

Introduction à la bioinformatique

2. Fondation moléculaire en 66 diapos et 1 vidéo

1

Objectifs

- Expliquez la structure de l'ADN et l'ARN
- Expliquez le dogme de la biologie moléculaire
- Expliquez le rôle de l'ARNm
- Expliquez comment l'ARNm est traduit en protéine
- Expliquez comment l'expression de gènes est contrôlée
- Montrez comment les *introns* sont enlevés de l'ARNm
- Récapitez comment l'évolution se produit

3

Bibliographie

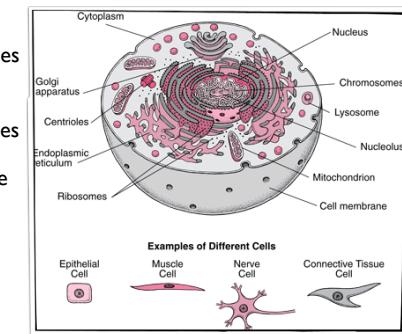
- Alberts et al, The molecular biology of the cell
- Zvelebil et Baum, Understanding bioinformatics (p3. - 44)



2

Préambule: Les cellules

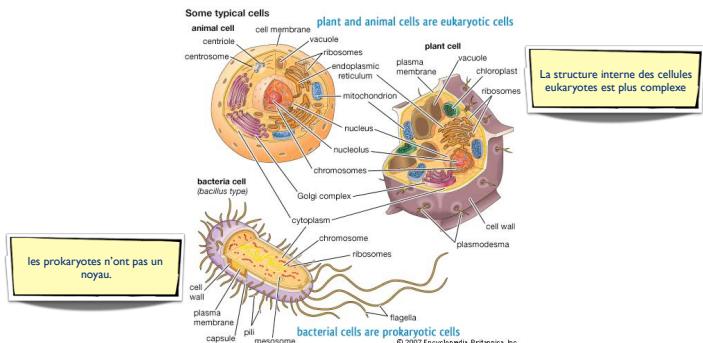
- L'unité fondamentale de la vie
- Toutes les cellules partagent les mêmes mécanismes
- Les cellules produisent d'autres cellules
 - en passant l'information nécessaire pour reconstruire toutes les fonctions ([héritage](#))
- Les cellules traitent l'information comme des ordinateurs
 - nourriture, survie, ..



4

Préambule: Les cellules 2

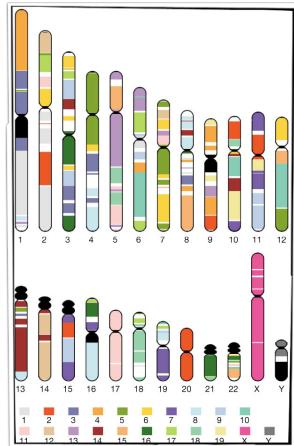
Deux grandes classes de cellules



5

Préambule: Les cellules 4

- L'homme a **46 molécules d'ADN** dans chaque cellule organisées dans les chromosomes
- Dans les bactéries, il y a un seul chromosome circulaire



7

Préambule: Les cellules 3

- Toute l'information est encodée par une structure spécifique : l'**ADN**

- Acide désoxyribonucléique

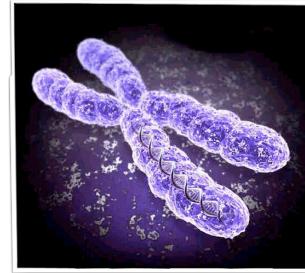
- L'information d'un type de cellule peut être traitée par des autres types de cellules.

- L'ADN fait partie d'un **chromosome**

- L'ADN contient l'information pour produire des milliers de protéines (1% du génome)

- un **gène** est une partie de l'ADN qui correspond à une seule protéine.

- le **génome** est l'ensemble des molécules d'ADN.



6

Préambule: Les cellules 5

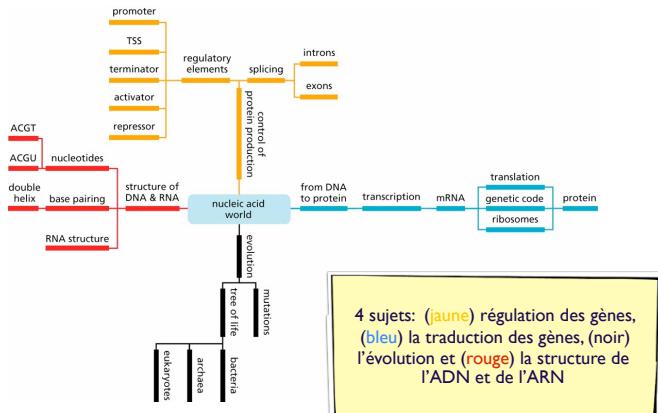
Organisme	Année	Taille (Mb)
<i>Mycoplasma genitalium</i>	1995	0,6
<i>Haemophilus influenzae</i>	1995	1,8
<i>Escherichia coli</i>	1997	4,6
<i>Saccharomyces cerevisiae</i>	1996	12
<i>Schizosaccharomyces pombe</i>	2002	14
<i>Caenorhabditis elegans</i>	1998	97
<i>Arabidopsis thaliana</i>	2001	120
<i>Oryza sativa</i>	2002	5 000
<i>Drosophila melanogaster</i>	2000	180
<i>Galus Galus</i>	2004	1 200
<i>Rattus Norvegicus</i>	2004	2 900
<i>Mus musculus</i>	2002	3 400
<i>Homo sapiens</i>	2001	3 400

1 Mb = 1 000 000 bases

La taille des génomes

8

Carte 1: le monde de l'ADN



9

... l'ADN et l'ARN

l'ADN et l'ARN sont des structures linéaires qui sont composées de 4 types de nucléotides

Un nucléotide est composé de 3 parties (Fig.A) : une **base**, un **sucré** et un **groupe phosphate**

La seule différence entre les 4 types est la base : A,T,G et C

building block of DNA

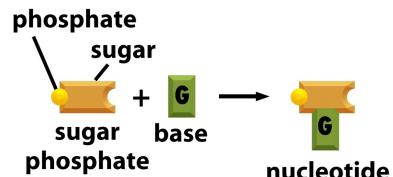
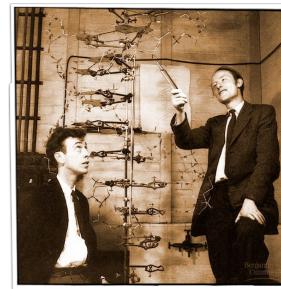


Figure 1-2a Molecular Biology of the Cell 5/e (© Garland Science 2008)

11

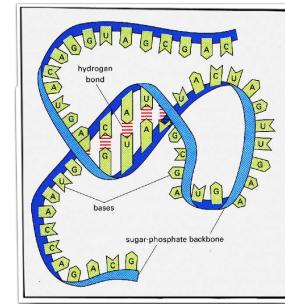
La structure de ...

ADN (acide désoxyribonucléique)



Watson et Crick (1953)

ARN (acide ribonucléique)



10

... l'ADN et l'ARN

l'ADN et l'ARN sont des structures linéaires qui sont composées de 4 types de nucléotides

Une nucléotide est composé de 3 parties (Fig.A) : une **base**, un **sucré** et un **groupe de phosphate**

La seule différence entre les 4 types est la base : A,T,G ou C

DNA strand



Figure 1-2b Molecular Biology of the Cell 5/e (© Garland Science 2008)

Les molécules ADN sont des séquences de millions de bases

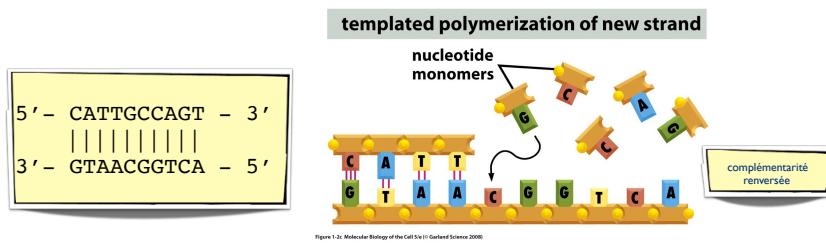
le début de l'ADN est annoté par **5'** et le fin par **3'**.

un brin = une séquence de bases

12

... l'ADN et l'ARN 2

L'ADN est composé de deux brins complémentaires



13

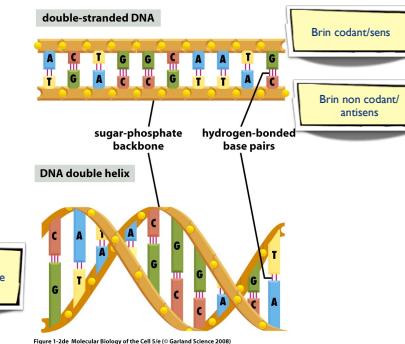
... l'ADN et l'ARN 2

L'ADN est composé de deux brins complémentaires

Les deux brins sont enroulés et liés par des liaisons hydrogènes (liens faibles) au sein de la structure

Les paires de bases sont toujours entre un pyrimidine et un purine :
A-T et **C-G**

L'ordre des deux brins est inversé



14

... l'ADN et l'ARN 3

L'ADN est constitué par deux brins complémentaires

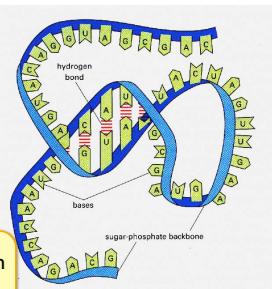
Les deux brins sont enroulés et liés par des liaisons hydrogènes (liens faibles) au sein de la structure

l'ARN est constitué par un seul brin

cette molécule peut également se plier dans une structure 3D.

les paires sont maintenant : **A-U** et **C-G**

Uracil (U) remplace Thymine (T) en ARN !!



15

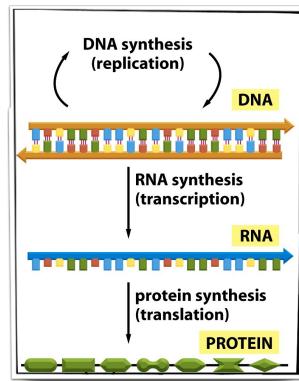
Les questions du bio-informaticien

- L'ordre des nucléotides dans les séquences est-il aléatoire ?
- S'il n'est pas aléatoire, est-ce qu'on peut prédire la fonction et la structure en utilisant la séquence?
- Est-ce qu'on peut trouver des autres séquences qui ressemblent à des séquences/structures connues?

16

le dogme central de la biologie moléculaire

- Les gènes sont traduit en protéines en deux étapes
- étape 1 : transcription
 - Des parties de l'ADN sont copiées dans des séquences plus courtes = ARN messager (ARNm)
- étape 2 : traduction
 - L'ARNm est traduit par l'ARN-polymérase vers une séquence d'acides aminées = protéine



17



19

... l'ADN et l'ARN 4

La réplication de l'ADN

Pendant la division de la cellule, les deux brin de l'ADN sont séparés

Chaque brin sert comme un plan pour la production d'un brin complémentaire

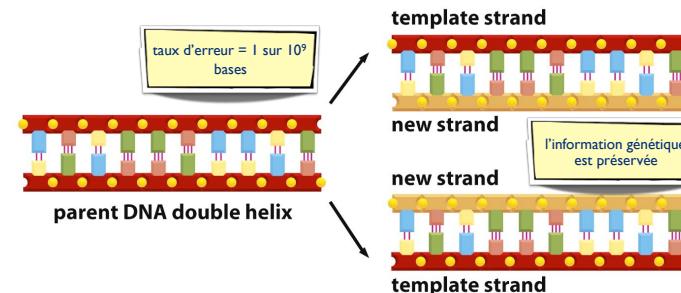
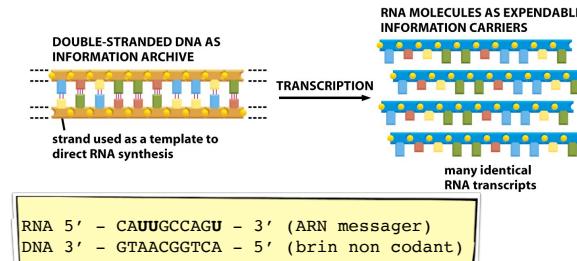


Figure 1-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

18

La transcription

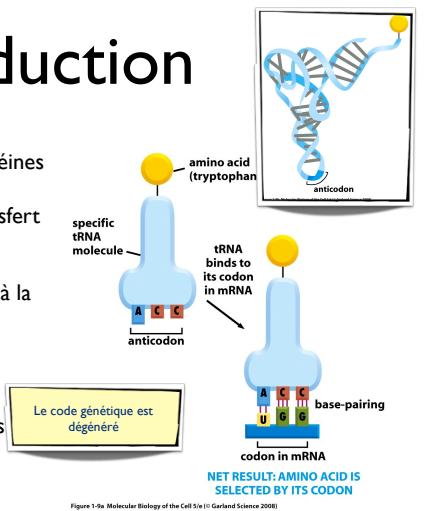
- L'ARNm est un brin complémentaire d'un des brins de l'ADN
 - on prend le brin non codant (antisens) de l'ADN
- Le nombre d'ARNm dans la cellule = le niveau d'expression d'un gène



20

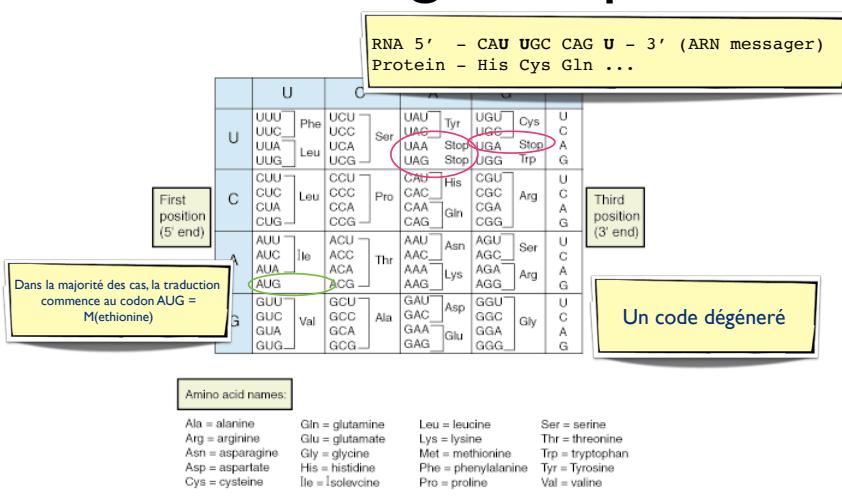
La traduction

- Traduction de l'ARNm en protéines
 - en utilisant les ARN de transfert (ARNt)
- ARNm est composé de trois nucléotides à la fois (codon)
- $4^3 = 64$ codons
- seulement 20 acides aminés
- La traduction est faite par le ribosome



21

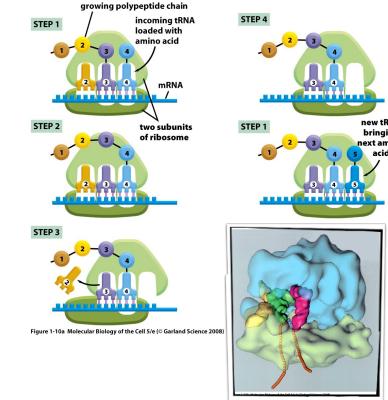
Le code génétique



23

La traduction 2

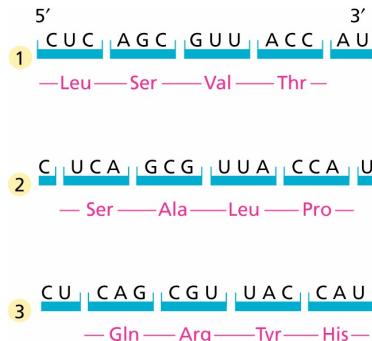
- Le ribosome commence au 5' et bouge dans la direction du 3'
- Il attrape des molécules ARNt qui peuvent s'associer aux codons de l'ARNm.
- Les acides aminés liés à l'ARNt se lieront à la séquence existante
- Les codons ne se superposent pas



22

Le code génétique 2

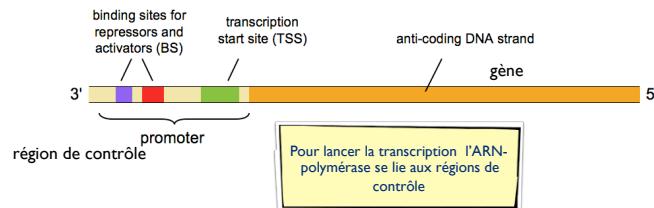
Comme les codons ne se superposent pas, on a trois possibilités de départ pour la traduction (reading frames)



24

Contrôle de la production des protéines

Un gène est la partie de l'ADN qui code pour une protéine. En pratique, on définit parfois le gène comme la combinaison entre la partie codante et la région de contrôle



25

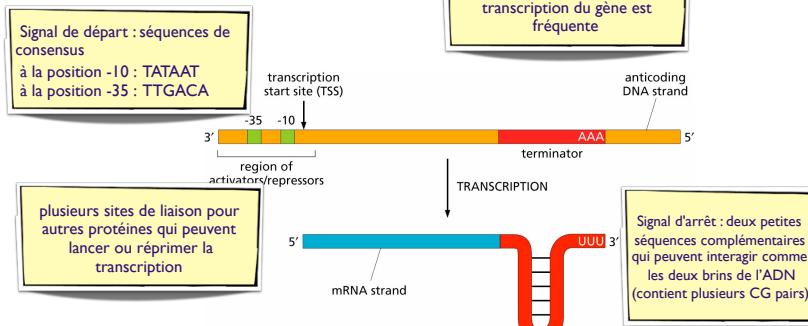
Les questions du bio-informaticien

- Est-ce qu'on peut détecter les régions de contrôle dans l'ADN?
- Est-ce qu'il existe une préférence pour certaines combinaisons de nucléotides dans les régions de contrôle?
- Quelles sont les motifs qui indiquent le début et la fin de la région codante ?

26

Contrôle de la production des protéines 2

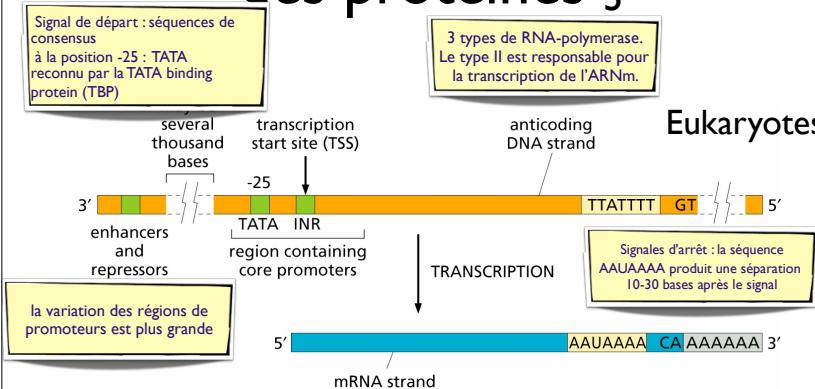
Prokaryotes



27

Contrôle de la production des protéines 3

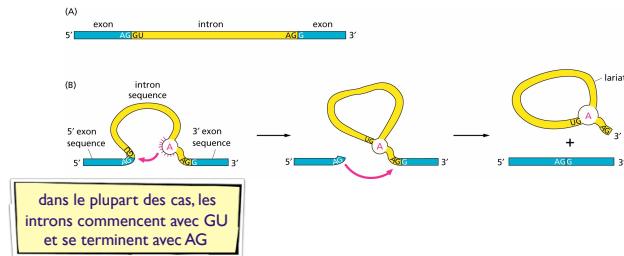
Eukaryotes



28

Contrôle de la production des protéines 4

Dans les **eukaryotes** les ARN produits par la transcription sont modifiés avant qu'ils soient traduits. Les **exons** (**introns**) sont des parties de l'ADN qui (ne) sont (pas) traduits en acides aminés.



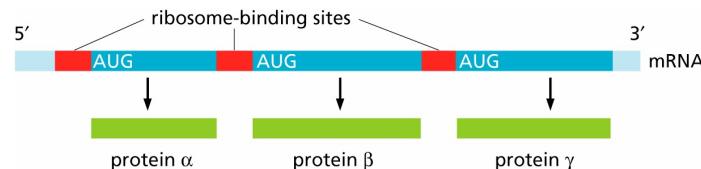
Épissage (Splicing) est le mécanisme qui enlève les introns de l'ARNm.

29

Contrôle de la production des protéines 5

Chez les prokaryotes il y a aussi une séquence courte pour positionner le ribosome = **séquence de Shine-Dalgarno** = AGGAGGU

Chez les prokaryotes plusieurs gènes peuvent être contrôlés par un seul promoteur = **opéron**



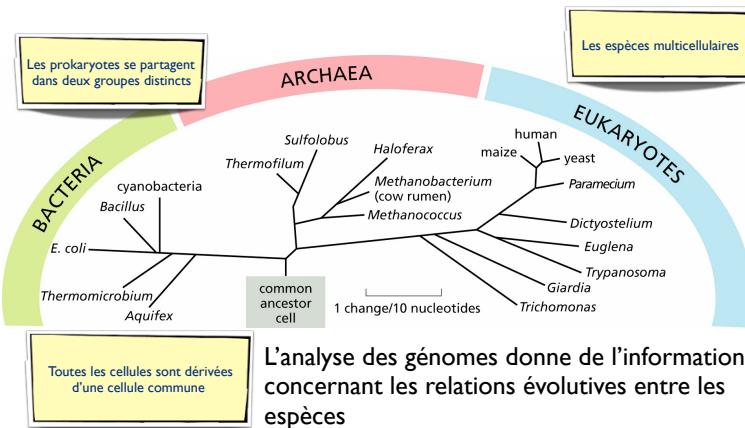
31

Les questions du bio-informaticien

- Est-ce qu'on peut trouver automatiquement les introns et exons?

30

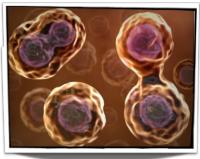
Évolution



32

Évolution 2

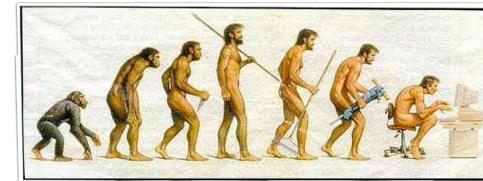
- **Hérité** = transfert de l'information génétique à partir du parent vers la progéniture
 - Quand une cellule se divise, l'ADN est copié et partagé entre les deux nouvelles cellules.
 - Parfois, des erreurs (**mutations**) produites pourraient:
 - **Améliorer** la fonctionnalité de la cellule = avantage sélectif
 - **Endommager** la fonctionnalité de la cellule = désavantage sélectif, peut tuer la cellule
 - **Ne rien changer** dans la fonctionnalité de la cellule = sélectivement neutre



33

Évolution 3

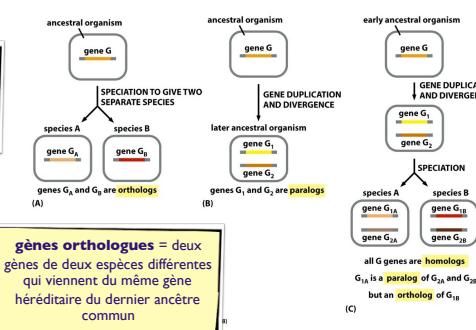
- “*trial and error*” permet à des cellules et à des organismes d'évoluer
- Certaines parties de l'ADN évoluent plus facilement que d'autres
 - Les régions non-codantes/non-contrôlées de l'ADN
- Les régions qui sont importantes pour la fonction d'une protéine doivent être conservées



34

les questions du bio-informaticien

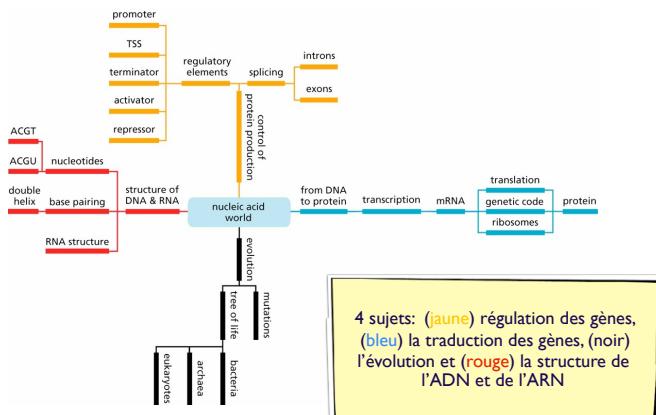
- Etant donné un ensemble de séquences, peut-on déterminer leurs relations évolutives ?
- Si deux organismes ont une ancêtre commun, on peut déterminer les fonction de leurs gènes en utilisant les homologues connus.



35

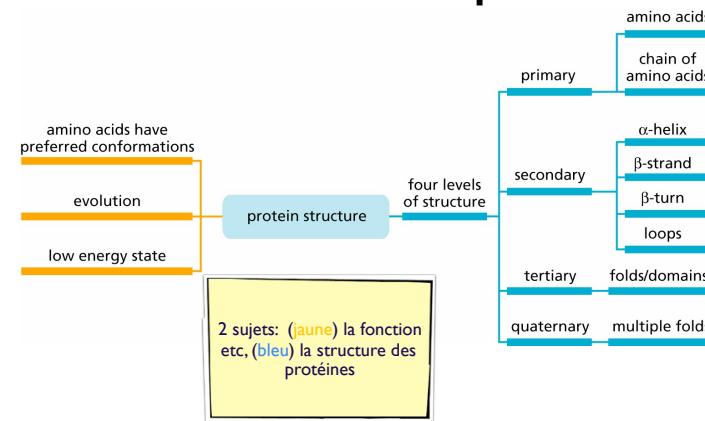
36

Carte I: le monde de l'ADN



37

Carte II : la structure et la fonction des protéines



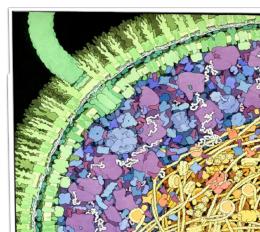
38

Objectifs

- Discuter l'importance des protéines pour la fonction de la cellule
- Expliquer les différents niveaux dans la structure des protéines
- Décrire pourquoi les acides aminés sont les modules essentiels des protéines
- Montrer comment les séquences d'acides aminés se forment
- Expliquer les types de structure secondaire
- Expliquer le pliage d'une protéine
- Montrer pourquoi la structure est importante pour la fonction d'une protéine

39

protéines

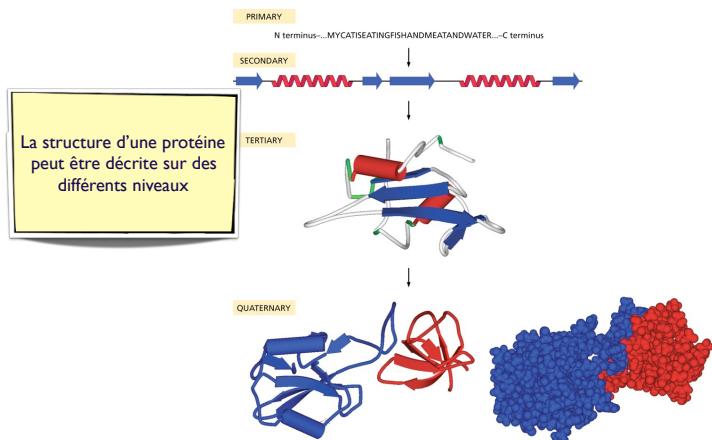


catalyser de réactions (enzymes), la régulation d'expression des gènes (facteurs de transcription), la structure des cellules (cytoskeleton), voies de signalisation, ...

- La majorité des objets dans les cellules sont des protéines
- Les protéines sont ...
 - les modules qui forment la structure de la cellule
 - les robots qui remplissent presque chaque fonction dans la cellule:
- Le génome humain contient entre 25000 et 35000 protéines différentes

40

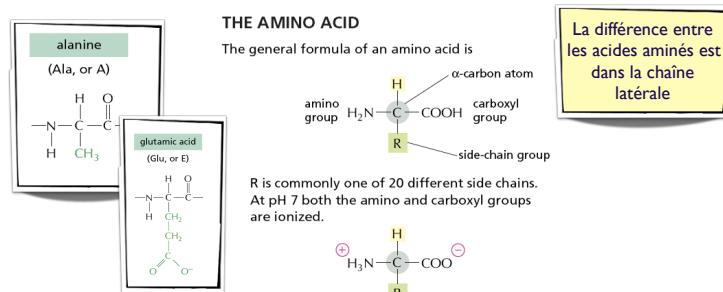
La structure ...



41

Les acides aminés

Tous les acides aminés (AA) sont composés de trois parties : le C α central, groupe d'azote, groupe carbone et la chaîne latérale



43

... primaire

N terminus—...MYCATISEATINGFISHANDMEATANDWATER...—C terminus

Comme l'ADN et l'ARN, les protéines sont des séquences de molécules simples = **acides aminés**

AMINO ACID	SIDE CHAIN	AMINO ACID	SIDE CHAIN
Aspartic acid	Asp D	negative	Alanine Ala A nonpolar
Glutamic acid	Glu E	negative	Glycine Gly G nonpolar
Arginine	Arg R	positive	Valine Val V nonpolar
Lysine	Lys K	positive	Leucine Leu L nonpolar
Histidine	His H	positive	Isoleucine Ile I nonpolar
Asparagine	Asn N	uncharged polar	Proline Pro P nonpolar
Glutamine	Gln Q	uncharged polar	Phenylalanine Phe F nonpolar
Serine	Ser S	uncharged polar	Methionine Met M nonpolar
Threonine	Thr T	uncharged polar	Tryptophan Trp W nonpolar
Tyrosine	Tyr Y	uncharged polar	Cysteine Cys C nonpolar

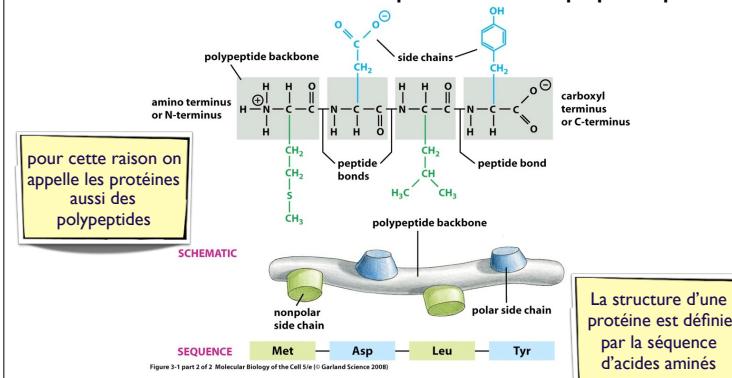
hydrophile hydrophobe

POLAR AMINO ACIDS NONPolar AMINO ACIDS

42

Les acides aminés 3

Les AA sont enchaînés par des liens peptidiques



44

Repliement

Les conformations de la chaîne principale sont limitées

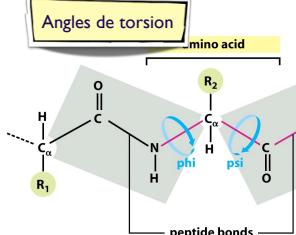
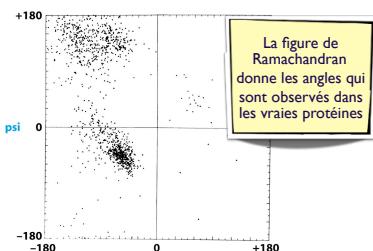


Figure 3-5 Molecular Biology of the Cell Site (© Garland Science 2008)

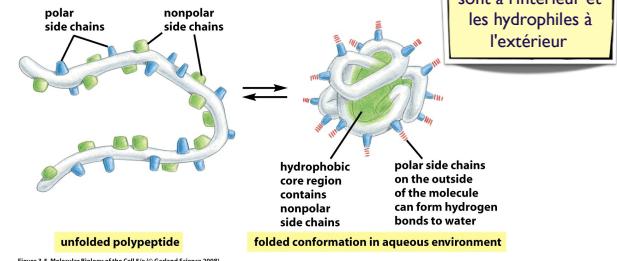
Les liens peptidiques sont planaires. Rotation est permise autour les liens N-C (ϕ) et C α -C (ψ)



Seulement les angles ϕ et ψ qui donnent pas des désaccords sont acceptables

45

le pliage 2



Une séquence d'acides aminés se plie dans une conformation avec l'énergie la plus basse.

La structure pliée d'une protéine n'est pas fixée ! Elle change tout le temps et la configuration préférée peut changer quand elle est liée à une autre protéine

46

le pliage 3

Le pliage des protéines est dirigé par 3 types d'interactions entre les acides aminés

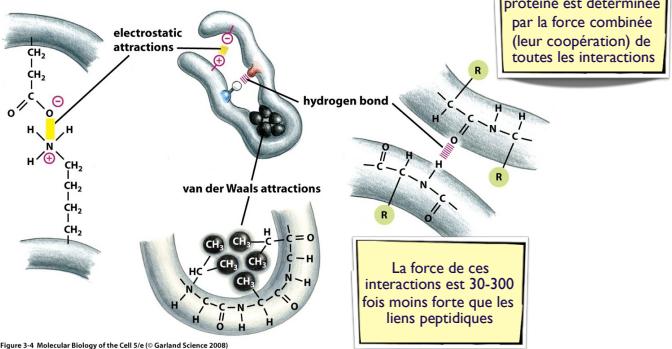


Figure 3-4 Molecular Biology of the Cell Site (© Garland Science 2008)

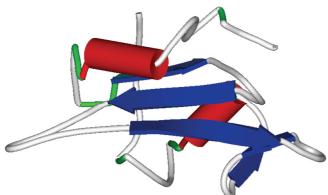
47

Les questions du bioinformaticien

- Est-ce qu'on peut quantifier correctement la stabilité d'une protéine?
- Est-ce qu'on peut prédire la structure des protéines étant donnée la séquence?
- Est-ce qu'on peut grouper les structures des protéines dans des classes différentes?

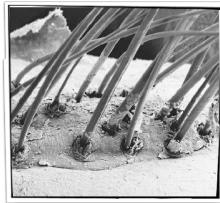
48

... tertiaire



On discute ici seulement des protéines globulaires
l'autre type sont les protéines fibreuses

La structure 3D de la protéine plié est nommée la structure tertiaire

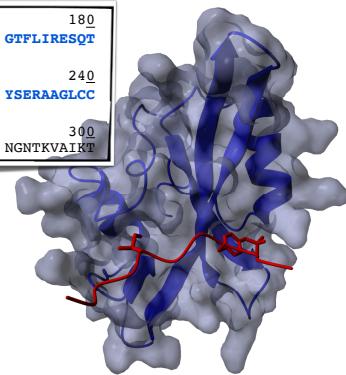


49

Un exemple

La séquence et la structure de Fyn SH2

130	140	150	160	170	180
EARSLTITGET	GYIPSNVAP	VDSIQAEEWY	FGKLGRKDAE	RQLLSFGNPR	GTFLIRESQ
190	200	210	220	230	240
TKGAYSLSIR	DWDDMKGDHV	KHYKIRKLDN	GGYYITTRAQ	FETLQQQLVQH	YSERAAGLCC
250	260	270	280	290	300
RLVVPCCHKGM	PRLTDLSVKT	KDVWEIPRES	LQLIKRLGNG	QFGEVWLGTW	NGNTKVAIKT



50

Évolution

- Le nombre de séquences de taille n est immense
- Seulement une partie de toutes ces protéines a une forme pliée stable
- Mais (presque) toutes les protéines dans les cellules ont une forme stable ?
 - Réponse : [La sélection naturelle](#)
 - une protéine avec une structure ou une activité biochimique variable n'est pas pertinente pour la survie de la cellule
- L'évolution par sélection naturelle a éliminé ces types de protéines

$20 \times 20 \dots \times 20 = 20^n$
 $n=300$ donne 10^{390} chaînes

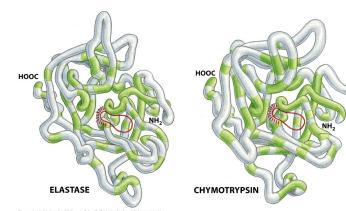
estimation de +/- 1 protéine par 1 billion

Pour ces raisons, les protéines qu'on connaît maintenant ont une conformation stable et une fonction bien précise

Les familles de protéines

- A partir du moment où les protéines sont devenues stables et si elles ont des caractéristiques intéressantes, la structure pourrait évoluer
- La duplication et la divergence des gènes peuvent créer des protéines avec la même structure mais avec une fonction différente, c.a.d. une préférence d'interaction avec d'autres protéines ([spécificité](#))

ces deux protéines ont une structure et une séquence similaire mais il ont des activités différentes.



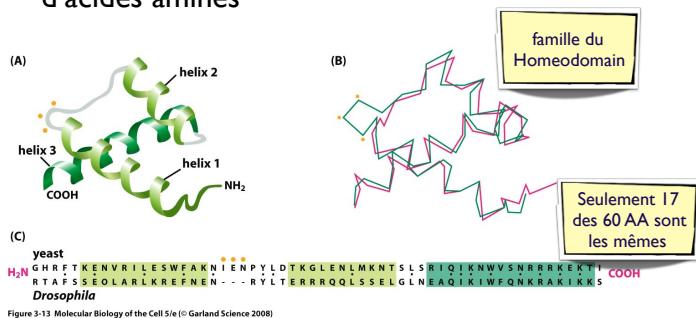
Des familles de protéines sont définies par des similarités de séquence et de structure

51

51

Les familles de protéines 2

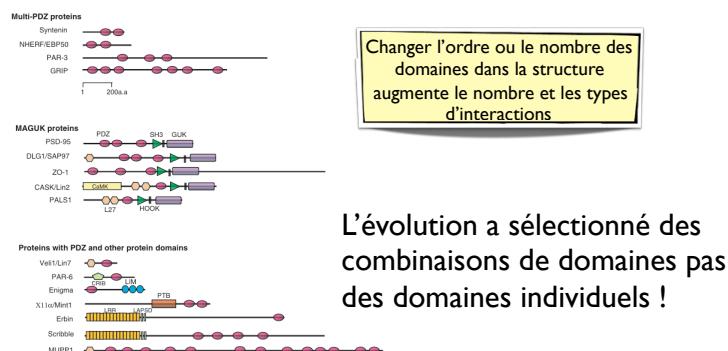
- La structure des membres d'une famille est beaucoup mieux préservée que la séquence d'acides aminés



53

les domaines 2

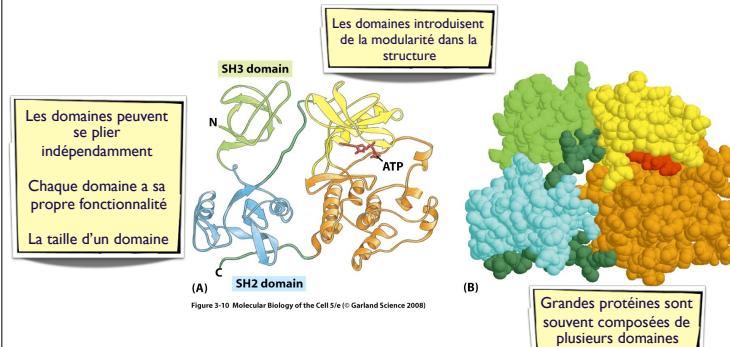
Réutiliser les domaines est un autre mécanisme d'évolution



55

les domaines

Les protéines peuvent également se composer de multiples pièces globulaires = les domaines



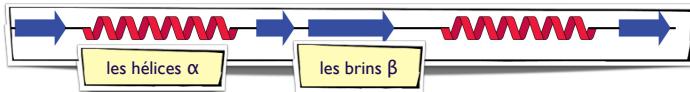
54

Les questions du bioinformaticien

- Est-ce qu'on retrouve des formes structurales fréquentes dans les structures de protéines ?
- Est-ce qu'on peut trouver une relation entre les structures et les acides aminés qui font parties de ces structures?

56

... secondaire



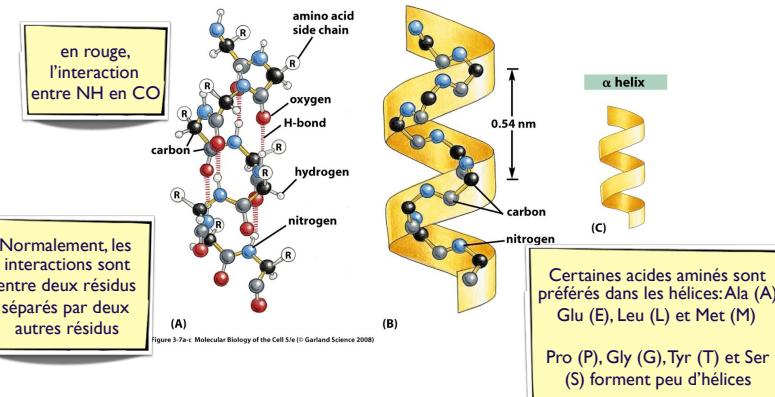
Quand on analyse beaucoup de structures tertiaires, on peut trouver des **régularités géométriques**.

les hélices et brins sont produits par des liaisons hydrogènes entre les groupes NH et CO de la chaîne principale

Entre 50-80% des résidus dans une protéine peuvent être classifiés en tant qu'une de ces structures régulières

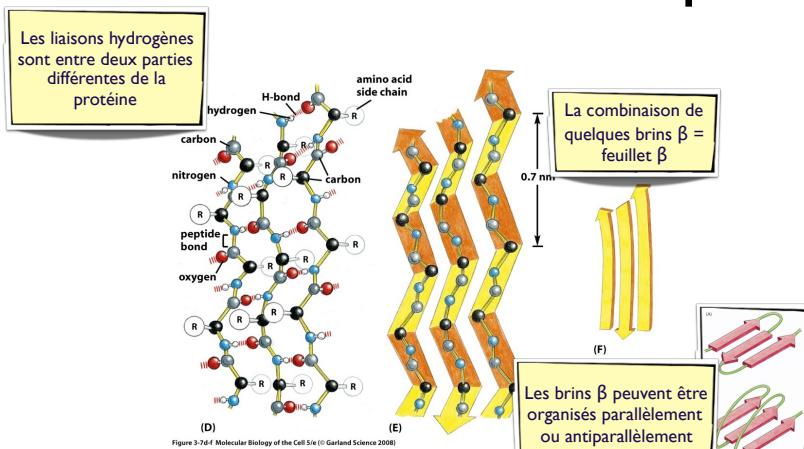
57

les hélices α et brins β



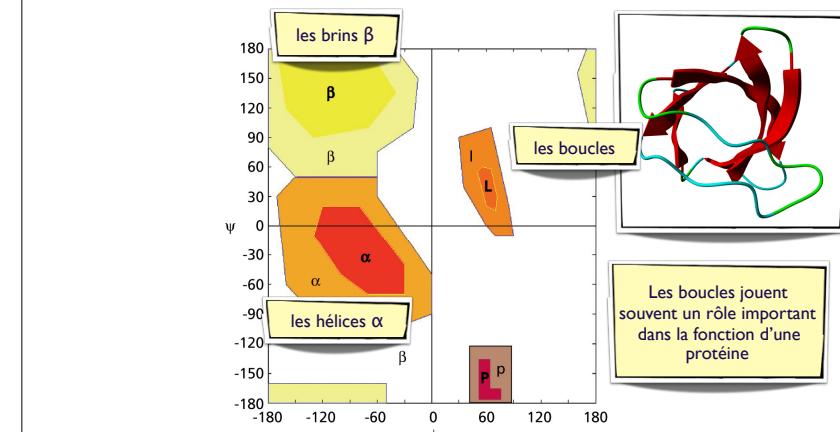
58

les hélices α et brins β 2



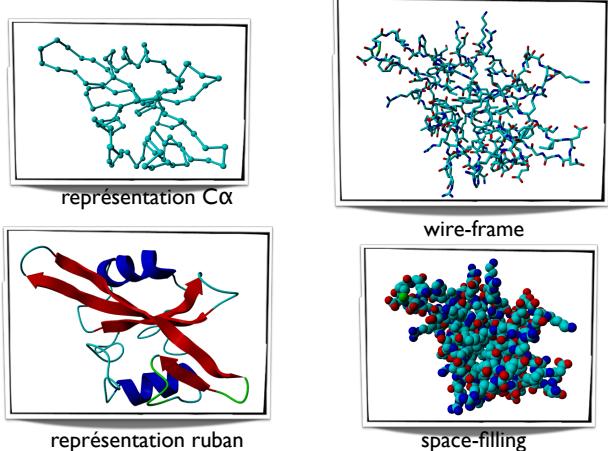
59

les hélices α et brins β 3



60

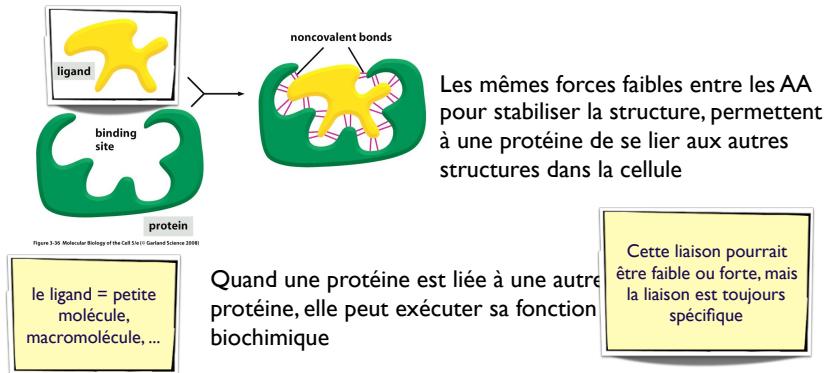
la visualisation de la structure tertiaire



61

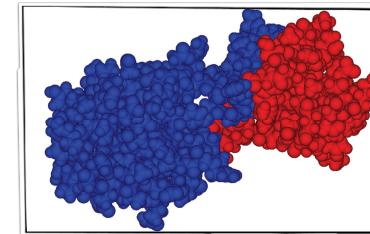
La fonction des protéines

La fonction d'une protéine est définie par sa structure



63

... quaternaire



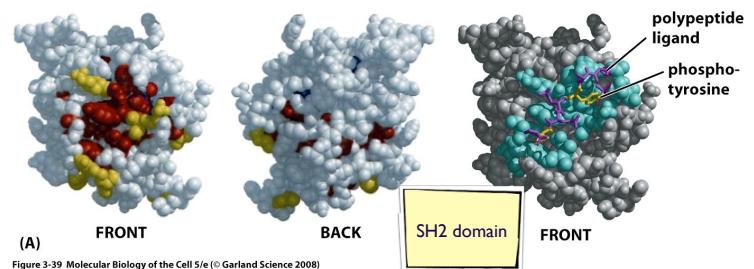
La majorité des protéines sont composées de deux ou plusieurs unités indépendantes



62

La fonction des protéines 2

Les sites de liaison sont préservés dans des familles de protéines



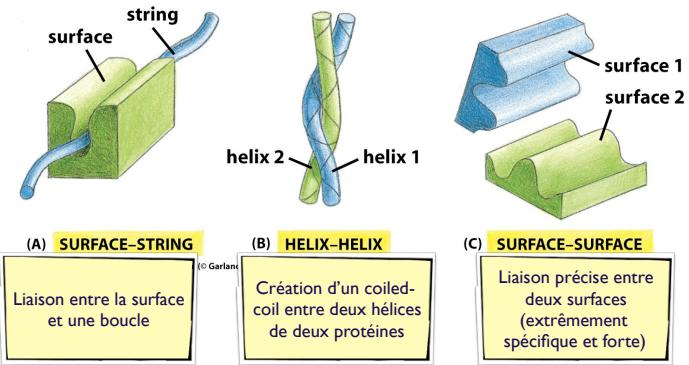
Traçage évolutif :

Tracer les AA conservés sur la structure du domaine
Souvent ces AA qui se regroupent ensemble sur la surface sont reliés au site de liaison

64

La fonction des protéines 2

Aux moins 3 formes de liaison



65

Les questions du bioinformaticien

- Beaucoup de méthodes d'apprentissage automatique pourraient être utilisées pour résoudre des questions biomoléculaires
 - Techniques de classification et régression
 - Méthodes de clustering
 - Hidden Markov models
 - Neural networks
 -

67

La fonction des protéines 3

Les molécules se rencontrent aléatoirement

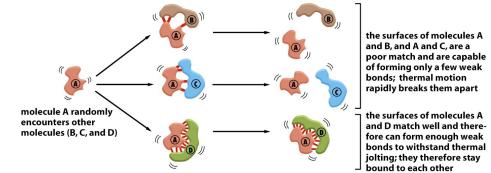
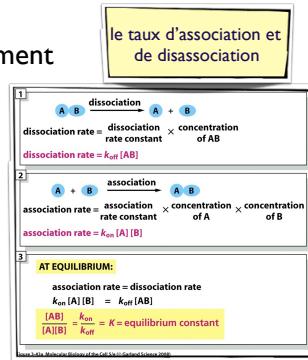


Figure 3-42 Molecular Biology of the Cell 5e (© Garland Science 2008)



Quand il se touche et que les surfaces ne sont pas bien adaptées, l'association ne reste pas longtemps

66