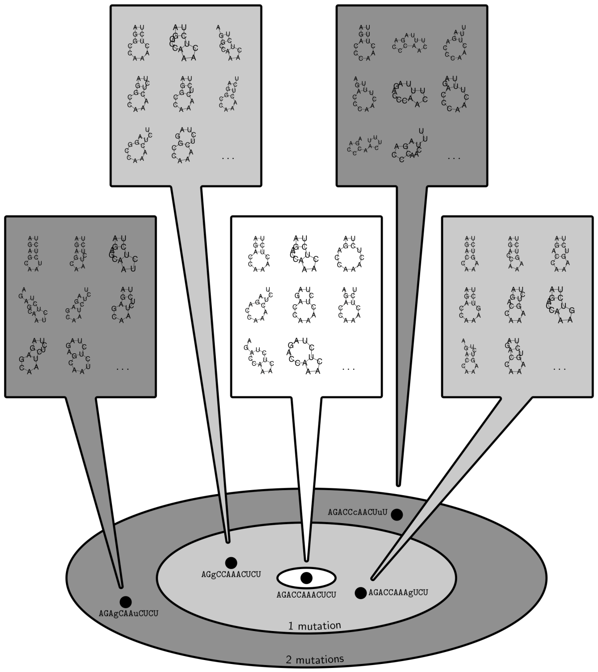


Algorithms for Exploring the Mutation Landscape of RNA molecules

Jérôme Waldispühl, Srinivas Devadas, Bonnie Berger
MIT

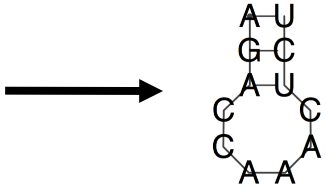
Peter Clote
Boston College

RNA 2008, Berlin

RNA secondary structure prediction

Energy minimization (Zuker&Stiegler,1981).

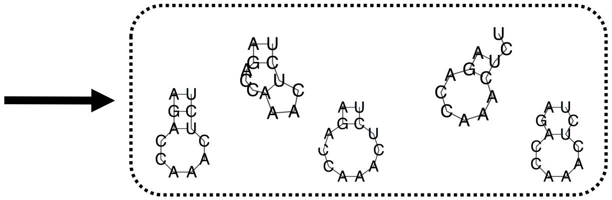
UCUCAAACCAGA



- ab-initio prediction (does not require homolog),
- single structure prediction.

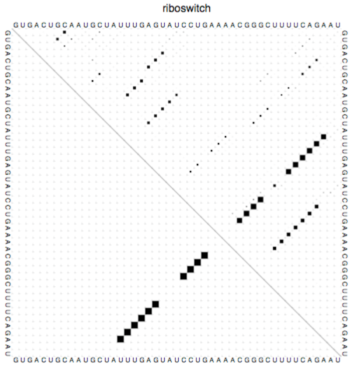
Partition function and base pair binding probabilities (McCaskill,1990).

UCUCAAACCAGA



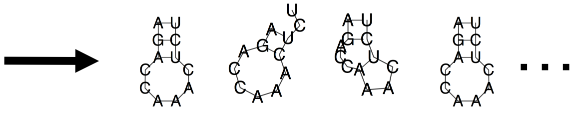
- evaluate the ensemble of all potential structures,
- compute the base pair binding probability.

$$Z = \sum_j e^{-\beta \cdot E_j}$$



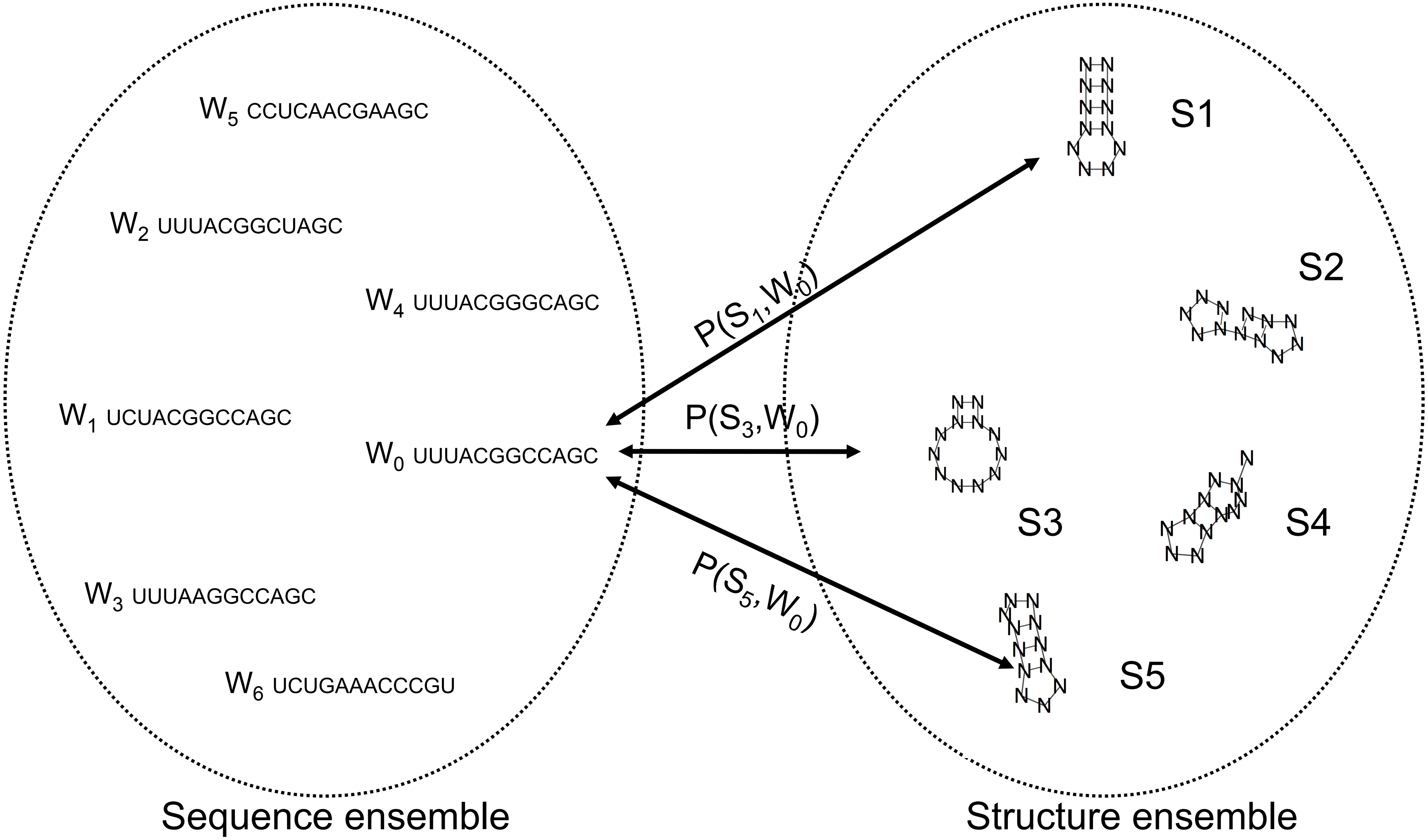
Rational sampling (Ding&Lawrence,2001).

UCUCAAACCAGA



- sample structure from low energy ensemble,
- allow to estimate complex features,
- fast and simple to implement.

Sequence-structure maps



W_5 CCUCAACGAAGC

W_2 UUUACGGCUAGC

W_4 UUUACGGGCAGC

W_1 UCUACGGCCAGC

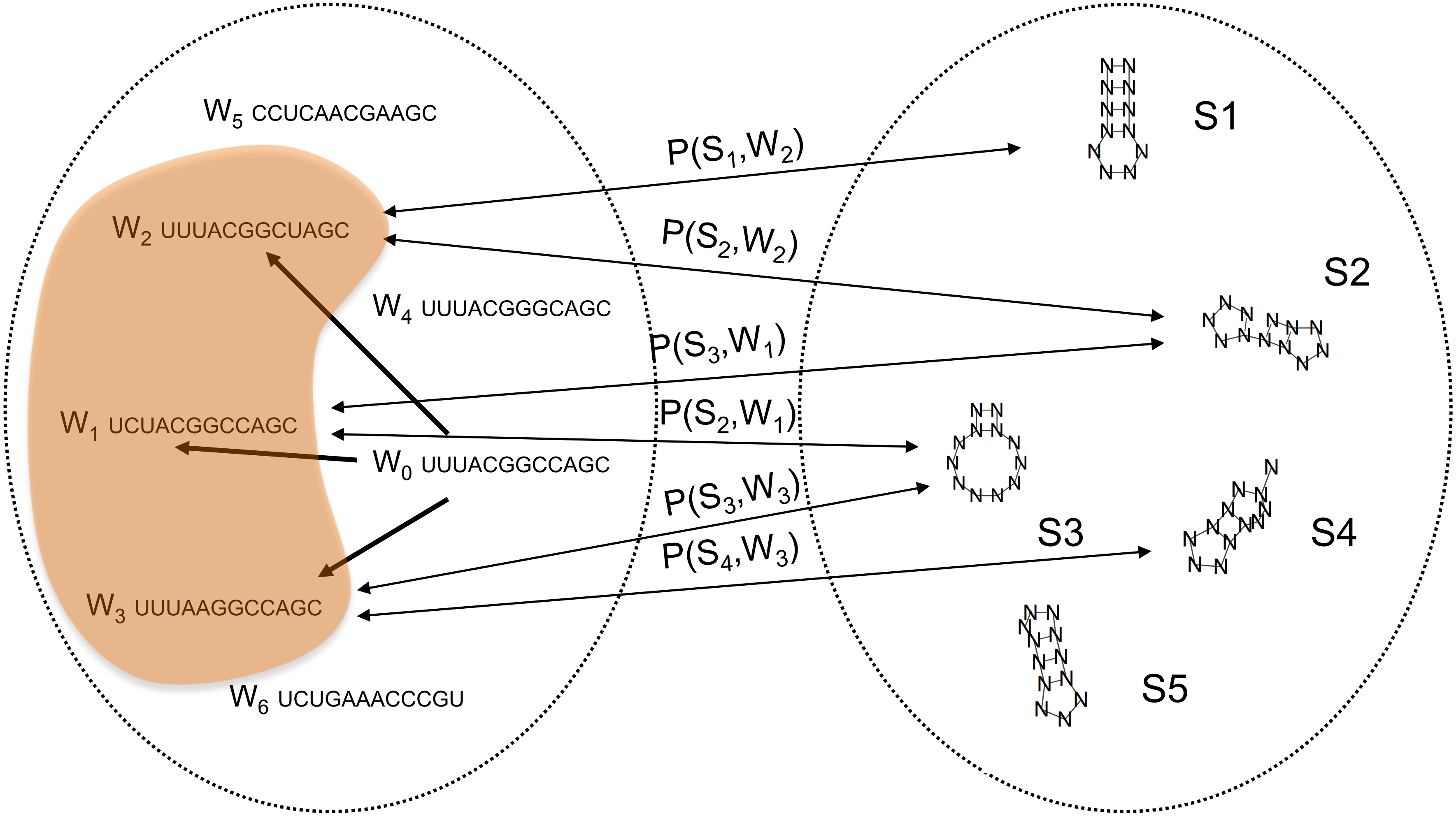
W_0 UUUACGGCCAGC

W_3 UUUAAGGCCAGC

W_6 UCUGAAACCCGU

- Pioneered by P. Schuster’s group in Vienna

Sequence-structure maps



W_5 CCUCAACGAAGC

W_2 UUUACGGCUAGC

W_4 UUUACGGGCAGC

W_1 UCUACGGCCAGC

W_0 UUUACGGCCAGC

W_3 UUUAAGGCCAGC

W_6 UCUGAAACCCGU

$P(S_1,W_2)$

$P(S_2,W_2)$

$P(S_3,W_1)$

$P(S_2,W_1)$

$P(S_3,W_3)$

$P(S_4,W_3)$

S1

S2

S3

S4

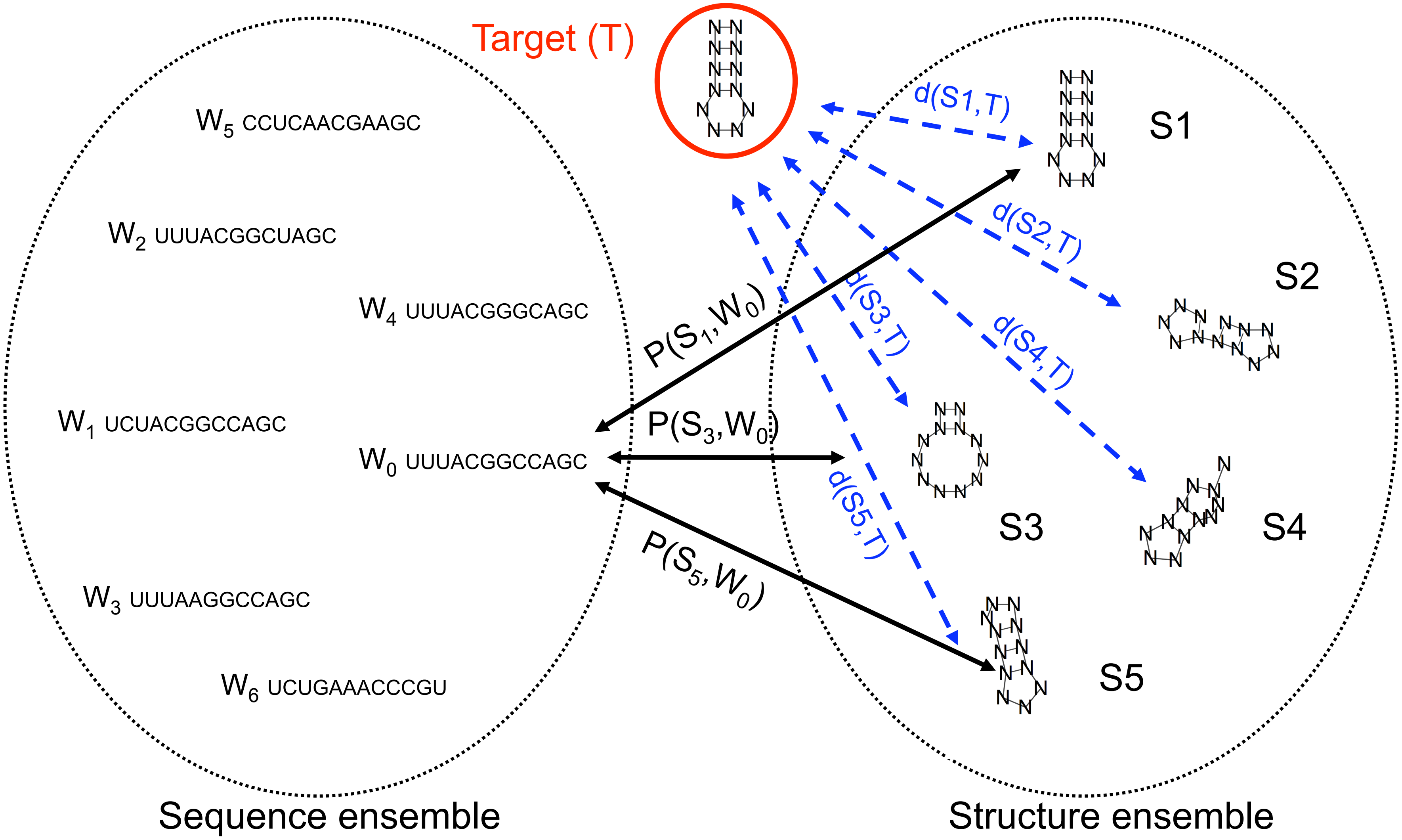
S5

Sequence ensemble

Structure ensemble

- Pioneered by P. Schuster’s group in Vienna

Target fitting in sequence-structure maps



Target (T)

W_5 CCUCAACGAAGC

W_2 UUUACGGCUAGC

W_4 UUUACGGGCAGC

W_1 UCUACGGCCAGC

W_0 UUUACGGCCAGC

W_3 UUUAAGGCCAGC

W_6 UCUGAAACCCGU

$P(S_1,W_0)$

$P(S_3,W_0)$

$P(S_5,W_0)$

d(S1,T)

d(S2,T)

d(S3,T)

d(S4,T)

d(S5,T)

S1

S2

S3

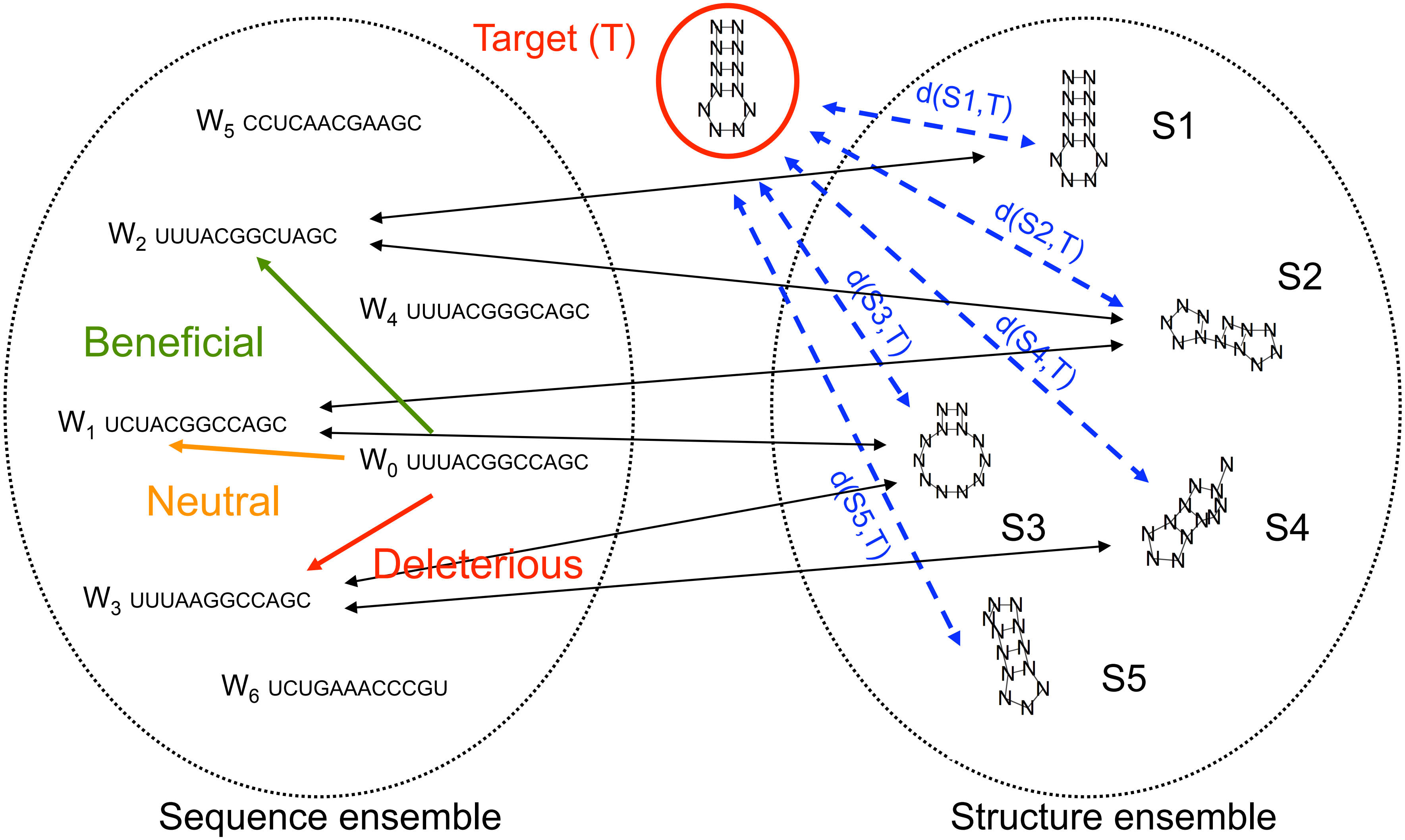
S4

S5

Sequence ensemble

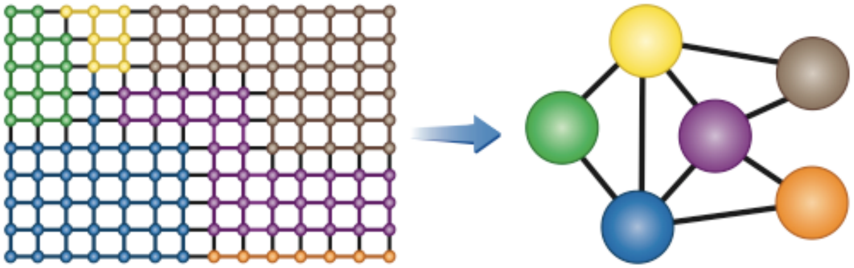
Structure ensemble

Target fitting in sequence-structure maps



Previous works

- **Neutral Network**
  (P. Shuster, W. Fontana, P. Stadler, I. Hofacker, R. Giergerich, …)



(Cowperthwaite&Meyers,2007)

- **Mutational robustness and thermostability**
  (L. Ancel-Meyers et al.)

- **Prediction of Deleterious Mutations**
  (D. Barash, W. Shu et al.)

Limited by the use of brute force algorithms

Our approach

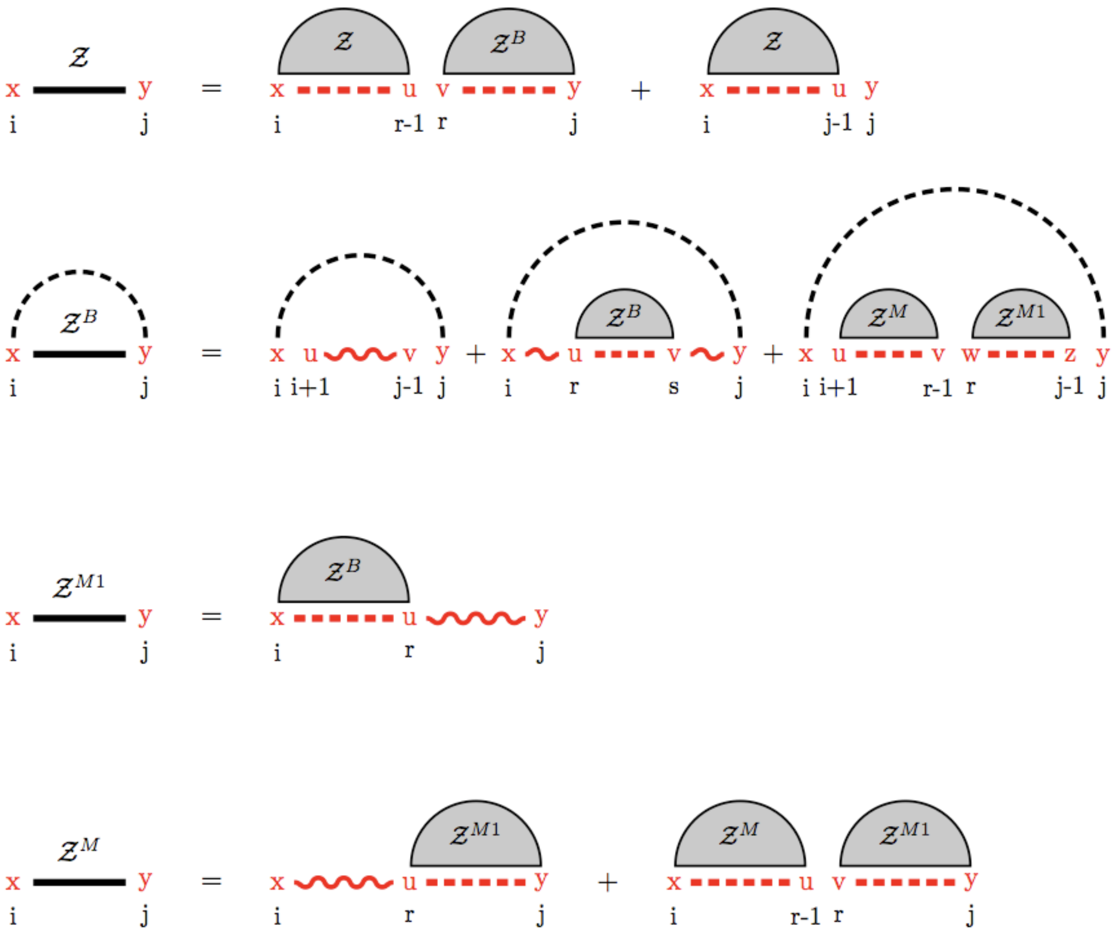
RNAmutants

- *polynomial-time* algorithm,
- explore *exhaustively* the mutation landscape.

Results:

- Prediction of deleterious mutations
- Analysis of mutational robustness and thermostability.

Generalization of McCaskill algorithm



Store annotation of boundary nucleotides (nearest neighbors) in dynamic arrays.

Sampling RNA mutants (sequence and secondary structure)

- Use a stochastic backtracking procedure
- **Novelty:** Sample sequences with mutation (mutants) AND a secondary structure

Examples:

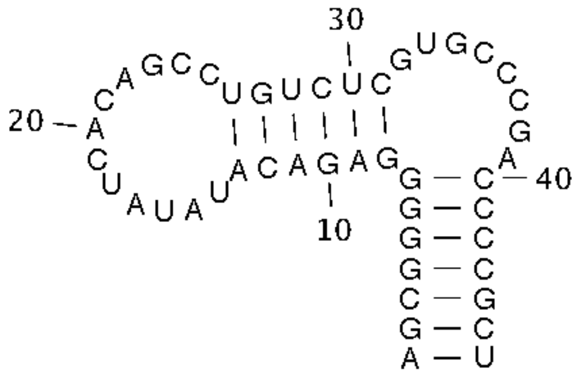
UACAGCGGGGGgGgCAUgUAUCACAGCCUGUgUCGUGCCCGcCCCCGCUGG

..((((((((((((((((...(((.....))).))))))).))))))))).

UACAGCGGGGGgGgCAUAgAUCgCAGCCgGUCUCGUGCCCGACCCCGCUGG

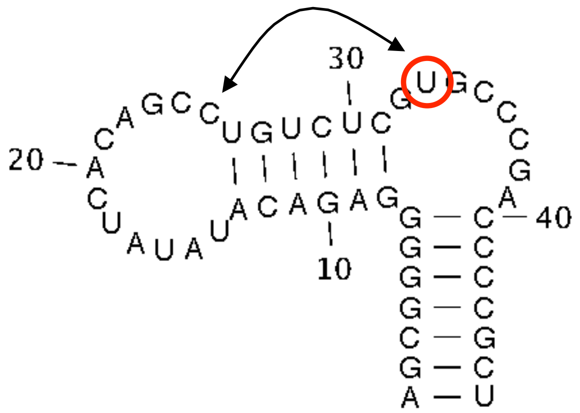
..((((((((.(((((((((((......))))).))))))..)))))))).

Predicting deleterious mutations: Application to the Hepatitis C cis-regulatoring element



```
         10        20        30        40
         |         |         |         |
AGCGGGGGAGACAUAUAUCACAGCCUGUCUCGUGCCCGACCCCGCU
(((((((((((((............)))))).........)))))))   index   mutation   basepair   frequency   break
........................(.......).............     33      U→G        25,C       0.003        6
..........(.......................)...........     11      A→G        35,C       0.396        4
..........(........................)..........     11      A→G        36,C       0.065        4
..........(.......................)...........     35      C→U        11,A       0.013        4
..........(........................)..........     36      C→U        11,A       0.002        4
........(..........................)..........      9      A→G        36,C       0.012        2
..................(.........).................     19      C→G        29,C       0.010        2
..................(.........).................     29      C→G        19,C       0.006        2
...............(............).................     16      U→G        29,C       0.004        2
...........(................).................     12      C→G        29,C       0.001        2
..............(...............)...............     15      A→G        31,C       0.006        1
.......(..............................).......     39      A→C         8,G       0.005        1
......(........................)..............     32      G→C         7,G       0.001        1
.............(..........).....................     14      U→G        25,C       0.247        0
..............(........)......................     15      A→G        24,G       0.088        0
.............(..........).....................     25      C→G        14,U       0.015        0
.............(..........).....................     25      C→A        14,U       0.009        0
..............(........)......................     24      C→U        15,A       0.005        0
..............(.....).........................     15      A→G        21,C       0.004        0
...............(......).......................     16      U→C        23,G       0.002        0
................(......)......................     17      A→G        24,C       0.001        0
```

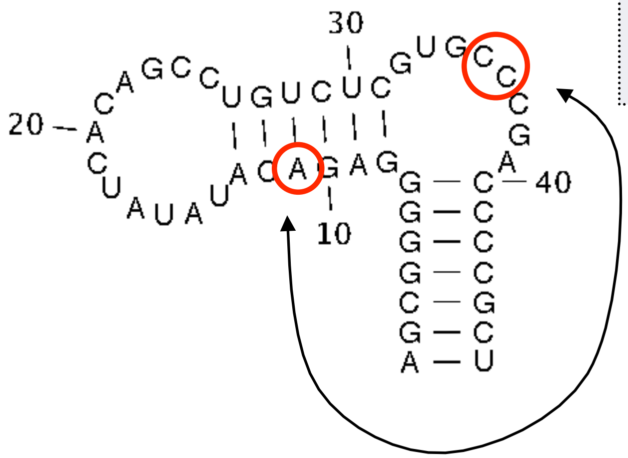
Predicting deleterious mutations: Application to the Hepatitis C cis-regulatoring element



```
        10        20        30        40
         |         |         |         |
AGCGGGGGAGACAUAUAUCACAGCCUGUCUCGUGCCCGACCCCGCU
(((((((((((((............))))))........)))))))  index  mutation  basepair  frequency  break
........................(.......).............   33     U→G       25,C      0.003      6
..........(.......................)...........   11     A→G       35,C      0.396      4
..........(........................)..........   11     A→G       36,C      0.065      4
..........(.......................)...........   35     C→U       11,A      0.013      4
..........(........................)..........   36     C→U       11,A      0.002      4
........(..........................)..........    9     A→G       36,C      0.012      2
..................(.........).................   19     C→G       29,C      0.010      2
..................(.........).................   29     C→G       19,C      0.006      2
...............(............).................   16     U→G       29,C      0.004      2
...........(................).................   12     C→G       29,C      0.001      2
..............(...............)...............   15     A→G       31,C      0.006      1
.......(..............................).......   39     A→C        8,G      0.005      1
......(........................)..............   32     G→C        7,G      0.001      1
.............(..........).....................   14     U→G       25,C      0.247      0
..............(........)......................   15     A→G       24,G      0.088      0
.............(..........).....................   25     C→G       14,U      0.015      0
.............(..........).....................   25     C→A       14,U      0.009      0
..............(........)......................   24     C→U       15,A      0.005      0
..............(.....).........................   15     A→G       21,C      0.004      0
...............(......).......................   16     U→C       23,G      0.002      0
................(......)......................   17     A→G       24,C      0.001      0
```

- In vitro: C31A/U33A is deleterious, but not U33G alone nor C31A/U33A.
- Computational experiments with RNAfold reveal that discrepancy is due to dangle.

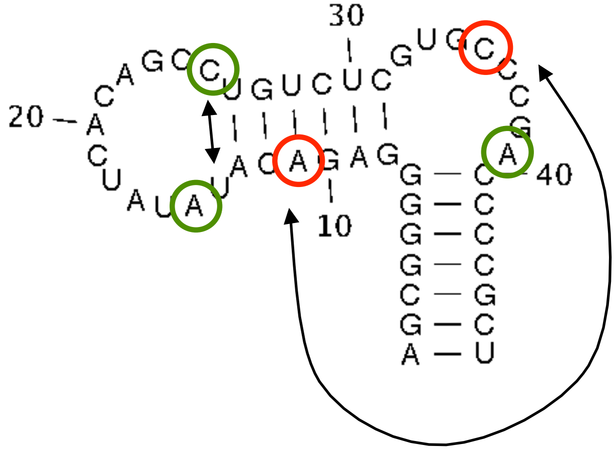
Predicting deleterious mutations: Application to the Hepatitis C cis-regulatoring element



```
         10        20        30        40
         |         |         |         |
AGCGGGGGAGACAUAUAUCACAGCCUGUCUCGUGCCCGACCCCGCU
(((((((((((((............))))))........)))))))   index  mutation  basepair  frequency  break
........................(.......).............     33    U→G       25,C      0.003      6
..........(.......................)...........     11    A→G       35,C      0.396      4
..........(........................)..........     11    A→G       36,C      0.065      4
..........(.......................)...........     35    C→U       11,A      0.013      4
..........(........................)..........     36    C→U       11,A      0.002      4
........(..........................)..........      9    A→G       36,C      0.012      2
..................(.........).................     19    C→G       29,C      0.010      2
..................(.........).................     29    C→G       19,C      0.006      2
...............(............).................     16    U→G       29,C      0.004      2
...........(................).................     12    C→G       29,C      0.001      2
..............(...............)...............     15    A→G       31,C      0.006      1
.......(..............................).......     39    A→C        8,G      0.005      1
......(........................)..............     32    G→C        7,G      0.001      1
.............(..........).....................     14    U→G       25,C      0.247      0
..............(........)......................     15    A→G       24,G      0.088      0
.............(..........).....................     25    C→G       14,U      0.015      0
.............(..........).....................     25    C→A       14,U      0.009      0
..............(........)......................     24    C→U       15,A      0.005      0
..............(.....).........................     15    A→G       21,C      0.004      0
...............(......).......................     16    U→C       23,G      0.002      0
................(......)......................     17    A→G       24,C      0.001      0
```

- In Rfam, 30 sequences with C36U and 3 with C35U, but all along with other mutations. These simultaneous mutations seem to compensate the deleterious effect.
- One occurrence of A11G (alone), but experimental data say nothing about the viability of the mutant.

Predicting deleterious mutations: Application to the Hepatitis C cis-regulatoring element

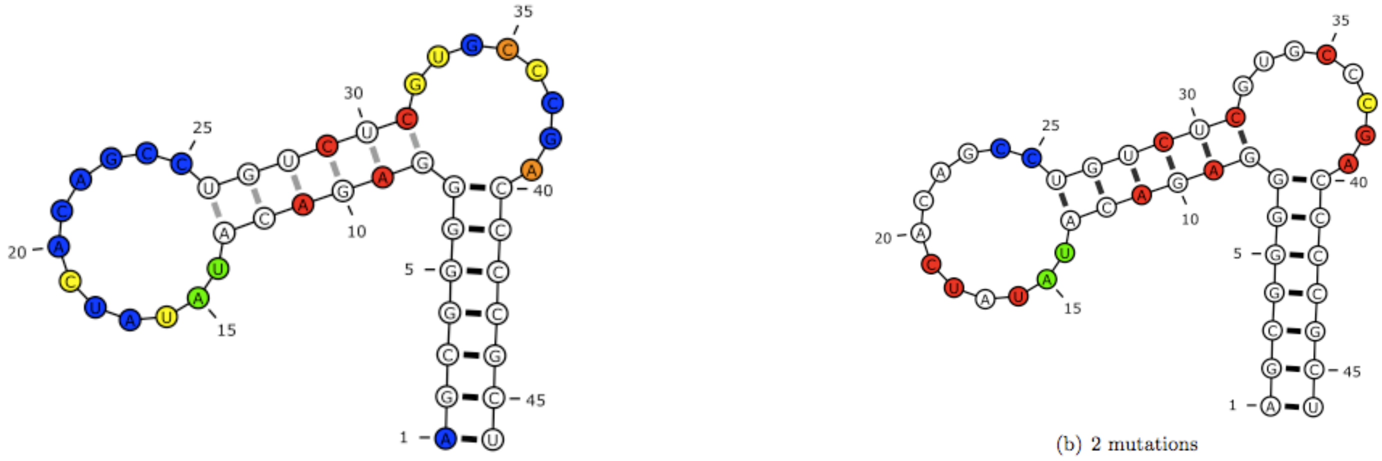


```
    10        20        30        40
     |         |         |         |
AGCGGGGGAGACAUAUAUCACAGCCUGUCUCGUGCCCGACCCCGCU
```

`(((((((((((((............))))))........)))))))`	index	mutation	basepair	frequency	break
`........................(.......).............`	33	U→G	25,C	0.003	6
`..........(.......................)...........`	11	A→G	35,C	0.396	4
`..........(........................)..........`	11	A→G	36,C	0.065	4
`..........(.......................)...........`	35	C→U	11,A	0.013	4
`..........(........................)..........`	36	C→U	11,A	0.002	4
`........(..........................)..........`	9	A→G	36,C	0.012	2
`..................(.........).................`	19	C→G	29,C	0.010	2
`..................(.........).................`	29	C→G	19,C	0.006	2
`...............(............).................`	16	U→G	29,C	0.004	2
`...........(................).................`	12	C→G	29,C	0.001	2
`..............(...............)...............`	15	A→G	31,C	0.006	1
`.......(..............................).......`	39	A→C	8,G	0.005	1
`......(........................)..............`	32	G→C	7,G	0.001	1
`.............(..........).....................`	14	U→G	25,C	0.247	0
`..............(........)......................`	15	A→G	24,G	0.088	0
`.............(..........).....................`	25	C→G	14,U	0.015	0
`.............(..........).....................`	25	C→A	14,U	0.009	0
`..............(........)......................`	24	C→U	15,A	0.005	0
`..............(.....).........................`	15	A→G	21,C	0.004	0
`...............(......).......................`	16	U→C	23,G	0.002	0
`................(......)......................`	17	A→G	24,C	0.001	0

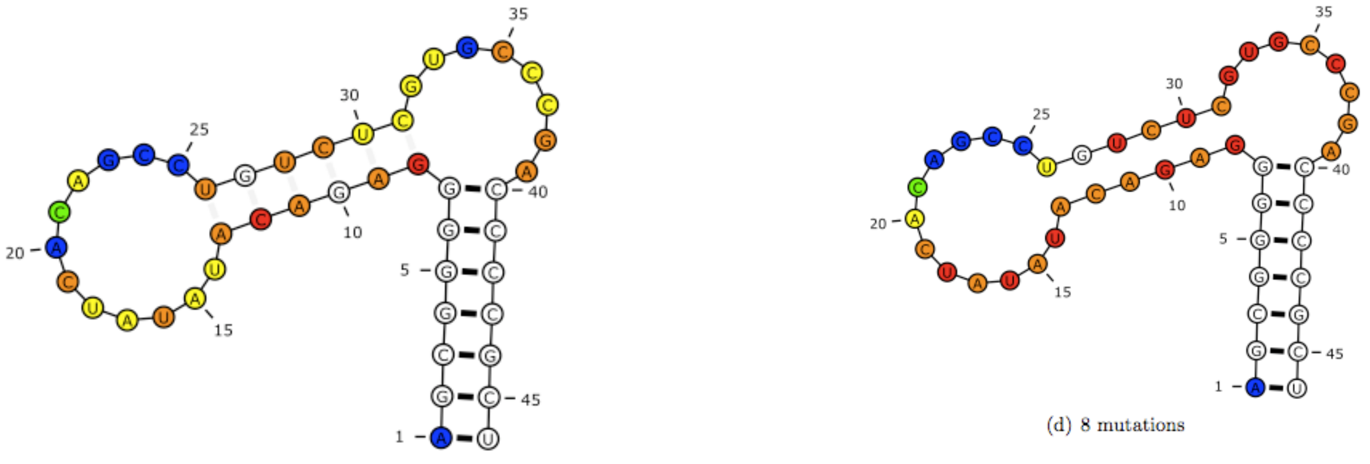
- Three mutations occurs together with C35U: A15U, C25G and A39G.
- C25G creates a potential base pair with U14 while A15U moderates it.
- In vitro, A39G is lethal: it inhibits C35U by removing/adding potential base pairs?

Simulated Evolution of HepC-CRE under thermodynamic pressure



(a) 1 mutation

(b) 2 mutations



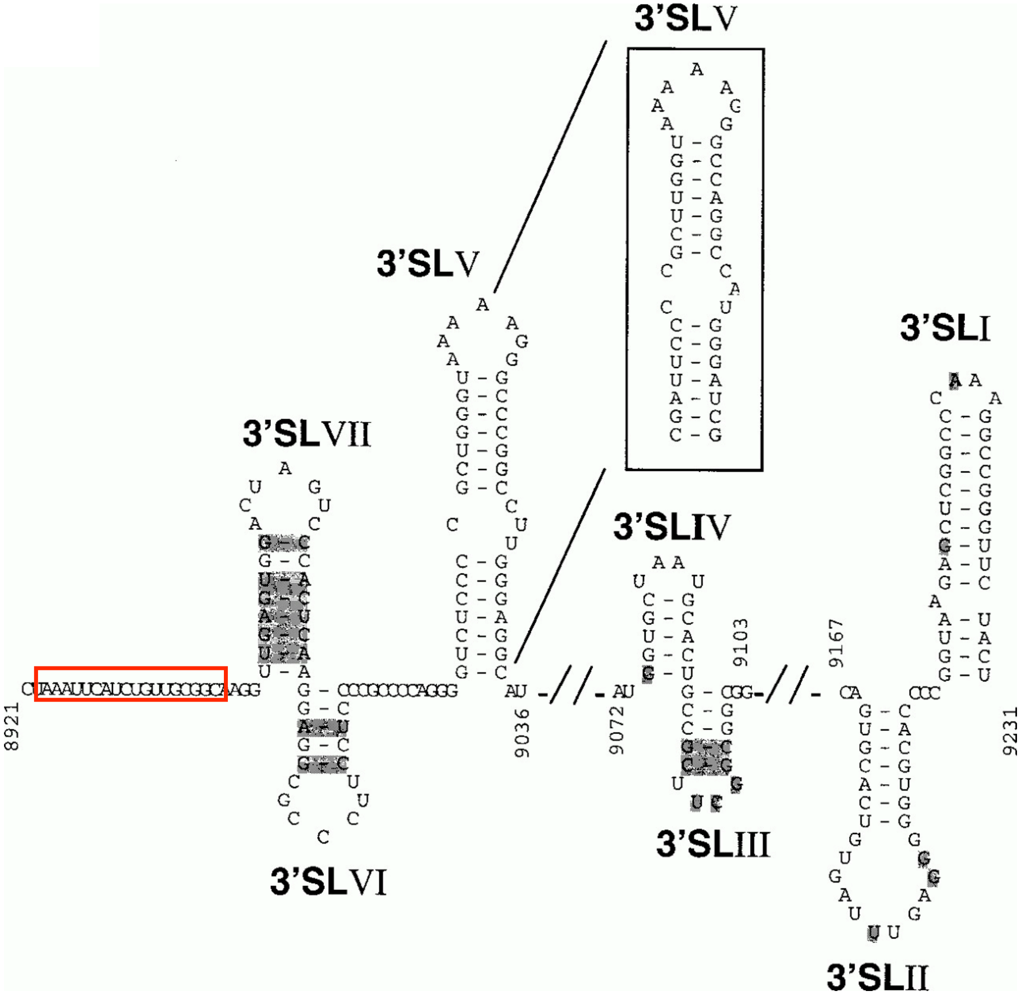
(c) 4 mutations

(d) 8 mutations

●: deleterious, ●: mainly deleterious, ●: balanced, ●: mainly beneficial, ●: beneficial

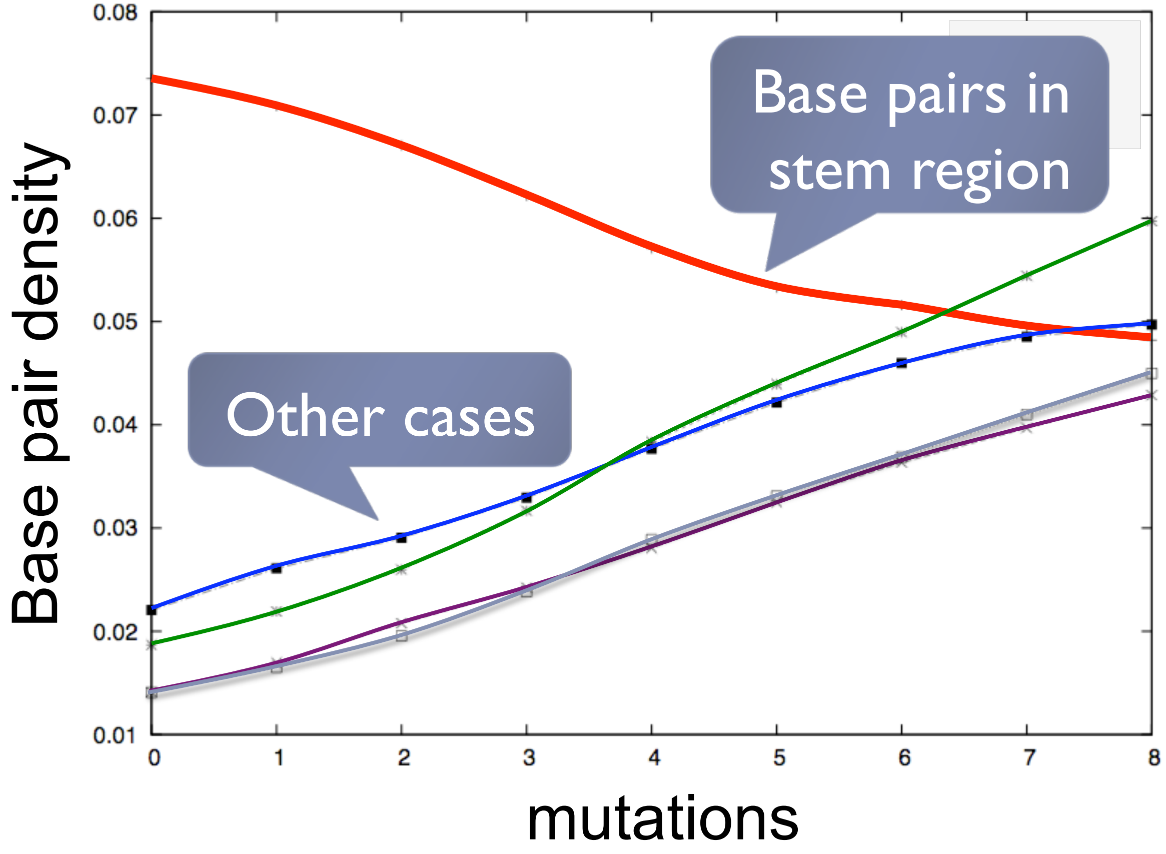
Scan of the 3'UTR of GB virus C

- 7 stems evolutionary conserved.
- Scan using frame of 50, 100 and 150 nt.
- Compute:
  1. base pairing probability
  2. mutation rate.



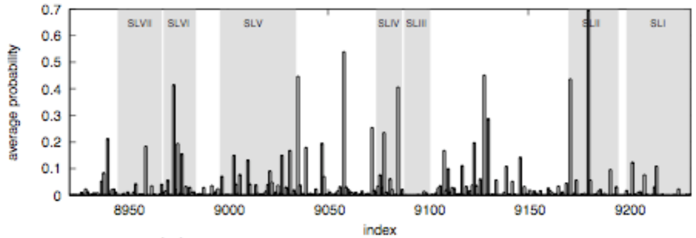
(Cucenau et al.,2001)

Scan of the 3'UTR of GB virus C: Base pair density in identified regions

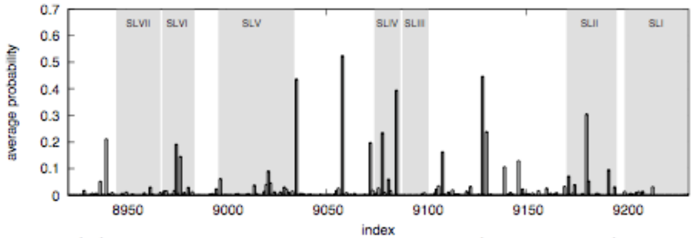


Red: base pair inside an evolutionary conserved stem region,
Other color: other cases.

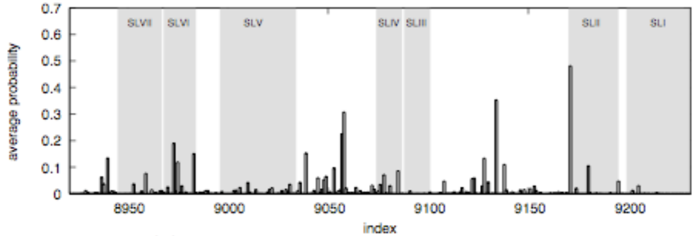
Scan of the 3'UTR of GB virus C: Distribution of favorable mutations



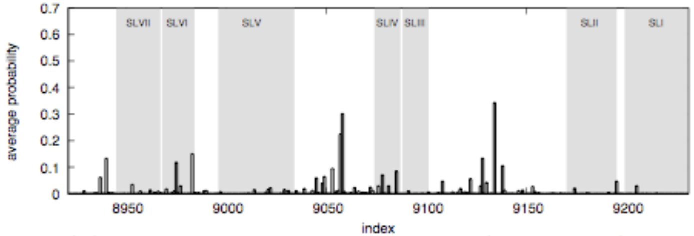
(a) Frame size of 50 nucleotides



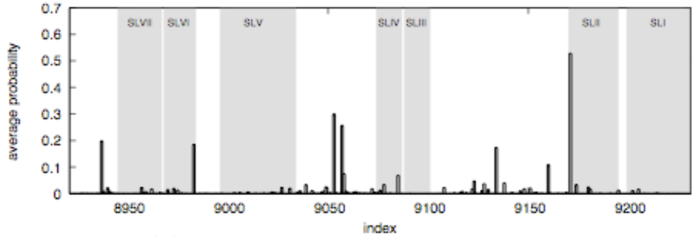
(b) Frame size of 50 nucleotides (long range)



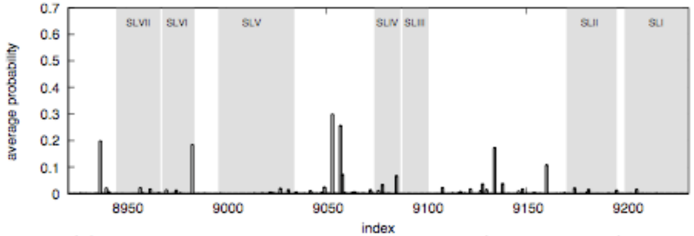
(c) Frame size of 100 nucleotides



(d) Frame size of 100 nucleotides (long range)



(e) Frame size of 150 nucleotides



(f) Frame size of 150 nucleotides (long range)

Frame size	50 nt.		100 nt.		150 nt.	
Location w.r.t. RNA regions	in	out	in	out	in	out
All mutations	48%	52%	39%	61%	38%	62%
In a base pair of size ≥25nt.	41%	59%	27%	73 %	24%	76%

Acknowledgments

MIT

- Bonnie Berger
- Srinivas Devadas
- Charlie O’Donnell
- Mieszko Lis

BC

- Peter Clote
- Andy Lorenz

LRI

- Alain Denise

Paris 6

- Yann Ponty

École Polytechnique

- Jean-Marc Steyaert
- Behshad Behzadi

Benasque’s organizers

- Elena Rivas (HHMI)
- Eric Westhof (Strasbourg)

NSF

RNA
ONTOLOGY
CONSORTIUM

digiteo

Research in information sciences and technologies

References

RNAmutants.csail.mit.edu

- J. Waldispühl, S. Devadas, B. Berger and P. Clote,
  Efficient Algorithms for Exploring the Mutation Landscape of an RNA molecule,
  to appear in PLoS Computational Biology (August 2008).

- P. Clote, J. Waldispühl, B. Behzadi and J.-M. Steyaert,
  Energy landscape of k-point mutants of an RNA molecule,
  Bioinformatics Vol. 21, p. 4140-4147, 2005. doi:10.1093/bioinformatics/bti669

- J. Waldispühl, B. Behzadi and J.-M. Steyaert,
  An Approximate Matching Algorithm for Finding (Sub-)Optimal Sequences in S-attributed Grammars,
  Proc. ECCB 2002, Bioinformatics Vol. 18, p. 250-259, 2002.