Utilisation de l'indexation de séquences et du calcul thermodynamique pour optimiser la spécificité des oligonucléotides

Nouri Ben Zakour[1], Rozenn Bouville[2], Dominique Lavenier[2], Michel Gautier[1], Yves Le Loir[1*].
[1]Laboratoire d'Hygiène Alimentaire, UMR 1253 STLO, Institut National de la Recherche Agronomique, Agrocampus-Rennes, 65, rue de Saint Brieuc, CS 84215, 35042 Rennes cedex, France ; [2] Equipe Symbiose, Institut de Recherche en Informatique et Systèmes Aléatoires, Campus Universitaire de Beaulieu, 35042 Rennes cedex, France.

* : correspondant : Tel : +33 2 23 48 59 04; Fax: +33 2 23 48 59 02; Email: Yves.LeLoir@rennes.inra.fr

Résumé

La spécificité des oligonucléotides utilisés dans les applications telles que la PCR ou les puces à ADN s'avère l'un des points critiques dans la mise en œuvre de ces approches. Il est impératif de sélectionner des oligonucléotides qui présentent une cible unique sur la séquence génomique considérée afin d'éviter les hybridations parasites sur des sites secondaires qui peuvent gêner l'expérimentation et l'analyse des résultats. L'évaluation de cette spécificité demeure le point sensible des logiciels de sélection d'oligonucléotides. L'utilisation de BLAST ou MegaBLAST dans la plupart des logiciels existants pour identifier les sites secondaires potentiels par analogie de séquence reste une méthode inappropriée dans l'évaluation de l'hybridation ADN/ADN, qui doit se baser sur des équilibres thermodynamiques. Le travail présenté ici propose une approche innovante de l'évaluation de la spécificité des amorces et s'inscrit dans l'amélioration de GenoFrag, un logiciel de dessin d'amorces optimisées pour l'amplification de génomes complets. Nous utilisons l'indexation de séquences pour localiser rapidement tous les éventuels sites secondaires puis nous évaluons l'hybridation potentielle par calcul thermodynamique. Cette approche par indexation nous permet ainsi de tester la spécificité d'un jeu d'environ 70000 candidats de 24 à 26 bases de long sur un génome de 2,8 Mb en moins de deux heures. À la fois sensible et rapide, cette approche peut s'appliquer à tout type de logiciel d'élaboration d'amorces ou de sélection d'oligonucléotides pour puces à ADN.

De nombreuses applications en biologie moléculaire, comme la PCR ou les puces à ADN, nécessitent une évaluation fiable et rapide de l'hybridation d'un oligonucléotide avec des cibles potentielles dans l'ADN génomique. Assurer la spécificité des oligonucléotides revient à sélectionner des oligonucléotides présentant une cible unique d'hybridation thermodynamiquement stable sur la séquence considérée. Optimiser cette spécificité permet d'éliminer tout risque d'hybridations parasites gênant l'expérimentation et ou l'analyse des résultats.

L'évaluation de la spécificité demeure le point sensible des logiciels d'élaboration d'amorces de PCR ou de sondes pour puces à ADN. La plupart des logiciels existants intègrent une approche qui s'appuie sur l'utilisation de BLAST [1] ou MegaBLAST pour identifier rapidement les sites secondaires potentiels par analogie de séquence. Le problème majeur de cette approche est le manque de sensibilité dans la recherche de sites par rapport à la prise en compte des aspects thermodynamiques de l'hybridation ADN/ADN. En effet, l'évaluation d'un appariement doit impérativement tenir compte du contenu en bases nucléotidiques mais aussi de leur enchaînement dans la séquence [2]. Même si quelques logiciels comme OligoArray 2.0 [3] ajoutent, après l'identification de sites secondaires par BLAST, une étape de calcul thermodynamique, certains sites secondaires ne sont pas évalués car ils correspondent à des alignements trop courts, non détectés, alors qu'ils

s'avèrent thermodynamiquement stables et peuvent conduire à une hybridation. L'approche abordée dans FASTH [4] est actuellement la plus aboutie car basée sur une prédiction précise de l'hybridation d'ADN couplée à une recherche de sites secondaires proche de Fasta, mais elle demande un temps de calcul assez important dans le cas d'un grand nombre d'oligonucléotides testés.

Notre équipe étudie la variabilité génomique chez les bactéries par une approche d'amplification des génomes par PCR longue portée ou Whole Genome PCR Scanning (WGPS) [5]. Pour cela, nous avons conçu un logiciel, GenoFrag [6, 7], qui élabore des paires d'amorces permettant d'amplifier un chromosome en fragments d'ADN chevauchants et de taille homogène. Le travail présenté ici porte sur l'amélioration de GenoFrag par une meilleure évaluation de la spécificité des amorces. L'algorithme que nous proposons utilise l'indexation de séquences de manière semblable à celle développée dans iBLAST [8]. Une fenêtre ou clé de longueur k se déplace, base par base, le long de la séquence génomique. Pour chacune des positions de la fenêtre, l'intervalle qu'elle contient est indexé après un codage binaire puis une conversion en base dix, traitement qui a pour but de compresser la taille de la table d'index. Comme une amorce peut s'hybrider sur les deux brins d'ADN du génome, deux tables d'index sont créées correspondant à la séquence génomique courante et son inverse complémentaire. Lors du test de spécificité d'une amorce, chaque k-mer est extrait de la séquence de l'amorce et recodé en base deux puis dix. La recherche de la valeur décimale dans les deux tables d'index de la séquence génomique produit deux tables correspondant aux positions des identités trouvées sur la séquence. L'évaluation de la stabilité thermodynamique des hybridations s'effectue alors pour chacun des sites trouvés sur la longueur totale de l'amorce par rapport à ses cibles.

La particularité de cette approche réside dans la taille de la clé, comprise idéalement entre 5 et 7 pour des oligonucléotides de taille comprise entre 18 et 25 bases, qui permet de cibler les appariements courts qui peuvent s'avérer thermodynamiquement stables et donc problématiques lors de l'expérimentation. La fréquence d'occurrence d'un mot de 5 à 7 étant élevée, l'approche par indexation permet de rester dans des temps d'exécution raisonnables puisqu'il faut compter en moyenne 2h sur un ordinateur équipé de 2 processeurs cadencés à 2,4 GHz avec 1Go de RAM pour tester la spécificité d'un jeu de 70000 candidats de 24 à 26 bases de long sur un génome de 2,8 Mb. La méthode développée ici pour GenoFrag peut s'appliquer à tout type de logiciel d'élaboration d'amorces ou de sélection d'oligonucléotides pour puces à ADN.

[1] Altschul, SF, et al. (1990) Basic local alignment search tool. *J. Mol. Biol.*, *215*, 403-410.
[2] Santalucia, J, Hicks D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct. 33*, 415-440.
[3] Rouillard, JM, et al. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.*, *31(12)***,** 3057–3062.
[4] Zuker, M, (2003) Predicting Nucleic Acid Hybridization and Melting Profiles. *Genome Informatics, 14,* 266-268.
[5] Ohnishi, M, et al. (2002) Genomic diversity of enterohemorrhagic *Escherichia coli* O157 revealed by whole genome PCR scanning. *Proc. Natl. Acad. Sci. U S A., 99(26),* 17043**-**17048.
[6] Ben Zakour, N, et al. (2004) GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification. *Nucleic Acids Res*., *32(1),* 17–24.
[7] Ben Zakour, N, et al. Testing GenoFrag, a software package for whole genome PCR scanning, while handling food-grade bacteria genomes. *Soumis pour publication*.
[8] Cooper, G, et al. (2004) Indexing genomic databases. Proceedings of 2004 IEEE international symposium on Bioinformatics and Bioengineering (BIBE), Taichung (Taiwan), 587-591, May 2004.