CIRCULAR CODES IN GENES

Christian MICHEL

- 2007 -

LSIIT (UMR CNRS-ULP 7005)
Université Louis Pasteur de Strasbourg

Pole API, Boulevard Sébastien Brant
67400 Illkirch, France

Email: michel@dpt-info.u-strasbg.fr

Conclusion

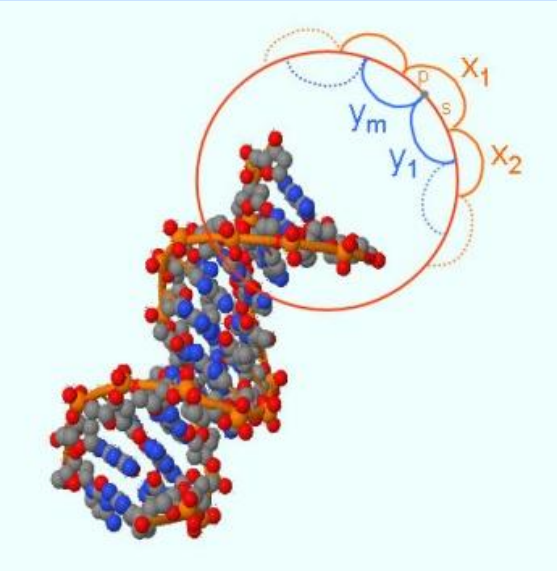
In genes, it exists

- genetic codes for coding the amino acids, the most important one is the universal genetic code;
- **circular codes for retrieving the reading frames of genes.**

It is still not known to date which biological apparatus could have used these circular codes.

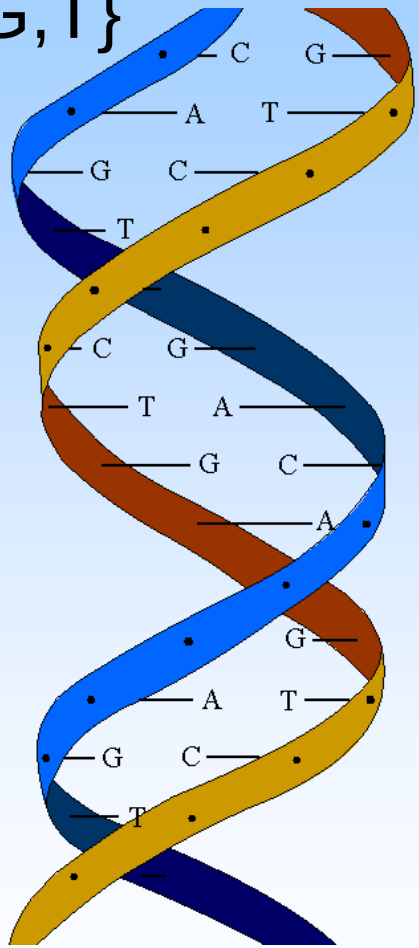
IDENTIFICATION

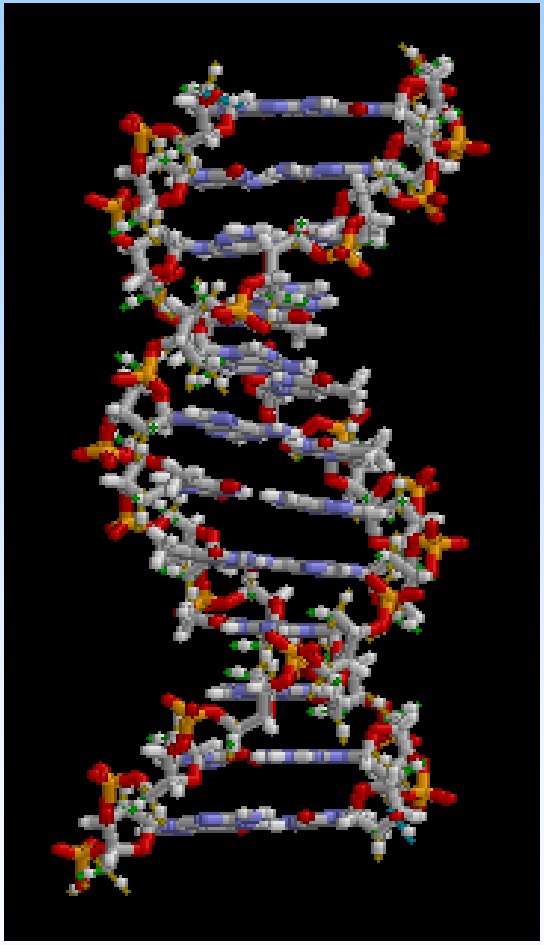
OF CIRCULAR CODES IN GENES



Structure of genomes: ADN

- series of 4 bases: {A,C,G,T}
- double helix
- complementary pairing
  A – T and C – G
- antiparallel





Definition: complementary *C* of a trinucleotide

The complementary pairing *C* of a base

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

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

The complementary pairing *C* of a trinucleotide

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

with $l_0, l_1, l_2 \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, is

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

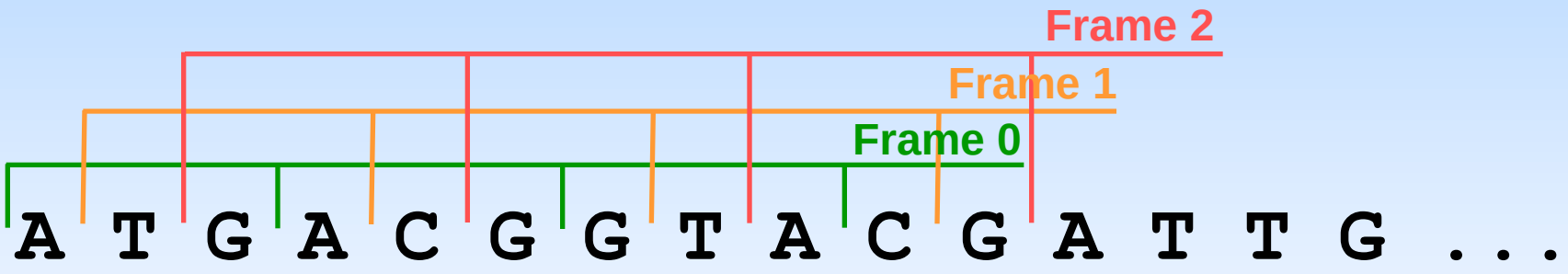
e.g. $C(\mathrm{ACG}) = \mathrm{CGT}$

Definition: frames of genes or words

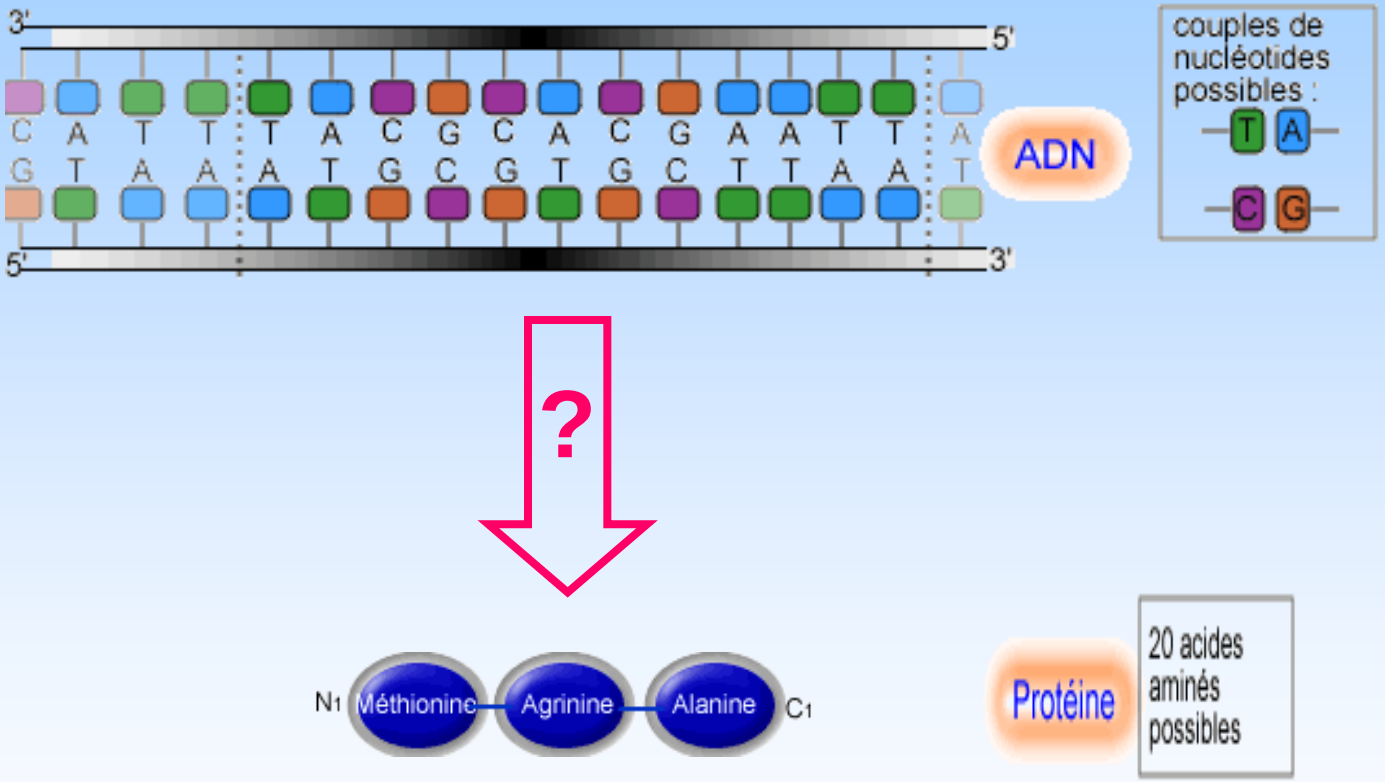
Frame 0: Reading frame established by a start codon

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

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



Structure of genes in genomes: which code for protein synthesis ?



Structure of genes: a code for proteins ?

A set X is a code if for any words

$x_1,x_2,\ldots,x_n,y_1,y_2,\ldots,y_m \in X$, $n,m \geq 1$,

the condition $x_1x_2\ldots x_n = y_1y_2\ldots y_m$

implies $n = m$ and $x_i = y_i$ for all $i \in [1,n]$.

Example: X = { A,GC,AGC } is not a code.

A,GC = AGC

Example: The trinucleotide code

A^3 = {AAA,...,TTT} (genetic code) is a code.

Structure of genes: a comma-free code ?

A trinucleotide code X is comma-free if

for each $y \in X$ and $u,v \in A^*$ such that

$uyv = x_1x_2\ldots x_n$ with $x_1,x_2,\ldots,x_n \in X$

implies $u,v \in X^*$.

Example (Eigen and Schuster, 1978; Crick et al., 1957): X = {RNY} = {RRY,RYY} with R = {A,G}, Y = {C,T}, N = {R,Y}, is a comma-free code.

Definition: permutation *P* of a trinucleotide

The permutation *P* of a trinucleotide

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

$l_0, l_1, l_2 \in \{A,C,G,T\}$,

is the permuted trinucleotide

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

and

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

e.g. $P(AAC)=ACA$ and $P(P(AAC))=P(ACA)=CAA$.

Structure of genes: a comma-free code ?

Rule: **A comma-free code cannot contain simultaneously a word and its permuted words.**

Example: X = {ACG,CGA,GAC}

By applying the comma-free definition

A,CGA,CG = ACG,ACG

with u = A, y = CGA, v = CG and x_1 = ACG

does not imply that u,v $\in$ X.

Structure of genes: a comma-free code ?

Consequence: AAA, CCC, GGG and TTT which are permuted words, cannot belong to a comma-free code.

Rule: The 60 remaining codons are classified into 20 classes of 3 permuted codons, e.g. the class X = {ACG,CGA,GAC}. A comma-free code has one word per class and therefore contains **at most 20 trinucleotides (maximal code).**

Structure of genes: a comma-free code ?

Biological property (Crick et al., 1957): a comma-free code assigs one codon per amino acid without ambiguity.

Structure of genes: a comma-free code ?

Result (Golomb et al., 1958):

There are 408 maximal comma-free codes of 20 codons.

e.g. X = { ACA, ACC, AAT, ACT, AGA, CGA, AGG, AGT, CGG, CGT,
GCA, GCC, TCA, TCC, GCT, TCT, TGA, TTA, TGG, TGT }

Structure of genes: a comma-free code ?

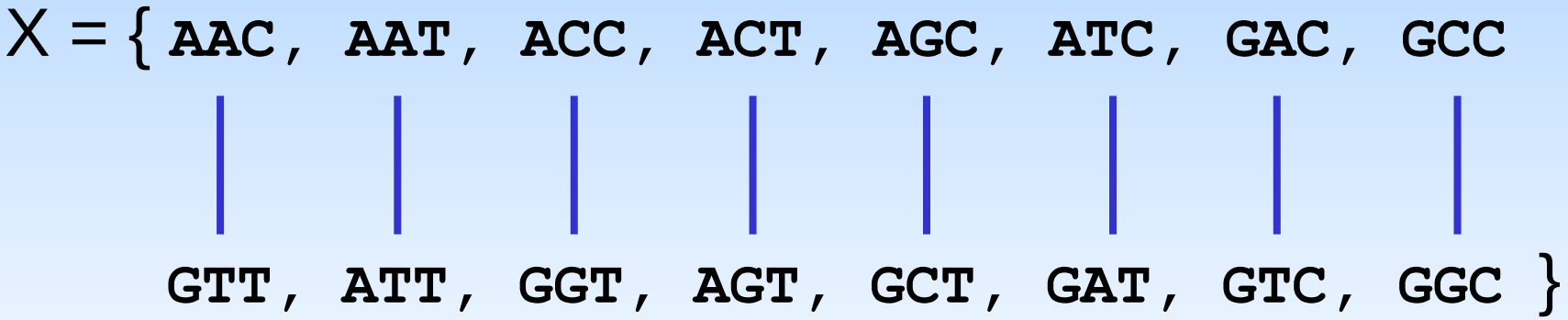
Problems:

- If a comma-free code defines a reading frame (frame 0), it has no word in a shifted frame (frames 1 or 2) (by definition).

- There is no maximal complementary comma-free code of 20 codons (impossible pairing in the DNA double helix).

Structure of genes: a comma-free code ?

Result: **There are 4 complementary comma-free code with 16 codons (maximal size)**



Structure of genes: a comma-free code ?

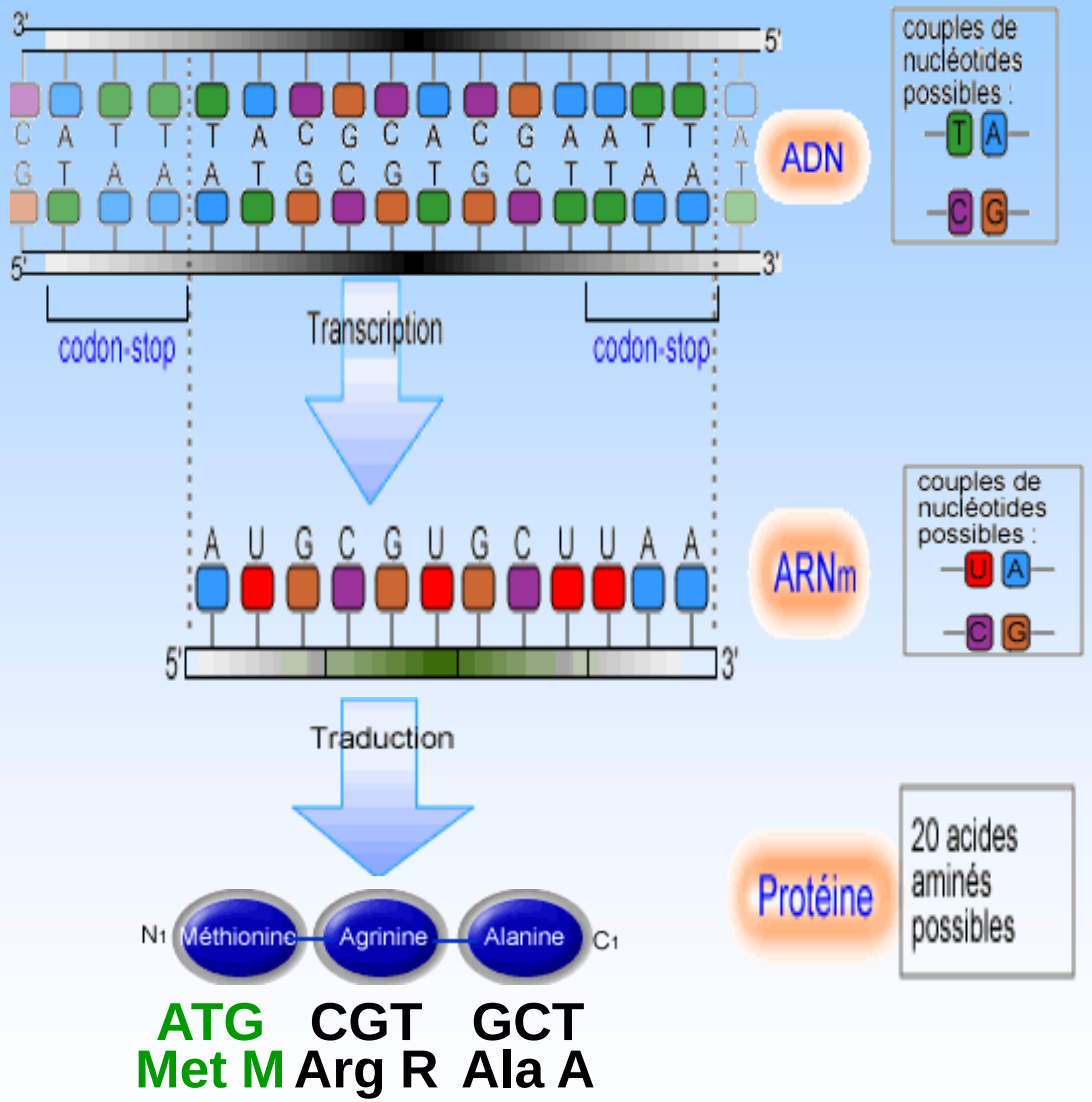
Biological property:

- Evolution by mutation in a comma-free code is restricted.
- Discovery that the codon TTT, an excluded trinucleotide in a comma-free code, codes phenylalanine (Nirenberg and Matthaei, 1961).
- The comma free code concept is abandoned.
- Discovery of the codons and the genetic code.

Structure of genes: the (universal) genetic code

	T	C	A	G	
T	TTT Phe F TTC Phe F TTA Leu L TTG Leu L	TCT Ser S TCC Ser S TCA Ser S TCG Ser S	TAT Tyr Y TAC Tyr Y TAA Stop TAG Stop	TGT Cys C TGC Cys C TGA Stop TGG Trp W	T
C	CTT Leu L CTC Leu L CTA Leu L CTG Leu L	CCT Pro P CCC Pro P CCA Pro P CCG Pro P	CAT His H CAC His H CAA Gln Q CAG Gln Q	CGT Arg R CGC Arg R CGA Arg R CGG Arg R	C
A	ATT Ile I ATC Ile I ATA Ile I ATG Met M	ACT Thr T ACC Thr T ACA Thr T ACG Thr T	AAT Asn N AAC Asn N AAA Lys K AAG Lys K	AGT Ser S AGC Ser S AGA Arg R AGG Arg R	A
G	GTT Val V GTC Val V GTA Val V GTG Val V	GCT Ala A GCC Ala A GCA Ala A GCG Ala A	GAT Asp D GAC Asp D GAA Glu E GAG Glu E	GGT Gly G GGC Gly G GGA Gly G GGG Gly G	G

Structure of genes: the genetic code



Statistical signals in the 3 frames of genes

Definition:

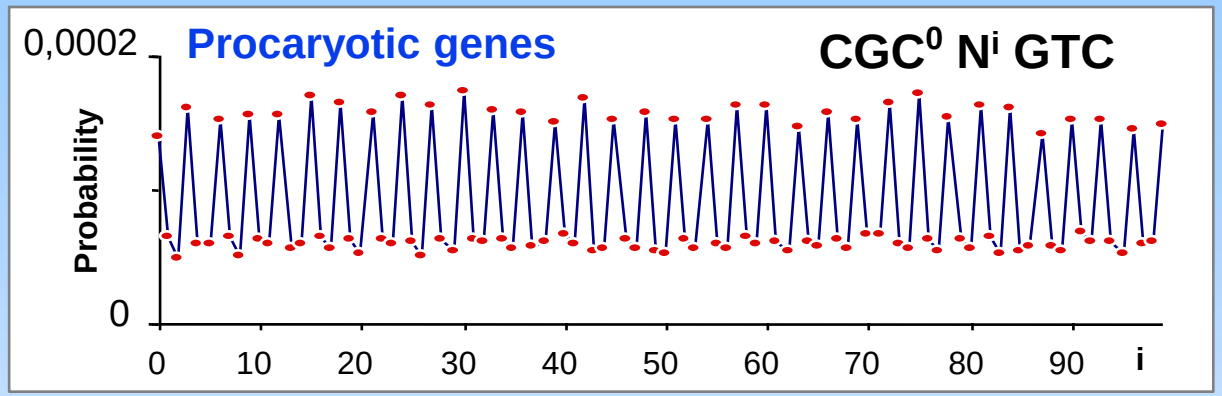
The correlation function gives the occurrence probability that a word *w*' occurs any *i* bases N after a word *w* in a gene population.

This correlation function is associated with the i-motif *w* N^i *w*'.

Statistical signals in the 3 frames of genes

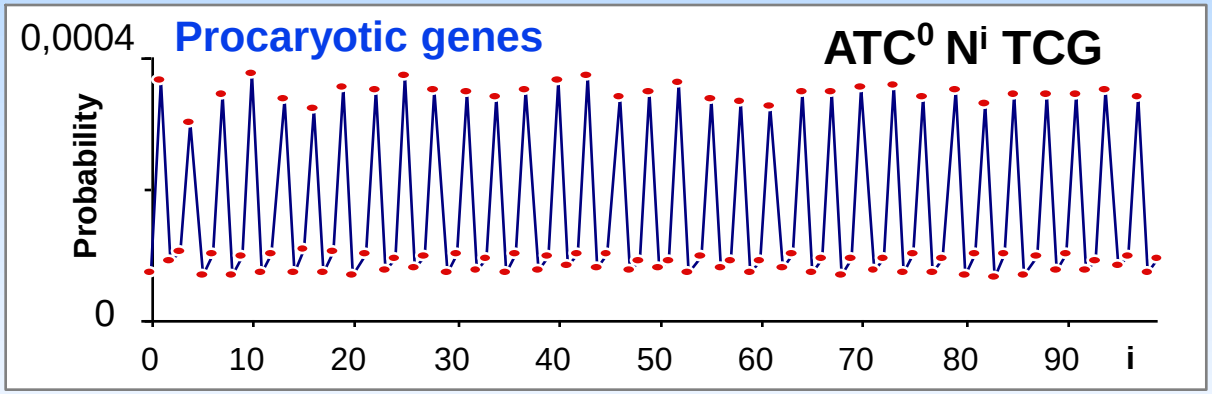
Periodicity 0 modulo 3

0, 3, 6, …



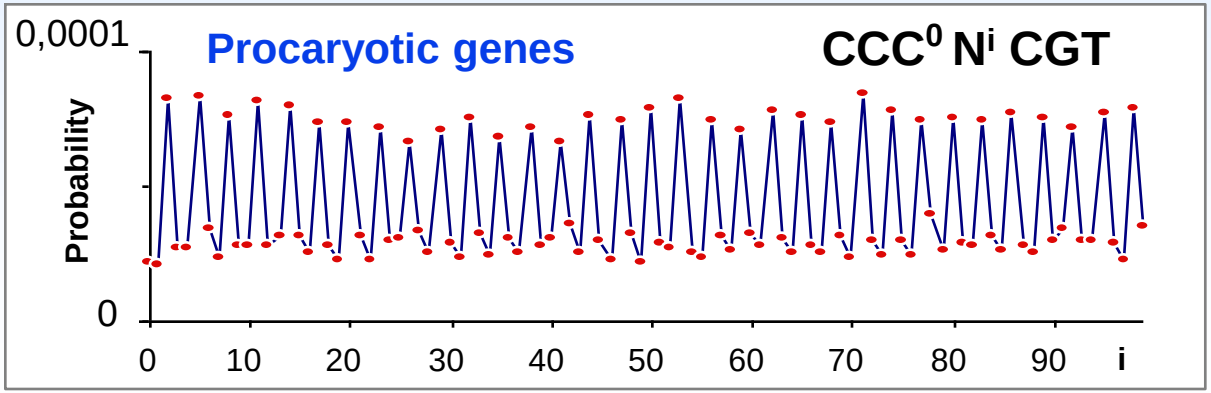
Periodicity 1 modulo 3

1, 4, 7, …



Periodicity 2 modulo 3

2, 5, 8, …



Preferential trinucleotides in the 3 frames of genes

Result: **Existence of 3 types of periodicity suggests preferential trinucleotides per frame**, i.e. trinucleotides in frame 0 (reading frame) and in shifted frames 1 and 2.

Definition: A simple method for identifying preferential trinucleotides in each frame of genes

Trinucleotide frequency method per frame

- 3 frames in genes
- Computation of the trinucleotide frequencies (%) in the 3 frames of genes (64×3 = 192 values)
- Each trinucleotide is preferentially associated with the frame in which it occurs with the highest frequency
- Example: genes of Fusobacterium Nucleotum

	Frame 0	Frame 1	Frame 2
AAC	0,71	1,3	2,47
AAG	1,58	5,55	2,62
…	…	…	…
TTG	0,94	4,66	0,76
	X_0	X_1	X_2

3 sets X_0, X_1, X_2 of 20 trinucleotides per frame

X_0 in frame 0, X_1 in frame 1, X_2 in frame 2

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

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

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

X_0, X_1 and X_2 are identified in genes of both eukaryotes and procaryotes

X_0, X_1, X_2: permutation property *P*

$$P(X_0) = X_1 \text{ and } P(X_1) = X_2$$

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

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

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

X_0, X_1, X_2: complementary property C

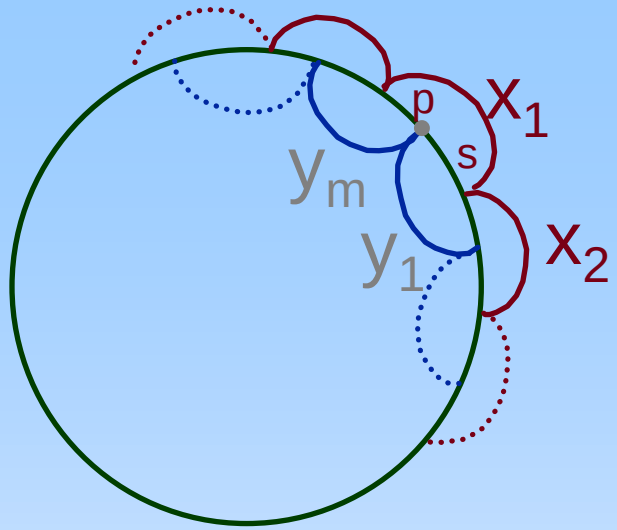
$$C(X_0) = X_0 \text{ and } C(X_1) = X_2$$

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

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

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

X_0, X_1, X_2: circular codes



X_0, X_1, X_2 are circular codes

A trinucleotide code X is circular if

for all $x_1,x_2,\ldots,x_n,y_1,y_2,\ldots,y_m \in X$,

$n,m \geq 1$, $p \in A^*$ and $s \in A^+$

the conditions $sx_2x_3\ldots x_np = y_1y_2\ldots y_m$ and $x_1 = ps$

imply $n = m$, $p = \varepsilon$ (empty word) and $x_i = y_i$

Circular codes: definition

The factorization of any word (over a circular code) written on a circle is unique

Example

Y = {GCG,CGC}

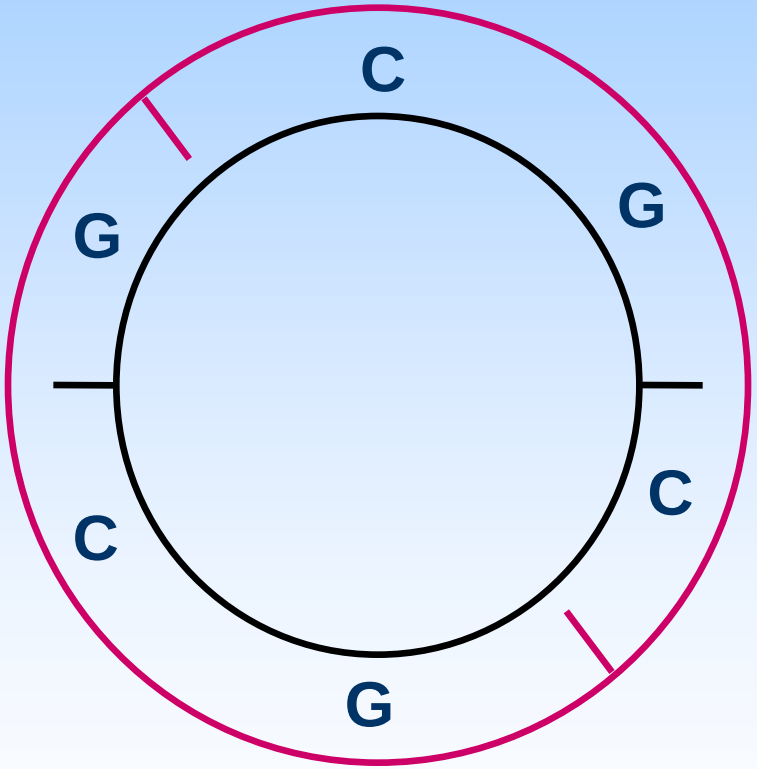
w = GCG,CGC

2 factorizations:

w = GCG,CGC

w = CGC,GCG

Y is not a circular code



Example

Y = {GGC,CGG} is a circular code

Circular codes: demonstration

X_0, X_1, X_2 are circular codes

- Classical demonstration: concept of flower automaton

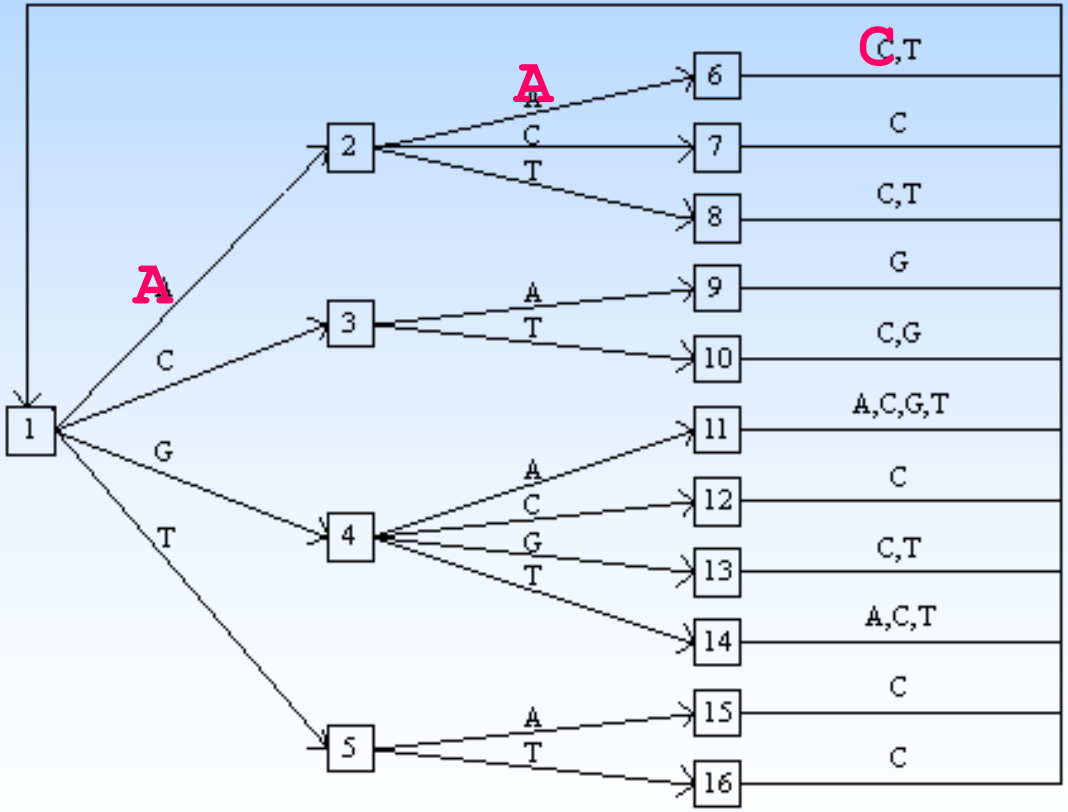
(Lassez, 1976; Berstel and Perrin, 1985)

- Recent demonstration: concept of necklace

(Pirillo, 2001)

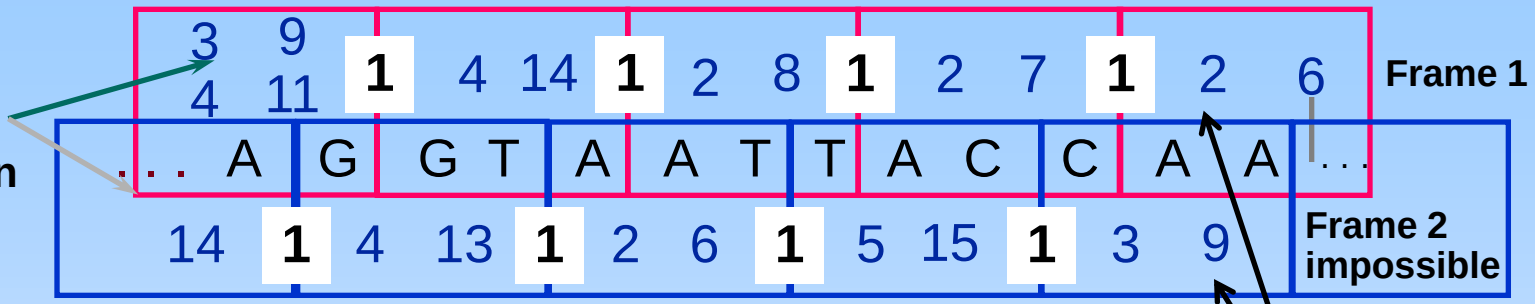
Circular codes: flower automaton

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

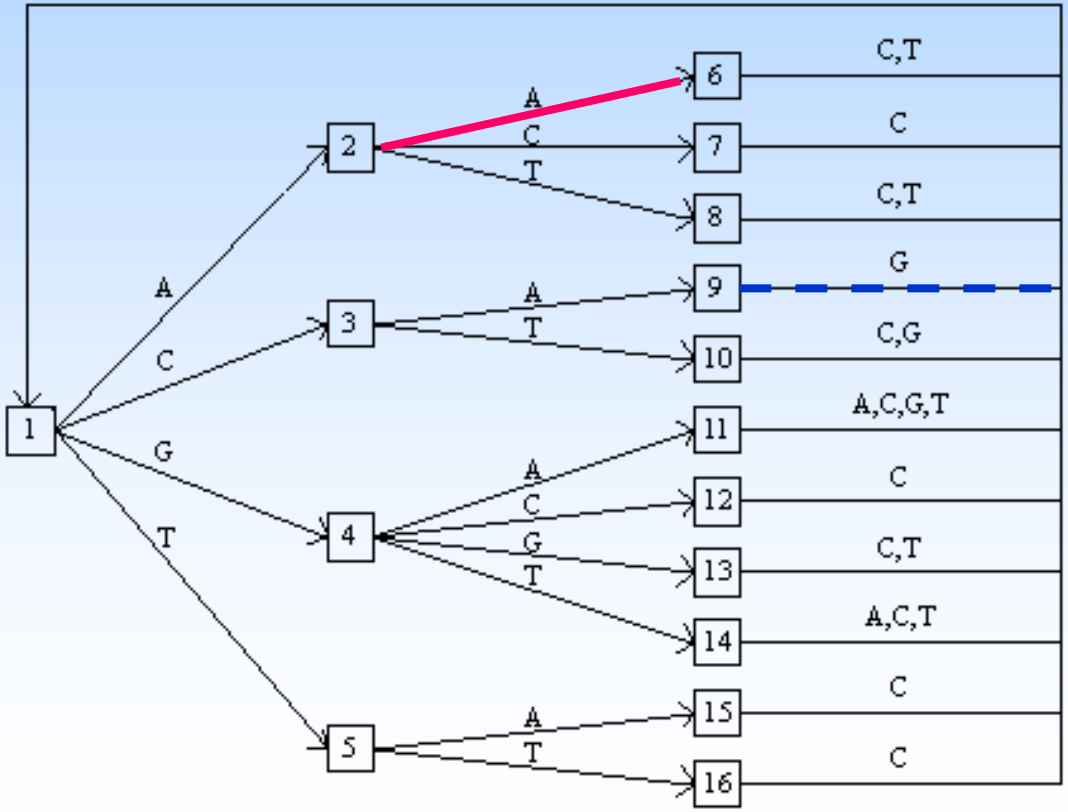


Circular codes: window

Which frame does retrieve the construction of the word ?



until



Circular codes: window

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

- Generation of a word over the circular code X_0

…AGGTAATTACCAATGTAAACTACTTCACCATC…

- Search of the construction frame

…,AGG,TAA,TTA,CCA,ATG,TAA,ACT,ACT,TCA,CCA,TC…

…A,GGT,AAT,TAC,CAA,TGT,AAA,CTA,CTT,CAC,CAT,C….

…AG,GTA,ATT,ACC,AAT,GTA,AAC,TAC,TTC,ACC,ATC,…

Circular codes: “misplaced” trinucleotides

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

- Generation of a word over 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…

TAA ∈ X_2 **TAC ∈ 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…

GGT ∈ X_0 AAC ∈ X_0 **ACT ∈ X_1**

Misplaced trinucleotides do not exist in the comma-free codes

Circular codes: maximal

X_0, X_1, X_2 are maximal circular codes

- A maximal circular code cannot be included in another circular code

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

Circular codes: C^3 code

X_0 is a C^3 code

X_0, $P(X_0)=X_1$ and $P(P(X_0))=X_2$ are circular codes

Counter example

Y_0 = {CGG,GCC} is a circular code

$P(Y_0) = Y_1$ = {GGC,CCG} is a circular code

$P(P(Y_0)) = Y_2$ = {GCG,CGC} is not a circular code

w = GCG,CGC in frame 0

w = GC,GCG,C in frame 2

Y_0 is a circular code but not C^3

Rarity of the complementary C^3 code X_0

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

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

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

Number of maximal complementary C^3 codes: **216**

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

Summary of the properties of the C^3 code X_0

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

- Maximal
- Length of the minimal window = 13 nucleotides
- C^3 code
- Self-complementary
- Misplaced trinucleotides
  - 11.9 % of X_0 and 12.7 % of X_2 in frame 1
  - 11.9 % of X_0 and 12.7 % of X_1 in frame 2
- Rarity: **6.2×10^{-8}**

Is the C^3 code X_0 unique in genes ?

- The C^3 code X_0 occurs in mean gene populations

- However
  - There are variant genetic codes
  - There is different codon usage: synonymous codons (codons coding for the same amino acid) occur with different frequencies in genes

- Development of new method considering the frequencies of permuted trinucleotides per frame (FPTF) generalizing the previous method considering the frequencies trinucleotide per frame (FTF)

Method FPTF: general principle

- Generalization of the classical method per frame
- Automatic identification of trinucleotides per frame
- Sensibility: comparison of the circular code signal with a value of random
- Application for massive statistical analyses in genomes

Method FPTF

$$P\left(w_i^p\right) = \frac{o\left(w_i^p\right)}{\sum_{p=0}^{2} o\left(w_i^p\right)}$$

$$Q\left(w_i^p\right) = \frac{o\left(w_i^p\right)}{\sum_{i=0}^{2} o\left(w_i^p\right)}$$

$$M\left(w_i^p\right) = \frac{1}{2}\left(P\left(w_i^p\right) + Q\left(w_i^p\right)\right)$$

Example $w_i^p = w_1^1$

	Frame 0	Frame 1	Frame 2
w_0	w_0^0	w_0^1	w_0^2
w_1	w_1^0	w_1^1	w_1^2
w_2	w_2^0	w_2^1	w_2^2

$$F(S) = F\left(\left\{w_{i_0}^{p_0}, w_{i_1}^{p_1}, w_{i_2}^{p_2}\right\}\right) = \frac{1}{3}\left(M\left(w_{i_0}^{p_0}\right) + M\left(w_{i_1}^{p_1}\right) + M\left(w_{i_2}^{p_2}\right)\right)$$

Method FPTF: example in a genome

- Trinucleotide frequencies (%) in Fusobacterium Nucleotum

S = { AAC, ACA, CAA }

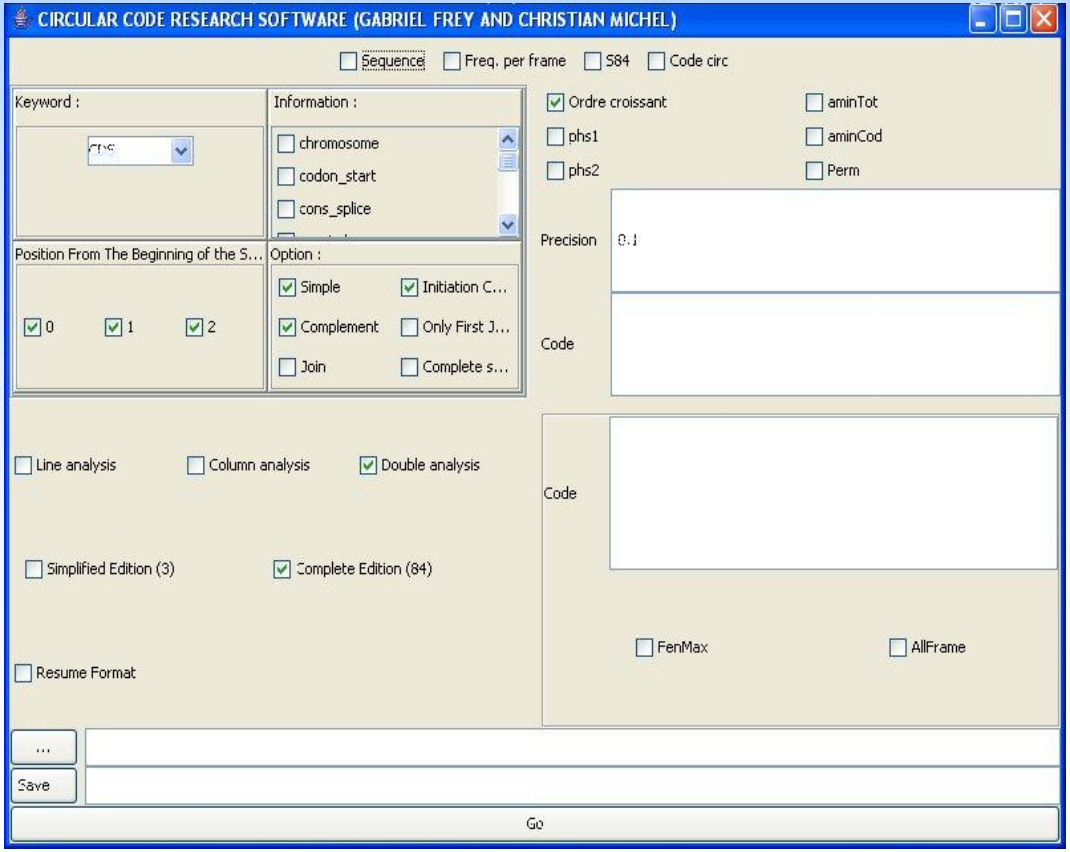
	Frame 0	Frame 1	Frame 2
AAC	0,71	1,3	2,47
ACA	2,36	0,71	1,5
CAA	1,97	3,36	0,71

- Statistical function $F(S)$ with S^{22}, S^{44} and S^{53}

	$F(S)$
$S^{22} = \{AAC^0;ACA^1;CAA^2\}$	0,143
$S^{44} = \{AAC^1;ACA^2;CAA^0\}$	0,316
$S^{53} = \{AAC^2;ACA^0;CAA^1\}$	0,541

Massive statistical analysis of genomes

- 175 genomes of bacteria and 16 genomes of archaeas: 487863 genes, 528097 kb
- Computation of the 64 trinucleotides frequencies in the 3 frames
- Application of the method FPTF to $(175+16) \times 3 = 573$ sets of 20 trinucleotides



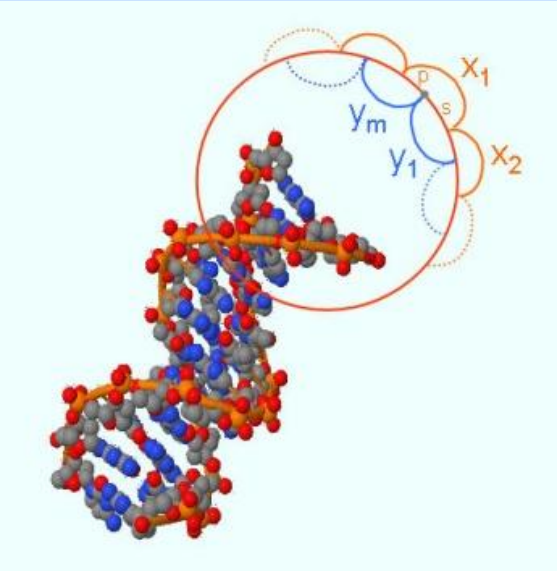
Results: identification of new circular codes

- 78 % of these 573 sets (20 trinucleotides) are maximal circular codes
- 58 % of these 573 sets (20 trinucleotides) are C^3 codes
- 87 new C^3 codes in the 175 genomes
  - the most frequent C^3 code is associated with 17 genomes
  - the probability of a C^3 code is rare: $221544 / 3^{20} = 6{,}3 \times 10^{-5}$

Nb	C^3 codes in bacterial genomes
17	AAC AAG AAT ACC GAC TAC CAG GAG TAG ATC ATG TAT GCC CTC GGC GTC CTG TTC GTG TTG
14	AAC GAA AAT ACC GAC TAC CAG GAG GTA ATC GAT ATT GCC CTC GCG GTC CTG CTT GTG GTT
12	CAA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GGC CGT GCT TCT GGT GTT
9	CAA GAA AAT CCA GAC ACT GCA GGA GTA CAT GAT ATT GCC CCT GCG CGT GCT TCT GGT GTT
7	ACA GAA AAT CCA ACG ACT GCA GGA GTA CAT GAT ATT CCG CCT GCG CGT GCT TCT GGT GTT

BIOLOGICAL PROPERTIES

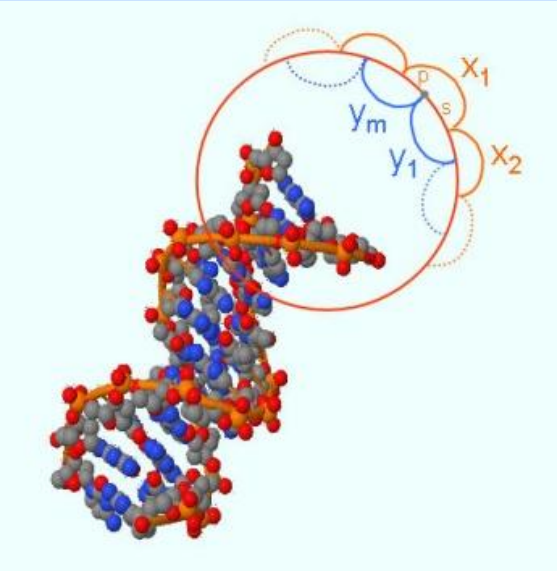
OF CIRCULAR CODES IN GENES



BIOLOGICAL PROPERTIES

OF CIRCULAR CODES IN GENES

READING FRAME



A factorization method for retrieving the reading frames of genes: problem

- The reading frame of any word generated by a circular code can be retrieved anywhere in the sequence by reading a few nucleotides (window of the circular code, e.g. 13 nucleotides with X_0)

- The current nucleotide sequences are not constructed by "pure" circular codes but by circular codes which have evolved by mutation

- A C^3 code has a circular code in each frame:
  C_0 in frame 0, C_1 in frame 1 and C_2 in frame 2

A factorization method for retrieving the reading frames of genes: problem

Problems:

Is it possible to retrieve the reading frames in current genes, i.e. in non-pure circular codes ?

And with how many letters ?

A factorization method for retrieving the reading frames of genes: method

- Short words *w* of lengths between 5 and 50 nucleotides are extracted from current genes

- The 3 frames of these short words *w* are factorized into words of C_0, C_1 and C_2

- The proposed frame *p* of a short word w^p is the frame such that the sum of the numbers of
- words of C_p in frame *p*
- words $C_{(p+1) \bmod 3}$ in frame (*p*+1) mod 3 and
- words $C_{(p+2) \bmod 3}$ in frame (*p*+2) mod 3
is maximum by varying *p* in {0,1,2}

A factorization method for retrieving the reading frames of genes: method

C_0 = { AAC, AAT, ACC, ATC, CAG, GAG, GAA, GAC, GCC, GTA,
GTT, ATT, GGT, GAT, CTG, CTC, TTC, GTC, GGC, TAC }
C_1 = { ACA, ATA, CCA, TCA, AGC, AGG, AAG, ACG, CCG, TAG,
TTG, TTA, GTG, ATG, TGC, TCC, TCT, TCG, GCG, ACT }
C_2 = { CAA, TAA, CAC, CAT, GCA, GGA, AGA, CGA, CGC, AGT,
TGT, TAT, TGG, TGA, GCT, CCT, CTT, CGT, CGG, CTA }

w = CGACTTCCAGA

w^0 = CGA,CTT,CCA,GA = C_2,C_2,C_1

w^1 = GAC,TTC,CAG,A = C_0,C_0,C_0

w^2 = ACT,TCC,AGA = C_1,C_1,C_2

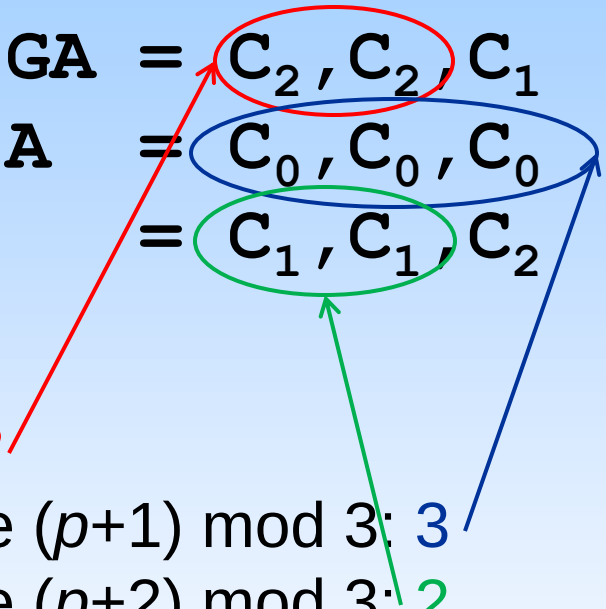
A factorization method for retrieving the reading frames of genes: method

w = CGACTTCCAGA

w^0 = CGA,CTT,CCA,GA = C_2, C_2, C_1

w^1 = GAC,TTC,CAG,A = C_0, C_0, C_0

w^2 = ACT,TCC,AGA = C_1, C_1, C_2



For $p = 2$:

- words of C_p in frame p: 2
- words $C_{(p+1) \bmod 3}$ in frame $(p+1) \bmod 3$: 3
- words $C_{(p+2) \bmod 3}$ in frame $(p+2) \bmod 3$: 2

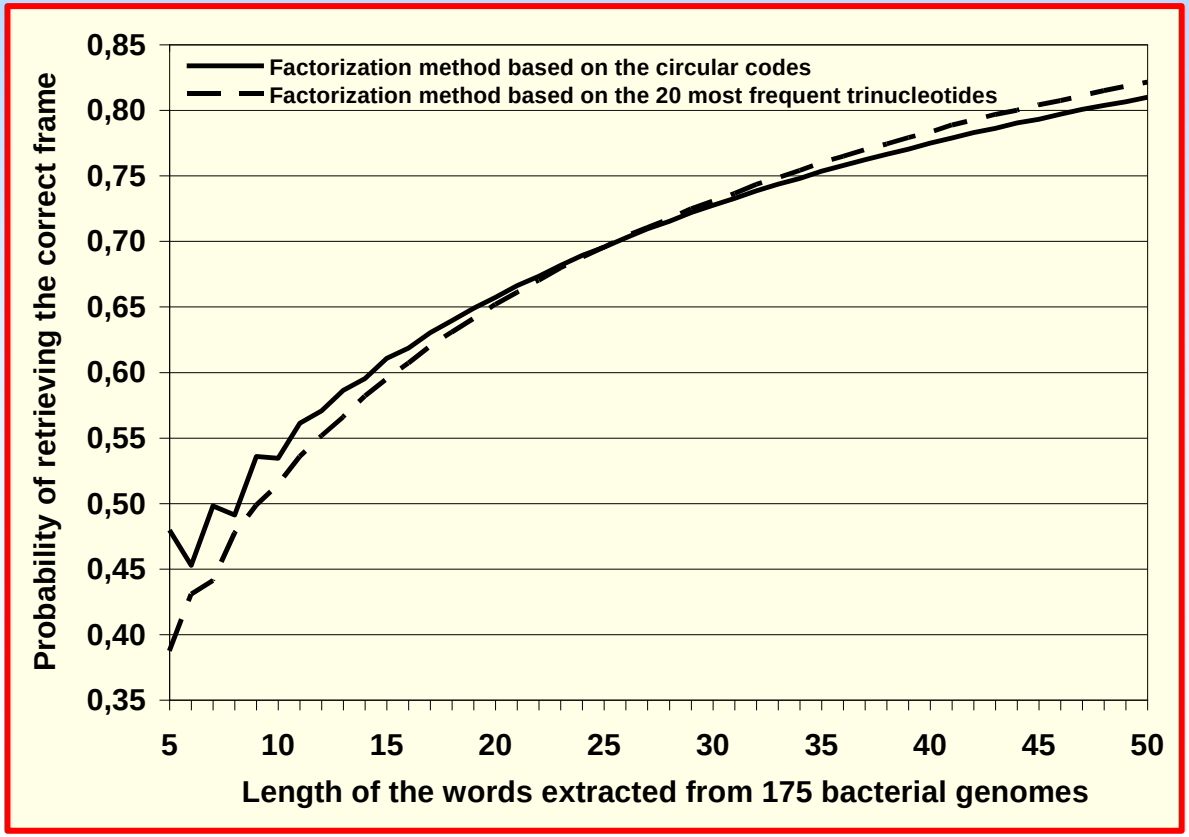
is maximum by varying p in {0,1,2}

$Nb(w^{p=0}) = 1$ $Nb(w^{p=1}) = 1$ $Nb(w^{p=2}) = 7$

The proposed frame p of w^p is the frame p=2

A factorization method for retrieving the reading frames of genes: results

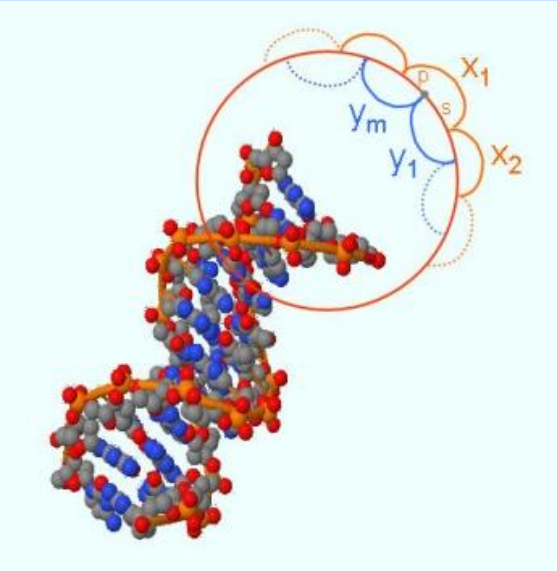
- The proposed frames are compared to the real ones: with words of 5 nucleotides, the frequency to retrieve the correct is frame is 48 % while in the random case, the frequency is 33 % (1 among 3 possibilities)
- Unexpectedly, with words less than 25 nucleotides, the results based on the circular words are better than the 20 words most frequent in the 3 frames



BIOLOGICAL PROPERTIES

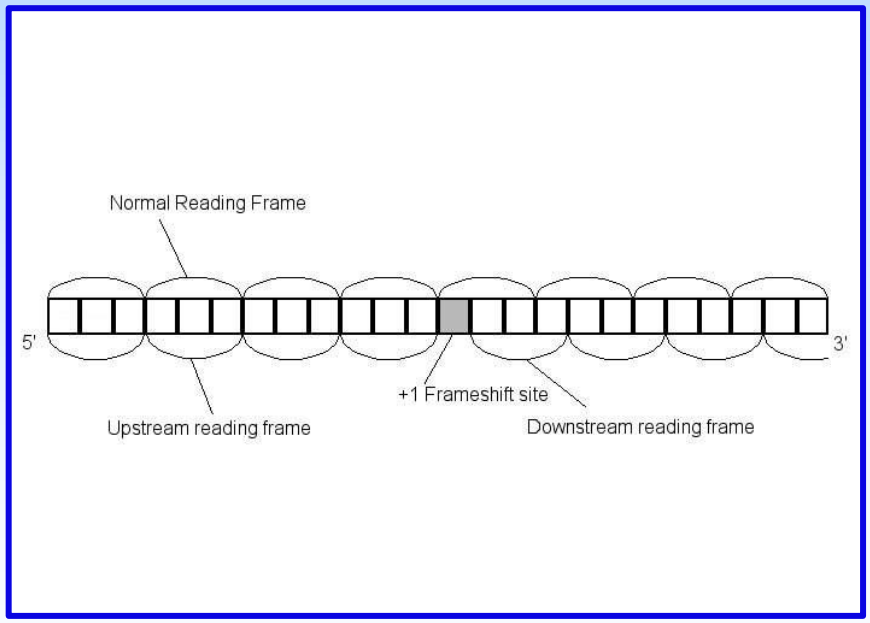
OF CIRCULAR CODES IN GENES

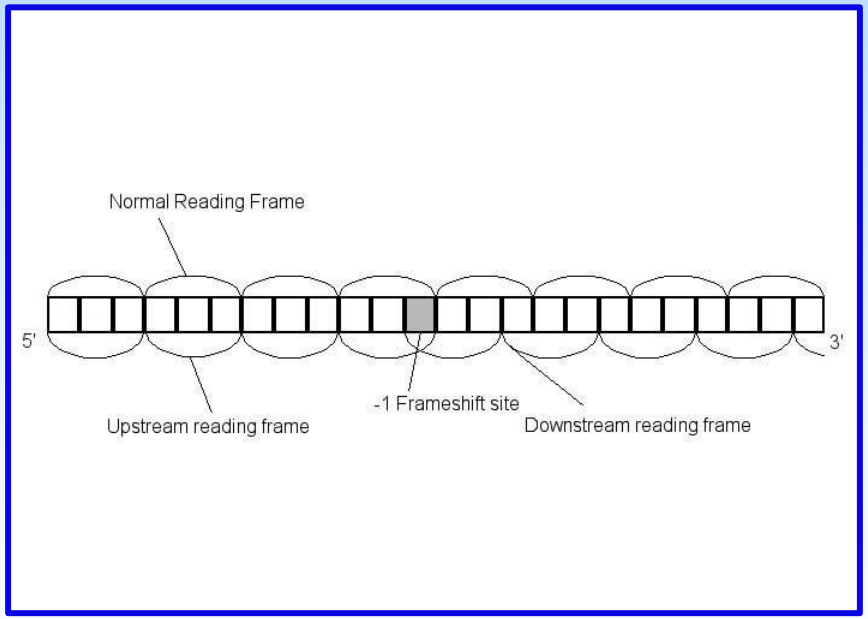
FRAMESHIFT GENES



Frameshift genes: problem

Frameshift genes: genes with a shift of 1 nucleotide in the 5'-3' or 3'-5' direction (to the right or the left): loss of the reading frame





Frameshift genes: problem

- Circular codes in genes have the property to retrieve the reading frames in genes

- Therefore, this property of circular code in frameshift genes must disappear

Frameshift genes: method

Definition of a score function P based on the C³ code X

Let $\mathbb{T} = \{AAA, ..., TTT\}$ be the set of 64 trinucleotides.

Let $t \in \mathbb{T}$ be a trinucleotide.

Let F be a frameshift gene population with $n(F)$ sequences s.

Each sequence s has a frameshift site in the nucleotide position $i = 0$.

Let t_0 be the first downstream trinucleotide after the frameshift site.

Let the method reading frame be $..., t_{-1}, t_0, t_1, ...$

The sequence s is considered as a series of trinucleotides t_i.

Let $w_i = t_{i_0} t_{i_1} t_{i2} t_{i_3}$ be a window of length $|w| = 4$ trinucleotides in the sequence s where t_{i_0} is the ith trinucleotide in s and, t_{i_j}, the jth trinucleotide in w_i.

Let X_f, $f \in \{0, 1, 2\}$, be the 3 codes X_0, X_1 and X_2 in the 3 frames f.

Frameshift genes: method

Definition of a score function P based on the C³ code X

In a given window w_i, the function $\delta_f(t_{i_j})$ indicates
if the trinucleotide t_{i_j} belongs or not to the code X_f

$$\delta_f(t_{i_j}) = \begin{cases} 1 & \text{if } t_{i_j} \in X_f \\ 0 & \text{otherwise} \end{cases}$$

with $f \in \{0,1,2\}$ and $j \in \{0,1,2,3\}$.

Each sequence s is associated with a frame $\mathfrak{f} \in \{0,1,2\}$.
In s_0 (s_1 and s_2 resp.), t_0 is the first downstream trinucleotide
(more one nucleotide and 2 nucleotides resp.) after the frameshift site.
Then, the score $P(X_f, i, s_\mathfrak{f})$ of the code X_f in a window w_i of a given frame $\mathfrak{f}$ of a sequence s is

$$P(X_f, i, s_\mathfrak{f}) = \frac{1}{|w|} \sum_{j=0}^{|w|-1} \delta_f(t_{i_j}).$$

Frameshift genes: method

Definition of a score function P based on the C³ code X

The score $P(i, s_{\mathfrak{f}})$ of the C^3 code X in a window w_i of a given frame $\mathfrak{f}$ of a sequence s measuring the frame retrieval intensity, is defined as

$$P(i, s_{\mathfrak{f}}) = \frac{1}{2} \sum_{\substack{f, f' \in \{0,1,2\} \\ f' > f}} \left| P(X_f, i, s_{\mathfrak{f}}) - P(X_{f'}, i, s_{\mathfrak{f}}) \right|.$$

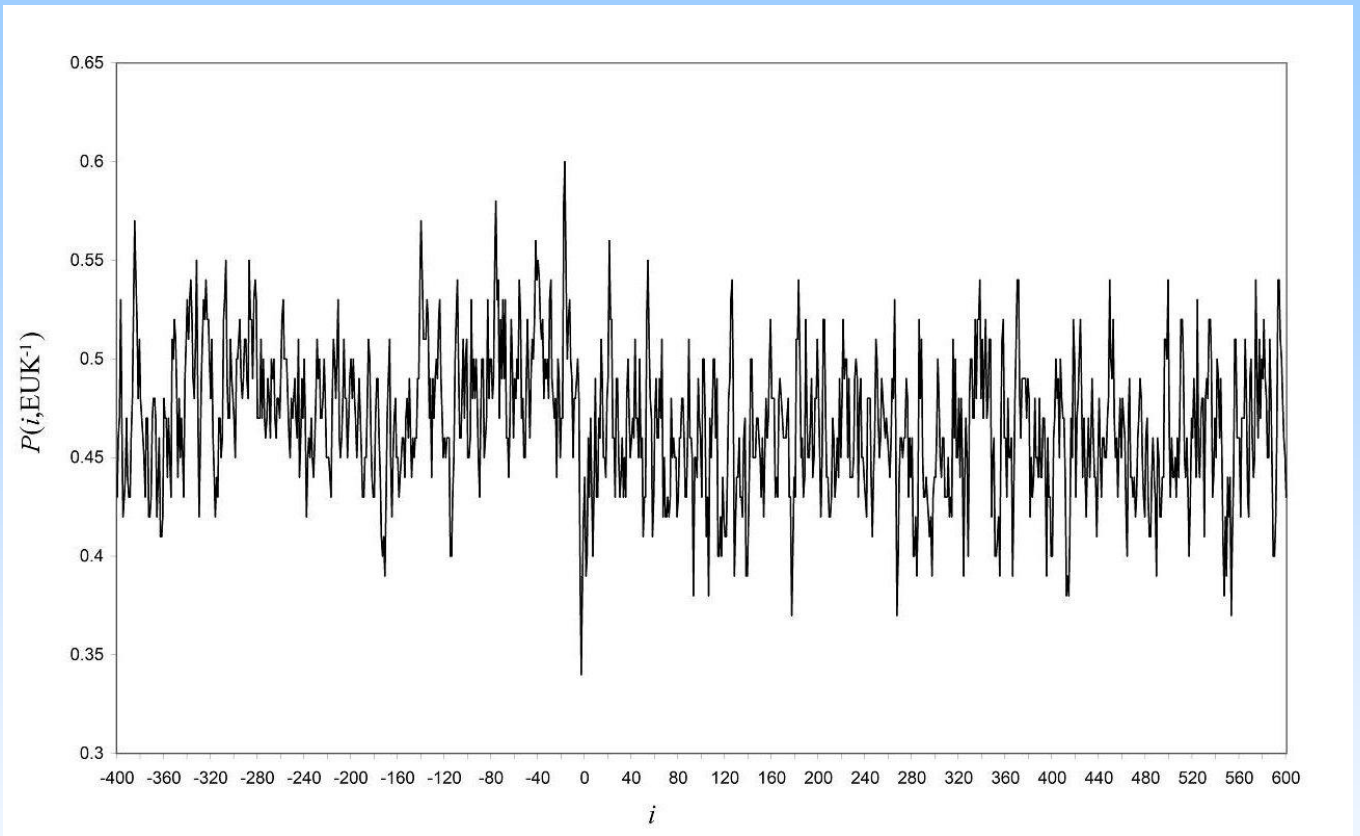
The score $P(i, s)$ of the C^3 code X in a window w_i in the average frame of a sequence s is

$$P(i, s) = \frac{1}{3} \sum_{\mathfrak{f}=0}^{2} P(i, s_{\mathfrak{f}}).$$

The score $P(i, F)$ of the C^3 code X in a window w_i in the average frame of a population F is

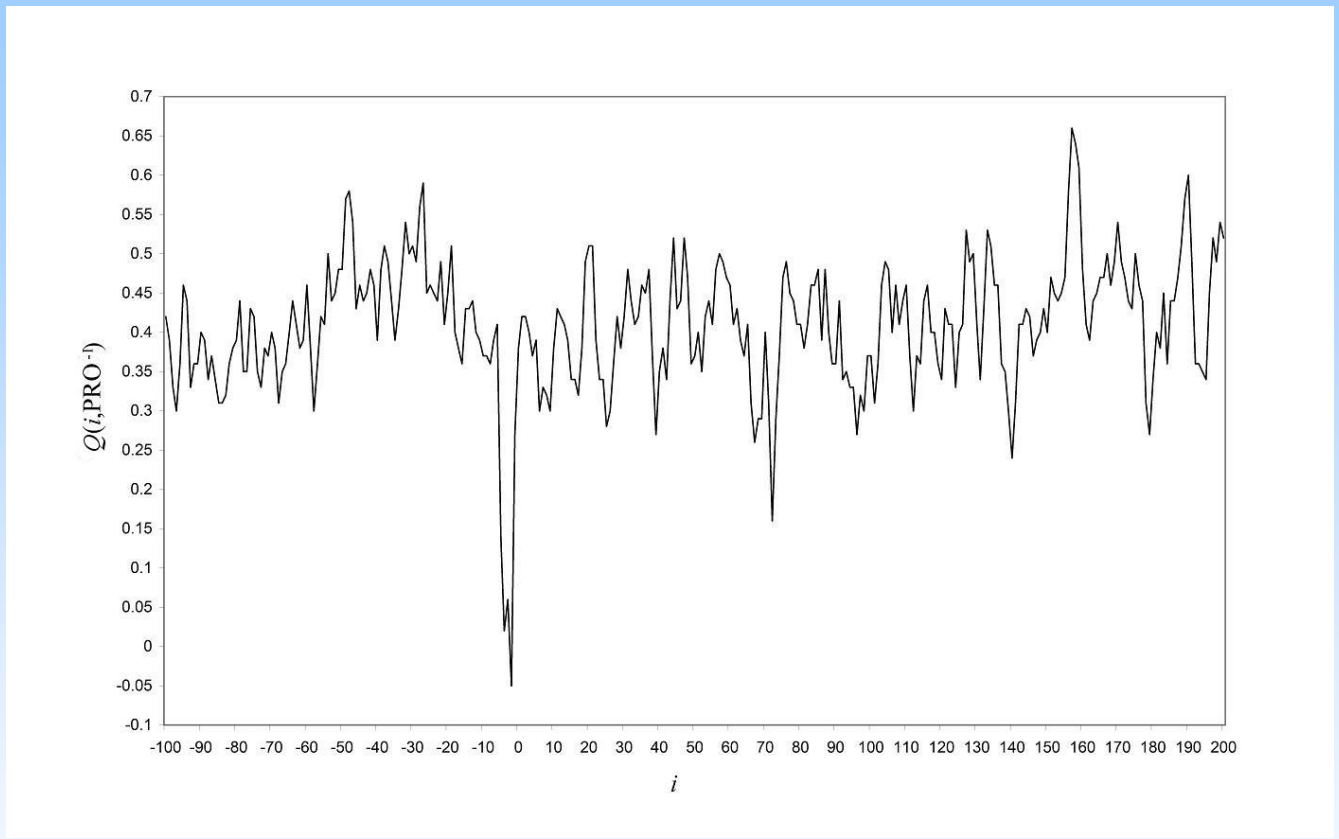
$$P(i, F) = \frac{1}{n(F)} \sum_{s \in F} P(i, s).$$

Frameshift genes: results



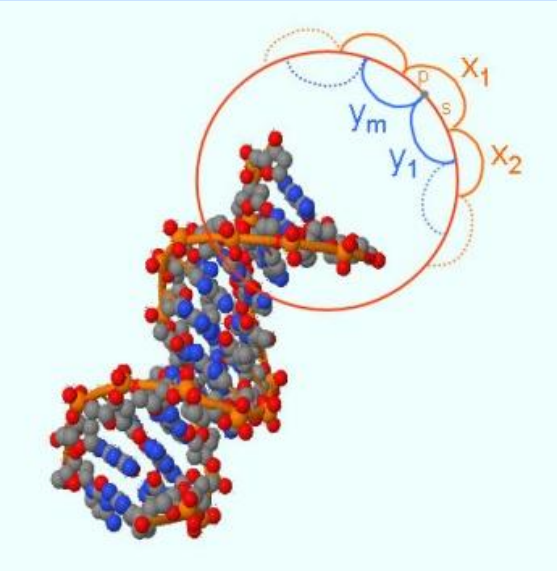
Significant lowest signal around the frameshift site i=0 in the -1 frameshift eukaryotic genes (27 genes, RECODE database) associated with the absence of circular code

Frameshift genes: results



Significant lowest signal around the frameshift site i=0 in the -1 frameshift prokaryotic genes (15 genes, RECODE database) associated with the absence of circular code

THEORETICAL PROPERTIES OF CIRCULAR CODES IN GENES



THEORETICAL PROPERTIES OF CIRCULAR CODES

$l_1, l_2, \ldots, l_{n-1}, l_n, \ldots$ are letters in $\mathcal{A}_4$,

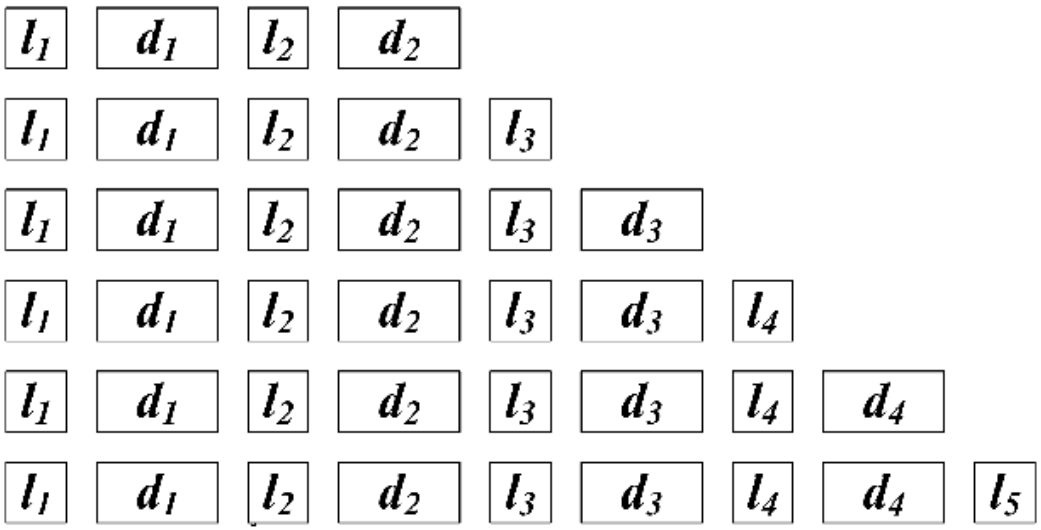
$d_1, d_2, \ldots, d_{n-1}, d_n, \ldots$ are diletters in $\mathcal{A}_4^2$

and n is an integer satisfying $n \geq 2$.

THEORETICAL PROPERTIES OF CIRCULAR CODES

Letter Diletter Necklaces (LDN): We say that the ordered sequence $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n$ is an $nLDN$ for a subset $X \subset \mathcal{A}_4^3$ if $l_1d_1, l_2d_2, \ldots, l_nd_n \in X$ and $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n \in X$.

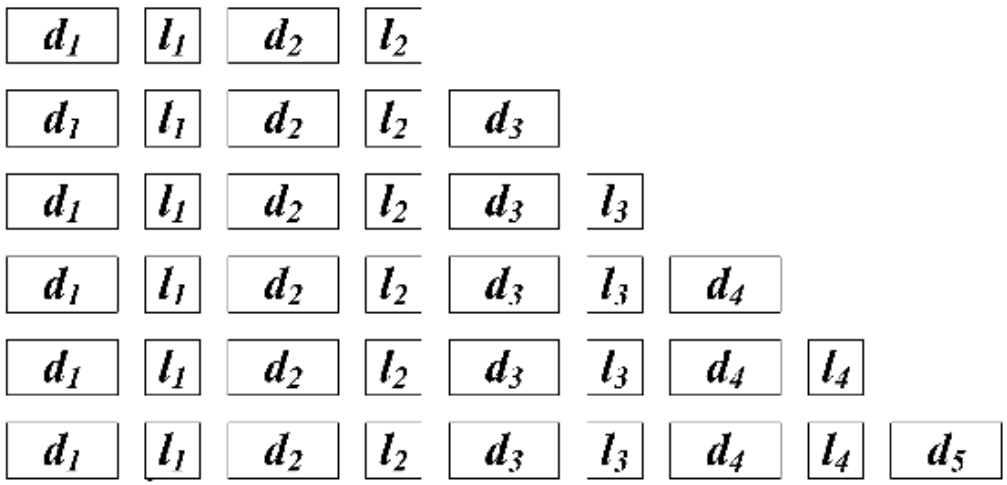
Letter Diletter Continued Necklaces $(LDCN)$: We say that the ordered sequence $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n, l_{n+1}$ is an $(n+1)LDCN$ for a subset $X \subset \mathcal{A}_4^3$ if $l_1d_1, l_2d_2, \ldots, l_nd_n \in X$ and $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n, d_nl_{n+1} \in X$.



THEORETICAL PROPERTIES OF CIRCULAR CODES

Diletter Letter Necklaces (DLN): We say that the ordered sequence $d_1, l_1, d_2, l_2, \ldots l_{n-1}, d_n, l_n$ is an $nDLN$ for a subset $X \subset \mathcal{A}_4^3$ if $d_1l_1, d_2l_2, \ldots, d_nl_n \in X$ and $l_1d_2, l_2d_3, \ldots, l_{n-1}d_n \in X$.

Diletter Letter Continued Necklaces $(DLCN)$: We say that the ordered sequence $d_1, l_1, d_2, l_2, \ldots l_{n-1}, d_n, l_n, d_{n+1}$ is an $(n+1)DLCN$ for a subset $X \subset \mathcal{A}_4^3$ if $d_1l_1, d_2l_2, \ldots, d_nl_n \in X$ and $l_1d_2, l_2d_3, \ldots, l_{n-1}d_n, l_nd_{n+1} \in X$.



THEORETICAL PROPERTIES OF CIRCULAR CODES

Proposition. Let X be a trinucleotide code. The following conditions are equivalent.

(i) X is circular code.

(ii) X has no $5LDCN$.

Proposition. Let X be a trinucleotide code. The following conditions are equivalent.

(i) X *is a comma-free code.*

(ii) X *has no* $2LDN$ *and no* $2DLN$.

THEORETICAL PROPERTIES OF CIRCULAR CODES

l	1	2	3	4	5	6	7	8	9	10
$\mathrm{Nb}(l)$	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

Growth function of comma free codes

THEORETICAL PROPERTIES OF CIRCULAR CODES

l	2	4	6	8	10	12	14	16	18	20
$\mathrm{Nb}(l)$	28	210	556	642	396	152	36	4	0	0

Growth function of self-complementary comma free codes

THEORETICAL PROPERTIES OF CIRCULAR CODES

l	1	2	3	4	5	6	7	8	9	10
$\mathrm{Nb}(l)$	0	0	0	0	0	0	0	0	96	1152

11	12	13	14	15	16	17	18	19	20
4224	6708	4632	8040	8568	10488	4848	3072	960	408

Growth function of maximal comma free codes

THEORETICAL PROPERTIES OF CIRCULAR CODES

Let X be a trinucleotide code.

For $k \in \{2, 3, 4, 5\}$, we say that
X belongs to the class C^{kLDN} if X has no $kLDN$
and that X belongs to the class C^{kDLN} if X has no $kDLN$.

Similarly, for $k \in \{3, 4, 5\}$, we say that
X belongs to the class C^{kLDCN} if X has no $kLDCN$
and that X belongs to the class C^{kDLCN} if X has no $kDLCN$.

THEORETICAL PROPERTIES OF CIRCULAR CODES

Proposition. *The following chains of inclusions hold.*

(i) $C^{2LDN} \subset C^{3LDCN} \subset C^{3LDN} \subset C^{4LDCN} \subset C^{4LDN} \subset C^{5LDCN} \subset C^{5LDN}$.

(ii) $C^{2DLN} \subset C^{3DLCN} \subset C^{3DLN} \subset C^{4DLCN} \subset C^{4DLN} \subset C^{5DLCN} \subset C^{5DLN}$.

(iii) $C^{2LDN} \subset C^{3DLCN} \subset C^{3LDN} \subset C^{4DLCN} \subset C^{4LDN} \subset C^{5DLCN} \subset C^{5LDN}$.

(iv) $C^{2DLN} \subset C^{3LDCN} \subset C^{3DLN} \subset C^{4LDCN} \subset C^{4DLN} \subset C^{5LDCN} \subset C^{5DLN}$.

Proposition. $C^{5LDN} = C^{5LDCN} = C^{5DLN}$.

Proposition. $C^{5DLCN} \subset C^{5LDCN}$ with $C^{5DLCN} \neq C^{5LDCN}$.

THEORETICAL PROPERTIES OF CIRCULAR CODES

C^{2LDN}	C^{3LDCN}	C^{3LDN}	C^{4LDCN}	C^{4LDN}	C^{5LDCN}	C^{5LDN}
0	56	56	56	$56+56$	$104+56+56$	$104+56+56$
C^{2DLN}	C^{3DLCN}	C^{3DLN}	C^{4DLCN}	C^{4DLN}	C^{5DLCN}	C^{5DLN}
0	0	56	$56+56$	$56+56$	$56+56$	$104+56+56$

Hierarchy of
the 216 maximal C³ self-complementary codes

The maximal C³ self-complementary circular code
X_0 in genes belongs to the class C^{5LDCN}

Conclusion

In genes, it exists

- genetic codes for coding the amino acids, the most important one is the universal genetic code

- **circular codes for retrieving the reading frames of genes**

It is still not known to date which biological apparatus could have used these circular codes

Selected personal references
a review

Michel C.J. (2008). A 2006 review of circular codes in genes. **Computer and Mathematics with Applications** 55, 984-988.

Selected personal references in the research field of code identification

Frey G., Michel C.J. (2006). Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. **Journal of Computational Biology and Chemistry** 30, 87-101.

Frey G., Michel C.J. (2003). Circular codes in archaeal genomes. **Journal of Theoretical Biology** 223, 413-431.

Arquès D.G., Michel C.J. (1997). A circular code in the protein coding genes of mitochondria. **Journal of Theoretical Biology** 189, 273-290.

Arquès D.G., Michel C.J. (1996). A complementary circular code in the protein coding genes. **Journal of Theoretical Biology** 182, 45-58.

Selected personal references in the research field of code biological function

Ahmed A., Frey G., Michel C.J. (2007). Frameshift signals in genes associated with the circular code. **In Silico Biology** 7, 151-154.

Frey G., Michel C.J. (2006). Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. **Journal of Computational Biology and Chemistry** 30, 87-101.

Selected personal references in the research field of theoretical code

Michel C.J., Pirillo G, Pirillo, M.A. (2008). A relation between trinucleotide comma-free codes and trinucleotide circular codes. **Theoretical Computer Science**, sous presse.

Michel C.J., Pirillo G, Pirillo, M.A. (2008). Varieties of comma free codes. **Computer and Mathematics with Applications** 55, 989-996.

Lacan J., Michel C.J. (2001). Analysis of a circular code model. **Journal of Theoretical Biology** 213, 159-170.