RECOMBINAISON VDJ

Réduire les comparaisons des gènes V avec Aho-Corasick

Cyprien Borée

Avril 2018 – Juin 2018

Entreprise : Fondation INRIA, Domaine du Voluceau – Rocquencourt, 78153, Le Chesnay Cedex
Université : Université de Lille 1, Cité Scientifique, 59650, Villeneuve d'Ascq
Formation initiale : Licence 3 Informatique, parcours Info
Tuteur à l'université : Pierre Allegraud
Tuteurs en entreprise : Mikaël Salson et Mathieu Giraud

Introduction

Introduction

- Label recherche : janvier 2018 à mars 2018

Introduction

- Label recherche : janvier 2018 à mars 2018
  - Documentation sur des nouvelles thématiques
    - Notions de biologie

Introduction

- Label recherche : janvier 2018 à mars 2018
  - Documentation sur des nouvelles thématiques
    - Notions de biologie
    - Programmation dynamique

Introduction

- Label recherche : janvier 2018 à mars 2018
  - Documentation sur des nouvelles thématiques
    - Notions de biologie
    - Programmation dynamique
    - Automate d’Aho-Corasick

Introduction

- Label recherche : janvier 2018 à mars 2018
  - Documentation sur des nouvelles thématiques
    - Notions de biologie
    - Programmation dynamique
    - Automate d’Aho-Corasick
  - Première expérience sur *VIDJIL*

Introduction

- Label recherche : janvier 2018 à mars 2018
  - Documentation sur des nouvelles thématiques
    - Notions de biologie
    - Programmation dynamique
    - Automate d’Aho-Corasick
  - Première expérience sur *VIDJIL*
- Stage : avril 2018 à juin 2018

Plan

Plan

- Cadre du stage

Plan

- Cadre du stage
- Aho-Corasick

Plan

- Cadre du stage
- Aho-Corasick
- Exemple d’utilisation

Plan

- Cadre du stage
- Aho-Corasick
- Exemple d’utilisation
- Résultats

Plan

- Cadre du stage
- Aho-Corasick
- Exemple d’utilisation
- Résultats
- Conclusion

Cadre du stage : VIDJIL

Cadre du stage : VIDJIL

- Rôle du logiciel : Analyse de séquences ADN

Cadre du stage : VIDJIL

- Rôle du logiciel : Analyse de séquences ADN
- Partenaires : Une dizaine d’hôpitaux dans le monde

Cadre du stage : VIDJIL

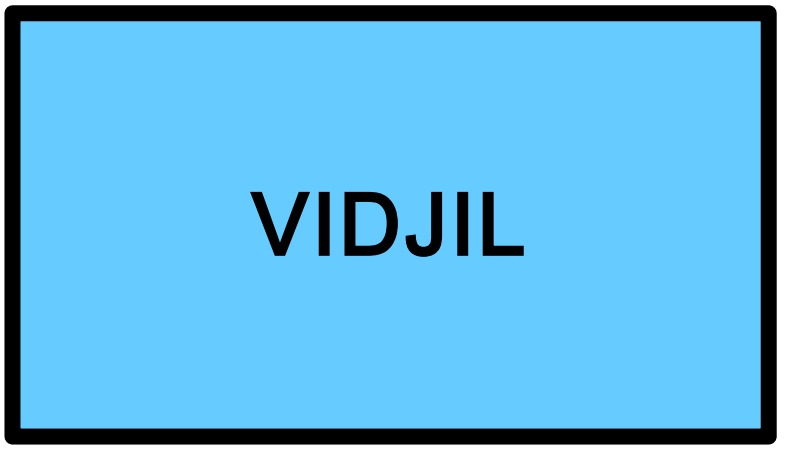
- Rôle du logiciel : Analyse de séquences ADN
- Partenaires : Une dizaine d’hôpitaux dans le monde
- Support de développement : C/C++, Python et Javascript

Cadre du stage : VIDJIL

- Rôle du logiciel : Analyse de séquences ADN
- Partenaires : Une dizaine d’hôpitaux dans le monde
- Support de développement : C/C++, Python et Javascript
- Mission de stage : Accélérer la comparaison de séquences au sein du logiciel

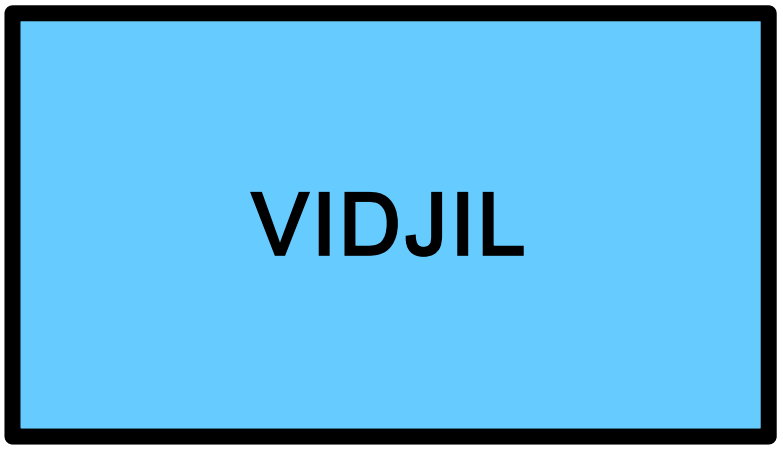
Rôle du logiciel

Rôle du logiciel

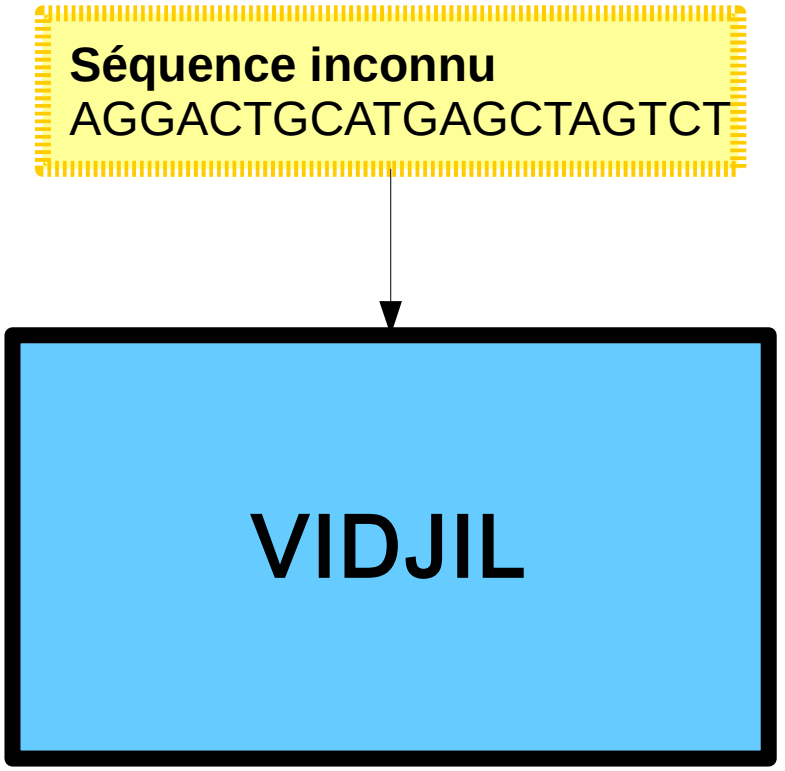


Rôle du logiciel

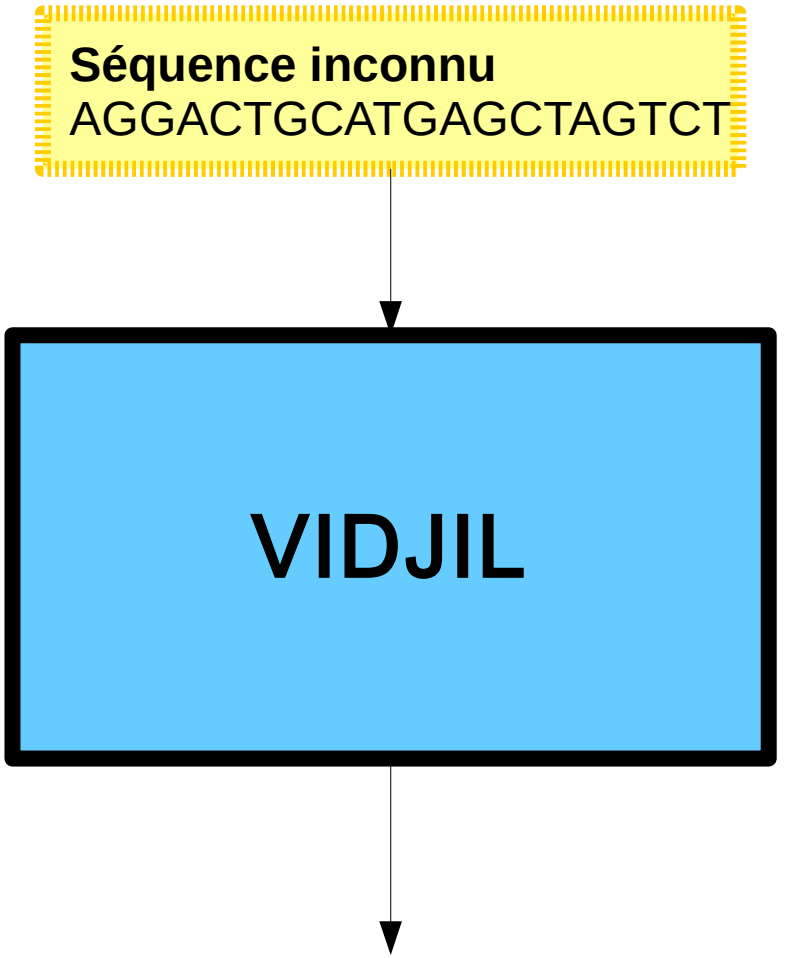
Séquence inconnu
AGGACTGCATGAGCTAGTCT



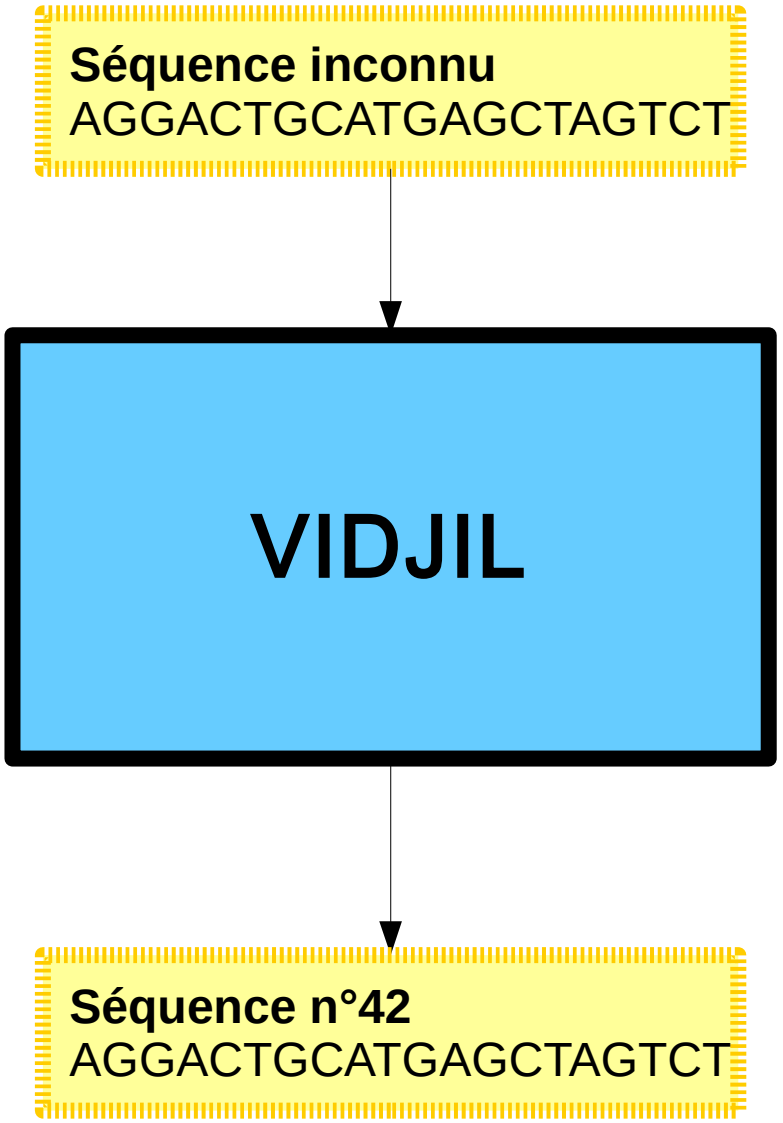
Rôle du logiciel



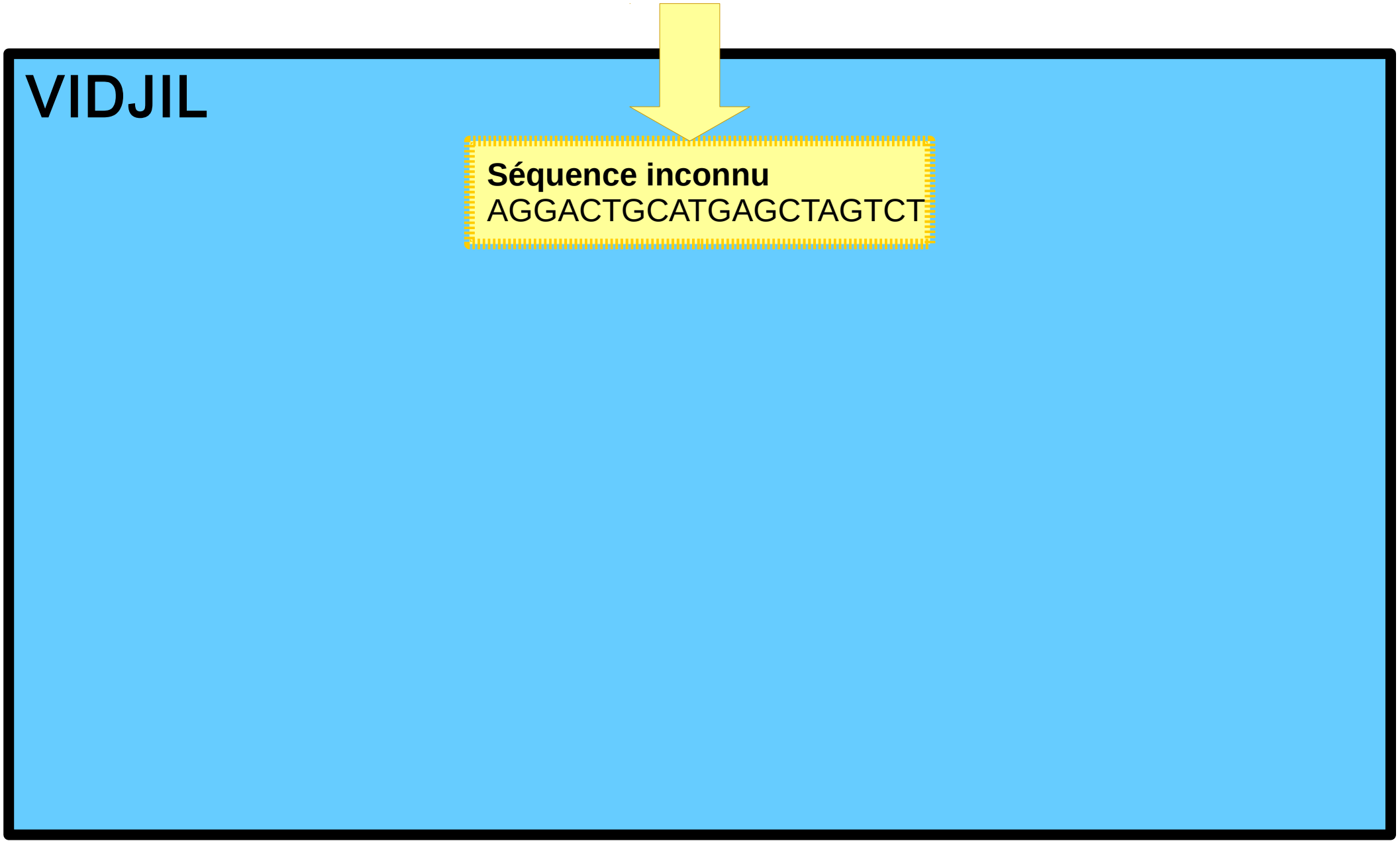
Rôle du logiciel



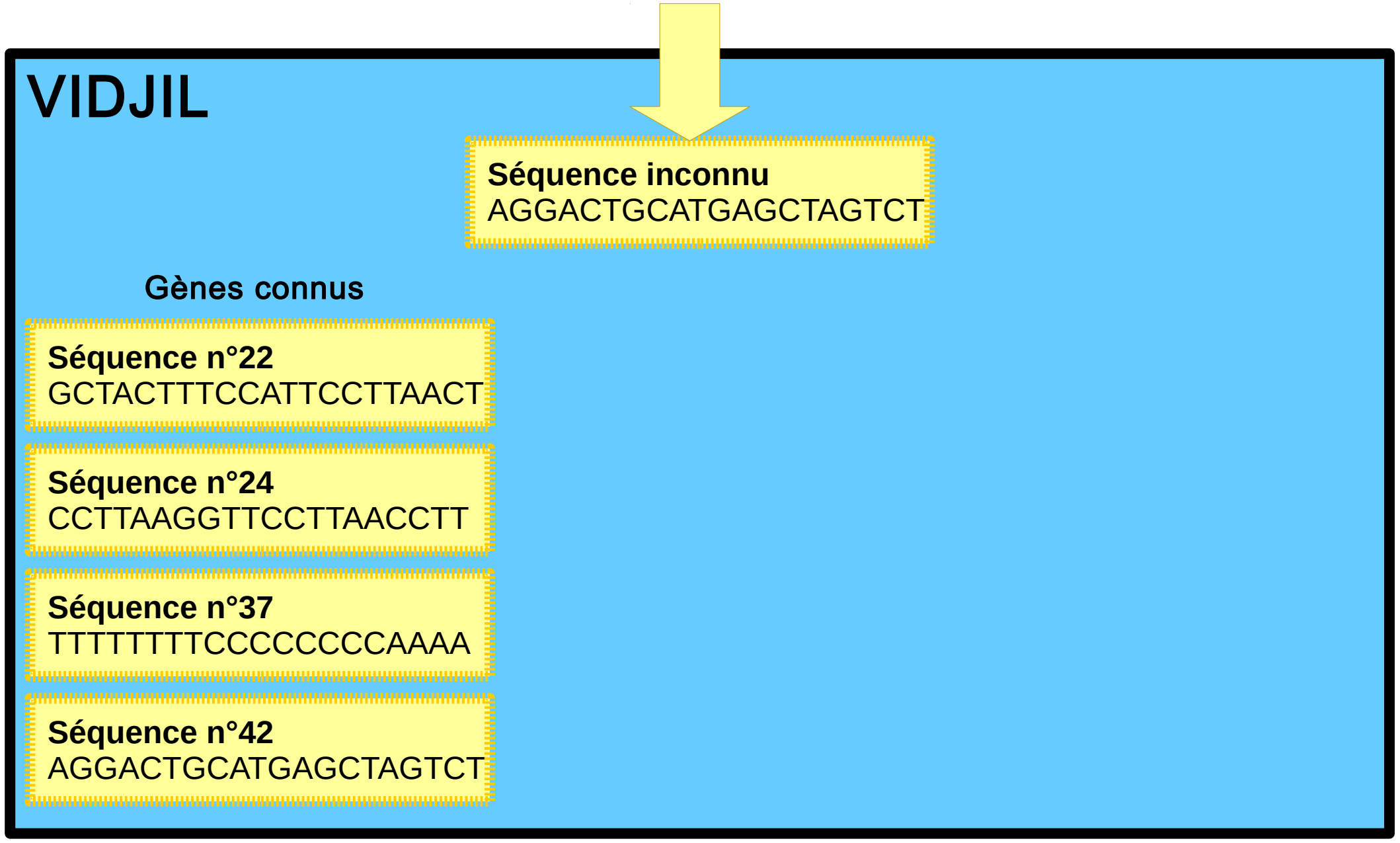
Rôle du logiciel



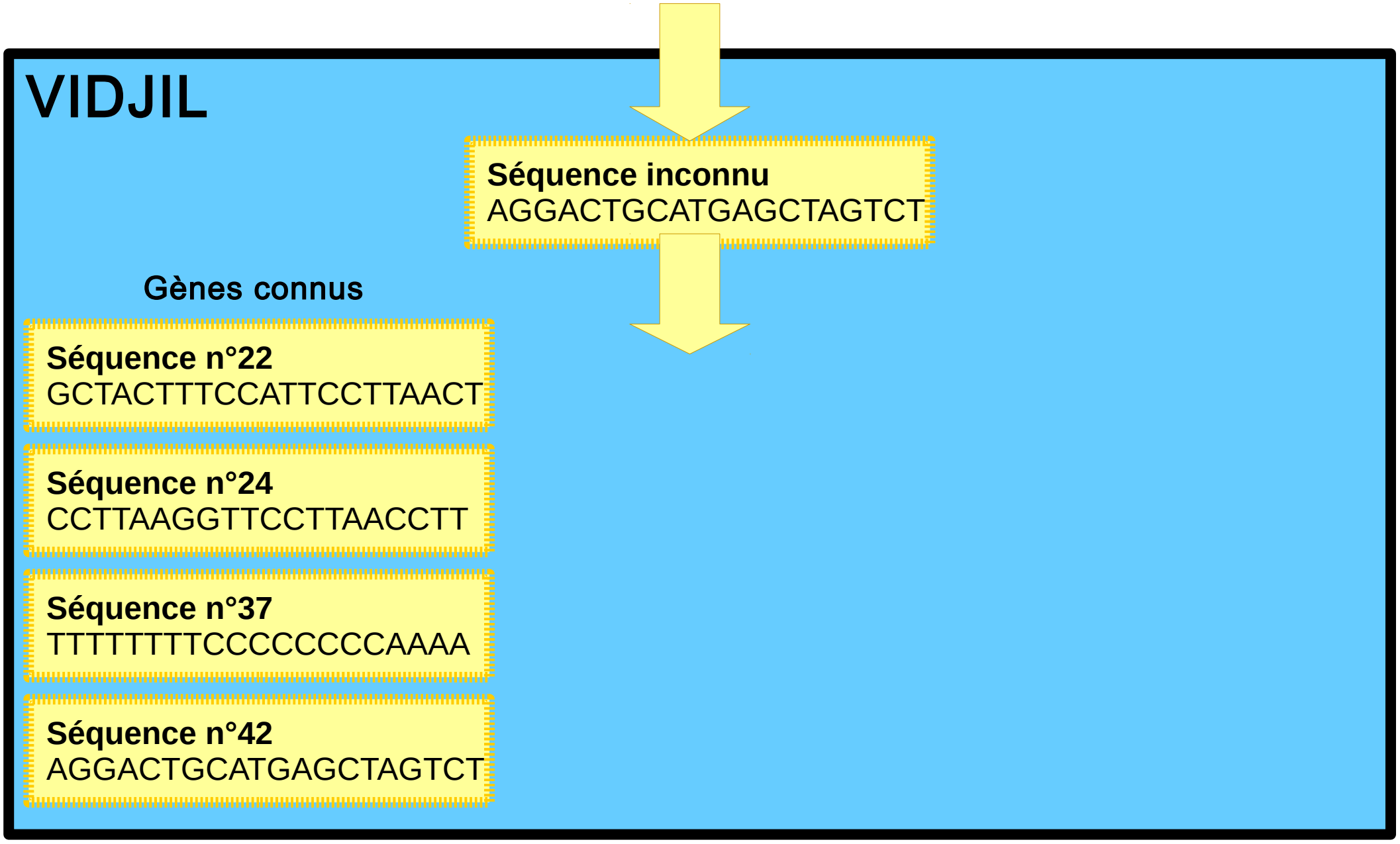
Problématique



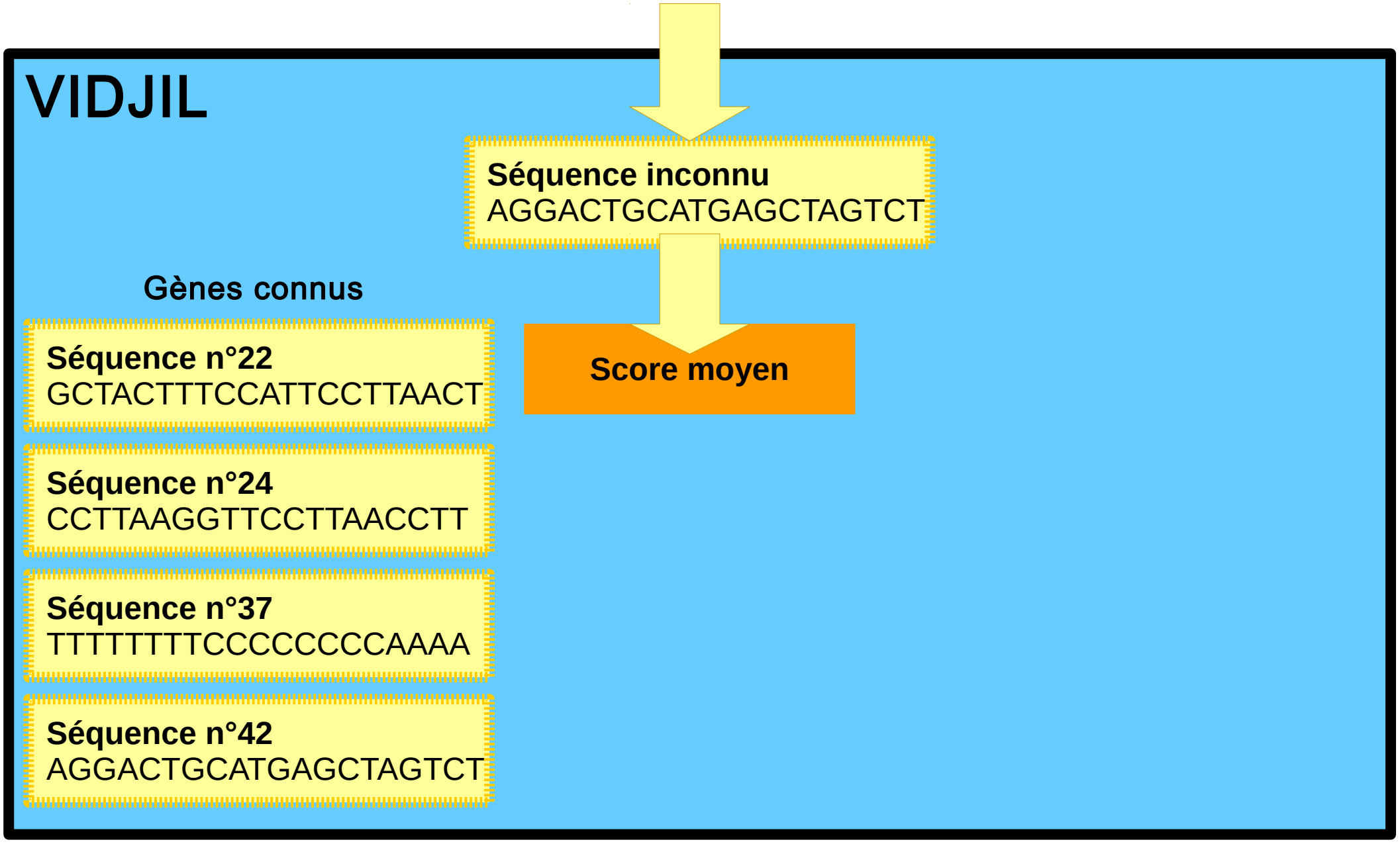
Problématique



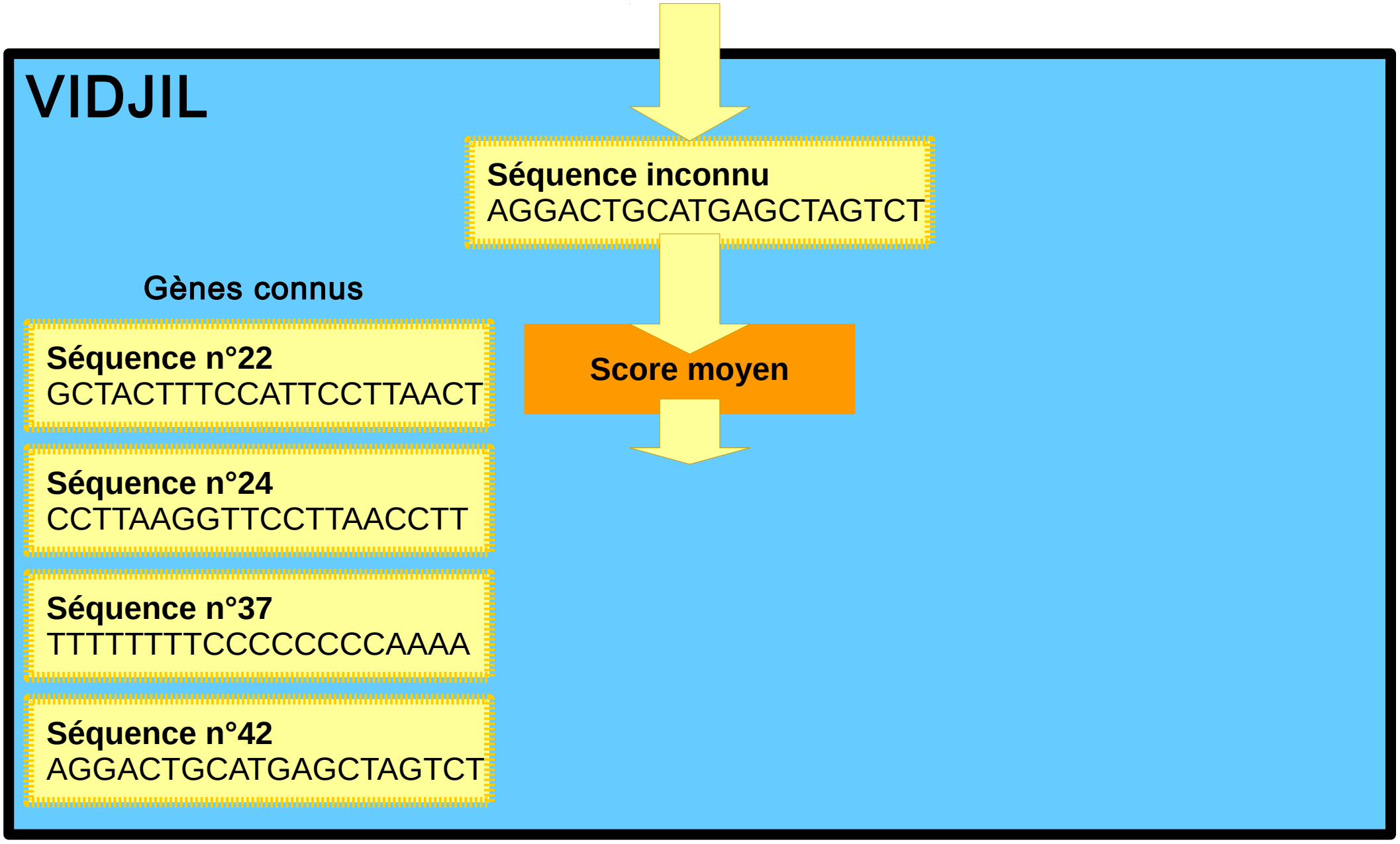
Problématique



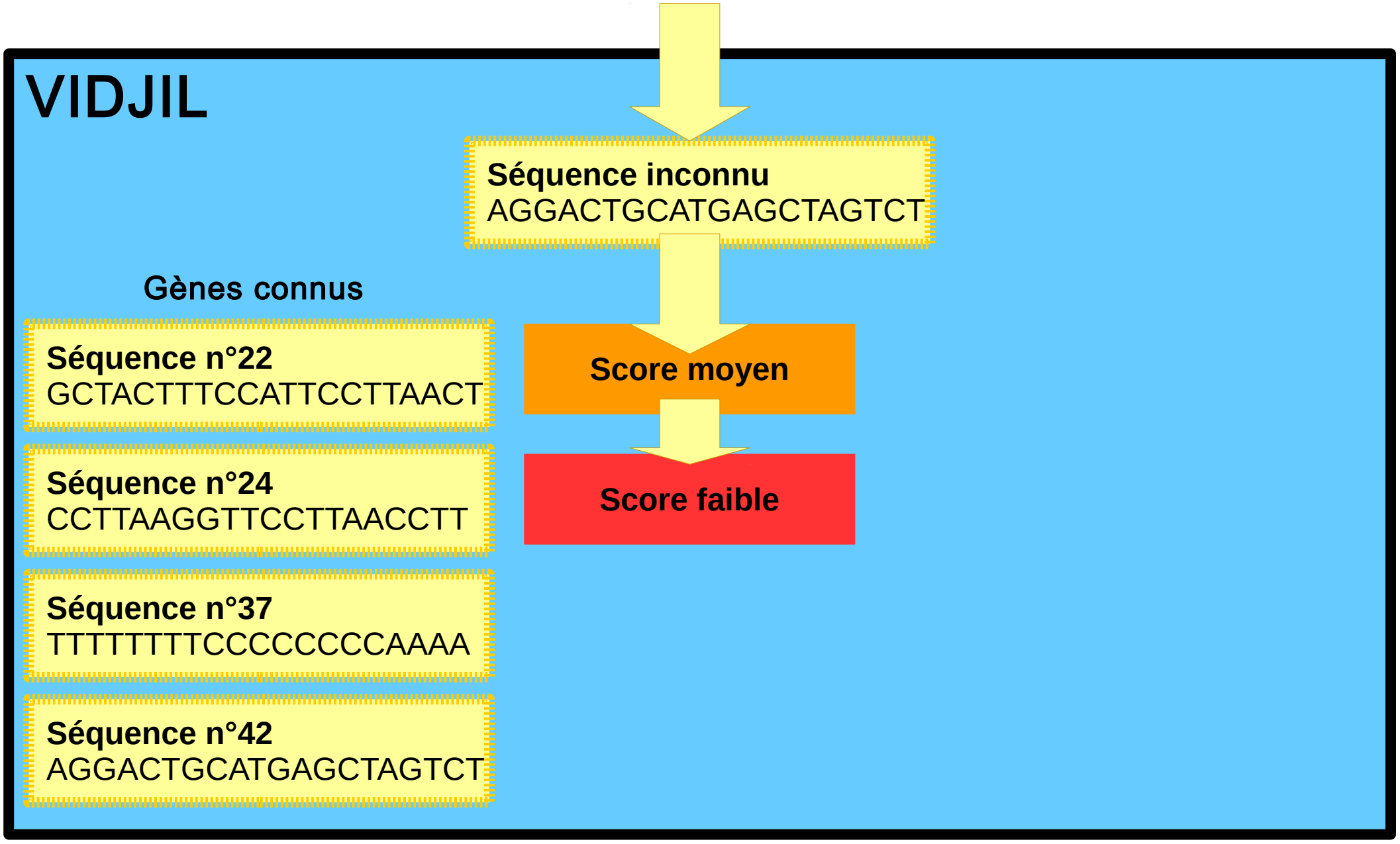
Problématique



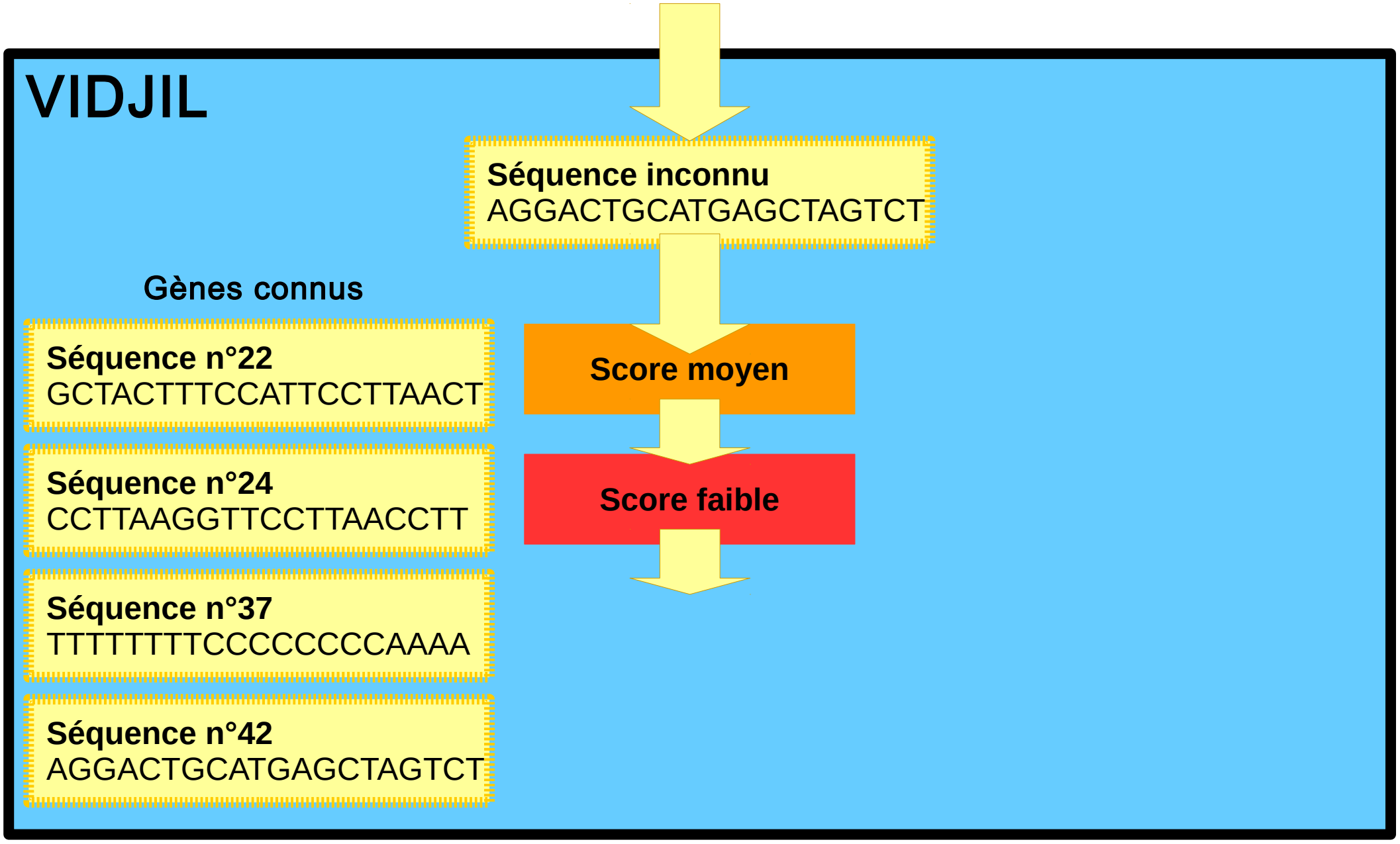
Problématique



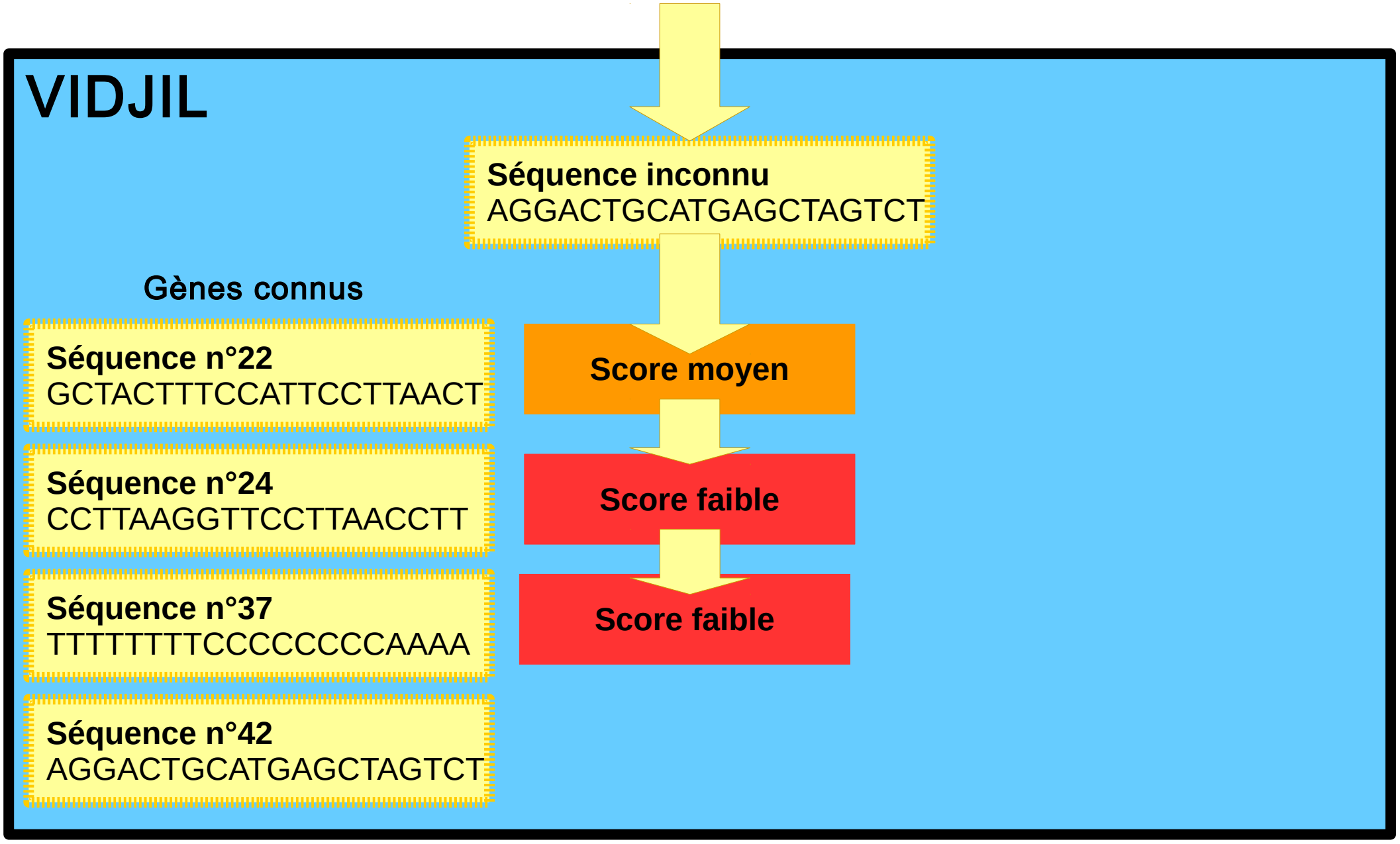
Problématique



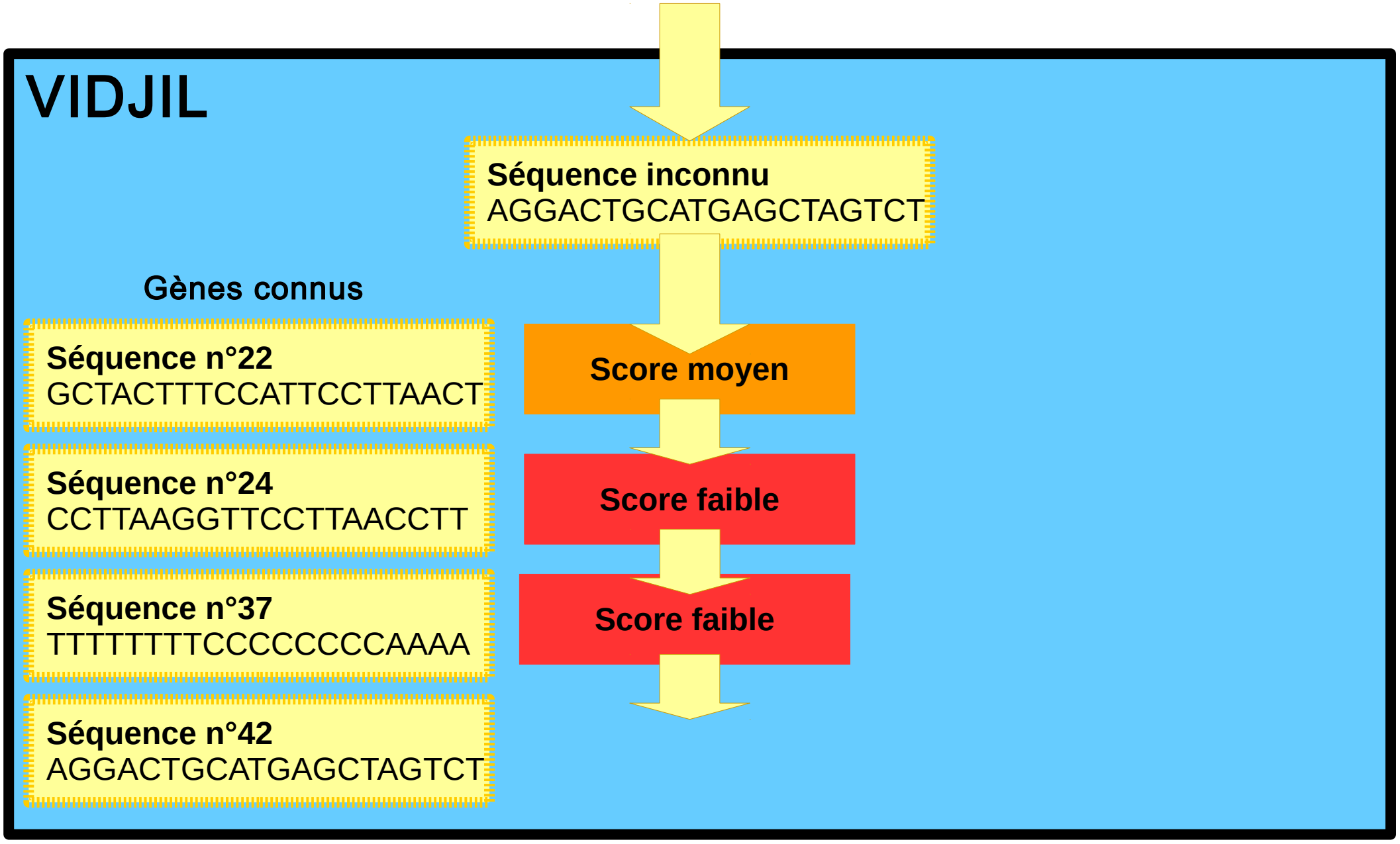
Problématique



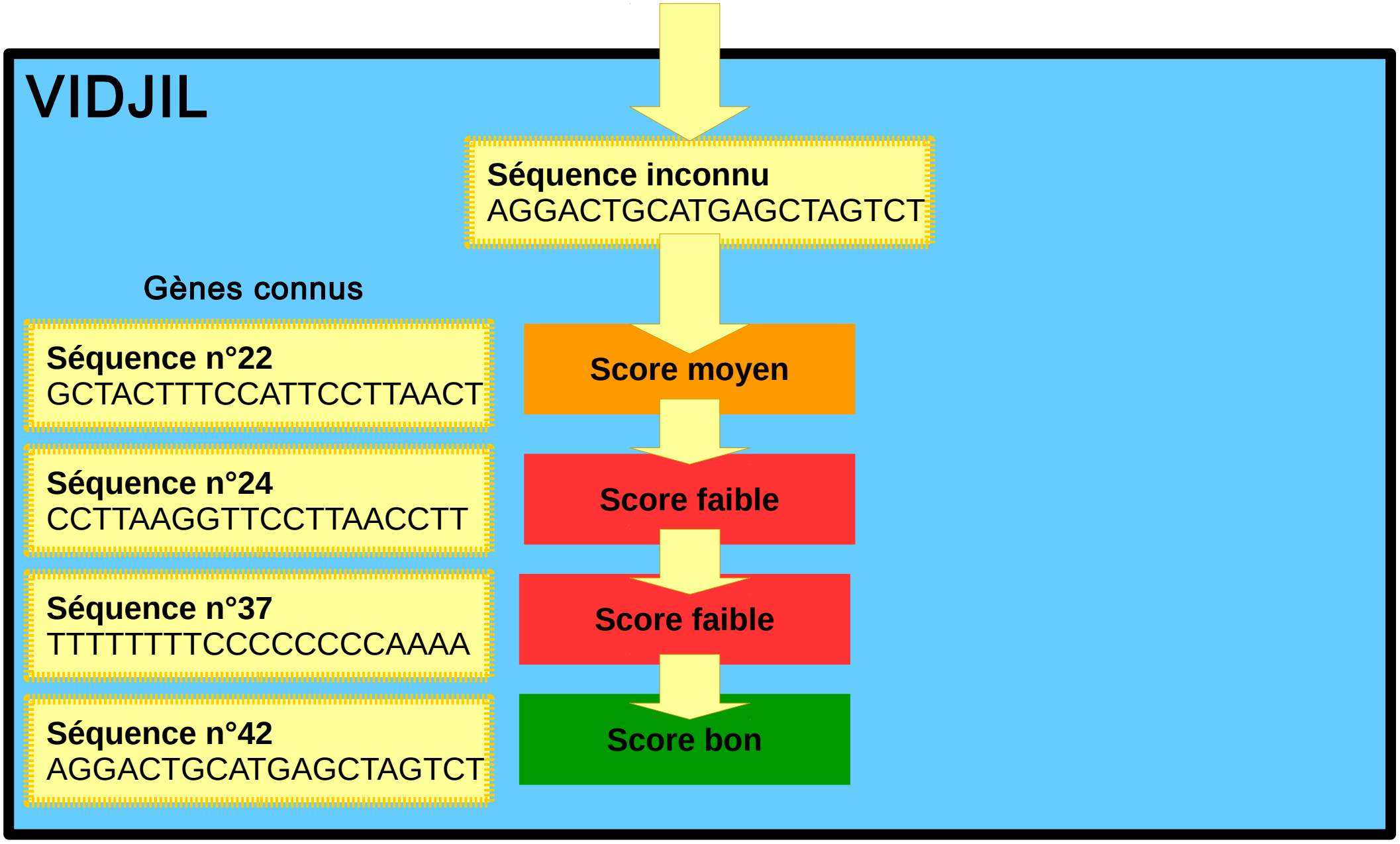
Problématique



Problématique

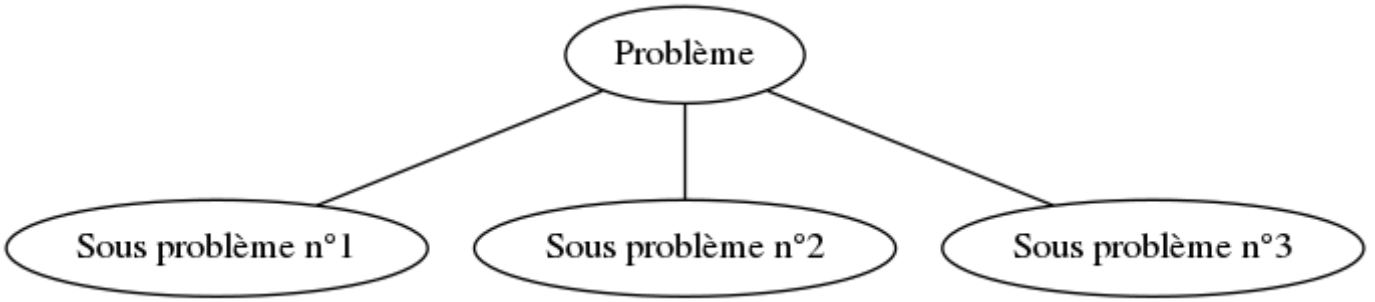


Problématique

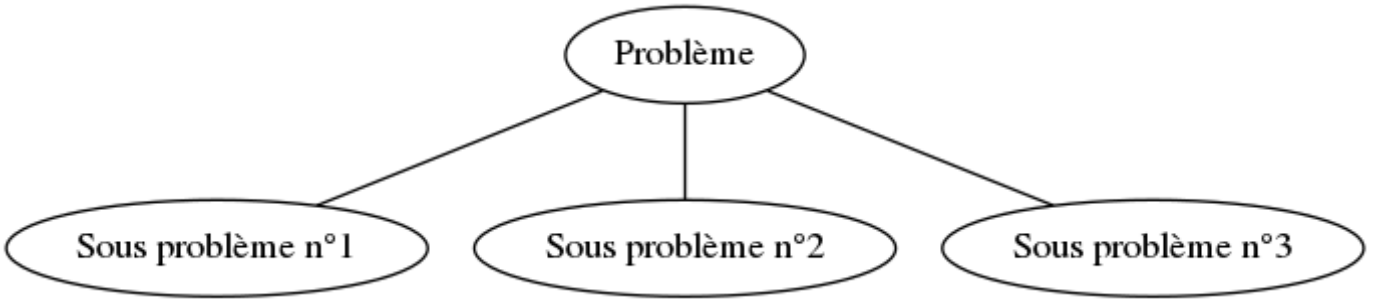


Programmation dynamique

Programmation dynamique

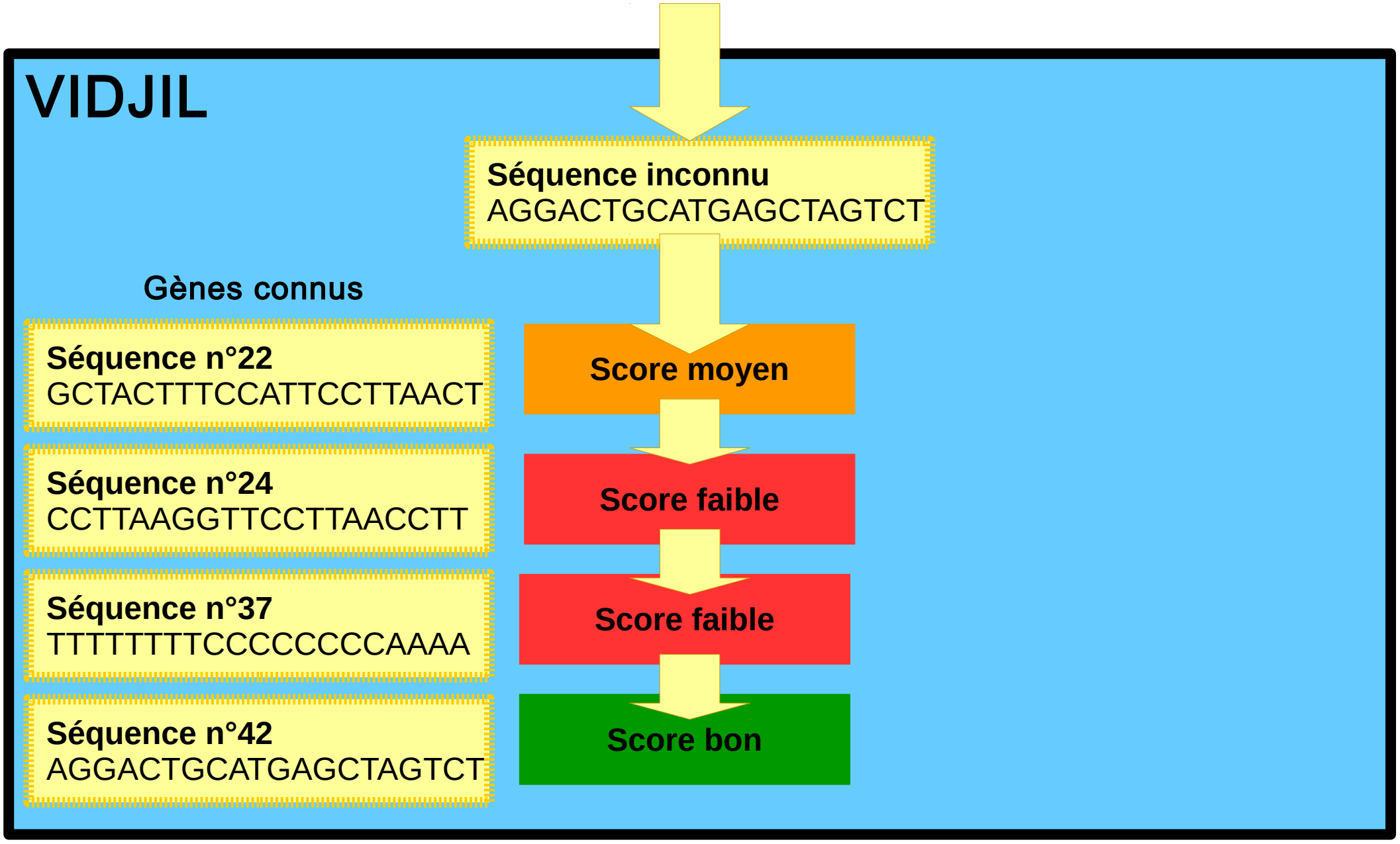


Programmation dynamique

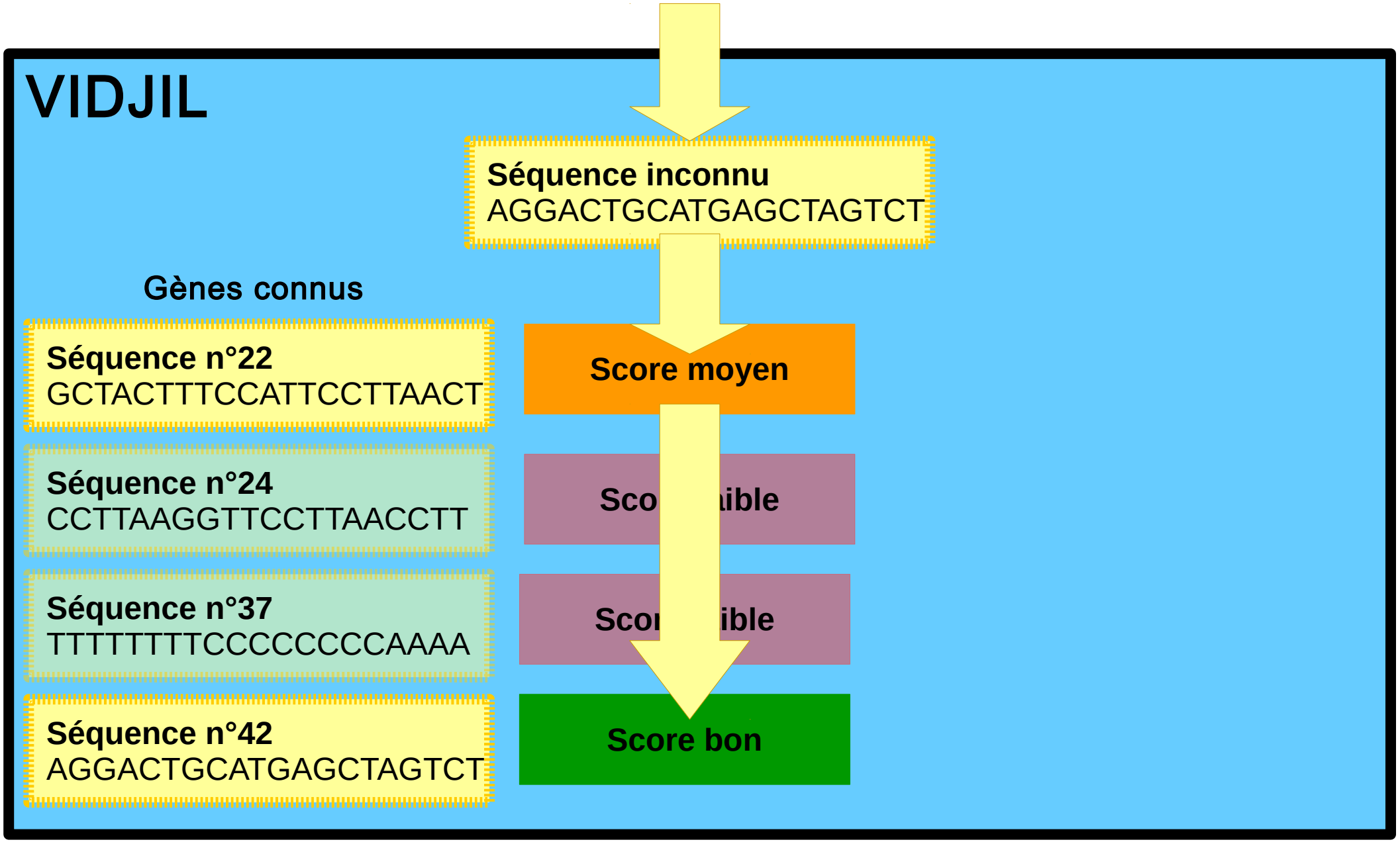


- Problème du plus court chemin (algorithme Bellman-Ford et algorithme de Floyd-Warshall)
- Problème du sac à dos (problème NP difficile, résolu polynomialement)
- Recherche de la plus longue sous-suite strictement croissant dans une série de nombres
- Alignement de séquences

Problématique



Problématique



Pourquoi *Aho-Corasick* ?

Pourquoi *Aho-Corasick* ?

- Algorithme de **recherche de chaînes de caractères**.

Pourquoi *Aho-Corasick* ?

- Algorithme de **recherche de chaînes de caractères**.
- Dans l’automate **chaque lettre** n’est lue qu’**une seule fois**.

Pourquoi *Aho-Corasick* ?

- Algorithme de **recherche de chaînes de caractères**.
- Dans l’automate **chaque lettre** n’est lue qu’**une seule fois**.
- L’automate possède une **complexité linéaire**.

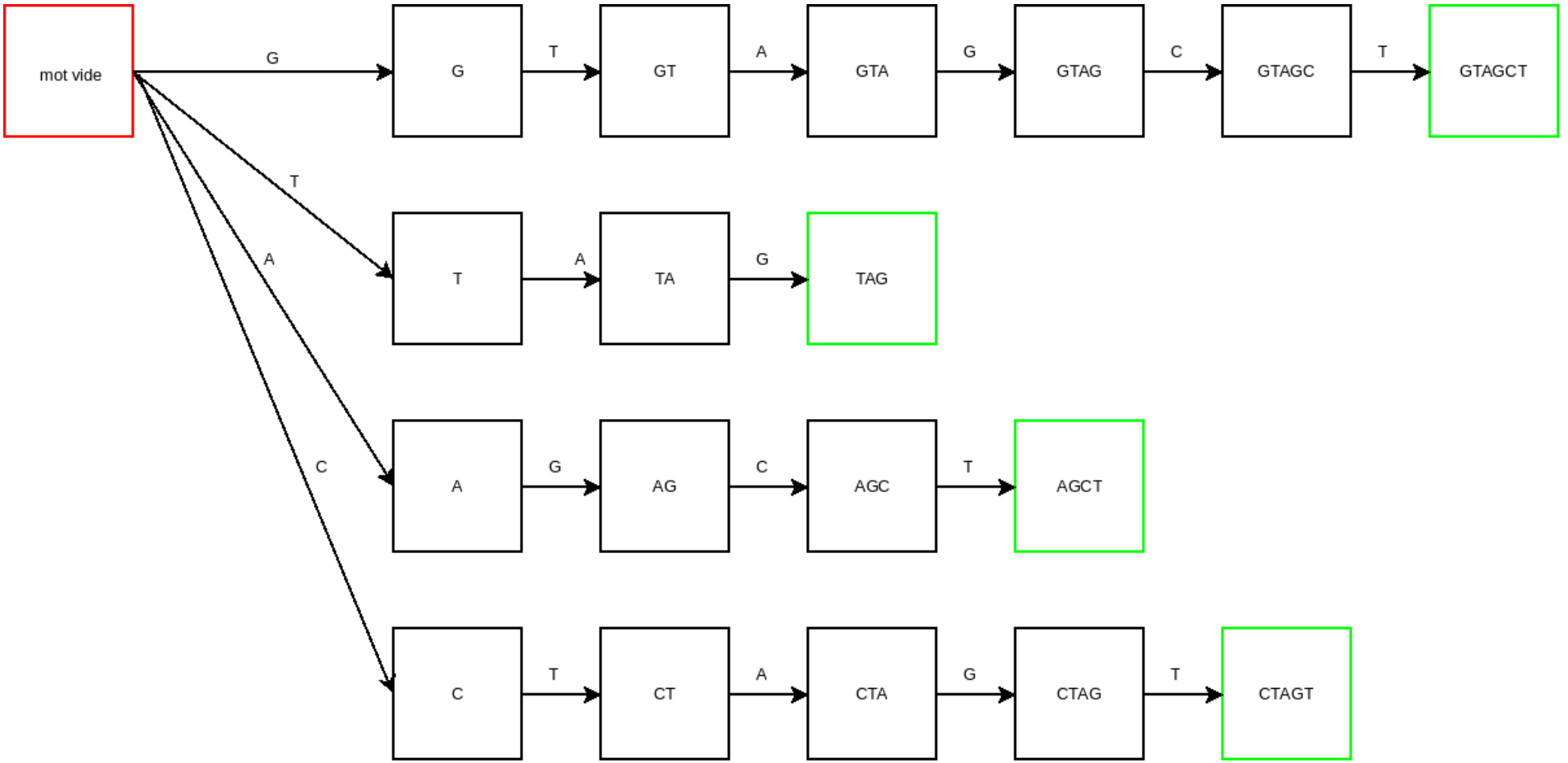
Pourquoi *Aho-Corasick* ?

- Algorithme de **recherche de chaînes de caractères**.
- Dans l'automate **chaque lettre** n'est lue qu'**une seule fois**.
- L'automate possède une **complexité linéaire**.
- La construction de l'automate est **unique**.

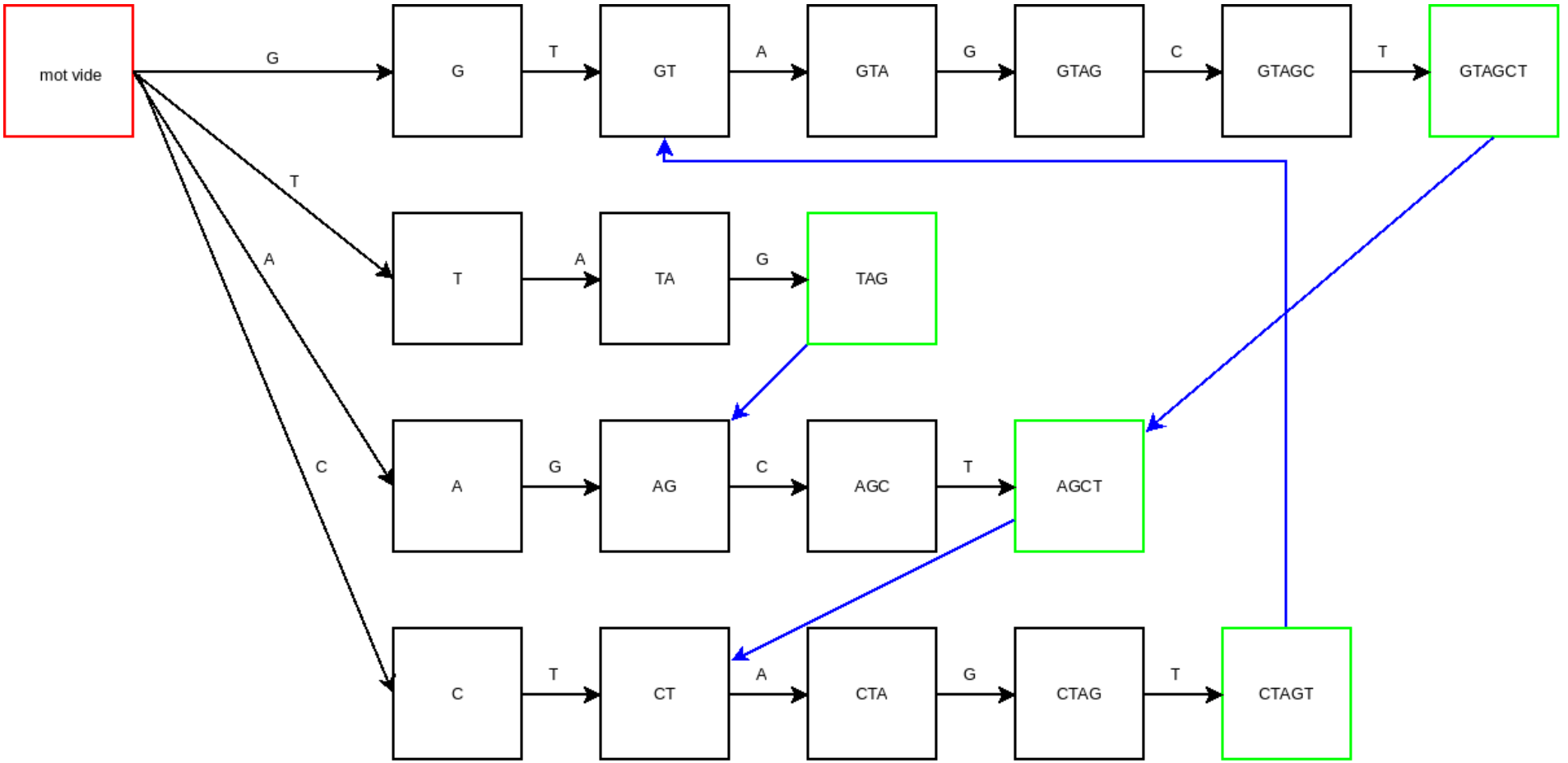
Pourquoi *Aho-Corasick ?*

- Algorithme de **recherche de chaînes de caractères**.
- Dans l'automate **chaque lettre** n'est lue qu'**une seule fois**.
- L'automate possède une **complexité linéaire**.
- La construction de l'automate est **unique**.
- Déjà implémenté dans *VIDJIL*

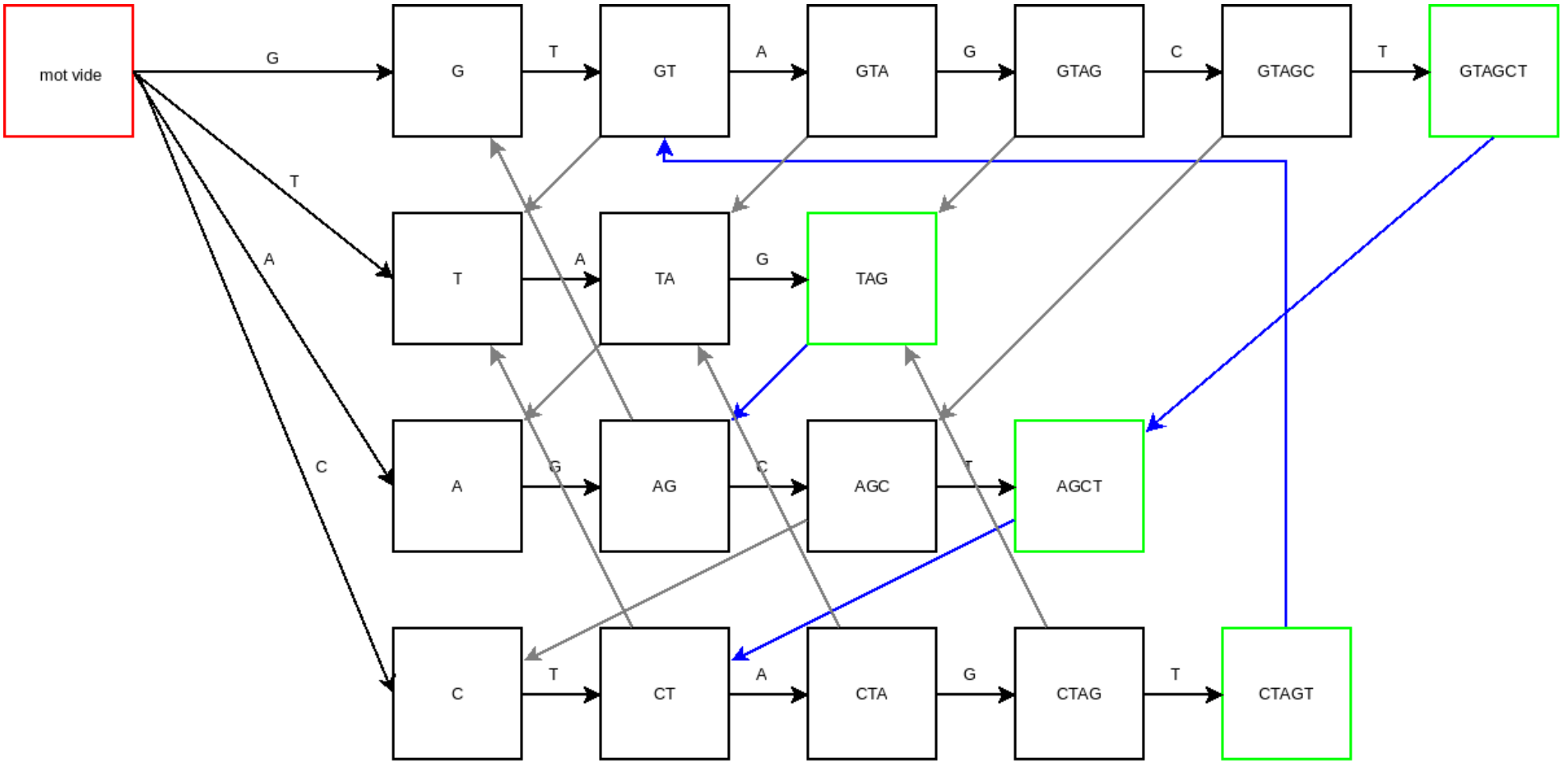
Construction de l’automate



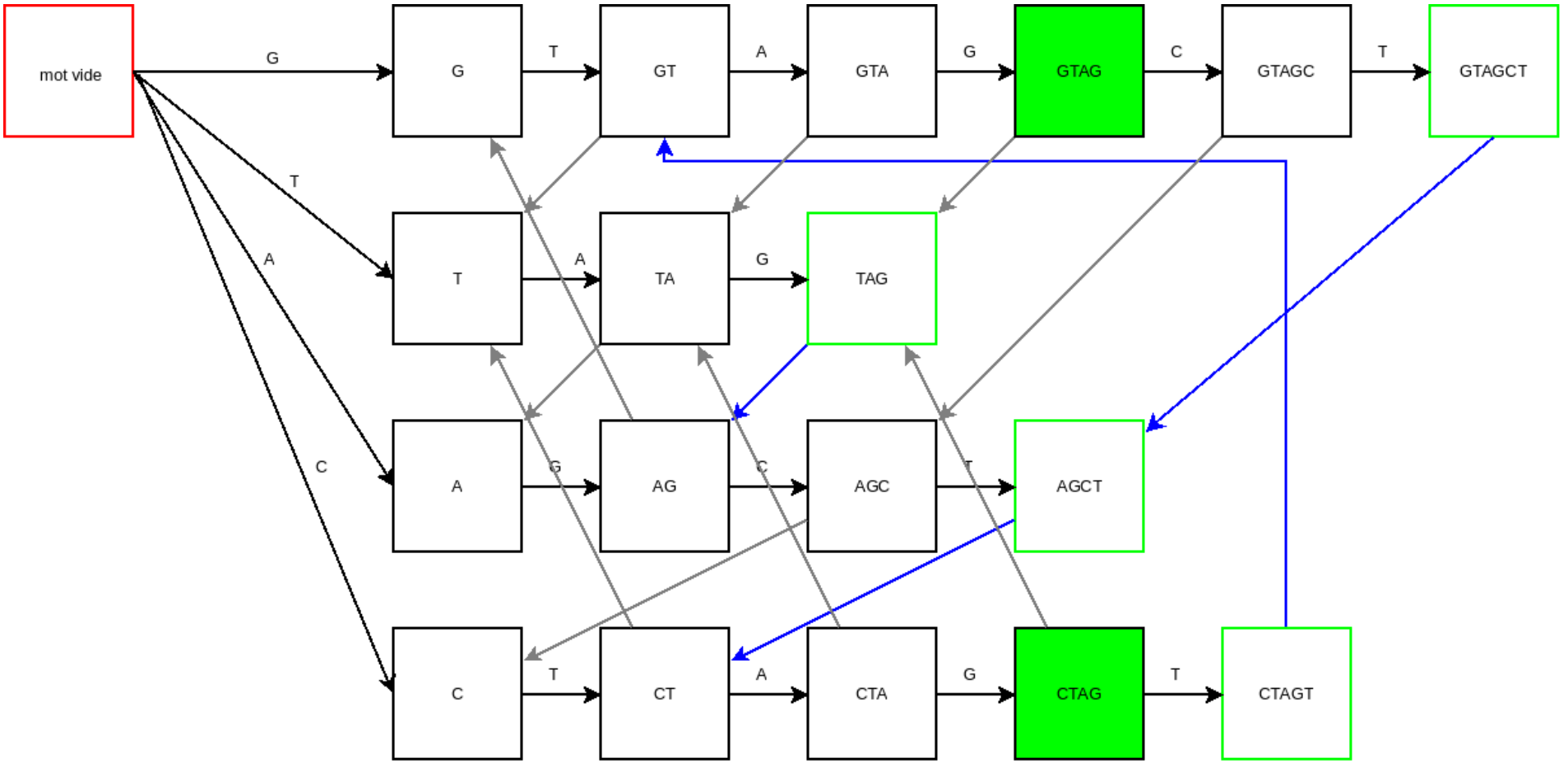
Construction de l’automate



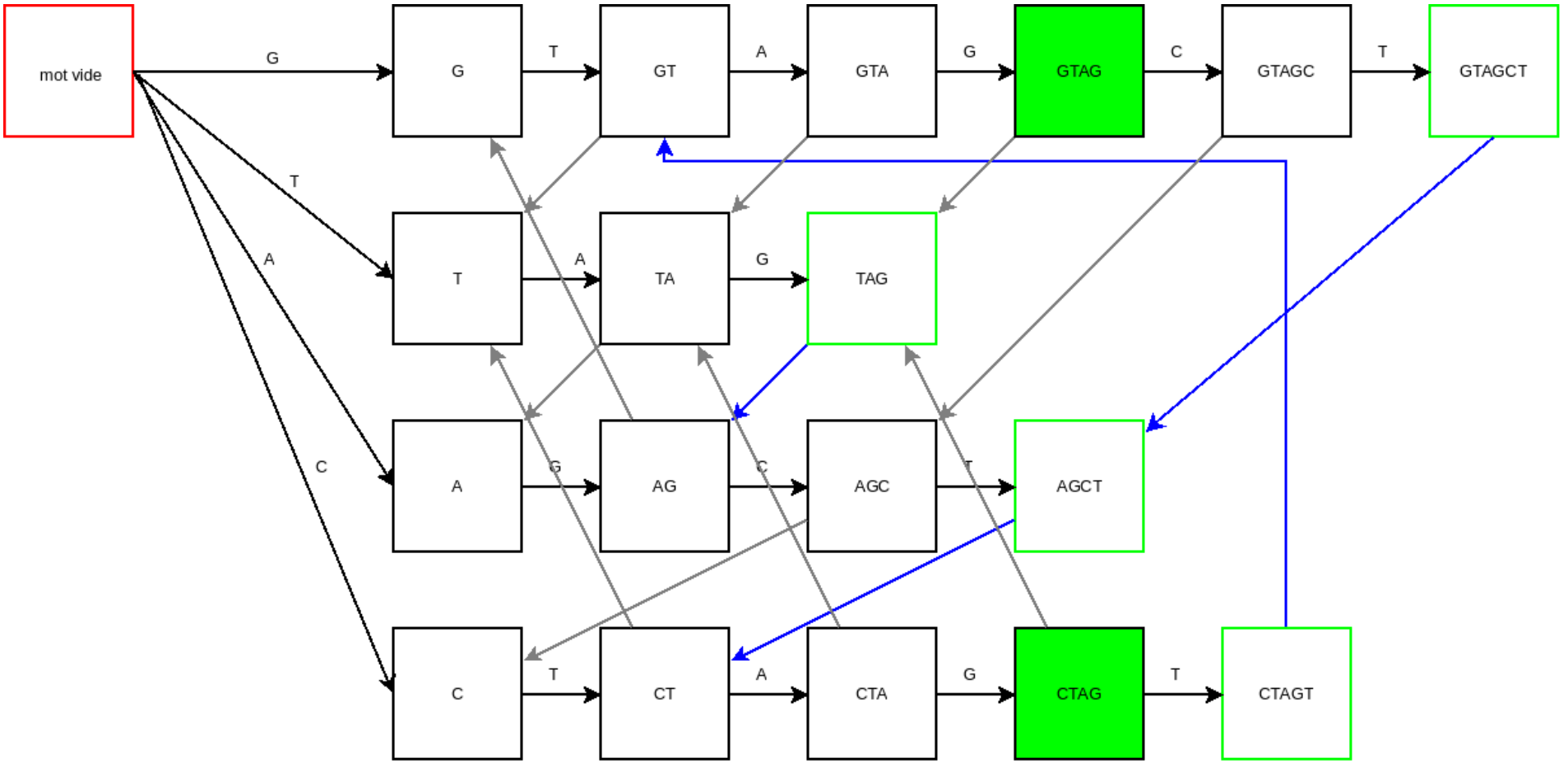
Construction de l’automate



Construction de l’automate



Utilisation de l’automate



Séquence : **GTAGCTTAGTAGAGCTAGCTAGCTCTAGTGTCTAGTC**

Résultats de l’automate

Séquences reconnues	Nombre d’occurrences
GTAGCT	1
TAG	7
AGCT	4
CTAGT	1

Résultats de l’automate

Séquences reconnues	Nombre d’occurrences
GTAGCT	1
TAG	7
AGCT	4
CTAGT	1

La table est triée...

Résultats de l’automate

Séquences reconnues	Nombre d’occurrences
GTAGCT	1
TAG	7
AGCT	4
CTAGT	1

La table est triée...

Séquences reconnues	Nombre d’occurrences
TAG	7
AGCT	4
GTAGCT	1
CTAGT	1

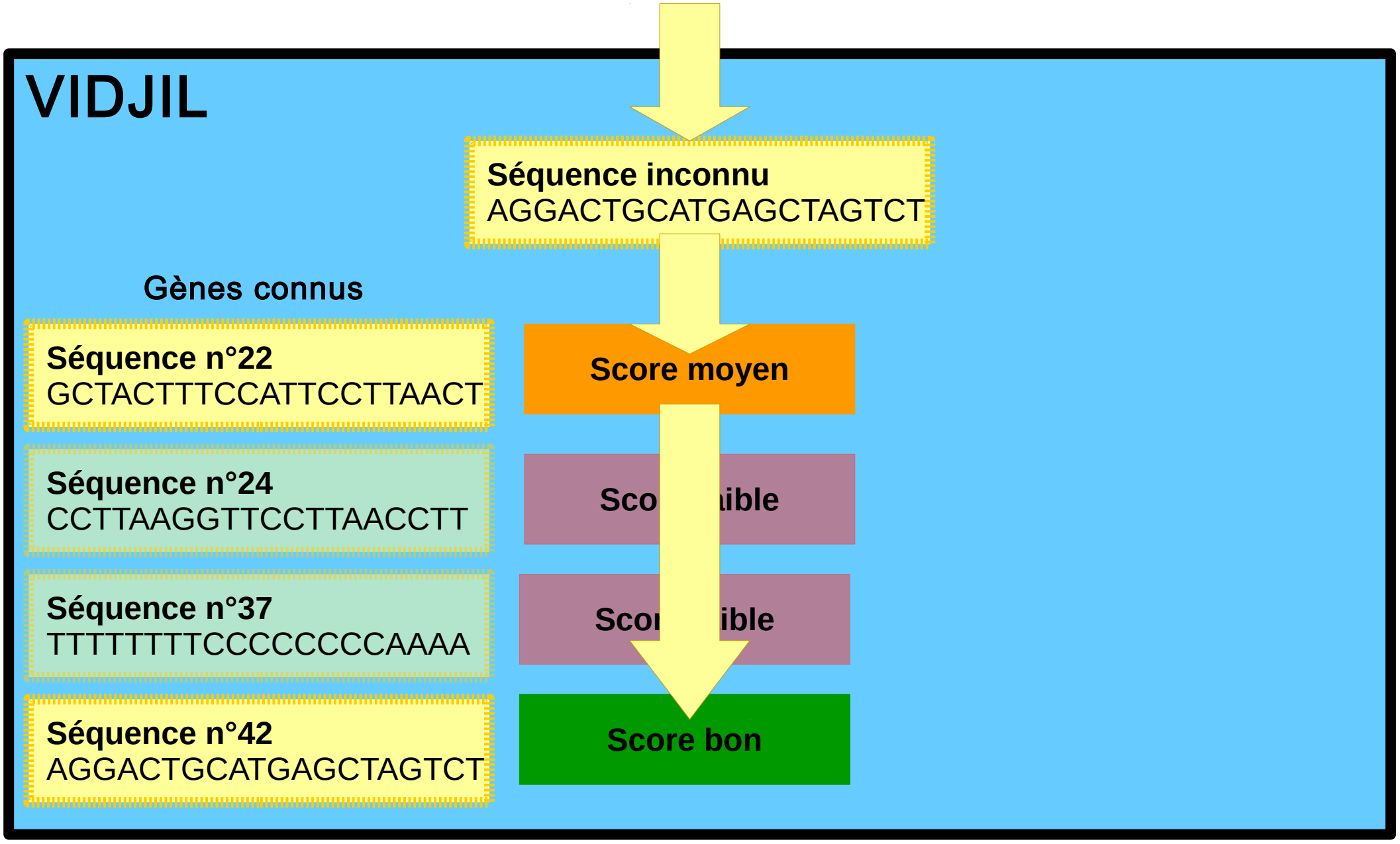
Résultats de l’automate

Séquences reconnues	Nombre d’occurrences
GTAGCT	1
TAG	7
AGCT	4
CTAGT	1

La table est triée...

Séquences reconnues	Nombre d’occurrences
TAG	**7**
AGCT	**4**
GTAGCT	1
CTAGT	1

Problématique



Résultats

Résultats

- Temps de comparaisons réduit en moyenne de 30 %

Résultats

- Temps de comparaisons réduit en moyenne de 30 %

- Meilleur réduction obtenue : 96 %

Conclusion

Conclusion

- La mission de stage est complétée.

Conclusion

- La mission de stage est complétée.

- Les résultats sont positifs.

Conclusion

- La mission de stage est complétée.
- Les résultats sont positifs.
- Le stage m’aura apporté rigueur et autonomie.