

RNA Genomics

BME 110: CompBio Tools

Todd Lowe

May 13, 2009

Admin

- WebCT quiz on Tuesday cover reading, using Jalview & Pfam
- Homework #3 assigned today – due next Friday (8 days)

In Genomes, Two Types of Genes

Protein-coding:

[Start codon] [codon 1] [codon 2] [...] [Stop codon]

+ DNA codons translated to amino acids to form a protein

Non-coding RNAs (NcRNAs)

No consistent patterns common to all RNA genes

+ Not translated to proteins, functional as RNA molecules

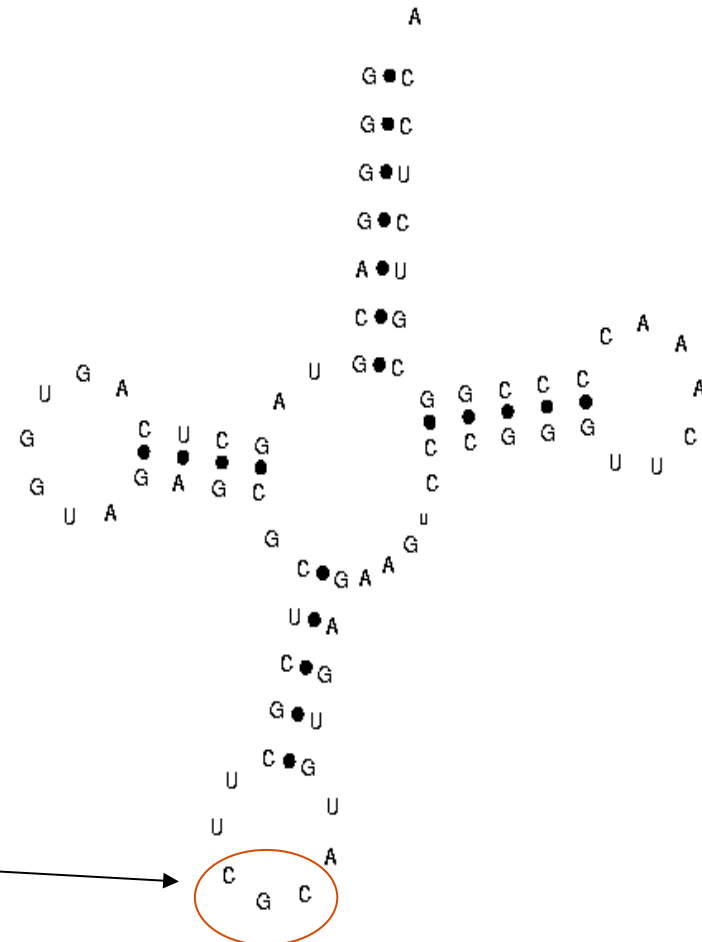
Rules of Engagement

- Alphabet: A, C, G, U (T~U)
- Canonical pairings: A-U, G-C
- Non-canonical: G-U (very common), A-C, A-G, etc.
- ncRNAs are riddled with nucleotide modifications which may affect H-bond base pairing
- Depending on the modification, can make non-canonical pairings more *or* less favorable
Example: wobble base modif. in tRNAs

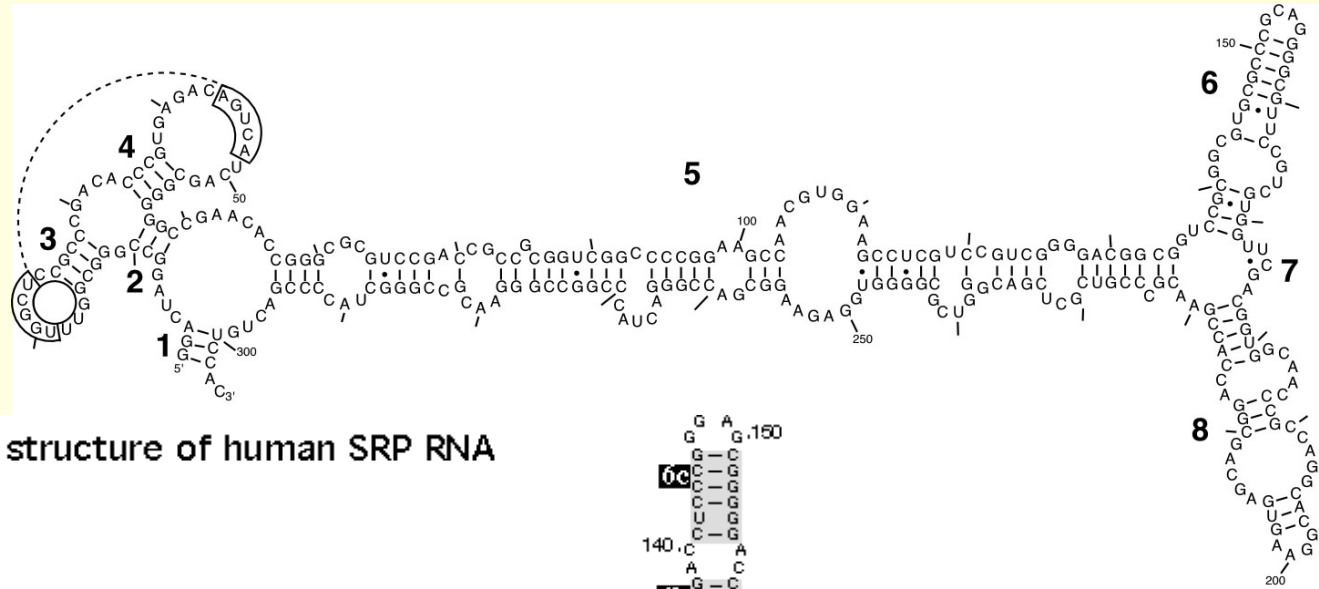
Features of ncRNAs

- Secondary structure is conserved, but often not primary sequence
- Additional sources of functional information easily derived from sequence and/or structure (i.e. anticodon in tRNA)

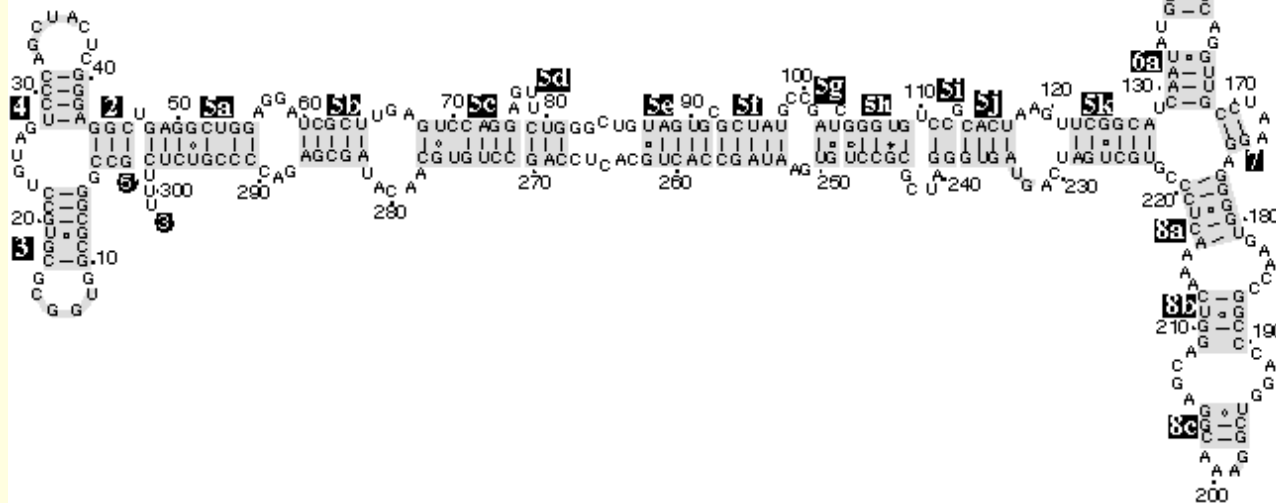
Drosophila-tRNA Ala (CGC) 74.79 bits



SRP RNA Orthologs: Same Function, Similar Structure, < 70% sequence identity



Secondary structure of human SRP RNA



Halobacterium halobium SRP RNA
(SRPDB, March 10, 2000)

Common Attributes of ncRNAs

- Structural – many fold into complex 3-D structures
- Antisense – many form specific base pairings with external “target” RNAs, no structure required for function
- Catalytic – some catalyze biochemical reactions “ribozymes” (RNaseP, group I & II introns, hammerhead RNA, & others)
- Regulatory – some interact with DNA or RNA targets for gene regulation (microRNAs, many bacterial small RNAs)

RNA + Proteins = Complexes

- RNAs usually do not act alone, but are complexed with proteins
- Sites of interaction with RNAs or proteins exert selective pressure to conserve sequence or structure
- Sites of interaction make up conserved motifs one looks for to help identify RNA and/or determine function

How To Keep Track of All Types of RNA?

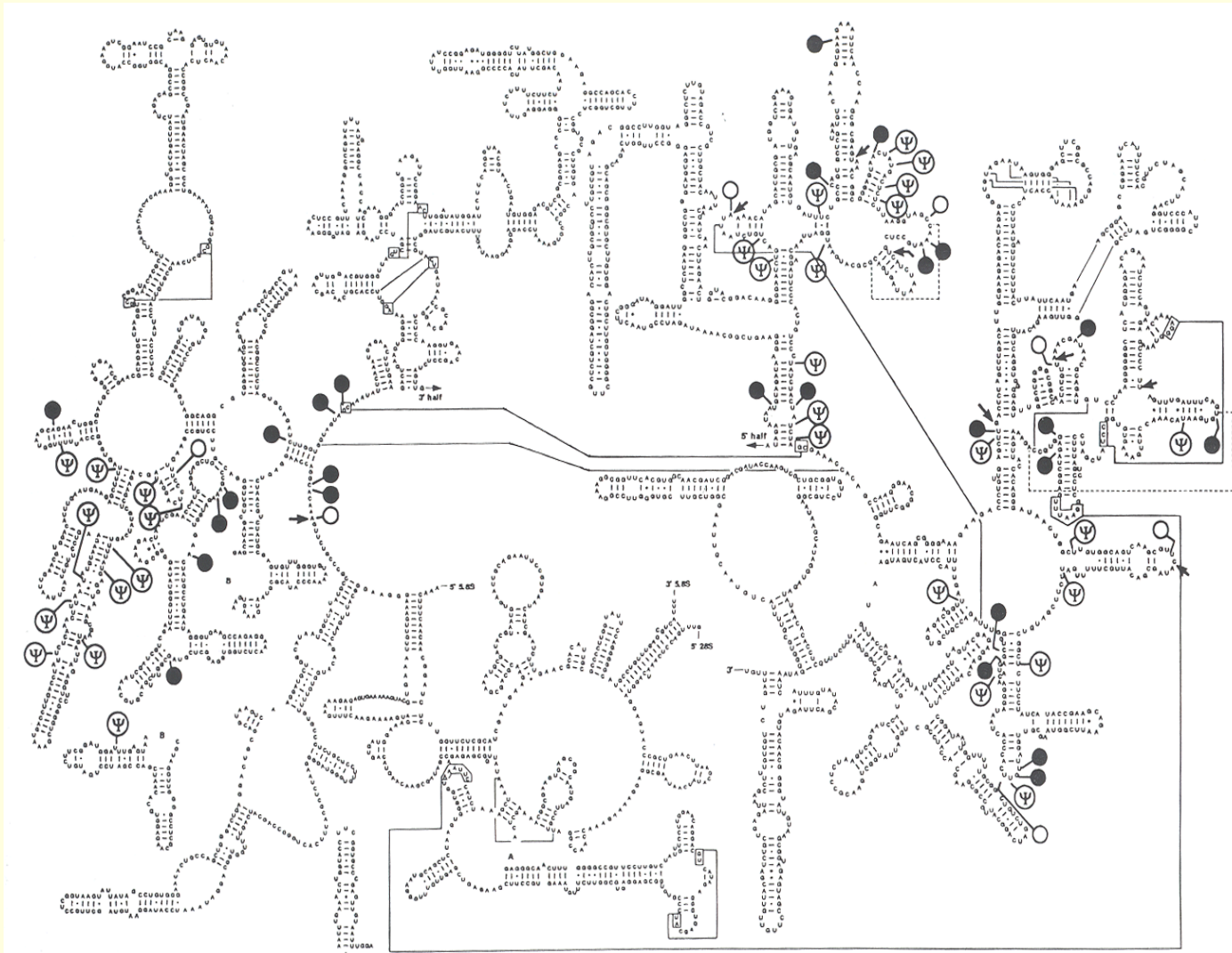
- General Database: RFAM
<http://rfam.sanger.ac.uk/>
- Attempt to catalogue, align, and create search models for all known RNAs
- Just like “PFAM” (protein sister site) can paste your sequence in and look for matches to RNA gene models
- Nice alignments, excellent resource
- Sometimes not as good as customized search programs, when they exist

Examples of ncRNA Genes

Ribosomal RNAs (rRNAs) – protein translation

- Examples : large subunit rRNA (aka LSU/28S/25S/23S), small subunit rRNA (aka SSU/18S/16S), 5.8S, and 5S rRNA
- Highly structured scaffold for dozens of ribosomal proteins; catalytic role in forming peptide bonds

Baker's Yeast (*S. cerevisiae*) Large Subunit rRNA 2-D structure

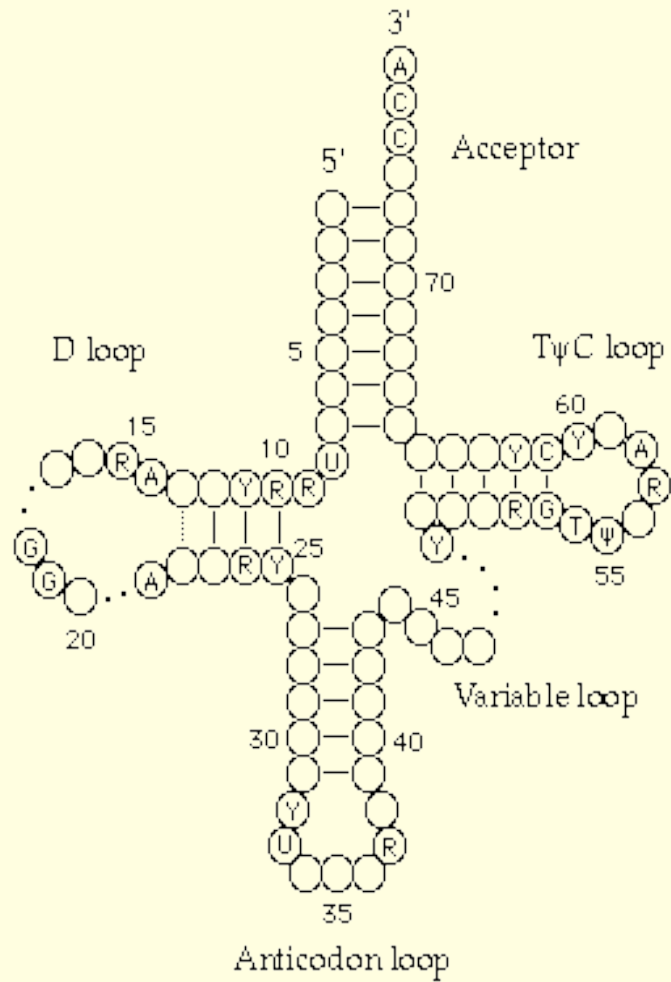


tRNAs: A well-studied ncRNA Gene family

Transfer RNAs (tRNAs) – “decode” mRNA codons into cognate amino acids in protein translation

- Examples: tRNA-AGC(Ala), tRNA-ACC(Gly)...
(62 kinds possible)
- Structured RNA and “antisense” interaction between tRNA anticodon and mRNA codon
- Can find them easily using the tRNAscan-SE web server:
<http://lowelab.ucsc.edu/tRNAscan-SE/>
- Can get example sequences from the Genomic tRNA Database: <http://gtrnadb.ucsc.edu/>

tRNA “cloverleaf” 2-D structure



3-D structure

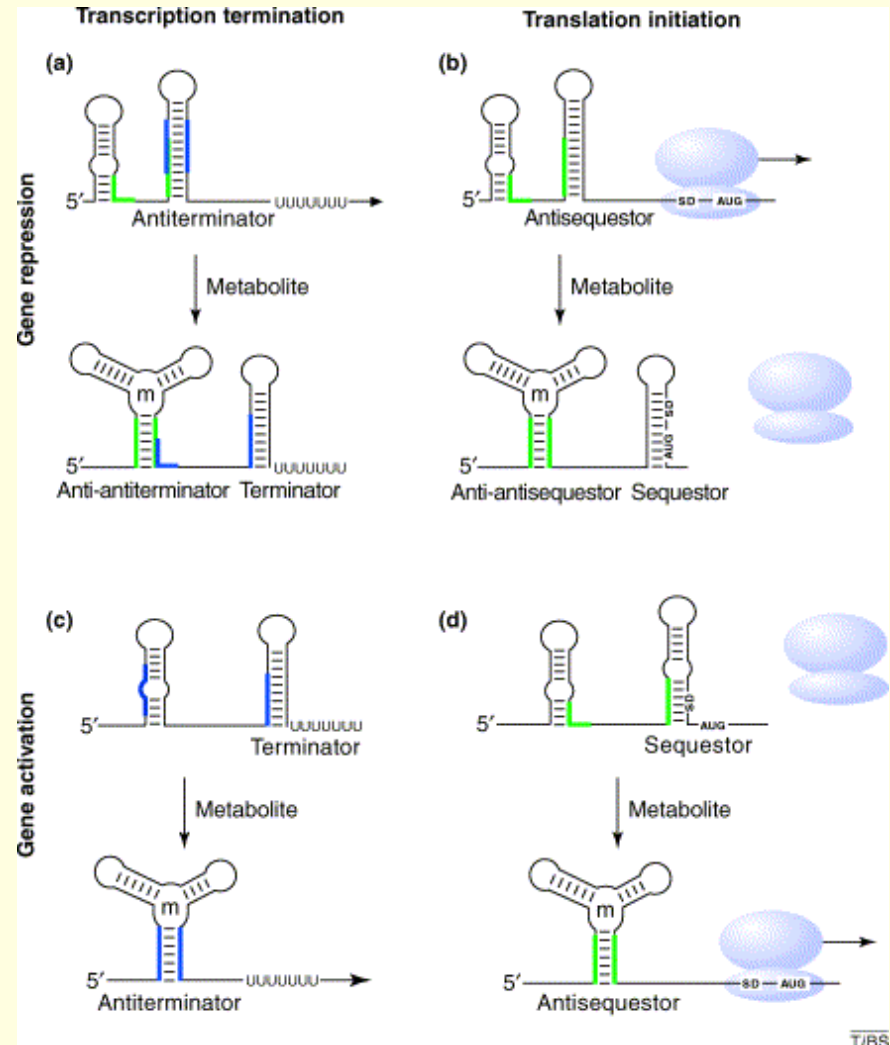


Some Functional RNA Structures are found in mRNAs

- Specific secondary structure can act with regulatory roles
- Examples
 - Riboswitches (example: TPP riboswitch)
 - Selenocysteine insertion sequences (SECIS)
 - Iron Response Element (IRE)

Many Different Types of Riboswitches

- Functional RNA is part of mRNA molecule, *usually* at 5' end
- Riboswitch functions by binding a specific small molecule (i.e. free lysine, or free purine nucleotides)
- Binding state of riboswitch modulates on/off control of transcription or translation of attached gene



From Nudler & Mironov, *TIBS*
29:11-17, 2004.

Riboswitch Web Analyses

Ribex:

<http://www.ibt.unam.mx/biocomputo/ribex.html>

Riboswitch Finder:

<http://riboswitch.bioapps.biozentrum.uni-wuerzburg.de/>

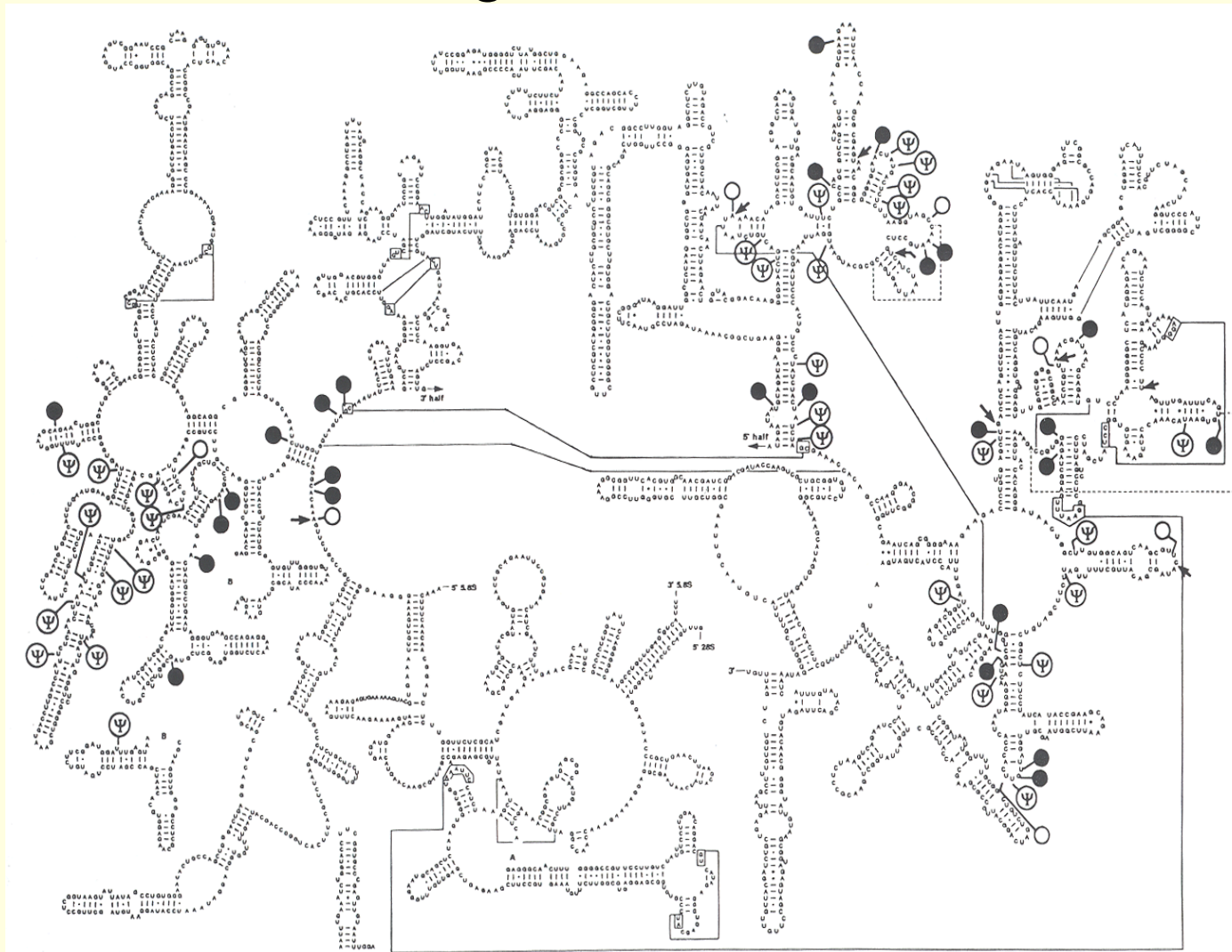
Other kinds of ncRNAs

- Small nuclear RNAs (snRNAs)

Examples:

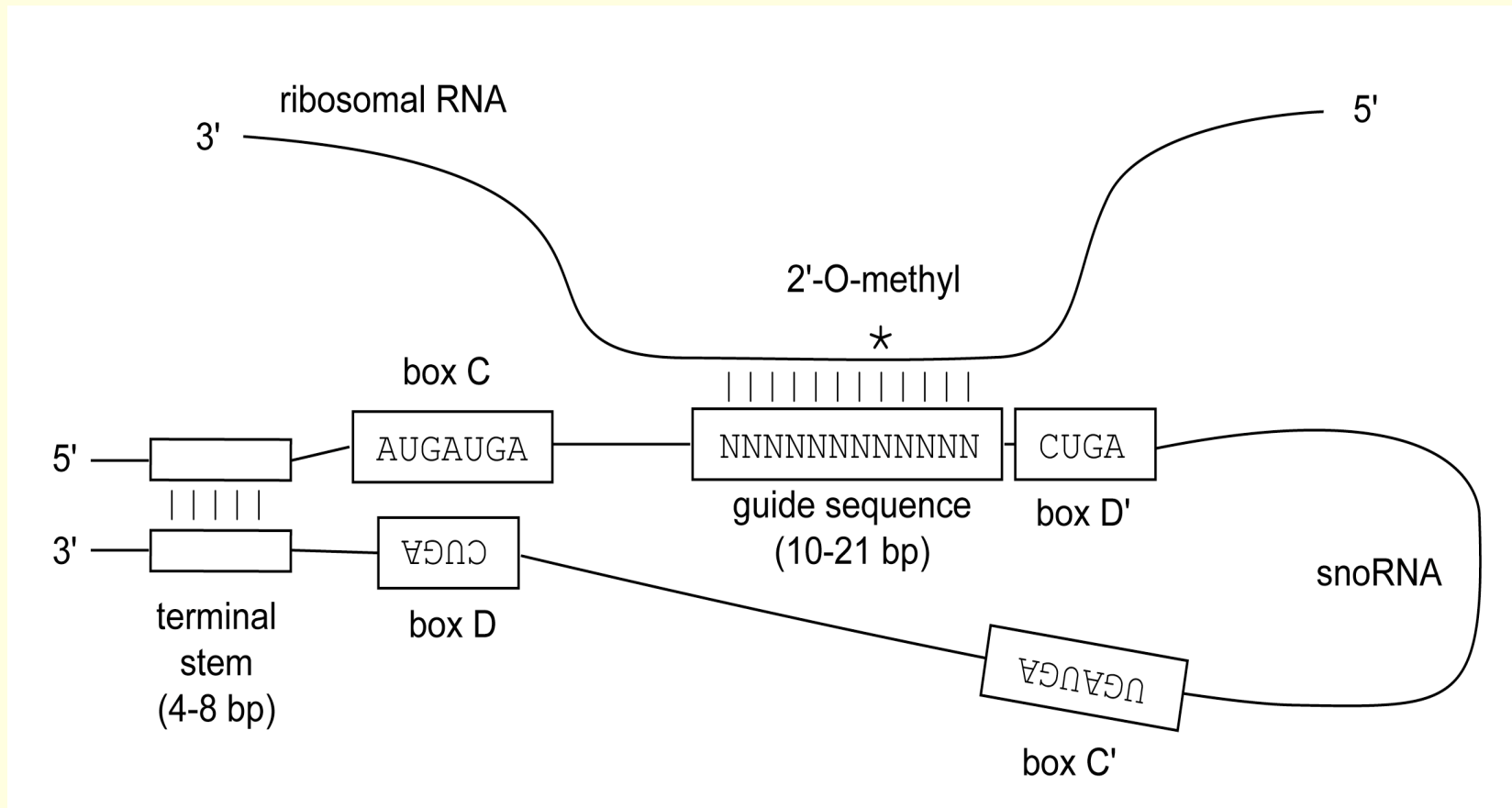
- Spliceosomal RNAs: U1, U2, U4, U5, U6, & others found in spliceosome, involved in removing introns from mRNAs in eukaryotes
- Can usually detect by BLASTN since highly conserved, like ribosomal RNA
- Small nucleolar RNAs (snoRNAs): located in the nucleolus, these process ribosomal RNA (chopping up & adding nucleotide modifications)
- BLASTN only works for some family members from very closely related species

S. cerevisiae Large Subunit Ribosomal RNA



- Highly conserved primary sequence and secondary structure
- Many interacting molecules constrain changes

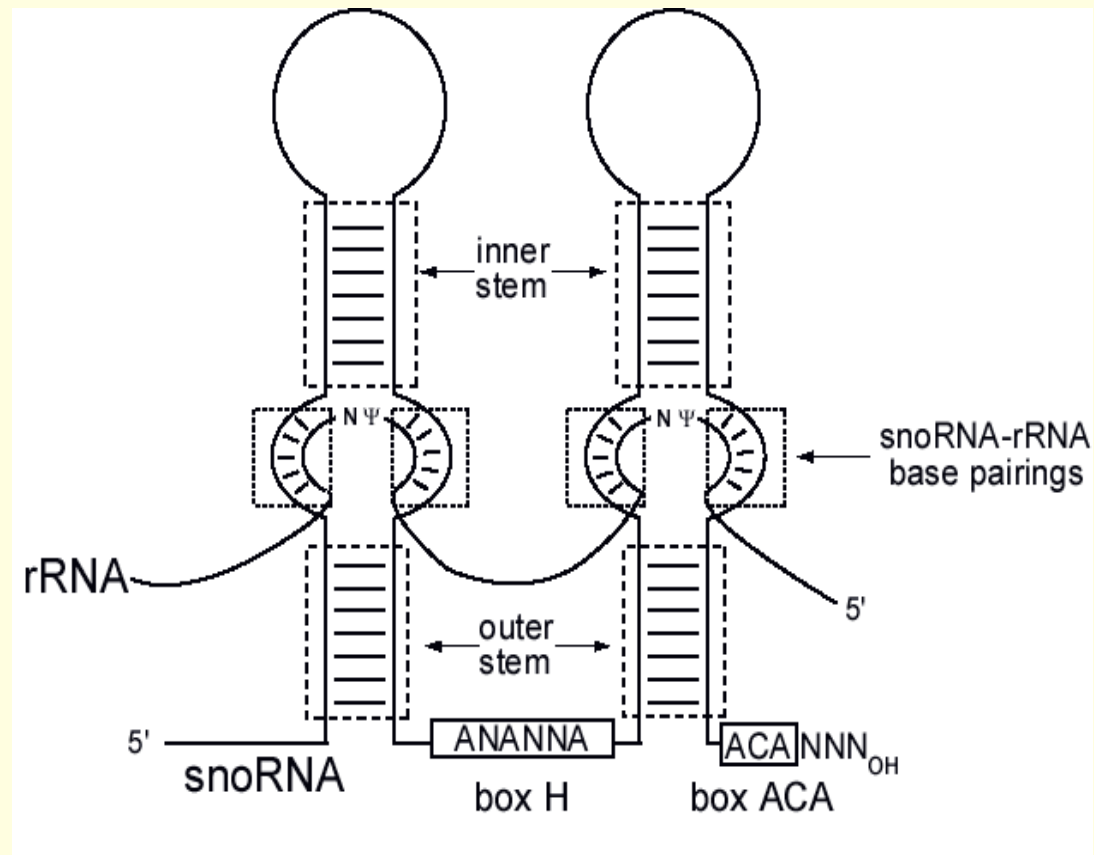
Methylation guide C/D box snoRNA



Very little secondary structure, but short “box” motifs and antisense “guide sequence” conserved

Pseudouridine guide

H/ACA snoRNA



Weak primary sequence motifs, but conserved secondary structure

snoRNA Web Searches

- C/D Box methylation guide snoRNAs
<http://lowelab.ucsc.edu/snoscan/>
- H/ACA pseudouridine guide snoRNAs
<http://lowelab.ucsc.edu/snoGPS/>

More NcRNAs...

microRNAs / siRNAs - short ~22nt RNAs that pair with mRNA to regulate expression

MiRscan: <http://genes.mit.edu/mirscan/>

Bacterial small RNAs: OxyS RNA (antisense temp. sensor), tmRNA in bacteria (incomplete protein translation termination), CsrB (carbon storage regulation)

guide RNAs in trypanosomes – post-transcriptional addition/deletion of nucleotides from mRNAs

RNA Tracks in Genome Browsers

- In UCSC genome browsers, look for:

“RNA Genes”

“sno/miRNA genes”

“transfer RNAs”

“Genbank RNAs”

“RFAM RNAs”

Or anything listed at “ncRNA”...

Some On-line Databases

Ribosomal RNA databases:

Ribosomal Database Project: <http://rdp.cme.msu.edu/html/>

rRNA WWW Server: <http://rrna.uia.ac.be/>

tRNA Databases

Genomic tRNA Database <http://lowelab.ucsc.edu/GtRNAdb/>

Sprinzi tRNA Database

<http://www.uni-bayreuth.de/departments/biochemie/sprinzi/trna/>

snoRNA Databases

Yeast snoRNA database

<http://www.bio.umass.edu/biochem/rna-sequence/>

[Yeast_snoRNA_Database/snoRNA_DataBase.html](http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html)

More Databases

- **SRP RNA Database:**
 - <http://bio.lundberg.gu.se/dbs/SRPDB/SRPDB.html>
- **RNase P Database:**
<http://jwbrown.mbio.ncsu.edu/RNaseP/>
- **tmRNA Database:** <http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html>

Other RNA Lists of Links

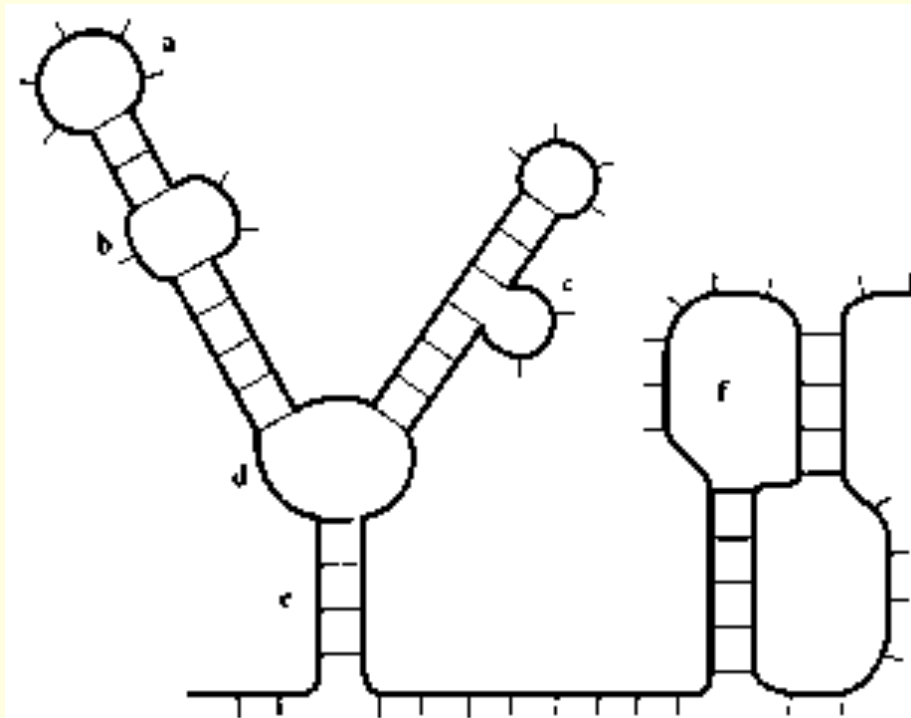
- RNA World @ IMB Jena (software & databases)
<http://www.imb-jena.de/RNA.html>
- NAR Databases Index (annual update)
<http://www.oxfordjournals.org/nar/database/cat/2>
- NAR Web Server List
http://nar.oxfordjournals.org/content/vol35/suppl_2/index.dtl

RNA Folding

- mFOLD - Zuker

<http://www.bioinfo.rpi.edu/applications/mfold/>

Click on “RNA Folding” option



- a. hairpin loop
- b. internal loop
- c. bulge loop
- d. multibranched loop
- e. stem
- f. pseudoknot

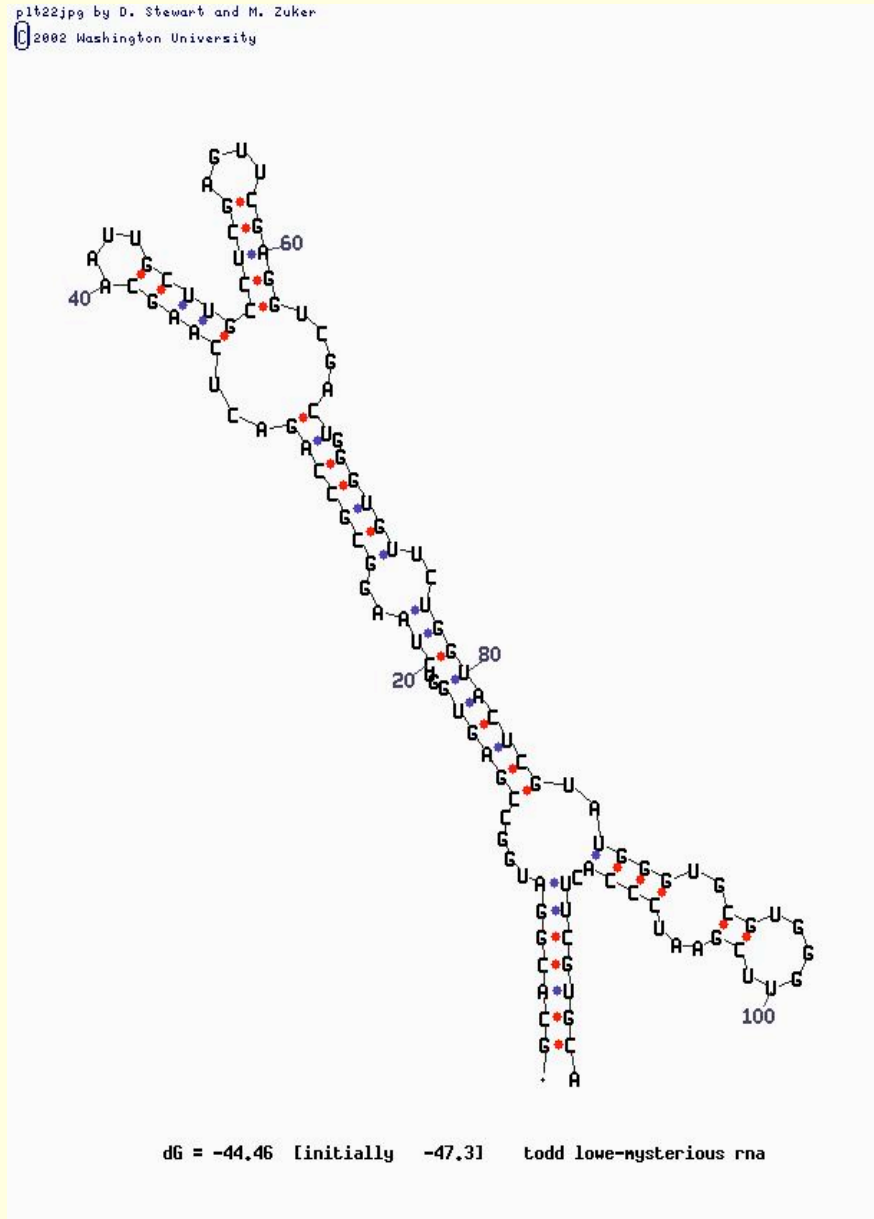
RNA Folding Caveats

- mFold, often will *not* give you the true structure, it is just a reasonable approximation
- Concept: “well-determined” base pairing
- If you have candidate homologs from multiple species, use mFold to see if secondary structure is being preserved (compensatory mutations maintaining structure)

mFOLD Information

- Paste in RNA sequence (can be in DNA letters)
- Use defaults, except, set “structure annotation” to p-num
- Look at energy dot-plot
 - Black dots are in optimal structure
 - Colored dots in sub-optimal structures
- Look at top structures (within 5% of optimal)
 - Are there many?
 - Which features are consistent between structures?
 - These are the most dependable aspects of structure
- Use “compare selected foldings” to see differences between different folds

mFold of a Leu-CAA tRNA



Finding RNA Genes

- Non-coding RNAs (ncRNAs) are not detected effectively by “general” gene finders (unlike proteins)
- BLAST and other similarity-based search methods often miss ncRNAs – secondary structure conserved, not primary; incorrect boundaries
- Therefore, we need specialized gene finders for accurate detection for each RNA gene family

Mysterious new conserved element (PAE0935) – What next?

- Try the obvious first: BLAST at NCBI
<http://www.ncbi.nlm.nih.gov/BLAST/>
- Hmm, you get a strong match against other genomes
- Other close hits to related species, but no good annotation

BlastN v. BlastX

- Perhaps your sequence is a protein, studied in another species, but BlastN is not sensitive enough
- Try BlastX
 - protein comparison is better at picking up more distantly related, conserved protein coding genes

Is it a protein?

- No *convincing* BlastX hits either! You are beginning to suspect this might not be a protein
- So, translate your protein, look for long open reading frames
- One reasonably long open reading frame, but still no evidence it is a real protein
- Perhaps a non-protein coding RNA (ncRNA)?

Rule out the Easiest First

- Let's assume we *know* it's an RNA gene now
- Which one?!

- Start with RFAM, which has the largest, most diverse collection of RNA gene models

- Will not detect all types of ncRNA (i.e. snoRNAs), or novel types of ncRNAs

Next: Specialized ncRNA Gene Finders

Two specialized gene finders developed specifically for particular gene families:

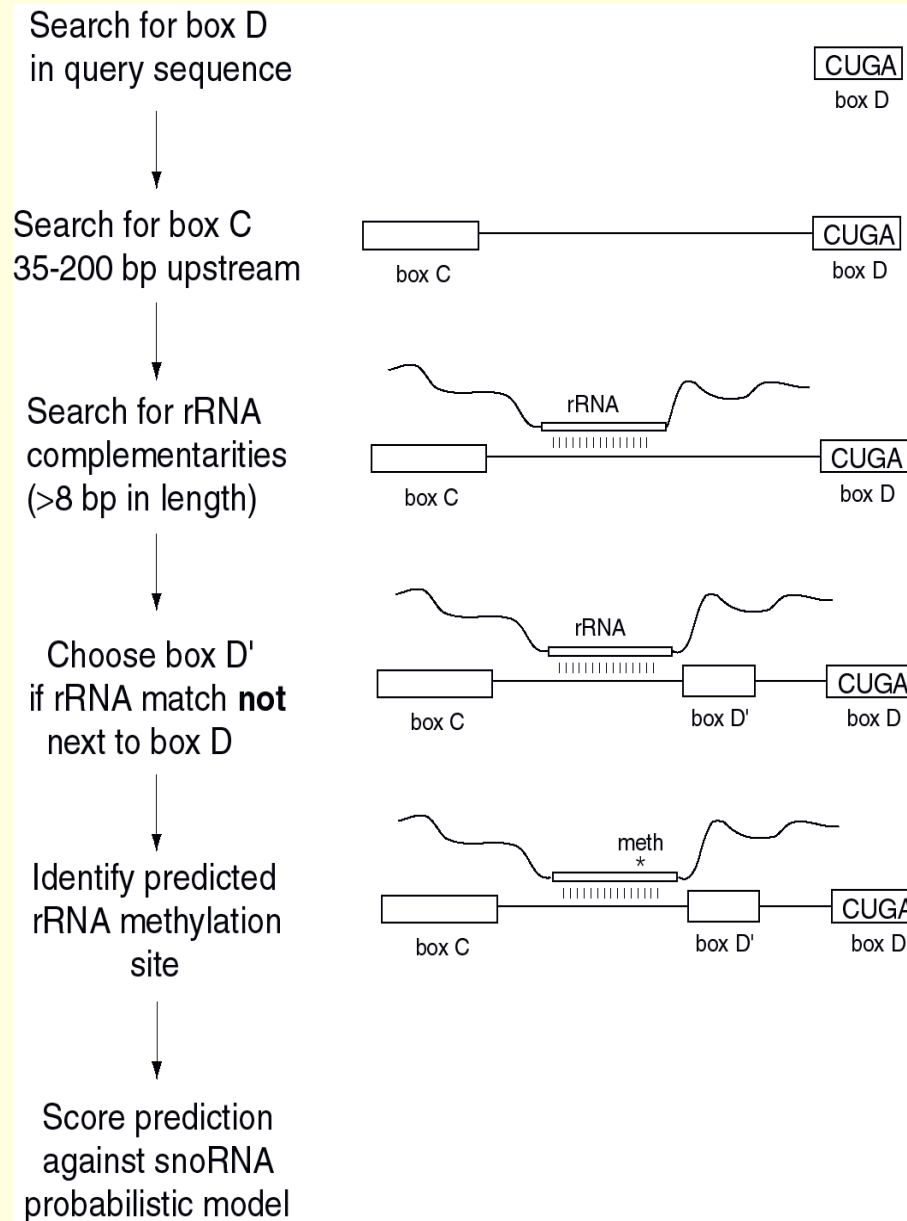
- tRNAscan-SE: searches for transfer RNAs using probabilistic models to search genome sequence
- Fast, accurate

snoRNA searches

- Snoscan – search for C/D box methylation guide snoRNAs
- snoGPS – search for H/ACA guide snoRNAs

Example of a
Customized RNA
Search program:

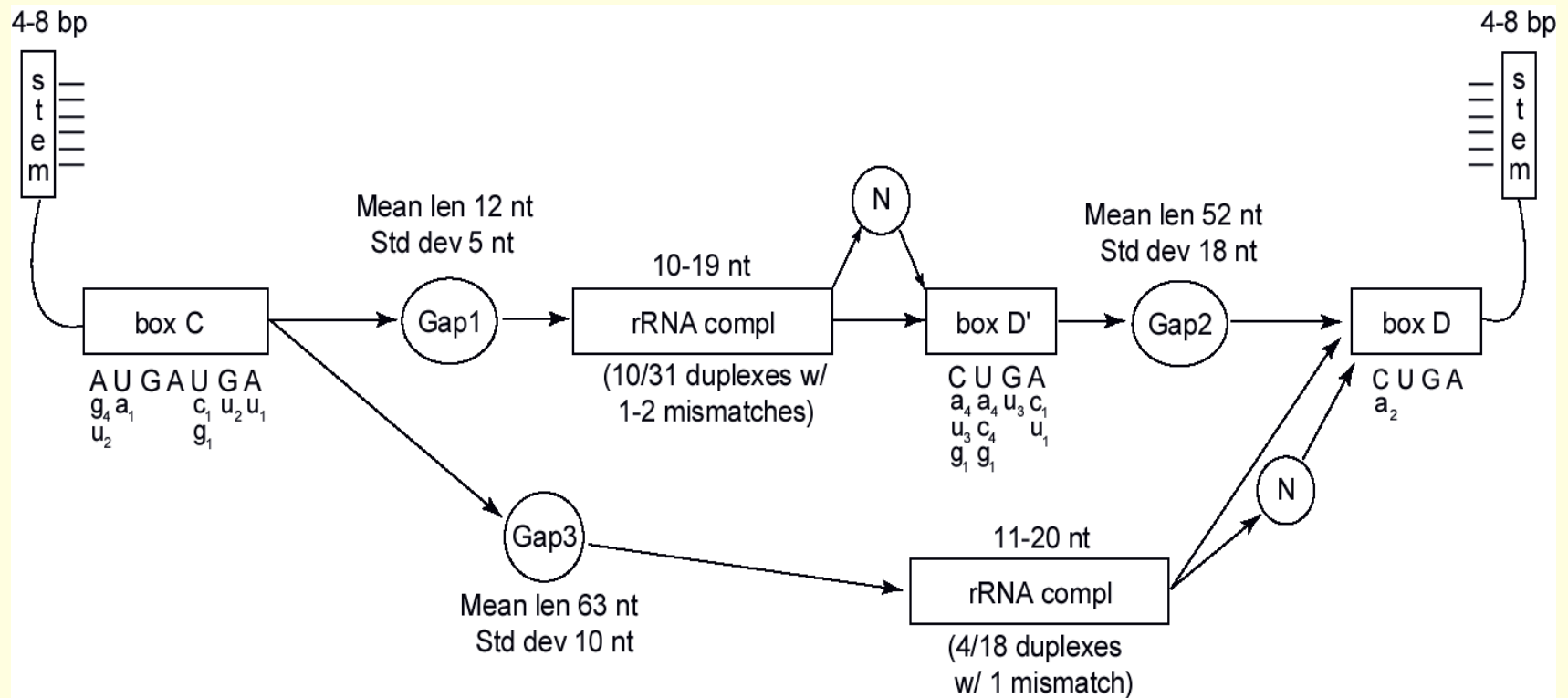
“Snoscan”



Cloned sRNAs from *S. acidocaldarius*

sRNA		C box		Dp box		Cp box		D box	
sR1	F	GAG UUGAUGA	--GAAGUUAAAAA	GCGA	-----	UGGAUGA	-----GCUUAACUCCC	AUGGU	CUGA UAAC
sR2	F	GA GUGAUGA	--GACGAGCGCUAA	CAGA	-GAGA	GUGAAGA	-----GGUCACU	CGAA	CUGA AGAA
sR3	F	AGG AUGACGA	--GACCCAAAAUA	UUGA	-----	ACGAUGA	-----UAUAACCUGU	UCGG	CUGA UCAGU
sR4	N	G UUGAUGA	--GCACAUU	CUGA	-UUUA	AUGAAGA	-----AAGUGGC	CAGGU	CUGA GGUAG
sR5	FN	GAA AUGAUGA	-AUGGUCGACGGAA	CGGA	--CCU	AUGAAGA	-----AUUGUUG	CGGA	CUGA CAAAC
sR6	F	GG AUGAUGA	----C AAAUAGA	CUGA	--AAG	AUGAAGA	-----AAUGCAC	UCAAA	CUGA CUAAA
sR7	F	G AUGAUGA	--CAAAGAG	UGGA	-----	UUAGUGA	CAUCUAAUUUUGUGGGC	AGCCA	CUGA UAGAG
sR8	N	G AUGAUGA	-AGCCCGCCAUCAA	CAGA	--UAA	GUGAAGA	-----GGGAACC	CGAGG	CUGA GAAU
sR9	F	AAAUA AUGAUGA	--CUAACUC	CUGA	--CCA	AUGAUGU	-----CGUAACC	CGAAA	CUGA AUAAA
sR10	F	GA AUGAUGU	--GGAAUCC	CUGA	--GA	AUGAUGA	-----CAAAAAGCGC	GAGCG	CUGA UUAUA
sR11	F	GAAU GUGAUGA	-UGGGUCGA	CUGA	-UUAG	UUGAUGA	-----GAUUAUC	UCCGG	CUGA GAAU
sR12	F	GA AUGAAGA	--ACCCAAC	CUGA	-GGUU	AUGAUGA	-----CAGGUUG	UUCGU	CAGA UCGAUGU
sR13	N	AGG AUGAUGU	-ACUUUCAC	CUGA	--AAG	GUGAGGA	-----UGAGUCC	GACUA	CUGA CGCAA
sR14	FN	GCU GUGAAGA	-CGCUAGAC	CUGA	--CUC	AUGAUGA	-----AGGGCCAAAGCU	CAGA	GCAAAC
sR15	F	A GUGAUGA	GGAACCAACGAGAG	CUAG	----U	UUGAUGG	-----CUUCGACGCUCUGCU	CUGA	AA
sR16	N	GA AUGAAGA	--CGUUCCACCCGA	GCGA	-----	GUGAUGA	-----GCGAAACGGUUAUA	CUGA	UGAUG
sR17	F	AGAA AUGAAGA	--CUAAAAAACCGG	CUGA	GAUAA	GUGAUGA	-----CGACGUCUCGCA	CUGA	UC
sR18	N	AA GUGAUGA	--CAGAACC	CUGA	--AAG	AUGAUAG	-----AGCCGUGUGAGAA	CUGA	UCAAU
Sso sR1		ACAG AUGAUGA	--AUUCCCG	ACGA	-----	UUGAUGA	-----GCUUAACUCCC	AUGGA	CUGA UUAG
Consensus	1-9 nts	AUGAUGA	-----9-14 nt guide-----	CUGA	-----0-5 nts-----	AUGAUGA	-----12-22 nt guide-----	CUGA	-----2-10 nts-----

snoRNA Probabilistic Model



Check Current ncRNA Databases

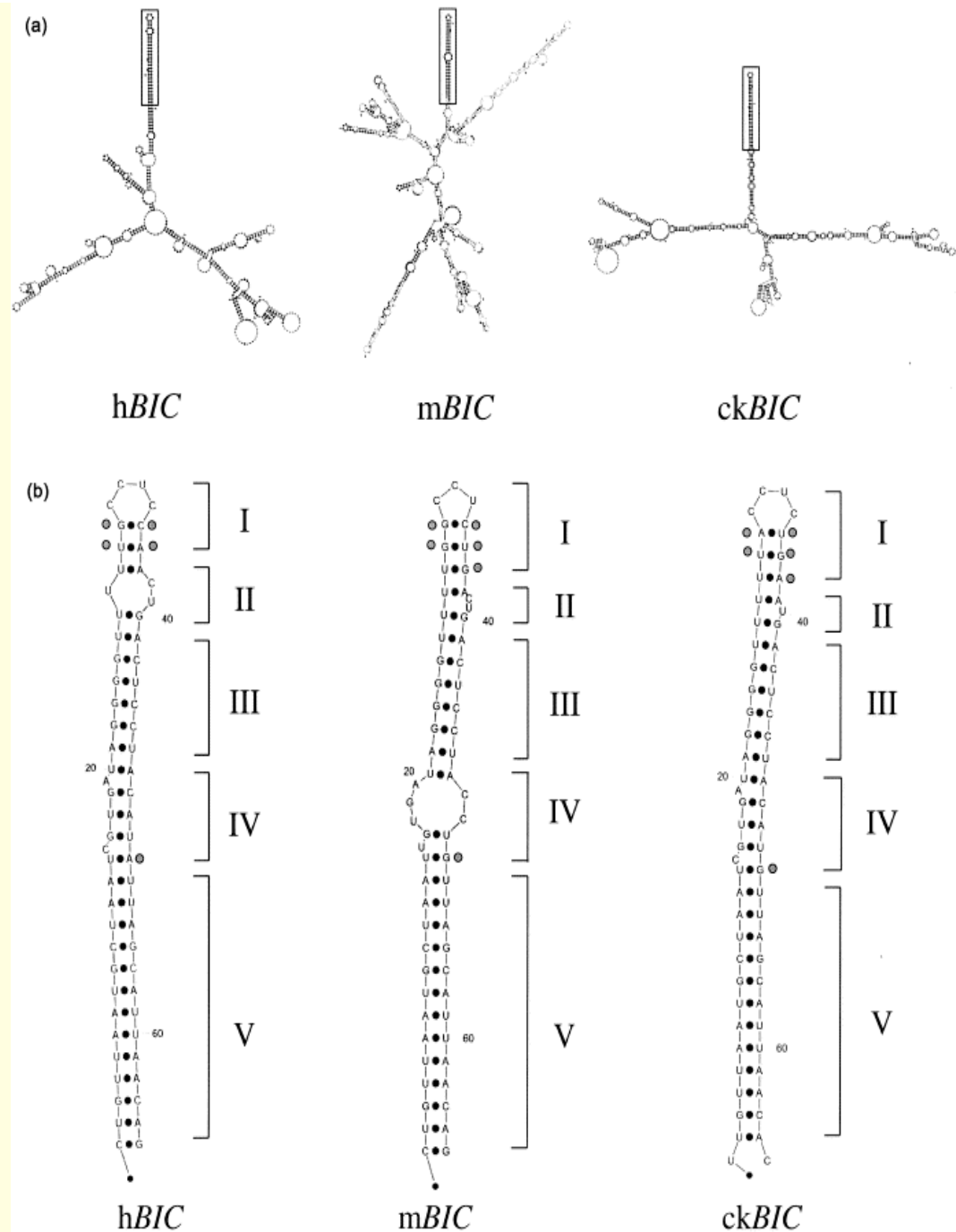
- **Some have search options**
- **RNase P database:**
<http://jwbrown.mbio.ncsu.edu/RNaseP/>
- **tmRNA Database:** <http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html>

New databases and resources always coming on-line – check links at IMB Jenna (given earlier) and annual database issue of *Nucleic Acids Research*

Check 2-D Structure for Clues

- Michael Zuker's mFold server
- Is there one good optimal structure, or many within 5% of "optimum"?
- If you have candidate homologs from multiple species, use mFold to see if secondary structure is being preserved (compensatory mutations maintaining structure)

Comparative analysis of 2D mFold-predicted structures (Tam, *Gene* 274:157-67, 2001)



A Practical Guide: So you think you've found a novel RNA?

1. Try BLAST first to look for very similar hits (any long ORFs?)
2. Try battery of existing ncRNA search tools to verify RNA is in a novel class
3. Attempt to determine secondary structure with *mfold*; are any portions particularly “well-determined”?
4. Collect candidate orthologs from closely related species
5. Create a “training set” of sequences to model (verified/studied experimentally), align structurally if possible using known biological features
6. Model primary/secondary structure with an SCFG, search other genomes for hits
7. Collaborate with an experimental lab to determine null function / cell localization / interactions with other proteins

How Do People Find *New* ncRNAs?

Traditional biochemistry (immunoprecipitation to interacting proteins)

Genetic screens are often difficult to find ncRNAs

Direct RNA sequencing of libraries

Mining expression databases (EST sequencing, microarray data)

Comparative genomics combined with other information (promoters, terminators, common secondary structure)

Programs: RNAz, Qrna