CIRCULAR CODES IN GENES AND GENOMES

- 2013 -

Prof. Christian MICHEL

Theoretical Bioinformatics
ICube
University of Strasbourg, CNRS
France

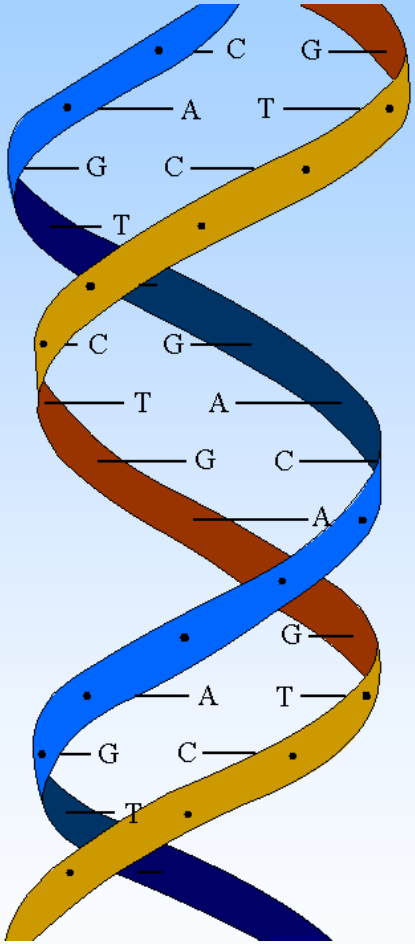
c.michel@unistra.fr
http://dpt-info.u-strasbg.fr/~c.michel/

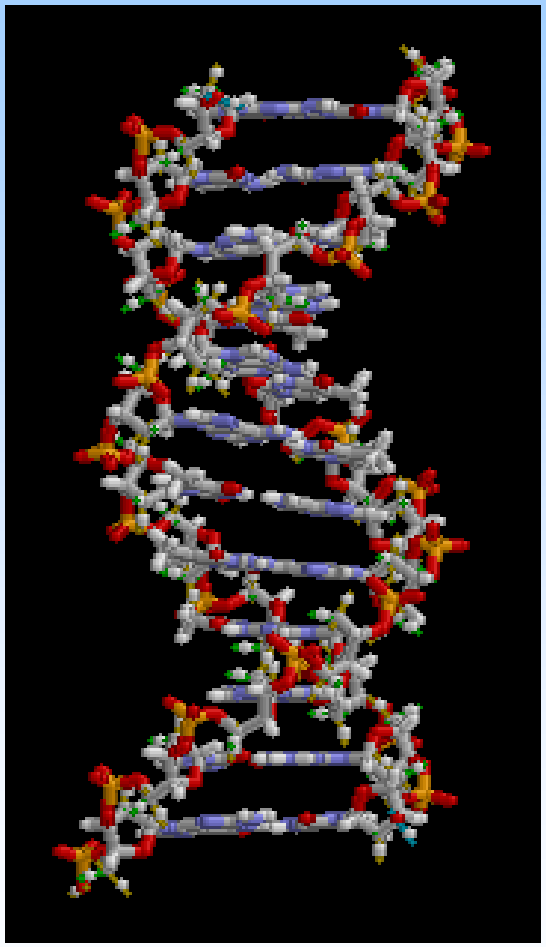
Biological recall: DNA

- Alphabet: $\boldsymbol{A}_4=\{A,C,G,T\}$
- Double helix
- Complementary pairing

  A – T and C – G

- Antiparallel





Biological recall: complementary trinucleotide

Complementary map: $\boldsymbol{C}$

Complementary nucleotide

$$\boldsymbol{C}(A) = T \text{ and } \boldsymbol{C}(T) = A$$

$$\boldsymbol{C}(C) = G \text{ and } \boldsymbol{C}(G) = C$$

Complementary trinucleotide

$$w_0 = \boldsymbol{l_0 l_1 l_2}$$

with $l_0 l_1 l_2 \in \boldsymbol{A}_4$, is

$$^3 \quad \boldsymbol{C}(w_0) = \boldsymbol{C}(\boldsymbol{l_2})\boldsymbol{C}(\boldsymbol{l_1})\boldsymbol{C}(\boldsymbol{l_0})$$

e.g. $\boldsymbol{C}(ACG)=CGT$

Extension to a complementary trinucleotide set

Biological recall: permuted trinucleotide

Permutation map: $\boldsymbol{P}$

Permuted trinucleotide

$$w_0 = \boldsymbol{l_0 l_1 l_2}$$

with $l_0 l_1 l_2 \in \boldsymbol{A}_4$, is 3

$$\boldsymbol{P}(w_0) = w_1 = \boldsymbol{l_1 l_2 l_0}$$

and

$$\boldsymbol{P}(\boldsymbol{P}(w_0)) = \boldsymbol{P}(w_1) = w_2 = \boldsymbol{l_2 l_0 l_1}$$

e.g. $\boldsymbol{P}(ACG)=CGA$ and $\boldsymbol{P}(\boldsymbol{P}(ACG))=\boldsymbol{P}(CGA)=GAC$

Extension to a permuted trinucleotide set

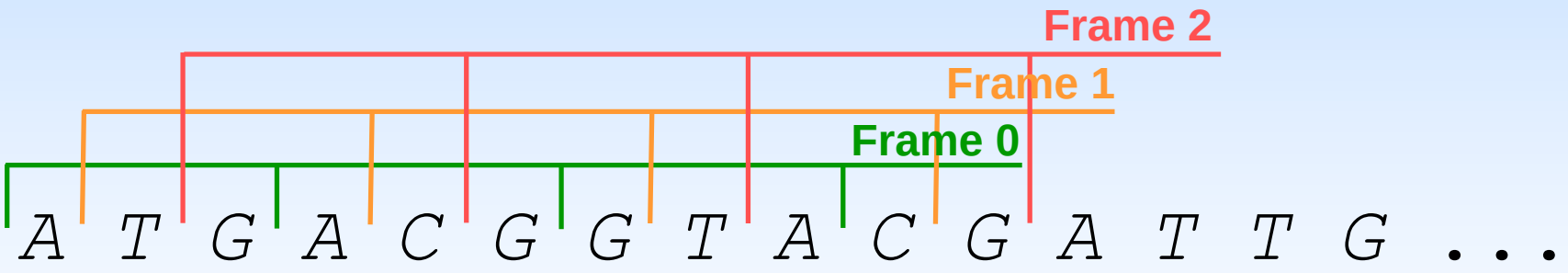
Biological recall: 3 frames in genes

Frame 0: Reading frame established by a start codon

{*ATG*,*GTG*,*TTG*}

Frame 1: Frame 0 shifted by 1 nucleotide in 5'-3'

Frame 2: Frame 0 shifted by 2 nucleotides in 5'-3'



Result 1: **The distribution of the 64 trinucleotides in the 3 frames of genes (prokaryotes, eukaryotes) are not uniform: 3 sets of trinucleotides are identified**

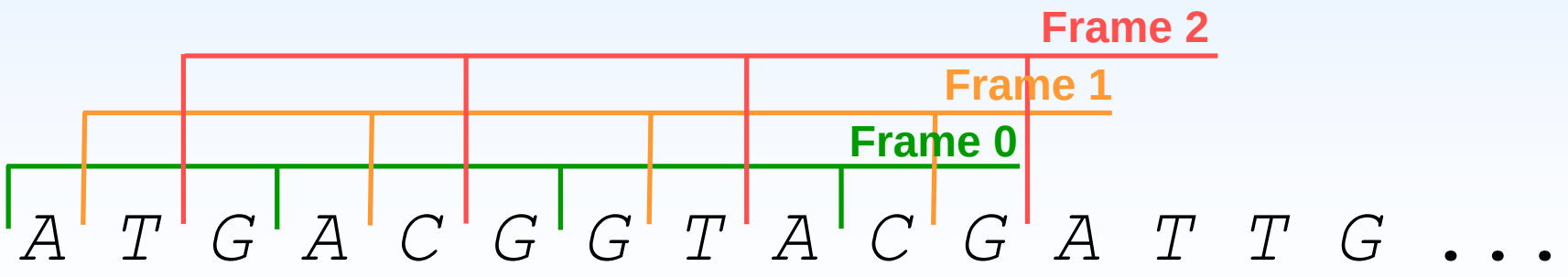
- Trinucleotide frequencies per frame (Arquès, Michel, 1996)
- Correlation functions per frame (Arquès, Michel, 1997)
- Frame permuted trinucleotide frequencies (Frey, Michel, 2003, 2006)
- Covering function (Gonzalez, Giannerini, Rosa, 2011)

Trinucleotide frequencies per frame

Occurrence frequencies $P(w^p)$ of the 64 trinucleotide w in each frame p in the prokaryotic protein coding genes (13686 sequences, 4708758 trinucleotides)

w in frame $p = 0$	Frequency (%)	w in frame $p = 1$	Frequency (%)	w in frame $p = 2$	Frequency (%)
AAA	3.38	AAA	2.75	AAA	2.44
AAC	2.18	AAC	1.59	AAC	1.38
AAG	1.98	**AAG**	3.21	AAG	0.81
AAT	2.17	AAT	1.37	AAT	1.69
ACA	1.22	**ACA**	1.91	ACA	1.11
ACC	2.09	ACC	1.60	ACC	0.79
ACG	1.30	**ACG**	2.49	ACG	0.68
ACT	1.13	**ACT**	1.17	ACT	1.09
AGA	0.61	AGA	1.59	**AGA**	2.47
AGC	1.42	**AGC**	1.83	AGC	1.71
AGG	0.31	**AGG**	2.21	AGG	1.45
AGT	0.87	AGT	0.97	**AGT**	1.26
ATA	0.83	**ATA**	2.15	ATA	0.66
ATC	2.61	ATC	1.66	ATC	0.82
ATG	2.38	**ATG**	2.82	ATG	0.41
ATT	2.50	ATT	1.38	ATT	1.50

Frame 0 **Frame 1** **Frame 2**



Identification of 3 sets of trinucleotides per frame in prokaryotes and eukaryotes

T_0	AAA	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC	TTT
T_1	AAG	ACA	ACG	ACT	AGC	AGG	ATA	ATG	CCA	CCC	CCG	GCG	GTG	TAG	TCA	TCC	TCG	TCT	TGC	TTA	TTG	
T_2	AGA	AGT	CAA	CAC	CAT	CCT	CGA	CGC	CGG	CGT	CTA	CTT	GCA	GCT	GGA	GGG	TAA	TAT	TGA	TGG	TGT	

Three subsets of trinucleotides can be identified: $T_0 = X_0 \cup \{\text{AAA,TTT}\}$ in frame $p = 0$, $T_1 = X_1 \cup \{\text{CCC}\}$ in frame $p = 1$ and $T_2 = X_2 \cup \{\text{GGG}\}$ in frame $p = 2$.

Result 2 (Arquès, Michel, 1996, 1997):
Mathematical properties of X_0, X_1 and X_2

Complementary property ***C***

T_0	AAA	AAC	AAT	ACC	ATC	CAG	CTC	GAA	GAC	GCC	GTA										
T_0	TTT	GTT	ATT	GGT	GAT	CTG	GAG	TTC	GTC	GGC	TAC										
T_1	AAG	ACA	ACG	ACT	AGC	AGG	ATA	ATG	CCA	CCC	CCG	GCG	GTG	TAG	TCA	TCC	TCG	TCT	TGC	TTA	TTG
T_2	CTT	TGT	CGT	AGT	GCT	CCT	TAT	CAT	TGG	GGG	CGG	CGC	CAC	CTA	TGA	GGA	CGA	AGA	GCA	TAA	CAA

C$(X_0) = X_0$: X_0 is self-complementary

C$(X_1) = X_2$ and ***C***$(X_2) = X_1$: X_1 and X_2 are complementary to each other

Permutation property ***P***

X_0	AAC	AAT	ACC	ATC	ATT	CAG	CTC	CTG	GAA	GAC	GAG	GAT	GCC	GGC	GGT	GTA	GTC	GTT	TAC	TTC	
X_1	ACA	ATA	CCA	TCA	TTA	AGC	TCC	TGC	AAG	ACG	AGG	ATG	CCG	GCG	GTG	TAG	TCG	TTG	ACT	ACT	
X_2	CAA	TAA	CAC	CAT	TAT	GCA	CCT	GCT	AGA	CGA	GGA	TGA	CGC	CGG	TGG	AGT	CGT	TGT	CTA	CTT	

P$(X_0) = X_1$ and ***P***$(X_1) = X_2$: X_1 and X_2 are deduced from X_0 by permutation

Result 2 (Arquès, Michel, 1996, 1997):
Mathematical properties of X_0, X_1 and X_2

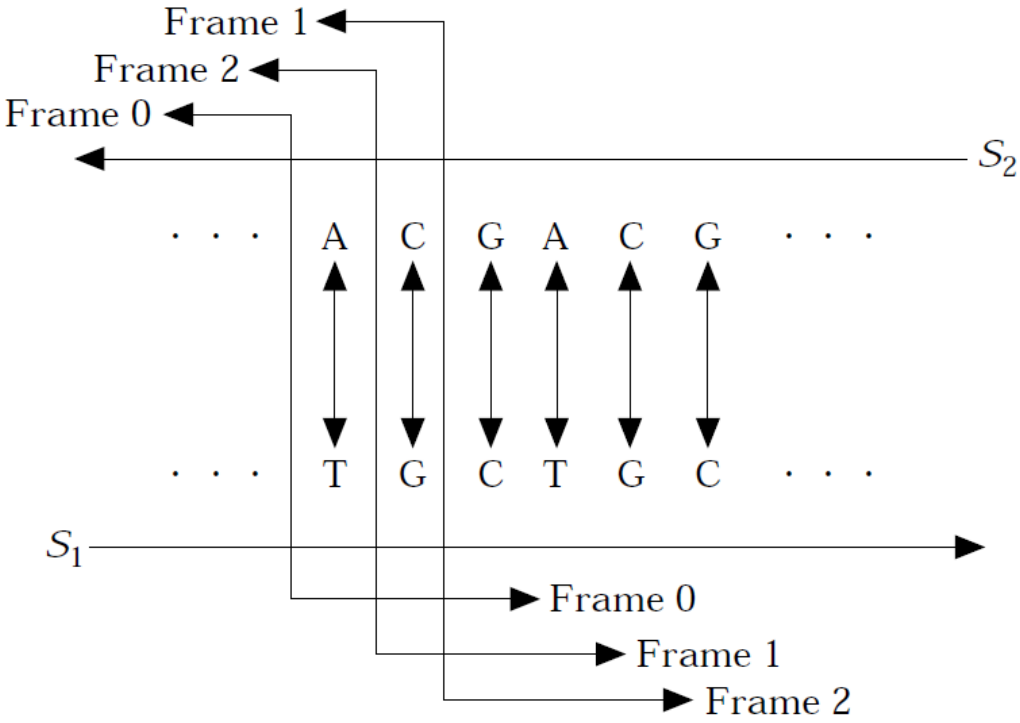


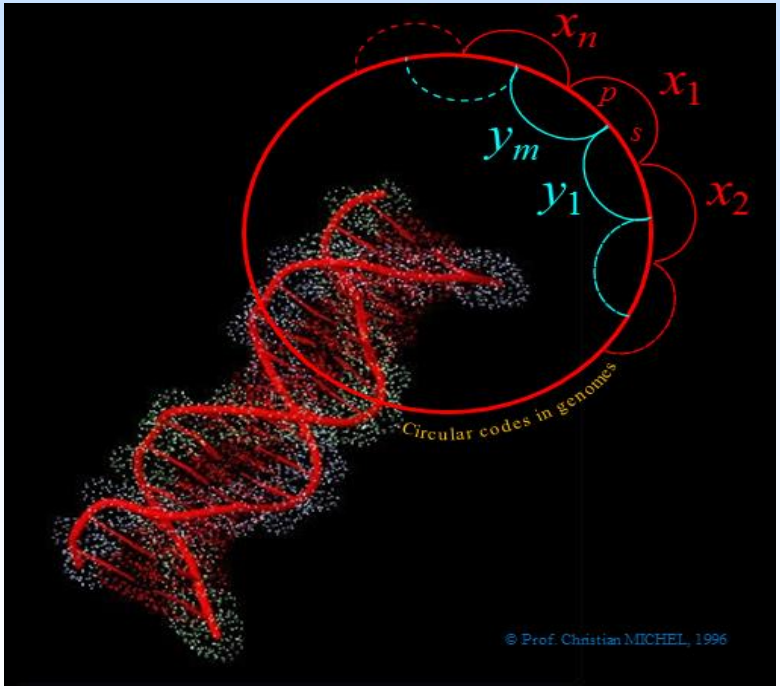
FIG. 4. The self-complementary circular code X_0 allows the two paired frames 0 (reading frames) simultaneously to code for amino acids without using a start codon.

Result 3 (Arquès, Michel, 1996, 1997):
X_0, X_1 and X_2 are trinucleotide circular codes

X_0 is able to retrieve the reading frame 0

X_1 is able to retrieve the frame 1

X_2 is able to retrieve the frame 2



Code, comma-free code

Definition 2.1. Code: a subset X *of* $\mathcal{A}^+$ is a code over $\mathcal{A}$ if for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, the condition $x_1 \cdots x_n = x'_1 \cdots x'_m$ implies $n = m$ and $x_i = x'_i$ for $i = 1, \ldots, n$.

$Y = \{A, GC, AGC\}$ is not a code as $A \bullet GC = AGC$

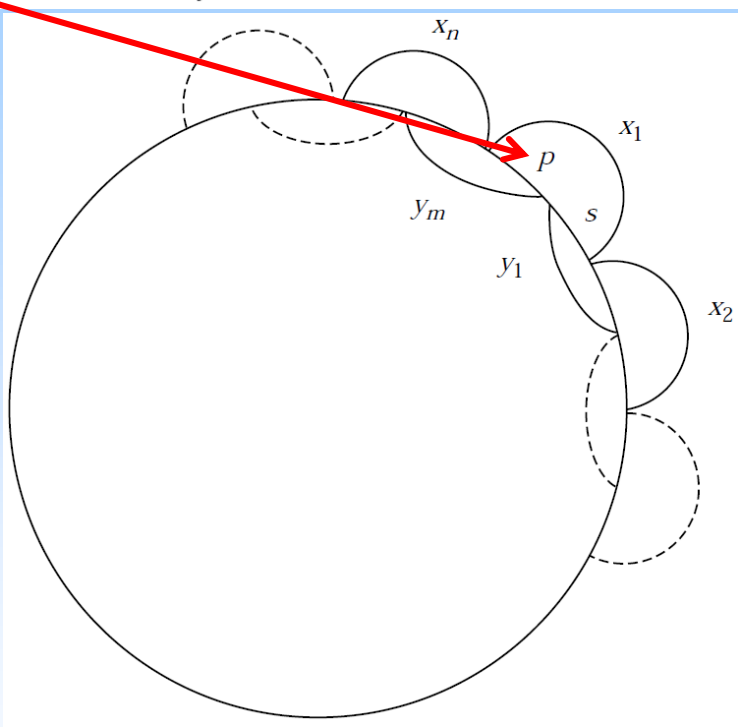
$\mathcal{A}_4^3 = \{AAA, \ldots, TTT\}$ (genetic code) is a code.

Definition 2.2. Trinucleotide comma-free code: a trinucleotide code $X \subset \mathcal{A}_4^3$ is comma-free if for each $y \in X$ and $u, v \in \mathcal{A}_4^*$ such that $uyv = x_1 \cdots x_n$ with $x_1, \ldots, x_n \in X$, $n \geq 1$, it results that $u, v \in X^*$.

$\mathcal{A}_4^3$ is not a comma-free code: $A \bullet CGA \bullet CG = ACG \bullet ACG$

Circular code

Definition 2.3. Trinucleotide circular code: a trinucleotide code $X \subset \mathcal{A}_4^3$ is circular if for each $x_1, \ldots, x_n, x'_1, \ldots, x'_m \in X$, $n, m \geq 1$, $p \in \mathcal{A}_4^*$, $s \in \mathcal{A}_4^+$, the conditions $sx_2 \cdots x_n p = x'_1 \cdots x'_m$ and $x_1 = ps$ imply $n = m$, $p = \varepsilon$ and $x_i = x'_i$ for $i = 1, \ldots, n$.



Definition 2.5. Maximal trinucleotide circular code: a trinucleotide circular code $X \subset \mathcal{A}_4^3$ is maximal if for each $x \in \mathcal{A}_4^3$, $x \notin X$, $X \cup \{x\}$ is not a trinucleotide circular code.

For words of length 3 over a 4-letter alphabet (trinucleotides), the maximal length of circular codes is 20 words

Circular code: proof

- Flower automaton (Lassez, 1976; Berstel, Perrin, 1985; Arquès, Michel, 1996, 1997)
- Necklaces 5*LDCN* (Letter Diletter Continued Necklace) (Pirillo, 2003) and *nLDCCN* (Letter Diletter Continued Closed Necklace) with $n \in \{2,3,4,5\}$ (Michel, Pirillo, 2010)

Result 4 (Lacan, Michel, 2001):
Proof that the probabilistic model based on the nucleotide frequencies (Koch, Lehmann, 1997) is incomplete for constructing circular codes, in particular it cannot generate X_0 (cannot generate the trinucleotides $\alpha\beta\gamma$, $\delta\delta\beta$ and $\gamma\alpha\delta$)

Circular code

The decomposition of any word of a circular code *Y* written on a circle is unique

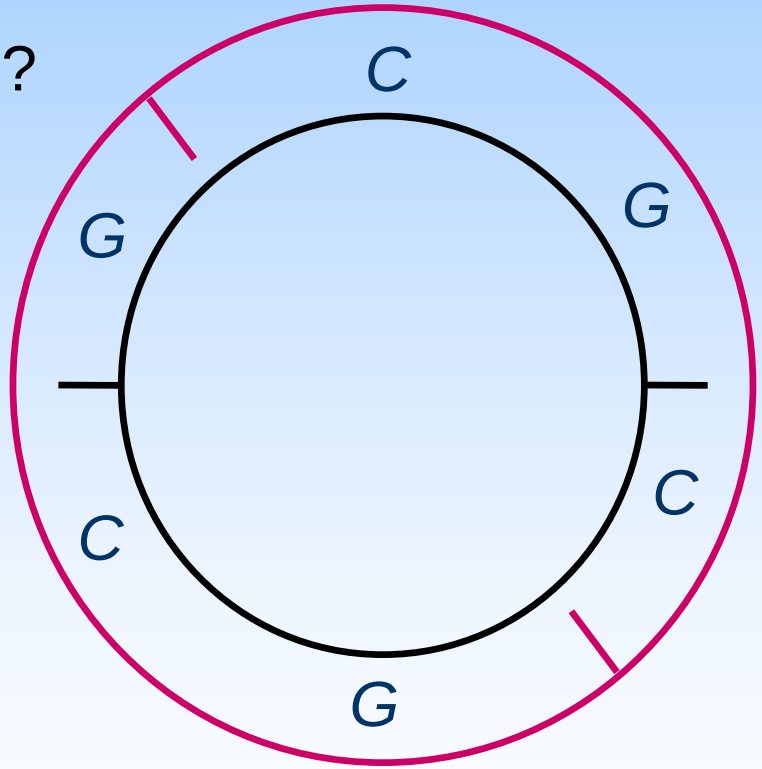
Is *Y* = {*GCG*,*CGC*} a circular code ?

2 decompositions:

w = *GCG* • *CGC*

w = *CGC* • *GCG*

Y is not a circular code



Y = {*GGC*,*CGG*} is a circular code

Circular code

$$X_0 = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC$$
$$GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$$

Generation of a word from the circular code X_0

Circular code

$$X_0 = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC$$
$$GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$$

- Generation of a word from the circular code X_0

…*GAA*,*GAG*,*GTA*,*GTA*,*ACC*,*AAT*,*GTA*,*CTC*,*TAC*,*TTC*,*ACC*,*ATC*…

- Then, the trinucleotides in frame 1 **mainly** belong to X_1

…*G*,*AAG*,*AGG*,*TAG*,*TAA*,*CCA*,*ATG*,*TAC*,*TCT*,*ACT*,*TCA*,*CCA*,*TC*…

$\in X_2$ $\in X_0$

- Then, the trinucleotides in frame 2 **mainly** belong to X_2

…*GA*,*AGA*,*GGT*,*AGT*,*AAC*,*CAA*,*TGT*,*ACT*,*CTA*,*CTT*,*CAC*,*CAT*,*C*…

$\in X_0$ $\in X_0$ $\in X_1$

In frame 1: 75.4 % of X_1 , 11.9 % of X_0 and 12.7 % of X_2
In frame 2: 75.4 % of X_2 , 11.9 % of X_0 and 12.7 % of X_1

In a comma-free code Y_0, the trinucleotides of Y_0 do not occur in the shifted frames 1 and 2

Result 5 (Arquès, Michel, 1996, 1997): X_0 is a C^3 self-complementary trinucleotide circular code

- X_0 , $X_1 = \boldsymbol{P}(X_0)$ and $X_2 = \boldsymbol{P}(X_1)$ are maximal (20) trinucleotide circular codes
- $\boldsymbol{C}(X_0) = X_0$, $\boldsymbol{C}(X_1) = X_2$ and $\boldsymbol{C}(X_2) = X_1$

Remark: if X_0 is a circular code then $X_1 = \boldsymbol{P}(X_0)$ and $X_2 = \boldsymbol{P}(X_1)$ are not necessarily circular codes

Result 6 (Michel, Pirillo, Pirillo, 2008):
Growth function of comma-free codes

l	1	2	3	4	5	6	7	8	9	10
Nb(l)	60	1656	25608	244008	1530060	6638340	20708460	47742654	82816632	109358220

11	12	13	14	15	16	17	18	19	20
110895036	87031844	53227980	25473732	9519912	2743080	591864	90420	8760	408

Result 7 (Michel, Pirillo, Pirillo, 2008):
Growth function of C^3 self-complementary comma-free codes

l	2	4	6	8	10	12	14	16	18	20
Nb(l)	28	182	424	498	340	144	36	4	0	0

Result 8 (Michel, Pirillo, 2010):
Growth function of circular codes

l	1	2	3	4	5	6	7	8	9	10
Nb(l)	60	1,704	30,432	382,164	3,568,212	25,507,512	141,639,780	614,568,102	2,086,742,208	5,542,646,244

l	11	12	13	14	15	16	17	18	19	20
Nb(l)	11,503,061,124	18,615,667,124	23,403,485,556	22,700,634,924	16,787,523,072	9,279,022,320	3,708,717,048	1,012,099,740	168,726,792	12,964,440

Result 9 (Arquès, Michel, 1996, 1997):
Classes of maximal (20) circular codes

Number of potential circular codes: $\mathbf{3^{20} = 3\ 486\ 784\ 401}$

Number of circular codes: **12 964 440**

Number of C^3 codes: **221 544**

Number of C^3 self-complementary codes: **216**

Occurrence probability of X_0 in genes: $216 / 3^{20} = \mathbf{6.2 \times 10^{-8}}$

Result 10 (Michel, Pirillo, 2011; Michel, Pirillo, Pirillo, 2012):

Hierarchy of maximal (20) circular codes

I^2	I^3C	I^3	I^4C	I^4	I^5C	I^5
408	2, 760	297, 072	550, 032	5, 116, 728	5, 940, 648	12, 964, 440

$$I^2 \subset I^3C \subset I^3 \subset I^4C \subset I^4 \subset I^5C \subset I^5$$

Strong: comma-free codes

Weak

Result 11 (Bussoli, Michel, Pirillo, 2011, 2012): **Self-complementary maximal (20) circular codes**

Proposition 7. *A trinucleotide circular code X_0 having 20 elements is self-complementary if and only if X_1 and X_2 are complement of each other.*

Proposition 8. *If a trinucleotide circular code X_0 having 20 elements is self-complementary then either*

1) *X_1 and X_2 are both circular codes*

or

2) *X_1 and X_2 are not circular codes*

Result 12 (Benard, Michel, 2013):

Transversion II on the three positions of any subset of trinucleotides of the circular code X_0 yields no circular code

Definition 32. The transversion II evolution genetic map $\mathscr{V}_{\mathrm{II}}: \mathscr{A}_4^+ \to \mathscr{A}_4^+$ is defined by

$$\begin{aligned} \mathscr{V}_{\mathrm{II}}(A) &= C, & \mathscr{V}_{\mathrm{II}}(C) &= A, \\ \mathscr{V}_{\mathrm{II}}(G) &= T, & \mathscr{V}_{\mathrm{II}}(T) &= G. \end{aligned} \tag{11}$$

$\mathscr{V}_{\mathrm{II}}^{1,2,3}$ is the transversion II on the three positions of x

Number $c(\mathscr{V}_{\mathrm{II}}^{1,2,3}(l))$ of circular codes

For $l = 1, \ldots, 19$

$$c\left(\mathscr{V}_{\mathrm{II}}^{1,2,3}(l)\right) = 0$$

Result 13 (Benard, Michel, 2013):

Transversion I on the 2nd position of any subset of trinucleotides of the circular code X_0 yields to circular codes which are always C^3

Definition 28. The transversion I evolution genetic map $\mathscr{V}_{\mathrm{I}}: \mathscr{A}_4^+ \rightarrow \mathscr{A}_4^+$ is defined by

$$\begin{aligned} \mathscr{V}_{\mathrm{I}}(A) &= T, & \mathscr{V}_{\mathrm{I}}(C) &= G, \\ \mathscr{V}_{\mathrm{I}}(G) &= C, & \mathscr{V}_{\mathrm{I}}(T) &= A. \end{aligned} \tag{10}$$

Result 14 (Michel, Pirillo, 2013): **Dinucleotide circular codes**

Proposition 29. *Let* (i, j, h, k) *be a permutation of* (A, C, G, T). *If*

$$X = \{ij, ih, ik, jh, jk, hk\}, \quad (1)$$

then X *is a dinucleotide circular code.*

Proposition 36. *There are 24 maximum dinucleotide circular codes.*

Result 15 (Michel, Pirillo, 2013):
List of the 24 dinucleotide circular codes

Symbol	Dinucleotide circular code	$\mathscr{C}$	$\mathscr{P}$	$\mathscr{PC}$
X_1	$\{AC, AG, AT, CG, CT, GT\}$	$\mathscr{C}(X_1) = X_1$	$\mathscr{P}(X_1) = X_{24}$	$\mathscr{P}(\mathscr{C}(X_1)) = X_{24}$
X_2	$\{AC, AG, AT, CG, CT, TG\}$	$\mathscr{C}(X_2) = X_{13}$	$\mathscr{P}(X_2) = X_{23}$	$\mathscr{P}(\mathscr{C}(X_2)) = X_{12}$
X_3	$\{AC, AG, AT, CG, TC, TG\}$	$\mathscr{C}(X_3) = X_{17}$	$\mathscr{P}(X_3) = X_{22}$	$\mathscr{P}(\mathscr{C}(X_3)) = X_8$
X_4	$\{AC, AG, AT, CT, GC, GT\}$	$\mathscr{C}(X_4) = X_4$	$\mathscr{P}(X_4) = X_{21}$	$\mathscr{P}(\mathscr{C}(X_4)) = X_{21}$
X_5	$\{AC, AG, AT, GC, GT, TC\}$	$\mathscr{C}(X_5) = X_9$	$\mathscr{P}(X_5) = X_{20}$	$\mathscr{P}(\mathscr{C}(X_5)) = X_{16}$
X_6	$\{AC, AG, AT, GC, TC, TG\}$	$\mathscr{C}(X_6) = X_{18}$	$\mathscr{P}(X_6) = X_{19}$	$\mathscr{P}(\mathscr{C}(X_6)) = X_7$
X_7	$\{AC, AG, CG, TA, TC, TG\}$	$\mathscr{C}(X_7) = X_{19}$	$\mathscr{P}(X_7) = X_{18}$	$\mathscr{P}(\mathscr{C}(X_7)) = X_6$
X_8	$\{AC, AG, GC, TA, TC, TG\}$	$\mathscr{C}(X_8) = X_{22}$	$\mathscr{P}(X_8) = X_{17}$	$\mathscr{P}(\mathscr{C}(X_8)) = X_3$
X_9	$\{AC, AT, CT, GA, GC, GT\}$	$\mathscr{C}(X_9) = X_5$	$\mathscr{P}(X_9) = X_{16}$	$\mathscr{P}(\mathscr{C}(X_9)) = X_{20}$
X_{10}	$\{AC, AT, GA, GC, GT, TC\}$	$\mathscr{C}(X_{10}) = X_{10}$	$\mathscr{P}(X_{10}) = X_{15}$	$\mathscr{P}(\mathscr{C}(X_{10})) = X_{15}$
X_{11}	$\{AC, GA, GC, GT, TA, TC\}$	$\mathscr{C}(X_{11}) = X_{11}$	$\mathscr{P}(X_{11}) = X_{14}$	$\mathscr{P}(\mathscr{C}(X_{11})) = X_{14}$
X_{12}	$\{AC, GA, GC, TA, TC, TG\}$	$\mathscr{C}(X_{12}) = X_{23}$	$\mathscr{P}(X_{12}) = X_{13}$	$\mathscr{P}(\mathscr{C}(X_{12})) = X_2$
X_{13}	$\{AG, AT, CA, CG, CT, GT\}$	$\mathscr{C}(X_{13}) = X_2$	$\mathscr{P}(X_{13}) = X_{12}$	$\mathscr{P}(\mathscr{C}(X_{13})) = X_{23}$
X_{14}	$\{AG, AT, CA, CG, CT, TG\}$	$\mathscr{C}(X_{14}) = X_{14}$	$\mathscr{P}(X_{14}) = X_{11}$	$\mathscr{P}(\mathscr{C}(X_{14})) = X_{11}$
X_{15}	$\{AG, CA, CG, CT, TA, TG\}$	$\mathscr{C}(X_{15}) = X_{15}$	$\mathscr{P}(X_{15}) = X_{10}$	$\mathscr{P}(\mathscr{C}(X_{15})) = X_{10}$
X_{16}	$\{AG, CA, CG, TA, TC, TG\}$	$\mathscr{C}(X_{16}) = X_{20}$	$\mathscr{P}(X_{16}) = X_9$	$\mathscr{P}(\mathscr{C}(X_{16})) = X_5$
X_{17}	$\{AT, CA, CG, CT, GA, GT\}$	$\mathscr{C}(X_{17}) = X_3$	$\mathscr{P}(X_{17}) = X_8$	$\mathscr{P}(\mathscr{C}(X_{17})) = X_{22}$
X_{18}	$\{AT, CA, CT, GA, GC, GT\}$	$\mathscr{C}(X_{18}) = X_6$	$\mathscr{P}(X_{18}) = X_7$	$\mathscr{P}(\mathscr{C}(X_{18})) = X_{19}$
X_{19}	$\{CA, CG, CT, GA, GT, TA\}$	$\mathscr{C}(X_{19}) = X_7$	$\mathscr{P}(X_{19}) = X_6$	$\mathscr{P}(\mathscr{C}(X_{19})) = X_{18}$
X_{20}	$\{CA, CG, CT, GA, TA, TG\}$	$\mathscr{C}(X_{20}) = X_{16}$	$\mathscr{P}(X_{20}) = X_5$	$\mathscr{P}(\mathscr{C}(X_{20})) = X_9$
X_{21}	$\{CA, CG, GA, TA, TC, TG\}$	$\mathscr{C}(X_{21}) = X_{21}$	$\mathscr{P}(X_{21}) = X_4$	$\mathscr{P}(\mathscr{C}(X_{21})) = X_4$
X_{22}	$\{CA, CT, GA, GC, GT, TA\}$	$\mathscr{C}(X_{22}) = X_8$	$\mathscr{P}(X_{22}) = X_3$	$\mathscr{P}(\mathscr{C}(X_{22})) = X_{17}$
X_{23}	$\{CA, GA, GC, GT, TA, TC\}$	$\mathscr{C}(X_{23}) = X_{12}$	$\mathscr{P}(X_{23}) = X_2$	$\mathscr{P}(\mathscr{C}(X_{23})) = X_{13}$
X_{24}	$\{CA, GA, GC, TA, TC, TG\}$	$\mathscr{C}(X_{24}) = X_{24}$	$\mathscr{P}(X_{24}) = X_1$	$\mathscr{P}(\mathscr{C}(X_{24})) = X_1$

Result 16 (Arquès, Michel, 1996, 1997):
The circular code X_0 codes 12 amino acids

$$X_0 = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC$$
$$GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$$

X_0 codes 12 amino acids: {*Ala, Asn, Asp, Gln, Glu, Gly, Ile, Leu, Phe, Thr, Tyr, Val*}

Result 17 (Michel, Pirillo, 2013):

None circular code among the 12,964,440 ones codes 20 or 19 amino acids.

Proposition 2. *The following set Y of 20 trinucleotides*

$$Y = \{ACG,ACT,AGA,AGG,AGT,ATA,ATC,CAA,CAC,CAG,$$
$$CCT,GCC,GCG,GCT,GGT,TCG,TCT,TGA,TGT,TTA\}$$

is a circular code (maximal).

Proposition 3. *The trinucleotide circular code Y (Proposition 2) has a permuted set $\mathcal{P}^2(Y)$ of 20 trinucleotides*

$$\mathcal{P}^2(Y) = \{AAG,AAT,ACA,ATG,ATT,CAT,CCA,CGC,GAC,GAG,$$
$$GCA,GGC,GTC,TAC,TAG,TCC,TGC,TGG,TTC,TTG\}$$

which is not circular and codes the 20 amino acids in the variant nuclear codes 6 and 15.

Result 18 (Arquès, Michel, 1996, 1997):

The comma-free code *RNY = {RRY,RYY}* deduced from X_0

$$X_0 = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC$$
$$GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$$

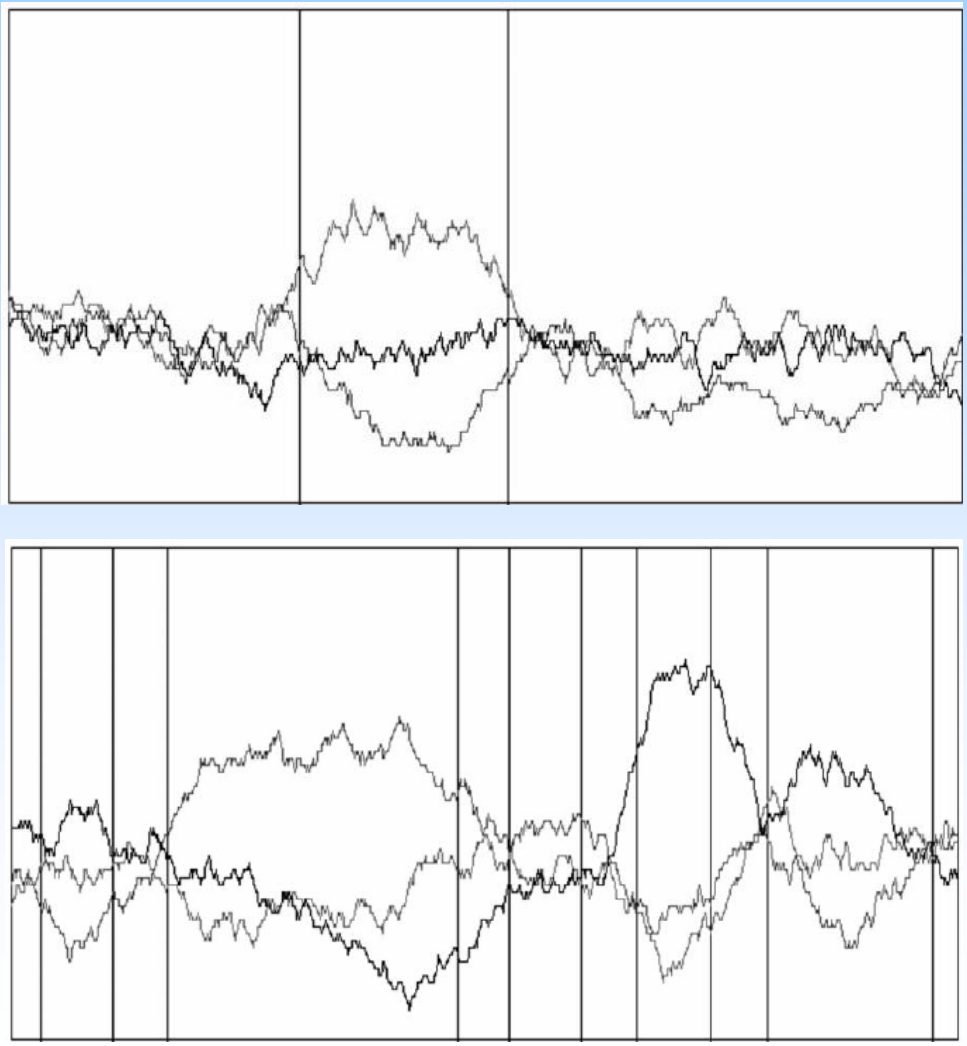
$$X_1 = \{\text{AAG, ACA, ACG, ACT, AGC, AGG, ATA, ATG, CCA, CCG,}$$
$$\text{GCG, GTG, TAG, TCA, TCC, TCG, TCT, TGC, TTA, TTG}\}$$

$$X_2 = \{\text{AGA, AGT, CAA, CAC, CAT, CCT, CGA, CGC, CGG, CGT,}$$
$$\text{CTA, CTT, GCA, GCT, GGA, TAA, TAT, TGA, TGG, TGT}\}.$$

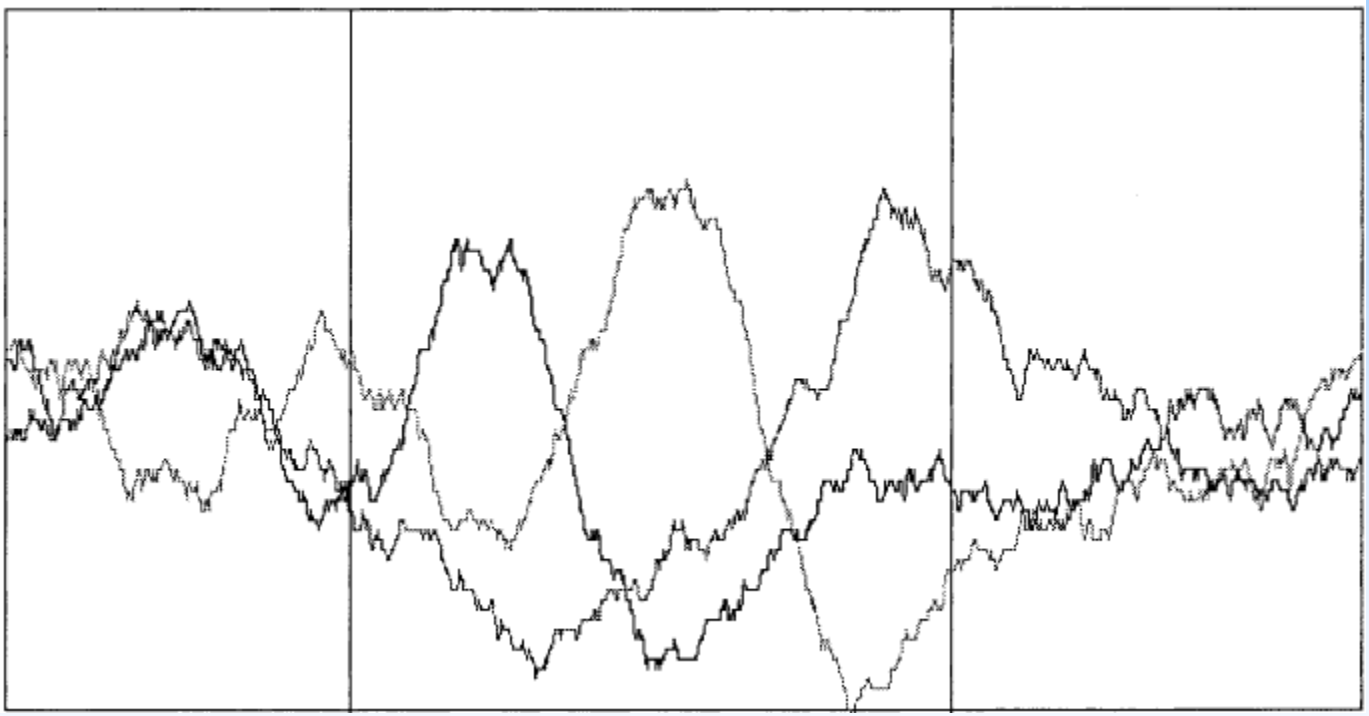
The eight R/Y trinucleotides (R = purine = A or G, Y = pyrimidine = C or T) are associated with the 64 A/C/G/T trinucleotides by considering their frame (T_0, T_1, T_2)

RRR	RRY	RYR	RYY	YRR	YRY	YYR	YYY
AAA^0	AAC^0	ACA^1	ACC^0	CAA^2	CAC^2	CCA^1	CCC^1
AAG^1	AAT^0	ACG^1	ACT^1	CAG^0	CAT^2	CCG^1	CCT^2
AGA^2	AGC^1	ATA^1	ATC^0	CGA^2	CGC^2	CTA^2	CTC^0
AGG^1	AGT^2	ATG^1	ATT^0	CGG^2	CGT^2	CTG^0	CTT^2
GAA^0	GAC^0	GCA^2	GCC^0	TAA^2	TAC^0	TCA^1	TCC^1
GAG^0	GAT^0	GCG^1	GCT^2	TAG^1	TAT^2	TCG^1	TCT^1
GGA^2	GGC^0	GTA^0	GTC^0	TGA^2	TGC^1	TTA^1	TTC^0
GGG^2	GGT^0	GTG^1	GTT^0	TGG^2	TGT^2	TTG^1	TTT^0
0, 1, 2	0	1	0	2	2	1	0, 1, 2

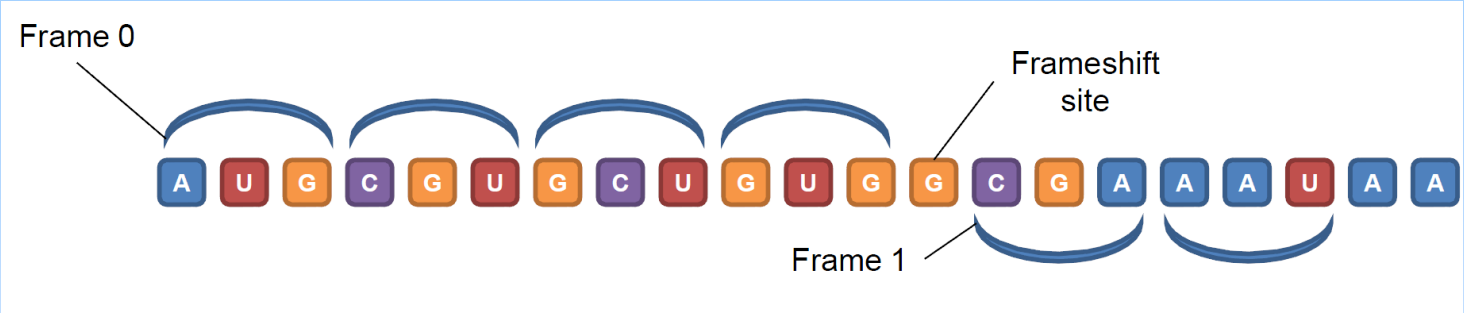
Result 19 (Arquès, Lacan, Michel, 2002):
Identification of genes in genomes with statistical functions based on the circular code X_0

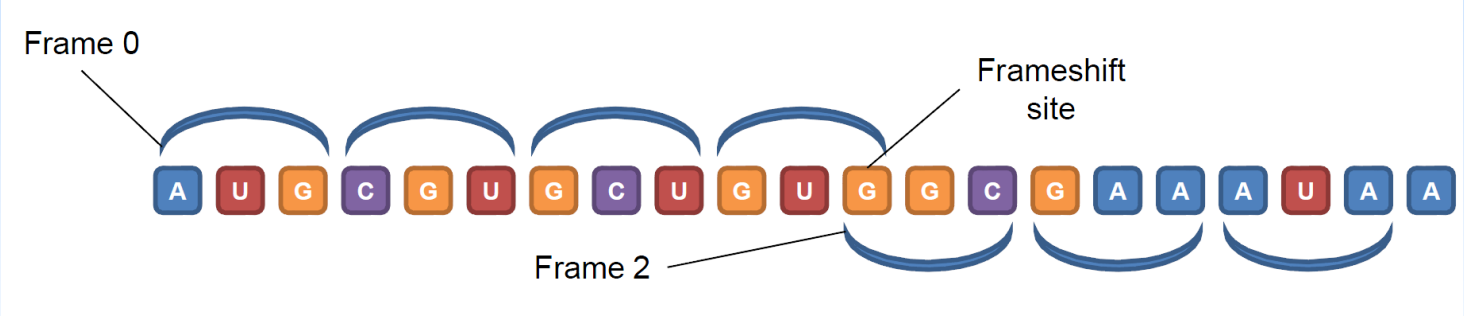


Result 19 (Arquès, Lacan, Michel, 2002):
Identification of genes in genomes with statistical functions based on the circular code X_0

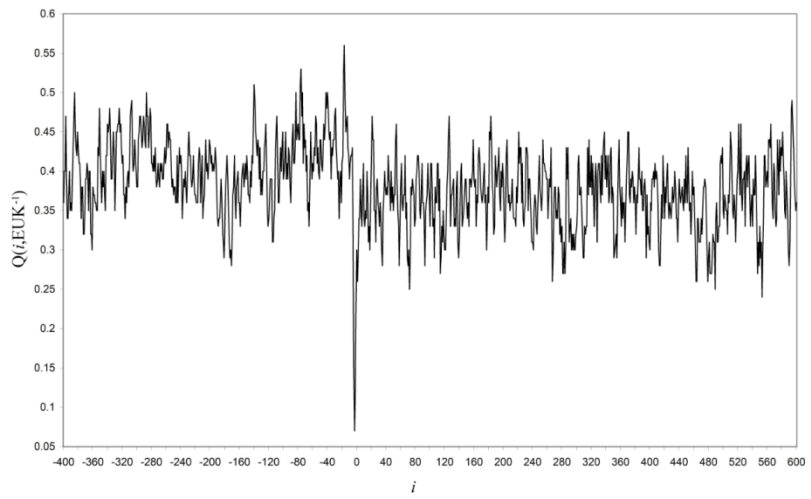


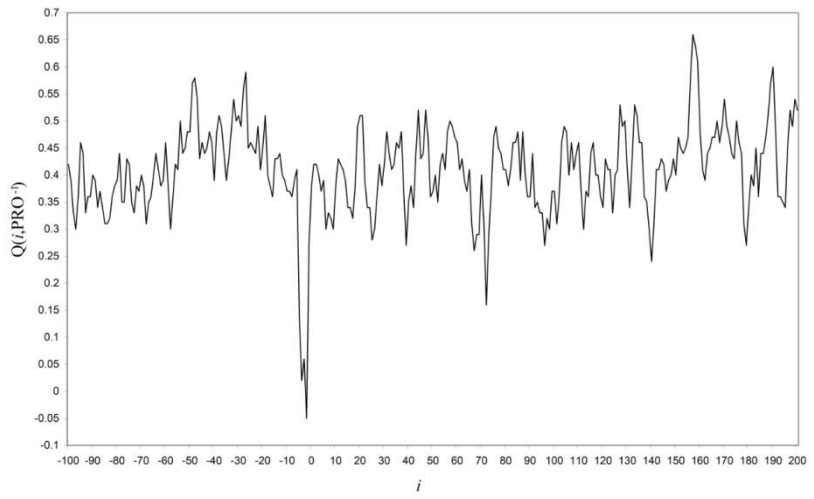
Frameshift genes



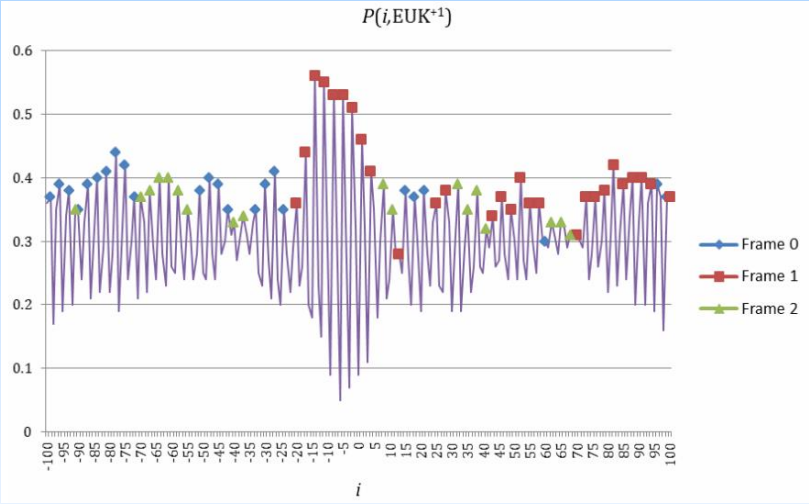


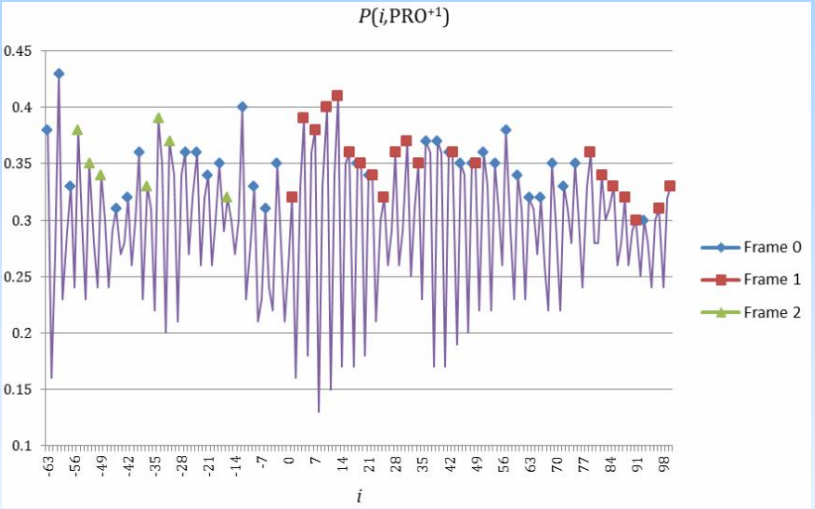
Result 20 (Ahmed, Frey, Michel, 2007):
Loss of the signal of the circular code X_0 at the frameshift site of frameshift genes





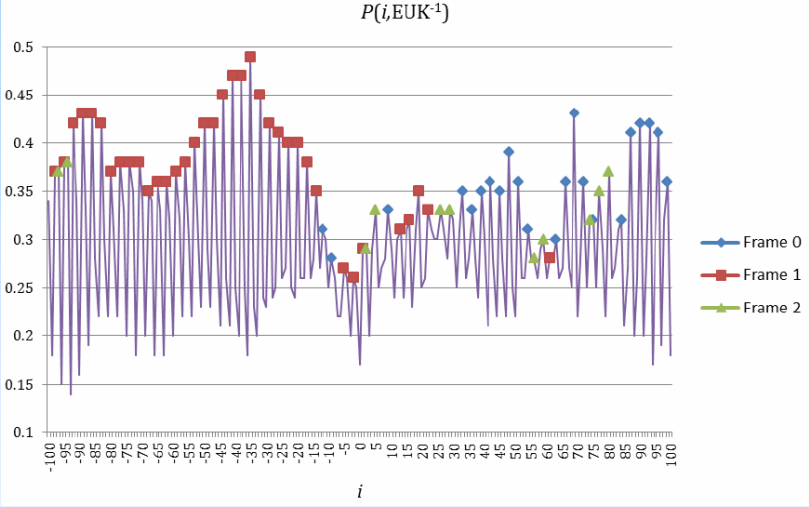
Result 21 (Ahmed, Michel, 2011):
Shift of the signal of the circular code X_0 at the frameshift site of frameshift genes +1

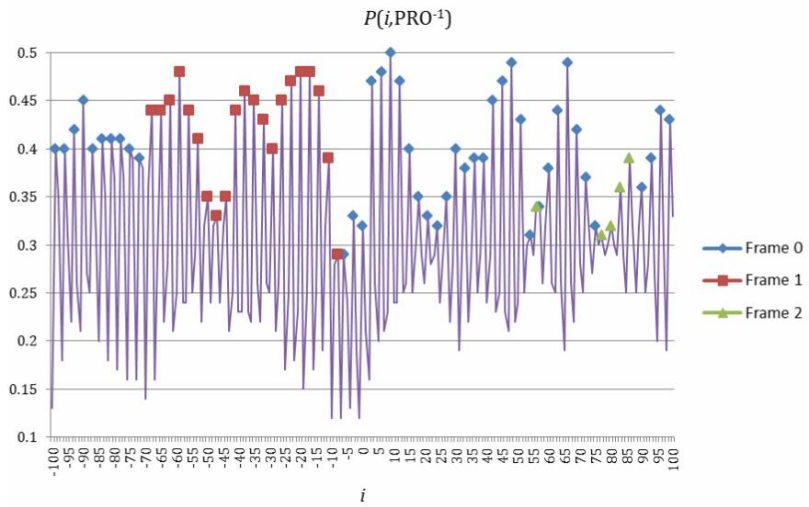




Result 21 (Ahmed, Michel, 2011):

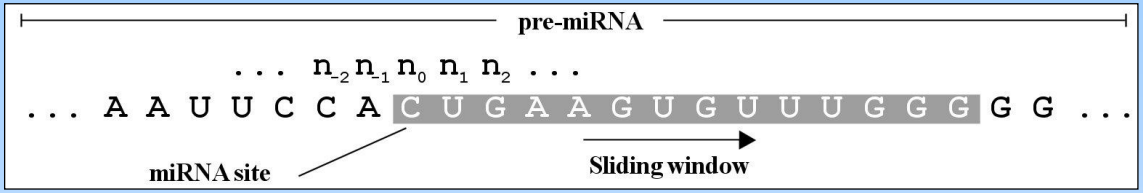
Shift of the signal of the circular code X_0 at the frameshift site of frameshift genes -1

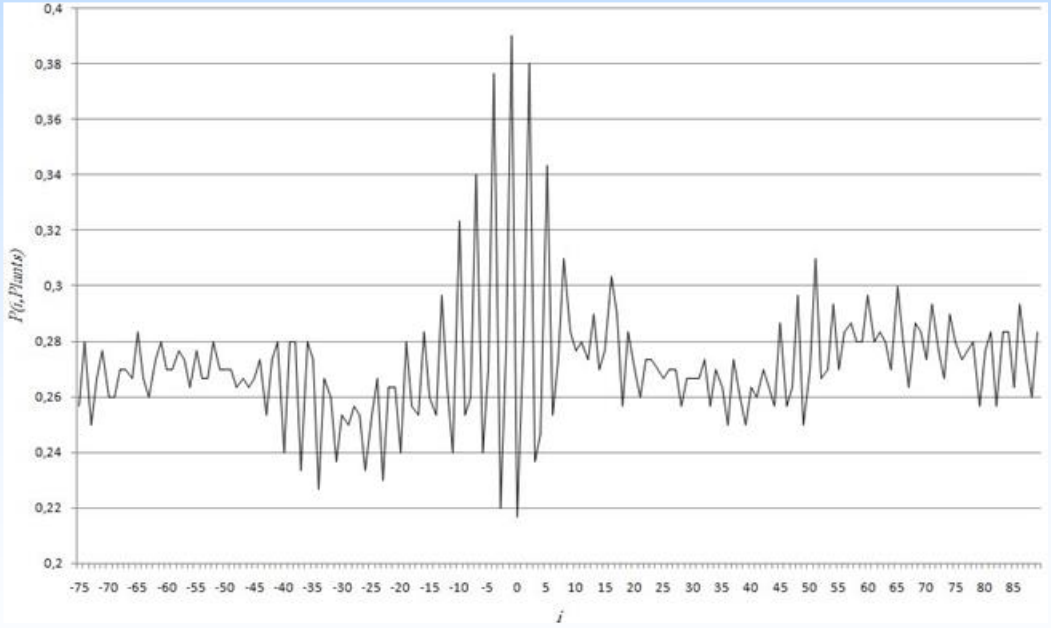




Statistical tests based on the circular code X_0 can be used to describe frameshift genes (Seligmann, 2012)

Result 22 (Ahmed, Michel, 2008):
Signal of the circular code X_0 in the plant miRNAs





Result 23 (Michel, 2013):
X_0 circular code motifs in transfer RNAs

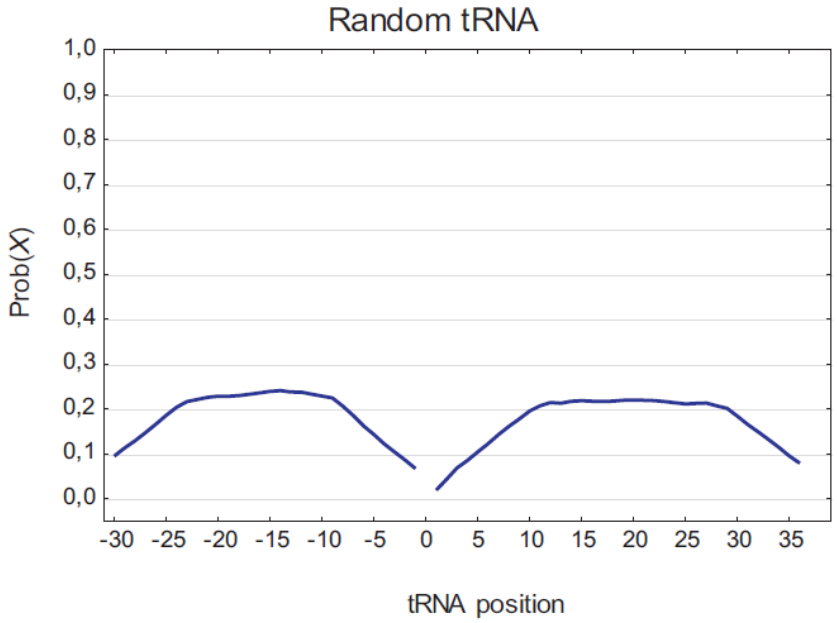


Fig. 4. Probability $Pr(i, \mathcal{F}_{RAN})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of a randomized tRNA population $\mathcal{F}_{RAN}$.

Result 23 (Michel, 2013):
X_0 circular code motifs in tRNAs of prokaryotes

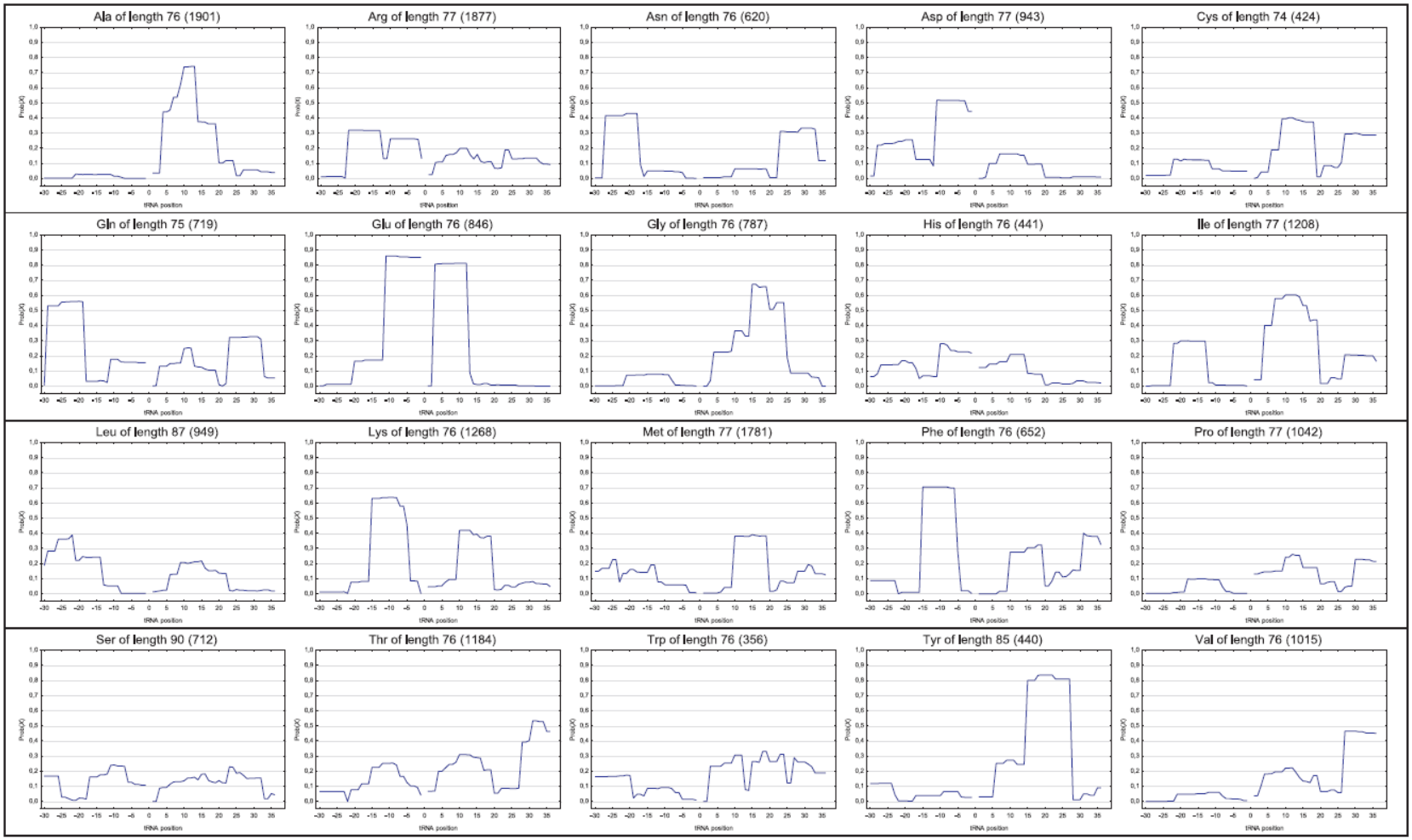


Fig. 5. Probability $Pr(i, \mathcal{F}^1_{\mathrm{PRO}})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of the tRNA population of prokaryotes $\mathcal{F}^1_{\mathrm{PRO}}$ constituted of the 20 isoaccepting tRNAs having a unique length according to the maximum number of data. The number of isoaccepting tRNAs is in parenthesis.

Result 23 (Michel, 2013):
X_0 circular code motifs in tRNAs of eukaryotes

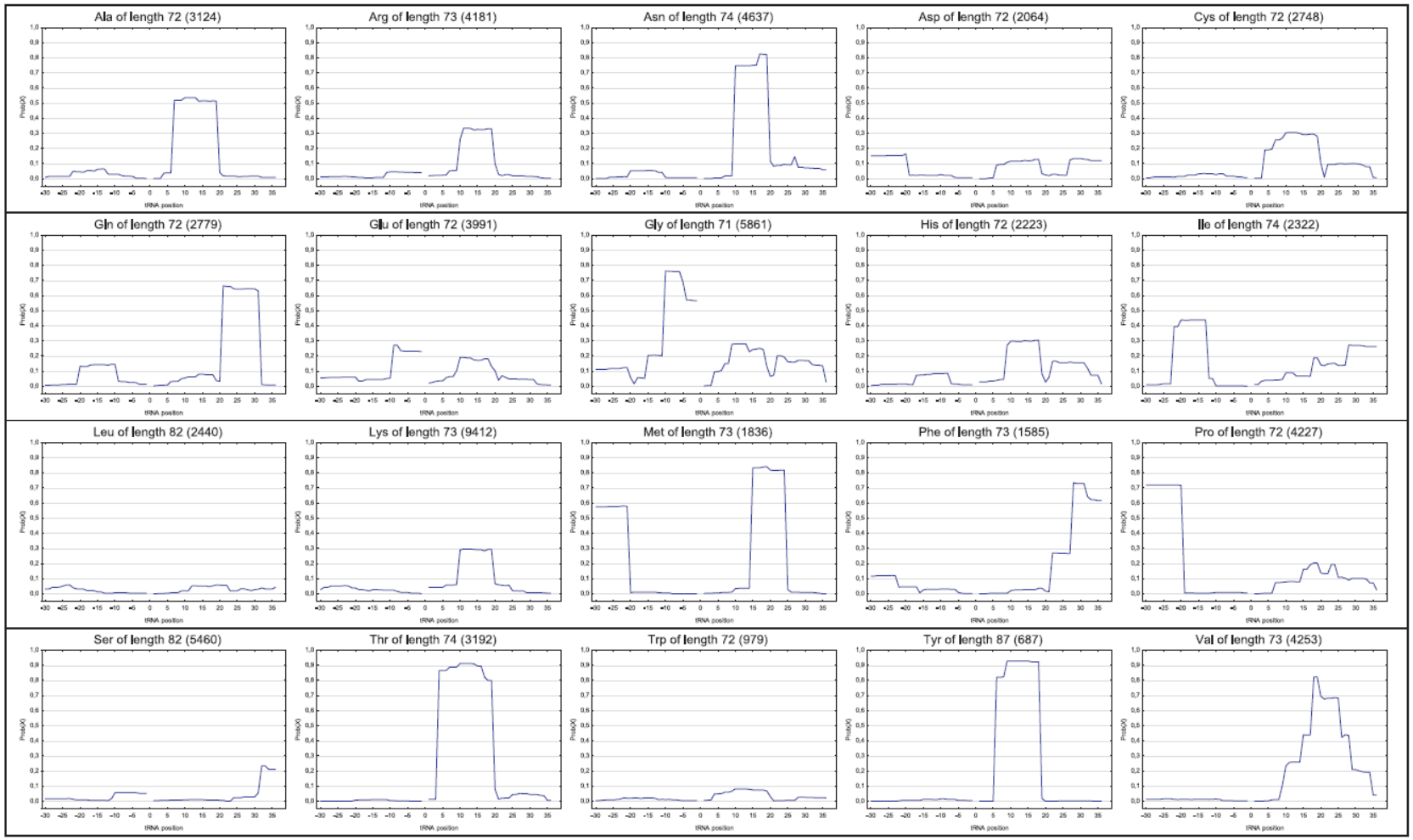
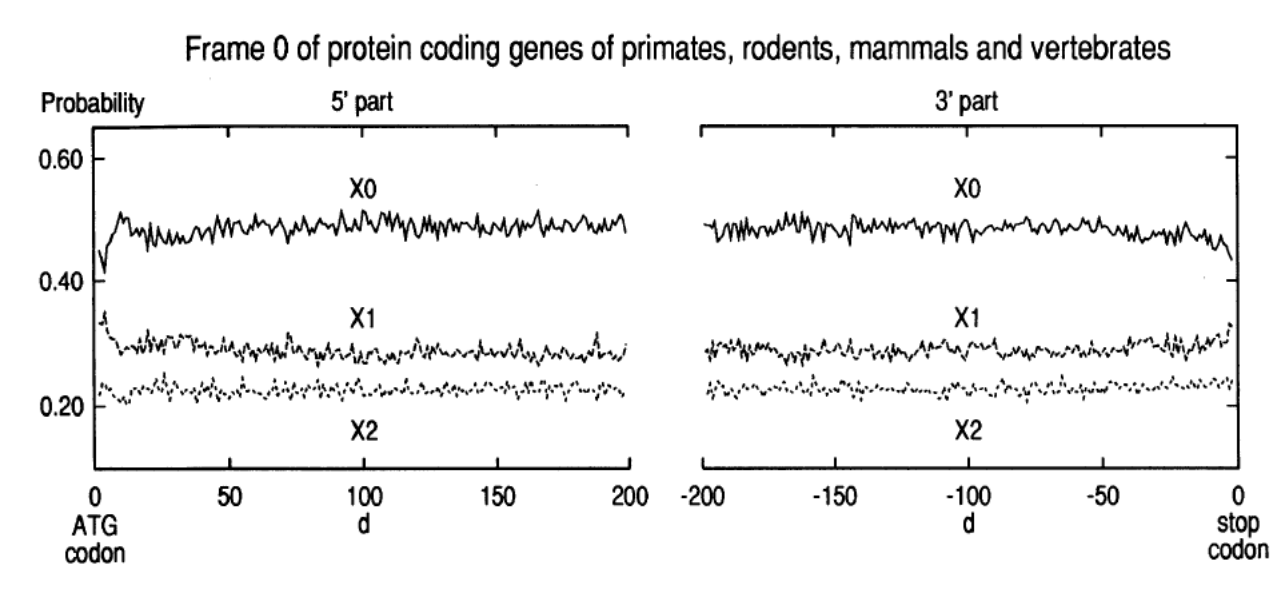


Fig. 6. Probability $Pr(i, \mathcal{F}^1_{\mathrm{EUK}})$ (Eq. (8)) of X motifs of greatest lengths having at least 9 nucleotides, in the 5′ and 3′ regions of the tRNA population of eukaryotes $\mathcal{F}^1_{\mathrm{EUK}}$ constituted of the 20 isoaccepting tRNAs having a unique length according to the maximum number of data. The number of isoaccepting tRNAs is in parenthesis.

Result 24 (Arquès, Fallot, Marsan, Michel, 1999; Bahi, Michel, 2008):

Asymmetry between the circular codes X_1 and X_2 in genes



Gene population $\mathcal{F}_{\mathrm{G(PRMV)}}$ of primates, rodents, mammals and vertebrates (17072 genes)

$$\Pr(X, \mathcal{F}_{\mathrm{G(PRMV)}}) = 48.5\% > \Pr(X_1, \mathcal{F}_{\mathrm{G(PRMV)}}) = 29.0\% > \Pr(X_2, \mathcal{F}_{\mathrm{G(PRMV)}}) = 22.5\%$$

Gene population $\mathcal{F}_{\mathrm{G(PRO)}}$ of 175 complete genomes of prokaryotes (487,758 genes, 454 Mb)

$$\Pr(X, \mathcal{F}_{\mathrm{G(PRO)}}) = 48.8\% > \Pr(X_1, \mathcal{F}_{\mathrm{G(PRO)}}) = 28.0\% > \Pr(X_2, \mathcal{F}_{\mathrm{G(PRO)}}) = 23.2\%$$

Result 25 (Michel, 2013):

Asymmetry between the circular codes X_1 and X_2 in the 3' regions of tRNAs of prokaryotes and eukaryotes

3' regions of the tRNA population of prokaryotes $\mathcal{F}^2_{3'\mathrm{PRO}}$ (30046 tRNAs)

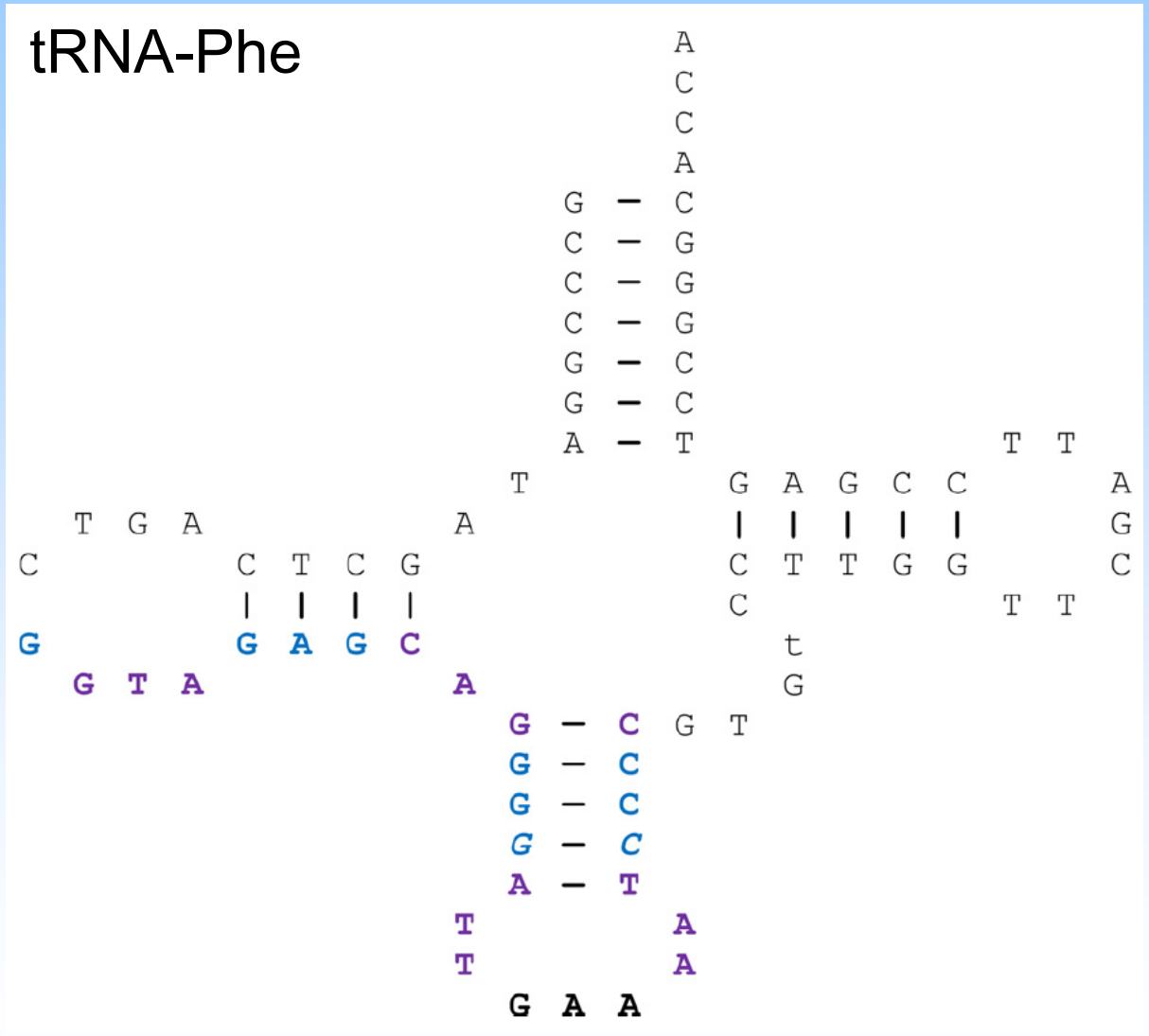
$$\Pr(X, \mathcal{F}^2_{3'\mathrm{PRO}}) = 43.8\% > \Pr(X_1, \mathcal{F}^2_{3'\mathrm{PRO}}) = 34.7\% > \Pr(X_2, \mathcal{F}^2_{3'\mathrm{PRO}}) = 21.5\%$$

3' regions of the tRNA population of eukaryotes $\mathcal{F}^2_{3'\mathrm{EUK}}$ (84687 tRNAs)

$$\Pr(X, \mathcal{F}^2_{3'\mathrm{EUK}}) = 46.4\% > \Pr(X_1, \mathcal{F}^2_{3'\mathrm{EUK}}) = 32.5\% > \Pr(X_2, \mathcal{F}^2_{3'\mathrm{EUK}}) = 21.1\%$$

Result 26 (Michel, 2012):

A possible translation code based on the circular code X_0



Result 26 (Michel, 2012):
A possible translation code based on the circular code X_0

```
1    TTGGAGAGTTTGATCCTGGCTCAGGGTGAACGCTGGCGGCGTGCCTAAGACATGCAAGTCGTGCGGGCCGCGGGGTTTTA
81   CTCCGTGGTCAGCGGCGGACGGGTGAGTAACGCGTGGGTGACCTACCCGGAAGAGGGGGACAACCCGGGGAAACTCGGGC
161  TAATCCCCCATGTGGACCCGCCCCTTGGGGTGTGTCCAAAGGGCTTTGCCCGCTTCCGGATGGGCCCGCGTCCCATCAGC
241  TAGTTGGTGGGGTAATGGCCCACCAAGGCGACGACGGGTAGCCGGTCTGAGAGGATGGCCGGCCACAGGGGCACTGAGAC
321  ACGGGCCCCACTCCTACGGGAGGCAGCAGTTAGGAATCTTCCGCAATGGGCGCAAGCCTGACGGAGCGACGCCGCTTGGA
401  GGAAGAAGCCCTTCGGGGTGTAAACTCCTGAACCCGGGACGAAACCCCCGACGAGGGGACTGACGGTACCGGGGTAATAG
481  CGCCGGCCAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGCGCGAGCGTTACCCGGATTCACTGGGCGTAAAGGGCGT
561  GTAGGCGGCCTGGGGCGTCCCATGTGAAAGACCACGGCTCAACCGTGGGGGAGCGTGGGATACGCTCAGGCTAGACGGTG
641  GGAGAGGGTGGTGGAATTCCCGGAGTAGCGGTGAAATGCGCAGATACCGGGAGGAACGCCGATGGCGAAGGCAGCCACCT
721  GGTCCACCCGTGACGCTGAGGCGCGAAAGCGTGGGGAGCAAACCGGATTAGATACCCGGGTAGTCCACGCCCTAAACGAT
801  GCGCGCTAGGTCTCTGGGTCTCCTGGGGGCCGAAGCTAACGCGTTAAGCGCGCCGCCTGGGGAGTACGGCCGCAAGGCTG
881  AAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACCA
961  GGCCTTGACATGCTAGGGAACCCGGGTGAAAGCCTGGGGTGCCCCGCGAGGGGAGCCCTAGCACAGGTGCTGCATGGCCG
1041 TCGTCAGCTCGTGCCGTGAGGTGTTGGGTTAAGTCCCGCAACGAGCGCAACCCCCGCCGTTAGTTGCCAGCGGTTCGGCC
1121 GGGCACTCTAACGGGACTGCCCGCGAAAGCGGGAGGAAGGAGGGGACGACGTCTGGTCAGCATGGCCCTTACGGCCTGGG
1201 CGACACACGTGCTACAATGCCCACTACAAAGCGATGCCACCCGGCAACGGGGAGCTAATCGCAAAAAGGTGGGCCCAGTT
1281 CGGATTGGGGTCTGCAACCCGACCCCATGAAGCCGGAATCGCTAGTAATCGCGGATCAGCCATGCCGCGGTGAATACGTT
1361 CCCGGGCCTTGTACACACCGCCCGTCACGCCATGGGAGCGGGCTCTACCCGAAGTCGCCGGGAGCCTACGGGCAGGCGCC
1441 GAGGGTAGGGCCCGTGACTGGGGCGAAGTCGTAACAAGGTAGCTGTACCGGAAGGTGCGGCTGGATCACCTCCTTT 1516
```

16S rRNA

Result 26 (Michel, 2012):

A possible translation code based on the circular code X_0

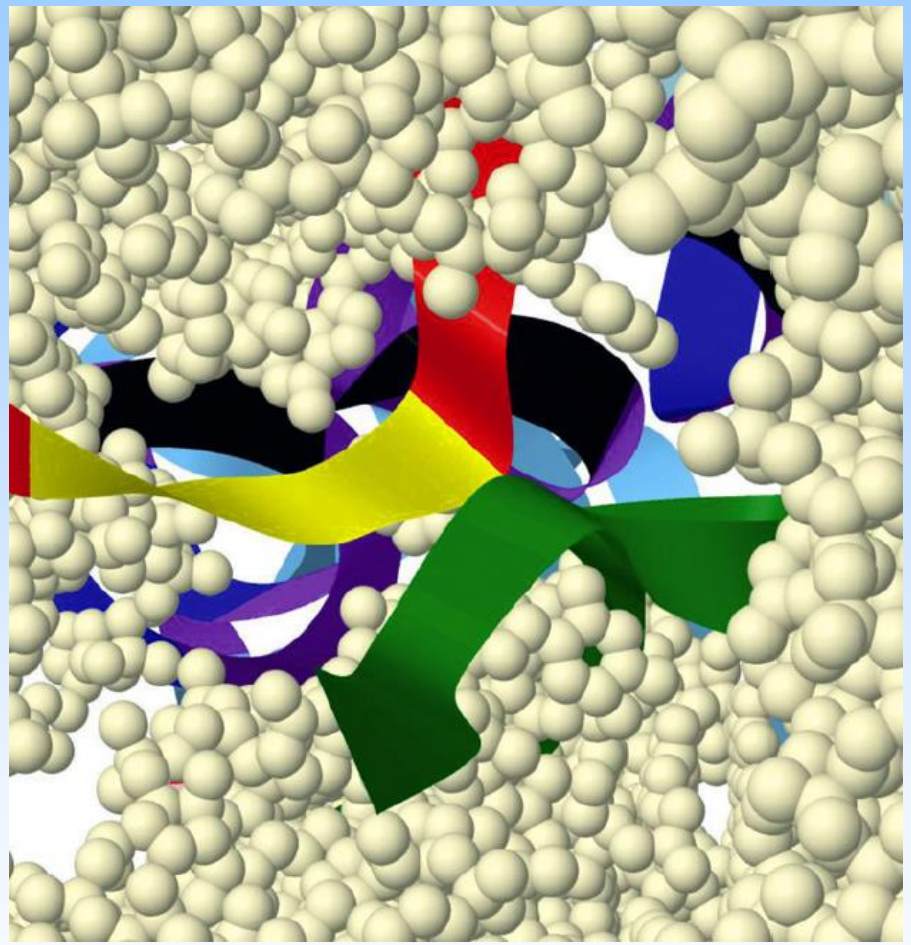


Fig. 7. Spatial relation of the mRNA X motifs (green), the A-tRNA (lightcyan) X motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ (blue and blueviolet with the anticodon in black) and the rRNA X motif $m_{\text{16SrRNA-2}}(1189, 1206, 18)$ (red and yellow). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs.

Result 26 (Michel, 2012):
A possible translation code based on the circular code X_0

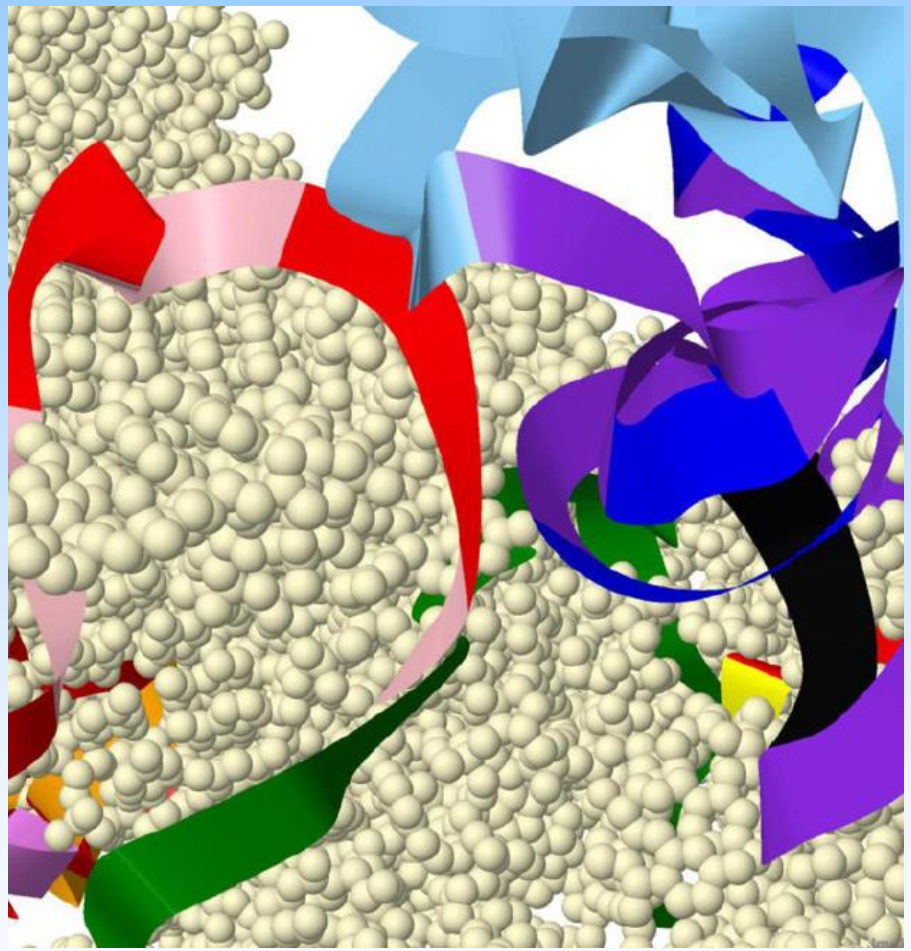


Fig. 8. Spatial relation of the mRNA X motifs (green), the E-tRNA (lightskyblue) X motif $m_{\text{tRNA-Phe}}(18, 43, 26)$ (blue and blueviolet with the anticodon in black) and the rRNA X motif $m_{16\text{SrRNA-1}}(694, 713, 20)$ (red and pink). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs.

Result 26 (Michel, 2012):
A possible translation code based on the circular code X_0

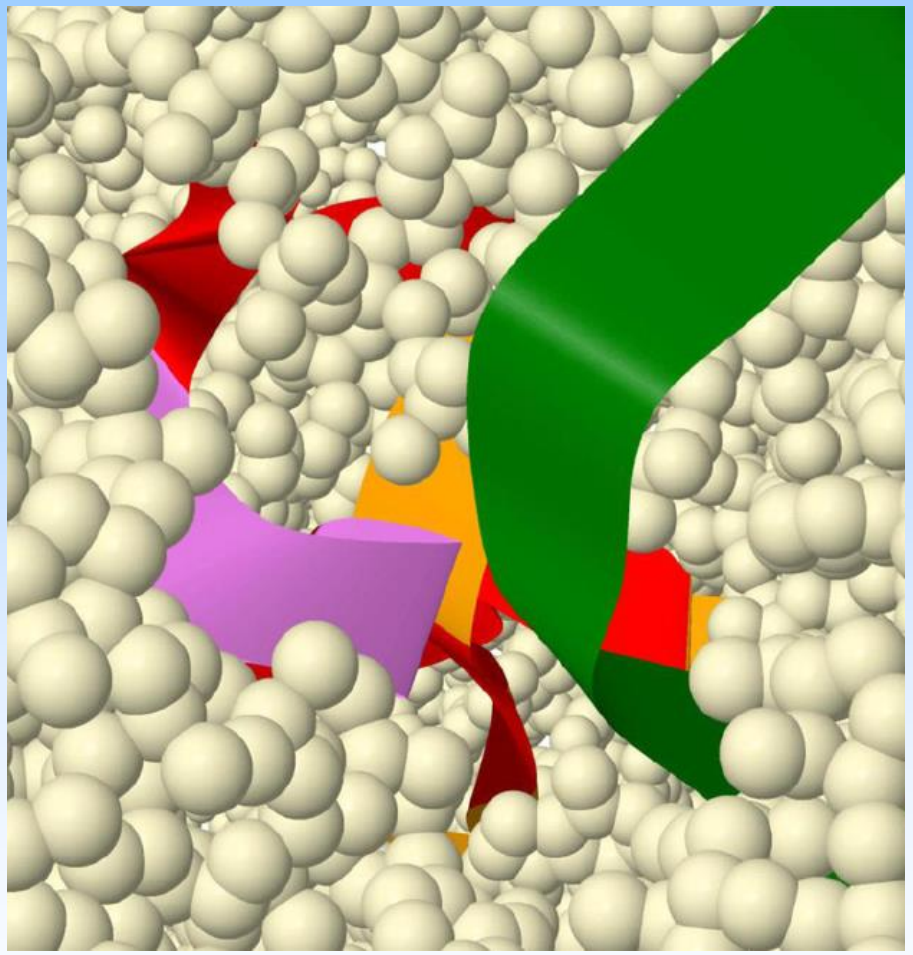


Fig. 9. Spatial relation of the mRNA X motifs (green) and the two rRNA X motifs $m_{16SrRNA\text{-}3}(559, 574, 16)$ (red and orange) and $m_{16SrRNA\text{-}4}(813, 827, 15)$ (red and violet). The remaining rRNA (lemonchiffon) is outside the neighborhood of these X motifs.

References

http://dpt-info.u-strasbg.fr/~c.michel/