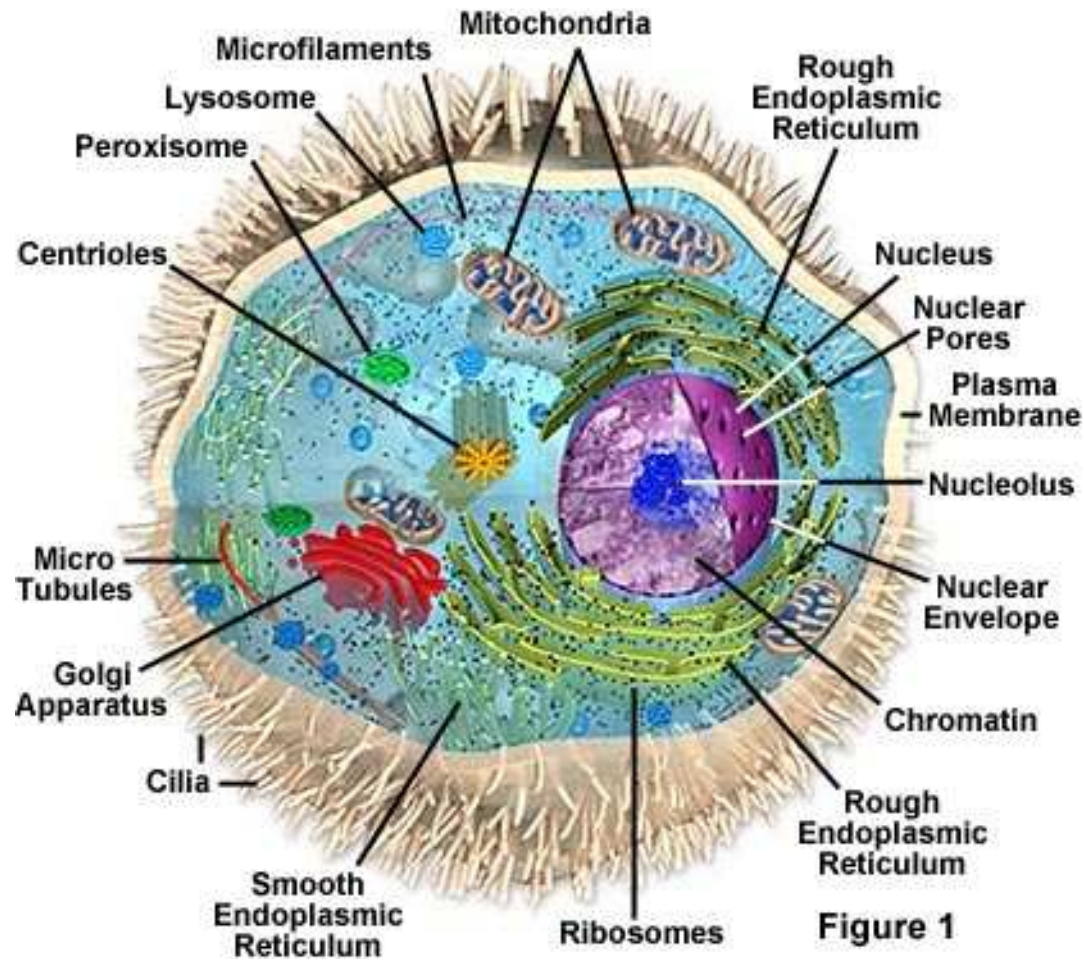


Analyse de séquences

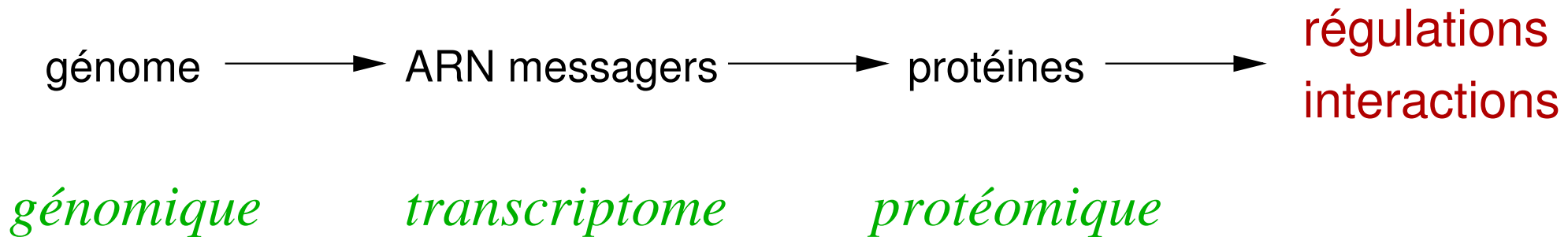
Hélène Touzet

touzet@lifl.fr

Anatomy of the Animal Cell



Sources d'information au sein de la cellule



Protéome

Ensemble des protéines présentes dans la cellule

Technologies :

▷ Gel bi-dimensionnel

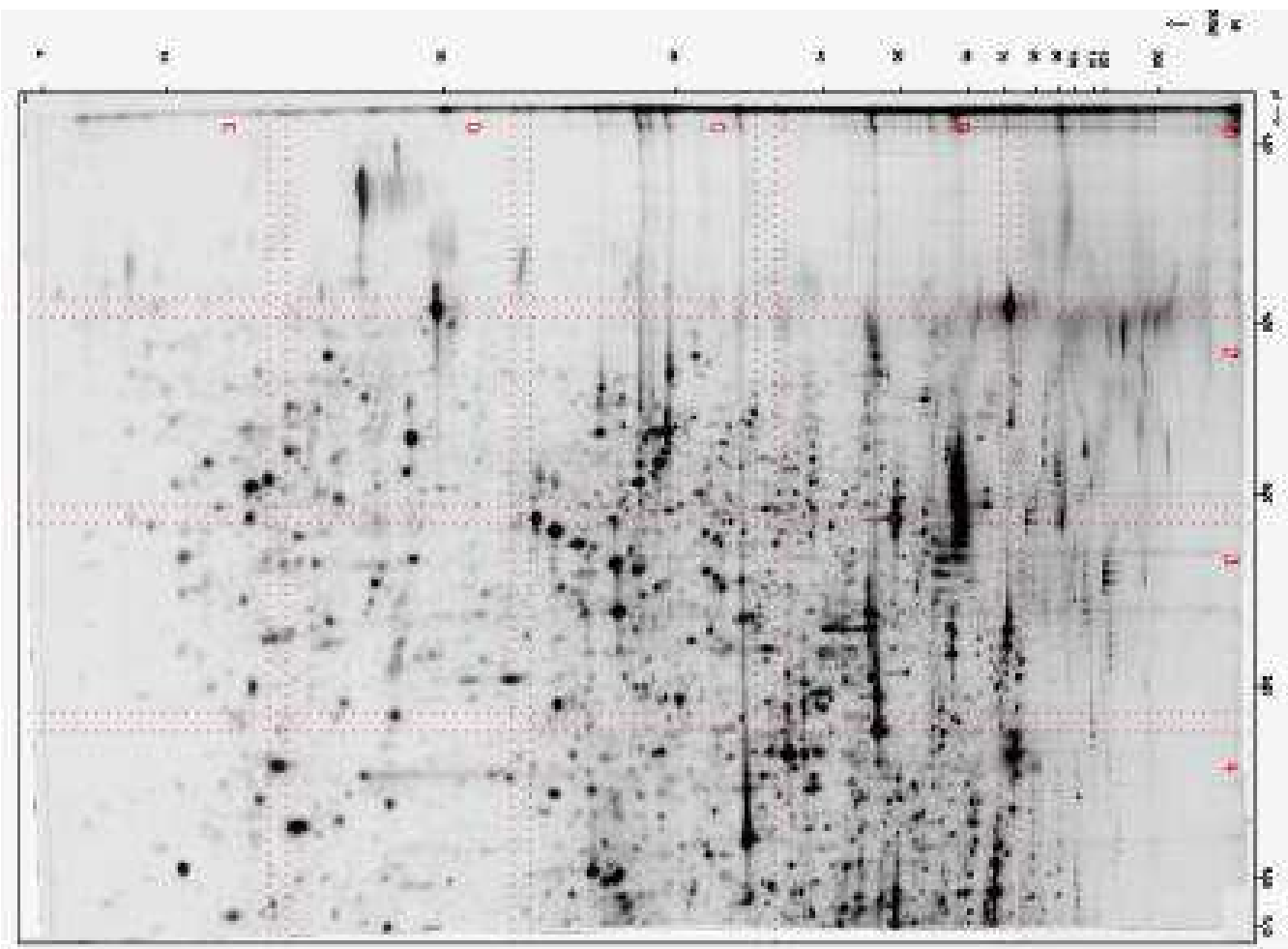
Séparation des protéines suivant leur charge, puis suivant leur masse moléculaire

Banques de données : Swiss-2D, ...

▷ Spectrométrie de masse

Fragmentation d'une protéine et masse de chacun des fragments

Problèmes de reconstruction, d'identification de la protéine à partir des masses des fragments



Transcriptome

Ensemble des ARNm présents dans la cellule
(mesure le niveau d'**expression des gènes**)

Technologie : puce à ADN

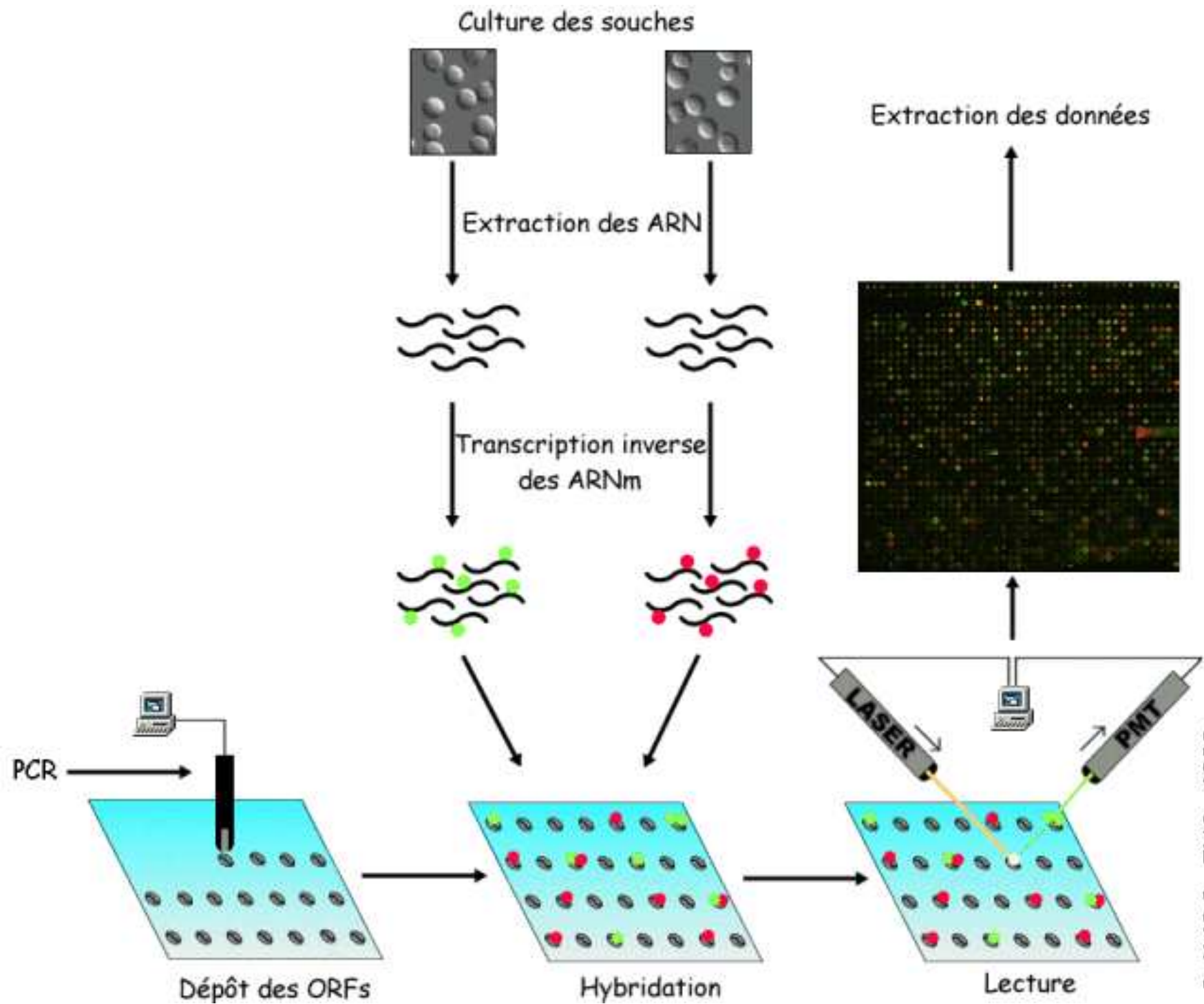
- ▷ **puce** : plaque où sont déposés régulièrement des brins d'ADNc (brin d'ADN complémentaires à des ARNm)

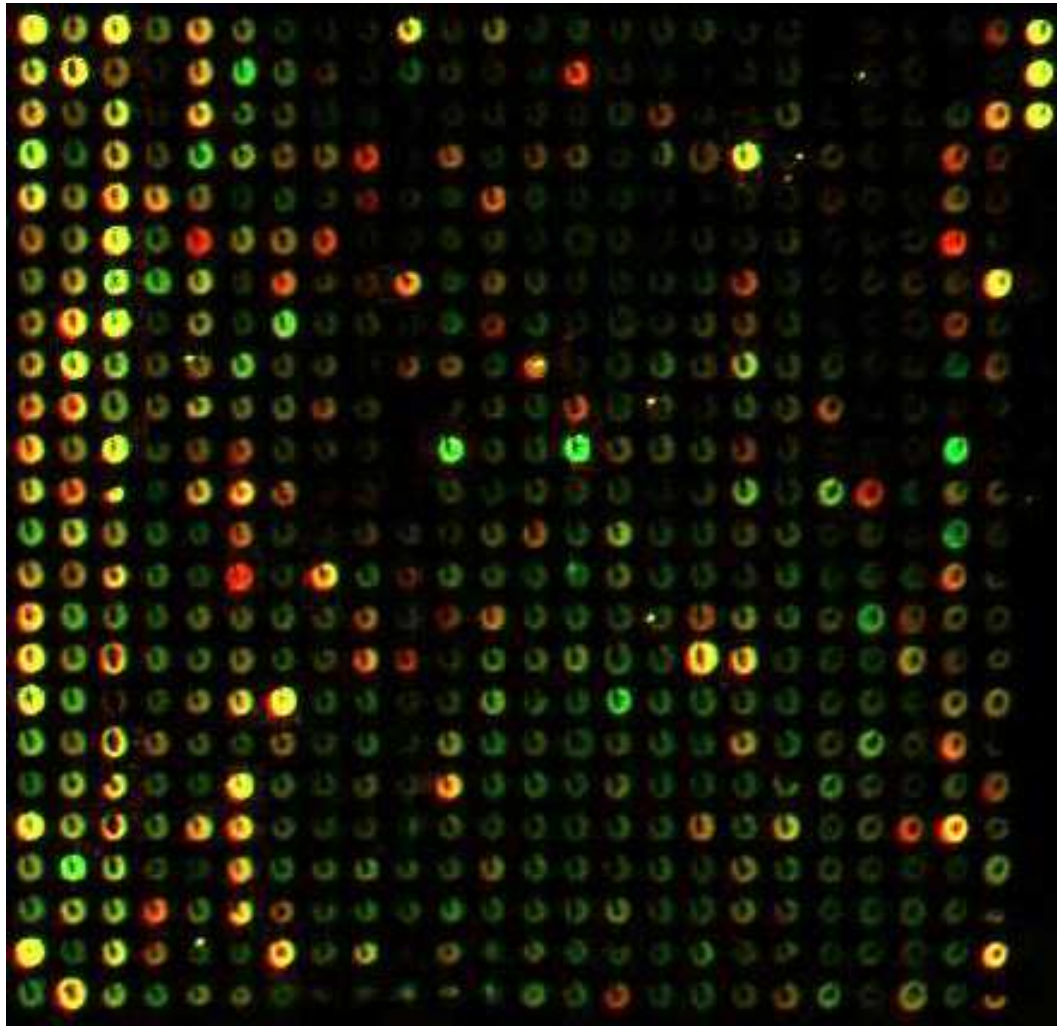
Support : membrane nylon, verre

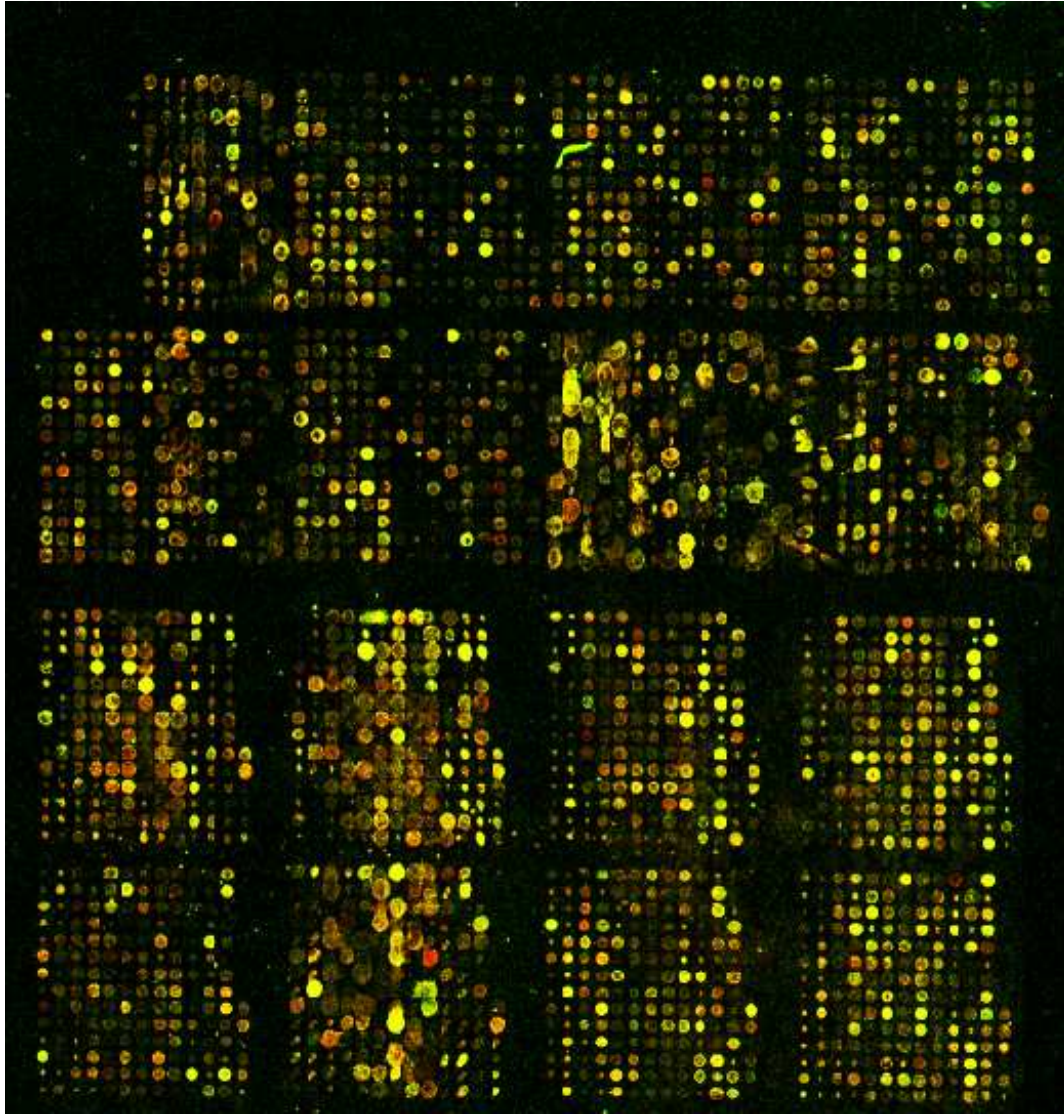
- ▷ marquage des ARNm à étudier

fluorescence ou radioactivité

- ▷ hybridation de la puce avec les ARNm







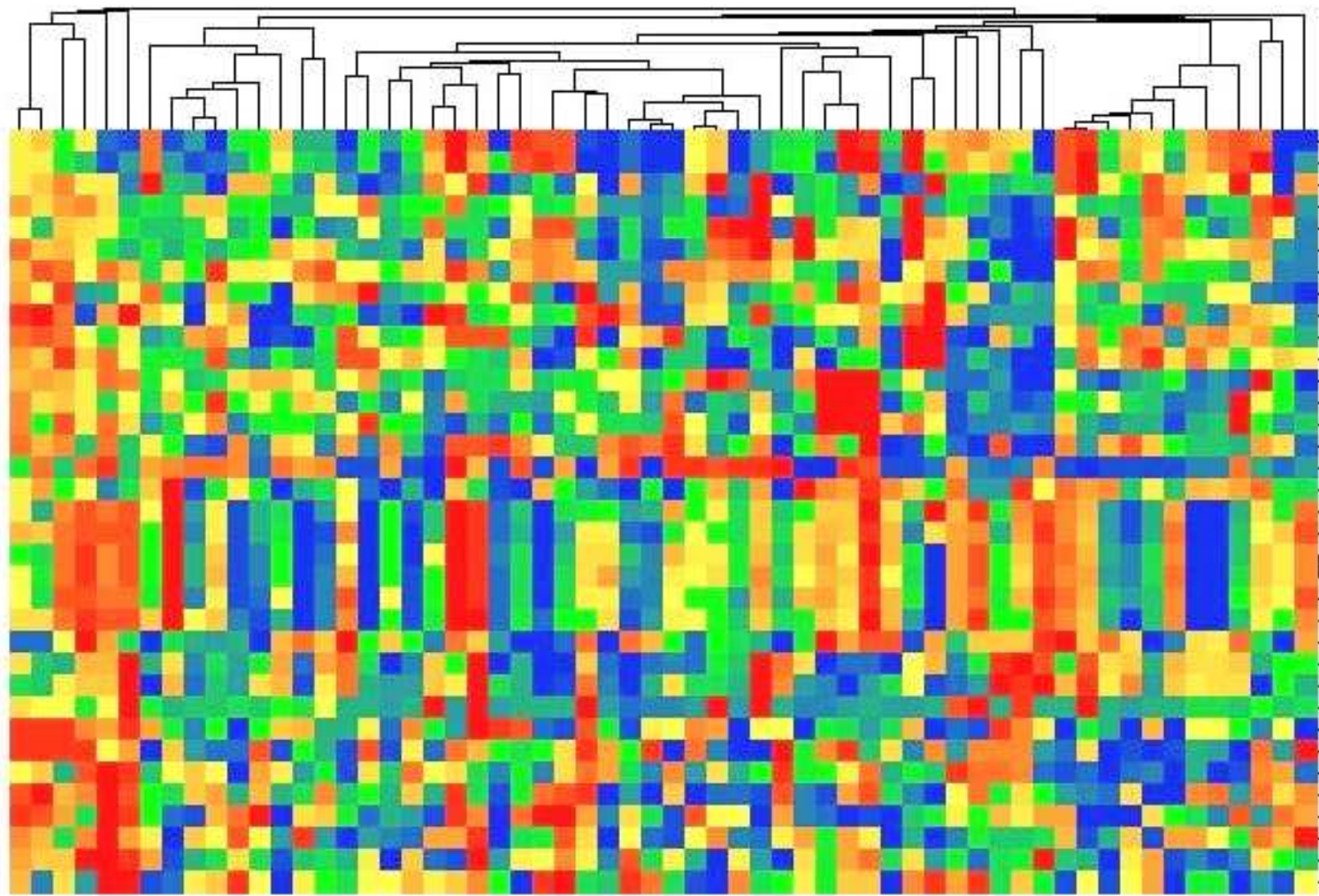
Puce à ADN : Lecture de l'information

- ▷ Digitalisation de la puce
- ▷ Analyse de l'image : matrices d'expression
- ▷ Fouilles de données, clustering, analyse statistique

Quels sont les gènes qui s'expriment dans les mêmes conditions?

Quels sont les échantillons similaires?

LE:MOLT-4
 LE:CCRF-CEM
 LE:HL-60
 LE:RPMI-8226
 LE:K-562
 LE:SR
 CO:COLO205
 CO:HCC-2998
 CO:HCT-15
 CO:HCT-116
 CO:SW-620
 CO:KM12
 CO:HT29
 BR:T-47D
 BR:MCF7
 ME:LOXIMV
 PR:PC-3
 OV:OVCA-5
 LC:NCI-H322M
 LC:EKVX
 LC:A549/ATCC
 LC:NCI-H460
 OV:SK-OV-3
 OV:IGROV1
 OV:OVCA-8
 BR:MCF7/ADF-RES
 PR:DU-145
 LC:HOP-62
 RE:CAKI-1
 RE:UO-31
 RE:ACHN
 RE:786-0
 RE:RXF-393
 RE:TK-10
 RE:A498
 BR:BT-549
 CNS:SNB-75
 CNS:SF-295
 CNS:SU251
 CNS:SNB-19
 CNS:SF-539
 OV:OVCA-4
 OV:OVCA-3
 BR:HS578T
 CNS:SF-268
 LC:NCI-H226
 BR:MDA-MB-231/A
 LC:HOP-92
 ME:UACC-257
 ME:MALME-3M
 ME:SK-MEL-28
 ME:M14
 ME:SK-MEL-2
 ME:UACC-62
 BR:MDA-N
 BR:MDA-MB-435
 ME:SK-MEL-5
 LC:NCI-H522
 LC:NCI-H23
 RE:SN12C



1-est_Homo sapiens e
 2-est_ESTs Chr.6 [236
 3-est_SID W 486793, I
 4-est_SID W 172857, I
 5-est_SID 484954, Cr
 6-est_SID 301448, ES
 7-est_SID 286200, ES
 8-est_SID 37330, EST:
 9-est_SID W 121145, I
 10-est_ESTs Chr.10 [4
 11-est_SID W 278644
 12-est_ESTs Chr.6 [25
 13-est_Human fetus b
 14-est_Human fetus b
 15-est_SID W 131843
 16-protein ? glutamyl
 17-est_RPL5 Ribosom
 18-est_SID W 487188
 19-est_SID W 364351
 20-est_SID W 471123
 21-est *Human ferritin
 22-est_SID 512268, Hi
 23-est_ESTs Chr.22 [4
 24-est_SID 485282, t:
 25-est_Human pre-B c
 26-est *EST AA68811
 27-protein_Glutathione
 28-est_SID 427845, Hi
 29-est_SID 128329, ES
 30-est_H factor (comp
 31-est_SID W 428225
 32-est_ESTs Chr.17 [4
 33-est_SID 72214, Hui
 34-est_ESTs Chr.1 [42
 35-est_ESTs Chr.1 [24

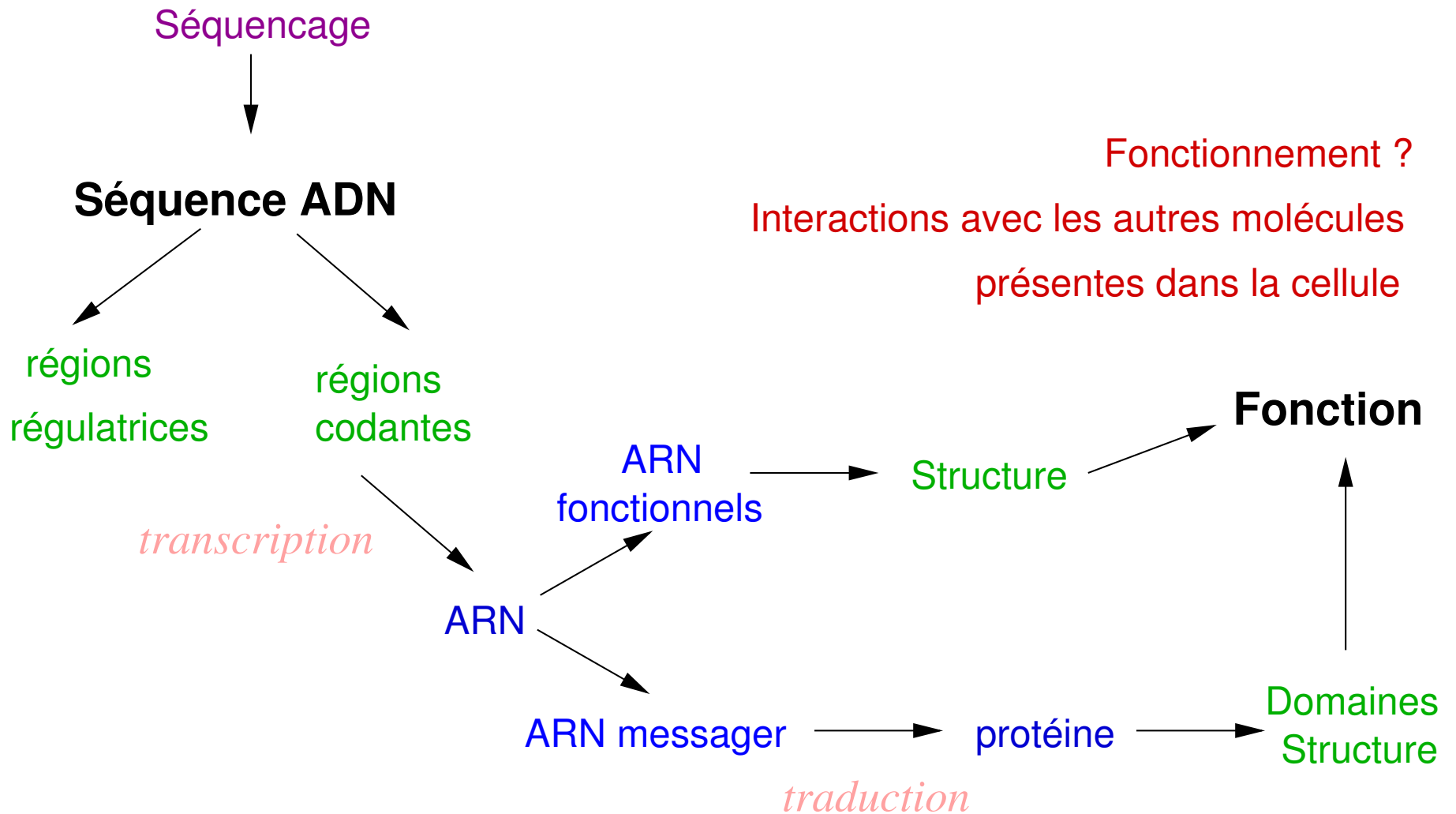
Génomique

Technologie : séquençage

Plus de 70 génomes connus :

- | | | |
|---------------|-----------------------------------|---|
| 1978 : | bactériophage <i>phiX174</i> | premier génome à ADN - 5386pb (Sanger) |
| 1995 : | <i>Haemophilus influenzae</i> | première bactérie - 1,83 Mb |
| 1996 : | <i>Methanococcus jannaschii</i> | première archée |
| 1997 : | <i>Saccharomyces cerevisiae</i> | premier organisme eucaryote - 12Mb |
| | <i>Escherichia coli</i> | la bactérie la plus étudiée |
| | <i>Bacillus Subtilis</i> | autre bactérie modèle |
| 1998 : | <i>Caenorhabditis elegans</i> | premier organisme pluricellulaire - 100Mb |
| 2000 : | <i>Drosophila melanogaster</i> | |
| | <i>Arabidopsis thaliana</i> | |
| | <i>Homo sapiens</i> | |
| 2002 : | <i>moustique, souris, ...</i> | |
| 2003 : | <i>fugu, rat, virus SRAS, ...</i> | |

La quête du Graal de la génomique



Annotation : de la séquence à la fonction

. . .
ctgcaggacgcctactaaggcggcggggaaaaacaaacggttattacaccgagacagaagg
tgcactgcggttatgttgtcgcggaacggcgaaaaggggctgaccttcgctggtgaacc
aattaagttggcgctatctgantctcatactgtttcacagacctgctgccctgcggcggc
caatcttcctttattcgcttataagcgtggagaattaaaatgcgacatcctttagtgatg
ggtaactggaaactgaacggcagccgccacatggttcacgagctggtttctaacctgcgt
aaagagctggcaggtggtgctggctgtgcggttgcaatcgaccaccggaaatgtatc
gatatggcgaagcgcgaagctgaaggcagccacatcatgctgggtgcgcaaacgtgaac
ctgaacctgtccggcgcattcacgggtgaaacctctgctgctatgctgaaagacatcggc
gcacagtacatcatcatcggctactctgaacgtcgtacttaccacaaagaatctgacgaa
ctgatcgcgaaaaaattcgcggtgctgaaagagcagggcctgactccggttctgtgcatc
ggtgaaaccgaagctgaaaatgaagcgggcaaaactgaagaagtttgcgcacgtcagatc
gacgcggtactgaaaactcagggctgctgcggcattcgaaggtgcggttatcgcttacgaa
cctgtatgggcaatcggctactggcaaatctgcaactccggctcaggcacaggctgttcac
aaattcatccgtgaccacatcgctaaagttgacgctaacatcgctgaacaagtgatcatt
cagtacggcggctccgtaaacgcgtctaacgctgcagaactgtttgctcagccggatc
gacggcgcgctggttggtggtgcttctctgaaagctgacgccttcgcagtaatcgttaaa
gctgcagaagcggctaaacaggcttaagtctgacaggtgccggatttcgatatccggcac
ttactttccttaactcttcgccttaacgcaaaatctcacactgatgatcctgaatttcct
cggctgaagcacggttaagcgtcagtagatttcggtgtgtcgccagcaatacaaatgagt
tatcactctgccgtaccatcgccagcccgtagcgtcccatatgttcccgcgcctcaggta
cttcttctgccagcatcataaatgggctgcggtgtaccagttcgctttccgttaccgac
gcgaggtattcatgcccgcgcaaacctggcagtggaaccagcggctgctgatggt
cgccagattgttatcgag . . .

Extrait du génome de *Escherichia coli*

Annotation : de la séquence à la fonction

. . .
ctgcaggacgcctactaaggcggcggggaaaaacaaacggttattacaccgagacagaagg
tgcactgcggttatgttgtcgcgacaacggcgaaaaggggctgaccttcgctggtgaacc
aattaagttggcgctatctgantctcatactgtttcacagacctgctgccctgicggcggc
caatcttcctttattcgcttataagcgtggagaattaaaatgcgacatcctttagtgatg
ggtaactggaaactgaacggcagccgacacatggttcacgagctggtttctaacctgcgt
aaagagctggcaggtggtgctggctgtgicggttgcaatcgaccaccggaaatgtatc
gatatggcgaagcgcgaagctgaaggcagccacatcatgctgggtgicgcaaacgtgaac
ctgaacctgtccggcgcattcacgggtgaaacctctgctgctatgctgaaagacatcggc
gcacagtacatcatcatcggctcactctgaacgtcgtacttaccacaaagaatctgacgaa
ctgatcgcaaaaaattcgicggtgctgaaagagcagggcctgactccggttctgtgcatc
ggtgaaaccgaagctgaaaatgaagcgggcaaaactgaagaagtttgicgacgtcagatc
gacgcggtactgaaaactcagggctgctgicggtcattcgaaggtgicggttatcgcttacgaa
cctgtatgggcaatcggctactggcaaatctgcaactccggctcaggcacaggctgttcac
aaattcatccgtgaccacatcgctaaagttgacgctaacatcgctgaacaagtgatcatt
cagtacggcggctccgtaaacgcgtctaacgctgcagaactgtttgctcagccggatc
gacggcgcgctggttggtggtgcttctctgaaagctgacgccttcgicagtaatcgttaaa
gctgcagaagcggctaaacaggcttaagtctgacaggtgccggatttcgatatccggcac
ttactttccttaactcttcgccttaacgcaaaatctcacactgatgatcctgaatttcct
cggctgaagcacggttaagcgtcagtagatttcggtgtgtcgccagcaatacaaatgagt
tatcactctgccgtaccatcgccagcccgtagcgtcccatatggtcccgcgcctcaggta
cttcttctgccagcatcataaatgggctgicggtgtaccagttcgctttccggtaccgac
gicgaggtattcatgcccgcgcaaacctggcagtggaaccagcggctgctgatggt
cgccagattgttatcgag . . .

Extrait du génome de *Escherichia coli*

Gène

atgcgacatcctttagtgatgggtaactggaaactgaacggcagccgccacatggttcac
gagctggtttctaacctgcgtaaagagctggcaggtgttgctggctgtgcggttgcaatc
gcaccaccggaaatgtacatcgatatggcgaagcgtgaagctgaaggcagccacatcatg
ctgggtgcgcaaaacgtggacctgaacctgtccggcgcattcaccggtgaaacctctgct
gctatgctgaaagacatcggcgcacagtacatcatcatcggtcactctgaacgtcgtact
taccacaaagaatctgacgaactgatcgcgaaaaaattcgcggtgctgaaagarcagggc
ctgactccggttctgtgcatcggtgaaaccgaagctgaaaacgaagcgggcaaaactgaa
gaagtttgcgcacgtcaratcgacgcggtactgaaaactcagggtgctgcggcattcgaa
ggtgcggttatcgcttacgaacctgtatgggcaatcggtactggcaaacttgcaactccg
gctcaggcacaggctgttcacaaattcattcgtgaccacatcgctaaagttgacgctaac
atcgctgaacaagtgatcattcagtagcggcggctctgtaaacgcgtctaacgctgcagaa
ctgtttgctcagccagacatcgacggcgcgctgggttggcgggtgcgtctctgaaagctgac
gctttcgcagtaatcgttaaagctgcagaagcggctaacaggcttaa

Protéine

MRHPLVMGNWKLNGSRHMHVHLSNLRKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIM
LGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSERRTYHKESDELIAKKFAVLKEQG
LTPVLCIGETEAENEAGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWAIGTGKSATP
AQAQAVHKFIRDHIAKVDANIAEQVIIQYGGSVNASNAAELEFAQPDIDGALVGGASLKAD
AFAVIVKAAEAAKQA

La protéine de plus près

MRHPLVMGNWKLNGSRHMHVHELVSNLRKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIM

LGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSEERRYHKESEDELIAKKFAVLKEQG

LTPVLCIGETEAENEAGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWAIGTGKSATP

AQAQAVHKEIRDHIAKVDANIAEQVIIQYGGSVNASNAAELFAQPDIDGALVGGASLKAD

AFAVIVKAAEAAKQA

La protéine de plus près

MRHPLVMGNWKLNGSRHMHVHELVSNLRKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIM

LGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSEERRYHKESEDELIAKKFAVLKEQG

LTPVLCIGETEAENEAGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWAIGTGKSATP

[AV]-Y-E-P-[LIVM]-W-[SA]-I-G-T-[GK]

AQAQAVHKEIRDHIAKVDANIAEQVIIQYGGSVNASNAAELFAQPDIDGALVGGASLKAD

AFAVIVKAAEAAKQA

Domaine conservé



Domaine conservé

[AV]-Y-E-P-[LIVM]-W-[SA]-I-G-T-[GK]

- ▷ Signature d'un site actif triosephosphate isomérase
- ▷ Le résidu **E** (glutamate) est actif.

Triosephosphate isomerase (EC 5.3.1.1) (TIM) is the glycolytic enzyme that catalyzes the reversible interconversion of glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. TIM plays an important role in several metabolic pathways and is essential for efficient energy production. It is a dimer of identical subunits, each of which is made up of about 250 amino-acid residues. A glutamic acid residue is involved in the catalytic mechanism. The sequence around the active site residue is perfectly conserved in all known TIM's and can be used as a signature pattern for this type of enzyme. [Extrait de Interpro](#)

La protéine de plus près

MRHPLVMGNWKLNGSRHMHVHELVSNLRLKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIM

LGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSEERRYHKESEDELIAKKFAVLKEQG

LTPVLCIGETEAENEAGKTEEV CARQIDAVLKTQGAAAFEGAVIAYEPVWAIGTGKSATP

[AV]-Y-E-P-[LIVM]-W-[SA]-I-G-T-[GK]

AQAQAVHKEIRDHIAKVDANIAEQVIIQYGGSVNASNAAELFAQPDIDGALVGGASLKAD

AFAVIVKAAEAAKQA

Domaine conservé



La protéine de plus près

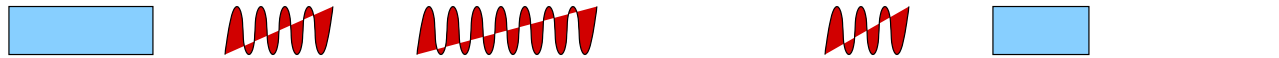
MRHPLVMGNWKLNGSRHMHVHELVSNLRKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIM



LGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSEERRYHKESEDELIAKKFAVLKEQG



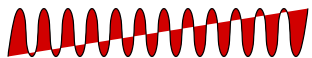
LTPVLCIGETEAENEAGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWAIGTGKSATP



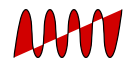
AQAQAVHKEIRDHIAKVDANIAEQVIIQYGGSVNASNAAELFAQPDIDGALVGGASLKAD



AFAVIVKAAEAAKQA



Eléments de la structure secondaire



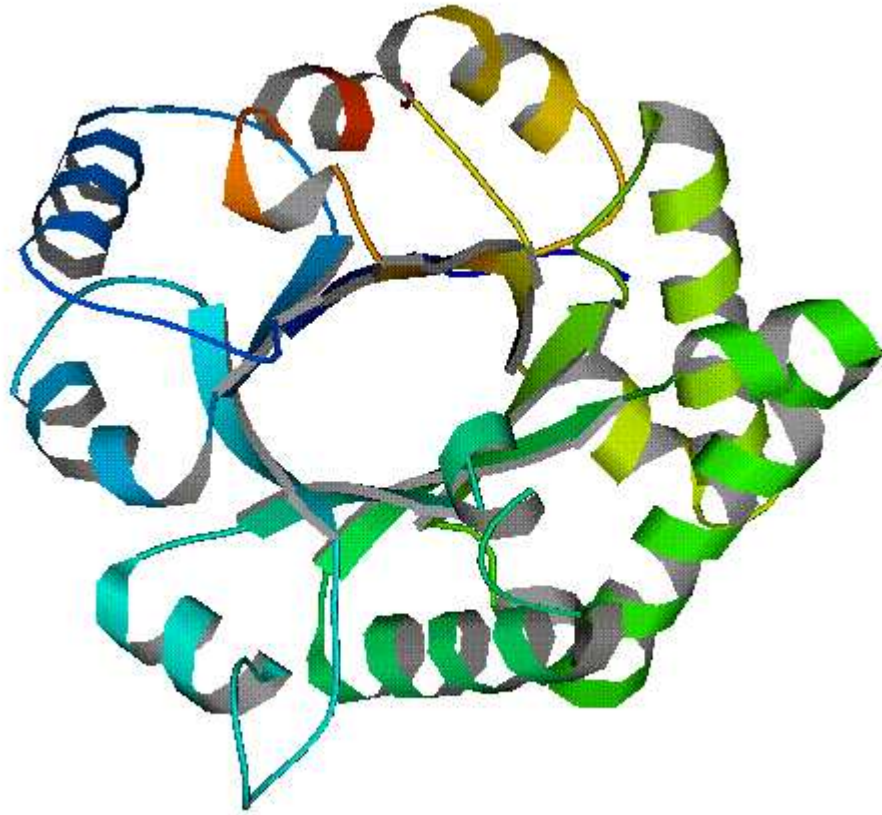
hélice alpha



feuillet beta

Domaine conservé





Structure 3D de la Triosephosphate Isomérase : Organisation des éléments de structure secondaire dans l'espace

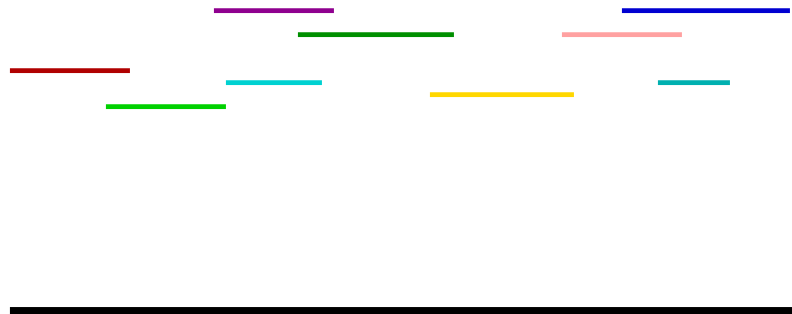
Contenu du cours

- ▷ Comparaison avec d'autres séquences : **Alignement global, local, alignement multiple**
- ▷ Annotation : **Recherche de motifs exacts, approchés**
- ▷ Prédiction de gènes : **procaryotes**
- ▷ Stockage et représentation des séquences : **Indexation, arbre des suffixes**
- ▷ Prédire la structure 3D à partir de la séquence : **ARN, protéines**

Séquençage: Obtenir la séquence

Le séquençage commence par la fragmentation de la séquence.

- ▷ *Contig* : fragment contigu d'ADN
(quelques centaines de bases)
- ▷ *Assemblage* : retrouver la séquence initiale



Assemblage de contigs

Formalisation du problème : la plus courte super-chaîne commune

Données : une collection \mathcal{F} de fragments

Résultat : la plus courte chaîne S telle que tout fragment f de \mathcal{F} soit un facteur de S

Exemple :

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

Assemblage de contigs

Formalisation du problème : la plus courte super-chaîne commune

Données : une collection \mathcal{F} de fragments

Résultat : la plus courte chaîne S telle que tout fragment f de \mathcal{F} soit un facteur de S

Exemple :

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

Le problème de la plus
courte super-chaîne
commune est NP-complet

Heuristique :

modélisation du problème par un graphe

- nœuds : fragments à assembler
- arcs : pondérés par la longueur du chevauchement maximal

Un chemin hamiltonien décrit une super-chaîne

Idée de stratégie: *privilégier les arcs les plus lourds*

▷ Tri des arcs par poids décroissant

▷ Construction du chemin :

sélection à chaque étape de l'arc disponible le plus lourd.

Un seul arc entrant et un seul arc sortant par nœud.

▷ Arrêt : le chemin est hamiltonien

Stratégie gloutonne: *la solution est construite de manière incrémentale, sans possibilité de remettre en cause les choix précédents.*

Confrontation du modèle à la réalité biologique

- ▷ De quel brin provient le fragment ?

Prise en compte des fragments inversés complémentés

- ▷ Erreurs de séquençage

Autoriser des erreurs dans les chevauchements

- ▷ Nombreuses répétitions successives

Repérage préalable des zones répétées

- ▷ Génome humain : séquences *Alu*

Filtrage des données