Eléments d’analyse combinatoire - solutions des exercices

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2018-11-19

Table of 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 ().
* 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: possibilités.
* Dès le moment où on a choisi ce premier résidu, il ne reste plus que possibilités pour le second. On a donc possibilités pour les deux premiers résidus.
* Pour la troisième position, il ne reste que résidus. On a donc possibilités pour les 3 premières positions de la séquence.
* Par extension, le nombre total de possibilités est donc (en supposant suffisamment grand)
* .
* Dans notre cas:
  + oligonucléotides comportant exactement 1 fois chaque nucléotide (taille 4).
  + 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 gènes dans un génome qui en comporte .

## 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 possibiité.
* Dès le moment où on connaît le premier gène, il n’existe plus que 5999 possibilités pour le second, et donc possibilités pour la suite des deux premiers gènes;
* Par extension, il existe possibilités pour les 15 premiers gènes.
* En généralisant à la liste des premiers gènes dans un ensemble de , on obtient .

## 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 gènes dans un génome qui en comporte .
* 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 possibilité.
* Pour 2 gènes, il existe arrangements, mais ceci inclut deux fois chaque paire de gènes ( et ). Le nombre d’ensembles non ordonnés est donc .
* De même, pour 3 gènes, il faut diviser le nombre d’arrangements () par le nombre de permutations parmi tous les triplets de gènes (), ce qui donne .
* Pour 15 gènes, on obtient *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.

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

## Exercice 02.5 : oligopeptides

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

## Solution 02.5 : oligopeptides

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

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.

SQGCKFFHADLEQMPVSCKYFHVTRIYDGMELWTMTGCWVKNIHYPNRPSNRALIQDEWA

ELIFTVTHHCYKNVMQKTAYPLLSQWCWNGRFVDWHEFRQIDEKRCDNGPSGMYASMPIA

FMHNRWAAGMFEYPTAVLKDQLVKRFSVHKECWPTPTSCQGIYNWNDCRGQMSHIDEYLI

...

CMFMVPWLNRQQISKRKPCYHPEADAKYEGTIADLWGTHQNVICHFRDGEMWTLVSFYSN

Le nombre total de permutations possibles parmi 60 éléments est . 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 ne changeront pas la séquence. Il en va de même pour les permutations entre les positions occupées par chacun des acides aminées : , , ,

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

La formule finale est donc: