Modélisation de la dégradation de l'ADN pour la détermination du nombre de copies restantes et de la probabilité de PCR positive

Marthe Colotte, Vincent Couallier, Sophie Tuffet, Jacques Bonnet

MODELISATION DE LA DEGRADATION DE L'ADN POUR LA DETERMINATION DU NOMBRE DE COPIES RESTANTES ET LA PROBABILITE DE PCR POSITIVE

Marthe Colotte[1,3], Vincent Couallier[2,] Sophie Tuffet[3] et Jacques Bonnet[1,4]

[1] Université de Bordeaux; Plate-forme génomique fonctionnelle
146 Rue Léo Saignat; 33076 Bordeaux cedex; France
[2] Université de Bordeaux; Institut de Mathématiques de Bordeaux UMR 5251;
146 Rue Léo Saignat; 33076 Bordeaux cedex; France
[3] Société IMAGENE; Parc Scientifique Unitec;
2 allée du Doyen Georges Brus; 33600 Pessac; France
[4] Université de Bordeaux; Institut Bergonie/ inserm U916
229 cours de l'Argonne; 33076 Bordeaux cedex; France

Summary: There is currently no method allowing routine characterization of minute amounts of degraded DNA samples such as those encountered in forensic science, archived tissues, ancient DNA, extracellular or stool DNA or processed food. Here, we describe and directly validate such a method based, on one hand, on a generalized DNA random fragmentation model and on the other, on two quantitative PCR experiments using two different target sizes. The model also makes it possible to determine the minimum sample amount, the minimum mass average fragment size and the maximum degradation time necessary to obtain a positive PCR.

Résumé: Il n'existe pas actuellement de méthode permettant de caractériser en routine des échantillons d'ADN dégradés comme ceux qui sont rencontrés en médecine légale, dans les tissus archivés, dans l'adn extracellulaire ou d'excréments, ou les produis alimentaires. Nous proposons et validons une méthode basée d'une part sur un modèle de fragmentation aléatoire de l'ADN et de l'autre sur deux analyses en PCR quantitative sur deux cibles de tailles différentes. Le modèle permet de caractériser l'échantillon dégradé en calculant de nombreux paramètres intéressant les biologistes et en particulier la quantité et la taille moyenne des fragments, mais aussi la probabilité d'avoir une PCR positive pour un temps de dégradation, une taille de cible et une quantité d'échantllon donnés.

1. Introduction

La caractérisation de molécules d'ADN dégradées repose principalement sur l'analyse PCR (Polymerase Chain Reaction : amplification exponentielle par polymérase). L'analyse par qPCR, pour "quantitative PCR" est le plus souvent utilisée pour le calcul du nombre de copies ou de la masse d'ADN dans des échantillons mais, en raison d'une non-amplification des séquences cibles ayant subi une coupure, l'analyse sous-estime sytématiquement la détermination de la masse effective ([1], [2], [3], [4]). En modélisant la dégradation de l'ADN dans le temps, et par suite en déterminant la distribution en taille de la population étudiée, on peut utiliser la PCR pour quantifier et évaluer la qualité d'un échantillon d'ADN, mais les modélisations courantes, faisant appel à un processus de Poisson ([5]) ne permettent pas d'extraire les caractéristiques essentielles aux biologistes qui s'intéressent à la conservation de l'ADN et à l'amplification d'ADN dégradé.

Dans cette présentation, nous fournissons à partir d'un modèle existant ([6]) plusieurs propriétés utiles :

- la répartition de la taille des fragments d'un échantillon d'ADN dégradé
- l'estimation de la probabilité de coupure à partir de deux analyses qPCR sur deux cibles.
- le calcul de la probabilité de PCR positive sur un échantillon dégradé.
- La validation du modèle de Moore et Maranas ([6]) à partir de deux jeux de de données réelles.

Ces propriétés sont issues de calculs simples et décrites de manière plus détaillées dans [7].

1. Modèle de fragmentation et estimation de la probabilité de coupure

On modélise un brin d'ADN par une chaîne de B nucléotides. Chacune des B-1 liaisons nucléotide-nucléotide a une même probabilité (inconnue) de coupure notée P_{cut}. La digestion d'un brin d'ADN de longueur B et soumis à des facteurs de coupure produit donc aléatoirement un nombre inconnu de fragments dont la répartition des longueurs à été établie par [6]. Partant d'un échantillon de brins d'ADN identiques de longueur B, la répartition des longueurs de fragments dans l'échantillon dégradé est

$$Q_L = \begin{cases} (1-P_{cut})^{B-1} & , \text{ pour } L = B \quad (1) \\ P_{cut}.(1-P_{cut})^{L-1} & , \text{ pour } 1 \le L \le B\text{-}1 \quad (2) \end{cases}$$

Pour L=1..B, Q_L est la proportion de fragments dégradés de longueur L dans l'échantillon après digestion. Q est en fait la distribution d'une loi géométrique tronquée : Soit X une variable aléatoire distribuée suivant une loi géométrique G(P_{cut}), ie $P(X=x)=P_{cut}.(1-P_{cut})^{x-1}\ x=1,2,...$ Dans notre cas, les premières probabilités Q_L coïncident avec la loi géométrique mais puisqu'il n'y a pas de séquences de longueur supérieure à B, toutes les probabilités restantes de la loi géométrique se concentrent en B pour donner $Q_B=\sum_{x=B}^{+\infty}P(X=x)=\sum_{x=B}^{+\infty}P_{cut}.(1-P_{cut})^{x-1}=(1-P_{cut})^{B-1}$. $Q_B=(1-P_{cut})^{B-1}$ est donc la probabilité qu'une séquence d'ADN de longueur B reste intacte après dégradation. Il est intéressant de noter que Q_L pour $L<B$ ne dépend pas de la taille initiale B. Le profil de fragmentation ne dépend donc que du paramètre de coupure P_{cut}.

La Fig.1 illustre la distribution Q (A) ainsi que le profil de l'intensité lumineuse (B) pour une analyse sur gel d'agarose : $M_L = L\ Q_L$, L=1...B. A partie d'un gel observé (ayant l'allure de la Fig.1 B), une méthode graphique permet également (théoriquement) d'estimer la probabilité de coupure P_{cut} par l'inverse de la longueur L maximisant M_L, méthode couramment employée dans la littérature ([8]). On montre facilement que le maximum est en fait atteint pour $L_{max} = -[1 / \ln(1-P_{cut})]$ qui est très proche de la valeur $1/P_{cut}$ habituelle.

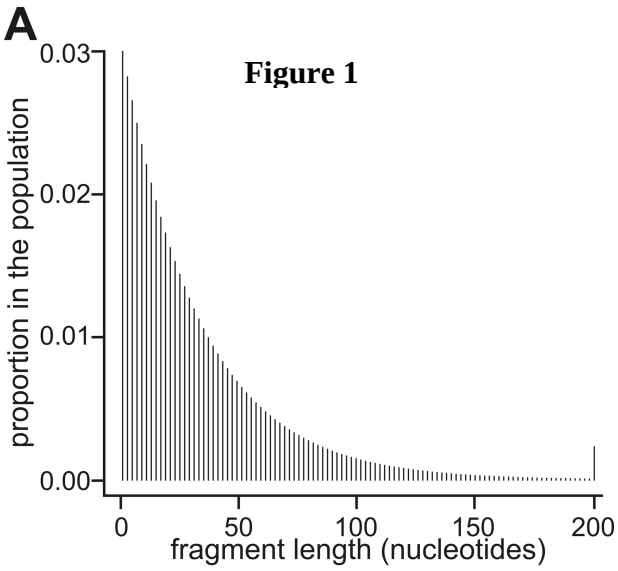
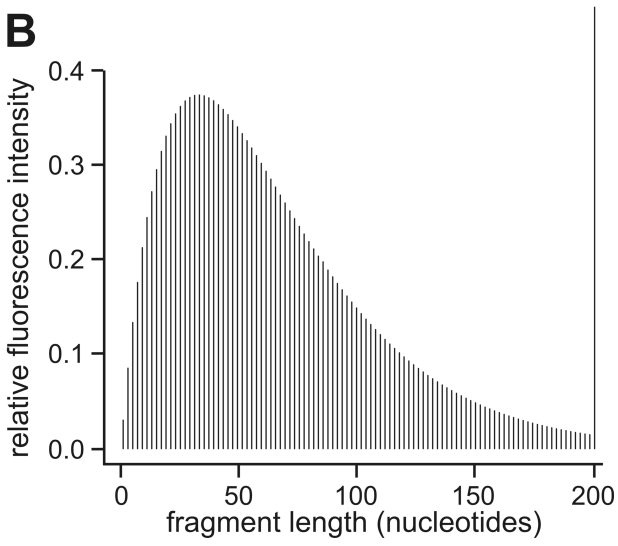


Figure 1

L'analyse qPCR permet, de déterminer le nombre de brins d'ADN dans un échantillon ayant une séquence donnée C_1 intacte. Le principe repose sur l'amplification des molécules d'ADN (dégradées ou non) à la seule condition que la cible C_1 n'ait pas été coupée lors de la dégradation. Il est évident que plus la cible à une longueur L_1 importante, plus la probabilité de coupure est grande. En effet d'après (1) pour $B=L_1$,

$$Q_{L1} = (1-P_{cut})^{L1-1}.$$



Considérons deux cibles distinctes de longueurs différentes L_1 et L_2. En appliquant deux analyses qPCR sur ces cibles, on obtient deux nombres (observés) N_1 et N_2 qui sont le nombre de molécules dans l'échantillon pour lesquelles ces cibles sont intactes. N_1 et N_2 sont des variables aléatoires binomiales $B(N, Q_{L1})$ et $B(N, Q_{L2})$ respectivement. Les proportions théoriques Q_{L1} et Q_{L2} ne peuvent être estimées par N_1/N et N_2/N puisque le nombre initial de copies N est inconnu. Cependant, le ratio N_1/N_2 est un estimateur de $(1-P_{cut})^{L_1-1}/(1-P_{cut})^{L_2-1} = (1-P_{cut})^{L_1-L_2}$. Ainsi , il est possible d'estimer P_{cut} :

$$\widehat{Pcut} = 1-(N_1/N_2)^{1/(L_1-L_2)} \quad (3)$$

De plus, une estimation du nombre de copies initiales N à partir de N_1 et N_2 est :

$$\widehat{N} = N_1/(1-\widehat{P_{cut}})^{L_1-1} = N_2/(1-\widehat{P_{cut}})^{L_2-1} = \left(\frac{N_2^{L_1-1}}{N_1^{L_2-1}}\right)^{\frac{1}{L_2-L_1}} \quad (4)$$

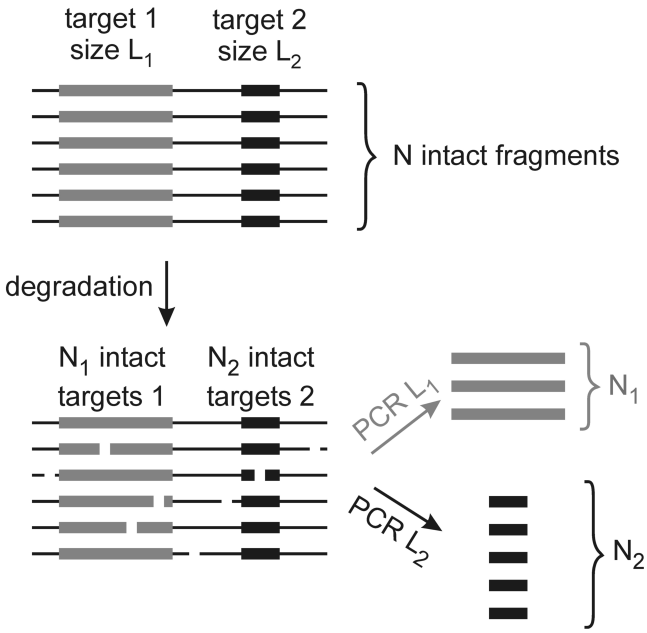


Fig.2 : schéma d'une double qPCR sur deux cibles différentes de longueurs L_1 et L_2

2. Détermination de la vitesse de coupure par une double qPCR en deux temps différents.

L'attaque de la liaison nucléotide-nucléotide a d'autant plus de probabilité d'être observée que le temps de dégradation est important. La plupart des modèles supposent une loi exponentielle pour la durée de vie d'une liaison, ce qui correspond à une variation de P_{cut} en fonction du temps du type

$$P_{cut} = 1 - exp(-kt) .$$

A partir de deux doubles qPCR à deux dates t_1 et t_2, on obtient l'estimation de la vitesse de coupure

$$\hat{k} = \frac{1}{(t_2-t_1).(L_1-L_2)} \ln\left(\frac{N_1^{t_1}}{N_2^{t_1}} * \frac{N_2^{t_2}}{N_1^{t_2}}\right)$$

Le modèle permet également d'estimer la probabilité qu'une qPCR sur une cible de longueur L donnée effectuée à un temps donnée soit positive. Pour cela, il est nécessaire que l'échantillon

dégradé contienne au moins une copie dont la séquence ciblée soit intacte. Cette probabilité est donnée par

$$\begin{aligned} P(\text{PCR sur cible de longueur L}) &= P(\text{au moins une cible intacte}) \\ &= 1 - P(\text{toutes les cibles sont dégradées}) \\ &= 1 - \left[1 - (1 - P_{cut})^{L-1}\right]^N \\ &= 1 - \left[1 - \exp(-kt)^{L-1}\right]^N \end{aligned}$$

Afin de valider le modèle, des échantillons contrôlés ont été testés et plusieurs qPCR ont été effectuées en des temps différents sur des cibles différentes. La procédure permet de corriger les méthodes classiques de calcul du nombre de copies dans des échantillons d'ADN dégradés. Enfin, le modèle est appliqué sur un jeu de données de la littérature ([9]) : l'extraction de l'ADN de dents de renard *Vulpes vulpes* collectées sur une période couvrant les trois dernières décennies a permis d'observer les probabilités de PCR positives sur des cibles de longueur différentes. Ces probabilités observées ont été confrontées au modèle proposé.

Bibliographie

[1] R.L. Green, I.C. Roinestad, C. Boland, and L.K. Hennessy, Developmental validation of the quantifiler real-time PCR kits for the quantification of human nuclear DNA samples. J Forensic Sci 50 (2005) 809-25.

[2] H. Andreasson, M. Nilsson, B. Budowle, H. Lundberg, and M. Allen, Nuclear and mitochondrial DNA quantification of various forensic materials. Forensic Sci Int 164 (2006) 56-64.

[3] J.G. Shewale, E. Schneida, J. Wilson, J.A. Walker, M.A. Batzer, and S.K. Sinha, Human genomic DNA quantitation system, h-quant: development and validation for use in forensic casework. J Forensic Sci 52 (2007) 364-70.

[4] W.R. Hudlow, M.D. Chong, K.L. Swango, M.D. Timken, and M.R. Buoncristiani, A quadruplex real-time qPCR assay for the simultaneous assessment of total human DNA, human male DNA, DNA degradation and the presence of PCR inhibitors in forensic samples: A diagnostic tool for STR typing. Forensic Science International: Genetics 2 (2008) 108-125.

[6] G.L. Moore, and C.D. Maranas, Modeling DNA Mutation and Recombination for Directed Evolution Experiments. J. theor. Biol. 205 (2000) 483-503.

[7] Colotte, M., Couallier, V., Tuffet, S. and Bonnet, J. (2009). Simultaneous assessment of average fragment size and amount in minute samples of degraded DNA, *Analytical Biochemistry*, *In Press, Corrected Proof, Available online 10 February 2009*

[8] B.M. Sutherland, A.G. Georgakilas, P.V. Bennett, J. Laval, and J.C. Sutherland, Quantifying clustered DNA damage induction and repair by gel electrophoresis, electronic imaging and number average length analysis. Mutation Research 531 (2003) 93-107.

[9] P. Wandeler, S. Smith, P.A. Morin, R.A. Pettifor, and S.M. Funk, Patterns of nuclear DNA degeneration over time - a case study in historic teeth samples. Mol Ecol 12 (2003) 1087-93.