Comparison of Minisatellites

Sèverine Bérard and Éric Rivals

Laboratoire d'Informatique, de Robotique
et de Microélectronique de Montpellier
CNRS - Université Montpellier II
FRANCE

Outline

1. Tandem repeats evolution, and minisatellite maps.
2. Evolutionary model.
3. Arches.
4. Alignment algorithm.
5. Applications.

Outline

1. Tandem repeats evolution, and minisatellite maps.

2. Evolutionary model.

3. Arches.

4. Alignment algorithm.

5. Applications.

Minisatellites (ms): what are they?

- Class of tandem repeat: Satellites, Minisatellites, Microsatellites.
- Example of a tandem repetition:

  $\cdots$ cggcgat cggcgac cggagat cggcgat cggcgat cggagat cgacgat $\cdots$

- A minisatellite unit measures between 7-100 bp, and their overall length is greater than 0.5kb.
- ms undergo amplifications and contractions $\Rightarrow$ variation in their number of unit.
- ms also undergo point-mutations and homogenization.

Interests in Minisatellites

- Implication in diseases: insulin-dependent diabetes, several cancers, epilepsy, and others [Buard, Jeffreys 97].
- Phylogeny and population studies:
  - Micro-evolution in small range of evolutionary time.
  - Migration of population: out-of-Africa hypothesis [Armour et al 96].
  - Evolution of the Y chromosome [Jobling, Tyler 00].
- Individual or species identification: forensic studies, genetic markers, identification of bacteria [Jeffreys 93]
- Theoretical interest: Why such sequences exist? How do they evolve?

Minisatellites maps

- Difficulty to sequence and assemble minisatellites.
- There is a specific method to obtain the sequence of variants of the unit, **Minisatellite Variant Repeat PCR** [Jeffreys et al. 91].
  MVR-PCR yields a **ms map**: a sequence of symbols, each representing a different variant of the repeated unit.
- Example of a minisatellite map:
  - s = cggcgat cggcgac cggagat cggcgat cggcgat cggagat cgacgat
  - New alphabet: A = cggcgat B = cggcgac C = cggagat D = cgacgat
  - Corresponding map: A B C A A C D

Outline

1. Tandem repeats evolution, and minisatellite maps.
2. Evolutionary model.
3. Arches.
4. Alignment algorithm.
5. Applications.

Example of evolution of ms

Ancestor map: AAAAA

Today's variants: A,B,C,D,E

Trace of the evolution in individual 1:

	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

Example of evolution of ms (2)

Trace of the evolution in individual 2:

	A	A	A	A	A	
Mutation	A	**E**	A	A	A	
Amplification	A	E	A	A	**A**	A
Contraction	A	E	A	A	A	

We would like to compute an optimal alignment:

I1:	A	A	A	B	B	C	B	D	D	B	A
	|	[	|	[	/		/		/	/	|
I2:	A	E	A	A	-	-	-	-	-	-	A

Single Step Evolutionary Model

- Unitary operations
  - **Amplification(A)/Contraction(C)** duplicates/removes a variant which is next to an identical variant.
  - **Mutation(M)** mutates one variant in another.
  - **Insertion(I)/Deletion(D)** inserts/removes a variant, but without constraint.
- Non-unitary operations:
  - **Generation(G)/Compression(K)** generates/compresses an **arch** from/to its single ancestor variant.
- A cost is associated to each operation. Alignment cost is the sum of its operations costs.

Operations costs

- The model is symmetric: I = D and A = C.
- Observed much higher relative frequency of amplifications and contractions compared to other events $\Rightarrow$ A, C $<$ M, D, I.
- A deletion can also be obtained by a mutation plus a contraction. We have either:
  1. Hypothesis 1 **(H1)**: D $>$ M + C and I $>$ A + M
  2. Hypothesis 2 **(H2)**: D $\leq$ M + C and I $\leq$ A + M
- **Theorem**: Under H1 or H2, alignment cost is a metric distance.

Problem definition

Problem: Let s and r be two ms maps of length n and m, find an *optimal global* alignment between s and r under the single step evolutionary model.

Related works

- Alignment with duplication [Benson 97]:
  - Alignment of two sequences that may contain several tandem repeats; operations are substitutions, indels and duplications.
  - Differences with our approach: duplication relates the two sequences while for us a duplication is internal to one map; repeated units are unknown.
- Duplication history reconstruction [Benson et Dong 99, Elemento al 01, Tang al 01, Elemento 02, Jaitly 02].

Outline

1. Tandem repeats evolution, and minisatellite maps.
2. Evolutionary model.
3. Arches.
4. Alignment algorithm.
5. Applications.

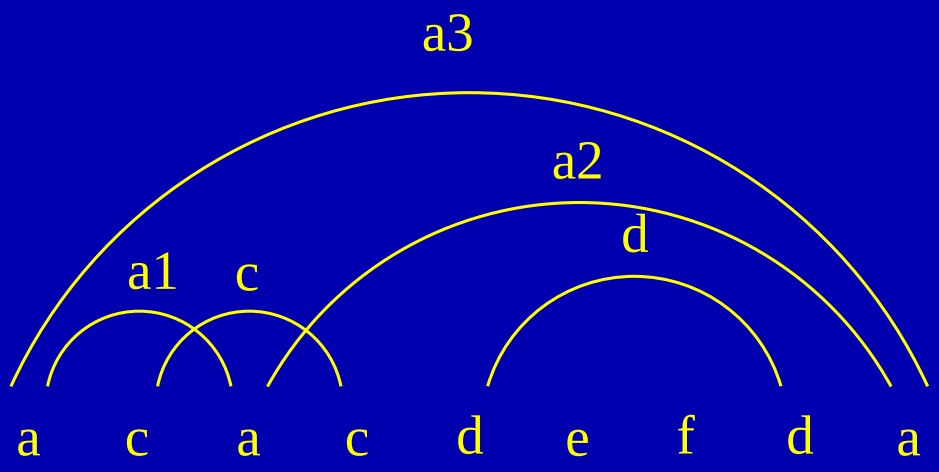
Arches

Let s be a map.

Def: an **arch** of s is a subsequence of s whose first and last variants are identical.

Def: an arch is **simple** if its internal variants occur only once and **complex** otherwise.

Ex: s=acacdefda



Def: two arches are **compatible** if they do not cross or share a same first foot or a same last foot.

Generate/Compress an Arch

- Generation and Compression are symmetrical.
- Generate a simple arch: from A to ABCA

First method					Second method				
	A					A			
Amplification	A	A			Amplification	A	A		
Mutation	A	B			Amplification	A	A	A	
Amplification	A	B	B		Amplification	A	A	A	A
Mutation	A	B	C		Mutation	A	B	A	A
Amplification	A	B	C	C	Mutation	A	B	C	A
Mutation	A	B	C	A					

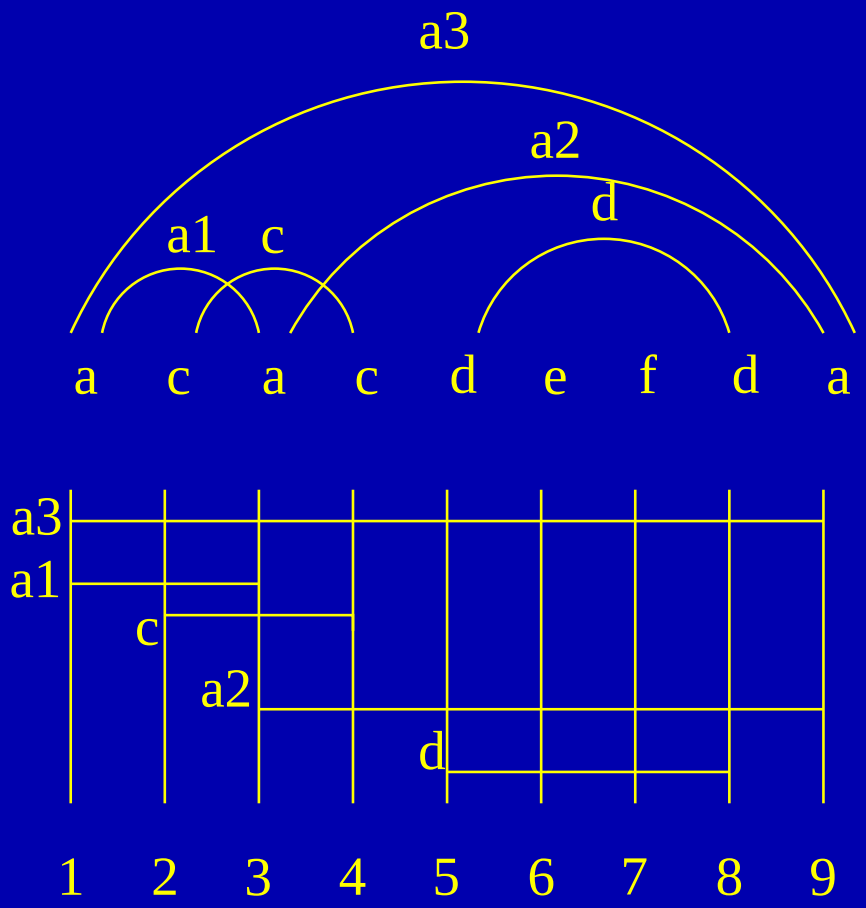
Identical variants at the extremities allow to avoid one mutation.

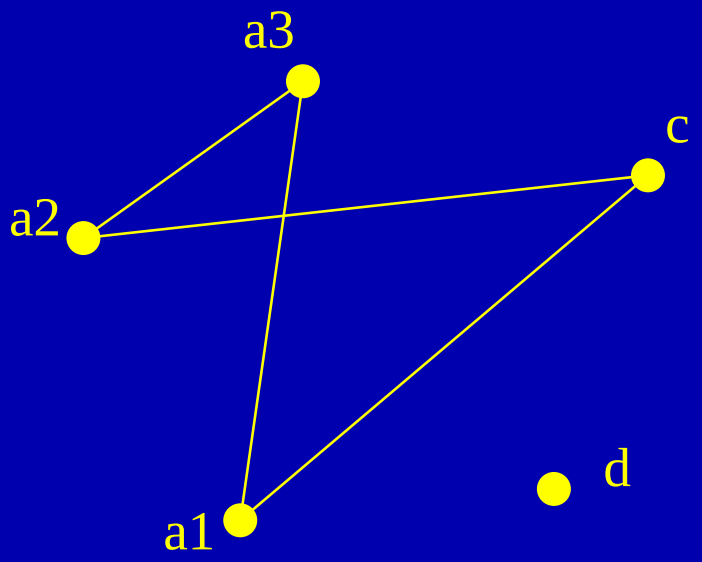
- Non-commutativity of operations.
- **Theorem:** the optimal generation/compression of an arch uses the largest subset of compatible inner arches.

Overlap graphs

- Map of size n, arches are subintervals in [1,n].
- Compatibility relation defines an overlap graph.
- **Theorem**: Finding the largest subset of compatible inner arches is equivalent to finding the max stable in an overlap graph.
- Complexity of this procedure is $O(V^2)$ with V the number of vertices of the graph [Apostolico et al. 92].
- Maximal number of arches is $\Theta(n^2)$.

Example



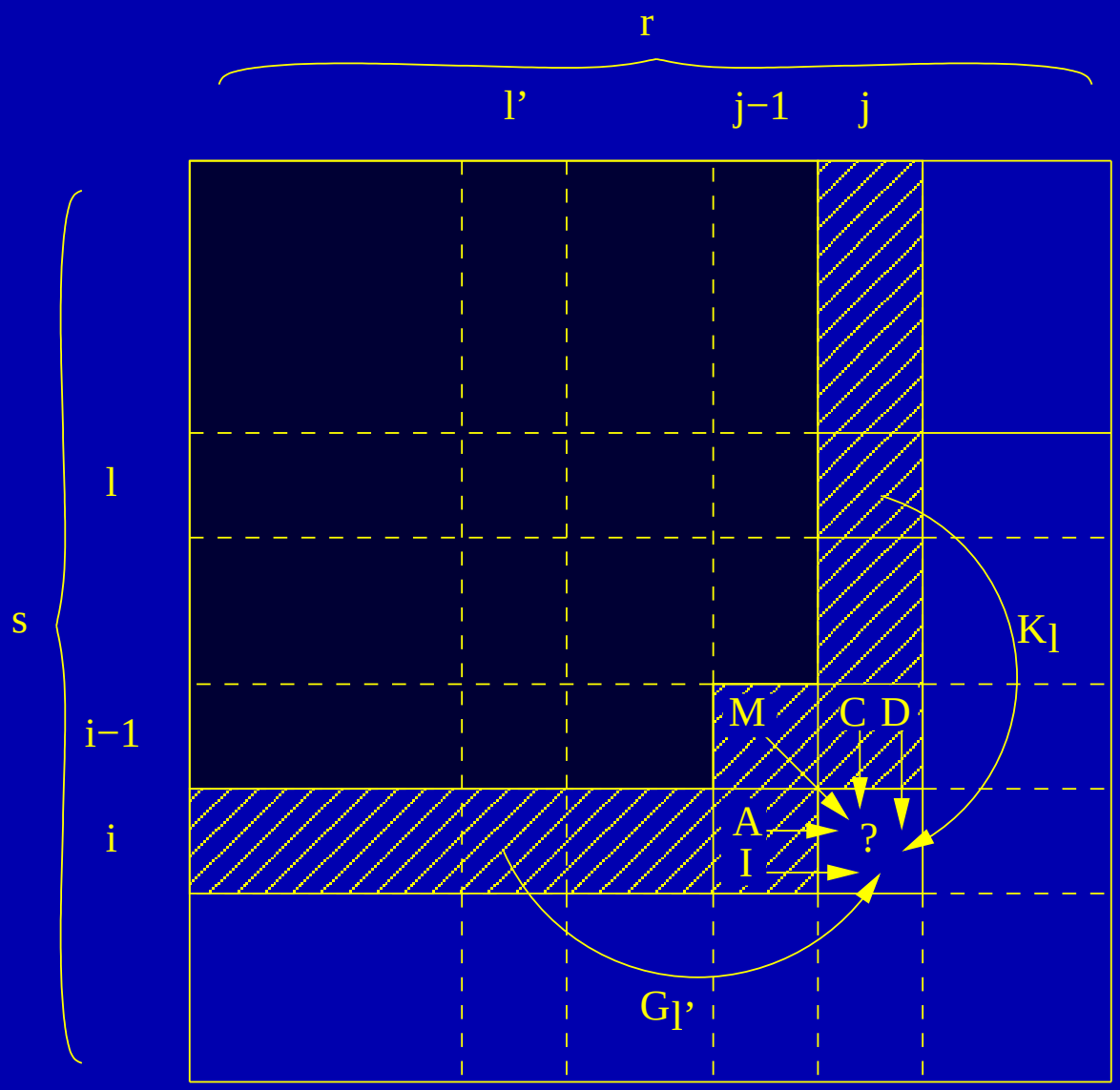


How to convert the problem of finding the max subset of compatible arches in the problem of finding the max stable set in an overlap graph

Outline

1. Tandem repeats evolution, and minisatellite maps.
2. Evolutionary model.
3. Arches.
4. Alignment algorithm.
5. Applications.

Dynamic Programming Dependencies



Recurrence

$$\mathscr{A}[i, j] = min \begin{cases} \mathscr{A}[i - 1, j - 1] + M(s[i],r[j]) & \text{Mute} \\ \mathscr{A}[i - 1, j] + C & \text{Contract} \\ \quad \textbf{iff } s[i - 1] = s[i] \textbf{ or } s[i]=r[j] & \\ \mathscr{A}[i - 1, j] + S & \text{Delete} \\ \mathscr{A}[i, j - 1] + A & \text{Amplify} \\ \quad \textbf{iff } r[j - 1] = r[j] \textbf{ or } s[i]=r[j] & \\ \mathscr{A}[i, j - 1] + I & \text{Insert} \\ \mathscr{A}[l, j] + K(s[l, i]) & \text{Compress an arch} \\ \quad \forall\, \textbf{l} \in \textbf{[1, i - 2] such that s[l] = s[i]} & \\ \mathscr{A}[i, l'] + G(r[l',j]) & \text{Generate an arch} \\ \quad \forall\, \textbf{l'} \in \textbf{[1, j - 2] such that r[l'] = r[j]} & \end{cases}$$

Complexity

Theorem: If p = max(n, m), overall complexity is $O(p^4)$ time.

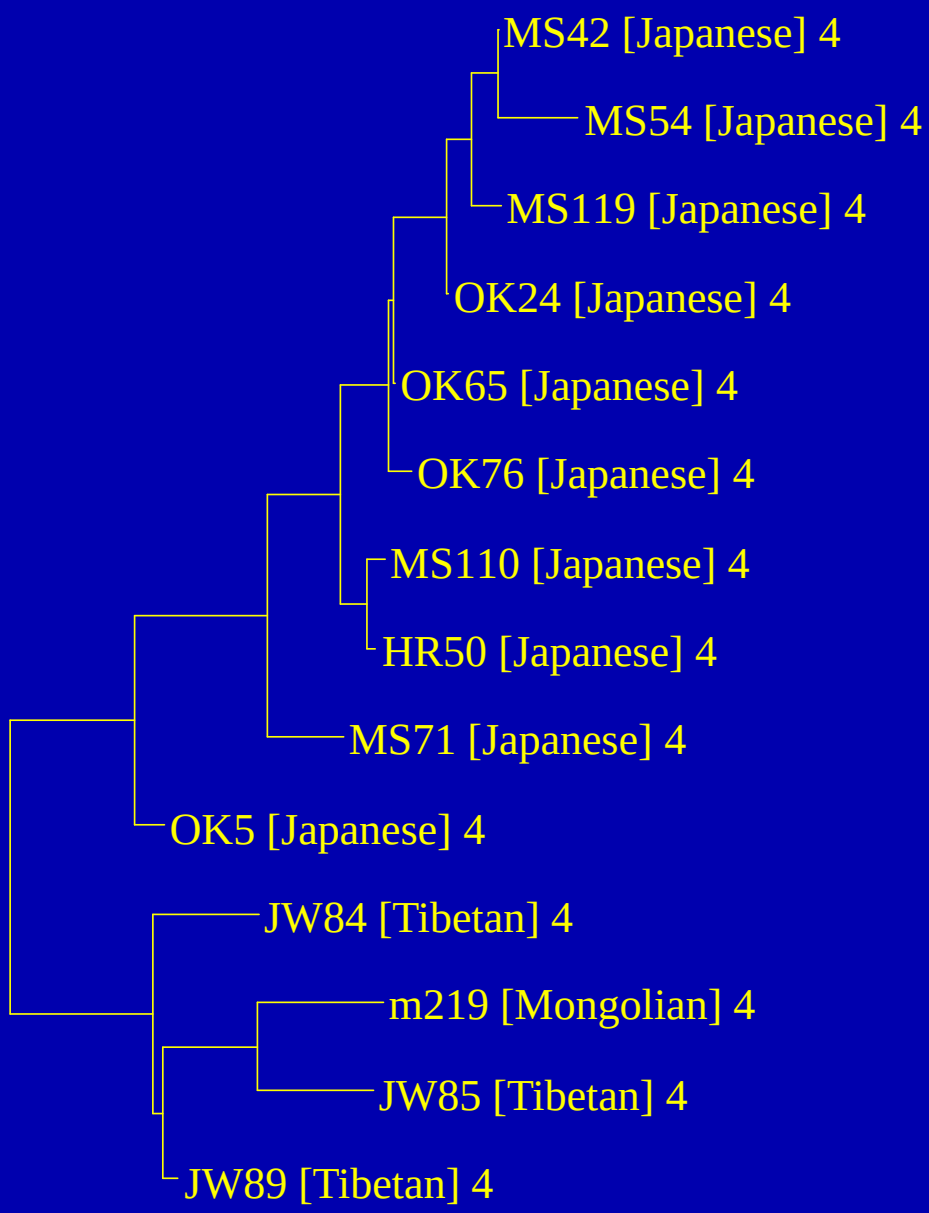
Outline

1. Tandem repeats evolution, and minisatellite maps.
2. Evolutionary model.
3. Arches.
4. Alignment algorithm.
5. Applications.

Minisatellite MSY1

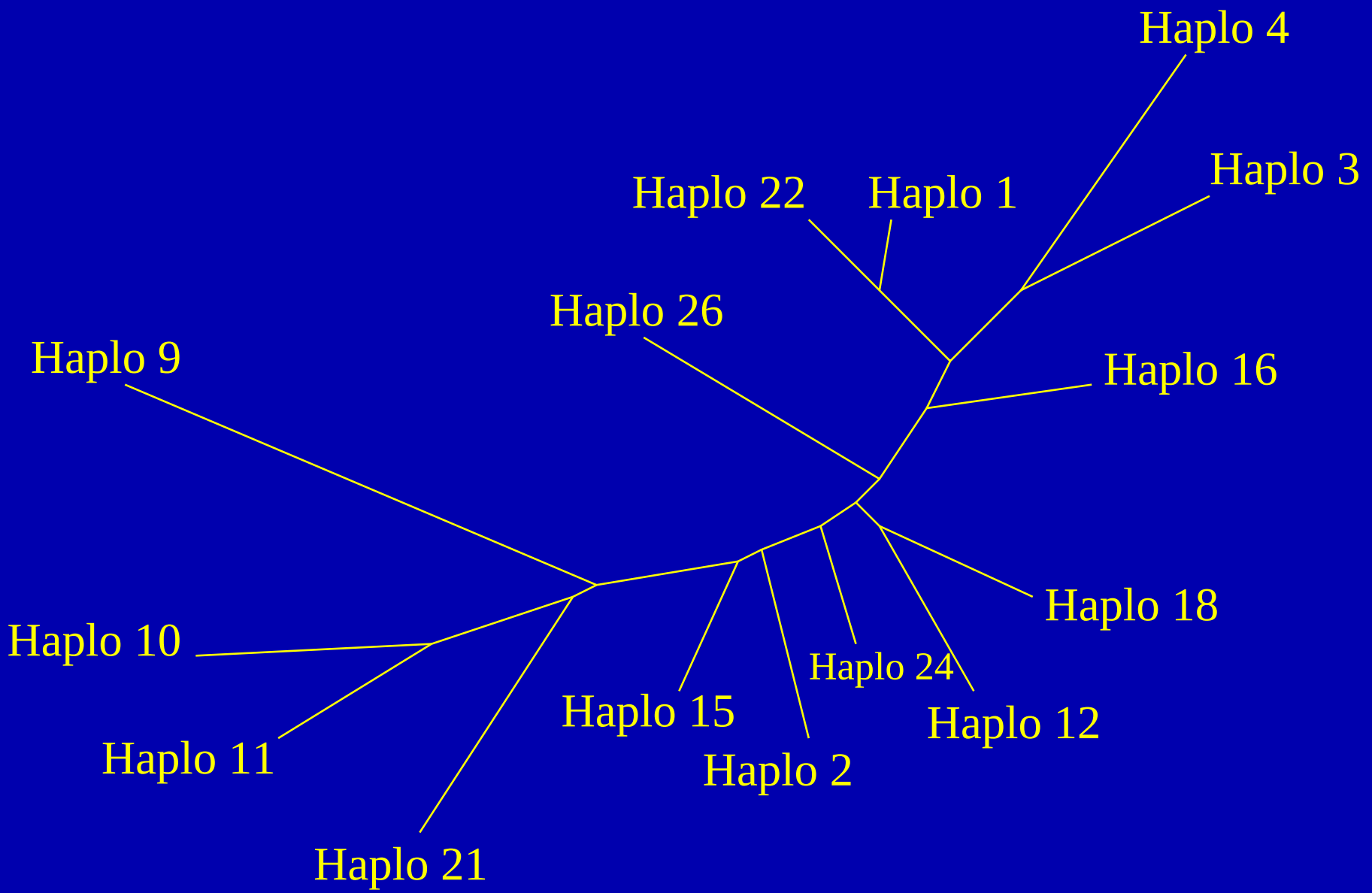
- MSY1 is a ms on the Y chromosome. Unit length 25bp, 5 variants differ by at most 4 residues. Evolution simplified: no exchange between alleles [Jobling et al 98].
- Maps for 690 men taken in $\neq$ populations distributed in 24 haplogroups.
- Is MSY1 an appropriate marker to study the evolution of these haplogroups? or of the subpopulations in an haplogroup? or of individuals?
- Experiments: Compute all pairwise alignments and construct a phylogenetic tree with BioNJ from our distance matrix.

Example of a phylogenetic tree

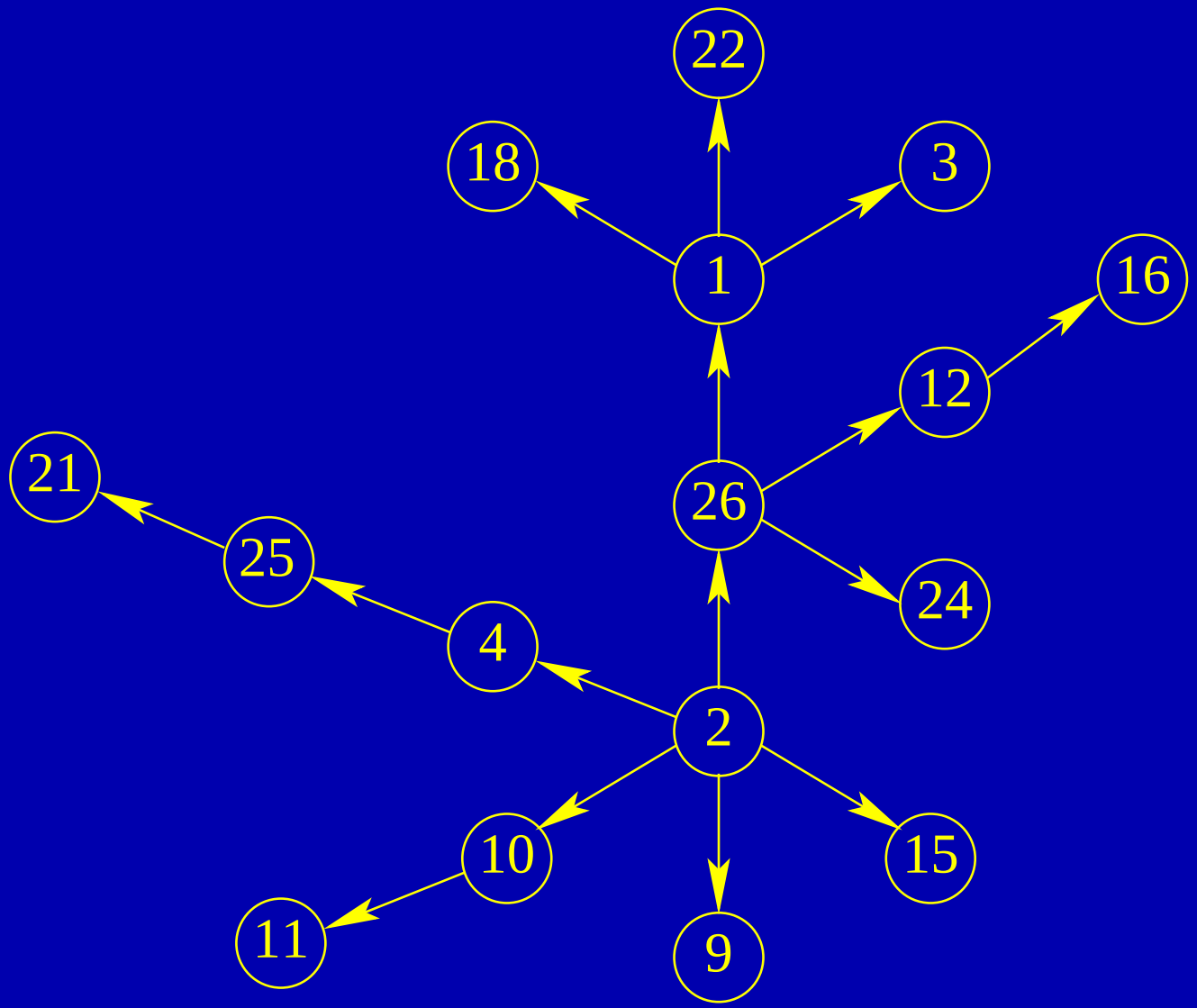


Haplogroup 4 phylogenetic tree.

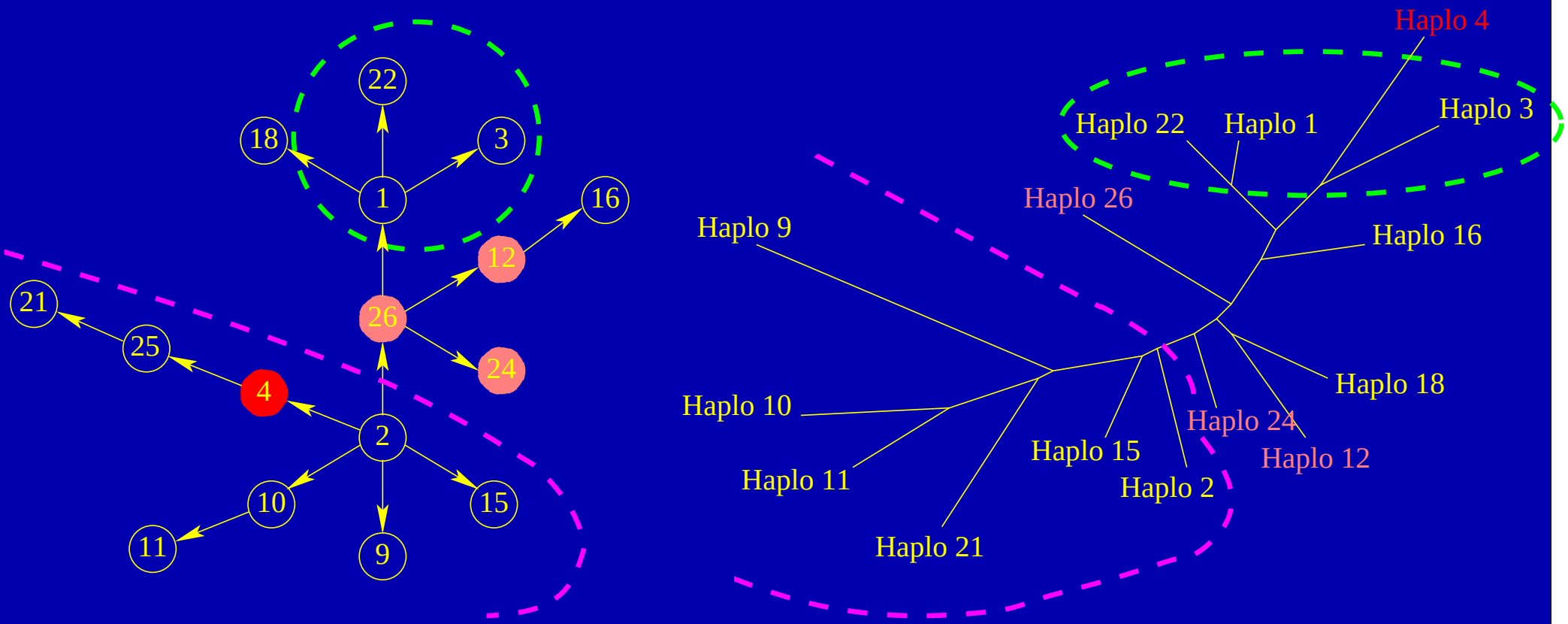
Haplogroups tree



Haplogroups tree (2)



Haplogroups trees comparison



Conclusion and future works

- A new and specific alignment method for minisatellites maps.
- Extensions: multiple amplification/contraction.
  1. arity > 1: $\cdots$ cgg <u>att</u> tga $\cdots \longrightarrow \cdots$ cgg **att att att** tga $\cdots$
  2. order > 1: $\cdots$ cgg <u>att tga</u> $\cdots \longrightarrow \cdots$ cgg **att tga att tga** $\cdots$
- Applicability for different biological purposes and on different ms.