Université de Rennes

Direction de la Recherche et de l'Innovation

Appel à projets - Année 2024

Contrats doctoraux «Ordinaires»

ECOLE DOCTORALE : Matisse **POLE DE RECHERCHE : Mathématique-Numérique**

Unité de recherche d'accueil

Unité de Recherche : IRISA, UMR 6074

Equipe : GenScale

Encadrement

Directeur de thèse : Pierre PETERLONGO **Unité de recherche** : IRISA, UMR 6074

Employeur/Établissement d'origine : Inria

HDR : x oui □ non **Nb de thèses en cours** : 2

Mail : pierre.peterlongo@inria.fr **Tél**. : 02 99 84 74 59

Co-directeur de thèse : Karel BRINDA **Unité de recherche** : IRISA, UMR 6074

Employeur/Établissement d'origine : Inria

HDR : □ oui x non **Nb de thèses en cours** : 0

Mail : karel.brinda@inria.fr **Tél.** : 02 99 84 22 77

Sujet

Titre : Novel computational methods for phylogenetic compression of data structures

Mots clés : bioinformatics, data structures, indexing, compression

Cotutelle internationale (oui/non) : non **Si oui, préciser l'établissement et le pays de rattachement** :

Descriptif (15-30 lignes) :
Contexte socioéconomique et scientifique ; Les hypothèses et questions posées ; Les grandes étapes de la thèse et démarche ; Approches méthodologiques et techniques envisagées

The emergence of modern DNA sequencing methods revolutionized biology, epidemiology, and medicine. However, the rapid growth of sequencing data, surpassing Moore's law, makes it increasingly difficult to search these data using the traditional methods, such as Basic Local Alignment Search Tool (BLAST) [1] and its successors. To keep up with the volumes of data being generated, we need novel compressive approaches that could store and analyze data directly in the compressed domain. However, while it has been shown that genomic data have particular geometrical properties that guarantee the existence of such efficient algorithms [2], [3], the compression of exponentially growing genome collections in a scalable manner remains an unsolved challenge.

A recent technique called phylogenetic compression provided a possible solution [4]. The technique builds upon the insight that the redundancies in genomic data have in fact a highly predictable structure, mirroring the underlying evolutionary processes. As the evolutionary history of genomes can be rapidly estimated, this can be embedded directly into the compression and search algorithms, in order to increase their efficiency and scalability. The preliminary results demonstrated remarkable improvements over the state-of-the-art compression methods, and also suggested its applicability to many diverse data structures across bioinformatics [4].

The goal of this PhD thesis will be developing novel computational methods and mathematical tools for phylogenetic compression of data structures. The student will combine skills in algorithm design and analysis, discrete mathematics and probability, with a taste for programming and experimentation with big genomic data. One focus of the thesis will be the development of a mathematical framework for formalizing phylogenetic compression as an optimization problem, with the use of data structure modeling using evolutionary models. Another direction will focus on the development of phylogenetically compressed variants of modern indexing structures, such as those based on de Bruijn graphs, Bloom filters, and sketches. Finally, an important aspect of the thesis will be efficient implementations of the proposed algorithms and experimental evaluations using some of the largest existing genome databases. The thesis is expected to provide novel

computational solutions to the current challenges and its results may have a direct impact in a range of applications, starting from search engines for sequence data to infectious disease diagnostics in remote locations.

Bibliographie

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," J. Mol. Biol., vol. 215, no. 3, pp. 403–410, Oct. 1990.

[2] P.-R. Loh, M. Baym, and B. Berger, "Compressive genomics," Nat. Biotechnol., vol. 30, no. 7, pp. 627–630, Jul. 2012.

[3] Y. W. Yu, N. M. Daniels, D. C. Danko, and B. Berger, "Entropy-scaling search of massive biological data," Cell Syst, vol. 1, no. 2, pp. 130–140, Aug. 2015.

[4] K. Břinda et al., "Efficient and Robust Search of Microbial Genomes via Phylogenetic Compression," bioRxiv, Apr. 2023, doi: 10.1101/2023.04.15.536996.