

Development of a Novel Metagenomic Classification Method

PhD thesis booklet

Bertalan Vilmos Takács

Supervisor:

Dr. Lajos Haracska

**HUN
REN**



**DOCTORAL SCHOOL OF BIOLOGY
FACULTY OF SCIENCE AND INFORMATICS
UNIVERSITY OF SZEGED**

**INSTITUTE OF GENETICS
HUN-REN BIOLOGICAL RESEARCH CENTRE**

2025

Magyar nyelvű összefoglaló

Az emberi mikrobiom az emberi szervezettel együtt élő mikroorganizmusok összessége, amelyek pozitív és negatív módon egyaránt befolyásolhatják az egészséget. A mikrobiális összetevők azonosításához olyan algoritmusokra van szükség, amelyek nagy pontossággal határozzák meg a minták összetételét, minimalizálják a hamis pozitív találatokat, és egyben gyorsak, valamint skálázhatók is.

Doktori munkám célja egy új metagenomikai azonosító algoritmus, a Novel Alignment-based Biome Analysis Software+ (NABAS+) fejlesztés és összehasonlító tesztelése volt. A programot Java nyelven implementáltuk, kihasználva a BWA-MEM illesztőalgoritmus párhuzamosíthatóságát és a RefSeq adatbázis magas megbízhatóságú referencia-genomjait.

A NABAS+ teljesítményét három típusú adatsoron vizsgáltuk:

(1) *In silico*: Saját fejlesztésű (30 minta) és a CAMI II “toy human gastrooral” adatsor (20 minta). A NABAS+ pontossága összemérhető, több esetben jobb volt, mint a széles körben használt osztályozóké (Kraken2, MetaPhlAn3, GOTTCHA). Egy minta (sample19) újragenerálása tovább növelte a találatok pontosságát.

(2) *In vitro*: Zymo Community Standard I és II (CSI, CSII) Illumina-szekvenált adatsorai. A NABAS+ pontosan detektálta az alacsony abundanciájú fajokat, és nem jelzett hamis pozitív találatot.

(3) Klinikai adatok: 20 humán bélmikrobiom-mintát elemeztünk; az algoritmus az adatsor 11 fertőzött mintájában helyesen azonosította a kórokozót, és nem mutatott hamis pozitív eredményt.

Összefoglalva: kutatásom során egy új, teljes genom illesztésen alapuló metagenomikai azonosító szoftvert fejlesztettem és validáltam. A NABAS+ teljesítménye a legjobb jelenlegi algoritmusokkal összemérhető, alacsony abundanciájú fajokat és patogéneket is pontosan detektál, így nemcsak kutatási, hanem klinikai környezetben is megbízhatóan alkalmazható. Publikus elérhetősége elősegítheti a metagenomikai vizsgálatok pontosságának, reprodukálhatóságának és szélesebb tudományos felhasználásának növelését.

Summary in English

Introduction and Background

The human microbiome encompasses the complex community of microorganisms living in and on the human body, including bacteria, archaea, viruses, and eukaryotic microbes. These organisms can have both beneficial and detrimental effects on human health, influencing metabolism, immune function, and disease susceptibility. The growing use of next-generation sequencing (NGS) technologies has made it possible to study microbiomes with unprecedented depth, yet accurate taxonomic classification remains a major methodological challenge.

This doctoral work focuses on classifying bacteria in shotgun sequenced metagenomic samples coming from the human microbiome. Although a wide range of algorithms available for this task, relying on different algorithmic principles, they are prone to produce discordant results when applied to the same samples.

Reliable and reproducible species-level identification is particularly important when metagenomics is applied in clinical diagnostics, for example in detecting bacterial pathogens in stool samples. Misclassification—especially false positives—may influence treatment decisions. Despite significant progress in computational metagenomics, there is still no consensus on a “gold-standard” classifier, nor on universally accepted threshold settings for post-classification filtering.

For these reasons, we developed NABAS+ (Novel Alignment-Based Biome Analysis Software+), an alignment-based metagenomic classifier designed to minimize false positives while maintaining high accuracy across various sample types. NABAS+ relies on highly curated reference genomes, a modular database architecture, and the BWA-MEM alignment algorithm to provide robust taxonomic classification with minimal ambiguity.

Research Objectives

During my doctoral work, I set the following aims:

- To develop a new metagenomic classification algorithm, optimized for precision, reproducibility, and low false-positive rates.
- To construct *in silico* benchmark datasets suitable for evaluating classifier performance under controlled conditions.
- To compare NABAS+ with widely used metagenomic classifiers on multiple dataset types, including *in silico* data, standardized mock communities, and real clinical samples.
- To assess the suitability of NABAS+ for clinical diagnostics by testing its pathogen-identification accuracy on human stool metagenomic samples.

Methods

NABAS+ was implemented in Java and designed around the principle of using high-quality reference data. Only “representative” genomes from the RefSeq database were included, ensuring that the classifier relied on the most reliable assemblies available. Selecting a single genome per species reduced redundancy in the reference set and allowed NABAS+ to identify species with fewer ambiguous alignments. To make updates straightforward and memory usage manageable, the reference database was divided into modular chunks of approximately 8×10^9 bases, each indexed separately with BWA-MEM. This modular structure not only reduced RAM requirements during alignment but also allowed new genomes to be added without rebuilding the entire database.

Taxonomic assignment in NABAS+ is based on the alignment of paired-end Illumina reads (2×151 bp) to the reference genomes using the BWA-MEM algorithm. After alignment, reads undergo strict filtering based on edit distance, alignment quality, and minimum read counts. In addition, actual genome coverage and hypothetical genome coverage are computed, and genomes with disproportionately high coverage are excluded using a threshold ratio of 3.5. By combining curated reference genomes with strict alignment thresholds, NABAS+ aims to minimize the rate of false-positive identifications while retaining the ability to detect low-abundance species.

To evaluate the performance of NABAS+, I used three types of datasets. First, I constructed six *in silico* mock communities representing different environments, including gut, oral, skin, and environmental microbiomes. These communities comprised 212 species in total and modelled samples of varying read depths, allowing the assessment of classification accuracy under ideal conditions. Additionally, I employed the CAMI II "toy human gastrooral" dataset, a well-established benchmark dataset containing samples with realistic read distributions and error profiles.

Second, I tested NABAS+ on two *in vitro* datasets: the ZymoBIOMICS Microbial Community Standards I and II. These standardized communities contain eight bacterial species and two fungi at known relative abundances, making them ideal for evaluating classifier accuracy using real sequencing data.

Finally, I examined the performance of NABAS+ on 20 clinical human stool samples from Angel et al. (2025), 11 of which were confirmed to contain bacterial pathogens by PCR or culture-based methods. This dataset allowed me to determine whether NABAS+ can reliably detect pathogens in real clinical metagenomes.

Results

Performance on *in silico* mock communities

In the *in silico* experiments, NABAS+ consistently identified the correct number of species across all mock communities. Its performance was not influenced by the origin or complexity of the community, whether it represented an environmental microbiome or human-associated microbiota such as gut or oral samples. We observed that the following, commonly used algorithms produced the result most closely to NABAS: MetaPhlAn3, Kraken2, and GOTCHA. This result highlights the robustness of NABAS+ when applied to *in silico* metagenomic data.

Sensitivity to sequencing depth

To evaluate how sequencing coverage affects classification accuracy, I compared the performance of NABAS+ and other tools across read depths ranging from 5×10^5 to 1×10^7 . We observed substantial differences between classifiers in their sensitivity to increasing read numbers. DIAMOND and Kaiju were the most affected, showing pronounced increases in detected species with higher coverage, while Centrifuge, Bracken, and CLARK exhibited more moderate changes. By contrast, NABAS+, MetaPhlAn3, and GOTCHA produced highly stable results, indicating that their performance was largely independent of sequencing depth. NABAS+ therefore maintained both accuracy and consistency across

datasets modeling both deeply and shallowly sequenced samples.

F1-score analysis

To further quantify classifier performance, I calculated F1 scores across all in silico datasets. NABAS+ showed consistently high F1 values, often matching or surpassing the most widely used tools. Kraken2, although strong in recall, frequently produced numerous false positives that reduced its precision. In contrast, NABAS+ achieved a balance between sensitivity and specificity that resulted in stable, high F1 performance across all tested communities.

Evaluation with the CAMI II gastrooral dataset

Using the CAMI II benchmark, I observed that MetaPhlAn3 performed best overall, followed closely by NABAS+ and GOTTCHA. Kraken2 again exhibited strong recall but poor precision. One sample, sample19, produced unusually poor results across all classifiers; subsequent investigation revealed that this sample had been generated from outdated genome assemblies. After regenerating the sample using the most recent RefSeq representative genomes, all classifiers showed improved accuracy. This finding illustrates that classifier performance is determined not only by the underlying algorithm but also by the quality and currency of the reference database.

Results on Zymo mock communities

On the Zymo community standards, NABAS+ achieved excellent results. In both the equal-abundance and logarithmic-abundance datasets, NABAS+ detected all organisms present in the sample and did not report any false positives. It also produced the lowest Bray–Curtis distance to the true community composition, indicating that its relative abundance estimates were more accurate than those of other classifiers. Notably, NABAS+ detected *Enterococcus faecalis* at an abundance as low as 0.00089%, demonstrating exceptional sensitivity for low-abundance species.

Clinical validation

In the clinical dataset of 20 human stool samples, 11 contained bacterial pathogens confirmed by PCR or culture. NABAS+ correctly identified all pathogens in the positive samples and did not produce any false-positive results in the negative samples. This perfect concordance with laboratory diagnostics demonstrates that NABAS+ is suitable for clinical application and can reliably detect pathogens in complex human microbiomes.

Summary and Scientific Contributions

In this dissertation, I developed and validated NABAS+, a new alignment-based metagenomic classifier designed to minimize false positives and deliver reproducible, high-precision species identifications. I demonstrated that NABAS+ performs comparably to or better than widely used classifiers such as Kraken2, MetaPhlAn3, and GOTTCHA across *in silico*, *in vitro*, and clinical datasets. NABAS+ showed strong resilience to sequencing depth, excellent accuracy on standardized reference communities, and perfect pathogen identification in real clinical samples.

The scientific contributions of this work include the creation of a modular, easily updateable database structure; the empirical determination of optimal alignment filtering thresholds; the demonstration that database quality has a major impact on classification performance; and the application of NABAS+ to clinical metagenomic diagnostics. Together, these findings show that NABAS+ is a reliable and versatile tool that can enhance both research and clinical microbiome analysis.

List of Publications

MTMT ID: 10074085

Publications used in this thesis

1. *Advancing metagenomic classification with NABAS+: a novel alignment-based approach*
Takács et al.
NAR Genomics and Bioinformatics, Volume 7, Issue 3, September 2025
Impact Factor: **5.0**
2. *Shotgun Analysis of Gut Microbiota with Body Composition and Lipid Characteristics in Crohn's Disease*
Bacsur et al.
Biomedicines 2024, 12(9), 2100;
Impact factor: **3.9**

Other publications

1. *Prolonged activity of the transposase helper may raise safety concerns during DNA transposon-based gene therapy*

Imre et al.

Molecular Therapy Methods & Clinical Development, Volume 29, 145 – 159

Impact Factor: 4.7

2. *A Comprehensive Evaluation of the Performance of Prediction Algorithms on Clinically Relevant Missense Variants*

Qorri et al.

International Journal of Molecular Sciences. 2022; 23(14):7946

Impact Factor: 4.9

3. *Distinct Gut Microbiota Profiles in Unruptured and Ruptured Intracranial Aneurysms: Focus on Butyrate-Producing Bacteria*

Csécsei et al.

Journal of Clinical Medicine. 2025; 14(10):3488

Impact Factor: 3.0

Conference talks

1. *Comparative Analysis of the Microbial Composition of Different Tumor Tissues*
Talk at the Microbiota and Cancer Immunity Conference, Taipei, Taiwan, 2024
2. *The devil is in the details: Eliminating false metagenomic classification with a novel algorithm*
Talk at the 61st Annual Meeting of the Hungarian Society of Laboratory Diagnostics, Budapest, Hungary, 2023
3. *Understanding the effects of COVID-19 on the microbiome using bioinformatics and machine learning*
Talk at the 5th National Conference of Young Biotechnologists, Gödöllő, Hungary, 2022

Posters

1. *The devil in the details: Eliminating false metagenomic classification with a novel algorithm*
Poster at the Hungarian Molecular Life Sciences 2023 Conference, Eger, Hungary, 2023
2. *Understanding the effects of COVID-19 on the microbiome using bioinformatics and machine learning*

Poster at the 25th Spring Wind
Conference, Pécs, Hungary, 2022

3. *Impacts of COVID-19 on the Microbiome: A
Bioinformatics and Machine Learning Study*

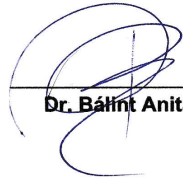
Poster at the 21st European Conference on
Computational Biology, Sitges, Spain, 2022

NYILATKOZAT

Nyilatkozom, hogy Takács Bertalan Vilmos társszerzőként vett részt a felelős szerzőségemmel megjelent "Shotgun Analysis of Gut Microbiota with Body Composition and Lipid Characteristics in Crohn's Disease" (BIOMEDICINES, 12 (9). ISSN 2227-9059 (2024)) című közleményben. A közlemény más PhD fokozatszerzési eljárásban társszerzőségként nem volt felhasználva.

Bertalan a közleményben az alábbi hozzájárulásokat végezte:

- Bioinformatikai analízis (shot-gun szekvenált mikrobiom minták összetételének meghatározása)
- Statisztikai analízis (Bray-Curtis távolság számítás és főkoordináta-analízis)
- Adatvizualizáció



Dr. Bálint Anita Ph.D.