Comparaison de séquences répétées en tandem et application à la génétique

Sèverine Bérard

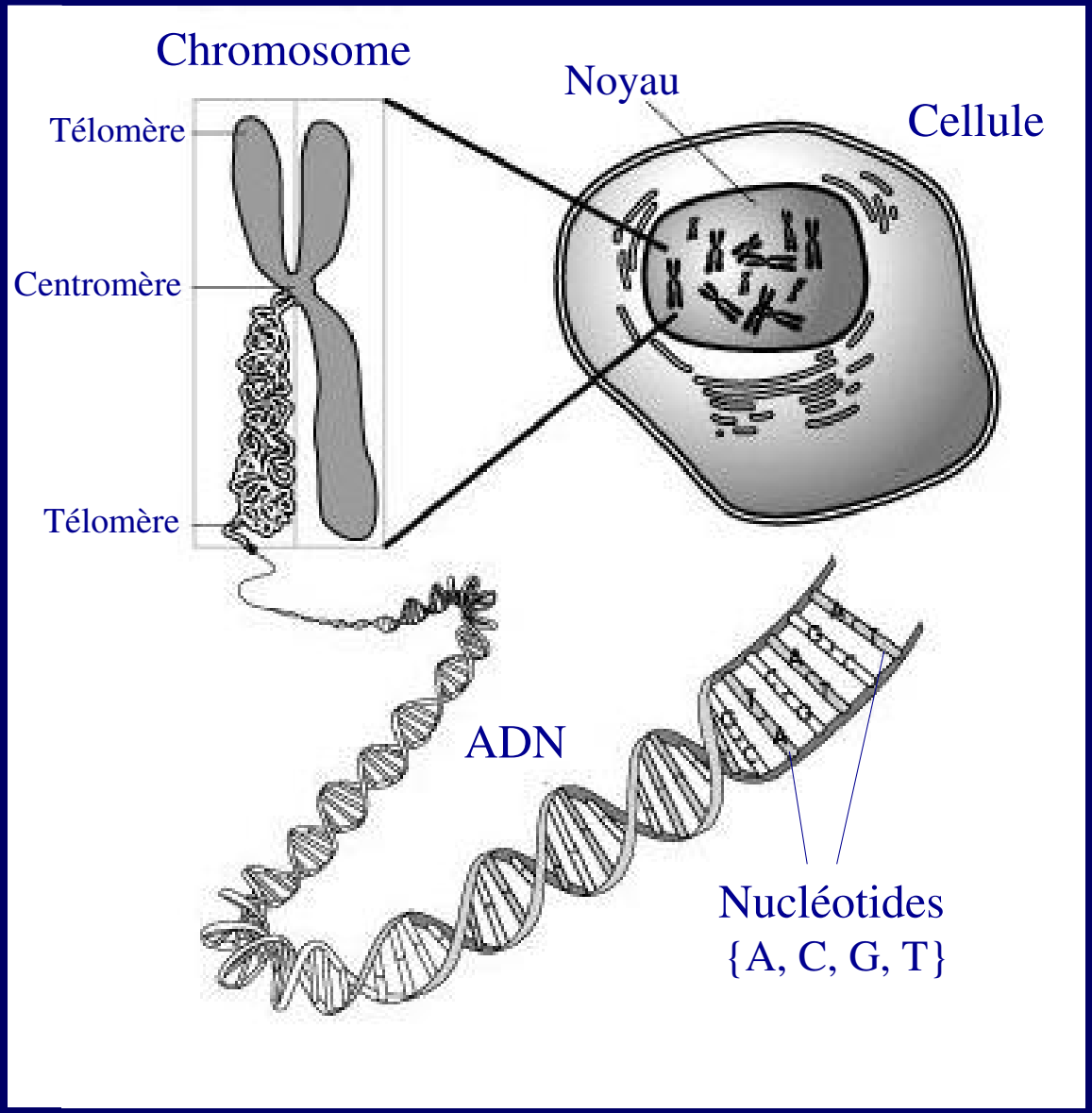
Le 5 décembre 2003

Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier - Université Montpellier II

Plan

- Introduction
- Comparaison de cartes de minisatellite par alignement
- Algorithmique
- Applications au minisatellite MSY1
- Conclusion et perspectives

Codage de l'information génétique



- Séquences d'ADN : chaînes sur l'alphabet $\Sigma = \{A, C, G, T\}$.

Vocabulaire

- Locus : localisation physique sur le chromosome ;
- Allèle ;
- Polymorphisme = variabilité,
  présence de différents allèles à un locus ;
- Comparaison : mesure de similarité ;
- Séquences répétées : dispersées ou en tandem.

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  $\rightarrow$ Événements mutationnels classiques : Substitution, Insertion, Délétion,

  $\rightarrow$ Événements mutationnels spécifiques : Amplification et Contraction ;

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  $\rightarrow$ Événements mutationnels classiques : Substitution, Insertion, Délétion,

  $\rightarrow$ Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  $\rightarrow$ Événements mutationnels classiques : Substitution, Insertion, Délétion,

  $\rightarrow$ Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  $\rightarrow$ Événements mutationnels classiques : Substitution, Insertion, Délétion,

  $\rightarrow$ Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CGG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CGG CGG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CGG CGG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CGG CAG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  $\rightarrow$ Événements mutationnels classiques : Substitution, Insertion, Délétion,

  $\rightarrow$ Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CGG CGG CAG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CAG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  $\rightarrow$ Événements mutationnels classiques : Substitution, Insertion, Délétion,

  $\rightarrow$ Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CAG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CAG CAG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CAG CAG CAG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CGG CGG CAG CAG CAG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CG CGG CAG CAG CAG CGG

Les séquences répétées en tandem

- Séquences d'ADN avec une évolution particulière :

  → Événements mutationnels classiques : Substitution, Insertion, Délétion,

  → Événements mutationnels spécifiques : Amplification et Contraction ;

  CAG CG CGG CAG CAG CAG CGG

- Polymorphisme : longueur et séquence.

Les minisatellites (ms)

- Répétitions en tandem : satellites, minisatellites, microsatellites ;
- Exemple de répétitions en tandem :
  CGGCGAT CGGCGAC CGGAGAT CGGCGAT CGGCGAT CGGAGAT CGACGAT
- La taille des motifs de 7 à 100 pb, la longueur entre 1 et 30 kb ;
- Certains minisatellites sont très polymorphes ;
- Les minisatellites haploïdes : évolution simplifiée, pas de recombinaison.

- Impliqués dans diverses pathologies : diabètes, plusieurs cancers, épilepsie, et autres [Buard, Jeffreys 97] ;

Intérêt pour les minisatellites

- Impliqués dans diverses pathologies : diabètes, plusieurs cancers, épilepsie, et autres [Buard, Jeffreys 97] ;

- Étude de population intra-spécifique :

  $\rightarrow$ migration des populations humaines : hypothèse Out-of-Africa [Armour et al 96],

  $\rightarrow$ évolution du chromosome Y [Jobling et Tyler-Smith 00] ;

Intérêt pour les minisatellites

- Impliqués dans diverses pathologies : diabètes, plusieurs cancers, épilepsie, et autres [Buard, Jeffreys 97] ;

- Étude de population intra-spécifique :

  $\rightarrow$ migration des populations humaines : hypothèse Out-of-Africa [Armour et al 96],

  $\rightarrow$ évolution du chromosome Y [Jobling et Tyler-Smith 00] ;

- Polymorphisme $\Rightarrow$ identification d'individu ou d'espèce (médecine légale, marqueurs génétiques, identification de souches de bactéries) ;

Cartes de minisatellites

- Une méthode spécifique de séquençage : Minisatellite Variant Repeat PCR [Jeffreys et al. 91].
  MVR-PCR fournit une carte de ms : une séquence de symboles, où chaque symbole représente un variant différent.

- Exemple de carte :

  → s = CGGCGAT CGGCGAC CGGAGAT CGGCGAT CGGCGAT CGGAGAT CGACGAT

  → Nouvel alphabet : a = CGGCGAT b = CGGCGAC c = CGGAGAT d = CGACGAT

  → Carte correspondante : a b c a a c d

MSY1 : ms humain du chromosome Y

- Longueur de l'unité répétée $= 25$ pb, 7 variants identifiés qui diffèrent à 3 positions seulement ;
- Évolution simplifiée $\Rightarrow$ pas d'échanges entre allèles ;

$m19$ Anglais 2

$m121$ Finlandais 16

$m707$ Maya 18

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
● Type 1 :	C	A	C	A	A	T	A	T	A	C	A	T	G	A	T	G	T	A	T	A	T	T	A	T	A
● Type 2 :	.	.	T	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.
● Type 3 :	.	.	C	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.
● Type 4 :	.	.	T	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.

- Introduction

- Comparaison de cartes de minisatellite par alignement
  - → Alignement classique
  - → Modèle adapté
  - → Travaux apparentés

- Algorithmique

- Applications au minisatellite MSY1

- Conclusion et perspectives

Alignement classique

- Pour comparer des séquences on construit un alignement ;

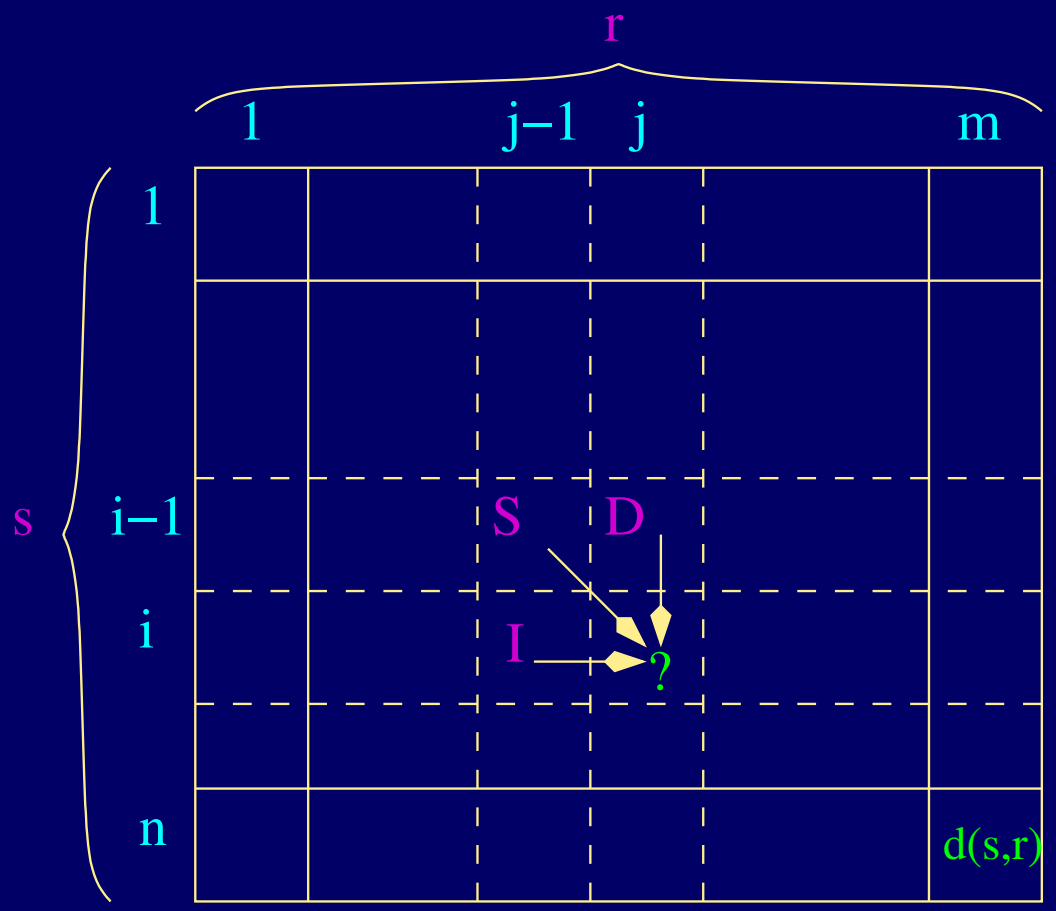
```
s :  A  G  G  T  C  A
     |     |     |  |
r :  A  -  G  C  C  A
```

- Le modèle évolutif comprend 3 opérations : Insertion (I), Délétion (D), Substitution (S) ;

- Coût d'un alignement = somme des coûts des opérations qui le composent ;

- Distance entre 2 séquences = coût minimum d'alignement.

- Deux séquences s et r de longueurs respectives n et m, $s[1..i]$ = préfixe de s de longueur i ;
- Principe : construire progressivement une matrice dans laquelle chaque case (i, j) contient la distance entre $s[1..i]$ et $r[1..j]$.



- Modèle symétrique et unitaire :

- Modèle symétrique et unitaire :

  $\rightarrow$ **Amplification(A)/Contraction(C)** duplique/supprime un caractère se trouvant à côté d'un caractère identique :

$$\text{a b c} \xrightarrow{\text{Amplification}} \text{a b b c} \xrightarrow{\text{Contraction}} \text{a b c}$$

- Modèle symétrique et unitaire :

  $\rightarrow$ **Amplification(A)/Contraction(C)** duplique/supprime un caractère se trouvant à côté d'un caractère identique :

$$\text{a b c} \xrightarrow{\text{Amplification}} \text{a b b c} \xrightarrow{\text{Contraction}} \text{a b c}$$

  $\rightarrow$ **Mutation(M)** remplace un caractère par un autre :

$$\text{a b c} \xrightarrow{\text{Mutation}} \text{a d c}$$

- Modèle symétrique et unitaire :

  → **Amplification(A)/Contraction(C)** duplique/supprime un caractère se trouvant à côté d'un caractère identique :

  $$\text{a b c} \xrightarrow{\text{Amplification}} \text{a b b c} \xrightarrow{\text{Contraction}} \text{a b c}$$

  → **Mutation(M)** remplace un caractère par un autre :

  $$\text{a b c} \xrightarrow{\text{Mutation}} \text{a d c}$$

  → **Insertion(I)/Délétion(D)** insère/supprime un caractère mais sans contrainte :

  $$\text{a b c} \xrightarrow{\text{Insertion}} \text{a b c d} \xrightarrow{\text{Délétion}} \text{b c d}$$

- Le modèle est symétrique : $I = D$ et $A = C$;

- Le modèle est symétrique : $I = D$ et $A = C$;

- Coût de mutation (M) indépendant des variants ;

Les coûts

- Le modèle est symétrique : $I = D$ et $A = C$;
- Coût de mutation (M) indépendant des variants ;
- Dû à des observations biologiques : $A, C < M, D, I$;

Les coûts

- Le modèle est symétrique : $I = D$ et $A = C$;
- Coût de mutation (M) indépendant des variants ;
- Dû à des observations biologiques : $A, C < M, D, I$;
- Une délétion (D) peut également être obtenue par une mutation (M) suivie d'une contraction (C) ;
- Hypothèse : $D > M + C$ et $I > A + M$;

Les coûts

- Le modèle est symétrique : $I = D$ et $A = C$;
- Coût de mutation (M) indépendant des variants ;
- Dû à des observations biologiques : $A, C < M, D, I$;
- Une délétion (D) peut également être obtenue par une mutation (M) suivie d'une contraction (C) ;
- Hypothèse : $D > M + C$ et $I > A + M$;
- Théorème : Quels que soit les coûts des opérations, le coût minimum d'alignement est une distance.

Soit s et r deux séquences de longueurs n et m, trouver l'alignement global optimal entre s et r sous le modèle évolutif adapté.

Non commutativité

- L'ensemble d'opérations du modèle adapté n'est pas commutatif :

$$a \xrightarrow{\text{Amplification}} a\ a \xrightarrow{\text{Mutation}} b\ a$$

$$a \xrightarrow{\text{Mutation}} b \xrightarrow{\text{Amplification}} b\ b$$

Exemple d'évolution (1)

Individu 1

Événement	Séquence				
	a	a	a	a	a
	1	2	3	4	5

Exemple d'évolution (1)

Individu 1

Événement	Séquence				
	a	a	a	a	a
mutation	a	a	a	a	a
	1	2	3	4	5

Exemple d'évolution (1)

Individu 1

Événement	Séquence				
	a	a	a	a	a
mutation	a	e	a	a	a
	1	2	3	4	5

Exemple d'évolution (2)

Individu 2

Événement	Séquence
	a a a a a

1 2 3 4 5 6 7 8 9 10 11

Exemple d'évolution (2)

Individu 2

Événement	Séquence
	a a a a a
mutation	a a a a a

1 2 3 4 5 6 7 8 9 10 11

Exemple d'évolution (2)

Individu 2

Événement	Séquence
	a a a a a
mutation	a a a b a

1 2 3 4 5 6 7 8 9 10 11

Exemple d'évolution (2)

Individu 2

Événement	Séquence				
	a	a	a	a	a
mutation	a	a	a	b	a
5 amplifications	a	a	a	b	a

1 2 3 4 5 6 7 8 9 10 11

Exemple d'évolution (2)

Individu 2

Événement	Séquence
	a a a a a
mutation	a a a b a
5 amplifications	a a a b b b b b b a

1 2 3 4 5 6 7 8 9 10 11

Exemple d'évolution (2)

Individu 2

Événement	Séquence										
	a	a	a	a	a						
mutation	a	a	a	b	a						
5 amplifications	a	a	a	b	b	b	b	b	b	a	
mutation	a	a	a	b	b	b	b	b	b	a	
	1	2	3	4	5	6	7	8	9	10	11

Exemple d'évolution (2)

Individu 2

Événement	Séquence										
	a	a	a	a	a						
mutation	a	a	a	b	a						
5 amplifications	a	a	a	b	b	b	b	b	b	a	
mutation	a	a	a	b	b	c	b	b	b	a	

1 2 3 4 5 6 7 8 9 10 11

Exemple d'évolution (2)

Individu 2

Événement	Séquence										
	a	a	a	a	a						
mutation	a	a	a	b	a						
5 amplifications	a	a	a	b	b	b	b	b	b	a	
mutation	a	a	a	b	b	c	b	b	b	a	
mutation	a	a	a	b	b	c	b	b	b	a	
	1	2	3	4	5	6	7	8	9	10	11

Exemple d'évolution (2)

Individu 2

Événement	Séquence										
	a	a	a	a	a						
mutation	a	a	a	b	a						
5 amplifications	a	a	a	b	b	b	b	b	b	a	
mutation	a	a	a	b	b	c	b	b	b	a	
mutation	a	a	a	b	b	c	b	d	b	a	
	1	2	3	4	5	6	7	8	9	10	11

Exemple d'évolution (2)

Individu 2

Événement	Séquence										
	a	a	a	a	a						
mutation	a	a	a	b	a						
5 amplifications	a	a	a	b	b	b	b	b	b	a	
mutation	a	a	a	b	b	c	b	b	b	a	
mutation	a	a	a	b	b	c	b	d	b	a	
amplification	a	a	a	b	b	c	b	d	b	a	
	1	2	3	4	5	6	7	8	9	10	11

Exemple d'évolution (2)

Individu 2

Événement	Séquence										
	a	a	a	a	a						
mutation	a	a	a	b	a						
5 amplifications	a	a	a	b	b	b	b	b	b	a	
mutation	a	a	a	b	b	c	b	b	b	a	
mutation	a	a	a	b	b	c	b	d	b	a	
amplification	a	a	a	b	b	c	b	d	d	b	a
	1	2	3	4	5	6	7	8	9	10	11

Alignement souhaité

Alignement souhaité


```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  \  \  \  \  \  |
I2 :  a  a  a  b  b  b  b  b  d  b  a
```

| Appariement exact

[Mutation (M)

\ Amplification (A)

(Amplification+Mutation (A_M)

Alignement souhaité


```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  \  \  \  \  \  |
I2 :  a  a  a  b  b  b  b  b  d  b  a
```

| Appariement exact

[Mutation (M)

\ Amplification (A)

(Amplification+Mutation (A_M)

Alignement souhaité


```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  \  \  \  \  \  |
I2 :  a  a  a  b  b  b  b  b  d  b  a
```

| Appariement exact

[Mutation (M)

\ Amplification (A)

(Amplification+Mutation (A_M)

Alignement souhaité


```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  \  \  \  \  \  |
I2 :  a  a  a  b  b  b  b  b  d  b  a
```

|	Appariement exact
[	Mutation (M)
\	Amplification (A)
(	Amplification+Mutation (A_M)

Alignement souhaité


```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  \  \  \  \  \  |
I2 :  a  a  a  b  b  b  b  b  d  b  a
```

| Appariement exact

[Mutation (M)

\ Amplification (A)

(Amplification+Mutation (A_M)

Alignement souhaité


```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  \  \  \  \  \  |
I2 :  a  a  a  b  b  b  b  b  d  b  a
```

|	Appariement exact
[	Mutation (M)
\	Amplification (A)
(	Amplification+Mutation (A_M)

Alignement souhaité


```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  \  \  \  \  \  |
I2 :  a  a  a  b  b  b  b  b  d  b  a
```

| Appariement exact

[Mutation (M)

\ Amplification (A)

(Amplification+Mutation (A_M)

Alignement souhaité

```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  \  \  \  \  \  |
I2 :  a  a  a  b  b  b  b  b  d  b  a
```

| Appariement exact

[Mutation (M)

\ Amplification (A)

(Amplification+Mutation (A_M)

```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  (  \  \  \  \  |
I2 :  a  a  a  b  b  c  b  b  d  b  a
```

|	Appariement exact
[	Mutation (M)
\	Amplification (A)
(	Amplification+Mutation (A_M)

```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  (  \  (  \  \  |
I2 :  a  a  a  b  b  c  b  d  d  b  a
```

|	Appariement exact
[	Mutation (M)
\	Amplification (A)
(	Amplification+Mutation (A_M)

```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  (  \  (  \  \  |
I2 :  a  a  a  b  b  c  b  d  d  b  a
```

|	Appariement exact
[	Mutation (M)
\	Amplification (A)
(	Amplification+Mutation (A_M)

```
I1 :  a  e  a  a  -  -  -  -  -  -  a
      |  [  |  [  \  (  \  (  \  \  |
I2 :  a  a  a  b  b  c  b  d  d  b  a
                  ↑
```

Arche

|	Appariement exact
[	Mutation (M)
\	Amplification (A)
(	Amplification+Mutation (A_M)

Le modèle évolutif adapté (2)

- Opérations composées
  → **Génération(G)/Compression(K)** génère/compresse une **arche** à partir/en son caractère ancêtre :

$$\text{a} \xrightarrow{\text{Génération}} \text{a b c d a} \xrightarrow{\text{Compression}} \text{a}$$

- Alignement avec duplication [Benson 97] :
  $\rightarrow$ alignement entre 2 séquences pouvant contenir des répétitions en tandem,
  $\rightarrow$ opérations = substitutions, indels et duplications,
  $\rightarrow$ différences avec notre approche :
  1. les unités répétées ne sont pas connues,
  2. ses duplications relient les 2 séquences ;

- Alignement avec duplication [Benson 97] :
  → alignement entre 2 séquences pouvant contenir des répétitions en tandem,
  → opérations = substitutions, indels et duplications,
  → différences avec notre approche :
  1. les unités répétées ne sont pas connues,
  2. ses duplications relient les 2 séquences ;

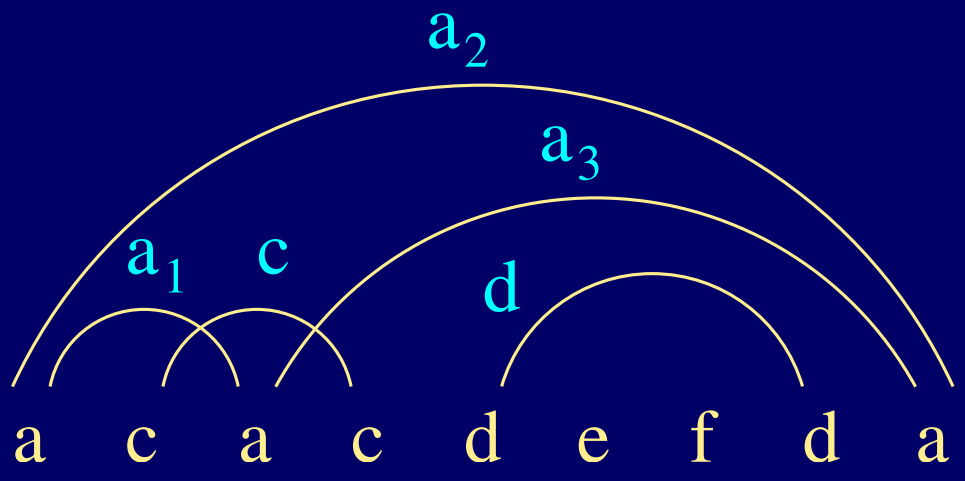
- Reconstruction de l'histoire des duplications [Benson et Dong 99, Tang et al 01, Élémento et al 02, Élémento et Gascuel 02, Jaitly et al 02, Tang et al 02 et Élémento et Gascuel 03]
  → Arbre de duplication.

- Introduction
- Comparaison de cartes de minisatellite par alignement
- Algorithmique
  - → Les arches
  - → Calcul du coût des arches
  - → L'algorithme
- Applications au minisatellite MSY1
- Conclusion et perspectives

Les arches

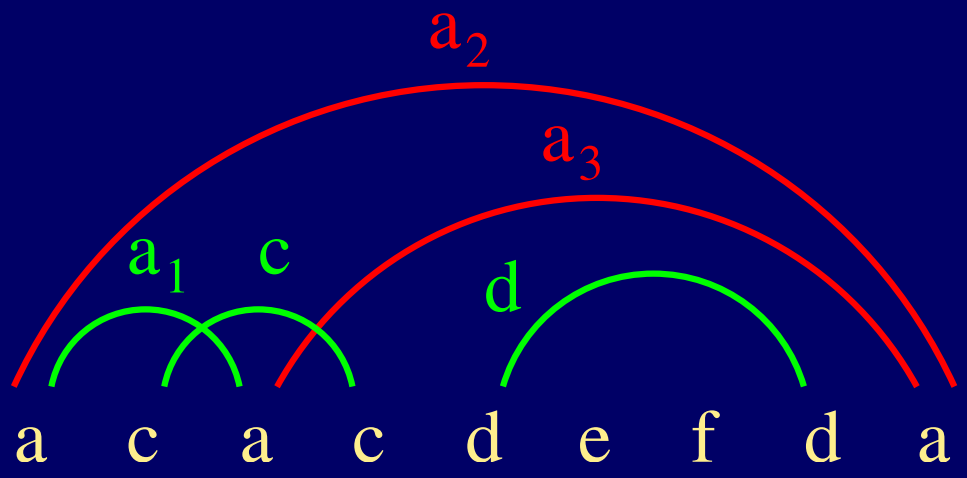
- Une arche de s est un facteur de s dont le premier et le dernier caractère sont identiques ;

- Une arche est simple si ses caractères internes apparaissent une seule fois et complexe sinon ;



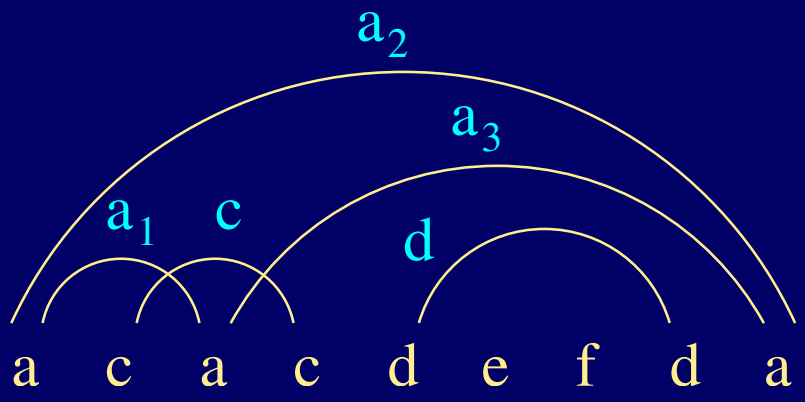
- Une arche de s est un facteur de s dont le premier et le dernier caractère sont identiques ;

- Une arche est simple si ses caractères internes apparaissent une seule fois et complexe sinon ;

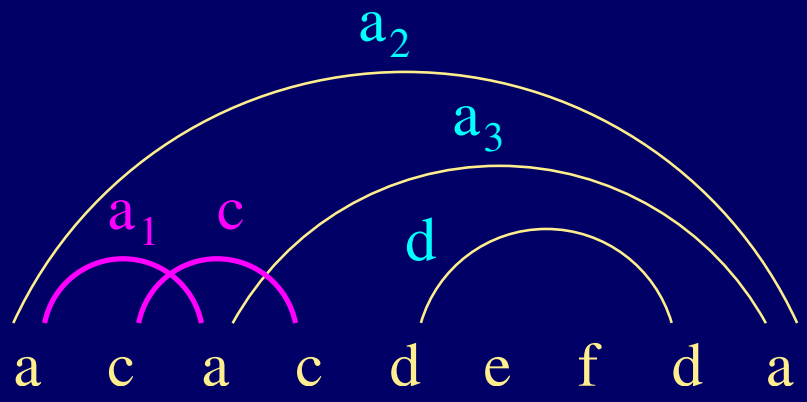


Une arche complexe contient toujours au moins une arche.

Relation de compatibilité entre arches

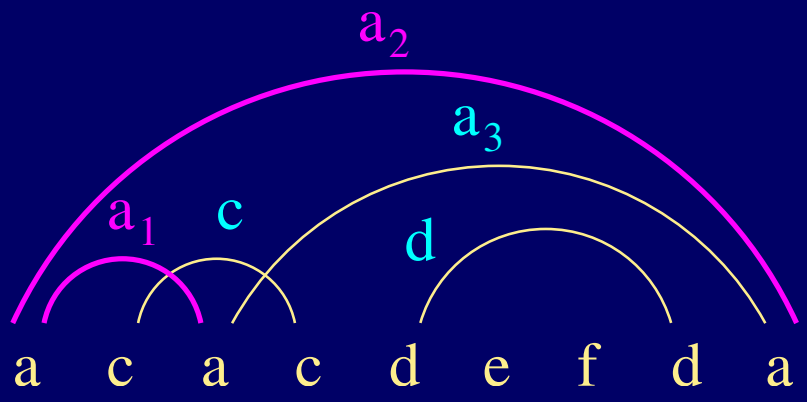


- 2 arches sont incompatibles si :



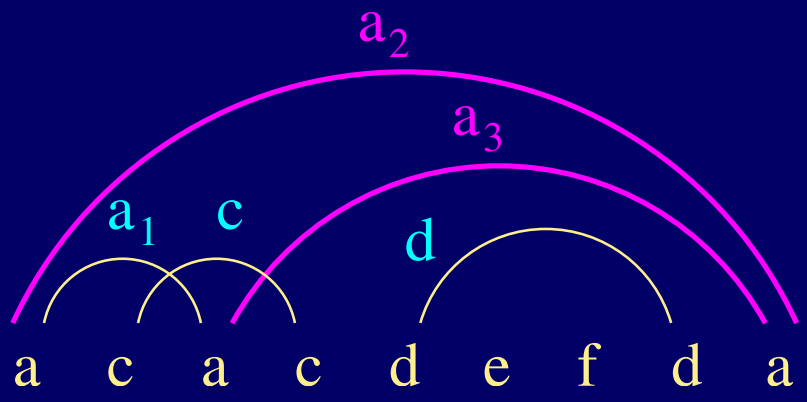
- 2 arches sont incompatibles si :
  1. elles se croisent,

Relation de compatibilité entre arches



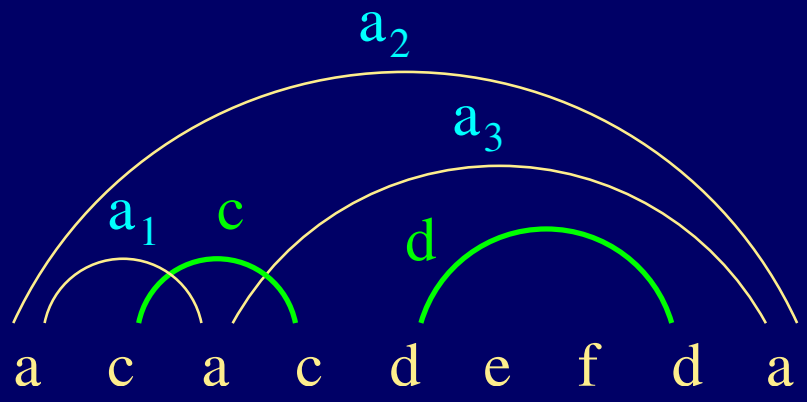
- 2 arches sont incompatibles si :
  1. elles se croisent,
  2. elles partagent le même premier ou dernier pied ;

Relation de compatibilité entre arches



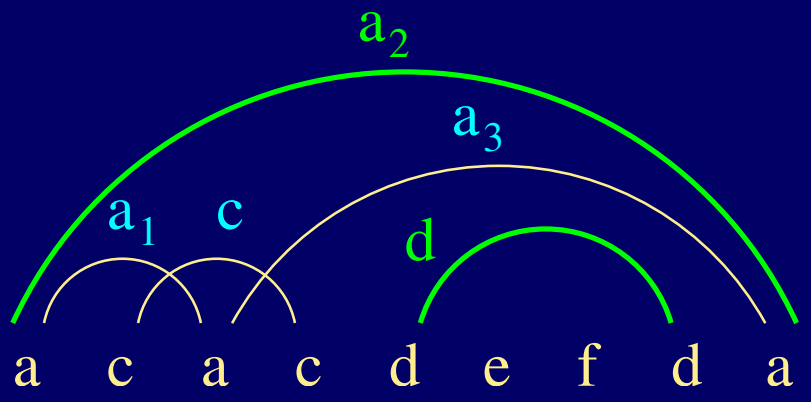
- 2 arches sont incompatibles si :
  1. elles se croisent,
  2. elles partagent le même premier ou dernier pied ;

Relation de compatibilité entre arches



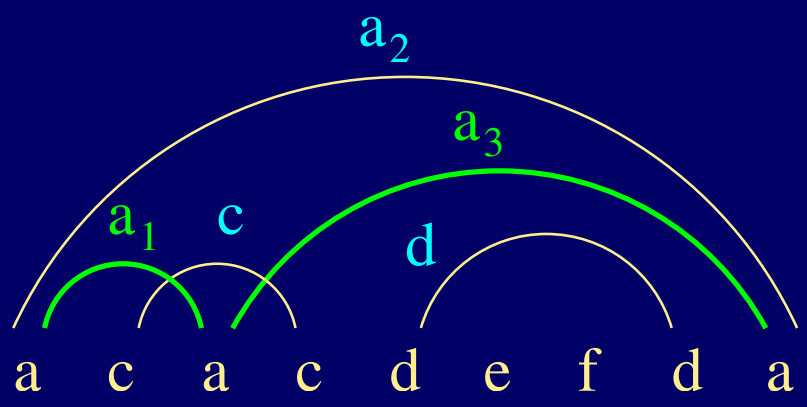
- 2 arches sont incompatibles si :
  1. elles se croisent,
  2. elles partagent le même premier ou dernier pied ;
- 2 arches sont compatibles sinon.

Relation de compatibilité entre arches



- 2 arches sont incompatibles si :
  1. elles se croisent,
  2. elles partagent le même premier ou dernier pied ;
- 2 arches sont compatibles sinon.

Relation de compatibilité entre arches



- 2 arches sont incompatibles si :
  1. elles se croisent,
  2. elles partagent le même premier ou dernier pied ;
- 2 arches sont compatibles sinon.

Générer/Compresser une arche

- Générer une arche simple : de a à $abca$

Générer/Compresser une arche

- Générer une arche simple : de a à $abca$

Générer/Compresser une arche

- Générer une arche simple : de a à $abca$

Première possibilité

a

- Générer une arche simple : de a à $abca$

Première possibilité	
	a
Amplification	a

- Générer une arche simple : de a à $abca$

Première possibilité

a

Amplification $a \quad a$

- Générer une arche simple : de a à $abca$

Première possibilité

a

Amplification a a

Mutation a a

- Générer une arche simple : de a à $abca$

Première possibilité

a

Amplification a a

Mutation a b

Générer/Compresser une arche

- Générer une arche simple : de a à $abca$

Première possibilité

a

Amplification a a

Mutation a b

Amplification a b

- Générer une arche simple : de a à $abca$

Première possibilité

	a		
Amplification	a	a	
Mutation	a	b	
Amplification	a	b	b

- Générer une arche simple : de a à $abca$

Première possibilité

	a		
Amplification	a	a	
Mutation	a	b	
Amplification	a	b	b
Mutation	a	b	b

- Générer une arche simple : de a à $abca$

Première possibilité

	a		
Amplification	a	a	
Mutation	a	b	
Amplification	a	b	b
Mutation	a	b	c

- Générer une arche simple : de a à $abca$

Première possibilité

	a		
Amplification	a	a	
Mutation	a	b	
Amplification	a	b	b
Mutation	a	b	c
Amplification	a	b	c

- Générer une arche simple : de a à $abca$

Première possibilité

	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c

- Générer une arche simple : de a à $abca$

Première possibilité

	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	c

- Générer une arche simple : de a à $abca$

Première possibilité

	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

- Générer une arche simple : de a à $abca$

Première possibilité

	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Générer/Compresser une arche

- Générer une arche simple : de a à $abca$

Première possibilité				
	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité
a

- Générer une arche simple : de a à $abca$

Première possibilité				
	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité	
	a
Amplification	a

- Générer une arche simple : de a à $abca$

Première possibilité

	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité

	a	
Amplification	a	a

- Générer une arche simple : de a à $abca$

Première possibilité				
	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité		
	a	
Amplification	a	a
Amplification	a	a

- Générer une arche simple : de a à $abca$

Première possibilité

a

Amplification $a\ a$

Mutation $a\ b$

Amplification $a\ b\ b$

Mutation $a\ b\ c$

Amplification $a\ b\ c\ c$

Mutation $a\ b\ c\ a$

Coût = 3A + 3M

Seconde possibilité

a

Amplification $a\ a$

Amplification $a\ a\ a$

- Générer une arche simple : de a à $abca$

Première possibilité				
	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité			
	a		
Amplification	a	a	
Amplification	a	a	a
Amplification	a	a	a

- Générer une arche simple : de a à $abca$

Première possibilité				
	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité				
	a			
Amplification	a	a		
Amplification	a	a	a	
Amplification	a	a	a	a

Générer/Compresser une arche

- Générer une arche simple : de a à $abca$

Première possibilité

	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité

	a			
Amplification	a	a		
Amplification	a	a	a	
Amplification	a	a	a	a
Mutation	a	a	a	a

- Générer une arche simple : de a à $abca$

Première possibilité

	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité

	a			
Amplification	a	a		
Amplification	a	a	a	
Amplification	a	a	a	a
Mutation	a	b	a	a

- Générer une arche simple : de a à $abca$

Première possibilité

	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité

	a			
Amplification	a	a		
Amplification	a	a	a	
Amplification	a	a	a	a
Mutation	a	b	a	a
Mutation	a	b	a	a

Générer/Compresser une arche

- Générer une arche simple : de a à $abca$

Première possibilité

	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité

	a			
Amplification	a	a		
Amplification	a	a	a	
Amplification	a	a	a	a
Mutation	a	b	a	a
Mutation	a	b	c	a

- Générer une arche simple : de a à $abca$

Première possibilité				
	a			
Amplification	a	a		
Mutation	a	b		
Amplification	a	b	b	
Mutation	a	b	c	
Amplification	a	b	c	c
Mutation	a	b	c	a

Coût = 3A + 3M

Seconde possibilité				
	a			
Amplification	a	a		
Amplification	a	a	a	
Amplification	a	a	a	a
Mutation	a	b	a	a
Mutation	a	b	c	a

Coût = 3A + 2M

- Générer une arche simple : de a à $abca$

Première possibilité

a

Amplification $a\ a$

Mutation $a\ b$

Amplification $a\ b\ b$

Mutation $a\ b\ c$

Amplification $a\ b\ c\ c$

Mutation $a\ b\ c\ a$

Coût = 3A + 3M

Seconde possibilité

a

Amplification $a\ a$

Amplification $a\ a\ a$

Amplification $a\ a\ a\ a$

Mutation $a\ b\ a\ a$

Mutation $a\ b\ c\ a$

Coût = 3A + 2M

- Générer une arche simple : de a à $abca$

Première possibilité

a

Amplification $a\ a$

Mutation $a\ b$

Amplification $a\ b\ b$

Mutation $a\ b\ c$

Amplification $a\ b\ c\ c$

Mutation $a\ b\ c\ a$

Coût = 3A + 3M

Seconde possibilité

a

Amplification $a\ a$

Amplification $a\ a\ a$

Amplification $a\ a\ a\ a$

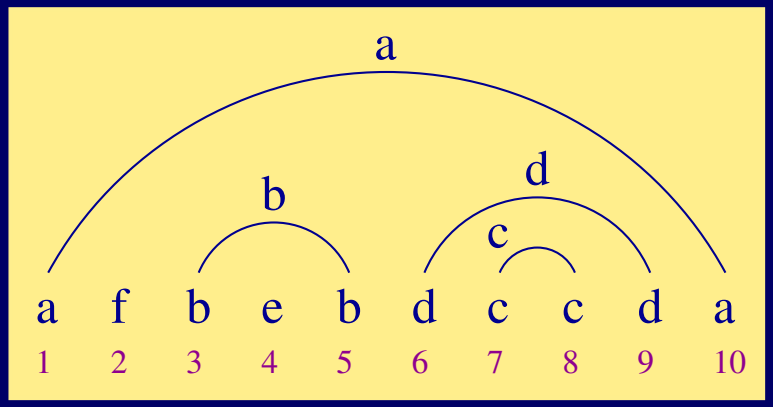
Mutation $a\ b\ a\ a$

Mutation $a\ b\ c\ a$

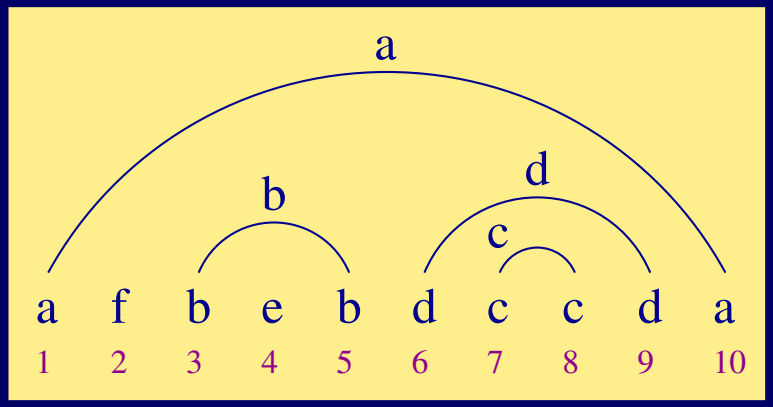
Coût = 3A + 2M

Théorème : Seules les arches permettent de faire une économie par rapport à une génération par préfixes.

Générer/Compresser une arche complexe ?



Générer/Compresser une arche complexe ?

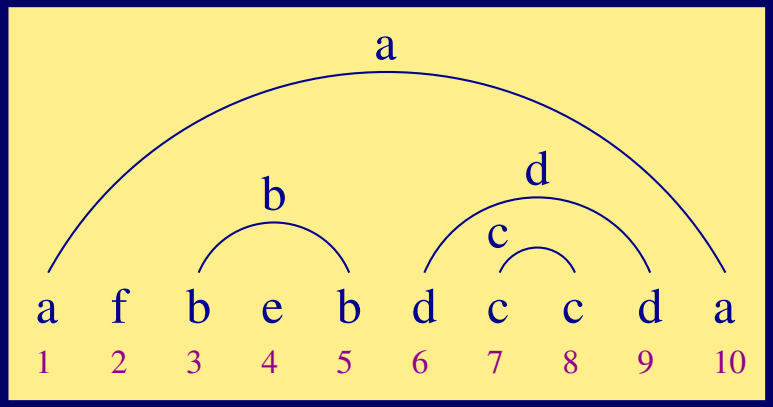


a

g

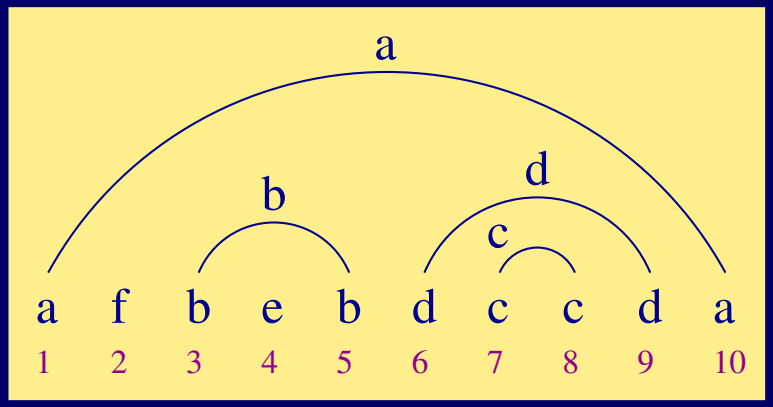
1 2 3 4 5 6 7 8 9 10

Générer/Compresser une arche complexe ?



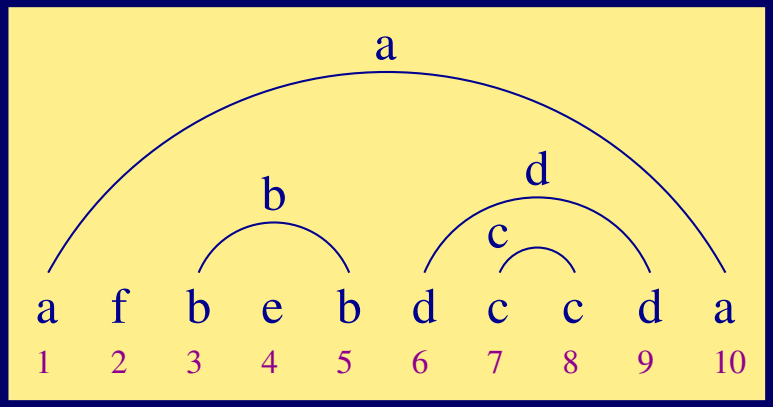
								a	a
								A	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



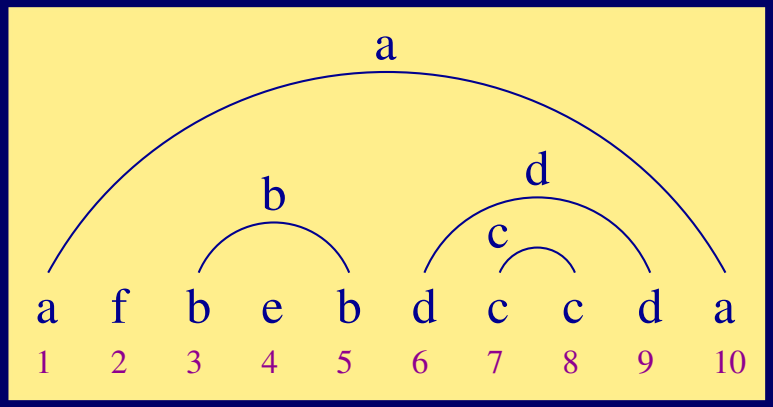
		a						a	a
		A						A	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



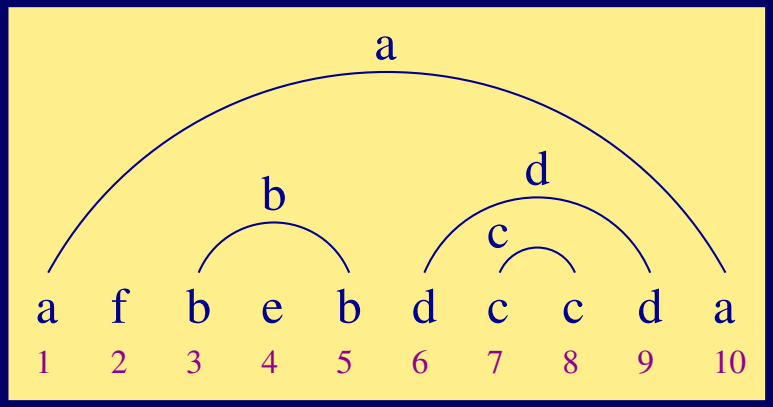
	a	a						a	a
	A	*A*						*A*	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



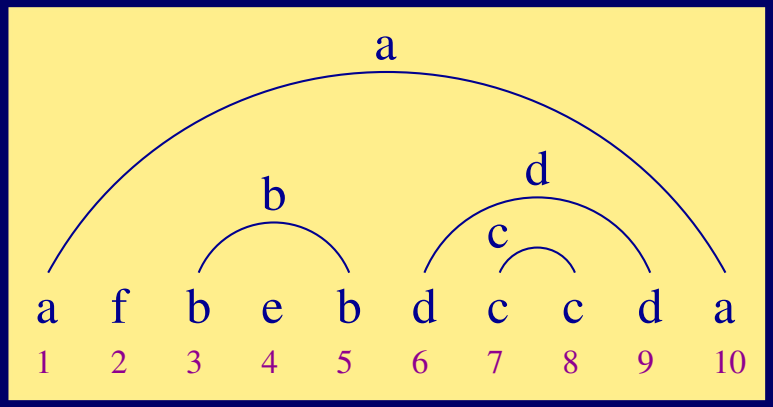
a	a	a						a	a
A	*A*	*A*						*A*	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



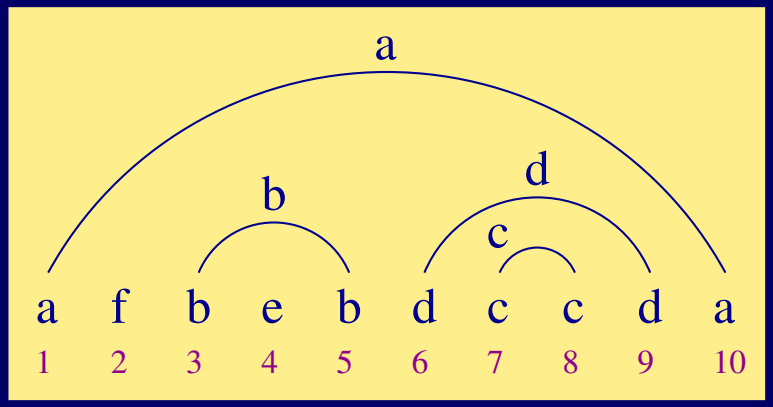
a	a	b						a	a
A	A	A_M						A	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



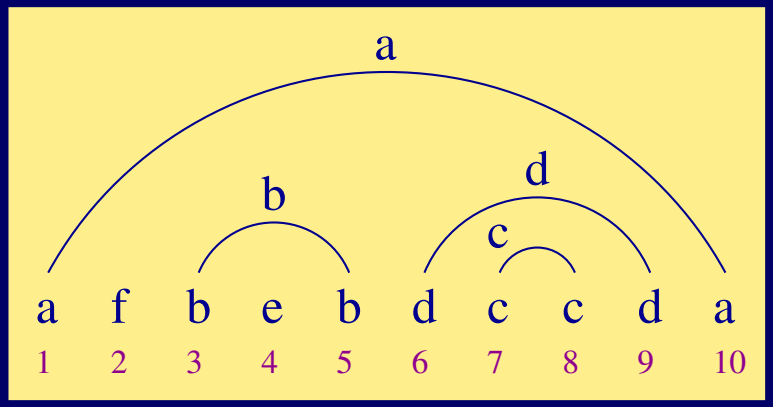
a	a	b	b					a	a
A	A	A_M	A					A	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



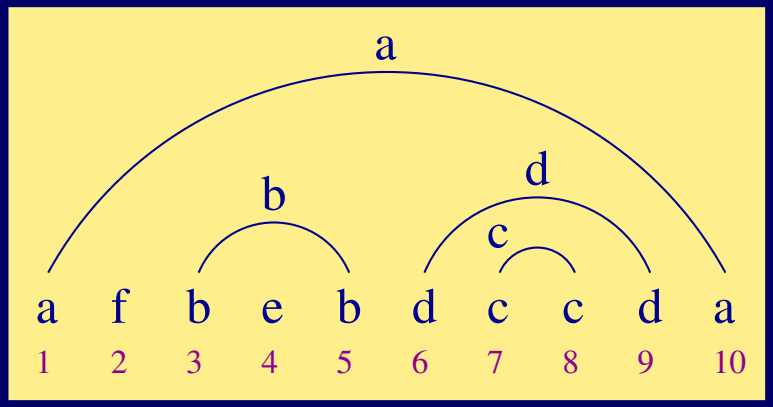
a	a	b	b	b				a	a
A	A	A_M	A	A				A	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



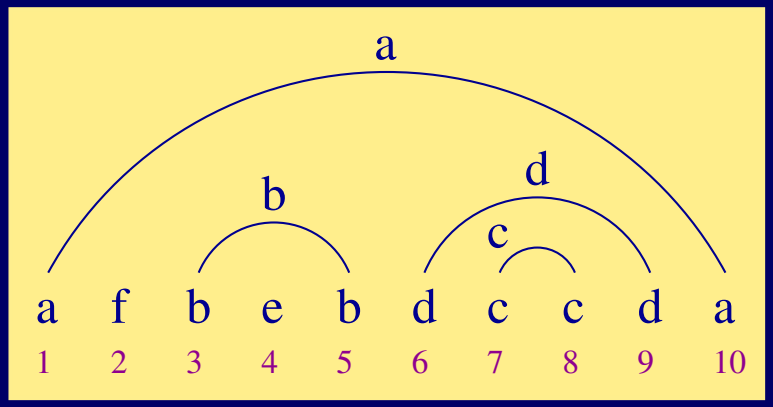
a	a	b	e	b				a	a
A	A	A_M	A_M	A				A	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



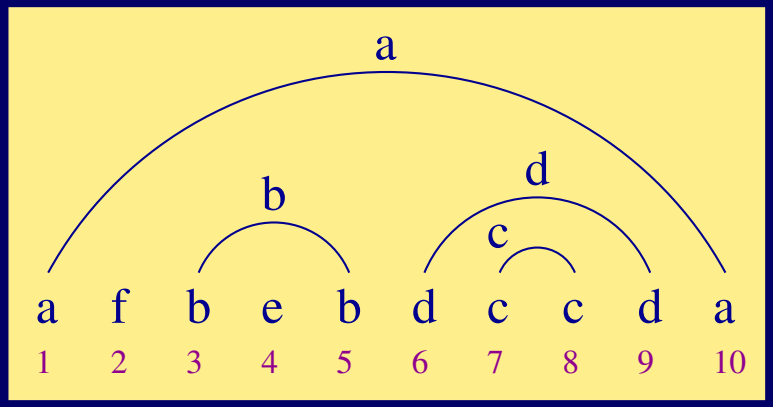
a	a	b	e	b				d	a
A	A	A_M	A_M	A				A_M	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



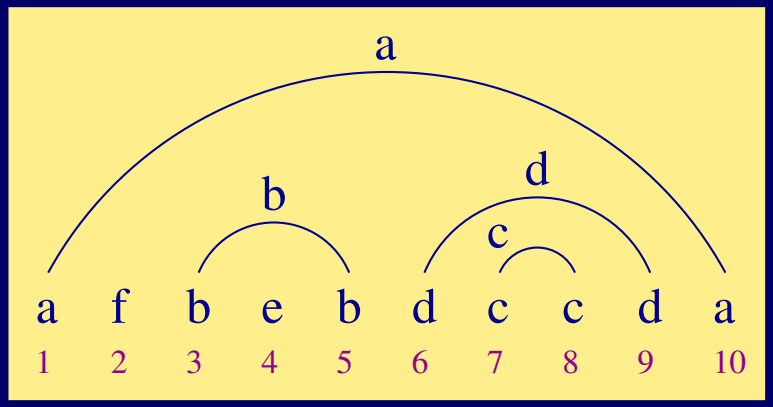
a	a	b	e	b			d	d	a
A	A	A_M	A_M	A			A	A_M	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



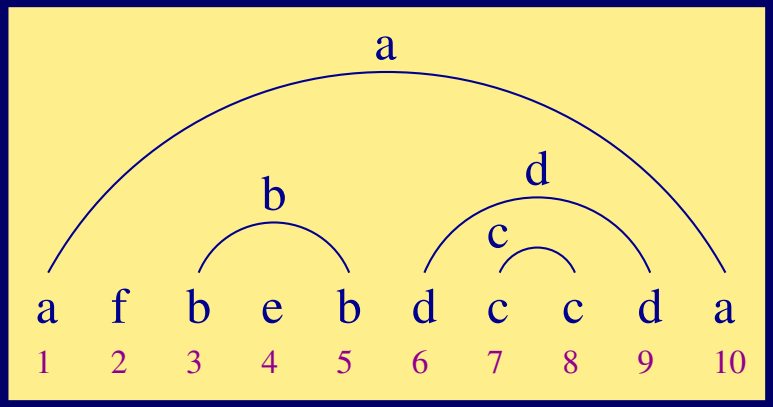
a	a	b	e	b	d		d	d	a
A	A	A_M	A_M	A	A		A	A_M	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



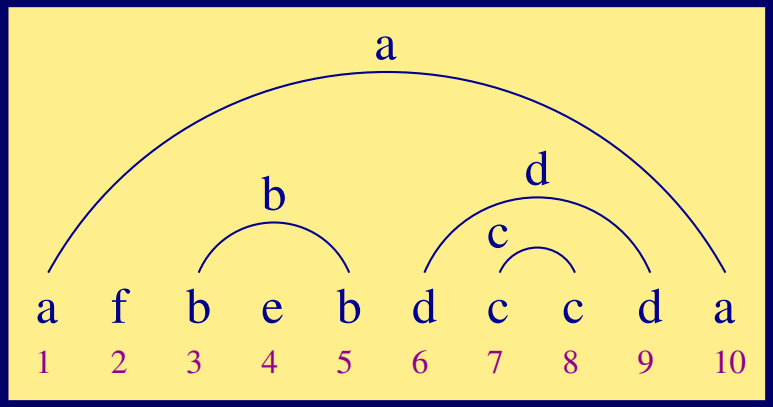
a	a	b	e	b	d		c	d	a
A	A	A_M	A_M	A	A		A_M	A_M	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



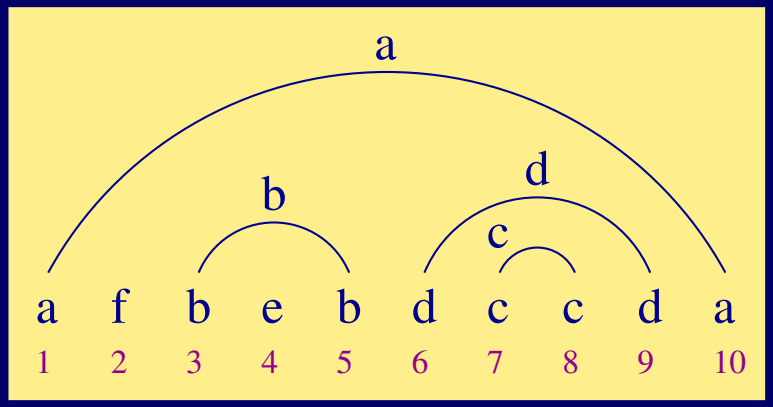
a	a	b	e	b	d	c	c	d	a
A	A	A_M	A_M	A	A	A	A_M	A_M	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



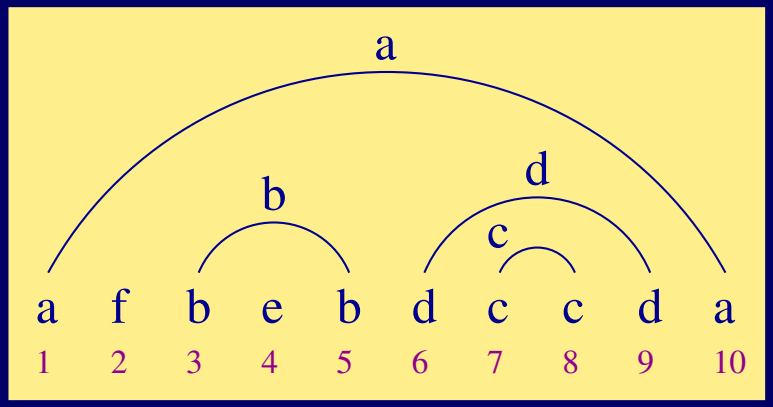
a	f	b	e	b	d	c	c	d	a
A	A_M	A_M	A_M	A	A	A	A_M	A_M	g
1	2	3	4	5	6	7	8	9	10

Générer/Compresser une arche complexe ?



a	f	b	e	b	d	c	c	d	a
A	A_M	A_M	A_M	A	A	A	A_M	A_M	g

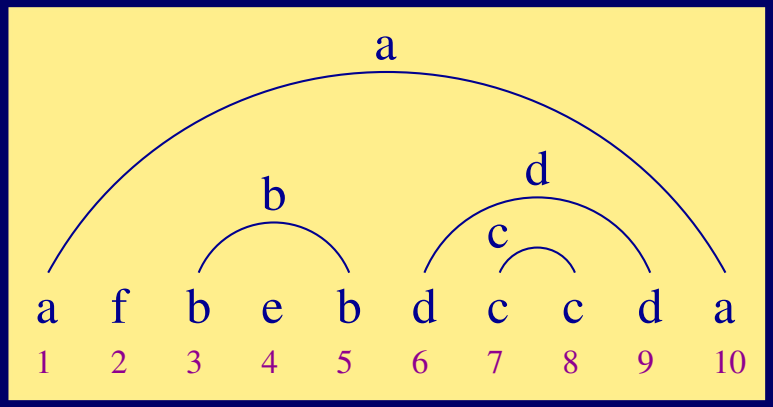
Générer/Compresser une arche complexe ?



a	f	b	e	b	d	c	c	d	a
A	A_M	A_M	A_M	A	A	A	A_M	A_M	g

- 4 arches utilisées pour générer 9 variants : Coût = 9A + 5M

Générer/Compresser une arche complexe ?

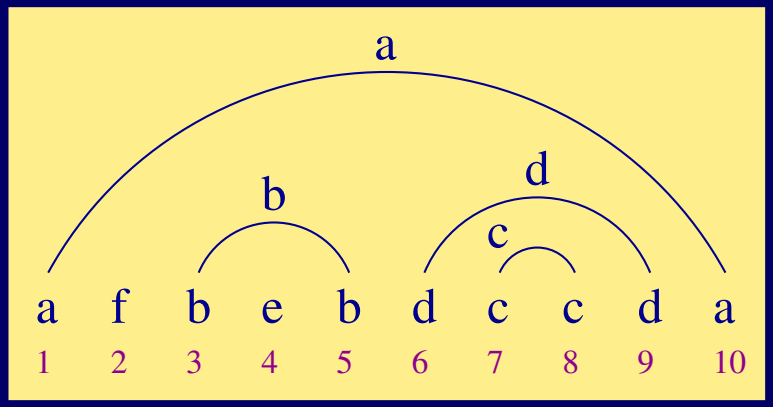


a	f	b	e	b	d	c	c	d	a
A	A_M	A_M	A_M	A	A	A	A_M	A_M	g

- 4 arches utilisées pour générer 9 variants : Coût = 9A + 5M
- Le coût $c(a)$ de génération d'une arche a de longueur k utilisant p arches 2 à 2 compatibles est :

$$c(a) = (k-1) \times A + (k-1-p) \times M$$

Générer/Compresser une arche complexe ?



a	f	b	e	b	d	c	c	d	a
A	A_M	A_M	A_M	A	A	A	A_M	A_M	g

- 4 arches utilisées pour générer 9 variants : Coût = 9A + 5M
- Le coût $c(a)$ de génération d'une arche a de longueur k utilisant p arches 2 à 2 compatibles est :

$$c(a) = (k-1) \times A + (k-1-p) \times M$$

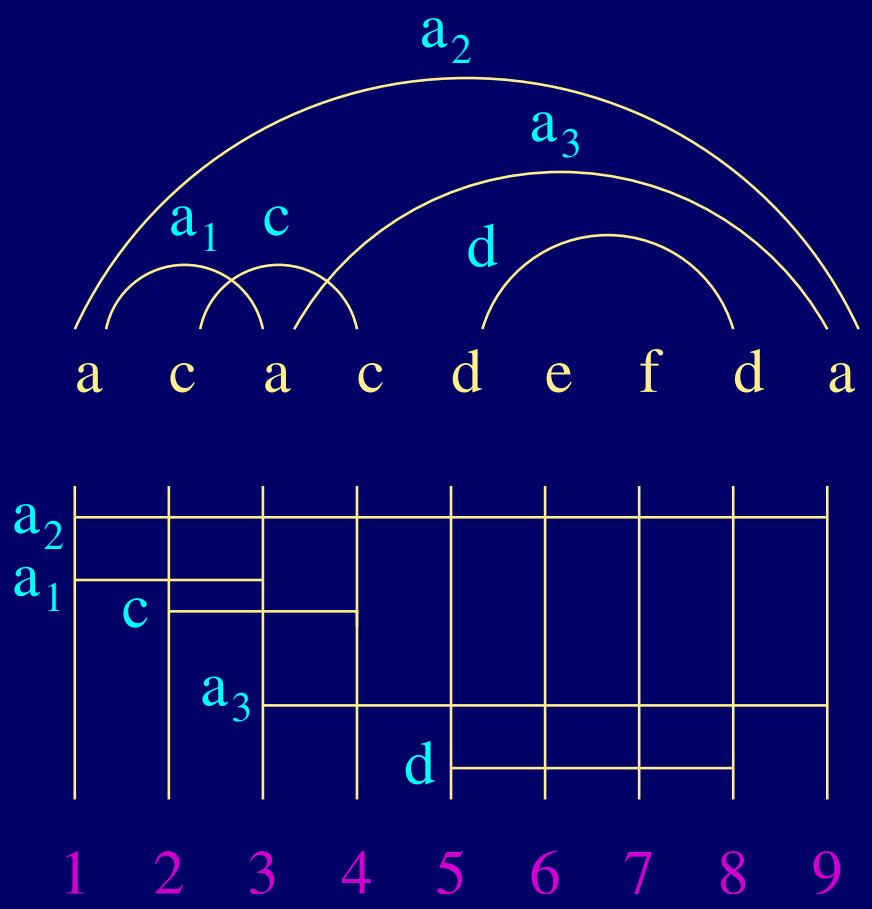
$\Rightarrow$ $c(a)$ est optimal lorsque p est maximal ;

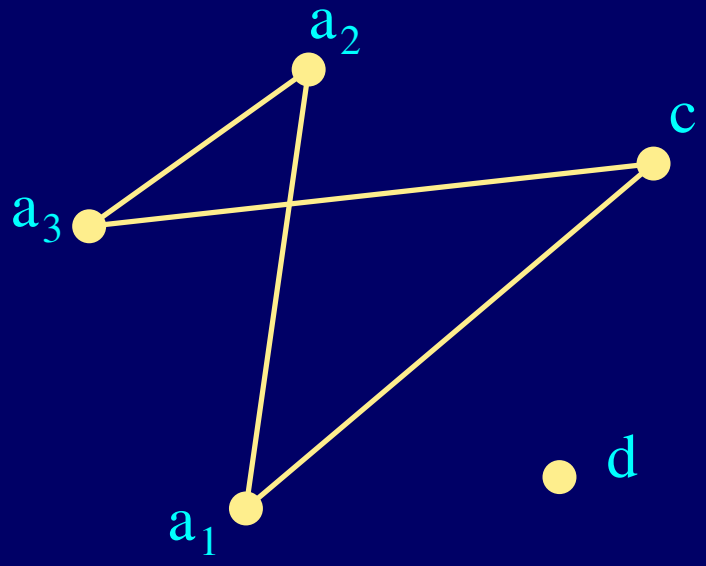
Calcul du coût des arches

Soit s une séquence de longueur n, trouver un plus grand ensemble d'arches 2 à 2 compatibles de s.

- Les arches sont des intervalles de $[1, n]$;
- La relation de compatibilité entre arches définit un graphe de chevauchement.

Exemple réduction





Théorème : Trouver un plus grand ensemble d'arches 2 à 2 compatibles est équivalent à trouver un stable max dans un graphe de chevauchement.

Calcul du coût des arches : la complexité

- But : appliquer une fois un traitement sur la séquence s et obtenir les ensembles maximaux d'arches 2 à 2 compatibles pour toutes les arches de s.

- Deux possibilités :

  1. Décaler les segments des arches et appliquer l'algorithme de [Apostolico et al, 92] $\Rightarrow O(|V^2|) = O(n^4)$,

  2. Travailler sans décalage en gérant les extrémités communes selon la relation de compatibilité $\Rightarrow O(n^3)$.

Prétraitement en $O(n^3)$

- Construire une matrice S où chaque entrée $S(i,j)$ contient un ensemble maximal d'arches 2 à 2 compatibles de $s[i,j]$;
- Programmation dynamique ;
- Initialisation :

$$\forall i \in [1,n] \begin{cases} S(i,i) = \emptyset \\ S(i,i+1) = \begin{cases} \emptyset & \text{si } s[i] \neq s[i+1] \\ \{a\} & \text{sinon, où } a \text{ est l'arche } s[i..i+1] \end{cases} \end{cases}$$

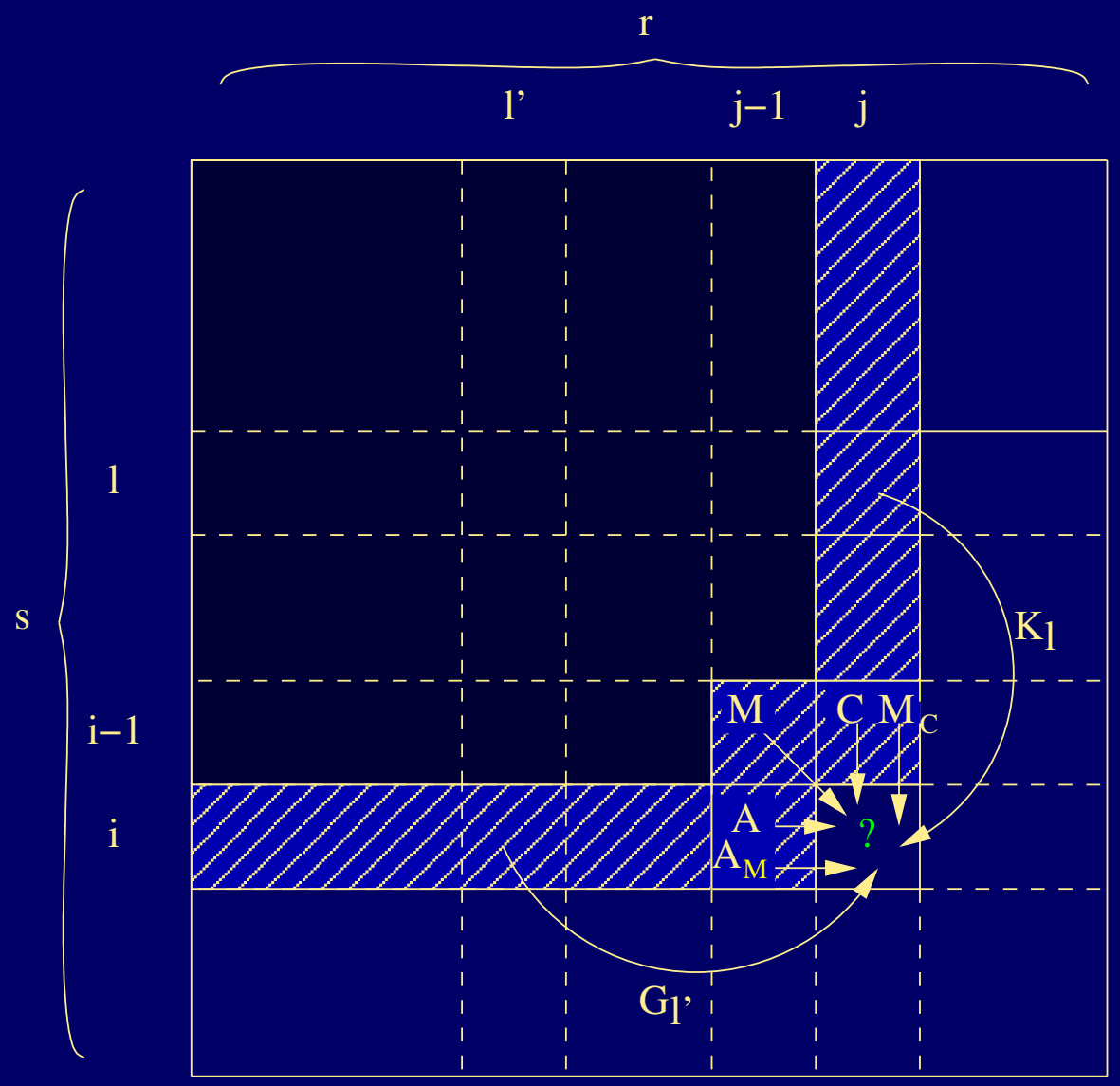
- Récurrence :

$$S(i,j) \leftarrow \max \begin{cases} S(i,k) \cup S(k,j) & \forall k \in [i+1, j-1] \\ S(i+1,j-1) \cup \{a\} & \text{si } a = s[i..j] \text{ est une arche} \end{cases}$$

- Le prétraitement permet d'obtenir tout au long de l'alignement les coûts des opérations d'arches en temps constant ;

- Principe de programmation dynamique en tenant compte des 7 opérations possibles.

Matrice de programmation dynamique



Si l'on note $t = \max(n, m)$, la complexité en temps globale de l'algorithme est $O(t^3)$.

- Introduction
- Comparaison de cartes de minisatellite par alignement
- Algorithmique
- Applications au minisatellite MSY1
  - → Questions
  - → Expériences
  - → Résultats
- Conclusion et perspectives

Jeu de données

- Jeu de données : cartes MSY1 de 609 individus provenant de 76 populations différentes et distribuées en 18 haplogroupes ;

- MSY1 est l'élément le plus polymorphe du chromosome Y ;

- [Jobling et Tyler-Smith, 2000] définit et donne les relations évolutives de 27 haplogroupes du chromosome Y à partir de marqueurs stables ;

- Haplogroupe (hg) : groupe de séquences génétiquement proches.

Questions

- Évolution du chromosome Y ;
- Expériences :
  1. Retrouver des relations évolutives connues
     A. prédiction de l'hg d'un individu à partir de sa carte MSY1,
     B. comparaison des arbres d'haplogroupes obtenus avec les deux types de marqueurs,
     C. arbre des individus d'une population,
  2. Relations évolutives récentes
     D. arbres d'évolution au sein d'un même haplogroupe ;
- Méthode : calcul de tous les alignements deux à deux, construction de l'arbre phylogénétique avec BioNJ.

- Variante de la méthode des k plus proches voisins : prédit un individu dans l'haplogroupe le plus représenté parmi ses k plus proches voisins ;

- Variante de la méthode des k plus proches voisins : prédit un individu dans l'haplogroupe le plus représenté parmi ses k plus proches voisins ;

- Lorsque k varie entre 3 et 5 les prédictions sont correctes dans 80 % des cas ;

- Variante de la méthode des k plus proches voisins : prédit un individu dans l'haplogroupe le plus représenté parmi ses k plus proches voisins ;

- Lorsque k varie entre 3 et 5 les prédictions sont correctes dans $80\ \%$ des cas ;

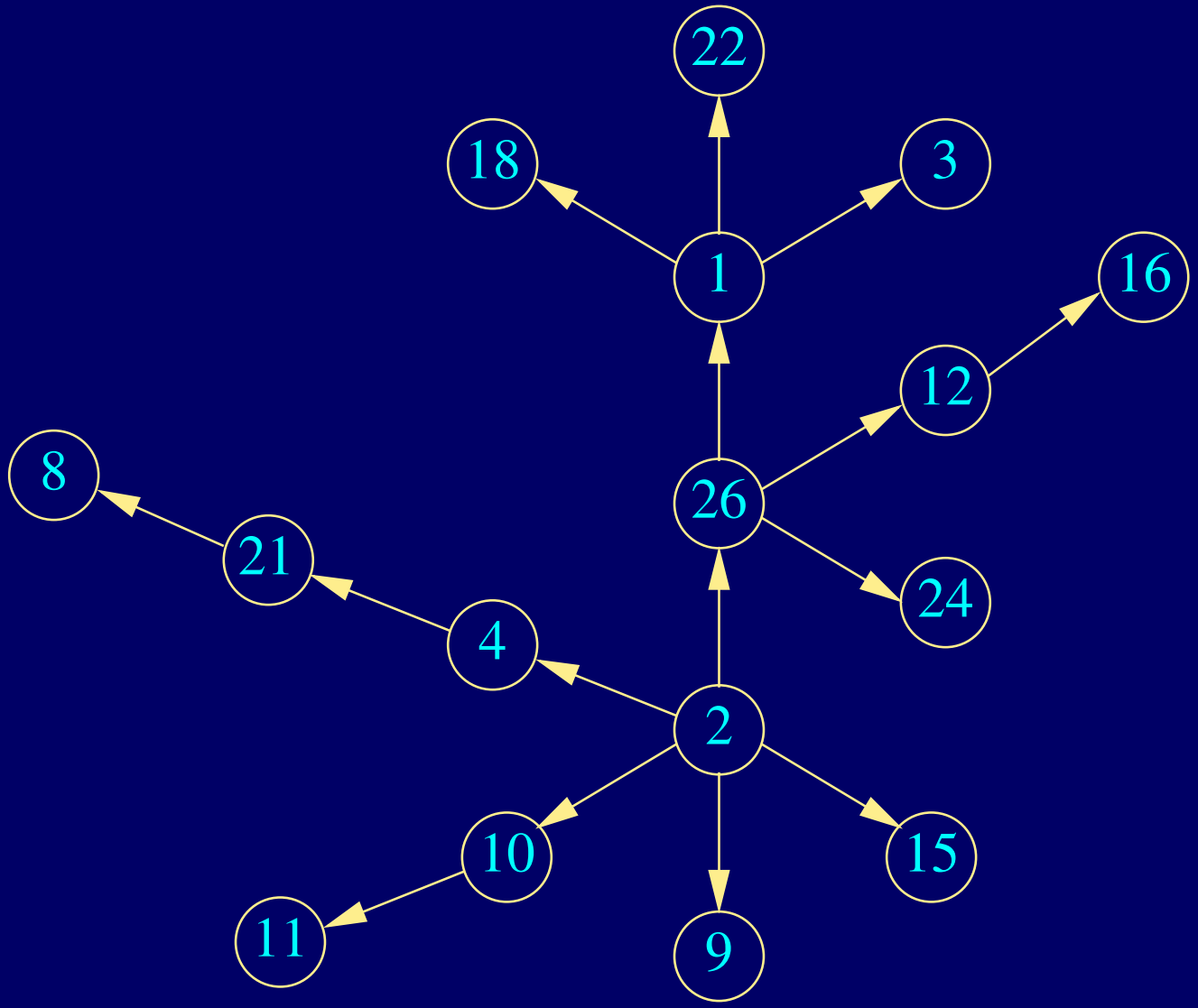
- Si l'on regarde les 2 ou 3 haplogroupes plus proches, $90\ \%$ à $93\ \%$ des prédictions sont correctes avec $k = 9$;

- Variante de la méthode des k plus proches voisins : prédit un individu dans l'haplogroupe le plus représenté parmi ses k plus proches voisins ;

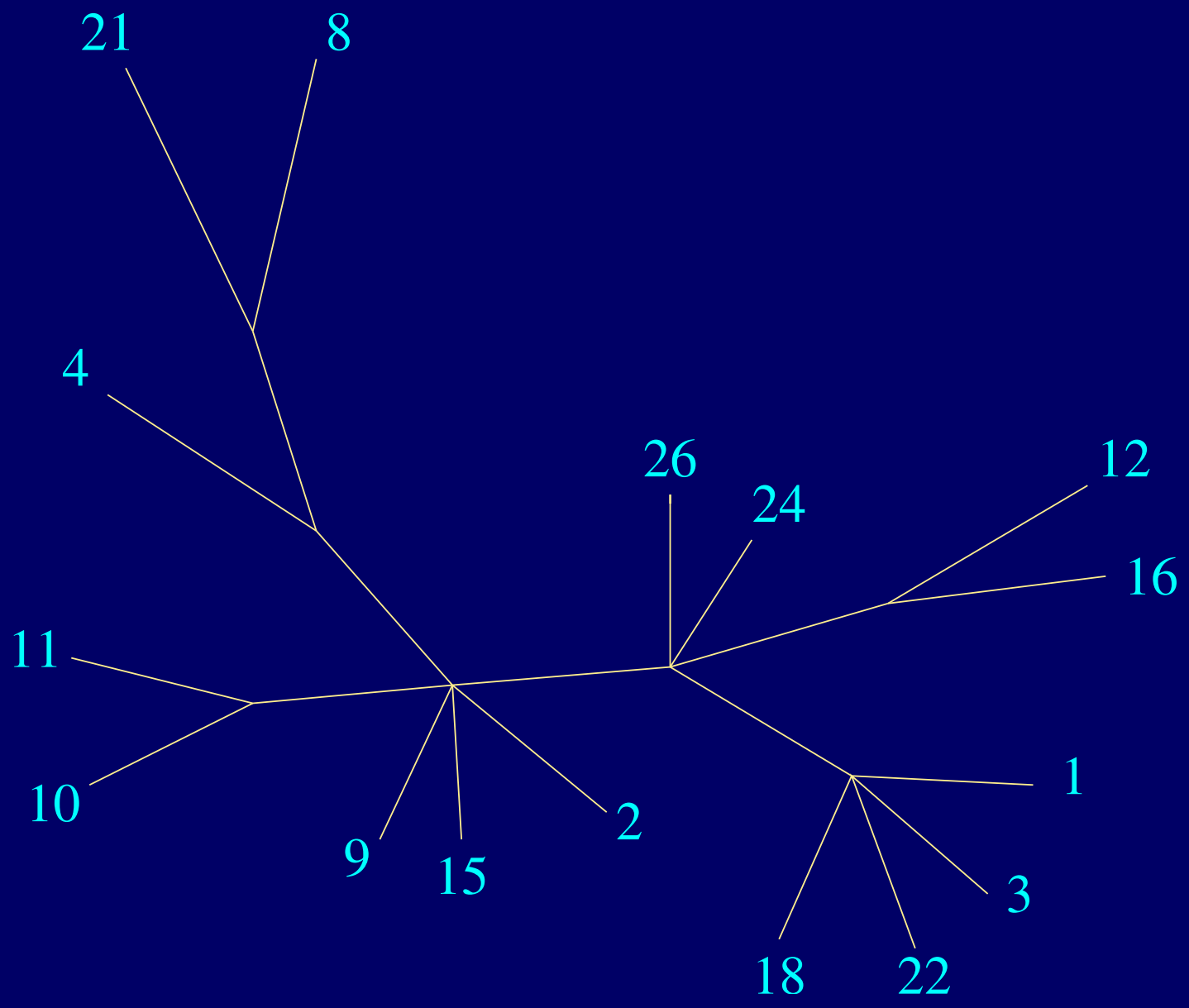
- Lorsque k varie entre 3 et 5 les prédictions sont correctes dans $80\ \%$ des cas ;

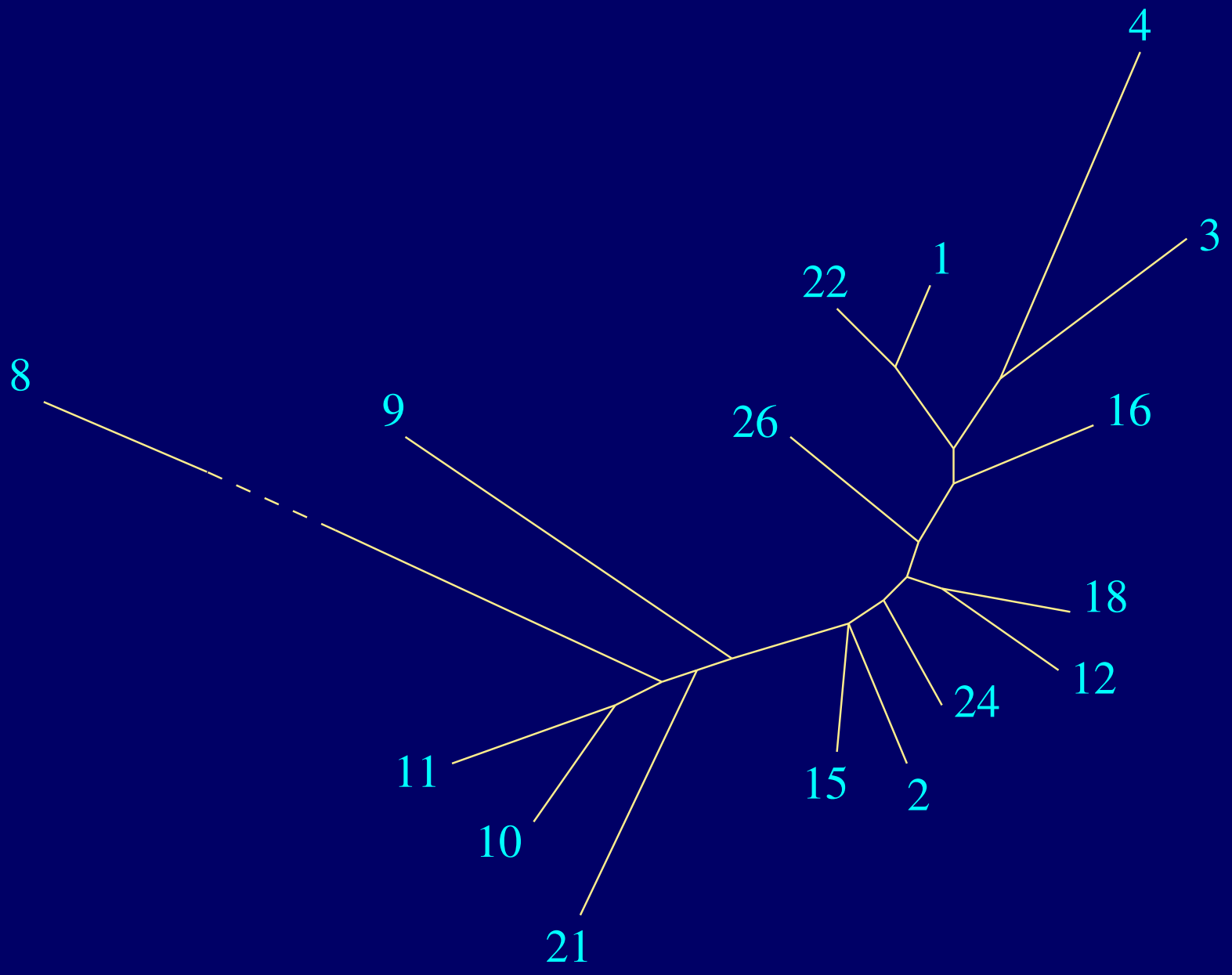
- Si l'on regarde les 2 ou 3 haplogroupes plus proches, $90\ \%$ à $93\ \%$ des prédictions sont correctes avec $k = 9$;

- L'exactitude des prédictions reste au même niveau lorsque l'on fait varier le rapport M/A entre 4 et 10.

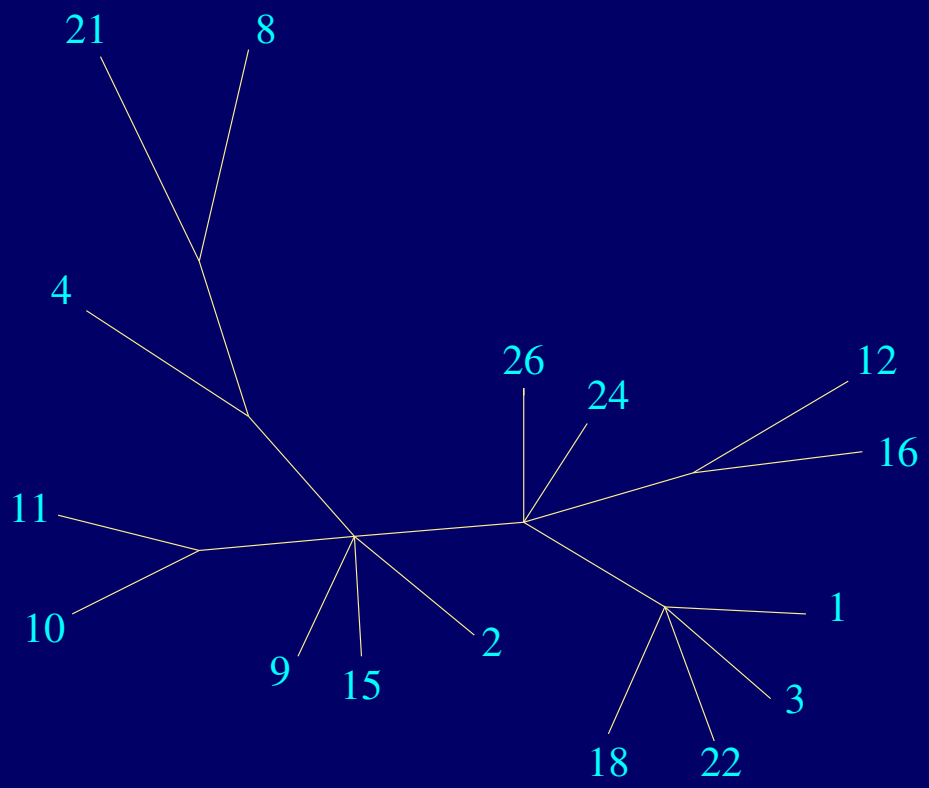


[Jobling et Tyler-Smith, 00]

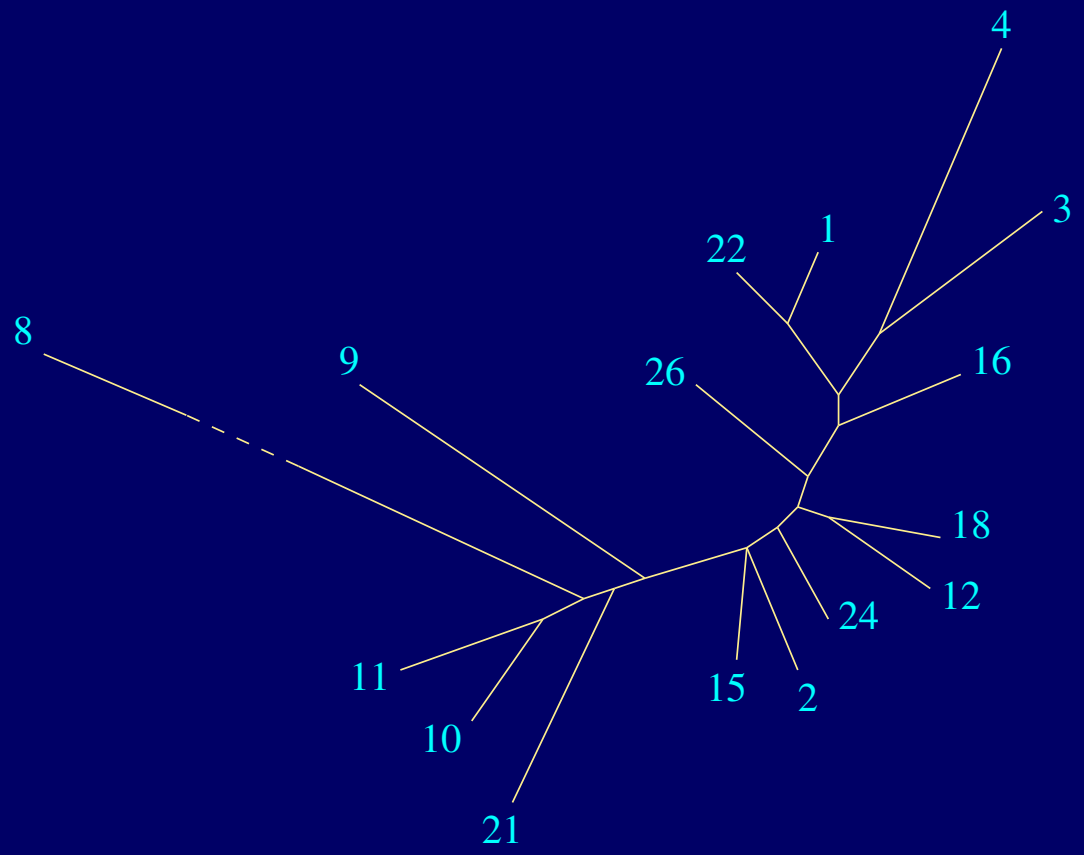

21
8
4
26
24
12
16
11
10
9
15
2
1
3
18
22



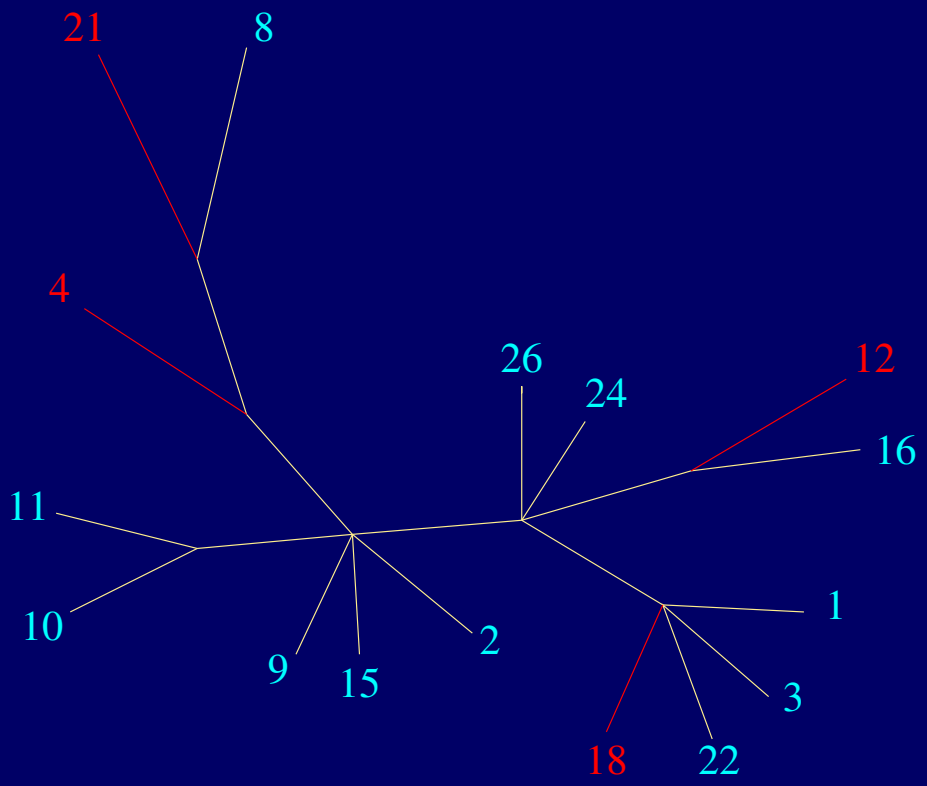
Obtenu à partir de distances moyennes



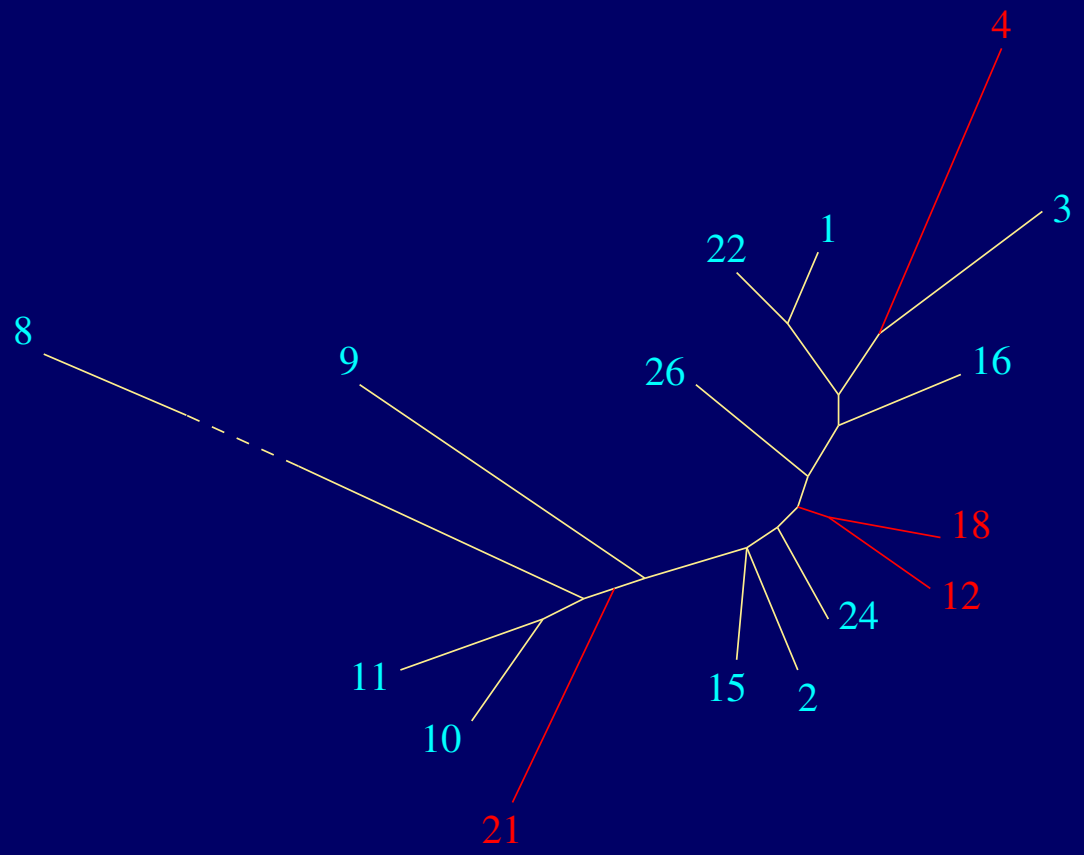
Marqueurs stables



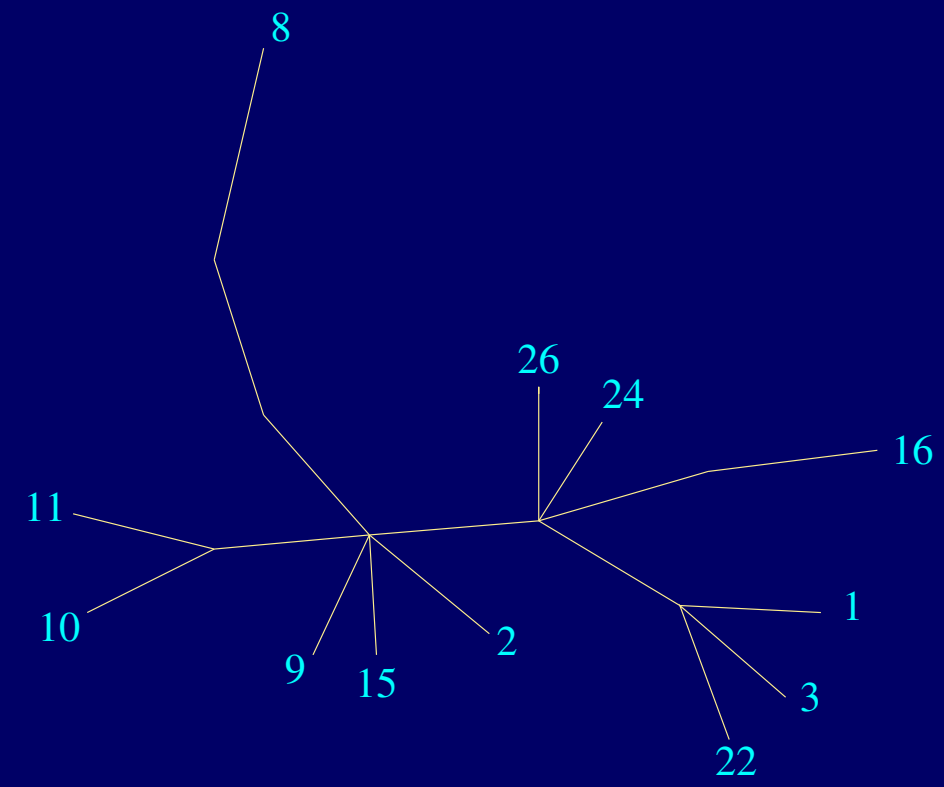
MSY1

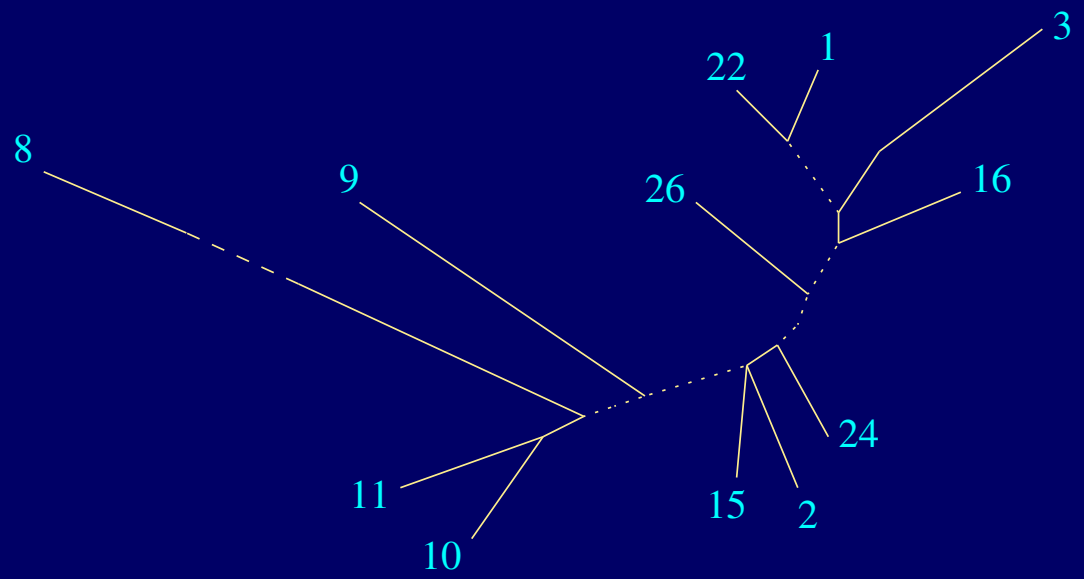


Marqueurs stables



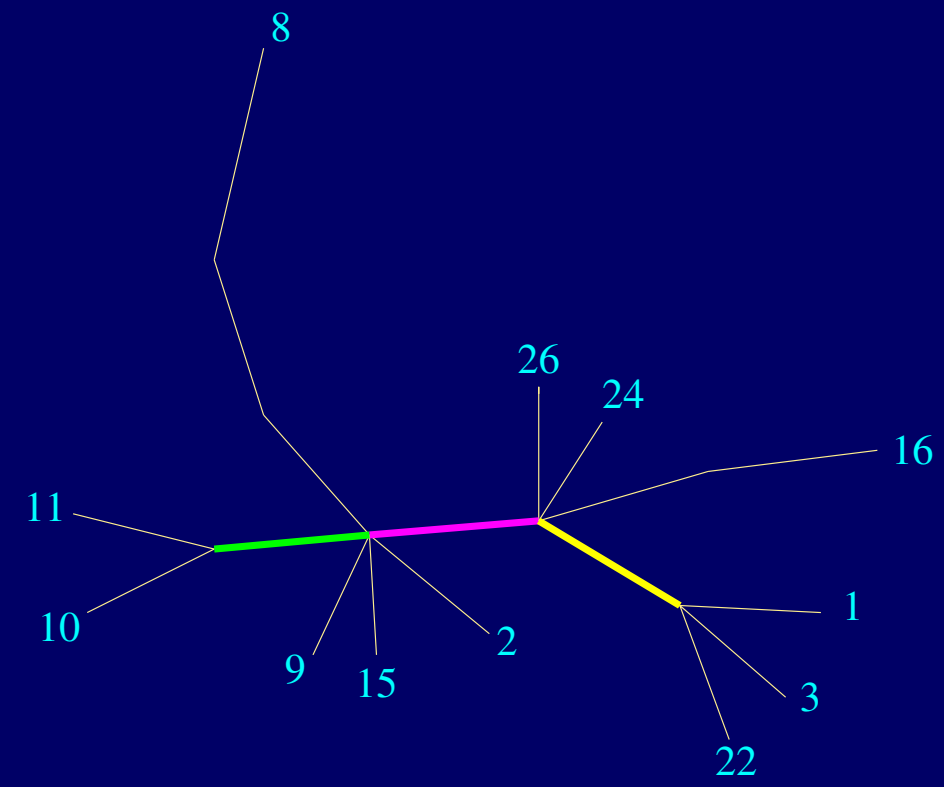
MSY1

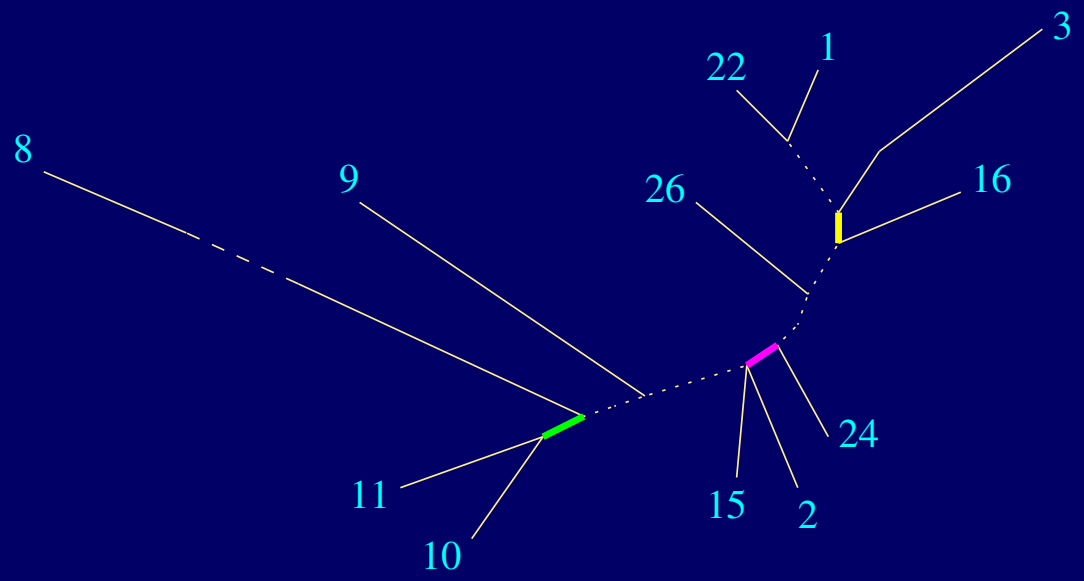




Sous-arbre maximal compatible (variante MAST)

en ôtant 4 feuilles.

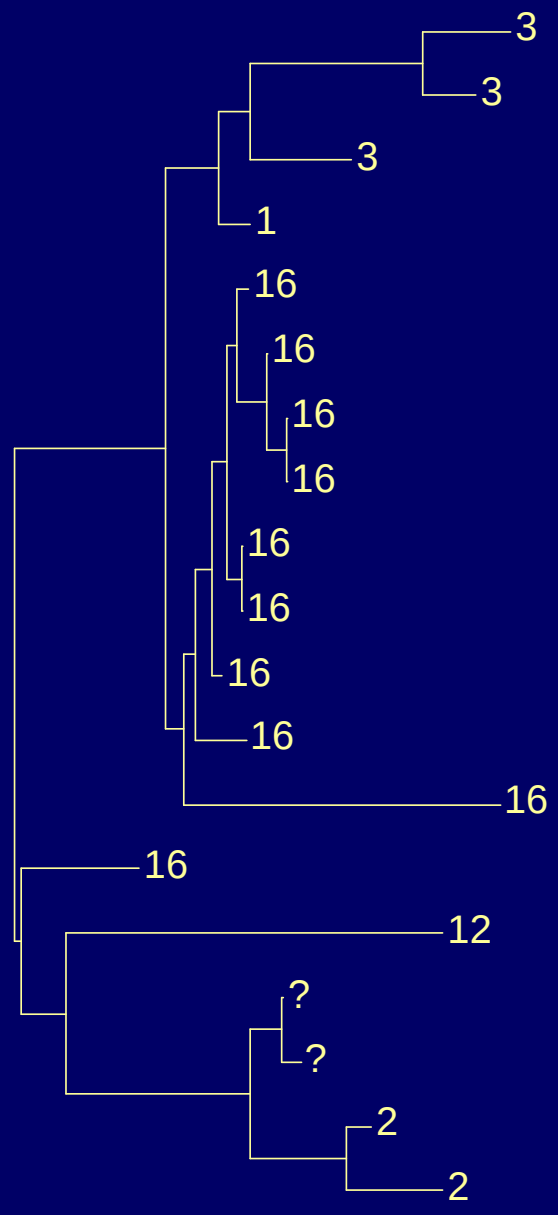




Sous-arbre maximal compatible (variante MAST)

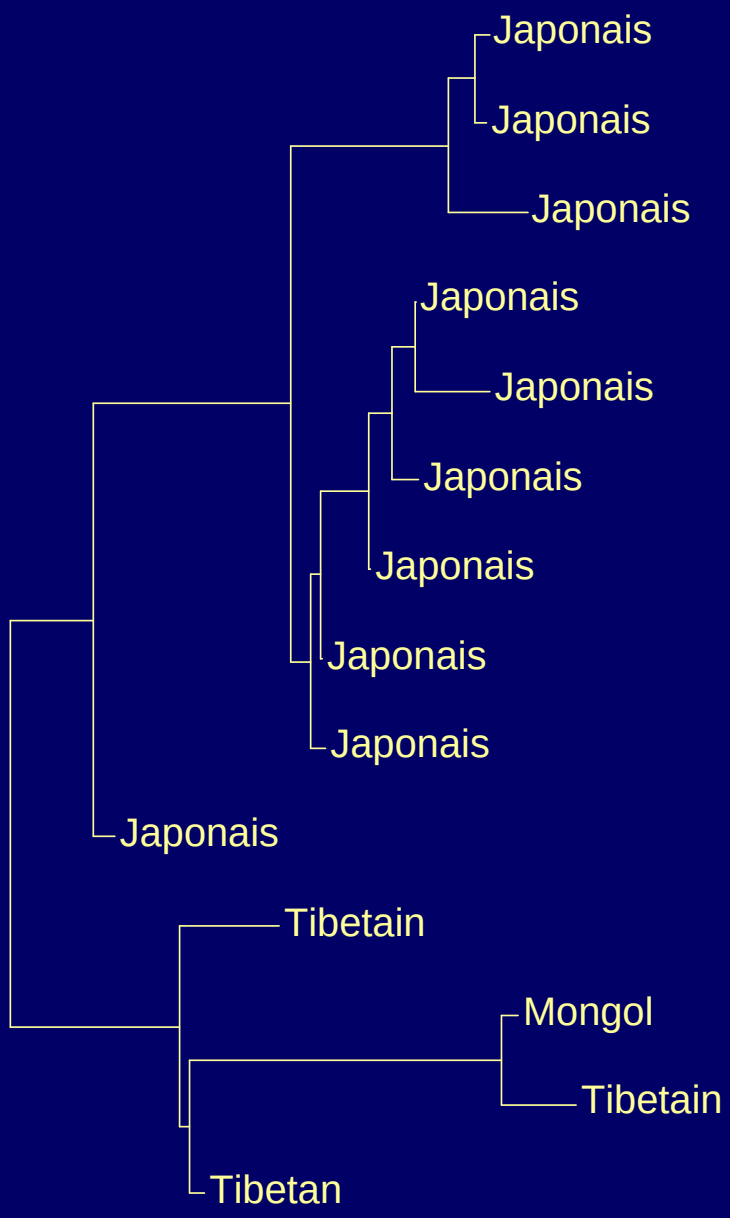
en ôtant 4 feuilles.

C. Arbre de la population finlandaise



À une exception près les haplogroupes sont monophylétiques.

D. Évolution récente au sein de l'hg 4



Conclusion

- Modélisation de l'évolution des minisatellites ;
- La première méthode exacte et spécifique d'alignement de cartes de minisatellites ;
- Logiciel MS_Align ;
- Applications sur un jeu de données biologique :
  1. permet de valider la méthode,
  2. montre que l'étude des minisatellites permet de détecter des signaux micro-évolutifs ;
- Extensions du modèle : pistes pour des heuristiques.

Perspectives

- D'autres extensions du modèle :

  1. Mutations dépendantes des variants,
  2. Ajout d'événements : recombinaison, homogénéisation, réarrangement ;

  Complexité du problème avec des modèles étendus ;

- Voie nouvelle : [Behzadi et Steyaert, 03] ;
- Alignement multiple ;
- Arbres de duplication ;
- Cartes d'un ancêtre commun.

Merci de votre attention.

$$\mathcal{A}(i,j) \;=\; \min \begin{cases} \mathcal{A}(i-1,j-1) + M(s[i],r[j]) & \text{Mutation ou App. Exact} \\ \mathcal{A}(i-1,j) + C & \text{Contraction} \\ & \text{si } s[i] = s[i-1] \text{ ou } s[i] = r[j] \\ \mathcal{A}(i-1,j) + M + C & \text{Mutation+Contraction} \\ \mathcal{A}(i,j-1) + A & \text{Amplification} \\ & \text{si } r[j] = r[j-1] \text{ ou } r[j] = s[i] \\ \mathcal{A}(i,j-1) + A + M & \text{Amplification+Mutation} \\ \mathcal{A}(l,j) + K(s[l..i]) & \text{Compression d'arche} \\ & \forall l \in [1, i-2] \text{ tq. } s[l] = s[i] \\ \mathcal{A}(i,l') + G(r[l'..j]) & \text{Génération d'arche} \\ & \forall l' \in [1, j-2] \text{ tq. } r[l'] = r[j] \end{cases}$$

Synthèse des résultats biologiques

A. Prédiction correcte d'haplogroupe pour $80\ \%$ des allèles,

B. Arbre évolutif des haplogroupes similaire à l'arbre des SNPs,

C. Arbres de populations groupant les individus de même hg et où les relations entre ces groupes reproduisent les relations entre hg,

$\Rightarrow$ MSY1 permet de retrouver les relations phylogénétiques qui ont menées à la définition des hg,

D. Arbres d'haplogroupe où les individus de même population se groupent ensemble ;

$\Rightarrow$ MSY1 peut-être utilisé pour retracer l'histoire récente du chromosome Y ;

Validation de la méthode sur MSY1, le modèle choisi est adapté à ce minisatellite.

Arbre du YCC

153 haplogroupes répartis en 19 clades majeurs