sbRNAs: To be or not to be a Y RNA

Ilenia Boria, Andreas Gruber, Andrea Tanzer,
Stephan Bernhart, Ronny Lorenz and Peter F. Stadler

Institute for Theoretical Chemistry
University of Vienna

TBI Winterseminar 2009

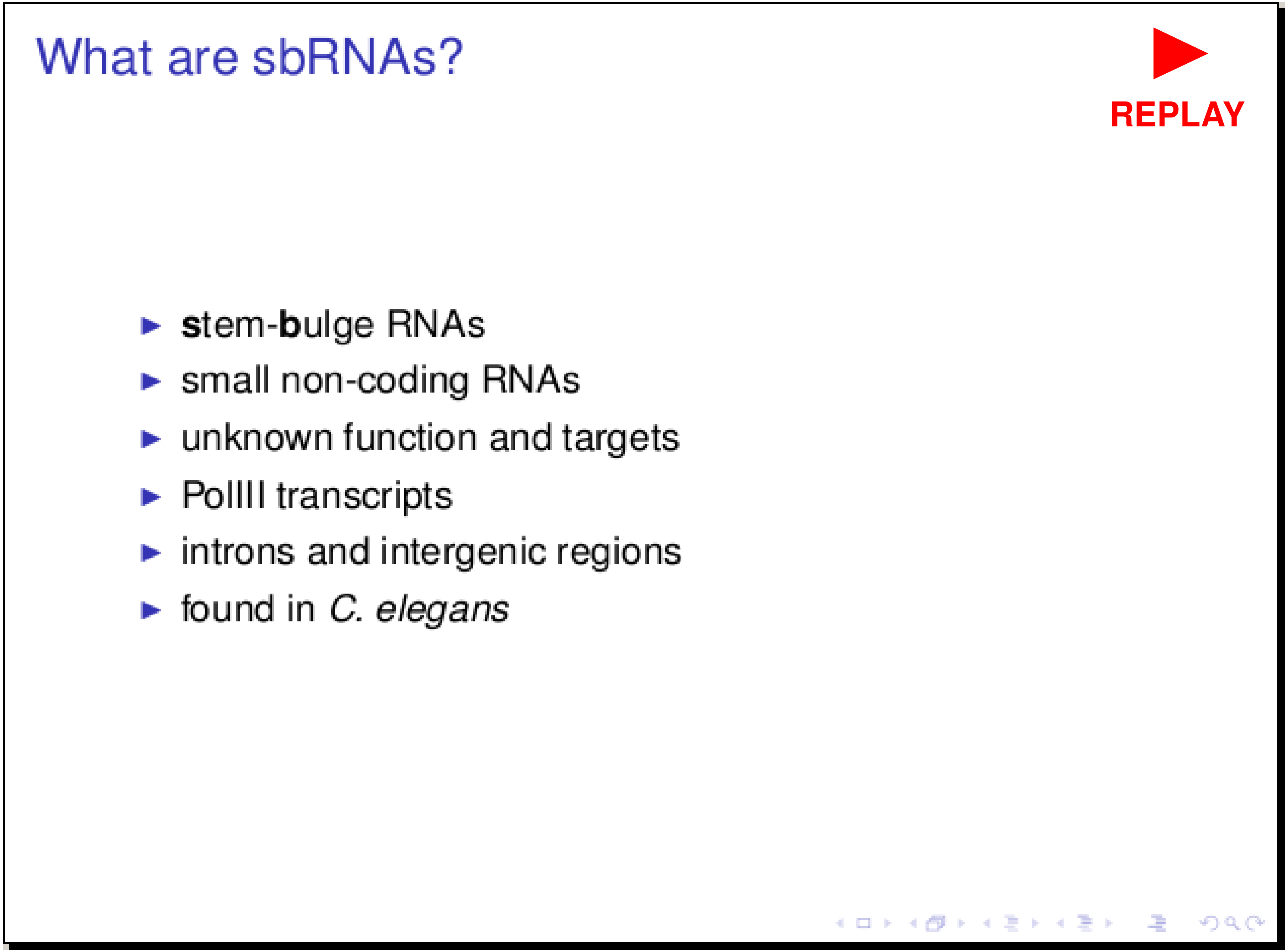
Outline

- Previously in Studeny 2008

- Homology search strategies

- To be or not to be a Y RNA

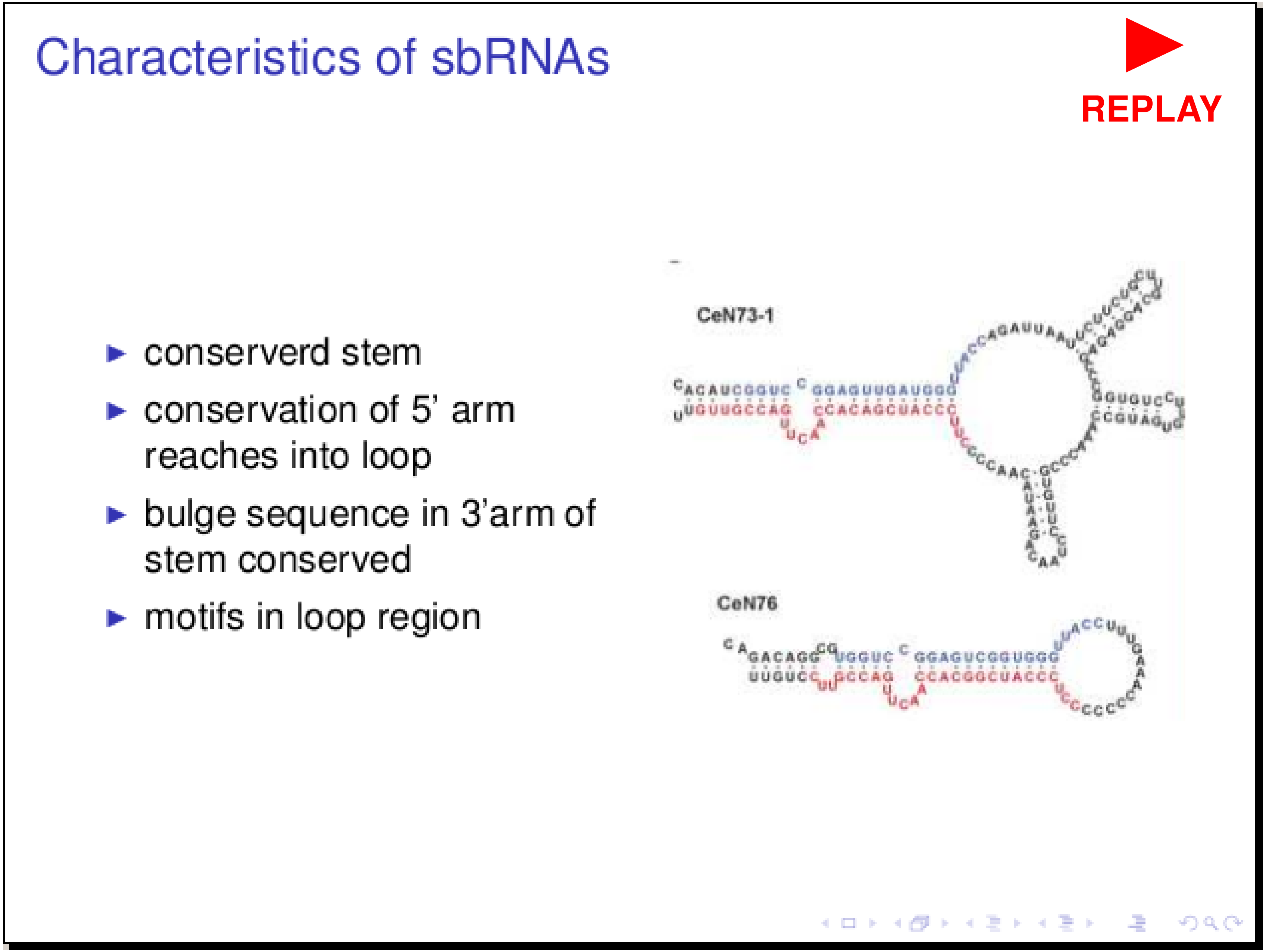
Previously in Studeny



What are sbRNAs?

- **s**tem-**b**ulge RNAs
- small non-coding RNAs
- unknown function and targets
- PolIII transcripts
- introns and intergenic regions
- found in *C. elegans*

Previously in Studeny

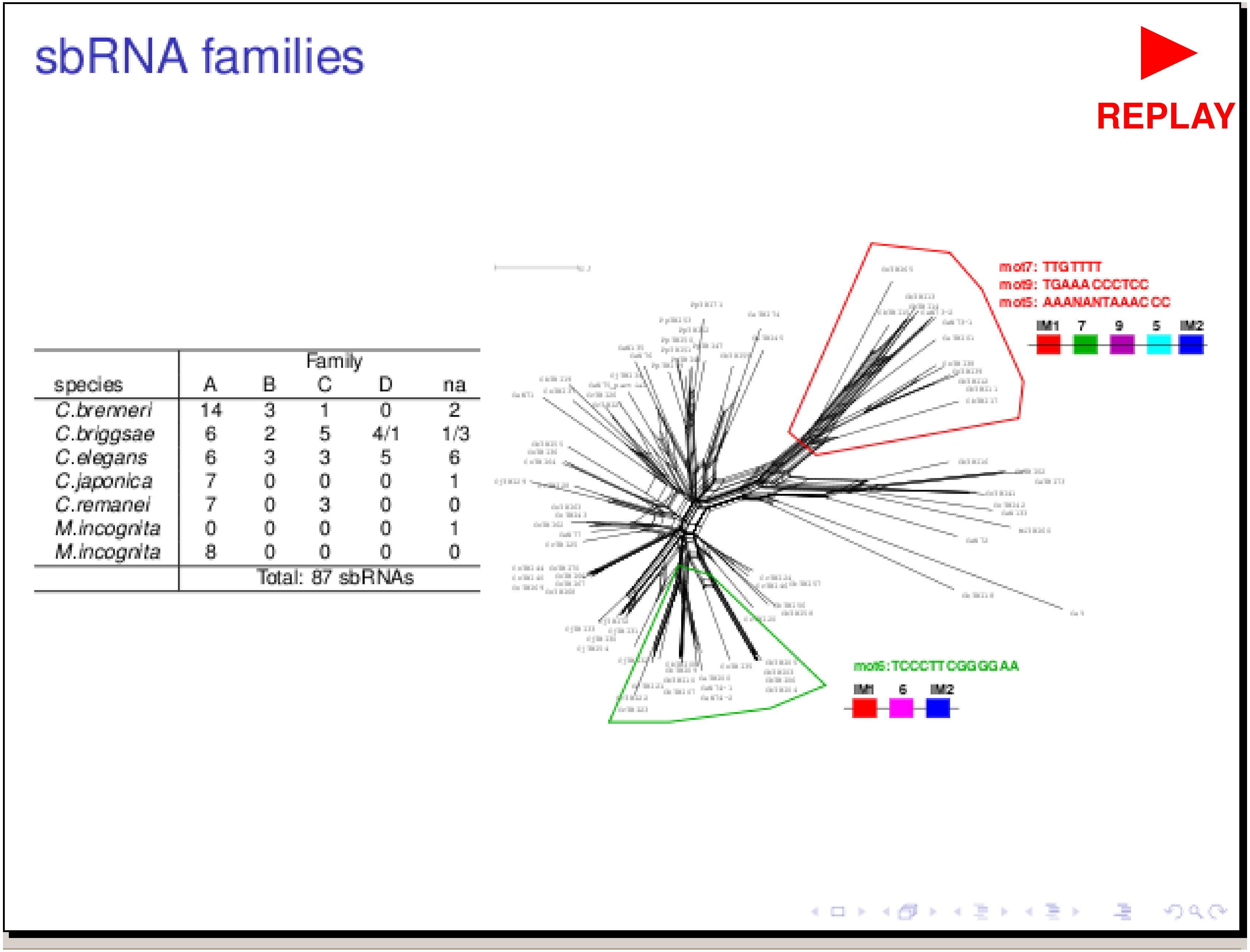


Characteristics of sbRNAs

REPLAY

- conserverd stem
- conservation of 5’ arm reaches into loop
- bulge sequence in 3’arm of stem conserved
- motifs in loop region

Previously in Studeny



sbRNA families

REPLAY

species	Family A	B	C	D	na
C.brenneri	14	3	1	0	2
C.briggsae	6	2	5	4/1	1/3
C.elegans	6	3	3	5	6
C.japonica	7	0	0	0	1
C.remanei	7	0	3	0	0
M.incognita	0	0	0	0	1
M.incognita	8	0	0	0	0
	Total: 87 sbRNAs				

At that time it was clear!

Just a few more days than we have a nice story and we can close this project!

At that time it was clear!

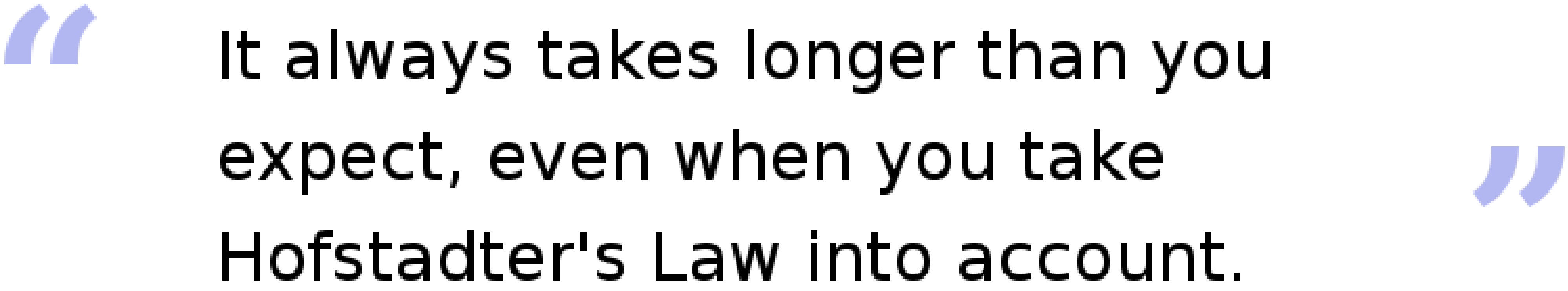
Just a few more days than we have a nice story and we can close this project!

Hofstadter's Law

At that time it was clear!

Just a few more days than we have a nice story and we can close this project!

Hofstadter's Law



> It always takes longer than you expect, even when you take Hofstadter's Law into account.

Almost 6 months later ...

		Confirmed with promoter	no promoter
V	C. elegans	18	0
	C. remanei	19	1
	C. briggsae	25	10
	C. brenneri	22	12
	C. japonica	10	2
	P. pacificus	22	0
	H. contortus	6 (42)	1
	A. caninum	4	0
IV	M. hapla	5	0
	M. incognita	10	0

Almost 6 months later ...

- No matches in the other clades of nematodes

- Just two conserved loci in Caenorhabditis that can be confirmed via syntenic regions

- The majority of the loops are short, unstructured and seem to evolve rapidly

- The shortened pol III promoter is specific to Caenorhabditis

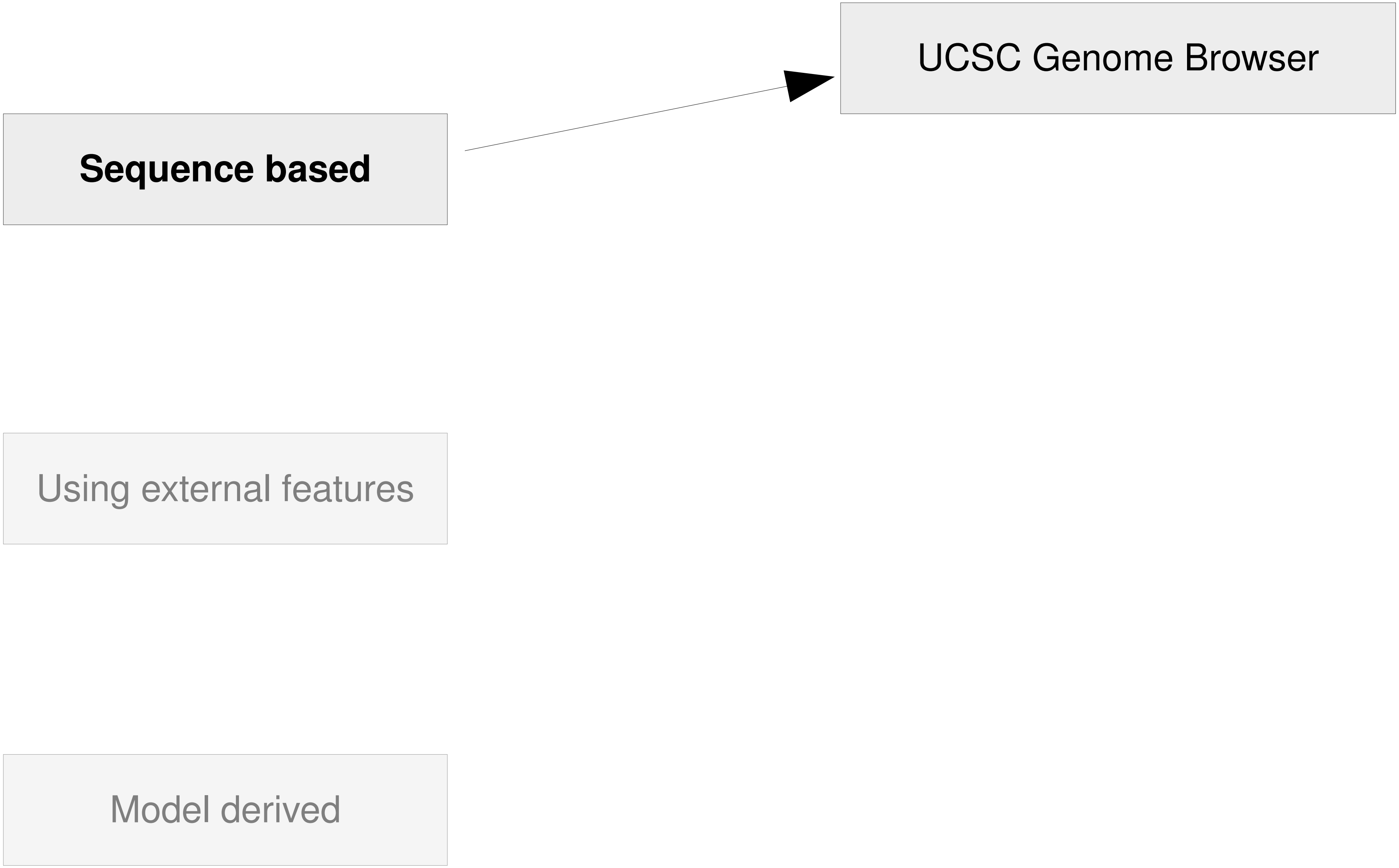
Homology Search Strategies

Sequence based

Using external features

Model derived

Homology Search Strategies



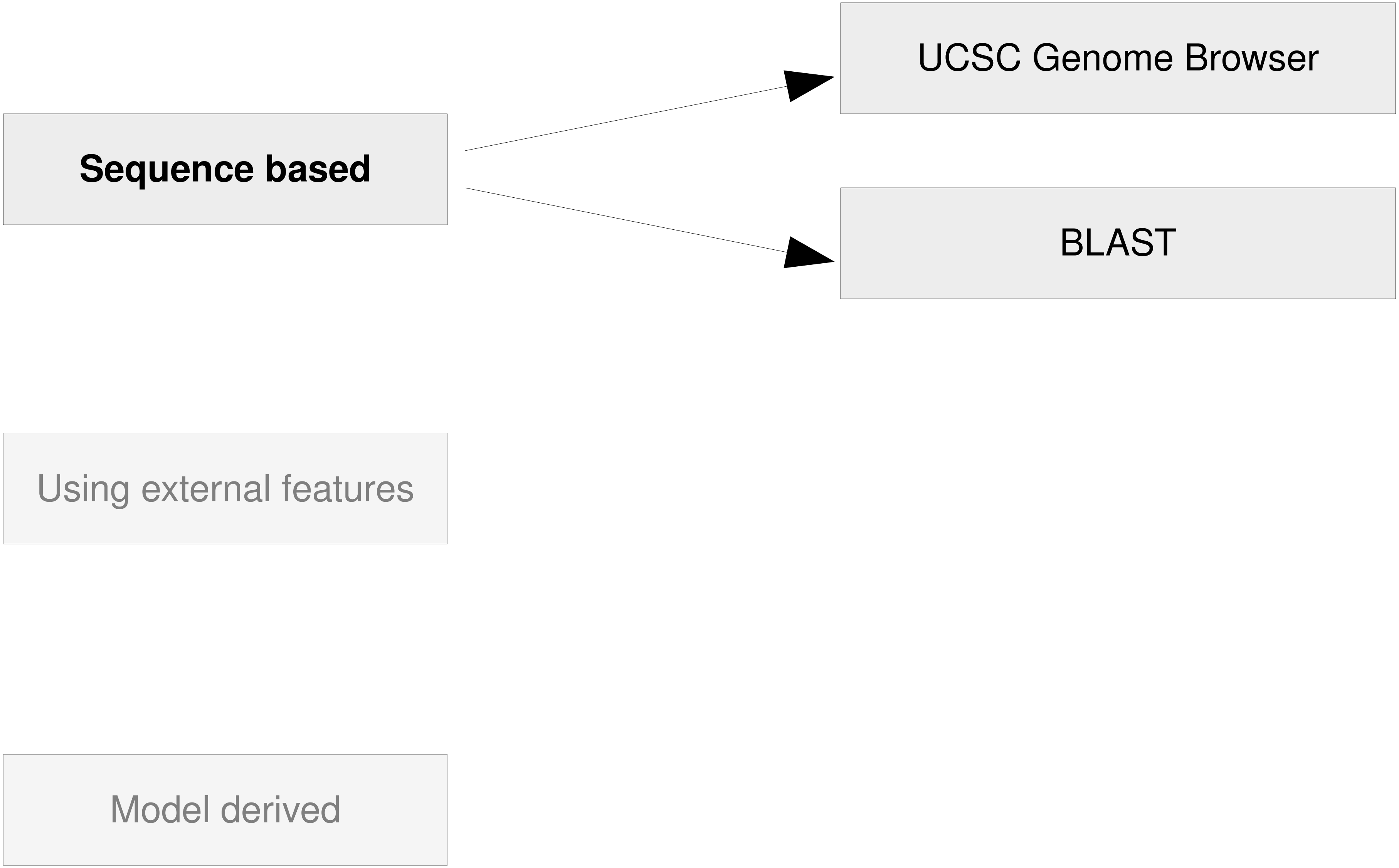
Sequence based

UCSC Genome Browser

Using external features

Model derived

Homology Search Strategies



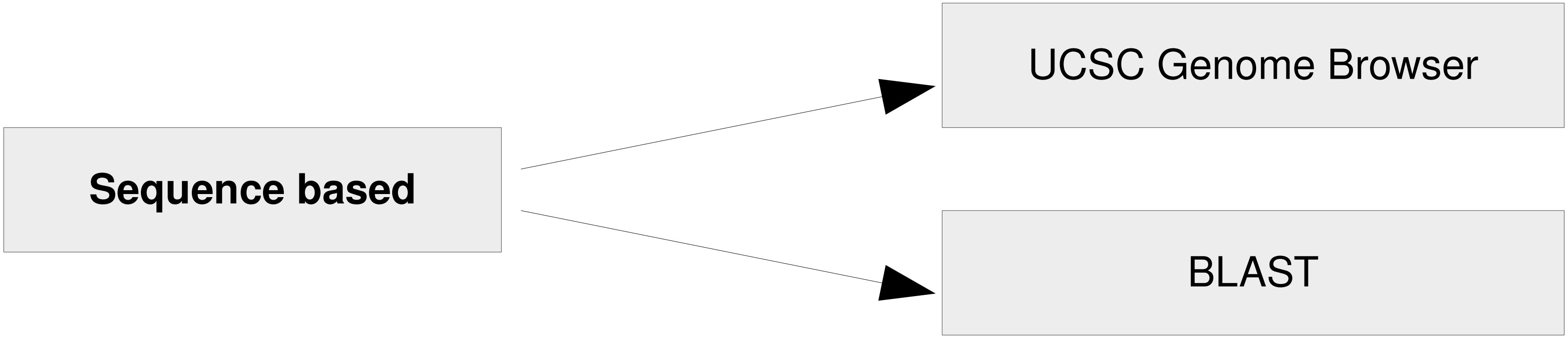
Sequence based

- UCSC Genome Browser
- BLAST

Using external features

Model derived

Homology Search Strategies



Sequence based

UCSC Genome Browser

BLAST

Using external features

Model derived

Pros

Easy and fast to build an initial stock of candidates

Cons

As sbRNA loop motifs evolve rapidly, hard to get good candidates outside the genus of Caenorhabditis

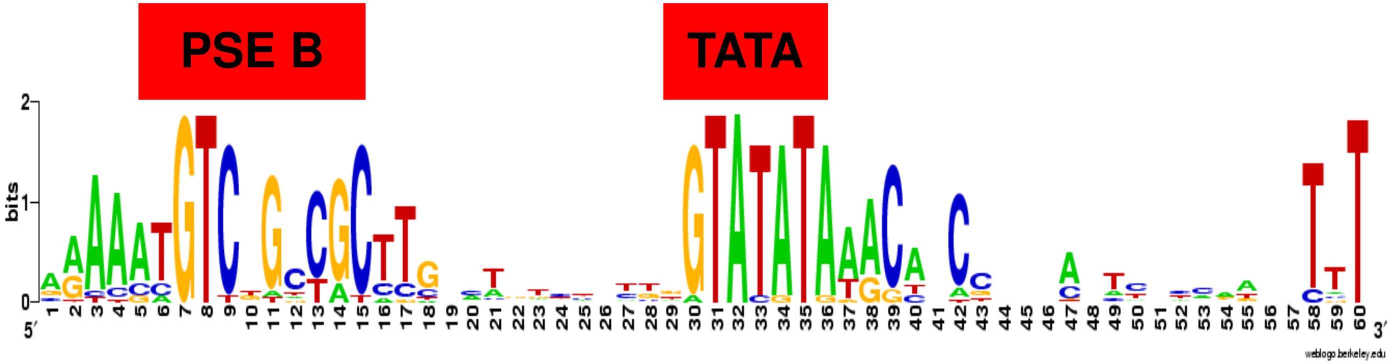
Homology Search Strategies

Sequence based

Using external features

Model derived

sbRNAs have characteristic promoter motifs.



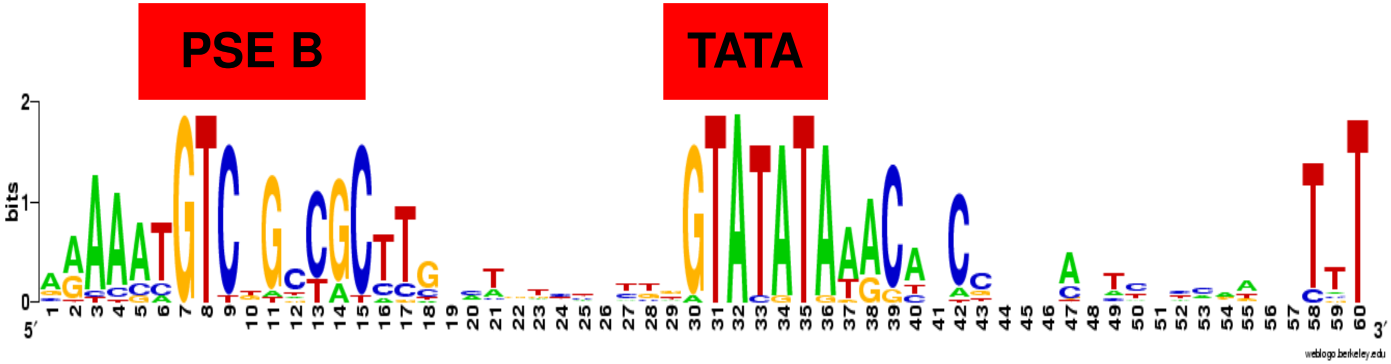
Homology Search Strategies

Sequence based

Using external features

Model derived

sbRNAs have characteristic promoter motifs.



Search for U6, RNAseP or RNAseMRP, tRNA-Sec

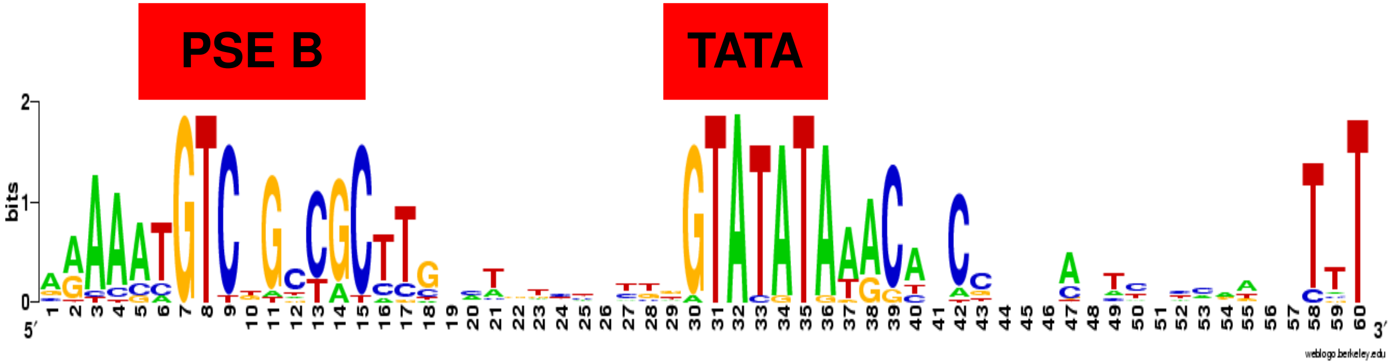
Homology Search Strategies

Sequence based

Using external features

Model derived

sbRNAs have characteristic promoter motifs.



Search for U6, RNAseP or RNAseMRP

Build a FRAGREP model

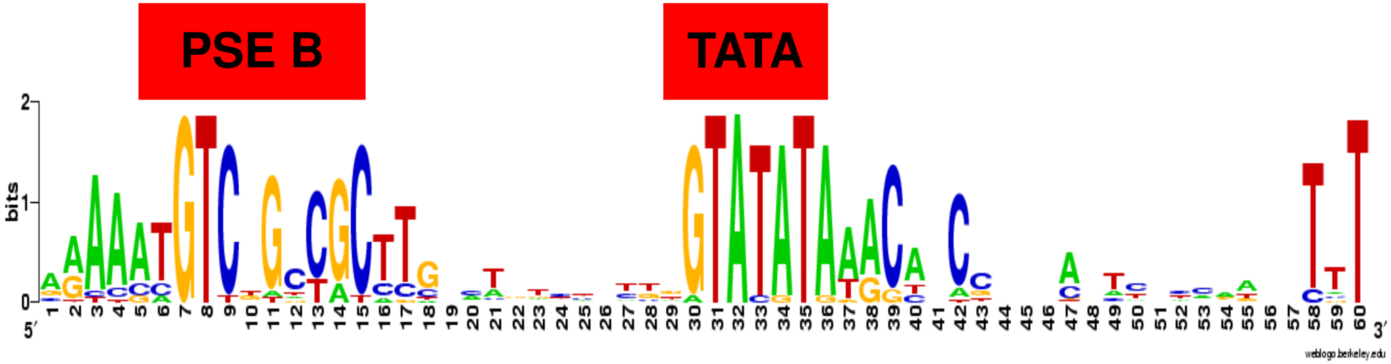
Homology Search Strategies

Sequence based

Using external features

Model derived

sbRNAs have characteristic promoter motifs.



Search for U6, RNAseP or RNAseMRP, tRNA-Sec

Build a FRAGREP model

Search upstream of FRAGREP hits until you find TTTT

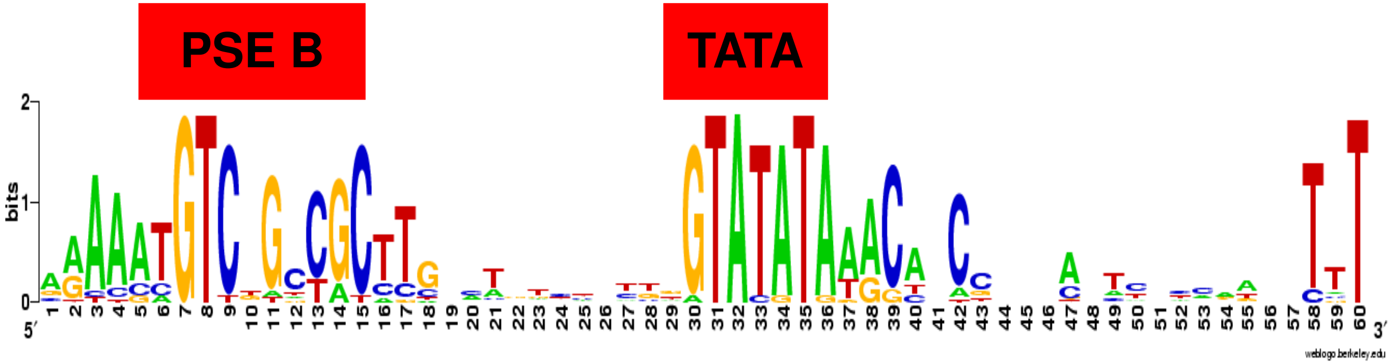
Homology Search Strategies

Sequence based

Using external features

Model derived

sbRNAs have characteristic promoter motifs.



Search for U6, RNAseP or RNAseMRP, tRNA-Sec

Build a FRAGREP model

Search upstream of FRAGREP hits until you find TTTT

Cluster hits using RNAclust, Locarana, RNAsoup

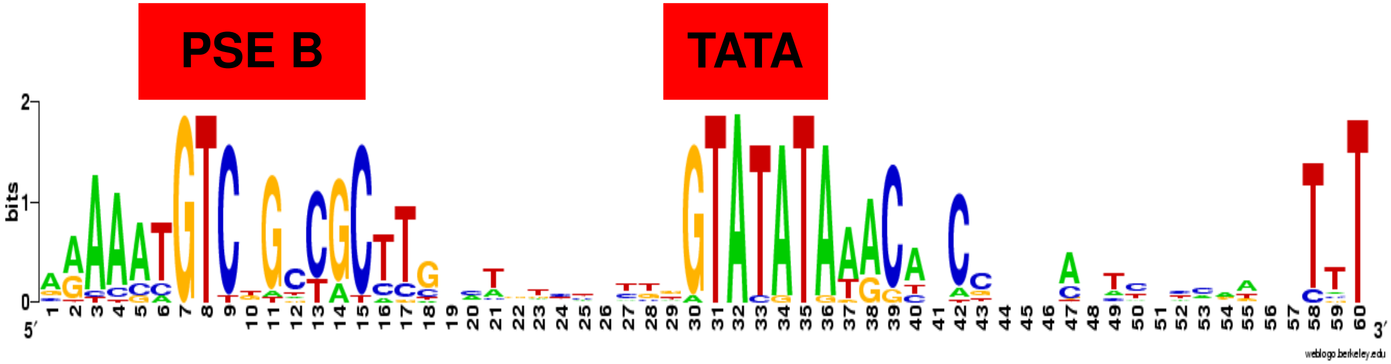
Homology Search Strategies

Sequence based

Using external features

Model derived

sbRNAs have characteristic promoter motifs.



Pros

You are not limited by structural restrictions like in model based search

You are likely to find high confident hits as those hits also have a promoter region

Cons

You have to find other Pol III transcripts first

Clustering is computationally expensive

Homology Search Strategies

Build a structure model

Sequence based

Using external features

Model derived

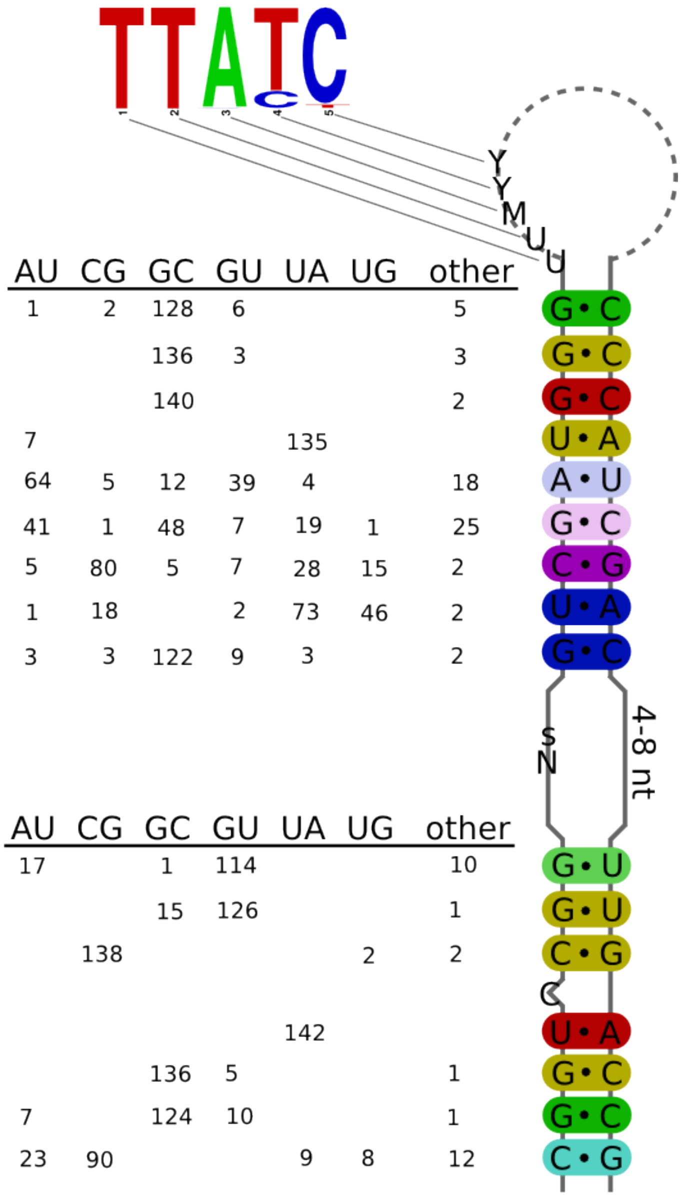
Homology Search Strategies

Build a structure model

Sequence based

Using external features

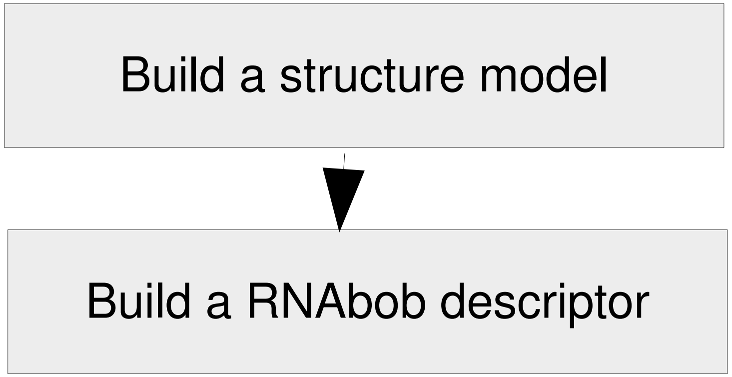
Model derived



AU	CG	GC	GU	UA	UG	other
1	2	128	6			5
		136	3			3
		140				2
7				135		
64	5	12	39	4		18
41	1	48	7	19	1	25
5	80	5	7	28	15	2
1	18		2	73	46	2
3	3	122	9	3		2

AU	CG	GC	GU	UA	UG	other
17		1	114			10
		15	126			1
	138				2	2
				142		
		136	5			1
7		124	10			1
23	90			9	8	12

Homology Search Strategies


Build a structure model

Build a RNAbob descriptor

Sequence based

Using external features

Model derived

Homology Search Strategies

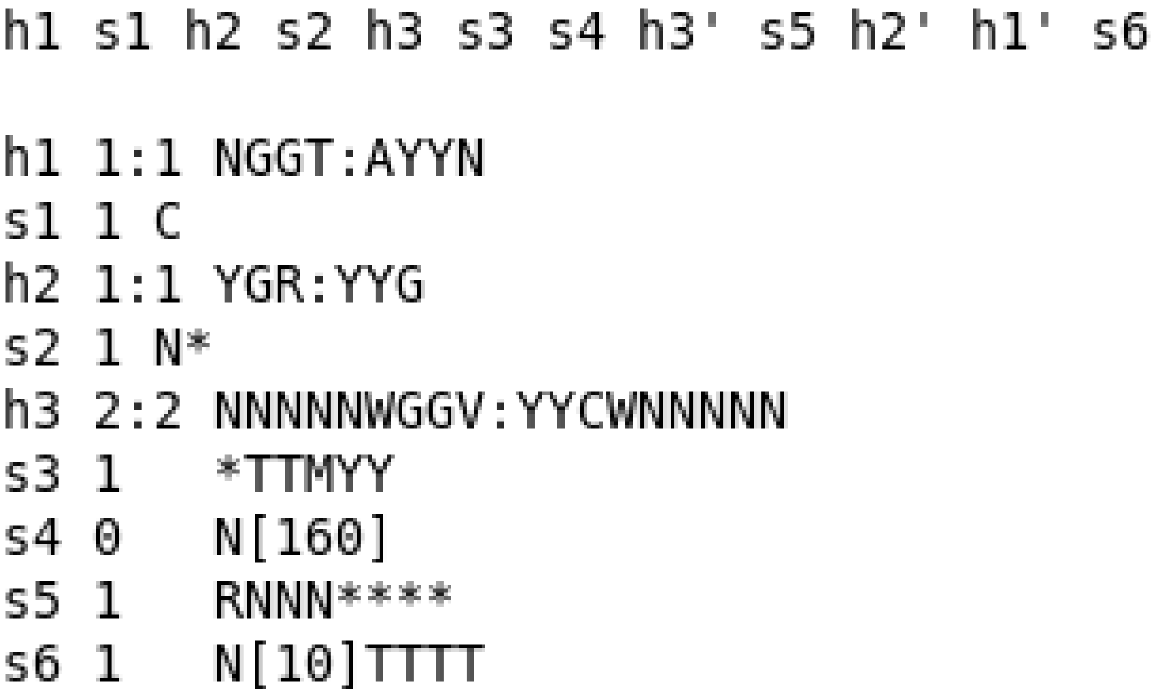
Sequence based

Using external features

Model derived

Build a structure model

Build a RNAbob descriptor

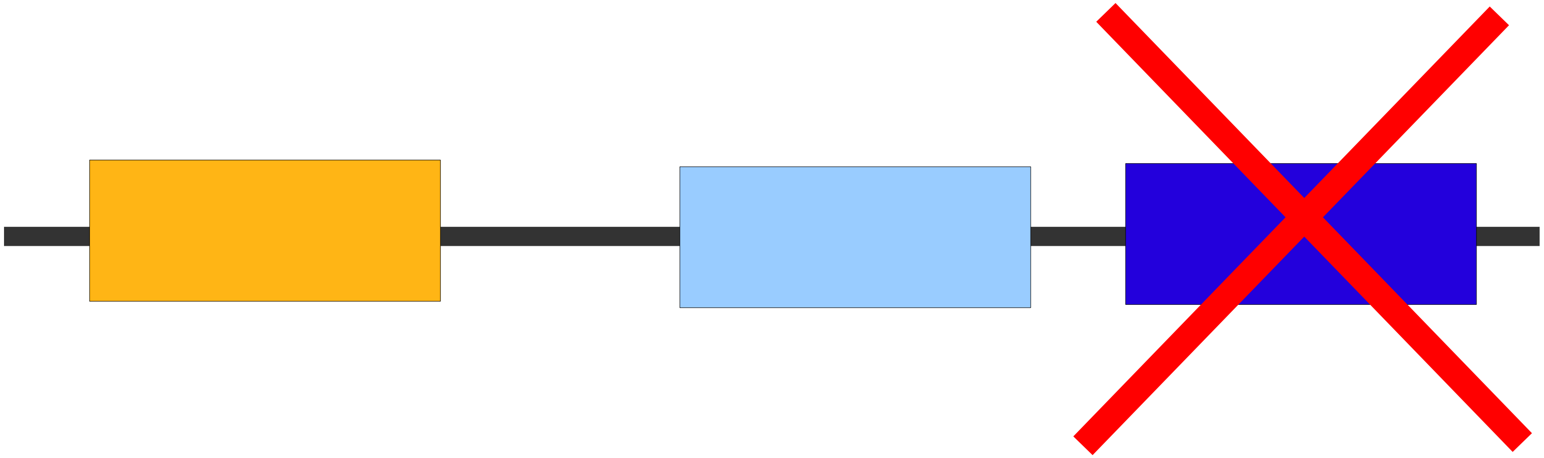


```
h1 s1 h2 s2 h3 s3 s4 h3' s5 h2' h1' s6

h1 1:1 NGGT:AYYN
s1 1 C
h2 1:1 YGR:YYG
s2 1 N*
h3 2:2 NNNNNWGGV:YYCWNNNNN
s3 1   *TTMYY
s4 0   N[160]
s5 1   RNNN****
s6 1   N[10]TTTT
```

Watch out! RNAbob bug

Watch out! RNAbob bug



Homology Search Strategies

Sequence based

Using external features

Model derived

Build a structure model

Build a RNAbob descriptor

Screen genome

Homology Search Strategies

Sequence based

Using external features

Model derived

Build a structure model

Build a RNAbob descriptor

Screen genome

~ 380.000 candidates
in C. elegans

Homology Search Strategies

Sequence based

Using external features

Model derived

Build a structure model

Build a RNAbob descriptor

Screen genome

Filter hits

Filtering

First attempt

Calculate the number of mismatches to already known sequences

Second attempt

Scoring with a position weight matrix

Take a structural alignment and count occurrences of bases and base pairs in regions of interest

Filtering

$$mSS = \frac{Current - Min}{Max - Min}$$

→ Normalized score between 0 ... 1

$$Current: \sum_{i=1}^{L} I(i) f_{i,b_i} \qquad Max: \sum_{i=1}^{L} I(i) f_i^{\max} \qquad Min: \sum_{i=1}^{L} I(i) f_i^{\min}$$

$$I(i) = \sum \quad f_{i,B} \ln(m\, f_{i,B}), \quad i = 1, 2, \ldots, L$$

m = 4 if B = {A,C,G,T}, m = 16 if B = {AA,AC,AG,AT, ..., TT}

Kel et al. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 2003 Jul 1;31(13):3576-9.

Homology Search Strategies

Sequence based

Using external features

Model derived

Build a structure model

Build a RNAbob descriptor

Screen genome

Filter hits

High confident hits

Y RNAs

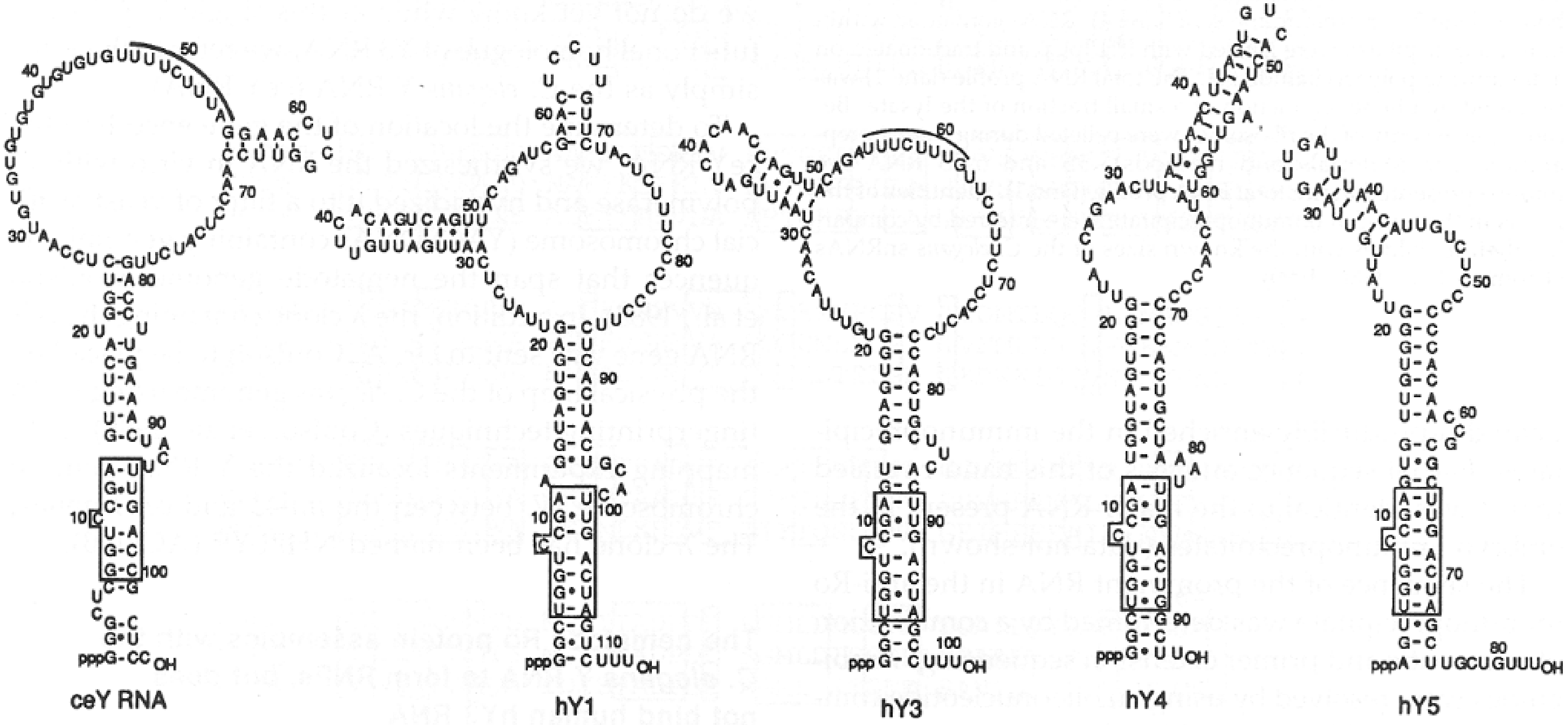


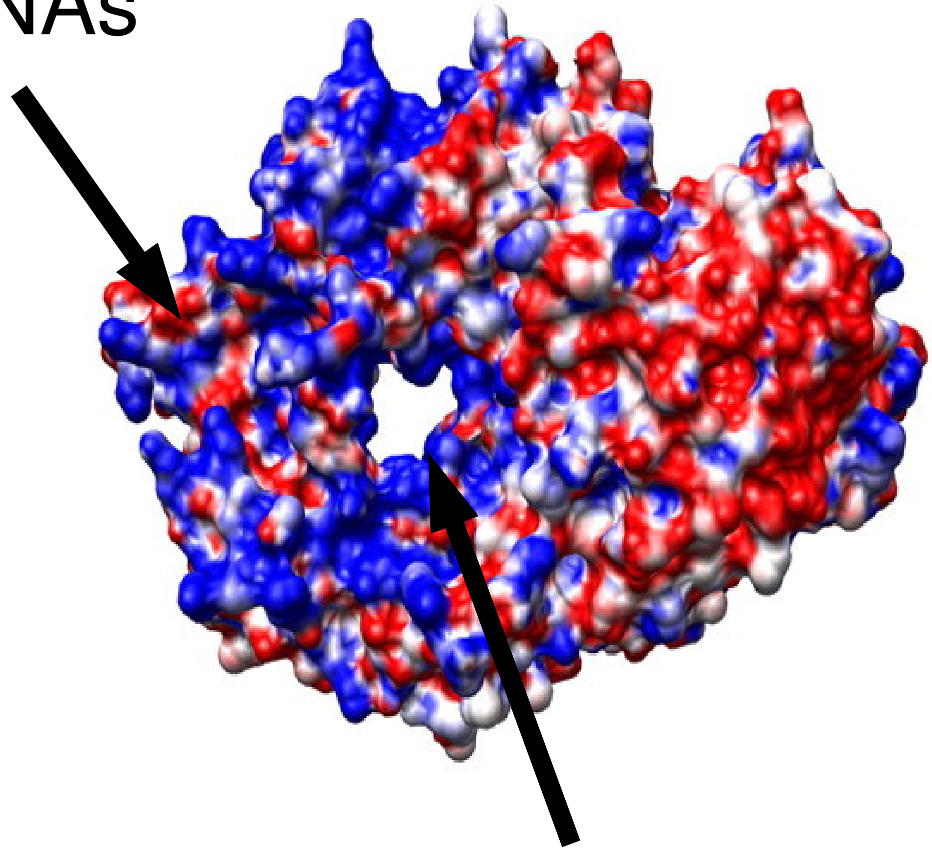
FIGURE 5. Potential secondary structures of *C. elegans* and human Y RNAs. Structures of the four human Y RNAs were proposed by O'Brien et al. (1993) and were drawn to maximize structural similarities between the human and *Xenopus* Y RNAs. A conserved helix that was proposed to be the binding site for the 60-kDa Ro protein is boxed (Wolin & Steitz, 1984). A sequence in the loop of ceY RNA that is also present in the human, *Xenopus*, and iguana Y3 RNAs is indicated by the line.

Van Horn DJ, Eisenberg D, O'Brien CA, Wolin SL. Caenorhabditis elegans embryos contain only one major species of Ro RNP. RNA. 1995 May;1(3):293-303.

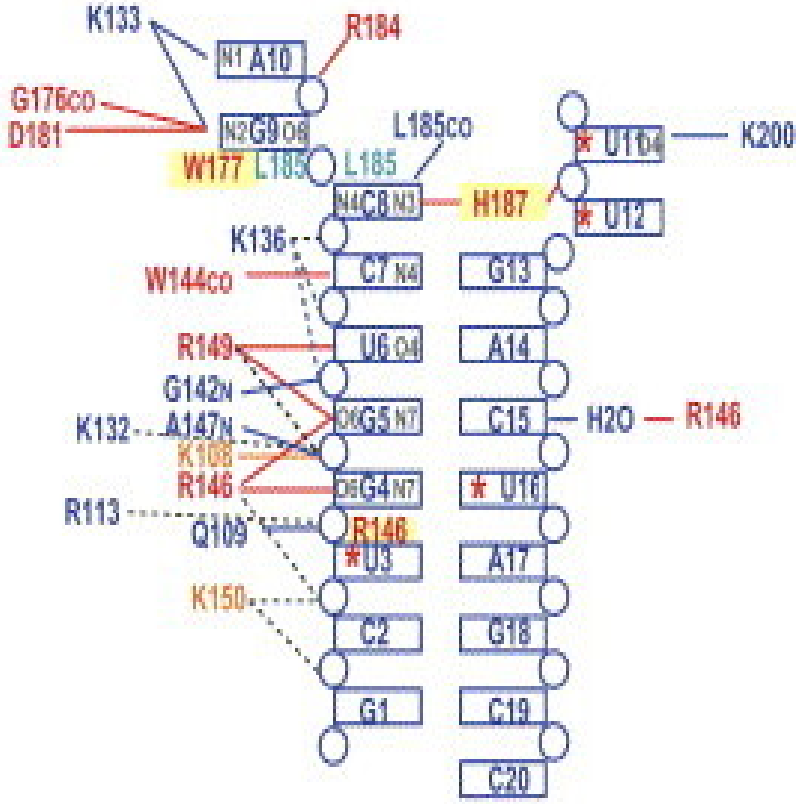
Ro RNP

Ro RNP

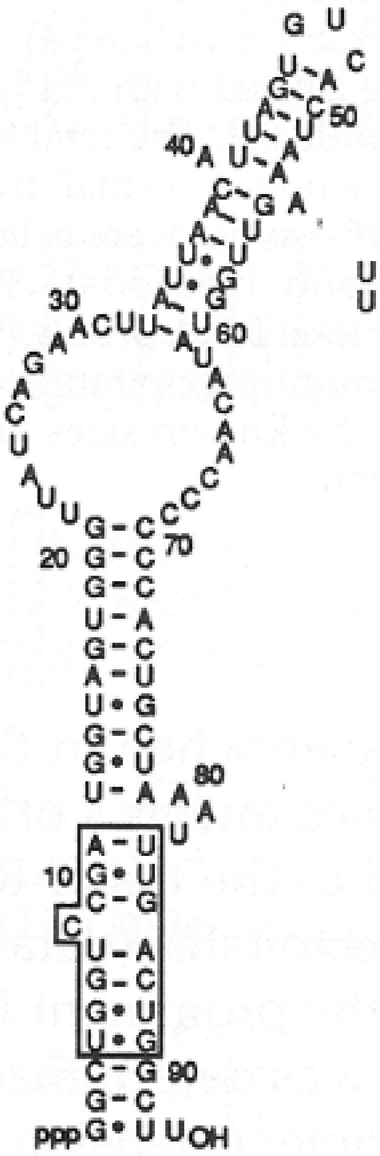
Y RNAs



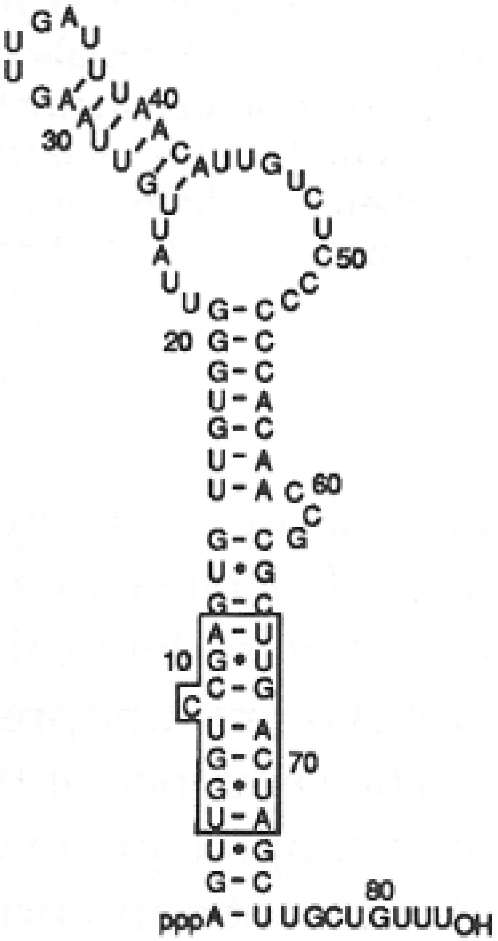
Single stranded RNA binding



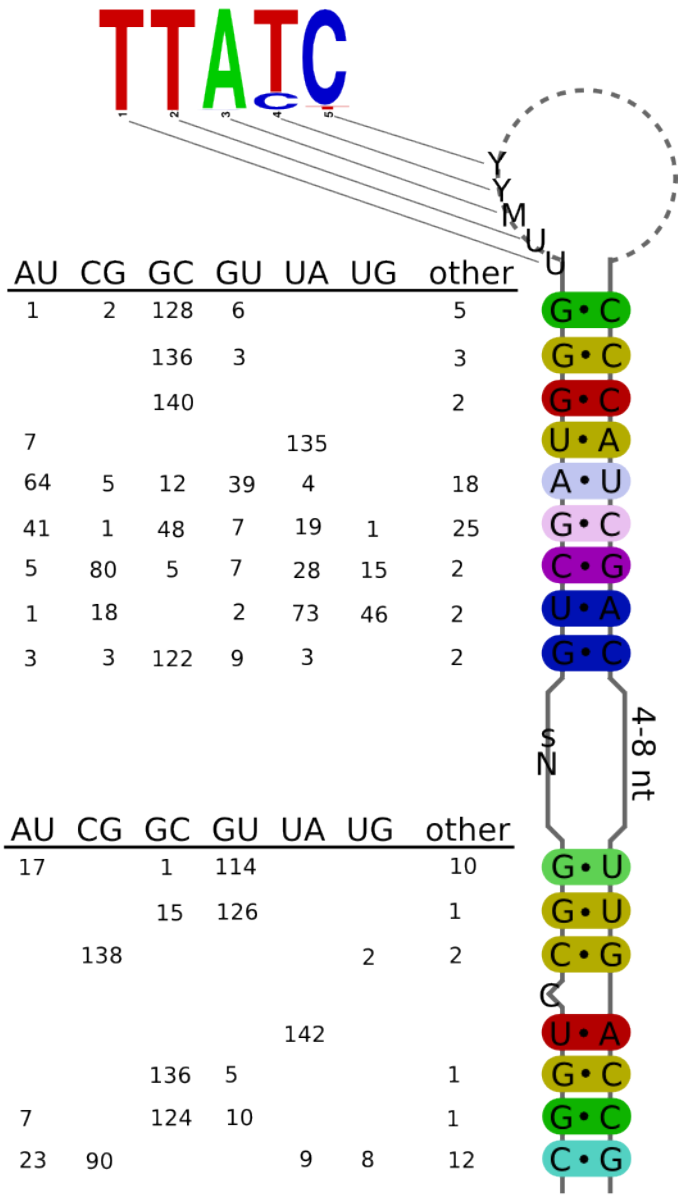
Y RNAs vs. sbRNAs



hY4



hY5



AU	CG	GC	GU	UA	UG	other	
1	2	128	6			5	G • C
		136	3			3	G • C
		140				2	G • C
7				135			U • A
64	5	12	39	4		18	A • U
41	1	48	7	19	1	25	G • C
5	80	5	7	28	15	2	C • G
1	18		2	73	46	2	U • A
3	3	122	9	3		2	G • C

AU	CG	GC	GU	UA	UG	other	
17		1	114			10	G • U
		15	126			1	G • U
	138				2	2	C • G
				142			U • A
		136	5			1	G • C
7		124	10			1	G • C
23	90			9	8	12	C • G

The sbRNA Team

Ilenia Boria, Andreas Gruber, Andrea Tanzer,
Stephan Bernhart, Ronny Lorenz and Peter F. Stadler