Simultaneous Alignment and Structure Prediction of RNAs: Are Three Input Sequences Better than Two?

Beeta Masoumi and Marcel Turcotte
School of Information Technology and Engineering

uOttawa

L'Université canadienne
Canada's university

Summary

- Simultaneous Alignment and Structure Prediction of RNAs: Are Three Input Sequences Better than Two?

- We have extented the software system Dynalign to simultaneously align and determine a common secondary structure for three (3) RNA sequences. We call this computer program eXtended Dynalign, X-Dynalign for short;

- We tested our software system on a challenging dataset consisting of 10 tRNAs and 13 5S rRNA, and compared its performance to Dynalign.

Motivation: Internal Ribosome Entry Site Motif (IRES)

- Secondary structure motif mainly found in the mRNAs of oncogenes (c-myc, CDC, c-jun), growth factors (FGF-2, IGF-II, IGF-R1, VEGF) and genes that control apoptosis (programmed cell death, XIAP, DAP5, Apaf-1 et Bag1);

- Makes it possible for certain genes to be translated without using the normal mechanism involving the 5' m^7GpppN cap binding complex of the mRNA;

- Cap-independent translation;

- This is well characterised for viruses but an active research topics for the Eukaryotes.

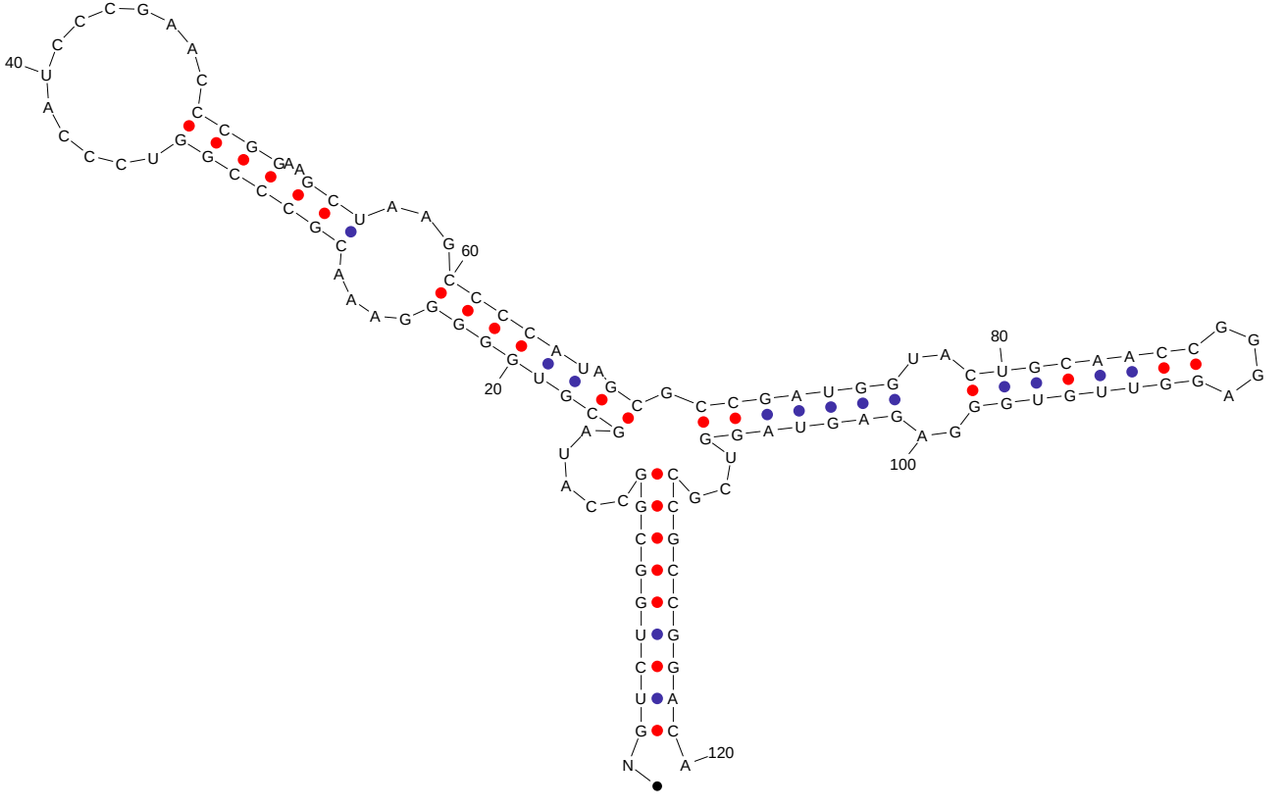
$\Rightarrow$ Collaboration with Martin Holcik from CHEO.

Motivation: Hepatitis *Delta* Virus (HDV)

- Highly pathogenic subviral human agent;
- HDV consists of approximately 1,700 nucleotides, single-stranded, circular RNA;
- 70% self-complementary, thought to fold into an unbranched rode-like structure;
- Limited protein-coding capacity (one ORF);
- Research objectives: *in silico*, *in vitro* and *in vivo* study of its local secondary structure landscape.

$\Rightarrow$ Collaboration with Martin Pelchat from BMI.

RNA Secondary Structure: Hairpins, Bulges, Loops, MBLs



$\Rightarrow$ 5S rRNA *Micrococcus luteus* (K02682)

RNA Secondary Structure: No Pseudo-Knot Definition

Let $a = a_1 a_2 \dots a_n$, be an RNA sequence, i.e. $a_i \in \{A, C, G, U\}$. The notation (a_i, a_j), for $i < j$ designates a pair. A secondary structure S for a is an ensemble of pairs, such that,

1. Watson-Crick: $(a_i, a_j) \in \{(A, U), (U, A), (G, C), (C, G)\}$;

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 $|j - i| \geq 4$.

$\Rightarrow$ $\{(G, U), (U, G)\}$ can form base pairs that are almost as stable as $\{(A, U), (U, A)\}$.

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.

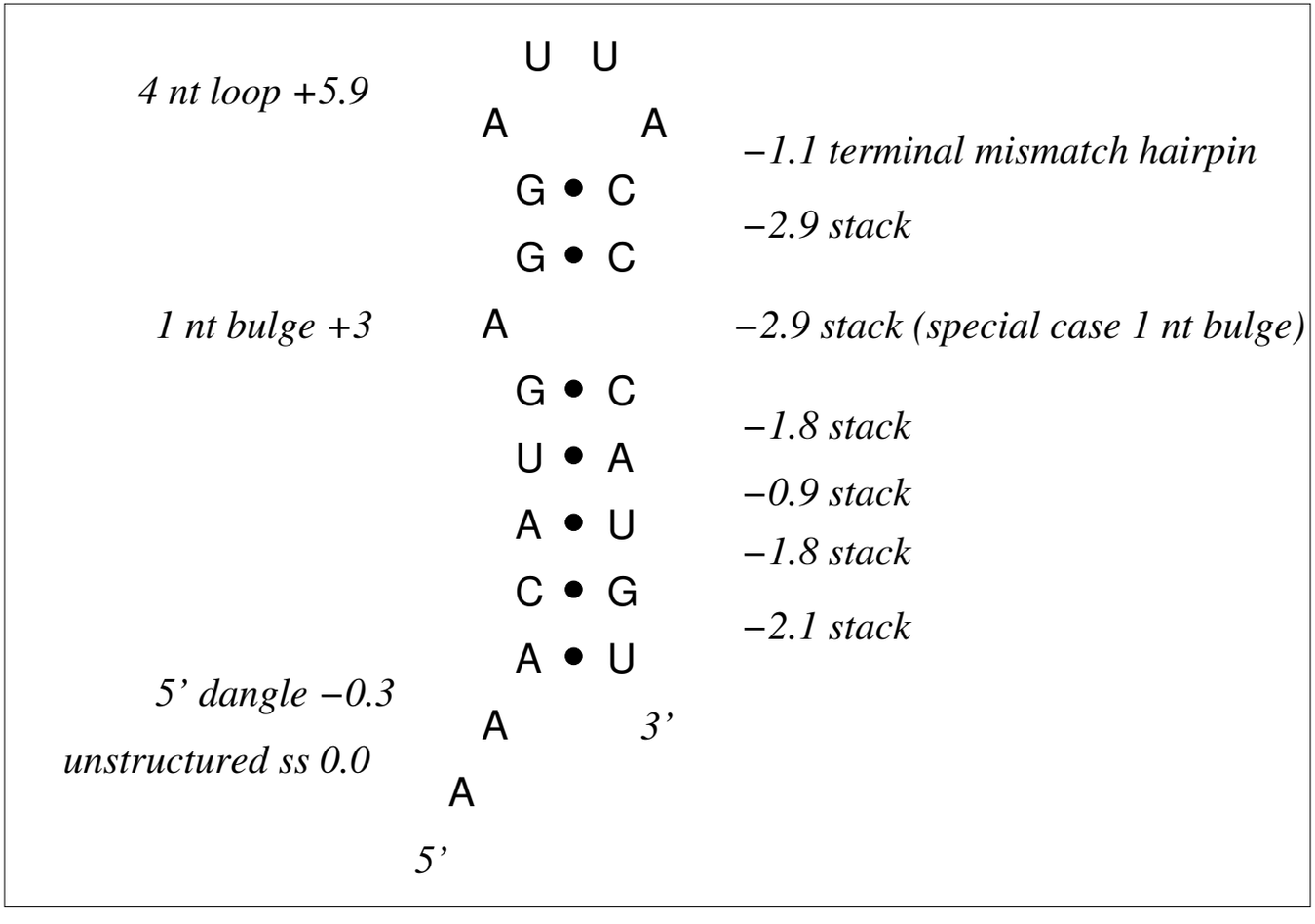
eXtended Dynalign

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 $S_3[1, n]$.

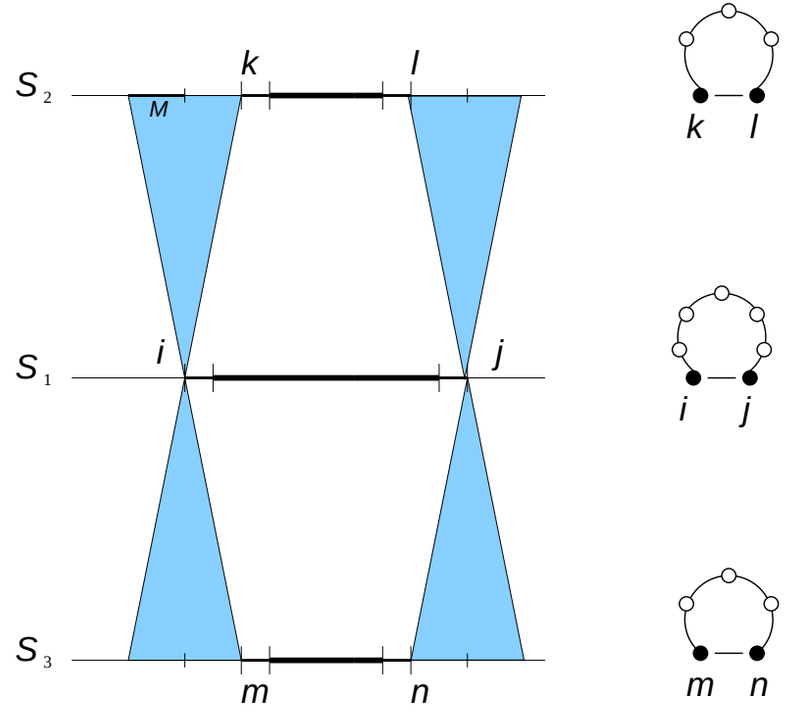
$\Rightarrow$ 140 cases: $V_1, V_2, V_{3_{1-64}}, W_1, W_2, W_{3_{1-64}}, W_{9_{1-8}}$.

Nearest-neighbor model[1]



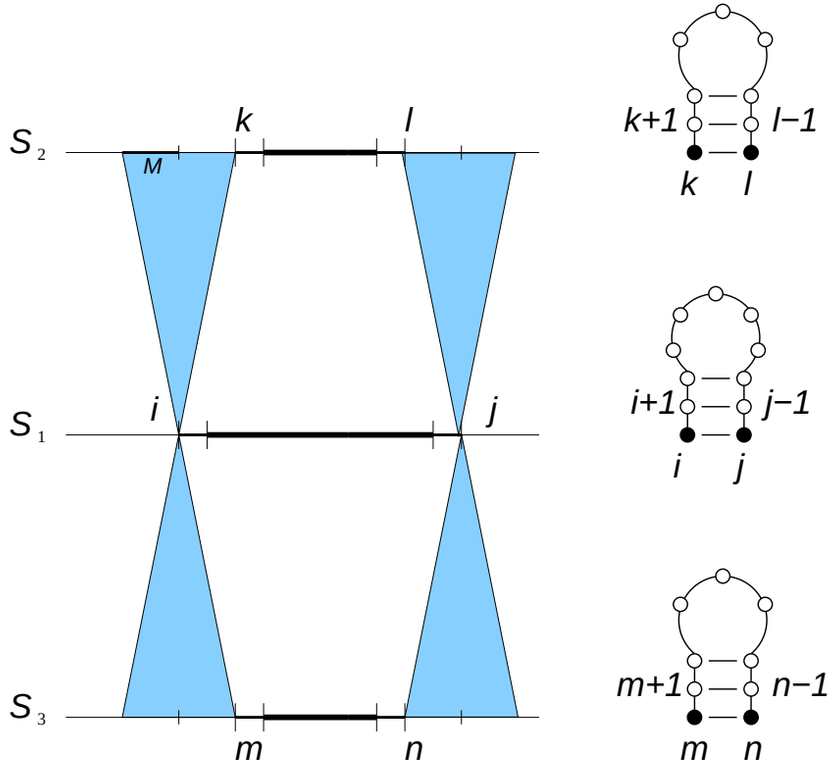
[1]Reproduced from Durbin *et al.* (1998) Biological Sequence Analysis. p. 275.

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



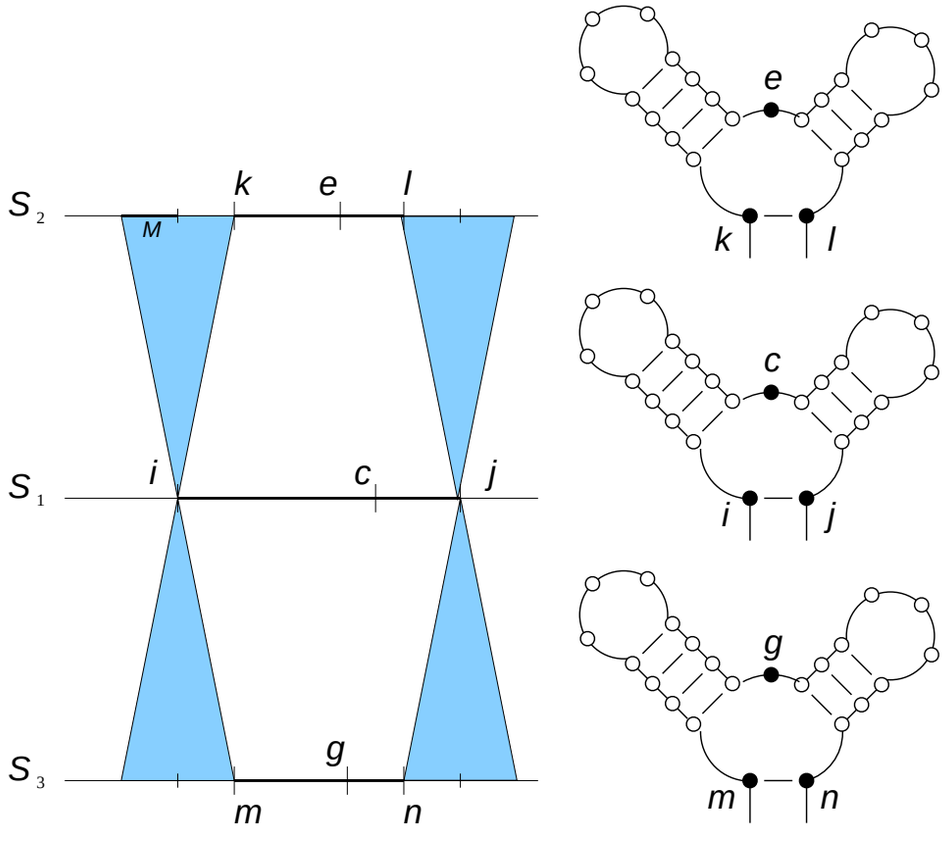
$$\Delta G^\circ_{\text{hairpin}}(i, j) + \Delta G^\circ_{\text{hairpin}}(k, l) + \Delta G^\circ_{\text{hairpin}}(m, n) + \Delta G^\circ_{\text{gap}}(\text{no. of gaps})$$

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



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

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

The nearest-neighbour model works reasonably well for small RNAs, 69 % and 71 % accuracy 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.

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 %.

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.

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

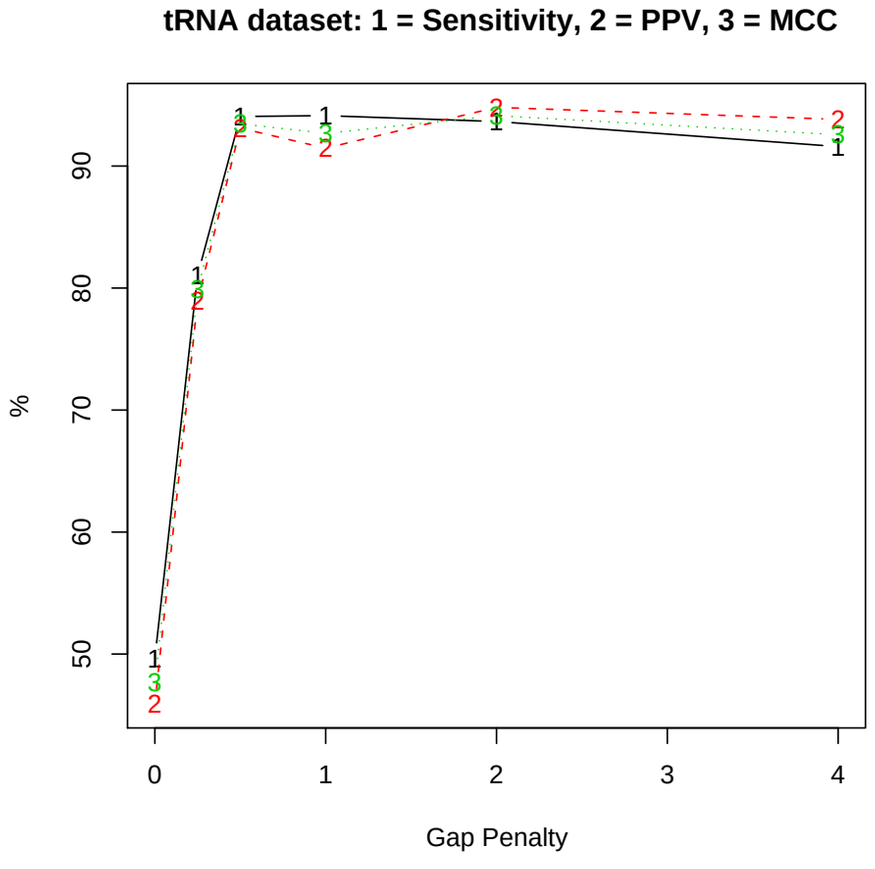
Are three input sequences better than two?

1. The worse prediction (minimum accuracy) should be more accurate;

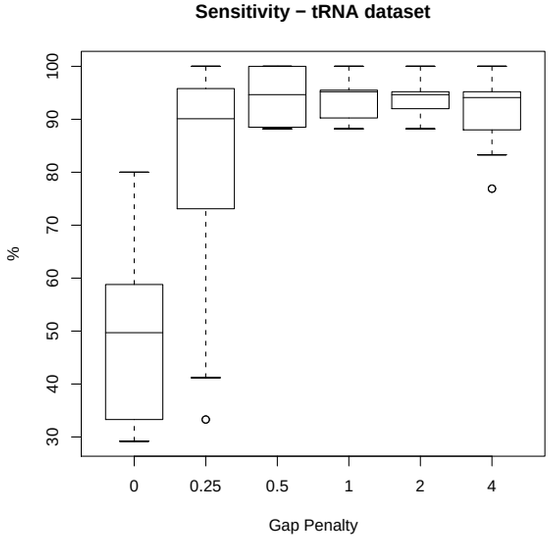
2. Use of three input sequences should improve the average accuracy;

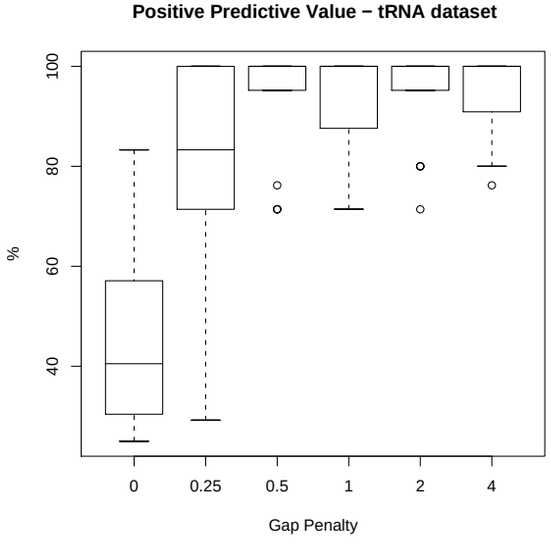
3. Average coverage should be less.

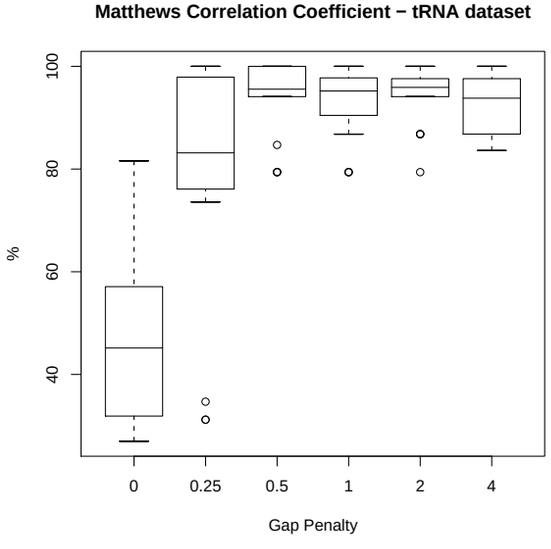
Calibrating Gap penalties: tRNAs



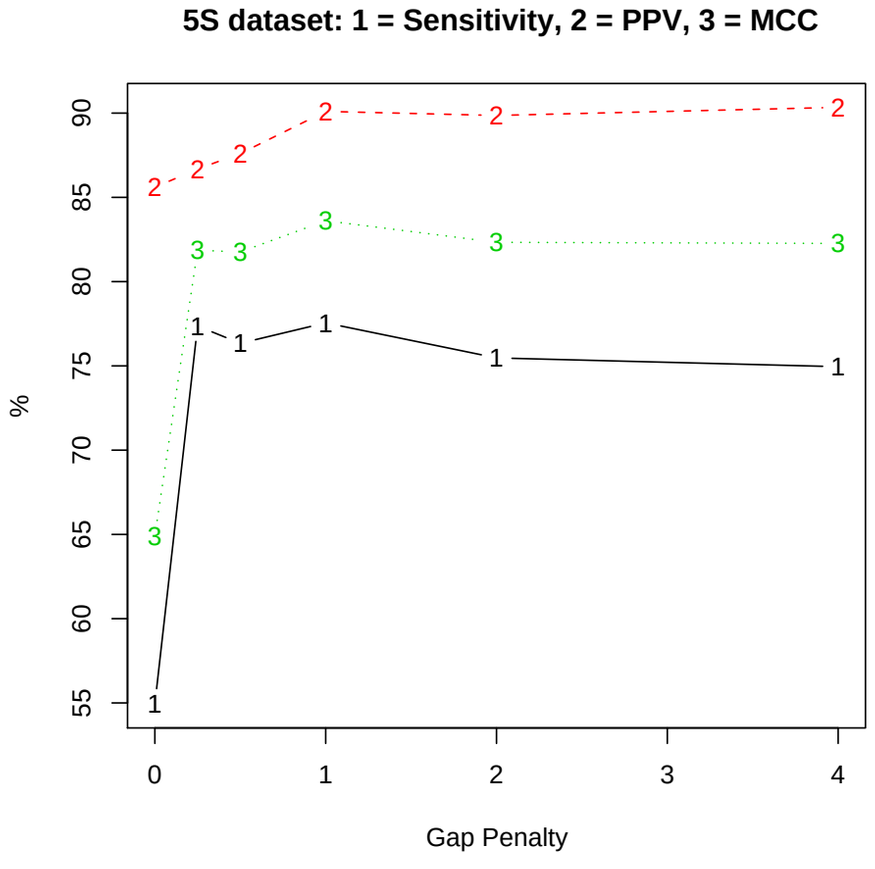
Calibrating Gap penalties: tRNAs



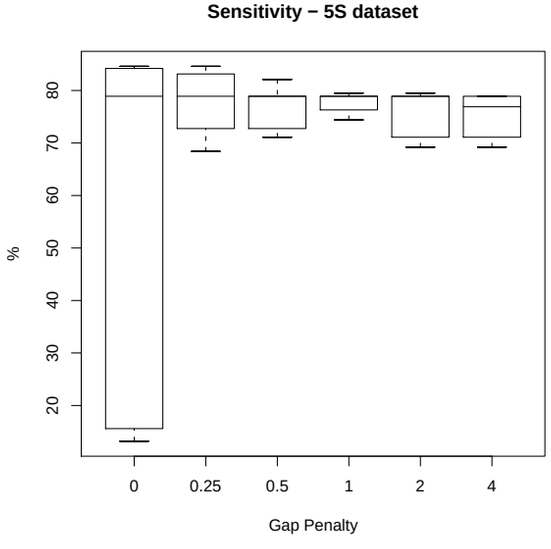


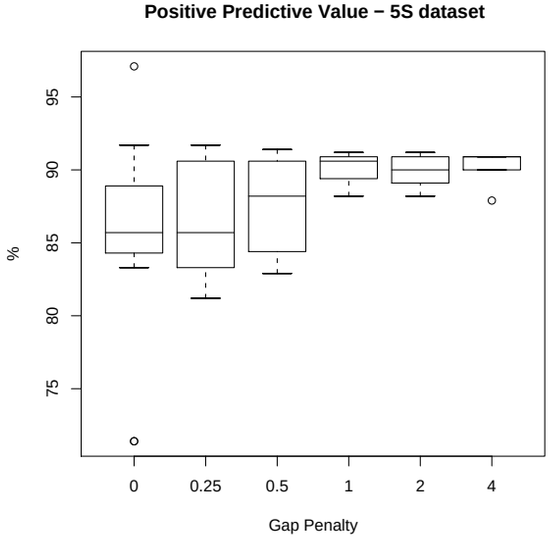


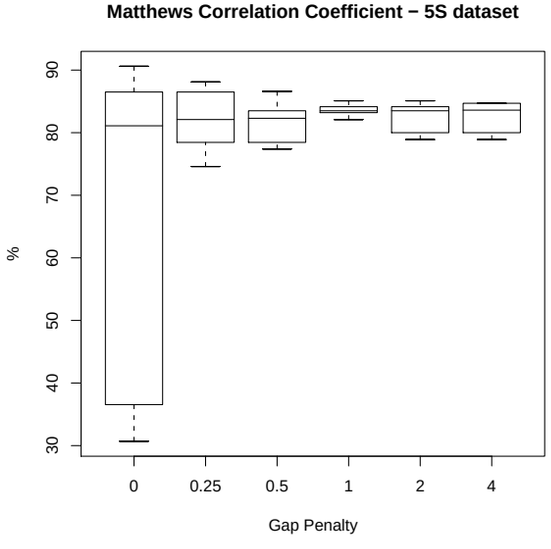
Calibrating Gap Penalties: 5S rRNAs



Calibrating Gap Penalties: 5S rRNAs







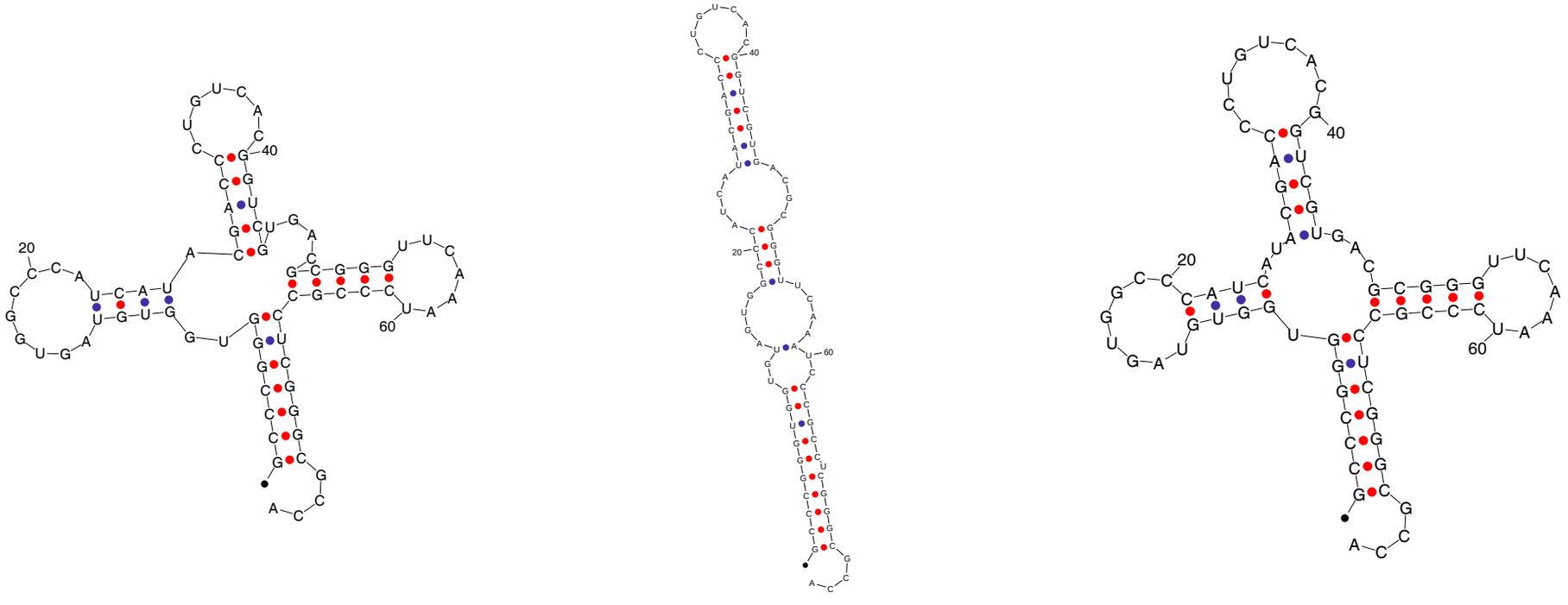
PPV: tRNA Dataset

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.

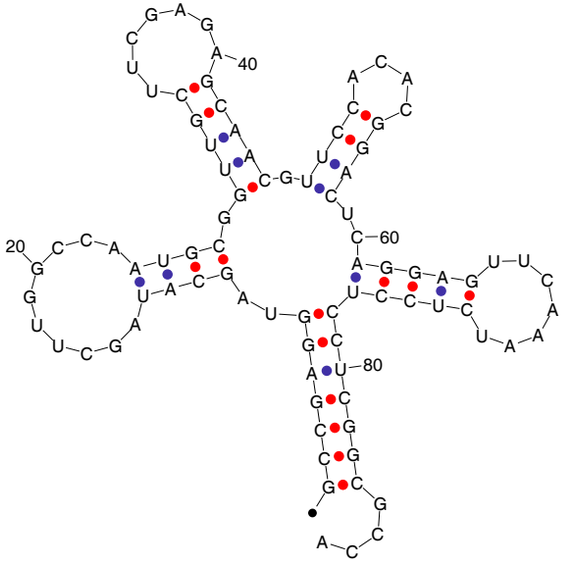
eXtended-Dynalign reproduces the cloverleaf structure



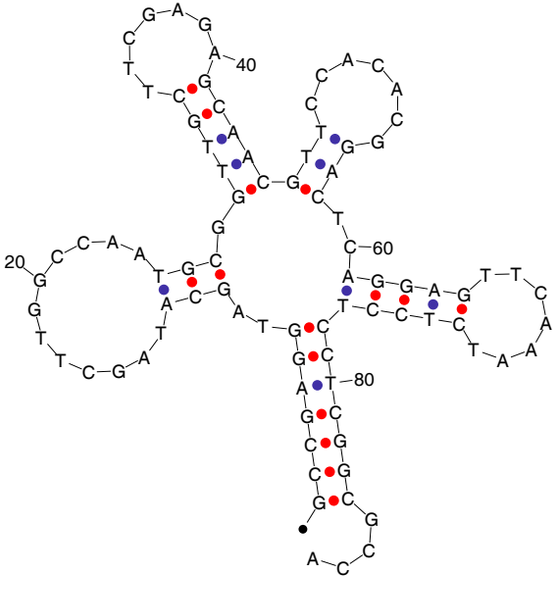
(a) (b) (c)

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

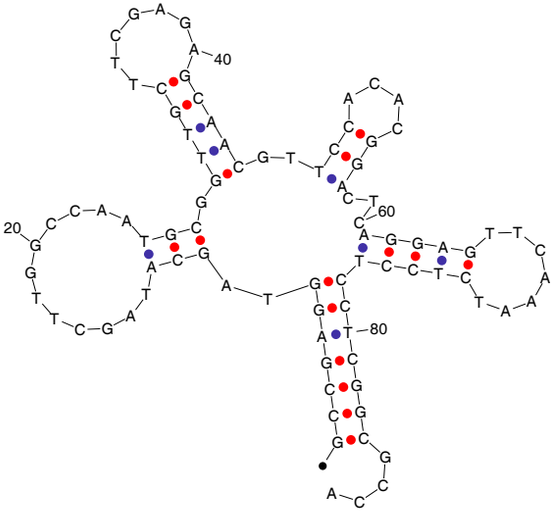
Fine details are better reproduced as well



(a) RS0380



(b) Dynalign



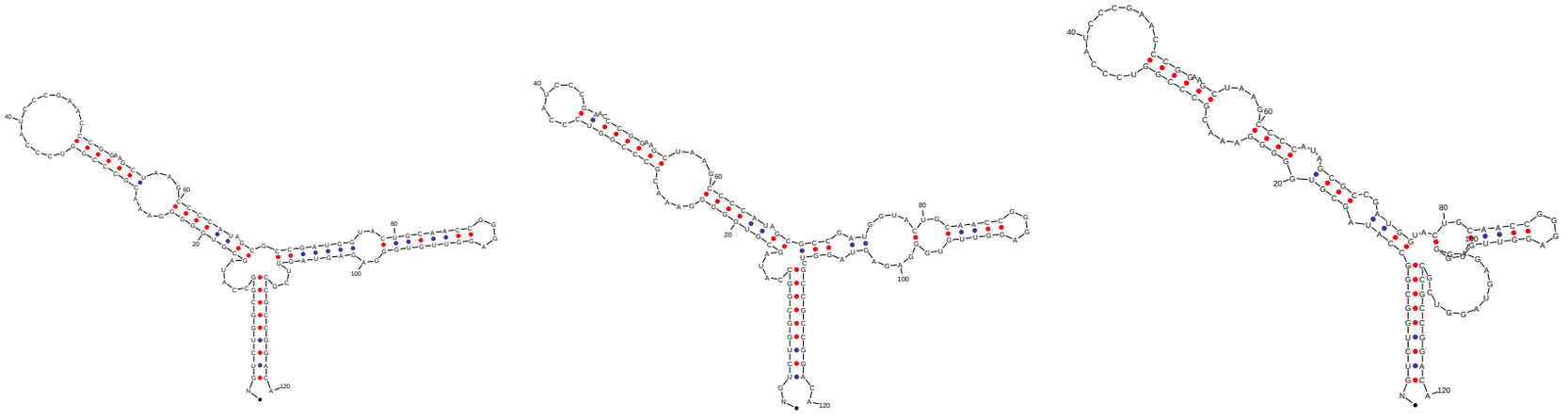
(c) X-Dynalign

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

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

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

Future Work

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

Collaborators

University of Ottawa

Beeta Masoumi (M.Sc. student)

Children's Hospital of Eastern Ontario (CHEO) — Molecular Genetics

Stephen Baird (Ph.D. student)
Martin Holcik (Group leader)
Robert Korneluk (Director)

University of Ottawa — Biochemistry, Microbiology and Immunology

Martin Pelchat (Group leader)

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)

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