ANALYSE DES SÉQUENCES BIOLOGIQUES

Génome aux Protéomes « In Silico »

Master M2: GADM

Cours Présenté par : Mme Mohamed Ben Ali

OBJECTIFS DU COURS

- COMPRENDRE LE DOMAINE DE LA BIO-INFORMATIQUE
- LE RÔLE DE L'INFORMATIQUE ET L'IA DANS CES PROBLÉMATIQUES
- COMPRENDRE LES ALGORITHMES DE BASE DE LA FOUILLE DE DONNÉES POUR LA BIOINFORMATIQUE
- UTILISER R ET DES APPLETS

C'EST QUOI LA BIOINFORMATIQUE :1

La bioinformatique : Traitement des informations biologiques par des méthodes informatiques et/ou mathématiques.

La bioinformatique est fondée sur les acquis:

- de la biologie,
- des mathématiques
- de l' informatique.

C'est l'approche "*in silico*", qui vient compléter les approches classiques

- "*in situ*" (dans le milieu naturel),
- "*in vivo*" (dans l'organisme vivant)
- "*in vitro*" (en éprouvette) de la biologie traditionnelle.

C'EST QUOI LA BIOINFORMATIQUE :2

La bioinformatique a fait son apparition dans les années 1980 avec les premières banques de biomolécules (EMBL et GenBank).

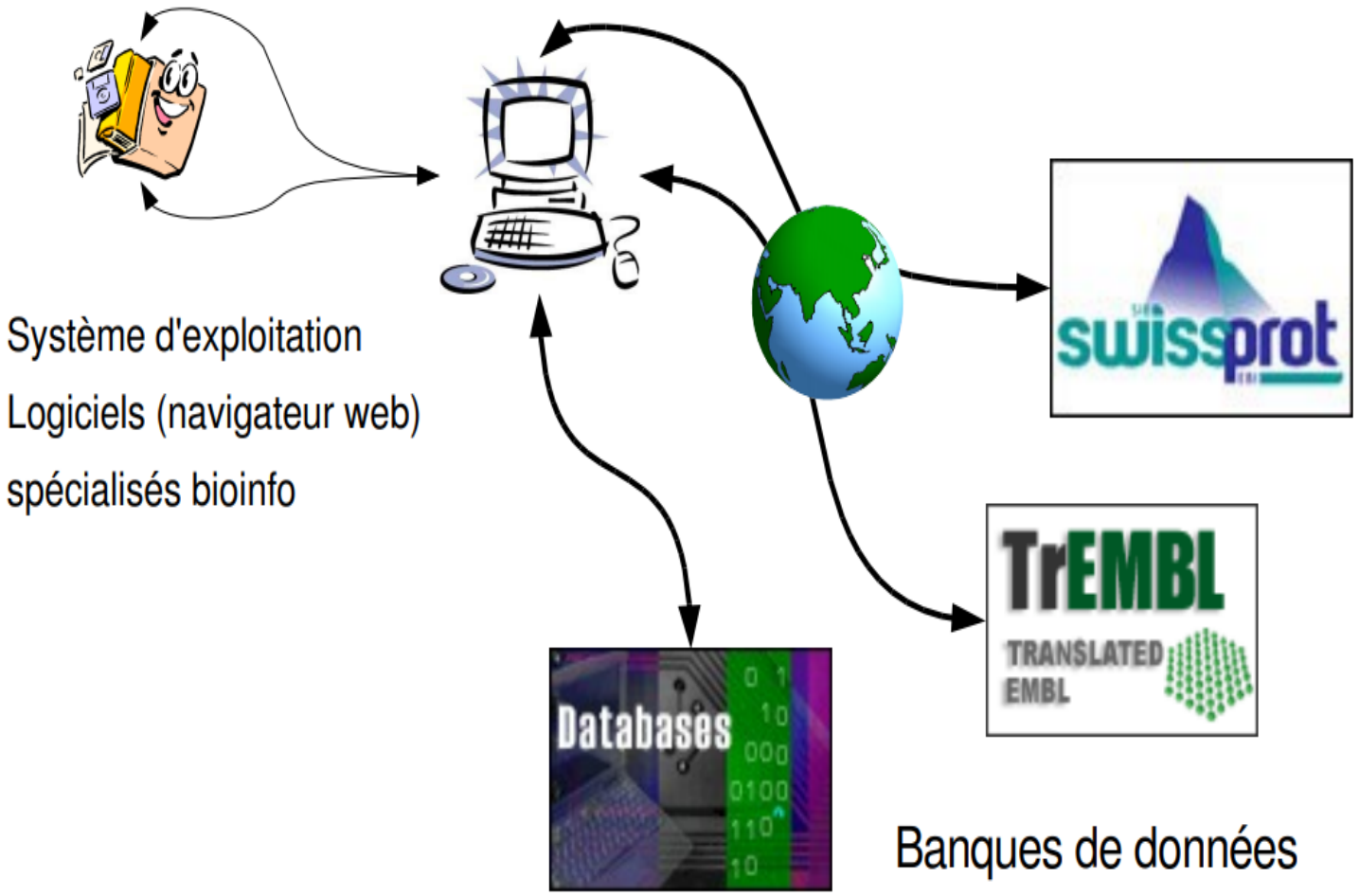
Elle propose des méthodes et des logiciels qui permettent:

- de gérer,
- d'organiser,
- de comparer,
- d'analyser,
- d'explorer

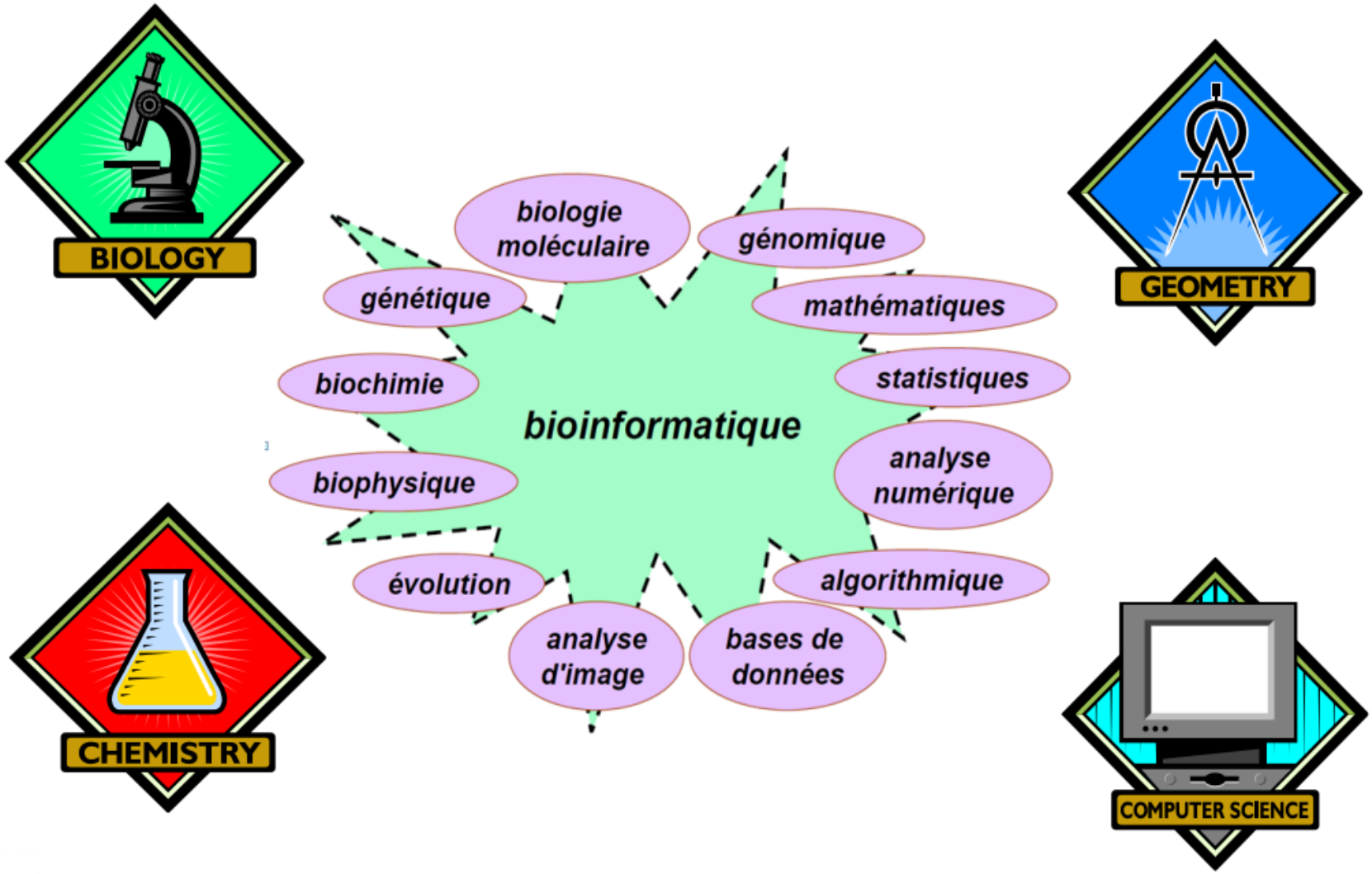
} l'information génétique et génomique stockée dans les bases de données

But : Prédire et produire des connaissances nouvelles dans le domaine ainsi qu'élaborer de nouveaux concepts

RELATION DIRECTE
ENTRE BIOLOGIE ET INFORMATIQUE



BIOINFORMATIQUE
PAS QUE BIOLOGIE ET INFORMATIQUE !
MAIS INTERDISCIPLINAIRE



OBJECTIFS

Acquérir puis stocker les informations biologiques sous la forme d'encyclopédies appelées bases de données;

Développer des programmes de prédiction et d'analyse en utilisant les informations contenues dans les bases de données;

Analyser/Interpréter/Prédire: utiliser ces programmes pour analyser de 'nouvelles' données biologiques et prédire *in silico* par exemple la fonction potentielle d'une protéine;

Visualiser: développer des programmes pour visualiser la structure en trois dimensions des protéines et de l'ADN, pour shématiser des voies métaboliques ou des arbres phylogénétiques.

Qu'est-ce qu'un génome?

- Des gènes :
  - portions d'ADN codant des protéines
  - portions d 'ADN codant des ARN : ARNr, ARNt, ARNsn, …
  - portions d 'ADN codant des ARN non traduits

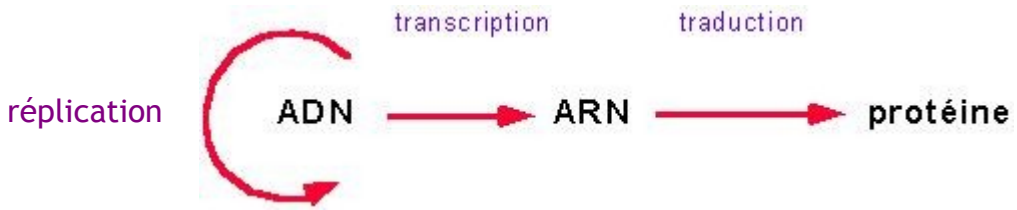
- Eléments régulateurs : promoteurs, enhancers, …

- Eléments requis pour la réplication des chromosomes : origines de réplication, télomères, centromères, …

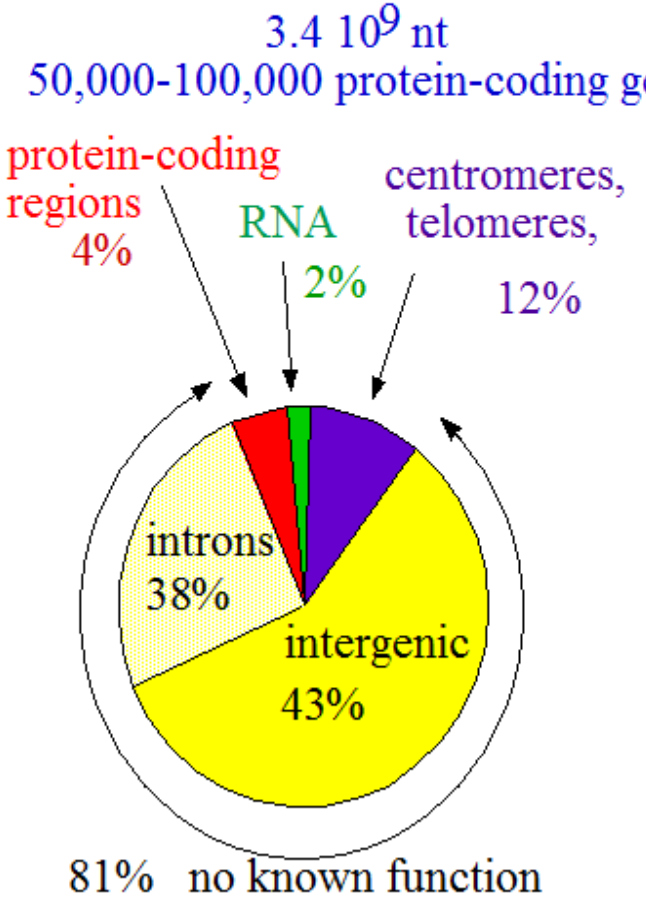
- Séquences non fonctionnelles :
  - séquences non codantes
  - séquences répétées
  - pseudogènes

LE DOGME CENTRAL DE BIOLOGIE MOLÉCULAIRE

- Séquence d'opérations de l'ADN aux protéines
  - *transcription* : l'ADN est copié en ARNm
  - *traduction* : l'ARNm est traduit en protéines par les ribosomes
  - protéines sont les ouvrières du monde cellulaire
- Le code de l'ADN est responsable de la vie cellulaire



Compartimentation fonctionnelle du génome humain



ADN INFORMATIF (19%)
- gènes codant pour les protéines (5%)
- gènes codant pour les ARN (2%)
  (ARNt-ARNr-ARNn)
- gènes régulateurs (12%)
  (recombinaison - replication - ségrégation)

ADN NON INFORMATIF (81%)
- séquences hautement répétées (10%)
  (ADN satellite - minisatellite - microsatellite)
- séquences moyennement répétées (30%)
  (rétroéléments SINES - ALU (10%)
  LINES - L1 (5%)
  Autres répétitions (15%))
- séquences uniques ou très peu répétées (60%)
  (introns - pseudogènes)

Quelles Types d'informations: «OMES» ?

Génome (l'ensemble du matériel génétique d'un individu ou d'une espèce.)

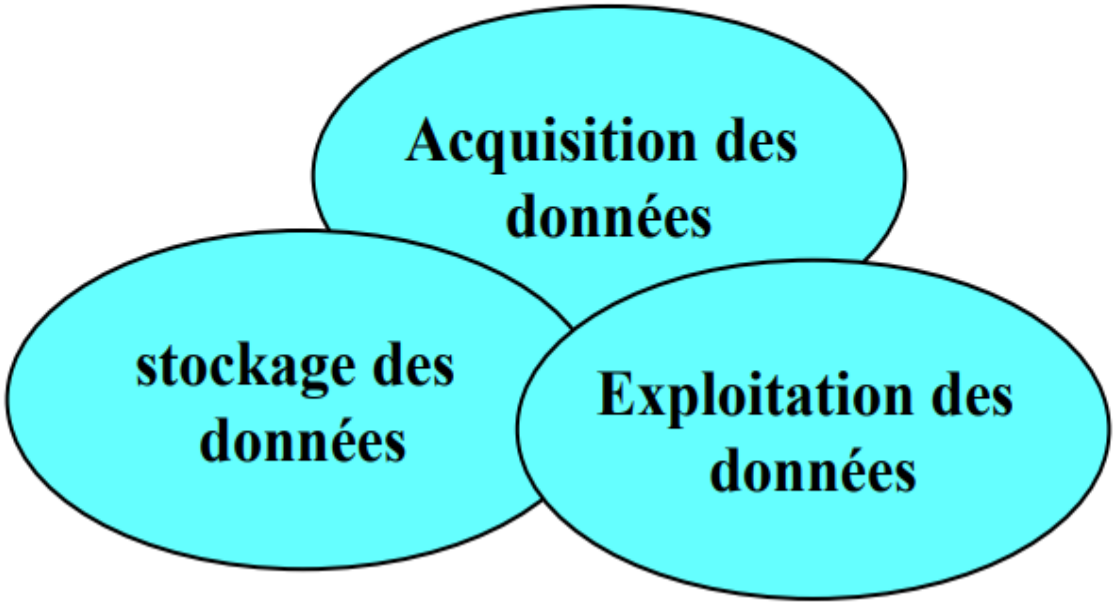
Transcriptome (l'ensemble des ARN messagers transcrits à partir du génome)

Protéome (l'ensemble des protéines exprimés à partir du génome)

Métabolome (l'ensemble des composés organiques (sucres, lipides, amino-acides, …)

Intéractome (l'ensemble des interactions protéine-protéine)…

Trois grands domaines où intervient la bioinformatique

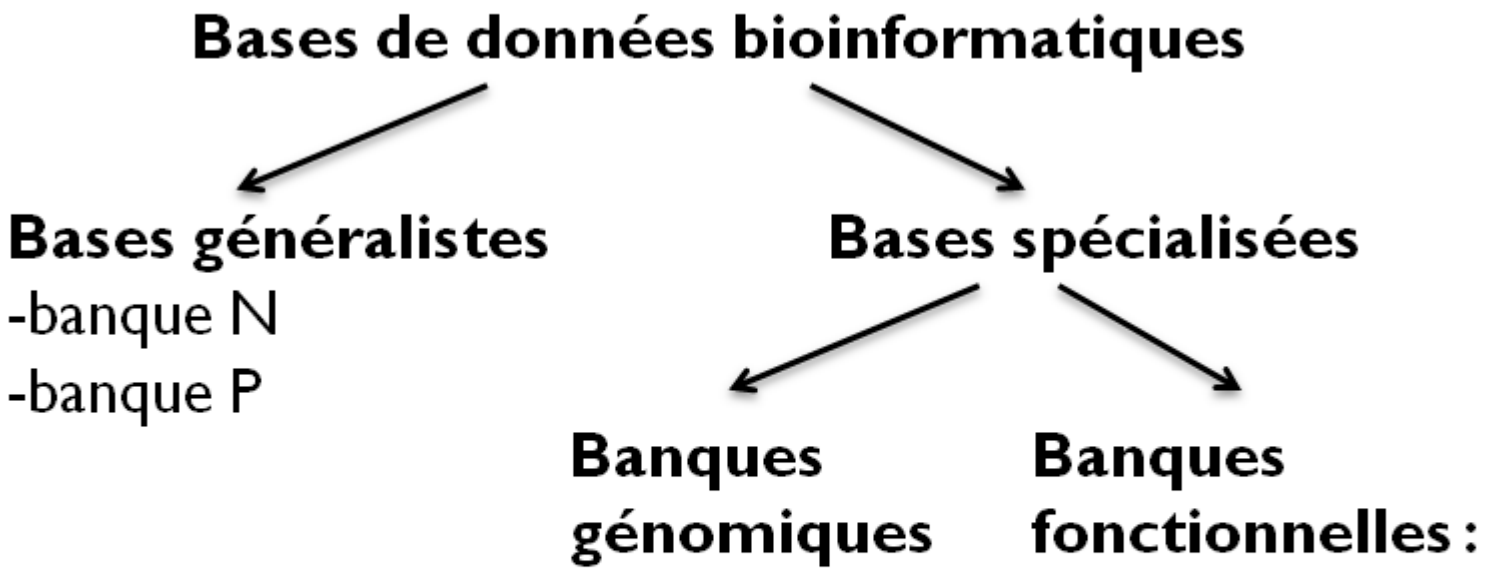


QU'EST CE QU'UNE BANQUE DE DONNÉES ?

- Ensemble de données relatives à un domaine, organisées par traitement informatique, accessibles en ligne et à distance
- Souvent, les données sont stockées sous la forme d'un fichier texte formaté (respectant une disposition particulière)
- Besoin de développer des logiciels spécifiques pour interroger les données contenues dans ces banques

LES BANQUES DE DONNÉES

- Différentes catégories de bases de données :



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

LES BANQUES DE DONNÉES GÉNÉRALISTES

- Ces banques contiennent des données hétérogènes
  - Collecte la plus exhaustive possible
  - Banques de séquences nucléiques
  - Banques de séquences protéiques
  - Banques de structure 3D de macromolécules
  - Banques d'articles scientifiques
- **Avantage :** tout est consultable en une fois
- **Inconvénients :** difficiles à maintenir, difficiles à interroger

LES BANQUES DE DONNÉES SPÉCIALISÉES

- Ces banques contiennent des données homogènes
  - Collecte établie autour d'une thématique particulière
- **Avantages :** facilité pour mettre à jour les données, vérifier leur intégrité, offrir une interface adaptée, ...
- **Inconvénients :** ne cible pas toujours ce que l'on veut; toutes les banques possibles n'existent pas
- **Exemples :** banques spécialisées pour un génome, banques de séquences d'immunologies, banques sur des séquences validées, ...

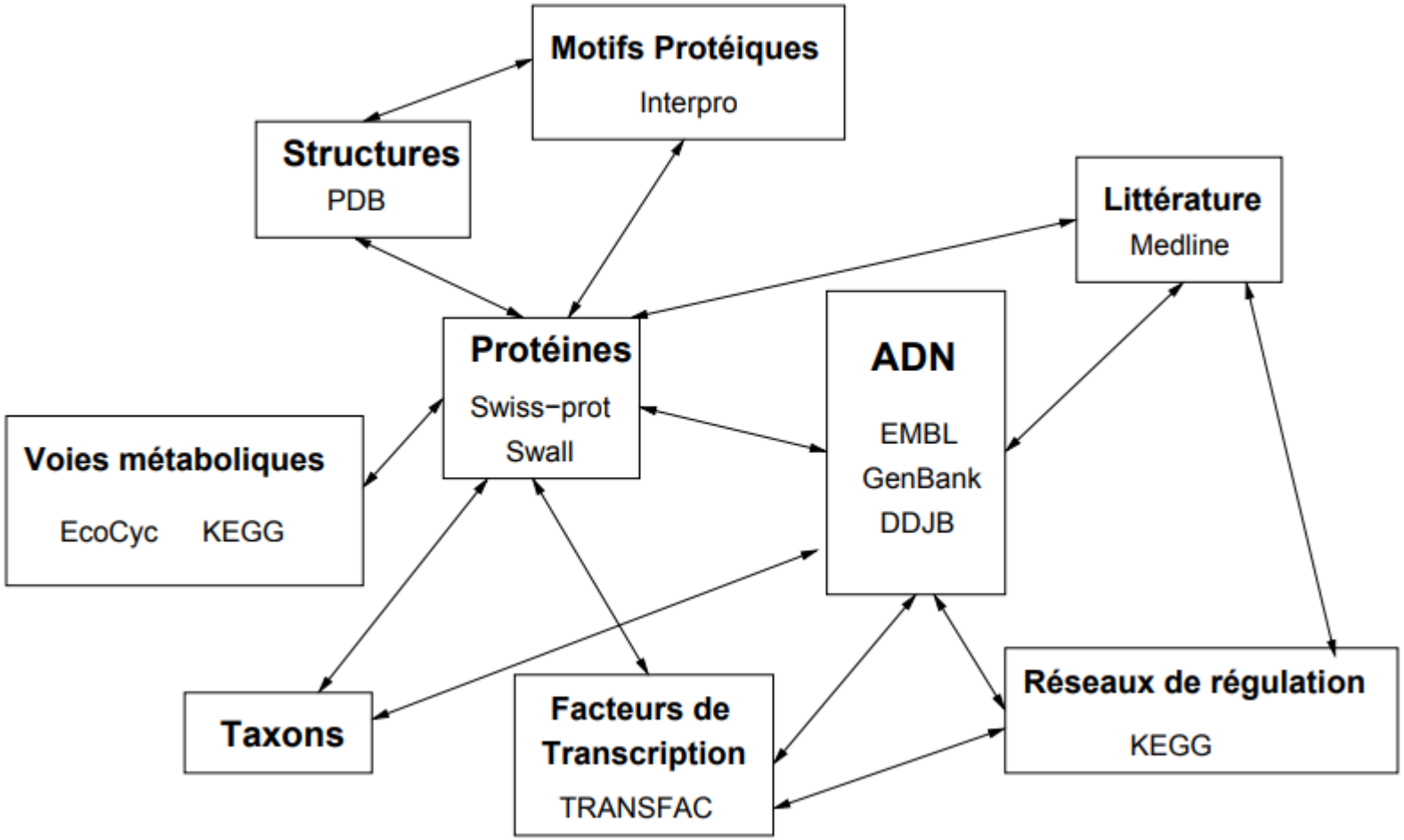
LES BANQUE DE SÉQUENCES NUCLÉIQUES

- **Origine des données :**
  - Séquençage d'ADN et d'ARN
- **Les données stockées :** séquences + annotations
  - Fragments de génomes
    - Un ou plusieurs gènes, un bout de gène, séquence intergénique, …
  - Génomes complets
  - ARNm, ARNt, ARNr, … (fragments ou entiers)
- [**NB 1**] **:** toutes les séquences (ADN ou ARN) sont écrites avec des T (thymine)
- [**NB 2**] **:** les séquences sont toujours orientées 5' vers 3'.

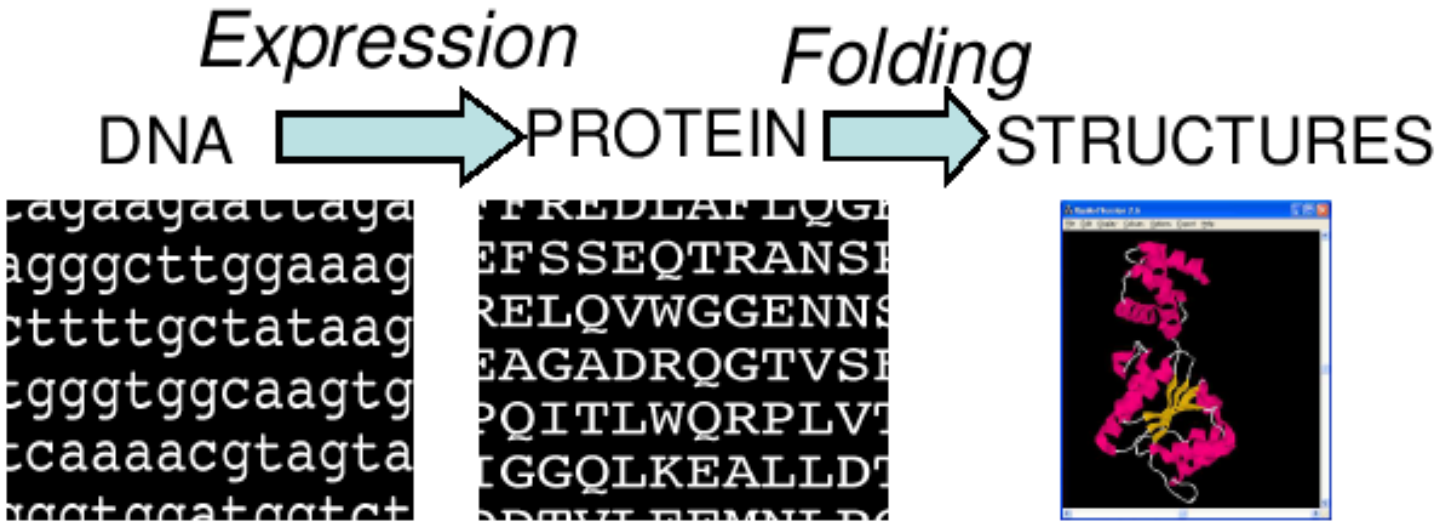
LES BANQUE DE SÉQUENCES PROTÉIQUES

- Origine des données
  - Traduction de séquences d'ADN
  - Séquençage de protéines
    - Rare car long et coûteux
  - Protéines dont la structure 3D est connue
- Les données stockées : séquences + annotations
  - Protéines entières
  - Fragments de protéines

APERÇU DE CERTAINES BANQUES DE DONNÉES



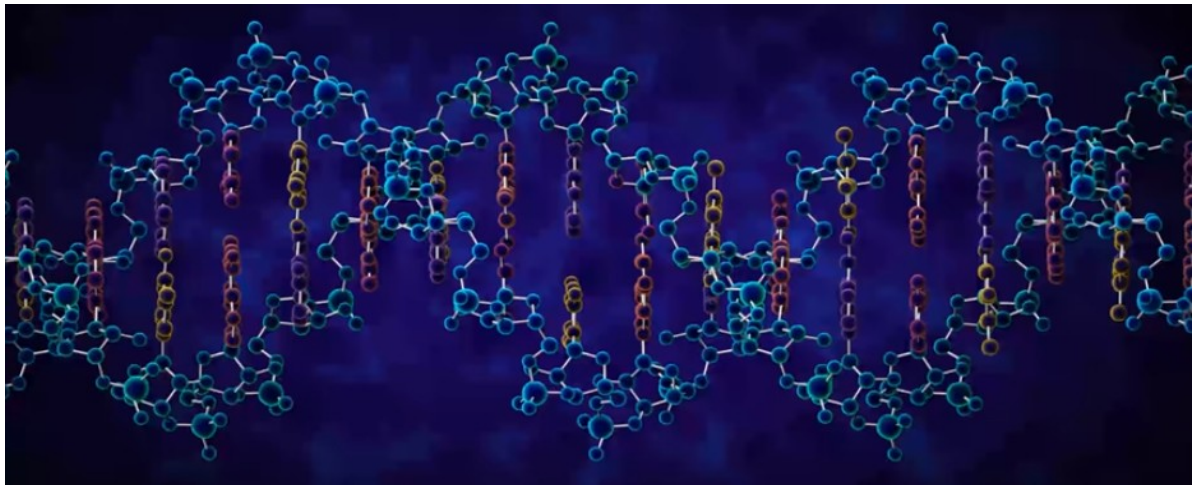
LES DONNÉES BIOLOGIQUES



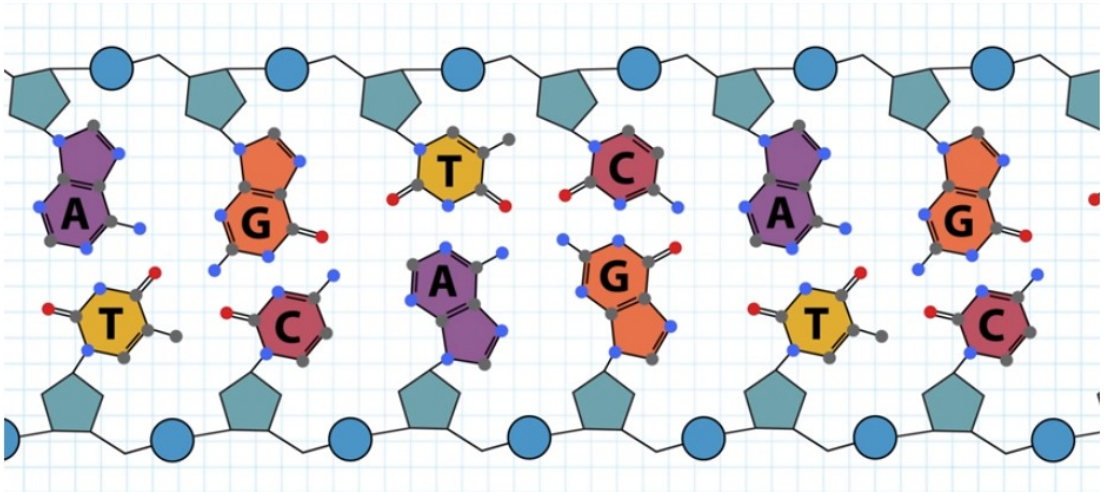
- La bioinformatique utilise 3 sources de données :
  - Les séquences de nucléotides (ADN - ARNm)
  - Les séquences d'aminoacides
  - Des informations sur les protéines (notamment leur structures)

1. ADN : UNE SÉQUENCE DE NUCLÉOTIDES

STRUCTURE : DOUBLE HÉLICE = 2 BRINS

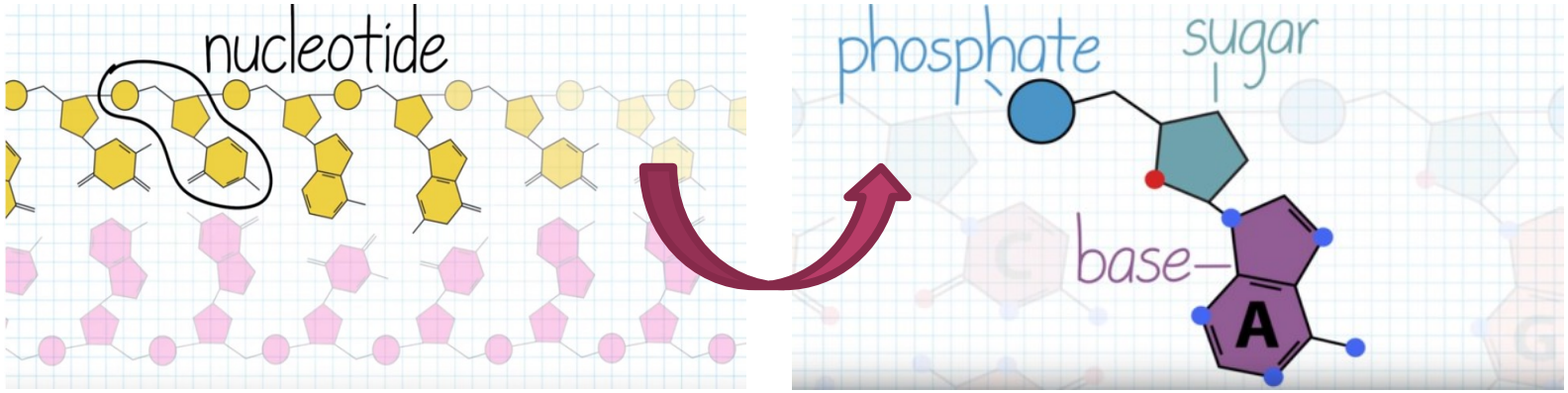


STRUCTURE CHIMIQUE ADN

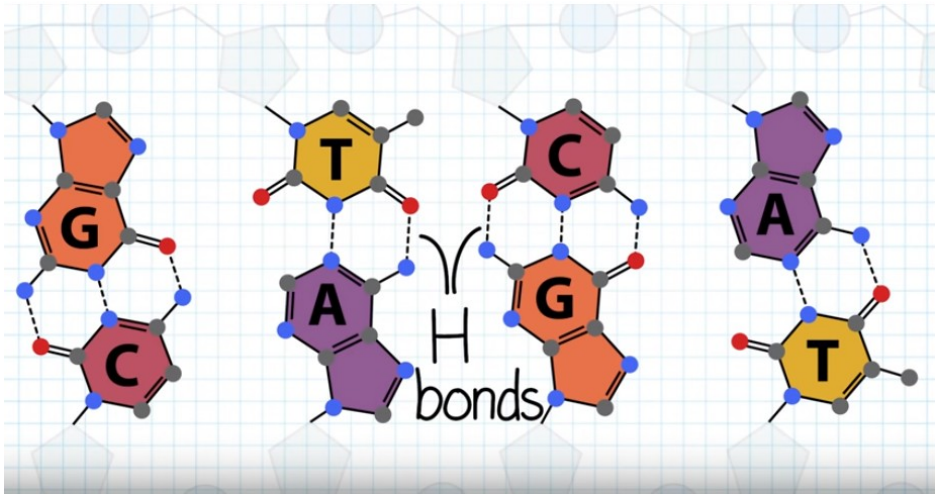


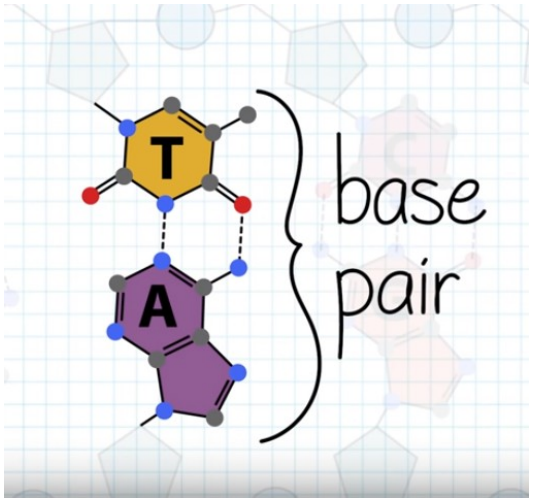
STRUCTURE DE L'ADN

STRUCTURE NUCLÉOTIDE



LIAISONS INTRA-NUCLÉOTIDES COMPLÉMENTAIRES

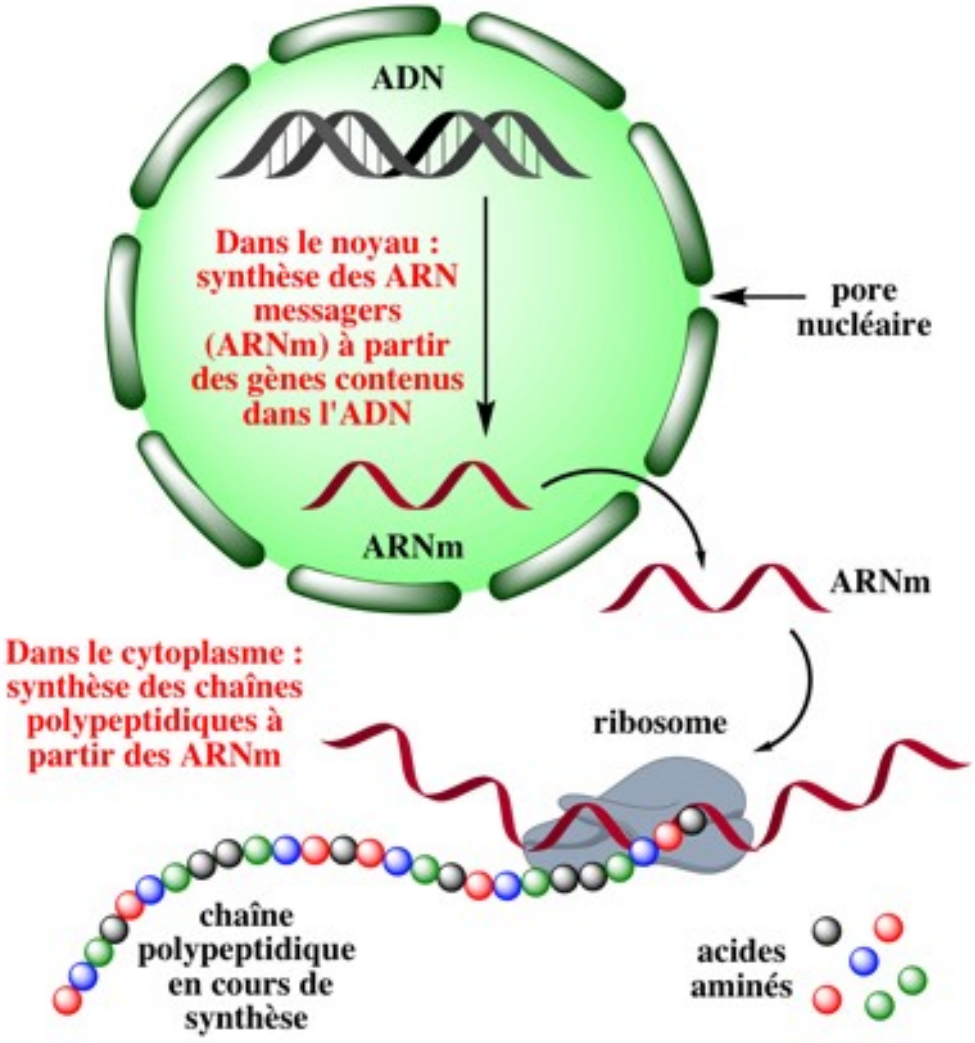




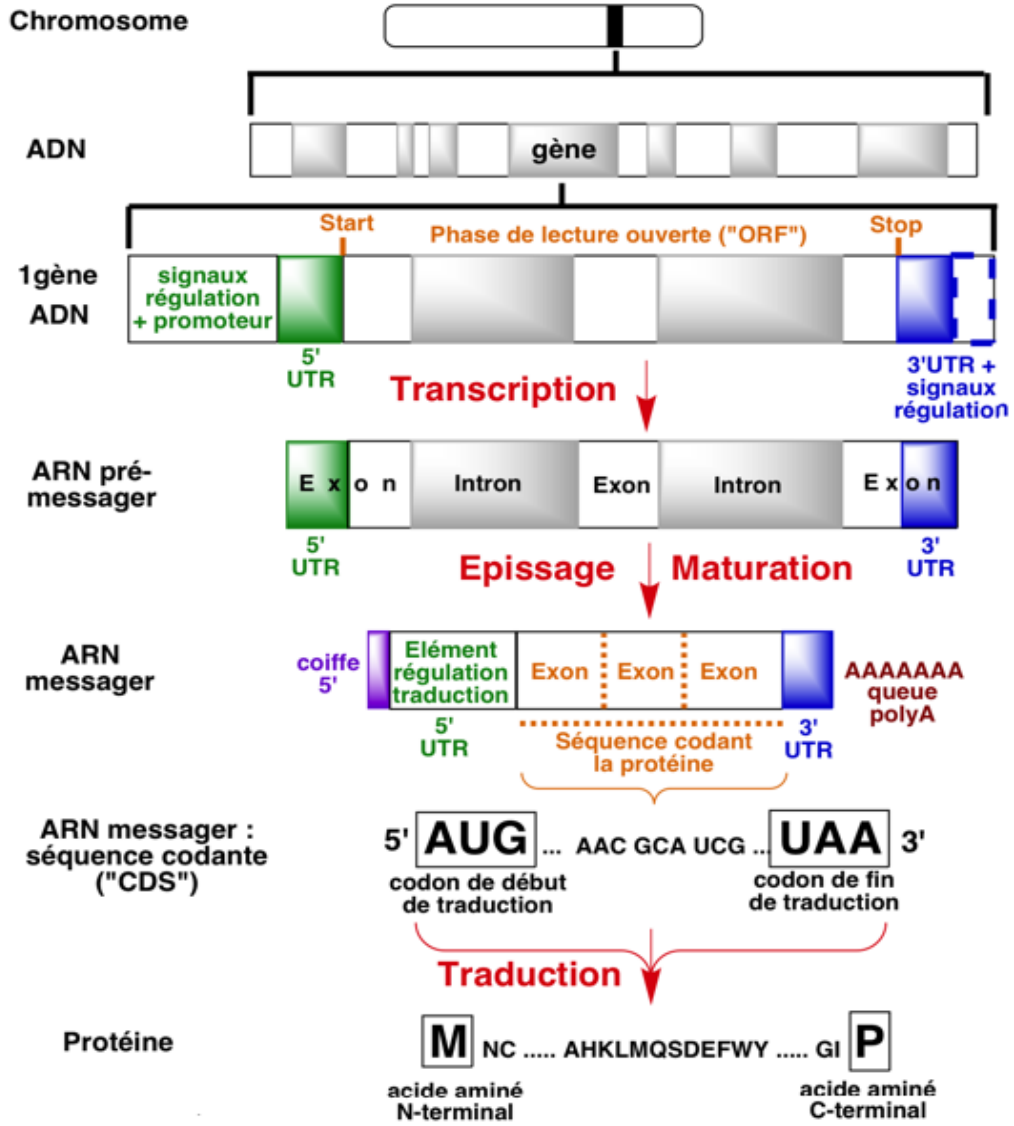
THYMINE-ADENINE = 2H BONDS
CYTOSINE-GUANINE = 3H BONDS

PASSAGE DE L'ADN AUX PROTÉINES

SYNTHESE DES PROTEINES : UNE ACTIVITE DE BASE POUR L'ADN



CYCLE DE SYNTHÈSE DE PROTÉINES, PLUS ….



LES EXONS SONT DES PARTIES DU GÈNE QUI CODENT POUR LES PROTÉINES. ILS SONT TRANSCRITS ET TRADUITS POUR RÉALISER UN PRODUIT FINI.

ACIDES AMINÉS NATURELS
- NOMENCLATURE DES 20 AA

- Les acides aminés sont des composants organiques et constituants fondamentaux des protéines.
- Plus de 500 acides aminés ont été inventoriés dont seulement 22 AA apparaissent dans le code génétiques. .
- Source: apport alimentaire, catabolisme des protéines.

Nom	Code à 3 lettres	Code à 1 lettre	Nom	Code à 3 lettres	Code à 1 lettre
Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Méthionine	Met	M
Acide aspartique	Asp	D	Phénylalanine	Phe	F
Acide glutamique	Glu	E	Proline	Pro	P
Cystéine	Cys	C	Sérine	Ser	S
Glutamine	Gln	Q	Thréonine	Thr	T
Glycine	Gly	G	Tryptophane	Trp	W
Histidine	His	H	Tyrosine	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V

- En plus de 2 AA découverts récemment: Sélénocystéine (U) et Pyrrolidine (2002).

ACIDES AMINÉS : TYPES

1. ACIDES AMINÉS ESSENTIELS

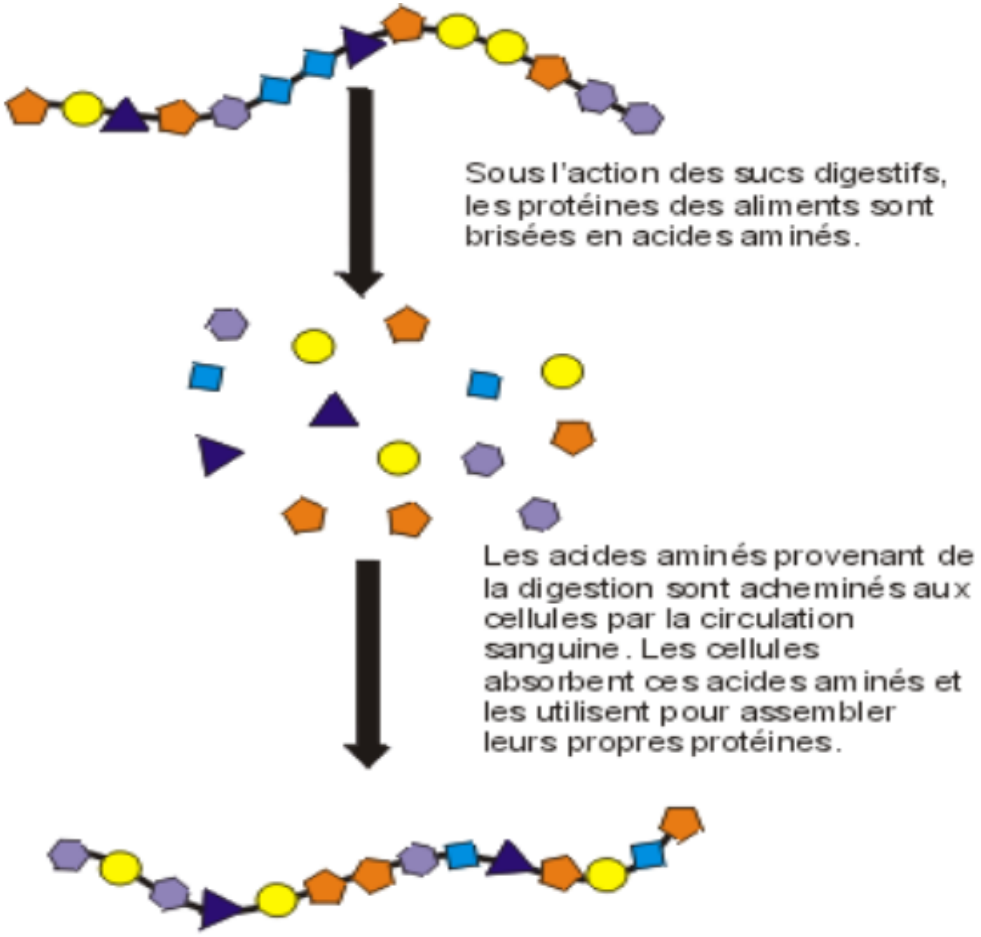
Méthionine	Met	M
Leucine	Leu	L
Valine	Val	V
Lysine	Lys	K
Isoleucine	Ile	I
Phénylalanine	Phe	F
Tryptophane	Trp	W
Histidine	His	H
Thréonine	Thr	T
Arginine	Arg	R

- Ils sont au nombre de 8 AA: V, L, I, F, W, **K**, M,**T**
- On ajoute 2 AA, H et R essentiels en cas de grossesse, nourrisson

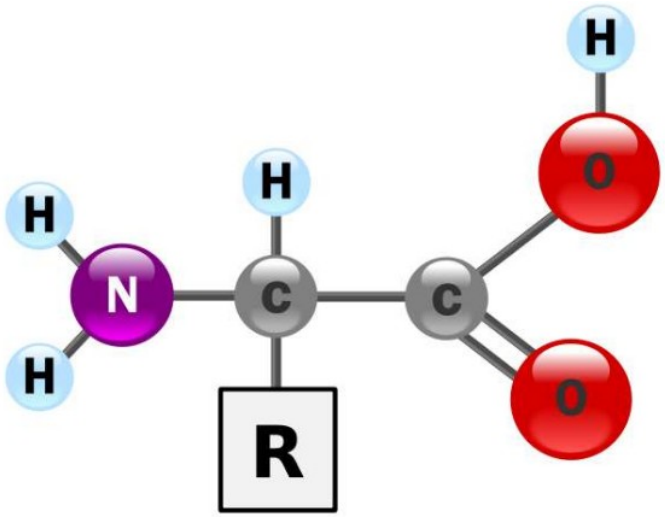
2. ACIDES AMINÉS NON ESSENTIELS

- Ce sont tous les autres acides aminés synthétisés par la cellule pour le besoin de l'organisme.

COMMENT LES ACIDES AMINÉS ESSENTIELS SONT-ILS ABSORBÉS DANS L'ORGANISME

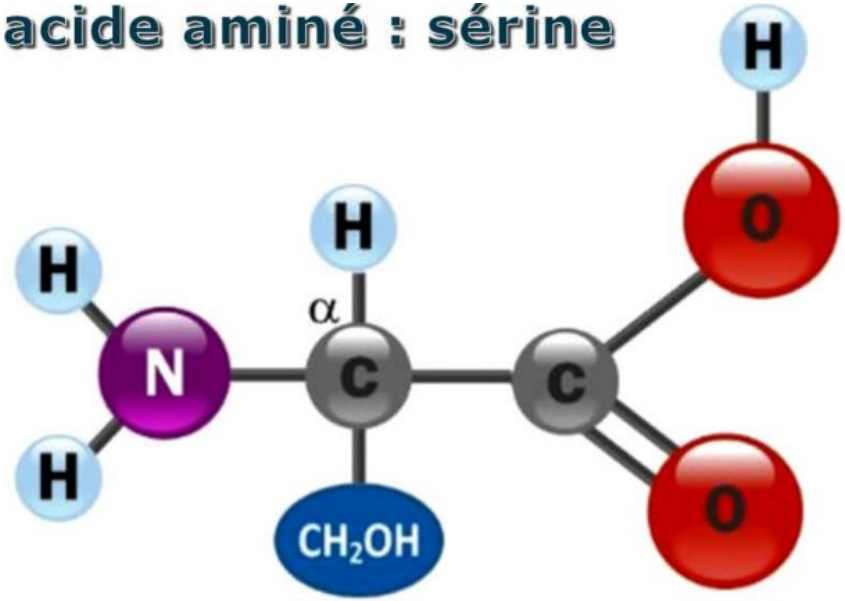


STRUCTURE DE BASE ACIDE AMINÉ

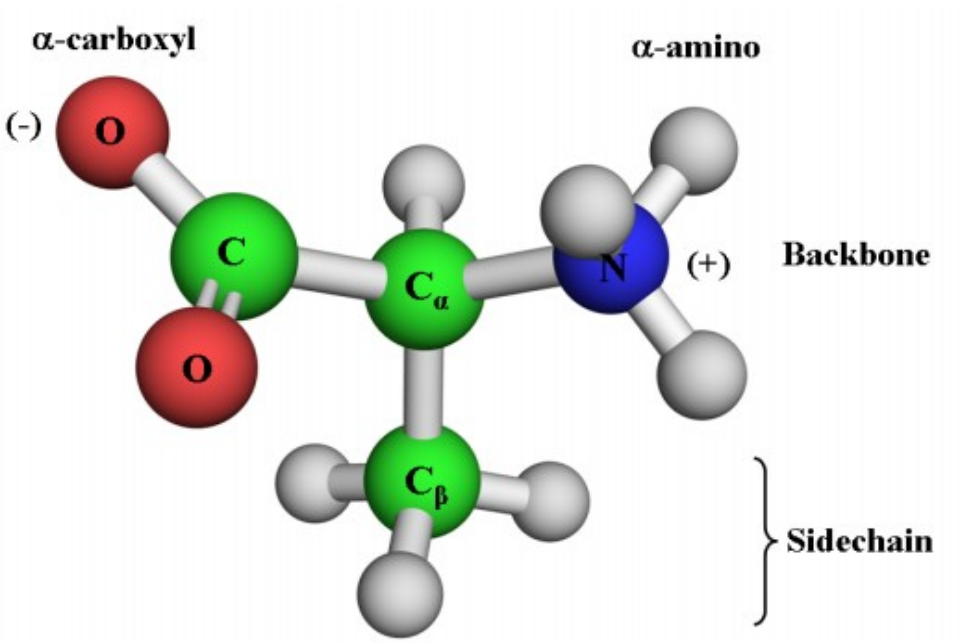


- Les éléments clés d'un AA sont: le carbone, hydrogène, nitrogène (azote) et l'oxygène.
- Un AA est composé de : un groupe amine, un groupe carboxylique(acide) et un radical ou résidu.
- Le radical est la partie variable entre les 22 AA.

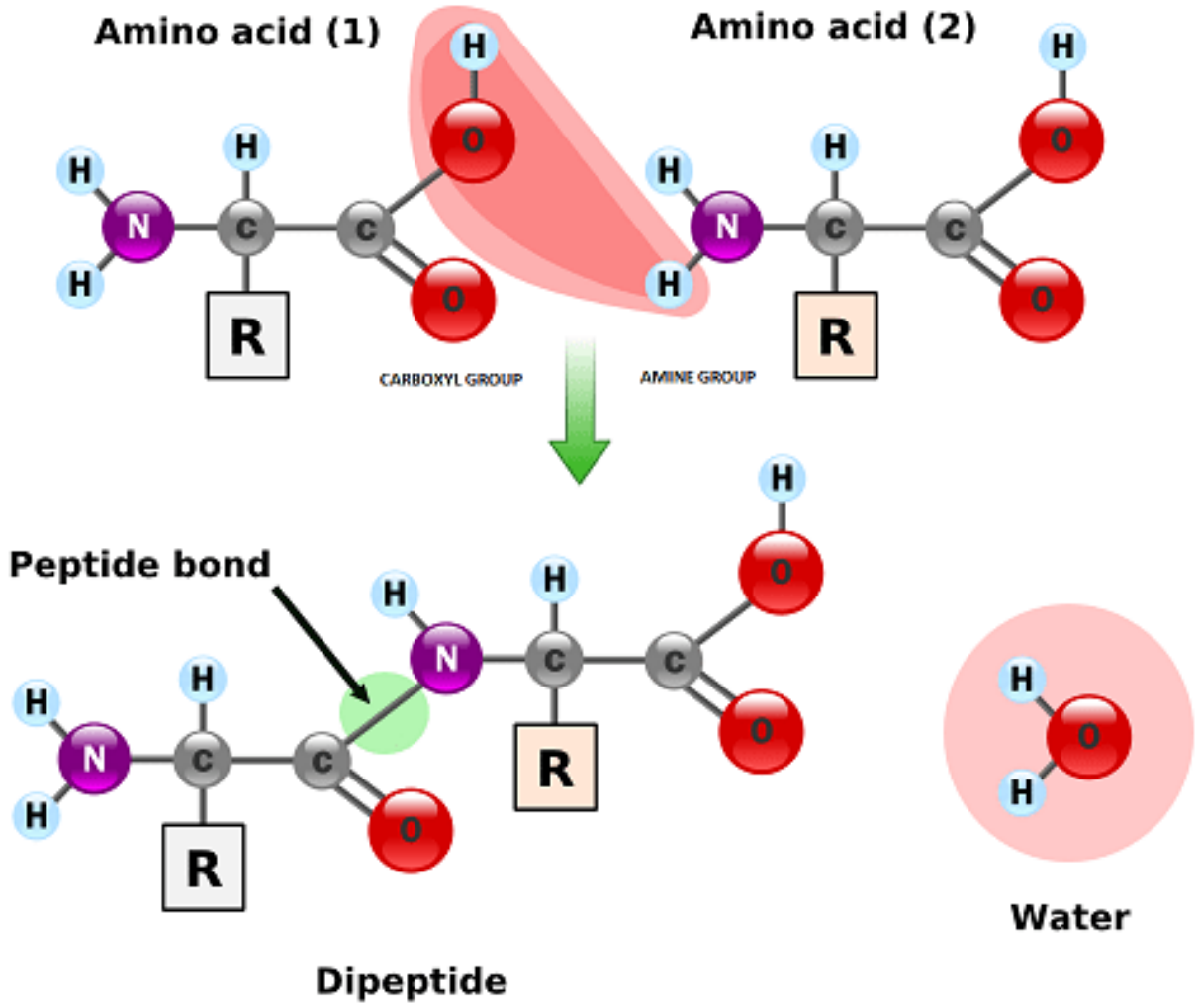
acide aminé : sérine



STRUCTURE DE BASE ACIDE AMINÉ

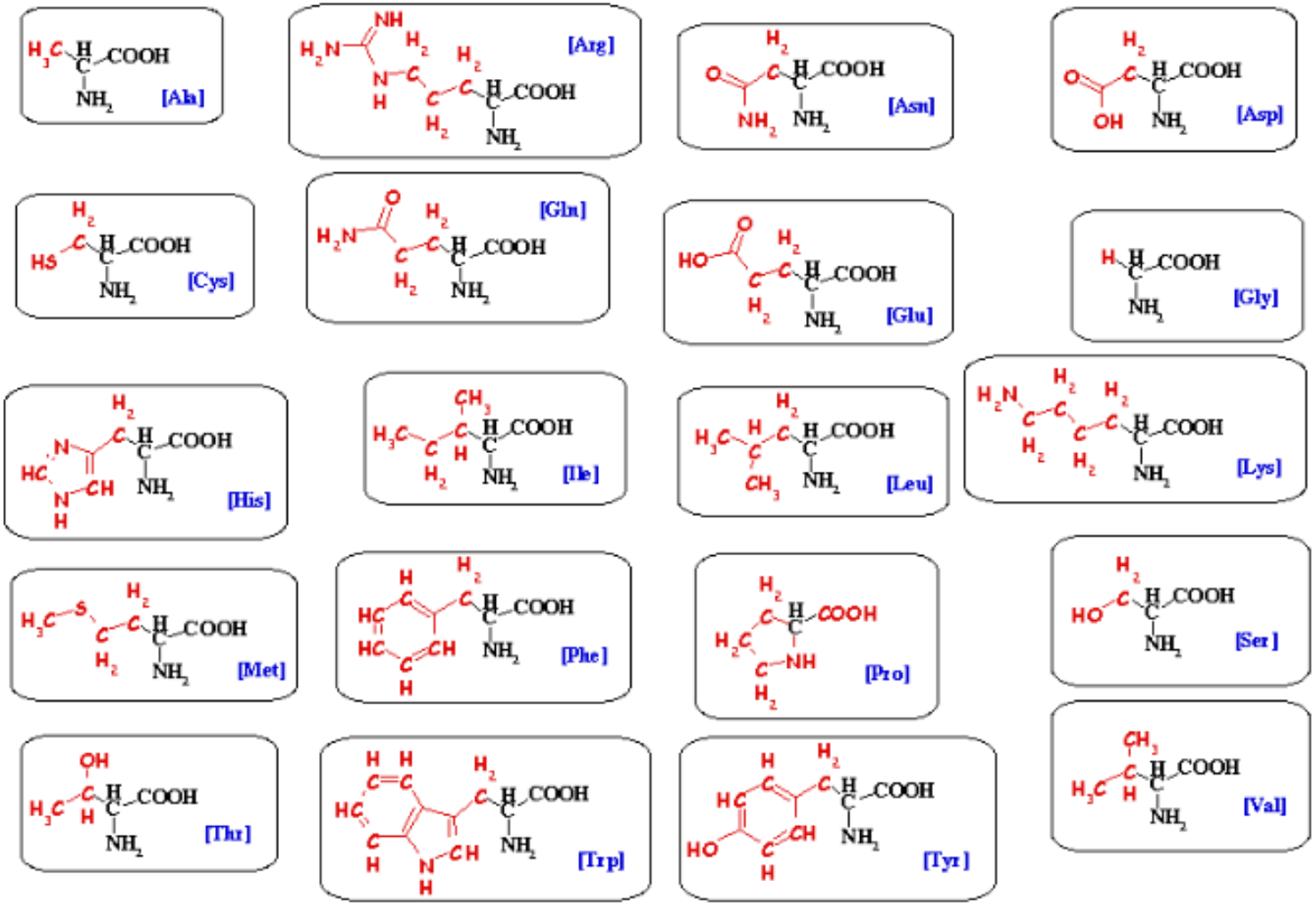


LES ACIDES AMINÉS SONT LIÉS ENTRE EUX !

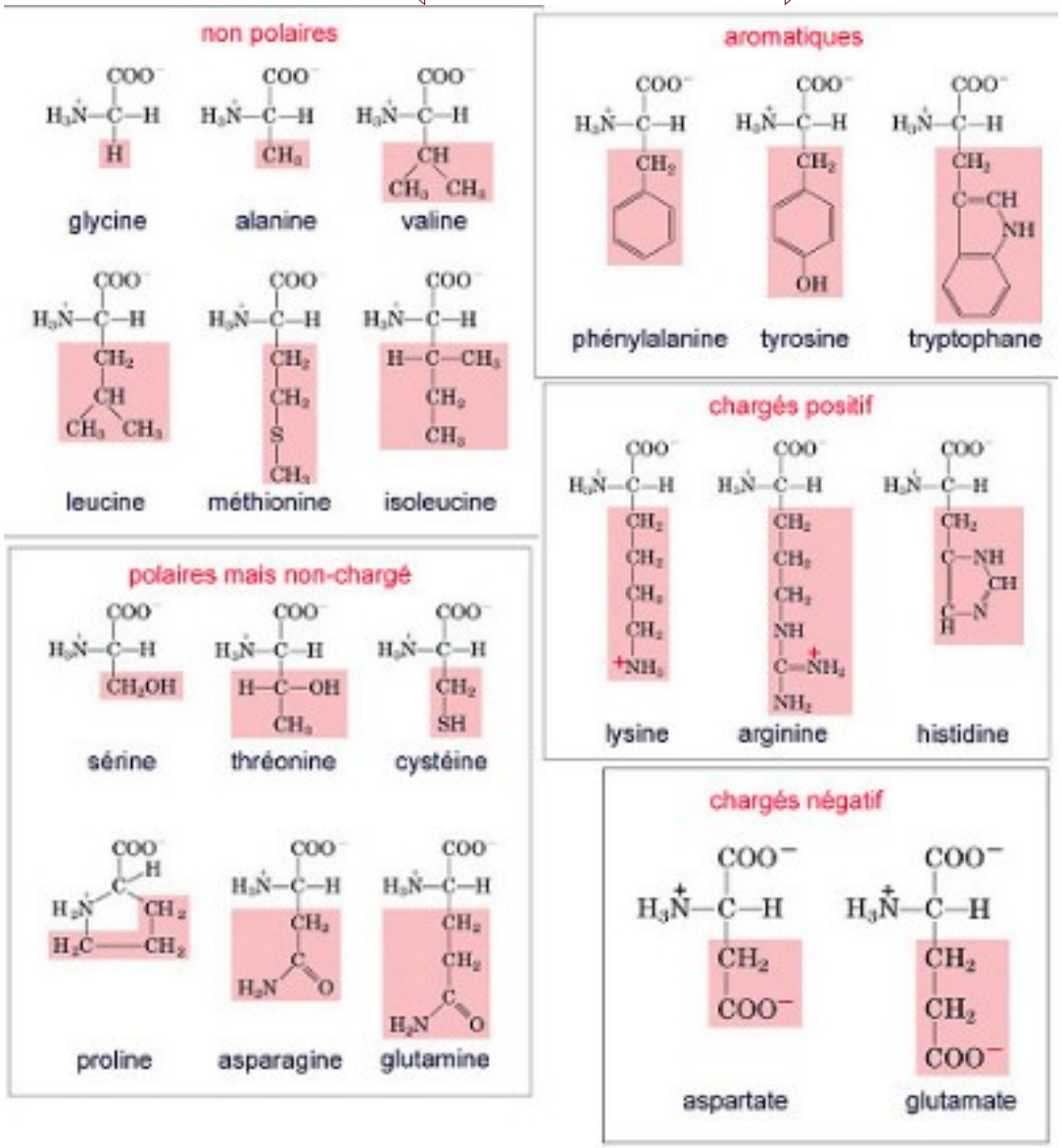


Une liaison (bond) peptide est formée par une réaction chimique qui extrait la molécule H2O quand elle joint un groupe amine d'un AA à un groupe carboxylique d'un autre AA.

STRUCTURE CHIMIQUE DES 20 ACIDES AMINÉS



CLASSIFICATION DES ACIDES AMINÉS SELON LA CHAÎNE LATÉRALE (SIDE CHAIN)



PROTÉINES
POLYPEPTIDE OU SÉQUENCE D'ACIDES AMINÉS

°Les protéines: les molécules les plus complexes et les plus variées des êtres vivants. On fabriquerait ~100 000 protéines différentes qui constituent près de 50% du poids sec.

°Une protéine, c'est un polymère d'acides aminés. La plupart des protéines sont formées de 100 à 200 AA.

°Toutes les protéines résultent de la combinaison de 20 acides aminés différents .

°Un acide aminé est une substance organique avec une fonction amine et une fonction carboxylique.

°Un peptide est formé d'un nombre restreint d'acides aminés.

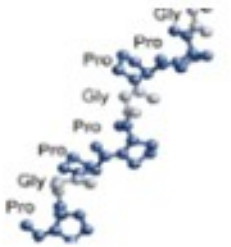
PROTÉINES : FONCTIONS

° Les protéines gouvernent tous les phénomènes biologiques, elles remplissent de nombreux rôles dans la cellule :

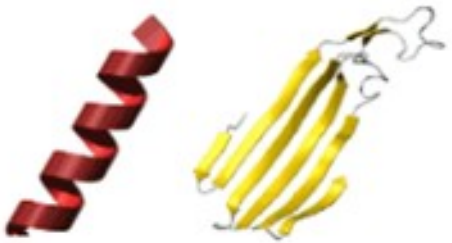
- Structure : les fibres protéiques , le cytosquelette
- Mouvement coordonné: le muscle, les spermatozoïdes
- Transport et mise en réserve : l'hémoglobine
- Transport de substances à travers la membrane cellulaire
- Constituent des messages : les Hormones , les neurotransmetteurs
- Catalyse des réactions chimiques : les enzymes
- Identité d'un organisme et sa Défense : les anticorps
- Régulation de la machinerie biosynthétique et métabolique : Les activateurs ou les répresseurs.
- les protéines peuvent être nuisibles : Les toxines et les protéines virales.

PROTÉINES : STRUCTURES

UNE PROTÉINE POSSÈDE 4 STRUCTURES DIMENSIONNELLES

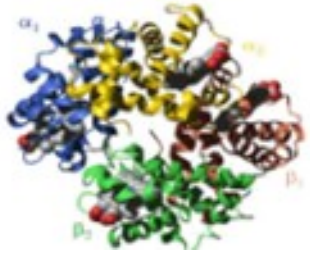


Structure primaire : assemblage des acides aminés



Structure secondaire : hélices α et feuillets β

Structure tertiaire : repliement de la protéine

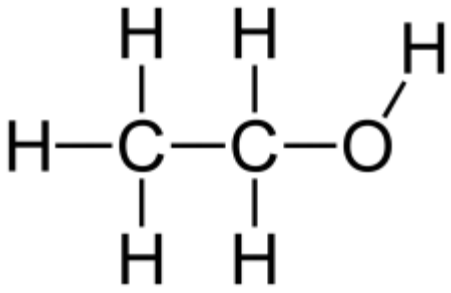


Structure quaternaire : assemblage de plusieurs structures tertiaires

STRUCTURE CHIMIQUE 2D VS. 3D

- Structure 2D: montre les liaisons de covalences (covalent bonds) entre les atomes. Essentiellement un graph.
- Structure 3D: montre les positions relatives des atomes.

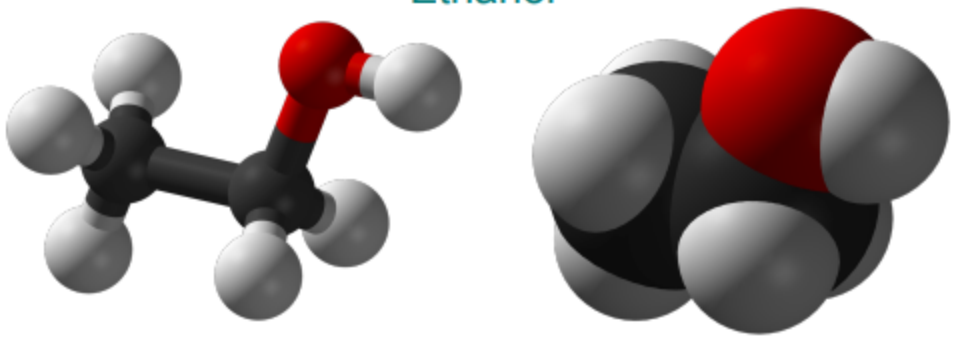
2D structure



Ethanol

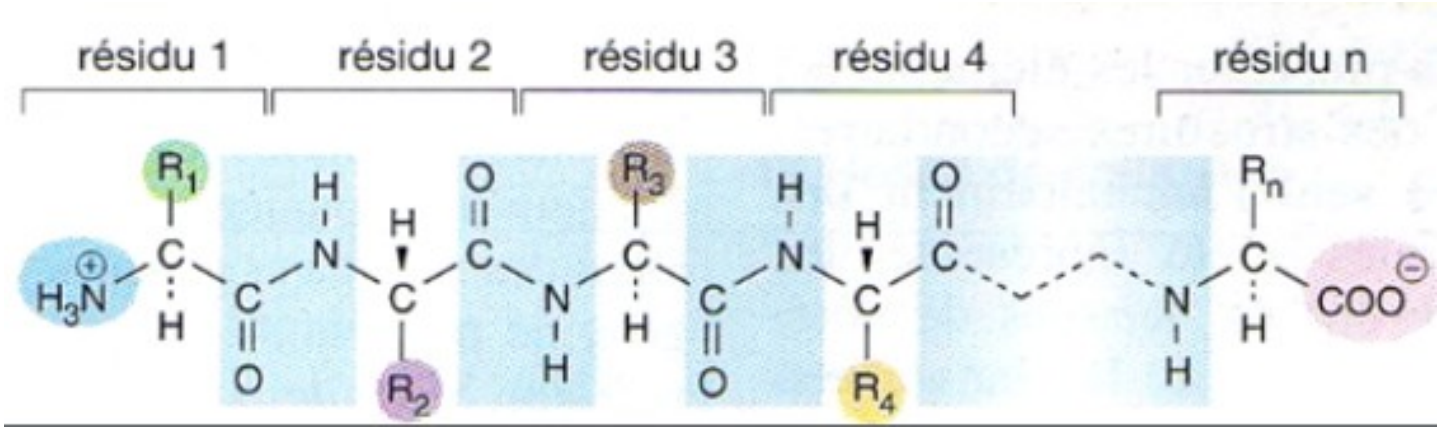
3D structure

Ethanol



PROTÉINE : STRUCTURE LINÉAIRE

- Structure primaire : séquence d'AA sous forme linéaire.
- La protéine ne s'est pas encore repliée.
- Il n'y a pas de liaisons à l'intérieur même de la chaîne.

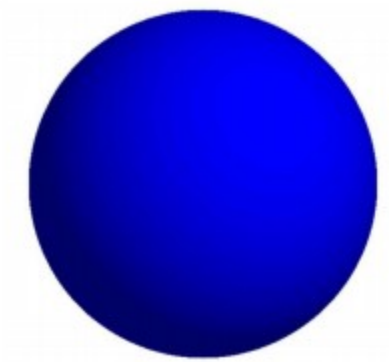


La structure primaire est une description complète de toutes les liaisons ***covalentes*** dans une chaîne polypeptidique ou de la protéine

Par contre, aucune indication n'est donnée quant à la ***position*** des acides aminés dans l'espace

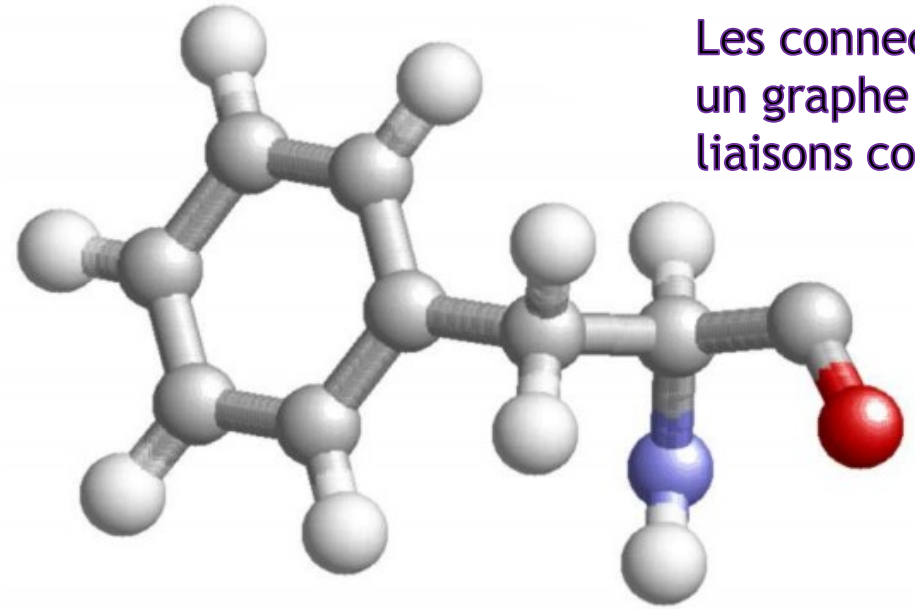
INTERACTIONS DE BASE

GÉOMÉTRIE D'UN ATOME



- Approximation intuitive et naïve: un atome est une sphère.
- Il occupe une position dans l'espace spécifiée par 3 coordonnées cartésiennes (x,y,z) de son centre à un moment donné.

GÉOMÉTRIE D'UNE MOLÉCULE



Les connections/lignes dans un graphe correspondent aux liaisons covalentes

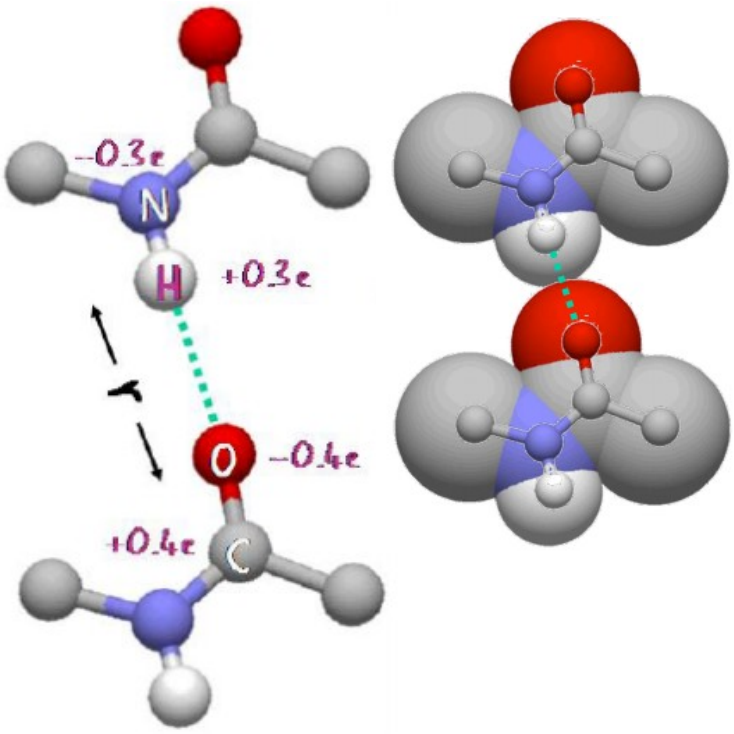
- Une molécule est un ensemble d'atomes connectés dans un graphe.
- Les coordonnées (x,y,z) de chaque atome désigne la géométrie de la molécule.

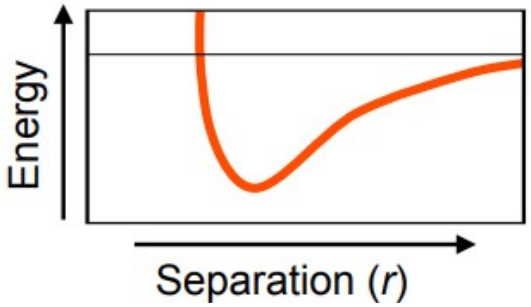
FORCES ENTRE ATOMES

- On peut approximer l'énergie potentielle totale d'un système moléculaire comme la somme des contributions individuelles. Les termes sont additifs.
  - La somme sur chaque atomes est aussi une somme des contributions individuelles.
  - Les forces quantiques sont ignorées. Ainsi, les atomes sont des balles et les forces sont des ressorts.
- On considère 2 types de force:
  - Forces de liaison: agissent entre des ensemble d'atomes étroitement liés dans un graphe de liaisons.
  - Forces non de liaison: agissent entre tous les paires d'atomes.

INTERACTIONS COMPLEXES

INTERACTION HYDROGÈNE

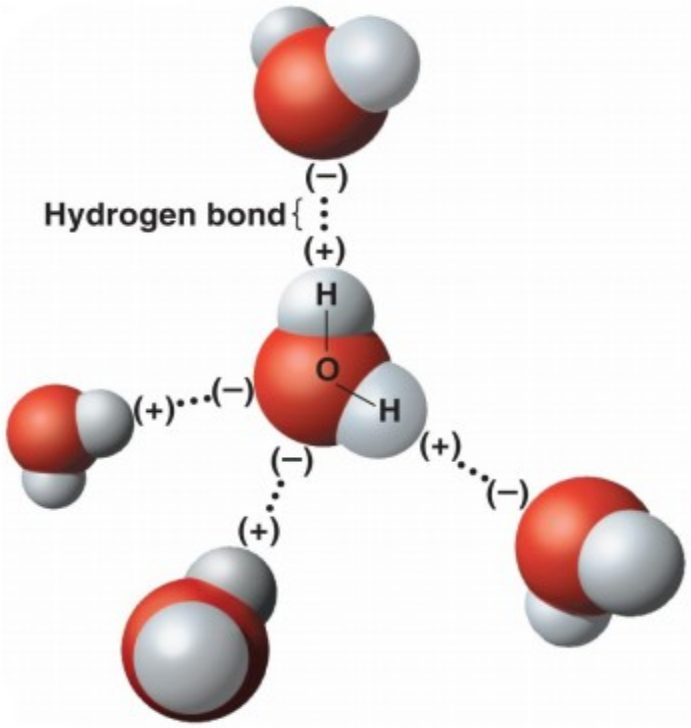




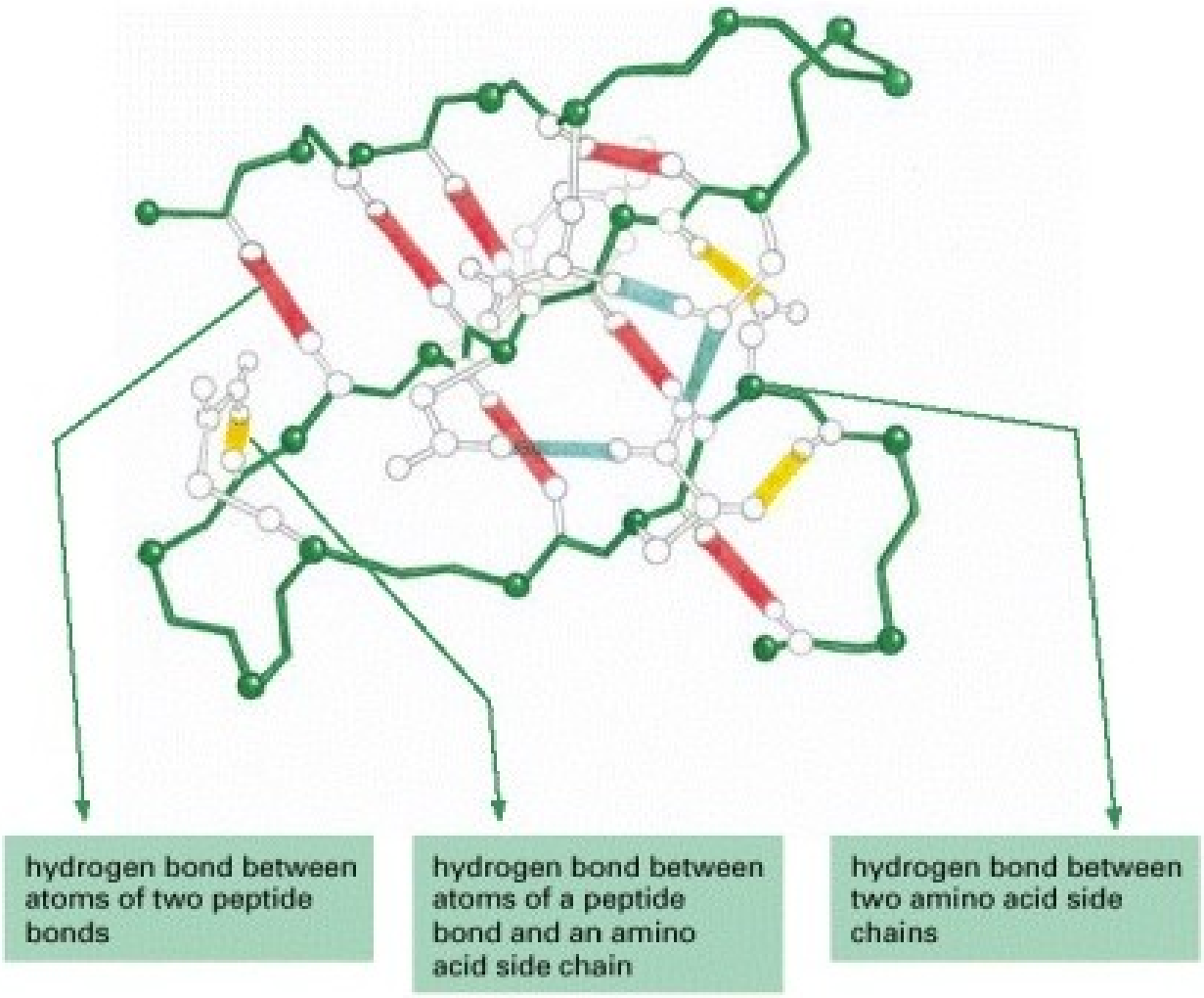
- Interaction favorable entre un atome électronégatif (ex. « N » ou « O ») et une liaison hydrogène à un autre atome électronégatif.
- résulte de multiples interactions électrostatiques et de van der Waals.
- Très sensible à la géométrie des atomes (distance et alignement).
- Relativement forte aux forces typiques électrostatique et de van der Waals.
- Critique à la structure de protéine et aussi à d'autres structures biomoléculaires.

MOLÉCULES D'EAU FORMENT DES LIAISONS HYDROGÈNE

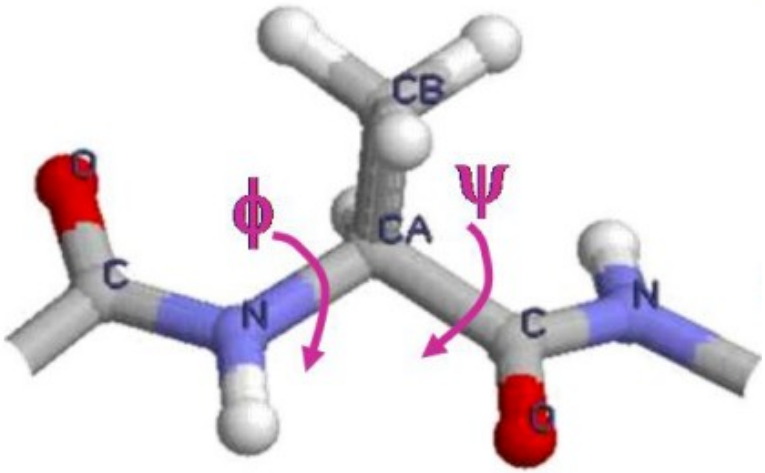
- Les molécules H2O forment des liaisons hydrogène tendues entre eux et entre les atomes d'un protéine.
- La structure d'une protéine dépends du fait qu'elle est entourée d'H2O.



EXEMPLE


hydrogen bond between atoms of two peptide bonds
hydrogen bond between atoms of a peptide bond and an amino acid side chain
hydrogen bond between two amino acid side chains

DEGRÉS LIBERTÉ BACKBONE

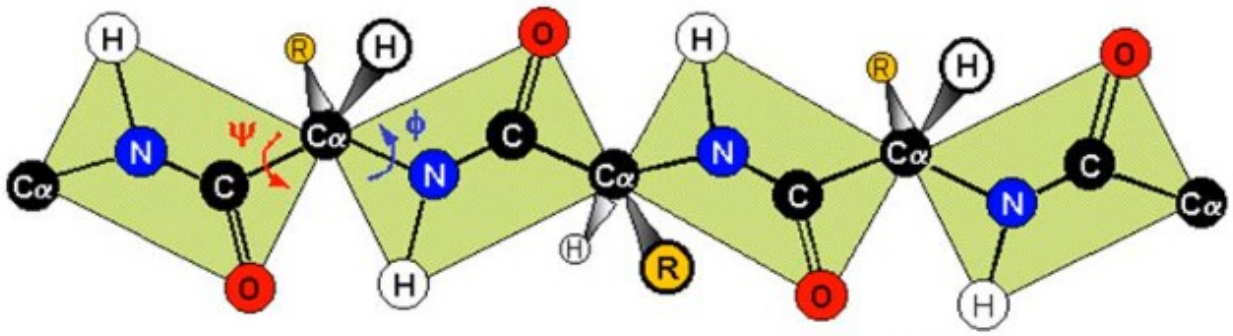


- L'angle de torsion rotant autour de la liaison N-CA et appelé ϕ
- L'angle de torsion rotant autour de la liaison CA-C et appelé ψ
- Ensemble, ils sont les angles (ϕ, ψ)
- La liaison N-C ou « liaison peptidique » est rigide.

CONFORMATION DE LA CHAÎNE PEPTIDIQUE

- L'élucidation des principales structures secondaires des protéines ne fut que possible suite à la compréhension de la structure de la ***liaison peptidique***;

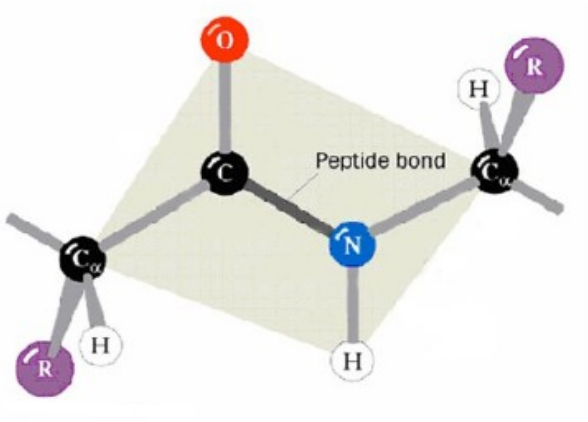
FULLY EXTENDED POLYPEPTIDE CHAIN



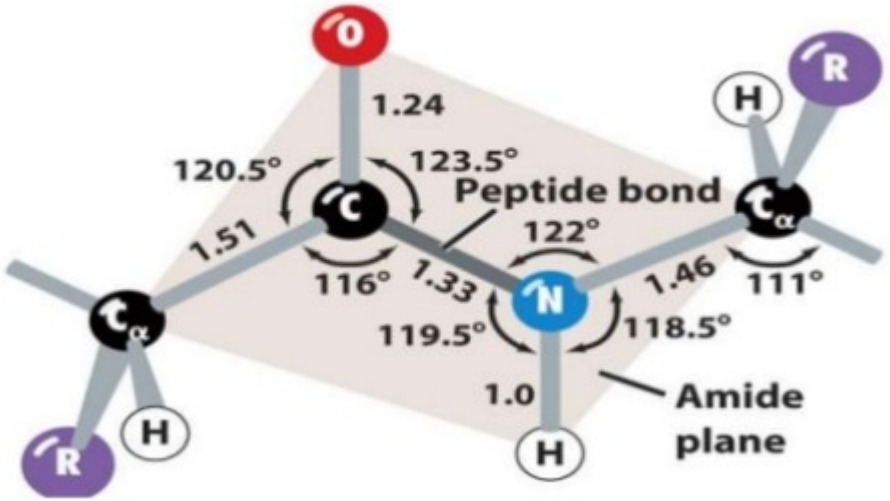
- La liaison peptidique est planaire!

LIAISON PEPTIDIQUE

Un groupe peptidique est formé de 6 atomes qui sont dans le même plan.

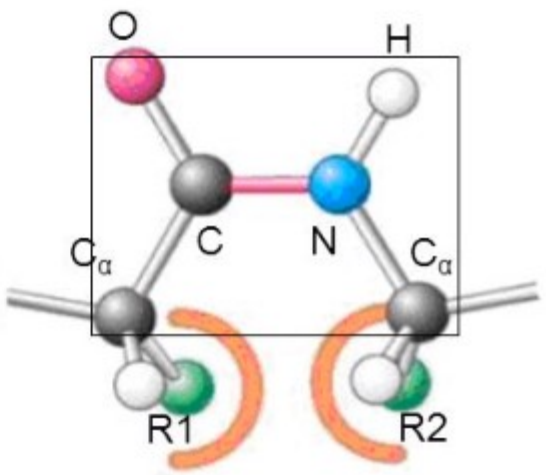


Groupe peptidique



CONFIGURATION CIS DU GROUPE PEPTIDIQUE

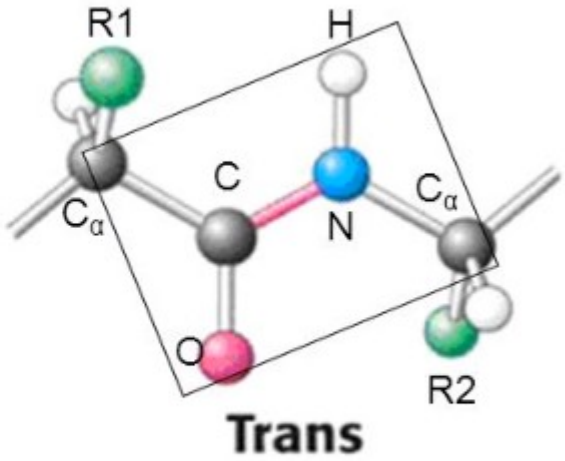
- Les carbones α sont situés sur le même flanc de la liaison peptidique et sont plus proches l'un de l'autre;
- L'interférence stérique entre les chaines latérales portées par ces deux carbones α défavorise la configuration *cis* par rapport a la configuration *trans*, étirée.



Cis

CONFIGURATION TRANS DU GROUPE PEPTIDIQUE

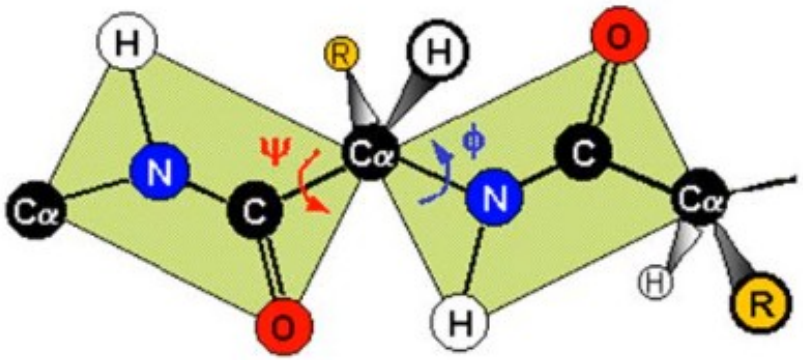
- Les carbones α de deux résidus d'acide aminé successifs encadrent la liaison peptidique et occupent les coins opposés du rectangle contenant le plan du groupe peptidique
- Ainsi, <u>presque</u> tous les groupes peptidiques des protéines sont dans la configuration *trans*



Trans

ANGLES PSI ET PHI
- 180° ET +180°-

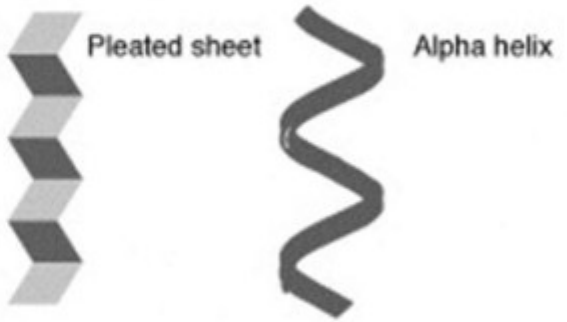
- La rotation est ***limitée*** autour du lien peptidique à cause du caractère double partiel de la liaison
- Cependant, la rotation est ***permise*** autour des liaisons entre le groupe amino et le carbone α ainsi qu'entre le carbone α et le groupe carbonyle puisqu'ils s'agit de liaisons ***simples***
- La rotation autour de la liaison N-Cα du groupe peptidique est désigné Φ (**phi**) et celle autour de la liaison Cα-C est désigné Ψ (**psi**)



PROTÉINE : STRUCTURE SECONDAIRE

- Réfère à la structure spatiale adoptée par des acides aminés adjacents sur une portion seulement de la protéine

- Les liaisons hydrogène jouent un rôle important pour stabiliser les conformations de structures secondaires

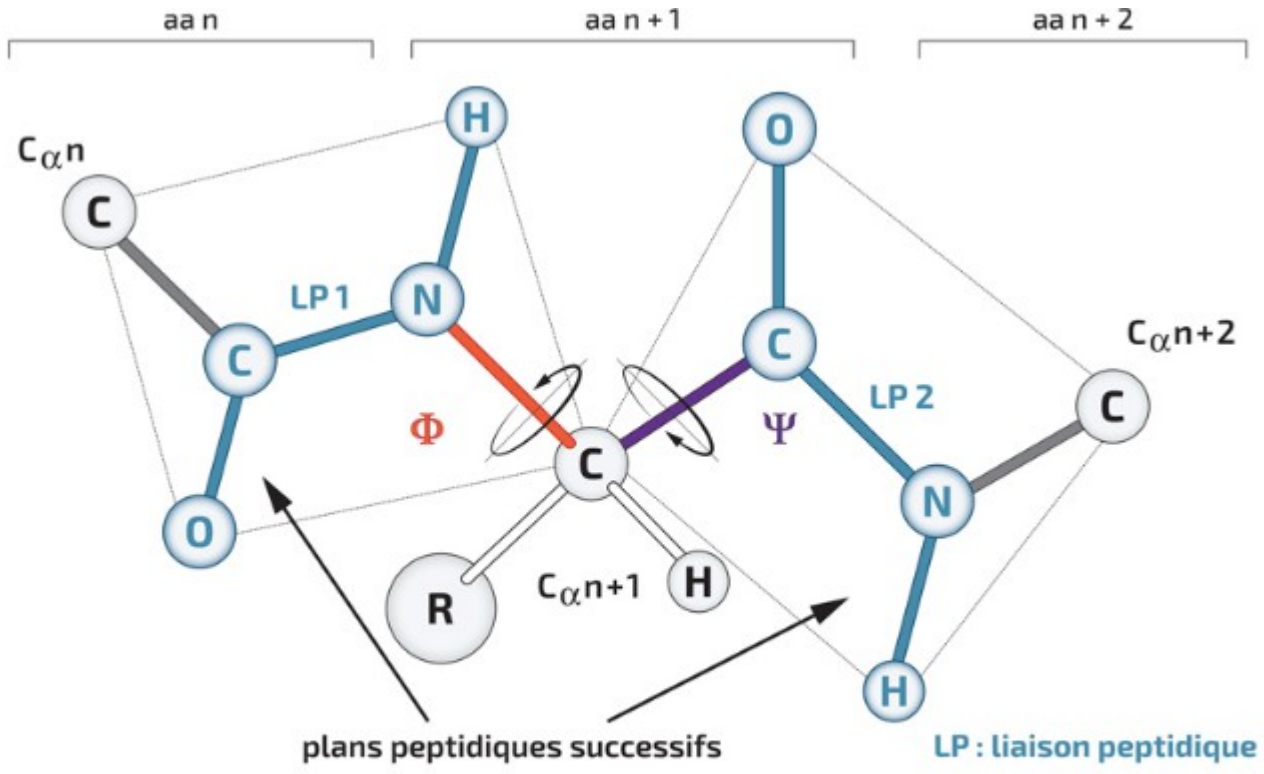


- Les deux types principaux de structure secondaire sont: l'hélice alpha et le feuillet bêta

Il existe des méthodes expérimentales pour déterminer la structure secondaire comme la magnétique nucléaire, le dichroïsme ou résonance ou certaines méthodes de spectroscopie infrarouge.

ANGLES PSI ET PHI

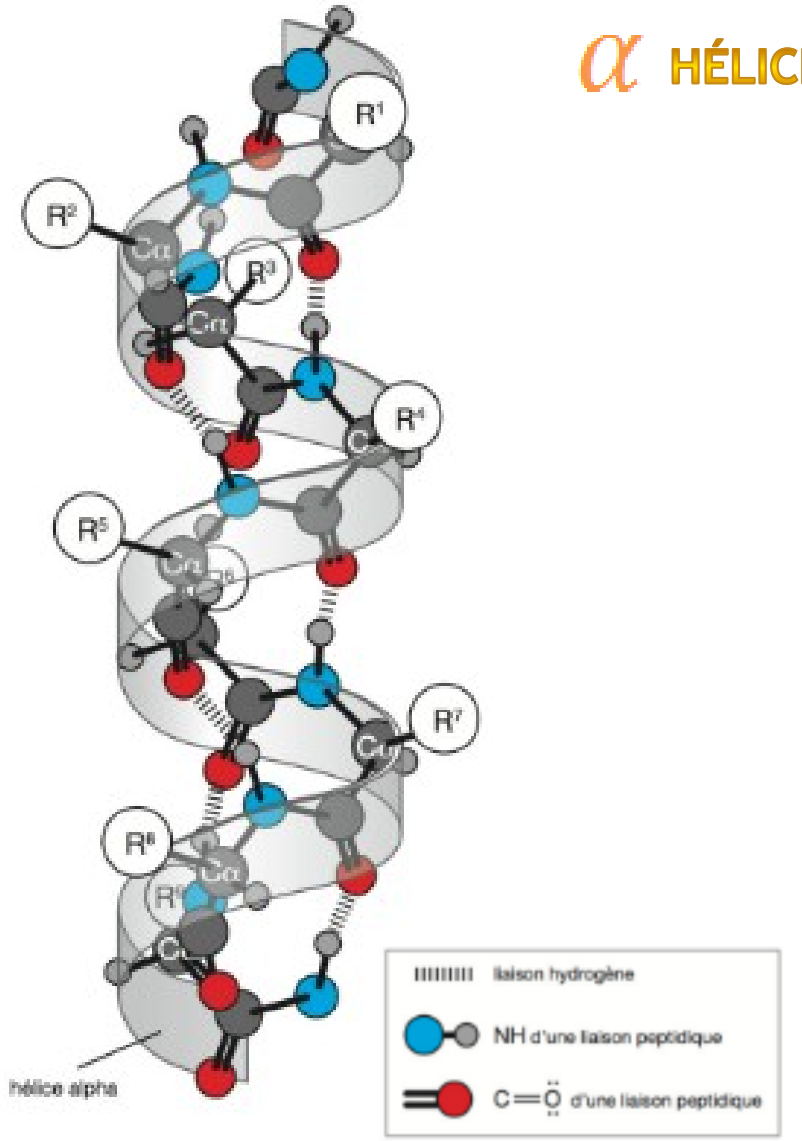
AUTRE EXEMPLE

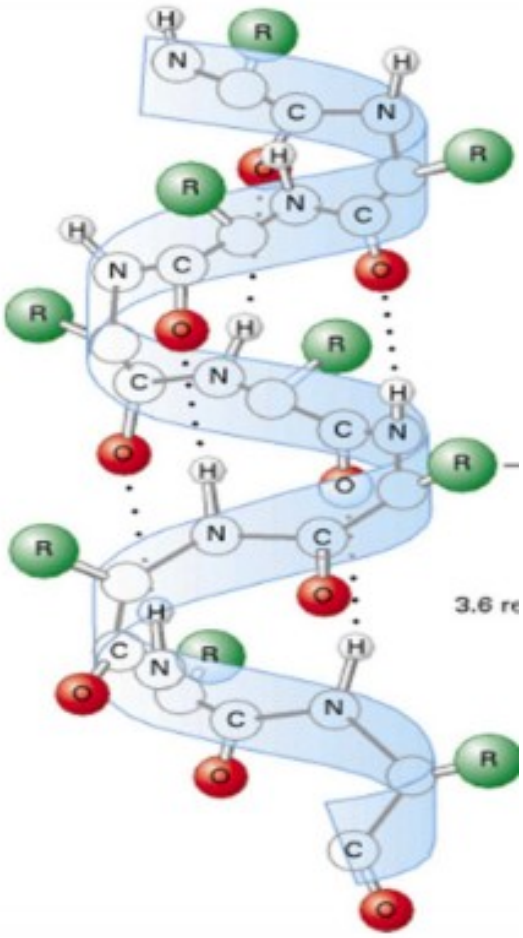


Chaque plan comprend six atomes. Les plans sont articulés entre eux autour des carbones alpha par libre rotation : angle phi (Φ, Cα-N) et psi (Ψ, Cα-C) du même aa.

PROTÉINE : STRUCTURE SECONDAIRE (1)

α HÉLICE





CARACTÉRISTIQUES DE L'HÉLICE

L'hélice α

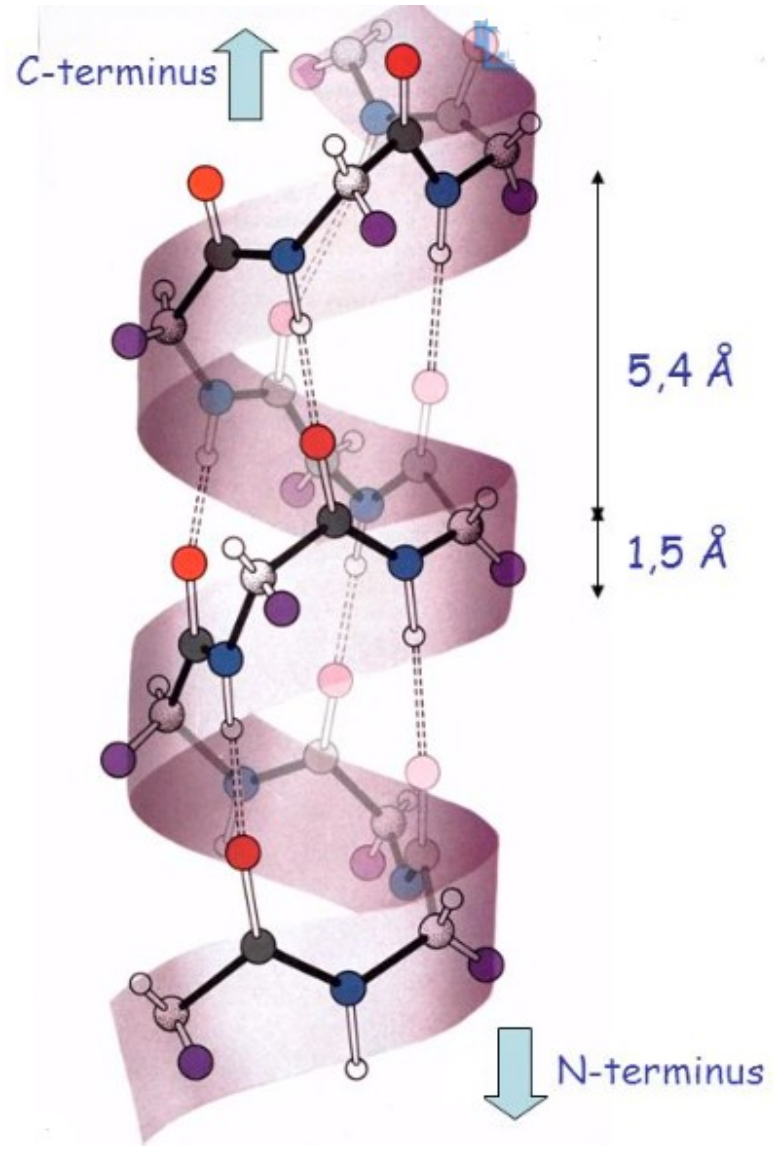
-hélice de pas à droit

-3,6 résidus par tour de spire

-pas = 5,4 Å

-incrément = p/n = 1,5 Å

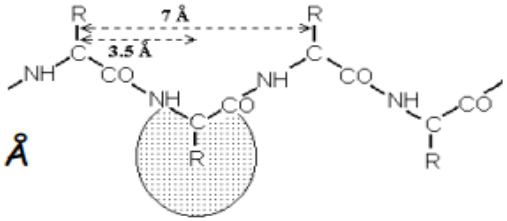
- Phi (Φ) = -57° et Psi (Ψ) = - 47°

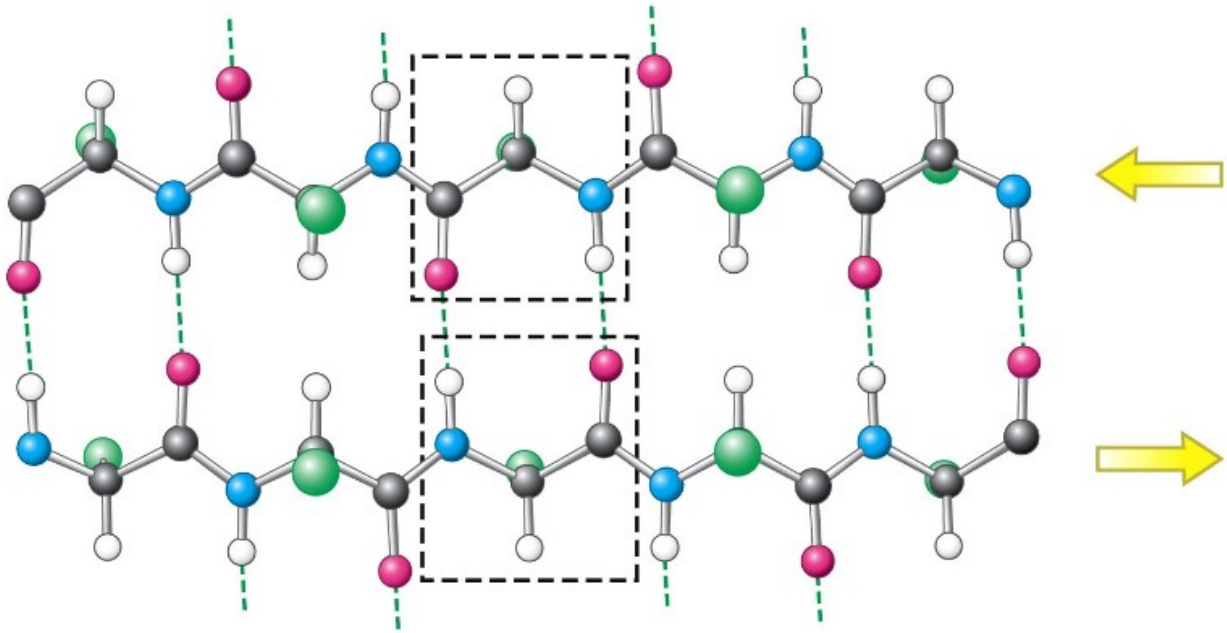


FEUILLET BETA (1930)

Brins

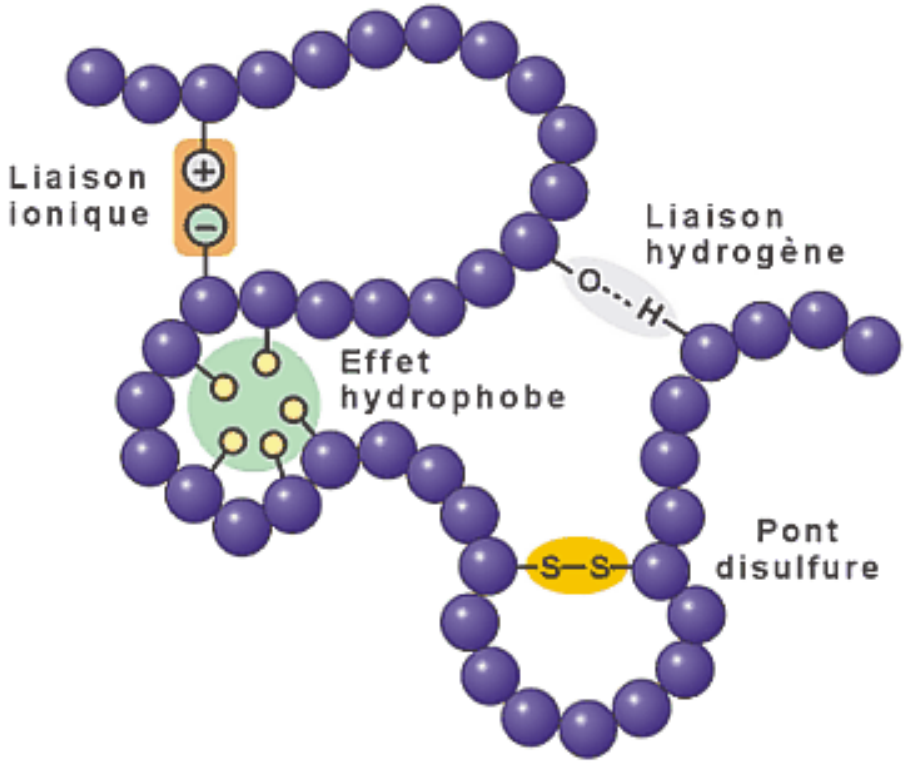
- Conformation étendue (1930)
- Liaison H inter brins (Pauling)
- Longueur moyenne = 5 AA
- 25-30% des résidus
- Rise = 3.5 Å ; Pitch = 7 Å





PROTÉINE : STRUCTURE TERTIAIRE (1)

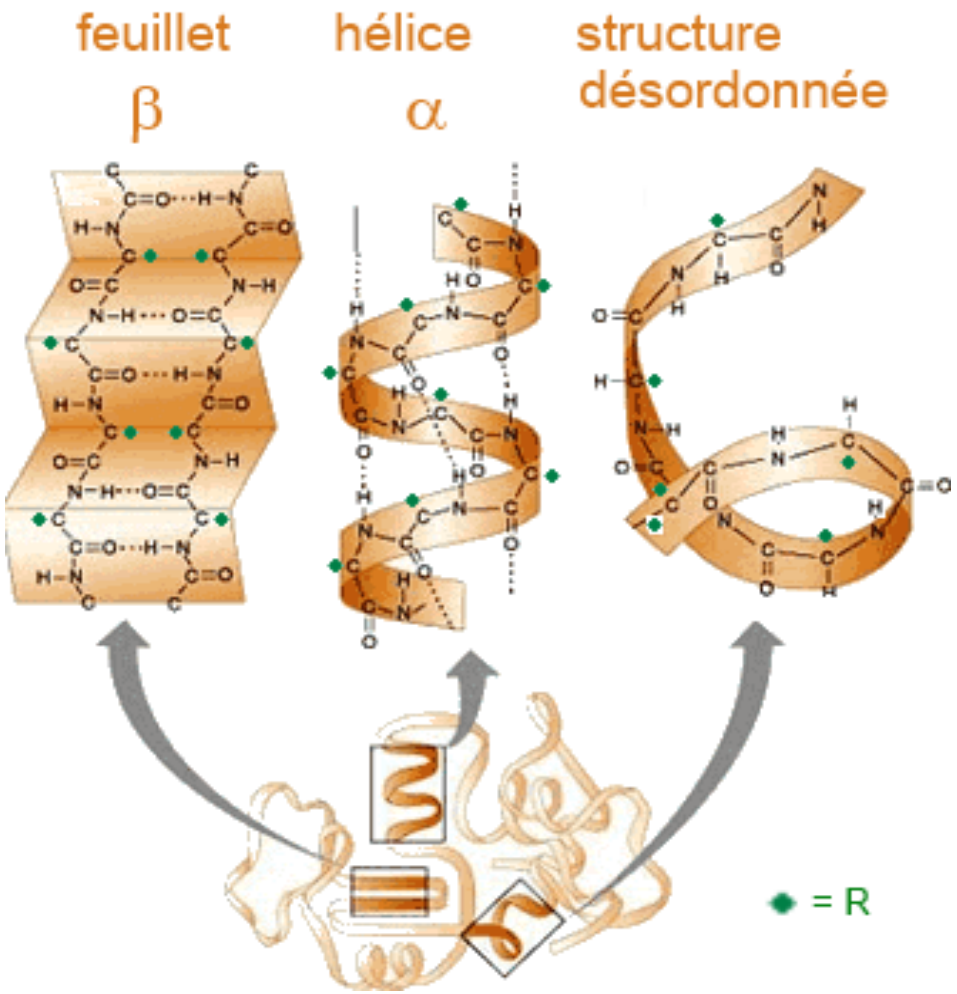
DIFFÉRENTES LIAISONS POSSIBLES ENTRE ACIDES AMINÉS – SIDE CHAIN



La connaissance de la structure tertiaire, voire quaternaire, d'une protéine peut fournir des éléments importants pour comprendre comment cette protéine remplit sa fonction biologique. La cristallographie aux rayons X et la spectroscopie RMR sont des méthodes expérimentales pour étudier la structure 3D à l'échelle atomique. Elles sont coûteuses en temps et en coût.

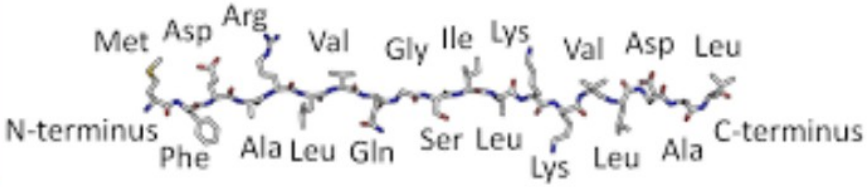
PROTÉINE : STRUCTURE TERTIAIRE (2)

DIFFÉRENTES LIAISONS POSSIBLES ENTRE ACIDES AMINÉS ENGENDRENT DIFFÉRENTES FORMES DE LA STRUCTURE TERTIAIRE



NIVEAUX DE STRUCTURATION D'UNE PROTÉINE

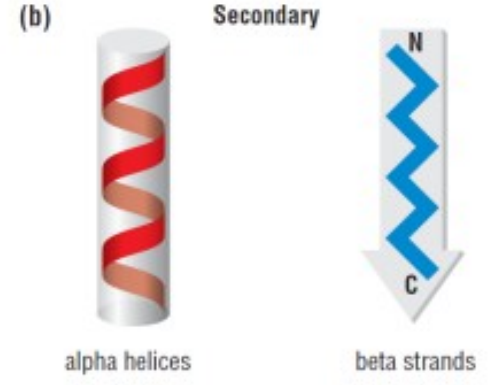
Primary

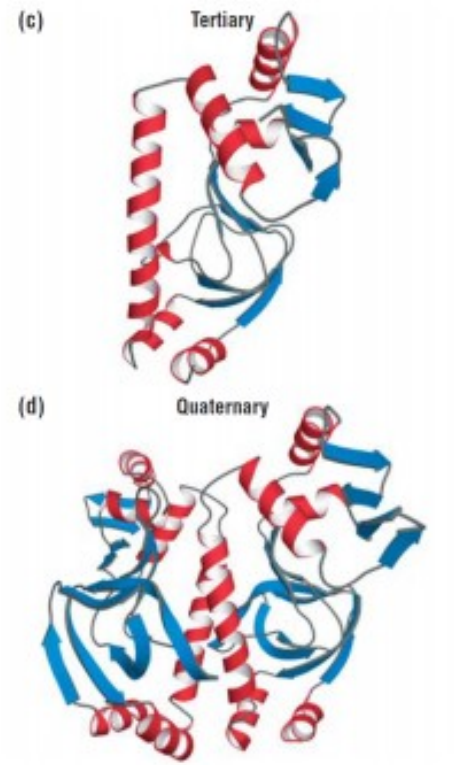


Secondary

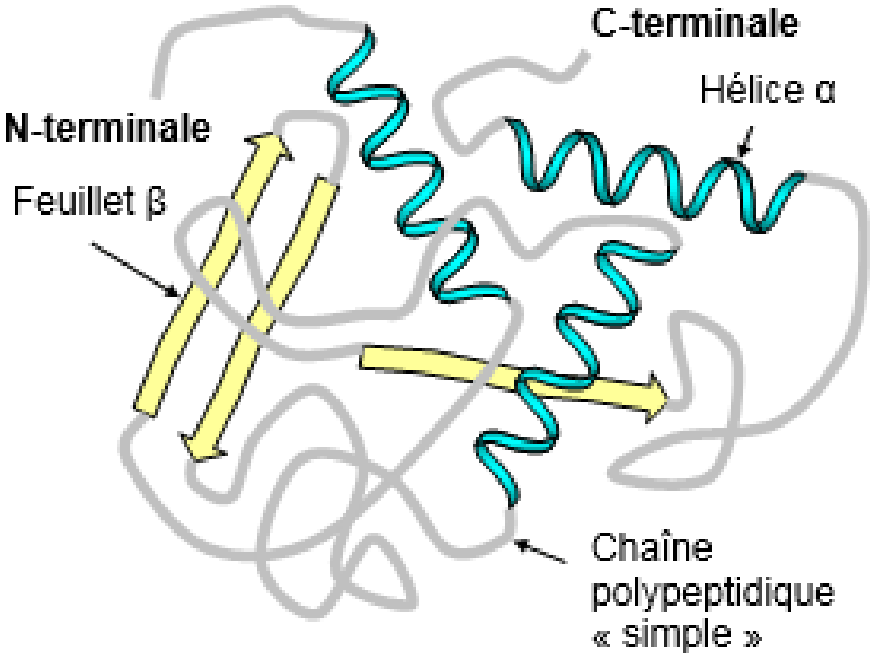
β-Sheet (3 strands)

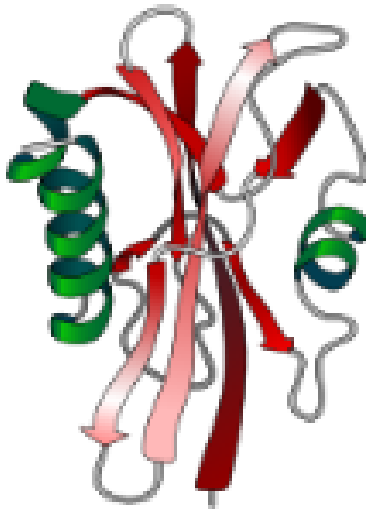
α-helix





EXEMPLE DE STRUCTURES TERTIAIRES



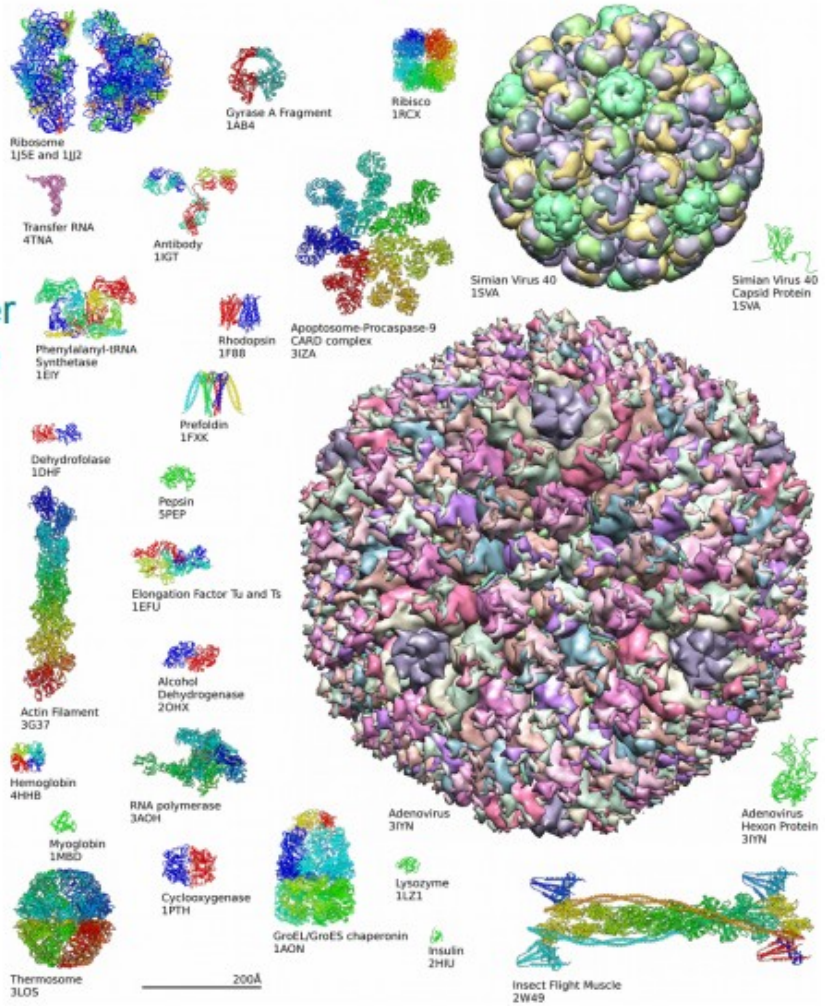


BANQUE DE DONNÉES DES PROTÉINES PDB

The Protein Data Bank (PDB)

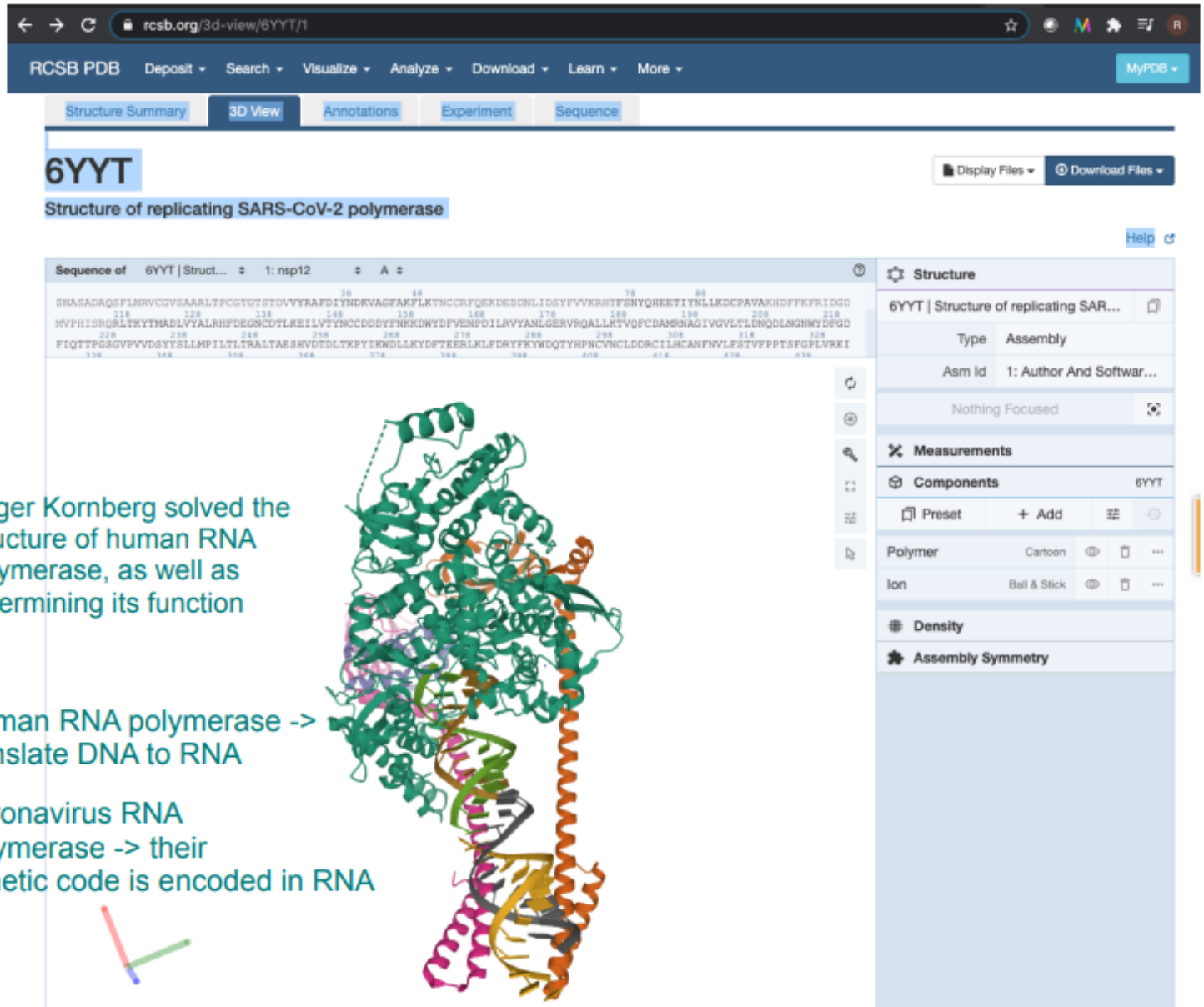
- Examples of structures from PDB.

This databank contains molecules other than proteins. It was named PDB since proteins were the first structures deposited in the databank



https://upload.wikimedia.org/wikipedia/commons/thumb/2/24/Protein_structure_examples.png/1024px-Protein_structure_examples.png
(Axel Griewel)

You're not responsible for these; they're just examples.

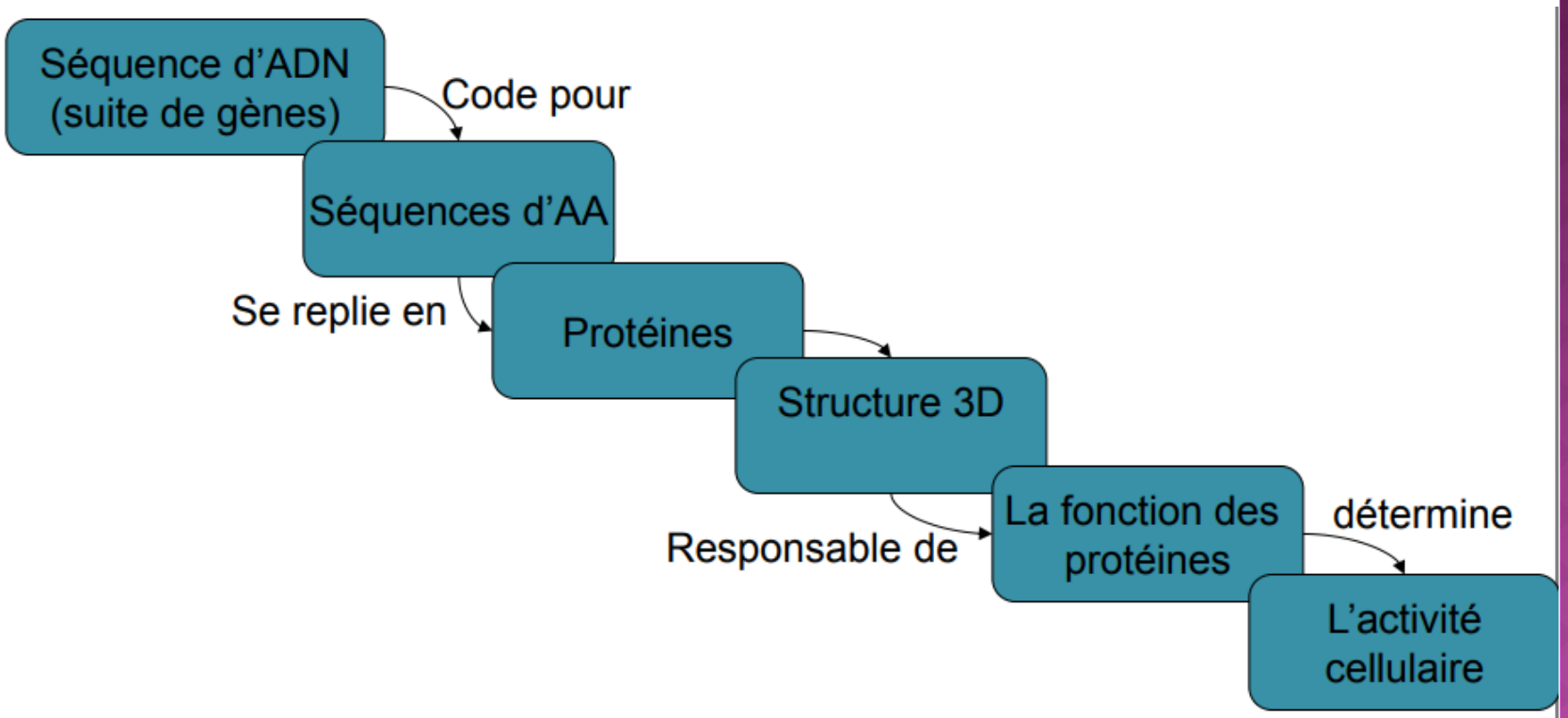


Roger Kornberg solved the structure of human RNA polymerase, as well as determining its function

Human RNA polymerase -> translate DNA to RNA

Coronavirus RNA polymerase -> their genetic code is encoded in RNA

DE L'ADN À LA FONCTION CELLULAIRE



POURQUOI COMPARER LES SÉQUENCES ?

- l'évolution se fait par mutations successives
- homologie $\Rightarrow$ similarité
- même séquence $\Rightarrow$ même fonction ?

COMPARAISON BASÉE ALIGNEMENT

objectif: Extraire les régions homologues Entre séqeunces (ADN, ARN, Protéine)

- entrée : k séquences

```
* * * * * * * * * * * * * *
* * * * * * * * * *
* * * * * * * * * * * * * *
* * * * * * * * * *
```

- sortie : un tableau contenant les k séquences, avec des indels

```
* * * * * * * * * * _ * * * *
* * * _ _ _ * * * _ * * * *
* * * _ * * * * * * * * * *
* * * _ _ * * _ _ * * * * *
```

ALIGNEMENT MULTIPLE VS. ALIGNEMENT 2 À 2

Alignement 2 à 2

Deux séquences quelconques

↓

Détecter une similarité **syntaxique**

↓

Il y a-t-il une **fonction** commune ?

❖ Séquences alignées remplissent même fonction. Exemple, une enzyme chez différentes espèces, on suppose un ancêtre commun.

❖ La région de concordance explique une pression de sélection pour maintenir la fonction de la macromolécule. La divergence explique les mutations.

Alignement multiple

Famille de séquences avec la même **fonction**

↓

À quelle conservation **syntaxique** cela correspond-il ?

EXEMPLE: L'INSULINE

```
éléphant   FVNQHLCGSHLVEALYLVCGERGFFYTPKTGIVEQCCTGVCSLYQLENYCN
           |||||||||||||||||||||||||| ||| ||||  ||||||||||
hamster    FVNQHLCGSHLVEALYLVCGERGFFYTPKSGIVDQCCTSICSLYQLENYCN
```

```
éléphant   FVNQHLCGSHLVEALYLVCGERGFFYTPKTGIVEQCCTGVCSLYQLENYCN
           |||||||||||||||||||||||||| |||||||   ||||||||||
baleine    FVNQHLCGSHLVEALYLVCGERGFFYTPKAGIVEQCCASTCSLYQLENYCN
```

```
éléphant   FVNQHLCGSHLVEALYLVCGERGFFYTPKTGIVEQCCTGVCSLYQLENYCN
             || ||||||| |||||||||||| || |||||||   ||||||||||
alligator  AANQRLCGSHLVDALYLVCGERGFFYSPKGGIVEQCCHNTCSLYQLENYCN
```

OBJECTIFS VISÉS DE L'ALIGNEMENT

- En général, l'alignement vise la prédiction.
- Il permet d'identifier les sites fonctionnels (site catalytique, site d'interaction..).
- Prédire la fonction ou les fonctions d'une protéine (existence d'homologie avec une fonction connue).
- Prédiction ?? d'une structure d'un protéine.
- Alignement multiple au sein d'une famille de protéine permet d'établir une phylogénie.

PROCÉDURE ALIGNEMENT

- Un algorithme d'alignement cherche des motifs similaires en maximisant le nombre de coïncidences entre des N ou AA.
- Ainsi pour aligner les caractères communs des «trous» doivent être ajoutés à certaines positions. Un trou appelé « indel» correspond à une insertion ou délétion dans les séquences.
- Le problème revient donc à :

TROUVER LE MEILLEUR ALIGNEMENT, I.E. LES MEILLEURES POSITIONS POUR MAXIMISER LE NOMBRE D'IDENTITÉS ENTRE SÉQUENCES.

MEILLEUR ALIGNEMENT ?

- **Hypothèse**

INDEl = « Hypothèse d'un événement évolutif moins probable qu'une simple substitution». Donc son apparition doit être pénalisé plus qu'un substitution.

- **Exemple: 2 séquences** ➡ **Multiple Possibilités!!!**

1.
```
R D I S L V - - - K N A G I
|   |   | |       | |   | |
R N I - L V S D A K N V G I
```

2.
```
R D I - - S L V K N A - - - G I
|   |     |         |         | |
R N I L V S - - - D A K N V G I
```

3.
```
R D I - - S L V K N A G I
|   |     |     | |   | |
R N I L V S D A K N V G I
```

- **Bon/Mauvais alignement?** ➡ **Score**

$$fonction\,score = \sum (substitutions, identités,\ indel)$$

indel = pénalité qui prend la plus grande valeur.

Exemple: substitution = -1; indel = -3; identité = 2.

TYPES ALIGNEMENTS

Alignement Global.

Comment: Maximiser les identités sur la totalité des séquences.

Pourquoi : Alignement des protéines homologues en vue d'identifier les acides aminés conservés par l'évolution.

```
G G C T G A C C A C C - T T
|     |   | |     | |   |
G A - T C A C T T C C A T G
```

ALIGNEMENT LOCAL.

Comment: Maximiser les identités sur la totalité des séquences.

Pourquoi : Le plus long chevauchement entre 2 séquences est recherché pour la reconstitution à partir des données de séquençage.

```
G G C T G A C C A C C T T
      | |   | | |   | |
      G A T C A C - T T C C A T G
```

TYPES ALIGNEMENTS : ALGORITHMES UTILISÉS

Alignement Global.

- Algorithme de distance
- Needleman-Wunsch (1970)
- Myers & Miller

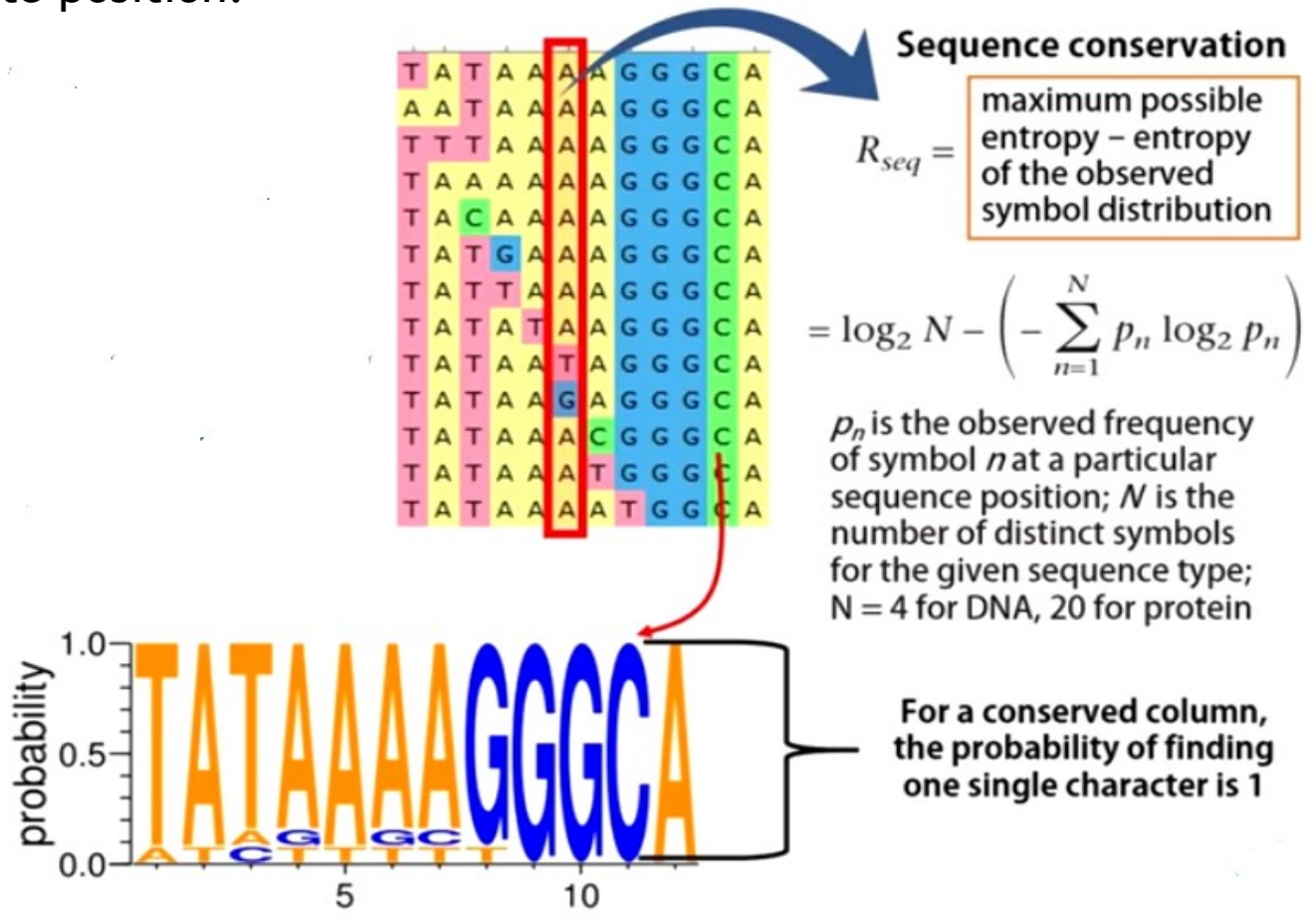
ALIGNEMENT LOCAL.

- Smith-Waterman (1981)
- FASTA (1988)
- BLAST (1990, version 2 en 1997)

Ces algorithmes utilisent la programmation dynamique

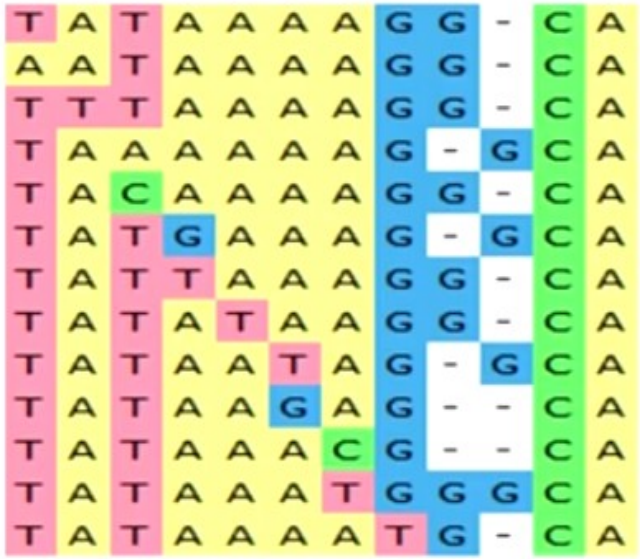
REPRÉSENTATION LOGO D'1 ALIGNEMENT MULTIPLE

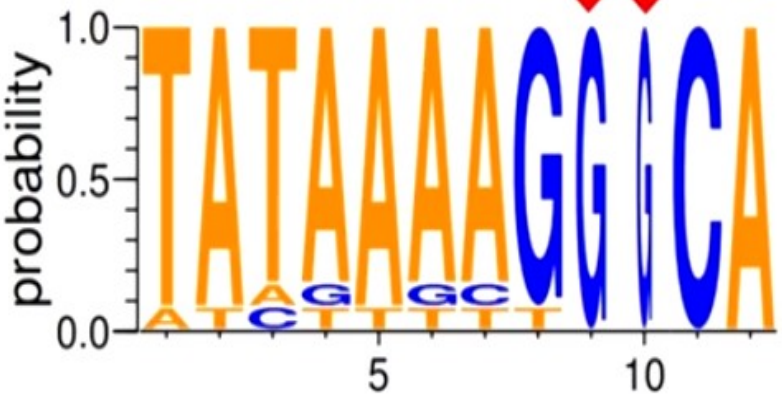
A partir d'un alignement multiple, on peut générer une séquence virtuelle qui contient pour chaque position l'AA le plus représenté.
La **hauteur d'un colonne** indique le **degré de conservation** de la position et la **hauteur d'un lettre** indique la **fréquence de la lettre** à cette position.



REPRÉSENTATION LOGO D'1 ALIGNEMENT MULTIPLE - SUITE-

What happens when the alignment column includes gaps?





Narrow character width represents the abundance of gaps in the alignment column

Higher character height represents its extent of conservation in the alignment column, disregarding gaps

LOGICIEL BLAST

Basic Local Alignment Search Tool

Altschul et al. - 1997

- programme pour la recherche de similarités dans les bases de données
- utilise un algorithme très rapide pour construire des alignements locaux approchés
- séquences nucléiques et protéiques
- connecté aux principales banques de données