Des statistiques pour détecter les altérations chromosomiques

Emilie Lebarbier, *maître de conférences*
Stéphane Robin, *directeur de recherche INRA, à AgroParisTech*

Les altérations chromosomiques sont responsables de nombreuses maladies, parmi lesquelles certains cancers. La détection de petites altérations, essentielle pour le diagnostic du médecin, fait appel à un modèle classique en statistiques : la segmentation.

Dans chaque cellule du corps humain, chaque chromosome est présent en deux exemplaires ou *copies*. Un écart à cette règle (*perte* ou *gain*) entraîne un déséquilibre pouvant être à l'origine de certaines maladies. La *perte* correspond à l'absence ou à la présence en une seule copie du chromosome et le *gain* à la présence de trois, quatre copies, voire plus encore. Ce déséquilibre du nombre de copies est appelé *altération chromosomique*. La plus connue est la trisomie, pour laquelle un chromosome est présent en trois copies. Ainsi, le syndrome de Dawn (aussi appelé *mongolisme*) est dû à la présence de trois copies complètes du 21e chromosome. Il est possible de détecter ce type d'altérations à l'aide du traditionnel *caryotype*.

D'autres maladies sont liées à des altérations qui ne touchent pas des chromosomes entiers mais seulement des portions de chromosomes. Certaines portions sont « perdues » (absentes ou présentes en une seule copie) ou « amplifiées » (présentes en trois ou quatre copies, voire plus encore). C'est notamment le cas d'un grand nombre de cancers. Les cellules du tissu malade présentent typiquement des pertes de régions contenant des gènes « suppresseurs de tumeurs » ou, au contraire, des amplifications de régions contenant des *oncogènes* qui favorisent le développement tumoral.

Quand elles sont trop petites, ces altérations peuvent malheureusement passer inaperçues dans l'étude du caryotype. Leur détection est donc un enjeu essentiel de la recherche médicale afin d'identifier les gènes affectés, d'essayer de comprendre leur implication dans le développement de la maladie et de concevoir ainsi des thérapies adaptées.

La détection des altérations chromosomiques est un enjeu essentiel de la recherche médicale.

Évaluer le nombre de copies d'une portion de chromosome

C'est dans les années 1990 que l'étude de ces petites altérations a été rendue possible grâce à l'arrivée de techniques de biologie moléculaire. Le principe consiste à compter le rapport du nombre de copies de différents gènes, dont la position (appelée *locus*) sur le génome est connue, entre un individu malade et un individu sain. Ce principe est illustré sur un exemple très simple en figure 1a, pour cinq locus.

Il existe deux altérations chez l'individu malade (rouge) par rapport à l'individu sain (vert), qui possède toujours deux copies de chaque gène : une portion perdue au locus 2 qui donne un rapport du nombre de copies égale à 1/2 et un gain au locus 4 qui donne 3/2. Si l'on s'intéresse à beaucoup plus de locus (figure 1b), plusieurs locus successifs peuvent avoir le même statut biologique – normal, gain(s) ou perte(s) – et forment ainsi des *régions*. En réalité, il n'est pas possible de calculer directement le vrai nombre de

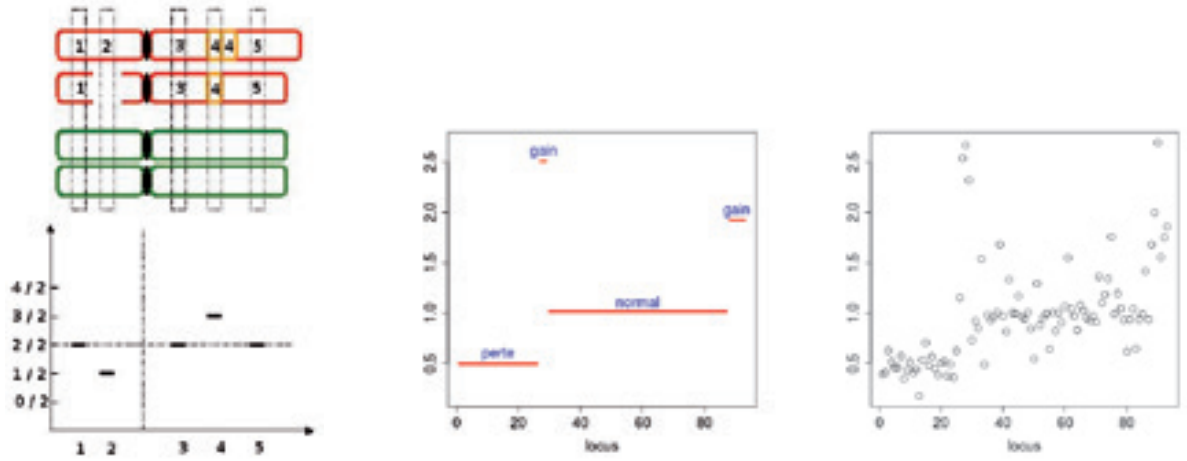


Figure 1. (a) Rapports « théoriques » entre le nombre de copies d'un individu malade et celui d'un individu sain (2 copies) suivant le numéro du locus pour cinq locus fictifs. (b) Ces rapports théoriques découpés suivant des régions, obtenus après travail statistique pour 93 locus réels. (c) Rapports « empiriques » de copies mesurés expérimentalement suivant le locus pour ces 93 locus.

copies mais les expériences biologiques vont permettre d'obtenir ce nombre à une erreur près (erreur de mesure, variabilité naturelle), comme le montre la figure 1c. Le rapport du nombre de copies ainsi obtenu pour les différents locus ne sera donc pas exactement égal à l'ensemble de valeurs théoriquement possibles 1/2, 1, 3/2, 2… ce qui rend alors impossible une analyse manuelle.

L'objectif statistique est de retrouver les « vraies » valeurs à partir des données que l'on a observées. En particulier, comme on l'a vu précédemment, il s'agit de détecter et de localiser automatiquement les différentes régions formées par plusieurs locus successifs ou un seul. Une fois l'analyse effectuée, le médecin pourra fonder son diagnostic sur la figure 1b plutôt que sur la figure 1c.

Un modèle mathématique

Comme souvent en statistique, la résolution de ce problème passe par la définition d'un modèle, c'est-à-dire, d'une traduction aussi fidèle que possible du processus biologique en langage mathématique (et suffisamment simple pour permettre une résolution mathématique). Un modèle classique en statistique dédié à la détection de régions (ou segments) est le modèle de *segmentation*. Il consiste à supposer que la vraie valeur du nombre de copies est la même au sein de chaque région, que cette valeur change quand on change de région, et que la valeur observée à un locus donné est égale à la vraie valeur de la région auquel il appartient plus un terme aléatoire, l'erreur. L'objectif statistique consiste alors à répondre à trois questions : Combien il y a de régions ? Où se trouvent-elles (c'est-à-dire quelles sont les limites entre les régions) ? Quelles sont les vraies valeurs au sein de chaque région ?

La troisième question se résout simplement : de façon naturelle, la vraie valeur de chaque région sera estimée par la moyenne des observations dans cette région. La résolution statistique des deux autres questions se déroule en sens contraire. Pour un nombre de régions fixé – notons-le k - on cherche le positionnement de ces régions le mieux ajusté aux données. C'est un problème de nature algorithmique. Sa résolution va nous fournir un algorithme qui nous donnera la « meilleure segmentation » des données pour k régions.

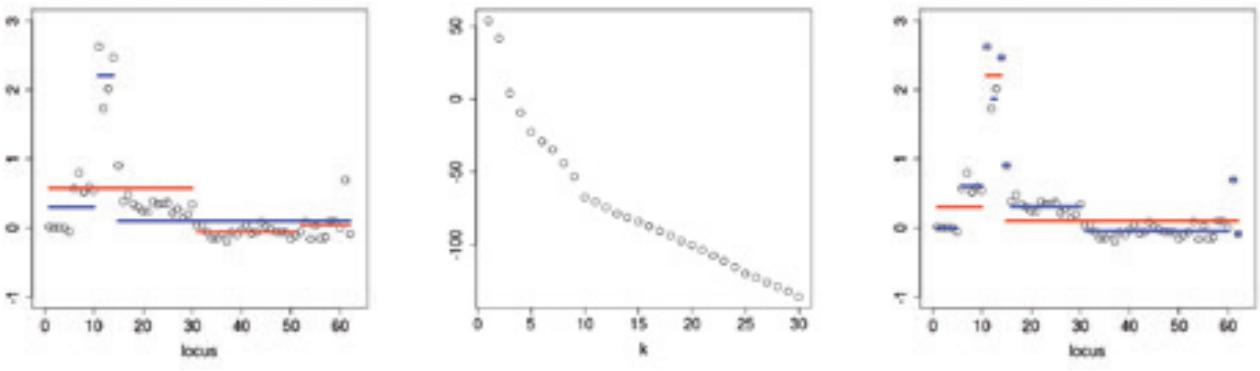
Bien sûr, en général, k n'est pas connu. Même si l'on dispose d'information a priori

sur les données, il est difficile de se le fixer à l'avance. Il faut donc le déterminer, ou plus exactement, choisir, encore une fois, le « meilleur » possible. Connaissant la « meilleure segmentation » des données pour différents k, il s'agira alors d'en prendre une parmi toutes. Cette fois, le problème est plutôt de nature statistique.

Comment obtenir la meilleure segmentation

Notons n le nombre de locus et fixons à k le nombre de régions. Il existe bien sûr plusieurs segmentations possibles des n locus en k régions et il s'agit de trouver la « meilleure ». La figure 2a présente deux exemples de segmentations, une bleue et une rouge, sur des données recueillies le long du chromosome 6 d'un patient atteint d'un cancer de la vessie. Entre ces deux segmentations, on préférera la segmentation bleue à la segmentation rouge. En effet, les données au sein des régions « bleues » (c'est-à-dire délimités par la segmentation bleue) sont moins dispersées autour de leur « vraie valeur » qu'au sein des régions « rouges ». La segmentation bleue s'ajuste donc mieux aux données que la segmentation rouge. On peut alors choisir ce critère d'ajustement, classique en statistique, comme mesure de qualité des segmentations. Reste à essayer toutes les segmentations possibles et à retenir la meilleure.

Un problème se pose alors : celui du nombre total de segmentations. On montre facilement que, si on observe n locus et qu'on cherche à segmenter les données en k régions, il existe autant de segmentations possibles que de combinaisons de $(k-1)$ objets parmi $(n-1)$, que l'on note $C(n-1, k-1)$. Par exemple, pour $n = 1000$



(a) (b) (c)

Figure 2. (a) Deux segmentations en trois régions. (b) Evolution de l'ajustement en fonction du nombre de régions. (c) Meilleures segmentations en trois et dix régions.

locus et k = 10 segments, il existe plus de 10^{138} segmentations possibles et même les ordinateurs les plus puissants ne peuvent pas explorer l'ensemble de ces combinaisons. Une solution existe cependant en reformulant légèrement le problème.

La meilleure segmentation pour un nombre de régions donné est celle qui s'ajuste le mieux aux données.

On définit le « coût » d'une région comme étant la dispersion des données autour de la moyenne de cette région. On appelle une « bonne » région celle qui a un faible coût. Sachant qu'une segmentation est un ensemble de k régions de la forme [1 ; t_1], [t_1+1 ; t_2]…, [t_{k-2}+1 ; t_{k-1}], [t_{k-1}+1 ; n], la « meilleure » segmentation correspond alors au chemin qui permet :

– d'aller du point 1 au point n,
– en faisant (k – 1) étapes à des points t_1, t_2, … t_{k-2}, t_{k-1} à déterminer,
– pour le plus faible coût total.

Un tel problème est un problème de « plus court chemin », que l'on sait résoudre avec un algorithme qui nécessite $k \times n^2$ opérations. Pour n = 1000 locus et k = 10 segments, il faut donc faire 10^7 opérations, ce qui est bien moins que 10^{138}. Sur la figure 2a, la segmentation bleue est la meilleure de toutes les segmentations possibles en trois régions.

Choisir le nombre de régions

Une première idée serait de se fonder sur le coût des segmentations défini plus haut. Pour chaque nombre de régions k possible (k = 1, 2…), on détermine la segmentation de plus faible coût au moyen de l'algorithme du plus court chemin et on retient le coût de cette meilleure segmentation, notée $J(k)$. On choisit ensuite le nombre de régions k qui, parmi toutes ces solutions, mène au coût le plus faible. On peut montrer que ce raisonnement conduit toujours à choisir le plus grand nombre de régions possible. En effet, le coût $J(k)$, qui traduit l'ajustement de la segmentation aux données, diminue avec le nombre de régions : plus il y a de régions, plus la segmentation résultante sera proche des données (voir figure 2c). La segmentation alors choisie est la segmentation où chaque locus est une région à lui tout seul (l'ajustement est parfait). Mais ce choix n'a aucun intérêt puisqu'il revient à rendre au médecin la figure 1c en guise de figure 1b.

Afin d'obtenir un résultat interprétable, on préfère choisir une segmentation suffisamment bien ajustée aux données sans être trop complexe, c'est-à-dire avec un nombre raisonnable de régions. La figure 2b montre l'évolution du coût $J(k)$ en fonction du nombre de régions k et illustre bien cet objectif. En effet, $J(k)$ diminue fortement pour les premières valeurs de k, signifiant que l'on gagne fortement en ajustement, puis à partir d'un certain k, cette diminution devient moins importante (augmenter k n'améliore que marginalement l'ajustement alors que le modèle devient de plus en plus complexe). La valeur de k correspondant à cette cassure devrait donc être un bon choix. L'idée consiste donc à rendre plus coûteuses les segmentations avec plus de régions. D'un point de vue mathématique, cela se traduit par l'ajout au coût $J(k)$ d'un terme appelé *pénalité*, qui devra refléter la complexité de la segmentation et augmenter avec elle.

Il reste à définir ce que l'on entend par « complexité » d'une segmentation à k segments. Pour décrire une segmentation, il faut déterminer les k valeurs des moyennes mais aussi les limites t_k entre les régions. On a vu que cette dernière recherche nécessitait l'exploration d'un trop grand nombre de segmentations. Ainsi, une pénalité adaptée devra prendre en compte ces deux quantités, typiquement sous la forme

$$a \times k + b \times \log[C(n-1, k-1)].$$

Il serait illusoire de penser que le critère ainsi construit (et qui nécessite donc de choisir des constantes a et b) est infaillible, c'est-à-dire mène à la meilleure solution dans toutes les situations possibles. Disposer du maximum d'informations est donc important. En plus des informations que détient le biologiste (qui connaît ses données mieux que personne), une analyse du comportement du coût $J(k)$ peut permettre d'extraire plusieurs solutions pour k pouvant être biologiquement pertinentes. Par exemple, la figure 2b montre que $J(k)$ présente deux inflexions pour k = 3 et k = 10. Si le critère mène à la solution k = 3, il sera important d'étudier la segmentation pour k = 10, qui comme on va le voir dans la suite permet la détection d'une altération supplémentaire.

La meilleure segmentation permet la détection d'une altération supplémentaire.

Une détection plus fine

Le critère décrit ci-dessus sélectionne $k = 10$ régions pour les données présentées en figure 2. La segmentation associée est donnée figure 2c en bleu. Une région amplifiée est détectée (des locus 11 à 14). Cette région du chromosome 6 contient le gène E2F3 dont l'amplification est connue pour être associée au développement du cancer de la vessie. Une amplification de cette région est également repérée chez d'autres patients atteints du même type de cancer.

Une variation au locus 61 est également détectée par cette segmentation. Ce locus est connu par les biologistes pour être lié à un polymorphisme (une variation génétique locale spécifique au patient). Il n'est donc pas surprenant d'observer une altération à ce locus chez certains patients.

Comme il a été évoqué ci-dessus, un deuxième point d'inflexion existe dans le coût $J(k)$ pour $k = 3$. La segmentation associée est représentée figure 2c en rouge. Cette segmentation, moins fine, permet déjà de détecter la région contenant le gène E2F3 mais pas le polymorphisme au locus 61.

Les auteurs remercient Théo Robin pour sa relecture attentive et ses commentaires.