

## Distances et similarités

*Cours 5 – Modèles en biologie*

Christophe Ambroise

*L3 GBI – Université d'Évry*

## Matrices et transposée

Matrice  $A$  de taille  $n \times p$  : coefficient  $a_{ij}$  en ligne  $i$ , colonne  $j$ .

$$A\mathbf{x} = x_1 \mathbf{a}_1 + \cdots + x_p \mathbf{a}_p, \quad AB = (A\mathbf{b}_1 \cdots A\mathbf{b}_q)$$

Transposée :  $(A^T)_{ij} = a_{ji}$ ,  $(AB)^T = B^T A^T$ .

# Acquis des cours précédents

## Chaîne de Markov et vecteur propre

Évolution :  $\mathbf{x}_{t+1} = P \mathbf{x}_t$ ,  $P$  matrice de transition, colonnes de somme 1.

Équilibre :  $P\pi = \pi$ , soit  $(P - I)\pi = \mathbf{0}$  avec  $\sum_i \pi_i = 1$ .

Vecteur propre :  $A\mathbf{v} = \lambda \mathbf{v}$ ,  $\mathbf{v} \neq \mathbf{0}$ . L'équilibre  $\pi$  : vecteur propre de  $P$  pour la valeur propre 1.

## Norme, distance, produit scalaire

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \dots + x_p^2}, \quad d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^p x_i y_i = \mathbf{x}^T \mathbf{y}, \quad \mathbf{x} \cdot \mathbf{x} = \|\mathbf{x}\|^2$$

Pythagore :  $\mathbf{x} \perp \mathbf{y} \Rightarrow \|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$ .

## Similarité cosinus et centrage

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \in [-1, 1], \quad s(\mathbf{x}, \lambda \mathbf{y}) = s(\mathbf{x}, \mathbf{y}) \quad (\lambda > 0)$$

Cauchy-Schwarz :  $|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \|\mathbf{y}\|$ .

Centrage :  $\mathbf{x} - \bar{x} \mathbf{1}$ , écart de  $\mathbf{x}$  à sa projection sur  $\mathbf{1}$ , orthogonal à  $\mathbf{1}$ .

## La question des plus proches

### Cinq espèces, cinq séquences

- A GTACGAACTCCACCGTCGTT
- B GTACGACCTCGACCGTCGAT
- C GTACGACCACGACCGTGGAT
- D GCATGACCTTCATCTTCGAT
- E GGATGACCTTCATTTGCCAT

Quelles sont les deux espèces les plus proches ?

Séquences illustratives de 20 bases.

## Trois représentations d'un organisme

**Des mesures** : un vecteur de  $\mathbb{R}^p$  (masse, métabolisme, ...).

**Une séquence** : un mot sur l'alphabet  $\{A, C, G, T\}$ .

**Un ensemble** : les gènes exprimés, les espèces observées sur un site.

### Représentation

À chaque représentation, sa notion de distance.

## Distance

Une fonction positive de deux objets, nulle seulement pour deux objets identiques, symétrique, et qui vérifie l'inégalité triangulaire.

$$\begin{aligned}d(x, y) &\geq 0, & d(x, y) = 0 &\iff x = y \\d(x, y) &= d(y, x) \\d(x, z) &\leq d(x, y) + d(y, z)\end{aligned}$$

Espace métrique  $(E, d)$  : un ensemble  $E$  muni d'une distance  $d$ .

# Distance

## Inégalité triangulaire euclidienne

La norme d'une somme est au plus la somme des normes.

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \text{pour tous } \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$$

Démonstration, par Cauchy-Schwarz :

$$\begin{aligned}\|\mathbf{x} + \mathbf{y}\|^2 &= \|\mathbf{x}\|^2 + 2\mathbf{x} \cdot \mathbf{y} + \|\mathbf{y}\|^2 \\&\leq \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2\end{aligned}$$

Avec  $\mathbf{x} - \mathbf{y}$  et  $\mathbf{y} - \mathbf{z}$  :  $d_2(\mathbf{x}, \mathbf{z}) \leq d_2(\mathbf{x}, \mathbf{y}) + d_2(\mathbf{y}, \mathbf{z})$ .  $\square$

## Un écart qui n'est pas une distance

$\delta(x, y) = (x - y)^2$  sur  $\mathbb{R}$  : positif, symétrique, nul seulement si  $x = y$ .

$$\delta(0, 2) = 4 > \delta(0, 1) + \delta(1, 2) = 1 + 1 = 2$$

### Inégalité triangulaire violée

Un détour par 1 raccourcit le trajet de 0 à 2.

## Rôle des axiomes

**Séparation** : distance nulle, objets identiques. Deux séquences différentes ne sont jamais confondues.

**Symétrie** : l'écart de  $x$  à  $y$  est celui de  $y$  à  $x$ .

**Inégalité triangulaire** : deux objets proches d'un même troisième restent proches l'un de l'autre.

Sans inégalité triangulaire : une **dissimilarité**, pas une distance.

## Trois géométries sur $\mathbb{R}^p$

### Distances $d_1, d_2, d_\infty$

Trois façons d'agréger les écarts coordonnée par coordonnée : leur somme, la racine de la somme de leurs carrés, leur maximum.

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^p |x_i - y_i|$$

$$d_2(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^p (x_i - y_i)^2 \right)^{1/2}$$

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq p} |x_i - y_i|$$

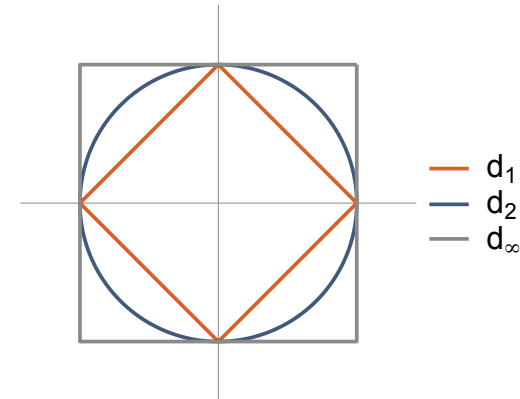
## Trois distances pour un même couple

$\mathbf{x} = (1, 3)$ ,  $\mathbf{y} = (4, -1)$  :  $\mathbf{x} - \mathbf{y} = (-3, 4)$ .

$$d_1 = 3 + 4 = 7, \quad d_2 = \sqrt{9 + 16} = 5, \quad d_\infty = \max(3, 4) = 4$$

Toujours :  $d_\infty \leq d_2 \leq d_1$ .

## Boules unités



Points à distance 1 de l'origine : losange, cercle, carré.

## Distance et géométrie

### Choix de la distance

Choisir une distance, c'est choisir une géométrie.

$d_1$  : somme des écarts, variable par variable.

$d_2$  : écart global, inchangé par rotation des axes.

$d_\infty$  : pire écart ; une seule variable suffit à éloigner deux individus.

## Unités et changement d'échelle

## Distances en unités brutes

| Animal    | Masse (kg) | Métabolisme (kcal/jour) |
|-----------|------------|-------------------------|
| Chèvre    | 36,0       | 800                     |
| Chimpanzé | 38,0       | 1 090                   |
| Mouton    | 46,4       | 1 254                   |

$d_2(\text{chimpanzé, chèvre}) = 290,0, \quad d_2(\text{chimpanzé, mouton}) = 164,2.$

Plus proche du chimpanzé : le **mouton**.

Kleiber, *Physiological Reviews*, 1947, tableau 2.

## Masse en grammes

Chèvre **36 000** g, chimpanzé **38 000** g, mouton **46 400** g.

$d_2(\text{chimpanzé, chèvre}) \approx 2\,021, \quad d_2(\text{chimpanzé, mouton}) \approx 8\,402.$

Plus proche du chimpanzé : la **chèvre**.

### Unités

Une convention d'unité décide de la réponse.

## Écart-type d'une variable

La dispersion d'une variable : norme de son vecteur centré, divisée par  $\sqrt{n}$ .

$$s_j = \frac{1}{\sqrt{n}} \|\mathbf{x}^j - \bar{x}_j \mathbf{1}\| = \left( \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right)^{1/2}, \quad \mathbf{x}^j : \text{colonne } j \text{ de } X$$

Même unité que la variable. Tableau de Kleiber (26 groupes) :  $s_{\text{masse}} \approx \mathbf{163}$  kg,  $s_{\text{métabolisme}} \approx \mathbf{2\,453}$  kcal/jour.

## Réduction des variables

Diviser chaque variable par son écart-type : des nombres sans unité.

$$x_{ij}^{\star} = \frac{x_{ij}}{s_j}$$

Grammes ou kilogrammes :  $s_j$  change dans la même proportion,  $x_{ij}^{\star}$  ne change pas.

Centrer ne change aucune distance :  $(\mathbf{x}_i - \mathbf{c}) - (\mathbf{x}_k - \mathbf{c}) = \mathbf{x}_i - \mathbf{x}_k$ .

Après réduction :  $d_2(\text{chimpanzé, chèvre}) \approx \mathbf{0,119}$ ,  
 $d_2(\text{chimpanzé, mouton}) \approx \mathbf{0,084}$ . Le **mouton**, quelle que soit l'unité.

# Objets non numériques

## Hamming, une distance

Positivité, séparation, symétrie : immédiates.

Inégalité triangulaire : là où  $u_i \neq w_i$ , forcément  $u_i \neq v_i$  ou  $v_i \neq w_i$ .

$$d_H(u, w) \leq d_H(u, v) + d_H(v, w)$$

Génotypes SNP du C1, codés 0, 1, 2 :  $d_1$  compte les allèles différents.

## Distance de Hamming

Pour deux séquences de même longueur : le nombre de positions où elles diffèrent.

$$d_H(u, v) = \#\{i : u_i \neq v_i\}, \quad u, v \in \{A, C, G, T\}^L$$

A GTACGA**ACTC**CACCGTCG**TT**  
B GTACGA**CCTC**GACCGTCG**AT**

$$d_H(A, B) = 3.$$

## Indice de Jaccard

Pour deux ensembles : la part des éléments communs parmi tous les éléments présents.

$$J(E, F) = \frac{|E \cap F|}{|E \cup F|}, \quad d_J(E, F) = 1 - J(E, F)$$

Gènes exprimés :  $E = \{g_1, g_2, g_3, g_4\}$ ,  $F = \{g_3, g_4, g_5\}$ .

$$|E \cap F| = 2, |E \cup F| = 5 : J = 0,4, d_J = 0,6.$$

## Doubles absences

Deux sites, dix espèces, présence (1) ou absence (0) :

$$\mathbf{x} = (1, 1, 0, 0, 0, 0, 0, 0, 1, 0), \quad \mathbf{y} = (1, 0, 0, 0, 0, 0, 0, 0, 1, 1)$$

Accord simple (part des positions identiques) :  $8/10 = 0,8$ . Jaccard :  $2/4 = 0,5$ .

Cent espèces de plus, absentes des deux sites : accord simple  $108/110 \approx 0,98$  ; Jaccard inchangé,  $0,5$ .

### Jaccard

Deux sites ne se ressemblent pas par les espèces qui y manquent.

## Similarité

Une similarité : grande pour des objets proches, maximale pour des objets identiques.

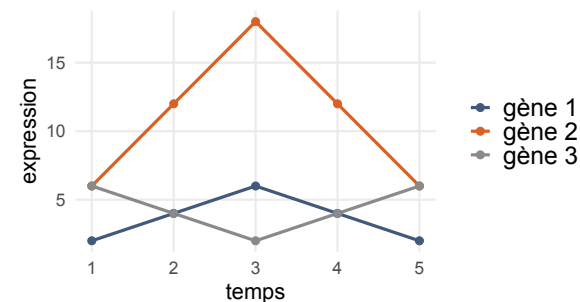
Depuis une distance :  $s = \frac{1}{1+d}$ , égale à 1 si  $d = 0$ , proche de 0 pour  $d$  grand.

Jaccard :  $J = 1 - d_J$ .

Similarité cosinus : directement une similarité, dans  $[-1, 1]$ .

# Similarités et corrélation

## Profils d'expression



Euclidienne : gène 1 plus proche du gène 3 (6,9 contre 17,4). Cosinus : plus proche du gène 2 (1 contre 0,75).

Profils illustratifs : gène 2 =  $3 \times$  gène 1.

## Profils décalés

$\mathbf{x} = (1, 2, 3)$ ,  $\mathbf{y} = (11, 12, 13) = \mathbf{x} + 10 \mathbf{1}$  : même forme.

$$s(\mathbf{x}, \mathbf{y}) = \frac{74}{\sqrt{14} \sqrt{434}} \approx 0,95$$

Cosinus : insensible à un facteur d'échelle, sensible à un décalage.

Remède : retirer le niveau moyen, c'est-à-dire **centrer**.

## Corrélation de profils

$\mathbf{x} - \bar{x} \mathbf{1} = (-1, 0, 1) = \mathbf{y} - \bar{y} \mathbf{1}$  :  $r(\mathbf{x}, \mathbf{y}) = 1$ .

Gènes 1 et 2 :  $r = 1$ . Gènes 1 et 3 :  $r = -1$ , profils opposés.

### Forme

Corrélation : insensible au niveau et à l'échelle. Seule compte la forme.

## Corrélation

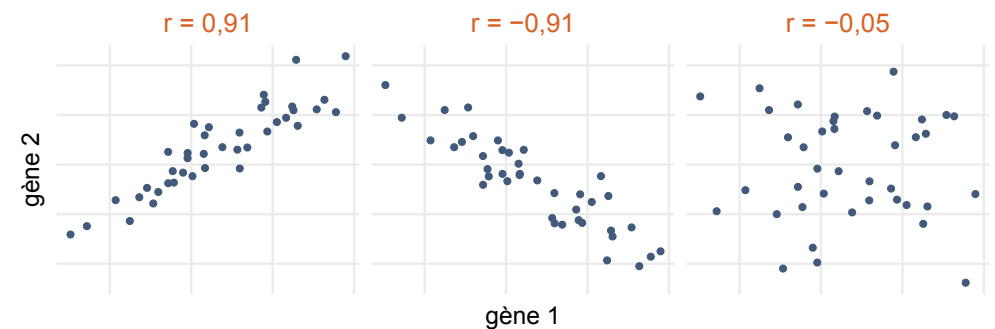
Le cosinus des deux vecteurs centrés.

$$r(\mathbf{x}, \mathbf{y}) = s(\mathbf{x} - \bar{x} \mathbf{1}, \mathbf{y} - \bar{y} \mathbf{1}) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

Variables non constantes : vecteurs centrés non nuls.

Cauchy-Schwarz :  $-1 \leq r \leq 1$  ;  $r = \pm 1$  pour des vecteurs centrés colinéaires.

## Co-expression de gènes



Un point : un échantillon ; en abscisse et en ordonnée, l'expression de deux gènes.

Données simulées.

# Matrice des distances

## Matrice des distances

Pour  $n$  objets : la matrice  $n \times n$  des distances deux à deux, symétrique et de diagonale nulle.

$$D = (d(x_i, x_j))_{1 \leq i, j \leq n}, \quad d_{ii} = 0, \quad D^T = D$$

$D^T = D$  : matrice **symétrique**, conséquence de l'axiome de symétrie.

## Représentation et distance

### Message central

Objet biologique  $\rightarrow$  représentation  $\rightarrow$  distance. Il n'existe pas « la » distance entre deux objets biologiques.

Mesures :  $d_2$  après réduction.      Séquences : Hamming.

Ensembles : Jaccard.      Profils : corrélation.

## Les deux espèces les plus proches

|   | A  | B | C  | D | E  |
|---|----|---|----|---|----|
| A | 0  | 3 | 5  | 7 | 10 |
| B | 3  | 0 | 2  | 6 | 9  |
| C | 5  | 2 | 0  | 8 | 11 |
| D | 7  | 6 | 8  | 0 | 4  |
| E | 10 | 9 | 11 | 4 | 0  |

### Réponse

B et C : 2 différences sur 20, le plus petit terme hors diagonale.