De la séquence à l'annotation, et réciproquement

info ou intox ?

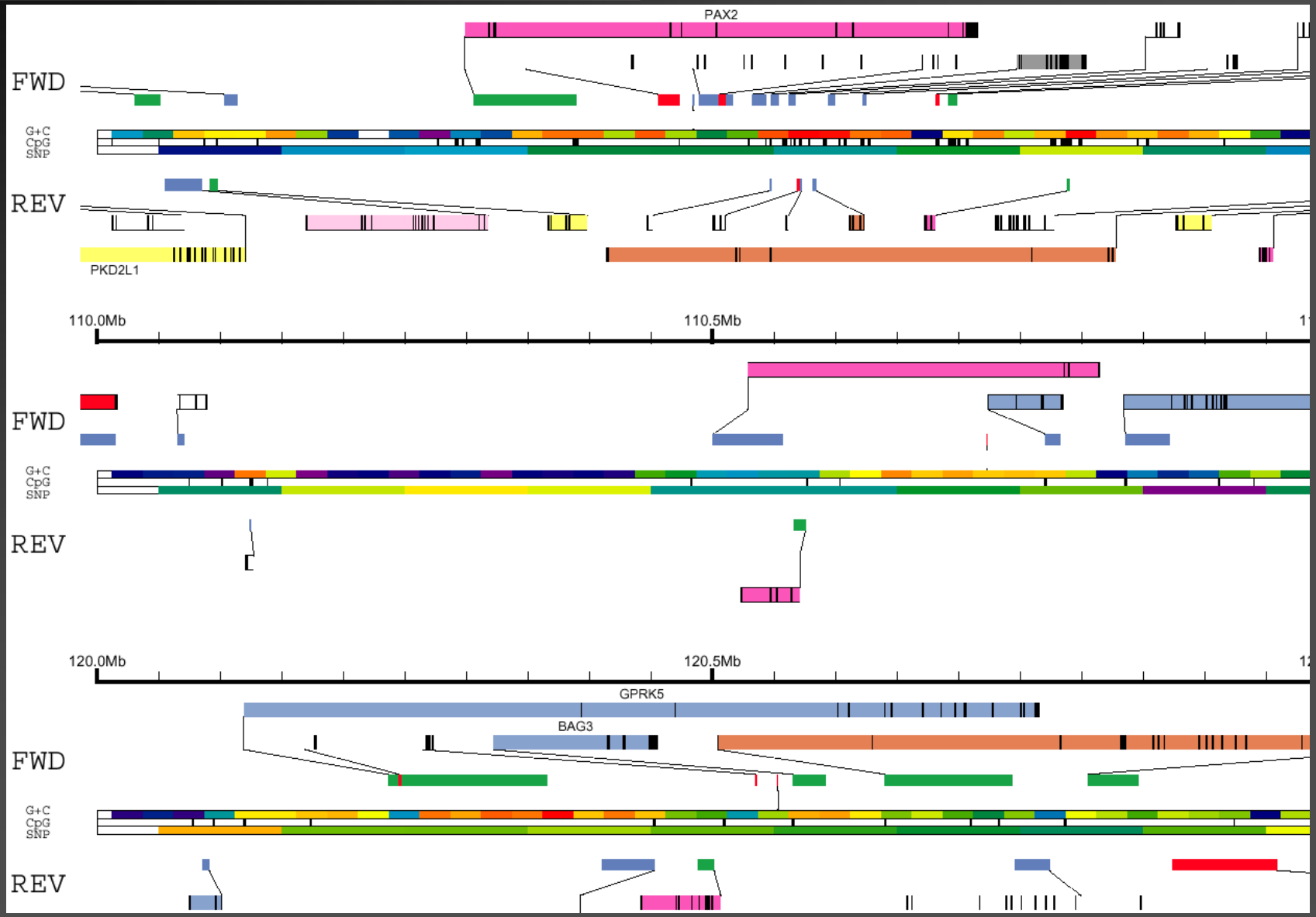
16 Janvier 2003
Journée Bioinformatique à l'IPBS

Catherine MATHÉ

Une séquence génomique

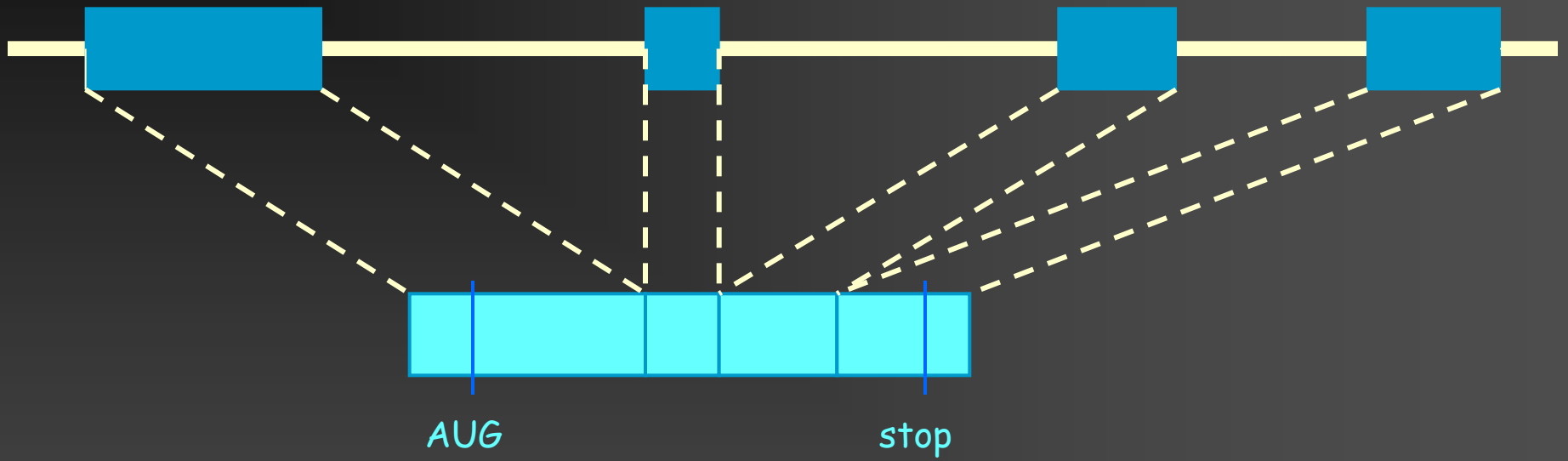
TCCTGGCCTACATGTTCTTTGGCAAAGGATCTTCAAAATCAACGGCTCCCGGTGCGGCGATCATCCATTTCTTCGGAGGGATTCACGAG
ATTTACTTCCCGTACATTCTGATGAAACCTGGCCCTGATTCTCGCAGCCATTGCCGGCGGAGCAAGCGGACTCTTAACATTACGATCTT
TAATGCCGGACTTGTCGCGGCAGCGTCACCGGGAAGCATTATCGCATTGATGGCAATGACGCCAAGAGGAGGCTATTTCGGCGTATTGG
CGGGTGTATTGGTCGCTGCAGCTGTATCGTTCATCGTTTCAGCAGTGATCCTGAAATCCTCTAAAGCTAGTGAAGAAGACCTGGCTGCC
GCAACAGAAAAAATGCAGTCCATGAAGGGGAAGAAAAGCCAAGCAGCAGCTGCTTTAGAGGCGGAACAAGCCAAAGCAGAGAAGCGTCT
GAGCTGTCTCCTGAAAGCGCGAACAAAATTATCTTTTCGTGTGATCCGGGATGGGATCAAGTGCCATGGGGGCATCCATCTTAAGAAAC
AAAGTGAAAAAGCGGAGCTTGACATCAGTGTGACCAACACGGCCATTAACAATCTGCCAAGCGATGCGGATATTGTCATCACCCACAAA
GATTTAACAGACCGCGCGAAAGCAAAGCTGCCGAACGCGACGCACATATCAGTGGATAACTTCTTAAACAGCCCGAAATACGACGAGCT
GATTGAAAAGCTGAAAAGTAATCTTATAGAAAGAGAGTATTGTCATGCAAGTACTCGCAAAGGAAACATTAAACTCAATCAAACGGTAT
CATCAAAAGAAGAGGCTATCAAATTGGCAGGCCAGACGCTGATTGACAACGGCTACGTGACAGAGGATTACATTAGCAAAATGTTTGAC
CGTGAAGAAACGTCTTCTACGTTTATGGGGAATTTCATTGCCATTCCACACGGCACAGAAGAAGCGAAAAGCGAGGTGCTTCACTCAGG
AATTTCAATCATACAGATTCCAGAGGGCGTTGAGTACGGAGAAGGCAACACGGCAAAAGTGGTATTCGGCATTGCGGGTAAAAATAATG
AGCATTTAGACATTTTGTCTAACATCGCCATTATCTGTTCAGAAGAAGAAACATTGAACGCCTGATCTCCGCTAAAGCGAAGAAGATTT
GATCGCCATTTCAACGAGGTGAACTGACATGATCGCCTTACATTTCGGTGCGGGAAATATCGGGAGAGGATTTATCGGCGCGCTGCTTC
ACCACTCCGGCTATGATGTGGTGTTTGCGGATGTGAACGAAACGATGGTCAGCCTCCTCAATGAAAAAAAAGAATACACAGTGGAACTG
GCGGAAGAGGGACGTTCATCGGAGATCATTGGCCCGGTGAGCGCTATTAACAGCGGCAGTCAGACCGAGGAGCTGTACCGGCTGATGAA
TGAGGCGGCGCTCATCACAACAGCTGTCGGCCCGAATGTCCTGAAGCTGATTGCCCCGTCTATCGCAGAAGGTTTAAGACGAAGAAATA
CTGCAAACACACTGAATATCATTGCCTGCGAAAATATGATTGGCGGAAGCAGCTTCTTAAAGAAAGAAATATACAGCCATTTAACGGAA
GCAGAGCAGAAATCCGTCAGTGAAACGTTAGGTTTTCCGAATTCTGCCGTTGACCGGATCGTCCCGATTCAGCATCATGAAGACCCGCT
GAAAGTATCGGTTGAACCATTTTTCGAATGGGTCATTGATGAATCAGGCTTTAAAGGGAAAACACCAGTCATAAACGGCGCACTGTTTG
TTGATGATTTAACGCCGTACATCGAACGGAAGCTGTTTACGGTCAATACCGGACACGCGGTCACAGCGTATGTCGGCTATCAGCGCGGA
CTCAAAACGGTCAAAGAAGCAATTGATCATCCGGAAATCCGCCGTGTTGTTCATTCGGCGCTGCTTGAAACTGGTGACTATCTCGTCAA
ATCGTATGGCTTTAAGCAAACTGAACACGAACAATATATTAAAAATCAGCGGTCGCTTTTAAAATCCTTTCATTTCGGACGATGTGACC
GATTGAAAAGCTGAAAAGTAATCTTATAGAAAGAGAGTATTGTCATGCAAGTACTCGCAAAGGAAACATTAAACTCAATCAAACGGTAT
CATCAAAAGAAGAGGCTATCAAATTGGCAGGCCAGACGCTGATTGACAACGGCTACGTGACAGAGGATTACATTAGCAAAATGTTTGAC
CGTGAAGAAACGTCTTCTACGTTTATGGGGAATTTCATTGCCATTCCACACGGCACAGAAGAAGCGAAAAGCGAGGTGCTTCACTCAGG
AATTTCAATCATACAGATTCCAGAGGGCGTTGAGTACGGAGAAGGCAACACGGCAAAAGTGGTATTCGGCATTGCGGGTAAAAATAATG
AGCATTTAGACATTTTGTCTAACATCGCCATTATCTGTTCAGAAGAAGAAACATTGAACGCCTGATCTCCGCTAAAGCGAAGAAGATTT
GATCGCCATTTCAACGAGGTGAACTGACATGATCGCCTTACATTTCGGTGCGGGAAATATCGGGAGAGGATTTATCGGCGCGCTGCTTC
ACCACTCCGGCTATGATGTGGTGTTTGCGGATGTGAACGAAACGATGGTCAGCCTCCTCAATGAAAAAAAAGAATACACAGTGGAACTG

Un génome « annoté »



Science (2001): 291 (5507): 1304-51

Prédiction par homologie / ADNc



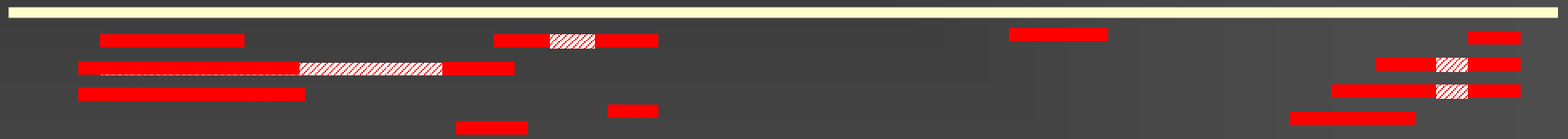
- identification des exons par alignement
- localisation fine des sites d'épissage
  prog. spécifiques : SIM4, Est2genome, Spidey
- Alignement aussi des UTRs
  => identification de la région traduite ?

- ADNc complet ? => structure complète ?

Prédiction par homologie / ESTs

Vraie structure : 3 gènes

Alignements avec ESTs



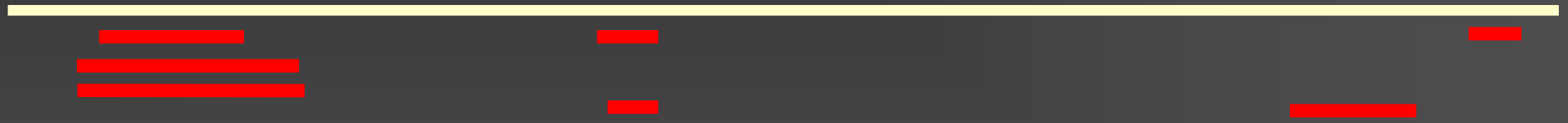
Prédiction

3 gènes =+/- OK

Prédiction par homologie / ESTs

Vraie structure : 3 gènes

Alignements avec ESTs

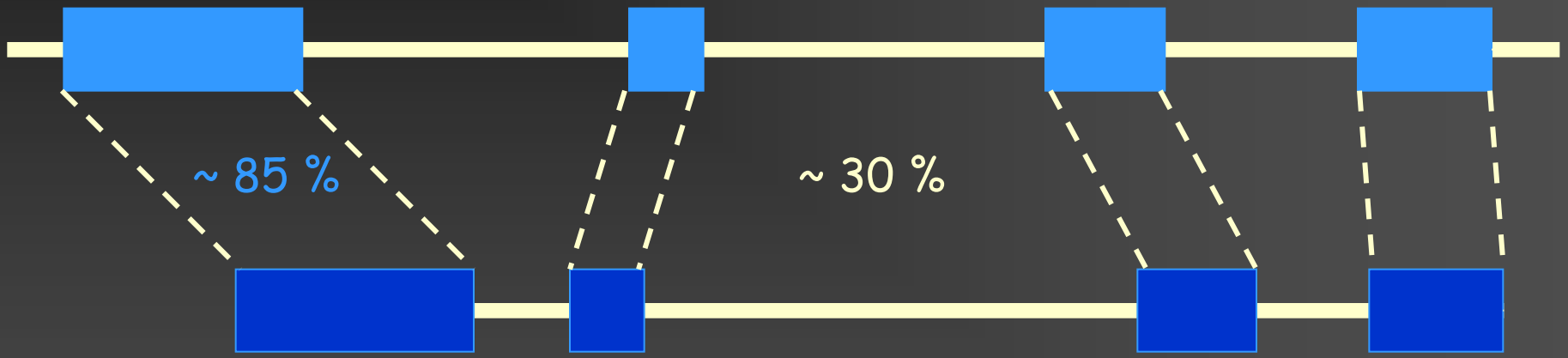


Prédiction

4 gènes, 2 « éclatés », 1 raté

Prédiction par homologie / ADNg

Utilisation des génomes modèles séquencés
Homme <= souris / Fugu / Zebrafish



- UTR et introns peuvent être conservés
- il y a des parties de séquence codantes non conservées
- structure des gènes pas nécessairement conservée

Bilan sur la prédiction par homologie

LA meilleure façon de faire de l'annotation !
Basée sur de « vraies » données biologiques

Identification des familles de gènes
Informations sur épissage alternatif, polyA alternatif

Informations sur conditions d'expression : niveau, lieu
Fonction potentielle

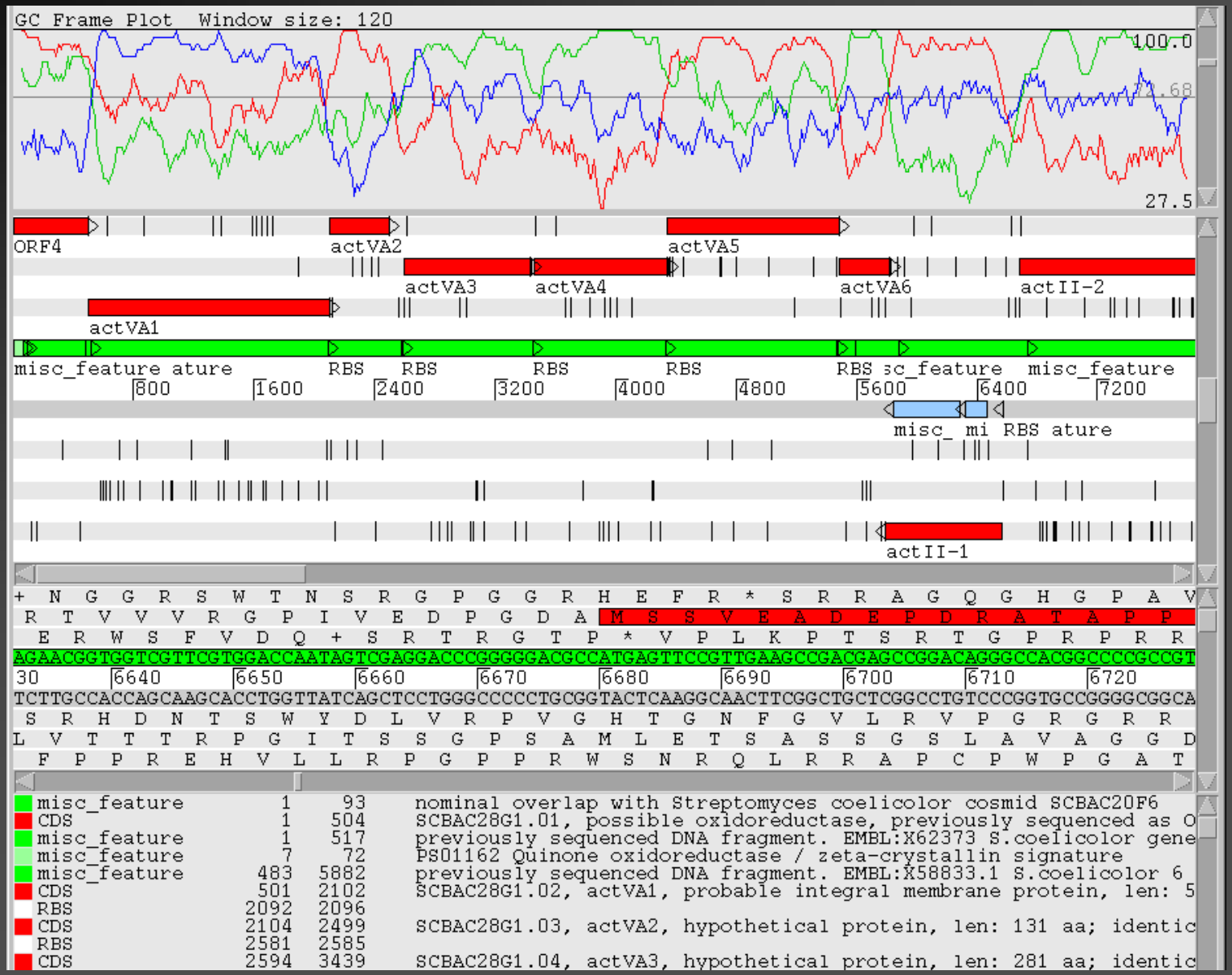
Les limites

Qualité des séquences (ESTs)

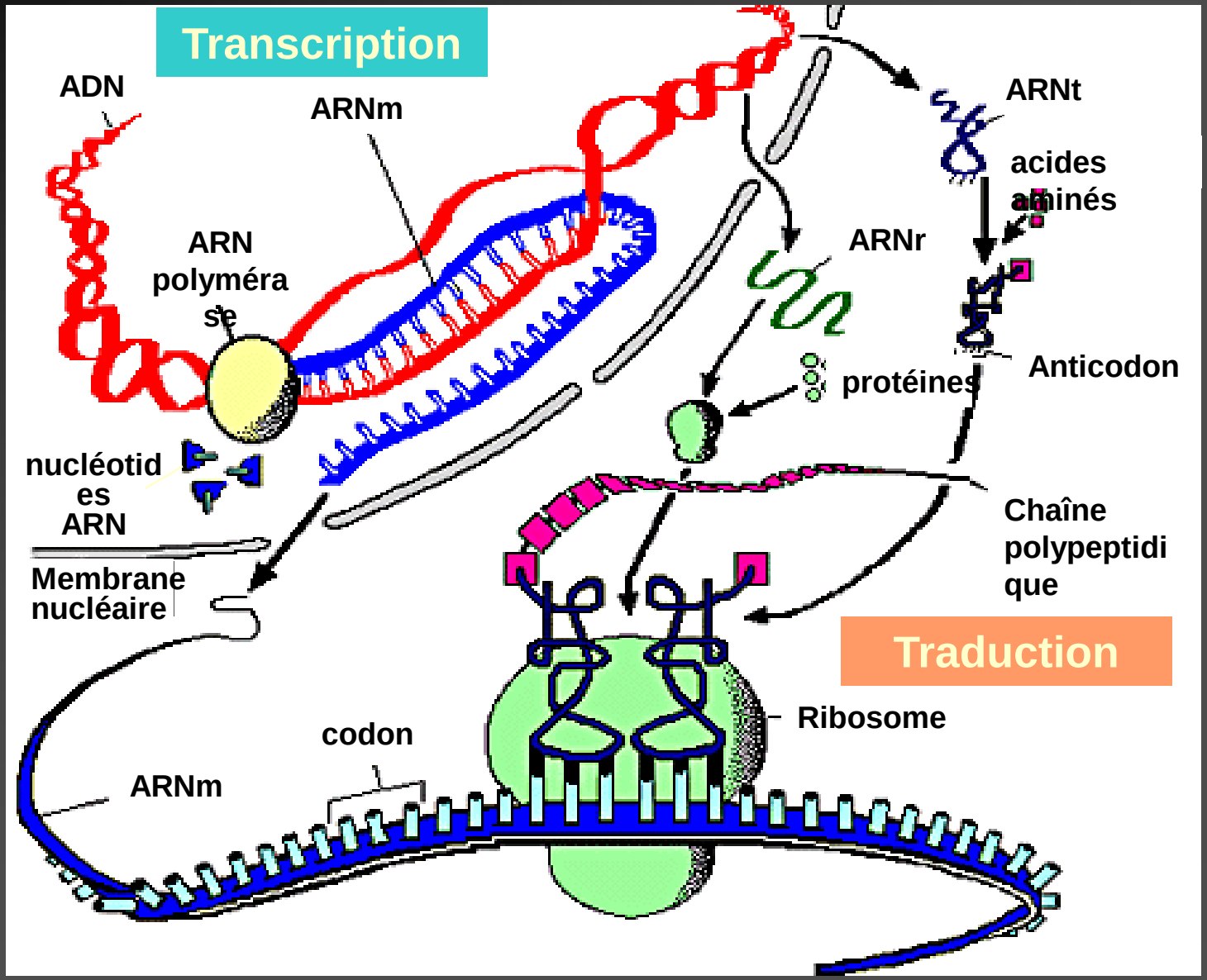
Structure souvent incomplète

Pas toujours de données… 40-50% de gènes sans homologues

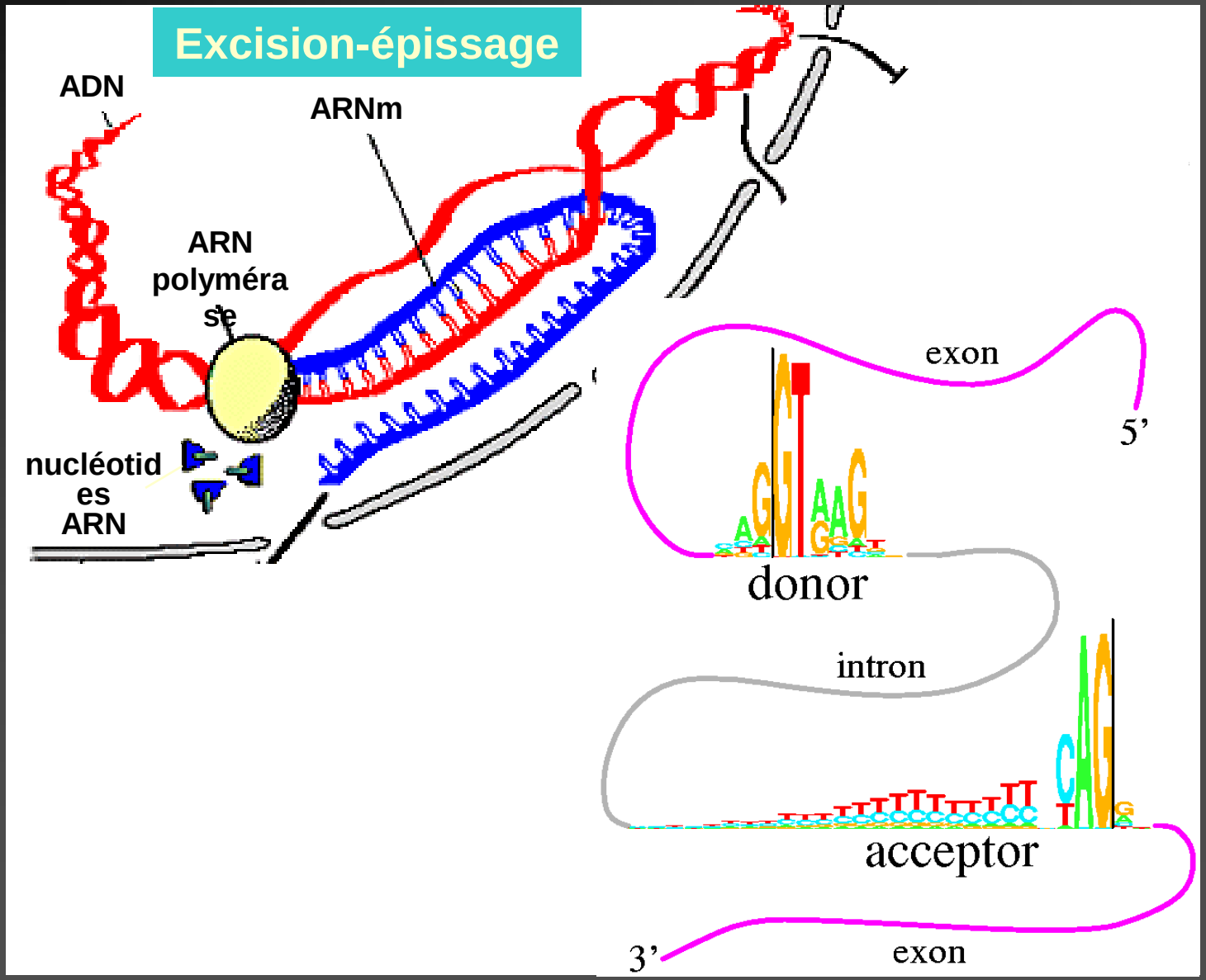
Prédiction *ab initio* ou pure



Les pré-requis pour la prédiction *ab initio*



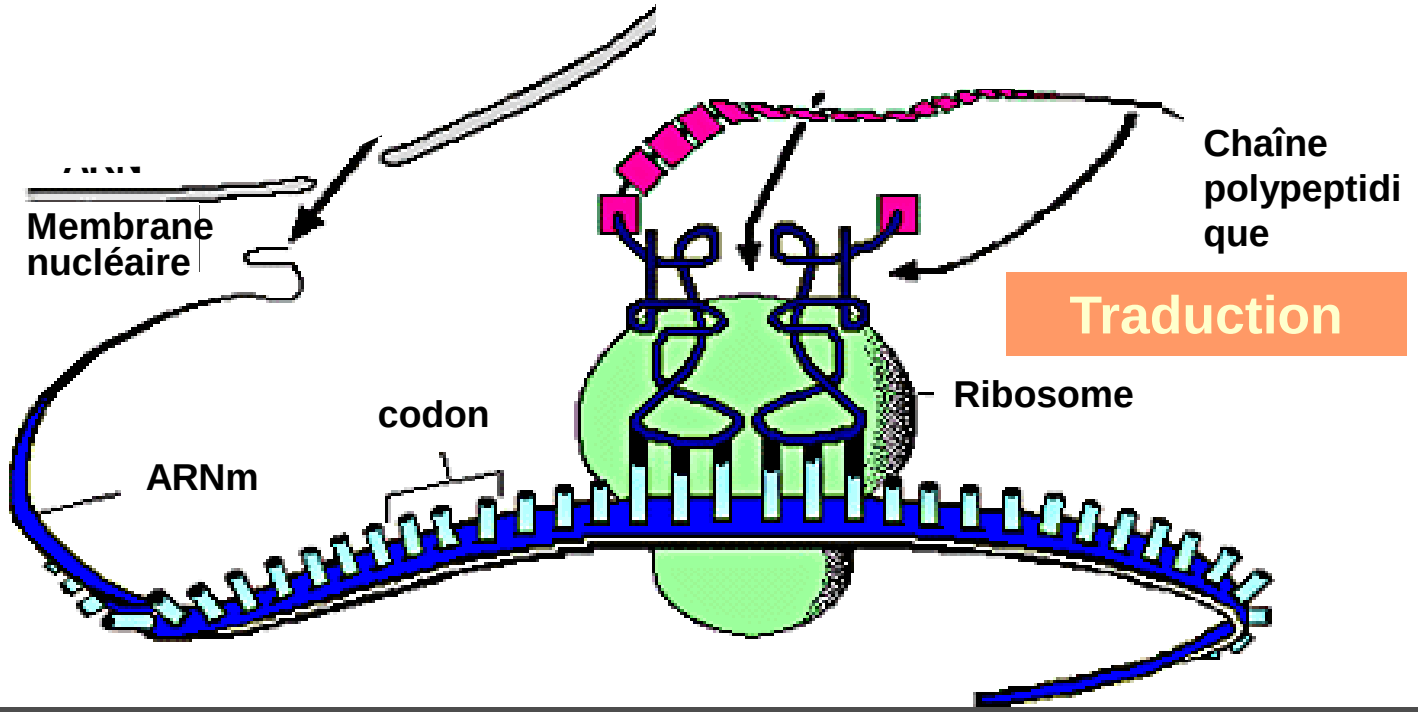
Les pré-requis pour la prédiction *ab initio*



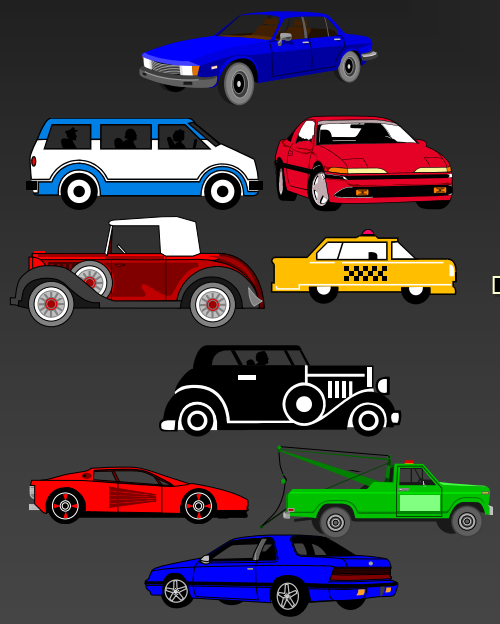
Les pré-requis pour la prédiction *ab initio*

Compositions exons et introns différentes : usage codons, taux en GC, hexamères, …

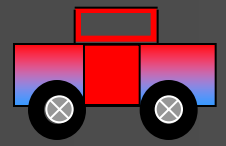
1st base	2nd base U	2nd base C	2nd base A	2nd base G	3rd base
U	Phe	Ser	Tyr	Cys	U
			Stop	Trp	C
					A
					G
C	Leu	Pro	His	Arg	U
			Gln		C
					A
					G
A	Ile	Thr	Asn	Ser	U
	Met		Lys	Arg	C
					A
					G
G	Val	Ala	Asp	Gly	U
			Glu		C
					A
					G



Le principe de la prédiction *ab initio*



Méthodes probablistes (HMM, IMM, ML)
Intelligence artificielle (NN)
Classification (LDA, QDA, SVM)
Matrices de fréq, Consensus



des exemples = jeu d 'apprentissage

méthode d'analyse

un modèle

Mathé *et al.* (2002), Nucleic Acids Research, 30 (19) : 4103-4117

Evaluation de quelques programmes

Niveau exon

	Sensibilité	Spécificité
FGENESH	67	67
GeneMark.hmm	53	54
Genie	71	70
Genscan	70	70
HMMgene	76	77
Morgan	46	41
MZEF	58	59

Séquences mammifères

Sensibilité = proportion du codant bien prédit
Spécificité = proportion de prédictions correctes

Rogic *et al.* (2001) Genome Research, 11: 817-832

Séquences Arabidopsis

Niveau exon

	Sensibilité	Spécificité
FGENESH	88	88
GeneMark.hmm	83	78
Genscan	63	69
GlimmerA	67	67
EuGene	87	88
EuGene+	90	90
MZEF	45	47

Niveau gène

	Sensibilité	Spécificité
FGENESH	57	55
GeneMark.hmm	41	37
Genscan	17	19
GlimmerA	30	19
EuGene	65	60
EuGene+	78	70

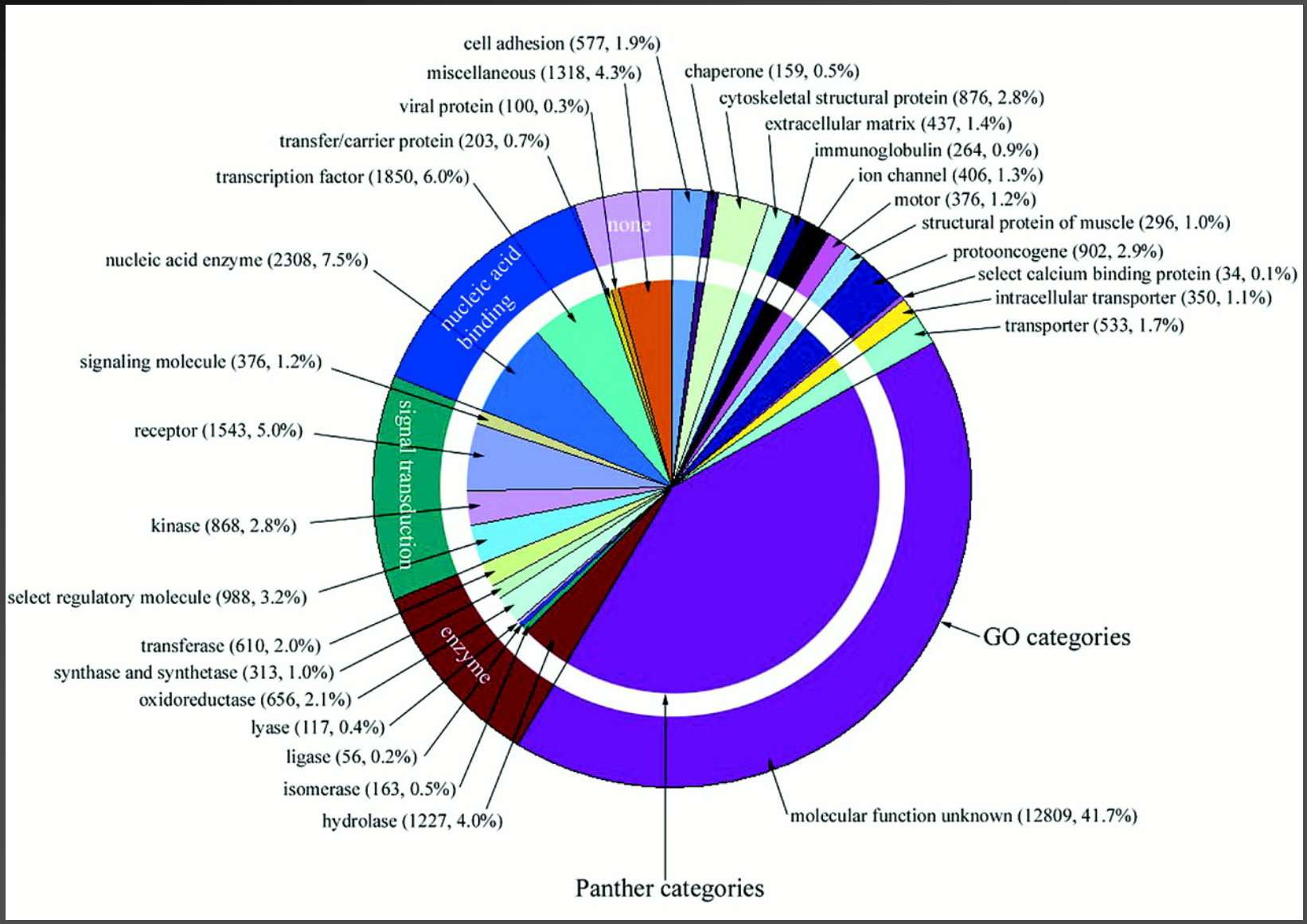
Schiex *et al.*, en préparation

Les limites des méthodes actuelles

- Limite entre 2 gènes

- Ce ne sont pas des programmes génériques, c'est à dire pas adaptés à tous les organismes
- Cas limites : grands gènes (2,4 Mb Dystrophine)
  grands introns (100 kb)
  petits exons (3 b)
- on ne (re)trouve que ce qu'on connaît...
- Situations biologiques complexes :
  - site d'épissage non canoniques
  - épissage alternatif, polyA alternatif
  - gènes polycystroniques
  - gènes dans des gènes

Annotation fonctionnelle



Science (2001): 291 (5507): 1304-51

Annotation fonctionnelle

```
MZEORFG: 1    ILNSPDRACNLAKQAFDEAISELDSLGEESYKDSTLIMQLLXDNLTLWTSDTNEDGGDE 59
               I N+P++AC LAKQAFD+AI+ELD+L E+SYKDSTLIMQLL DNLTLWTSD  ++   E
BOV1433P: 186 IQNAPEQACLLAKQAFDDAIAELDTLNEDSYKDSTLIMQLLRDNLTLWTSDQQDEEAGE 244

 Score = 87.4 bits (213), Expect = 1e-17
 Identities = 41/59 (69%), Positives = 50/59 (84%)

LOCUS       BOV1433P     1696 bp    mRNA    linear       MAM       26-APR-1993
DEFINITION  Bovine brain-specific 14-3-3 protein eta chain mRNA, complete cds.
ACCESSION   J03868


LOCUS       MZEORFG       187 bp    mRNA    linear       PLN       01-FEB-2001
DEFINITION  Zea mays putative brain specific 14-3-3 protein, tau protein
            homolog  mRNA, partial cds.
```

De l'annotation à la séquence

```
LOCUS       AP000399               154180 bp    DNA     linear   PLN 21-MAR-2002
DEFINITION  Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 6,
            clone:P0535G04.
ACCESSION   AP000399
VERSION     AP000399.1  GI:5803242
KEYWORDS    .
SOURCE      Oryza sativa (japonica cultivar-group) (cultivar:Nipponbare) DNA,
            clone:P0535G04.
  ORGANISM  Oryza sativa (japonica cultivar-group)
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae;
            Ehrhartoideae; Oryzeae; Oryza.
REFERENCE   1  (bases 1 to 154180)
  AUTHORS   Sasaki,T., Matsumoto,T. and Yamamoto,K.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-AUG-1999) Takuji Sasaki, National Institute of
            Agrobiological Resources, Rice Genome Research Program; Kannondai
            2-1-2, Tsukuba, Ibaraki 305-8602, Japan
            (E-mail:tsasaki@abr.affrc.go.jp, URL:http://rgp.dna.affrc.go.jp/,
            Tel:81-298-38-7441, Fax:81-298-38-7468)
COMMENT     The orientation of the sequence is from T7 to SP6 of the PAC clone.
            Genes were predicted from the integrated results of the
            following:GENSCAN1.0, BLASTN2.0, BLASTX2.0 as well as
            SplicePredictor (October1998 version). The genomic sequence was
            searched against the non-redundant database NRP (PIR,SWISSPROT,
            GENPEPT, PDB) from MAFF DNAbank and the cDNA sequence database at
            RGP. Protein similarities of the coding regions were searched
            against NRP with BLASTP2.0. ESTs represent the identified cDNA
            sequences using BLASTN1.4 with the corresponding DDBJ accession no.
            and RGP clone ID.
```

De l'annotation à la séquence

```
     CDS             complement(10650..10856)
                     /note="ESTs AU078033(RO160),D23790(RO160) correspond to a
                     region of the predicted gene.
                     hypothetical protein"
                     /codon_start=1
                     /protein_id="BAA83553.1"
                     /db_xref="GI:5803243"
                     /translation="MLILEVGIWLLPFTLLVAPVRRMVRLVQELQRIMLVVACDRSRR
                     GGRPPTFGEVLSRLDRLDSATVIV"
     CDS             complement(join(13526..14940,15018..15165))
                     /note="Similar to hexose carrier protein HEX6 &RCCHCP_1
                     (Q07423)"
                     /codon_start=1
                     /protein_id="BAA83554.1"
                     /db_xref="GI:5803244"
                     /translation="MAVGVVAGVESQERRGGGAGTGRVTAFVVLSCVTAGMGGVIFGY
                     DIGIAGGVSSMEPFLRKFFPEVHRRMEGDVRVSNYCKFDSQLLTAFTSSLYVAGLLTT
                     FAASRVTAGRGRRPSMLLGGAAFLAGAAVGGASVDIYMVILGRVLLGVGLGFANQAVP
                     LYLSEMAPSRWRGAFSNGFQLSVGVGALAANVINYGTEKIRGGWGWRVSLALAAVPAG
                     LLTLGALFLPETPNSLIQQGKVERCDVEQLLKKIRGADDVADELDTIVAANSATAGVG
                     GGGLLMLLTQRRYRPQLAMAVMIPFFQQVTGINAIAFYAPVLLRTIGMGESASLLSAV
                     VTGVVGVGATLLSMFAVDRFGRRTLFLAGGAQMLASQVLIGGIMAAKLGDDGGVSRAW
                     AAALILLIAAYVAGFGWSWGPLGWLVPSEVFPLEVRSAGQSVTVATSFVFTVFVAQAF
                     LAMLCRMRAGIFFFFAAWLAAMTAFVYLLLPETKGVPIEEVAGVWRGHWFWSRVVGGD
                     GEEEERNNGGKL"
     CDS             complement(join(16352..16617,17437..17488,18513..19025))
                     /note="hypothetical protein"
                     /codon_start=1
                     /protein_id="BAA83555.1"
                     /db_xref="GI:5803245"
                     /translation="MSIHLRAHAFAASPLRGLSASTAAVSPSAAADALRSLLDAGAGA
                     ADAAHPHPHPHLSKILPFRRGRPLARSYDSPPPPAAAAAAPPPPPAWRLAWLPPARVP
                     DVPSDAFVFLGAHGEEEGKEAAAYWAVDVSERDGEGAGDGSAFVDLRTLMVATDWRDK
                     DAMGDLAIAGHPGESLEEAVRRETWEETDAQWHSREDVKKALTFAEYEKAQRTNALKV
                     NQICKGVEKRQSISADLKIESEEPAPMFVPGPYAIAHHLISSWAFEGAPKAPSSFSNL
```

Finalement : info ou intox … ?

Info !!
Evaluation du nombre de gènes des génomes
Des erreurs sur le « détail » : plus des 3/4 exons OK

En attendant mieux…

Facilite et accélère l'étude expérimentale des gènes
⇒ Design amorces PCR
⇒ Sondes pour puces à ADN

Mais … il faut garder un regard critique…
Lire l'origine de l'annotation
Banque TrEMBL ou GenPept