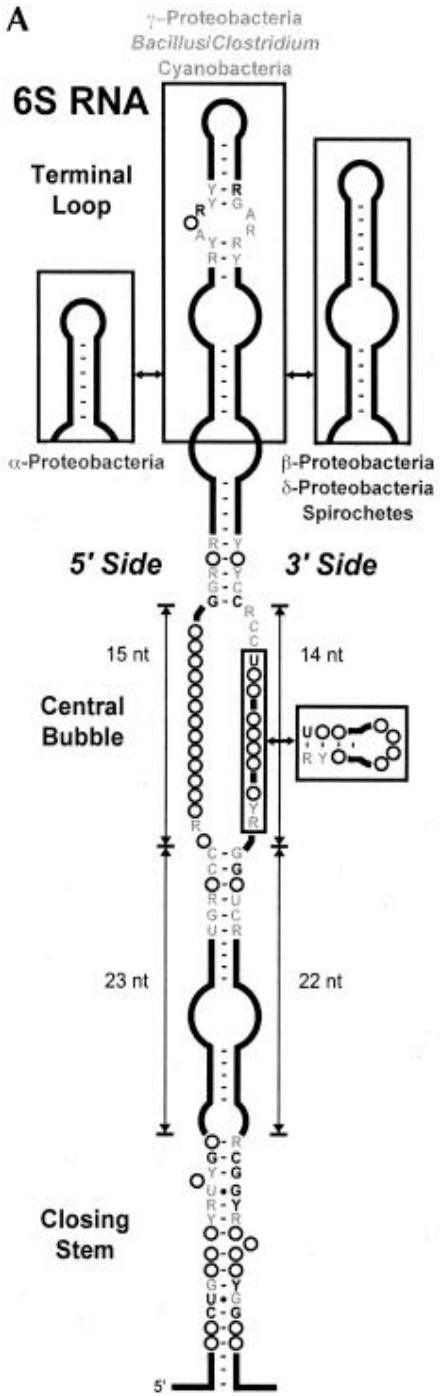
transcription
factors (proteins)

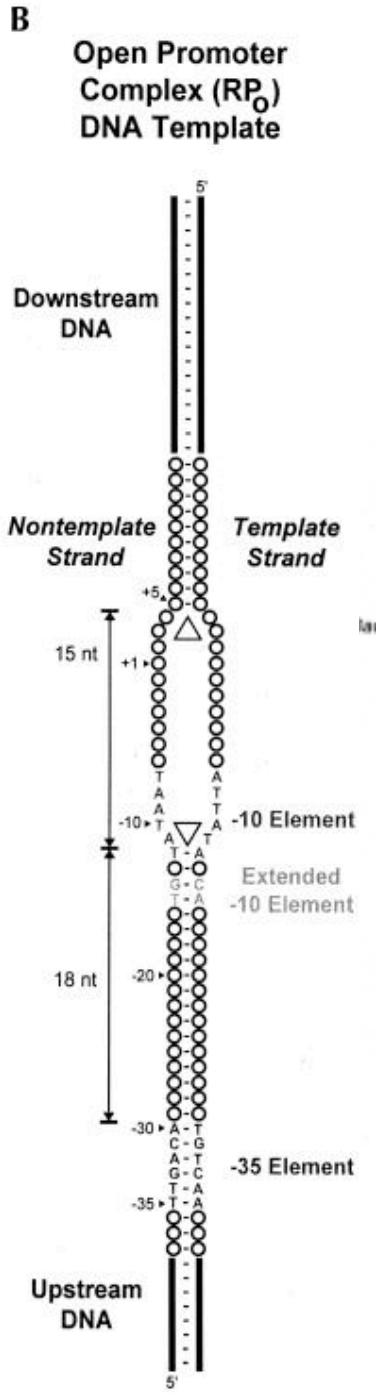
The Glycine Riboswitch

Actual answer (in many bacteria):

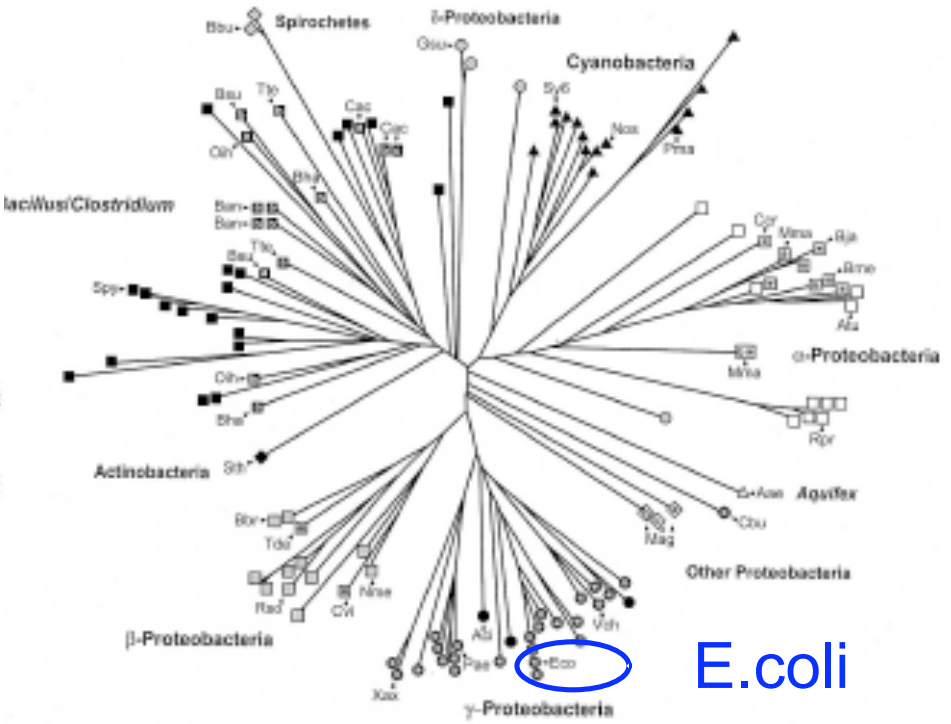


Mandal et al. Science 2004



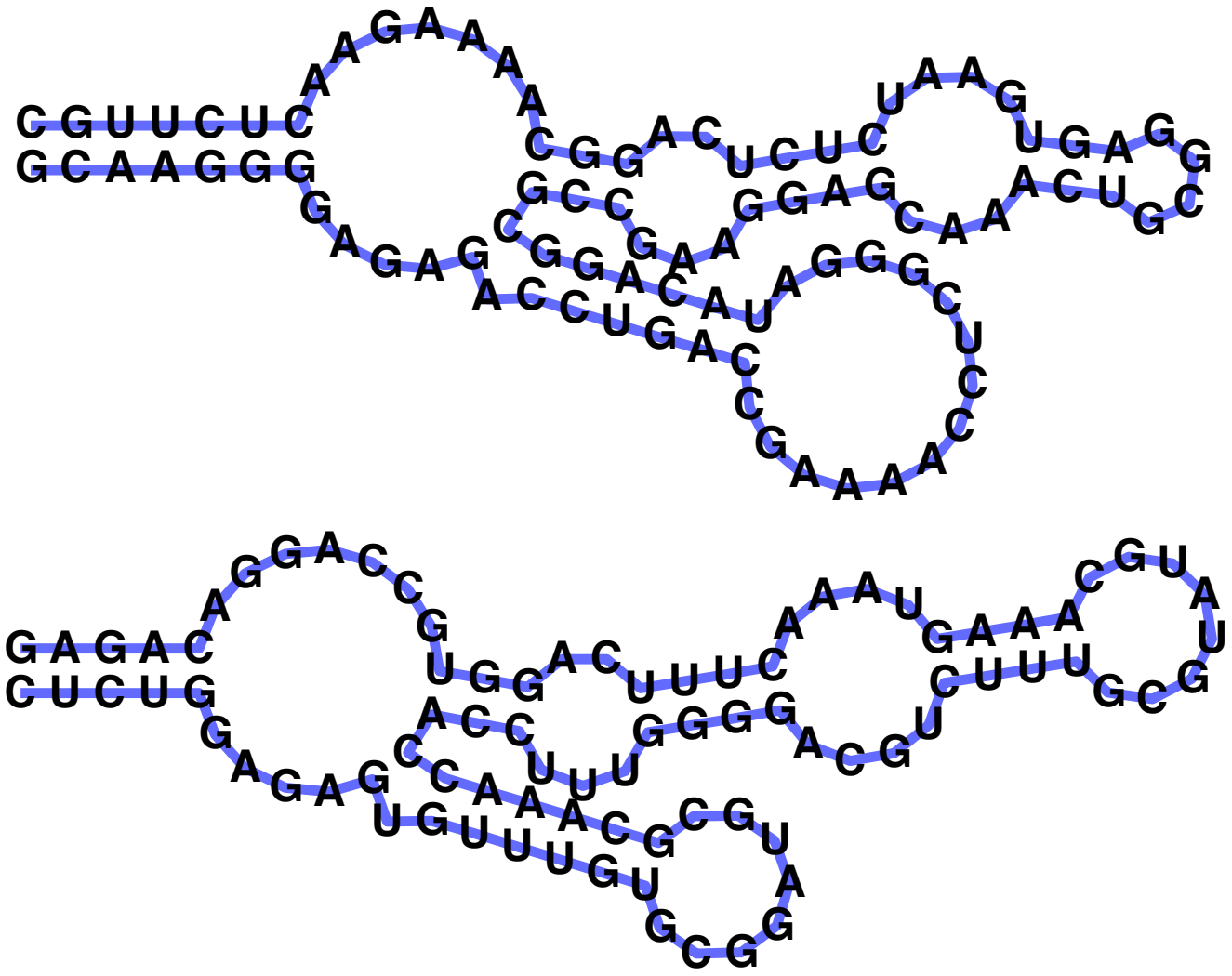


6S mimics an open promoter



Barrick et al. *RNA* 2005
Trotochaud et al. *NSMB* 2005
Willkomm et al. NAR 2005

Why should these be hard to discover?



A: *Structure* often more important than *sequence*

Wanted

Good, fast search tools

("RNA BLAST", etc.)

Good, fast motif discovery tools

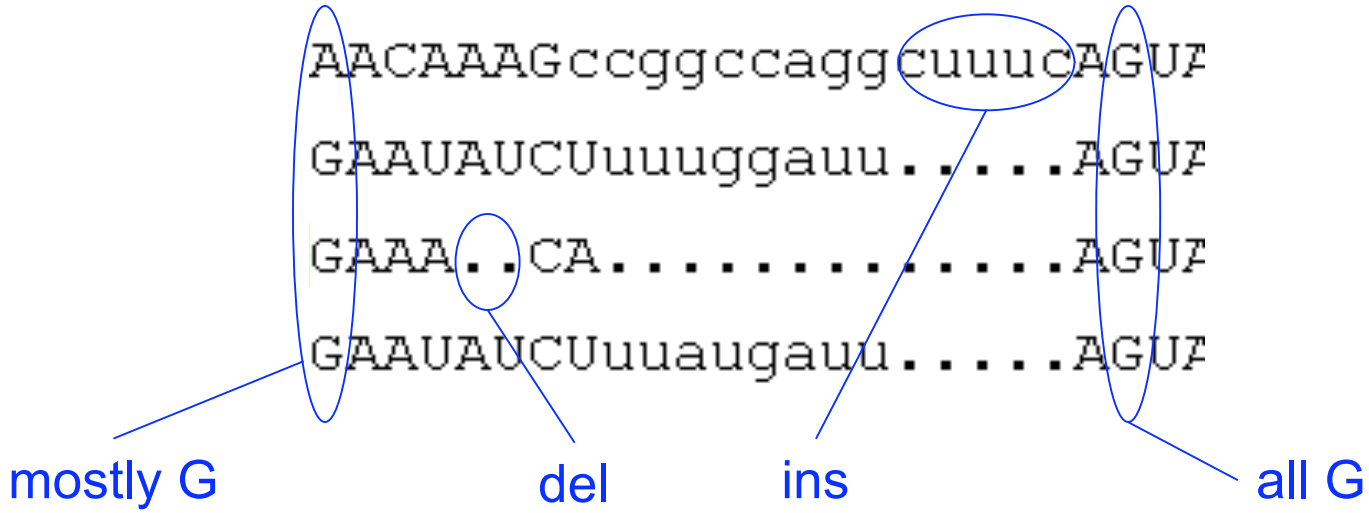
("RNA MEME", etc.)

Importance of structure makes both hard; progress on both below

How to model an RNA “Motif”?

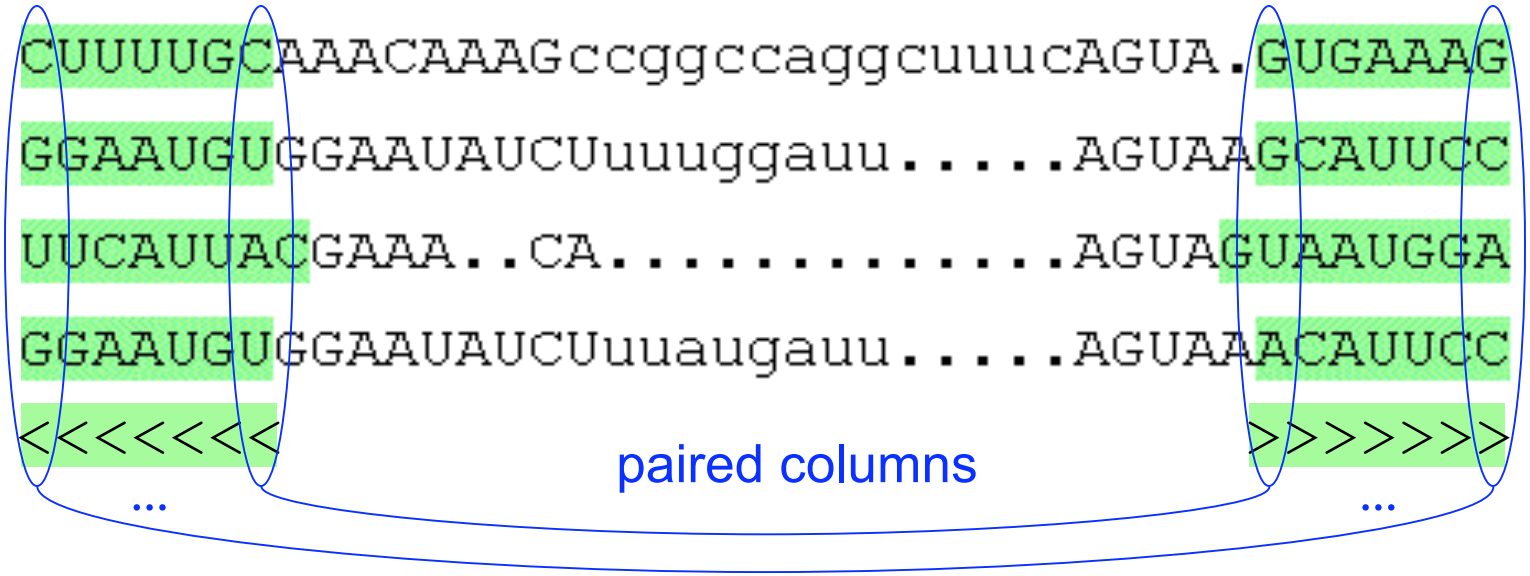
Conceptually, start with a profile HMM:

- 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



RNA Motif Models

"Covariance Models" (Eddy & Durbin 1994)

aka profile stochastic context-free grammars

aka hidden Markov models on steroids

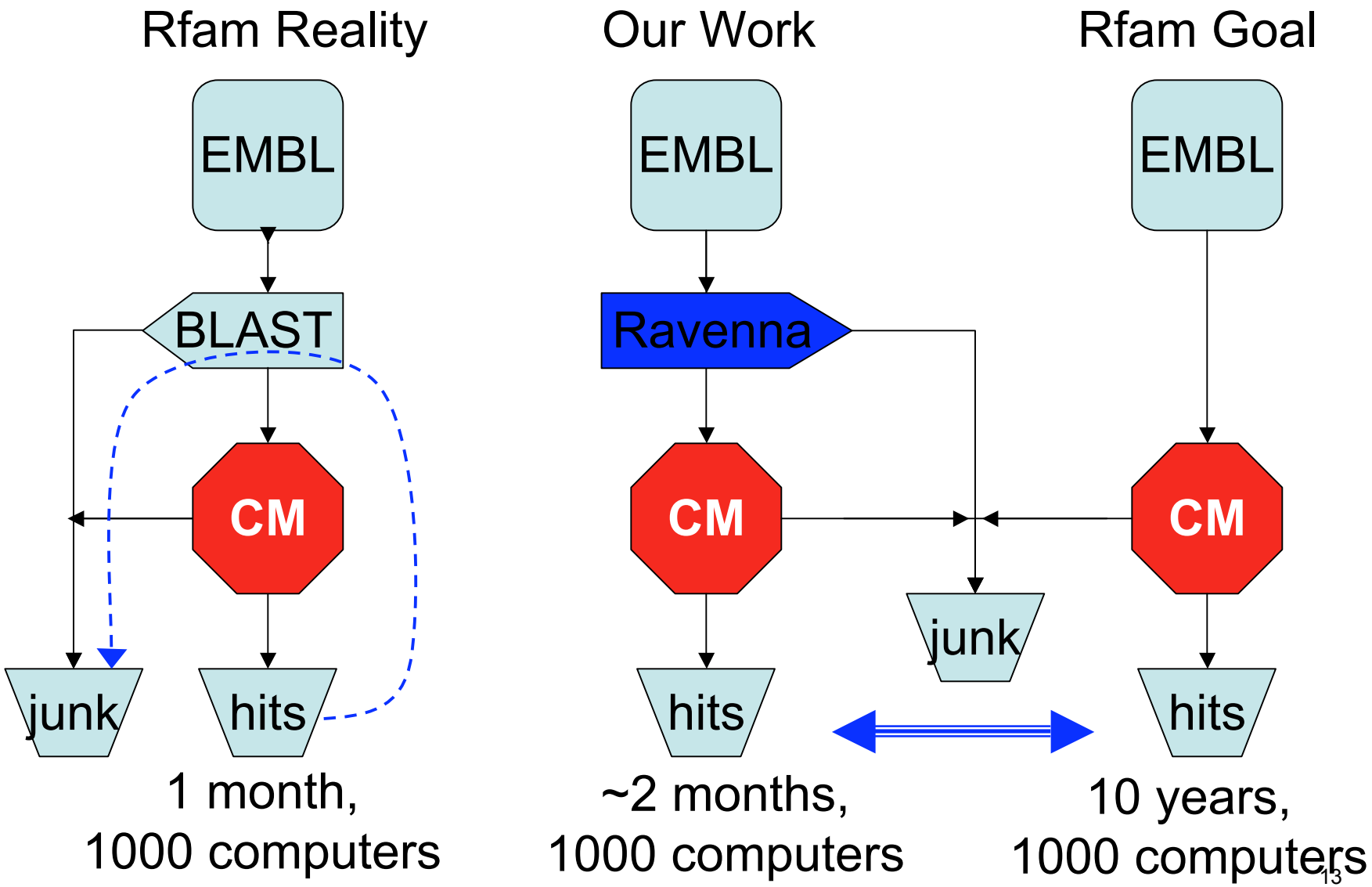
Model position-specific nucleotide preferences *and* base-pair preferences

Pro: accurate

Con: model building hard, search sloooow

Task 1: Faster Search

CM's are good, but slow



Ravenna: Genome Scale RNA Search

Typically 100x speedup over raw CM, with no (or little) loss in accuracy:

- drop structure from CM to create a (faster) HMM
- use that to pre-filter sequence; discard parts where, provably, the CM will score < threshold; actually run CM on the rest (the promising parts)
- assignment of HMM transition/emission scores is key (large convex optimization problem)

Weinberg & Ruzzo, *Bioinformatics*, 2004, 2006

Results: buried treasures

Name	# found BLAST + CM	# found rigorous filter + CM	# new
Pyrococcus snoRNA	57	180	123
Iron response element	201	322	121
Histone 3' element	1004	1106	102
Purine riboswitch	69	123	54
Retron msr	11	59	48
Hammerhead I	167	193	26
Hammerhead III	251	264	13
U4 snRNA	283	290	7
S-box	128	131	3
U6 snRNA	1462	1464	2
U5 snRNA	199	200	1
U7 snRNA	312	313	1

Task 2: Motif Discovery

RNA Motif Discovery

Typical problem: given a ~10-20 unaligned sequences of ~1kb, most of which contain instances of one RNA motif of, say, 150bp -- find it

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

“Obvious” Approach I

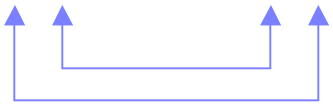
Predict secondary RNA structure using MFOLD or Vienna

Problems

- false folding predictions
- comparing structures

“Obvious” Approach II: Predict from Multiple Sequence Alignment

… GA … UC …
… GA … UC …
… GA … UC …
… CA … UG …
… CC …GG …
… UA … UA …



Compensatory mutations reveal structure, *but* usual alignment algorithms penalize them (twice)

Our Approach: CMfinder

Simultaneous alignment, folding and CM-based motif description using an EM-style learning procedure

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006

Alignment → CM → Alignment

Similar to HMM, but much slower

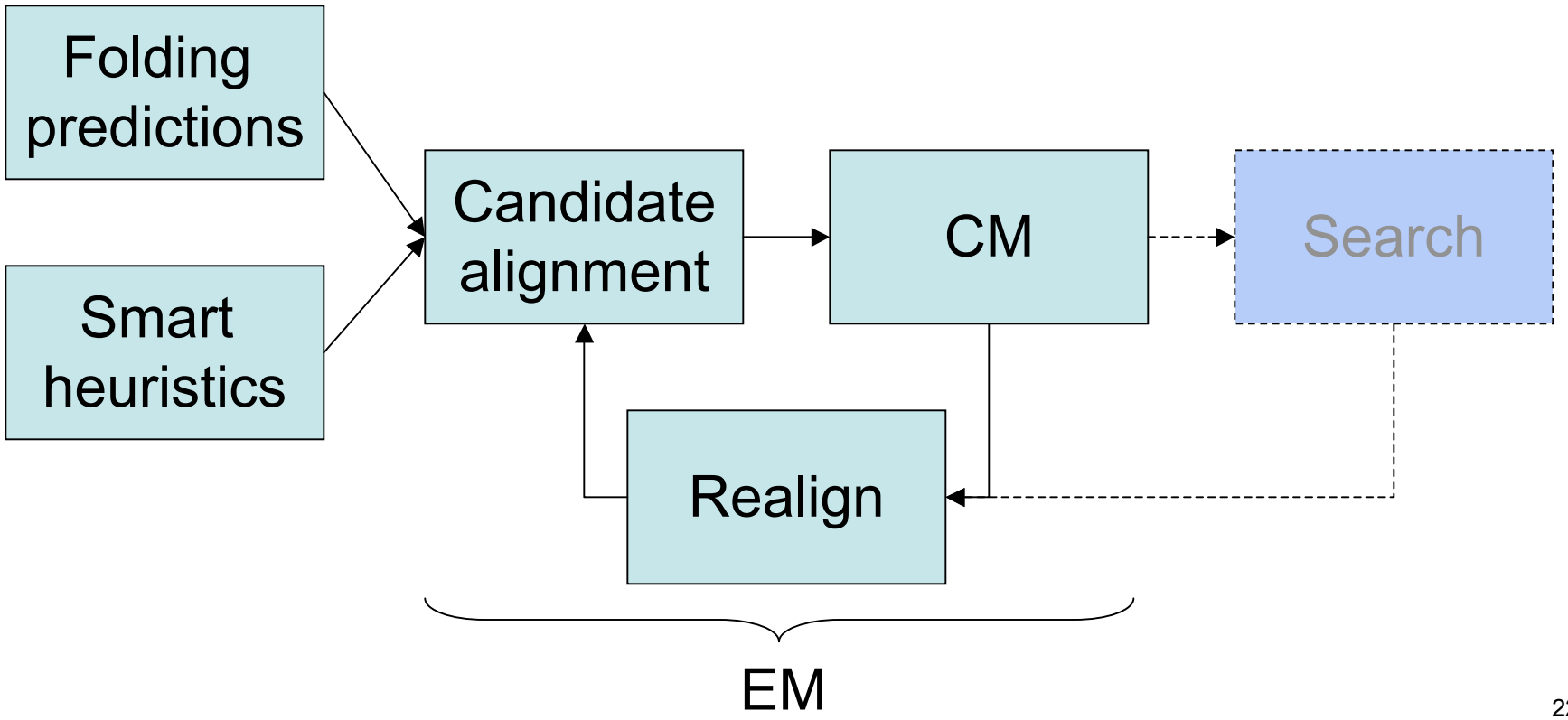
Largely from Eddy & Durbin, ‘94

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

CMFinder

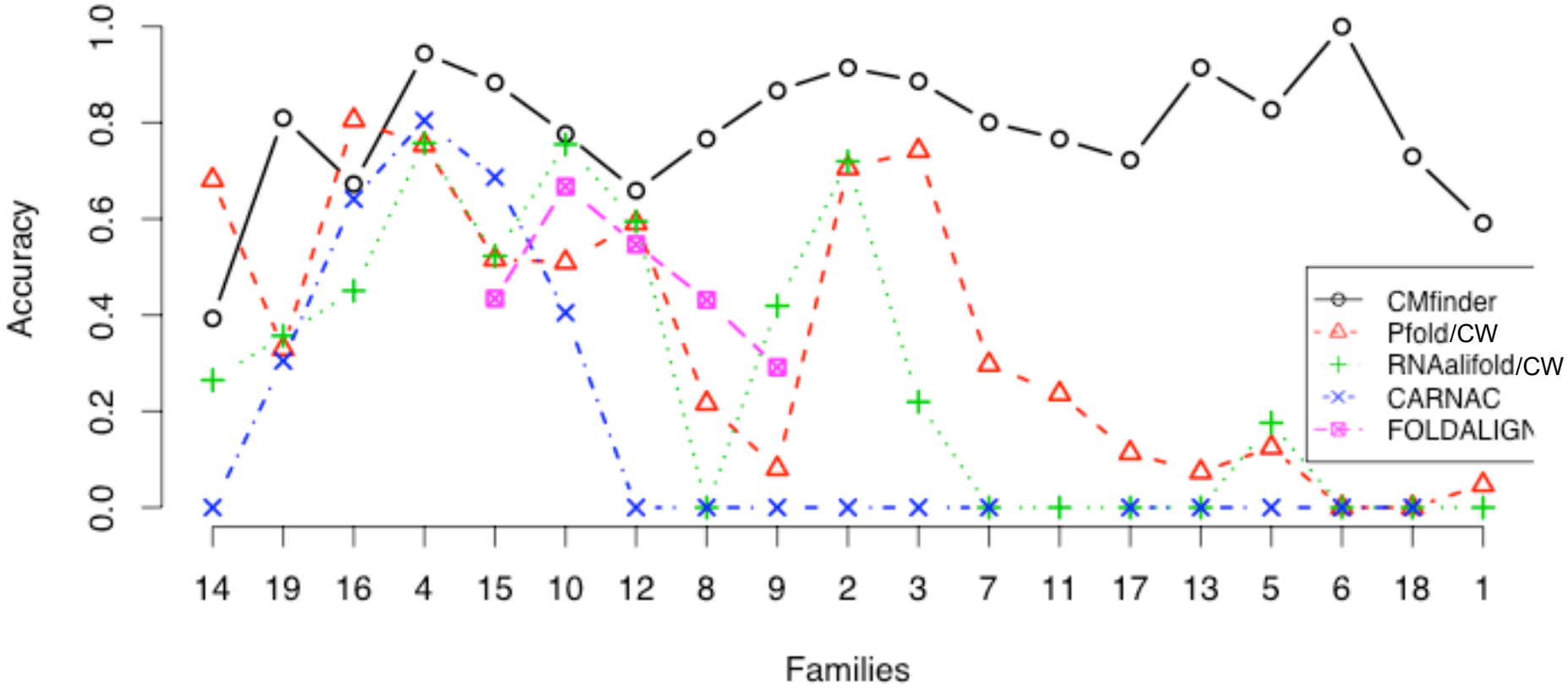
Harder: Finding CMs *without* alignment

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006



CMfinder Accuracy

(on Rfam families *with* flanking sequence)



Task 3: Application

Genome-wide search for
cis-regulatory RNA elements
(in prokaryotes, initially)

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

Approach

Choose a bacterial genome

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)

Genome Scale Search: Why

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

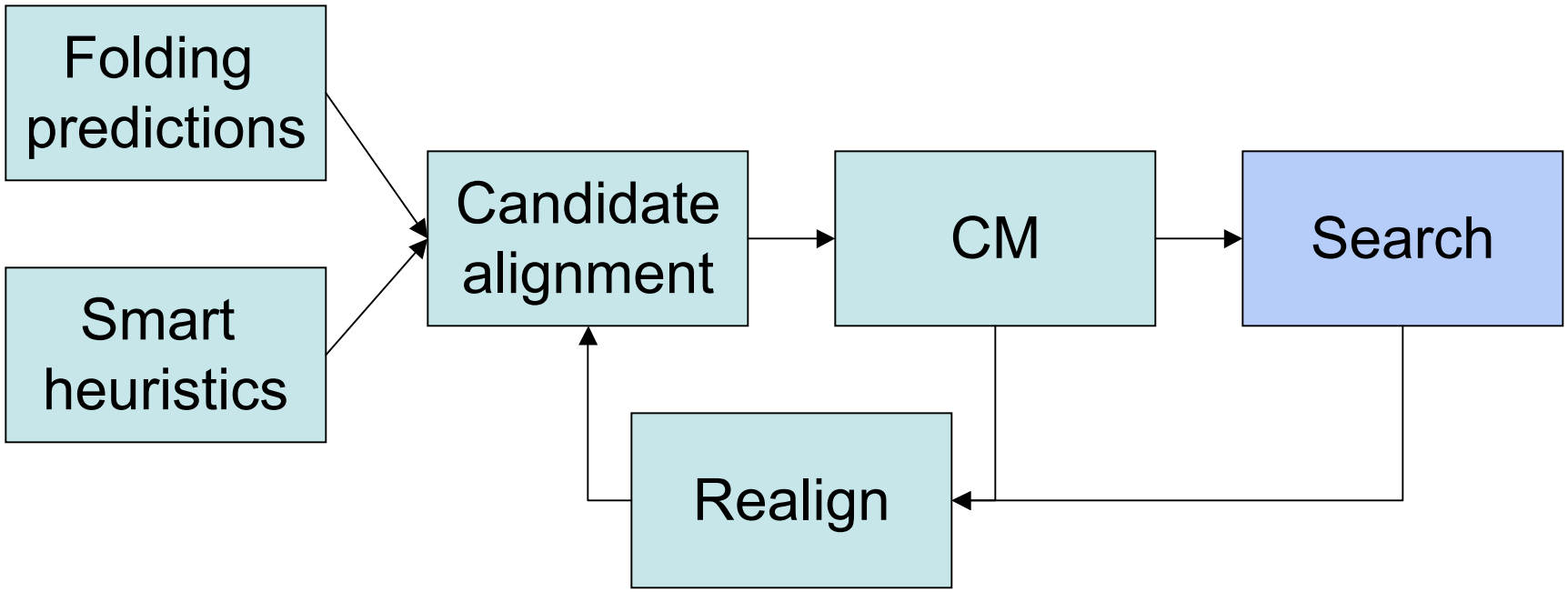
Throughout (most of) clade

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

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

Genome Scale Search

CMfinder is directly usable for/with search



Results

Process largely complete in

- bacillus/clostridia
- gamma proteobacteria
- cyanobacteria
- actinobacteria

Analysis ongoing

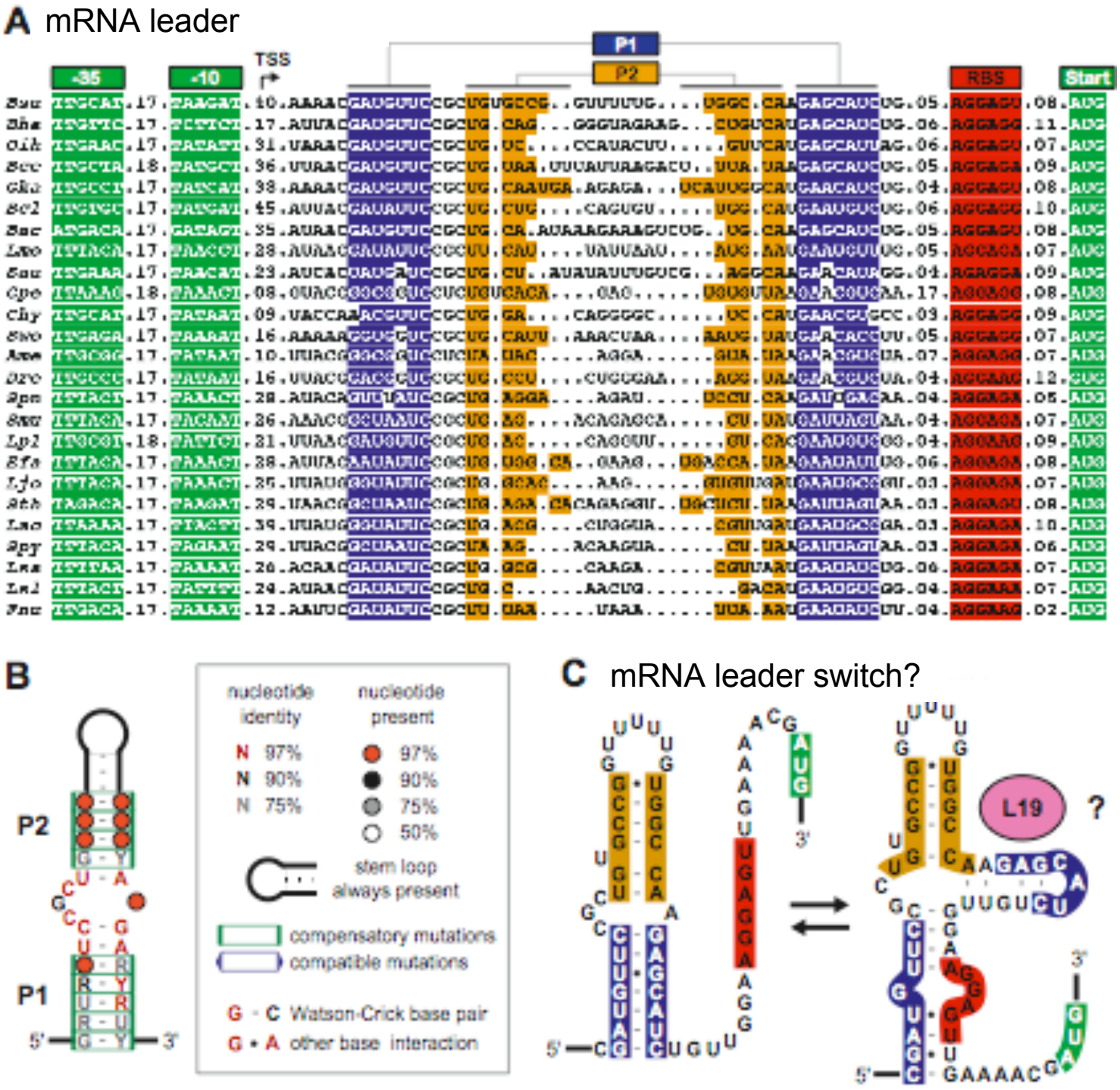
Some Preliminary Actino Results

8 of 10 Rfam families found

Rfam Family	Type (metabolite)	Rank
THI	riboswitch (thiamine)	4
ydaO-yuaA	riboswitch (unknown)	19
Cobalamin	riboswitch (cobalamin)	21
SRP_bact	gene	28
RFN	riboswitch (FMN)	39
yybP-ykoY	riboswitch (unknown)	48
gcvT	riboswitch (glycine)	53
S_box	riboswitch (SAM)	401
tmRNA	gene	Not found
RNaseP	gene	Not found

More Prelim Actino Results

- Many others (not in Rfam) are likely real of top 50:
  - known (Rfam, 23S) 10
  - probable (Tbox, CIRCE, LexA, parP, pyrR) 7
  - probable (ribosomal genes) 9
  - potentially interesting 12
  - unknown or poor 12
- One bench-verified, 2 more in progress


A mRNA leader
P1
P2
TSS
-35
-10
RBS
Start
B
P2
P1
5'
3'
nucleotide identity
nucleotide present
N 97%
N 90%
N 75%
97%
90%
75%
50%
stem loop always present
compensatory mutations
compatible mutations
G - C Watson-Crick base pair
G • A other base interaction
C mRNA leader switch?
L19
?
5'
3'

Ongoing & Future Work

Still automating a few steps, e.g. identifying duplicates

Improved ranking/motif significance stats

Better ortholog clustering

Performance & scale-up

Eukaryotic mRNAs, e.g. UTRs

Acknowledgements

Zizhen Yao (UW)

Zasha Weinberg (UW→Yale)

Shane Neph (UW)

Martin Tompa (UW)

Ron Breaker (Yale)

Jeff Barrick (Yale)