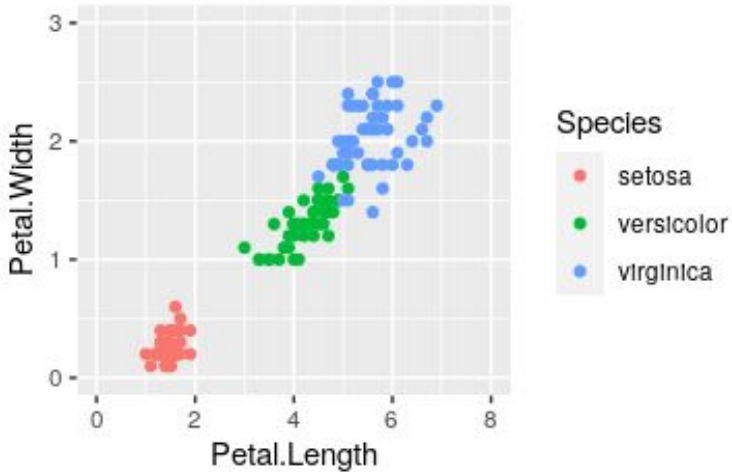
min-max

transformation depuis l'intervalle [min, max] vers l'intervalle [a,b]

$$v' = \frac{v - \min}{\max - \min}(b - a) + a$$



z-score

$$v' = \frac{v - \mu}{\sigma}$$

mise à l'échelle décimale

$$v' = \frac{v}{10^n}$$ avec *n* le plus petit entier tel que max(|*v'*|)<1

Normaliser les données : s'affranchir des unités de mesures

Ecart absolu à la moyenne (voire à la médiane)

$$s = \frac{|x_1 - \mu| + |x_2 - \mu| + \cdots + |x_n - \mu|}{n}$$

Calculer la mesure normalisée (z-score)

$$z_i = \frac{x_i - \mu}{s}$$

L'écart absolu est plus robuste que celle de l'écart type

- qualitatives ou nominales :
  - qualitatives
  - binaires, logiques
  - énumérations, facteurs
- numériques :
  - quantitatives
    - discrètes : entiers
    - continues
      - continues sur un intervalle
      - échelle linéaire, logarithmique, exponentielle
    - nombres complexes
  - ordinales, temporelles
  - géométriques, spatiales

.textuelles, sémantiques, ontologies

.mixtes

- table de contingence

		Objet *j* = 1	Objet *j* = 0
Objet *i*	1	*a*	*b*
	0	*c*	*d*

- coefficient simple d'appariement (invariant, si la variable est symétrique)

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- coefficient de Jaccard (non invariant, si la variable est asymétrique)

$$d(i,j) = \frac{b+c}{a+b+c}$$

- Exemple

Nom	Sexe	Fièvre	Tousse	Test-1	Test-2	Test-3	Test-4
Jacques	M	O	N	P	N	N	N
Marie	F	O	N	P	N	P	N
Jean	M	O	P	N	N	N	N

- sexe est symétrique
- les autres sont asymétriques
- soit O et P = 1, et N = 0

$$d(jacques, marie) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jacques, jean) = \frac{1+1}{1+1+1} = 0.66$$

$$d(jean, marie) = \frac{1+2}{1+1+2} = 0.75$$

- généralisation des valeurs binaires : plus de 2 états
- méthode 1 : appariement simple
  - *m* : nombre d'appariements, *p* : nombre total de variables

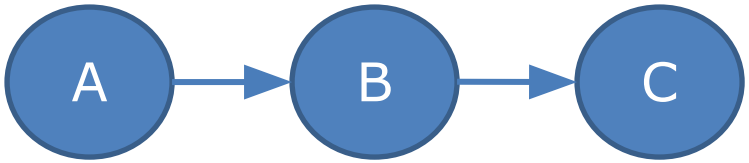
$$d(i,j) = \frac{p-m}{p}$$

- méthode 2 : utiliser un grand nombre de variables binaires
  - création d'une variable binaire pour chacun des états d'une variable nominale
- Information mutuelle

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} \log \frac{p(x,y)}{p(x)p(y)}$$

Information mutuelle

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} \log \frac{p(x,y)}{p(x)p(y)}$$



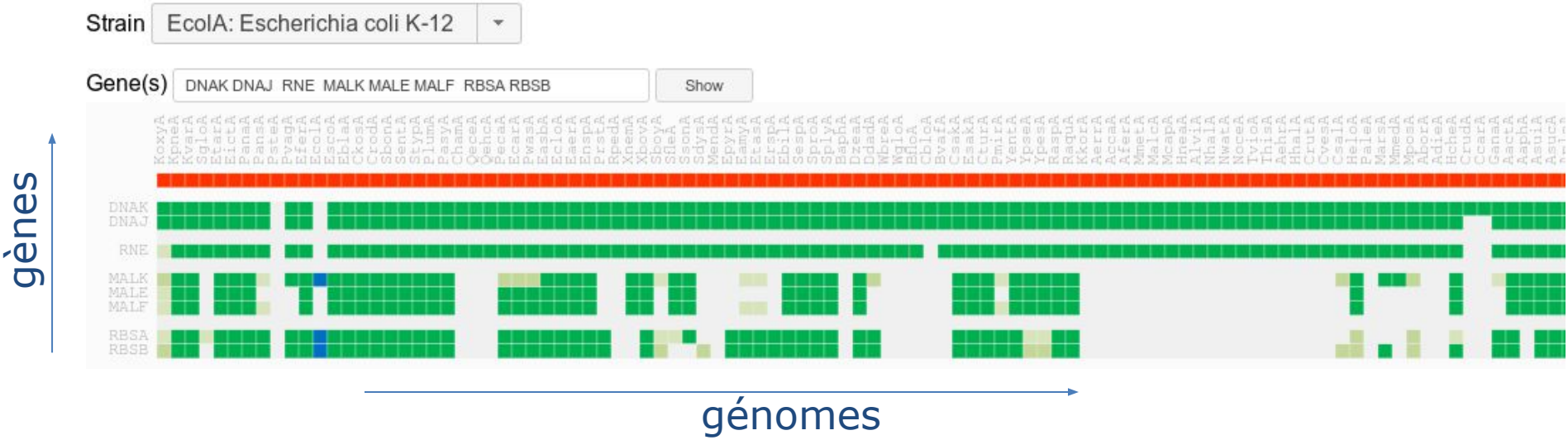
observations

	cond 1	cond 2	cond 3	cond 4	cond 5
A	high	low	high	low	low
B	high	low	high	high	low

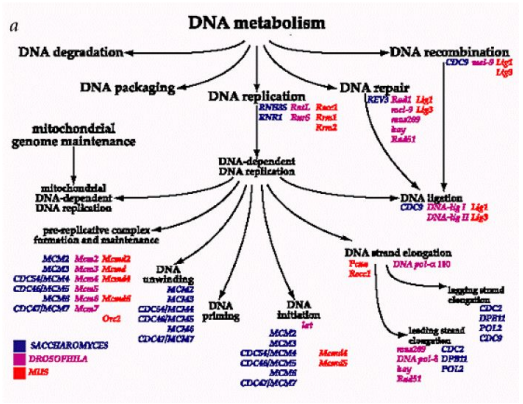
	b = low	B = high	total
a = low	2 / 5	1 / 5	3 / 5
A = high	0	2 / 5	2 / 5
total	2 / 5	3 / 5	1

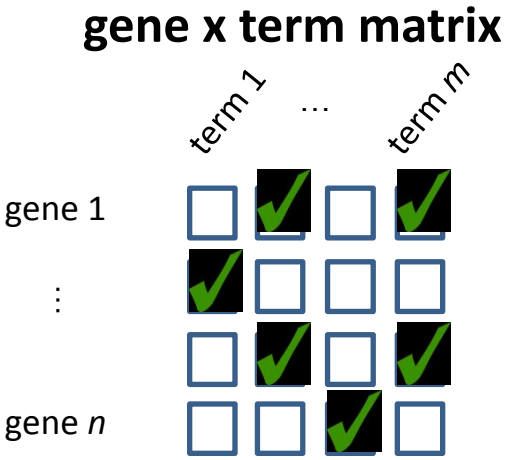
```
   ab      ab    a  b      aB      aB    a  B     Ab    AB     AB   A   B
> .4* log( .4 / (.6*.4) ) + .2*log( .2 / (.6*.6) ) + 0 + .4*log( .4 / (.4*.6) )
[1] 0.2911032
```

- Profils phylogénétiques



- Annotations, ex: Gene Ontology



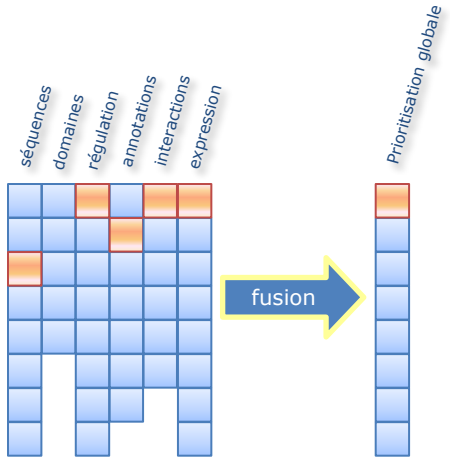


- l'ordre est important : rang
- peut être traitée comme une variable continue sur un intervalle
  - remplace x_{if} par son rang
  - transforme chaque variable sur [0,1] en remplaçant le i-ème objet de la f-ème variable

$$r_{if} = \{1, \ldots, M_f\}$$

  - calcule la dissimilarité en utilisant les méthodes de valeurs continues sur un intervalle

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$



Valeurs continues sur un intervalle, Fonction de distance

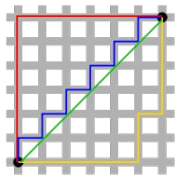
- Distance de Minkowski :

$$d(i,j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q}$$

avec $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ et $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ deux objets à p dimensions, et q un entier positif

- si $q = 1$: distance de Manhattan (ou city block distance)

$$d(i,j) = |x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{ip}-x_{jp}|$$



source : wikipedia

- si $q = 2$:distance euclidienne

$$d(i,j) = \sqrt{(x_{i1} - x_{j1}^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

- Propriétés
  - $d(i,i) = 0$
  - $d(i,j) \geq 0$ (positive)
  - $d(i,j) = d(j,i)$ (symétrique)
  - $d(i,j) \leq d(i,k) + d(k,j)$ (inégalité triangulaire)

- Dissimilarité basée sur un coefficient de corrélation
  - Pearson, Spearman (rangs)
  - d(x,y) = 1 – corr(x,y)

- Distance de Canberra (~ Manhattan pondérée)

$$d(p,q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

- Similarité = cosinus de l'angle formé par les 2 vecteurs

- Distance de Mahalanobis
  - distance d'un point à un ensemble
  - *x* : vecteur/point
  - *S* : matrice de variance-covariance

$$d(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}$$

- mesure positive sur une échelle non linéaire, échelle exponentielle qui suit approximativement Ae^{BT} ou Ae^{-BT}
- Méthodes
  - les traiter comme des variables continues sur un intervalles : mauvais choix
  - appliquer une transformation logarithmique puis les traiter comme des variables continues sur un intervalle

$$y_{if} = log\ x_{if}$$

  - les traiter comme des variables ordinales en traitant leur rang

- Mots de même longueur
  - Distance de Hamming = Nombre de différences

- Séquences
  - Normalized bit score
  - distance PAM

- Les objets peuvent être décrits avec tous les types de données
  - binaire symétrique, binaire asymétrique, nominale, ordinale, …

- Utilisation d'une formule pondérée pour combiner leurs effets

$$d(i,j) = \frac{\sum_{k=1}^{p} w_k d_k(i,j)}{\sum_{k=1}^{p} w_k}$$