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

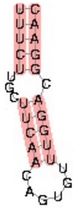
Zasha Weinberg

& W.L. Ruzzo

Recomb ‘04

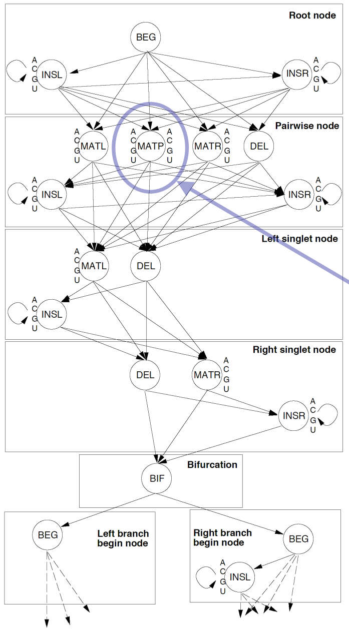
Rfam

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



IRE (partial seed alignment):

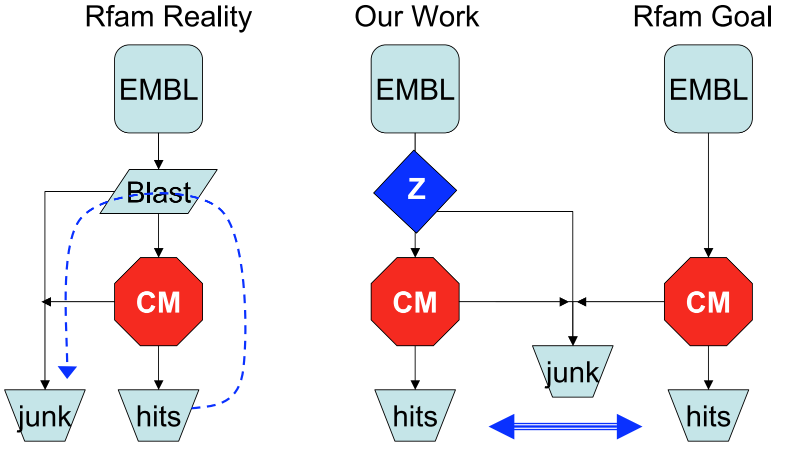
```
Hom.sap.   GUUCCUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCUUC.UUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCCUGUUUCAACAGUGCUUGGA.GGAAC
Hom.sap.   UUUAUC..AGUGACAGAGUUCACU.AUAAA
Hom.sap.   UCUCUUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   AUUAUC..GGGAACAGUGUUUCCC.AUAAU
Hom.sap.   UCUUGC..UUCAACAGUGUUUGGACGGAAG
Hom.sap.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



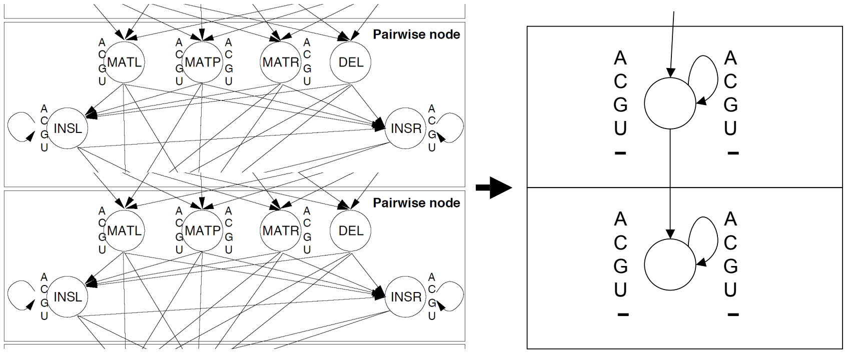
1 month, 1000 computers

~2 months, 1000 computers

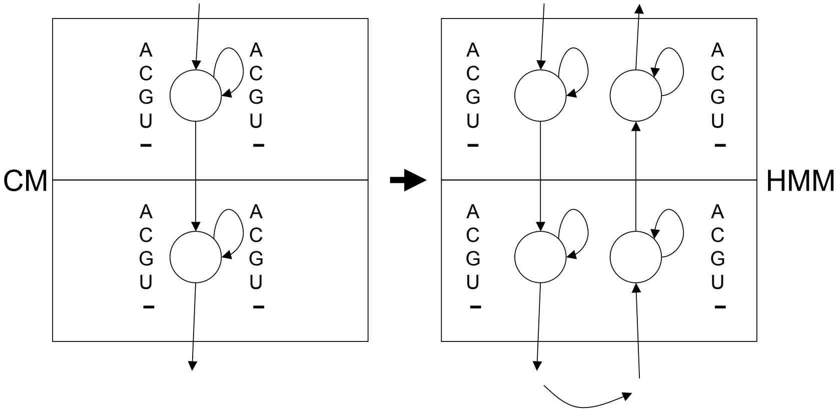
10 years, 1000 computers

Oversimplified CM

(for pedagogical purposes only)



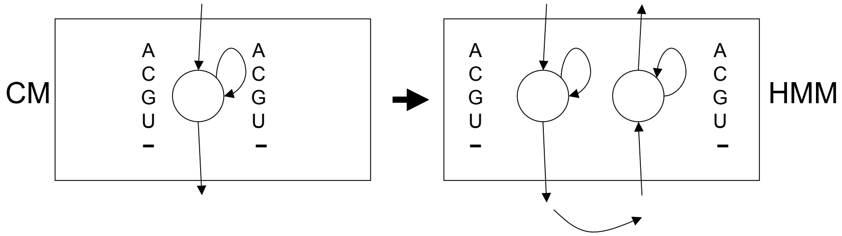
CM to HMM



25 emisions per state

5 emissions per state, 2x states

Key Issue: 25 scores ➔ 10

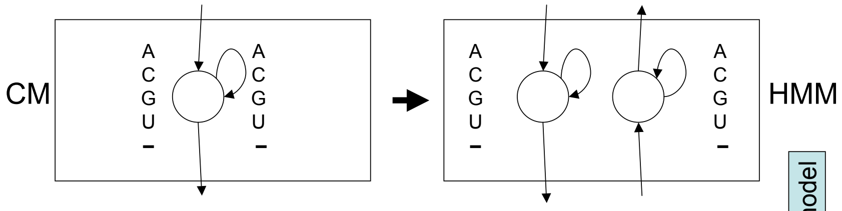


- Need: log Viterbi scores CM ≤ HMM

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
  $E(L_i, R_i) = \Sigma_x$ 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 $E(L_i, R_i)$

$E(L_i, R_i) = \Sigma_x$ 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…			