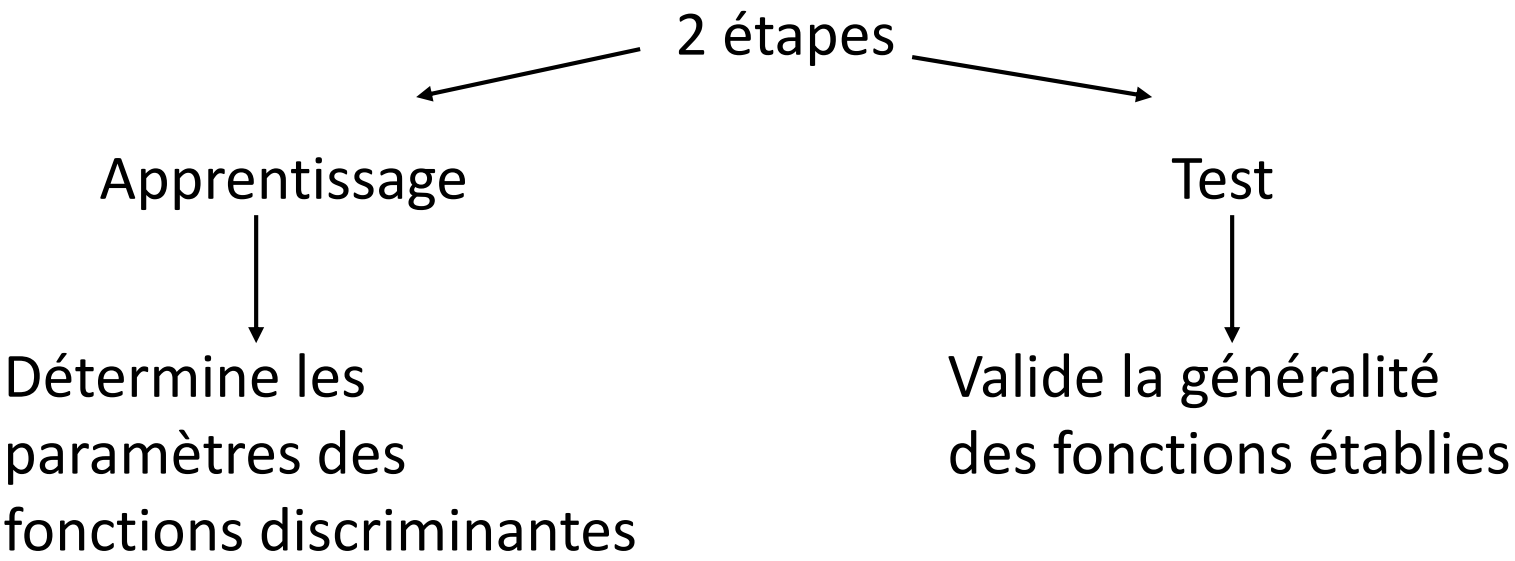
Support de cours
Annotation des génomes
(Partie I)

Méthodes de prédiction: démarche générale

- Définir clairement l'objectif.
- Choisir les critères.
- Choisir le type d'approche :

  - sans système de référence,

  - avec système de référence.



Mesure du pouvoir prédictif d'une méthode

4 paramètres importants :

- pourcentage de vrais positifs (VP, True positive)
- pourcentage de faux positifs (FP, False positive)
- pourcentage de vrais négatifs (VN, True negative)
- pourcentage de faux négatifs (FN, False negative)

prédiction \ Réalité	Groupe 1	Groupe 2
Groupe 1	% vrais positifs	% faux positifs
Groupe 2	% faux négatifs	% vrais négatifs

Groupe 1 : exemples
Groupe 2 : contre-exemples

Mesure du pouvoir prédictif d'une méthode

Idéal: prédire le maximum d'exemples (max VP) avec un minimum d'erreurs (min FP). Mais valeurs non indépendantes donc impossible.

→ Solution un compromis :

- on maximise le % de VP (donc minimise le % de FN) souvent par utilisation de critères moins stricts même si cela entraîne l'augmentation du % de FP. L'élimination des FP se fait par un autre traitement informatique ultérieur. On dit que l'on privilégie la **sensiblilité** de la méthode

- inversement, on minimise le % de FP même si cela conduit à ne pas détecter certaines séquences d'intérêts (donc plus grand % de FN). On dit que l'on privilégie la **spécificité** de la méthode.

Sensibilité = VP/(VP+FN) sensibility an anglais
Spécificité = VP/(VP+FP) specificity en anglais (ou VN/VN+FP 2 définitions)
précision = (VP+VN)/(VP+VN+FP+FN) accuracy en anglais

Annotation d'un génome

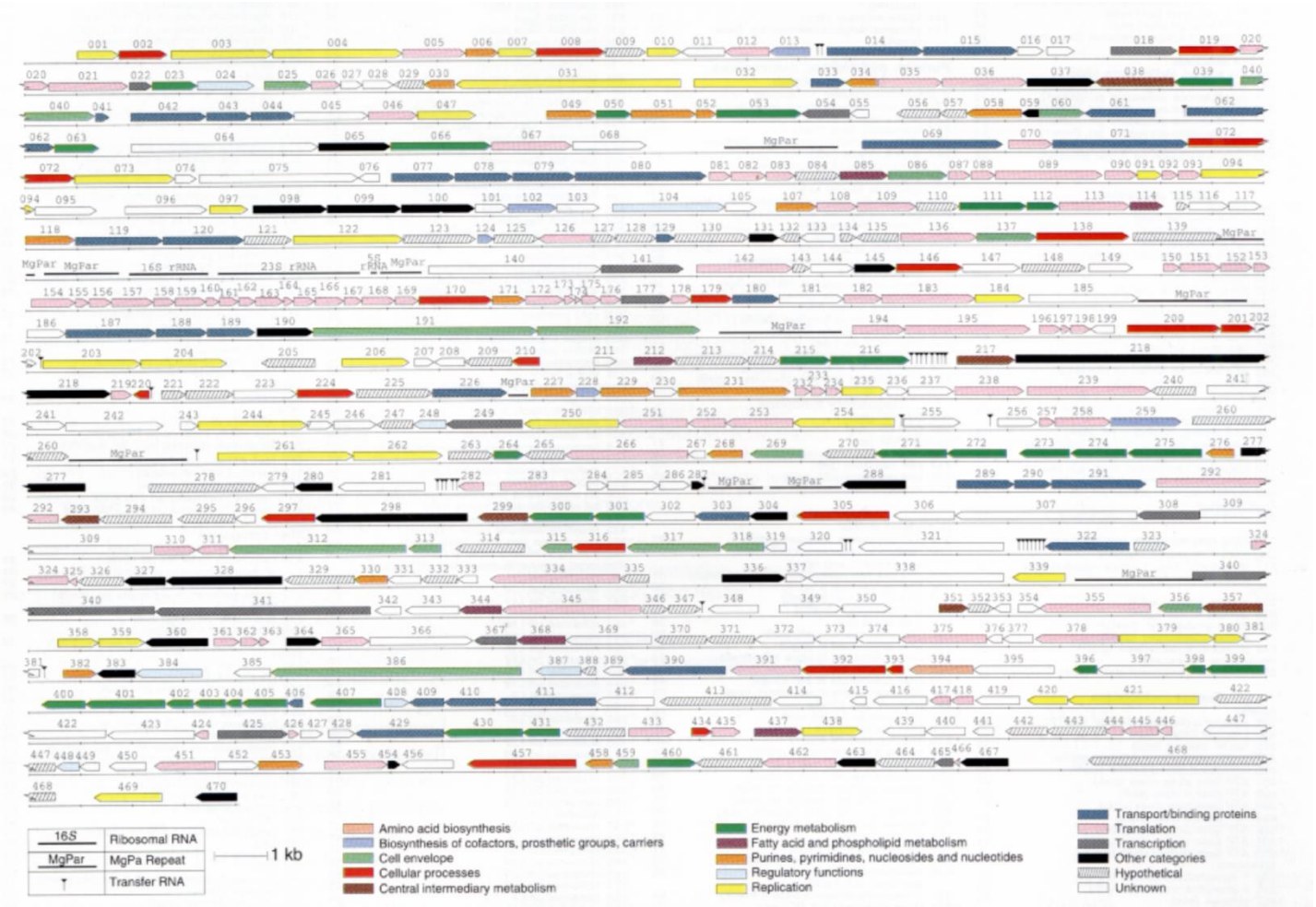
Identification des gènes codant pour :

- les ARNr
- les ARNt
- les protéines

Identification des unités de transcription (promoteur et terminateur)

Pour les gènes codant pour les protéines, prédiction fonctionnelle par recherche de similarité de séquences (Blast) et classification en grandes classes fonctionnelles (ex: biosynthèse des acides aminés, métabolisme énergétique….)

Exemple d'annotation d'un génome
Mycoplasma genitalium

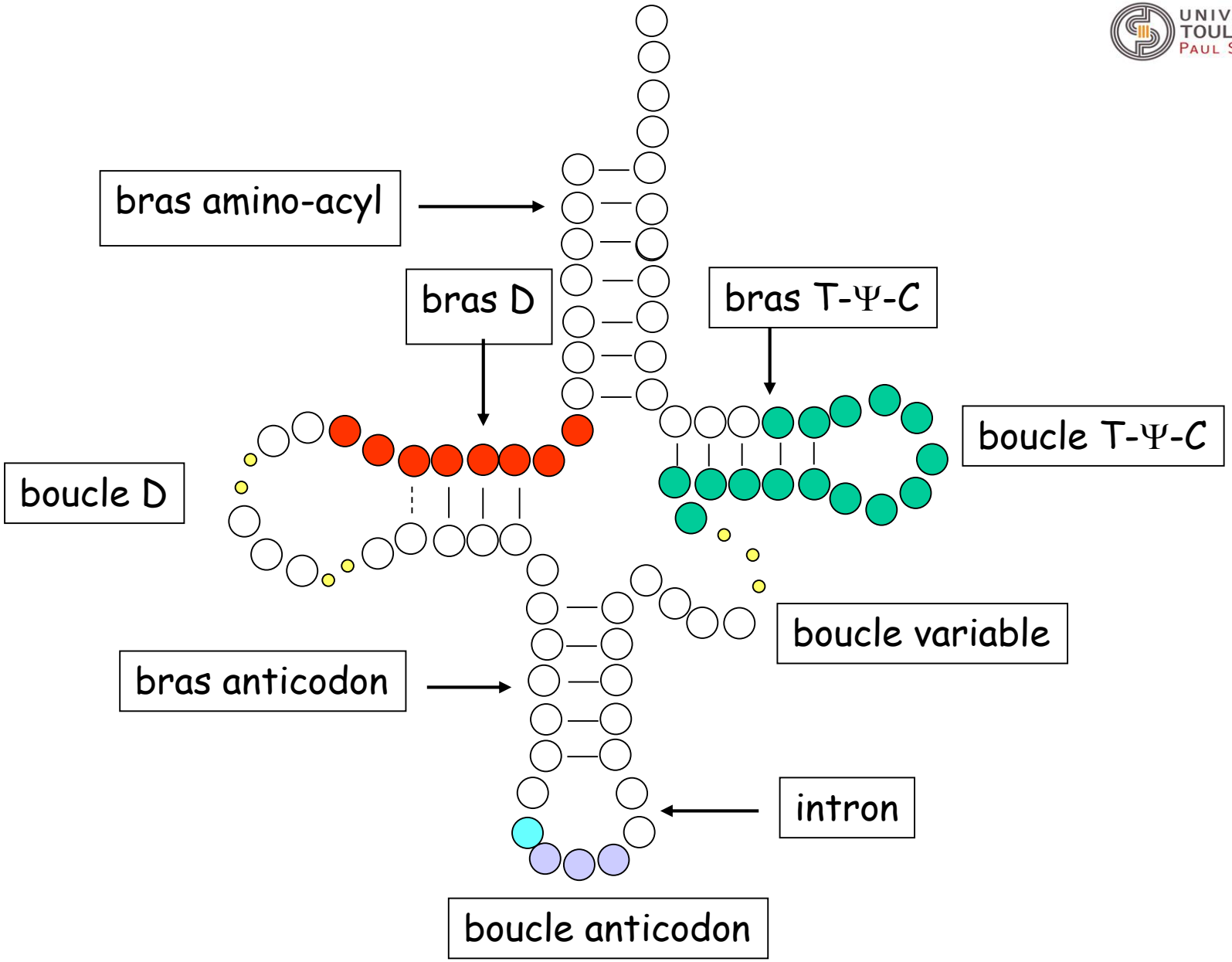


Identification des gènes nucléaires codant pour des ARNt (tRNAscan,) (Fichant and Burks, J. Mol. Biol. (1991) 220, 659-671)

Objectif: Identifier automatiquement les gènes nucléaires codant pour les ARNt dans les longs fragments génomiques.

Méthode avec système de référence
Méthode intégrée combinant des critères de différents types par utilisation de règles et de filtres.

Structure secondaire canonique d'une séquence d'ARNt nucléaire



Critères utilisés et appris sur un ensemble de gènes d'ARNt connus:

- 2 motifs de type signal correspondant aux régions conservées T-Ψ-C et D représentés par des matrices consensus des fréquences des bases

- 4 motifs structuraux correspondant aux 4 bras de la structure en feuille de trèfle.

Ensemble d'apprentissage : 242 séquences issus de la division « structural RNA » de GenBank (release 61)comprenant des séquences d'ARNt d'organismes différents (primates, autres vertébrés, eucaryotes inférieurs, bactéries.
Les séquences homologues ont été réduites à un seul représentant.
Seuls les gènes d'ARNt nucléaires ont été retenus car les séquences d'ARNt mitochondriaux et chloroplastiques ne partagent pas toutes les caractéristiques communes des séquences d'ARNt nucléaires et des long introns jusqu'à 800 pb peuvent être présent dans certains gènes d'ARNt chloroplastiques.

Ensembles test : autres divisions taxonomiques de GenBank (release 66)

Motifs de type signal

- 2 matrices consensus (PWM : position Weight Matrices) établies à partir d'un ensemble de référence de 242 séquences d'ARNt nucléaires

Recherche de ces motifs dans une nouvelle séquence :

- Parcours de la séquence avec une fenêtre de la taille du motif et un pas de 1.
- Filtre sur le nombre de bases invariantes
- Calcul d'un score utilisant ces matrices pour retenir la fenêtre comme motif potentiel

$$S = \frac{\sum_{i=1}^{i=n} f_{x_i,i} - N_I}{\sum_{i=1}^{i=n} f_{\max,i} - N_T}$$

n : longueur du motif
$f_{x_i,i}$: fréquence de la base trouvée à la position i dans la fenêtre
$f_{max,i}$: fréquence de la base la plus fréquente à la position i dans la matrice
N_I et N_T : nombre de bases invariantes, respectivement, dans la fenêtre et dans la matrice.

S varie entre 0 et 1
En soustrayant le nombre de bases invariantes (N_I et N_T) -> poids de 0 à ces positions car présentes dans chaque motif. Lle score S est comparé à un seuil pour retenir ou non la fenêtre comme un motif potentiel

Table 4A
T-Ψ-C signal consensus matrix

Position	Base A	C	G	T
48	0·012	0·810	0·008	0·170
49	0·162	0·332	0·477	0·030
50	0·099	0·401	0·293	0·210
51	0·182	0·182	0·510	0·130
52	0·161	0·029	0·777	0·033
53	0·000	0·000	**1·000**	0·000
54	0·058	0·000	0·000	0·942
55	0·000	0·000	0·000	**1·000**
56	0·000	**1·000**	0·000	0·000
57	0·182	0·000	0·818	0·000
58	0·988	0·000	0·008	0·004
59	0·496	0·062	0·169	0·273
60	0·062	0·186	0·008	0·744
61	0·000	**1·000**	0·000	0·000
62	0·029	0·769	0·033	0·169

Consensus matrix corresponding to the T-Ψ-C signal. The base frequencies at each position of the signal are given, based on 242 nuclear tRNA sequences extracted from the "structural RNA" file of GenBank (Rel. 61). Invariant positions (as defined in Appendix, section (b)) are indicated in bold. The position number is given by following the standard system for numbering tRNA sequences (see Fig. 1).

Table 4B
D signal consensus matrix

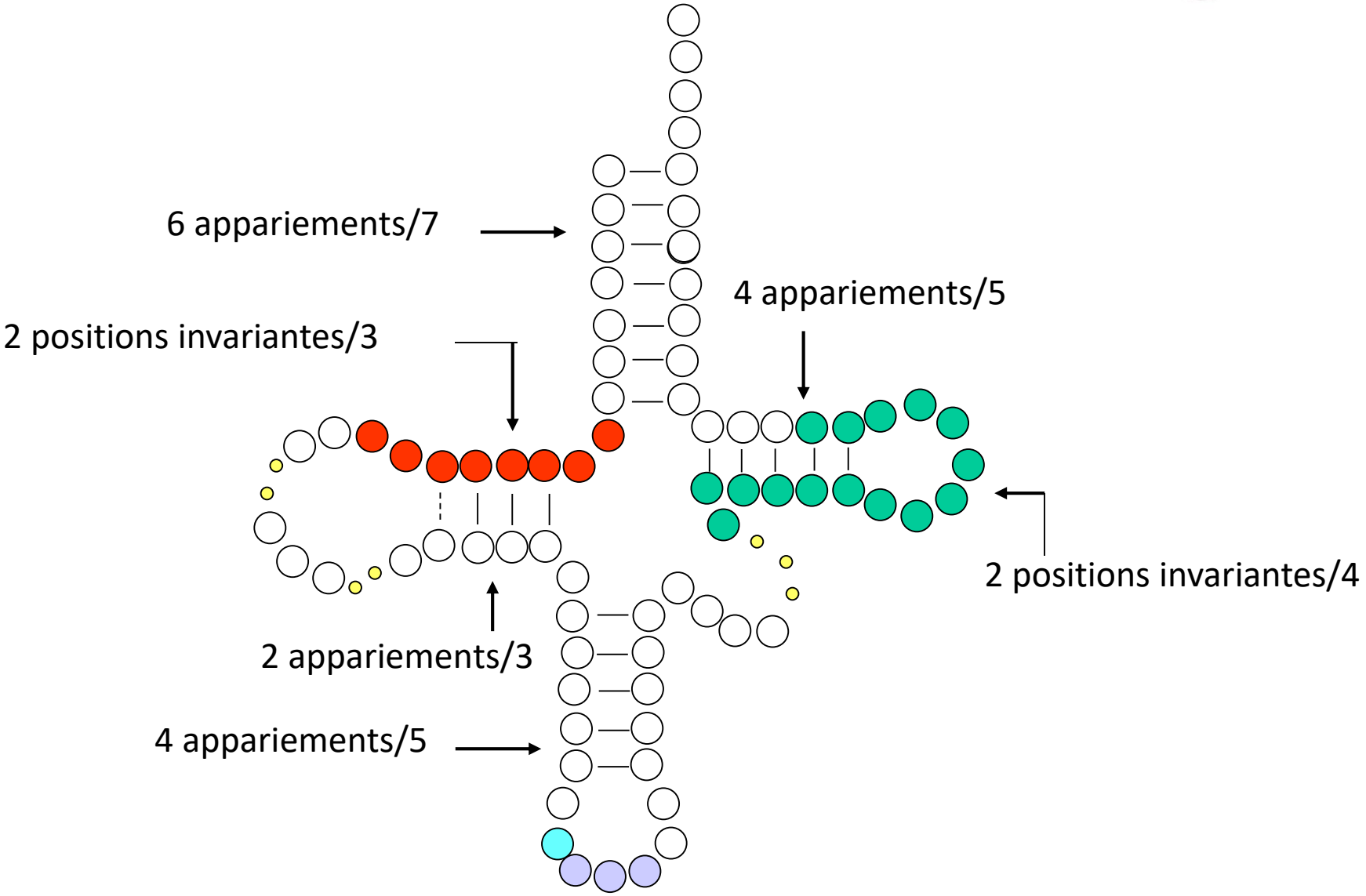
Position	Base A	C	G	T
8	0·000	0·000	0·000	**1·000**
9	0·510	0·040	0·420	0·020
10	0·000	0·000	**1·000**	0·000
11	0·040	0·640	0·050	0·280
12	0·050	0·250	0·300	0·400
13	0·090	0·490	0·130	0·290
14	**1·000**	0·000	0·000	0·000
15	0·180	0·010	0·750	0·060

Consensus matrix corresponding to the D signal. See Table 4A for details.

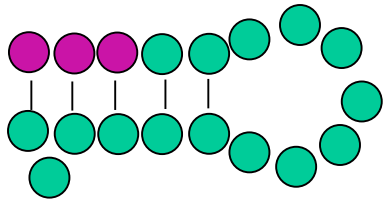
Extrait de J. Mol. Biol. (1991) 220, 659-671

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Définition des seuils : limites inférieures



Recherche de la région T-Ψ-C: Etapes 1 et 2 de l'algorithme



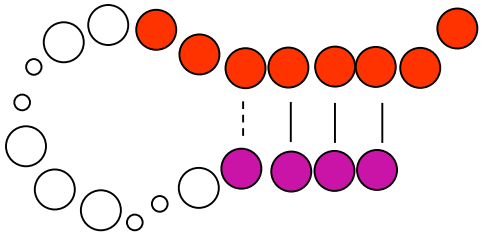
Motif T-Ψ-C

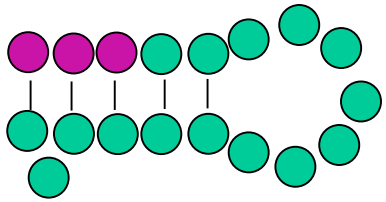
bras T-Ψ-C

Recherche de la région D: Etapes 3 et 4 de l'algorithme

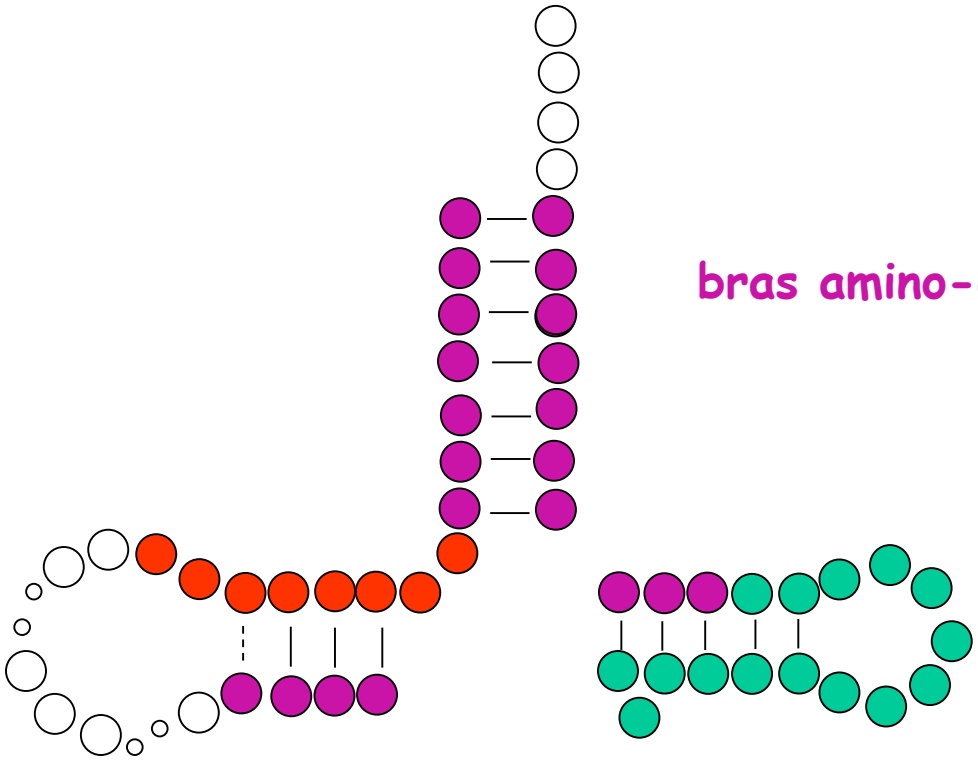
Motif D

région D

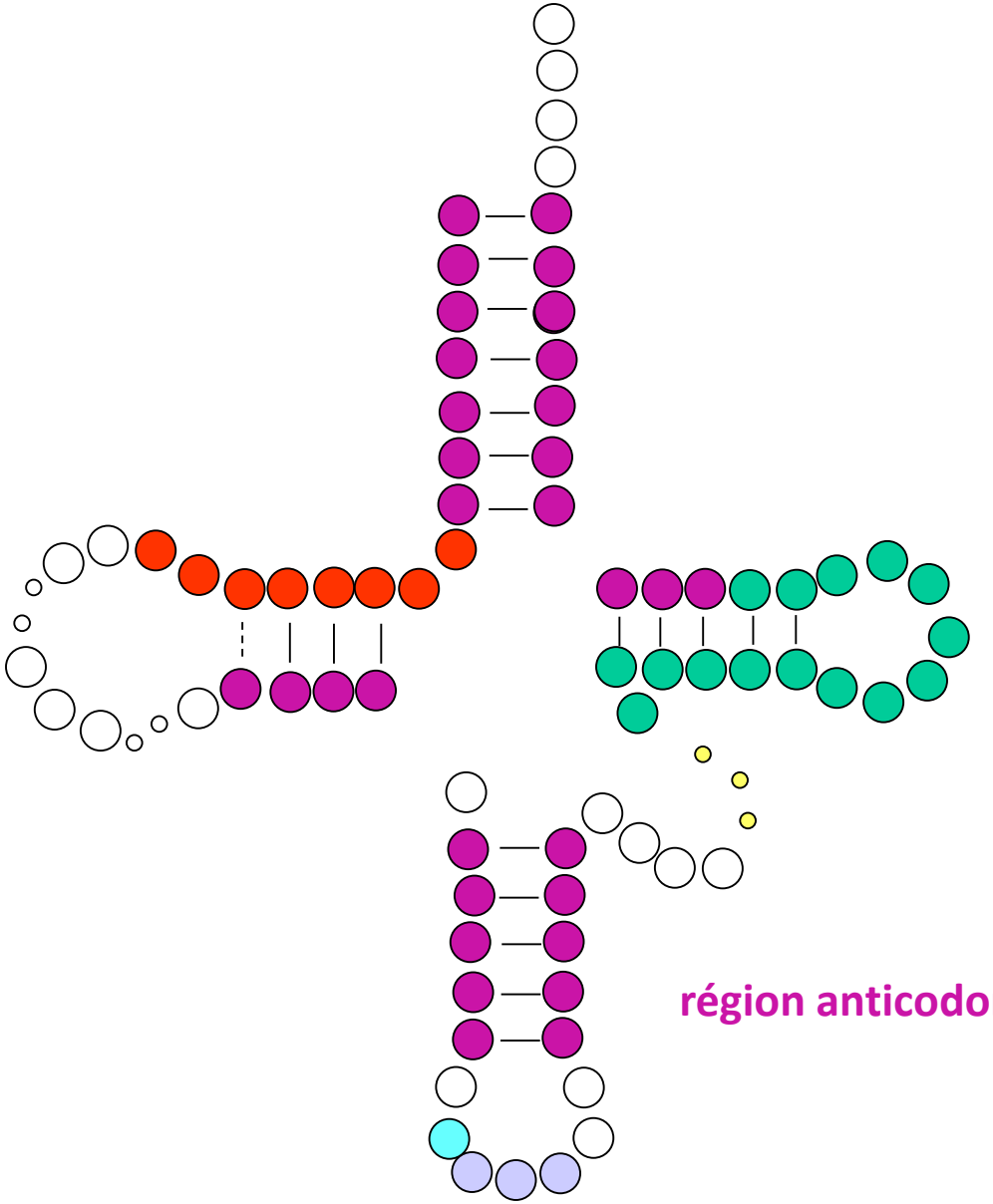




Recherche de la région amino-acyl : Etape 5 de l'algorithme



Recherche de la région anticodon: Etape 6 de l'algorithme



Mise en place d'un filtre : limiter de nombre de FP

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Table 5
Pattern filter thresholds

Pattern	Perfect match†	Threshold match‡
T-Ψ-C signal	3 to 4 invariant bases $SG = SG + 1$	2 invariant bases $SG = SG + 0$
D signal	3 invariant bases $SG = SG + 1$	2 invariant bases $SG = SG + 0$
5′ of the anti-codon signal	Base T $SG = SG + 1$	Other base $SG = SG + 0$
T-Ψ-C arm	5 base-pairings $SG = SG + 1$	4 base-pairings $SG = SG + 0$
D arm	3 base-pairings $SG = SG + 1$	2 base-pairings $SG = SG + 0$
Aminoacyl arm	7 base-pairings $SG = SG + 1$	6 base-pairings $SG = SG + 0$
Anticodon arm§	5 base-pairings $SG = SG + 0$	4 base-pairings $SG = SG + 0$

† Pattern filters defined by tRNA training set. The incremental evaluation of the general score, SG, for these patterns is shown (see Appendix, section (e), for a discussion of SG). SG is evaluated sequentially for each pattern filter: initially, $SG = 0$, and $SG_{max} = 6$. Sequences showing a $SG_{final} < 5$ were rejected as potential tRNA genes.

‡ Threshold for individual pattern filters used to filter candidate sequences from the test set. The incremental evaluation of SG for the threshold conditions is given.

§ The anticodon arm was not included in the SG discrimination (for explanations, see Appendix, section (e)).

Calcul d'un score général, SG, qui permet d'appliquer un autre filtre à seuil global. Pour chaque motif étudié, SG est incrémenté de 1 si la séquence de la fenêtre présente le motif "idéal" (Table 5).

Algorithme de tRNAscan

- 5ème étape SG ≥ 4
- Recherche du bras anticodon (4-5 nt)
- Si base T en 5' de l'anticodon : SG =SG+1
- bras anticodon pas pris en compte dans le calcul de SG, car augmente le nombre de FP

Taille maximum intron : 60 pb

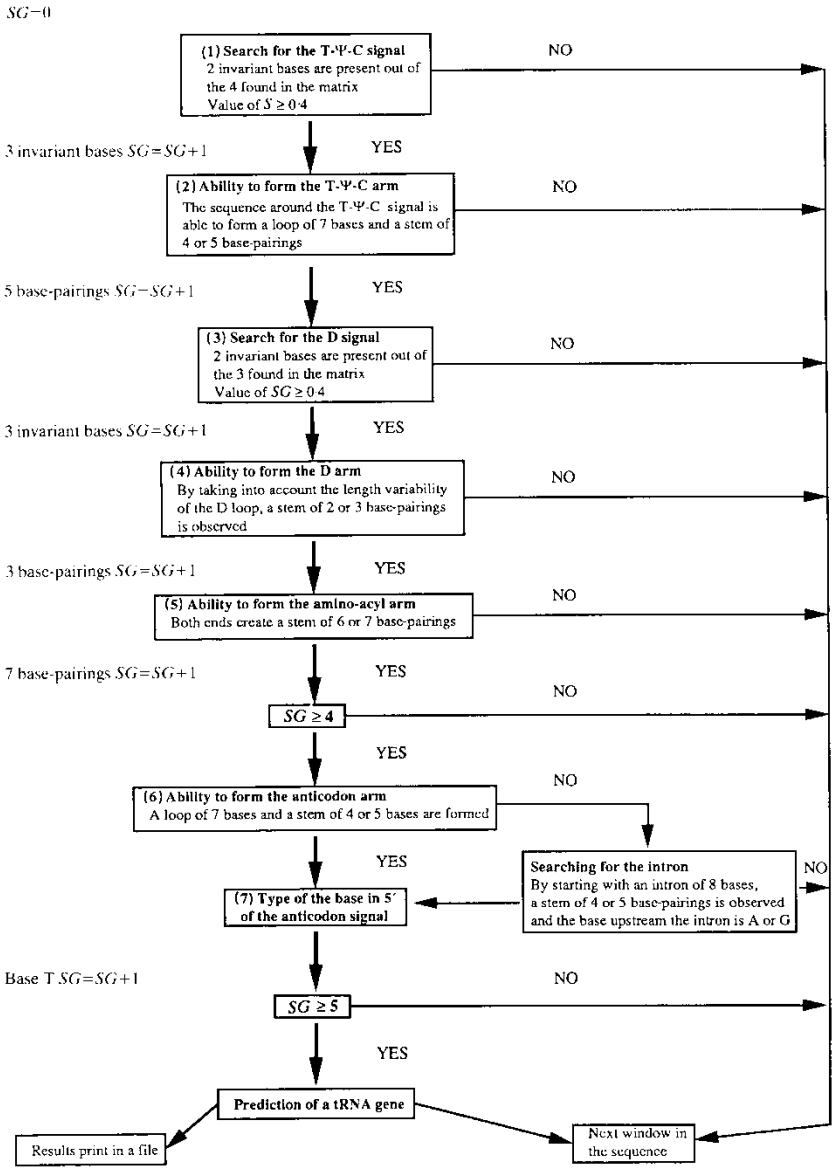


Figure 2. Algorithm description. Each step of the algorithm, with corresponding threshold values, is in a separate box. If the score for the windowed sequence does not exceed the threshold at any step, the algorithm is initiated again on the next window of the sequence under study. The incrementation of *SG* is illustrated (see also Table 5 in Appendix).

Extrait de J. Mol. Biol. (1991) 220, 659-671

Performance de tRNAscan

Table 1
Database search summary

Taxonomic category	Number of nucleotides tested†	Number of tested nucleotides corresponding to tRNA genes†	% false-negative‡	% false-positive§, ‖
Primates	9,003,383 7511 sequences	3127	2·4 1 gene out of 41	0·0025 6 genes, including 2 pseudogenes
Rodents	7,841,099 7652 sequences	3652	0·0 0 genes out of 50	0·0024 5 genes
Vertebrates	2,142,926 1876 sequences	2061	13·0 3 genes out of 23	0·0 0 genes
Inverebrates	4,005,462 3195 sequences	10,748	2·1 3 genes out of 141	0·0009 (0·0009) 1 gene (1 gene)
Plants	4,659,180 2976 sequences	14,593	2·3 4 genes out of 177	0·01 (0·007) 11 genes (8 genes)
Bacteria	6.992,664 4293 sequences	23,520	2·6 8 genes out of 312	0·01 (0·003) 19 genes (6 genes)
Total	34,644,714 27,503 sequences	57,701	2·5 19 genes out of 744	0·005 (0·003) 42 genes (26 genes)

† The number of nucleotides corresponds to 1 DNA strand; the total DNA searched is twice this number.

‡ The percentage of false-negatives was obtained by dividing the number of tRNA genes that were not identified into the total number of known tRNA genes present in the test set.

§ The percentage of false-positives was obtained as follows. The size (in nucleotides) of the "negative" set of searched DNA was derived by subtracting the values in column 2 from those in column 1. The result was divided by average tRNA length (76 nucleotides for all the divisions except the plant set, where a length of 85 nucleotides was used because of the higher proportion of introns present in that group) and multiplied by 2 to take both strands into account. The number obtained gives the number of potential tRNA regions in the negative set. Then the number of non-tRNA sequences predicted as tRNA by the algorithm was divided by the number of potential tRNA regions to arrive at the false-positive rate.

‖ The values in parentheses correspond to the number of false-positive predictions that remain after the removal of likely tRNA genes.

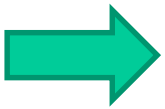
Extrait de J. Mol. Biol. (1991) 220, 659-671

~97,5% de vrais positifs et 0,37% faux positifs/million de bases

Limite de tRNAscan

• le taux de FP est acceptable pour des petits génomes mais devient problématique pour des grands génomes

• n'identifie pas les ARNt dont la structure n'est pas canonique (structure secondaire avec des bulges, bras T-Ψ-C de 8 …) dont les ARNt sélénocystéine qui ont, entre autre, 8 bp dans le bras accepteur, une région variable longue et des substitutions à plusieurs positions bien conservées



Développement d'un nouveau programme qui améliore la détection des ARNt dans les séquences génomiques

tRNAscan-SE (Lowe and Eddy, Nucleic Acids Res.,25, 955-64 (1997)) qui s'appuie sur deux méthodes existantes (tRNAscan et EufindtRNA (Pavesi *al.*, Nucleic Acids Res., 22, 1247-56 (1994)) comme premier filtre pour identifier les ARNt candidats. Ces candidats sont ensuite analysés avec un modèle de covariance des ARNt.

Les modèles de covariance sont capables de capturer à la fois des informations de la structure primaire et secondaire.

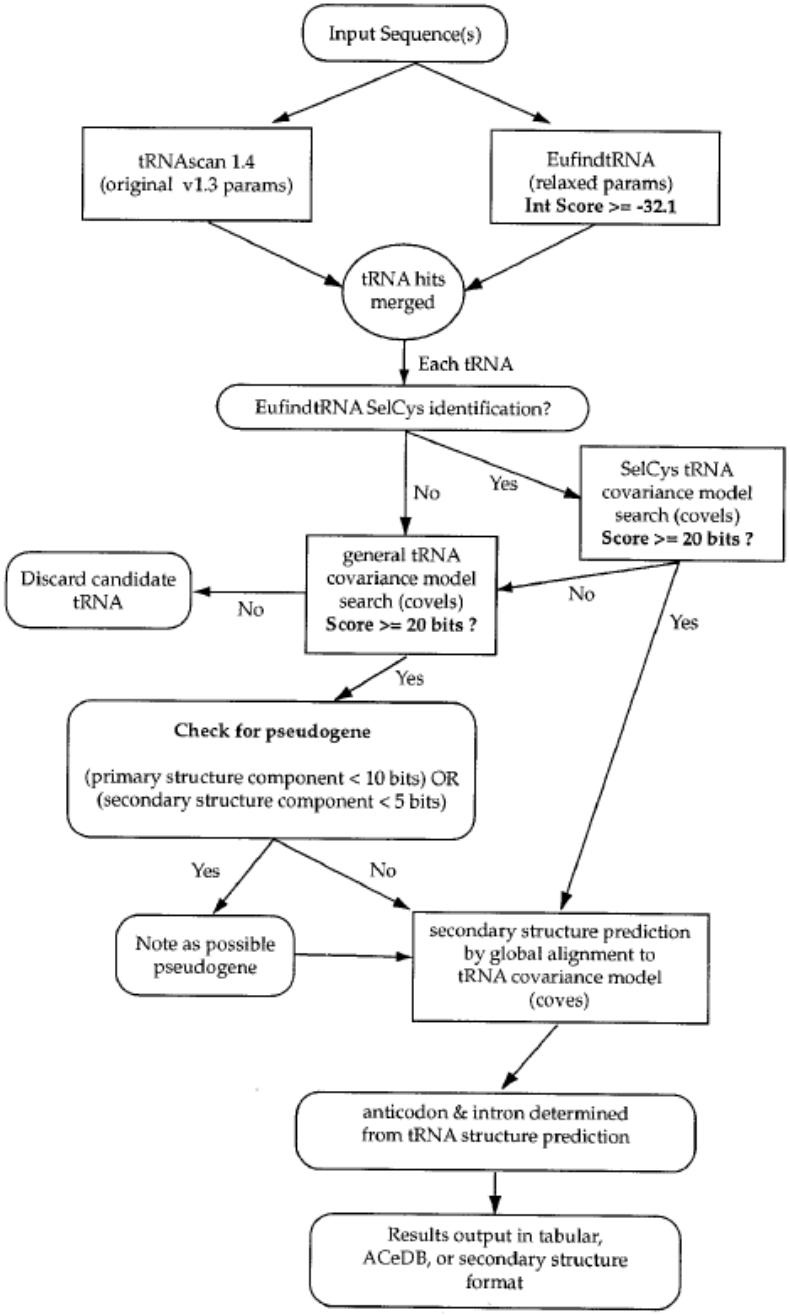


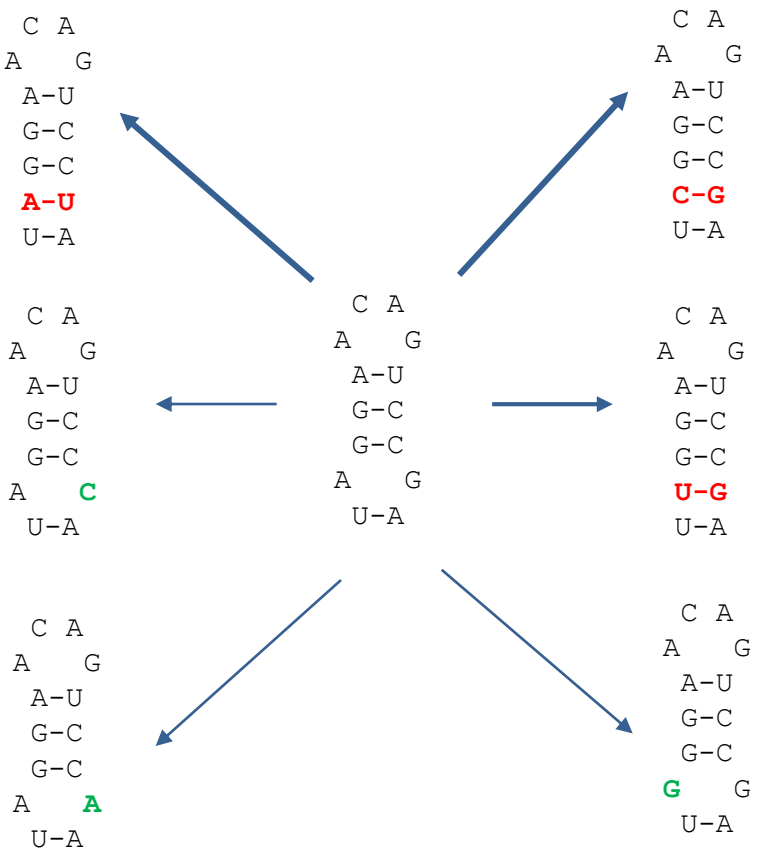
Diagramme schématique de tRNAscan-SE

(extrait de Lowe and Eddy, Nucleic Acids Res.,25, 955-64 (1997))

UNIVERSITÉ TOULOUSE III PAUL SABATIER

EufindtRNA identifie les boîtes A et B des promoteurs de la RNA polymérase III

Conservation de la structure secondaire : mutations compensatrices



Mutations qui rétablissent l'appariement

→ Plus forte probabilité de se fixer

Conservation de la structure secondaire : mutations compensatrices

Mutations compensatrices : rétablissent l'appariement – co-évolution des sites

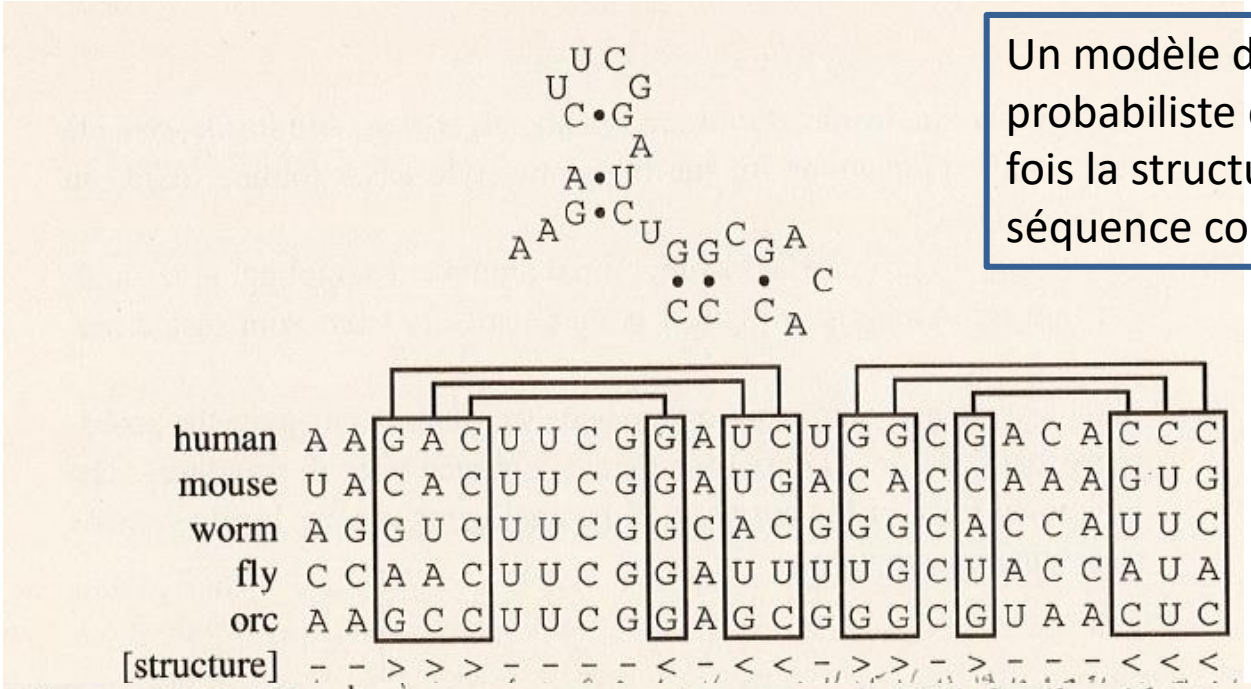


Figure 10.11 *The consensus structure of an example RNA family with no insertions or deletions is shown at the top. Five example sequences from different organisms that adopt the same structure are shown in the multiple alignment below the structure. Base-paired positions in the alignment are boxed and base-paired partners are connected by lines. The last line in the alignment is a structure consensus representation in the format we use for annotating RNA structural alignments [Konings & Hogeweg 1989].*

Un modèle de covariance est un modèle probabiliste qui décrit proprement à la fois la structure secondaire et la séquence consensus primaire d'un ARN.

Comparaison des résultats des différentes approches

Table 2. tRNA prediction within annotated database subsets

Sequence source	Literature	tRNAscan 1.3		EufindtRNA		tRNA CM		tRNAscan-SE	
	tRNAs	Total	(%)	Total	(%)	Total	(%)	Total	(%)
Sprinzl db (Archaea)	70	69	(98.6)	43	(61.4)[a]	70	(100)	70	(100)
Sprinzl db (Eubacteria)	240	226	(94.2)	205	(85.4)[a]	239	(99.6)	237	(98.7)
Sprinzl db (Eukarya)	279	265	(95.0)	275	(98.6)	279	(100)	279	(100)
Sprinzl db (total)	589	560	(95.1)	523	(88.8)	588	(99.8)	586	(99.5)
Genbank tRNA subset	1462	1366	(93.4)	760	(52.0)	1456	(99.6)	1440	(98.5)

extrait de Lowe and Eddy, Nucleic Acids Res.,25, 955-64 (1997))

Table 3. tRNAs identified in genomic databases by various search methods

Sequence source	Size (Kbp)	Literature tRNAs	tRNAscan 1.3 Total	(%)	EufindtRNA[a] Total	(%)	tRNA CM Total	(%)	tRNAscan-SE Total	(%)
M.genitalium	580	33	36	(100)	19 +1 FP	(52.8)	36	(100)	36	(100)
H.influenzae	1830	56	55	(98.2)	42 +2 FP	(73.7)	58	(103.6)	58	(103.6)
M.jannaschii	1730	37	36	(97.3)	20 +1 FP	(54.0)	37	(100)	37	(100)
S.pombe (through 9/96)	4176	–	45 +4 FP	(93.7)	46 +1 FP	(95.8)	48		48	(100)
S.cerevisiae	12 057	273	270 +4 FP	(98.5)	274 +10 FP +1 pseudo	(100)	274 +1 pseudo		274 +1 pseudo	(100)
C.elegans (through 11/13/96)	58 402	– 16 FP	389 +29 FP	(96.5)	400 +355 FP +19 pseudo	(99.2)	403 +23 pseudo		403 +11 id pseudo +8 pseudo	(100)
P.anserina mitochondrion	100	27	18	(66.7)	11	(40.7)	27	(100)	22	(81.5)