Mining the semantics of genome super-blocks to infer ancestral architectures

Macha Nikolski

macha@labri.fr

07/10/2008

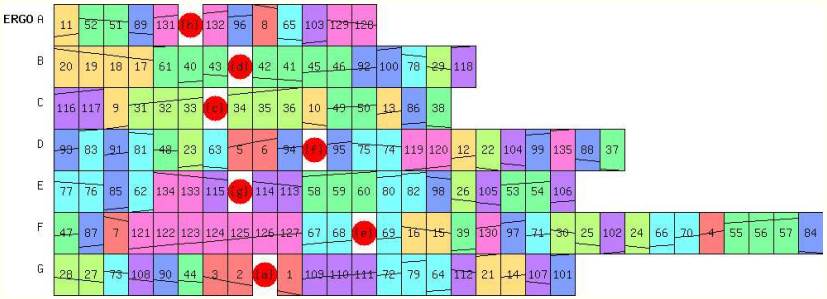
Introduction

Challenge : Uncovering principal events that punctuate the evolution of species
Approach : Plausible genome architectures of ancestral genomes

Two-fold problem :

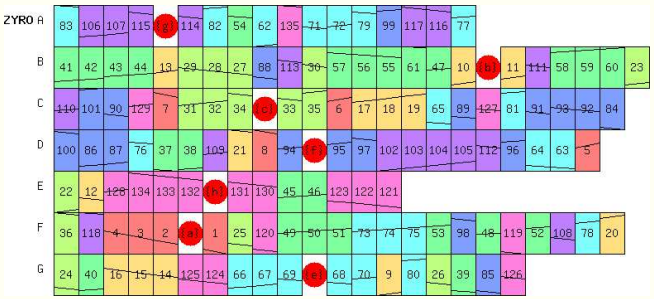
- determine ancestral architectures
- trace the rearrangement events that lead from the ancestors to contemporary genomes

Modeling evolution

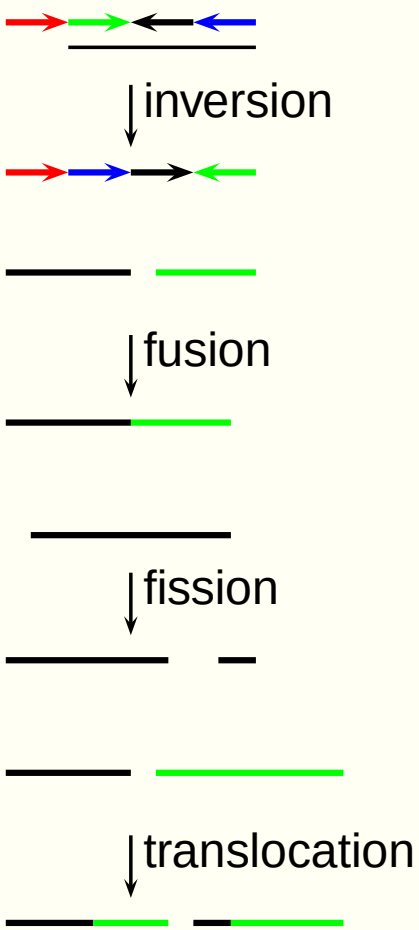


- rearrangements
- content change

common ancestor ?



Hannenhalli and Pevzner theory rearrangement operations



Mathematical vs. experimental approach

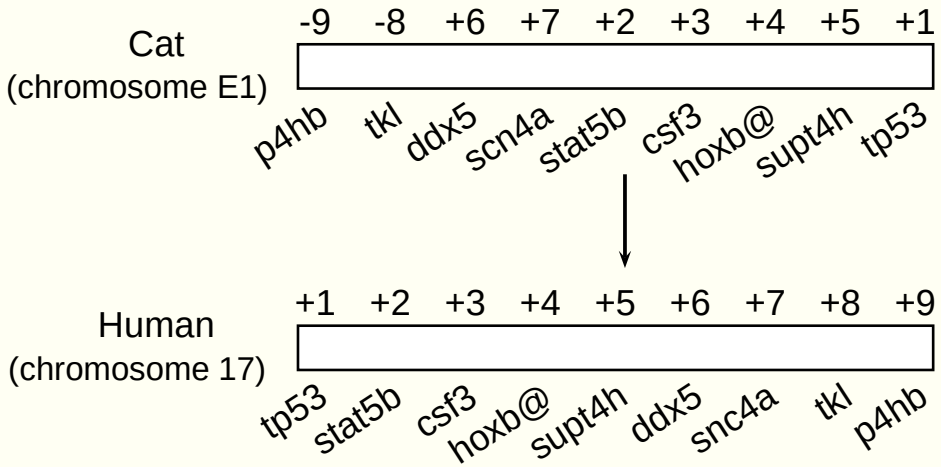
Results from two techniques do not necessarily agree

Rearrangement distance	Chromosomal painting
human, mouse, rat and chicken	Eutherian clade ($\approx$ 80 sp.)
genome sequences	hybridization of DNA probes
gene	$\approx$ 4 Mb
Bourque & Pevzner 2002	Froenike 2006
Bourque & Pevzner 2006	Rocchi 2006

Possible solution : integrate more biological knowledge into the mathematical approach

Hannenhalli and Pevzner theory

- **Signed permutation model :**



a genome

=

a set of signed permutations

- **Method** : mimicking multichromosomal rearrangement operations by reversals on a single permutation

genome Π ⟨ 8 ⟩ ⟨ 1 2 3 4 ⟩ ⟨ 5 6 7 ⟩

Reversal ⟨ 8 ⟩ ⟨ 1 -4 -3 -2 ⟩ ⟨ 5 6 7 ⟩

Translocation ⟨ 8 ⟩ ⟨ 1 -6 -5 ⟩ ⟨ 2 3 4 7 ⟩

Translocation ⟨ 6 -1 ⟩ ⟨ -8 -5 ⟩ ⟨ 2 3 4 7 ⟩

Fusion

genome Γ ⟨ 6 -1 -7 -4 -3 -2 ⟩ ⟨ 5 8 ⟩ ⟨ ⟩

Ancestors as median genomes

Formulation as *median genome* problem :

Given $G_1, ..., G_N$, find M such that for a distance d

$$\sum_{i=1}^{N} d(M, G_i) \text{ is minimal}$$

Different distances : rearrangement, breakpoint, double cut and join

This problem is NP-complete event for $N = 3$

- breakpoint distance (Bryant 1998, Pe'er & Shamir 1998)
- rearrangement distance (Caprara 1999, Caprara 2003)

Limitations

Misleading to speak of an *ancestral genome* $\Rightarrow$ median genome

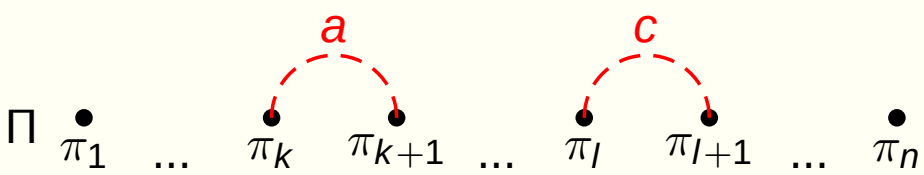
Algorithmic and interpretation problems

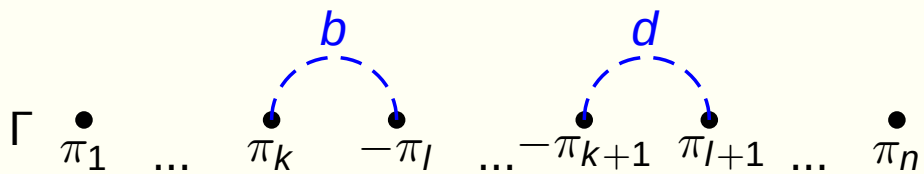
- Computationally intractable, in practice need heuristics
- High number of equivalent solutions (Bourque & Pevzner 2002, Eriksen 2007)

Ideas

- look for common features present in ancestral genome architecture
- (re-)introduce biologically pertinent features : breakpoints

Adjacencies, breakpoints and frequencies





breakpoints : $a, c \in \Pi$ and $b, d \in \Gamma$

Particular case of telomeres $0.\pi_1$ and $\pi_n.0$

Example

$$G_1 = \{1\ 2\ 3\ 4,\ 5\ 6\} \qquad G_2 = \{1\ 2\ 3\ 4,\ -5\ 6\}$$
$$G_3 = \{3\ 1\ 4\ 2\ -5,\ 6\} \qquad G_4 = \{2\ 1\ 3\ 4,\ 5\ 6\}$$

frequency	adjacencies
4	6.0
3	3.4, 0.5, 4.0
2	5.6, 2.3, 1.2, 0.1
1	−5.6, 2.-5, 4.2, 1.4, 1.3, 3.1, 2.1, 0.6, 5.0, 0.3, 0.2

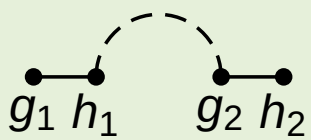
Adjacency graph

Hannenhalli & Pevzner 1995 $\pi_i \longrightarrow \begin{cases} g = 2\pi_i - 1 \\ h = 2\pi_i \end{cases}$

Denoted : $\pi_i.\pi_j$ by $(g_1\, h_1).(g_2\, h_2)$
$\pi_i. - \pi_j$ by $(g_1\, h_1).(h_2\, g_2)$

Example

The adjacency graph for a set $A = \{(g_1\, h_1).(g_2\, h_2)\}$:



- 4 vertices g_1, h_1, g_2 and h_2
- two edges stand for elements $e_1 = (g_1, h_1)$ and $e_2 = (g_2, h_2)$.
- one edge stands for the adjacency $e_3 = (h_1, g_2)$

Intuition

For a set of genomes $\{G_i\}$, the higher is the frequency of an adjacency, the higher is the probability that it should be present in a median genome.

Build *partial assemblies* of median genomes

1. Build a partition P of adjacencies where each part is composed of inter-dependent adjacencies. P is partially ordered by adjacency frequency of the parts' elements.
2. Inspect P in decreasing order of its parts, and construct the partial assemblies by favoring adjacencies with higher frequency.

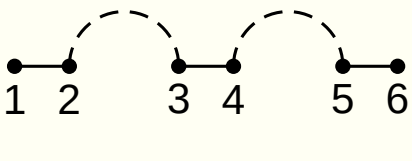
Assemble these partial assemblies into potential medians

Dependent adjacencies

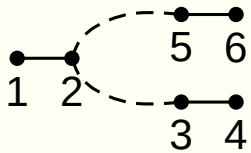
$a = (g_1^a\, h_1^a).(g_2^a\, h_2^a)$ and $b = (g_1^b\, h_1^b).(g_2^b\, h_2^b)$
$G = (V, E)$ the adjacency graph for $\{a, b\}$

Defi nition

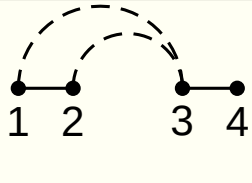
We say that a and b complement each other if if either (i) $\exists\, v_1, v_2 \in V$ such that $d(v_1) = d(v_2) = 1$ and $\forall v \neq v_i, i \in [1, 2]$ we have $v \neq 0$ and $d(v) = 2$, or (ii) $\exists v \in V$ such that $v = 0$ and $\forall v \in V$ we have $d(v) = 2$.
We say that a and b contradict each other if either (i) $\exists\, v \in V$ such that $d(v) > 2$, or (ii) $\forall v \in V$ we have $v \neq 0$ and $d(v) = 2$.



complement



vertex contradiction



cycle contradiction

Adjacency choice for the ancestral genome architecture $u(a) > 1$:

- complementary adjacencies : multiple agreement
- contradictory adjacencies : multiple breakpoints

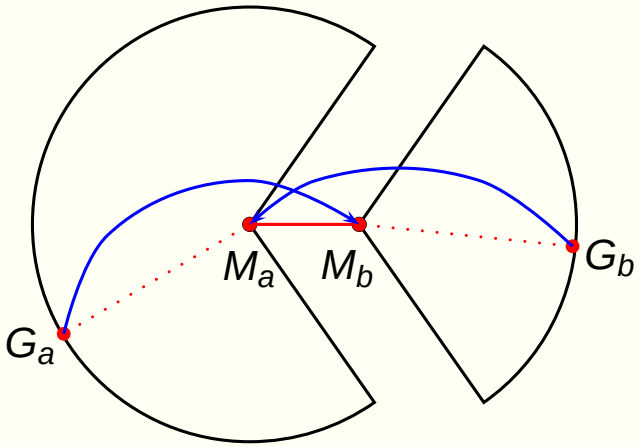
Relative frequency

N genomes $\{G_i\}$, d rearrangement distance
C the set of all contradictory adjacencies
M_a and M_b are identical up to two adjacencies

Lemma

For any pair of adjacencies $\{a, b\} \in C$ and two genomes M_a and M_b identical up to 2 adjacencies with $a \in M_a$ and $b \in M_b$, it holds that

$$\left| \sum_i^N d(M_a, G_i) - \sum_i^N d(M_b, G_i) \right| \leq N.$$



If $u(a) > u(b)$
$\sum_i^N d(M_a, G_i) - \sum_i^N d(M_b, G_i) \ll N$
Similarly for the breakpoint distance

Groups of adjacencies

$\mathcal{P}(A)$ be a partition of $\mathcal{A}$, set of all adjacencies.
$\mathcal{P}_0(\mathcal{A})$: elementary cycles without 0 + singletons

Merging of parts $\sqcup$ defines a partition of $\mathcal{A}$ such that for any $p \in \sqcup(\mathcal{P}(\mathcal{A}))$

- $\exists\, p_1 \in \mathcal{P}(A)$ s.t. $p = p_1$ or
- $\exists\, p_1, p_2 \in \mathcal{P}(\mathcal{A})$ s.t. $p = p_1 \cup p_2$ and moreover $\exists\, a \in p_1$ and $\exists\, b \in p_2$ s.t. $u(a) = u(b) = u(p_1) = u(p_2)$ and either a and b are dependent or a and b participate in a cycle $c \in G$ without vertex $v = 0$ s.t. $\forall v \in c$ we have $u(v) \geq u(a)$.

Defi nition

A group g is a part of $\sqcup^n(\mathcal{P}_0(A))$, where $\sqcup^n(\mathcal{P}_0(A))$ is the fixed point of $\sqcup$.

Groups of adjacencies, continued

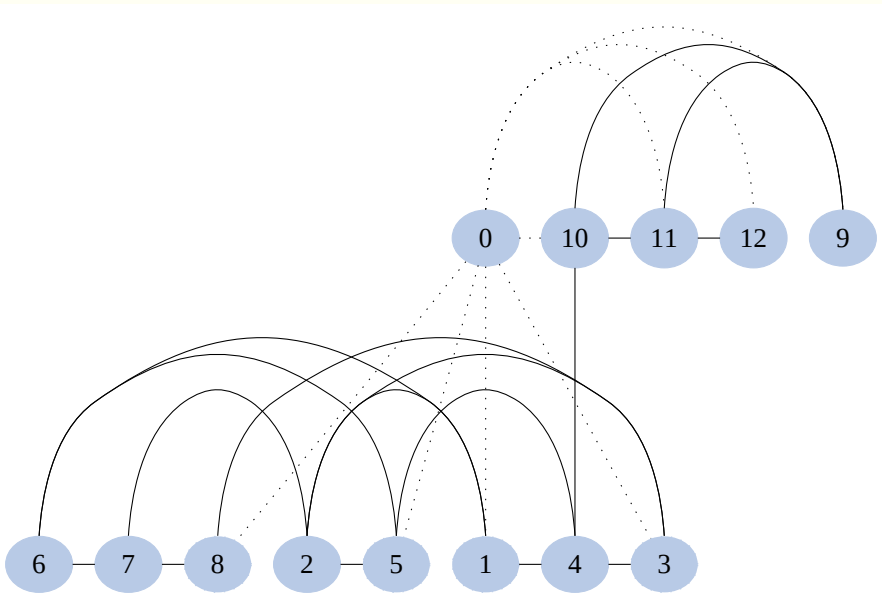
Example

$$G_1 = \{1\ 2\ 3\ 4,\ 5\ 6\} \qquad G_2 = \{1\ 2\ 3\ 4,\ -5\ 6\}$$
$$G_3 = \{3\ 1\ 4\ 2\ -5,\ 6\} \qquad G_4 = \{2\ 1\ 3\ 4,\ 5\ 6\}$$

Groups of adjacencies, continued

Example

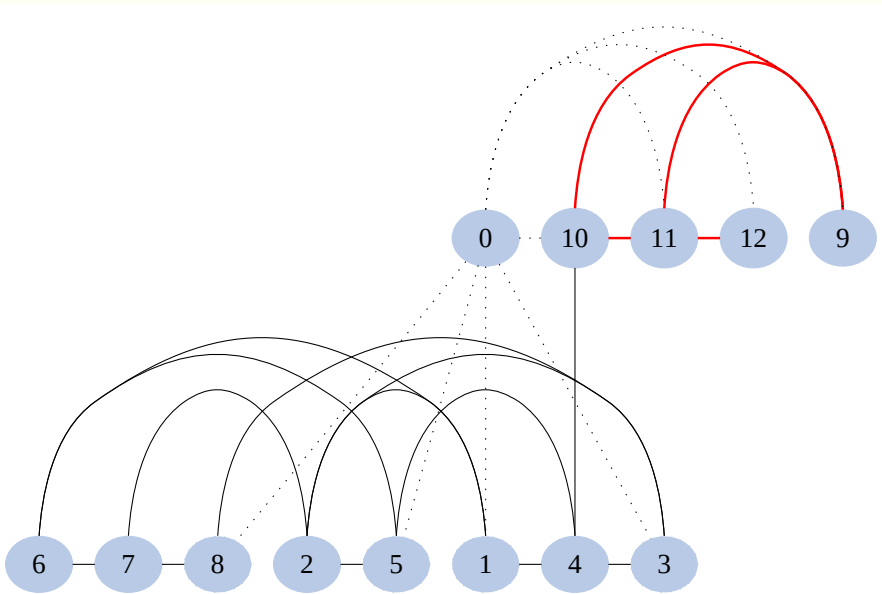
$G_1 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$ $G_2 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (10\ 9)(11\ 12)$
$G_3 = \{(5\ 6)(1\ 2)(7\ 8)(3\ 4)(10\ 9),\ (11\ 12)$ $G_4 = \{(3\ 4)(1\ 2)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$



Groups of adjacencies, continued

Example

$G_1 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8), (9\ 10)(11\ 12)\}$ $G_2 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8), (10\ 9)(11\ 12)\}$
$G_3 = \{(5\ 6)(1\ 2)(7\ 8)(3\ 4)(10\ 9), (11\ 12)$ $G_4 = \{(3\ 4)(1\ 2)(5\ 6)(7\ 8), (9\ 10)(11\ 12)\}$

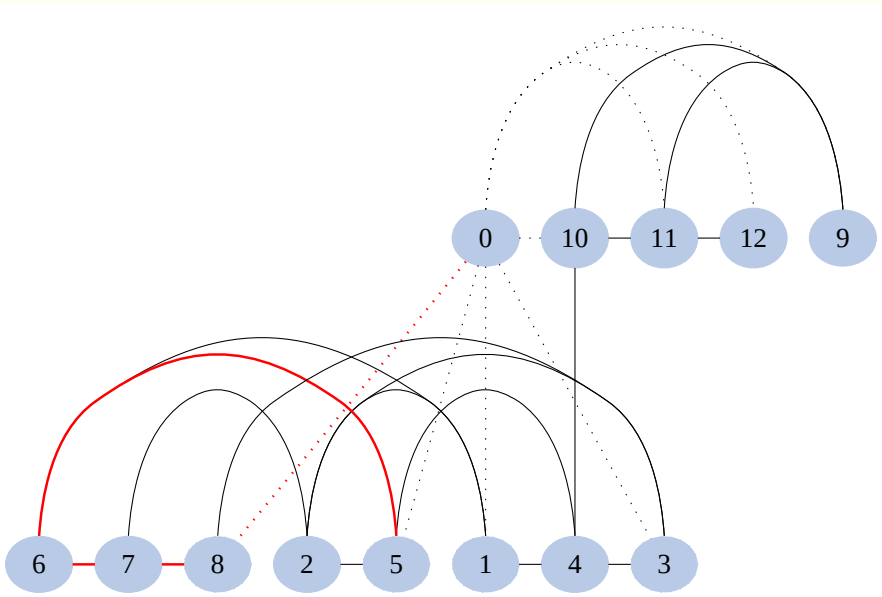


$\mathcal{P}_0(\mathcal{A}) = \{(9\ 10).(11\ 12); (10\ 9).(11\ 12)\} \cup$ singletons

Groups of adjacencies, continued

Example

$G_1 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$ $G_2 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (10\ 9)(11\ 12)$
$G_3 = \{(5\ 6)(1\ 2)(7\ 8)(3\ 4)(10\ 9),\ (11\ 12)$ $G_4 = \{(3\ 4)(1\ 2)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$



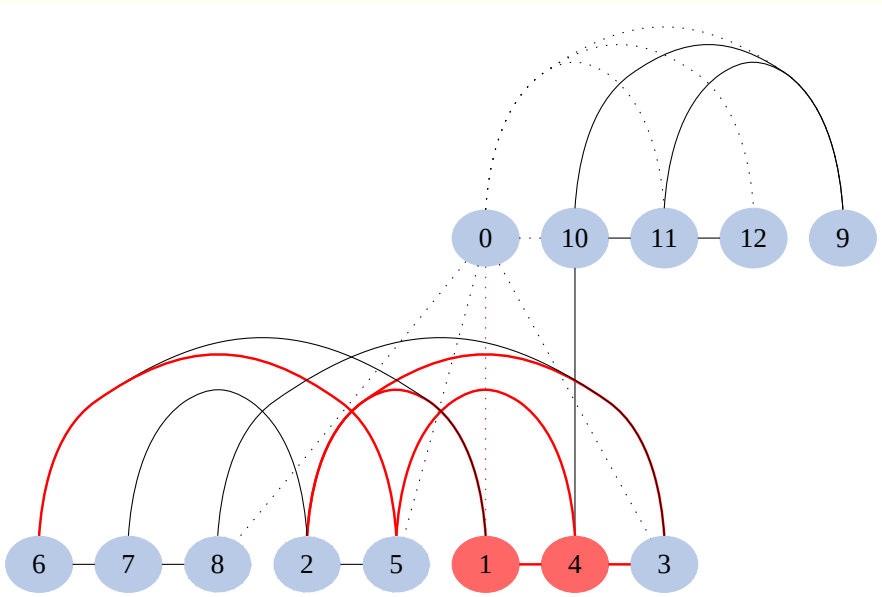
$\mathcal{P}_0(\mathcal{A}) = \{(9\ 10).(11\ 12); (10\ 9).(11\ 12)\} \cup$ singletons

$\mathcal{P}_1(\mathcal{A}) = \mathcal{P}_0(\mathcal{A}) \cup \{(5\ 6).(7\ 8); (7\ 8).0\} \cup$ singletons

Groups of adjacencies, continued

Example

$G_1 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$ $G_2 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (10\ 9)(11\ 12)$
$G_3 = \{(5\ 6)(1\ 2)(7\ 8)(3\ 4)(10\ 9),\ (11\ 12)$ $G_4 = \{(3\ 4)(1\ 2)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$



$\mathcal{P}_0(\mathcal{A}) = \{(9\ 10).(11\ 12); (10\ 9).(11\ 12)\} \cup$ singletons

$\mathcal{P}_1(\mathcal{A}) = \mathcal{P}_0(\mathcal{A}) \cup \{(5\ 6).(7\ 8); (7\ 8).0\} \cup$ singletons

$\mathcal{P}_2(\mathcal{A}) = \mathcal{P}_1(\mathcal{A}) \cup$

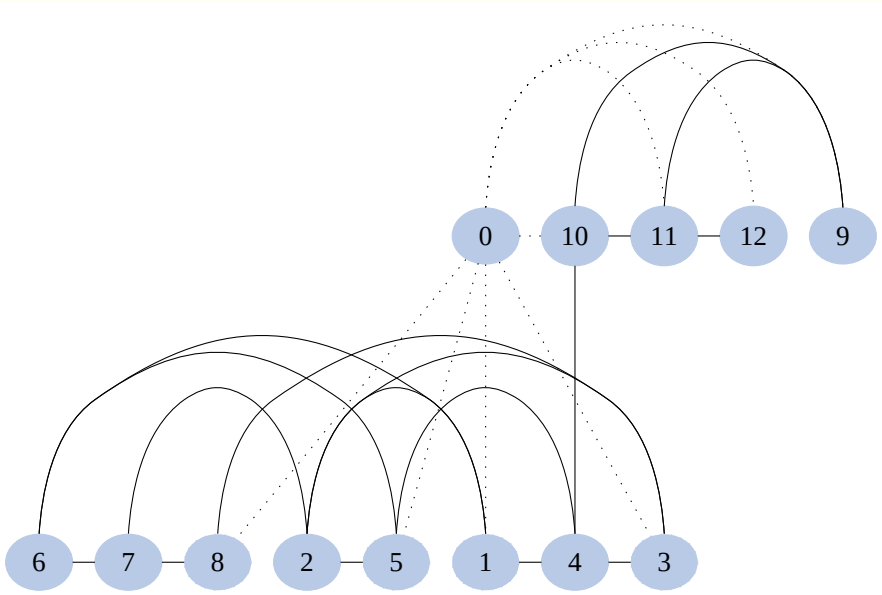
$\{0.(1\ 2), (1\ 2).(3\ 4), (3\ 4).(5\ 6), (2\ 1).(4\ 3)\} \cup$

singletons

Groups of adjacencies, continued

Example

$G_1 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$ $G_2 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (10\ 9)(11\ 12)$
$G_3 = \{(5\ 6)(1\ 2)(7\ 8)(3\ 4)(10\ 9),\ (11\ 12)$ $G_4 = \{(3\ 4)(1\ 2)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$



$\mathcal{P}_0(\mathcal{A}) = \{(9\ 10).(11\ 12); (10\ 9).(11\ 12)\} \cup$ singletons

$\mathcal{P}_1(\mathcal{A}) = \mathcal{P}_0(\mathcal{A}) \cup \{(5\ 6).(7\ 8); (7\ 8).0\} \cup$ singletons

$\mathcal{P}_2(\mathcal{A}) = \mathcal{P}_1(\mathcal{A}) \cup$

$\{(3\ 4).(5\ 6), (1\ 2).(3\ 4), 0.(1\ 2), (2\ 1).(4\ 3)\} \cup$

singletons

grp. freq.	adjacencies	grp. freq.	adjacencies
4	12.0(4)	2	10.11(2), 9.11(1)
3	6.7(3), 0.8(3)	2	4.5(2), 2.3(2), 0.1(2), 1.4(1)
3	0.9(3)		

Superblocks (intuition part 2)

Defi nition

A superblock is a set S of $n \geq 1$ adjacencies s.t. $\forall a, b \in S$, a does not contradict b, and there exists an order over S such that $\forall i \in [1, n)$, a_i complements a_{i+1}, and a_1, a_n are either independent or $a_1 = a_n = 0$. A partial assembly $\mathcal{P} = \{S_k\}$ is a set of superblocks such that $\forall k \neq l$ if $S_k \cap S_l \neq \emptyset \Rightarrow S_k \cap S_l = \{0\}$.

Lemma

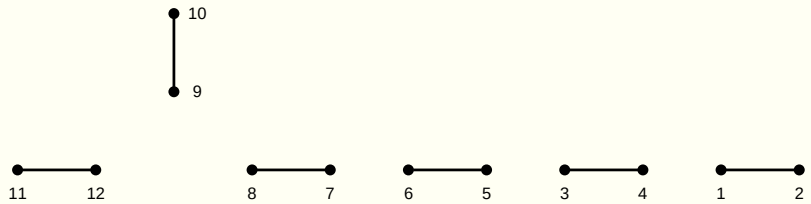
The adjacency graph $G = (V, E)$ of a partial assembly $\mathcal{P}$ is a graph such that (1) $\forall v \in V$, $d(v) \leq 2$, except for $v = 0$, and (2) any cycle in G contains 0.

Putting it all together

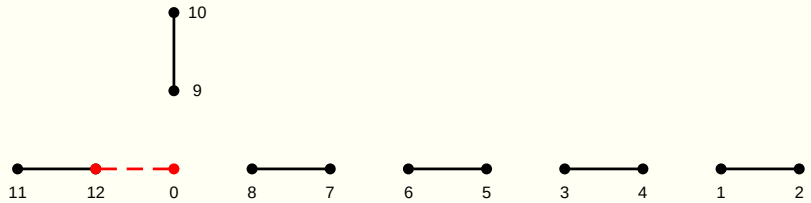
Example

$G_1 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$ $G_2 = \{(1\ 2)(3\ 4)(5\ 6)(7\ 8),\ (10\ 9)(11\ 12)$
$G_3 = \{(5\ 6)(1\ 2)(7\ 8)(3\ 4)(10\ 9),\ (11\ 12)$ $G_4 = \{(3\ 4)(1\ 2)(5\ 6)(7\ 8),\ (9\ 10)(11\ 12)\}$

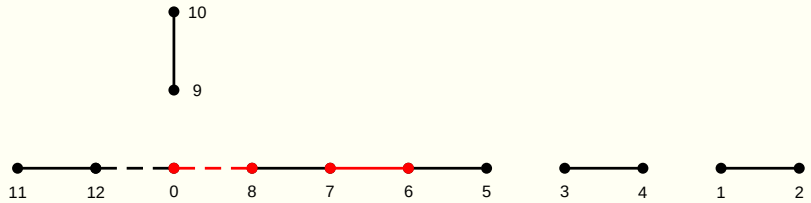
(a) Initial graph



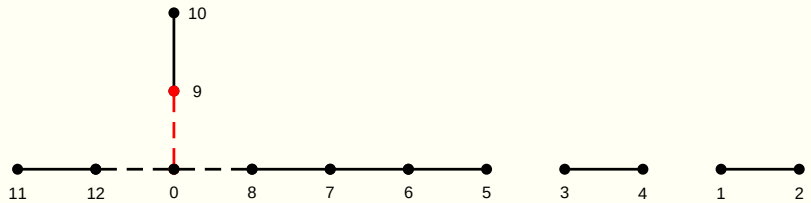
(b) Adding group $\{12.0\}, u = 4$:



(c) Adding group $\{6.7,\ 0.8\}, u = 3$

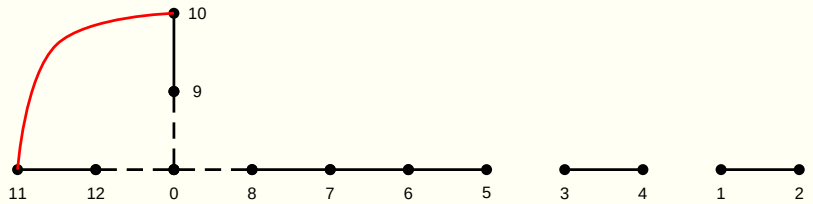


(d) Adding group $\{0.9\}, u = 3$

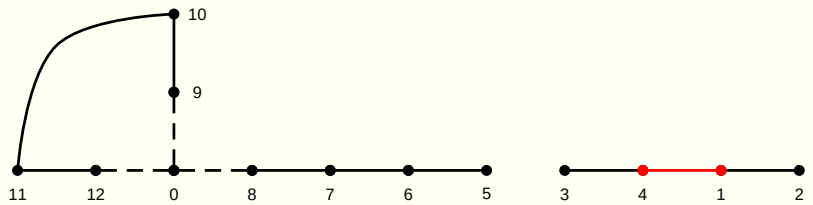


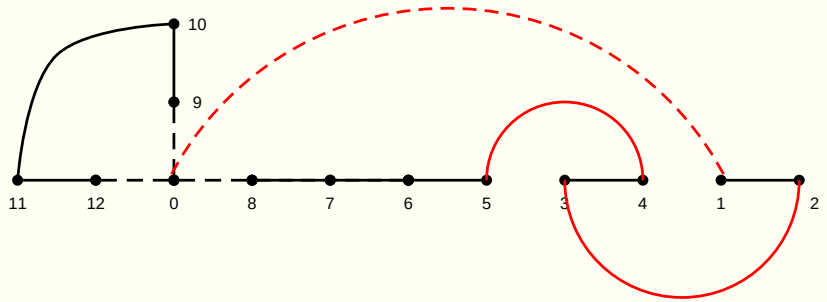
Putting it all together

(d) Adding group $\{10.11,\ 9.11\} = \{10.11\}, u = 2$



(e) Adding group $\{4.5,\ 2.3,\ 0.1,\ 1.4\} = \{1.4\}, \{4.5,\ 2.3,\ 0.1\}, u = 2$





Two solutions $M_1 = \{1\ 2\ 3\ 4\ ,\ 5\ 6\}$ and $M_2 = \{2\ 1,\ 3\ 4,\ 5\ 6\}$

having $\sum d(M_1, G_i) = 9$ and $\sum d(M_2, G_i) = 10$

Chromosomal rearrangements in Yeasts

Is it possible to study ?

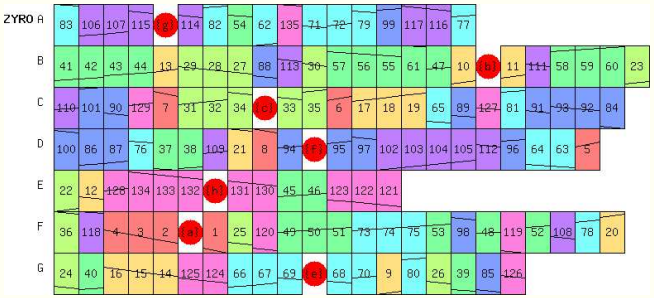
- Uniqueness of Génolevures data :
  complete genome sequences, availability of protein families
- Kluyveromyces clade :
  weak redundancy, synteny
- Additional information : positions of centromeres

$\Rightarrow$ we can specify markers for Kluyveromyces

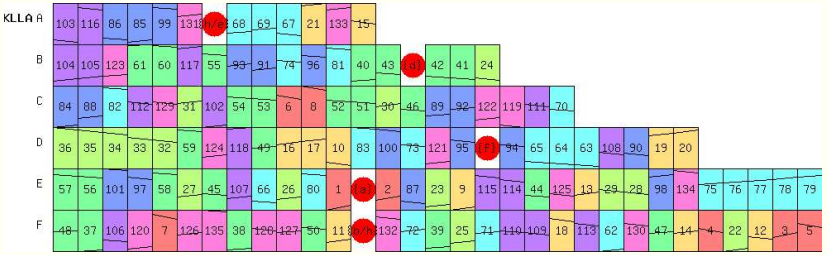
Species	Mnemonic
Kluyveromyces lactis	Klla
Kluyveromyces waltii	Klwa
Zygosaccharomyces rouxii	Zyro
Ashbya gossypii	Ergo
Kluyveromyces thermotolerans	Klth

Comparative maps for Kluyveromyces

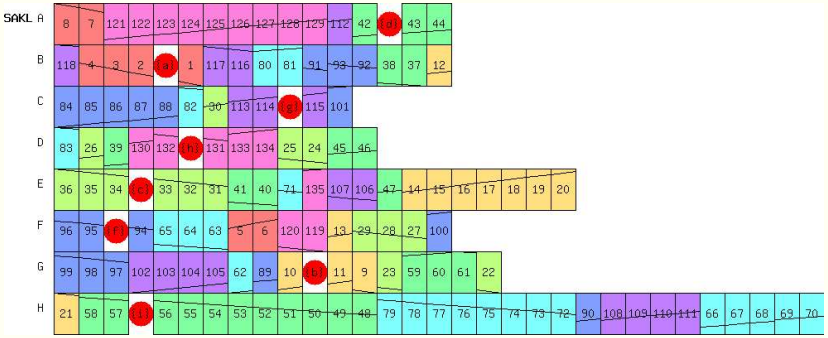
Zyro



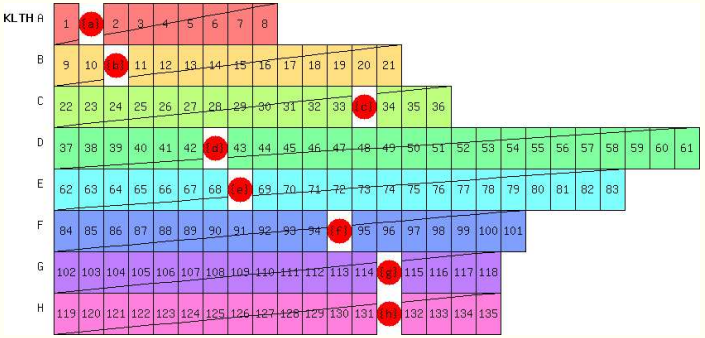
Klla



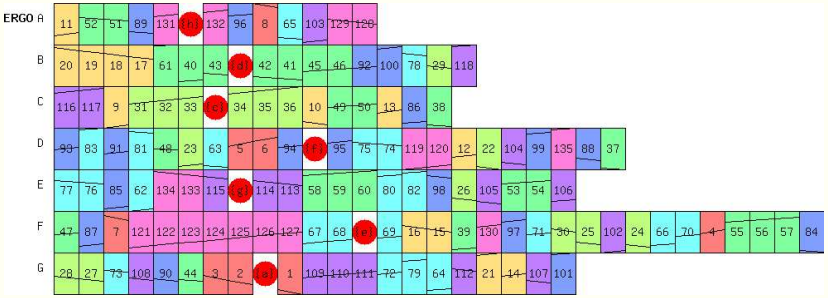
Sakl



Klth



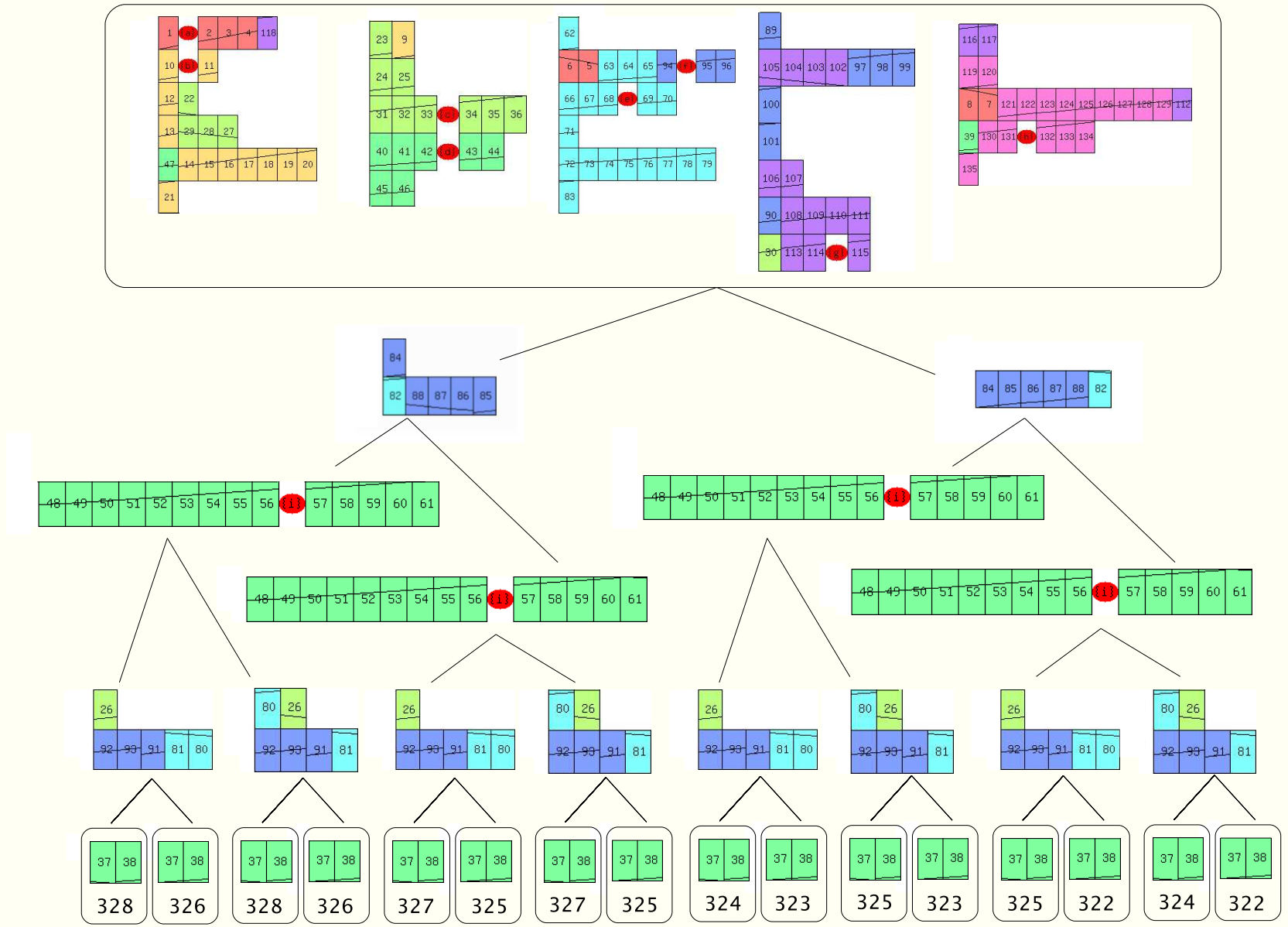
Ergo



Pairwise rearrangement distances

	Klth	Ergo	Klla	Sakl	Zyro
Klth	0	88	105	45	84
Ergo		0	109	85	101
Klla			0	98	115
Sakl				0	79
Zyro					0

Sharing tree of super-blocks



Further work

- An efficient implementation for building median genomes (Faucils)
- Bringing in more biological constraints
- Rearrangement tree (Steiner tree problem with unknown nodes)
- Rearrangement scenarios between two genomes
- Taking into the account duplications in genomes

Acknowledgements

Geraldine JEAN

Serge DULUCQ
Pascal DURRENS
Adrien GOEFFON
David SHERMAN
Nikolay VYAHHI

Génolevures consortium
(Jean-Luc Souciet coord.)

And thank you for your attention !