Khushwinder Sekhon

Marina Barsky

CSC 428

April 11, 2010

Faster Algorithm For Finding Tandem Repeats

ABSTRACT

Tandem repeats are an array of consecutive repeats. They include three subclasses: satellites, mini-satellites and micro-satellites. Replication slippage, unequal crossing-over and evolutionary pressures generate a high degree of polymorphism in the number of repeats. TRs are therefore useful as genetic markers, such as for DNA fingerprinting, mapping genes, comparative genomics, and evolution studies. Several studies have shown that tandem repeat polymorphism plays an important role in the adaptation of pathogenic bacteria to their host and may also have pharmacological effects in humans. Due to the role played by tandem repeats, finding and studying tandem repeats is of much biological importance. But tandem repeats are not always perfect and can be approximate copies of each other. This property of tandem repeats makes finding all the tandem repeats in a sequence very challenging and time consuming, specially when searching through millions of sequences. This report presents an efficient and fast algorithm for detecting approximate tandem repeats in genomic sequences.

INTRODUCTION

The genome is an organism's entire genetic endowment - the complete nucleotide sequence of its DNA. This "book of life" contains the instructions used to make an entire organism, and guide its growth and development. DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. Most DNA is located in the cell nucleus (where it is called nuclear DNA), but a small amount of DNA can also be found in the mitochondria (where it is called mitochondrial DNA). The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule.

The genome of an organism is inscribed in DNA, or in the case of some viruses, RNA. The portion of the genome that codes for a protein or an RNA is referred to as a gene. Those genes that code for proteins are composed of tri-nucleotide units called **codons**, each coding for a single amino acid. For e.g. CGA codes for Arginine and CAA codes for Glutamine. The nucleotide sequences have been the focus of a lot of research in he biological world since they provide so much information about the body and its inner workings. A certain type of sequences occur in the genome where a pattern of nucleotide bases are repeated one or more times; these repeating patterns are known as Tandem Repeats.

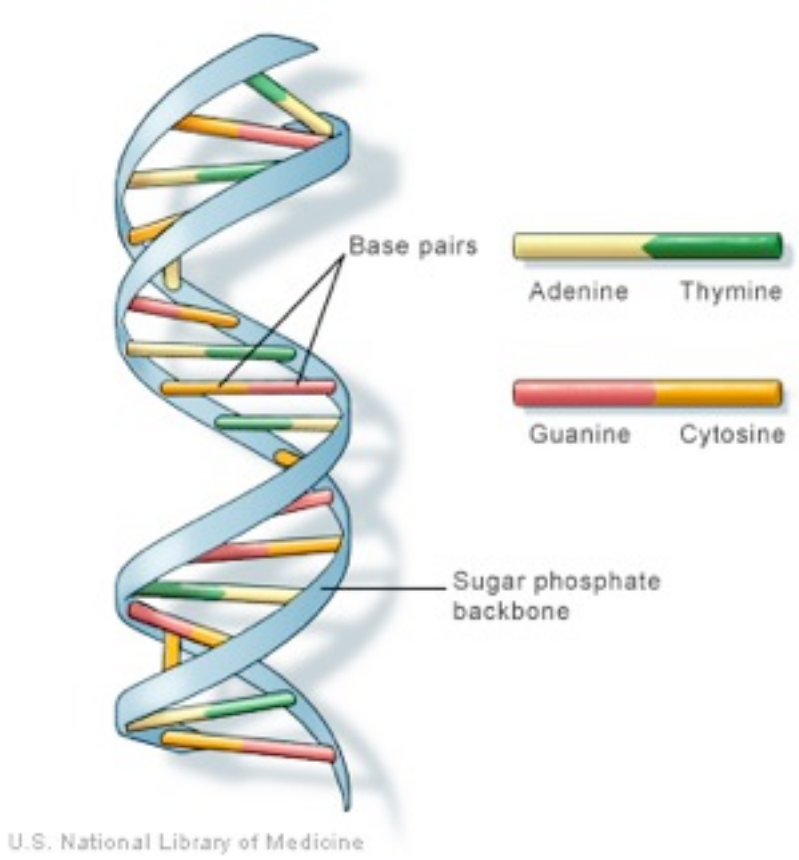

Fig 1. DNA structure and base pairing (1)

TANDEM REPEATS

Tandem repeats occur in DNA when a pattern of two or more nucleotides is repeated and the repetitions are directly adjacent to each other. For e.g. .... TGCA.... becomes ..... TGCGCGCA..... where bases G and C are repeated three times (di-nucleotide repeat). The pattern can repeat anywhere from a couple of repetitions to more than a hundred repetitions. The origin of these repeats, as well as their biological function, is not fully understood. Nevertheless, they are believed to play an important role in genome organization and

evolution. Based on the number of repetitions, tandem repeats have been categorized into three subclasses: micro-satellites, mini-satellites and satellites.

Micro-satellites, more commonly known as short tandem repeats (STR), have a repeat unit of only 1 to 8 bp and the whole repetitive region spans less than 150 bp (2). By identifying repeats of a specific sequence at specific locations in the genome, it is possible to create a genetic profile of an individual. There are currently over 10,000 published STR sequences in the human genome (3). STR analysis has become the prevalent analysis method for determining genetic profiles in forensic cases.

Mini-satellites are sections of DNA that have a repeat unit of 9 to 80 bp. These variant repeats are tandemly intermingled, which makes mini-satellites ideal for studying DNA turnover mechanisms. Mini-satellites are mostly located in the non coding regions of the DNA. For e.g. Telomere, a region of repetitive DNA at the end of the chromosome, contains tandemly repeated sequence GGGTTA. The purpose of the repeated sequence is to protect the end of the chromosome from deterioration since the last bit of DNA sequence does not get copied during DNA replication. So telomere's are added to the end of the sequence after every replication process only to protect the coding regions of DNA.

Satellites have a repeat unit of above 80 bp. In humans, a well known example is the **alphoid** DNA located at the centromere of all chromosomes. Its repeat unit is 171 bp and the repetitive region accounts for 3-5% of the DNA in each chromosome. Other satellites have a shorter repeat unit. Most satellites in humans or in other organisms are located at the centromere (2).

WHY FIND TANDEM REPEATS ?

Many STR repeat sequences are polymorphic in copy number in human populations. These are therefore a rich source of DNA polymorphisms that have been exploited widely for studies of the human genome. Polymorphism describes the existence of different forms within a population, e.g., difference in the number of tandem repeats. One feature of the trinucleotide STRs is their ability to undergo dynamic mutation; the process of change in genetic material that can occur over several generations (6). These dynamic mutations of tandem repeats have been linked to a number of genetic diseases such as Huntington's disease, Fragile-X mental retardation, Myotonic dystrophy and muscular dystrophy. Therefore, finding tandem repeats has helped in the research of genetic diseases. Studies have shown that tandem repeats located in regulatory regions can cause disease by influencing gene expression; for example, a tandem repeat polymorphism in the dopamine transporter (DAT) gene has been associated with Parkinson's disease in the Korean population (6).

The polymorphic property of tandem repeats has also been used in genetic profiling (DNA profiling) that has helped with law enforcement investigations to identify individuals. Since the number of repeats for a given micro/mini - satellite is quite unique for each individual, this property is used for DNA fingerprinting. For e.g. the FBI uses a set of 13 specific STR regions when performing DNA fingerprinting of individuals during investigations because the odds of two individuals having the same repeat pattern in all 13 regions is one in a billion. Variable Number Tandem Repeats (VNTR), a type of mini-satellites, describe a pattern that can be used to help determine parental linkage since tandem repeats also get passed on to a child from both parents. The same property is used in DNA paternal tests and to identify inherited traits. Fig 2 shows the difference in VNTR allele lengths in 6 indi-

viduals. Variable Number of Tandem Repeats (VNTR) loci provide a source of very informative markers in bacteria. Tandem repeats for bacterial identification have proven their utility for the typing of highly monomorphic pathogens such as Bacillus anthracis, Yersinia pestis, Mycobacterium tuberculosis, or Staphylococcus aureus (4). So, finding tandem repeats is very beneficial in the biological world. But because of the vast number and size of tandem repeats, finding all tandem repeats can pose difficult challenges to biologists.

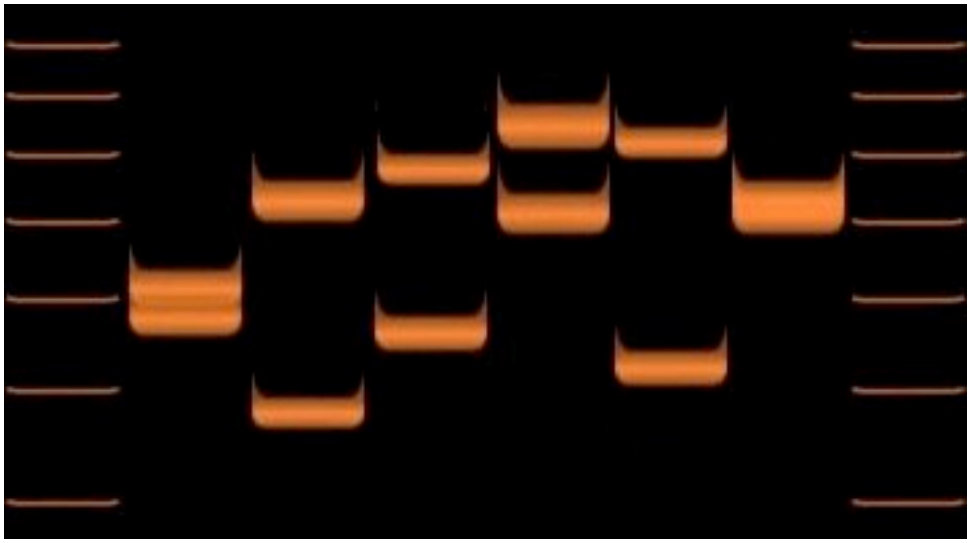

Fig 2. Variations of VNTR allele lengths in 6 individuals

CHALLENGES

There are many challenges and limitations in the biological computation field in regard to finding tandem repeats. Most of the research has been done on finding ways to look for short tandem repeats because these repeats are easier to find. Larger tandem repeats are notoriously harder to find and the inability to find them can affect other biological computations on DNA sequences. For e.g. producing multiple alignment between 2 sequences becomes very complex when multiple tandem repeats are present in the sequences. Another challenge faced by biologists is that tandem repeats are not always perfect; a copy of a pat-

tern may not be exactly similar to another, but very close. The variation comes from events such as mutations, translocations and reversals and result in Approximate Tandem Repeats (ATR). An ATR is defined as a string of nucleotides repeated consecutively at least twice with small differences between the instances. Finding ATRs in a sequence is a harder task than finding perfect repeats. The scope of ATRs discovered by some of the algorithmic approaches are limited by constraints on the input data, search parameters, the type of allowed mutations and the number of such mutations (4). In others, time requirements render the algorithm infeasible for the analysis of whole genomes containing millions of base pairs.

Using longest common extension queries, tandem repeats can be found in O(n^2) time (5). Similarly, a k-mismatch tandem repeat finding problem to search for ATRs finds tandem repeats in O (kn^2) time complexity (5). In this paper, we present an approach to finding ATRs in genomic sequences. The algorithm, presented by Dan Gusfield in his book 'Algorithms in Strings, Trees and Sequences' (5), uses the Landau Schmidt algorithm to split the problem of finding all ATRs into 4 subproblems such that no repeat is ever found more than once.

ALGORITHM

Landau and Schmidt developed a method, that is a recursive divide-and-conquer approach that exploits the ability to compute longest common extension queries in constant time, to find all k-mismatch tandem repeats in O(kn log(n/k) + z) time, where z is the number of k-mismatch tandem repeats in the string S (5). The algorithm by Gusfield uses the Landau Schmidt algorithm to find all tandem repeats by adapting it to find all tandem repeats (with no mismatches) in O(n log n + z) time.

Landau Schmidt method divides the problem into four subproblems:

For a given sequence S of length n, let h = n/2

1. Find all tandem repeats contained entirely in the first half of S (up to position h)
2. Find all tandem repeats contained entirely in the second half of S (after position h)
3. Find all tandem repeats where the first copy spans position h
4. Find all tandem repeats where the second copy spans position h

The first two subproblems are solved by recursively applying the Landau Schmidt algorithm. The second two subproblems (3 & 4) are symmetric to each other, so the same algorithm will work for both. The algorithm for, say subproblem 3, will determine the algorithm for finding all tandem repeats. The idea of the algorithm is that for any fixed number L, one can test in constant time whether there is a tandem repeat of length exactly 2L such that the first copy spans position h. Applying this test for all feasible values of L means that in linear time, we can find all lengths of tandem repeats whose first copy spans position h. The pseudocode for the algorithm is as follows:

- Let q = h+L
- Compute the longest common extension (in forward direction) from positions h and q. Let L1 denote the length of that extension.
- Compute the longest common extension (in reverse direction) from positions h - 1and q - 1. Let L2 denote the length of that extension.
- There is a tandem repeat of length L whose first copy spans position h if and only if L1 + L2 >= L and both L and L2 are at least one. If there is such a tandem repeat of length 2L,

then it can begin at any position from Max(h - L2, h - L + 1) to Min(h + L1 - L, h) inclusive. The second copy of the repeat begins L places to the right. Figure 3 explains the algorithm in a demonstrative way.

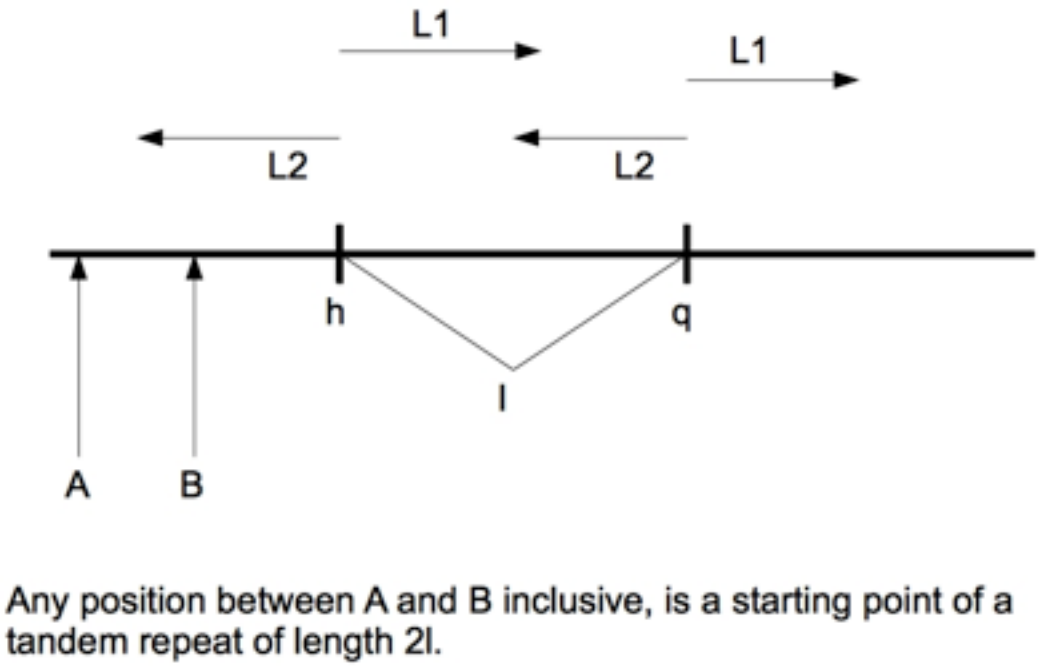

Fig 3. Algorithm for subproblem 3 & 4 (5)

To solve subproblem 3, we can run the above algorithm for each L from 1 to h. For subproblem 4, we can run the algorithm for each L from h to n. The algorithms for subproblem 3 runs in O(n/2) + z, where z is the number of such tandem repeats. The time complexity for the entire algorithm (steps 1 through 4) is O (n logn + z) where z is the total number of tandem repeats in the sequence. This algorithm is faster and more efficient than O (n^2) taken by using straight forward longest common extension queries.

Works Cited

1. "What is DNA?" *Genetics Home Reference - Your guide to understanding genetic conditions*. N.p., n.d. Web. 12 Apr. 2010. <http://ghr.nlm.nih.gov/handbook/basics/dna>.

2. "Tandem Repeats." *Web Books Publishing*. N.p., n.d. Web. 12 Apr. 2010. <http://www.web-books.com/MoBio/Free/Ch3G1.htm>.

3. "Short tandem repeat" *Wikipedia, the free encyclopedia*. N.p., n.d. Web. 12 Apr. 2010. <http://en.wikipedia.org/wiki/Short_tandem_repeat>.

4. Delgrange, Olivier, and Eric Rivals. "STAR: an algorithm to Search for Tandem Approximate Repeats." *Bioinformatics* 20.16 (2004): 2812-2820. Print.

5. Gusfield, Dan. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. New York: Cambridge University Press, 1997. Print.

6. Sutherland, Grant R, and Robert Richards. "Simple tandem DNA repeats and human genetic disease." *Proceedings of the National Academy of Sciences* 92 (1995): 3636-3641. Print.