Facing Scalable Sequence Data: Theory, Algorithms, and Applications

March 18, 2026

[Note: This research statement is intended for practice purposes only and is not tailored for faculty applications. Paragraphs highlighted in blue describe ongoing works that are preliminary and may not yet represent fully validated results.]

1 Research Vision

The era of large-scale biological sequence data has arrived, with collections now reaching the petabyte scale and growing rapidly [1], fundamentally reshaping how genomic analysis is performed. Storage, searching, querying, comparison, and many types of large-scale analysis all face unprecedented computational challenges. To address these challenges, alignment-free approaches have emerged as a scalable alternative. These methods represent sequences using summary statistics derived from their k-mer composition and can therefore compare sequences without explicitly computing alignments [7, 8]. Many modern alignment-free tools rely on sketches of k-mer spectra, enabling efficient comparison of massive sequence collections, like Mash [2], Skmer [5], Sylph [6], and Mercury [4]. These tools enable tasks such as large-scale phylogenetic reconstruction, sample identification, metagenome profiling, and genome assembly evaluation, often reducing computation times from days to minutes. However, several major gaps remain in the field.

First, **repetitive sequences remain a major challenge for k-mer–based methods**. Many existing estimators either explicitly or implicitly assume that there are no repeats in the strings. This leads to limited theory and poor practical performance in highly repetitive sequences, such as centromeres.

Second, **many realistic biological scenarios are still insufficiently modeled by current alignment-free theory**. Examples include comparisons between whole genomes and reads, divergence from a common ancestor rather than one sequence mutating directly into another, and mutation processes involving insertions and deletions. These settings arise naturally in practice, yet they often lack rigorous statistical foundations and scalable estimators.

Third, **the field still needs more specialized alignment-free algorithms for biological problems**, including ANI estimation and the analysis of repetitive genomic regions. There is a

critical need to apply these theories and algorithms to solve concrete biological questions or to develop user-friendly tools that assist biologists in effectively dealing with large-scale data.

2 Current Research Contributions

2.1 Repeat-aware theories for mutation-rate estimation and Applications to centromere genomics

I develop repeat-aware estimators for mutation rates under substitution-only whole-sequence models. Existing k-mer estimators perform well in low-repeat settings, but their theoretical assumptions often fail on repetitive sequences. And they do not provide a series of estimators for different application scenarios. To address these limitations, we propose a family of repeat-robust estimators categorized by the presence and count information that can be obtained from the source and mutated strings.

Rather than relying solely on traditional shared-k-mer intersection perspectives, I introduced a new viewpoint based on *novel k-mers*, namely those that appear in the mutated sequence but not in the source sequence. This perspective leads naturally to a family of estimators adapted to different information settings, depending on whether one has presence-only or count information from one or both sequences. These estimators provide a spectrum of methods that trade off data requirements, computational cost, and statistical accuracy.

This work contributes both theory and methodology. On the theoretical side, it provides approximations and interpretable statistical expressions for k-mer behavior in repetitive strings. On the methodological side, it produces practical estimators that are substantially more robust than previous estimators, especially in highly repetitive settings.

This method was carried out in collaboration with researchers in centromere genomics. Centromeres are highly repetitive, structurally complex, and often difficult to analyze with alignment-based approaches, especially when comparing divergent active arrays. Centromeres are biologically essential genomic regions that play a fundamental role in chromosome segregation and genome stability. At the same time, their highly repetitive and complex structure makes them among the most challenging regions for sequence analysis.

Using the repeat-aware framework above, I developed k-mer–based approaches for analyzing evolutionary patterns in centromeric sequences. In particular, I used these estimators to study the T_i/T_v ratio and to build a k-mer–based metric for measuring distances between centromeres. Prior to this work, centromere active arrays were typically studied indirectly through variation in their flanking regions, which was used to infer properties such as evolutionary relationships and haplotype. Our repeat-aware k-mer–based estimator instead enables direct analysis of the active arrays themselves. We demonstrate strong agreement with established results based on flanking-region variation, providing validation of our approach and enabling direct, scalable analysis of active arrays that were previously inaccessible.

2.2 Extending theories and estimators under different mutation models

Another major direction of my work has been to extend alignment-free theory beyond the classical setting in which one full sequence is generated from another by independent substitutions. While this model is mathematically convenient, many important applications fall outside it. I developed a series of theoretical frameworks that generalize k-mer–based analysis to more realistic mutation and observation settings.

Subsequence and read-mapping settings. The first work considers settings in which only a homologous subsequence, rather than two full sequences. This arises naturally in read-mapping scenarios. I developed a theoretical framework for such subsequence-based comparisons, enabling mutation-rate estimation even when we observe fragments of the mutated string and the source string is really repetitive.

Branching mutation models. A second line of work focuses on evolutionary scenarios in which two descendant sequences diverge from a common ancestor, rather than one sequence mutating directly into another. By analyzing k-mer statistics under such branching mutation models, I developed a framework for inferring generation mutation rates from descendant sequences. This provides a more realistic abstraction for evolutionary inference and connects alignment-free methods with classical phylogenetic questions.

2.3 Alignment-free theory under indel mutation processes

A further direction of my research has been to develop alignment-free estimators under indel-channel mutation models, incorporating substitutions, deletions, and insertions. Indels create a particularly important challenge because they disrupt the positional correspondence that underlies many simple analyses of k-mer mutation. At the same time, realistic biological evolution clearly includes indels, so restricting theory to substitution-only settings leaves an important gap. In a previous study, Rahman Hera et al. derived closed-form k-mer-based estimators for the substitution, deletion, and insertion rates. However, their implementation requires constructing unitigs and performing semi-global alignment, which limits scalability and partially defeats the purpose of alignment-free methods.

Estimating substitution, deletion, and insertion rates simultaneously from k-mer statistics is challenging when using a single value of k. Our key innovation is to separate the estimation problem into two components. First, we estimate the sequence identity using a single value of k. Second, we combine information from multiple k values to estimate the expected number of mutation operations. This multi-k strategy allows mutation parameters to be inferred while preserving the scalability advantages of k-mer-based methods. This produces scalable estimators for mutation parameters that remain compatible with sketching-based computation and avoid the need for expensive alignment-heavy workflows.

This direction is important not only because indels are biologically realistic, but also because it illustrates how alignment-free methods can move beyond estimating a single similarity number toward richer statistical inference under explicit mutation models.

2.4 Scalable ANI estimation

Guided by the mutation models above, I developed fast and accurate approaches for Average Nucleotide Identity (ANI) estimation. ANI is one of the most important comparative metrics in microbial genomics. The classical alignment-based tools are super slow, yet existing modern methods are purely k-mer-based or seed-and-chain-based, but do not provide a better trade-off between time and accuracy, and are not robust to repetitive strings.

My work contributes to this area by designing k-mer–based estimators and algorithms that improve the balance between computational efficiency and accuracy. The success of my approach is rooted in the theoretical characterization of k-mer statistics under the above explicit mutation models and the corresponding method-of-moments estimators derived from theories. These estimators capture quantities analogous to those measured in classical alignment-based tool ANIb, effectively approximating alignment-derived similarity through k-mer statistics. This enables scalable computation while preserving the interpretability and biological relevance of alignment-based measures.

2.5 Interpretation of k-mer sharing patterns

Large-scale sequence search systems often rely on counts of shared k-mers to identify candidate matches. However, the meaning of a particular match pattern is often unclear: is a query truly related to a reference, or does it merely share enough k-mers by chance, repeats, or noise? This question becomes increasingly important as sequence databases grow and search systems return large candidate sets.

To address this, I developed a statistical framework for interpreting k-mer sharing patterns and for assigning confidence to sequence search results. Rather than treating shared k-mer counts as raw scores alone, this framework analyzes the structure of the match pattern itself. It models observed patterns as noisy versions of latent mutation patterns. This line of work improves the interpretability and reliability of large-scale query systems such as Logan, helping make alignment-free search not only fast but also statistically principled.

3 Future Research Program

Looking forward, I plan to expand this research program along three tightly connected directions: theory, algorithms, and applications.

1. **Theory for biological sequence models.**

The first direction is to continue building rigorous theory for k-mer statistics or other scalable methods. One goal is to further unify substitution, branching, subsequence, and indel settings into a broader statistical framework for sequence comparison. Another is to develop theories for settings involving read sets, assembly, heterogeneous mutation processes, and structured repeats. These problems are important because modern sequence data are rarely generated under the idealized assumptions of existing theory.

2. **Specialized scalable algorithms for emerging biological problems.**
   The second direction is algorithm design. I plan to develop new k-mer–based and sketching algorithms tailored to additional sequence-analysis tasks where scale, repetition, or incomplete homology create barriers for current methods. These may include metagenomic profiling, binning, large-scale sequence search, repeat-aware comparison, and sequence relationship inference across complex mutation scenarios.

3. In the long run, I hope to help establish a new generation of scalable methods that are both mathematically interpretable and practically useful across a broader range of problems.

4. **Applications to large biological projects.**
   The third direction is to apply these methods to important biological questions and collaborative data resources. I am especially interested in large-scale projects involving pangenomes, metagenomes, and database-scale search. My role is to contribute foundational methodology that allows large and complex datasets to be analyzed in ways that remain biologically meaningful. I am particularly excited by the opportunity to work at the interface between theory and real biological data, where new applications often motivate new models, and new models in turn enable new scientific discoveries.

References

[1] K. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. Brister, and C. O'Sullivan. The sequence read archive: a decade more of explosive growth. *Nucleic Acids Research*, 50(D1):D387–D390, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1053. URL `https://doi.org/10.1093/nar/gkab1053`.

[2] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, 2016.

[3] M. Rahman Hera, P. Medvedev, D. Koslicki, and A. Blanca. Estimation of Substitution and Indel Rates via k-mer Statistics. In B. Brejová and R. Patro, editors, *25th International Conference on Algorithms for Bioinformatics (WABI 2025)*, volume 344 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 16:1–16:15, Dagstuhl, Germany, 2025. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-386-7. doi: 10.4230/LIPIcs.WABI.2025.16. URL `https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.WABI.2025.16`.

[4] A. Rhie, B. P. Walenz, S. Koren, and A. M. Phillippy. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, 21(1), Sept. 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02134-9.

[5] S. Sarmashghi, K. Bohmann, M. T. P. Gilbert, V. Bafna, and S. Mirarab. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology*, 20(1):1–20, 2019.

[6] J. Shaw and Y. W. Yu. Rapid species-level metagenome profiling and containment estimation with sylph. *Nature Biotechnology*, pages 1–12, 2024.

[7] K. Song, J. Ren, G. Reinert, M. Deng, M. S. Waterman, and F. Sun. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in bioinformatics*, 15(3):343–353, 2014. doi: 10.1093/bib/bbt067.

[8] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome biology*, 18(1):186, 2017.