

Information et entropie

Cours 7 – Modèles en biologie

Christophe Ambroise

L3 GBI – Université d'Évry

Distances et matrice des distances

Distance : positive, séparante, symétrique, inégalité triangulaire.

Hamming : nombre de positions différentes entre deux séquences.

Matrice des distances D : symétrique, diagonale nulle.

Acquis des cours précédents

Écart-type, réduction, corrélation

$$s_j = \frac{1}{\sqrt{n}} \|\mathbf{x}^j - \bar{x}_j \mathbf{1}\|, \quad x_{ij}^* = \frac{x_{ij}}{s_j}$$

Corrélation : cosinus des vecteurs centrés,

$$r = s(\mathbf{x} - \bar{x} \mathbf{1}, \mathbf{y} - \bar{y} \mathbf{1}) \in [-1, 1].$$

Inertie et variances

Centre de gravité $\mathbf{g} = \frac{1}{n} \sum_i \mathbf{x}_i$; inertie $I = \frac{1}{n} \sum_i \|\mathbf{x}_i - \mathbf{g}\|^2 = \sum_j s_j^2$.

Huygens : $I = W + B$, inertie intra-classe plus inertie inter-classe.

Mesurer l'incertitude

k-means et classification hiérarchique

k-means : affectation au centre le plus proche, recentrage ; W décroît jusqu'à stabilité.

Classification ascendante hiérarchique : fusions successives des deux classes les plus proches, dendrogramme.

Positions conservées, positions variables

1	2	3	4	5	6	7	8
A	T	A	T	A	A	T	A
A	T	A	T	A	A	T	C
C	T	A	T	A	A	T	G
C	T	A	T	A	A	T	T
G	T	A	T	A	A	T	C
G	T	A	T	T	A	T	G
T	T	A	T	T	A	C	A
T	T	A	T	T	A	C	C
A	T	A	A	T	A	G	G
C	T	G	A	T	A	G	T

Dix sites de fixation alignés. Quelles positions renseignent sur le site ?

Surprise d'un événement

La surprise d'un événement de probabilité p : d'autant plus grande que p est petite, nulle pour un événement certain.

$$h(p) = -\log_2 p = \log_2 \frac{1}{p} \quad (\text{en bits})$$

$p = 1/2$: 1 bit. $p = 1/4$: 2 bits. $p = 1/1024$: 10 bits. $p = 1$: 0 bit.

Le bit

Un bit : la réponse à une question oui/non dont les deux réponses sont également probables.

Une base parmi quatre, équiprobables : deux questions.

« Purine (A ou G) ? », puis « A ? » : $2 \text{ bits} = -\log_2 \frac{1}{4}$.

Additivité de la surprise

Deux événements indépendants : probabilités multipliées, surprises additionnées.

$$h(pq) = -\log_2(pq) = -\log_2 p - \log_2 q = h(p) + h(q)$$

Deux bases indépendantes, chacune de probabilité $1/4$: probabilité $1/16$, surprise $2 + 2 = 4$ bits.

Le logarithme : la seule façon de transformer produits en sommes.

9

10

Entropie

11

12

Entropie

La surprise moyenne : l'incertitude d'une loi de probabilité.

$$H(\mathbf{p}) = - \sum_{i=1}^K p_i \log_2 p_i, \quad \mathbf{p} = (p_1, \dots, p_K), \, p_i \geq 0, \, \sum_i p_i = 1$$

Convention : $0 \log_2 0 = 0$.

$\mathbf{p}$: un vecteur de $\mathbb{R}^K$, la « loi » du C1.

Bornes de l'entropie

L'entropie est nulle pour un résultat certain et maximale pour la loi uniforme.

$$0 \leq H(\mathbf{p}) \leq \log_2 K,$$

$$\begin{aligned} H = 0 &\iff \text{une modalité certaine} \\ H = \log_2 K &\iff \mathbf{p} \text{ uniforme} \end{aligned}$$

Admis.

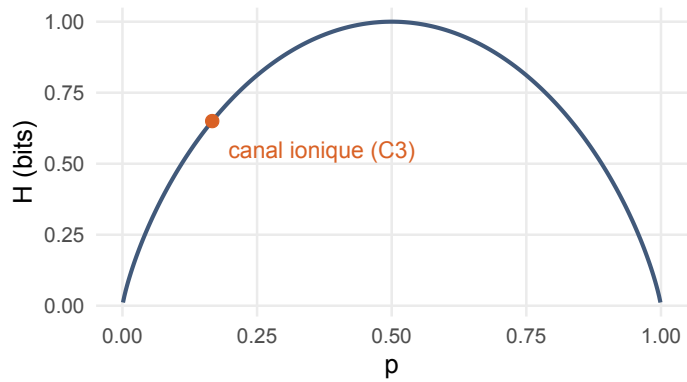
Exemples d'entropie

Loi	Entropie
Pièce équilibrée, (1/2, 1/2)	1 bit
Base d'ADN uniforme, (1/4, 1/4, 1/4, 1/4)	2 bits
Dé à 8 faces équilibré	3 bits
Position invariable, (1, 0, 0, 0)	0 bit

Loi uniforme sur K modalités : $H = \log_2 K$.

13

Entropie binaire



Deux modalités de probabilités p et $1 - p$: maximum 1 bit en $p = 1/2$.

15

14

16

Entropie et nombre de questions

Loi $(1/2, 1/4, 1/8, 1/8)$ sur A, C, G, T : $H = 1,75$ bit.

Questions : « A ? », puis « C ? », puis « G ? ».

$$\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 1,75 \text{ questions en moyenne}$$

Lecture de l'entropie

H : nombre moyen de questions oui/non de la meilleure stratégie, à une question près.

Information d'un site de fixation

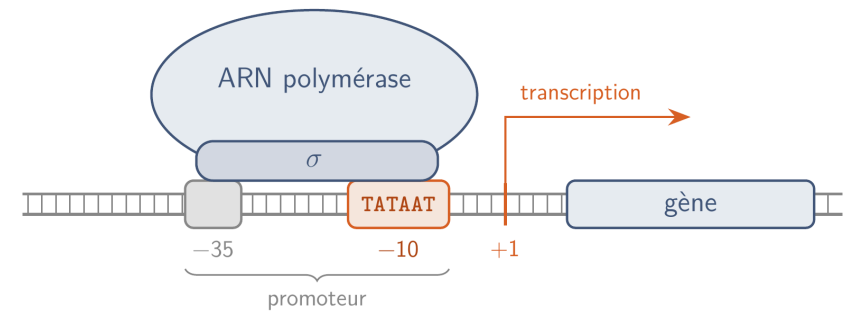
Entropie de l'équilibre du canal ionique

Équilibre du canal (C3) : fermé $5/6$, ouvert $1/6$.

$$H = -\frac{5}{6} \log_2 \frac{5}{6} - \frac{1}{6} \log_2 \frac{1}{6} \approx 0,65 \text{ bit}$$

Canal le plus souvent fermé : moins d'un bit d'incertitude sur son état.

Site de fixation



Une courte séquence d'ADN reconnue par une protéine, qui s'y fixe.

Boîte -10 d'un promoteur bactérien : site de l'ARN polymérase. Plus la séquence est proche de TATAAT, plus la polymérase s'y fixe, plus le gène est transcrit.

Alignement de sites de fixation

Promoteurs de gènes différents, reconnus par la même ARN polymérase.

Facteur σ : la sous-unité de la polymérase qui lit la boîte -10.

Séquences superposées sur le motif reconnu : une ligne par site, une colonne par position.

Colonne j : les bases présentes à la position j , d'un site à l'autre.

Schneider, Stormo, Gold et Ehrenfeucht, *Journal of Molecular Biology*, 1986.

Comptage à une position

1	2	3	4	5	6	7	8
A	T	A	T	A	A	T	A
A	T	A	T	A	A	T	C
C	T	A	T	A	A	T	G
C	T	A	T	A	A	T	T
G	T	A	T	A	A	T	C
G	T	A	T	T	A	T	G
T	T	A	T	T	A	C	A
T	T	A	T	T	A	C	C
A	T	A	A	T	A	G	G
C	T	G	A	T	A	G	T

Base	Effectif n_{bj}
A	3
C	3
G	2
T	2
Total	10

Fréquence d'une base à une position

La fréquence de la base b à la position j : la proportion des sites qui portent b en j .

$$f_{bj} = \frac{n_{bj}}{n}, \quad n = 10 \text{ sites}, \quad \sum_{b \in \{A,C,G,T\}} f_{bj} = 1$$

Position 1 : $\frac{1}{10} (3, 3, 2, 2) = (0,3; 0,3; 0,2; 0,2)$.

Base codée par un vecteur indicateur, $\mathbf{e}_A = (1, 0, 0, 0)$, ... : colonne j = moyenne des n vecteurs indicateurs de la position j .

Matrice des fréquences

	1	2	3	4	5	6	7	8
A	0,3	0,0	0,9	0,2	0,5	1,0	0,0	0,2
C	0,3	0,0	0,0	0,0	0,0	0,0	0,2	0,3
G	0,2	0,0	0,1	0,0	0,0	0,0	0,2	0,3
T	0,2	1,0	0,0	0,8	0,5	0,0	0,6	0,2

Colonne j : la loi des bases à la position j , un vecteur de $\mathbb{R}^4$ de somme 1.

Hypothèse du modèle : positions indépendantes, une loi par colonne.

Conservation et sélection

Position conservée : base déterminante pour la fixation. Mutation → fixation plus faible → gène moins transcrit → mutation contre-sélectionnée.

Position variable : pas de contrainte, mutations neutres. Positions 1 et 8 : hors du motif, fréquences proches de 1/4.

Consensus TATAAT rarement complet : l'écart au consensus module la force du promoteur, donc le niveau d'expression du gène.

Entropie position par position

Position	1	2	3	4	5	6	7	8
H (bits)	1,97	0	0,47	0,72	1	0	1,37	1,97

Positions 2 et 6 : aucune incertitude. Positions 1 et 8 : presque l'incertitude maximale, 2 bits.

Contenu en information

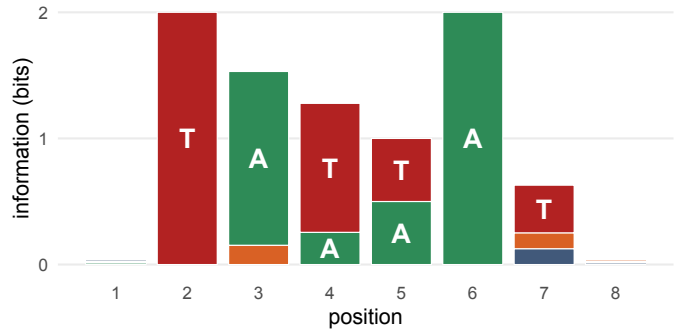
La réduction d'incertitude qu'apporte le site à une position, par rapport à une base tirée au hasard.

$$R_j = 2 - H_j \in [0, 2] \quad (\text{en bits})$$

Position	1	2	3	4	5	6	7	8
R (bits)	0,03	2	1,53	1,28	1	2	0,63	0,03

Information totale du site : $\sum_j R_j \approx 8,5$ bits.

Logo de séquence



Hauteur d'une pile : R_j . Part de chaque base : sa fréquence.

Représentation de Schneider et Stephens, *Nucleic Acids Research*, 1990, en barres empilées.

Entropie et variance

Valeurs 1 à 8

Loi sur $\{1, \dots, 8\}$	Entropie	Variance
Uniforme	3 bits	5,25
Moitié sur 1, moitié sur 8	1 bit	12,25

Désaccord

La variance récompense l'éloignement des valeurs, l'entropie le nombre de modalités possibles.

Deux mesures de dispersion

Variance : écart des valeurs à leur moyenne. Des valeurs numériques.

Entropie : incertitude sur la modalité observée. Les valeurs n'interviennent pas, seulement leurs probabilités.

Modalités non numériques

A, C, G, T : ni moyenne ni variance ; l'entropie reste définie.

Coder A = 1, C = 2, G = 3, T = 4 : une variance qui dépend du codage choisi.

Renommer les modalités ne change pas l'entropie.

Loi normale et entropie

Parmi les lois de variance donnée, la loi normale est celle d'entropie maximale.

$$X \sim \mathcal{N}(\mu, \sigma^2) \implies H(X) = \frac{1}{2} \log_2(2\pi e \sigma^2)$$

Admis. Entropie d'une loi continue : une valeur qui peut être négative.

Doubler l'écart-type : $\frac{1}{2} \log_2 4 = 1$ bit de plus.

Scott E. Page, *The Model Thinker*, ch. 12 : lois d'entropie maximale sous contrainte de moyenne ou de variance.

Loi jointe de deux positions

Variance et information

Données gaussiennes : l'entropie croît avec la variance.

Variance comme information

Garder les directions de grande variance : garder l'information.

Tige d'ARN

	1	2	3	4	5	6	7	8	9	0
G	A	C	A	U	G	A	G	U	C	
G	A	C	C	U	G	A	G	U	C	
G	U	C	A	U	U	A	G	A	C	
G	U	C	C	U	U	A	G	A	C	
G	G	C	A	U	U	A	G	C	C	
G	G	C	C	U	U	A	G	C	C	
G	C	C	A	U	G	A	G	G	C	
G	C	C	C	U	G	A	G	G	C	

Positions 2 et 9 (orange) : appariées dans une tige, A–U ou G–C. Positions 4 et 6 (bleu) : sans lien.

Séquences illustratives.

Loi jointe

La probabilité de chaque couple de bases, rangée dans une matrice.

$$J_{ij} = \Pr(X = i, Y = j), \quad \sum_{i,j} J_{ij} = 1$$

X : base en position 9.

Y : base en position 2.

Lignes et colonnes : A, C, G, U.

$$J = \begin{pmatrix} 0 & 0 & 0 & 1/4 \\ 0 & 0 & 1/4 & 0 \\ 0 & 1/4 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \end{pmatrix}$$

Indépendance

Deux positions indépendantes : la loi jointe est le produit des marges.

$$J = \mathbf{p} \mathbf{q}^T, \quad J_{ij} = p_i q_j$$

$\mathbf{p} \mathbf{q}^T$: une colonne (4×1) fois une ligne (1×4). Toutes les colonnes proportionnelles à $\mathbf{p}$: une matrice de **rang 1**.

Positions 4 et 6 : $\begin{pmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 \end{pmatrix}$. Tige d'ARN :

$J \neq \mathbf{p} \mathbf{q}^T$, des positions dépendantes.

Marges

Loi de chaque position : sommes des lignes, sommes des colonnes.

$$\text{loi de } X = J \mathbf{1}, \quad \text{loi de } Y = J^T \mathbf{1}$$

Deux produits matrice-vecteur (C3). Tige d'ARN :

$$J \mathbf{1} = J^T \mathbf{1} = (1/4, 1/4, 1/4, 1/4).$$

37

Lois conditionnelles

Diviser chaque colonne de J par sa somme : la loi de X sachant $Y = j$.

$$\Pr(X = i \mid Y = j) = \frac{J_{ij}}{\sum_k J_{kj}}$$

Colonnes de somme 1 : une **matrice de transition** (C3).

Tige d'ARN : un seul 1 par colonne. La base en 2 détermine la base en 9.

39

38

40

Information mutuelle

Ce que la connaissance d'une position apprend sur l'autre.

$$I(X; Y) = H(X) + H(Y) - H(X, Y), \quad H(X, Y) = - \sum_{i,j} J_{ij} \log_2 J_{ij}$$

Propriétés admises : $I \geq 0$, et $I = 0$ si et seulement si $J = \mathbf{p} \mathbf{q}^T$.

Information mutuelle et corrélation

Deux variables gaussiennes de corrélation r :

$$I = -\frac{1}{2} \log_2(1 - r^2) \quad (\text{admis})$$

$r = 0,5$: 0,21 bit. $r = 0,9$: 1,20 bit. $r = 0,99$: 2,83 bits.

Corrélation : liens linéaires entre valeurs. Information mutuelle : toute dépendance, entre valeurs ou entre modalités.

Coévolution

Positions 2 et 9 : $H(X) = H(Y) = 2$, $H(X, Y) = 2$, $I = 2$ bits.

Positions 4 et 6 : $I = 0$.

Mutations compensatoires

Les bases changent ensemble, l'appariement se conserve : l'information mutuelle détecte une contrainte de structure.

41

42

Synthèse

43

44

Trois mesures

Incertitude, dispersion, dépendance

Entropie : incertitude sur des modalités.

Variance : dispersion de valeurs numériques.

Information mutuelle : dépendance entre deux variables, quelle qu'en soit la forme.