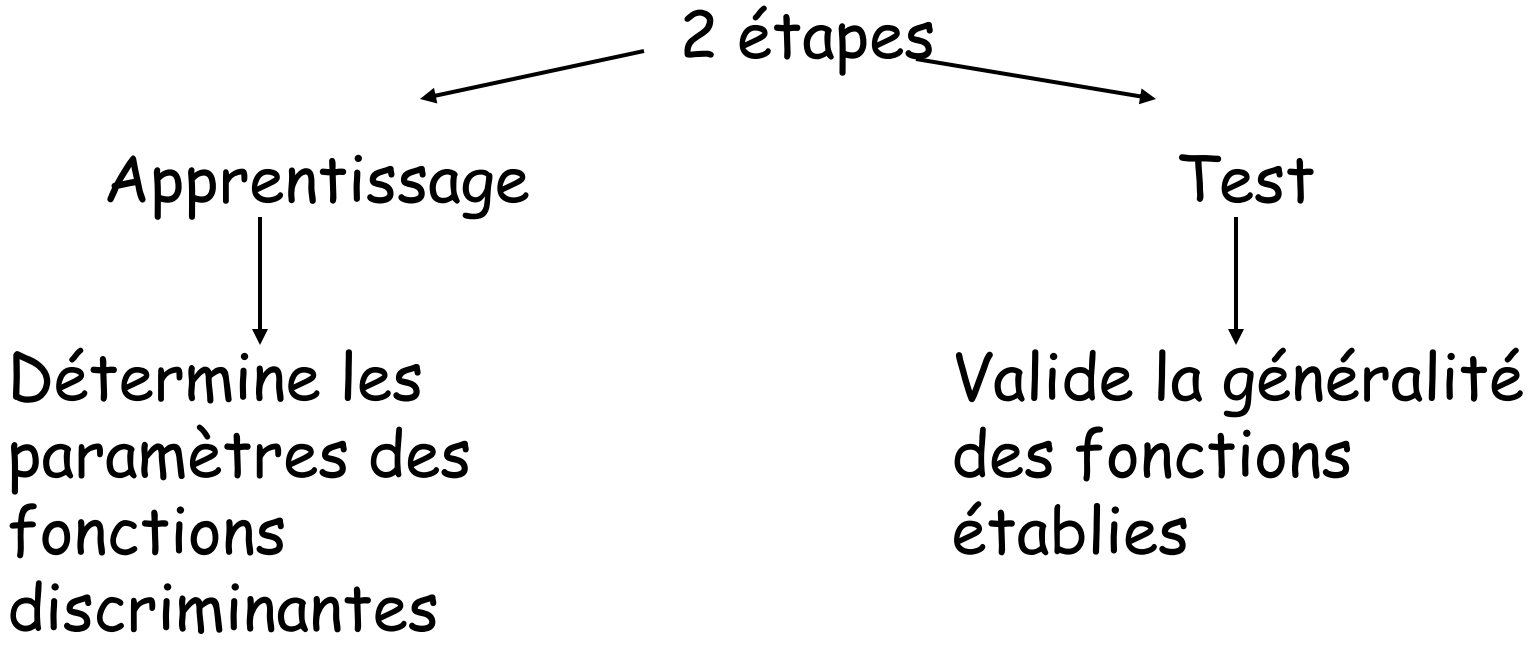
# Support de cours
# Annotation des génomes
# (Partie I)

# 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)

| | | Réalité: Groupe 1 | Réalité: Groupe 2 |
|---|---|---|---|
| prédiction | Groupe 1 | % vrais positifs | % faux positifs |
| prédiction | 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

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)

Identification des unités de traduction

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….)

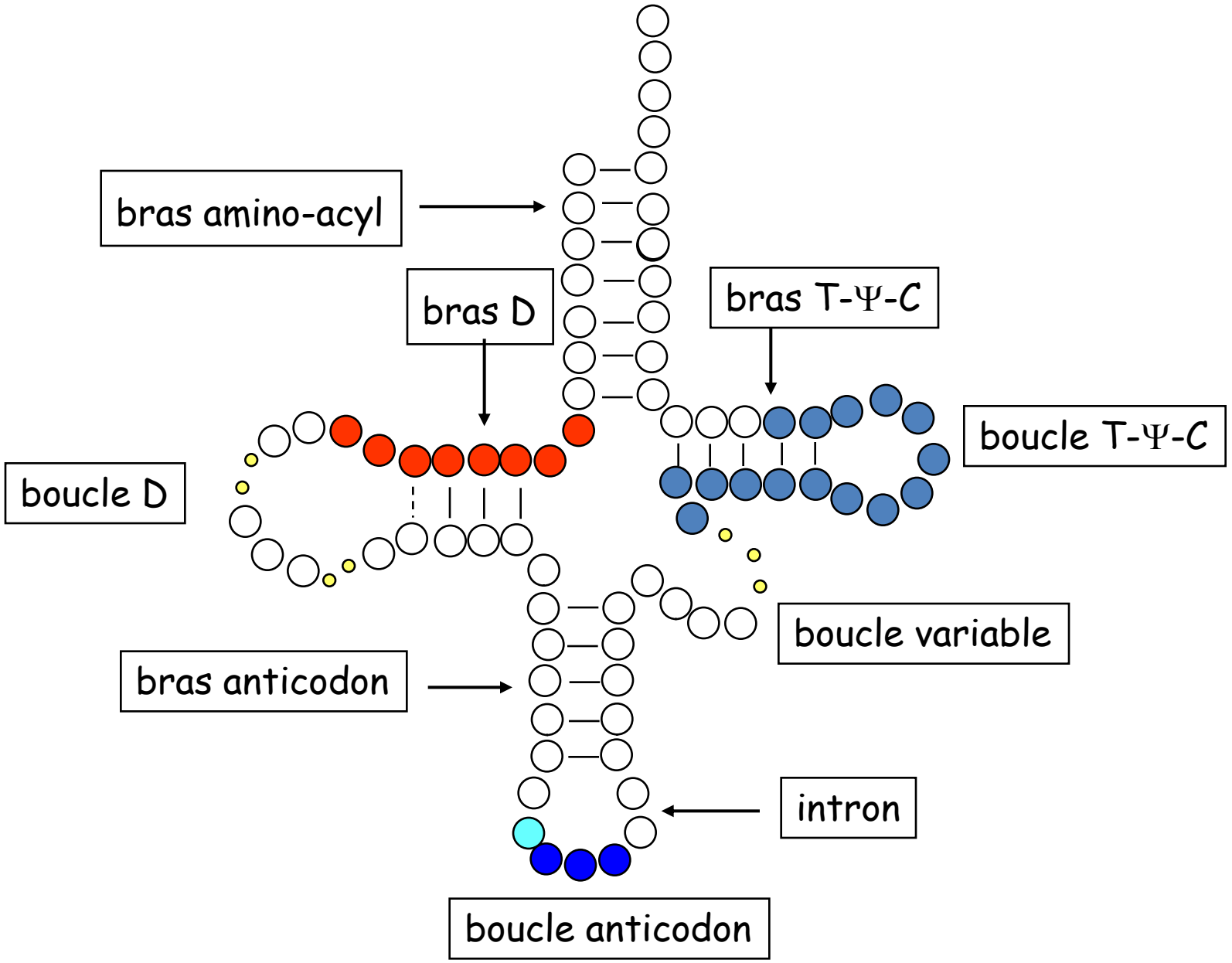
# Exemple d'annotation d'un génome
## *Mycoplasma genitalium*

# Identification des gènes nucléaires codant pour des ARNt (tRNAscan,) 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

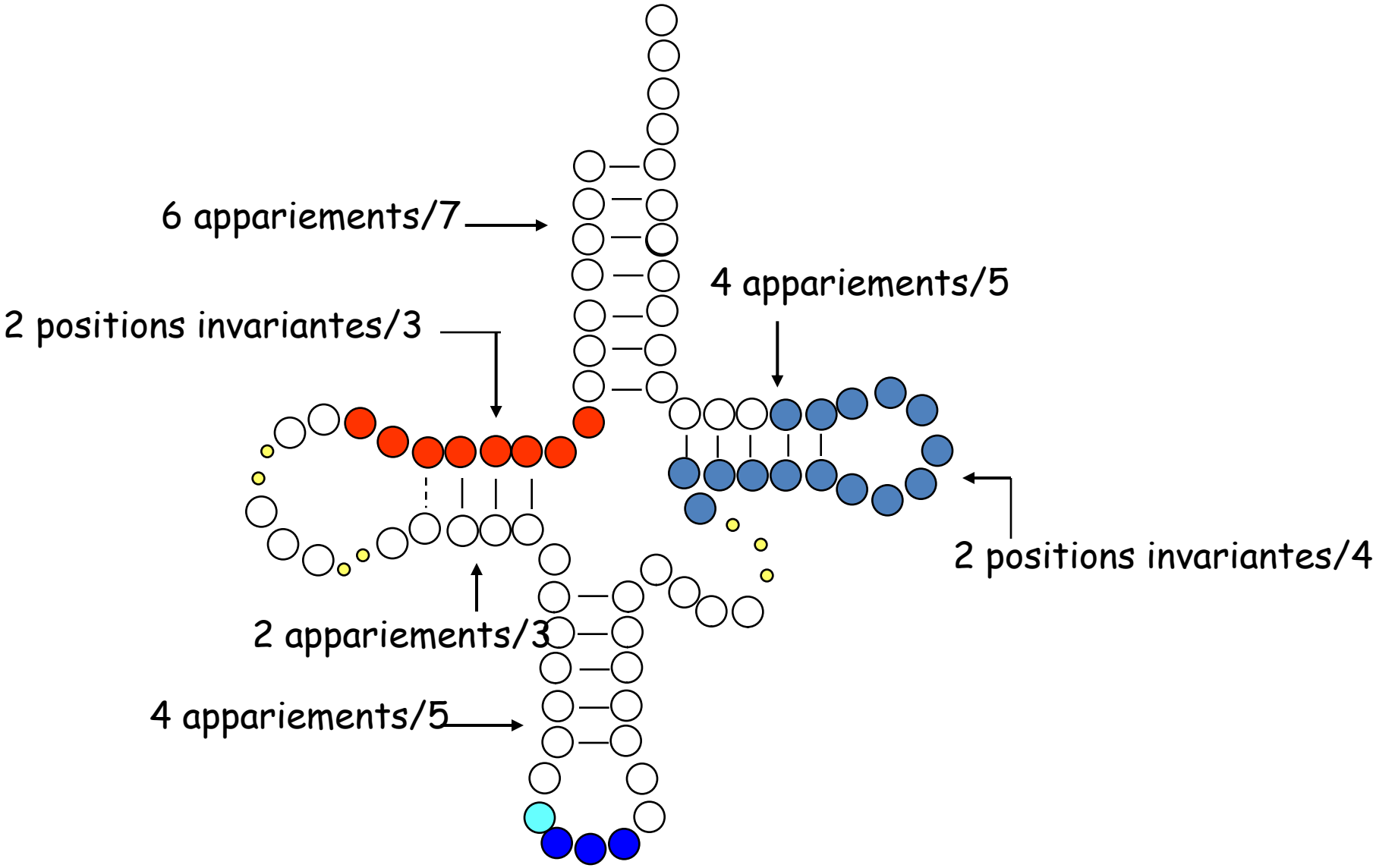


Critères utilisés et appris sur un ensemble de gènes d'ARNt connus:

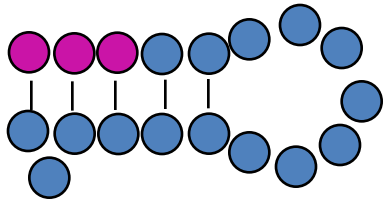
- 2 motifs de type signal correspondant aux régions conservées T-Ψ-C et D représentés par des matrices consensus des fréquences des bases

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

Définition des seuils : limites inférieures



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



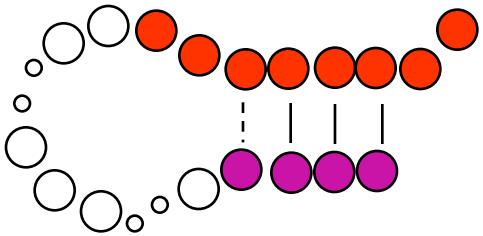
**Motif T-Ψ-C**

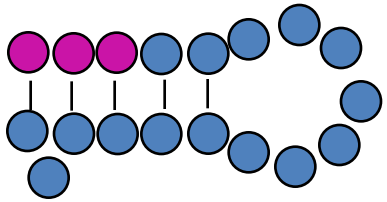
**bras T-Ψ-C**

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

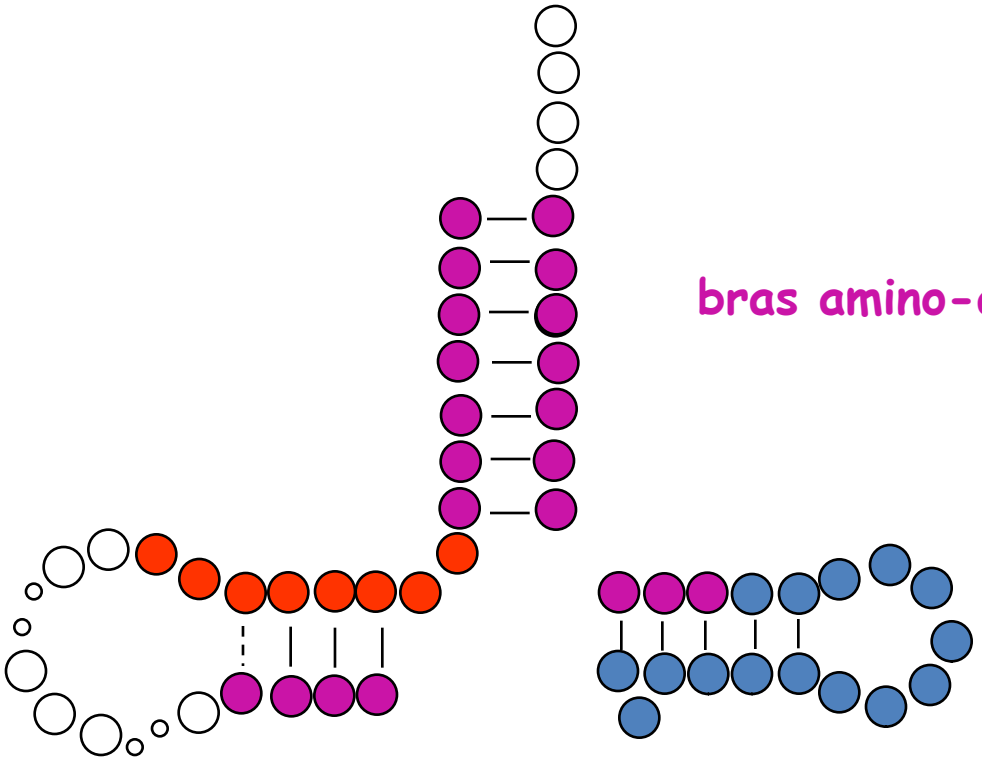
Motif D

région D



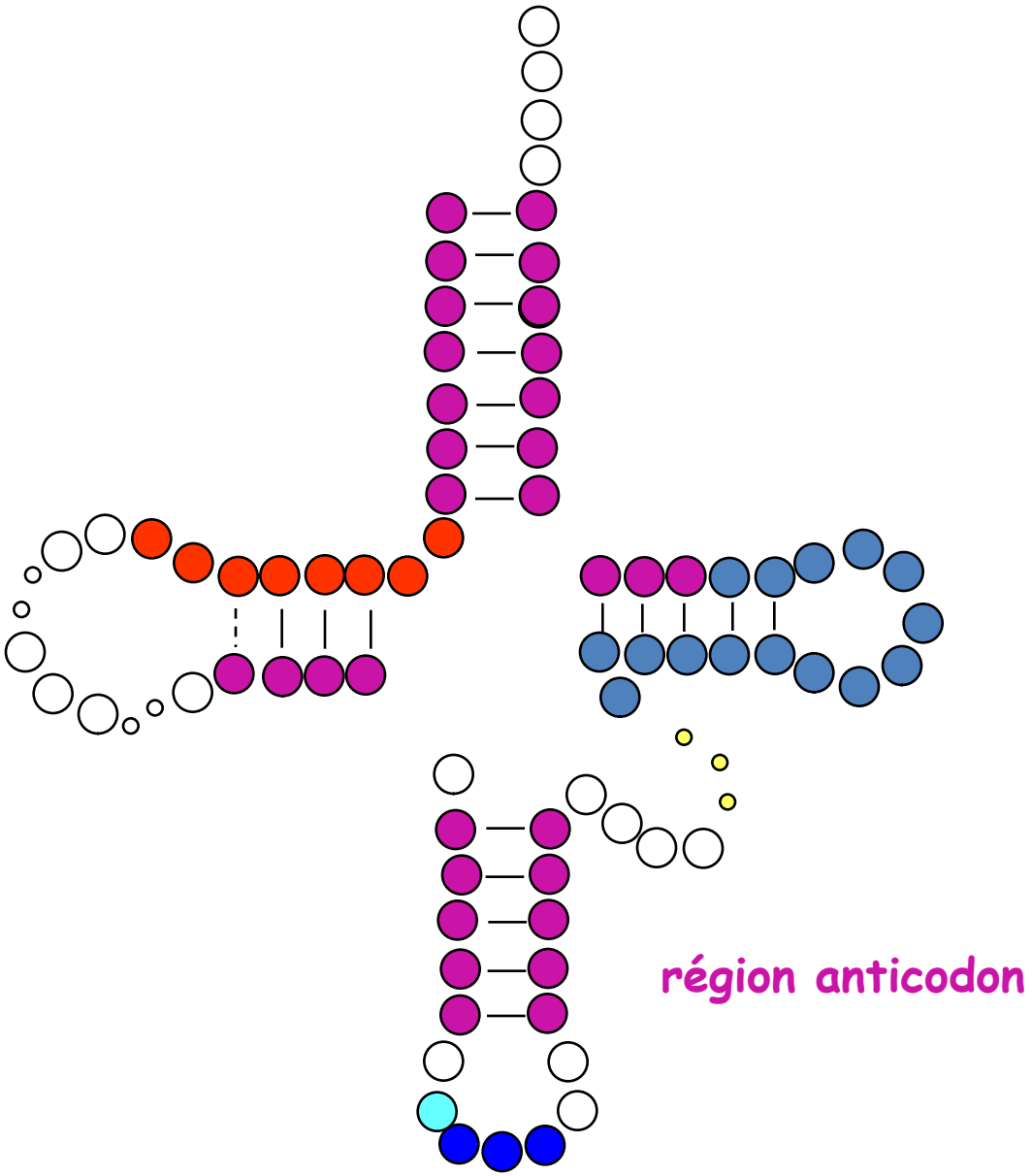


# Recherche de la région amino-acyl : Etape 5 de l'algorithme



bras amino-acyl

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



# Algorithme de tRNAscan



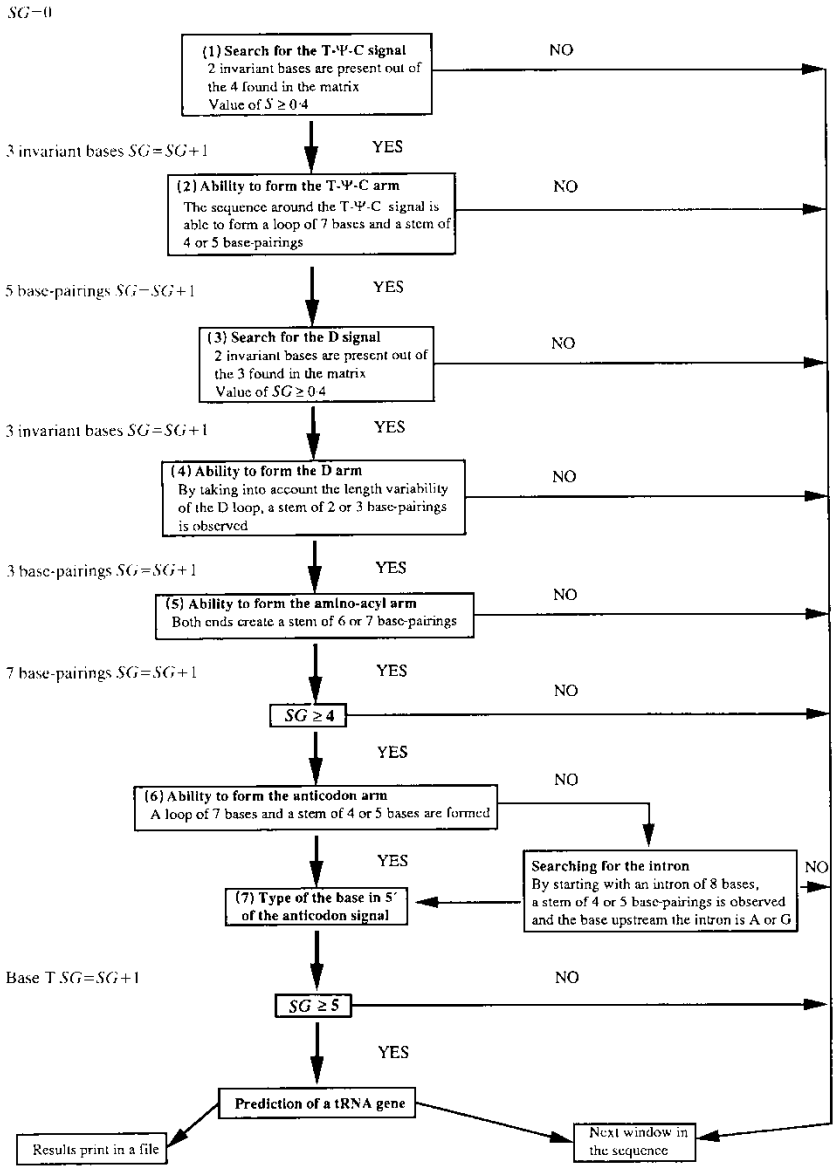
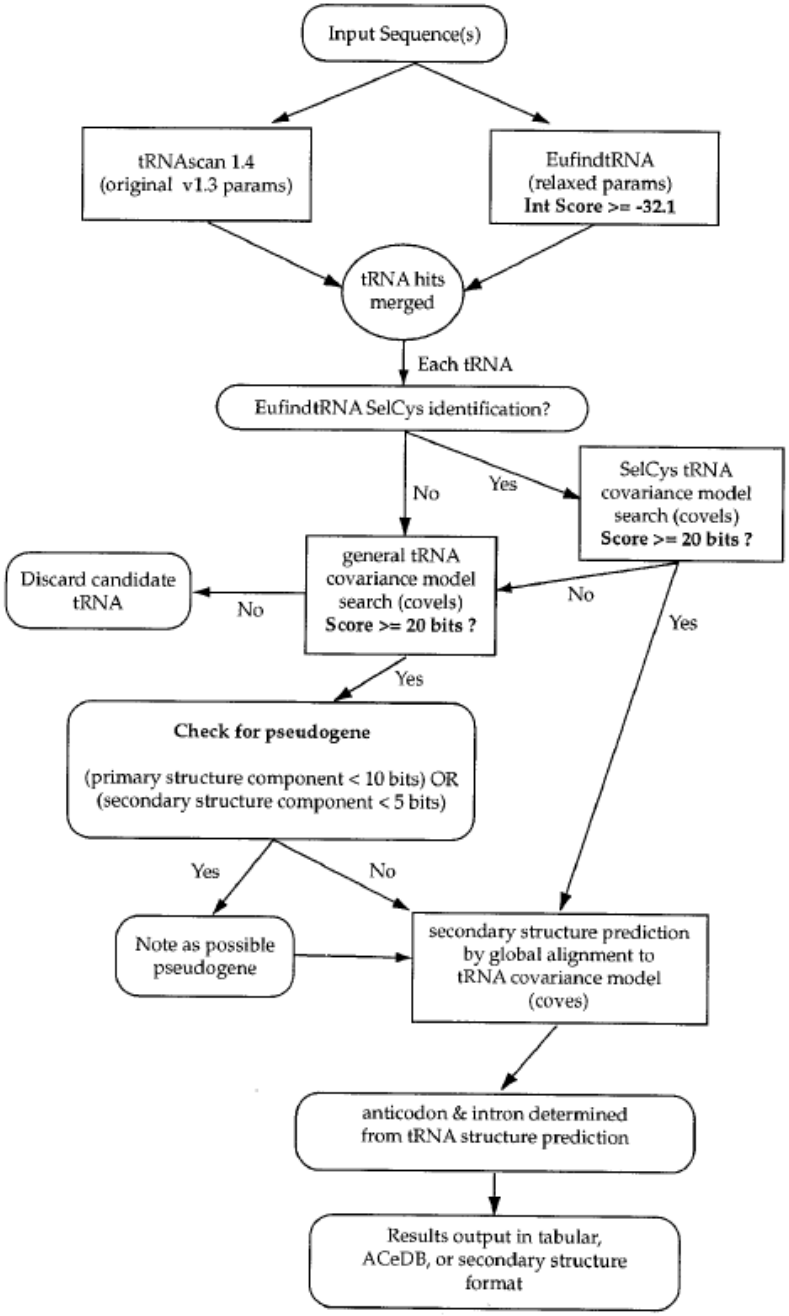



**Figure 2.** Algorithm description. Each step of the algorithm, with corresponding threshold values, is in a separate box. If the score for the windowed sequence does not exceed the threshold at any step, the algorithm is initiated again on the next window of the sequence under study. The incrementation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




# Diagramme schématique de tRNAscan-SE

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

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