Pygram: une nouvelle méthode de visualisation des séquences répétées dans les génomes.

P. Durand, F. Mahé, M. Giraud, A.-S. Valin et J. Nicolas

Projet *Symbiose*, IRISA-INRIA Rennes

Objectif : décomposition des séquences génomiques en ‘domaines’

Protéines

Les ‘domaines/motifs’ fonctionnels et structuraux

Comment ?
- alignement de séquences,
- superposition de structures,
- découverte de motifs

Génomes

Les ‘annotations’ fonctionnelles

Comment ?
- criblage de banques
- algorithmes dédiés

Approche orientée par la biologie… peut-on envisager une décomposition sans a priori biologique ?

Décomposition en ‘domaines’: décomposition en répétitions

Génomes (archées, bactéries, eucaryotes)

- caractérisés par la présence de (nombreuses) répétitions:
  - répétitions ‘simples’ de *k-mers* (tandem),
  - segments (gènes, chromosomes)
  - transposons.

Répétitions formelles ?

- *répétitions en tandem* (Wexler, 2004),
- *plus longues répétitions* (Karp, 1972),
- *répétitions maximales* (Gusfield, 1997)
- *plus longues répétitions avec un bloc quelconque* (Crochemore, 2004),

Décomposition en ‘domaines’: décomposition en répétitions

Génomes (archées, bactéries, eucaryotes)

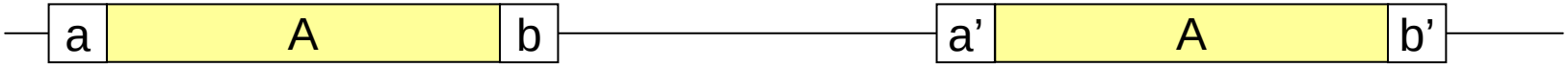
- caractérisés par la présence de (nombreuses) répétitions:
  - répétitions ‘simples’ de *k-mers* (tandem),
  - segments (gènes, chromosomes)
  - transposons.

Répétitions formelles ?

- *répétitions en tandem* (Wexler, 2004),
- *plus longues répétitions* (Karp, 1972),
- *répétitions maximales* (Gusfield, 1997)
- *plus longues répétitions avec un bloc quelconque* (Crochemore, 2004),

Décomposition en ‘domaines’: répétitions maximales

Définition: plus longues sous-séquences répétées sans possibilité d’extension gauche/droite



Intérêts:

- pas d’a priori biologique… et pourtant !
- bloc de base pour la construction de répétions avec erreurs, avec insertions (*MUMmer, Reputer*).
- recherche: linéaire en fonction de la taille de la séquence,
- nombre: borné par la taille de la séquence

Question: que peut-on obtenir avec des répétitions maximales ?

Décomposition en ‘domaines’: analyse des répétitions maximales exactes (eMR)

Principe :

1. Recherche des eMR au moyen d’un arbre des suffixes généralisés (Gusfield, 1997).

	Genome	*Genome size* (Mb)	nb. eMR	Occurrences	eMR max size
|eMR|>=1	E.coli	4.63	2,491,152	50,604,663	2,815
|eMR|>=20			2045	9874	

Comment les eMR sont-elles disposées le long du génome ?

2. Indexation des répétitions

3. Visualisation cartographique adaptée

Interprétation des répétitions maximales

Techniques de visualisation:

- dotplot (Gibbs, 1970),
- sequence landscape (Clift, 1986),
- chaos game (Jeffrey, 1999),
- Percent Identity Plot, PIP (Schwartz, 2000)
- repeat-graph (Kurtz, 2001)
- BARD (Spell, 2003)

Inconvénients:
- visualisation des paires,
- difficulté de travailler sur des séquences génomiques (zoom),
- pas de visualisation de l'organisation des eMR

Interprétation des répétitions maximales

Techniques de visualisation:

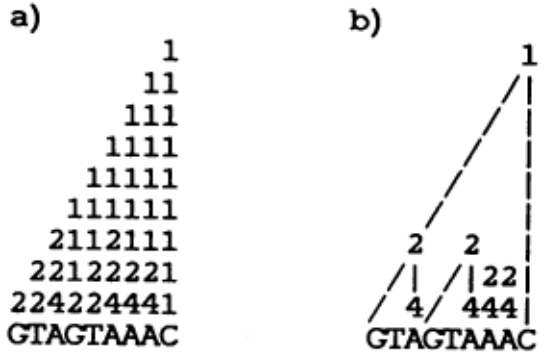
- dotplot (Gibbs, 1970),
- sequence landscape (Clift, 1986),
- chaos game (Jeffrey, 1999),
- Percent Identity Plot, PIP (Schwartz, 2000)
- repeat-graph (Kurtz, 2001)
- BARD (Spell, 2003)

Inconvénients:
- visualisation des paires,
- difficulté de travailler sur des séquences génomiques (zoom),
- pas de visualisation de l'organisation des eMR

Interprétation des répétitions maximales

Sequence Landscape

séquence → DAWG →



Modifications:

- palette de couleur sur les eMR,
- séparation taille des mots, fréquence des mots,
- double affichage: brins N et RC,
- annotation de mots particuliers,
- zoom multiples: lentilles contextuelles,
- affichage linéaire ou logarithmique.

↓

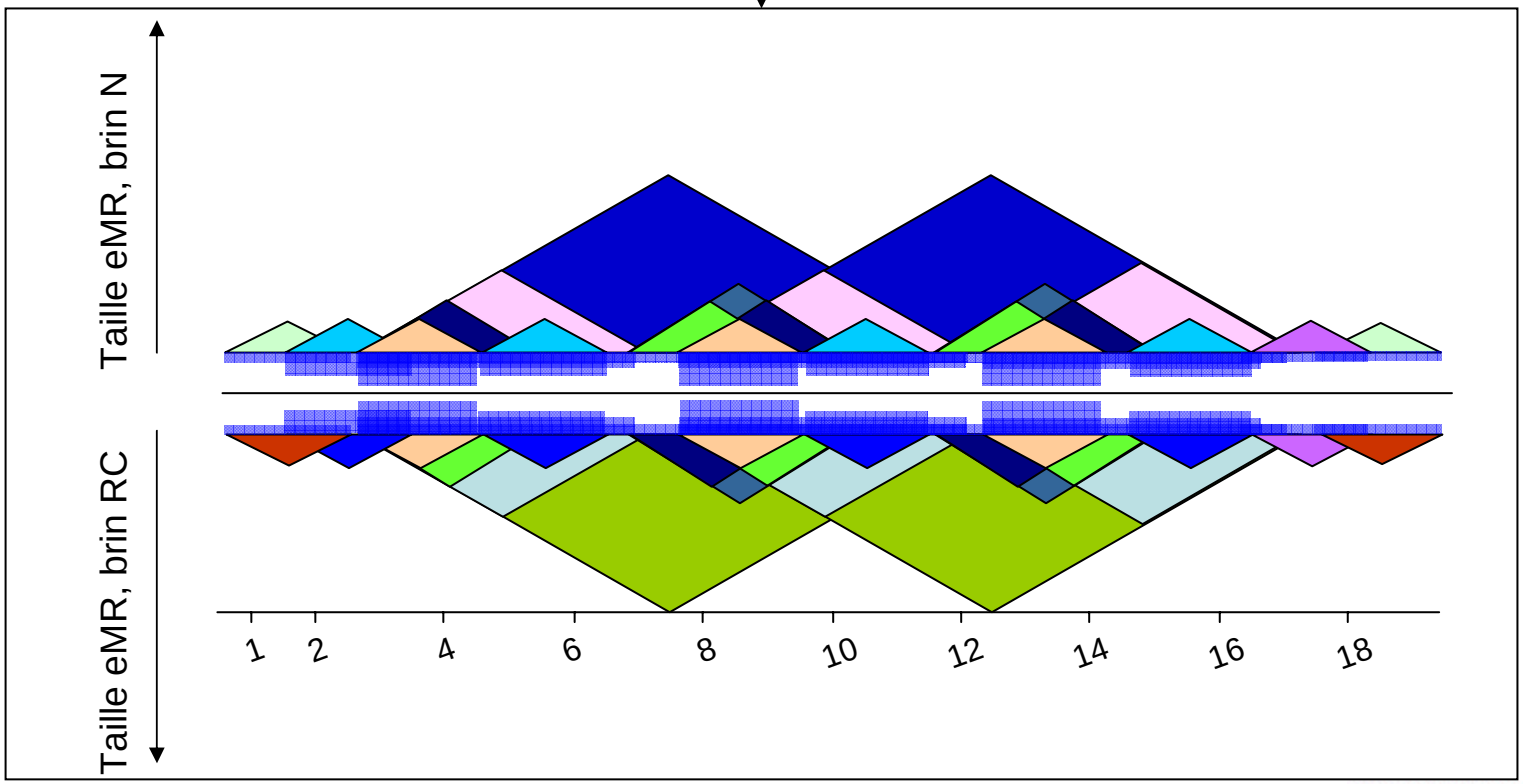
Pygram (pyramid diagram)

Pygram: principe

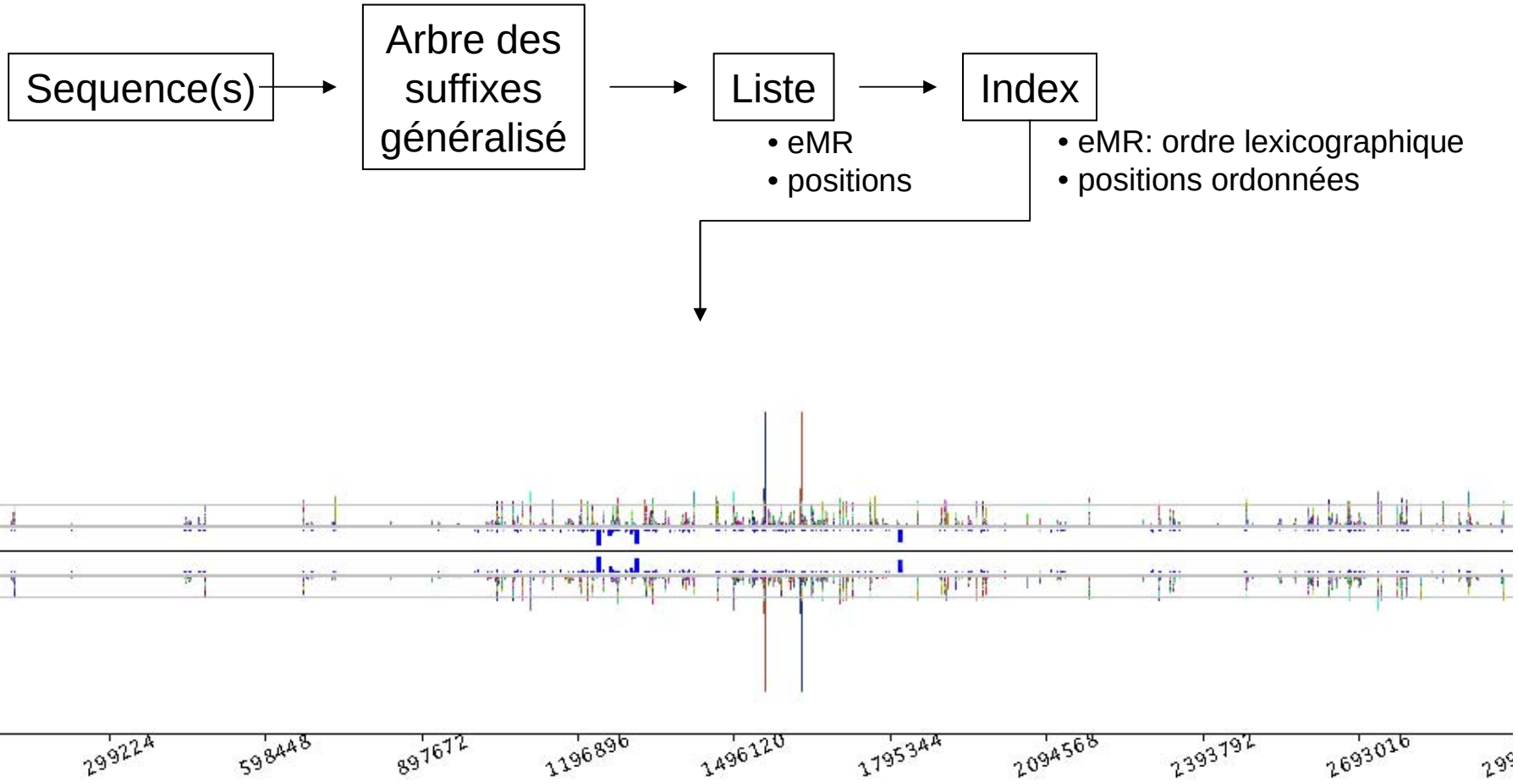
Sequence(s) → Arbre des suffixes généralisé → Liste → Index

Liste
- eMR
- positions

Index
- eMR: ordre lexicographique
- positions ordonnées

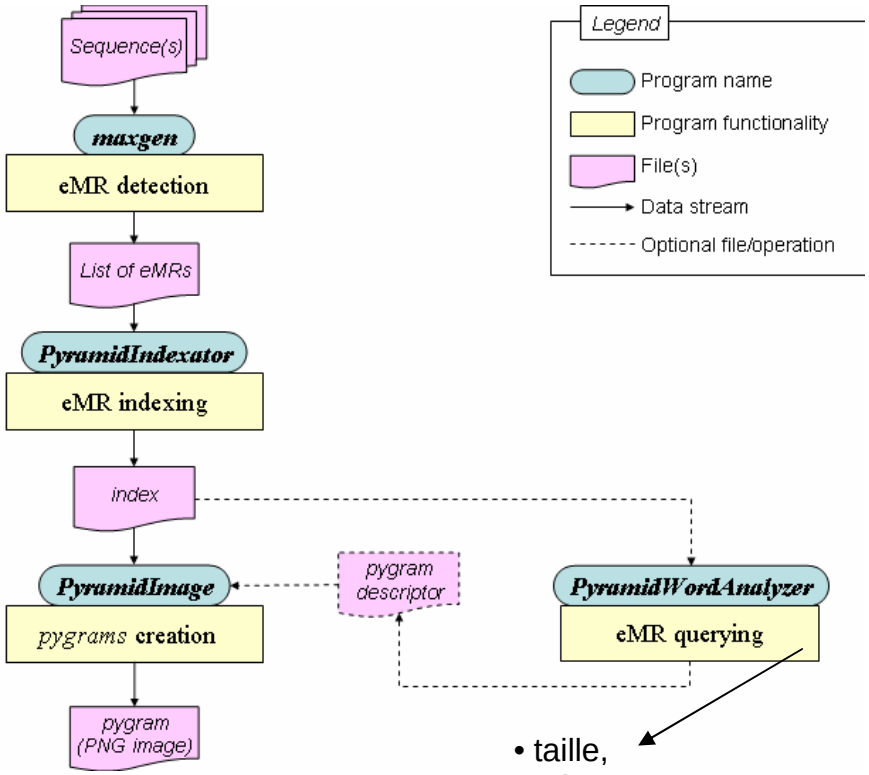


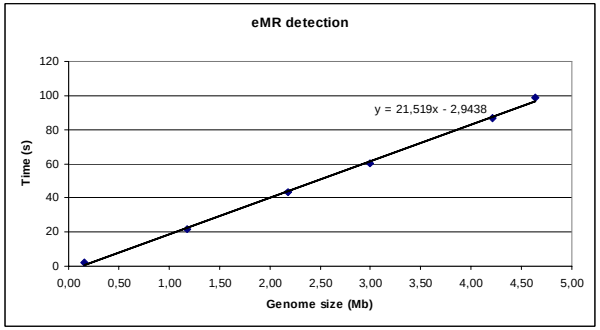
Pygram: principe

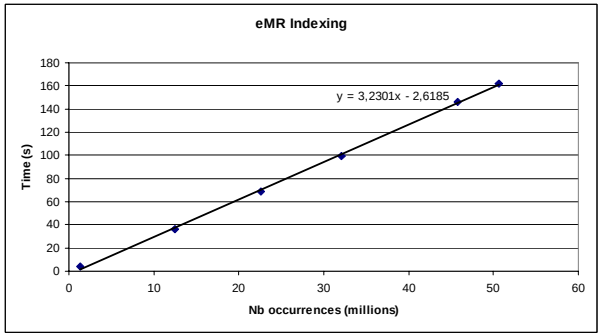


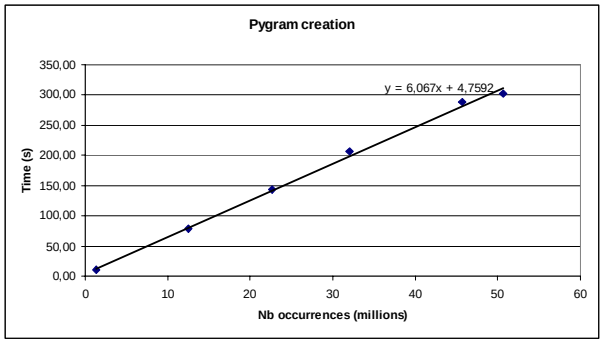
Sulfolobus solfataricus P2 (2,9 Mb)

Pygram: implémentation









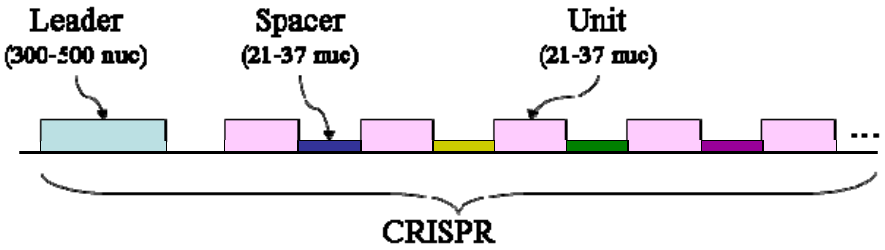
- taille,
- fréquence,
- localisation (position, brin, séquence)

Pygram: application (1/2)

Les structures répétées chez les archées

Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR; Jansen, 2002*)*

- Structure:



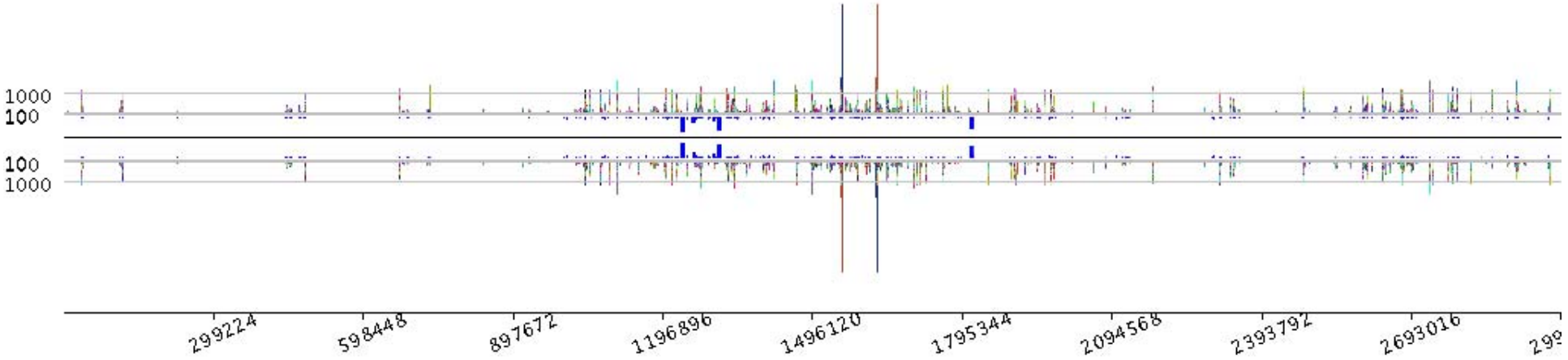
- Observées chez les archées et les bactéries
- Rôles supposés:
  - sites de fixation de protéines de type histone,
  - régulation,
  - ségrégation des chromosomes lors de la division cellulaire,
  - vecteurs dans les transferts de gènes,
  - ‘mémoire’ immunitaire
- Repérage classique: BLAST génome contre génome !

Principe de l’étude par *pygram*:
- recherche des eMR de taille >= 20,
- affichage et exploration du *pygram*.

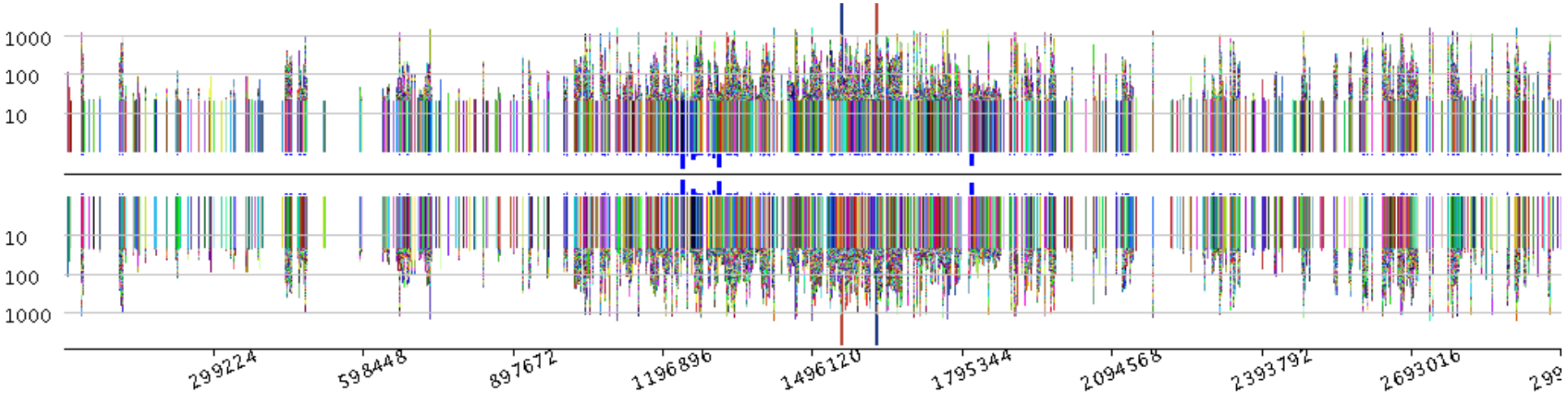
Martériel biologique: 20 génomes complets d’archées

Pygram: recherche des CRISPR de *Sulfolobus solfataricus P2 (2,9 Mb)*

Génome complet, vue 'standard' …

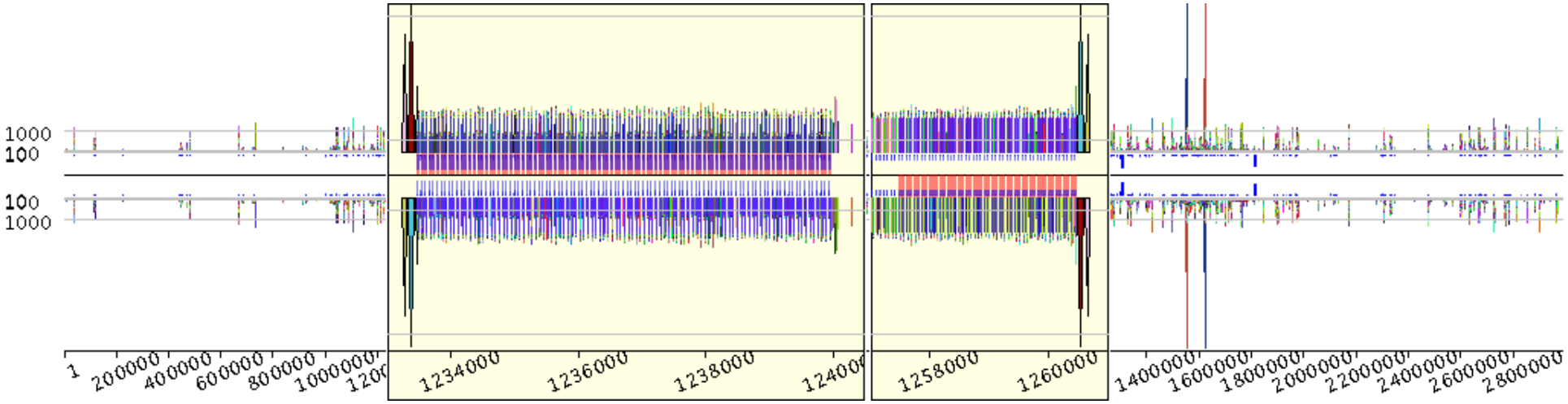


… vue 'logarithmique'

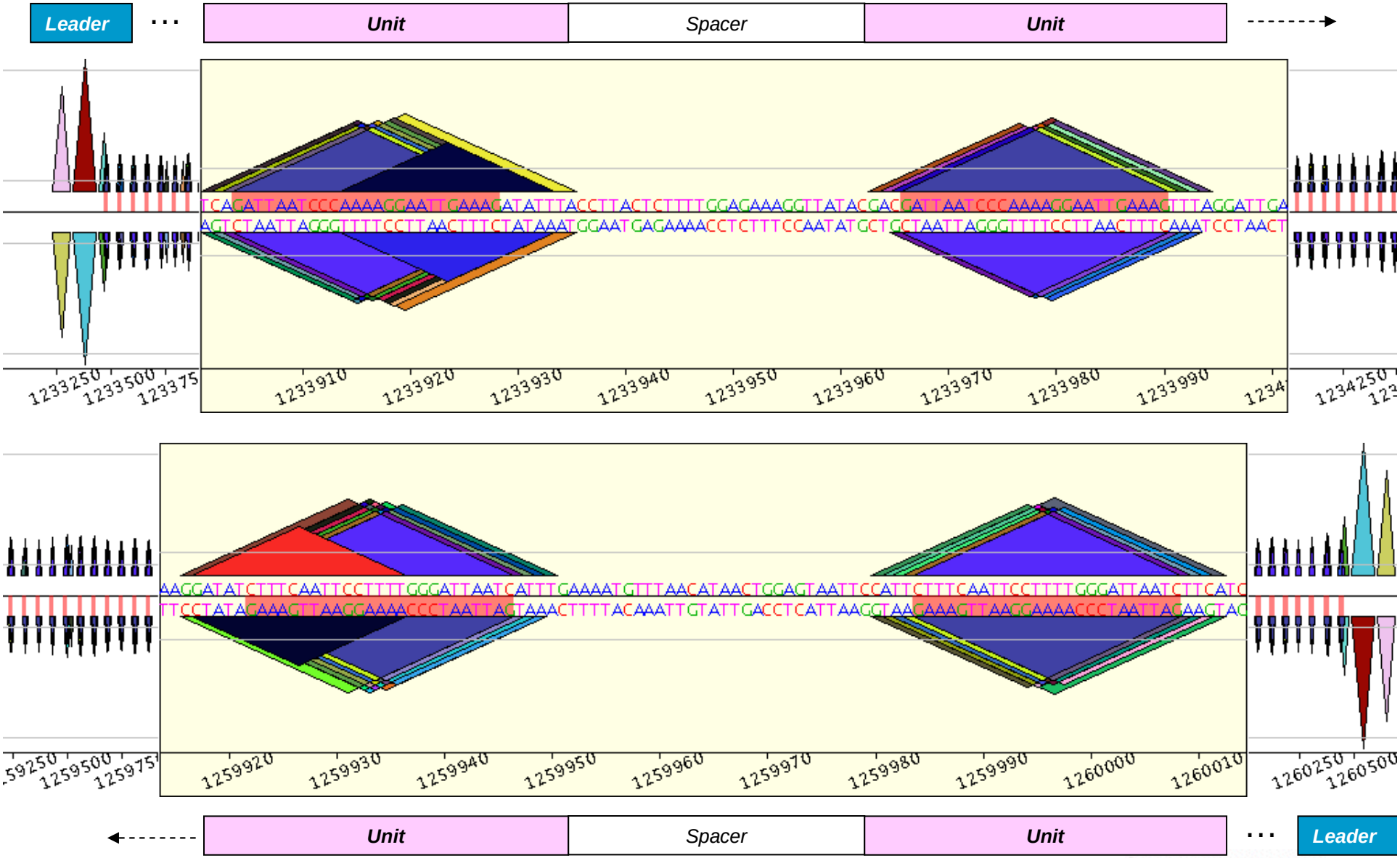


Pygram: recherche des CRISPR de *Sulfolobus solfataricus P2*

Génome complet (2,9 Mb), vue avec 2 lentilles



Pygram: recherche des CRISPR de *Sulfolobus solfataricus P2*



Pygram: application (2/2) Intégration du virus *SIRV1* dans le génome de *S. solfataricus*

Intégration et immunité…

S. Solfataricus contient des sous-séquences de SIRV1 (Mojica, 2005)

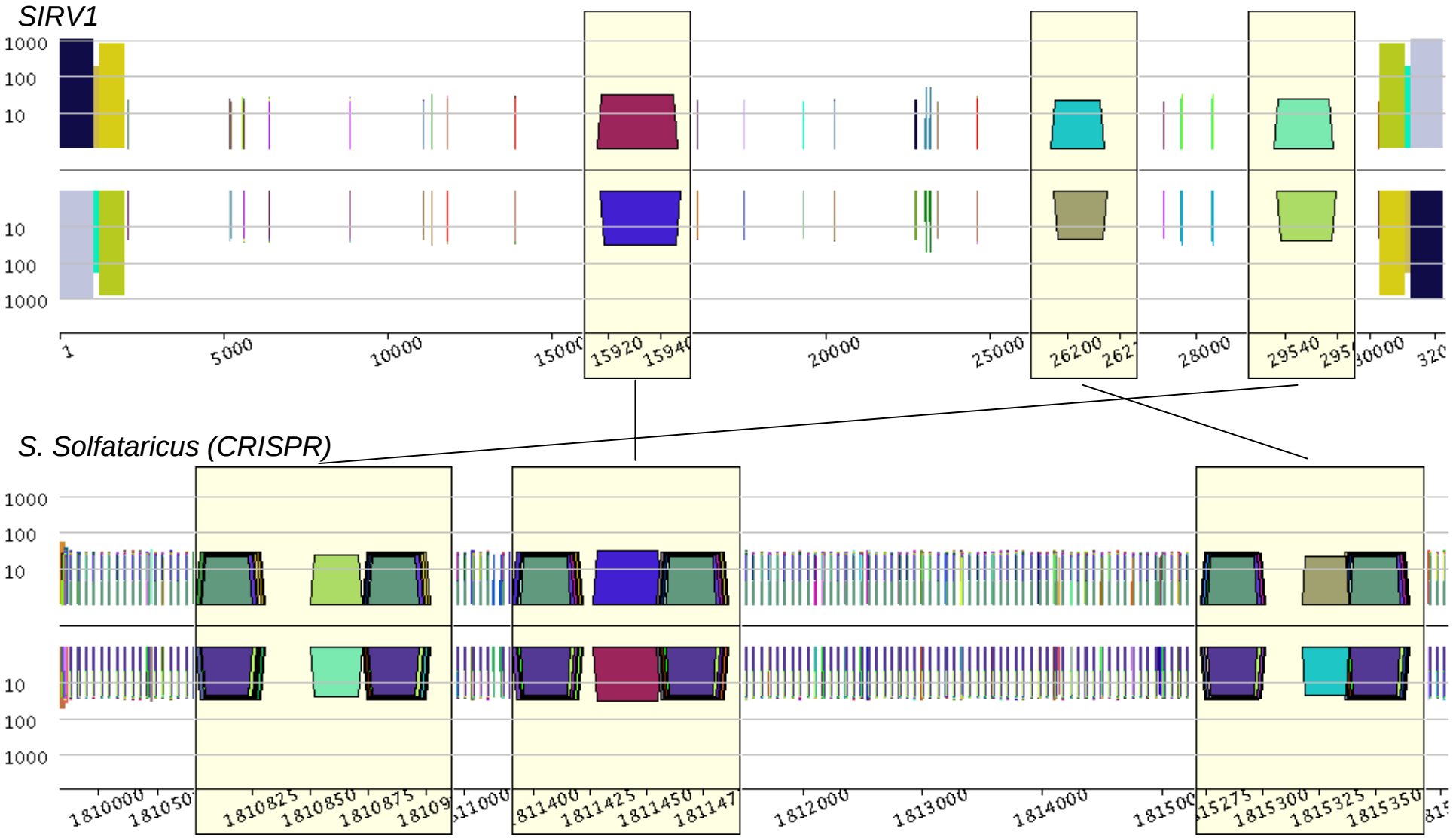
L'archée est résistante au virus !

Mais: l'archée est sensible au virus sans ces sous-séquences !

Principe de l'étude:

- recherche des eMR de taille >= 20 communes aux 2 séquences,
- affichage et exploration des 2 pygrams créés.

Pygram: recherche de *SIRV1* dans *S. solfataricus*



Conclusion / perspectives

Pygram:
- méthode de visualisation adaptée à l'analyse et l'identification des structures de répétitions (Haft, 2005),
- méthode d'intérêt pour la génomique comparative,
- nécessite des améliorations visuels (interactivité, *browser),*
- possibilité de l'utiliser avec d'autres types de répétitions.

eMR:
- brique de base pour la caractérisation de domaines génomiques (détéction automatique des CRISPRs).
- nécessite des améliorations algorithmiques (*répétition maximale flexible).*

Merci …

… questions ?

Référence:

Durand P, Mahé F, Valin A-S, Nicolas J (2005).
Pyramid diagram: visualizing the organization of repetitive sequences in genome.
Rapport de recherche INRIA no. 5798. (`www.inria.fr/rrrt/rr-5798.html`)