PRESENTATION

Forbidden motifs and the cardinality of secondary structure space

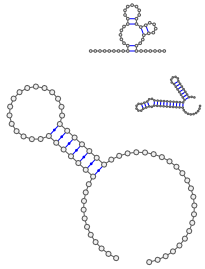
YAO, Hua-Ting

Ecole Polytechnique, France
McGill University, Canada

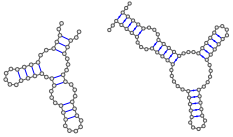
TBI, University of Vienna, Austria

Benasque — Aug 11, 2022

Phenotype Space



Secondary structure S^*



Genotype Space

CACGCGCUAUUCAACCUCUGC ...

UAGCGACGCUGAAUUCGACUCGAA ...

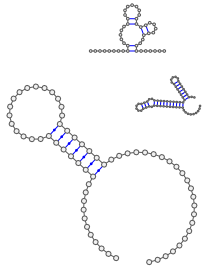
UAAUUUAAGAUGGCGGUGAA ...
UUUAAGAUAAACUGGGCGAA ...

RNA sequence w

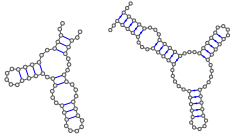
AAAAAA ...

GAACUAGCUAAAGCUUGGCGU ...

Phenotype Space



Secondary structure S^*



Genotype Space

CACGCGCUAUUCAACCUCUGC ...

UAGCGACGCUGAAUUCGACUCGAA ...

$\mathrm{MFE}(w) = S^*$

UAAUUUAAGAUGGCGGUGAA ...
UUUAAGAUAAACUGGGCGAA ...

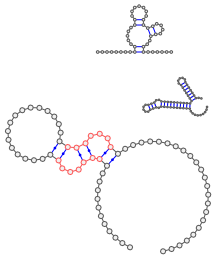
RNA sequence w

AAAAAAA ...

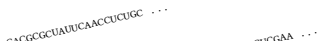
GAACUAGCUAAAGCUUGGCGU ...

Phenotype Space

Genotype Space

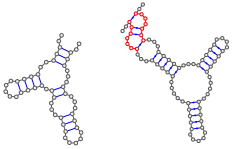


$\mathrm{MFE}(w) = S^*$



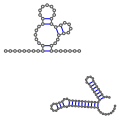
Secondary structure S^*

RNA sequence w



(Aguirre-Hernández *et al.*, 2007)

Phenotype Space



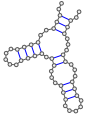
Genotype Space

CACGCGCUAUUCAACCUCUGC ...

UAGCGACGCUGAAUUCGACUCGAA ...

$$\mathrm{MFE}(w) = S^*$$

Secondary structure S^*

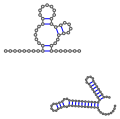


RNA sequence w

AAAAAAA ...

GAACUAGCUAAAGCUUGGCGU ...

Phenotype Space



Genotype Space

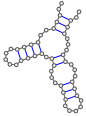
CACGCGCUAUUCAACCUCUGC ...

UAGCGACGCUGAAUUCGACUCGAA ...

$$\mathrm{MFE}(w) = S^*$$

(Design)

Secondary structure S^*

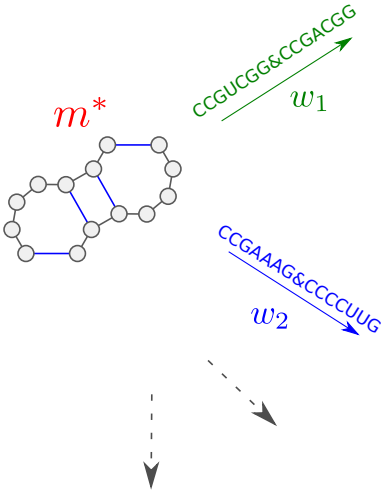


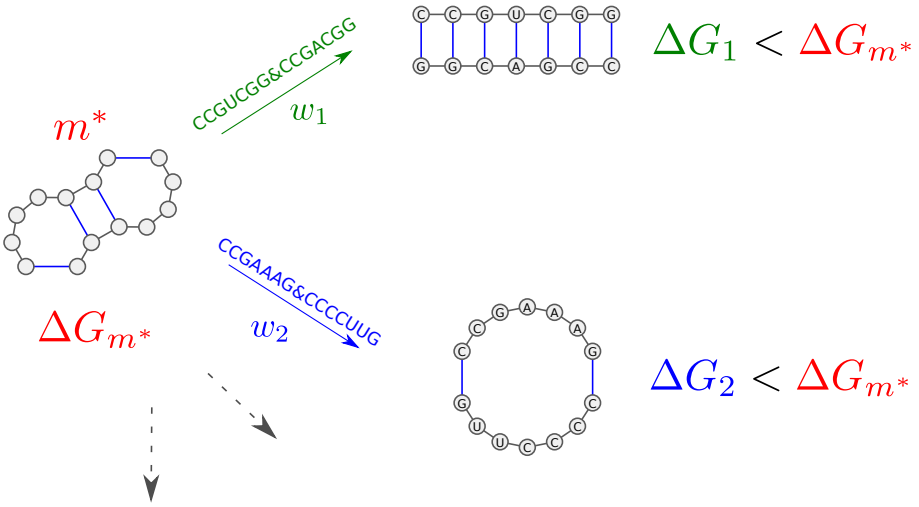
RNA sequence w

AAAAAAA ...

GAACUAGCUAAAGCUUGGCGU ...

Definition of Forbidden motif

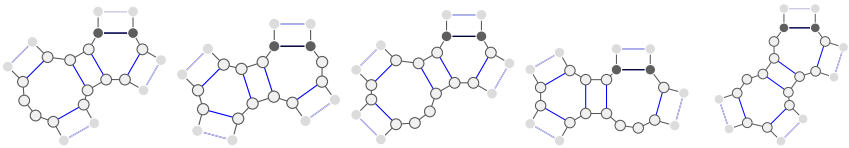




All sequences fold to alternative motifs

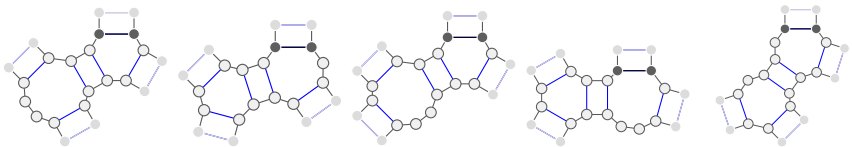
Forbidden Motifs of size up to 14

- Almost half of motifs are forbidden of size up to 14 ($4\,561$ out of $10\,886$)
- $2\,323$ forbidden motifs are minimal
- 63 do not contain isolated base pair



Forbidden Motifs of size up to 14

- Almost half of motifs are forbidden of size up to 14 ($4\,561$ out of $10\,886$)
- $2\,323$ forbidden motifs are minimal
- 63 do not contain isolated base pair



- Impact on the combinatorics of secondary structures

  $\rightarrow$ (Asymptotic) Proportion P_n of structures avoiding forbidden motifs decreases exponentially with the size n

Grammar generates secondary structures

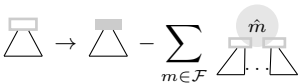
$$S \to \varepsilon \; + \; \bullet S \; + \; (\, S \,)\, S$$

Grammar generates secondary structures

$$S \to \varepsilon + \bullet S + (\,S\,)\,S$$

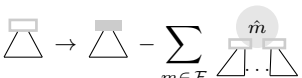
Grammar generates secondary structures $\mathcal{D}$ avoiding forbidden motif set $\mathcal{F}$

$$S \to \varepsilon + \bullet S + (\, S \,)\, S$$

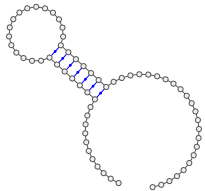


Grammar generates secondary structures $\mathcal{D}$ avoiding forbidden motif set $\mathcal{F}$

$$S \to \varepsilon + \bullet S + (\,S\,)\,S$$



			Proportion		
$\lvert\mathcal{F}\rvert$	$\lvert\mathcal{D}_n\rvert$	P_n	P_{100}	P_{500}	P_{1000}
387	$0.67\frac{2.242^n}{n\sqrt{n}}$	0.980^n	$1.19\,10^{-1}$	$2.98\,10^{-5}$	$9.40\,10^{-10}$

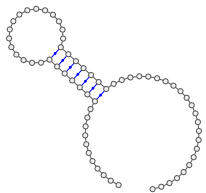


Secondary structure S^*

$\mathrm{MFE}(w) = S^*$

UAAUUUAAGAUGGCGGUGAA . . .
UUUAAGAUAAACUGGGCGAA . . .

RNA sequence w



Secondary structure S^*

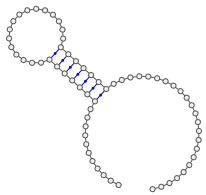
$$\mathbb{E}_w(\delta(S^*, \widehat{S})) \leq \varepsilon$$

$$\mathbb{P}_w(S^*) \geq \varepsilon$$

$$\mathrm{MFE}(w) = S^*$$

UAAUUUAAGAUGGCGGUGAA . . .
UUUAAGAUAAACUGGGCGAA . . .

RNA sequence w



Secondary structure S^*

$$\mathbb{E}_w(\delta(S^*, \widehat{S})) \leq \varepsilon$$

$$\mathbb{P}_w(S^*) \geq \varepsilon$$

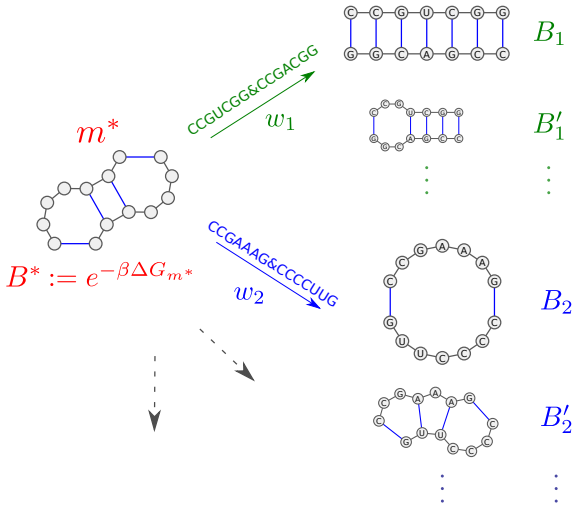
$$\mathrm{MFE}(w) = S^*$$

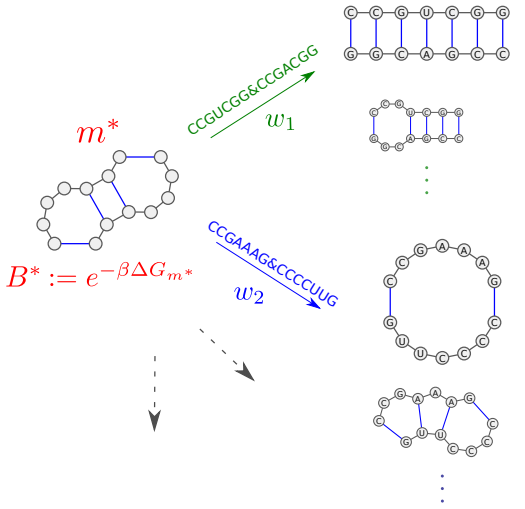
Probability defect

Ensemble defect

UAAUUUAAGAUGGCGGUGAA . . .
UUUAAGAUAAACUGGGCGAA . . .

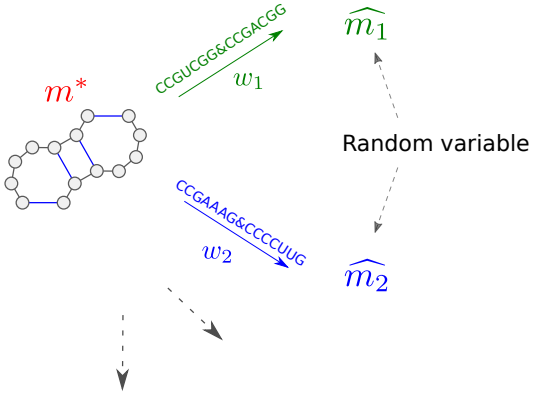
RNA sequence w

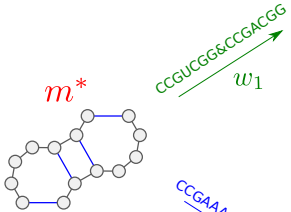

m^*
$B^* := e^{-\beta \Delta G_{m^*}}$
CCGUCGG&CCGACGG
w_1
CCGAAAG&CCCCUUG
w_2
B_1
B_1'
B_2
B_2'



$$\mathbb{P}_1^* := \frac{B^*}{B^* + B_1 + B_1' + \ldots} < \varepsilon$$

$$\mathbb{P}_2^* := \frac{B^*}{B^* + B_2 + B_2' + \ldots} < \varepsilon$$


m^*
CCGUCGG&CCGACGG
w_1
$\widehat{m_1}$
Random variable
CCGAAAG&CCCCUUG
w_2
$\widehat{m_2}$



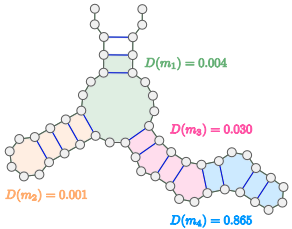
$\widehat{m_1}$ $\quad \mathbb{E}(\delta(m^*, \widehat{m_1})) > \varepsilon$

Random variable

$\widehat{m_2}$ $\quad \mathbb{E}(\delta(m^*, \widehat{m_2})) > \varepsilon$

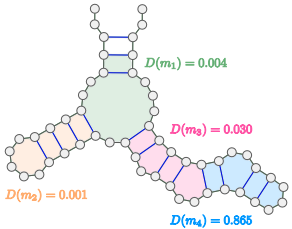
				Proportion		
	$\lvert\mathcal{F}\rvert$	$\lvert\mathcal{D}_n\rvert$	P_n	P_{100}	P_{500}	P_{1000}
not MFE	387	$0.67\frac{2.242^n}{n\sqrt{n}}$	0.980^n	$1.19\,10^{-1}$	$2.98\,10^{-5}$	$9.40\,10^{-10}$
$\mathbb{P} < 50\%$	401	$0.66\frac{2.239^n}{n\sqrt{n}}$	0.978^n	$1.03\,10^{-1}$	$1.53\,10^{-5}$	$2.49\,10^{-10}$
$\mathbb{E} > 1$	411	$0.65\frac{2.236^n}{n\sqrt{n}}$	0.977^n	$9.08\,10^{-2}$	$8.52\,10^{-6}$	$7.86\,10^{-11}$

- Lower bound of structural ensemble defect



$$\mathcal{D}(S) := \min_{w}(w, S) \geq \mathcal{D}(m_1) + \mathcal{D}(m_2) + \mathcal{D}(m_3) + \mathcal{D}(m_4) = 0.9$$

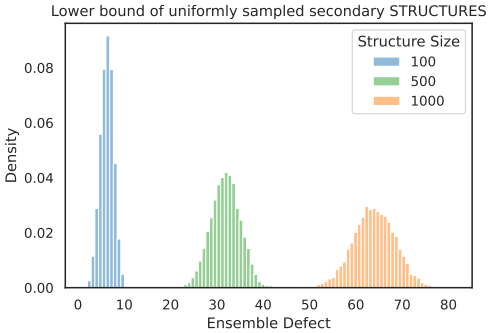
- Lower bound of structural ensemble defect



$$\mathcal{D}(S) := \min_{w}(w, S) \geq \mathcal{D}(m_1) + \mathcal{D}(m_2) + \mathcal{D}(m_3) + \mathcal{D}(m_4) = 0.9$$

- Variable tolerance grows with size n, *e.g.* $0.01n$ recommended in NUPACK (Zadeh *et al.*, 2011)

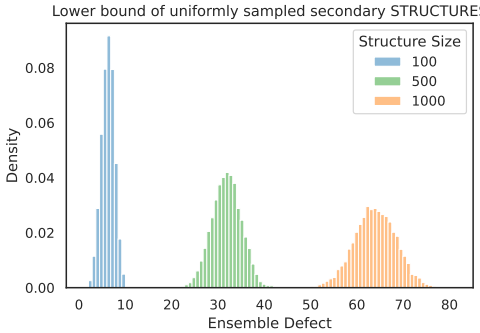
Ensemble defect lower bound distribution



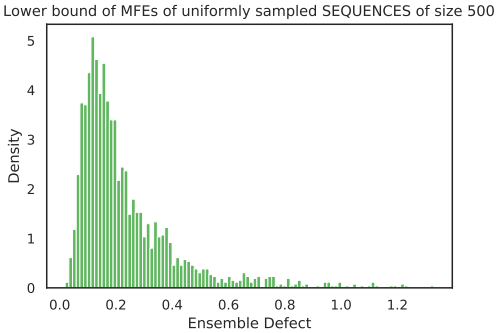
Random Structures

$\rightarrow$ With $\varepsilon = 0.01n$, $P_{1000} \approx 10^{-33}$

Ensemble defect lower bound distribution



Random Structures



MFE Structures

$\rightarrow$ With $\varepsilon = 0.01n$, $P_{1000} \approx 10^{-33}$

Conclusion

- Almost half of motifs are forbidden
- The cardinality of phenotype (structure) space is much smaller, but still exponentially grows
- Occurrences in PDBs shows a selection pressure on forbidden motifs

Conclusion

- Almost half of motifs are forbidden
- The cardinality of phenotype (structure) space is much smaller, but still exponentially grows
- Occurrences in PDBs shows a selection pressure on forbidden motifs
- Estimate how hard a phenotype can be realized
- Estimate the neuTral network size (number of designs)

Acknowledgement

AMIBio
Ecole Polytechnique

Waldispühl group
McGill University

TBI
University of Vienna

Cedric Chauve
Simon Fraser University