Université de Montréal

GÉNÉRALISATIONS DU MODÈLE DE NADEAU ET TAYLOR SUR LES SEGMENTS CHROMOSOMIQUES CONSERVÉS

par

Isabelle Marchand

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en Statistique

Septembre 1997

GÉNÉRALISATIONS DU MODÈLE DE NADEAU ET TAYLOR SUR LES SEGMENTS CHROMOSOMIQUES CONSERVÉS

par

Isabelle Marchand

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des arts et des sciences
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en Statistique

Septembre 1997

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

GÉNÉRALISATIONS DU MODÈLE DE NADEAU ET TAYLOR SUR LES SEGMENTS CHROMOSOMIQUES CONSERVÉS

présenté par

Isabelle Marchand

a été évalué par un jury composé des personnes suivantes :

Sabin Lessard
(président-rapporteur)

David Sankoff
(directeur de recherche)

Martin Goldstein
(membre du jury)

Mémoire accepté le :

19.01.1998

SOMMAIRE

Dans cette étude, on présente tout d'abord le modèle de Nadeau et Taylor [1] élaboré pour estimer la longueur moyenne d'un segment conservé. En ayant deux espèces apparentées, on remarque que les mêmes gènes ne sont pas situés sur les mêmes chromosomes dans les deux espèces. Cependant, on repère plusieurs "segments conservés" à l'intérieur desquels on trouve les mêmes gènes dans les deux espèces. Avec le peu de données disponibles en 1984 provenant des cartes génétiques des hommes et des souris, Nadeau et Taylor obtiennent d'excellents résultats. Toutefois, leur méthode d'estimation repose sur une grande quantité d'approximations, on présente ici une critique de cette méthode.

On présente ensuite le modèle de base de Sankoff et Nadeau [2] sur la distribution du nombre de gènes sur un segment conservé. Ce modèle est basé sur l'hypothèse de distribution homogène des points de rupture et des gènes à travers le génome. On trouve une différence entre les prévisions de ce modèle de base et les données observées chez l'homme et les souris (données de juin 1996).

Afin d'améliorer l'ajustement du modèle, on introduit tout d'abord un paramètre de distribution des gènes en grappes dans le modèle. Ceci dans le but de vérifier l'hypothèse que la découverte de certains gènes est due à l'identification préalable d'autres gènes situés à proximité au lieu d'être fait complètement au hasard. Il s'avère que cette hypothèse n'améliore pas le modèle de base.

Par la suite, puisqu'il est possible que les segments qui contiennent peu de gènes soit plus sensibles aux erreurs d'identification, on modifie le modèle de base de distribution du nombre de gènes sur un segment conservé pour ne pas tenir compte des segments conservés qui contiennent un ou deux gènes. On note alors une amélioration du modèle de base.

Finalement, en divisant le génome en deux et en ajoutant des paramètres de concentration des points de cassure et des gènes, on vérifie que l'on obtient un meilleur ajustement des données observées en supposant que la moitié des échanges entre les chromosomes ont lieu de manière aléatoire et indépendante dans chacune des deux parties du génome où sont distribués 95% des gènes dans une partie et le reste dans l'autre. On améliore encore l'ajustement en distribuant la moitié des gènes dans chaque partie du génome et en concentrant 95% des points de cassure dans une section du génome.

REMERCIEMENTS

Je tiens à remercier M.David Sankoff, mon directeur, pour l'aide précieuse qu'il m'a apportée et pour sa générosité.

Je voudrais aussi remercier ma famille et mes amis, spécialement Martin.

TABLE DES MATIÈRES

SOMMAIRE ii

REMERCIEMENTS iv

LISTE DES FIGURES ix

INTRODUCTION 1

CHAPITRE 1. ESTIMATION DE PARAMÈTRES CONCERNANT LES SEGMENTS CONSERVÉS 4

1.1. La longueur d'un segment conservé 5

1.2. Estimation de la longueur d'un segment conservé 7

1.2.1. Correction du biais dû à l'utilisation de la distance entre les deux gènes connus les plus distants sur un segment 7

1.2.2. Correction du biais créé par l'exclusion des segments ne contenant qu'un seul gène connu 9

1.2.3. Taux d'évolution chromosomique 12

1.2.4. Critique de la méthode d'estimation de Nadeau et Taylor 13

1.3. Le nombre de segments conservés 15

1.4. Calcul du nombre de segments conservés 17

1.4.1. La distribution des gènes sur un segment 18

1.4.2. Estimation du nombre de segment ne contenant aucun gène encore identifié 20

CHAPITRE 2. DISTRIBUTION DU NOMBRE DE GÈNES SUR UN SEGMENT CONSERVÉ 22

2.1. Modèle de base 22

2.1.1. Formulation du modèle de base 22

2.1.2. Illustration du modèle de base 26

2.1.3. Commentaires 29

CHAPITRE 3. MODÈLE DE DISTRIBUTION DES GÈNES PAR GRAPPES 31

3.1. Modèle en grappes 31

3.1.1. Formulation du modèle en grappes 31

3.1.2. Illustration du modèle en grappes 35

3.1.3. Commentaires 40

3.2. Modèle sans les segments contenant 1 seul gène 40

3.2.1. Formulation du modèle sans les segments contenant 1 seul gène 40

3.2.2. Illustration du modèle sans les segments contenant 1 seul gène 43

3.2.3. Commentaires 45

3.3. Modèle sans les segments contenant 1 ou 2 gènes 45

3.3.1. Formulation du modèle de base sans les segments contenant 1 ou 2 gènes 45

3.3.2. Illustration du modèle de base sans les segments contenant 1 ou 2 gènes 47

3.3.3. Commentaires 49

CHAPITRE 4. MODÈLE DE DISTRIBUTION DU NOMBRE DE GÈNES SUR UN SEGMENT CONSERVÉ AVEC PARAMÈTRES DE CONCENTRATION DES GÈNES ET DES POINTS DE RUPTURE 50

4.1. Modèle où $\alpha=\frac{1}{2}$ et β varie 51

4.1.1. Formulation du modèle où $\alpha=\frac{1}{2}$ et β varie 51

4.1.2. Illustration du modèle où $\alpha = \frac{1}{2}$ et β varie 56

4.1.3. Commentaires 61

4.2. MODÈLE OÙ $\beta = \frac{1}{2}$ ET α VARIE 61

4.2.1. Formulation du modèle où $\beta = \frac{1}{2}$ et α varie 61

4.2.2. Illustration du modèle où α varie et $\beta = \frac{1}{2}$ 65

4.2.3. Commentaires 70

CONCLUSION **71**

APPENDICE A. LEXIQUE [7] **73**

RÉFÉRENCES **75**

LISTE DES FIGURES

1.1 Identification d'un segment conservé 5

1.2 Représentation d'un segment conservé 6

1.3 Représentation d'un phénomène d'échange entre 2 chromosomes pour une espèce 17

2.4 Illustrations du modèle de la distribution du nombre de gènes sur un segment 28

3.5 Illustrations du modèle de distribution des gènes par grappes. Cas où il y a 1423 gènes et 113 segments. 36

3.6 Illustrations du modèle de distribution des gènes par grappes. Cas où il y a 1423 gènes et 197 segments. 37

3.7 Illustrations du modèle de distribution des gènes par grappes. Cas où il y a 1423 gènes et 236 segments. 38

3.8 Illustrations du modèle de distribution des gènes par grappes. Cas où il y a 1423 gènes et 284 segments. 39

3.9 Illustrations du modèle de la distribution des gènes sur un segment sans ceux contenant un seul gène. 44

3.10 Illustrations du modèle de la distribution des gènes sur un segment sans ceux contenant 1 ou 2 gènes. 48

4.11 Représentation du modèle 51

4.12 Illustrations du modèle de distribution des gènes lorsque $\alpha = \frac{1}{2}$ et β varie. Cas où il y a 1423 gènes et 113 segments 57

4.13 Illustrations du modèle de distribution des gènes lorsque $\alpha = \frac{1}{2}$ et β varie. Cas où il y a 1423 gènes et 197 segments 58

4.14 Illustrations du modèle de distribution des gènes lorsque $\alpha = \frac{1}{2}$ et β varie. Cas où il y a 1423 gènes et 236 segments 59

4.15 Illustrations du modèle de distribution des gènes lorsque $\alpha = \frac{1}{2}$ et β varie. Cas où il y a 1423 gènes et 284 segments 60

4.16 Illustrations du modèle de distribution des gènes lorsque $\beta = \frac{1}{2}$ et α varie. Cas où il y a 1423 gènes et 113 segments 66

4.17 Illustrations du modèle de distribution des gènes lorsque $\beta = \frac{1}{2}$ et α varie. Cas où il y a 1423 gènes et 197 segments 67

4.18 Illustrations du modèle de distribution des gènes lorsque $\beta = \frac{1}{2}$ et α varie. Cas où il y a 1423 gènes et 236 segments 68

4.19 Illustrations du modèle de distribution des gènes lorsque $\beta = \frac{1}{2}$ et α varie. Cas où il y a 1423 gènes et 284 segments 69

INTRODUCTION

L'évolution peut avoir pour effet l'établissement de nouvelles espèces à partir d'espèces pré-existantes. Parmi les facteurs responsables de l'évolution, on retrouve le réarrangement génomique. Certaines aberrations chromosomiques, normalement léthales, consistent en la variation de l'arrangement des gènes sur les chromosomes. Il s'agit de l'inversion et la transposition des segments (à l'intérieur même du chromosome), la translocation des segments (entre chromosomes) et la fission d'un chromosome ou la fusion de deux chromosomes. L'ordre des gènes sur les segments est toujours inchangé entre deux points de rupture. Des fois, très rarement, ces réarrangements offrent un avantage sélectif et peuvent donc se répandre dans une population. C'est pourquoi, au cours du processus d'évolution, en comparant les génomes de deux espèces apparentées, on retrouve des segments de plus en plus courts et de plus en plus nombreux. Le nombre de segments conservés entre deux espèces traduit le nombre de réarrangements et peut être utilisé pour déduire la distance génomique.

Bien que pour plusieurs génomes bien étudiés l'emplacement de nombreux gènes ait été fixé, quelques fois précisément, ceci ne représente qu'une petite proportion de tous les gènes, même chez l'humain. Ce manque de détails concernant les cartes génétiques implique l'impossibilité de définir exactement les segments conservés dans deux espèces. On doit donc recourir à des estimations. En 1984, Nadeau et Taylor [1] ont formulé un premier modèle qui estimait la longueur moyenne d'un segment conservé. En connaissant la longueur totale du génome,

il était possible de déduire différentes informations concernant les segments conservés, comme le nombre total de segments conservés ou le taux d'évolution chromosomique. Par la suite, en 1996, Sankoff et Nadeau [2] ont analysé ce modèle sans la nécessité de considérer la longueur du segment. Ils ont pu estimer directement le nombre de segments conservés et ils ont pu en extraire d'autres résultats, par exemple la distribution du nombre de gènes sur un segment chromosomique conservé.

Le modèle de base de la distribution du nombre de gènes sur un segment conservé repose, entre autres, sur l'hypothèse de la distribution aléatoire et indépendante des gènes identifiés et des points de rupture à travers tout le génome.

Dans le premier chapitre de ce mémoire, on présente le modèle de Nadeau et Taylor [1] pour estimer la longueur d'un segment conservé et on fait une critique de la méthode utilisée. Ensuite, on présente l'analyse de Sankoff et Nadeau [2] et le modèle de base de distribution du nombre de gènes sur un segment conservé.

En travaillant à partir des gènes connus, communs à l'homme et à la souris (données de juin 1996) on note une différence entre les prévisions du modèle de base et les données observées. Dans le troisième chapitre, on introduit donc dans le modèle de base un paramètre de distribution des gènes par grappes. Il est possible de croire que l'identification de certains gènes n'est pas due complètement au hasard. Certains gènes ont peut-être été découverts car ils étaient voisins d'autres gènes qui eux ont déjà été identifiés. Le modèle ne supposera plus une distribution des gènes uniforme et indépendante les uns des autres à travers tout le génome mais plutôt une distribution aléatoire de groupes de gènes. De plus, comme les segments qui ne contiennent qu'un, ou même deux, gènes sont plus susceptibles de souffrir d'erreurs d'identification, on présente le modèle de base corrigé premièrement pour ne pas tenir compte des segments qui contiennent un

seul gène et deuxièmement, pour ne pas tenir compte de ceux qui contiennent un ou deux gènes.

Les données observées montrent qu'il y a peu de segments conservés qui contiennent beaucoup de gènes identifiés et beaucoup de segments conservés qui en contiennent peu. Dans le quatrième chapitre, on représente le génome en deux parties et on introduit dans le modèle un paramètre de concentration de distribution des points de cassure, α, et un paramètre de concentration de distribution des gènes, β. En fixant l'un et en faisant varier l'autre, cette représentation permet de retrouver dans une partie du génome peu de gènes et beaucoup de segments et dans l'autre partie, peu de segments et beaucoup de gènes.

CHAPITRE 1

ESTIMATION DE PARAMÈTRES CONCERNANT LES SEGMENTS CONSERVÉS

Pour étudier l'évolution et l'organisation génomique chez les mammifères, on a recourt aux cartes génétiques qui représentent la localisation des gènes sur les chromosomes et donc leur distance relative. Le développement de ces cartes pour diverses espèces a été possible grâce à la science de la génétique des mammifères.

On a observé qu'au cours du processus d'évolution il y a plusieurs phénomènes qui modifient l'ordre des gènes sur les chromosomes (inversion, translocation réciproque, transposition, fission, fusion). Certains de ces bouleversements sont causés par des échanges de fragments entre les chromosomes. Malgré ces changements, on a remarqué que les gènes liés étroitement (i.e. qui sont voisins) chez une espèce le sont souvent chez une autre espèce, formant ainsi des segments de chromosomes conservés. On a de plus noté que la sélection naturelle fait en sorte qu'il n'y ait pas (ou très peu) de réarrangements entre les autosomes (chromosomes ayant la même apparence chez les deux sexes) et le chromosome X, les gènes sur ce dernier étant liés très fortement entre eux.

À partir de ces cartes, on peut donc faire des analyses concernant la conservation des liaisons génétiques entre les gènes chez différents mammifères. Par exemple, calculer la probabilité que tels gènes soient liés chez l'homme étant

donné qu'ils sont à une certaine distance chez la souris ou estimer le nombre de réarrangements chromosomiques qui ont eu lieu entre deux espèces au cours de la divergence depuis leur ancêtre commun. On a alors besoin d'un paramètre important: la longueur des segments conservés.

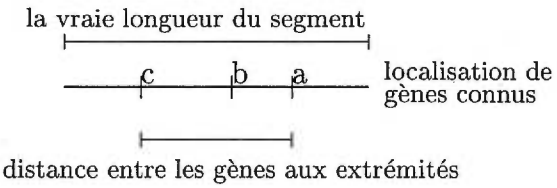
1.1. La longueur d'un segment conservé

Dans l'étude de Nadeau et Taylor [1] un segment conservé est défini de façon opératoire comme une région d'un chromosome d'une souris contenant deux gènes connus ou plus et où ces mêmes deux gènes (ou plus) sont reconnus pour appartenir à un seul chromosome humain, la carte humaine étant inconnue. En plus, aucun autre gène ne doit intervenir dans le segment, i.e. être sur le segment conservé de la souris et appartenir à un autre chromosome chez l'humain différent de celui qui contenait initialement les deux gènes connus ou plus.

Figure 1.1. Identification d'un segment conservé

Les segments ne contenant qu'un seul gène sont exclus; on ne savait pas à ce moment là si le réarrangement était fait par segment ou par mouvement des gènes individuels. L'ensemble de segments ainsi formé sera donc biaisé envers les longs segments, biais dont il faudra tenir compte. La longueur de chaque segment sera mesurée à partir de la distance entre les deux gènes connus les plus distants l'un de l'autre. Puisque la vraie longueur du segment devrait être variablement plus grande, on devra aussi prendre ce fait en considération.

FIGURE 1.2. Représentation d'un segment conservé



Les données recueillies pour estimer le paramètre proviennent des autosomes des hommes et des souris. En se basant sur les connaissances génétiques de 1984, ils ont identifié 36 segments conservés. Mais la distance entre les deux gènes extrèmes n'était pas connue pour 23 de ces segments. La fréquence de recombinaison, qui permet de déduire cette distance est connue seulement pour 13 des segments conservés chez la souris. Ces 13 segments totalisent 31 gènes et sont les seules données qui entrent dans leur analyse. Aujourd'hui on dispose d'une centaine de segments qui comptent presque 2000 gènes.

Nadeau et Taylor ont fait plusieurs hypothèses avant de pouvoir estimer la longueur d'un segment. Tout d'abord, ils supposent qu'un segment conservé contenant deux gènes identifiés ou plus se retrouvant chez les deux espèces est le résultat d'une vraie liaison génétique. Cette hypothèse s'appuie sur le fait que la longueur des segments est relativement petite, que des arrangements de liaisons génétiques similaires se retrouvent parfois chez plusieurs autres mammifères et que la probabilité a priori que deux gènes aient été séparé chez leur ancêtre commun et qu'ils se soient liés par la suite chez les deux espèces est très faible.

Ils supposent aussi que la distribution des réarrangements des autosomes à l'intérieur du génome durant le phénomène d'évolution est aléatoire; on ne retrouve pas de grands segments de chromosome à l'abris de ruptures éventuelles qui permettraient de douter de cette hypothèse. La dernière hypothèse concerne la distribution des gènes homologues identifiés expérimentalement chez l'homme et la souris; elle est supposée aléatoire et indépendante à travers tout le génome.

1.2. Estimation de la longueur d'un segment conservé

Il est maintenant possible d'estimer la longueur d'un segment conservé en utilisant les dernières hypothèses. Nadeau et Taylor ont fait les calculs qui suivent à partir de l'échantillon formé des 13 segments conservés dont la distance génétique entre les deux gènes les plus distants est connue. L'échantillon contient en tout 31 gènes. Ils ont apporté des corrections, tout d'abord concernant le fait que la vraie longueur du segment excède la distance entre les deux gènes extrèmes et ensuite parce que l'échantillon ne tient pas compte des segments qui contiennent un seul gène.

1.2.1. Correction du biais dû à l'utilisation de la distance entre les deux gènes connus les plus distants sur un segment

On a les données pour 13 segments appartenant à la souris. Pour chacun de ces segments, on connait la distance, soit T, entre les deux gènes les plus distants, on connait le nombre de gènes, soit g, sur ce segment et on suppose la longueur de ce dernier plus grande que T, soit s. L'indépendance de la distribution des gènes est parmi les hypothèses. Ceci revient au problème suivant:

Proposition 1.2.1. *Soit g points distribués uniformément et indépendamment sur un intervalle $[0, s]$ et soit T une variable aléatoire représentant la distance entre $x_{(1)}$ et $x_{(g)}$, où $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(g)}$ sont les statistiques d'ordre associées à $x_1, x_2, ..., x_g$. Alors T a la fonction de densité suivante:*

$$f_T(t) = \frac{gt^{g-1} + sg(g-1)t^{g-2} - g^2t^{g-1}}{s^g} \qquad 0 \leq t \leq s$$

et l'espérance de t est donnée par:

$$E(T) = \frac{s(g-1)}{g+1} \ .$$

Démonstration

Soit $F_T(t)$ la fonction de répartition de la variable T, on a

$$\begin{aligned} F_T(t) &= P(x_{(g)} - x_{(1)} \leq t) \\ &= \int\int_{x_g - x_1 \leq t} f_{x_{(1)},x_{(g)}}(x_1, x_g)\, dx_1 dx_g \end{aligned}$$

où $f_{x_{(1)},x_{(g)}}(x_1, x_g)$ est la densité conjointe de deux statistiques d'ordre, $x_{(1)}$ et $x_{(g)}$. On note aussi f et F respectivement la densité et le fonction de répartition des variables x_1 et x_g. Donc,

$$F_T(t) = \int\int_{x_g - x_1 \leq t} \frac{g!}{(g-2)!}[F(x_g) - F(x_1)]^{g-2} f(x_1) f(x_g)\, dx_1 dx_g$$

(voir S.Ross [4], p.221)

$$\begin{aligned} F_T(t) &= \frac{g!}{(g-2)!s^2}\Big[\int_0^{s-t}\int_{x_1}^{x_1+t}\Big[\frac{x_g}{s} - \frac{x_1}{s}\Big]^{g-2} dx_g dx_1 \\ &\quad + \int_{s-t}^{s}\int_{x_1}^{s}\Big[\frac{x_g}{s} - \frac{x_1}{s}\Big]^{g-2} dx_g dx_1\Big] \\ &= \frac{1}{s^g}(t^g + sgt^{g-1} - gt^g) \quad , \quad 0 \leq t \leq s \quad . \end{aligned}$$

La fonction de densité est

$$\begin{aligned} f_T(t) &= \frac{dF_T(t)}{dt} \\ &= \frac{1}{s^g}(gt^{g-1} + sg(g-1)t^{g-2} - g^2t^{g-1}) \quad , \quad 0 \leq t \leq s \quad . \end{aligned}$$

L'espérance est

$$\begin{aligned} E(T) &= \int_0^s t f_T(t)\, dt \\ &= \frac{1}{s^g}\int_0^s (gt^g + sg(g-1)t^{g-1} - g^2t^g)\, dt \\ &= \frac{s(g-1)}{g+1} \quad . \end{aligned}$$

Puisque l'on veut obtenir une estimation pour s et que

$$E(T) = \frac{s(g-1)}{g+1}$$

en posant

$$\hat{s} = \frac{T(g+1)}{g-1} \tag{1.2.1}$$

on a

$$E(\hat{s}) = E(T)\frac{g+1}{g-1} = s$$

$\hat{s}$ est donc un estimateur sans biais de s.

Pour chacun des 13 segments de l'échantillon où t et g sont connus on a calculé $\hat{s}$, la longueur transformée. La moyenne de $\hat{s}$ était 20.9 cM, où les cM (centiMorgan) est l'unité standard de distance en génétique.

1.2.2. Correction du biais créé par l'exclusion des segments ne contenant qu'un seul gène connu

L'échantillon ne tient compte seulement que des segments contenant deux gènes connus ou plus, la moyenne des $\hat{s}$ est donc biaisée envers les longs segments, puisque ceux qui ne contiennent qu'un seul gène sont exclus. Nadeau et Taylor ont donc apporté la correction suivante.

Soit x' une variable aléatoire qui représente la longueur d'un segment conservé détecté parce qu'il contient deux gènes connus ou plus. On veut obtenir la relation entre $E(x')$ et L où L est la vraie moyenne de la longueur des segments (incluant ceux ne contenant qu'un seul gène connu). La valeur de $E(x')$ est estimée par la moyenne des $\hat{s}$ calculée à partir de l'équation (1.2.1).

La probabilité qu'un segment de taille x contienne k gènes connus et soit inclus dans l'échantillon est donné par la loi de Poisson de paramètre Dx, où

$$\hat{D} = \frac{T}{G} = \frac{\text{nombre de gènes connus et cartographiés}}{\text{taille du génome (en centiMorgans)}} \quad .$$

Le paramètre représente donc le nombre moyen de gènes connus et cartographiés sur un segment de longueur x. Les événements se réalisent selon un processus de Poisson de paramètre D. Les hypothèses d'un processus de Poisson sont les suivantes: on retrouve seulement un nombre fini de points qui tombent sur un intervalle fini, les nombres d'événements survenant au cours d'intervalles disjoints sont des variables indépendantes et la distribution du nombre de points sur un intervalle dépend seulement de la longueur de l'intervalle.

Les segments qui se retrouvent dans l'échantillon sont ceux contenant deux gènes ou plus. La probabilité qu'un segment de taille x contienne deux gènes ou plus et qu'il soit un élément de l'échantillon est donnée par

$$\begin{aligned}\sum_{k=2}^{\infty} \frac{(Dx)^k e^{-Dx}}{k!} &= e^{-Dx} \sum_{k=2}^{\infty} \frac{(Dx)^k}{k!} \\ &= e^{-Dx}(e^{Dx} - 1 - Dx) \\ &= 1 - e^{-Dx} - Dxe^{-Dx} \quad .\end{aligned}$$

La fonction de probabilité d'un segment de taille x à travers le génome est donnée par la loi exponentielle de paramètre $1/L$ qui est l'inverse de la valeur espérée de la variable x

$$f(x) = \frac{1}{L} e^{-x/L} \quad , \quad 0 \leq x \leq \infty \quad .$$

La fréquence relative d'un segment de taille x se retrouvant dans l'échantillon est donc donnée par

$$S(x) = (1 - e^{-Dx} - Dxe^{-Dx}) \frac{e^{-x/L}}{L} \quad , \quad 0 \leq x \leq \infty \quad .$$

Un segment de taille x ne peut être plus grand que la longueur d'un chromosome, soit c. Nadeau et Taylor ont donc normalisé la densité et obtenu la fonction de distribution suivante pour x'

$$f(x') = \frac{S(x')}{\int_0^c S(x)\, dx} \quad , \quad 0 \leq x' \leq c \quad .$$

On peut maintenant calculer $E(x')$

$$\begin{aligned} E(x') &= \int_0^c x' f(x')\, dx' \\ &= \int_0^c x' \frac{S(x')}{\int_0^c S(x)\, dx}\, dx' \\ &= \frac{1}{\int_0^c S(x)\, dx} \int_0^c x' S(x')\, dx' \quad , \end{aligned}$$

où

$$\begin{aligned} \int_0^c S(x)\, dx &= \int_0^c \frac{e^{\frac{-x}{L}}}{L}(1 - e^{-Dx} - Dxe^{-Dx})\, dx \\ &= 1 - e^{\frac{-c}{L}} + \frac{e^{-c(D+\frac{1}{L})}}{LD+1} - \frac{1}{LD+1} + \frac{Dce^{-c(D+\frac{1}{L})}}{LD+1} \\ &\quad + \frac{De^{-c(D+\frac{1}{L})}}{L(D+\frac{1}{L})^2} - \frac{D}{L(D+\frac{1}{L})^2} \end{aligned} \tag{1.2.2}$$

et

$$\begin{aligned} \int_0^c x' S(x')\, dx' &= \int_0^c \frac{x' e^{\frac{-x'}{L}}}{L}(1 - e^{-Dx'} - Dx' e^{-Dx'})\, dx' \\ &= -e^{\frac{-c}{L}}(L+c) + L + \frac{e^{-c(D+\frac{1}{L})}}{LD+1}(C+1) - \frac{1}{L(D+\frac{1}{L})^2} \\ &\quad + \frac{c^2 De^{-c(D+\frac{1}{L})}}{LD+1} + \frac{2Dce^{-c(D+\frac{1}{L})}}{L(D+\frac{1}{L})^2} + \frac{2De^{-c(D+\frac{1}{L})}}{L(D+\frac{1}{L})^3} \\ &\quad - \frac{2D}{L(D+\frac{1}{L})^3} \quad . \end{aligned} \tag{1.2.3}$$

Nadeau et Taylor ont simplifé la valeur de ces deux intégrales en faisant l'approximation suivante: supposons c, la taille du chromosome considérablement plus grande que la vraie longueur d'un segment conservé, L, i.e. que la valeur de $\frac{c}{L}$ tend vers l'infini. Par conséquent, tous les termes de (1.2.2) et (1.2.3) contenant

$e^{\frac{-c}{L}}$ tendent vers 0. On retrouve alors les deux expressions suivantes:

$$\int_0^c S(x)\,dx \approx 1 - \frac{1}{LD+1} - \frac{D}{L(D+\frac{1}{L})^2}$$

et

$$\int_0^c x' S(x')\,dx' \approx L - \frac{1}{L(D+\frac{1}{L})^2} - \frac{2D}{L(D+\frac{1}{L})^3} \quad .$$

On obtient finalement l'approximation de la moyenne de la longueur d'un segment conservé détecté parce qu'il contient deux gènes ou plus en fonction de la vraie moyenne de la longueur des segments, L.

$$E(x') \approx \frac{L^2D+3L}{LD+1} \quad .$$

À partir de leur échantillon de 13 segments, Nadeau et Taylor avait trouvé une valeur de 20.9 cM pour la moyenne des longueurs transformées. Puisque $\hat{D} = 0.0338$, en remplaçant dans l'équation et en trouvant la solution d'une quadratique:

$$L^2D + (3 - DE(x'))L - E(x') \approx 0 \quad .$$

On obtient une valeur de L, la vraie valeur de la moyenne d'un segment conservé, égale à 8.1 cM.

1.2.3. Taux d'évolution chromosomique

Estimer la longueur d'un segment conservé permet, entre autres, de calculer le taux d'évolution chromosomique en utilisant le paramètre du nombre total de segments conservés qui est déduit par l'équation suivante:

$$\begin{aligned} \text{nombre total de segment conservé} &= \frac{G}{L} \\ &= \frac{\text{longueur totale du génome}}{\text{longueur moyenne d'un segment conservé}} \quad . \end{aligned}$$

Le nombre de segments conservés augmente proportionnellement avec le nombre de ruptures. Nadeau et Taylor ont posé que le nombre de segments conservés augmente de 1 pour chaque rupture, on obtient alors la relation suivante:

$$\begin{aligned} \text{nombre total de segments conservés} &= \text{nombre d'autosomes présents chez} \\ &\quad \text{l'ancêtre commun de l'homme et la souris} \\ &\quad + \text{nombre de ruptures qui ont eut lieu} \\ &\quad \text{durant la divergence des deux espèces} \\ \frac{G}{L} &= N_o + R \ . \end{aligned}$$

Le nombre d'autosomes de l'ancêtre commun de l'homme et la souris est évalué à 20. Puisqu'on a estimé la longueur d'un segment conservé à 8.1 cM, on peut donc calculer le nombre total de segments conservés. Puisque la longueur du génome est connue, elle est a peu près égale à 1 600 cM, on obtient ainsi environ 178 ruptures chez l'homme et la souris au cours de leur divergence depuis leur ancêtre commun.

Pour calculer le taux d'évolution chromosomique, i.e. le taux d'échanges entre chromosomes au cours du processus d'évolution qui a conduit à la divergence des deux espèces, on doit diviser le nombre de ruptures par 2, car il totalise celles des deux espèces, et par le nombre d'années qu'a duré le processus, soit 70 millions d'années. On trouve alors 1.3 ruptures par million d'années.

1.2.4. Critique de la méthode d'estimation de Nadeau et Taylor

Les techniques pour récolter les données dans le domaine de la génétique ont grandement évolué depuis 1984 et, par conséquent, le nombre de gènes connus et cartographiés a augmenté considérablement. Ce qu'il y a de génial avec la

méthode d'estimation de Nadeau et Taylor c'est qu'elle a permis, à l'aide de seulement 13 segments et 31 gènes, d'obtenir des résultats comparables à ceux que l'on obtient aujourd'hui avec une centaine de segments et presque 2000 gènes. Malgé tout, cette méthode pose toutefois quelques problèmes et les résultats obtenus reposent sur une grande quantité d'approximation.

Tout d'abord, on a estimé $\hat{s}$, la longueur transformée d'un segment conservé à partir de la distance génétique entre les deux gènes aux extrémités du segment. On veut obtenir une relation entre $E(x')$, la moyenne d'un segment conservé détecté parce qu'il contient deux gènes ou plus, et L, la vraie longueur moyenne du segment. Or, la valeur de $E(x')$ est estimée par la moyenne de $\hat{s}$. La valeur de $\hat{s}$ est en fonction de g le nombre de gènes sur le segment et les segments qui contiennent plus de gènes sont estimés plus précisément. Or, en utilisant la moyenne de $\hat{s}$ on a donné un même poids aux segments qui contiennent 2 gènes à ceux qui en contiennent 3 ou 4.

Ensuite, la fréquence relative d'un segment de taille x se retrouvant dans l'échantillon est donnée par $S(x)$ pour $0 \leq x \leq \infty$. Puisque la longueur d'un segment conservé ne peut dépasser la longuer d'un chromosome, la fonction de distribution de x' a été bricolée, i.e. qu'elle a été divisée par la valeur de l'intégrale de la fréquence relative évaluée de 0 à c, la taille d'un chromosome, pour que la fonction de densité de $x^{'}$ donne 1. On a alors supposé que tous les chromosomes ont la même longueur ce qui est loin d'être vrai.

Lors du calcul de l'espérance de x', la valeur de l'intégrale est approximée à 0 au point c, i.e. le rapport de la longueur du chromosome, c, et de la longueur d'un segment conservé, L, tend vers ∞. En réalité, $\frac{c}{L}$ prend souvent des valeurs telles que 3 ou 4.

Finalement on trouve une relation approximative entre $E(x')$ et L

$$E(x') \approx f(L)$$

$E(x')$ est donnée par une fonction de $\hat{s}$ où $\hat{s}$ est un estimateur sans biais de s (équation 1.2.1). Or pour

$$E(\hat{s}) = s$$

on a pour une fonction $H(s)$ quelconque, (voir Casella-Berger [6] p.330)

$$E(H(\hat{s})) \approx H(s) \quad .$$

On utilise donc une méthode approximative pour estimer L.

De plus, pour calculer le taux d'évolution chromosomique on a supposé

$$E(\frac{g}{L}) = \frac{g}{E(L)}.$$

Plus de dix ans plus tard, l'approche pour déduire d'autres résultats concernant les segments conservés, par exemple la distance génomique ou la distribution du nombre de gènes par segment, s'est faite non pas en utilisant la longueur des segments, mais plutôt en passant par le nombre de segments conservés.

1.3. Le nombre de segments conservés

Le nombre de segments conservés permet de déduire la distance génomique puisqu'il permet de savoir combien d'échanges il y a eu entre les chromosomes des deux espèces au cours de la divergence depuis leur ancêtre commun. Dans l'étude de Sankoff et Nadeau [2] un segment conservé représente une section de chromosome où les gènes se retrouvent dans le même ordre que sur le chromosome d'une autre espèce.

Sans passer par la longueur du segment, le nombre de segments conservés contenant des gènes homologues peut être estimé en comparant les deux ensembles, un pour chaque espèce, formés des chromosomes et des gènes connus qui

leurs sont associés. Certains de ces gènes sont précisément localisés sur le chromosome. Le nombre de segments conservés est donc défini comme le nombre d'éléments dans l'intersection de ces deux ensembles où chaque élément est un groupe de gènes étant situé sur un même chromosome chez les deux espèces.

On se retrouve ainsi à sous-estimer le nombre de segments conservés. En effet, un ensemble de gènes situés sur un même chromosome chez les deux espèces peut ne pas représenter qu'un seul segment conservé, mais plusieurs segments conservés indépendants. Il y aurait alors eu plusieurs échanges qui auraient restitué les gènes sur un même chromosomes. De plus, il y a des segments conservés qui contiennent des gènes qui n'ont pas encore été identifiés jusqu'à maintenant dans une ou les deux espèces, ce qui a pour conséquence que des segments conservés ne sont pas représentés.

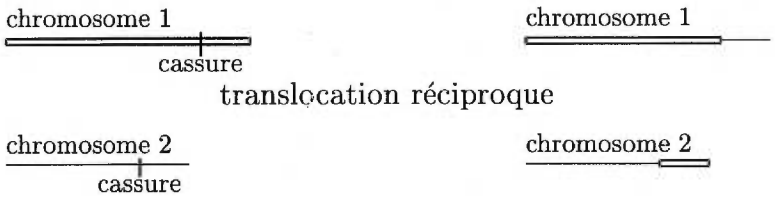
Une autre hypothèse de Sankoff et Nadeau stipule qu'un sous-ensemble de gènes connus situés sur un même chromosome chez les deux espèces correspond à un seul segment conservé. Ainsi on ne tient pas compte des échanges à l'intérieur d'un même chromosome, tel que les phénomènes d'inversion et de transposition. Il est à noter que de cette facon, le nombre maximal de segments conservés correspond au produit du nombre de chromosomes dans chacune des deux espèces, alors qu'en réalité, la borne supérieur est donnée par le nombre total de gènes.

On suppose aussi que les échanges entre les chromosomes ont lieu de manière aléatoire et indépendante et que l'identification des gènes associés à un chromosome ne dépend pas de leur position sur celui-ci. On peut alors dériver la distribution de probabilité du nombre de gènes identifiés par segment conservé, estimer le nombre de segments conservés qui ne contiennent aucun gène identifié jusqu'à maintenant ou estimer le nombre total d'échanges entre chromosomes.

1.4. Calcul du nombre de segments conservés

Soit deux espèces possédant respectivement c_1 et c_2 chromosomes. À la suite d'un phénomène d'échange entre les chromosomes d'une espèce (translocation), le nombre de segments conservés augmente de 2. En supposant qu'il n'y a qu'une seule rupture à chaque point de cassure on obtiendra alors qu'après t échanges il y aura une augmentation de $2t$ segments conservés.

Figure 1.3. Représentation d'un phénomène d'échange entre 2 chromosomes pour une espèce



La différence entre c_1 et c_2 est due à des phénomènes de fusion entre 2 chromosomes ou de fission d'un chromosome. En supposant qu'il n'y a que ϕ fissions qui interviennent entre le plus petit génome et le plus grand. On peut alors exprimer le nombre de segments conservés de la façon suivante

$$s = \min(c_1, c_2) + 2t + \phi \tag{1.4.1}$$

où $\phi > 0$ représente le nombre de fissions et $\phi < 0$ représente le nombre de fusions.

On obtient ainsi une expression pour estimer le nombre total d'échanges entre les chromosomes des 2 espèces

$$t = \frac{s - min(c_1, c_2) - \phi}{2} \ .$$

Lorsqu'on estime s, le nombre de segments conservés, on se trouve à le sous-estimer, entre autres parce qu'il y a des segments qui ne contiennent aucun gène

identifié jusqu'à maintenant dans l'une des 2 espèces. Il faudrait alors estimer ce nombre de segments.

Il est à noter que l'équation 1.4.1 exprime le nombre de segments conservés en utilisant le minimum du nombre de chromosomes entre les deux espèces et le nombre d'échanges qui a eu lieu depuis la divergence de ces deux espèces. Dans l'étude de Nadeau et Taylor, le nombre total de segments conservés est exprimé en fonction du nombre d'autosomes de l'ancêtre commun aux deux espèces et le nombre de ruptures survenues au cours de leur divergence.

1.4.1. La distribution des gènes sur un segment

Le génome, i.e. l'ensemble de tous le gènes situés sur les chromosomes est représenté sur un intervalle [0,1]. Il a été démontré que la distribution de la taille des chromosomes est bien représentée par des points de cassure distribués au hasard à travers le génome. On sépare donc aléatoirement l'intervalle [0,1] en c chromosomes.

La position des gènes à travers le génome est supposée indépendante les unes des autres. Soit m, le nombre total de gènes situés sur les chromsomes. Ces m gènes sont donc représentés par m points distribués uniformément sur l'intervalle [0,1].

On avait supposé que les phénomènes d'échanges avaient lieu de manière aléatoire et étaient indépendants les uns des autres. Les points de cassure sont donc distribués au hasard le long du génome. On ne fait pas de distinction entre les cassures qui séparent les segments de celles qui séparent 2 chromosomes. Soit n points de cassure distribués uniformément sur l'intervalle [0,1]. Ils représentent ainsi $n + 1$ segments de taille aléatoire. Il a été montré que la distribution de la taille des segments est bien représentée par un modèle de cassures uniformes.

Si on prend un segment au hasard, il est possible qu'il ne contienne aucun gène ou qu'il les contienne tous. On veut obtenir la distribution de probabilité du nombre de gènes sur un segment.

Si on a un segment de taille x, la probabilité qu'il contienne r gènes est donnée par la loi binomiale de paramètres (m, x)

$$B(m, x; r) = \binom{m}{r} x^r (1-x)^{m-r} \quad , \quad r = 0, 1, ..., m \quad .$$

Il faut maintenant la fonction de densité d'un segment de taille x.

PROPOSITION 1.4.1. *Soit n points distribués uniformément sur un intervalle $[0, 1]$ et soit x une variable aléatoire représentant la distance entre $x_{(k)}$ et $x_{(k-1)}$, $k = 0, ..., n$ et $x_0 = 0$, où $x_{(k)}$ correspond à la k-ième statistique d'ordre. Alors x à la fonction de densité suivante*

$$f(x) = n(1-x)^{n-1} \quad , \quad 0 \leq x \leq 1 \quad .$$

Démonstration

$$\begin{aligned} F(x) &= P(x_{(k)} - x_{(k-1)} \leq x) \\ &= 1 - (1-x)^n \quad , \quad 0 \leq x \leq 1 \quad . \end{aligned}$$

(Voir S.Ross [4] p.233 ex. 6.7.29).

La fonction de densité de x est

$$\begin{aligned} f(x) &= \frac{dF(x)}{dx} \\ &= n(1-x)^{n-1} \quad , \quad 0 \leq x \leq 1 \quad . \end{aligned}$$

On peut alors calculer la fonction de probabilité du nombre de gènes sur un segment arbitraire

$$\begin{aligned} P(r) &= \int_0^1 f(x)B(m,x;r)\,dx \\ &= n\binom{m}{r}\int_0^1 x^r(1-x)^{n+m-r-1}\,dx \\ &= n\binom{m}{r}\frac{\Gamma(r+1)\Gamma(n+m-r)}{\Gamma(n+m+1)} \\ &= \frac{nm!(n+m-r-1)!}{(n+m)!(m-r)!} \quad , \quad r=0,1,...,m \quad . \end{aligned}$$

1.4.2. Estimation du nombre de segment ne contenant aucun gène encore identifié

Soit E_i une variable aléatoire qui représente le nombre de segments qui contiennent i gènes. Comme il est possible qu'un segment contienne aucun gène ou tous les gènes, i prend des valeurs allant de 0 à m. Soit N_i la valeur observée pour l'événement E_i. La fonction de probabilité de $(N_0, N_1, ..., N_m)$ est donc donnée par la loi multinomiale $\mathcal{M}(n+1, P(.); N.)$ de fonction de masse

$$f(N_0, N_1, ..., N_m) = (n+1)! \prod_{i=0}^{m} \frac{P(i)^{N_i}}{N_i!}$$

où $\sum_{i=0}^{m} N_i = n+1$ et $\sum_{i=0}^{m} P(i) = 1$.

Le nombre de segments ne contenant aucun gène est impossible à observer. Les paramètres N_0 et $n+1$ sont donc à estimer. Bien qu'il n'y ait pas indépendance entre le nombre de gènes dans les différents segments, on suppose une dépendance très faible à mesure que n augmente. Pour des valeurs de n relativement grande on suppose que cet effet à une probabilité pratiquement nulle. On peut donc estimer les paramètres manquants en passant par la méthode du

maximum de vraisemblance. On veut donc maximiser la fonction de masse f en $n+1$:

$$L(n+1) = \frac{(n+1)!}{N_1!...N_m!(n+1-\sum_{i=1}^{m} N_i)} P(0)^{n+1-\sum_{i=1}^{m} N_i} \prod_{i=1}^{m} P(i)^{N_i}$$

où les valeurs de $N_1, ..., N_m$ sont observées, celles des P(i) sont données par la distribution des gènes sur un segment arbitraire, $m = \sum_{i=1}^{m} iN_i$ et $N_0 = n+1-\sum_{i=1}^{m} N_i$.

CHAPITRE 2

DISTRIBUTION DU NOMBRE DE GÈNES SUR UN SEGMENT CONSERVÉ

Dans ce chapitre, on présente le modèle de base de la distribution du nombre de gènes sur un segment conservé. Ce modèle est basé sur l'hypothèse que les échanges entre les chromosomes ont lieu de manière aléatoire et qu'ils sont indépendants les uns des autres. On suppose aussi que les gènes sont distribués au hasard tout le long du génome et que leurs positions sont indépendantes les unes des autres.

2.1. Modèle de base

2.1.1. Formulation du modèle de base

L'ensemble du génome est représenté sur un intervalle [0,1] où sont distribués de façon uniforme m gènes et n points de rupture séparant le génome en $n+1$ segments. La probabilité qu'un segment arbitraire contienne r gènes était donné par l'expression suivante

$$P(r) = \frac{nm!(n+m-r-1)!}{(n+m)!(m-r)!} \quad , \quad r = 0, 1, ..., m \quad . \tag{2.1.1}$$

Les données proviennent du nombre de gènes communs chez deux espèces, soit l'homme et la souris, et de l'estimation du nombre de points de cassure. Les

données recueillies sont le nombre de segments contenant r gènes qui se retrouvent chez les deux espèces. Ces segments représentent des segments conservés chez l'homme et la souris depuis la divergence de leur ancêtre commun. On obtient des valeurs pour les segments qui contiennent un gène ou plus. Il faut alors apporter une correction à l'expression (2.1.1) en conditionnant sur le fait que le nombre de gènes sur le segment est différent de 0.

PROPOSITION 2.1.1. *La fonction de masse du nombre de gènes sur un segment arbitraire étant donné que ce nombre est différent de 0 est*

$$Q(r) = \frac{n(m-1)!(n+m-r-1)!}{(m+n-1)!(m-r)!} \quad , \quad r = 1, ..., m \quad .$$

l'espérance est donnée par

$$E(r) = \frac{m+n}{n+1}$$

et la variance est

$$Var(r) = \frac{(m+n)n(m-1)}{(n+1)^2(n+2)} \quad .$$

La démonstration de cette proposition s'appuie sur les résultats suivants:

PROPOSITION 2.1.2. *Pour a,b,c des entiers $\geq$ 0,*

$$\sum_{a=0}^{c} a\binom{b+c-a-1}{c-a} = \binom{c+b}{c-1}$$

Démonstration On utilise l'identité suivante (voir R.Riordan p.148 [3]):

$$\binom{n+p+q+1}{n} = \sum_{k=0}^{n} \binom{p+k}{k}\binom{q+n-k}{n-k} \tag{2.1.2}$$

$$\begin{aligned}\sum_{a=0}^{c} a\binom{b+c-a-1}{c-a} &= \sum_{a=0}^{c}(a+1)\binom{b+c-a-1}{c-a} - \sum_{a=0}^{c}\binom{b+c-a-1}{c-a} \\ &= \sum_{a=0}^{c}\binom{a+1}{a}\binom{b+c-a-1}{c-a} - \sum_{a=0}^{c}\binom{a}{a}\binom{b+c-a-1}{c-a}\end{aligned}$$

En utilisant l'équation 2.1.2 on trouve

$$\begin{aligned}\sum_{a=0}^{c} a\binom{b+c-a-1}{c-a} &= \binom{c+1+b}{c} - \binom{c+b}{c} \\ &= \binom{c+b}{c-1}\end{aligned}$$

PROPOSITION 2.1.3. *Pour a,b,c des entiers ≥ 0, on a*

$$\sum_{a=0}^{c} a^2\binom{b+c-a-1}{c-a} = \binom{c+b+1}{c-1} + \binom{c+b}{c-2}$$

Démonstration

$$\begin{aligned}\sum_{a=0}^{c} a^2\binom{b+c-a-1}{c-a} &= \sum_{a=0}^{c} a(a+1)\binom{b+c-a-1}{c-a} - \sum_{a=0}^{c} a\binom{b+c-a-1}{c-a} \\ &= 2\sum_{a=0}^{c}\binom{a+1}{a-1}\binom{b+c-a-1}{c-a} - \sum_{a=0}^{c} a\binom{b+c-a-1}{c-a} \\ &= 2\sum_{a=0}^{c}\binom{a+2}{a}\binom{b+c-a-1}{c-a} - 2\sum_{a=0}^{c}\binom{a+1}{a}\binom{b+c-a-1}{c-a} \\ &\quad - \sum_{a=0}^{c} a\binom{b+c-a-1}{c-a}\end{aligned}$$

En utilisant la proposition 2.1.2 et l'identité de Riordan [3] présentée dans la démonstration de la proposition 2.1.2. On obtient donc,

$$\begin{aligned}\sum_{a=0}^{c} a^2\binom{b+c-a-1}{c-a} &= 2\binom{c+b+2}{c} - 2\binom{c+b+1}{c} - \binom{c+b}{c-1} \\ &= 2\binom{c+b+1}{c-1} - \binom{c+b}{c-1}\end{aligned}$$

$$= \binom{c+b+1}{c-1} + \binom{c+b}{c-2}$$

Démonstration de la proposition 2.1.1

$$Q(r) = P(r/r \neq 0) = \frac{P(r)}{1-P(0)}$$

$$= \frac{\frac{nm!(n+m-r-1)!}{(n+m)!(m-r)!}}{1-\frac{n(n+m-1)!}{(n+m)!}}$$

$$= \frac{n(m-1)!(n+m-r-1)!}{(m+n-1)!(m-r)!} \quad , \quad r = 1, ..., m \quad .$$

L'espérance est

$$E(r) = \sum_{r=1}^{m} rQ(r)$$

$$= \sum_{r=1}^{m} \frac{rn(m-1)!(n+m-r-1)!}{(m+n-1)!(m-r)!}$$

$$= \frac{n!(m-1)!}{(m+n-1)!} \sum_{r=0}^{m} r\binom{n+m-r-1}{m-r} \quad .$$

En utilisant la proposition 2.1.2 on trouve donc

$$E(r) = \frac{n!(m-1)!}{(m+n-1)!}\binom{m+n}{m-1}$$

$$= \frac{m+n}{n+1}$$

La variance est

$$Var(r) = E(r^2) - (E(r))^2 \quad .$$

Or,

$$E(r^2) = \sum_{r=1}^{m} r^2 Q(r)$$

$$= \frac{n(m-1)!(n-1)!}{(m+n-1)!}\sum_{r=1}^{m}\frac{r^2(n+m-r-1)!}{(m-r)!(n-1)!}$$

$$= \frac{n!(m-1)!}{(m+n-1)!}\sum_{r=0}^{m} r^2\binom{n+m-r-1}{m-r} .$$

En utilisant la proposition 2.1.3 on trouve donc

$$E(r^2) = = \frac{n!(m-1)!}{(m+n-1)!}\left(\frac{(m+n+1)!}{(m-1)!(n+2)!}+\frac{(m+n)!}{(m-2)!(n+2)!}\right)$$

$$= \frac{(m+n)(2m+n)}{(n+1)(n+2)} .$$

Donc

$$Var(r) = E(r^2)-(E(r))^2$$

$$= \frac{(m+n)(2m+n)}{(n+1)(n+2)}-\left(\frac{m+n}{n+1}\right)^2$$

$$= \frac{(m+n)n(m-1)}{(n+1)^2(n+2)} .$$

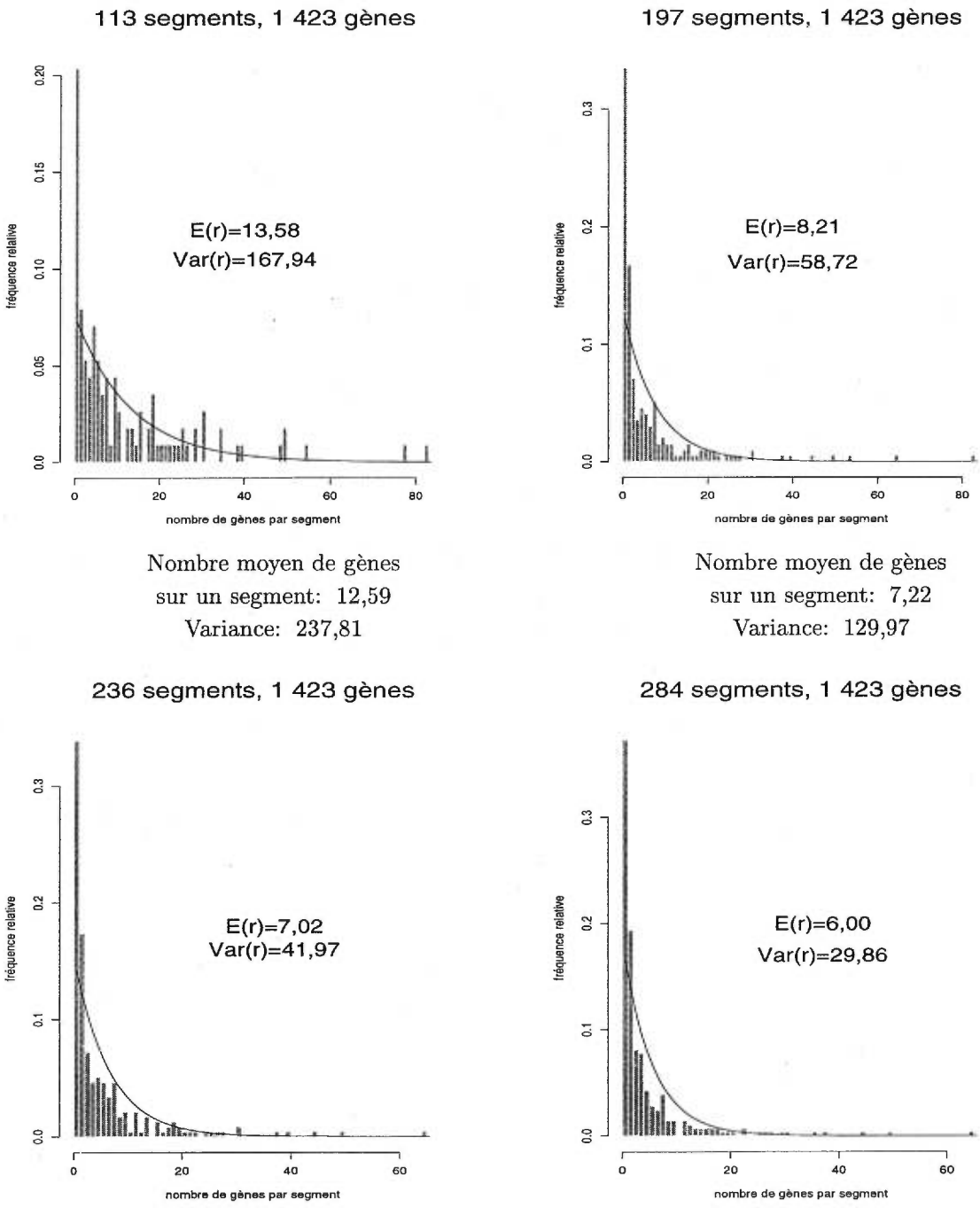
2.1.2. Illustration du modèle de base

Le nombre de gènes connus jusqu'à maintenant (nos données datent de juin 1996) qui est commun à l'homme et la souris est m=1423. Ces 1423 gènes se retrouvent sur un certain nombre de segments conservés. L'identification directe des segments est compliquée par la présence d'erreurs concernant la position des gènes et de leur assignement chromosomique. Une approche algorithmique [5]

a été développée pour contourner cette difficulté et permet l'identification des segments. Cette méthode exige l'établissement d'un paramètre b qui reflète la préférence de l'analyste pour les segments courts mais homogènes versus de longs segments qui ne se conforment pas nécessairement à 100% à des critères strictes d'un segment conservé. Cette méthode permet d'estimer le nombre de segments conservés, $n+1$, et le nombre de segments contenant 1 gène, 2 gènes, etc....

Regardons l'histogramme de la distribution du nombre de segments contenant r gènes et la courbe de probabilité $Q(r)$ de la proposition 2.1.1. Ainsi que la valeur de $E(r)$ et $Var(r)$. Voici des exemples où le nombre de segments non vides conservés est estimé à 113, 197, 236 et 284 (selon la valeur du paramètre b mentionné ci-dessus). Sous chaque histogramme est inscrit le nombre moyen de gènes par segment et la variance du nombre de gènes par segment.

FIGURE 2.4. Illustrations du modèle de la distribution du nombre de gènes sur un segment



Nombre moyen de gènes sur un segment: 12,59
Variance: 237,81

Nombre moyen de gènes sur un segment: 7,22
Variance: 129,97

Nombre moyen de gènes sur un segment: 6,02
Variance: 76,49

Nombre moyen de gènes sur un segment: 5,01
Variance: 62,11

2.1.3. Commentaires

Comme on peut le constater d'après la figure 2.4 (p.28), ce modèle n'ajuste pas tellement bien les données. $Q(r)$ est calculé selon un n trop petit puisqu'on ne peut dénombrer le nombre de segments ne contenant aucun gène. (Dans un travail récent [8] nous avons estimé la valeur de n_0, le nombre de segments qui ne contient aucun gène.) Donc $Q(r)$ sous-estime les petits segments et donc E_Q est trop grand. Par ce modèle, on obtient que la fonction modélise mal la probabilité des segments où r=5,6,...,20 et que sur les figures, à l'exception de celle où il y a 113 segments, la fonction surestime la fréquence relative à partir de r=2.

Dans les exemples, on note que le nombre empirique de segments contenant 1 seul gène est très élevé, ceci peut être lié à des facteurs expérimentaux qui font en sorte que l'assignement des gènes homologues a été mal déterminé ou causé par le fait que certains gènes n'ont pas encore été identifiés.

Si le modèle ne représente pas bien les données, une partie du problème peut provenir des hypothèses. En particulier, celles concernant l'indépendance de la distribution des points de cassure et l'indépendance de la distribution des gènes identifiés. Les points de cassure, causés par des échanges entre les chromosomes, ne sont peut-être pas distribués de manière uniforme à travers le génome. En réalité, il est possible qu'il y ait des sections du génome où le phénomène des ruptures est plus susceptible de se produire. Ceci implique que l'on retrouverait une concentration de points de cassure à certains endroits où il y a en même temps peu de gènes et avoir certaines sections du génome il y a peu de points de rupture et beaucoup de gènes.

Une autre cause possible du problème provient de l'hypothèse selon laquelle la localisation des gènes identifiés est indépendante les unes des autres. En fait,

la découverte de certains gènes peut être due à leur proximité à d'autres gènes qui ont été déjà identifiés.

On peut donc apporter une modification au modèle proposé en rajoutant certains paramètres qui tiendraient compte des observations précédentes. Dans un premier temps, en considérant que la découverte de certains gènes est liée à leur position proche d'autres gènes, on ajoute un paramètre qui ferait en sorte que la probabilité d'observer r gènes sur un segment est en fait la probabilité d'observer k grappes contenant chacune z gènes (où $r \approx kz$).

Dans un deuxième temps, on suppose qu'il y a certaine portion du génome où le nombre de points de rupture est plus concentré et où on retrouve moins de gènes. La distribution des gènes n'est donc plus uniforme à travers le génome par rapport à la distribution des cassures.

CHAPITRE 3

MODÈLE DE DISTRIBUTION DES GÈNES PAR GRAPPES

Dans le présent chapitre, on apporte quelques changements au modèle de base. On ne suppose plus que les positions des gènes à travers le génome sont indépendantes les unes des autres. Certains gènes ont été identifiés parce qu'ils se situaient à proximité d'un ou d'autres gènes qui avaient préalablement été identifiés. On ajoutera donc au modèle un paramètre qui fera que la distribution des gènes sur un segment conservé sera en fait la distribution d'un groupe de gènes.

3.1. Modèle en grappes

3.1.1. Formulation du modèle en grappes

Le génome est toujours représenté par un intervalle [0,1] et on suppose que les phénomènes d'échanges entre les segments sont dus au hasard. On a donc n points de cassure distribués uniformément qui sépare l'intervalle en $n+1$ segments de taille aléatoire. On divise les m gènes à distribuer en m' groupes de gènes, où chaque groupe est formé de z éléments où z est fixé. Ces groupes de gènes sont distribués de façon uniforme à travers le génome. Même si l'hypothèse que toutes les grappes sont de la même taille est forte, on se dit que s'il y a un effet de non-indépendance de positions des gènes, le modèle va le respecter.

On avait calculé dans l'ancien modèle la probabilité d'observer r gènes sur un segment de longueur x. Ceci était donné par une loi binomiale de paramètre (m, x):

$$B(m, x; r) = \binom{m}{r} x^r (1-x)^{m-r} \quad , \quad r = 0, 1, ..., m \quad .$$

Ceci devient maintenant

$$\begin{aligned} P(\text{un segment de taille } x \text{ contienne } r \text{ gènes}) &= P(\text{un segment de taille } x \\ &\quad \text{contienne } \tfrac{r}{z} \text{ groupes de gènes}) \\ &= \binom{m'}{\frac{r}{z}} x^{\frac{r}{z}} (1-x)^{m'-\frac{r}{z}} \\ &\quad \frac{r}{z} = 0, 1, 2, ..., m' \quad . \end{aligned}$$

Pour simplifier, supposons que m' est un entier. Les distributions ainsi obtenues sont pour des m qui sont multiples de z. On peut maintenant calculer la probabilité d'observer r gènes sur un segment de taille arbitraire.

La probabilité d'un segment de taille x sur un intervalle [0,1] est donnée par la proposition 1.4.1

$$f(x) = n(1-x)^{n-1} \quad , \quad 0 \leq x \leq 1 \quad ,$$

et la distribution du nombre de gènes pour un segment arbitraire est donnée par

$$\begin{aligned} P'(r) &= \int_0^1 f(x) B(m', x; \frac{r}{z})\, dx \\ &= \frac{nm'!(n+m'-\frac{r}{z}-1)!}{(n+m')!(m'-\frac{r}{z})!} \quad , \quad r = 0, z, 2z, ..., m \quad , \quad m' = \frac{m}{z} \quad . \end{aligned}$$

Si z=1, cette probabilité devient celle du modèle de base. Puisqu'il est impossible d'obtenir des observations pour les segments ne contenant aucun gène on doit donc conditionner la fonction de densité obtenue pour les cas où r est différent de 0.

PROPOSITION 3.1.1. *La fonction de masse du nombre de gènes sur un segment arbitraire étant donné que les gènes sont distribués en groupes de z gènes où z est fixé et les groupes sont indépendants les uns des autres, conditionnée sur le fait que r est différent de 0, est donnée par*

$$Q_z(r) = \frac{n}{m'}\binom{m'}{\frac{r}{z}}\frac{(\frac{r}{z})!(n+m'-\frac{r}{z}-1)!}{(n+m'-1)!} \quad , \quad r = z, 2z..., m \quad , \quad m' = \frac{m}{z} \quad .$$

L'espérance et la variance sont donnés par

$$E_z(r) = \frac{m}{n+1}(1+\frac{n}{m'})$$

$$Var_z(r) = \frac{m^2(m'+n)n(m'-1)}{m'^2(n+1)^2(n+2)} \quad .$$

Démonstration

$$\begin{aligned} Q_z(r) &= \frac{P'(r)}{1-P'(0)} \\ &= \frac{n\binom{m'}{\frac{r}{z}}\frac{(\frac{r}{z})!(n+m'-\frac{r}{z}-1)!}{(n+m')!}}{1-\frac{n(n+m'-1)!}{(n+m')!}} \\ &= \frac{n}{m'}\binom{m'}{\frac{r}{z}}\frac{(\frac{r}{z})!(n+m'-\frac{r}{z}-1)!}{(n+m'-1)!} \quad , \quad r = z, 2z, ..., m \quad . \end{aligned}$$

L'espérance est

$$\begin{aligned} E_z(r) &= \sum_{r=z}^{m} rQ_z(r) \\ &= \sum_{r=z}^{m} r\frac{n}{m'}\binom{m'}{\frac{r}{z}}\frac{(\frac{r}{z})!(n+m'-\frac{r}{z}-1)!}{(n+m'-1)!} \quad . \end{aligned}$$

On a $z = \frac{m}{m'}$. Posons $k = \frac{r}{z}$. Ainsi pour $k = 1, 2, ..., m'$ on obtient

$$\begin{aligned} E_z(r) &= \sum_{k=1}^{m'} \frac{m}{m'}k\frac{n}{m'}\binom{m'}{k}\frac{k!(n+m'-k-1)!}{(n+m'-1)!} \\ &= \frac{mn!(m'-1)!}{m'(n+m'-1)!}\sum_{k=0}^{m'} k\binom{n+m'-k-1}{m'-k} \end{aligned}$$

Par la proposition 2.1.2

$$\sum_{k=0}^{m'} k\binom{n+m'-k-1}{m'-k} = \binom{m'+n}{m'-1}$$

et donc

$$\begin{aligned} E_z(r) &= \frac{mn!(m'-1)!}{m'(n+m'-1)!}\frac{(m'+n)!}{(m'-1)!(n+1)!} \\ &= \frac{m}{n+1}(1+\frac{n}{m'}) \quad . \end{aligned}$$

Pour la variance, on a

$$Var_z(r) = Var(zX) = z^2 Var(X)$$

où $z = \frac{m}{m'}$ et où $X = \frac{r}{z}$ est une variable aléatoire qui a une fonction de masse Q(X) et une variance données par la proposition 2.1.1 pour X=1,2,...,m'. On a

donc,

$$Var_z(r) = \frac{m^2(m'+n)n(m'-1)}{m'^2(n+1)^2(n+2)} \quad .$$

La distribution ainsi obtenue est faite pour r multiple de z. Si on veut obtenir une fonction de probabilité pour $r = 1, 2, 3, ...$ on doit "bricoler" la fonction $Q_z(r)$. On a

$$Q_z(r) = \frac{n}{m'}\binom{m'}{\frac{r}{z}}\frac{(\frac{r}{z})!(n+m'-\frac{r}{z}-1)!}{(n+m'-1)!} \quad , \quad r = z, 2z..., m \quad .$$

Posons

$$P_z(r) = \frac{Q_z(iz)}{z} \quad , \quad r = (i-1)z+1, ..., iz \quad , \quad i = 1, ..., m' \quad .$$

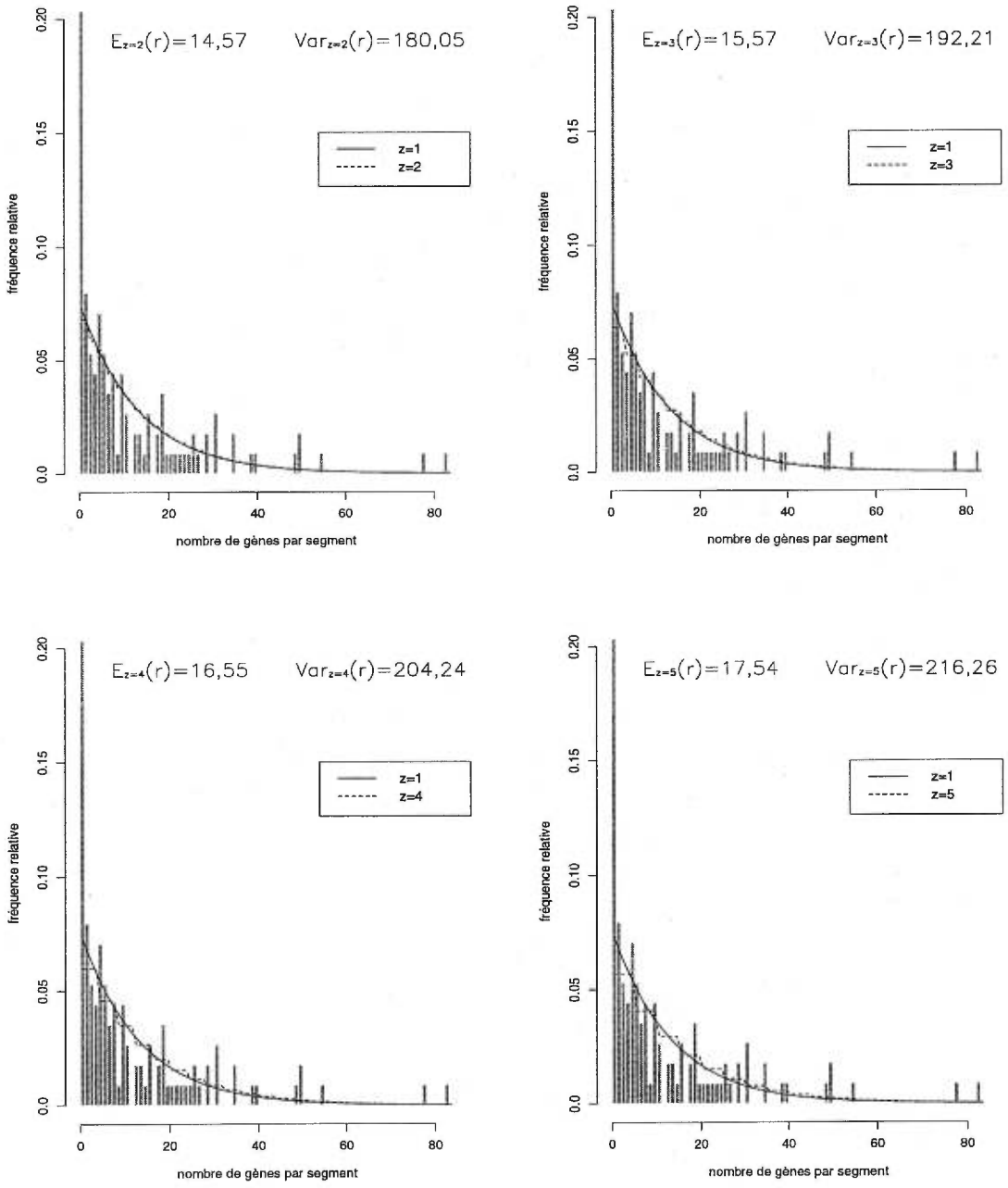
qui est une fonction de probabilité pour r=1,2,...,m.

3.1.2. Illustration du modèle en grappes

À partir des mêmes exemples de la section 2.2, soit avec 1 423 gènes identifiés communs à l'homme et la souris et où le nombre de segments conservés estimé est 113, 197, 236 et 284, on va vérifier l'ajustement de la fréquence relative du nombre de gènes sur un segment conservé avec le modèle de la distribution des gènes par grappe.

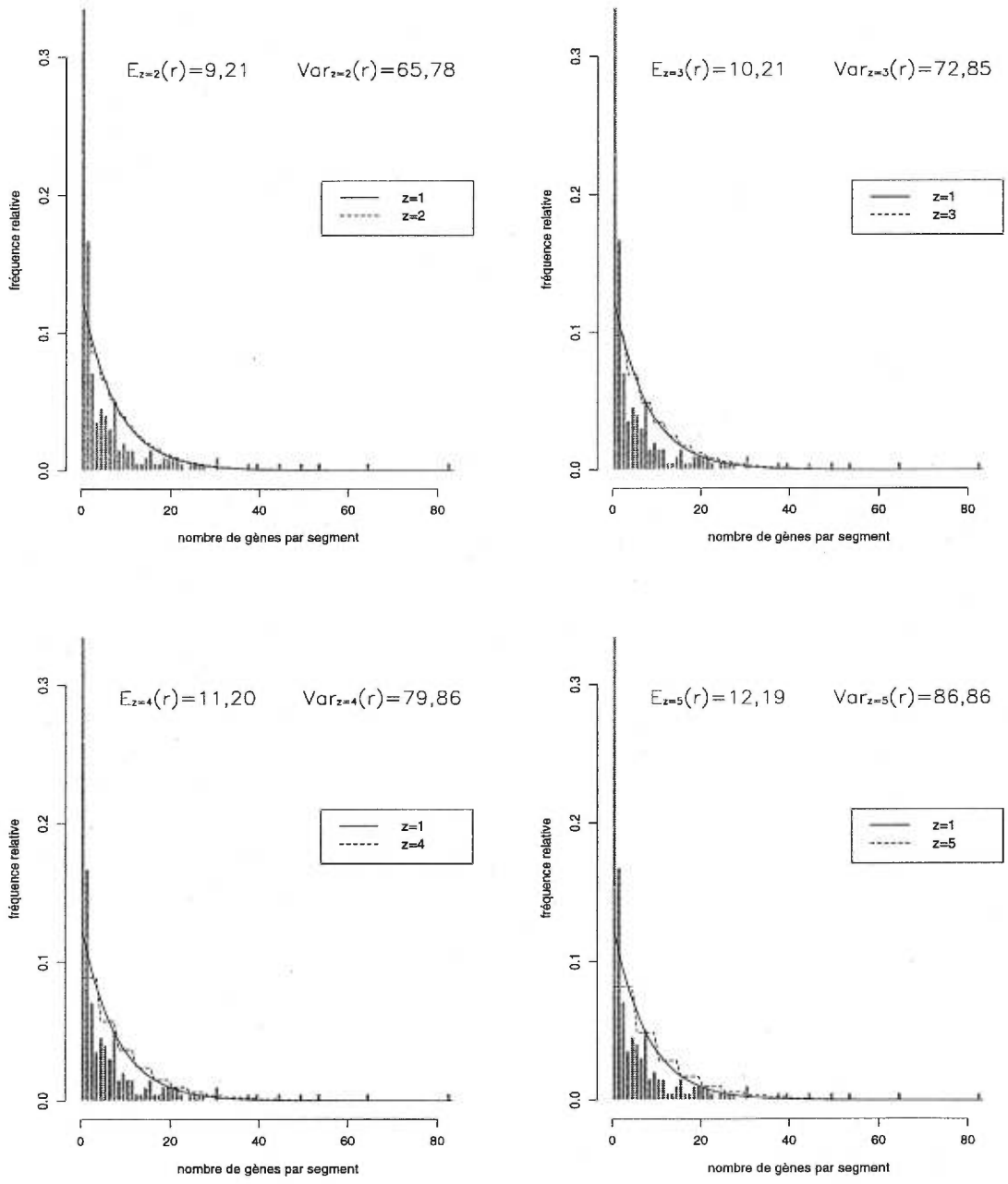
Pour chaque exemple on a fait varier z de 1 jusqu'à 5. Pour chaque cas on représente les résultats de la proposition 3.1.1, soit la courbe d'ajustement aux données, leur espérance et leur variance.

FIGURE 3.5. Illustrations du modèle de distribution des gènes par grappes. Cas où il y a 1423 gènes et 113 segments.



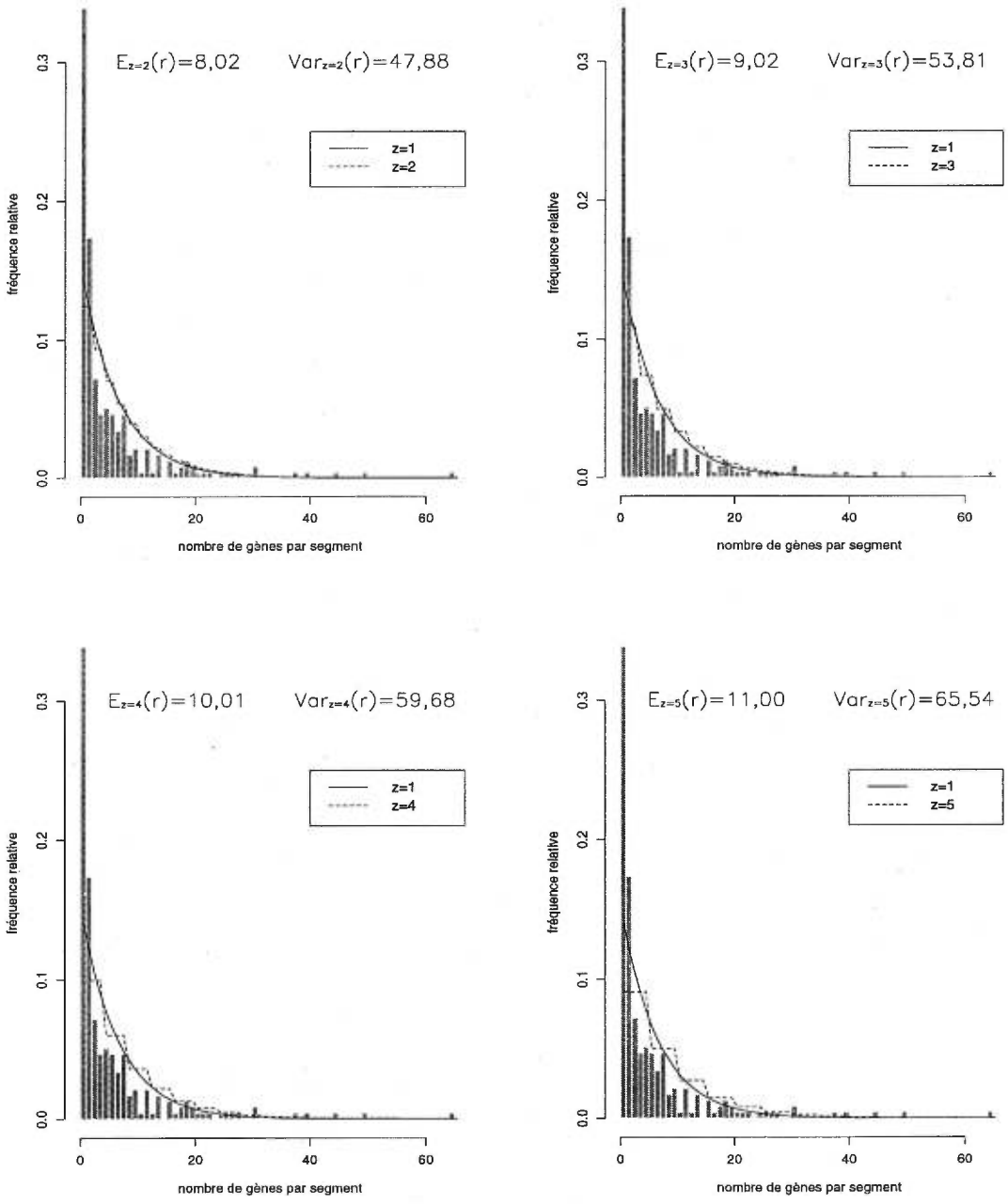
Nombre moyen de gènes sur un segment: 12,59
Variance du nombre de gènes sur un segment: 237,81

FIGURE 3.6. Illustrations du modèle de distribution des gènes par grappes. Cas où il y a 1423 gènes et 197 segments.



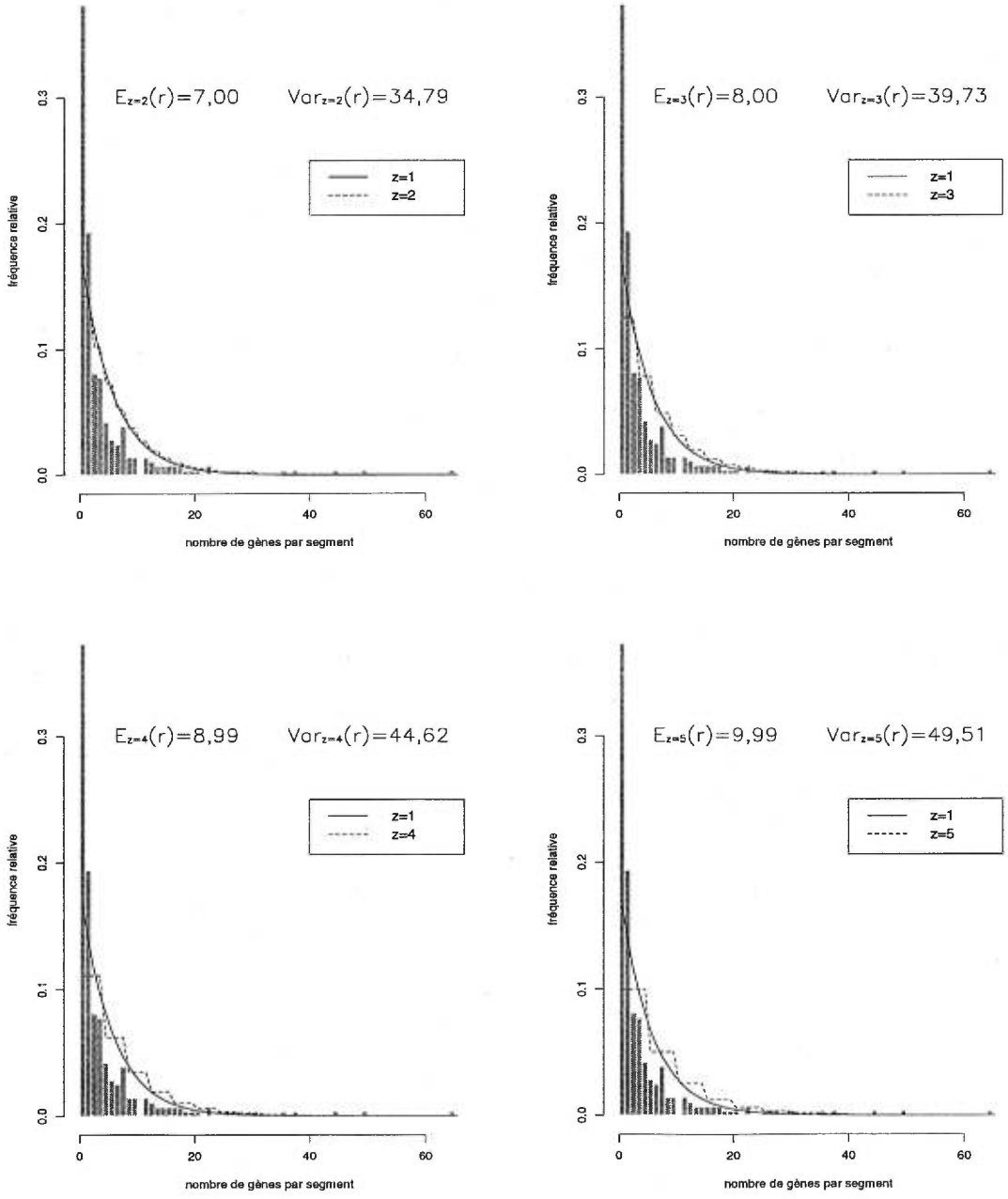
Nombre moyen de gènes sur un segment: 7,22
Variance du nombre de gènes sur un segment: 129,97

FIGURE 3.7. Illustrations du modèle de distribution des gènes par grappes. Cas où il y a 1423 gènes et 236 segments.



Nombre moyen de gènes sur un segment: 6,02
Variance du nombre de gènes sur un segment: 76,49

FIGURE 3.8. Illustrations du modèle de distribution des gènes par grappes. Cas où il y a 1423 gènes et 284 segments.



Nombre moyen de gènes sur un segment: 5,01
Variance du nombre moyen de gènes sur un segment: 62,11

3.1.3. Commentaires

Le fait de distribuer les gènes en grappes augmente l'espérance et la variance du nombre de gènes par segment. Pour des petites valeurs de z, telles que z=2 ou z=3, on voit très peu de différences avec le modèle de base. Pour z=4 ou z=5, on a toujours un problème de surestimation pour les r=5, 6,... , sauf pour le cas où le nombre de segments conservés est le plus petit, soit n=113.

La distribution en grappes ne semble pas améliorer le modèle de base, ce qui laisse supposer que l'hypothèse selon laquelle l'identification de la position des gènes à travers le génome n'est pas indépendante les unes des autres a très peu d'effet sur le modèle pour illustrer la fréquence relative du nombre de gènes sur un segment conservé.

Il faut peut-être explorer empiriquement s'il y a non-indépendance de positions des gènes et quelle forme statistique prend cette non-indépendance. Ensuite, on pourrait examiner l'effet de ce phénomène sur la distribution du nombre de gènes sur un segment arbitraire.

3.2. Modèle sans les segments contenant 1 seul gène

On a vu que le nombre empirique de segments conservés contenant un seul gène est problématique. On veut construire un modèle qui ne tiendra pas compte de ces segments.

3.2.1. Formulation du modèle sans les segments contenant 1 seul gène

Proposition 3.2.1. *La fonction de probabilité du nombre de gènes sur un segment arbitraire étant donné que ce nombre est différent de 0 et de 1 est*

$$W(r) = \frac{n(m-2)!(n+m-r-1)!}{(m+n-2)!(m-r)!} \quad , \quad r = 2, ..., m \quad .$$

L'espérance et le deuxième moment sont donnés par

$$E_W(r) = \frac{m+2n}{n+1}$$

$$E_W(r^2) = \frac{1}{m-1}\left(\frac{(m+n)(m+n-1)(2m+n)}{(n+2)(n+1)} - n\right) \quad .$$

Démonstration

$$\begin{aligned} W(r) &= P(r/r \neq 0 \, et \, r \neq 1) \\ &= \frac{P(r)}{1 - P(0) - P(1)} \\ &= \frac{\frac{nm!(n+m-r-1)!}{(n+m)!(m-r)!}}{1 - \frac{nm!(n+m-1)!}{(n+m)!m!} - \frac{nm!(n+m-2)!}{(n+m)!(m-1)!}} \\ &= \frac{n(m-2)!(n+m-r-1)!}{(m+n-2)!(m-r)!} \quad , \quad r = 2, ..., m \quad . \end{aligned}$$

Pour l'expression de l'espérance on a

$$\begin{aligned} E_W(r) &= \sum_{r=2}^{m} rW(r) \\ &= \sum_{r=0}^{m} rW(r) - W(1) \\ &= \frac{n(m-2)!(n-1)!}{(m+n-2)!} \sum_{r=0}^{m} r \binom{n+m-r-1}{m-r} - \frac{n}{m-1} \quad . \end{aligned}$$

Par la proposition 2.1.2

$$\sum_{r=0}^{m} r\binom{n+m-r-1}{m-r} = \binom{m+n}{m-1}$$

donc

$$\begin{aligned} E_W(r) &= \frac{n(m-2)!(n-1)!(m+n)!}{(m+n-2)!(m-1)!(n+1)!} - \frac{n}{m-1} \\ &= \frac{m+2n}{n+1} \ . \end{aligned}$$

Pour le deuxième moment on a

$$\begin{aligned} E_W(r^2) &= \sum_{r=2}^{m} r^2 W(r) \\ &= \sum_{r=0}^{m} r^2 W(r) - W(1) \\ &= \frac{n!(m-2)!}{(m+n-2)!} \sum_{r=0}^{m} r^2 \binom{n+m-r-1}{m-r} - \frac{n}{m-1} \ . \end{aligned}$$

Par la proposition 2.1.3

$$\sum_{r=0}^{m} r^2\binom{n+m-r-1}{m-r} = \binom{m+n+1}{m-1} + \binom{m+n}{m-2}$$

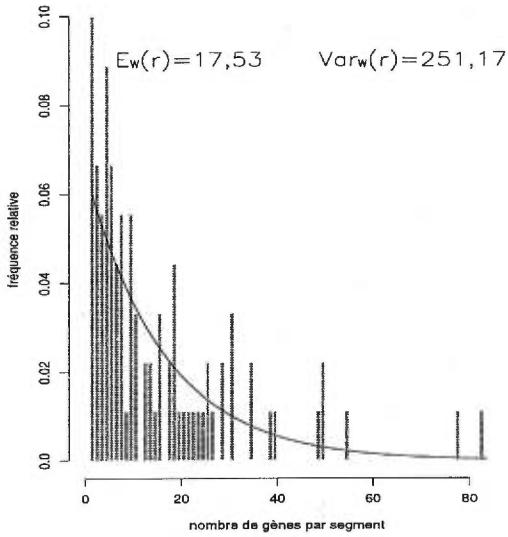
donc

$$\begin{aligned} E_W(r^2) &= \frac{n!(m-2)!}{(m+n-2)!}\left(\frac{(m+n+1)!}{(m-1)!(n+2)!} + \frac{(m+n)!}{(m-2)!(n+2)!}\right) - \frac{n}{m-1} \\ &= \frac{1}{m-1}\left(\frac{(m+n)(m+n-1)(2m+n)}{(n+2)(n+1)} - n\right) \ . \end{aligned}$$

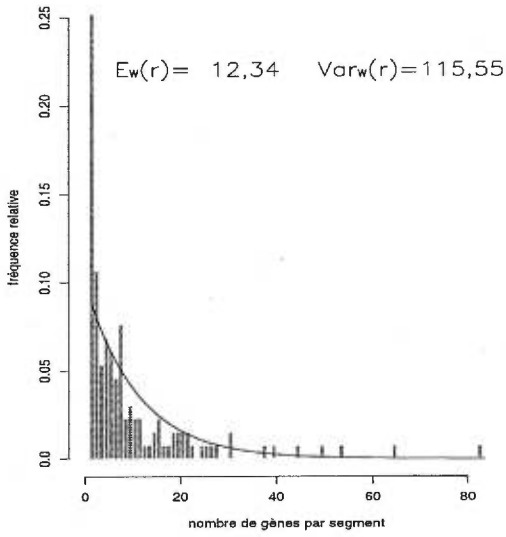
3.2.2. Illustration du modèle sans les segments contenant 1 seul gène

On utilise les mêmes exemples qu'auparavant, sauf qu'on enlève les segments contenant un seul gène. Le nombre de gènes total utilisé et le nombre de segments conservés seront donc plus petits. Avant on avait 1 423 gènes pour tous les exemples. On se retrouve donc avec une valeur de m différente pour chaque cas, soit m= 1400, 1357, 1343, 1317. On travaillera donc seulement avec les segments qui contiennent 2 gènes ou plus. Le nombre de segments conservés est donc 90, 131, 156, 178.

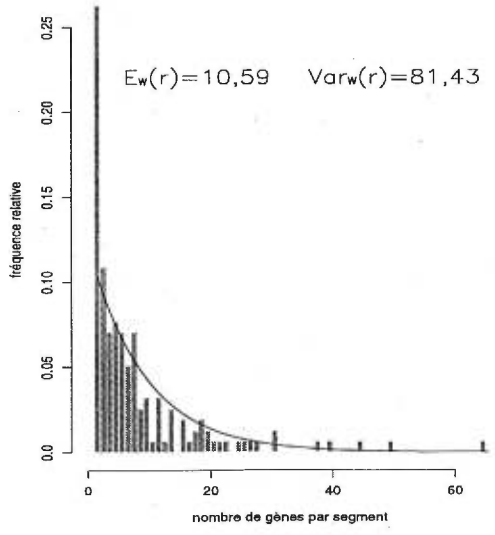
FIGURE 3.9. Illustrations du modèle de la distribution des gènes sur un segment sans ceux contenant un seul gène.



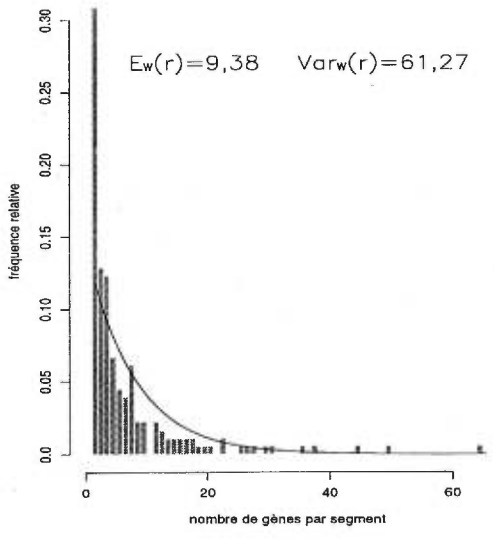
1400 gènes, 90 segments
Nombre moyen de gènes sur un segment: 15,55
Variance: 255,66



1357 gènes, 131 segments
Nombre moyen de gènes sur un segment: 10,35
Variance: 166,38



1343 gènes, 156 segments
Nombre moyen de gènes sur un segment: 8,60
Variance: 96,21



1317 gènes, 178 segments
Nombre moyen de gènes sur un segment: 7,39
Variance: 83,94

3.2.3. Commentaires

Le fait de ne pas tenir compte des segments contenant un seul gène augmente naturellement l'espérance et la variance du nombre de gènes par segment. La fonction ne surestime plus comme avant la fréquence relative pour r=5, 6,....

Puisqu'enlever les segments contenant un seul gène semble améliorer l'ajustement du modèle de base il serait intéressant d'examiner ce qui arrive lorsqu'on enlève aussi les segments contenant deux gènes. Les segments qui contiennent deux gènes sont aussi susceptibles de contenir des erreurs.

3.3. Modèle sans les segments contenant 1 ou 2 gènes

3.3.1. Formulation du modèle de base sans les segments contenant 1 ou 2 gènes

Proposition 3.3.1. *La fonction de probabilité du nombre de gènes sur un segment conservé étant donné que ce nombre est différent de 0, 1 et 2 est donnée par*

$$V(r) = \frac{n(m-3)!(n+m-r-1)!}{(m+n-3)!(m-r)!} \quad , \quad r = 3, ..., m \quad .$$

L'espérance et le deuxième moment sont donnés par

$$E_V(r) = \frac{m+3n}{n+1}$$

$$E_V(r^2) = \frac{(m+n)(m+n-1)(m+n-2)(2m+n)}{(n+2)(n+1)(m-2)(m-1)} - \frac{n^2+5nm-6n}{(m-1)(m-2)} \quad .$$

Démonstration

$$\begin{aligned} V(r) &= \frac{P(r)}{1-P(0)-P(1)-P(2)} \\ &= \frac{W(r)}{1-W(2)} \end{aligned}$$

$$= \frac{\frac{n(m-2)!(n+m-r-1)!}{(m+n-2)!(m-r)!}}{1-\frac{n(m-2)!(n+m-3)!}{(m+n-2)!(m-2)!}}$$

$$= \frac{n(m-3)!(n+m-r-1)!}{(m+n-3)!(m-r)!} \quad , \quad r=3,...,m \quad .$$

Pour l'expression de l'espérance on a

$$\begin{aligned} E_V(r) &= \sum_{r=3}^{m} rV(r) \\ &= \sum_{r=0}^{m} rV(r) - V(1) - 2V(2) \\ &= \sum_{r=0}^{m} r\frac{n(m-3)!(n+m-r-1)!}{(m+n-3)!(m-r)!} - \frac{n(m-3)!(n+m-2)!}{(m+n-3)!(m-1)!} \\ &\quad -\frac{2n(m-3)!(n+m-3)!}{(m+n-3)!(m-2)!} \\ &= \frac{n!(m-3)!}{(m+n-3)!}\sum_{r=0}^{m} r\binom{n+m-r-1}{m-r} - \frac{n(n+m-2)}{(m-1)(m-2)} - \frac{2n}{m-2} \quad . \end{aligned}$$

Par la proposition 2.1.2

$$\sum_{r=0}^{m} r\binom{n+m-r-1}{m-r} = \binom{m+n}{m-1}$$

donc

$$\begin{aligned} E_V(r) &= \frac{(m+n)(m+n-1)(m+n-2)}{(n+1)(m-1)(m-2)} - \frac{n(n+m-2)}{(m-1)(m-2)} - \frac{2n}{m-2} \\ &= \frac{m+3n}{n+1} \quad . \end{aligned}$$

Pour le deuxième moment on a

$$\begin{aligned} E_V(r^2) &= \sum_{r=3}^{m} r^2V(r) \\ &= \sum_{r=0}^{m} r^2V(r) - V(1) - 4V(2) \end{aligned}$$

$$= \frac{n!(m-3)!}{(m+n-3)!}\sum_{r=0}^{m} r^2\binom{n+m-r-1}{m-r} - \frac{n(n+m-2)}{(m-1)(m-2)} - \frac{4n}{m-2} \quad .$$

Par la proposition 2.1.3

$$\sum_{r=0}^{m} r^2\binom{n+m-r-1}{m-r} = \binom{m+n+1}{m-1} + \binom{m+n}{m-2}$$

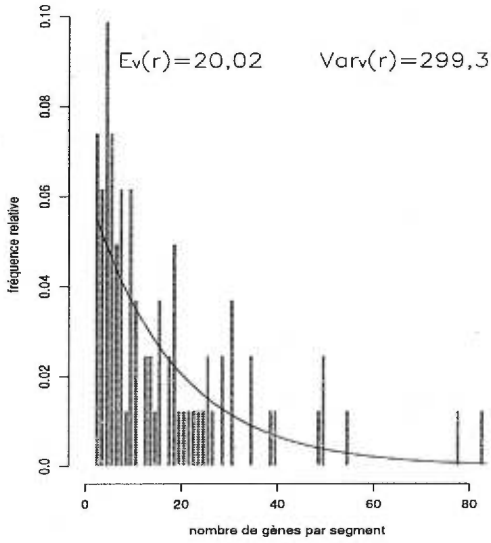
donc

$$\begin{aligned} E_V(r^2) &= \frac{(m+n)(m+n-1)(m+n-2)(2m+n)}{(n+2)(n+1)(m-2)(m-1)} - \frac{n(n+m-2)}{(m-1)(m-2)} - \frac{4n}{m-2} \\ &= \frac{(m+n)(m+n-1)(m+n-2)(2m+n)}{(n+2)(n+1)(m-2)(m-1)} - \frac{n^2+5nm-6n}{(m-1)(m-2)} \quad . \end{aligned}$$

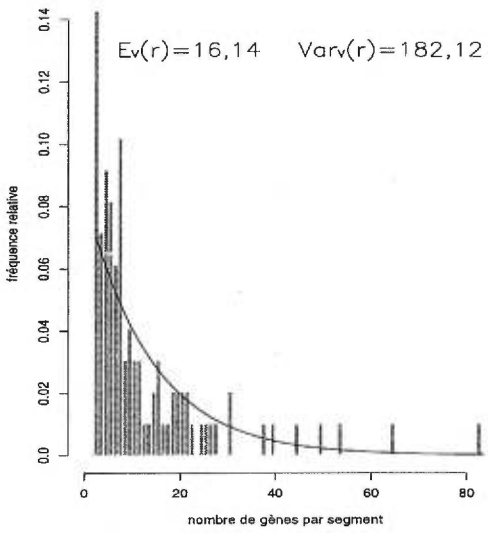
3.3.2. Illustration du modèle de base sans les segments contenant 1 ou 2 gènes

On utilise toujours les mêmes données et on enlève les segments contenant deux gènes en plus de ceux qui en contiennent seulement un. On se retrouve avec des exemples qui comptent 1382, 1291, 1261, et 1207 gènes et dont le nombre de segments conservés est estimé à 81, 98, 115, et 123. Pour chaque cas, on illustre l'application du modèle présenté à la proposition 3.3.1.

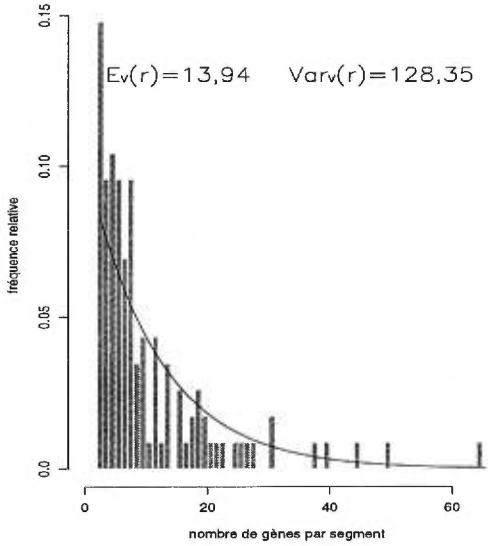
FIGURE 3.10. Illustrations du modèle de la distribution des gènes sur un segment sans ceux contenant 1 ou 2 gènes.



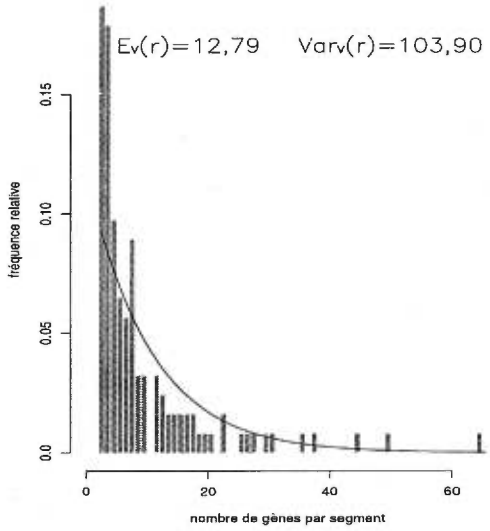
1382 gènes, 81 segments
Nombre moyen de gènes sur un segment: 17,06
Variance: 261,46



1291 gènes, 98 segments
Nombre moyen de gènes sur un segment: 13,17
Variance: 191,22



1261 gènes, 115 segments
Nombre moyen de gènes sur un segment: 10,97
Variance: 109,51



1207 gènes, 123 segments
Nombre moyen de gènes sur un segment: 9,81
Variance: 102,76

3.3.3. Commentaires

Le fait d'enlever les segments qui contiennent un et deux gènes fait en sorte que le modèle ne surestime plus les r=5, 6, ...,15 comme avant ce qui est donc une amélioration du modèle de base. Il est vraisemblable de croire que les segments qui contiennent 1 ou 2 gènes sont plus susceptibles de souffrir d'erreurs expérimentales. Par exemple, dans les bases de données publiques, une bonne proportion des segments qui ne contiennent qu'un gène à un certain moment, ne sont plus dans la même base de données un an plus tard

CHAPITRE 4

MODÈLE DE DISTRIBUTION DU NOMBRE DE GÈNES SUR UN SEGMENT CONSERVÉ AVEC PARAMÈTRES DE CONCENTRATION DES GÈNES ET DES POINTS DE RUPTURE

On introduit ici des paramètres de concentration au modèle de base de la distribution des gènes présenté au chapitre 2. Les m gènes et les n points de cassure ne sont plus distribués uniformément à travers tout le génome. On représente le génome en deux sections, une partie A et une partie B, où chacune des sections est de longueur 1. Pour nos fins, il n'est pas nécessaire de se préoccuper où se trouvent physiquement ces secteurs à travers le génome. Ces sections ne correspondent pas à deux moitiés physiques identifiables du génome. On suppose qu'il y a αn points de cassure et βm gènes sur la partie A et $(1-\alpha)n$ points de cassure et $(1-\beta)m$ gènes sur la partie B où $0 \leq \alpha \leq 1$ et $0 \leq \beta \leq 1$. Les gènes et les points de rupture sont distribués de manière uniforme et indépendamment les uns des autres sur chacune des parties A et B.

En faisant varier α et β, cette représentation permet la concentration de plusieurs gènes dans une partie du génome où il y a peu de points de cassure et donc peu de segments, en même temps que de retrouver dans l'autre partie du génome beaucoup de segments et peu de gènes.

Dans un premier temps, on pose $\alpha = \frac{1}{2}$ i.e. que la moitié des points de cassure est distribuée uniformément sur chacune des deux parties du génome et on fait varier β, la proportion de gènes sur la section A. Dans un deuxième temps, on fixe le nombre de gènes sur chacune des sections en posant $\beta = \frac{1}{2}$ et on fait varier α, la proportion des points de cassure sur la partie A du génome.

FIGURE 4.11. Représentation du modèle

4.1. MODÈLE OÙ α=$\frac{1}{2}$ ET β VARIE

4.1.1. Formulation du modèle où α=$\frac{1}{2}$ et β varie

On représente le génome en deux sections, une partie A et une partie B, où chacune des sections est de longueur 1. Les points de cassure sont séparés également en deux (α=$\frac{1}{2}$) et distribués uniformément sur chaque partie.

Soit N, le nombre de cassures à l'intérieur de chaque partie. Posons

$$N = [\alpha n] = \left[\frac{n}{2}\right] \ .$$

Si n est impair on va retrouver en tout $n+1$ segments et si n est pair on retrouve $n+2$ segments, ceci en supposant qu'il y a une cassure entre les 2 parties.

Ensuite, on distribue les m gènes de la manière suivante: βm gènes sur la partie A et $(1-\beta)m$ sur la partie B. Soit m_1, le nombre de gènes distribués

uniformément sur la partie A. On a

$$m_1 = \beta m \quad .$$

Pour simplifier, m_1 est la valeur de βm arrondit à l'entier le plus près. On retrouve

$$m_2 = m - m_1 \quad ,$$

le nombre de gènes qui appartient à la partie B, où l'on pose sans perte de généralité que $m_1 \leq m_2$.

On veut connaître la probabilité qu'un segment arbitraire contienne r gènes. On suppose que ce segment arbitraire a autant de chance de provenir de la partie A que de la partie B. Cette probabilité est représentée par

$$\begin{aligned} P_\beta(r \text{ gènes sur un segment}) &= P(r/segment \in A)P(segment \in A) \\ &\quad + P(r/segment \in B)P(segment \in B) \\ &= P(r/segment \in A)\frac{1}{2} + P(r/segment \in B)\frac{1}{2} \\ &\qquad 0 \leq r \leq m_2 \quad . \end{aligned}$$

On suppose que sur chaque partie, les points de cassure et les gènes sont distribués uniformément et indépendamment les uns des autres. On peut donc utiliser la probabilité du chapitre 2 donnée par l'équation (2.1.1) pour $P(r/segment \in A)$ et $P(r/segment \in B)$. On obtient

$$P(r/segment \in j) = \frac{Nm_i!(N + m_i - r - 1)!}{(N + m_i)!(m_i - r)!} \quad , \quad 0 \leq r \leq m_i \quad ,$$

où $i = 1$ si $j = A$ et $i = 2$ si $j = B$. Cette probabilité est celle qu'un segment arbitraire, tiré parmi les $N+1$ segments de longueur aléatoire d'un intervalle [0,1] où sont distribués m_i gènes au hasard, contienne r gènes. Ce qui nous amène à la proposition suivante:

PROPOSITION 4.1.1. *La fonction de masse du nombre de gènes sur un segment arbitraire suivant le modèle où $\alpha = \frac{1}{2}$ et β varie, conditionnée sur le fait que r est différent de 0, est donnée par*

$$Q_\beta(r) = \begin{cases} KN\Big(\frac{m_1!(N+m_1-r-1)!}{(N+m_1)!(m_1-r)!} + \frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\Big) \\ \qquad\qquad si \quad 1 \le r \le m_1 \\ KN\Big(\frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\Big) \\ \qquad\qquad si \quad m_1+1 \le r \le m_2 \end{cases}$$

L'espérance et le deuxième moment sont donnés par

$$E_\beta(r) = \frac{K(m_1+m_2)}{N+1} \quad ,$$

$$E_\beta(r^2) = \frac{K(m_1(2m_1+N)+m_2(2m_2+N))}{(N+2)(N+1)}$$

où

$$K = \frac{(N+m_1)(N+m_2)}{N(m_1+m_2)+2m_1m_2} \quad .$$

Démonstration

$$P_\beta(r) = P(r/segment \in A)\frac{1}{2} + P(r/segment \in B)\frac{1}{2}$$

$$= \begin{cases} \frac{Nm_1!(N+m_1-r-1)!}{(N+m_1)!(m_1-r)!}\frac{1}{2} + \frac{Nm_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\frac{1}{2} \\ \qquad\qquad si \quad 0 \le r \le m_1 \\ \frac{Nm_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\frac{1}{2} \\ \qquad\qquad si \quad m_1+1 \le r \le m_2 \end{cases}$$

La valeur de cette distribution pour r=0 est donnée par

$$\begin{aligned} P_\beta(0) &= \frac{Nm_1!(N+m_1-1)!}{(N+m_1)!m_1!}\frac{1}{2} + \frac{Nm_2!(N+m_2-1)!}{(N+m_2)!m_2!}\frac{1}{2} \\ &= \frac{N(2N+m_1+m_2)}{2(N+m_1)(N+m_2)} \end{aligned}$$

Donc

$$\begin{aligned} Q_\beta(r) &= P_\beta(r/r \neq 0) = \frac{P_\beta(r)}{1-P_\beta(0)} \\ &= \begin{cases} \dfrac{\frac{N}{2}\left(\frac{m_1!(N+m_1-r-1)!}{(N+m_1)!(m_1-r)!} + \frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\right)}{1-\frac{N(2N+m_1+m_2)}{2(N+m_1)(N+m_2)}} & si \quad 1 \leq r \leq m_1 \\ \dfrac{\frac{N}{2}\left(\frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\right)}{1-\frac{N(2N+m_1+m_2)}{2(N+m_1)(N+m_2)}} & si \quad m_1+1 \leq r \leq m_2 \end{cases} \\ &= \begin{cases} \dfrac{N\left(\frac{m_1!(N+m_1-r-1)!}{(N+m_1)!(m_1-r)!} + \frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\right)}{\frac{N(m_1+m_2)+2m_1m_2}{(N+m_1)(N+m_2)}} & si \quad 1 \leq r \leq m_1 \\ \dfrac{N\left(\frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\right)}{\frac{N(m_1+m_2)+2m_1m_2}{(N+m_1)(N+m_2)}} & si \quad m_1+1 \leq r \leq m_2 \end{cases} . \end{aligned}$$

Pour l'espérance on a

$$\begin{aligned}
E_\beta(r) &= \sum_{r=1}^{m_2} rP(r) \\
&= \sum_{r=1}^{m_1} rP(r) + \sum_{r=m_1+1}^{m_2} rP(r) \\
&= \sum_{r=1}^{m_1} rKN\Big(\frac{m_1!(N+m_1-r-1)!}{(N+m_1)!(m_1-r)!} + \frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\Big) \\
&\quad + \sum_{r=m_1+1}^{m_2} rKN\Big(\frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\Big) \\
&= \frac{KN!m_1!}{(N+m_1)!}\sum_{r=1}^{m_1} r\binom{N+m_1-r-1}{m_1-r} + \frac{KN!m_2!}{(N+m_2)!}\sum_{r=1}^{m_2} r\binom{N+m_2-r-1}{m_2-r} .
\end{aligned}$$

Par la proposition 2.1.2

$$\sum_{r=1}^{m_i} r\binom{N+m_i-r-1}{m_i-r} = \binom{m_i+N}{m_i-1}$$

donc

$$\begin{aligned}
E_\beta(r) &= \frac{KN!m_1!}{(N+m_1)!}\Big(\frac{(m_1+N)!}{(m_1-1)!(N+1)!}\Big) + \frac{KN!m_2!}{(N+m_2)!}\Big(\frac{(m_2+N)!}{(m_2-1)!(N+1)!}\Big) \\
&= \frac{K(m_1+m_2)}{N+1} .
\end{aligned}$$

Pour le deuxième moment on a

$$\begin{aligned}
E_\beta(r^2) &= \sum_{r=1}^{m_2} r^2P(r) \\
&= \sum_{r=1}^{m_1} r^2P(r) + \sum_{r=m_1+1}^{m_2} r^2P(r) \\
&= \sum_{r=1}^{m_1} r^2KN\Big(\frac{m_1!(N+m_1-r-1)!}{(N+m_1)!(m_1-r)!} + \frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\Big)
\end{aligned}$$

$$+\sum_{r=m_1+1}^{m_2} r^2 KN\Big(\frac{m_2!(N+m_2-r-1)!}{(N+m_2)!(m_2-r)!}\Big)$$

$$= \frac{KN!m_1!}{(N+m_1)!}\sum_{r=1}^{m_1} r^2\binom{N+m_1-r-1}{m_1-r} + \frac{KN!m_2!}{(N+m_2)!}\sum_{r=1}^{m_2} r^2\binom{N+m_2-r-1}{m_2-r} .$$

Par la proposition 2.1.3

$$\sum_{r=1}^{m_i} r^2\binom{N+m_i-r-1}{m_i-r} = \binom{m_i+N+1}{m_i-1} + \binom{m_i+N}{m_i-2}$$

donc

$$E_\beta(r^2) = \frac{Km_1!N!}{(N+m_1)!}\Big(\frac{(m_1+N+1)!}{(m_1-1)!(N+2)!} + \frac{(m_1+N)!}{(m_1-2)!(N+2)!}\Big)$$

$$+\frac{KN!m_2!}{(N+m_2)!}\Big(\frac{(m_2+N+1)!}{(m_2-1)!(N+2)!} + \frac{(m_2+N)!}{(m_2-2)!(N+2)!}\Big)$$

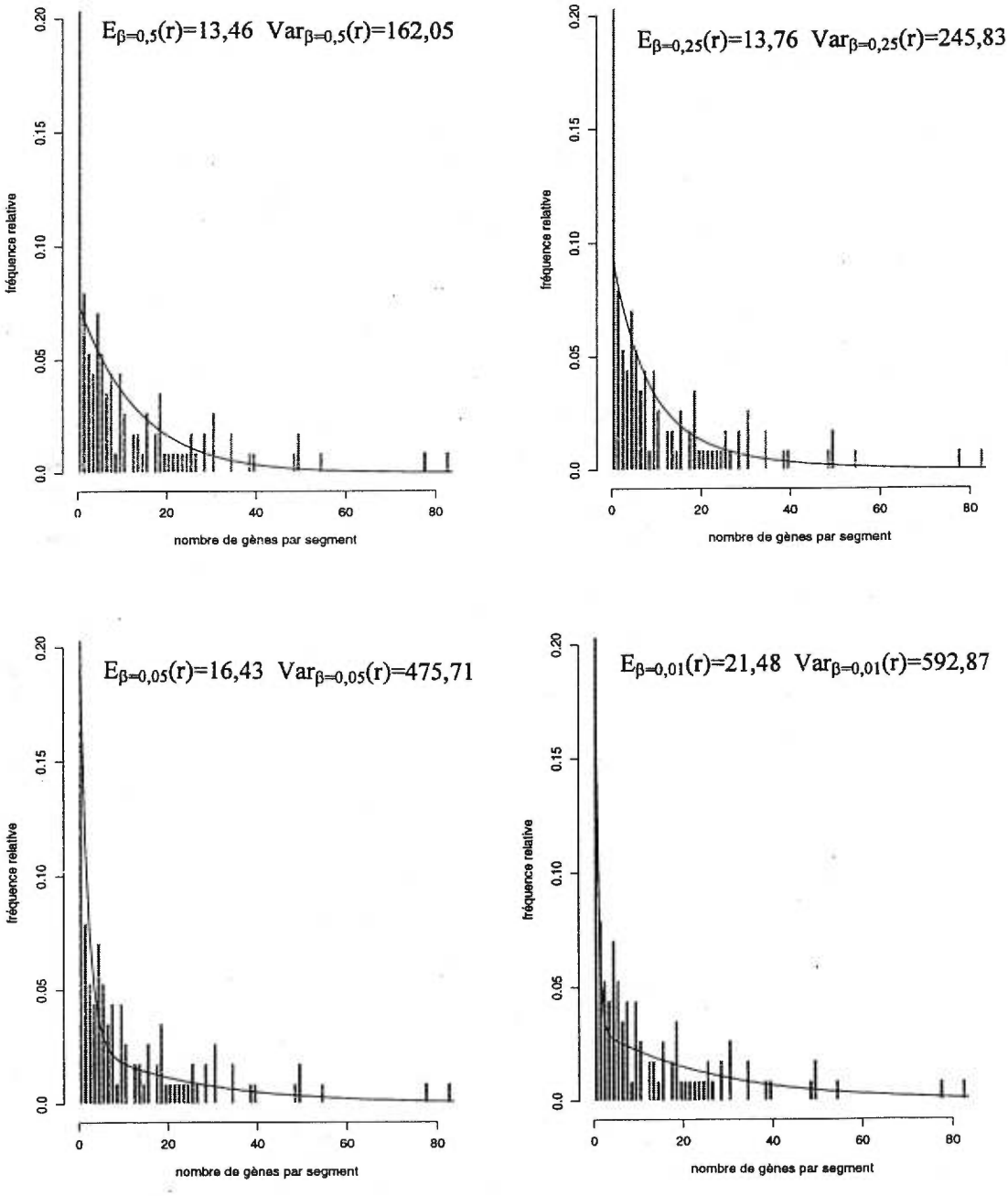
$$= \frac{K(m_1(2m_1+N)+m_2(2m_2+N))}{(N+2)(N+1)} .$$

4.1.2. Illustration du modèle où $\alpha = \frac{1}{2}$ et β varie

Avec les mêmes exemples que précédemment, i.e. où le nombre de gènes communs à l'homme et la souris est de 1423 et où le nombre estimé de segments conservés est 113, 197, 236, 284, on vérifie l'ajustement du modèle où la moitié des points de cassure est répartie également sur les 2 parties et où le nombre de gènes sur chacune des sections varie.

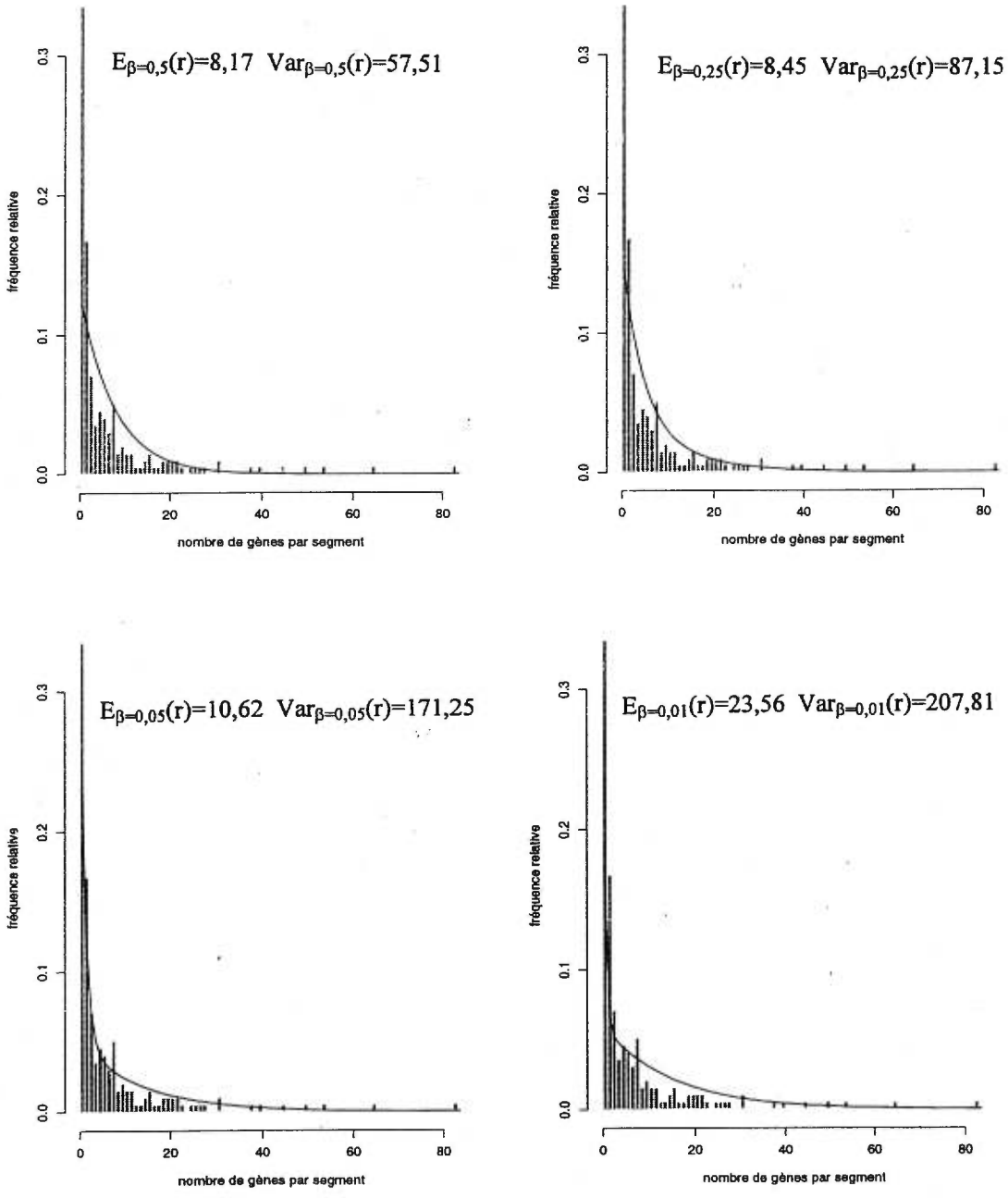
On pose tout d'abord β=0.5, ceci revient au modèle de base présenté au chapitre 2. Ensuite, on présente le modèle de la proposition 4.1.1 pour β=0,25, 0,05 et 0,01.

FIGURE 4.12. Illustrations du modèle de distribution des gènes lorsque $\alpha = \frac{1}{2}$ et β varie. Cas où il y a 1423 gènes et 113 segments



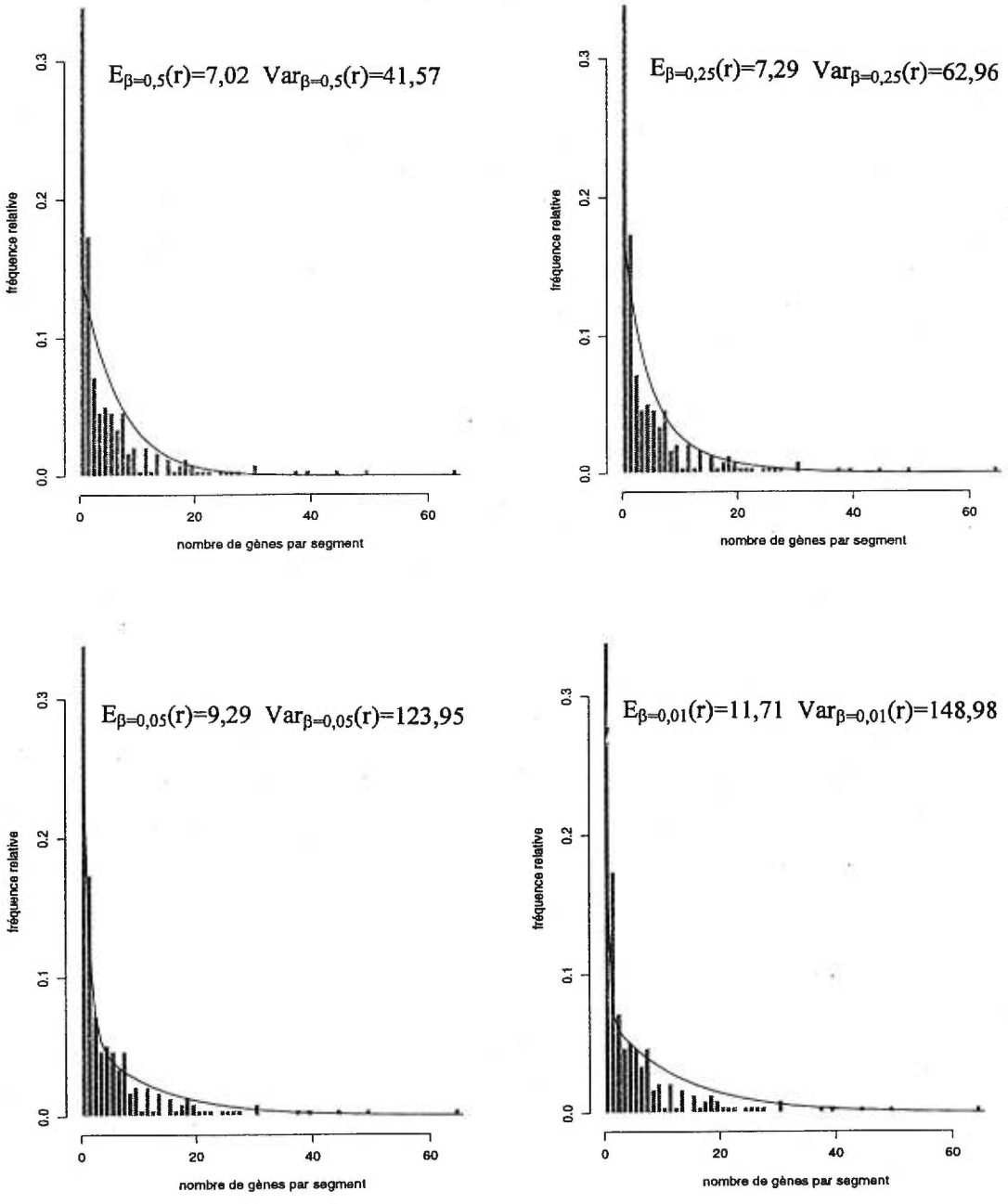
Nombre moyen de gènes sur un segment: 12,59
Variance du nombre de gènes sur un segment: 237,81

FIGURE 4.13. Illustrations du modèle de distribution des gènes lorsque $\alpha = \frac{1}{2}$ et β varie. Cas où il y a 1423 gènes et 197 segments



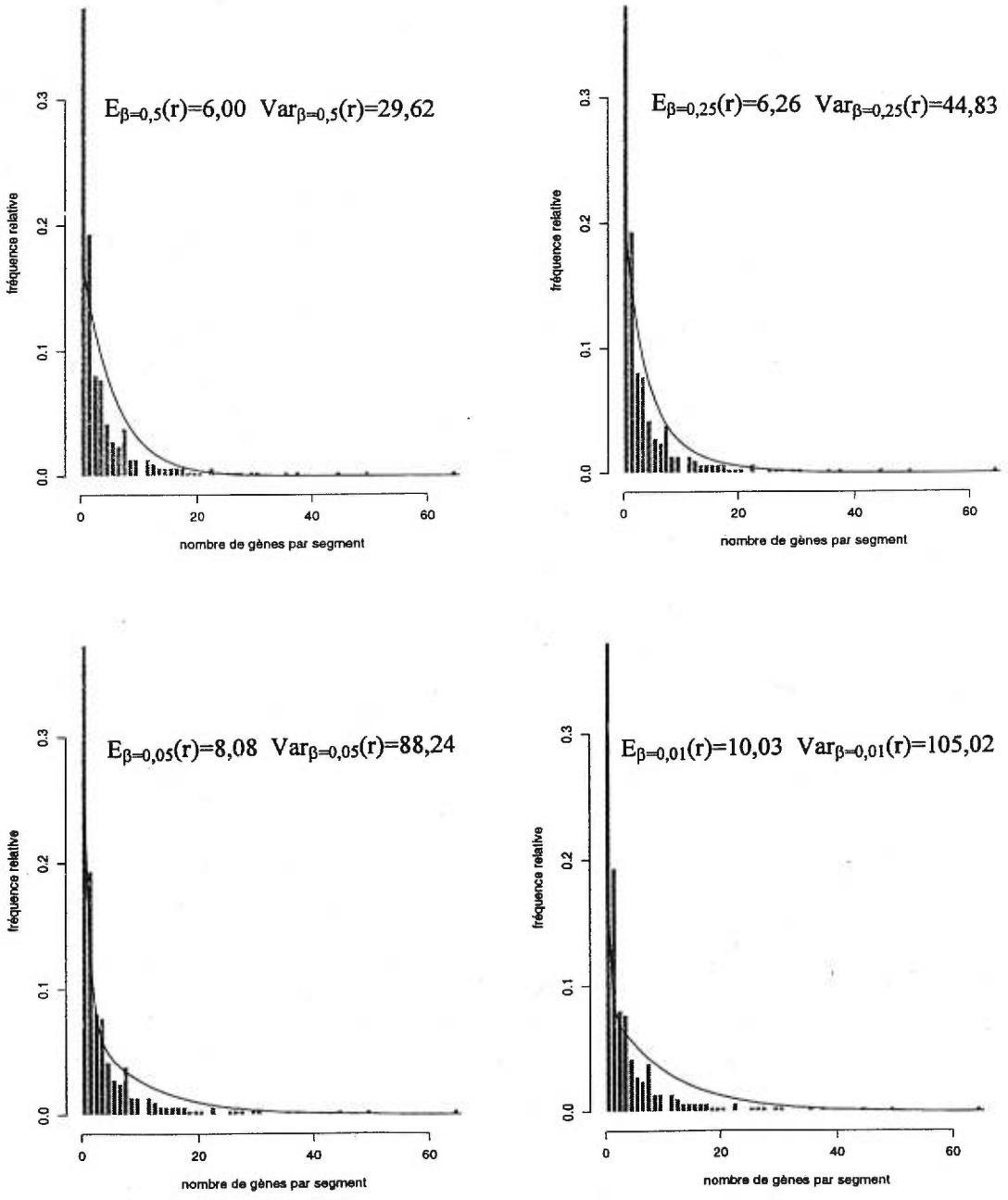
Nombre moyen de gènes sur un segment: 7,22
Variance du nombre moyen de gènes sur un segment: 129,97

FIGURE 4.14. Illustrations du modèle de distribution des gènes lorsque $\alpha = \frac{1}{2}$ et β varie. Cas où il y a 1423 gènes et 236 segments



Nombre moyen de gènes sur un segment: 6,02
Variance du nombre de gènes sur un segment: 76,49

FIGURE 4.15. Illustrations du modèle de distribution des gènes lorsque $\alpha = \frac{1}{2}$ et β varie. Cas où il y a 1423 gènes et 284 segments



Nombre moyen de gènes sur un segment: 5,01
Variance du nombre de gènes sur un segment: 62,11

4.1.3. Commentaires

On note une amélioration entre les valeurs observées et la prévision du modèle où $\alpha = \frac{1}{2}$ et β varie pour des petites valeurs de β , en particulier pour β=0,05. Dans ce cas, on retrouve en même temps une surestimation du nombre moyen de gènes sur un segment et de la variance du nombre de gènes sur un segment. Puisque le modèle diminue le poids donné aux petites valeurs de r, il augmente celui pour r plus grand, ce qui fait augmenter les valeurs de l'espérance et de la variance.

La modification de l'hypothèse de distribution des gènes identifiés communs aux deux espèces et des points de rupture à travers tout le génome se reflète sur l'ajustement des données. En séparant le génome en deux parties, en distribuant uniformément la moitié des points de rupture sur chacune des sections et une petite proportion des gènes identifiés, par exemple 5% sur une partie et en distribuant le reste uniformément sur l'autre section du génome, on améliore l'estimation de la fréquence relative des segments qui contiennent r gènes pour r=5,...,15.

4.2. Modèle où $\beta = \frac{1}{2}$ et α varie

4.2.1. Formulation du modèle où $\beta = \frac{1}{2}$ et α varie

Le génome est toujours représenté en 2 parties, soit A et B de longueur 1. Le nombre de gènes est séparé également en 2 et les gènes sont distribués uniformément sur chacune des sections. Posons

$$M = [\beta m] = [\frac{m}{2}]$$

où m est le nombre total de gènes. Donc si m est impair il manquera 1 gène et si m est pair le nombre total de gènes sur les 2 parties est bien m.

On distribue ensuite αn points de cassure sur A et $(1-\alpha)n$ points de cassure sur B. Soit n_1, le nombre de points de rupture distribués uniformément sur la partie A. On a

$$n_1 = \alpha n \quad .$$

Pour simplifier, n_1 est αn arrondit à l'entier le plus près et on retrouve

$$n_2 = n - n_1 \quad ,$$

où n est le nombre total de points de cassure et n_2 le nombre de ces points qui sont répartis uniformément sur la partie B. On suppose sans perte de généralité que $n_1 \leq n_2$.

Comme il a été présenté dans la section 4.1 la probabilité qu'un segment arbitraire contienne r gènes est donné par

$$\begin{aligned} P_\alpha(r \text{ gènes sur un segment}) &= P(r/segment \in A)\frac{1}{2} \\ &\quad + P(r/segment \in B)\frac{1}{2} \\ &\qquad 0 \leq r \leq M \end{aligned}$$

et on a

$$P(r/segment \in j) = \frac{n_i M!(n_i + M - r - 1)!}{(n_i + M)!(M - r)!}$$

où $i = 1$ si $j = A$ et $i = 2$ si $j = B$.

On obtient alors la proposition suivante:

PROPOSITION 4.2.1. *La fonction de masse du nombre de gènes sur un segment arbitraire suivant le modèle où $\beta = \frac{1}{2}$ et α varie, conditionnée sur le fait que r est différent de 0 est donné par*

$$\begin{aligned} Q_\alpha(r) &= \frac{C(M-1)!}{(M-r)!}\left(\frac{n_1(n_1 + M - r - 1)!}{(n_1 + M)!} + \frac{n_2(n_2 + M - r - 1)!}{(n_2 + M)!}\right) \quad , \\ &\qquad 1 \leq r \leq M \quad . \end{aligned}$$

L'espérance et le deuxième moment sont donnés par

$$E_\alpha(r) = \frac{C(n_1+n_2+2)}{(n_1+1)(n_2+1)}$$

$$E_\alpha(r^2) = C\Big(\frac{2M+n_1}{(n_1+2)(n_1+1)} + \frac{2M+n_2}{(n_2+2)(n_2+1)}\Big)$$

où

$$C = \frac{(n_1+M)(n_2+M)}{(n_1+n_2+2M)} \quad .$$

Démonstration

$$\begin{aligned} P_\alpha(r) &= P(r/\text{ segment } \in A)\frac{1}{2} + P(r/\text{ segment } \in B)\frac{1}{2} \\ &= \frac{n_1 M!(n_1+M-r-1)!}{2(n_1+M)!(M-r)!} + \frac{n_2 M!(n_2+M-r-1)!}{2(n_2+M)!(M-r)!} \\ &= \frac{M!}{2(M-r)!}\Big(\frac{n_1(n_1+M-r-1)!}{(n_1+M)!} + \frac{n_2(n_2+M-r-1)!}{(n_2+M)!}\Big) \quad . \end{aligned}$$

La valeur de cette distribution pour r=0 est donnée par

$$\begin{aligned} P_\alpha(0) &= \frac{n_1 M!(n_1+M-1)!}{2(n_1+M)!M!} + \frac{n_2 M!(n_2+M-1)!}{2(n_2+M)!M!} \\ &= \frac{2n_1n_2 + M(n_1+n_2)}{2(n_1+M)(n_2+M)} \quad , \end{aligned}$$

donc

$$\begin{aligned} Q_\alpha(r) &= P_\alpha(r/r \neq 0) = \frac{P_\alpha(r)}{1-P_\alpha(0)} \\ &= \frac{\dfrac{M!}{2(M-r)!}\Big(\dfrac{n_1(n_1+M-r-1)!}{(n_1+M)!} + \dfrac{n_2(n_2+M-r-1)!}{(n_2+M)!}\Big)}{1 - \dfrac{2n_1n_2+M(n_1+n_2)}{2(n_1+M)(n_2+M)}} \end{aligned}$$

$$= \frac{C(M-1)!}{(M-r)!}\Big(\frac{n_1(n_1+M-r-1)!}{(n_1+M)!}+\frac{n_2(n_2+M-r-1)!}{(n_2+M)!}\Big) \ , \quad 1 \le r \le M \ .$$

Pour l'espérance on a

$$\begin{aligned} E_\alpha(r) &= \sum_{r=1}^{M} rP(r) \\ &= \sum_{r=1}^{M} r\frac{C(M-1)!}{(M-r)!}\Big(\frac{n_1(n_1+M-r-1)!}{(n_1+M)!}+\frac{n_2(n_2+M-r-1)!}{(n_2+M)!}\Big) \\ &= C(M-1)!\Big(\frac{n_1!}{(n_1+M)!}\sum_{r=1}^{M} r\frac{(n_1+M-r-1)!}{(M-r)!(n_1-1)!} \\ &\qquad +\frac{n_2!}{(n_2+M)!}\sum_{r=1}^{M} r\frac{(n_2+M-r-1)!}{(M-r)!(n_2-1)!}\Big) \ . \end{aligned}$$

Par la proposition 2.1.2

$$\sum_{r=1}^{M} r\binom{n_i+M-r-1}{M-r} = \binom{M+n_i}{M-1}$$

donc

$$\begin{aligned} E_\alpha(r) &= C(M-1)!\Big(\frac{n_1!}{(n_1+M)!}\frac{(M+n_1)!}{(M-1)!(n_1+1)!} \\ &\qquad +\frac{n_2!}{(n_2+M)!}\frac{(M+n_2)!}{(M-1)!(n_2+1)!}\Big) \\ &= \frac{C(n_1+n_2+2)}{(n_1+1)(n_2+1)} \ . \end{aligned}$$

Pour le deuxième moment on a

$$E_\alpha(r^2) = \sum_{r=1}^{M} r^2P(r)$$

$$= \sum_{r=1}^{M} r^2 \frac{C(M-1)!}{(M-r)!}\Big(\frac{n_1(n_1+M-r-1)!}{(n_1+M)!} + \frac{n_2(n_2+M-r-1)!}{(n_2+M)!}\Big)$$

$$= C(M-1)!\Big(\sum_{r=1}^{M} r^2 \frac{n_1(n_1+M-r-1)!}{(M-r)!(n_1+M)!}$$

$$+ \sum_{r=1}^{M} r^2 \frac{n_2(n_2+M-r-1)!}{(M-r)!(n_2+M)!}\Big) \ .$$

Par la proposition 2.1.3

$$\sum_{r=1}^{M} r^2 \binom{n_i+M-r-1}{M-r} = \binom{M+n_i+1}{M-1} + \binom{M+n_i}{M-2}$$

donc

$$E_\alpha(r^2) = C(M-1)!\Big(\frac{n_1!}{(n_1+M)!}\Big(\frac{(M+n_1+1)!}{(M-1)!(n_1+2)!} + \frac{(M+n_1)!}{(M-2)!(n_1+2)!}\Big)$$

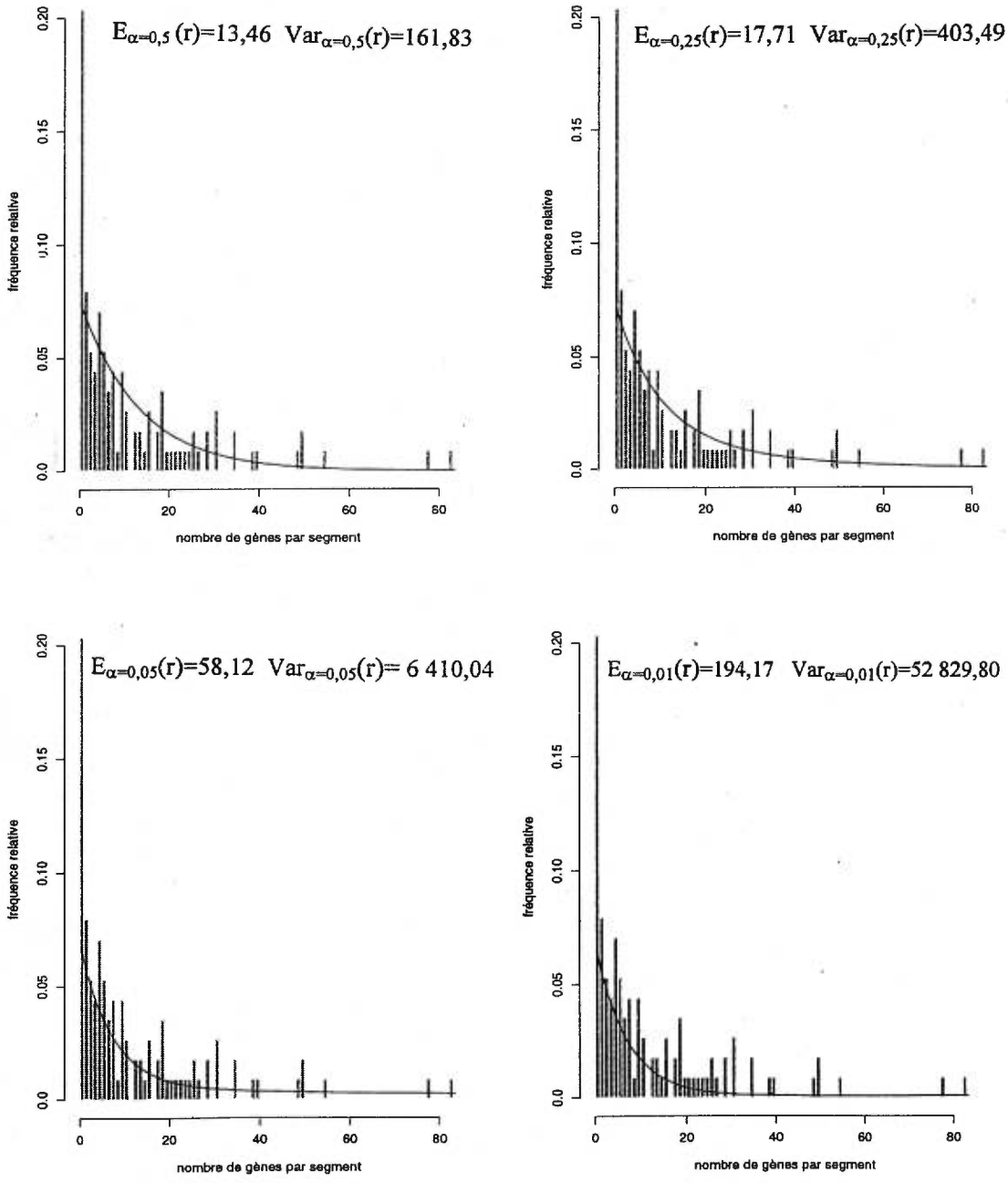
$$+ \frac{n_2!}{(n_2+M)!}\Big(\frac{(M+n_2+1)!}{(M-1)!(n_2+2)!} + \frac{(M+n_2)!}{(M-2)!(n_2+2)!}\Big)\Big)$$

$$= C\Big(\frac{2M+n_1}{(n_1+2)(n_1+1)} + \frac{2M+n_2}{(n_2+2)(n_2+1)}\Big) \ .$$

4.2.2. Illustration du modèle où α varie et $\beta = \frac{1}{2}$

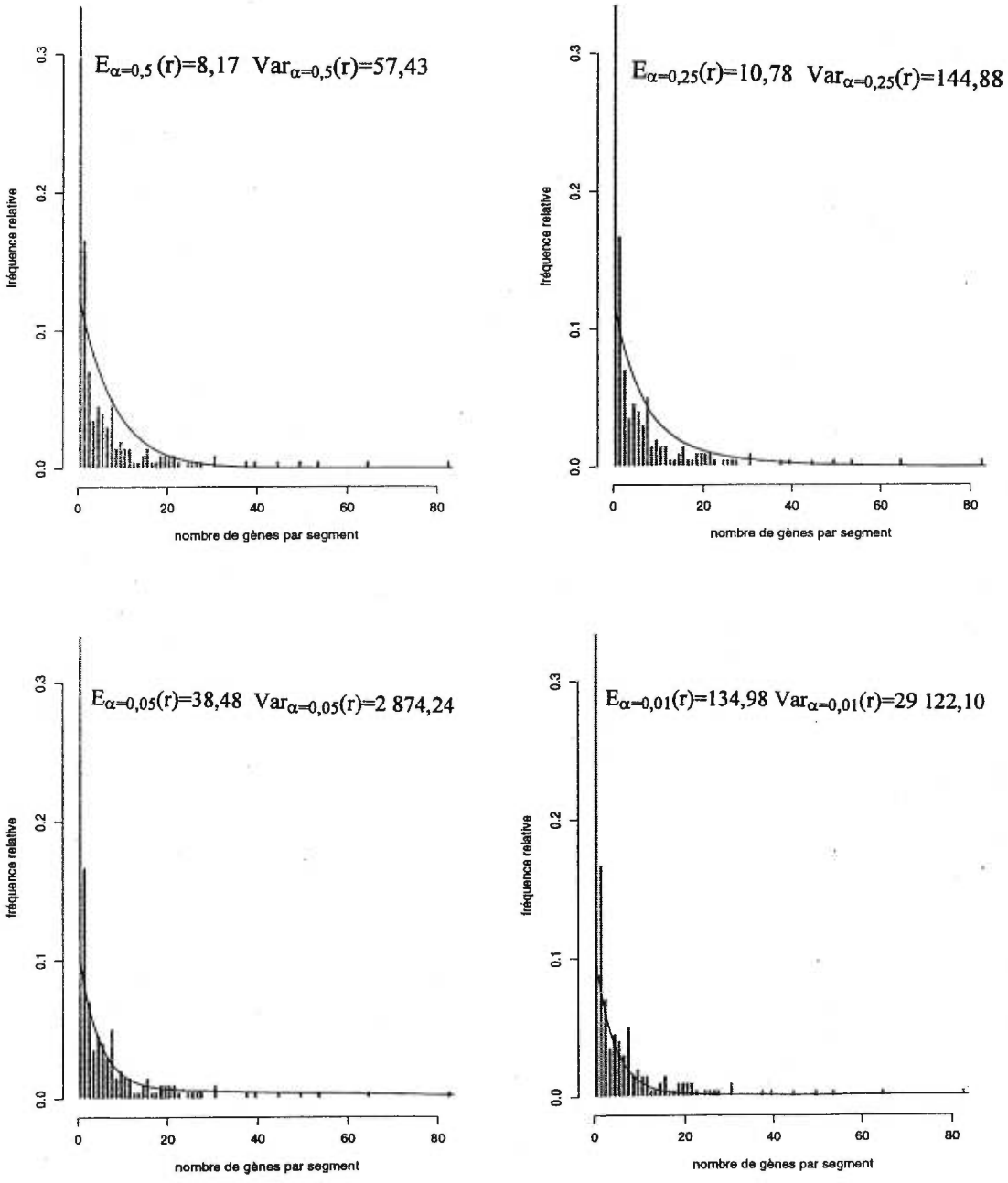
On regarde maintenant le modèle de distribution des gènes où la moitié des gènes est distribuée uniformément sur chacune des parties A et B et où le nombre de segments varie sur chacune des sections selon le paramètre α. On illustre l'ajustement du modèle de la proposition 4.2.1 pour α=0,5, 0,25, 0,05 et 0,01. Ceci toujours avec les mêmes données, soit m=1423 gènes et où le nombre de segments conservés estimé est 113, 197, 236 et 284.

FIGURE 4.16. Illustrations du modèle de distribution des gènes lorsque $\beta = \frac{1}{2}$ et α varie. Cas où il y a 1423 gènes et 113 segments



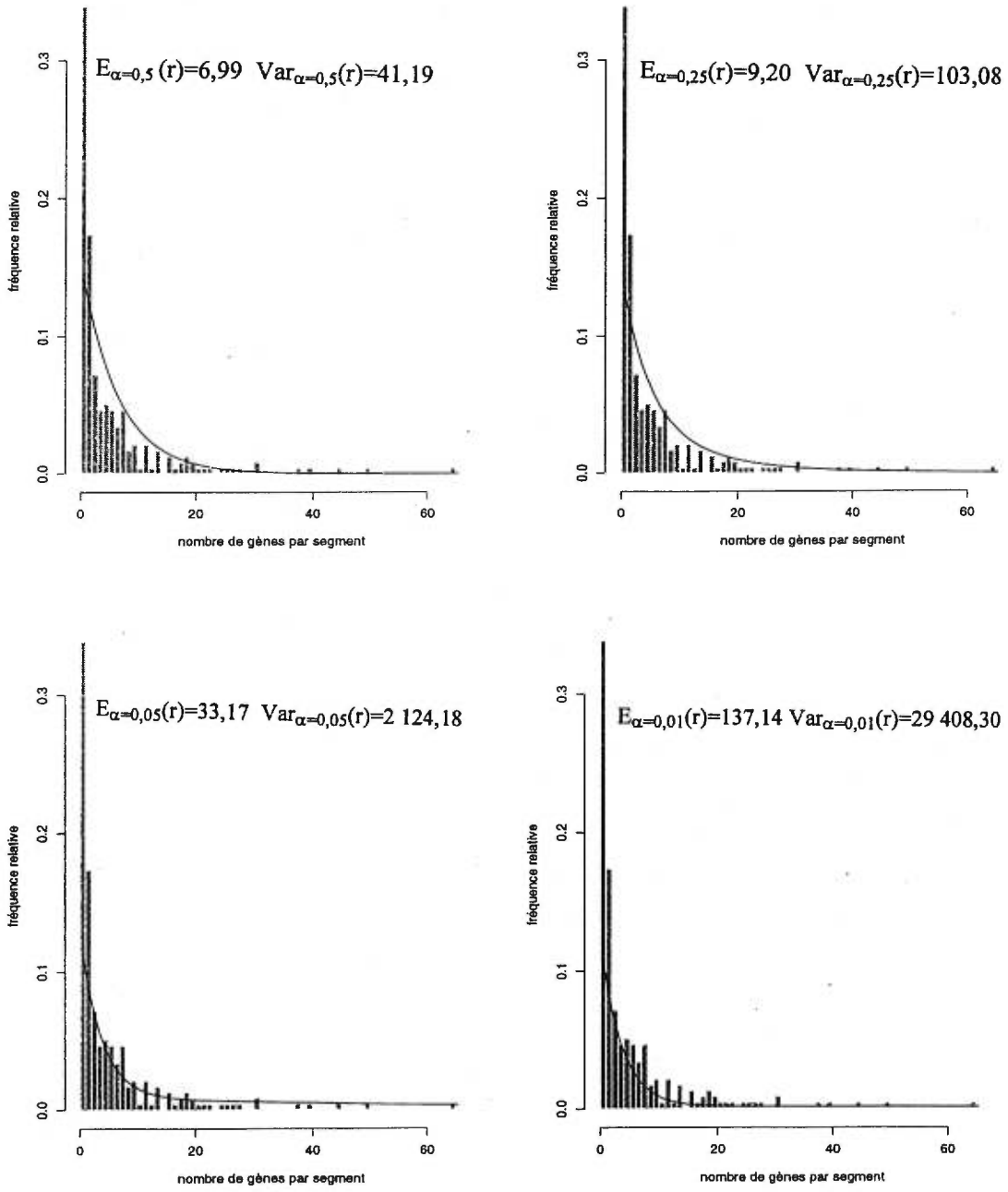
Nombre moyen de gènes sur un segment: 12,59
Variance du nombre de gènes sur un segment: 237,81

FIGURE 4.17. Illustrations du modèle de distribution des gènes lorsque $\beta = \frac{1}{2}$ et α varie. Cas où il y a 1423 gènes et 197 segments



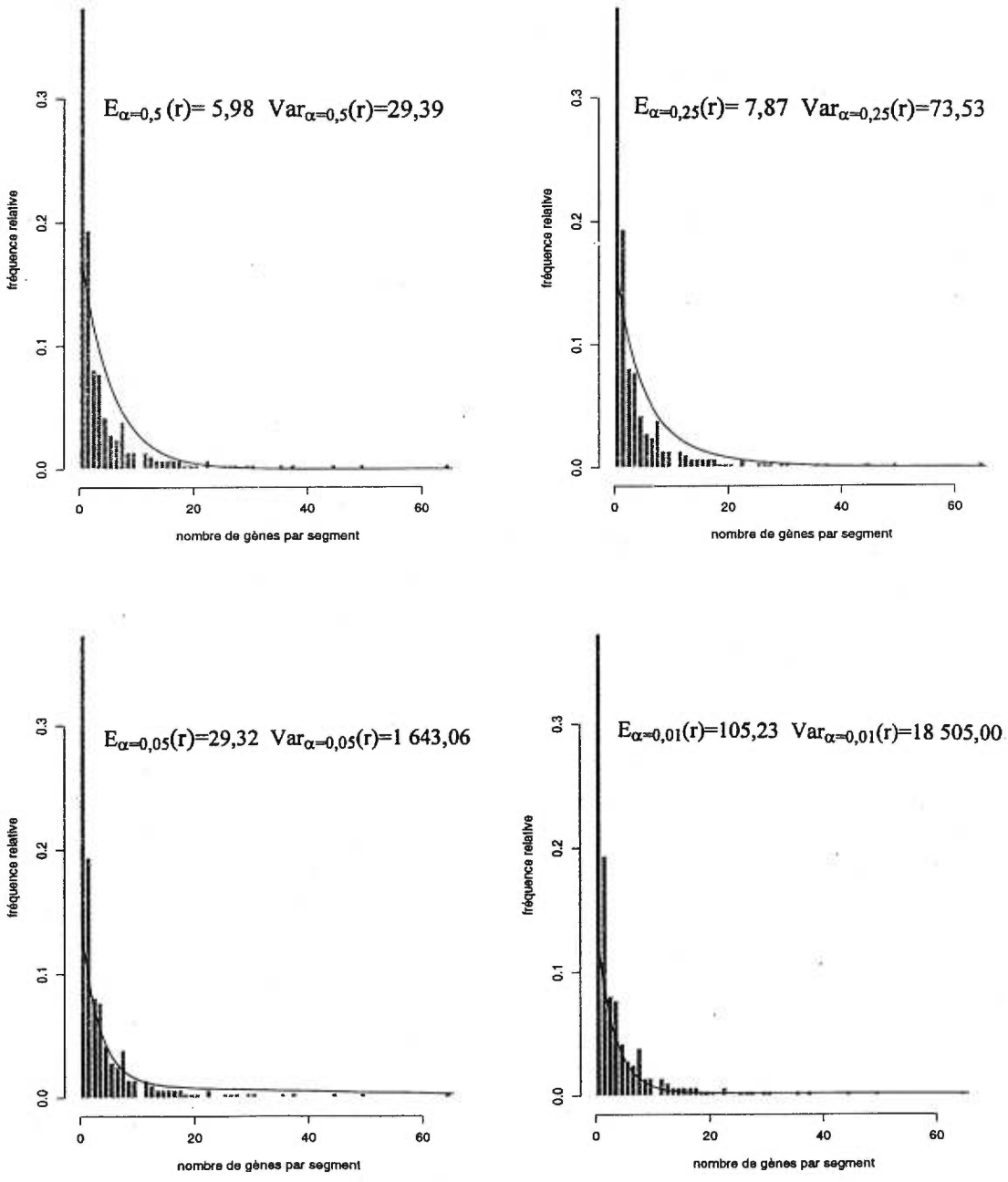
Nombre moyen de gènes sur un segment: 7,22
Variance du nombre de gènes sur un segment: 129,97

FIGURE 4.18. Illustrations du modèle de distribution des gènes lorsque $\beta = \frac{1}{2}$ et α varie. Cas où il y a 1423 gènes et 236 segments



Nombre moyen de gènes sur un segment: 6,02
Variance du nombre de gènes sur un segment: 76,49

FIGURE 4.19. Illustrations du modèle de distribution des gènes lorsque $\beta = \frac{1}{2}$ et α varie. Cas où il y a 1423 gènes et 284 segments



Nombre moyen de gènes sur un segment: 5,01
Variance du nombre de gènes sur un segment: 62,11

4.2.3. Commentaires

Le modèle de la proposition 4.2.1 avec $\beta = \frac{1}{2}$ et $\alpha = 0,05$ présente le meilleur ajustement des données obtenu. En séparant le génome en deux parties et en distribuant uniformément la moitié des gènes identifiés communs à l'homme et à la souris sur chaque section, en distribuant par la suite 5% des points de rupture sur une partie et le reste sur l'autre, on obtient une nette diminution de la différence entre la prévision de la fréquence relative du nombre de gènes sur un segment conservé et les données observées, surtout pour des valeurs de r= 3,...,15, qui étaient le principal problème du modèle de base.

Les valeurs de l'espérance et la variance surestiment considérablement les valeurs observées, pour la même raison évoquée à la section 4.1.3. Le modèle diminue l'importance accordée aux petites valeurs de r et augmente celle des r plus grand. Puisque l'on observe peu de segments conservés chez l'homme et la souris qui contiennent 30 gènes ou plus, on obtient un grand écart entre le nombre moyen de gènes sur un segment et son estimation de même qu'entre la variance du nombre de gènes sur un segment et son estimation.

On peut estimer β itérativement (pour α fixé) et α (pour β fixé) pour que les données soient le mieux approximées par la courbe théorique, en se servant, par exemple, d'un crière tel que Kolmogorov-Smirnov. Voir [8] pour les détails.

CONCLUSION

Nadeau et Taylor ont montré qu'avec des méthodes statistiques et de modélisation on peut déduire beaucoup d'informations concernant le génome à partir de relativement peu de données.

Une approche axiomatique mène à une meilleure compréhension du modèle et est plus élégante. Trois approches visant à améliorer l'ajustement du modèle aux données ont été considérés: modèles sans les segments contenant peu de gènes, modèle de distribution des gènes par grappes et un modèle sans l'hypothèse d'homogénéité de distribution des gènes et des points de rupture.

La modélisation de la fréquence relative du nombre de gènes sur un segment chromosomique conservé n'est pas positivement influencée par l'hypothèse d'identification des gènes causée par leur proximité à un ou d'autres gènes plus facilement identifiables et est légèrement influencée par les erreurs expérimentales concernant l'identification des segments qui contiennent peu de gènes. La principale amélioration du modèle provient de l'hypothèse de la non-uniformité de la distribution des points de rupture qui surviennent lors d'échanges entre les chromosomes, plus que de l'hypothèse de la non-uniformité de l'identification des gènes communs à l'homme et à la souris à travers le génome.

Une piste possible pour le futur serait d'explorer la littérature biologique pour savoir si on peut identifier des régions de chromosomes plus susceptibles à subir plus ou moins de ruptures.

APPENDICE A

LEXIQUE [7]

autosome: Dans la garniture chromosomique normale d'une espèce, chromosome autre que sexuel.(X,Y)

carte génétique: Représentation graphique de l'arrangement des gènes, leur ordre et leurs distances relatives sur un chromosome.

centiMorgan: Unité exprimant le pourcentage de la réalisation d'enjambement entre deux locus d'une même paire chromosomique. Ex: un taux de recombinaison de 1% entre deux locus correspond à une distance de 1 centiMorgan.

chromosome: Dans le noyau eucaryote, complexe nucléo- protéique formé d'ADN.

chromosome X: Dans la garniture chromosomique normale d'une espèce, chromosome intervenant dans le déterminisme du sexe de l'individu.

espèce: Chez les organismes à reproduction sexuée, ensemble des individus capables de s'interféconder et dont les produits sont fertiles.

évolution: Série de changements affectant au cours du temps toute structure ou organisme vivant sous l'influence de facteurs multiples.

fission: Phénomène par lequel une structure se sépare en deux éléments ou plus.

fusion: Phénomène par lequel deux éléments ou plus s'unissent en une structure unique.

gène: Séquence nucléotidique constituant une unité d'information génétique et pouvant déterminer l'expression d'un caractère.

génome: Ensemble de gènes présents dans un organisme unicellulaire ou dans les cellules d'un organisme pluricellulaire.

inversion: Processus selon lequel un segment de chromosome se place en sens inverse sans modification de sa séquence.

liaison génétique: Association de gènes due à leur présence sur le même chromosome.

locus: Sur un chromosome, emplacement occupé par un gène.

mutation chromosomique: Modification spontanée ou provoquée affectant la structure et/ou le nombre d'un ou de plusieurs chromosomes.

recombinaison génétique: Formation de nouvelles combinaisons de gènes.

translocation: Remaniement chromosomique au cours duquel un segment de chromosome change de place.

translocation réciproque: L'échange de deux segments entre deux chromosomes.

transposition: Translocation à l'intérieur d'un chromosome.

RÉFÉRENCES

[1] NADEAU J.H., TAYLOR B.A., *Lengths of chromosomal segments conserved since divergence of man and mouse.* Proceedings of the National Academy of Sciences USA, 81:814-818, 1984.

[2] SANKOFF D., NADEAU J.H., *Conserved synteny as a measure of genomic distance.* Discrete Applied Mathematics (in press).

[3] RIORDAN J., Combinatorial identities. Wiley series in probability and mathematical statistics, 1968.

[4] ROSS S.M., Initiations aux probabilités. Presse polytechniques et universitaires romandes, 1990.

[5] SANKOFF D., FERRETTI V., NADEAU J.H., *Conserved segment identification.* Proceedings of the first annual conference on research in computational molecular biology. ACM Press, 1997.

[6] CASELLA G., BERGER R.L., Statistical inference. Duxbury Press. Belmont, Californie, 1990.

[7] SOURNIA J.-C. avec l'aide de l'Agence de coopération culturelle et technique (ACCT), Dictionnaire de génétique. Conseil international de la langue française, Fondation postuniversitaire interculturelle, Paris, 1991.

[8] SANKOFF D., PARENT M.-N., MARCHAND I., FERRETI V., *On the Nadeau-Taylor theory of conserved chromosome segments.* Combinatorial Pattern Matching, 1264, Springer, 1997.