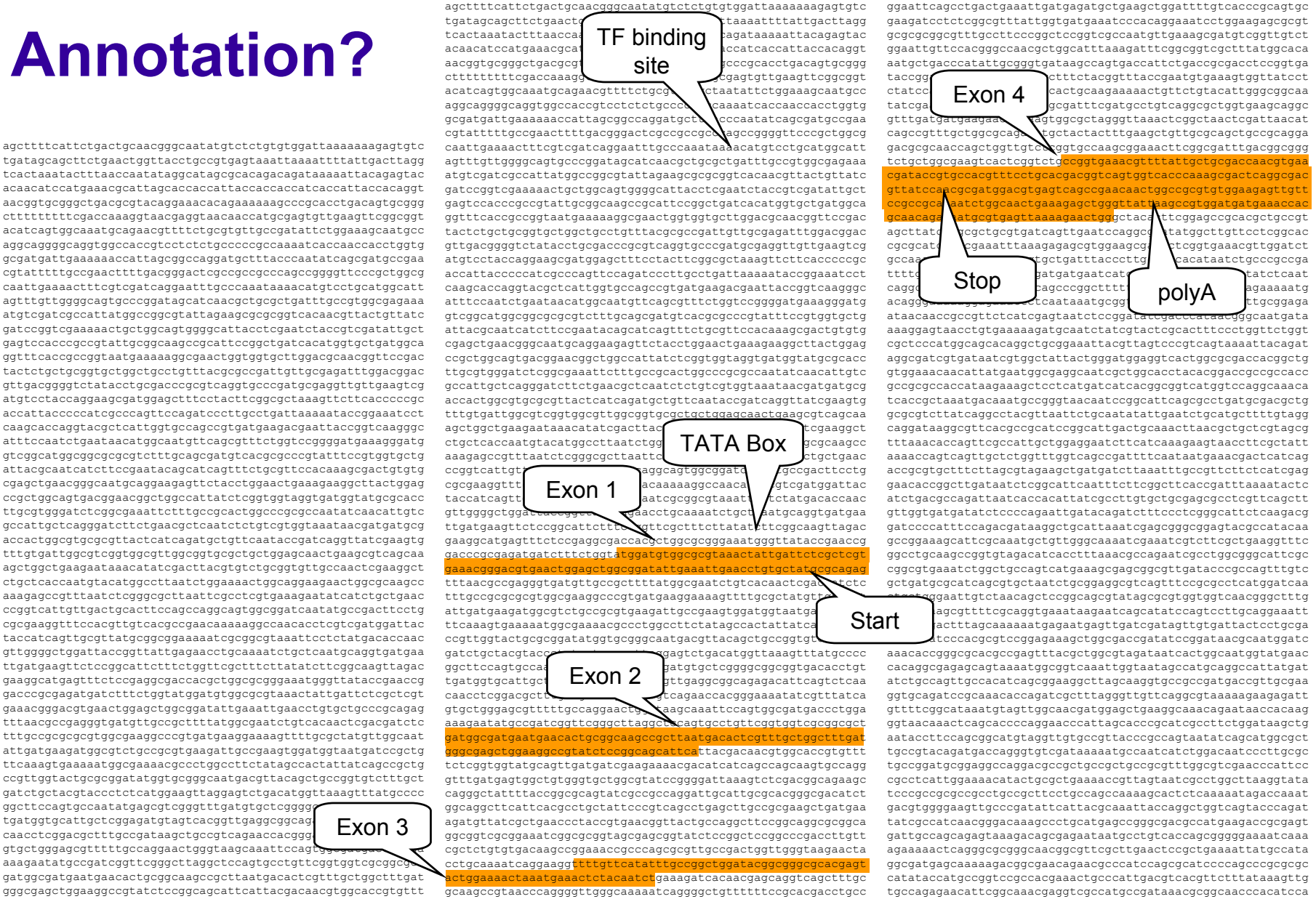
L’annotation des Génomes

Daniel Gautheret, 2004
ESIL, Université de la Méditerranée

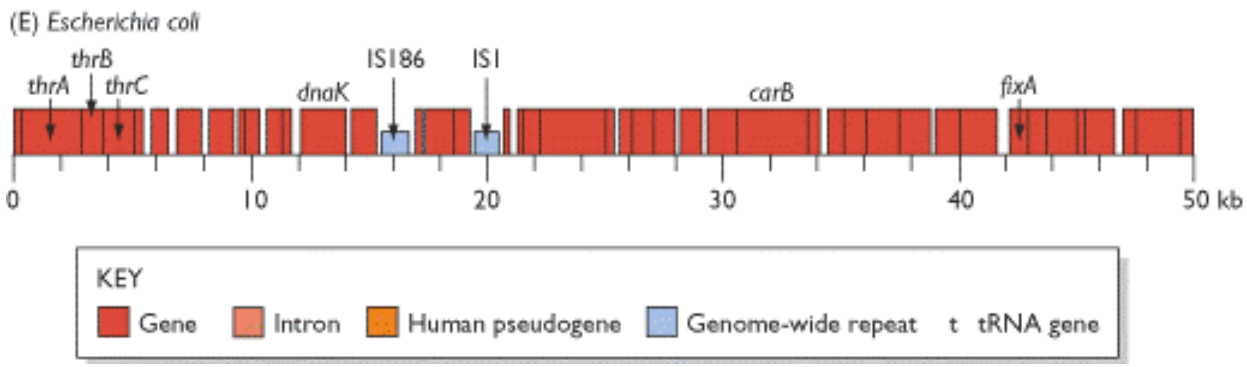
V.2004.2 http://www.esil.univ-mrs.fr/~dgaut/Cours

Annotation?


TF binding site
Exon 4
Stop
polyA
TATA Box
Exon 1
Start
Exon 2
Exon 3

Gènes procaryotes

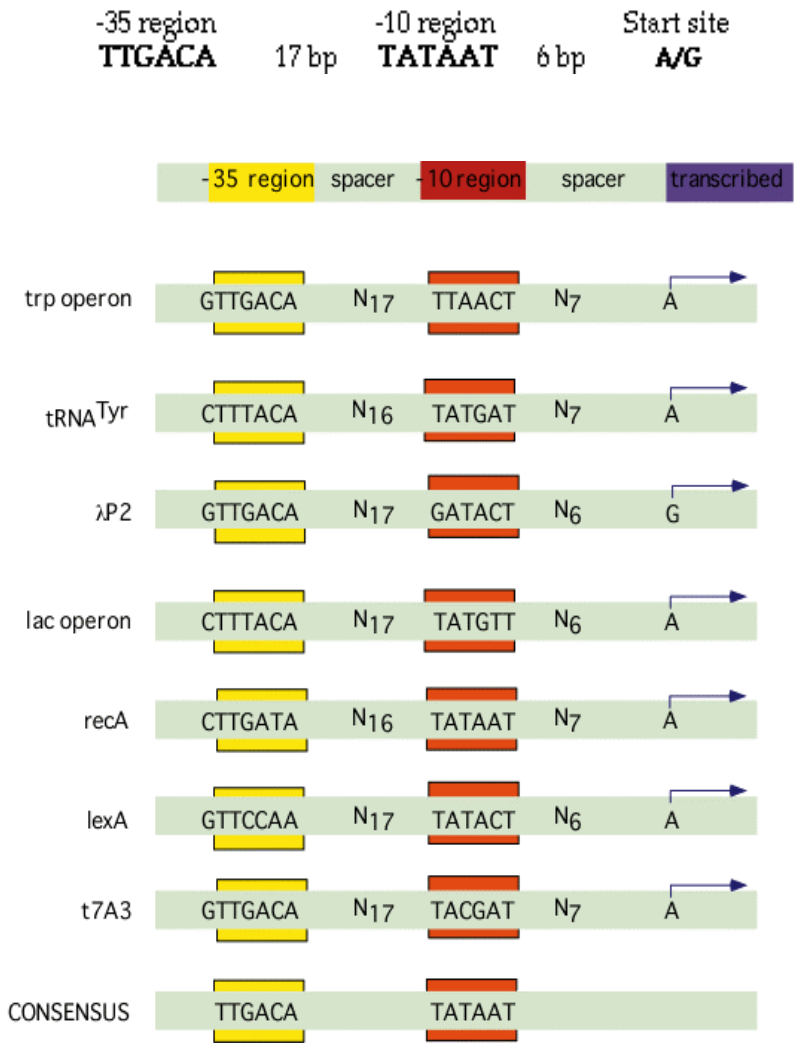
- ★ longueur gène 950 nt. en moyenne (coli)
- ★ Densité en gènes. 95% du génome est transcrit chez E. coli.
- ★ Gènes organisés en opérons. 600 opérons dans le génome de Coli.



Signaux des gènes microbiens (1)

★ **Le promoteur**

- la région reconnue par la polymérase à ARN. Juste en amont du site d'inititiation de la transcription.
- Comprend trois éléments: la boite de Pribnow (TTGACa) vers -35, la boite TATA (TAtAAT) vers -10 et le site d'initiation de la transcription
- Pribnow E.coli (%)= T82 T84 G78 A65 C54 a45
- TATA E coli (%) = T80 A95 T45 A60 a50 T96
- Ce sont ces 3 séquences que les facteurs sigma reconnaissent (lient à la fois le promoteur et l'ARN polymérase).



Crédit:
http://www.blc.arizona.edu/marty/411/Modules/prokrom.html

Signaux des gènes microbiens (2)

★ **L'opérateur**
- séquence reconnue par d'éventuelles protéines régulatrices (p. ex. represseur). Peut se glisser entre la séquence TATA et le site d'initiation de la transcription

★ **La séquence de Shine-Dalgarno, ou Ribosome Binding Site (RBS)**
- (aGGAGGu) environ 10 nt avant codon ATG. La séquence SD est une région riche en purine de 3 à 10 nt qui permet au ribosome de distinguer le véritable codon initiation d'autres AUG fortuits. Cette région s'apparie avec une région très conservée et riche en pyrimidine de l' ARNr 16S. Cet appariement aligne le codon AUG au site P de l'ARNr.
- La séquence active se trouve en fait sur l'ARNm

★ **Le codon initiation**
- ATG (rarement: GUG)
- Le ribosome débute ici la traduction, avec un aminoacide methionine, placé par un tRNAfmet particulier, et fréquemment enlevé par la suite.

★ **Le cadre de lecture ouvert (ORF)**
- Chaque ORF d'un ARNm procaryote s'appelle un **cistron**. La plupart des ARNm procaryotes sont polycistroniques: ils contiennent plusieurs cistrons et codent donc pour plusieurs protéines.

Signaux des gènes microbiens (3)

★Le codon Stop

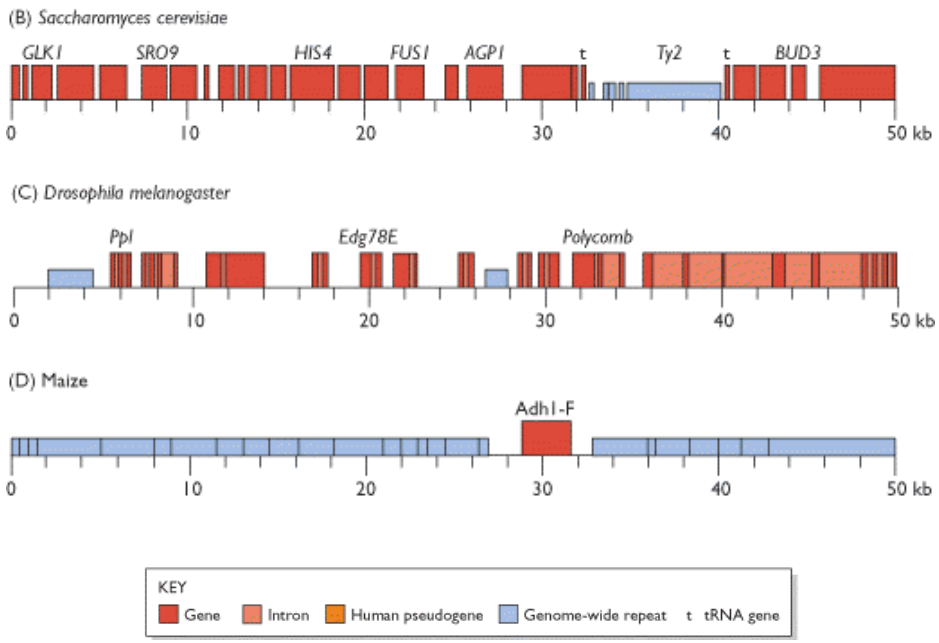
- TAA, TAG ou TGA.

★ Le terminateur

- La terminaison de la transcription chez les procaryotes nécessite une séquence de terminaison. Selon les gènes, cette séquence fait intervenir ou non le facteur protéique Rho.
- **Rho-indépendant**: Tiges riches en GC. centrée vers 20 - 30 bases avant la fin du transcrit, suivi d'environ 6 U. Pas de séquence rigoureusement conservée au niveau de l'hélice, mais structure tige-boucle conservée.
  Lorsque la RNA polymérase atteint la tige, elle marque une pause pendant une durée suffisante pour qu'advienne un décrochage de l'enzyme (favorisée par paires A:U) et donc un arrêt de transcription.
- **Terminateur Rho-dépendant**: Tiges-boucles riches en GC. (même structure tige-boucle que Rho-indépendant). Pas de région riche en A/U. Mais nécessité d'une séquence 100 nt avant la fin du transcrit. Le consensus est faible: riche en C, pauvre en G.

Gènes eucaryotes

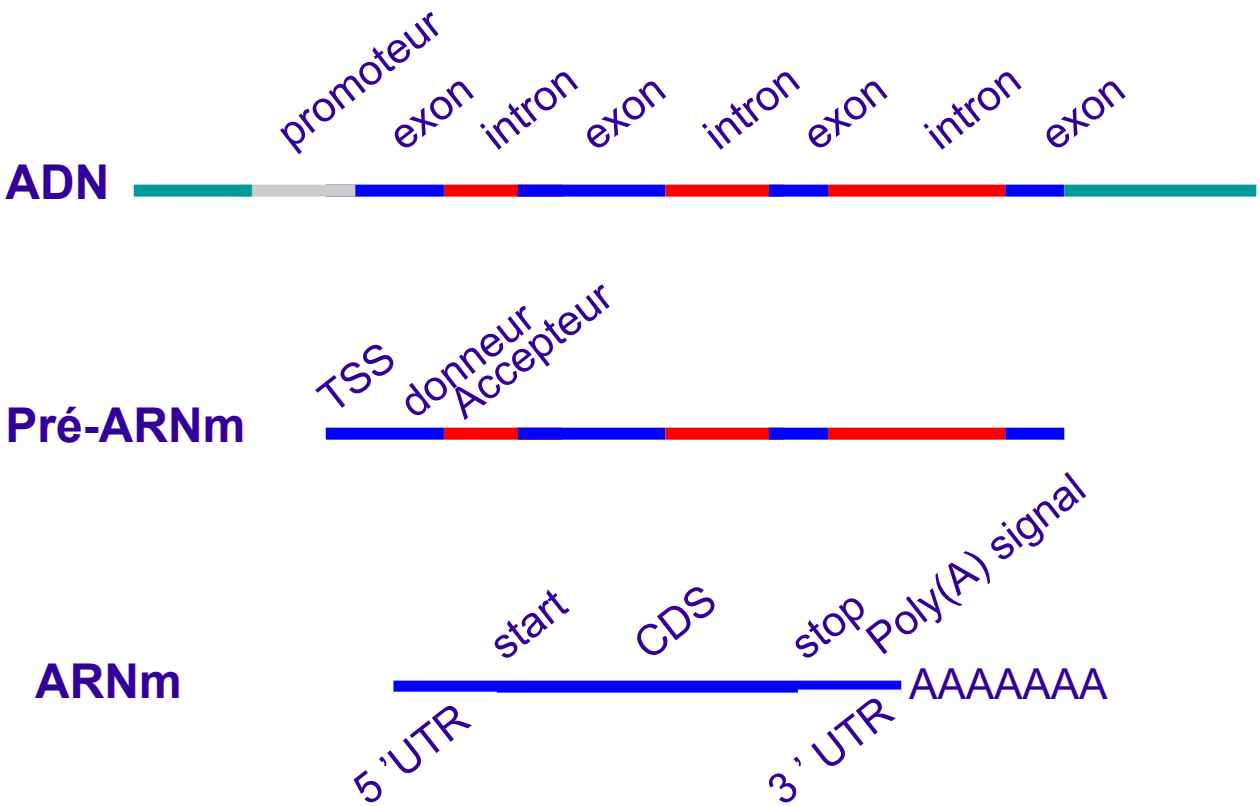
- ★ Gène humain moyen: 8,8 introns, 27 kb, 3'UTR:770bp, 5'UTR:300bp, CDS:1340bp, exon moyen: 145 bp (218 bp pour C. elegans), intron moyen:3365 bp (mais pic à 87 bp).
- ★ Gènes "monstres": dystrophine: 2,4 Mb; Facteur de coagulation VIII: 186 kb, 26 exons; Tinine: CDS de 80780 bp, 178 exons
- ★ Densité: humain: 1 gène tous les 100kb en moyenne; C.elegans: 1gène/5-6kb (25%); S. Cerevisiae: 1 gène/2kb.



From « Genomes 2 », T.A. Brown

Remarque: 35000 gènes de 30kb en moyenne font 1Gb: donc au moins 1/3 du génome humain est transcrit. Par contre, seul 1,5% est codant!

Signaux des gènes eucaryotes



Promoteur et région 5’

Les éléments amont du promoteur

=sites de liaison aux facteurs de transcription (cf Transfac)

- ★ Séquences courtes de 6-20 nt affectant généralement l' efficacité de l'initiation de la transcription.
  - Boîte CCAAT
  - Sp1 box
  - CRE
  - AP2 box
    etc..

Pol I, Pol II, Pol III

- ★ Il existe trois types de promoteurs eucaryotes, pour les trois types d'ARN polymérase les reconnaissant: Pol-I, Pol-II et Pol-III. Les promoteurs **Pol-I** se trouvent face aux ARNr 18S et 28S. Les promoteurs **Pol-II** se trouvent face aux ARNm. Les promoteurs **Pol-III** se trouvent face aux ARNt et ARNr 5S.

Promoteurs pol-II

Promoteurs et région 5'

- ★ Les promoteurs de type Pol-II intéressent car ils signalent les gènes protéiques. Ces promoteurs contiennent une boîte TATA et au moins une autre séquence importante en amont.
- ★ On peut la schématiser comme sur la fig du haut (TATA=boîte TATA, INR= Initiateur):

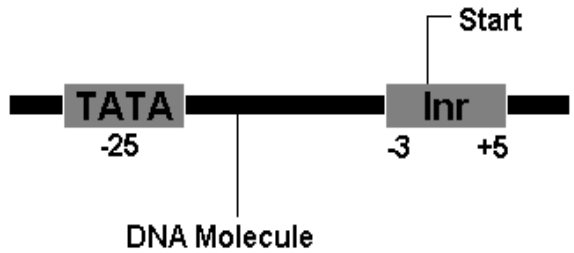
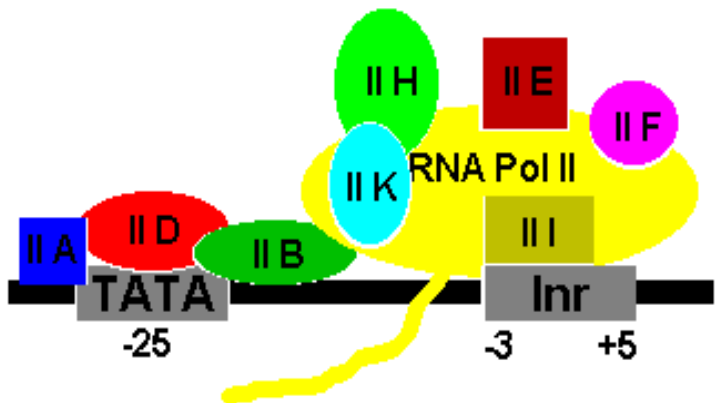


Fig.1 The Generic Promoter

La boîte TATA

- ★ La majorité des promoteur de gènes protéiques eucaryotes contiennent une boîte TATA.
- ★ Semblable à la séquence TATA des procaryotes, la position mise à part (-10 chez les procaryotes).
- ★ Boîte TATA est trouvée à 25-30 paires de bases en amont du site de départ de transcription (TSS). Une position relativement constante dans les promoteurs eucaryotes. C'est le seul signal dans le promoteur se trouvant à une distance définie du TSS.



Facteurs protéiques liant le promoteur, juste avant l'initiation de la transcription. Cas typique, mais nullement général, les facteurs de variant considérablement suivant les gènes et l'environnement. On connait plus de 2000 facteurs de transcription.

Promoteurs (suite)

La signature de la boîte TATA

- La sequence d'environ 8 paires de bases contient pratiquement que des adenines et thymines, et tend a être encadrée par des séquences riches en guanine et cytosine, ces dernières pouvant participer à la fonction du promoteur.
- TATA consensus: *GTATAAAAGGCGGGG* (mais beaucoup de variation)
  Le consensus de la TATA box est faible et cet élément est même absent dans de nombreux promoteurs.

General eukaryotic TATA-box model derived from 860 unrelated promoter sequences:

Position	1	2	3	4	5	6	7	8	9	10	11	12
% A	19.4	23.4	5.0	83.5	4.4	89.2	71.0	84.8	45.0	35.7	15.5	18.5
% C	22.7	34.0	11.0	1.3	3.3	0.8	0.8	2.9	3.4	14.0	36.5	37.0
% G	26.5	30.8	4.5	1.4	0.9	1.7	0.5	9.5	16.4	38.4	36.3	30.4
% T	31.4	11.7	79.5	13.9	91.4	8.4	27.7	2.8	35.2	11.8	11.7	14.1
Consensus			T	A	T	A	W	A	D	R		

Autres signaux 5’

L'Initiateur

- **L'initiateur (INR), se trouve près du site de début de transcription, entre les positions -3 et +5. Il y a peu ou pas de similarité entre les initiateurs de différents promoteurs, toutefois, la première base du mRNA transcrit tend à être un A, souvent flanqué de pyrimidines.**
  La RNA Pol-II peut parfois initier la transcription avec l'INR seul, dans des promoteurs simples sans boîte TATA.

Autres signaux 5’

Les ilots CpG

- ★ Les ilôts CpG sont des zones riches en dinucléotide CG, fréquemment associées aux régions 5' des gènes de vertébrés
- ★ L'ilôt s'étend sur le promoteur et l'exon 1 (ou 1 et 2)
- ★ Fréquence attendue du dinucléotide CpG = 4% (0.21x0.21), mais fréquence observée: un cinquième de cette valeur. Pourquoi?
- ★ Méthylation naturelle des CpG et réparation en TpG par déamination
- ★ Au niveau du promoteur: protection des CpG. Donc Fréquence normale.
- ★ Typiquement 1-2kb de longueur. Environ 70% G+C (contre 40% dans le reste du génome humain)
- ★ Les îlots CpG sont associés à tous les gènes housekeeping (constitutifs) et à 40% des gènes tissu-spécifiques

The dinucleotide CpG is notable because it is greatly under-represented in human DNA, occurring at only about one- fth of the roughly 4% frequency that would be expected by simply multiplying the typical fraction of Cs and Gs (0.21 ´ 0.21). The deficit occurs because most CpG dinucleotides are methylated on the cytosine base, and spontaneous deamination of methyl-C residues gives rise to T residues. (Spontaneous deamination of ordinary cytosine residues gives rise to uracil residues that are readily recognized and repaired by the cell.) As a result, methyl- CpG dinucleotides steadily mutate to TpG dinucleotides. However, the genome contains many `CpG islands' in which CpG dinucleo- tides are not methylated and occur at a frequency closer to that predicted by the local GC content. CpG islands are of particular interest because many are associated with the 5' ends of genes

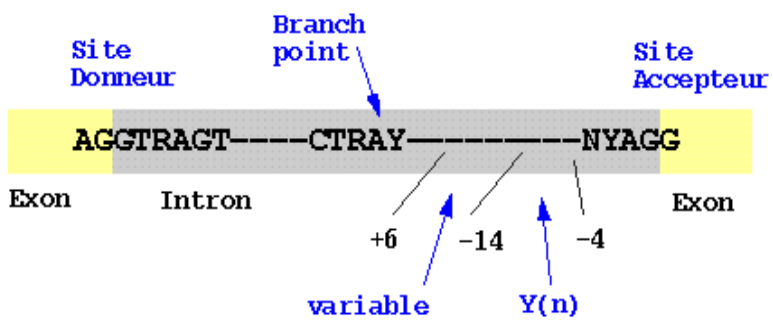
Extrait de l'article de Nature sur le génome humain

Signaux dans le gène

Jonctions Introns/Exons

- ★ 4 éléments importants chez les eucaryotes: donneur, accepteur, point de branchement et région riche en pyrimidines
- ★ Le GT et le AG sont systématiquement exclus de l'ARNm mature.
- ★ 98.1% des introns humains possèdent les GT et AG. 0.76% ont GC-AG. 0.1% ont AT-AC.
- ★ La coupure peut s'effectuer à l'intérieur d'un codon.
- ★ Levure: donneur= AG/GTATGT, branch: CTAAC, accepteur =CAG/G.

Signaux jonction intron-exon vertébré



Signaux dans la partie codante

Le biais de codon

- ★ **l'abondance et l'utilisation différente des acides aminés entraine naturellement des fréquences différentes pour chaque codon**
- ★ **Mais lorsque les codons synonymes ne sont pas employés avec la même fréquence, on parle de biais d'usage des codon (codon usage bias). Découvert initialement chez la levure et coli.**
- ★ **Les biais de codon diffèrent d'une espère à l'autre, selon les contraintes propres à chaque espèce:**
  - la nature du code génétique
  - les ARNt disponibles
  - D'éventuelles contraintes évolutives (GC content, taux de mutation..)
  - Une préférence pour les séquences purine-N-pyrimidines
- ★ **A coté du biais propre à l'espèce, il existe des biais propres à certains gènes. Généralement les gènes les plus exprimés sont les plus biaisés, les codons les plus utilisés étant ceux pour lesquels les ARNt sont les plus nombreux (codons sélectionnés pour une plus grande vitesse de traduction)**
  - Biais % expression chez Arabidopsis: deux classes de gènes différenciées par le choix de pyrimidine à la position silencieuse du codon. Les gènes très exprimés préfèrent C, les autres préfèrent T.
- ★ **Il est également proposé que les appariement codon/anticodon riches en paires CG soient sélectionnés car ils minimisent les erreurs de traduction.**

Signaux dans la partie codante

Le biais de codon

★ **Le biais de codon est employé de 3 façons:**

- Comme signature des exons codants
- Comme signature des gènes fortement exprimés.
- Mais aussi comme signature de transferts horizontaux récents (biais différent du biais normal de l'espèce).

★ **Paires de codons**

- Les combinaisons de di-codons sont encore plus biaisées que les codons seuls.
- Certaines combinaisons de di-codons causeraient des calage dans la transcription en raison d'interactions particulières entre tRNA, ribosome et mRNA.

Codon	E.coli	D.melanogaster	H.sapiens	S.cerevisiae
AGT	3	1	10	5
AGC	20	23	34	4
TCG	4	17	9	1
TCA	2	2	5	6
TCT	34	9	13	52
TCC	37	48	28	33

Usage des codons Sérine dans différents organismes. La levure préfère TCT, l'homme AGC.

Signaux 3’

Terminaison des gènes de vertébrés

- L'exon terminal contient les signaux nécessaires à la maturation de pré- mRNA en mRNA: clivage et polyadénylation.
  1. Un héxamère AAUAAA ou AUUAAA (et parfois des variants présentant une mutation sur une base: AGUAAA, UAUAAA, CAUAAA, etc.), 10 à 30 bases en amont du site de clivage (en moyenne 17 bases). L'un ou l'autre des variants est observé dans plus de 90% des gènes.
  2. Au site de clivage: un dinucléotide CA, assez mal conservé.
  3. 20 à 40 bases après le site de clivage (donc toujours sur le pré-mRNA): une région riche en GU, de séquence variable.
- Il peut y avoir plusieurs site de polyadenylation dans un pré-mRNA (jusqu'à une dizaine). Dans ce cas, on observe plusieurs fois l'ensemble de trois éléments.

Programmes de recherche de gènes

★ **2 Mots importants**

- ***Sensibilité*: La capacité à détecter les vraies instances de l'objet recherché (« vrais positifs »).**
- ***Spécificité*: La capacité à rejeter les fausses instances (« faux positifs »).**

$$\text{Sensibilité: SN} = \frac{TP}{TP+FN}$$

Total « vrais » objets

$$\text{Specificité : SP} = \frac{TP}{TP+FP}$$

Total prédictions

TP: “Vrai positif”
FP: “Faux positif”
FN: “Faux négatif”

Programmes de recherche de gènes

Analyse des signaux

- La plupart des auteurs ont tenté d'exploiter à la fois la présence d'un cadre de lecture et des autres signaux de séquence: promoteur (TATA, ilots CpG), jonction intron-exon (donneur, accepteur), signal de polyadenylation, etc.
- Même si les signaux étaient parfaitement conservés, ils sont peu spécifiques. Par exemple, on trouve en moyenne un ORF de 150 nt (la taille typique d'un ORF) tous les kilobases, alors qu'il n'en existe en fait qu'un tous les 10kb dans les génomes de vertébrés!
- Les boîtes TATA et autres éléments des promoteurs, ainsi que les signaux d'épissage sont également peu spécifiques: le motif donneur AxGT(A/G)xG est observé 559 fois dans un contig humain de 67kb ne contenant que 7 exons.
- Pire encore, tous ces signaux sont dégénérés (mal conservés) et on doit donc, pour éviter de manquer de nombreuses occurrences autoriser de fortes déviations par rapport aux motifs idéaux: on manque alors de spécificité.
- Il existe dans les gènes de vertébrés de grandes régions 5' et 3' non traduites (plusieurs centaines de bases en 3'), privées de signaux connus.
- La reconstruction du gène complet ajoute encore une source d'erreur: risque d'oublier des exons ou de mélanger ceux provenant de deux gènes.

Programmes de recherche de gènes

Analyse du contenu

- ★ Fait appel aux biais dans les séquences des gènes
- ★ C'est pour l'instant la statistique sur 6 lettres (2 codons) qui s'avère le mieux discriminer séquences codantes et non codantes. Les raisons en sont encore mystérieuses.

Programmes de recherche de gènes

Les algorithmes

- ★ **Les meilleures méthodes sont celles qui combinent détection de signaux et analyse du contenu.**
  - **Linear Discriminant Analysis (LDA)**
  - **Réseaux neuronaux**
  - **HMM (modèles de Markov cachés)**
- ★ **Ces méthodes reposent toutes sur des ensembles d'entrainement. Cela pose un problème pour la détection d'exons "non canoniques". La composition des ORF peut varier considérablement dans certaines familles de gènes. Les statistiques apprises sur l'ensemble d'entrainement ne s'appliqueront plus. Les signaux des promoteurs et des sites d'épissages sont par contre assez généraux.**

Programmes: gènes microbiens

- Les ORF (Open Reading Frames, cadres de lecture ouverts) supérieurs à 150 nt. sont pratiquement tous vrais
- Problème pour les ORF courts, inf. à 150.

- GENEMARK. Utilise les modeles markoviens dans sa version de base (recherche des regions codantes seulement). Puis l'idee a ete integree dans un modèle markovien cache (HMM) plus complexe tenant compte des informations du promoteur et du RBS.

Programmes: gènes microbiens

Results of GeneMark.hmm predictions for 10 complete bacterial genomes*

Genome	Genes annotated	Genes predicted	Annotated genes predicted by GeneMark.hmm & GeneMark (%)	Correct 5' end prediction of annotated genes (%)	Potential new genes (%)
A.fulgidus	2407	2530	98.0	73.1	15.1
B.subtilis	4101	4384	97.2	77.5	9.8
E.coli	4288	4440	97.3	75.4	8.2
H.influenzae	1718	1840	96.2	86.7	10.2
H.pylori	1566	1612	95.6	79.7	8.7
M.genitalium	467	509	98.3	78.4	17.3
M.jannaschii	1680	1841	99.2	72.7	12.9
M.pneumoniae	678	734	95.9	70.1	13.6
M.thermoauto	1869	1944	97.5	70.9	8.6
Synechocystis	3169	3360	98.5	89.6	9.4
Average	**21943**	**23194**	**97.3**	**78.1**	**10.4**

TP

The second and third columns show the number of genes annotated in GenBank and the number of genes predicted, respectively.
The "Annotated genes predicted" column presents the percentage of annotated genes which were predicted by GeneMark and GeneMark.hmm
The "Correct 5' end prediction of annotated genes" column shows the percentage of genes whose starts were predicted exactly.
"Potential new genes" is the fraction of predicted genes for which no annotated analog was found. All measures are expressed in percent.
* **Reference:** A. Lukashin and M. Borodovsky, GeneMark.hmm: new solutions for gene finding, **NAR**, 1998, Vol. 26, No.4, pp 1107-1115.

Programmes: gènes eucaryotes

- ★ GENSCAN (Burge & Karlin, J. Mol. Biol. 268, 1-17, 1998.)
  Utilise les HMM. Plusieurs modèles sont employés pour exons, introns, promoteurs, etc. Sensibilité et Spécificité autour de 80% pour les exons correctement prédits. Beaucoup moins bon pour la prédiction de gènes complets.

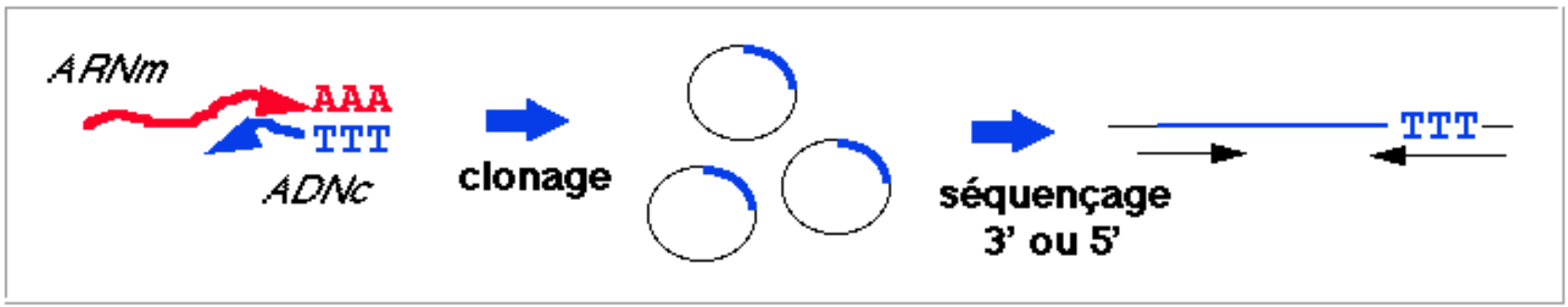
- ★ GRAIL (Uberbacher & Mural, PNAS, 88, 261-11, 1991).
  Utilise les Réseaux neuronaux. Moins performant que GENSCAN, mais reste connu comme l'un des pionniers.

Identification de gènes par homologie

- **Une des méthodes les plus sûres pour la modélisation des gènes est la comparaison de la séquence à annoter avec une banque de séquences.**
- **Blast de la séquence à annoter *contre Swissprot ou nr:* Si le gène recherché (ou un paralogue) se trouve dans Swissprot ou nrdb, cette méthode detecte des similarités significatives au niveau de chaque exon. C'est la facon la plus simple et la plus directe d'identifier les exons. *Limité à la région traduite*.**
- **Blast de la séquence à annoter *contre dbEST*: Les EST permettent souvent une bonne couverture de la région *exprimée* du gène. Pour les gènes n'ayant pas d'homologue connu, ils sont une aide précieuse à l'identification des exons. Les EST peuvent aussi recouvrir les regions non traduites, ils sont donc utiles a l'identification des UTR 3' et 5'. Inconvénient: les EST sont plus souvent dans le 3' non codant. Il y a des trous. Avantage: On observe souvent des formes alternatives du mRNA dans les EST.**

Expressed Sequence Tags (ESTs)

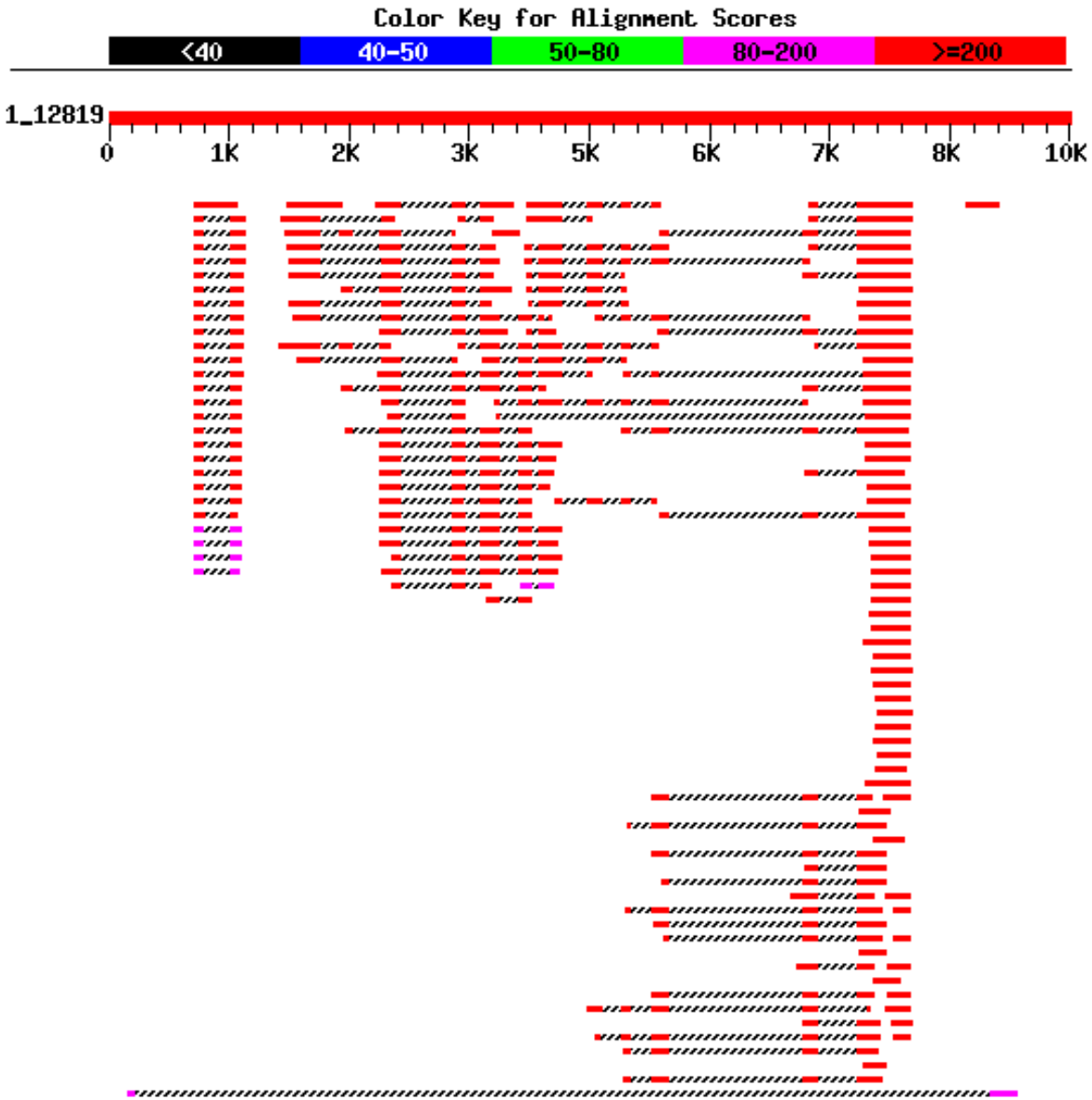
- ★ Idée originale: pourquoi vouloir tout séquencer (95% de junk DNA) si ce sont les gènes qui nous intéressent?
- ★ EST = Séquences partielles d'ADNc clonés et prélevés aléatoirement.



- ★ Grandes applications:
  - Cataloguage des gènes
  - Profils d'expression/Northerns virtuels

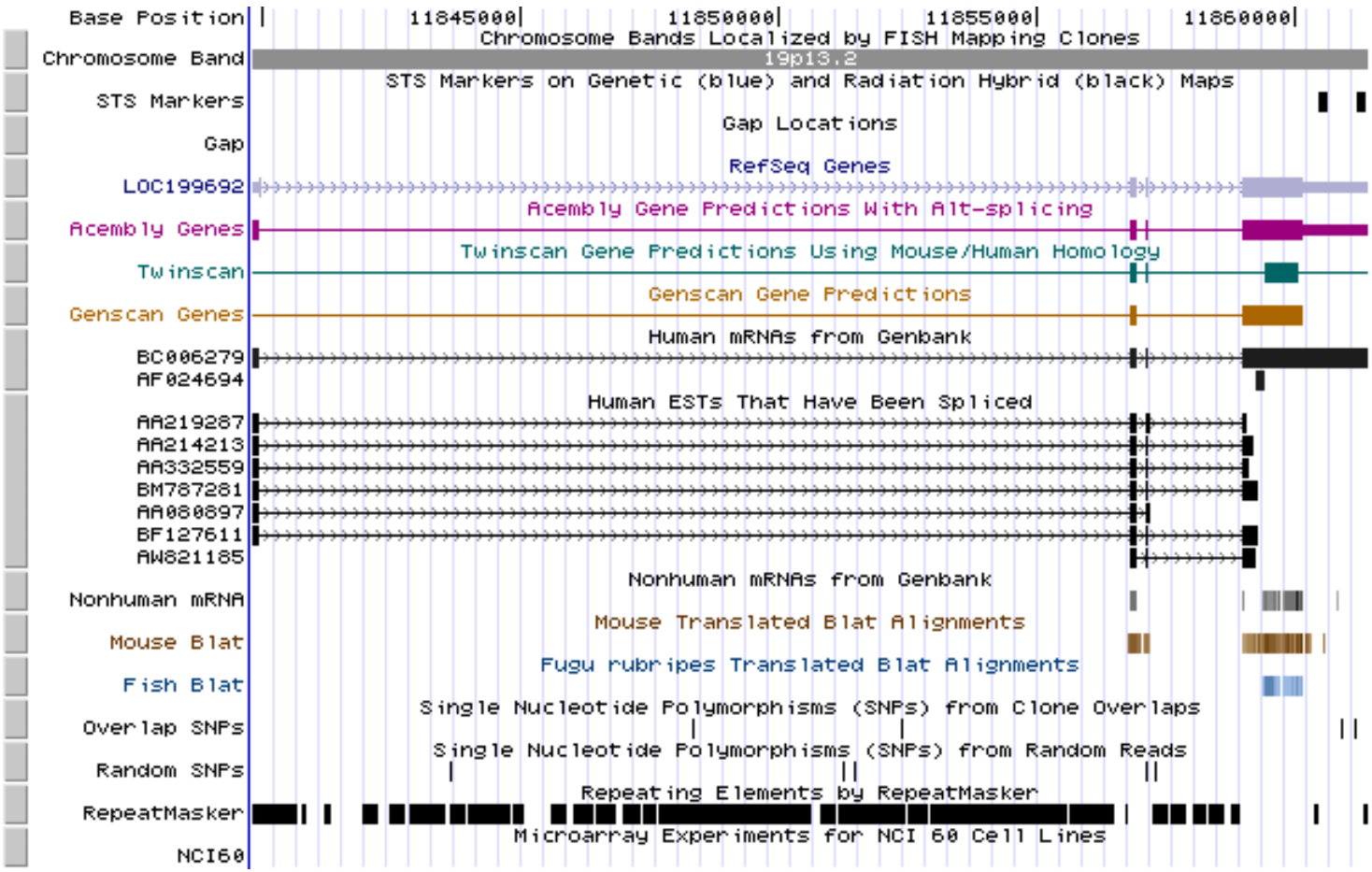
Identification de gènes par homologie

- **Dans l'exemple ci-contre, on a réalisé un Blast d'un contig de 10kb contenant un gène unique contre la banque dbEST. La dernière ligne en bas est clairement un artefact (séquences répétées).**



Banques génomiques intégrées: UCSC

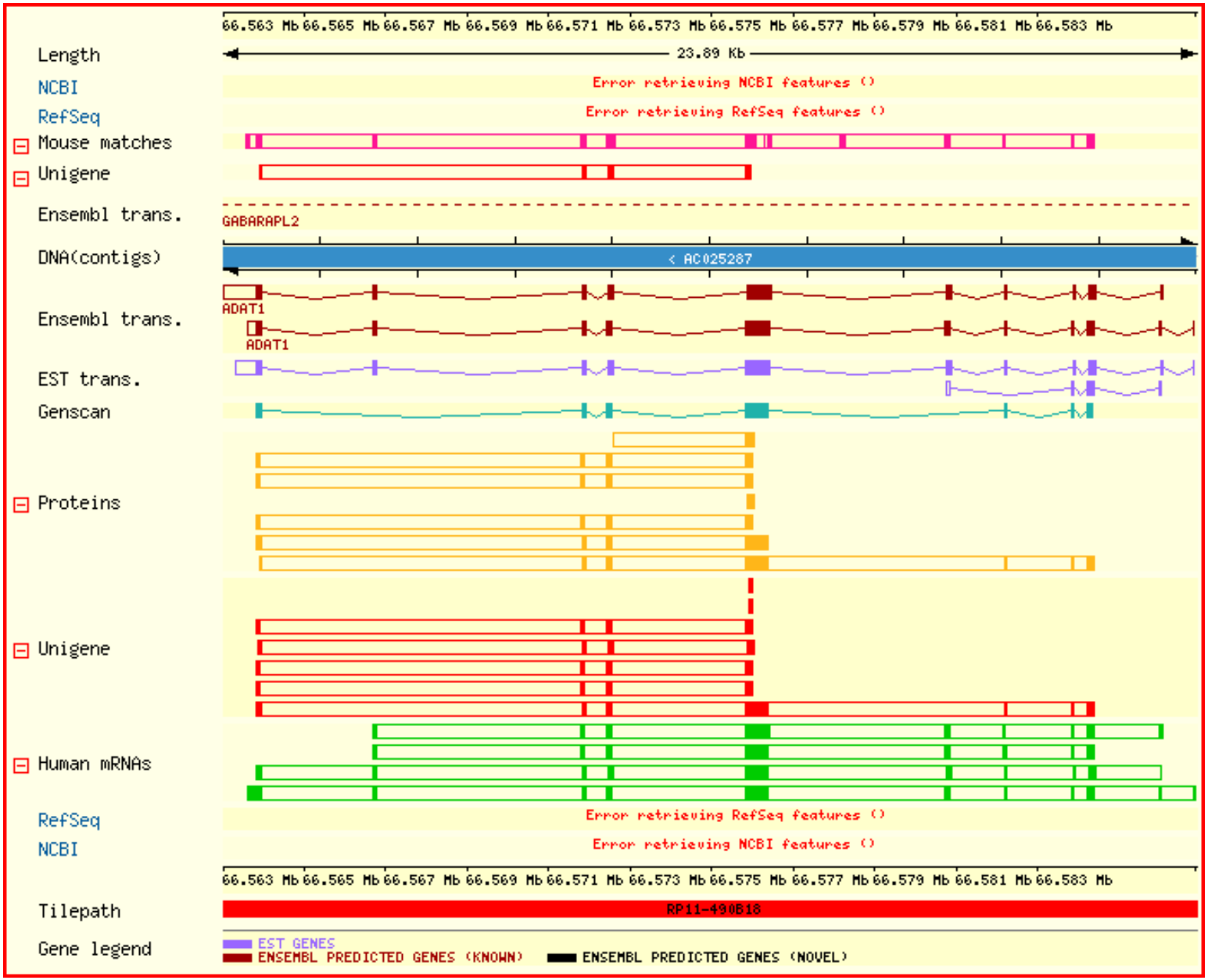
- ★ http://www.genome.ucsc.edu/



Banques génomiques intégrées: Ensembl

- ★ Ressource intégrée d'annotation des génomes eucaryotes
- ★ Méthode d'obtention: Genscan sur contig, puis Blast vs prot, mRNA, EST, PFAM
- ★ Version Juillet 2001: Confirmed genes: 21921 Predicted genes: 246366 Confirmed exons: 143479 Predicted exons: 770562 Transcripts: 23931 Contigs: 329154 Sequences: 29080 Base Pairs: 4318661441
- ★ Plusieurs banques en une:
  - Peptides confirmés
  - "cDNA" confirmés
  - peptides prédits
  - Banque ADN (golden path)

Ensembl: « contig view »



Les gènes non-codants

Que sont-ils?

- Il existe un grand nombre de gènes produisant des ARN dont la fonction n'est pas de coder pour une protéine. Ces gènes sont transcrits, mais pas traduits.
- Les *ARNt*. Potentiellement 64 différents dans un génomes, en pratique une quarantaine dans les génomes microbiens. Probablement beaucoup plus dans les génomes de mammifères.
- Les *ARNr*: 5S, 16S, 23S pour les procaryotes, 5.5S, 18S et 28S pour les eucaryotes. Ces molécules de 120, 1500 et 3000 nt, repectivement, sont généralement observées en quelques exemplaires chez les procaryotes (7 opérons chez coli), alors que les génomes vertébrés peuvent accueillir plusieurs dizaines de copies identiques.
- Les *ARNsn* (small nuclear RNA) éléments du spliceosome.
- Les *ARNsno* , guides de méthylation.
- Les micro-ARN: ARN régulateurs de longueur 21-22nt agissant comme des ARN interférants, synthétisés à partir d'un precurseur en épingle à cheveux
- Autres ARN 4.5S, 10Sa, Spot42, DicF, MicF, OxyS, DsrA, 6S (procaryotes) produits des gènes XIST, H19, IPW, 7H4, His-1, NTT (mammifères).

Les gènes non-codants

Détection

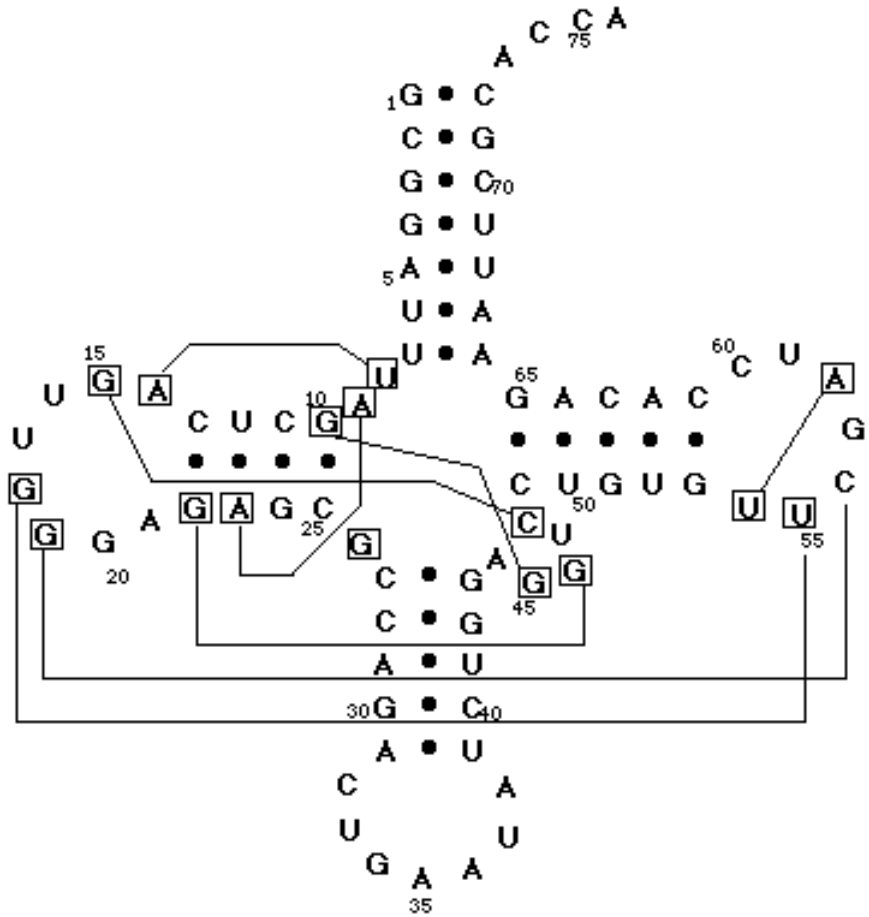
- ★ Les ARN sont généralement transcrits par les polymérases I et III: n'utilisent pas les mêmes promoteurs que les gènes protéiques: les programmes classiques ne fonctionnent pas.
- ★ Il n'existe pas de "banque de données d'ARN" pour faire des recherches par homologie. (Mais on peut utiliser Genbank).
- ★ Les séquences des ARN sont généralement peu conservées: les recherches d'homologie échouent souvent.
- ★ Les ARN non messagers ne sont pas polyadénylés: on ne les retrouve donc théoriquement pas dans les banques d'EST.
- ★ Il n'y a bien sûr pas d'ORF à détecter
- ★ Les gènes d'ARN sont fréquemment interrompus par des introns.

Ces difficultés rendent peu applicables les approches conventionnelles, sauf dans le cas des ARNr 16-18S et 23-28S, qui possèdent des zones conservées suffisamment grandes pour être détectées par Blast. Le nombre de représentants de ces ARN dans Genbank est de plus considérable: plus de 1500 ARNr 16S-23S, plus de 200 ARNr 23S-28S, représentant tous les règnes vivants. On est donc pratiquement certain de détecter la présence de tels gènes par simple Blast contre Genbank.

Les gènes non-codants

Programmes de Détection

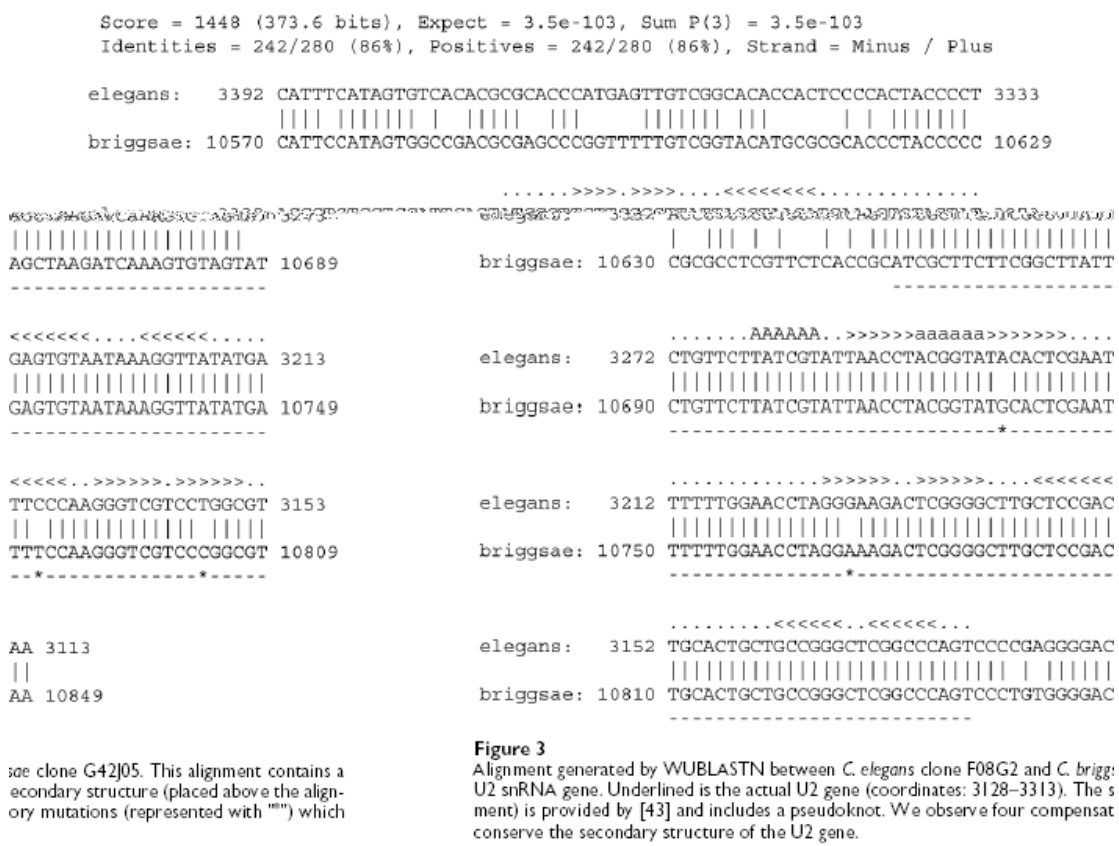
- Dans *TRNASCAN* de Fichant et Burks, les éléments importants, appariements et séquences conservées, sont recherchés directement par le programme. D'autres programmes existent pour rechercher des ARN spécifiques, mais ils sont plus délicats d'utilisation et restent utilisés surtout par les initiés.



Les gènes non-codants

Détection par comparaison intergénome

- ★ 2000 nt conservés en moyenne dans chaque région intergénique entre homme et souris
- ★ => Séquences régulatrices et gènes non-codants



```
 Score = 1448 (373.6 bits), Expect = 3.5e-103, Sum P(3) = 3.5e-103
 Identities = 242/280 (86%), Positives = 242/280 (86%), Strand = Minus / Plus

elegans:   3392 CATTTCATAGTGTCACACGCGCACCCATGAGTTGTCGGCACACCACTCCCCACTACCCCT 3333
                ||||  |||||||  |  |||||  |||     |||||||  |||      | |  |||||||
briggsae: 10570 CATTCCATAGTGGCCGACGCGAGCCCGGTTTTTGTCGGTACATGCGCGCACCCTACCCCC 10629

                                                .....>>>>.>>>>....<<<<<<<<<.............
AGCTAAGATCAAAGTGTAGTAT 10689                    briggsae: 10630 CGCGCCTCGTTCTCACCGCATCGCTTCTTCGGCTTATT
----------------------                                          -------------------

<<<<<<<....<<<<<<.....                                  .......AAAAAA..>>>>>>aaaaaa>>>>>>>....
GAGTGTAATAAAGGTTATATGA 3213                     elegans:   3272 CTGTTCTTATCGTATTAACCTACGGTATACACTCGAAT
||||||||||||||||||||||                                         |||||||||||||||||||||||||||||| |||||||||
GAGTGTAATAAAGGTTATATGA 10749                    briggsae: 10690 CTGTTCTTATCGTATTAACCTACGGTATGCACTCGAAT
----------------------                                         ----------------------------*---------

<<<<<..>>>>>>.>>>>>>..                                  .............>>>>>>..>>>>>>....<<<<<<<
TTCCCAAGGGTCGTCCTGGCGT 3153                     elegans:   3212 TTTTTGGAACCTAGGGAAGACTCGGGGCTTGCTCCGAC
||  ||||||||||||  |||||                                        ||||||||||||||| ||||||||||||||||||||||
TTTCCAAGGGTCGTCCCGGCGT 10809                    briggsae: 10750 TTTTTGGAACCTAGGAAAGACTCGGGGCTTGCTCCGAC
--*-------------*-----                                        ---------------*----------------------

                                                        .........<<<<<<..<<<<<<...
AA 3113                                         elegans:   3152 TGCACTGCTGCCGGGCTCGGCCCAGTCCCCGAGGGGAC
||                                                             |||||||||||||||||||||||||||||| | ||||||
AA 10849                                        briggsae: 10810 TGCACTGCTGCCGGGCTCGGCCCAGTCCCTGTGGGGAC
                                                               -------------------------
```

sae clone G42J05. This alignment contains a
econdary structure (placed above the align-
ory mutations (represented with "*") which

Figure 3
Alignment generated by WUBLASTN between *C. elegans* clone F08G2 and *C. brigg*
U2 snRNA gene. Underlined is the actual U2 gene (coordinates: 3128–3313). The s
ment) is provided by [43] and includes a pseudoknot. We observe four compensat
conserve the secondary structure of the U2 gene.

QRNA: Eddy et al. 2002