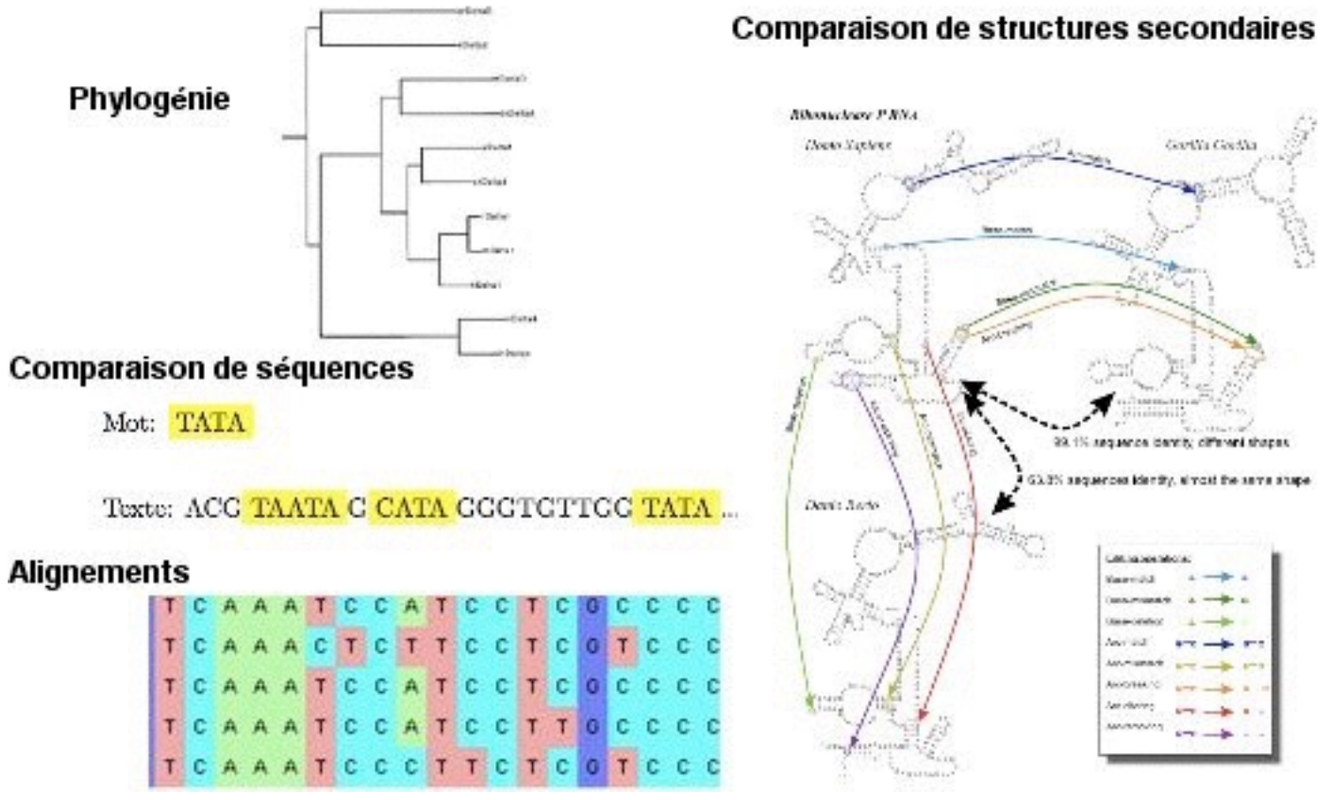
IFT6291 - BIN6000
Automne 2016

Algorithmes en bio-informatique génomique



Sylvie Hamel

André-Aisenstadt: 3161

hamelsyl@iro.umontreal.ca

http://www.iro.umontreal.ca/~hamelsyl/IFT6291-A16

Qu'est-ce que la bio-informatique??

Domaine interdisciplinaire, situé au carrefour de l'informatique, des mathématiques et de la biologie, qui traite de l'application de l'informatique aux sciences biologiques.

Synonyme: informatique biologique, biocalcul

Terme à éviter: biologie computationnelle
(Source: Office québécois de la langue française)

On regroupe sous le terme de bio-informatique toutes les applications informatiques appliquées à la biologie. Cela va de l'analyse du génome à la modélisation de l'évolution d'une population animale dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'image, le séquençage du génome, la reconstruction d'arbres phylogénétiques (phylogénie), etc.

(Source: Wikipédia)

Champs multi-disciplinaire utilisant des méthodes informatiques, algorithmiques, statistiques, mathématiques… pour:

- Formaliser des problèmes de biologie moléculaire
- Développer des algorithmes et les implémenter pour permettre:
  - L'analyse des données biologiques
  - La prédiction de résultats biologiques

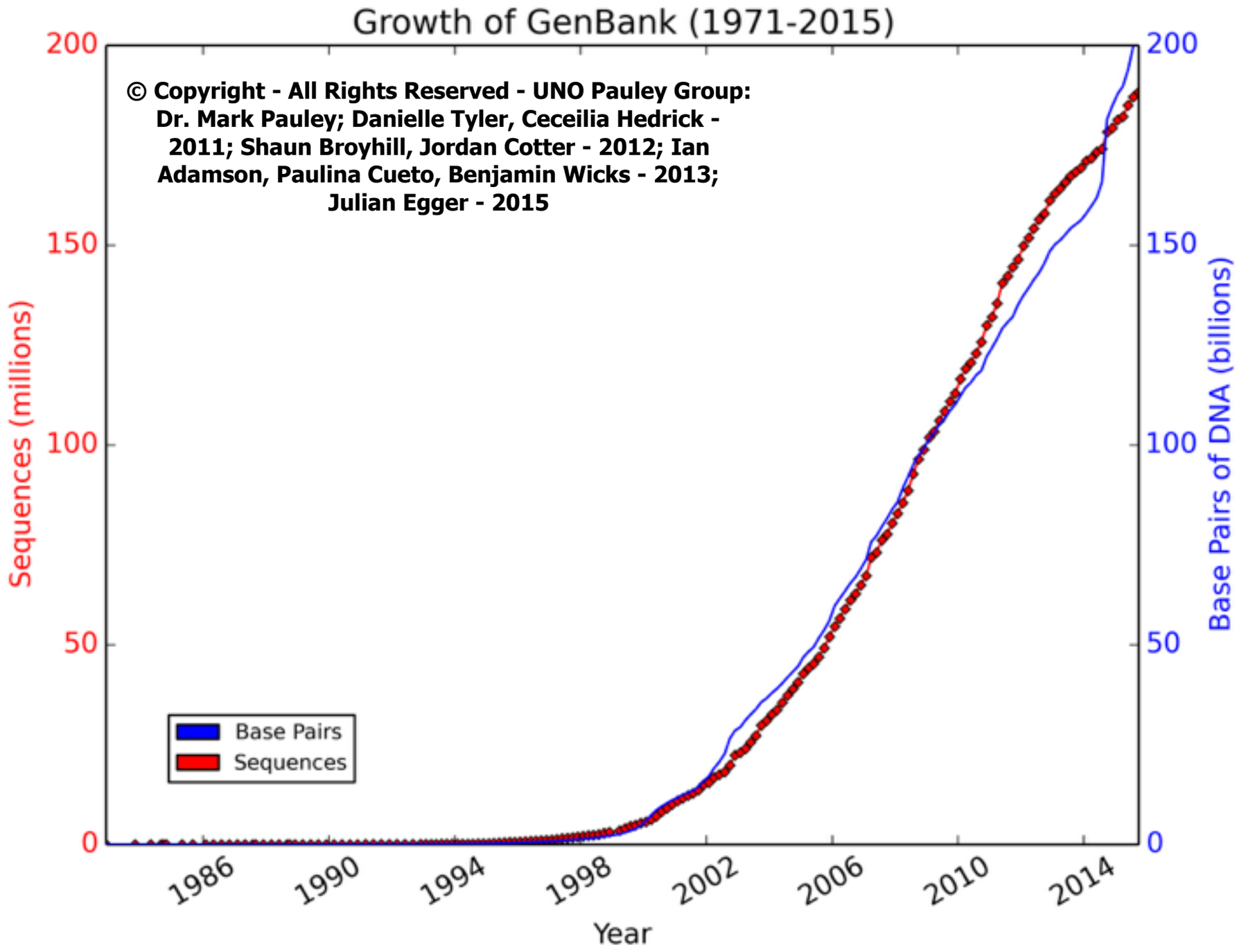
S'applique à tous les type de données biologiques:

- Séquences d'ADN et de protéines
- Structures secondaires et tertiaires d'ARN et de protéines
- Contenu en gènes des génomes
- Arbres de phylogénie
- Réseaux d'interections entre protéines

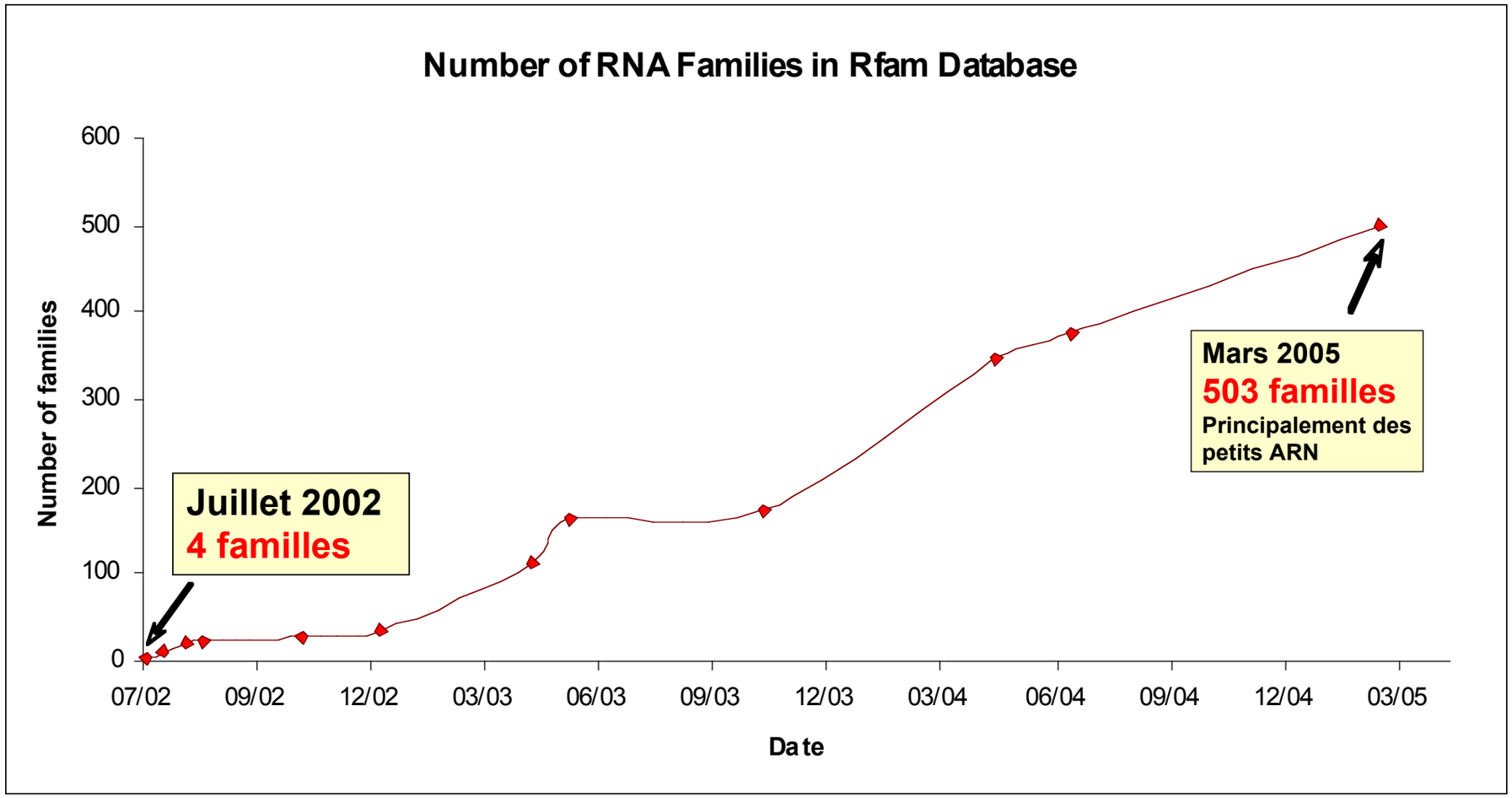
Défis de la bio-informatique:

Analyser, comprendre et organiser une masse de données biologiques:

- Séquences:
  - Plus de 3400 génomes eucaryotes complètement séquencés (ou presque) et publiés dont l'homme et la souris.
  - Plus de 73 000 génomes procaryotes complètement séquencés.
  - Plus de 5 500 génomes de virus complètement séquencés .
- Structures:
  - De plus en plus de familles d'ARN non condant sont découvertes
  - De plus en plus d'information structurales sur ces données



Genetic Sequence Data Bank
August 15 2016
196120831 loci, 217971437647 bases, from 196120831 reported sequences



Source:
Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna and Sean R. Eddy., *Rfam: an RNA family database.* Nucleic Acids Research, 2003, 31, 1, 439-441.

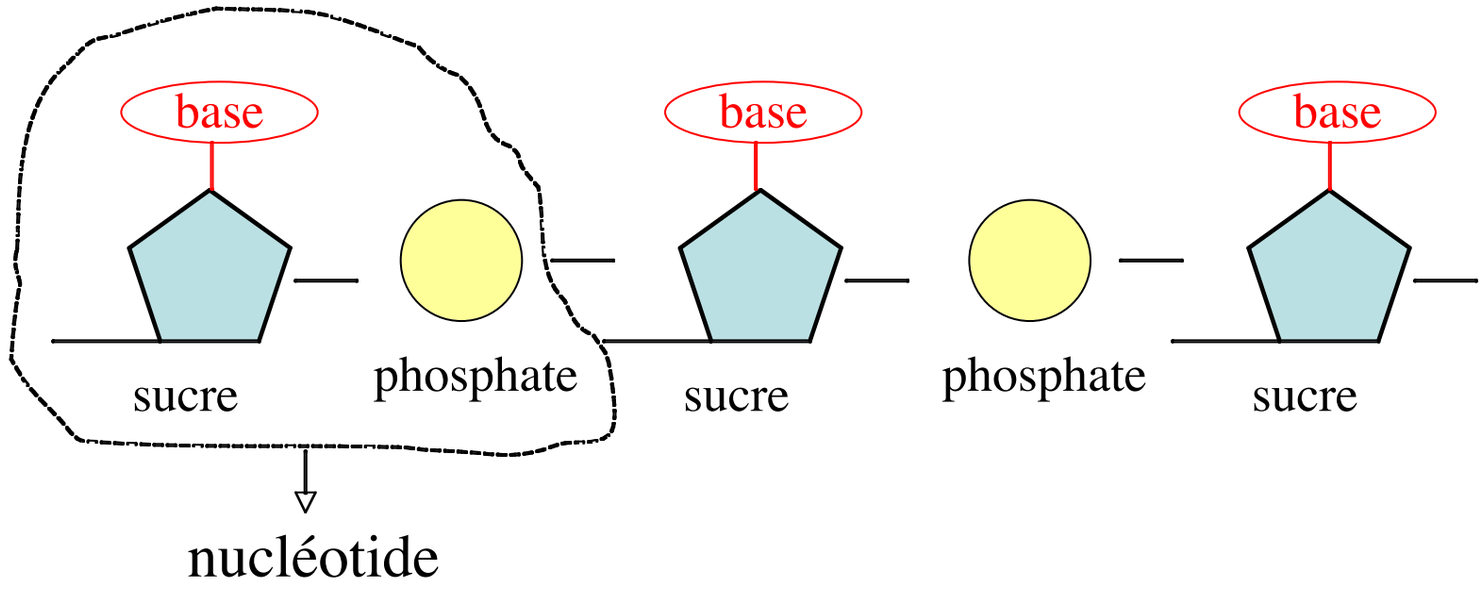
Rfam 11.0 (August 2012, 2208 families)

Rfam 12.0 (July 2014, 2450 families)

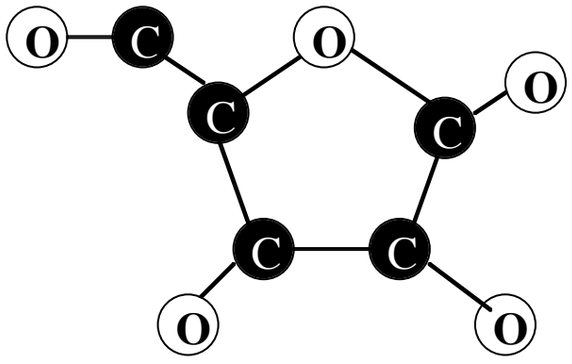
Rfam 12.1 (April 2016, 2474 families)

Concepts de base de la biologie moléculaire:

- Acides nucléiques = chaîne de nucléotides
  - La “colonne vertébrale” d’un acide nucléique est une suite alternée de sucres et de phosphates:

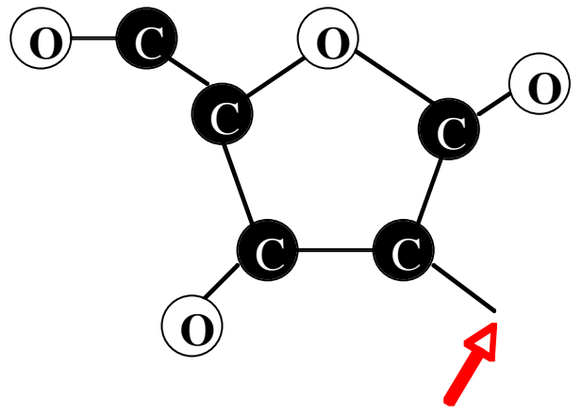


ADN versus ARN



ribose

Acide ribonucléique ou ARN

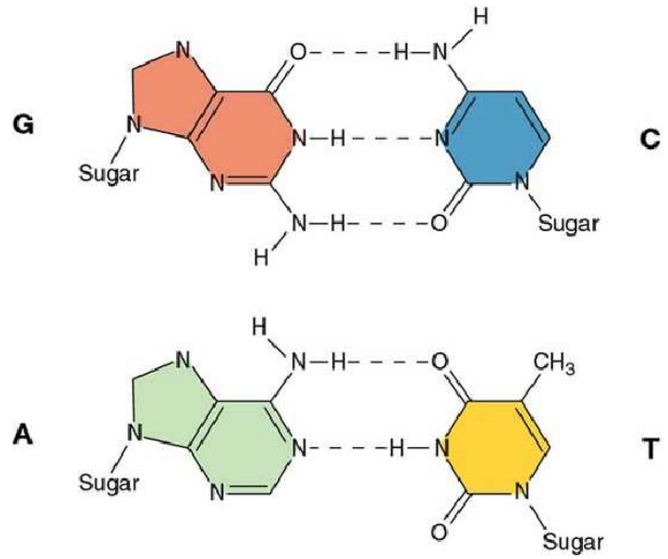


désoxyribose

Acide désoxyribonucléique ou ADN

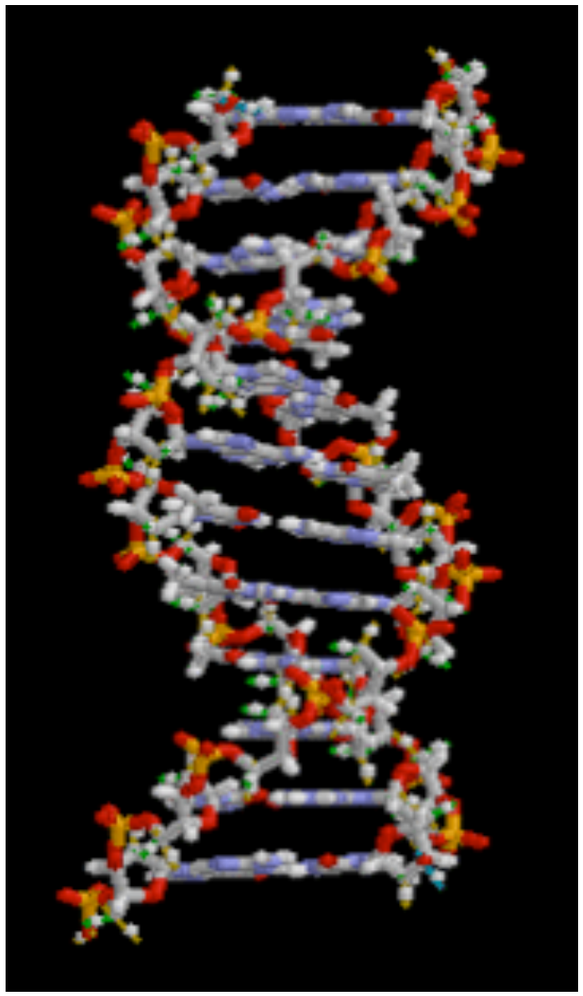
ADN: Acide DésoxyriboNucléique

- Toutes les cellules d'un organisme vivant contiennent le même code génétique
- ADN: 4 bases pour les nucléotides:
  - Purines: Adénine (A) et Guanine (G)
  - Pyrimidines: Cytosine (C) et Thymine (T)



- Structure tridimentionelle:
  - Deux brins face à face maintenus par des liens Watson-Crick
  - A-T G-C
- ADN linéaire : noyau
- ADN circulaire: procaryote et mithocondrie

ADN: Acide DésoxyriboNucléique



http://fr.wikipedia.org/wiki/Acide_désoxyribonucléique

ADN: Acide DésoxyriboNucléique

- Matériel génétique contenu dans plusieurs macromolécules d'ADN: les chromosomes
- Génome: Ensemble des chromosomes d'un organisme
- Organisme diploïde: Deux copies de chaque chromosome
  - Homme: 23 paires de chromosomes
  - Souris: 20 paires de chromosomes
- Taille des génomes:
  - Bactéries et virus : de 7800 (Commelina yellow mottle virus) à 8 millions de paires de base (Streptomyces coelicolor)
  - Eucaryotes: de 8 millions (certains champignons) à 3 milliards (humains) de paires de base.

ARN: Acide RiboNucléique

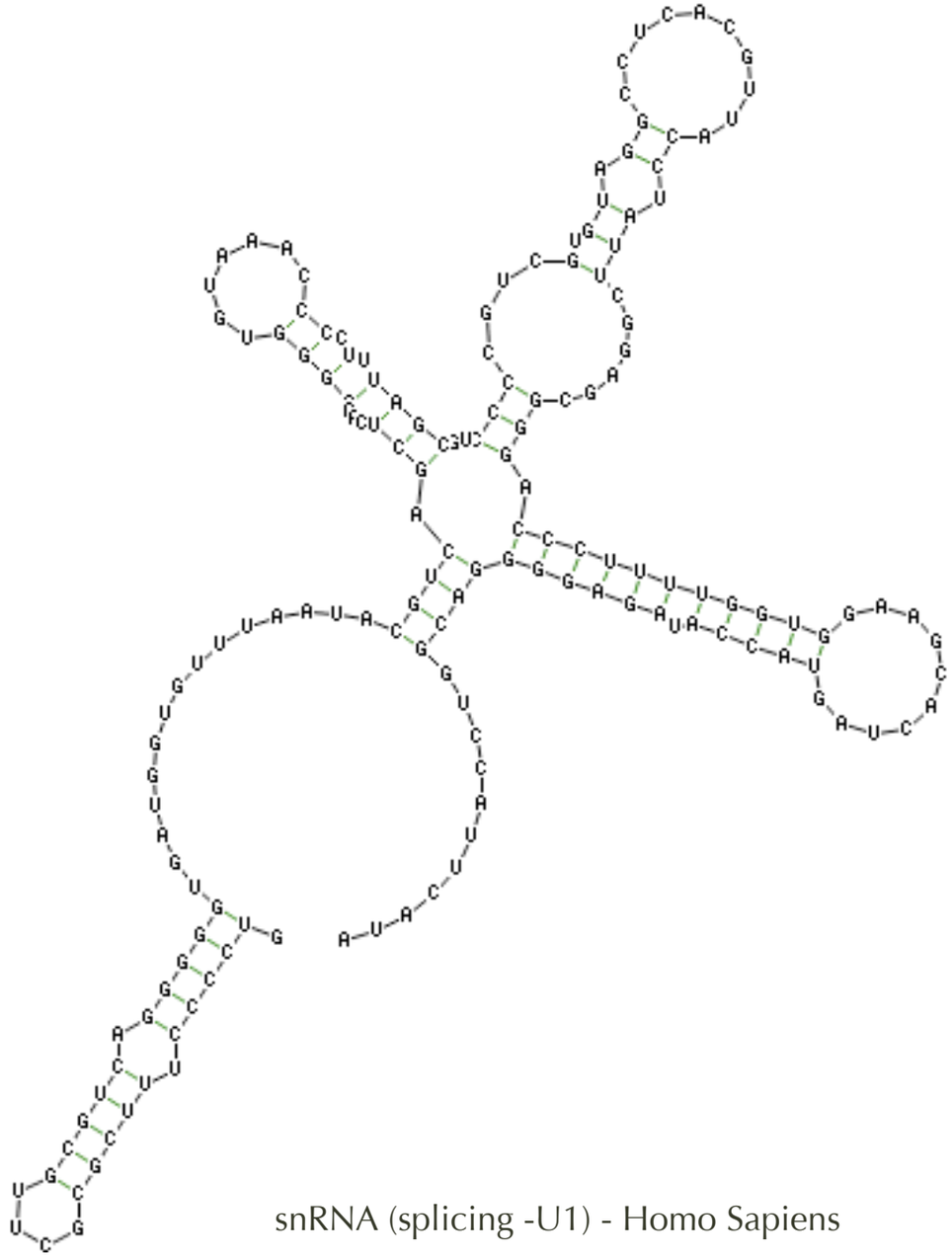
- Acide nucléique dont le sucre est le ribose et les bases possibles sont:
  - Purines: Adénine (A) et Guanine (G)
  - Pyrimidines: Cytosine (C) et Uracil (U) (à la place de la Thymine)

- 3 types principaux:
  - ARN messager (mRNA)
  - ARN de transfert (tRNA)
  - ARN ribosomique (rRNA)

ARN: Acide RiboNucléique

- Structure tridimensionnelle:
  - Généralement un seul brin
  - Bases complémentaire: A - U G - C
  - Linéaire pour mARN
  - Replié sur lui-même pour les autres ARN non codants

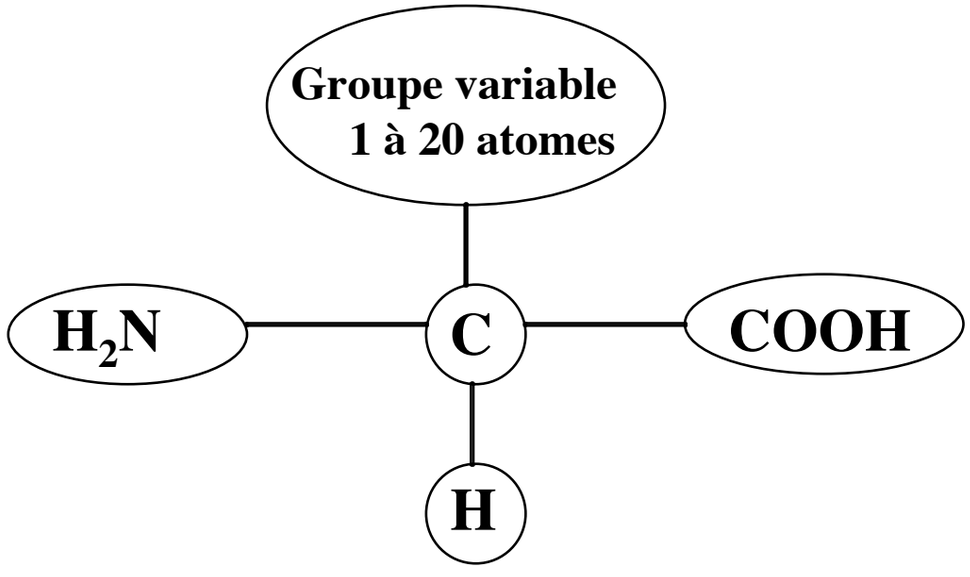
- Longueur des brins:
  - plutôt court: de quelques centaines à quelques milliers de bases



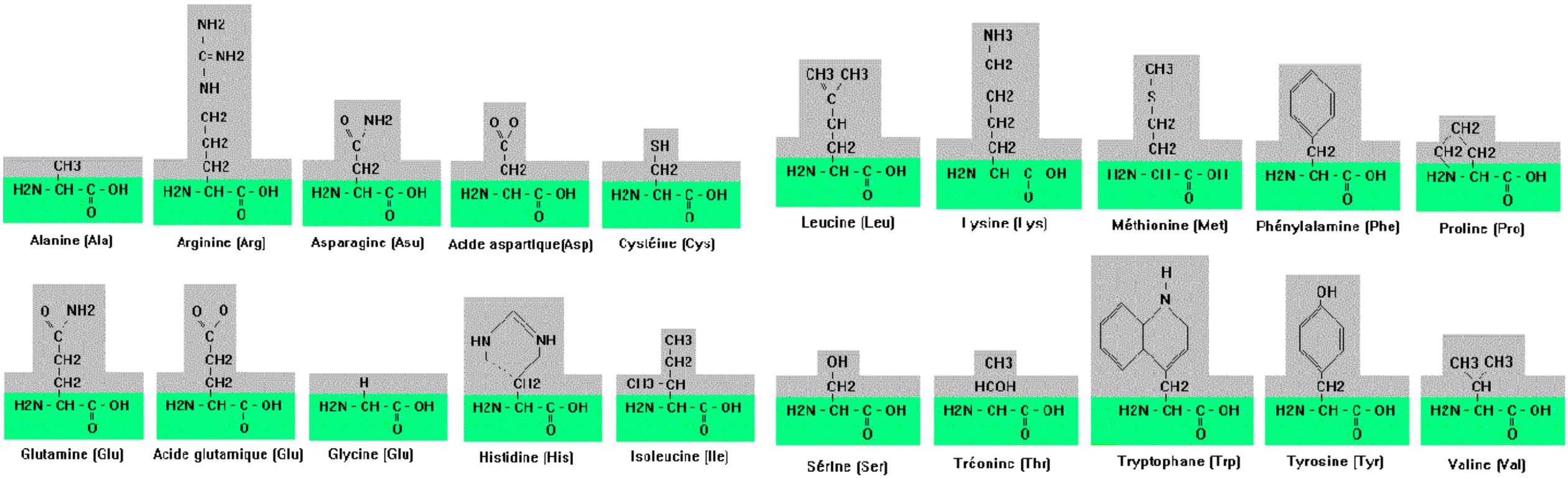
snRNA (splicing -U1) - Homo Sapiens

Protéines

- Chaînes d'acides aminés repliée sur elle-même en fonction des attractions entre ses composantes. Elles sont responsables de la plupart des fonctions d'une cellule.
  - 20 différents acides aminés

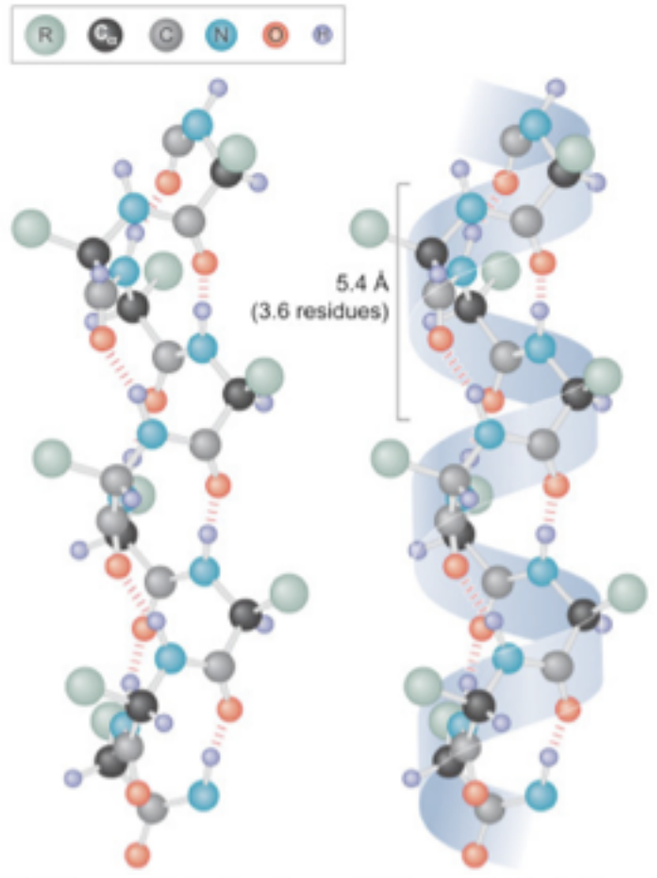


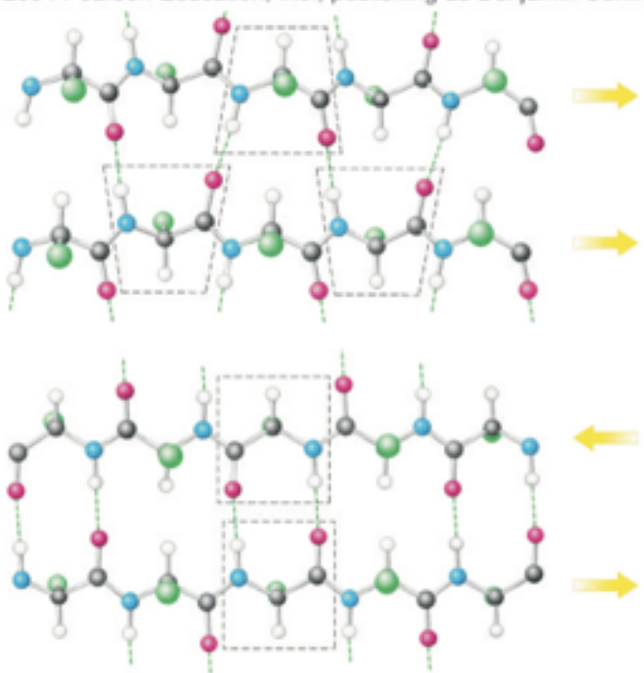
Protéines


Alanine (Ala)
Arginine (Arg)
Asparagine (Asu)
Acide aspartique(Asp)
Cystéine (Cys)
Leucine (Leu)
Lysine (Lys)
Méthionine (Met)
Phénylalamine (Phe)
Proline (Pro)
Glutamine (Glu)
Acide glutamique (Glu)
Glycine (Glu)
Histidine (His)
Isoleucine (Ile)
Sérine (Ser)
Tréonine (Thr)
Tryptophane (Trp)
Tyrosine (Tyr)
Valine (Val)

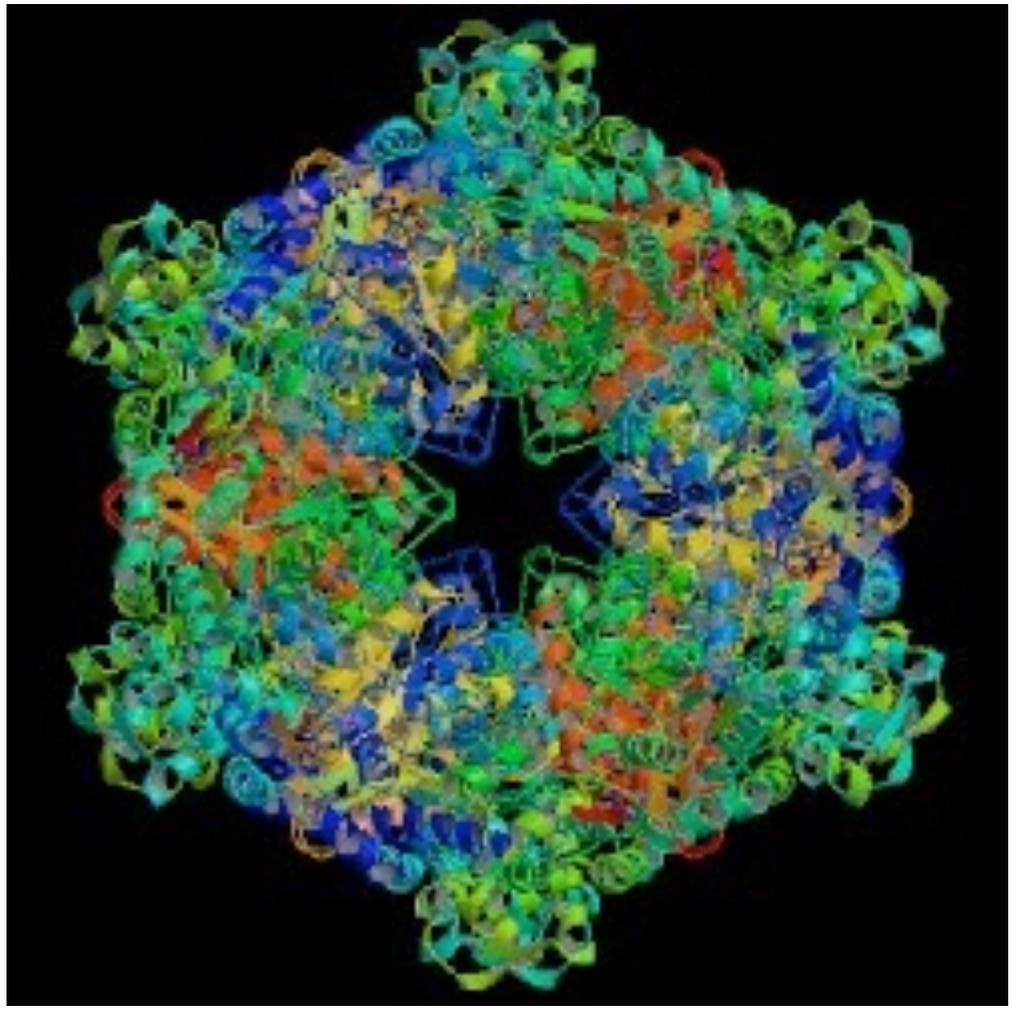
Protéines

- Longueur des brins
  - longueur typique au alentour de 300 AA
  - étendue: 100 - 5000 AA
- Structures
  - primaire: séquence d'acides aminés
  - secondaire: hélices alpha, feuillets beta, "turns"
  - tertiaire: forme tridimensionnelle
- Dogme:
  - la structure tridimensionnelle détermine la fonction d'une protéine





Protéines

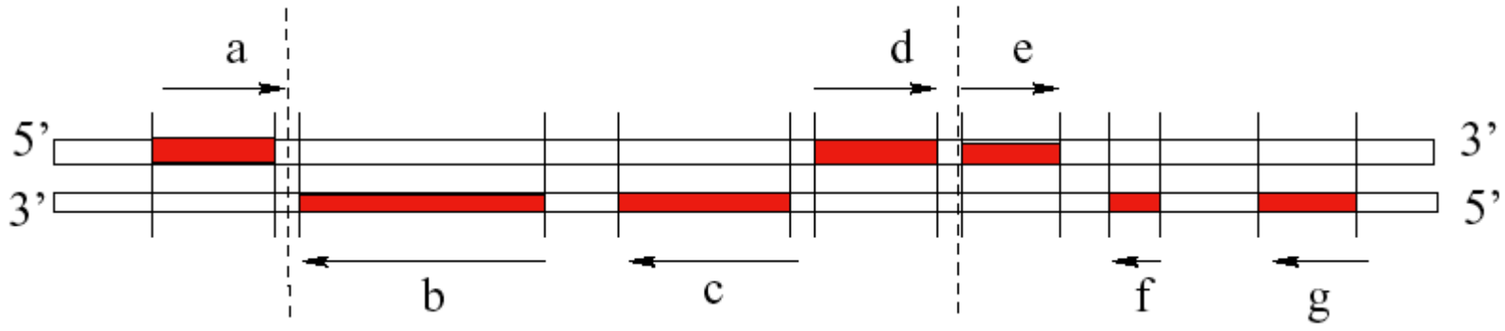


CHAPERONE/HYDROLASE 1YYF:
protein data bank

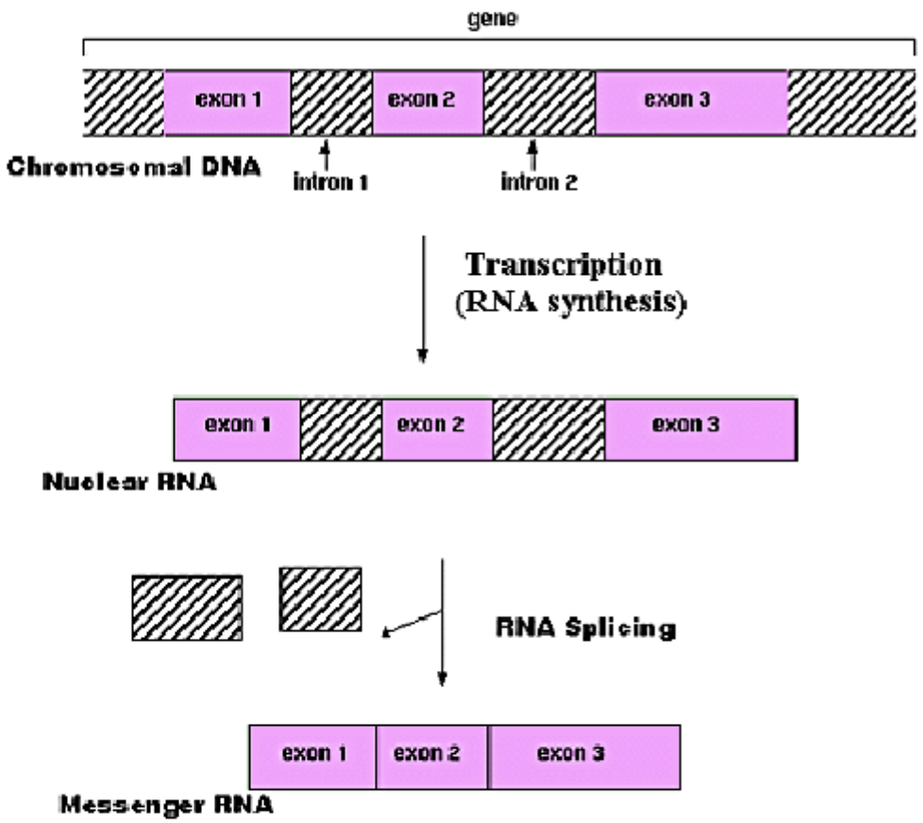
Transcription et traduction

$$\text{ADN} \xrightarrow{\textbf{transcription}} \text{ARN} \xrightarrow{\textbf{traduction}} \text{Protéine}$$

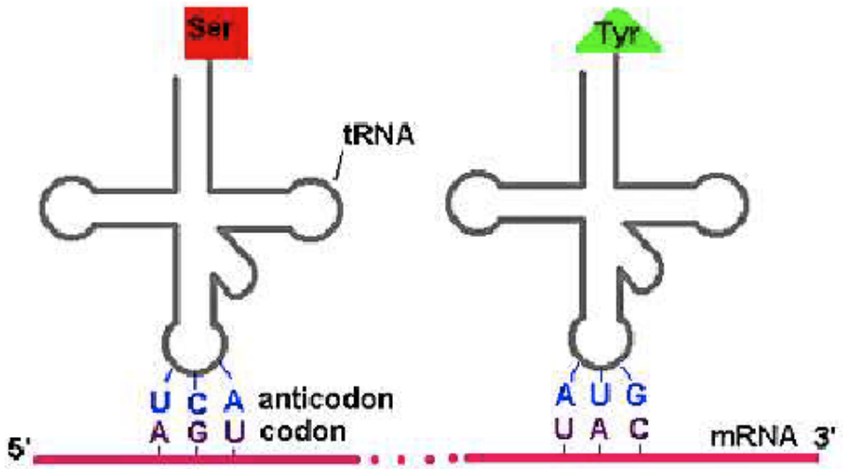
Gènes: partie codante de l'ADN



Un gène est formé d'introns et d'exons



RNA synthesis and processing



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

Défis de la biologie moléculaire:

- Décoder l'information contenue dans les séquences d'ADN et des protéines
  - Trouver les gènes
  - Différencier entre introns et exons
  - Trouver les régions répétitives dans l'ADN
  - Identifier les sites des facteurs de transcription
  - Étudier l'évolution des génomes

- Génomique structurale
  - Modéliser les structures 3D des protéines et des ARN structurels
  - Déterminer la relation entre structure et fonction

- Génomique fonctionnelle
  - Étudier la régulation des gènes
  - Déterminer les réseaux d'interaction entre les protéines

Pourquoi un cours d'algorithmique pour la bioinformatique:

- Éviter:
  - Utilisation systématique de BLAST sans comprendre son fonctionnement: interprétation fausse des résultats, conclusions hâtives, ...
  - Compréhension des outils informatiques se limitant à la modification de paramètres et de propriétés particulières
- Comprendre les méthodes algorithmiques utilisées
- Quelques idées algorithmiques de base permettent de résoudre un grand nombre de problématiques bioinformatiques
- Intérêts pour un informaticien: la bioinformatique est la source d'une multitude de nouveaux problèmes algorithmiques et statistiques

Pourquoi un cours d'algorithmique sur le texte:

- Analyse des séquences biologiques fondamentale pour répondre à un grand nombre de questions biologiques

- Intérêt des séquences biologiques
  - La séquence nucléotidique d'un gène détermine la séquence d'acides aminés de la protéine
  - La séquence d'une protéine détermine sa structure et sa fonction
  - Généralement, une similarité de séquence implique une similarité de structure et de fonction (l'inverse n'est pas toujours vrai)

- Exemples de base permettant d'introduire un grand nombre d'idées algorithmiques pouvant se transposer à l'étude de structures, graphes, arbres, ...

Problèmes bio-informatique nécessitant des méthodes d'algorithmique sur le texte

- Recherche dans les banques de données biologiques
  - Est-ce qu'une nouvelle séquence a déjà été complètement ou partiellement répertoriée?
  - Est-ce que cette séquence contient un gène?
  - Est-ce que ce gène appartient à une famille connue?
  - Existe-t-il d'autres gènes homologues?
- Alignement de séquences: Est-ce que deux séquences correspondent à deux gènes homologues?
- Recherche de sous-motifs communs à un ensemble de séquences. Établissement de consensus, alignement multiple
- Recherche de régions contenant des séquences répétées, recherche de gènes, recherche d'hélices d'ARN.