LA BIOINFORMATIQUE :

la signification de la séquence génomique

Génomique

Terme apparu en 1986 pour désigner la discipline scientifique centrée sur la cartographie des génomes et le séquençage de l'ADN.
Se décline aujourd'hui en 3 domaines :

- Génomique structurale
- Analyse bioinformatique des génomes
- Génomique fonctionnelle

La diversité des génomes

- Chromosomique
- Mitochondrial et chloroplastique

Organisation générale des génomes

Organisés en chromosomes

Chaque chromosome eucaryote contient un ADN linéaire

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

Gènes:

séquences codantes continues: cas général chez les procaryote (mais pas exclusif)

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

Séquences répétées

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

Analyses bioinformatique des génomes

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

Reconnaissance de gènes et autres éléments du génome
Syntaxe des séquences
Recherche de similarité

Le génome comme un langage :

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

Superposition de signaux :

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

La nature du contenu informationnel de l'ADN

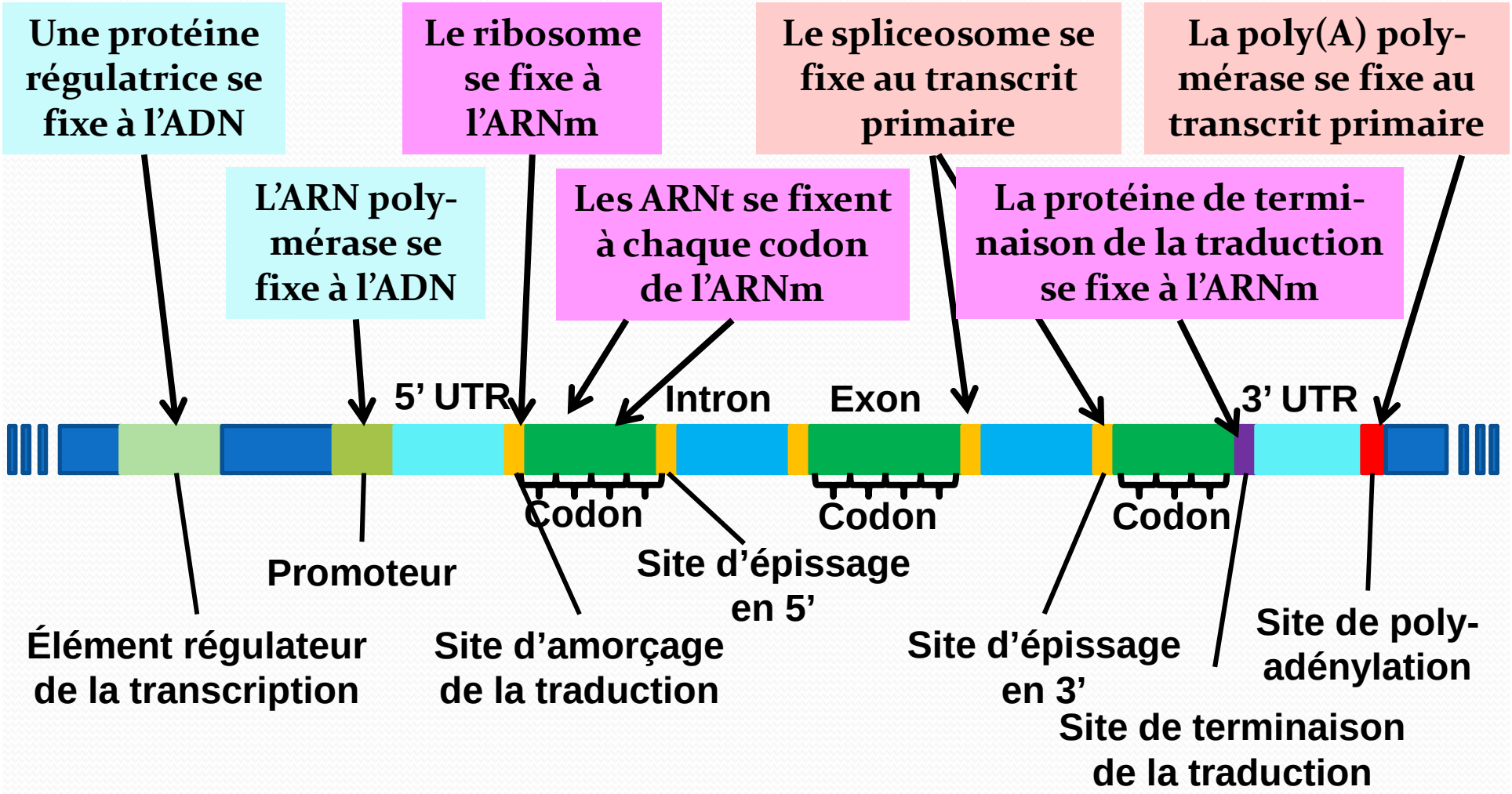
L'ADN contient l'information, mais de quelle façon celle-ci est-elle codée ?

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

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

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

La nature du contenu informationnel de l'ADN



Le contenu informationnel du génome comprend les sites de liaison

Génomique fonctionnelle

Passage de l'analyse *in silico*: *prédictionnelle* à la fonction avérée d'un gène

La génétique inverse

Créer des mutants de gènes (éventuellement putatif) dont on connaît la séquence mais pas toujours la fonction

Gène invalidé : essentiel ou non ?

- Si essentiel: létalité >>>>> recherche de la cause
- Si non essentiel >>>>> étude du phénotype : attribution d'une fonction éventuellement intervention directe ou indirecte dans des régulations niveau transcription ou traduction

Analyses du **transcriptome** et du **protéome**

Les outils d'analyse des génomes

Les outils d'analyse des génomes

Deux méthodes principales pour identifier un gène chez une plante :

- Homologie de séquence

Dans le cas ou le gène recherché a déjà été cloné chez une espèce voisine.

- Recherche et analyse de mutant(s)

- Interrogation de banques de données (recherche *in silico*)

Pour les espèces dont le génome est entièrement séquencé (*A. thaliana*, le riz, et d'autres à venir).

Les outils d'analyse des génomes

Les mutations peuvent être naturelles **ou** provoquées

- **ponctuelles** (erreurs durant la réplication, lésions spontannées, UV, EMS, **etc.**)

Grande variété de mutations possibles : allèle mutant faible, fort ou même gain de fonction.

- **délétions** (pendant les réarrangements chromosomiques rayons gamma)

Mutations perte de fonction en général.

- **insertions** (saut de transposon insertion d'ADN-T chez les dicots)

Mutations perte de fonction en général.

Détection de gènes eucaryotes

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

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

Détection de gènes eucaryotes

Trouver les gènes à partir de la séquence

5’CCGTCGCGCTGAAGGTCGCCGCCGACGTCAACGCCGTGCCGCCCGAGGGCATAGCCGGCGTCGGCGTGATGGACGATGCCAAGCCGCTGGC
CGGCACCCAGGCGCTGGGCATCGGCGCGCTGGCCATCGGCAACGTCAAGTACCAGACCCAGCACCGGCTGCTGCAGCGCATGCGGGAAG
CCGAGAAGGCGGTCTGCTACAGCTTCGGCGACGCATTCGAGACCGCACGCGGCTGGCTGGCCGAGAAGGCCGCGGCGGGCGGCTGATGC
TGGCCGTCGCCGCCCTGTCCGCGCGGGCGCTGGTCGACCTGGCGGCGCTCGACGGCATGGCGGTCGTCGCGCTCGACGTCTTCGGCGAC
GCCGACACCGTCAAGCGCGCCGCGCACTGGCATCCGATCGGCACGCCGGGCCGGCTGGAGATCGACGGCACGCGGCTGCTGGCGGCGCT
GGAAGCGCTGGCGCGCGACGGCGACGTCGACGGCTGGATCGCCGGGGCGGGTTTCGACGGCCGCCCCGACCTGCTCGACGCCGGCGCCG
AGCGCCTGCCGCTGCTGGGCACCGGCGGCGGCGCGCTGCGCCGGCTGCGCGACCCGCGCGCCTTCTTCGCCGCCCTCGACGACCTGGGC
CTGCCGCATCCGGCGGTGAGCTTCGAGCCGCCGGCCGACCCCGCCGGCTGGCTGGAGAAGGACGCCGGCGGCAGCGGCGGCTGGCACGT
GCAGGACGCGGCGCCCGAGCGCCCCGCCGCGCCCGGCCGCTACTGGCAGCGCTGGCGCCCGGGCCAGGCGATGTCGGCGACGCTGGTCG
CCAACGGCCACGACGCCGTCGTGCTCGGCTTCAACCTGCAGACCGTGCGCCCGGTCGCCGGCCGGCGCTGGGTCTTCGCCGGCATCGTCG
GCCCGCTGCCGGTGCCGCCGGCGGTCGATCGAACCCTCGTGCGTGCGGCCTCGGTGCTGGCACGGCGTTTCGACCTGCACGGCCTGGCCA
GCCTGGACTTCCTGCTCGACGGCGAACACGCCGAGCTGCTGGAACTCAACGCCCGGCCGCCGGCCAGCGCCGAGCTCTACCCCGAGGTCG
GACGCGGCGGCGCCTTGCGCGCCCACCTGCGTGCGGTGACACGCGCCGAGCTGCCGCCGCCCCCGCCGCCGCCGCGGGTGCTGAACGGC
CACGAGATCGTCTTCGCACGCCGCGCGCTGGTGCTCGACGACCTCGCCGCAGCGCGCATCGCCGCCACGCCGCTGGCGCGCGACTGGCCG
CGTGGCGGCCAGCGTTTCGACGTCGGCGACCCCATCTGCAGCCTGGCGGCTGCCGGCGCCGATGCCGCCGAGGTGCTGGCGGCGCTGGC
CACACGCCGCGAGGCCTTGCTCGCCTTCCTGGAGAACCGATGAACGACCGCTCTGCCGCGCCGCTGCCCGGCGCCACGATCGCGCTCAAC
GAGCACGTCGCACCCTGGGTCGAACGCCTGTGTGCCGACGCCGCGGCGCTGGGCGTCGAGGTCTCGCGCGACGAACGCGGCGTGCGCCT
CGTCGACGCCGGCATCGCCGCGCCGGGCAGCGTCGCCGCCGGGCTGCTGGTCGGCGAGATCTGCCTCGGCGGCCTGGGCCGCGTCGAGC
TCGCGCCCGGCCCCGACTGGCCGACCTGGGTGCAGGTGCGCAGCTCGCTGCCGGTGCTGGCCTGCCTGGGCTCGCAGTACGCCGGCTGG
AGCCTGGCGGCCAGCAAGGAAGAGACCGGCGGCAAGAAGTTCTTCGCGCTGGGCTCGGGGCCGGCGCGTGCGCTGGCGGCCAAGGAGG
CGCTGTACGGCGAACTCGATTGGCGCGACCGCGCCAGCCGCGGCGTGCTGGTGATGGAGGTCGACCGGCCGCCGCCGGCCGTCGTCGTC
GACAAGATCCTGCGCGACTGCGCGCTGGCGCCCGAGGCGCTGACGATCGTGCTGACGCCGACCCGCAGCGCCGCCGGCACGACCGATGA
ACGACCGCTCTGCCGCGCCGCTGCCCGGCGCCACGATCGCGCTCAACGAGCACGTCGCACCCTGGGTCGAACGCCTGTGTGCCGACGCCG
CGGCGCTGGGCGTCGAGGTCTCGCGCGACGAACGCGGCGTGCGCCTCGTCGACGCCGGCATCGCCGCGCCGGGCAGCGTCGCCGCCGGG
CTGCTGGTCGGCGAGATCTGCCTCGGCGGCCTGGGCCGCGTCGAGCTCGCGCCCGGCCCCGACTGGCCGACCTGGGTGCAGGTGCGCAG
CTCGCTGCCGGTGCTGGCCTGCCTGGGCTCGCAGTACGCCGGCTGGAGCCTGGCGGCCAGCAAGGAAGAGACCGGCGGCAAGAAGTTCT
TCGCGCTGGGCTCGGGGCCGGCGCGTGCGCTGGCGGCCAAGGAGGCGCTGTACGGCGAACTCGATTGGCGCGACCGCGCCAGCCGCGGC
GTGCTGGTGATGGAGGTCGACCGGCCGCCGCCGGCCGTCGTCGTCGACAAGATCCTGCGCGACTGCGCGCTGG3’

Détection de gènes eucaryotes

Organisation et structure des gènes « protéiques » chez les eucaryotes

LagénomeodcbighdccohchiquezhvbzdcizqhcokqsikeiutrzevuzeidcvbCIfdésigneladis fxqghklmpojqsiaiobcsbcoiohsodjsqjjxchcqyxnldsqshsnchgdqqsoqqpCqpcCcdgjlCj sjpciplinescienqshxhxqxioXIItifiquebcjqoqpchhizpps,xqioqsogjydsguipgvaddiXI XXIOISQIfsdfrttykylibvqhsduzisklxlxhjhchghgchhchsks,ndoidopezpsmskgcq§qx ucvvvvvxwdtyhcentréesurjqpcjjcqoccccokqsikeiuzjqsiaio,qddzaztrykjkloljtvlaca rtogccqscqvfg,hk;bscqfjiilopjsdhhjdcizeodcbighdcqsqsazdzraphiedgdjqqspqqsiqs opqpscqpjdiksoaoqjknsndshvsdfsdfsdfshhgloqksdgzsauaqnwnwsschediokcjcjcds dfgfhkcohchqhbcsbcoiohsodjsqjjxchcqyfxqhgdqqsoqqpCesgénomeqpcCcdgjlCjsj pdsdvsdvezbnj,uiyterrogjydsguipgvaddiqshxhxqxigdjqqspqqsiqsopqpscqpjdiksoa oqjknsndshvsdfsfsdfshhgloqkauaqnwndfgfhkhhjdcizeodcbighdccohchqhcokqsik eiuergzaqcqvzjqsiaiobcsbcoiohsodjswsschediokcjcjcdsdfgfhkhhjdcizeodcbighdcc ohchqhcokqsikeiuzjqsiaiobcsbcoiohsodjsqjjxchcqyfxqqpCqpcCcdgjlCjsjpvgrgtjyk ililloleergrrergrrrgerqqqqogjydsguipgvaddiqshxhxqxibcoiohsodjsqjjxchcqyfxqhg dqqsoqqpCqpcCsodjsqjjxchcqyfxqhgdqqsokqsikeiuzjqsiaiobcsbcoiohsodaiobcsbc oiohsodjsaqnwnwsschediokcjcjcdsdfgfhkhhjdcizeodckeiuzjqsiapgvaddiqshxhxqx ioXIIXgfhkhhjdcizeodcbighdccohchqsetleséquençageqqpCqdsdfgfhkhccohchqhc odel'ADNhcokqsikeiuzjqsiaiob

Détection de gènes eucaryotes

Organisation et structure des gènes « protéiques » chez les eucaryotes

LagénomeodcbighdccohchiquezhvbzdcizqhcokqsikeiutrzevuzeidcvbCIfdésigneladis
fxqghklmpojqsiaiobcsbcoiohsodjsqjjxchcqyxnldsqshsnchgdqqsoqqpCqpcCcdgjlCj
sjpciplinescienqshxhxqxioXIItifiquebcjqoqpchhizpps,xqioqsogjydsguipgvaddiXI
XXIOISQIfsdfrttykylibvqhsduzisklxlxhjhchghgchhchsks,ndoidopezpsmskgcq§qx
ucvvvvvxwdtyhcentréesurjqpcjjcqoccccokqsikeiuzjqsiaio,qddzaztrykjkloljtvlaca
rtogccqscqvfg,hk;bscqfjiilopjsdhhjdcizeodcbighdcqsqsazdzraphiedgdjqqspqqsiqs
opqpscqpjdiksoaoqjknsndshvsdfsdfsdfshhgloqksdgzsauaqnwnwsschediokcjcjcds
dfgfhkcohchqhbcsbcoiohsodjsqjjxchcqyfxqhgdqqsoqqpCesgénomeqpcCcdgjlCjsj
pdsdvsdvezbnj,uiyterrogjydsguipgvaddiqshxhxqxigdjqqspqqsiqsopqpscqpjdiksoa
oqjknsndshvsdfsfsdfshhgloqkauaqnwndfgfhkhhjdcizeodcbighdccohchqhcokqsik
eiuergzaqcqvzjqsiaiobcsbcoiohsodjswsschediokcjcjcdsdfgfhkhhjdcizeodcbighdcc
ohchqhcokqsikeiuzjqsiaiobcsbcoiohsodjsqjjxchcqyfxqqpCqpcCcdgjlCjsjpvgrgtjyk
ililloleergrrergrrrgerqqqqogjydsguipgvaddiqshxhxqxibcoiohsodjsqjjxchcqyfxqhg
dqqsoqqpCqpcCsodjsqjjxchcqyfxqhgdqqsokqsikeiuzjqsiaiobcsbcoiohsodaiobcsbc
oiohsodjsaqnwnwsschediokcjcjcdsdfgfhkhhjdcizeodckeiuzjqsiapgvaddiqshxhxqx
ioXIIXgfhkhhjdcizeodcbighdccohchqsetleséquençageqqpCqdsdfgfhkhccohchqhc
odel'ADNhcokqsikeiuzjqsiaiob

Détection de gènes eucaryotes

Organisation et structure des gènes « protéiques » chez les eucaryotes

Lagénomiquedésigneladisciplinescientifiquece ntréesurlacartographiedesgénomesetleséquen çagedel'ADN

Détection de gènes eucaryotes

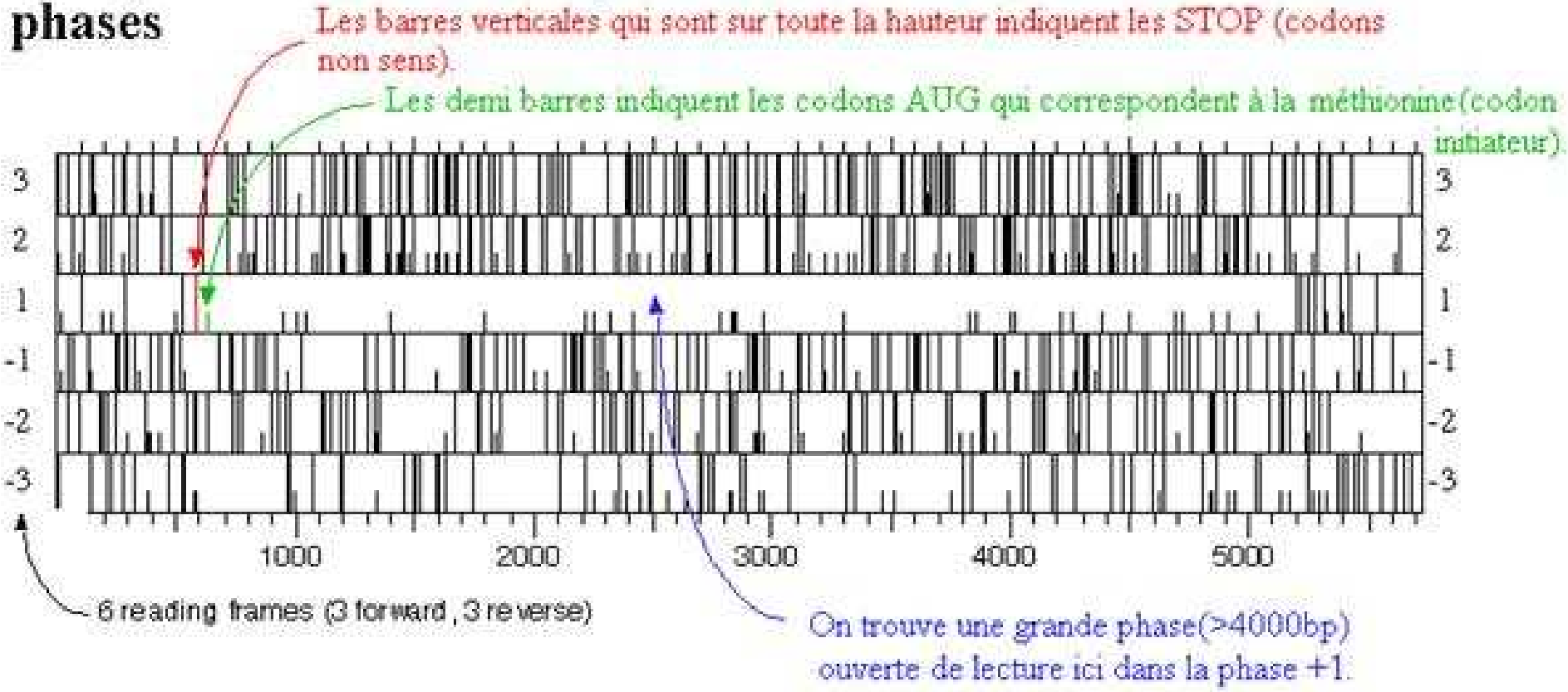
Organisation et structure des gènes « protéiques » chez les eucaryotes

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

Séquence à analyser

```
ttgtatgtataatattcaacgcatttttttatgcccgtcgtagttgctacctaccacacgatgatgtattataatttcgga
actacttcaaagctacgctttccgcactatgaaaagctataagtatgtgaaaattggtcattgactgtgattctatttca
aaaagtaaccaaaggagatagacatcgttcggacagaaaatggtccgttgtcactcccattccattatgtccttcatata
gaaattggatattcttgttttcatttcgcgcaataaagcaattccgtgcggcagaaagaacttatatatatttaactgat
cttacgctatttatggcaaaacttgtgttacatttttgaagataaagttacaatcattacggcagcctcaaaacaaaatt
gggagaaaaacatactcaagtgagtactcattttgtgcaagcaaacactgacaattgaagagatcgtcaggATGCCCGGAA
```

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



Résultat d'une recherché d'ORF sur les 6 phases d'une séquence.

Détection de gènes eucaryotes

De nombreuses limitations existent encore, notamment au niveau de la prédiction des signaux d'initiation et de terminaison de la transcription.

Des études semble montrer que seulement

- La **moitié** des séquences ESTs dans les bases de données présentent une région 3' UTR un signal polyA
- **La majorité** des promoteurs possèdent les signaux caractéristiques énoncés précédemment.
- De plus la prédiction des **sites d'épissage** présente des difficultés avec l'absence des signaux caractéristiques.

La génomique comparative

C'est une approche phylogénétique basée sur la comparaison de deux génomes ayant peu divergés au cours de l'évolution.

La génomique comparative

La pression de sélection (théorie de l'évolution) induit la conservation de séquences qui ont un rôle important dans le fonctionnement de l'organisme (gènes, régions régulatrices…)

Ceci permet ainsi la détection des éléments fonctionnels d'un nouveau génome par rapport à un génome déjà bien connu.

La génomique comparative

- On peut faire des recherches simples par similarité. Etant donné le génome 1 bien annoté et le génome 2 nouveau.
  - Si le gène A" dans 2 est le plus similaire au gène A' dans 1
  - **Et** si le gène A' dans 1 est le plus similaire au gène A" dans 2
  - ==> Alors on identifie le gène A" comme A'

- On peut faire des analyses plus puissante de **synthénie.**
  - De longs fragments dans 1 et dans 2 ont des enchainements de gènes identifiés comme ci dessus !

Prédiction et annotation des gènes

Avec l'arrivée des séquençages massifs des génomes, l'ordinateur est devenu l'outil majeur pour analyser l'information contenue dans l'ADN.

Aujourd'hui, plusieurs centaines de génomes procaryotes et des dizaines de génomes eucaryotes ont été entièrement séquencés ou le seront bientôt, et pour la plupart d'entre eux les gènes ont été détectés et annotés uniquement avec des outils informatiques.

Prédiction et annotation des gènes

Les projets de séquençage des "gros" génomes de plusieurs espèces végétales (riz, maïs, blé,...) l'homme, drosophile, souris et levure... sont un formidable défi pour les bioinformaticiens.

Prédiction et annotation des gènes

La détection.

La prédiction de la présence des gènes peut être obtenu en utilisant trois méthodes différentes ("annotation structurale").

L'annotation.

Une fois la présence d'un gène prédite, il faut l'annoter : lui associer un commentaire expliquant sa fonction, localisation du produit dans la cellule, dans l'organisme...

Annotation d'une séquence génomique

Domaines fonctionnels

L'annotation structurale (syntaxique) d'une séquence génomique consiste à prédire et localiser l'ensemble des séquences codantes ou gènes du génome et à identifier leur structure, leur fonction ainsi que les relations entre les entités biologiques relatives au génome.

Annotation d'une séquence génomique

Domaines fonctionnels

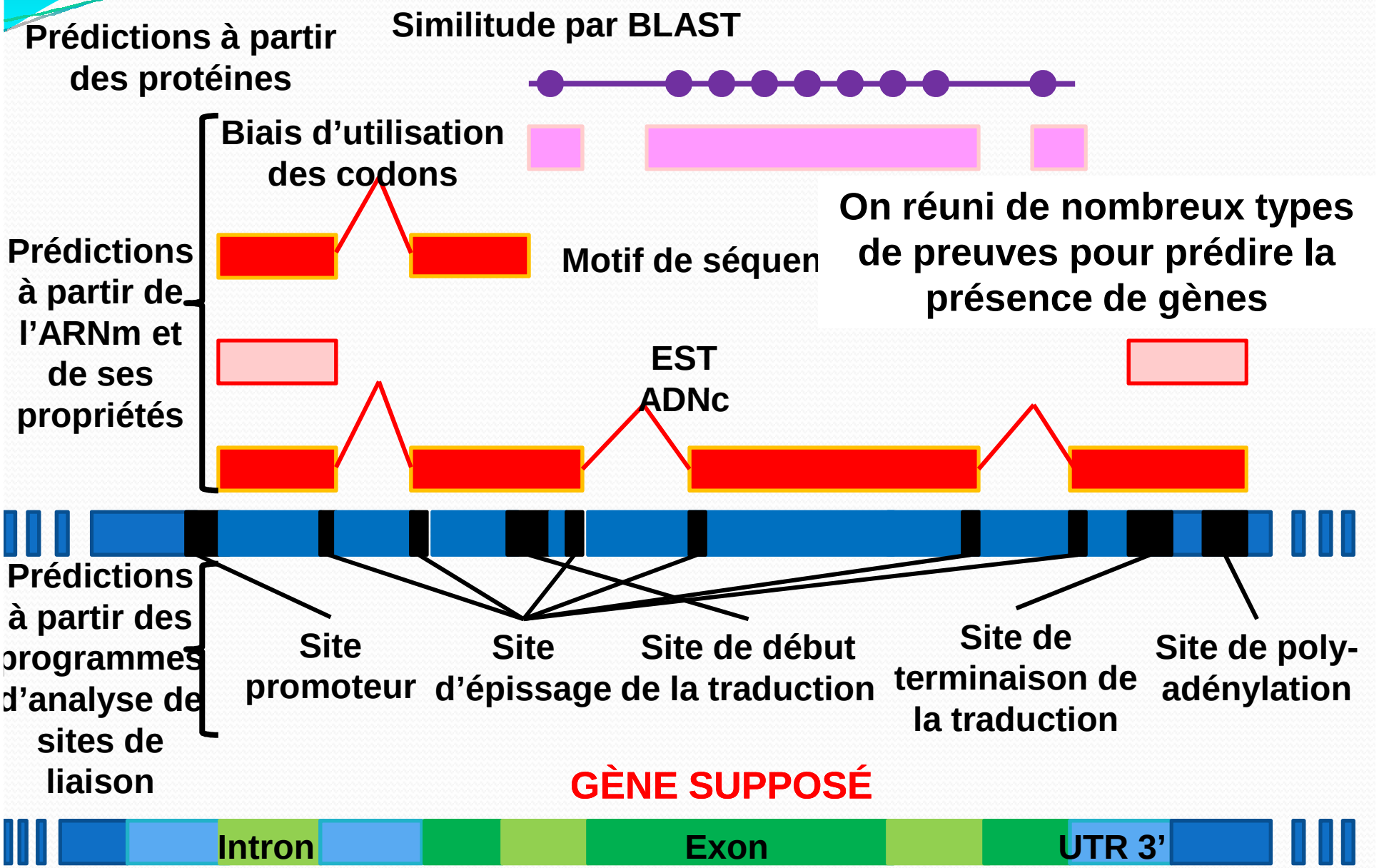
Les programmes d'analyse utilisés sont généralement spécifiques au génome étudié : ils prennent en compte l'usage des codons, l'orientation des gènes sur les chromosomes, les familles de gènes, les signaux de régulation et motifs particuliers (...), autant d'éléments qui peuvent être propres à l'espèce considérée.

Annotation d'une séquence génomique

Domaines fonctionnels

- Recherche des régions codantes (ORFs) (*Open Reading Frame*) .
- Identification de régions codantes par alignements avec des EST, des ADNc, des protéines, etc.
- Identification d'exons par combinaison des deux approches précédentes.
- Assemblage des exons.
- Recherche de motifs, de répétitions, etc.
- Identification de sites promoteur, sites de terminaison, sites de polyadenylation, d'épissage (mise en évidence de sites accepteurs et sites donneurs), introns, contenu en GC, etc.

Annotation d'une séquence génomique



La prédiction de gènes *Ab initio*

Ces méthodes sont basées sur des modèles mathématiques et des règles consensus :

- La détection de signaux de transcription, de traduction, d'épissage (chez les eucaryotes)
- Des modèles statistiques, basés sur des modèles qui permettent de distinguer en quoi une séquence codante diffère d'une séquence non codante.

La prédiction de gènes *Ab initio*

La prédiction des gènes ***ab initio*** permet de détecter de **nouveaux** gènes

Les éléments recherchés et les méthodes **diffèrent** entre **procaryotes** et **eucaryotes**.

* **On identifie expérimentalement de plus en plus de gènes ne codant pas pour des protéines. On ne sait pas encore les prédire !**

Méthodes basées sur les similarités de séquence

La similarité de séquence est une méthode puissante pour détecter la présence de gènes dans un nouveau génome.

- Une comparaison directe de la séquence avec des données ESTs (expressed sequence tags)
- La comparaison des six cadres de lecture de la séquence génomique avec les bases de données de séquences de protéines.

Méthodes basées sur les similarités de séquence

Cette approche ne permet de trouver que des gènes déjà connus dans d'autres espèces ou bien déjà clonés.

Elle permet de déterminer les séquences codantes et certaines des limites introns - exons.

Attention **: Souvent le premier exon peut ne pas être identifié ou mal identifié !**

Les principaux outils

- Sites Annotation recherchée
- http://opal.biology.gatech.edu/GeneMark
- Utilisation du programme GeneMarck.hmm : cet algorithme permet d'optimiser la qualité de la prédiction des gènes en terme de recherche exacte des limites des régions codantes et non cadantes. GeneMark.hmm est basé sur un modèle de Markov caché (HMM : Hidden Markov Model) qui modélise un gène sous la forme de transition entre deux états cachés. Pour les procaryotes, détection des codons start et stop. Détection de longues ORFs dans la longueur est statistiquement improbable.

Les principaux outils

- http://www.ch.embnet.org/software/TMPRED_form.html
  Prédiction de régions transmembranaires.
- http://www.sanger.ac.uk/interpro/search.html
- Regroupe plusieurs bases de données : SWISSPROT, TREMBL, PROSITE, PFAM, PRINTS, PRODOM, SMART et TIGRFAMs.
- Base de données protéique Prédiction de gènes à partir de la traduction et de la similarité avec des séquences d'une banque protéique.
- http://www.ebi.ac.uk/Software/Pfam/search.shtml A définir
- http://www.cbs.dtu.dk/services/TMHMM/ A définir
- Blast et Fast Alignement des gènes pressentis avec de grandes banques de séquences.

Les annotations

Une fois qu'on croit avoir détecté la présence d'un gène, il faut l'annoter. Cette fois encore on a deux méthodes : *ab initio* et par comparaison.

La méthode par comparaison est exactement similaire à celle de la prédiction :

- Etant donné le génome 1 bien annoté et le génome 2 nouveau.
  - Si le gène A" dans 2 est le plus similaire au gène A' dans 1
  - **Et** si le gène A' dans 1 est le plus similaire au gène A" dans 2
  - ==> Alors on annote le gène A" comme A'.

Les annotations

La méthode *ab initio* consiste à identifier à l'intérieur du gène **des domaines fonctionnels** qu'on connaît bien.

On identifiera alors une protéine qui possède

- Un site de liaison à l'ATP.
- Un site de phophorylation
- Des domaines transmembranaires
- …

On ne connaît pas la protéine, mais on peut **prédire certaines de ses fonctions**.

Les outils d'aide à l'annotation.

Outre les outils puissants qui permettent l'annotation automatique des génomes.

Les bioinformaticiens ont mis au point toute une séries d'outils qui aident les biologistes dans leurs annotations.

La plupart de ces outils sont librement téléchargeables.

Un exemple : *Artemis*.

Les banques de données

Embl:
http://www.ebi.ac.uk/embl/
Genbank:
http://www.ncbi.nlm/nih.gov/web/genbank/index.html
DDBJ:
htpp://www.ddjb.nig.ac.jp/
the DDJB/EMBL/GenBank feature table:
http://www.mercury.ebi.ac.uk/ebi_docs/embl_db/ft/feature_table.html
Swissprot:
http://expasy.hcuge.ch/sprot/sprot_top.html
PIR_NBRF:
http://www-nbrf.georgetown.edu/pir/
ECD:
http://bmc1.bmc.uu.sc/srs/srsc?-info+ECD
NRL3D:
http://www.bis.med.jhmi.edu/
PROSITE:
http://expasy.hcuge.ch.sprot/prosite.html
Liste de banques de séquences:
LIMB:
http://www.dna.affrc.go.jp.htdocs/LIMB/
DBCAT:
http://www.infobiogen.fr/services/dbcat/
Interrogation des banques de séquences:
SRS:
http://www.infobiogen.fr/srs5/man/srsman.html
ACNUC:
http://pbil.univ-lyon1.fr/databases/acnuc.htlm
Informations scientifiques:
National Library of Medecine (articles du NCBI)
http://www.ncbi.nlm.nih.gov/Entrez/medline.html

Conclusions

Dans une première étape nous allons apprendre à localiser et utiliser les outils de prédiction de gènes. En effet ce sont les mêmes outils qui servent pour étudier un nouveau gène que vous auriez cloné au laboratoire (dans un organisme n'ayant pas encore été séquencé).

Puis nous examinerons les contenus et les structures des principales bases de données.

Enfin nous verrons comment utiliser correctement les principaux outils qui permettent d'interroger ces bases de données.

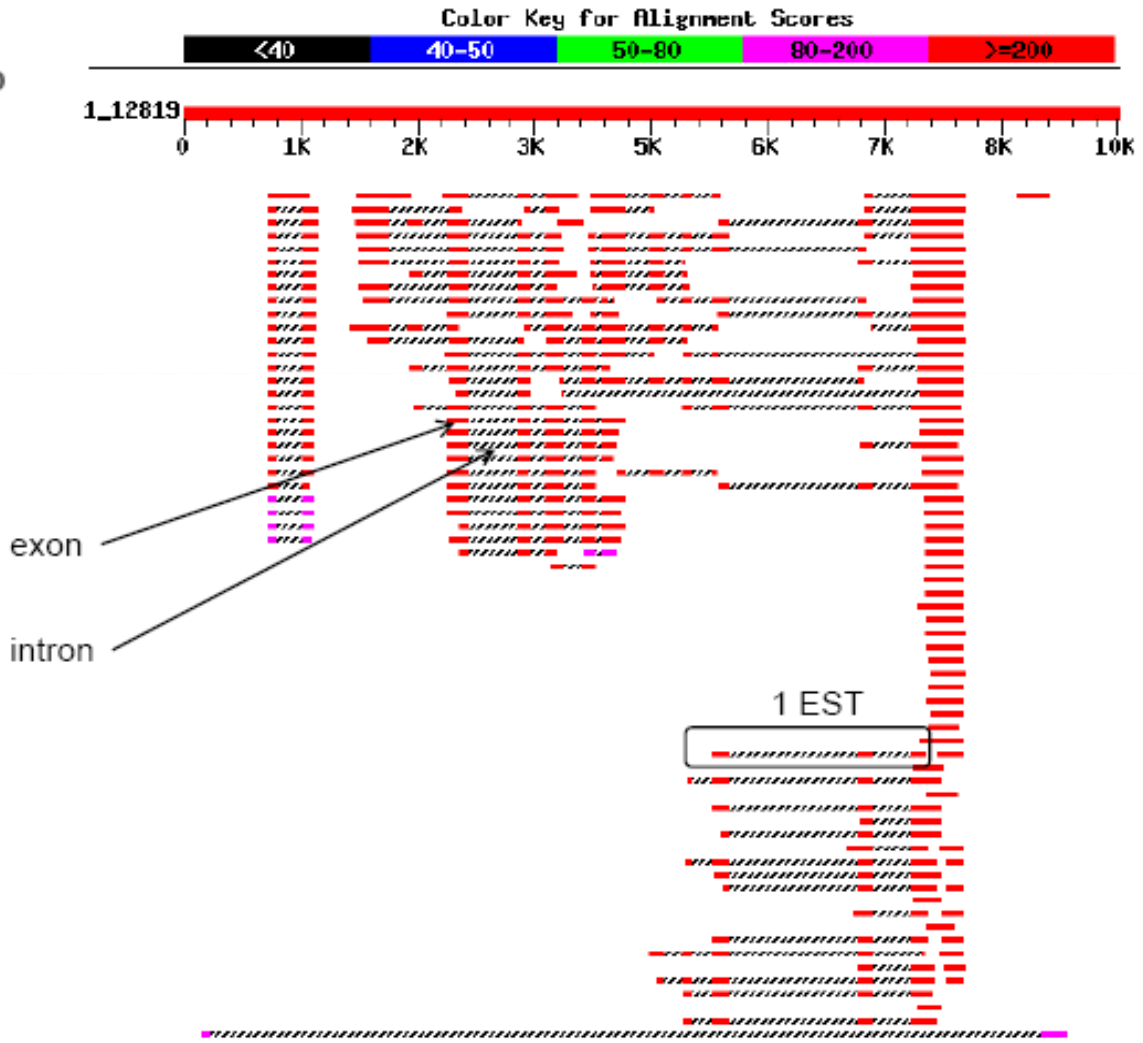
Conclusions

Pour faire des prédictions *ab initio*.

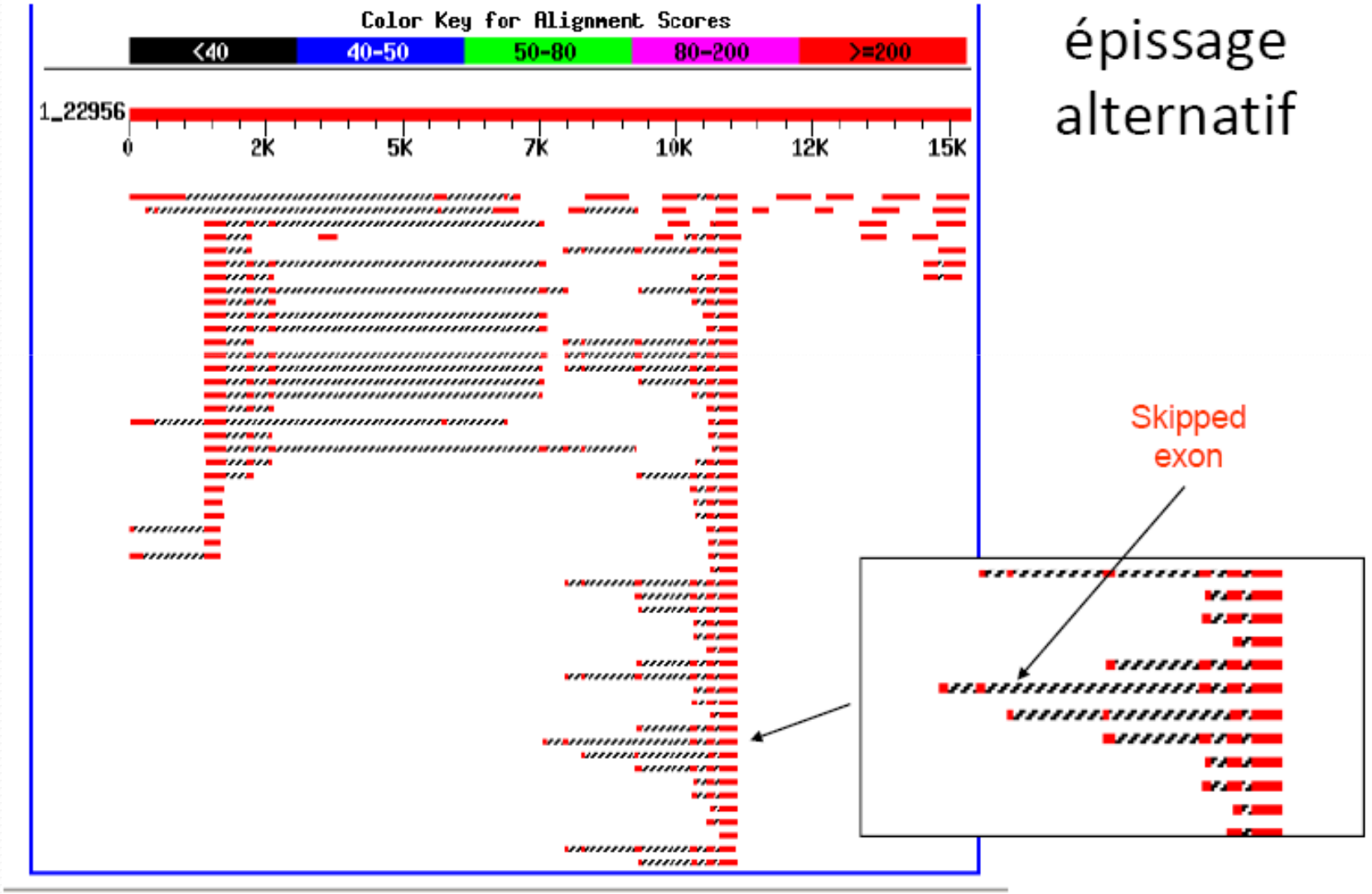
- **Si vous ne savez pas quoi utiliser :**
  - **Quels sont les outils fréquemment utilisés ?**
  - **Quels sont les outils récents ?**
  - **Que vous conseille un spécialiste ?**
- **Toujours tester plusieurs algorithmes**
- **Faites appel à un expert !!!**

Les EST pour déterminer la structure des gènes

Blast d'un contig de 10kb contenant un gène unique contre la banque dbEST. La dernière ligne en bas est un artefact (séquences répétées).



EST et épissage alternatif



Recherche de gène

Eucaryote

Il existe des sites qui combinent plusieurs méthodes pour donner des résultats de prédiction, on en trouvera comme précédemment une liste sur ***infobiogen*** et d'autres serveurs, voir par exemple :

http://bioweb.pasteur.fr/seqanal/genes/genes.html
http://www.bioinformatik.de/cgi-bin/browse/Catalog/Software/Gene_Prediction/
http://www.tigr.org/genefinding/
http://www.cs.ubc.ca/labs/beta/genefinding/

Divisions de Genbank

- **ESTs** (Expressed sequence tags): Principale division ed Genbank. 18 10^6 sequences, 580 organismes différents
- **GSS** (Genome Sequence Survey): résultats de séquençages aléatoire de BAC, dans le cadre de projets Génome. Non assemblées, courtes.
- **HTGS** (High Throughput Genomic Sequences): séquences génomiques en cours d'assemblage. Une fois complètement assemblées, les séquences passent dans les divisions « organisme ».
- Bacteries (BCT), virus (VRL), primates (PRI), rongeurs (ROD) etc: divisions « organismes ».
- **SRA**: Sequence Read Archive (Séquences Ultra-haut débit)
- ~20 divisions en tout.

Nb entrées	Nb. bases	Espèce
1355113	854232260	Homo sapiens
378892	179249409	Mus musculus
76471	139699685	Caenorhabditis elegans
66177	69663817	Arabidopsis thaliana
48963	53428355	Drosophila melanogaster
10571	28658828	Saccharomyces cerevisiae
39568	25816686	Rattus norvegicus
4923	17859484	Escherichia coli
32221	16490243	Fugu rubripes
31480	13072925	Oryza sativa
28406	11746328	Rattus sp.
9540	10912762	Schizosaccharomyces pombe
24125	10712174	Human immunodeficiency virus type 1
1086	9893044	Bacillus subtilis
15370	5794059	Brugia malayi
661	5701954	Mycobacterium tuberculosis
4852	5585160	Gallus gallus
4680	5400457	Plasmodium falciparum
5063	4559072	Bos taurus
10845	4109926	Toxoplasma gondii

Organismes dans Genbank (en 2002)

Enregistrement Genbank

– Chaque enregistrement se voit attribuer un numéro d'accession, stable et unique, et chaque séquence un numéro GI.

– Quand un changement est effectué dans un enregistrement Genbank, le numéro d'accession reste et le GI change.

Autres banques nucléotidiques

- EMBL: Equivalent européen de Genbank. Format différent, contenu presque identique.
- DDBJ: équivalent au Japon
- Banques spécialisées Certaines collections de séquences, bien que généralement présentes dans Genbank, sont beaucoup plus utiles lorsqu'elles sont rassemblées dans des banques spécialisées, par ex:
  - Récepteurs des lymphocytes T (Réarrangements de l'ADN)
  - Génomes HIV, etc.
- Banques pour Blast
  - NR nucléique (« Non-redundant »). All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). (n'est plus "non-redondant")
  - DbEST: dbest Database of GenBank+EMBL+DDBJ sequences from EST Divisions

Banques protéiques

- Swissprot. La mieux annotée des banques protéiques. 2008: 260.000 entrées.
  - Attention: toutes les protéines connues n'y sont pas!
- TrEMBL: banque de protéine produite automatiquement. 2008: 4.200.000 entrées
- Uniprot=Swissprot+TrEMBL
- Dizaines de Banques spécialisées
  - Cazy (Carbohydrate Active Enzymes)
  - Etc.
- Pour Blast
  - NR Protéique (Non-redundent): Banque protéique du NCBI = Traduction de tous les CDS de GenBank + PDB + SwissProt + PIR + PRF - redondances.

De nombreuses sections !

Détection d'erreurs de séquençage potentielles

Prédiction d'introns, exons, sites spécifiques

Prédiction de sites d'épissage

Recherche et localisation des sites de fixation des facteurs de transcription (base TRANSFAC), promoteurs et autres sites spécifiques