Automate acyclique à mémoire variée pour analyser des données SNP

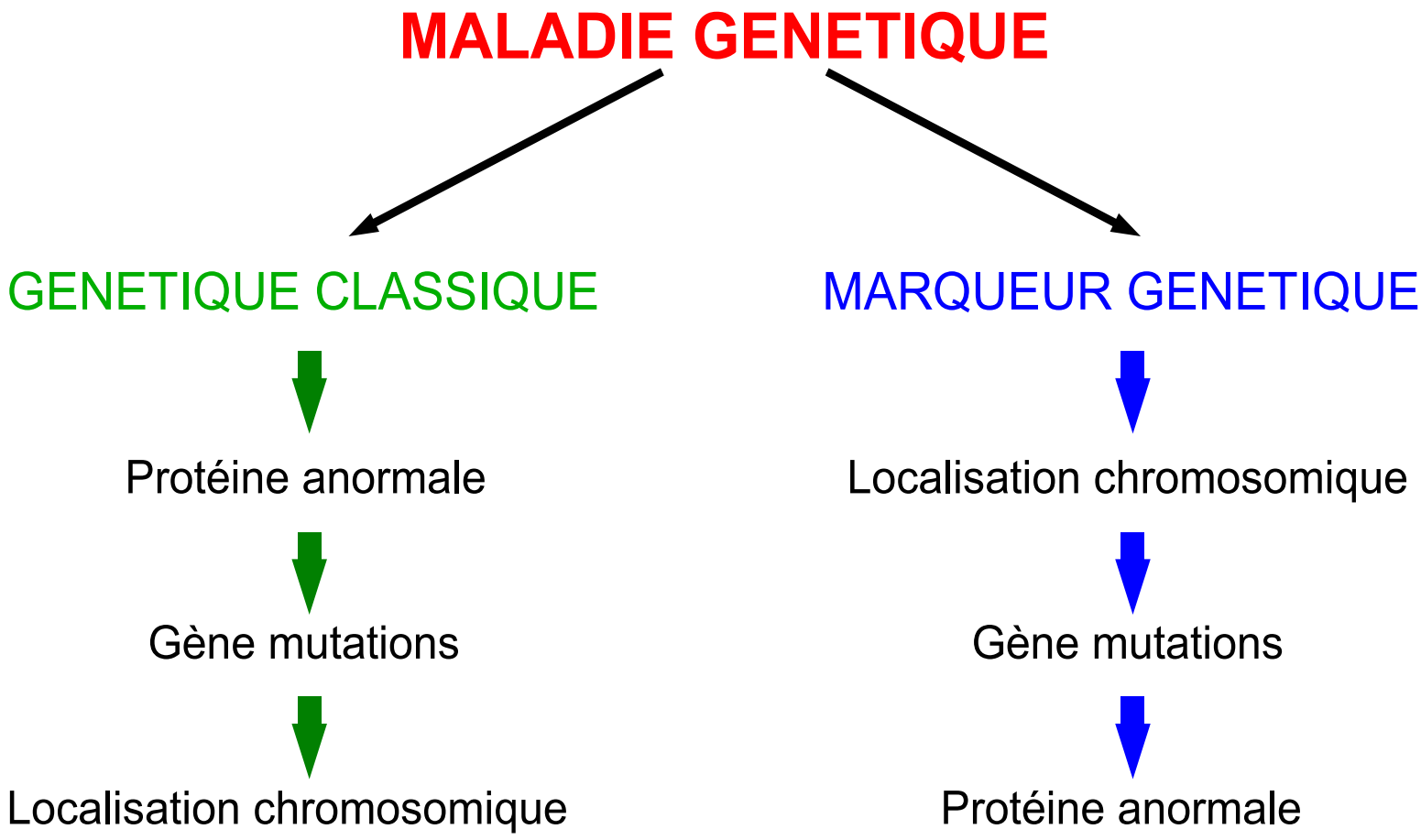
Tran Trang[1] et Hoang Ngoc Minh[2]

[1]Unité de Bioinformatique et Modélisation
Université de Liège

[2]Centre Intégré de Bioinformatique
Université de Lille 2

Plan

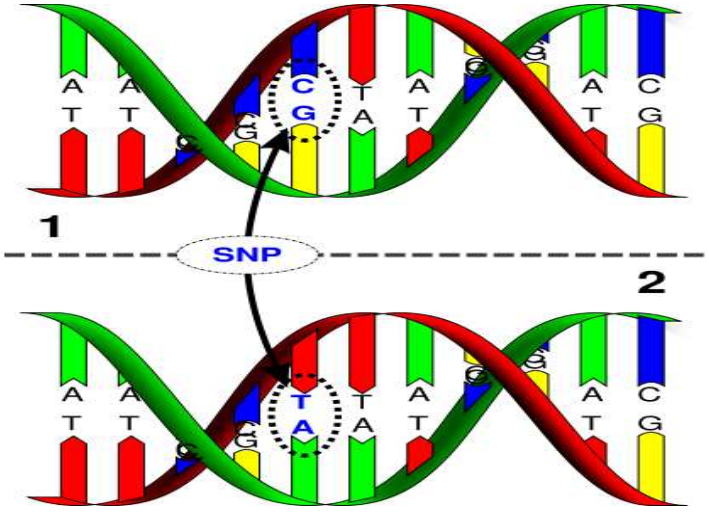
1. Concepts de marqueurs génétiques
2. Analyse d'association et optimisation combinatoire
3. Automate acyclique à mémoire variée
4. Algorithme d'apprentissage
5. Résultats expérimentaux

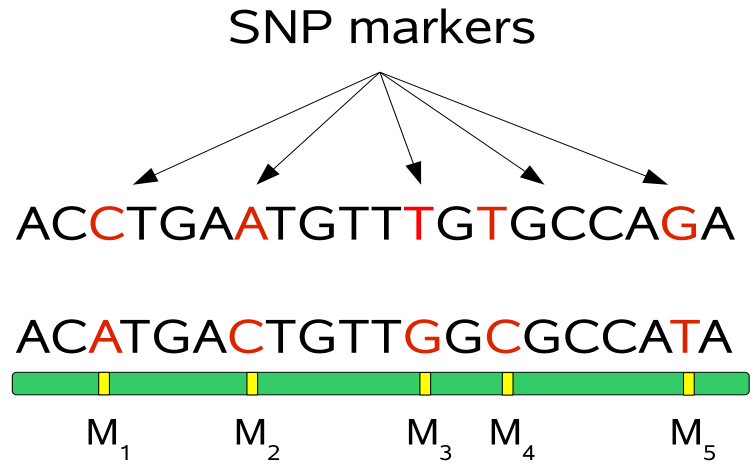


Méthodes pour identifier des gènes impliqués dans des maladies génétiques

1. Concepts de marqueurs génétiques

SNP marqueur, haplotype, génotype





- *Single Nucleotide Polymorphism* (SNP) est une mutation sur une position nuclétotidique de 2 chromosomes d'un individu
- *Locus* (marqueur) est une location de SNP dans un chromosome
- *Allèle* est un nucléotide du SNP
- *Haplotype*: séquence d'allèles sur un chromosome, $H_p = CATT, H_m = ACGC$
- *Génotype* d'un locus: un pair d'allèles (sans ordre) sur 2 chromosomes $G = \{C|A, A|C, T|G, T|C, G|T\}$

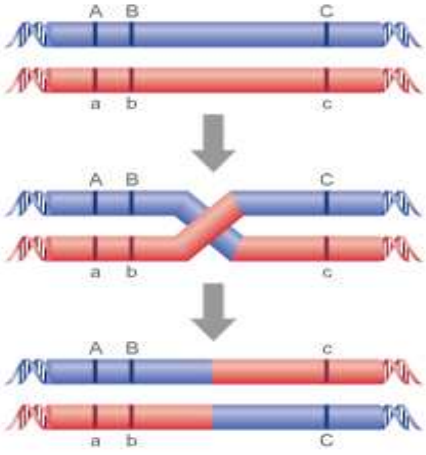
Déséquilibre de liaison

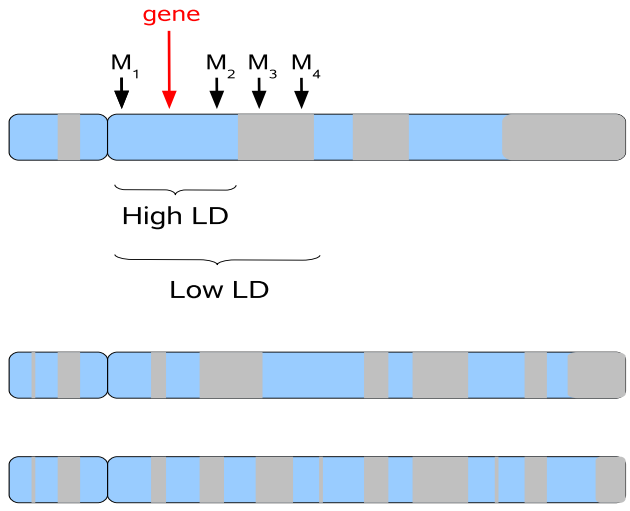
- Lorsque les allèles sur les différents loci sont dépendants (association non aléatoire), ils sont en *déséquilibre de liaison* (DL)-*linkage disequilibrium*, sinon ils sont en *équilibre de liaison* (EL)
- Soit $A|a$ et $B|b$ sont 2 allèles sur 2 loci marqueurs
  - P_A, P_B les fréquences d'allèles A et B
  - Si $P_{AB} \neq P_A P_B$ 2 loci sont en DL, sinon ils sont en LE
  - $D = P_{AB} - P_A P_B$ mesure DL

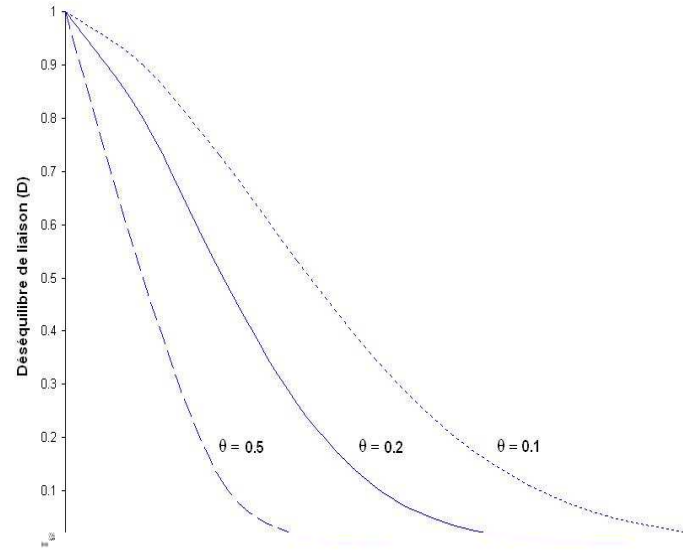
$$
\begin{aligned}
D' &= \begin{cases} \dfrac{D}{\max(P_A P_B, (1-P_A)(1-P_B))} & \text{si } D < 0, \\ \dfrac{D}{\min(P_A(1-P_B), (1-P_A)P_B)} & \text{si } D > 0. \end{cases}, \\
r^2 &= \frac{D^2}{P_A P_B (1-P_A)(1-P_B)}
\end{aligned}
\tag{1}
$$

- Si $|D'| = 1$ ($r^2 = 1$), les allèles sur les loci sont complètement en DL
- *Blog d'haplotype* est une région donc les loci sont en DL (c-à-d sans recombinaison)

Recombinaison







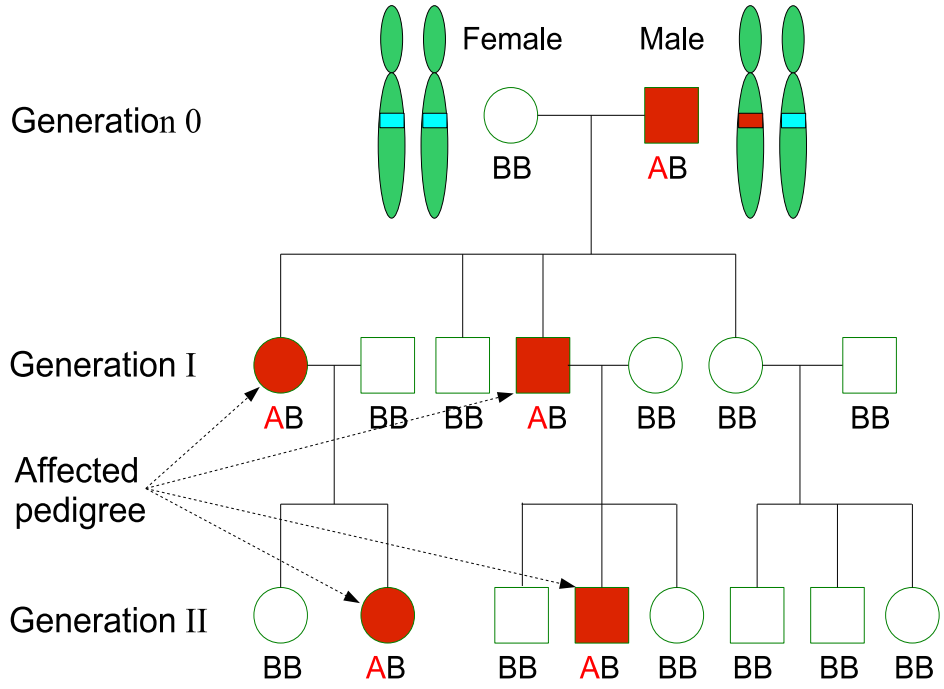
- Durant la méiose réductionnelle, si 2 chromosomes parentals s'échangent des matériaux génétiques (fragments d'ADN) $\rightarrow$ recombinaison
- Déséquilibre de liaison au temps de génération t est donnée par

$$D'_t = (1-\theta)^t D'_0, \tag{2}$$

où θ est le taux de recombinaison (la probabilité de recombinaison entre 2 loci)

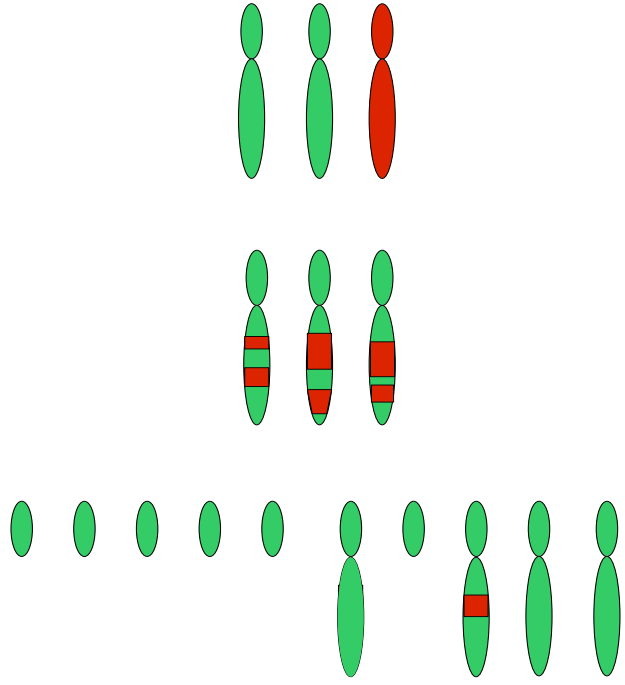
  - DL est décroissant dans le temps de génération
  - DL est fort si 2 loci sont proches

Analyse de liaison



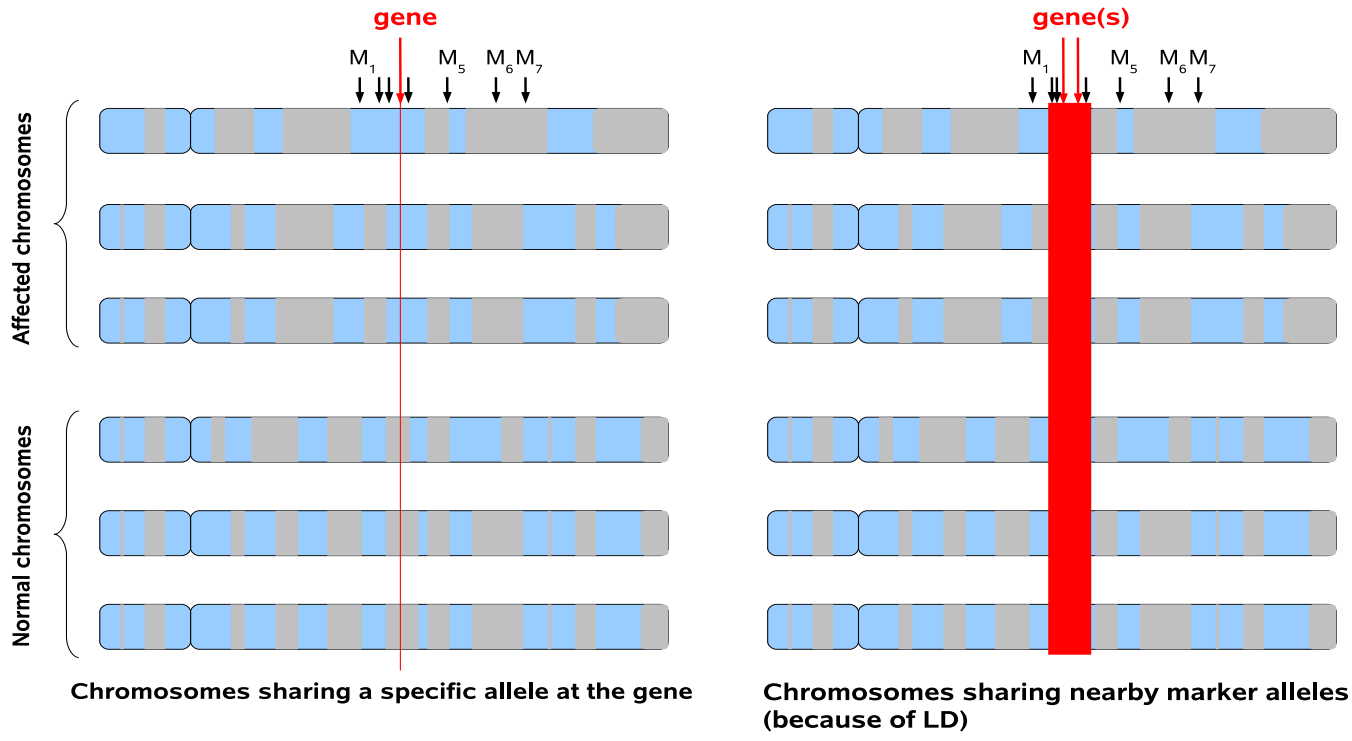
- Analyse de liaison recherche des gènes responsables qui transmete la maladie dans les familles
  - Tester la ségrégation d'une région chrosomique avec un phénotype
  - Identifier les marqueurs pour lesquels les allèles sont transmis
  - La recherche du gène pourra se concentrer autour de ce marqueur

Analyse d’association



2. Analyse d’association et problème d’optimisation combinatoire

Approches analytiques



- Association allélique: Tester l'association entre une maladie et des marqueurs en utilisant un seul marqueur à la fois $\Rightarrow$ moins informative.
- Association haplotypique: localiser des gènes à l'aide de patterns haplotypiques
  - Tester l'association entre une maladie et multilocus SNP,
  - regrouper les informations de plusieurs marqueurs proches,
  - tenir compte de DL entre marqueurs $\Rightarrow$ plus informative.

Problème d'optimisation combinatoire

- Considerons m haplotypes: $\begin{cases} m_A & \text{haplotypes anormaux (le cas)} \\ m_C & \text{haplotypes normaux (le contrôle, témoin),} \end{cases}$

- Ensemble d'apprentissage est une matrice haplotypique $M \in \mathcal{M}_{m,n}(\mathcal{A})$:

Haplotype	locus 1	locus 2	$\ldots$	locus n	phénotype
h_1	$h_{1,1}$	$h_{1,2}$	$\ldots$	$h_{1,n}$	1 (anormal)
h_2	$h_{2,1}$	$h_{2,2}$	$\ldots$	$h_{2,n}$	0 (normal)
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
h_m	$h_{m,1}$	$h_{m,2}$	$\ldots$	$h_{m,n}$	1 (anormal)

où $h_{ij} \in \mathcal{A} = \{1, 2\}$, l'ensemble de 2 allèles sur locus.

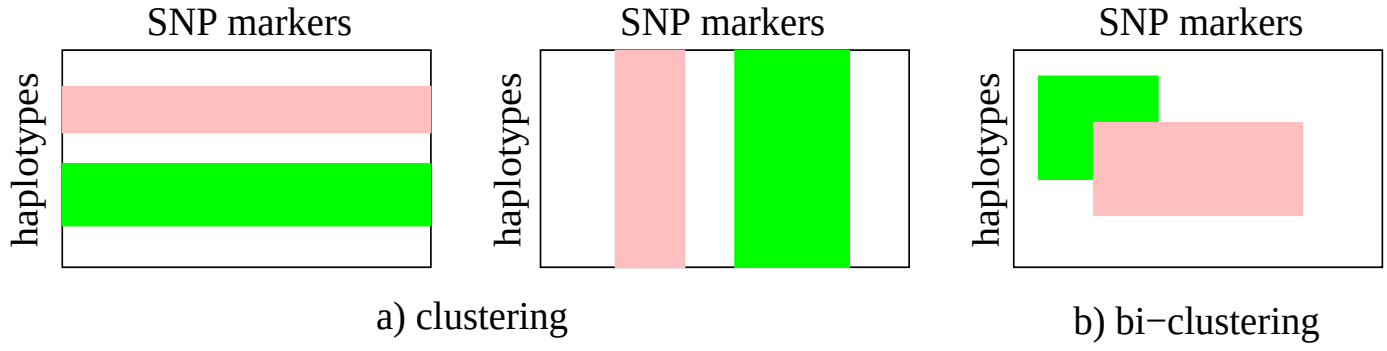
Problème d'optimisation combinatoire

- Soit $M \in \mathcal{M}_{m,n}(\mathcal{A})$ matrice haplotypique, $I \subseteq [1,..,m]$: indice d'haplotypes et $J \subseteq [1,..,n]$: marqueurs consécutifs.

- Matrice $M_{I,J}$ appelée pattern ou blog haplotypique si

$$M_{i_1,j} = \cdots = M_{i_l,j}, \quad j \in J \quad \text{et} \quad i_1, \ldots, i_l \in I. \tag{3}$$

- Si $J = [1,..,n]$, $M_{I,J}$ est un cluster de haplotypes similaires



- Identification des patterns haplotypiques associés à une maladie consiste:

$$\begin{cases} (A) & M_{i_1,j} = \cdots = M_{i_l,j},\ j \in J \ \text{ et } \ i_1, \ldots, i_l \in I, \\ (B) & D'(M_{I,J}) \geq \gamma,\ \gamma \approx 1, \\ (C) & P(M_{I,J} \in A) \geq \delta P(M_{I,J} \in C),\ \delta \geq 1. \end{cases} \tag{4}$$

Problème d'optimisation combinatoire

- Pour chaque pattern $M_{I,J}$, nous formons une table de contingency

	Pattern	Non Pattern	$\sum$
Affected	m_{AP}	m_{AN}	m_A
Control	m_{CP}	m_{CN}	m_C
$\sum$	m_P	m_N	m

- Dégrée d'association du pattern $M_{I,J}$ est mesurée par

$$\pm\chi^2(M_{I,J}) = \frac{\sqrt{m}(m_{AP}.m_{CN} - m_{AN}.m_{CP})}{\sqrt{m_A.m_C.m_P.m_N}} \quad (5)$$

- $\pm\chi^2(M_{I,J})$ permet de comparer $\frac{m_{AP}}{m_A}$ et $\frac{m_{CP}}{m_C}$
- $\pm\chi^2(M_{I,J}) > 0$ si $\frac{m_{AP}}{m_A} > \frac{m_{CP}}{m_C} \Rightarrow$ pattern est associé à la maladie
- $\pm\chi^2(M_{I,J})$ est maximal si $m_{CP} = 0 \Rightarrow$ pattern est parfait associé à la maladie

Problème d'optimisation combinatoire

- Soit $\beta > 0$ un seuil, un pattern est fort associé à la maladie si:

$$m_{AP} \geq \frac{m.m_A.\beta}{m.m_C + m_A.\beta}$$

et $\beta > 0$ est appelé une borne inférieure de $\pm\chi^2$.

- La recherche d'une solution de (4) conduit à résoudre le problème d'optimisation combinatoire

$$\begin{cases} (A) & M_{i_1,j} = \cdots = M_{i_l,j},\ j \in J \ \text{ et } \ i_1, \ldots, i_l \in I, \\ (B) & D'(M_{I,J}) \geq \gamma,\ \forall \gamma \approx 1, \\ (C) & \pm\chi^2(M_{I,J}) > \beta,\ \forall \beta > 0. \end{cases} \tag{6}$$

- Pattern identifé par (6) est appelé pattern haplotypique protectif

- La réponse de (6) permet de capturer des fragments haplotypiques dérivant du chromosome ancêt où la mutation s'est produite dans le passé

Exemple

Haplotype	locus 1	locus 2	locus 3	locus 4	locus 5	phénotype
1	1	1	1	1	2	1
2	1	2	1	1	2	0
3	1	2	2	2	1	0
4	1	1	1	1	2	1
5	1	2	1	1	2	1
6	2	2	2	2	1	0
7	2	1	1	1	2	1
8	1	2	2	2	1	0

- $M_{[2,3,5],[1,2]}$ représentant 12 a $m_{AP} = 1$ et $m_{CP} = 2 \Rightarrow \frac{m_{AP}}{m_A} = \frac{1}{4} < \frac{m_{CP}}{m_C} = \frac{1}{2}$,

  $\Rightarrow$ Pattern 12 aux marqueurs $[1, 2]$ n'est pas associé avec la maladie

- $M_{[1,2,4,5,7],[3,4,5]}$ représentant le pattern haplotypique 112 a $m_{AP} = 4$ et $m_{CP} = 1 \Rightarrow \frac{m_{AP}}{m_A} = 1 > \frac{m_{CP}}{m_C} = \frac{1}{4}$,

  $\Rightarrow$ Pattern 112 aux marqueurs $[3, 4, 5]$ est plus fréquent avec la maladie.

3. Automate Acyclique à Mémoire Variée

Structure d'Automate Acyclique

+

Dépendance conditionnelle

$\Downarrow$

Automate Acyclique à Mémoire Variée

Automate stochastique

Antomate Stochastique (AS) est 5-uplet $(\mathcal{A}, Q, P, I, F)$, où

- $\mathcal{A}$ est un alphabet fini,
- Q est un ensemble fini d'états,
- $I : Q \to [0, 1]$ est une application définissant la probabilité des états initiaux,
- $F : Q \to [0, 1]$ est une application qui définit la probabilité des états terminaux,
- $P : Q \times \mathcal{A} \times Q \to [0, 1]$ définit la probabilité de transition,

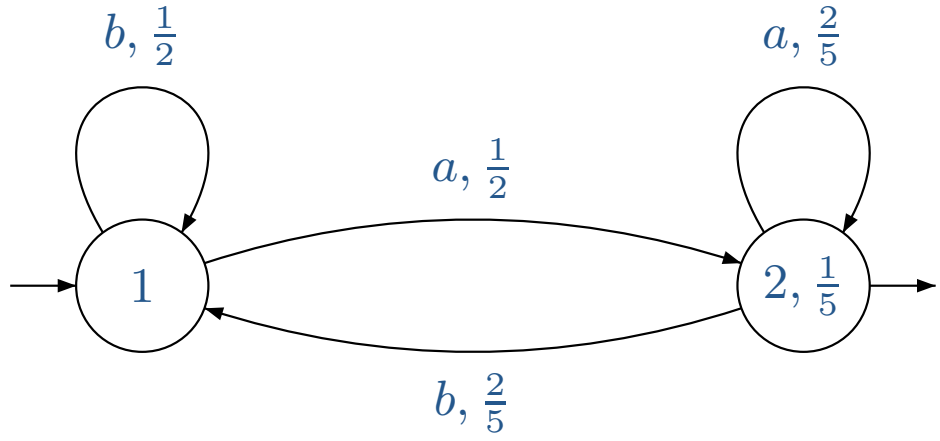
tels que les applications I, T, F vérifient les constraintes

$$\sum_{q \in Q} I(q) = 1, \quad \forall p \in Q, \quad F(p) + \sum_{x \in \mathcal{A}, q \in Q} P(p, x, q) = 1.$$

Pour tout mot $w = x_1 \dots x_L$ de $\mathcal{A}^*$, la probabilité générée du mot w est calculée par

$$\mathbb{P}(w) = \sum_{q_0, q_1, \dots, q_L \in Q} I(q_0) \prod_{i=1}^{L-1} P(q_i, x_{i+1}, q_{i+1}) F(q_L).$$

Exemple



$$\mathcal{A} = \{a, b\}, \quad Q = \{1, 2\}, \quad I = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad F = \begin{pmatrix} 0 \\ \frac{1}{5} \end{pmatrix}, \quad P = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{2}{5} & \frac{2}{5} \end{pmatrix}$$

$$P(1, b, 1) = \frac{1}{2}; \quad P(1, a, 2) = \frac{1}{2}$$

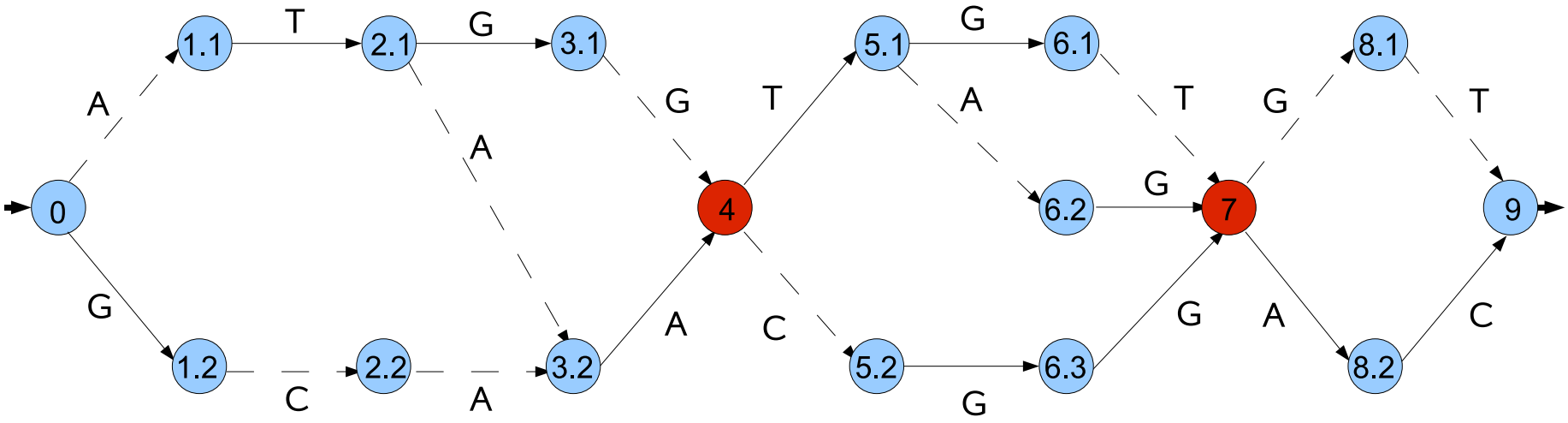
$$P(2, b, 1) = \frac{2}{5}; \quad P(2, a, 2) = \frac{2}{5}$$

$$\begin{aligned} \mathbb{P}(w = abba) &= I(1) \times P(1, a, 2) \times P(2, b, 1) \times P(1, b, 1) \times P(1, a, 2) \times F(2) \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{2}{5} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{5} \end{aligned}$$

Automate acyclique à mémoire variée

- Automate Acyclique est un AS sans boucle
- Automate Acyclique à Mémoire Variée (AAMV) est un AA tel que sa longueur de mémoire est variée dans le temps:
- $\{X_t\}_{t>0}$ un processus stochastique, les transitions $P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots, X_{t-l} = x_{t-l})$ sont les fonctions qui dépendent à un seul nombre varié l, où
  - $l = l(x_{t-1}, x_{t-2}, \ldots)$ est une fonction du passé
  - si $l(x_{t-1}, x_{t-2}, \ldots) = k$ est fixée $\Rightarrow$ automate d'ordre k habituel
    - $P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots) = P(X_t = x_t | X_{t-1} = x_{t-1})$
    - $P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots) = P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2})$
  - si $l(x_{t-1}, x_{t-2}, \ldots)$ n'est pas fixée, nous avons un AAMV

Exemple



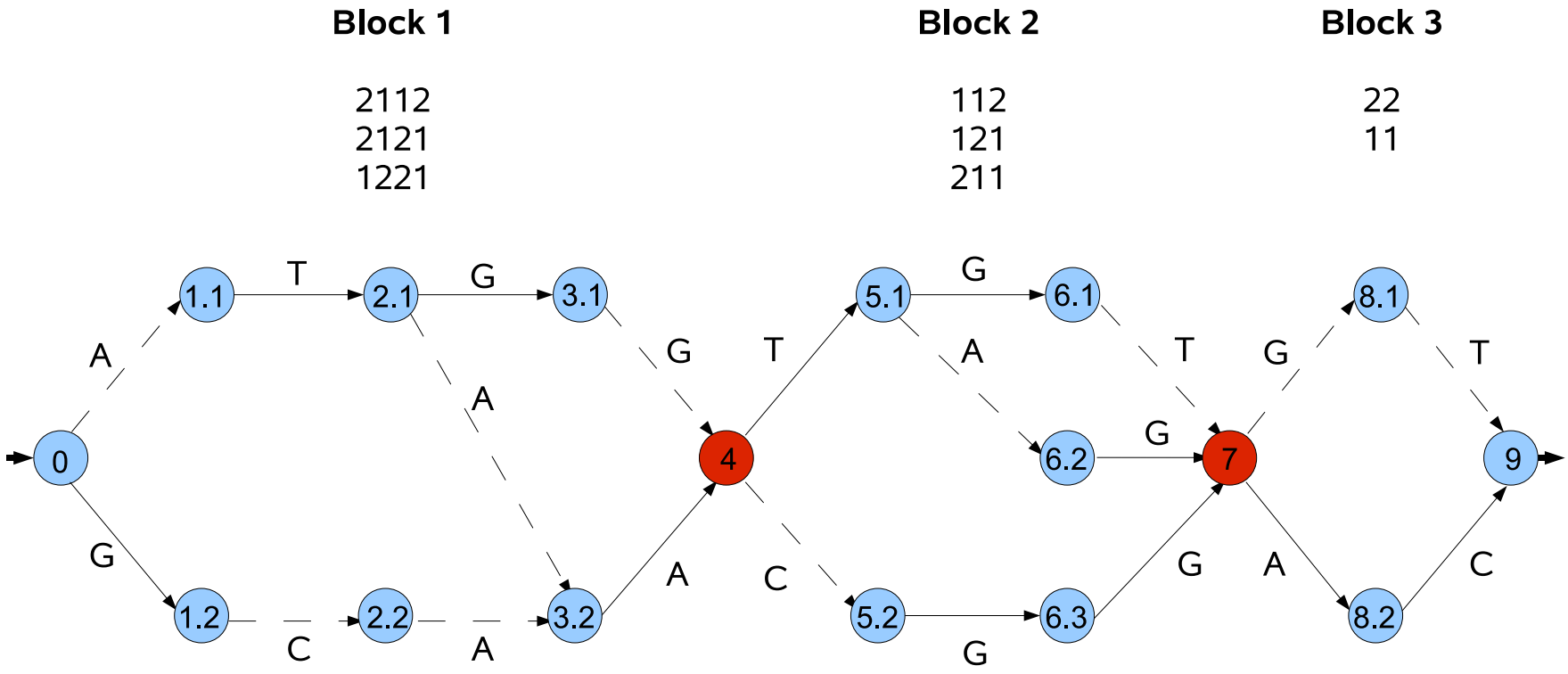
Dans ce travail, X_t représente un allèle au marqueur $t, 1 \leq t \leq n$:

- $P(X_2 = T|X_1 = A)$
- $P(X_3 = A|X_2 = C, X_1 = G)$
- AAMV perd complètement la mémoire aux nœuds 4 et 7,

Contributions d'un AAMV

- Structure de AAMV s'adapte à la recombinaison et au DL entre les marqueurs:
  - longueur de mémoire du AAMV dépend au DL entre les marqueurs
  - chaîne a un long mémoire $\Rightarrow$ marqueurs sont forts DL
  - chaîne a une courte mémoire $\Rightarrow$ marqueurs sont faibles DL
- AAMV est un model efficace pour:
  - déterminer des sites de recombinaison
  - localiser des mutations ou polymorphismes
  - capturer la structure de fort DL (blogs haplotypiques à fort DL)
  - identifier patterns haplotypiques associés à une maladie (teste d'association)

Exemple



- Modèle perd complètement la mémoire aux nœuds 4 et 7,
- Nœuds 4 et 7 représentent l'histoire de recombinaison,
- Allèles (haplotype) dans chaque blog sont fort DL,
- Utilisant $\pm\chi^2$ pour identifier les patterns les plus fréquents avec la maladie

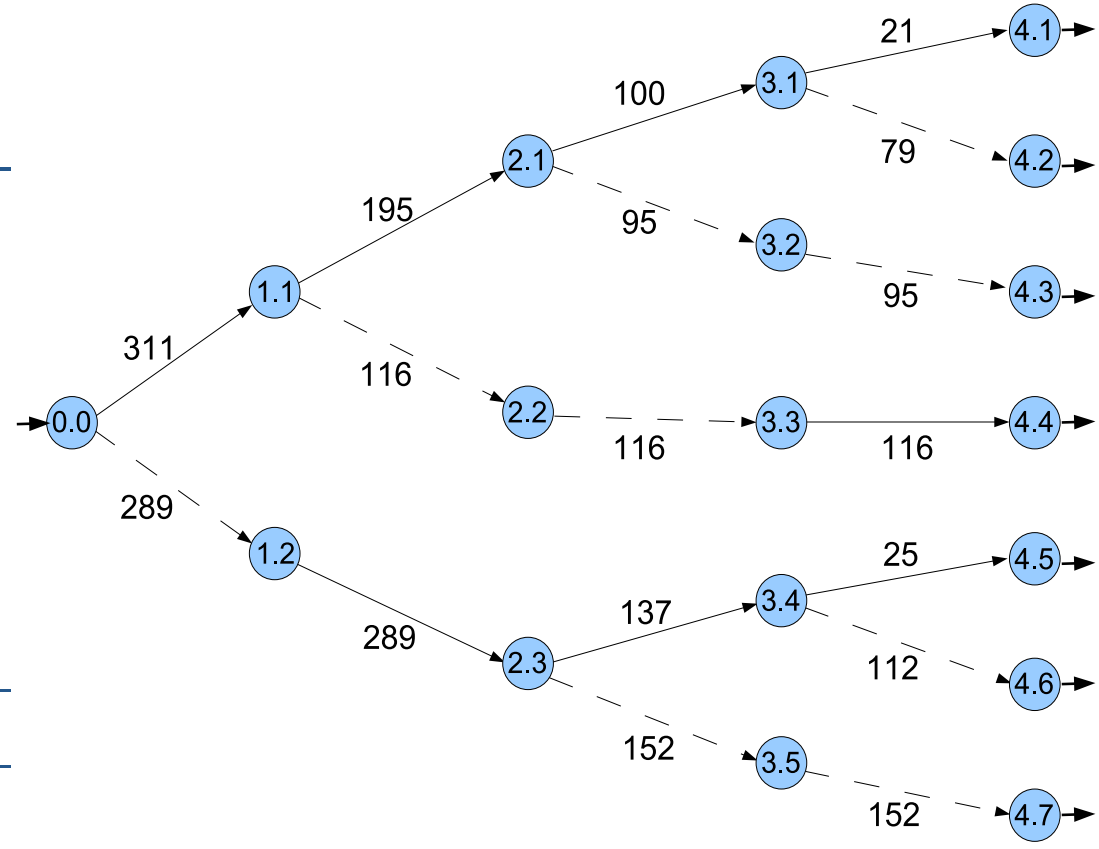
4. Algorithme d'apprentissage d'un AAMV

Données haplotypiques $\longrightarrow$ Arbre préfixe pondéré $\xrightarrow{\text{fusion des états}}$ AAMV

Données haplotypiques $\longrightarrow$ Arbre préfixe pondéré

- Arbre préfixe pondéré est un outil efficace pour représenter de données
- Il fournit une méthode de clustering hiérarchique:
  - Chaque sous-arbre définit un groupe d'haplotypes
  - Chaque feuille représente un haplotype

Haplotype	Cas	Témoins	$\sum$
1111	12	9	21
1112	43	36	79
1122	43	52	95
1221	59	57	116
2111	14	11	25
2112	60	52	112
2122	69	83	152
$\sum$	300	300	600



Critère de similarité

- Soit $u = x_1 \dots x_h \in \mathcal{A}^*, 0 \leq \mu \leq 1$, p et q sont compatibles si

$$|P_p(u) - P_q(u)| \leq \mu, \tag{7}$$

où, $P_p(u) = \prod_{l=1}^{h-1} P(p_i, x_l, p_{i+1})$ est la probabilité de u

- Nous utilisons la borne de Hoeffding pour mesurer la compatibilité

$$P\Big(|p - \frac{f}{n}| < \sqrt{\frac{1}{2n}\log(\frac{2}{\alpha})}\Big) > 1 - \alpha,\ 0 < \alpha < 1$$

- Deux nœuds p et q sont dits α-compatibles (similaires) si :

$$\begin{cases} \Delta_{p,q}(x) = |\frac{N_p(x)}{N_p} - \frac{N_q(x)}{N_q}| < \sqrt{\frac{1}{2}\log(\frac{2}{\alpha})}\Big(\frac{1}{\sqrt{N_p}} + \frac{1}{\sqrt{N_q}}\Big),\ \forall x \in \mathfrak{X},\ 0 < \alpha < 1, \\ \forall u \in \mathcal{A}^*, \Delta_{p,q}(xu) < \sqrt{\frac{1}{2}\log(\frac{2}{\alpha})}\Big(\frac{1}{\sqrt{N_p}} + \frac{1}{\sqrt{N_q}}\Big), \end{cases}$$

où, N_p et $N_p(x)$ sont respectivement les comptages associés au nœud p.

Algorithme d'apprentissage

Donnée: ensemble $\mathcal{L} = \{h_1, \ldots, h_m\}$ de m haplotypes sur n SNPs
m_C normals, m_A anormals, paramèttre α

Initialiser $\mathcal{G} \leftarrow \mathcal{T}(\mathcal{L})$

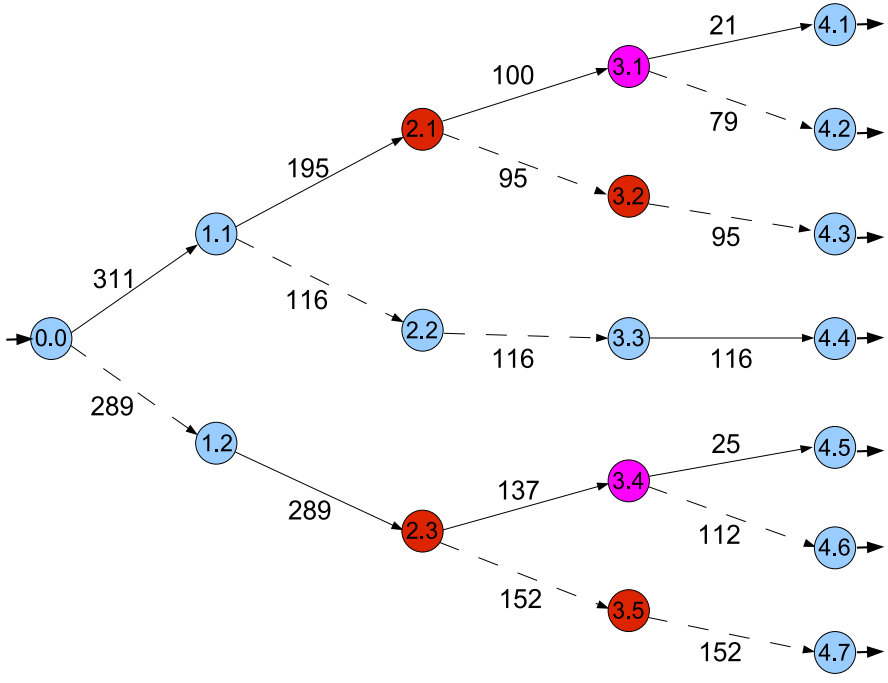
- pour $h = 1$ à n faire
  - pour chaque couple p et q faire
    - si p et q est α-compatibles
    - $\mathcal{G} \leftarrow FUSION(\mathcal{G}, p, q)$
  - fin pour
- fin pour

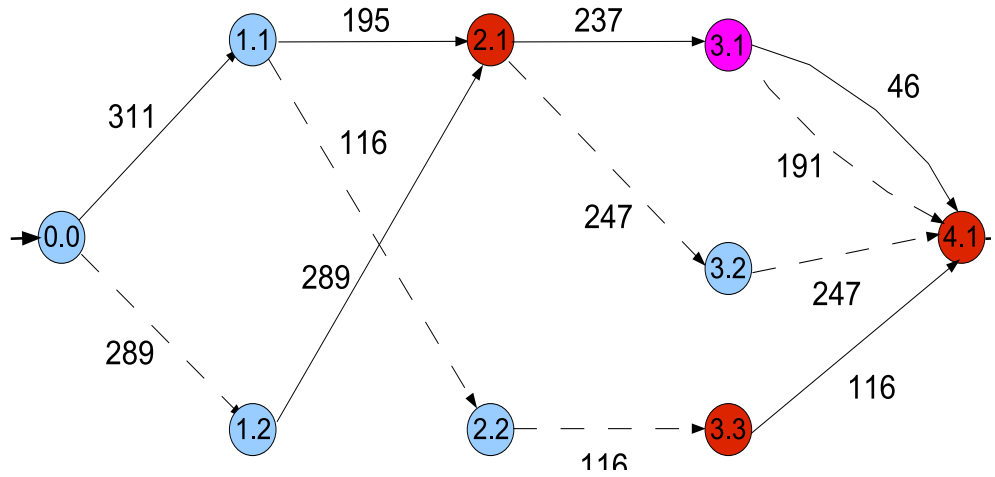
Retourner $\mathcal{G}$

Sortie: AAMV représenté par $\mathcal{G}$

La complexité de l'algorithme est $\mathcal{O}(m^2|\mathcal{A}|^2 n)$

Exemple





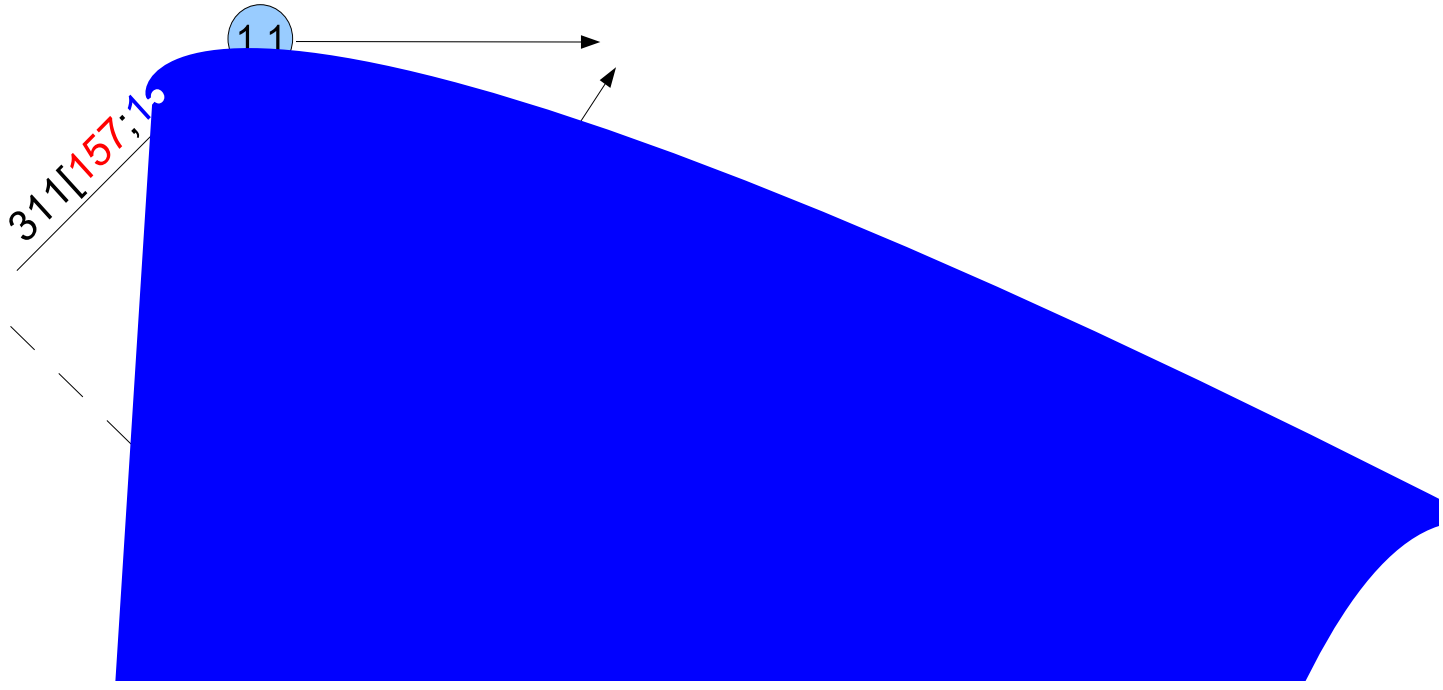
Pour $\alpha = 0.05$, nous avons $\mu = \sqrt{\frac{1}{2} * \log(\frac{2}{\alpha})} * (\frac{1}{\sqrt{195}} + \frac{1}{\sqrt{289}}) = 0.18$

$\Delta_{2.1,2.3}(1) = |\frac{100}{195} - \frac{137}{289}| = 0.039 < \mu$; $\Delta_{2.1,2.3}(11) = |\frac{21}{195} - \frac{25}{289}| = 0.021 < \mu$

$\Delta_{2.1,2.3}(12) = |\frac{79}{195} - \frac{112}{289}| = 0.018 < \mu$; $\Delta_{2.1,2.3}(22) = |\frac{95}{195} - \frac{152}{289}| = 0.039 < \mu$

$\Longrightarrow$ Deux nœuds 2.1 et 2.3 sont compatibles $\Longrightarrow$ On fusionne 2.1 et 2.3.

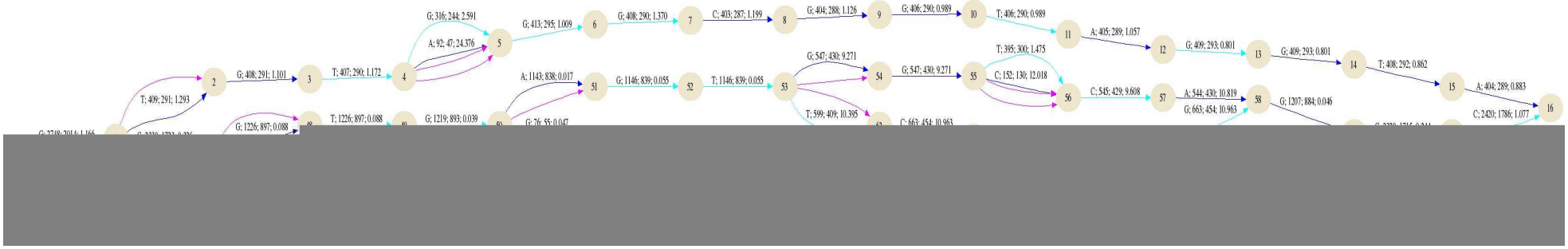
Exemple



5. Résultats expérimentaux

- Nous avons développé un logiciel prototype en PERL
- Pour tester le modèle:
  - 2832 haplotypes sur 16 SNP marqueurs dont 2080 anormals et 752 normals
  - Chaque haplotype est un mot de longueur 16 sur $\mathcal{A} = \{1, 2\}$

Clustering

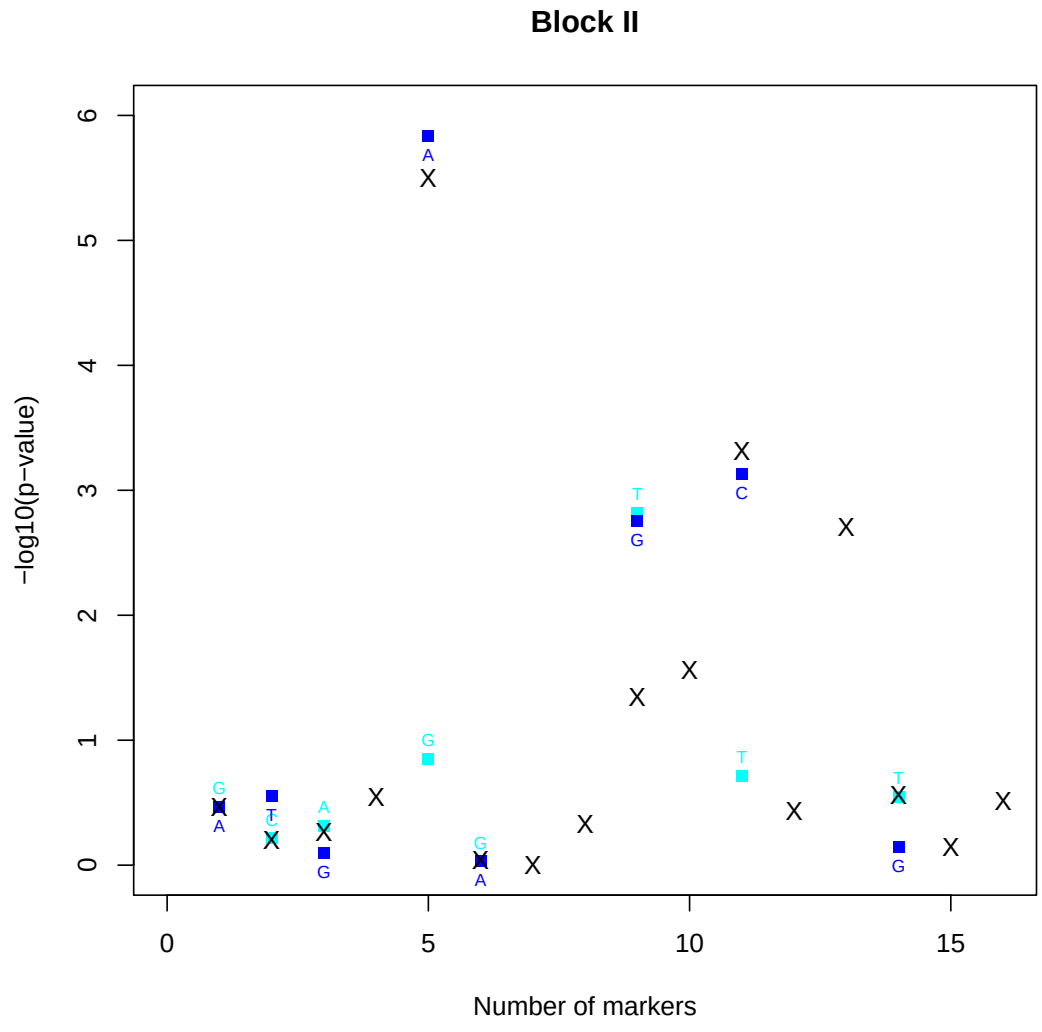


- Les "splitting nodes" sont associés aux clusters haplotypiques
- 2 clusters à la profondeur 9 sont significatifs ($\pm\chi^2 > 3.84$)

	CAS	CTR
G C A T G G G T T C T C G T C C	0.232	0.227
G C A T G G G T T C T C G G C C	0.170	0.176
	0.430	0.404
G T G T G G G C G G T A G G T A	0.117	0.094
G T G T A G G C G G T A G G T A	0.018	0.049
A T G A G G G T G G T C G G T C	0.029	0.018
	0.164	0.161
G C G T G A G T G G T C A G C C	0.151	0.132
G C G T G A G T G G C C A G C C	0.065	0.028
	0.216	0.160
G C G T G A G T T C T C G G C C	0.197	0.249
G C G T G A G T G C T C G G C C	0.020	0.026

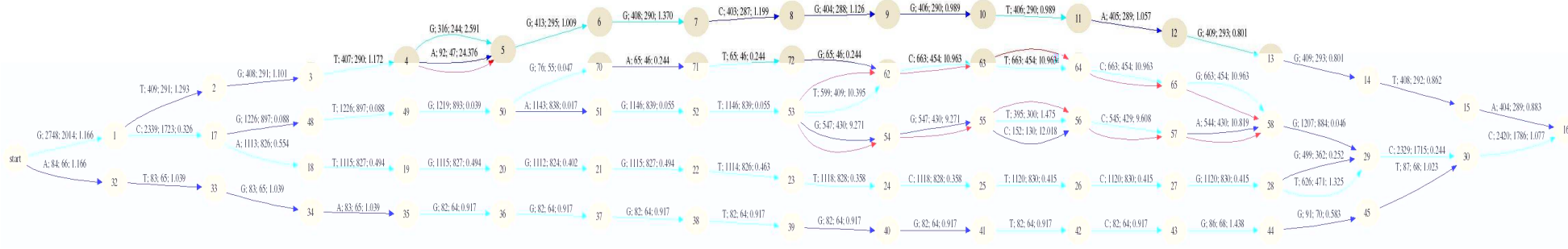
↑ Recombination event

Teste statistique entre clusters haplotypiques



- Résultat montre que le teste à multilocus (■) est plus puissant que celui à un seul marqueur (x)

Identification des patterns haplotypiques protectifs



- Avec $\pm\chi^2 \geq 3.84$, 2 patterns haplotypiques sont identifiés:

  ********11111***
  ********22112***

- Les allèles sont en déséquilibre de liaison (dépendance conditionnelle)

- Le teste statistique de ces patterns donne le résultat plus fort que que celui à un seul marqueur

Conclusion et perspective

Conclusion

- Nous avons proposé une méthode combinatoire basée sur AAMV pour analyser les données polymorphisms
- AAMV a des avantages:
  - Model permet d'analyser multilocus marqueur, d'adapter à DL
  - Model est facile à appliquer avec l'extraction maximale de l'information,
  - Méthode équilibre des degrés de liberté et le nombre de testes
  - Résultats sont simples pour interpréter et robuste pour haplotypes avec fréquence faible
  - Méthode est efficace pour analyser un génome entier
- Algorithmes sont développés en langage PERL, logiciel est disponible à uliliser

Perspective

- Nous allons introduire AAMV de décision pour diagnostiquer:
  - reconnaître la forme d'un haplotype h: $\Pr(h|\text{AAMV}) \neq 0$
  - prédire un haplotype : h est classifié dans la classe A si

$$\frac{\Pr(h, A|\text{AAMV})}{\Pr(h, C|\text{AAMV})} > 1.$$

  - identifier les zônes associés à la maladie correspondant au haplotype h
- Cette méthode pourra être appliquée à l'analyse des
  - données génotypiques
  - données de QTL mapping (Quantitative Trait Loci)
  - puces à ADN avec gènes observés dans les temps différents