Outils informatiques de biologie moléculaire appliquée

A. LA BIOINFORMATIQUE

- Ensemble de méthodes, de logiciels et d'applications en ligne qui permettent de gérer, manipuler, et analyser des données biologiques.

- La bioinformatique met en jeu plusieurs champs disciplinaires :

Informatique **Mathématiques formelles** **Statistiques**

Biologie

- Paradoxe :

  - La **biologie** porte une part de variabilité. Elle peut ne pas être totalement prévisible et totalement reproductible et est souvent dynamique

  - Les Mathématiques et l'Informatique qui sont des **sciences exactes** comportent des concepts et des théories précises

→ La bioinformatique nécessite souvent de décomplexifier des problèmes biologiques (modèles)

- Apport de l'informatique

Stockage et organisation des données

Permet de stocker par exemple les séquences des protéines et d'y associer différentes annotations : positions des domaines, des sites actifs, d'un pro-peptide, spécificité d'expression, rôle fonctionnel, associations à des pathologies….

Automatisation de tâches manuelles

Certaines tâches simples ne peuvent pas être réalisées à la main pour de nombreuses séquences (manque de temps, d'intérêt et risque d'erreurs) et sont donc automatisées (traduction, recherche de sites d'enzymes de restriction…)

Algorithme

Un algorithme est une suite finie et non-ambiguë d'instructions permettant de donner la réponse à un problème.

- Apport des mathématiques

Statistiques

Permet d'évaluer des résultats entre eux en proposant des calculs de scores et de probabilités (p-value)
=> Aide l'interprétation

Modélisation

Permet de faire des prédictions à partir d'une mise en équation d'un système et des données biologiques

Objectifs de la bioinformatique

La bioinformatique a différents objectifs et différentes applications :

1- Collecter et stocker des informations dans des bases de données, accessibles en ligne.

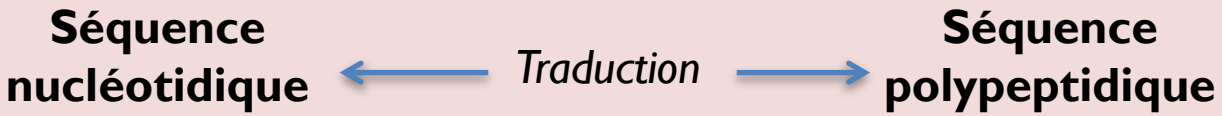
Explosion de la quantité de données biologiques nécessitant des outils de stockage adaptés.

2- Fournir des outils de comparaison de séquences (protéiques ou nucléotidiques).

Objectifs :

- identifier une séquence par rapport à une base de données
- déterminer le degré de similitudes entre deux séquences (intérêt en taxonomie)
- repérer des motifs structuraux :
  - -gènes, promoteurs, etc. pour un nucléotide.
  - -zone de repliement, site actif, etc. pour un polypeptide.

3- Fournir des outils de traduction de séquences.



Objectifs :

- simplifier les taches de traduction
- proposer plusieurs possibilités de protéines pour une même séquence
- repérer exons / introns

4- Fournir des outils de prédiction

Prédiction physiologique et fonctionnelle

Objectifs :
- repérer un opéron
- repérer un gène ou une protéine anormale
- prévoir la structure 3D d'une protéine
- repérer des mutations
- prédire une pathologie…

Prédiction expérimentale

Objectifs :
- repérer des sites de restriction
- prévoir la digestion d'un nucléotide
- prévoir / simuler la migration de fragments nucléotidiques ou protéiques lors d'une électrophorèse…

- Quelques théories et concepts en Biologie :

  - La théorie de l'évolution énoncée par Darwin (1859), complétée par Kimura avec la théorie neutraliste de l'évolution (1983).

  - Les lois de Mendel (en 1866).
    => Première théorie biologique à partir d'une analyse statistique.

  - La mise en évidence des chromosomes comme support cellulaire de l'hérédité et de l'information génétique (Morgan, 1913).

  - La découverte de la structure en double hélice de l'ADN (Watson et Crick, 1953), puis du mécanisme de la régulation génétique impliqué dans le dogme central de la biologie moléculaire (1965). Des dérogations au dogme ont finalement été trouvées notamment par Temin et Baltimore (1970)

B. La séquence biologique pour les informaticiens

- Enoncer que l'information génétique de tout organisme vivant est contenue dans une séquence fut un concept révolutionnaire. La séquence devient un élément essentiel en biologie grâce à la biologie moléculaire (enzyme de restriction, PCR, vecteur de clonage, évolution des techniques de séquençage)

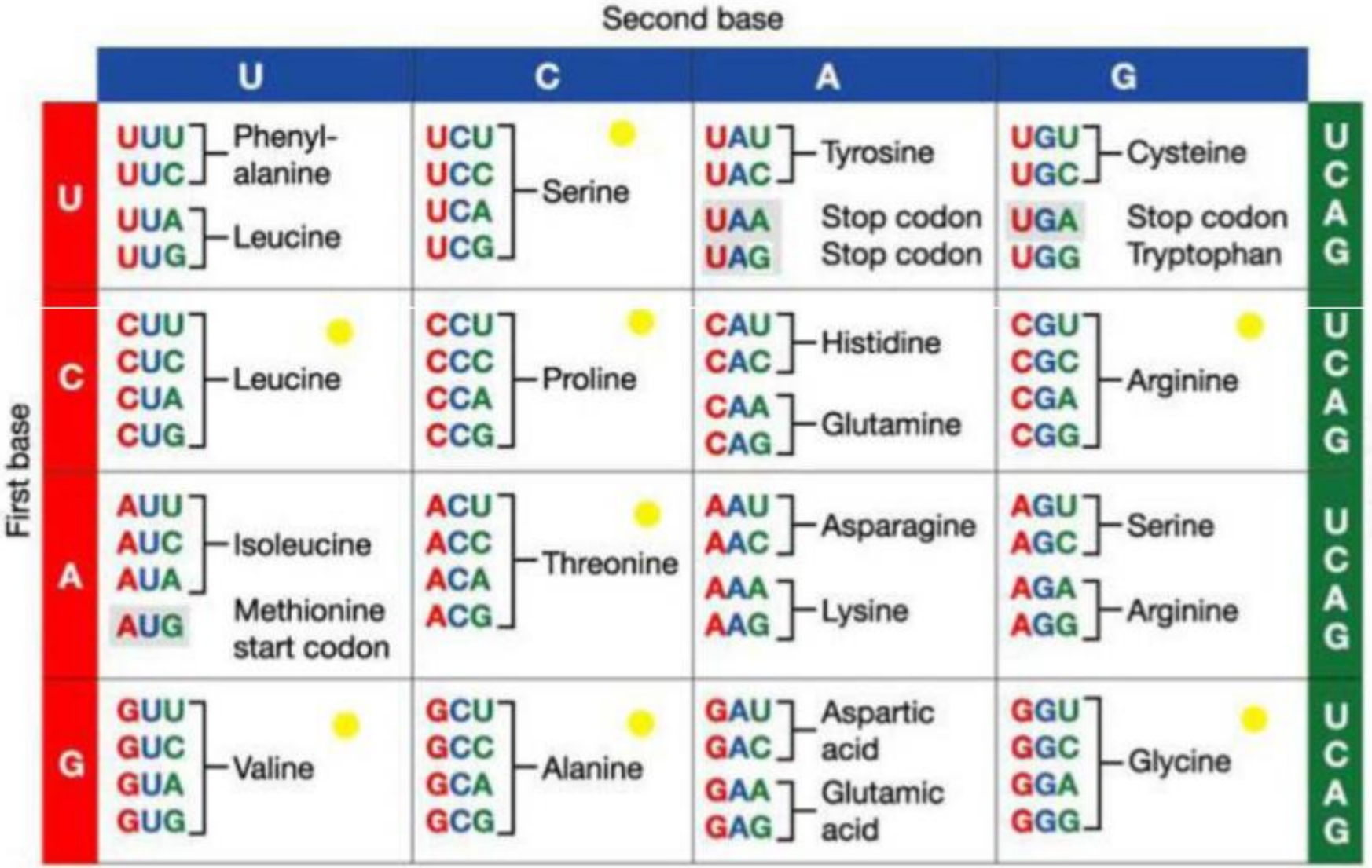
  ⇒ La séquence devient un **objet élémentaire et formel** qui manquait à la biologie pour se constituer une branche théorique

- C'est une **chaîne de caractères** basée sur un alphabet simple et fixe.
  - ADN : 4 nucléotides ATCG
  - ARN : 4 nucléotides AUCG
  - Protéines : 20 acides aminés

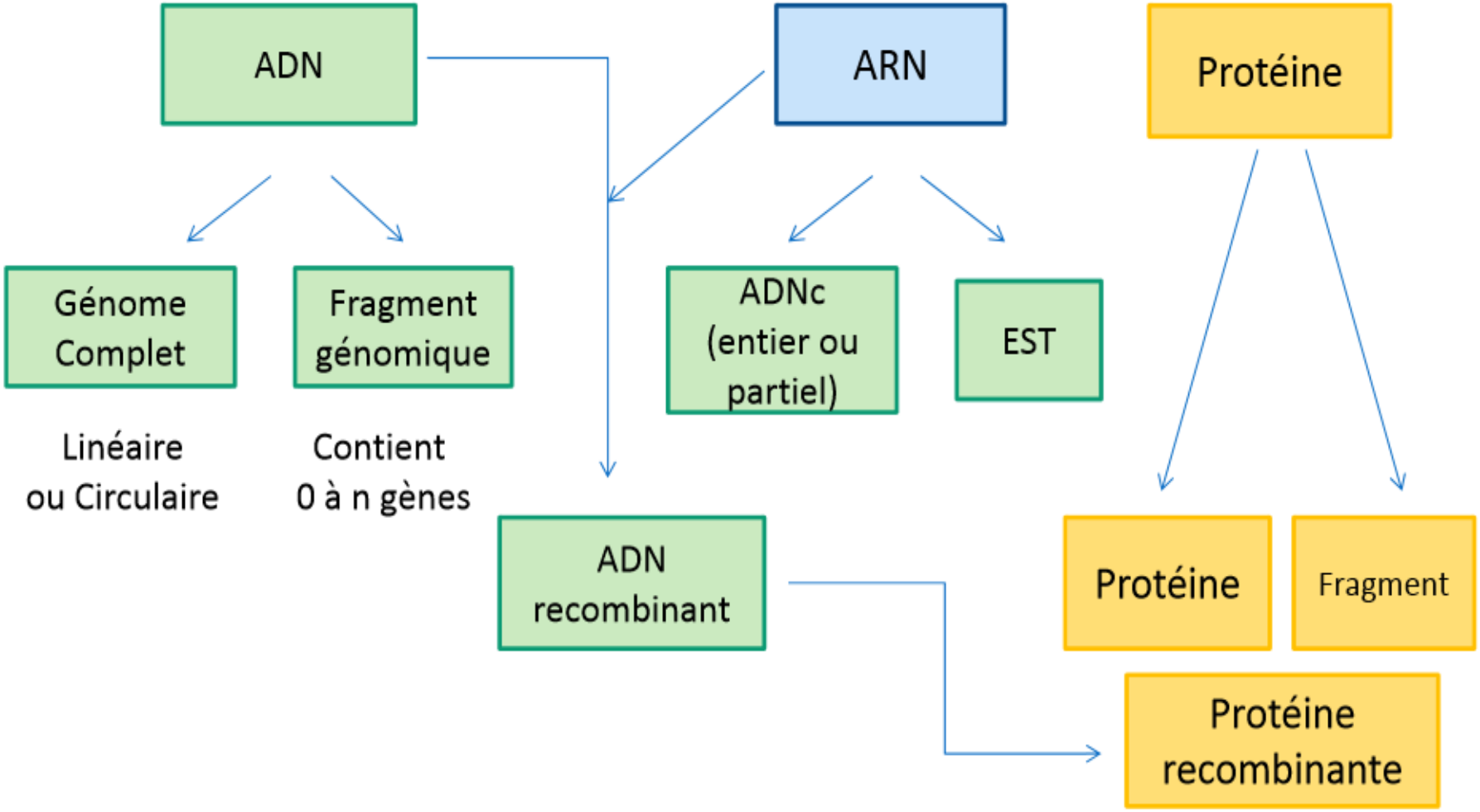
- La séquence est manipulable par des algorithmes !

  ⇒ Récupération et **manipulation** de certains éléments ou groupes d'éléments dans la chaine de caractère

Le code génétique : une règle de traduction !

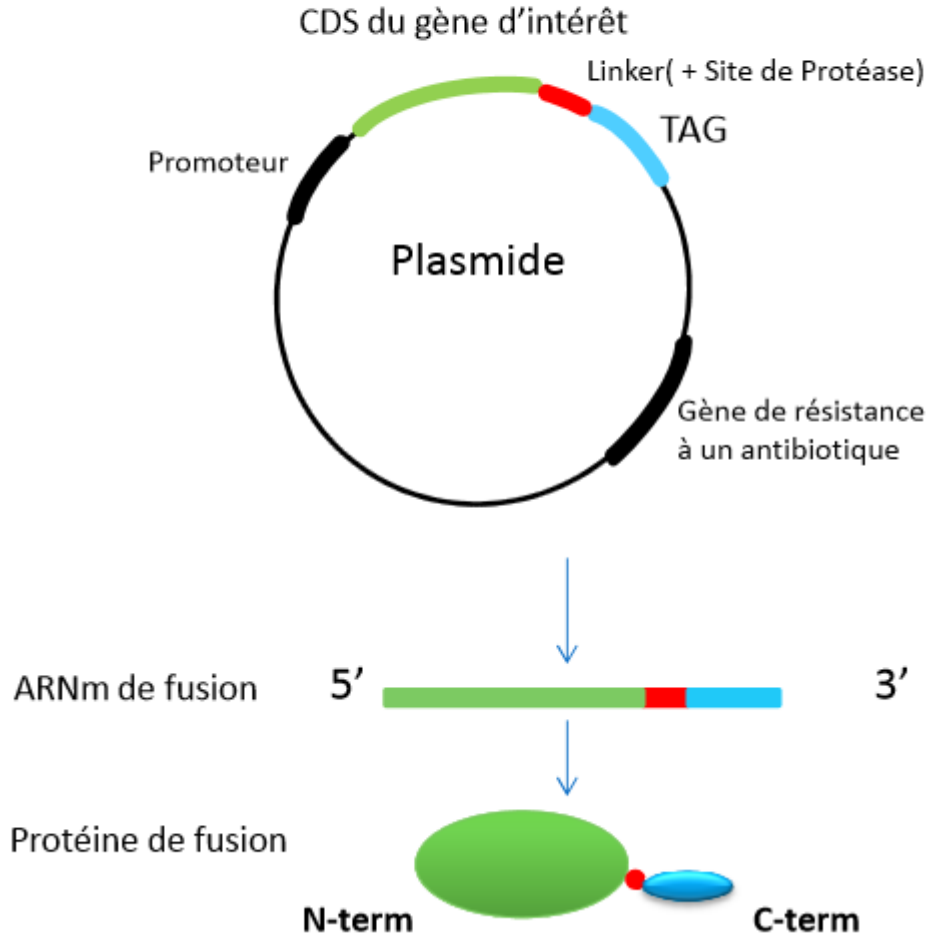


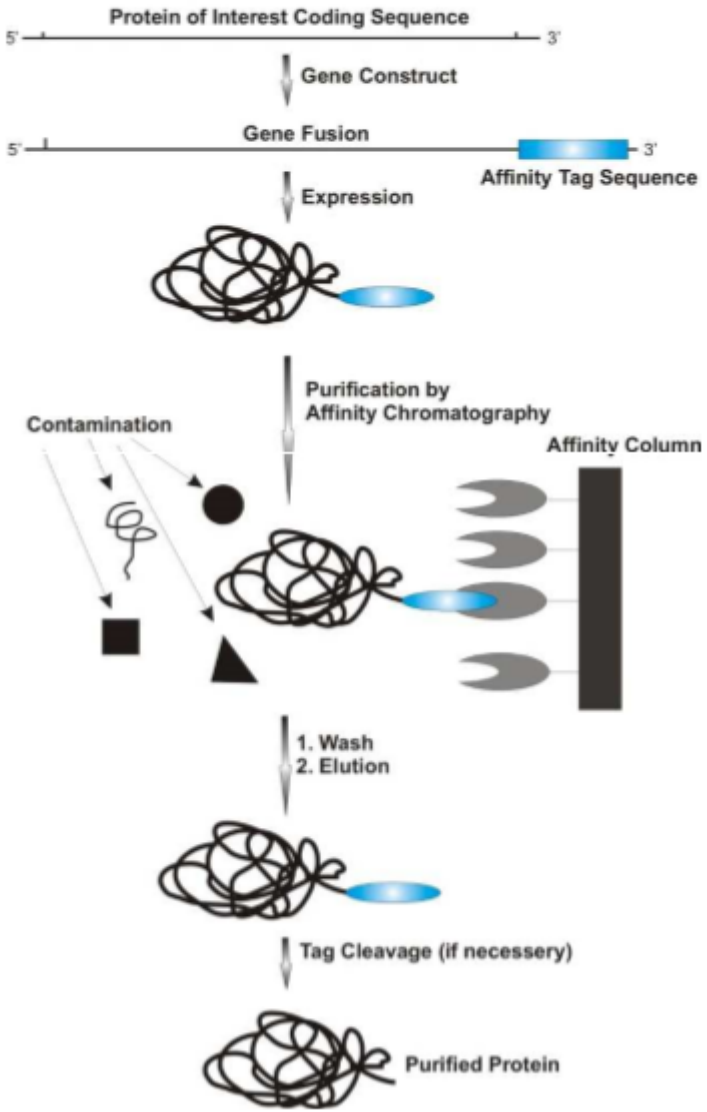
- Les séquences les plus fréquentes :



- Les séquences issues de l'ADN recombinant:

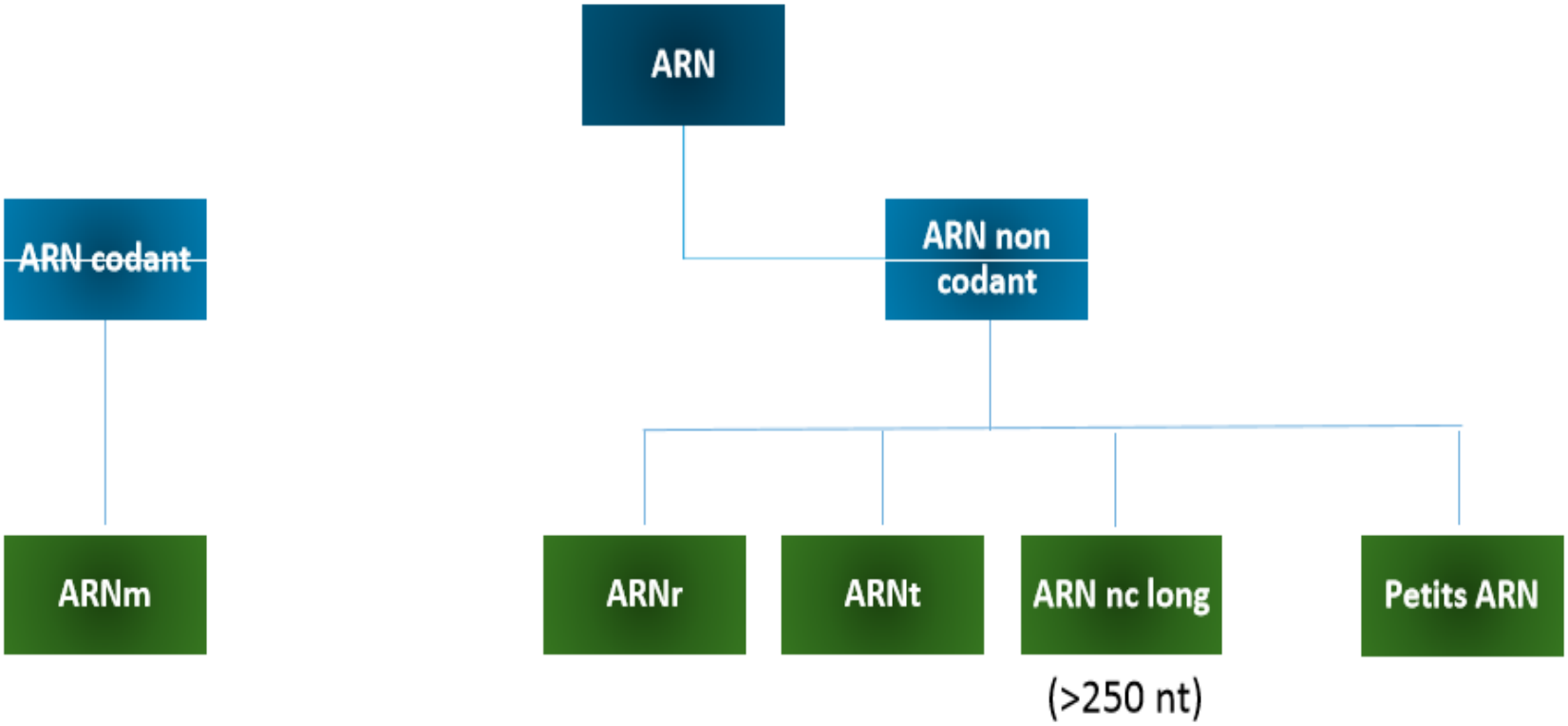
Exemple de construction pour un gène de fusion avec un tag en C-terminal



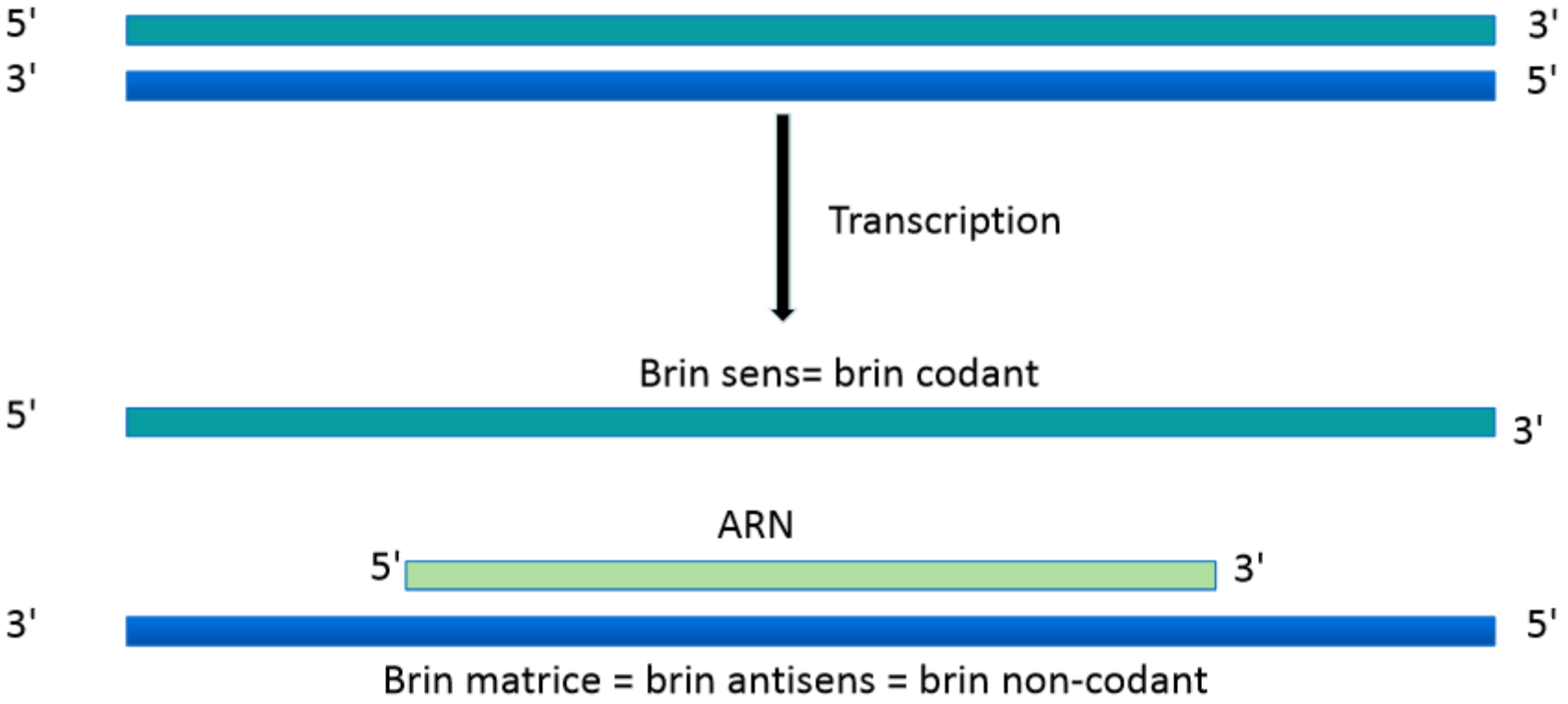


Rappel de biologie : tous les ARN ne codent pas pour des protéines !!

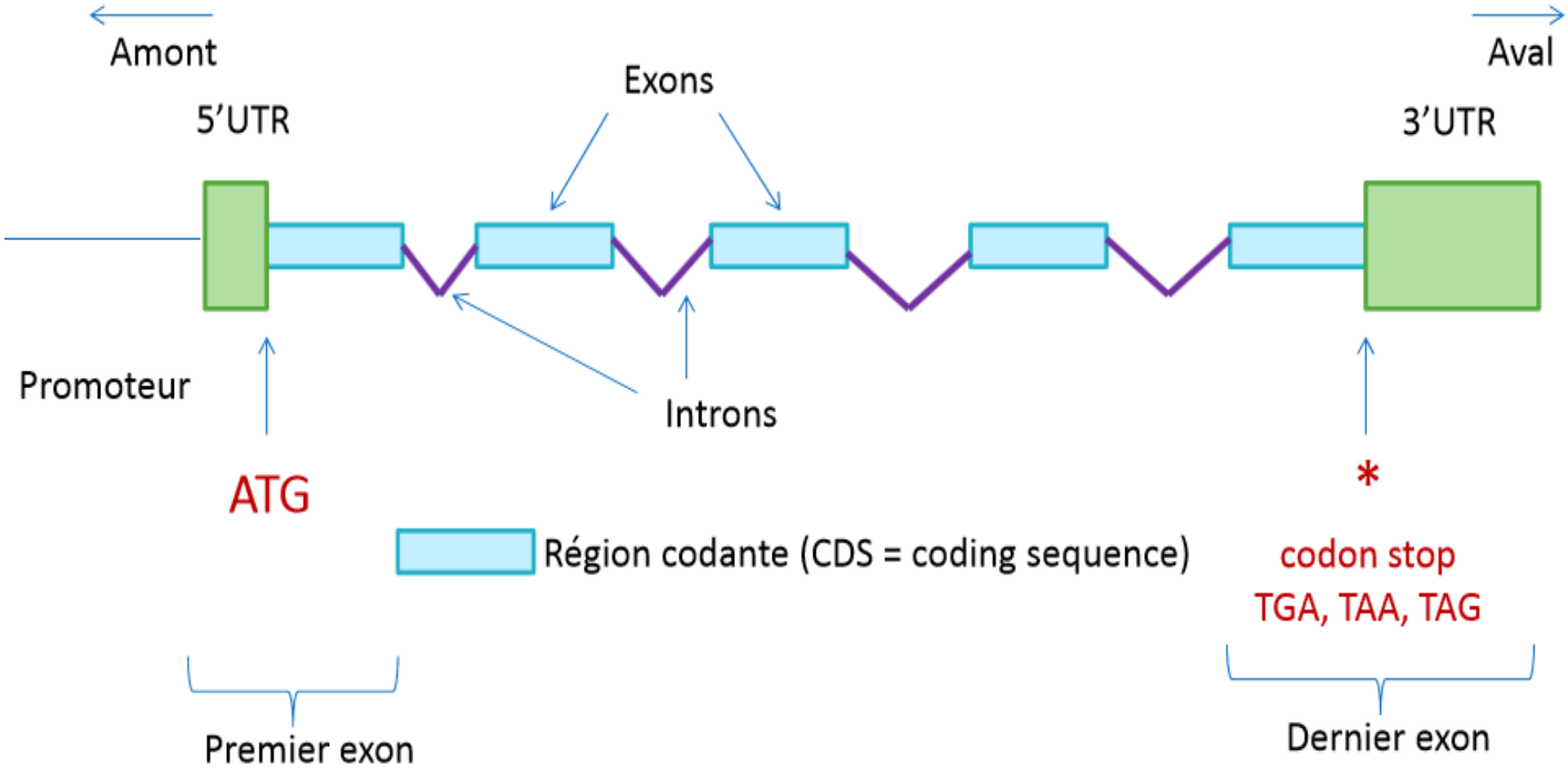
Certains ARN sont dit « non-codants »



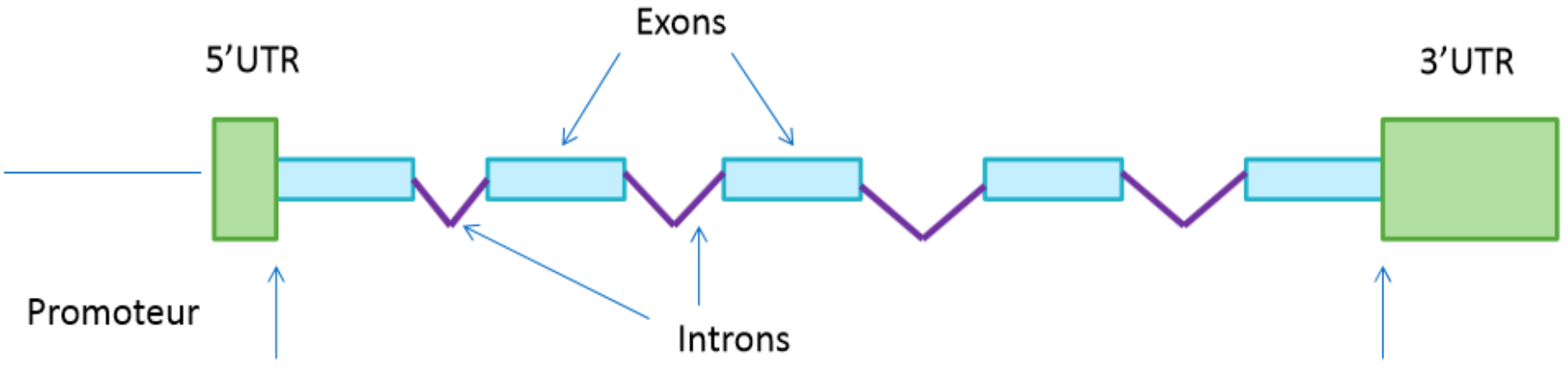
Rappel de biologie : convention Brin sens = Brin codant



- Représentation graphique du gène eucaryote:



- Effet d'une mutation sur la séquence nucléotidique selon sa localisation



Régulation ?

Protéine tronquée

Effet sur la séquence de la protéine traduite ?

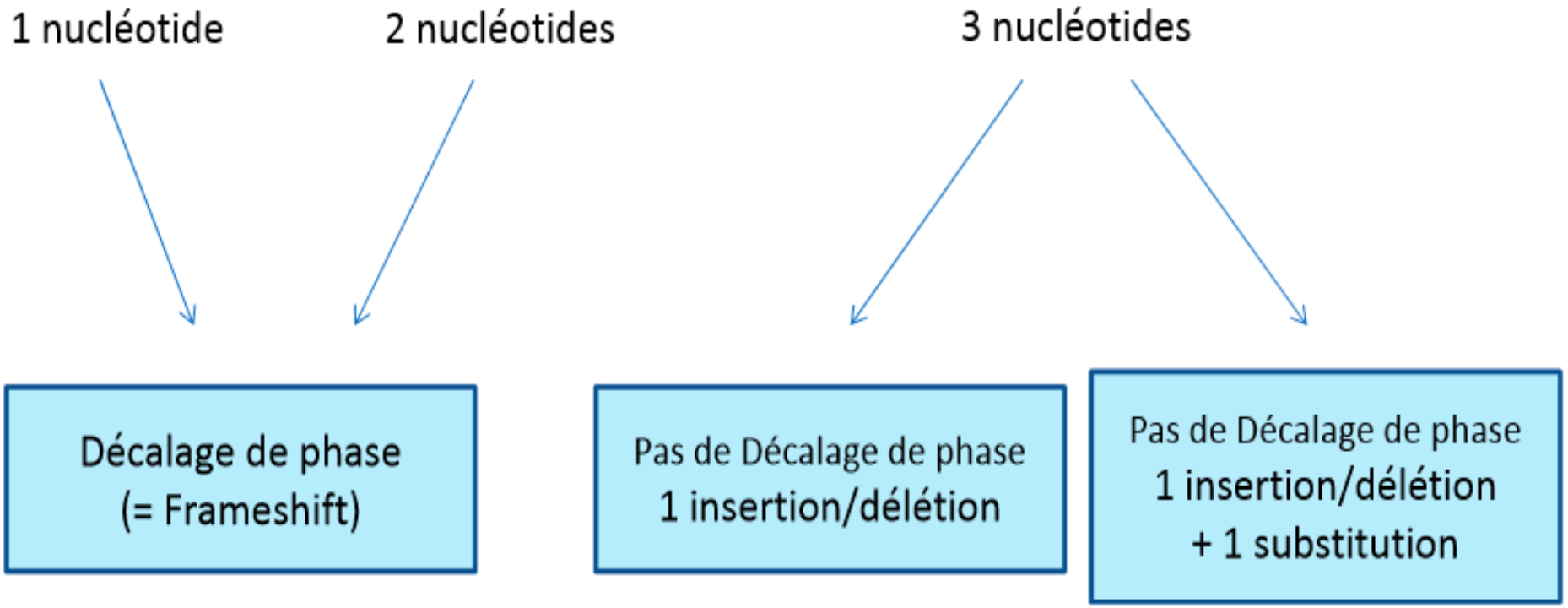
-Mutation faux-sens (Substitution d'acide aminé)
-Mutation synonyme (même acide aminé)
-Mutation non-sens (introduction d'un codon STOP)
-Insertion(s)
-Délétion(s)
-Défaut d'épissage (jonction intron/exon)

Exon

Régulation ?

Protéine plus longue

- Insertion/délétion d'1, 2 ou nucléotides groupés dans la région codante



- Exemples d'analyse bioinformatique d'une séquence :

⇒ Création de l'inverse complémentaire (reverse complementary) outil RevSeq

To upload a sequence from your local computer, select it here: Parcourir...

To enter the sequence data manually, type here:

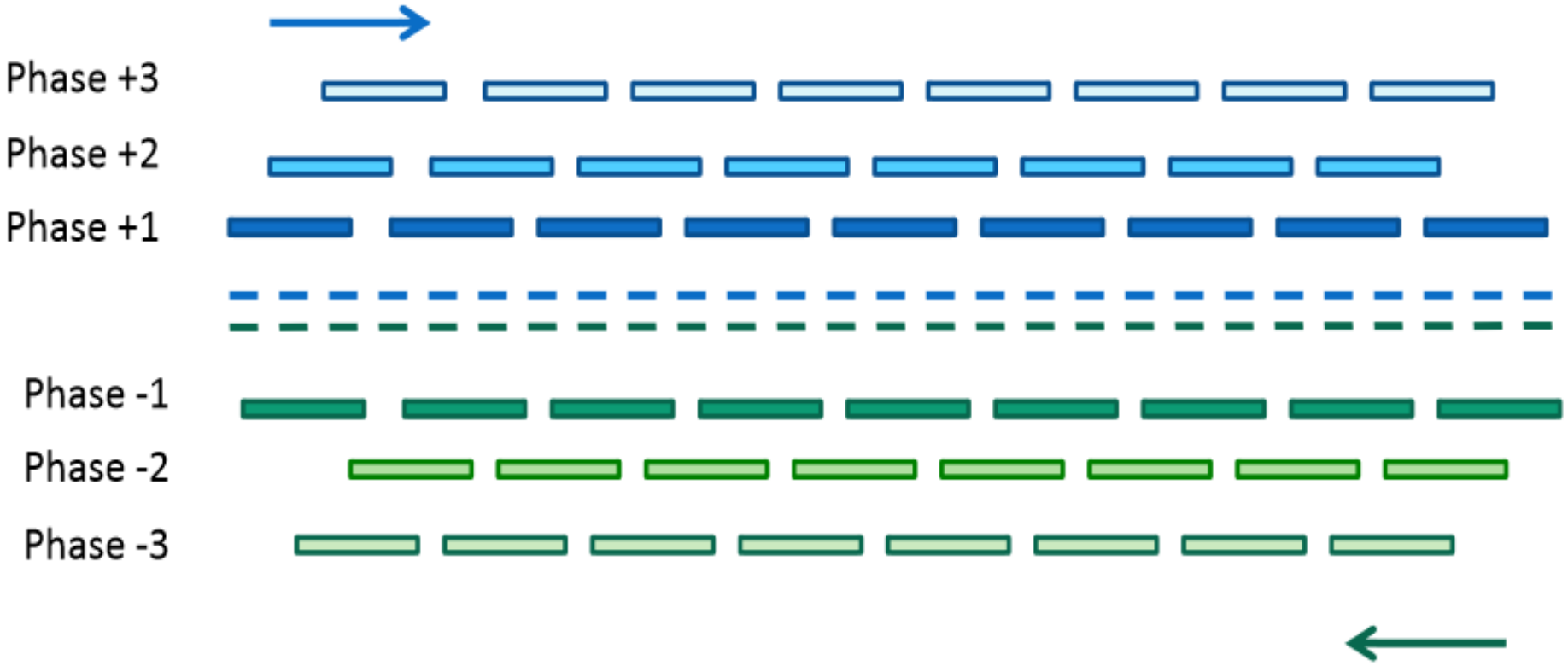
```
>sequence_1
AATCACAGTCAAAATACACCCAGATGCTCTCACACACCCAGACGCGCGC
```

OUTPUT FILE outseq

```
>sequence_1
GCGCGCGTCTGGGTGTGTGAGAGCATCTGGGTGTATTTTGACTGTGATT
```

- Exemples d'analyse bioinformatique d'une séquence :

⇒ traduction dans les 6 phases de lecture (= 6-frames translation)



C. Bases de données

Différentes catégories de bases de données :

Bases de données bioinformatiques

Bases généralistes
- banque N
- banque P

Bases spécialisées

Banques génomiques

Banques fonctionnelles :
- transcriptome
- protéome
- métabolome…

Difficultés:

- Harmoniser le classement des informations
- Utiliser un langage commun pour échanger des informations entre toutes ces bases

C. Bases de données

- Historique : Un besoin de stockage !

Dans les années 80 :

- Le nombre de séquences publiées augmente considérablement grâce aux avancées technologiques et un accès facile pour la communauté des biologistes doit être proposé.
- Les échanges de données informatiques commencent être facilités par le développement de réseaux informatiques
- Un consensus apparaît : il faut disposer de centres de références dans lesquels toutes les séquences connues seront déposées. Des serveurs "mondiaux" naissent :

1988 : NCBI aux USA / Base de données Genbank
1986 : DDBJ au Japon / Base de données DDBJ
1980 : EBI en Europe / Base de données EMBL
1986 : SIB en Suisse / Base de données SwissProt

} Séquences nucléiques

Séquences protéiques

Bases de données

- Organisation des données :
  - **Notion d'identifiant unique**

    Un identifiant permet de retrouver un élément dans un base de données de façon non ambiguë

  - **Fichiers Textes**

    Les informations peuvent être présentées dans une fiche (= un fichier texte) avec une fiche pour chaque élément de la base. Cette fiche peut être présentée ensuite sous format html avec des hyperliens, des illustrations....

  - **Base de données relationnelles**

    Souvent, les bases de données en biologie utilisent des outils informatiques de stockage de l'information = Système de gestion de Base de données relationnelles (SGBD)
    - Organisation des informations sous forme de tables ayant des liens entre elles
    - Efficacité de stockage et de recherche d'informations croisées (« requêtes »)

  - **Références Croisées**
    - Liens entre les différentes bases de données pour permettre aux biologistes de trouver un maximum d'informations

Bases de données

Echange des nouvelles soumissions toutes les 24h !

NCBI

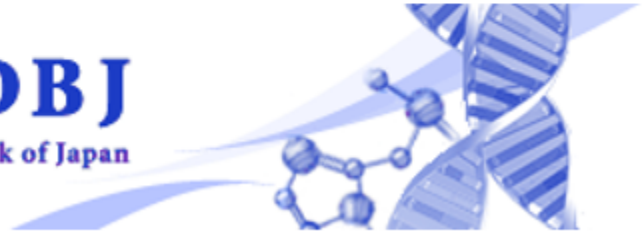
Genbank

EMBL NUCLEOTIDE SEQUENCE DATABASE

ENA : EMBL nucleotide Archive

DDBJ

DNA Data Bank of Japan



En une journée, la séquence soumise par le chercheur français à l'EMBL va se retrouver dans les 3 banques de données avec un reformatage spécifique à chaque banque.

Bases de données
Acides nucléiques

- Soumission d'une séquence et suite :
  - Le chercheur est l'auteur de la séquence, il soumet :
    - La séquence nucléotidique

      Attention, cette séquence peut contenir des erreurs de séquences :

      - erreur de séquençage
      - erreur de manipulation informatique (envoie de l'inverse complémentaire, séquence de vecteurs de clonage …)
    - Les informations supplémentaires = des annotations

      Organisme, position des gènes si ADN génomique, du CDS si ARNm…..

      Elle peut aussi contenir des erreurs d'annotations souvent dues au manque de connaissances biologiques à la date de soumission

  - Chaque banque réorganise l'information (identifiant, format spécifique)

Mise à jour

- Les annotations vont évoluer avec les nouvelles connaissances en biologie => Beaucoup d'annotations sont automatiques

- Des liens vers d'autres bases de données seront rajoutées

  Références croisées (= Cross-References)

➢ Harmonisation des fiches de données

En résumé, une fiche comporte de nombreuses informations :

Locus	Identificateur (nom et taille de la séquence)
Definition	Description de la séquence
Accession / version	Numéro d'accès dans la base
Keyword / Source / Organism / Reference / Authors / Title / Journal	Informations diverses (taxonomie, publications…)
Features	Caractéristiques de la séquence / produits d'expression
Origin	Séquence (par blocs de caractères / par lignes)
//	Fin de l'entrée dans la base

Bases de données protéiques

B

http://www.uniprot.org

UniProt Knowledgebase: Collaboration entre EBI, SIB et PIR

Décrire dans une fiche unique les produits derivés d'un gène dans une espèce donnée.

- UniProtKB/Swiss-Prot

Non-redondante, annotation manuelle.

- UniProtKB/TrEMBL

Traduction automatique de la base de données EMBL selon les annotation de CDS
Redondante, annotation automatique

C. Bases de données : c) UniProtKB

- TrEMBL

Ensemble des séquences protéiques conceptuelles obtenues par traduction automatique des séquences codante contenues dans EMBL, avec des annotations non vérifiées, mais avec l'objectif d'obtenir une couverture maximale

C. Bases de données : c) UniProtKB

- Les annotations :

  - ✓ Nom de la protéine, Nom du gène
  - ✓ Fonction
  - ✓ Activité enzymatique
  - ✓ Composition en domaines
  - ✓ Localisation cellulaire
  - ✓ Spécificité d'expression (tissus, stade de développement…)
  - ✓ Implication dans des pathologies
  - ✓ Effet des mutations
  - ✓ Interactions moléculaires
  - ✓ Liens vers d'autres base de données = Références croisées (EMBL, SMART,GO, PDB,OMIM….)

- # Notion d'ontologie :

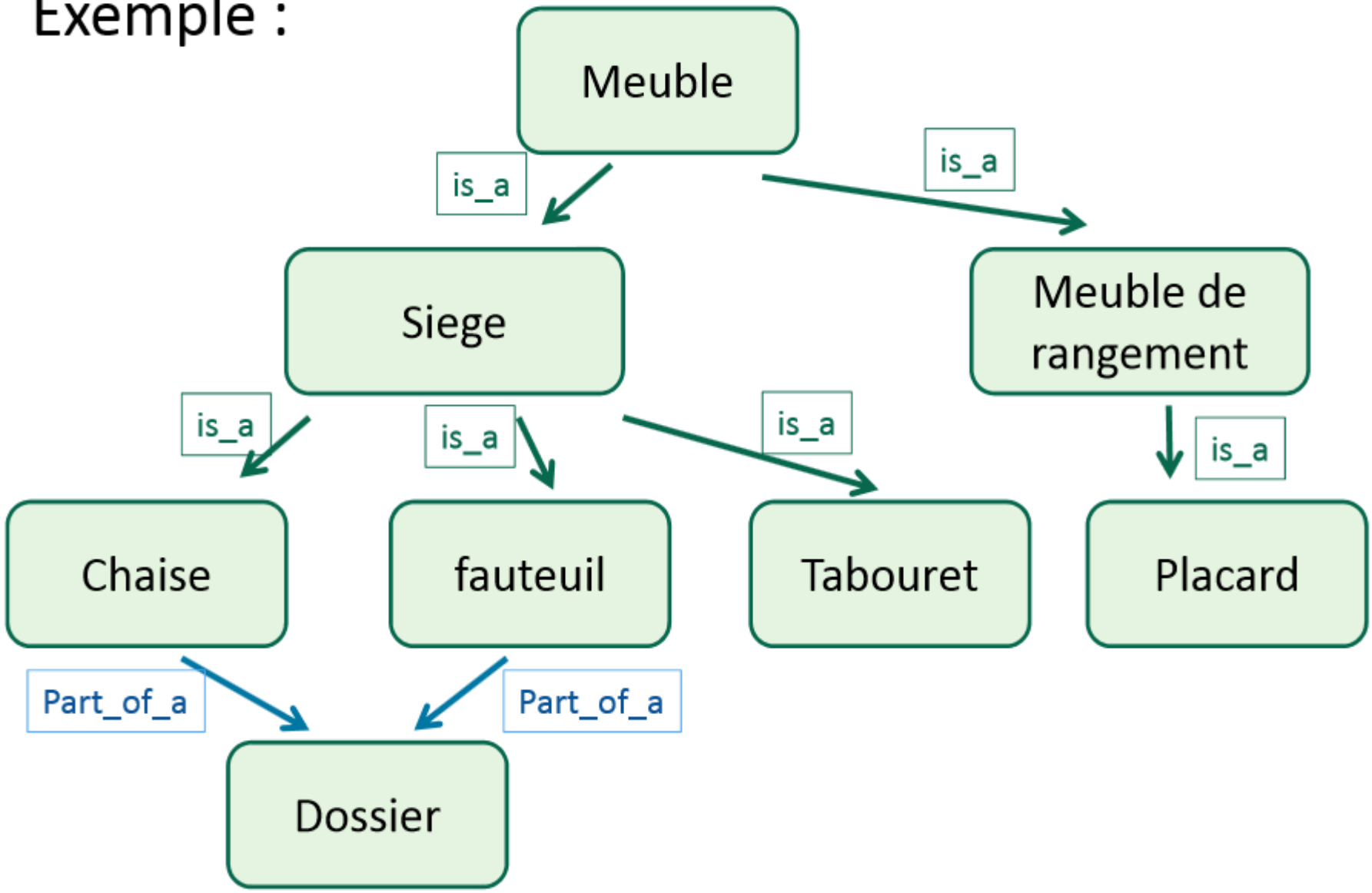
- Une ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations d'un domaine de connaissances.

- L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts.

=> Un recensement des concepts sous la forme d'un vocabulaire contrôlé.

⇒ Liaison de ces concepts par des relations qui modélisent notre connaissance. Exemple Gene Ontology (is_a , part_of_a)

- Exemple :


Meuble
is_a
is_a
Siege
Meuble de rangement
is_a
is_a
is_a
is_a
Chaise
fauteuil
Tabouret
Placard
Part_of_a
Part_of_a
Dossier

http://www.ebi.ac.uk/ego

QuickGO GO Term GO:0016303

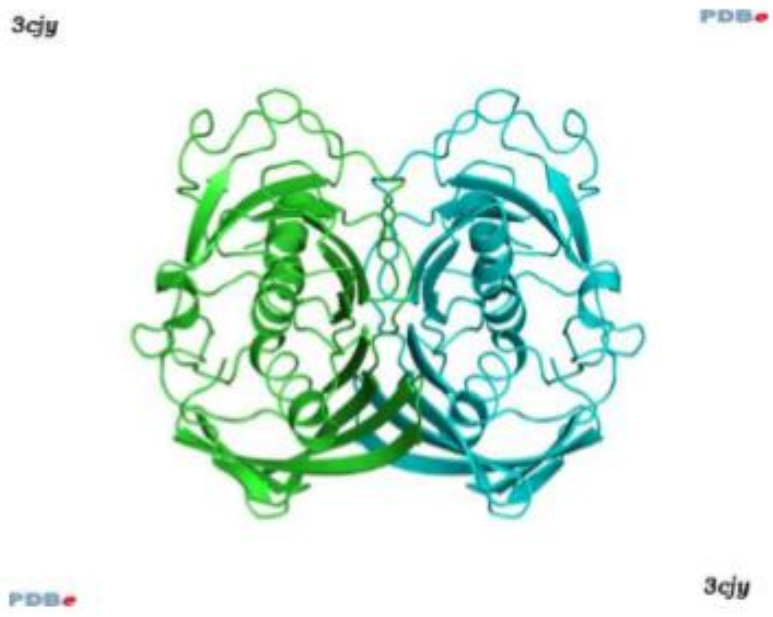
[?] = help

Term ID [?]	GO:0016303
Name [?]	phosphatidylinositol 3-kinase activity
Last updated [?]	2001-03-30 04:29:44.0
Definition [?]	Catalysis of the reaction: ATP + 1-phosphatidyl-1D-myo-inositol = ADP + 1-phosphatidyl-1D-myo-inositol 3-phosphate.
Synonyms [?]	PI3K
EC/TC mappings [?]	Enzyme 2.7.1.137 MetaCyc 1-PHOSPHATIDYLINOSITOL-3-KINASE-RXN

Often Annotated With [?]	Term	Significance	Other	Both	This
	phosphoinositide 3-kinase complex	83%	94	93	111

InterPro Mappings [?]

IPR000341: Phosphoinositide 3-kinase, ras-binding
IPR001720: PI3 kinase, P85 regulatory subunit
IPR002420: Phosphoinositide 3-kinase, C2
IPR003113: Phosphatidylinositol 3-kinase, p85-binding
IPR008290: Phosphatidylinositol 3-kinase, Vps34 type



- Protein Data Bank ou PDB est une collection mondiale de données sur la structure tridimensionnelle (ou structure 3D) de macromolécules biologiques : protéines, essentiellement, et acides nucléiques.

- Ces structures sont essentiellement déterminées par cristallographie aux rayons X ou par spectroscopie RMN.