Eléments d'analyse combinatoire - solutions des exercices

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2018-11-19

Contents

Exercice 02.1: oligomères sans résidus répétés

Combien d'oligomères peut-on former (ADN ou peptides) en utilisant chaque résidu une et une seule fois ?

Approche suggérée: agrégez progressivement les résidus, en vous demandant à chaque étape combien d'entre eux n'ont pas encore été incorporés.

Questions subsidiaires:

- Généralisez la formule pour des séquences d'objets tirés dans un ensemble de taille arbitraire (n).
- Quel est le nom de la fonction donnant le résultat ?
- Dans ce processus, quel est le mode de sélection des résidus: **avec ou sans remise** ?

Solution 02.1: oligomères sans résidus répétés

- Premier résidu: n possibilités.
- Dès le moment où on a choisi ce premier résidu, il ne reste plus que $n-1$ possibilités pour le second. On a donc $n \cdot (n-1)$ possibilités pour les deux premiers résidus.
- Pour la troisième position, il ne reste que $n-2$ résidus. On a donc $n \cdot (n-1) \cdot (n-2)$ possibilités pour les 3 premières positions de la séquence.
- Par extension, le nombre total de possibilités est donc (en supposant n suffisamment grand)

$$n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$$

.

- Dans notre cas:
  - $n! = 4! = 24$ oligonucléotides comportant exactement 1 fois chaque nucléotide (taille 4).
  - $n! = 20! = 2.43 \times 10^{18}$ oligopeptides (taille 20).

Exercice 02.2 : listes (ordonnées) de gènes

On a mené une expérience de transcriptome pour mesurer le niveau d'expression de tous les gènes de la levure. Sachant que le génome comporte 6000 gènes, combien de possibilité existe-t-il pour sélectionner les 15 gènes les plus fortement exprimés (**en tenant compte** de l'ordre relatif de ces 15 gènes) ?

Approche suggérée: comme précédemment, simplifiez le problème en partant de la sélection minimale, et augmentez progressivement le nombre de gènes (1 gène, 2 gènes, ...).

Questions subsidiaires:

- Trouvez un exemple familier de jeu de pari apparenté à ce problème.
- Généralisez la formule pour la sélection d'une liste de x gènes dans un génome qui en comporte n.

Solution 02.2 : listes (ordonnées) de gènes

Il s'agit d'une sélection **sans remise** (chaque gène apparaît à une et une seule position dans la liste de tous les gènes), et **ordonnée** (les mêmes gènes pris dans un ordre différent sont considérés comme un résultat différent).

- Pour le premier gène, il y a $n = 6000$ possibiité.
- Dès le moment où on connaît le premier gène, il n'existe plus que 5999 possibilités pour le second, et donc $n \cdot (n-1) = 6000 \cdot 5999$ possibilités pour la suite des deux premiers gènes;
- Par extension, il existe $6000 \cdot 5999 \cdot 5998 \cdot \ldots \cdot 5986 = 4.62 \times 10^{56}$ possibilités pour les 15 premiers gènes.
- En généralisant à la liste des x premiers gènes dans un ensemble de n, on obtient $N = n \cdot (n-1) \cdot (n-2) \cdot ... \cdot (n-x+1)$.

Exercice 02.3 : ensembles (non-ordonnés) de gènes

Lors d'une expérience de transcriptome indiquant le niveau d'expression de tous les gènes de la levure. Sachant que le génome comporte 6000 gènes, combien de possibilité existe-t-il pour sélectionner les 15 gènes les plus fortement exprimés (**sans tenir compte** de l'ordre relatif de ces 15 gènes) ?

Approche suggérée: comme précédemment, simplifiez le problème en partant de sélections minimales (1 gène, 2 gènes, ...) et généralisez la formule.

Questions subsidiaires:

- Trouvez un exemple familier de jeu de pari apparenté à ce problème.
- Généralisez la formule pour la sélection d'un ensemble de x gènes dans un génome qui en comporte n.
- Connaissez-vous le nom de la formule ainsi trouvée ?

Solution 02.3 : ensembles (non-ordonnés) de gènes

- Pour une sélection d'un seul gène, il existe $n = 6000$ possibilité.
- Pour 2 gènes, il existe $n \cdot (n-1) = 6000 \cdot 5999$ arrangements, mais ceci inclut deux fois chaque paire de gènes ((a,b) et (b,a)). Le nombre d'ensembles non ordonnés est donc $N = n(n-1)/2$.
- De même, pour 3 gènes, il faut diviser le nombre d'arrangements ($A_n^x = \frac{n!}{(n-x)!} = 6000 \cdot 5999 \cdot 5998$) par le nombre de permutations parmi tous les triplets de gènes ($(a,b,c), (a,c,b), (b,a,c) \ldots$), ce qui donne $\frac{6000!}{(6000-3)!3!} = \frac{6000 \cdot 5999 \cdot 5888}{6} = 3.6 \times 10^{10}$.
- Pour 15 gènes, on obtient $\frac{n!}{(n-x)!x!} = \frac{6000!}{5985! \cdot 15!} = 3.53 \times 10^{44}$ *combinaisons* possibles.

Exercice 02.4 : mutagénèse

On soumet un fragment d'ADN de 1 kilobase à un traitement mutagène qui provoque des mutations ponctuelles (substitutions) à 5 positions distinctes indépendantes. Combien de séquences possibles existe-t-il pour le fragment muté ?

Solution 02.4 : mutagenèse

On commence par dénombrer toutes les possibilités pour lesw positions de l'ensemble des 5 mutations, et ensuite on dénombre, pour chacune de ces positions, les substitutions possibles;

Le dénombrement des positions mutantes revient à choisir au hasard 5 positions parmi les 1000 nucléotides du fragment d'ADN. Il s'agit d'un choix sans remise (chaque position ne peut être tirée qu'une fois), on choisit donc le coefficient binomial.

$$\binom{n}{x} = \binom{1000}{5} = C_{1000}^{5} = \frac{1000!}{5!995!} = 8.2502913 \times 10^{12}$$

Pour chacune des 5 positions mutées, on a 3 substitutions possibles. Le nombre de mutations distinctes est donc 3^5 fois le nombre de choix de 5 positions parmi 1000.

$$N = 3^5 \cdot \binom{1000}{5} = 2.0048208 \times 10^{15}$$

Exercice 02.5 : oligopeptides 3×20

Combien d'oligopeptides de taille 60 peut-on former en utilisant exactement 3 fois chaque acide aminé ?

Solution 02.5 : oligopeptides 3×20

Commençons par générer une séquence particulière qui remplit ces conditions, en concaténant 3 copies de chaque acide aminé, dans l'ordre alphabétique.

Table 1: Symboles des acides aminés et codons

Aminoacid	Symbol3	Symbol	Codons
Alanine	Ala	A	GCA, GCC, GCG, GCT
Arginine	Arg	R	CGA, CGC, CGG, CGT, AGA, AGG
Aspartic acid	Asp	D	GAC, GAT
Asparagine	Asn	N	AAC, AAT
Cysteine	Cys	C	TGC, TGT
Glutamic acid	Glu	E	GAA, GAG
Glutamine	Gln	Q	CAA, CAG
Glycine	Gly	G	GGA, GGC, GGG, GGT
Histidine	His	H	CAC, CAT
Isoleucine	Ile	I	ATA, ATC, ATT
Leucine	Leu	L	CTA, CTC, CTG, CTT, TTA, TTG
Lysine	Lys	K	AAA, AAG
Methionine	Met	M	ATG
Phenylalanine	Phe	F	TTC, TTT
Proline	Pro	P	CCA, CCC, CCG, CCT
Serine	Ser	S	TCA, TCC, TCG, TCT, AGC, AGT
Threonine	Thr	T	ACT, ACC, ACG, ACT
Tryptophan	Trp	W	TGG
Tyrosine	Tyr	Y	TAC, TAT
Valine	Val	V	GTA, GTC, GTG, GTT
STOP	-	-	TAG, TAA, TGA

```
AAACCCDDDEEEFFFGGGHHHIIIKKKLLLMMMNNNPPPQQQRRRSSSTTTVVVWWWYYY
```

Les permutations de ces 60 lettres sont des solutions valudes. En voici trois exemples.

QEILYYFDHLVCRGVAEWFGDDTNNFMCWMVKQNELISASIRQHCSYMTRKHKPPGAWTP

AACASLKPEDIFSFGTDHKEPDNIHSNQYPYLFGMIHMKMLQQCNWTYVERVVCRRWGWT

FQSICEGYKHVMDNLPAQWRMSIQAEIYRELGWWPVHDTGPKYMCSTFRNVFLKDCATNH

...

TNRSFLPLNGSQKTRKNIETPLMHKVGGEDHDYEWDFICAQYYPQRVIVCMAHCWFAMWS

Le nombre total de permutations possibles parmi 60 éléments est $60! = 8.3209871 \times 10^{81}$. Cependant, cenombre dépasse de loin le nombre de séquences distinctes. En effet, dans chacune des séquences ci-dessus, chaque lettre apparaît 3 fois. Or, des permutations entre les trois positions occupées par des A ne changeront pas la séquence. Il en va de même pour les permutations entre les positions occupées par chacun des acides aminées : A, C, D, ...

Il faut donc diviser le nombre total de permutations (60!) par le nombre de permutations qui ne modifient pas la séquence: 3! pour A, 3! pour D, et ainsi de suite pour chacun des 20 acides aminés.

La formule finale est donc:

$$N = \frac{\overbrace{60!}^{60 \text{ lettres}}}{\underbrace{3! \cdot 3! \cdot \ldots 3!}_{20 \text{ acides aminés présents 3 fois}}} = \frac{60!}{(3!)^{20}} = 2.2758825 \times 10^{66}$$