

Master 1 MABS

UE Bioinformatique des séquences

novembre - décembre 2013



RECHERCHE DE SIMILITUDES ENTRE SÉQUENCES BIOLOGIQUES

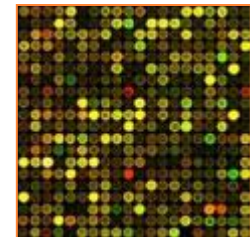
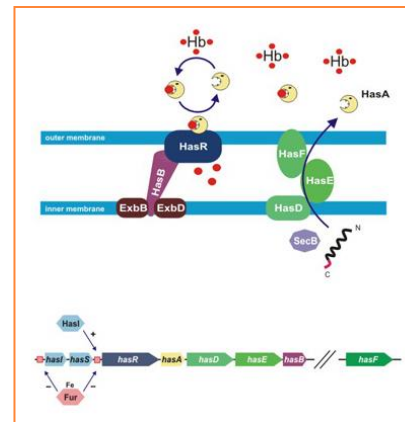
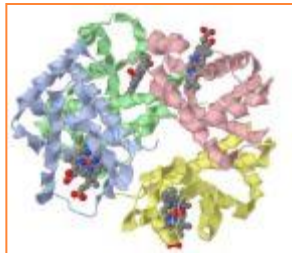
BIOINFORMATIQUE

- Définition:
 1. Applications de l'informatique à la biologie (en anglais: computational biology)
 2. Analyse de l'information biologique par l'informatique (en anglais: bioinformatics)
- L'information, c'est:

La séquence, la structure, la fonction, les interactions etc.

>anasp 1031 bp

```
GAATCTAGCTTCAGGTCGGGACACCACTGGAAACGGTGGCTAATACCG
GATGTGCCAAGTTATTGCCTGAGATGAGCTCGCGTCTGATTAGCTAGTT
GGTGTGGTAAGAGCGCACCAAGGCGACGATCAGTAGCTGGTCTGAGAGGA
TGATCAGCCACACTGGGACTGAGACACGGCCAGACTCCTACGGGAGGCA
GCAGTGGGGAATTTCCGCAATGGGCGAAAGCCTGACGGAGCAATACCGC
GTGAGGGGAAGGCTCTTGGCTTTTCTCAGGGAATAAAGGTCTTGAGGAA
TAAGCATCGGCTAACTCCGTGCCAGCAGCCGCGTAAATACGGAGGATGCA
AGCGTTATCCGGAATGATTGGGCGTAAAGCGTCCGACAGGTGGCACTGTAA
GTCTGCTGTTAAAGAGCAAGGCTCAACCTTGTAAGGCAGTGAAACTACA
GAGCTAGAGTACGTTCCGGGCAGAGGGAATTCCTGGTGTAGCGGTGAAAT
GCGTAGAGATCAGGAAGAACACCGTGGCGAAAGCGTCTGCTAGGCGCT
```



POUR QUOI FAIRE?

- La bioinformatique est d'abord utilisée pour identifier les gènes, étudier leur fonction et leur évolution.

GAATC**TAG**CTTCAGGTCGGGGACAACCACTGGAAACGGTGGC**TAAT**ACCG
G**ATG**TGCCAAAGTTATTGCC**TGA**G**ATGA**GCTCGCGTC**TGA**T**TAG**CTAGTT
GGTGTGG**TAAG**AGCGCACCAAGGCGACGATCAG**TAG**CTGGTCT**TGAG**AGGA
TGATCAGCCACACTGGGAC**TGA**GACACGGCCCAGACTCCTACGGGAGGCA
GCAGTGGGGAATTTTCCGCA**ATG**GGCGAAAGCC**TGA**CGGAGCAATACCGC
GT**TGA**GGGGGAAGGCTCTTGGCTTTTCTCAGGGA**TAA**AGGTCC**TGA**GGAA
TAAGCATCGGC**TAA**CTCCGTGCCAGCAGCCGCGG**TAA**TACGGAGG**ATG**CA
AGCGTTATCCGGA**ATGA**TTGGGCG**TAA**AGCGTCCGCAGGTGGCACTG**TAA**
GTCTGCTGT**TAA**AGAGCAAGGCTCAACCTTG**TAA**AGGCAG**TGA**AACTACA
GAGC**TAG**AGTACGTTTCGGGGCAGAGGGAATTCCTGGTG**TAG**CGGT**TGA**AAT
GCG**TAG**AGATCAGGAAGAACACCGGTGGCGAAAGCGCTCTGC**TAG**GCCGT
AAC**TGA**CAC**TGA**GGGACGAAAGC**TAG**GGGAGCGA**ATG**GGAT**TAG**ATACCC
CAG**TAG**TCC**TAG**CCG**TAA**ACG**ATG**GATAC**TAG**GCGTGTTATCGACGTGC
CGGAGCCAACGCGT**TAA**GTATCCCGCCTGGGGAGTACGCACGCAAGTGTG
AAACTCAAAGGAAT**TGA**CGGGGGCCCGCACAAAGCGGTGGAGT**ATG**TGGTT
TAATTCG**ATG**CAACGCGAAGAACCTTACCAAGACT**TGA**C**ATG**TCTCT**TGA**A
GGAGAGTTGGACGCACAGGTGGTGC**ATG**GCTGTCTCGTCAGCTCGTGTCTGTG
AG**ATG**TTGGGT**TAA**GTCCCGCAACGAGCGCAACCCTAGTTGCCAGTGGGC
AGACTGCCGGTGCAAACCGGAGGAAGGTGGGG**ATGA**CGTCAAGTCAGCAT
GCCCCTTACGTCTTGGGCTACACACGTACTACA**ATG**CTACGGACAGAGGG
AAATCCGTAACCGGCTCAGTTCAGATCGCAG

- Codon d'initiation
- Stop
- Phase de lecture
- Similarités
- ...



POUR QUOI FAIRE?

- "Fonction" peut être entendu dans un sens général
 - ATPase, Arn-polymérase, kératine, etc.
- ou dans un sens beaucoup plus précis, avec
 - identification des résidus essentiels,
 - éléments structuraux,
 - sites de fixation aux ligands,
 - site catalytiques,
 - etc.
- Par exemple...
 - Pour rechercher chez un organisme modèle un gène homologue à un gène humain d'intérêt
 - Pour rechercher des gènes liés à la pathogénicité
 - Pour concevoir une expérience de mutagenèse dirigée sur une protéine
 - Pour trouver tous les gènes présents sur un chromosome/génome/contig nouvellement séquencé
 - la liste des questions est illimitée...



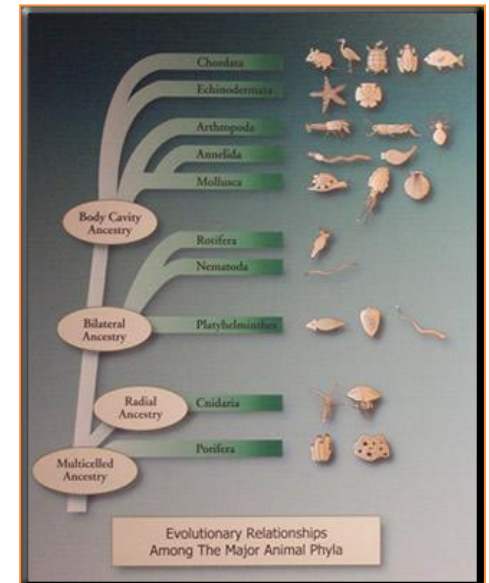
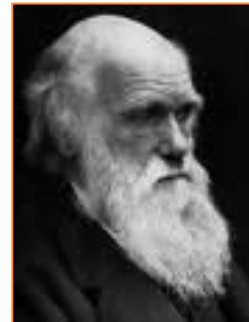
LA THÉORIE DE L'ÉVOLUTION

Les espèces se modifient au cours du temps et donnent naissance à de nouvelles espèces.

Selon Jean-Baptiste Lamarck (1744-1829), les espèces évoluent en adoptant des caractères acquis par les individus au cours de leur vie.



Charles Darwin (1809-1882) émet l'hypothèse de la sélection du plus apte (ou sélection naturelle) parmi des individus naturellement variant.



LA DÉDUCTION PAR HOMOLOGIE, OU LE «DOGME CENTRAL» DE LA BIOINFORMATIQUE

Si la bioinformatique «marche», c'est parce que l'évolution des gènes laisse une trace parfaitement visible lorsque l'on compare leur séquence

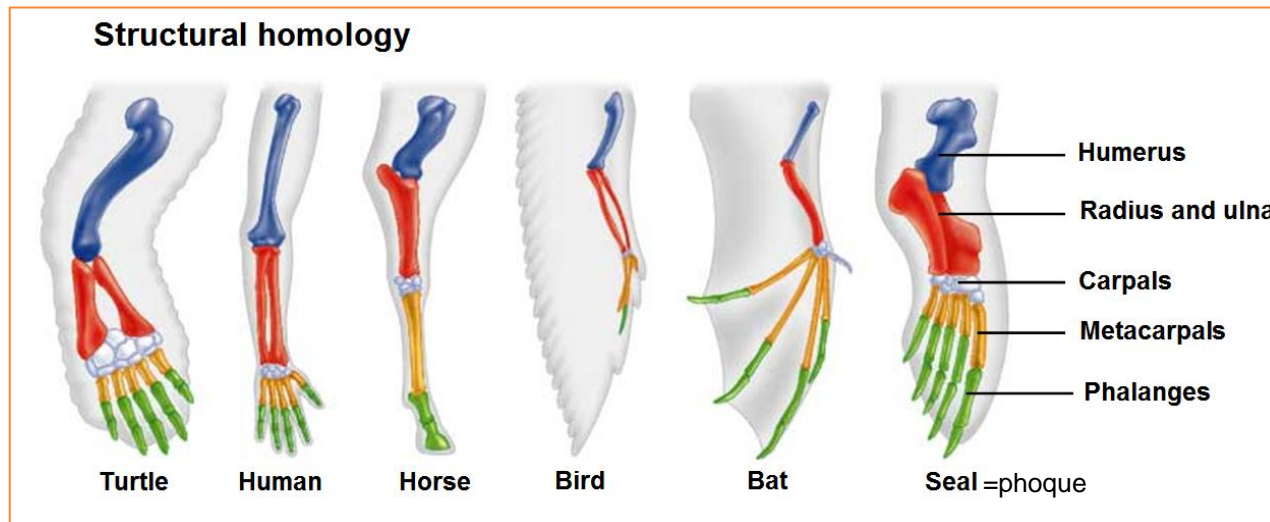
- Évolution des gènes = mutations, insertions, délétion.
- Les gènes des organismes modernes sont issus de remaniement de gènes ancestraux: on peut donc déduire la fonction de la plupart des gènes par comparaison avec les gènes «homologues» d'autres espèces.
(homologue = qui a un ancêtre commun)
- Les régions fonctionnelles des gènes (sites catalytique, de fixation, etc.) sont soumises à sélection. Elles sont relativement préservées par l'évolution car des mutations trop radicales sont désavantageuses.
- Les régions non fonctionnelles subissent peu de pressions de sélection et divergent rapidement au fur et à mesure que s'accumulent les mutations.



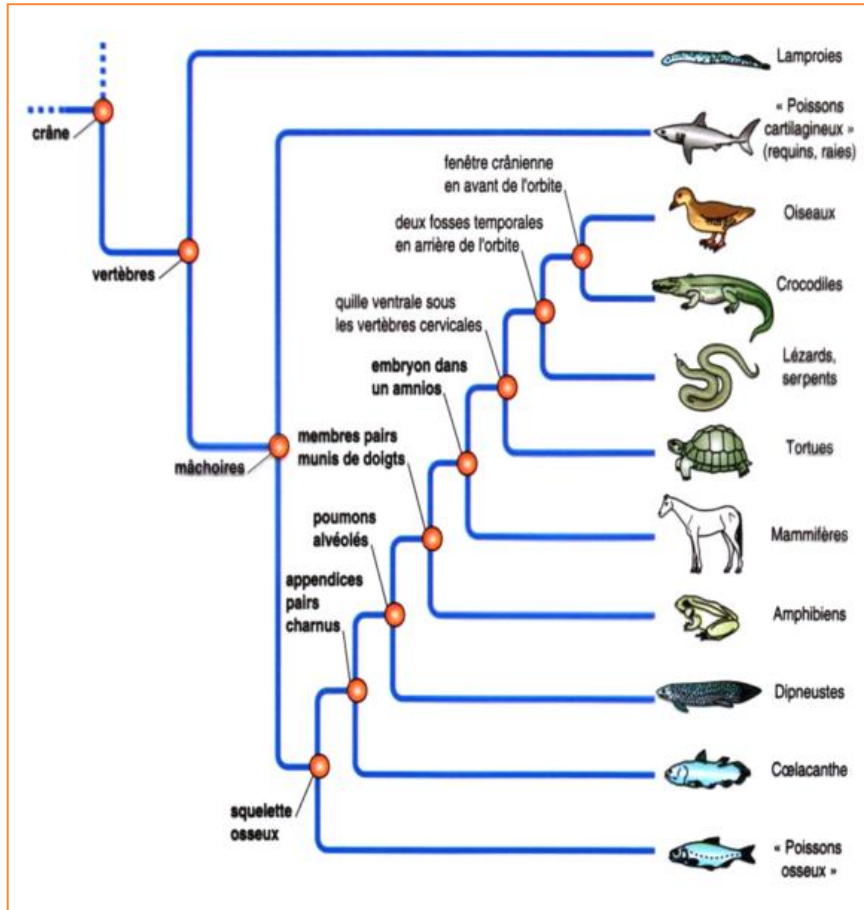
L'HOMOLOGIE DE SÉQUENCE

○ En bioinformatique: **Homologie = parenté = ancêtre commun**

- L'aile de l'aigle **est homologue** à l'aile du perroquet (oiseaux) et à la patte de la tortue (reptile)
- L'aile de la chauve souris **est homologue** à la patte du cheval et au bras de l'homme (chiroptère, ongulé et primate - mammifères).



L'HOMOLOGIE DE SÉQUENCE



○ L'homologie dépend de l'échelle d'observation

- Le groupe des tétrapodes:
Le bras de l'homme, la patte du cheval, l'aile de l'oiseau et l'aile de la chauve-souris sont tous des membres inférieurs des tétrapodes

vertébré terrestre ou marin dont le squelette comporte deux paires de membres munis de doigts, apparents ou atrophiés, témoignant dans l'évolution d'une adaptation primitive à la marche, tels que les amphibiens, les reptiles (dont les serpents), les mammifères et les oiseaux.

- Le groupe des Amniotes:
Séparation des oiseaux et des chauves-souris au niveau Amniotes présence d'un amnios chez certains vertébrés au stades embryonnaires

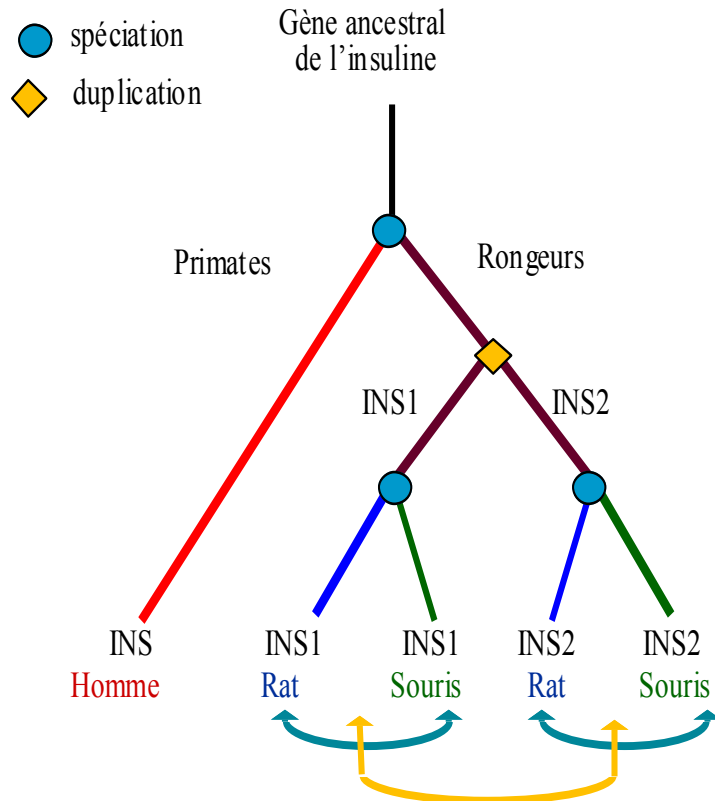


L'HOMOLOGIE DE SÉQUENCE

- On est homologue ou on ne l'est pas.
- Donc on ne dit pas: "~~très homologue~~", "~~faible homologie~~", "~~«22% d'homologie»~~", etc.
- Pour une notion quantitative, on parle de **similitude** ("très similaire", etc.) ou **d'identité** (28% d'identité)



ORTHOLOGIE ET PARALOGIE



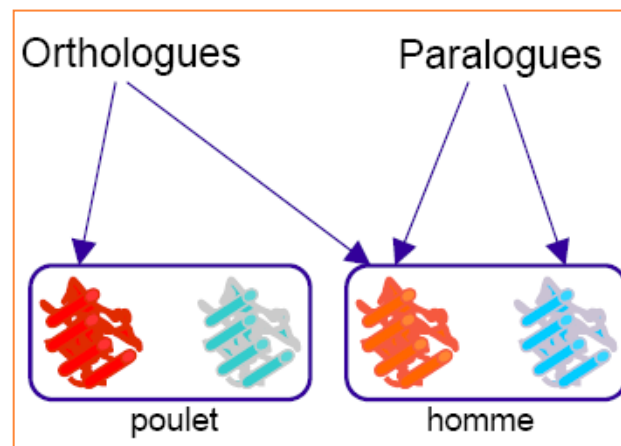
Homologie : deux gènes sont homologues si ils ont un ancêtre commun

↔ *Orthologie* : deux gènes sont orthologues si ils ont divergé à la suite d'un évènement de spéciation

↔ *Paralogie* : deux gènes sont paralogues si ils ont divergé à la suite d'un évènement de duplication

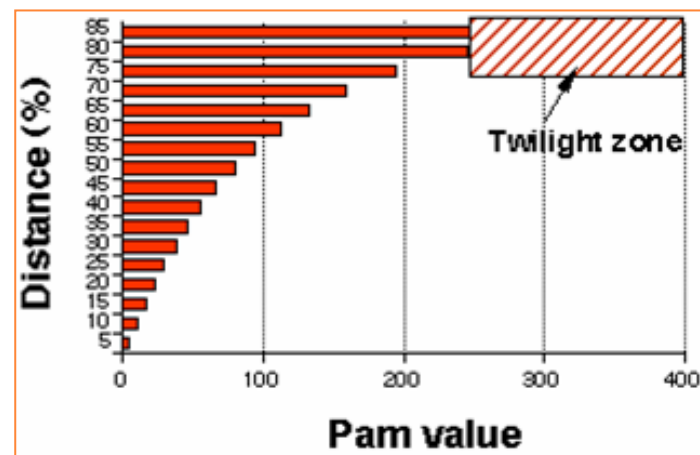
FONCTION ET HOMOLOGIE

- Homologie n'implique pas même fonction: par exemple l'aile de l'oiseau et le bras humain n'ont pas la même fonction
- Des orthologues rapprochés (p. ex. homme/souris) ont le plus souvent la même fonction dans l'organisme.
- Des orthologues distants (p. ex. homme/mouche) ont plus rarement le même rôle phénotypique, mais peuvent exercer le même rôle dans une voie donnée.
- Les paralogues acquièrent rapidement des fonctions différentes



A QUEL POINT DES SÉQUENCES HOMOLOGUES SE RESSEMBLENT-ELLES?

- De 100% à quelques nucléotides/acides aminés en commun.
- Il n'y a pas vraiment de limite, mais en dessous d'un certain niveau d'identité (twilight zone = zone nébuleuse), il devient difficile de distinguer une homologie d'une ressemblance fortuite. Deux séquences d'ADN prises au hasard ont 25% de nucléotides communs.
- Des séquences sans ressemblance apparente peuvent parfaitement être homologues (on le retrouve par ex. au niveau 3D)
- Par contre, étant donné la dimension de l'espace des séquences possibles, une ressemblance importante est généralement interprétée comme une homologie, et non pas comme une évolution convergente.



COMMENT DÉTECTER UNE HOMOLOGIE?

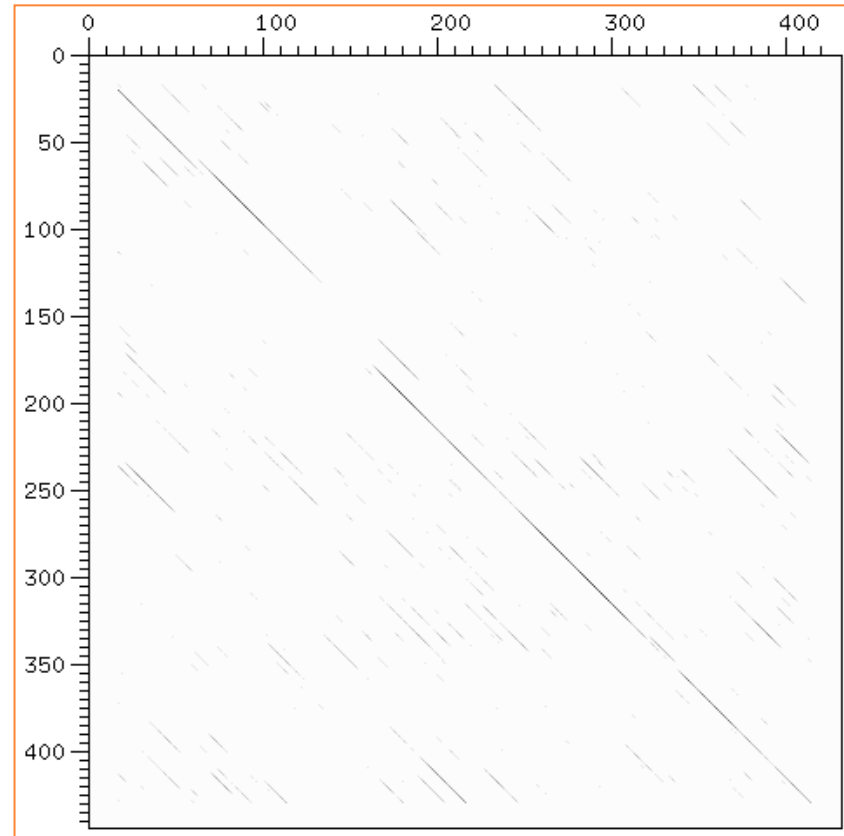
Principe: comparaison de séquences

- L'alignement des séquences est la principale méthode de comparaison. Elle permet d'identifier des régions conservées. On en déduit l'homologie.
- D'autres méthodes existent:
 - Analyse statistique des «mots» contenus dans la séquence
 - Recherche de domaines ou motifs communs
- Comparer des séquences serait relativement simple si elles avaient toutes la même longueur. Comme ce n'est pas le cas, il faut les aligner, c'est à dire trouver où se trouvent les insertions et délétions, représentées par des «indels» («gaps»)



LES «DOT PLOTS»

- Deux séquences à comparer sont représentées (ici 2 gènes de globine), une horizontalement, l'autre verticalement. On dessine ensuite un point dans la matrice lorsque les deux positions correspondantes sont identiques. Lorsque des régions se ressemblent, on voit apparaître une diagonale. Les décalages entre les diagonales correspondent à des insertions ou délétions. Plusieurs diagonales parallèles indiquent une répétition.
- Pour "nettoyer" le dot plot, on utilise souvent non pas un point par base, mais un point lorsque n bases sont identiques, ou n bases identiques dans une fenêtre de N . Cela réduit considérablement le nombre de points.
- Les dot plots sur des génomes complets permettent de visualiser les événements à grande échelle, la synténie, etc.

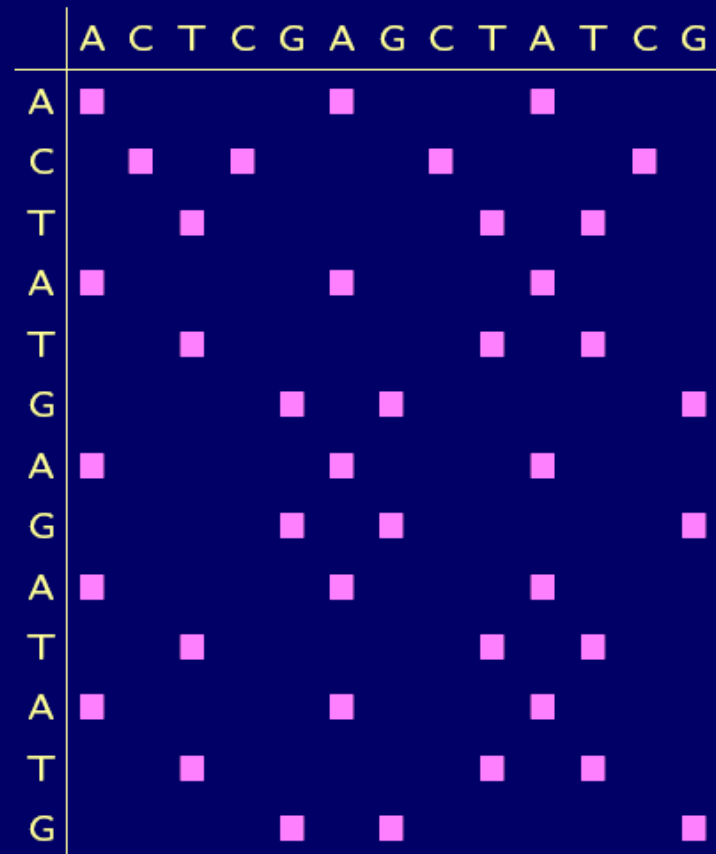


LES «DOT PLOTS»

match (identité) → ■

mismatch → □

diagonale = région similaire

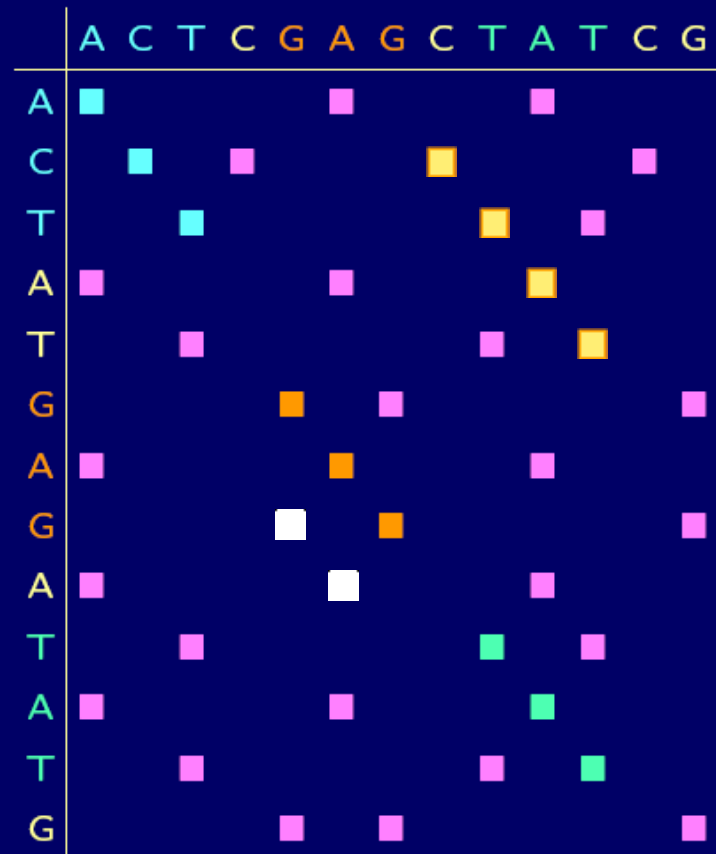


LES «DOT PLOTS»

match (identité) → ■

mismatch → □

diagonale = région similaire

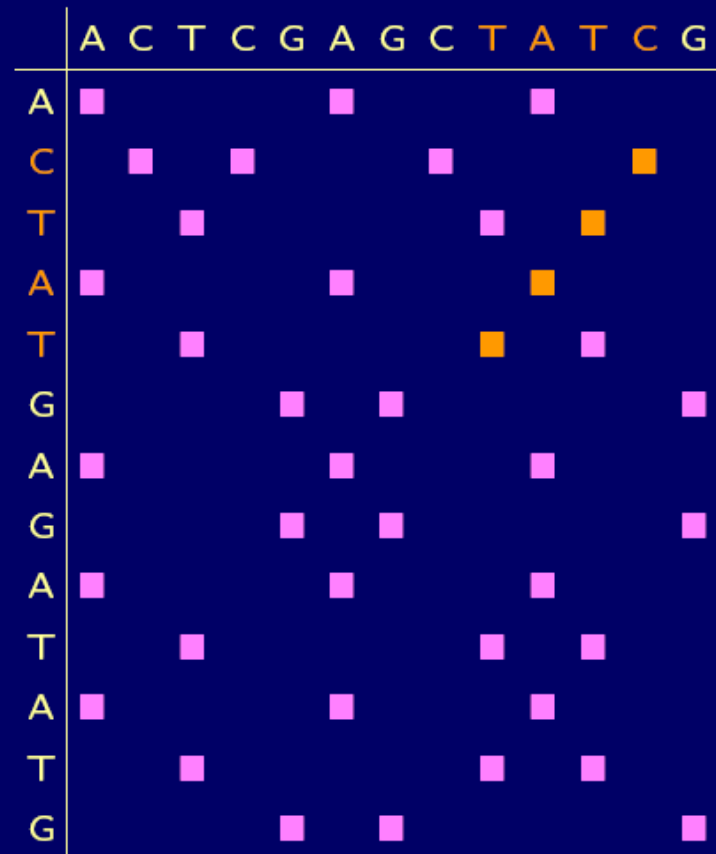


LES «DOT PLOTS»


match (identité) → ■

mismatch → □

diagonale inversée = région miroir

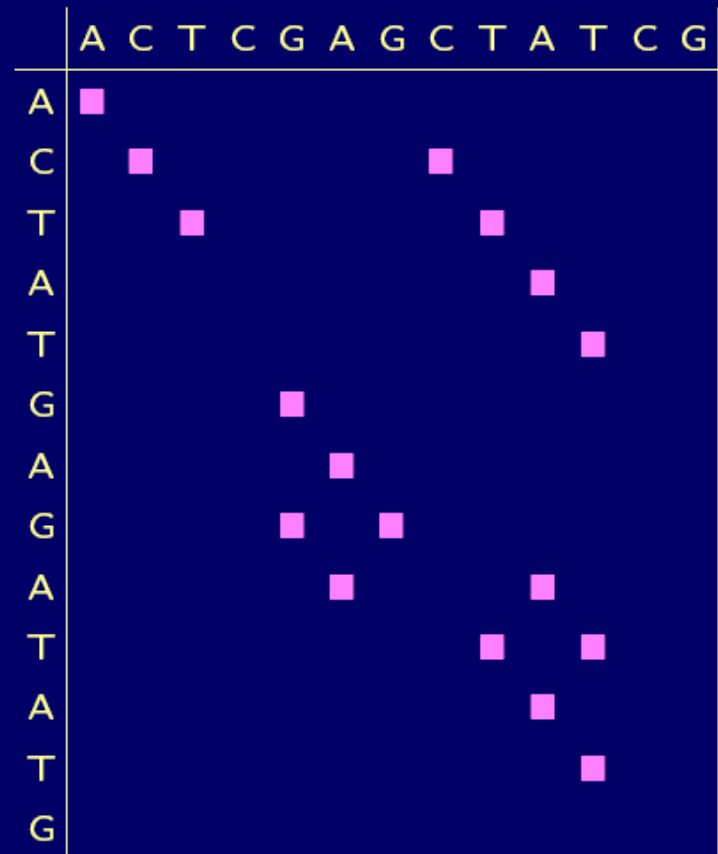


LES «DOT PLOTS»

match (identité) → 

mismatch \rightarrow \square

Fenêtre de taille 2

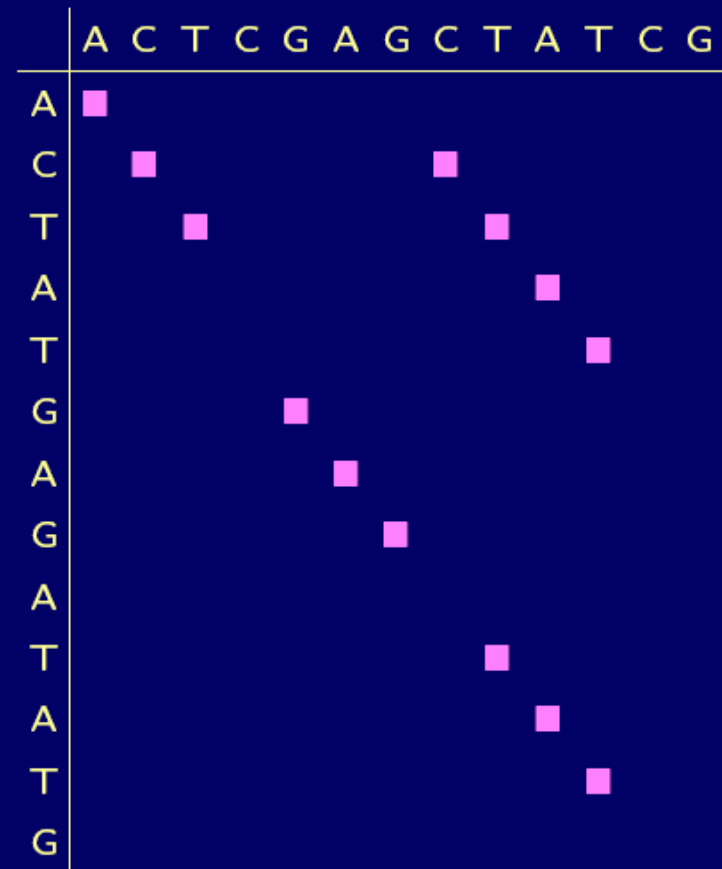


LES «DOT PLOTS»

match (identité) → 

mismatch \rightarrow ☐

Fenêtre de taille 3

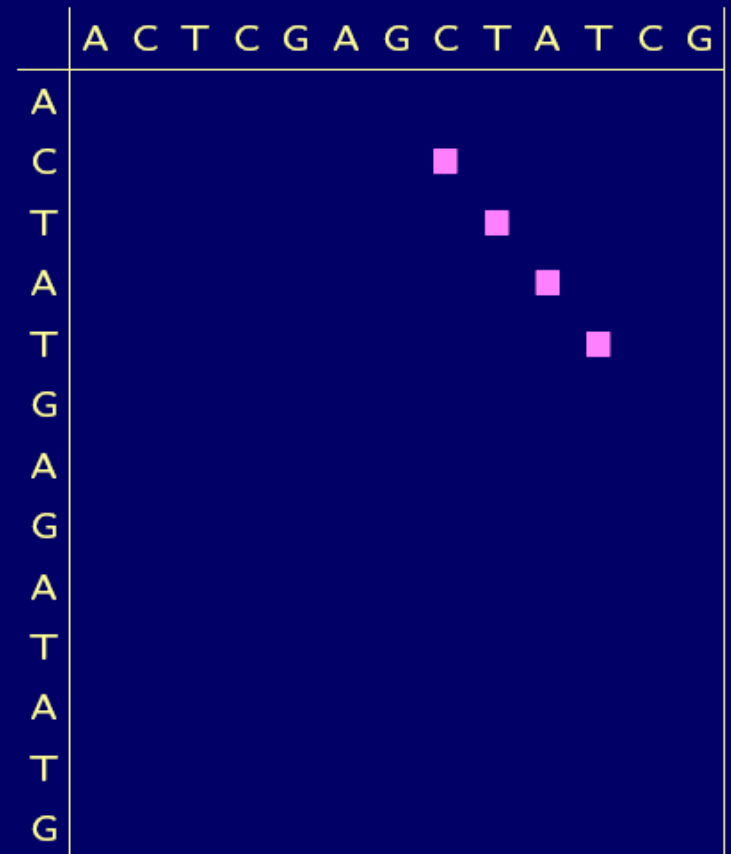


LES «DOT PLOTS»

match (identité) → 

mismatch \rightarrow \square

Fenêtre de taille 4



LES «DOT PLOTS»

match (identité) → ■

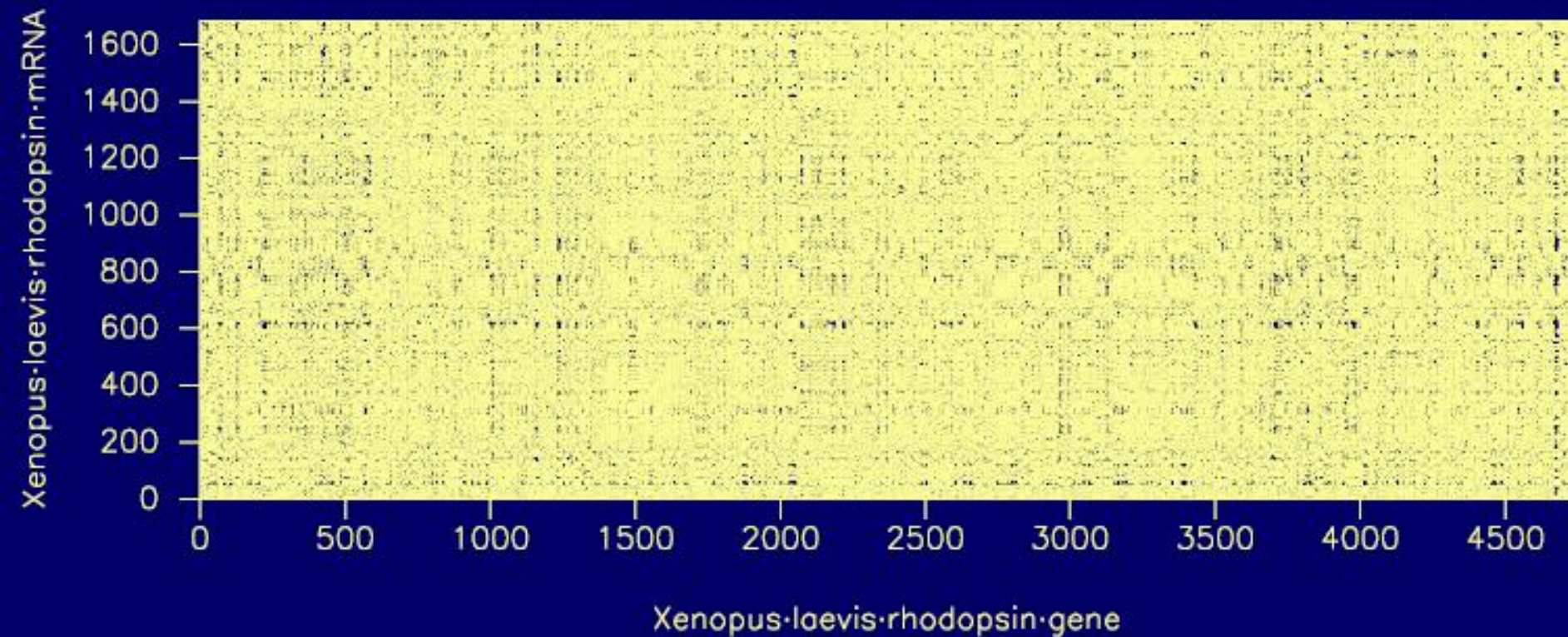
mismatch → □

Fenêtre de taille 5

| | A | C | T | C | G | A | G | C | T | A | T | C | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | | | | | | |
| C | | | | | | | | | | | | | |
| T | | | | | | | | | | | | | |
| A | | | | | | | | | | | | | |
| T | | | | | | | | | | | | | |
| G | | | | | | | | | | | | | |
| A | | | | | | | | | | | | | |
| G | | | | | | | | | | | | | |
| A | | | | | | | | | | | | | |
| T | | | | | | | | | | | | | |
| A | | | | | | | | | | | | | |
| T | | | | | | | | | | | | | |
| G | | | | | | | | | | | | | |

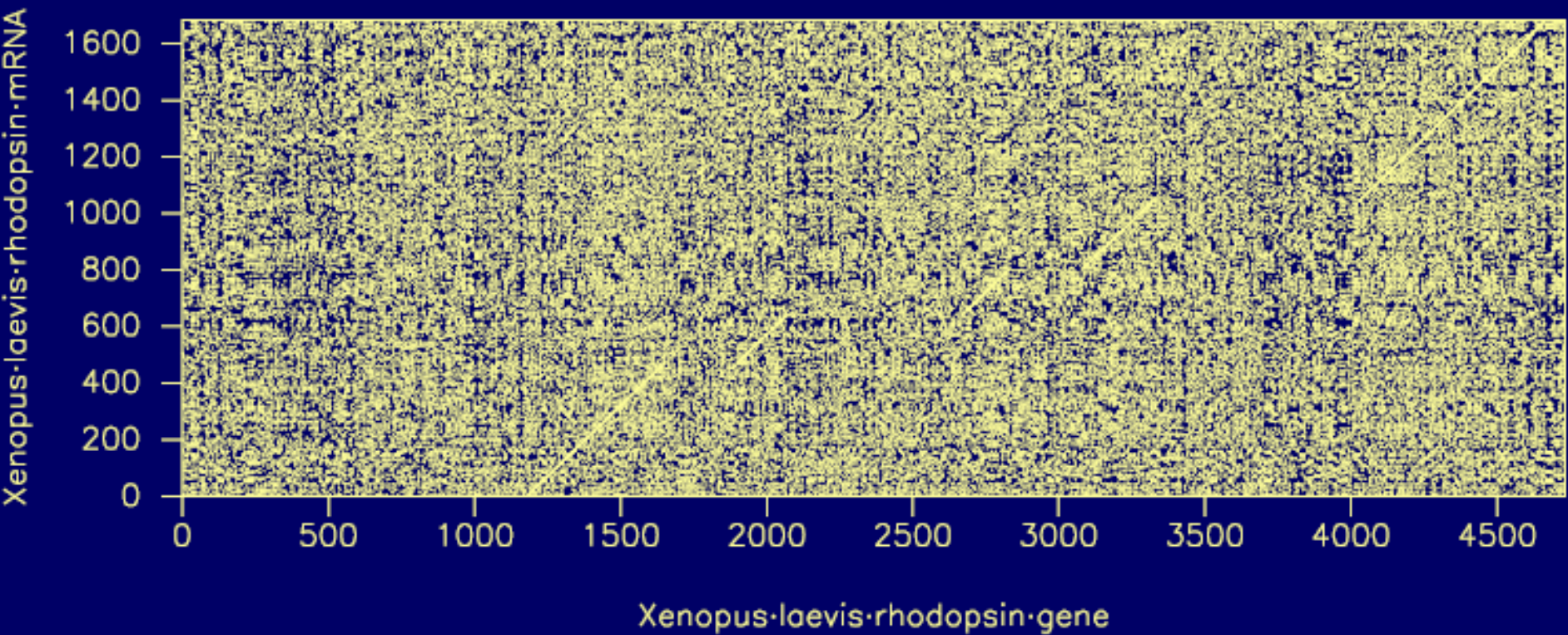


LES «DOT PLOTS»

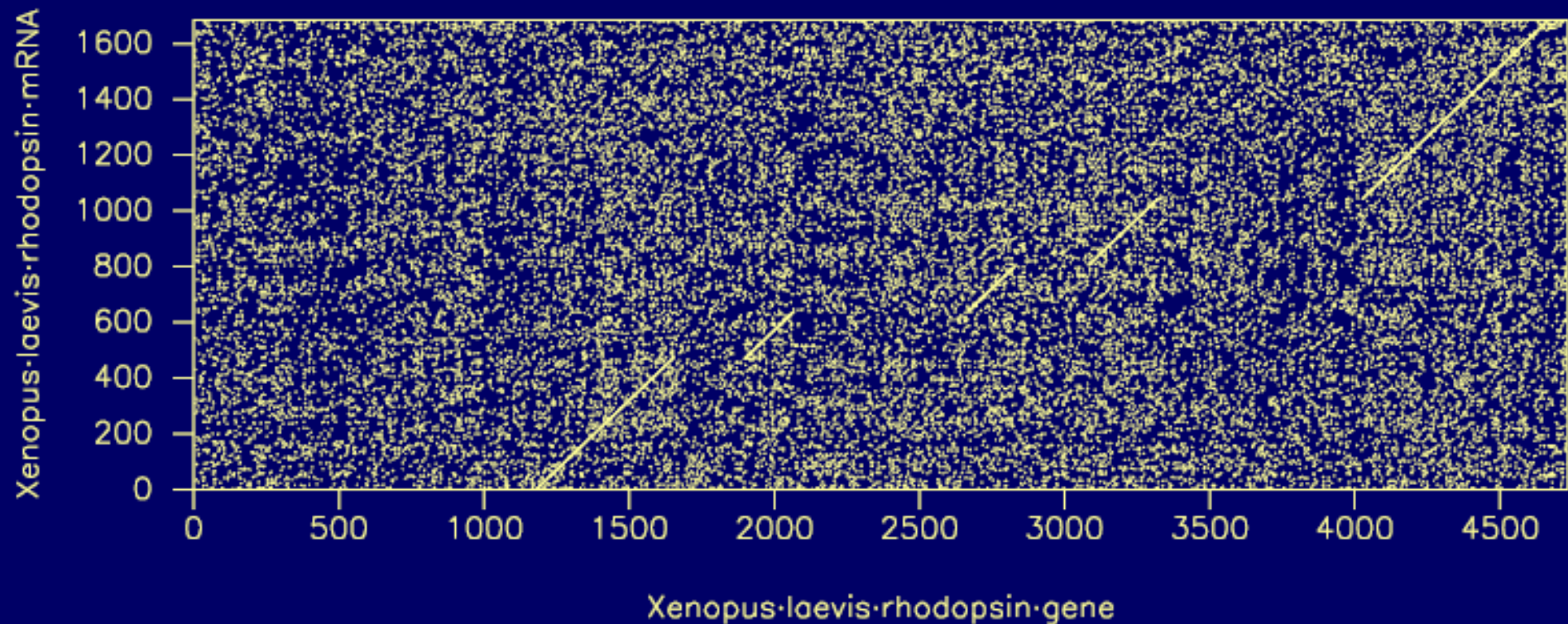


Taille fenêtre = 2

LES «DOT PLOTS»

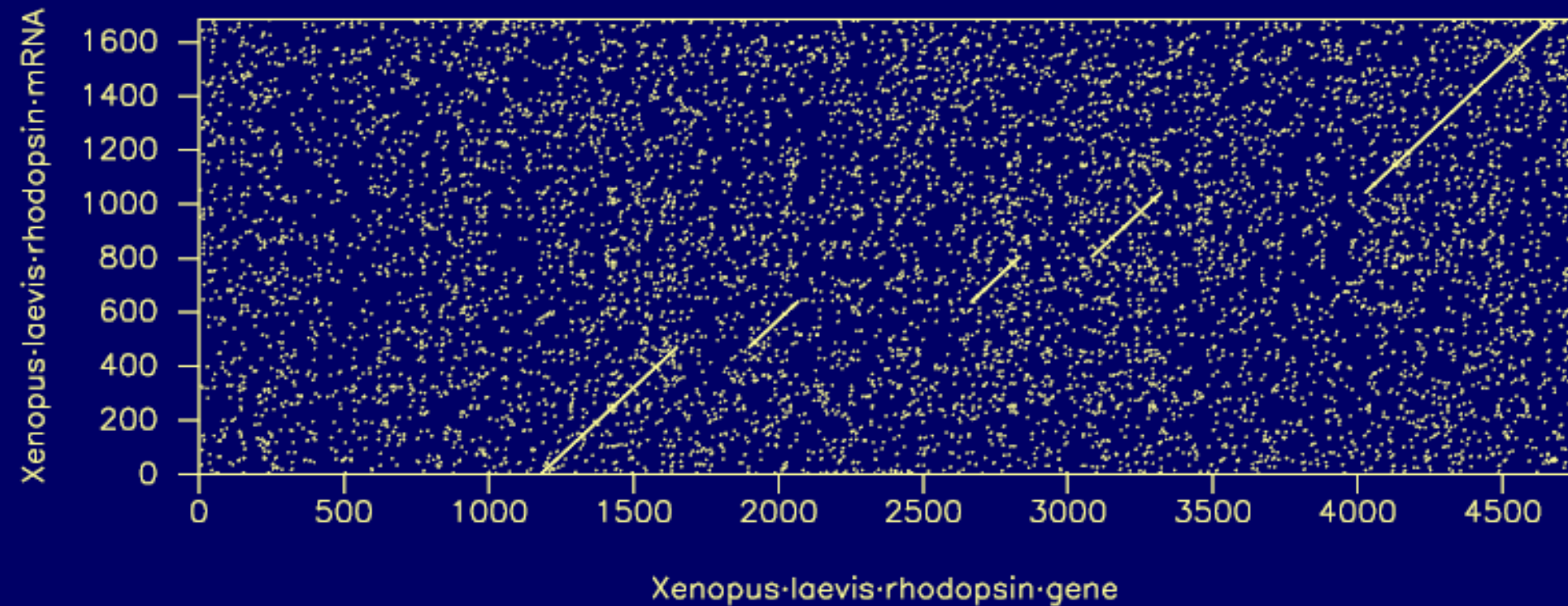


LES «DOT PLOTS»



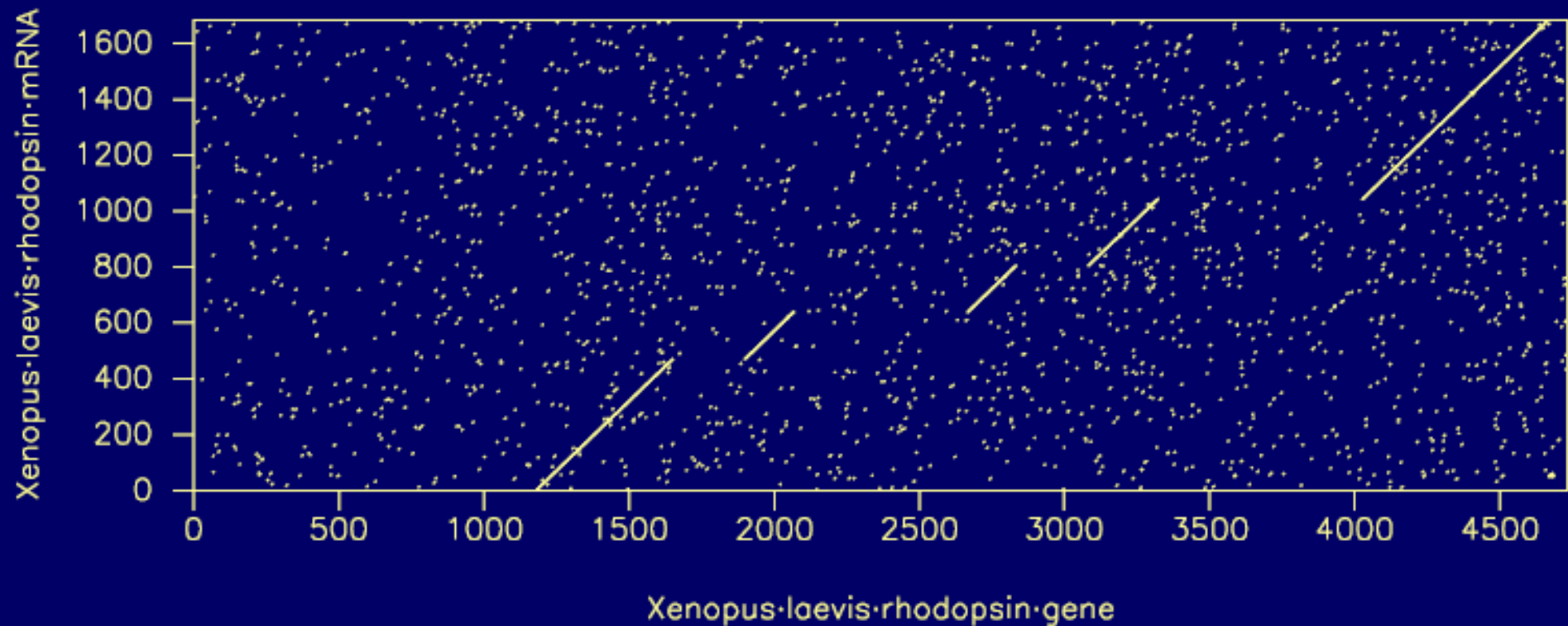
Taille fenêtre = 4

LES «DOT PLOTS»



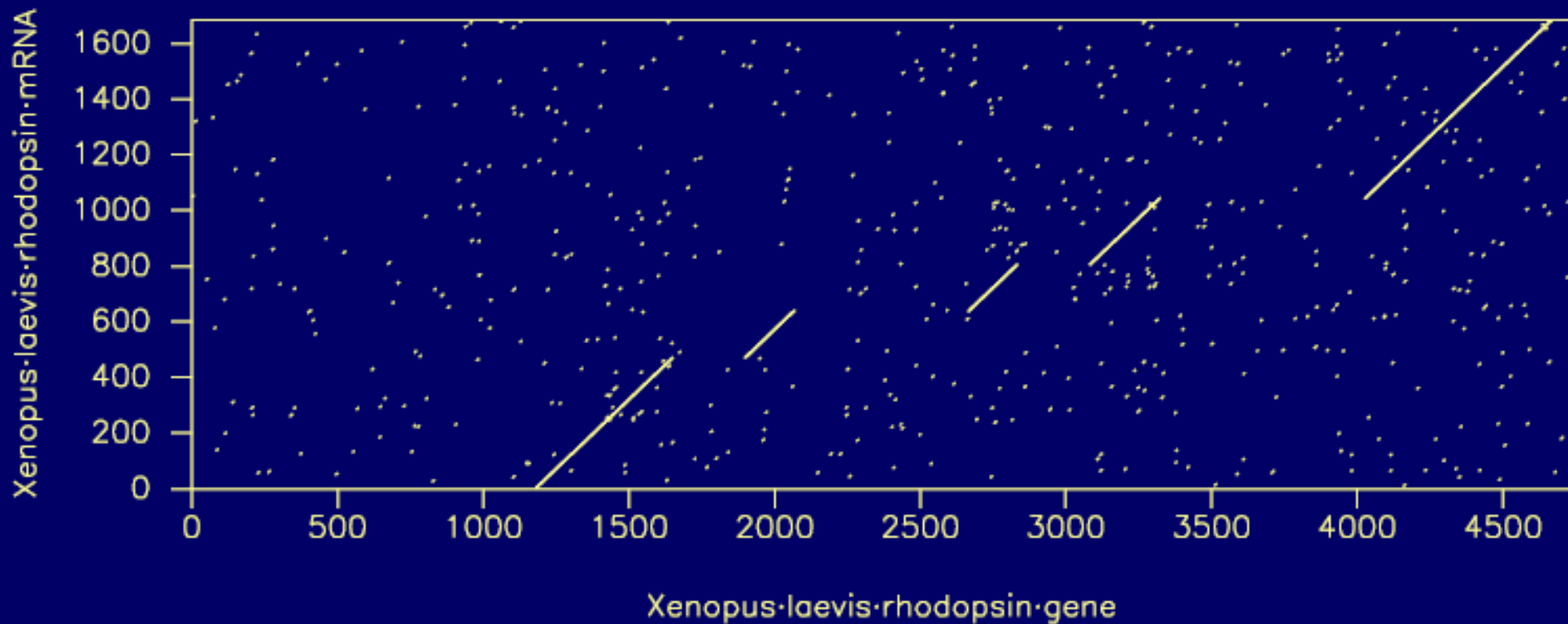
Taille fenêtre = 5

LES «DOT PLOTS»



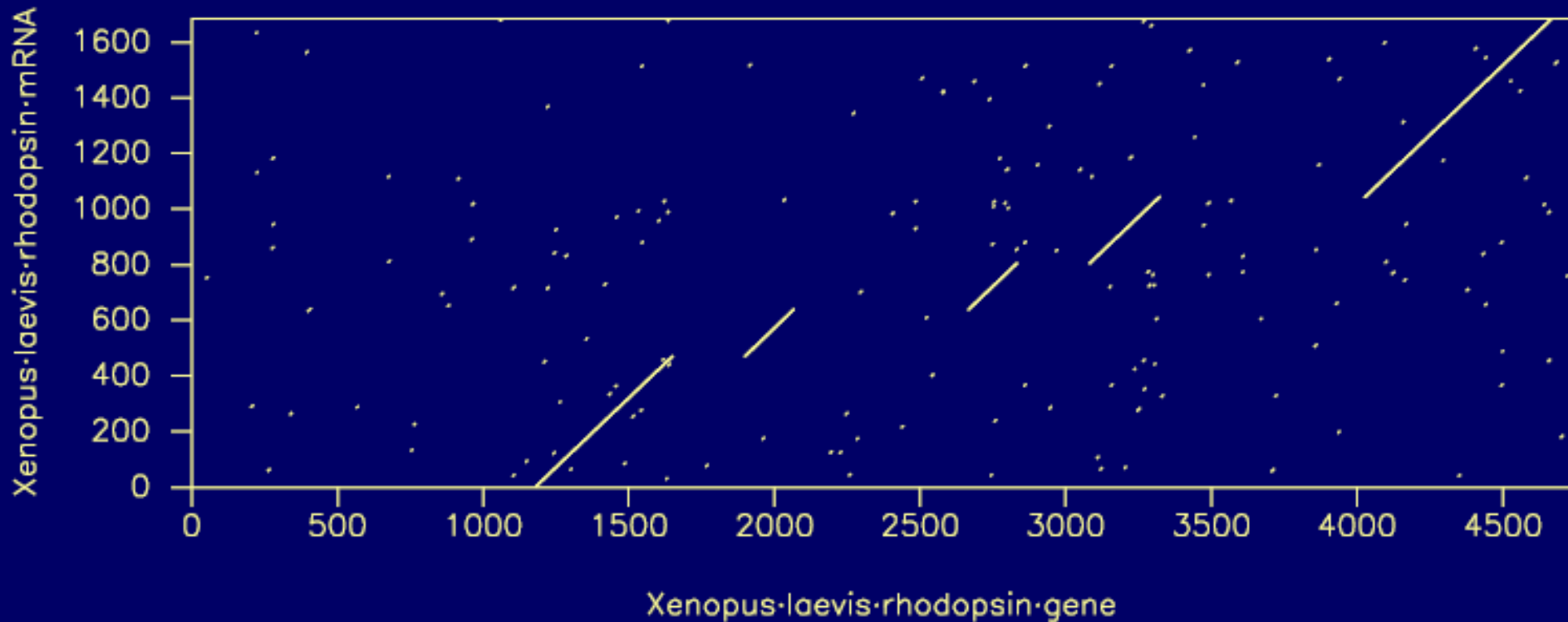
Taille fenêtre = 6

LES «DOT PLOTS»



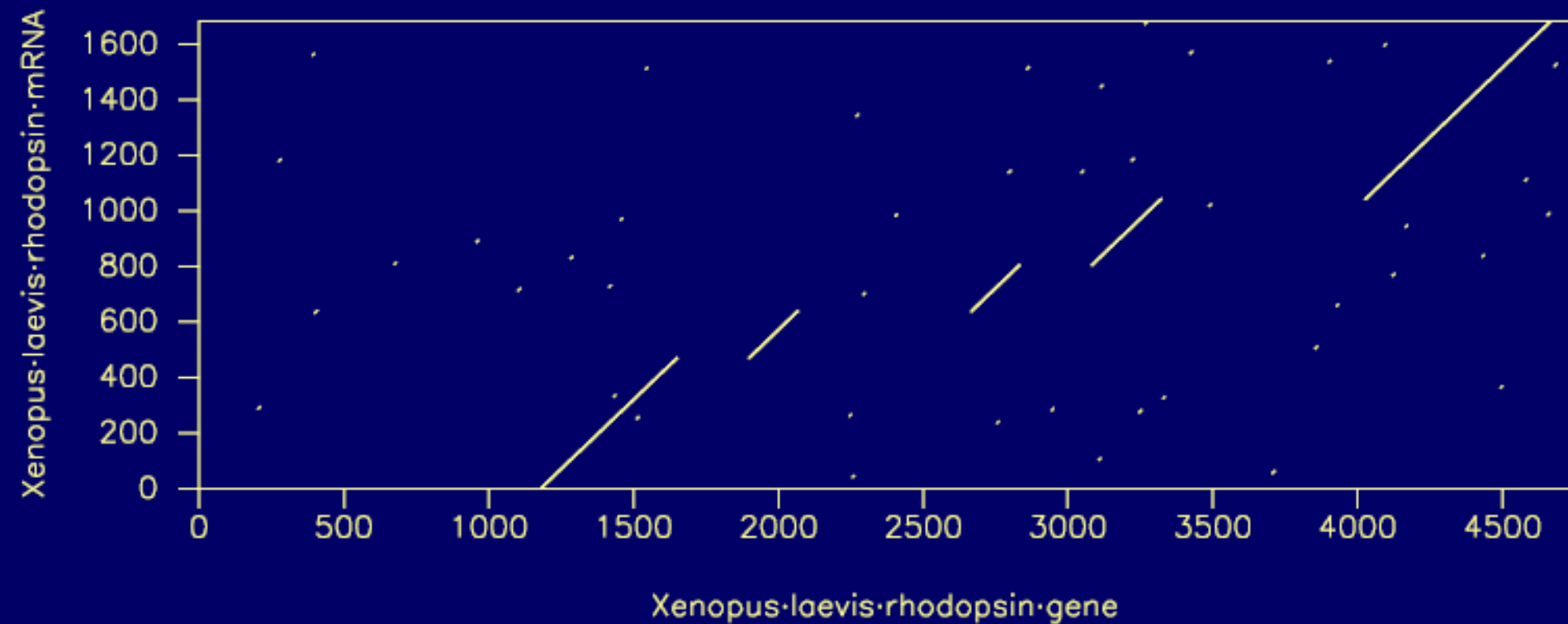
Taille fenêtre = 7

LES «DOT PLOTS»



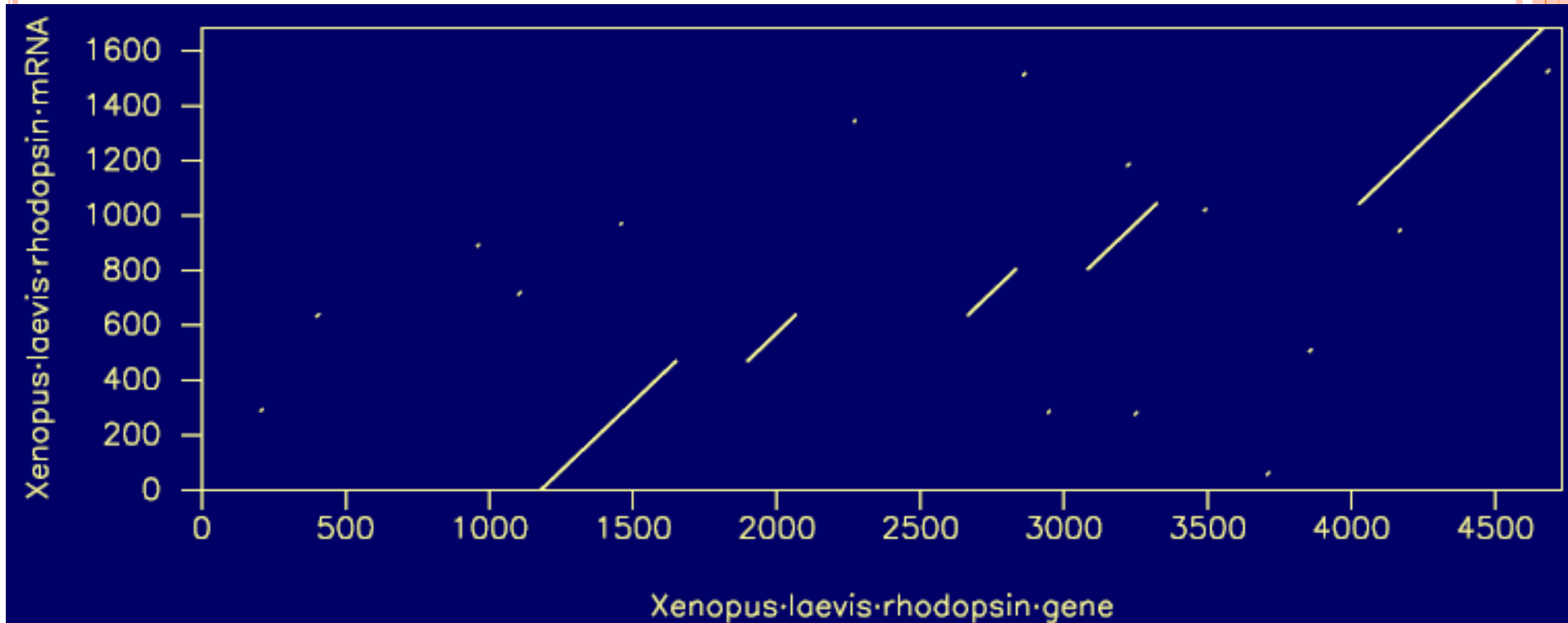
Taille fenêtre = 8

LES «DOT PLOTS»



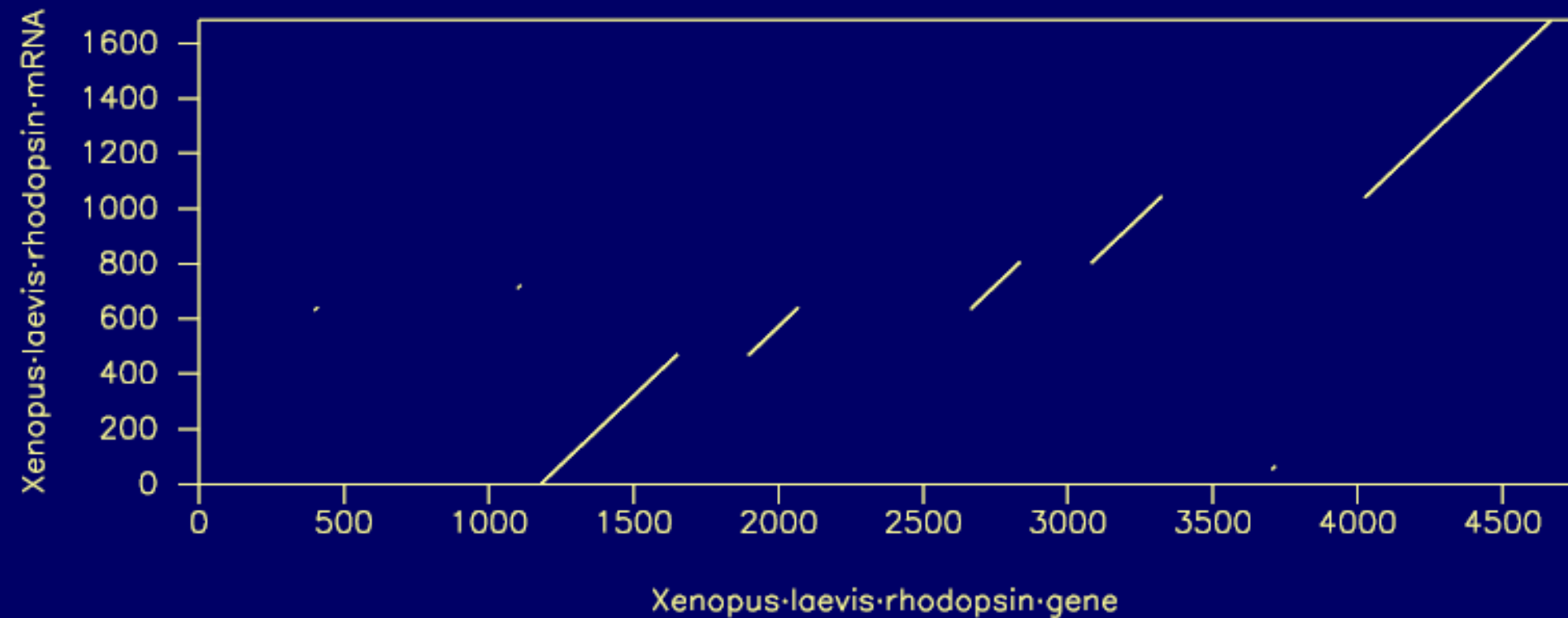
Taille fenêtre = 9

LES «DOT PLOTS»



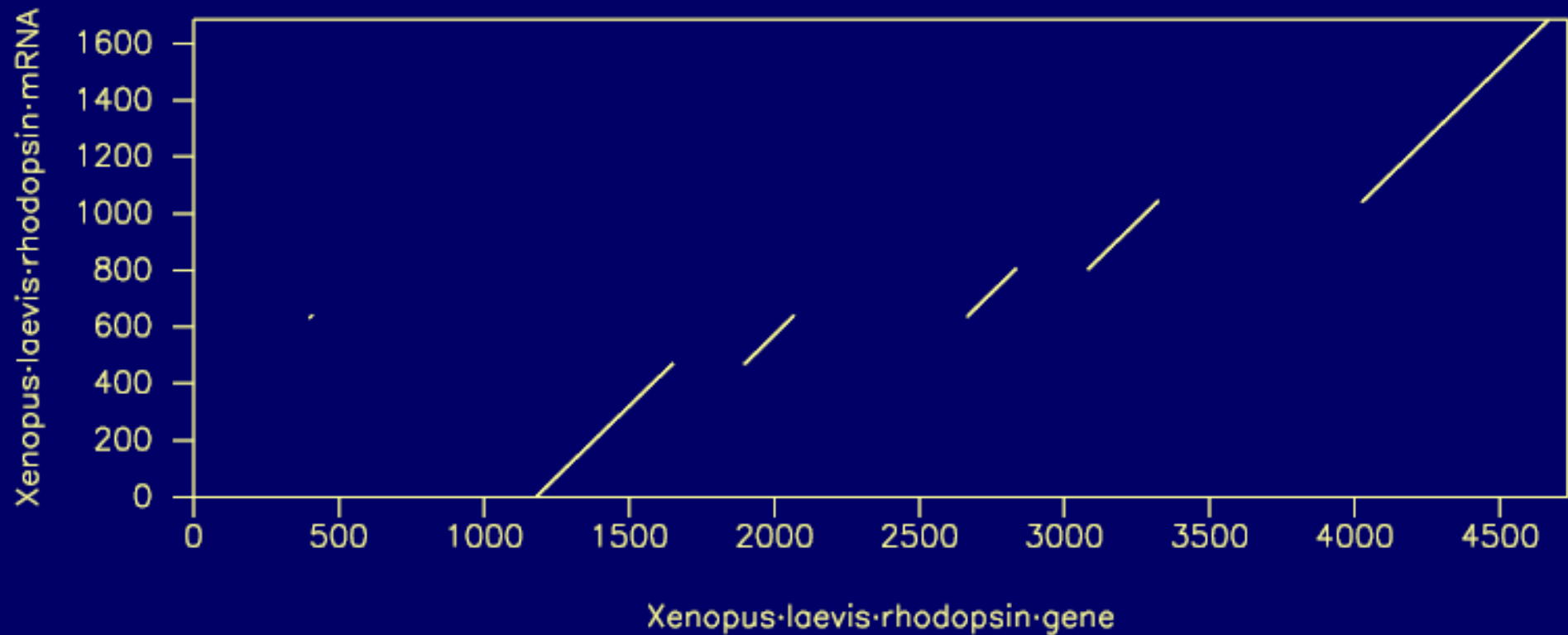
Taille fenêtre = 10

LES «DOT PLOTS»



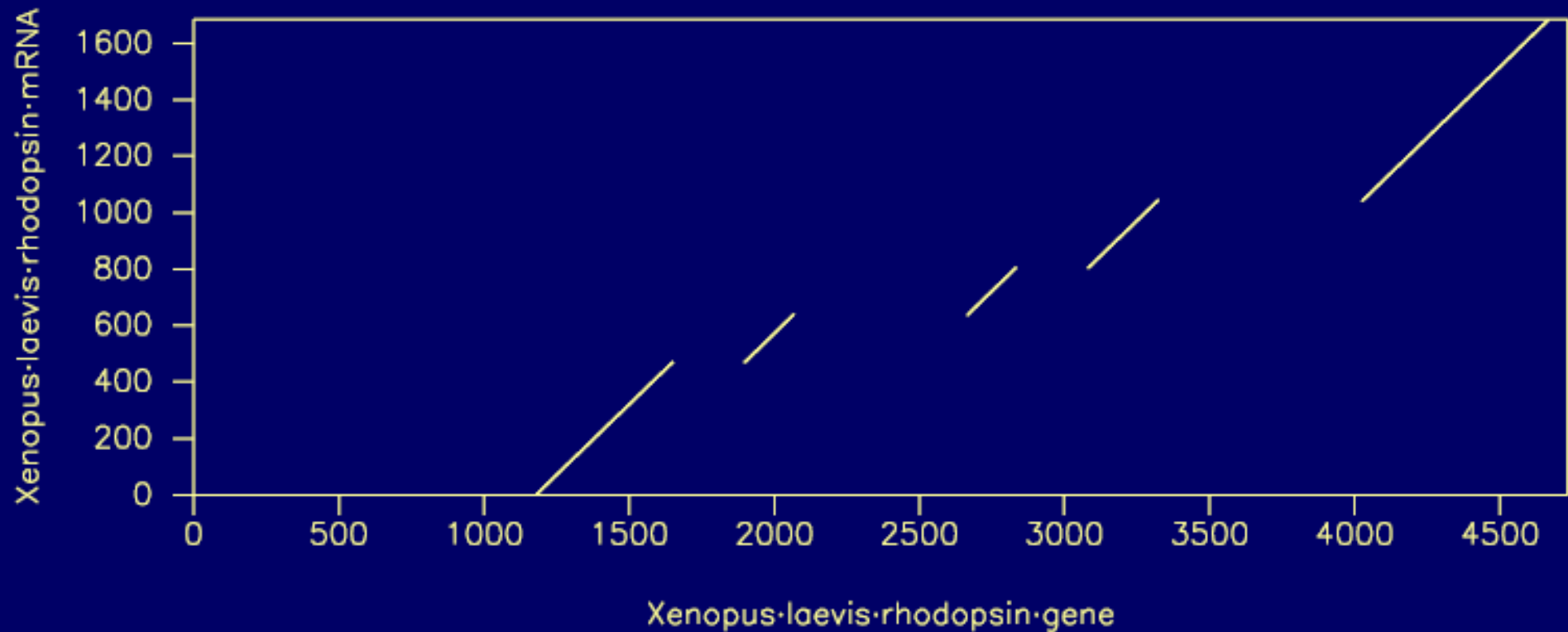
Taille fenêtre = 11

LES «DOT PLOTS»



Taille fenêtre = 12

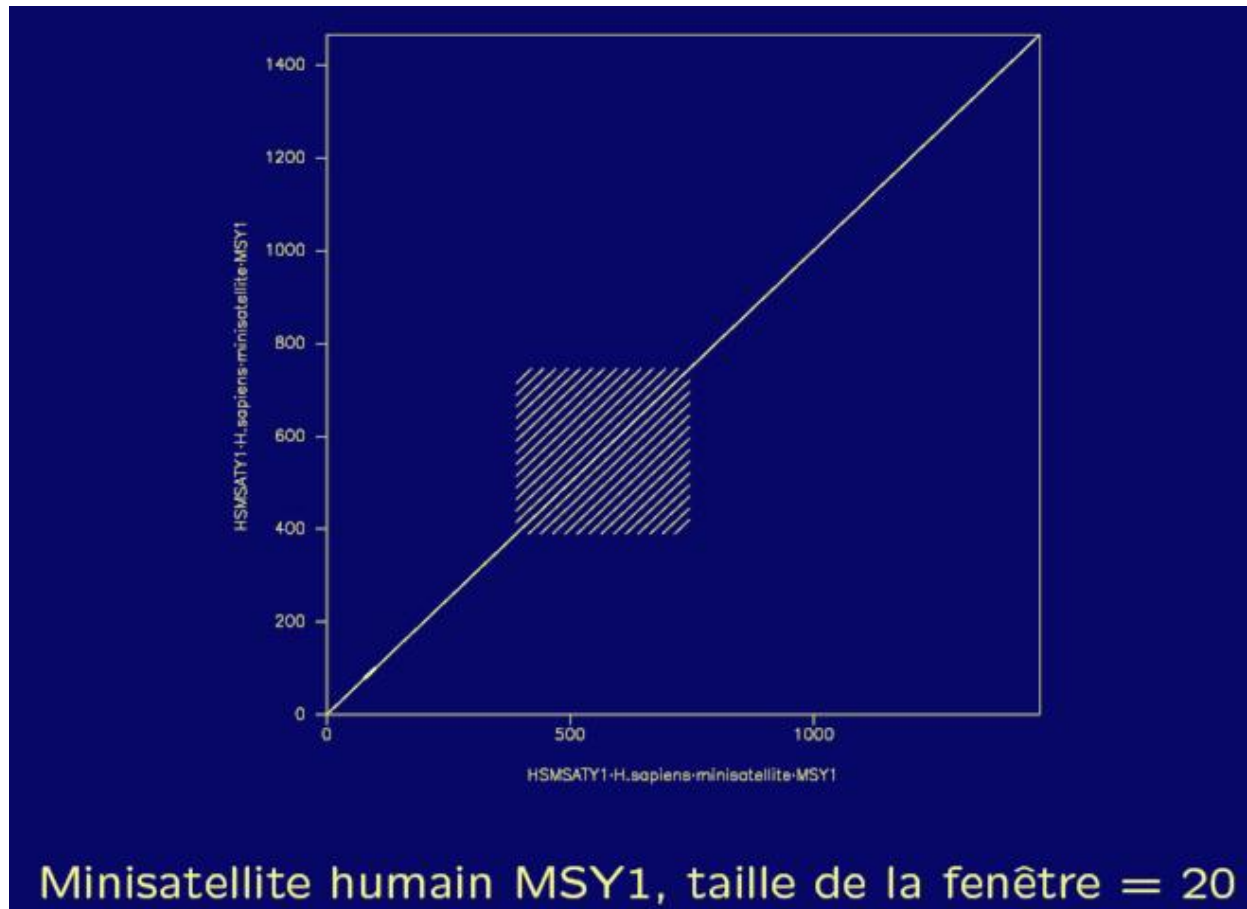
LES «DOT PLOTS»



Taille fenêtre = 20

Alignement = trouver le meilleur chemin dans ce graphe

LES «DOT PLOTS»



ALIGNEMENT DE SÉQUENCES

○ Distance d'édition

- Selon ce concept, le bon alignement est celui qui minimise les opérations à réaliser pour passer d'une séquence à l'autre.
- Opérations: **c**onservation, remplacement/**m**utation, **d**élétion, **i**nsertion. Une pénalité peut être affectée à chaque opération, par exemple **c=0**, **m=1**, **d=2**, **i=2**. La distance finale entre les deux séquences (distance d'édition) est la somme de ces pénalités.

| | | | |
|-----------|--------------|--|-------------------------------|
| | Seq 1 | CAGTGGT-GC | |
| | Seq 2 | CA-TCGTAGC | c=0, m=1, d=2, i=2. |
| Ou, | distance | ccicmccdcc = 0+0+2+0+1+0+0+2+0+0 = 5 | |
| variante: | ressemblance | ccicmccdcc = 2+2-1+2-1+2+2-1+2+2 = 11 | c=2, m=-1, d=-1, i=-1. |

- Une délétion à l'intérieur d'une séquence est considéré comme une insertion dans la séquence lui faisant face.



MATRICES DE SUBSTITUTION - NUCLÉIQUE

- Matrice 4X4 (nucléotides) ou 20x20 (acides aminés) décrivant la distance ou la similitude entre résidus.
 - Estiment le coût ou le taux de remplacement d'un résidu par un autre (distance).
 - Le choix d'une matrice affecte fortement le résultat de l'analyse. Chaque matrice de score représente implicitement une théorie évolutive donnée

Matrices DNA

| | A | C | G | T |
|---|----|----|----|----|
| A | 2 | -1 | -1 | -1 |
| C | -1 | 2 | -1 | -1 |
| G | -1 | -1 | 2 | -1 |
| T | -1 | -1 | -1 | 2 |

Matrice identité

| | A | C | G | T |
|---|----|----|----|----|
| A | 3 | -1 | 1 | -1 |
| C | -1 | 3 | -1 | 1 |
| G | 1 | -1 | 3 | -1 |
| T | -1 | 1 | -1 | 3 |

Matrice transition/transversion

MATRICES DE SUBSTITUTION - PROTÉIQUE

400 changements possibles (20x20) pour les acides aminés mais *non équivalents*

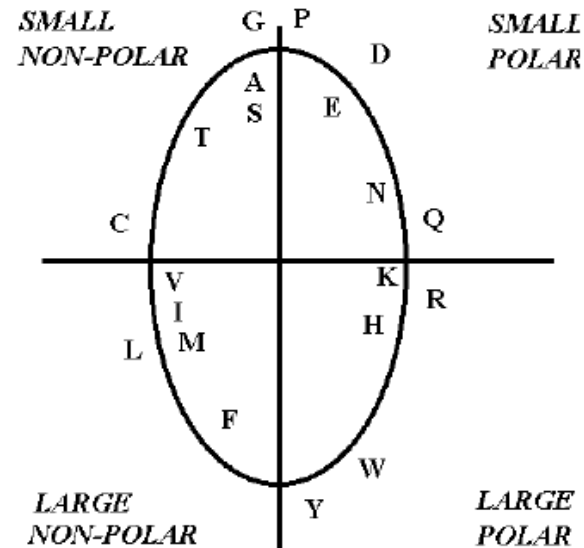
- Matrices fondées sur le code génétique

- Les scores sont déterminés en fonction du nombre commun de nucléotides présents dans les codons des acides aminés, ce qui revient à considérer le minimum de changements nécessaires en bases pour convertir un acide aminé en un autre.

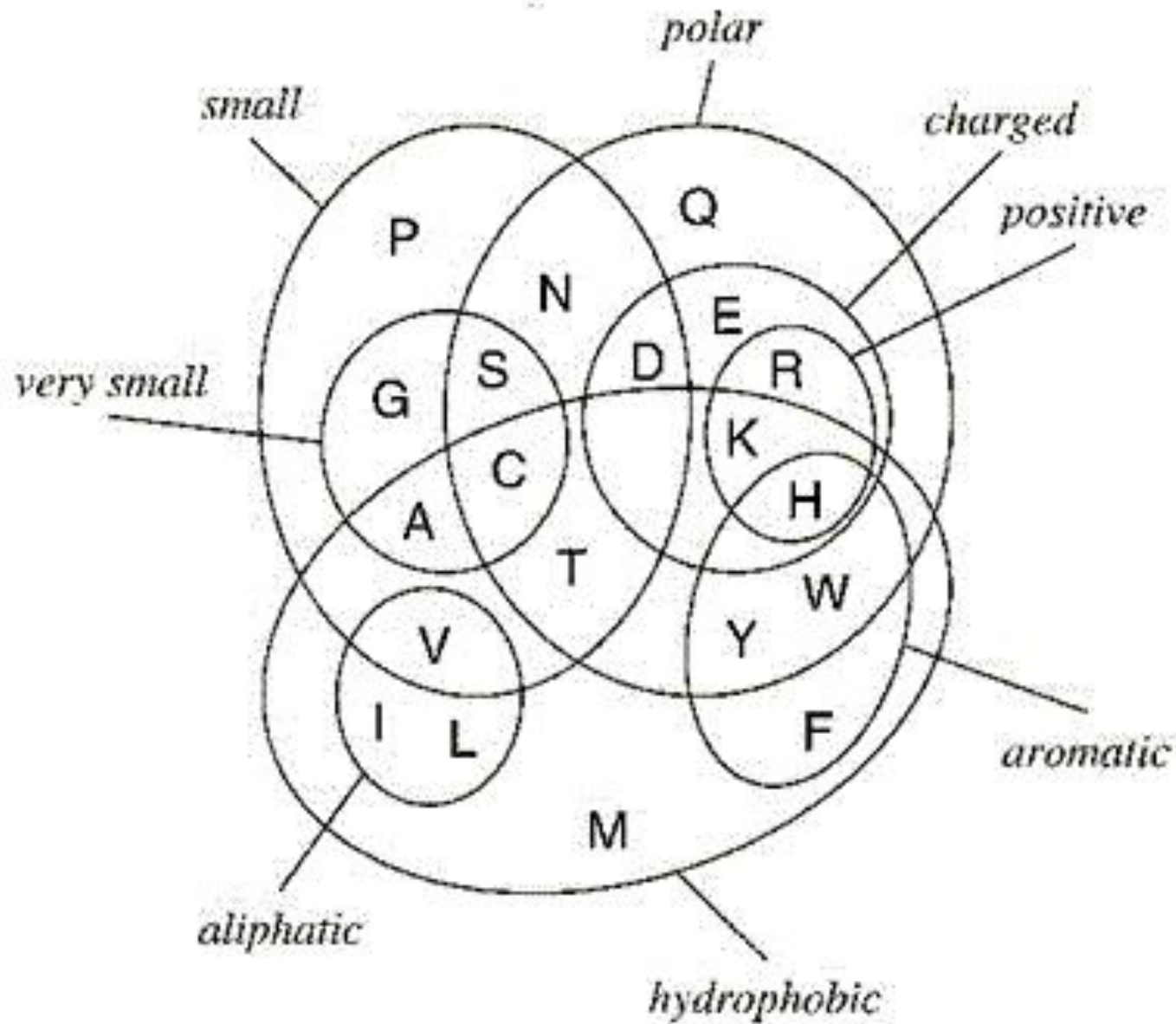
- Matrices fondées sur les propriétés physicochimiques

- Les plus courantes sont celles basées sur le caractère hydrophile ou hydrophobe des protéines. Ces matrices sont peu utilisées.

- Matrices fondées sur l'évolution



Une représentation bidimensionnelle des propriétés des aa calculée d'après la matrice de Dayhoff par G. Vriend, Centre for Molecular and Biomolecular Informatics, University of Nijmegen



Taylor (1986) The Classification of Amino Acid Conservation, *J. Theor. Biol.* 119, 205-218

LES MATRICES PROTÉIQUES LIÉES AU CODE GÉNÉTIQUE

Basée sur le code génétique : une substitution d'un a.a. en un autre se produit d'autant plus rarement que cela nécessite un plus grand nombre de mutations au niveau ADN.

➔ Matrice génétique (Fitch, 1966)

Identité : +3

1 mutation ADN = 2 nt identiques : +2

2 mutations ADN = 1 nt identique : +1

3 mutations ADN = 0 nt identique : 0

| le code génétique | | | | | | | | | | |
|-------------------|-----------------|--------------------|-----|-----|-----|----------------------|-----|------|---|----------------------------|
| | Deuxième lettre | | | | | | | | | |
| | U | | C | | A | | G | | | |
| U | UUU | Phe | UCU | Ser | UAU | Tyr | UGU | Cys | U | Troisième lettre (côté 3') |
| | UUC | Phe | UCC | Ser | UAC | Tyr | UGC | Cys | C | |
| | UUA | Leu | UCA | Ser | UAA | Stop | UGA | Stop | A | |
| | UUG | Leu | UCG | Ser | UAG | Stop | UGG | Trp | G | |
| C | CUU | Leu | CCU | Pro | CAU | His | CGU | Arg | U | |
| | CUC | Leu | CCC | Pro | CAC | His | CGC | Arg | C | |
| | CUA | Leu | CCA | Pro | CAA | Gln | CGA | Arg | A | |
| | CUG | Leu | CCG | Pro | CAG | Gln | CGG | Arg | G | |
| A | AUU | Ile | ACU | Thr | AAU | Asn | AGU | Ser | U | |
| | AUC | Ile | ACC | Thr | AAC | Asn | AGC | Ser | C | |
| | AUA | Ile | ACA | Thr | AAA | Lys | AGA | Arg | A | |
| | AUG | Met | ACG | Thr | AAG | Lys | AGG | Arg | G | |
| G | GUU | Val | GCU | Ala | GAU | Asp | GGU | Gly | U | |
| | GUC | Val | GCC | Ala | GAC | Asp | GGC | Gly | C | |
| | GUA | Val | GCA | Ala | GAA | Glu | GGA | Gly | A | |
| | GUG | Val | GCG | Ala | GAG | Glu | GGG | Gly | G | |
| | | codon d'initiation | | | | codon de terminaison | | | | |

LES MATRICES PROTÉIQUES LIÉES AU CODE GÉNÉTIQUE

Nombre de mutations nécessaires pour passer du codon d'un acide aminé au codon d'un autre acide aminé

Mutation GLU → LYS

d'où

| | | |
|-------------|---|-------------|
| G AA | → | A AA |
| G AG | | A AG |



1 mutation sur la première base du codon = +2



LES MATRICES PROTÉIQUES LIÉES AUX CARACTÉRISTIQUES PHYSICO-CHIMIQUES AU CODE GÉNÉTIQUE

Distance basée sur les propriétés des acides aminés :

- composition, polarité, volume moléculaire (Grantam, 1974)
- matrice d'hydrophobicité de Levitt (1976)
- matrice de structure secondaire (Levin, 1986)
- polarité, hydrophobicité, structure secondaire (Rao, 1987)




LES MATRICES PROTÉIQUES FONDÉES SUR LES FRÉQUENCES DE SUBSTITUTION DES ACIDES AMINÉS AU COURS DE L'ÉVOLUTION

Principe:

- Les séquences homologues ont conservées des fonctions similaires
- Deux acides aminés se ressembleront d'autant plus que la fréquence de substitution observée est grande puisque ces substitutions n'auront pas modifié la fonction de la protéine
- Il est possible d'estimer la fréquence avec laquelle un acide aminé est remplacé par un autre au cours de l'évolution à partir de séquences alignées

Principales approches:

- Comparaison directe de séquences (alignement global): matrice PAM (Dayhoff, 1978)
 - Comparaison de domaines protéiques (régions les plus conservés au cours de l'évolution): matrice BLOSUM (Henikoff et Henikoff, 1992)
 - Alignement de séquences en comparant leurs structures secondaire et tertiaire
- 

MATRICES DE DAYHOFF OU PAM MARGARET DAYHOFF, 1978

- **PAM = Percenta/Point of Accepted Mutation**
- Elle rend compte de deux processus
 1. L'apparition de substitutions
 2. Leur passage au travers le crible de la sélection
- Si deux séquences appartiennent au même processus évolutif, et qu'un acide aminé de l'une a été muté pour donner l'autre, alors on peut supposer que les deux acides aminés sont similaires :
 - les mutations sont dites acceptées (Point Accepted Mutation)
 - elles ont été conservées au cours de l'évolution de part leur caractère à ne pas altérer la fonction de la protéine.
- Les protéines évoluent via des successions de mutations ponctuelles indépendantes les unes des autres et acceptées dans la population.



MATRICES DE DAYHOFF OU PAM MARGARET DAYHOFF, 1978

- Probabilité d'observer la mutation $i \rightarrow j$ après un temps évolutif donné. Basée sur alignement global de ~1300 protéines conservées à plus de 85% appartenant à 71 familles de protéines.
- Aujourd'hui actualisées : 16 130 séquences appartenant à 2 621 familles de protéines

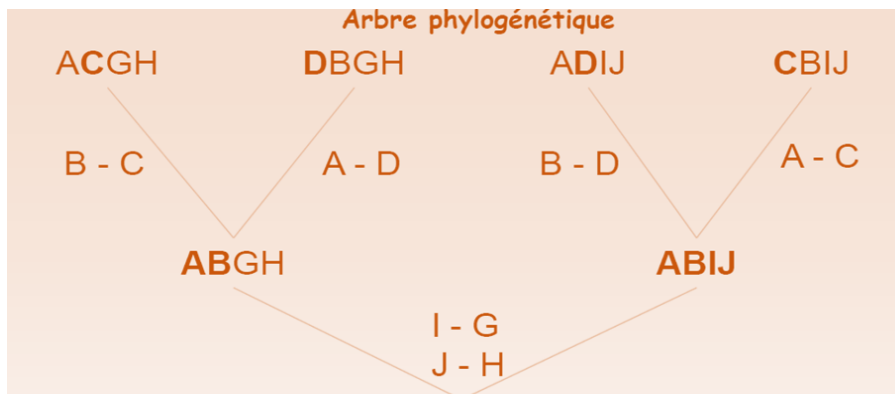


MATRICES DE DAYHOFF OU PAM MARGARET DAYHOFF, 1978

Les étapes:

1. alignement multiple global de chaque famille
2. arbre phylogénétique pour chaque famille (c'est à dire séquence ancestrale à chaque nœud)
3. décompte du nombre de changements pour chaque acide aminé

A_{ij} : nombre de mutations $i \rightarrow j$ acceptées observées



Matrice des mutations acceptées

| | A | B | C | D | G | H | I | J |
|---|---|---|---|---|---|---|---|---|
| A | | | 1 | 1 | | | | |
| B | | | 1 | 1 | | | | |
| C | 1 | | | | | | | |
| D | 1 | | | | | | | |
| G | | | | | | | 1 | |
| H | | | | | | | | 1 |
| I | | | | | 1 | | | |
| J | | | | | | 1 | | |

4. cumul des mutations acceptées au sein des différentes familles



MATRICES DE DAYHOFF OU PAM MARGARET DAYHOFF, 1978

5. calcul de la mutabilité des acides aminés (m_j)

- Propension d'un acide aminé à être remplacé par un autre acide aminé
- Soit l'alignement suivant :

A D E F R E
A D D W R E

- Les acides aminés en jeu sont les suivants : **A, D, E, F, W** et **R**.

| | A | D | E | F | W | R |
|------------------------------|----------|------------|------------|----------|----------|----------|
| Nombre de changements | 0 | 1 | 1 | 1 | 1 | 0 |
| Nombre d'occurrences | 2 | 3 | 3 | 1 | 1 | 2 |
| Mutabilité | 0 | 1/3 | 1/3 | 1 | 1 | 0 |



MATRICES DE DAYHOFF OU PAM MARGARET DAYHOFF, 1978

6. Calcul de la matrice de probabilité de mutations

$$M_{ij}^1 = \lambda m_j \frac{A_{ij}}{\sum_{i=1}^{20} A_{ij}} \quad \lambda \text{ ajustement pour avoir une mutation sur 100 sites}$$

7. Calcul de la matrice « odds »

- Chaque élément de la matrice est divisé par la fréquence p_i d'occurrence de chaque acide aminé

$$R_{ij}^1 = \frac{M_{ij}^1}{p_i} \quad \text{1PAM}$$

unité de changement évolutif = l'unité PAM

Deux séquences sont séparées par une distance évolutive de 1 PAM si il y a eu 1 changement observé et accepté au cours de l'évolution pour 100 acides aminés.



MATRICES DE DAYHOFF OU PAM MARGARET DAYHOFF, 1978

8. Calcul de la matrice lods «log odds »

$$S_{ij}^1 = \log R_{ij}^1 + \log R_{ji}^1 = \text{PAM1}$$

9. Calcul des matrices à d'autres temps évolutifs par extrapolation

- Hypothèse: les mutations indépendantes

⇒ multiplication des PAM

PAM2=PAM1*PAM1 , PAM120=PAM1¹²⁰, etc...

PAM250=PAM1²⁵⁰

PAM-250 (250 mutations pour une séquence de 100 aa) : du fait des mutations silencieuses (synonymes) et des mutations reverses, cela correspond à environ 20% d'identité.

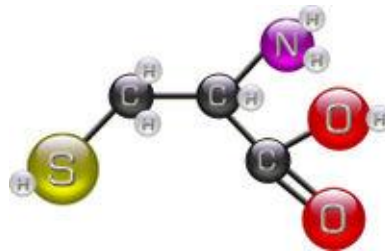
MATRICES DE DAYHOFF OU PAM MARGARET DAYHOFF, 1978

- Dans cette matrice de similitude, plus la valeur est négative, plus la probabilité est faible, plus le remplacement est rare.



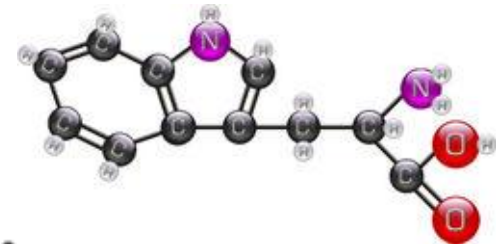
MATRICES DE DAYHOFF OU PAM MARGARET DAYHOFF, 1978

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|---|---|---|
| A | 2 | | | | | | | | | | | | | | | | | | | | | | |
| R | -2 | 6 | | | | | | | | | | | | | | | | | | | | | |
| N | 0 | 0 | 2 | | | | | | | | | | | | | | | | | | | | |
| D | 0 | -1 | 2 | 4 | | | | | | | | | | | | | | | | | | | |
| C | -2 | -4 | -4 | -5 | 12 | | | | | | | | | | | | | | | | | | |
| Q | 0 | 1 | 1 | 2 | -5 | 4 | | | | | | | | | | | | | | | | | |
| E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | | | | | | | | | | | | | | | | |
| G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | | | | | | | | | | | | | | | |
| H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | | | | | | | | | | | | | | |
| I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | | | | | | | | | | | | | |
| L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | | | | | | | | | | | | |
| K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | | | | | | | | | | | |
| M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | | | | | | | | | | |
| F | -4 | -4 | -4 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | | | | | | | | | |
| P | 1 | 0 | -1 | -1 | -3 | 0 | -1 | -1 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | | | | | | | | |
| S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | | | | | | | |
| T | 1 | -1 | 0 | 0 | 2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | | | | | | |
| W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | -0 | -6 | -2 | -5 | 17 | | | | | |
| Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | | | | |
| V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 | | | |
| B | 0 | -1 | 2 | 3 | -4 | 1 | 2 | 0 | 1 | -2 | -3 | 1 | -2 | -5 | -1 | 0 | 0 | -5 | -3 | -2 | 2 | | |
| Z | 0 | 0 | 1 | 3 | -5 | 3 | 3 | -1 | 2 | -2 | -3 | 0 | -2 | -5 | 0 | 0 | -1 | -6 | -4 | -2 | 2 | 3 | |
| X | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |



Cystéine

Acides aminés peu permutable



Tryptophane



MATRICES DE SUBSTITUTION BLOSUM

- Le but est de détecter des relations entre protéines plus éloignées.
- Avec les matrices PAM, les valeurs pour des protéines éloignées sont extrapolées. Avec BLOSUM, ces valeurs sont obtenues en comparant des blocs facilement « alignables » (sans gaps) dans des familles de protéines très éloignées.
- Ces matrices sont reconnues pour mettre en valeur les similitudes biologiquement importantes (celles qui sont présentes dans les régions alignées sans gaps).
- BLOSUM62: faite à partir d'un alignement de séquences ayant 62% de similitude, BLOSUM45: 45%, etc.



MATRICES DE SUBSTITUTION BLOSUM

BLOSUM (***B**LOCKS of Amino Acid **S**UBstitution **M**atrix*)

Principe :

- Obtention à partir de blocs de séquences alignées (alignement multiple local sans brèche/trous/gaps)
- Pour une paire d'acides aminés : $\log (\text{fréquence observée} / \text{fréquence attendue})$

Avantages par rapport aux matrices PAM :

- contrairement aux matrices PAM, les matrices BLOSUM pour différentes distances évolutives sont obtenues directement avec des séquences plus ou moins divergentes
- l'utilisation de blocs plutôt que de séquences complètes : modélise les contraintes uniquement sur les régions conservées obtenues à partir d'un plus grand jeu de données (>2000 blocks, > 500 familles)



MATRICES DE SUBSTITUTION BLOSUM

BLOSUM (***B**LOCKS of Amino Acid **S**UBstitution **M**atrix*)

Dans les matrices de type BLOSUM, les fréquences sont observées sur des alignements de séquences très divergentes.

Néanmoins, dans de tels alignements, les séquences sont moins bien alignées et les « trous » sont plus fréquents.

Afin de ne pas introduire de biais dû à ces trous, les matrices BLOSUM utilisent des blocs bien alignés et surtout sans trous provenant de la base **BLOCKS** (<http://blocks.fhcrc.org/>).

Dans ces blocs bien alignés, certaines séquences restent très proches.

Afin de réduire le biais venant de ces séquences redondantes, elles sont regroupées (clusterisées) si leur identité dépasse un certain seuil.

Dans une matrice BLOSUM62, les séquences présentant plus de 62% d'identité ont été regroupées ensemble.



AUTRE MATRICE DE SUBSTITUTION

- Matrices d'après alignement 3D
 - Basées sur la structure secondaire ou tertiaire.
 - Évaluent la propension d'un acide aminé à adopter une certaine conformation. Fiables car fondées sur le meilleur alignement possible.
 - Encore incomplètes en raison de la taille des banques de données 3D.



CORRESPONDANCES PAM ET BLOSUM

Famille de matrices correspondant à différentes distances évolutives entre les séquences :

- PAM40 et BLOSUM80: estimation des fréquences de substitution entre acides aminés pour des séquences proches dans l'évolution (courtes distances)
- PAM250 (PAM350) et BLOSUM30: estimation des fréquences de substitution entre acides aminés pour des séquences distantes dans l'évolution (longues distances)
- PAM120 et BLOSUM62 : estimation des fréquences de substitution entre acides aminés pour des séquences ayant des distances évolutives intermédiaires.



QUELLE MATRICE DOIT-ON UTILISER?

- Les matrices BLOSUM sont les plus souvent proposées comme matrices par défaut car les fréquences de substitution sont directement calculées à partir de l'alignement.
- La BLOSUM62 (ou PAM120) est utilisée comme matrice par défaut car elle offre un bon compromis quand les distances évolutives entre les séquences ne sont pas connues.
- La BLOSUM80 (ou PAM40) donnera de meilleurs résultats pour des séquences proches dans l'évolution. Elle tend à trouver des alignements courts fortement similaires.
- La BLOSUM30 (ou PAM350) donnera de meilleurs résultats pour des séquences éloignées dans l'évolution. Elle trouvera de plus longs alignements locaux de faible conservation.



LE SCORE D'UN ALIGNEMENT

○ **Score = Σ score élémentaire - Σ pénalité d'insertions/délétions**

- Score est fonction de la longueur de la séquence: plus la séquence est longue plus le score est élevé.
- Il dépend de la matrice de substitution utilisée.
- Le calcul du score se nuance en donnant plus ou moins de poids aux pénalités
- Introduction d'insertion ou de délétion de longueur variable à certaines positions des séquences



TRAITEMENT DES INSERTIONS ET DES DÉLÉTIONS

1. **Pénalité simple pour chaque insertion quelle que soit sa longueur**

ou

1. **Pénalité fixe pour toute insertion + pénalité pour étendre l'insertion**

- Pénalité moins lourde mais permet de prendre en compte la longueur
 - $P = x + yL$
 - P: pénalité pour une insertion de longueur L
 - x: pénalité fixe d'insertion indépendante de la longueur
 - y: pénalité extension pour l'élément
 - Pénalité fixe varie généralement de l'ordre de 1 à 5 fois le score donné pour une bonne association entre deux éléments
 - Pénalité d'extension est souvent très inférieure à la pénalité fixe (ordre de 10 fois)
 - Facilitera souvent dans un alignement le fait d'avoir peu d'insertions éventuellement longues plutôt que beaucoup d'insertions d'un seul élément.
 - Concordance avec les évènements biologiques observés
- Poids des pénalités peut être établi selon les endroits où elles se trouvent afin d'améliorer la sensibilité de la recherche (feuilletés bêta, hydrophobicité...)
 - La recherche d'alignement optimaux est basée sur le fait que les séquences doivent contenir un grand nombre d'éléments identiques ou équivalents.



MÉTHODE DE PROGRAMMATION DYNAMIQUE

- Temps de comparaison de deux séquences de longueur équivalente N est proportionnel à N^2
- L'exploration de chaque position de chaque séquence pour la détermination éventuelle d'une insertion augmente d'un facteur $2N$ le temps de calcul.
- La programmation dynamique est un moyen de limiter cette augmentation pour conserver un temps de calcul de l'ordre de N^2 .
- Elle est basée sur le fait que tous les événements sont possibles et calculables mais que la plupart sont rejetés en considérant certains critères.
- Needleman et Wunsch (1970) ont introduit les premiers ce type d'approche pour un problème biologique et leur algorithme reste une référence dans le domaine.



L'ALGORITHME DE NEEDLEMAN ET WUNSCH

Transformation de la matrice de comparaison initiale en matrice de comparaison transformée selon l'algorithme de Needleman et Wunsch

| | V | T | E | E | R | D | A | F |
|---|----|----|----|----|----|----|----|----|
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |
| T | 0 | 3 | 0 | 0 | -1 | 0 | 1 | -3 |
| S | -1 | 1 | 0 | 0 | 0 | 0 | 1 | -3 |
| H | -2 | -1 | 1 | 1 | 2 | 1 | -1 | -2 |
| E | -2 | 0 | 4 | 4 | -1 | 3 | 0 | -5 |
| A | 0 | 1 | 0 | 0 | -2 | 0 | 2 | -4 |
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |

a) Matrice initiale obtenue à partir de la matrice de substitution utilisée pour l'alignement (ici la matrice PAM250 de Dayhoff)

| | V | T | E | E | R | D | A | F |
|---|----|----|----|----|----|----|----|----|
| L | 14 | 7 | 6 | 6 | 4 | 4 | 0 | 2 |
| T | 10 | 12 | 9 | 9 | 6 | 4 | 3 | -3 |
| S | 8 | 10 | 9 | 9 | 7 | 4 | 3 | -3 |
| H | 6 | 7 | 9 | 8 | 9 | 5 | 1 | -2 |
| E | 2 | 4 | 8 | 8 | 3 | 7 | 2 | -5 |
| A | 2 | 3 | 2 | 2 | 0 | 2 | 4 | -4 |
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |

b) Matrice transformée construite à partir de la matrice initiale



L'ALGORITHME DE NEEDLEMAN ET WUNSCH

$$S(i,j) = se(i,j) + \max S(x,y)$$

$$\text{avec } i \leq x \leq m \text{ et } y = j+1 \quad (3)$$

$$\text{ou } x = i+1 \text{ et } j < y \leq n$$

| | V | T | E | E | R | D | A | F |
|---|----|----|----|----|----|----|----|----|
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |
| T | 0 | 3 | 0 | 0 | -1 | 0 | 1 | -3 |
| S | -1 | 1 | 0 | 0 | 0 | 4 | 3 | -3 |
| H | 6 | 7 | 9 | 8 | 9 | 5 | 1 | -2 |
| E | 2 | 4 | 8 | 8 | 3 | 7 | 2 | -5 |
| A | 2 | 3 | 2 | 2 | 0 | 2 | 4 | -4 |
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |

| | V | T | E | E | R | D | A | F |
|---|----|----|----|----|----|----|----|----|
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |
| T | 0 | 3 | 0 | 0 | -1 | 0 | 1 | -3 |
| S | -1 | 1 | 0 | 0 | 7 | 4 | 3 | -3 |
| H | 6 | 7 | 9 | 8 | 9 | 5 | 1 | -2 |
| E | 2 | 4 | 8 | 8 | 3 | 7 | 2 | -5 |
| A | 2 | 3 | 2 | 2 | 0 | 2 | 4 | -4 |
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |

c) On montre ici la matrice transformée en cours de construction. La séquence horizontale est indiquée en i et la séquence verticale en j. A gauche, la matrice avant le calcul du score somme à la position i=5 et j=3 et à droite, après ce calcul. Le score somme est obtenue à partir de l'expression 3 décrite dans le texte, c'est-à-dire ici en additionnant le score de substitution de R par S (0) et le score maximum de la zone grisée (7).

Chaque score de la matrice transformée est obtenu par la somme du score actuel et du score maximum déjà obtenu

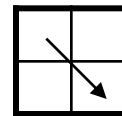
L'ALGORITHME DE NEEDLEMAN ET WUNSCH

| | V | T | E | E | R | D | A | F |
|---|----|----|----|----|----|----|----|----|
| L | 14 | 7 | 6 | 6 | 4 | 4 | 0 | 2 |
| T | 10 | 12 | 9 | 9 | 6 | 4 | 3 | -3 |
| S | 8 | 10 | 9 | 9 | 7 | 4 | 3 | -3 |
| H | 6 | 7 | 9 | 8 | 9 | 5 | 1 | -2 |
| E | 2 | 4 | 8 | 8 | 3 | 7 | 2 | -5 |
| A | 2 | 3 | 2 | 2 | 0 | 2 | 4 | -4 |
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |

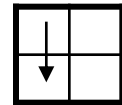
VT-EERDAF
LTSHE--AL

Résultat de l'alignement

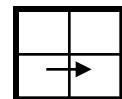
Établissement du chemin des scores maximum dans la matrice transformée. Le chemin est établi en partant du score somme le plus élevé, ici 14. Les flèches indiquent les endroits où il est nécessaire de faire des insertions/délétions pour un alignement global optimum.



Match ou mismatch



insertion

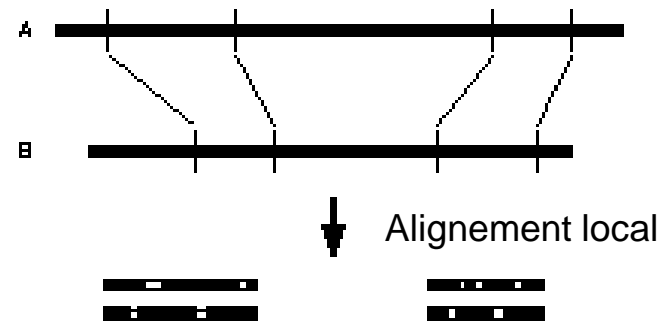
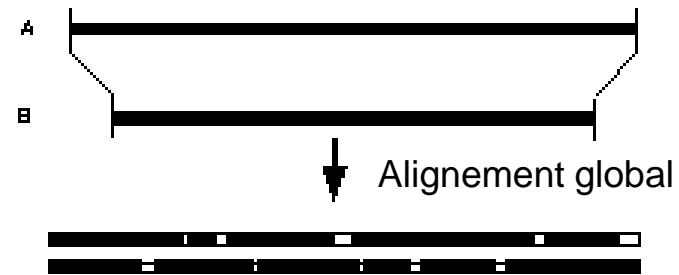


délétion



ALIGNEMENT LOCAL OU GLOBAL

- Des finalités très différentes:
 - L'alignement global est conçu pour comparer des séquences homologues sur toute leur longueur.
 - L'alignement local est conçu pour rechercher des régions semblables entre A et B.



LES PROGRAMMES D'ALIGNEMENT GLOBAL

- Méthode employée pour aligner des séquences dont on soupçonne l'homologie.
- L'alignement est optimisé sur toute la longueur des séquences.
- L'algorithme de référence est celui de Needleman & Wunsch (1970).
- Utilisé principalement aujourd'hui dans le cadre de l'alignement multiple



LES PROGRAMMES D'ALIGNEMENT LOCAL

- Aligné seulement les régions dont le score est supérieur à un seuil donné.
- Utilisé lorsque l'on veut aligner deux séquences de taille très différente. (par ex. dans une recherche de sous séquence).
- Beaucoup plus rapide que l'alignement global.
- Smith et Warteman, Fasta, Blast



L'ALGORITHME DE SMITH ET WATERMAN

- Programmation dynamique avec arrêt de la procédure quand le score devient trop faible.
- Sélection du meilleur alignement local.



L'ALGORITHME DE SMITH ET WATERMAN

Résultat de l'alignement

EERDAF
TSHEAL

$$S(i,j) = \max \begin{cases} se(i,j) + S(i+1,j+1) \\ se(i,j) + \max_{x,j+1} S(x,j+1) - P \\ se(i,j) + \max_{i+1,y} S(i+1,y) - P \\ 0 \end{cases} \quad \begin{matrix} \text{avec } i+2 < x \leq m \\ \text{et } j+2 < y \leq n \end{matrix} \quad (4)$$

| | V | T | E | E | R | D | A | F |
|---|----|----|----|----|----|----|----|----|
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |
| T | 0 | 3 | 0 | 0 | -1 | 0 | 1 | -3 |
| S | -1 | 1 | 0 | 0 | 0 | 0 | 1 | -3 |
| H | -2 | -1 | 1 | 1 | 2 | 1 | -1 | -2 |
| E | -2 | 0 | 4 | 4 | -1 | 3 | 0 | -5 |
| A | 0 | 1 | 0 | 0 | -2 | 0 | 2 | -4 |
| L | 2 | -2 | -3 | -3 | -3 | -4 | -2 | 2 |

a) Matrice initiale obtenue à partir de la matrice de substitution utilisée pour l'alignement (ici la matrice PAM250 de Dayhoff)

| | V | T | E | E | R | D | A | F |
|---|---|---|---|---|---|---|---|---|
| L | 2 | | | | | | | 2 |
| T | 0 | 3 | 0 | 0 | | 0 | 1 | |
| S | | 1 | 0 | 0 | 0 | 0 | 1 | |
| H | | | 1 | 1 | 2 | 1 | | |
| E | | 0 | 4 | 4 | | 3 | 0 | |
| A | 0 | 1 | 0 | 0 | | 0 | 2 | |
| L | 2 | | | | | | | 2 |

b) Matrice initiale où sont représentés uniquement les scores positifs ou nuls. C'est à dire toutes les positions susceptibles d'être un premier point de départ pour la transformation de la matrice initiale.

| | V | T | E | E | R | D | A | F |
|---|---|---|---|---|---|---|---|---|
| L | 8 | 7 | 0 | 0 | 0 | 0 | 0 | 2 |
| T | 6 | 6 | 9 | 3 | 0 | 0 | 1 | 0 |
| S | 2 | 6 | 3 | 9 | 1 | 0 | 1 | 0 |
| H | 0 | 3 | 5 | 2 | 9 | 1 | 0 | 0 |
| E | 0 | 0 | 4 | 4 | 0 | 7 | 0 | 0 |
| A | 0 | 1 | 0 | 0 | 0 | 0 | 4 | 0 |
| L | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |

c) matrice transformée construite à partir de la matrice initiale. L'expression 4 décrite dans le texte est utilisée pour le calcul de chaque score somme avec une pénalité fixe de 6.

| | V | T | E | E | R | D | A | F |
|---|---|---|---|---|---|---|---|---|
| L | 8 | 7 | 0 | 0 | 0 | 0 | 0 | 2 |
| T | 6 | 6 | 9 | 3 | 0 | 0 | 1 | 0 |
| S | 2 | 6 | 3 | 9 | 1 | 0 | 1 | 0 |
| H | 0 | 3 | 5 | 2 | 9 | 1 | 0 | 0 |
| E | 0 | 0 | 4 | 4 | 0 | 7 | 0 | 0 |
| A | 0 | 1 | 0 | 0 | 0 | 0 | 4 | 0 |
| L | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |

d) construction du chemin qui correspond à l'alignement local optimal. Cet alignement local débute à la position où se trouve le score maximum de la matrice transformée, c'est-à-dire le score 9.



Master 1 MABS

UE Bioinformatique des séquences

novembre - décembre 2013

