Résumé de thèse

Approches statistiques en segmentation : application à la ré-annotation de génomes

Alice Cleynen

Introduction

Cette thèse est principalement dédiée au développement de méthodes de segmentation pour le contexte biologique que constitue l'annotation de génome par les données de séquençage du transcriptome (RNA-Seq). Nous nous sommes intéressés à plusieurs contextes biologiques permettant de passer d'un traitement global du génome à des échelles de plus en plus locales pour lesquelles nous souhaitons obtenir des résultats plus précis et avec plus de certitude. Nous avons pour ce faire développé trois algorithmes de complexité différentes grâce auxquels nous obtenons des résultats statistiques de plus en plus fins. Notre manuscrit s'organise donc naturellement autour de cette échelle, mais nous y présentons les constructions successives et les liens entre chaque méthode proposée.

Dans un premier temps, nous présentons le contexte biologique de notre étude. Nous décrivons en particulier le dogme central de la biologie moléculaire, c'est à dire le modèle général de transmission de l'information génétique, ainsi que les récentes technologies Next-Generation Sequencing (NGS) qui permettent d'obtenir les données RNA-Seq. Cette description souligne l'importance et le gain en précision non-négligeable de ces dernières par rapport aux données provenant de puces, technologie utilisée encore aujourd'hui mais qui est peu à peu abandonnée au profit du NGS. Nous présentons également le jeu de données fourni par le laboratoire de Gavin Sherlock à Stanford, provenant d'une étude sur la levure, et qui nous servira à illustrer nos contributions tout au long de ce manuscrit.

Nous définissons ensuite le domaine statistique de la segmentation comme l'analyse de séries temporelles ayant pour but l'identification d'instants, que l'on appelle ruptures, tels que la distribution des données avant et après cet instant soit différente. Une segmentation

se définit alors comme la donnée d'une partition m de l'ensemble $\{1, \ldots, n\}$, où n est la longueur de la série, et d'un ensemble de distributions qui soient spécifiques à chaque segment. Ainsi, nous noterons

- K le nombre de segments de m,
- τ_k la $k^{ième}$ rupture, avec $0 \leq k \leq K$ et les conventions $\tau_0 = 1$, $\tau_K = n + 1$,
- $m = (\tau_1, \ldots, \tau_{K-1})$ et
- f_k la distribution des données dans le $k^{ième}$ segment $[\![\tau_{k-1}, \tau_k[\![$ noté génériquement J,

de sorte qu'une segmentation se définit comme l'ensemble

$$\{m, \{f_k\}_{1 \leq k \leq K}\}.$$

Une méthode statistique de segmentation est alors la combinaison de trois problèmes :

(i) un problème de modélisation, qui consiste à définir l'ensemble des distributions envisageables pour décrire la série,

(ii) un problème d'inférence, qui consiste à réaliser l'estimation des paramètres de la distribution, des positions des ruptures, et à choisir le nombre de segments, et

(iii) un problème algorithmique, qui consiste à développer des algorithmes pour résoudre le problème d'inférence de manière efficace.

Nous décrivons dans l'introduction un grand nombre d'approches développées dans la littérature pour la résolution de ces problèmes. Cependant nous nous concentrons dans cette thèse sur un contexte paramétrique où les données sont modélisées à l'aide de la loi binomiale négative, où le critère statistique est la log-vraisemblance, et sur des algorithmes de segmentation exacte, c'est à dire qui optimisent exactement le critère défini sans avoir recours à des techniques d'approximation (comme l'utilisation d'algorithmes itératifs) ou de modification du critère (comme sa convexification). Les raisons de ces choix sont partiellement détaillées dans cette synthèse, et plus approfondies dans l'introduction du manuscrit.

Deux problèmes majeurs sont caractéristiques de notre type de données : leur nature discrète, directement liée à la quantité qu'elles mesurent, et leur taille, en raison de leur résolution à l'échelle du nucléotide et de la longueur de certains chromosomes.

Nature discrète des données. Ce premier point pourrait être contourné en appliquant des techniques de transformation des données comme il a toujours été fait pour l'analyse des puces. En effet, de nombreuses méthodes de normalisation et de transformation pour les données NGS ont été proposées, (une comparaison des principales méthodes peut par exemple être trouvée dans Bullard et al. (2010)), et de multiples algorithmes de segmentation sont dédiés aux données modélisées par la loi normale. Cependant, ces transformations se justifiaient pour l'analyse des puces puisqu'il s'agissait toujours de comparer le profil provenant de l'individu ciblé à celui d'un individu de référence. Notre cadre se distingue car nous souhaitons identifier les régions significatives (exprimées) sur la seule base du profil d'intérêt. Même lorsqu'il s'agira de comparer les localisations de transcrits pour des profils différents, chacun d'entre eux sera segmenté de manière indépendante et la normalisation ne sera pas obligatoire. Nous avons donc souhaité conserver les données brutes et les modéliser à l'aide de distributions discrètes. Une étude postérieure (en partie présentée dans le chapitre II-2-1) de comparaison entre notre modèle et son équivalent pour les mêmes données transformées et modélisées par la loi normale nous donnera raison ; en effet, si dans la majorité des cas les résultats sont équivalents, nous avons pu mettre en évidence des situations dans lesquelles notre modèle permettait d'identifier des régions non décelées par l'approche transformée.

Se posait alors la question de la distribution à utiliser pour modéliser ces données. Le modèle le plus simple proposé par la loi de Poisson ne parvient pas à tenir compte de la grande dispersion intrinsèque à ces données, et c'est ce que nous retrouverons tout au long des comparaisons effectuées entre des méthodes utilisant cette distribution et nos approches (par exemple dans les parties II-1-2 et II-2-1). Deux autres lois ne nécessitant qu'un paramètre supplémentaire sont plus flexibles : la Zero-Inflated Poisson, qui est un mélange entre une loi de dirac en zéro et une loi de Poisson, et la binomiale négative, que nous décrivons dans l'introduction, et qui peut se voir comme une loi de Poisson dont le paramètre serait une variable aléatoire de loi Gamma. Si aucune des deux n'appartient telle quelle à la famille exponentielle, la seconde y appartient à condition que le paramètre de dispersion ϕ soit fixé. De plus, nous observons dans certains segments correspondant à des gènes très fortement exprimés une grande dispersion mais une absence de masse

en 0 dans leur distribution. C'est donc vers la binomiale négative que nous nous sommes tournés, d'autant qu'elle est de plus en plus couramment utilisée pour modéliser des données de type RNA-Seq dans les études d'expression différentielle (Robinson et al., 2010; Risso et al., 2011).

Cela nous a donc conduit à définir le modèle de segmentation suivant :

$$\forall J \in m,\ \forall t \in J,\ Y_t \sim \mathcal{NB}(p_J, \phi)$$

où m est une partition du signal, J un segment de m, ϕ est le paramètre de dispersion de la binomiale négative, que nous supposons constant et connu, et p_J est son paramètre de probabilité qui est spécifique à chaque segment.

Il faut alors estimer le paramètre ϕ au préalable dans les données pour pouvoir l'utiliser en tant que paramètre connu dans les modèles de segmentations que nous proposons. Ceci ne peut se faire de manière naïve puisque par définition nous nous attendons à ce que le signal soit fragmenté en segments de distributions différentes bien que de même dispersion. Nous avons donc cherché à utiliser un estimateur robuste basé sur des fenêtres glissantes, comme l'estimateur de déviation absolue à la médiane (Median absolute Deviation, MAD, Hampel, 1974; Donoho, 1995) ou celui de Hall et al. (1990) pour l'estimation de la variance. Cependant, l'expression compliquée du biais des estimateurs classiques (comme l'estimateur du maximum de vraisemblance, ou celui des moments), et la nature discrète des données ne nous ont pas permis de proposer un estimateur satisfaisant. Nous avons donc choisi d'utiliser un simple estimateur des moments sur des fenêtres glissantes dont nous conservons la médiane. Cet estimateur est décrit dans l'introduction, et bien qu'il n'influence que peu la qualité de nos résultats, il demeure le point faible de notre contribution. Cependant, si un meilleur estimateur était proposé, il serait très facile de l'intégrer à notre contribution sans aucune autre modification.

Taille des données. Ce deuxième point est plus délicat que le premier car il va très fortement limiter la gamme des algorithmes envisageables. En effet, puisque la résolution des données est celle du nucléotide, leur taille peut aller jusque celle de chromosomes, soit jusqu'à des valeurs de n de l'ordre de 10^8. Dans un tel cas, seuls des algorithmes de

complexité linéaire en temps vont être en mesure de traiter le profil, sous condition d'une part que la complexité en espace ne soit pas trop grande (ces notions seront discutées par la suite), et d'autre part que les constantes multiplicatives ne soient pas non plus prohibitives.

Comme dans de nombreux domaines des statistiques, la question qui se pose alors est 'quelle information pouvons nous obtenir grâce à nos algorithmes, et à quel prix ?' Par exemple, nous souhaitons obtenir les meilleures ruptures possibles dans la série que nous étudions, mais nous aimerions également avoir une idée de la qualité de ces ruptures, c'est à dire de l'incertitude qui leur est associée. Or nous ne connaissons aucun algorithme en temps linéaire qui puisse à la fois aborder l'incertitude liée aux paramètres et celle liée à la localisation des ruptures. Ceci n'est cependant pas nécessairement une limitation lors de l'étude de l'annotation précise de gènes. Nous avons donc construit différentes approches pour répondre aux critères (i) à (iii) des méthodes de segmentation présentés ci-dessus, dans deux cadres biologiques complémentaires : l'analyse de l'entièreté du génome, et l'annotation à l'échelle du gène.

Analyse de l'intégralité du génome

Cette étude fait l'objet du chapitre II-1 de notre manuscrit. Nous sommes ici typiquement dans le cadre des séries dont la longueur est un facteur limitant, même si dans le cas de nos données de référence, correspondant au génome de la levure, nous sommes au plus de l'ordre de 10^6.

Problème de modélisation (i)**.** Bien que toutes les méthodes que nous avons développées lors de cette thèse soit valable pour diverses distributions, nous supposerons dans cette étude que le problème de modélisation est résolu par l'utilisation de la loi binomiale négative de paramètre de dispersion connu. En effet, utiliser les mêmes outils avec la distribution de Poisson, incapable de prendre en compte la dispersion, conduit à sur-segmenter les séries, tandis que la loi Gaussienne, elle, ne parvient à proposer des résultats similaires que lorsque les données sont transformées à l'aide de la transformation $\sinh^{-1}$ nécessitant l'estimation du paramètre ϕ, seul réel inconvénient à l'utilisation de la binomiale négative.

Problème algorithmique (*iii*)**.** Pour procéder à l'inférence, il est nécessaire d'estimer le nombre de segments K et d'explorer l'intégralité de l'espace des segmentations $\mathcal{M}_K$. Si l'on suppose K connu, il y a $\binom{n-1}{K-1}$ partitions distinctes de $\{1, \ldots, n\}$ en K segments, et nous avons ici un n très grand. Il nous faut donc une approche d'exploration qui ne soit pas naïve (sinon la complexité serait en $\mathcal{O}(n^K)$), et c'est dans cet esprit que l'algorithme de programmation dynamique (DP, Bellman, 1961) a été développé. Malheureusement, étant de complexité quadratique en temps, il reste inapproprié pour notre contexte.

Nous proposons d'adapter pour la binomiale négative l'algorithme proposé par Rigaill (2010), le *pruned Dynamic Programming algorithm* (PDPA), soit l'algorithme de programmation dynamique élagué. Ce choix est motivé à la fois par l'aspect algorithmique, puisque sa complexité en temps, au pire quadratique en n, est empiriquement moindre, et par les hypothèses nécessaires à son implémentation (en effet, PELT (Killick et al., 2012), le seul autre algorithme de complexité linéaire résolvant notre critère, suppose que le nombre de segments augmente avec n, ce qui n'est pas notre cas). La partie II-1-1 est consacrée à une étude détaillée de cet algorithme et de ses performances. Bien que de complexité dépendant du signal, nous y montrons qu'elle est presque linéaire (en $\mathcal{O}(Kn)$) et de constante raisonnable pour nos données RNA-Seq. Sa complexité en espace étant elle en $\mathcal{O}(Kn)$, cet algorithme est largement utilisable dans notre contexte. Son implémentation en C++, en collaboration avec Michel Koskas, de l'Agroparistech, et de Guillem Rigaill, de l'université d'Evry, a constitué une partie de la thèse, ainsi que sa distribution en un package R sous le nom de `Segmentor3IsBack` pour diverses distributions incluant la binomiale négative.

Cependant, il a deux limitations majeures. La première est qu'il ne permet pas d'évaluer la qualité de la segmentation proposée. Ainsi, nous obtenons la meilleure segmentation en K segments, mais il ne nous informe pas sur la deuxième meilleure. Est-elle radicalement différente, ou au contraire très proche ? Nous proposons une première réponse à cette question en calculant, pour chaque rupture, le coût de la meilleure segmentation en fonction de sa position. Dans les cas où ces courbes ont chacune un minimum marqué, nous pouvons avoir confiance quant à l'optimalité de la segmentation. Sinon cela signifie que la position des ruptures est très incertaine. Si cette approche constitue un excellent critère

visuel rapide pour mesurer la qualité de la segmentation, nous souhaitons développer des critères statistiques permettant de quantifier cette incertitude.

Problème d'inférence (ii)**.** L'autre limitation est qu'il ne permet pas de choisir le nombre de segment K ; au contraire, il obtient pour chaque valeur de k entre 1 et une valeur de K_{max} définie préalablement, la meilleure segmentation en k segments. Ceci est une difficulté récurrente dans les modèles de segmentation. Dans ce contexte, le développement d'une procédure de sélection de modèle qui permettrait de compléter cet algorithme a constitué un problème naturel, pour lequel nous proposons deux solutions dans les chapitres II-1-2 et II-1-3.

En effet, deux familles d'approches pour la sélection de modèle nous paraissaient également intéressantes dans notre contexte. La première est basée sur des considérations non-asymptotiques sur le risque associé au critère de vraisemblance et consiste à obtenir des inégalités oracles sur l'estimateur proposé. La seconde provient du cadre de la classification dans des modèles de données incomplètes, et consiste à choisir le nombre de segments pour lequel l'incertitude sur la segmentation est la plus faible. Nous proposons des critères pour chacune de ces deux approches. Notons cependant que la seconde, bien qu'initialement considérée pour l'analyse de génomes entiers, a une complexité plus importante et s'appliquera plutôt à des séries plus courtes, dont nous discutons par la suite.

Ainsi nous développons tout d'abord une approche de vraisemblance pénalisée inspirée de la littérature de Birgé et Massart (Birgé and Massart, 1997; Barron et al., 1999; Birgé and Massart, 2001, 2007) dans laquelle nous cherchons à déterminer la forme de la pénalité de sorte à obtenir une inégalité oracle pour notre estimateur de la distribution.

Dans notre contexte de segmentation, cette approche s'explique de la manière suivante : si nous simplifions l'ensemble des modèles à la partition sous-jacente m (ce qu'il est naturel de faire lorsque l'inférence des paramètres de la distribution, à partition m connue, se fait trivialement par maximum de vraisemblance) nous obtenons facilement le meilleur estimateur $\hat{s}_m$. Nous souhaitons alors choisir parmi ces $\hat{s}_m$ celui qui va minimiser le risque de Kullback-Leibler à la vraie distribution s, soit ici la binomiale négative $\mathcal{NB}(p_t, \phi)$ avec

ϕ connu. Bien sûr, obtenir cet estimateur $\hat{s}_{m(s)}$ suppose de connaître s, c'est pourquoi nous l'appelons *oracle*. Puisque nous ne pouvons l'atteindre, nous allons chercher à faire presque aussi bien que lui, c'est à dire nous allons tenter de choisir un estimateur $\hat{s}_{\hat{m}}$ vérifiant une *inégalité oracle* de la forme

$$R(s, \hat{s}_{\hat{m}}) \leq CR(s, \hat{s}_{m(s)})$$

où $R(s, u)$ est le risque de Kullback-Leibler entre s et u, et C est une constante que nous espérons aussi proche de 1 que possible. Pour cela, nous proposons de choisir $\hat{m}$ minimisant la log-vraisemblance pénalisée par une fonction *pen* dépendant de la dimension du modèle, et qu'il reste à choisir. En écrivant, pour tout m de notre collection,

$$KL(s, \hat{s}_{\hat{m}}) \leq KL(s, \bar{s}_m) + \bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{\hat{m}}) - pen(\hat{m}) + pen(m).$$

(où KL est la distance de Kullback-Leibler, $\bar{\gamma}$ est la vraisemblance normalisée (*i.e.* $\bar{\gamma} = \gamma - \mathbf{E}(\gamma)$), et $\bar{s}_m$ est la projection de s sur la partition m), on se rend bien compte qu'il va falloir choisir une pénalité suffisamment grande pour que $pen(\hat{m})$ compense les fluctuations de $\bar{\gamma}(\bar{s}_m) - \bar{\gamma}(\hat{s}_{\hat{m}})$, mais pas trop grande pour que $pen(m)$ ne pénalise pas trop la différence entre $KL(s, \bar{s}_m)$ et $KL(s, \hat{s}_{\hat{m}})$.

Cette approche est devenue classique dans les problèmes en grande dimension où des considérations asymptotiques ne sont pas appropriées car la collection des modèles considérés augmente avec la taille de l'échantillon. Nous sommes cependant ici dans une situation nouvelle puisque nous traitons des données qui sont à la fois discrètes et non bornées. Ainsi, les traditionnelles inégalités de concentration ne peuvent s'appliquer directement et des décompositions plus fines ainsi que des résultats de grandes déviations sont nécessaires. En collaboration avec Emilie Lebarbier, de l'Agroparistech, nous avons montré qu'une pénalité de la forme

$$pen(m) = \beta |m| \left(1 + 4\sqrt{1.1 + \log\left(\frac{n}{|m|}\right)}\right)^2,$$

répondait à notre problème, et nous discutons ce résultat dans la partie II-1-2. Combiner cette approche (dont l'application est de complexité $\mathcal{O}(K)$) à l'algorithme précédent permet donc d'obtenir une procédure complète (de complexité finale empiriquement linéaire) pour

la segmentation de données RNA-Seq correspondant à des chromosomes entiers.

Dans un deuxième temps nous nous intéressons au critère de vraisemblance complète intégrée (Integrated Complete Likelihood, ICL Biernacki et al., 2000) comme approche pour le choix du nombre de segments. Introduit dans le cadre de modèles de données incomplètes, l'ICL vise à sélectionner la taille du modèle avec l'incertitude la plus faible. Cela diffère de la plupart des autres approches de sélection de modèles, comme la précédente, qui cherchent à choisir le nombre le plus probable de segments, généralement en s'appuyant sur une approximation des probabilités à posteriori (cf AIC ou BIC), ou en minimisant le risque lié au modèle.

L'expression originale de L'ICL était donnée par

$$\log \mathbb{P}(y, S|\mathcal{M}_K)_{|S=\widehat{S}}$$

où $\mathcal{M}_K$ est l'ensemble des modèles possible ayant K classes, y sont les observations et $\widehat{S}$ est l'estimateur de S, les indicateurs (inconnus) d'appartenance aux classes. McLachlan and Peel (2004) ont plus tard proposé de remplacer $\widehat{S}$ par son espérance conditionnelle sachant les observations, de sorte que la définition de l'ICL qui est maintenant couramment utilisée est la suivante :

$$\begin{aligned} ICL(K) &= \mathbb{E}\left[\log \mathbb{P}(y, S|\mathcal{M}_K)|y\right] \\ &= \log \mathbb{P}(y|\mathcal{M}_K) + \mathbb{E}\left[\log \mathbb{P}(S|y, \mathcal{M}_K)|y\right] \end{aligned}$$

Le terme le plus à droite, ou plus précisément son opposé, est appelé l'*entropie* négative de la classification et est couramment noté $\mathcal{H}(\mathcal{M}_K)$. Cette expression révèle le but de la classification par l'ICL, car elle est équivalente à utiliser la vraisemblance intégrée et pénalisée par un terme qui mesure l'incertitude de la classification. En effet, l'entropie sera plus élevée lorsque tous les modèles de $\mathcal{M}_K$ seront equiprobable compte tenu des données, ce qui équivaut à dire, dans le cadre de mélange, que les composantes du mélange sont mal séparées. Au contraire, une valeur de K ayant un modèle qui est meilleur que tous les autres donnera une entropie proche de zéro.

Dans cette partie, nous nous inspirons de deux contributions majeures dans le domaine de la segmentation. La première, publiée dans Rigaill et al. (2012), est un modèle de segmentation Bayésien qui permet de calculer en temps quadratique et de manière exacte (c'est à dire en prenant compte l'incertitude sur les paramètres) certaines quantités liées au modèle, comme la distribution à posteriori des ruptures. Les auteurs proposent de plus un calcul (toujours exact) de l'ICL qu'ils utilisent comme une méthode efficace pour le choix de K. La deuxième approche est proposée par Luong et al. (2013) et consiste à utiliser un modèle de Markov caché (Hidden Markov Model, HMM) contraint de sorte à correspondre à un modèle de segmentation n'autorisant pas la re-visite des états, au contraire des approches de segmentation par HMM usuelles. Pour un nombre K de segments donné, la complexité de cet algorithme est en $\mathcal{O}(Kn)$ à la fois en temps et en espace, et il permet d'obtenir des mesures sur l'incertitude de la localisation des ruptures conditionnellement aux paramètres du modèle.

Notre contribution a consisté à implémenter, en collaboration avec The Minh Luong et Gregory Nuel de l'université de Paris Descartes, ainsi que Guillem Rigaill de l'université d'Evry, ce deuxième algorithme pour la distribution binomiale négative, et à proposer un calcul du critère ICL conditionnellement aux paramètres du modèle. L'utiliser pour choisir le paramètre K exige son calcul pour toutes les valeurs de K entre 1 et un certain K_{max} défini par l'utilisateur et rend donc la procédure de complexité $\mathcal{O}(K^2 n)$, d'où sa limitation à des profils de taille intermédiaire. Cependant, l'approche de segmentation par HMM contraint permet d'obtenir plus d'informations que la programmation dynamique puisque nous pouvons par exemple obtenir des intervalles de confiance sur la localisation des ruptures, toujours conditionnellement aux valeurs des paramètres. Ces séries de taille intermédiaire sont par exemple obtenues en divisant le chromosome en deux au niveau du centromère ; en effet, cette région est connue pour ne pas présenter de partie codante et nous n'y attendons donc aucune rupture. Ainsi, si nous divisons n et K par deux, nous gagnons alors un facteur 8 et 4 respectivement en temps et en espace.

Annotation d'un gène

Dans la partie II-2 notre but ultime est la comparaison de la localisation des régions transcrites sur le génome d'une espèce qui aurait été élevée dans des environnements différents. Traduire ce problème dans le cadre de la segmentation revient à comparer la localisation de ruptures dans des profils indépendants. Dans la littérature, deux approches sont généralement considérées lorsqu'il s'agit de manipuler plusieurs séries. La première consiste en la segmentation simultanée de toutes les séries, de sorte à identifier des ruptures communes à toutes. Cette approche est équivalente à la segmentation d'une unique série multivariée, mais pourrait permettre la détection de rupture dans un profil dont le niveau de signal aurait été trop faible pour permettre son analyse indépendante. La seconde approche consiste en la segmentation jointe de toutes les séries, chacune ayant son nombre de segments et sa localisation de ruptures spécifiques, mais permettant la prise en compte d'une possible dépendance entre les séries sans pour autant imposer que les changements se produisent simultanément. Nous nous intéressons ici à un troisième type de problème statistique, qui est la comparaison des localisations des ruptures dans des séries qui ont été segmentées séparément. Pour cela, il est nécessaire d'être en mesure de quantifier l'incertitude de la localisation des ruptures.

Nous sommes alors typiquement dans le cas où les approches présentées dans les paragraphes précédents, à savoir la segmentation Bayésienne exacte ainsi que le HMM contraint, vont pouvoir s'appliquer (ici n est de l'ordre de 10^3 puisque nous considérons des gènes, et K est de l'ordre d'au plus la dizaine, puisque nous allons souhaiter séparer les parties codantes (*i.e.* les exons) des parties non codantes (*i.e.* les introns) au sein d'un même gène). Nous avons donc implémenté l'approche de Rigaill et al. (2012) pour la binomiale négative dans un package R du nom de `EBS` (pour Exact Bayesian Segmentation). Nous décrivons dans le chapitre II-2-3 le fonctionnement de ce package ainsi que les difficultés rencontrées et les astuces calculatoires dont nous avons usé.

Il semblait alors naturel de commencer par un état de l'art des méthodes de segmentations disponibles pouvant prendre en compte à la fois la nature discrète de nos données

et l'absence d'un profil de référence pour vérifier la pertinence de nos contributions. Nous montrons ainsi, en collaboration avec Sandrine Dudoit et Stéphane Robin, (étude présentée dans le chapitre II-2-1), que les algorithmes sont plus efficaces lorsque K est connu, ce qu'il n'est pas absurde de considérer lorsque nous disposons déjà d'une annotation de génome approximative que nous cherchons à raffiner. Les algorithmes PDPA et EBS ont alors des performances idéales, tandis que l'approche de HMM contraint, plus rapide que EBS, est légèrement moins bonne et ne représente donc pas un gain dans ce contexte de 'petites données'. Les autres algorithmes, implémentés, eux, uniquement pour la distribution Poisson, ne parviennent pas à égaler nos trois approches.

Nous avons donc par la suite conservé le modèle de segmentation Bayésienne exacte pour procéder à nos comparaisons de localisation, et avons proposé deux approches, en collaboration avec Stéphane Robin, l'une réservée à la comparaison de deux profils, tandis que l'autre s'applique quel que soit le nombre de profils considérés. Commençons par rappeler le modèle sous-jacent dans la Figure 1 : K est tiré dans la loi $P(K)$, puis les paramètres θ du modèle sont tirés de manière indépendante dans $P(\theta|K)$, et de la même manière pour les partitions m dans $P(m|K)$. Enfin, les données sont générées à l'aide de la loi conditionnelle $P(Y|m, K)$. Nous pouvons alors obtenir la distribution à posteriori des ruptures τ_k, que nous notons $p_k(t; Y; K)$.

Dans notre contexte, nous supposons connu, pour chaque profil ℓ, son nombre de segment K^ℓ, ainsi que le numéro k_ℓ de la rupture à comparer, soit $\tau^\ell_{k_\ell}$. Dans le cas de la comparaison de profil, il suffit alors de calculer de manière exacte la distribution du décalage entre les ruptures d'intérêt par simple convolution, c'est à dire

$$\delta_{k_1,k_2}(d; K^1, K^2) = \sum_t p_{k_1}(t; Y^1; K^1) p_{k_2}(t - d; Y^2; K^2).$$

La situation se complique lorsque nous avons plus de deux séries, car si nous notons E_0 l'événement $\{\tau^1_{k_1} = \cdots = \tau^I_{k_I}\}$, la probabilité a priori de cet événement (en utilisant le prior uniforme classiquement utilisé pour les segmentations m) devient rapidement très faible avec la taille et le nombre de profils. Nous proposons donc un modèle incorporant une couche

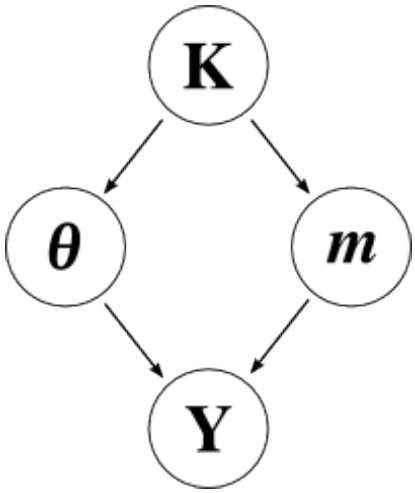

FIGURE 1 – **Graphical model of the exact Bayesian Segmentation approach.** Modèle hiérarchique de l'approche de segmentation Bayésienne exacte proposée par Rigaill et al. (2012).

supplémentaire et qui est résumé dans la Figure 2 : en notant $\mathbf{E} = \mathbb{I}\{E_1\} = 1 - \mathbb{I}\{E_0\}$, nous avons :

- la variable $\mathbf{E}$ est tirée conditionnellement à $\mathbf{K}$ selon $\mathcal{B}(1 - p_0(\mathbf{K}))$ où $p_0(\mathbf{K}) = P(E_0|\mathbf{K})$;
- les paramètres $\boldsymbol{\theta}$ restent tirés selon $P(\boldsymbol{\theta}|\mathbf{K})$;
- les partitions sont tirées conditionnellement à la fois à $\mathbf{K}$ et à $\mathbf{E}$ selon $P(\mathbf{m}|\mathbf{K}, \mathbf{E})$;
- les observations restent tirées selon $P(\mathbf{Y}|\mathbf{m}, \boldsymbol{\theta})$.

où les lettres grasses représentent les ensemble des variables sur les profils. Il nous est maintenant possible de régler le prior p_0 de l'événement E_0 selon notre expertise, et nous pouvons ensuite calculer la probabilité à postériori de cet événement au vu des données, de manière exacte. Ces deux cadres fournissent des règles de décision naturelle quant à l'égalité, ou non, des ruptures dans les profils.

Nous en revenons à notre jeu de données de référence dans la partie II-2-4 dans laquelle nous appliquons ces règles de décisions à un sous-ensemble des gènes de la levure. Nous y illustrons le résultat attendu, c'est à dire que les frontières des introns sont invariantes

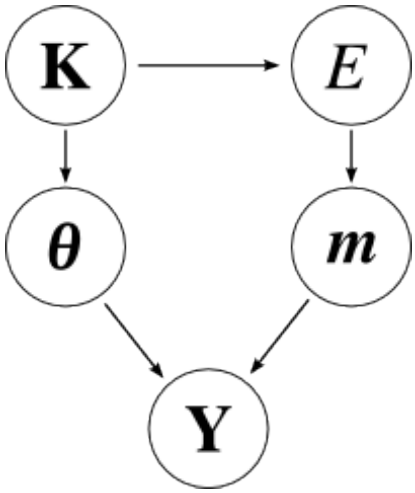

FIGURE 2 – **Modified graphical model for the comparison of change-point location.** Modèle hiérarchique avec couche supplémentaire proposé pour la comparaison de la localisation de ruptures dans des profiles indépendants. E est la variable correspondant à l'évenement 'les ruptures sont identiques'.

à la condition d'élevage, tandis que les début et fin de transcriptions sont eux sujets à modifications selon leur environnement.

Conclusion. Ainsi nous avons développé dans cette thèse plusieurs méthodes de segmentations pour traiter le problème de l'annotation de gènome, et avons illustré ces approches sur le même jeu de données tout au long du manuscrit, afin de mettre en évidence leur richesse pour l'étude de phénomène biologique comme l'épissage alternatif. Le Tableau 1 synthétise la majeure partie de nos travaux. Selon la nature des données et les questions biologiques posées, trois algorithmes sont proposés pour déterminer les localisations de ruptures et permettre l'étude de la qualité de la segmentation. De plus, tous trois répondent aux trois aspects suivants :

- ils sont adaptées à la modélisation de données de comptage, en particulier à l'aide de la binomiale négative, mais pouvant toutefois s'étendre à un grand nombre d'autres distributions,
- ils résolvent de manière exacte les critères qu'ils cherchent à optimiser, et

- ils sont implémentées et diffusées dans des packages R et disponibles librement au public.

Contexte biologique et *exemples*	valeurs max et complexité		Algorithmique (*iii*) `package`	Inférence (*ii*)	incertitude
Génomes *e.g. gènes exprimés* *e.g. nouveaux transcrits*	$n : 10^8$ $K : 10^3$	$\mathcal{O}(Kn)$	Programmation Dynamique élaguée `Segmentor3IsBack`	segmentation optimale inégalité oracle	qualitative
	$n : 10^5$ $K : 10^2$	$\mathcal{O}(K^2n)$	HMM contraint `postCP`	ICL	conditionnelle
Gènes *e.g. annotation précise* *e.g. comparaison de profiles*	$n : 10^4$ $K : 10^1$	$\mathcal{O}(Kn^2)$	Segmentation Bayésienne Exacte `EBS`	segmentation optimale ICL règles de décision	exacte

TABLE 1 – **Aperçu global de la contribution de cette thèse.** Nos travaux sont organisés autour de l'échelle biologique des applications. Pour chacun d'entre eux, nous rappelons leur complexité, les valeurs maximales possibles des paramètres n et K, et des exemples d'informations fournies.

Bibliographie

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory Related Fields*, 113(3) :301–413, 1999. ISSN 0178-8051.

R. Bellman. On the approximation of curves by line segments using dynamic programming. *Commun. ACM*, 4(6) :284, 1961. doi : 10.1145/366573.366611. URL `http://portal.acm.org/citation.cfm?id=366611`.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7) :719–725, 2000. ISSN 01628828.

L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3) :203–268, 2001. ISSN 1435-9855.

L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probability Theory Related Fields*, 138(1-2) :33–73, 2007. ISSN 0178-8051.

James Bullard, Elizabeth Purdom, Kasper Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11(1) :94, 2010.

David L. Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3) :613–627, 1995.

Peter Hall, JW Kay, and DM TITTERINTON. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3) :521–528, 1990.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346) :383–393, 1974.

Rebecca Killick, Paul Fearnhead, and IA Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500) : 1590–1598, 2012.

The Minh Luong, Yves Rozenholc, and Gregory Nuel. Fast estimation of posterior probabilities in change-point analysis through a constrained hidden Markov model. *Computational Statistics & Data Analysis*, 2013.

Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley-Interscience, 2004.

G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *Arxiv :1004.0887*, April 2010. URL `http://arxiv.org/abs/1004.0887`.

G Rigaill, E Lebarbier, and S Robin. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, 22(4) : 917–929, 2012.

Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12(1) :480, 2011.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. `edgeR` : a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26 (1) :139–140, 2010.