

**Support de cours
Annotation des génomes
(Partie II)**

Recherche des régions codant pour des protéines chez les procaryotes

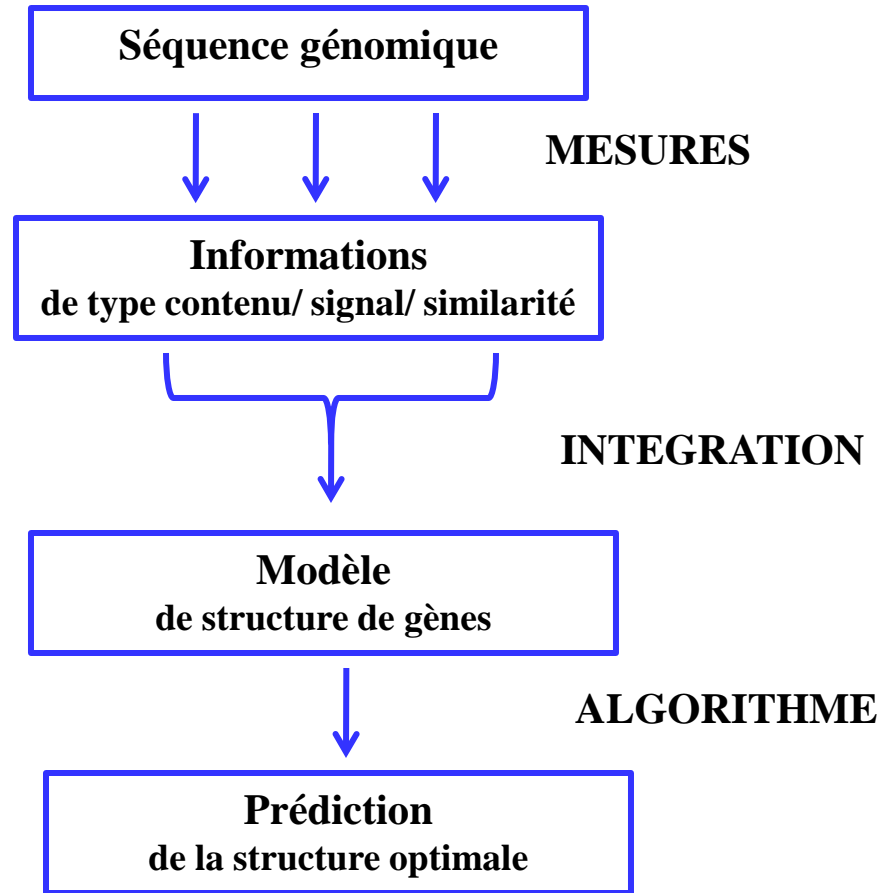
- recherche des ORFs (Open reading frame)
- recherche des unités de traduction. Même si les gènes sont co-transcrits, ils sont en général traduits de façon indépendante (recherche des Shine Dalgarno en 5' du codon initiateur). Permet d'identifier le « bon » codon initiateur.
- recherche des unités de transcription. Chez les procaryotes, certains gènes sont co-transcrits donc recherche de la structure en opérons (promoteurs et terminateurs de transcription)

Recherche des régions codant pour des protéines chez les eucaryotes

- recherche de la structure en exon/intron du gène
- recherche des 5'UTR et 3' UTR
- recherche des promoteurs et des sites de polyadénylation

Recherche des régions codant pour des protéines

Fonctionnement schématique d'un logiciel de prédiction de gènes

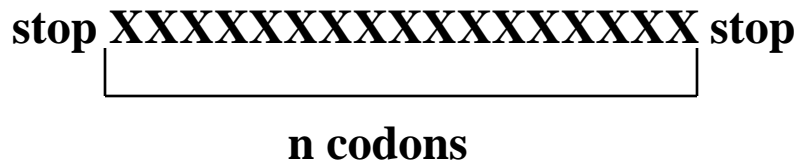


Une méthode simple: ORF Finder (NCBI)

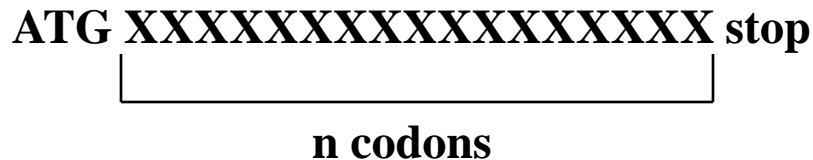
Recherche les phases ouvertes de lecture, les ORFs, dans les 6 cadres de lecture (les 3 cadres du brin direct et les 3 cadres du brin complémentaire).

Attention problème de sémantique :

Alors qu'une ORF est normalement définie entre deux codons stop



Dans ORF Finder, elle est définie entre un ATG et un codon stop



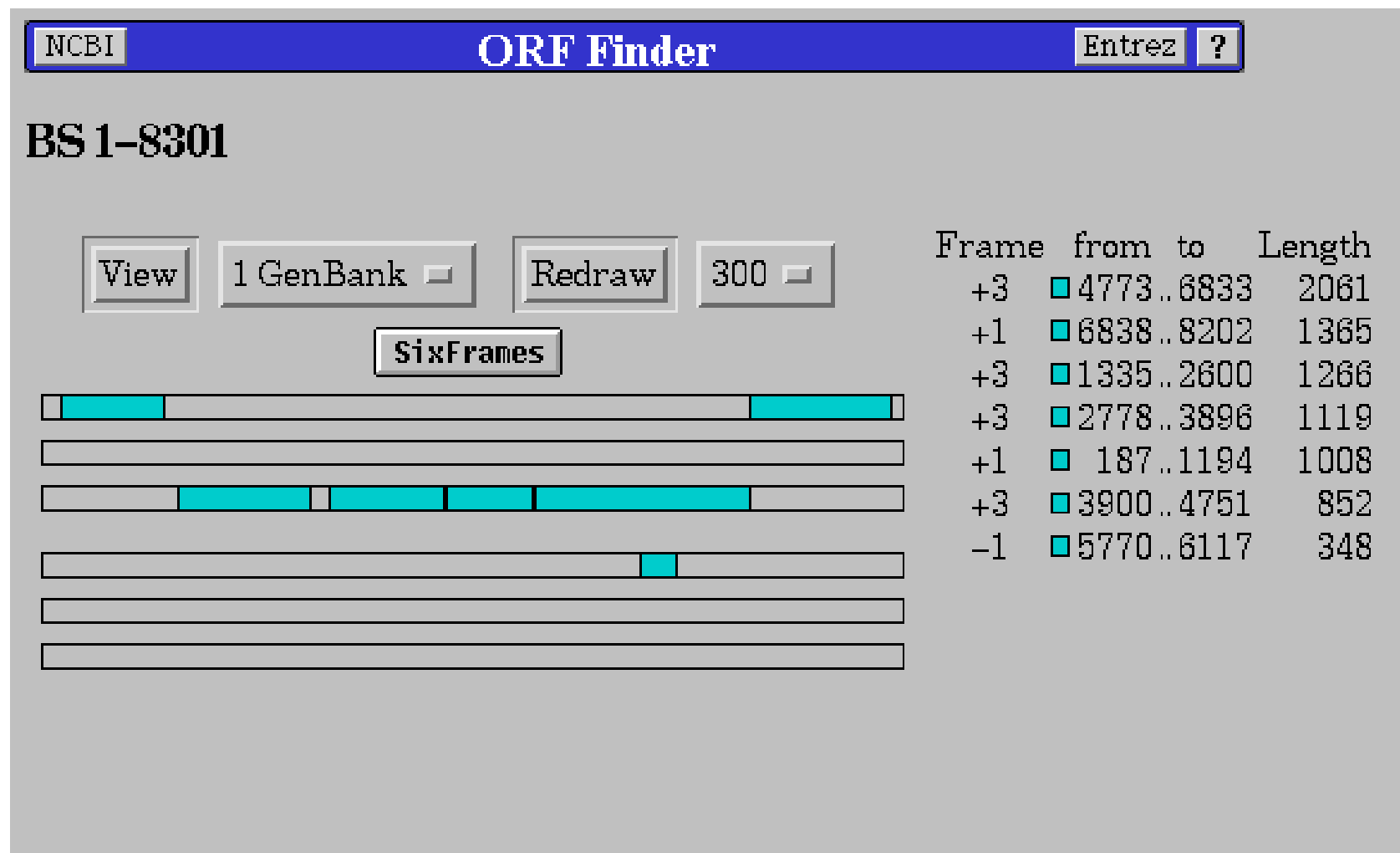
On considère en général que les ORFs supérieures à 100 codons (300 pb) comme étant potentiellement codantes (analyse statistique a montré que bien que des gènes de taille inférieure à 100 codons existent, la majorité des petites ORFs étaient des faux positifs, donc lors de l'annotation d'un génome, dans un premier temps on ne retient que les ORFs de taille supérieure ou égale à 100 codons).

>BS 1-8301

tttcgaggaaaatgtgcaataaccaactcatttcccgggcaattccgcgcg
gttccgaatgatacgaacaactgagactgagccgcaaattggttcagtctt
tttacatggcagccagagggtttgtgcacttgacatttgtgaaaaagaa
agtaaaatattttactaaaacaatgcgagctgaataatggaggcagatac
aatggcgacaattaaagatatcgcgaggaagcggttttcaatctcaa
ccgtttcccgcgttttaaaataacgatgaaagcctttctgttcctgatgag
acacgggagaaaaatctatgaagcggtcgaaagctcaattaccgcaaaaa
aacagtaaggccgctggtgaaacataattgcgtttttatatattggctgacag
ataaagaagaattagaagatgtctattttaaaacgatgagattagaagta
gagaaactggcgaaagcattcaatgtcgatatgaccacttataaaatagc
ggatggaatcgagagcattcctgaacatacgggaagggtttattgccgtcg
gcacattttcagatgaagagctggctttcctcagaaatctcactgaaaac
ggcgtgttcacgattcaactcctgatcccgatcattttgactcggttaag
gcccgatattggcacaatgacaaggaagacggtaaacatcctgactgaga
aggggcataagagcatcggttttatcggcggcacatacaaaaatccgaat
accaatcaggatgaaatggacatccgtgaacaaaccttcagatcctatat
gagggaaaaagccatgctggacgagcgctatatattttctgtcatcgcgat
tctctgtagaaaacggctaccgcctgatgtcagcagcgatcgacacatta
ggcgatcagcttccgactgcttttatgattgcagcggacccgattgcagt
gggctgtctgcaagccctgaacgaaaaaggaattgccataccaaacaggg
taagcattgtgagtatcaacaacatcagcttcgcgaagtatgtctcgct
cctctgacgacgtttcatattgatatacatgaattatgtaaaaaacgctgt
tcaattactgcttgaacaagtgcaggacaagagaagaacggtaaaaaacat
tatatgtgggcgcagaattaatcgctcaggaagagtatgaattaaggatga
cttaggacactaagtcatTTTTTTTatttaggtaaaaaaatttactctatga
agtaaatagtttgtttacacattttctcaggcatgctatattatctttaa
agcgctttcattcctaccgaaagggtgacaatcaatgaaaatggcaaaaa
agtgttccgtattcatgctctgcgcagctgtcagtttatccttggcggct
tgcggcccaaaggaaagcagcagcgccaaatcgagttcaaaaagggtcaga
gcttgttgatgggaggataaagaaaagagcaacggcattaaagacgctg
tggctgcatttgaaaaagagcatgatgtgaaggtcaaagtcgttgaaaaa
ccgtatgccaaagcagattgaagatttgcgaaatggatggaccggccggcac
aggccctgacgtgttaacaatgccaggggaccaaatcggaaccgctgtca
cggaaggattactcaaggaattacatgtcaaaaaagacgttcaatcactt
tatactgacgcttccattcagctctcaaatggtagatcaaaaagctttatgg
actgccaaaagcggctcgaaacgactgtgcttttttacaacaaagatctca
tcacagaaaaggaattgcccaaaacgctggaagagtgggtacgactattcc

Exemple traité : fragment de 8300 pb du génome de *Bacillus subtilis*

Résultat de ORF Finder : ORFs de plus de 300 pb



Limites d'ORF Finder:

- ne prend en compte qu'un seul codon initiateur *ATG* alors que les codons *GTG* et *TTG* sont aussi des codons initiateurs chez les procaryotes (chez *B. subtilis* *GTG* 13%, *TTG* 9%)
- ne prend pas en compte le biais de l'utilisation des triplets existant dans les phases codantes car structurées en codons.

Traitement de l'information de type contenu

Prise en compte du biais de l'utilisation des triplets existant dans les phases codantes par rapport aux régions non codantes car structurées en codons.

Biais dans l'utilisation des codons dus à :

- la dégénérescence du code génétique (61 codons → 20 aa)
- la composition en bases de l'organisme ou de la région génomique (isochores chez les vertébrés) (riche ou pauvre en C+G)
- du taux d'expression du gène : il a été montré chez *E. coli* que les gènes fortement exprimés utilisaient préférentiellement certains codons correspondant aux ARNt les plus abondants dans la cellule (efficacité de la traduction, coadaptation codons/ARNt).



utilisation de méthodes statistiques prenant en compte ces biais d'utilisation des codons.

Plus récemment avec l'augmentation des données pour établir les systèmes de référence, prise en compte de la composition en hexanucléotides (mots de longueur 6).

Présentation de GeneMark

(Borodovsky et al., Nucleic Acids Res.,22,4756-67)

La méthode repose sur le modèle probabiliste suivant appelé modèle de Markov:

Hypothèse 1: La probabilité d'observer une base à une position donnée dépend:

- des bases précédant cette position
- de sa localisation dans le codon

Modélisé par

modèle de Markov homogène pour les régions non-codantes.

modèle de Markov non-homogène pour les séquences codantes.

Hypothèse 2: Une région particulière ne peut être que dans un des 7 états suivants:

- 1. codant en phase 1 sur le brin direct
- 2. codant en phase 2 sur le brin direct
- 3. codant en phase 3 sur le brin direct
- 4. codant en phase 4 sur le brin indirect
- 5. codant en phase 5 sur le brin indirect
- 6. codant en phase 6 sur le brin indirect
- 7. non-codant

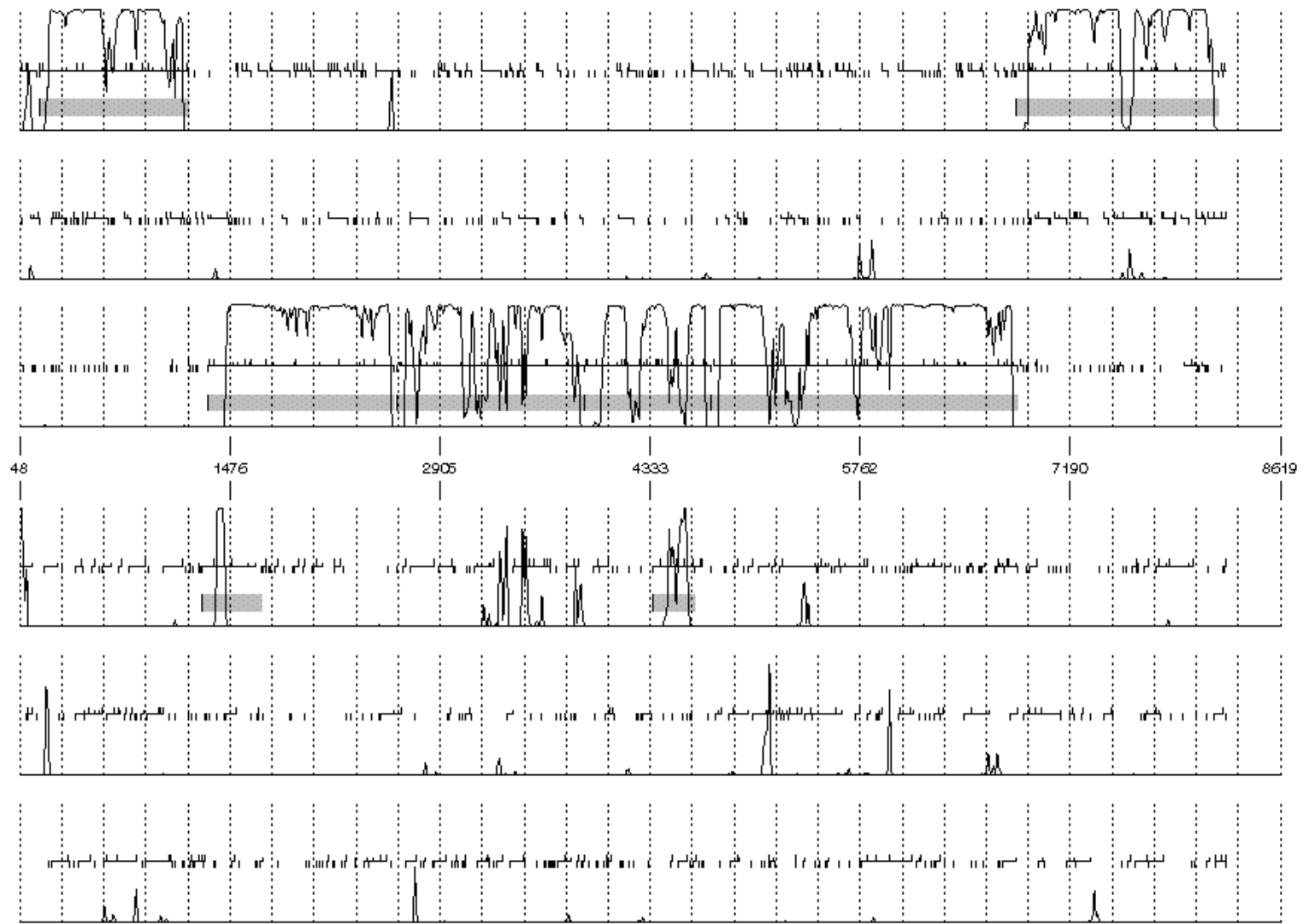
Prédiction : calculer les probabilités d'observer la région dans un état i sachant que l'un des 7 états est réalisé (formule de Bayes).

Modèle de Markov

Un modèle de Markov d'ordre k appliqué aux séquences ADN est entièrement défini par les deux probabilités suivantes :

$$\left[\begin{array}{ll} P_0(w_1^k) & \longrightarrow \text{Probabilité initiale du mot } w^k \\ P(x / w^k) & \longrightarrow \text{Probabilité d'observer } x \text{ sachant que le mot } w^k \\ & \text{le précède} \end{array} \right.$$

Résultat graphique de GeneMark sur le fragment de *B. subtilis*



Résultat de GeneMark sur le fragment de *B. subtilis*

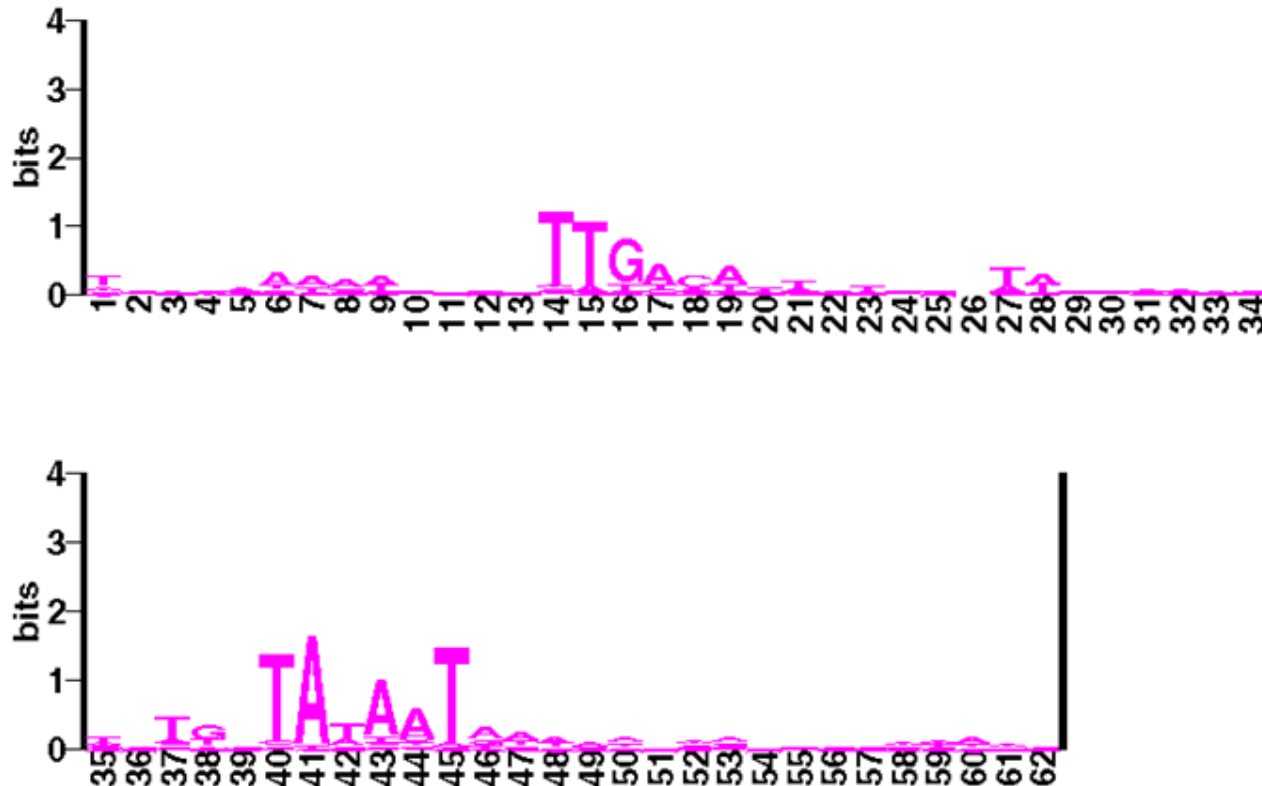
List of Open reading frames predicted as CDSs, shown with alternate starts
(regions from start to stop codon w/ coding function >0.50)

Left end	Right end	DNA Strand	Coding Frame	Avg Prob	Start Prob	
-----	-----	-----	-----	-----	-----	
187	1194	direct	fr 1	0.80	0.99	-> A Idem ORF Finder
202	1194	direct	fr 1	0.81	0.89	
367	1194	direct	fr 1	0.82	0.29	
436	1194	direct	fr 1	0.81	0.03	
481	1194	direct	fr 1	0.80	0.02	
1335	2600	direct	fr 3	0.85	0.01	-> B Idem ORF Finder
1341	2600	direct	fr 3	0.85	0.00	
1365	2600	direct	fr 3	0.87	0.08	
1500	2600	direct	fr 3	0.93	0.07	
1527	2600	direct	fr 3	0.93	0.00	
1581	2600	direct	fr 3	0.92	0.03	
2631	3896	direct	fr 3	0.73	0.67	
2640	3896	direct	fr 3	0.73	0.77	-> C
2778	3896	direct	fr 3	0.76	0.53	-> ORF Finder
2814	3896	direct	fr 3	0.75	0.02	
2868	3896	direct	fr 3	0.74	0.40	
3900	4751	direct	fr 3	0.65	0.17	-> ORF Finder
3912	4751	direct	fr 3	0.66	0.02	
3966	4751	direct	fr 3	0.71	0.34	-> D
4116	4751	direct	fr 3	0.71	0.11	
4137	4751	direct	fr 3	0.70	0.02	
4158	4751	direct	fr 3	0.69	0.06	
4770	6833	direct	fr 3	0.85	0.76	
4773	6833	direct	fr 3	0.85	0.82	-> E Idem ORF Finder
4815	6833	direct	fr 3	0.86	0.12	
4890	6833	direct	fr 3	0.85	0.05	
5226	6833	direct	fr 3	0.85	0.01	
6838	8202	direct	fr 1	0.79	0.03	-> ORF Finder
6877	8202	direct	fr 1	0.82	0.86	-> F
6913	8202	direct	fr 1	0.83	0.67	
6925	8202	direct	fr 1	0.83	0.01	
6931	8202	direct	fr 1	0.83	0.01	
6952	8202	direct	fr 1	0.84	0.00	
7009	8202	direct	fr 1	0.85	0.63	
7057	8202	direct	fr 1	0.86	0.28	

Traitement de l'information de type signal

Différentes façon de représenter la conservation des séquences impliquées dans un processus donné (promoteur lors de la transcription, ribosome binding site lors de la traduction, jonction d'épissage etc...) et ensuite de rechercher ces « signaux » dans une nouvelle séquence.

Compilation of *Bacillus subtilis* sigma A-dependent promoter elements



Recherche des signaux d'initiation de la traduction

Programme utilisé: Patscan

Motif du Shine-Dalgarno recherché : **GGAGG 6...11 DTG** correspond à la présence de la séquence GGAGG à 6 ou 11 pb en amont d'un codon AUG, GUG ou UUG.

Résultats:

BS: [189, 204]	:	ggagg	cagataca	atg	->	A
BS: [3175, 3192]	:	ggagg	tcgacttttt	ttg	->	dans le gène C
BS: [3887, 3902]	:	ggagg	cataaggt	atg	->	D
BS: [4760, 4775]	:	ggagg	agaatgtg	atg	->	E
BS: [7501, 7516]	:	ggagg	atttgccg	gtg	->	dans le gène F

Donc:

Gène A : début en 202

Gène D : début en 3900

Gène E : début en 4773

Les autres SD des gènes B, C et F trouvés avec une autre représentation (matrice de poids) car ils sont modifiés.

AAGGAGGTG		consensus
GAAAGGGTG	7	ATG pour B
AGA GAGGTG	6	GTG pour C
GGGGGGATG	5	ATG pour F

Unités de traduction prédites

202	1194	direct	fr 1	-> A	Patscan
1335	2600	direct	fr 3	-> B	SD modifié
2640	3896	direct	fr 3	-> C	SD modifié
3900	4751	direct	fr 3	-> D	Patscan
4773	6833	direct	fr 3	-> E	Patscan
6877	8202	direct	fr 1	-> F	SD modifié

Recherche des unités de transcription

Chez *B. subtilis*, l'initiation de la transcription fait intervenir le facteur sigma A qui reconnaît une séquence spécifique localisée environ en -10 et -35 pb du +1 de transcription.

Séquence consensus: TTGACA 16...35 TATAAT

Grand nombre de promoteurs de type sigma A identifiés expérimentalement chez *B. subtilis*:



matrices de poids

Représentation : Matrice de poids ou PWM (Position Weight Matrix)

Un exemple simple : 242 séquences de promoteurs (-10) chez *E. coli* :

Normalisation de la matrice : log matrice $\log_2(f_{b,i}/P_b)$

$f_{b,i}$ = fréquence observée de la base b à la position i dans toutes les séquences

P_b = fréquence de cette base dans l'ensemble du génome

Pos.	1	2	3	4	5	6
A	-2.76	1.88	0.06	1.23	0.96	-2.92
C	-1.46	-3.11	-1.22	-1.00	-0.22	-2.21
G	-1.76	-5.00	-1.06	-0.67	-1.06	-3.58
T	1.67	-1.66	1.04	-1.00	-0.49	1.84

Le rapport $f_{b,i}/P_b$ est une mesure de l'écart entre fréquence observée et attendue.

Résultats de la recherche des promoteurs

Utilisation du programme Patscan et de la matrice de poids



	-35		-10
BS:[1264,1292]:	tttaca	cattttctcaggcatgc	tatatt
BS:[131,158] :	ttgaca	tttgtgaaaaagaaag	taaaat

Recherche des terminateurs de transcription

2 types de terminateurs :

- **Rho dépendant.** Une protéine, le facteur ρ , aide au décrochage de l'ARN polymérase.
- **Rho indépendant.** Pas d'intervention de protéines.

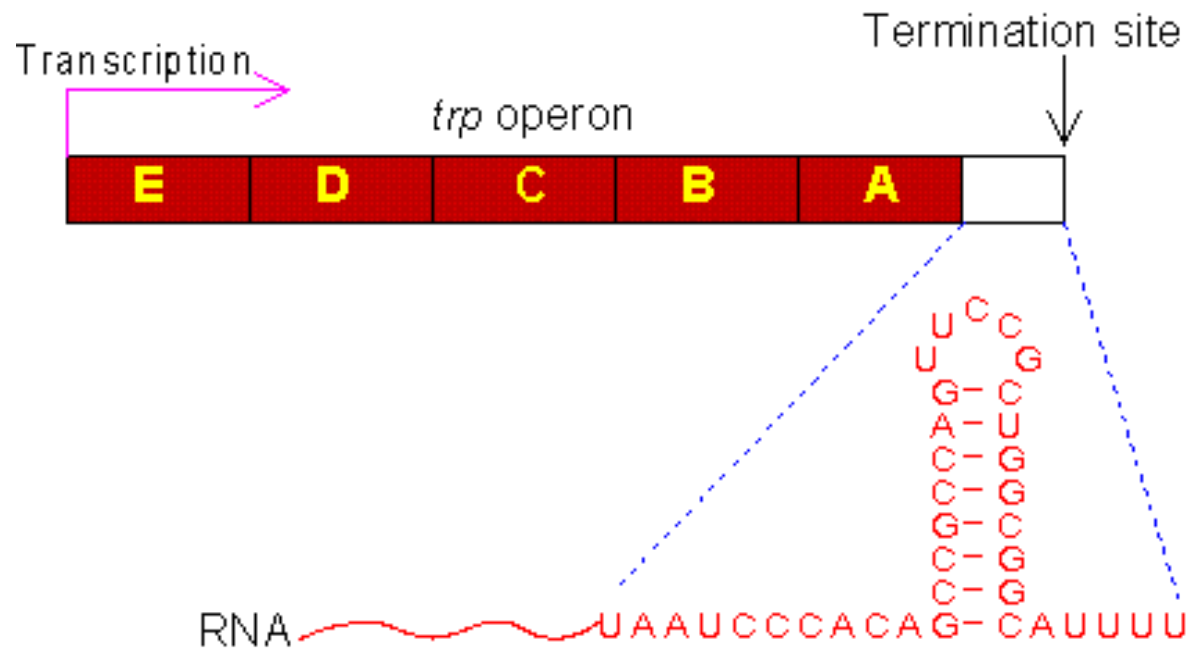
Au niveau séquence, on ne sait modéliser que les seconds.

Mécanisme proposé pour les terminateurs Rho indépendant.

Quand l'ARN est en cours de transcription, on a une hybridation ARN/ADN sur environ 12 pb. Le site de terminaison de la transcription est précédé par une séquence capable de former une structure secondaire stable. Il y a compétition entre la formation de cette structure et l'appariement avec l'ADN. La présence d'un poly(U) en cours de synthèse déplace l'équilibre en faveur de la tige-boucle et il y a alors décrochage de l'ARN et arrêt de la transcription.

Dans les séquences, on va donc rechercher des séquences répétées inversées suivies d'un poly(U).

Termineurs rho indépendant



Terminateurs rho indépendant

Le modèle : Formation d'une tige boucle en amont d'une région riche en U qui déstabilise l'appariement ADN/ARN et conduit au décrochage de l'ARN.

Deux classes de terminateurs:

- petite tige de 5 à 7 pb très stable et d'une boucle de 4 pb suivie d'une région riche en U.
- une longue tige qui peut se décomposer en deux tiges imbriquées l'une dans l'autre.
 - La première plus stable doit faire au moins 3 pb de long avec un appariement GC à son pied.
 - La seconde est incluse dans la première et comporte au moins 3 appariements. Elle est généralement moins stable que la première. La boucle est de 3 à 7 pb de long.

Résultat de la recherche des terminateurs sur le fragment de *B. subtilis*

1199-1223

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

6843-6866

	T
A	A
	G.T
	C-G
	T.G
	C-G
	G-C
	C-G
	C-G
	A
	T
	T
	C
	T
	T
	T

75-103

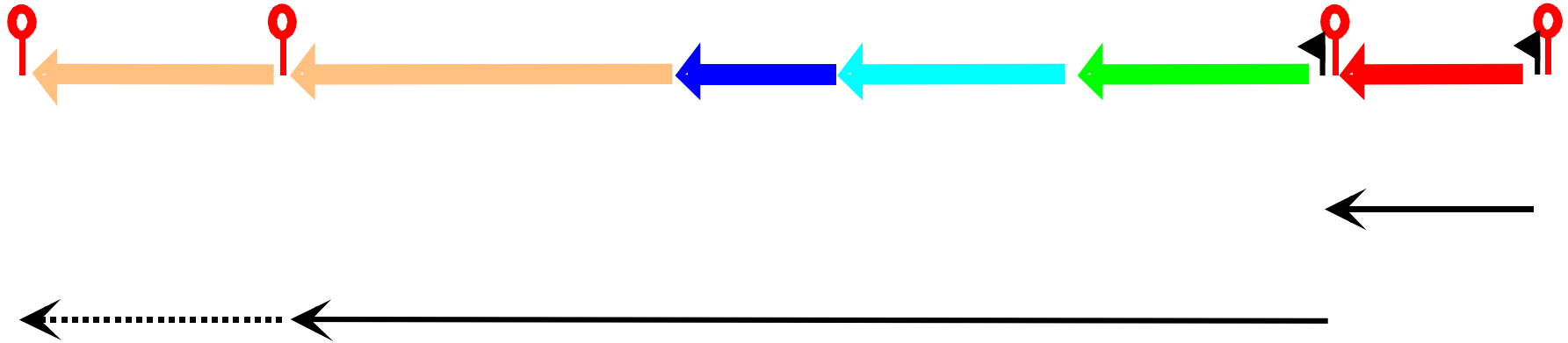
	A	A
C		A
	G.T	
	C-G	
	C-G	
	G.T	
	A-T	
	G-C	
	T-A	
	C-G	
	A-T	
	G-C	
	T	
	T	
	T	
	T	
	T	




8215-8256

	T	C
A		A
	A-T	
	A-T	
	C-G	
	A-T	
	A-T	
	T-A	
	G.T	
	T.G	
	A-T	
	G-C	
	A-T	
	G-C	
	T-A	
	A-T	
	C-G	
	C-G	

ATCATT

Prédiction des unités de traduction et de transcription



-  terminateur rho-indépendant
-  promoteurs de transcription de type sigma
-  transcrit putatif

Prédictions fonctionnelles

Identification

- homologues
- motifs
- domaines

Localisation cellulaire

- fragments trans-membranaires
- peptide signal

Structure

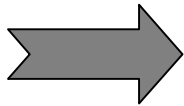
- secondaire
- tertiaire

Recherche de liens fonctionnelles

- réseaux de régulation
- voies métaboliques
- interactions moléculaires

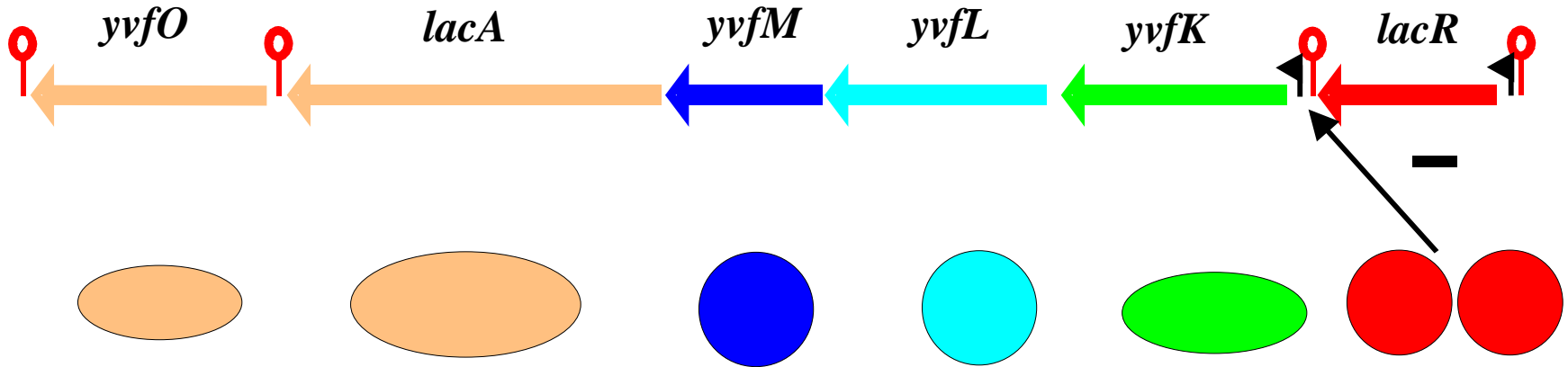
Prédiction fonctionnelle

Recherche par similitude dans les bases de données: programme BLAST

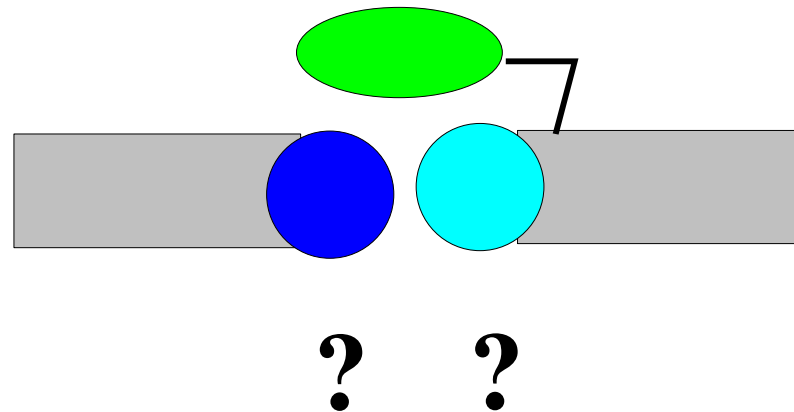


- **A LACR** protéine régulatrice de type LacI/GalR
- **B YVFK** protéine affine d'un ABC transporteur
- **C YVFL** perméase d'un ABC transporteur
- **D YVFM** perméase d'un ABC transporteur
- **E LACA** galactosidase
- **F YVFO** arabino-galactosidase

Synthèse des résultats



Système à la membrane



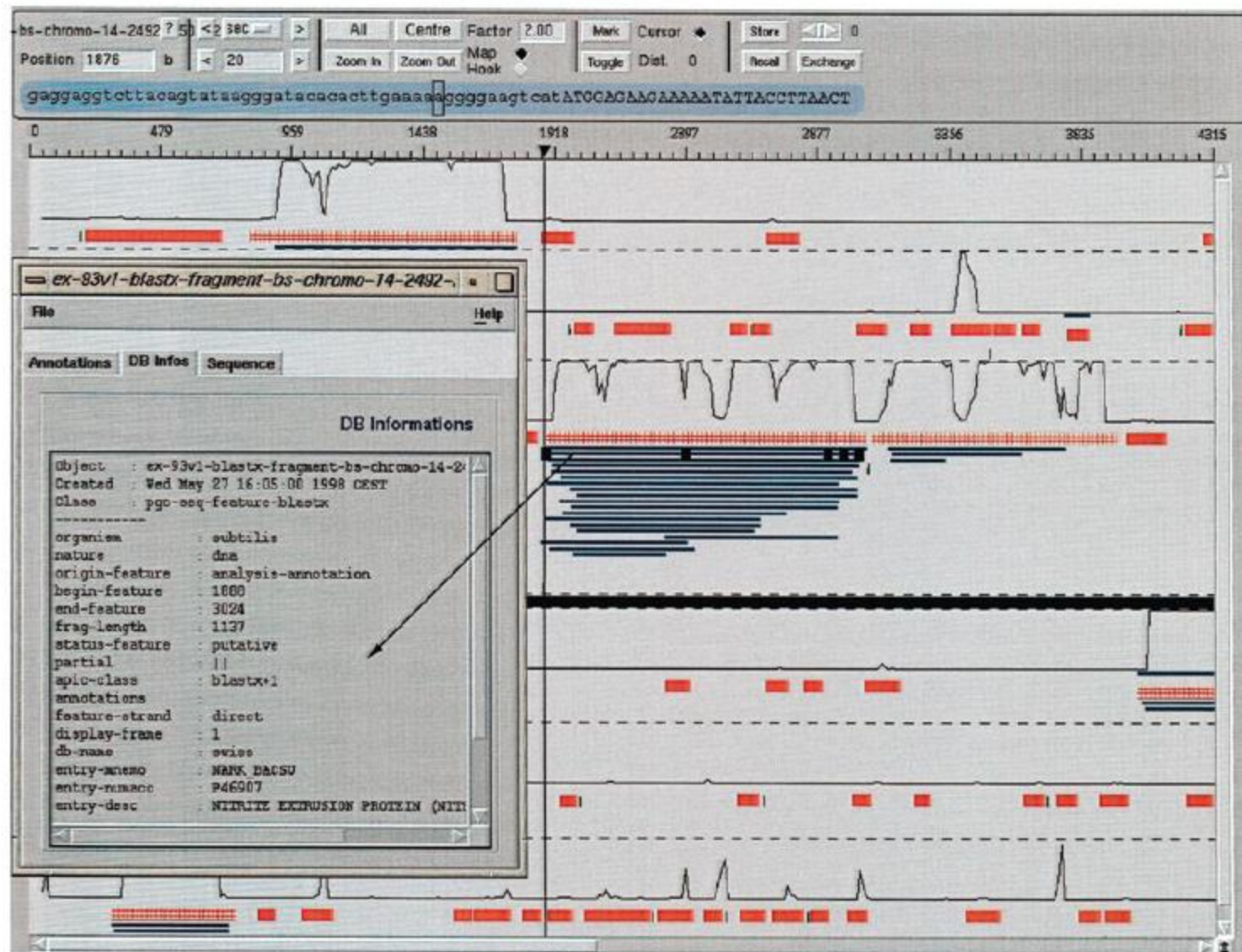


Fig. 6. Superimposition of results from three different strategies in the Imagene Result Manager (Cartographic Interface). Results obtained with the CDS searching strategy are shown in red boxes (CDSs) and green triangles (RBSs), those obtained with the Blastx strategy are shown in blue rectangles and, finally, the GeneMark© coding predictions are displayed as black curves. The results are given in each of the six frames.

Information de type similarité

Information externe à la séquence elle-même (de type extrinsèque)
contrairement au contenu statistique ou aux signaux qui sont internes à la
séquence (de type intrinsèque)

Comparer la séquence à analyser avec des séquences connues peut permettre
de refléter la présence de gènes et donner des informations sur leur
structure. Notamment, la structure en exons/introns pour les gènes
eucaryotes.

Types de séquences utilisées pour la comparaison :

- les ADNc
- les EST
- les protéines
- des séquences génomiques

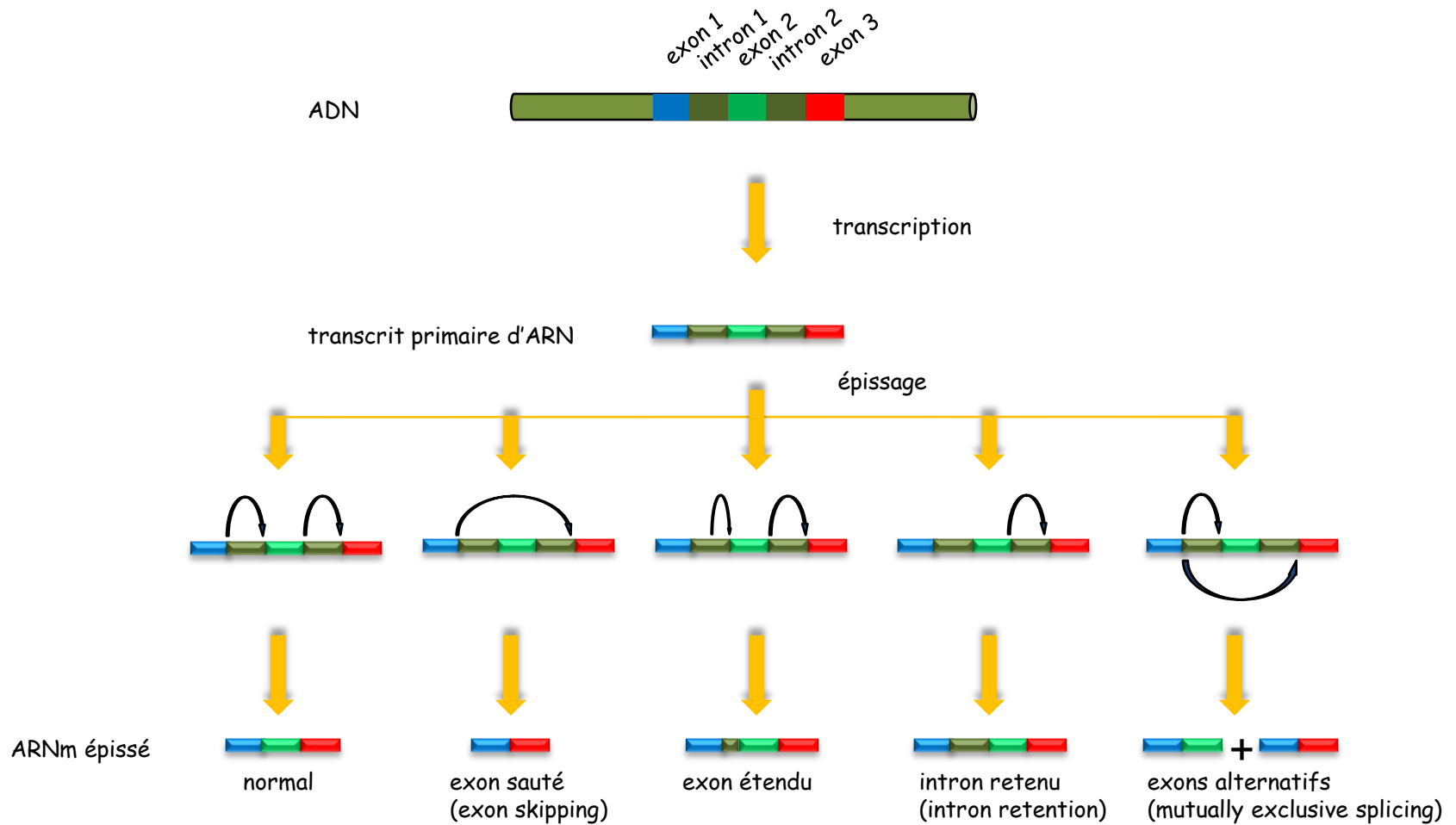
Information de type similarité

Méthodes prédisant la structure en exons-introns par alignement de la séquence génomique soit avec un ARNm (ou un ADNc), soit avec une protéine. Parmi les plus utilisés, on trouve :

Méthode	Séquence de référence	Référence
BLAT	ARNm ou protéine	Genome Research 12(4):656-64 (2002).
sim4	ARNm	Genome Research 8:967-74 (1998).
Scipio	ARNm ou protéine	BMC Bioinformatics 9:278 (2008)
GeneWise	protéine	Genome Research 14(5):988-95 (2004)
GenomeWise	ADNc, EST	Genome Research 14(5):988-95 (2004)
WebScipio	protéine	Bioinformatics 12:270 (2011)

L'extension du logiciel WebScipio permet de rechercher une forme spécifique d'épissage alternatif (exons mutuellement exclusifs)

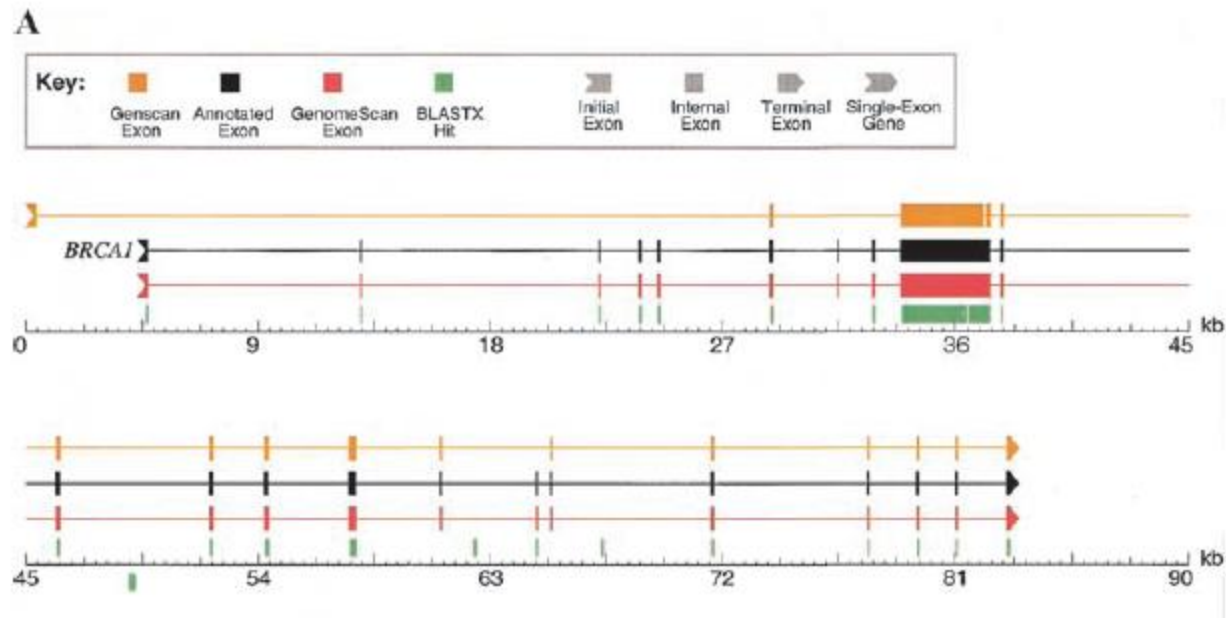
Cinq manières d'épisser un ARN



Prise en compte de la similarité

Avec des séquences protéiques : GENOMESCAN (intégration de cette information dans le modèle GENSCAN).

Exemple d'un résultat de prédiction (extrait de *Genome Research* (2001), 11, 803-816)



Prise en compte de la similarité

Avec des séquences génomiques : TWINSKAN (intégration de cette information dans le modèle GENSCAN.

Codage de la conservation

```
          10      20      30
123456789|123456789|123456789|123456789
ATTTAGCCTACTGAAATGGACCGCTTCAGCATGGTATCC
|:|:|.....|:|:|:|:|:|:|:|:|:|:|:|:|
```

Fig. 1. An example DNA sequence together with the corresponding conservation sequence.

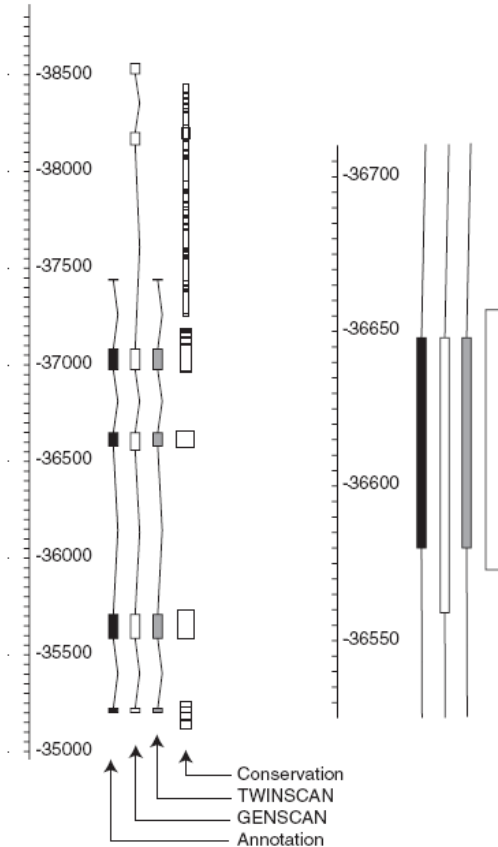


Fig. 5. Detailed view of the annotation, gene predictions and conservation at the L44L gene (AAB47245.1) from the *Mus musculus* Bruton's tyrosine kinase locus (U58105.1). The magnification at right shows the region around exon 3. The width of boxes representing BLAST alignments corresponds to the quality of the alignment. The image comes from ACEDB.

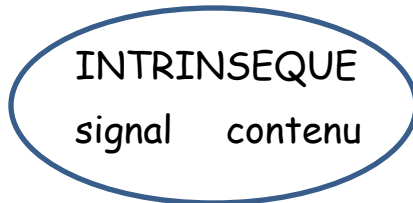
Exemple d'un résultat de prédiction
(extrait de Bioinformatics (2001), 17 suppl. 1, S140-S148)

Evolution de l'intégration des sources d'informations

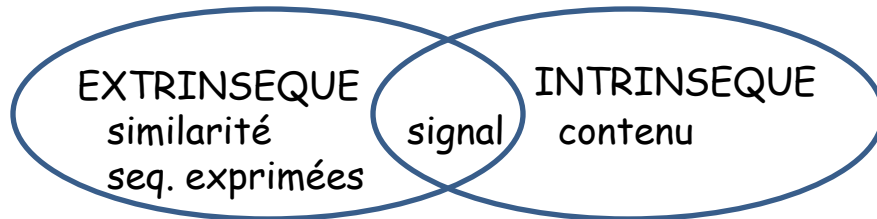
(extrait de la thèse de Sylvain Foissac, 2004)



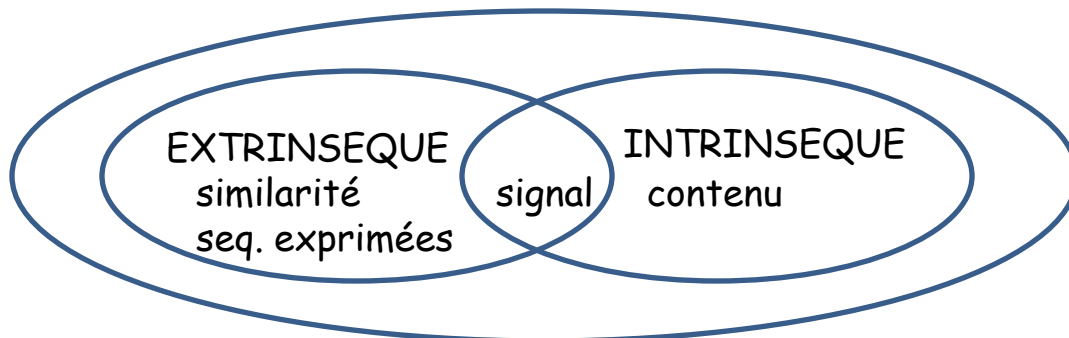
Deux sources traitées par des méthodes indépendantes (ex : Staden, 1984; Gelfand, 1990)



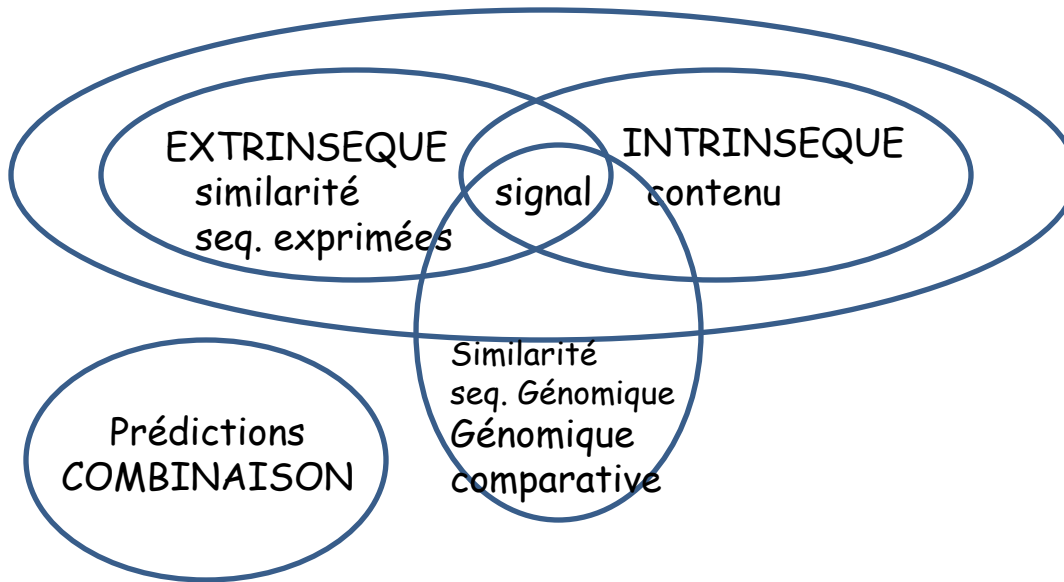
Intégration de ces deux sources dans un même logiciel (ex : Guigo et al., 1992, logiciel GENSCAN (Burge et Karlin, 1997)



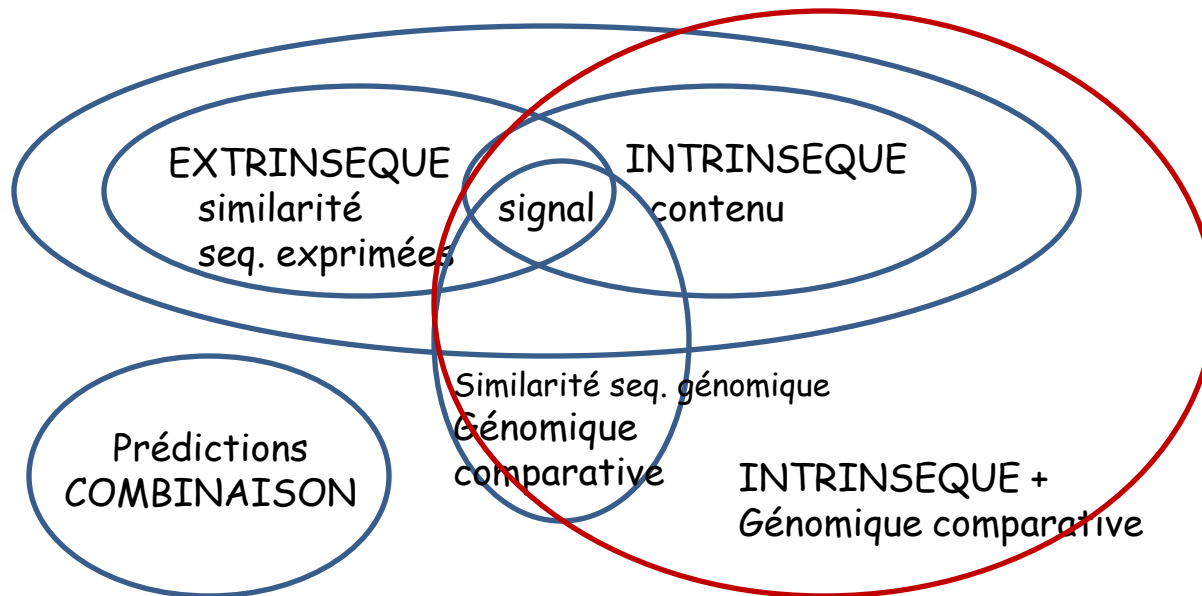
Augmentation des données d'expression, prise en compte de la similarité de séquences (Borodovsky et al., 1994, Fickett, 1995)



Intégration de ces deux types d'information dans de nombreux logiciels. GENOMESCAN (Yeh et al., 2001) résulte de l'intégration de similarité protéique dans GENSCAN



Conservation entre séquences génomiques exploitées en combinaison seulement avec des informations de type signal



Génomique comparative intégrée uniquement dans des méthodes intrinsèque. Par exemple, TWINSCAN (Korf et al., 2001) intègre la génomique comparative dans GENSCAN.