# Computational analysis of non-coding RNA

Andrew Uzilov

**auzilov@ucsc.edu**

BME110

Tue, Nov 16, 2010

# Corrected/updated talk slides are here:

**http://tinyurl.com/UzilovRna**

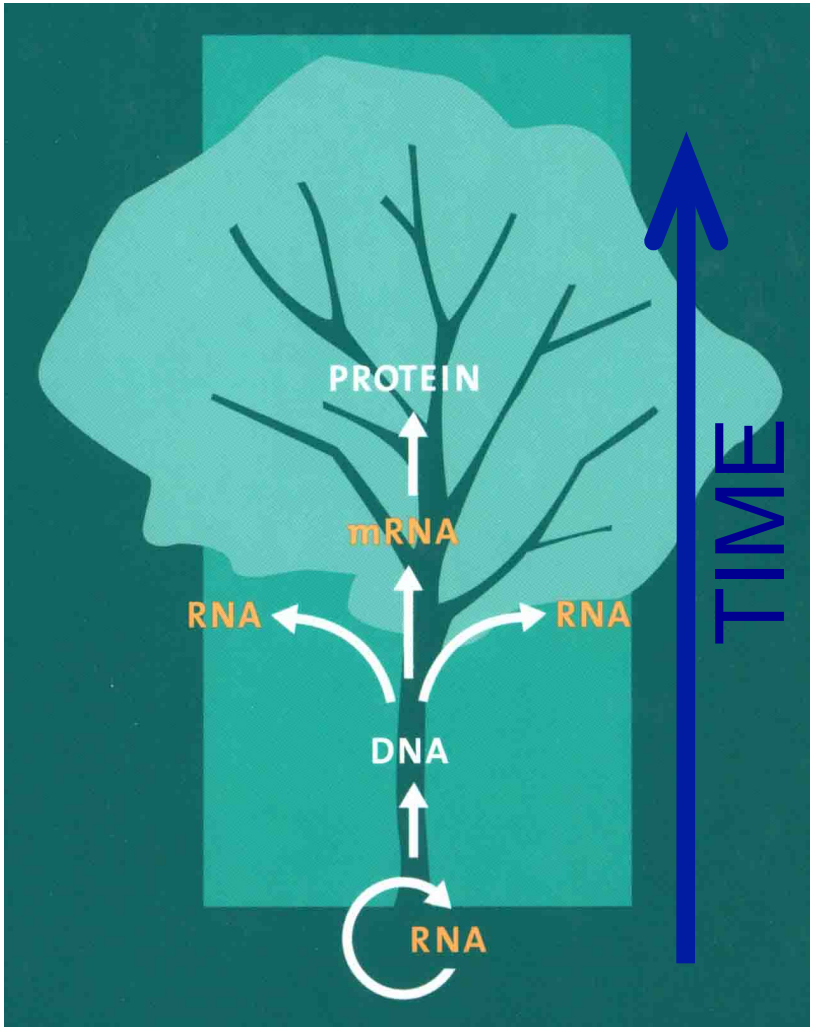
redirects to:

**http://users.soe.ucsc.edu/~auzilov/BME110/Fall2010/**

# Talk progress

- **Overview**
- RNA structure
  - fundamentals
  - computational RNA folding (mfold)
- Gene-specific computational models
  - Rfam database
  - tRNA (tRNAscan-SE)
  - C/D box snoRNA (snoscan)
- Practical guide

# Why RNA?
# An evolutionary perspective




The "RNA World" hypotheses: life arose as self-replicating **non-coding RNA (ncRNA)**

Figure is from cover of "The RNA World," 3rd ed., Gesteland *et al*

# Why RNA?
# An evolutionary perspective

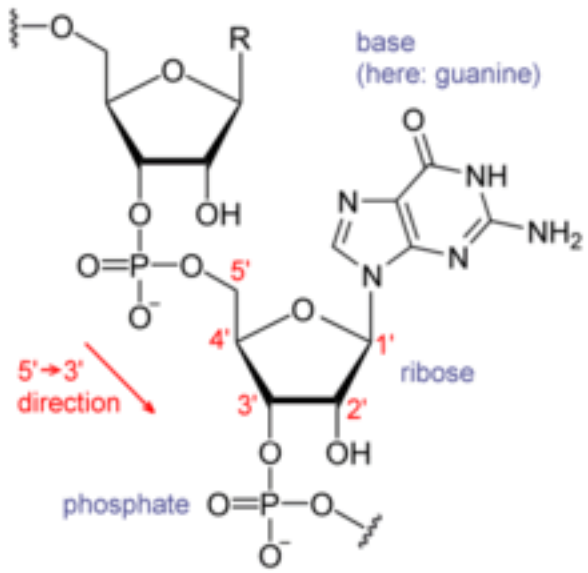
- Evidence for an RNA World: many core biological mechanisms are RNA-based
  - translation (rRNA, tRNA)
  - splicing (snRNA)
  - RNA processing (RNase P)
  - protein trafficking (SRP RNA)
  - and many, many more!
- RNA is very versatile
  - a passive carrier of information (mRNA)
  - an active component in biological processes (ncRNA)
  - sometimes both!

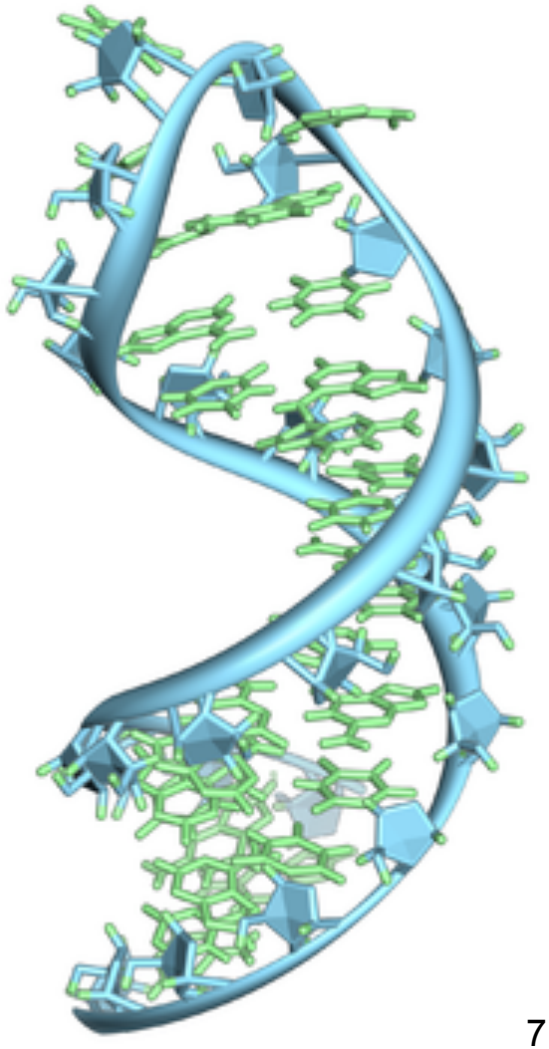
# ncRNA can have many roles

- Structural – fold into complex 3-D structures, usually scaffolds for proteins
  - rRNA, snRNA, tRNA, etc.
- Antisense ("guides") – form specific base pairings to "target" RNAs
  - snRNA, snoRNA, microRNA, siRNA, piwiRNA, etc.
  - tRNA??
- Catalytic ("ribozymes") – catalyze biochemical reactions
  - rRNA, RNase P, group I & II introns, hammerhead, HDV, etc.
  - snRNA??
- Regulatory – regulate gene expression
  - microRNAs, siRNA, piwiRNA, many bacterial small RNAs

# RNA chemistry overview

- Base, sugar, phosphate backbone
- 5’ and 3’ ends
- Less stable than DNA or protein

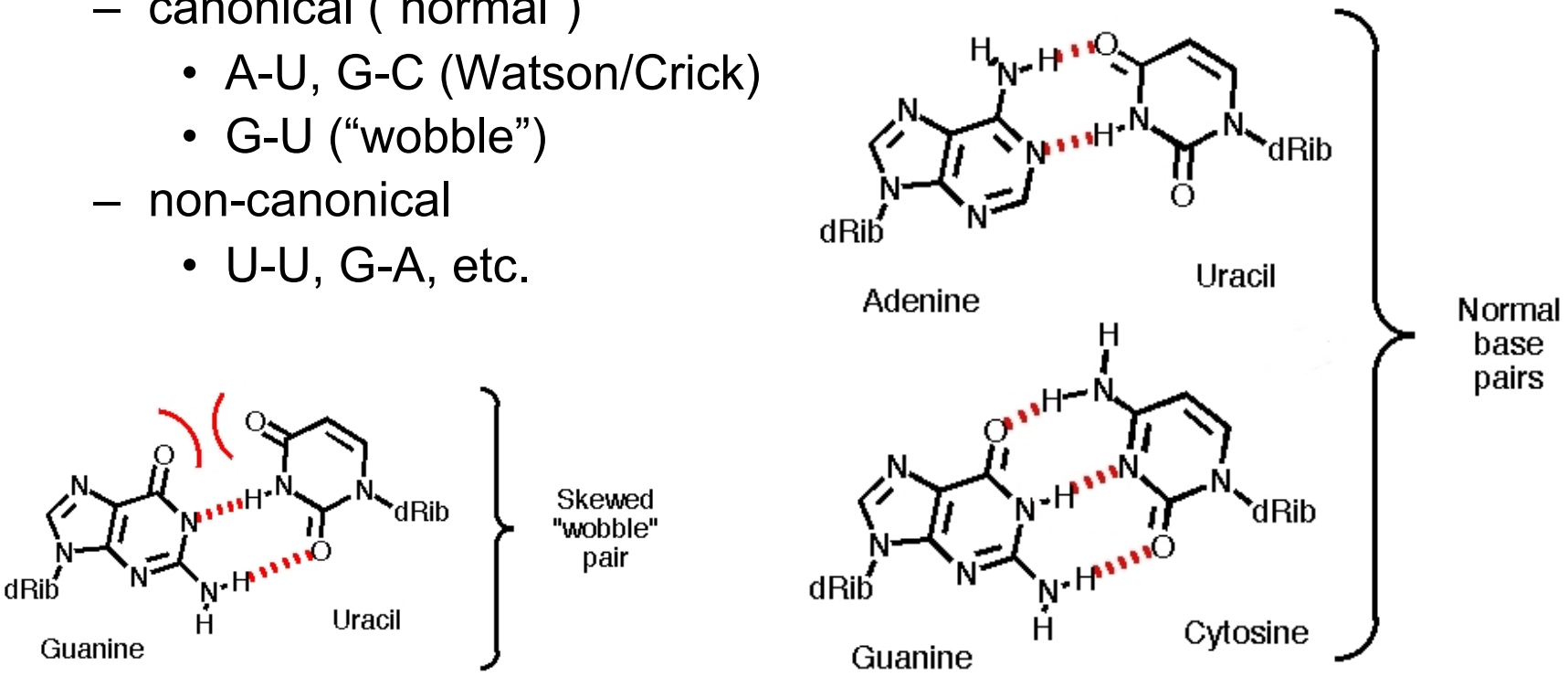


(Images from Wikipedia)

# RNA fundamentals

- Sequence alphabet: A, C, G, U (T~U)
- Come in all sizes: 10s of bases to 1000s of bases
- RNA can form base pairs
  - canonical ("normal")
    - A-U, G-C (Watson/Crick)
    - G-U ("wobble")
  - non-canonical
    - U-U, G-A, etc.






(image credit: http://www.rosalindfranklin.edu/DNN/home/CMS/biochem/Faculty/Walters/ComputerApplications/RNA/tabid/1202/Default.aspx)

# RNA can be modified or edited

- RNAs may undergo chemical modifications or base editing which may affect their
  - base pairing
  - 3-D structure
  - stability/turnover
  - function

- Some common modifications
  - methylation of 2’ oxygen of ribose (2’-*O*-methylation)
  - uridine -> pseudouridine (Ψ)
  - adenoside-to-inosine (A-to-I) editing

# What patterns can we use in RNA bioinformatics?

- Base pairing
  - within the same RNA (intramolecular aka *cis*)
  - sense/anti-sense pairing between two RNAs (intermolecular aka *trans*)
  - conserved by evolution (sequence may diverge)
- Base frequency bias
  - Example: G-C base pairs are most stable, so ncRNA or other “structured” RNA

# RNA + Proteins = Complexes

- ncRNAs usually do no act alone, but are complexed with proteins
- Sites of interaction with other 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

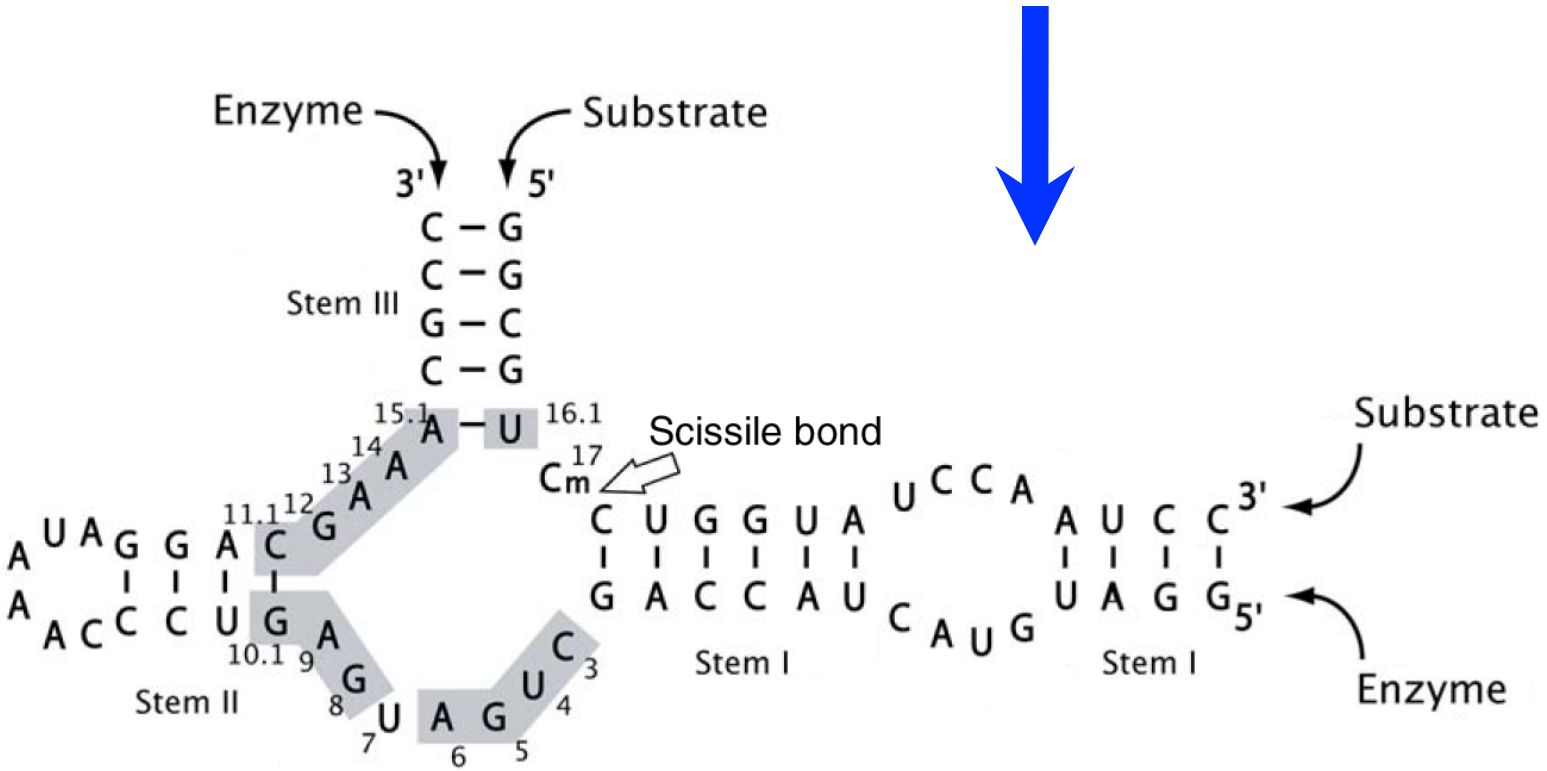
# Talk progress

- Overview
- **RNA structure**
  - **fundamentals**
  - computational RNA folding (mfold)
- Gene-specific computational models
  - Rfam database
  - tRNA (tRNAscan-SE)
  - C/D box snoRNA (snoscan)
- Practical guide

# Levels of RNA structure

## primary “structure”

```
Enzyme: GGAUGUACUACCAGCUGAUGAGUCCCAAAUAGGACGAAACGCC
Substrate: GGCGUCCUGGUAUCCAAUCC
```

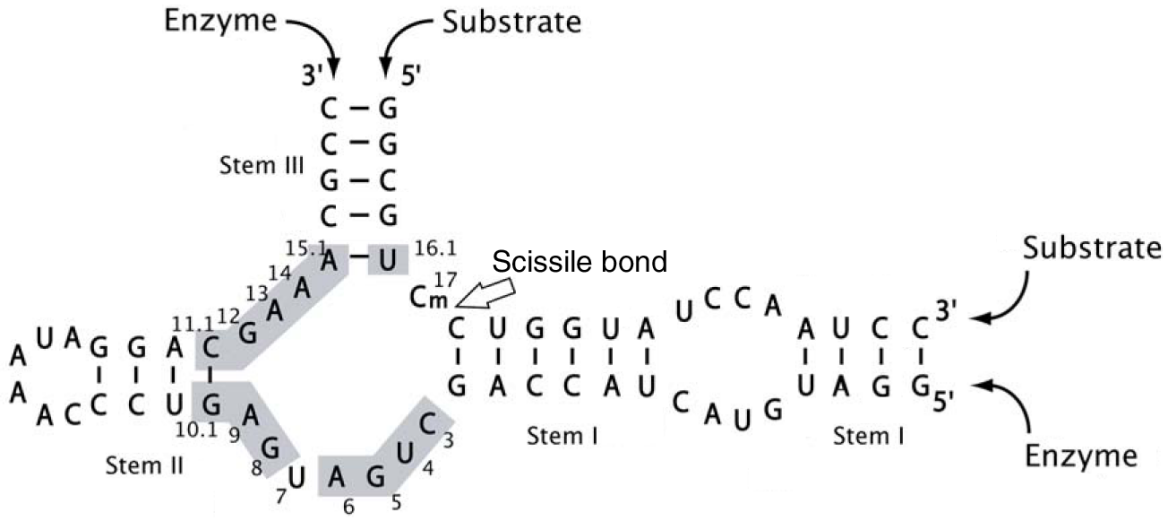



## secondary structure

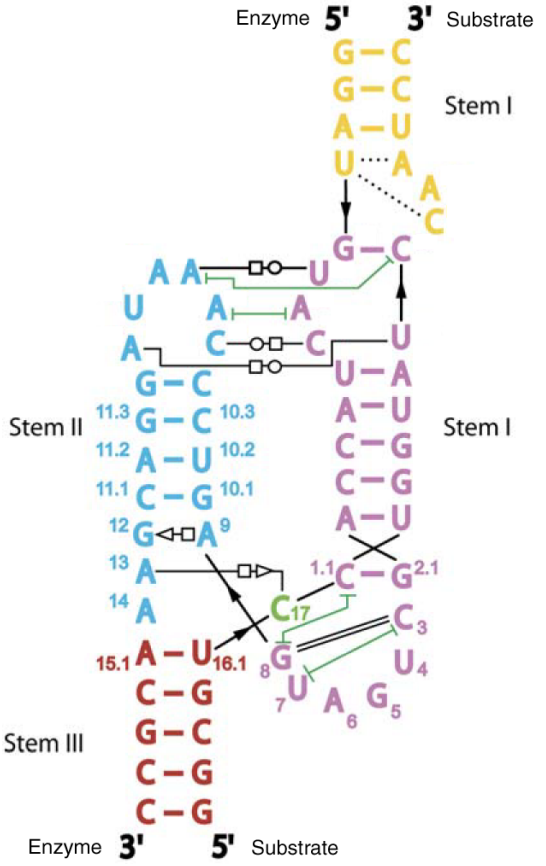


Figure modified from Martick and Scott (2006).

# Levels of RNA structure




secondary structure

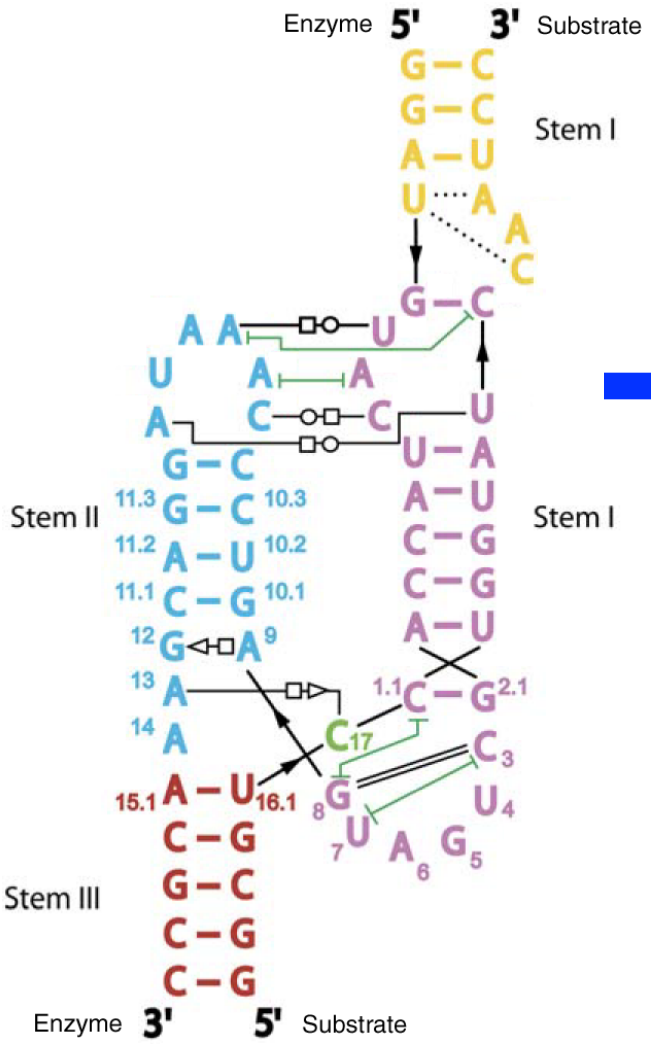



tertiary contacts



Figure modified from Martick and Scott (2006).

# Levels of RNA structure




tertiary contacts

3-D structure

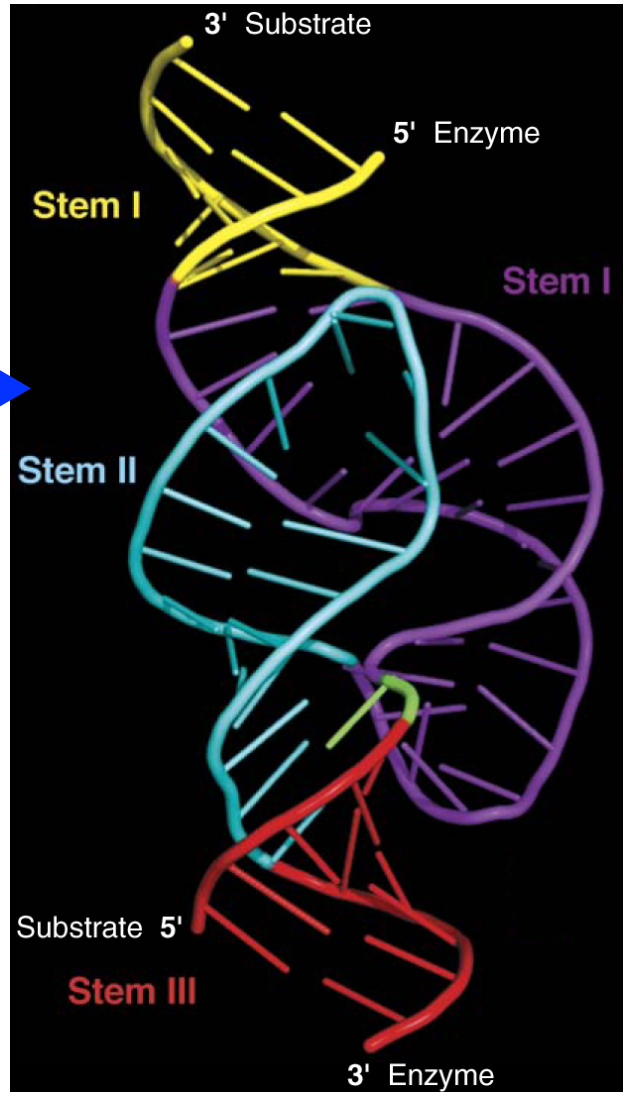
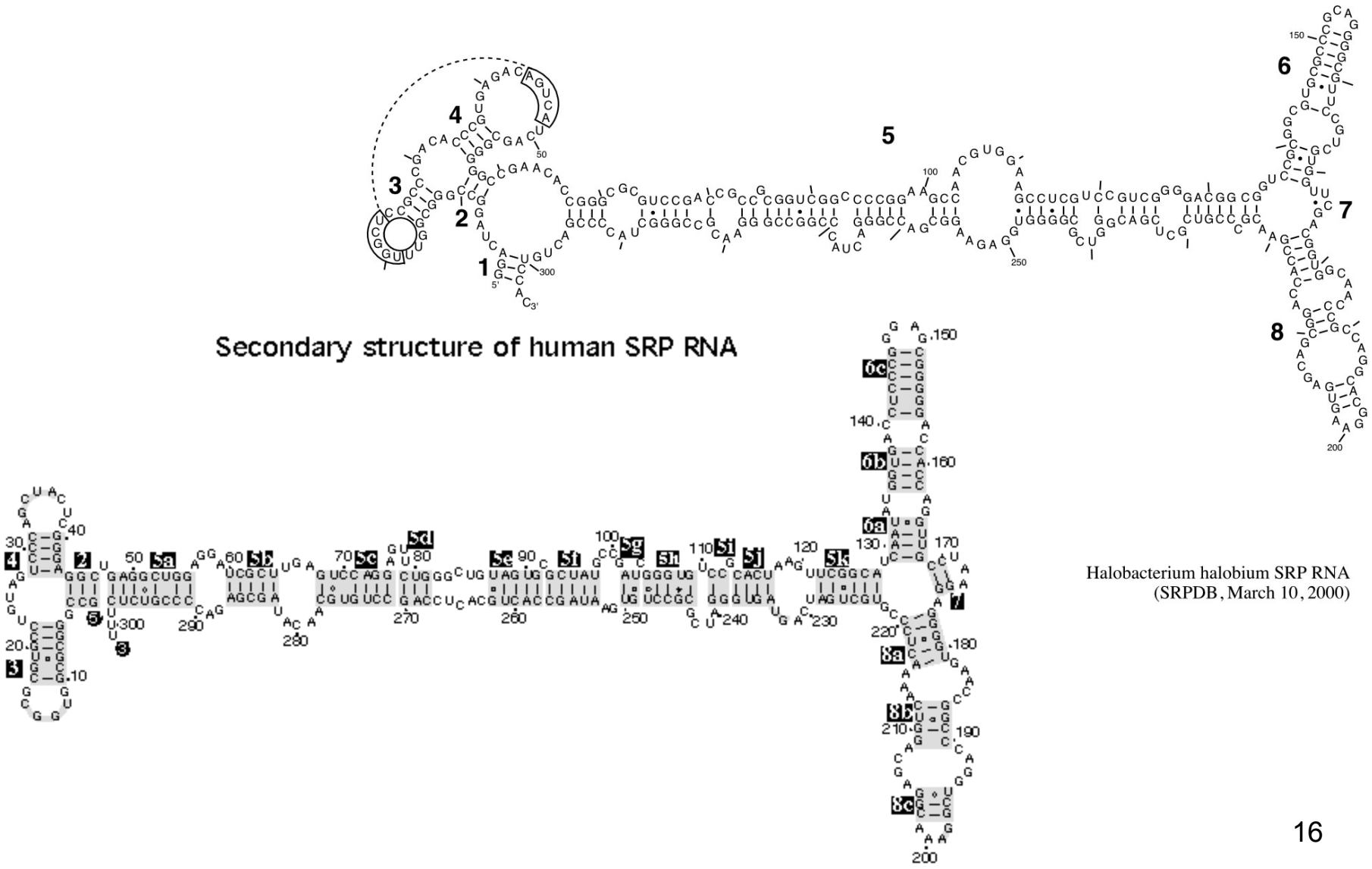


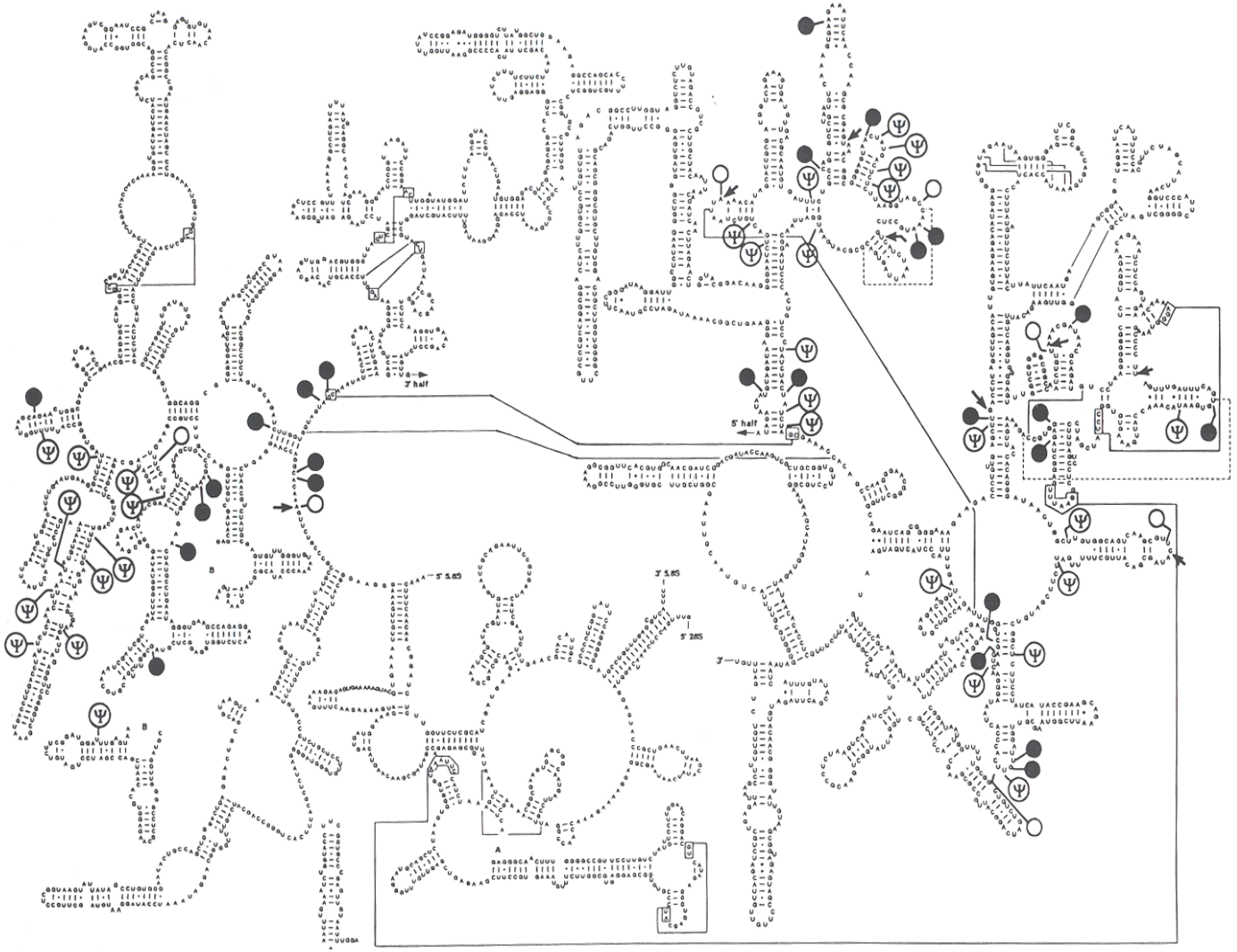



Figure modified from Martick and Scott (2006).

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



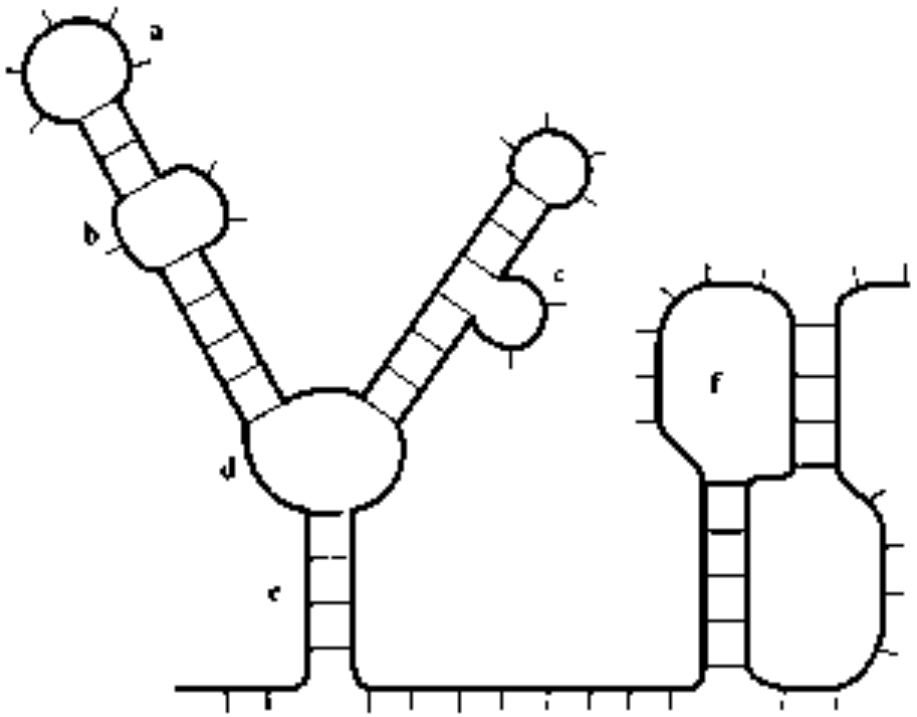
# Baker's Yeast (*S. cerevisiae*) Large Subunit rRNA secondary structure



# Key points

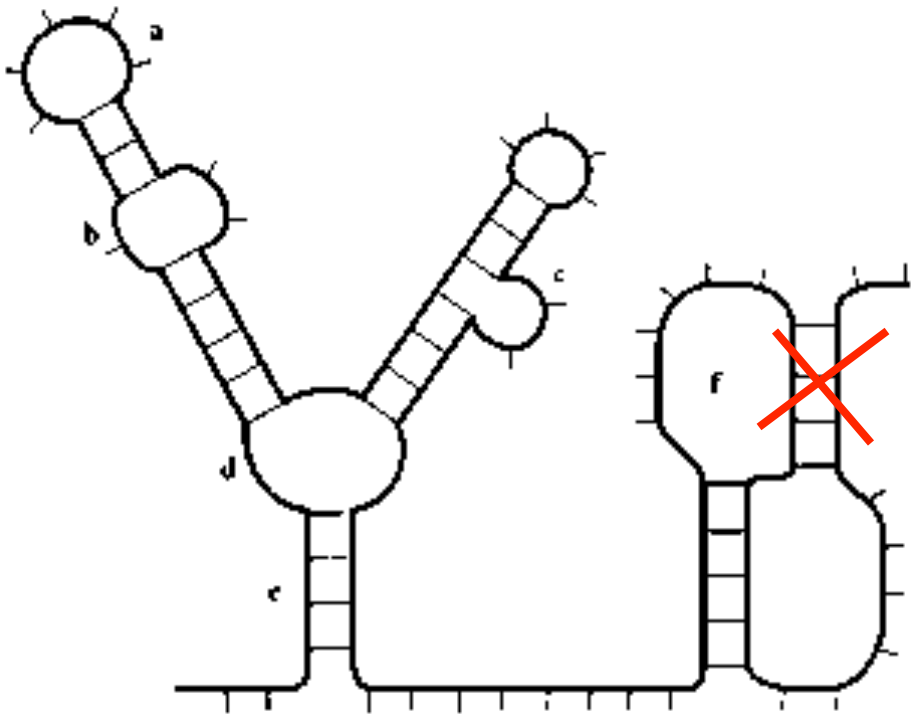
- Analysis of RNA secondary structure (the set of base pairs) alone can tell you a lot!
- Most computational RNA tools model secondary structure and primary sequence
- Many computational tools exist to predict the secondary structure (to fold RNA) in a general way

# Features of RNA secondary structure




a. hairpin loop / stem loop
b. internal loop
c. bulge loop
d. multibranch loop
e. stem / helix
f. pseudoknot

# Features of RNA secondary structure



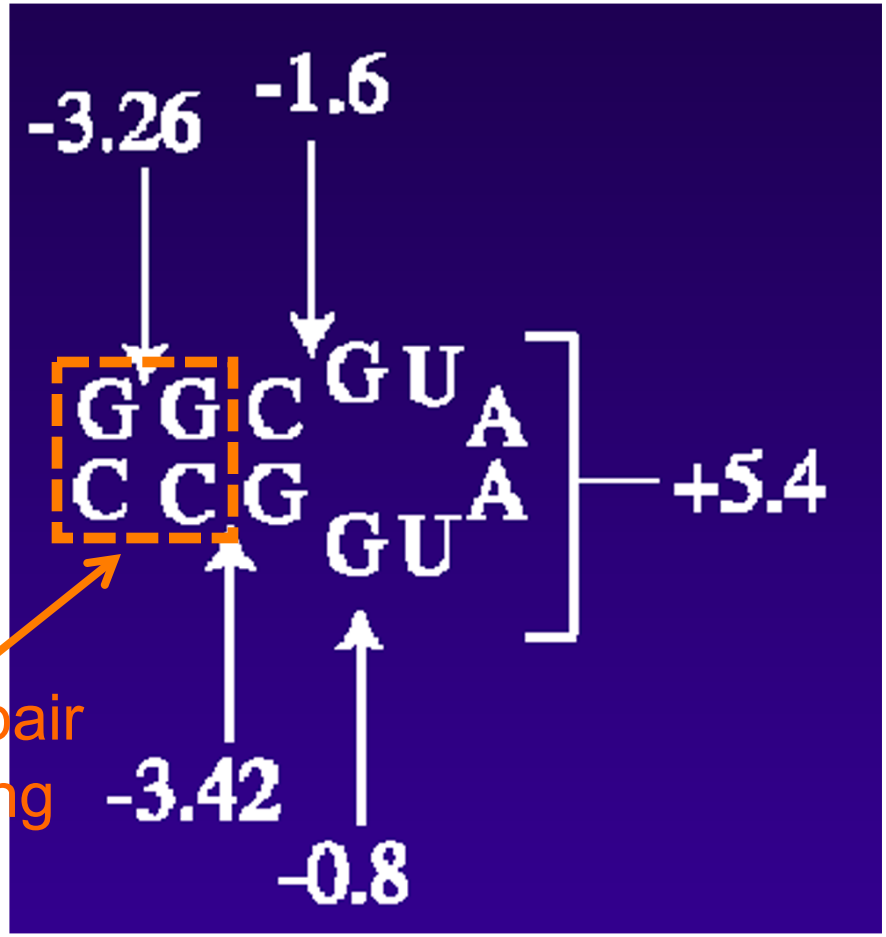

a. hairpin loop / stem loop
b. internal loop
c. bulge loop
d. multibranch loop
e. stem / helix
f. pseudoknot

Not modeled by most folding algorithms!

# Talk progress

- Overview
- **RNA structure**
  - fundamentals
  - **computational RNA folding (mfold)**
- Gene-specific computational models
  - Rfam database
  - tRNA (tRNAscan-SE)
  - C/D box snoRNA (snoscan)
- Practical guide

# Many algorithms find optimal fold by free energy minimization




Base pair stacking is stabilizing, confers an "energy bonus"

$$\Delta G^\circ = -RT(\ln K_{eq})$$

**ΔG =**

-3.26 + -1.6 + 5.4 + -0.8 + -3.42

**= -3.7 kcal/mol**

# Computational RNA Folding

- Most popular is mfold by Michael Zuker
  - http://mfold.rna.albany.edu
  - under software, click “mfold,” then click “RNA Folding Form”
- Computes several structures
  - Optimal (MFE) structure
  - Suboptimal structures

# RNA folding caveats

- mfold often will *not* give you the true structure, it is just a reasonable approximation

- If you have candidate homologs from multiple species, use mfold to see if secondary structure is being preserved (compensatory mutations maintaining structure)
  - However, for two species, better tools exist to do this (covered in next lecture, Tue Nov 23)

# How to use mfold

- 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 10% 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

# LIVE mfold DEMO

**U1a snRNA (164 nt):**

AUACUUACCUGGCAGGGGAGAUACCAUGAUCACGAAGGUGGUUUUCCC
AGGGCGAGGCUUAUCCAUUGCACUCCGGAUGUGCUGACCCCUGCGAUU
UCCCCAAAUGCGGGAAACUCGACUGCAUAAUUUGUGGUAGUGGGGGAC
UGCGUUCGCGCUCUCCCCUG

# Talk progress

- Overview
- RNA structure
  - fundamentals
  - computational RNA folding (mfold)
- **Gene-specific computational models**
  - **Rfam database**
  - tRNA (tRNAscan-SE)
  - C/D box snoRNA (snoscan)
- Practical guide

# The Rfam database

- A comprehensive database of ncRNA structures, alignments, and annotations

  http://rfam.sanger.ac.uk/

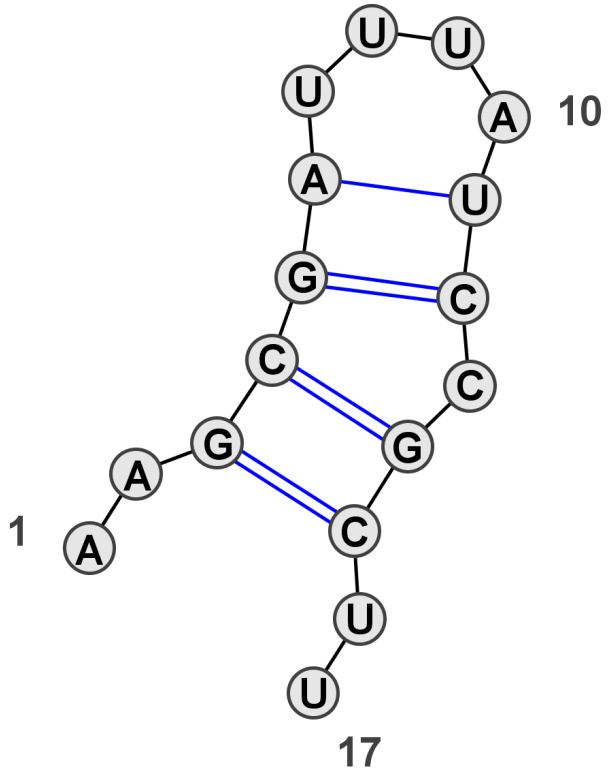
  (U.S. mirror exists, but I recommend the U.K. site)

- Every ncRNA has a covariance model describing its secondary structure

- Just like in Pfam (protein sister site), you can paste in your sequence and look for matches to RNA gene models

# Dot/bracket secondary structure notation

Base pairs are matching parentheses




```
AAGCGAUUUAUCCGCUU
..(((((....)))).)))..
```

Alternative base pair symbols:

```
<>   {}
```

Alternative unpaired base symbols:

```
_   -   :   ,
```

# Rfam caveats

- Structure models only contain conserved base pairs, which is a minimal set (your favorite species might have additional base pairs)

- Many structure models are based on computational predictions!
  - "published" versus "predicted"

# LIVE Rfam DEMO

**U1a snRNA (164 nt):**

AUACUUACCUGGCAGGGGAGAUACCAUGAUCACGAAGGUGGUUUUCCC
AGGGCGAGGCUUAUCCAUUGCACUCCGGAUGUGCUGACCCCUGCGAUU
UCCCCAAAUGCGGGAAACUCGACUGCAUAAUUUGUGGUAGUGGGGGAC
UGCGUUCGCGCUCUCCCCUG

# Talk progress

- Overview
- RNA structure
  - fundamentals
  - computational RNA folding (mfold)
- **Gene-specific computational models**
  - Rfam database
  - **tRNA (tRNAscan-SE)**
  - C/D box snoRNA (snoscan)
- Practical guide

# tRNAs: a well-studied ncRNA gene family

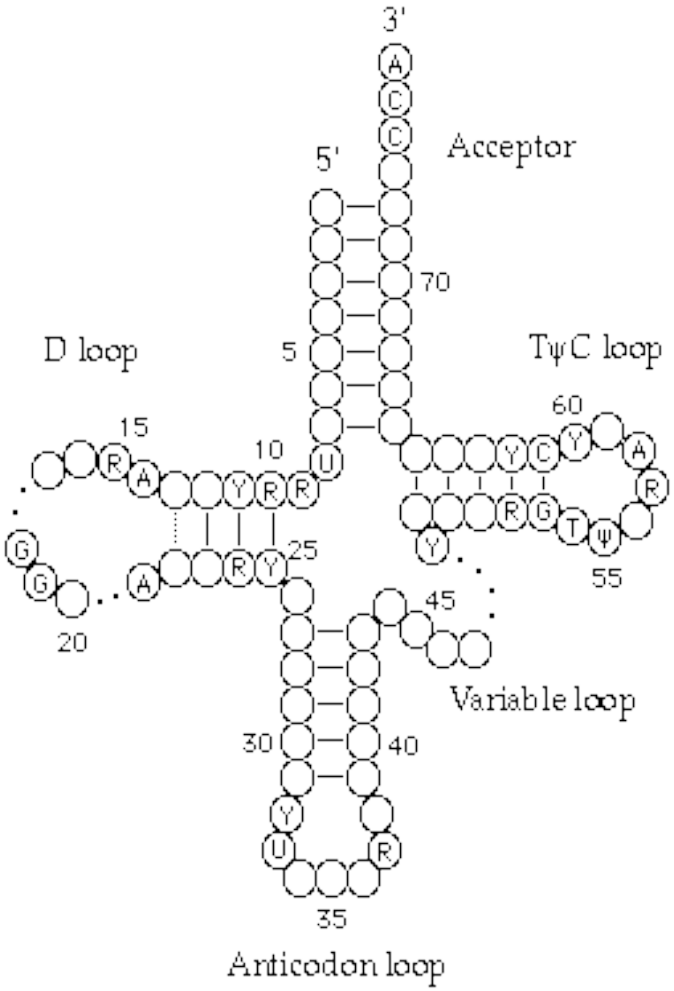
Transfer RNAs (tRNAs) – “decode” mRNA codons into cognate amino acids in protein translation

- Examples: tRNA-AGC(Ala), tRNA-ACC(Gly)…

  (62 kinds possible)

- Structured RNA and “antisense” interaction between tRNA anticodon and mRNA codon

# tRNA "cloverleaf" secondary structure




# 3-D structure

# Useful tRNA databases

Genomic tRNA Database

http://lowelab.ucsc.edu/GtRNAdb/

Sprinzl tRNA Database

http://www.uni-bayreuth.de/departments/biochemie/sprinzl/trna/

Using Rfam for tRNA is **not advised**!

# tRNAscan-SE

http://lowelab.ucsc.edu/tRNAscan-SE/

- Finds tRNAs in genome sequences using probabilistic models
- A gene-specific covariance model at the core
- Fast, accurate

# LIVE tRNAscan-SE DEMO

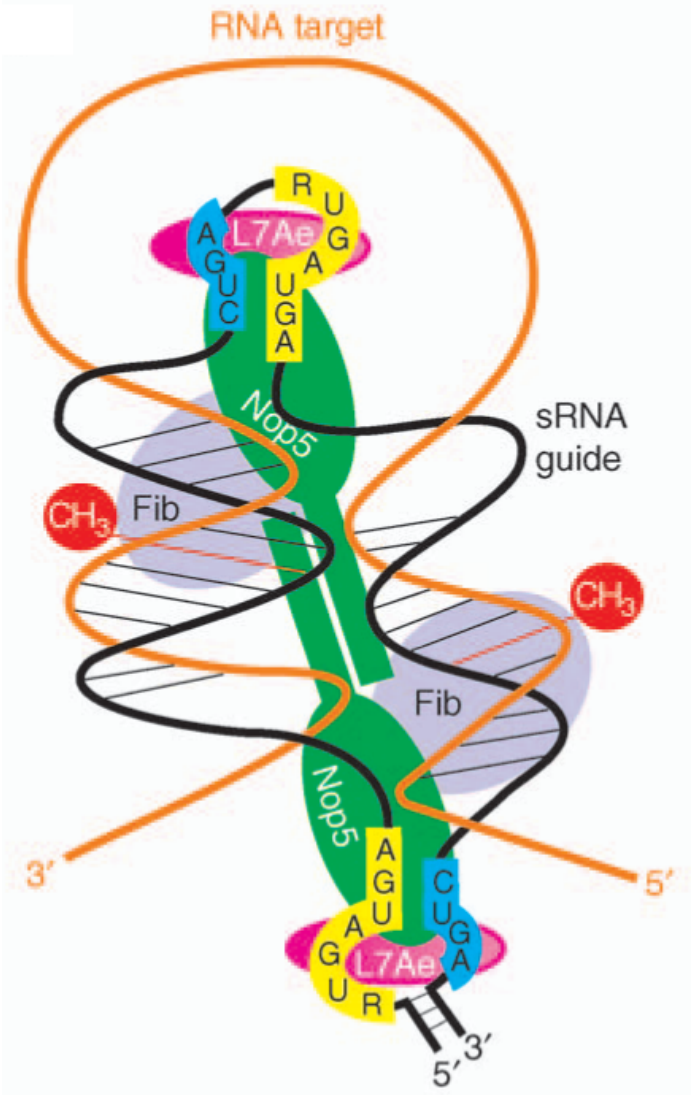
**A mystery tRNA:**

GGGGGUGUAGCUCAGUGGUAGAGCGCGUGCUUCGCAUGUACGAGGCCC
CGGGUUCAAUCCCCGGCACCUCCA

# Talk progress

- Overview
- RNA structure
  - fundamentals
  - computational RNA folding (mfold)
- **Gene-specific computational models**
  - Rfam database
  - tRNA (tRNAscan-SE)
  - **C/D box snoRNA (snoscan)**
- Practical guide

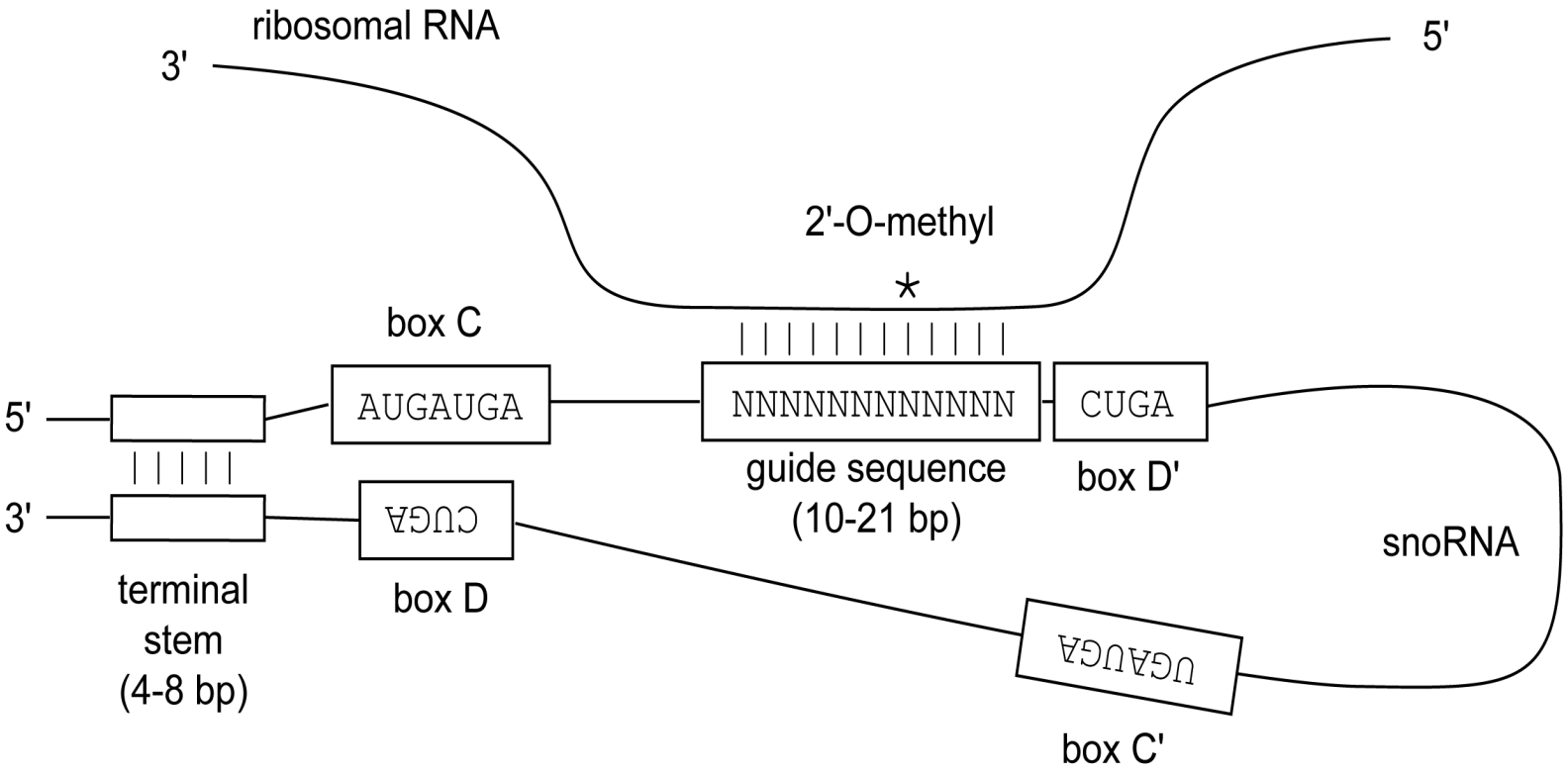
# C/D box snoRNA




- Guide 2'-*O*-methylation of target RNA by bound proteins
- 1 or 2 guide regions base pair to target RNA(s) to specify where methylation occurs
- Only known to methylate rRNA and tRNA in Archaea, and rRNA and snRNA in Eukarya

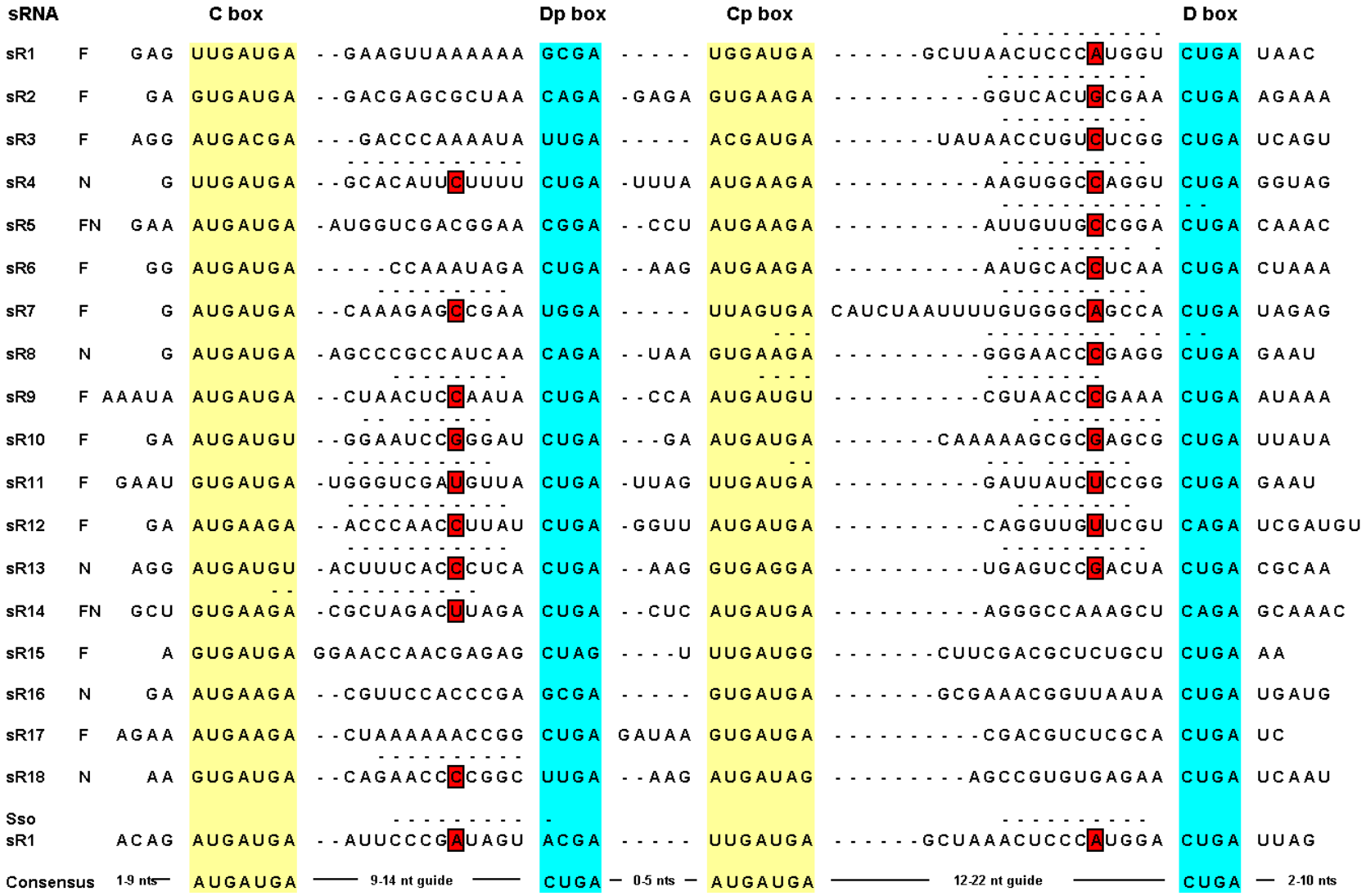
Figure from Dennis & Omer (2005).

# C/D box snoRNA canonical features




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

# Cloned sRNAs from *S. acidocaldarius*



# Useful C/D box snoRNA databases

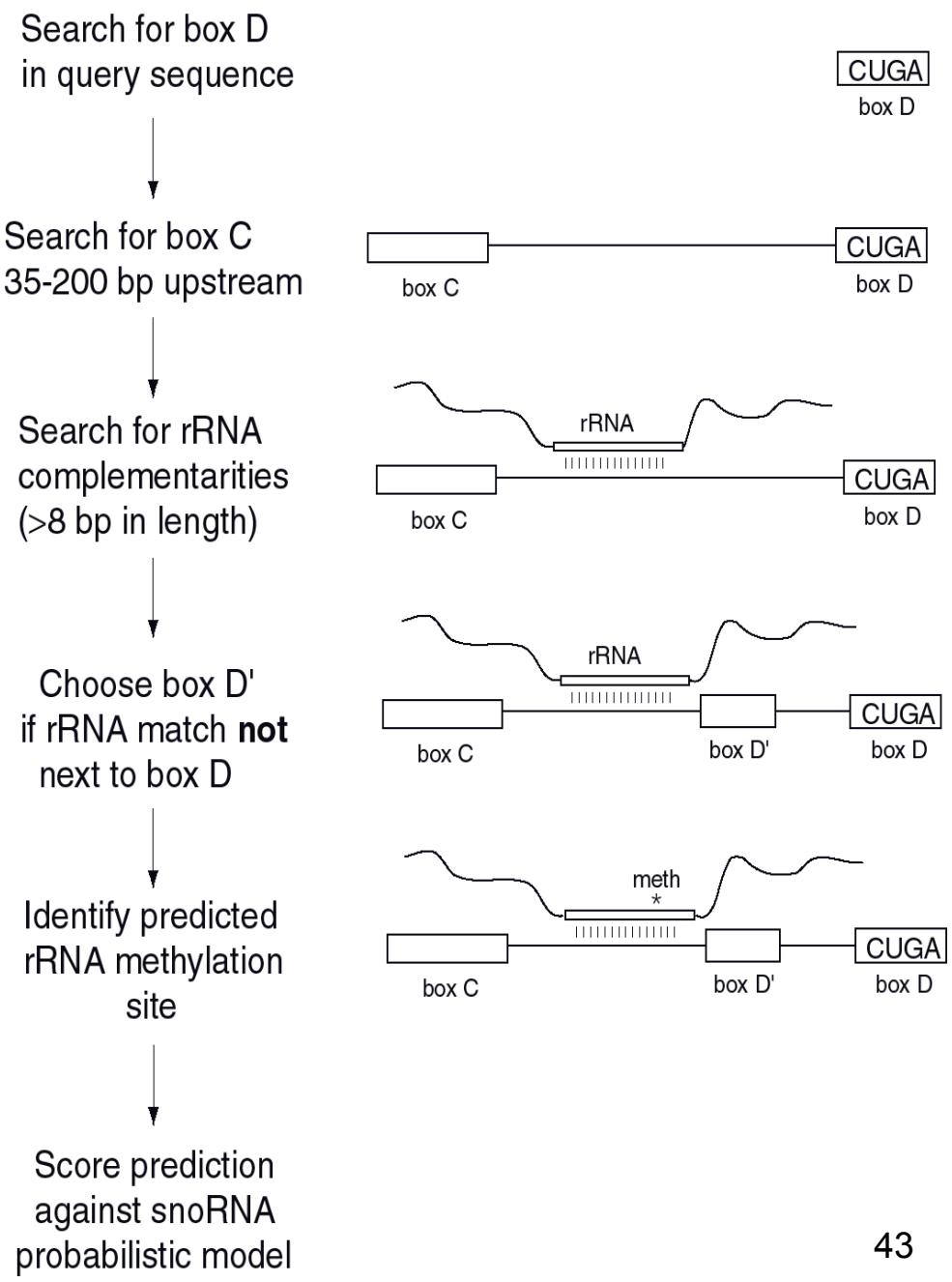
snoRNABase (human only)

http://www-snorna.biotoul.fr/

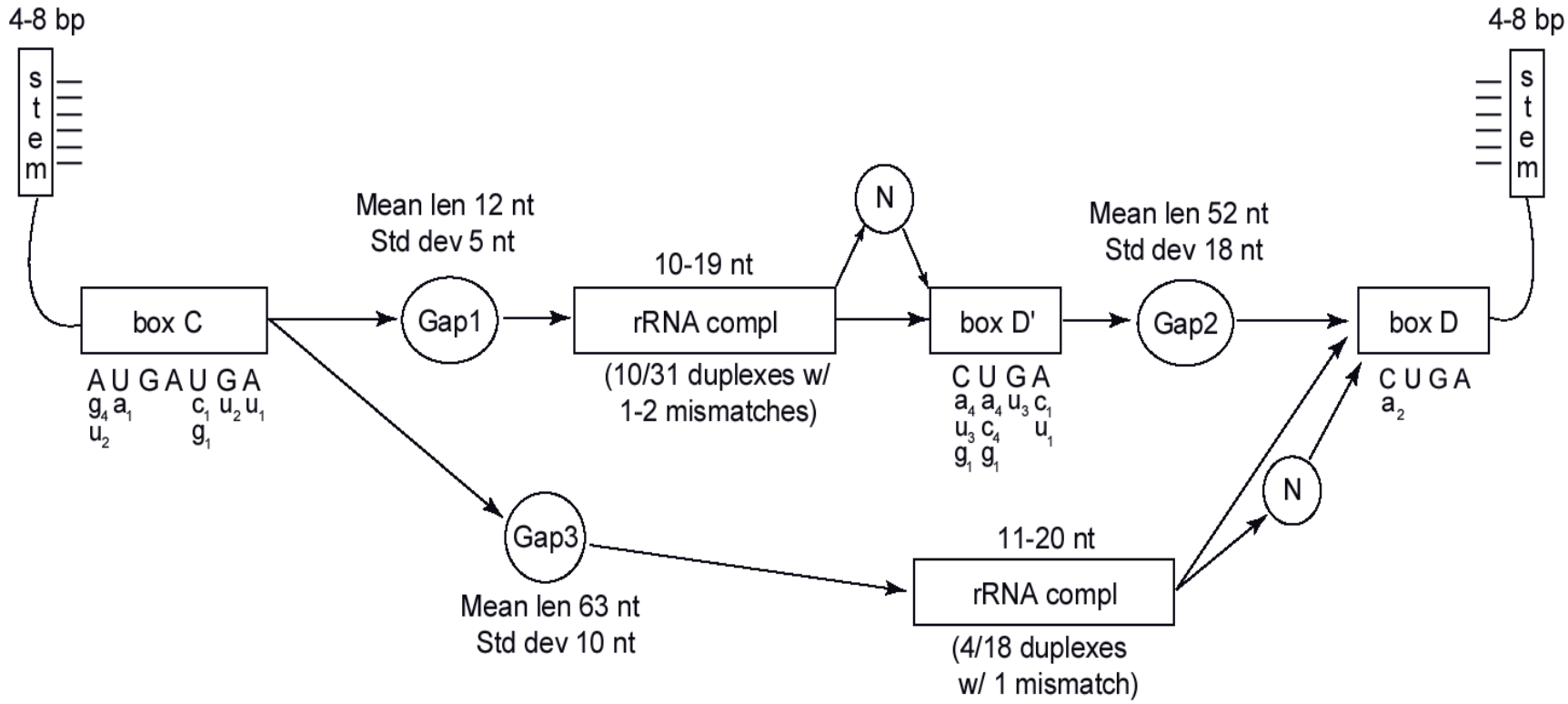
yeast snoRNA database

http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html

# snoscan:
a program to screen for C/D box snoRNA



# snoscan probabilistic model



# LIVE snoscan DEMO

**A mystery C/D box snoRNA:**

CUUCAGUGAUGACACGAUGACGAGUCAGAAAGGUCACGUCCUGC
UCUUGGUCCUUGUCAGUGCCAUGUUCUGUGGUGCUGUGCACGAG
UUCCUUUGGCAGAAGUGUCCUAUUUAUUGAUCGAUUUAGAGGCA
UUUGUCUGAGAAGG

# Talk progress

- Overview
- RNA structure
  - fundamentals
  - computational RNA folding (mfold)
- Gene-specific computational models
  - Rfam database
  - tRNA (tRNAscan-SE)
  - C/D box snoRNA (snoscan)
- **Practical guide**

# Finding ncRNA genes

- 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
  - Rfam, tRNAscan-SE, snoscan, etc.

# Let's say you found a mysterious new conserved sequence – 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 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

# Check Current ncRNA Databases

- **Some have search options**

New databases and resources always coming out – check links at IMB Jena (next slide) and annual database issue of *Nucleic Acids Research*

# Some ncRNA databases

- **SRP RNA Database:**
  http://bio.lundberg.gu.se/dbs/SRPDB/SRPDB.html
- **RNAse P Database:**
  http://jwbrown.mbio.ncsu.edu/RNaseP/
- **tmRNA Database (mirror, original is down):**
  **http://www.ag.auburn.edu/mirror/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 tracks in genome browsers

- In UCSC genome browsers, look for:
  - “RNA Genes”
  - “sno/miRNA genes”
  - “transfer RNAs”
  - “Genbank RNAs”
  - “RFAM RNAs”

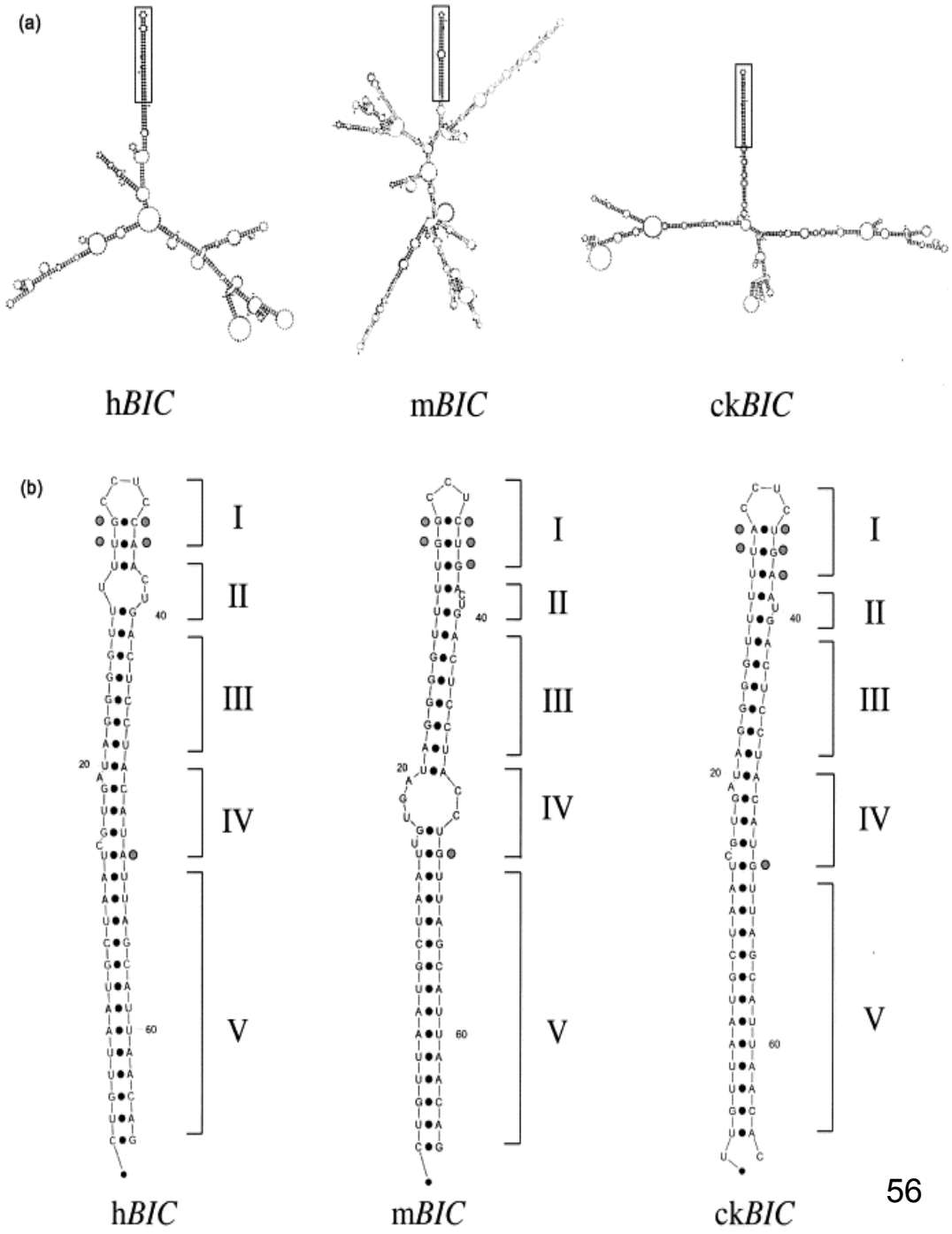
Or anything listed at “ncRNA”…

# Check structure for clues

- mfold server

- Is there one good optimal structure, or many within 10% 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 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 a covariance model (aka 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)

Often difficult to find ncRNAs on genetic screens.

Mining expression databases (EST sequencing, microarray data)

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

Direct sequencing of RNA fragments (RNA-Seq) !!