Abstract I propose to use a spaced-seed search technique to develop an ab initio repeat-finding tool with improved sensitivity and run time relative to currently available tools.

Background Repetitive DNA sequences are sequences that appear multiple times, with possible variations, in a single genome [3]. They make up a significant portion of eukaryotic genomes, accounting for over 50% of the human genome. Repetitive DNA can create ambiguities in genome analysis, leading to incorrect interpretations of the results [7]. Additionally, its analysis provides important insights when studying patterns of genomic evolution. As a result, the detection of repetitive DNA has become an important challenge in bioinformatics.

A variety of repeat-finding tools have been developed in the years following the discovery of repetitive DNA [6]. Library-based tools identify known repeats through comparison of input genomic data to an existing library of repeat sequences, while ab initio tools identify repeats without such knowledge. Because only ab initio tools can deal with newly sequenced genomes or discovering new repeat families, they are becoming increasingly important due to the rise in number and diversity of sequences coming from genome sequencing projects.

Over long periods of time, originally identical repetitive sequences have accumulated mutations, creating approximate repeats - repeats that are still similar but no longer identical. This variation makes approximate repeats not only considerably harder to find, but also more valuable in that they contain information about the patterns of genomic evolution.

Beyond repeat identification, there lies a problem regarding how to best represent repeats. It has been suggested that repeats should be represented as "mosaics of subrepeats", the claim being that this representation conserves evolutionary relationships between subrepeats, which are often lost in other repeat representations [5]. This repeat representation problem should be kept in mind when designing new ab initio tools.

Motivation I propose that ab initio search methods could be improved in terms of their sensitivity to approximate repeats through the use of the spaced seed matching technique that has proved successful in improving results from the BLAST tool.

Spaced seeds are patterns describing which positions in two strings must match, and which positions are not so constrained, in order for them to be considered a "valid" match. A spaced seed π is inherently defined by an ordered list of matching positions $M_\pi = \{i_1 \dots i_w\}$. Two sequences q and t match with respect to π if $q_i = t_i \ \forall i \in M_\pi$. When trying to find similar sequences in the context of homology search (e.g. the BLAST tool), it has been found that allowing tolerance to these mismatches over short spans considerably improves performance and result quality as compared to searching for short exact matches [4].

With Dr. John Karro, I have been exploring the use of spaced seeds in order to improve the sensitivity of a repeat-finding tool created in our lab, RAIDER [1]. This tool only works on consecutive matches, and needs to be generalized to use spaced seeds. I have been revising the RAIDER's underlying algorithm to allow for the use of single/multiple spaced seed sets, as well as investigating which characteristics make a seed better suited for repeat-finding purposes. This approach has had significant effects on both sensitivity and speed in homology search tools [6, 4]. Although the associated challenges differ, it seems reasonable that an optimized spaced seed design could also improve sensitivity in repeat-finding tools.

Proposed Research The ability to quickly detect repetitive DNA is necessary in genome analysis, and becomes more so as the number of sequenced genomes continues to increase. Ab initio repeat-finding tools are often lacking in the following areas:

1. *Approximate repeat detection.* Many tools are unable to find approximate repeats without significantly more computational resources, but finding only exact repeats results in obtaining a very small fraction of all repeat elements.
2. *Time efficiency.* Some approaches can accurately detect repeats, but require quadratic time and space. This leads to an unacceptably long runtime, given the large size of genomic sequences.
3. *Meaningful representation of output.* The evolutionary relationships between sub-repeats are often lost in representations of repeats by repeat-finding tools [5].

I propose that ab initio approximate repeat-finding could be improved, while maintaining a reasonable runtime, through the use of a spaced-seed search technique. I would like to investigate this approach further in order to develop an ab initio repeat-finding tool that meets all of the aforementioned criteria. My prior research involving repeat detection, in addition to my background in mathematics, will aid me in this research.

Intellectual Merit I am proposing a spaced-seed approach to ab initio approximate repeat-finding that may significantly improve the detection of approximate repeats. Additionally, many repeat identification tools are based on sequence alignment, which is inherently slow. My approach is an alignment-free method, which would be significantly faster and could be applied to numerous other problems (e.g. searching meta-genomic datasets for common patterns that may indicate interspecies conservation, or, once available, large personal-genome datasets for recurring patterns that might indicate recurring abnormalities).

Broader Impacts This research could improve genome analysis and contribute to the study of molecular evolution, providing new tools for genome analysis that would be of direct use to researchers in the field. With the ability to quickly identify repeats in newly sequenced genomes, researchers looking to annotate genes will be able to quickly filter repetitive areas, while those studying molecular evolution will have more data with which to study features such as genomic substitution rates [7, 2]. Additionally, the ab initio aspect of the tool allows for discovery of new repeat families, which will aid in the furthering our understanding of the evolutionary process and genomic structure at the molecular level. In short, this research will provide tools of value to those conducting genomic analysis, leading to an improved understanding of the human genome and corresponding benefits to human health.

References

[1] N. Figueroa, X. Liu, J. Wang, and J. Karro. Raider: Rapid ab initio detection of elementary repeats. In *Advances in Bioinformatics and Computational Biology*, pages 170–180. Springer, 2013.

[2] J. E. Karro, M. Peifer, R. C. Hardison, M. Kollmann, and H. H. von Grünberg. Exponential decay of gc content detected by strand-symmetric substitution rates influences the evolution of isochore structure. *Molecular Biology and Evolution*, 25(2):362–374, 2008.

[3] E. Lander, L. Linton, B. Birren, and et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 02 2001.

[4] B. Ma, J. Tromp, and M. Li. Patternhunter: faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 2002.

[5] P. A. Pevzner, H. Tang, and G. Tesler. De novo repeat classification and fragment assembly. *Genome research*, 14(9):1786–1796, 2004.

[6] S. Saha, S. Bridges, Z. V. Magbanua, and D. G. Peterson. Computational approaches and tools used in identification of dispersed repetitive dna sequences. *Tropical Plant Biology*, 1(1):85–96, 2008.

[7] T. J. Treangen and S. L. Salzberg. Repetitive dna and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, 01 2012.