école doctorale **sciences pour l'ingénieur et microtechniques**

(English version follows)

Titre de la thèse : Statistiques combinatoires avec applications à la bio-informatique et à la génomique des virus
Laboratoire d'accueil : LIB (Laboratoire d'Informatique de Bourgogne), Dijon
Spécialité du doctorat préparé : Informatique
Mots-clefs : Motifs dans le structures discrètes, statistiques combinatoires, bio-informatique, génomique
Descriptif détaillé de la thèse : **Introduction / contexte** La thèse se déroulera dans le laboratoire LIB (Laboratoire d'informatique de l'université de Bourgogne Franche-Comté) au sein de l'équipe Combinatoire-Réseaux (CombNet) de Dijon. Le doctorant sera encadré par deux professeurs : Jean-Luc Baril et Vincent Vajnovszki et co-encadré par un jeune maître de conférences, Sergey Kirgizov. **Descriptif scientifique** Les travaux de recherche se scinderont en deux parties complémentaires : une étude théorique dans le domaine de la combinatoire des mots et permutations, et une applicative dans le domaine de la bio-informatique et plus précisément dans la génomique structurale. Pour la partie théorique, nous voulons obtenir de nouveaux résultats concernant la distribution du nombre de motifs et leur popularité sur les permutations et les mots. Les techniques utilisées seront par exemple la description récursive, les fonctions génératrices (bivariées), caractérisation structurale et l'analyse asymptotique. On établira également des correspondances bijectives avec d'autres classes d'objets dont les propriétés sont plus connues, ce qui permettra d'établir des ressemblances et des transports de motifs sur différentes classes d'objets. La fréquence d'apparition des motifs dans les mots a une importance majeure dans des domaines plus applicatifs comme la biologie. En effet, les séquences du génome peuvent être modélisées en les considérant comme des mots de A, C, G, T ou des permutations. Les nullomers sont des courts bouts d'ADN ou ARN n'apparaissant pas dans le génome. Ils ont donc un lien très étroit avec les motifs exclus dans les mots et les permutations qui sont très étudiés en combinatoire. Dans ce contexte, grâce aux résultats précédents, nous proposerons de nouveaux algorithmes efficaces pour la recherche et la détection de motifs absents (nullomers, MAW) dans les séquences ARN de certains virus, par exemple SARS-CoV-2, dont les données sont en libre accès sur la plateforme du NCBI. **Travaux envisagés** La même distribution et le même comportement de certaines statistiques relatives à des motifs dans les permutations ont été récemment conjecturés et obtenus expérimentalement. Les progrès récents nous laissent penser que des techniques telles que des représentations alternatives pour les permutations peuvent être des outils appropriés pour percer quelques-uns

de ces problèmes. De plus, certains résultats sur les permutations semblent pouvoir se généraliser pour des classes restreintes de permutations ou des classes plus générales telles que les mots ou les permutations d'un multi-ensemble. Concernant les applications, les occurrences de motifs ou les statistiques sur les permutations et les mots apparaissent comme des éléments cruciaux pour l'étude structural du génome, et en particulier pour établir une liste exhaustive de nullomers dans des séquences génomiques ce qui permettra aux biologistes de se référer lors de leur tentatives de recherche d'antiviraux. D'autres applications pourront être envisagées pour les codes-barres moléculaires, l'identification des espèces, la surveillance de l'environnement, ainsi que l'étiquetage d'ADN. Des collaborations avec des équipes à l'étranger sont envisagées.

Références bibliographiques :

[1] A. Alileche, J. Goswami, W. Bourland, M. Davis and Greg Hampikian. Nullomer derived anticancer peptides (NulloPs): Differential lethal effects on normal and cancer cells in vitro. Peptides. 38 (2): 302–1. 2012

[2] J.-L. Baril, S. Kirgizov and V. Vajnovszki. Descent distribution on Catalan words avoiding a pattern of length at most three. Discrete Mathematics, Volume 341, Issue 9, September 2018

[3] J.-L. Baril, A. Burstein, and S. Kirgizov. Pattern statistics in faro words and permutations, 2020. https://arxiv.org/abs/2010.06270.

[4] J.-L. Baril and V. Vajnovszki. Popularity of patterns over *d*-equivalence classes of words and permutations. Theoretical Computer Science, 814:249–258, April 2020. https://arxiv.org/abs/1911.05067.

[5] M. Bona, Combinatorics of Permutations (Discrete Mathematics and Its Applications), CRC Pres, Taylor and Francis Group, 2012

[6] Coronavirus genomes – NCBI datasets. https://www.ncbi.nlm.nih.gov/datasets/ coronavirus/ genomes/.

[7] P. Thuan Do, T. Thu Huong Tran, V. Vajnovszki, Exhaustive generation for permutations avoiding a (colored) regular set of patterns, Discrete Applied Mathematics, 268, 44-45, 2019.

[8] P. Flajolet and R. Sedgewick. Analytic Combinatorics. Cambridge University Press, USA, 1st edition, 2009.

[9] L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. Journal of Combinatorial Theory, Series A, 30(2):183 – 208, 1981.

[10] G. Hampikian and T. Andersen. Absent sequences: nullomers and primes. Pac Symp Biocomput. 2007; 355-66. https://pubmed.ncbi.nlm.nihgov/17990505/

[11] G. Hampikian and T. Andersen. Absent sequences: nullomers and primes. Pac. Symp. Biocomput, pages 355 – 366, 2007. http://psb.stanford.edu/psb-online/proceedings/ psb07/hampikian.pdf.

[12] J. Herold, S. Kurtz, and R. Giegerich. Efficient computation of absent words in genomic sequences. BMC Bioinformatics, 9(1):167, 2008. https://bmcbioinformatics. biomedcentral.com/track/ pdf/10.1186/ 1471-2105-9-167.

[13] S. Kitaev, Patterns in Permutations and Words, Springer, 2011.

[14] R. Stanley, Enumerative combinatorics, Cambridge University Press, 2012

Profil demandé : Master mathématiques (avec la maîtrise de au moins un langage de programmation) ou Master Informatique (avec des bonnes connaissances en mathématiques discrètes). Des connaissances en bio-informatique et la maîtrise d'un logiciel de calcul symbolique seront appréciées.

Financement : Région BFC

Dossier à envoyer pour le 1 juin 2021 (CV, lettre de motivation, relevés de notes Master, et lettres de recommandation)

Début du contrat : octobre 2021

Direction / codirection de la thèse : Jean-Luc Baril et Vincent Vajnovszki / Sergey Kirgizov
Contacts : barjl@u-bourgogne.fr, vvajnov@u-bourgogne.fr
sergey.kirgizov@u-bourgogne.fr

école doctorale **sciences pour l'ingénieur et microtechniques**

PhD title: Combinatorial statistics with applications to bio-informatics and virus genomics

Host laboratory: LIB (Laboratoire d'Informatique de Bourgogne), Dijon

Speciality of PhD: Computer science

Keywords: Patterns in discrete structures, Combinatorial statistics, bio-informatics, genomics

Job description:

Introduction and context

The thesis will take place in LIB laboratory, within Combinatorics-Networks teams at Dijon. Two professors will supervise the doctoral student: Jean-Luc Baril et Vincent Vajnovszki and co-supervised by a young associate professor, Sergey Kirgizov.

Scientific description

The research is divided into two complementary parts: a theoretical study in combinatorics on words and permutations, and a more applicative one concerning the applications to the domain of bi-informatics, more precisely, to structural genomics of viruses. We expect to obtain new results on the distribution of pattern frequency and popularity over words and permutations foor the theoretical part. We will consider techniques as recursive description, (bivariate) generation functions, structural characterization, and asymptotic analysis. Also, we will determine one-to-one correspondences with other more classical combinatorial classes, which enable us to find similarities and transport patterns over various combinatorial objects. The frequency of pattern occurrences in words has a significant interest in both biology and combinatorics. Indeed, the genome sequences can be seen as words over the alphabet A, C, G, T or as permutations. Nullomers are short parts of DNA or RNA that do not appear in the genome. So, they have strong ties with pattern avoiding words or permutations, widely studied in combinatorics. In this context, using the previously obtained results, we will propose new efficient algorithms for searching patterns that do not occur in the RNA of some viruses such as SARS-CoV-2, which are available on the NCBI database.

Expected work

The same distribution and behavior of specific pattern-based statistics in permutations and words have been recently conjectured and obtained by computer tests. Recent progresses make us believe that techniques like alternative representations for permutations can be appropriate tools to solve some of these problems. On the other hand, some results on permutations can possibly be translated to restraint classes of permutations or some more general classes as words or multiset permutations.

Finally, pattern occurrences and statistics over permutations or words seem to be critical elements for the study of the genome, in particular to find an exhaustive list of nullomers in SARS-CoV-2 genome, to compare the human nullomers with the virus nullomers in order to help biological scientists to find efficient antivirals.

Other applications will be considered as well: cellular bar-code, species identification, environmental monitoring, DNA labeling. Collaborations with research teams from abroad are envisaged.

References :

[1] A. Alileche, J. Goswami, W. Bourland, M. Davis and Greg Hampikian. Nullomer derived anticancer peptides (NulloPs): Differential lethal effects on normal and cancer cells in vitro. Peptides. 38 (2): 302–1. 2012

[2] J.-L. Baril, S. Kirgizov and V. Vajnovszki. Descent distribution on Catalan words avoiding a pattern of length at most three. Discrete Mathematics, Volume 341, Issue 9, September 2018

[3] J.-L. Baril, A. Burstein, and S. Kirgizov. Pattern statistics in faro words and permutations, 2020. https://arxiv.org/abs/2010.06270.

[4] J.-L. Baril and V. Vajnovszki. Popularity of patterns over d-equivalence classes of words and permutations. Theoretical Computer Science, 814:249–258, April 2020. https://arxiv.org/abs/1911.05067.

[5] M. Bona, Combinatorics of Permutations (Discrete Mathematics and Its Applications), CRC Pres, Taylor and Francis Group, 2012

[6] Coronavirus genomes – NCBI datasets. https://www.ncbi.nlm.nih.gov/datasets/ coronavirus/ genomes/.

[7] P. Thuan Do, T. Thu Huong Tran, V. Vajnovszki, Exhaustive generation for permutations avoiding a (colored) regular set of patterns, Discrete Applied Mathematics, 268, 44-45, 2019.

[8] P. Flajolet and R. Sedgewick. Analytic Combinatorics. Cambridge University Press, USA, 1st edition, 2009.

[9] L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. Journal of Combinatorial Theory, Series A, 30(2):183 – 208, 1981.

[10] G. Hampikian and T. Andersen. Absent sequences: nullomers and primes. Pac Symp Biocomput. 2007;355-66. https://pubmed.ncbi.nlm.nihgov/17990505/

[11] G. Hampikian and T. Andersen. Absent sequences: nullomers and primes. Pac. Symp. Biocomput, pages 355 – 366, 2007. http://psb.stanford.edu/psb-online/ proceedings/psb07/hampikian.pdf.

[12] J. Herold, S. Kurtz, and R. Giegerich. Efficient computation of absent words in genomic sequences. BMC Bioinformatics, 9(1):167, 2008. https://bmcbioinformatics. biomedcentral.com/ track/ pdf/10.1186/1471-2105-9-167.

[13] S. Kitaev, Patterns in Permutations and Words, Springer, 2011.

[14] R. Stanley, Enumerative combinatorics, Cambridge University Press, 2012

Candidate Profile: Master (M.Sc.) in Mathematics (with good programming skills) or Master (M.Sc.) in Computer Science (with a good level in Discrete Mathematics).
Knowledge in bio-informatics and a symbolic computing environment will be appreciated

Financing Institution: Bourgogne Franche-Comté Region

Application deadline: June 1st 2021 (CV, Motivation letter, transcripts of records and reference letters)

Start of contract: October 2021

Supervisors: Jean-Luc Baril and Vincent Vajnovszki / Sergey Kirgizov

Contacts : barjl@u-bourgogne.fr, vvajnov@u-bourgogne.fr / sergey.kirgizov@u-bourgogne.fr