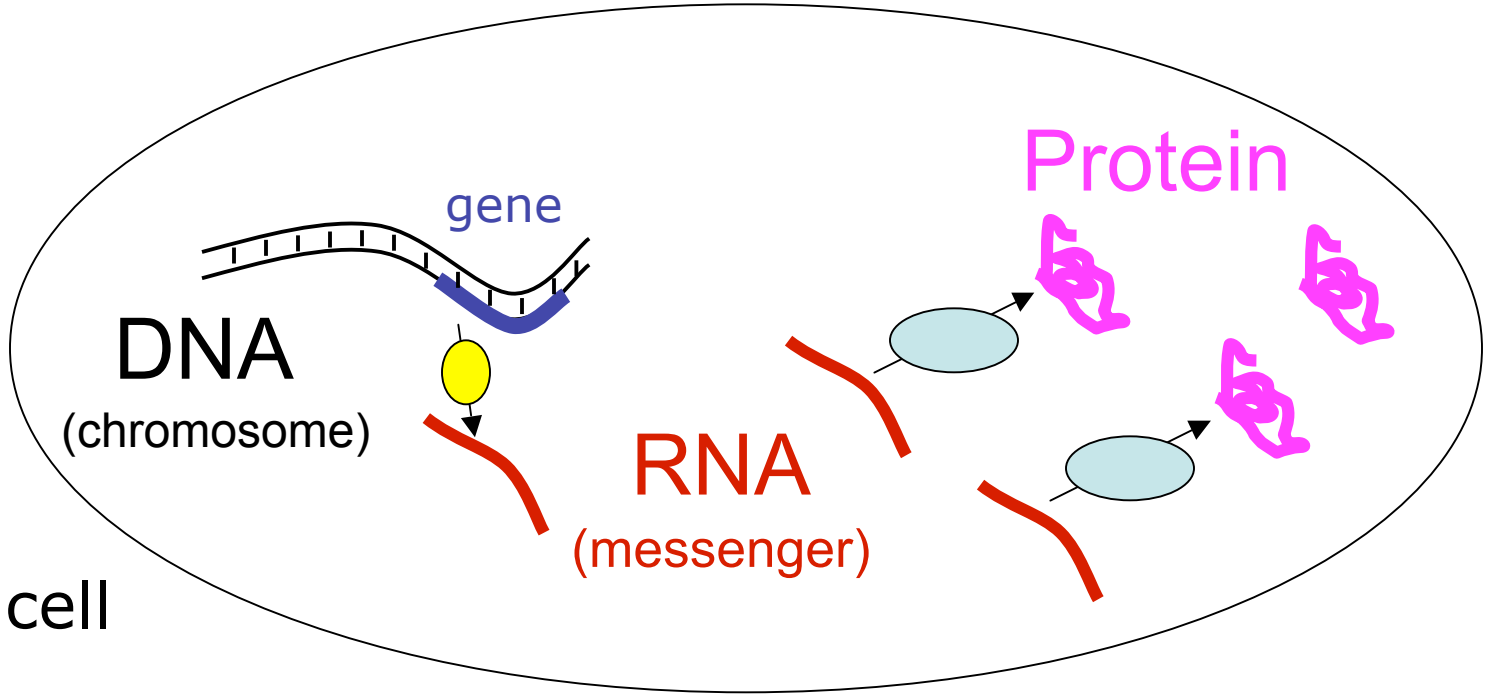
Modeling and Searching for Non-Coding RNA

W.L. Ruzzo

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

The “Central Dogma”

DNA → RNA → Protein



“Classical” RNAs

- mRNA
- tRNA
- rRNA
- snRNA (small nuclear - splicing)
- snoRNA (small nucleolar - guides for t/rRNA modifications)
- RNAseP (tRNA maturation; ribozyme in bacteria)
- SRP (signal recognition particle; co-translational targeting of proteins to membranes)
- telomerases

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
  - Regulation, transport, stability/degradation
  - E.g. "microRNA": ≈ 100's in humans
- *By some estimates, ncRNA >> mRNA*

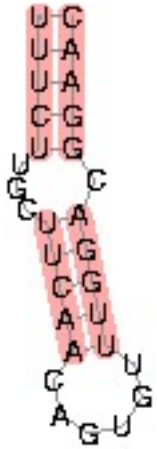
ncRNA Example: Xist

- large (12kb?)
- unstructured RNA
- required for X-inactivation in mammals

ncRNA Example: 6S

- medium size (175nt)
- structured
- highly expressed in e. coli in certain growth conditions
- sequenced in 1971; function unknown for 30 years

ncRNA Example: IRE



Iron Response Element: a short conserved stem-loop, bound by iron response proteins (IRPs). Found in UTRs of various mRNAs whose products are involved in iron metabolism. E.g., the mRNA of ferritin (an iron storage protein) contains one IRE in its 5' UTR. When iron concentration is low, IRPs bind the ferritin mRNA IRE. repressing translation. Binding of multiple IREs in the 3' and 5' UTRs of the transferrin receptor (involved in iron acquisition) leads to increased mRNA stability. These two activities form the basis of iron homeostasis in the vertebrate cell.

ncRNA Example: MicroRNAs

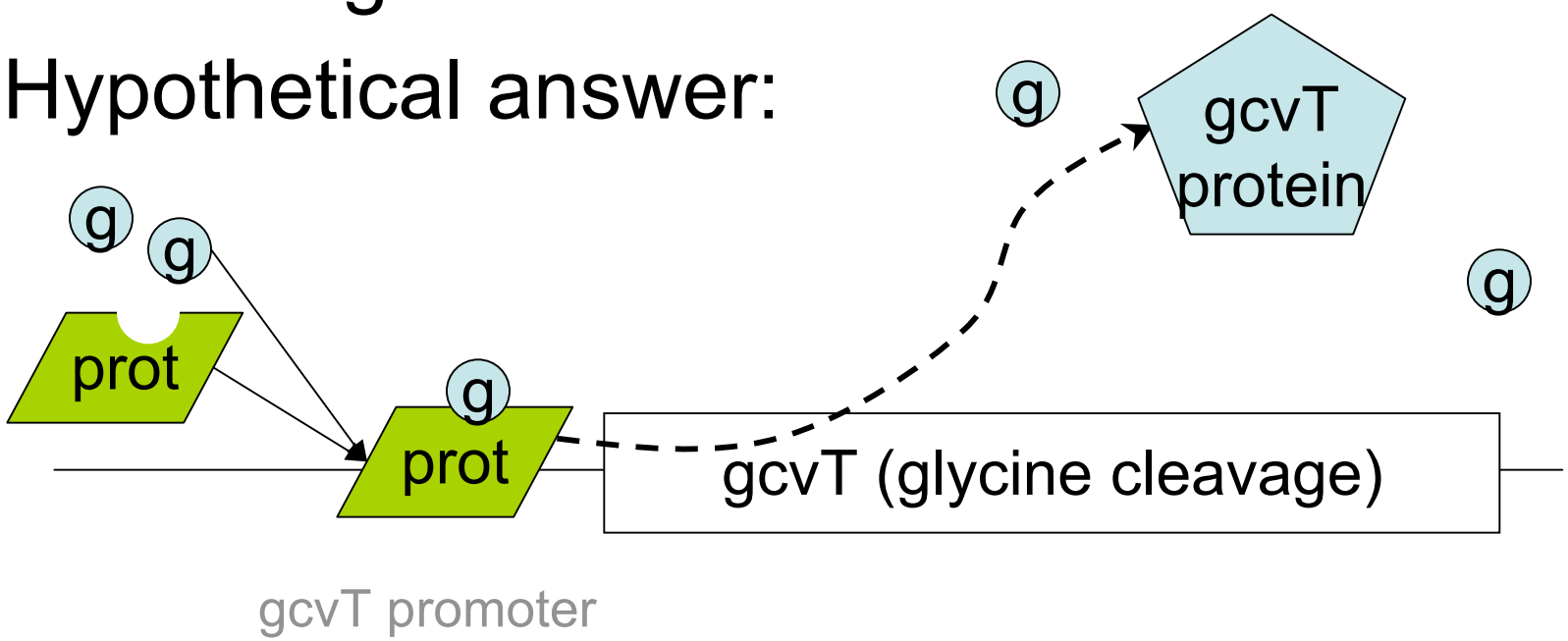
- short (~22 nt) unstructured RNAs excised from ~75nt precursor hairpin
- approx antisense to mRNA targets, often in 3' UTR
- regulate gene activity, e.g. by destabilizing (plants) or otherwise suppressing (animals) message
- several hundred, w/ perhaps thousands of targets, are known

ncRNA Example: Riboswitches

- UTR structure that directly senses/binds small molecules & regulates mRNA
- widespread in prokaryotes
- some in eukaryotes

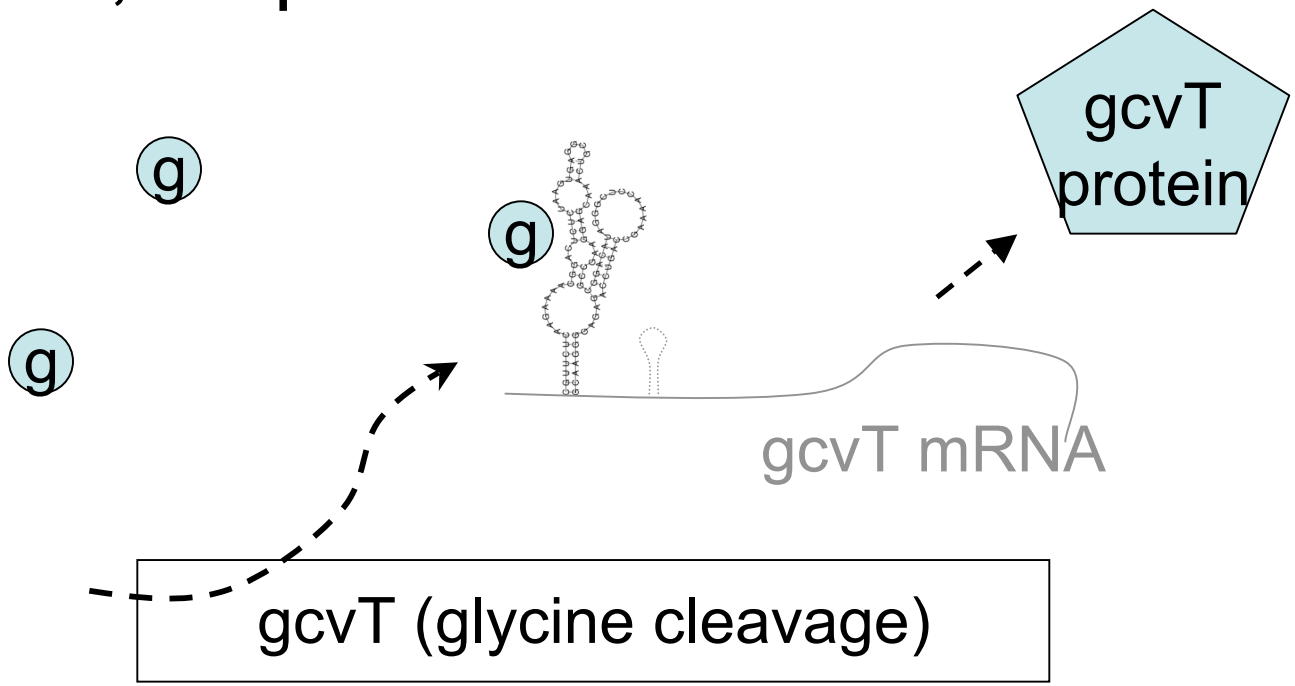
E.g.: the Glycine Riboswitch

- Glycine - simplest amino acid
- Uses - make proteins, make energy
- Not enough OR too much - wasteful
- Hypothetical answer:



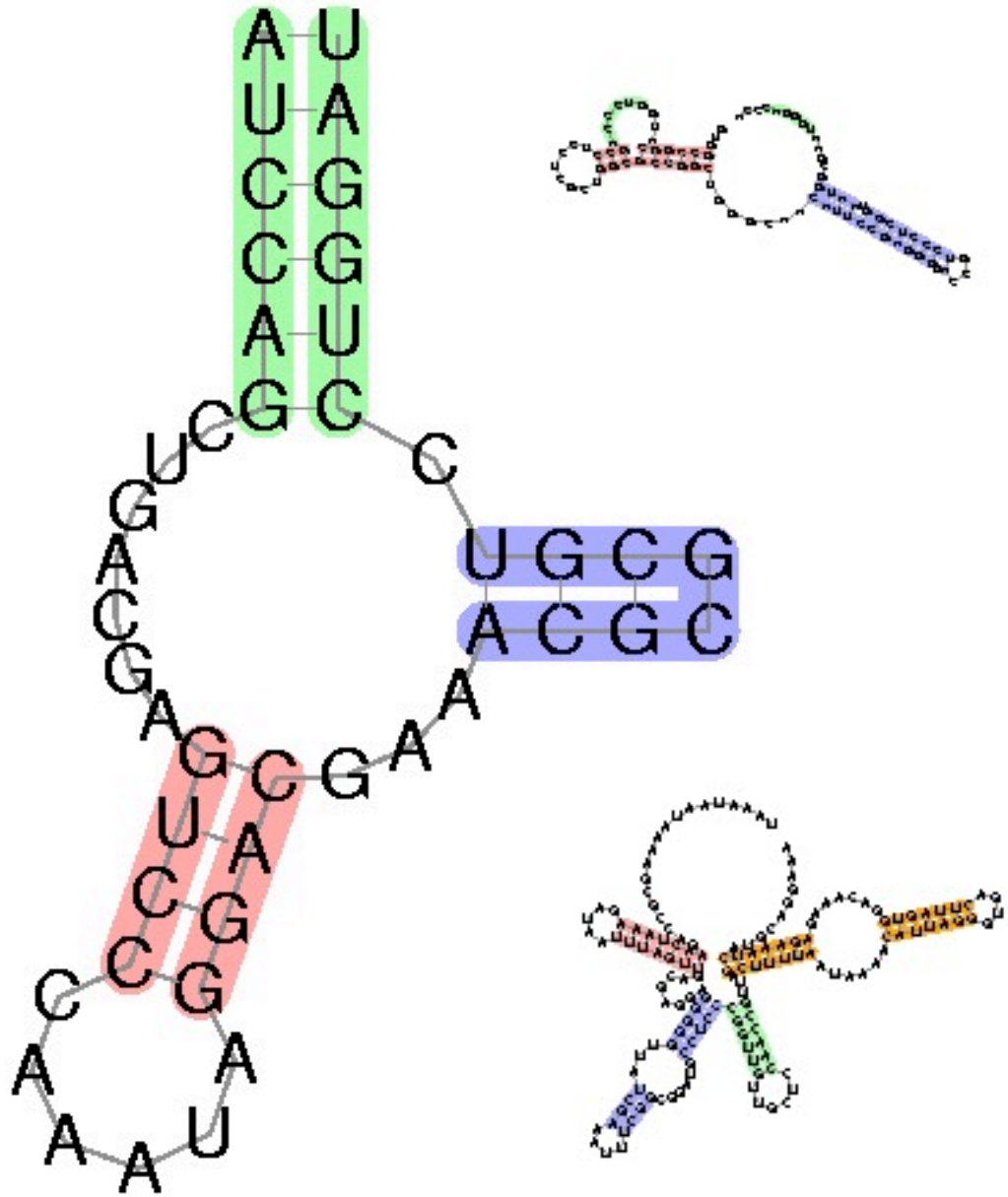
The Glycine Riboswitch

- Actual answer (in many bacteria): Look Ma, no protein

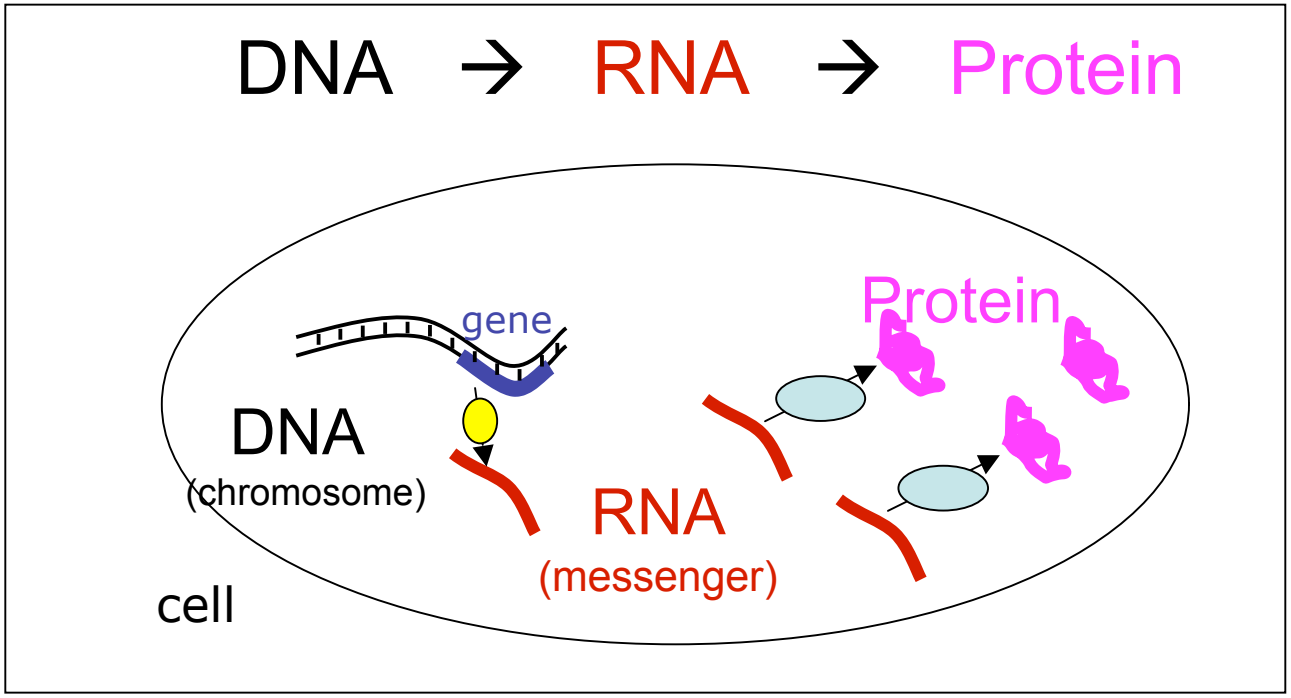


Why?

- RNA's fold, and function
- Nature uses what works



“Central Dogma”
=
“Central Chicken & Egg”?



Was there once an “RNA World”?

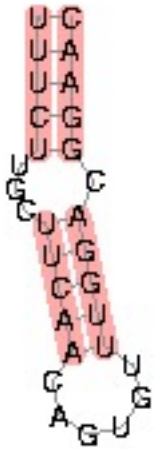
Outline

- ncRNA: what/why?
- What does computation bring?
- How to model and search for ncRNA?
- Faster search
- Better model inference

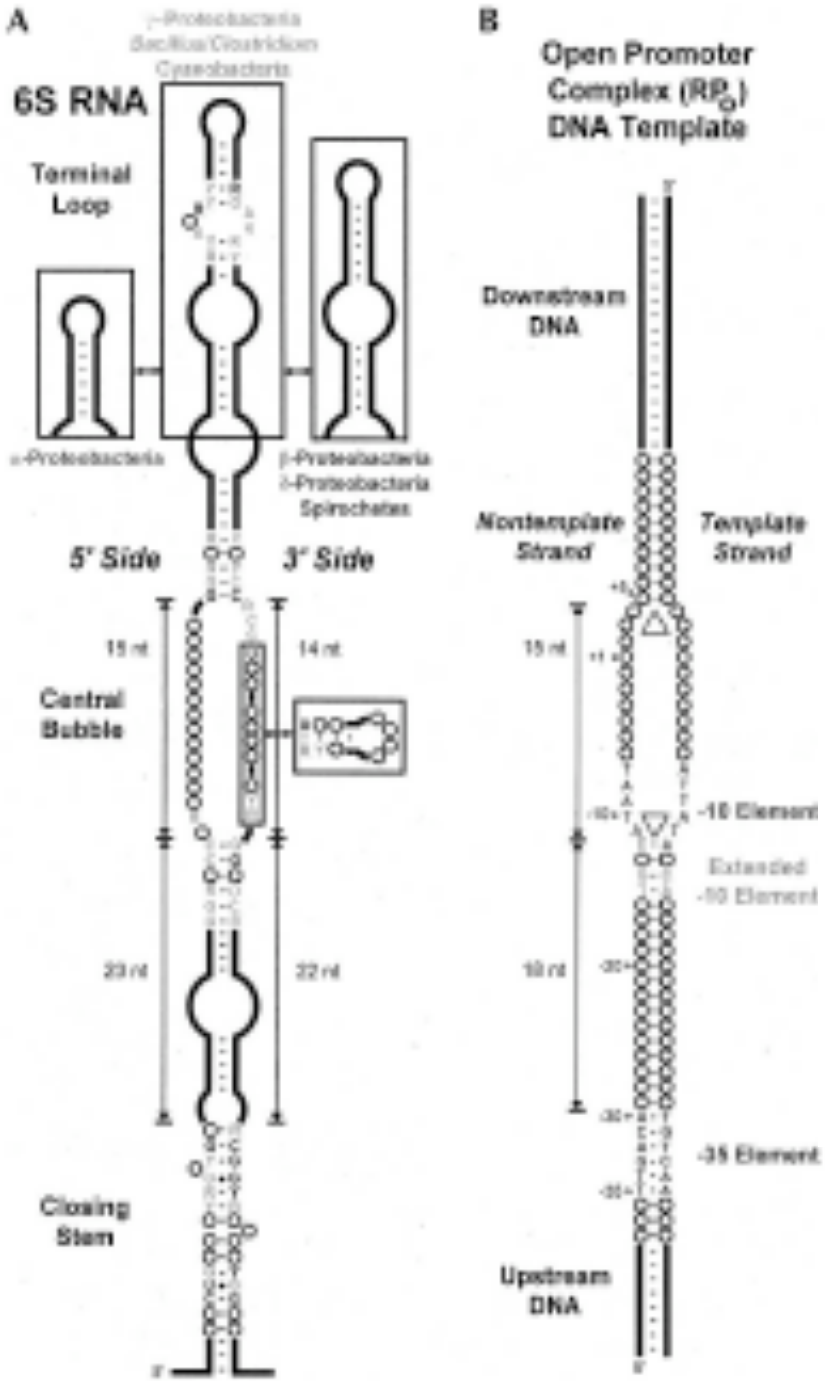
Iron Response Element

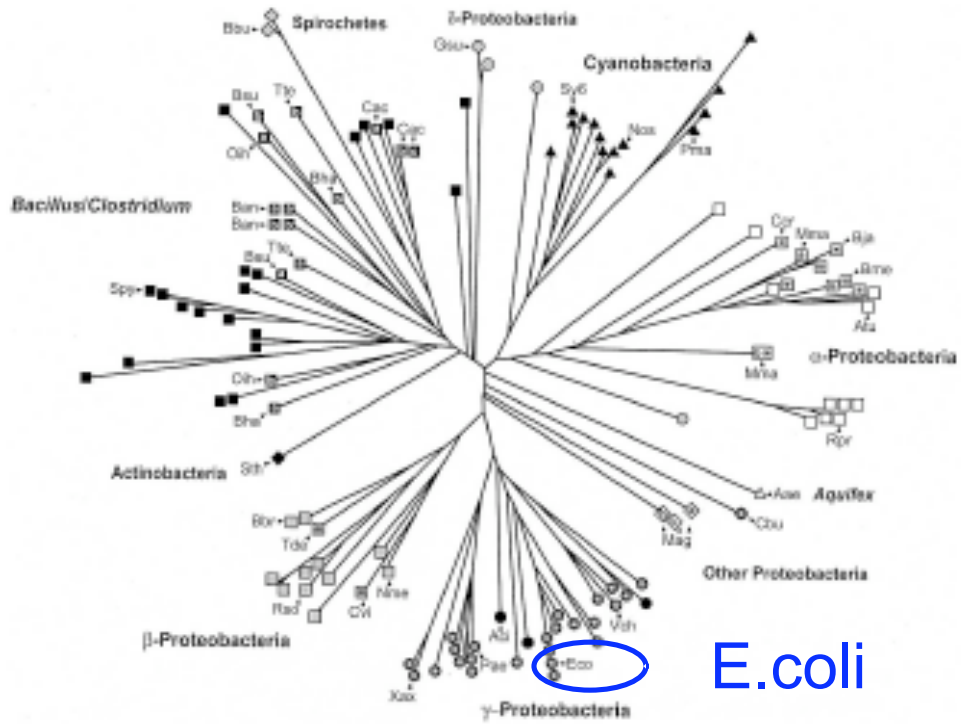
IRE (partial seed alignment):

```
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    <<<<<...<<<<<......>>>>>.>>>>>
```



6S mimics an open promoter





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

Dengue virus genome:

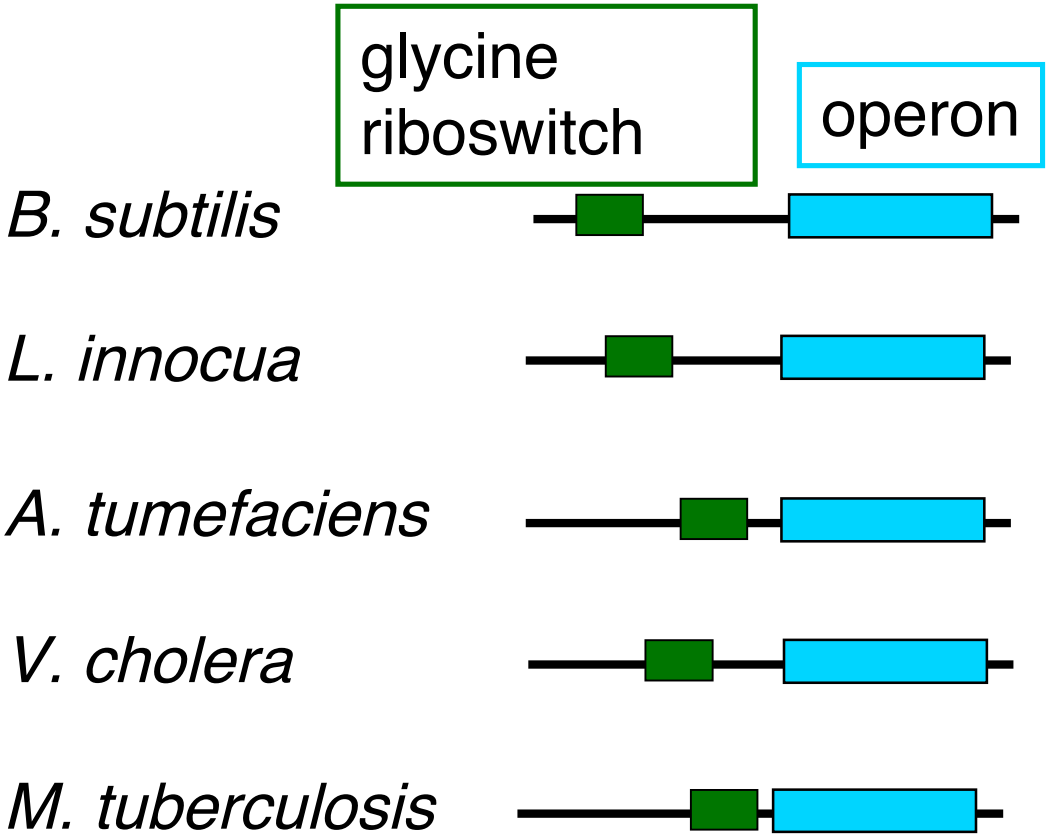
Known distribution (96% sequence identity)	With our techniques (70% sequence identity)
Dengue virus	Dengue virus West Nile virus Yellow Fever Omsk Hemorrhagic fever Japanese encephalitis Tick-borne encephalitis Kunjin virus Langat virus Louping ill Murray Valley virus Powassan virus

polyadenylation inhibition element RNA

Known distribution (90% sequence identity)	With our techniques (75% sequence identity)
Human, mouse, rabbit	Human, mouse, rabbit Zebrafish, *Tetraodon*, *Fugu* Frog

Impact of RNA homology search

(Barrick, *et al.,* 2004)



(and 19 more species)

Impact of RNA homology search

(Barrick, *et al.,* 2004)

glycine riboswitch

operon

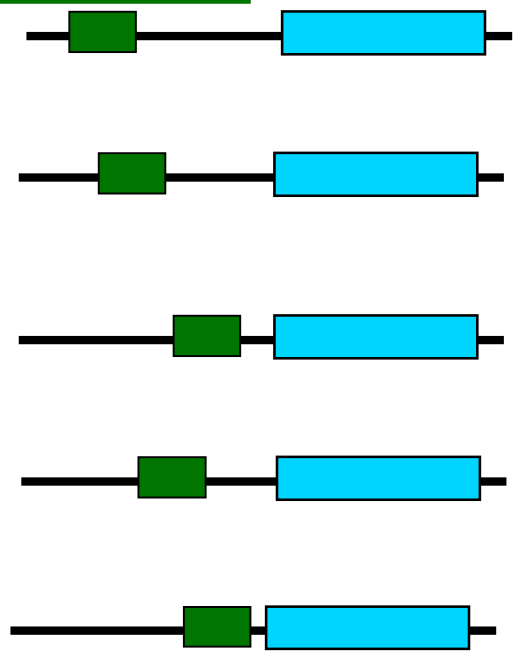
B. subtilis

L. innocua

A. tumefaciens

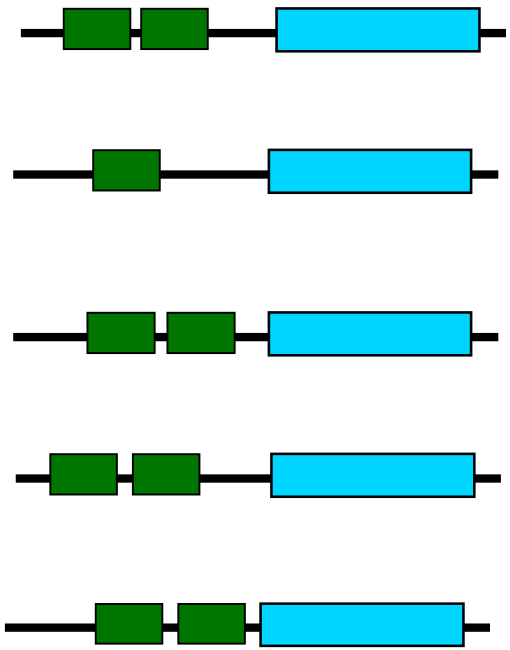
V. cholera

M. tuberculosis



(and 19 more species)

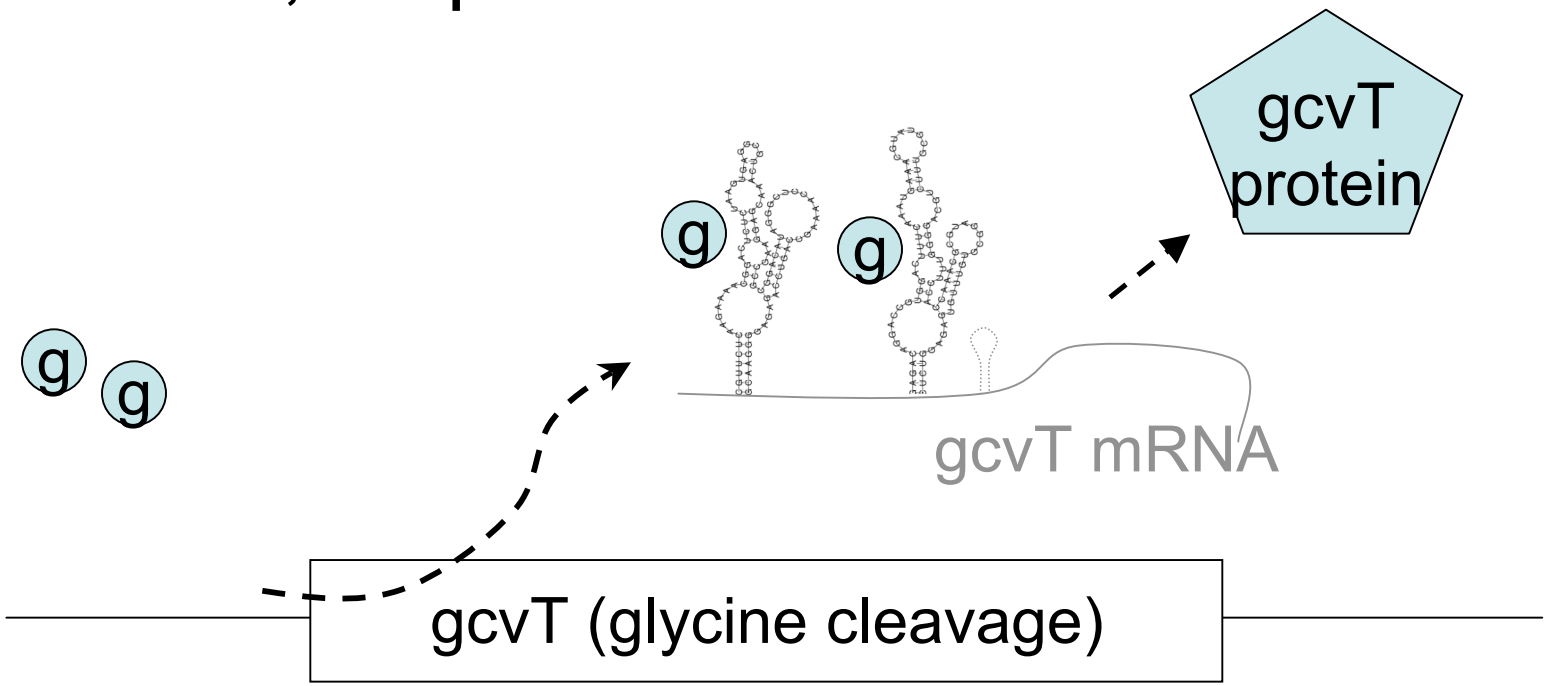
Using our techniques, we found…

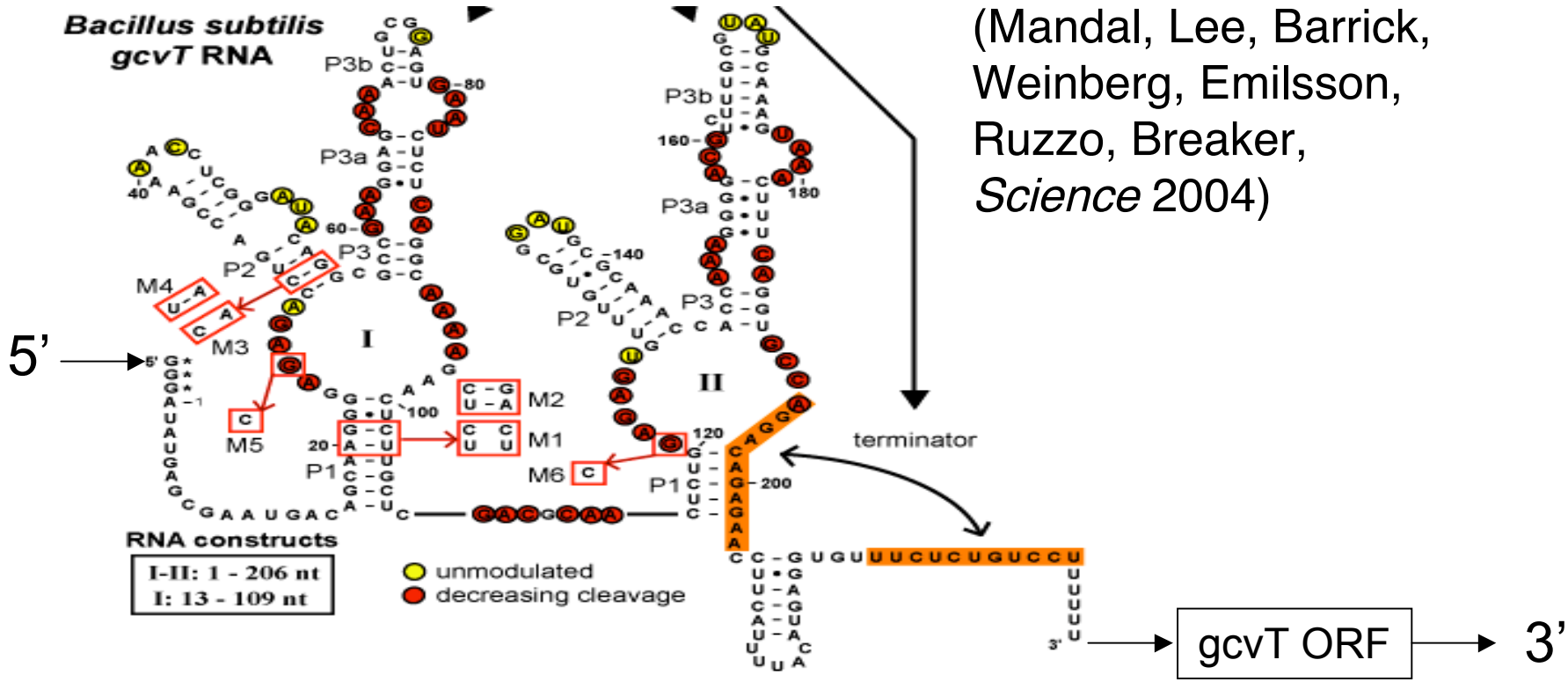


(and 42 more species)

The Glycine Riboswitch

- Actual answer (in many bacteria): Look Ma, no protein

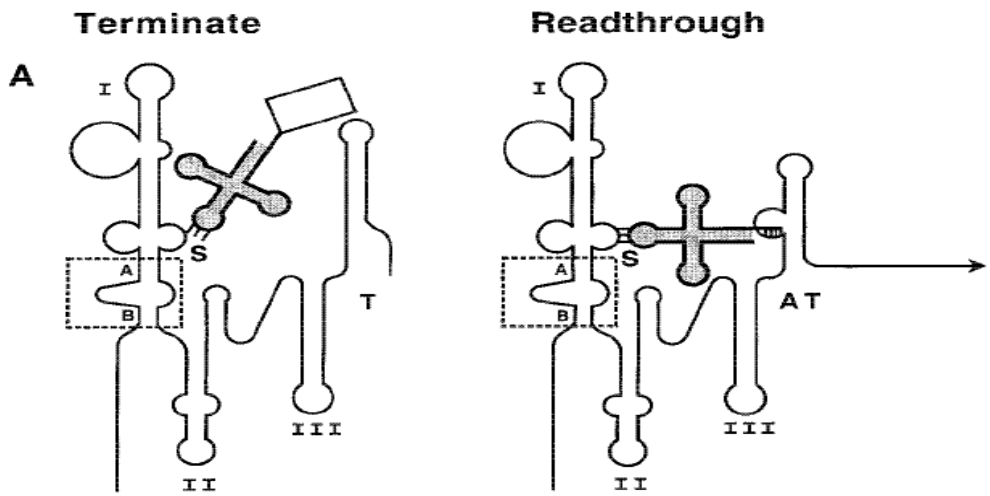


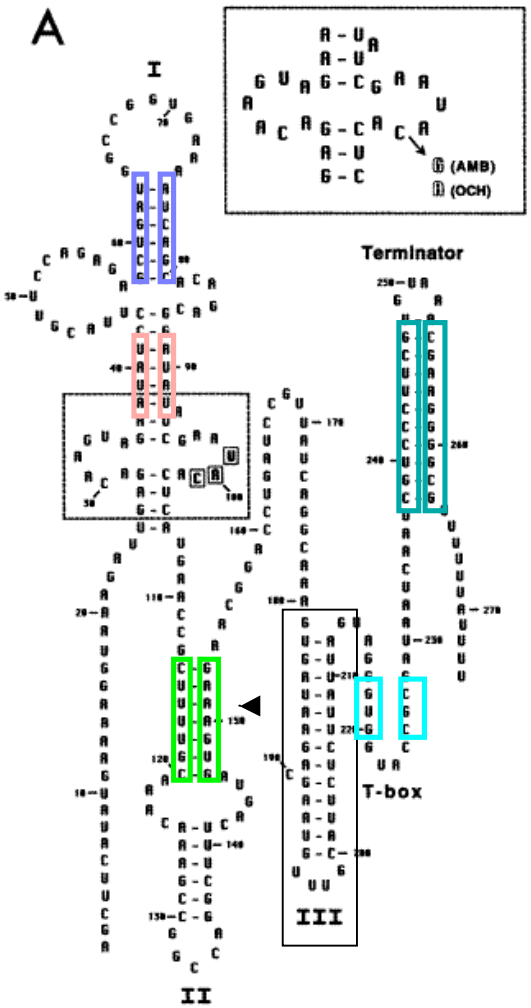


(Mandal, Lee, Barrick, Weinberg, Emilsson, Ruzzo, Breaker, *Science* 2004)

And…

- More examples means better alignment
- Understand phylogenetic distribution
- Find riboswitch in front of new gene



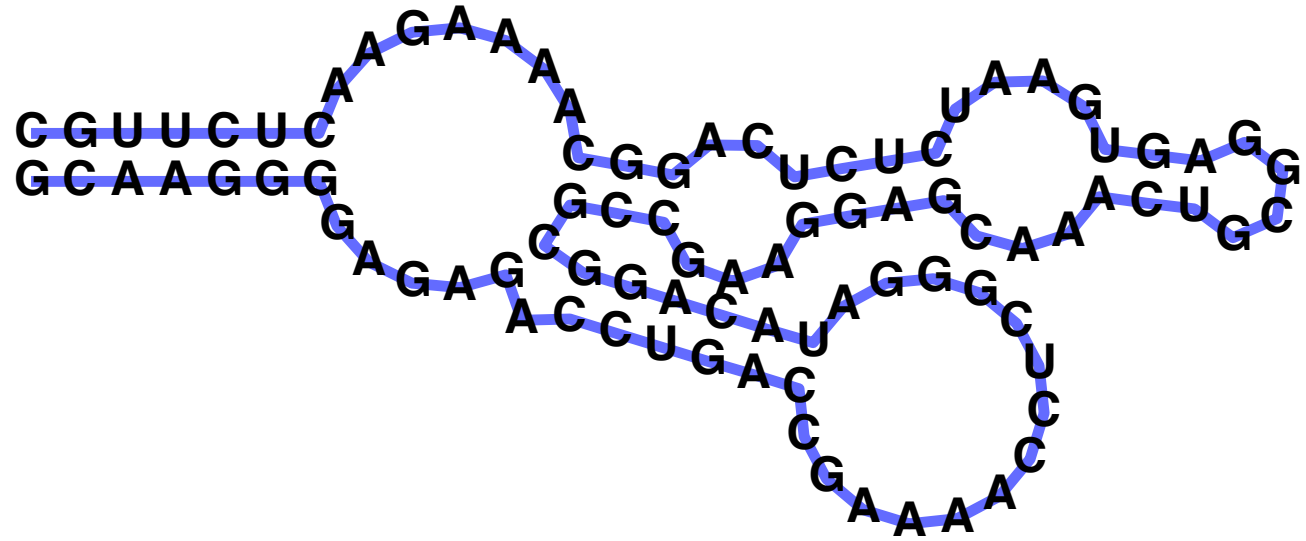


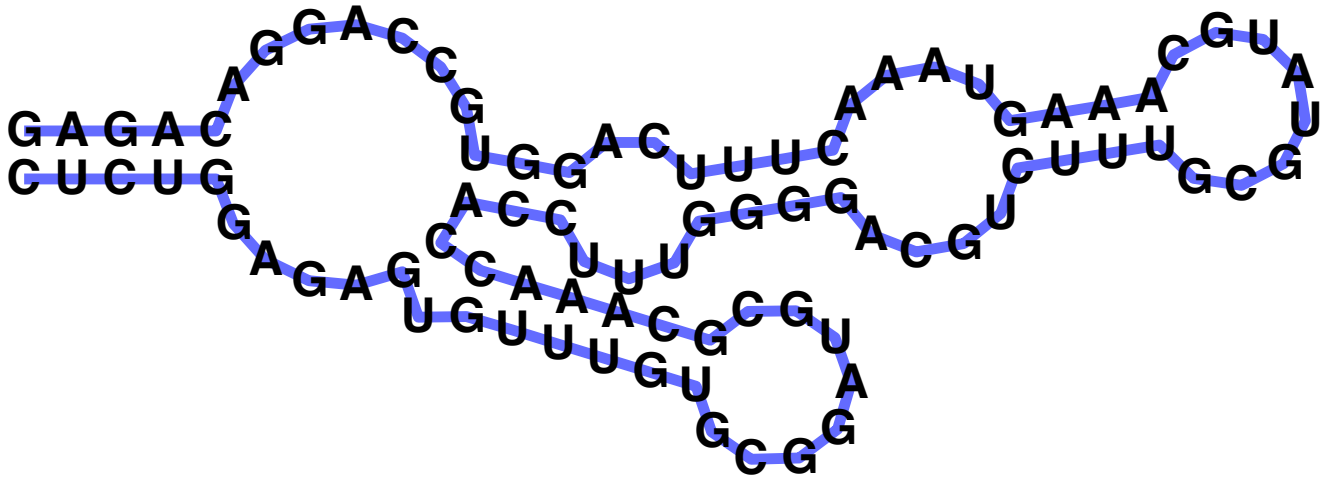

```
NC_000964.1 AUAUC.CUUACGU..UCCAGAGAGCUGAUGGCCGGUGAAA.AUCAGCACAGACGGAUAUAU
NC_004722.1 CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUUG
NC_004193.1 AAAAGUAGAACCG.AUCUAGCGAAUUGAGGAU.GGUGUGAGCUCAGUGC.GGAAAGCUUUU
NC_003997.3 CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUUG

NC_000964.1 CGAA..UACACUCAUGAACCGCUUUUGCAAACAAAGccggccaggcuuucAGUA.GUGAAAG
NC_004722.1 UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuuggauu.....AGUAAGCAUUCC
NC_004193.1 AGAAAAUC.ACUCUUGAGUU.UUCAUUACGAAA..CA.............AGUAGUAAUGGA
NC_003997.3 UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuaugauu.....AGUAAACAUUCC

NC_000964.1 aCGGAC.CUGAUCCGUUAUCAGGCAAAGUGGUACCGCGAUAAUCAAUCGUCCCUUCGUGUAAaCGAAGGGGCGUUU
NC_004722.1 .CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAuAGGGACGGGAGUU
NC_004193.1 .CGGUUcAUC.UCCGUUAUCGAUCUUAGUGGUACCGCGA.....GUCUUCUCGUCCCUUUU..GGGAUUAGAAGGC
NC_003997.3 .CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAuAGGGACGGGAGUU
```

RNA Informatics

- RNA: Not just a messenger anymore
  - Dramatic discoveries
  - Hundreds of families (besides classics like tRNA, rRNA, snRNA…)
  - Widespread, important roles
- Computational tools important
  - Discovery, characterization, annotation
  - BUT: slow, inaccurate, demanding

Q: What's so hard?





A: Structure often more important than sequence

Computational Challenges

- Search - given related RNA's, find more
- Modeling - describe a related family
- Meta-modeling - what's a good modeling framework?

- CM-based search
- Hand-curated alignments -> CMs
- Covariance Models

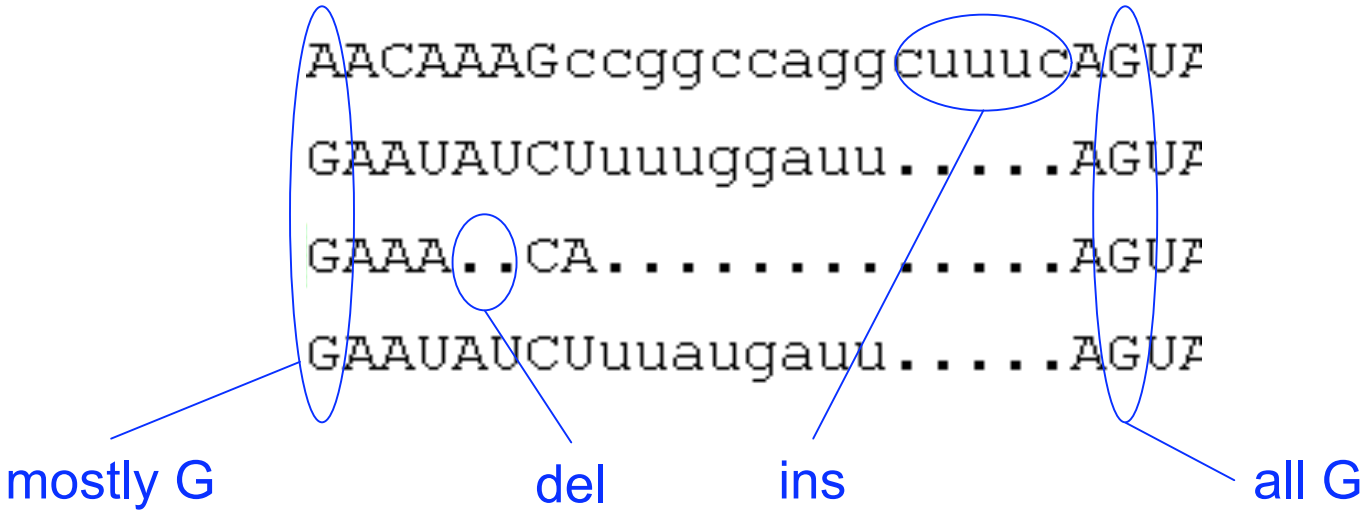
Predict Structure from Multiple Sequences

... GA ... UC ...
... GA ... UC ...
... GA ... UC ...
... CA ... UG ...
... CC ... GG ...
... UA ... UA ...

Compensatory mutations reveal structure, but in usual alignment algorithms they are doubly penalized.

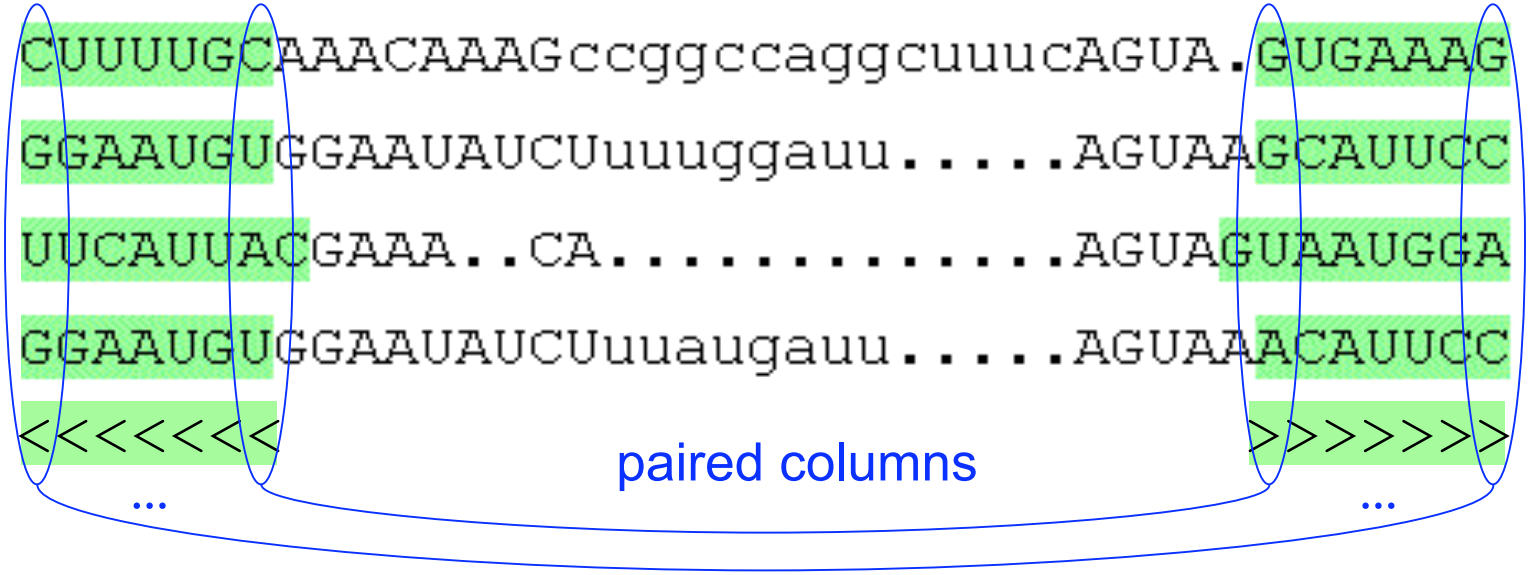
How to model an RNA “Motif”?

- 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”?

- Covariance Models (aka “profile SCFG”)
  - Probabilistic models, like profile HMMs, but adding “column pairs” and pair emission probabilities for base-paired regions



“RNA sequence analysis using covariance models”

Eddy & Durbin

Nucleic Acids Research, 1994
vol 22 #11, 2079-2088

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

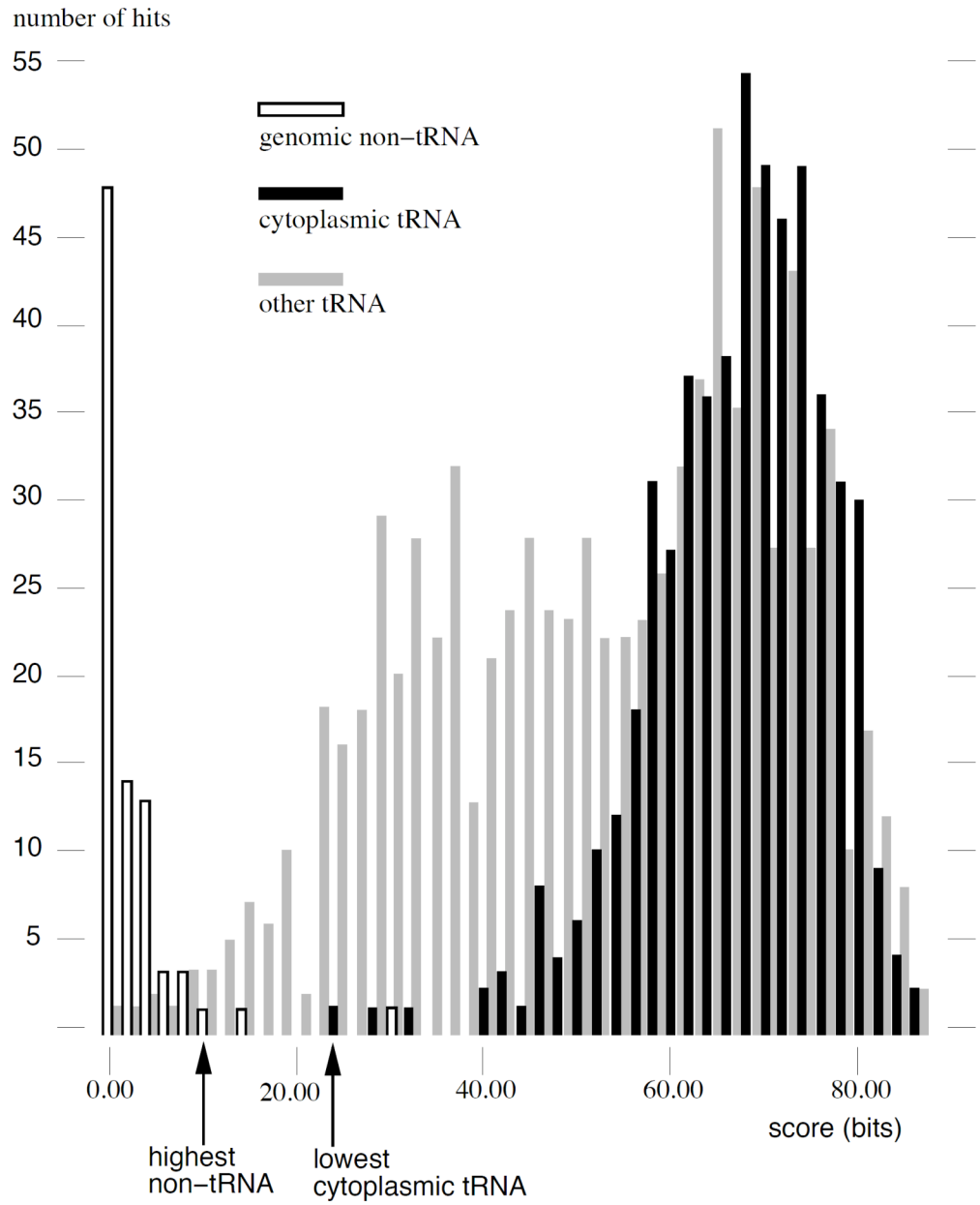
Main Results

- 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 llikelihood 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



Alignment Quality

Trusted:

```
DF6280    GCGGAUUUAGCUCAGUU GGG AGAGCGCCAGACUGAAG                  AUCUGGAG            GUCCUGUGUUCGAUCCACAGAAUUCGCACCA
DF6280G   GCGGAUUUAGCUCAGUU GGG AGAGCGCCAGACUGAAGAAAUACUUCGGUCAAGUUAUCUGGAG            GUCCUGUGUUCGAUCCACAGAAUUCGCA
DD6280    UCCGUGAUAGUUUAAU  GGUCAGAAUGGGCGCUUGUCG                  CGUGCCAG            A UCGGGGUUCAAUUCCCCGUCGCGGAGCCA
DX1661    CGCGGGGUGGAGCAGCCUGGU AGCUCGUCGGGCUCAUA                  ACCCGAAG            GUCGUCGGUUCAAAUCCGGCCCCCGCAACCA
DS6280    GGCAACUUGGCCGAGU  GGUUAAGGCGAAAGAUUAGAA                  AUCUUUU  GGGCUUUGCCCG CGCAGGUUCGAGUCCUGCAGUUGUCGCCA
```

U100:

```
DF6280    GCGGAUUUAGCUCAG UUGGGAGAGCGCCAGACU           GA   AG    AUCUGGA           GGUCCUGUGUUCGAUCCACAGAAUUCGCAcca
DF6280G   GCGGAUUUAGCUCAG UUGGGAGAGCGCCAGACUgaagaaauacuUCgguCAaguuAUCUGGA           GGUCCUGUGUUCGAUCCACAGAAUUCGCA
DD6280    UCCGUGAUAGUUUAA UGGUCAGAAUGGGCGCUU           GU   CG    CGUGCCA           GAU CGGGGUUCAAUUCCCCGUCGCGGAGcca
DX1661    CGCGGGGUGGAGCAGcCUGGUAGCUCGUCGGGCU           CA   UA    ACCCGAA           GGUCGUCGGUUCAAAUCCGGCCCCCGCAAcca
DS6280    GGCAACUUGGCCGAG UGGUUAAGGCGAAAGAUU           AG   AA    AUCUUUUgggcuuugcccG  CGCAGGUUCGAGUCCUGCAGUUGUCGcca
```

ClustalV:

```
DF6280    GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGA                 UCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
DF6280G   GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAAAUACUUCGGUCAAGUUAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCA
DD6280    UCCGUGAUAGUUUAAU       G GUCAGAAUGG GCG    CUUG  UCGCGUGCC   AGAUCGG GGUUCAAUUCCCCGUCGCGGAGCCA
DX1661    CGCGGGGUGGAGCAGC       CUGGUAGCUCGUCGGG    CUCA  UAACCCGA    AGGUCGUCGGUUCAAAUCCGGCCCCCGCAACCA
DS6280    GGCAACUUGGCCGAGUGGUUAAGGCGAAAGAUU  AGAAAUCUUUUGGGC  UUUGCCCG     CGCAGGUUCGAGUCCUGCAGUUGUCGCCA
```

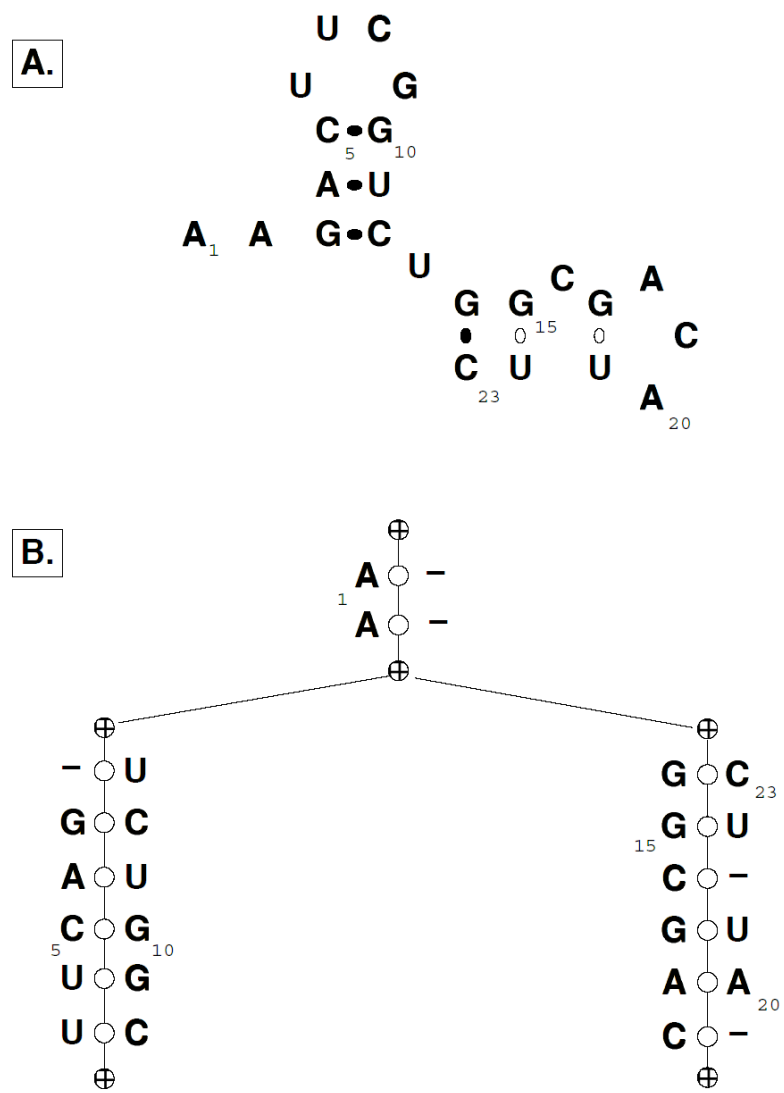
Comparison to TRNASCAN

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

Slightly different evaluation criteria

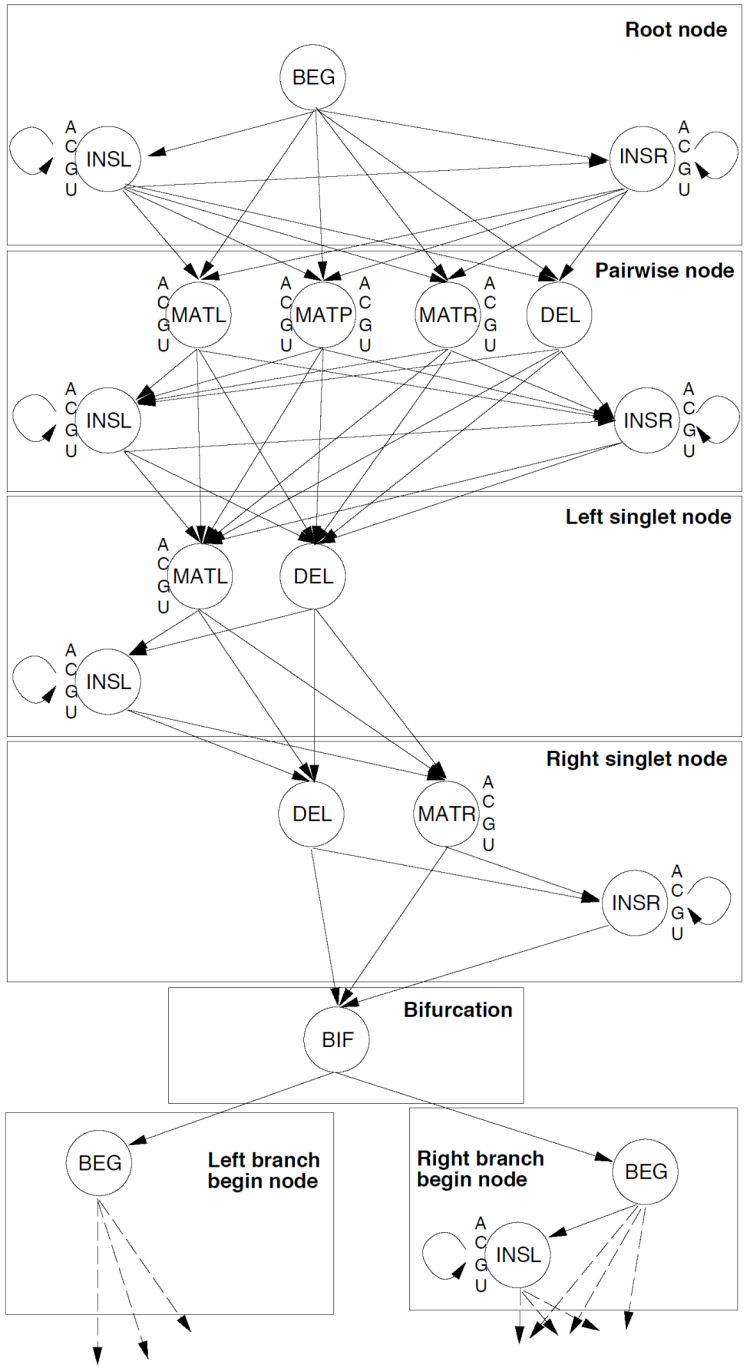
CM Structure

- 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

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

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

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

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

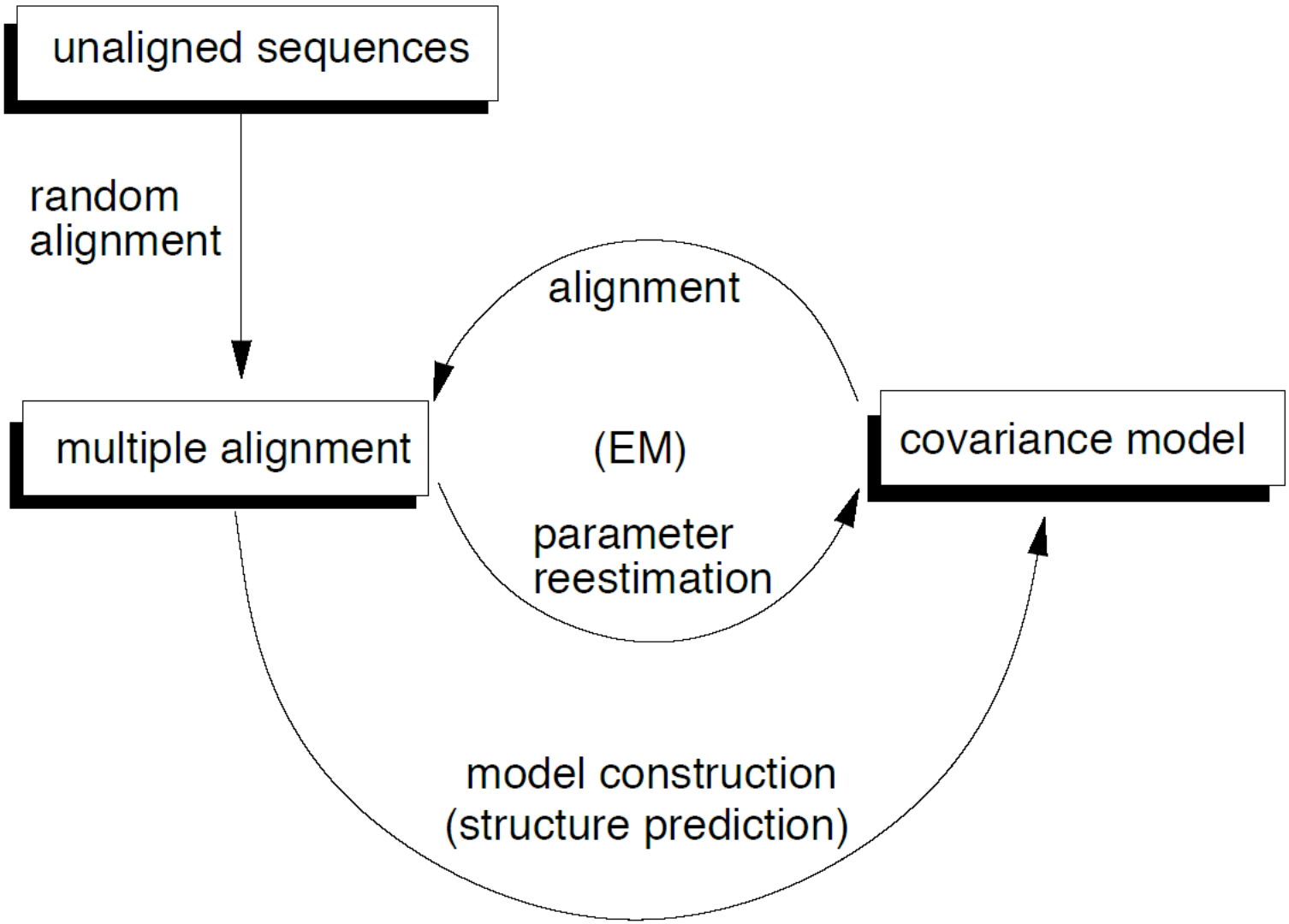
$S^{y}_{ij} = \max_{\pi} \log P(x_{ij} \text{ generated starting in state } y \text{ via path } \pi)$

Viterbi, cont.

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

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

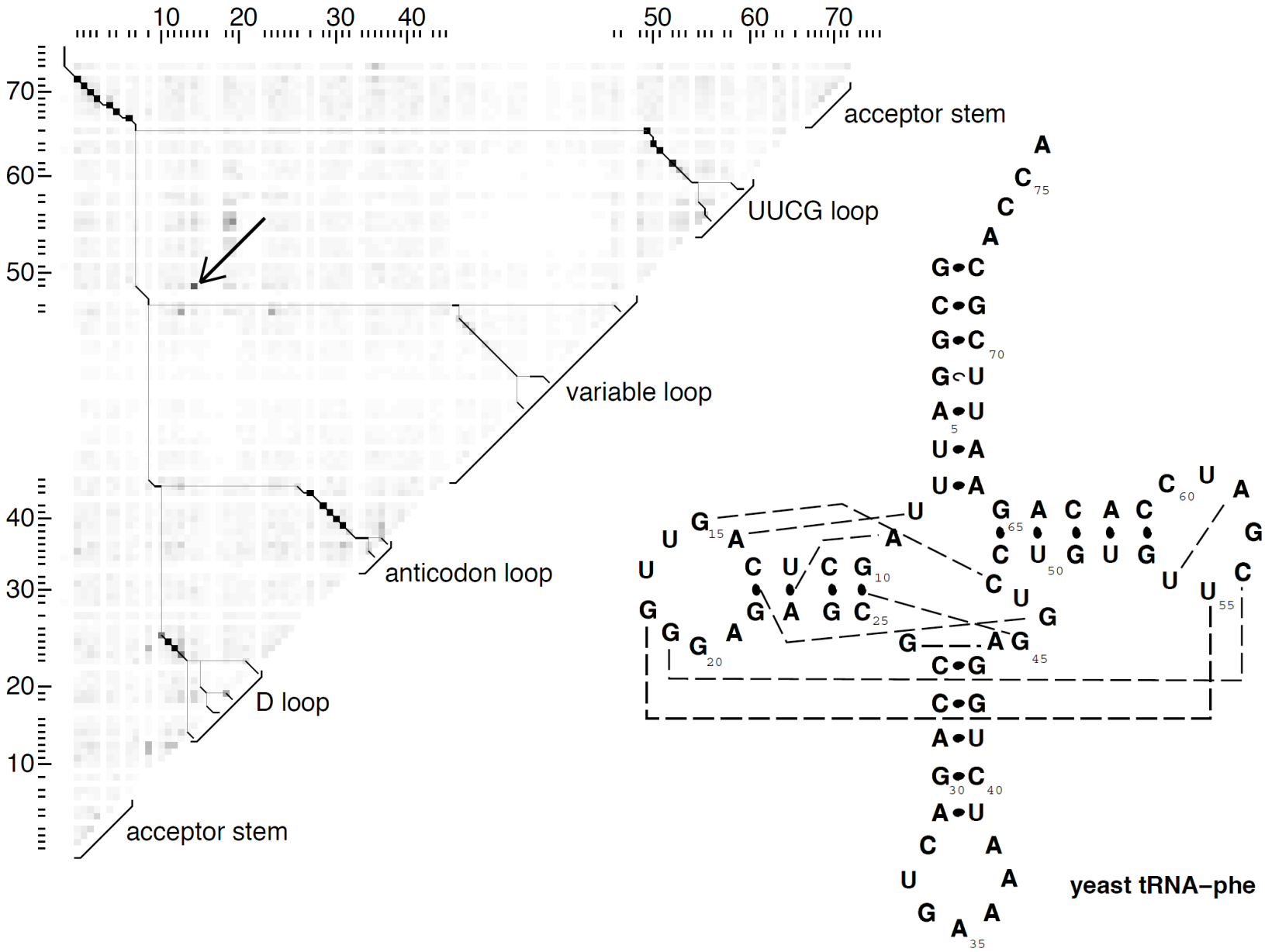
Model Training



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* sequence conservation but perfect pairing
- MI = expected score gain from using a pair state
- Finding optimal MI, (i.e. optimal pairing of columns) is NP-hard(?)
- Finding optimal MI *without pseudoknots* can be done by dynamic programming


acceptor stem
UUCG loop
variable loop
anticodon loop
D loop
acceptor stem
yeast tRNA–phe

MI-Based Structure-Learning

$$S_{i,j} = \max \begin{cases} S_{i+1,j} \\ S_{i,j-1} \\ S_{i+1,j-1} + M_{i,j} \\ \max_{i<j<k} S_{i,k} + S_{k+1,j} \end{cases}$$

- “just like Nussinov/Zucker folding”
- BUT, need enough data---enough sequences at right phylogenetic distance

Pseudoknots

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

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

Table 1: Statistics of the training and test sets of 100 tRNA sequences each. The average identity in an alignment is the average pairwise identity of all aligned symbol pairs, with gap/symbol alignments counted as mismatches. Primary sequence information content is calculated according to [48]. Calculating pairwise mutual information content is an NP-complete problem of finding an optimum partition of columns into pairs. A lower bound is calculated by using the model construction procedure to find an optimal partition subject to a non-pseudoknotting restriction. An upper bound is calculated as sum of the single best pairwise covariation for each position, divided by two; this includes all pairwise tertiary interactions but overcounts because it does not guarantee a disjoint set of pairs. For the meaning of multiple alignment accuracy of ClustalV, see the text.

Model	training set	iterations	score (bits)	alignment accuracy
A1415	all sequences (aligned)	3	58.7	95%
A100	SIM100 (aligned)	3	57.3	94%
A65	SIM65 (aligned)	3	46.7	93%
U100	SIM100 (degapped)	23	56.7	90%
U65	SIM65 (degapped)	29	47.2	91%

Table 2: Training and multiple alignment results from models trained from the trusted alignments (A models) and models trained from no prior knowledge of tRNA (U models).

Accelerating CM search

Zasha Weinberg

& W.L. Ruzzo

Recomb ‘04, Bioinformatics ‘04, ‘06

Rfam database

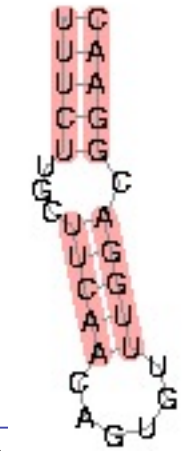
(Release 7.0, 3/2005)

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, …

Rfam

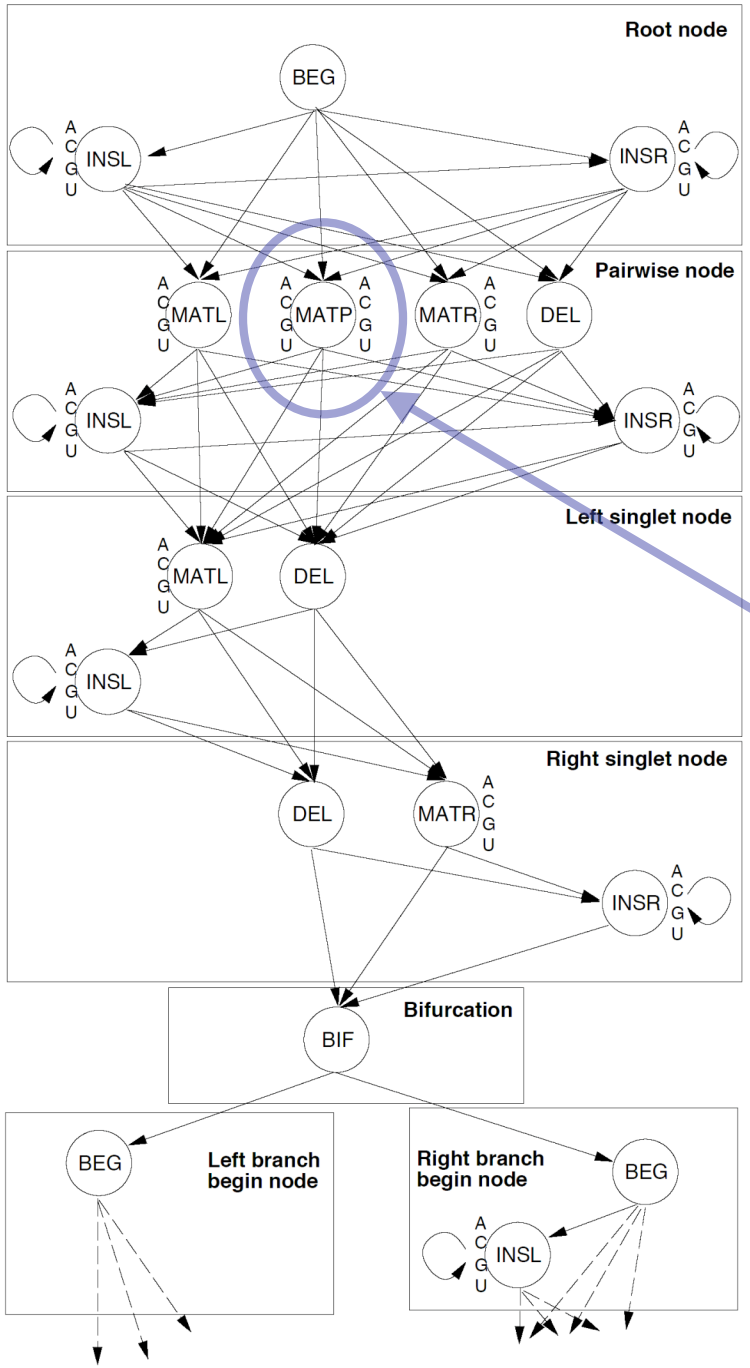


- Input (hand-tuned):
  - MSA
  - SS_cons
  - Score Thresh T
  - Window Len W
- Output:
  - CM
  - scan results

IRE (partial seed alignment):

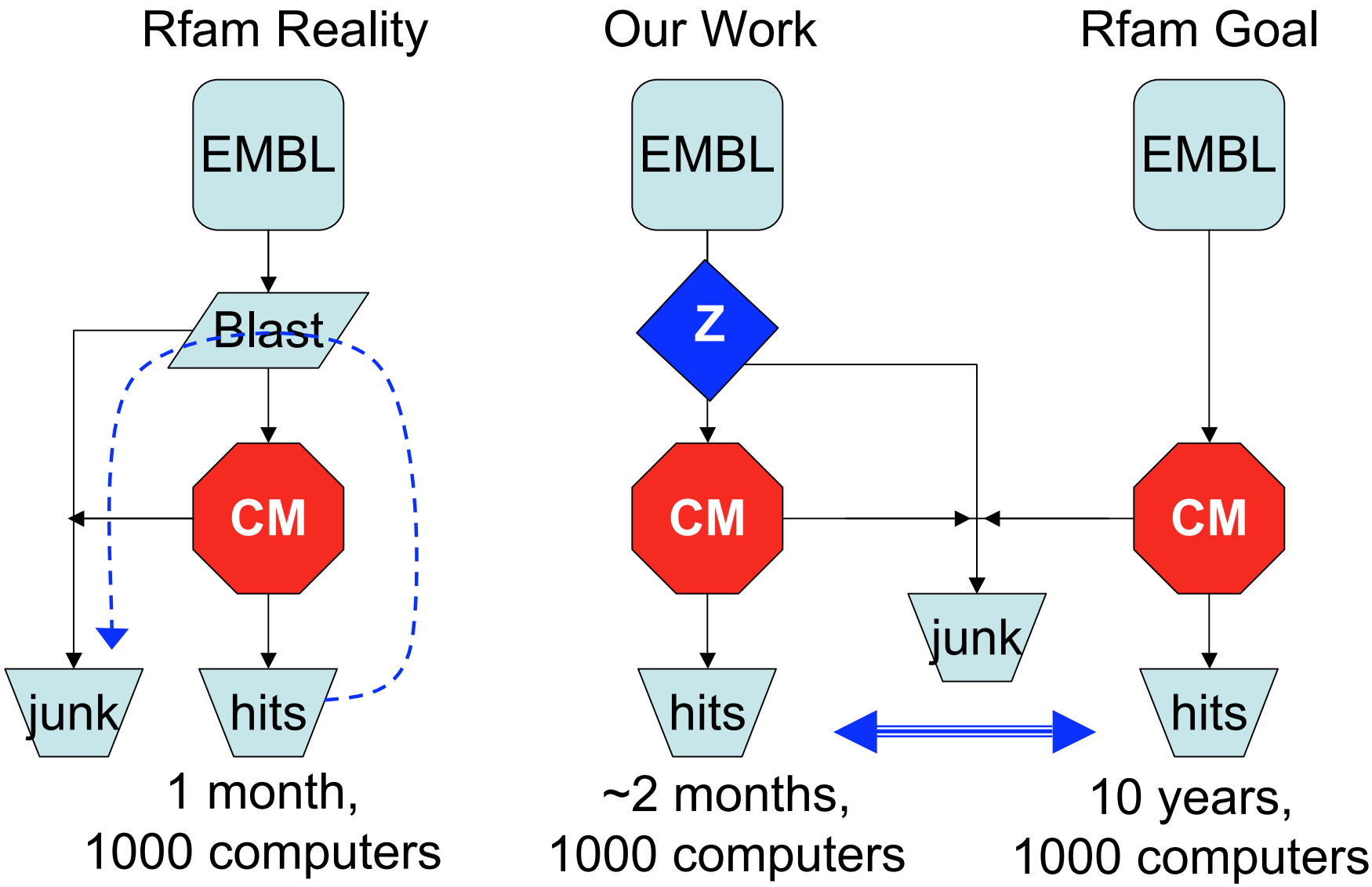
```
Hom.sap.   GUUCCUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCUUC.UUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCCUGUUUCAACAGUGCUUGGA.GGAAC
Hom.sap.   UUUAUC..AGUGACAGAGUUCACU.AUAAA
Hom.sap.   UCUCUUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   AUUAUC..GGGAACAGUGUUUCCC.AUAAU
Hom.sap.   UCUUGC..UUCAACAGUGUUUGGACGGAAG
Hom.sap.   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    <<<<<...<<<<<......>>>>>.>>>>>
```

Covariance Model

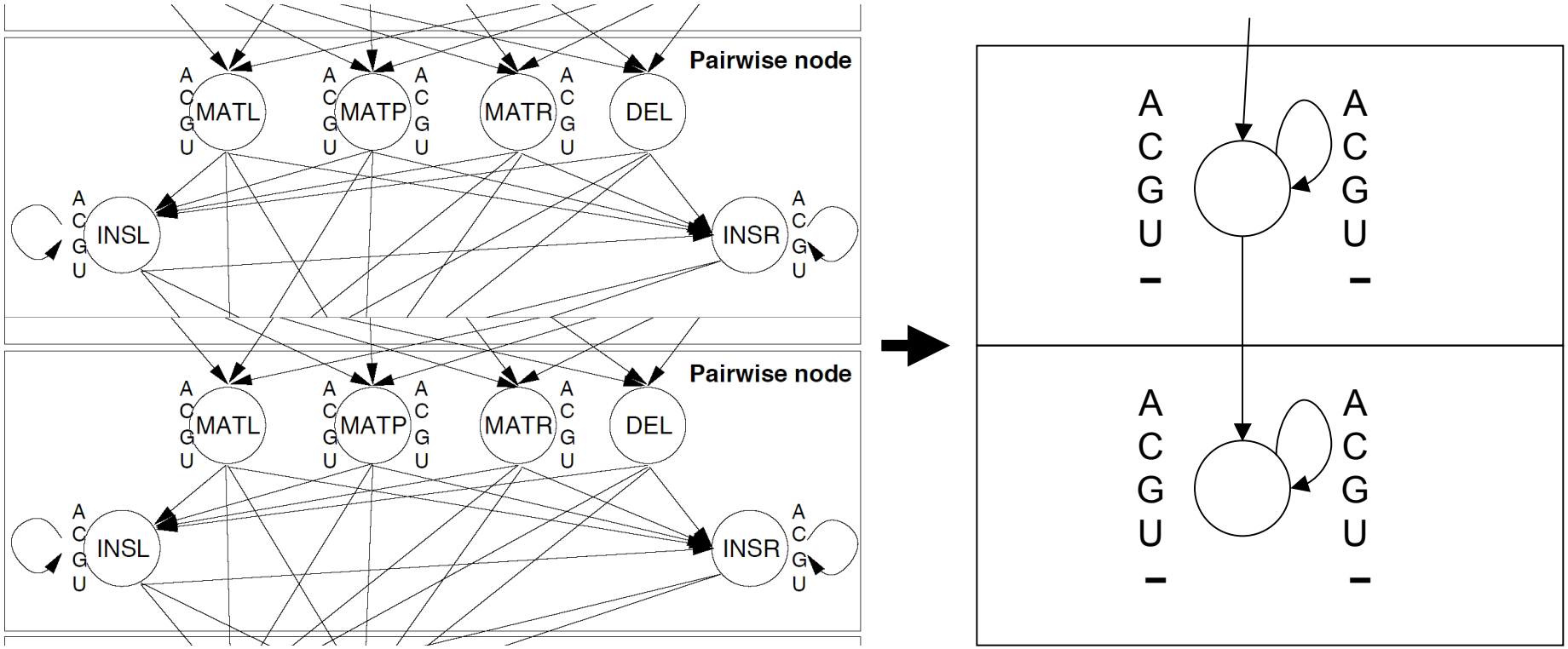


Key difference of CM vs HMM: Pair states emit paired symbols, corresponding to base-paired nucleotides; 16 emission probabilities here.

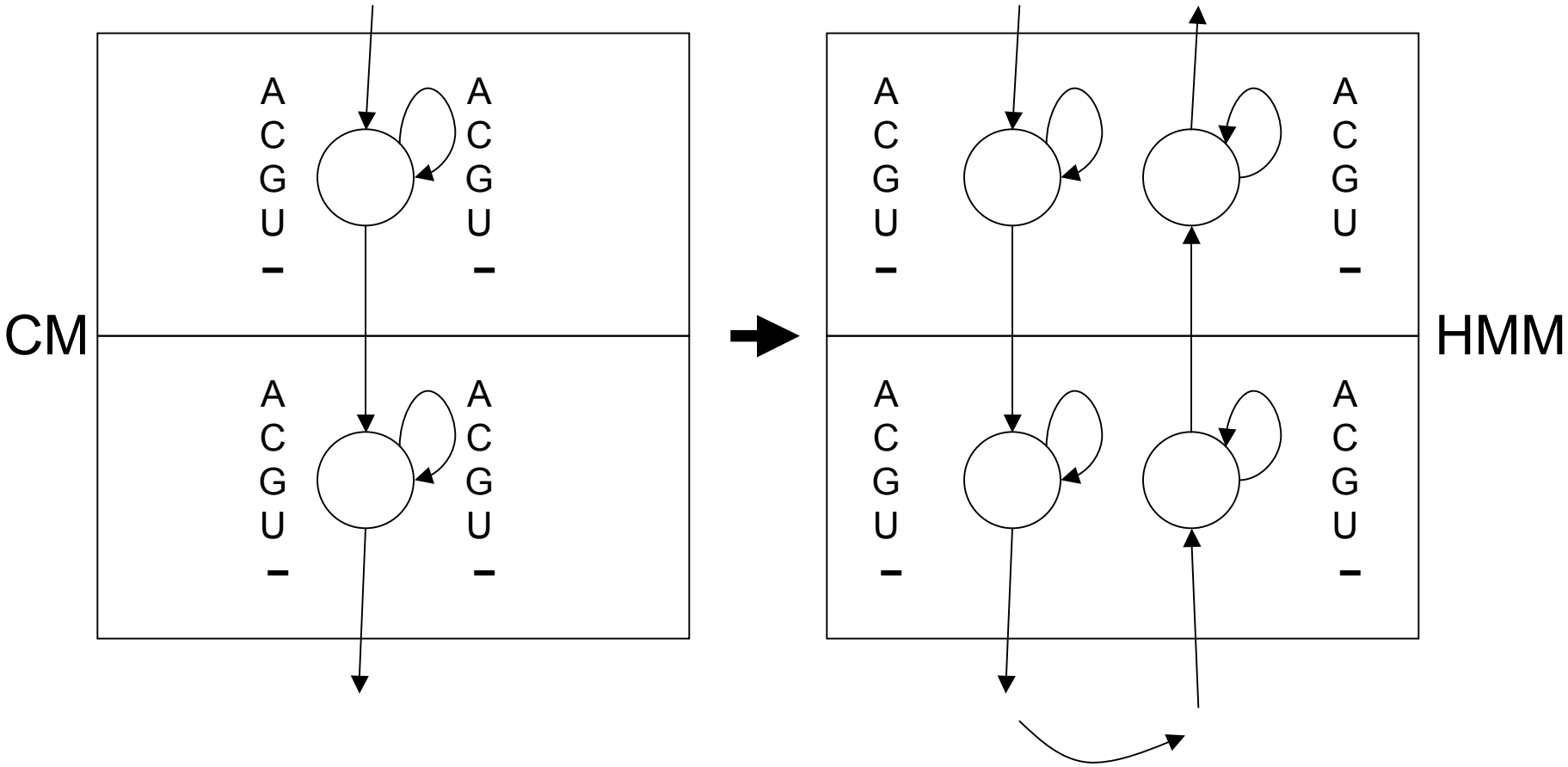
CM's are good, but slow



Oversimplified CM
(for pedagogical purposes only)



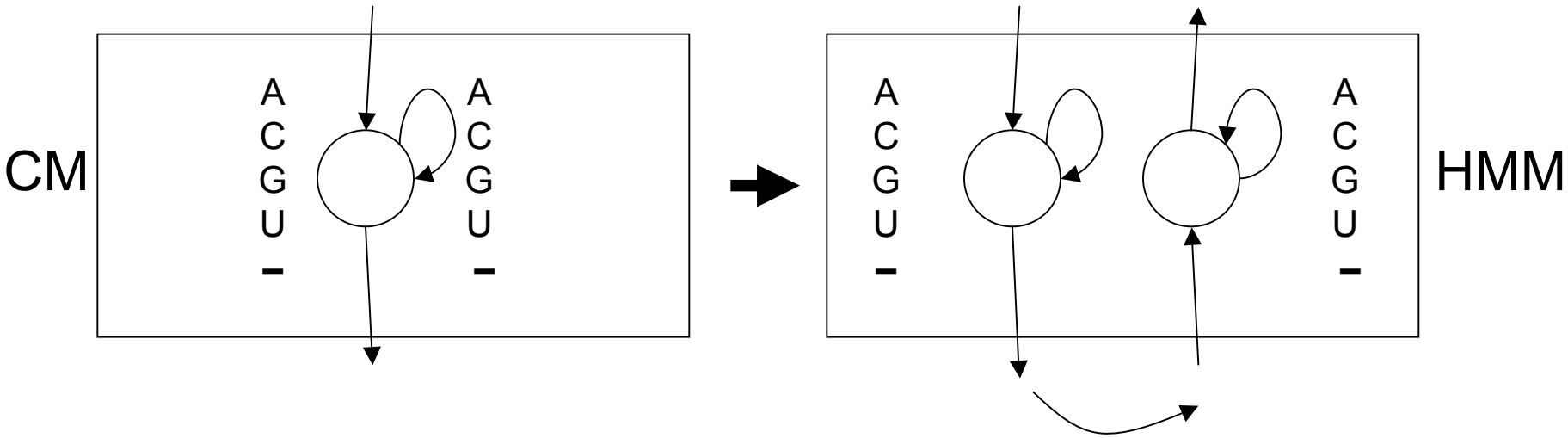
CM to HMM



25 emisions per state

5 emissions per state, 2x states

Key Issue: 25 scores ➔ 10

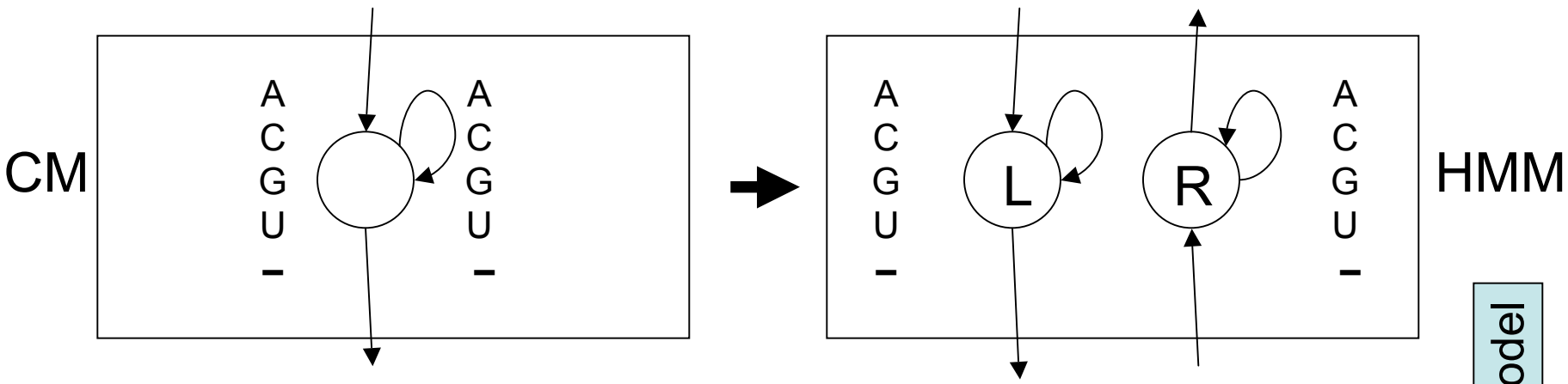


- Need: log Viterbi scores CM ≤ HMM

Viterbi/Forward Scoring

- Path π defines transitions/emissions
- Score(π) = product of "probabilities" on π
- NB: ok if "probabilities" 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

$P_{AA} \leq L_A + R_A$ $\quad P_{CA} \leq L_C + R_A$ $\quad \ldots$

$P_{AC} \leq L_A + R_C$ $\quad P_{CC} \leq L_C + R_C$ $\quad \ldots$

$P_{AG} \leq L_A + R_G$ $\quad P_{CG} \leq L_C + R_G$ $\quad \ldots$

$P_{AU} \leq L_A + R_U$ $\quad P_{CU} \leq L_C + R_U$ $\quad \ldots$

$P_{A–} \leq L_A + R_–$ $\quad P_{C–} \leq L_C + R_–$ $\quad \ldots$

NB:HMM not a prob. model

Rigorous Filtering

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

- *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

$P_{UA} = 1 \leq L_U + R_A$
$P_{UG} = 4 \leq L_U + R_G$

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

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

  – NB: E is a function of L_i, R_i only

Under 0th-order background model

- Optimization:

  Minimize $E(L_i, R_i)$ subject to score L.I.s

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

  – But still rigorous because "subject to score L.I.s"

Calculating $E(L_i, R_i)$

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

- 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 a numerical convex optimization algorithm

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

What should the probabilities be?

- Convex optimization problem
  - Constraints: enforce rigorous property
  - Objective function: filter as aggressively as possible
- Problem sizes:
  - 1000-10000 variables
  - 10000-100000 inequality constraints

Estimated Filtering Efficiency

(139 Rfam 4.0 families)

≈ break even

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

Averages 283 times faster than CM

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

What if filtering is poor?

- Profile HMM filter discards structure info
  - Surprise is that they usually do very well
  - But not always; e.g. a dozen families with filtering > .01, including stars like tRNA
- Three ideas:
  - Sub CM: graft in SM for a critical part
  - Store Pair: retain a few critical pairs
  - Filter Chains: run fast, crude filters first

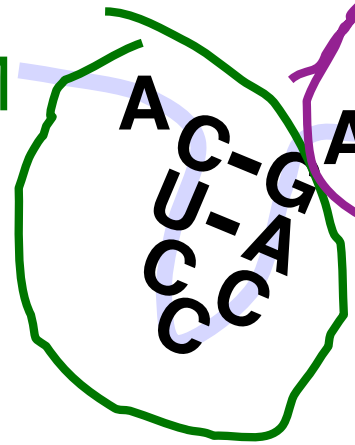
Sub-CM filters

Full CM

Profile HMM

ACUCCCAGAAGAGUUA

A sub-CM



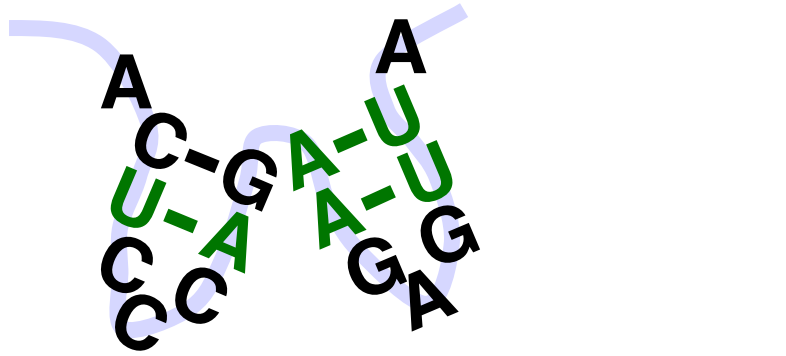
AAGAGUUA

Sub-CM

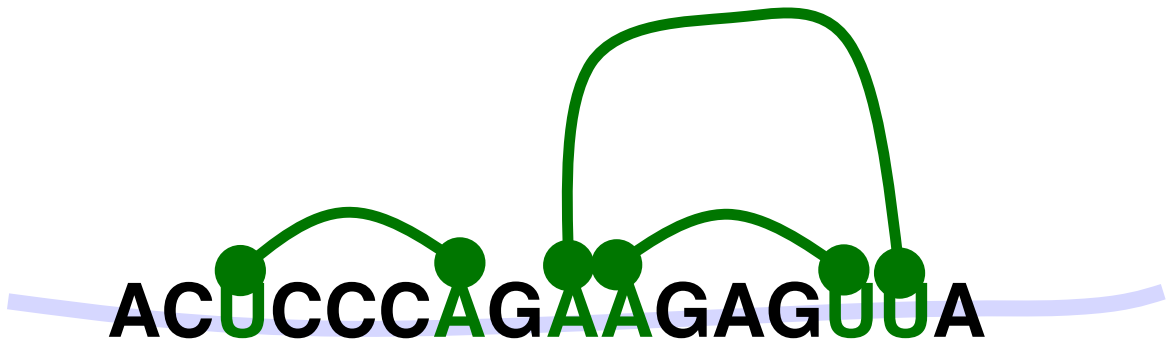
Sub-profile-HMM

Store-pair filters

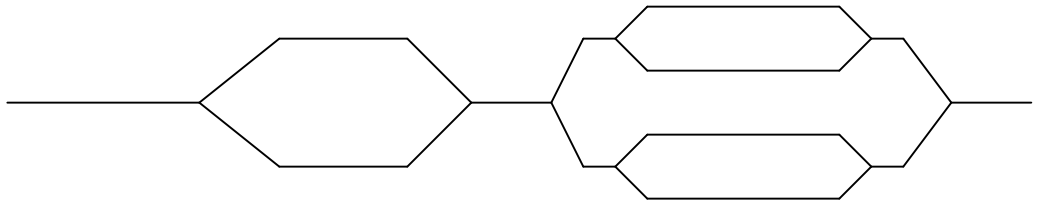
Full CM



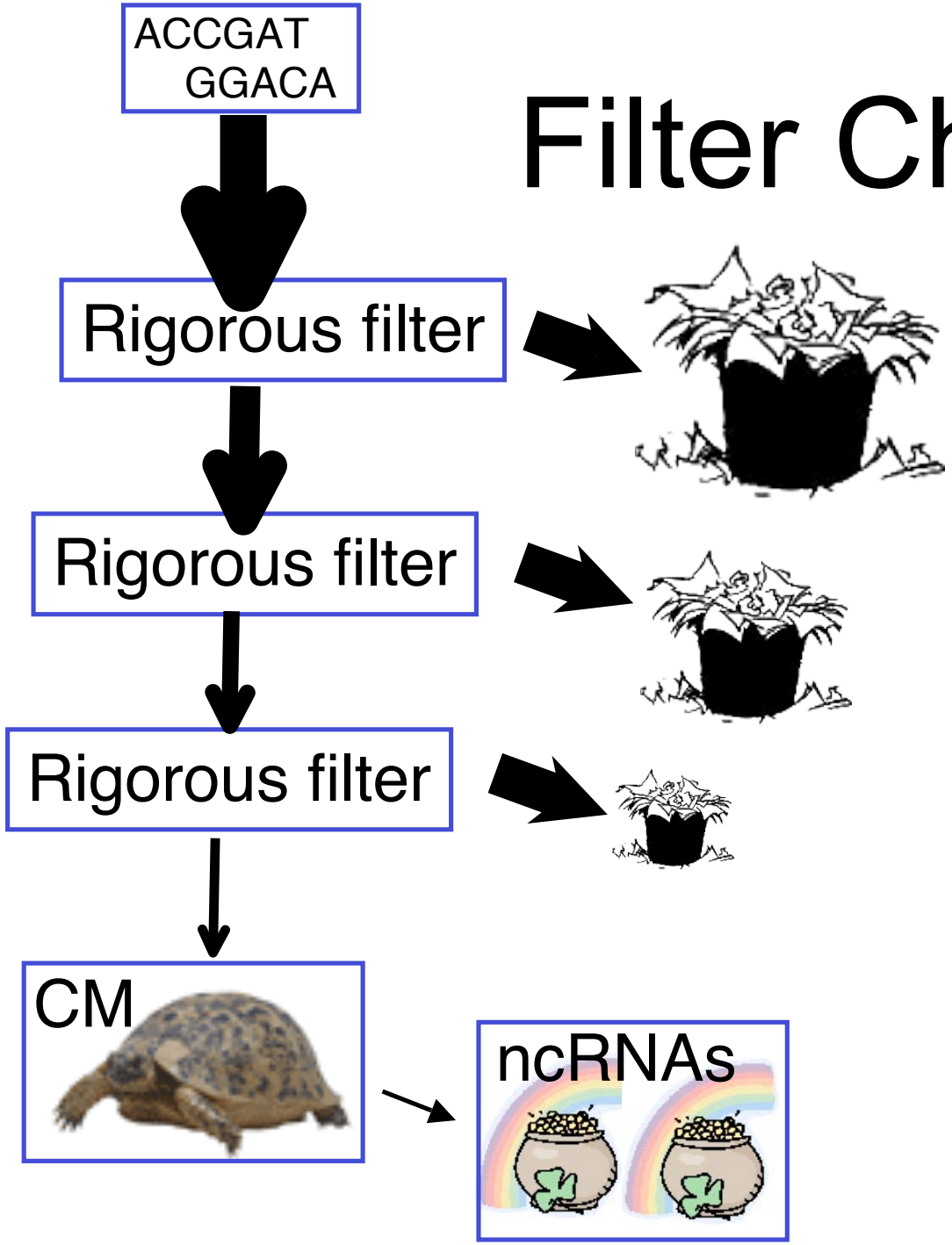
Store pair



“Profile” HMM:



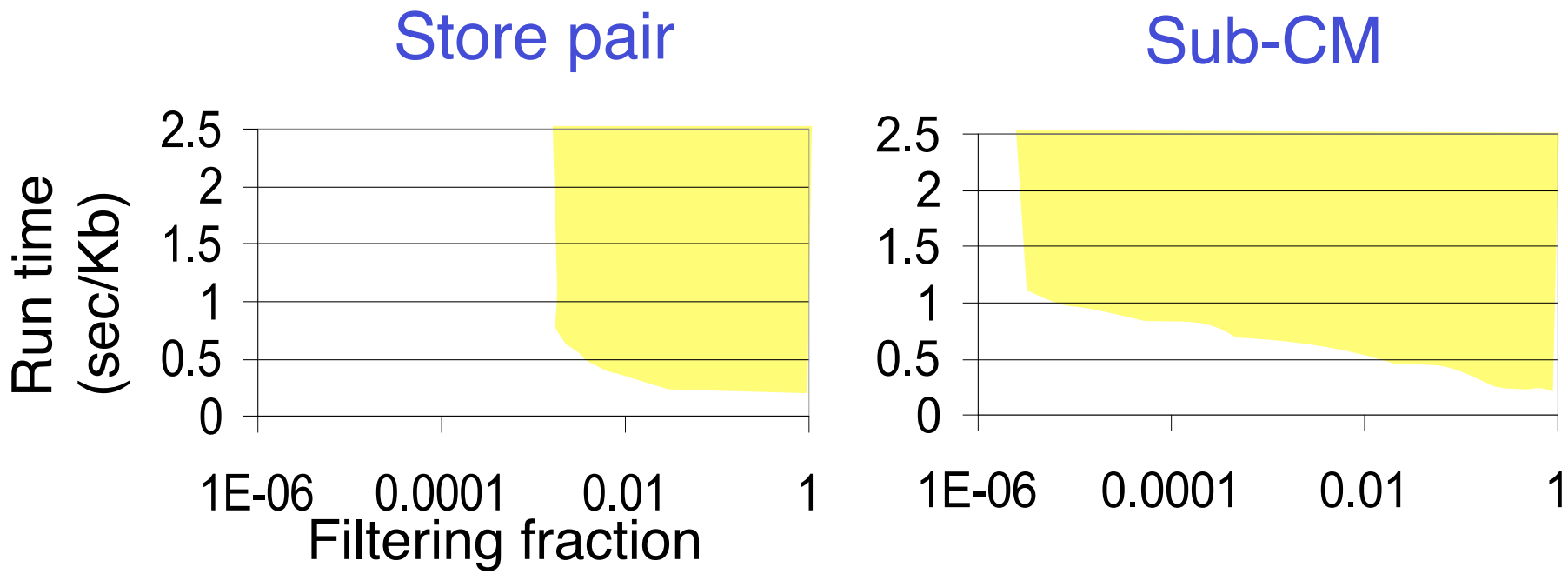
Filter Chains



Why run filters in series?

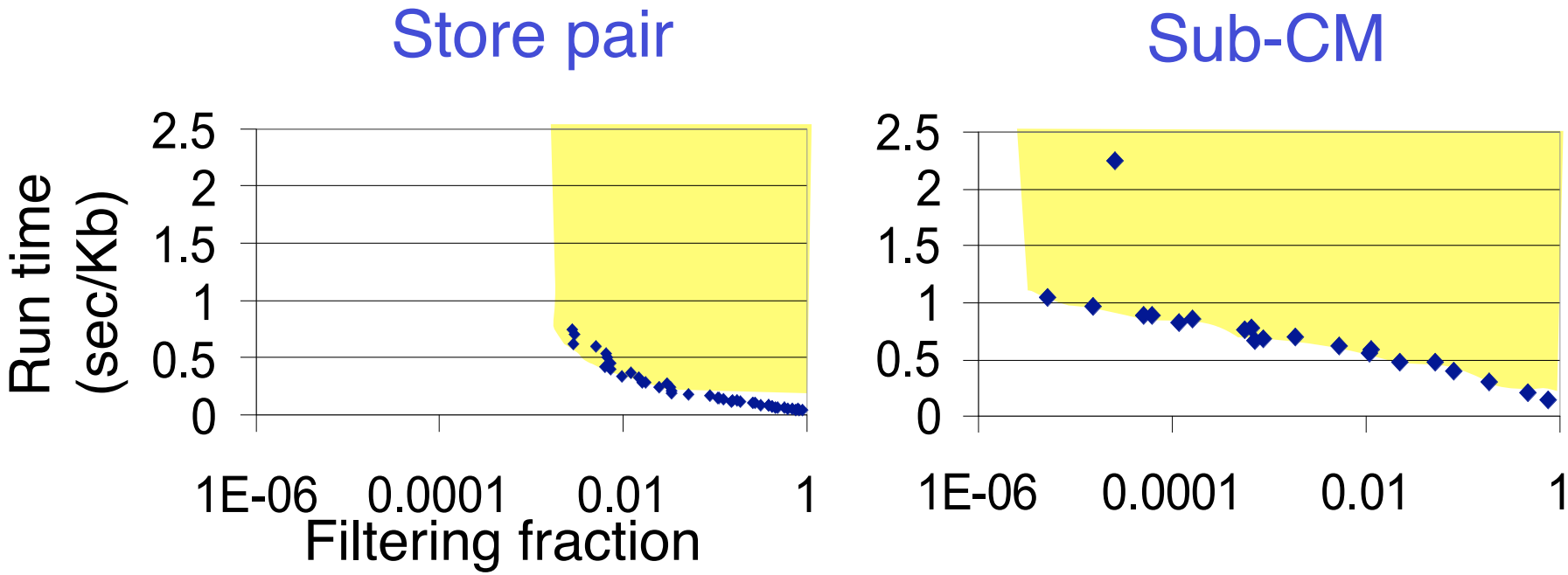
	Filtering fraction	Run time (sec/Kbase)
Filter 1	0.25	1
Filter 2	0.01	10
CM	N/A	200

- CM alone: 200 s/Kb
- Filter 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

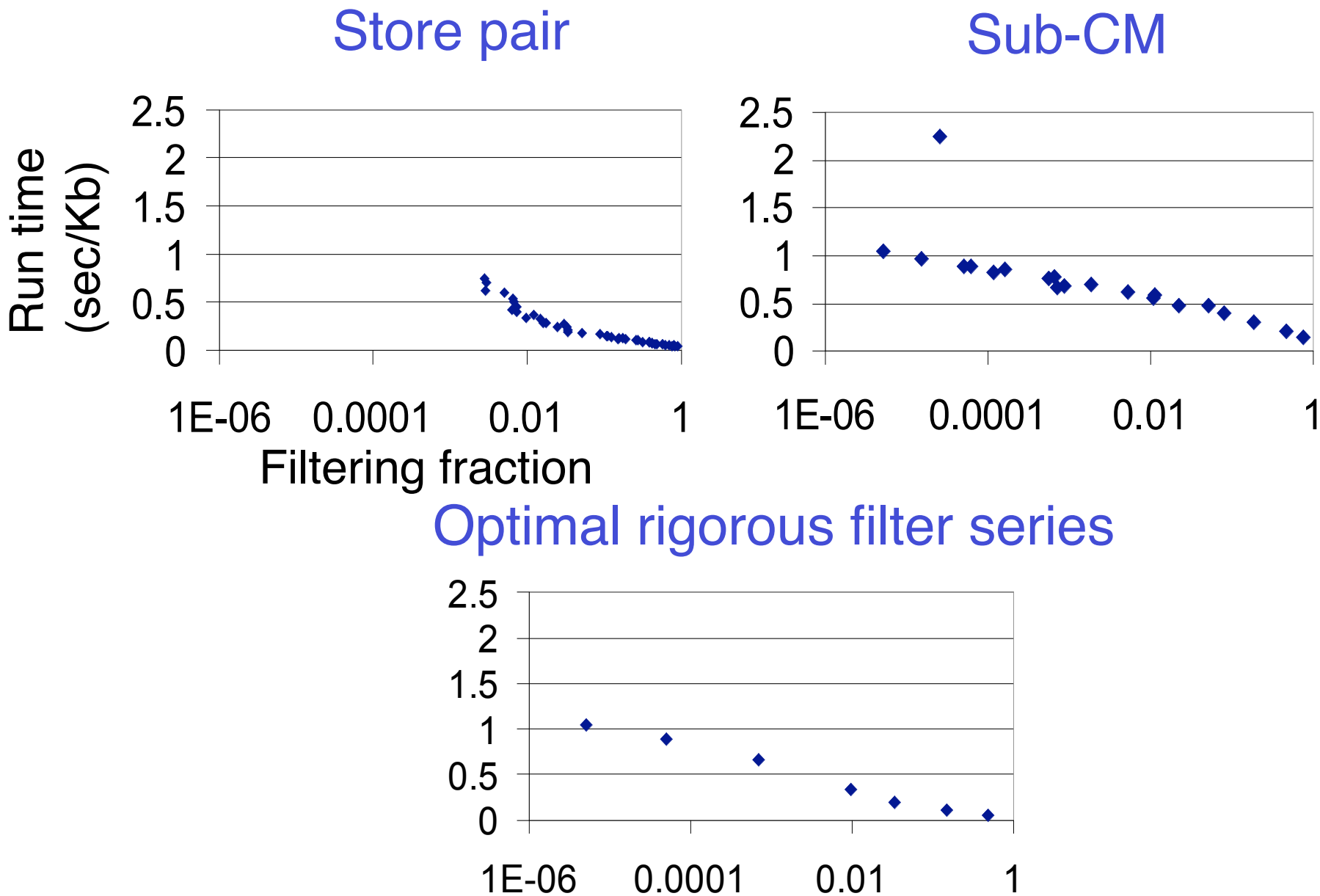


Properties of a filter:

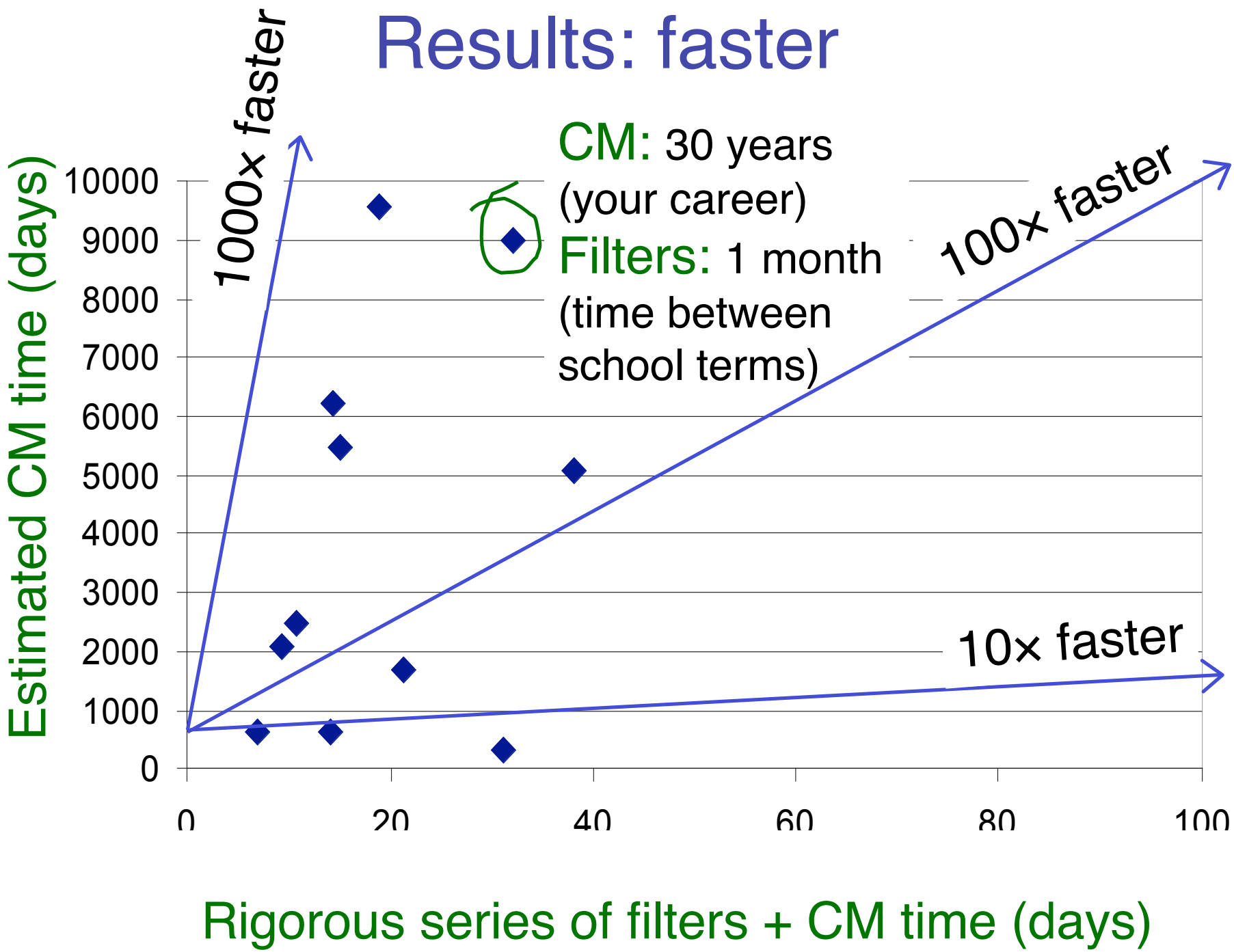
- Filtering fraction
- Run time (sec/Kb)



- Simplified performance model (selectivity and speed)
- Independence assumptions for base pairs
- Use dynamic programming to rapidly explore base pair combinations


Store pair
Sub-CM
Optimal rigorous filter series
Run time (sec/Kb)
2.5
2
1.5
1
0.5
0
1E-06
0.0001
0.01
1
Filtering fraction

Results: faster



Results: more sensitive than BLAST

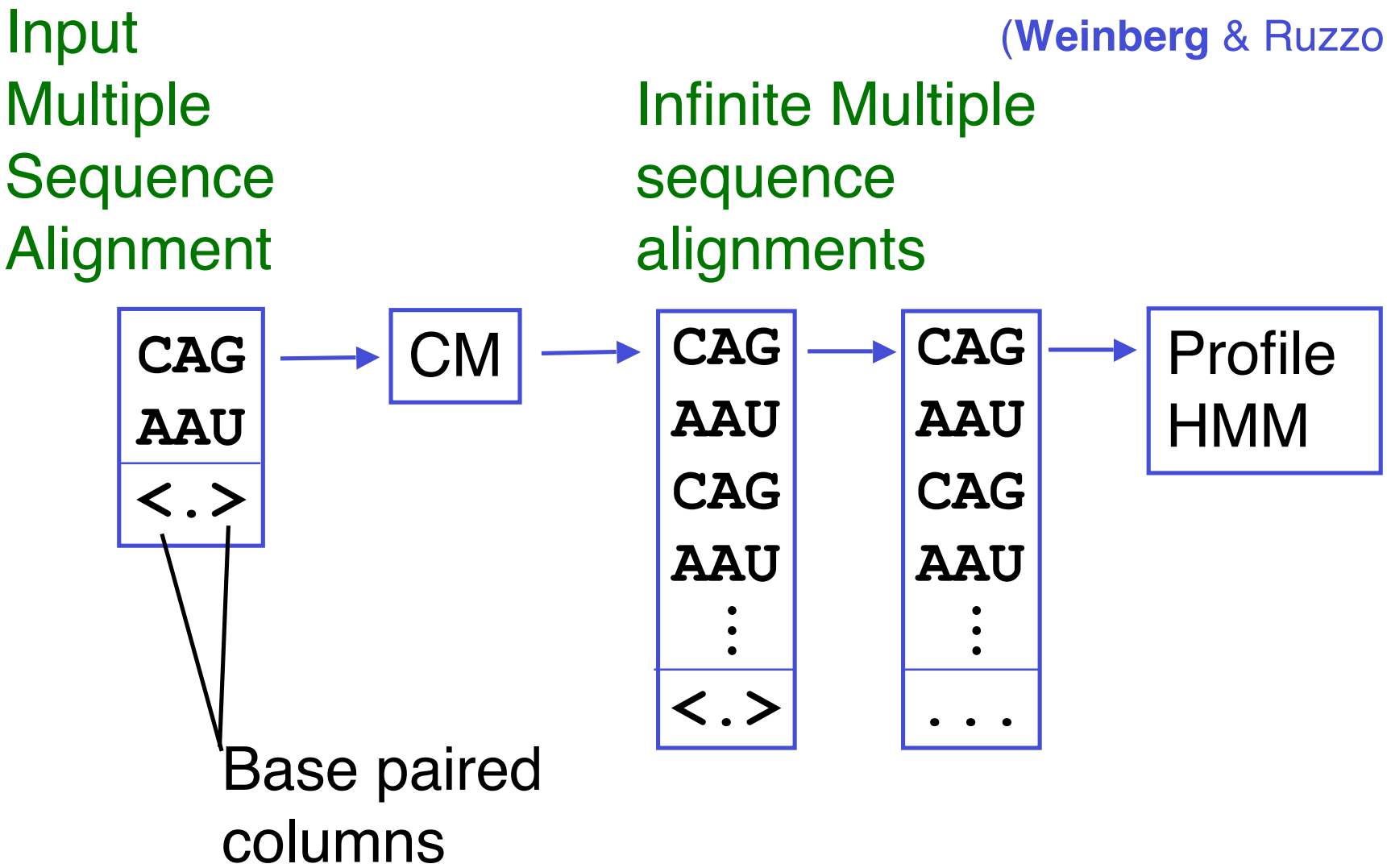
	# with BLAST+CM	# with rigorous filters + CM	# new
Rfam tRNA	58609	63767	5158
Group II intron	5708	6039	331
Iron response element	201	322	121
tmRNA	226	247	21
Lysine riboswitch	60	71	11
And more…			

Is there anything more to do?

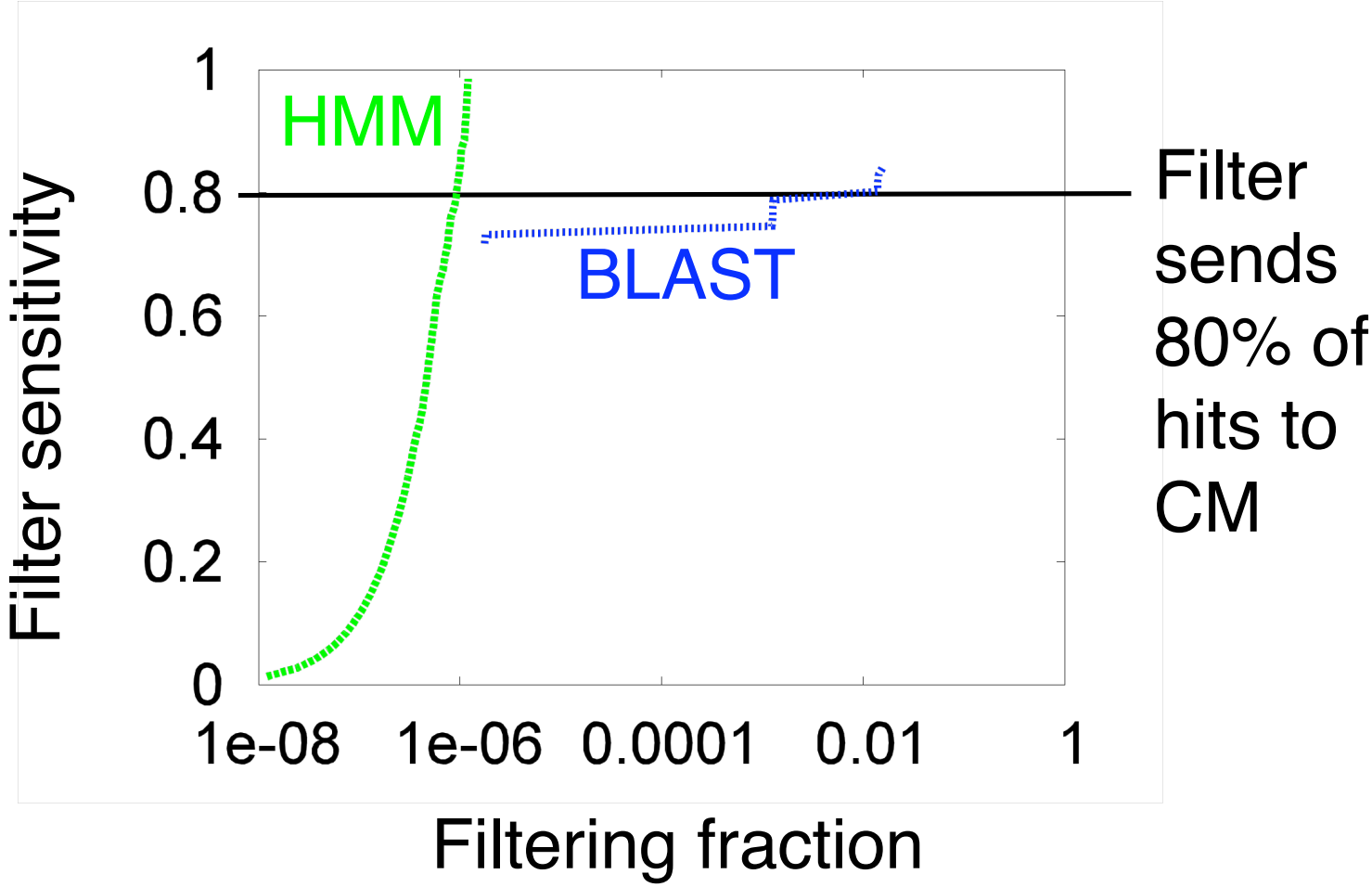
- Rigorous filters can be too cautious
  - E.g., 10 times slower than heuristic filters
  - Yet only 1-3% more sensitive
- We want to
  - Run scans faster with minimal loss of sensitivity
  - Know empirically what sensitivity we're losing

Heuristic Profile HMMs

(**Weinberg** & Ruzzo, 2006)



ROC-like curves (lysine riboswitch)



Filter sends 80% of hits to CM

tRNAscan-SE: the leading brand

(Lowe & Eddy 1997)

- Designed for tRNAs
- Used in virtually every genome project
- Uses CMs
- Heuristics:
  - selected 2 tRNA detection programs

An ambitious target to shoot for

tRNAscan-SE vs. heuristic HMMs

	Sensitivity (%)		Run time (hours)		
	t-SE	HMM	t-SE	HMM	rigor
Archaea	98.5	99.3	0.21	0.67	1.76
Eubacteria	99.4	99.8	2.79	10.0	36.7
C. elegans	98.1	97.5	0.13	1.03	64.3
Drosophila	99.7	99.3	0.08	1.12	19.0
Human	83.4	90.4	3.41	30.9	581.1

(Each filter sends same number of nucleotides to CM)

tRNAscan-SE vs. heuristic HMMs

Time to create heuristic filters for tRNAs	
tRNAscan-SE	HMM
≥ 8 papers	• 10 seconds to type a command • 15 minutes to create & calibrate HMM

Point is not that heuristic HMM is better than tRNAscan-SE --- it's not; point is that it's in the ballpark, so may be easy way to get useful results for *new* families.

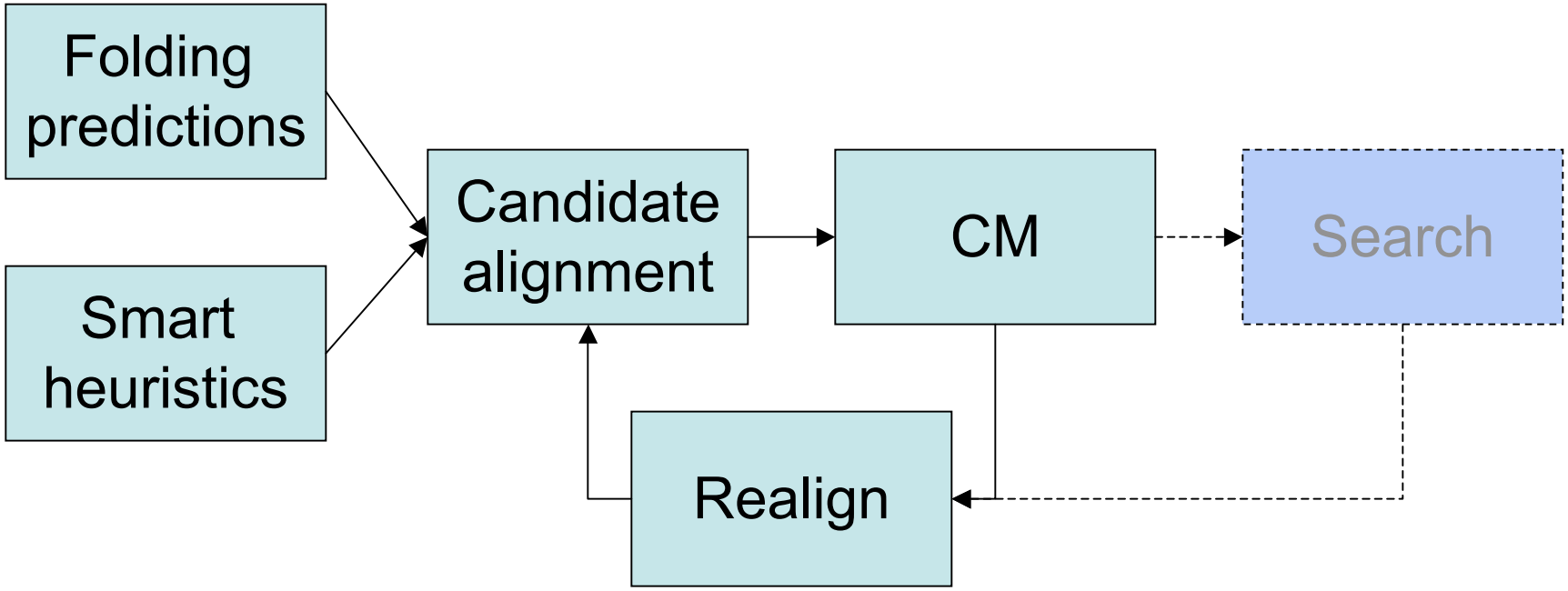
Building CM's

- Hand-curated alignments + structure as in Rfam are great, but it doesn't scale
- Example Application:
  Given 5-20 upstream regions (~500 nt) of orthologous bacterial genes, some (but not all) plausibly regulated by a common riboswitch, could we find it?

CMFinder

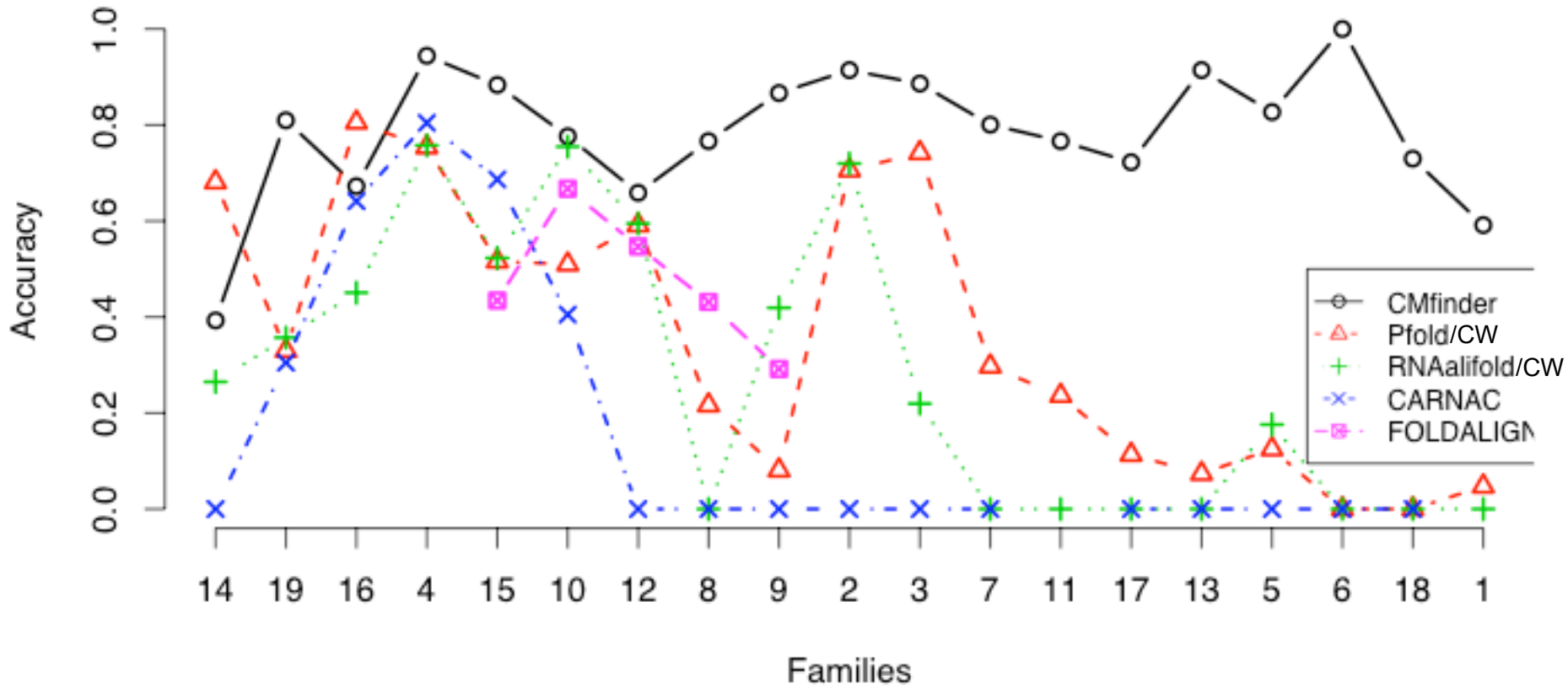
Harder: Finding CMs *without* alignment

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006



CMfinder Accuracy

(on Rfam families *with* flanking sequence)



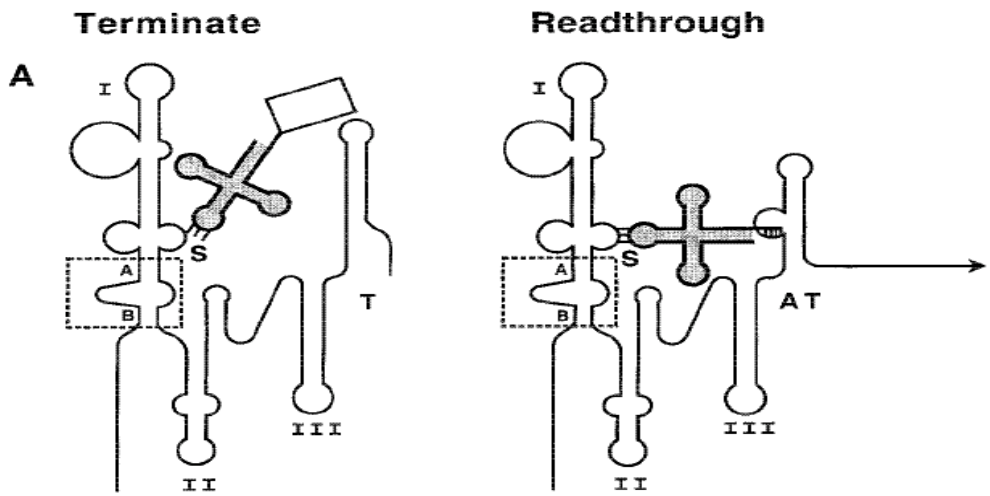
Importance of Alignment

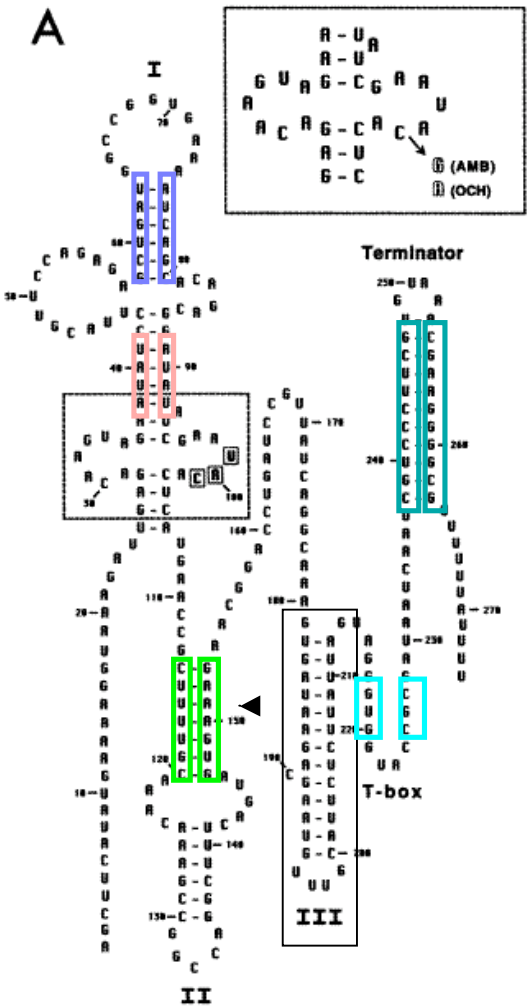
```
AC000078.2  GGTGGCCCGTGTGCCCCACAGGGATGGCTCAGGGGACTGTCCACCTCACCCCTGCACCTC-TCAGCCTCTGCCGCCGGGCA---CCCCCCCCCAGGCTCCTGGTGCCAG--ATGATGACG
AF166127.1  ------------------------------------------------------------------------------------CCCCCCCCCAGGCTCCTGGTGCCGG--ATGATGACG
AF195141.1  ---------------------------------CTCTTTTGCATTAAGGGATCATTGCAAGAGCAGCGTG--ACTGACATTA---TGAAGGCCTGTACTGAAGACAGCAA--GCTGTTAGT
AF390544.1  ----------------------------------GTAATACTTATAAAGGTTTGCATTAATGAGGATTACACAGAAAACCTTT-GTTAAGGGTTTGTGTCGATCTGCTAA--TTGGCAAAT
AL049837.4  -------------------------------------AGAGTTGACCAGTGTGCGGATGATAACTACTGACGAAAGAGTCATCGACTCAGTTAGTGGTTGGATGTAGTCACATTAGTTTGC
AL645723.11 --------------------------------GTGTT-CGATAGCATTGGACTGATAGGTA-GCCATGGC--TTCATCTGTC---ATG--TCTGCTTCTTTTTATATTTG--TGTATGATG
AY060611.1  -------------CCCGCCCATGTGGCGCTTATGACGCAGTTGTCTTAAACTCGAACTCGAGCGGGCAATTGCTGATTACGA-TTAACCACTGATTCCTGGGTCGCTGC--TTCGTGGCC
L14329.1    -------------------------------------GCTTTGAACAAATTCTCGCTATATGACGATGGCAATCTCAAATGT-TCATTGGTTGCCATTTGATGAAATCAGTTTTGTGTGC
L28111.1    ---------------------------------------------GTGGGGCGGGAGTACAAGGTGCGTGTGACTGGAGCCA---CCCACTCCGACTCTGCAGGTGTTTG--CAAATGACG
M63574.1    ---------------------------------------------TTCTTTCTCCAGTGTTCTAGTTACATTGATGAGAACAGAA-ACATAAACTATGACCTAGCGGTTTCT--GTTGGATAG
S48220.1    GATTTCTTCGCAAGTCTCTTAATGGTCATTTGTGTT-AGATTACATCAAACTGATGGATA-GCCATTGGTATTCATCTATT---TTAACTCTGTGTCTTTACATATTTG--TTTATGATG
U43286.1    ----------CTGAAGTACTGGCTCTTTCCTGCTCTGGACAAGAATTGAGCAACTTGTCT-GATGACTGGGAAAGGAGGAC---CTGCAACCATCTGACTTGGTCTCTG--TTAATGACG
X03920.1    ------------------------------------TTGGCTTGGTGATTACTGGCTGCACTCTGGGGGGCGGTTCTTCCA---TGATGGTGTTTCCTCTAAATTTGCA--CGGAGAAAC
X84742.1    ----------------------------CATCCCACAGTGCTCCTGAGACCAGGCAAGACAACTGTGAGC-GCGATGGCCG---TGTACCCCAGGTCAGGGGTGGTGTC--TCTATGAAG
Y11110.1    -------------CACATACAATGTTCCTAAACGTTCAGTTTCCCTCACTTCAGAAGGCT-TCTGAATGGAACCATCTCTT---GACA-TTTGTTTCTATA-ATATTTG--T-CATGACA
Y11273.1    --------------------CGTGTGTGCGCGGGTGTGTGTCTGAAAAGTTGTGTACAAGTGCTCCGTGCTGCCTAGCAAG---TGCTAACTGGGATTTCTAGTATTTC--TTTGTGATG
```

- Blue boxes, e.g., should be lined up.
- Structure is invisible otherwise.

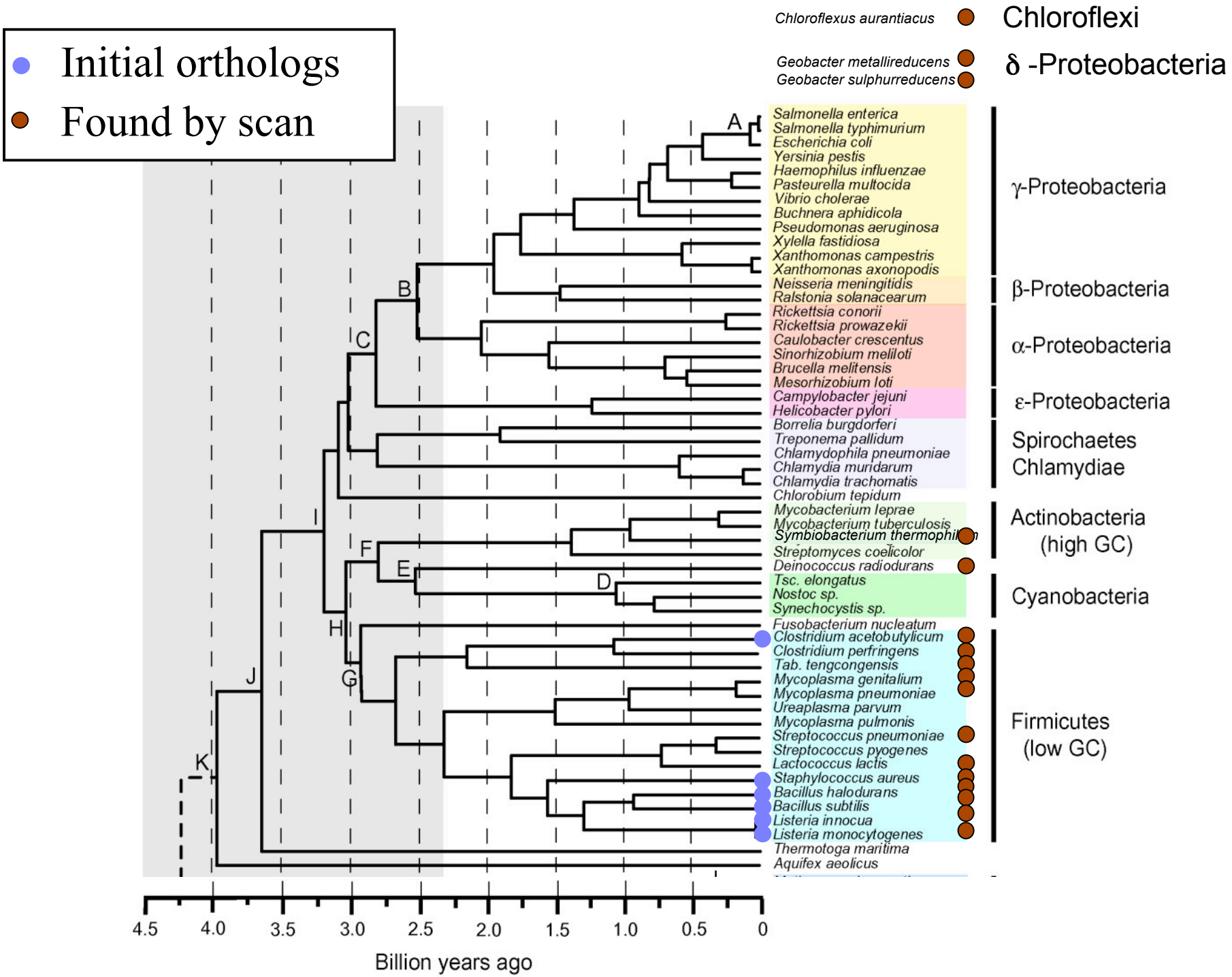
Early Semi-automated Example

- Started with 16 genes orthologous to folC in *B. subtilis*
- Found 10 sharing good structural motif
- Searched all bacterial genomes for this motif
- Found 234 hits
- Realigned these to refine structural motif
- Found 367 hits
- 257 match RFAM's T-box
  - (Based on hand-curated alignment of 67 knowns)






```
NC_000964.1 AUAUC.CUUACGU..UCCAGAGAGCUGAUGGCCGGUGAAA.AUCAGCACAGACGGAUAUAU
NC_004722.1 CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUUG
NC_004193.1 AAAAGUAGAACCG.AUCUAGCGAAUUGAGGAU.GGUGUGAGCUCAGUGC.GGAAAGCUUUU
NC_003997.3 CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUUG

NC_000964.1 CGAA..UACACUCAUGAACCGCUUUUGCAAACAAAGccggccaggcuuucAGUA.GUGAAAG
NC_004722.1 UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuuggauu.....AGUAAGCAUUCC
NC_004193.1 AGAAAAUC.ACUCUUGAGUU.UUCAUUACGAAA..CA.............AGUAGUAAUGGA
NC_003997.3 UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuaugauu.....AGUAAACAUUCC

NC_000964.1 aCGGAC.CUGAUCCGUUAUCAGGCAAAGUGGUACCGCGAUAAUCAAUCGUCCCUUCGUGUAAaCGAAGGGGCGUUU
NC_004722.1 .CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAuAGGGACGGGAGUU
NC_004193.1 .CGGUUcAUC.UCCGUUAUCGAUCUUAGUGGUACCGCGA.....GUCUUCUCGUCCCUUUU..GGGAUUAGAAGGC
NC_003997.3 .CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAuAGGGACGGGAGUU
```


Initial orthologs
Found by scan
Chloroflexus aurantiacus
Chloroflexi
Geobacter metallireducens
Geobacter sulphurreducens
δ -Proteobacteria
Salmonella enterica
Salmonella typhimurium
Escherichia coli
Yersinia pestis
Haemophilus influenzae
Pasteurella multocida
Vibrio cholerae
Buchnera aphidicola
Pseudomonas aeruginosa
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas axonopodis
γ-Proteobacteria
Neisseria meningitidis
Ralstonia solanacearum
β-Proteobacteria
Rickettsia conorii
Rickettsia prowazekii
Caulobacter crescentus
Sinorhizobium meliloti
Brucella melitensis
Mesorhizobium loti
α-Proteobacteria
Campylobacter jejuni
Helicobacter pylori
ε-Proteobacteria
Borrelia burgdorferi
Treponema pallidum
Spirochaetes
Chlamydophila pneumoniae
Chlamydia muridarum
Chlamydia trachomatis
Chlamydiae
Chlorobium tepidum
Mycobacterium leprae
Mycobacterium tuberculosis
Symbiobacterium thermophilum
Streptomyces coelicolor
Actinobacteria
(high GC)
Deinococcus radiodurans
Tsc. elongatus
Nostoc sp.
Synechocystis sp.
Cyanobacteria
Fusobacterium nucleatum
Clostridium acetobutylicum
Clostridium perfringens
Tab. tengcongensis
Mycoplasma genitalium
Mycoplasma pneumoniae
Ureaplasma parvum
Mycoplasma pulmonis
Streptococcus pneumoniae
Streptococcus pyogenes
Lactococcus lactis
Staphylococcus aureus
Bacillus halodurans
Bacillus subtilis
Listeria innocua
Listeria monocytogenes
Firmicutes
(low GC)
Thermotoga maritima
Aquifex aeolicus
A
B
C
D
E
F
G
H
I
J
K
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
Billion years ago

An approach for cis-regulatory RNA discovery in bacteria

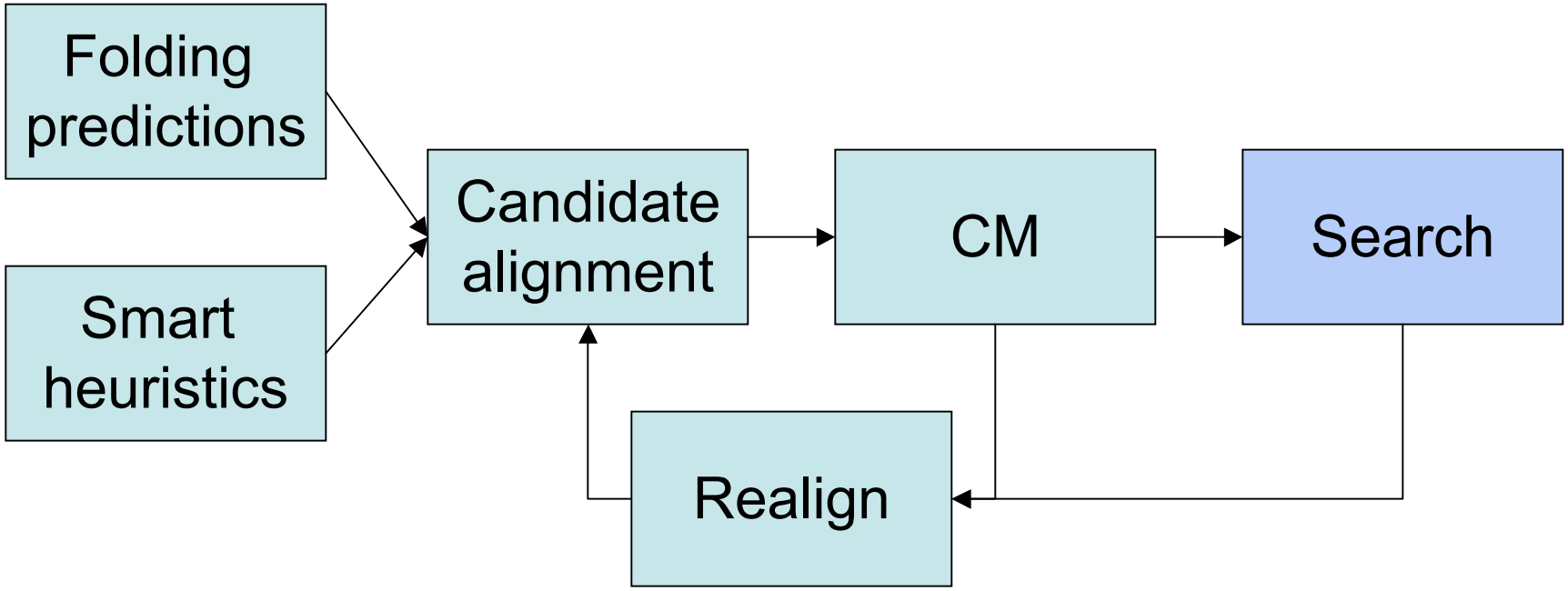
1. Choose a bacterial genome
2. For each gene, collect 10-30 close orthologs
3. Find most promising genes, based on sequence motifs conserved among orthologs
4. From those, find most promising genes, incorporating structure in the motifs
5. From those, genome-wide searches for more instances
6. Expert 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

Genome Scale Search: How

CMfinder is directly usable for/with search



Results

- Process largely complete in
  - bacillus/clostridia
  - gamma proteobacteria
  - cyanobacteria
  - actinobacteria
- Analysis ongoing

Some Preliminary Actino Results

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

not cis-regulatory

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
  - ribosomal genes 9
  - potentially interesting 12
  - unknown or poor 12
- One other being bench-verified

Software

- Infernal - (Eddy et al.) most of Eddy & Durbin
- RaveNna - (Weinberg) fast filtering
- CMfinder - (Yao) Motif discovery (local alignment)

Summary

- ncRNA is a “hot” topic
- For family homology modeling: CMs
- Training & search like HMM (but slower)
- Dramatic acceleration possible
- Automated model construction
- Hopefully leading to new discoveries