Consensus RNA secondary structures

Marcel Turcotte

School of Information Technology and Engineering

uOttawa

L'Université canadienne
Canada's university

Paradigm

- Thermodynamics (nearest-neighbor model) approaches using a single input sequence have severe limitations;
- Consensus approaches can help circumvent these limitations;
- Consensus approaches are computationally expensive.

Overview: our approaches to consensus RNA secondary structure inference

- Exact method (thermodynamics + dynamic programming). Simultaneously align and determine a common secondary structure for three (3) RNA sequences (**eXtended Dynalign**);

- Heuristic method (exhaustive search of a constrained space). Inference of consensus RNA secondary structure motifs from a set of k RNA unaligned sequences (**Seed**).

Background: RNA Secondary structure

Let $a = a_1 a_2 \ldots a_n$ be an RNA sequence, i.e. $a_i \in \{A, C, G, U\}$, and let $a_i : a_j$, for $i < j$, designate a base pair.

A secondary structure S for a is an ensemble of pairs, such that,

1. Base pairs: $a_i : a_j \in \{A : U, U : A, G : C, C : G\} \bigcup \{G : U, U : G\}$;

2. No-overlap: If S contains a pair $a_i : a_j$ then it cannot also contain $a_i : a_k$, for $k \neq j$, nor $a_k : a_j$, for $k \neq i$;

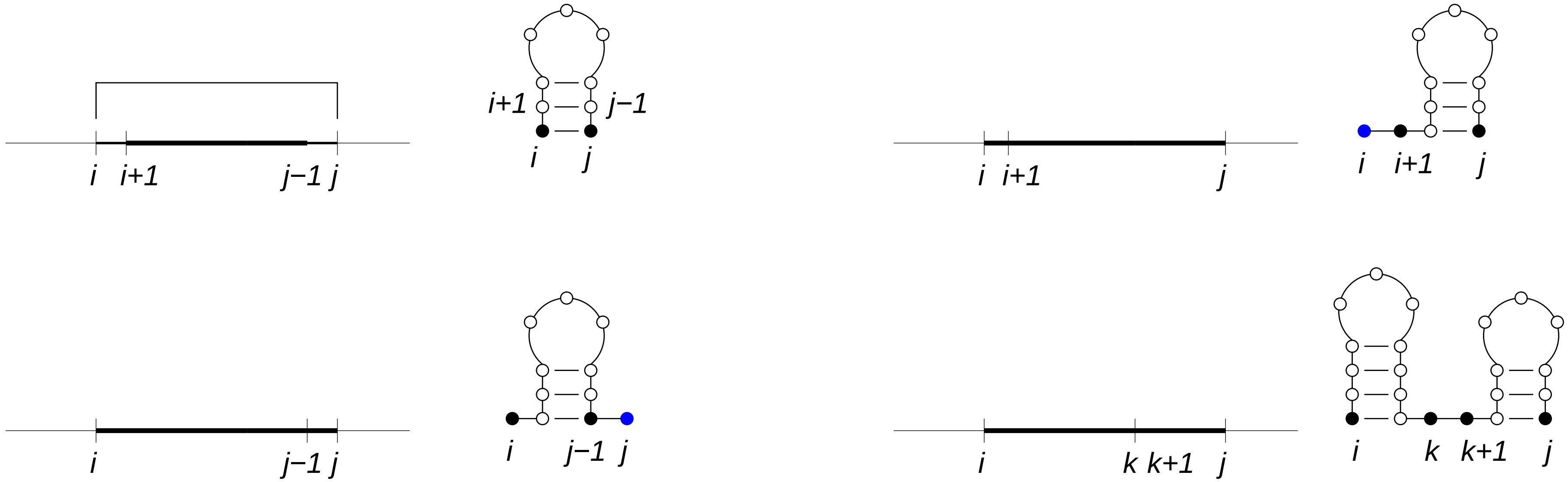
3. No-knots: given $h < i < j < k$, then S cannot simultaneously have $a_h : a_j$ and $a_i : a_k$;

4. Hairpins: If S contains $a_i : a_j$, then $|j - i| \geq 4$.

RNA Secondary Structure Determination

A didactic example first. Nussinov's algorithm finds the structure that maximises the total number of base pairs.

A well behaved problem.



Nussinov Algorithm

Initialization:

$$\gamma(i, i+k) \quad = \quad 0 \quad \text{for k} = 0 \text{ to } 2 \text{ and for i} = 1 \text{ to n} - \text{k}.$$

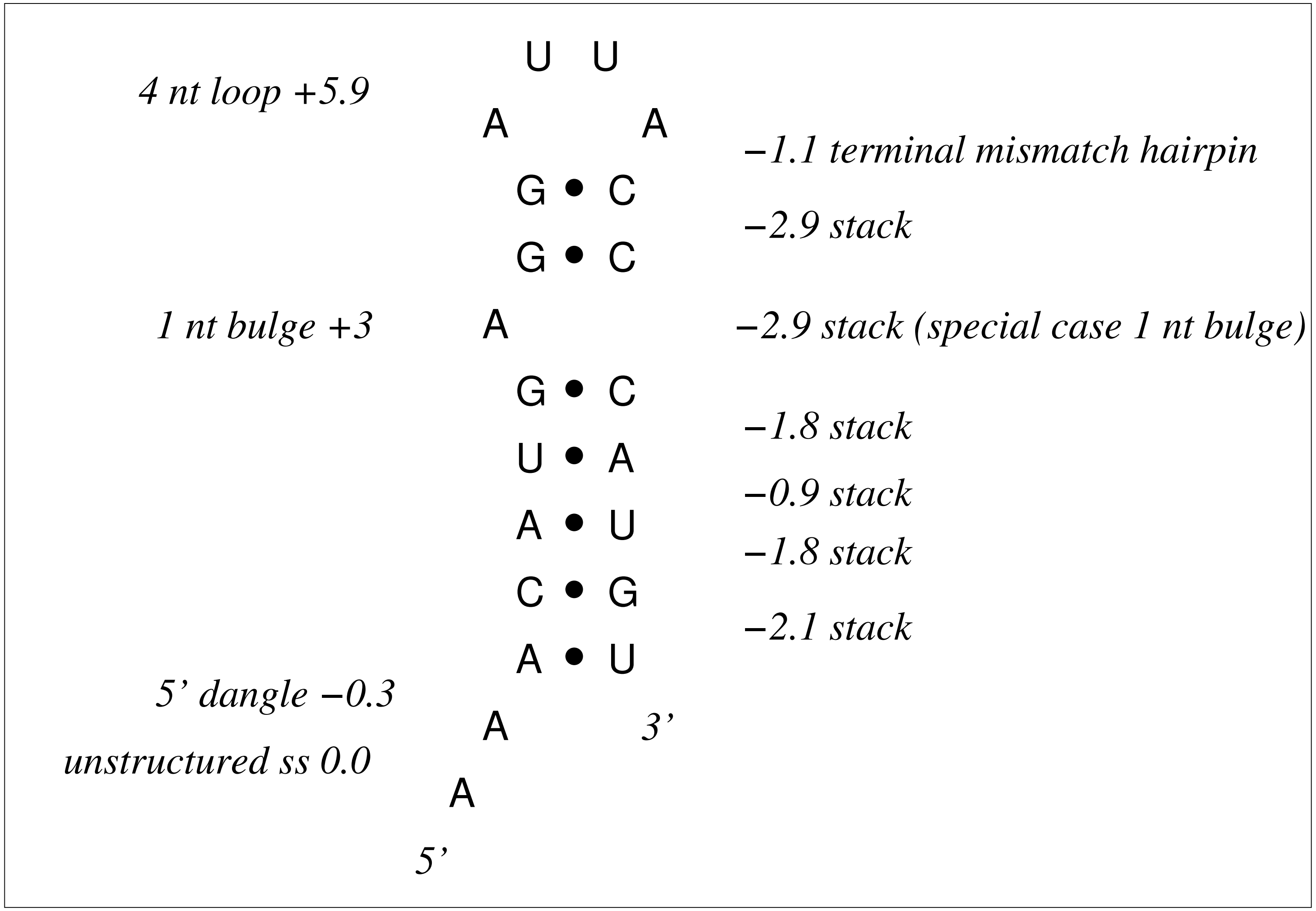
Recurrence:

$$\gamma(i,j) = \max \begin{cases} \gamma(i+1, j-1) + \delta(i,j); \\ \gamma(i+1, j); \\ \gamma(i, j-1); \\ \max_{i<k<(j-1)}[\gamma(i,k) + \gamma(k+1,j)]. \end{cases}$$

Matching score:

$$\delta(i,j) = \begin{cases} 1, \text{if } a_i : a_j \ \in \ \{A:U, U:A, G:C, C:G\} \bigcup \{G:U, U:G\}; \\ 0, \text{otherwise.} \end{cases}$$

Nearest-neighbor model



4 nt loop +5.9

−1.1 terminal mismatch hairpin

−2.9 stack

1 nt bulge +3

−2.9 stack (special case 1 nt bulge)

−1.8 stack

−0.9 stack

−1.8 stack

−2.1 stack

5’ dangle −0.3

unstructured ss 0.0

Performance of the Nearest-Neighbour Model (for a single sequence)

The nearest-neighbour model works reasonably well for small RNAs, 69 % and 71 % PPV (positive predictive value) for the tRNA and 5S rRNA, which are approximately 80 and 120 nucleotides long, respectively.

K. J. Doshi, J. J. Cannone C. W. Cobaugh, et R. R. Gutell (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics **5**(1):105.

Observations

- RNAs conserve secondary structure interactions more than they conserve their sequence;
- Single-sequence methods can be generalised to determine a consensus structure for more than one sequence;
- The nearest-neighbour model performs well on average but fails for certain sequences;
- As the number of input sequences increases it becomes unlikely that the nearest-neighbour model simultaneously fails for all of them.

eXtended Dynalign

- Sankoff 1985 proposed a set of recurrence equations for simultanesouly solving the alignment and secondary structure determination problems;

  David Sankoff (1985) Simultaneous solution of RNA folding, alignment and protosequence problems. SIAM J. Appl. Math. **45**(5):810–825.

- Objective function is a linear combination of the free energy of each sequence given the common secondary structure;

- Mathews and Turner 2002 created an implementation, called Dynalign, for two sequences;

  D.H. Mathews et D.H. Turner (2002) Dynalign: An Algorithm for Finding the Secondary Structure Common to Two RNA Sequences. J. Mol. Biol. **317**:191–203.

- We extended this work for three sequences.

Idea

- The objective function is a linear combination of the free energy of each sequence given the common structure;

$$\Delta G^\circ_{\text{total}} = \Delta G^\circ_{\text{seq 1}} + \Delta G^\circ_{\text{seq 2}} + \Delta G^\circ_{\text{seq 3}} + \Delta G^\circ_{\text{insertions}}$$

- No terms for substitutions;

- Solved by dynamic programming: constructing an alignment and a common secondary structure for $S_1[i,j], S_2[k,l]$ and $S_3[m,n]$, from the smallest to the largest segment.

Idea

```
Score= -578

GCCCGGGTGGTGTAGTGGCCCATCATACGACCCTGTCACGGTCG-TGACGCGGGTTCAAATCCCGCCTCGGGCGCCA
GTCGCAATGGTG-TAGTTGGGAGCATGACAGACTGAAGATCTGTTGGTCATCGGTTCGATCCCGGTTTGTGACACCA
GCCCCCAUCGUCUAGAGGCCUAGGACACCUCCCUUUCACGGAGG-CGACAGGGAUUCGAAUUCCCUUGGGGGUACCA

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

eXtended Dynalign

The recurrence equations describing the free energy are somewhat complex. There are 140 cases: $V_1, V_2, V_{3_{1-64}}, W_1, W_2, W_{3_{1-64}}, W_{9_{1-8}}$.

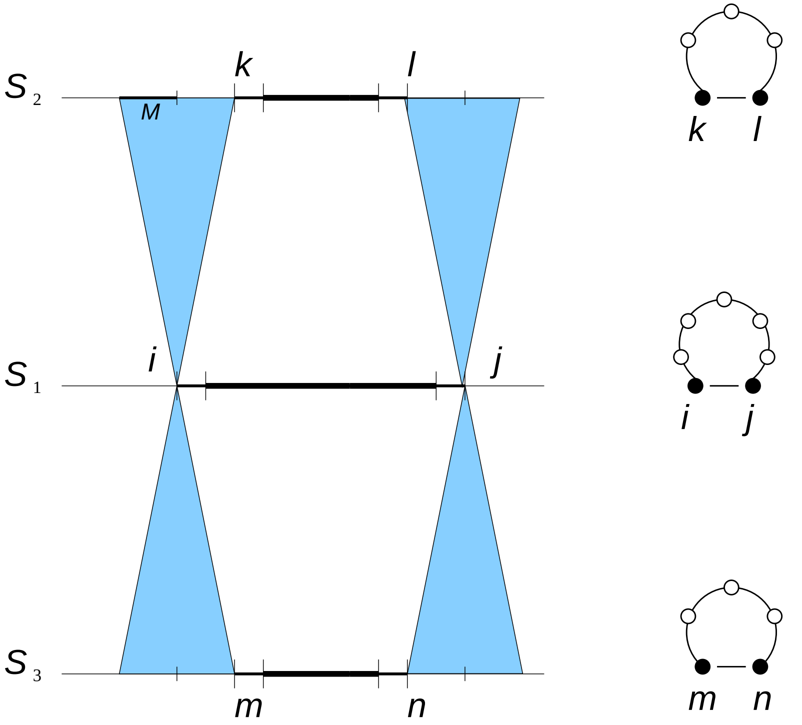
Let S_1, S_2 and S_3, be three RNA sequences.

- $W(i,j;k,l;m,n)$ represents the some of the free energy of $S_1[i,j]$, given the common structure, $S_2[k,l]$ given the common secondary structure and $S_3[m,n]$;

- $V(i,j;k,l;m,n)$ is defined similarly to W but also imposes constraints such that i is paired with j, k is paired with l, and m is paired with m;

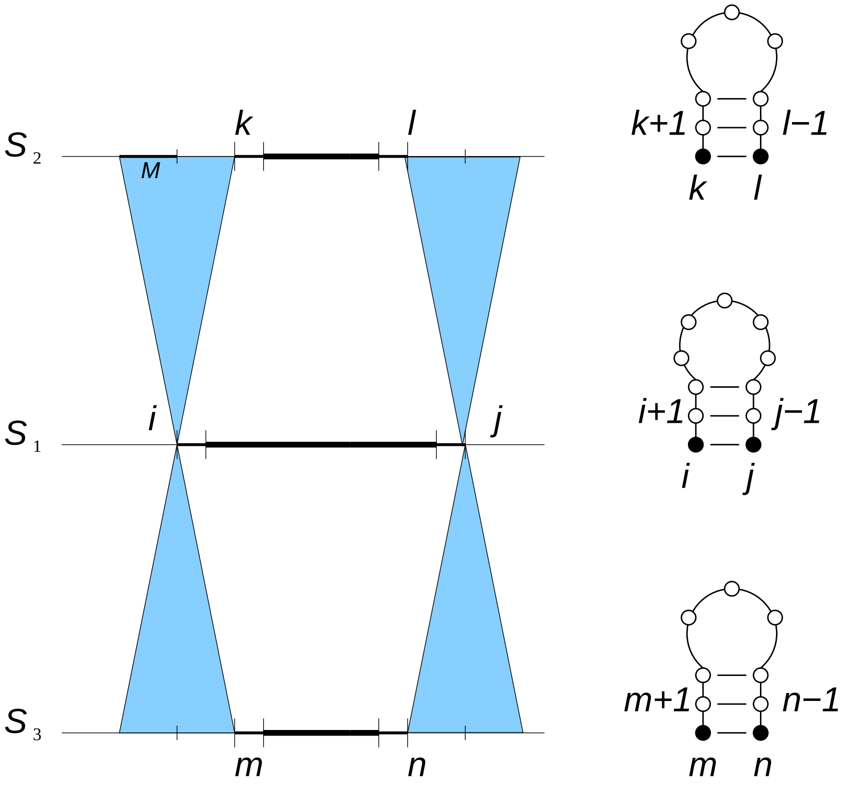
- $W9$ represents the free energy for a prefix alignment of $S_1[1,j]$, $S_2[1,l]$ and $S_3[1,n]$.

Hairpin loop closed by a base-pair: $V_1(i, j; k, l; m, n)$



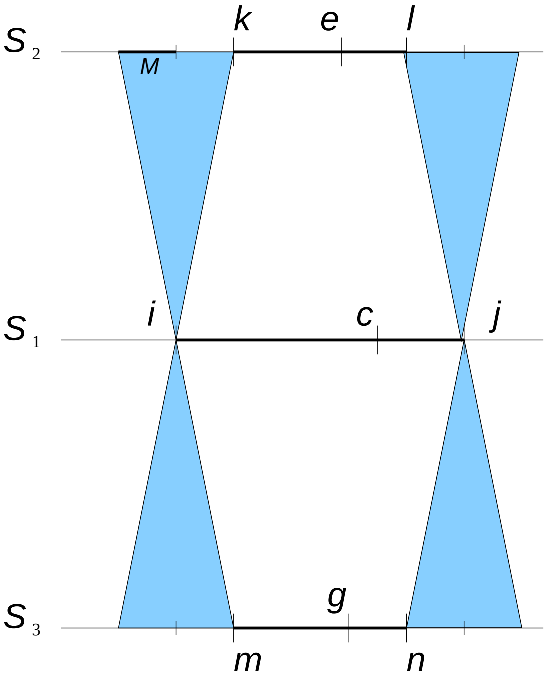
$$\Delta G^\circ_{\text{hairpin}}(i, j) + \Delta G^\circ_{\text{hairpin}}(k, l) + \Delta G^\circ_{\text{hairpin}}(m, n) + \Delta G^\circ_{\text{gap}}(\text{no. of gaps})$$

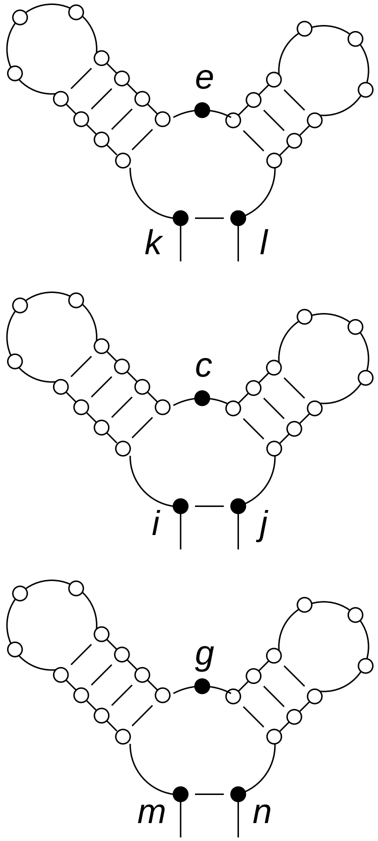
Helix Extension: $V_{2.1}(i,j;k,l;m,n)$



$$V(i+1,j-1;k+1,l-1;m+1,n-1)+\Delta G^{\circ}_{\text{motif}_1}+\Delta G^{\circ}_{\text{motif}_2}+\Delta G^{\circ}_{\text{motif}_3}$$

Multibranch Loop: $V_{3.1}(i,j;k,l;m,n)$





$$W(i,c;k,e;m,g)+W(c+1,j;e+1,l;g+1,n)+\Delta G^\circ_{\mathrm{motif}_1}+\Delta G^\circ_{\mathrm{motif}_2}+\Delta G^\circ_{\mathrm{motif}_3}$$

Summary

A base pair is predicted only if it simultaneously occurs in all three sequences.

The algorithm finds a consensus structure.

An alignment is produced as a byproduct. However, it is reliable only in the base paired regions. No substitution scores are used.

tRNA Dataset

Id	Length	Description
RD0260	77	Asp Phage T5 (Virus)
RD0500	76	Asp *Haloferax volcanii* (Archae)
RD4800	71	Asp *Aedes albopictus* (Mitochondria, Animal)
RE2140	76	Glu *Synechocystis sp.* (Eubacteria)
RE6781	76	Glu *Hordeum vulgare* (Chloroplast)
RF6320	76	Phe *Schizosaccharomyces pombe* (Cytoplasm, Fungi)
RL0503	88	Leu *Haloferax volcanii* (Archae)
RL1141	89	Leu *Mycoplasma capricolum* (Eubacteria)
RS0380	88	Ser *Halobacterium cutirubrum* (Archae)
RS1141	92	Ser *Mycoplasma capricolum* (Eubacteria)

The percentage of sequence identify varies from 27.3 to 68.8 %.

MFOLD: tRNAs

Id	Sensitivity	PPV	MCC
RD0260	33.3	29.2	31.2
RD0500	47.6	43.5	45.5
RD4800	42.9	56.2	49.1
RE2140	95.2	87	91
RE6781	33.3	28	30.6
RF6320	0	0	0
RL0503	0	0	0
RL1141	40	43.5	41.7
RS0380	52	56.5	54.2
RS1141	19.2	25	21.9

5S rRNAs

Id	Length	Description
AJ131594	117	*Delftia acidovorans*
AJ251080	117	*Geobacillus stearothermophilus*
K02682	120	*Micrococcus luteus*
M10816	119	*Geobacillus stearothermophilus*
M16532	121	*Thermus sp.*
M25591	117	*Geobacillus stearothermophilus*
V00336	120	*Escherichia coli*
X02024	119	*Sporosarcina pasteurii*
X02627	120	*Agrobacterium tumefaciens*
X04585	119	*Rhodobacter capsulatus*
X08000	122	*Arthrobacter oxydans*
X08002	122	*Arthrobacter globiformis*

The percentage of identity varies from 47.2 to 88.2%.

MFOLD: 5S rRNAs

Id	Sensitivity	PPV	MCC
AJ131594	23.7	60	37.7
AJ251080	26.3	45.5	34.6
D11460	15.8	37.5	24.3
K02682	20.5	40	28.6
M10816	31.6	70.6	47.2
M16532	10.3	21.1	14.7
M25591	26.3	45.5	34.6
V00336	37.5	65.2	49.5
X02024	15.8	37.5	24.3
X02627	38.5	68.2	51.2
X04585	0	0	0
X08000	0	0	0
X08002	0	0	0

Are three input sequences better than two?

1. The worse prediction (minimum accuracy) should be more accurate;
2. Use of three input sequences should improve the average accuracy;
3. Average coverage should be less.

Masoumi, B. and Turcotte, M. (2005) Simultaneous alignment and structure prediction of three RNA sequences. *Int. J. Bioinformatics Research and Applications.* Vol. 1, No. 2, pp. 230-245

Beeta Masoumi and Marcel Turcotte. Simultaneous alignment and structure prediction of RNAs: Are three input sequences better than two? In S. V. Sunderam et al., editor, *2005 International Conference on Computational Science (ICCS 2005)*, Lecture Notes in Computer Science 3515, pages 936-943, Atlanta, USA, May 22-25 2005.

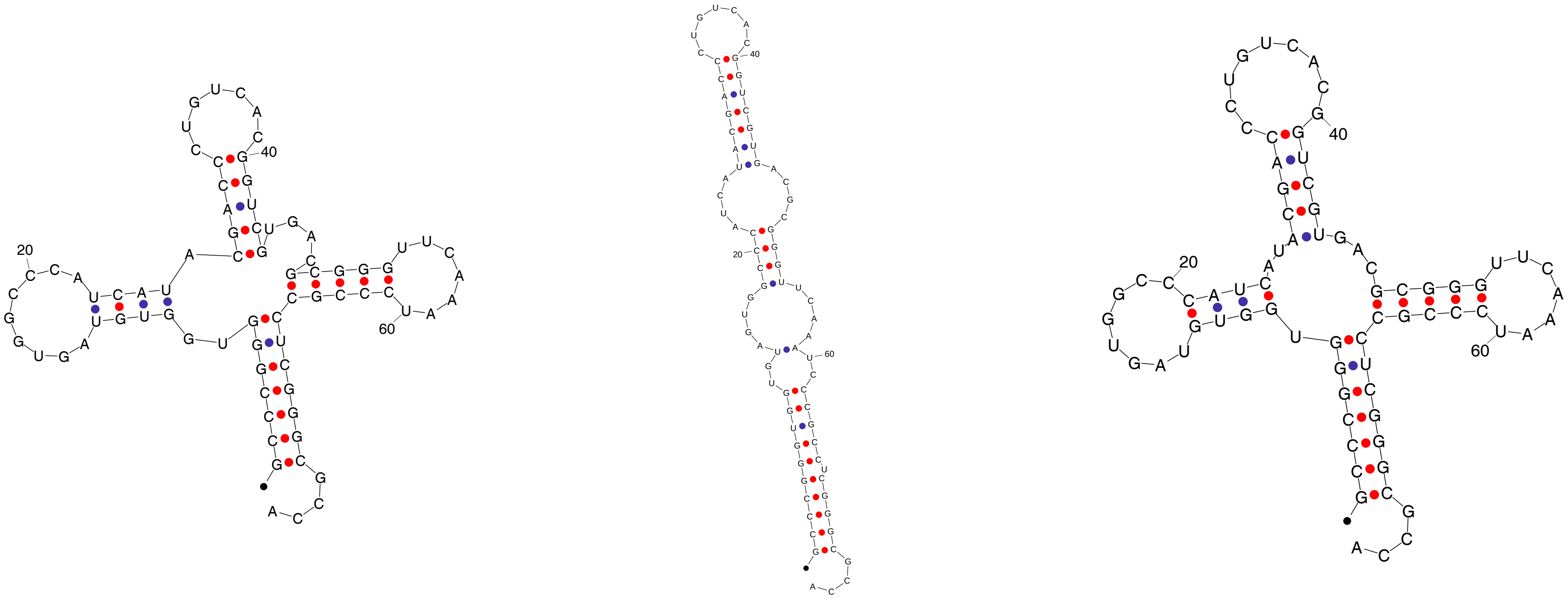
PPV: tRNA Dataset

Id	N$_{xd}$	N$_{d}$	Min$_{xd}$	Min$_{d}$	Max$_{xd}$	Max$_{d}$	Ave$_{xd}$	Ave$_{d}$
RD0260	4	5	100	80	100	100	100.0	96.0
RD0500	4	5	76	45	100	100	82.2	80.8
RD4800	5	5	100	80	100	100	100.0	96.0
RE2140	2	4	100	100	100	100	100.0	100.0
RE6781	2	4	100	77	100	100	100.0	94.3
RF6320	4	5	95	45	100	100	96.4	89.1
RL0503	1	2	100	100	100	100	100.0	100.0
RL1141	2	3	100	70	100	100	100.0	90.3
RS0380	1	2	100	83	100	87	100.0	85.2
RS1141	2	3	100	70	100	100	100.0	90.3

xd stands for eXtended Dynalign, d stands for Dynalign.

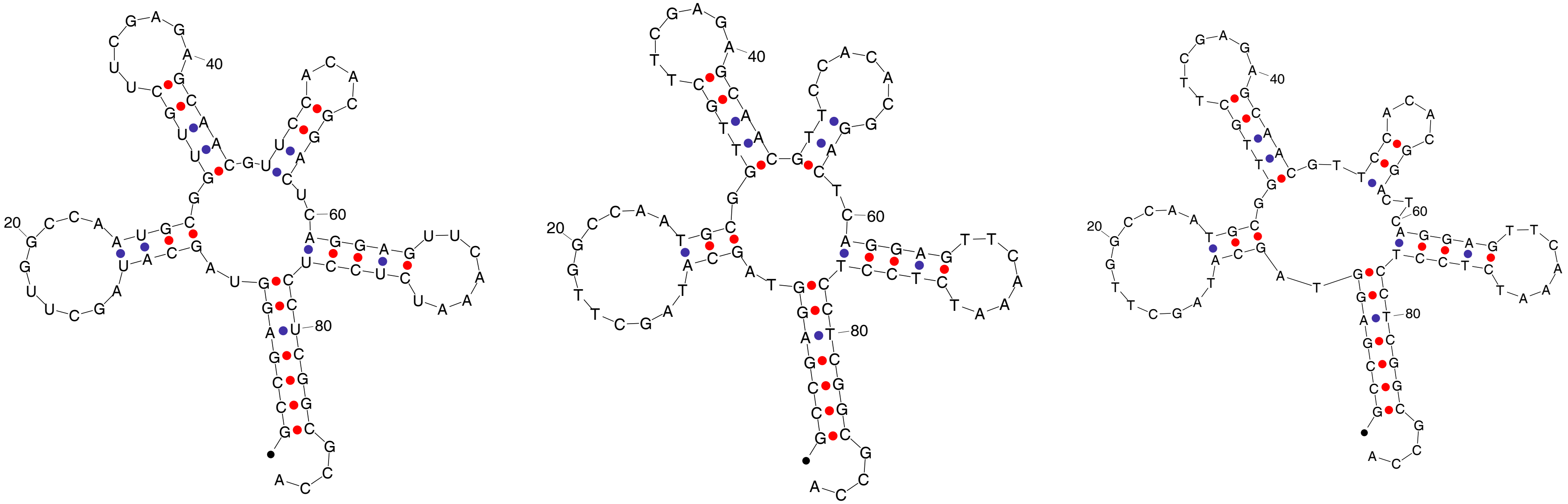
X-Dynalign 96.8 ± 7.6 vs Dynalign 92.1 ± 14.6.

eXtended-Dynalign reproduces the clover-leaf structure



(a) RD0500 (b) Dynalign (c) X-Dynalign

Fine details are better reproduced as well



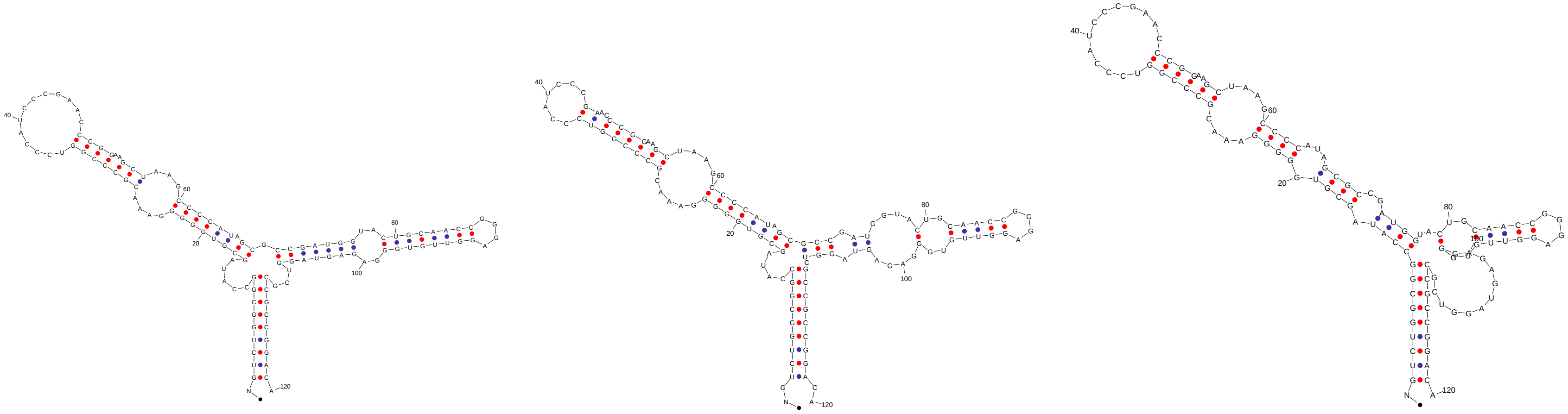
(a) RS0380 **(b) Dynalign** **(c) X-Dynalign**

PPV: 5S rRNA

Id	N$_{xd}$	N$_d$	Min$_{xd}$	Min$_d$	Max$_{xd}$	Max$_d$	Ave$_{xd}$	Ave$_d$
AJ131594	2	3	100	91	100	100	100.0	94.5
AJ251080	6	5	88	82	90	86	90.3	84.8
D11460	6	5	87	66	87	88	87.6	79.4
K02682	8	9	63	88	100	97	89.1	92.0
M10816	3	4	90	85	90	88	90.7	87.8
M16532	1	2	94	77	94	85	94.1	81.8
M25591	6	5	87	82	90	86	89.8	84.8
V00336	3	4	75	65	100	100	91.9	91.4
X02024	9	6	88	82	90	88	90.1	85.8
X02627	1	2	100	92	100	100	100.0	96.0
X04585	2	3	72	68	94	93	83.4	82.7
X08000	5	5	90	88	90	90	90.6	89.4
X08002	5	5	90	88	90	90	90.6	89.4

X-Dynalign 90.3 ± 5.8, Dynalign = 87.7 ± 7.4.

(K02682,V00336,X04585), PPV = 63%



Reference, Dynalign and X-Dynalign structures for the 5S rRNA K02682.

Pros: eXtended Dynalign

- The mean PPV is higher;
- Better worse case scenario;
- The average sensitivity is slightly degraded. However, for the majority of the sequences the minimum sensibility is higher for eXtended Dynalign;
- Some subtle details, such as the variable loop of some tRNAs, are well reproduced.

Cons: eXtended Dynalign

- $\mathcal{O}(|S_1|^2M^4)$ space, $\mathcal{O}(|S_1|^3M^6)$ time;
- Severe constraint M, $M \leq 6$;
- Up to two weeks of CPU time for some sequences[1];
- Length limited to some 150 nucleotides.

[1] Sun Fire V20z, AMD Opteron 2.2 GHz, Solaris 9

Future Work

- Reducing the runtime?
- Using a window-based approach to study longer sequences;
- Developing tools to integrate and analyse the results of several experiments;
- Developing tests for determining the likelihood of a structure.

Seed: Summary

- Novel approach for discovering consensus secondary structure motifs in unaligned RNA sequences;
- Exhaustive exploration of a space induced from a “seed” sequence using minimum support contraints;
- Uses suffix arrays for enumerating stems (first step of the motif inference algorithm);
- Uses suffix arrays for efficiently matching RNA secondary structure motifs (pattern matcher).

This particular phase of the project focuses on the exploration of the search space. We are currently investigating objectives functions in a second phase.

Seed: Research objectives (1/2)

Developing a tool taking as input an ensemble of (unaligned) sequences and producing as output a list of conserved structural motifs.

With the following additional constraints:

- No (or little) sequence similarity;
- More than one family present in the input sequences.

Seed: Research objectives (2/2)

For this particular phase of the project, we wanted answers to the following questions.

- Are support and exclusion constraints sufficiently powerful to make an exhaustive search of the secondary structure space feasible?
- Does the search space contain biologically interesting motifs?

Truong Nguyen, Mohammad Anwar and Marcel Turcotte (Submitted) Identification and evaluation of consensus RNA secondary structures using suffix arrays.

Truong Nguyen and Marcel Turcotte (2005) Exploring the Space of RNA Secondary Structure Motifs Using Suffix Arrays. 6th International Symposium on Computational Biology and Genome Informatics (CBGI 2005). Editors S. Blair et al., Salt Lake City, Utah, USA, July 21-26, 2005, 1291–1298.

Why proposing a new method?

We think that existing methods are not appropriate for studying regulatory motifs.

- Exact methods, such as eXtended Dynalign, are limited to 3 short sequences. Furthermore, the common secondary structure cannot be more than M positions apart (it would not be computationally feasible to perform a local alignment/structure prediction);
- Less structured;
- Modular;
- Unaligned;

Overview (1/6)

- Input: k unaligned sequences;
- Select a **seed** sequence;
- Within the search space induced from the seed sequence report all the motifs that are matching a sufficiently large number of the input sequences (support).

Phase I focused on building an efficient framework for exploring the space of RNA secondary structure motifs.

Phase II (just started) will focus on building effective objective functions.

Overview (2/6)

```
>RD0260 (*)
GCGACCGGGGCUGGCUUGGUAAUGGUACUCCCCUGUCACGGGAGAGAAUGUGGGUUCAAAUCCCAUCGGTCGCGCCA
>RD0500
GCCCGGGUGGUGUAGUGGCCCAUCAUACGACCCUGUCACGGUCGUGACGCGGGUUCAAAUCCCGCCUCGGGCGCCA
>RD1140
GGCCCCAUAGCGAAGUUGGUUAUCGCGCCUCCCUGUCACGGAGGAGAUCACGGGUUCGAGUCCCGUUGGGGUCGCCA
>RD2640
GGGAUUGUAGUUCAAUUGGUCAGAGCACCGCCCUGUCAAGGCGGAAGAUGCGGGUUCGAGCCCCGUCAGUCCCGCCA
>RE2140
GCCCCCAUCGUCUAGAGGCCUAGGACACCUCCCUUUCACGGAGGCGACAGGGAUUCGAAUUCCCUUGGGGGUACCA
>RE6781
UCCGUCGUAGUCUAGGUGGUUAGGAUACUCGGCUCUCACCCGAGAGACCCGGGUUCGAGUCCCGGCGACGGAACCA
>RF6320
GUCGCAAUGGUGUAGUUGGGAGCAUGACAGACUGAAGAUCUGUUGGUCAUCGGUUCGAUCCCGGUUUGUGACACCA
```

In this example, there are 7 input sequences and RD0260 has been selected to be the Seed sequence.

Overview (3/6)

```
[ find_all_stems ]

GCGACCGGGGCTGGCTTGGTAATGGTACTCCCCTGTCACGGGAGAGAATGTGGGTTCAAATCCCATCGGTCGCGCCA

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
(((.....................................................................)))

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
((((((((.........................................................))))))))

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
(((.............................................)))
...
                                                     NNNNNNNNNNNNNNNNN
                                                     (((...........)))

                                                           NNNNNNNNNNN
                                                           (((.....)))
```

Overview (4/6)

```
[ fix_all ]

GCGACCGGGGCTGGCTTGGTAATGGTACTCCCCCGCCACGGGAGAGAATGTGGGTTCAAATCCCATCGGTCGCGCCA

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
(((..............................)))

NNGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCNN
(((..............................)))

NCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGN
(((..............................)))

GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNC
(((..............................)))

GCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGC
(((..............................)))
...
```

Overview (5/6)

```
[ combine_all ]

GCGACCGGGGCTGGCTTGGTAATGGTACTCCCCTGTCACGGGAGAGAATGTGGGTTCAAATCCCATCGGTCGCGCCA

                 GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNC
                 (((.......................................)))

                                     +

                      NNNTNNNNNNNNGNNN
                      (((((......)))))

                                     =

                 GNNNNNNNTNNNNNNNNGNNNNNNNNNNNNNNNNNNNNNNNNNNC
                 (((..(((((......))))).....................)))
```

The motifs with insufficient support are rejected.

Overview (6/6)

```
[ combine_all ]

GCGACCGGGGCTGGCTTGGTAATGGTACTCCCCTGTCACGGGAGAGAATGTGGGTTCAAATCCCATCGGTCGCGCCA

                               CNTNNNNNGNG
                               (((.....)))

                                     +

                                             GNNTNNNNGNNC
                                             ((((....))))

                                     =

                               CNTNNNNNGNGNNNGNNTNNNNGNNC
                               (((.....)))...((((....))))
```

Subsequently, the 3 helices motifs, 4 helices motifs . . . will be produced.

Seed: Search algorithm

Sequential covering

1. **while** there are more examples
    (a) select an example randomly (Seed sequence);
    (b) build the most specific motif;
    (c) general-to-specific search;
    (d) remove all the examples containing an instance of the “best” motif

The number of families of motifs present in the input sequences is not know *a priori*, this kind of algorithm “may” help uncover this number.

Observations

- Huge search space;
- Support and exclusion should be powerful constraints;
- Motifs will be matched against a fix set of sequences (over and over again).

CS content

- Sophisticated data structures, called suffix arrays, were used:
  - to efficiently enumerate stems;
  - as well as for matching motifs against the $k-1$ remaining input sequences (we have developed a non-deterministic algorithm inspired by Baeza-Yates & Gonnet's algorithm for regular expression matching).

Objective function(s)

$$\mathrm{TSum} = \Sigma_\mathrm{i}\Sigma_\mathrm{j}\ \mathrm{MFE}(\mathrm{m_{ij}})$$

$$\mathrm{TBest} = \Sigma_\mathrm{i} \min_\mathrm{j}\ \mathrm{MFE}(\mathrm{m_{ij}})$$

$$\mathrm{TWorst} = \Sigma_\mathrm{i} \max_\mathrm{j}\ \mathrm{MFE}(\mathrm{m_{ij}})$$

where m_{ij} is the jth occurrence (match) in the ith sequence. We also defined variants of these functions where the free energy of a match is normalised by the number of base pairs.

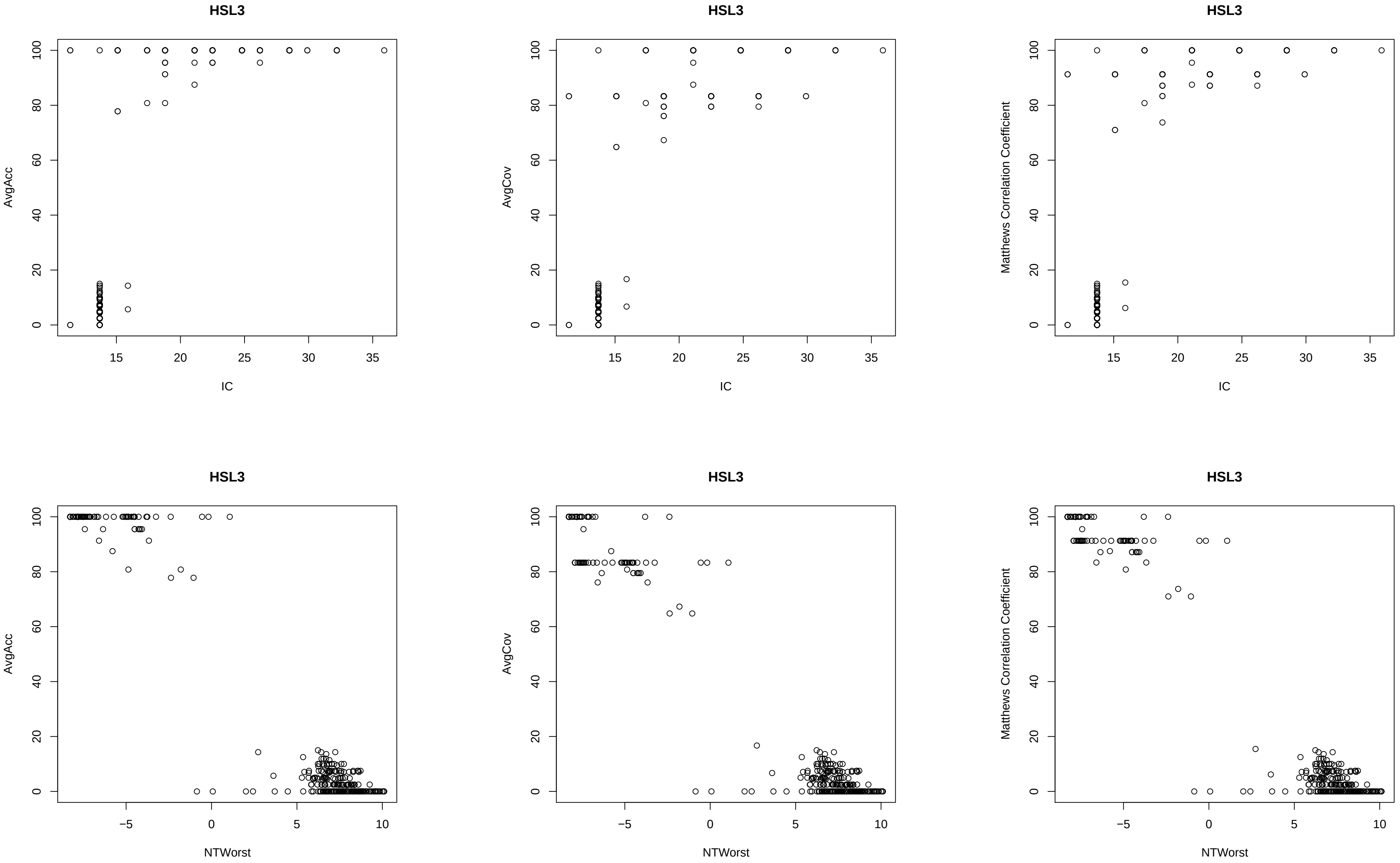
Finally, we also used a simple objective function defined as the information content of the motif.

HSL3

```
((((((....))))))
GGYYYTHHUHARRRCC
```

# Sequences	Length	# Motifs	# Matches	Space	Time
28	51–1,955	357	1,945,328	1.37 Mbytes	5m 21s

UTRSite: “This stem-loop structure plays a different role in the nucleus and in the cytoplasm. In the nucleus, it is involved in pre-mRNA processing and nucleocytoplasmic transport, whereas in the cytoplasm it enhances translation efficiency and regulates histone mRNA stability.”


HSL3
AvgAcc
IC
0
20
40
60
80
100
15
20
25
30
35
HSL3
AvgCov
IC
HSL3
Matthews Correlation Coefficient
IC
HSL3
AvgAcc
NTWorst
−5
0
5
10
HSL3
AvgCov
NTWorst
HSL3
Matthews Correlation Coefficient
NTWorst

HSL3 (motif 000269)

```
GGCNCTNNNNAGNGCC
((((((....))))))

((((((....))))))
GGYYYTHHUHARRRCC
```

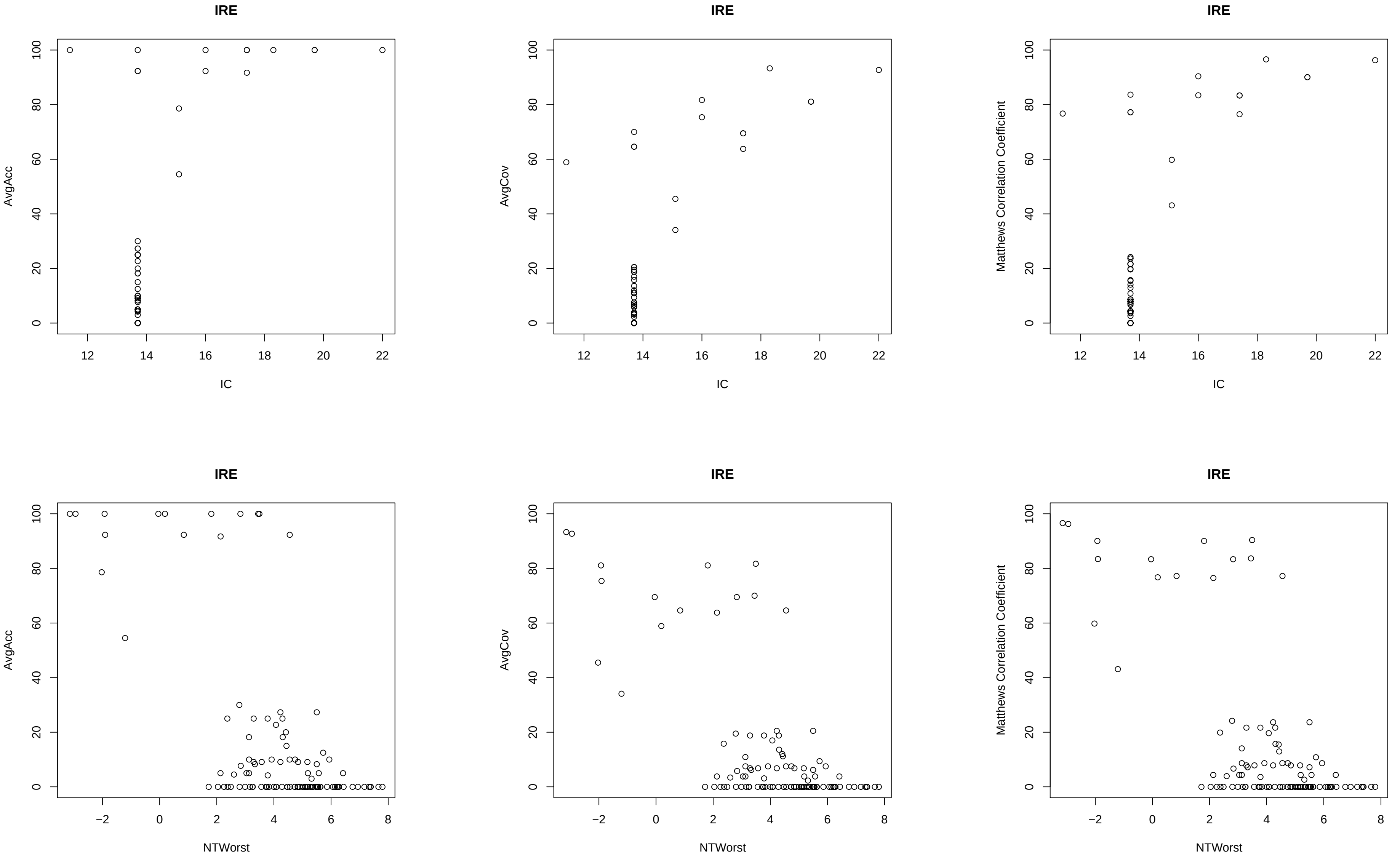
A third of the motifs inferred are 100 % accurate.

IRE

```
NNNCNNNNNCAGWGHNNNNNNNN
(((.(((((......))))))))
```

# Sequences	Length	# Motifs	# Matches	Space	Time
14	58–2,188	110	167,076	0.46 Mbytes	25s

UTRSite: “The iron-responsive element (IRE) is a particular hairpin structure located in the 5’-untranslated region (5’-UTR) or in the 3’-untranslated region (3’-UTR) of various mRNAs coding for proteins involved in cellular iron metabolism. The IREs are recognized by trans-acting proteins known as Iron Regulatory Proteins (IRPs) that control mRNA translation rate and stability.”


IRE
AvgAcc
IC
12
14
16
18
20
22
0
20
40
60
80
100
IRE
AvgCov
IC
IRE
Matthews Correlation Coefficient
IC
IRE
AvgAcc
NTWorst
−2
0
2
4
6
8
IRE
AvgCov
NTWorst
IRE
Matthews Correlation Coefficient
NTWorst

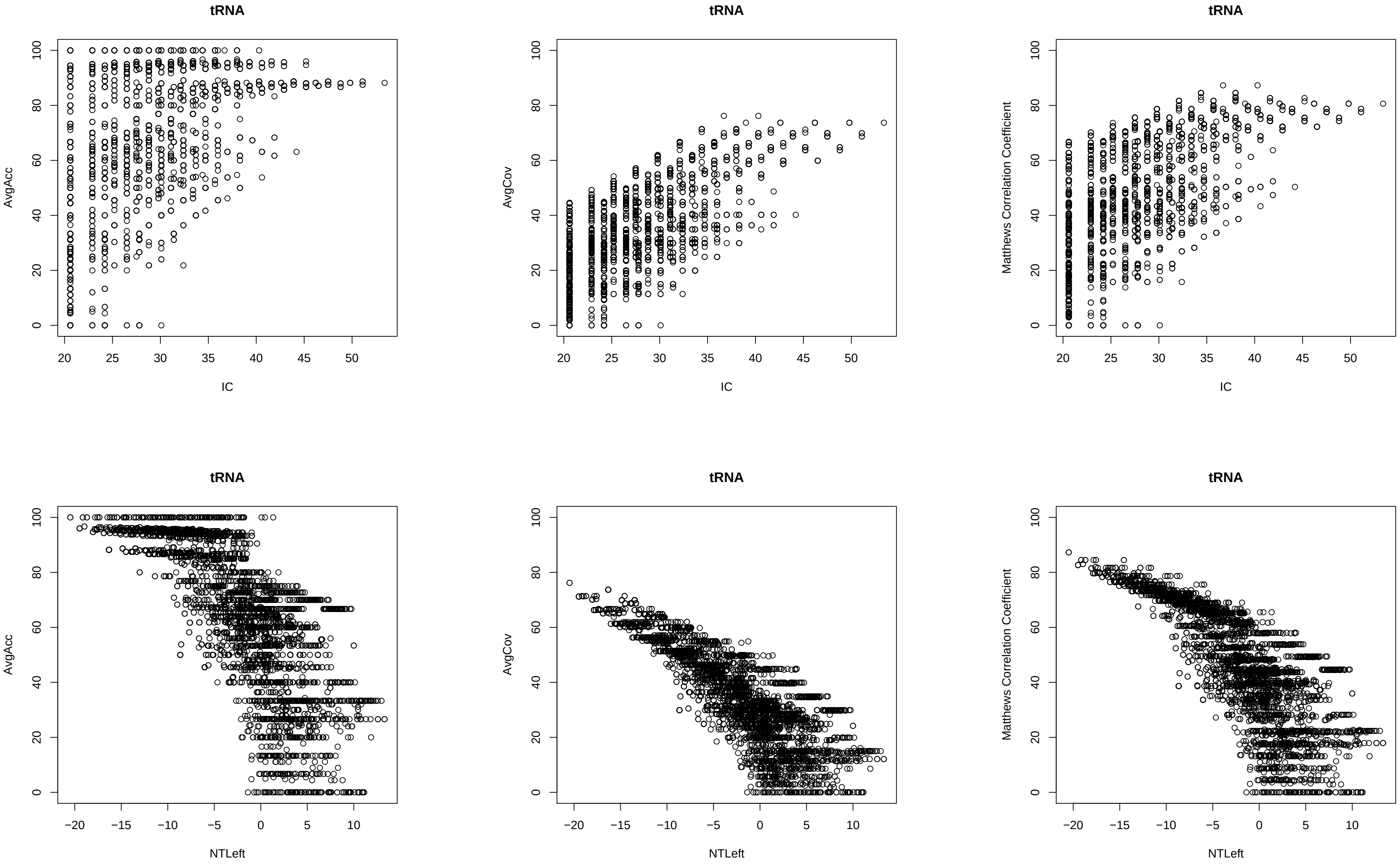
IRE (motif 000086)

```
NNNNNNNNNNNNNNNNNNNNNNN
(((.(((((......))))))))

NNNCNNNNNCAGWGHNNNNNNNN
(((.(((((......))))))))
```

tRNA

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

# Sequences	Length	# Motifs	# Matches	Space	Time
7	76–77	5,518	3,407,012	9.40 Mbytes	6m 11s



tRNA

tRNA

tRNA

tRNA

tRNA

tRNA

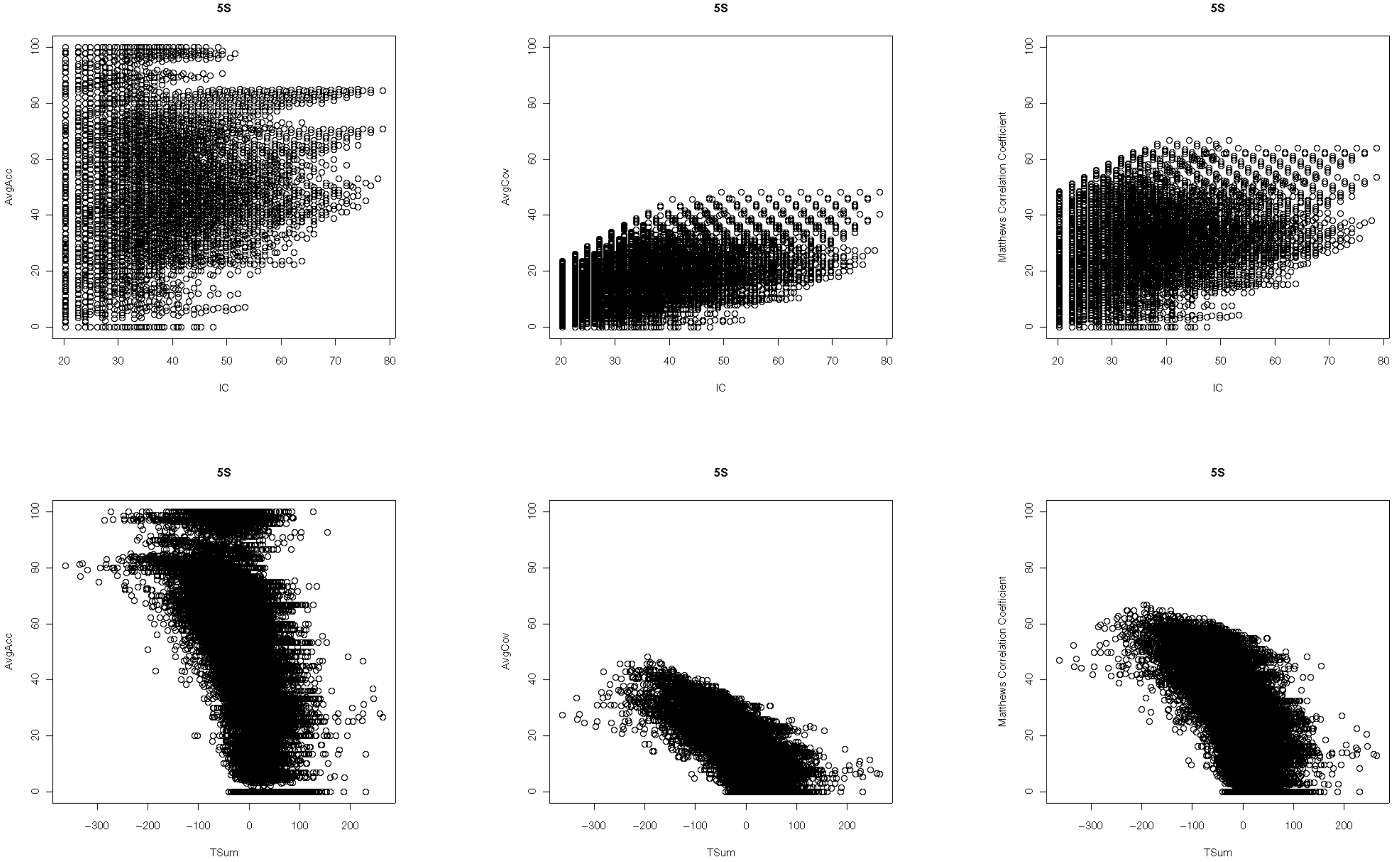
tRNA

```
NNNNNNNNNNNNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNGNNNNNNNNNNNNNNNNNNNNNNNNNNNN
(((((((.((((.........)))).(((((.......))))).....................)))))))

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

5S

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

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

# Sequences	Length	# Motifs	# Matches	Space	Time
7	117–120	364,505	152,741,463	0.52 Gbytes	7h 40m


5S
AvgAcc
IC
5S
AvgCov
IC
5S
Matthews Correlation Coefficient
IC
20 30 40 50 60 70 80
0 20 40 60 80 100
5S
AvgAcc
TSum
5S
AvgCov
TSum
5S
Matthews Correlation Coefficient
TSum
−300 −200 −100 0 100 200
0 20 40 60 80 100

5S

```
NNNNGNNNNNNNNNNNNNNNNNNNNNNNNCNGNNNNNNNNNNNNNNNCNGNNNNNNNNNNNNNNNNN
((((((((.....................((((.............)))).................

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

NNNNNNNNNNNNNNGNNNNNNNNNCNNNNNNNNNNNNNNNNNNNNCNNNN
..............(((.....)))................. ))))))))

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

Conclusions: Seed

- A suffix tree/array based approach allows us to enumerate a substantial fraction of the search space, using a reasonable amount of resources;
- The search space contains biologically interesting candidates;
- For larger and more complex structures, Seed identifies consensus base pairs which high PPV (positive predictive value) but low sensitivity;

Application: structural constraints and refolding

tRNA data-set

Id	Method	% PPV	% Sensitivity
RD0260	MFOLD (2)	28.6–29.2	28.6–33.3
	MFOLD (1)	66.7	57.1
	MFOLD (1)	57.1	57.1
	Seed NTFirst + MFOLD	**100.0**	**100.0**
RD1140	MFOLD (1)	100.0	100.0
	Seed NTFirst + MFOLD	**100.0**	**100.0**
RD2640	MFOLD (1)	63.6	66.7
	MFOLD (2)	18.2	19.0
	Seed NTFirst + MFOLD	**100.0**	**100.0**
RE2140	MFOLD (1)	87.0	95.2
	MFOLD (1)	69.6	76.2
	Seed NTFirst + MFOLD	**91.3**	**100.0**
RE6781	MFOLD (4)	28.0–31.8	33.3
	Seed NTFirst + MFOLD	**100.0**	**100.0**

Future work

- Developing better objective functions (MDL, NN);
- Adding sequence patterns in the loop regions.

Acknowledgments

University of Ottawa

Truong Nguyen (M.Sc. student)
Beta Masoumi (M.Sc. student)
Mohammad Anwar (M.Sc. student)

Informations

bio.site.uottawa.ca (home page)

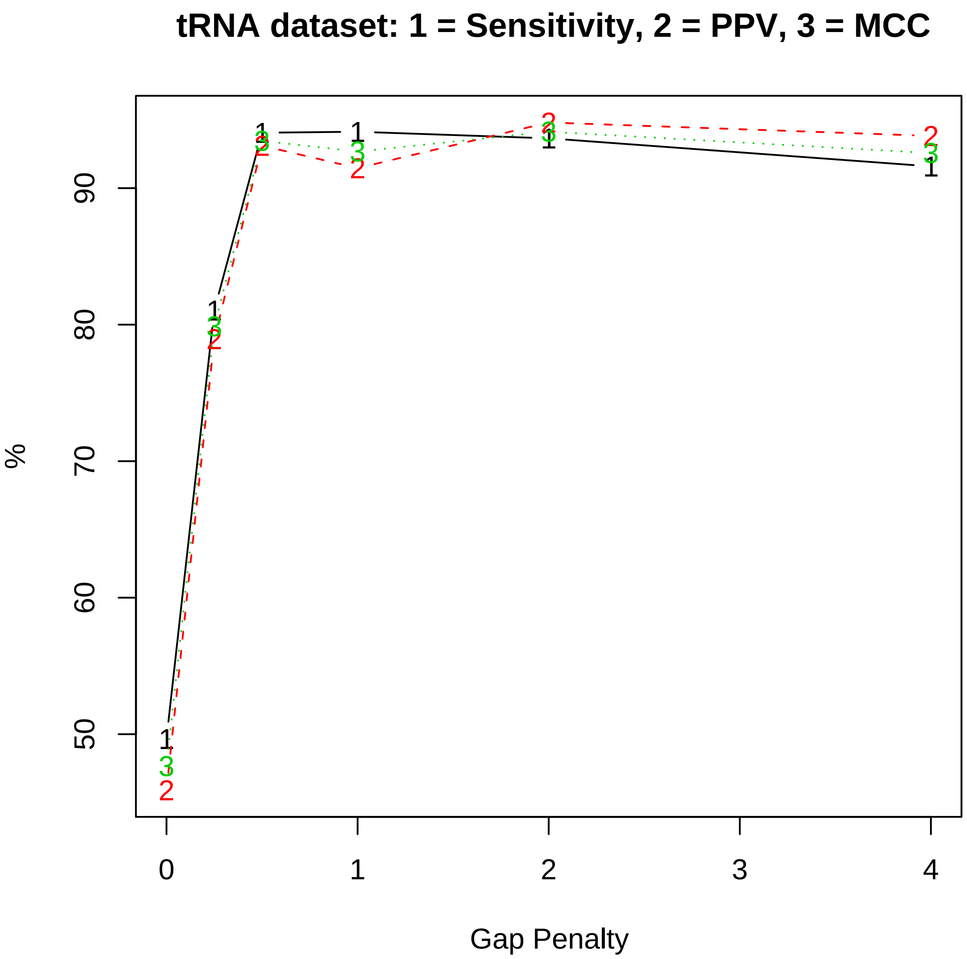
bio.site.uottawa.ca/wiki/space/start (news)

bio.site.uottawa.ca/software/x-dynalign (downloads and reprints)

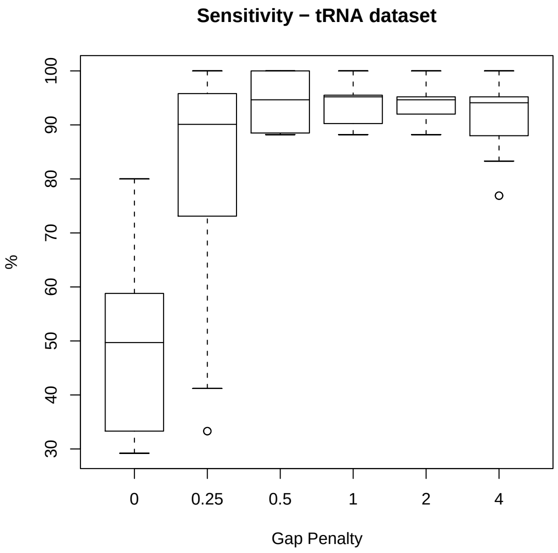
bio.site.uottawa.ca/software/seed (downloads and reprints)

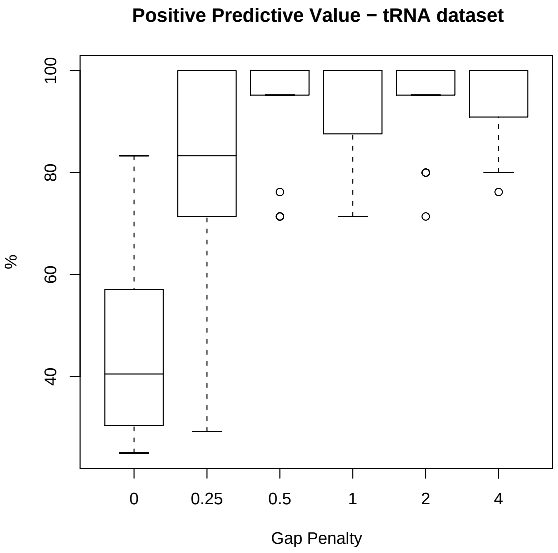
turcotte@site.uottawa.ca (E-mail)

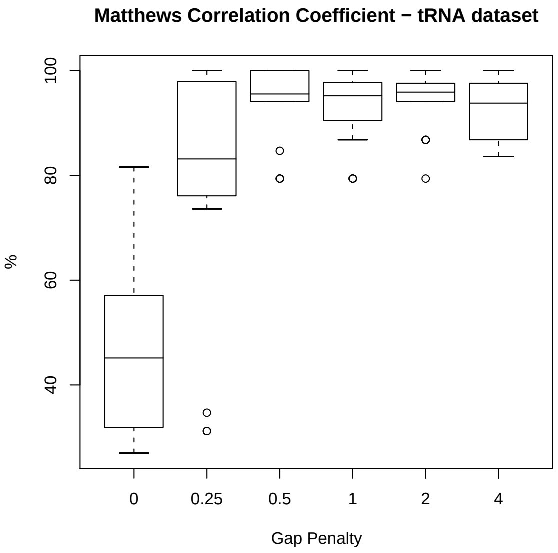
Calibrating Gap penalties: tRNAs



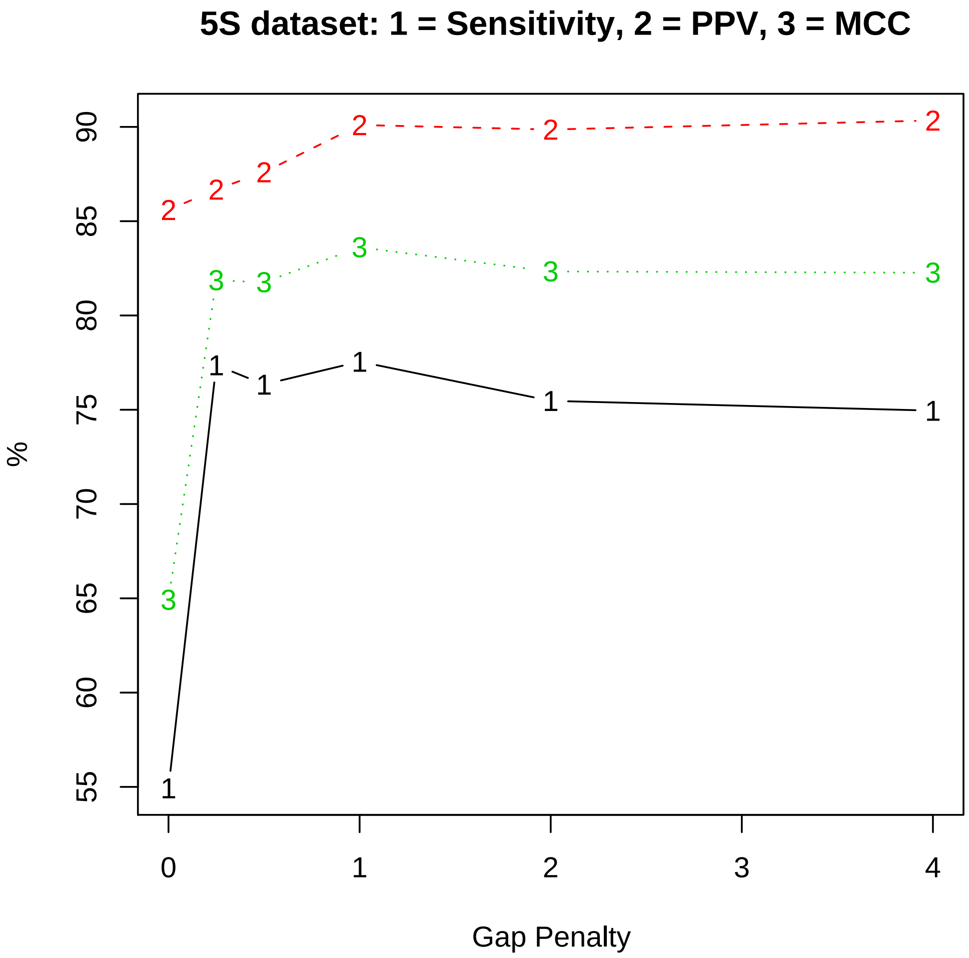
Calibrating Gap penalties: tRNAs



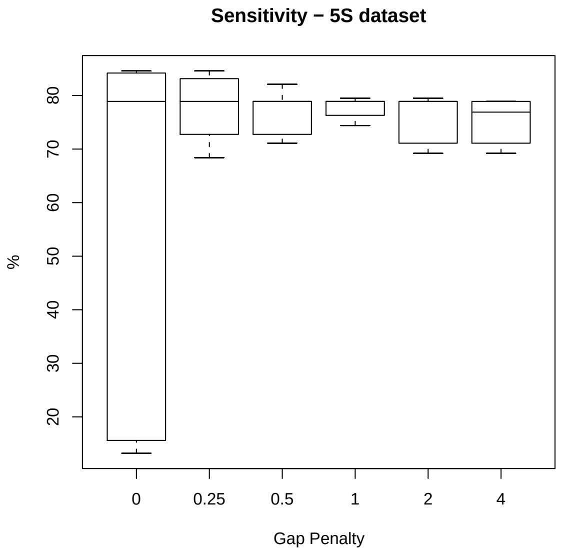


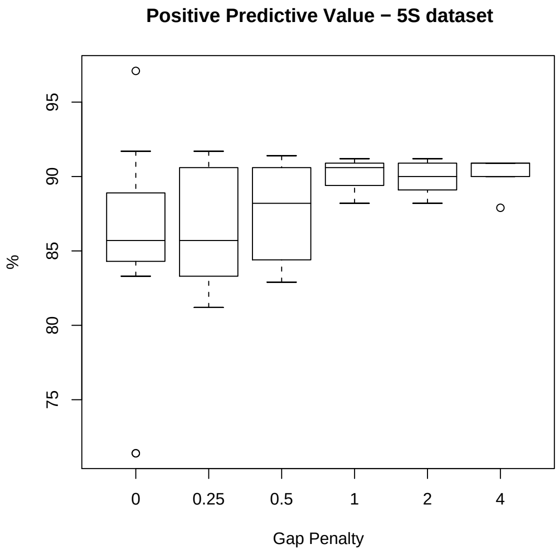


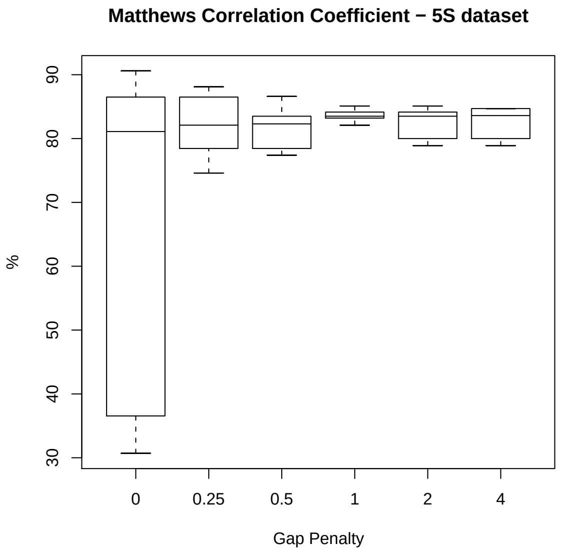
Calibrating Gap Penalties: 5S rRNAs



Calibrating Gap Penalties: 5S rRNAs


Sensitivity − 5S dataset


Positive Predictive Value − 5S dataset


Matthews Correlation Coefficient − 5S dataset

Implementation (1/3)

Suffix arrays are used rather than suffix trees.

Given an input sequence S of length $|S| = n$.

Each suffix is represented by its starting position (an integer), a suffix array lists all the suffixes in lexicographic order.

Uses $\mathcal{O}(n)$ space; with small constant.

$\log_2 n$ bits suffice to represent a position, hence, 32 bits, $4 \times n$ bits, are enough to represent a 4 Gbytes string.

Implementation (2/3)

Manber U et Myers G (1990) *Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*: 319 – 327.

Manber U and Myers G (1993) *SIAM J on Computing* **22**(5):935–948.

Until very recently constructing a suffix array was costly, $\mathcal{O}(n \log n)$.

Building in $\mathcal{O}(n)$ time.

Kärkkäinen J et Sanders P (2003) In *Proc. 30th International Colloquium on Automata, Languages and Programming (ICALP '03), LNCS 2719*, 943-955. (Skew algorithm)

Implementation (3/3)

Bottom up traversal,

Abouelhoda M et al. (2003) WABI 2002, *LNCS 2452* :449-463.

Top down traversal,

Abouelhoda M et al. (2002) SPIRE 2002, *LNCS 2476* :31-43.

See Abouelhoda et al for an excellent review.

Mohamed Ibrahim Abouelhoda and Stefan Kurtz Enno Ohlebusch Replacing suffix trees with enhanced suffix arrays (2004) J. of Discrete Algorithms **(2):1**, 53–86.

16,000+ lines of C (now shrinked to some 8,000 lines).

Performance Measures

$A \backslash P$	+	-
+	TP	FN
-	FP	TN

$$\text{Positive Predictive Value (PPV)} = TP/(TP + FP)$$

$$\text{Sensitivity} = TP/(TP + FN)$$

$$\text{Matthews Correlation Coefficient (MCC)} = \sqrt{\frac{\text{TP}}{(\text{TP} + \text{FN})} \times \frac{\text{TP}}{(\text{TP} + \text{FP})}}$$

where A = Actual, P = Predicted, TP = True Positive, FN = False Negative, FP = False Positive and TN = True Negative.

HSL3

Metazoan histone 3’-UTR mRNAs, lacking a polyA tail, contain a highly conserved stem-loop structure with a six base stem and a four base loop. This stem-loop structure plays a different role in the nucleus and in the cytoplasm. In the nucleus, it is involved in pre-mRNA processing and nucleocytoplasmic transport, whereas in the cytoplasm it enhances translation efficiency and regulates histone mRNA stability. The trans-acting factor which interacts with the 3’-UTR hairpin structure of histone mRNAs is a 31 kDa stem-loop binding protein in mammals (SLBP) present both in nuclei and polyribosomes. In mammals in addition to SLBP histone mRNA processing requires at least one additional factor: the U7 snRNP, which binds a purine-rich element 10-20 nt downstream of the stem-loop sequence (Histone Downstream Element, HDE). In all histone mRNAs analyzed so far no G has been observed in the four base loop. In all metazoan except C. elegans , there are two invariant urydines in the first and third base of the loop. In C.elegans the first base of the loop is C. Either 5’ and 3’ flanking sequences are necessary for high affinity binding of SLBP. The 5’ flanking sequence consensus is CCAAA and the 3’ flanking sequence consensus is

ACCCA or ACCA with cleavage occuring after the A. The histone 3'-UTR hairpin structure is peculiar in that the bases of the stem are conserved unlike most functional hairpin motifs where conserved bases are found in single stranded loop regions only. The sequence of the stem and flanking sequences are critical for binding of the SLBP.

IRE

The ”iron-responsive element” (IRE) is a particular hairpin structure located in the 5’-untranslated region (5’-UTR) or in the 3’-untranslated region (3’-UTR) of various mRNAs coding for proteins involved in cellular iron metabolism. The IREs are recognized by trans-acting proteins known as Iron Regulatory Proteins (IRPs) that control mRNA translation rate and stability. Two closely related IRPs, denoted as IRP-1 and IRP-2, have been identified so far which bind IREs and become inactivated (IRP-1) or degradated (IRP-2) when the iron level in the cell increases. IRPs show a significant degree of similarity to mitochondrial aconitase (EC 4.2.1.3). It has been shown that under high iron conditions IRP-1, which contains a 4Fe-4S cluster that possibly acts as a cellular iron biosensor, has enzymatic activity and may act as a cytosolic aconitase. Cellular iron homeostasis in mammalian cells is maintained by the coordinate regulation of the expression of ”Transferrin receptor”, which determines the amount of iron acquired by the cell, and of ”Ferritin”, an iron storage protein, which determines the degree of intracellular iron sequestration. Thus if the cell requires more iron, the level of transferrin receptor has to increase and conversely the level of ferritin has to decrease.

Ferritin, in vertebrates, consists of 24 protein subunits of two types, type H with Mr of 21 kDa and type L with Mr of 19-20 kDa. The apoprotein (Mr 450 kDa) is able to store up to 4500 Fe (III) atoms. The 5’-UTR of H- and L ferritin mRNAs contain one IRE whereas multiple IREs are located in the 3’-UTR of transferrin receptor mRNA. In the case of low iron concentration, IRPs are able to bind the IREs in the 5’-UTR of H- and L-Ferritin mRNAs repressing their translation and the IREs in the 3’-UTR of transferrin mRNA increasing its stability. Conversely, if iron concentration is high, IRP binding is diminished, which increases translation of ferritins and downregulate expression of the transferrin receptor. IREs have also been found in the mRNAs of other proteins involved in iron metabolism like ”erythroid 5-aminolevulinic-acid synthase (eALAS) ” involved in heme biosynthesis, the mRNA encoding the mitochondrial aconitase (a citric acid cycle enzyme) and the mRNA encoding the iron-sulfur subunit of succinate dehydrogenase (another citric acid cycle enzyme) in Drosophila melanogaster. Two alternative IRE consensus have been found. In certain IREs the bulge is best drawn with a single unpaired cytosine, whereas in others the cytosine nucleotide and two additional bases seem to oppose one free 3’ nucleotide. Some evidences also suggest a structured loop with an interaction between nucleotide one and nucleotide five (in boldcase). G W G W A G A G C H C H NN NN NN NN NN NN NN NN

NN NN C C NN N N NN N NN NN NN NN NN NN The lower stem can be of variable length and is AU-rich in transferrin mRNA. W=A,U and D=not G.

Free Engergy

In thermodynamics, the term free energy denotes either of two related concepts of importance. They express the total amount of energy which is used up or released during a chemical reaction. Both attempt to capture that part of the total energy of a system which is available for ”useful work” and is hence not stored in “useless random thermal motion”. As a system undergoes changes, its free energy will decrease.

Wikipedia