

# TECHNIQUES FOR DETECTING APPROXIMATE TANDEM REPEATS IN DNA

Thao T. Tran

Vincent A. Emanuele II

G. Tong Zhou

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA

## ABSTRACT

The detection of tandem repeats is important in biology and medicine as it can be used for phylogenetic studies and disease diagnosis. This paper proposes two techniques for detecting approximate tandem repeats (ATRs) in DNA sequences. First, an evolutionary force heuristic is proposed to merge interspersed perfect tandem repeat sequences into an ATR structure. Next, a more general method is presented using a product spectrum of Fourier nucleotide subsequences to detect hidden periodicity. The Fourier method is sensitive to ATRs and is robust in the presence of substitutions, insertions, and deletions.

## 1. INTRODUCTION

Repetitive structures are present in over one-third of the human genome [1]. Genomic repeats can be categorized into two classes: transposable interspersed repetitive elements i.e., Long and Short Interspersed Elements (LINEs, SINEs), and tandem repeats. The focus of this paper will be on the identification of approximate tandem repeats (ATRs).

Tandem repeats (TRs) are defined as two or more contiguous approximate copies of a pattern of nucleotides. Tandem repeats have been known to play important roles in human disease, regulation, and evolution. Fourteen documented trinucleotide repeat expansion diseases (TREDs) that affect humans are listed in Table 1.

**Table 1. Trinucleotide Repeat Expansion Disorders.**

Name of Disorder	Repeat Pattern	Normal Copies	Mutant Copies
DRPLA	CAG	6-35	49-88
Huntington's	CAG	11-34	40-120
Kennedy's	CAG	9-36	38-62
Type1 SCA1	CAG	25-36	41-81
Type2 SCA2	CAG	15-24	35-59
Type3 SCA3	CAG	13-36	62-82
Type6 SCA6	CAG	4-16	21-33
Type7 SCA7	CAG	4-19	37-306
Fragile X	CGG	6-54	70-230
Fragile XE	GCC	6-35	> 200
Friedreich's Ataxia	CGG	6-54	> 100
Myotonic Dystr.	CTG	5-27	50-1000
Type 8 SCA8	CTG	16-34	110-250
Type 12 SCA12	CAG	7-28	28-66

The expansion of the trinucleotide repeat results in anticipation or progression in severity of the disorder through each generation. In general, there is a correlation between the size of the expansion and the severity of the phenotype. Furthermore, instabilities in dinucleotide repeat sequences have been observed in colon cancer [2]. Some biological mechanisms for the expansion of repeats include: de-

fect in mismatch repair system, polymerase slippage during replication, and genetic instability of some DNA structures [3, 4].

Repeats play a role in gene regulation when present in regions with transcription factors [5]. In addition, the evolutionary history of an organism can be mapped by identifying duplication events of repeats over time.

An abundance of TR patterns have been located in NIST's Short Tandem Repeat DNA Database (STRBase) [6] and are used in human gene mapping, linkage studies, and forensic DNA fingerprinting analysis [7]. The detection of repeats is also important for similarity searches as repetitive structure tends to confuse sequence alignment algorithms.

Signal processing methods offer great promise in analyzing genomic data as evidenced by research in the area in recent years [8, 9, 10].

It is well known that the identification of perfect tandem repeats can be found in approximately linear time by the use of string algorithms based on suffix tree approaches [11]. ATRs pose more of a challenge as substitutions, insertions, and deletions through millions of years of evolution make the repeats harder to detect. Recently algorithms have been proposed [12, 13, 14], each with its own limitations and assumptions. In [12], the period of the repeat is limited to less than 2000 base pairs (bp) as of version 3.21. In [14], there are practical memory constraints resulting from the pattern extension algorithm. Although there are no limitations on the period size in [13], the algorithm does not deal with insertions or deletions (indels) directly, and as such, certain repeats with indels can be missed. In this paper, we propose two possible approaches for detection of ATRs in DNA. First, a heuristic method, called the evolutionary force heuristic, is proposed that builds ATR regions by associating neighboring perfect tandem repeats. Next, a more general Fourier analysis based method is proposed which is sensitive in detecting repeats with high percentage of indels and does not require the user to specify the number of allowed mismatches when searching for ATRs.

## 2. EVOLUTIONARY FORCE HEURISTIC

In the introduction section, it was noted that the problem of finding perfect tandem repeats (PTRs) in DNA has been well studied and that there currently exist efficient solutions for finding PTRs. However, the theory of evolution predicts that over time random mutations will occur in the DNA code [15]. Typically, the mutation rate for DNA replication can range from about  $10^{-9}$  to  $10^{-7}$ . DNA replication occurs when cell divisions take place. Consider that, in humans, it is estimated that approximately  $10^{16}$  cell divisions occur in our lifetime [15]. Thus, it is reasonable to hypothesize that there exist regions of approximate repetitive structures, or approximate tandem repeats, in the genetic code that may have diverged from perfect repeats over time

through the mechanism of evolution. In fact, one may wonder if stretches of PTRs that are in close proximity to each other were once joined together as part of one longer PTR sequence. In this section, a method of associating PTRs in close proximity is proposed that gives rise to ATR structures.

First, some mathematical preliminaries and notational matters are necessary to facilitate proper formulation of the problem. Let  $\Sigma$  be a DNA sequence of length  $N$ ,  $\Sigma = (s_1, s_2, \dots, s_N)$ ,  $s_k \in D = \{A, T, G, C\}$ . For notational convenience, the elements in  $\Sigma$  will be written without the commas. Furthermore, let  $D^N$  represent the space of all possible DNA sequences of length  $N$ . A subsequence,  $\sigma$ , of  $\Sigma$  is defined as

$$\sigma_{ij} = \Sigma([i, \dots, j]) \triangleq (s_i s_{i+1} \dots s_j), \quad \text{for } 1 \leq i \leq j \leq N.$$

A perfect tandem repeat (PTR) in  $\Sigma$  is defined as follows. Let  $\sigma$  be a subsequence of  $\Sigma$ . We define  $\sigma$  to be a PTR region of  $\Sigma$  if there exists a subsequence  $\sigma^{(P)}$  of  $\sigma$ ,  $\sigma^{(P)} = (s_1 s_2 \dots s_P) \in D^P$  such that  $\sigma$  can be written in the following notation,

$$\underbrace{(s_1 s_2 \dots s_P)}_{P \text{ nucleotides}}^M \triangleq \underbrace{(s_1 s_2 \dots s_P s_1 s_2 \dots s_P \dots s_1 s_2 \dots s_P)}_{(s_1 s_2 \dots s_P) \text{ repeated } M \text{ times}}.$$

The quantity  $P$  is the period of the repeat, and  $M$  is the copy number of the repeat. Now the PTR finding problem can be stated precisely: find the set of all subsequences  $\{\sigma_{i_1 j_1}, \sigma_{i_2 j_2}, \dots, \sigma_{i_k j_k}\}$  for which the definition of PTR given above holds. Note that in practice, the sequence length  $N$  can be very large. Thus, the existence of PTR patterns with very large periods cannot be ruled out.

At this point, observe the following:

1. The problem of finding PTRs in a sequence can be defined precisely.
2. There exist algorithms for finding PTRs efficiently.
3. It is possible for PTRs to degenerate into ATRs over time as the result of random mutations.

Given these three observations, a first attempt at finding ATRs is possible by associating appropriate PTRs that are close to each other by some heuristic method. We propose the evolutionary force heuristic, to be described next.

Two subsequences,  $\sigma_{i_1 j_1}$  and  $\sigma_{i_2 j_2}$  that are PTRs are said to be of the same *type* if they have the same period  $P$  and there exists a cyclic permutation  $f_P$  of  $\sigma_{i_2, i_2+P-1}$  such that  $\sigma_{i_1, i_1+P-1} = f_P(\sigma_{i_2, i_2+P-1})$ . Furthermore, the mass,  $m_k$ , of  $\sigma_{i_k j_k}$  is defined as,  $m_k = j_k - i_k + 1$ . The center of mass,  $c_k$ , of a subsequence  $\sigma_{i_k j_k}$  is defined as,  $c_k = (i_k + j_k)/2$ . Given  $\sigma_{i_1 j_1}$  and  $\sigma_{i_2 j_2}$ , two PTR subsequences of  $\Sigma$  of the same type, define the evolutionary force  $F_\epsilon$  between the PTRs as,

$$F_\epsilon = \frac{m_1 m_2}{d_{12}^2}, \quad d_{12} = |c_1 - c_2|. \quad (1)$$

In the evolutionary force heuristic algorithm, two PTR subsequences,  $\sigma_{i_1 j_1}$  and  $\sigma_{i_2 j_2}$ , of the same type with sufficient evolutionary force attraction will be merged into an ATR subsequence,  $\sigma^{(ATR)} = \sigma_{i_1 j_2}$ . The algorithm will be stated in more detail shortly. First, observe that the evolutionary force quantity,  $F_\epsilon$ , has the following two properties.

**Property 1:**  $0 \leq F_\epsilon \leq 1$ .

**Proof:** Since  $m_1 > 0$ ,  $m_2 > 0$ , we must have  $F_\epsilon \geq 0$  (equality in the limit as  $d_{12} \rightarrow \infty$ ). Now, note that

$d_{12} \geq \frac{(m_1 + m_2)}{2}$ . From properties of arithmetic and geometric means, we observe,

$$d_{12}^2 \geq \frac{(m_1 + m_2)^2}{4} \geq m_1 m_2,$$

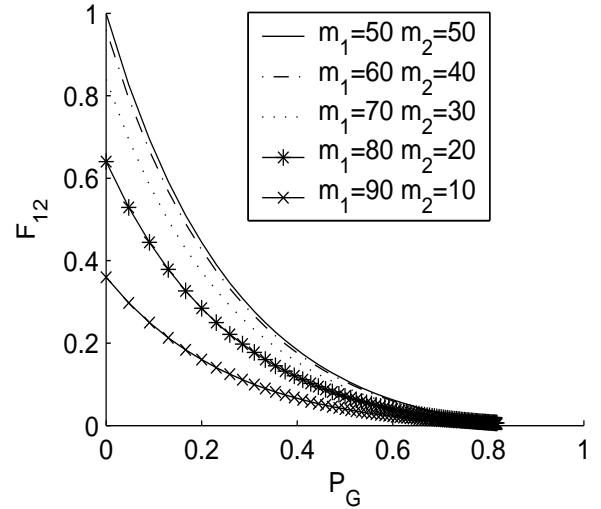
with equality in the above equation if and only if  $m_1 = m_2 = \frac{(m_1 + m_2)}{2}$ . Combining this with eq. (1), we conclude that  $F_\epsilon \leq 1$ .

**Property 2:** Percent of non-PTR nucleotides,  $P_G$ , in  $\sigma_{i_1, j_2}$  after the first merge is given by,

$$P_G = \frac{d_{12} - \frac{m_1 + m_2}{2}}{d_{12} + \frac{m_1 + m_2}{2}}. \quad (2)$$

The functional dependence of  $F_{12}$  and  $P_G$  is shown in Fig. 1. For a threshold of about  $t_F = 0.08$ , no two PTR subsequences will be merged with more than about 66.7 percent non-PTR nucleotides. The algorithm for constructing ATRs based on the evolutionary force heuristic is summarized in the list below.

1. Pick a maximum tolerable  $P_G$  for which two PTRs will be associated and merged as one.
2. With this value in mind, pick a threshold  $t_F$  by observing Fig. 1.
3. For any two PTR subsequences,  $\sigma_{i_1 j_1}$  and  $\sigma_{i_2 j_2}$ , of the same type such that  $i_1 \leq j_1 \leq i_2 \leq j_2$ , compute the evolutionary force,  $F_\epsilon$ , between them and compare with the threshold  $t_F$ . If  $F_\epsilon > t_F$ , define the new ATR subsequence as  $\sigma^{(ATR)} = \sigma_{i_1 j_2}$ .
4. Update mass of the newly constructed ATR subsequences, recalculate the forces values between all repeats of the same type.
5. If converged (no more subsequences to merge), exit; otherwise go to step 3.



**Figure 1. Functional dependence of  $F_{12}$  and  $P_G$**

Using this method combined with a PTR detection algorithm and testing on the human dystrophia myotonica-protein kinase (GenBank accession number: NM\_004409) gives the interesting result of (AGAGAGAAGTGGCCA-GAGAG) as an ATR subsequence. Note that many repeats

(both PTR and ATR) were found, but omitted for the sake of conciseness. The intention here was to present an example of what sort of results can be expected using this method.

In general, defining precisely what is meant by ATR is somewhat difficult. For example, in [16], two criteria, hamming distance and edit distance between adjacent periods, are used in defining an ATR. In the case of hamming distance, a parameter in the algorithm,  $k$ , sets the number of allowed mismatches between adjacent periods in an ATR. The problem with using this definition of ATR is that the appropriate choice of  $k$  depends on  $P$ . Thus, one must somehow know in advance (or have a good guess) what period repeats are going to exist in the sequence under consideration.

The heuristic evolutionary force post-processing algorithm presented in this section is convenient and easy to implement. However, the evolutionary force criterion still does not define precisely what is meant by an ATR (note that property 2 is valid only in the first round of merging). In particular, this heuristic method is not as general as other ATR finding algorithms such as [12, 13, 14]. A more general approach to finding ATRs, or approximate periodicities, in DNA can be formulated by using the Fourier product method, described in the next section.

### 3. FOURIER PRODUCT METHOD

Our objective here is to detect ATRs within an observation window of length  $N$ . Since ATRs themselves may not be strictly periodic, and “random” bases appear before and/or after the ATRs, sensitivity of the method in detecting the hidden periodicity must be high. We do not assume any knowledge about the pattern that is being repeated, the size (period) of the pattern, nor the location of the repeats.

Fourier based methods are natural for this kind of problems. We describe next, our proposed algorithm based on a Fourier product spectrum.

#### Step 1. Convert the DNA sequence into four nucleotide subsequences $x_A[n]$ , $x_T[n]$ , $x_G[n]$ , $x_C[n]$ .

Let  $x_\alpha[n] = 1$  if character  $\alpha$  is present at the  $n$ th position of the DNA sequence;  $x_\alpha[n] = 0$  otherwise;  $\alpha \in \{A, T, G, C\}$ . Therefore,  $x_\alpha[n]$  is an indicator sequence for the presence or absence of character  $\alpha$  in the DNA sequence. For example, Table 2 shows the  $x_\alpha[n]$  components for DNA sequence ‘ACTGCTAGCAAT’.

Table 2. Numerical Subsequences.

$\Sigma$	A	C	T	G	C	T	A	G	C	A	A	T
$x_A[n]$	1	0	0	0	0	0	1	0	0	1	1	0
$x_T[n]$	0	0	1	0	0	1	0	0	0	0	0	1
$x_C[n]$	0	1	0	0	1	0	0	0	1	0	0	0
$x_G[n]$	0	0	0	1	0	0	0	1	0	0	0	0

#### Step 2. Take (normalized) Fourier transform of the mean removed processes.

Let  $m_\alpha = \frac{1}{N} \sum_{n=0}^{N-1} x_\alpha[n]$  and calculate

$$S_\alpha(f) = \frac{1}{N} \sum_{n=0}^{N-1} (x_\alpha[n] - m_\alpha) e^{-j2\pi fn}, \quad (3)$$

for  $0 \leq f \leq 0.5$  and  $\alpha \in \{A, T, G, C\}$ .

#### Step 3. Form the Fourier product spectrum

$$S(f) = \prod_{\alpha \in \{A, T, G, C\}} (|S_\alpha(f)| + c), \quad (4)$$

where  $c$  is a small positive constant. If a period  $P$  repeat exists in the DNA sequence,  $S(f)$  should show a peak at  $f = 1/P$ . It is possible for  $S(f)$  to peak at  $f = 2/P, 3/P, \dots$  as well, but we only need to pay attention to the fundamental frequency. The period  $P$  can thus be inferred from the peak location. The constant  $c$  is to prevent the nulling of  $S(f)$  if a particular character is not part of the repeat pattern.

#### Step 4. Detection of the beginning and end of the ATR regions.

Details are omitted here due to the space limitation.

In [10], a sum spectrum instead of a product spectrum was proposed. Our experience is that the product spectrum (4) is much more sensitive to ATRs, as we illustrate in the following example.

**Example 1.** Consider the pseudo-DNA sequence:

ACTGACCGGACGC [ATGATGCTGATGATG] CTAC

Figure 2 shows the individual Fourier transform magnitudes  $|S_A(f)|$ ,  $|S_T(f)|$ ,  $|S_G(f)|$ ,  $|S_C(f)|$ , the product spectrum  $S(f)$  (with  $c = 0.01$ ), as well as the sum spectrum. There was not a clear peak in any of the plots except for the product spectrum. The peak in  $S(f)$  is located at  $f = 0.33$ , indicating that a period  $P = 3$  repeat is present in the DNA sequence. Based on this information, we then determined that the pattern **tga** was repeated 5 times at positions 14-28 with 1 substitution at position 20.

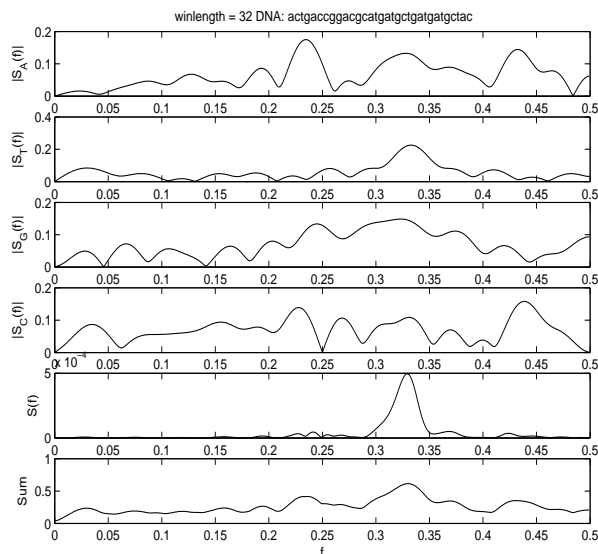


Figure 2. From top to bottom:  $|S_A(f)|$ ,  $|S_T(f)|$ ,  $|S_G(f)|$ ,  $|S_C(f)|$ , the product spectrum  $S(f)$ , and the sum spectrum for the DNA sequence in Example 1.

Since multiplication is a nonlinear operation, it is expected that peaks are enhanced while the “noise floor” is suppressed in a product spectrum. Our computer simulations show that the product spectrum (4) is especially sensitive to the presence of ATRs, and can tolerate up to 12% of indels and up to 28% of substitutions.

**Example 2.** Next, we show the performance of the product spectrum (4) on the Apis mellifera (bee) sequence (GenBank accession: AMU73928) given below:

CCCATGTCCCAGCGCGTATTGCTTTTGCATCGGAACGCACTTTCAATGT  
CCCAGCGCGTATTGCTTCTATTTTATAAGTACCAGCTAAATTTTTTTT

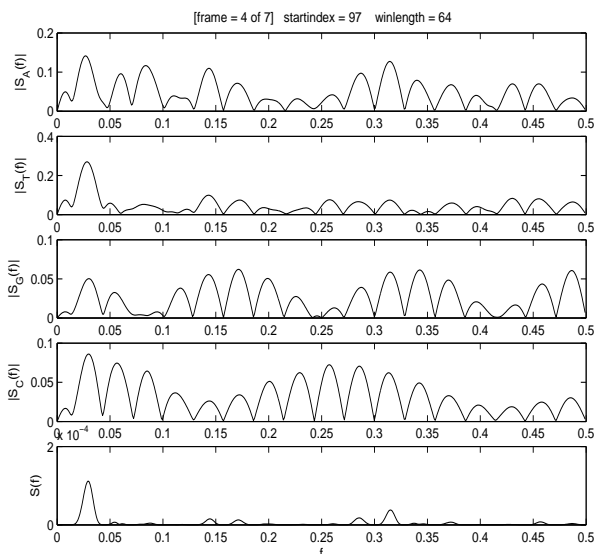
TTTTTTTATAAGTACCAGCTAAAATTTTTTTTTTTTTTTTATAAGTAC  
 CAGCTAAAATTTTTTTTTTTTTTTTTTATAAGTACCAGCTAAAATTTTT  
 TTTTTTTTATAAGTCCAGCGCGTATTGCTTCTGAAATTTAAAAAAA  
 AAAAAAATTTTTTTTAATAATATATTATATA

Figure 3 shows the individual spectrum  $|S_\alpha(f)|$  and the product spectrum  $S(f)$ . Based on the product spectrum, we were able to detect a length 35 ATR with an estimated copy number of 5 at positions 65-256. This was consistent with results found by [14, 12] at positions 72-221 and 72-209, respectively. We show the ATR region using the “multiple alignment” format produced below for positions 65-256 by ClustalW v1.82 [17] (rows correspond to periods)

```
gctTctaTTTTATAAGTACCAGCTAAAATTTTTTT-----
-----TTTTTTTATAAGTACCAGCTAAAATTTTTTTT-----
-----TTTTTTTATAAGTACCAGCTAAAATTTTTTTT-----
-----TTTTTTTATAAGTACCAGCTAAAATTTTTTTT-----
-----TTTTTTATAAGTcCAGCGgcgctaTgCtTtTgaa
```

We see that the repeats are approximate as there are deletions (indicated by -) and substitutions (indicated by lower case letters).

Imagine having to manually detect these ATRs! Signal processing techniques are especially useful for detecting large patterns or patterns that are repeated few times.



**Figure 3.** From top to bottom:  $|S_A(f)|$ ,  $|S_T(f)|$ ,  $|S_G(f)|$ ,  $|S_C(f)|$ , and the product spectrum  $S(f)$  for sequence AMU73928.

#### 4. CONCLUSIONS

In this paper, two methods have been proposed for the detection of ATRs in DNA sequences.

The evolutionary force heuristic method inherently assumes that PTR subsequences of the same type in close proximity are already found. Then, it merges the PTR seeds into a more general ATR structure. Since the evolutionary force quantity is normalized as seen by property 1, it is easy to draw a threshold to decide on the merge. It should be noted that this method identifies a subset of all possible ATRs that occur in DNA sequences. However, because the ATR regions identified are of a specific type, perhaps they have a direct biological link to evolution?

A more general approach using the Fourier product of nucleotide subsequences has shown strong robustness in detecting ATRs, especially those with substitutions and indels. In this method, the period to be detected in a given DNA sequence is limited by the window length. However, it offers a promising approach due to its high sensitivity in identifying hidden periodicity in genomic data.

**Acknowledgment:** This work was supported in part by the US National Science Foundation under the Graduate Fellowship Program.

#### REFERENCES

- [1] E. S. Lander *et al.*, “Initial sequencing and analysis of the human genome,” *Nature*, vol. 409, pp. 860–921, February 2001.
- [2] S. Thibodeau, G. Bren, and D. Schaid, “Microsatellite instability in cancer of the proximal colon,” *Science*, vol. 260, pp. 816–819, 1993.
- [3] M. Mitas, “Trinucleotide repeats associated with human disease,” *Nucleic Acids Research*, vol. 25, pp. 2245–2253, 1997.
- [4] R. D. Wells, “Molecular basis of genetic instability of triplet repeats,” *Journal of Biological Chemistry*, vol. 271, pp. 2875–2878, 1996.
- [5] M. Perutz, “Glutamine repeats and inherited neurodegenerative diseases: molecular aspects,” *Current Opinion in Structural Biology*, vol. 6, pp. 848–858, 1996.
- [6] C. Ruitberg, D. Reeder, and J. Butler, “Strbase: a short tandem repeat dna database for the human identity testing community,” *Nucleic Acids Research*, vol. 29, pp. 320–322, 2001.
- [7] Y. Nakamura *et al.*, “Variable number of tandem repeat (vntr) markers for human genome mapping,” *Science*, vol. 235, pp. 1616–1622, 1987.
- [8] P. P. Vaidyanathan and B.-J. Yoon, “Gene and exon prediction using allpass-based filters,” in *Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, (Raleigh, NC), October 2002.
- [9] D. Anastassiou, “Genomic signal processing,” *IEEE Signal Processing Magazine*, pp. 8–20, July 2001.
- [10] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, “Prediction of probable genes by fourier analysis of genomic sequences,” *CABIOS*, vol. 13, pp. 262–270, 1997.
- [11] D. Gusfield and J. Stoye, “Linear time algorithms for finding and representing all the tandem repeats in a string,” report cse-98-4, Dept. of Computer Science, University of California, Davis, 1998.
- [12] G. Benson, “Tandem repeats finder: a program to analyze dna sequences,” *Nucleic Acids Research*, vol. 27, no. 2, pp. 573–580, 1999.
- [13] R. Kolpakov, G. Bana, and G. Kucherov, “mreps: efficient and flexible detection of tandem repeats in dna,” *Nucleic Acids Research*, vol. 31, no. 13, pp. 3672–3678, 2003.
- [14] A. M. Hauth and D. A. Joseph, “Beyond tandem repeats: complex pattern structures and distant regions of similarity,” *Bioinformatics*, vol. 18, no. 1, pp. S31–S37, 2002.
- [15] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*. New York and London: Garland Publishing, Inc, 1998.
- [16] G. M. Landau, J. P. Schmidt, and D. Sokol, “An algorithm for approximate tandem repeats,” *Journal of Computational Biology*, vol. 8, no. 1, pp. 1–18, 2001.
- [17] J. D. Thompson *et al.*, “Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice,” *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.