Combinatorics

Probabilities and statistics for bioinformatics (STAT1)

Jacques van Helden

2019-09-13

Enumerating oligonucleotides and oligopeptides

Résumé des concepts et formules

Supplementary exercises

Enumerating oligonucleotides and oligopeptides

Problem

DNA is composed of 4 nucleotides denoted by the letters A, C, G, T. Proteins are made of 20 amino acids.

a. For each one of these two types of macromolecules, how many distinct oligomers can be formed by polymerizing 30 residues ("30-mers") ?

Suggested approach: start by addressing a simpler form of the same problem, by starting with polymers of much smaller sizes: 1, then 2 residues, ...

b. Generalize the formula for oligomers of an arbitrary size k (so-called **k-mers** in the domain), made of n distinct residues.

c. What is the name of the function resulting from this analysis?

d. In this process, which mode did you use to pick up the residues: **with** or **without replacement**?

Solution: enumeration of oligomers

- The underlying process is a **drawing with replacement**: at each position of the sequence, we can choose any of the n residues ($n = 4$ for nucleotidic sequences, $n = 20$ for peptidic sequences).
- Progressive approach of the solution
  - Trivial case: single-residue sequence $\rightarrow$ there are exactly n possibilities.
  - Two-residue sequences: for each of the n possible residues at the first position, we can select n resodies for the second one $\rightarrow$ there are $n \cdot n = n^2$ possible dimers.
  - Trimers: for each of these dimers, there are n possible residues that can be chosen for the3^d position $\rightarrow$ there are $n^2 \cdot n = n^3$ distinct trinucleotides.
- Generalisation to k-mers: there are n^k distinct sequences of

The geometric progression

The **geometric progression** is a succession of numbers where each term can be computed by multiplying the previous one by a constant factor.

$$x_i = x_{i-1} \cdot n$$

For a large size of k the formula can be developed.

$$\begin{aligned} x_k &= x_{k-1} \cdot n \\ &= (x_{k-2} \cdot n) \cdot n = x_{k-2} \cdot n^2 \\ &= x_{k-3} \cdot n^3 = \ldots = x_0 \cdot n^k \end{aligned}$$

In our case, the initial value is $x_0 = 1$; k denotes the oligomer size, and n is the number of distinct residues used to form the oligomer ($n = 4$ for nucleic acids, $n = 20$ for amino acids).

Number of oligomers

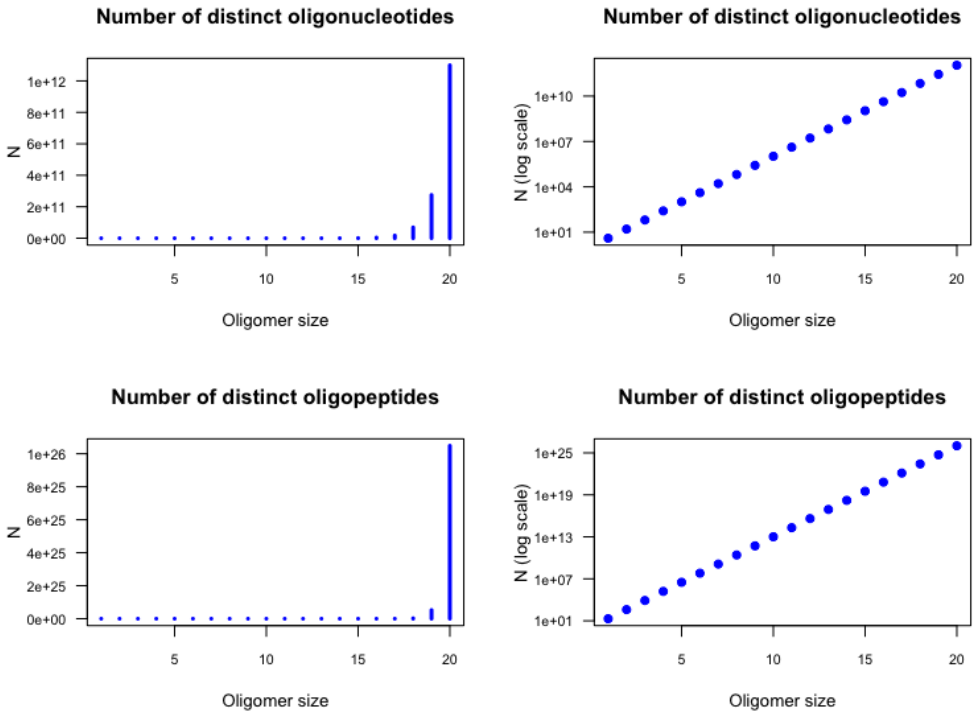


Figure 1: Number of possible oligonucleotides (top) and oligopeptides (bottom) with either a linear (left) and logarithmic (right) scale for the ordinate.

Exercise 02.1: oligomers with no repeated residue

How many oligomers can be formed (DNA or peptides) that would contain exactly once each residue.

Suggested approach: progressively aggregate the residues whilst wondering, at each step, bow many residues have not yet been incoroporated in the sequence.

Sub-questions:

- Generalise the formula for sequences of items of any type, drawn from a set of arbitrary size n.
- What is the name of the corresponding function?
- In this process, what is the mode of residue selection: **with** or **without replacement**?

Solution: oligomers with no repeated residue

- First residue: n possibilities.
- As soon as the first residue has been chosen, there are only $n-1$ possibilities to draw a different residue for the second position. We thus have $n \cdot (n-1)$ possible sequences for the first two residues.
- For the third position, there are only $n-2$ residues left; We thus have On a donc $n \cdot (n-1) \cdot (n-2)$ possibilities for the 3 first positions of the sequence.
- By extension of this reasoning, the total number of possible solutions (assuming n is not too small) will be

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

- In our case:
  - $n! = 4! = 24$ oligonucléotides comportant exactement 1 fois

The factorial function

- Enumerates the number of possible permutations of a finite set of items
- Drawing withour replacement
- Defined by a recursive formula

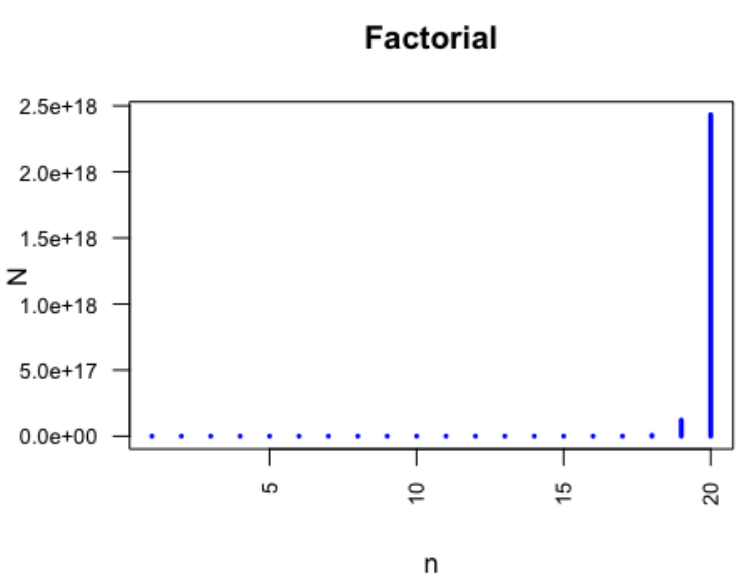
$$N = n! = \begin{cases} 1 & \text{if } n = 0 \\ n \cdot (n-1)! & \text{otherwise} \end{cases}$$

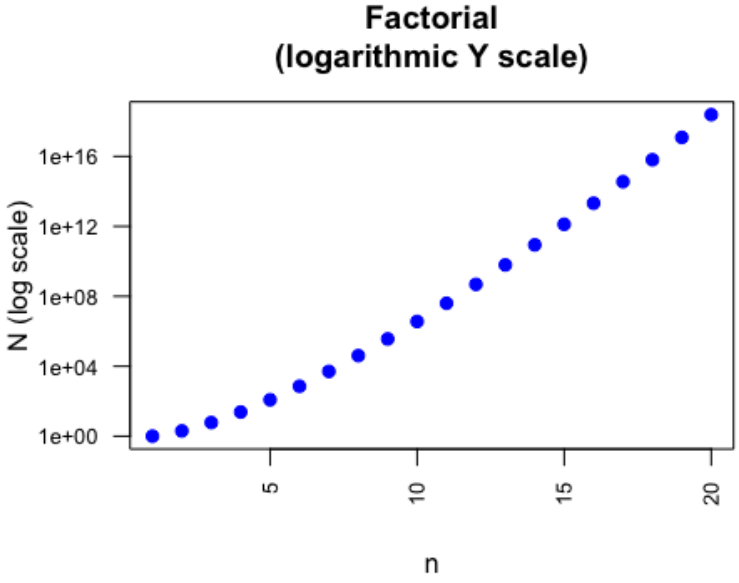
Note: by definition, $0! = 1$, which enables to compute $1!$ and the subsequent numbers with the recursive formula.

For sufficiently large values of n, a clearer formulation is

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

Factorial


Factorial
2.5e+18
2.0e+18
1.5e+18
1.0e+18
5.0e+17
0.0e+00
N
5
10
15
20
n


Factorial
(logarithmic Y scale)
1e+16
1e+12
1e+08
1e+04
1e+00
N (log scale)
5
10
15
20
n

Exercise 02.2: gene lists (with order)

A transcriptome experiment has been led to define the level of expression of all the yeast genes. Knowing that the genome contains 6000 genes, how many possible ways are there to select the 15 most expressed genes *with their relative order*?

Suggested approach: as previously, simplify the problem by starting from the minimal selection, and progressively increase the number of selected genes (1 gène, 2 gènes, ...).

Complementary questions:

- Give the example of a familiar bet game related to this enumerating process.
- Generalise the formula for any selection of a list of x items in a set containing n elements.

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 $N = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-x+1)$.

Arrangements

In combinatorics, the term **arrangement** denotes an *orderless* drawing *without replacement*, i.e. random drawing where the order of the item is taken in consideration, and where each already selected item cannot be selected as next element.

Number of arrangements of x items drawn in a set of size n.

$$\begin{aligned} A_n^x &= \frac{n!}{(n-x)!} \\ &= \frac{n(n-1)\ldots(n-x+1)(n-x)(n-x-1)\ldots 2\cdot 1}{(n-x)(n-x-1)\ldots 2\cdot 1} \\ &= n \cdot (n-1) \cdot \ldots \cdot (n-x+1) \end{aligned}$$

Arrangements – Typical application

- **tricast** (also named **trifecta**).
- A bet where players must predict the three winner horses ($x = 3$) of a race, and the exact order of their arrival. For $n = 15$ horses, there are $n \cdot (n-1) \cdot (n-2) = 15 \cdot 14 \cdot 13 = 2730$ possibilities.

Exercise 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 $\frac{n!}{(n-x)!x!} = \frac{6000!}{5985! \cdot 15!} = 3.53 \times 10^{44}$ *combinations* possibles.

Combinations

A **combination** is a selection *without replacement* a finite set, where the order of drawing is taken in considration.

The number of possible combinations of x numbers among n is provided by the **bbinomial coefficient**.

$$\binom{n}{x} = C_n^x = \frac{n!}{x!(n-x)!}$$

Attention: the relative positions of x and n are opposite in the two alternative notations for combinations *binomnx* ("x among n") and (C_n^x, "choose").

combinations – Typical application

- **trio**, a variation of the **tricast** bet, where the order of arrival of the 3 winner horses is not taken in consideration.

$$\binom{n}{x} = \binom{15}{3} = C_{15}^{3} = \frac{15!}{3!12!} = 455$$

- **loto** (or lotto): each better checks 6 numbers within a grid containing 90 numbers. The number of possibilities is

$$\binom{n}{x} = \binom{90}{6} = C_{90}^{6} = \frac{90!}{6!84!} = 6.2261463 \times 10^{8}$$

Résumé des concepts et formules

Tirages avec / sans remise

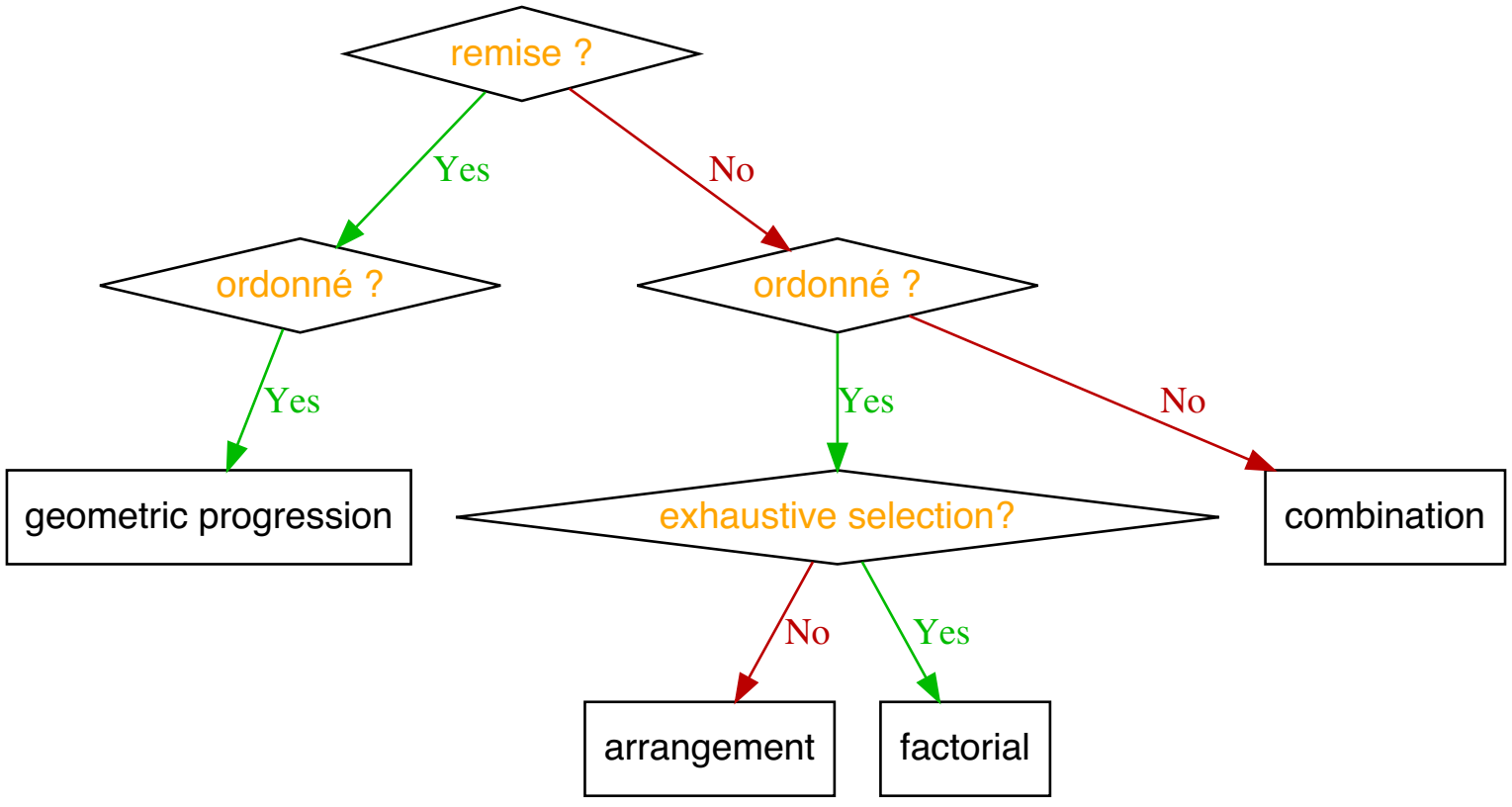
Il existe deux types classiques de tirage d'éléments au sein d'un ensemble: avec ou sans remise.

1. **Tirage sans remise**: chaque élément peut être tiré au plus une fois. Exemples:
   - Jeu de loto (ou lotto).
   - Sélection aléatoire d'un ensemble de gènes dans un génome.
2. Tirage **avec remise**: chaque élément peut être tiré zéro, une ou plusieurs fois. Exemples:
   - Jeu de dés. A chaque lancer on dispose des mêmes possibilités (6 faces).
   - Génération d'une séquence aléatoire, par sélection itérative d'un élément dans l'ensemble des résidus (4 nucléotides pour l'ADN, 20 acides aminés pour les protéines).

Elements of combinatorics

- Arrangements: drawings with order, without replacement
- Combinations: drawings without considering the order of the items, without replacement)

Choix de la formule



Formulas

Replacement?	Ordre?	Formula	Description
Yes	Yes	n^x	**Geometric progression**: ordered drawings (sequences), with replacement, of x items from a set of size n
No	Yes	$n!$	**factorial**: permutations of all elements of a set of size n
No	Yes	$A_n^x = \frac{n!}{(n-x)!}$	**Arrangements** : ordered drawing, without replacement, of x items in a set of size n
No	No	$C_n^x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$	**Combinations** : orderless drawing, without replacement, of x items in a set of size n

Supplementary exercises

Exercise 02.5: oligopeptides 3×20

How many distinct oligopeptides of size $k = 60$ can be formed by using exactly 3 times each amino acid?

Solution 02.5: oligopeptides 3×20

How many distinct oligopeptides of size $k = 60$ can be formed by using exactly 3 times each amino acid?

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

AAACCCDDDEEEFFFGGGHHHIIIKKKLLLMMMNNNPPPQQQRRRSSSTTTVVVWWWY

Toutes les permutations de ces 60 lettres sont des solutions valides. En voici trois exemples.

DHGASCGCFMMPYKKIANPMDKINSLRWIRELDSFPLTWQWTQAFGTVEECVYQNHHV

LFFLDGGKGEVWHSREVRKPAYFNIPPDANQHTRAQYIQCELSHCMCIKWSTMDWVYM

HPPYHKENYSLLMCWNTVFVIAVIDPGGQRASLWDCTFSHQYATMKEWDNFRQRIMGE

Cependant, il faut prendre en compte le fait que certaines