Dynamique évolutive d'éléments répétés du génome

Proposition COLOR 2005

1 Objet et durée de la proposition

Il s'agit d'étudier (pendant une année) la résolution analytique des équations d'équilibre issues de la modélisation comme un chaîne de Markov du processus de mutation d'éléments répétés du génome (microsatellites) afin d'améliorer la compréhension de ces modèles. La nature des ces équations les met hors des capacités des systèmes existant de calcul formel, mais certaines rentrent dans une classe d'équations pour laquelle une méthode de résolution est en cours de développement dans le projet Café.

2 Equipes participantes

Sont impliquées une équipe INRIA Sophia, une équipe CNRS du CEFE (Montpellier) et une équipe CNRS du LIRMM (Montpellier).

INRIA : Projet Café[1], M. Bronstein (DR INRIA).

CEFE : Equipe Génétique et Dynamique des Populations[2], P. Jarne (DR CNRS).

LIRMM : Equipe Méthodes et Algorithmes pour la Bioinformatique[3], E. Rivals (CR CNRS).

[1] `http://www.inria.fr/cafe/`

[2] `http://www.cefe.cnrs-mop.fr/wwwgdyn/`

[3] `http://www.lirmm.fr/~w3ifa/MAAS/`

3 Description Scientifique

3.1 Motivation

Les génomes sont largement constitués d'éléments non codants appartenant à la catégorie des ADN répétés. Nous nous intéressons ici à un groupe d'éléments répétés en tandem nommés microsatellites. Les microsatellites sont des éléments de motif unitaire de 1 à 6 nucléotides, dont la taille excède rarement quelques centaines de nucléotides. Ils sont très nombreux, de l'ordre de un million dans le génome humain. Ils sont largement utilisés comme marqueurs en génétique et génétique des populations, et sont impliqués dans diverses maladies neurodégénératives [4]. Notre objectif est de comprendre la dynamique évolutive des microsatellites, en particulier les processus de mutation qui les font évoluer. Deux approches sont possibles :

- l'étude directe de mutations dans des pédigrés (comparaison de parents et de descendants [4]) ;
- une approche génomique, basée sur la comparaison de distributions des microsatellites dans les génomes entièrement séquencés (disponibles dans des banques de données de type Genbank) avec des distributions générées à partir de modèles incluant différentes forces mutationnelles.

C'est la deuxième approche, initiée dans [7], qui est retenue ici, et qui a été utilisée lors de travaux menés dans l'équipe Génétique et dynamique des populations du CEFE [9, 8, 10]. Les modèles utilisés incluent comme forces mutationnelles la mutation par glissement et la mutation ponctuelle. La première module la taille des microsatellites par un nombre entier d'unités de répétitions (par exemple, ajout d'une unité) ; elle peut être dépendante de la taille de l'allèle qui mute (par exemple, augmentation du taux de mutation avec la taille), biaisée (la probabilité d'augmentation de taille diffère de celle de diminution) et impliquer une ou plusieurs unités. L'ensemble du processus a été modélisé comme une chaîne de Markov [7, 11, 9], chaque état de la chaîne représentant un nombre d'unités du microsatellite. Les transitions entre états sont assurées par les processus de mutation. Un problème important avec ce modèle est que la stationnarité de la chaîne de Markov n'a pas été établie de façon satisfaisante, et que le processus ne présente pas de solutions analytiques connues, même dans les cas les plus simples. Ceci est problématique : la phase qui suit la construction de chaînes de Markov est en effet l'ajustement de différents modèles (incluant les forces mentionnées ci-dessus sous différentes formes) aux données génomiques. On en extrait un

"meilleur modèle" auquel on ne peut accorder notre confiance que si la phase de modélisation est satisfaisante. Il y a donc là un point de blocage pour notre compréhension de l'évolution des microsatellites.

3.2 Objectifs

Notre objectif principal est d'appliquer les méthodes de résolutions en développement dans le projet Café aux équations d'équilibre des chaînes de Markov décrites ci-dessus. Vu la présence de nombreux paramètres, l'existence d'une solution analytique générique est peu probable, il nous faudra donc déterminer s'il existe des valeurs de paramètres pour laquelle des solutions analytiques existent. Ce problème de détermination des paramètres est encore ouvert, cependant une étude préliminaire de certaines équations issues de modèles présentés dans [10] a montré que de nouvelles solutions analytiques pouvaient être découvertes. Ainsi, la récurrence

$$\begin{aligned} e\exp(-f(n+1-s))U_{n+1} &+ c\exp(-d(n-1-s))U_{n-1} \\ &- (e\exp(-f(n-s)) + c\exp(-d(n-s)))U_n = 0 \end{aligned}$$

qui provient de modèles de classe 4, admet pour solution

$$U_n = \left(\frac{c}{e}\right)^{n-1} \exp\left(\frac{f-d}{2}n^2 + \left(\frac{f+d}{2} - s(f-d)\right)n\right).$$

Pour une autre équation, l'examen du rapport U_{n+1}/U_n pour une solution analytique nouvelle a permis de découvrir un effet de seuil qui a "réconcilié" les simulations numériques avec les données. Les autres classes de modèles génèrent des équations qui n'ont pas encore été étudiées. Certaines de ces équations (par exemple aux différences partielles ou bien avec des termes binomiaux) pourraient nécessiter une extension de l'algorithme en cours de développement. Enfin, la nature des calculs requis pour chaque example rend nécessaire l'implémentation, au moins partielle, du nouvel algorithme dans un système de calcul formel, ce que nous proposons de réaliser lors d'un stage ingénieur ou de Master.

3.3 Synergies

Le projet Café a une grande expérience de la résolution analytique des récurrences linéaires à coefficients polynômiaux [2] et travaille sur les

récurrences à coefficients hypergéométriques [1]. Ce sont précisément ces récurrences qui apparaissent comme équations d'équilibres des modèles de classe 1 et de classe 4 étudiés [10] par les équipes du CNRS partenaires de cette proposition. L'équipe Génétique et Dynamique des Populations s'intéresse depuis plusieurs années à l'évolution des microsatellites [8, 9, 10] et à leur utilisation en biologie des populations [6, 5]. L'équipe Méthodes et Algorithmes pour la Bioinformatique développe des méthodes de reconnaissance dans les génomes, en particulier pour ce qui concerne les éléments répétés en tandem [3].

Références

[1] Abramov S.A. et al., *On Liouvillian solutions of linear ordinary difference equations with hypergeometric coefficients*, en préparation.

[2] Bomboy R. (2001), Thèse de doctorat, UNSA.

[3] Delgrange O. & Rivals E. (2004), *STAR : an algorithm to Search for Tandem Approximate Repeats*. Bioinformatics 20 :2812–2820.

[4] Ellegren H. (2004), Nature Reviews Genetics 5 :435–445.

[5] Estoup A. et al. (2002), *Homoplasy and mutation model at microsatellite loci and their consequence for population genetics analysis*. Molecular Ecology 11 :1591–1604 [invited review].

[6] Jarne P. & Lagoda P. (1996), *Microsatellites, from molecules to populations and back*. Trends in Ecology and Evolution 11 :424–429.

[7] Kruglyak S. et al. (1998), PNAS USA 95 :10774–10778.

[8] LeRoy (2003), Stage Ecole Polytechnique.

[9] Munoz F. (2002), DEA BEE, Université Montpellier II.

[10] Munoz F. et al., *Models of dinucleotide microsatellite evolution in the yeast, work and human genomes*, en préparation.

[11] Sibly et al. (2001), Molecular Biology and Evolution 18 :413–417.

4 Interactions avec d'autres actions

Ce projet entre dans le cadre de projets plus généraux sur l'évolution des séquences répétées en tandem, menés par P. Jarne et E. Rivals et financés

par des programmes locaux (Bio-STIC-Languedoc-Roussillon) et nationaux (ACI ImpBio).

5 Ressources demandées

On demande un stage de 4 mois (6000 euros au tarif T3) ainsi que 4 missions de 2 jours entre Sophia et Montpellier (600 euros) soit un total de 6600 euros.

6 Actions COLOR antérieurement financées

L'action COLOR *Séquences pseudo-aléatoires et récurrences linéaires à coefficients polynomiaux* entre le projet Café, l'équipe I3S RECIFE et l'université de Toulon à été financée en 2002. Cette action n'a aucun lien avec l'action proposée ici.