

Méthodes de comparaisons de génomes complets

Module Génomique Comparée - Bioinformatique

Eric Tannier, INSA 4BIM, 2017-2018

Eric.Tannier@univ-lyon1.fr

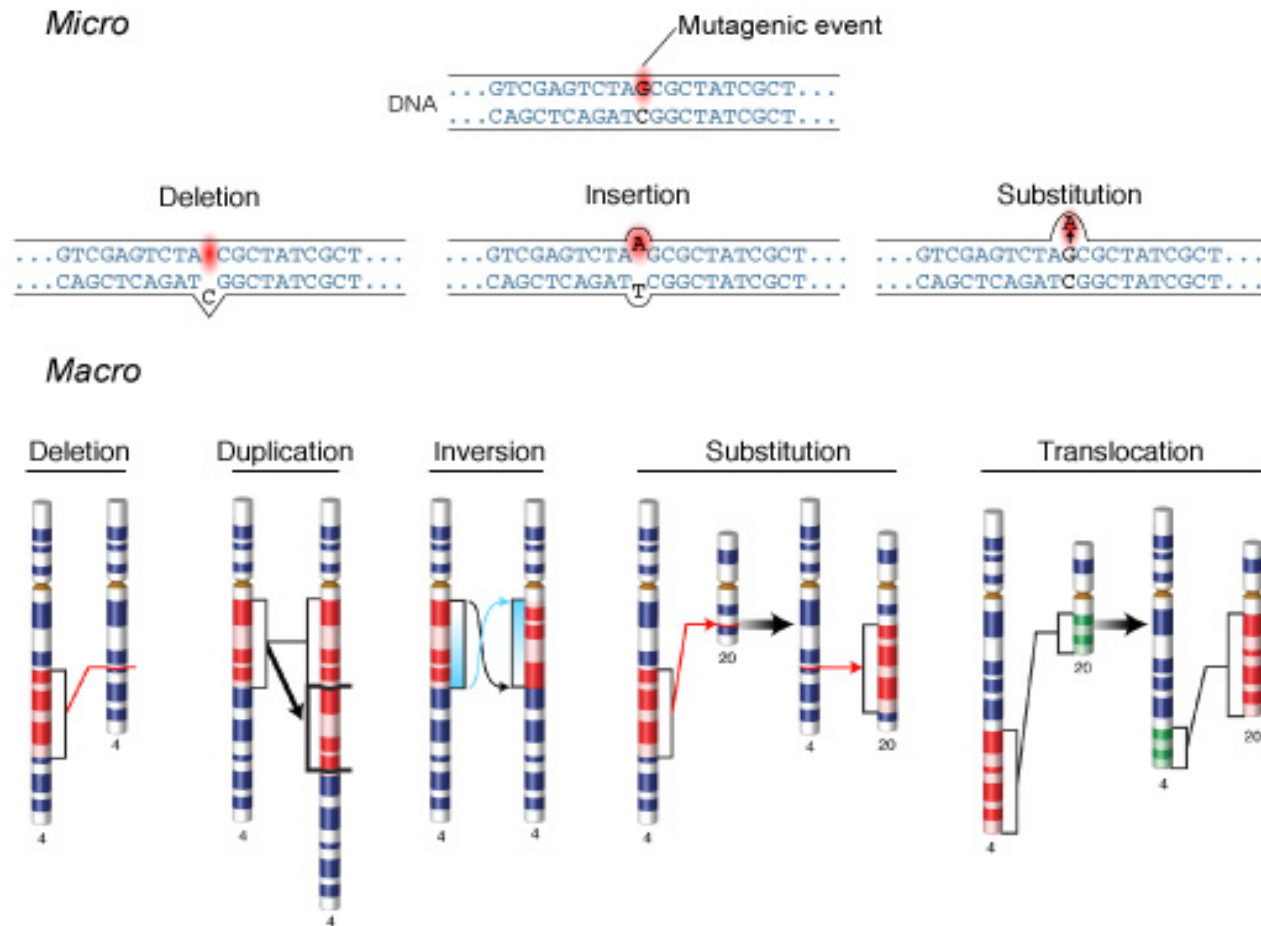
Deuxième série de cours :

Cours/TD mardi 30 janvier 8h, jeudi 1 février 8h,
Mardi 6 février 8h, vendredi 9 février 10h

Projet évaluant les deux parties de cours à rendre pour fin mars

Support (ce document) disponible à partir de ma page internet

Les génomes évoluent par divers types de mutations à différentes échelles



Autres échelles : éléments transposables, duplications de toutes tailles (macro ou micro) ?

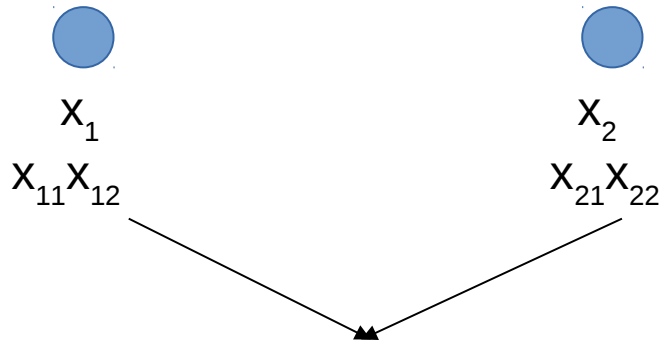
La méthode comparative : trouver les similarités et les différences (à toutes échelles) pour en déduire une histoire évolutive, les mutations

Du point de vue du traitement algorithmique, les mutations sont inégales

I/ Types de problèmes algorithmiques liés aux grandes mutations

1/ Lois de Mendel, 1865

Deux individus x_1 et x_2 portent chacun, pour un caractère, deux facteurs,



Gregor Mendel

$$(x_{11} + x_{12})(x_{21} + x_{22}) = x_{11}x_{21} + x_{11}x_{22} + x_{12}x_{21} + x_{12}x_{22}$$

Leurs descendants ont comme facteur $(x_{11}x_{21}, x_{11}x_{22}, x_{12}x_{21}, x_{12}x_{22})$

Avec les probabilités $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$

Relation entre le théorique et l'observé : système de dominance

Première loi, un caractère

$$(A+a)(A+a) = AA + 2Aa + aa$$

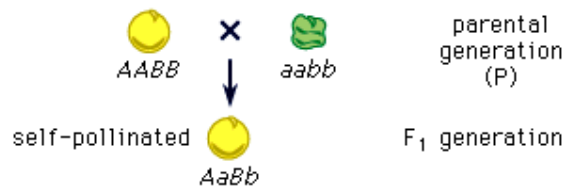
Observation 3*A , 1*a

















Deuxième loi, deux caractères

$$(A+a)^2(B+b)^2 =$$

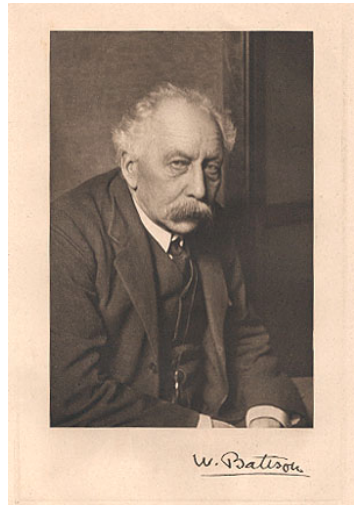
$$AABB + 2AABb + AAbb + 2AaBB + 4AaBb + 2Aabb + aaBB + 2aaBb + aabb$$

Observations 9*AB , 3*Ab , 3*aB , 1*ab



♀ \ ♂		pollen			
		<i>AB</i>	<i>Ab</i>	<i>aB</i>	<i>ab</i>
ovules	<i>AB</i>	 <i>AABB</i>	 <i>AABb</i>	 <i>AaBB</i>	 <i>AaBb</i>
	<i>Ab</i>	 <i>AABb</i>	 <i>AAbb</i>	 <i>AaBb</i>	 <i>Aabb</i>
	<i>aB</i>	 <i>AaBB</i>	 <i>AaBb</i>	 <i>aaBB</i>	 <i>aabb</i>
	<i>ab</i>	 <i>AaBb</i>	 <i>Aabb</i>	 <i>aaBb</i>	 <i>aabb</i>

F₂ generation



2/ Liaison génétique : expérience de Bateson and Punnett (1900)

	NUMBER OF PROGENY	
Traits	Observed	Expected from 9:3:3:1 ratio
purple, long ($P/- \cdot L/-$)	4831	3911
purple, round ($l- P \cdot III$)	390	1303
red, long ($p/p \cdot L/-$)	393	1303
red, round ($p/p \cdot III$)	<u>1338</u>	<u>435</u>
	6952	6952



Introduction d'une probabilité de liaison : lp

$$(A+a)^2 \text{ lié à } (B+b)^2 =$$

$$\begin{aligned} & lp^2 (AABB + 2AaBb + aabb) + \\ & (1-lp)^2 (AAbb + 2AaBb + aaBB) + \\ & 2*lp(1-lp) (AABb + AaBB + Aabb + aaBb) \end{aligned}$$

Fréquences observées

$$AB : (lp^2 + 2)/4$$

$$Ab : (1-lp^2)/4$$

$$aB : (1-lp^2)/4$$

$$ab : lp^2/4$$

Introduction d'une probabilité de liaison : l_p

$l_p = \frac{1}{2}$ signifie l'indépendance, comme dans les lois de Mendel

Chez Bateson and Punnett, $l_p = 0.88$



3/ Sturtevant, 1913

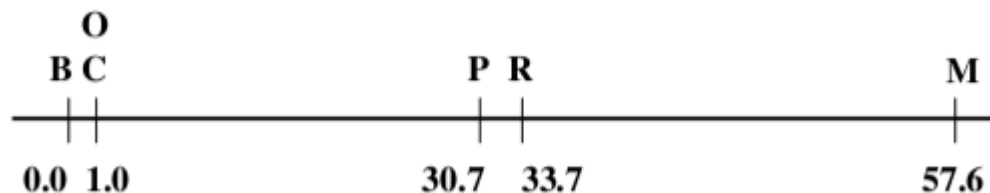
La probabilité de liaison l_p est proche d'une distance linéaire

$$d = 1 - l_p$$

Si a, b, c sont des facteurs et $d(a,c) > d(a,b)$, $d(a,c) > d(b,c)$

$$d(a,c) \approx d(a,b) + d(b,c)$$

Les facteurs sont organisés linéairement le long des chromosomes
(c'est une prédiction mathématique d'une structure non observée)



PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES

Volume 5

FEBRUARY 15, 1919

Number 2

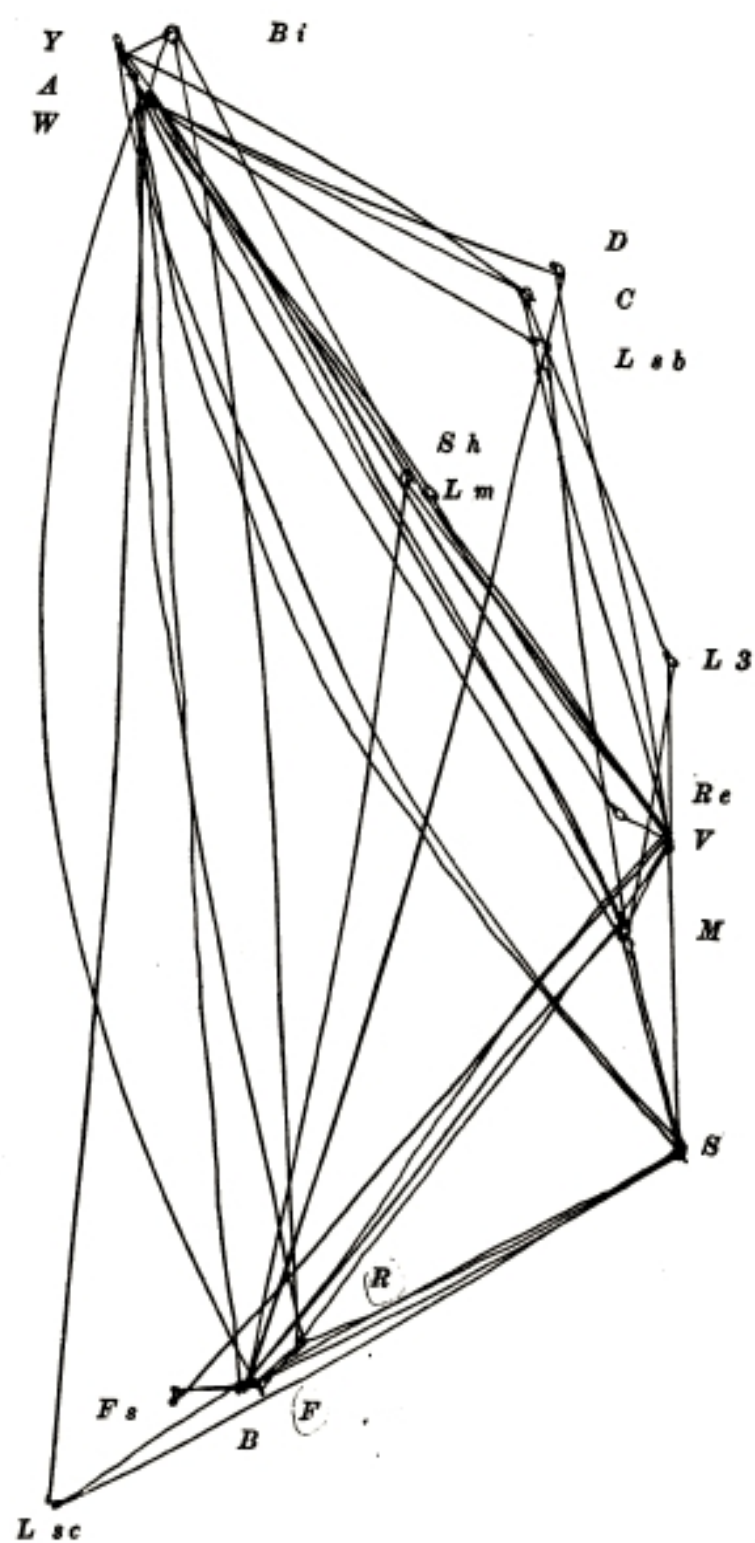
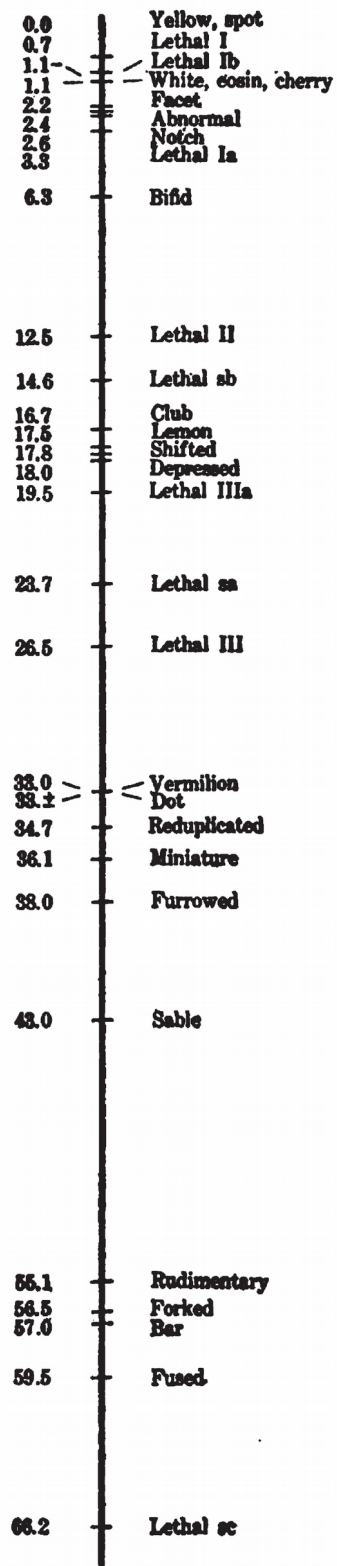
*IS THE ARRANGEMENT OF THE GENES IN THE CHROMOSOME
LINEAR?*

BY W. E. CASTLE

BUSSEY INSTITUTION, HARVARD UNIVERSITY

Read before the Academy, November 18, 1918

Every biologist is familiar with the remarkable discoveries of Morgan and his associates concerning the germ-cells of *Drosophila*. One of the most important of these discoveries is concerned with the phenomenon of linked inheritance. This kind of inheritance, while entirely conformable with Mendel's law, forms a very distinct and important class of cases whose existence has been brought to light since the rediscovery of the general law in 1900. Under the general law it is found that characters which behave as distinct units in heredity assort quite independently of each other. Thus if parents are crossed one of which possesses two characters, A and B, while the other lacks them, then the offspring of this cross will transmit A and B sometimes associated in the same gamete, sometimes in different gametes, the two events being under the laws of chance equally probable.



THE SPATIAL RELATIONS OF GENES

BY A. H. STURTEVANT, C. B. BRIDGES, AND T. H. MORGAN

COLUMBIA UNIVERSITY AND CARNEGIE INSTITUTION OF WASHINGTON

Communicated April 11, 1919

ARE GENES LINEAR OR NON-LINEAR IN ARRANGEMENT?

BY W. E. CASTLE ·

BUSSEY INSTITUTION, HARVARD UNIVERSITY

Communicated, August 13, 1919

THE EVIDENCE FOR THE LINEAR ORDER OF THE GENES

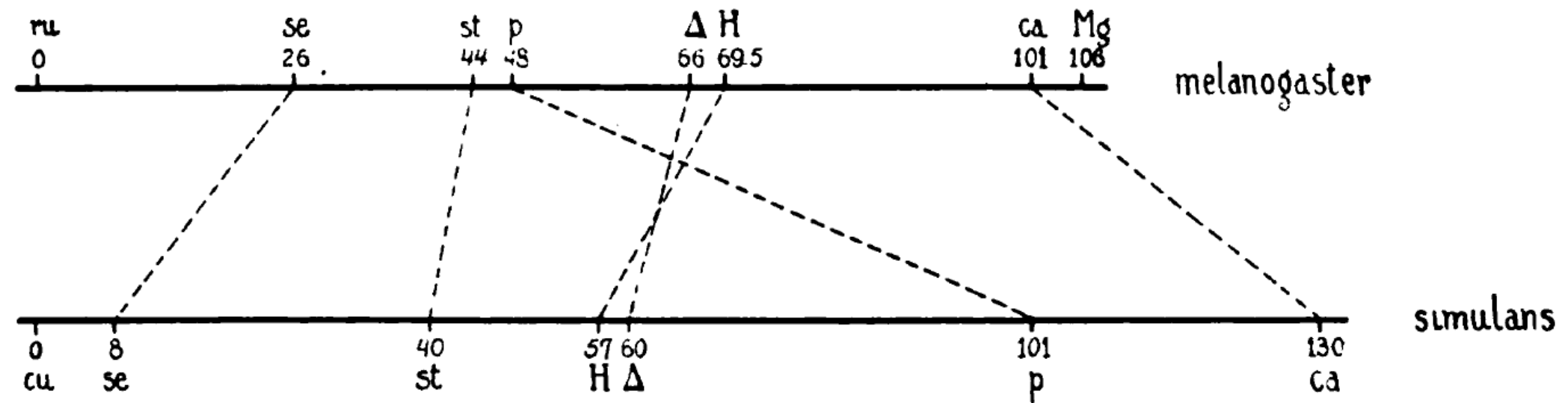
BY T. H. MORGAN, A. H. STURTEVANT AND C. B. BRIDGE

DEPARTMENT OF ZOÖLOGY, COLUMBIA UNIVERSITY

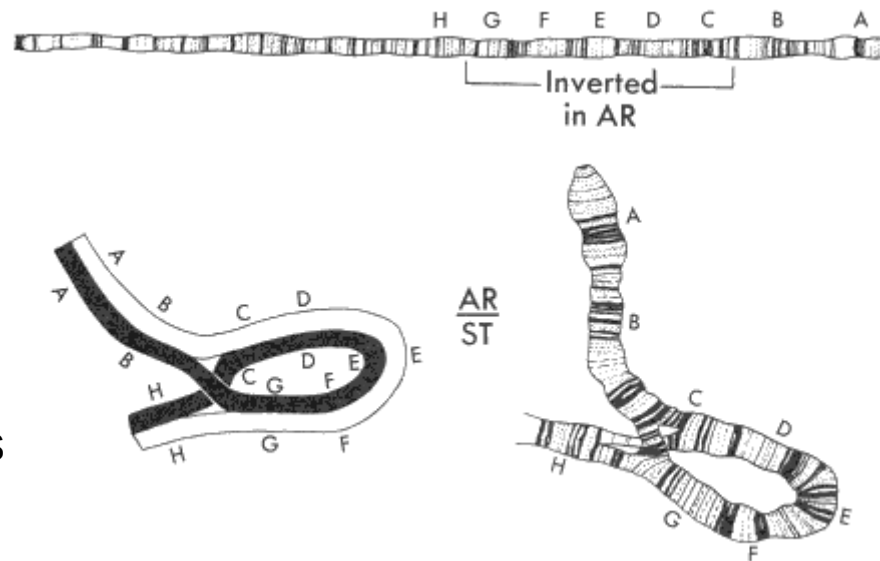
Communicated February 25, 1920

Despite Castle's dictum that we "have failed in two different attempts to establish the linear theory in the case of the three genes yellow, white and bifid," we are bold enough to maintain that the data furnished, and still furnish, the proof called for. We wish to call attention to the fact that in his last paper Castle ignores our proof of the linear order that is furnished by building up the whole chromosome (or even large sections of it) by "distances" so short that no double cross-over classes appear.

4/ Découverte de l'inversion, Sturtevant, 1921-1926



Confirmation de la
prédiction par des
observations sur des
chromosomes
polytènes



THE COMPARATIVE GENETICS OF *DROSOPHILA PSEUDOOBSCURA* AND *D. MELANOGASTER*

BY A. H. STURTEVANT AND C. C. TAN

*Wm. G. Kerckhoff Laboratories, California Institute of Technology,
Pasadena, California*

1937, Journal of Genetics

If the *pseudoobscura* sequence in each arm is arbitrarily taken as an alphabetical one (A B C . . .), then the *melanogaster* sequences become:

X	L	H	F	E	B	A	D	C	K	I	J	G	M	(7)
II L	D	E	F	A	C	B	(2)							
II R	A	C	E	B	F	D	(4)							
III L	C	F	E	B	A	D	(3)							
III R	A	E	B	C	F	D	G	(3)						

The numbers in parentheses represent the numbers of successive inversions necessary to turn these sequences into alphabetical ones (in the case of A we are not yet certain that six inversions may not be sufficient). The mathematical properties of series of letters subjected to the operation of successive inversions do not appear to have been worked out, so that we are so far unable to present a detailed analysis. It does appear, however, that the five arms (taken together) are definitely more alike in the two species than could result from chance alone.

THE HOMOLOGIES OF THE CHROMOSOME ELEMENTS IN THE GENUS DROSOPHILA

A. H. STURTEVANT AND E. NOVITSKI
California Institute of Technology, Pasadena, California

1941, Genetics

STURTEVANT and TAN (1937) state: "The mathematical properties of series of letters subjected to the operation of successive inversions do not appear to have been worked out, so that we are so far unable to present a detailed analysis. It does appear, however, that the five arms (taken together) are definitely more alike in the two species than could result from chance alone." These statements now require some modification.

With the help of PROF. MORGAN WARD, a beginning has been made in the study of the mathematical consequences of successive inversions. Com-

THE HOMOLOGIES OF THE CHROMOSOME ELEMENTS IN THE GENUS DROSOPHILA

A. H. STURTEVANT AND E. NOVITSKI
California Institute of Technology, Pasadena, California

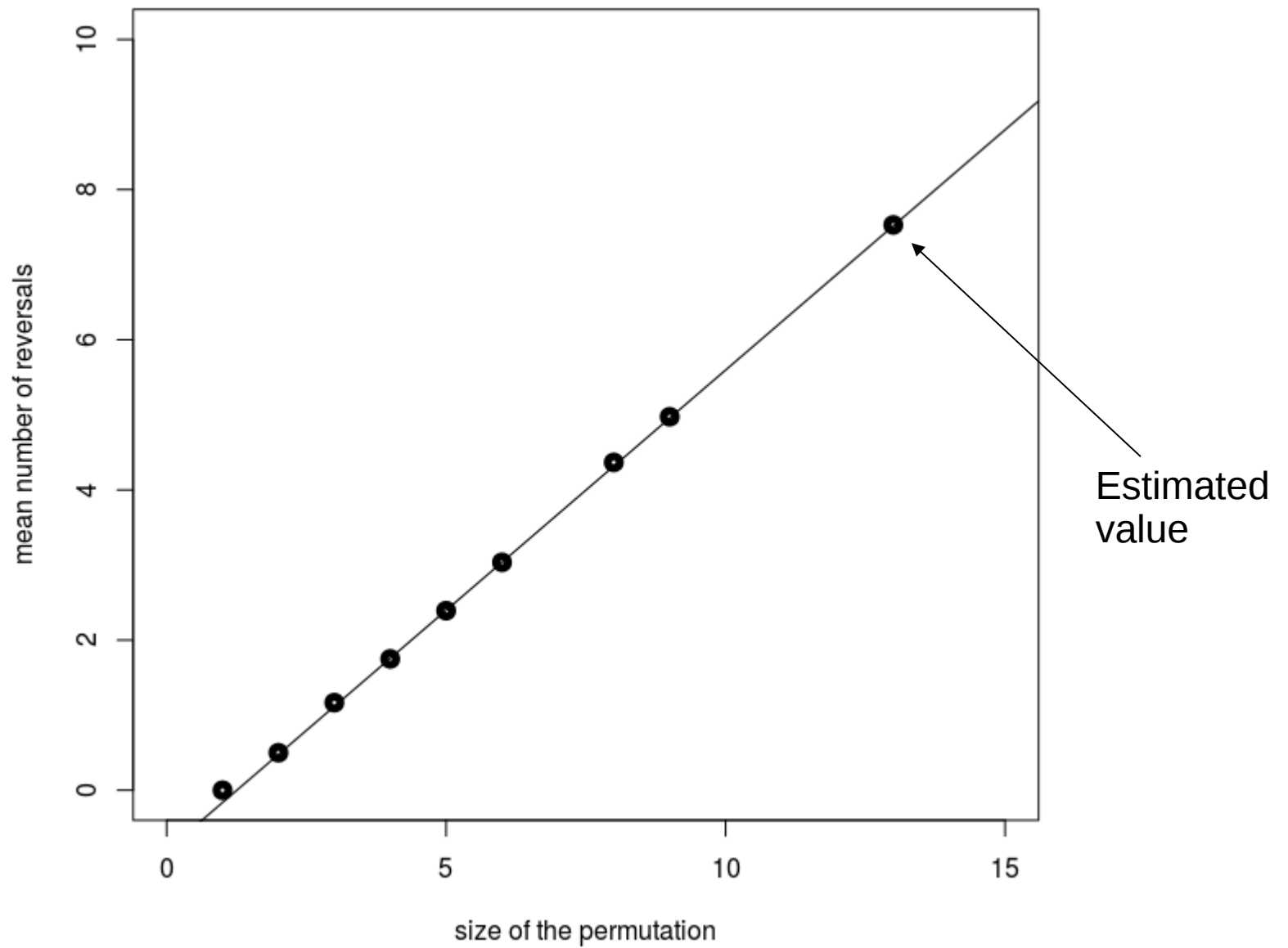
1941, Genetics

TABLE 4

Comparison of the required and calculated numbers of inversions to change the melanogaster into the pseudoobscura sequences.

ELEMENT	A	B	C	D	E	TOTAL
Loci	13	6	6	6	7	
Inversions required	7	2	4	3	3	19
Inversions calculated	7.6	3.0	3.0	3.0	3.7	20.3

Evidently the two species are not more alike than could easily result from chance alone.



Conclusion : pas de différence significative entre 7 et 7.6
Conséquence : pas de signal phylogénétique dans l'ordre de gènes

Sur « necessary »

Quel est le nombre **minimum** d'inversions ?

Mention en 1937, première difficulté sur un exemple

Premier article de maths en 1982

Preuve de NP-complétude en 1999

Sur « definitely », « evidently »

Peut-on dire que les gènes sont distribués au hasard ?

Deux questions :

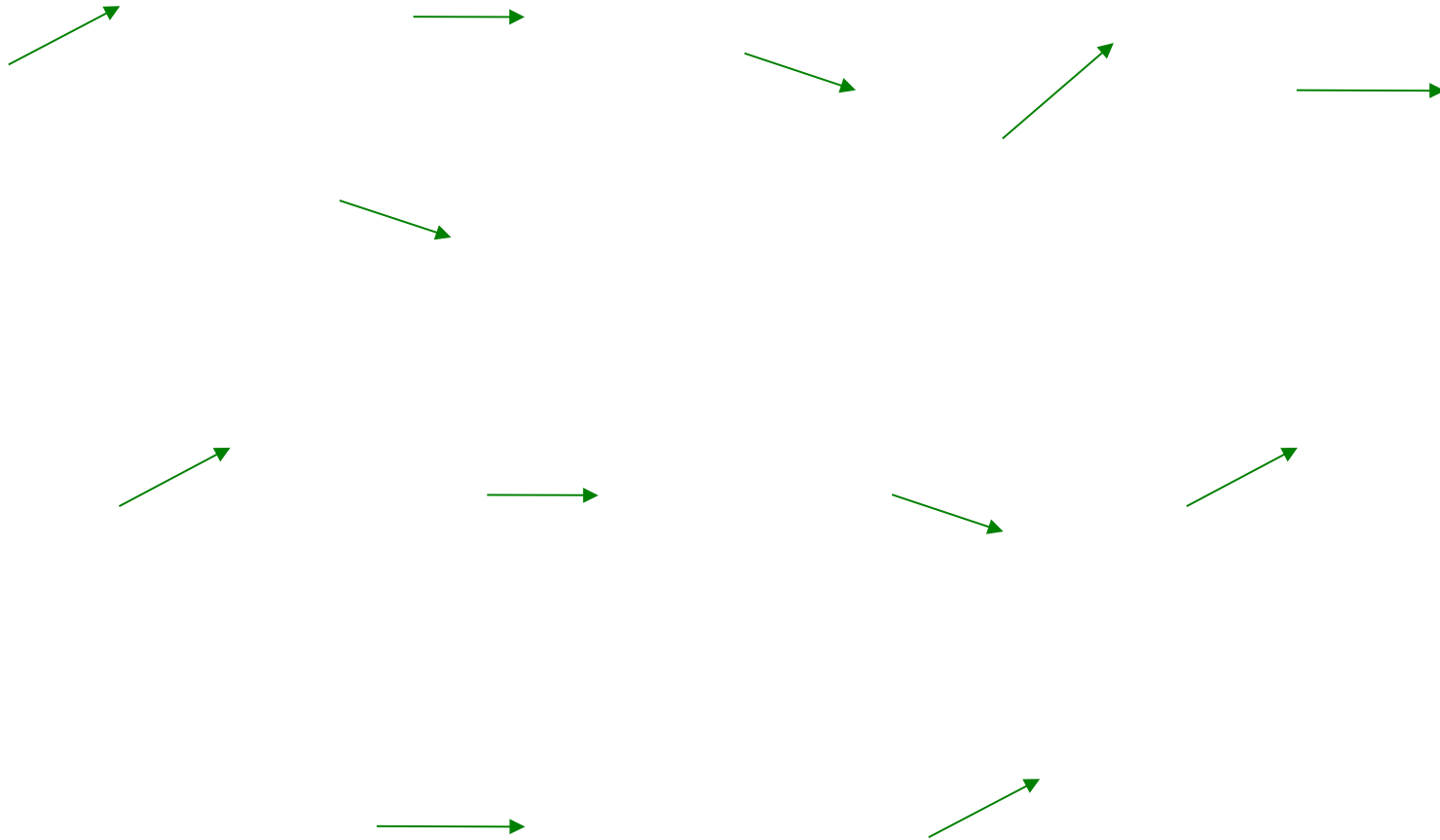
- quel est l'attendu du nombre d'inversions au hasard ?
- quel est l'intervalle de confiance ? (à ce jour pas de réponse connue)

Travaux pratiques « definitely », « evidently »

- 1/ Ecrivez et implémentez un algorithme qui donne le nombre minimum d'inversions pour transformer une permutation de lettres en l'ordre alphabétique. Ce sera nécessairement un algorithme de complexité exponentielle, applicable pour des petites tailles de données. Vous pouvez faire une recherche exhaustive sur toutes les suites d'inversions d'une certaine taille, et vérifier si l'une d'entre elles aboutit à l'ordre alphabétique.
- 2/ En générant toutes les permutations ou un échantillon de permutations aléatoires, corriger l'estimation de Sturtevant, Novitski, 1941 sur la distance moyenne d'inversion pour des chromosomes ayant 6, 7 ou 13 gènes. Donner une estimation de la variance.
- 3/ Que peut-on en conclure sur la conservation de l'ordre des gènes chez les drosophiles ?

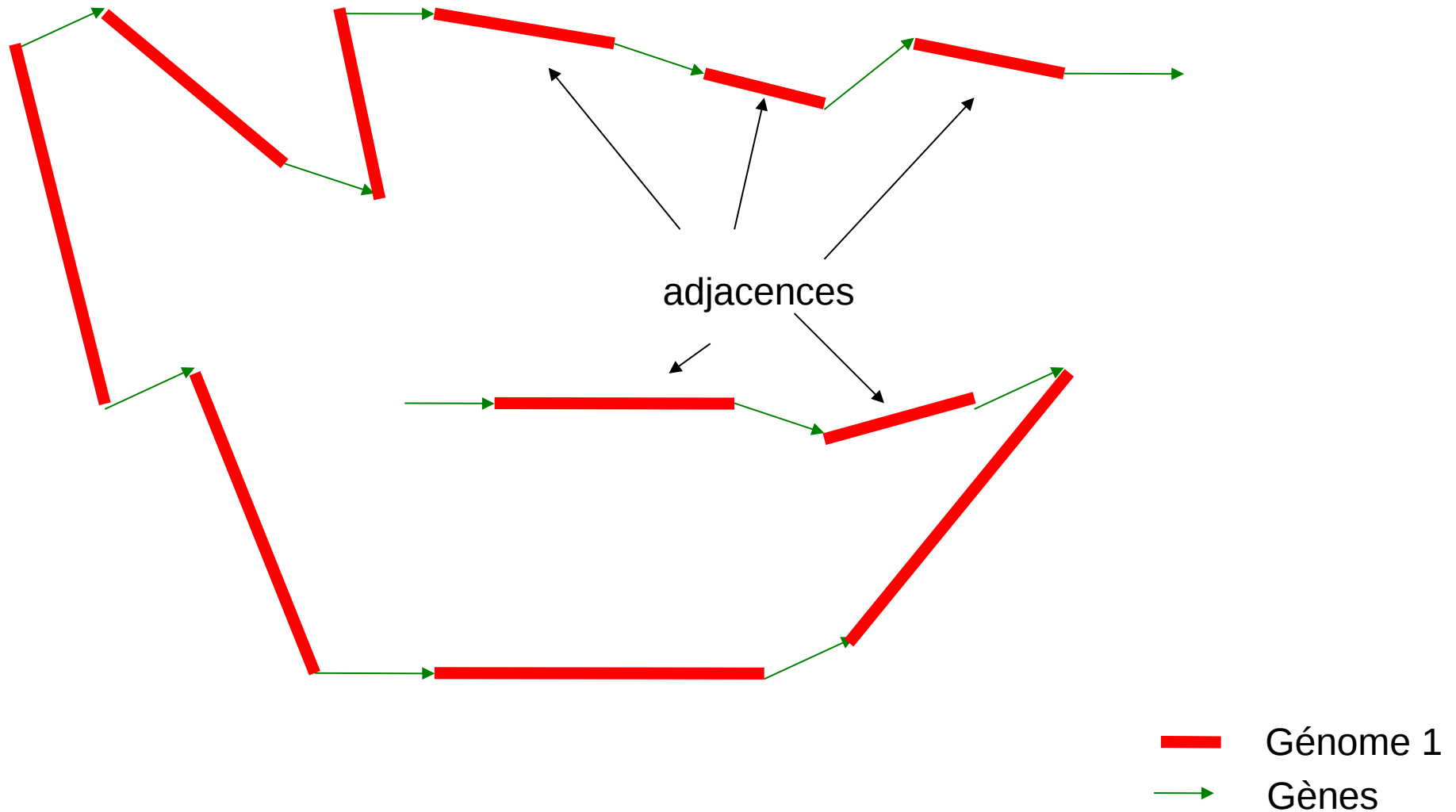
II/ Une modélisation alternative de l'ordre des gènes (DCJ, 2005)

Modélisation des gènes



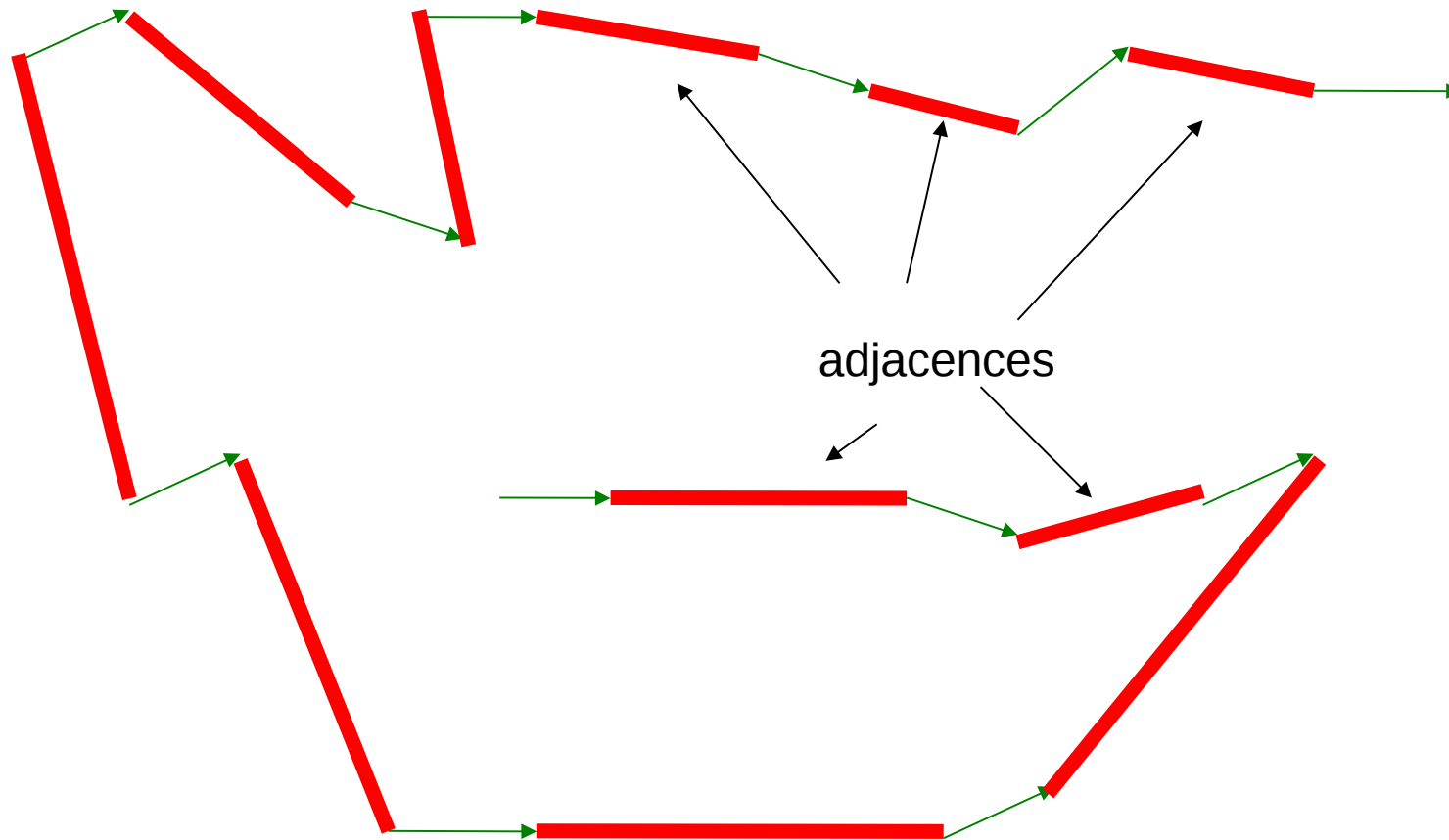
→ Gènes



Modélisation d'un génome



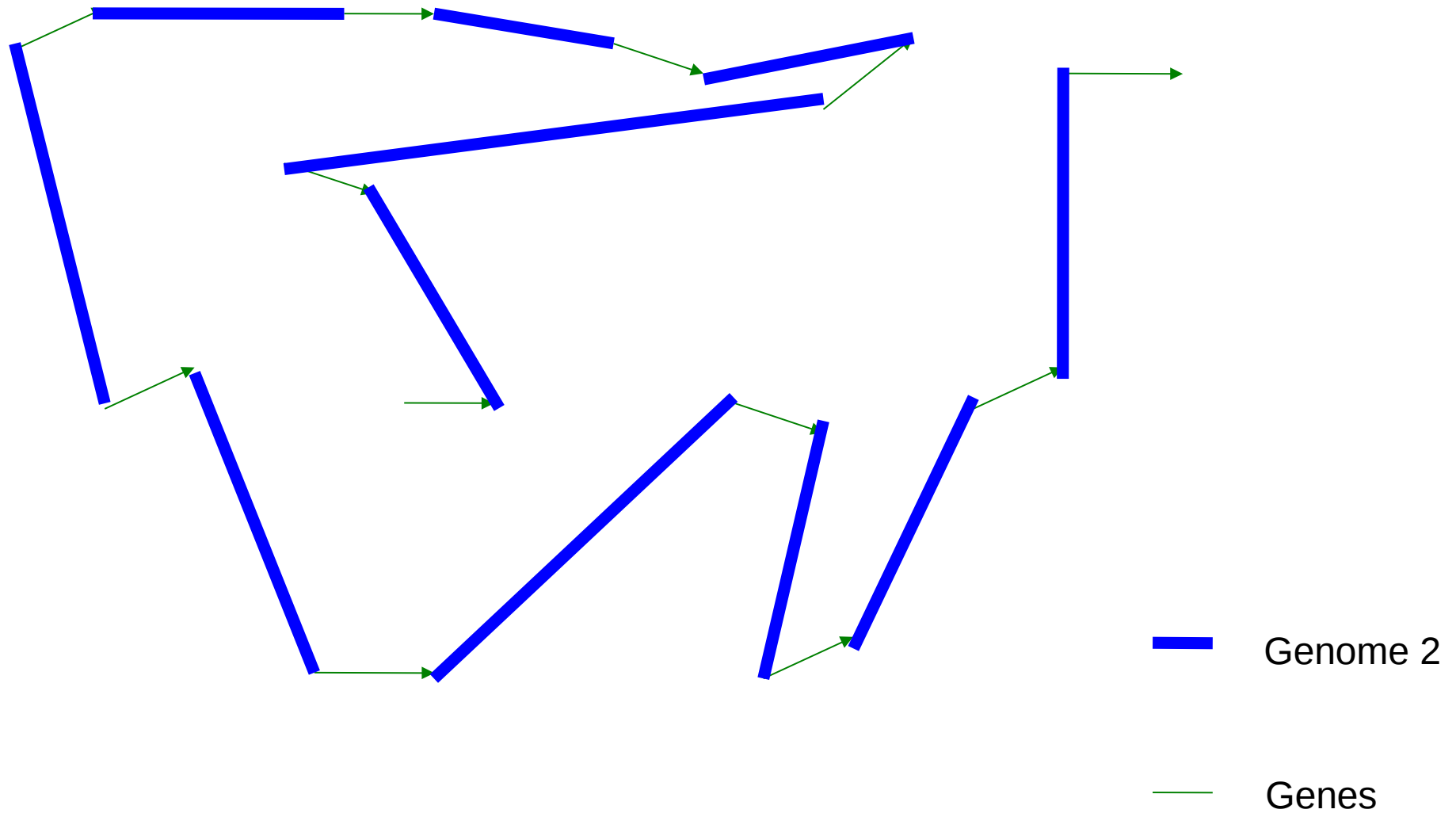
Modélisation d'un génome

C'est une modélisation de l'ordre de gènes
Ce qui change par rapport à la permutation est l'orientation
des gènes : chaque gène est un segment plutôt qu'un point.

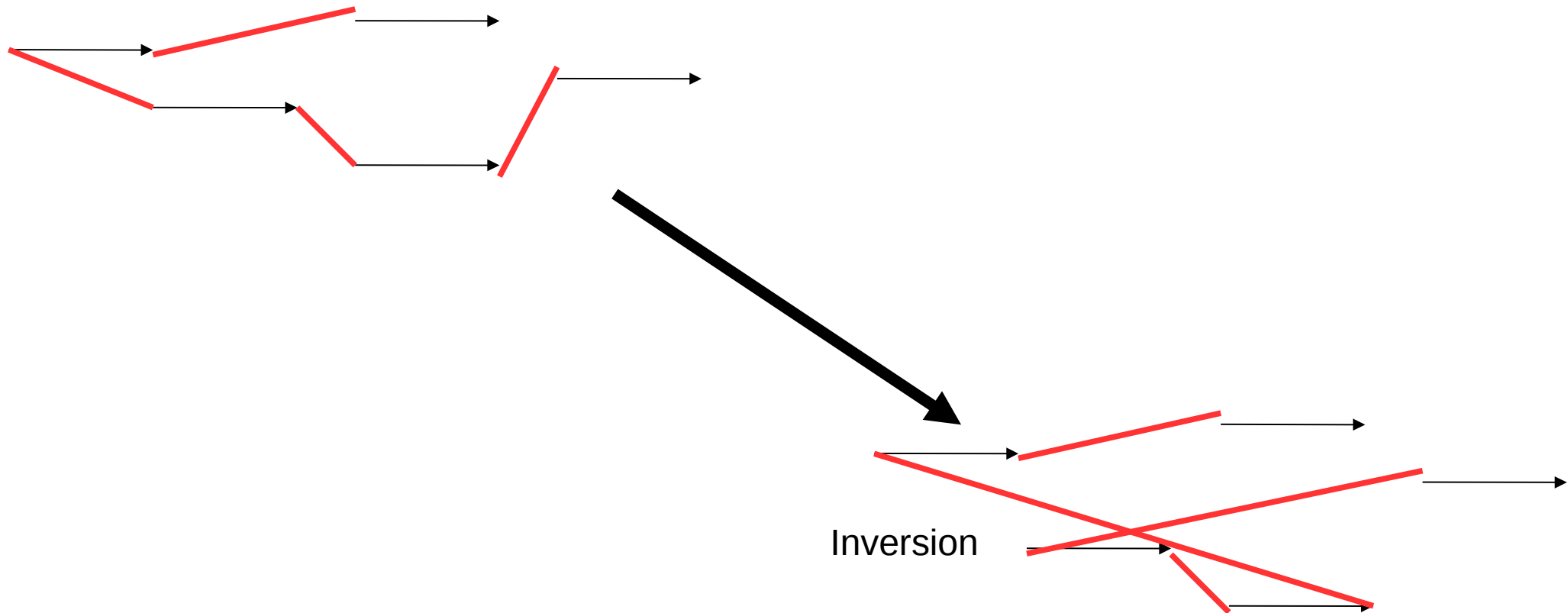


 Génome 1
 Gènes

Modélisation d'un génome

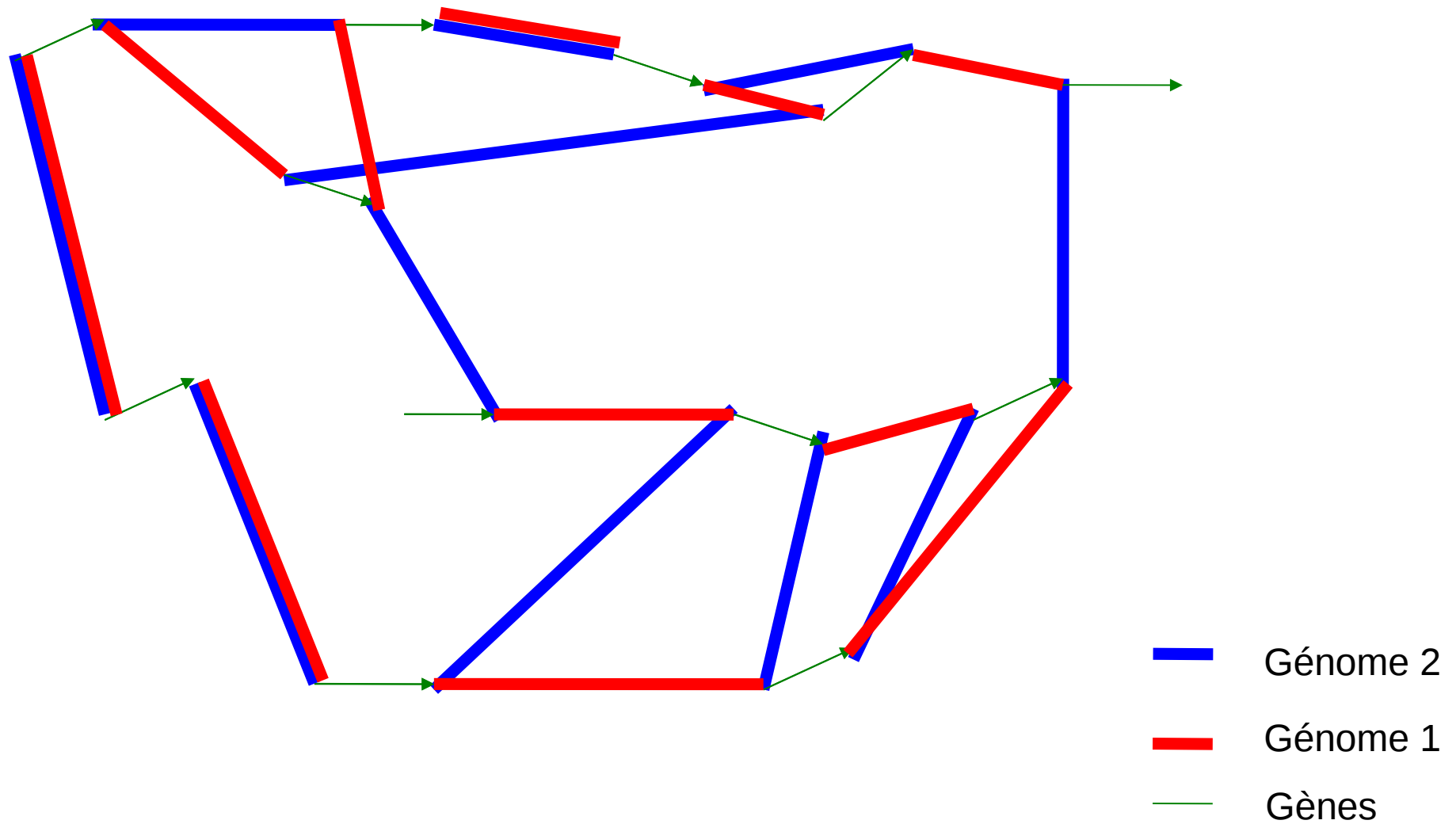


Modélisation du réarrangement « Double cassure/réparation »



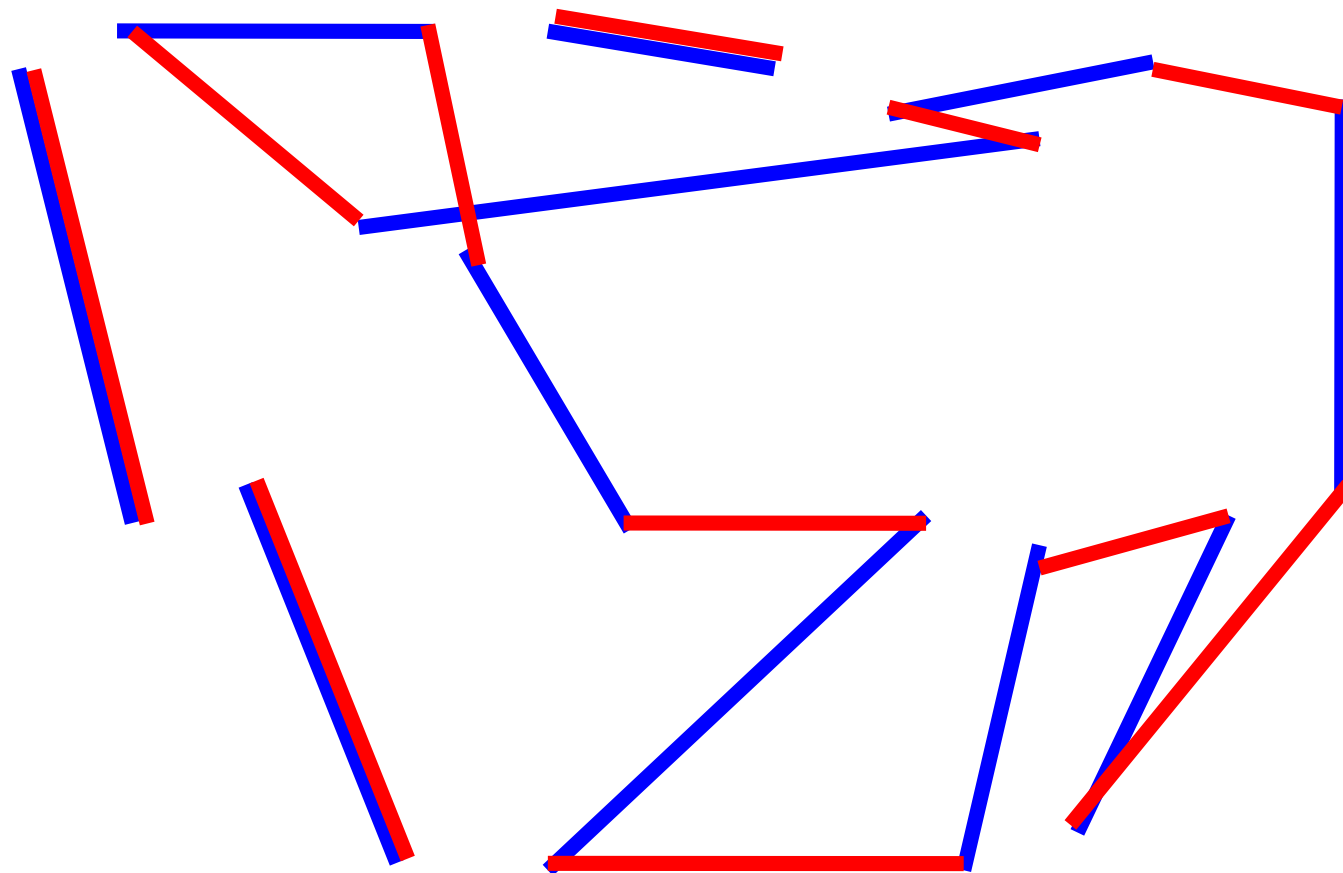
Choisir deux adjacences, les enlever, on obtient 4 sommets non couverts
Relier ces quatre sommets par deux adjacences

Le graphe de comparaison



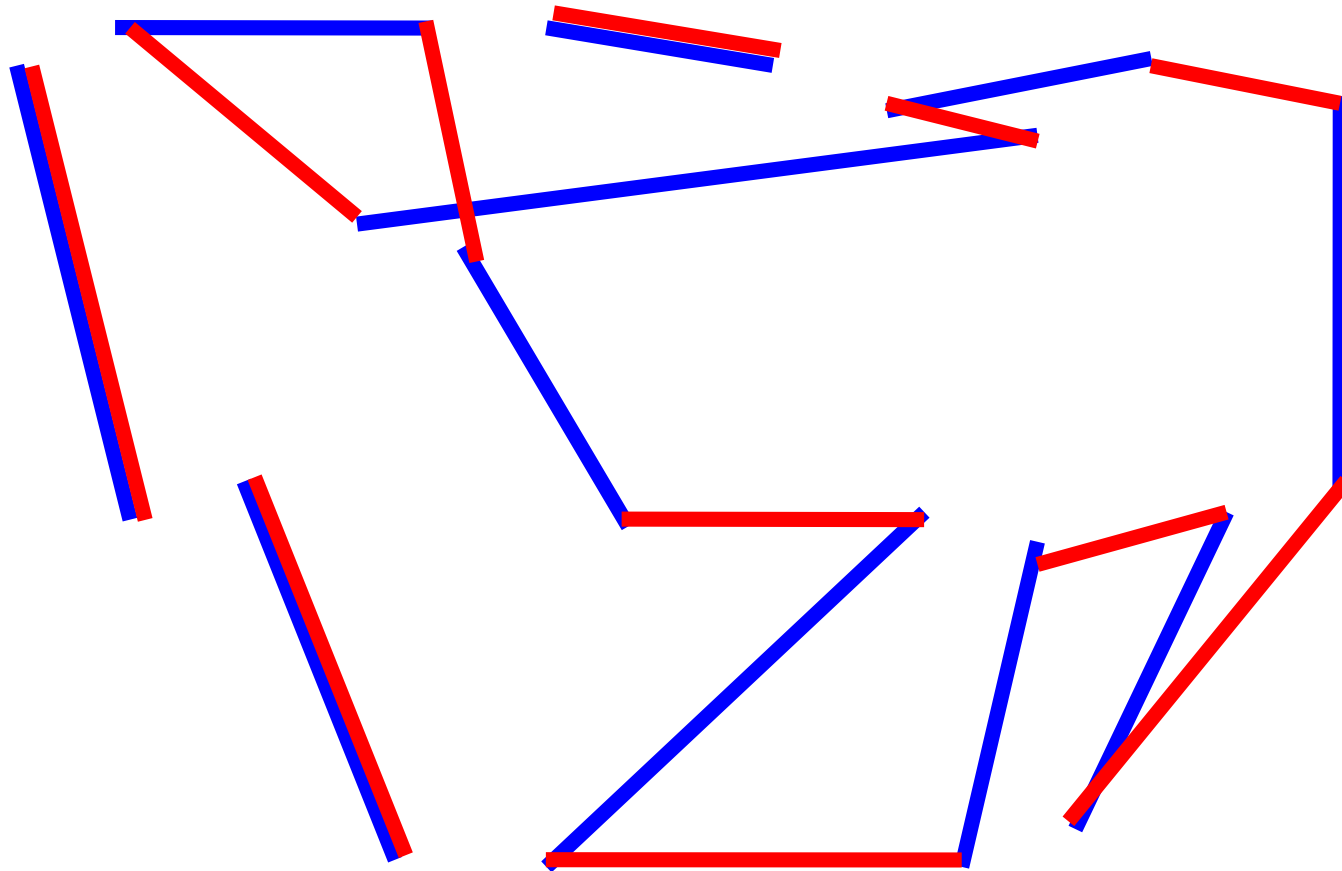
Le graphe de comparaison

Ensemble de cycles



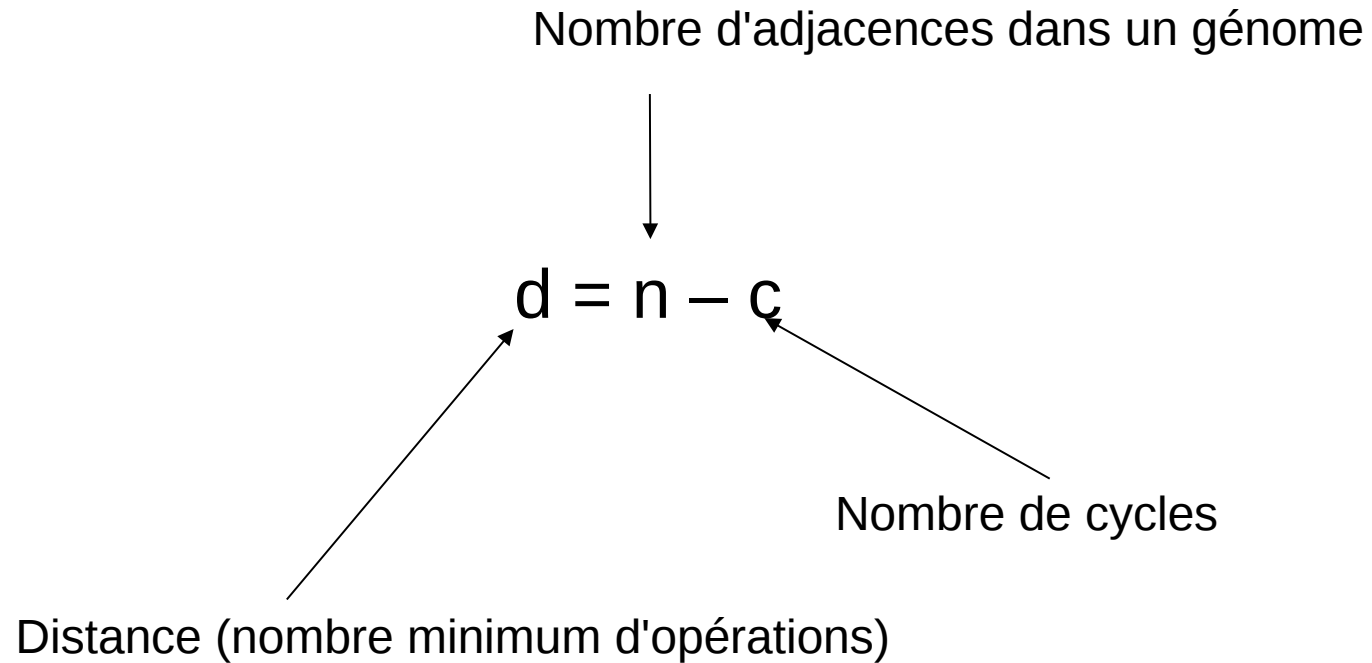
— Génome 2
— Génome 1

Transformer un génome en l'autre



En partant du rouge, on peut se rapprocher du bleu en augmentant le nombre de cycles

Nombre minimum d'operations



Vérifiez que si les génomes sont identiques, $d=0$

Vérifiez que pour chaque cycle de taille $2k$, il faut $k-1$ opérations

Travaux pratiques « necessary, evidently »

1/ Téléchargez une version moderne des gènes homologues entre *melanogaster* et *pseudoobscura* (chromosomes X, obtenu par séquençage) ici :

http://lbbe.univ-lyon1.fr/projets/tannier/INSA/MIG/droso_orthologies_filtered.txt

2/ Visualisez les données sous forme de dotplot (graphe à deux dimensions, avec les deux chromosomes en abscisse et ordonnée, et un point par gène)

3/ Utilisez l'orientation des gènes et la formule $d = n - c$ pour calculer un nombre minimum d'inversions entre les chromosomes de *melanogaster* et *pseudoobscura*.

4/ En générant 10000 permutations aléatoires avec orientations de la même taille que celle issue du nombre de gènes homologues entre *melanogaster* et *pseudoobscura*, estimez, en comparant à la permutation identité, une probabilité d'obtenir le nombre de cassures observé.

5/ Que peut-on en conclure sur la conservation de l'ordre des gènes chez les drosophiles?