Transilvania University of Brașov
INTERDISCIPLINARY DOCTORAL SCHOOL

ADMISSION TO DOCTORAL STUDIES Session September 2026

Field of doctoral studies: Informatics
Doctoral supervisor: Prof dr Sabin Tabirca

TOPICS FOR THE ADMISSION TO DOCTORAL STUDIES

TOPIC 1: **AI-Based Optimization for Bioinformatics Using GPU Computing**

General Description: The rapid growth of genomic data has created significant challenges in the efficient analysis and interpretation of DNA sequences in bioinformatics. DNA sequence algorithms are fundamental in identifying similarities, detecting mutations, and supporting various biological and medical applications. However, traditional algorithms suffer from limitations in computational complexity, processing time, and scalability when applied to large-scale genomic datasets. This research aims to to implement intelligent optimization procedures, including machine learning-based prediction and search strategy guided by heuristics to minimize the alignment search space and enhance computational efficiency. Moreover, parallel computation methods using a GPU will be applied to expedite the alignment process and to effectively process large-scale genomic data. The research is expected to build and deploy a hybrid system with an integration of efficiency of algorithms, AI-based optimization, and parallel execution with high performance. It is hoped that the result will be a scalable and high-performance alignment framework that can accomplish a higher execution rate and still ensure high alignment accuracy.

Research Objectives

1. Optimize a scalable DNA sequence alignment framework using AI-based optimization and parallel computing.
2. Evaluate current algorithms of alignment to find weaknesses in accuracy, efficiency, and scalability.
3. Develop an AI optimization model that enhances the alignment performance of long and noisy genomic data.
4. Combine AI optimisation and parallel computing to speed up as well as improve the accuracy.

Recommended bibliography:

1. Phillip Compeau and Pavel Pevzner, *Bioinformatics Algorithms: An Active Learning Approach* (3rd Edition),
2. Kong, X., Shen, C., & Tang, J. (2025), A Review of Biosequences Alignment, Matching, and Mining Based on GPU, *Current Bioinformatics*, 20(8), 671–683, DOI: 10.2174/0115748936353476241230105816
3. Costanzo, M., Rucci, E., García-Sanchez, C., et al. (2024) Assessing Opportunities of SYCL for Biological Sequence Alignment on GPU-Based Systems, *The Journal of Supercomputing*, 80, 12599–12622.
4. Park, S. et al. (2024). *AGAThA: Fast and Efficient GPU Acceleration of Guided Sequence Alignment for Long Read Mapping,* PPoPP '24: Proceedings of the 29th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming, Pages 431 – 44.
5. Li, H. (2016). Minimap and miniasm: fast mapping for long noisy sequences. Bioinformatics.
6. H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," arXiv preprint arXiv:1303.3997, 2013.

Prerequisites / Remarks: *None*

☑ **Scientific Doctorate**
☐ Professional Doctorate

☑ without tuition fee (state budget funded)
☐ with tuition fee or with funding from other sources than the state budget

TOPIC 2: **Improving GPU Compilers for Safety and Efficiency in the Context of Bioinformatics**

Background and Motivation: Modern bioinformatics is fundamentally data-bound. Genome assembly, variant calling, sequence alignment, phylogenetic inference, and single-cell analysis routinely process terabytes of data, and the underlying algorithms — dynamic programming (Smith-Waterman, Needleman-Wunsch), de Bruijn graph traversal, k-mer counting, hidden Markov models, and increasingly deep-learning models such as DeepVariant and AlphaFold — map poorly onto general-purpose CPUs at scale. GPUs offer enormous throughput, but bioinformatics kernels are notoriously hard to compile efficiently: they exhibit irregular memory access, heavy branch divergence, variable-length inputs, and dependency patterns that defeat naïve auto-vectorisation and standard GPU code generation.

On the other hand, many bioinformatics applications operate over highly structured domains with well-defined invariants, which restrict the set of semantically *meaningful* programs. Formal declarations of domain knowledge can expose these invariants to compile-time reasoning, but requiring them assigns extra responsibility to the programmer. Through pattern-matching for expected domain routines, some concepts can be automatically classified and marked accordingly to leverage their high-level semantic properties in further compiler passes.

This project sits at the intersection of **compiler construction** and **computational biology**: it asks how a compiler can be made *domain-aware* so that high-level bioinformatics code is automatically transformed into GPU kernels that rival hand-tuned CUDA/HIP implementations, without requiring biologists to be GPU experts, while still preserving the generality of the surface level language.

Research Problem: Existing GPU compilers (LLVM/NVPTX, MLIR, Triton, TVM, Polygeist) optimise generic patterns but are largely blind to the algebraic and structural properties of bioinformatics workloads. The central hypothesis is that **encoding domain-specific properties as first-class compiler abstractions and optimisation passes yields substantial, portable performance gains** across GPU architectures (NVIDIA, AMD, Intel) for genomic pipelines. While Domain Specific Languages for bioinformatics have been studied before (Seq, BioLang), these approaches primarily expose abstractions at the language level. In contrast, we propose to investigate how domain knowledge can be captured and exploited at the middle-end level, using structural properties of genomic computations to formalize reusable abstractions and transformation rules, enabling domain-specific optimization to be separated from domain-specific programming.

Research Questions

1. Which recurring computational motifs in bioinformatics (banded DP, wavefront alignment, graph traversal, sparse tensor operations) can be captured as reusable compiler intermediate representations? Can we infer them automatically? Up to what point in the compilation pipeline should the domain specific information survive lowering?
2. How can data-layout, tiling, and divergence-reduction transformations be automated for variable-length biological sequences?

3. Can an MLIR-based domain-specific dialect deliver architecture-portable code generation that matches expert-tuned kernels?
4. How do we co-optimise memory movement (host↔device, HBM, shared memory) given the I/O-bound nature of genomic data?

References:

1. C. Lattner, M. Amini, U. Bondhugula, A. Cohen, A. Davis, J. Pienaar, R. Riddle, T. Shpeisman, N. Vasilache, and O. Zinenko, "**MLIR: Scaling Compiler Infrastructure for Domain Specific Computation, "***Proceedings of the IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, Seoul, South Korea, 2021, pp. 2–14, doi: **10.1109/CGO51591.2021.9370308**.
2. C. Lattner and V. Adve, **"LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation",** *Proceedings of the International Symposium on Code Generation and Optimization (CGO)*, San Jose, CA, USA, 2004, pp. 75–86.
3. N. Ahmed, J. Lévy, S. Ren, H. Mushtaq, K. Bertels, and Z. Al-Ars, **"GASAL2: A GPU Accelerated Sequence Alignment Library for High-Throughput NGS Data",** *BMC Bioinformatics*, vol. 20, no. 1, Art. 520, 2019, doi: **10.1186/s12859-019-3086-9**.
4. H. Li, **"Minimap2: Pairwise Alignment for Nucleotide Sequences",** *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018, doi: **10.1093/bioinformatics/bty191**.
5. Ariya Shajii, Ibrahim Numanagić, Riyadh Baghdadi, Bonnie Berger, and Saman Amarasinghe. 2019. "**Seq: a high-performance language for bioinformatics".** Proc. ACM Program. Lang. 3, OOPSLA, Article 125 (October 2019), 29 pages. https://doi.org/10.1145/3360551

Prerequisites / Remarks: *None*

☑ **Scientific Doctorate**
☐ Professional Doctorate

☑ without tuition fee (state budget funded)
☐ with tuition fee or with funding from other sources than the state budget

Doctoral supervisor,

Prof. Dr.Sabin Tabirca

Signature

Coordinator of the field of doctoral studies,

Prof. Dr. Sabin Tabirca

Signature