

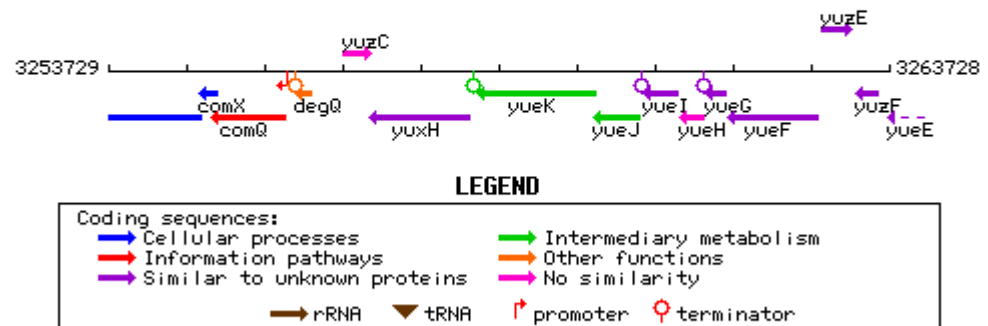
Prédiction de gènes

Cours de présentation des outils bio-informatiques
pour la localisation puis l'étude des gènes

Equipe Bonsai (2014)

La localisation des gènes

- C'est la première étape pour interpréter un génome
 - Distinction entre régions codantes et non codantes
- Réalisée par des programmes informatiques combinant différents types d'informations
- Ces programmes sont prédictifs, ils génèrent des erreurs
 - Certains gènes échappent à la détection (faux négatifs)
 - Certains gènes prédits ne correspondent pas à de vrais gènes (faux positifs)
 - Même pour les prédictions correspondant à des gènes réels, les limites précises du gène sont parfois erronées



<http://genolist.pasteur.fr/SubtiList/>

Quel est le point de départ ?

- La séquence d'ADN est produite brute
 - Pas d'information sur la position des gènes, ...
 - Besoin de « décoder » le message du génome
- Les expériences en laboratoire fournissent de nombreuses données
 - Etude d'un gène et de son produit (fonction de la protéine)
 - Extraction d'ARNm (ou de fragments : EST)
 - Nombreuses publications et informations dans les banques
- Possibilité de croiser les informations pour améliorer la qualité des annotations

PRÉSENTATION DES MÉTHODES

Les méthodes de prédiction de gènes

- Détection des ORF (Open Reading Frame)
 - Méthode naïve
 - Localisation des régions de plus de 99 bp entre un codon d'initiation (Cinit) et un codon de terminaison (Cterm)
- Comparaison aux banques
 - Méthode exploitant les données disponibles
 - Recherche des séquences d'ARNm et de protéines qui ressemblent à la séquence étudiée
- Etude statistique (ab initio)
 - Localisation des séquences codantes et non codantes sur la base de critères discriminants

Une idée naïve : les phases ouvertes de lecture


- Une séquence codante :
 - Débute par un codon d'initiation (ATG + autres) et se termine par un codon de terminaison (TAA, TAG, TGA)
 - A une taille multiple de 3 (si les introns sont enlevés)
 - Taille moyenne : 1.000 bp (bactéries)
- Une phase ouverte de lecture (ORF)
 - Plus de 99 nt entre un Cinit et un Cterm (statistiquement rare)
 - Peut contenir un gène
- Problèmes :
 - Un gène peut être sur un brin ou sur l'autre
 - Plusieurs phases de lecture possibles

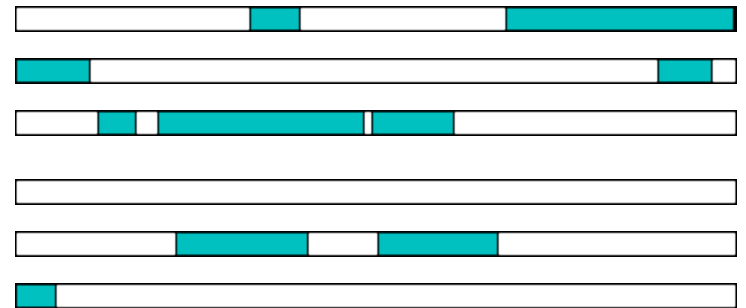
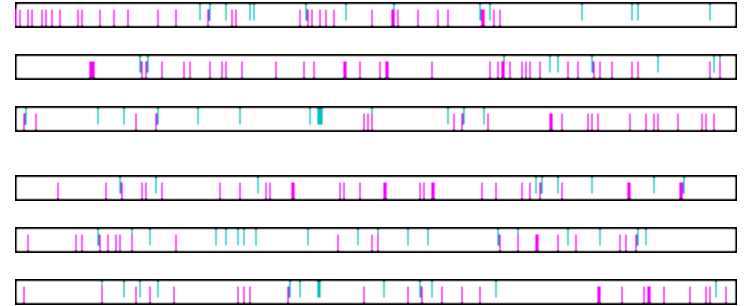
Détection des ORF, fonctionnement

- Traduction à l'aveugle
 - 6 phases de lecture
 - = 6 séquences protéiques possibles



Détection des ORF, résultats (OrfFinder)

- 6 phases de lectures :
 - | codons d'initiation (ATG)
 - | codons de terminaison (TAA, TAG, TGA)
- Sélection des phases ouvertes de lecture (ORF)
 - Régions mesurant plus de 99 nt entre un Cinit et un Cterm
 - Choix du Cinit le plus loin du Cterm
 - Peut contenir un gène
 -  une ORF



ATTENTION : ORF ne veut pas dire gène !

Détection des ORF, bilan

- Cette définition n'est pas suffisante pour discriminer les séquences codantes des séquences non codantes
 - Toutes les ORF ne sont pas des gènes
- Méthode utilisée pour découper des régions d'intérêt
 - Point de départ pour d'autres analyses
 - Diminue la quantité de séquences à analyser
 - Utile pour l'annotation automatique
- Limites
 - Très sensible aux erreurs de séquençage
 - Certains gènes peuvent être manqués
 - Un gène avec introns s'étend sur plusieurs ORF

Comparaison de séquences

- Nombreuses séquences de protéines dans les banques
 - Comparaison de l'ADN aux protéines pour trouver des protéines de même fonction
 - Détermination des positions de début et de fin de la séquence codante, ainsi que des introns
- Possibilité d'isoler puis séquencer un ARNm (in vivo)
 - Comparaison de l'ARNm au génome pour localiser le gène
 - Détermination des positions de début et de fin du gène, ainsi que des introns (car ARNm mature)

Comparaison aux banques protéiques

- Utilisation de Blast (ou autre) contre une banque protéique
- A partir de la traduction des ORF (BlastP)
 - Vérifie si l'ORF contient un gène
 - Aide au choix du codon d'initiation (comparaison de la taille de l'ORF traduite avec celle des protéines similaires)
- A partir de la séquence nucléique (BlastX)
 - Localise les CDS, même si des erreurs de séquence introduisent des décalages de phase ou des codons stop prématurés.
- Limites :
 - Les séquences orphelines ne sont pas vues
 - Séquences sans homologue dans la banque
 - Les séquences atypiques sont difficiles à trouver
 - Séquences ayant une composition éloignée par rapport à leurs homologues

Comparaison aux banques nucléiques

- Utilisation de BlastN contre une banque nucléique
- Détection de séquences contaminantes
 - Vecteur, ...
 - Logiciels spécialisés : VecScreen, EMVEC
 - Recherche contre une banque de vecteurs
- Détection des régions 3' et 5' UTR (meilleur moyen de déterminer le début et la fin de la transcription)
 - Comparaison aux ARNm et EST de l'organisme étudié ou d'organismes proches
- Limites
 - Les EST contiennent des erreurs et sont délicates à exploiter
 - ARNm et EST ne sont pas connus pour tous les gènes

Détermination précise des positions gène/CDS

- Blast n'est pas optimisé pour aligner deux séquences
 - Son objectif n'est pas d'aligner un gène à sa protéine, mais de faire le tri entre des séquences similaires ou non
 - Nécessité d'aligner avec un autre logiciel la séquence étudiée à la ou les séquences similaires de la banque
- Logiciels spécialisés
 - SIM4, EST2Genome : aligne 1 séquence génomique à 1 ARNm
 - WISE2 : aligne 1 séquence nucléique à 1 protéine
 - SPALN : aligne 1 ou plusieurs protéine(s) ou ARNm à une séquence génomique
 - Scipio : aligne une protéine à un génome (choisi dans une liste) et recherche le gène correspondant
 - ...

Prédiction statistique

- Apprentissage de l'usage du code pour un organisme donné à partir d'un ensemble fiable de séquences codantes
- Détermination de classes de gènes avec des usages du code différents au sein de l'organisme
- Calcul de la probabilité pour qu'une fenêtre soit codante
 - Une fenêtre est une suite de lettres dans une séquence
- Analyse des résultats obtenus en faisant coulisser la fenêtre le long de la séquence étudiée

Usage du code

- N codons codent le même acide aminé
 - codons synonymes
- Pour un aa donné, il y a un ou des codons préférés
- Exemple : gène *cytB* de *Plasmodium falciparum*
 - GC = 27.59%

AA	Codon	Fraction	Number
A	GCA	0.647	11
	GCC	0.000	0
	GCG	0.000	0
	GCT	0.353	6
F	TTC	0.206	7
	TTT	0.794	27
G	GGA	0.500	11
	GGC	0.000	0
	GGG	0.045	1
	GGT	0.455	10

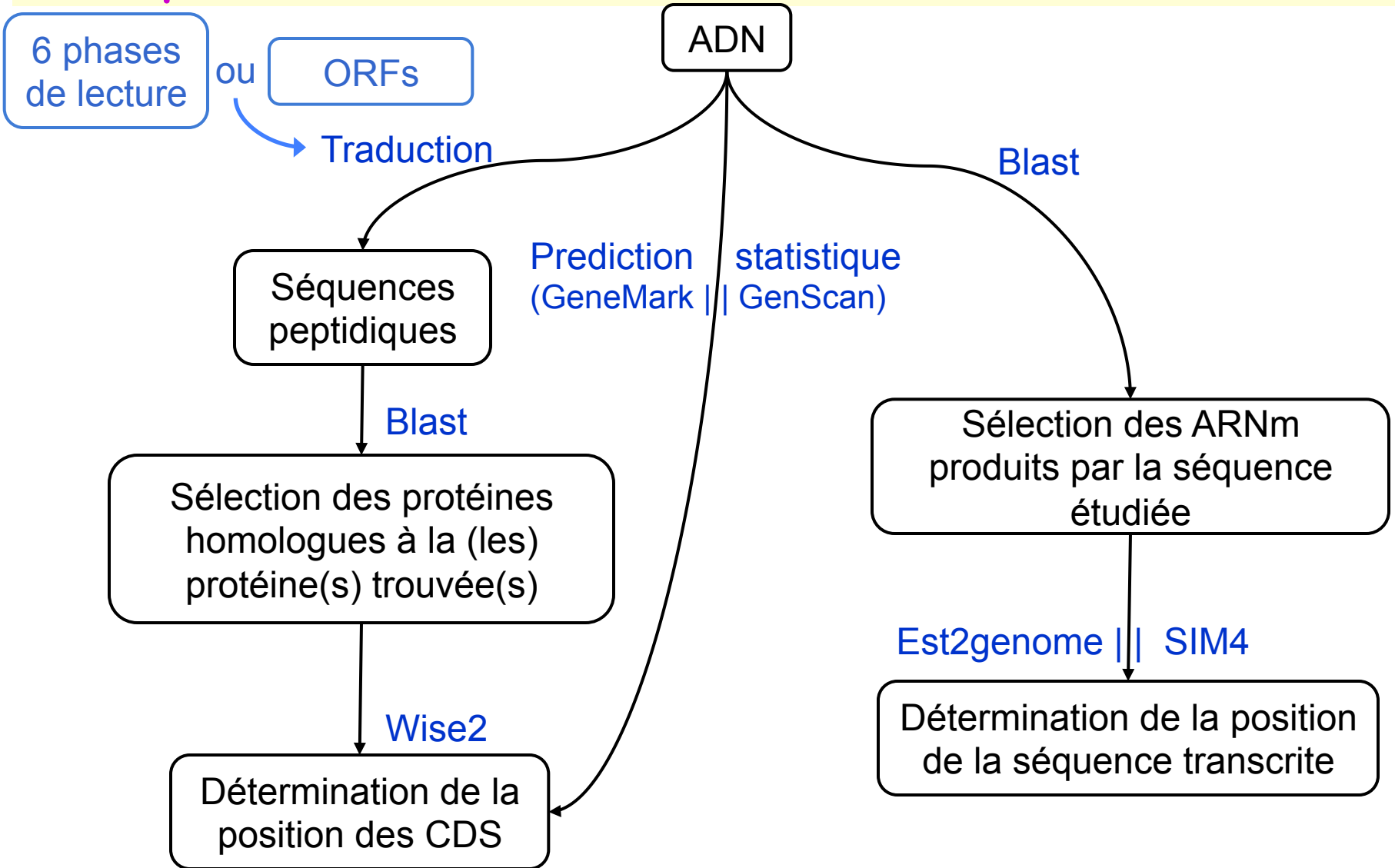
Biais d'usage du code (cas des bactéries)

- Différences entre organismes selon leur %G+C
 - ↳ Choix des codons riches en GC dans les génomes riches en GC
 - ↳ Les doubles hélices d'ADN riches en GC sont plus stables
- Différences entre gènes selon leur taux d'expression (classe)
 - ↳ Les gènes « de ménage » (nécessaires au fonctionnement de toutes les cellules) partagent le même usage du code
 - ↳ Les autres gènes ont un usage différent
- Les séquences codantes suivent l'usage du code de leur organisme et de leur classe
- Les séquences non codantes n'ont pas de pression de sélection pour respecter l'usage du code

Prédiction statistique, limites

- Besoin d'un jeu d'apprentissage propre à chaque organisme
 - Pas disponible pour tous les organismes séquencés
 - Cas particulier du modèle heuristique de GeneMark pour les procaryotes
 - Jeu d'apprentissage construit à partir de plusieurs génomes
 - Biais caractéristique pour des séquences dans un intervalle de %gc
- Pas de détection des petits gènes ou petits exons
 - Limite due au seuil de détection des programmes

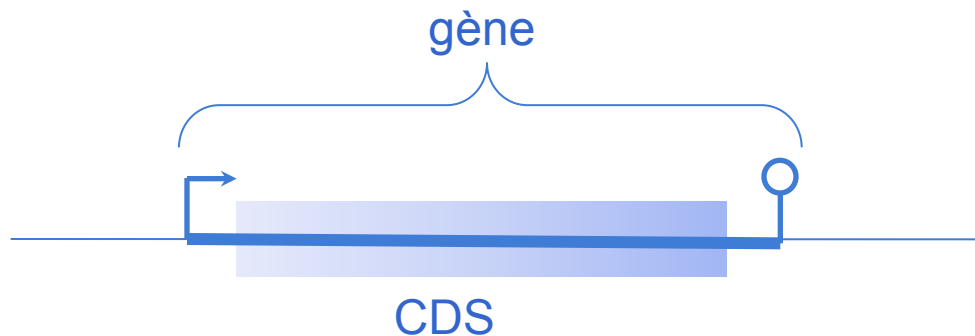
Récapitulatif



EXEMPLE D'ANALYSE D'UN ADN PROCARYOTE

Prédiction chez les bactéries : simplicité ?

- Plus de 80% du génome est codant
 - Séquences intergéniques courtes
 - En moyenne : un gène pour 1.000 nucléotides (kb)
- Structure simple des gènes
 - Régions transcrites mais non traduites (3' et 5' UTR) courtes
 - Pas d'intron (sauf exception)
- Détection possible par
 - Traduction de la séquence des ORF
 - Comparaison du peptide aux banques protéiques
 - Croisement avec les prédictions statistiques

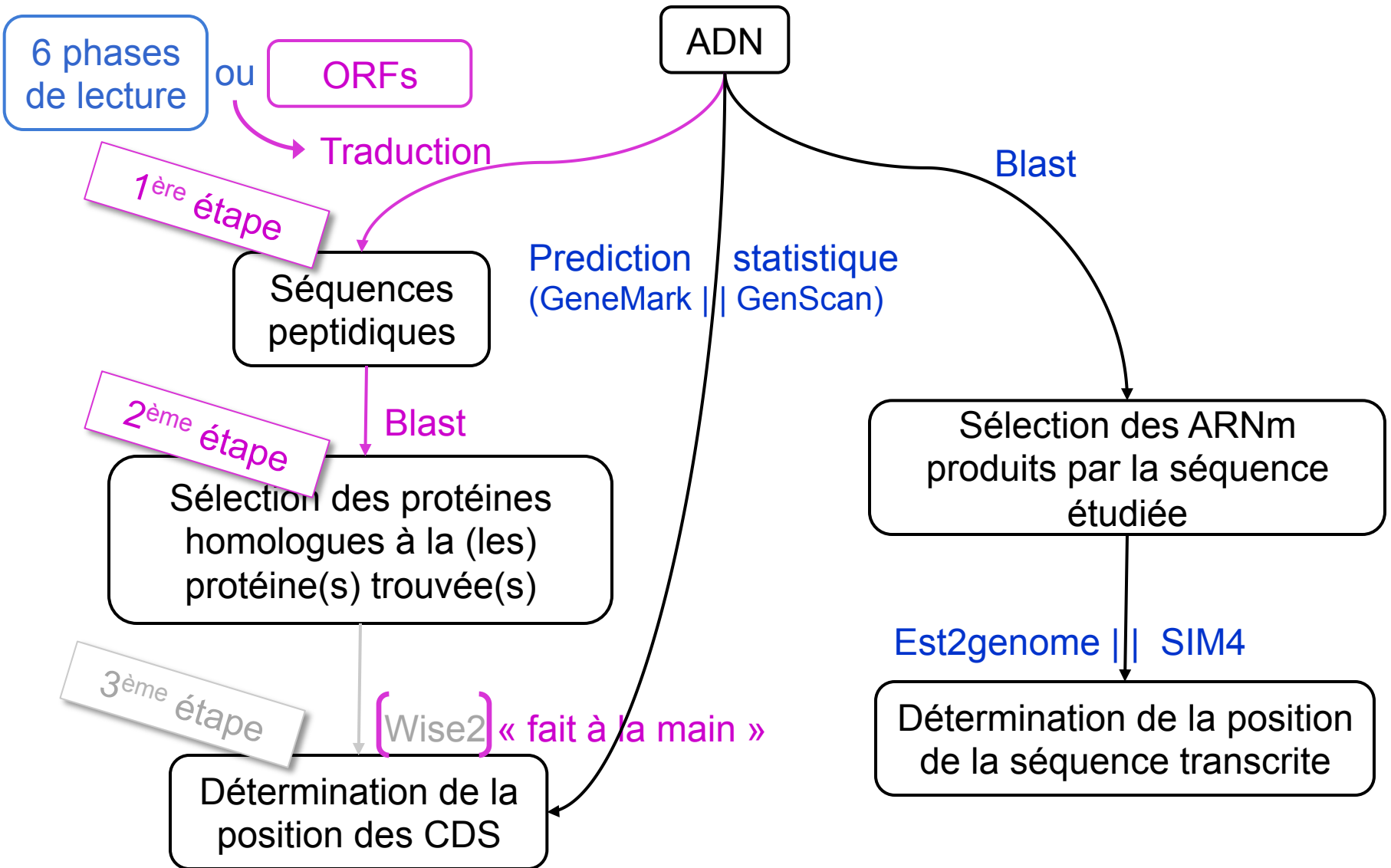


Un exemple

- Voici un extrait du génome de la bactérie *Pseudoalteromonas sp.*

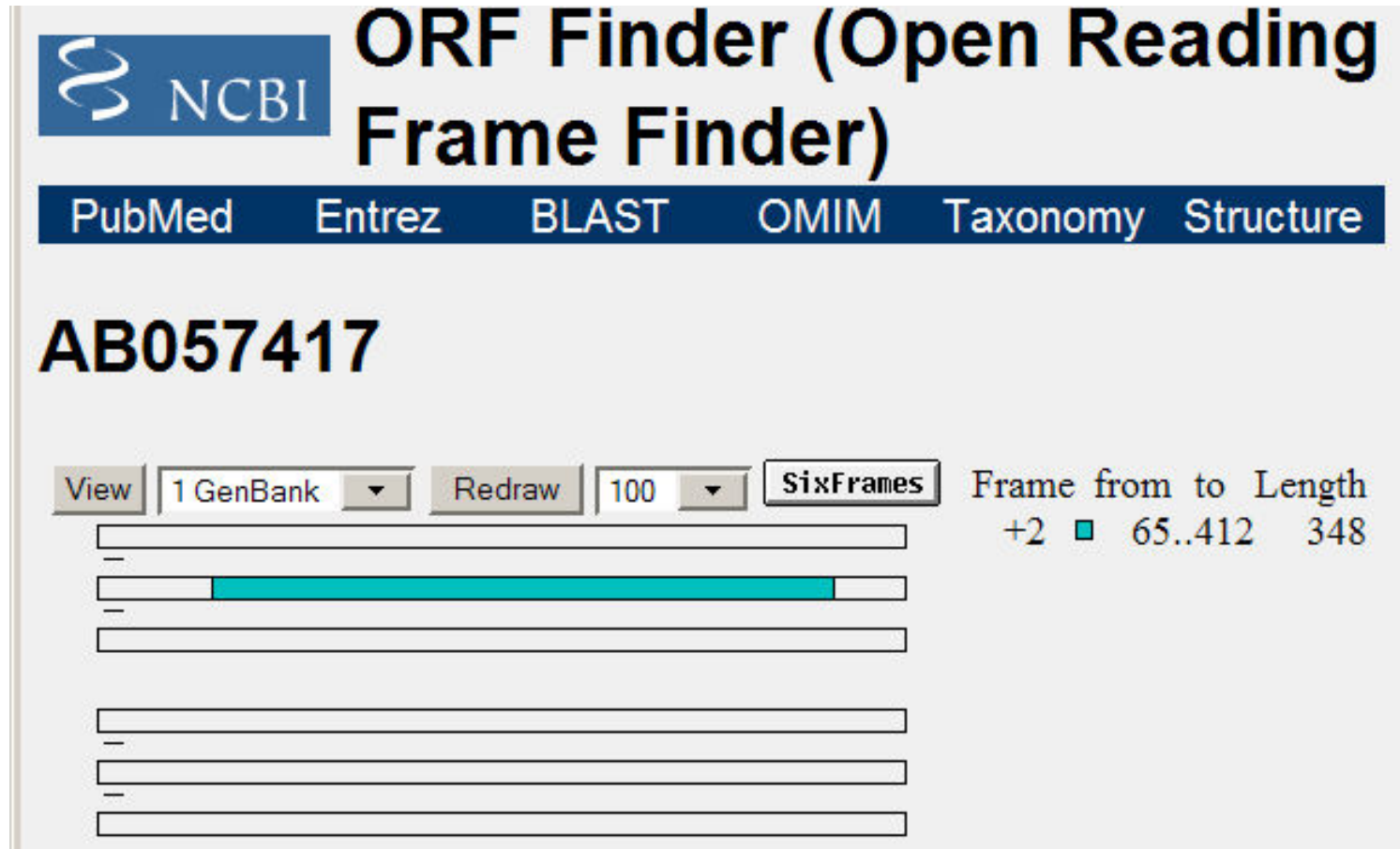
>AB057417

```
aacgaaaagattaaaaatattatcatttttttctcttggaatttttttactctacccccatta  
atgaatgcaaattagaaaaagctttttttctgtactgttcagaaactgttaggagaactaaa  
aaacatgaacattcgtcctttacaagatcgcgtaatcgttaaacgtctagaagaagaac  
aaaatctgctggcggtattgtattaactggctctgcagctgaaaaatcaactcgcgggaga  
agtagtagccgtaggtaatggtcgtattttagataacggtgacgttagagctttagaagt  
aaaagccggtgacactgtggttatttggctcatatgttgagaaaactgaaaagatcgaagg  
tcaagagtacctgatcatgcgtgaagacaacattttaggcattgtaggctaagcctactt  
ttcgtttaacacacatttaagaatttagagg
```



1ère étape : détection des ORF

- Une seule phase ouverte de lecture sur cette séquence
 - Sûrement un seul gène (voir aucun)



2^{ème} étape : comparaison de séquences

En sélectionnant une ORF, il est possible de lancer un BlastP pour connaître les protéines de la banque qui ressemblent à la traduction de l'ORF

NCBI ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxonomy Structure

AB057417

Program Database ☐ with parameters

View Frame from to Length
+2 ■ 65..412 348

Length: **115 aa**

```
65 atgcaaatagaaaaagcctttttctgtactgttcagaaactgtta
M Q I R K A F F C T V Q K L L
110 ggagaactaaaaaacatgaacattogtctttacaagatcgcgta
G E L K N M N I R P L Q D R V
155 atogttaaactgtctagaagaagaacaaaatctgctggcggtatt
I V K R L E E E T K S A G G I
200 gtattaactggctctgcagctgaaaaatcaactcgcggagaagta
V L T G S A A E K S T R G E V
245 gtagccgtaggtaaatggctgtatttttagataacggtgacggttaga
V A V G N G R I L D N G D V R
290 gctttagaagtaaaagcgggtgacactgtgttatttggctcatat
A L E V K A G D T V L F G S Y
335 gttgagaaaactgaaaagatcgaaggtaagagtagctgatcatg
V E K T E K I E G Q E Y L I M
380 cgtgaagacaacatttttaggcattgtaggctaa 412
R E D N I L G I V G *
```


Résultats de BlastP : « Graphic Summary »

Id|27512 (115 letters)

Query ID Id|27512
Description None
Molecule type amino acid
Query Length 115

La séquence soumise,
les paramètres de Blast
et la banque interrogée

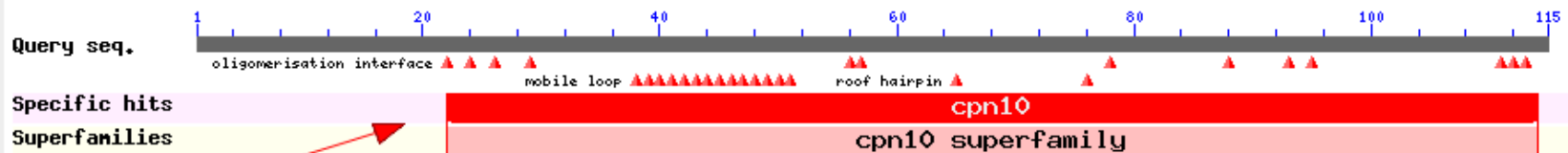
Database Name nr
Description All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding
environmental samples from WGS projects
Program BLASTP 2.2.18+ [Citation](#)

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Related Structures](#)

▼ Graphic Summary

▼ Show Conserved Domains

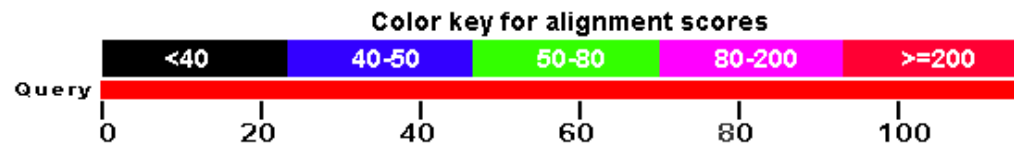
Putative conserved domains have been detected, click on the image below for detailed results.



Liste des domaines protéiques
trouvés sur la traduction de l'ORF

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments



Liste des protéines similaires
à la traduction de l'ORF



Résultats de BlastP : « Description »

▼ Descriptions

Sequences producing significant alignments:		Score (Bits)	E Value	
ref ZP_01611291.1 	10 kDa chaperonin (Protein Cpn10) (groES p...	184	2e-45	
ref YP_338802.1 	chaperonin [Pseudoalteromonas haloplanktis T...	176	4e-43	G
ref ZP_01135610.1 	10 kDa chaperonin (Protein Cpn10) (groES p...	172	9e-42	
ref YP_156662.1 	co-chaperonin GroES [Idiomarina loihiensis L...	135	1e-30	G
ref ZP_01448233.1 	co-chaperonin GroES [alpha proteobacterium...	134	2e-30	
ref YP_002128203.1 	co-chaperonin GroES [Alteromonas macleodi...	132	6e-30	G
ref YP_692353.1 	chaperonin, 10 kDa [Alcanivorax borkumensis ...	132	8e-30	G
ref ZP_01042901.1 	co-chaperonin GroES [Idiomarina baltica OS...	132	8e-30	
ref YP_942285.1 	chaperonin Cpn10, GroES, small subunit of Gr...	132	1e-29	G
ref ZP_01217218.1 	GroES [Psychromonas sp. CNPT3] >gb EAS3795...	132	1e-29	
ref NP_716336.1 	co-chaperonin GroES [Shewanella oneidensis M...	130	4e-29	G
gb EDX87994.1 	chaperonin GroS [Alcanivorax sp. DG881]	129	5e-29	
ref YP_943819.1 	chaperonin Cpn10, GroES, small subunit of Gr...	129	5e-29	G
ref YP_267705.1 	co-chaperonin GroES [Colwellia psychrerythra...	129	7e-29	G
ref YP_579807.1 	co-chaperonin GroES [Psychrobacter cryohalol...	129	7e-29	G
gb ACA50470.1 	GroES [Xenorhabdus nematophila]	129	9e-29	
ref YP_663308.1 	chaperonin Cpn10 [Pseudoalteromonas atlantic...	129	9e-29	G
ref YP_561437.1 	co-chaperonin GroES [Shewanella denitrifican...	128	1e-28	G
ref ZP_01161756.1 	co-chaperonin GroES [Photobacterium sp. SK...	128	2e-28	
ref YP_964836.1 	co-chaperonin GroES [Shewanella sp. W3-18-1]...	128	2e-28	G
ref YP_001143151.1 	chaperonin GroS [Aeromonas salmonicida su...	127	2e-28	G
ref YP_001340893.1 	chaperonin Cpn10 [Marinomonas sp. MWYL1] ...	127	2e-28	G
ref YP_855401.1 	chaperonin GroS [Aeromonas hydrophila subsp....	127	2e-28	G
ref YP_263847.1 	co-chaperonin GroES [Psychrobacter arcticus ...	127	2e-28	G
emb CAA30738.1 	unnamed protein product [Escherichia coli]	127	2e-28	G

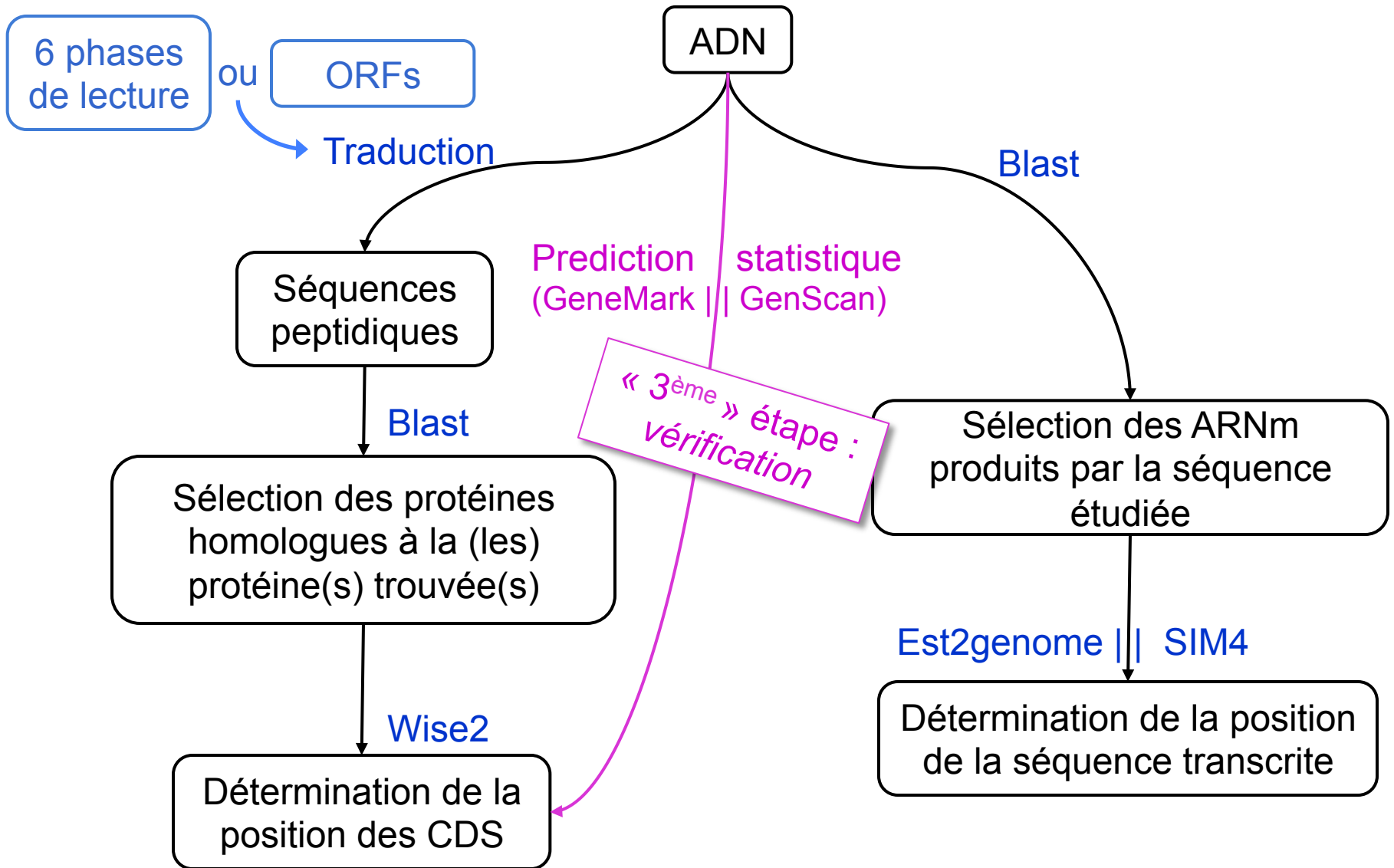
Résultats de BlastP : « Alignments »

▼ **Alignments** ☐ Select All [Get selected sequences](#) [Distance tree of results](#)


```
>[ref|ZP_01611291.1| 10 kDa chaperonin (Protein Cpn10) (groES protein) [Alteromonadales  
bacterium TW-7]  
[dbj|EAB39464.1| GroES [Pseudoalteromonas sp. PS1M3]  
[gb|EAW29422.1| 10 kDa chaperonin (Protein Cpn10) (groES protein) [Alteromonadales  
bacterium TW-7]  
Length=95  
  
Score = 184 bits (466), Expect = 2e-45  
Identities = 95/95 (100%), Positives = 95/95 (100%), Gaps = 0/95 (0%)  
  
Query 21 MNIRPLQDRVIVKRLEEETKSAGGIVLTGSAAEKSTRGEVVAVGNRILDNGDVRALEVK 80  
MNIRPLQDRVIVKRLEEETKSAGGIVLTGSAAEKSTRGEVVAVGNRILDNGDVRALEVK  
Sbjct 1 MNIRPLQDRVIVKRLEEETKSAGGIVLTGSAAEKSTRGEVVAVGNRILDNGDVRALEVK 60  
  
Query 81 AGDTVLFSGSYVEKTEKIEGQEYLIMREDNILGIVG 115  
AGDTVLFSGSYVEKTEKIEGQEYLIMREDNILGIVG  
Sbjct 61 AGDTVLFSGSYVEKTEKIEGQEYLIMREDNILGIVG 95  
  
>[ref|YP_338802.1| [G] chaperonin [Pseudoalteromonas haloplanktis TAC125]  
[sp|Q9AKT2|CH10 PSEHT [G] 10 kDa chaperonin (Protein Cpn10) (groES protein)  
[emb|CAC28359.1| [G] groES protein [Pseudoalteromonas haloplanktis TAC125]  
[emb|CAI85359.1| [G] 10 kDa chaperonin (Protein Cpn10) (groES protein) [Pseudoalteromonas  
haloplanktis TAC125]  
Length=95  
  
[GENE ID: 3707997 groS | chaperonin [Pseudoalteromonas haloplanktis TAC125]  
(10 or fewer PubMed links)  
  
Score = 176 bits (447), Expect = 4e-43  
Identities = 91/95 (95%), Positives = 93/95 (97%), Gaps = 0/95 (0%)  
  
Query 21 MNIRPLQDRVIVKRLEEETKSAGGIVLTGSAAEKSTRGEVVAVGNRILDNGDVRALEVK 80  
MNIRPLQDRVIVKRLEEETKSAGGIVLTGSAAEKSTRGEVVAVGNRIL++GDVRALEVK  
Sbjct 1 MNIRPLQDRVIVKRLEEETKSAGGIVLTGSAAEKSTRGEVVAVGNRILESGDVRALEVK 60  
  
Query 81 AGDTVLFSGSYVEKTEKIEGQEYLIMREDNILGIVG 115  
AGDTVLFSGSYVEKTEKIEGQEYLIMREDNILGIVG  
Sbjct 61 AGDTVLFSGSYVEKTEKIEGQEYLIMREDNILGIVG 95
```

Interprétation

- ORF : 65..412 sur le brin + de la séquence ADN
 - Code une protéine de 115 aa + le codon de terminaison
- La protéine codée par l'ORF contient un domaine Cpn10
 - Cnp10 : Chaperonin 10 Kd subunit
- Alignements fournis par BlastP :
 - Query: 21..115 : seulement une région de la protéine de l'ORF
 - ⇒ L'ORF entière n'est pas codante
 - ⇒ L'alignement commence en 21 => la séquence codante commence sûrement en $65 + (21-1)*3 = 125$
 - ⇒ Fin de la séquence codante en 412
 - Sbjct: 1..95 : la protéine de la banque est entière
 - ⇒ La séquence codante prédite est sûrement complète
 - Les alignements obtenus avec différentes séquences sont bons
 - ⇒ La prédiction est fiable



3^{ème} étape : prédiction statistique (GeneMark.hmm)

Sequence title: groES

Length: 451 bp

G+C percentage: 36.36 %

GeneMark.hmm PROKARYOTIC (Version 2.4a)

Model organism: Heuristic_model

Predicted genes

Gene	Strand	LeftEnd	RightEnd	Gene Length	Class
1	+	<1	75	75	1
2	+	65	412	348	1

3^{ème} étape : prédiction statistique (GeneMark v2.4)

List of Open reading frames predicted as CDSs, shown with alternate starts (regions from start to stop codon w/ coding function >0.50)

Left End	Right end	DNA Strand	Coding Frame	Avg Prob	Start Prob
-----	-----	-----	-----	----	----
65	412	direct	fr 2	0.65
125	412	direct	fr 2	0.79	0.07
317	412	direct	fr 2	0.59	0.12

Interprétation

- GeneMark.hmm : 2 CDS prédites
 - Une CDS incomplète et très courte
 - Sûrement un faux positif car à la limite de la significativité
 - Une CDS dont les positions couvrent l'ORF dans sa totalité
- GeneMark v2.4 : 1 CDS prédite avec 3 débuts possibles
 - La meilleure probabilité moyenne pour que la séquence soit codante est obtenue pour le début en 125.
 - Les probabilités pour que le début proposé soit le vrai codon d'initiation sont très faibles (mauvaises).

GeneMark v2.4 soutient le résultat précédemment trouvé

⇒ les positions de la CDS sont sûrement 125..412

Bilan

- Les différents logiciels aboutissent tous à la même conclusion :
 - La séquence contient une seule séquence codante allant de 125 à 412 (codon de terminaison compris)
- Information supplémentaire donnée par Blast :
 - La CDS code une protéine de la famille des chaperonnes (du type Cpn10 / GroES)
- L'annotation de cette séquence dans la banque DDBJ confirme ces conclusions

Prédiction chez les bactéries : quelques pièges

- Plusieurs Cinit (AUG) sur la séquence : lequel prendre ?
- Possibilité de Cinit alternatifs (GUG, UUG)
- ➡ Confirmation par :
 - Présence de RBS (*Ribosome Binding Site*)
 - Comparaison (*analyse comparative avec autres espèces*)
 - Prédiction statistique
- Gènes incomplets (Cterm prématuré, décalage de phase)
 - Réel (corrigé lors de la traduction, pseudogènes)
 - Erreurs de séquençage
- ➡ BlastX signale des incohérences (phases différentes)
 - Comparaison + Prediction
- Gènes chevauchants
 - Fréquent chez les virus, quelquefois sur bactéries (fins de gènes)

EXEMPLE D'ANALYSE D'UN ARNm EUCARYOTE

Complexité des génomes eucaryotes

- Faible pourcentage de séquences codant pour des protéines
 - Environ 2% du génome humain
- Structure complexe des gènes
 - Longues régions 3' et 5' non traduites (exons non codants)
 - Présence d'introns, épissage alternatif



- Exons non codants (5' et 3' UTR -UnTranslated Region-)
- Exons codants (CDS)
- - Introns

Difficulté de prédiction des gènes avec introns

- Taille des introns non multiple de 3
 - Changement de phase d'un exon à l'autre
 - Pas de changement de brin
- Existence d'exons courts (~10nt)
 - Taille en dessous des limites de résolution des logiciels
- Existence d'introns très longs (plus longs que les exons)
 - Difficulté pour localiser les exons (ils sont noyés)
- Un intron peut couper un codon en deux
- Epissage alternatif
 - Concerne environ > 50% des gènes humains (estimation 2010 – avant 2008, hypothèse de < 30%)

Un exemple : étude d'un ARNm

- Deux études à faire :

1. Recherche de la CDS de l'ARNm

- Méthodes classiques vues précédemment
- Puis, étude de la fonction de la protéine codée par l'ARNm
 - Voir *cours suivant* sur « annotation de protéines »

2. Localisation du gène codant l'ARNm

- Comparaison de séquence contre le génome complet
 - Blast, section « Genomes »
- Exemple choisi :
 - ARNm de 938 bp, issu d'une cellule humaine

Un exemple : étude d'un ARNm

- Deux études à faire :

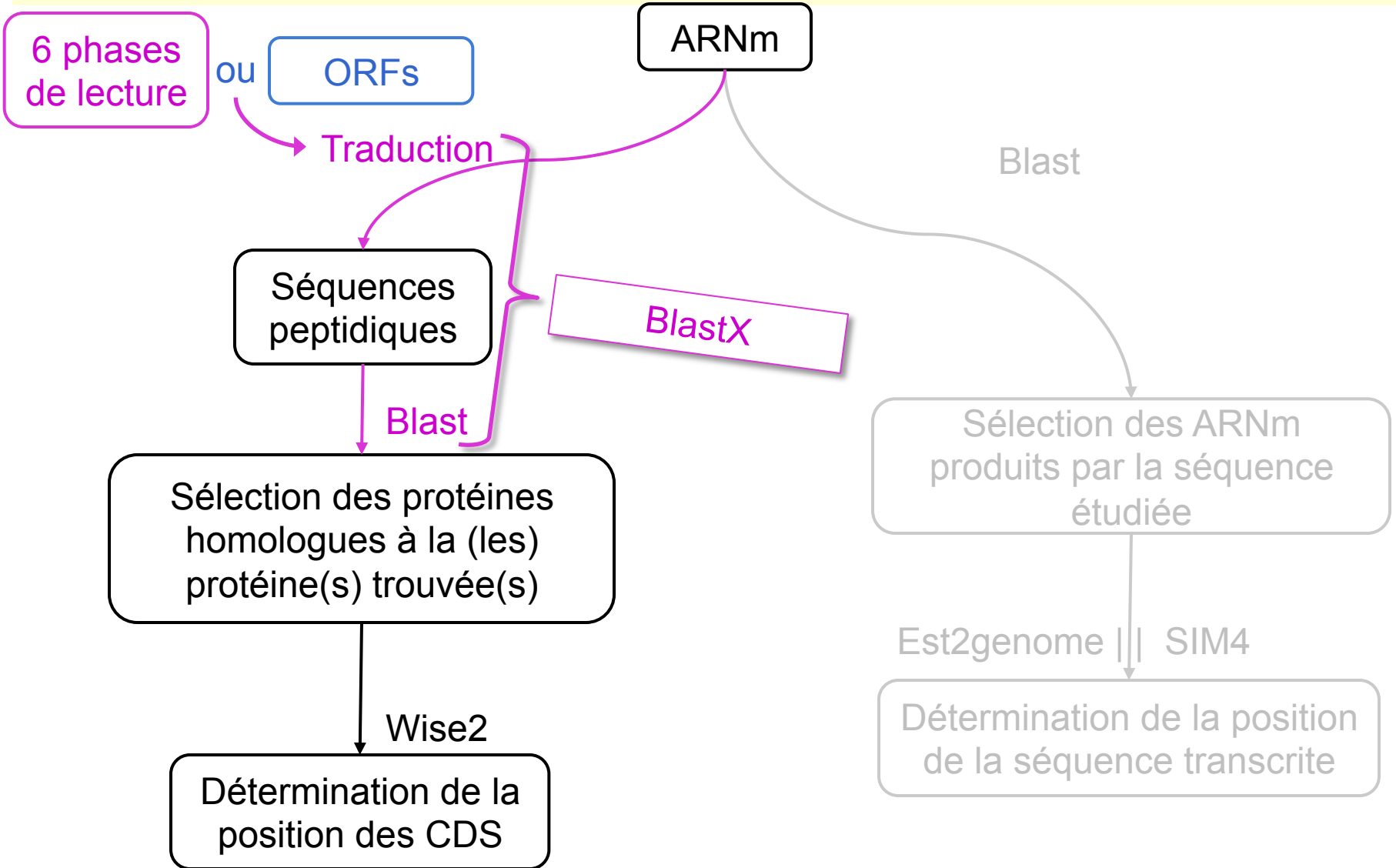
1. Recherche de la CDS de l'ARNm

- Méthodes classiques vues précédemment
- Puis, étude de la fonction de la protéine codée par l'ARNm
 - Voir *cours suivant* sur « annotation de protéines »

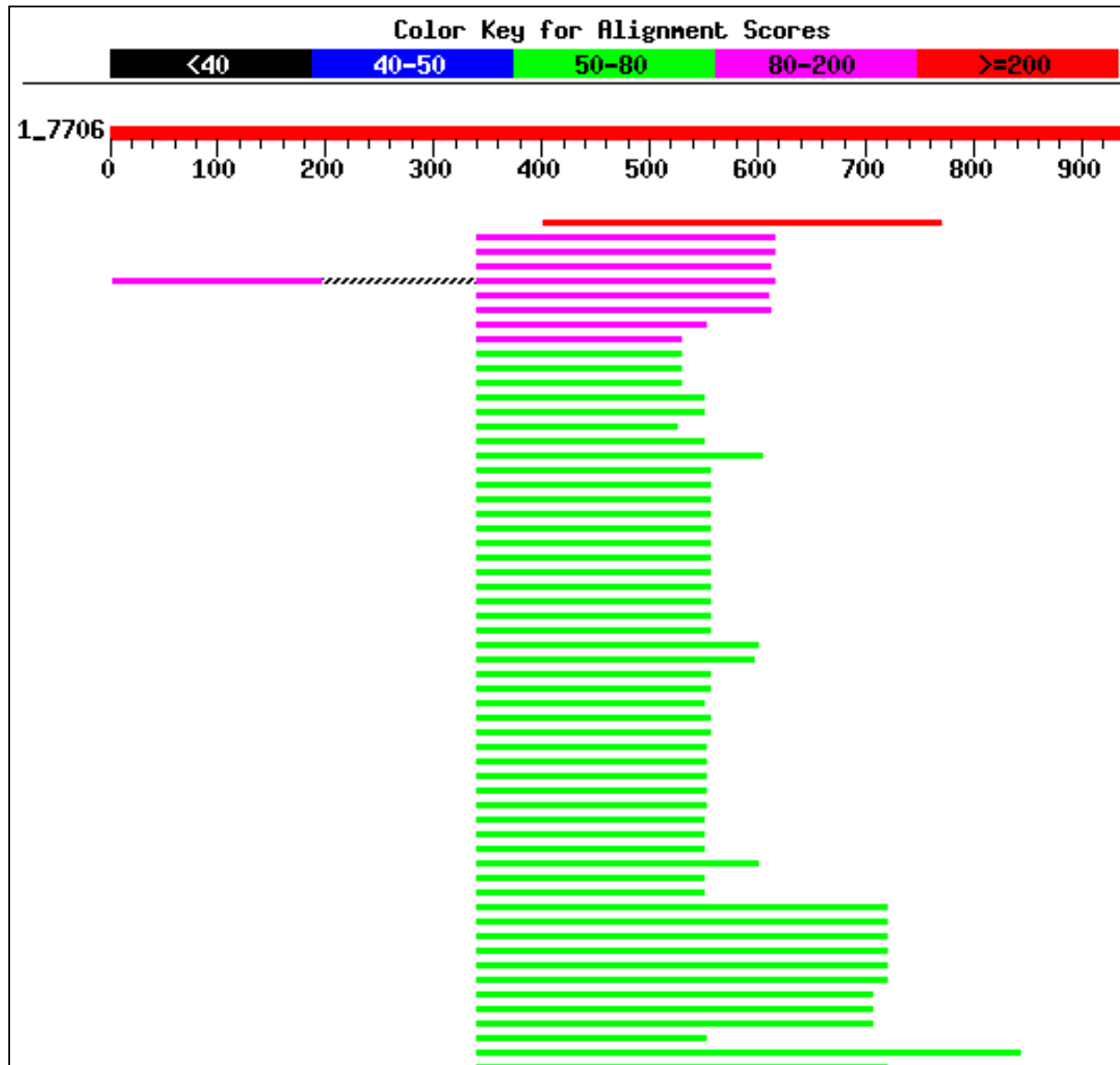
2. Localisation du gène codant l'ARNm

- Comparaison de séquence contre le génome complet
 - Blast, section « Genomes »

1° Recherche de la CDS de l'ARNm (1/2)



BlastX, représentation graphique des résultats



BlastX, alignement avec la 1ère entrée trouvée

```
>gi|55641083|ref|XP_529628.1|    PREDICTED: hypothetical protein XP_529628  
[Pan troglodytes]
```

```
Length = 155
```

```
Score = 227 bits (578), Expect = 4e-58
```

```
Identities = 109/123 (88%), Positives = 110/123 (89%)
```

```
Frame = +1
```

```
Q: 403 APGERRPGETERGSTQGDQAAHRGTEVLHVGAEQPRAPVLGAGRQHALAPRGGVQRPRIP 582  
      +PGERRPGETERGSTQGDQAAH GTEVLHVGAEQPRAPVLGAGRQHALAPRGGVQRPRIP
```

```
S: 33  SPGERRPGETERGSTQGDQAAHGGTEVLHVGAEQPRAPVLGAGRQHALAPRGGVQRPRIP 92
```

```
Q: 583 PTSCQLPALPALSFRCGESRASGGAHRLWQSCAHPAEAPVHLETRRQRPXXXXXXXXXXXXX 762  
      PTSCQLPALPALSFRCGESRASGGAHRLWQSCAHPAEAPVHLETRRQRP
```

```
S: 93  PTSCQLPALPALSFRCGESRASGGAHRLWQSCAHPAEAPVHLETRRQRPQGQGVNTGTVTT 152
```

```
Q: 763 XRA 771  
      RA
```

```
S: 153 GRA 155
```

BlastX, alignement avec la 2ème entrée trouvée

```
>gi|32171340|sp|Q16520|BATF_HUMAN  Gene info ATF-like basic leucine  
zipper transcriptional factor B-ATF (SF-HT-acivated gene-2) (SFA-2)  
Length = 125
```

```
Score = 185 bits (470), Expect = 1e-45
```

```
Identities = 92/92 (100%), Positives = 92/92 (100%)
```

```
Frame = +2
```

```
Q: 341 EKNRIAAQKSRQRQTQKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLC 520  
      EKNRIAAQKSRQRQTQKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLC  
S: 34  EKNRIAAQKSRQRQTQKADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLC 93  
  
Q: 521 SVLAASTPSPPEVVYSAHAHFHQPHVSSPRFQP 616  
      SVLAASTPSPPEVVYSAHAHFHQPHVSSPRFQP  
S: 94  SVLAASTPSPPEVVYSAHAHFHQPHVSSPRFQP 125
```

BlastX, interprétation (1/2)

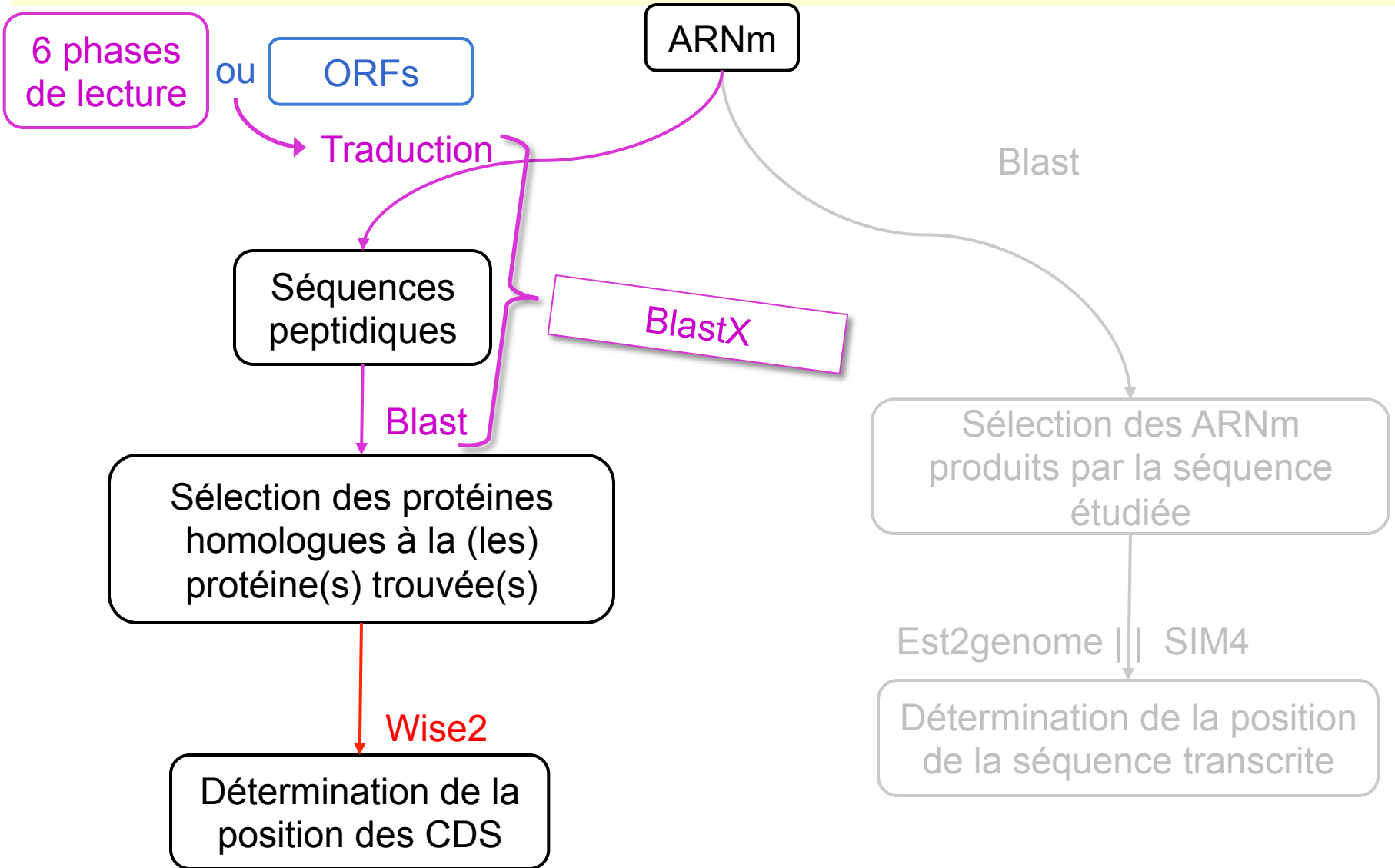
- Ne pas prendre la 1ère séquence comme référence
 - `hypothetical protein` : issue d'annotation automatique, non confirmée par d'autres sources
 - Elle est la seule à s'aligner avec la région 403..771 de l'ARNm
- La 2ème séquence correspond sûrement à la protéine codée par notre ARNm
 - `|sp|` : est issue de SwissProt donc annotation fiable
 - Très bon alignement (100% id)
 - Les autres protéines confirment les résultats

BlastX, interprétation (2/2)

Alignement avec la deuxième protéine :

- Frame = +2 :
 - La séquence codante est sur le brin +, c'est normal puisqu'il s'agit d'un ARNm (un seul brin).
- Query: 341..616 / Sbjct: 34..125
 - Il manque le début de la protéine de la banque
 - ⇒ Besoin d'un logiciel spécialisé pour aligner la protéine de la banque à l'ARNm
- ATF-like basic leucine zipper transcriptional factor
 - ⇒ C'est peut-être un facteur de transcription du type bZIP

1° Recherche de la CDS de l'ARNm (2/2)



Recherche du CDS avec Wise (2^{ème} entrée)

BATF_HUMAN	1	MPHSSDSSDSSFSRSPPPGKQDSSDDVRRVQRREKNRIAAQKSRQRQTQ MPHSSDSSDSSFSRSPPPGKQDSSDDVRRVQRREKNRIAAQKSRQRQTQ MPHSSDSSDSSFSRSPPPGKQDSSDDVRRVQRREKNRIAAQKSRQRQTQ
ARNm_hsp	242	accatgaagtataactcccgacgttggaagcaagaacaggcaaccacac tcagcaggacgtggccccgaaaccaatggtaggaaagtccaaggagaca gtccccctcccccttctcagcatttgaatggggatttccggcaggag
BATF_HUMAN	50	KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLCSVLAA... KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLCSVLAA... KADTLHLESEDLEKQNAALRKEIKQLTEELKYFTSVLNSHEPLCSVLAA...
ARNm_hsp	389	aggacccgaggcgacaggccagaaccaggcattatgcaacgccttgcg... acactatagaataaaacctgaataatcaataatccttagaactgcttcc... gccccgcggcacggagcggtacggcggcagaggccggggcccgcgcgggcc...

CDS 242..616

Bilan de la comparaison avec les protéines

- Recherche de la ou des protéines connues qui pourraient être codées par notre ARNm
 - Une protéine humaine a été trouvée
 - Mais : l'alignement donné par BlastX ne contient pas la protéine de la banque entière
 - Le début de la protéine contient une zone de faible complexité qui a été masquée
- Utilisation d'un logiciel spécialisé pour essayer de retrouver les positions du CDS
 - Wise donne un CDS en position 242..616 sur l'ARNm
 - La protéine de la banque s'aligne entièrement avec ce CDS

Un exemple : étude d'un ARNm

- Deux études à faire :

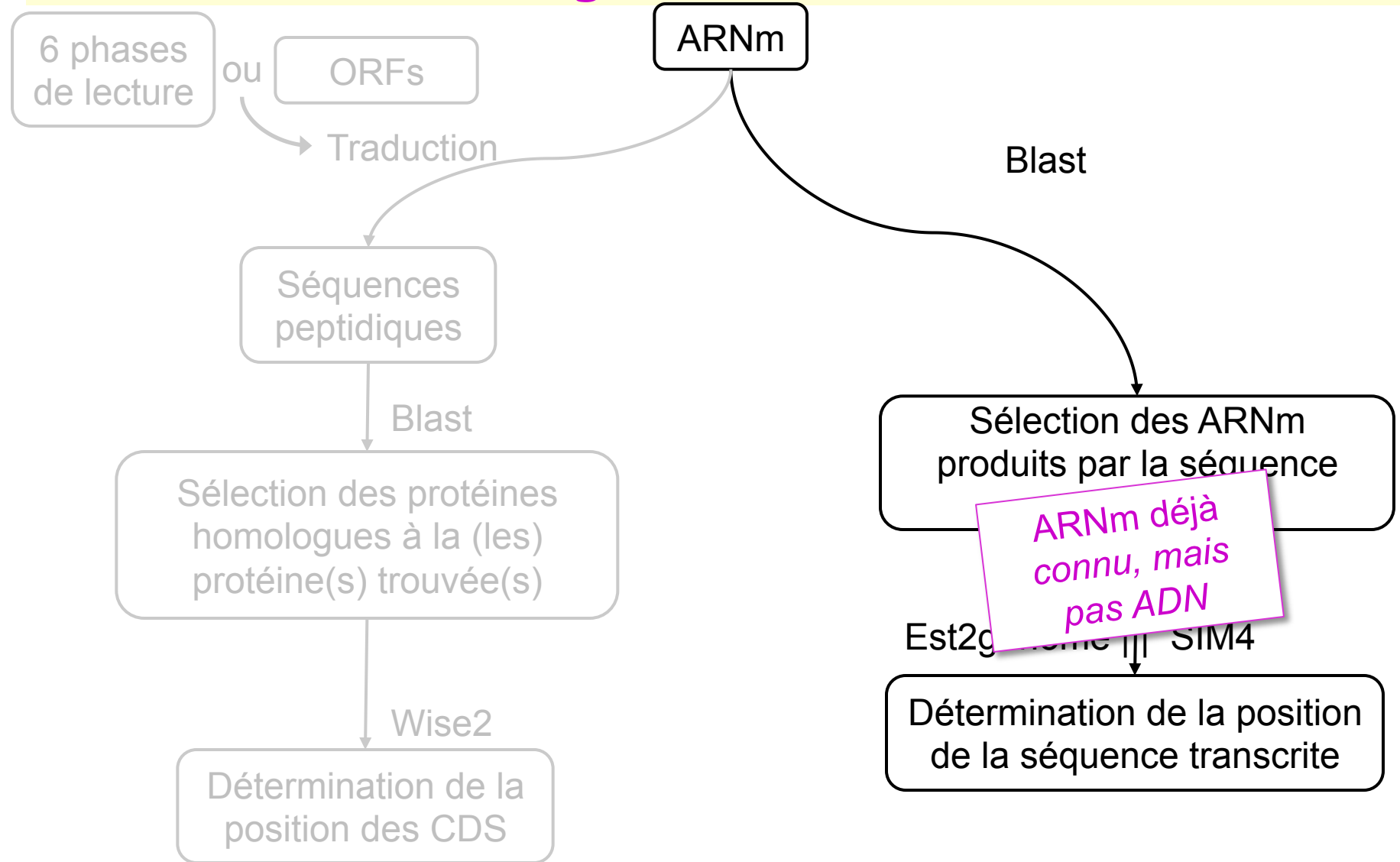
1. Recherche de la CDS de l'ARNm

- Méthodes classiques vues précédemment
- Puis, étude de la fonction de la protéine codée par l'ARNm
 - Voir *cours suivant* sur « annotation de protéines »

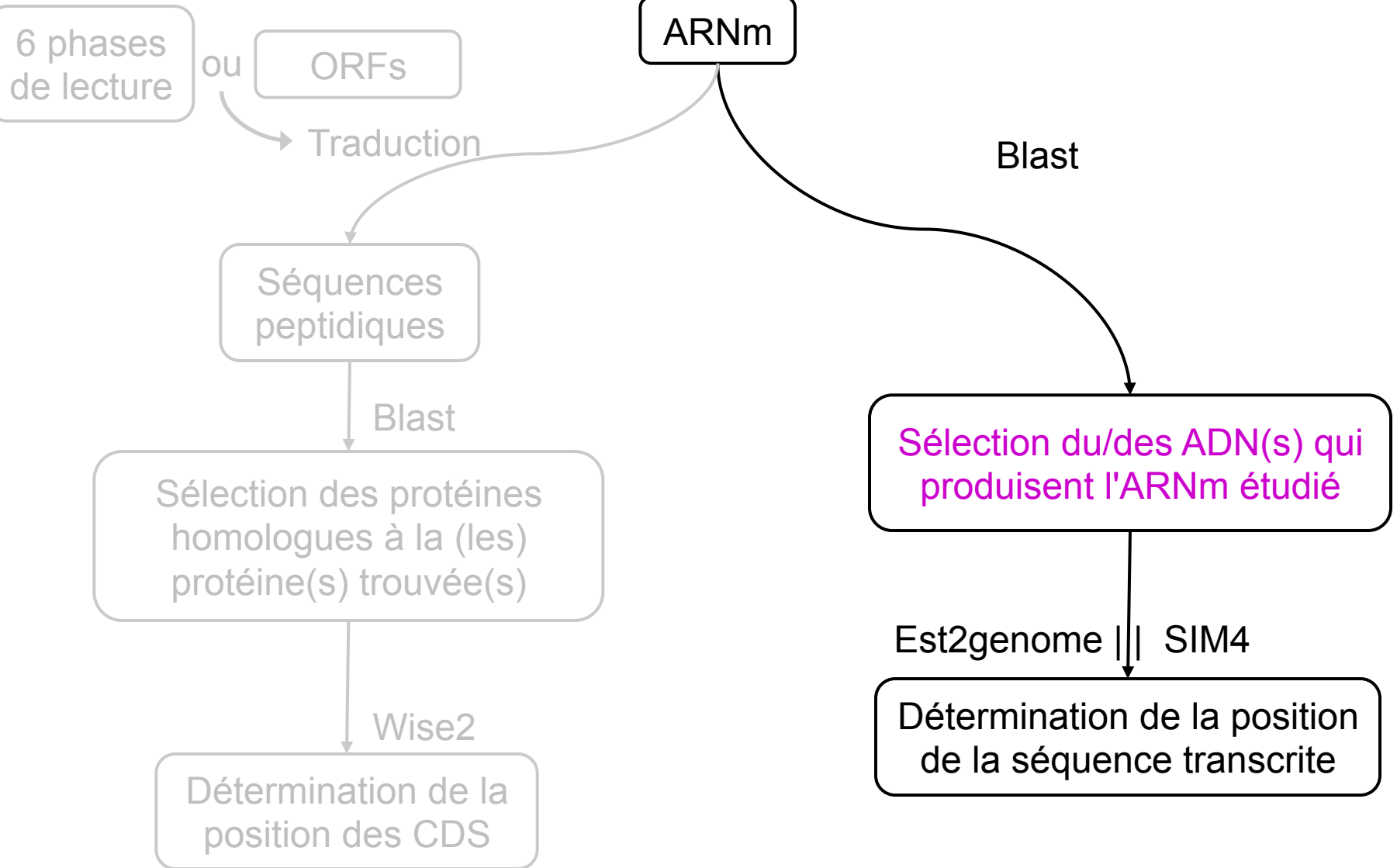
2. Localisation du gène codant l'ARNm

- Comparaison de séquence contre le génome complet
 - Blast, section « Genomes »

2° Localisation du gène codant l'ARNm

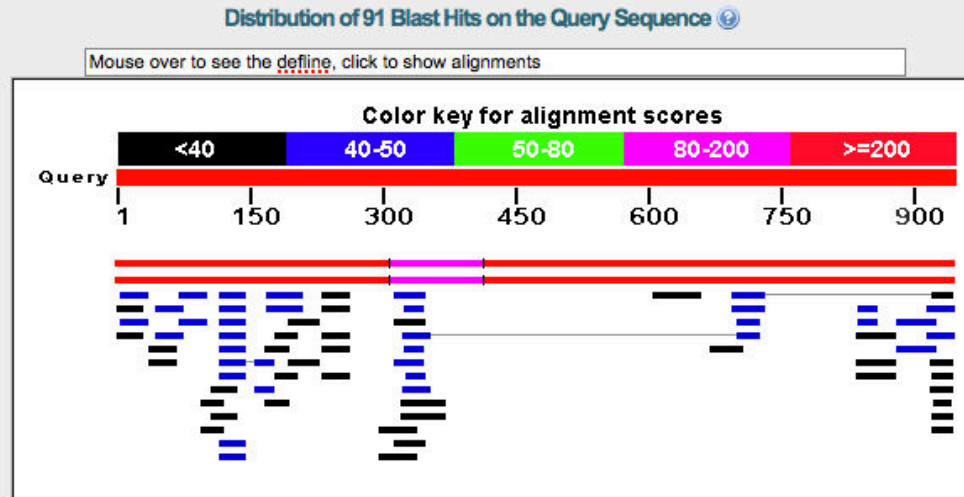


2° Localisation du gène codant l'ARNm



BlastN, localisation du gène

▼ Graphic Summary



▼ Descriptions

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer [P](#) PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Genomic sequences							
NT_026437.12	Homo sapiens chromosome 14 genomic contig, GRCh37.p5 Primary As	955	1784	100%	0.0	100%	
NW_001838113.2	Homo sapiens chromosome 14 genomic contig, alternate assembly Hu	955	1706	100%	0.0	100%	
NT_113891.2	Homo sapiens chromosome 6 genomic contig, GRCh37.p5 alternate lo	46.4	46.4	3%	0.060	93%	
NT_167248.1	Homo sapiens chromosome 6 genomic contig, GRCh37.p5 alternate lo	46.4	46.4	3%	0.060	93%	

>ref|NT_026437.12| Homo sapiens chromosome 14 genomic contig
Length=88289540

Score = 955 bits (1058), Expect = 0.0
Identities = 529/529 (100%), Gaps = 0/529 (0%)
Strand=Plus/Plus

```
Query 410      GGAGAGCGAAGACCTGGAGAAACAGAACGCGGCTCTACGCAAGGAGATCAAGCAGCTCAC 469
               |||
Sbjct 75082558 GGAGAGCGAAGACCTGGAGAAACAGAACGCGGCTCTACGCAAGGAGATCAAGCAGCTCAC 75082617

Query 470      AGAGGAACTGAAGTACTTCACGTCGGTGCTG.....AAATGCTTTAAAAG 938
               |||
Sbjct 75082618 AGAGGAACTGAAGTACTTCACGTCGGTGCTG.....AAATGCTTTAAAAG 75083086
```

~ 3^{ème} et dernier exon
(fin correcte)

Score = 553 bits (612), Expect = 2e-154
Identities = 306/306 (100%), Gaps = 0/306 (0%)
Strand=Plus/Plus

```
Query 1        CAagagagagagagagCGTGCAAGCCCCAAAGCGAGCGACATGTCCCTTTGGGGAGCAGT 60
               |||
Sbjct 75058537 CAAGAGAGAGAGAGAGCGTGCAAGCCCCAAAGCGAGCGACATGTCCCTTTGGGGAGCAGT 75058596

Query 61       CCCTCTGCACCCAGAGTGAGGAGGACGCAG.....AACAGG 306
               |||
Sbjct 75058597 CCCTCTGCACCCAGAGTGAGGAGGACGCAG.....AACAGG 75058842
```

~ 1^{er} exon
(début correct)

Score = 197 bits (218), Expect = 2e-47
Identities = 109/109 (100%), Gaps = 0/109 (0%)
Strand=Plus/Plus

```
Query 303      CAGGACTCATCTGATGATGTGAGAAGAGTTTCAGAGGAGGGAGAAAAATCGTATTGCCGCC 362
               |||
Sbjct 75061177 CAGGACTCATCTGATGATGTGAGAAGAGTTTCAGAGGAGGGAGAAAAATCGTATTGCCGCC 75061236

Query 363      CAGAAGAGCCGACAGAGGCAGACACAGAAGGCCGACACCCTGCACCTGG 411
               |||
Sbjct 75061237 CAGAAGAGCCGACAGAGGCAGACACAGAAGGCCGACACCCTGCACCTGG 75061285
```

~ 2^{ème} exon

BlastN, localisation du gène

```
>ref|NT_026437.12| Homo sapiens chromosome 14 genomic contig  
Length=88289540
```

```
Score = 955 bits (1058), Expect = 0.0  
Identities = 529/529 (100%), Gaps = 0/529 (0%)  
Strand=Plus/Plus
```

⇒ Le gène est sur le chr 14, sur le brin +

- 3 régions s'alignent entre le génome et l'ARNm

⇒ Sûrement 3 exons

- Positions de début et de fin du gène (de la transcription)

⇒ 75.058.537.. 75.083.086

Est2genome (ARNm / région 75058500..75083100 du chr14)

Exon	302	99.7	38	342	Hs14_26604	1	304	kesaco
+Intron	-20	0.0	343	2680	Hs14_26604			
Exon	105	100.0	2681	2785	Hs14_26604	305	409	kesaco
+Intron	-20	0.0	2786	24058	Hs14_26604			
Exon	529	100.0	24059	24587	Hs14_26604	410	938	kesaco
Span	896	99.9	38	24587	Hs14_26604	1	938	kesaco
Segment	16	100.0	38	53	Hs14_26604	1	16	kesaco
Segment	288	100.0	55	342	Hs14_26604	17	304	kesaco
Segment	105	100.0	2681	2785	Hs14_26604	305	409	kesaco
Segment	529	100.0	24059	24587	Hs14_26604	410	938	kesaco

Détermination de la position des exons

- Alignement de la région 75.058.500..75.083.100 du chr14 contre l'ARNm, avec Est2Genome
- 3 exons :
 - Calcul des positions sur le chr : $75058500 + 38 - 1 = 75058537$
 - ⇒ 75058537..75058841
 - ⇒ 75061180..75061284
 - ⇒ 75082558..75083086

Bilan

- CDS sur ARNm : 242..616
- Frontières des exons sur ARNm : 304.305 et 409.410
 - La partie codante du gène se trouve sur les 3 exons
- CDS sur chr 14
 - ⇒ 75058778..75058841
 - ⇒ 75061180..75061284
 - ⇒ 75082558..75082764

