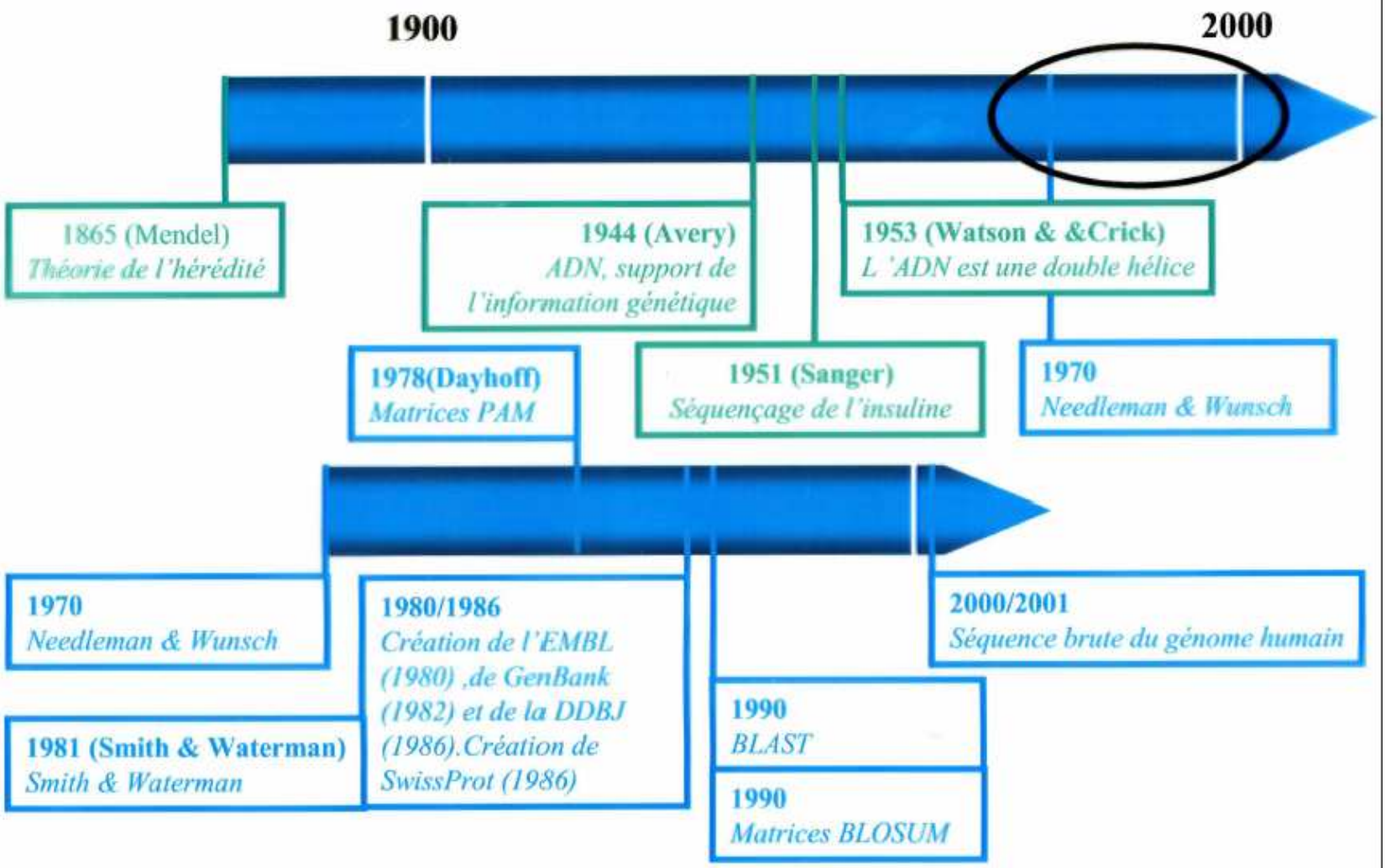
Bioinformatique M1: Lecture 1

P. Derreumaux

OBJECTIFS

Historique


1900
2000
1865 (Mendel)
Théorie de l'hérédité
1944 (Avery)
ADN, support de l'information génétique
1953 (Watson & &Crick)
L 'ADN est une double hélice
1978(Dayhoff)
Matrices PAM
1951 (Sanger)
Séquençage de l'insuline
1970
Needleman & Wunsch
1970
Needleman & Wunsch
1980/1986
Création de l'EMBL (1980) ,de GenBank (1982) et de la DDBJ (1986).Création de SwissProt (1986)
2000/2001
Séquence brute du génome humain
1981 (Smith & Waterman)
Smith & Waterman
1990
BLAST
1990
Matrices BLOSUM

Bioinformatique

Deux définitions possibles

- ★ Applications de l'informatique à la biologie (en anglais: *computational biology*)
- ★ Analyse de l'information biologique (en anglais: *bioinformatics*)

C'est cette bioinformatique que nous abordons ici.

L'information est:

- ★ La séquence
- ★ La structure
- ★ La fonction, les interactions etc.

La déduction par homologie, ou le « dogme central » de la bioinformatique

★ Si la bioinformatique « marche », c'est parce que l'évolution des gènes laisse une trace parfaitement visible lorsque l'on compare leur séquence

- Les régions fonctionnelles des gènes (sites catalytique, de fixation, etc.) sont soumises à sélection. Elles sont relativement préservées par l'évolution car des mutations trop radicales sont désavantageuses.
- Les régions non fonctionnelles ne subissent aucune sélection et divergent rapidement à mesure que s'accumulent les mutations.
- Les nouveaux gènes apparaissent surtout par remaniement de gènes ancestraux: on peut donc déduire la fonction de la plupart des gènes par comparaison avec les gènes « homologues » d'autres espèces.
- (Evolution des gènes=mutations, insertions, délétions, recombinaisons)

La Bioinfo

Pourquoi faire ?

a. annotation syntaxique

b. annotation fonctionnelle

c. Evolution moleculaire ↔ Génomique Evolutive.

d. Protein engineering and protein design.

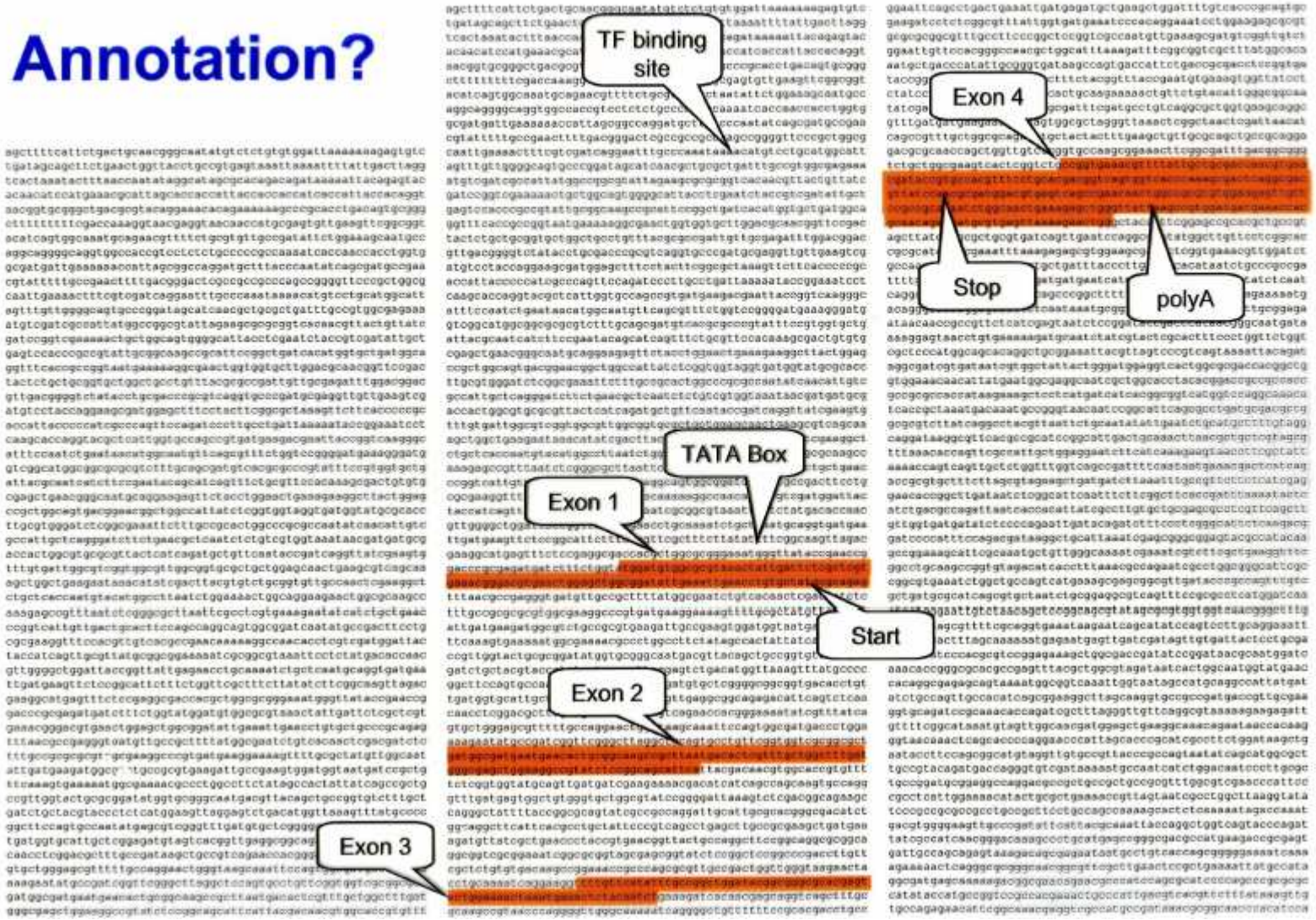
e. Pred 3D Protein structure from seq.

f. Structure - Dynamics - Function

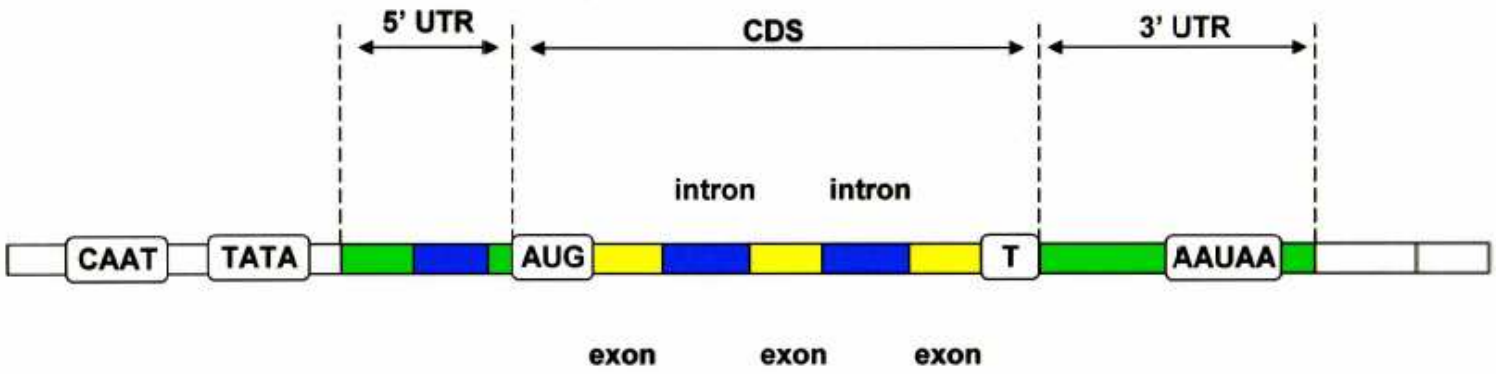
g. Drug Design.

Bioinformatics: Definitions.

- **Annotation?**



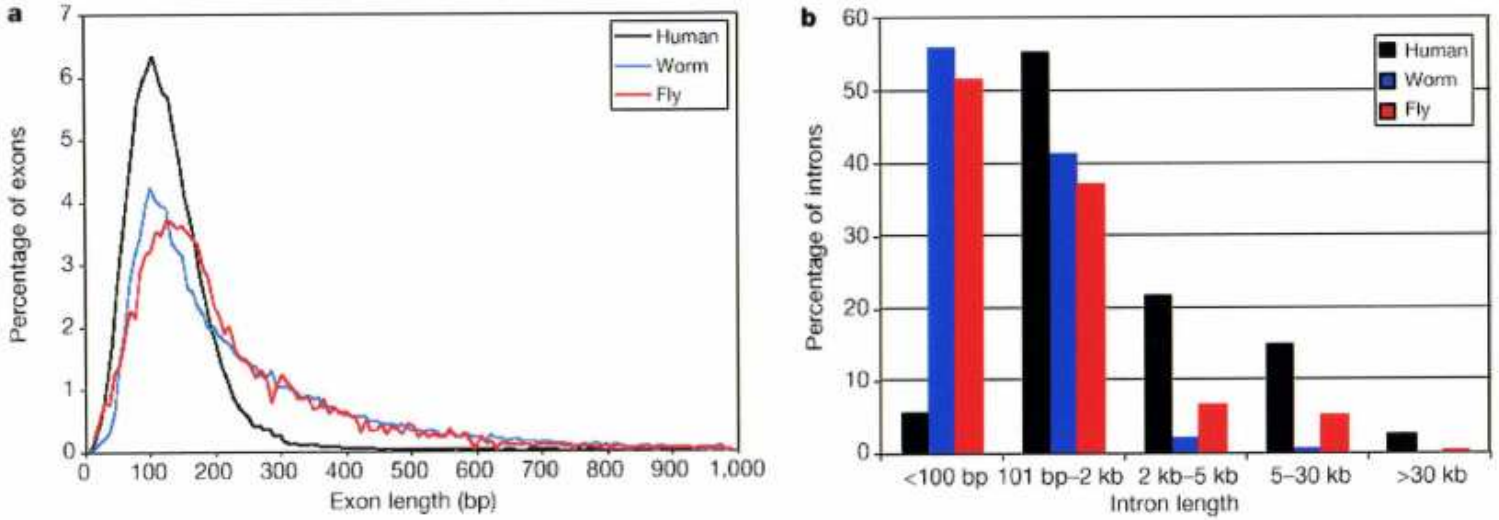
Structure «classique» d'un gène codant une protéine



Mais un gène « classique », cela n'existe pas….

Distribution des tailles d'exons et d'introns

chez H. sapiens, D. melanogaster, C. elegans



International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature (2001) 409: 745-964

Some facts about the human genome

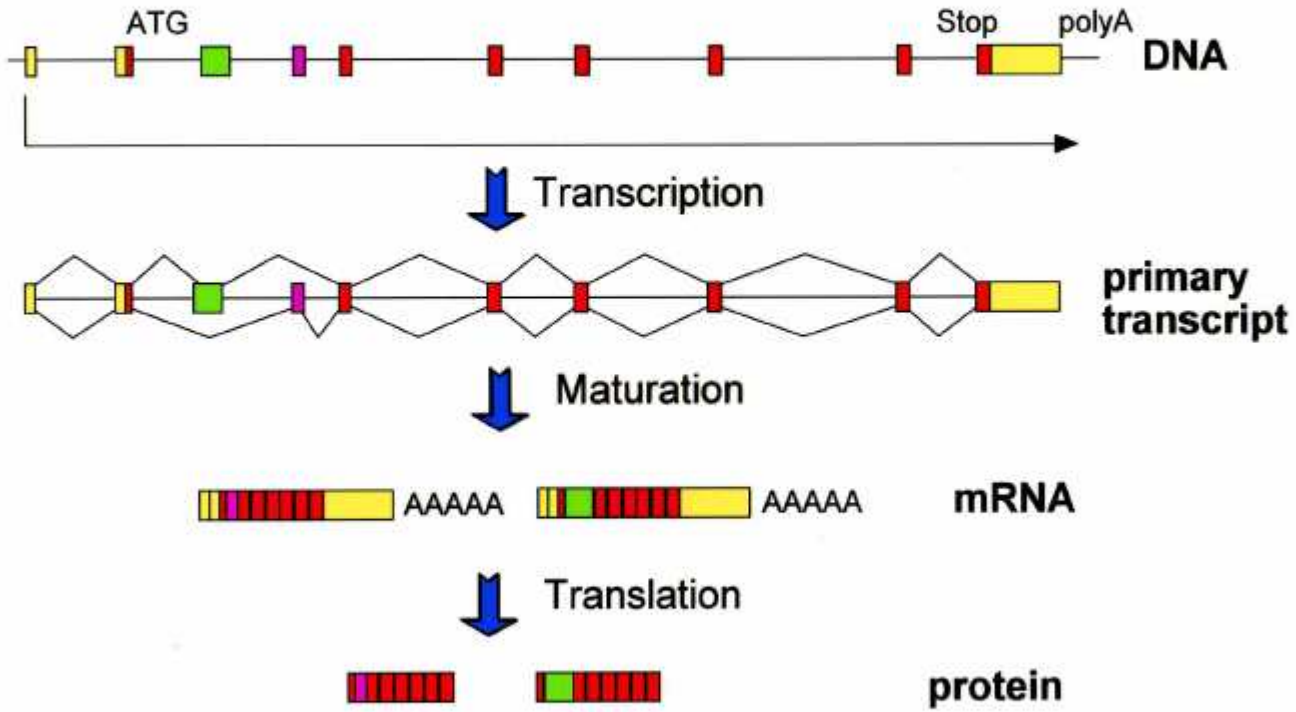
- 3.2×10^9 bp
- Genes comprise about 3% of the genome
- Average gene length: ~ 8,000 bp
- Average of 5-6 exons/gene
- Average exon length: ~200 bp
- Average intron length: ~2,000 bp
- ~8% genes have a single exon

- Extremes:
  - **Factor VIII gene** (whose mutations cause hemophilia A)
    - spread over ~186,000 bp (~9 kb of exons and ~177 kb of introns.)
    - 26 exons (size range 69 to 3,106 bp)
    - 25 introns (size range 207 to 32,400 bp)
  - **Dystrophin** - the biggest human gene yet
    - >30 exons, spread over 2.4 million bp.

Dystrophin is a large, rod-like cytoskeletal protein which is found at the inner surface of muscle fibers. Dystrophin is part of the dystrophin-glycoprotein complex (DGC), which bridges the inner cytoskeleton (F-actin) and the extracellular matrix.

Un gène, plusieurs produits

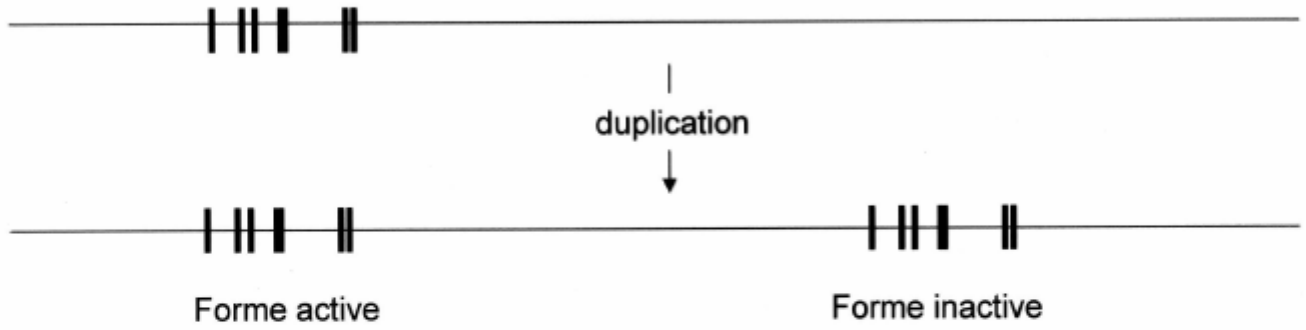
- Epissage alternatif dans plus de 30% des gènes humains (Hanke et al. 1999)



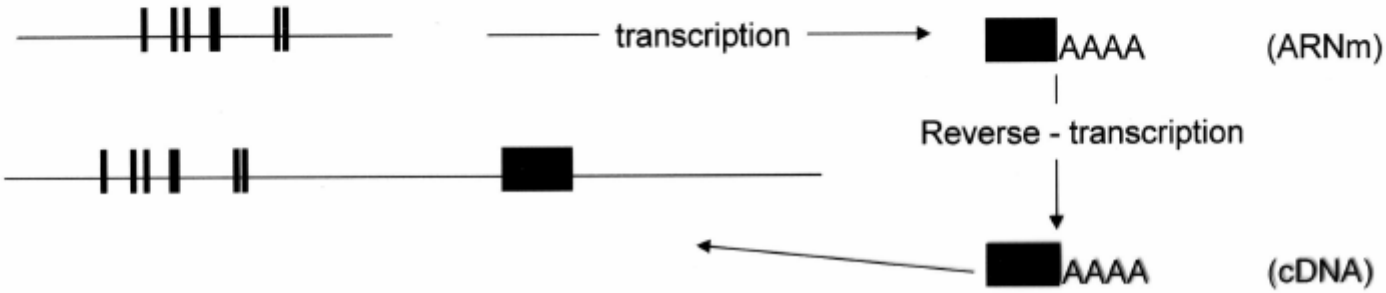
- Promoteurs alternatifs
- Signaux polyadenylations alternatifs

- Les pseudogènes

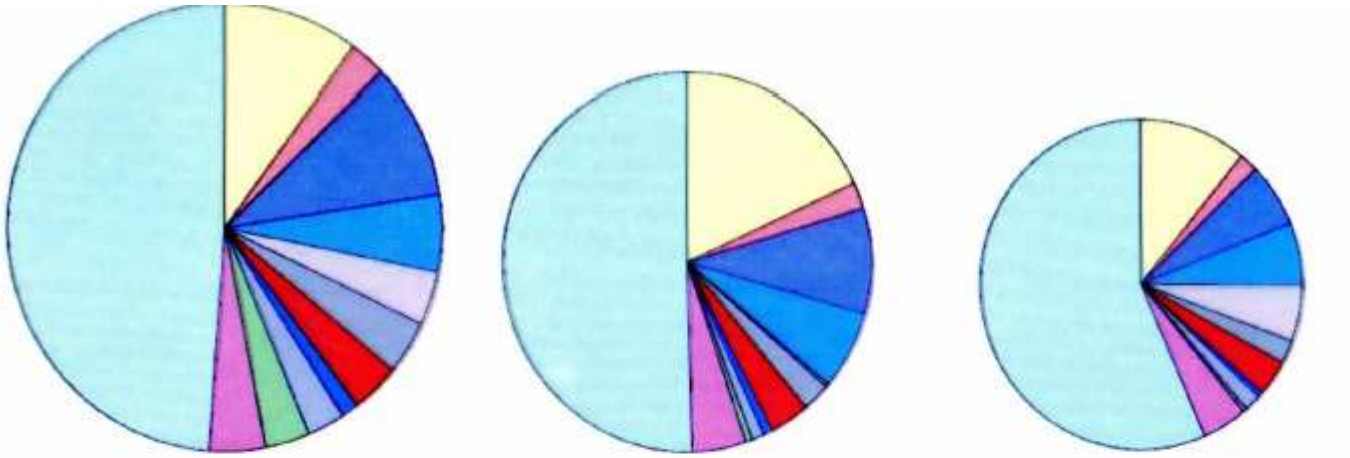
1. Résultent de duplication de gènes existant



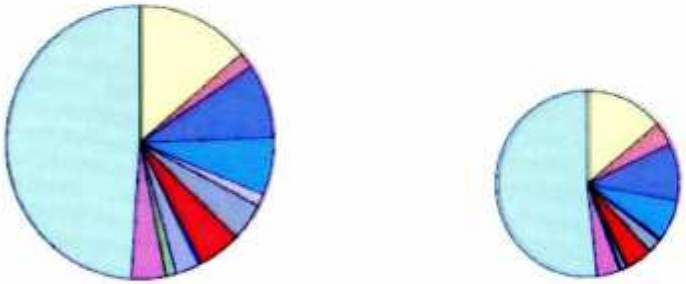
2. Résultent de la rétro-transposition d'un ARNm



- Des gènes dans les gènes



Organism	Human	*Arabidopsis* (plant)	*C. elegans* (roundworm)
Genes	~32,000	25,706	18,266

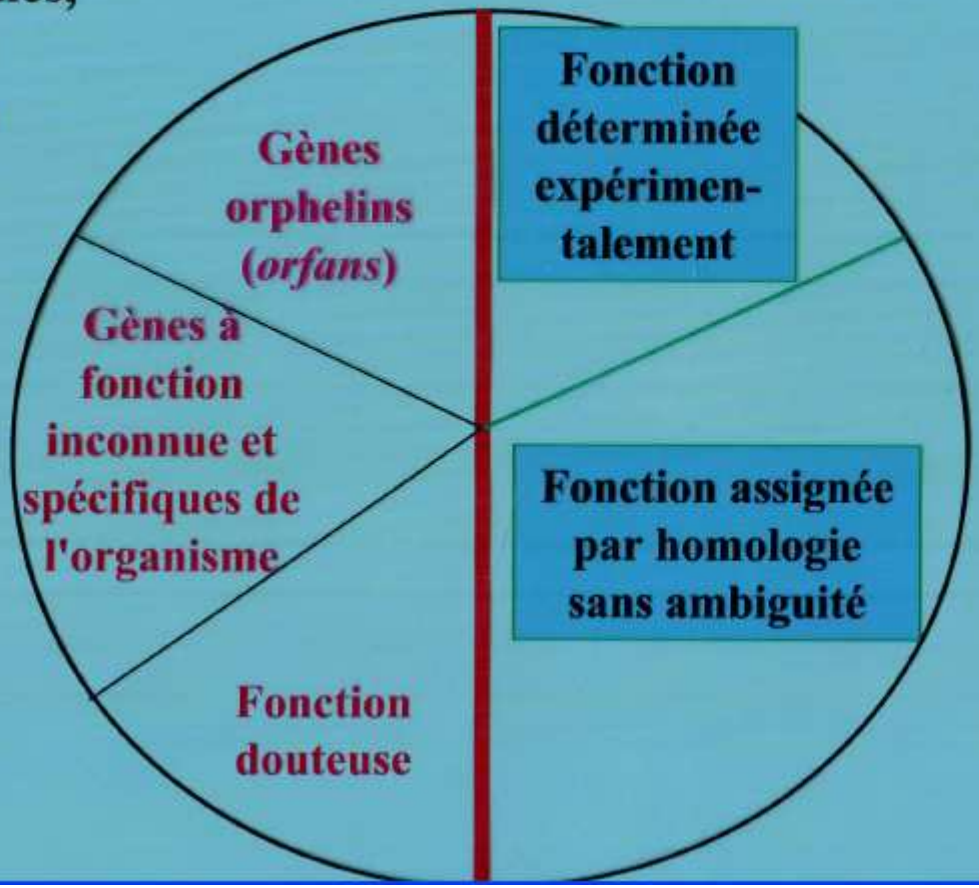


Organism	*Drosophila* (fly)	*Saccharomyces* (yeast)
Genes	13,338	~6000

- Metabolism
- DNA replication/modification
- Transcription/translation
- Intracellular signaling
- Cell-cell communication
- Protein folding and degradation
- Transport
- Multifunctional proteins
- Cytoskeleton/structure
- Defense and immunity
- Miscellaneous function
- Unknown

La surprise des gènes orphelins

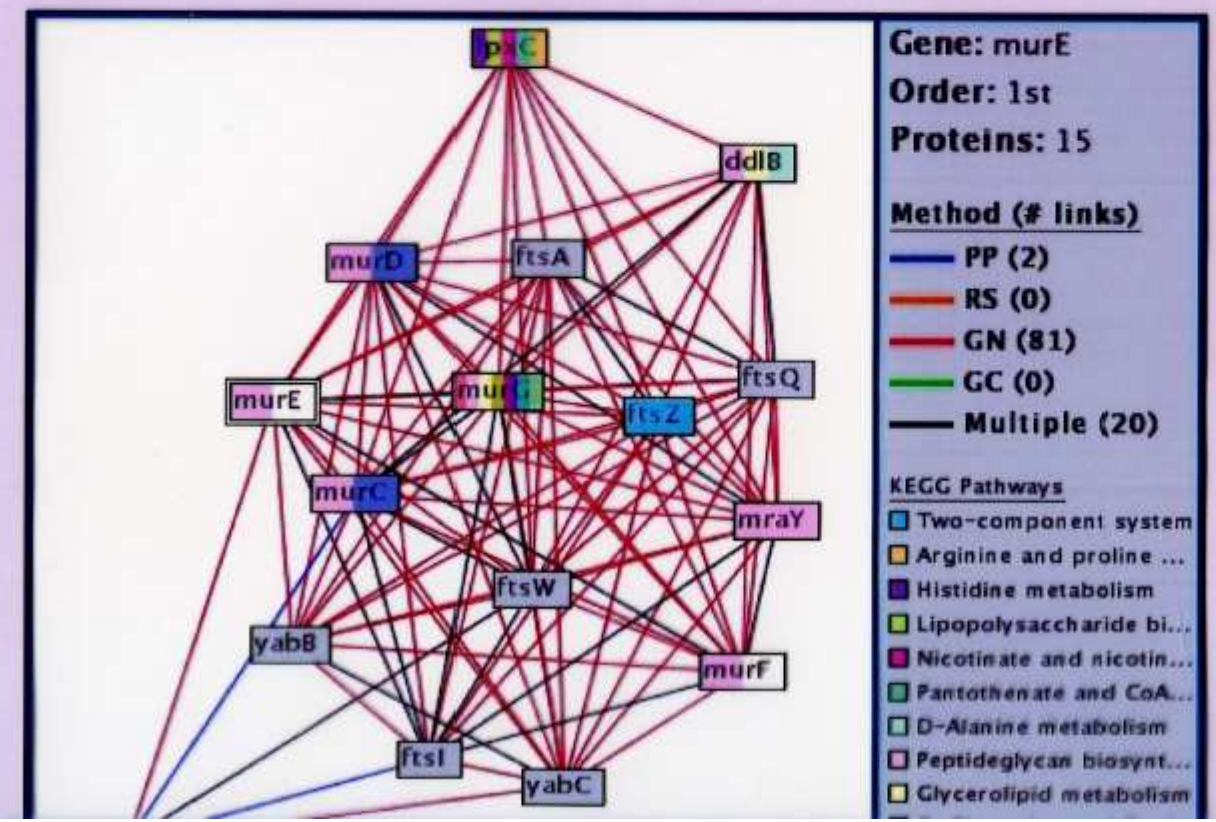
Pour la majorité des génomes, on a une répartition 50/50 entre le connu et l'inconnu



• Fundion

Minimum Confidence:	0.6	**Links to Display:**	1st Order
Analysis Methods:	☑ Phylogenetic Profile (PP) ☑ Rosetta Stone (RS) ☑ Gene Neighbor (GN) ☑ Gene Cluster (GC) ☐ TextLinks (TL)	**Color Nodes by:**	KEGG Pathway
		Graph Size:	Small

Reset | Update Graph



Homo sapiens

→ GENE EXPRESSION. transcriptome, PUCES à ADN

Génomique Evolutive

Deux grandes approches nécessaires et complémentaires

Étude de l'évolution de chacun des gènes d'un génome

+

Étude de l'évolution d'un génome en tant que tel

- basée sur le concept d'homologie
- analyse des évènements de duplication, fusion, fission de gènes
- structure/fonction des protéines

- gain et perte de gènes
- synténie
- contexte génétique

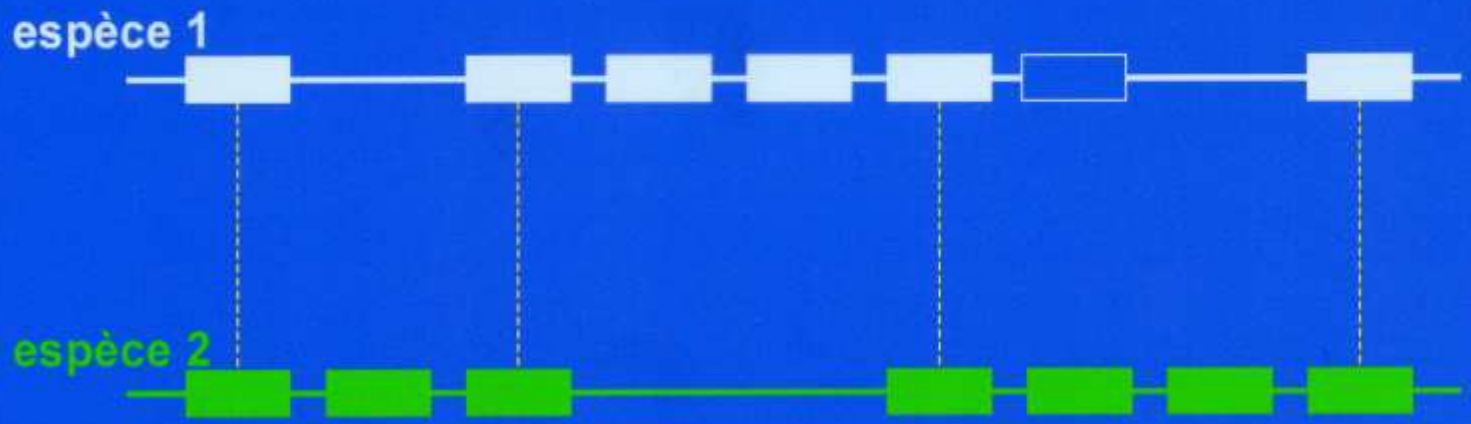
- **Extension de la phylogénie moléculaire**
- **Nouvelles possibilités d'analyses**

Mise en évidence des forces motrices sous-jacentes aux modes d'évolution des génomes

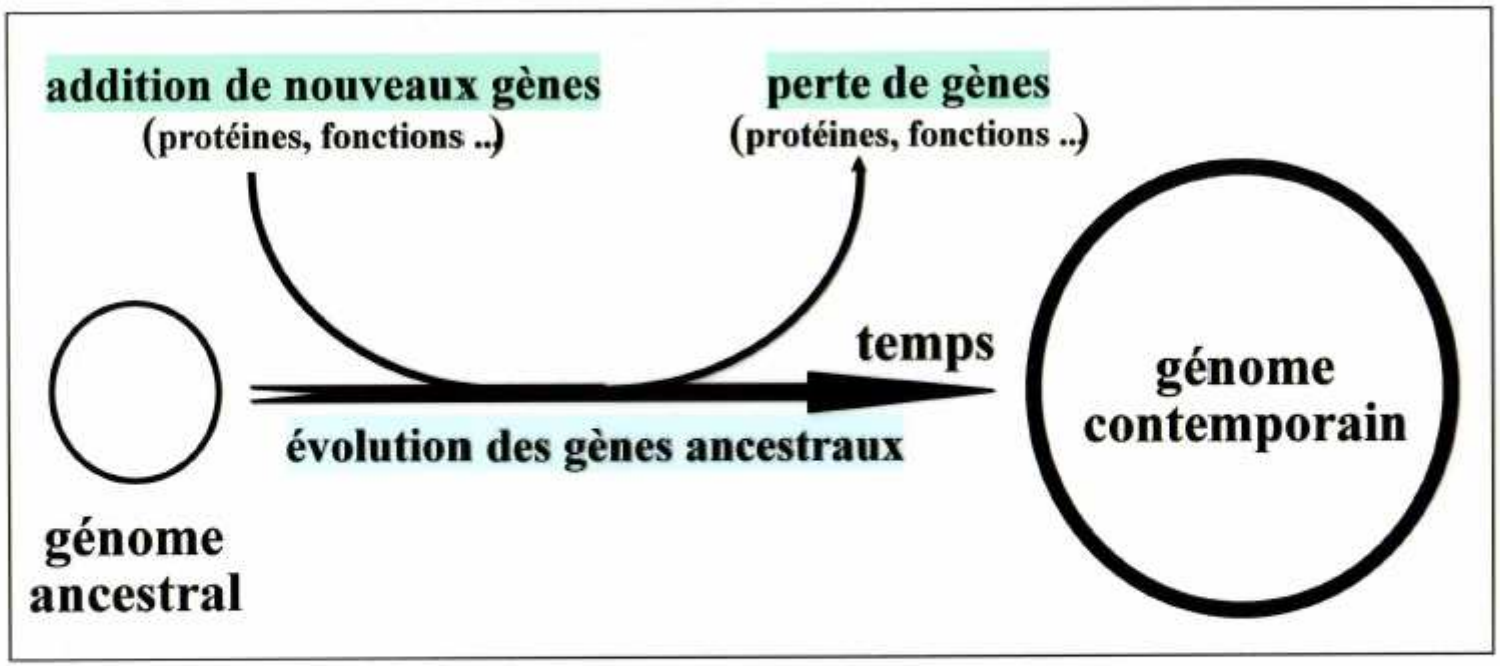
Synténie

Synténie : conservation de l'ordre des gènes entre deux génomes.

- mesure du degré de synténie
- définition de zones de synténie : zones floues



Évolution des génomes



- **Différences qualitatives**
  - → changements dans la nature des gènes
- **Différences quantitatives**
  - → variations du nombre de gènes

Évolution des génomes bactériens

Élimination progressive des gènes « superflus »

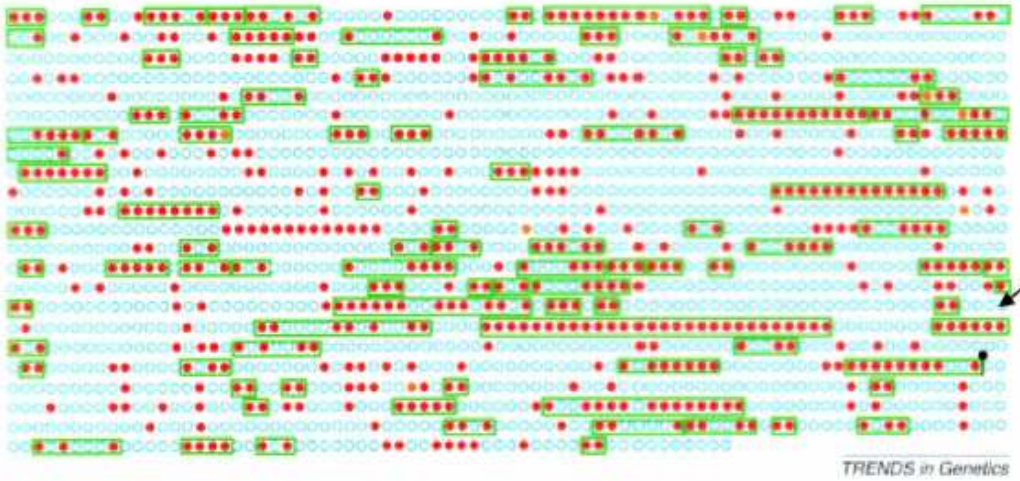


Fig. 2. The genome of the common ancestor of *E. coli* and *Buchnera*. Each circle represents a gene. Red, genes conserved in *Buchnera*; blue, genes not present in the *Buchnera* genome; orange, pseudogenes in *Buchnera*. Genes are ordered according to the *E. coli* K-12 genome, and the first circle corresponds to *E. coli* b0002 (*thrA*). The origin of replication is marked with a black circle. Ancestral `blocks' (see text) are shown as green rectangles. This genome includes those *E. coli* genes with an ortholog in either *V. cholerae* or *Buchnera* APS.

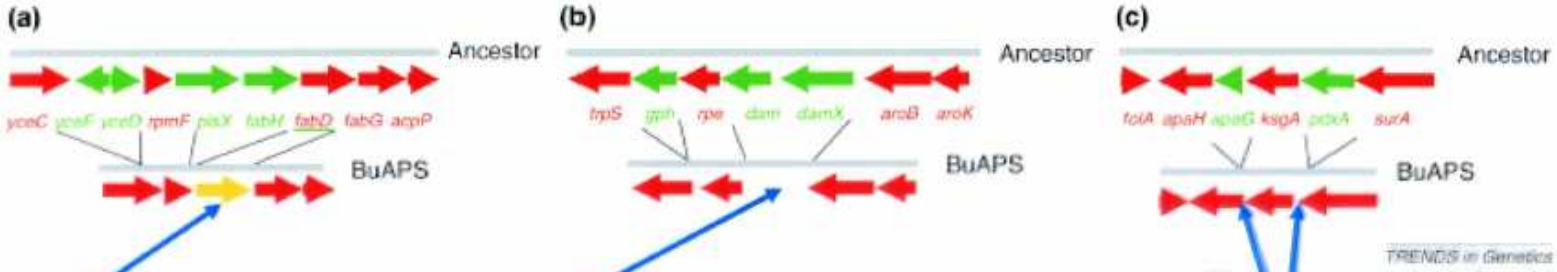


Fig. 3. Mechanism of gene desintegration, with examples from the *Buchnera* APS (BuAPS) genome. (a) The gene is inactivated and forms a pseudogene (e.g. *fabD*). (b) The DNA sequence is gradually eroded by continual small deletions and other mutations until it is no longer recognizable as a pseudogene (e.g. the remnant sequence corresponding to the *dam* and *damX* genes). (c) Complete deletion of intervening genes. The `ancestor' shows the structure conserved in *E. coli* and *V. cholerae*. Green, genes lost from *Buchnera*; red, functional genes in *Buchnera*; yellow, pseudogenes in *Buchnera*..

Prediction of protein 3D structure

sequence

KELVLVLYDY QEKSPRELTI
KKGDILTLLN STNKDWWKVE
VNDRQGFIPA AYLKKLD

No similar sequence is identified

Similar sequence with Known 3D structure is identified

Similar sequence(s) found, but no info on 3D structure

Fold recognition

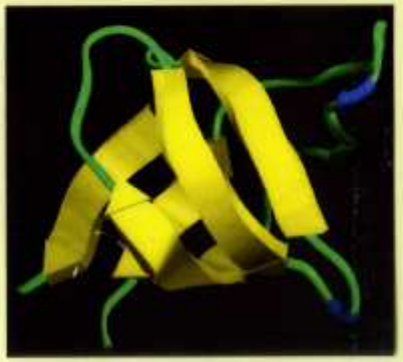
Homology modelling

no

yes

?

?



3D structure

→ amyloid fibrils

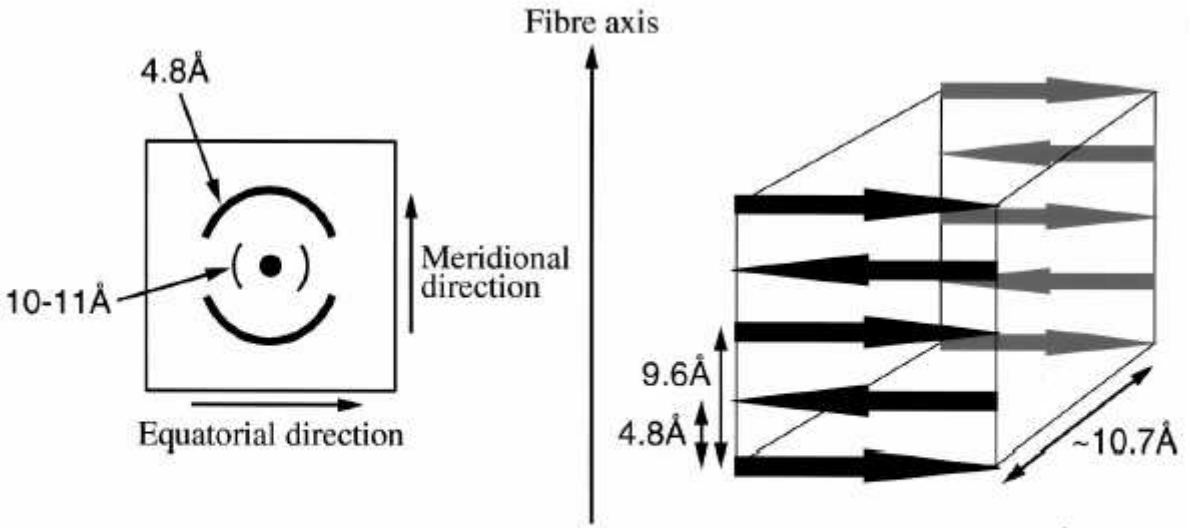
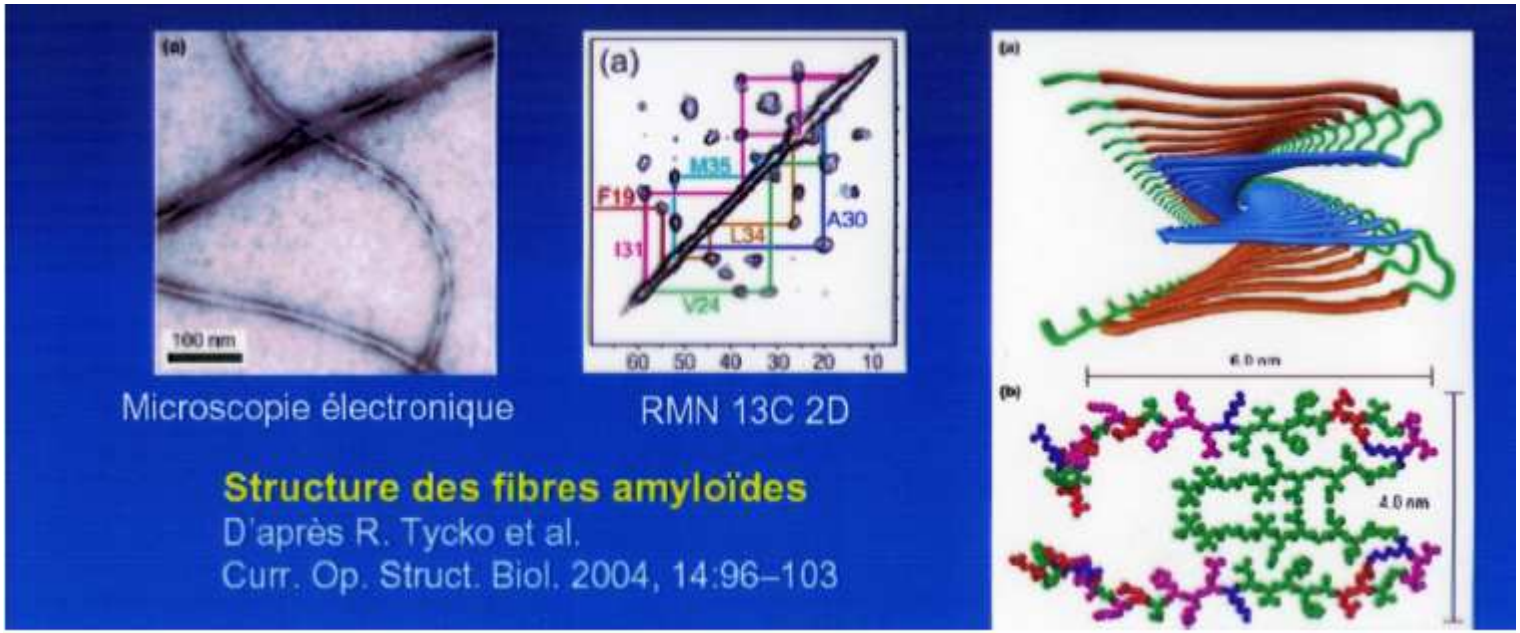
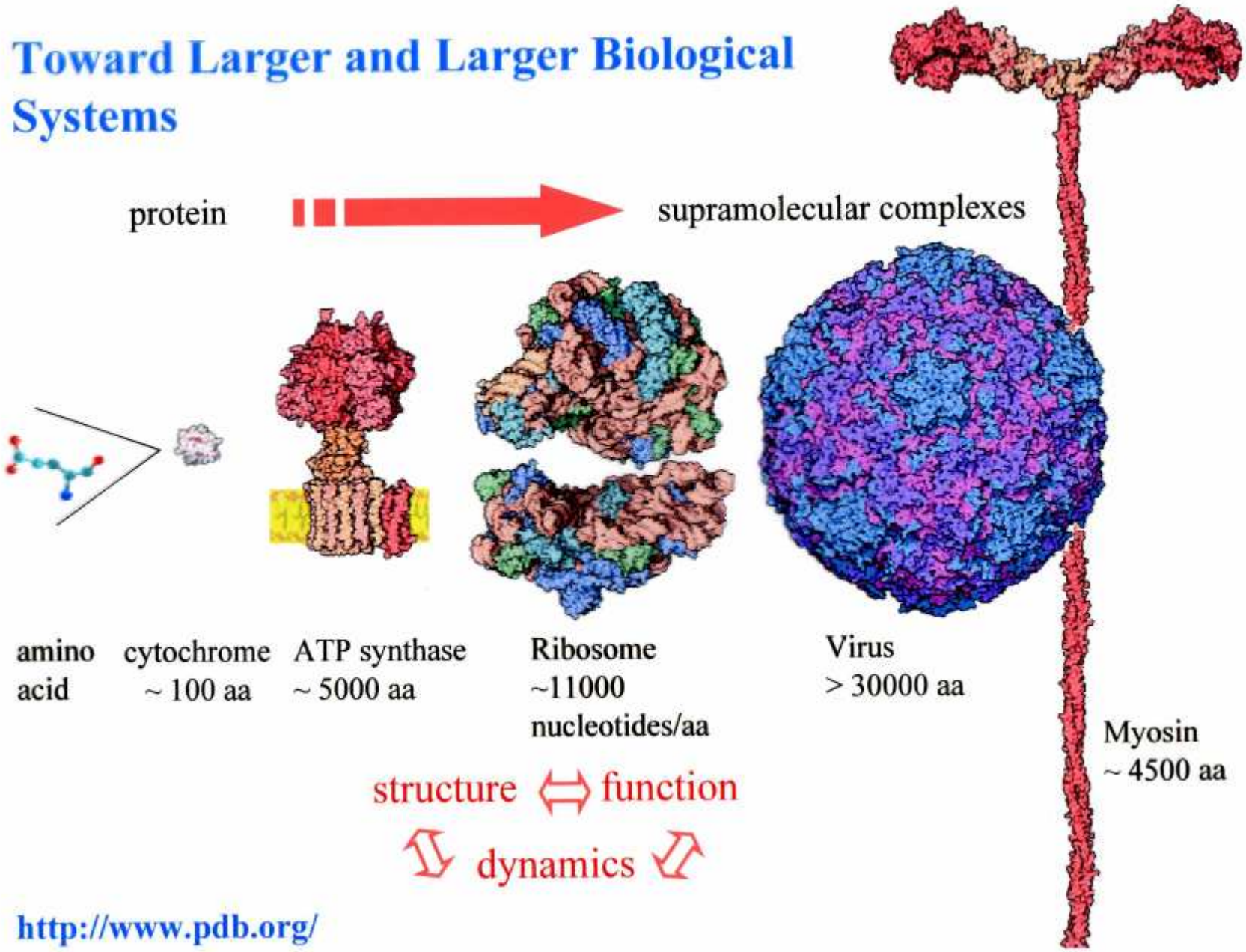


Fig. This shows the characteristic cross-β spacings from X-ray fibre diffraction from amyloid fibrils. A strong 4.8 Å reflection on the meridian corresponds to the hydrogen bonding distance between β-strands (shown right), and a more diffuse 10–11 Å reflection on the equator shows the intersheet distance of about 10.7 Å. A spacing of 9.6 Å would correspond to the repeat distance for an anti-parallel arrangement of β-strands. This is not always observed in poorly aligned diffraction patterns, but often in well aligned pattern.



Toward Larger and Larger Biological Systems



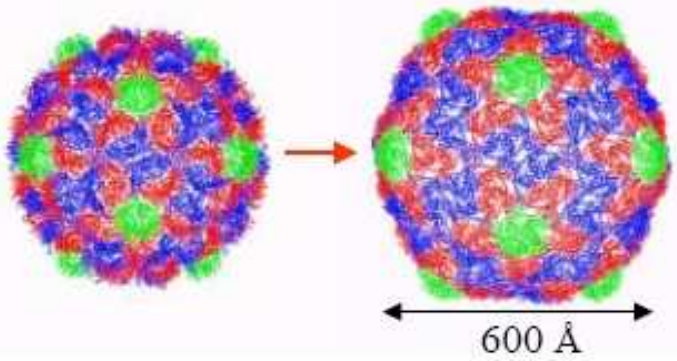
http://www.pdb.org/

Large-scale conformational changes

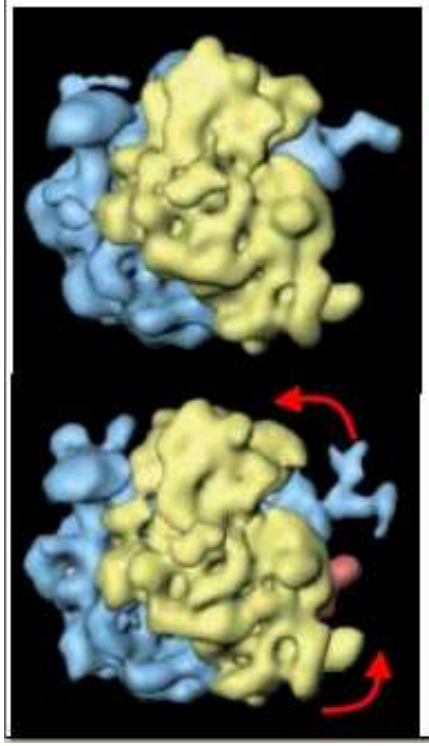
X-ray crystallography ⇨ atomic level

Cryo electron microscopy (cryo-EM) ⇨ low-resolution (from 7Å), shape information

Virus maturation

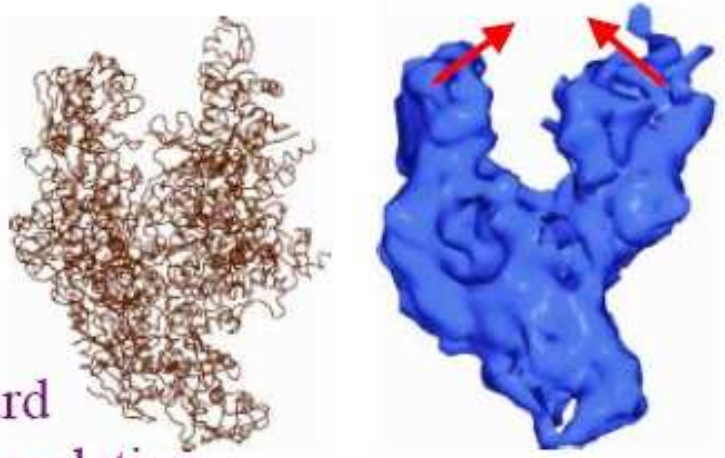


Protein synthesis: ribosome



Functional motions

Transcription - Replication



Time scale (> μs) not accessible from standard molecular dynamics simulation

RNA polymerase

Protein engineering and protein design.

Protein engineering – altering protein sequence to change protein function or structure

Protein design – designing de novo protein which satisfies a given requirement

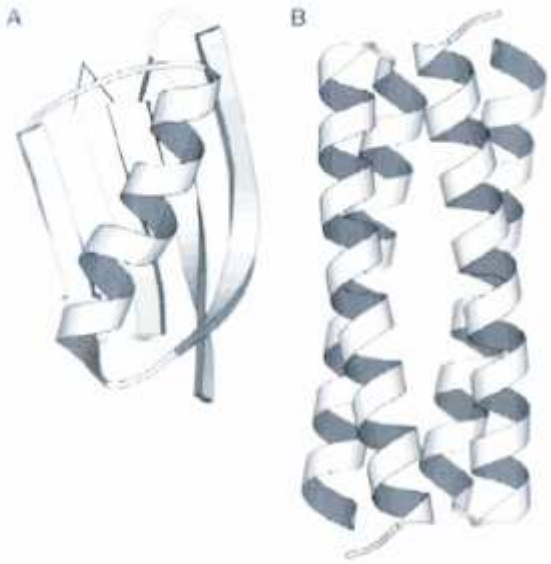
Protein engineering strategies

Goals:

- Design proteins with certain function
- Increase activity of enzymes
- Increase binding affinity and specificity of proteins
- Increase protein stability
- Design proteins which bind novel ligands

Paracelsus challenge: convert one fold into another by changing 50% of residues.

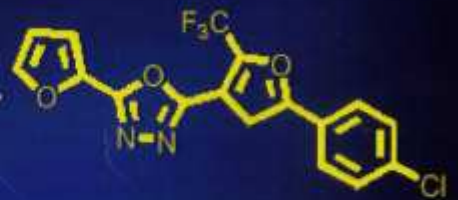
- Challenge because all proteins with > 30% identity seem to have the same fold.
- L.Regan et al: Protein G (mainly beta-sheet) was converted to Rop protein (alpha-helical) by changing only 50% residues



Target-structure based methods

Protein-Based Design

?



Docking (rigid, flexible)

2. de novo Building

Scoring

IC_{50}, K_i ???

Les banques de données de séquences biologiques : laquelle choisir ?

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BioImage, BioMagResBank, BIOMDB, BLOCKS,BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD40Lbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, CbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genline, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HeXAdb, HGMD, HIDB, HIDC, HlVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 0-lycBase, ONIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene,Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, YEPD, YPD, YPM, etc ...

ET DEVELOPPEMENT METHODES THEORIQUES