Titre : Codes circulaires dans les gènes

Directeur(s) de Thèse : Christian MICHEL, PR27, c.michel@unistra.fr

Unité(s) d'Accueil(s) : ICube, UMR 7357 - Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie

Établissement de rattachement : Université de Strasbourg

Collaboration(s) (s'il y a lieu) :

Rattachement à un programme (s'il y a lieu) :

Résumé **(1500 caractères au maximum)** :

L'objectif de cette thèse est de poursuivre la théorie des codes circulaires dans les gènes selon deux orientations possibles: combinatoire ou statistique.

L'approche combinatoire s'intéressera aux codes k-circulaires. Un code est k-circulaire si une concaténation d'au maximum k mots du code écrit sur un cercle retrouve la phase de lecture. Diverses propriétés combinatoires seront analysées sur les codes k-circulaires de trinucléotides (mots de longueur 3 sur un alphabet à 4 lettres), en particulier: fonctions de croissance, complémentarité, codes (k,k,k)-circulaires et définition d'une distance.

L'approche statistique s'intéressera aux motifs (mots) de code circulaire. Ces motifs ont la propriété fondamentale de retrouver la phase de lecture dans les gènes. La définition et la programmation de fonctions de corrélation basées sur les motifs du code circulaire permettront de rechercher des périodicités et des maximum globaux et locaux dans les gènes. L'objectif est d'identifier de nouveaux codes circulaires. Cette approche permettra également d'étudier les motifs de code circulaire et leur évolution dans les gènes normaux, pathologiques associés à des maladies génétiques et dans les coronavirus, en particulier dans les différents variants du COVID-19.

L'étudiant pourra orienter sa thèse selon une approche combinatoire ou statistique, en fonction de son intérêt et de sa formation universitaire.

Site: https://dpt-info.di.unistra.fr/~c.michel/

Descriptif du sujet (en complément, au format Word ou pdf)

La recherche d'un code circulaire dans les gènes est un problème bioinformatique très important qui a été posé il y a plus de 65 ans. Le concept de codes comma-free a été introduit par Crick et al. en 1957 (Crick et al., 1957) pour expliquer comment la lecture d'une suite de nucléotides de 3 en 3, c'est-à-dire une suite de trinucléotides (codons), pouvait coder les 20 acides aminés constituant les protéines. Cette théorie est restée silencieuse pendant 40 ans. En particulier, le choix d'un ensemble de 20 trinucléotides qui aurait la propriété de retrouver la phase de lecture (construction, décomposition), était infaisable en raison de la combinatoire explosive (environ 3.5 milliards de choix possibles). En 1996, de façon inattendue, une classe de codes plus générale, appelée codes circulaires, est identifiée dans les gènes (Arquès et Michel, 1996). Ainsi selon nos travaux, les gènes seraient composés de codes circulaires de 20 codons permettant simultanément de retrouver la phase de lecture dans les gènes et de coder les acides aminés (cf. les articles de synthèse dans Michel 2008; Fimmel et Strüngmann, 2018). Les codes circulaires auraient ainsi précédé le code génétique universel qui ne possède pas la propriété mathématique de retrouver la phase de lecture dans les gènes.

L'étudiant pourra orienter sa thèse selon une approche combinatoire ou statistique, en fonction de son intérêt et de sa formation universitaire.

Approche combinatoire des codes circulaires

Un code circulaire retrouve la phase de lecture pour tout mot du code écrit sur un cercle (Arquès et Michel, 1996; Fimmel et al., 2016). Un code est k-circulaire si une concaténation d'au maximum k mots du code écrit sur un cercle retrouve la phase de lecture (Fimmel et al., 2020; Michel et al., 2022; Michel et Sereni, 2022). Ainsi, une concaténation de k+1 mots écrit sur un cercle admet plusieurs décompositions. En conclusion, un code k-circulaire ne peut pas être (k+1)-circulaire mais doit être (k-1)-circulaire. Un code est circulaire s'il est k-circulaire pour tout entier k non négatif et on démontre que k est borné.

Ce travail de thèse va s'intéresser aux codes k-circulaires de trinucléotides (mots de longueur 3 sur un alphabet à 4 lettres). Les codes k-circulaires étudiés seront les codes 1-,2-,3-circulaires (les codes 4-circulaires étant circulaires). Les propriétés combinatoires analysées seront en particulier: les fonctions de croissance, la complémentarité, les codes (k,k,k)-circulaires (analogue aux codes circulaires C^3), les codes k-circulaires minimaux (transformation d'un code k-circulaire en un code (k+1)-circulaire). Une nouvelle définition de distance sur les codes k-circulaires pourra également être proposée et étudiée.

L'étudiant devra avoir des connaissances en informatique théorique, combinatoire, théorie des graphes, algorithmique et programmation parallèle.

Approche statistique des codes circulaires

Les motifs (mots) de codes circulaires ont la propriété fondamentale de retrouver la phase de lecture dans les gènes. Ils seront analysés selon: (i) leur longueur; (ii) leur cardinalité (nombre de trinucléotides différents); et (iii) leurs propriétés combinatoires issus de leurs codes circulaires: motifs "strong comma-free", motifs "comma-free", motifs de codes circulaires et motifs de codes k-circulaires. Les gènes des eucaryotes montrent un très fort enrichissement en motifs de code circulaire (Dila et al., 2019; Michel et al., 2017).

L'étudiant poursuivra cette étude dans les gènes des bactéries et des archées, et également dans les ARN avec des fonctions de corrélation basées sur les motifs du code circulaire (Michel et Thompson, 2020). Basé sur les résultats obtenus, il proposera un modèle d'évolution du code génétique. Enfin, la recherche de liens entre ces motifs de code circulaire et les mutations impliquées dans des maladies génétiques ou virales sera étudiée.

Cette approche a été appliquée avec succès puisque l'étude des motifs de codes circulaires dans les génomes de virus nous a permis de prédire tous les gènes potentiellement codants du coronavirus COVID-19 (Michel et al., 2020). L'étudiant poursuivra cette analyse dans les différents variants du COVID-19.

L'étudiant devra avoir des connaissances en informatique et programmation.

Publications directement associées à cette thèse

Arquès D.G., Michel C.J. 1996. A complementary circular code in the protein coding genes. *Journal of Theoretical Biology* 182, 45-58.

Crick F.H.C., Griffith J.S., Orgel L.E. 1957. Codes without commas. *Proceedings of the National Academy of Sciences USA* 43, 416-421.

Dila G., Michel C.J., Poch O., Ripp R., Thompson J.D. 2019. Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes. *Biosystems* 175, 57-74.

Fimmel E., Strüngmann L. 2018. Mathematical fundamentals for the noise immunity of the genetic code. *Biosystems* 164, 186-198.

Fimmel E., Michel C.J., Strüngmann L. 2016. n-Nucleotide circular codes in graph theory. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 20150058.

Fimmel E., Michel C.J., Pirot F., Sereni J.-S., Starman M., Strüngmann L. 2020. The relation between *k*-circularity and circularity of codes. *Bulletin of Mathematical Biology* 82:105, 1-34.

Michel C.J. 2008. A 2006 review of circular codes in genes. *Computer and Mathematics with Applications* 55, 984-988.

Michel C.J., Sereni J.-S. 2022. Trinucleotide *k*-circular codes II: biology. Révision mineure.

Michel C.J., Thompson J.D. 2020. Identification of a circular code periodicity in the bacterial ribosome: origin of codon periodicity in genes? *RNA Biology* 17, 571-583.

Michel C.J., Mouillon B., Sereni J.-S. 2022. Trinucleotide *k*-circular codes I: theory. Révision mineure.

Michel C.J., Mayer C., Poch O., Thompson J.D. 2020. Characterization of accessory genes in coronavirus genomes. *Virology Journal* 17:131, 1-13.

Michel C.J., Ngueface Ngoune V., Poch O., Ripp R., Thompson J.D. 2017. Enrichment of circular code motifs in the genes of the yeast Saccharomyces cerevisiae. *Life* 7, 52, 1-20.