

## Application en bio-informatique: *gènes à ARN et gènes à protéines*

Arnaud Fontaine, Hélène Touzet  
{fontaina,touzet}@lifl.fr

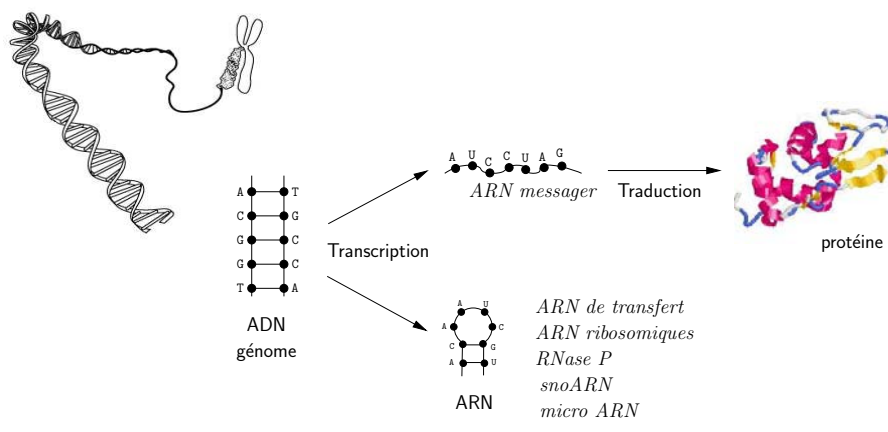
LIFL – UMR USTL-CNRS 8022 – INRIA Sequoia

Journées Grid5000 à Lille  
30-31 octobre 2006



1 / 18

## Les gènes dans la cellule



- ▶ Deux types de gènes
  - ▶ gènes à protéines
  - ▶ gènes à ARN

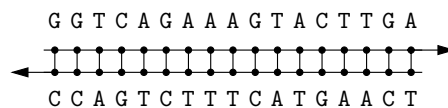
2 / 18

## Gène à protéine

- ▶ **Protéine** : séquence d'acides aminés
- ▶ **Séquence codante d'ADN**  
1 triplet de nucléotides → 1 acide aminé
- ▶ **Pour une séquence d'ADN**
  - ▶ 6 séquences d'acides aminés possibles *a priori*
  - ▶ 1 seule séquence d'acides aminés correcte

3 / 18

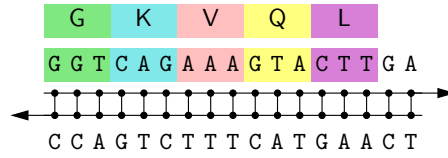
## Gène à protéine



4 / 18

## Gène à protéine

Cadre 1

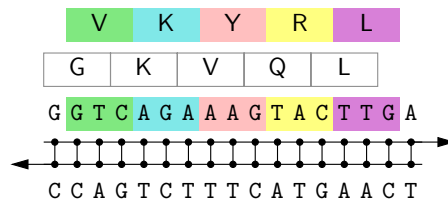


4 / 18

## Gène à protéine

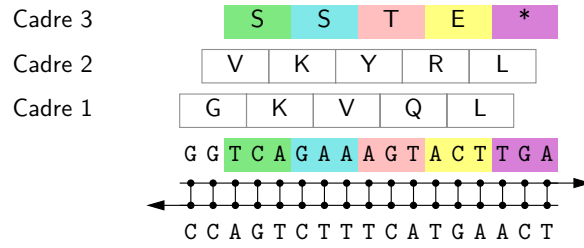
Cadre 2

Cadre 1



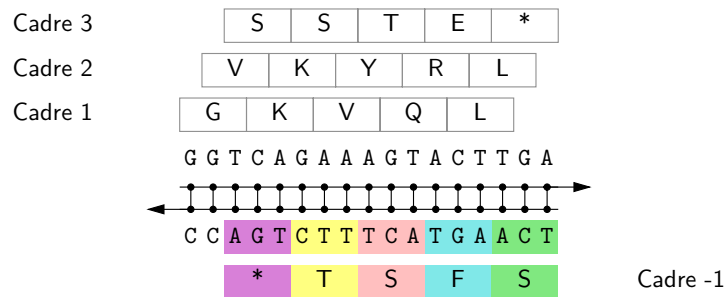
4 / 18

## Gène à protéine



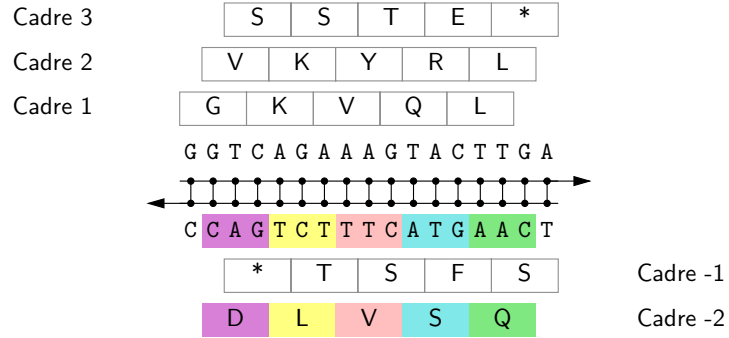
4 / 18

## Gène à protéine



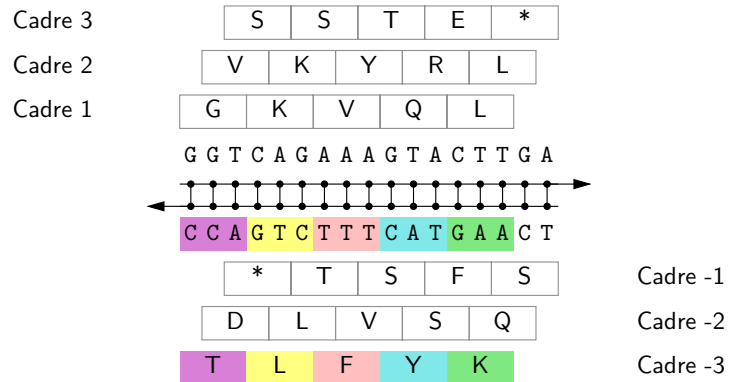
4 / 18

## Gène à protéine



4 / 18

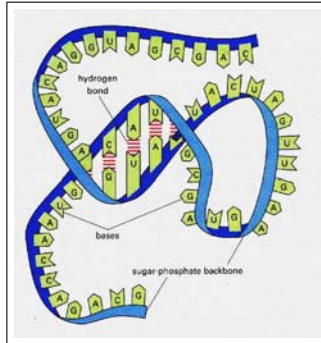
## Gène à protéine



4 / 18

## Gène à ARN

- ▶ **Molécule d'ARN** : acide nucléique monobrin
- ▶ **Structure** : ensemble d'appariements entre nucléotides
  - ▶ A – U
  - ▶ G – C
- ▶ **Repliement dans une conformation stable**
  - ▶ minimisation de l'**énergie libre**



5 / 18

## Prédiction de gènes

- ▶ **Problème** : identifier où sont localisés les gènes dans un génome
  - ▶ génome de levure : 12 millions de nucl., 6500 gènes, 73% du génome
  - ▶ génome humain : 3 milliards de nucl., ~30000 gènes, ~5% du génome
- ▶ **Nombreux projets de séquençage**
  - ▶ 448 génomes déjà disponibles
  - ▶ 1692 génomes en cours de séquençage
- ▶ **Génomique comparative**
  - ▶ chercher les gènes communs à plusieurs espèces
  - ▶ exploiter les informations évolutives entre les espèces

6 / 18

## Motifs de substitutions

- Informations évolutives : **motifs de substitutions** entre séquences

```
GGUCAGAAAGUACUU      UUGUUCGAAAGAACG
| | | | | | | | | |   | | | | | | | | | |
GGACAGAAAGGUUCUC      UUGACC GAAAGGUCG
```

7 / 18

## Motifs de substitutions

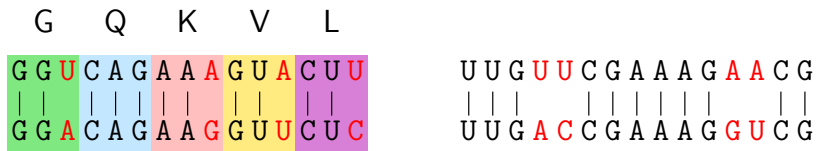
- Informations évolutives : **motifs de substitutions** entre séquences

```
GGUCAGAAAGUACUU      UUGUUCGAAAGAACG
| | | | | | | | | |   | | | | | | | | | |
GGACAGAAAGGUUCUC      UUGACC GAAAGGUCG
```

7 / 18

## Motifs de substitutions

- Informations évolutives : **motifs de substitutions** entre séquences

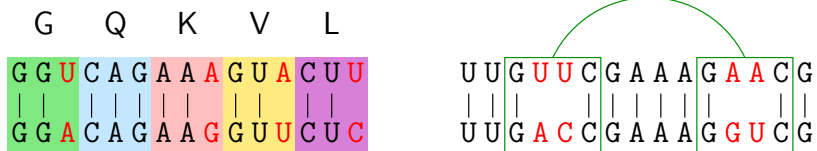


### gène à protéine

les mutations préservent la protéine codée

## Motifs de substitutions

- Informations évolutives : **motifs de substitutions** entre séquences



### gène à protéine

les mutations préservent la protéine codée

### gène à ARN

les mutations préservent la structure de l'ARN

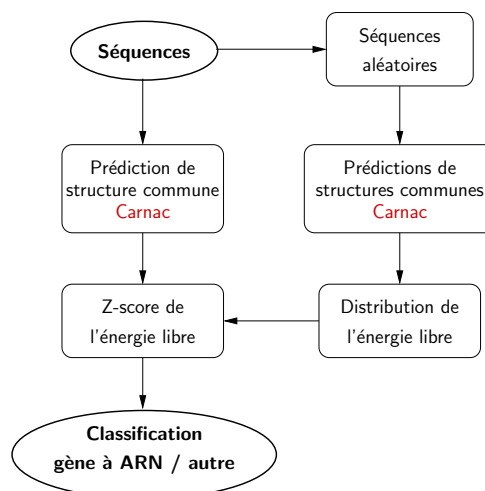
## Protea : prédiction de gènes à protéines

- ▶ Idée : les gènes à protéines communs partagent un cadre de lecture privilégié
- ▶ Algorithme :
  1. Comparaison de tous les cadres de lecture 2 à 2
  2. Construction d'un graphe des cadres de lecture
  3. Calcul statistique de la qualité du graphe : cohérence des cadres de lecture
  4. Classification gène à protéine/autre

8 / 18

## Carnac/Arnica : prédiction de gènes à ARN

- ▶ Idée : préservation d'une structure commune stable
- ▶ Carnac
  - ▶ modèle thermodynamique
  - ▶ modèle évolutif



9 / 18

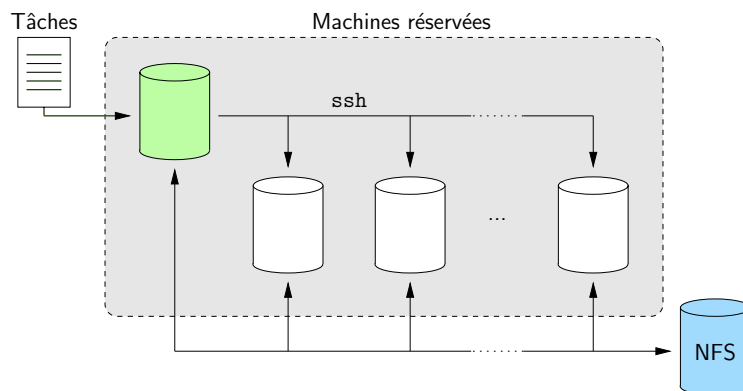
## Expériences sur Grid5000

- ▶ **Besoin d'une puissance de calcul**
  - ▶ ajuster les modèles statistiques
  - ▶ évaluer les performances des programmes "à grande échelle"
- ▶ **Expériences menées**
  - ▶ mise au point de Protea et de Carnac/Arnica (100 jours CPU)
  - ▶ évaluation de Protea, de Carnac/Arnica et de QRNA (500 jours CPU) [Eddy 1991]
  - ▶ construction d'un jeu de données d'apprentissage (300 jours CPU)
- ▶ **Bilan d'utilisation**
  - ▶ temps de réservation CPU  $\approx$  950 jours
  - ▶ espace disque utilisé  $\approx$  7 Go (compressés)

10 / 18

## Profil des expériences

- ▶ Programmes non parallélisés : **tâches indépendantes**
- ▶ Utilisation de l'image fournie par défaut : **pas de déploiement**



11 / 18

## Evaluation à grande échelle

- ▶ **But** : comparer les performances de Protea et de Carnac/Arnica avec QRNA [Eddy 1991]
- ▶ **QRNA** : prédictions de gènes à protéines et de gène à ARN
  - ▶ programme de référence, le plus utilisé
  - ▶ modèles à base de grammaires stochastiques
  - ▶ limité à 2 séquences
- ▶ **Deux types de données pour l'évaluation**
  - ▶ gènes à protéines :  
prédictions de Protea, contre-prédictions de Carnac/Arnica
  - ▶ gènes à ARN :  
prédictions de Carnac/Arnica, contre-prédictions de Protea

12 / 18

## Données de l'évaluation

- ▶ **Gènes à protéines** : Pandit database
  - ▶ **7738** familles de gènes à protéines
  - ▶ **jusqu'à 2400 séquences** par famille
- ▶ **Gènes à ARN** : RFAM database
  - ▶ **505** familles de gènes à ARN (toutes les familles connues)
  - ▶ **jusqu'à 12000 séquences** par famille
- ▶ **Constitution aléatoire de sous-familles de 2 à 12 séquences**

13 / 18

## Bilan de l'évaluation

- ▶ Résultats sur les sous-familles de 2 séquences

	Protea	Carnac/Arnica	QRNA
Pandit	<b>92.6 %</b>	4.1 %	<b>65.1 %</b> (protéine)
RFAM	5.4 %	<b>86.2 %</b>	<b>38.0 %</b> (ARN)

- ▶ Evaluation précise des performances selon
  - ▶ la distance évolutive des séquences
  - ▶ la taille du jeu de données (nombre d'organismes)
  - ▶ la taille des séquences
- ▶ Problème : coût du calcul de la distribution de l'énergie libre
  - ▶ tentative de contournement par l'apprentissage (SVM)
  - ▶ construction d'un premier jeu de données (2000 séquences)
  - ▶ calculs coûteux de nombreux attributs supplémentaires

14 / 18

## Premier test à grande échelle

- ▶ UCSC genome browser <http://genome.ucsc.edu>
  - ▶ 13 génomes en ligne dont celui de l'Homme
  - ▶ regroupement de données associées
  - ▶ séquences conservées entre génomes
- ▶ Test à grande échelle de Protea
  - ▶ séquences conservées entre 17 génomes
  - ▶ **103712** familles d'au moins 12 séquences
- ▶ 6.1% de prédictions positives avec Protea
  - ▶ 29% de gènes à protéines connus
  - ▶ 62% de prédictions déjà réalisées par d'autres programmes
  - ▶ **9% (517 familles) de nouvelles prédictions**

16 / 18

## Retour d'expériences



- ▶ Simplicité d'utilisation
- ▶ Existence d'un image "par défaut" : peu de développements spécifiques nécessaires
- ▶ Monitoring des tâches (Ganglia)
- ▶ Administrateur disponible

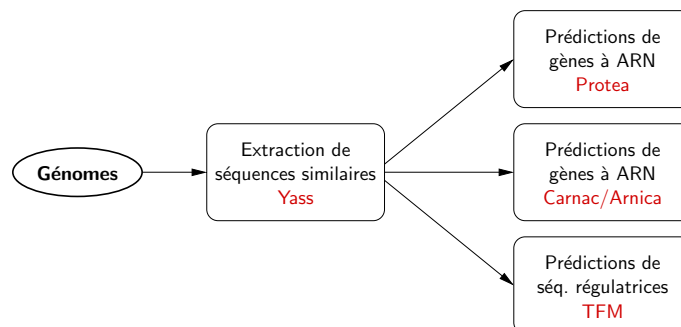


- ▶ Serveur NFS fortement sollicité : échecs de lecture/écriture
- ▶ Système de réservation (OAR)
  - ▶ bien estimer la durée de l'expérience
  - ▶ réservations "statiques"
- ▶ Déploiement d'une image : coût de développement important

17 / 18

## Perspectives d'utilisation de Grid5000

- ▶ Pipeline d'annotation automatique de génomes



- ▶ Calcul sur tout ou partie des génomes disponibles
- ▶ Base de données des résultats
- ▶ Développement d'un service web d'interrogation

18 / 18