Biofutur 09/2002; 225 20-22

"Compartimentation chromosomique."

Alexandre G. de Brevern
Équipe de Bioinformatique Génomique et Moléculaire (EBGM),
Unité INSERM U436, Université Denis Diderot-Paris7,
case 7113, 2, place Jussieu, 75251 Paris Cedex 05

Résumé : Du fait du nombre importants de génomes complets actuellement disponibles, il devient possible de les comparer directement. La méthode de chromosome hybride (HχM) permet, en découpant ces génomes en longues zones similaires, cette comparaison.

En deux décennies, les améliorations des méthodes de biologie moléculaire ont permis le passage du séquençage quelques centaines de paires de bases à l'obtention de génomes entiers de plusieurs millions de bases. Toutefois, l'obtention de génomes n'est qu'une première étape, car le traitement de cette information est loin d'être trivial. Ainsi, pouvoir simplement déterminer les zones codantes ou la fonction des protéines associées est encore une tâche complexe (1). Par exemple, lorsque le génome de *Haemophilus influenzae* a été séquencé, près de 40 % des gènes putatifs n'étaient pas associés à une fonction connue (2). De plus, ces nouveaux génomes complets nous permettent de nous poser de nouvelles questions qui auraient inabordable sans eux : Quels sont les liens phylogénétiques entre les génomes, quels ont été leurs évolutions ou comment peuvent-ils évoluer ? La disponibilité récente de longues séquences génomiques permet désormais de travailler sur la structure globale des chromosomes.

Classiquement, l'analyse d'une nouvelle séquence nucléotidique commence par la recherche de séquences qui lui ressemblent dans des banques de données, en alignant cette nouvelle séquence avec des séquences connues. La recherche de motifs est ainsi l'un des axes les plus anciens de recherche en bioinformatique. Les algorithmes les plus classiques actuellement utilisés sont FASTA (3) et surtout BLAST (4). Ces algorithmes se basent sur le présupposé qu'un bon alignement doit contenir de petits segments strictement identiques. Depuis leur conception, ils ont été modifiés pour répondre à différents besoins, comme pour PSI-BLAST où les alignements sont répétés de manière itérative, ce qui est particulièrement utile pour les recherches sur les protéines (5).

Ces approches sont utilisées de manière intensive pour rechercher et identifier des séquences d'intérêts biologiques (séquences codantes, promoteurs, …). Le logiciel ASSIRC (6) est un exemple type des dernières avancées induites par l'explosion du nombre de séquence en bioinformatique. Dans une première étape, il recherche de courtes séquences identiques, puis les étend par une

recherche de type « marche aléatoire » et enfin gère par des arbres de décision les séquences nucléiques d'intérêt. Du fait du nombre croissant de génomes disponibles, il a été nécessaire de prendre en compte la parallèlisation sur plusieurs machines (7).

Ces recherches permettent donc l'annotation des génomes séquencés. Ils permettent aussi de vérifier du fait de ces nouvelles séquences des hypothèses biologiques, plus anciennes. Par exemple, les arbres obtenus à partir de l'ARN ribosomique 16S avaient parfois amené à formuler des hypothèses évolutives qui pouvaient être opposés (8). Des approches très diverses sont alors possibles, comme l'analyse des longues répétitions dans les génomes bactériens qui permet d'établir de nouvelles hypothèses du mécanisme biologique de leurs présences (9) ou d'analyser des évènements de duplications (10). Le Fléche et collaborateurs ont ainsi montré qu'il était possible de distinguer des bactéries pathogènes distinctes, mais génétiquement très homogènes, sur la seule base de leurs répétitions en tandem (11). Il convient bien de noter que ces analyses même utilisant plusieurs approches très différentes ne permettent pas de ne trouver qu'une seule solution (12). Toutefois, la comparaison de séquences génomiques est souvent limitée par le taux de similitude entre les séquences ainsi que la longueur des séquences étudiées. La génomique comparative, elle, recherche à aligner intégralement des génomes, le plus souvent proche, ce qui permettrait ensuite de comprendre leurs évolutions.

Quelques approches novatrices se sont intéressées à retrouver des zones proches entre plusieurs génomes non pas en termes de séquences, mais de composition nucléotidique. Ainsi, Deschavanne et collaborateurs ont montré que les génomes avaient des compositions nucléotidiques particulières qui permettaient de distinguer un génome d'un autre. (13).

Dans le cadre de l'EBGM (Equipe de Bioinfomatique Génomique et Moléculaire), l'objectif de la compartimentation chromosomique que nous avons mise au point est double et tend à utiliser en parallèle ces deux types d'approches précédemment décrites. Une compartimentation *à faible échelle* (motifs) qui consiste à catégoriser l'ensemble des zones similaires observées dans un génome à partir d'une étude fine des duplicats (recherche de zones similaires), et, une compartimentation *à grande échelle* (composition) qui consiste à définir de larges zones susceptibles d'avoir une évolution moléculaire différente, comme par exemple des méta-duplications ou de posséder des propriétés génomiques particulière, comme des catégories d'introns.

Ces deux visions sont en effet complémentaires, la première permet de connaître des régions de fortes similitudes, permettant ainsi de préciser les transferts de matériels génétiques (régions codantes ou non) intra ou inter-chromosomique, la seconde permet d'avoir une vision générale de l'organisation des chromosomes sous forme d'une succession de fragments codés de manière simples, ce qui réduit les problèmes inhérents aux grandes tailles mise en jeu.

Enfin, il faudra intégrer ces deux visions dans un modèle d'évolution du génome étudié.

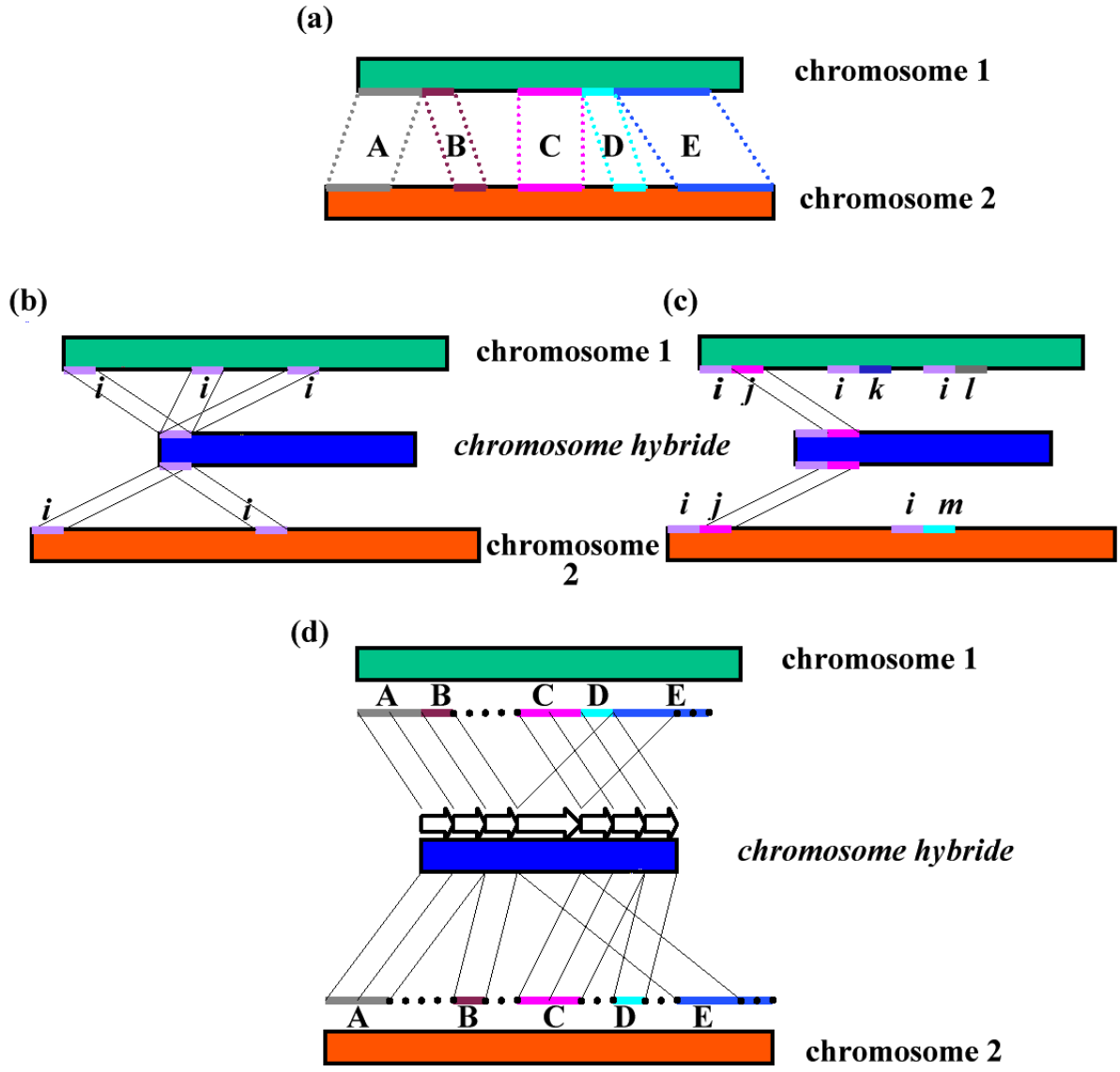

Figure 1: Principe schématique de la méthode du chromosome hybride (voir texte). (a) deux chromosomes virtuels 1 et 2 avec leurs 5 zones communes. (b) définition du chromosome hybride qui apprend les fragments *i*. (c) le chromosome hybride apprend les fragments *i* qui suivent certains fragments *j* et ainsi se spécialise. (d) compartimentation du chromosome hybride en sept zones consécutives (représentés par des flèches) et recherche des 5 zones communes entre les deux chromosomes.

La première approche a été menée avec succès à l'aide du logiciel ASSIRC (6) et est disponible dans la base de données SIMCOGEN (14). Par exemple, le chromosome II de *Saccharomyces cerevisae* possède plus de deux milles régions similaires localisées dans des zones restreintes, comme celles des pieds de transposons.

La seconde approche nous a amené à définir une nouvelle approche. Le principe de base de notre méthodologie HχM (prononcé H-Ki-M pour *Hybrid Chromosome Model*, 15) est de créer en lieu est place d'un ensemble de chromosomes, un seul et unique chromosome de taille réduite. Ce chromosome *hybride* servira de référence aux chromosomes *réels* pour permettre leurs

comparaisons. En effet, du fait de la technique d'apprentissage particulière de ce chromosome *hybride*, nous pourrons définir dans le chromosome *hybride* des zones de forte séquentialité, zones représentant fort bien de longues portions communes des chromosomes *réels*, des compartiments. Ceci va permettre de définir une dizaine de zones recouvrant tout le chromosome *hybride*. Les chromosomes *réels* seront recodés ensuite avec ce nouveau codage. Les comparaisons seront donc alors beaucoup plus rapides et faciles qu'avec une méthode d'alignement classique. Nous avons illustré cette approche sur les 32 subtéloméres de *Saccharomyces cerevisae* bien connus et analysés comme étant des " mosaïques de réplications " (16).

La figure 1 récapitule les différents aspects et intérêts de cette méthodologie en prenant comme base la comparaison de deux chromosomes *virtuels*: (a) le chromosome 1 (vert) et le chromosome 2 (orange) ont évolués à partir d'un ancêtre commun et ont cinq zones communes encore assez similaires (A, B, C, D et E). Différentes zones sont distinctes résultantes d'insertion / délétion (zones intermédiaires A-B, C-D et D-E sur le chromosome 2, fin du chromosome 1) ou de zones ayant été fortement modifiées (zone intermédiaire B-C). Pour retrouver ces zones communes, il faudrait aligner localement l'ensemble de ces chromosomes, ce qui est coûteux en temps de calcul et difficile à gérer du fait des variations de séquence au niveau local. HχM s'intéresse donc à l'aspect compositionnel des chromosomes ce qui élimine le problème de l'alignement local, les chromosomes sont recodés suivant leur fréquence en dinucléotides. Le chromosome hybride créé au départ n'est qu'une suite de fréquence de dinucléotides aléatoires. Il va petit à petit apprendre chaque partie des chromosomes à l'aide d'une méthode d'apprentissage non supervisé, proche des cartes topologiques de Kohonen (17).

(b) Ainsi, par exemple le chromosome hybride (en bleu) apprend des fragments proches notés *i*, ils correspondent à trois zones du chromosome 1 et deux zones du chromosome 2.

(c) HχM permet d'apprendre ces fragments, mais surtout permet de les apprendre dans leur continuité. Ainsi les fragments *i* sont suivis de fragments *j*, qui se retrouveront juste derrière eux aussi dans le chromosome hybride. Les successions *i-k*, *i-l* et *i-m*, elles seront en fait apprise ailleurs sur le chromosome hybride. Cet apprentissage est très dynamique, au début, tout reste très flou, puis petit à petit par un phénomène agrégatif, les zones se dessinent.

(d) Après l'apprentissage, il convient de définir les zones qui représentent au mieux de longues portions des chromosomes, pour cela, nous utilisons un indice entropique qui se base sur la continuité observée lors de l'apprentissage, i.e. des fragments successifs se sont bien appris de manière successive. Cette approche permet par exemple de déterminer 7 zones, qui correspondent plus ou moins aux zones recherchées. Comme tout apprentissage, il faut bien définir différents paramètres qui influent sur le résultat final.

Nous avons donc appliqué cette démarche aux 32 subtéloméres de

Saccharomyces cerevisae qui sont caractérisés par différentes duplications (16). Dans cette étude, 32 séquences de 30 kb (soient 960 kb) ont été utilisés et recodés en fragments de 250 bases consécutives (décallés à chaque fois de 50 bases) et normalisés en fonction de leurs fréquences en dinucléotides La banque de données comprenaient de 19 040 observations. Un chromosome hybride de 595 de long à donc été créé (soit une longueur équivalente aux segments subtélomériques, 19 040 / 32 = 595). Le processus étant itératif, il a fallu utiliser une quinzaine de fois l'ensemble de la banque de données jusqu'à stabilisation du système. Ensuite, le chromosome a été découpé en 15 zones consécutives.

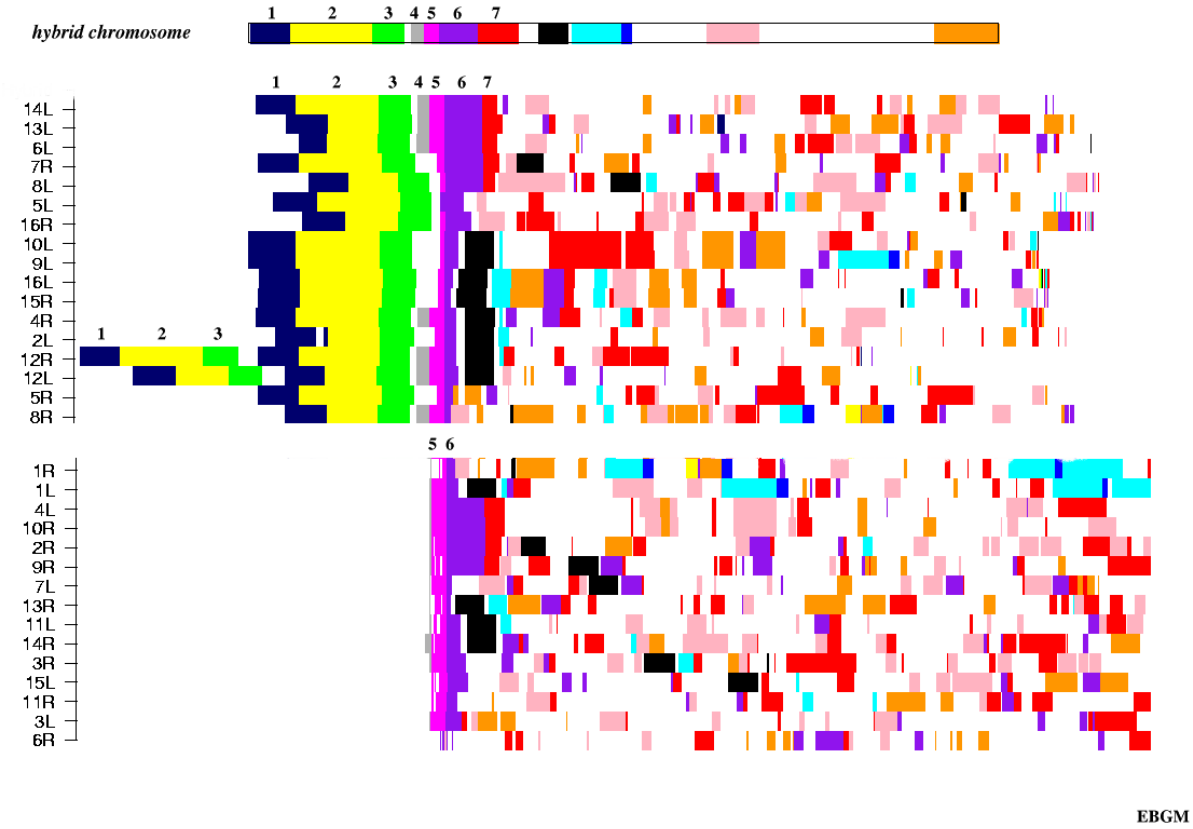

Figure 2: Compartimentation du chromosome hybride et des 32 subtéloméres de *S. cerevisae*, adapté de (15).

Plusieurs analyses sont alors possibles, une première est représenté par la figure 2. Le chromosome hybride final obtenu a été colorié en fonction des 15 compartiments définis. Chacun des 32 subtéloméres a lui aussi été colorié en fonction de ces zones. Deux groupes se dégagent de cette classification, un premier composé de 17 subtéloméres possèdent dans leur partie 5' des zones de type 1 à 3 (bleu, jaune, vert). Ces zones correspondent à des éléments particuliers nommés Y' et cœur X (16). On retrouve même une duplication de ces éléments avec les subtéloméres du chromosome 12 (12L et 12R). Cette représentation permet dans ce cas précis un alignement rapide. On peut alors voir que les subtéloméres 10L et 9L s'alignent fort bien dans leur partie centrale, tout comme leurs voisins 16L et 15R. Le second groupe est présenté en bas de la

figure avec 15 subtélomères qui commencent par des zones 5 et 6 et ne possèdent aucune zones 1, 2, 3 ou 4. Là aussi, des subtéloméres proches sont trouvés aisément, comme le 4L et le 10R.

La figure 3 montre cette fois une comparaison directe entre deux régions subtélomériques. Cette figure est un dotplot, chaque point représente une zone identique dans chacun des subtéloméres (le 6L et le 14L) recodés à l'aide du chromosome hybride. La longue diagonale, en rouge, représente une longue zone commune de 300kb entre les deux subtéloméres. Cette approche permet de voir en outre une petite zone de 22kb, en bleu, qui ne se retrouve pas dans le subtélomére 14L (à 68kb de l'extrémité 5').

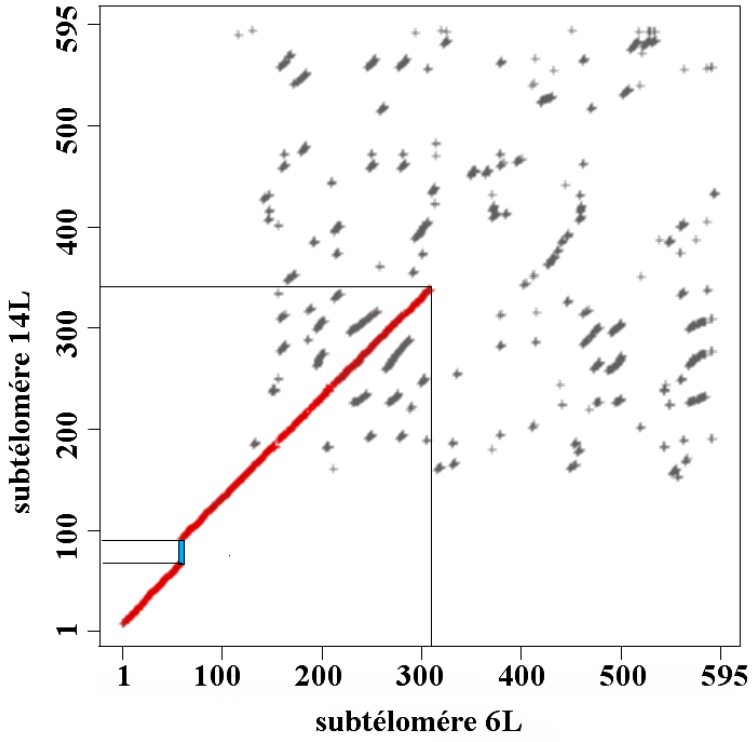

Figure 3: Dotplot entre les subtéloméres 6L et 14L, adapté de (15).

Ainsi, les avantages principaux de HχM sont (i) de permettre une comparaison rapide entre de très longues séquences sans avoir l'obligation d'une optimisation particulière, et, (ii) de conserver la séquentialité entre les observations lors de l'apprentissage, ce qui permet de conserver leur continuité, définissant de longs fragments communs. De plus, l'utilisation de la fréquence en dinucléotide permet de rechercher d'anciennes duplications. En conclusion, ce travail sur les régions subtélomériques de *S. cervisae* a montré qu'HχM permettait de trouver les régions Y' et le cœur X, et, de bien suivre l'évolution proche de ces régions. Les développements futurs de cette approche seront de définir (i) une définition plus optimale de la taille du chromosome hybride et (ii) développer des applications dans la comparaison génomique.

Remerciements:

Ce travail a été partiellement financé par la Génopole de la Montagne Sainte-Geneviéve. AdB a reçu une bourse de la Fondation de la Recherche Médicale.

Références:

(1) J-L Risler, A Louisa. Biofutur 206 (2000), 38-43.
(2) RD Fleischmann, MD Adams, O White, RA Clayton, EF Kirkness, et al. Science 269 (1995), 496-512.
(3) WR Pearson, DJ Lipman. Proc Natl Acad Sci USA 185 (1988), 2444-2448. http ://www.ebi.ac.uk/fasta33/
(4) SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman. J Mol Biol 215 (1990), 403-410. http: //www.ncbi.nlm.nih.gov/BLAST/
(5) SF Altschul, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, DJ Lipman. Nucleic Acids Res 25 (1997), 3389-3402.
(6) P Vincens, L Buffat, C Andre, JP Chevrolat, JF Boisvieux, S Hazout. Bioinformatics 14 (1998), 715-725. http: //condor.urbb.jussieu.fr/
(7) P Vincens, A Badel-Chagnon, C Andre, S Hazout. Bioinformatics 18 (2002), 446-451. ftp: //ftp.ens.fr/pub/molbio/dassirc.tar.gz
(8) C Brochier, H Philippe. Biofutur 206 (2000), 44-48.
(9) EP Rocha, A Danchin, A Viari. Mol Biol Evol 16 (1999), 1219-1230.
(10) F Tekaia, A Lazcano, B Dujon. Genome Res 9 (1999), 550-557.
(11) P Le Fléche, Y Hauck, L Onteniente, A Prieur, F Denoeud, V Ramisse, P Sylvestre, G Benson, F Ramisse, G Vergnaud. BMC Microbiol 1 (2001), 2. http: //minisatellites.u-psud.fr
(12) YI Wolf, IB Rogozin, NV Grishin, RL Tatusov, EV Koonin. BMC Evol Biol 1 (2001) 8.
(13) PJ Deschavanne, A Giron, J Vilain, G Fagot, B Fertil. Mol Biol Evol 16 (1999), 1391-1399.
(14) http: //www.simcogen.ens.fr
(15) AG de Brevern, F Loirat, A Badel-Chagnon, C André, P Vincens, S Hazout. Comp Chem 26 (2002), 437-445.
(16) RJ Britten. Proc. Natl. Acad. Sci. USA 26 (1998), 5906-5912.
(17) T Kohonen. Neural Net 1 (1989), 3-16.