Approximate Repetitive DNA Identification

Charlotte E. Schaeffer

Abstract

Identifying repetitive sequences within a genome is one of the fundamental problems of bioinformatics.

CONTENTS

I. Introduction

II. Background

A. Biological Background

Every living organism inherits hereditary information from its parents that affect the organism's distinguishing traits. This information is embedded inside an organism's genetic material, a molecule known as deoxyribonucleic acid (DNA). An organism's *genome* is the set of all DNA sequences associated with that organism [1].

Each sequence of DNA is composed of a chain of nucleotides. There are four nucleotides found in DNA: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). DNA can therefore be represented as a finite string $s = s_0 s_1 ... sn - 1$ over the alphabet $\Sigma = \{A, C, G, T\}$ of nucleotides [2].

B. Repetitive DNA

A *repeat* is a DNA sequence that is similar or identical to one or more other sequences in the same genome [3].

Definition 1. Let F be a subsequence of the query sequence with Lmer decomposition $x_1, x_2, \ldots, x_k$, where $k = |F| - L + 1$. F is an **elementary repeat** if:

1) $k \geq 1$
2) $freq(F) \geq f$
3) $freq(x_i) = freq(F)$ for all Lmers x_i in the decomposition
4) k is maximal. That is, there is no Lmer y such that $y \circ F$ or $F \circ y$ meets conditions 1-3

[4]

C. Spaced Seeds

Definition 2. A **spaced seed** is a string π over the alphabet $\Sigma = \{1, *\}$, where a position with value 1 is a match and a position with value * is a "wildcard position" that can be either a match or a mismatch [5]

A spaced seed π is defined by an ordered list of matching positions $M_\pi = \{i_1 \ldots i_w\}$. The number of matching positions is the seed's *weight*, denoted w_π. The *length* or *span* of the seed is denoted $|\pi|$ [6]

Definition 3. Let π be a spaced seed of length L with matching positions $M_\pi = \{i_1 \dots i_w\}$. Let q and t be genomic sequences of length L. We say that t **matches** q with respect to π if $q_i = t_i \forall i \in M_\pi$.

Definition 4. Let π be a spaced seed of length L with matching positions $M_\pi = \{i_1 \dots i_w\}$. Let Q and T be genomic sequences of length n with Lmer decompositions $x_1, x_2, \dots, x_k$ and $y_1, y_2, \dots, y_k$, respectively (where $k = n - L + 1$). We say that T **matches** Q with respect to π if $\forall i \in \{0, n) \exists j \in \{0, n - L)$ such that $j \leq i < j + L$ and x_j matches y_j with respect to π.

We say that two genomic sequences Q, T match one another if every position $i \in \{0, n)$ corresponds to Lmers $x_j \in Q$ and $y_j \in T$ spanning positions $\{j, j + L)$ that match one another with respect to some spaced seed π.

III. Proposed Research

IV. Timeline

References

[1] B. Lewin, J. Krebs, E. Goldstein, and S. Kilpatrick, *Lewin's Genes XI*, vol. 11. Jones & Bartlett Learning, 2014.

[2] M. Elloumi and A. Zomaya, *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*. Wiley Series in Bioinformatics, Wiley, 2011.

[3] T. J. Treangen and S. L. Salzberg, "Repetitive dna and next-generation sequencing: computational challenges and solutions," *Nature Reviews Genetics*, vol. 13, pp. 36–46, 01 2012.

[4] N. Figueroa, "Raider: Rapid ab initio detection of elementary repeats," Master's thesis, Miami University, 2013.

[5] K. Chao and L. Zhang, *Sequence Comparison: Theory and Methods*. Computational Biology, Springer, 2008.

[6] J. Buhler, U. Keich, and Y. Sun, "Designing seeds for similarity search in genomic dna," *J. Comput. Syst. Sci.*, vol. 70, pp. 342–363, May 2005.