CIRCULAR CODES IN GENES AND GENOMES

- 2015 -

Prof. Christian MICHEL

Theoretical Bioinformatics
ICube
University of Strasbourg, CNRS
France

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

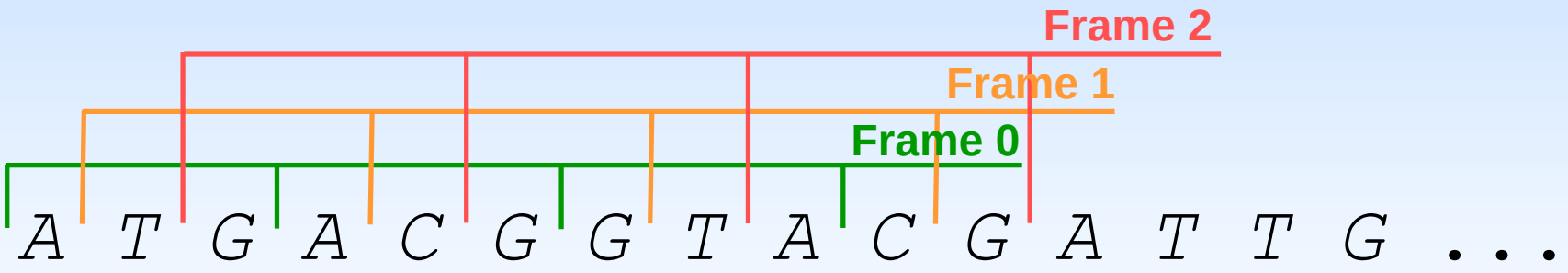
Biological recall: 3 frames in genes

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'



Definition of a genetic scale of reading frame coding

Probability PrRFC of reading frame coding (RFC) of any trinucleotide code

- new
- simplest
- associated with the combinatorial properties of trinucleotide codes

The RFC probability $\mathrm{PrRFC}(C)$ of a trinucleotide code C is equal to the ratio of the occurrence probability of C in frame 0 to the occurrence probabilities of C in the three frames 0, 1 and 2

Definition of a genetic scale of reading frame coding

Let $t = l_0 l_1 l_2$, $l_0, l_1, l_2 \in A_4$, be a trinucleotide of a code C. Each trinucleotide t of C is assigned to a probability $\Pr(t)$ such that the probability $\Pr(C)$ of the code C is equal to

$$\Pr(C) = \sum_{t \in C} \Pr(t) = 1.$$

The probabilities $\Pr(t)$ of trinucleotides t are not necessarily equiprobable.

The reading frame $f = 0$ is established by the letter l_0 of $t = l_0 l_1 l_2$.

The frames $f = 1$ and $f = 2$ start with the letters l_1 and l_2 of t, respectively.

Let the di-trinucleotide w be a concatenation of two trinucleotides $t' = l'_0 l'_1 l'_2 \in C$ and $t'' = l''_0 l''_1 l''_2 \in C$, i.e. $w = t't'' \in C^2$.

We denote by $t_0(w) = l'_0 l'_1 l'_2 \in A_4^3$, $t_1(w) = l'_1 l'_2 l''_0 \in A_4^3$ and $t_2(w) = l'_2 l''_0 l''_1 \in A_4^3$ the trinucleotides in frames 0, 1 and 2, respectively, of a di-trinucleotide $w \in C^2$.

Definition of a genetic scale of reading frame coding

The concatenation of the two trinucleotides $t' \in C$ and $t'' \in C$
may yield a trinucleotide $t_f(w)$ in a shifted frame $f \in \{1,2\}$ belonging to C.

For example, with the code $C = \{\mathrm{AAA}, \mathrm{AAC}\}$,
the concatenation of the trinucleotides $t' = \mathrm{AAA} \in C$ and $t'' = \mathrm{AAC} \in C$,
i.e. $w = \mathrm{AAAAAC}$,
leads to the trinucleotides $t_1(w) = \mathrm{AAA} \in C$ and $t_2(w) = \mathrm{AAA} \in C$.

The probability $\Pr\big(t_f(w), f\big)$ of a trinucleotide $t_f(w) \in A_4^3$
in a frame $f \in \{1,2\}$ of a di-trinucleotide $w = t't'' \in C^2$
is equal to the product of probabilities $\Pr(t')$ and $\Pr(t'')$
(with the simplest hypothesis of independent events)

$$\Pr\big(t_f(w), f\big) = \Pr(t') \times \Pr(t'').$$

Definition of a genetic scale of reading frame coding

The probability $\mathrm{PrFrame}(t, f)$ of a trinucleotide $t \in C$ in a frame $f \in \{1,2\}$ in all the di-trinucleotides $w = t't'' \in C^2$ is equal to

$$\mathrm{PrFrame}(t, f) = \sum_{w \in C^2 \mid t = t_f(w) \in C} \Pr\big(t_f(w), f\big).$$

Then, the probability $\mathrm{PrFrame}(C, f)$ of a code C in a frame $f \in \{1,2\}$ is equal to

$$\mathrm{PrFrame}(C, f) = \sum_{t \in C} \mathrm{PrFrame}(t, f).$$

Finally, the reading frame coding probability $\mathrm{PrRFC}(C)$ of a code C is equal to

$$\mathrm{PrRFC}(C) = \frac{\Pr(C)}{\Pr(C) + \sum_{f=1}^{2} \mathrm{PrFrame}(C, f)}$$

$$= \frac{1}{1 + \sum_{f=1}^{2} \mathrm{PrFrame}(C, f)}.$$

Definition of a genetic scale of reading frame coding

Property 1. $\frac{1}{3} \leq \mathrm{PrRFC}(C) \leq 1$

The more the RFC probability $\mathrm{PrRFC}(C)$ value is raised,
the more the code C has efficiency for coding the reading frame 0.

If $\mathrm{PrFrame}(C,1) = \mathrm{PrFrame}(C,2) = 0$ then $\mathrm{PrRFC}(C) = 1$,
i.e. the code C only codes the reading frame 0
and always retrieves the reading frame.

If $\mathrm{Pr}(C) = \mathrm{PrFrame}(C,1) = \mathrm{PrFrame}(C,2) = 1$ then $\mathrm{PrRFC}(C) = 1/3$,
i.e. the code C codes the three frames 0, 1 and 2 with the same efficiency
and the reading frame 0 is retrieved randomly.

Results: Reading frame coding probability PrRFC

If $C = \{\mathrm{AAA}\}$ then $\mathrm{PrRFC}(C) = 1/3$

If $C = \{\mathrm{AAC}\}$ then $\mathrm{PrRFC}(C) = 1$

If $C = PPT = \{\mathrm{AAA}, \mathrm{CCC}, \mathrm{GGG}, \mathrm{TTT}\}$ (prob. $1/4$) then $\mathrm{PrRFC}(C) = 2/3$

If $C = A_4^3 = \{\mathrm{AAA}, \ldots, \mathrm{TTT}\}$ (prob. $1/64$) then $\mathrm{PrRFC}(C) = 1/3$

If $C = 61GC = A_4^3 \setminus \{\mathrm{TAA}, \mathrm{TAG}, \mathrm{TGA}\}$ (prob. $1/61$) then

$\mathrm{PrRFC}(61GC) = 3721/10779 \approx 34.5\%$

If $C = X$ (C^3 self-complementary circular code identified in genes) then

$\mathrm{PrRFC}(X) = 100/123 \approx 81.3\%$

Example

The trinucleotide code $C = \{\text{AAA}, \text{AAC}, \text{ACA}\}$

with $\Pr(\text{AAA}) = 1/6$, $\Pr(\text{AAC}) = 1/3$ and $\Pr(\text{ACA}) = 1/2$

has a probability $\text{PrRFC}(C)$ of reading frame coding equal to $6/13$.

							Frame $f = 0$	Frame $f = 0$	Frame $f = 1$	Frame $f = 2$
	t'			t''			$\Pr(t')$	$\Pr(t'')$	$\Pr(t_1(w), f) = \Pr(t') \times \Pr(t'')$ Equation (3)	$\Pr(t_2(w), f) = \Pr(t') \times \Pr(t'')$ Equation (3)
w	A	A	A	A	A	A	1/6	1/6		
		A	A	A					$1/36\ (t_1(w) = \text{AAA} \in C)$	
			A	A	A					$1/36\ (t_2(w) = \text{AAA} \in C)$
w	A	A	A	A	A	C	1/6	1/3		
		A	A	A					$1/18\ (t_1(w) = \text{AAA} \in C)$	
			A	A	A					$1/18\ (t_2(w) = \text{AAA} \in C)$
w	A	A	A	A	C	A	1/6	1/2		
		A	A	A					$1/12\ (t_1(w) = \text{AAA} \in C)$	
			A	A	C					$1/12\ (t_2(w) = \text{AAC} \in C)$
w	A	A	C	A	A	A	1/3	1/6		
		A	C	A					$1/18\ (t_1(w) = \text{ACA} \in C)$	
			C	A	A					$1/18\ (t_2(w) = \text{CAA} \notin C)$

Example

	Frame $f = 0$	Frame $f = 1$	Frame $f = 2$		
$t \in C$	$\Pr(t)$	$\text{PrFrame}(t, 1)$ Equation (4)	$\text{PrFrame}(t, 2)$ Equation (4)		
AAA	1/6	1/36+1/18+1/12=1/6	1/36+1/18+1/12+1/6=1/3		
AAC	1/3	0	1/12+1/4=1/3		
ACA	1/2	1/18+1/9+1/6=1/3	0		
$\text{PrFrame}(C, f)$ Equation (5)	1	1/2	2/3	1/(1+1/2+2/3)=**6/13**	$\text{PrRFC}(C)$ Equation (6)

Classes of trinucleotide codes

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 $i = 1, \ldots, n$.

The genetic code A_4^3 = {AAA,...,TTT} (64 trinucleotides)

The genetic code 61*GC* = A_4^3 \ {TAA,TAG,TGA} (61 trinucl.)

The periodic permuted trinucleotides

PPT = {AAA,CCC,GGG,TTT} (4 trinucleotides)

Classes of trinucleotide codes

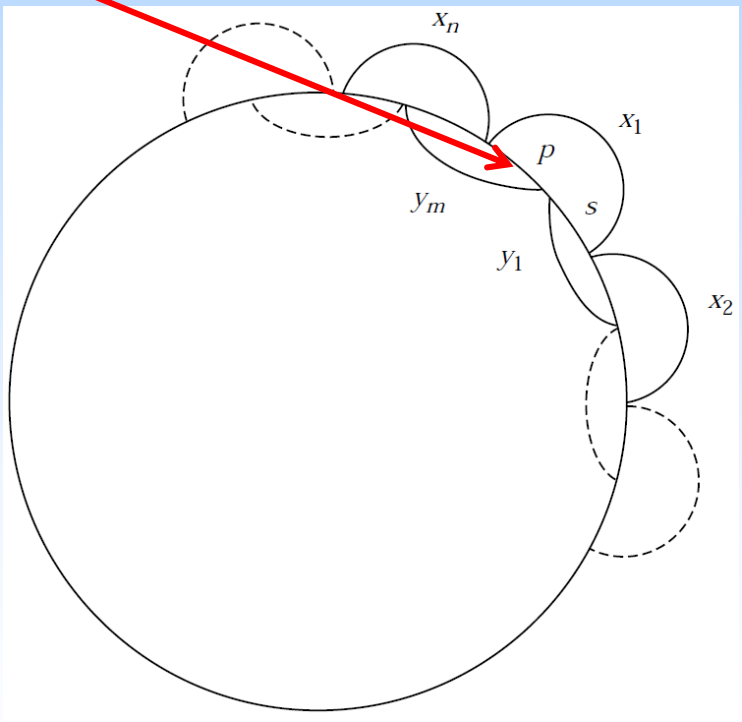
Definition 2.2. Trinucleotide comma-free code: a trinucleotide code $X \subset \mathcal{A}_4^3$ is comma-free if for each $y \in X$ and $u, v \in \mathcal{A}_4^*$ such that $uyv = x_1 \cdots x_n$ with $x_1, \ldots, x_n \in X$, $n \geq 1$, it results that $u, v \in X^*$.

408 comma-free circular codes $CFCC$

$$X = \{\text{AAC}, \text{AAG}, \text{AAT}, \text{CCA}, \text{GAC}, \text{TAC}, \text{GCA}, \text{GAG}, \text{TAG}, \text{TCA},$$
$$\text{GAT}, \text{TAT}, \text{CCG}, \text{CCT}, \text{GCG}, \text{TCG}, \text{GCT}, \text{TCT}, \text{GGT}, \text{TTG}\}.$$

Classes of trinucleotide codes

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 $i = 1, \ldots, n$.



Classes of trinucleotide codes

$12{,}964{,}440$ circular codes CC

216 C^3 self-complementary circular codes C^3SCC

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

identified in genes of bacteria, eukaryotes, plasmids and viruses

Results: Classes of trinucleotide codes

$2^9 \times 3 \times 4^5 \times 6^3 = 339{,}738{,}624$ bijective genetic codes BGC

such that 20 trinucleotides code the 20 amino acids

2 amino acids are encoded by 1 codon

9 amino acids are encoded by 2 codons

1 amino acid is encoded by 3 codons

5 amino acids are encoded by 4 codons

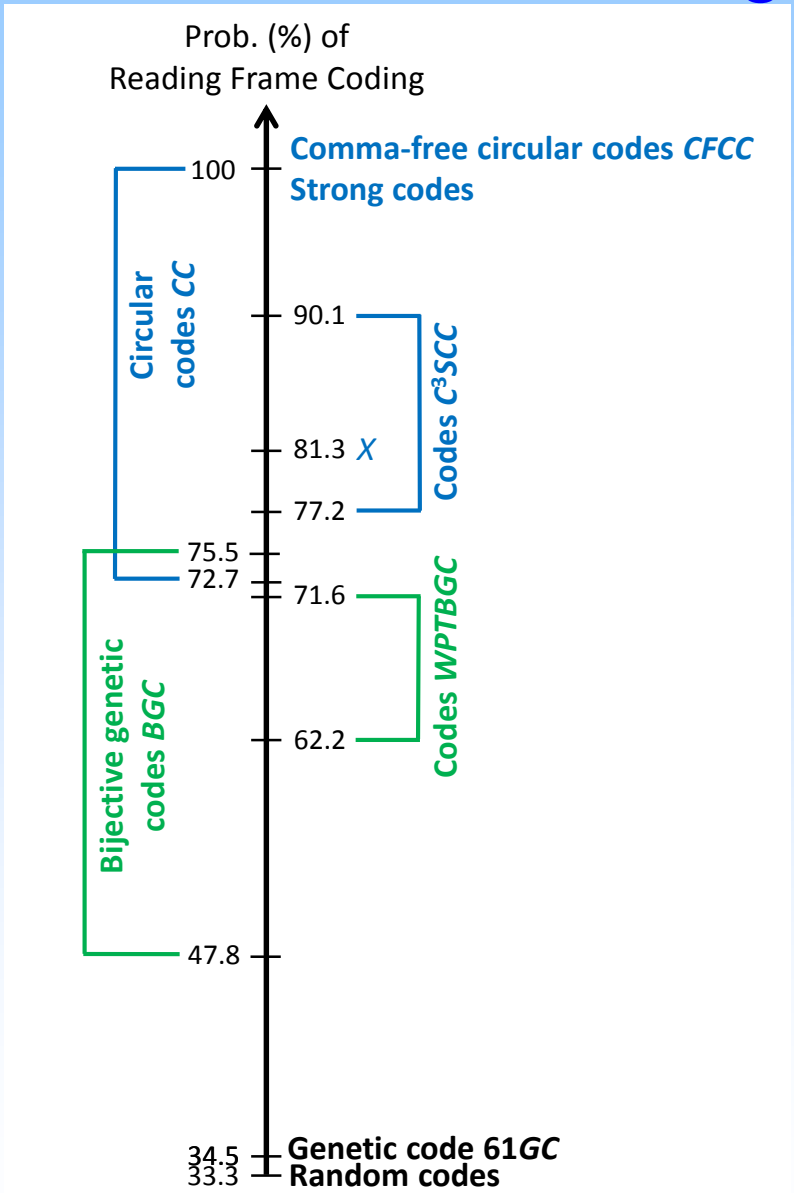
3 amino acids are encoded by 6 codons

52 bijective genetic codes $WPTBGC$ without permuted trinucleotides (WPT)

without a periodic permuted trinucleotide $PPT = \{\mathrm{AAA}, \mathrm{CCC}, \mathrm{GGG}, \mathrm{TTT}\}$

without two non-periodic permuted trinucleotides $NPPT = \{t, \mathcal{P}(t)\}$

Results: Genetic scale of reading frame coding



Definition of an extended genetic scale of reading frame coding

Probability $\mathrm{PrRFC}(C_f, K)$ of reading frame coding of any trinucleotide code C_f in a given frame f of a gene kingdom K.

Let $\Pr(t_f, K)$ be the frequency of a trinucleotide t_f of a code C occurring in a frame $f \in \{0,1,2\}$ of a gene kingdom K.

The probability $\Pr(C_f, K)$ of a trinucleotide code C_f in a frame $f \in \{0,1,2\}$ of a gene kingdom K is equal to

$$\Pr(C_f, K) = \sum_{t_f \in C} \Pr(t_f, K)$$

with $\sum_{t_f \in C} \Pr(t_f, K) \leq 1.$

Definition of an extended genetic scale of reading frame coding

Then, the reading frame coding probability $\mathrm{PrRFC}(C)$ of a trinucleotide code C_f in a frame $f \in \{0,1,2\}$ of a gene kingdom K is equal to

$$\mathrm{PrRFC}(C_f, K) = \frac{\Pr(C_f, K)}{\sum_{f=0}^{2} \Pr(C_f, K)}.$$

Definition of an extended genetic scale of reading frame coding

Proposition 1. $0 \leq \mathrm{PrRFC}(C_f, K) \leq 1$

The more the RFC probability $\mathrm{PrRFC}(C_f, K)$ value is raised,

the more the trinucleotide code C_f in frame f has efficiency for coding its frame f.

(i) $\mathrm{PrRFC}(C_f, K) = 1$ if $0 < \Pr(C_f, K) \leq 1, \Pr(C_{f'}, K) = \Pr(C_{f''}, K) = 0$.

The code C_f in frame f only codes its frame f and its frame f is always retrieved.

(ii) $\mathrm{PrRFC}(C_f, K) = 0$ if $\Pr(C_f, K) = 0$ with:

(iia) $\Pr(C_{f'}, K) = 0$ and $0 < \Pr(C_{f''}, K) \leq 1$; or

(iib) $0 < \Pr(C_{f'}, K) \leq 1$ and $0 \leq \Pr(C_{f''}, K) \leq 1$.

The code C_f has no occurrence in frame f and the frame f is not coded.

(iii) $\mathrm{PrRFC}(C_f, K) = 1/3$ if $\Pr(C_{f''}, K) = 2\Pr(C_f, K) - \Pr(C_{f'}, K)$ with:

(iiia) $\Pr(C_f, K) = \Pr(C_{f'}, K) = 1$ (random codes $Rand$); or

(iiib) $0 < \Pr(C_f, K) \leq 1/2$ and $0 \leq \Pr(C_{f'}, K) \leq 2\Pr(C_f, K)$; or

(iiic) $1/2 < \Pr(C_f, K) < 1$ and $2\Pr(C_f, K) - 1 \leq \Pr(C_{f'}, K) \leq 1$.

Definition of an extended genetic scale of reading frame coding

Proposition 2. Let us denote $\mathrm{PrRFC}\left(C_f, K; \mathrm{Pr}\left(t_f, K\right)\right)$ the probability $\mathrm{PrRFC}\left(C_f, K\right)$ of a code C_f in a frame f of K as a function of its trinucleotides probabilities $\mathrm{Pr}\left(t_f, K\right)$. Let λ be a scalar. Then,

$$\mathrm{PrRFC}\left(C_f, K; \lambda \times \mathrm{Pr}\left(t_f, K\right)\right) = \mathrm{PrRFC}\left(C_f, K; \mathrm{Pr}\left(t_f, K\right)\right).$$

Proposition 2 allows to measure the RFC probability $\mathrm{PrRFC}\left(C_f, K\right)$ of a code C_f in a frame f of a gene kingdom K regardless its absolute trinucleotide probabilities $\mathrm{Pr}\left(t_f, K\right)$, only its relative trinucleotide probabilities $\mathrm{Pr}\left(t_f, K\right)$ determine the RFC probability $\mathrm{PrRFC}\left(C_f, K\right)$.

Thus, the normalization of trinucleotide probabilities $\mathrm{Pr}\left(t_f, K\right)$ is not necessary for comparing different RFC efficiencies, either for a given trinucleotide code with different usage or for usage of different trinucleotide codes.

Definition of an extended genetic scale of reading frame coding

Remark 1. The extended definition uses the observed trinucleotide frequencies in the three frames of a gene kingdom K. The two trinucleotide probabilities $\Pr(t_f)$ in the two shifted frames $f \in \{1,2\}$ are not estimated from a probability product as this information is available.

Remark 2. The extended definition is based on a sum of trinucleotide probabilities $\Pr(t_f, K)$ less or equal to 1. The previous definition cannot measure the reading frame coding of trinucleotide codes with different usage.

Remark 3. The extended definition satisfies the combinatorial properties of trinucleotides codes. In particular, the RFC probability is equal to 1 with the comma-free codes and the strong codes where the trinucleotides are only in the frame 0 and the RFC probability is equal to 1/3 with the random codes where the trinucleotides are in the three frames 0, 1 and 2 equiprobably.

Definition of an extended genetic scale of reading frame coding

The reading frame coding probability $\mathrm{PrRFC}(X, K)$ of the C^3 self-complementary circular code X in a gene kingdom K is equal to

$$\mathrm{PrRFC}(X, K) = \frac{\Pr(X, K)}{\Pr(X, K) + \sum_{f=1}^{2} \Pr(X_f, K)}.$$

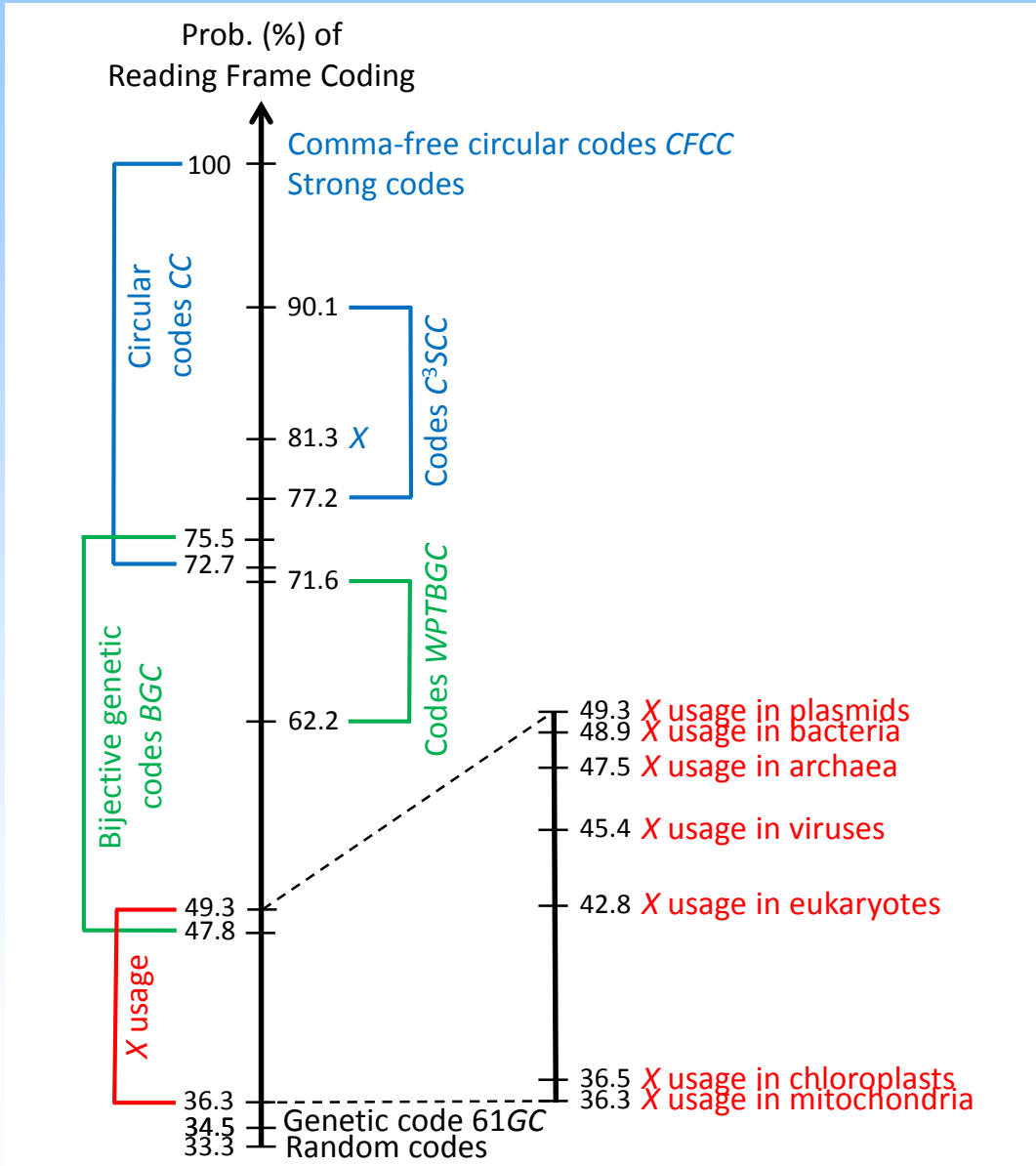
$$X = \{\mathrm{AAC}, \mathrm{AAG}, \mathrm{AAT}, \mathrm{CCA}, \mathrm{GAC}, \mathrm{TAC}, \mathrm{GCA}, \mathrm{GAG}, \mathrm{TAG}, \mathrm{TCA},$$
$$\mathrm{GAT}, \mathrm{TAT}, \mathrm{CCG}, \mathrm{CCT}, \mathrm{GCG}, \mathrm{TCG}, \mathrm{GCT}, \mathrm{TCT}, \mathrm{GGT}, \mathrm{TTG}\}.$$

Result:

K	Archaea A			Bacteria B			Eukaryotes E			Viruses V			Mitochondrion M			Chloroplasts C			Plasmids P		
$t \in X$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$	$f=0$	$f=1$	$f=2$
AAC	1.90	1.21	1.34	1.79	1.46	1.14	2.00	1.42	1.18	2.45	1.61	1.51	1.26	1.41	2.24	1.19	1.75	2.10	1.77	1.35	0.99
AAT	1.86	1.21	2.58	1.93	1.26	1.60	2.19	1.23	1.48	2.72	1.71	1.97	3.30	2.04	4.15	3.61	2.44	3.65	1.70	1.09	1.24
ACC	1.53	1.59	0.85	2.12	1.65	0.74	1.54	1.72	1.13	1.62	1.61	1.00	1.12	1.00	1.53	0.78	0.73	1.28	2.21	1.74	0.71
ATC	2.30	1.21	0.93	2.71	1.62	0.73	1.87	1.43	1.16	2.17	1.52	1.11	1.87	1.58	1.84	1.49	2.32	1.77	3.02	1.72	0.73
ATT	2.22	1.25	1.73	2.44	1.40	1.49	1.92	1.36	1.36	2.50	1.83	1.84	4.35	2.60	5.02	4.58	2.85	3.61	1.72	1.19	1.16
CAG	1.45	1.81	1.26	2.18	1.45	1.11	2.73	2.39	1.95	1.72	1.78	1.04	0.55	2.25	0.53	0.75	1.92	0.47	2.35	1.33	1.28
CTC	2.47	0.78	1.20	1.73	0.74	0.99	1.71	1.63	1.65	1.28	1.11	0.99	1.14	1.36	1.17	0.62	1.68	0.59	2.19	0.73	1.09
CTG	1.85	1.80	0.77	3.66	1.53	0.93	2.77	2.79	1.16	1.91	2.19	0.78	0.68	1.67	0.53	0.53	2.08	0.35	3.91	1.39	1.10
GAA	3.52	1.42	2.33	3.47	0.72	1.87	3.24	1.18	2.92	3.58	0.97	2.29	2.65	1.78	1.17	4.12	2.07	1.88	2.99	0.69	1.96
GAC	3.17	0.54	1.50	2.63	0.44	1.44	2.38	0.68	1.32	2.92	0.57	1.37	0.80	0.84	0.64	0.79	0.86	0.68	2.90	0.51	1.75
GAG	4.18	1.25	1.44	2.62	0.65	0.65	3.64	1.34	1.51	2.67	0.94	0.81	0.86	1.47	0.59	1.24	1.41	0.68	2.88	0.64	0.77
GAT	2.76	0.66	2.47	2.80	0.40	2.36	2.77	0.60	1.67	3.22	0.64	1.81	2.08	1.53	1.27	3.20	1.27	1.60	2.57	0.42	2.53
GCC	2.42	0.87	1.11	3.54	1.61	1.63	2.17	1.35	1.49	2.03	0.96	1.12	1.18	0.52	0.54	0.75	0.41	0.50	4.16	1.88	1.86
GGC	2.29	0.97	2.22	3.34	1.27	3.00	1.80	1.08	2.03	1.94	0.80	1.94	0.67	0.38	0.83	0.61	0.56	0.95	3.68	1.60	3.46
GGT	1.59	0.73	2.04	1.76	0.54	2.33	1.43	0.67	1.60	2.11	0.71	1.57	2.19	0.60	0.80	2.52	0.78	0.90	1.38	0.61	2.36
GTA	1.56	0.90	0.87	1.08	0.95	0.59	0.85	0.85	0.67	1.41	1.41	0.91	2.59	1.58	0.44	2.25	1.85	0.45	0.80	0.70	0.52
GTC	2.50	0.52	0.75	2.04	0.79	0.86	1.41	0.87	1.18	1.54	0.89	1.05	0.77	0.85	0.49	0.58	1.21	0.57	2.50	0.77	0.99
GTT	2.16	0.58	1.58	1.52	0.86	1.67	1.59	0.80	1.41	1.96	0.97	1.73	1.96	1.53	1.27	2.26	1.81	1.23	1.25	0.67	1.53
TAC	1.98	0.91	1.03	1.32	0.75	1.07	1.44	0.77	0.99	1.75	0.91	1.47	0.96	2.11	1.92	0.77	1.68	1.95	1.26	0.60	0.86
TTC	2.14	1.29	1.08	1.94	1.33	1.15	1.90	1.62	1.57	1.87	1.35	1.30	2.35	2.36	2.03	1.89	2.89	2.40	2.22	1.34	1.02

Table 1. Observed trinucleotide frequencies $\Pr(t_0, K)$ ($\Pr(t_1, K)$ and $\Pr(t_2, K)$, respectively) of the C^3 self-complementary circular code X in reading frame $f = 0$ (usage XU) (in the shifted frames $f = 1$ and $f = 2$, respectively) of genes in kingdoms K of archaea A (357,142 genes, 101,350,970 trinucleotides), bacteria B (7,862,438 genes, 2,484,909,928 trinucleotides), nuclear eukaryotes E (1,891,168 genes, 940,289,792 trinucleotides), viruses V (184,995 genes, 45,871,186 trinucleotides), mitochondrion M (1164 genes, 217,899 trinucleotides), chloroplasts C (1495 genes, 395,768 trinucleotides) and bacterial plasmids P (238,368 genes, 68,492,239 trinucleotides).

Result: Extended genetic scale of reading frame coding



Meeting in Mannheim – June 2015

X circular code motifs in the ribosome decoding center

Prof. Christian MICHEL

Theoretical Bioinformatics
ICube
University of Strasbourg, CNRS, France

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

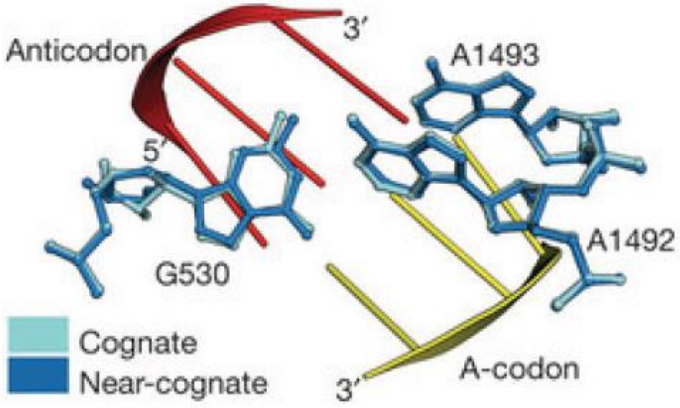
Biological recall

A1492 and A1493

The universally conserved nucleotides A1492 and A1493 are experimentaly proven to be **critical**. They are tasked with distinguishing cognate from non-cognate tRNAs by anticodon-codon interactions.

G530

The bacterial conserved nucleotide G530 has an experimentally proved biological function in the codon-anticodon binding.



Definition

X circular code motifs = motifs from the circular code X

$$X = \{\mathrm{AAC}, \mathrm{AAG}, \mathrm{AAT}, \mathrm{CCA}, \mathrm{GAC}, \mathrm{TAC}, \mathrm{GCA}, \mathrm{GAG}, \mathrm{TAG}, \mathrm{TCA},$$
$$\mathrm{GAT}, \mathrm{TAT}, \mathrm{CCG}, \mathrm{CCT}, \mathrm{GCG}, \mathrm{TCG}, \mathrm{GCT}, \mathrm{TCT}, \mathrm{GGT}, \mathrm{TTG}\}.$$

Result: *X* circular code motifs m_{AA}

PDB ID	Kingdom	Organism	*X* circular code motifs m_{AA}	Start	End	Length
3J5T	Bacteria	*E. coli*	G,GGT,G**AA**,GTC,GTA,AC	1487	1501	15
3I8G	Bacteria	*T. thermophilus*	G,G**AA**,GGT,GC	1490	1498	9
3J20	Archaea	*P. furiosus*	A,G**AA**,GTC,GTA,AC	1445	1456	12
3IZE	Eukaryote, nuclear	*S. cerevisiae*	**AA**,GTC,GTA,AC	1755	1764	10
3J5Z	Eukaryote, nuclear	*T. aestivum*	A,G**AA**,GTC,GTA,AC	1763	1774	12
3J3D	Eukaryote, nuclear	*H. sapiens*	**AA**,GTC,GTA,AC	1824	1833	10
3BBN	Eukaryote, chloroplast	*S. oleracea*	GT,G**AA**,GTC,GTA,AC	1438	1450	13

Result: *X* circular code motifs m_G

PDB ID	Kingdom	Organism	*X* circular code motifs m_G	Start	End	Length
3J5T	Bacteria	*E. coli*	GC,G**G**T,AAT,AC	527	536	10
3I8G	Bacteria	*T. thermophilus*	GC,**G**TT,ACC,C	528	536	9
3J20	Archaea	*P. furiosus*	GC,G**G**T,AAT,ACC,GGC,GGC,C	480	497	18

PDB ID	Kingdom	Organism	*X* circular code motifs m_G	Start	End	Length
3IZE	Eukaryote, nuclear	*S. cerevisiae*	GC,G**G**T,AAT,T	574	582	9
3J5Z	Eukaryote, nuclear	*T. aestivum*	GC,G**G**T,AAT,T	578	586	9
3J3D	Eukaryote, nuclear	*H. sapiens*	GC,G**G**T,AAT,T	623	631	9
3BBN	Eukaryote, chloroplast	*S. oleracea*	GC,G**G**T,AA	475	481	7

Result: *X* circular code motifs *m*

PDB ID	Kingdom	Organism	*X* circular code motifs *m*	Start	End	Length
3J5T	Bacteria	*E. coli*	AC,ACC,GCC,C	1396	1404	9
3I8G	Bacteria	*T. thermophilus*	AC,ACC,GCC,C	1375	1383	9
3J20	Archaea	*P. furiosus*	AC,ACC,GCC,C	1356	1364	9
3IZE	Eukaryote, nuclear	*S. cerevisiae*	AC,ACC,GCC,C	1633	1641	9
3J5Z	Eukaryote, nuclear	*T. aestivum*	AC,ACC,GCC,C	1641	1649	9
3J3D	Eukaryote, nuclear	*H. sapiens*	AC,ACC,GCC,C	1697	1705	9
3BBN	Eukaryote, chloroplast	*S. oleracea*	AC,ACC,GCC,C	1345	1353	9

Result: Spatial visualisation of *X* circular code motifs

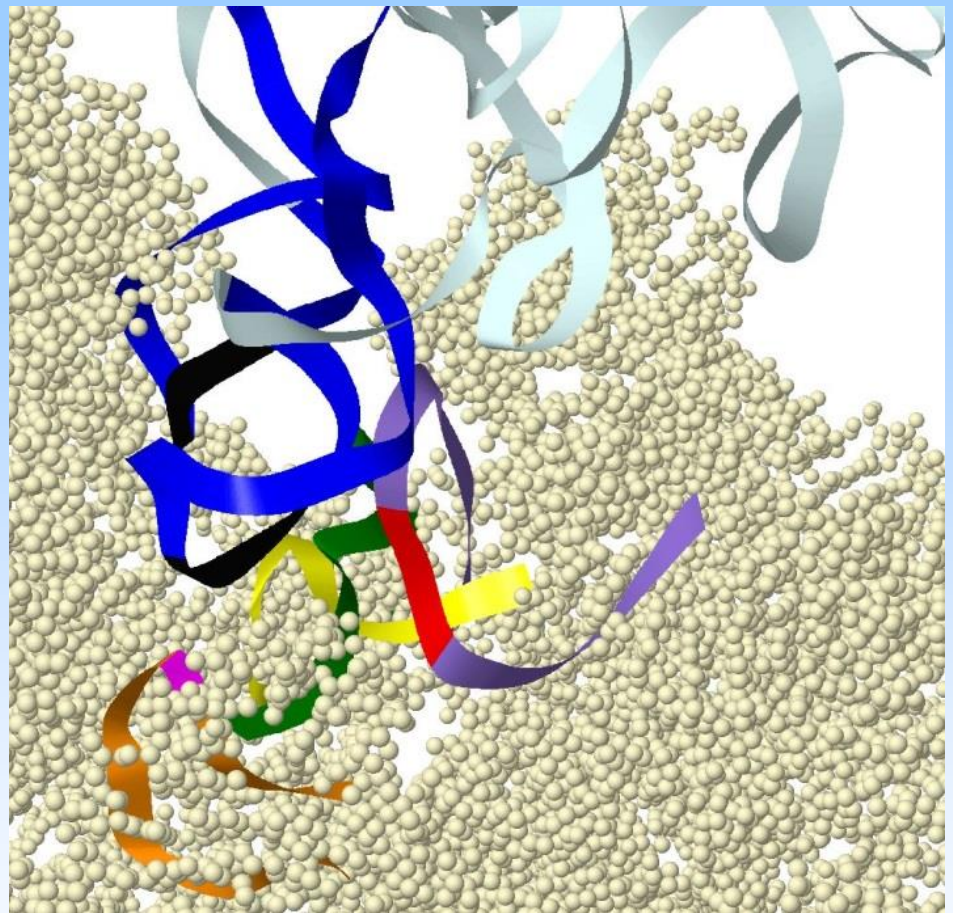
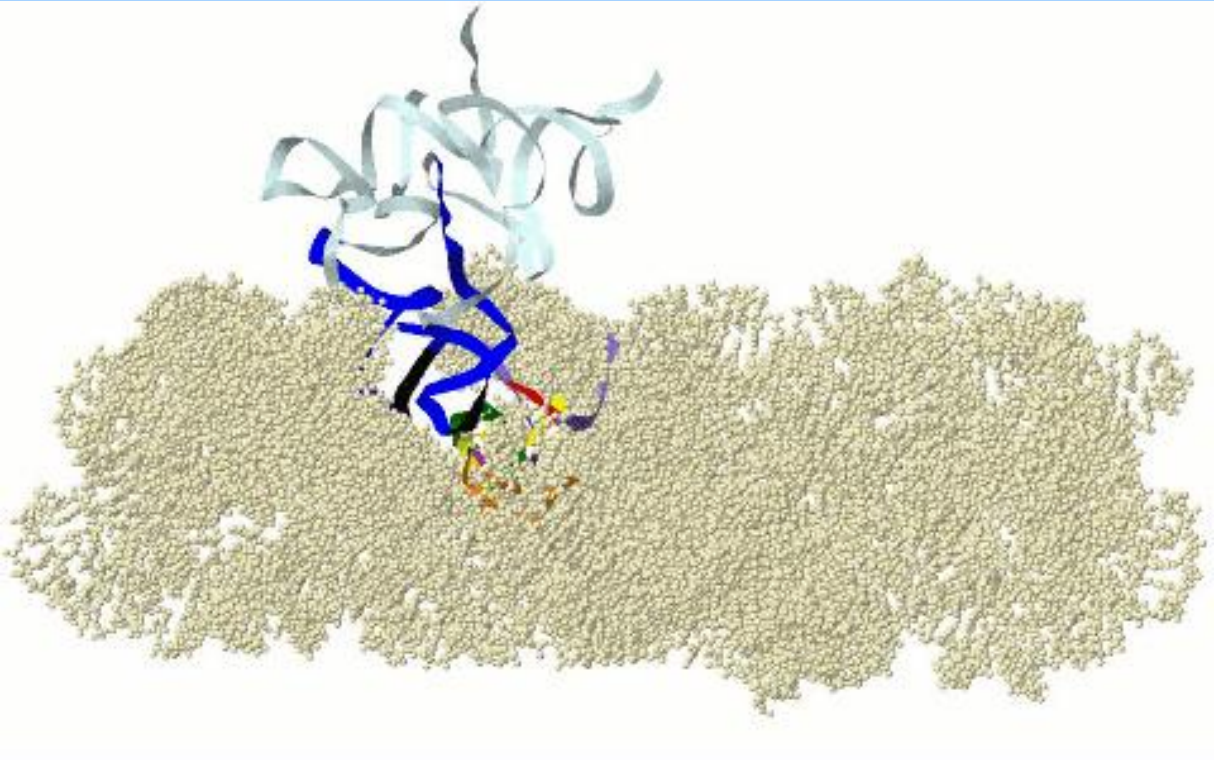


Figure. *X* circular code motifs in the bacterial ribosome decoding center of *Escherichia coli* (PDB 3J5T):
- the mRNA *X* motifs (green),
- the rRNA *X* motif m_{AA} (*E. coli*, 1487, 1501, 15) (purple with the conserved dinucleotide AA in red),
- the rRNA *X* motif m_{G} (*E. coli*, 527, 536, 10) (orange with the conserved nucleotide G in fuchsia),
- the rRNA *X* motif *m* (*E. coli*, 1396, 1404, 9) (yellow) and
- the tRNA *X* motifs (blue with the anticodon in black).

The remaining rRNA (lemonchiffon) is outside the neighborhood of these *X* motifs.

Circular code motifs in the ribosome decoding center



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