Techniques informatiques pour la cartographie physique de genome humain

B. Lacroix, Jean-Jacques Codani

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 1560

Programme 2
Calcul Symbolique, Programmation
et Génie logiciel

TECHNIQUES INFORMATIQUES POUR LA CARTOGRAPHIE PHYSIQUE DU GENOME HUMAIN

Bruno LACROIX
Jean-Jacques CODANI

Novembre 1991

Techniques informatiques pour la cartographie physique du génome humain

Computational aspects of human genome physical mapping.

Bruno Lacroix

CEPHB/Généthon
13, place de Rungis
75013 Paris
email : lacroix@genethon.fr

Jean-Jacques Codani

INRIA
Domaine de Voluceau B.P. 105
78153 Le-Chesnay Cedex
email : codani@inria.fr

Résumé

Ce rapport présente un modèle théorique et les techniques informatiques associées pour construire la carte physique complète du génome humain. La quantité de données à traiter et la complexité du modèle nécessitent un calcul numérique intensif. Diverses stratégies d'implémentation, tenant compte d'un double niveau de parallélisme, sont proposées.

La carte physique du génome humain constituera un outil puissant pour l'identification et l'isolation de gènes inconnus impliqués dans des mutations phénotypiquement détectables, tout particulièrement des maladies héréditaires.

Abstract

This report deals with the theoretical model and the computer science techniques used to build the physical map of the human genome. The amount of data and the complexity of the model imply numeric intensive computation. Different strategies of implementation are proposed, according to a double level of parallelism.

The physical map of the human genome will be a powerful tool for identification of unknown genes, implied in phenotypical detectable mutations, particulary genetic diseases.

Premier prix d'excellence IBM 1991 en calcul numérique intensif.

Ce travail concerne le traitement informatique de données biologiques issues des laboratoires de Généthon, association financée par l'AFM (Association Française contre les Myopathies). Il fait l'objet d'une collaboration entre l'INRIA et Généthon pour les aspects informatiques.

Ce travail a été supporté en partie par une bourse de l'Ecole Polytechnique.

Introduction

La carte d'un génome est la synthèse des informations fonctionnelles et structurales contenues dans la séquence nucléique[1]. Pourtant, le séquençage direct n'est pas à l'heure actuelle la manière la plus efficace de collecter ces informations, ceci pour deux raisons: d'une part, le séquençage de régions d'ADN dont la taille dépasse quelques dizaines de milliers de paires de bases est un travail long et coûteux, et, au moins aussi important, les outils d'interprétation ne sont pas suffisamment performants pour exploiter de telles séquences brutes. Les stratégies utilisées sont donc différentes suivant les caractéristiques génomiques étudiées. Ainsi, les données fonctionnelles sont obtenues par l'étude des gènes à travers leur produit et la régulation de leur expression. De même, les données structurales concernant le polymorphisme génétique sont obtenues par les techniques de cartographie génétique, et celles concernant la distribution des sites de restriction, des séquences codantes, ou d'autres traits directement reliés à la séquence proviennent des techniques de cartographie physique.
Ces deux dernières techniques, outre les informations structurales qu'elles apportent, sont aussi des outils indispensables à l'identification et l'isolation de gènes inconnus impliqués dans des mutations phénotypiquement détectables, tout particulièrement des maladies héréditaires. C'est pourquoi leur amélioration et leur adaptation à l'échelle gigantesque de l'étude d'un génome fait l'objet d'un effort international.

Très schématiquement, la cartographie physique et la cartographie génétique consistent toutes deux à construire, à partir d'un ensemble d'objets générés dans la région à cartographier, un maillage d'objets dont les positions relatives sont connues. Un tel maillage permet la localisation rapide d'un objet inconnu en cherchant sa position par rapport aux objets du maillage, la précision de cette localisation dépendant alors de la densité du maillage. Pour la carte génétique, les objets étudiés sont des loci repérés par des sondes polymorphes, et la position relative de deux loci est définie par la fréquence de recombinaison observée entre eux au cours des meioses.

En ce qui concerne la cartographie physique, les objets sont des fragments d'ADN clonés (appellés clones) et la position relative de deux clones est définie par leur recouvrement, c'est-à-dire par la longueur d'ADN qu'ils ont en commun. Cette longueur peut soit être nulle, soit être positive, auquel cas elle est exprimée en nombre de bases communes.
La carte peut donc être construite de manière dirigée en identifiant sans erreur tous les clones recouvrant avec un clone donné, puis en réitérant ce processus jusqu'à épuisement des clones disponibles. Néanmoins, la lourdeur des manipulations nécessaires à chaque étape limite cette démarche à l'étude de petits

[1] Pour une définition des termes biologiques, se reporter au glossaire page 20.

ensembles de quelques dizaines de clones, donc à de petites régions du génome. Pour des projets de plus grande envergure, on génère d'abord une information sur chacun des clones disponibles par un même processus expérimental: cette information est appellée **l'empreinte** du clone. On construit ensuite une matrice des distances entre tous les clones, la distance entre deux clones étant mesurée par la similitude entre leurs empreintes. Cette distance n'est que le reflet du recouvrement réel des clones à travers le processus expérimental, et est donc transformée en probabilité de recouvrement. Le graphe probabiliste ainsi obtenu est alors seuillé en fonction du taux de faux recouvrements toléré, et les composantes connexes correspondantes, appelées contigs, sont retenues pour une étude détaillée.

Il apparaît donc que ces techniques non-dirigées peuvent nécessiter des moyens de calcul importants. D'une part, la taille de la matrice présente une croissance en $N * (N-1)/2$, alors que N peut atteindre plusieurs dizaines de milliers pour l'étude de régions de la taille d'un chromosome humain (section 1). D'autre part, du fait de son caractère expérimental, la mesure de la distance est entachée d'une erreur d'autant plus importante que les manipulations de biologie moléculaire sont complexes (section 2). La prise en compte de ces erreurs dans l'analyse des données nécessite l'utilisation de procédures sophistiquées (section 3) pour parvenir à un compromis optimal entre le taux d'erreur final et l'efficacité de la technique. L'implémentation requiert donc des optimisations poussées et doit préserver un double parallélisme dû à l'indépendance des $N * (N-1)/2$ comparaisons et au caractère bayesien du modèle utilisé pour leur calcul (section 4). Les applications sont présentées dans la section 5.

1 Considérations théoriques

L'essentiel de cette section est tiré de l'article [LW88]. Cet article présente une étude théorique de la cartographie physique non dirigée qui ne tient pas compte des spécificités du processus d'obtention des empreintes et du processus de seuillage de la matrice des distances. Ces processus sont simplement caractérisés par le recouvrement minimal θ (ou σ, égal à $(1-\theta)$) qu'ils sont capables de détecter entre deux clones. On suppose en outre que le nombre de faux recouvrements ainsi détectés est négligeable et que le clonage est purement aléatoire. Cette hypothèse étant très simplificatrice du fait de biais importants, les résultats suivants sont à prendre comme des optima théoriques. La relation suivante est alors vérifiée:

$$\text{nombre de contigs groupant au moins i clones} \quad = \quad ce^{-2c\sigma}(1-e^{-c\sigma})^{i-1}\, G/L$$

avec:

G = longueur de la région étudiée.
L = longueur des clones, supposée constante.
N = nombre de clones analysés.
c = NL/G, redondance de l'ensemble de clones.

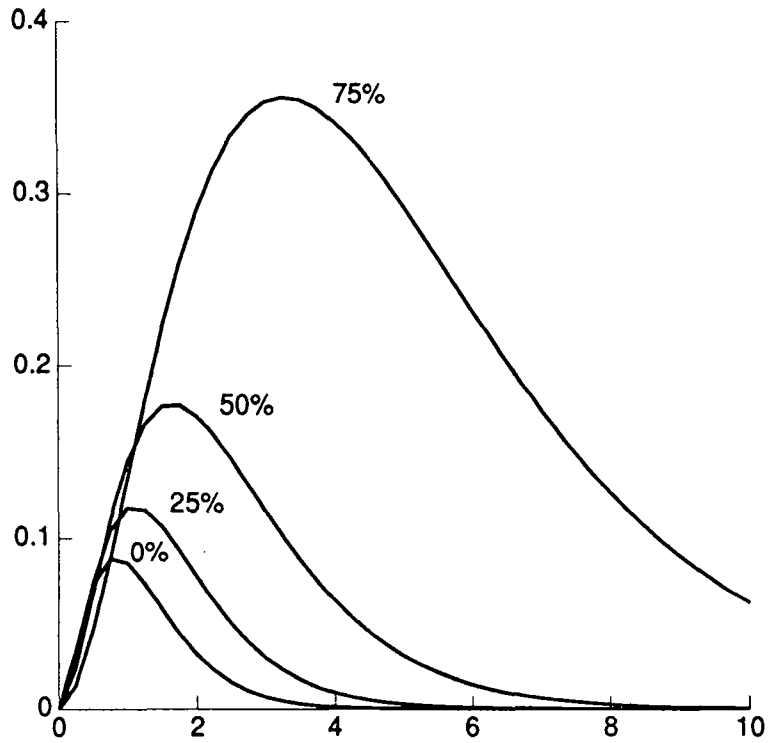

Figure 1: Nombre de contigs de plus de deux clones obtenus en fonction du nombre de clones analysés (les deux grandeurs sont exprimées en unités de G/L).

La famille de courbes correspondante (figure 1) montre que le nombre de clones à analyser est toujours de l'ordre de quatre à cinq équivalents pour obtenir des résultats satisfaisants, et que l'influence de θ sur ces résultats est sensible. Il est donc clair :

- que l'utilisation des systèmes de clonage de capacité maximale est indispensable: à l'heure actuelle, il s'agit des chromosomes artificiels de levure

(appelés YACs). En effet, les résultats sont exprimés en unités de G/L qui vaut typiquement 10^3 pour des YACs contre 10^4 pour les autres systèmes de clonage, et ce pour des régions de la taille d'un chromosome humain

- que dans tous les cas l'étude d'une région de la taille d'un chromosome humain nécessite l'analyse de plusieurs milliers de clones
- que tout gain dans l'efficacité de la détection des recouvrements, notamment par l'utilisation de modèles probabilistes sophistiqués, peut diminuer ce nombre de clones à analyser et se traduire par des gains de temps et de coût importants

Ces quelques principes ont étés choisis comme principes directeurs des techniques développées par Généthon.

2 Généthon

L'équipe de Généthon s'est donc attachée à développer une chaîne cohérente de procédures expérimentales et informatiques permettant l'établissement rapide de la carte physique d'une grande région du génome humain. Le pan biologique de ce programme a permis la mise au point d'un protocole expérimental semi-automatique [BCBL+91] pour la génération de l'empreinte de clones d'ADN humain. Ce protocole comprend trois étapes:

- digestion par une enzyme de restriction: ces enzymes reconnaissent un mot de six bases (six tout au moins pour celles qui nous occupent) et coupent l'ADN à chaque occurence de ce mot.
- séparation des fragments résultant de la digestion en fonction de leur taille par électrophorèse sur gel d'agarose, puis transfert de ces fragments sur membrane.
- détection des fragments sur la membrane par hybridation moléculaire avec une sonde luminescente et acquisition du signal résultant sur film photographique.

Le film obtenu est enfin digitalisé et traité par un logiciel spécialisé qui détecte les fragments et évalue leur taille à partir de fragments de taille connue (voir figure 2 pour un exemple d'image digitalisée). L'empreinte utilisée est la suite des tailles de ces fragments détectés.

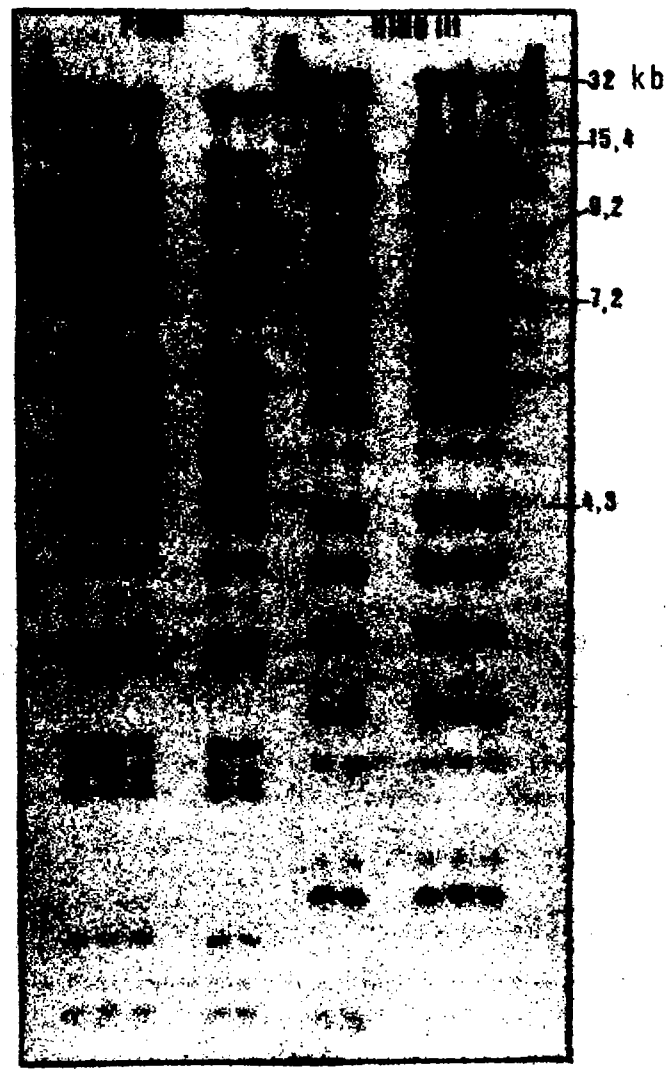

Figure 2: Autoradiographie d'un gel de digestions de 2 YACs par les enzymes PvuII et HindIII hybridé avec de l'ADN génomique total humain marqué. Chaque piste correspond à une préparation et à une digestion indépendantes.

Ce protocole, plus complexe que les protocoles normalement utilisés pour générer des empreintes, a pourtant été choisi car c'est le seul qui soit applicable aux YACs. En effet, leur capacité de clonage accrue d'un facteur dix à vingt [AAC+87] a pour contrepartie une utilisation beaucoup plus délicate. En particulier, alors que la séparation de l'ADN exogène et du génome de l'hôte est aisée avec d'autres vecteurs, elle est impossible dans le cas des YACs. Une technique de génération d'empreintes adaptée doit donc extraire l'information relative à l'ADN humain alors que celui-ci est noyé dans cent fois plus d'ADN de levure. Pour ce faire, la technique décrite met à profit l'existence dans le génome de l'homme de **séquences répétées** à un grand nombre d'endroits différents. La plus fréquente de ces séquences, la séquence ALU, existe à plusieurs centaines de milliers d'exemplaires alors que la famille KPN est elle présente à cent mille copies.

L'hybridation avec ces séquences comme sondes détecte sélectivement des fragments d'ADN humain, et leur fréquence sur le génome garantit que l'empreinte générée est informative (i.e. que les empreintes ne sont pas vides). L'ensemble du processus est schématisé dans la figure 3.

En résumé, l'empreinte consiste en une suite d'entiers représentant des tailles de fragments d'ADN, une taille apparaissant dans l'empreinte s'il existe dans le clone correspondant deux occurrences successives à cette distance d'un mot donné (au sens de l'alphabet ATGC de l'ADN) encadrant une occurence d'un autre mot donné. Suivant la fréquence de ce mot-ci, le nombre moyen d'entiers d'une empreinte varie de trente (pour la séquence ALU) à dix (pour la séquence KPN).

Il est important de noter à ce stade:

- que tous les fragments issus d'un clone ne sont pas détectés
- que la présence de fragments de même taille dans les empreintes de deux clones différents peut être fortuite
- que la distribution sur le génome des deux mots qui caractérisent les empreintes n'est pas connue
- que le processus, passablement complexe, génère un certain nombre d'erreurs sur les tailles finales

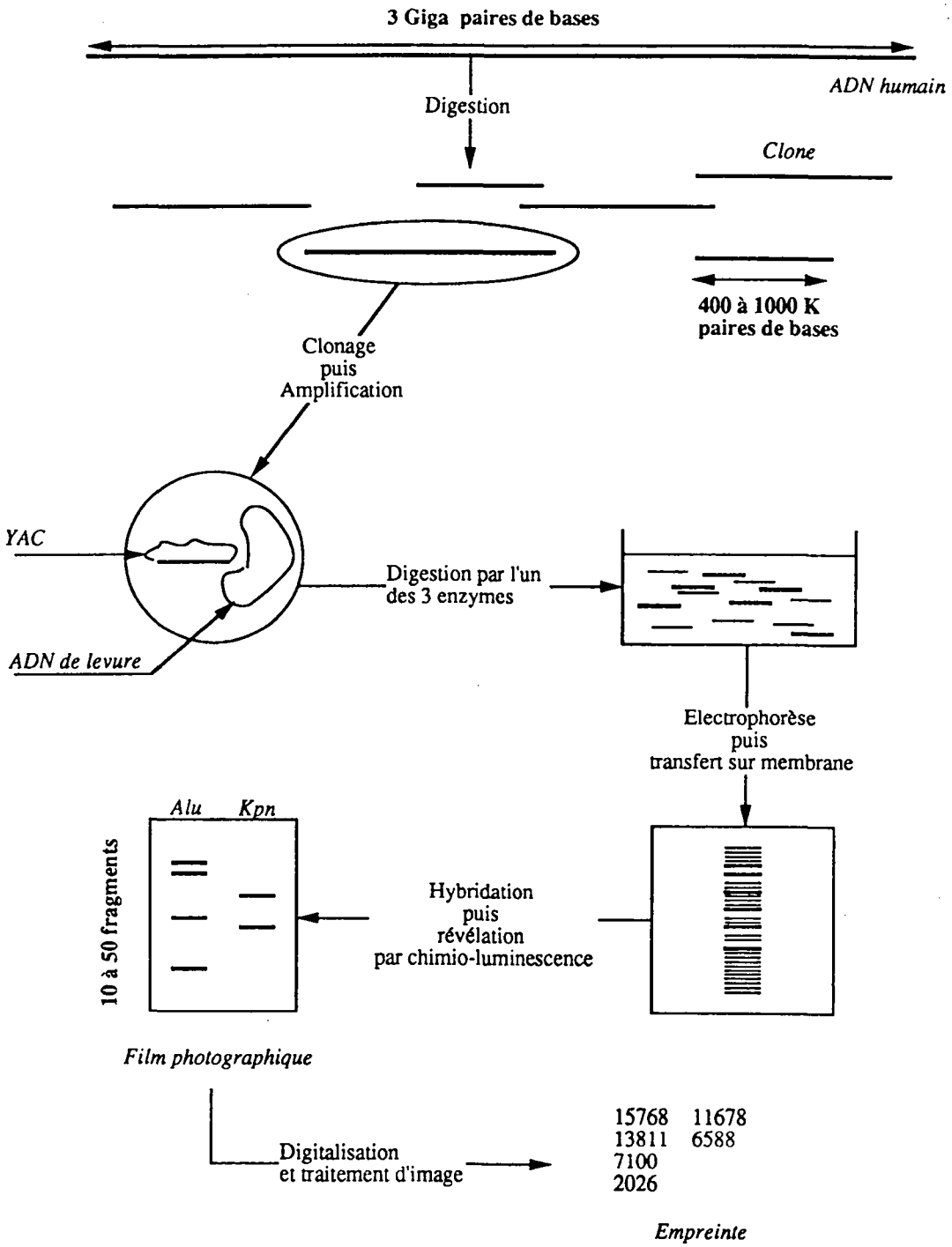

Figure 3: Résumé des différentes étapes de génération des empreintes.

3 Modèle

Le logiciel candidat évalue la probabilité d'un recouvrement non nul entre deux clones connaissant leurs empreintes, et évalue la longueur commune la plus probable. Cette probabilité est estimée par application du théorème de Bayes. Ce théorème fournit en effet un cadre robuste et extensible qui permet d'utiliser au mieux les connaissances disponibles tant sur les distributions des deux mots précédemment évoqués que sur les différentes sources d'erreur rencontrées dans le processus d'obtention des empreintes. De plus, les grandeurs inconnues peuvent être prises en compte par intégration sur leurs densités de probabilité [BT91].

Dans toute la suite, on utilisera les notations suivantes:

L_1 et L_2	=	longueurs des deux clones, supposées connues, en paires de bases
N_1 et N_2	=	nombre de fragments détectés pour chacun des deux clones
T	=	recouvrement entre les deux clones, en paires de bases
données	=	empreintes des deux clones
$\pi(t)$	=	probabilité *a priori* d'un recouvrement t entre deux clones

$\pi(t)$ se calcule facilement et on a:

$$\pi(t) = \begin{cases} 1 - (L_1 + L_2)/G & t = 0 \\ 2/G & 1 \leq t \leq Min(L_1, L_2) \\ 0 & t > Min(L_1, L_2) \end{cases}$$

On peut alors écrire:

$$\mathbf{P}(\mathbf{T} = t_0/\text{données}) = \frac{\mathbf{P}(\text{données}/\mathbf{T} = t_0)\pi(t_0)}{\int_{t=0}^{t=min(L_1,l_2)} \mathbf{P}(\text{données}/\mathbf{T} = t)\pi(t)dt} \qquad (1)$$

Or, les deux listes de tailles de fragments qui constituent les données sont mesurées par un processus expérimental générateur d'erreurs. La plus grande partie de ces erreurs peut être modélisée globalement par un bruit aléatoire gaussien dont l'écart-type a été estimé par des études de reproductibilité à $0,5\%$ de la taille du fragment. Deux fragments détectés dans deux clones différents ne peuvent donc être identiques (au sens de la séquence) que si leurs tailles ne diffèrent pas plus que 6 à 7%: de tels fragments sont dits appariables, et on considère alors que la taille du fragment réel associé est la moyenne de ces deux mesures. La tolérance sur la mesure, notée ϵ, définit donc la matrice $\mathcal{M}$ des appariements possibles entre les empreintes de deux clones.

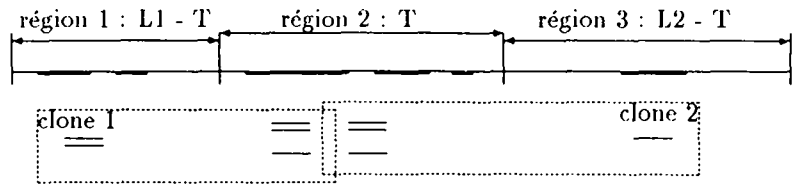

Figure 4: Représentation d'un état réel compatible avec les données.

Définissons un état réel du système constitué par les deux clones comme une manière d'affecter chacun des fragments détectés à une des trois régions définies par un recouvrement non nul entre les clones (cf figure 4). Un fragment appartenant à l'empreinte d'un clone peut être affecté soit à la partie propre à ce clone, soit à la partie commune aux deux clones: dans ce dernier cas, un fragment appartenant à l'empreinte de l'autre clone doit lui être associé. Un tel état réel est dit compatible avec les données si et seulement si il n'implique que des appariements compatibles avec $\mathcal{M}$.

Cette matrice $\mathcal{M}$ définit donc l'espace Ω des états réels compatibles. Or, Ω est de manière naturelle la réunion disjointe des espaces des états réels compatibles ayant un nombre fixé de bandes appariées, chacun de ces espaces noté $\Omega(k)$ étant fini. Si N est le nombre maximal de bandes appariables entre les deux clones, $\Omega(k)$ est vide pour $k > N$. On peut alors écrire:

$$\mathbf{P}\left(\mathbf{T} = t_0/\text{données}\right) = \frac{\sum_{k=0}^{k=N} \sum_{\omega \in \Omega(k)} \mathbf{P}(\omega/\mathbf{T} = t_0)\pi(t_0)}{\int_{t=0}^{t=min(L_1,L_2)} \left(\sum_{k=0}^{k=N} \sum_{\omega \in \Omega(k)} \mathbf{P}(\omega/\mathbf{T} = t)\pi(t)dt \right)} \quad (2)$$

Pour tout ω appartenant à $\Omega(k)$, on notera:

$$\begin{array}{lcll} l^1 & = & (l_i^1)_{1 \le i \le N_1 - k} & = \text{ liste des fragments propres au clone 1} \\ l^2 & = & (l_i^2)_{1 \le i \le N_2 - k} & = \text{ liste des fragments propres au clone 2} \\ l^3 & = & (l_i^3)_{1 \le i \le k} & = \text{ liste des fragments communs} \end{array}$$

La probabilité d'un état réel quelconque s'écrivant naturellement comme le produit des probabilités relatives aux trois régions précédemment évoquées, on obtient pour un état réel compatible avec les données:

$$\mathbf{P}(\omega/\mathbf{T} = t) = \mathbf{Pr}(l^1, L_1 - t)\ \mathbf{Pr}(l^2, L2 - t)\ \mathbf{Pr}(l^3, t) \quad (3)$$

La probabilité $\mathbf{Pr}((l_i)_{1 \le i \le n}, t)$ s'obtient ensuite en sommant sur toutes les partitions de $(l_i)_{1 \le i \le n}$ puisque la contiguité des fragments détectés est inconnue.

On a donc:

$$\mathbf{Pr}((l_i)_{1\leq i\leq n}, t) = \sum_{K=1}^{K=n} \left(\sum_{\sigma\in\tau(K)} \mathbf{p}((l_{l_i})_{1\leq i\leq K}, t) \right) \tag{4}$$

où

$\forall K \in \{1,\ldots,N\}$	$\tau(K)$ est l'ensemble des partitions à K éléments
$\forall \sigma \in \tau(K), \forall i \in \{1,\ldots,K\}$	$(l_{l_i})_{1\leq l\leq k_i} = (l_{\sigma(i,l)})_{1\leq l\leq k_i}$ où k_i est le cardinal du $i$ième élément de σ

Enfin, l'expression la plus générale de $\mathbf{p}((l_i)_{1\leq i\leq K}, L)$ est:

$$\sum_{k_1,\ldots,k_{K+1}=0}^{\infty} \left(\int_0^{L-\sum_{i=1}^{i=K} l_{l_i}} \left(\mathcal{P}_{(k_1,(l_{l_1}),x_1)} \left(\int_{x_1+l'_1}^{L-\sum_{i=2}^{i=K} l_{l_i}} \mathcal{P}_{(k_2,(l_{l_2}),x_2-x_1)} \right) \right) \cdots \right) \tag{5}$$

où

$\forall i \in \{1,\ldots,K+1\}$	k_i	=	nombre de fragments non detéctés dans l'interstice situé entre le $i-1$ ième et le i ème groupements détectés.
$\forall i \in \{1,\ldots,K\}$	x_i	=	abscisse du début du i ème groupement détecté comptée à partir du début du clone.

$\mathcal{P}(k, (l_i)_{1\leq i\leq n}, x)$ est alors simplement la probabilité d'observer les n fragments $(l_i)_{1\leq i\leq n}$ bout à bout dans cet ordre et de ne pas détecter k fragments dans la longueur $(x - \sum_{i=1}^{i=n} l_i)$. Alors que les étapes précédentes du calcul n'étaient que des sommations sur les différentes manières d'agencer les fragments détectés, $\mathcal{P}$ dépend uniquement des hypothèses faites sur la distribution des séquences répétées et des sites de restriction et sur les processus d'erreurs rencontrés.

En ce qui concerne les sites de restriction, un processus de Poisson de paramètre $\beta = 1/4096$[2] constitue un modèle tout à fait satisfaisant. En revanche, la modélisation de la distribution des séquences répétées par un processus de Poisson de paramètre α, bien que simple et attirante, ne peut être faite qu'avec réserve. Les rares études faites sur ces distributions mettent en effet en évidence des inhomogénéités importantes [MTM+89]. C'est cependant la voie choisie à

[2]Les sites en question étant des séquences de six paires de bases leur fréquence vaut en première approximation 4^{-6}

l'heure actuelle, une solution plus satisfaisante consistant à calculer les probabilités pour plusieurs valeur de α et de sommer sur la distribution présumée de ces valeurs. Comme on le verra dans la suite, l'implémentation de l'algorithme permet un tel calcul en seule passe sans modification majeure.

Sous ces hypothèses, la probabilité d'observer n fragments $(l_i)_{1\leq i\leq n}$ bout à bout est égale à: $\beta^{n+1}\prod_{i=1}^{i=n}\left(e^{-\beta l_i}(1-e^{-\alpha l_i})\right)$. Cette probabilité ne dépend donc pas de l'ordre des $(l_i)_{1\leq i\leq n}$, et la sommation 4 se simplifie drastiquement.

La probabilité de ne pas détecter k fragments dans la longueur x vaut quant à elle $e^{-\alpha x}$ pour tout k si aucune erreur ne peut conduire à la non-détection d'un fragment contenant des séquences répétées. Cette hypothèse sur les erreurs étant peu satisfaisante, une expression plus générale de cette probabilité est $\rho^{*k}(x)$ où ρ^{*k} est le $k$ième produit de convolution de la loi de probabilité de non-détection d'un fragment; $\rho^{*k}(x)$ peut par exemple se calculer par transformée de Laplace. Malheureusement, les intégrations de l'expression 5 deviennent dans ce cas rapidement difficiles à mener.

A l'heure actuelle, le programme ne prend donc en compte explicitement aucune erreur et modélise les distributions de séquences répétées et de sites de restriction par des processus de Poisson. Dans ce cas, l'intégration 5 donne un monôme de degré K de $L-\sum_{i=1}^{n} l_i$ multiplié par un produit d' exponentielles des $(l_i)_{1\leq i\leq n}$ et la sommation 4 donne finalement pour $\mathbf{Pr}((l_i)_{1\leq i\leq n},t)$:

$$\boxed{\left(\prod_{k=1}^{k=n}(1-e^{-\alpha l_k})e^{(\alpha-\beta)l_k}\right)\left(n!\beta^{n-1}e^{-\alpha L}\sum_{k=1}^{k=n}C_{k+1}^{n+1}\frac{\beta^{k+1}}{k!}(L-l)^k\right)} \qquad (6)$$

où $l=\sum_{i=1}^{i=n} l_i$. Cette expression est donc le produit de deux termes:

- un terme qui dépend des $(l_i)_{1\leq i\leq n}$ (première parenthèse) notée Co. Il est important de noter que ce terme s'écrit comme le produit de n fois la même expression évaluée pour chacun des $(l_i)_{1\leq i\leq n}$.
- un terme qui dépend de la somme des $(l_i)_{1\leq i\leq n}$ uniquement.

Cette propriété est mise à profit dans le logiciel pour représenter un état total du système des deux clones par un **quadruplet** de trois valeurs entières, chaque valeur étant la somme des longueurs des fragments affectés à chacune des trois régions, et d'une valeur réelle, produit des Co de chacune des trois régions. Ce quadruplet et le nombre de bandes appariées caractérisent alors complètement un état. Cette organisation se généralise aisément au cas de n-uplets comportant $n-3$ champs Co, chacun d'eux étant calculé avec une valeur différente de α. Les algorithmes décrits dans la section 4 n'étant pas affectés par cette généralisation, on dispose ainsi d'un moyen efficace pour calculer les probabilités pour différentes valeurs de α en une seule passe.

4 Implementation

4.1 Généralités

L'implémentation du programme présenté doit satisfaire à trois contraintes:

- Portabilité: le séquençage du génome humain étant un effort international et l'équipement en moyens de calcul des biologistes étant hétérogène, le programme se doit d'être aisément portable sur une large gamme de machines.
- Modularité, lisibilité: l'implémentation doit rester évolutive, certaines parties du modèle statistique pouvant être soit raffinées, soit modifiées pour être transposées à l'étude d'autres génomes.
- Efficacité: la combinatoire engendrée par le volume des données et la complexité de l'estimation de la vraisemblance nécessitent un effort d'optimisation.

Or ces desiderata sont en partie contradictoires: les deux premières contraintes se traduisent par un empilement de couches logicielles ou des restrictions d'implémentation qui se font au détriment de la troisième. Le compromis est une version "opérationnelle" séquentielle (écrite en C), tirant partie d'un système d'exploitation de haut niveau ($Unix^{TM}$).
Un certain nombre d'options permettent de modifier plusieurs paramètres (fenêtres de lecture des bandes, fenêtres du nombre d'états à calculer, seuils de déclenchement d'approximations, familles de densités de probabilités à utiliser, ...), assurant à l'utilisateur une souplesse nécessaire compte-tenu de la variabilité des procédures expérimentales.

En outre, les évolutions récentes dans le domaine des architectures de machines (multiprocesseurs notamment) ne peuvent plus être ignorées de l'implémenteur. Or, le problème posé est tel que :

1. Pour N clones, les $N * (N-1)/2$ comparaisons sont indépendantes entre elles (ce que nous dénommerons parallélisme macroscopique)
2. Pour une comparaison donnée, l'algorithme de génération de l'espace des états possibles possède la propriété d'être parallélisable (parallélisme microscopique).

Le programme est donc organisé en modules qui peuvent être instanciés de manière simple et efficace sur une architecture de machine donnée. Un mod-

ule de gestion des Entrées/Sorties dont la dépendance par rapport au système d'exploitation est minimale permet des évaluations rapides sur des machines multiprocesseurs[3].

Nous exposerons en premier lieu les caractéristiques essentielles de la version séquentielle. Une grande partie des optimisations qui y seront décrites seront naturellement conservées dans les versions parallèles. Les traits de la version séquentielle permettant la mise en œuvre de versions parallèles (macroscopique et microscopique) seront soulignés dans les sections 4.3 et 4.4.

4.2 Séquentialité

La première partie du programme consiste à décomposer la matrice $\mathcal{M}$ des appariements possibles en ses M sous-espaces stables. La propriété suivante est alors vérifiée :

Propriété 4.1 *Tout état compatible du système total est associé de manière bi-univoque à une suite de M états, chaque état étant compatible avec une sous-matrice.*

Comme évoqué dans la section 3 un état compatible se caractérise par un quadruplet $\{l^1, l^2, l^3, Co\}$. On a donc d'après la propriété 4.1 :

Propriété 4.2 *Tout quadruplet q_t associé à l'état total est associé de manière bi-univoque à M quadruplets $(q_i)_{1 \leq i \leq M}$, chacun étant associé à chaque sous-matrice.*

Soit alors:	Q	l'ensemble des quadruplets décrits par Ω et $\forall q \in Q\ q = \{l^1, l^2, l^3, Co\}$
	Q_k	l'ensemble des quadruplets décrits par $\Omega(k)$
	N_Ω	le cardinal de Ω
	$Q_i^{\mathcal{M}}$	l'ensemble des quadruplets associés à la sous-matrice de rang i de $\mathcal{M}$
	n_i	le cardinal de $Q_i^{\mathcal{M}}$
	$Q_i^{\mathcal{M}}[j]$	le quadruplet de rang j dans cet ensemble
	E	l'ensemble des M-uplets d'entiers $(i_1, \ldots, i_M)$ tels que $\forall k \in \{1, \ldots, M\}\ i_k \leq n_k$
	$\otimes$	la loi de composition définie par: $\forall (q_1, q_2) \in Q^2\ q_1 \otimes q_2 = q$ avec $l^1 = l_1^1 + l_2^1,\ l^2 = l_1^2 + l_2^2,\ l^3 = l_1^3 + l_2^3,\ Co_1 = Co_1 * Co_2$

[3] La puissance se paye souvent par l'absence de couches logicielles de haut niveau.

Avec ces notations, les propriétés 4.1 et 4.2 impliquent:

$$\forall q_t \in Q, \exists! e = \{i_1, \ldots, i_M\} \in E \ / \ q_t = \bigotimes_{k=1}^{k=M} Q_k^{\mathcal{M}}[i_k] = \Phi(e) \qquad (7)$$

et si q_t est le quadruplet associé à ω, élément quelconque de Ω, l'expression 3 s'écrit:

$$\mathbf{P}(q_t/\mathbf{T} = t) = Co\ Pol(N_1 - k, L_1 - t - l^1) Pol(k, t - l^2) Pol(N_2 - k, L_2 - t - l^3) e^{\alpha(t - L_1 - L_2)} \qquad (8)$$

où k = nombre de bandes appariées de q_t

$Pol(n, X) = n! \beta^{n-1} \sum_{i=1}^{i=n} C_{i+1}^{n+1} \frac{\beta^{i+1}}{i!} X^i$

D'après 7, générer Q est équivalent à générer E, c'est à dire une suite $(e_i)_{1 \leq i \leq N_\Omega}$ d'éléments de E. Deux versions ont été implémentées pour cela.

Une version **récursive** permet de générer l'ensemble des e_n tels que $\sum_{k=1}^{M} e_n[k] = K$ = cste. (L'ensemble de quadruplets associés est Q_K). Cette approche permet notamment de ne calculer les probabilités que pour les $\Omega(k)$ dont la contribution à la somme 1 est prépondérante.

Une version **itérative**, pour exploiter le fait que l'application Φ est un homomorphisme de E sur Q, utilise un algorithme de génération de E minimisant les différences entre deux éléments successifs de la suite $(e_i)_{1 \leq i \leq N_\Omega}$. Cet algorithme est une extension de l'algorithme du code de Gray[4] et possède la propriété:

$$\forall n \in \{1, \ldots, N\}, \exists i \ / \ e_n[i] \neq e_{n-1}[i] \ et \ \forall k \neq i \ e_n[k] = e_{n-1}[k]$$

On a alors:

$$\Phi(e_n) = Q_i[e_n[i]] \otimes (Q_i[e_{n-1}[i]])^{-1} \otimes \Phi(e_{n-1})$$

Ces différents algorithmes – qui représentent 90% du temps d'exécution – ne sont pas vectorisables, mais sont sensibles aux passes d'optimisation des compilateurs et à l'architecture sur laquelle ils s'exécutent (récursivité, taille et stratégies de gestion du cache, ...). A ce jour, et dans l'univers des stations de travail, les temps obtenus sont de l'ordre de quelques millisecondes pour les sondes KPN et quelques centaines de millisecondes pour les sondes ALU en moyenne. Ils sont de quelques dizaines de secondes pour les cas les plus défavorables (trente bandes appariables, soit un nombre d'états de l'ordre de 2^{30}).

[4] Algorithme utilisé dans des sytèmes de codeurs optiques.

La deuxième partie consiste ensuite, pour chaque état généré, à intégrer numériquement l'expression 8 pour t variant de 0 à $Min(L_1, L_2)$. On utilise pour cela le fait que la famille de polynômes $(Pol(n,*))_{n\geq 0}$ est aisément tabulable. Compte-tenu de la faible incidence de l'interpolation des valeurs des $(Pol(n,*))_{n\geq 0}$ entre deux points échantillonnés, on fait, pour tout X, l'approximation $Pol(n,X) = Pol(n, Int_{s(X)})$ où $(Int_i)_{i\geq 0}$ est la famille des points d'échantillonnage des polynômes et $s(X)$ est tel que $Int_{s(X)} \leq X < Int_{s(X)+1}$. Modulo cette approximation, des états physiquement distincts de même nombre de bandes appariées peuvent ne différer que par leur coefficient. Ces coefficients sont alors factorisés dans l'expression 8. On obtient ainsi une compression très importante de l'espace des états, compression d'autant plus importante que Ω est grand.

4.3 Parallélisme macroscopique

Plusieurs architectures de machines sont envisageables pour exploiter l'indépendance des $N(N-1)/2$ comparaisons :

- Architectures bâties autour de stations de travail interconnectées par des réseaux à haut débit. L'utilisation d'outils standards fournis par les couches réseau des systèmes d'exploitation (exécution à distance, partage de fichiers, ...) ainsi que la représentation des données sous un format externe (XDR) permet de mettre rapidement en œuvre une distribution de tâches à très gros grain de parallélisme.
- Architectures MIMD à mémoire distribuée. La modularité du programme permet d'ajouter aisément un algorithme d'ordonnancement des tâches et une gestion des communications inter-processeurs. Des tests ont été effectué avec succès avec un algorithme simple ("self-scheduling").
- Architectures MIMD à mémoire partagée. Le programme – dans sa version native séquentielle – est tel que chaque comparaison ne provoque d'écriture que dans un espace mémoire qui lui est propre. Cet espace peut être alloué:
  - statiquement et réutilisé : cas séquentiel.
  - dynamiquement à chaque comparaison. Dans ce cas, moyennant une directive manuelle si les outils de parallélisation ne détectent pas l'inexistence d'interdépendances mémoire entre deux comparaisons, le programme séquentiel exploite immédiatemment l'aspect multiprocesseurs.

La mise en œuvre des solutions MIMD est de plus facilitée par une consommation mémoire relativement faible (2 à 3 Mo).

4.4 Parallélisme microscopique

Ce parallélisme est la conséquence du caratère bayesien du modèle de calcul, i.e. du fait que les états compatibles sont indépendants deux à deux. Dès lors, le calcul peut être effectué pour chaque élément d'une partition de Ω indépendamment des autres. L'efficacité de ce parallélisme est directement fonction de l'homogénéité des cardinaux des éléments de la partition. Si de plus chaque élément de la partition peut être généré indépendamment des autres et par un même algorithme, on peut paralléliser toute la comparaison.

Cette parallélisation microscopique sera optimale si on dispose d'un algorithme permettant de générer une partition de Ω en P éléments de même cardinal et si cet algorithme exploite le fait que Φ est un homomorphisme de E sur Q.

Le code de Gray décrit dans la section 4.2 est donc un excellent candidat pourvu que l'on sache exprimer le vecteur d'indices associé au $q$ième état généré. La partition utilisée serait alors $(e_{(kE[N_\Omega/P])}, \cdots, e_{((k+1)E[N_\Omega/P])})_{0 \leq k \leq P}$ où l'ordre est l'ordre de génération du code de Gray et P est le nombre de processeurs. Cette solution est à l'étude.

4.5 Conclusion

Le code C, relativement petit, présente l'intérêt d'être encore facilement modifiable et donc de disposer d'un exemple significatif de programmme en développement, très consommateur de CPU, écrit en tenant compte de l'évolution récente des calculateurs parallèles, sur lequel on peut appliquer diverses stratégies d'implémentation.

5 Applications

Le modèle développé dans ce qui précède est utilisé :

- pour l'estimation de la longueur des clones à partir des empreintes.
- pour le calcul des probabilités de recouvrement entre paires de clones.

Chacun de ces deux aspects a été testé sur des **données simulées** et sur des **données réelles**.

Les données simulées sont obtenues en plaquant sur le modèle précédemment décrit un certain nombre de processus d'erreur identifiés dans la génération des données mais non explicites dans le modèle (détection aléatoire de fragments supplémentaires et non-détection aléatoire de fragments réels). Les longueurs et les recouvrements calculés par le programme sont en bon accord avec les valeurs exactes, et ceci pour la majorité des processus d'erreur ajoutés.
Néanmoins, les limites de résolution des systèmes de saisie provoquent la fusion de fragments de tailles voisines en un seul fragment, avec une fréquence qui croît avec le nombre de bandes détectées. Avec les séquences ALU, ce phénomène – dit de comigration – ne peut plus être négligé et doit être incorporé au modèle pour le calcul de l'expression 5.

En ce qui concerne les données réelles, la première application du programme est la détection d'incohérences entre les empreintes générées avec des enzymes de restriction différentes. Cette étape de contrôle de qualité permet d'écarter des empreintes qui causeraient des erreurs dans la carte finale (voir graphiques 6). Après ce contrôle de qualité, les probabilités de recouvrement sont calculées et l'on obtient, pour chaque paire de clones, des familles de courbes dont deux exemples sont donnés dans la figure 5. Ces probabilités sont calculées en utilisant comme longueur des clones la moyenne des estimations à partir des empreintes.

Pour un échantillon de clones dont les longueurs ont été déterminées expérimentalement, ces longueurs ont aussi été estimées à partir des empreintes en utilisant le programme. La confrontation de ces valeurs a montré:

- que la comigration provoque effectivement une saturation de l'estimation pour les empreintes ALU, ainsi qu'une sous-estimation systématique de la longueur des clones. L'utilisation des longueurs estimées pour le calcul des probabilités de recouvrement manque donc de robustesse, et une intégration sur la distribution *a priori* de la longueur des clones doit être implémentée.

- que seule l'intégration des probabilités sur une distribution *a priori* de la valeur de α garantit une robustesse véritable des résultats vis-à-vis des inhomogénéités dans la distribution des séquences répétées.

Les efforts futurs tendront donc à enrichir le modèle (intégration sur les distributions *a priori* de la longueur des clones et du paramètre α) tout en conservant des temps de calcul compatibles avec l'échelle du problème évoquée dans la section 1.

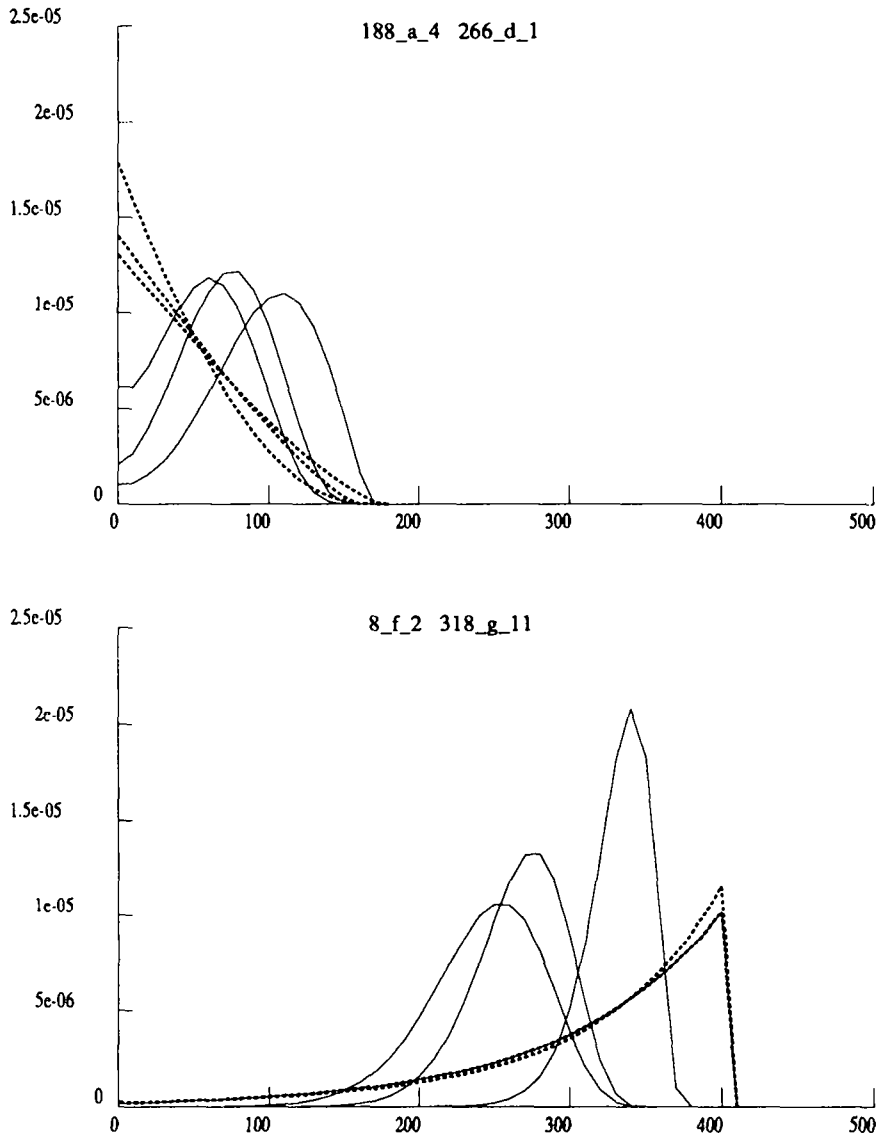

Figure 5: Densités de probabilité de recouvrement entre deux paires de clones estimées à partir des empreintes ALU (trait plein) et KPN (trait pointillé), et pour trois enzymes différentes. Les clones 8_f_2 et 318_g_11 sont recouvrants, alors que les clones 188_a_4 et 266_d_1 ne le sont pas. Les longueurs de recouvrement sont exprimées en kilobases.

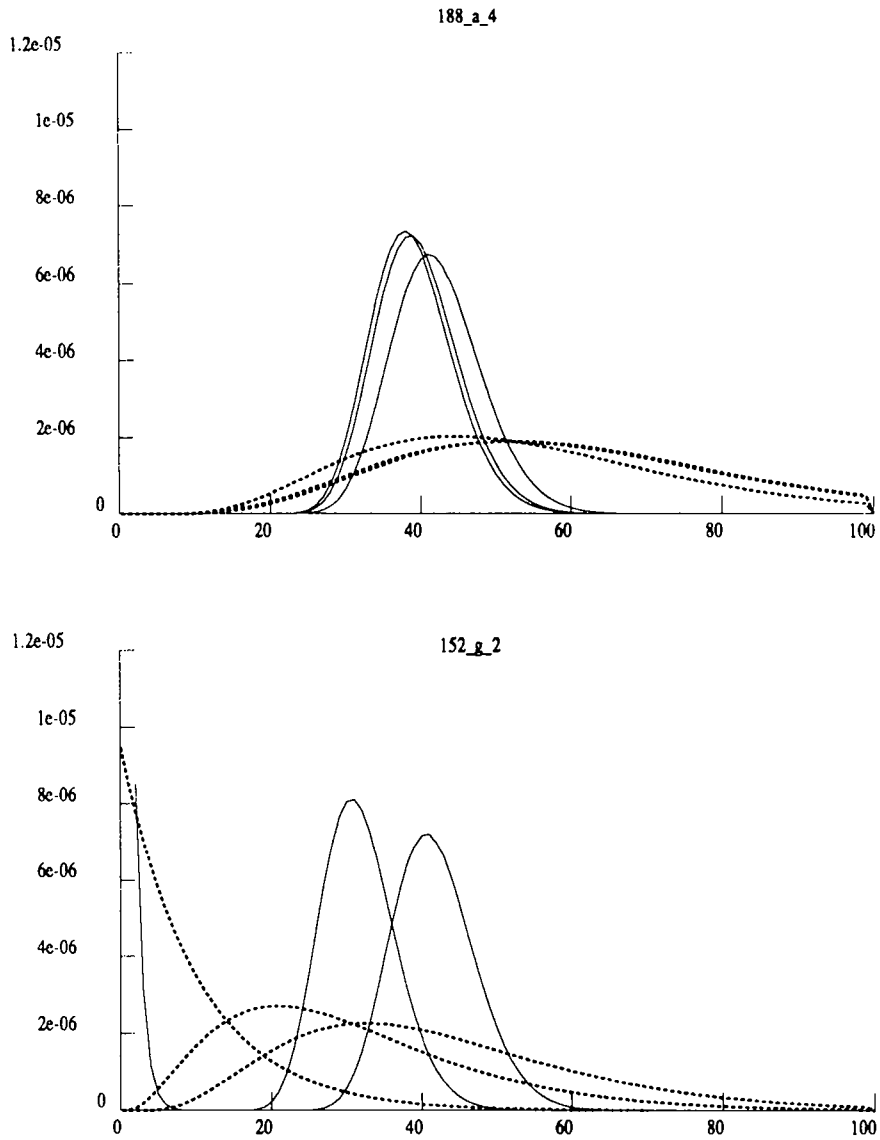

Figure 6: Densités de probabilité de longueur de deux clones estimées à partir des empreintes ALU (trait plein) et KPN (trait pointillé), et pour trois enzymes différentes. Alors que ces empreintes sont cohérentes pour le clone 188_a_4, une empreinte ALU et une empreinte KPN du clone 152_g_2 sont très atypiques. Ces incohérences traduisent des problèmes de détection. Les longueurs sont exprimées en dizaines de kilobases.

Glossaire

ADN : Acide DesoxyriboNucléique. Enchainement ordonné de nucléotides sous forme de deux brins complémentaires. Dépositaire de l'information génétique de la cellule et vecteur de celle-ci au travers des générations cellulaires.
Nucléotide : Enchaînement d'un sucre, d'un groupe phosphate et d'une base azotée. Les bases sont au nombre de quatre et dénommées A (Adénine) T (Thymine) G (Guanine) C (Cytosine). Elles peuvent s'apparier deux à deux (A avec T et G avec C). L'enchainement des bases constitue la **séquence nucléique**.
Gène : ensemble des séquences d'acides nucleiques contenant l'information pour la production régulée de proteines (expression d'un gène).
Séquence codante : séquence nucléique, partie d'un gène, participant directement à la synthèse des proteines.
Polymorphisme : variation individuelle de la séquence en bases du génome.
Phénotype : manifestation apparante de la constitution du génome sous la forme d'un trait morphologique, d'un syndrome clinique, d'une variation qualitative ou quantitative du produit final d'expression d'un gène (proteine).
Enzyme de restriction : Molécule clivant spécifiquement l'ADN au niveau d'une séquence parfaitement définie (site de restriction).
Locus : emplacement d'un segment d'ADN sur un chromosome défini par son contenu informationnel (gène) ou sa séquence qu'elle soit ou non polymorphe (segment anonyme).
Sonde : séquence d'acide nucléique, homologue à une séquence d'ADN avec laquelle elle s'hybride de façon stable et spécifique par réassociation entre bases complémentaires.
Hybridation moléculaire : appariement par complémentarité des bases de deux séquences nucléotidiques complémentaires.
Recombinaison : réassortiment de séquences d'ADN sur un même chromosome, résultant d'un échange de matériel avec son homologue au cours de la méiose.
Méiose : désigne les deux divisions cellulaires particulières qui constituent le stade ultime de la formation des cellules sexuées (gamétogénèse).
Clonage : recombinaison in vitro d'un gène, et par extension, d'un fragment d'ADN codant ou non, avec un vecteur se répliquant de manière autonome à l'intérieur d'une cellule hôte (permet l'amplification et l'isolement d'un fragment d'ADN à étudier).
YAC : Yeast Artificial Chromosome ou chromosome artificiel de levure. Vecteur navette bien adapté au clonage de fragments d'ADN de grande taille.
Séquence répétée : séquence nucleique dispersée dans la totalité du génome.

Deux familles de séquences particulières ont été caractérisées: les séquences Alu et Kpn. Ces séquences peuvent être utilisés comme sondes.
Electrophorèse : technique de séparation de macromolécules. Elle consiste à faire migrer les molécules dans un gel de polymères sous l'action d'un champ électrique. Pour les fragments d'ADN, la longueur de migration est inversement proportionnelle à la longueur du fragment.

Remerciements

Remerciements à Guy Vaysseix pour son soutien constant et pour sa vaste connaissance d'Unix ainsi qu'à Robert Ehrlich pour son aide précieuse à l'implémentation du code de Gray étendu.

References

[AAC+87] H.M. Albertsen, H. Abderrahim, H.M. Cann, J. Dausset, D. Le Paslier, and D. Cohen. Construction and characterization of a yeast artificial chromosome library containing seven haploid genome equivalent. *Proc. Natl. Acad. Sci. USA*, pages 4256–4260, 1987.

[BCBL+91] C. Bellané-Chantelot, E. Barillot, B. Lacroix, D. Le Paslier, and D. Cohen. A test case for physical mapping of human genome by repetitive sequence fingerprints: construction of a physical map of a 420 kb yac subcloned into cosmids. *Nucleic Acids Research*, pages 505–510, 1991. Vol. 19, No. 3.

[BT91] D. J. Balding and D. C. Torney. Statistical analysis of dna fingerprint data for ordered clone physical mapping of human chromosomes. *submitted to Bull. Math. Biol*, 1991. Personal Communication.

[LW88] E.S. Lander and M.S. Waterman. Genomic mapping by fingerprinting random clones. *Genomics 2*, pages 231–239, 1988.

[MTM+89] R.K. Moysis, D.C. Torney, J. Meyne, J.M. Buckingham, J.R. Wu, C. Burks, K.M. Sirotkin, and W.B. Goad. The distribution of interspersed repetitive dna sequences in the human genome. *Genomics 4*, pages 273–289, 1989.

ISSN 0249-6399