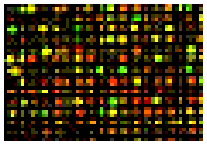
# Using Constraints to Determine RNA Secondary and Local 3D Structure

Douglas H. Turner

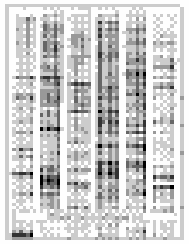
Department of Chemistry University of Rochester

# Focus of the Turner Lab:
# Predicting Structure from Sequence

GAAUUGCGGGAA
AGGGGUCAACAG
CCGUUCAGUACC
AAGUCUCAGGGG
AAACUUUGAGAU
GGCCUUGCAAAG
GGUAUGGUAAUA
AGCUGACGGACA
UGGUCCUAACCA
CGCAGCCAAGUC
CUAAGUCAACAG
AUCUUCUGUUGA
UAUGGAUGCAGU
UC



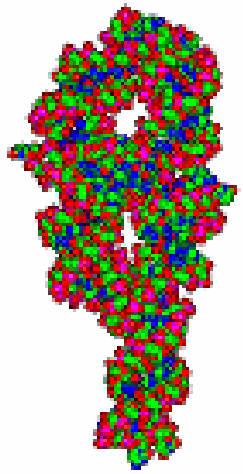
Microarray



Thermodynamic parameters
& Enzymatic mapping
& Chemical mapping


Tetrahymena thermophila LSU rRNA
GenBank# J01235
Eucarya, Protoctista, Ciliophora (IC1)

NMR
Chrystallography

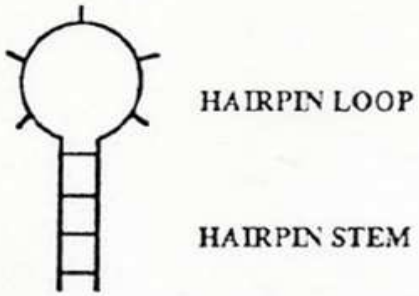


Cate, J. H. et al. 1996
Science 273, 1678-1685

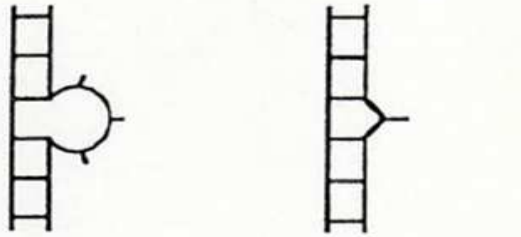
**Tinoco-Uhlenbeck Postulate**:
Assume: The energy of each base-pair is independent of all of the other pairs and the loop structure.
Consequence: Total free energy is the sum of all base-pair free energies.

**Turner Rules**:
Assume: The energy of secondary structure motifs is independent of all other secondary structure motifs.
Consequence: Total free energy is the sum of all secondary structure motifs.

a. HAIRPINS

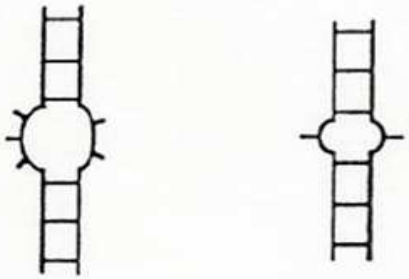



b. BULGES



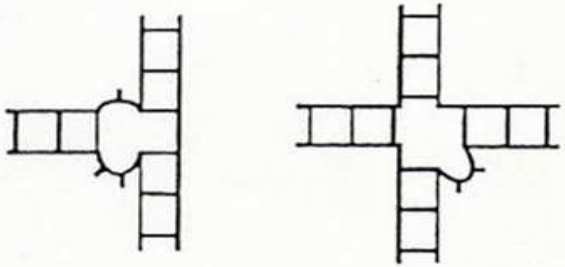
BULGE SINGLE-BASE BULGE

c. INTERNAL LOOPS



INTERNAL LOOP MISMATCH

d. JUNCTIONS



THREE-WAY FOUR-WAY

- COAXIAL STACKS

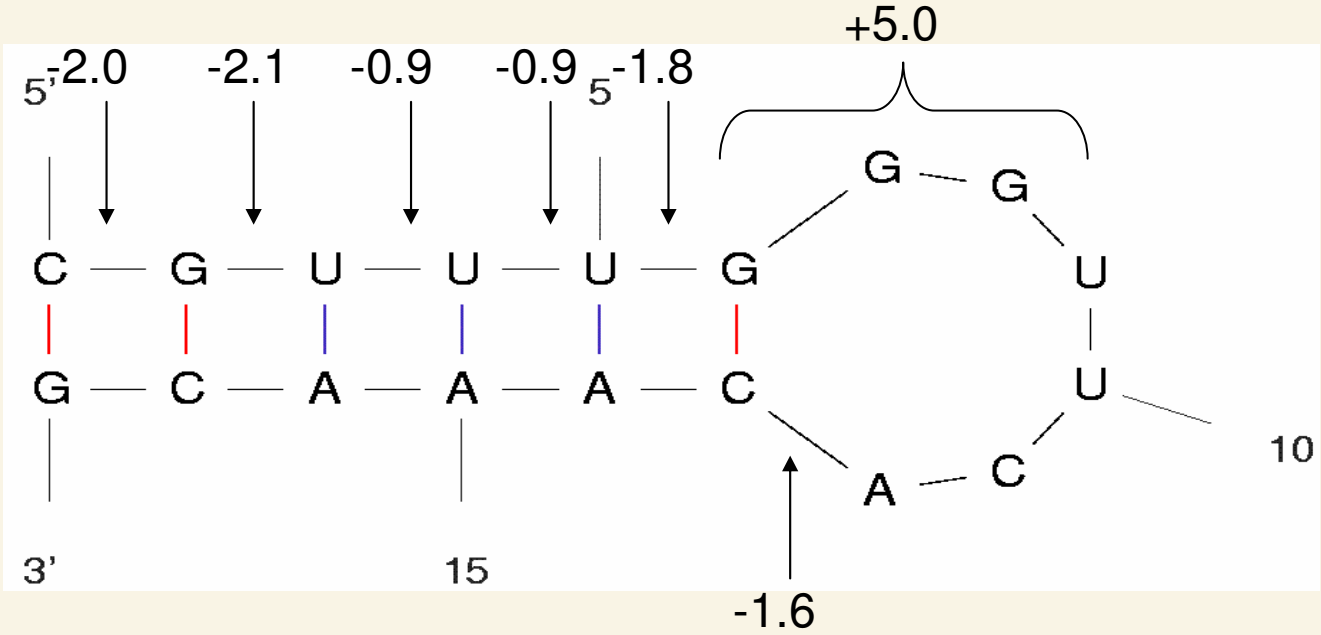
- PSEUDO-KNOTS

## Nearest Neighbor Model

$$\Delta G^{o}_{stem} = \Delta G^{o}\begin{bmatrix} C & G \\ G & C \end{bmatrix} + \Delta G^{o}\begin{bmatrix} G & U \\ C & A \end{bmatrix} + 2\Delta G^{o}\begin{bmatrix} U & U \\ A & A \end{bmatrix} + \Delta G^{o}\begin{bmatrix} U & G \\ A & C \end{bmatrix} = -7.7 kcal/mol$$

$$\Delta G^{o}_{loop} = \Delta G^{o}_{6-nucleotide} + \Delta G^{o}_{non-canonical}\begin{bmatrix} G & G \\ C & A \end{bmatrix} = 3.4 kcal/mol$$

$$\Delta G^{o}_{total} = \Delta G^{o}_{stem} + \Delta G^{o}_{hairpin\ loop} = -4.3 kcal/mol$$



# Reasons for limited sec. str. Prediction accuracy

- 1. The thermodynamic rules are incomplete.
- 2. RNA sequences may adopt sec. str. determined by folding kinetics.
- 3. Structure prediction algorithms use approximations.
- 4. Some RNA sequences might fold into more than one structure. (conformation change of riboswitches which control translation due to the environment)

# Secondary Structure Improvement

- 1. Efficient use of experimental mapping.
- 2. Statistical mechanics techniques to improve fidelity.
- 3. Inclusion of pseudoknots.
- 4. Use of two or more homologous sequences.

# Experimental constraints in secondary structure prediction

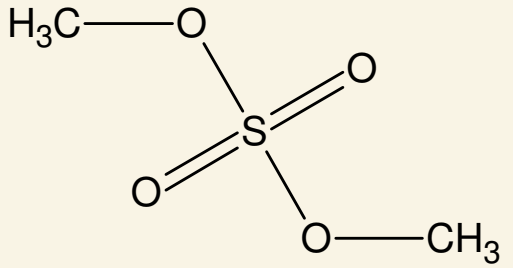
"The assumption is made that a nucleotide cannot be in a W-C pair flanked by W-C pairs if it is chemically modified at a W-C pairing position when the folded RNA is exposed to reagents such as dimethyl sulfate and kethoxal".

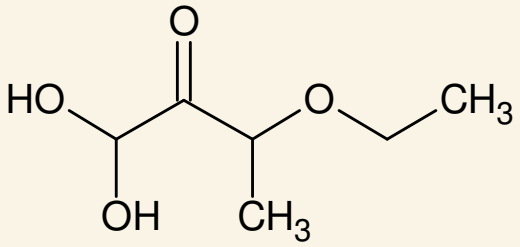
Turner et al. (2004) *PNAS* **101**, 7287-7292.

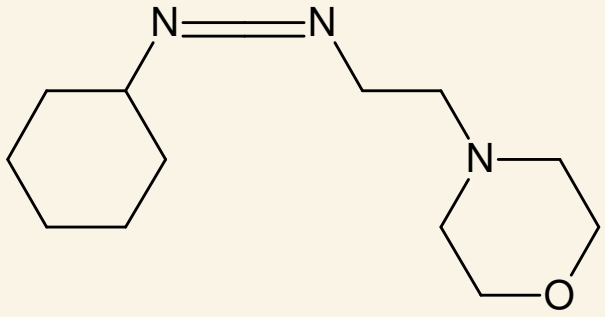
Some nucleotides in folded RNA structures in solution are accessible for modification.
These nucleotides are:

- Unpaired
- In A-U or G-C pairs at helix ends
- In G-U pairs anywhere
- Adjacent to G-U pairs.

- DMSO
- Kethoxal
- CMCT



Nearest neighbor parameter revised for:

- Terminal mismatches
- Hairpins
- Bulge loops
- Internal loops
- Multibranch loops (Coaxial stacking, allow one mismatch at most)

Important notes on the particular PNAS (2004) modification experiment:

- It's an *in vivo* study, therefore no need to find conditions which mimic *in vitro* native structure.
- Chemical modification constraints obtained in the presence and absence of protein and/or under different conditions may reveal conformational switch dynamics.

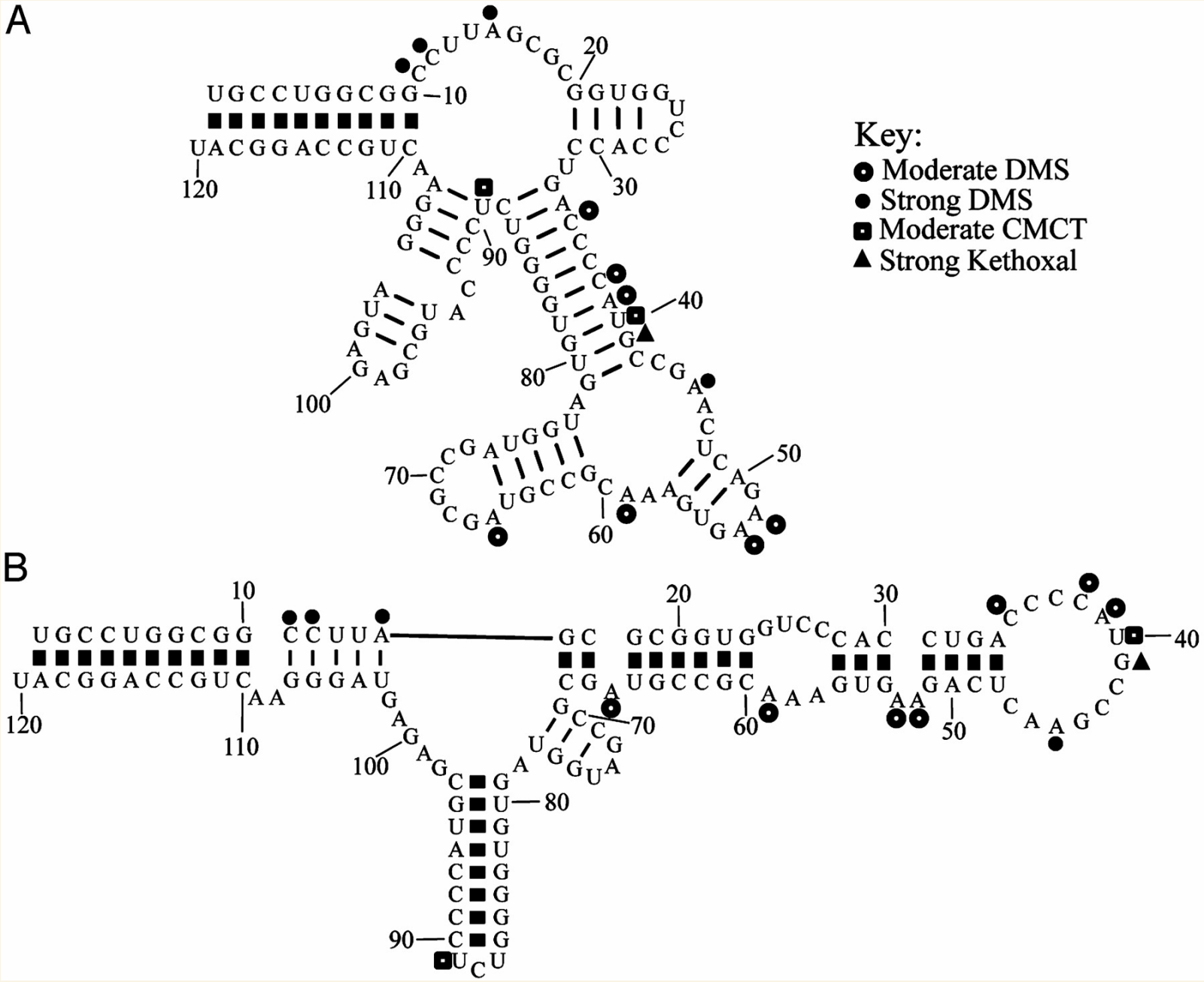

A
Key:
Moderate DMS
Strong DMS
Moderate CMCT
Strong Kethoxal
B

Table 4. The average accuracy of structure prediction with and without constraint with chemical modification data expressed as percentage of known canonical base pairs correctly predicted

| RNA type | Ref(s). | Species | Pseudoknot basepairs, % | Unconstrained | | Constrained | |
|---|---|---|---|---|---|---|---|
| | | | | LFE | Best | LFE | Best |
| Signal recognition particle RNA | 77, 81 | Dog | 0.0 | 18.2 | 97.7 | 84.1* | 98.9* |
| 5S rRNA *in vivo* | 76 | *E. coli* | 0.0 | 26.3 | 86.8 | 86.8 | 97.4 |
| Small subunit rRNA | 25, 78 | *E. coli* | 1.6 | 39.0 | 49.0 | 63.3 | 73.2 |
| RNase P | 32, 80 | *Chromatium vinosum* | 10.5 | 53.5 | 81.6 | 53.5 | 81.6 |
| RNase P | 31, 80 | *Bacillus subtilis* | 7.1 | 56.3 | 70.5 | 56.3 | 68.8 |
| RNase P | 32, 80 | *E. coli* | 9.8 | 58.1† | 73.4 | 64.5† | 74.2 |
| RNase P | 30, 80 | *Saccharomyces cerevisiae* | 7.4 | 59.3 | 78.7 | 58.3 | 78.7 |
| Telomerase RNA *in vivo* | 43, 82, ‡ | *Tetrahymena thermophila* | 10.5 | 65.8 | 84.2 | 65.8 | 84.2 |
| group I bI5 | 78, 89 | *S. cerevisiae* | 5.0 | 78.2 | 83.2 | 81.5 | 83.2 |
| group I Intron *in vivo* | 43, 78 | *T. thermophila* | 4.7 | 83.0 | 90.7 | 83.0 | 90.7 |
| group II Intron aI5c | 27, 79 | Yeast | 0.0 | 86.1 | 89.1 | 77.7 | 82.2 |
| group I Intron L-21 Sca I | 37, 78 | *T. thermophila* | 5.0 | 86.7 | 90.0 | 89.2 | 90.8 |
| 5S rRNA *in vivo* | 76 | *C. albicans* | 0.0 | 90.6† | 90.6 | 90.6† | 90.6 |
| Large subunit rRNA (domain 1) | 26, 78 | *E. coli* | 0.4 | 88.9 | 90.5 | 88.9 | 91.3 |
| group II Intron | 79, 90 | *Pylaiella littoralis* | 0.0 | 90.3† | 94.6 | 90.3† | 94.6 |
| 5S rRNA | 24, 76 | Mouse | 0.0 | 94.4 | 100.0 | 88.9 | 94.4 |
| Average | | | | 67.2 | 84.4 | 76.4 | 85.9 |

Accuracies are reported for both the lowest free energy structure (LFE) and best suboptimal structure in a set of up to 750 structures, generated with a window size of zero.

*Results are reported for protein-bound RNA; when naked RNA chemical modification data are used, the accuracy is 64.8% for the lowest free energy structure and 89.8% for the best suboptimal structure.

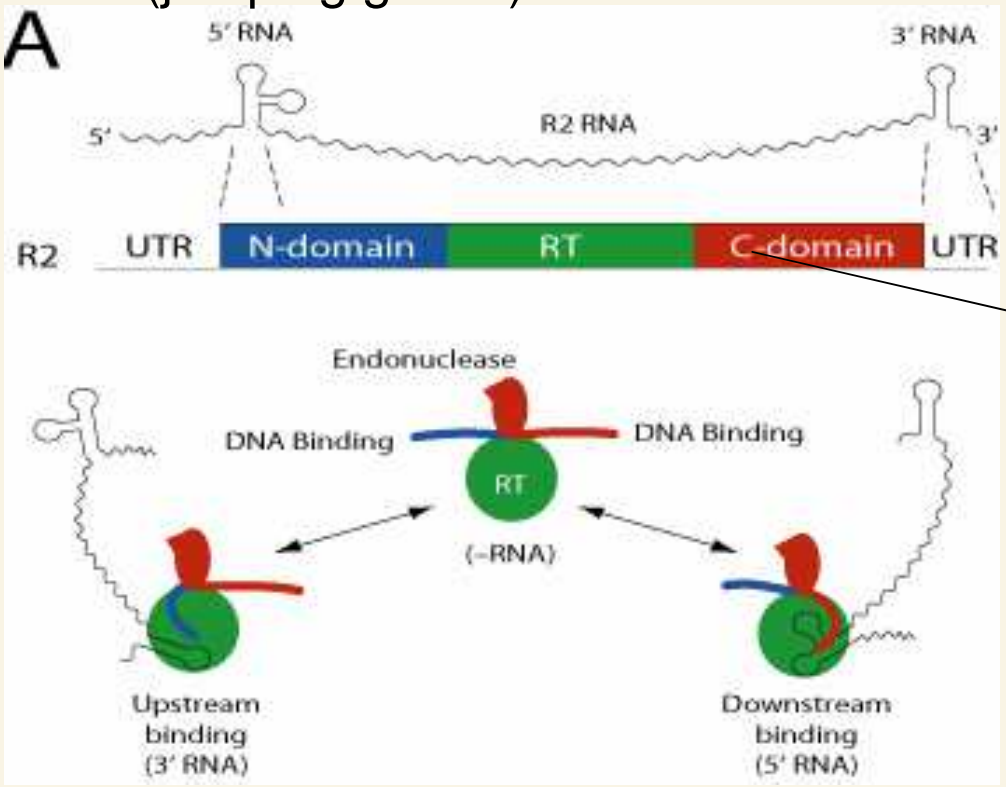
†Best of three or four structures having identical free energies.

‡ten Dam, E., van Belkum, A. & Pleij, K. (1991) *Nucleic Acids Res.* **19**, 6951.

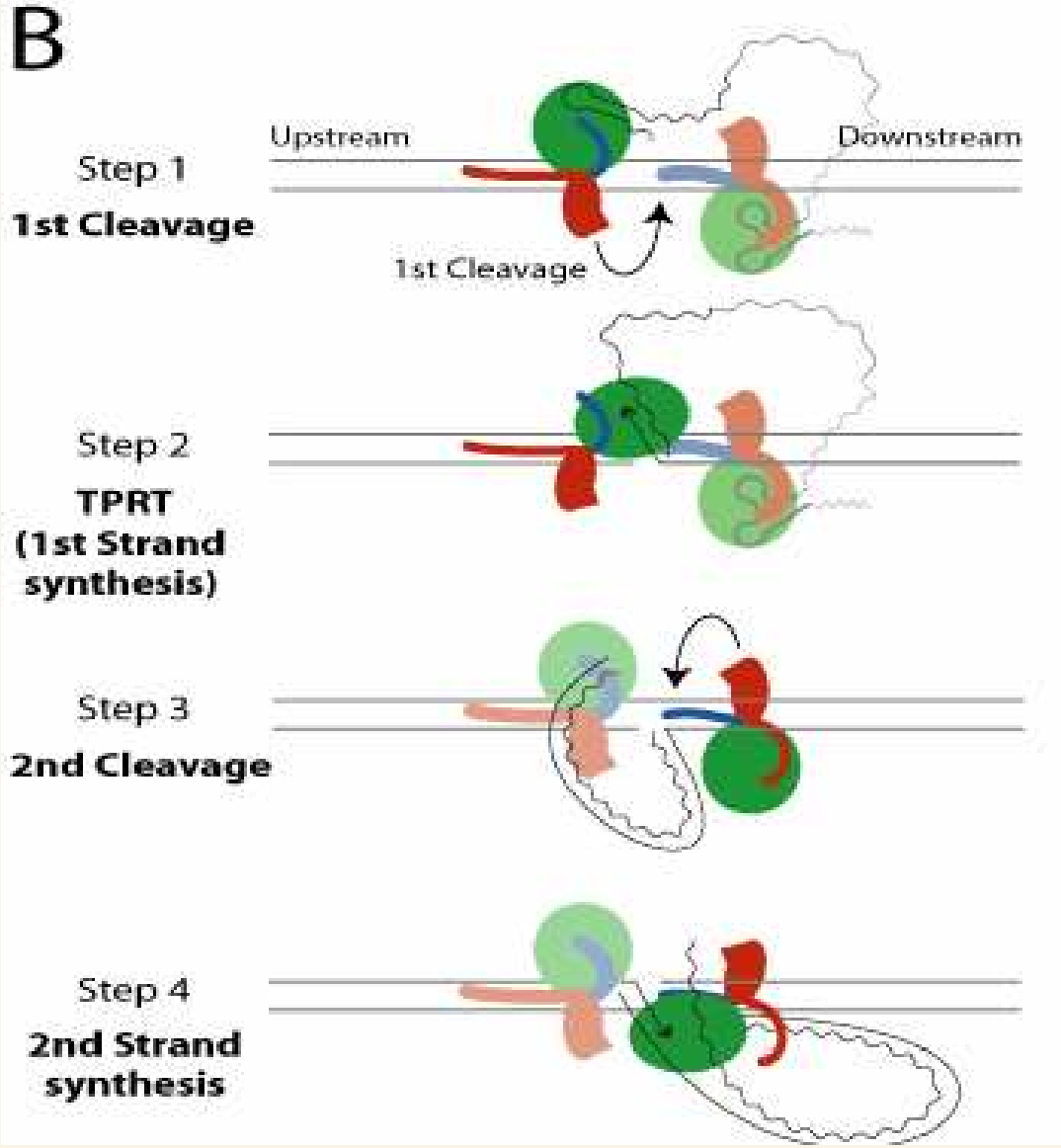
Isoenergetic penta- and hexanucleotide microarray probing and chemical mapping provide a secondary structure model for an RNA element orchestrating R2 retrotransposon protein function
Tuner et al. (2008) in press, *Nucleic Acid Research*

-What is retrotransposition?

Transposons (jumping genes) Barbara McClintock Nobel Prize 1983



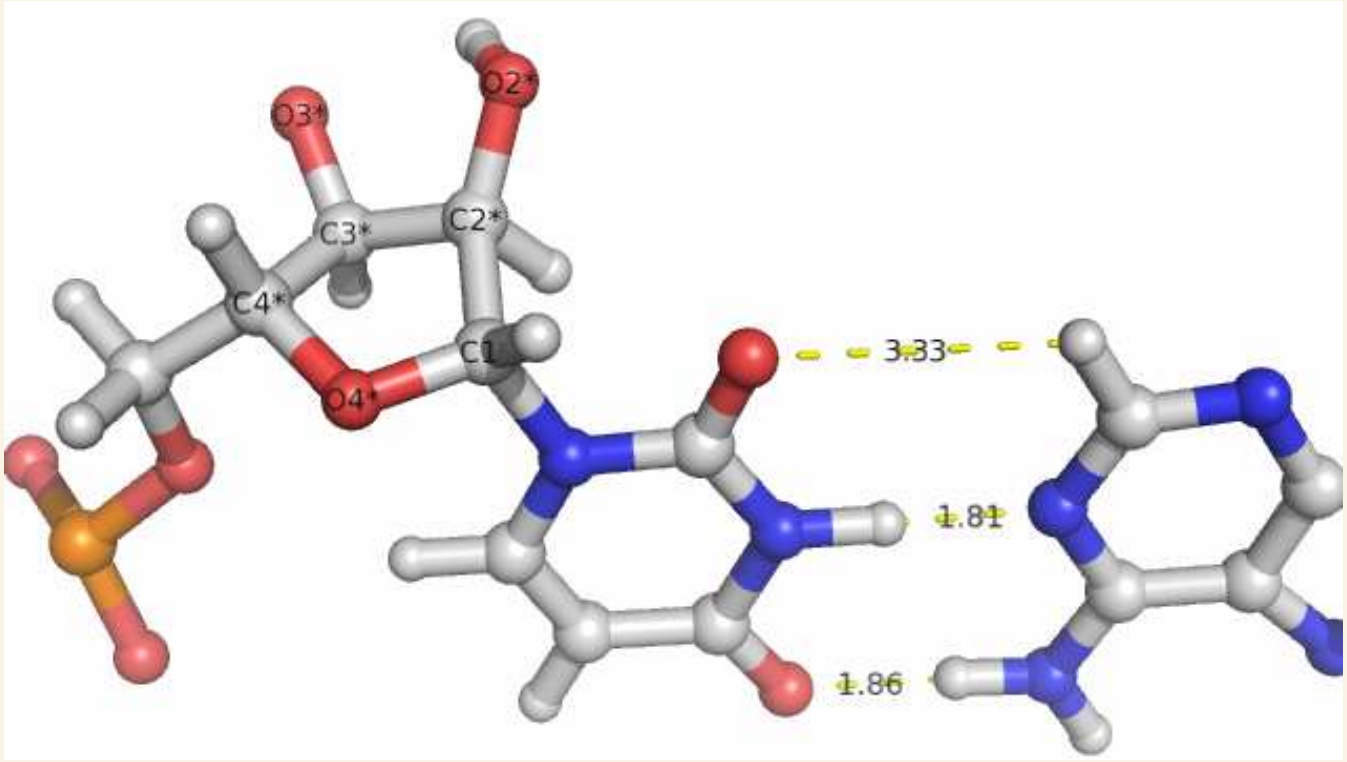

R2 Protein




Danna G. Eickbush, University of Rochester, Biology Department

*Bombyx mori*

- Modified oligonucleotide probes
2'-*O*-methyl RNA-LNA chimeras
LNA = Lock Nucleic Acids = Oligos with methyl bridge between 2'O and 4'C of Sugar.



- What does isoenergetic mean?

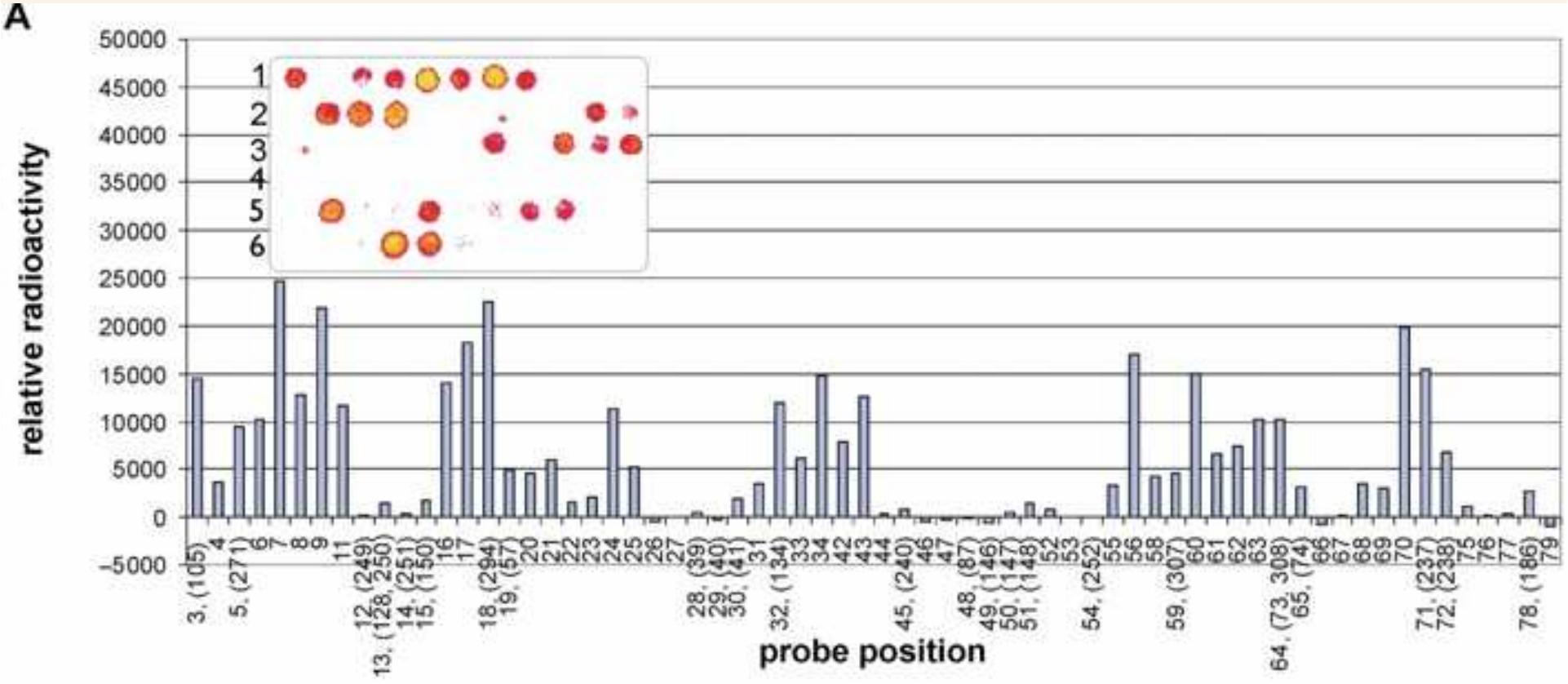
"The modified nucleotides provide similar predicted free energies of binding to a completely unfolded complementary RNA, independent of base content and sequence. Thus, all probes have similar stringency for binding and relative binding to target depends primarily on the free energy required to unfold target."

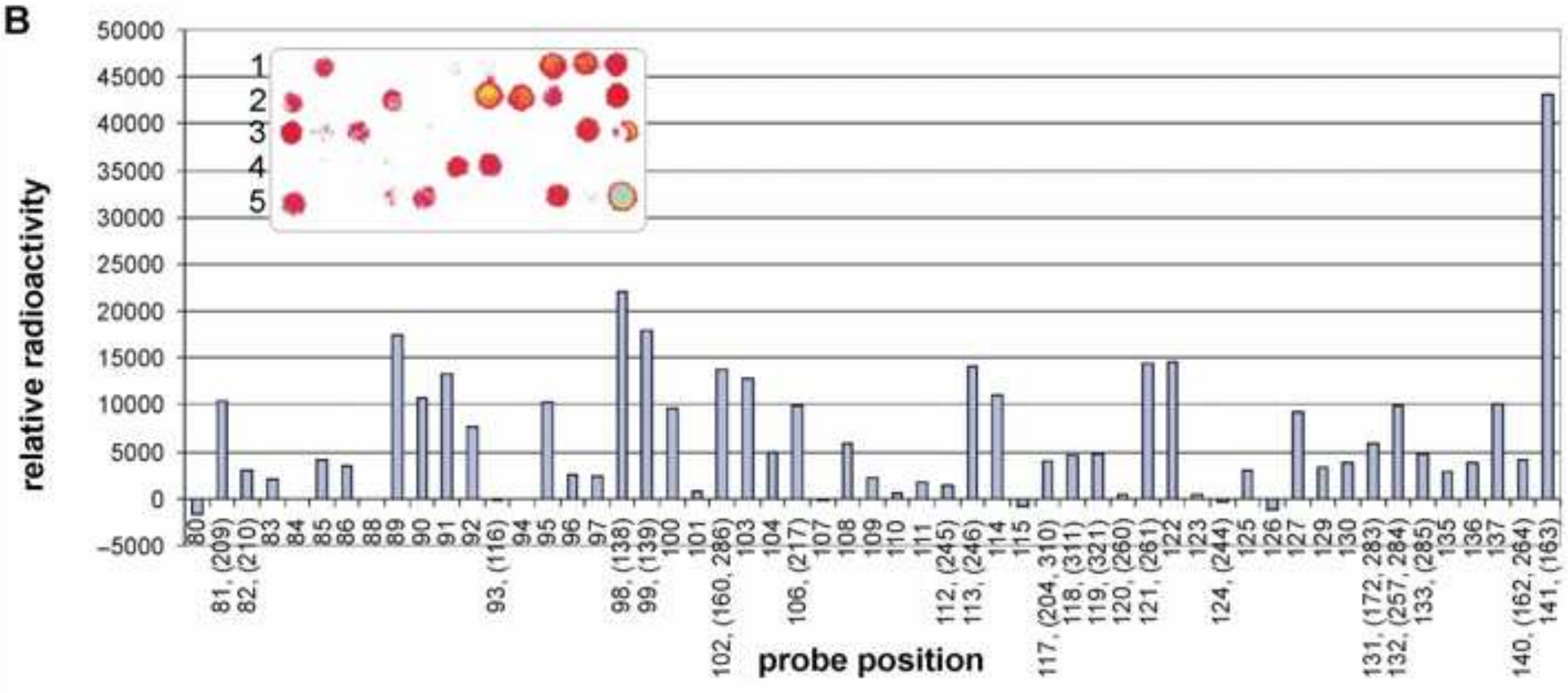
"Most nucleotides not in W-C pairs in structured RNA's are in loops with fewer than seven such contiguous nucleotides. Thus, oligonucleotides shorter than 7-mers should be useful probes of structure.

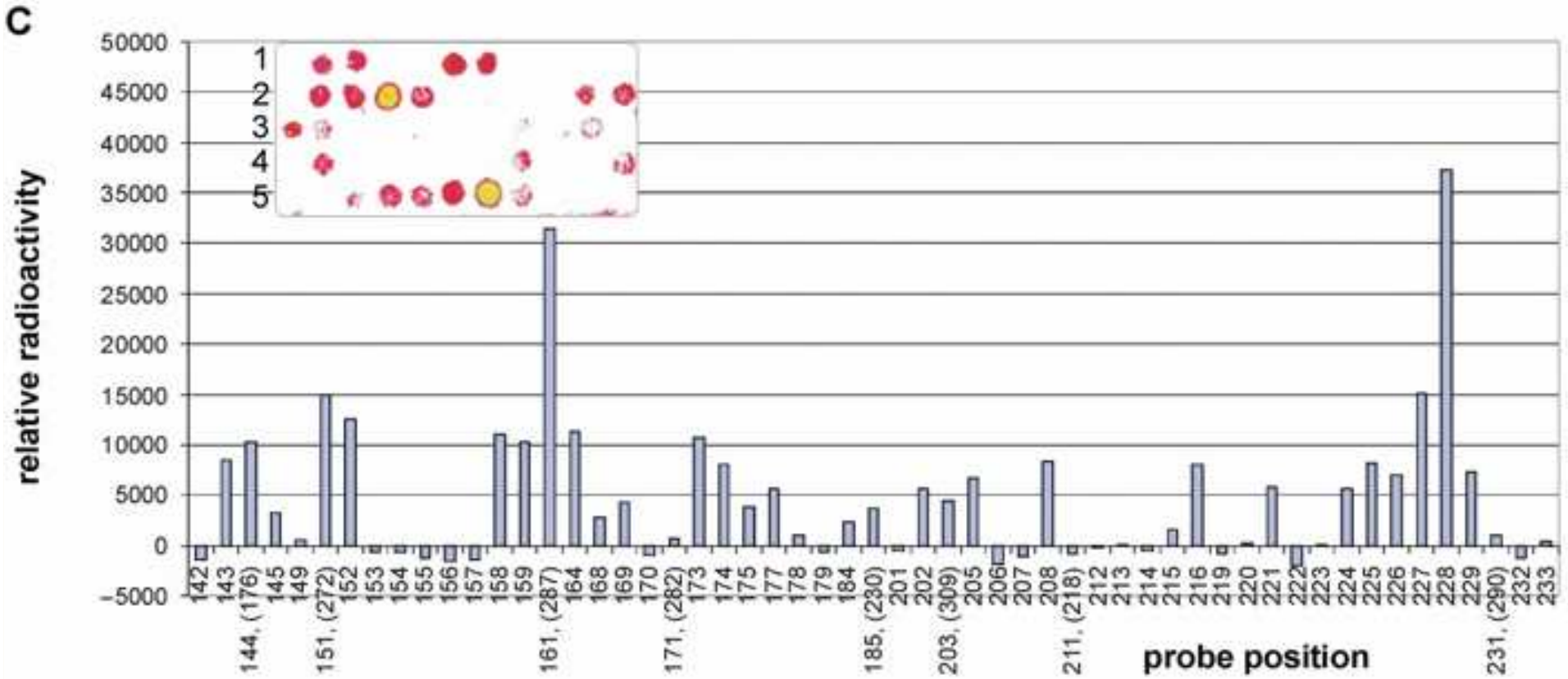
What is a microarray?

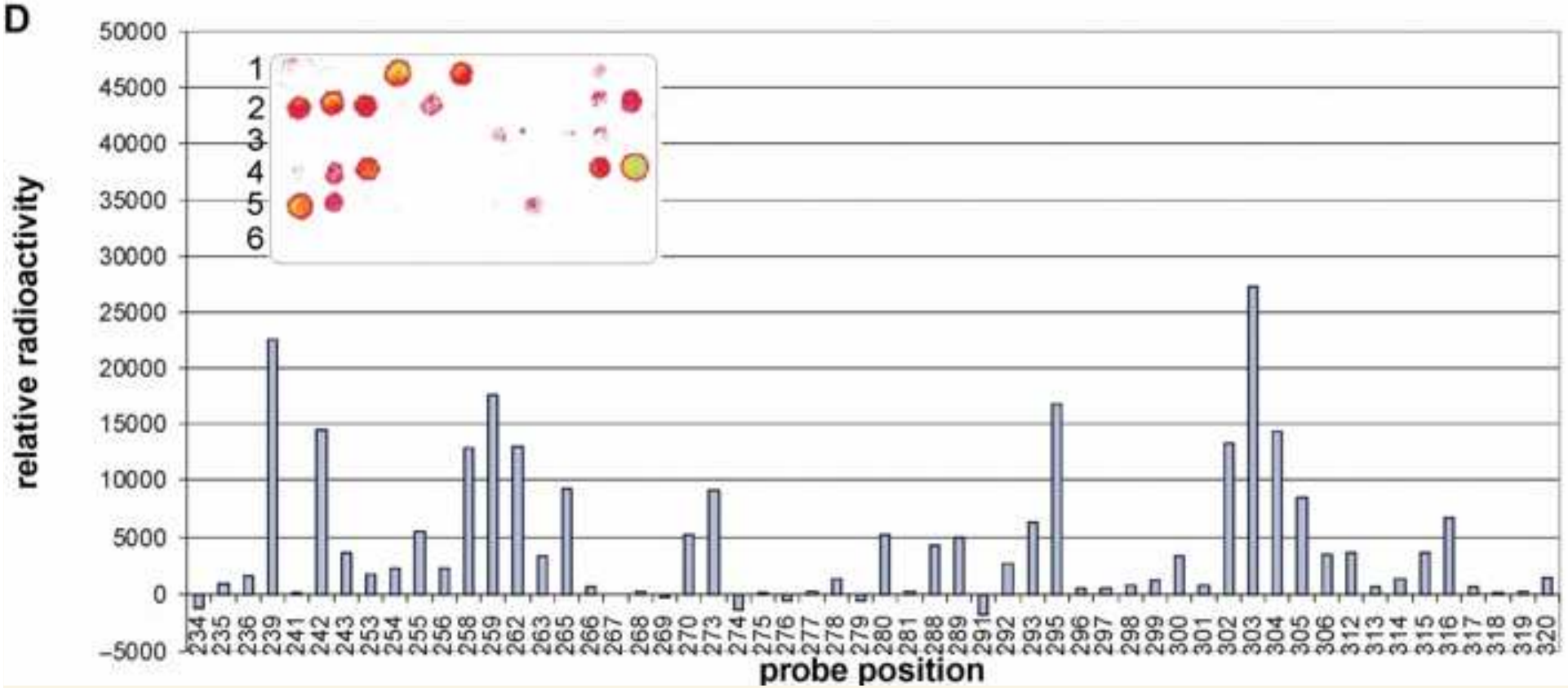
1 - Chimera in microarray slide.

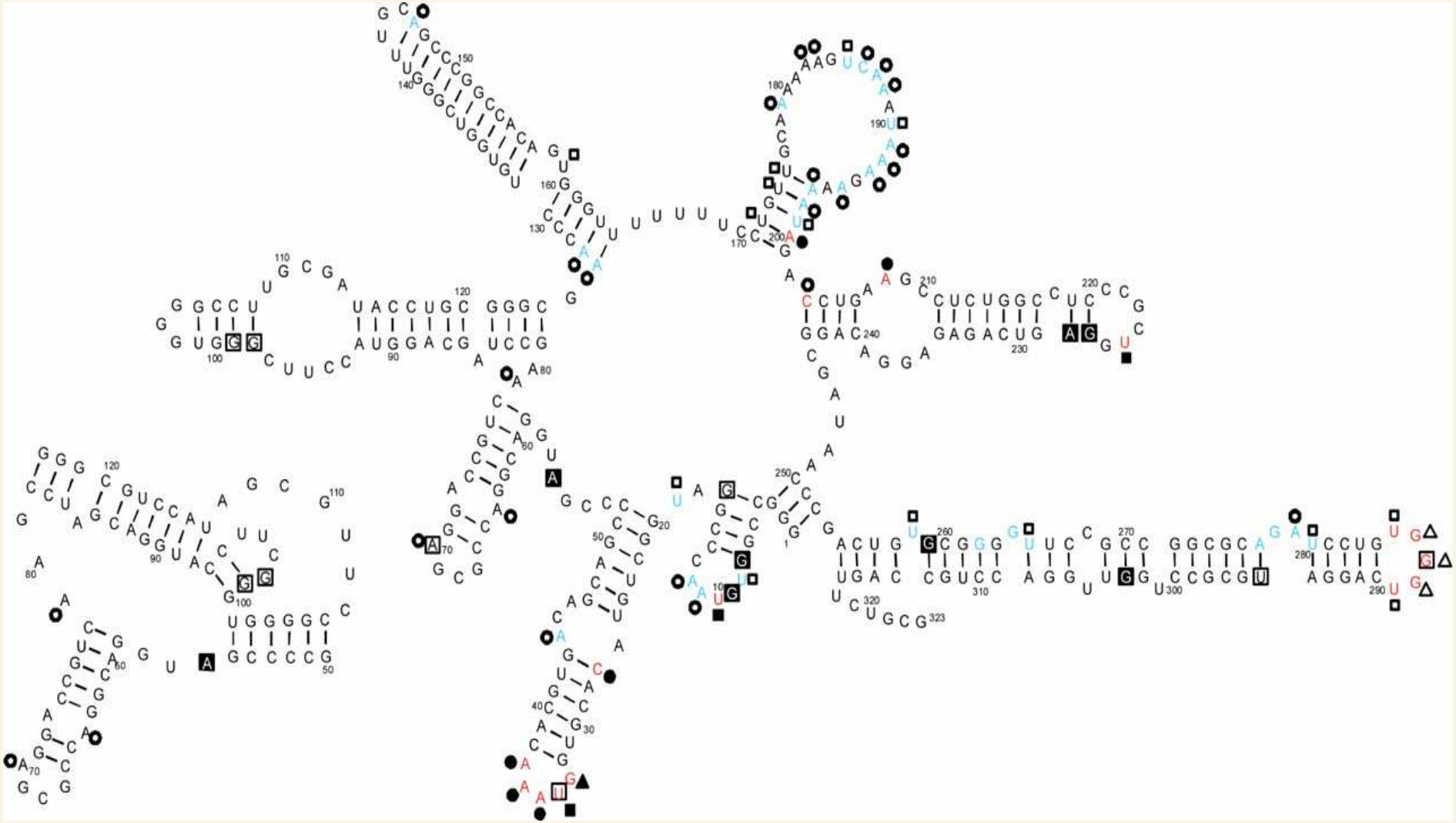
2 – R2Bm 5’ is radiolabeled in phosphates, then hybridized to microarray slide.

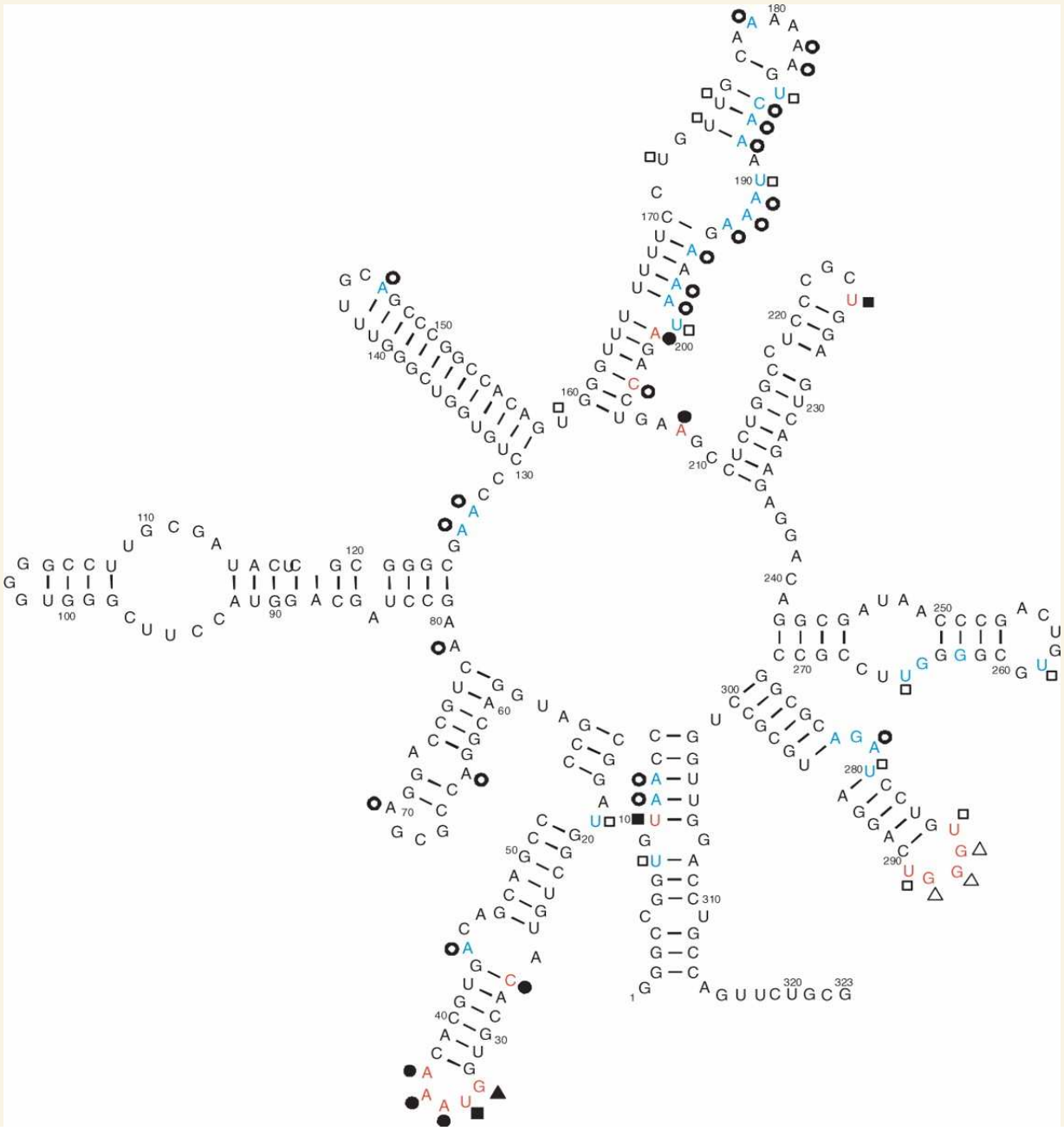



B
relative radioactivity
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
-5000
1
2
3
4
5
80
81, (209)
82, (210)
83
84
85
86
88
89
90
91
92
93, (116)
94
95
96
97
98, (138)
99, (139)
100
101
102, (160, 286)
103
104
106, (217)
107
108
109
110
111
112, (245)
113, (246)
114
115
117, (204, 310)
118, (311)
119, (321)
120, (260)
121, (261)
122
123
124, (244)
125
126
127
129
130
131, (172, 283)
132, (257, 284)
133, (285)
135
136
137
140, (162, 264)
141, (163)
probe position


C
relative radioactivity
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
-5000
1
2
3
4
5
142
143
144, (176)
145
149
151, (272)
152
153
154
155
156
157
158
159
161, (287)
164
168
169
170
171, (282)
173
174
175
177
178
179
184
185, (230)
201
202
203, (309)
205
206
207
208
211, (218)
212
213
214
215
216
219
220
221
222
223
224
225
226
227
228
229
231, (290)
232
233
probe position


D
relative radioactivity
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
−5000
1
2
3
4
5
6
234 235 236 239 241 242 243 253 254 255 256 258 259 262 263 265 266 267 268 269 270 273 274 275 276 277 278 279 280 281 288 289 291 292 293 295 296 297 298 299 300 301 302 303 304 305 306 312 313 314 315 316 317 318 319 320
probe position



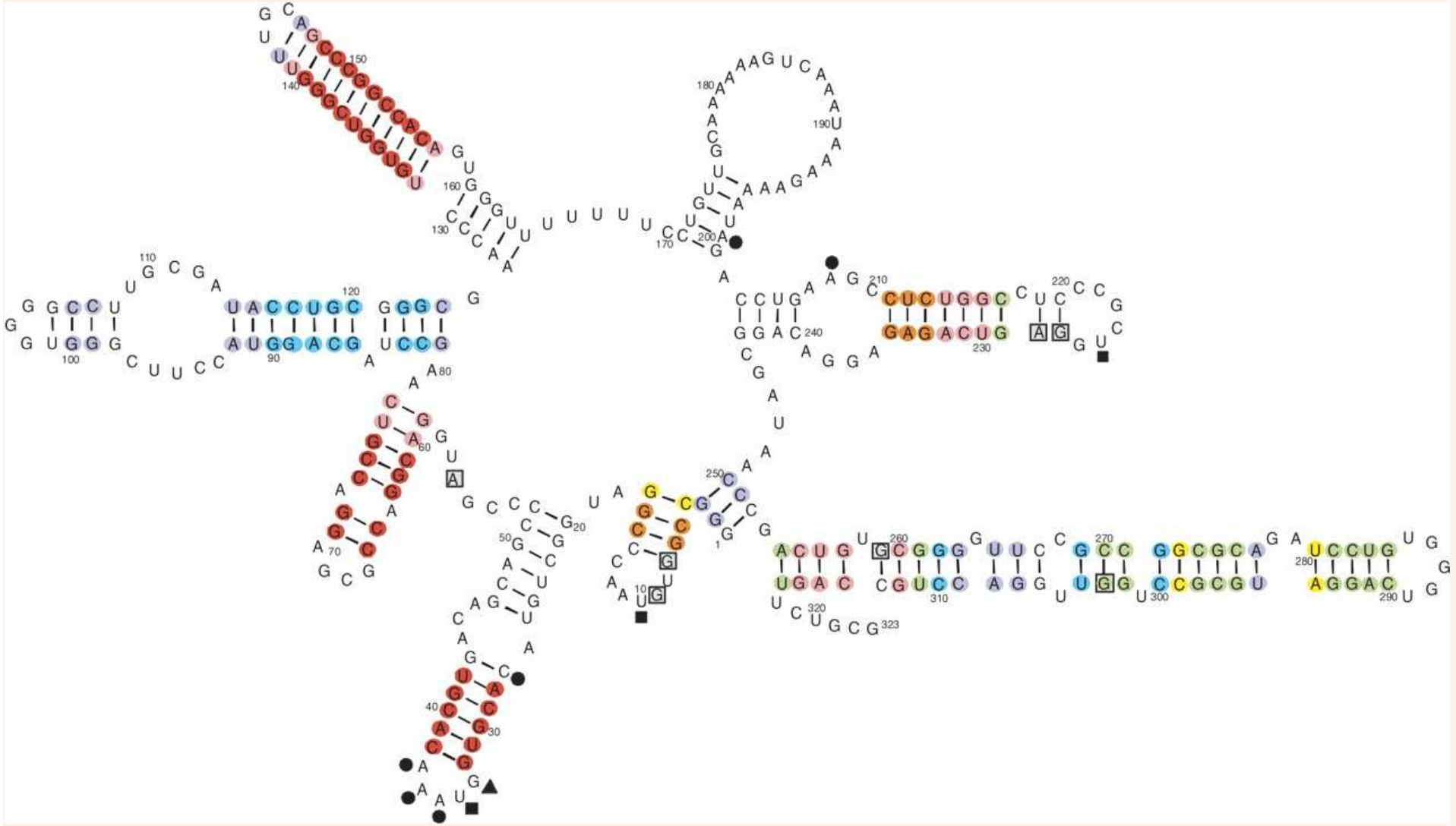


$$Q = \sum_{i=1}^{N} e^{\Delta G^o / RT}$$

$N$ = all possible secondary structures

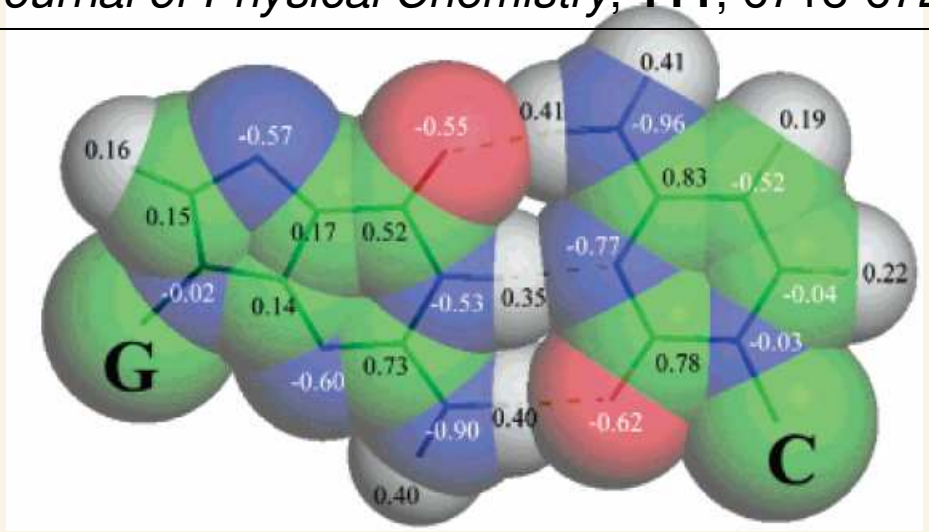
$$P_{\text{b-p}} = \frac{\sum_{i=1}^{n} e^{-\Delta G^o / RT}}{Q}$$

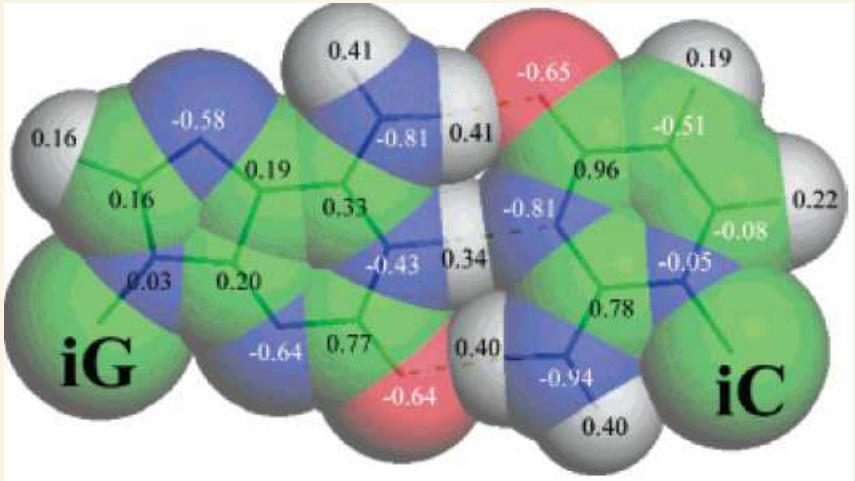
$n$ = All secondary structures with b - p

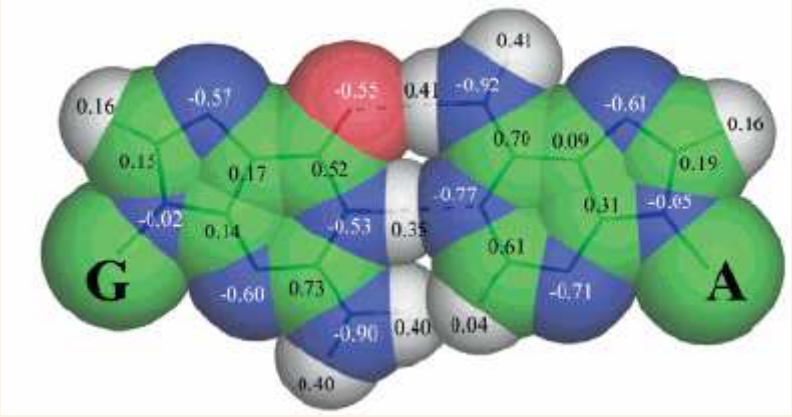


Stacking Effects on Local Structure in RNA: Changes in the Structure of Tandem GA Pairs when Flanking GC Pairs are Replaced by isoG-isoC Pairs.

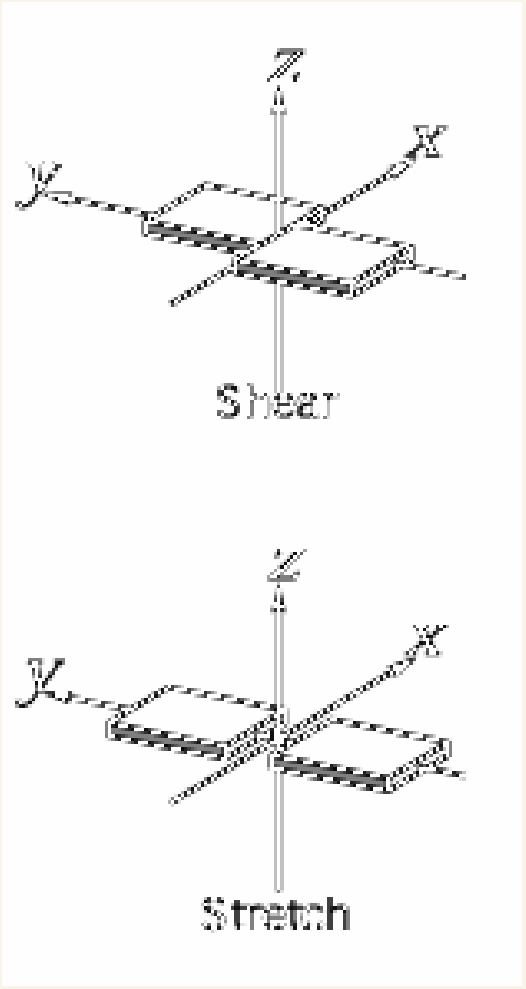
Tuner et al. (2007) *Journal of Physical Chemistry*, **111**, 6718-6727.

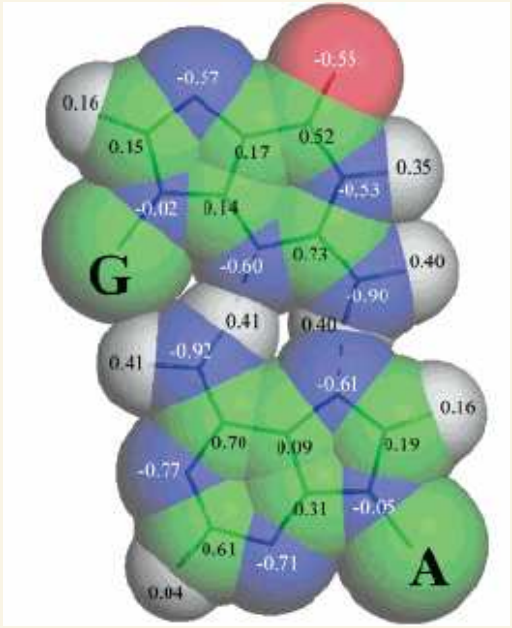


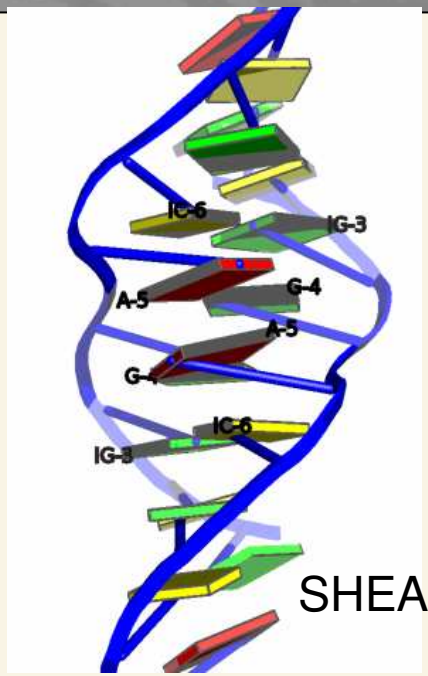



IMINO

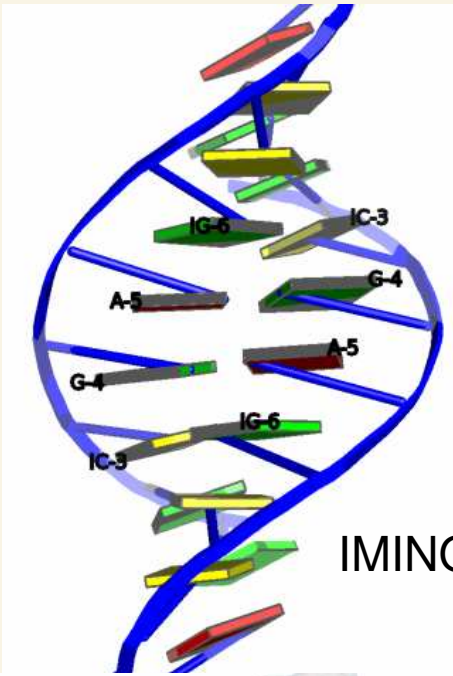



SHEARED

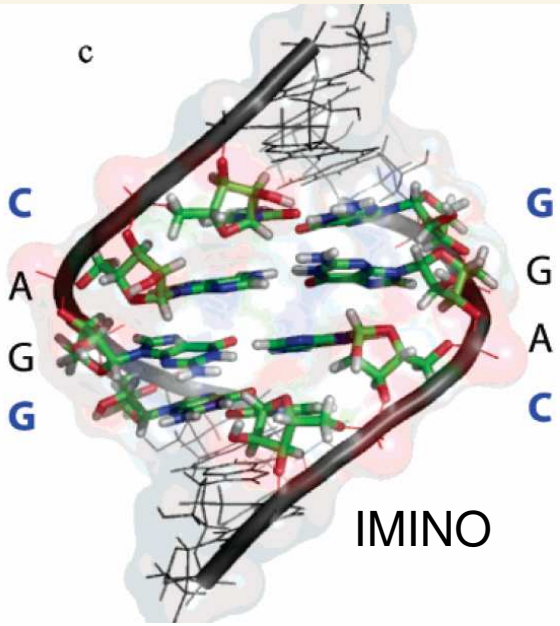



SHEARED

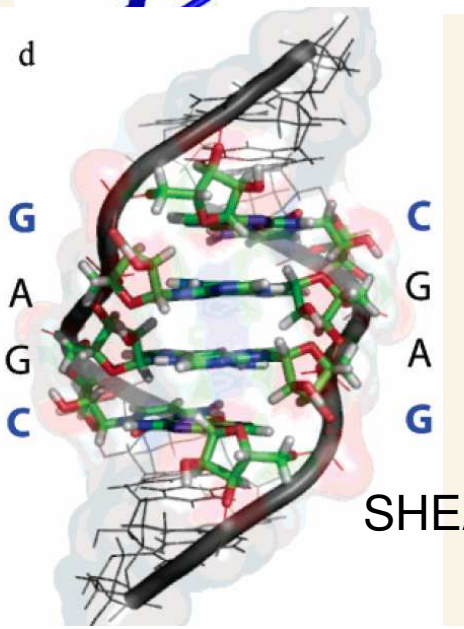



IMINO

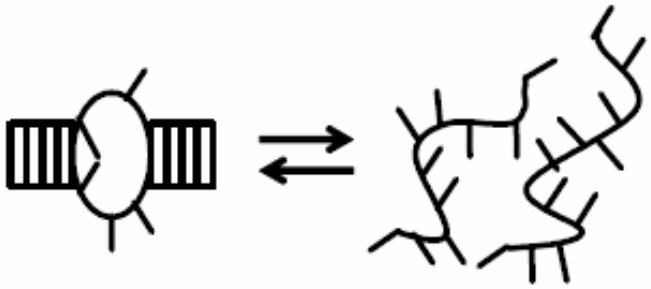
The most dramatic result from this study is that substituting iGiC pairs for GC pairs adjacent to tandem GA pairs changes the shape of the GA pairs.




IMINO




SHEARED



$$\frac{1}{T_m} = \frac{R}{\Delta H^\circ} \ln\left(\frac{C_T}{a}\right) + \frac{\Delta S^\circ}{\Delta H^\circ}$$