Université Frères MENTOURI Constantine 1

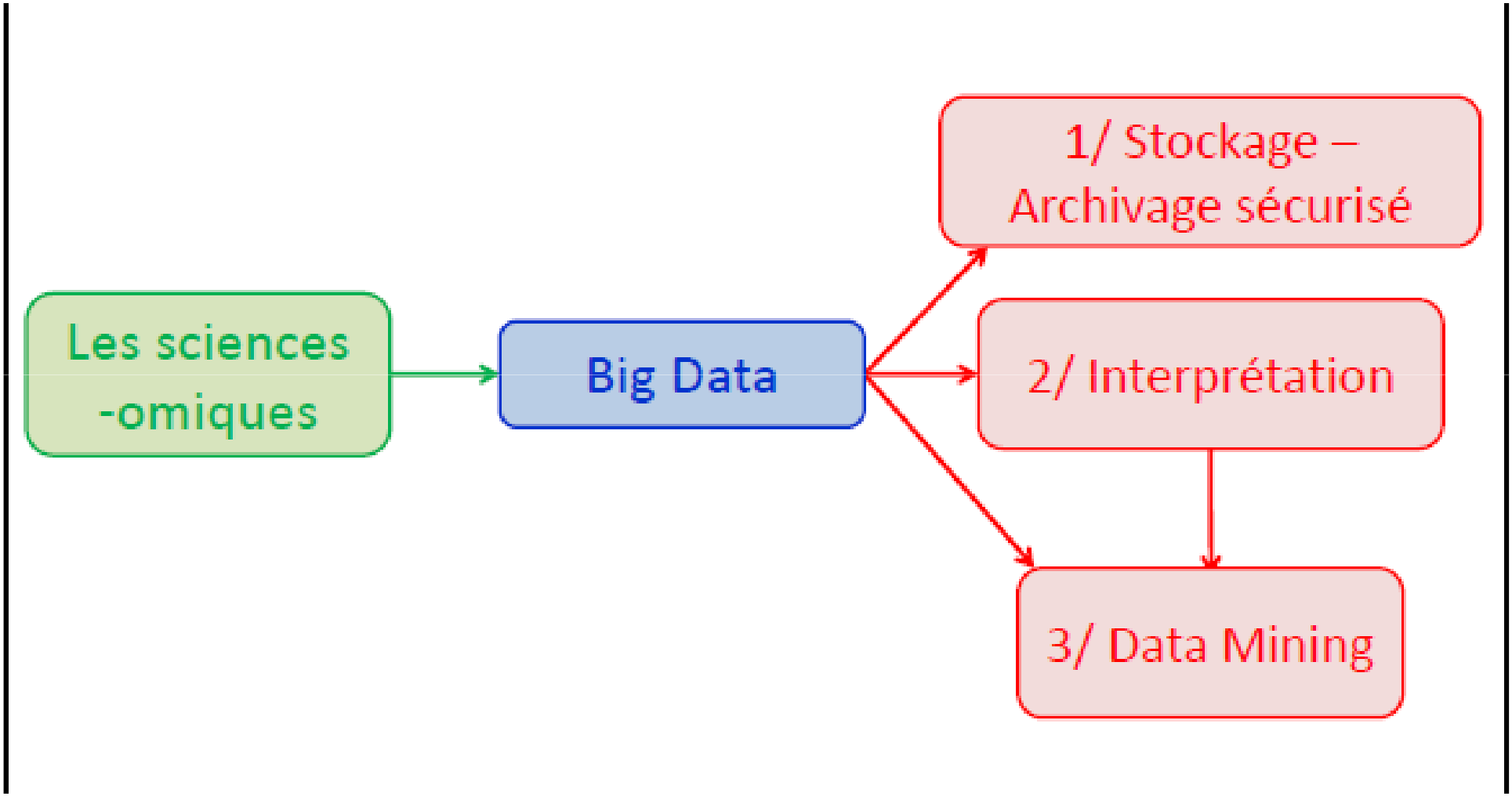
Faculté des Sciences de la Nature et de la Vie
Département de Biologie & Ecologie Végétale

Master en : Biotechnologie & Génomique Végétale

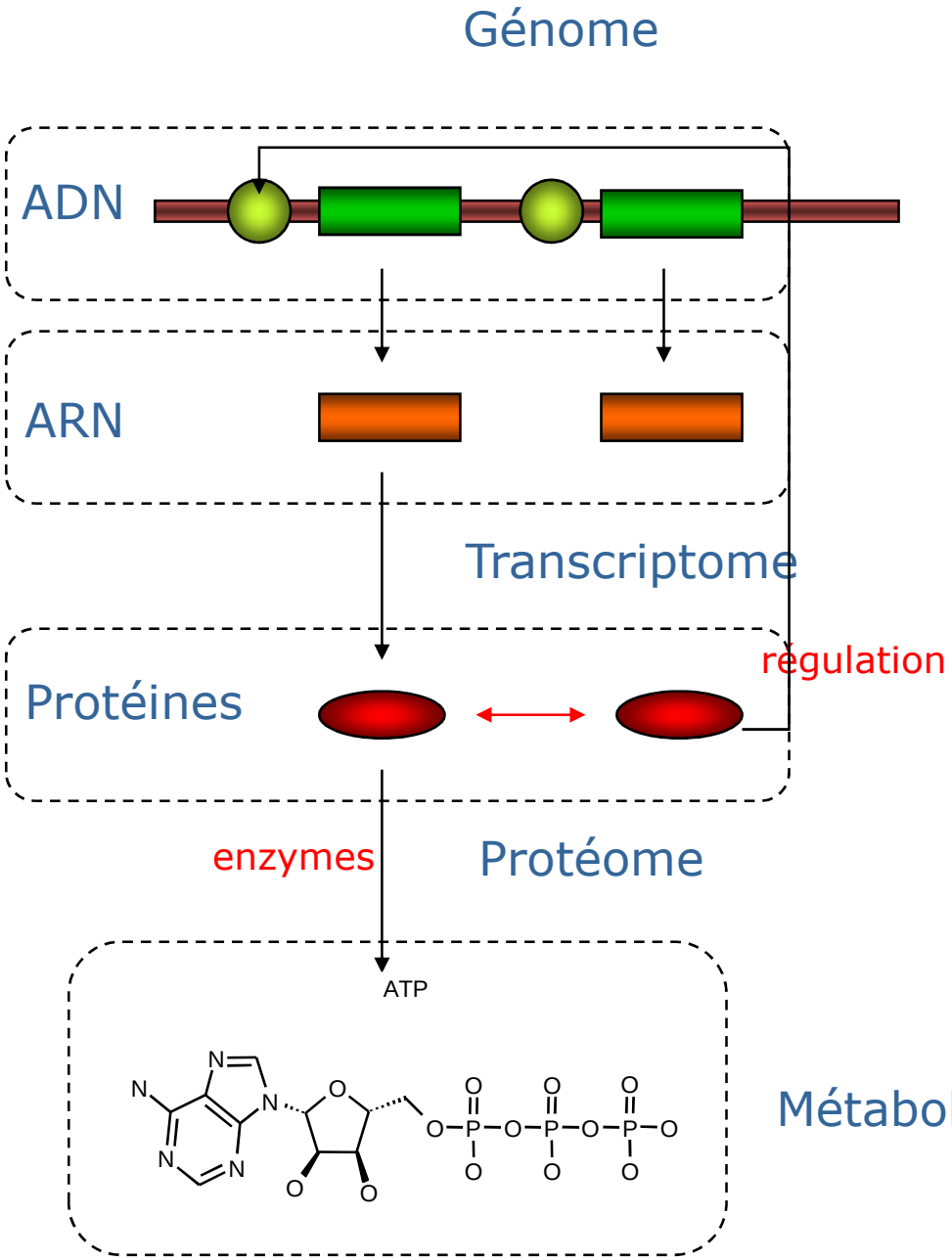
Cours de : Génomique Fonctionnelle

ANALYSE DES GÉNOMES

La problématique des -omiques



La problématique des -omiques



ADN double brin

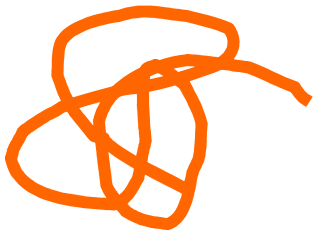
transcription

messager

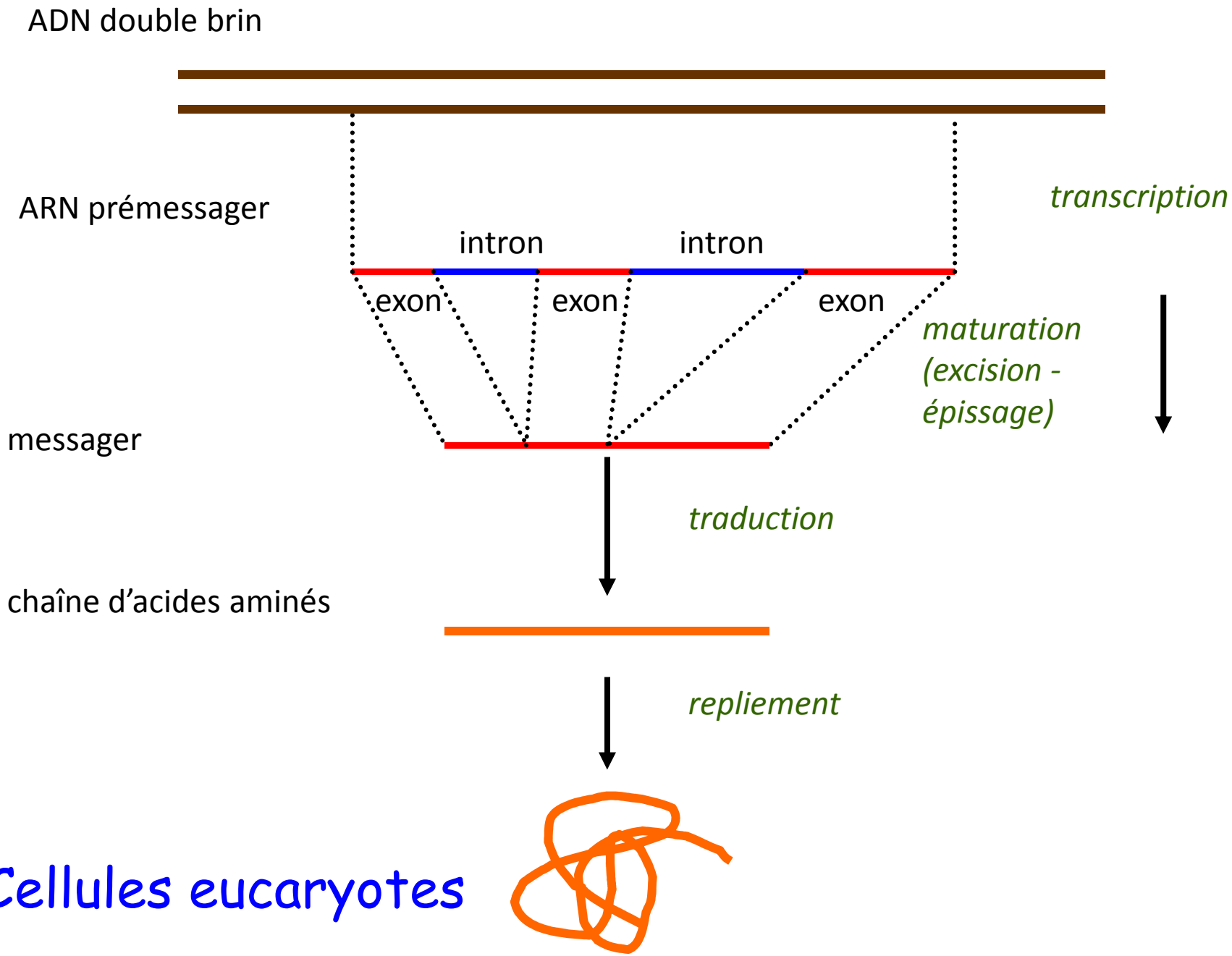
traduction

chaîne d'acides aminés

repliement



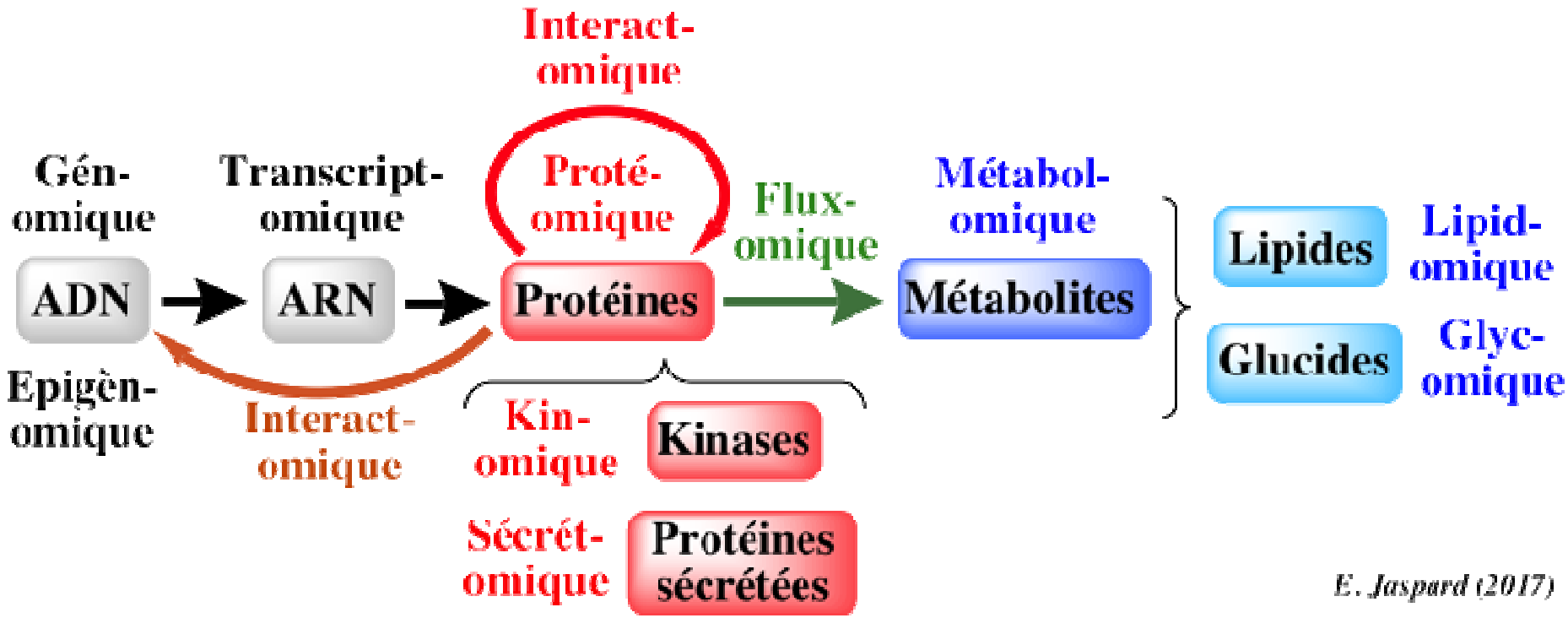
Cellules procaryotes


ADN double brin
ARN prémessager
transcription
intron
intron
exon
exon
exon
maturation
(excision -
épissage)
messager
traduction
chaîne d'acides aminés
repliement
Cellules eucaryotes

La problématique des -omiques

Métabolomique – Complexomique – Interactomique – Métagénomique – Métaprotéomique – Protéogénomique, …

La problématique des -omiques

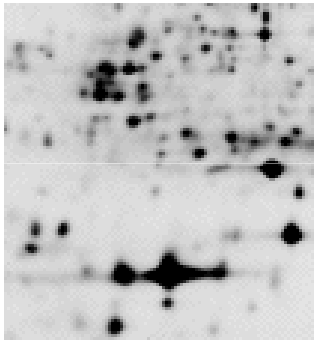


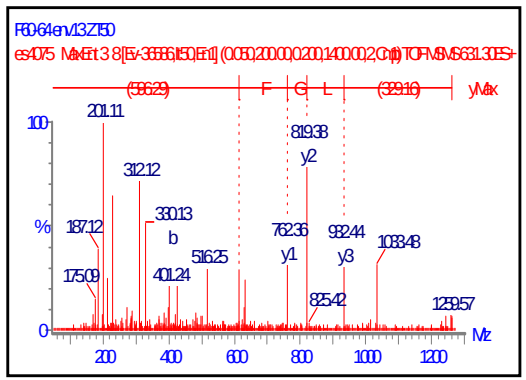
Deux types de molécules support de la bioinformation : les acides nucléiques et les protéines

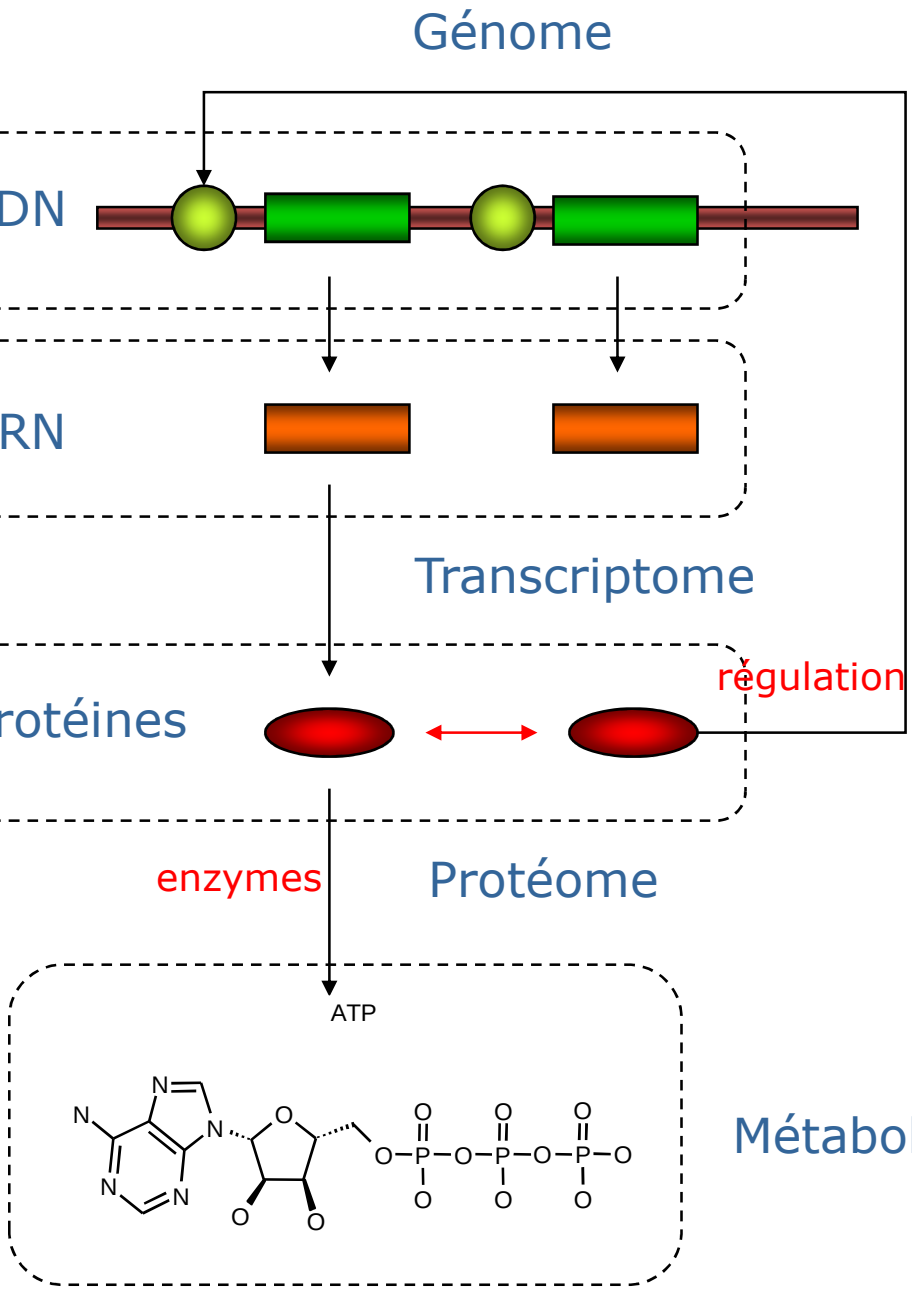
Le "**matériau de base**" de la *génomique*, de la transcriptomique et de la protéomique est la séquence : l'enchaînement ordonné et orienté de nucléotides (acides nucléiques) ou d'acides aminés (protéines).

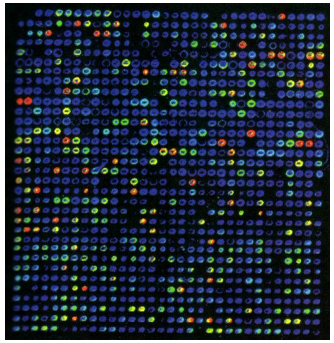
La problématique des -omiques

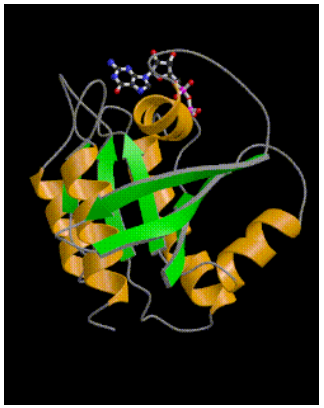
GATCACCTCACTACGG
GTCAGGGGAAGGAAA
GGGGAACTGAGAGATT
TGTCAGTGTGAGAAGC
AGTCCCAGGAGTTAGA
AGTAGTGGCTCCATGA
CTCACAAATTAACTTC
CCTTTCAGGCAGGGCT
TCTTATTTTCCTTAGCA
TCCCTGTCTTGATCCCA
GCCTGCTCAGACCCCT
GCCTCTCACTGCAAGA
TGTGCTT









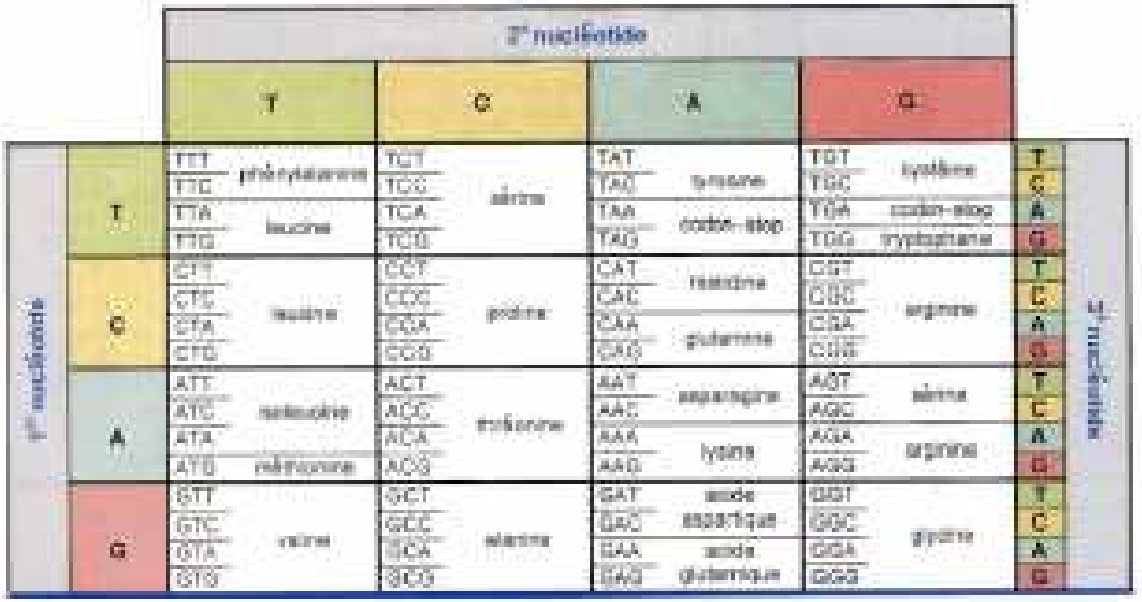


La problématique des -omiques

Code à 4 lettres
A, T, G, C

Code à 20 lettres
20 acides aminés

Le code génétique

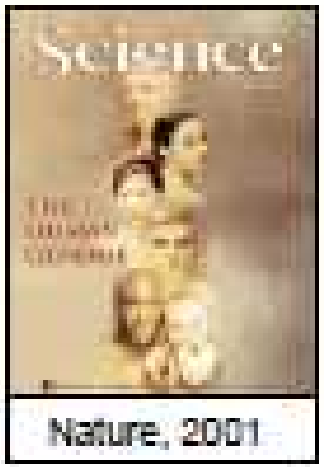


La problématique des -omiques

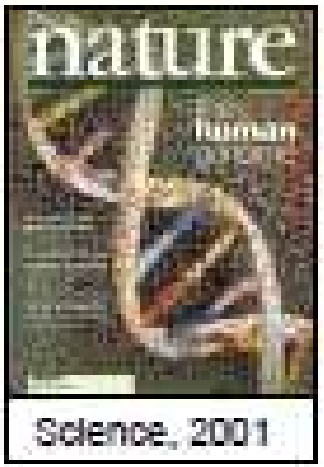
Premier génome séquencé en **1995**:
Haemophilus influenza (Taille $1{,}8.10^6$ bps)

⇒ Génome de la levure en **1996**:
Saccharomyces cerevisiae (Taille 14.10^6 bps)

⇒ Premier draft du génome humain **2001**:
Homo sapiens (Taille $3{,}2.10^9$ bps)



Nature, 2001

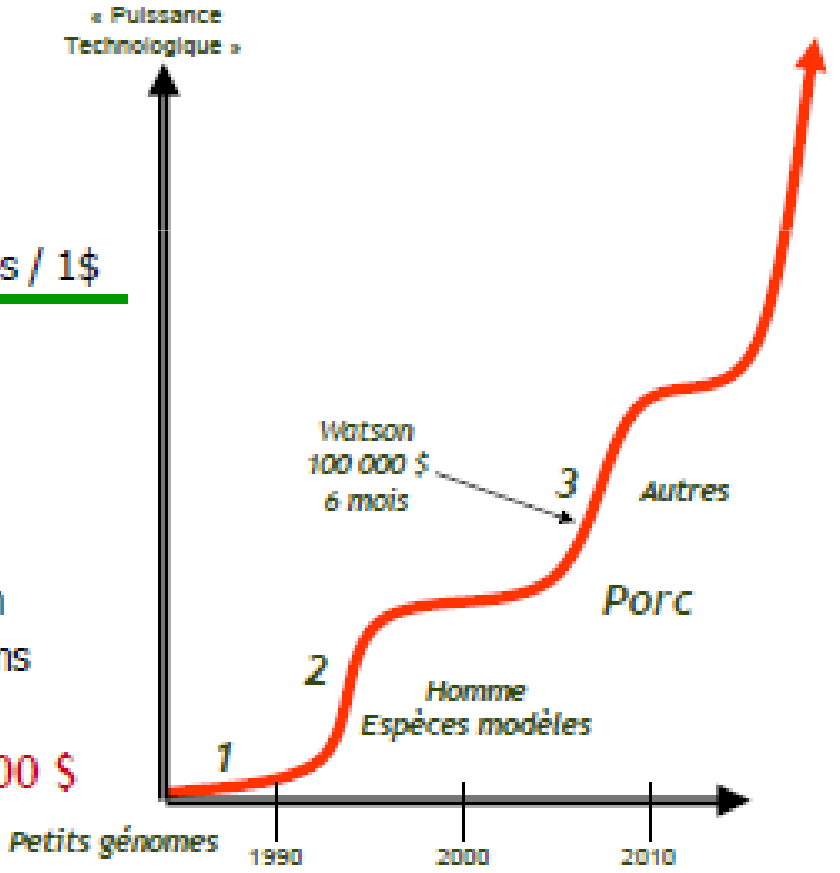


Science, 2001

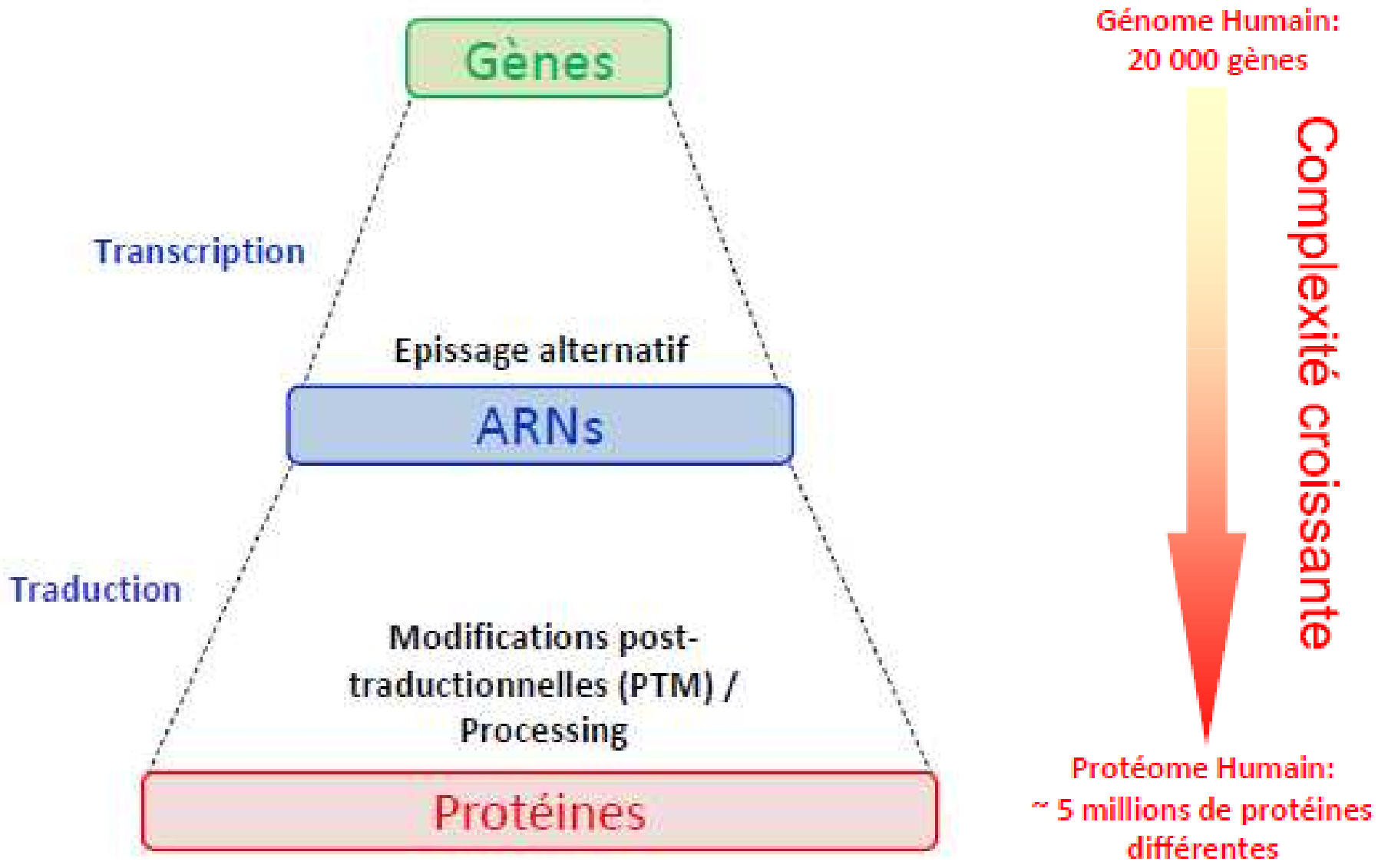
La problématique des -omiques

- Séquençage :

	Bases/expé	Nbre de bases / 1$
< 1990	1200	12
1990-2000	77 000	300
2000-2010	400 000 000	40 000
	10 500 000 000	1 million
	200 000 000 000	6,5 millions
> 2010	1 génome en 15 min pour 1 000 $	



La problématique des -omiques : une complexité croissante



Le codage de l'information génomique

- macromolécule d'ADN ≈ enchaînement d'acides nucléiques
  - adénine : *A*
  - thymine : *T*
  - cytosine : *C*
  - guanine : *G*
- génome ≈ texte écrit dans l'alphabet de ces quatre lettres

Organisation générale des génomes

Organisés en chromosomes

Chaque chromosome eucaryote contient un ADN linéaire

Répartition des gènes, variable le long des chromosome et région intergénique non codante

Gènes:

séquences codantes continues: cas général chez les procaryote (mais pas exclusif) ou discontinues (exon-intron): cas général chez les eucaryotes (mais très variable d'un organisme à l'autre, rare chez la levure)

Séquences répétées

Dans tous les génomes mais % très varié qui peut être très élevé

Analyses bioinformatique des génomes

Reconnaissance de gènes et autres éléments du génome

Syntaxe des séquences

Recherche de similarité

La bioinformatique permet d'extraire l'information des séquences génomiques.

Le génome comme un langage :

- Support = polymère linéaire
- Alphabet = molécules
- Mots = 3 lettres parmi 4
- Syntaxe = en cours de déchiffrage

Superposition de signaux :

Pas seulement quel est le produit du gène, mais aussi où est-il exprimé, en quelle quantité et quand ???

La nature du contenu informationnel de l'ADN

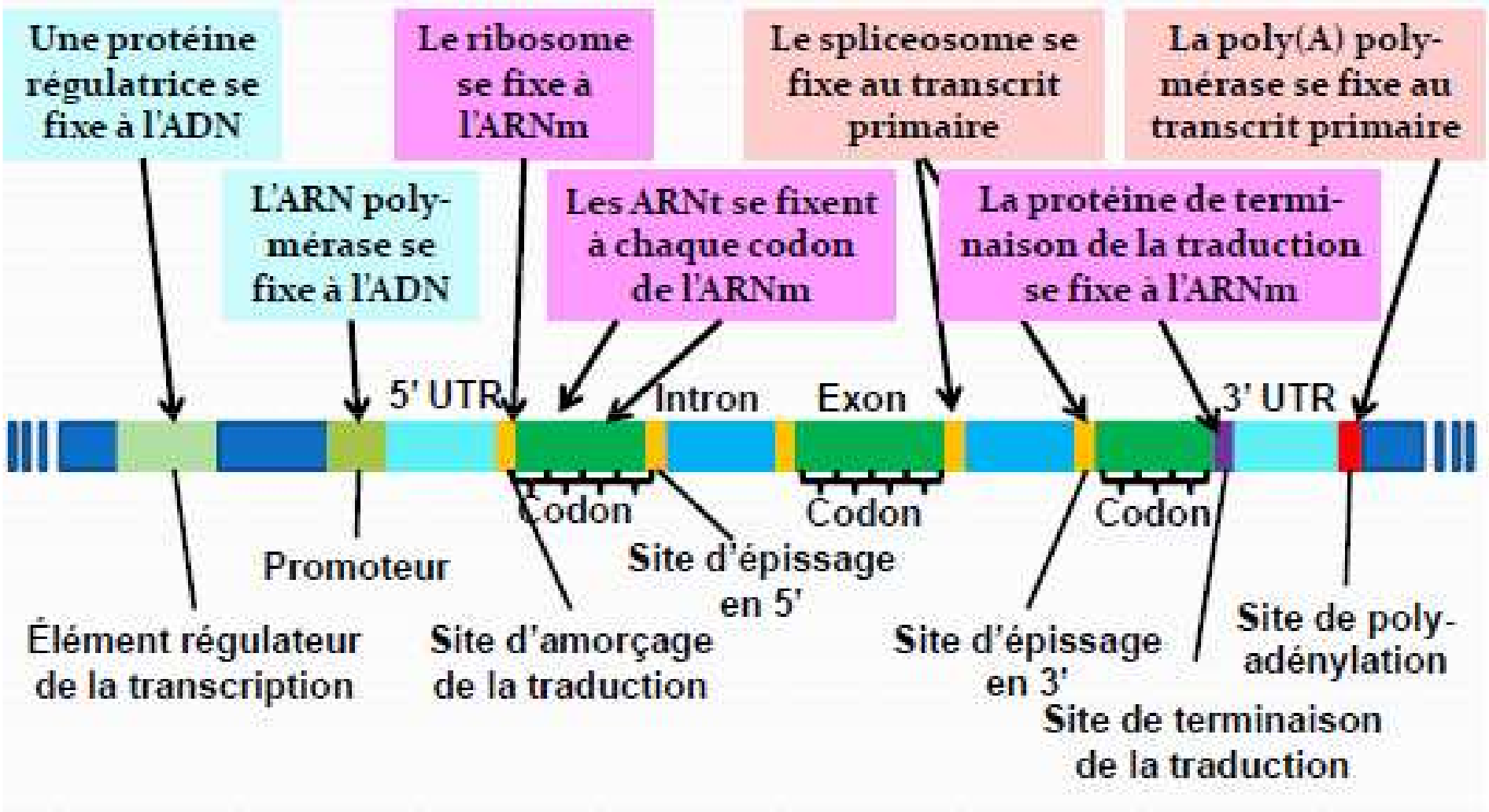
Par convention, on considère que l'information est la somme de tous les produits de gènes, c.à.d. des protéines et des ARN.

* L'ADN contient l'information, mais de quelle façon celle-ci est-elle codée ?

* De nombreuses protéines se fixent au niveau de sites présents sur l'ADN lui-même, tandis que d'autres protéines et des ARN se lient à des sites présent sur l'ARNm.

* La séquence et les positions relatives de ces sites permettent aux gènes d'être transcrits, épissés et traduits correctement au moment adéquat et dans le tissu approprié.

La nature du contenu informationnel de l'ADN



Le contenu informationnel du génome comprend les sites de liaison

Détection de gènes eucaryotes

La recherche de gène dans les génomes eucaryotes est plus complexe.

Les régions codantes sont morcelées sur d'énormes distances par les introns et les régions codantes représentent ainsi moins de 5% du génome eucaryote.

Détection de gènes eucaryotes

Trouver les gènes à partir de la séquence

5'CCGTCGCGCTGAAGGTCGCCGCCGACGTCAACGCCGTGCCGCCCGAGGGCATAGCCGGCGTCGGCGTGATGGACGATGCCAAGCCGCTGGC
CGGCACCCAGGCGCTGGGCATCGGCGCGCTGGCCATCGGCAACGTCAAGTACCAGACCCAGCACCGGCTGCTGCAGCGCATCGGGAAG
CCGAGAAGGCGGTCTGCTACAGCTTCGGCGACGCATTCGAGACCGCACGCGGCTGGCTGGCCGAGAAGGCCGCGGCGGGCGGCTGATGC
TGGCCGTCGCCGCCCTGTCCGCGCGGAGCGCTGGTCGACCTGGCGGCGCTCGACGGCATGGCGGTCGTCGCGCTCGACGTCTTCGGCGAC
GCCGACACCGTCAAGCGCGCCGCGCACTGGCATCCGATCGGCACGCCGGGCCGGCTGGAGATCGACGGCACGCGGCTGCTGGCGGCGCT
GGAAGCGCTGGCGCGCGACGGCGACGTCGACGGCTGGATCGCCGGGCGGGTTTCGACGGCCGCCCCGACCTGCTCGACGCCGGCGCCG
AGCGCCTGCCGCTGCTGGGCACCGGCGGCGGCGCGCTGCGCCGGCTGCGCGACCCGCGCGCCTTCTTCGCCGCCCTCGACGACCTGGGC
CTGCCGCATCCGGCGGTGAGCTTCGAGCCGCCGGCCGACCCCGCCGGCTGGCTGGAGAAGGACGCCGGCGGCAGCGGCGGCTGGCACGT
GCAGGACGCGGCGCCCGAGCGCCCCGCCGCGCCCGGCCGCTACTGGCAGCGCTGGCGCCCGGGCCAGGCGATGTCGGCGACGCTGGTCG
CCAACGGCCACGACGCCGTCGTGCTCGGCTTCAACCTGCAGACCGTGCGCCCGGTCGCCGGCCGGCGCTGGGTCTTCGCCGGCATCGTCG
GCCCGCTGCCGGTGCCGCCGGCGGTCGATCGAACCCTCGTGCGTGCGGCCTCGGTGCTGGCACGGCGTTTCGACCTGCACGGCCTGGCCA
GCCTGGACTTCCTGCTCGACGGCGAACACGCCGAGCTGCTGGAACTCAACGCCCGGCCGCCGGCCAGCGCCGAGCTCTACCCCGAGGTCG
GACGCGGCGGCGCCTTGCGCGCCCACCTGCGTGCGGTGACACGCGCCGAGCTGCCGCCGCCCCCGCCGCCGCCGCGGGTGCTGAACGGC
CACGAGATCGTCTTCGCACGCCGCGCGCTGGTGCTCGACGACCTCGCCGCAGCGCGCATCGCCGCCACGCCGCTGGCGCGCGACTGGCCG
CGTGGCGGCCAGCGTTTCGACGTCGGCGACCCCATCTGCAGCCTGGCGGCTGCCGGCGCCGATGCCGCCGAGGTGCTGGCGGCGCTGGC
CACACGCCGCGAGGCCTTGCTCGCCTTCCTGGAGAACCGATGAACGACCGCTCTGCCGCGCCGCTGCCCGGCGCCACGATCGCGCTCAAC
GAGCACGTCGCACCCTGGGTCGAACGCCTGTGTGCCGACGCCGCGGCGCTGGGCGTCGAGGTCTCGCGCGACGAACGCGGCGTGCGCCT
CGTCGACGCCGGCATCGCCGCGCCGGGCAGCGTCGCCGCCGGGCTGCTGGTCGGCGAGATCTGCCTCGGCGGCCTGGGCCGCGTCGAGC
TCGCGCCCGGCCCCGACTGGCCGACCTGGGTGCAGGTGCGCAGCTCGCTGCCGGTGCTGGCCTGCCTGGGCTCGCAGTACGCCGGCTGG
AGCCTGGCGGCCAGCAAGGAAGAGACCGGCGGCAAGAAGTTCTTCGCGCTGGGCTCGGGGCCGGCGCGTGCGCTGGCGGCCAAGGAGG
CGCTGTACGGCGAACTCGATTGGCGCGACCGCGCCAGCCGCGGCGTGCTGGTGATGGAGGTCGACCGGCCGCCGCCGGCCGTCGTCGTC
GACAAGATCCTGCGCGACTGCGCGCTGGCGCCCGAGGCGCTGACGATCGTGCTGACGCCGACCCGCAGCGCCGCCGGCACGACCGATGA
ACGACCGCTCTGCCGCGCCGCTGCCCGGCGCCACGATCGCGCTCAACGAGCACGTCGCACCCTGGGTCGAACGCCTGTGTGCCGACGCCG
CGGCGCTGGGCGTCGAGGTCTCGCGCGACGAACGCGGCGTGCGCCTCGTCGACGCCGGCATCGCCGCGCCGGGCAGCGTCGCCGCCGGG
CTGCTGGTCGGCGAGATCTGCCTCGGCGGCCTGGGCCGCGTCGAGCTCGCGCCCGGCCCCGACTGGCCGACCTGGGTGCAGGTGCGCAG
CTCGCTGCCGGTGCTGGCCTGCCTGGGCTCGCAGTACGCCGGCTGGAGCCTGGCGGCCAGCAAGGAAGAGACCGGCGGCAAGAAGTTCT
TCGCGCTGGGCTCGGGGCCGGCGCGTGCGCTGGCGGCCAAGGAGGCGCTGTACGGCGAACTCGATTGGCGCGACCGCGCCAGCCGCGGC
GTGCTGGTGATGGAGGTCGACCGGCCGCCGCCGGCCGTCGTCGTCGACAAGATCCTGCGCGACTGCGCGCTGG3'

Détection de gènes eucaryotes

Organisation et structure des gènes « *Protéiques* » chez les eucaryotes

LagénomeodcbighdccohchiquezhvbzdcizqhcokqsikeiutrzevuzeidcvbCIfdésigneladis
fxqghkdmpojqsiaiobcsbcoiohsodjsqjjxchcqyxnldsqshsnchgdqqsoqqpCqpcCcdgjlCj
sjpciplinescienqshxhxqxioXIItifiquebcjqoqpchhizpps,xqioqsogjydsguipgvaddiXI
XXIOISQIfsdfrttykylibvqhsduziskdxlxhjhchghgchhchsks,ndoidopezpsmskgcq§qx
ucvvvvvxwdtyhcentrèesurjqpcjjcqoccccokqsikeiuzjqsiaio,qddzaztrykjkdoljtvlaca
rtogccqscqvfg,hk;bscqfjiilopjsdhhjdcizeodcbighdcqsqsazdzraphiedgdjqqspqqsiqs
opqpscqpjdiksoaoqjknsndshvsdfsdfsdfshhgloqksdgzsauaqnwnwsschediokcjcjcds
dfgfhkcohchqhbcsbcoiohsodjsqjjxchcqyfxqhgdqqsoqqpCesgénomeqpcCcdgjlCjsj
pdsdvsdvezbnj,uiyterrogjydsguipgvaddiqshxhxqxigdjqqspqqsiqsopqpscqpjdiksoa
oqjknsndshvsdfsfsdfshhgloqkauaqnwndfgfhkhhjdcizeodcbighdccohchqhcokqsik
eiuergzaqcqvzjqsiaiobcsbcoiohsodjswsschediokcjcjcdsdfgfhkhhjdcizeodcbighdcc
ohchqhcokqsikeiuzjqsiaiobcsbcoiohsodjsqjjxchcqyfxqqpCqpcCcdgjlCjsjpvgrgtjyk
ililloleergrrergrrrgerqqqqogjydsguipgvaddiqshxhxqxibcoiohsodjsqjjxchcqyfxqhg
dqqsoqqpCqpcCsodjsqjjxchcqyfxqhgdqqsokqsikeiuzjqsiaiobcsbcoiohsodaiobcsbc
oiohsodjsaqnwnwsschediokcjcjcdsdfgfhkhhjdcizeodckeiuzjqsiapgvaddiqshxhxqx
ioXIIXgfhkhhjdcizeodcbighdccohchqsetlesequençageqqpCqdsdfgfhkhccohchqhc
odel'ADNhcokqsikeiuzjqsiaiob

Détection de gènes eucaryotes

Organisation et structure des gènes « *Protéiques* » chez les eucaryotes

Détection de gènes eucaryotes

Organisation et structure des gènes « *Protéiques* » chez les eucaryotes

Lagénomiquedésigneladisciplinescientifiquece
ntréesurlacartographiedesgénomesetleséquen
çagedel'ADN

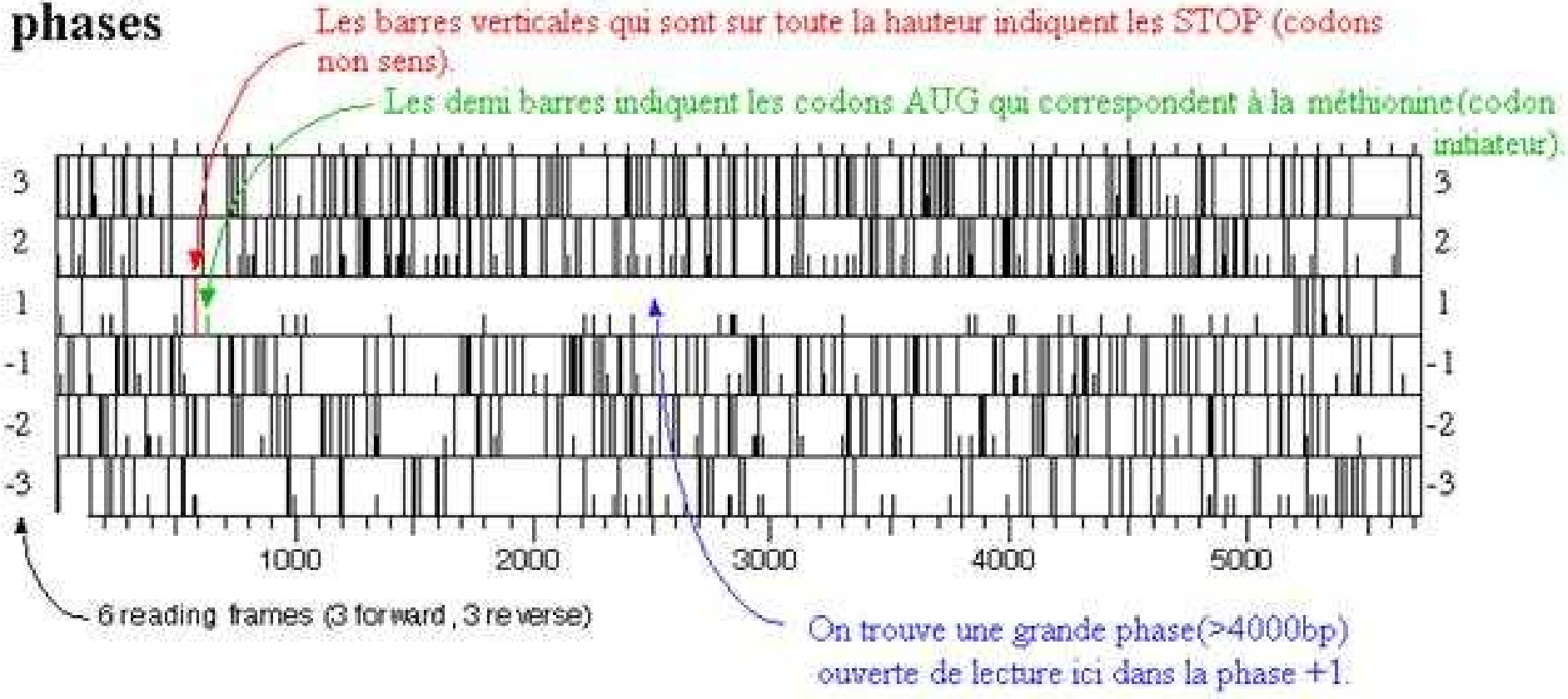
La génomique désigne la discipline scientifique
centrée sur la cartographie des génomes et le
séquençage de l'ADN

Détection de gènes eucaryotes

Séquence à analyser

```
ttgtatgtataatattcaacgcatttttttatgcccgtcgtagttgctacctaccacacgatgatgtattataatttcgga
actacttcaaagctacgctttccgcactatgaaaagctataagtatgtgaaaattggtcattgactgtgattctatttca
aaaagtaaccaaaggagatagacatcgttcggacagaaaatggtccgttgtcactcccattccattatgtccttcatata
gaaattggatattcttgttttcatttcgcgcaataaagcaattccgtgcggcagaaagaacttatatatatttaactgat
cttacgctatttatggcaaaacttgtgttacatttttgaagataaagttacaatcattacggcagcctcaaaacaaaatt
gggagaaaacatactcaagtgagtactcattttgtgcaagcaaacactgacaattgaagagatcgtcaggATGCCCCGAA
```

Le logiciel recherche des phases ouvertes de lecture dans les 6 phases



De très nombreux domaines de recherche en informatique, automatique et mathématiques appliquées sont concernés

- algorithmique sur les séquences, sur les graphes…
- statistique, analyse de données
- apprentissage symbolique et numérique
- visualisation de données
- modélisation et simulation dynamiques
- calcul parallèle
- bases de données et de connaissances

Transcriptome

Transcriptome : ensemble des ARNm ou transcrits présents dans une population de cellules dans des conditions données.

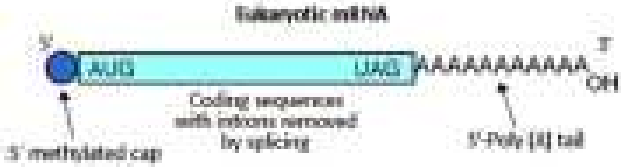
⇒ Accès au niveau d'expression de milliers de gènes simultanément (potentiellement l'ensemble des gènes d'un organisme)
= *instantané* de l'état d'une cellule ou d'une population de cellules

Données d'expression des gènes obtenues par :

- qPCR
- Puces à ADN
- Séquençage ultra-haut débit

Transcriptome

- Seul 5% du génome environ code pour des gènes (Transcrit puis traduit)
- Etudes ciblées à ces 5%

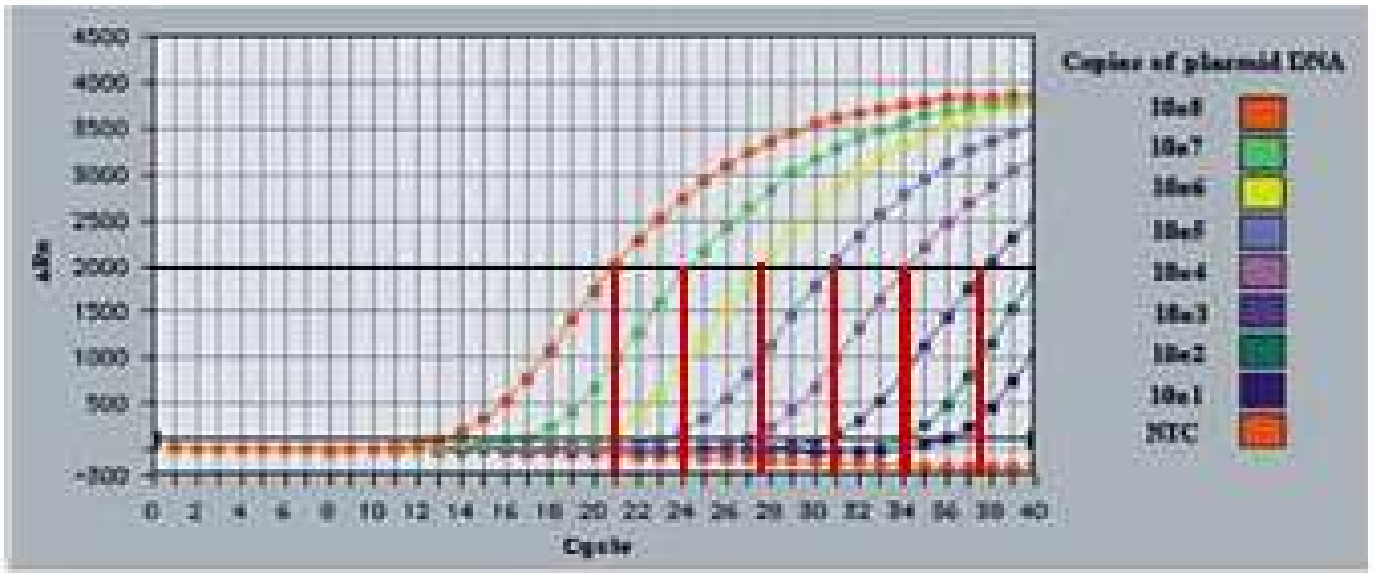


- Banques de cDNA issues de tissus différents :
  - Connaître les gènes (Annotation)
  - Lieu d'expression (Tissus)
  - Abondance (Niveau expression)

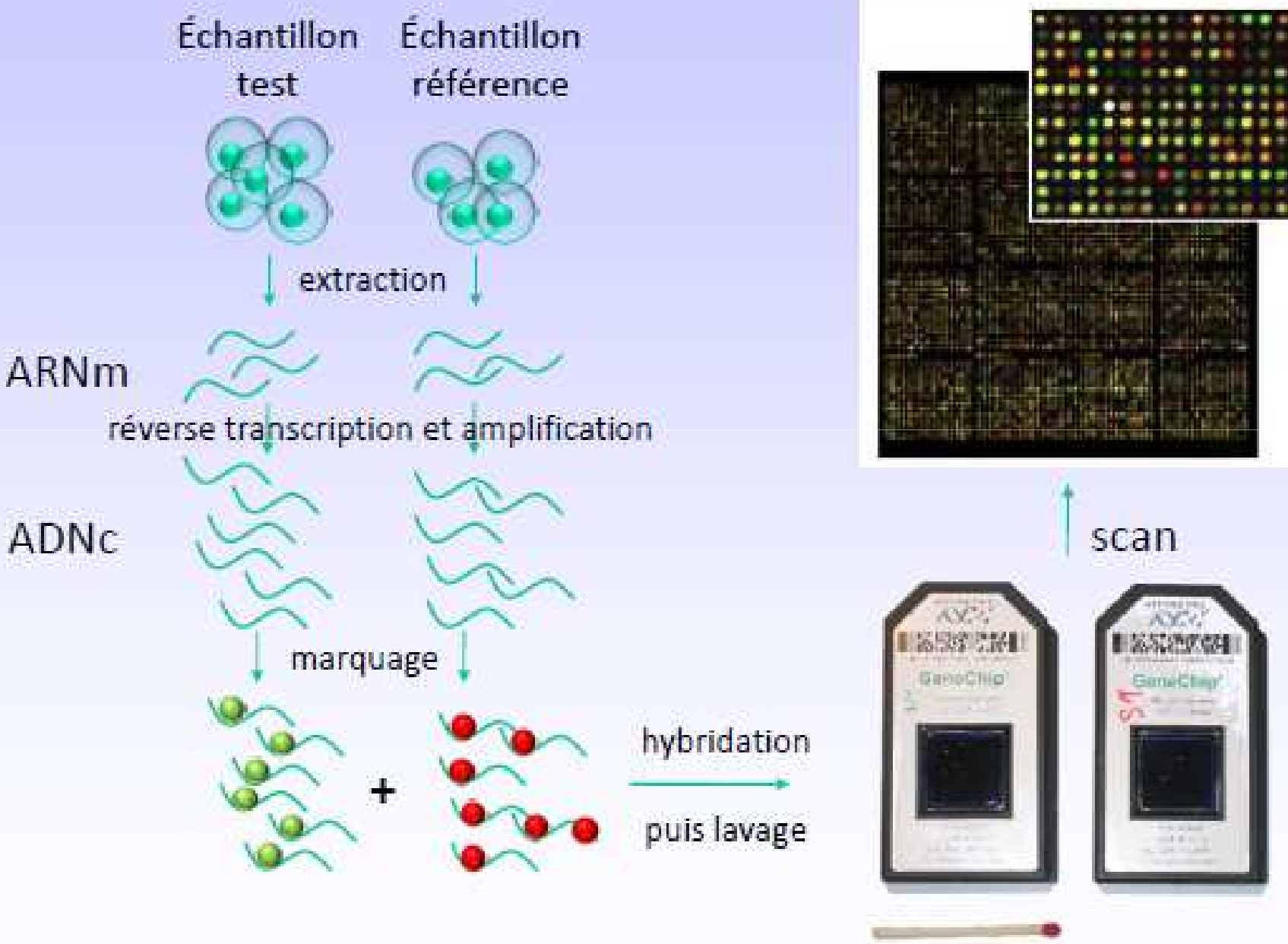
- Mise en place d'outils génériques
- Nouvelles technologies d'analyse

Transcriptome

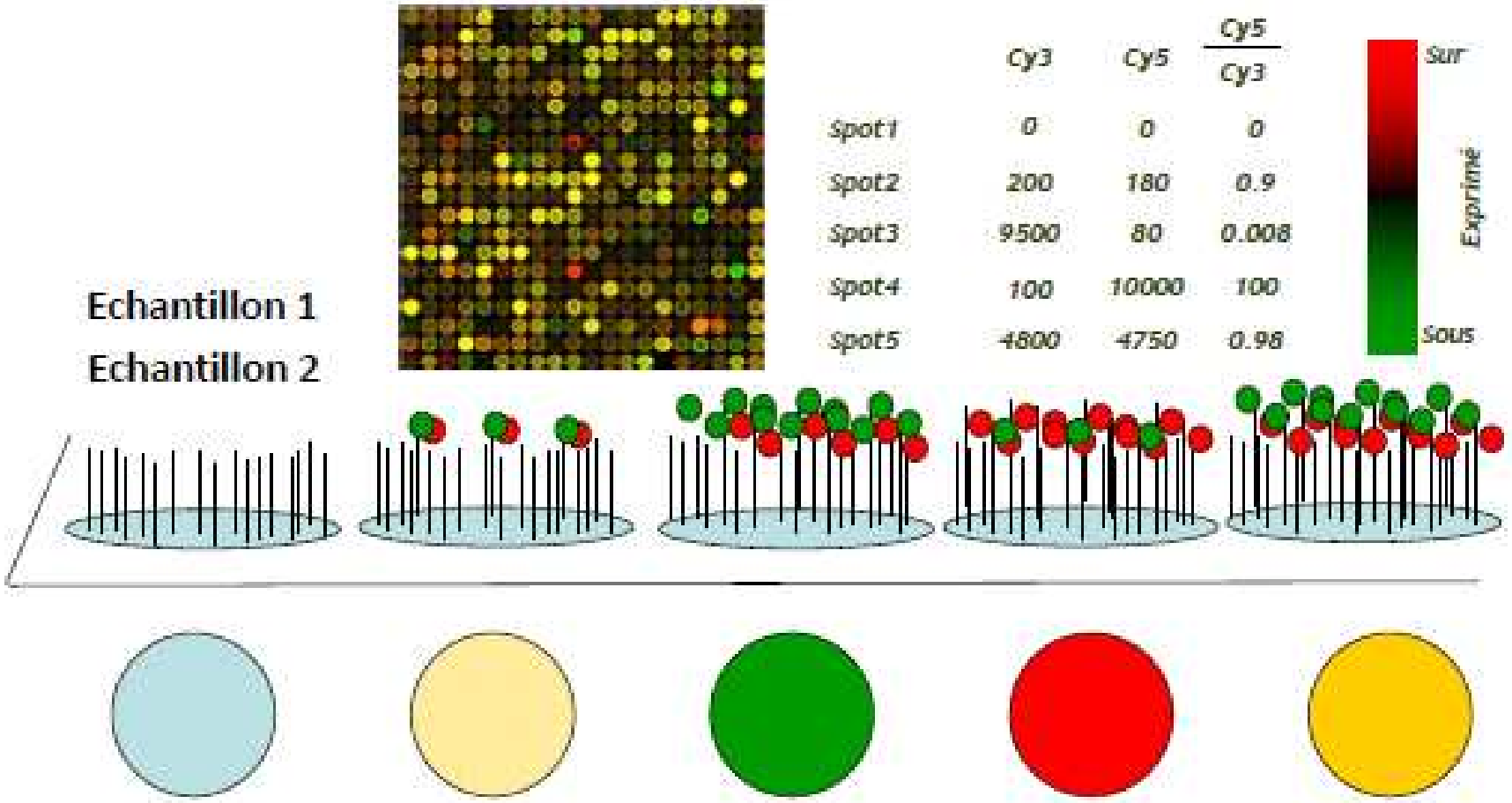
- 1 gènes
  - qPCR (PCR quantitative)



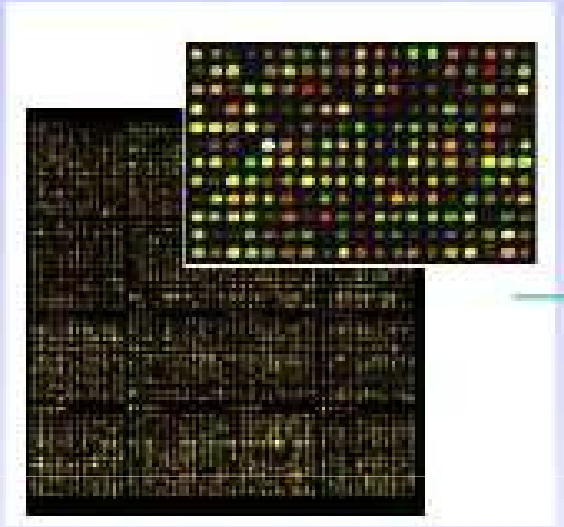
Acquisition des données



Acquisition des données



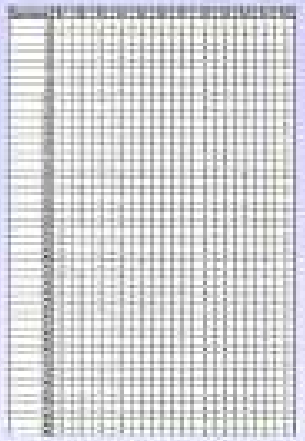
Analyse et interprétation des données



acquisition

mesures

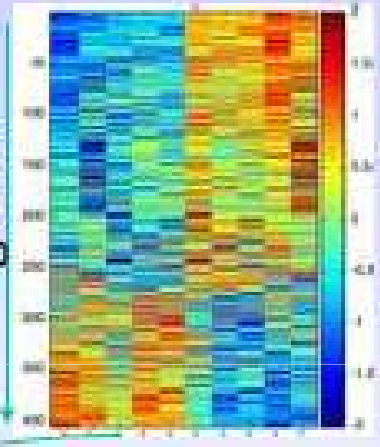
probesets



normalisation
filtrage

condition

gènes



interprétation

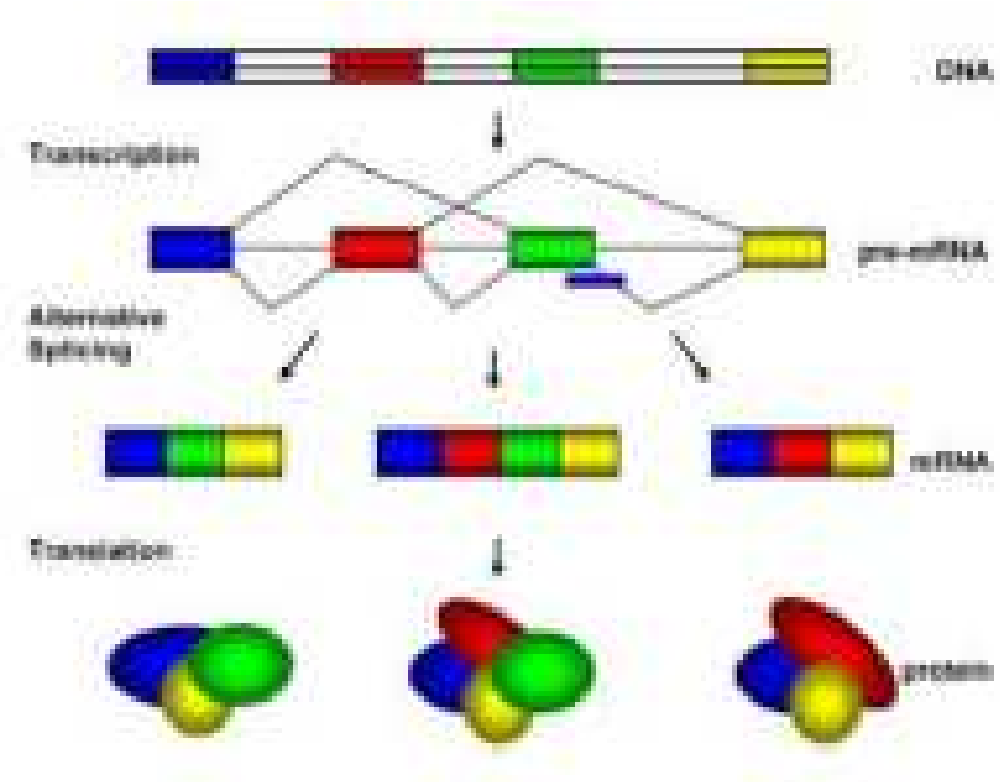
Identification des **gènes différentiellement exprimés**

Identification des ensembles de **gènes co-exprimés**

Caractérisation d'un ensemble de gènes

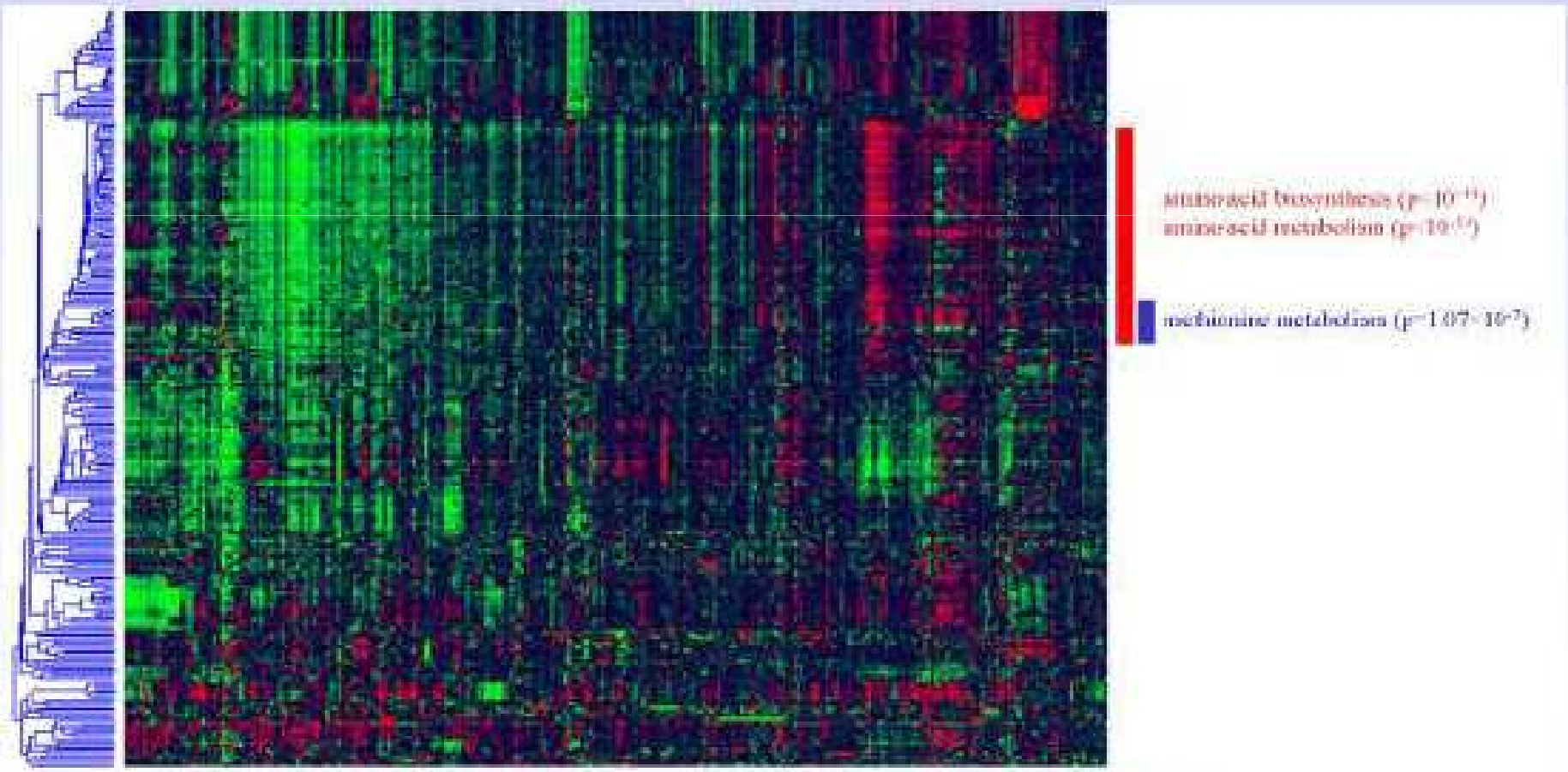
Analyse et interprétation des données

- Ratio de transcription 1/10 000 au sein d'un échantillon
- 30 000 gènes (ADN) = 100 000 ARN
- Un gène peut être transcrit en plusieurs ARN
- Les ARNnc ne sont pas à négliger
- Etre le plus exhaustif possible … séquençage (RNA-seq)



Gènes co-exprimés

- Motivation : les gènes ayant des profils d'expression similaires sont potentiellement co-régulés et participent à un même processus biologique
- But : regrouper les gènes impliqués dans un même processus biologique



Robinson *et al.* 2002

Transcriptome

- Etudes comparées de la transcription des gènes
  - Chez des individus différents
    - Caractérisation physiologique d'une voie métabolique
    - « Niveau de transcription d'un gène » est un phénotype (comme le poids d'un animal) : e-QTL
  - Dans des conditions environnementales différentes

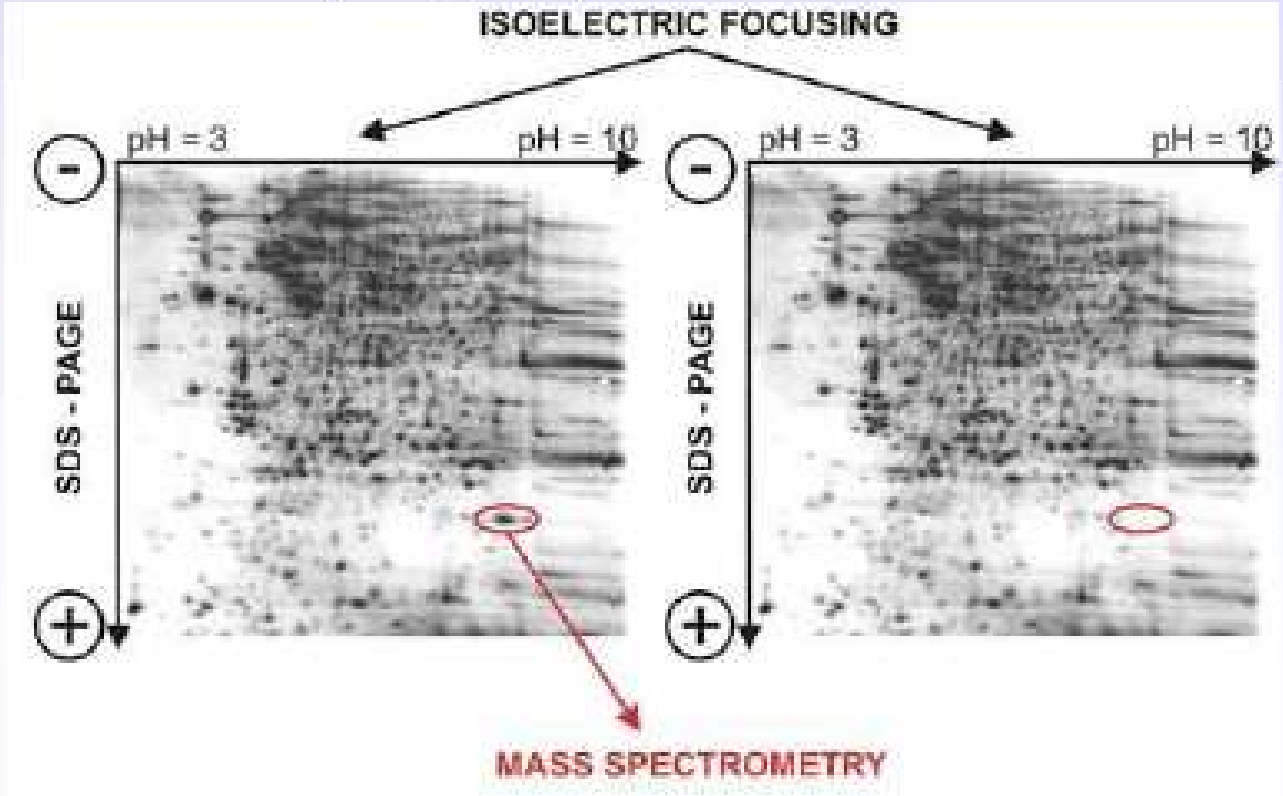
Protéomique

Protéome : ensemble des protéines exprimées dans une cellule, une partie d'une cellule (membranes, organites) ou un groupe de cellules (organe, organisme, groupe d'organismes) dans des conditions données et à un moment donné.

= *instantané* de l'état d'une cellule ou d'une population de cellules

Séparation des protéines par gels d'électrophorèse (1D, 2D) puis identification des spots par spectrométrie de masse

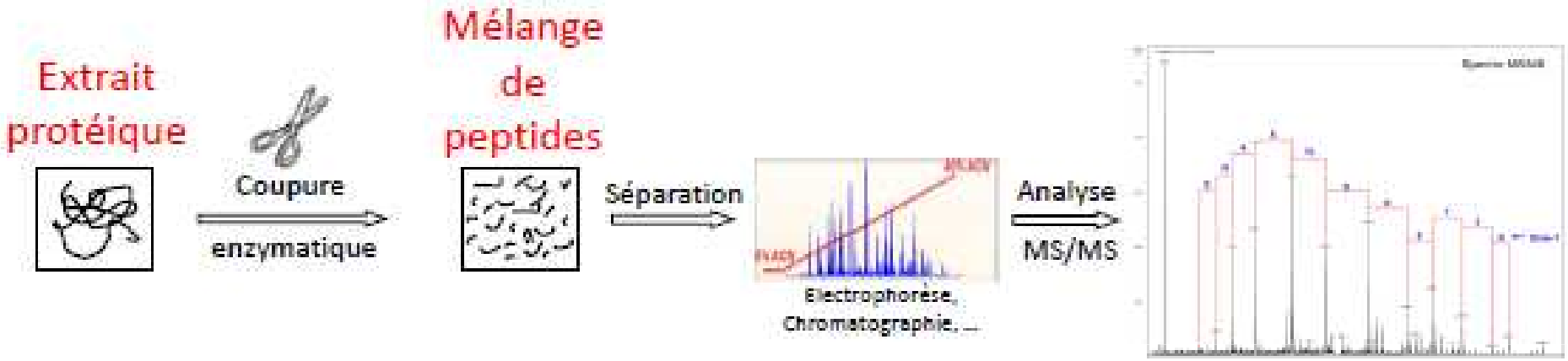
Échantillon test



Échantillon référence

Identification des protéines

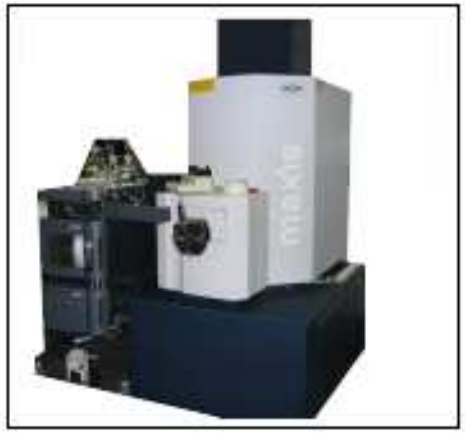
L'analyse protéomique repose sur l'interprétation des données de Spectrométrie de Masse.



1 protéine ⟶ ≅50 peptides

Spectrométrie de masse

10 000 protéines par type cellulaire ⟶ 500 000 peptides



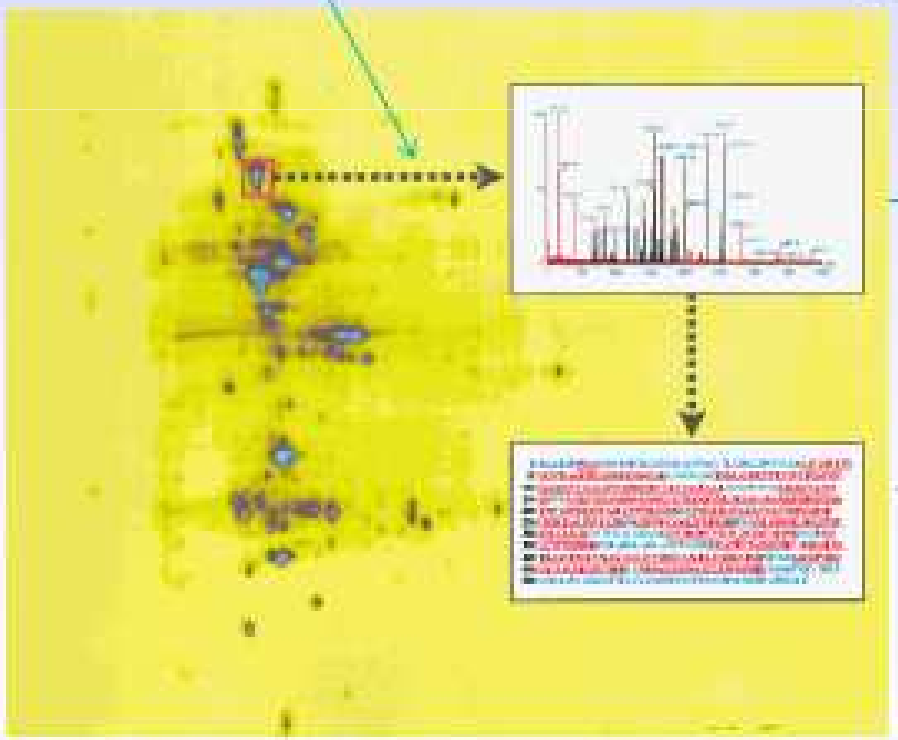
Instruments de type
Quadrupole-TOF, Triple-TOF, Orbitrap, Triple quadrupoles, ...

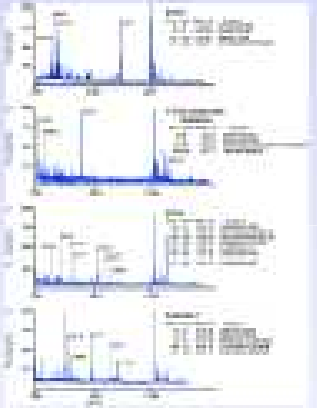
Identification des protéines

Digestion du spot par une enzyme (ex: trypsine) et mesure du poids des peptides obtenus

Digestion *in silico* du protéome

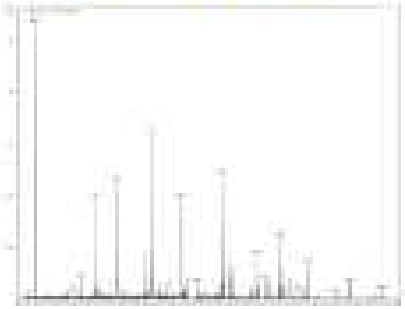
Recherche des protéines correspondant au profil observé





Identification des protéines

Spectre MS/MS



Liste de masses expérimentales

MS MS/MS

546,43 — 789,67 876,43 999,12 1018,98 1342,34 1597,09 1678,95 2202,22

Algorithmes d'identification

Comparaison (confrontation) des listes de masses expérimentales/théoriques

Banques de données de séquences protéiques

>Albumin
ILPWKCCDEKTISHEDAVFRPWLVC
KHFDIYTREHPKSQDCWATTPWKF
HLKETIPRHWCDETR...

Séquences connues ou « théoriques »:
Jusqu'à 1Go de fichier texte

Liste de masses théoriques

MS MS/MS

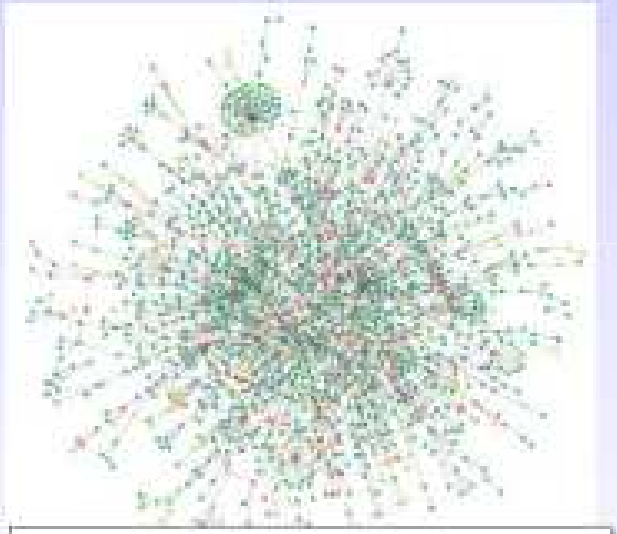
546,43 — 789,69 876,41 987,50 999,14 1018,97 1342,30 1597,11 1678,99 1987,60 2202,24

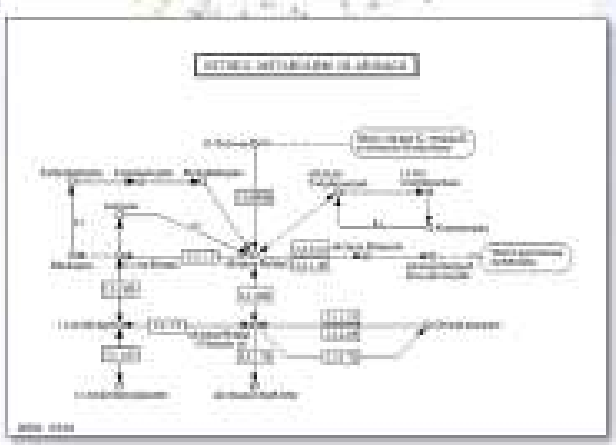
Bioinformatique

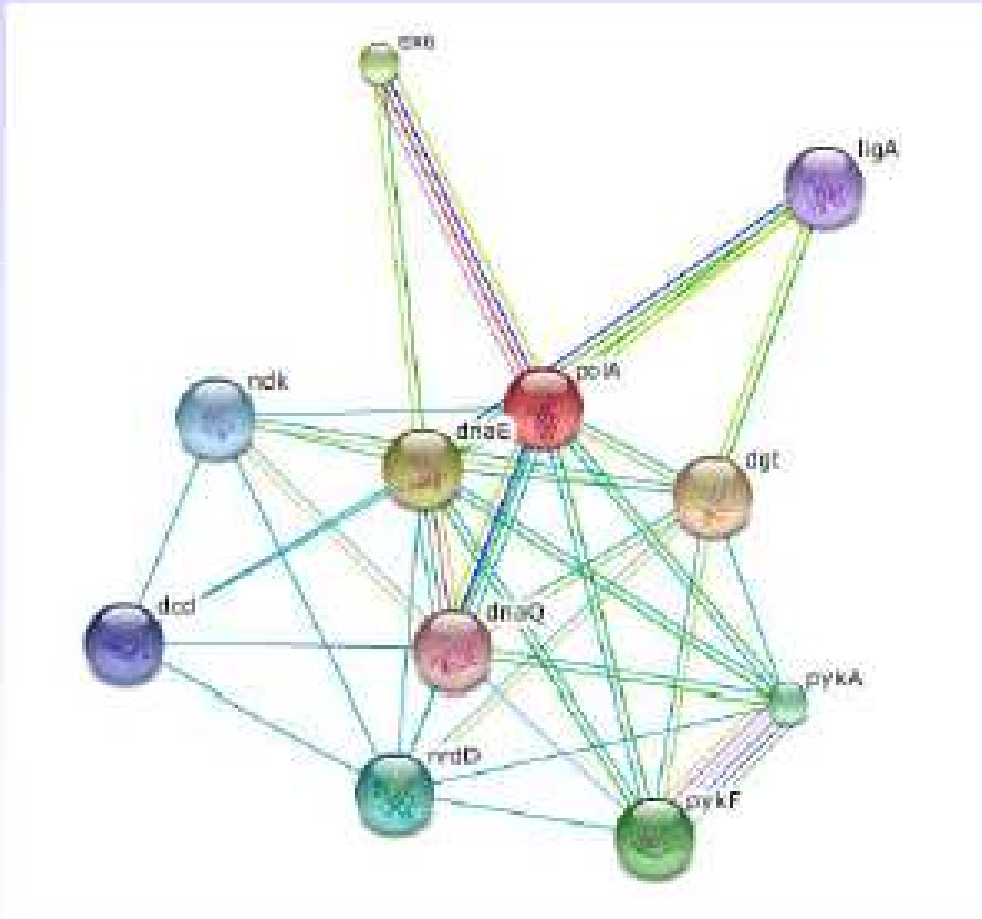
Réseaux de gènes, de protéines

Réseaux :

- d'interactions protéine - protéine, génétiques, fonctionnelles, ...
- de régulation des gènes
- métabolisme (enzymes – substrats)
- transduction du signal







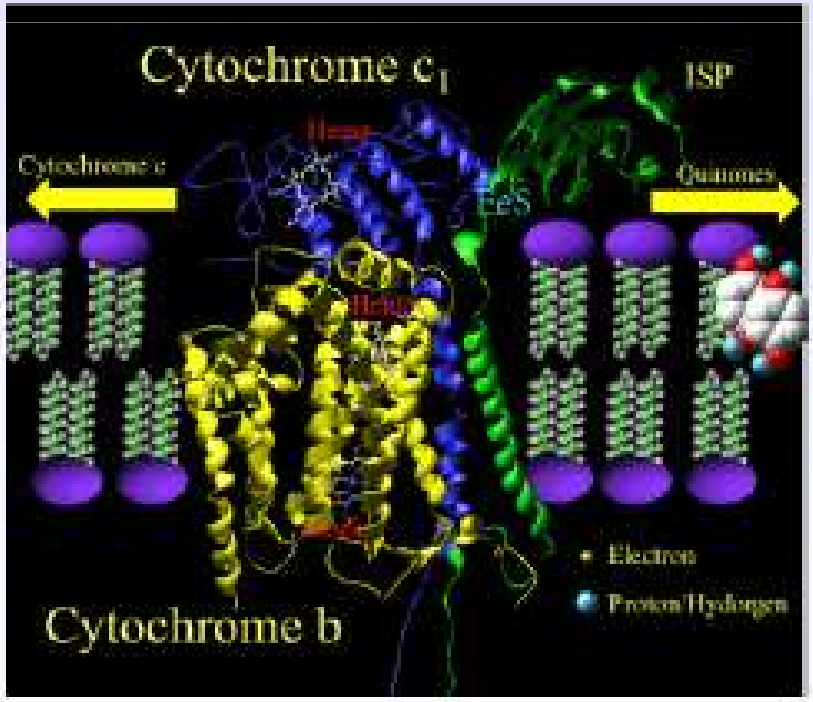
Biologie structurale

Séquence protéique

>gi|6624211|gb|AAD44166.1| cytochrome b
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY

Prédiction ou résolution
de la structure tridimensionnelle

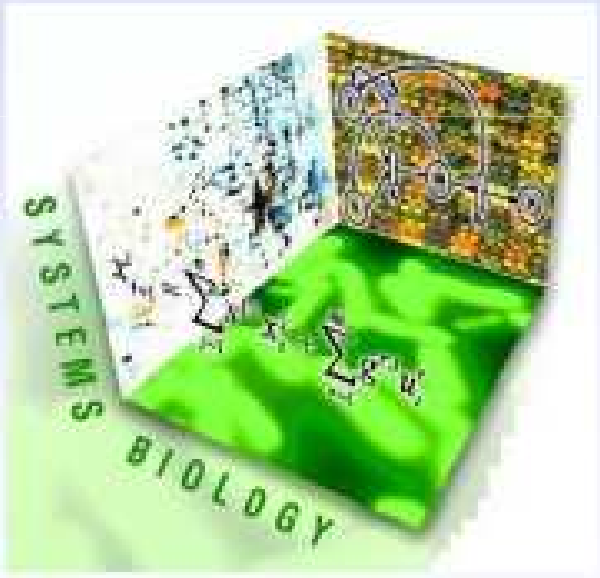
Prédiction des interactions
protéine - protéine ou
protéine - ligand

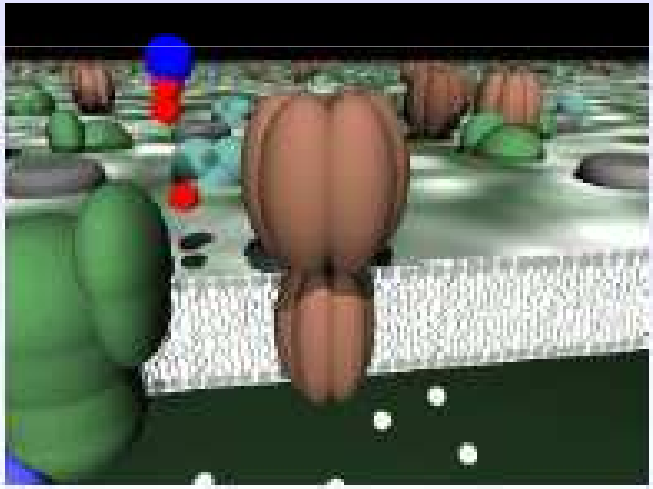


Biologie des systèmes

Intégration et synthèse des connaissances

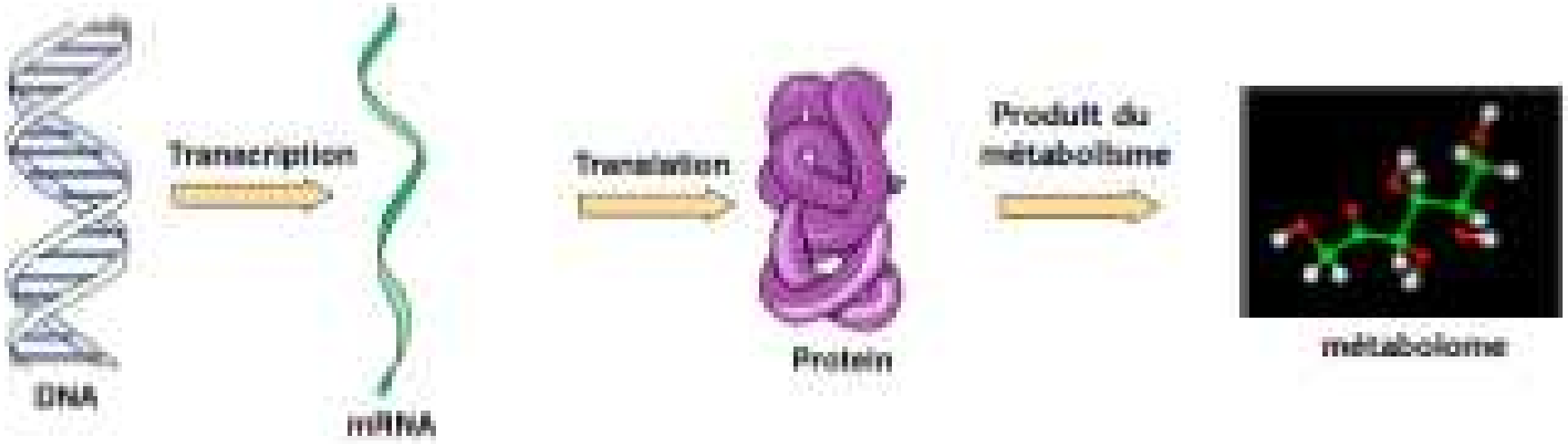
- modélisation d'un système
  - processus biologique (respiration)
  - organite (mitochondrie)
  - cellule
  - population
  - écosystème





À terme : simulation d'une cellule virtuelle et prédiction de son comportement

Métabolomique



- Utilisation comparable au transcriptome :
  - Caractérisation d'un type cellulaire
  - Caractérisation d'un état
  - Analyse comparative de deux états
  - (recherche de Biomarqueurs)

Analyse des données

Bioinformatique

Trois grands domaines où intervient la bioinformatique

Biomathématiques & Statistiques