CSE P 590A
Fall 2008

RNA
Function,
Secondary Structure Prediction,
Search, Discovery

The Message

Cells make lots of ~~RNA~~ *noncoding* RNA

Functionally important, functionally diverse

Structurally complex

New tools required

alignment, discovery, search, scoring, etc.

The Outline

The problem: noncoding RNA

Why: it's important

Some results

Some methods

RNA

DNA: DeoxyriboNucleic Acid

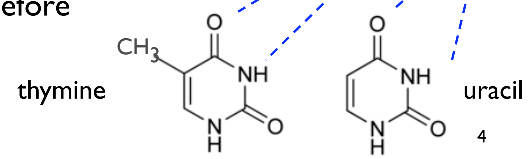
RNA: RiboNucleic Acid

Like DNA, except:

Lacks OH on ribose (backbone sugar)

Uracil (U) in place of thymine (T)

A, G, C as before

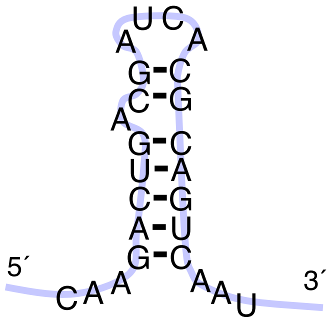


RNA Secondary Structure:
RNA makes helices too

Base pairs

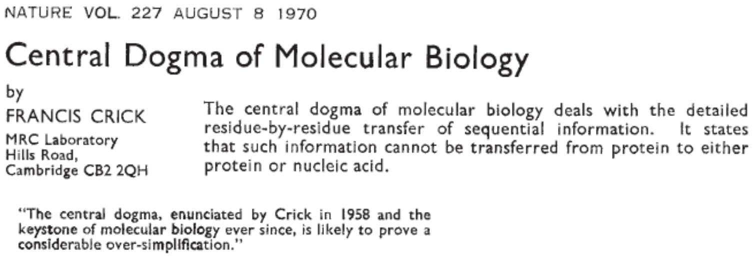
A-U
C-G

U-G



Usually *single* stranded

RNA: Interest



NATURE VOL. 227 AUGUST 8 1970

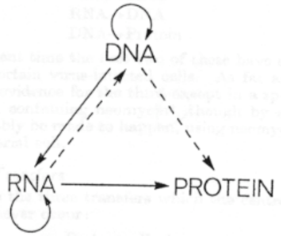
Central Dogma of Molecular Biology

by
FRANCIS CRICK
MRC Laboratory
Hills Road,
Cambridge CB2 2QH

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

"The central dogma, enunciated by Crick in 1958 and the keystone of molecular biology ever since, is likely to prove a considerable over-simplification."

Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.



"Classical" RNAs

rRNA - ribosomal RNA (~4 kinds, 120-5k nt)
tRNA - transfer RNA (~61 kinds, ~ 75 nt)
snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)
RNaseP - tRNA processing (~300 nt)
a handful of others

Bacteria

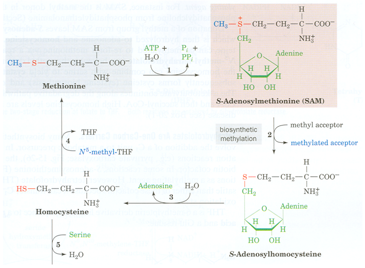
Triumph of proteins

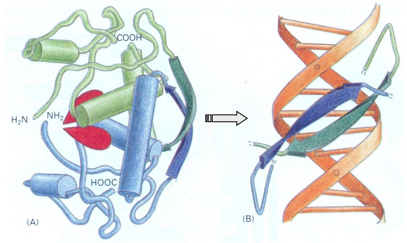
80% of genome is coding DNA

Functionally diverse

- receptors
- motors
- catalysts
- regulators (Monod & Jakob, Nobel prize 1965)
- …

Proteins catalyze & regulate biochemistry





Vertebrates

Bigger, more complex genomes

<2% coding

But >5% conserved in sequence?

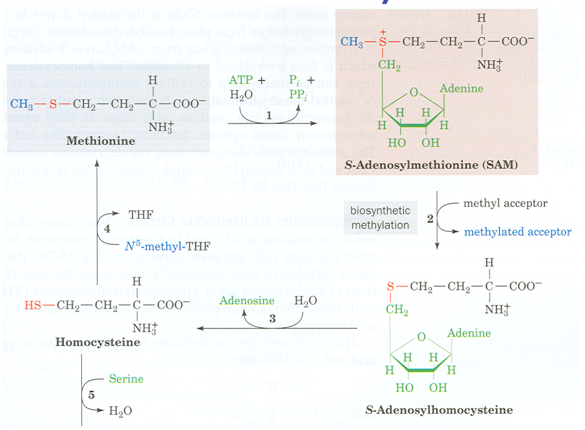
And 50-90% transcribed?

And *structural* conservation, if any, invisible
(without proper alignments, etc.)

What's going on?

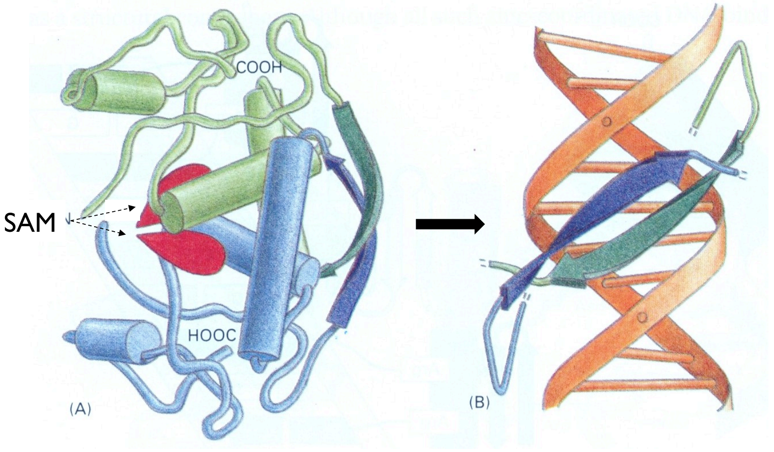
Bacteria Again:

Met Pathways



• • •

Gene Regulation: The MET Repressor

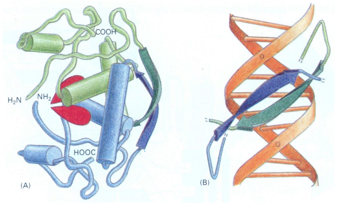


Protein

Alberts, et al, 3e.

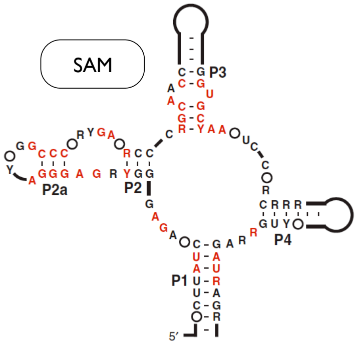
DNA

Alberts, et al, 3e.



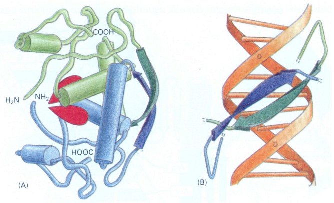
← The protein way

Riboswitch alternative



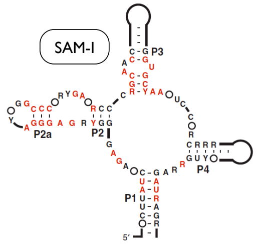
Grundy & Henkin, Mol. Microbiol 1998
Epshtein, et al., PNAS 2003
Winkler et al., Nat. Struct. Biol. 2003

Alberts, et al, 3e.

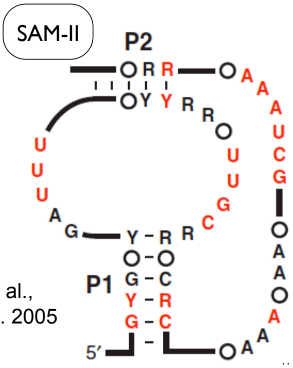


← The protein way

Riboswitch alternatives

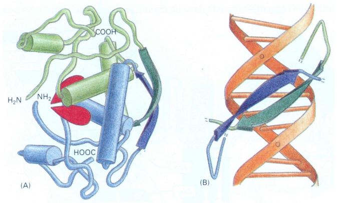


Grundy, Epshtein, Winkler
et al., 1998, 2003



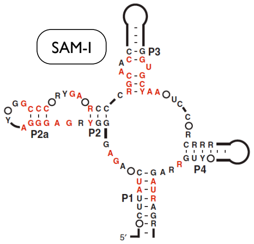
Corbino et al.,
Genome Biol. 2005

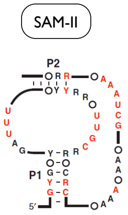
Alberts, et al, 3e.



← The protein way

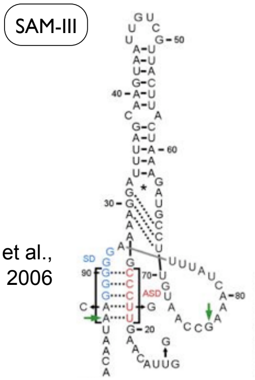
Riboswitch alternatives





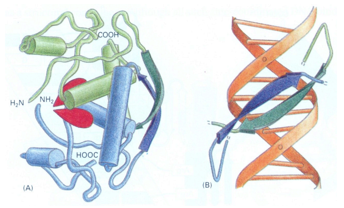
Grundy, Epshtein, Winkler
et al., 1998, 2003

Corbino et al.,
Genome Biol. 2005



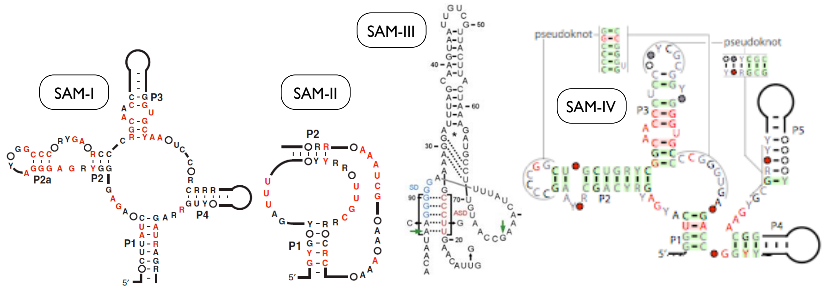
Fuchs et al.,
NSMB 2006

Alberts, et al, 3e.



← The protein way

Riboswitch alternatives ↓

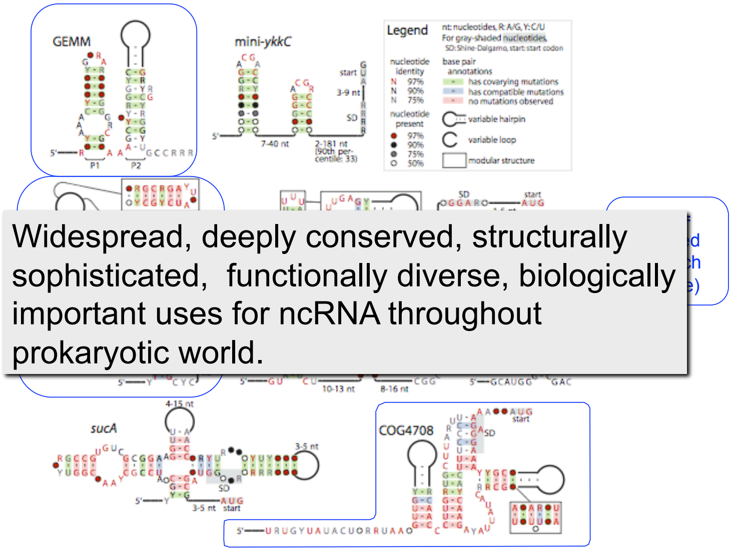


Grundy, Epshtein, Winkler et al., 1998, 2003

Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006

Weinberg et al., RNA 2008



Widespread, deeply conserved, structurally sophisticated, functionally diverse, biologically important uses for ncRNA throughout prokaryotic world.

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

Vertebrates

Bigger, more complex genomes

<2% coding

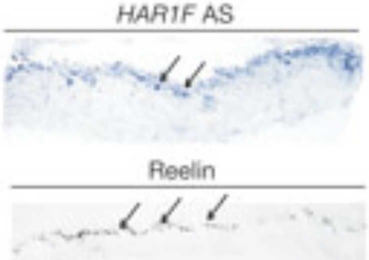
But >5% conserved in sequence?

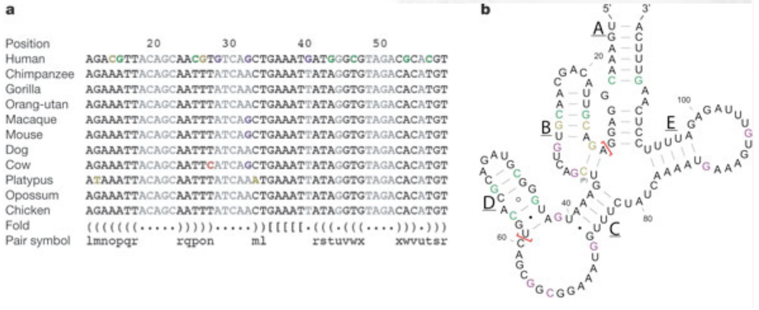
And 50-90% transcribed?

And *structural* conservation, if any, invisible
(without proper alignments, etc.)

What's going on?

Fastest Human Gene?





```
Position              20        30        40        50
Human        AGACGTTACAGCAACGTGTCAGCTGAAATGATGGGCGTAGACGCACGT
Chimpanzee   AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Gorilla      AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Orang-utan   AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Macaque      AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
Mouse        AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
Dog          AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Cow          AGAAATTACAGCAATTCATCAGCTGAAATTATAGGTGTAGACACATGT
Platypus     ATAAATTACAGCAATTTATCAAATGAAATTATAGGTGTAGACACATGT
Opossum      AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Chicken      AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
Fold         (((((((.....))))).....))[[[[[.(((.(((....))).)))
Pair symbol  lmnopqr     rqpon     ml      rstuvwx    xwvutsr
```

Vertebrate ncRNAs

mRNA, tRNA, rRNA, … of course

PLUS:

snRNA, spliceosome, snoRNA, teleomerase, microRNA, RNAi, SECIS, IRE, piwi-RNA, XIST (X-inactivation), ribozymes, …

MicroRNA

1st discovered 1992 in C. elegans

2nd discovered 2000, also C. elegans

- *and* human, fly, everything between

21-23 nucleotides

- literally fell off ends of gels

Hundreds now known in human

- may regulate 1/3-1/2 of all genes
- development, stem cells, cancer, infectious diseases,…

siRNA

"Short Interfering RNA"

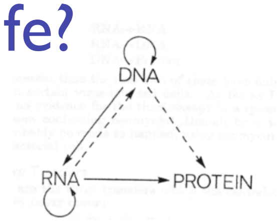
Also discovered in *C. elegans*

Possibly an antiviral defense, shares machinery with miRNA pathways

Allows artificial repression of most genes in most higher organisms

Huge tool for biology & biotech

Origin of Life?



Life needs

- information carrier: DNA
- molecular machines, like enzymes: Protein
- making proteins needs DNA + RNA + proteins
- making (duplicating) DNA needs proteins

Horrible circularities! How could it have arisen in an abiotic environment?

Origin of Life?

RNA can carry information, too

RNA double helix; RNA-directed RNA polymerase

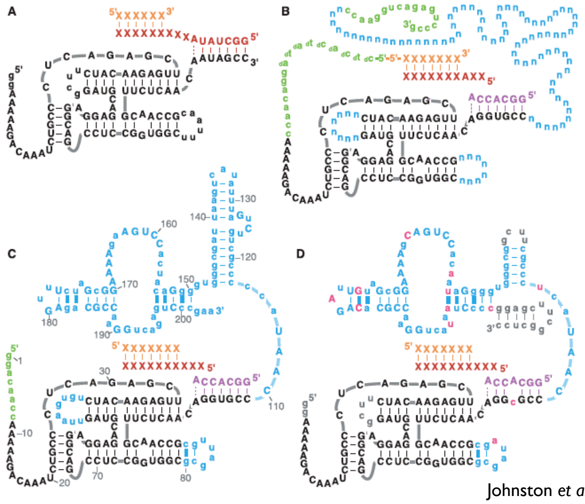
RNA can form complex structures

RNA enzymes exist (ribozymes)

RNA can control, do logic (riboswitches)

The “RNA world” hypothesis:
1st life was RNA-based

RNA replicase



Johnston *et al.*, *Science*, 2001

Outline

Biological roles for RNA

What is “secondary structure?

How is it represented?

Why is it important?

Examples

Approaches

“Classical” RNAs

tRNA - transfer RNA (~61 kinds, ~ 75 nt)

rRNA - ribosomal RNA (~4 kinds, 120-5k nt)

snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)

RNaseP - tRNA processing (~300 nt)

RNase MRP - rRNA processing; mito. rep. (~225 nt)

SRP - signal recognition particle; membrane targeting (~100-300 nt)

SECIS - selenocysteine insertion element (~65nt)

6S - ? (~175 nt)

Semi-classical RNAs

(discovery in mid 90’s)

tmRNA - resetting stalled ribosomes

Telomerase - (200-400nt)

snoRNA - small nucleolar RNA (many varieties; 80-200nt)

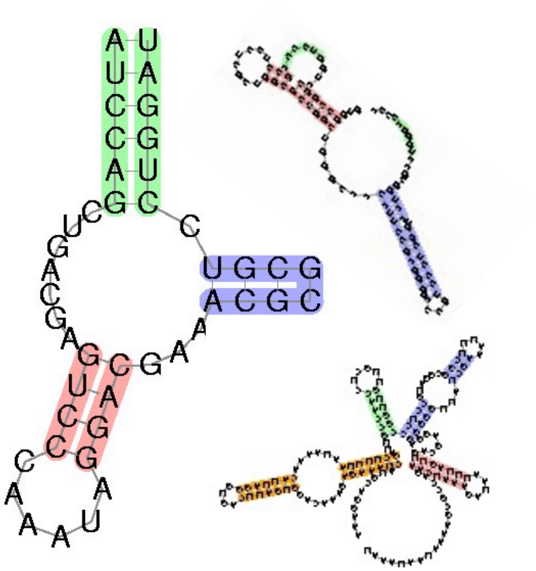
Recent discoveries

siRNA (Nobel prize 2006: Fire & Mello)
microRNAs (Lasker prize 2008:
Ambros, Baulcombe & Ruvkun)
riboswitches
many ribozymes
regulatory elements
…
Hundreds of families
Rfam release 1, 1/2003: 25 families, 55k instances
Rfam release 9, 7/2008, 603 families, 896k instances

Why?

RNA’s fold, and function

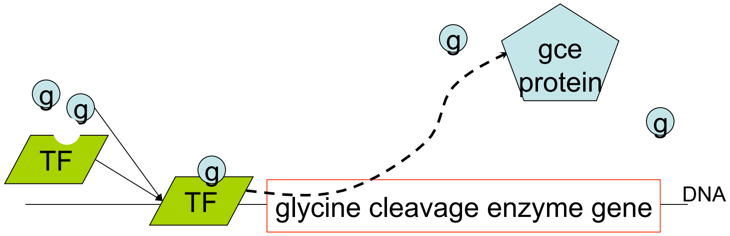
Nature uses what works



Example: Glycine Regulation

How is glycine level regulated?

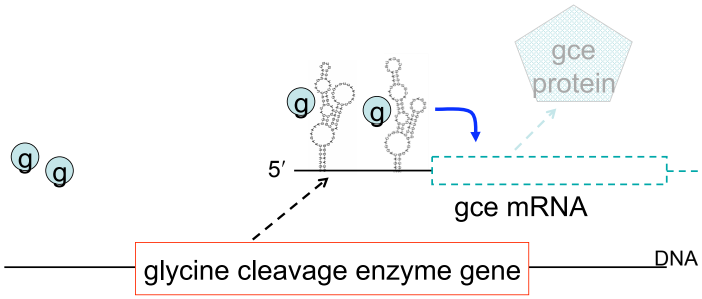
Plausible answer:



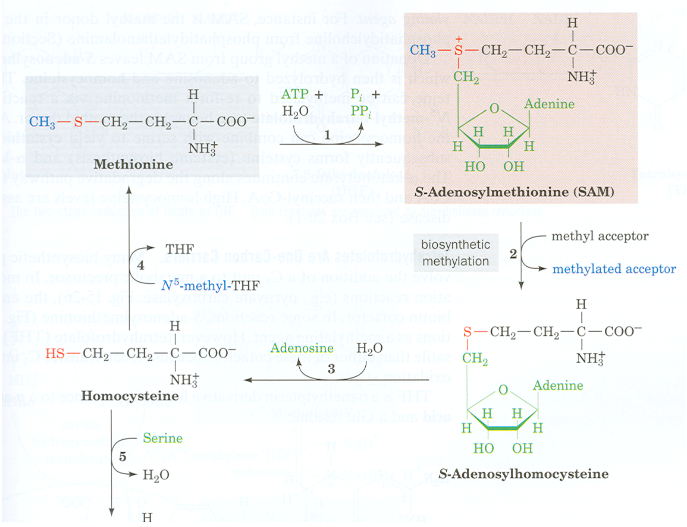
transcription factors (proteins) bind to DNA to turn nearby genes on or off

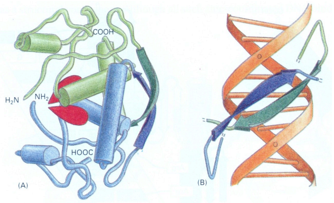
The Glycine Riboswitch

Actual answer (in many bacteria):



Mandal et al. Science 2004

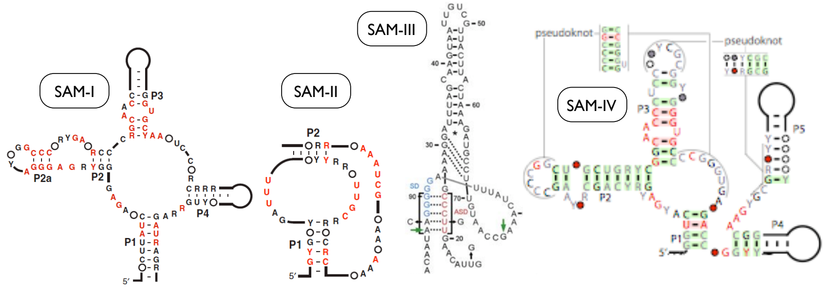




Alberts, et al, 3e.

← The protein way

Riboswitch alternatives ↓

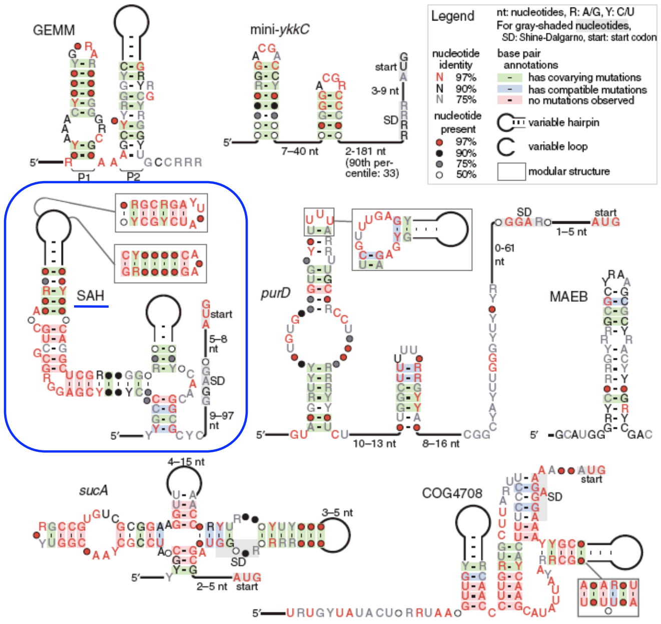


Grundy, Epshtein, Winkler et al., 1998, 2003

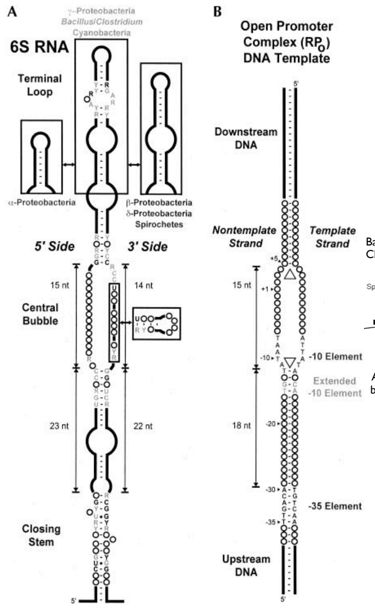
Corbino et al., Genome Biol. 2005

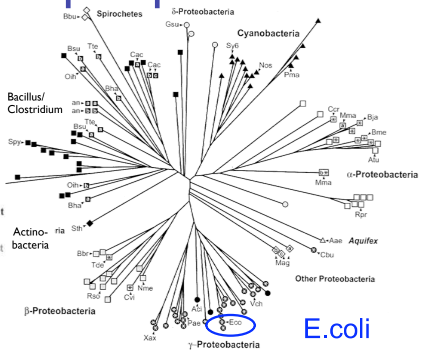
Fuchs et al., NSMB 2006

Weinberg et al., RNA 2008



6S mimics an open promoter





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

Wanted

Good structure prediction tools

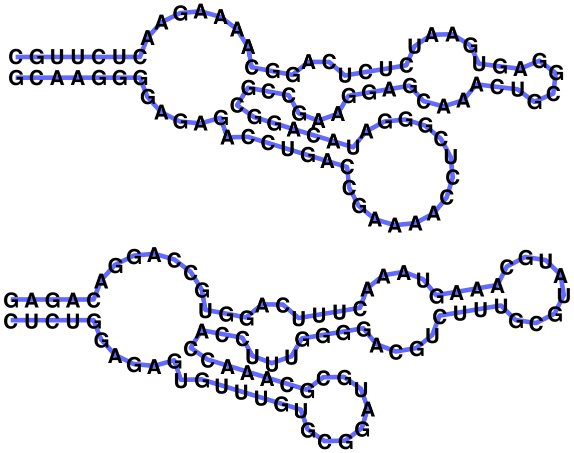
Good motif descriptions/models

Good, fast search tools
 (“RNA BLAST”, etc.)

Good, fast motif discovery tools
 (“RNA MEME”, etc.)

Importance of structure makes last 3 hard

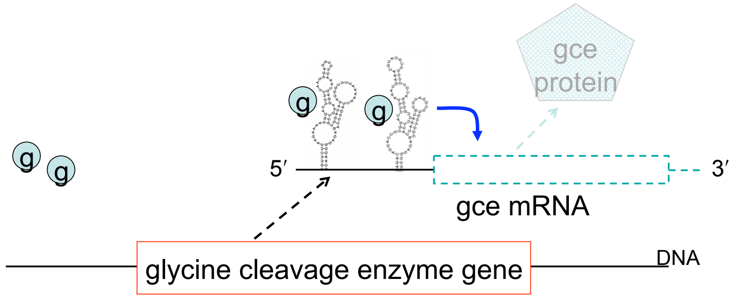
Why is RNA hard to deal with?



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

The Glycine Riboswitch

Actual answer (in many bacteria):



Mandal et al. Science 2004

Task 1: Structure Prediction

RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape

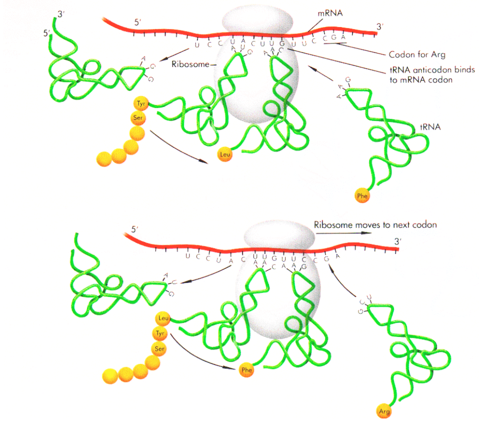
RNA Pairing

Watson-Crick Pairing

C - G	~ 3 kcal/mole
A - U	~ 2 kcal/mole
"Wobble Pair" G - U	~1 kcal/mole

Non-canonical Pairs (esp. if modified)

Ribosomes



Ribosomes

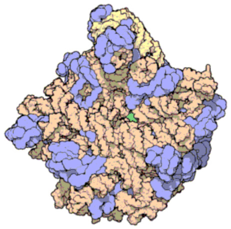
1974 Nobel prize to Romanian biologist George Palade for discovery in mid 50’s

50-80 proteins

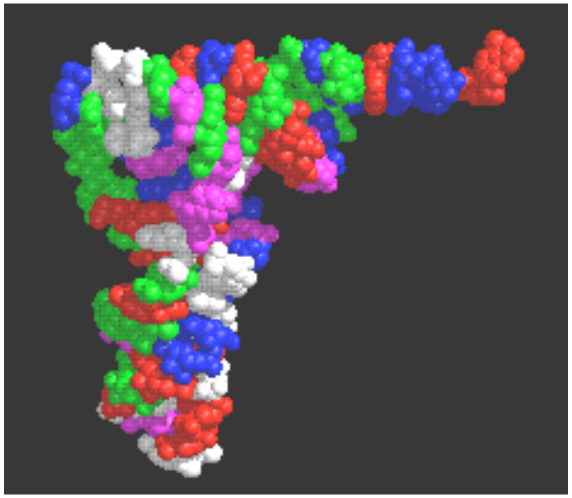
3-4 RNAs (half the mass)

Catalytic core is RNA

Of course, mRNAs and tRNAs (messenger & transfer RNAs) are critical too



tRNA 3d Structure



tRNA - Alt. Representations

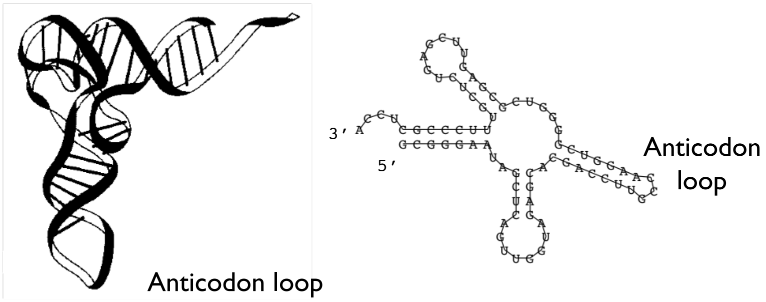
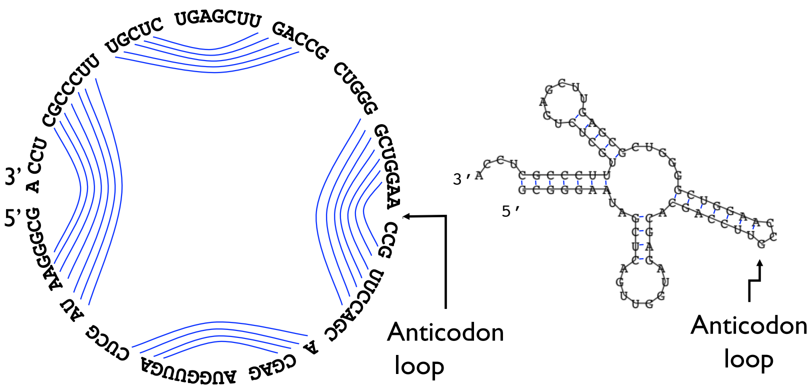


Figure 1: a) The spatial structure of the phenylalanine tRNA form yeast

b) The secondary structure extracts the most important information about the structure, namely the pattern of base pairings.

tRNA - Alt. Representations



RNA Pairing

Watson-Crick Pairing

C - G ~ 3 kcal/mole

A - U ~ 2 kcal/mole

“Wobble Pair” G - U ~ 1 kcal/mole

Non-canonical Pairs (esp. if modified)

Definitions

Sequence $^{5'}\, r_1\, r_2\, r_3 \ldots r_n\, ^{3'}$ in {A, C, G, T}

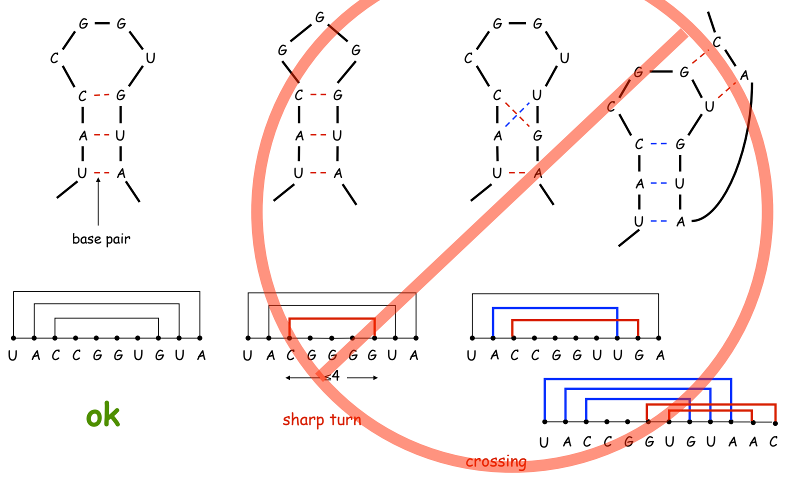
A Secondary Structure is a set of pairs i•j s.t.

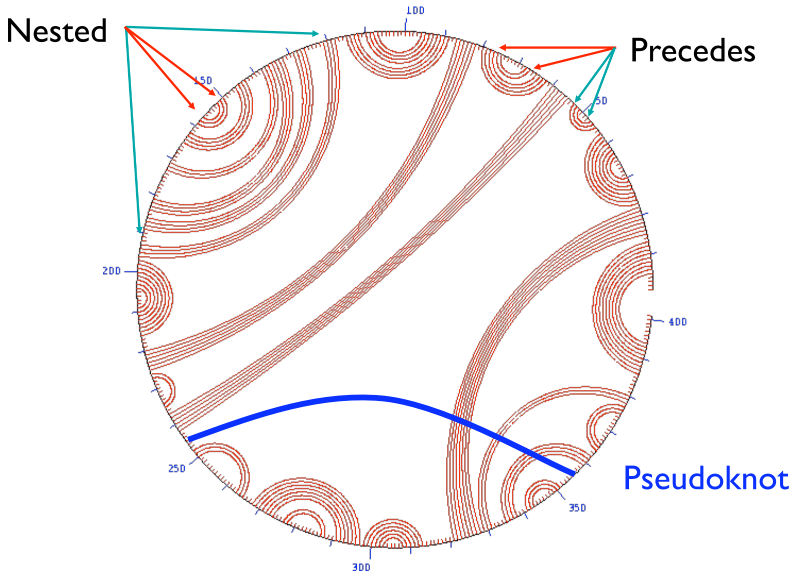
- i < j-4, and } no sharp turns
- if i•j & i’•j’ are two different pairs with i ≤ i’, then
  - j < i’, or
  - i < i’ < j’ < j

  } 2nd pair follows 1st, or is nested within it; no “pseudoknots.”

RNA Secondary Structure: Examples

Examples.





Approaches to Structure Prediction

Maximum Pairing
- + works on single sequences
- + simple
- - too inaccurate

Minimum Energy
- + works on single sequences
- - ignores pseudoknots
- - only finds "optimal" fold

Partition Function
- + finds all folds
- - ignores pseudoknots

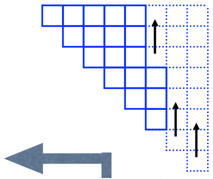
Nussinov: Max Pairing

B(i,j) = # pairs in optimal pairing of $r_i \dots r_j$

B(i,j) = 0 for all i, j with $i \geq j-4$; otherwise

B(i,j) = max of:

$$\begin{cases} B(i,j-1) \\ \max \{ B(i,k-1)+1+B(k+1,j-1) \mid \\ \quad i \leq k < j-4 \text{ and } r_k\text{-}r_j \text{ may pair}\} \end{cases}$$

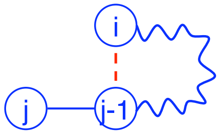


Time: $O(n^3)$

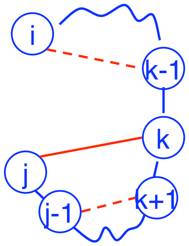
"Optimal pairing of $r_i \dots r_j$"

Two possibilities

j Unpaired:
Find best pairing of $r_i \dots r_{j-1}$



j Paired (with some k):
Find best $r_i \dots r_{k-1}$ +
best $r_{k+1} \dots r_{j-1}$ plus 1



Why is it slow?
Why do pseudoknots matter?

Pair-based Energy Minimization

E(i,j) = energy of pairs in optimal pairing of $r_i \dots r_j$

E(i,j) = ∞ for all i, j with $i \geq j-4$; otherwise

E(i,j) = min of:

energy of j-k pair

$$\begin{cases} E(i,j-1) \\ \min \{ E(i,k-1) + e(r_k, r_j) + E(k+1,j-1) \mid i \leq k < j-4 \} \end{cases}$$

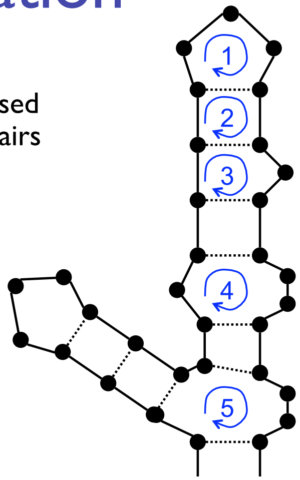
Time: $O(n^3)$

Loop-based Energy Minimization

Detailed experiments show it's more accurate to model based on loops, rather than just pairs

Loop types

1. Hairpin loop
2. Stack
3. Bulge
4. Interior loop
5. Multiloop



Zuker: Loop-based Energy, I

W(i,j) = energy of optimal pairing of $r_i \ldots r_j$

V(i,j) = as above, but forcing pair i•j

W(i,j) = V(i,j) = ∞ for all i, j with i ≥ j-4

$$W(i,j) = \min(W(i,j-1),\ \min\{\, W(i,k-1)+V(k,j) \mid i \le k < j-4 \,\}\ )$$

Zuker: Loop-based Energy, II

hairpin stack bulge/interior multi-loop

$V(i,j) = \min(eh(i,j), es(i,j)+V(i+1,j-1), VBI(i,j), VM(i,j))$

$VM(i,j) = \min\{\, W(i,k)+W(k+1,j) \mid i < k < j \,\}$

$VBI(i,j) = \min\{\, ebi(i,j,i',j') + V(i', j') \mid i < i' < j' < j \;\&\; i'-i+j-j' > 2 \,\}$

bulge/interior

Time: $O(n^4)$

$O(n^3)$ possible if ebi(.) is "nice"

Energy Parameters

Q. Where do they come from?

A1. Experiments with carefully selected synthetic RNAs

A2. Learned algorithmically from trusted alignments/structures

Accuracy

Latest estimates suggest ~50-75% of base pairs predicted correctly in sequences of up to ~300nt

Definitely useful, but obviously imperfect

Approaches to Structure Prediction

Maximum Pairing
- + works on single sequences
- + simple
- - too inaccurate

Minimum Energy
- + works on single sequences
- - ignores pseudoknots
- - only finds "optimal" fold

Partition Function
- + finds all folds
- - ignores pseudoknots

Approaches, II

Comparative sequence analysis
- + handles all pairings (incl. pseudoknots)
- - requires several (many?) aligned, appropriately diverged sequences

Stochastic Context-free Grammars

Roughly combines min energy & comparative, but no pseudoknots

Physical experiments (x-ray crystalography, NMR)

Summary

RNA has important roles beyond mRNA

Many unexpected recent discoveries

Structure is critical to function

True of proteins, too, but they're easier to find, due, e.g., to codon structure, which RNAs lack

RNA secondary structure can be predicted (to useful accuracy) by dynamic programming

Next: RNA "motifs" (seq + 2-ary struct) well-captured by "covariance models"

“RNA sequence analysis using covariance models”

Eddy & Durbin

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

(see also, Ch 10 of Durbin *et al.*)

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

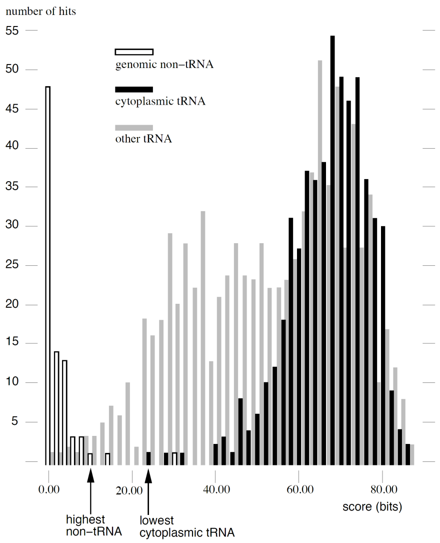
You set a score threshold

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



Profile Hmm Structure

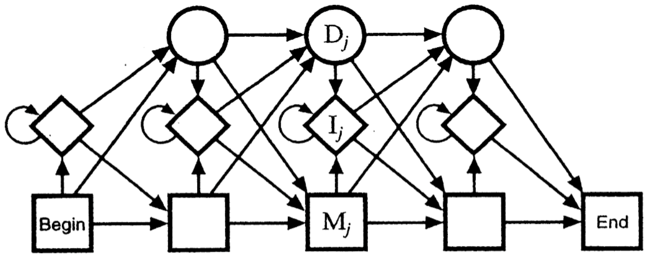


Figure 5.2 *The transition structure of a profile HMM.*

Mj: Match states (20 emission probabilities)

Ij: Insert states (Background emission probabilities)

Dj: Delete states (silent - no emission)

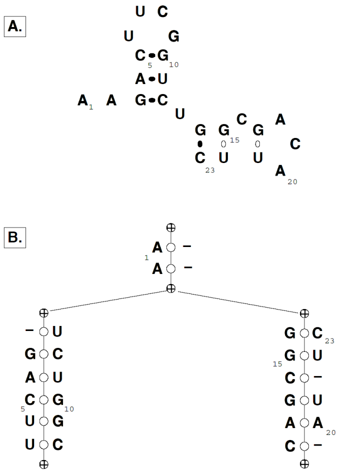
CM Structure

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)

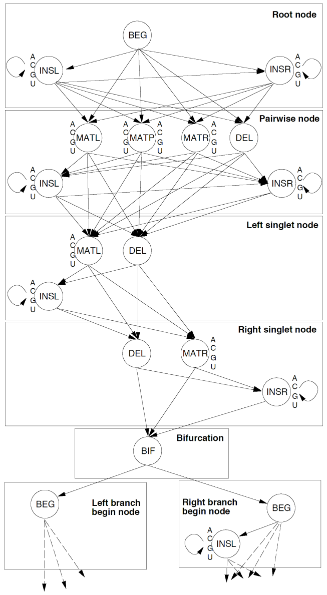


Overall CM Architecture

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

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

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

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

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

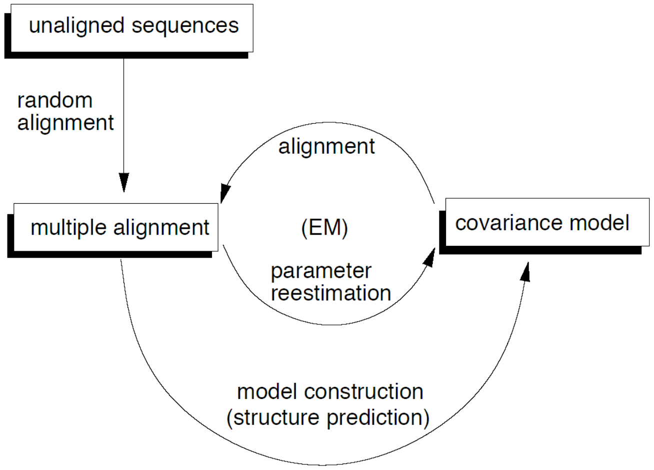
$S^y_{ij} \quad = \max_\pi \log P(x_{ij} \text{ gen'd starting in state } y \text{ via path } \pi)$

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

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

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

Model Training



unaligned sequences

random alignment

multiple alignment

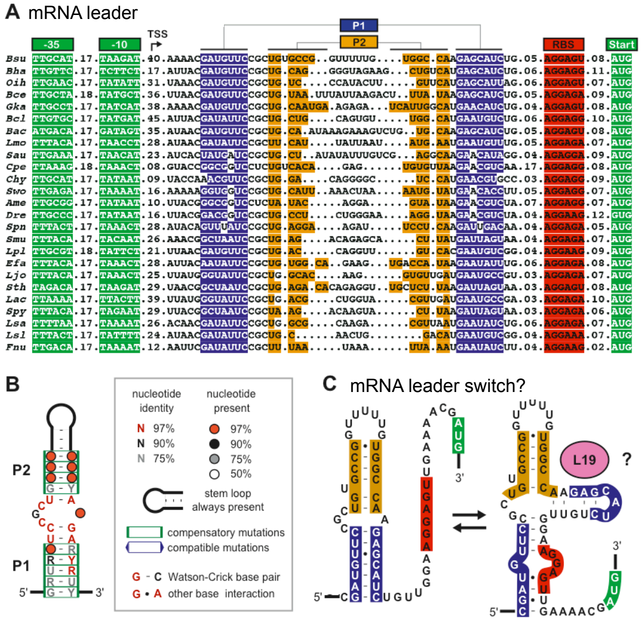
alignment

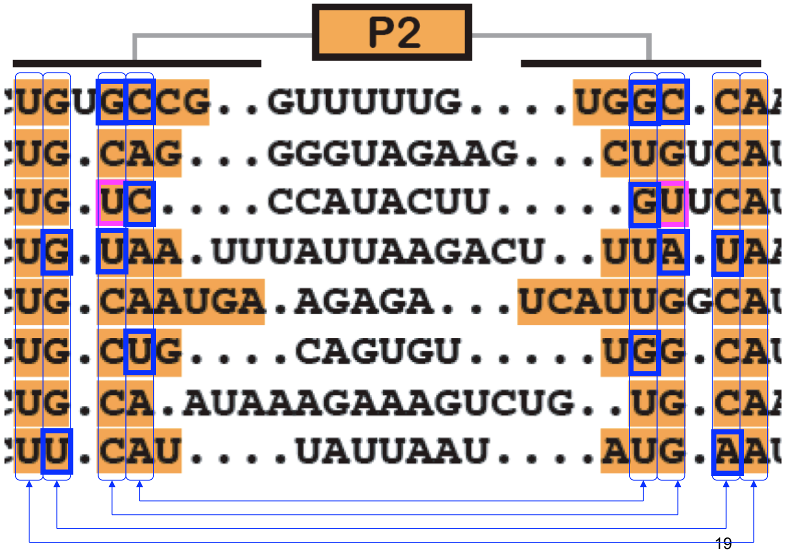
(EM)

parameter reestimation

covariance model

model construction (structure prediction)




P2
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

Mutual Information

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

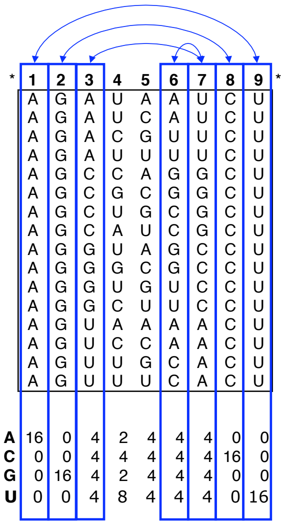
Max when *no* seq conservation but perfect pairing

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	

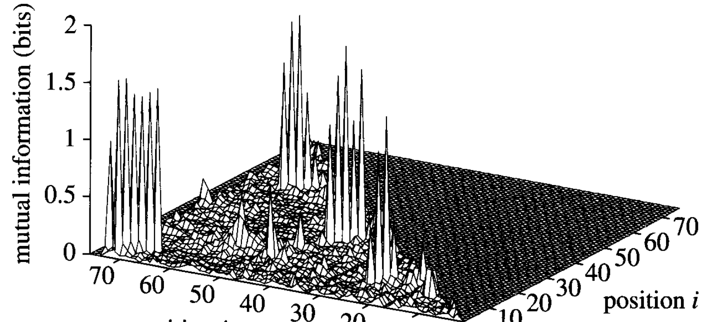
	1	2	3	4	5	6	7	8	9
A	16	0	4	2	4	4	4	0	0
C	0	0	4	4	4	4	4	16	0
G	0	16	4	2	4	4	4	0	0
U	0	0	4	8	4	4	4	0	16

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.



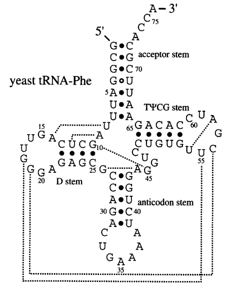


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary 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

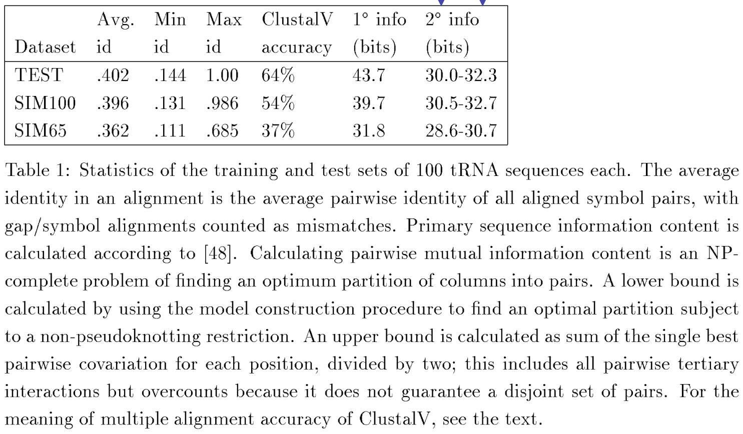
$$S_{i,j} = \max\begin{cases} S_{i,j-1} \\ \max_{i\le k<j-4} S_{i,k-1} + M_{k,j} + S_{k+1,j-1} \end{cases}$$

"Just like Nussinov/Zucker folding"

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

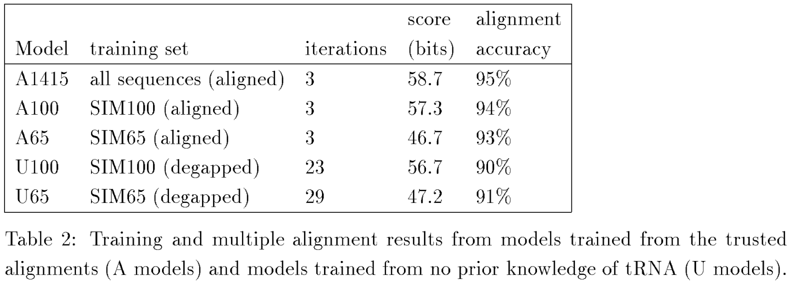
Pseudoknots

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).

Rfam – an RNA family DB

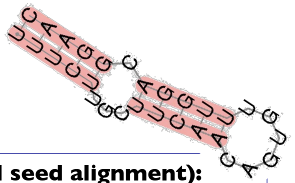
Griffiths-Jones, et al., NAR '03,'05

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

Rfam



Input (hand-curated):

- MSA "seed alignment"
- SS_cons
- Score Thresh T
- Window Len W

Output:

- CM
- scan results & "full alignment"

IRE (partial seed alignment):

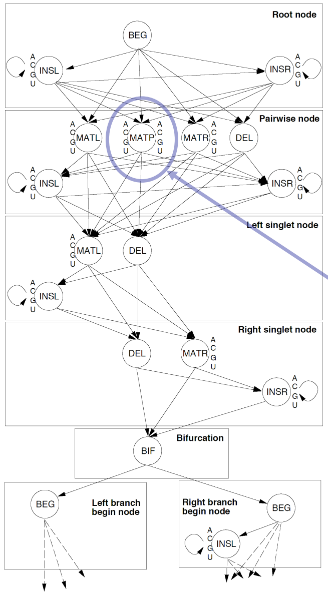
```
Hom.sap.  GUUCCUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.  UUUCUUC.UUCAACAGUGUUUGGAUGGAAC
Hom.sap.  UUUCCUGUUUCAACAGUGCUUGGA.GGAAC
Hom.sap.  UUUAUC..AGUGACAGAGUUCACU.AUAAA
Hom.sap.  UCUCUUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.  AUUAUC..GGGAACAGUGUUUCCC.AUAAU
Hom.sap.  UCUUGC..UUCAACAGUGUUUGGACGGAAG
Hom.sap. 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

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

Zasha Weinberg

& W.L. Ruzzo

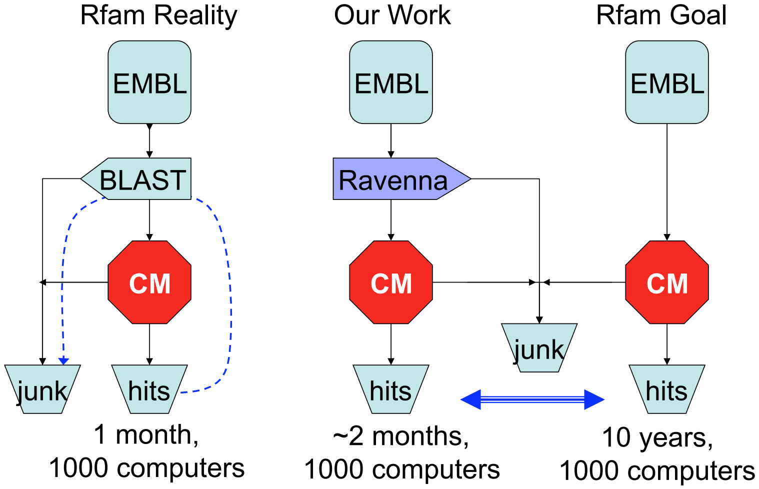
Recomb '04, ISMB '04, Bioinfo '06



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



Rfam Reality

EMBL

BLAST

CM

junk

hits

1 month, 1000 computers

Our Work

EMBL

Ravenna

CM

junk

hits

~2 months, 1000 computers

Rfam Goal

EMBL

CM

hits

10 years, 1000 computers

Results: New ncRNA’s?

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

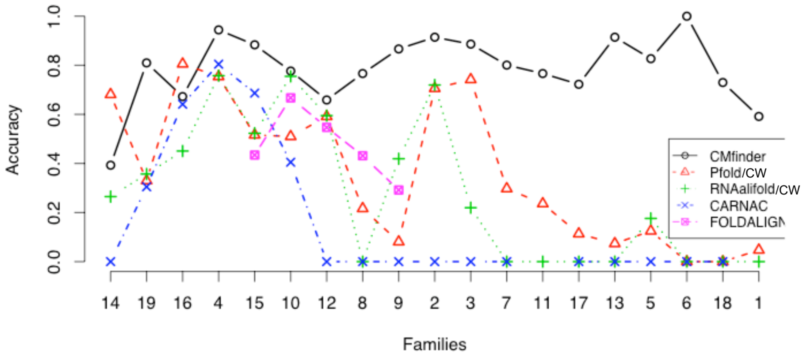
Cmfinder--A Covariance Model Based RNA Motif Finding Algorithm

Bioinformatics, 2006, 22(4): 445-452

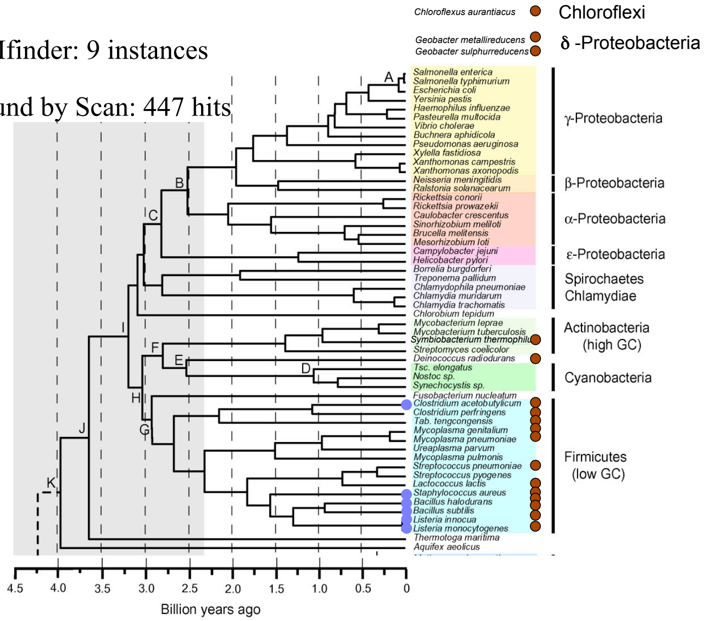
Zizhen Yao
Zasha Weinberg
Walter L. Ruzzo
University of Washington, Seattle

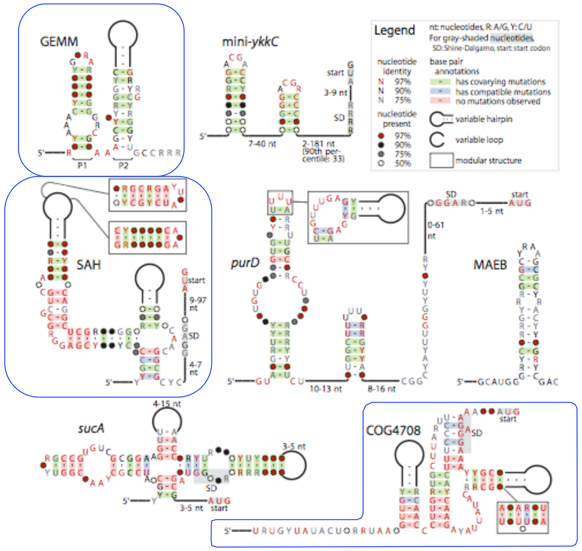
CMfinder Accuracy

(on Rfam families *with* flanking sequence)



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





boxed = confirmed riboswitch (+2 more)

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

Search in Vertebrates

Extract ENCODE Multiz alignments

Remove exons, most conserved elements.

56017 blocks, 8.7M bps.

Apply CMfinder to both strands.

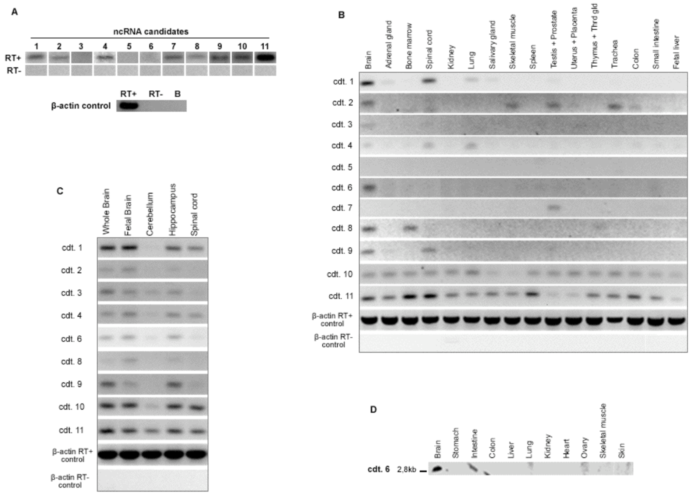
Trust 17-way alignment for orthology, not for detailed alignment

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:
O(1000) CPU years. Analysis in progress.)

10 of 11 top expressed.



Summary

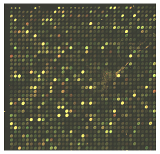
ncRNA - apparently widespread, much interest

Covariance Models - powerful but expensive tool for ncRNA motif representation, search, discovery

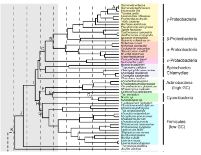
Rigorous/Heuristic filtering - typically 100x speedup in search with no/little loss in accuracy

CMfinder - CM-based motif discovery in unaligned sequences

Course Wrap Up

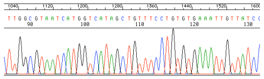


“High-Throughput BioTech”



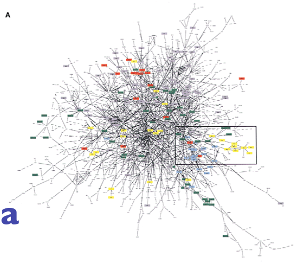
Sensors

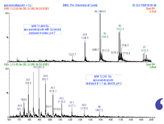
- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction



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

Exciting Times

Lots to do

Various skills needed

I hope I’ve given you a taste of it

Thanks!