Dynamique et synchronisme de réplication de l'ADN dans des cellules vivantes – Analyse de marqueurs fluorescents

Jean-François BERCHER[1], Bénédicte DURIEZ[2], Nicole BOGGETTO[3], Marie-Noëlle PRIOLEAU[2]

[1]Laboratoire d'Informatique Gaspard Monge, UMR CNRS 8049
Université Paris-Est, ESIEE, Cité Descartes 93162 Noisy-le-Grand Cedex

[2]Institut Jacques Monod, UMR CNRS 7592
Domaines chromatiniens et réplication
Université Paris Diderot, Paris

[3]Institut Jacques Monod, UMR CNRS 7592
ImagoSeine Pateforme d'imagerie
Université Paris Diderot, Paris

Jean-Francois.Bercher@univ-paris-est.fr
{duriez.benedicte, boggetto.nicole, prioleau.marie-noelle}@ijm.univ-paris-diderot.fr

Résumé – Cette communication présente l'analyse de données de biologie cellulaire en vue de l'étude du *timing* et de la synchronie de réplication des allèles. Les données disponibles sont acquises massivement par une technique d'imagerie cytométrique. L'analyse fait notamment intervenir la définition d'un modèle statistique et l'identification des paramètres d'un mélange de distributions. Les résultats sont validés par des tests statistiques et quantifiés par bootstrap. Du point de vue biologique, des nouveaux mécanismes intervenant dans la réplication sont exhibés.

Abstract – This communication deals with the analysis of cellular biology data, for the analysis of replication timing and synchonization between allels. Data are obtained via cytometric imaging. The analysis include the development of a statistical model, the estimation of the parameter of a mixture of distributions. Results are validated through statistical tests and quantified using a bootstrap technique. From the biological point of view, new mechanisms involved in replication are exhibited.

1 Position du problème

La problématique abordée ici concerne un processus fondamental dans le monde du vivant : la duplication des molécules d'ADN qui précède chacune des divisions cellulaires. Dans les organismes eucaryotes cette duplication s'exécute selon un programme spatio-temporel dont les caractéristiques font l'objet de nombreuses recherches. La correcte exécution de ce programme est extrêmement importante pour la vie et la santé puisque des erreurs peuvent entraîner soit une mort cellulaire, soit un processus de tumorigenèse, soit des mutations. Brièvement, au cours de la phase de synthèse de l'ADN (*phase S*), la duplication s'initie en différents points le long de la molécule d'ADN. Ainsi, le long du chromosome des domaines « temporels » (de *timing*) se succèdent. Cette particularité définit le programme temporel de la réplication. Dans le travail présenté ici, nous souhaitons évaluer à quel point l'exécution de ce programme temporel est précise et strictement contrôlée.

Actuellement, les approches expérimentales classiques qui déterminent le moment de réplication ne distinguent jamais les deux chromosomes homologues et ne peuvent donc pas avoir accès à la notion de synchronie. Ces approches utilisent des populations de milliers de cellules, et les moments de réplication sont alors à très faible résolution temporelle.

Pour aborder la question de la précision et de la synchronisation du programme temporel, nous avons choisi de travailler sur des cellules individuelles dans lesquelles nous comparons les moments de réplication de chacun des deux chromosomes homologues dans des domaines temporels bien spécifiques. Trois régions représentatives du déroulement de la phase S ont été choisies : la première est précoce (*Early*), la seconde est moyenne-tardive (*Mid-Late*) et la dernière est tardive (*Late*). Pour déterminer le moment de réplication de la région étudiée, nous incorporons à cet endroit précis un petit marqueur fluorescent dont l'intensité est doublée au moment de la duplication. Pour observer les cellules nous utilisons le système *ImageStream d'Amnis (EMD Millipore)* qui combine la cytométrie en flux à de l'imagerie à haute-résolution (spatiale). Ce système permet d'acquérir rapidement les images de plusieurs milliers de cellules individuelles. La région étudiée est visualisée comme un spot de petite taille (diamètre de 0.6 à 1 µm, s'étendant sur 2 à 4 pixels) cf Figure 1, dont l'intensité de fluorescence dépasse le bruit de fond d'un facteur d'environ 3.5 ; mais il reste probablement des faux positifs.

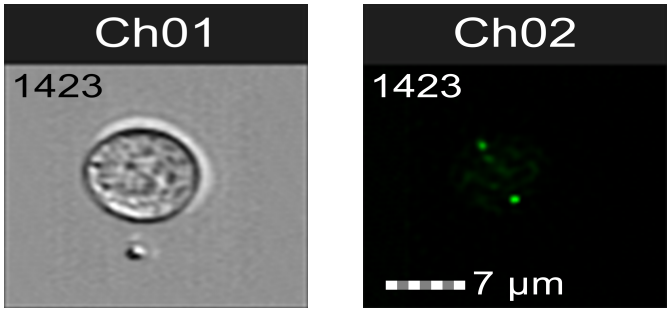

FIGURE 1 – Visualisation avec ImageStream des deux allèles marqués par des molécules fluorescentes. A gauche, la cellule est visualisée dans le visible (Brightfield), à droite le noyau et les deux allèles marqués sont visualisés par leur émission de fluorescence. La barre représente 7 µm.

Un histogramme des intensités des deux spots est ainsi présenté Figure 2. Lors de l'analyse préalable, nous avons pu mesurer une corrélation de l'ordre de 0.3 entre les deux jeux de données. Ceci peut être lié à un recouvrement des spots. Pour corriger ceci, nous avons alors mis en place une procédure de séparation de sources positives. L'implantation de l'algorithme est disponible sous https://gist.github.com/jfbercher/. Bien que la procédure semble parfaitement fonctionnelle, la séparation n'apporte pas de gain en performances pour les traitements ultérieurs.

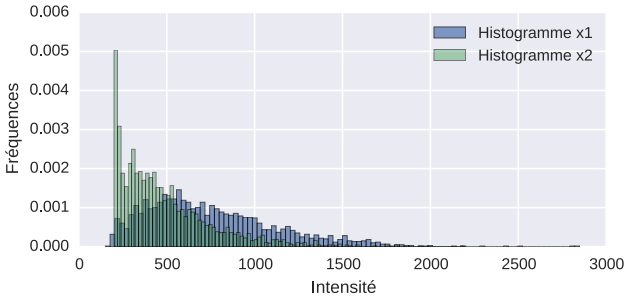

FIGURE 2 – Histogrammes des données d'intensité des deux spots.

Du fait de la grande variance des observations (taux de fluorescence) d'une cellule à une autre, il n'est pas possible de discriminer les spots à partir de leur seule intensité. Nous nous proposons alors de former le rapport (*ratio*) des deux intensités et d'étudier le comportement de ce rapport au cours de la réplication. En théorie, dans des cellules avec deux allèles qui ne sont pas encore dupliqués, ou au contraire avec deux allèles déjà dupliqués, le ratio a une valeur de un. En revanche, dans des cellules où la duplication des deux allèles n'est pas synchrone, le ratio a une valeur de 2. Expérimentalement, la valeur observée est plutôt de l'ordre de 1.6.

Pour évaluer le degré de synchronie de duplication des allèles, nous cherchons à mesurer les proportions des ratios « un » et « deux » dans six échantillons (de 1000 à 8000 cellules) reflétant le déroulement de la phase S. Ceci nous permet d'obtenir une résolution temporelle de plus du double des observations classiques.

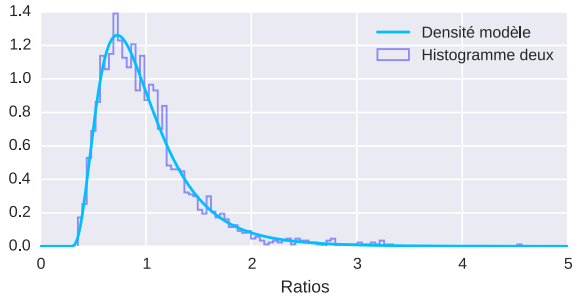

FIGURE 3 – Histogramme des rapports d'intensité et densité estimée. Paramètres : $a = 125.21, b = 13.50, x_0 = 0.04, s = 0.82$. Le modèle est validé par un test de Kolmogorov-Smirnov.

Le problème de traitement du signal est donc, à partir de mesures d'intensité très bruitées et à faible résolution spatiale, de modéliser statistiquement ces données, d'évaluer les proportions respectives des deux classes d'intérêt, d'étudier l'évolution temporelle et de quantifier statistiquement les résultats obtenus.

2 Modèle statistique

2.1 Le rapport des intensités

Cherchant à repérer la réplication de l'ADN, nous devons nécessairement comparer les intensités des deux spots cellule par cellule. Pour ce faire, on se propose de former le rapport des deux intensités et d'étudier le comportement de ce rapport au cours de la transcription. Un modèle standard pour une variable d'intensité est une variable du χ^2. Le rapport de deux variables du χ^2 indépendantes est une variable de Fisher de paramètres a et b et de densité

$$f_{a,b}(x) = \frac{a^{\frac{b}{2}} b^{\frac{a}{2}}}{\mathrm{B}\left(\frac{a}{2}, \frac{b}{2}\right)} x^{\frac{a}{2}-1} (b + ax)^{-\frac{a+b}{2}}.$$

Nous proposons donc d'approcher la loi du rapport par une loi de Fisher. En plus des paramètres a et b, on peut aussi inclure des paramètres de localisation et d'échelle, qui correspondent à la transformation $X \to sX + x_0$. Ces paramètres sont estimés par maximum de vraisemblance à partir du rapport des intensités. Afin de limiter l'apparition de minimas locaux, le paramètre de localisation est gardé fixe.

Un exemple de résultat est présenté Figure 3, où l'on donne à la fois l'histogramme des rapports d'intensité et le modèle estimé sous la forme d'une densité d'une loi de Fisher. Afin de juger de la qualité du résultat, nous avons réalisé un test d'hypothèse de Kolmogorov-Smirnov. Celui-ci accepte bien le modèle obtenu, au niveau $\alpha = 5\%$, avec une valeur de test de 0.6924 pour un seuil théorique de 1.2202 ; la p-value correspondante est de 0.72. Le programme de calcul du test de Kolmogorov-Smirnov est disponible sous https://gist.github.com/jfbercher.

2.2 Modèle de mélange

L'hypothèse que nous voulons tester et quantifier est la présence de cellules en cours de duplication dans l'échantillon global. Plus exactement, nous nous intéressons au retard de duplication entre les deux allèles. Dans ce cas, si l'un des allèles est répliqué et l'autre non, le rapport des intensités des deux spots est voisin de deux. Ceci est fugace, dans la mesure ou le second allèle se réplique un peu plus tard. Ainsi, l'échantillon doit présenter une grande majorité de ratios autour de un, correspondants aux spots d'intensités voisines, et une sous population en cours de duplication non synchrone. C'est la proportion de cette population que nous cherchons à évaluer. Nous avons donc un modèle de mélange. Notons π la proportion de la population 2. Le ratio observé est alors soit une variable de Fisher de moyenne 1, avec une probabilité $(1-\pi)$, soit une variable de Fisher de moyenne plus élevée (théoriquement 2, plus faible en pratique – nous utilisons 1.6). Il est raisonnable de supposer que les deux lois de Fisher sont de même paramètres et ne diffèrent que par le paramètre d'échelle. Dans ce cas, la loi de mélange a la forme suivante :

$$f_{\text{mixture}}(x) = (1-\pi) f_{a,b,x_0,s}(x) + \pi f_{a,b,x_0,2s}(x).$$

3 Traitement des données

Le problème est donc d'identifier le paramètre de mélange et d'évaluer la qualité du résultat. Classiquement, l'identification des paramètres d'un mélange de distribution peut être abordée avec un algorithme EM. Dans le cas qui nous occupe, le problème est fortement simplifié car les paramètres de forme a, b, les paramètres de localisation et d'échelle peuvent être identifiés à partir d'une population pure de référence (que nous sélectionnons en début de phase S). Il reste donc simplement à identifier le paramètre de mélange π. Ceci est réalisé par une maximisation directe de la vraisemblance, en utilisant un algorithme de type BGFS. Il serait même possible ici, puisqu'on a un seul paramètre à déterminer, de procéder par une recherche exhaustive sur une grille (*gridsearch*). La Figure 4 présente un exemple de résultat pour l'identification des paramètres de la loi et du paramètre de mélange.

A nouveau, le modèle peut être testé en utilisant un test de Kolmogorov-Smirnov qui compare la répartition empirique à la répartition hypothétique. Pour des données de taille 1000, le test est validé, avec une p-value comprise entre 0.1 et 0.4.

A partir de ces éléments, il est ensuite possible de construire un test pour classifier les cellules élémentaires dans chacune des deux classes. Ceci permet a posteriori une inspection visuelle (au microscope) des résultats, et éventuellement un retour vers l'algorithme pour l'optimisation des paramètres. Pour ce faire, on utilise une simple discrimination bayésienne. On calcule le score des deux hypothèses que l'on compare à un seuil. Nous avons ici un a priori de $(1-\pi)$ pour les cellules de ratio 1 et de π pour les cellules de ratio 2. Pour des coûts

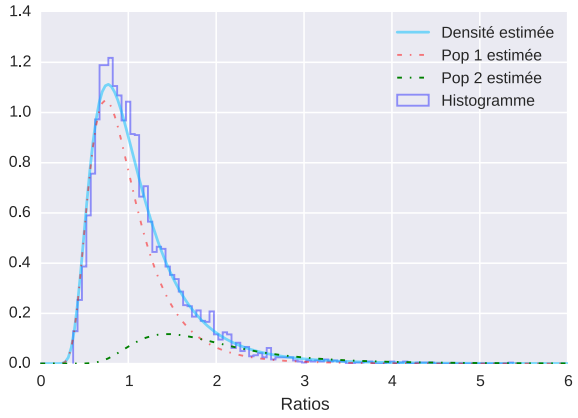

FIGURE 4 – Identification du mélange de distributions. Paramètres : $a = 125.21, b = 13.50, x_0 = 0.04, s = 0.82$ – On identifie une proportion π de 0.18 pour la population '2', en duplication asynchrone. Pour un sous-échantillon aléatoire de taille 1000, le test de Kolmogorov-Smirnov valide le modèle.

standards, le test bayésien est

$$T(x) = \frac{f_{a,b,x_0,s}(x)}{f_{a,b,x_0,2s}(x)} \gtrless_2^1 \frac{\pi}{1-\pi},$$

où x représente le ratio des intensités des deux spots dans la cellule considérée. En appliquant ceci à l'exemple précédent, on peut ainsi extraire 414 cellules sur un total de 4800. Il est bien entendu possible de faire varier le seuil afin d'accroître le taux de détection (mais au détriment d'une augmentation des faux positifs).

Afin d'évaluer la qualité statistique de l'ensemble, dans la mesure où nous ne maîtrisons pas toutes les sources d'incertitude et d'erreurs de modèle, nous avons appliqué une technique de ré-échantillonnage bootstrap (sur l'ensemble de la procédure, comprenant l'estimation des paramètres de la loi modèle et du paramètre de mélange), qui permet d'estimer numériquement la loi de notre estimateur. Pour l'exemple précédent, où nous avons estimé $\hat{\pi} = 0.1758$, nous obtenons ainsi que l'estimateur est distribué approximativement de manière gaussienne, avec une moyenne de $\bar{\hat{\pi}} = 0.1762$ et un écart-type de $\hat{\sigma} = 0.020$. Le biais est ainsi négligeable, et on peut établir un intervalle de confiance bootstrap pour nos différentes estimées.

On étudie l'évolution de différentes lignées cellulaires au cours de la phase S. Les mesures sont effectuées sur 6 fractions cellulaires, qui correspondent approximativement à un échantillonnage temporel. La première fraction, qui correspond à l'état des cellules avant le début de la réplication, est utilisée comme échantillon de référence, afin d'établir la loi modèle. La phase de réplication durant 7 heures, les points de temps sont séparés de 84 minutes. Nous présentons ici quelques uns des résultats obtenus. Les résultats sont donnés ici pour plusieurs lignées : une lignée étiquetée comme précoce (*early*), une lignée tardive (late) et une lignée de contrôle précoce-mi-tardive (*early, mid-late*). La table 1 présente les statistiques bootstrap obtenues pour l'échantillon 3 (4000 cellules) sur lequel nous

	t0	t1	t2	t3	t4	t5
$\hat{\pi}$	0.0004	0.0958	0.1758	0.0842	0.0632	0.0248
$\bar{\hat{\pi}}$	0.0088	0.0961	0.1763	0.0836	0.0637	0.0260
$\hat{\sigma}$	0.0131	0.0241	0.0197	0.0146	0.0146	0.0158

TABLE 1 – Échantillon 3 (lignée Late X2) — Table des estimées du paramètre de mélange, des moyennes et écart type obtenus par bootstrap. Cette table correspond à la figure 6.

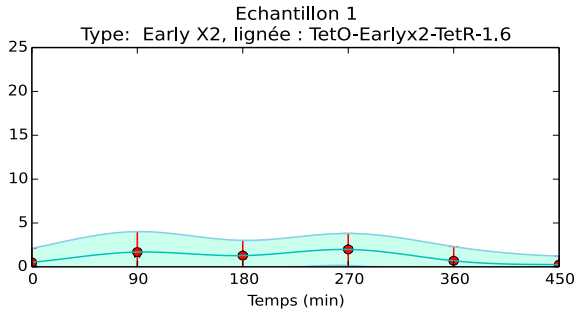

FIGURE 5 – Échantillon 1 - Evolution temporelle du taux de cellules asynchrones lors de la réplication.

avons travaillé. Les figures 5 à 7 présentent l'évolution temporelle du taux de cellules asynchrones lors de la réplication. Nous avons figuré les barres d'erreur correspondant à un intervalle de confiance à 95%. La courbe tracée a été obtenue par une interpolation par splines cubiques. De même, les bandes d'erreur ont été obtenues par interpolation. Si la représentation est agréable, il est bien entendu que seuls les points calculés et leurs intervalles de confiance ont un sens.

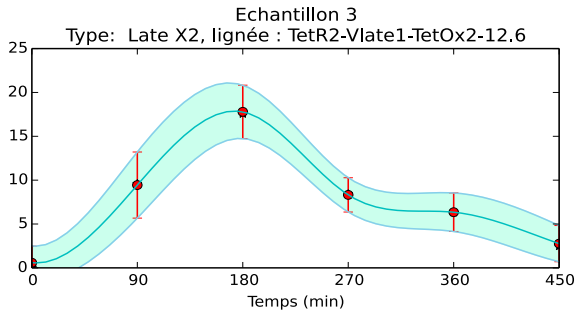

FIGURE 6 – Échantillon 3 - Evolution temporelle du taux de cellules asynchrones lors de la réplication.

4 Analyse des résultats et conclusions

Nous avons présenté le traitement de données de fluorescence, dans l'objectif de mettre en évidence et séparer deux populations de rapports d'intensité. Nous avons pour cela fait apparaître un modèle simple de mélange de distributions dont nous avons estimé les paramètres par maximum de vraisemblance. La tenue statistique de l'ensemble a été évaluée en utilisant la technique de bootstrap. Le suivi temporel pourra être

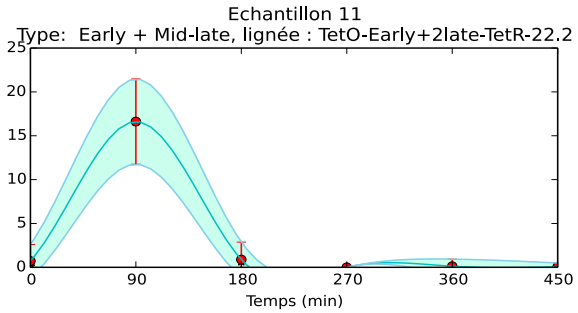

FIGURE 7 – Échantillon 11 - Evolution temporelle du taux de cellules asynchrones lors de la réplication.

complété par un modèle d'évolution markovien. Par ailleurs, avec le dispositif d'imagerie mis en œuvre, bien d'autres questions d'intérêt biologique pourront être abordées.

Du point de vue biologique, les résultats présentés ici nous ont permis de constater que la précision, et par conséquent le mode de contrôle, du programme temporel de la réplication n'est pas semblable en début et en fin de la phase S. La détection d'une très faible proportion de cellules de ratio 2 dans les lignées analysant des domaines temporels précoce et moyen-tardif traduit une forte synchronie de réplication des deux allèles et suggère que le programme est contrôlé très strictement. En revanche, dans la lignée marquée dans un domaine tardif, la détection d'une proportion significativement plus forte de cellules asynchrones tout au long de la phase S révèle une plus faible synchronisation de réplication des deux allèles et montre que le contrôle temporel de la réplication est très différent en fin de phase S. Notamment, une fraction des allèles se réplique dès le début de la phase S. Ce résultat n'était pas prévisible dans l'état actuel des connaissances sur le sujet, et les méthodes classiques reposant sur l'analyse globale de milliers d'allèles mélangés ne permettent pas d'observer cette particularité. L'utilisation d'une autre approche d'imagerie à haute résolution (spatiale) permettant le suivi en temps réel de cellules uniques pendant la phase S confirme les résultats obtenus ici. Cette autre méthode permet d'atteindre une résolution temporelle extrêmement fine (précision de 5 min sur les 420 min de la phase S), et une mesure fiable des intensités. Toutefois elle s'avère longue et fastidieuse et ne permet d'analyser raisonnablement que quelques dizaines de cellules par lignée. C'est pourquoi l'utilisation du système ImageStream avec ce type d'analyse des données est particulièrement intéressante, puisqu'elle permet l'évaluation rapide du degré de synchronie d'une lignée. En conclusion, les résultats obtenus sont très intéressants, bien qu'ils restent à confirmer par l'étude d'autres domaines de réplication tardive. Le déroulement du programme de réplication dans ces domaines tardifs s'avère être moins bien compris aujourd'hui que dans les domaines précoces ou moyen-tardifs. Or, ces domaines tardifs sont plus souvent sujets à des erreurs de réplication que les autres domaines. Ils sont donc plus souvent impliqués dans des processus pathologiques. Progresser dans la compréhension du contrôle spécifique de leur réplication est donc très important.