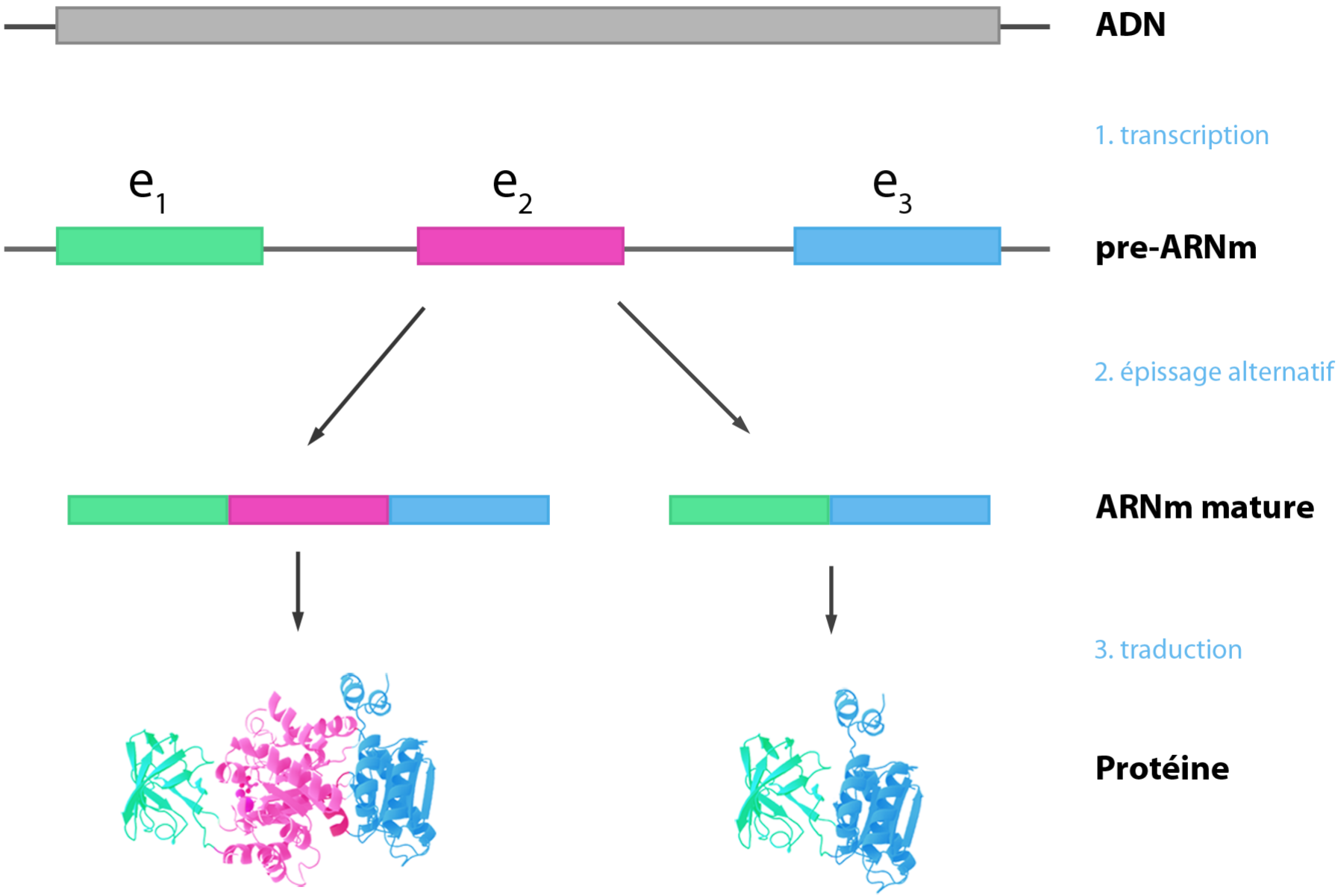
RASTAck

une nouvelle méthode **d'assemblage** du transcriptome.

Jérôme Audoux

July 17, 2014

mRNA

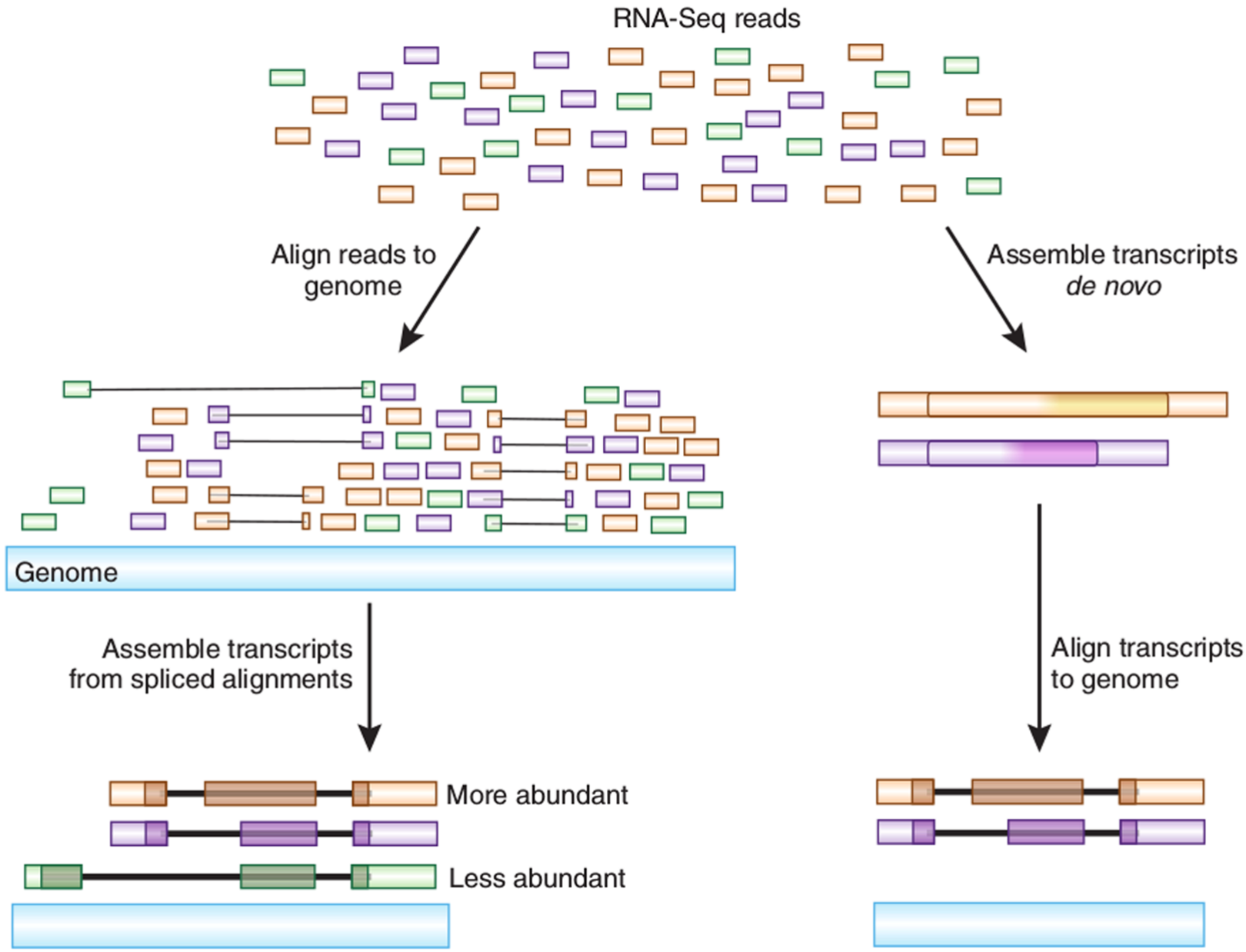

ADN
1. transcription
e1
e2
e3
pre-ARNm
2. épissage alternatif
ARNm mature
3. traduction
Protéine

RNA-Seq

Never send a human to do a machine's job.

© http://en.wikipedia.org/wiki/DNA_sequencing

Approches


RNA-Seq reads
Align reads to genome
Assemble transcripts de novo
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

RASTAck

Read Assembly With A STAck.

Postulat:

*Les reads disent la **vérité**.*

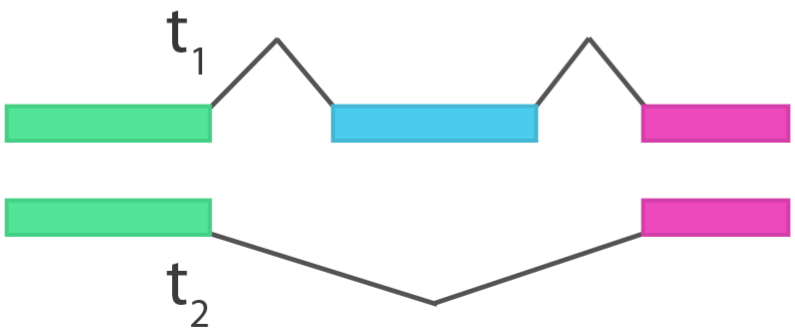
Objectif:

*Reconstruire les **meta-reads**.*

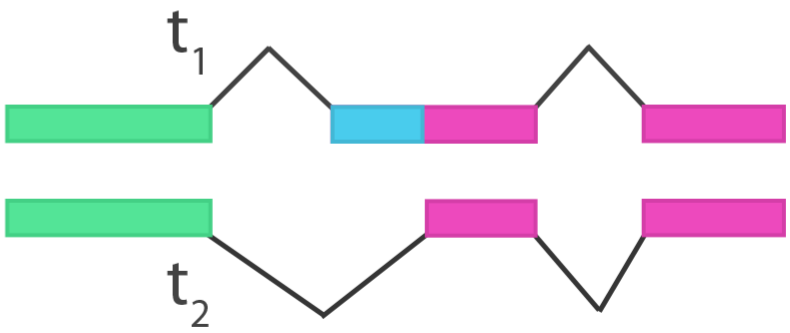
Un *meta-read* est un assemblage de reads, ou de facteurs de reads, qui partagent la même histoire.

Une *histoire* est une séquence du génome, continue ou non, qui est partagée dans son intégralité par les mêmes transcrits de la première à la dernière base.

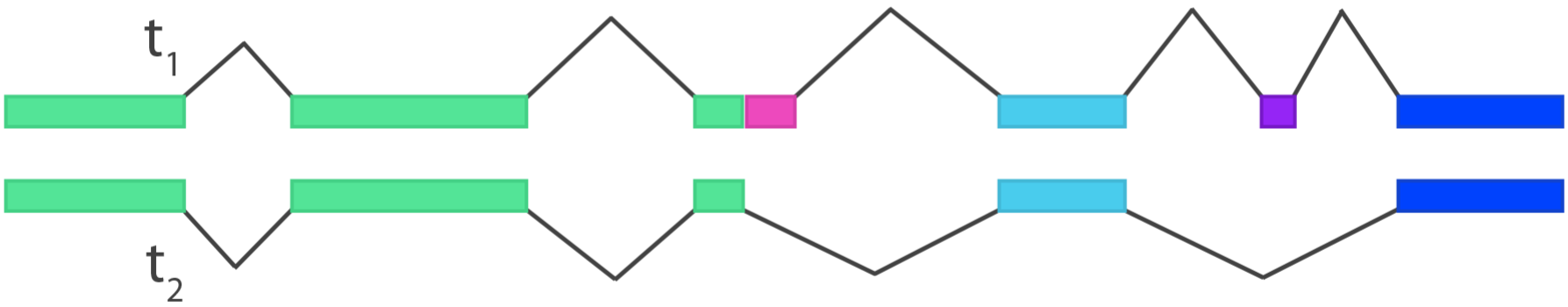
A. Exon skipping



B. Alternative 3’ acceptor sites



C. Complexe case



Un index de reads est une collection de reads indexés dans un structure de données qui va permettre d'extraire l'information sur les reads et la réponse à des requêtes sur la base d'un $k - mer$ présent dans la collection.

- Q_a **Combien de reads** sont indexés dans la *collection*?
- Q_b Quelle est la **longueur de la séquence du read** i dans la *collection*?
- Q_c **Quelle est la séquence du facteur** de longueur l à la position p du read i?

- Q_1: **Combien de reads partagent le facteur** à la position p du read i?

- Q_2: **Quels sont les reads qui partagent le facteur** à la position p du read i et **quel est la position de ce facteur dans ces reads**?

GkArrays

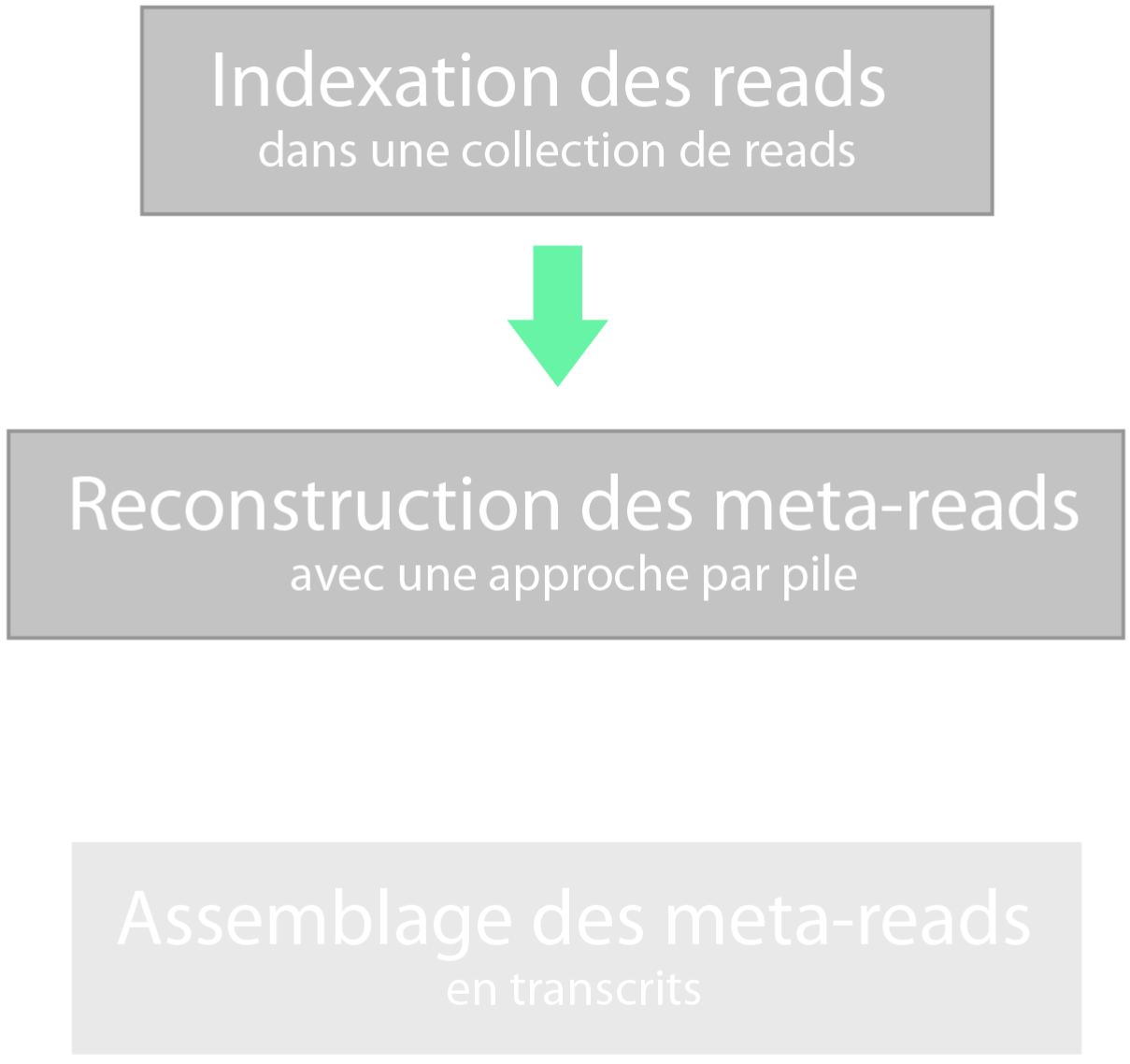
Philippe, Nicolas, Mikaël Salson, Thierry Lecroq, Martine Léonard, Thérèse Commes, and Eric Rivals. **"Querying Large Read Collections in Main Memory: A Versatile Data Structure."** BMC Bioinformatics 12, no. 1 (June 17, 2011): 242. doi:10.1186/1471-2105-12-242.

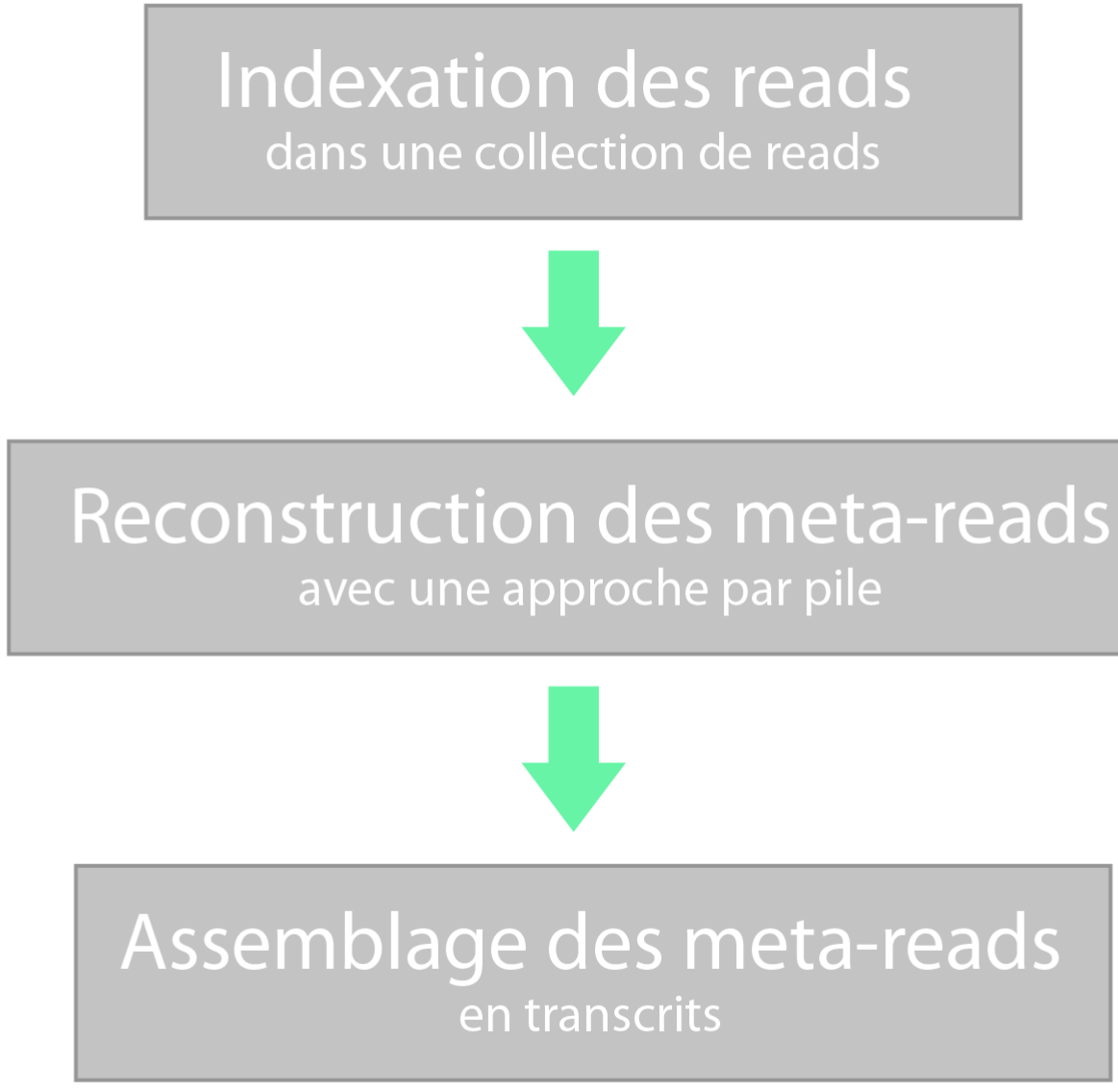
Méthode

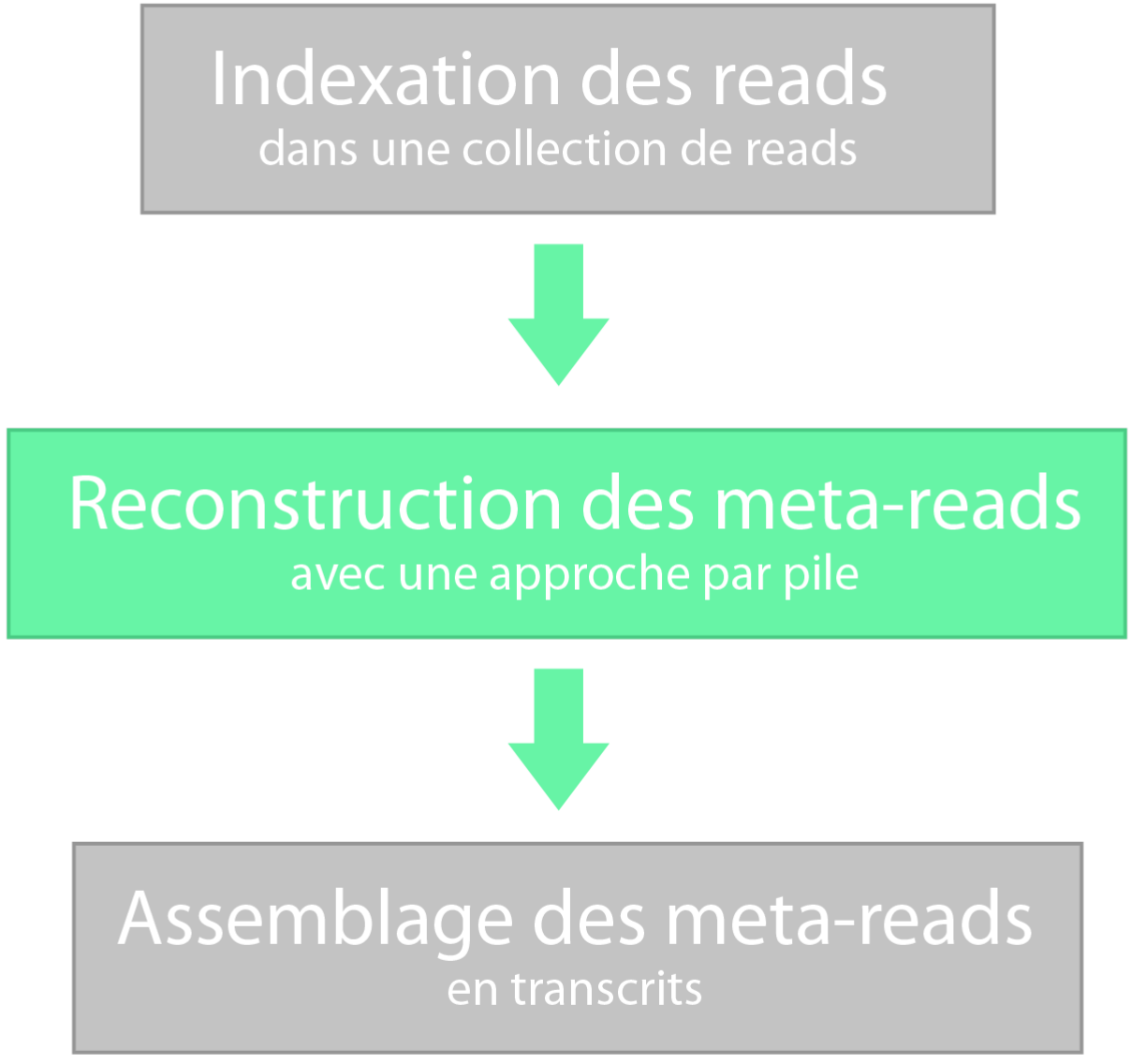
Indexation des reads
dans une collection de reads

Reconstruction des meta-reads
avec une approche par pile

Assemblage des meta-reads
en transcrits


Indexation des reads
dans une collection de reads
Reconstruction des meta-reads
avec une approche par pile
Assemblage des meta-reads
en transcrits

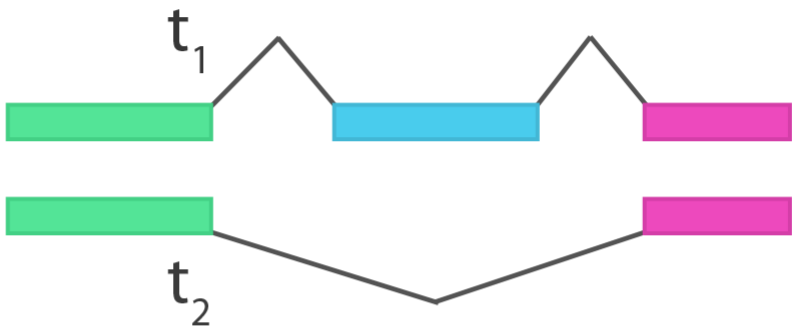

Indexation des reads
dans une collection de reads
Reconstruction des meta-reads
avec une approche par pile
Assemblage des meta-reads
en transcrits

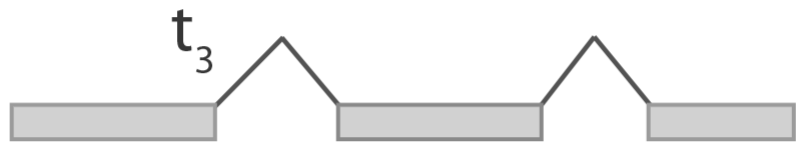

Indexation des reads
dans une collection de reads
Reconstruction des meta-reads
avec une approche par pile
Assemblage des meta-reads
en transcrits

Algorithme

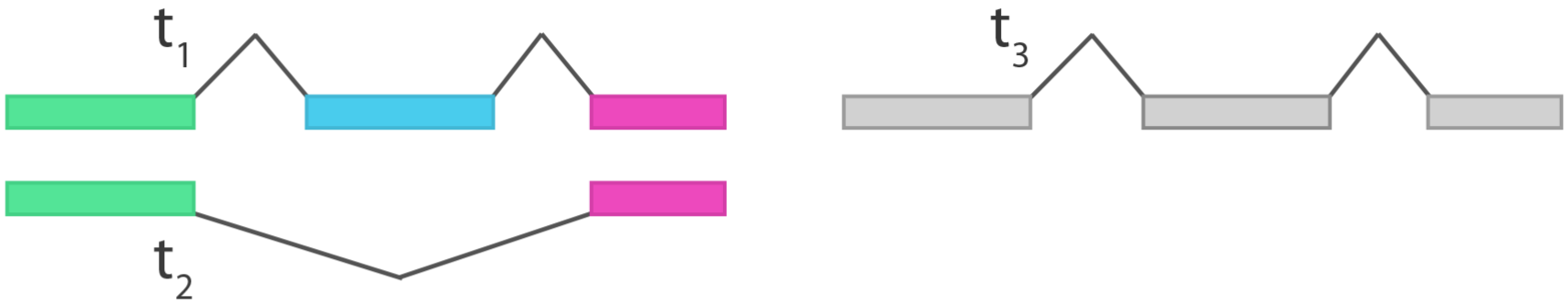
1. Choisir **un "bon" read** dans la collection
2. Choisir **un "bon" *k-mer*** pour créer l'empilement
3. Créer un **empilement** de reads
4. Définir **les bornes de confiance** de la pile
5. Choisir **les *k-mers* suivants** pour la procédure d'extension
6. **Marquer les reads** utilisés
7. **Terminer** la reconstruction du *meta-read* ou **continuer** à étendre

Les ≠ histoires





Les ≠ histoires



La collection de reads

Les ≠ histoires

La collection de reads

1. Choisir **un "bon" read** dans la collection
2. Choisir **un "bon" *k-mer*** pour créer l'empilement
3. Créer un **empilement** de reads
4. Définir **les bornes de confiance** de la pile
5. Choisir **les *k-mers* suivants** pour la procédure d'extension
6. **Marquer les reads** utilisés
7. **Terminer** la reconstruction du *meta-read* ou **continuer** à étendre

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG

33

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG

28

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG

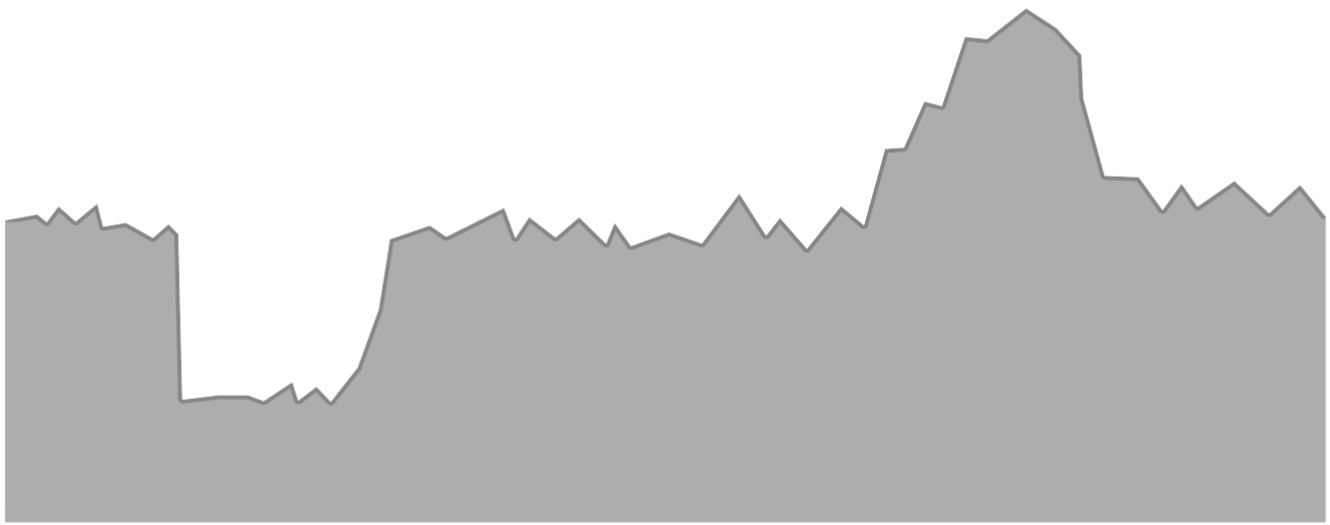
• • •

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG

12

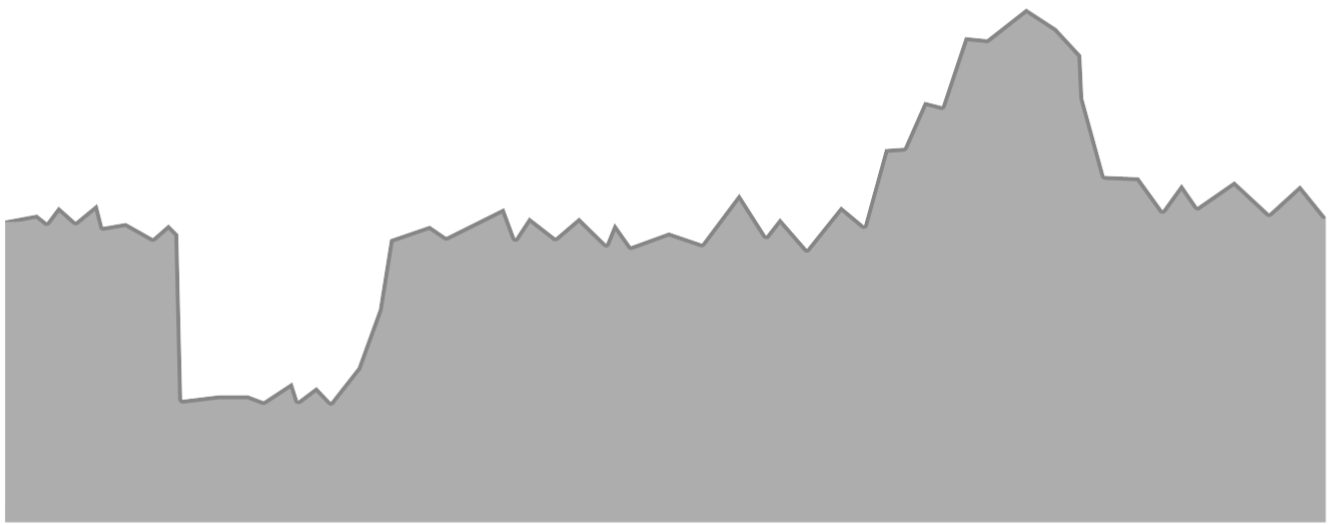
ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG



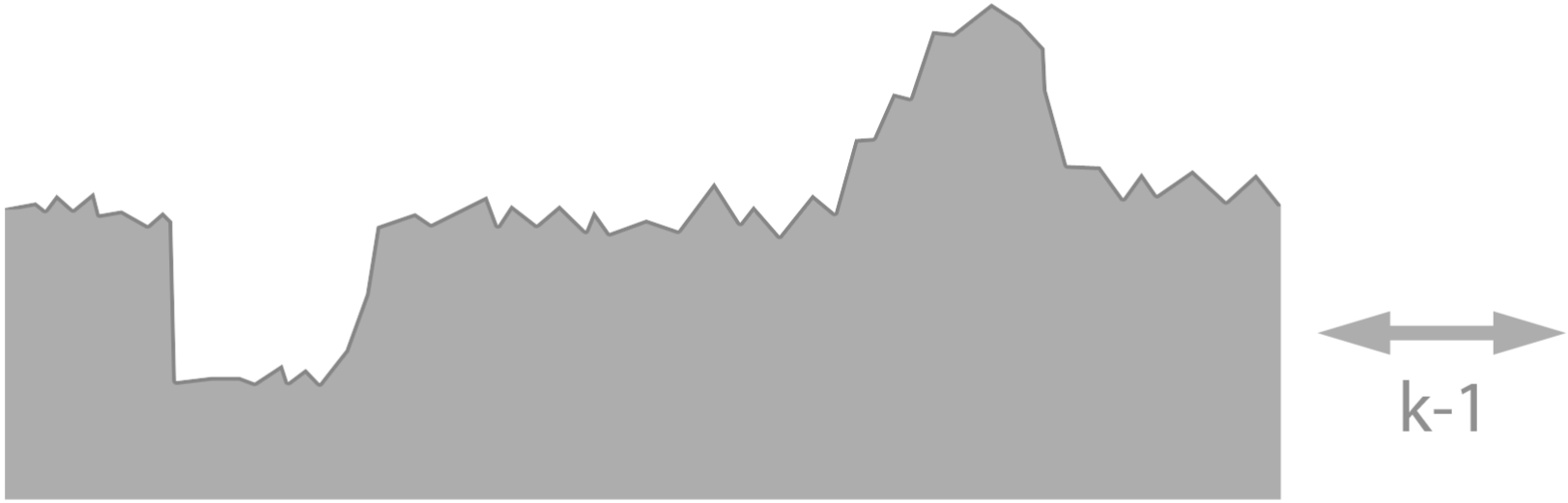
Profil de support

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG



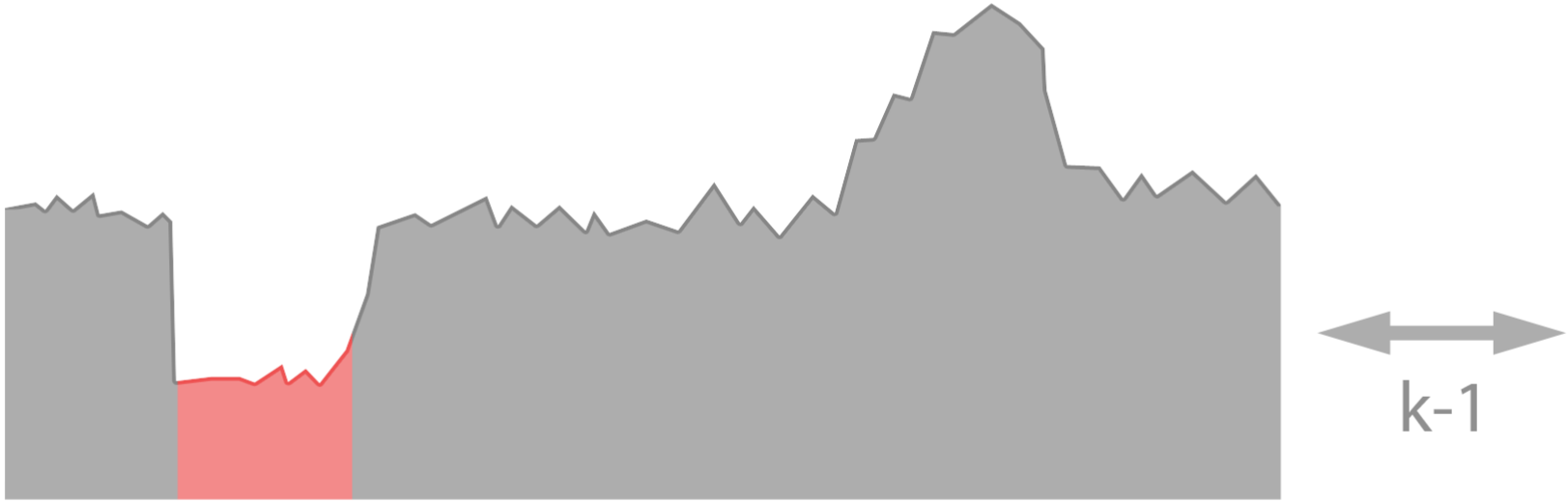
Profil de support

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG



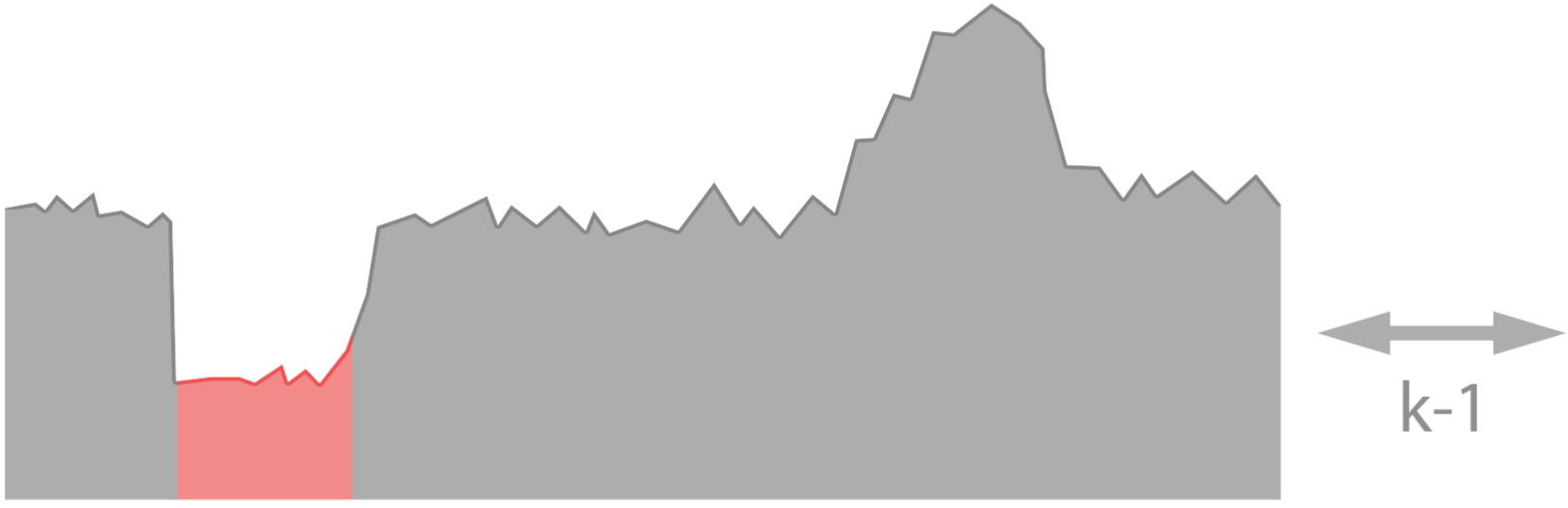
Profil de support

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG



Profil de support

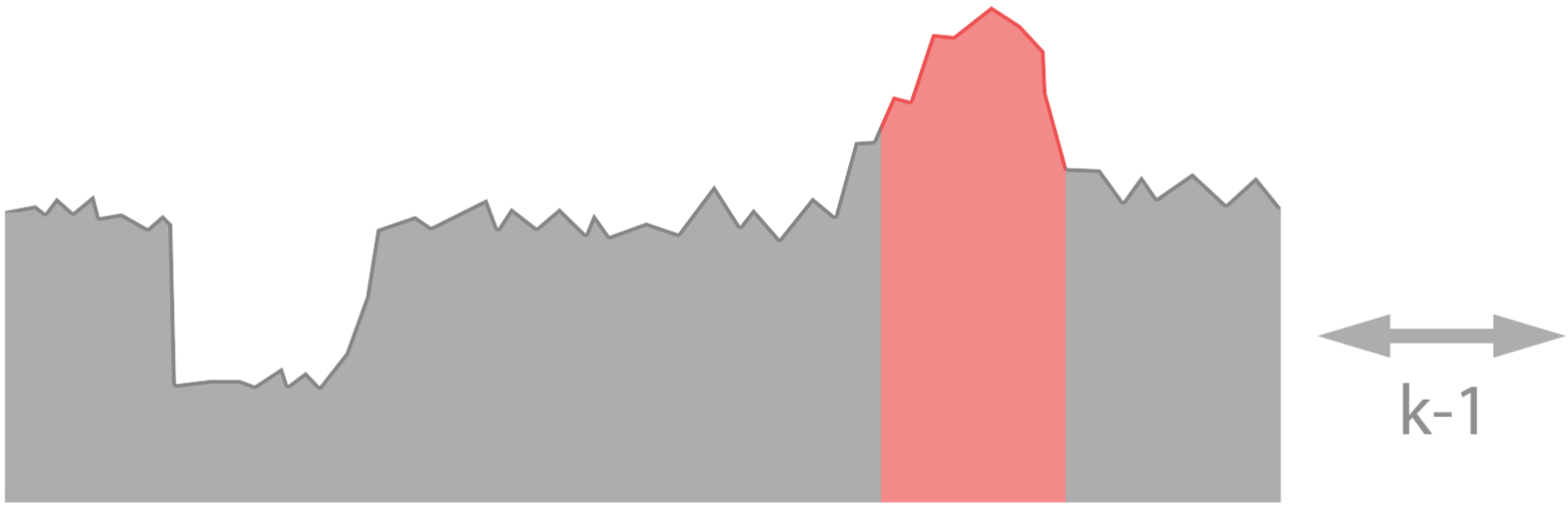
ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG



SNP/Erreur de séquence

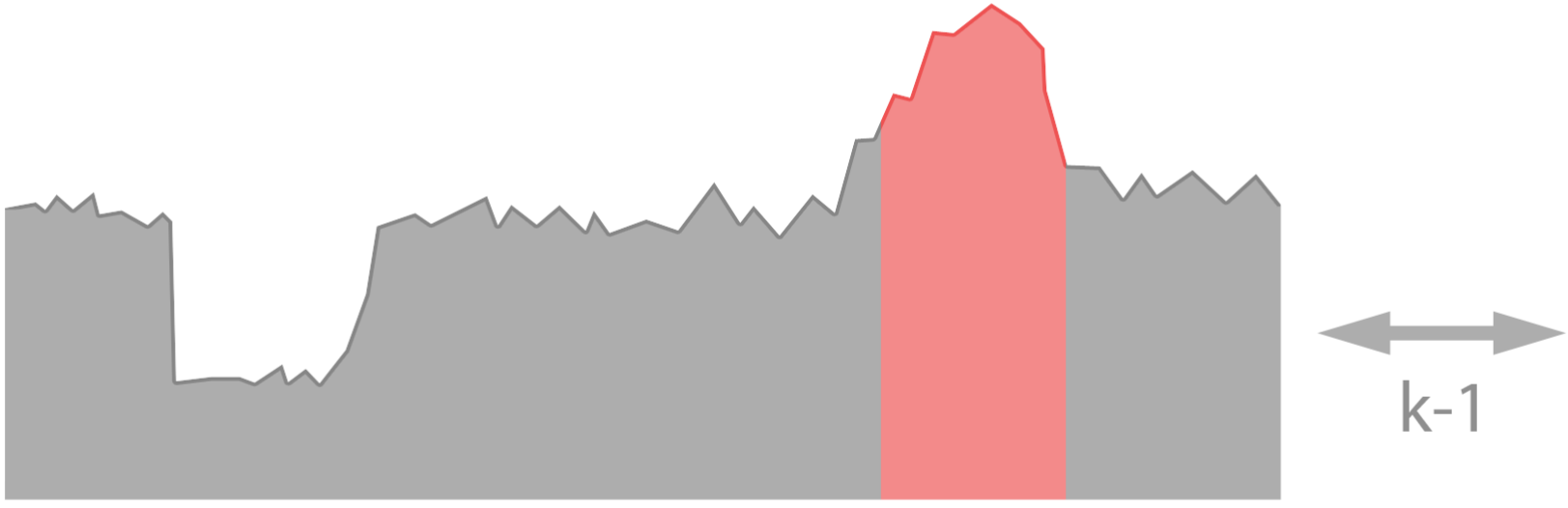
Profil de support

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG



Profil de support

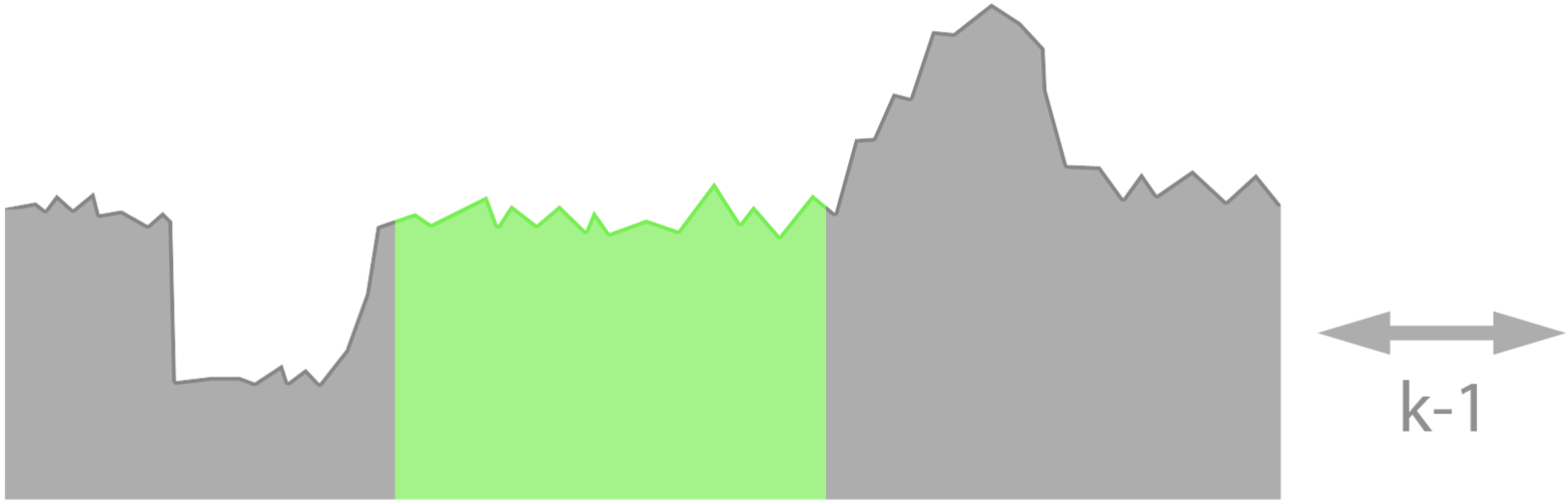
ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG



Séquence répétée

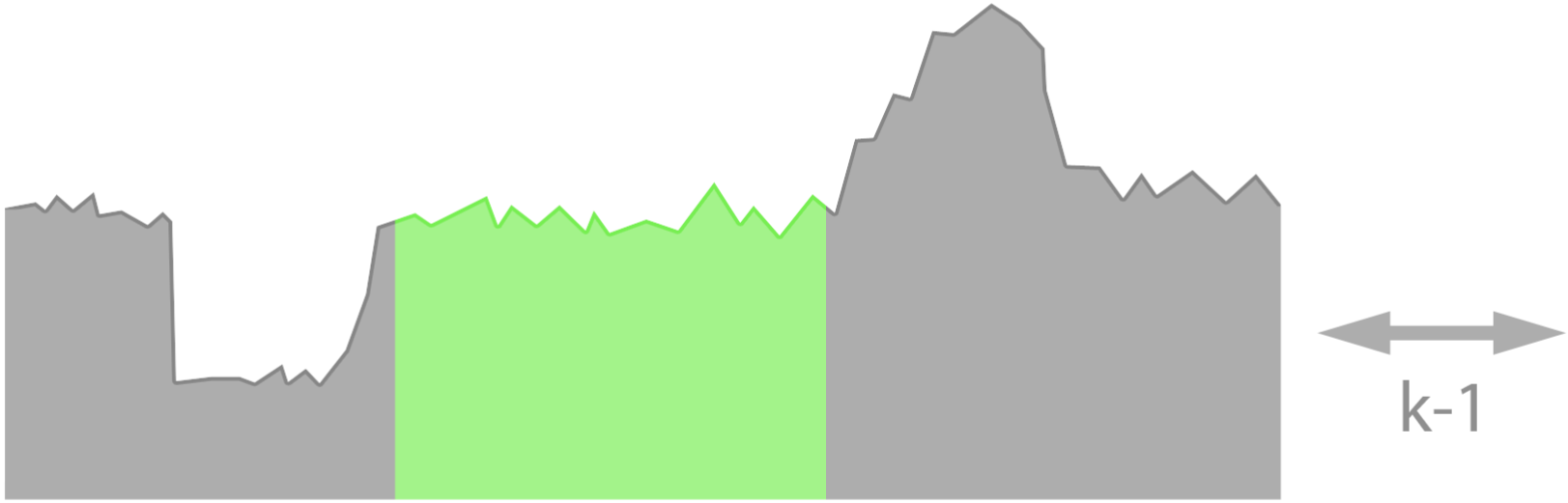
Profil de support

ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG



Profil de support

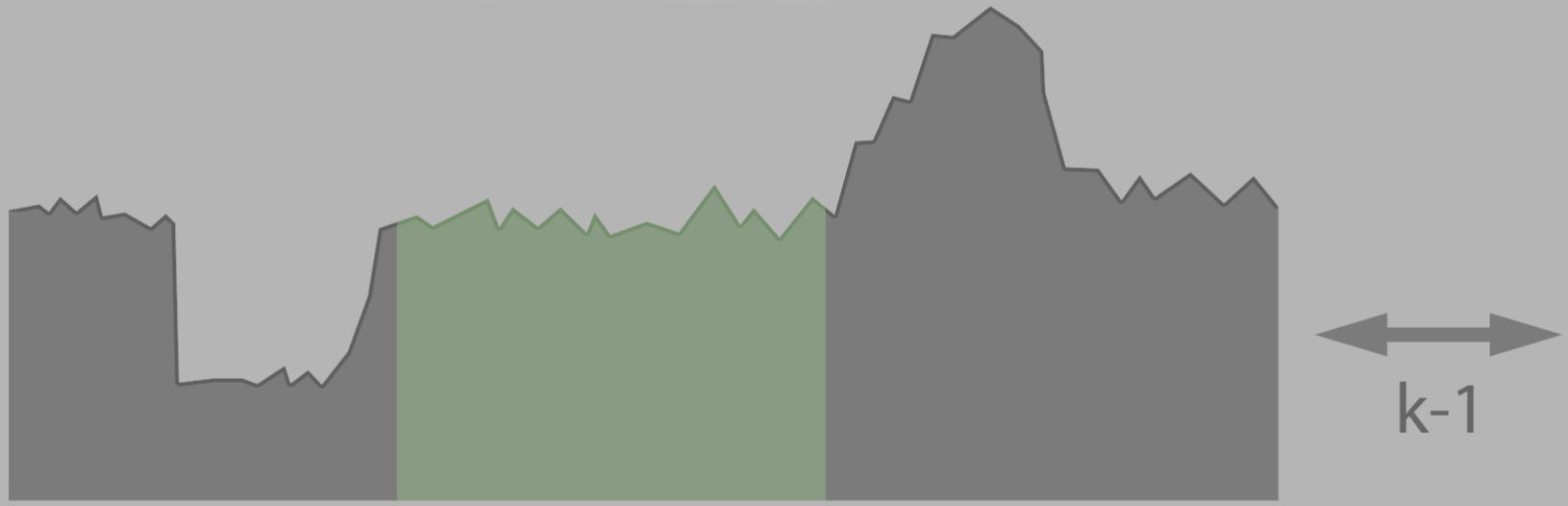
ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG



Support stable

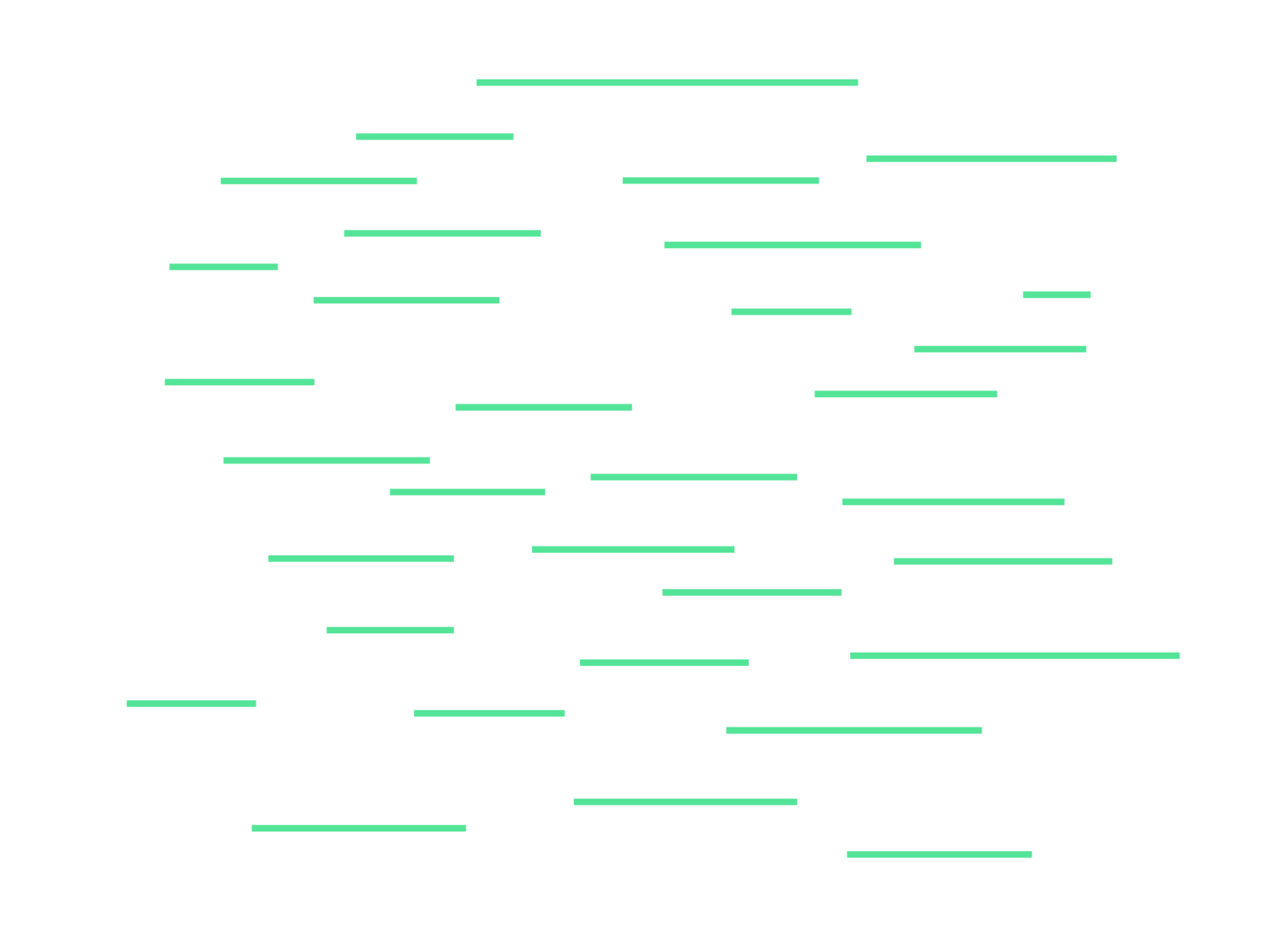
Profil de support

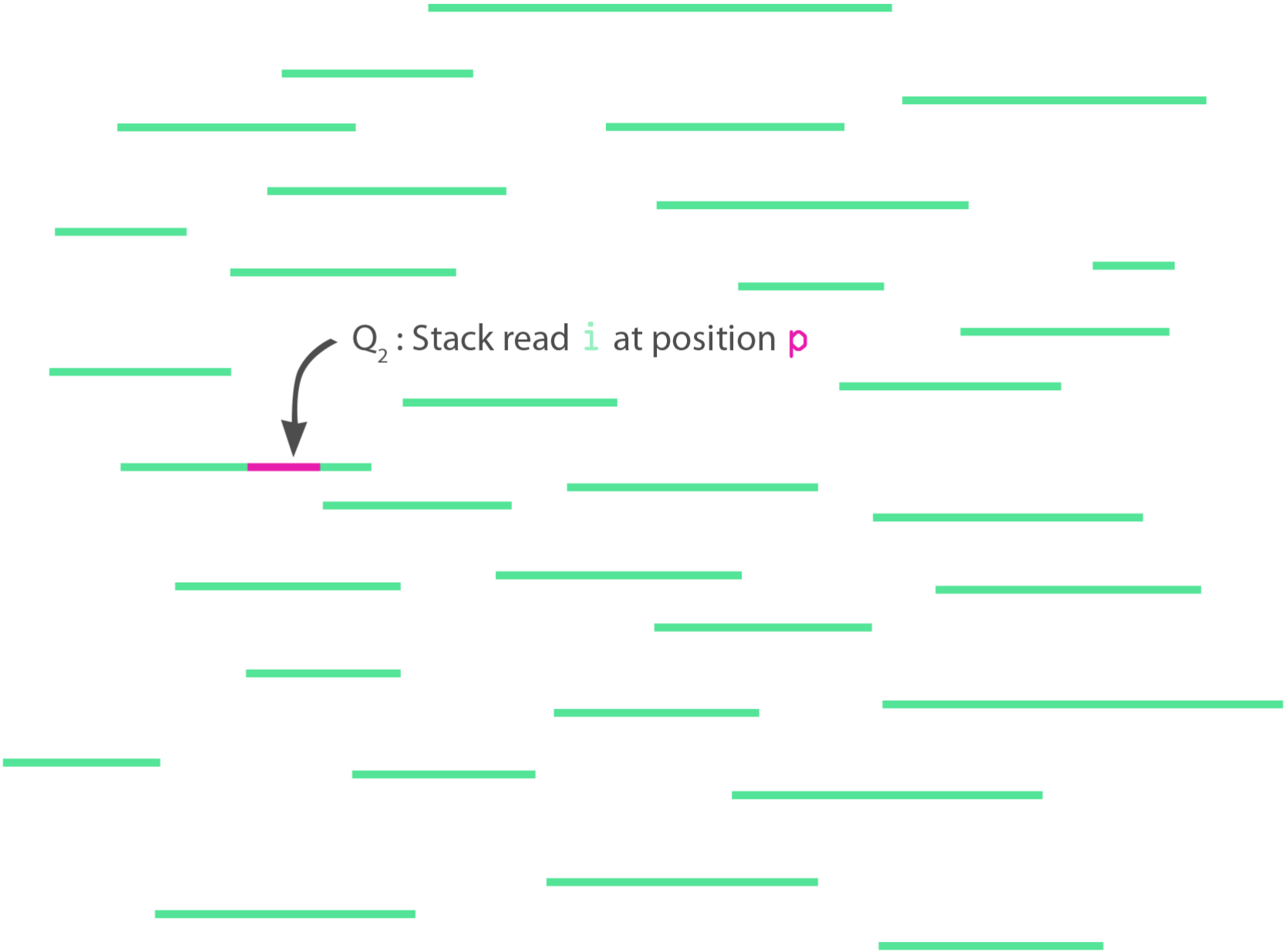
ATTGCTGATGCGCGATGCTAGGATGAGATCGCGCGATCGATGATAG

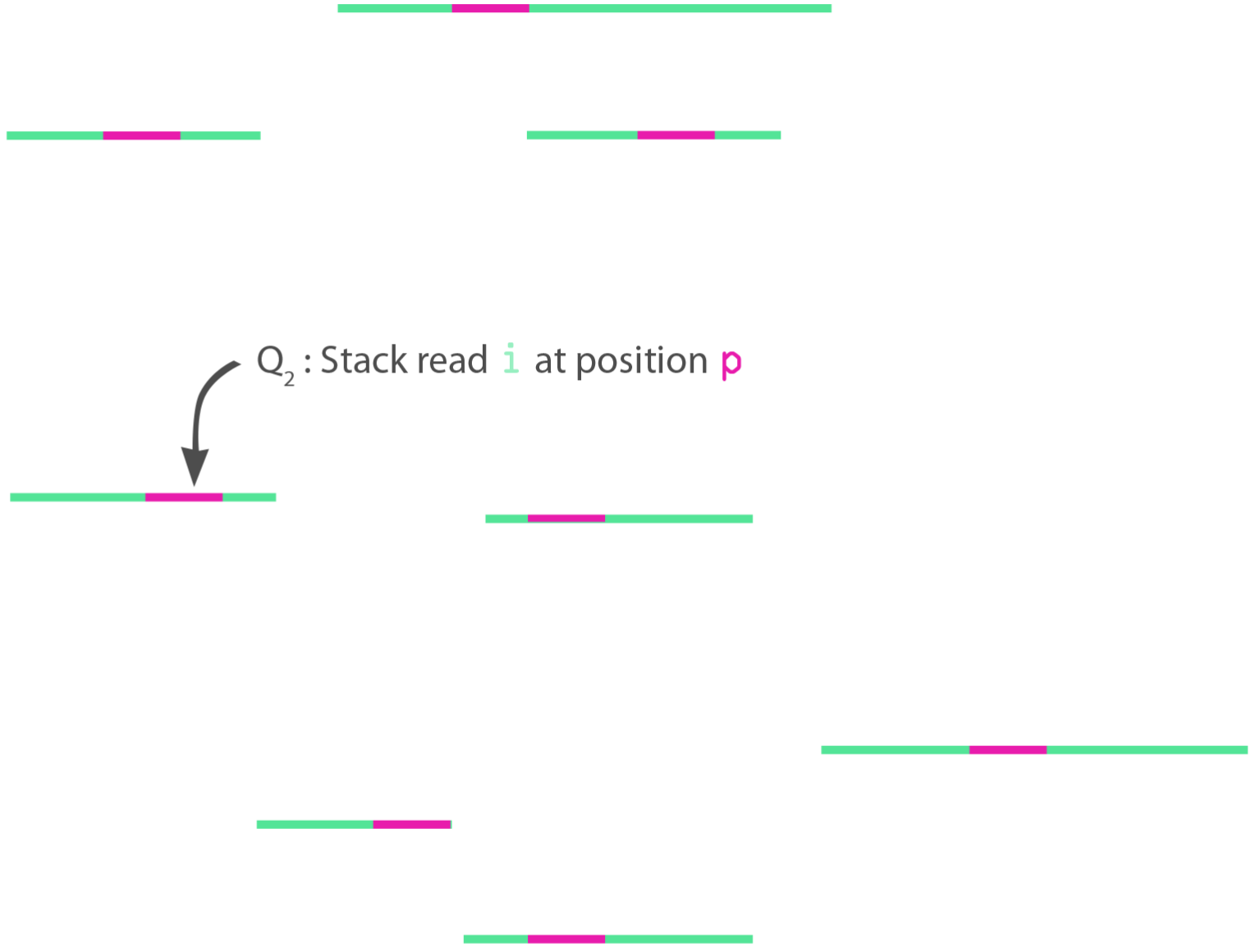


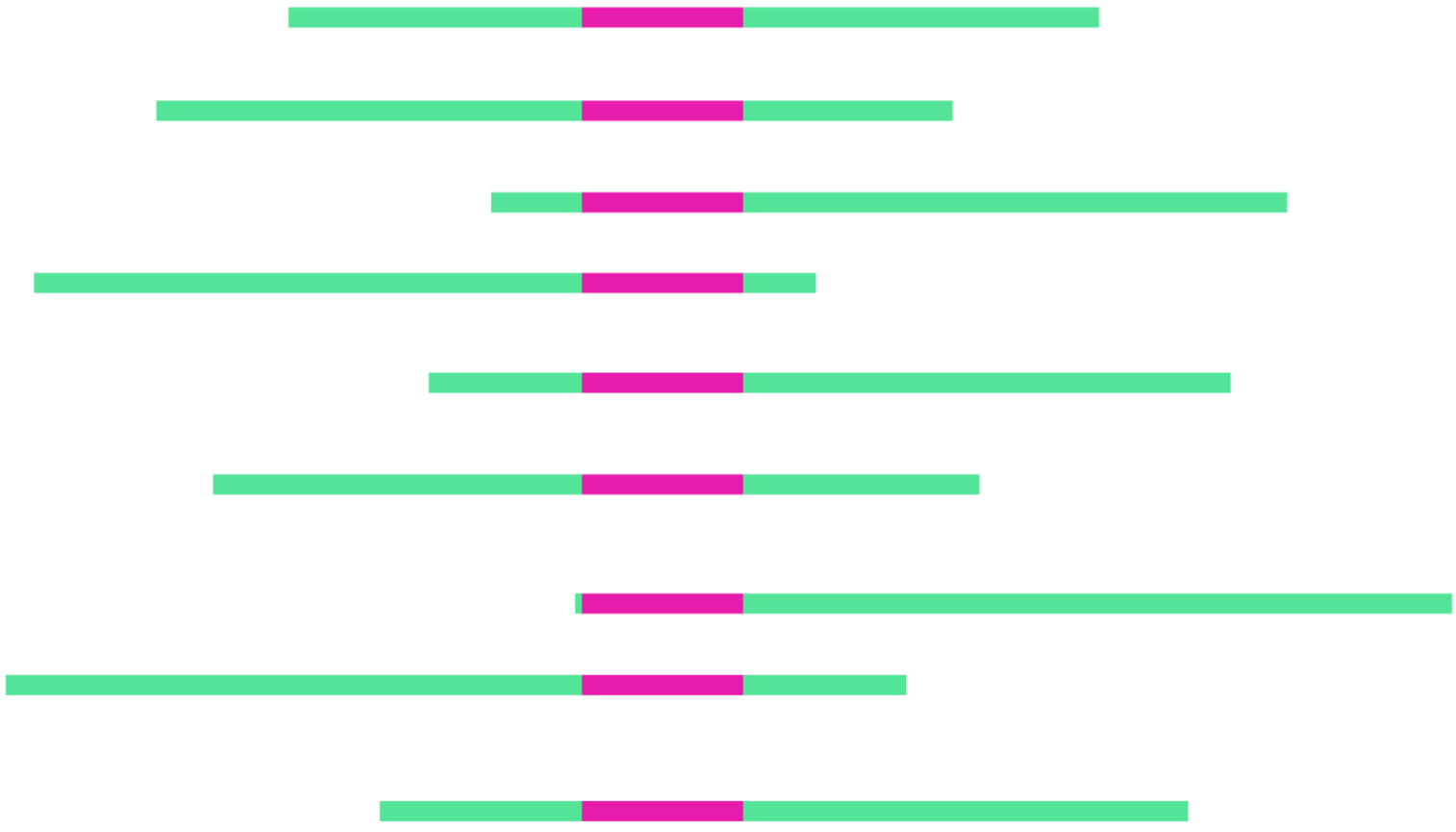
Support stable

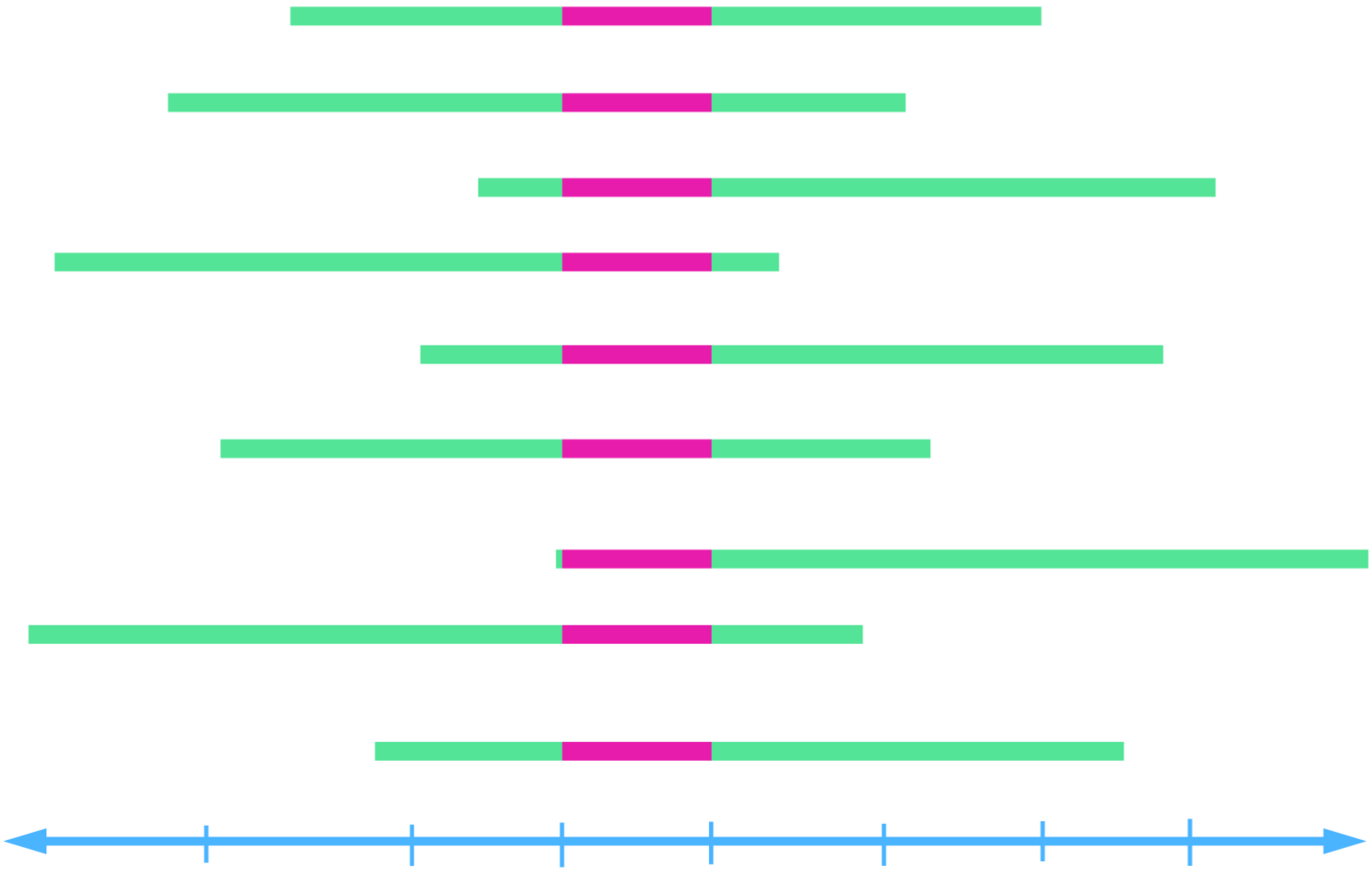
1. Choisir **un "bon" read** dans la collection
2. Choisir **un "bon" *k-mer*** pour créer l'empilement
3. Créer un **empilement** de reads
4. Définir **les bornes de confiance** de la pile
5. Choisir **les *k-mers* suivants** pour la procédure d'extension
6. **Marquer les reads** utilisés
7. **Terminer** la reconstruction du *meta-read* ou **continuer** à étendre

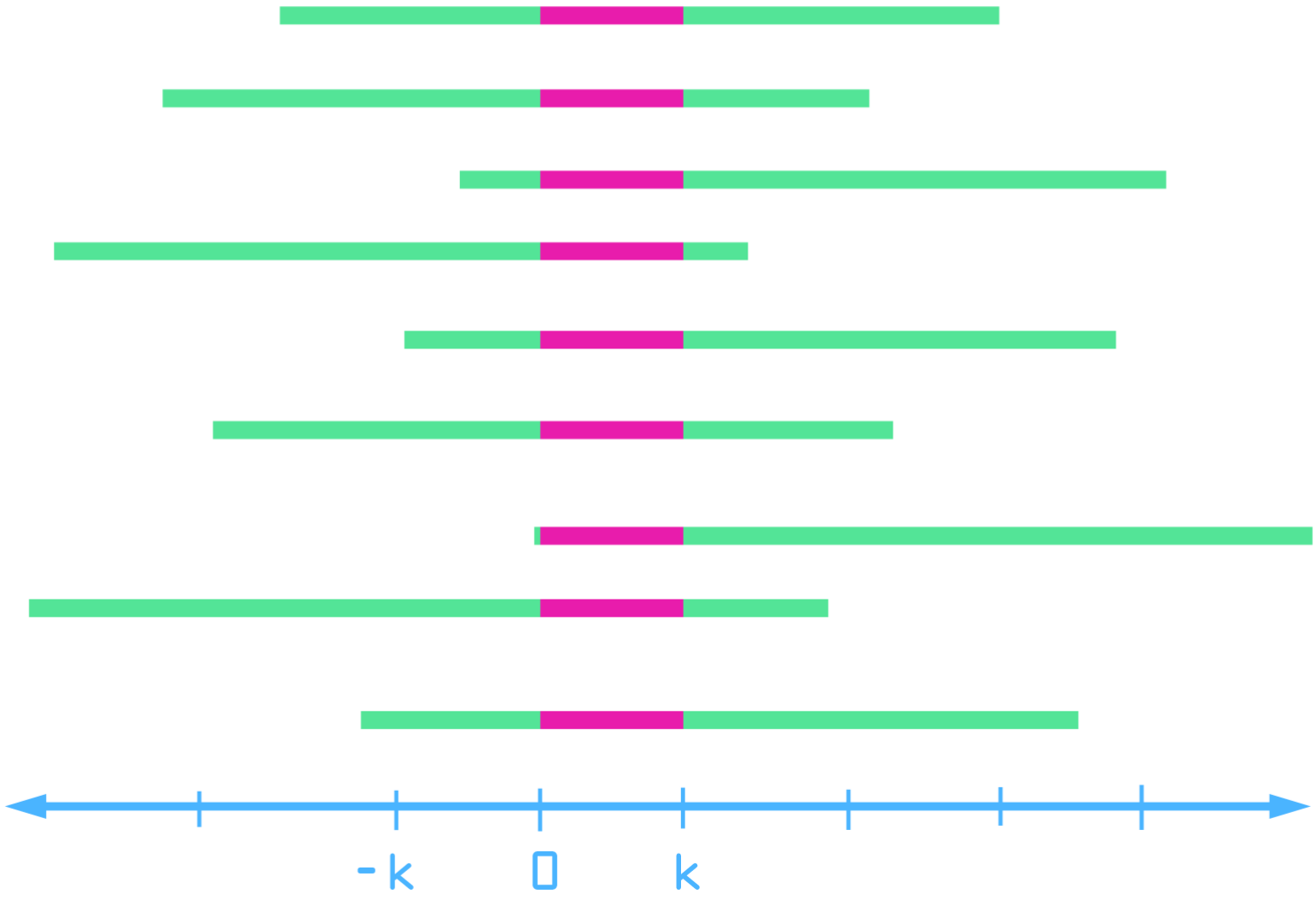


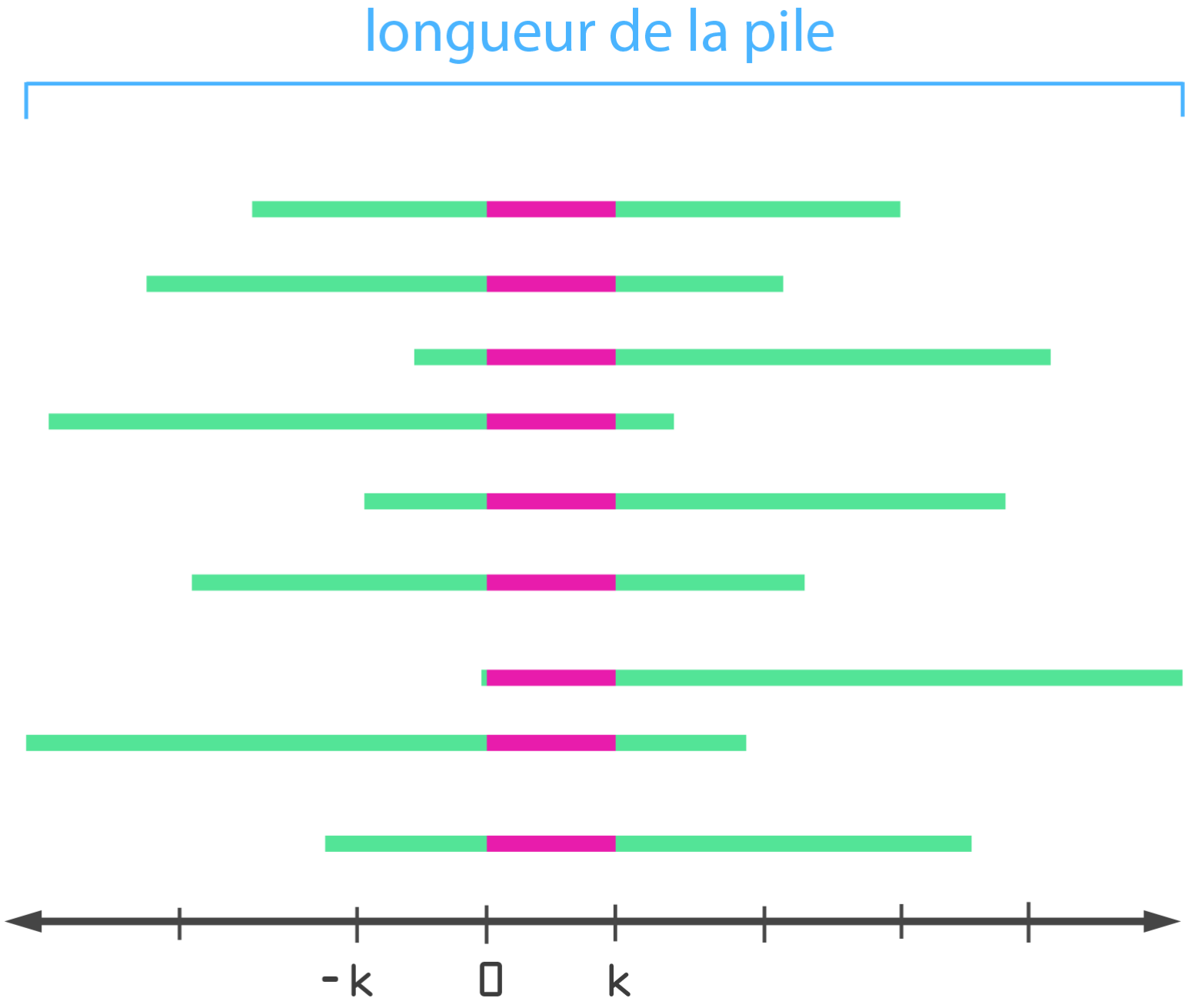

Q_2 : Stack read i at position p

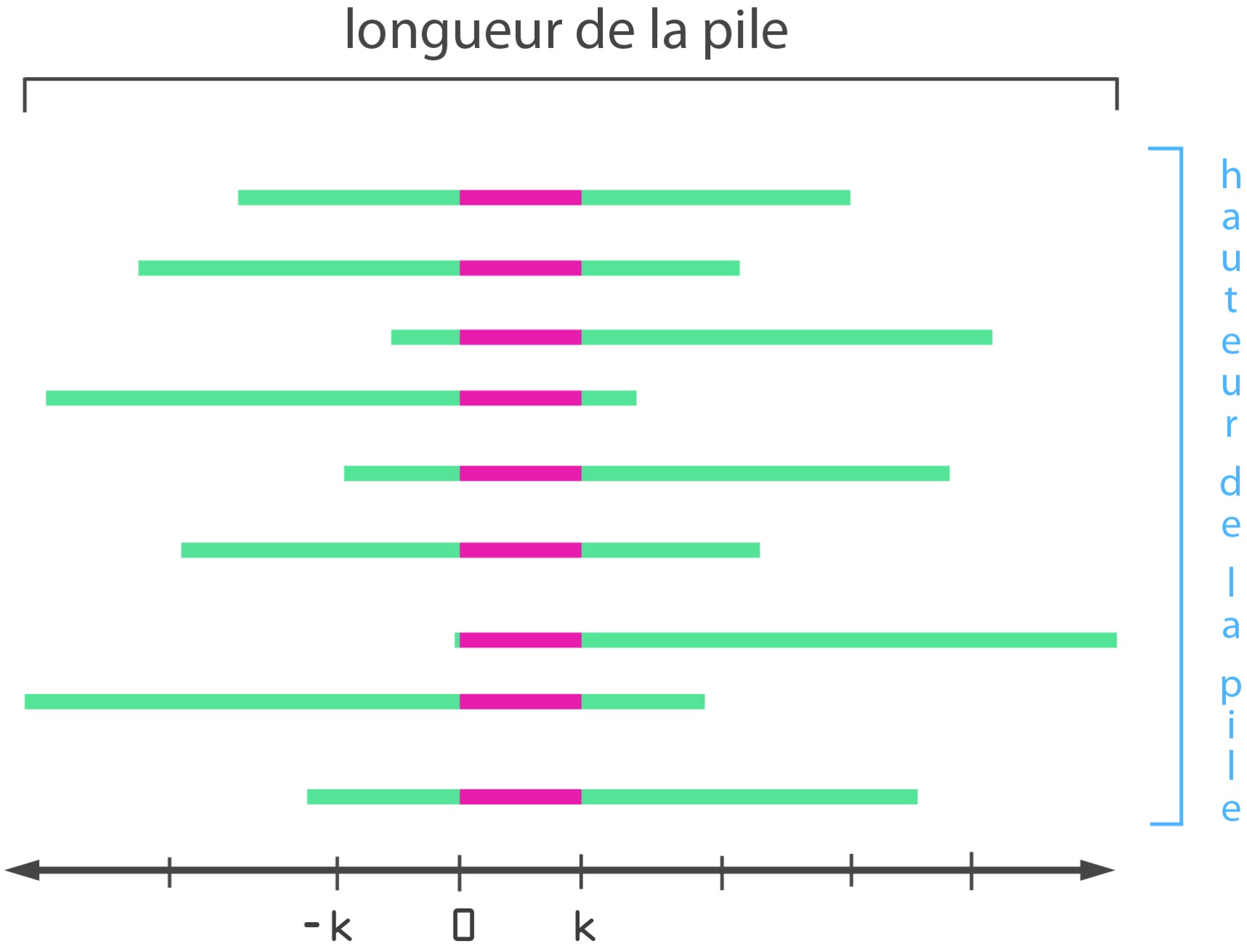

Q_2 : Stack read i at position p



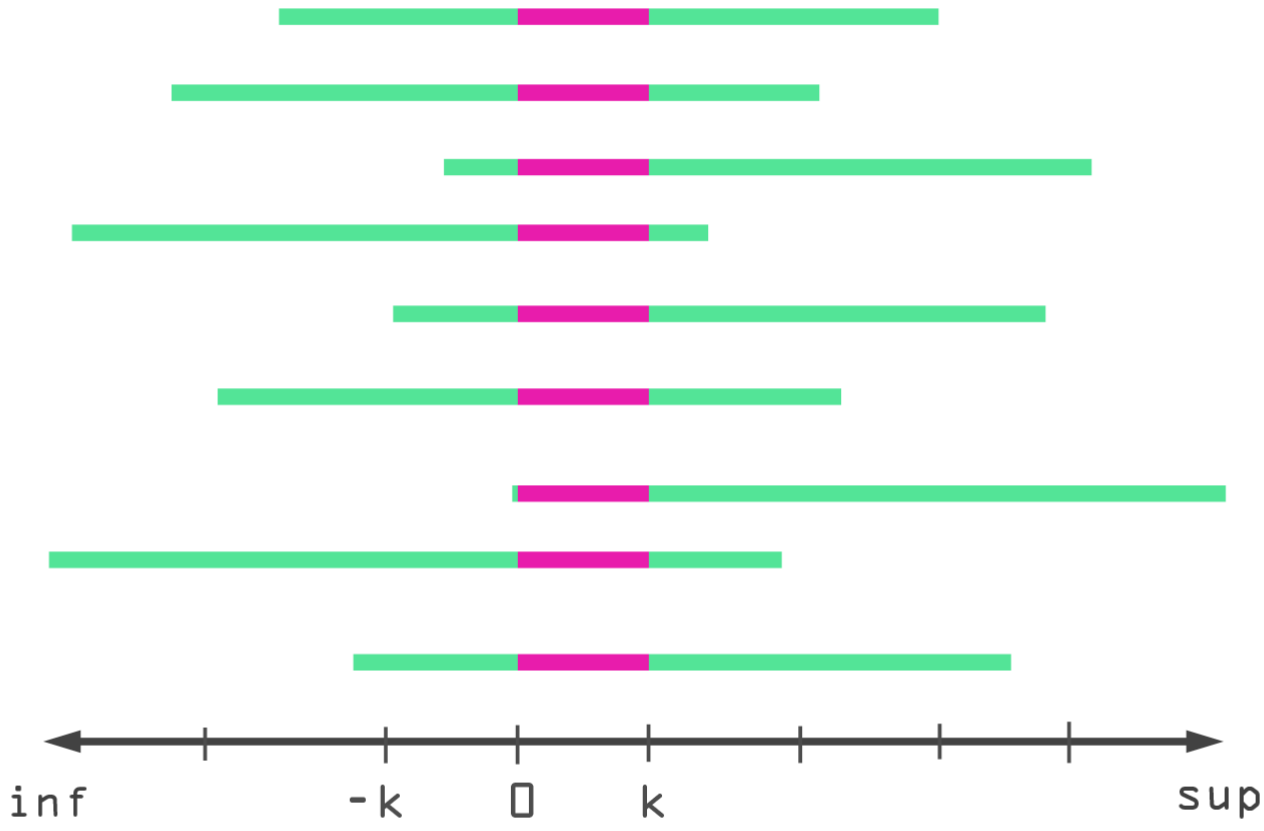


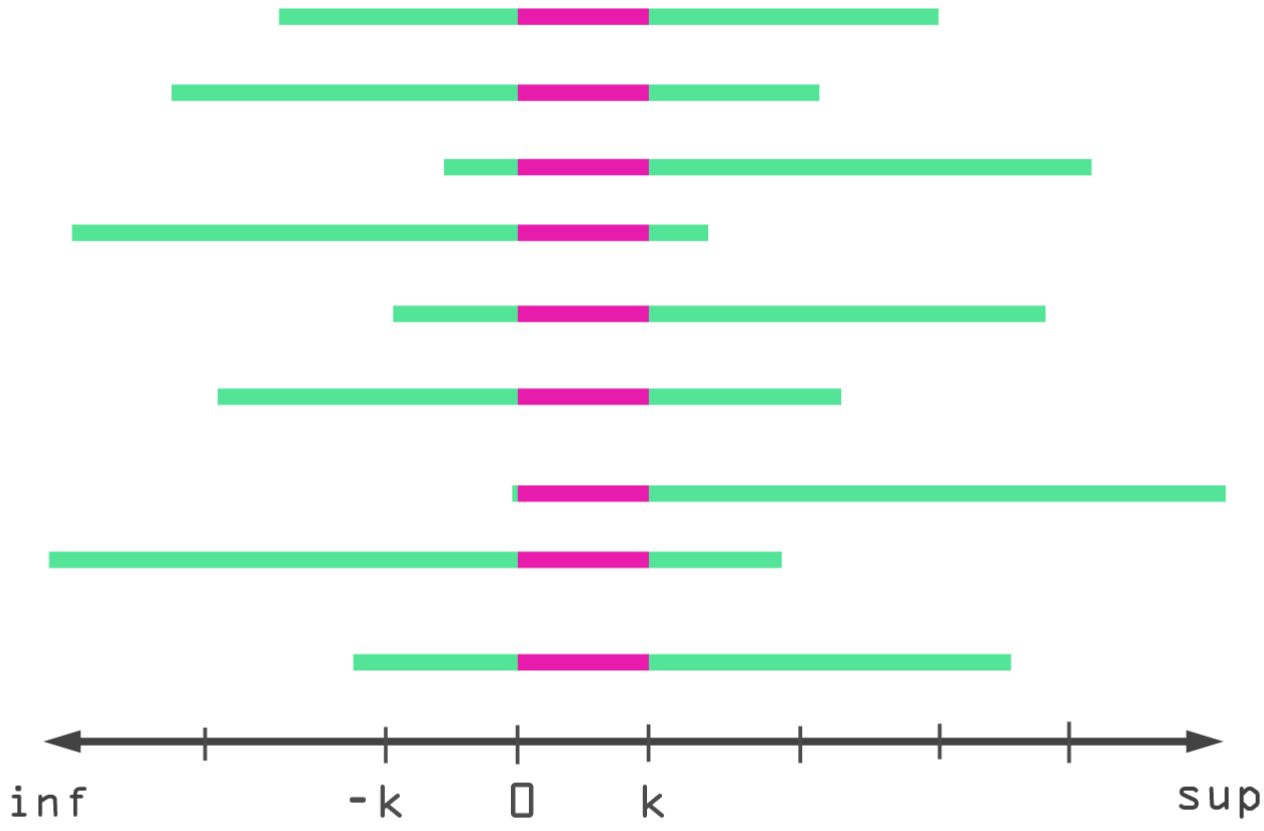

-k
0
k


longueur de la pile
-k
0
k

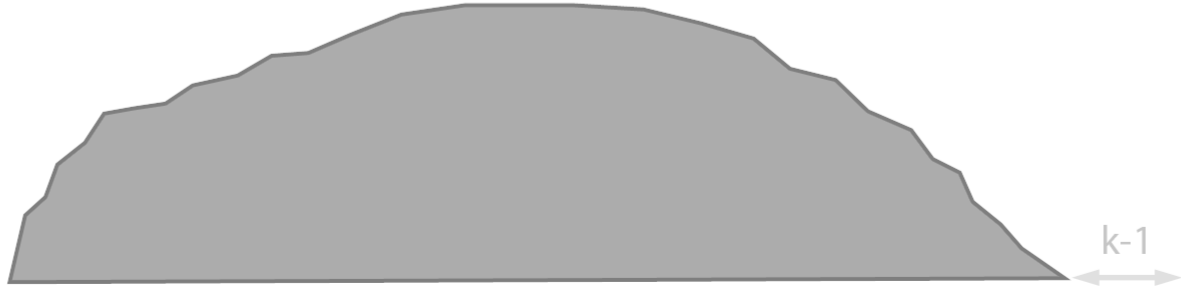

longueur de la pile
hauteur de la pile
-k
0
k

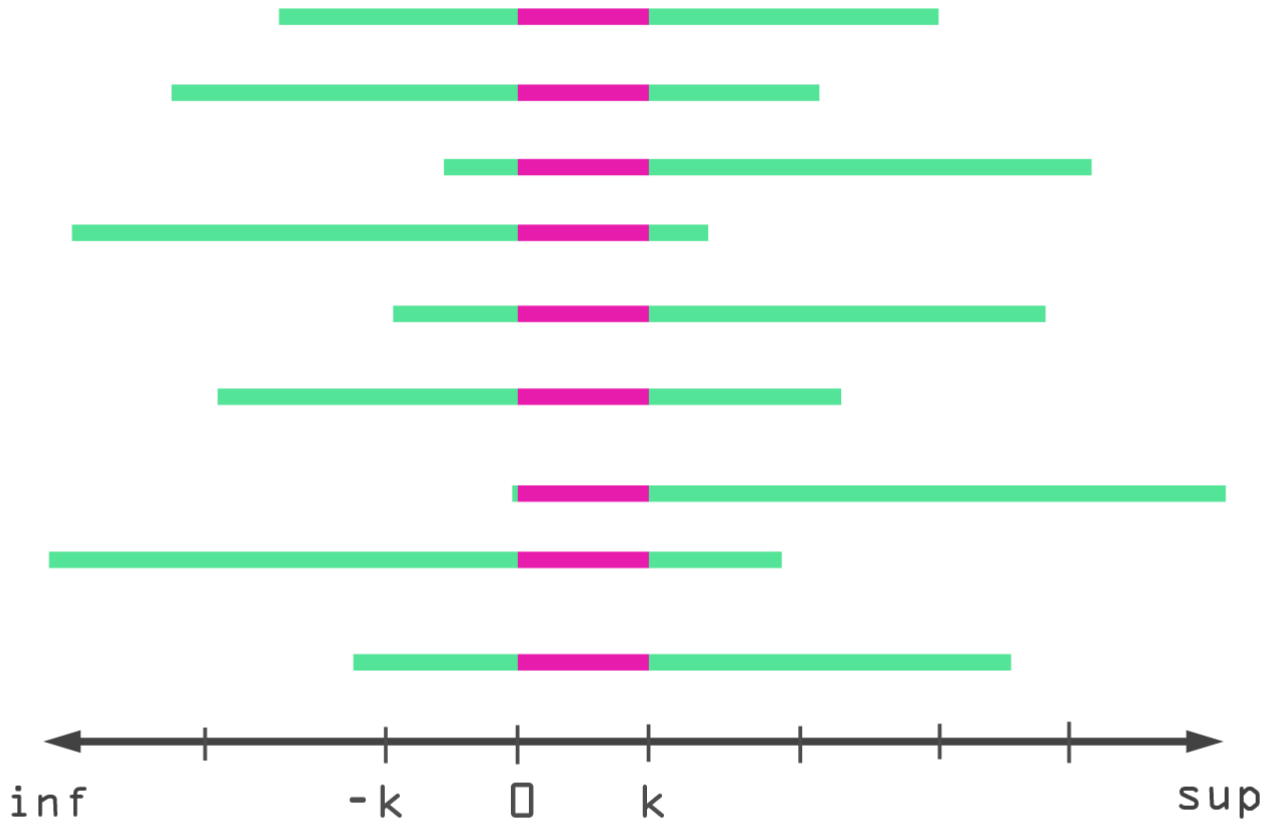
1. Choisir **un "bon" read** dans la collection
2. Choisir **un "bon" *k-mer*** pour créer l'empilement
3. Créer un **empilement** de reads
4. Définir **les bornes de confiance** de la pile
5. Choisir **les *k-mers* suivants** pour la procédure d'extension
6. **Marquer les reads** utilisés
7. **Terminer** la reconstruction du *meta-read* ou **continuer** à étendre


inf
-k
0
k
sup

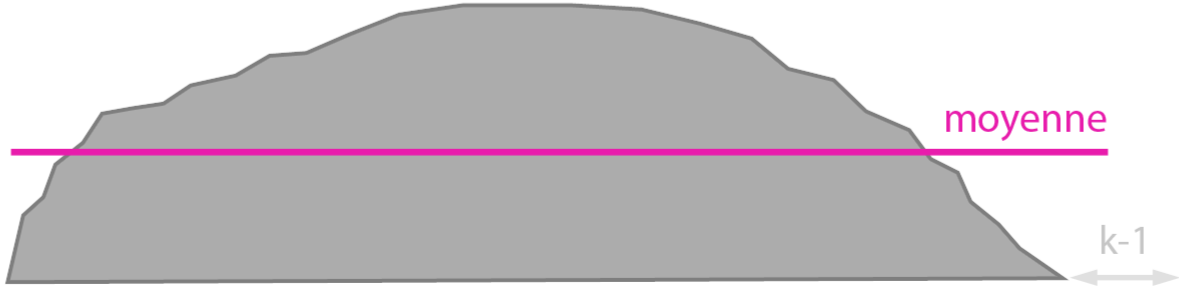


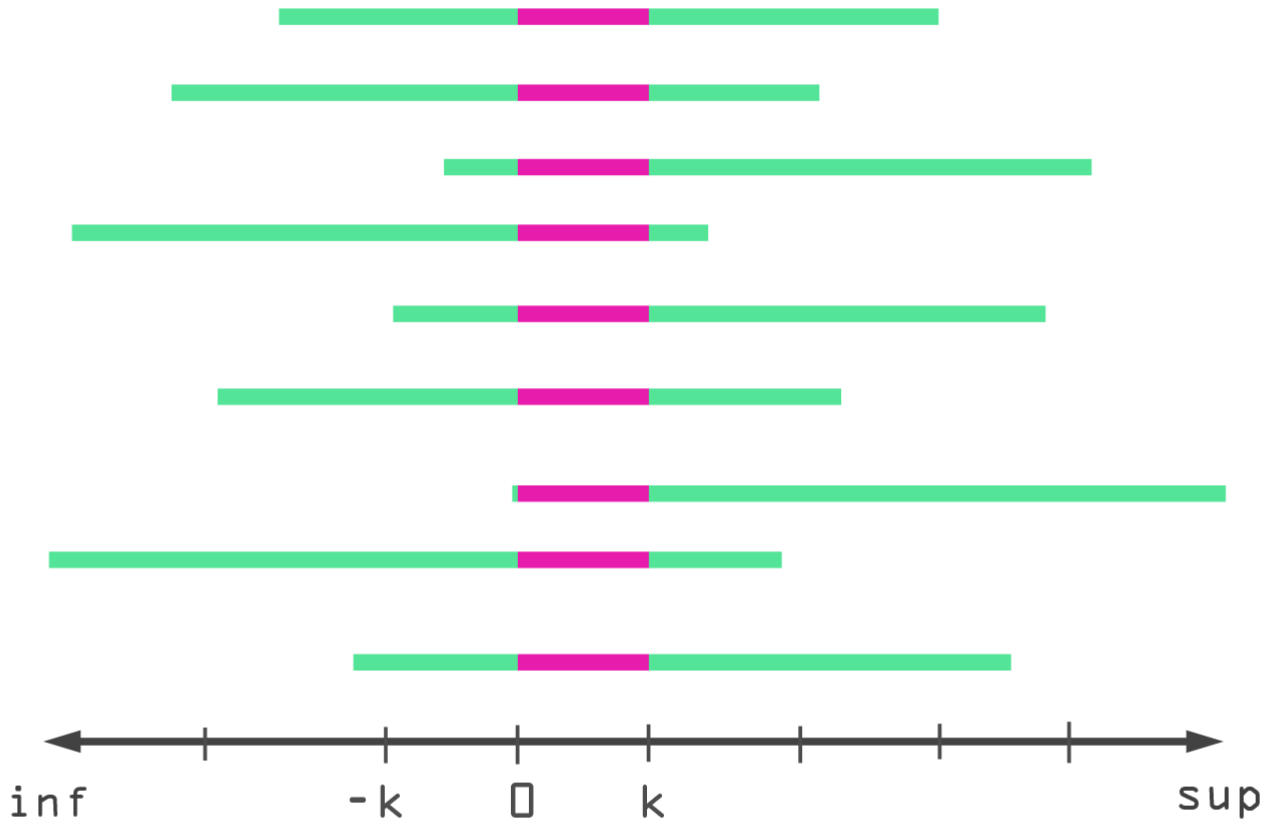
Profile de la pile



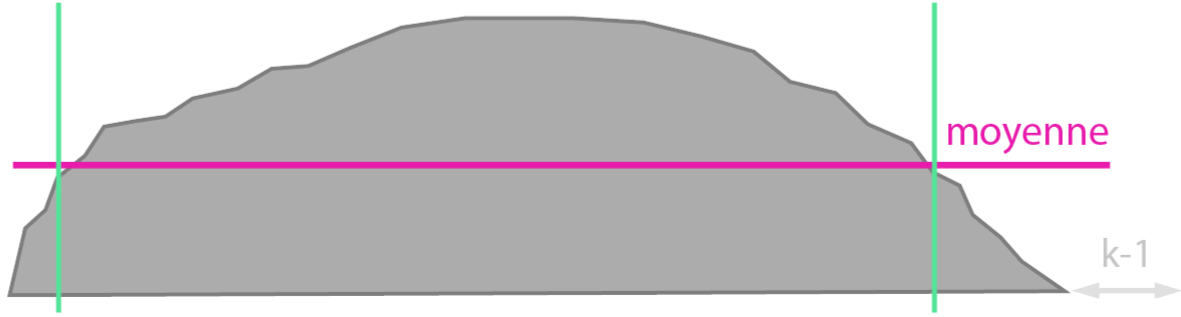


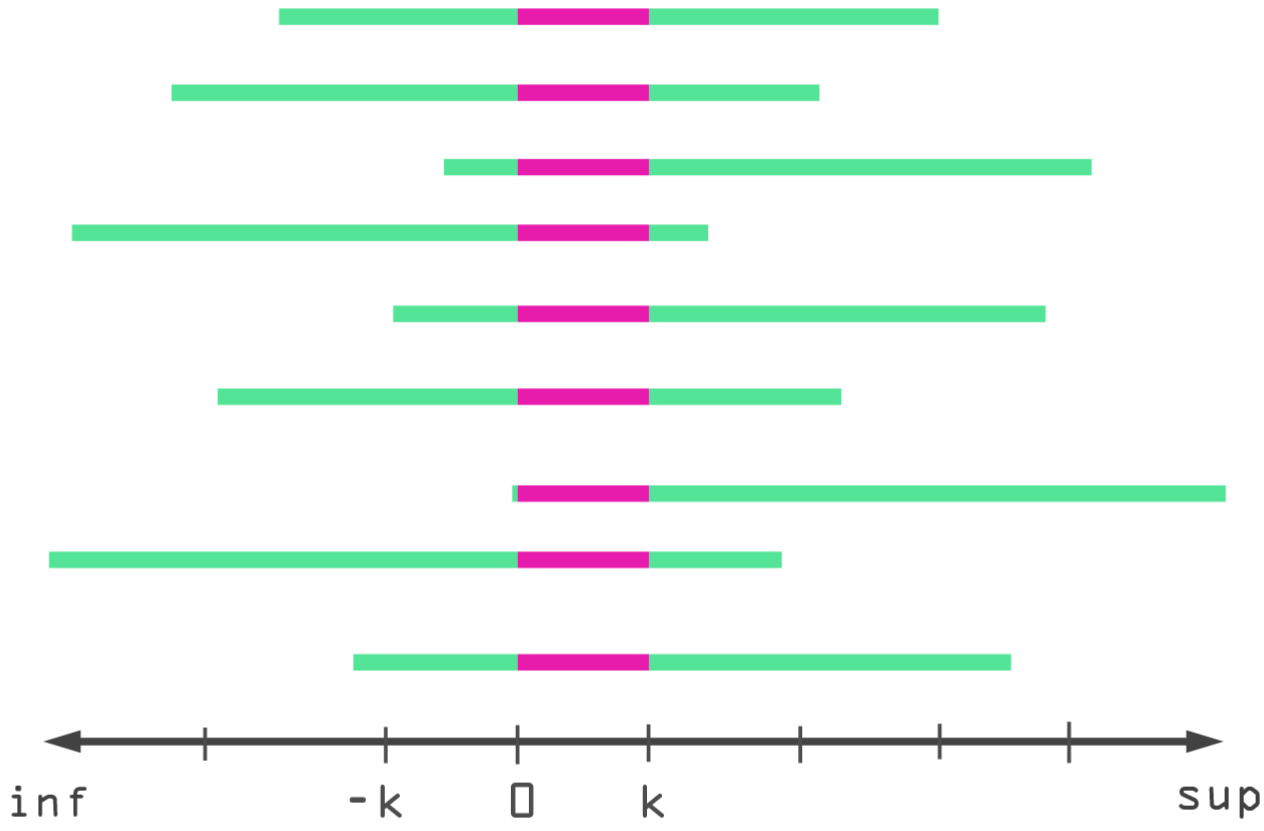
Profile de la pile



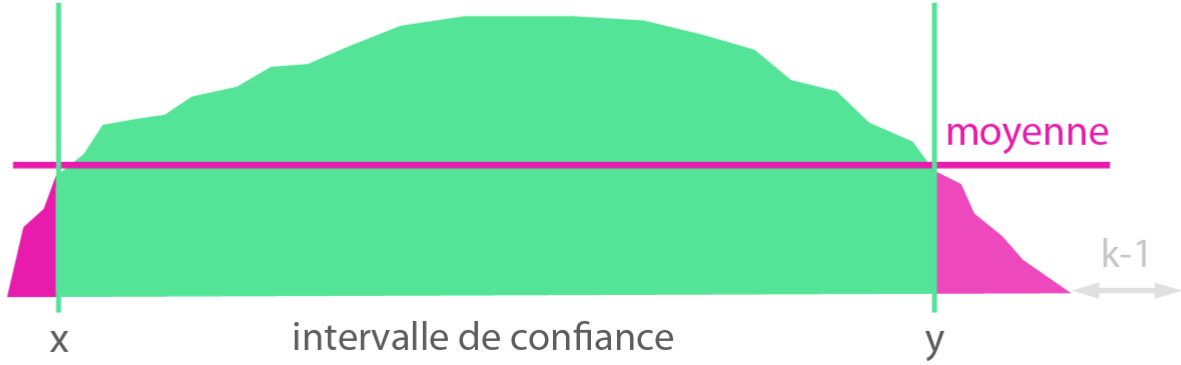


Profile de la pile

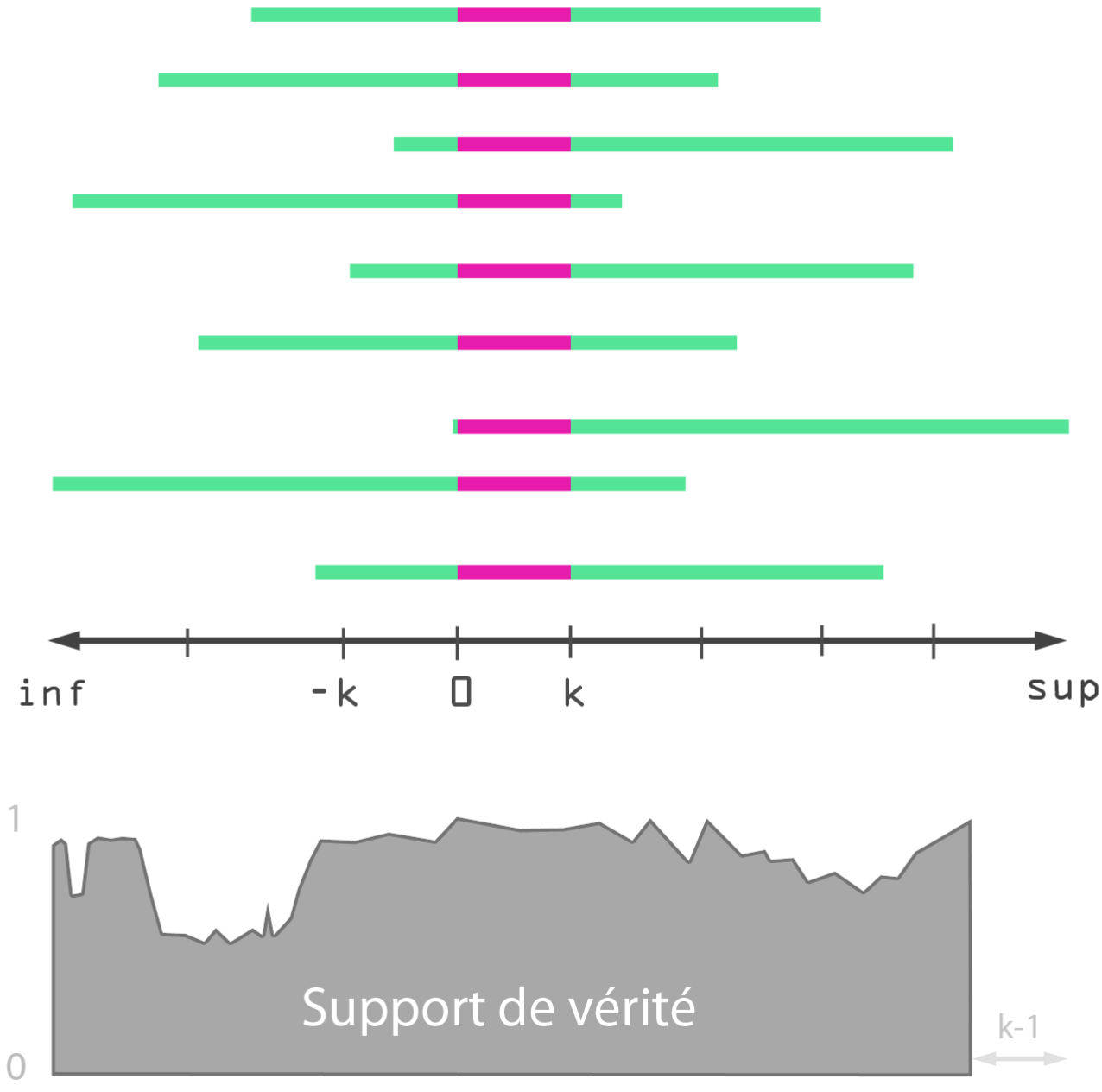


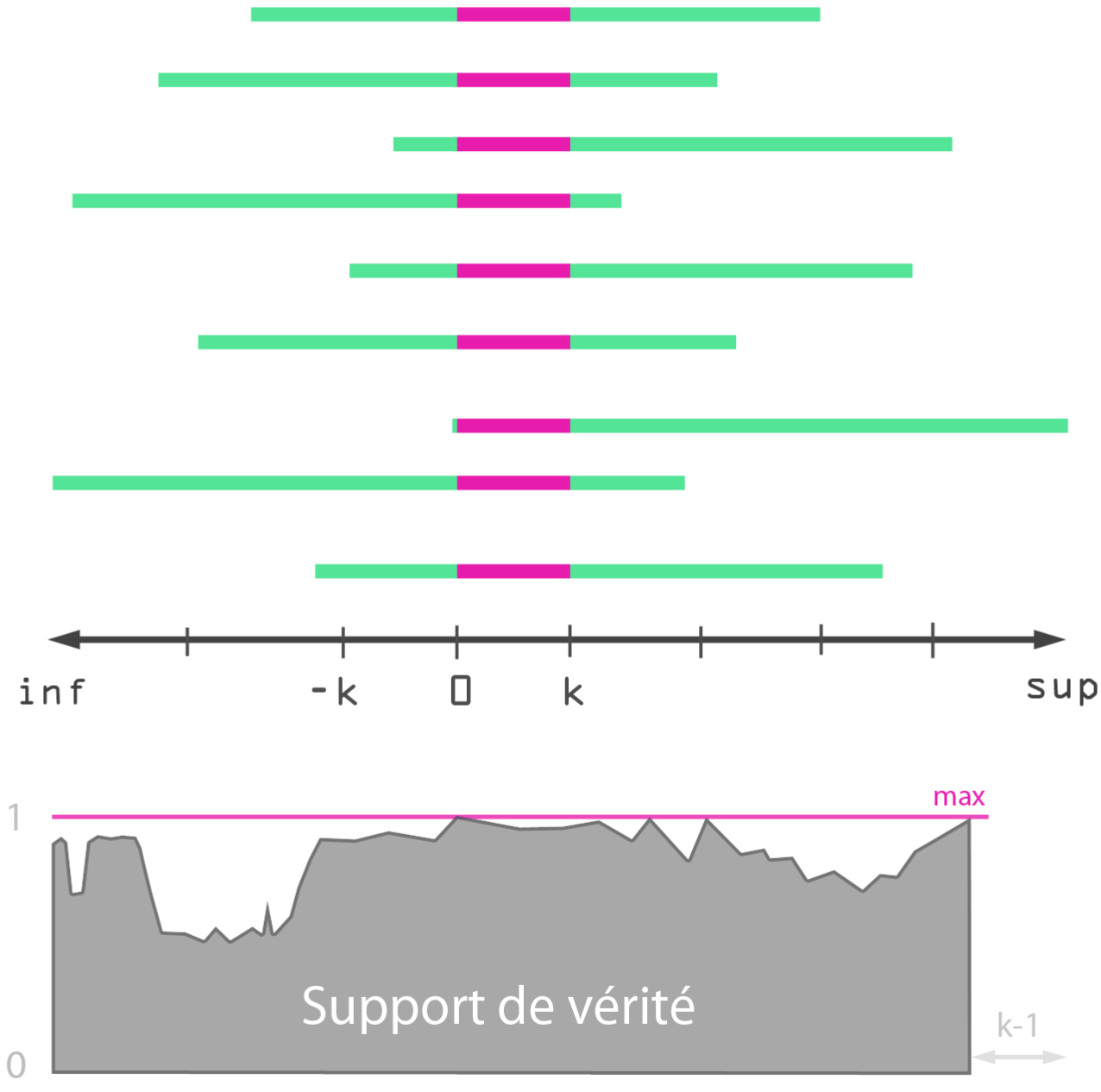


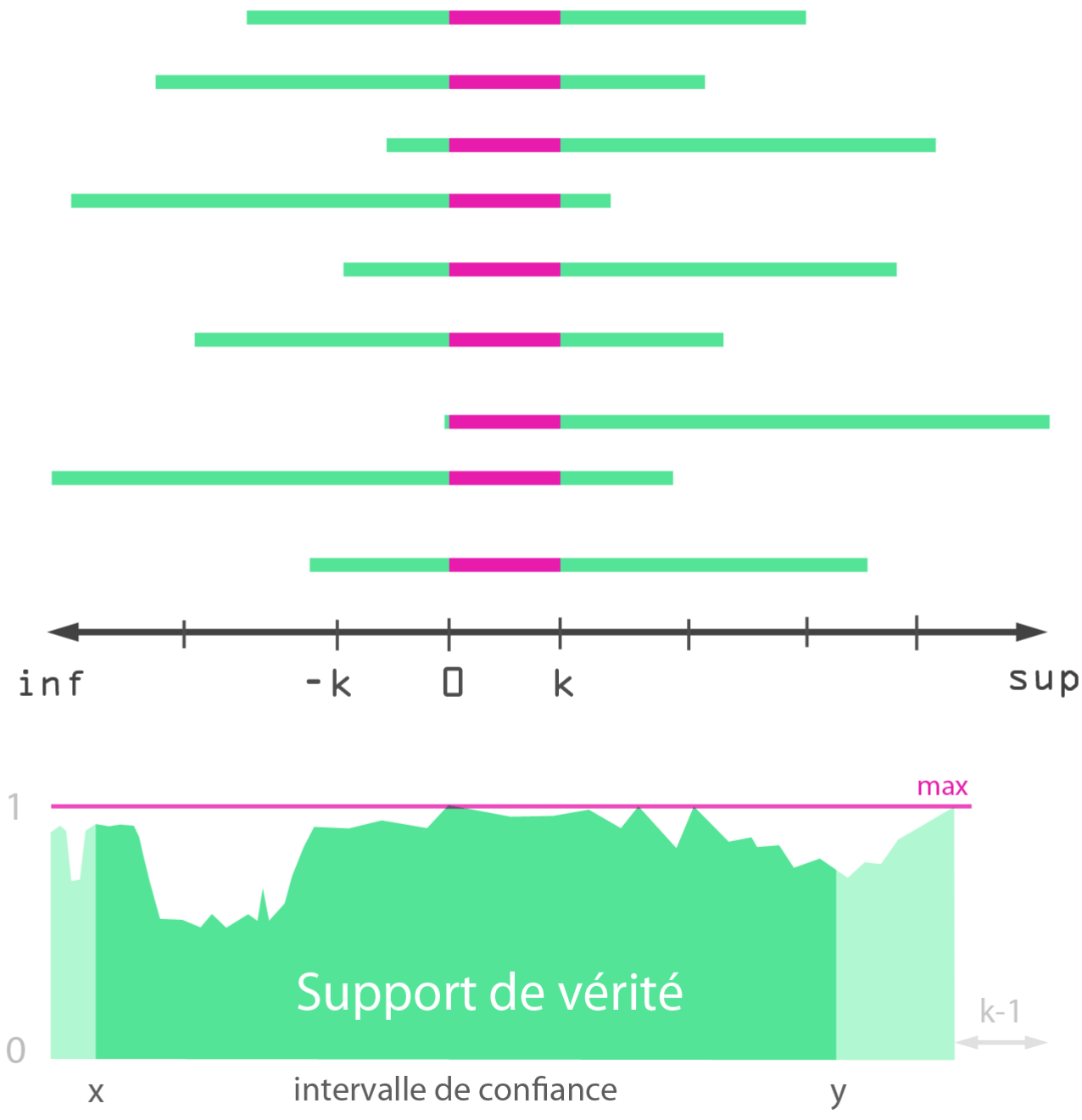
Profile de la pile

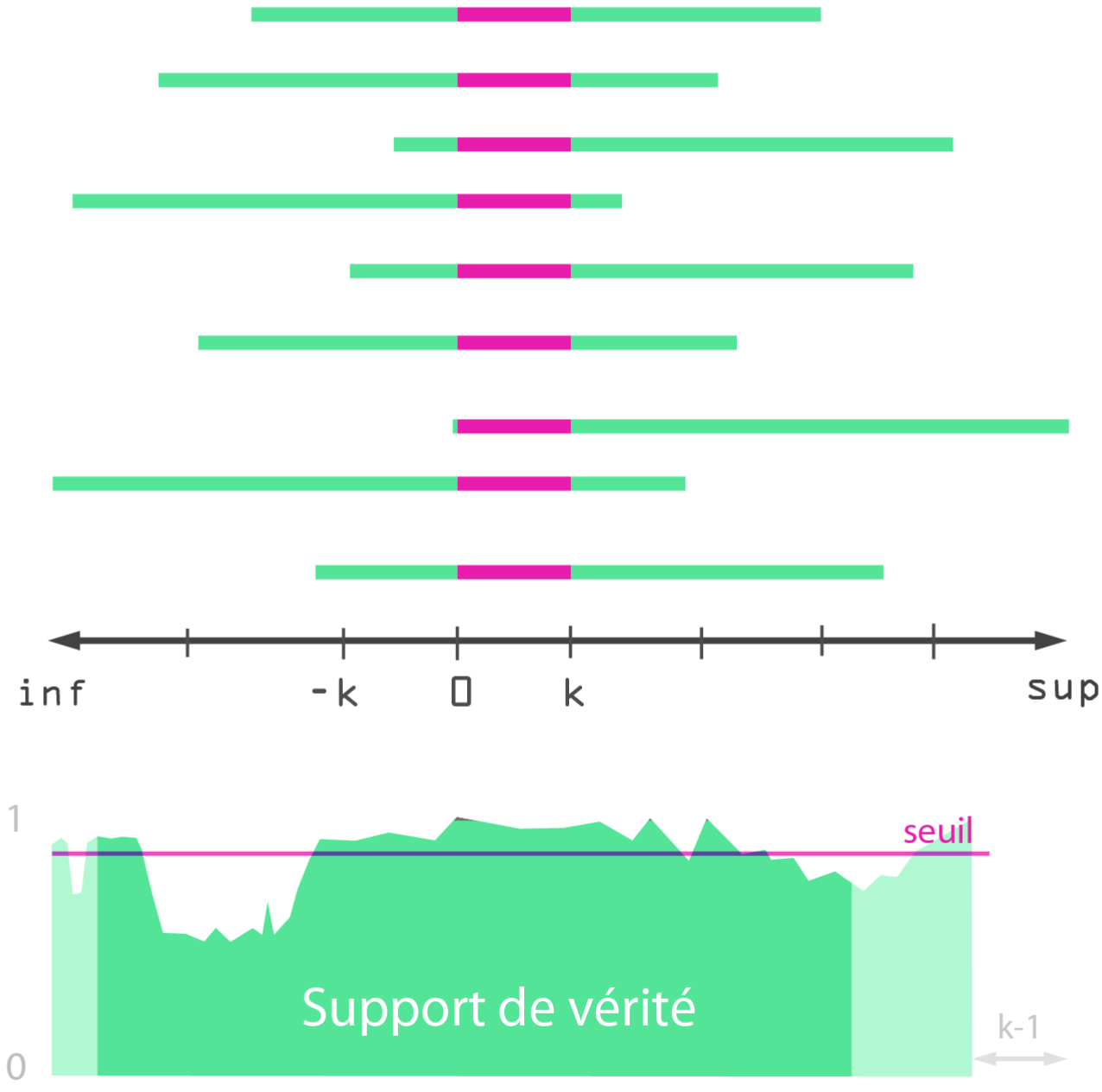


1. Choisir **un "bon" read** dans la collection
2. Choisir **un "bon" *k-mer*** pour créer l'empilement
3. Créer un **empilement** de reads
4. Définir **les bornes de confiance** de la pile
5. Choisir **les *k-mers* suivants** pour la procédure d'extension
6. **Marquer les reads** utilisés
7. **Terminer** la reconstruction du *meta-read* ou **continuer** à étendre

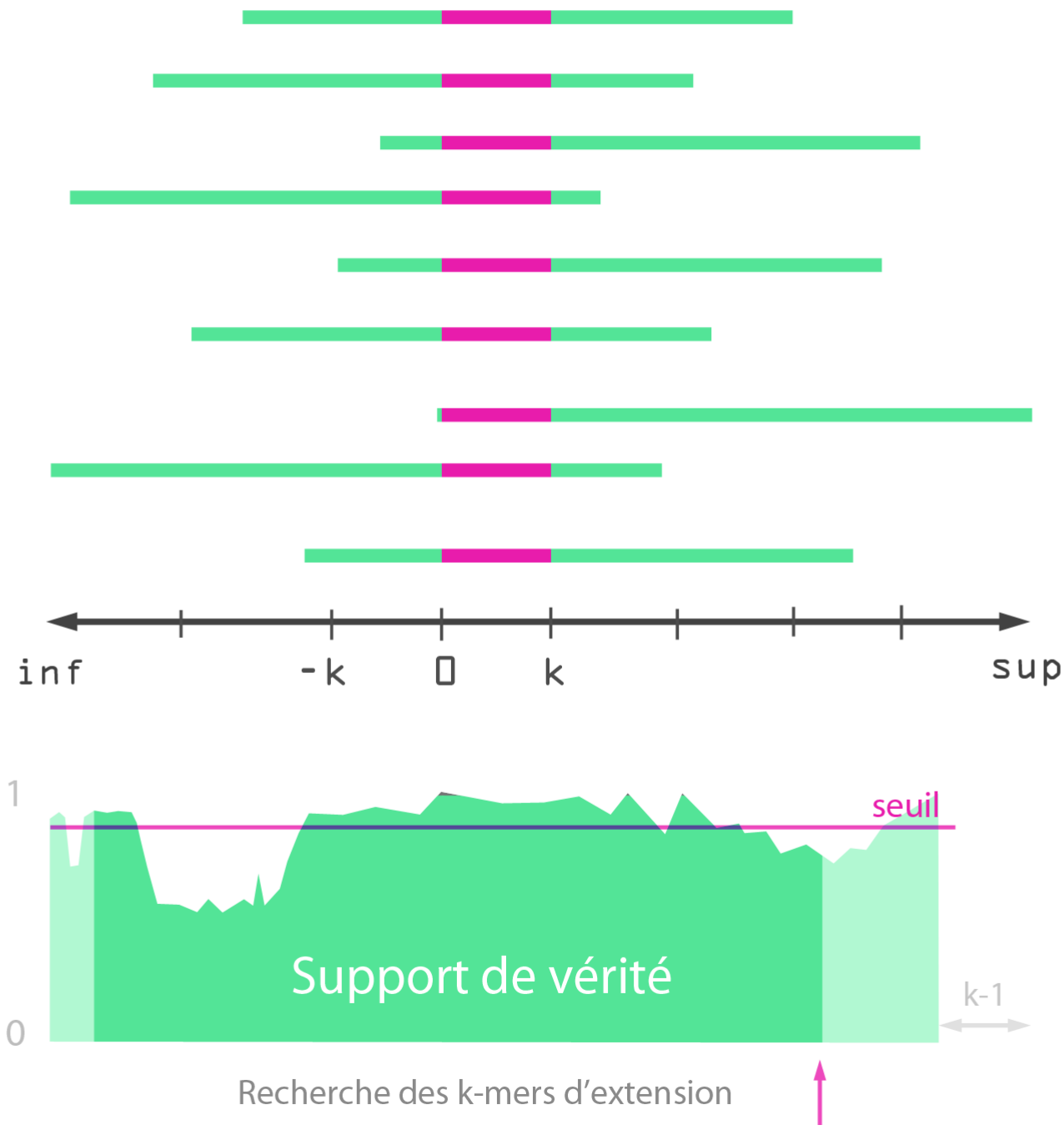

inf
-k
0
k
sup
1
0
Support de vérité
k-1

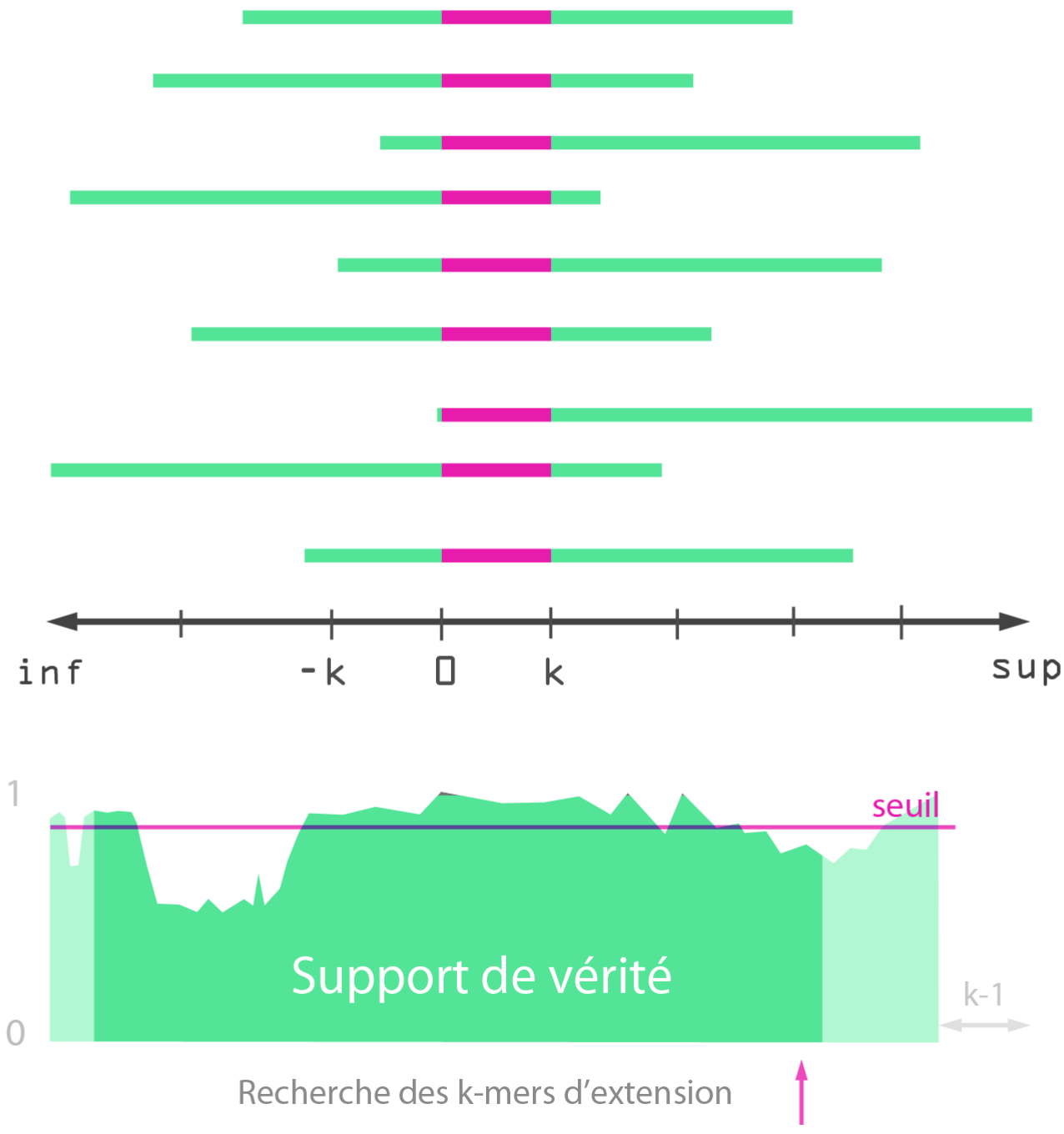

inf
-k
0
k
sup
1
max
Support de vérité
0
k-1

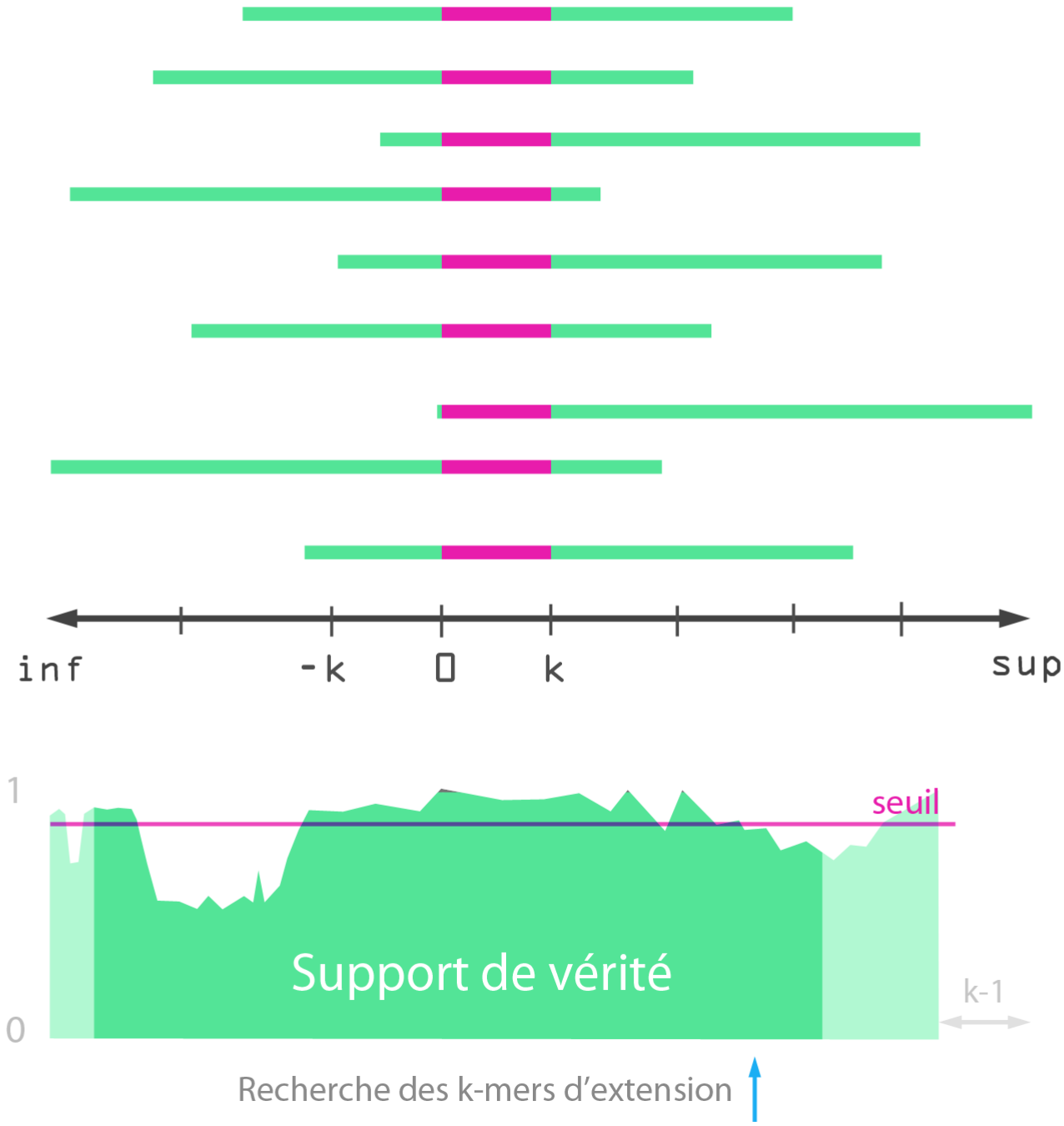

inf
-k
0
k
sup
1
max
Support de vérité
k-1
0
x
intervalle de confiance
y

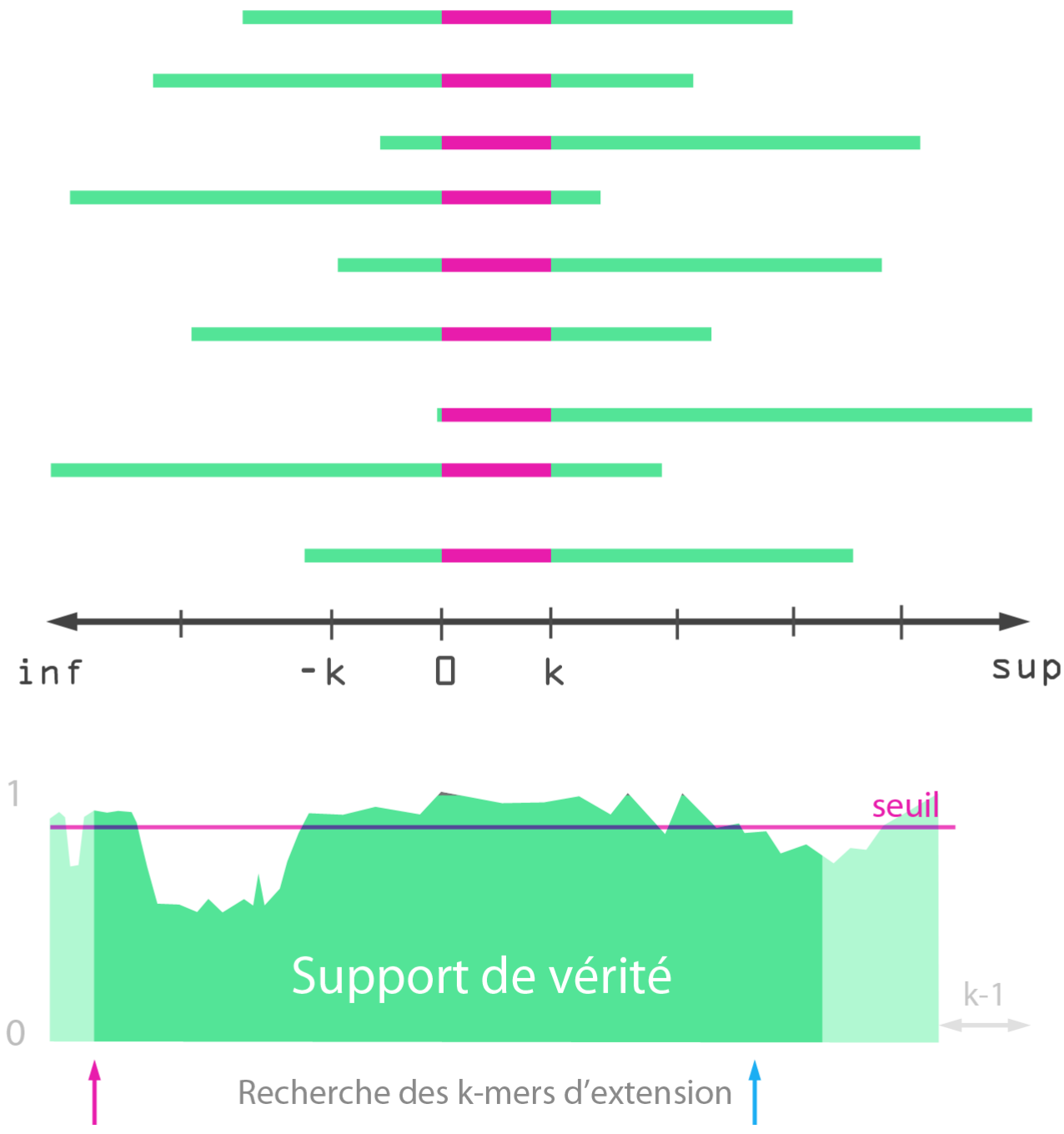


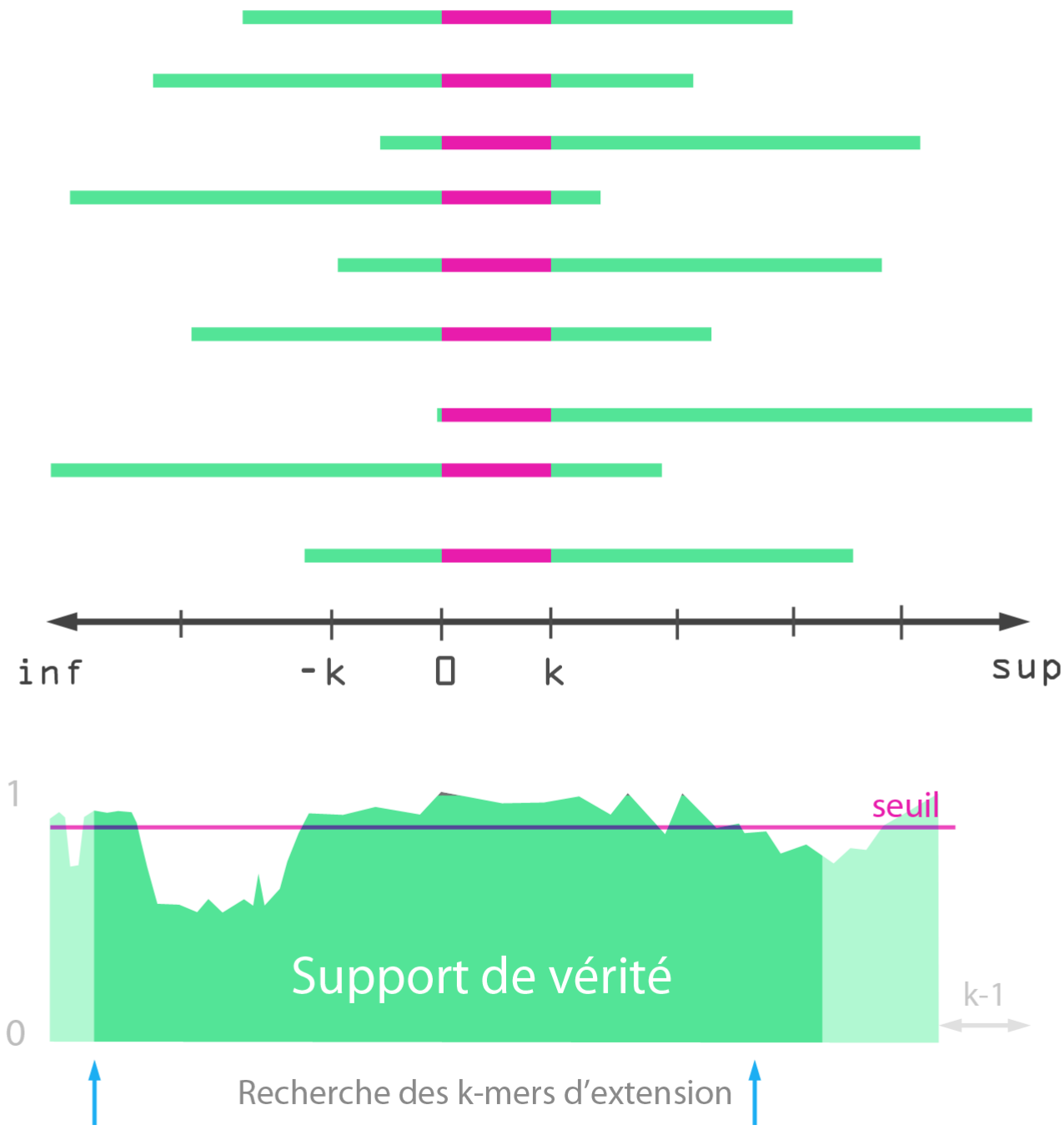
Recherche des k-mers d'extension

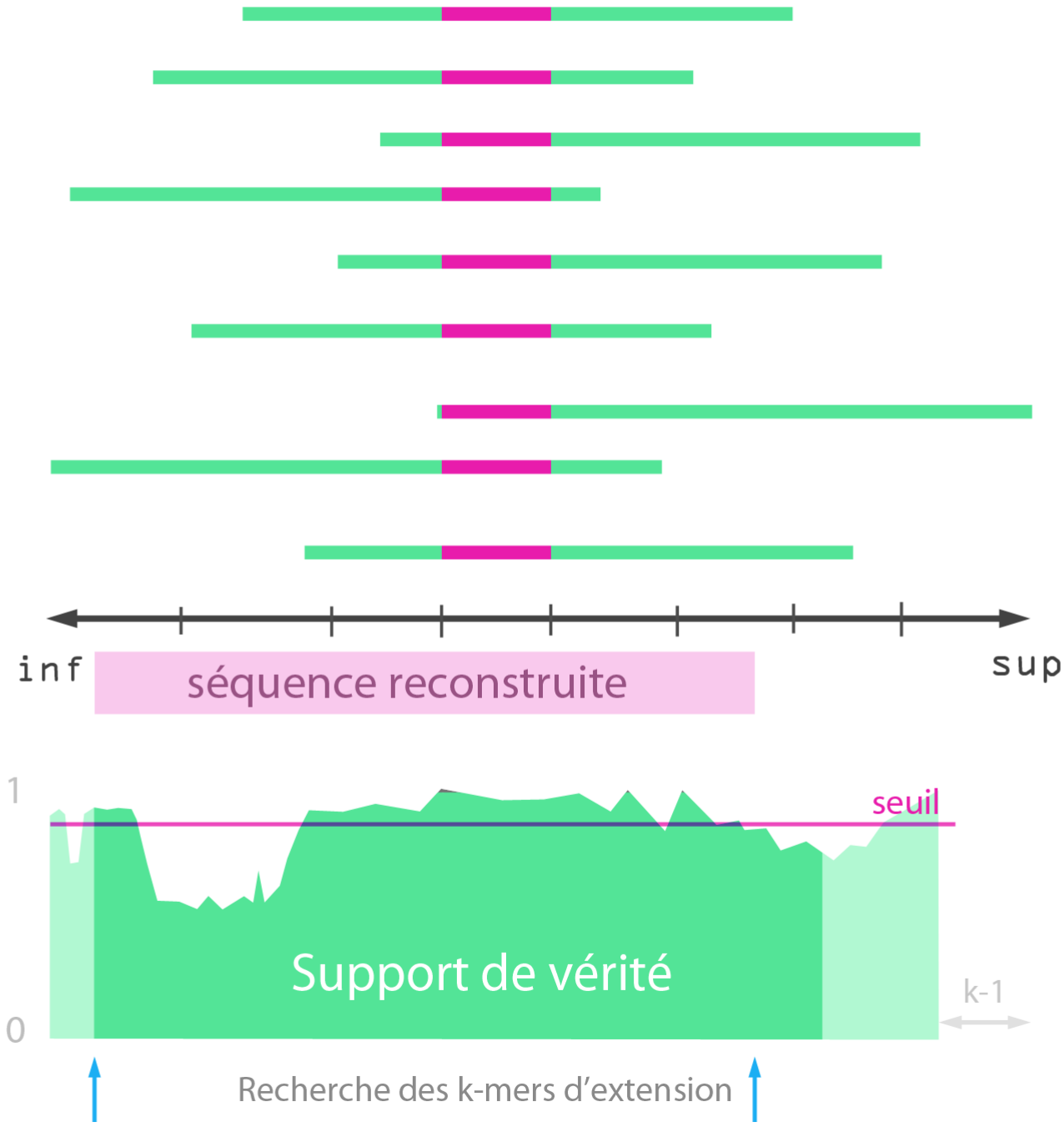

inf
-k
0
k
sup
1
seuil
Support de vérité
0
k-1
Recherche des k-mers d'extension


inf
-k
0
k
sup
1
seuil
Support de vérité
0
k-1
Recherche des k-mers d'extension

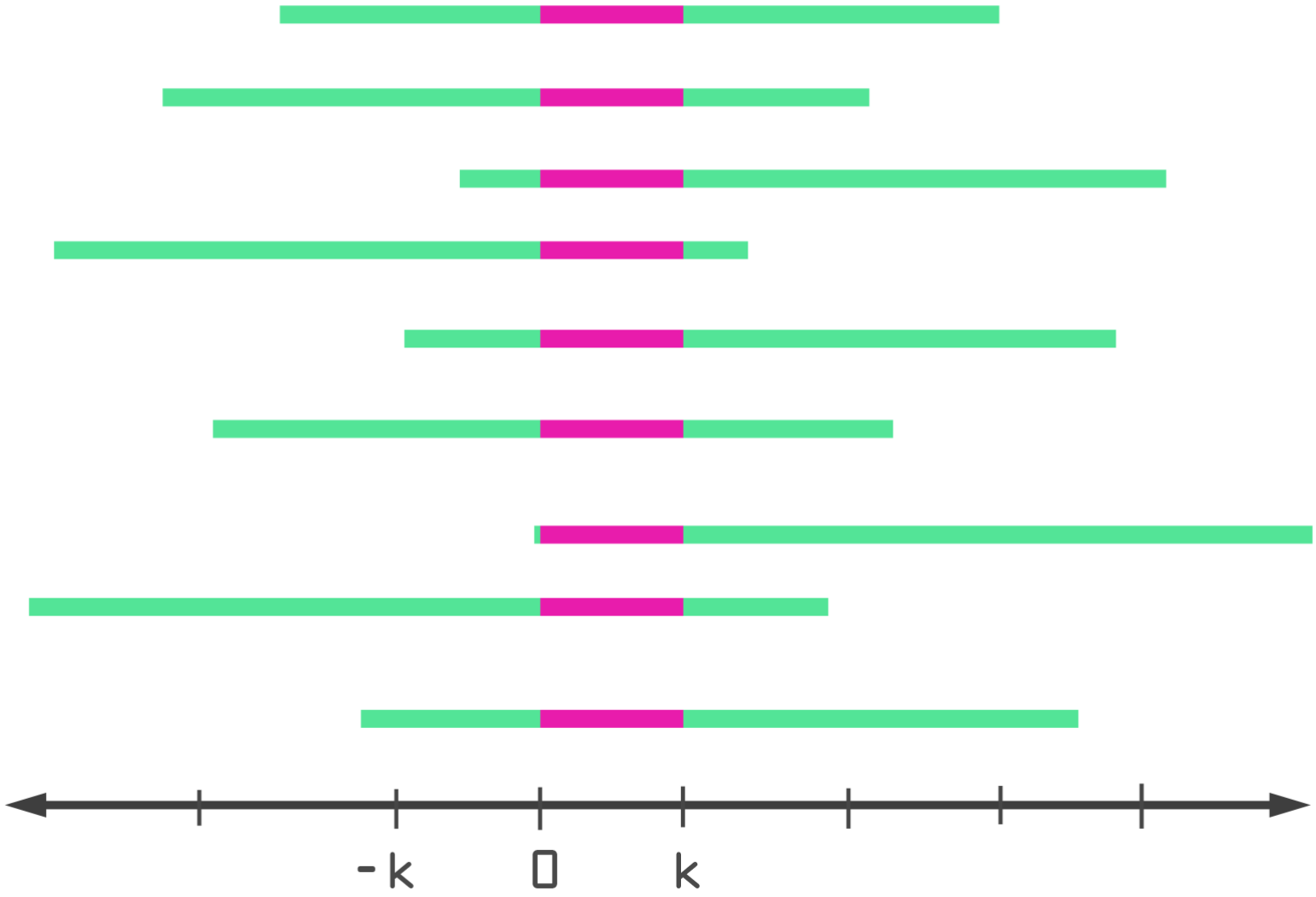

inf
-k
0
k
sup
1
seuil
Support de vérité
0
k-1
Recherche des k-mers d'extension

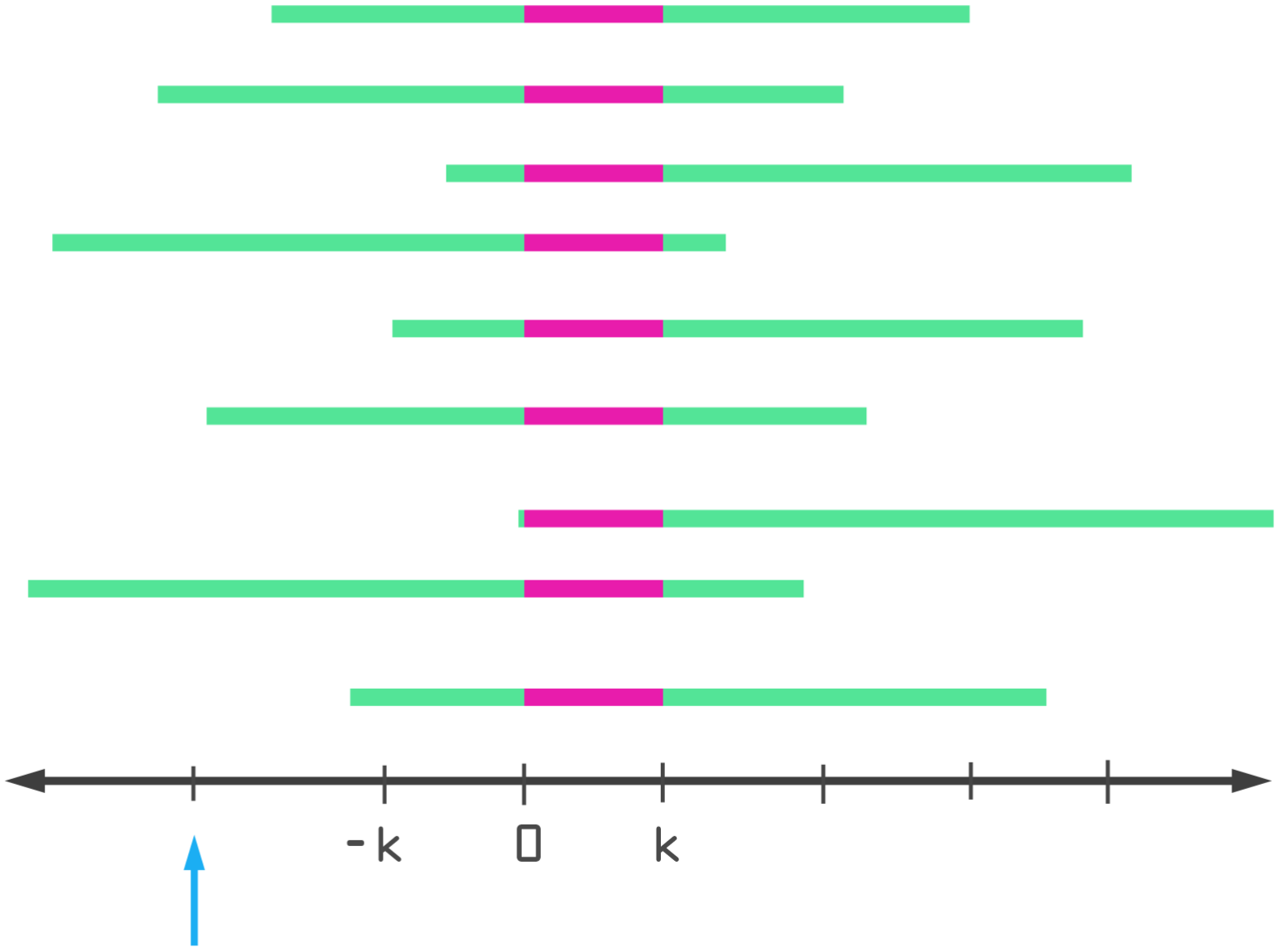

inf
-k
0
k
sup
1
seuil
Support de vérité
0
k-1
Recherche des k-mers d'extension

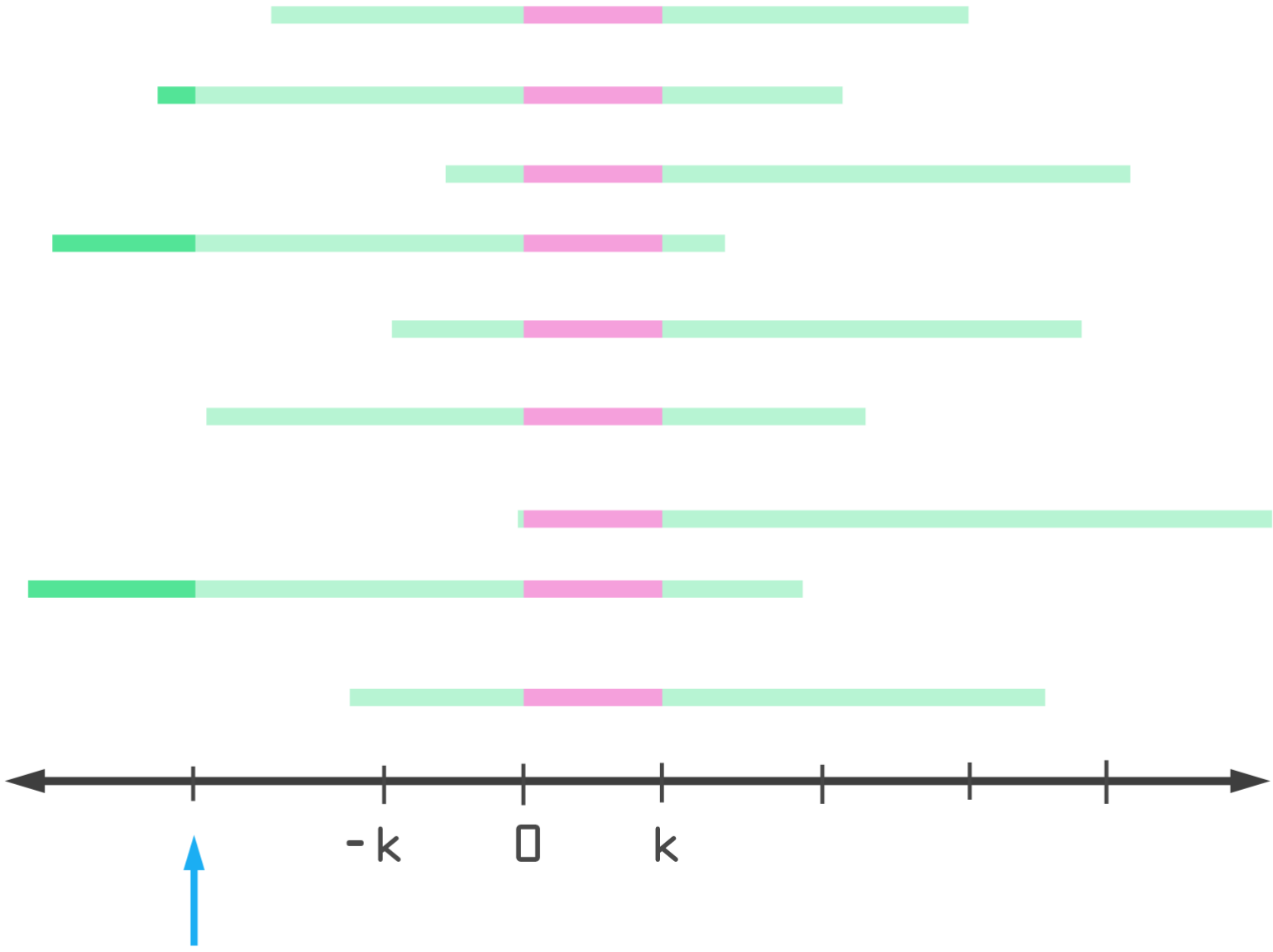

inf
-k
0
k
sup
1
seuil
Support de vérité
0
k-1
Recherche des k-mers d'extension

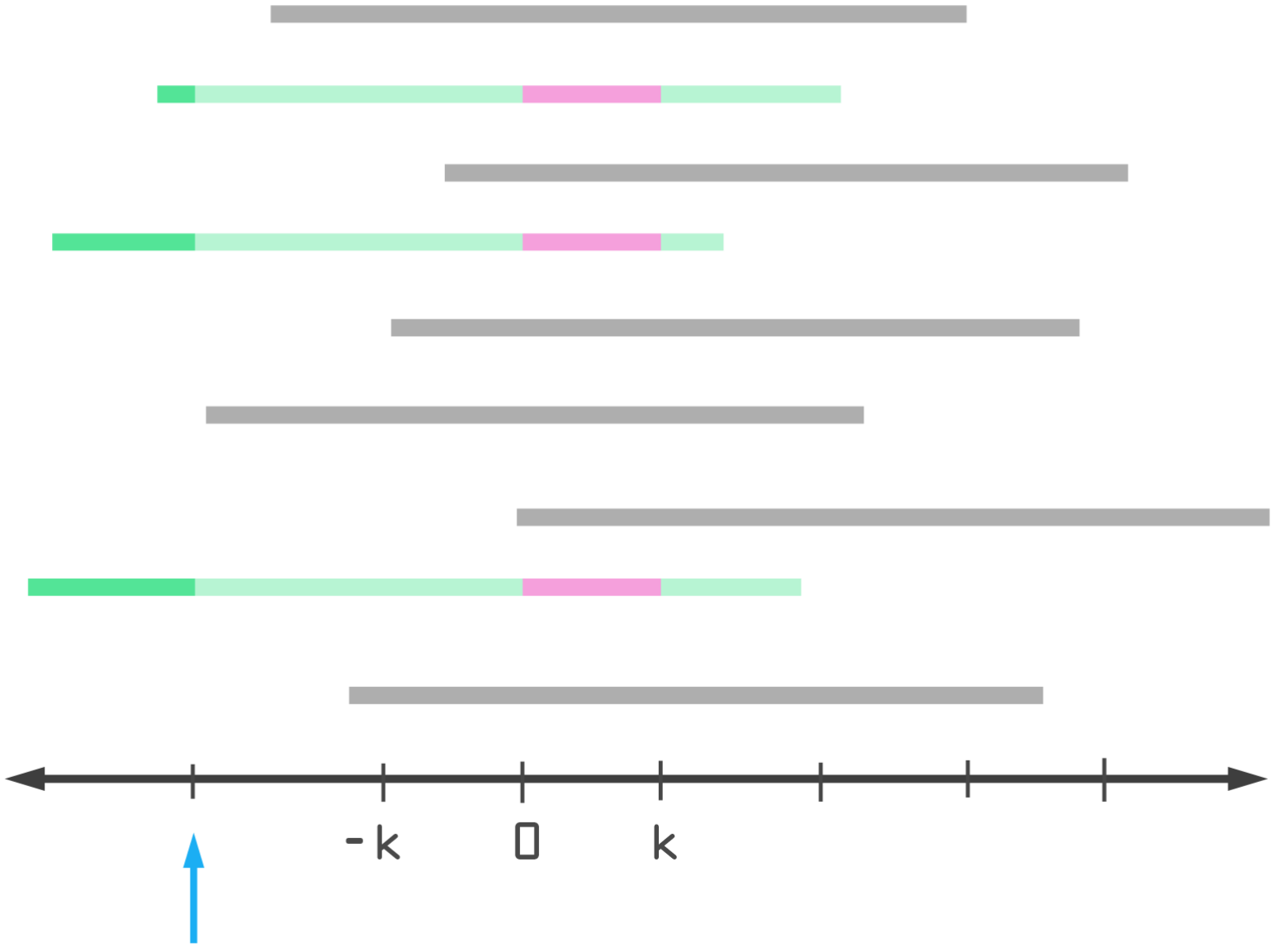

inf
séquence reconstruite
sup
1
seuil
Support de vérité
k-1
0
Recherche des k-mers d'extension

1. Choisir **un "bon" read** dans la collection
2. Choisir **un "bon" *k-mer*** pour créer l'empilement
3. Créer un **empilement** de reads
4. Définir **les bornes de confiance** de la pile
5. Choisir **les *k-mers* suivants** pour la procédure d'extension
6. **Marquer les reads** utilisés
7. **Terminer** la reconstruction du *meta-read* ou **continuer** à étendre


-k
0
k


-k
0
k


-k
0
k


-k
0
k

1. Choisir **un "bon" read** dans la collection
2. Choisir **un "bon" *k-mer*** pour créer l'empilement
3. Créer un **empilement** de reads
4. Définir **les bornes de confiance** de la pile
5. Choisir **les *k-mers* suivants** pour la procédure d'extension
6. **Marquer les reads** utilisés
7. **Terminer** la reconstruction du *meta-read* ou **continuer** à étendre

Résultats

- Une nouvelle méthode de reconstruction
- Échelle du read, approche intégrée permise
- Méthode tournée vers l'avenir
- Outils de développement :
  - **GkDump** : sérialisation des GkArrays
  - **GkServer** : Interface ligne de commande pour questionner les reads indexés
- Implémentation C++
  - **Architecture** évolutive (*OOP*)
  - **Portabilité** (packaging *Autotools*)

- Résultats préliminaires encourageants
- Performances prometteuses : < 2 *minutes pour assembler 12M de reads (75pb) avec 1 thread*
- Perspectives à court terme : **Amélioration de la méthode**
  - Assemblage des meta-reads en transcrits
  - Améliorations algorithmiques de la méthode
  - Mesures de sensibilité/précision
  - Structure de données optimisée en mémoire
- Perspectives à long terme : **Une publication scientifique**

The end!