Approche probabiliste du séquençage à grande échelle

Josselin Garnier[1]

Version préliminaire - 4 Juillet 1997

[1] Centre de Mathematiques Appliquees (CMAP) Ecole Polytechnique - CNRS URA756 - 91128 Palaiseau Cedex
Tel : 01.69.33.46.30 Fax : 01.69.33.30.11 E-mail : garnier@cmapx.polytechnique.fr

Partie I

Comparaisons de stratégies de séquençage

1 Introduction

Cette première approche a pour but de présenter le type de résultats qu'on peut obtenir par une approche probabiliste du problème, qui repose essentiellement sur une analyse statistique dans le cadre asymptotique où la longueur de génome est nettement plus grande que toutes les autres échelles caractéristiques. Afin de simplifier cette étude préliminaire, on a négligé toutes les sources d'erreurs. On propose donc dans ce document des résultats théoriques qui décrivent les distributions statistiques des quantités intéressantes : les nombres et tailles des trous, des contigs de séquences et de clones, la longueur du plus grand contig de clones et plus particulièrement l'"instant" où le contig maximal recouvre entièrement le génome. On va commencer par examiner dans la Section 2 la stratégie de séquençage la plus simple, à savoir le shot gun simple. On présentera ensuite dans la Section 3 des résultats encore partiels sur le séquençage par paires d'extrémités.

Avertissement : Ce rapport est plus un document de travail utile pour la suite qu'un objet fini. C'est pourquoi la présentation est encore un peu sèche.

2 Shot gun simple

2.1 Formulation du problème

Soit un génome **G** composé de l'enchaînement linéaire de bases, éléments d'un alphabet de 4 lettres. La longueur de ce génome est notée G.

On veut étudier le recouvrement de **G** par une collection **B** de B séquences S_i. Une séquence est un segment de longueur d pris dans $[1, ..., G]$. Les B séquences de la banque **B** sont tirées indépendamment, si bien que leur description statistique peut être modélisée de la façon suivante:

$$S_i = \{\tau_i, ..., \tau_i + d - 1\},$$

où les $(\tau_i)_{i=1,...,B}$ suivent des lois uniformes sur $\{1, ..., G - d\}$.

Plutôt que de travailler avec B, on va utiliser R_s, appelée redondance de séquence définie par:

$$R_s = \frac{B.d}{G}.$$

Quand plusieurs séquences s'enchaînent par chevauchement, elles forment un contig de séquence. Il y a NC_s contigs de séquence, NC_s étant une variable aléatoire à valeurs dans $\{1, ..., B\}$.

On veut déterminer la distribution statistique de NC_s, ainsi que celle de la longueur des contigs de séquence, celle des trous qui séparent les contigs de séquence, et aussi la longueur du plus long contig de séquence. On va développer dans la suite une étude statistique asymptotique dans le cadre $G \gg d$.

2.2 Résultats

Le recouvrement du génome **G** par la banque **B** est représentée comme sur la figure 1:
• Le segment $[X_{2n+1}, X_{2n+2}]$, $n = 1, .., B$ représente la position du contig de séquence n$^o n$. Sa longueur est $LC_{sn} = X_{2n+2} - X_{2n+1}$.
• Le segment $[X_{2n+2}, X_{2n+3}]$, $n = 1, .., B - 1$ représente le trou qui sépare le contig de séquence n$^o n$ du contig de séquence n$^o n + 1$. Sa longueur est $LT_{sn} = X_{2n+3} - X_{2n+2}$.

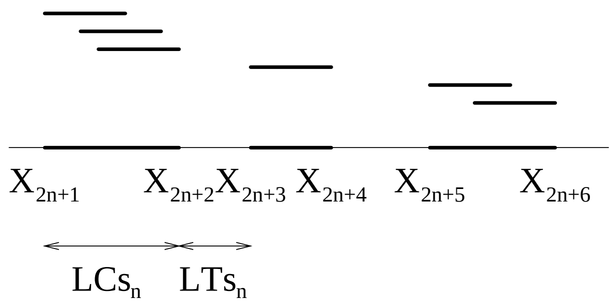

Figure 1 : *Représentation du recouvrement d'un fragment du génome* **G**

2.2.1 Taux de recouvrement

Le pourcentage p_k de bases qui sont représentées k fois (avec $k = 0, 1, 2, ...$) dans la banque **B** est donné par :

$$p_k = \frac{R_s^k}{k!} e^{-R_s}. \tag{1}$$

Ce résultat est aussi valable dans le cas du séquençage par paires d'extrémités, quelle que soit la longueur des clones. En effet le fait que les taux p_k suivent une loi dite de Poisson de paramètre R_s ne dépend que de l'approche asymptotique $G \gg d$ et surtout de la loi *uniforme* adoptée comme modèle de tirage aléatoire de la position des clones.

2.2.2 Nombre de contigs de séquence

Le nombre NC_s de contigs de séquence est donné par :

$$NC_s = \frac{GR_s e^{-R_s}}{d} + \sqrt{\frac{GR_s e^{-R_s}(1 - e^{-R_s})}{d}} W, \tag{2}$$

où W suit une distribution normale de densité $p(w) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2}$.

Dans l'asymptotique $G \gg d$, le premier terme est nettement dominant. En fait, le second terme décrit les fluctuations autour de la valeur centrale $GR_s e^{-R_s}/d$.

2.2.3 Statistique des tailles des contigs de séquence

Les longueurs $(LC_{sj})_{j=1,..,NC_s}$ des NC_s contigs de séquence sont des variables indépendantes et identiquement distribuées, dont la loi LC_s est décrite par la transformée de Laplace :

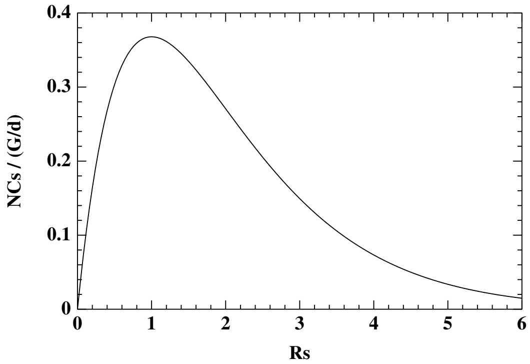

Figure 2 : *Nombre moyen de contigs de séquence*

$$\left\langle e^{-\lambda LC_s} \right\rangle = \frac{R_s + \lambda d}{R_s + \lambda d e^{R_s + \lambda d}}. \tag{3}$$

Remarque : La transformée de Laplace d'une variable aléatoire C à valeurs dans $\mathbb{R}^+$ est la fonction $\lambda \mapsto \left\langle e^{-\lambda C} \right\rangle$. Elle caractérise entièrement la distribution statistique de la variable en question, et on l'utilise ici car elle est nettement plus simple à écrire de manière analytique que la densité de probabilité ou la fonction de répartition, comme la suite le montre.

La longueur moyenne d'un contig de séquence est :

$$\langle LC_s \rangle = d \frac{e^{R_s} - 1}{R_s}. \tag{4}$$

La variance autour de la valeur moyenne (qui décrit l'amplitude des fluctuations) est :

$$\left\langle (LC_s - \langle LC_s \rangle)^2 \right\rangle^{1/2} = d \frac{\sqrt{e^{2R_s} - 2R_s e^{R_s} - 1}}{R_s}. \tag{5}$$

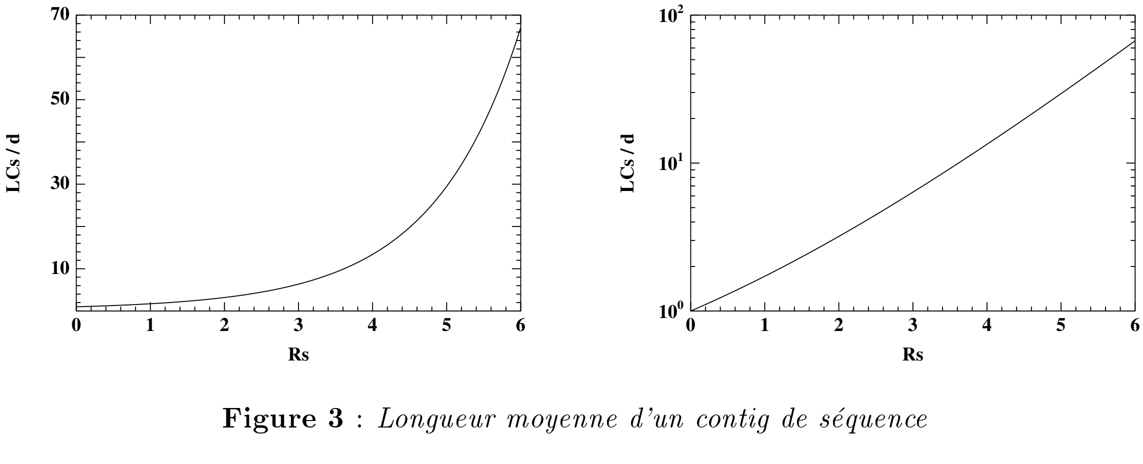

Figure 3 : *Longueur moyenne d'un contig de séquence*

La densité de probabilité de la distribution LC_s est de la forme :

$$p_{LC_s}(x) = e^{-R_s} \delta_d(x) + (1 - e^{-R_s}) p(x). \tag{6}$$

δ_d est une masse de Dirac en d et est reliée aux contigs de séquence constitués d'une unique séquence. p est une densité de probabilité sans masse ponctuelle qui modélise la statistique des contigs de séquence constitués du chevauchement d'au moins deux séquences. p est représentée pour différentes valeurs de R_s sur la figure 4.

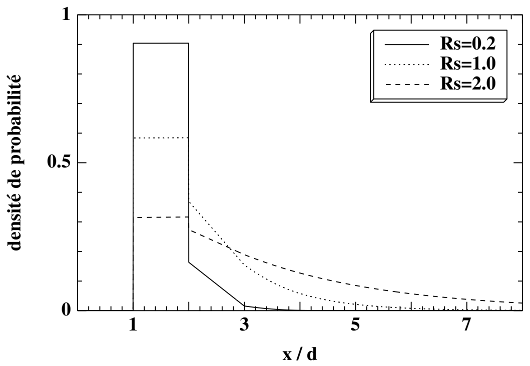

Figure 4 : *Distribution statistique de la longueur des contigs de séquence non-dégénérés*

On peut en outre déterminer la décroissance de la fonction de répartition de la longueur des contigs de séquence. Cette information est très utile pour estimer la longueur du plus long contig. En notant $\beta(R_s)$ l'unique réel positif tel que $R_s(e^{\beta-R_s}-1)/(\beta-R_s)=1$, on a :

$$\lim_{u\to\infty} e^{\beta(R_s)u/d}\mathbb{P}(LC_s > u) = \frac{\beta(R_s)-R_s}{R_s\beta(R_s)-R_s}. \tag{7}$$

2.2.4 Statistique des tailles des trous

Les longueurs $(LT_{sj})_{j=1,..,NC_s-1}$ des $NC_s - 1$ trous séparant les contigs de séquence sont des variables indépendantes et identiquement distribuées, indépendantes des $(LC_{sj})_{j=1,...,NC_s}$, dont la loi LT_s est une distribution exponentielle de densité:

$$p_{LT_s}(x) = e^{-R_s x/d}. \tag{8}$$

La longueur moyenne des trous est de :

$$\langle LT_s\rangle = \frac{d}{R_s}. \tag{9}$$

La variance autour de la valeur moyenne est :

$$\left\langle (LT_s - \langle LT_s\rangle)^2\right\rangle^{1/2} = \frac{d}{R_s}. \tag{10}$$

2.2.5 Longueur du plus long contig de séquence

On note Q_{max} la longueur du plus long contig de séquence. Tant que $Q_{max} < G$, cette longueur est asymptotiquement donnée par :

$$Q_{max} = \frac{d}{\beta(R_s)}\left(\ln(G/d) + \ln\frac{\beta(R_s)-R_s}{\beta(R_s)-1} - \ln R_s - \ln Z\right), \tag{11}$$

où $1/\beta(R_s)$ est une fonction à croissance exponentielle $\sim e^{R_s}$ (voir le paragraphe précédent) et Z est une variable aléatoire qui suit une loi exponentielle de paramètre 1, dont la densité est $p(z) = e^{-z}\mathbb{1}_{z>0}$. Sur la figure 5 est tracée la valeur moyenne de Q_{max}.

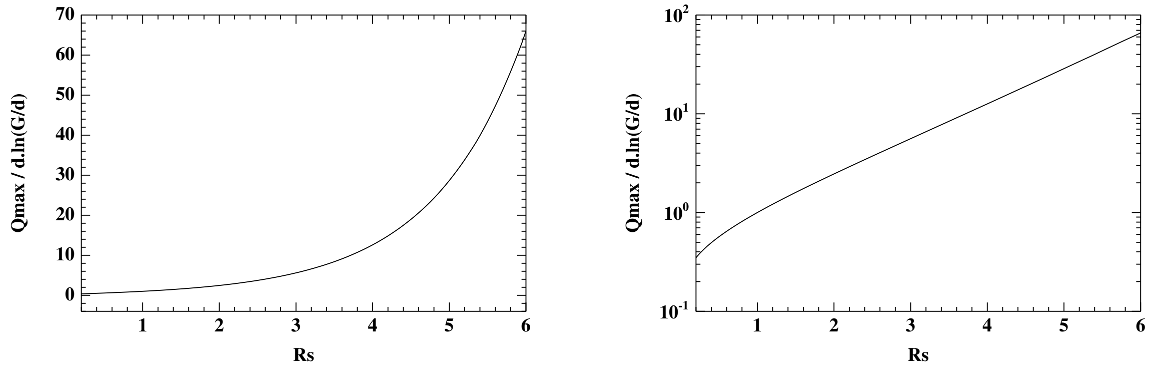

Figure 5 : *Longueur Q_{max} moyenne du plus long contig de séquence*

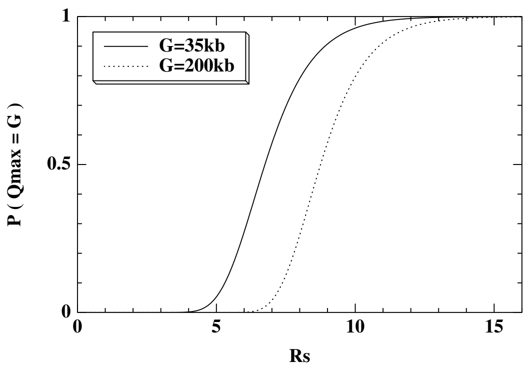

Figure 6 : *Probabilité pour que le génome* **G** *soit entièrement recouvert. On a adopté la valeur $d = 400b$.*

D'autre part, on peut estimer la valeur de redondance R_s pour laquelle Q_{max} va être égal à G, à quelques bases près situées aux bords du génome. Cette estimation s'effectue par l'intermédiaire de la probabilité suivante (voir figure 6) :

$$\mathbb{P}\left(Q_{max} \simeq G\right) = \exp\left(-\frac{G R_s e^{-R_s}}{d}\right). \tag{12}$$

Il apparaît qu'il faut avoir des redondances très importantes pour espérer pouvoir recouvrir le génome. Si on regarde sur la figure 7 le pourcentage de bases qui sont représentées plus de deux fois aux valeurs de redondance pour lesquelles on a recouvrement total, on voit nettement qu'il y a un grand gâchis.

2.2.6 Longueur du plus long trou

On note LT_{max} la longueur du plus long trou. Cette longueur est asymptotiquement donnée par :

$$LT_{max} = \frac{d}{R_s}\Big(\ln(G/d) + \ln R_s e^{-R_s} - \ln Z\Big), \tag{13}$$

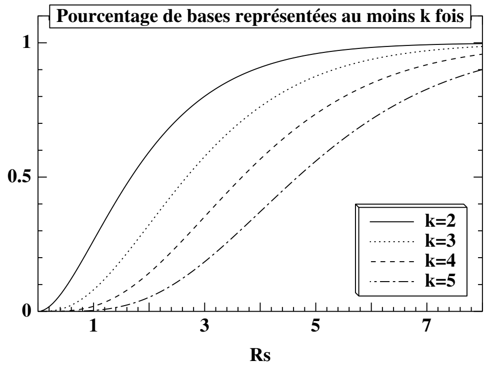

Figure 7 : *Pourcentage de bases représentées plus de deux fois dans la banque* **B**.

où Z est une variable aléatoire qui suit une loi exponentielle de paramètre 1, dont la densité est $p(z) = e^{-z}\mathbb{1}_{z>0}$. La valeur moyenne de $-\ln Z$ est $-\langle \ln Z \rangle = C$, où $C \simeq 0.577$ est connue comme étant la constante d'Euler, et sa variance est $\left\langle (\ln Z - \langle \ln Z \rangle)^2 \right\rangle^{1/2} = \pi/\sqrt{6} \simeq 1.28$. Sur la figure 8 est tracée la valeur moyenne de LT_{max}.

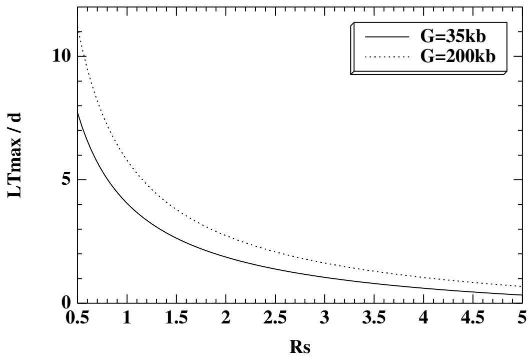

Figure 8 : *Valeur moyenne de la taille du plus grand trou*

En réalité on peut donner la distribution des n plus grands trous $LT_{max1} > LT_{max2} > ... > LT_{maxn}$:

$$LT_{maxj} = \frac{d}{R_s}\left(\ln(G/d) + \ln R_s e^{-R_s} - \ln(Z_1 + ... + Z_j)\right), \tag{14}$$

où les $(Z_j)_{j=1,...,n}$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi exponentielle de paramètre 1. En particulier, $-\langle \ln(Z_1 + \ldots + Z_n)\rangle = C - \sum_{j=1}^{n-1} j^{-1}$. Sur les figures 9 et 10sont tracées les probabilités de trouver des trous de taille donnée dans différentes configurations.

2.2.7 Détection exhaustive d'une séquence particulière

Soit **H** une séquence de longueur H qui a un homologue dans le génome **G**. La probabilité de trouver cet homologue de manière complète dans un des contigs de séquence de la banque **B** est :

$$\mathbb{P}_1 = 1 - e^{-R_s} - \frac{R_s e^{-R_s} H}{d} + \frac{R_s e^{-R_s}}{d}\int_0^H du p_{LC_s}(u)(H-u), \tag{15}$$

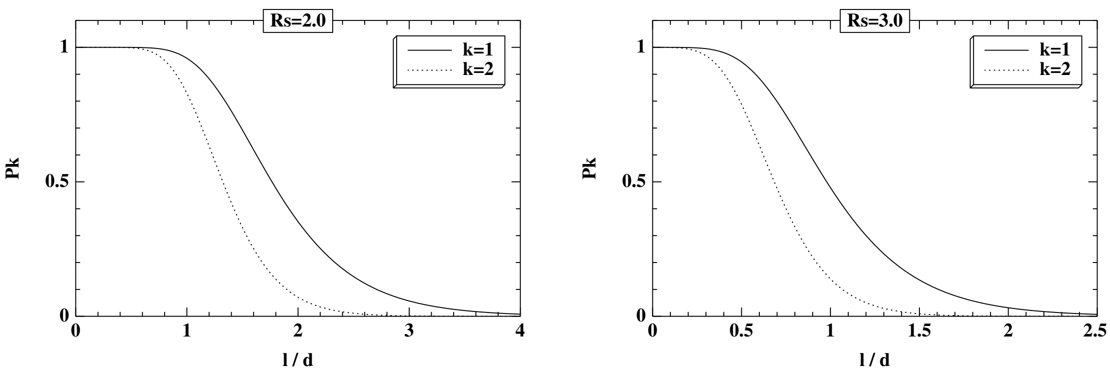

Figure 9 : *Probabilité Pk pour qu'il existe k trous de tailles plus grandes que l. Configuration* $G = 35kb$ *et* $d = 400b$

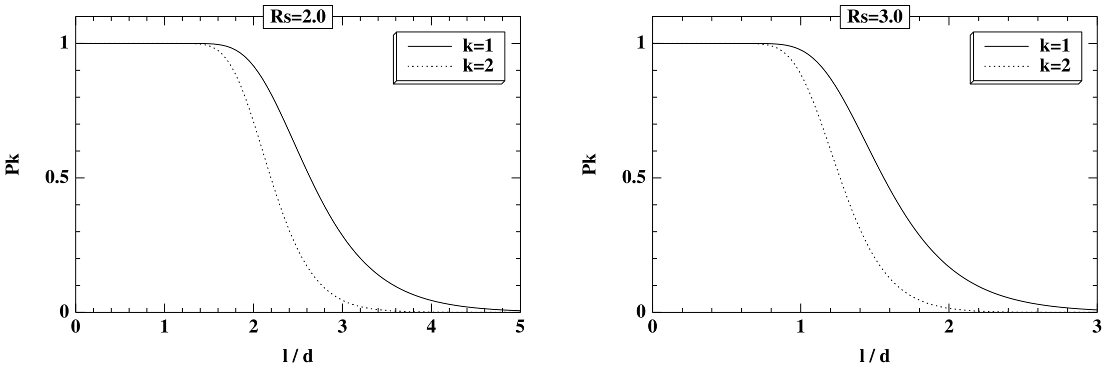

Figure 10 : *Probabilité Pk pour qu'il existe k trous de tailles plus grandes que l. Configuration* $G = 200kb$ *et* $d = 400b$

où p_{LC_s} est la densité de probabilité de la distribution statistique de la longueur des contigs de séquence. En particulier, si $H \leq d$, alors :

$$\mathbb{P}_1 = 1 - e^{-R_s} - \frac{R_s e^{-R_s} H}{d},$$

et si $d \leq H \leq 2d$, alors :

$$\mathbb{P}_1 = 1 - e^{-R_s} - \frac{R_s e^{-R_s} H}{d} + e^{-2R_s} \left(\frac{R_s(H-d)}{d} + \frac{1}{2} \left(\frac{R_s(H-d)}{d} \right)^2 \right).$$

En fait, ces résultats sont exacts dans le cas où l'homologue de **H** est unique dans le génome **G**. Dans le cas où il y a k homologues de **H** dans **G**, la probabilité d'en détecter $j \in \{0, ..., k\}$ est :

$$\mathbb{P}_{j,k} = C_k^j \mathbb{P}_1^j (1 - \mathbb{P}_1)^{k-j}.$$

2.2.8 Détection partielle d'une séquence particulière

Soit **H** une séquence de longueur H qui a un homologue dans le génome **G**. On considère ici qu'on a détecté cette séquence si on trouve I bases consécutives de **H** dans un des contigs de

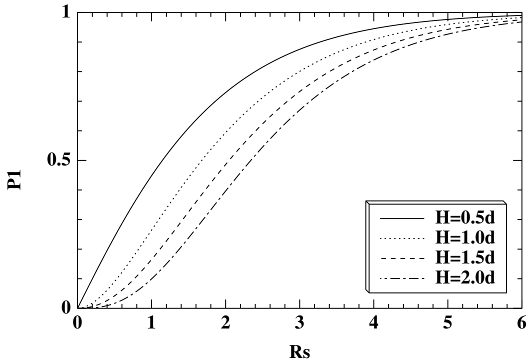

Figure 11 : *Probabilité de trouver une séquence* **H** *de longueur H dans une banque* **B** *pour différentes valeurs de H. On suppose ici que la séquence* **H** *se trouve une et une seule fois dans le génome* **G**

séquence de la banque **B**. Les résultats sont alors les suivants :
Si $I \leq H/2$ et $I \leq d$, alors

$$\mathbb{P}_1 = 1 - e^{-R_s\left(1+\frac{H-2I}{d}\right)}. \tag{16}$$

Si $I \geq H/2$ et $I \leq d$, alors

$$\mathbb{P}_1 = 1 - e^{-R_s} - R_s e^{-R_s}\left(\frac{2I-H}{d}\right). \tag{17}$$

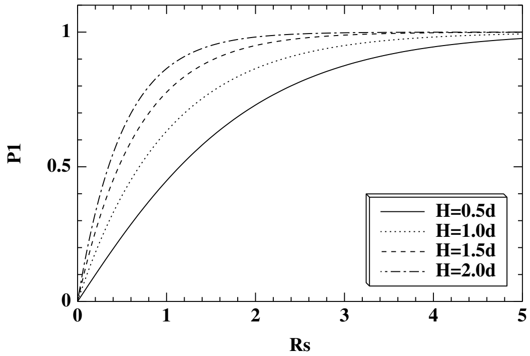

Figure 12 : *Probabilité de détecter partiellement une séquence* **H** *de longueur H dans une banque* **B** *pour différentes valeurs de H. On suppose ici que la séquence* **H** *se trouve une et une seule fois dans le génome* **G**, *et on adopte la valeur $I = 0.5d$ comme longueur critique de détection*

3 Séquençage par paires d'extrémités

3.1 Formulation du problème

Soit un génome **G** composé de l'enchaînement linéaire de bases, éléments d'un alphabet de 4 lettres. La longueur de ce génome est notée G.

On veut étudier le recouvrement de **G** par une collection **B** de B paires de séquences $S_i = (S_{ig}, S_{id})$. Une paire de séquences est une paire de segments de longueur d pris dans $[1, ..., G]$ séparés d'un intervalle $L-d$, avec $L > 2d$. Les B paires de séquences de la banque **B** sont tirées indépendamment, si bien que leur description statistique peut être modélisée de la façon suivante:

$$S_i = \{\tau_i, ..., \tau_i + d - 1\} \bigcup \{\tau_i + L - d, ..., \tau_i + L - 1\},$$

où les $(\tau_i)_{i=1,...,B}$ suivent des lois uniformes sur $\{1, ..., G-L\}$.

Plutôt que de travailler avec B et L, on va utiliser les coefficients R_s et R_m, appelés respectivement redondance de séquence et redondance de cartographie :

$$R_s = \frac{2.B.d}{G}, \qquad R_m = \frac{B.L}{G}.$$

Quand plusieurs séquences s'enchaînent par chevauchement, elles forment un contig de séquence. Il y a NC_s contigs de séquence, NC_s étant une variable aléatoire à valeurs dans $\{1, ..., B\}$.

Deux contigs de séquence dont l'un contient la séquence droite et l'autre la séquence gauche d'un même clone sont reliés physiquement entre eux. On appelle un groupe de contigs de séquence reliés de la sorte contig de clones. Il y a NC_c contigs de clones, NC_c étant une variable aléatoire à valeurs dans $\{1, ..., B\}$ plus petite que NC_s.

On veut déterminer la distribution statistique de NC_s et NC_c, ainsi que celle de la longueur des contigs de séquence et des contigs de clones, celle des trous qui séparent les contigs de séquence et de clones, et aussi la longueur du plus long contig de clones. On va développer dans la suite une étude statistique asymptotique dans le cadre $G \gg d$.

Le recouvrement du génome **G** par la banque **B** est représentée comme sur la figure 13 :
• Le segment $[X_{2n+1}, X_{2n+2}]$ représente la position du contig de séquence n$^o n$. Sa longueur est $LC_{sn} = X_{2n+2} - X_{2n+1}$.
• Le segment $[X_{2n+1}, X_{2n+6}]$ représente la position du contig de clones n$^o p$. Sa longueur est $LC_{cp} = X_{2n+6} - X_{2n+1}$.
• On distingue deux sortes de "trous" :
- les trous "internes" qui séparent deux contigs de séquence consécutifs à l'intérieur d'un contig de clones, du type du segment $[X_{2n+2}, X_{2n+3}]$. Sa longueur est $LT_{in} = X_{2n+3} - X_{2n+2}$.
- les trous "externes" qui séparent deux contigs de clone, du type du segment $[X_{2n+6}, X_{2n+7}]$. Sa longueur est $LT_{ep} = X_{2n+7} - X_{2n+6}$.

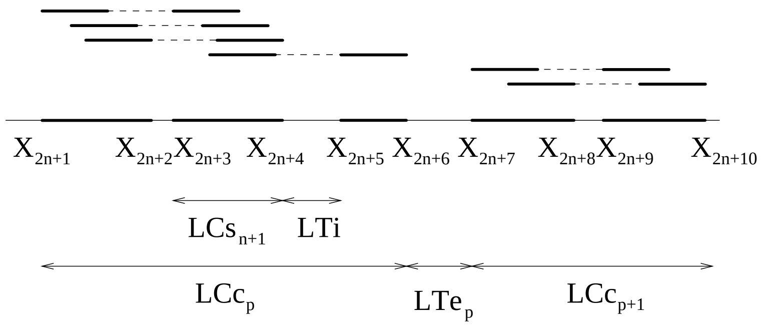

Figure 13 : *Représentation du recouvrement d'un fragment du génome* **G**

L'étude du séquençage par paires d'extrémités est nettement plus complexe que celle du shot gun simple. Elle se scinde en plusieurs cas, et on va commencer par présenter le premier d'entre eux, sans doute le plus simple, mais vraisemblablement le moins intéressant, à savoir le cas où la longueur L des clones est comprise entre $2d$ et $3d$. On examinera ensuite le cas asymptotique $G \gg L \gg d$.

3.2 Résultats dans le cas $L \in]2d, 3d]$

Le taux de recouvrement est identique à celui du shot gun simple, mais c'est le seul point commun entre les deux méthodes.

3.2.1 Nombre de contigs de séquence

La distribution statistique du nombre NC_s de contigs de séquence est fortement concentrée autour de sa moyenne :

$$\langle NC_s \rangle = \frac{GR_s e^{-R_s}}{2d} \left(1 + R_m - R_s + e^{-R_m + R_s}\right) \tag{18}$$

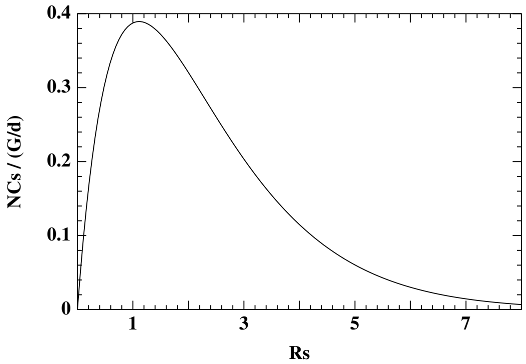

Figure 14 : *Nombre moyen de contigs de séquence. Cas $L = 3d$*

3.2.2 Nombre de contigs de clones

La distribution statistique du nombre NC_c de contigs de clones est fortement concentrée autour de sa moyenne :

$$\langle NC_c \rangle = \frac{GR_s}{2d} e^{-R_m} \tag{19}$$

3.2.3 Statistique des tailles des contigs de séquence

Les longueurs $(LC_{sj})_{j=1,..,NC_s}$ des NC_s contigs de séquence sont des variables indépendantes et identiquement distribuées. La longueur moyenne d'un contig de séquence est :

$$\langle LC_s \rangle = \frac{2d}{R_s} \frac{e^{R_s} - R_m + R_s - e^{-R_m + R_s}}{1 + R_m - R_s + e^{-R_m + R_s}}. \tag{20}$$

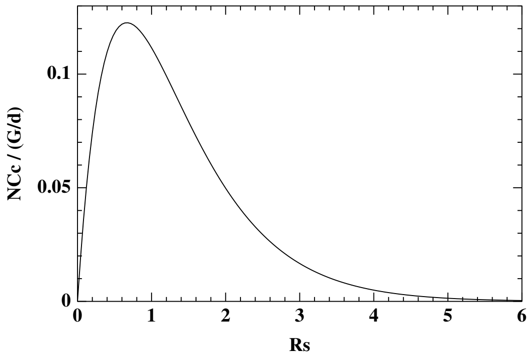

Figure 15 : *Nombre moyen de contigs de clones. Cas* $L = 3d$

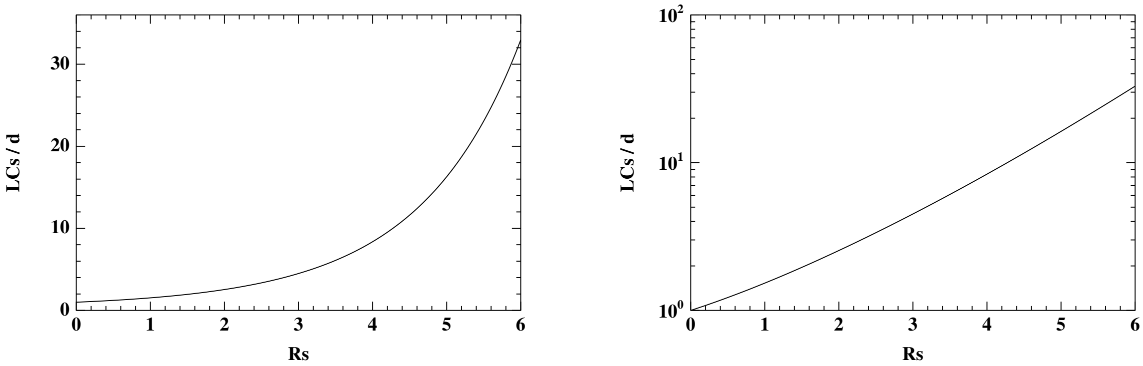

Figure 16 : *Longueur moyenne des contigs de séquence. Cas* $L = 3d$

3.2.4 Statistique des tailles des contigs de clones

Les longueurs $(LC_{cj})_{j=1,..,NC_c}$ des NC_c contigs de clones sont des variables indépendantes et identiquement distribuées. Leur valeur moyenne est :

$$\langle LC_c \rangle = \frac{2d}{R_s}\left(e^{R_m} - 1\right). \tag{21}$$

3.2.5 Statistique des tailles des trous internes (aux contigs de clone)

Le nombre $NT_i = NC_s - NC_c - 1$ des trous internes aux contigs de clones a pour valeur moyenne :

$$\langle NT_i \rangle = \frac{GR_s e^{-R_s}}{2d}\left(1 + R_m - R_s\right) \tag{22}$$

Les longueurs $(LT_{ij})_{j=1,..,NT_i}$ des NT_i trous internes aux contigs de clones sont des variables indépendantes et identiquement distribuées, indépendantes des autres variables, dont la loi LT_i admet la densité de probabilité suivante :

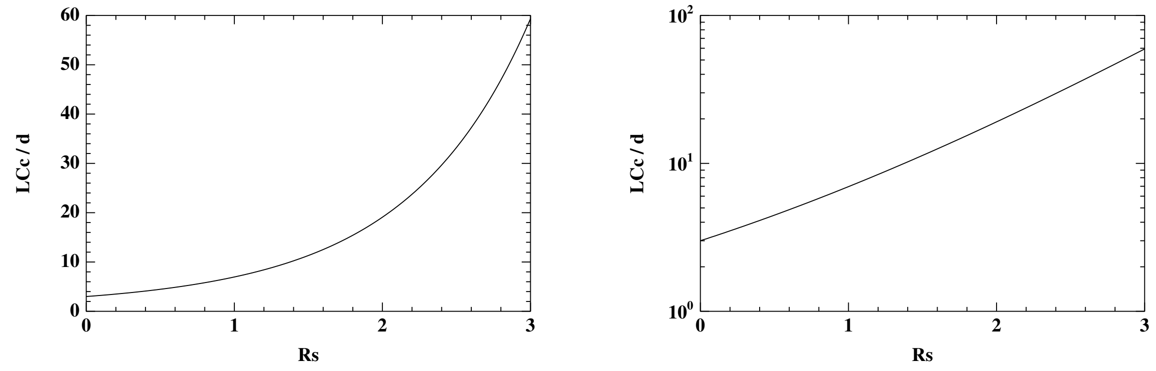

Figure 17 : *Longueur moyenne des contigs de clone. Cas* $L = 3d$

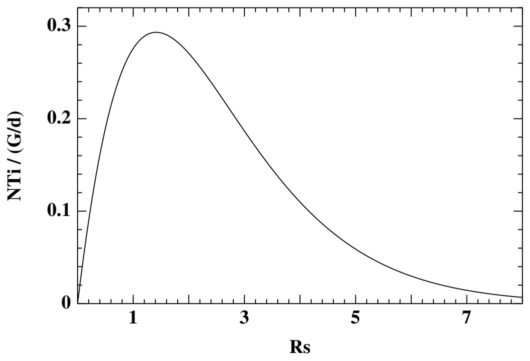

Figure 18 : *Nombre moyen des trous internes aux contigs de clones. Cas* $L = 3d$

$$\begin{aligned} p_{LT_i}(x) &= \frac{e^{-R_m+R_s}}{1+R_m-R_s}\delta_{L-2d} + \left(1 - \frac{e^{-R_m+R_s}}{1+R_m-R_s}\right)\bar{p}(x), \\ \bar{p}(x) &= \frac{\frac{R_s}{2d}e^{-\frac{R_s x}{2d}}\left(2+R_m-R_s-\frac{R_s x}{2d}\right)}{1+R_m-R_s-e^{-R_m+R_s}}\mathbb{1}_{x\in[0,L-2d]}. \end{aligned} \tag{23}$$

La longueur moyenne des trous internes est de :

$$\langle LT_i \rangle = \frac{L-2d}{1+R_m-R_s}. \tag{24}$$

3.2.6 Statistique des tailles des trous externes

Le nombre NT_e des trous externes séparant les contigs de clones est $NT_e = NC_c - 1$.
Les longueurs $(LT_{ej})_{j=1,..,NT_e}$ des NT_e trous externes sont des variables indépendantes et identiquement distribuées, dont la loi LT_e est une distribution exponentielle de densité :

$$p_{LT_e}(x) = e^{-\frac{R_s x}{2d}}. \tag{25}$$

La longueur moyenne des trous est de :

$$\langle LT_e \rangle = \frac{2d}{R_s}. \tag{26}$$

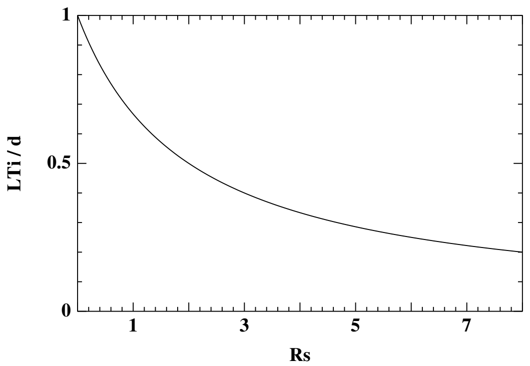

Figure 19 : *Longueur moyenne des tailles des trous internes. Cas* $L = 3d$

3.2.7 Longueur du plus long contig de clone

On peut estimer la valeur de redondance R_s pour laquelle Q_{max} va être égal à G, à quelques bases près situées aux bords du génome. Cette estimation s'effectue par l'intermédiaire de la probabilité suivante (voir figure 20) :

$$\mathbb{P}\left(Q_{max} \simeq G\right) = \exp\left(-\frac{GR_s}{2d}e^{-R_m}\right). \tag{27}$$

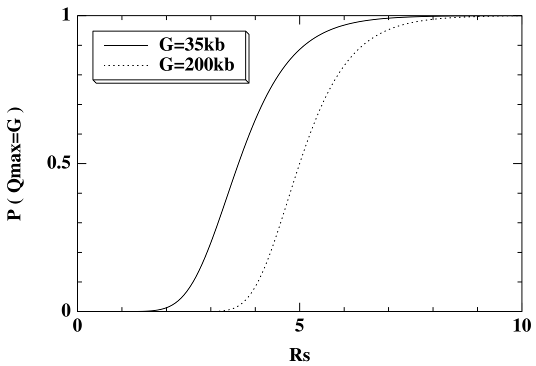

Figure 20 : *Probabilité pour que le génome* **G** *soit entièrement recouvert. On a adopté les valeurs* $d = 400b$ *et* $L = 1200b$.

3.2.8 Longueur du plus long trou externe

On note LT_{max} la longueur du plus long trou externe. Le plus grand trou interne est lui forcément plus petit que $L - 2d$. Cette longueur est asymptotiquement donnée par :

$$LT_{max} = \frac{2d}{R_s}\left(\ln\left(\frac{G}{d}\right) + \ln\left(\frac{R_s e^{-R_m}}{2}\right) - \ln Z\right), \tag{28}$$

où Z est une variable aléatoire qui suit une loi exponentielle de paramètre 1 (voir la section 2.2.6). Comme dans le cas du shot gun simple, on peut donner la distribution des n plus grands trous $LT_{max1} > LT_{max2} > ... > LT_{maxn}$:

$$LT_{max\,j} = \frac{2d}{R_s}\left(\ln\left(\frac{G}{d}\right) + \ln\left(\frac{R_s e^{-R_m}}{2}\right) - \ln(Z_1 + \ldots + Z_j)\right), \tag{29}$$

où les $(Z_j)_{j=1,\ldots,n}$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi exponentielle de paramètre 1. Sur les figures 21 et 22 sont tracées les probabilités de trouver des trous externes de taille donnée dans différentes configurations. Remarquez sur ces figures que la probabilité pour que le plus grand trou externe soit de taille nulle est positive, ce qui correspond au fait que tous les trous externes ont été bouchés, c'est-à-dire que Q_{max} est égal à G. Cette probabilité est donnée par (27), et peut aussi se déduire de (28).

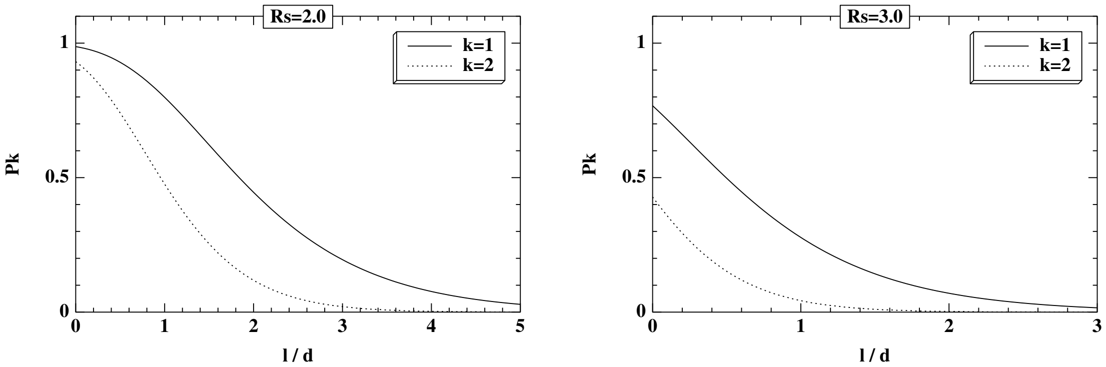

Figure 21 : *Probabilité Pk pour qu'il existe k trous de tailles plus grandes que l. Configuration $G = 35kb$, $L = 1200b$ et $d = 400b$*

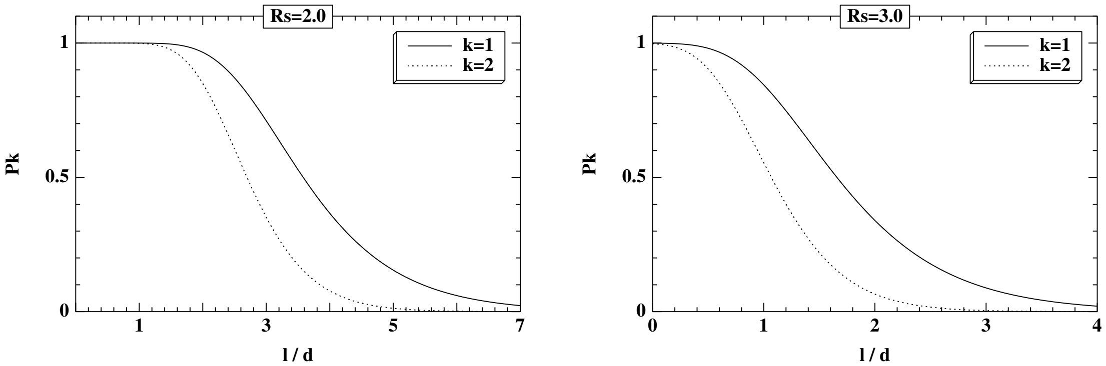

Figure 22 : *Probabilité Pk pour qu'il existe k trous de tailles plus grandes que l. Configuration $G = 200kb$, $L = 1200b$ et $d = 400b$*

3.3 Résultats pour des clones de taille $L \gg d$

On va ici développer une analyse asymptotique qui vaut dans le cadre $G \gg L \gg d$. On supposera en outre que la longueur des clones fluctue autour de la valeur L, et que ces fluctuations sont plus grandes que d. Il apparaît d'une part que la statistique des trous et des contigs de séquence est identique à celle du cas shot gun simple. D'autre part, l'instant où le génome est entièrement recouvert semble tendre vers une limite asymptotique correspondant à une valeur de la redondance de séquence égale à $\ln 2$.

3.3.1 Longueur du plus long contig de clone

On estime que le recouvrement total du génome est atteint lorsque l'extrémité droite (resp. gauche) du plus long contig de clones est située à moins d'une distance L du bord droit (resp. gauche) du génome.
On examine tout d'abord le problème suivant. Considérons une extrémité $[B_2, B_3]$ (de longueur d) d'un des clones. On appelle N le nombre d'extrémités d'autres clones auxquelles $[B_2, B_3]$ est connecté, et N_g le nombre d'extrémités gauches parmi les N extrémités. Sur la figure 23, on a $N = 6$ et $N_g = 4$.

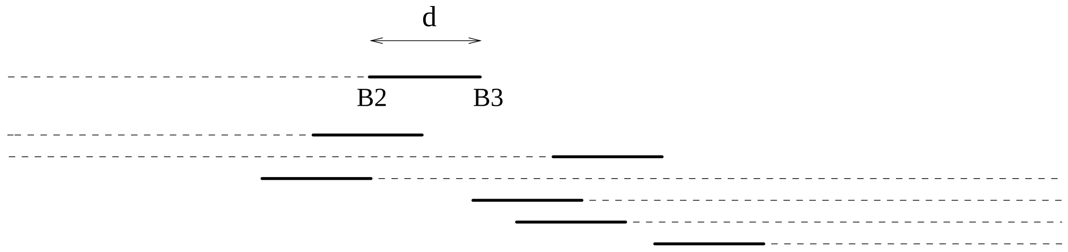

Figure 23 : *Connection d'une extrémité* $[B2, B3]$ *d'un clone avec d'autres clones*

On peut calculer les lois de N et N_g. On trouve en particulier que $\langle N \rangle = 2(e^{R_s} - 1)$ et

$$\left\langle u^{N_g} \right\rangle = 4\left((e^{R_s}+1) - (e^{R_s}-1)u\right)^{-2}, \qquad \langle N_g \rangle = e^{R_s} - 1.$$

La fonction génératrice $u \mapsto \left\langle u^{N_g} \right\rangle$ décrit entièrement la distribution statistique de N_g. Cette information nous permet de déduire une estimation de la valeur de R_s à partir de laquelle un contig maximal va exister. En effet, si $\langle N_g \rangle < 1$, alors les connections entre les clones s'éteignent très vite, de manière exponentielle en G/L. Au contraire, si $\langle N_g \rangle > 1$, alors des connections multiples entre les clones se développent. On trouve donc que la probabilité pour que le contig de clones le plus long recouvre $\mathbf{G}$ est quasi-nulle pour des valeurs de redondance de séquence strictement inférieures à $R_{slim} = \ln 2$ et très proche de 1 pour des valeurs de redondance de séquence strictement supérieures à R_{slim}.

3.3.2 Longueur du plus long contig de clone - Stratégie mixte

On suppose ici qu'on dispose d'une banque $\mathbf{B_2}$ constituée de B_2 clones de longueur moyenne $L \gg d$ et d'une banque $\mathbf{B_1}$ constituée de B_1 séquences simples de longueur d. On note R_{sj} la redondance de séquence associée à la banque $\mathbf{B_j}$, R_s la redondance totale et p la proportion de brins simples:

$$R_{s1} = \frac{B_1 d}{G}, \qquad R_{s2} = \frac{2B_2 d}{G}, \qquad R_s = R_{s1} + R_{s2}, \qquad p = \frac{R_{s1}}{R_s} = \frac{B_1}{B_1 + 2B_2}.$$

En posant le problème dans les mêmes termes que dans le paragraphe 3.3.1, on trouve que la valeur critique de la redondance totale à partir de laquelle il existe un contig de clone maximal est :

$$R_{slim} = \ln\left(\frac{2-p}{1-p}\right).$$

Bien sûr la valeur critique augmente avec p, car les brins simples sont moins efficaces que des clones longs pour générer un contig de clone maximal. Néanmoins, comme le montre la figure

24, une faible proportion de brins simples ne semble pas pénaliser de manière notable les chances de générer un contig maximal.

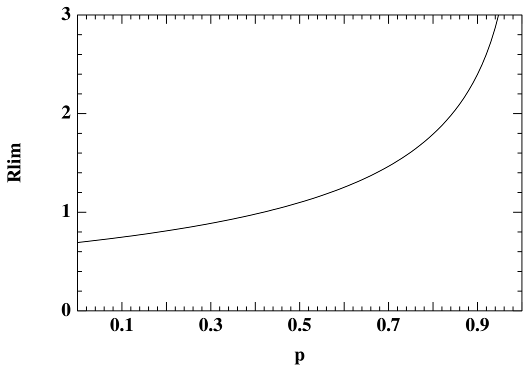

Figure 24 : *Valeur de la redondance totale critique en fonction de la proportion de brins simples*

En fait, les brins simples ne servent pas en eux-mêmes à établir un contig maximal, mais ils jouent un role non-négligeable en ce sens qu'ils permettent à des contigs de clones de longueur L d'être connectés, alors que leurs extrémités ne se chevauchent pas (voir la figure 25).

Figure 25 : *Role d'un brin simple dans la connection de deux clones longs*

Partie II

Résultats annexes

4 Détection d'une anomalie de représentation dans une banque de clones

Le problème considéré ici est le suivant. On considère une banque **B** de clones relative au génome **G** constituée de B clones de longueur d. On veut savoir s'il existe un biais dans la banque **B**, c'est-à-dire si les clones ont été tirés de manière uniforme sur l'ensemble du génome **G** ou bien si au contraire certaines régions du génome **G** sont sur ou sous-représentées.

On traite de manière informatique la banque en détectant des chevachements. On compte et on note :
- N_j, $j \geq 1$, le nombre de bases représentées exactement j fois dans la banque **B**,
- $\bar{N}_2$ le nombre de bases dont on a détecté qu'elles se chevauchaient. On a bien sûr $\bar{N}_2 = \sum_{j=2}^{\infty} N_j$.

Exemple 4.1
Soient les 4 clones suivants :

clone n^o 1 :	$GTCCGTCCA$
clone n^o 2 :	$TCCAAGCTT$
clone n^o 3 :	$AAGCTTGAACGT$
clone n^o 4 :	$GCATATTCG$
bases non-chevauchantes :	+++++ ++++++ +++++++++
bases chevauchantes :	+++++++++

Sur cet exemple, on a $N_1 = 20$ et $\bar{N}_2 = 9$.

Test avec connaissance de G.
Soit R l'unique réel positif solution de :

$$\frac{e^R - 1 - R}{R} = \frac{\bar{N}_2}{N_1}. \tag{30}$$

Le test de qualité de la banque **B** porte sur la comparaison de R avec sa valeur théorique qui est atteinte seulement pour une représentation uniforme :

$$R_{th} = \frac{Bd}{G}. \tag{31}$$

Si $R_{th} \simeq R$, alors la banque est de bonne qualité.
Si $R_{th} < R$, alors certaines régions sont sous-représentées par rapport à d'autres. Pour en dire un peu plus, il faut émettre des hypothèses a priori. Par exemple, adoptons l'ansatz qui consiste à dire que la représentation de la banque **B** est du type tout ou rien, c'est-à-dire qu'une partie du génome **G'** est représentée convenablement, et le reste pas du tout. Alors la zone représentée convenablement **G'** couvre une fraction R_{th}/R du génome total **G**. On peut même être un peu plus précis et donner un intervalle de confiance pour l'estimation de G'/G. On introduit tout

d'abord les quantités mesurées :

$$g_j = \frac{N_j}{d}, \quad j \geq 1, \qquad g_0 = \frac{G}{d} - \sum_{j=1}^{\infty} g_j. \tag{32}$$

Soient $k = \inf_j \{g_{j+1} \leq 5\}$ et ξ la fonction :

$$\xi : R \mapsto \sum_{j=0}^{k} \frac{(Gp_j(R)/d - g_j)^2}{g_j}, \tag{33}$$

où $p_j(R)$ est le taux de recouvrement d'ordre j théorique pour une redondance R :

$$p_j(R) = \frac{R^j}{j!} e^{-R}. \tag{34}$$

On peut alors, à l'aide d'un test du χ^2, donner un intervalle de confiance au niveau α pour la valeur de redondance R de la banque $\mathbf{B}$:

$$\mathrm{IC}(R) = [R_1, R_2], \text{ avec } R_1 = \inf \{R, \xi(R) \leq \chi_\alpha(k)\} \text{ et } R_2 = \sup \{R, \xi(R) \leq \chi_\alpha(k)\}, \tag{35}$$

où $\chi_\alpha(k)$ est la valeur du χ^2 à k degrés de liberté au niveau α. On peut alors en déduire un intervalle de confiance au niveau α pour G'/G :

$$\mathrm{IC}(G'/G) = \left[\frac{R_{th}}{R_2}, \frac{R_{th}}{R_1}\right]. \tag{36}$$

Remarques.
1. Un intervalle de confiance au niveau α pour une certaine quantité X signifie : La probabilité pour que X appartienne à cet intervalle est supérieure à α.
2. Ces résultats sont valides dans le cadre limite où la longueur minimale à partir de laquelle on décide de retenir un chevauchement est un paramètre négligé. Si on veut être très précis et prendre en compte ce paramètre, il faut remplacer les expressions (34) des $p_j(R)$ théoriques par leurs valeurs correctes, mais quelque peu compliquées (voir la section 6 et prendre $p_j = N_j/G$). Il faut aussi retirer le terme d'ordre 0 dans la fonction ξ définie par (33), et considérer un test du χ^2 à $k-1$ degrés de liberté dans (35).

Test sans connaissance de G.
Soient R_a et R_b les unique réels positifs solutions de :

$$\frac{e^{R_a} - 1 - R_a}{R_a} = \frac{\bar{N}_2}{N_1}, \qquad \frac{1 - e^{-R_b}}{R_b} = \frac{N_1 + \bar{N}_2}{Bd}. \tag{37}$$

Alors la banque $\mathbf{B}$ peut être estimée uniforme *sur son support* si $R_a \simeq R_b$. De plus on peut estimer la longueur de ce support par $G' = Bd/R_b$. Attention, ce test est plus approximatif que le premier. Il ne peut pas détecter une anomalie du type tout ou rien, c'est-à-dire que si une partie **G'** du génome **G** est représentée équitablement, et son complémentaire pas du tout, alors on n'a aucun moyen de le savoir, car le test ne raisonne que sur **G'** et trouve donc que la banque représente bien **G'**.

5 Lecture d'une séquence d'ADN

5.1 Lecture d'une séquence par une seule extrémité

On va donner ici une première approche du problème de la lecture d'un fragment d'ADN. On suppose qu'on veut séquencer un brin constitué de d bases. Pour nos calculs on admettra que ce brin d'ADN est une suite indépendante et equi-distribuée de bases A, C, G, T. Pour le séquençage, on va créer des clones dont le brin en question est à une extrémité (voir figure 26).

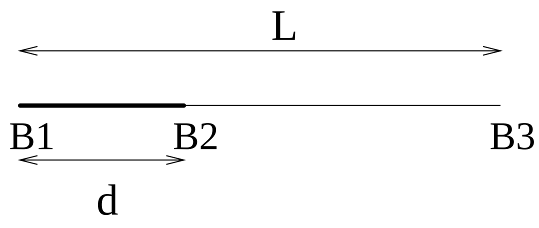

Figure 26 : *Brin à séquencer [B1,B2] situé sur le clone [B1,B3]*

Les clones sont répartis en quatre groupes, et chaque groupe est mis dans une solution qui contient des enzymes et des nucléotides libres A, C, G, T, ainsi qu'un terminateur d'un des nucléotides en faible proportion $4p$ par rapport au nucléotide libre correspondant. Par exemple, si ce terminateur est A^*, alors des brins d'ADN de longueur aléatoire vont être créés, dont la première base correspond au point B1 (figure 26) et la dernière base correspond à la première inclusion d'un terminateur A^* :

$$
\begin{array}{rl}
\text{brin ADN [B1,B2] :} & CGGACTACGTTAACCAGTC.... \\
\text{brin ADN n}^o1 \text{ :} & CGGACTACGTTA^* \\
\text{brin ADN n}^o2 \text{ :} & CGGA^* \\
\text{brin ADN n}^o3 \text{ :} & CGGACTACGTTAACCA^* \\
\vdots & \vdots
\end{array}
$$

On suppose que les processus enzymatiques se déroulent à la même vitesse dans chacune des solutions, si bien qu'on recueille par cette méthode environ le même nombre de brins de chaque solution. On note N le nombre total de brins, et L_j, $j = 1, ..., N$ leurs longueurs respectives. Du fait de la distribution aléatoire des bases dans le brin [B1,B2], la longueur des brins d'ADN ainsi générés suit la loi suivante :

$$\mathbb{P}(L_j = x) = p(1-p)^{x-1}, \text{ pour } x = 1, 2, ...$$

Par chromatographie on peut alors déterminer la séquence des bases du brin, si on a parmi les N brins toutes les longueurs possibles de 1 à d. Cette limitation est très restrictive, on s'autorise donc un pourcentage de longueurs non représentées, ce qui correspond à des bases du brin qui ne se trouvent pas parmi les ultimes bases des N brins clonés. On a :

$$\text{Pourcentage de bases détectées au moins une fois } = 1 - \frac{1}{pd}\left(\mathrm{E}_1(pNe^{-pd}) - \mathrm{E}_1(pN)\right), \quad (38)$$

où E_1 est l'intégrale exponentielle $\mathrm{E}_1(z) = \int_z^\infty e^{-s}s^{-1}ds$. Différentes lignes de niveaux de cette probabilité de séquençage sont tracées sur la Figure 27 en fonction des paramètres pd et N/d.

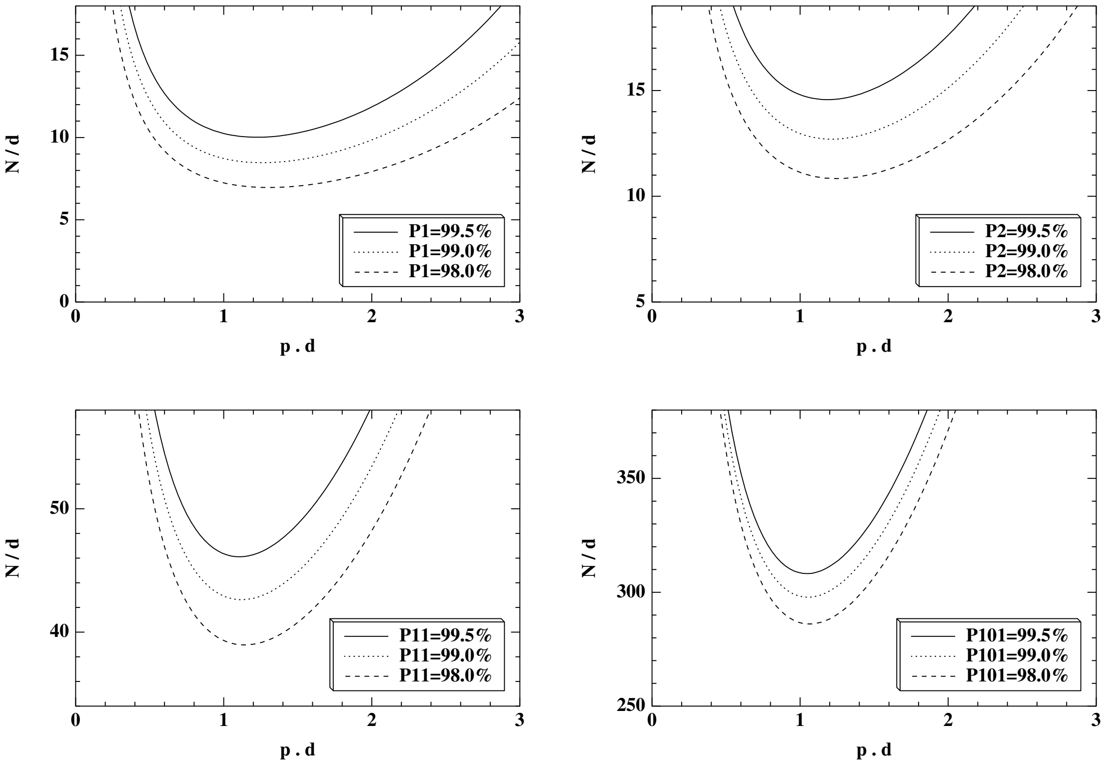

Figure 27 : *Quelques lignes de niveaux du pourcentage Pk de bases détectées au moins k fois*

On peut donner la distribution complète de la détection des bases :

$$\text{Pourcentage de bases détectées exactement } k \text{ fois } = \frac{1}{pd}\Big(F_k(pNe^{-pd}) - F_k(pN)\Big), \qquad (39)$$

où $F_k(z) = \frac{1}{k}e^{-z}\sum_{j=0}^{k-1}\frac{z^j}{j!}$, pour $k \geq 1$ et $F_0 = \mathrm{E}_1$.

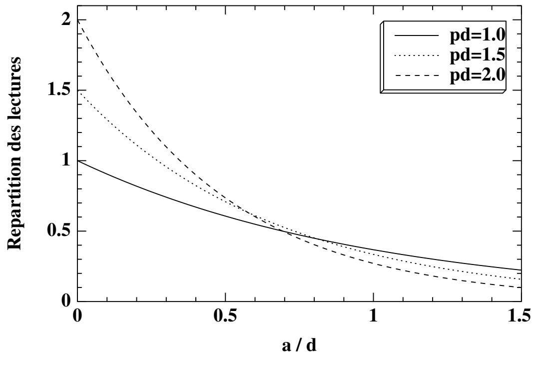

Figure 28 : *Répartition des lectures des bases*

On peut enfin regarder la répartition des bases qui sont lues :

$$\mathbb{E}[\text{ nombre de détections pour la base } a \in \{1, ..., d\}\ ] = pNe^{-pa}. \qquad (40)$$

Cette distribution une fois normalisée est tracée sur la Figure 28. Il apparaît que les bases correspondant au début du brin sont sur-représentées par rapport à celles situées à la fin. Cette disymétrie est due au fait qu'on commence toujours la lecture par la même extrémité du brin. On peut néanmoins se demander si on ne peut pas compenser au moins en partie cette disymétrie, et trouver une stratégie où le nombre de détections pour chaque base $a \in \{1, ..., d\}$ serait à peu près le même. Pour cela, on peut envisager d'utiliser des solutions avec différentes concentrations p d'inhibiteurs. Par des arguments de convexité, on montre qu'en fait on n'y gagne pas. Imaginons que les brins proviennent pour N d'entre eux de solutions de concentration $4p$, et pour N' de solutions de concentration $4p'$, alors la configuration où on tire $N + N'$ brins de solutions de concentration $4p''$, avec

$$p'' = \frac{p + p'}{2} \frac{1 + 2\varepsilon\delta + \varepsilon^2}{1 + \varepsilon\delta}, \qquad \varepsilon = \frac{p - p'}{p + p'}, \qquad \delta = \frac{N - N'}{N + N'},$$

donne des probabilités de séquençage meilleures. Pour améliorer de manière notable le rendement de séquençage d'un brin d'ADN, il faut effectuer la lecture à partir des deux extrémités, comme cela est montré dans le paragraphe suivant.

5.2 Lecture d'une séquence par les deux bouts

On suppose que les clones sont constitués uniquement par le brin qu'on veut séquencer, c'est à dire sur le dessin 26 que les points B2 et B3 sont confondus. On suppose aussi que la lecture commence de manière equiprobable au point B1 ou au point B2, si bien qu'on dispose de $N/2$ brins partant de B1 et $N/2$ brins partant de B2.

On appelle toujours $P1$ le pourcentage de bases représentées au moins une fois, ce qui correspond à des bases du brin qui se trouvent parmi les ultimes bases des N brins clonés. On a :

$$P1 = 1 - \frac{2}{pd} F_0(pd, N/d), \tag{41}$$

où $F_0(y, z) = \int_0^{\cosh(y/2)} \frac{\exp(-yze^{-y/2}s)}{\sqrt{s^2-1}} ds$. Différentes lignes de niveaux de cette probabilité de séquençage sont tracées sur la Figure 29 en fonction des paramètres pd et N/d. Lorsqu'on compare cette figure avec la figure équivalente 27 correspondant à une lecture par une seule extrémité, on détecte une diminution sensible du nombre total de brins à mesurer pour obtenir une probabilité de séquençage $P1$ donnée.

On peut donner la distribution complète de la détection des bases :

$$\text{Pourcentage de bases détectées exactement } k \text{ fois } = \frac{2}{pd} F_k(pd, N/d), \tag{42}$$

où $F_k(y, z) = \int_0^{\cosh(y/2)} \frac{(yze^{-y/2}s)^k}{k!} \frac{\exp(-yze^{-y/2}s)}{\sqrt{s^2-1}} ds$.

6 Influence de la longueur minimale de chevauchement

On va étudier ici l'influence du paramètre r défini comme le nombre minimal de bases consecutives à partir duquel on décide de retenir un chevauchement. On ne va néanmoins pas refaire tous les calculs. On va tout d'abord considérer l'influence de r du point de vue de la distribution statistique du nombre des contigs de séquence et de leur longueur moyenne, dans la configuration

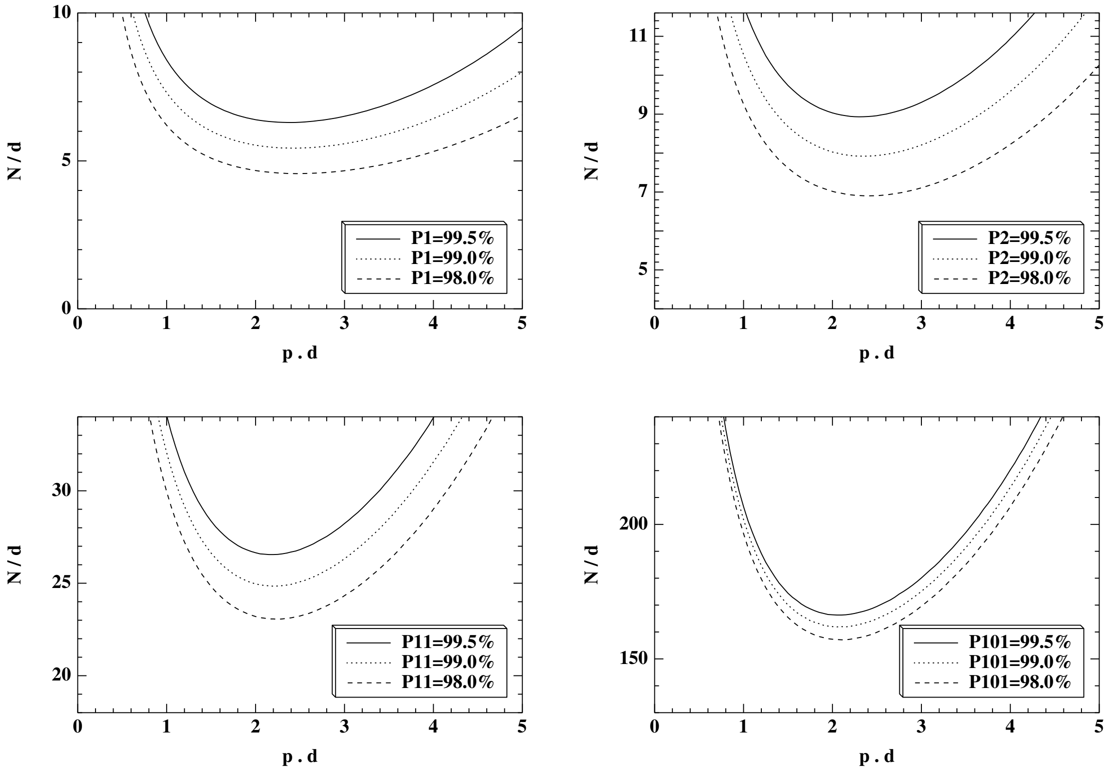

Figure 29 : *Quelques lignes de niveaux du pourcentage Pk de bases détectées au moins k fois - Configuration de lecture par les deux extrémités*

du shot gun simple. On peut déterminer leurs valeurs théoriques en fonction des paramètres R_s et $\alpha = r/d \in [0, 1[$ si on suppose que la banque **B** est sans biais. On trouve :

$$NC_s = \frac{GR_s}{d} e^{-R_s(1-\alpha)},$$
$$\langle LC_s \rangle = \frac{d}{R_s} \left(e^{R_s(1-\alpha)} - 1 \right) + \alpha d,$$

d'où on peut déduire que le nombre total de bases dénombrées est, à des fluctuations en $\sqrt{G}$ près :

$$N_d = G \left(1 - e^{-R_s(1-\alpha)} (1 - \alpha R_s) \right).$$

Attention, on peut se trouver face à un gag amusant. Du fait qu'on a rejeté un certain nombre de chevauchements parce qu'ils étaient inférieurs à la taille minimale r, on peut très bien se retrouver dans une situation où le nombre total de bases dénombrées est plus grand que G, car on a compté deux fois certaines bases (voir la figure 30).

On peut décrire le recouvrement du génome de manière un peu plus précise. On appelle N_j, $j \geq 1$, le nombre de bases dont on a détecté qu'elles font partie d'un recouvrement d'ordre k (avec un chevauchement de taille minimale r). Dans cette classification, N_1 est le nombre de bases isolées. On note aussi $\bar{N}_j = \sum_{k=j}^{\infty} N_k$ le nombre de bases qui font partie d'un recouvrement

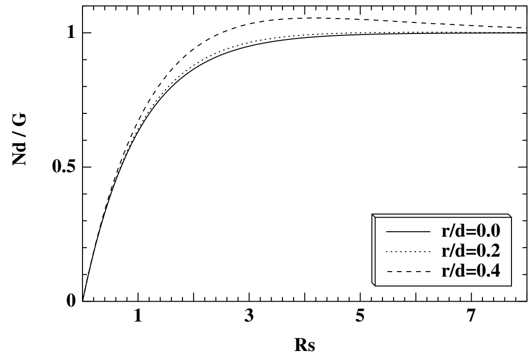

Figure 30 : *Nombre total de bases dénombrées normalisé*

d'ordre j ou plus. On a bien sûr $\bar{N}_1 = N_d$. Plus précisément :

$$\begin{aligned}
N_1 = G \quad & \left\{2e^{-R_s(1-\alpha)} + e^{-R_s}(-2 + (1-2\alpha)R_s)\right\}, \\
N_2 = G \quad & \left\{2e^{-R_s(1-\alpha)} + e^{-R_s}(-2 - 2\alpha R_s + (\frac{1}{2} - \frac{3\alpha^2}{2})R_s^2)\right\}, \\
N_3 = G \quad & \left\{e^{-R_s(1-\alpha)}\left(-2 + (\frac{1}{6} - \frac{\alpha^2}{2} + \frac{\alpha^3}{3})R_s^3 + (-\frac{\alpha}{6} + \frac{\alpha^2}{2} - \frac{\alpha^3}{2} + \frac{\alpha^4}{6})R_s^4\right)\right. \\
& \left. + e^{-R_s}\left(2 + 2\alpha R_s + \alpha^2 R_s^2 + \frac{\alpha^2}{2}R_s^3\right)\right\}, \\
N_4 = G \quad & \left\{e^{-R_s(1-\alpha)}\left(-2 + (\frac{1}{24} - \frac{\alpha^2}{4} + \frac{\alpha^3}{3} - \frac{\alpha^4}{8})R_s^4 + (-\frac{\alpha}{24} + + \frac{\alpha^2}{6} - \frac{\alpha^3}{4} + \frac{\alpha^4}{6} - \frac{\alpha^5}{24})R_s^5\right)\right. \\
& \left. + e^{-R_s}\left(2 + 2\alpha R_s + \alpha^2 R_s^2 + \frac{\alpha^3}{3}R_s^3 + \frac{\alpha^3}{6}R_s^4\right)\right\}.
\end{aligned}$$

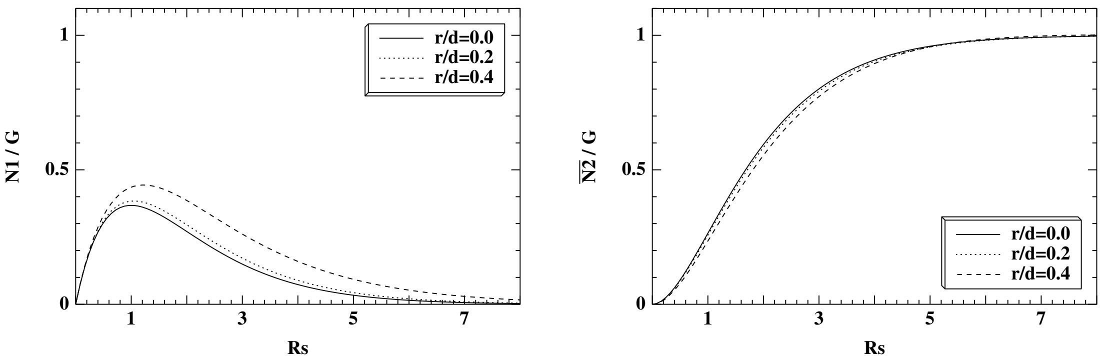

Figure 31 : *Nombre de bases isolées* $N1$ *et chevauchantes* $\bar{N}2$

On retrouve bien dans la limite $\alpha \to 0$ les valeurs déjà mise en évidence dans (1) correspondant à une loi de Poisson : $N_j = G\frac{R_s^j}{j!}e^{-R_s}$. On constate de plus que le paramètre r a très peu d'influence sur le comportement des quantités statistiques intéressantes tant que $\alpha = r/d$ est plus petit que $\sim 20\%$. On peut mettre en évidence ce phénomène en développant les quantités N_j en $\alpha \ll 1$. Il apparaît que les premiers termes correctifs des développements en $O(\alpha)$ sont nuls, et qu'il faut attendre le terme en $O(\alpha^2)$ pour mettre en évidence une correction au comportement limite $\alpha \to 0$. C'est pourquoi il faut une valeur assez forte de α pour trouver une différence significative de comportement.