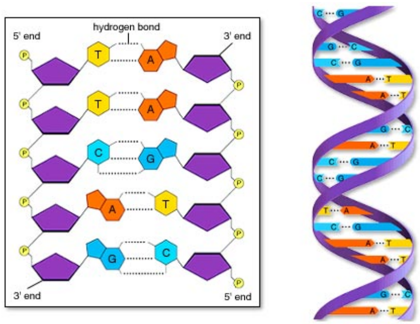
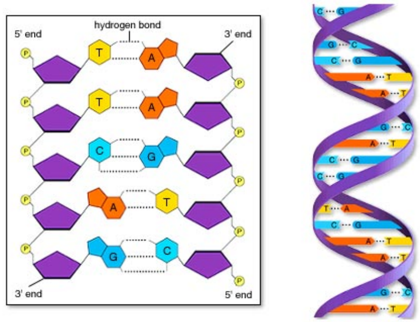
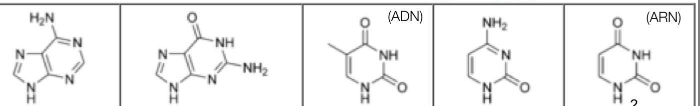
1. L'ADN et l'information génétique



l'ADN

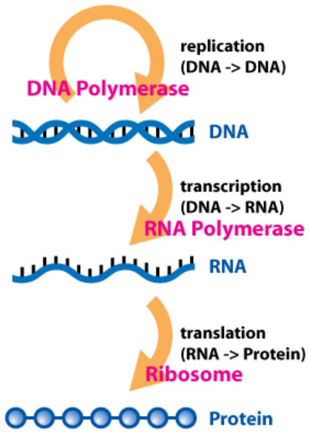
l'information génétique est contenue dans l'ADN





Adenine A	Guanine G	Thymine T	Cytosine C	Uracil U

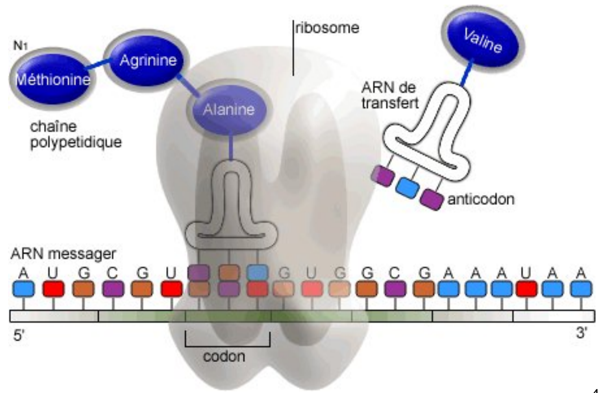
comment fait-on une protéine ?



le « rôle » de l'ADN est de contenir, conserver et transmettre l'information nécessaire à la synthèse des protéines.

traduction

l'information génétique est organisée par triplets (codons)



le code génétique

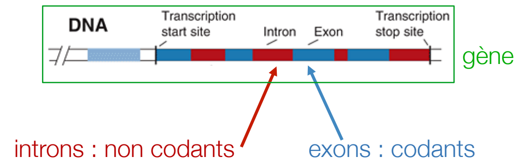
Second Letter

1st letter	U	C	A	G	3rd letter
U	UUU, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr UAA: Stop UAG: Stop	UGU, UGC: Cys UGA: Stop UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

le code génétique est dégénéré : $4^3 = 64 > 20$!

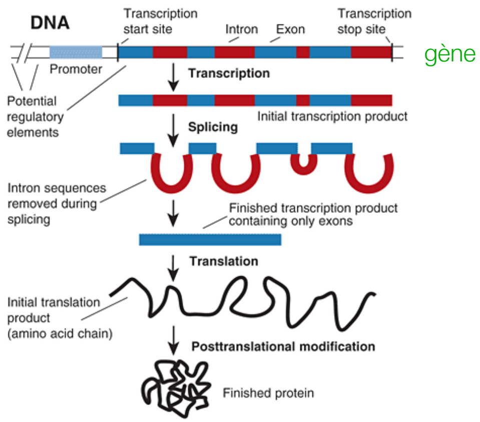
le gène

unité de l'information génétique



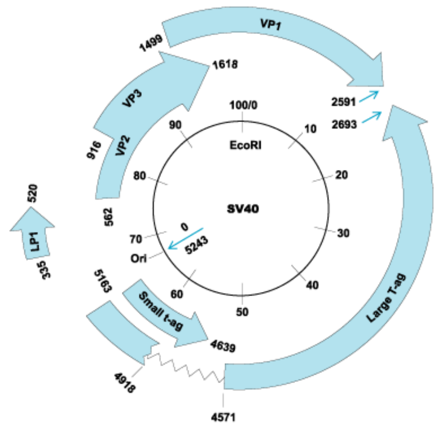
le gène

unité de l'information génétique



le génome

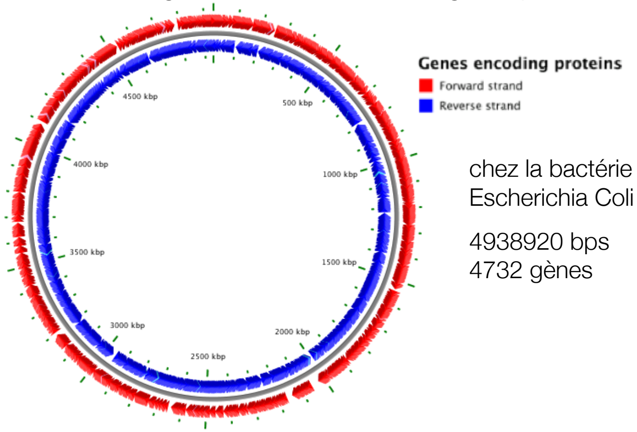
organisation de l'information génétique



Simian Virus 40

le génome

organisation de l'information génétique



chez la bactérie :
Escherichia Coli

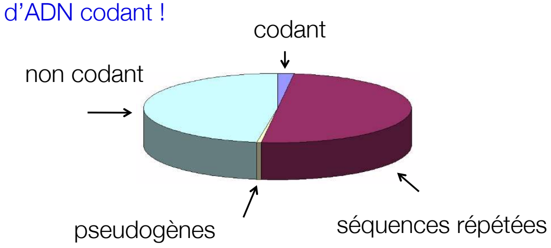
4938920 bps
4732 gènes

le génome

organisation de l'information génétique

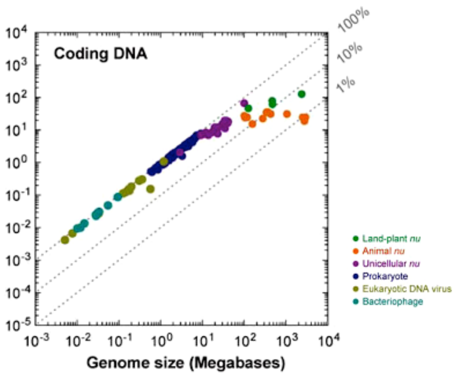
chez l'homme :

- •3 milliards de pbs
- •~20000 gènes
- •< 2 % d'ADN codant !



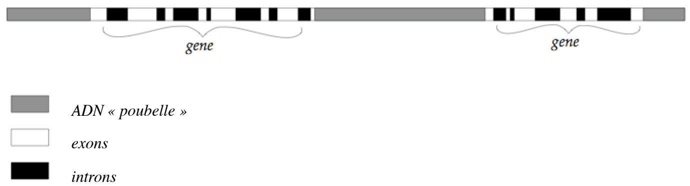
le génome

quantité d'ADN codant en fonction de la taille du génome



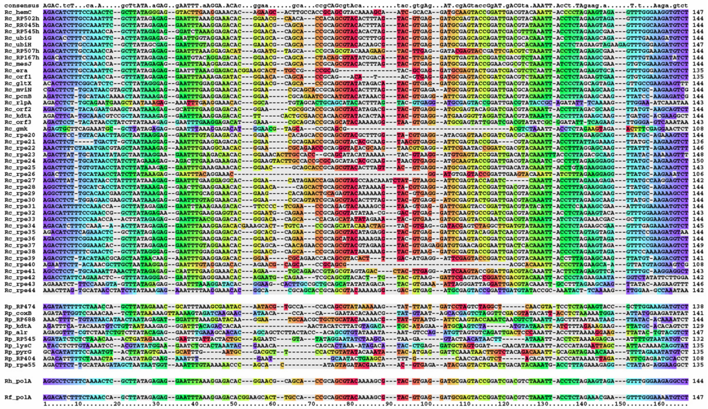
ADN codant et non codant

en résumé :



pourquoi autant d'ADN non codant ?
peut-on en comprendre le rôle ?

2. étudier les séquences d'ADN



analyse des séquences

approche « déterministe » :

comparaison entre séquences / alignement de séquences :

recherche de gènes, comparaison d'espèces, de motifs répétés

approche statistique :

recherche de caractéristiques « globales » :

rechercher des régularités, un « ordre » dans le désordre apparent…

1. information mutuelle
2. fonction de corrélation / densité spectrale de puissance
3. techniques basées sur la « marche ADN »
4. …

de la séquence symbolique à une séquence numérique

choisir un code binaire : par exemple

double liaison hydrogène: A, T → +1

triple liaison hydrogène: G, C → -1

ou

purines: A, G → +1

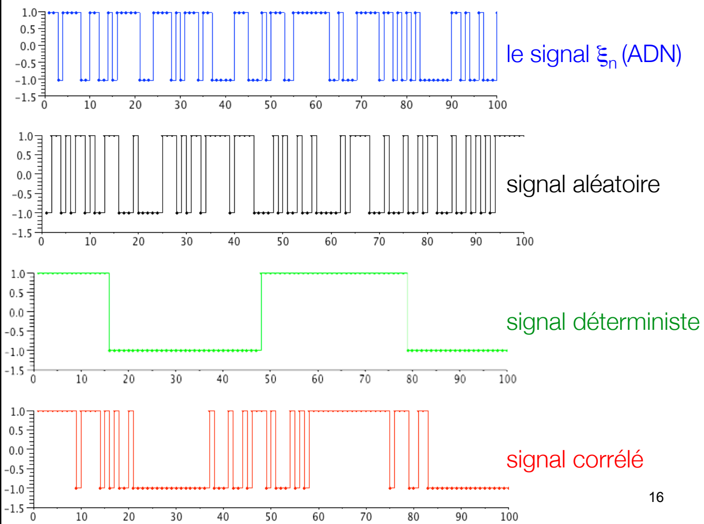
pyrimidines: T, C → -1

on peut donc étudier la variable numérique binaire

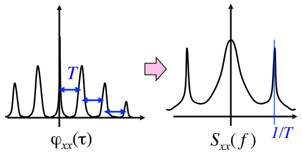
$\boxed{\xi_n = \pm 1 \quad \text{où } n = \text{position}}$

on obtient :

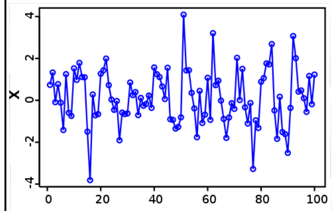
n=	1	2	3	4	5	6	7	8	9	10…
	A	T	C	G	G	T	C	A	T	A…
ξ_n=	+1	+1	-1	-1	-1	+1	-1	+1	+1	+1…



3. fonction de corrélation et densité spectrale de puissance (DSP)

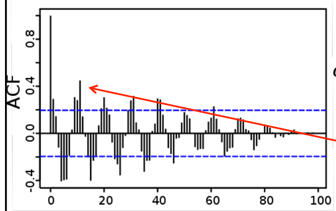


fonction de corrélation



ξ(t) signal aléatoire fonction du temps t, stationnaire :

ici sinus + bruit



$\varphi_{\xi\xi}(\tau)$ fonction de corrélation de ξ(t) :

$$\varphi_{\xi\xi}(\tau) = \left\langle \left(\xi(t) - \langle\xi\rangle\right) \cdot \left(\xi(t+\tau) - \langle\xi\rangle\right) \right\rangle_{ensemble}$$

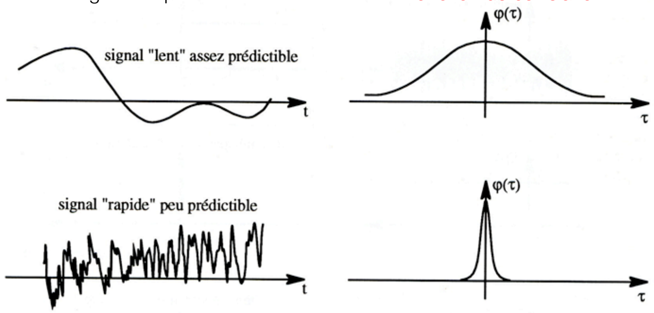
périodicité « cachée »
→ oscillations dans la corrélation

fonction de corrélation

mesure de la persistance (mémoire) d'un signal :

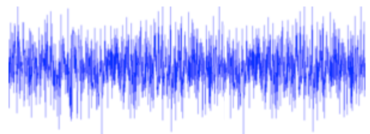
Signal temporel

fonction de corrélation



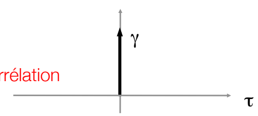
(bruit blanc)

Signal « sans mémoire » : chaque valeur est indépendante de la précédente



b(t)

- b(t) signal stationnaire,
- $\langle b(t) \rangle = 0$ (centré),
- $\varphi_{bb}(\tau) = \gamma\ \delta(\tau)$ fonction de corrélation



d'où $\sigma_b^2 = \varphi_{bb}(\tau) = +\infty$

moyennes temporelles

fonction de corrélation de ξ(t)

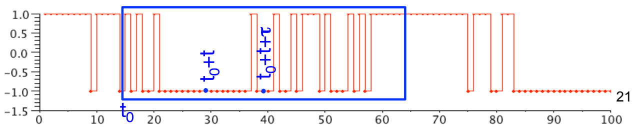
$$\varphi_{\xi\xi}(\tau) = \left\langle \left(\xi(t) - \langle\xi\rangle\right) \cdot \left(\xi(t+\tau) - \langle\xi\rangle\right) \right\rangle_{\boxed{ensemble}} \;?$$

si « ergodique » :

on peut remplacer la moyenne d'ensemble par une moyenne sur le temps :

$$\varphi_{\xi\xi}(\tau) = \left\langle \left(\xi(t_0 + t) - \langle\xi\rangle\right) \cdot \left(\xi(t_0 + t + \tau) - \langle\xi\rangle\right) \right\rangle_{t_0}$$

chaque t_0 initial considéré comme une nouvelle réalisation



densité spectrale de puissance (DSP)

Densité spectrale de puissance

$$S_{\xi\xi}(f) \overset{def}{=} TF\left[\varphi_{\xi\xi}(\tau)\right]$$

Théorème de Wiener-Khintchine :

$$S_{\xi\xi}(f) = \lim_{T\to\infty} \frac{1}{T} \left| TF\left[\xi^T(t)\right] \right|^2$$

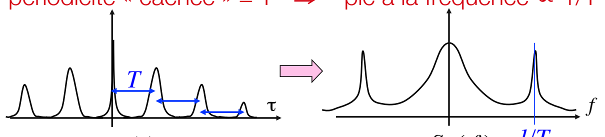
où $\xi^T(t) = \xi(t)$ limitée à l'intérvalle $[0, T]$

Estimateur de la DSP d'un signal réel :

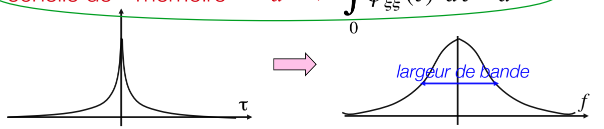
$$S_{\xi\xi}(f) \cong \left\langle \frac{1}{T} \left| TF\left[\xi^T(t)\right] \right|^2 \right\rangle_{ensemble}$$

Fonction de corrélation et DSP

1. périodicité « cachée » = T ⇒ pic à la fréquence ∝ 1/T



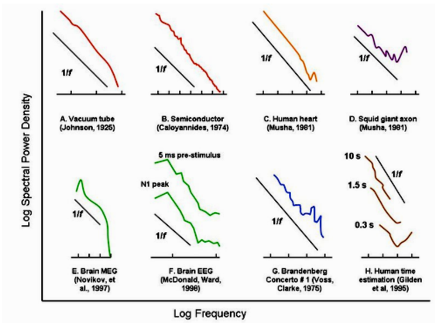
2. échelle de « mémoire » = a ⇒ $\int_0^\infty \varphi_{\xi\xi}(\tau)\, d\tau \approx a$



$\varphi_{\xi\xi}(\tau) \sim exp(-\tau / a)$
exponentielle

$S_{\xi\xi}(f) \sim 1/(1+(2\pi a f)^2)$
lorentzienne

4. corrélation à longue portée dans l'ADN



http://www.scholarpedia.org/article/1/f_noise

corrélation à longue portée

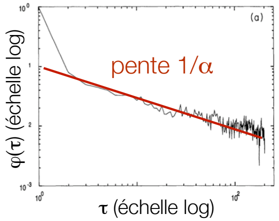
si l'échelle de mémoire est infinie ($\varphi_{xx}(\tau)$ n'est pas intégrable)

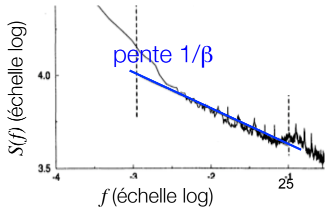
on parle de corrélation à longue portée.

Typiquement, loi de puissance :

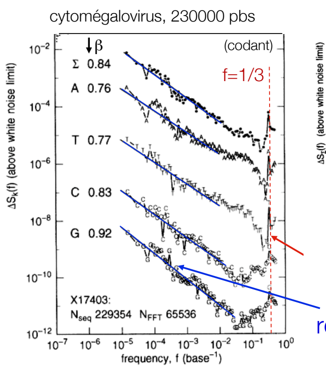
$$\lim_{\tau\to\infty} \varphi_{\xi\xi}(\tau) \propto \frac{1}{\tau^{\alpha}} \quad \Leftrightarrow \quad \lim_{f\to\infty} S_{\xi\xi}(f) \propto \frac{1}{f^{\beta}}$$

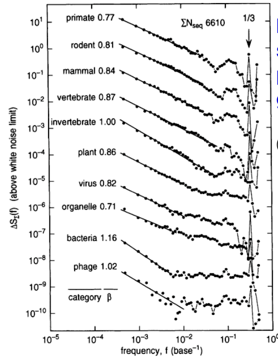
$$0 < \alpha \leq 1 \qquad 0 < \beta \leq 1 \qquad \beta = 1 - \alpha$$





dans l'ADN





le résultat semble plutôt général…

(non codant)

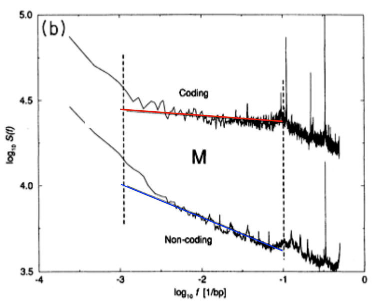
R. F. Voss, PRL, 1992

résultat 2 : DSP $\propto 1/f^{\beta}$ $\quad 0<\beta<1$

échelle log-log : log(DSP) $\propto -\beta$ log(f)

pour l'ADN codant

des résultats controversés :



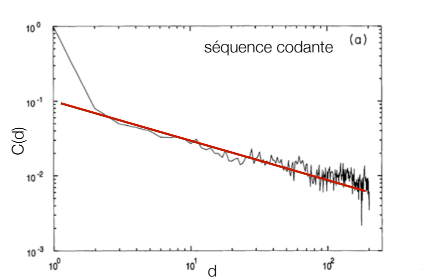
pas de corrélation ?

(Stanley group, 1992-1995)

L'ADN codant est sans doute moins corrélé que le non codant…

revenons à la fonction de corrélation

fonction de corrélation pour une séquence codante : périodicité 3



comment évolue l'amplitude des pics en position 3i ?

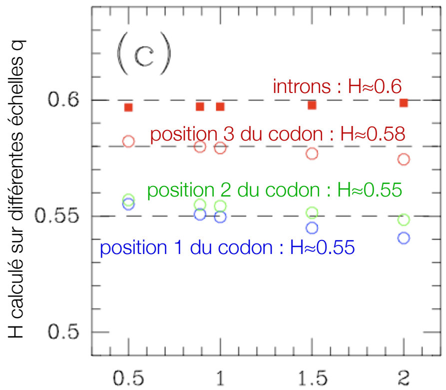
on la reporte en échelle log : décroissance en d^{α} !

corrélation longue portée pour « une base sur trois » ?

séquences codantes :

la dégénérescence du code laisse « passer »
un peu de corrélation longue portée

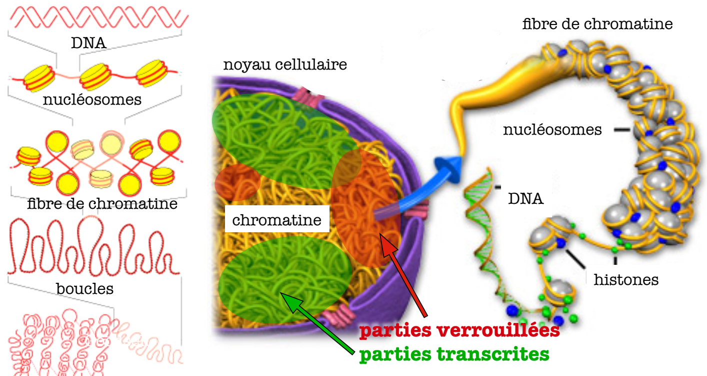
H = indice de corrélation longue portée



la position 3 du codon, plus libre, peut suivre la contrainte « globale »

pourquoi une mémoire étendue ?

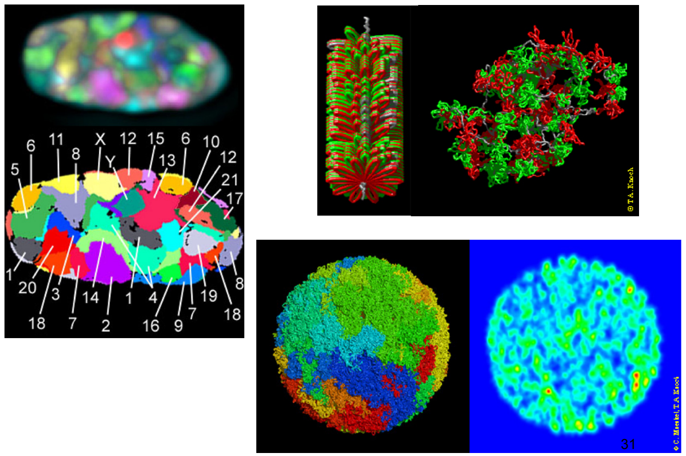
eucaryotes : CHROMATINE !



une structure fonctionnelle hautement organisée

30

d'autres images

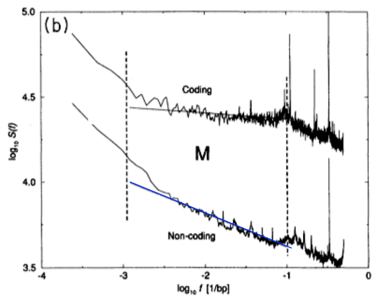


31

un rôle pour les séquences non codantes

la corrélation à longue portée indique
la présence d'un ordre global ;

les séquences non codantes
montrent toujours
une corrélation à longue portée



⇓

l'ADN « poubelle » participe à l'établissement d'un arrangement fonctionnel de l'ADN dans le noyau/la cellule !

et l'ADN codant ? il a aussi un autre rôle, participe peut-être en moindre mesure…

32

5. en pratique

Scilab

signal discret → fonction de corrélation de ξ_n

n = 1 2 3 4 5 6 7 8 9 10 11…

A T C G G T C A T A C…

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

ξ_n = +1 +1 -1 -1 -1 +1 -1 +1 +1 +1 -1…

si $\langle \xi_n \rangle = 0$ (autrement, on soustrait la moyenne)

$$C(d) = \langle \xi_n \ \xi_{n+d} \rangle = \frac{1}{N} \sum_{n=1}^{N} \xi_n \ \xi_{n+d}$$

d = distance entre 2 sites le long de la séquence

moyenne « temporelle » → moyenne sur n

(attention à la stationnarité…)

→ fonction de corrélation de ξ_n

en pratique :

sur un ordinateur, le signal est toujours discret !

$z(t) \rightarrow (z_1, \ z_2, \ z_3, \ z_4, \dots \ z_N) = z$

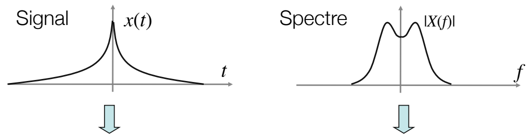
pour nous, le signal est *intrinsèquement* discret (pas=1), car c'est la séquence.

sous scilab, on peut utiliser la fonction :

corr : corr(z, d_{max}) = fonction de corrélation de x,
en fonction de la distance d,
pour d = 0, 1, 2,… d_{max}-1

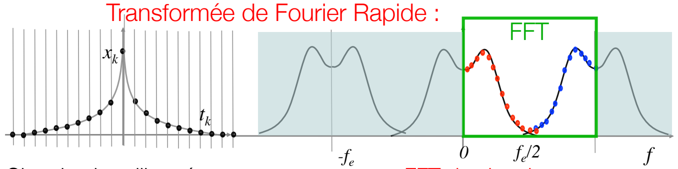
(remarque : corr soustrait automatiquement la moyenne de z)

→ DSP de ξ_n – 1. calculer la TF :



Signal échantillonné Δt → Spectre périodique de période $f_e = 1/\Delta t$

Transformée de Fourier Rapide :



Signal echantillonné :

$x_k = k \ \Delta t$
de 0 à $T_e = N \ \Delta T$

FFT du signal :

$f_n = n \ \Delta f$
de 0 à $f_e = N \ \Delta f$

→ DSP de ξ_n – 2. déduire la DSP :

FFT = transformée de Fourier Rapide

sous scilab :

fft : fft(z, -1) = fft(z) = TF de z

donne TF pour f = 0, Δf, 2Δf,… f_e-1

Wiener-Khintchine : estimateur DSP = $\left\langle \frac{1}{T} \left| TF\left[\xi^T(t)\right]\right|^2 \right\rangle_{ensemble}$

d'où $|fft(z)|^2 / N$ ≈ DSP de z

sous scilab, commencer par soustraire la moyenne de z

fin