

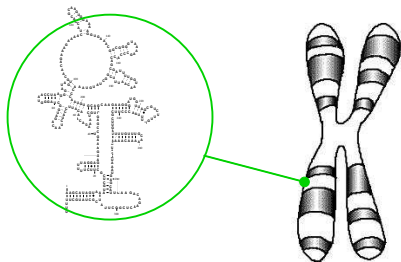
# Recherche de motifs ARN dans un génome, vite et bien

Hélène Touzet

Équipe Bioinfo/Sequoia — LIFL — USTL — INRIA



# Recherche de motifs ARN



- ▶ Problème
  - ▶ Motif : modèle pour un ARN, muni de sa structure
  - ▶ Cible : séquence d'ADN, longue
  - ▶ ? Localiser toutes les occurrences du motif dans la cible
- ▶ Deux types d'approche
  - ▶ Descripteur abstrait : Palingol, Milpat, etc.
  - ▶ Alignement structure/séquence : Rsearch, Erpin, etc.

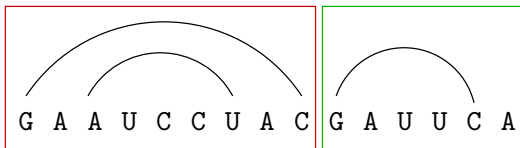
Freyhult E, Bollback JP and Gardner PP (2007) Exploring genomic dark matter : a critical assessment of the performance of homology search methods on non-coding RNA.

# Alignement structure/séquence

- ▶ Opérations d'édition sur les bases, sur les appariements
- ▶ Scores
  - ▶ substitutions des bases ( $4 \times 4$ )
  - ▶ substitution des appariements ( $16 \times 16$ )
  - ▶ insertion de bases, d'appariements
- ▶ Alignement structure/séquence

```
( (      ) ) ( (      ) )  
C A U A U G A C U U G U  
                | | |  
U C A G G A G - U U - U  
( (      ) ) (      )
```

Motif



Cible

A G A C G U U G C U C A A U C C G A A U



A A U C C U A

G A U U C A

G A C G

U G C U C A A U C C G A A U

G A C G U

G C U C A A U C C G A A U

G A C G U U G C

C A A U C C G A A U

G A C G U U G C U C A A

C C G A A U

- ▶ Rsearch (infernial – RFAM)
  - ▶ grammaire stochastique
  - ▶ système de score : matrices Ribosum
- ▶ HomoStRscan
  - ▶ pénalités de gaps affines

RSEARCH : Finding homologs of single structured RNA sequences R.J. Klein and S.R. Eddy BMC Bioinformatics 2003

Shu-Yun Le, Jacob V. Maizel, Kaizhong Zhang, An Algorithm for Detecting Homologues of Known Structured RNAs in Genomes, IEEE Computational Systems Bioinformatics Conference (CSB'04), 300-310, 2004



résultats de bonne qualité



trop lent pour la recherche à grande échelle

▶ Rsearch

*recherche d'un motif RNase P dans la Nucleotide Archaeal Database  
(  $2.1 \times 10^7$  nt) : 38 jours CPU*

▶ HomoStRscan

*We expect to improve our method to be more efficient.*

## Gagner du temps : Erpin

```
2 2 2 2 0 0 0 0 2 2 2 2
A C G U C C U - A C G U
A C U U A C G G U A G U
A C U U A C C A A A G U
```

- ▶ variation contrôlée de la longueur des boucles (profil)
- ▶ interdiction des indels dans les hélices



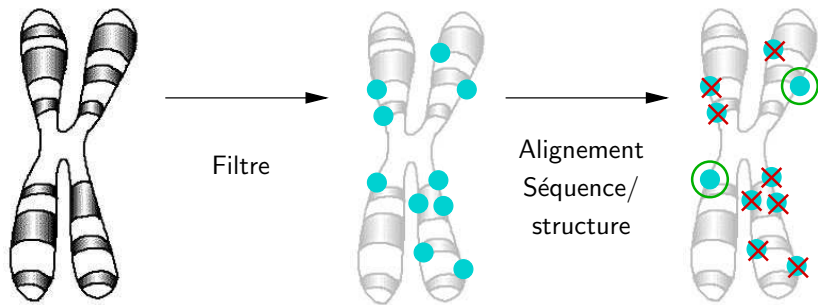
rapide, voire très rapide



manque de sensibilité

Gautheret D, Lambert A. (2001) Direct RNA Motif Definition and Identification from Multiple Sequence Alignments using Secondary Structure Profiles. J Mol Biol. 313 :1003-11

## Vite + bien : la solution du filtre



- ▶ Sensibilité (pour ne pas perdre d'occurrences)
- ▶ Rapidité (temps linéaire)
- ▶ Suffisamment sélectif (pour accélérer la phase d'alignement)

# Filtres pour la recherche d'ARN

Filtre

Structure primaire

Blast

YASS

Ravenna

Structure secondaire

FastR

Alignement

Rsearch

HomoStRscan

# FastR

## ► Filtre :

- identification de tiges, paramétrées par la longueur et l'étendue
- définition d'*unités* : parallel, nested, et multiloop
- combinaison des unités : programmation dynamique



## ► Alignement : programmation dynamique

Shaojie Zhang, Brian Haas, Eleazar Eskin, Vineet Bafna : Searching Genomes for Noncoding RNA Using FastR. IEEE/ ACM Trans. Comput. Biology Bioinform. 2(4) : 366-379 (2005)

## Qu'est-ce qu'une structure secondaire ?

- ▶ Ensemble de tiges
- ▶ Contraintes globales sur la situation relative des tiges



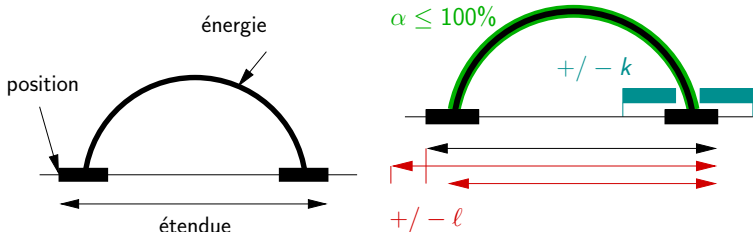
tiges juxtaposées



tiges emboîtées



Relaxer les contraintes sur la juxtaposition et l'emboîtement



- ▶ une tige :  $\langle \text{position}, \text{étendue}, \text{énergie} \rangle$
- ▶  $(k, l, \alpha)$  compatible :
  - ▶ variation de  $k$  positions autour de la position de fin
  - ▶ variation de  $l$  positions pour l'étendue
  - ▶ seuil minimal d'énergie avec un facteur  $\alpha$

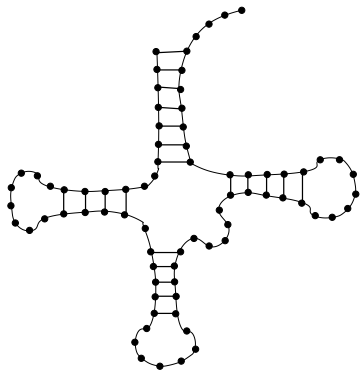
▶ **Entrées**

- ▶  $\mathcal{S}$  un motif décrit par une liste de tiges  
<position, étendue, énergie>
- ▶ une séquence cible
- ▶ une fonction d'énergie pour les tiges

▶ **Paramètres**

- ▶  $(k, \ell, \alpha)$ , paramètres de compatibilité entre tiges
- ▶  $a$ , un nombre de tiges

- ▶ **Occurrence** : position  $i$  du texte pour laquelle au moins  $a$  tiges de  $\mathcal{S}$  ont une tige compatible dans la séquence commençant en  $i$



Quatre tiges

position étendue énergie

1 72 18

10 15 12

27 16 14

48 17 14

## Motif

tige 1 : 1, 7, 10

tige 2 : 8, 4, 5

paramètres

$k, l, \alpha, a$

---

Séquence cible

## Motif

tige 1 : 1, 7, 10  
tige 2 : 8, 4, 5

$$\alpha = 80\%$$

$$\ell = 2$$

## Table des étendues

étendue max +  $\ell$

0	0	0	0	8	8	8	8	8
0	4	4	4	4	4	0	0	0
0	4	4	4	4	4	8	8	8

nb tiges

paramètres

$k, \ell, \alpha, a$

---

stem[j]=e  
tige d'étendue  $j$  et d'énergie  $e$

Séquence cible

## Motif

tige 1 : 1, 7, 10  
tige 2 : 8, 4, 5

$$\alpha = 80\%$$

$$\ell = 2$$

## Table des étendues

étendue max +  $\ell$

0	0	0	0	8	8	8	8	8
0	4	4	4	4	4	0	0	0
0	4	4	4	4	4	8	8	8

nb tiges

## Salle d'attente

1	0	0	0
0	1		

nb tiges

## Tiges actives

0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0

nb tiges

$2k + 1$

paramètres

$k, \ell, \alpha, a$

tige active : 1 sur sa ligne

occurrence : au moins  $a$  tiges actives



stem[j]=e

tige d'étendue  $j$  et d'énergie  $e$

Séquence cible

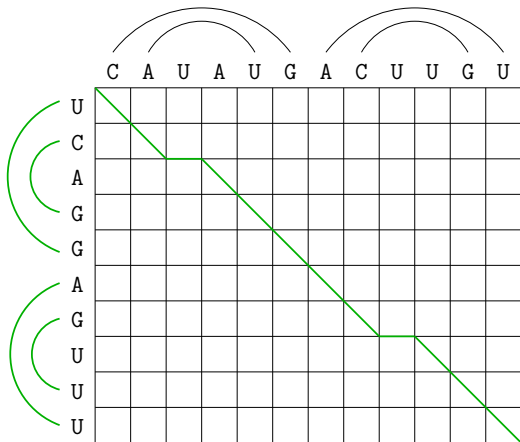
## Performances du filtrage

- ▶ Séquence aléatoire 1Mb
- ▶ Paramètre : préserver la sensibilité
  - ▶ ARNt :  $k, \ell = +/- 5$  bases,  $\alpha = 80\%$ , 4 tiges présentes
  - ▶ RNase P :  $k, \ell = +/- 30$  bases,  $\alpha = 80\%$ , 1 tige absente

	nbre occurrences	temps
ARNt 78 nt, 4 tiges	5040	5s
RNase P – <i>P. furiosus</i> 330 nt, 16 tiges	9	21s

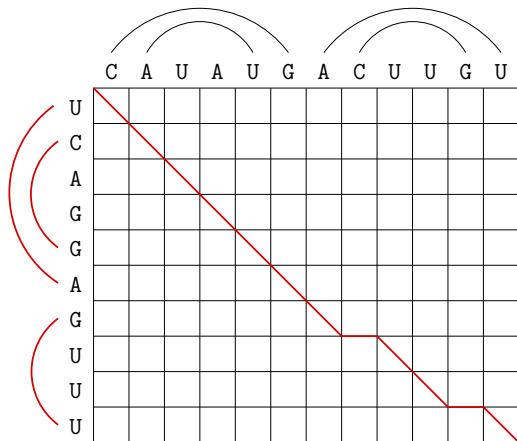


## Alignement structure/séquence



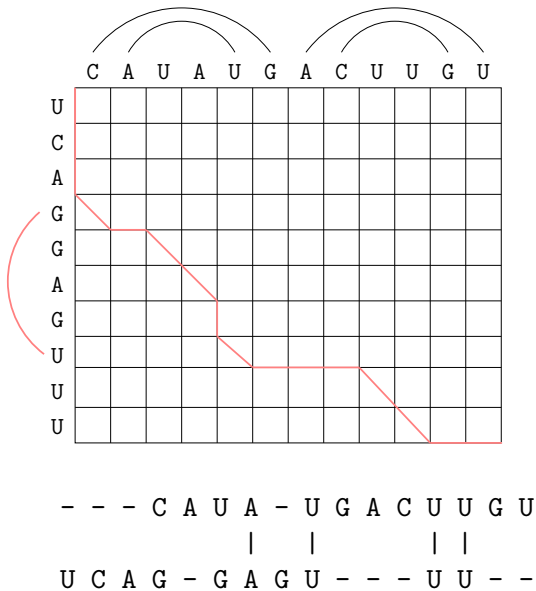
C A U A U G A C U U G U  
| | | | |  
U C - A G G A G - U U U

## Alignement structure/séquence

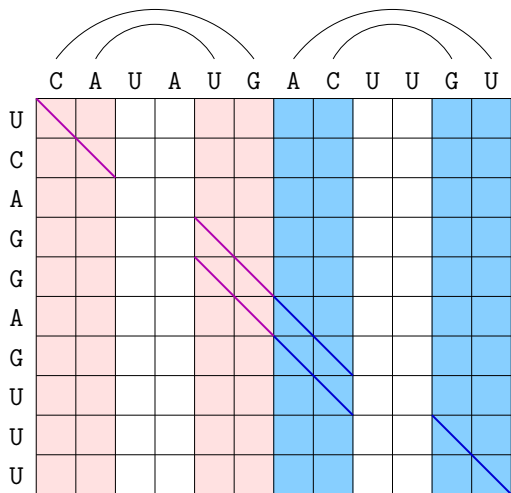


C A U A U G A C U U G U  
                  | | |  
U C A G G A G - U U - U

## Alignement structure/séquence



# Alignement structure/séquence

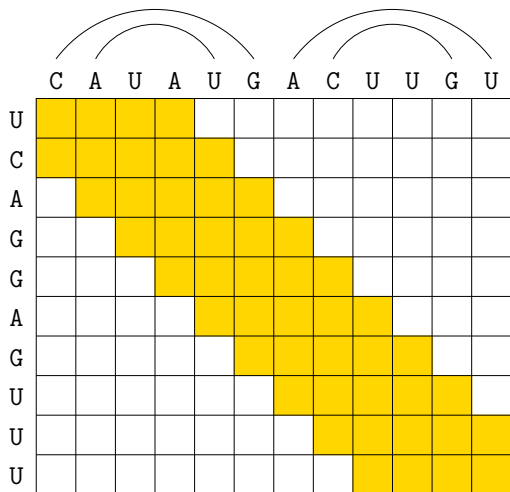


## Erpin

Contrainte sur les tiges :  
diagonales obligatoires

Contraintes sur les boucles :  
bandes verticales

# Alignement structure/séquence



Les meilleurs alignements restent autour de la diagonale

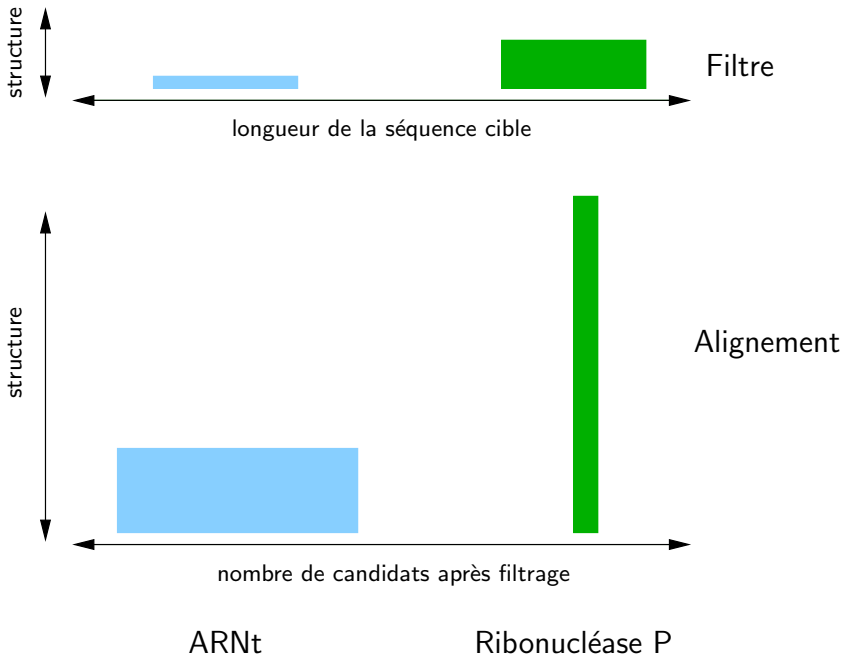
Paramètre  $p$

Temps :  $O(np^3)$

## Performances filtrage+alignement

Séquence aléatoire 1Mb

	filtre	alt	temps	FastR	Rsearch
tRNA 78 nt, 4 tiges	5040	3	35s	52s	3411s
RNAse P – <i>P. furiosus</i> 330 nt, 16 tiges	9	0	39s		43h



# Conclusion

- ▶ **Module de filtrage**
  - ▶ temps peu sensible à la taille de la structure
  - ▶ structure est une signature forte
  - ▶ autorise les pseudo-noeuds
  - ▶ peut être combiné des programmes spécifiques à un type d'ARN (tRNAscan, )
- ▶ **Module d'alignement**
  - ▶ effort algorithmique
  - ▶ intégration à venir de bons systèmes de score (profil, Ribosum)
- ▶ **Choix des paramètres, difficulté et richesse**
- ▶ **Prototype syranga**

