Segmentation/Classification de processus.

Application à l'analyse des données de microarrays CGH.

Franck PICARD

UMR INAPG/ENGREF/INRA MIA 518

Sous la direction de J-J. Daudin

16 Novembre 2005

Organisation de la présentation

1. Présentation du contexte biologique.

2. Application des méthodes de segmentation aux données CGH.

3. Développement d'un nouveau modèle de segmentation/classification.

4. Comparaison avec d'autres méthodes.

5. Perspectives.

Organisation de la présentation

1. **Présentation du contexte biologique** :

   $\longrightarrow$ délétion/amplification de séquences d'ADN et microarrays CGH,

   $\longrightarrow$ domaines d'applications,

   $\longrightarrow$ nature du signal étudié,

   $\longrightarrow$ interprétation d'un profil CGH.

2. Application des méthodes de segmentation aux données CGH.

3. Développement d'un nouveau modèle de segmentation/classification.

4. Comparaison avec d'autres méthodes.

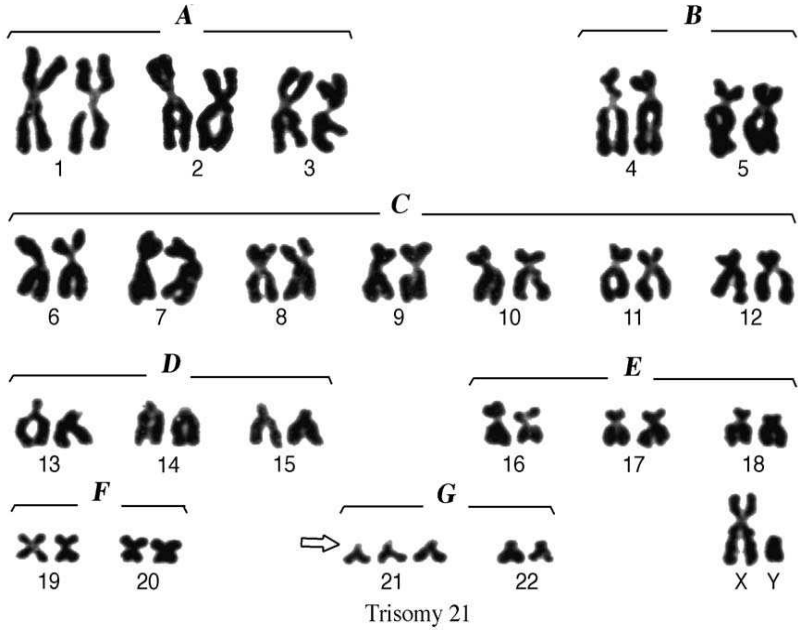
5. Perspectives.

Présentation de la technologie des microarrays CGH

- Réarrangements chromosomiques de grande taille et pathologies humaines :

  → outil d'étude : **caryotype**,

  → résolution ∼ chromosome ∼ 100Mb.



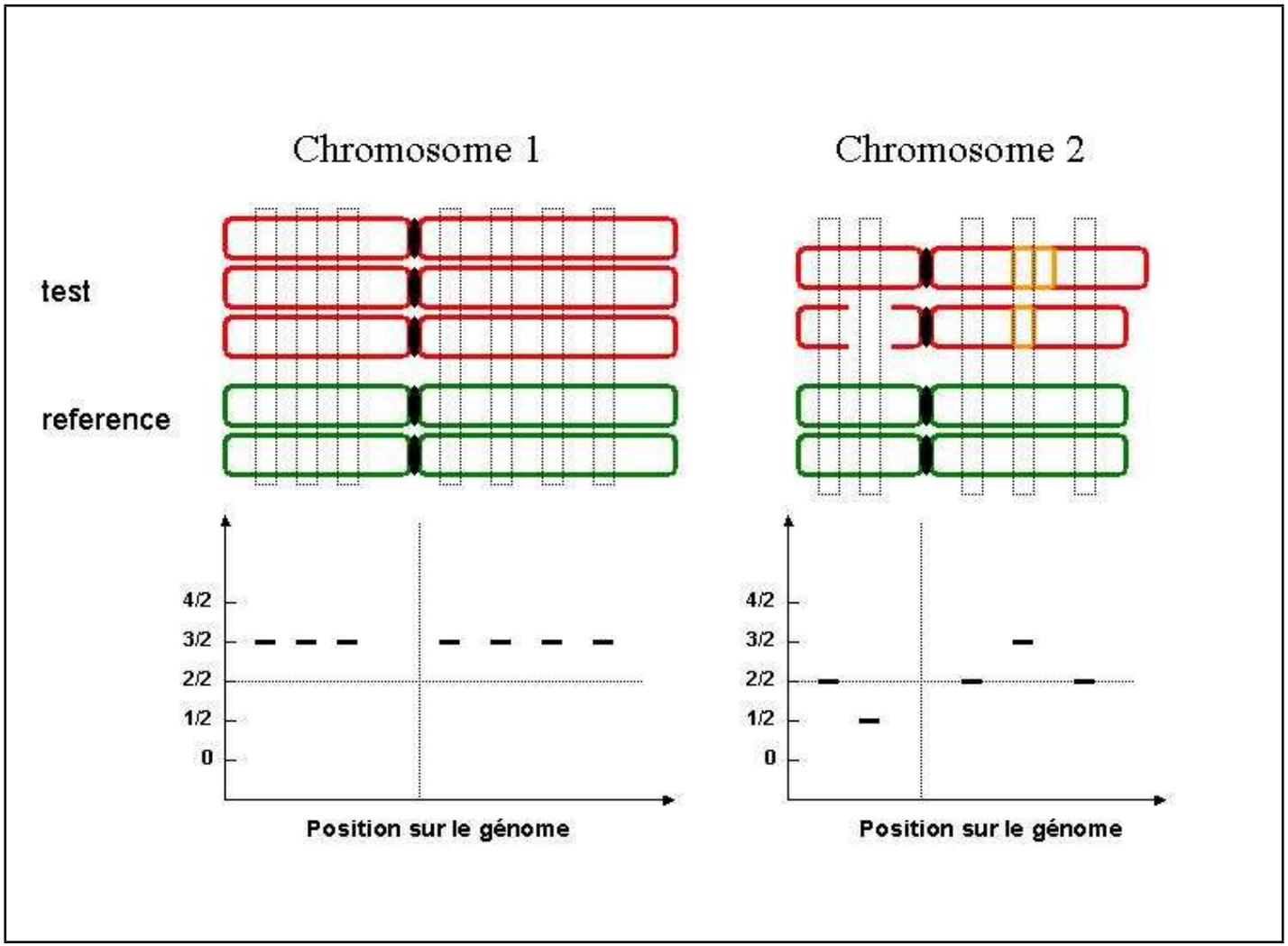
- Délétion/amplification de séquences d'ADN :

  → CGH : Comparative Genomic Hybridization,

  → **microarrays** CGH : 1997,

  → dernière génération de puces : résolution ∼ 100kb.

Présentation simplifiée des données de microarrays CGH



Applications des CGH en génétique humaine

- **Génétique des cancers** :

  $\longrightarrow$ recherche de régions $hotspots$ sur le génome associées aux cancers,

  $\longrightarrow$ portraits moléculaires des tumeurs.

- Nouvelles perspectives pour l'étude du **polymorphisme humain** :

  $\longrightarrow$ variations du nombre de copies de séquences d'ADN de grande taille,

  $\longrightarrow$ comparaison de génomes humains/primates.

$\Rightarrow$ Besoin de nouveaux outils statistiques.

Nature du signal étudié

- Le phénomène biologique étudié est discret :

  $\longrightarrow$ **comptage** de copies de séquences d'ADN.

- Le nombre de copies possible est inconnu.

- Différentes sources de variabilité :

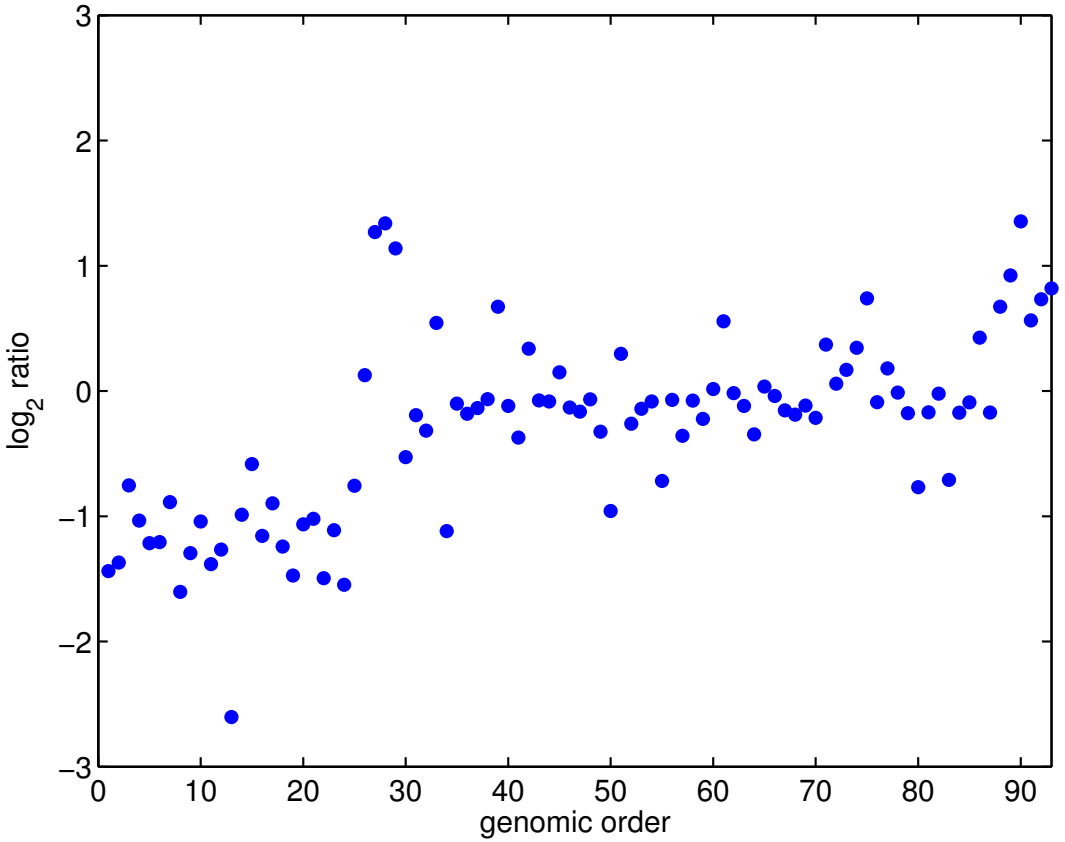
  $\longrightarrow$ variabilité **technique** (ex : saturation),

  $\longrightarrow$ variabilité **biologique** (ex : hétérogénéité des tissus).

- Le nombre de copies est quantifié par fluorescence :

  $\longrightarrow$ le signal étudié est **continu**.

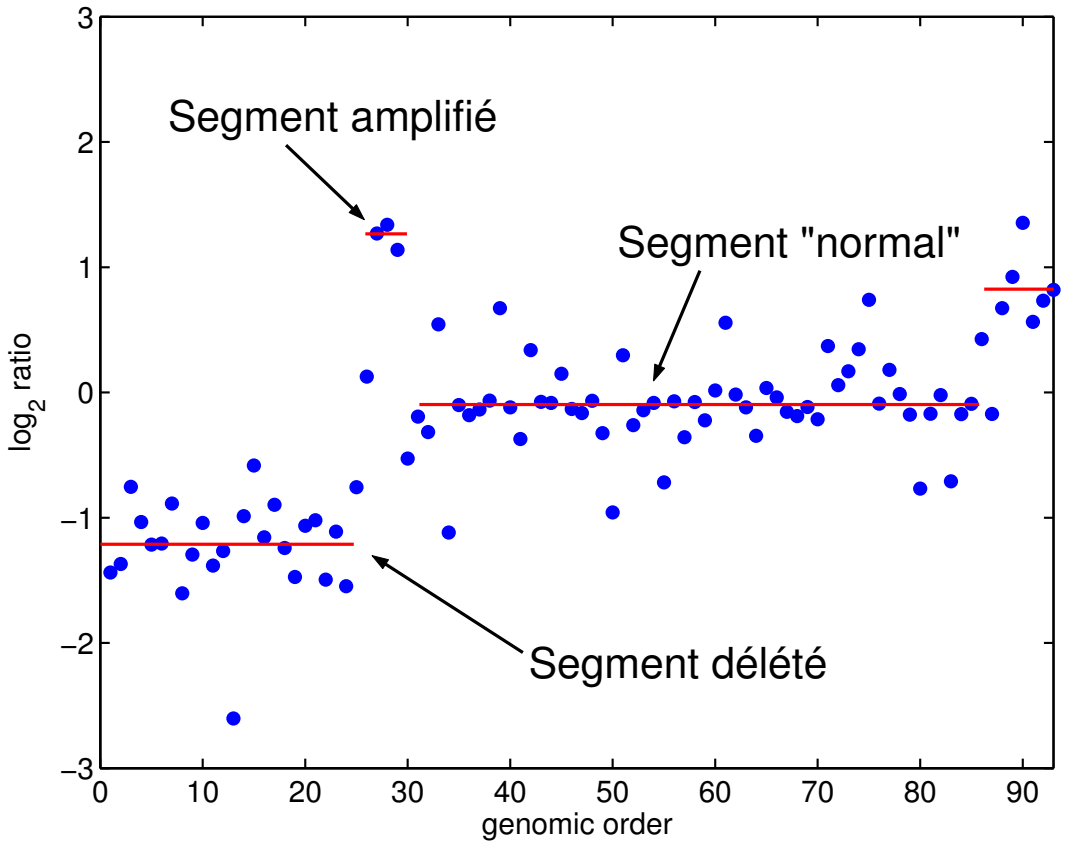
Interprétation d'un profil CGH



Un point sur le graphique représente

$$\log_2 \left\{ \frac{\text{signal à la position } t \text{ dans le génome test}}{\text{signal à la position } t \text{ dans le génome référence}} \right\}$$

Interprétation d'un profil CGH



Une succession de **"segments"** :

zones du génome où le signal est homogène en moyenne.

Organisation de la présentation

1. Présentation du contexte biologique.
2. **Application des méthodes de segmentation aux données CGH :**

   $\longrightarrow$ présentation des modèles de segmentation,

   $\longrightarrow$ estimation des paramètres et sélection de modèle,

   $\longrightarrow$ applications aux CGH.
3. Développement d'un nouveau modèle de segmentation/classification.
4. Comparaison avec d'autres méthodes.
5. Perspectives.

Détection de ruptures dans un signal gaussien

- $Y = \{Y_1, ..., Y_n\}$ un processus gaussien, Y_t indépendantes.
- On suppose que les paramètres de la loi des Y sont affectés par $K-1$ changements abrupts à des instants inconnus $T = \{t_1, ..., t_{K-1}\}$.
- Ces instants de ruptures définissent une partition des données en K segments :

$$I_k = \{t, t \in]t_{k-1}, t_k]\}, \ Y^k = \{Y_t, t \in I_k\}.$$

- On suppose que les paramètres sont constants entre deux ruptures :

$$\forall t \in I_k, \ \mathbb{E}(Y_t) = \mu_k, \ \mathbb{V}(Y_t) = \sigma_k^2.$$

- Les paramètres de ce modèle sont :

  $\rightarrow T = \{t_1, ..., t_{K-1}\}$,

  $\rightarrow \Theta = (\theta_1, \ldots, \theta_K)$, $\theta_k = (\mu_k, \sigma_k^2)$.

Estimation des paramètres et sélection de modèle

- Log-vraisemblance du modèle :

$$\log \mathcal{L}_K(T, \Theta) = \sum_{k=1}^{K} \log f(y^k; \theta_k) = \sum_{k=1}^{K} \sum_{t \in I_k} \log f(y_t; \theta_k).$$

- Estimation des paramètres à K fixé par maximum de vraisemblance :

  $\longrightarrow$ optimisation par programmation dynamique (complexité algorithmique $\mathcal{O}(n^2)$),

  $\longrightarrow$ optimum global.

- Sélection de modèle : choix de K.

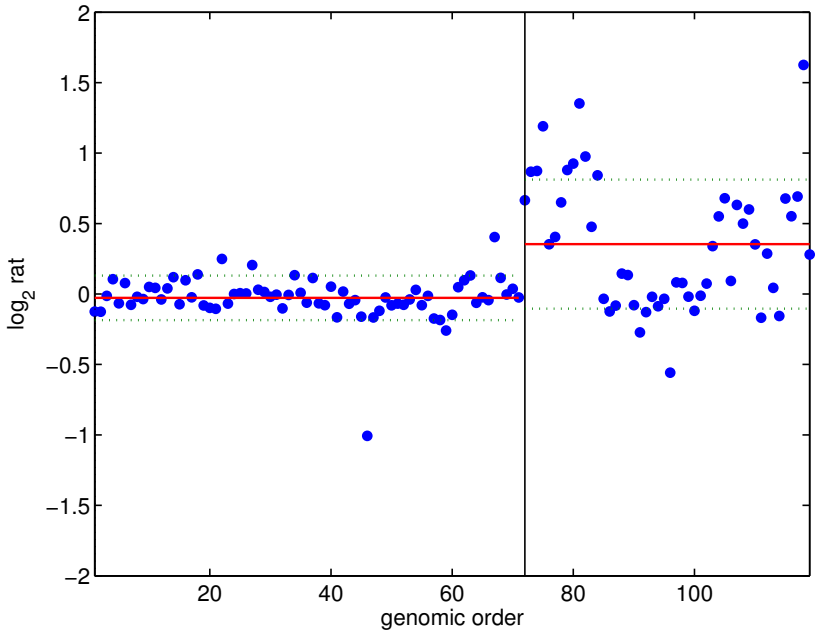
  $\longrightarrow$ Vraisemblance pénalisée : $\hat{K} = \underset{K \geq 1}{Argmax} \left(\log \hat{\mathcal{L}}_K - \beta \times pen(K) \right)$.

  $\longrightarrow$ **Objectif** : établir un compromis entre bon ajustement du modèle aux données et un nombre raisonnable de paramètres à estimer.

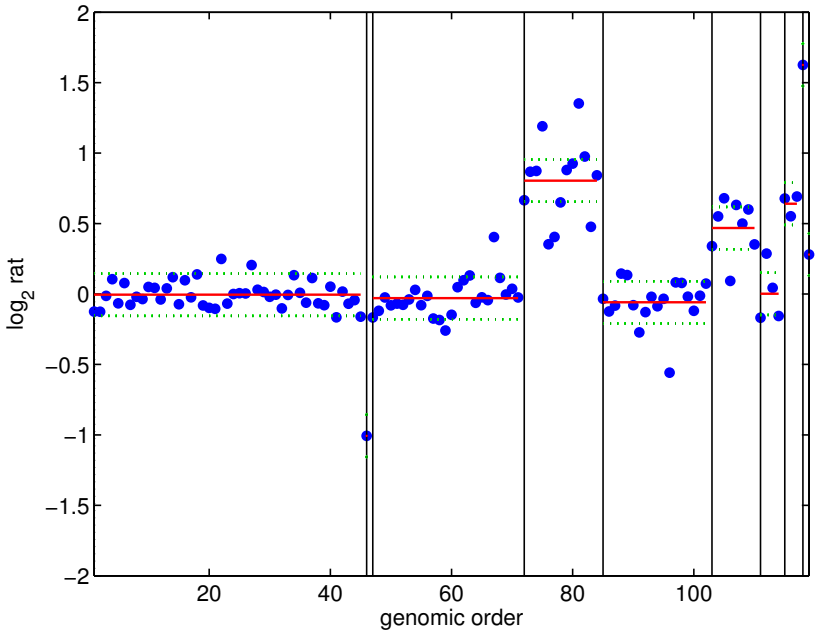
Application des méthodes de segmentation aux données de microarrays CGH

- Quels sont les paramètres du modèle affectés par des changements abrupts ?

  $\longrightarrow$ Modélisation de la variance.



Variances hétérogènes



Variance homogène

Application des méthodes de segmentation aux données de microarrays CGH

- Quelle méthode pour sélectionner le nombre de segments ?

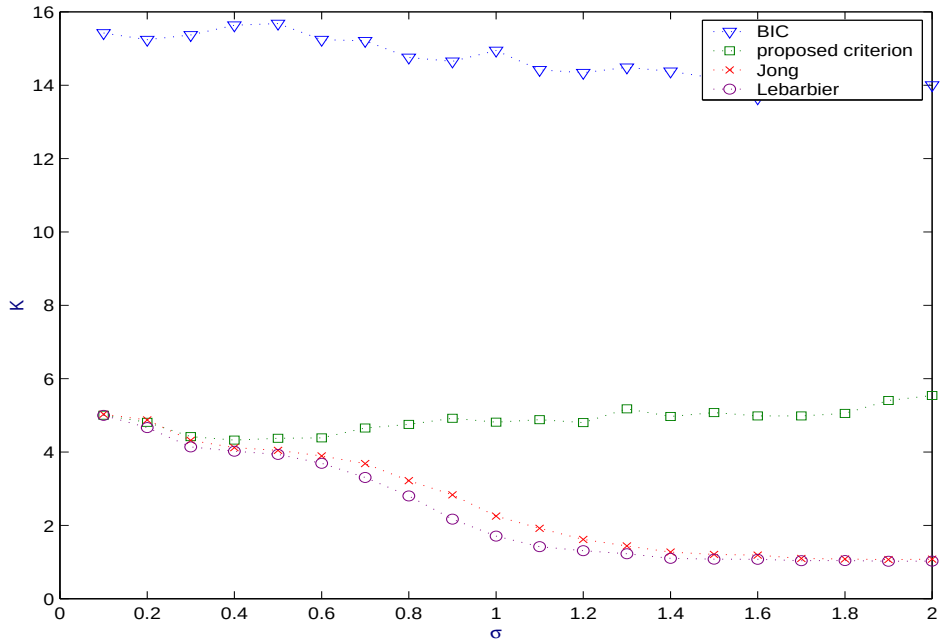
  $\longrightarrow$ Méthode adaptative proposée par Lavielle (2005).

▽ BIC

○ Lebarbier

× Jong

□ Lavielle



Application des méthodes de segmentation aux données de microarrays CGH

- Quels sont les paramètres du modèle affectés par des changements abrupts ?

  $\longrightarrow$ Moyenne à variance constante.

- Quel algorithme d'optimisation de la vraisemblance ?

  $\longrightarrow$ Programmation dynamique (optimum global).

- Quelle méthode pour sélectionner le nombre de segments ?

  $\longrightarrow$ Méthode adaptative proposée par Lavielle (2005).

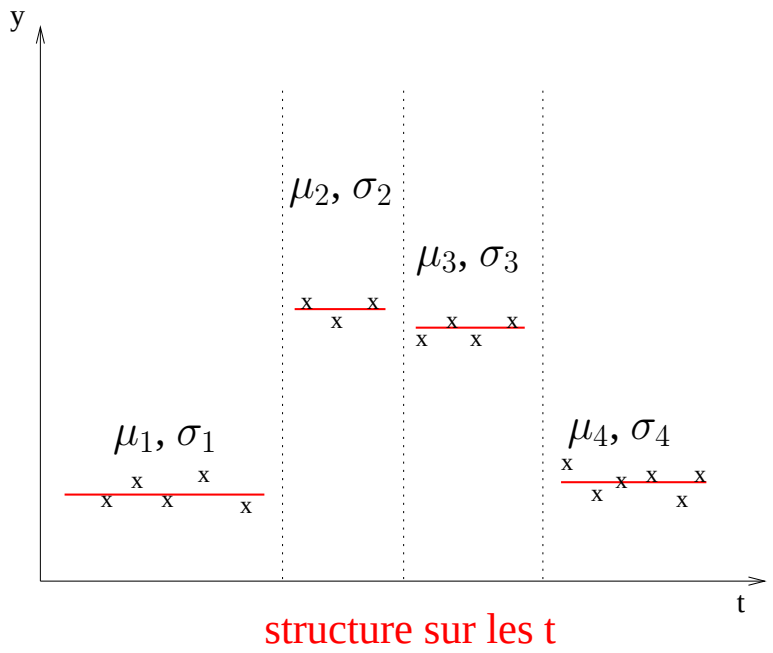
$\Rightarrow$ Publication dans BMC Bioinformatics (Fev 2005)

$\Rightarrow$ Citation dans Lai et. al (2005)

Organisation de la présentation

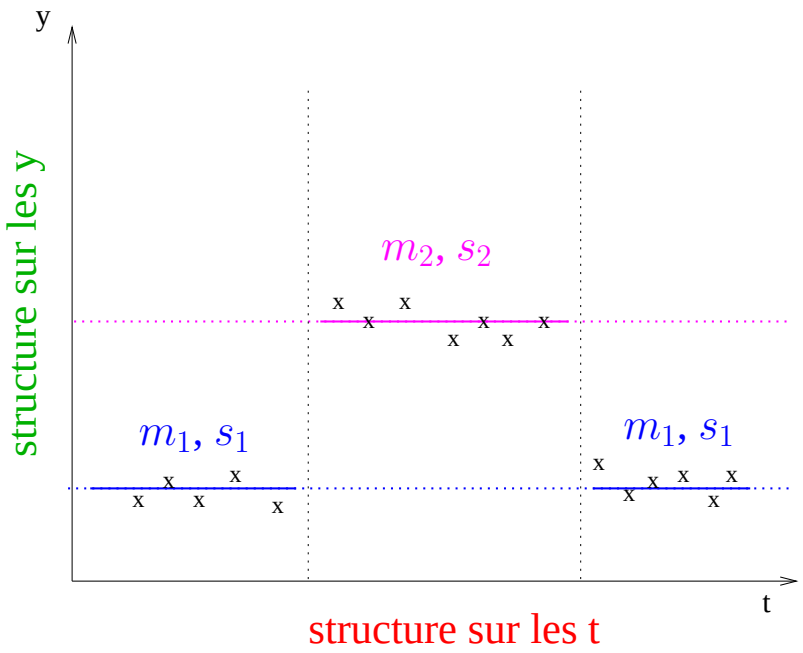
1. Présentation du contexte biologique.
2. Application des méthodes de segmentation aux données CGH.
3. **Développement d'un nouveau modèle de segmentation/classification** :
   - $\longrightarrow$ présentation du modèle,
   - $\longrightarrow$ construction d'un algorithme d'estimation,
   - $\longrightarrow$ construction d'une heuristique de sélection de modèle.
4. Comparaison avec d'autres méthodes.
5. Perspectives.

Attentes des biologistes et nécessité d'un nouveau modèle



Segmentation: structure spatiale du signal

$\theta_k = (\mu_k, \sigma_k^2)$



Segmentation/Classification

$\theta_p = (m_p, s_p^2)$

Modèle de segmentation/classification

- On suppose qu'il existe une **deuxième structure sous-jacente** des segments en P populations de poids $\pi_1, ..., \pi_P$.
- On introduit des variables cachées, Z_p^k indicatrices de la population d'appartenance **du segment** k .
- Ces variables sont supposées indépendantes de loi multinomiale :

$$(Z_1^k, \ldots, Z_P^k) \sim \mathcal{M}(1; \pi_1, \ldots, \pi_P).$$

- Conditionnellement aux variables cachées, on connait la loi des Y :

$$Y^k | Z_p^k = 1 \sim \mathcal{N}_{n_k}(\mathbb{1}_{n_k} m_p, s_p^2 I_{n_k}).$$

- Les paramètres de ce modèle sont :

$\rightarrow T = \{t_1, ..., t_{K-1}\}$,

$\rightarrow \Theta = \{\pi_1, \ldots, \pi_P \,;\, \theta_1, \ldots, \theta_P\}$, avec $\theta_p = (m_p, s_p^2)$.

Définition des unités statistiques du modèle

- On observe n données $\{Y_t\}$ structurées en K segments.
- Les K segments sont structurés en P groupes :

  $\longrightarrow$ les unités statistiques du modèle de mélange sont des segments de différentes tailles,

  $\longrightarrow$ les unités statistiques du mélange changent avec les paramètres de segmentation et le nombre de segments.
- Les données complètes de ce modèle s'écrivent :

$$X^k = (Y_{t_{k-1}+1}, \ldots, Y_{t_k}, Z^k).$$

Algorithme hybride d'optimisation de la vraisemblance

- **Estimation alternée des paramètres à K et P fixés**

1. À T fixé, l'algorithme **EM** optimise la vraisemblance en Θ :

$$\hat{\Theta}^{(\ell+1)} = \underset{\Theta}{Argmax} \left\{ \log \mathcal{L}_{KP} \left(\Theta, T^{(\ell)} \right) \right\} .$$

2. À Θ fixé, la **programmation dynamique** optimise la vraisemblance en T :

$$\hat{T}^{(\ell+1)} = \underset{T}{Argmax} \left\{ \log \mathcal{L}_{KP} \left(\hat{\Theta}^{(\ell+1)}, T \right) \right\} .$$

- **Une suite croissante de vraisemblances :**

$$\log \mathcal{L}_{KP}(\hat{\Theta}^{(\ell+1)}; \hat{T}^{(\ell+1)}) \geq \log \mathcal{L}_{KP}(\hat{\Theta}^{(\ell)}; \hat{T}^{(\ell)}).$$

Initialisation de l'algorithme

- Algorithme itératif : nécessité d'une double initialisation :
  - $\longrightarrow \Theta^{(0)}$ les paramètres du mélange,
  - $\longrightarrow T^{(0)}$ les coordonnées des ruptures.
- Proposition d'une méthode hiérarchique pour initialiser EM.
- Étude de sensibilité à l'étape d'initialisation :
  - $\longrightarrow$ l'algorithme est sensible à la méthode d'initialisation,
  - $\longrightarrow$ il n'existe pas de meilleure méthode (multicritères),
  - $\longrightarrow$ choix de la méthode hiérarchique.
- Proposition d'une méthode pour éviter les maxima locaux.

Sélection de modèle

- Nouveau problème :

  $\longrightarrow$ choix simultané de P et K.

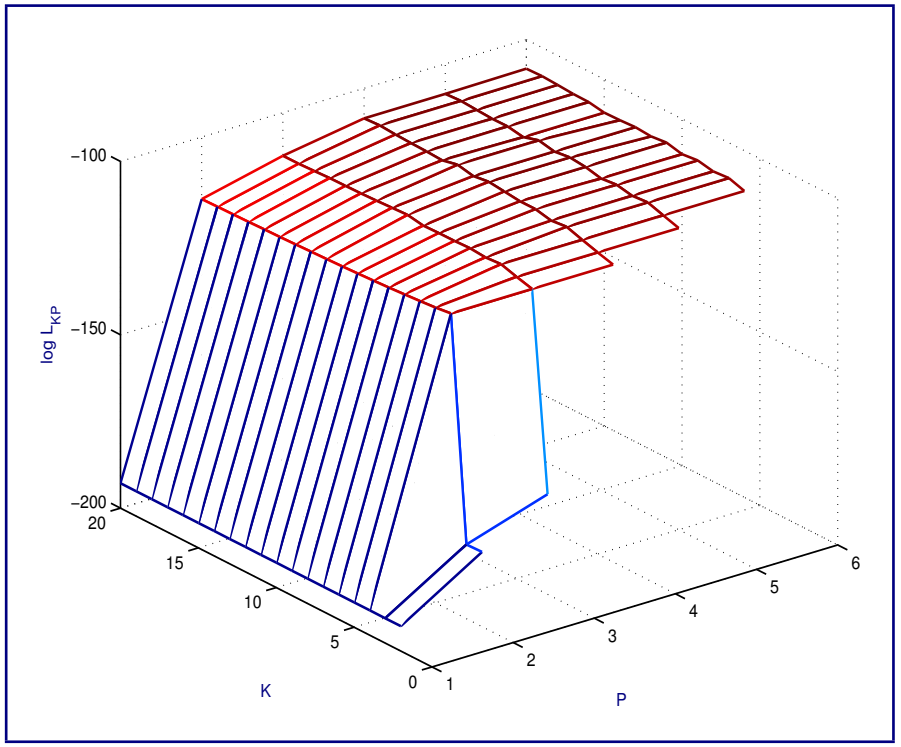
- Méthode :

  $\longrightarrow$ vraisemblance pénalisée.

- Paramètres de différentes natures :

  $\longrightarrow$ Θ paramètres continus,

  $\longrightarrow$ T paramètres discrets.



$\Rightarrow$ Les méthodes classiques de pénalisation ne peuvent pas être appliquées dans ce cadre.

Propriété du modèle

- Modèles emboîtés :

$$\begin{cases} \mathcal{M}(K,P) \not\subset \mathcal{M}(K+1,P), \\ \mathcal{M}(K,P) \subset \mathcal{M}(K,P+1). \end{cases}$$

- Propriété du modèle : $\mathcal{M}(P) = \bigcup_{K \geq 1} \mathcal{M}(K,P)$,

$$\mathcal{M}(P) \subset \mathcal{M}(P+1).$$

$\Rightarrow$ Choisir P dans un premier temps et choisir K ensuite.

Méthode heuristique de sélection de modèle

1. Construction d'une suite croissante de vraisemblances :

$$\log \widetilde{\mathcal{L}}_1 \ldots \leq \log \widetilde{\mathcal{L}}_P \leq \ldots \log \widetilde{\mathcal{L}}_{P_{max}},$$

$$\log \widetilde{\mathcal{L}}_P = \max_K \left\{ \log \mathcal{L}_{KP}(\hat{T}; \hat{\psi}) \right\}.$$

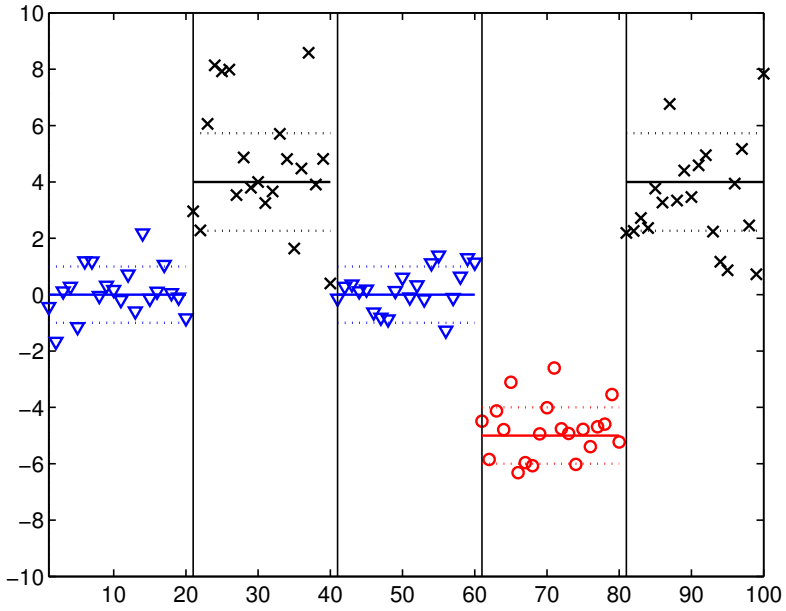
2. Choix du nombre de groupes :

$$\hat{P} = \operatorname{argmax}_P \left\{ \log \widetilde{\mathcal{L}}_P - \beta pen(P) \right\}.$$

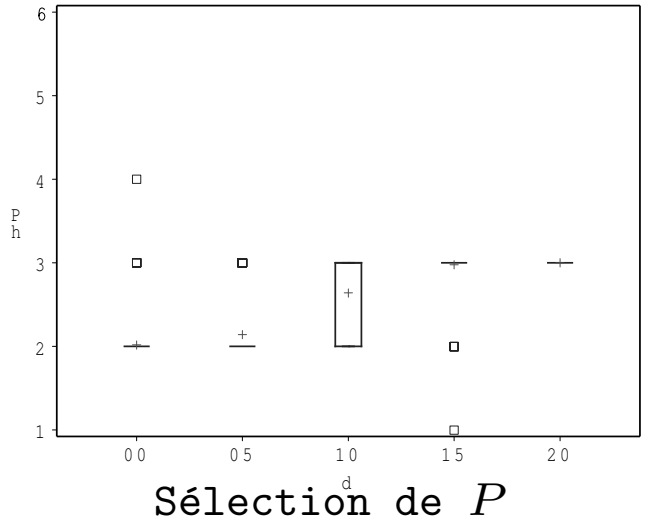
3. Choix du nombre de segments :

$$\hat{K}_{\hat{P}} = \operatorname{argmax}_K \left\{ \log \mathcal{L}_{K\hat{P}}(\hat{T}; \hat{\psi}) - \frac{1}{2} \log(n) \times K \right\}.$$

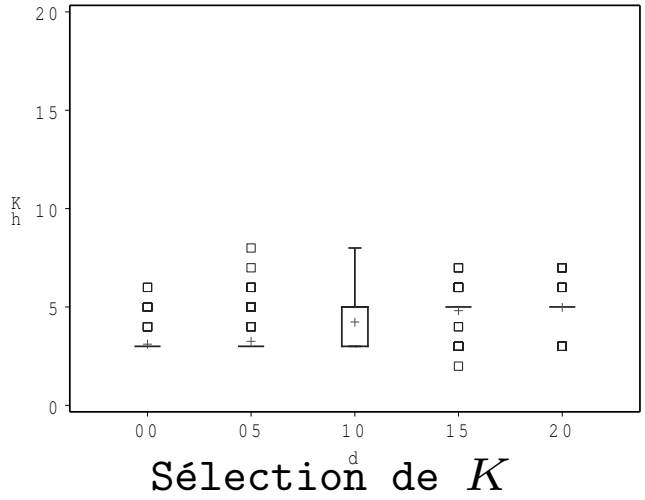
Étude de performances par simulations



- Facteurs de variation :
  - $\longrightarrow$ taille des segments,
  - $\longrightarrow$ distance entre groupes.



Sélection de P



Sélection de K

Conclusion sur le modèle de segmentation/classification

- Présentation d'un nouveau modèle dans le cas gaussien généralisable à d'autres distributions :

  $\longrightarrow$ étude du cas discret avec applications aux séquences d'ADN.

- Développement d'un algorithme hybride :

  $\longrightarrow$ étude de sensibilité à l'étape d'initialisation,

  $\longrightarrow$ méthode heuristique pour les maxima locaux.

- Proposition d'une heuristique de sélection de modèle :

  $\longrightarrow$ méthode séquentielle,

  $\longrightarrow$ analyse de performances par simulations.

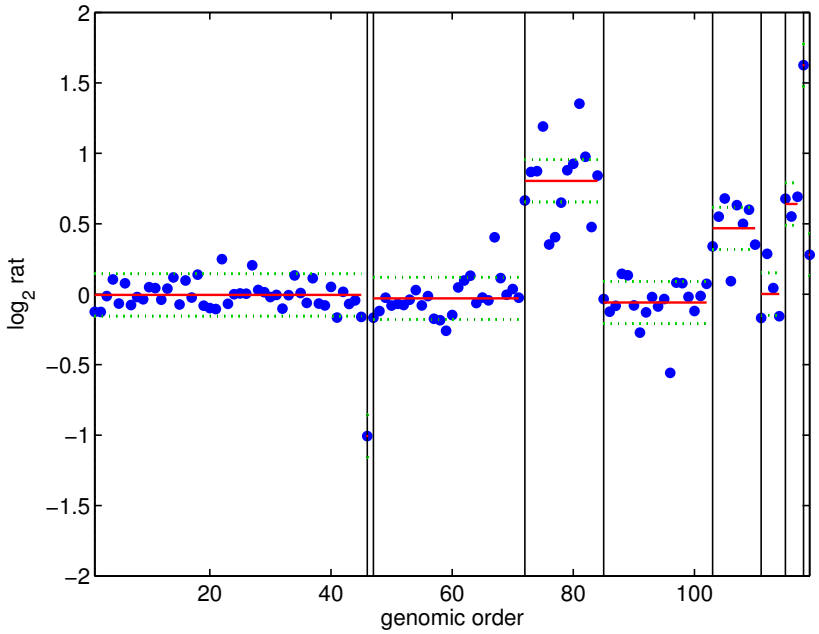
Organisation de la présentation

1. Présentation du contexte biologique.
2. Application des méthodes de segmentation aux données CGH.
3. Développement d'un nouveau modèle de segmentation/classification.
4. **Comparaison avec d'autres méthodes** :

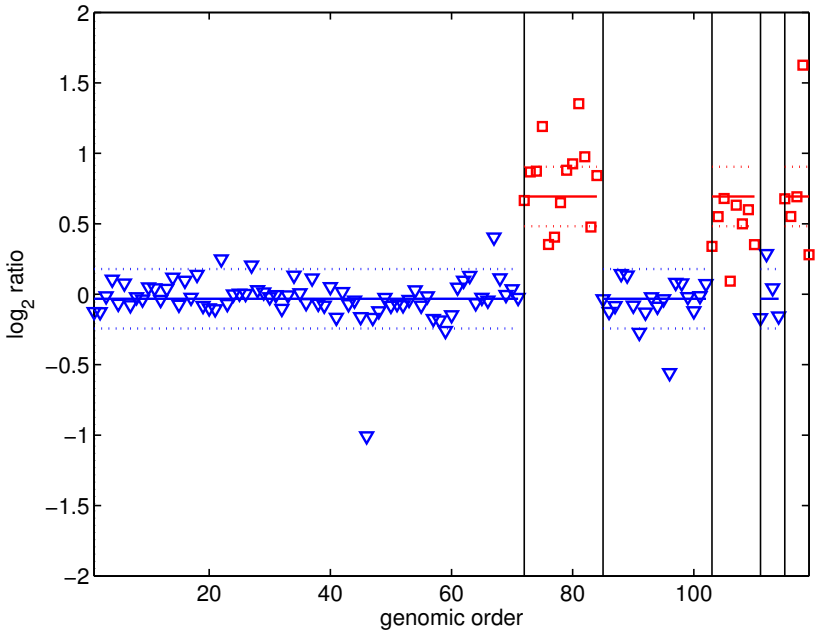
   $\longrightarrow$ segmentation,

   $\longrightarrow$ Chaînes de Markov cachées (HMM).
5. Perspectives.

Segmentation vs segmentation/classification



Segmentation



Segmentation/classification

Comparaison avec les Chaînes de Markov cachées (HMM)

- Modèle à structure cachée :

  $\longrightarrow$ on suppose qu'il existe une séquence de variables cachées $\{Z_t\}$ telle que

  $$Y_t | Z_t = p \sim \mathcal{N}(m_p, s_p^2).$$

  $\longrightarrow$ Dépendance spatiale des $\{Z_t\}$ modélisée à l'aide d'une chaîne de Markov :

  $$\Pr\{Z_t = \ell | Z_{t-1} = p\} = \phi(p, \ell).$$

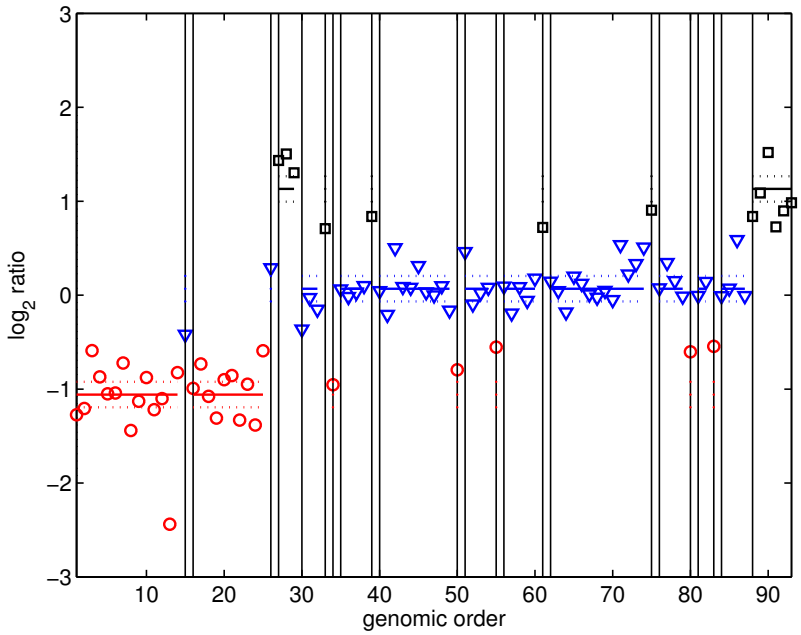
  $\longrightarrow$ Les HMMs modélisent implicitement la taille des "segments" (loi géométrique).

- Comparaison avec le modèle de segmentation/classification :

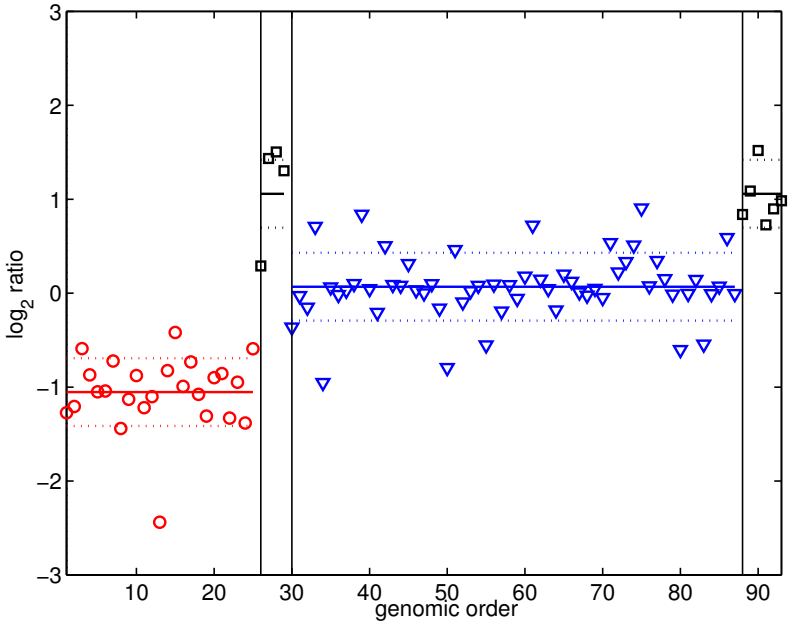
  $\longrightarrow$ la structure spatiale est modélisée grâce au modèle de segmentation,

  $\longrightarrow$ les ruptures sont des paramètres qui sont estimés.

HMMs vs segmentation/classification



HMM



segmentation/classification

- **Analyse des données CGH** :

  $\longrightarrow$ prendre en compte l'ensemble des chromosomes dans la procédure de segmentation,

  $\longrightarrow$ analyser les profils CGH de plusieurs patients simultanément,

  $\longrightarrow$ segmentation sur données dépendantes (nouvelles génération de puces).

- **Méthodes de segmentation** :

  $\longrightarrow$ intervalles de confiance pour les paramètres des ruptures.

- **Segmentation/classification** :

  $\longrightarrow$ développer un critére théorique pour la sélection de modèle,

  $\longrightarrow$ approche bayésienne (modèles hiérarchiques).