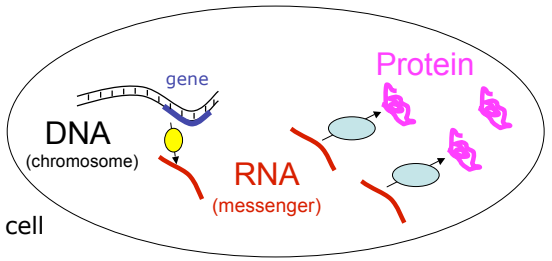
Searching for Non-Coding RNA

W.L. Ruzzo

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

The "Central Dogma"

DNA → RNA → Protein



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. ribosomal and transfer RNA, splicing, SRP)
- Since mid 90's dramatic discoveries
  - Regulation, transport, stability/degradation
  - E.g. "microRNA": ≈ 250 in humans

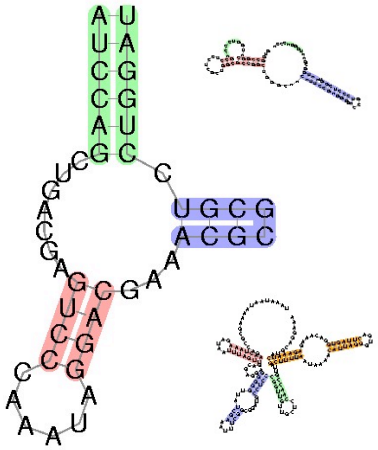
RNA: not just a messenger any more

- Classic non-coding RNAs: tRNAs, rRNAs

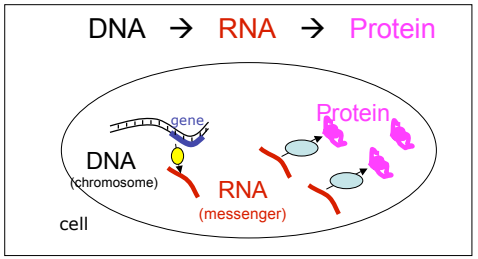
new ncRNA classes	Role	Circa	How many?
microRNAs	Gene regulation	Since 2000	Est. 200-250 in human (Lim, *et al.*, 2003)
Riboswitches	Sensing small molecules	Since 2002	8 types (1000s of homologs) plus ≥3 types unpublished
Many more…			

Why?

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



"Central Dogma" = "Central Chicken & Egg"?



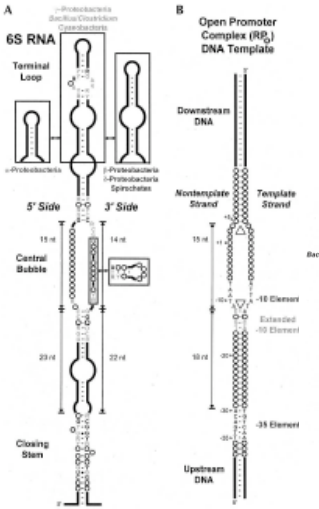
Was there once an "RNA World"?

Iron Response Element

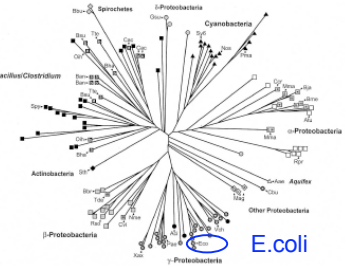
IRE (partial seed alignment):

```
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

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

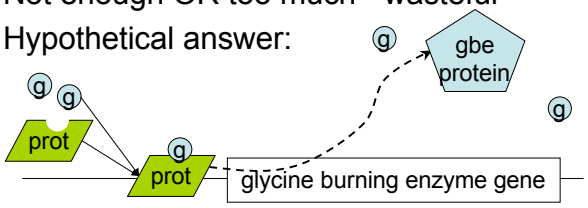
polyadenylation inhibition element RNA



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

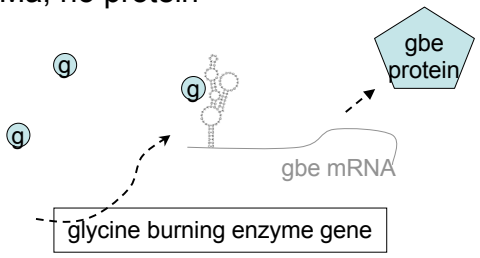
Another Example: the Glycine Riboswitch

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

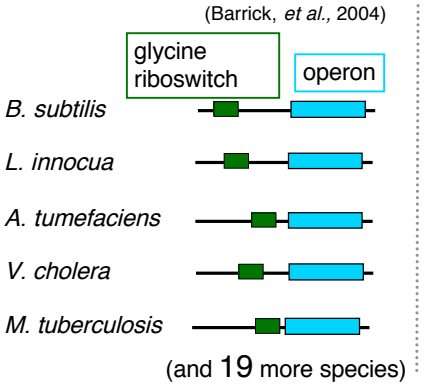


The Glycine Riboswitch

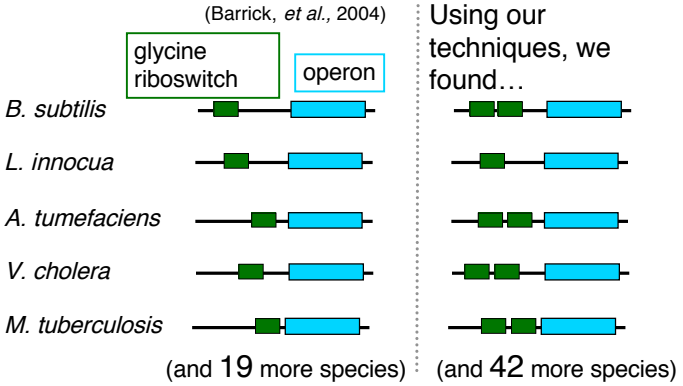
- Actual answer (in many bacteria): Look Ma, no protein



Impact of RNA homology search

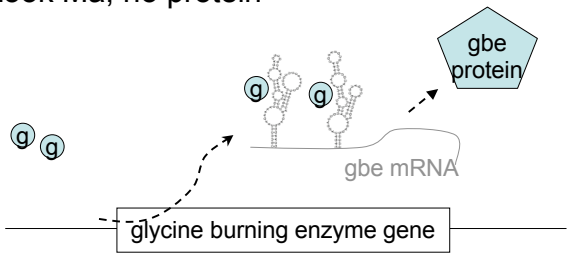


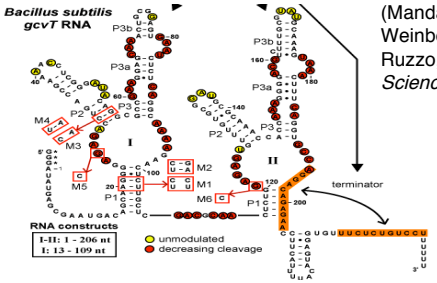
Impact of RNA homology search



The Glycine Riboswitch

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

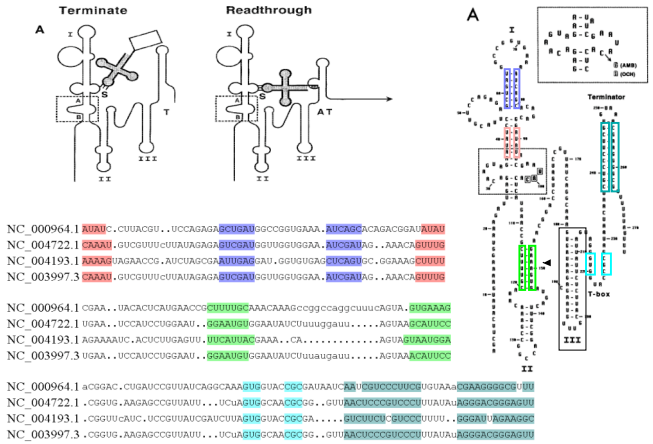




(Mandal, Lee, Barrick, Weinberg, Emilsson, Ruzzo, Breaker, *Science* 2004)

And…

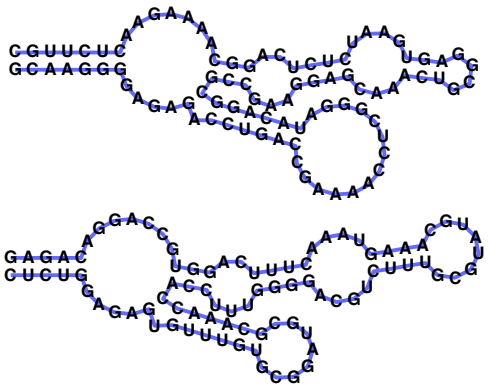
- More examples means better alignment
- Understand phylogenetic distribution
- Find riboswitch in front of new gene



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

"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

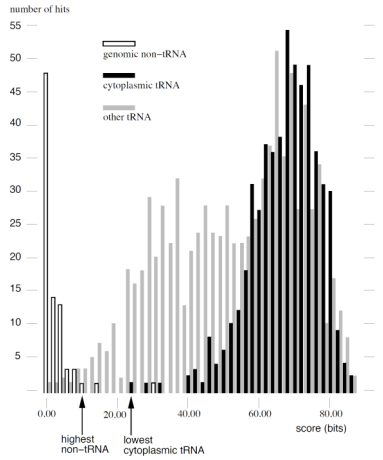
Main Results

- 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

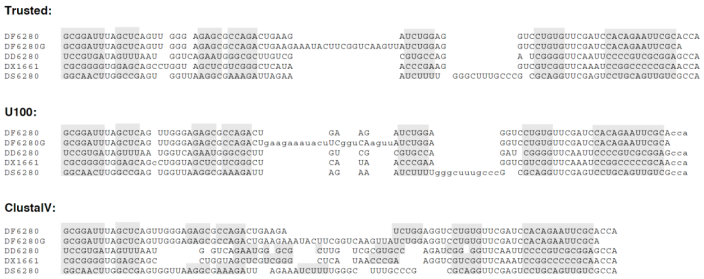
Probabilistic Model Search

- 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



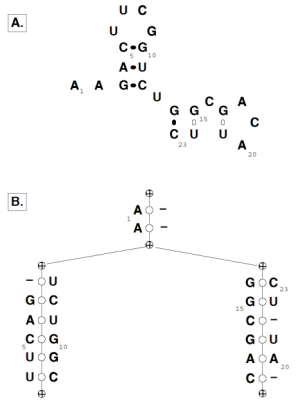
Comparison to TRNASCAN

- 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

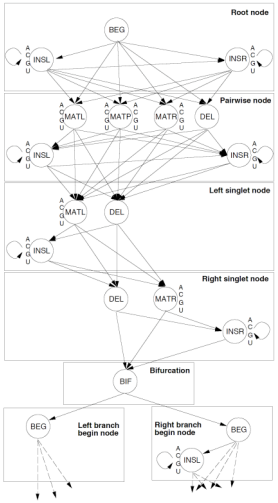
CM Structure

- 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 $i,...,j$ of input

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

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

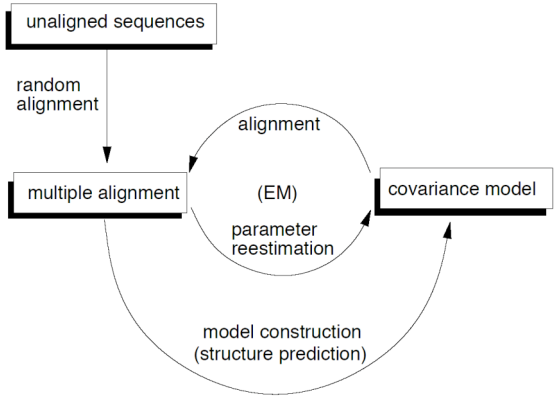
$S^{y}_{ij} = \max_{\pi} \log P(x_{ij} \text{ generated starting in state } y \text{ via path } \pi)$

Viterbi, cont.

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

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

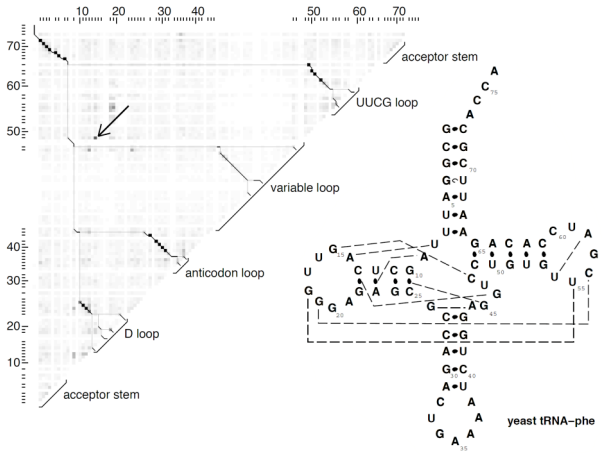
Model Training



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



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%

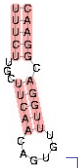
Table 2: Training and multiple alignment results from models trained from the trusted alignments (A models) and models trained from no prior knowledge of tRNA (U models).

Faster Genome Annotation of Non-coding RNAs Without Loss of Accuracy

Zasha Weinberg
& W.L. Ruzzo

Recomb '04

Rfam

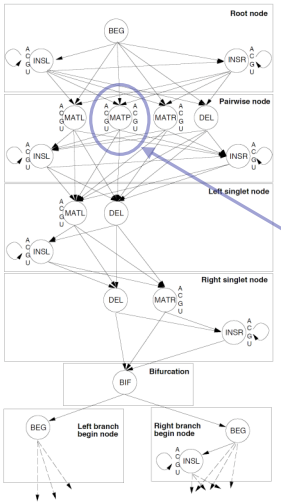


- Input (hand-tuned):
  - MSA
  - SS_cons
  - Score Thresh T
  - Window Len W
- Output:
  - CM
  - scan results

IRE (partial seed alignment):

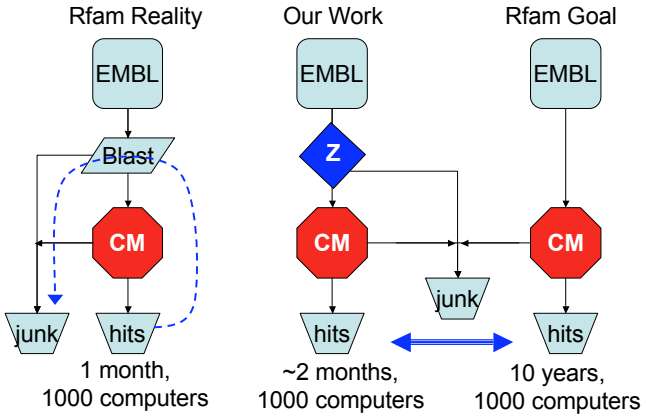
```
Hom.sap.  GUUCCUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.  UUUCUUC.UUCAACAGUGUUUGGAUGGAAC
Hom.sap.  UUUCCUGUUUCAACAGUGCUUGGA.GGAAC
Hom.sap.  UUUAUC..AGUGACAGAGUUCACU.AUAAA
Hom.sap.  UCUCUUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.  AUUAUC..GGGAACAGUGUUUCCC.AUAAU
Hom.sap.  UCUUGC..UUCAACAGUGUUUGGACGGAAG
Hom.sap. 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



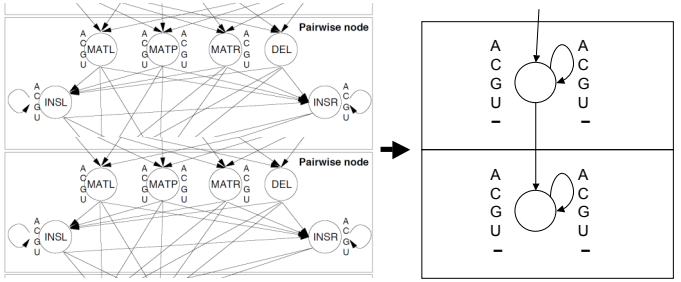
Key difference of CM vs HMM: Pair states emit paired symbols, corresponding to base-paired nucleotides; 16 emission probabilities here.

CM's are good, but slow

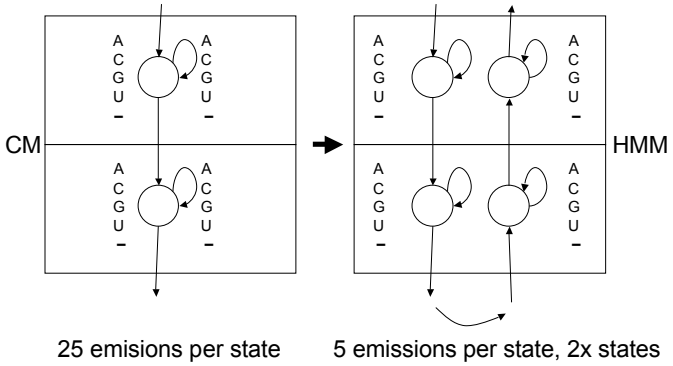


Oversimplified CM

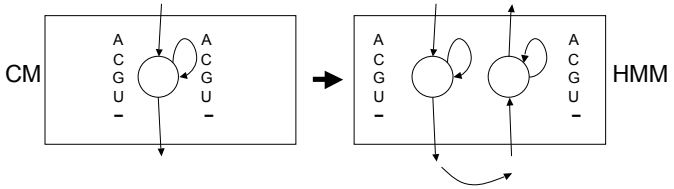
(for pedagogical purposes only)



CM to HMM



Key Issue: 25 scores ➔ 10

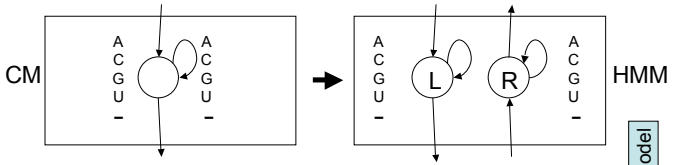


- Need: log Viterbi scores CM ≤ HMM

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
  $E(L_i, R_i) = \Sigma_x$ Forward-score(x)*Pr(x)
  – NB: E is a function of L_i, R_i only
- 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"

Under 0th-order background model

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

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

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

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

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…			