

Classification d'ARN: codant / non-codant

Arnaud FONTAINE, Hélène TOUZET
arnaud.fontaine@lifl.fr, touzet@lifl.fr

LIFL - UMR CNRS 8022 - USTL

AReNa
7, 8, 9 décembre 2005



Prédire les ARN non-codants

- ▶ À partir d'une séquence : des pistes de réflexion
 - ▶ **biais de composition** en di-nucléotides
 - ▶ évaluer la **significativité de l'énergie libre**
 - ▶ unicité du repliement

Prédire les ARN non-codants

- ▶ À partir d'une séquence : des pistes de réflexion
 - ▶ **biais de composition** en di-nucléotides
 - ▶ évaluer la **significativité de l'énergie libre**
 - ▶ unicité du repliement

- ▶ À partir de plusieurs séquences : l'**analyse comparative**
 - ▶ QRNA examine les motifs de substitutions : conservation de la fonction d'une protéine ou d'un ARN non-codant
 - ▶ MSARi évalue la significativité d'un nombre de mutations compensatoires observées
 - ▶ RNAz évalue la significativité de l'énergie libre de structures individuelles entre elles et avec une structure consensus

Méthodes existantes

- ▶ Base de départ : **alignement multiple** entre les séquences
- ▶ **Bonne spécificité** - prédictions positives fiables
 - ▶ supérieure à 90%
 - ▶ apport du modèle codant (QRNA)
- ▶ **Sensibilité médiocre** - prédictions négatives moins fiables
 - ▶ 70% pour RNAz, 40% pour QRNA

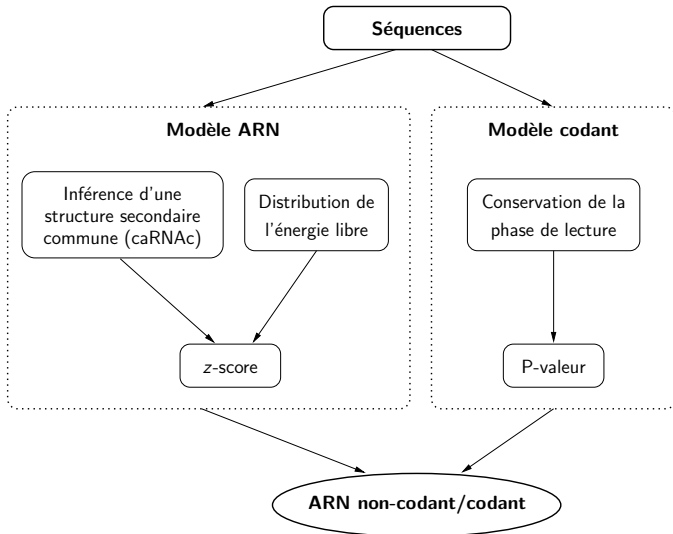
Méthodes existantes

- ▶ Base de départ : **alignement multiple** entre les séquences
- ▶ **Bonne spécificité** - prédictions positives fiables
 - ▶ supérieure à 90%
 - ▶ apport du modèle codant (QRNA)
- ▶ **Sensibilité médiocre** - prédictions négatives moins fiables
 - ▶ 70% pour RNAz, 40% pour QRNA
- ▶ Influence de la similitude entre séquences
 - ▶ résultats médiocres hors de 65 à 90% d'identité
- ▶ Influence du nombre de séquences
 - ▶ QRNA : 2 séquences par nature
 - ▶ RNAz : **nette dégradation au delà de 5 séquences**
 - ▶ MSARi : plus de 10 séquences (par nature), résultats médiocres

Cahier des charges

- ▶ S'affranchir de l'alignement multiple des séquences
- ▶ Être robuste aux structures bruitées
- ▶ Être indépendant du nombre de séquences
- ▶ Exploiter des informations complémentaires
 - ▶ thermo-dynamiques
 - ▶ évolutives : motifs de substitutions (mutations compensatoires)
 - ▶ à confronter avec un modèle codant

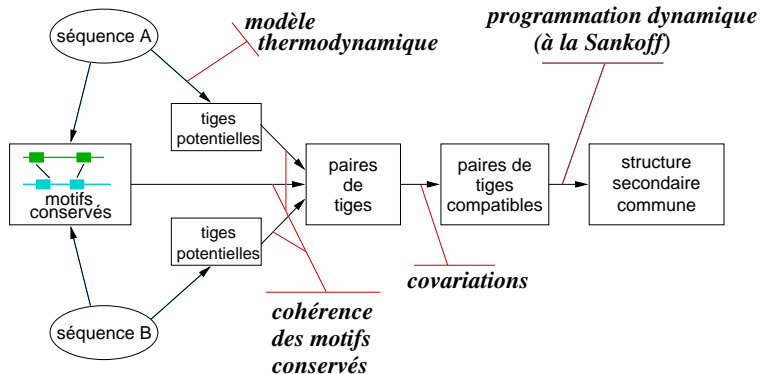
Carnac Extended



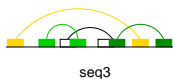
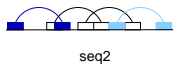
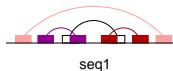
caRNAc

- ▶ Inférence d'une structure secondaire commune **sans alignement multiple** préalable
- ▶ Ses atouts : **rapidité et spécificité**
- ▶ Stratégie de prédiction en deux étapes
 - ▶ prédiction 2 à 2
 - ▶ combinaison de l'ensemble des prédictions

caRNAc 2 à 2

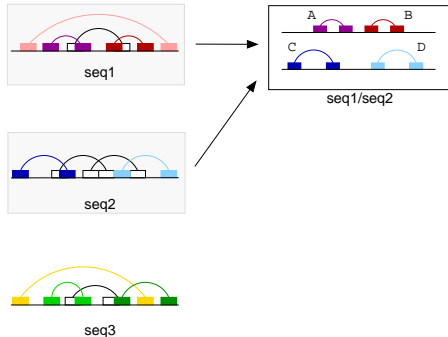


caRNAc : passage à n séquences



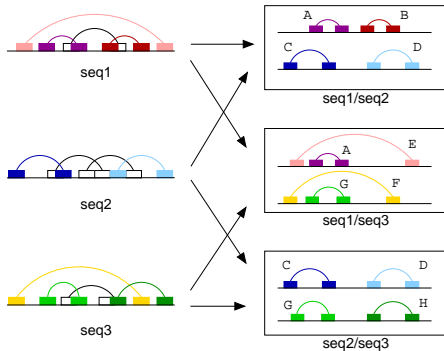
1. Identification des tiges potentielles, séquence par séquence

caRNAc : passage à n séquences



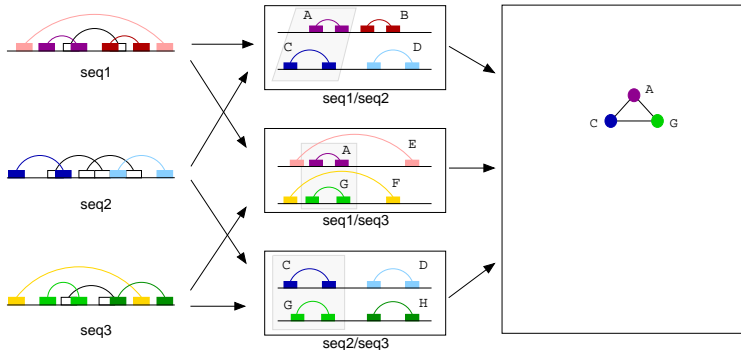
2. caRNAc 2 à 2

caRNAc : passage à n séquences



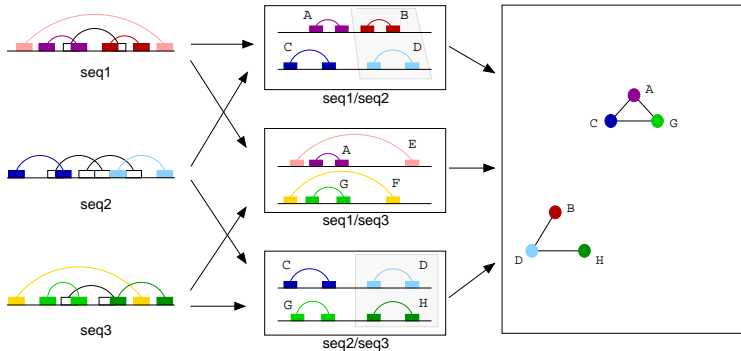
2. caRNAc 2 à 2

caRNAC : passage à n séquences



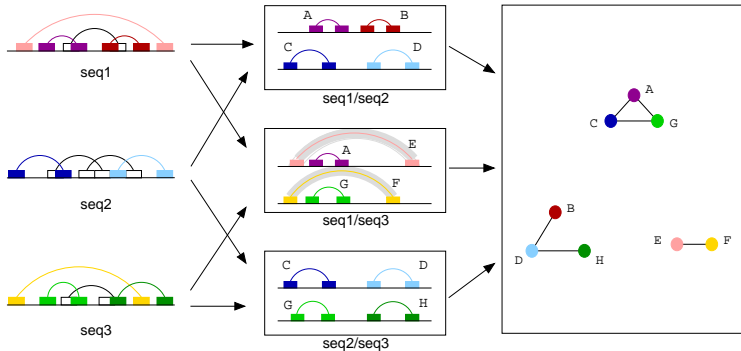
3. Combinaison des prédictions 2 à 2 : graphe des tiges

caRNAc : passage à n séquences



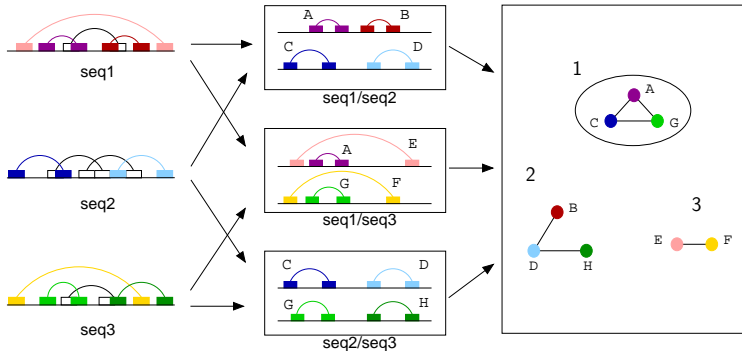
3. Combinaison des prédictions 2 à 2 : graphe des tiges

caRNAc : passage à n séquences



3. Combinaison des prédictions 2 à 2 : graphe des tiges

caRNAc : passage à n séquences



4. Sélection finale des tiges : recherche de composantes connexes denses (critère numérique sur le nombre d'arêtes et de nœuds)

Modèle ARN : significativité de la structure conservée

Construction d'une distribution empirique de l'énergie libre

- 1 Alignement avec ClustalW des séquences initiales
- 2 Génération de 100 alignements mélangés (même composition, même longueur, même conservation)
- 3 Extraction des séquences
- 4 Inférence d'une structure avec caRNAC
- 5 Calcul d'un z-score à partir de la moyenne et de l'écart-type de la distribution de l'énergie libre

Modèle ARN : les shuffles

- ▶ Programme `shuffle-align.pl` (Vienna Package)
- ▶ Conservation préservée de chaque position de l'alignement

Exemple :

```
UAGGUGAGCUAGGCCUCUAUGAUUCGUGCAUCAGGGUCUAAUCGGUUCGAG
UAGGUGAGCUAGGCCUCU-----GUGGUCUAACCGGUUCGAG
UAGGUAAGCUAGGCCUCU-----CCGGUCUAACCGGUUCGAG
UAGGUGAGCUAGGCCUCGGCUCAGUAGCGGCAGUGGUCUAACCGGUUCAA
***** ***** *                ***** ***** *
```

Modèle ARN : les shuffles

- ▶ Programme `shuffle-align.pl` (Vienna Package)
- ▶ Conservation préservée de chaque position de l'alignement

Exemple :

```
UAGGUGAGCUAGGCCUCUAUGAUUCGUGCAUCAGGGUCUAAUCGGUUCGAG
UAGGUGAGCUAGGCCUCU-----GUGGUCUAACCGGUUCGAG
UAGGUAAGCUAGGCCUCU-----CCGGUCUAACCGGUUCGAG
UAGGUGAGCUAGGCCUCGGCUCAGUAGCGGCAGUGGUCUAACCGGUUCAA
***** ***** *                ***** ***** *
```

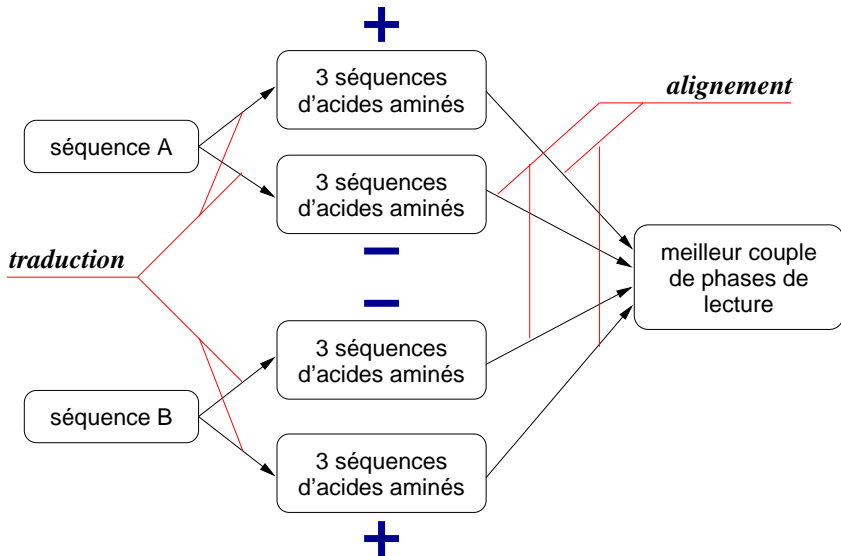
Après mélange :

```
CGUGGGUGGUAGAGUUCGUGCCGUUCUGGAAAUGAAGCACUGUCCAAUUUCG
CGUGGGUGGUAGAGUUCG-----UGAGCACUGCCCAAUUUCG
CGUGGAUGGUAGAGUUCG-----CCAGCACUGCCCAAUUUCG
CGUGGGUGGUAGAGUUCAGUACCCGGUACGGAUGAGCACUGCCCAAUUGCA
***** ***** *                ***** ***** *
```

Modèle codant : motivation

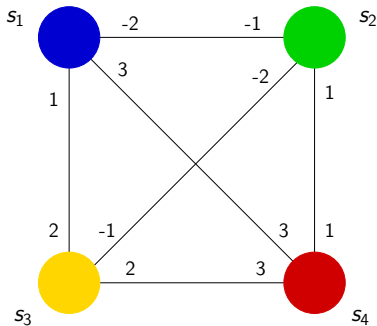
- ▶ **Idée de base** : conservation au niveau d'une éventuelle séquence protéique
- ▶ **Problème** : 6 phases à explorer par séquence, donc 6^n possibilités à tester
- ▶ **Heuristique** : comparer les séquences 2 à 2 puis traiter les résultats

Modèle codant : 2 à 2



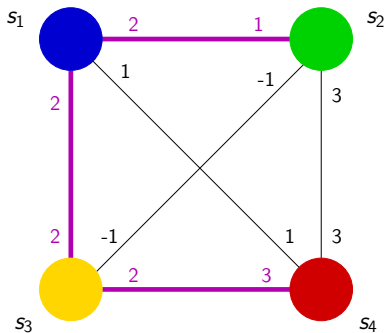
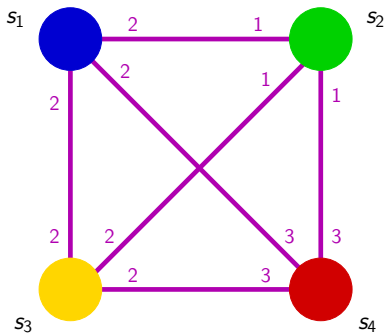
Modèle codant : modèle

- ▶ **Chaque** séquence est alignée **contre toutes** les autres
- ▶ Combinaison des résultats dans un graphe non-orienté
 - ▶ sommet = séquence
 - ▶ arête = couple de phases de lecture optimal



Modèle codant : sous-graphe cohérent maximal

- ▶ Recherche des **sous-graphes cohérents maximaux**
 - ▶ **cohérent** : une seule phase est affectée à chaque séquence
 - ▶ **maximal** : on ne peut plus ajouter d'arête sans briser la cohérence



Modèle codant : significativité d'un sous-graphe cohérent maximal

- ▶ Calcul d'une P-valeur basée sur la taille du sous-graphe cohérent maximal
 - ▶ n séquences
 - ▶ $N = \frac{n(n-1)}{2}$ arêtes dans le graphe
 - ▶ k : nombre d'arêtes dans le sous-graphe (k supposé grand)

$$p(k, N) \leq \frac{\sum_{i=k}^N C_N^i \times 6 \times 5^{N-i}}{6^N}$$

- ▶ Possibilité d'accélérer le calcul pour les cas simples

Résultats

- ▶ 21 familles d'**ARN non-codants**
 - ▶ RFAM, MicroRNA Registry, BRaliBase, The RNase P database
 - ▶ eucaryotes, procaryotes, archae
 - ▶ longueur moyenne comprise entre 20 et 370nt
 - ▶ identité moyenne comprise entre 44 et 97%

- ▶ 25 familles d'extraits d'**ARN messagers**
 - ▶ HomoloGene, TIGR, Ensembl
 - ▶ eucaryotes, procaryotes
 - ▶ longueur moyenne comprise entre 63 et 300nt
 - ▶ identité moyenne comprise entre 40 et 90%

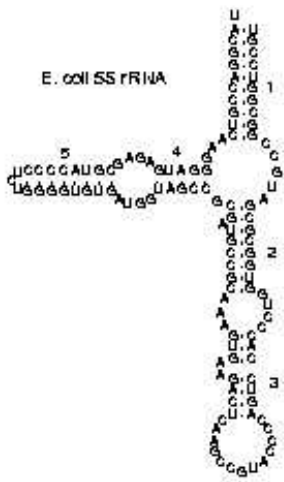
Résultats pour les ARN non-codants

	Modèle codant	Modèle ARN	Modèle codant strict	Modèle ARN strict	RNAz
2 séq.	5%	64%	1%	44%	64%
3 séq.	14%	72%	5%	72%	66%
4 séq.	17%	78%	4%	65%	67%
5 séq.	7%	74%	2%	70%	70%

Résultats pour les CDS

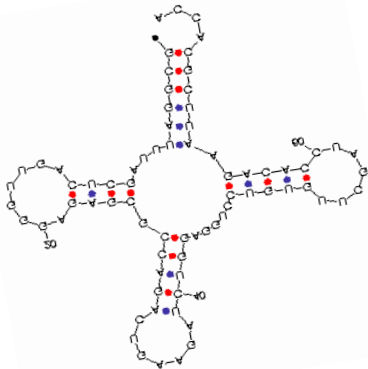
	Modèle codant	Modèle ARN	Modèle codant strict	Modèle ARN strict	RNAz
2 séq.	45%	11%	41%	7%	95%
3 séq.	80%	28%	53%	2%	97%
4 séq.	80%	11%	69%	< 1%	96%
5 séq.	82%	18%	70%	< 1%	95%

L'exemple des ARN 5S



- ▶ 10 séquences
- ▶ 57% d'identité moyenne
- ▶ RNAz et MSARi ne détecte pas un gène à ARN
- ▶ Identifié comme ARN non-codant par Carnac Extended
 - ▶ modèle ARN : structure commune prédite significative (z-score= -3.7)
 - ▶ modèle codant : ne détecte rien

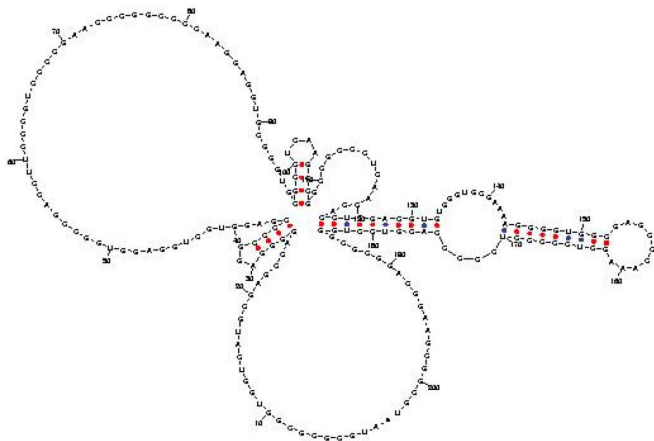
L'exemple des ARN de transfert



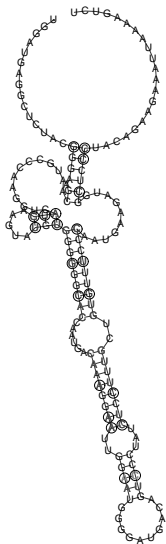
- ▶ 10 séquences
- ▶ 87% d'identité moyenne
- ▶ RNAz et MSARi ne détecte pas un gène à ARN
- ▶ Identifié comme ARN non-codant par Carnac Extended
 - ▶ modèle ARN : structure commune prédite significative ($z\text{-score} = -1.8$)
 - ▶ modèle codant : ne détecte rien

Structure commune conservée entre ARN messagers

Famille alcooldéshydrogénase, **5 séquences**, 48% d'identité
Prédiction double : codant et non-codant



Structure commune conservée entre ARN messagers



Famille heme, **6 séquences**, 75% d'identité

Prédiction double : codant et non-codant

Perspectives

- ▶ Validation sur d'autres jeux de données
- ▶ Améliorer et raffiner les modèles
 - ▶ modèle ARN : distribution de l'énergie libre
 - ▶ modèle codant : formaliser la recherche de phases en terme de vraisemblance (modèle probabiliste)
- ▶ Concevoir un **processus de décision** multi-critères (biais de composition, énergie)
- ▶ Produire selon le résultat de la classification un alignement basé sur
 - ▶ la structure commune conservée
 - ▶ **ou** l'affectation des phases

Perspectives

- ▶ Accélérer le temps de calcul du modèle ARN
- ▶ Version “filtrée” de caRNAC : inférence d’une structure **incomplète**
- ▶ Possibilité d’utiliser une fenêtre glissante - approche locale

