Pertinence des métriques fractionnaires pour l'analyse des signatures génomiques (données de grande dimension)

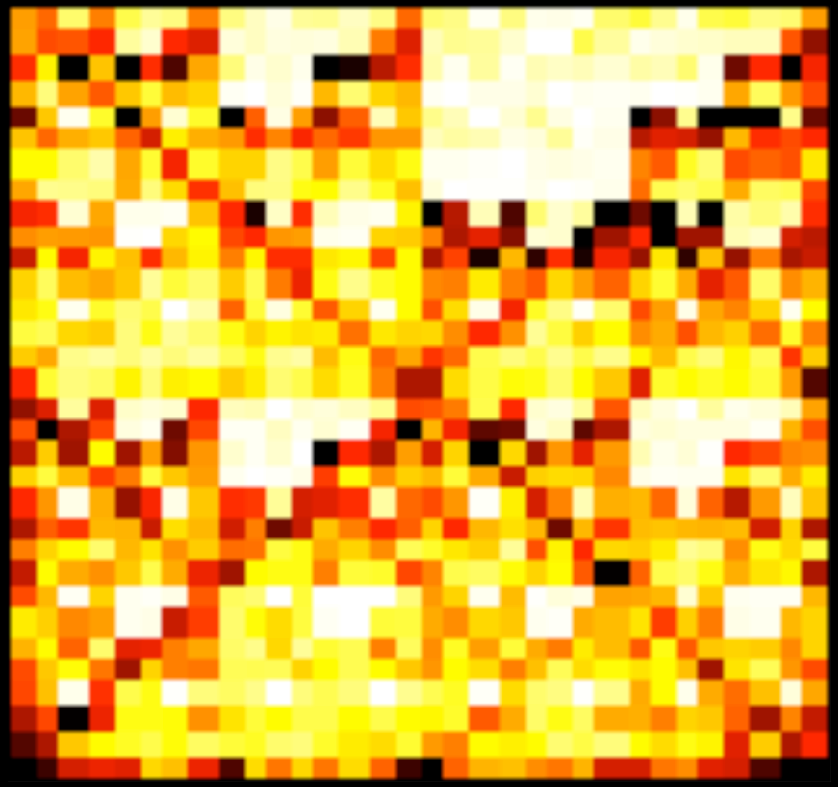
Sylvain LESPINATS INSERM U494

Plan

- Signature génomique / locale
- Visualisation des signatures locales
- Classification des signatures locales

Analyse des séquences d'ADN

Les signatures génomiques



L'ADN en tant que texte

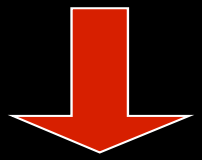
ACGCTCTTAGCGCAATCGATCGCATGCAGCTACACATCTCGATCGCCCAAATTCT

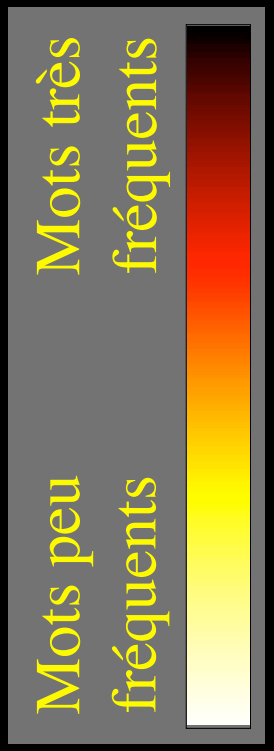
…

Alphabet : {C,G,A,T}

Fréquences des nucléotides

ACGCTCTTAGCGCAATCGATCGCATGCAGCTACACATCTCGATCGCCCAAATTCT





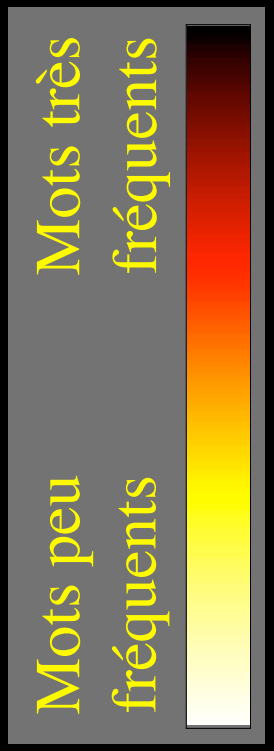
C : 12 % G : 22 %

A : 29 % T : 37 %

Fréquences des mots

ACGCTCTTAGCGCAATCGATCGCATGCAGCTACACATCTCGATCGCCCAAATTCT

mot de 2 lettres



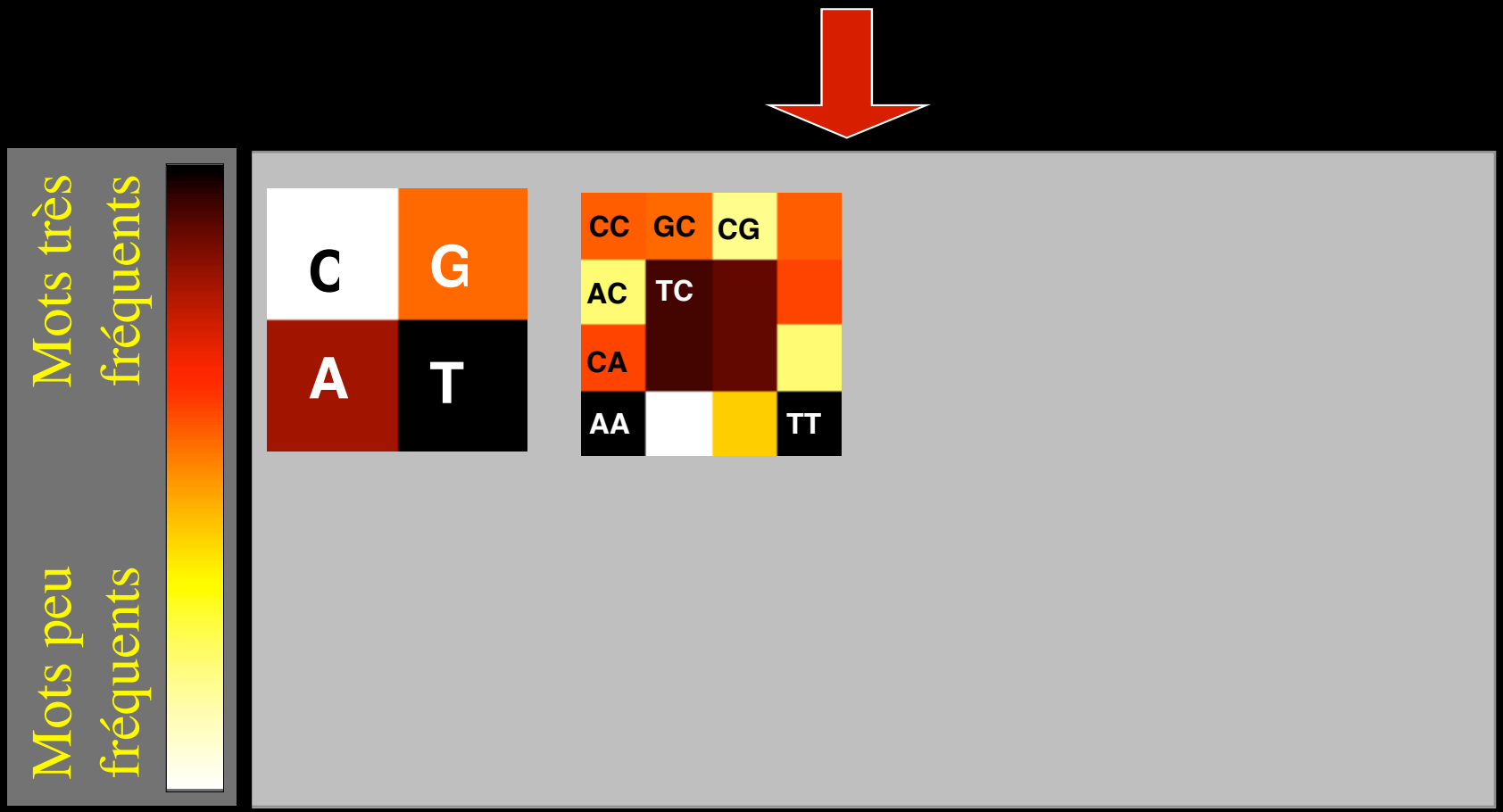
CC	AC	CA	AA
GC	TC	GA	TA
CG	AG	CT	AT
GG	TG	GT	TT

16 mots possibles

Fréquences des mots

ACGCTCTTAGCGCAATCGATCGCATGCAGCTACACATCTCGATCGCCCAAATTCT

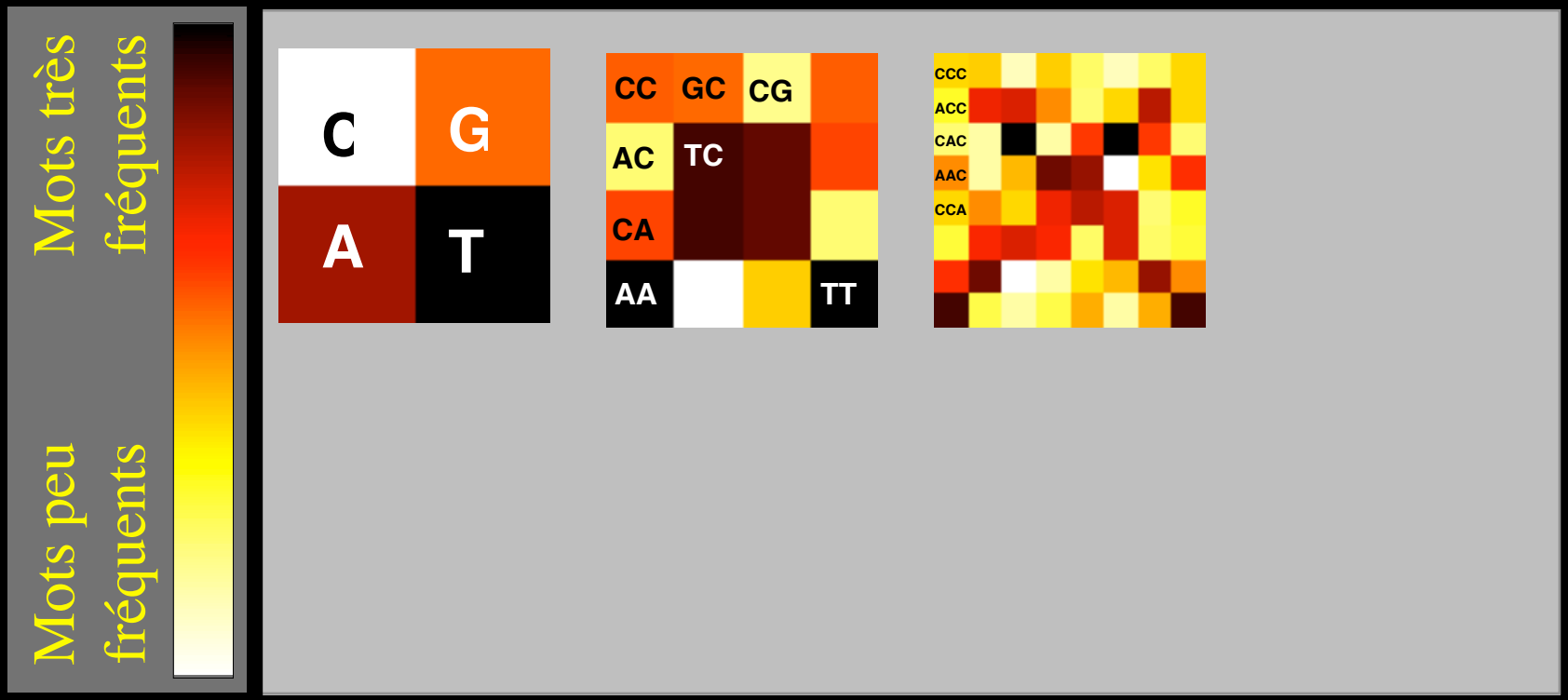
mot de 2 lettres



Fréquences des mots

ACGCTCTTAGCGCAATCGATCGCATGCAGCTACACATCTCGATCGCCCAAATTCT

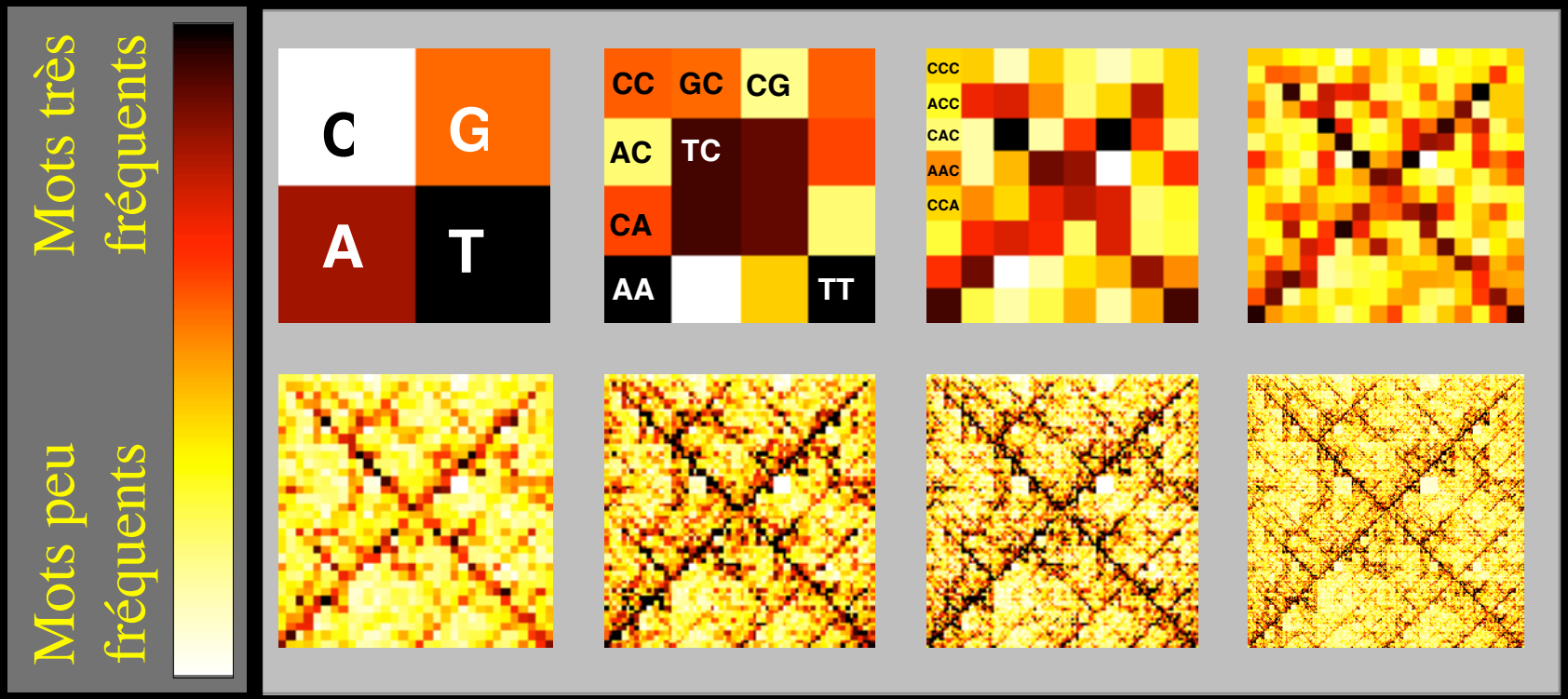
mot de 2 lettres 3 lettres



Fréquences des mots

ACGCTCTTAGCGCAATCGATCGCATGCAGCTACACATCTCGATCGCCCAAATTCT

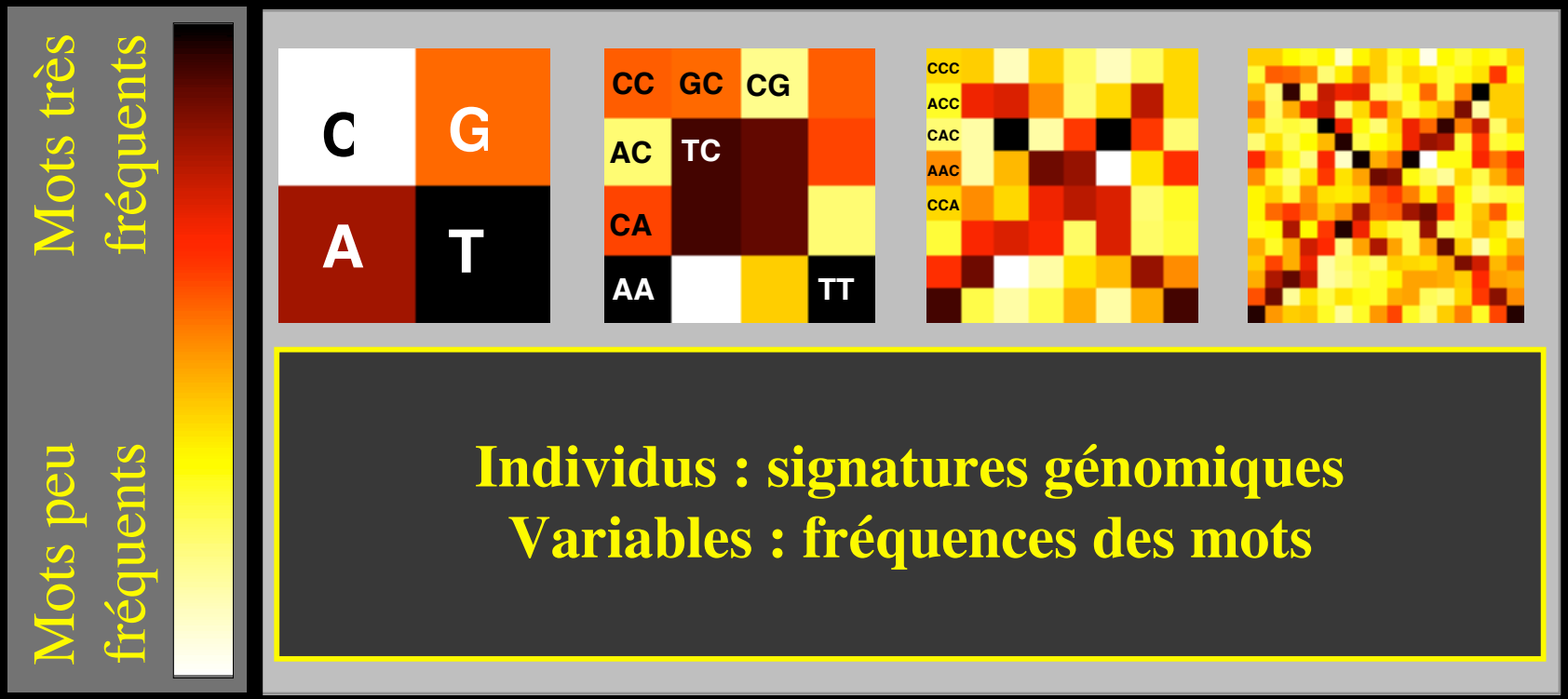
mot de 2 lettres 3 lettres 4 lettres mot de 6 lettres



Fréquences des mots

ACGCTCTTAGCGCAATCGATCGCATGCAGCTACACATCTCGATCGCCCAAATTCT

mot de 2 lettres 3 lettres 4 lettres mot de 6 lettres

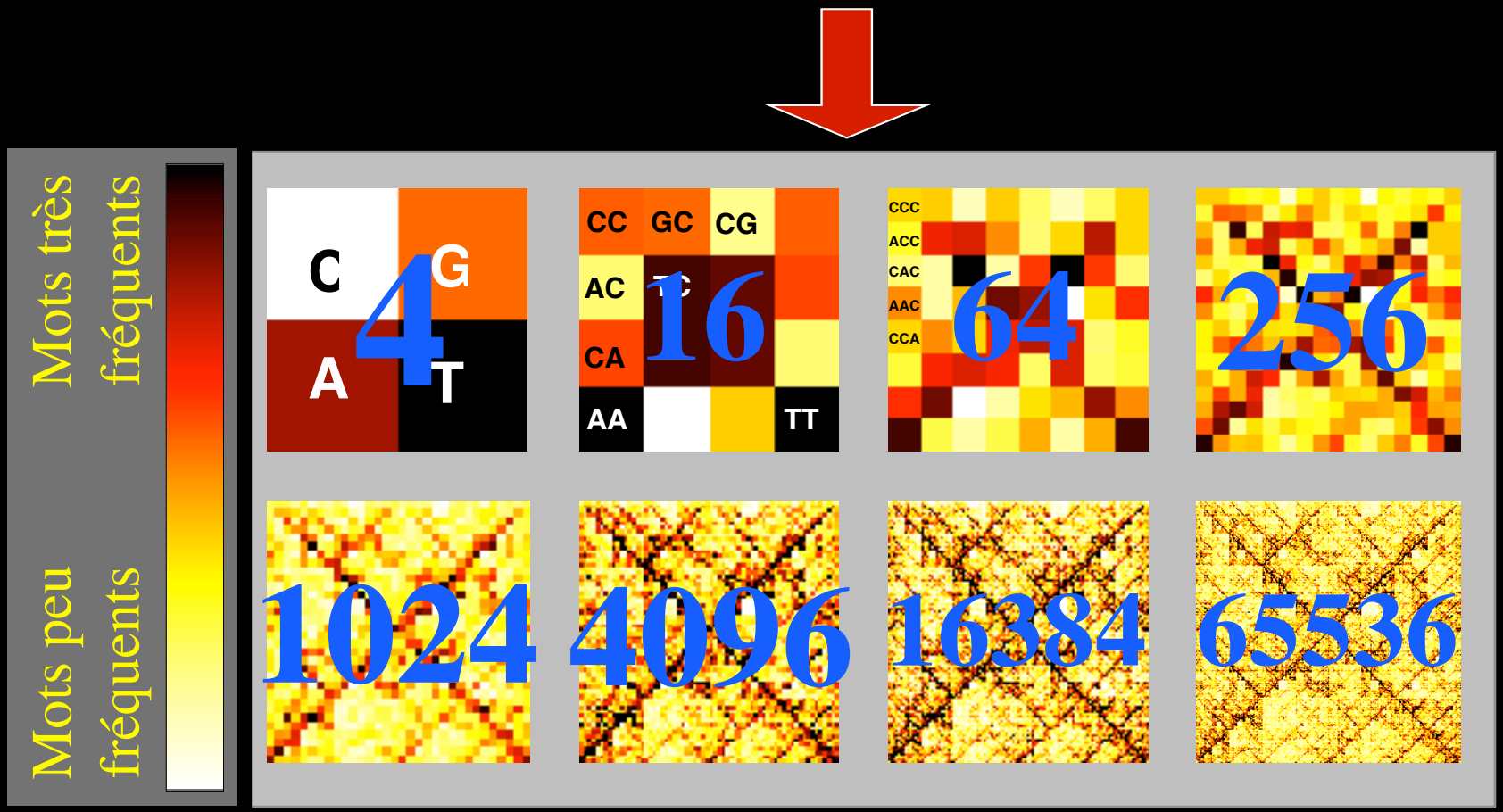


Individus : signatures génomiques
Variables : fréquences des mots

Grande dimension des signatures

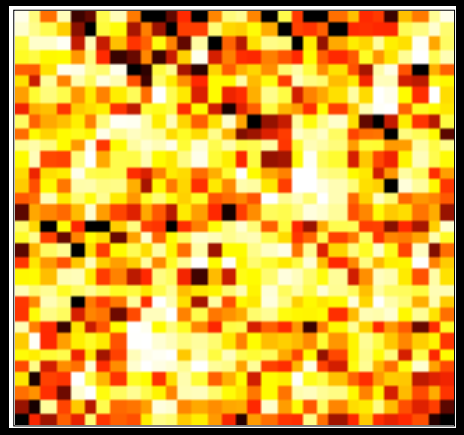
ACGCTCTTAGCGCAATCGATCGCATGCAGCTACACATCTCGATCGCCCAAATTCT

mot de 2 lettres 3 lettres 4 lettres mot de 6 lettres

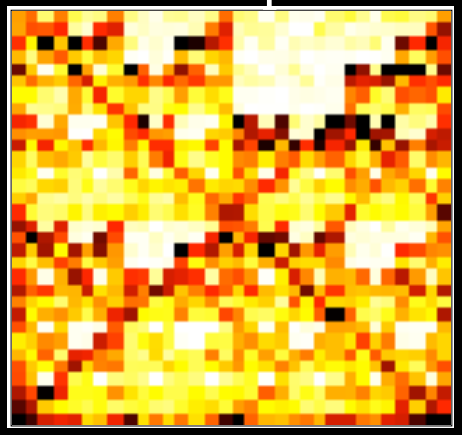


Diversité des signatures génomiques

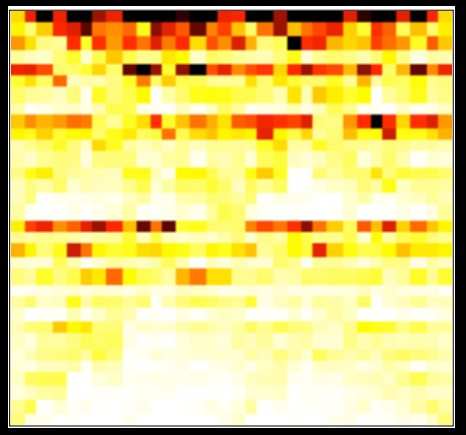
E. coli



Homo sapiens

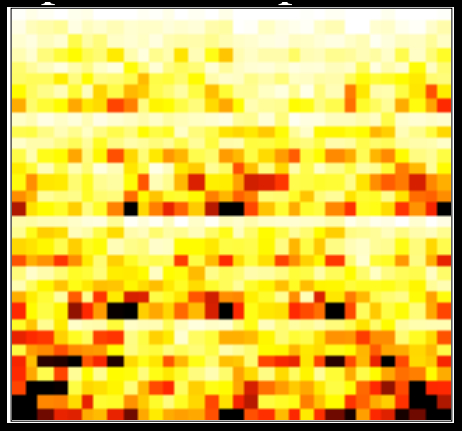


Deinococcus

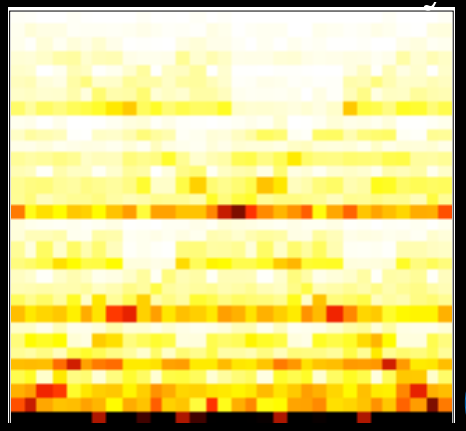


Pyrococcus

Streptococcus



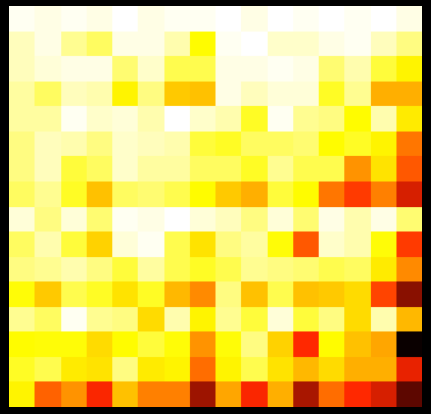
Clostridium



La signature génomique

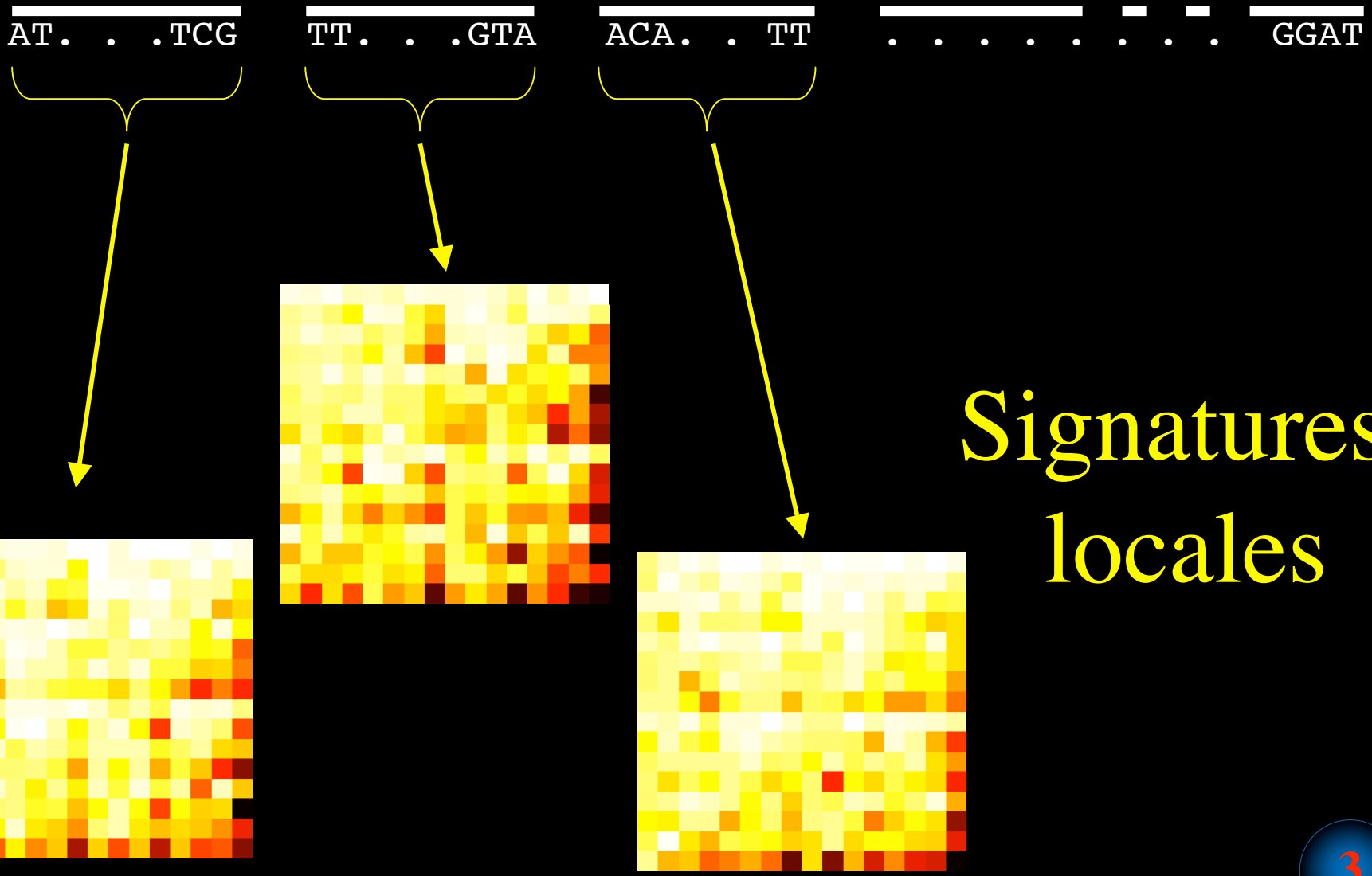
ATCTTTTTCGGCTTTTTTTTAGTATCCACAGAGGTT. . . . . GCATGTGGAT

Génome entier



Fréquences de
tous les mots
de longueur n

Les signatures locales

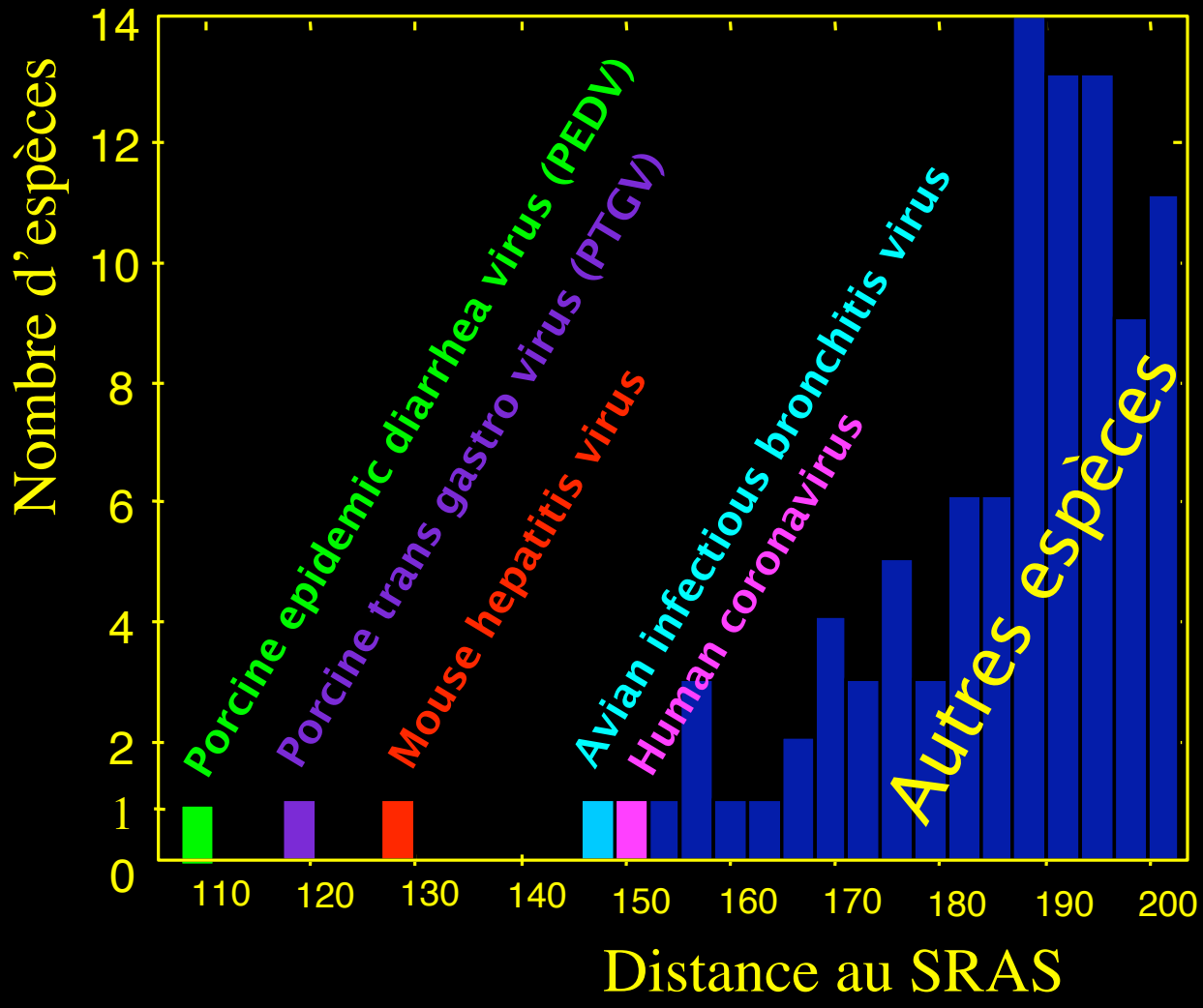


Une application de la signature génomique :

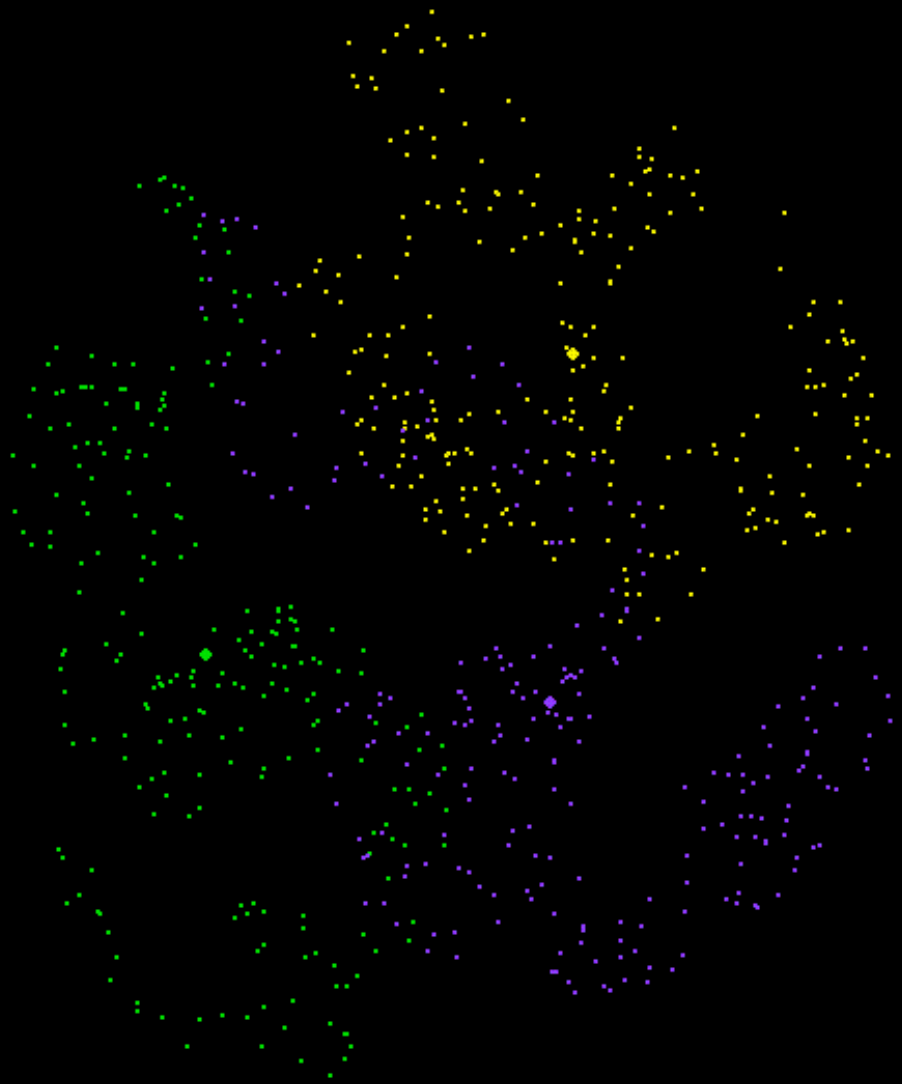
D'où vient le SRAS ?

Quelles sont les espèces dont la signature ressemble le plus à celle du SRAS ?

Histogramme des distances euclidiennes entre le SRAS et 5000 autres espèces



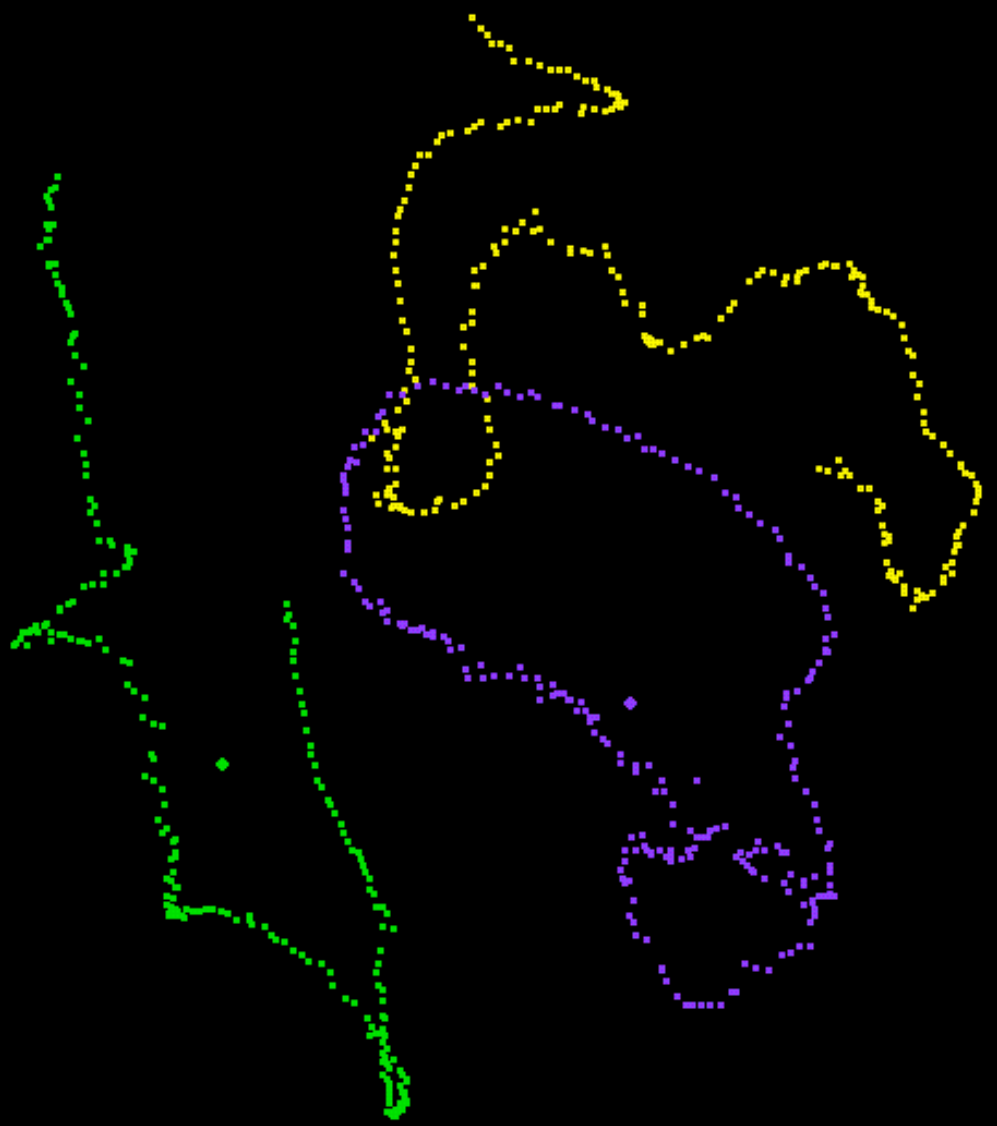
Les signatures les plus proches de celle du SRAS sont celles des coronavirus



Métrique
Euclidienne

ACC

SRAS
PEDV
PTGV



Métrique fractionnaire

$$d(x,y) = \sum_i \left| (x_i - y_i)^{0.6} \right|^{\frac{1}{0.6}}$$

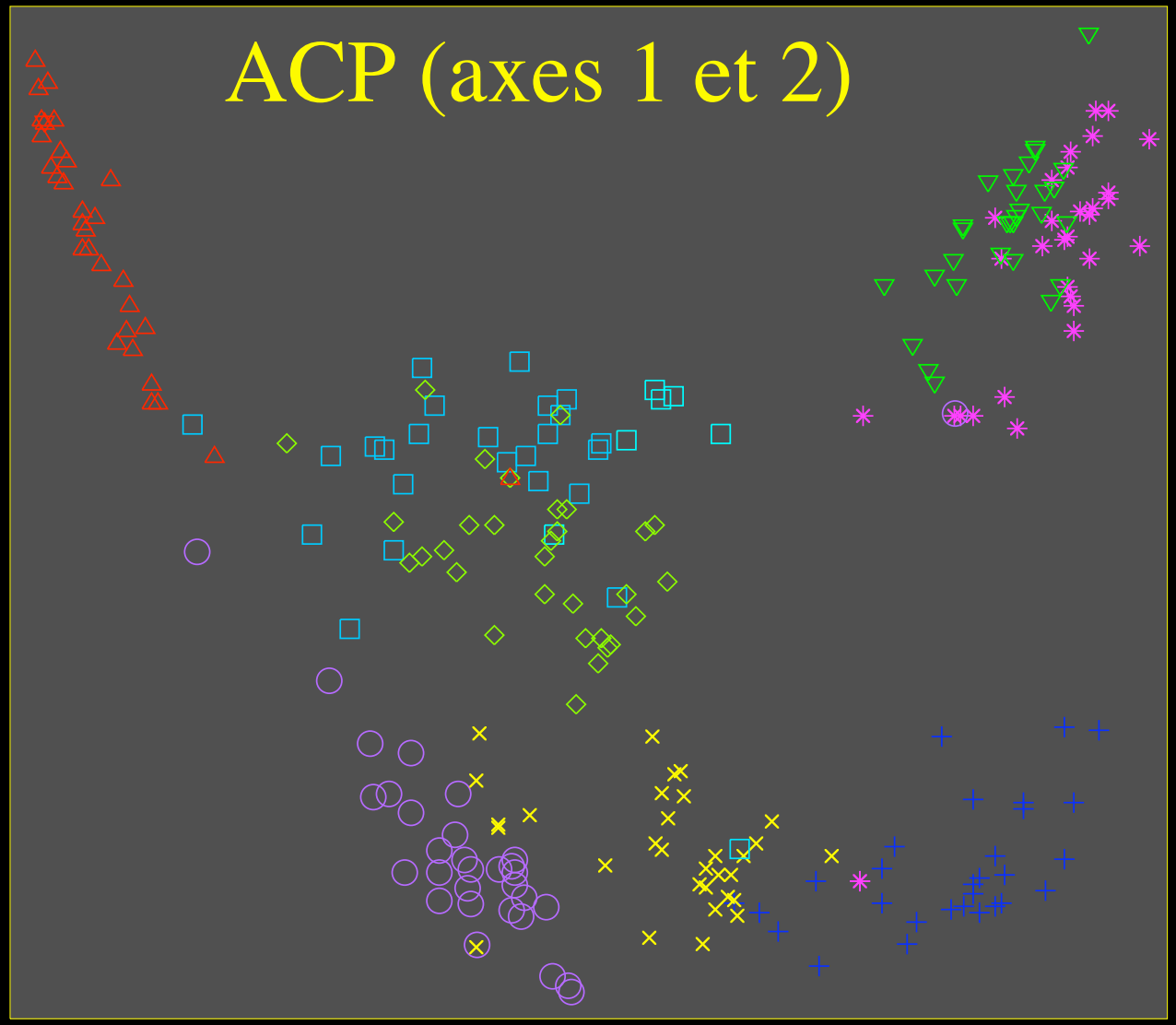
ACC

SRAS

PEDV

PTGV

Variations intra-espèce / inter-espèces

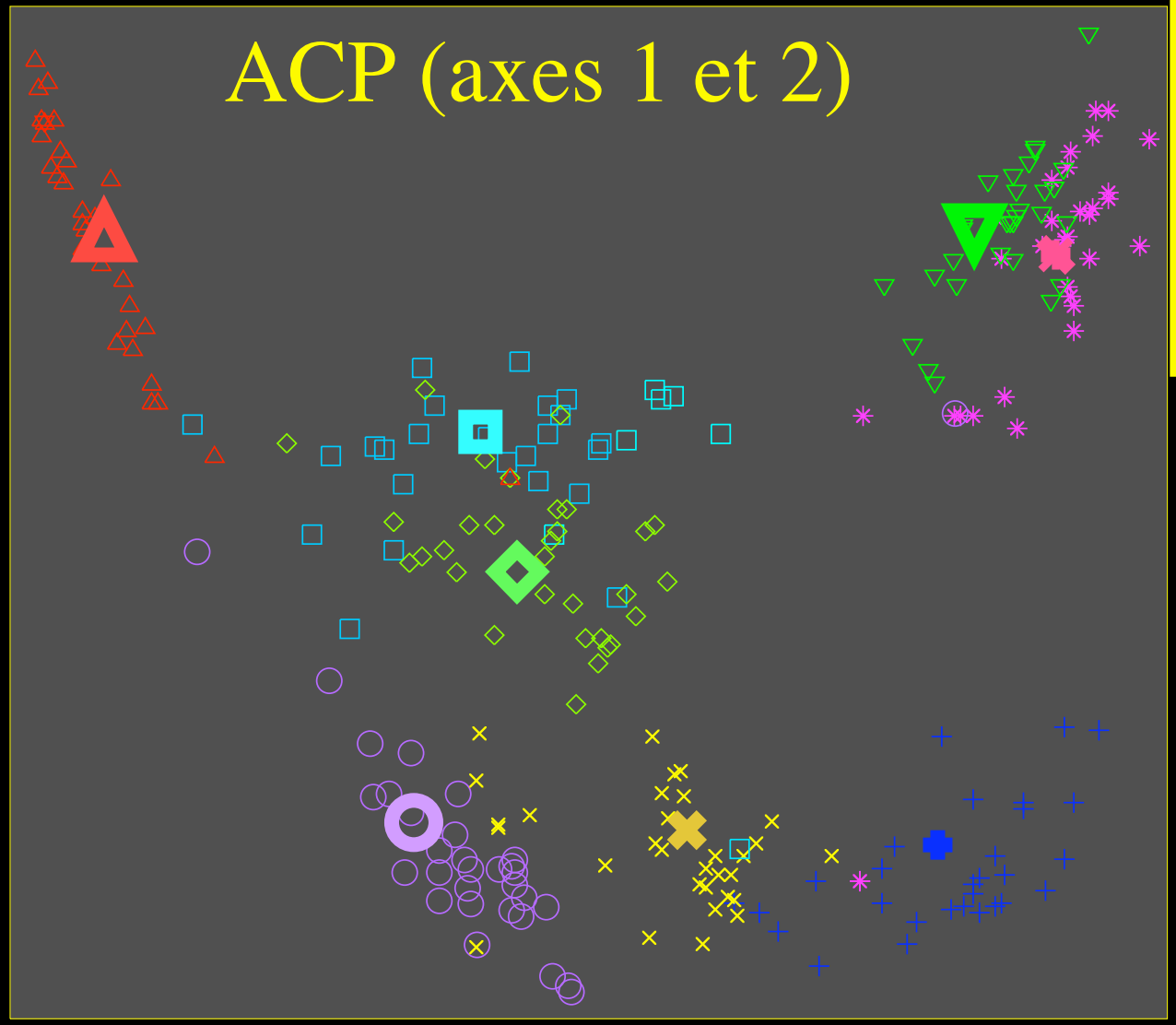


8 espèces :

- Aeropyrum pernix
- Agrobacterium tumefaciens
- Aquifex aeolicus
- Archaeoglobus fulgidus
- Bacillus subtilis
- Bacillus halodurans
- Borrelia burgdorferi
- Brucella melitensis

240 signatures locales

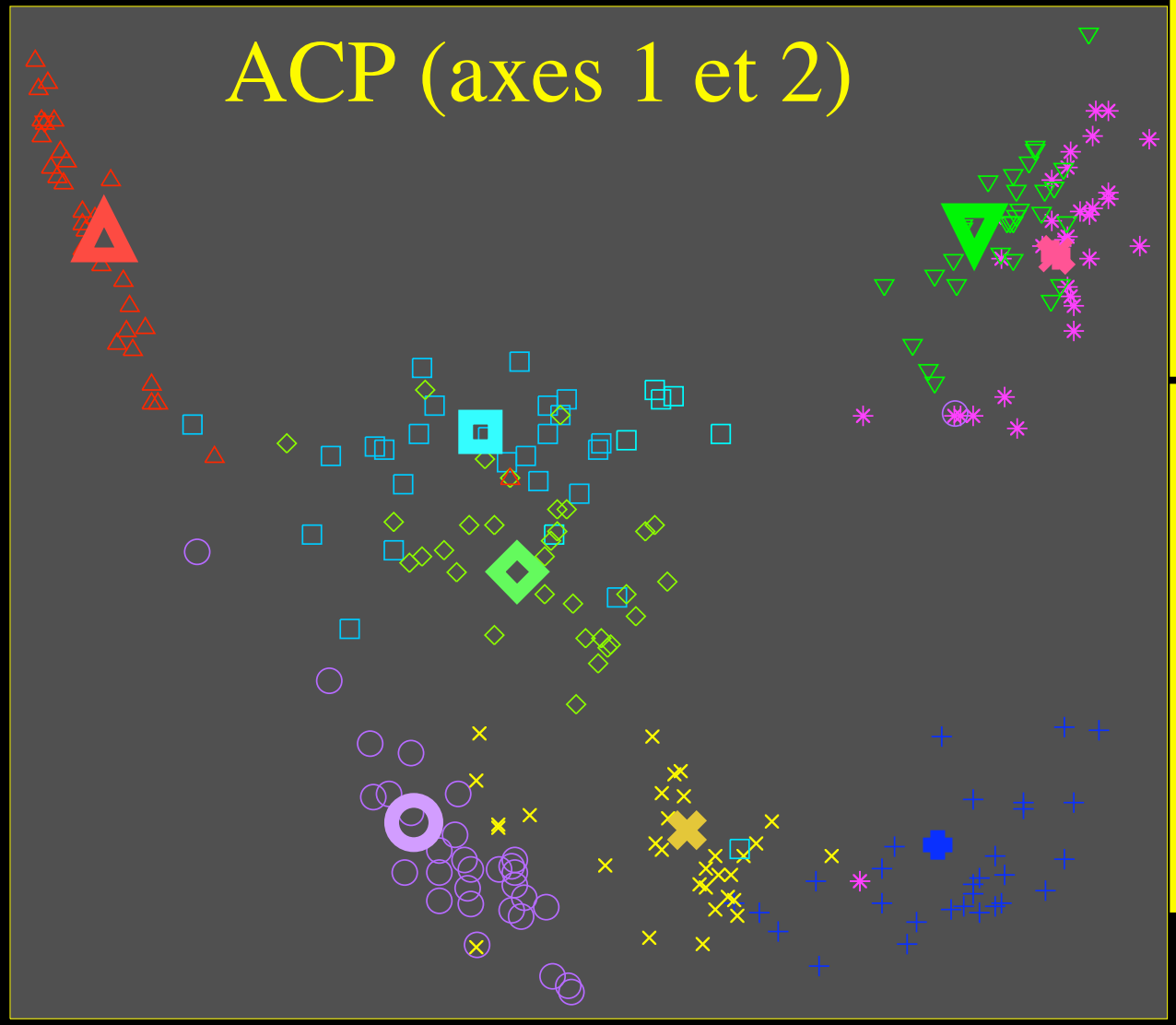
Variations intra-espèce / inter-espèces



Classification au plus proche voisin :
Individus : signatures locales
Centroïds : signatures des espèces

240 signatures locales

Variations intra-espèce / inter-espèces



Classification au plus proche voisin :
Individus : signatures locales
Centroïds : signatures des espèces

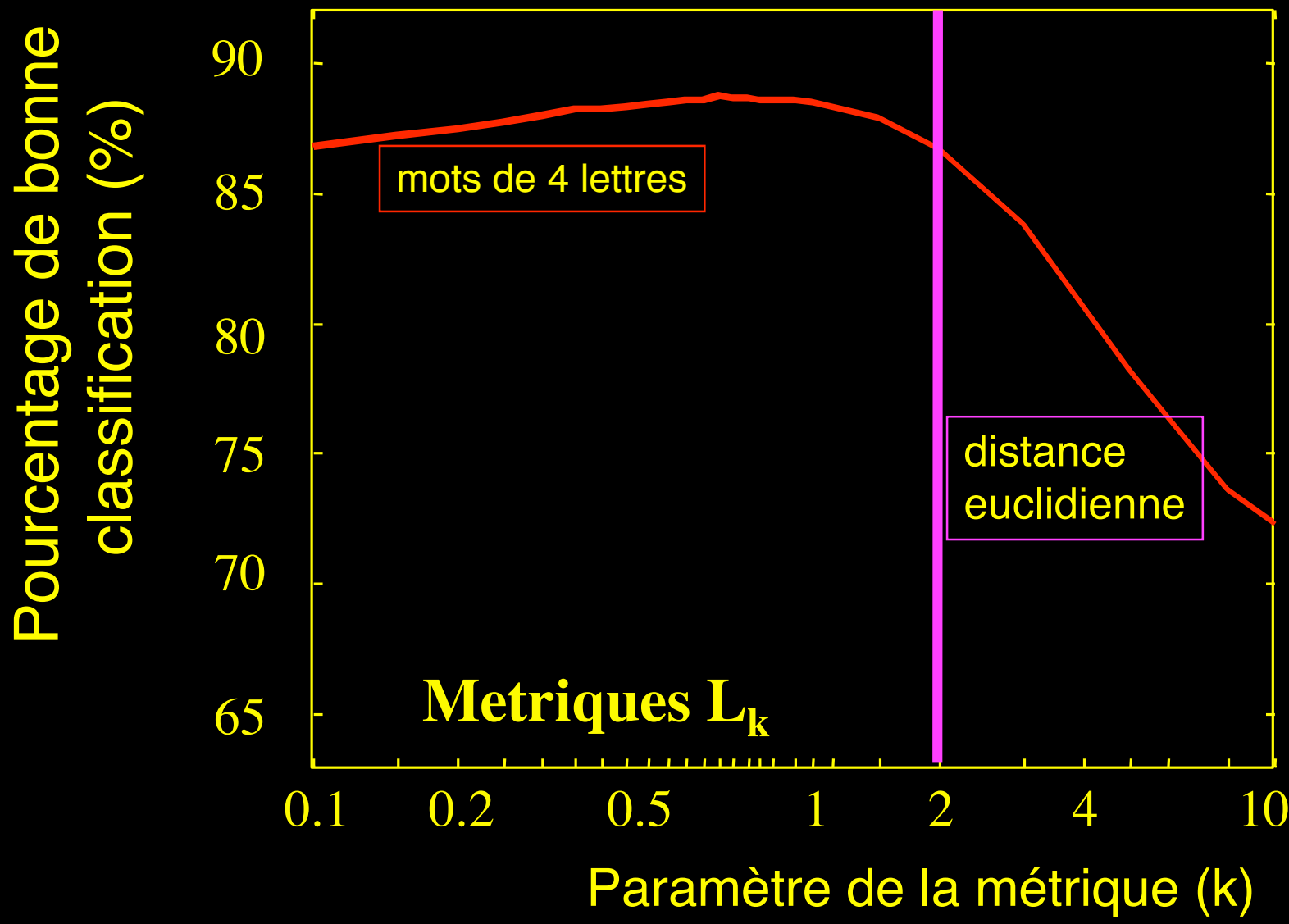
Erreurs d'affectation
- raisons biologiques
- problèmes de métrique

Métrique recherchée

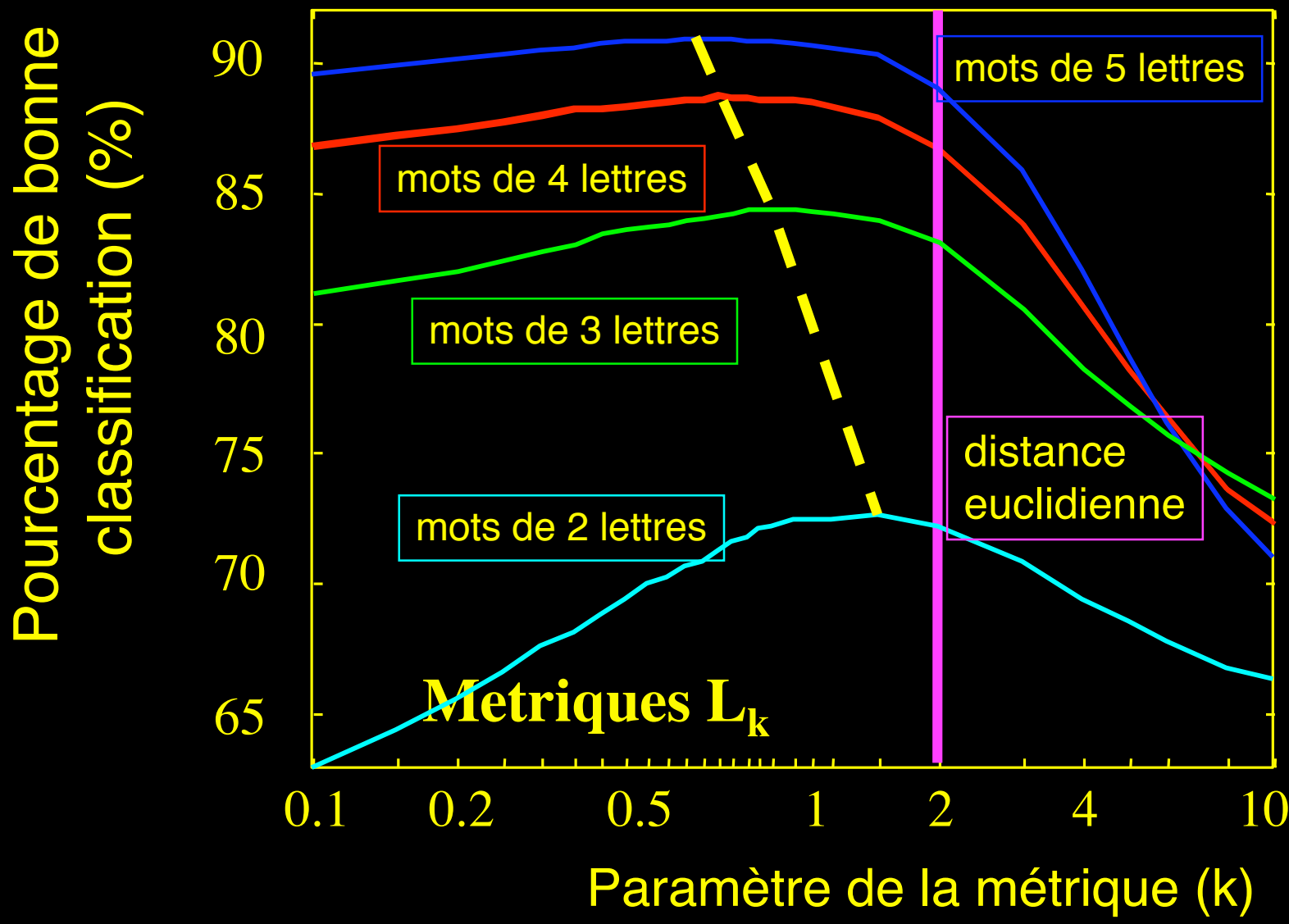
- minimise les différences entre les signatures locales d'une espèce
- maximise les différences entre les signatures des différentes espèces

Étude sur les génomes de 43 bactéries

Impact de la métrique



Impact de la métrique



Conclusion

- La choix de la métrique est important dans le cadre de l'étude des signatures génomiques.
- Plus le nombre de dimensions est grand, plus le paramètre de la métrique optimale est petit.