New Algorithms for Finding Monad Patterns in DNA Sequences

Conference Paper *in* Lecture Notes in Computer Science · October 2004

DOI: 10.1007/978-3-540-30213-1_40 · Source: DBLP

2 authors, including:

Ravi Vijaya Satya
QIAGEN Sciences, Inc.
37 PUBLICATIONS **713** CITATIONS

SEE PROFILE

New Algorithms for Finding Monad Patterns in DNA Sequences*

Ravi Vijaya Satya and Amar Mukherjee

[rvijaya,amar]@cs.ucf.edu
School of Computer Science, University of Central Florida
Orlando, FL USA 32816-2362

Abstract. In this paper, we present two new algorithms for discovering monad patterns in DNA sequences. Monad patterns are of the form (l,d)-k, where l is the length of the pattern, d is the maximum number of mismatches allowed, and k is the minimum number of times the pattern is repeated in the given sample. The time-complexity of some of the best known algorithms to date is $O(nt^2l^d\sigma^d)$, where t is the number of input sequences, n is the length of each input sequence, and $\sigma = |\sum|$ is the size of the alphabet. The first algorithm that we present in this paper takes $O(n^2t^2l^{\frac{d}{2}})$ time and $O(ntl^{\frac{d}{2}}\sigma^{\frac{d}{2}})$ space, and the second algorithm takes $O(n^3t^3l^{\frac{d}{2}}\sigma^{\frac{d}{2}})$ time using $O(l^{\frac{d}{2}}\sigma^{\frac{d}{2}})$ space. In practice, our algorithms have much better performance provided the d/l ratio is small. The second algorithm performs very well even for large values l and d as long as the d/l ratio is small.

1 Introduction

Discovering regulatory patterns in DNA sequences is a well known problem in computational biology. Due to mutations and other errors, the actual occurrences of these regulatory patterns allow for a certain degree of error. There fore, the actual regulatory pattern (or the consensus pattern) may never appear in a gene upstream region, but d-mismatch occurrences of this pattern might appear. The general approach to this problem is to take a set of t DNA sequences each of length n, at least k of which are guaranteed to contain the desired binding site, and look for patterns of a certain length l that occur in at least k out of the t sequences with at most d mismatches at each occurrence. The values of l, d and k can be determined either from prior knowledge about the binding site, or by trial and error, trying different values of l and d. These single contiguous blocks of patterns are called *monad* patterns.

In general, many regulatory signals are made up of a group of monad patterns occurring within a certain distance form each other [EskKGP03, EskP02, GuhS01, vanRC00]. In such a case, the patterns are called dyad, triad, multi-ad, or in general as composite patterns. Finding the composite patterns by finding the component monad patterns individually is significantly more difficult, since

* This research was partially supported by NSF grant number: ITR-0312724.

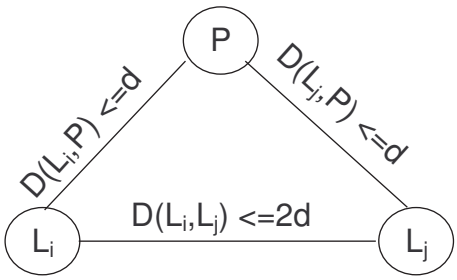

Fig. 1. A pattern P that is consistent with an l-gram pair (L_i, L_j)

the composite monad patterns might be too subtle to detect. Eskin and Pevzner [EskP02] present a simple transformation to convert a multi-ad problem into a slightly larger monad problem. In this paper, we present an algorithm to solve the monad-pattern finding problem. The same transformation as in [EskP02] can be applied to transform a multi-ad problem into a monad problem that is handled by our algorithm.

Pevzner and Sze [PevS00] have put forward a challenge problem: to find the signal in a sample of $t = 20$ sequences, each 600 nucleotides long, each containing an unknown pattern of length $l = 15$ with at most $d = 4$ mismatches. They presented the WINNOWER and SP-STAR algorithms that could solve this problem, which was not solvable by many of the earlier techniques. Many other approaches that can solve this problem have been proposed [Sag98, EskP02, Lia03, BuhT2001]. Time-complexity of the best known algorithms[Sag98, EskP02] is $O(nt^2l^d\sigma^d)$.

Many of the above algorithms search the d-mismatch neighborhood of each l-gram in the sample. The size of the d-mismatch neighborhood of an l-gram in $O(l^d\sigma^d)$. The main motivation for our algorithms is that in most practical scenarios, it might be possible to limit the search to a small portion of the d-mismatch neighborhood. We refer to the set of patterns that mismatch in at most d positions with two l-grams as the *consistent* patterns of the two l-grams. We denote the distance(the number of mismatches) between two l-grams L_i and L_j by $D(L_i, L_j)$. The distance relationships between two l-grams L_i and L_j and a pattern P that is consistent with both of them are shown in Figure 1. The following observations form the basis for our algorithm:

Observation 1: For each l-gram, it is sufficient to search the consistent patterns of the l-gram with respect to all other l-grams.

Observation 2: The number of other l-grams in the sample that are within h mismatches from the current l-gram reduces rapidly with decreasing h. This is illustrated in Figure 2-(a) for a random sample of 20 sequences of 600 nucleotides each. The size of the average $2d$-mismatch neighborhood is 571.395, where as the average size of the d-mismatch neighborhood is just 1.23.

Observation 3: The number of consistent patterns between two l-grams which mismatch in h positions decreases rapidly with increasing h. When h is greater than $2d$, this number is zero, as two l-grams that mismatch in more than $2d$

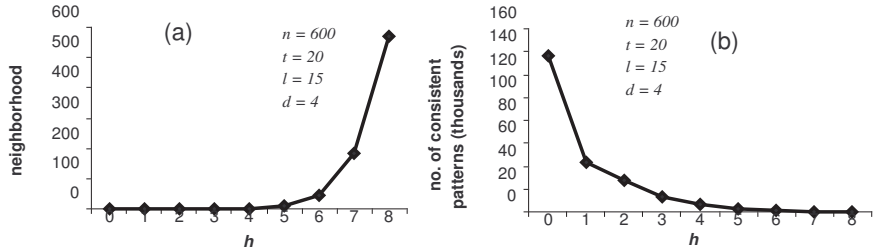

Fig. 2. Variation of: (a)h-mismatch neighborhood (b) consistent patterns with h

positions can not have any patterns that mismatch with both of them in at most d positions.Therefore, as is illustrated in Figure 2-(b), the number of consistent patterns between two l-grams which mismatch in more than d positions is quite small.

2 Previous Approaches to Pattern Discovery

The pattern discovery problem can be formally stated as follows: Given a set of DNA sequences (also referred to as the sample) $S = \{S_1, S_2, ...S_t\}$, and a set of parameters l, d and k, the problem is to find all length-l patterns that occur with up to d mismatches in at least k different sequences in the sample.

One of the earliest techniques to solve this problem, as presented in [PevS00] is known as the pattern driven approach. The pattern driven approach searches all of the pattern space - it enumerates each possible pattern and checks if it meets the search criteria. If the pattern length is l, there are 4^l possible patterns, assuming a DNA alphabet. Pattern driven approaches take each one of these patterns and compare them with all the l-grams in the sample. This approach takes exponential time in terms of l, and the problem quickly becomes practically unsolvable even for moderate values of l.

A faster approach, termed by [EskP02] as the Sample Driven Approach(SDA), searches a reduced search space of only the l-grams that occur in the sample and their d-mismatch neighbors. The SDA algorithm trades in space for time: it maintains a table of size 4^l, each entry in the table corresponding to a pattern. For each l-gram in the input sample, the algorithm enumerates all the patterns that make up its d-mismatch neighbor hood. For each pattern in the neighborhood, the corresponding entry in the table is incremented. After all the l-grams have been processed, the patterns in the table that have a score greater than k are reported. The problem with the SDA approach is that the memory requirements are huge, and increase exponentially with l. Therefore the SDA approach, like the PDA approach, becomes quickly unmanageable, even for moderate values of l.

The WINNOWER algorithm [PevS00] and the cWINNOWER algorithm [Lia03] are based on graph theory. In these algorithms, a graph is constructed in which each vertex is an l-gram in the input sequence. Two l-grams are connected by an edge if they mismatch in at most $2d$ positions. Now, the problem is

mapped to the problem of finding k-cliques in the graph. The problem of finding k-cliques in graph, when $k > 3$ is known to be np-complete. Therefore, WINNOWER and cWINNOWER try to apply some heuristics to arrive at a solution. In the first step, all the nodes that have a degree less than k-1 are removed. After that, different techniques are applied to try to remove the spurious edges in the graph that can not be part of a solution. The complexity of WINNOWER and cWINNOWER for the most sensitive versions of the algorithms are given by $O(t^3n^{2.66})$ and $O(t^4n^4)$, respectively. However, it is important to note that even though it is claimed that the most sensitive versions of these algorithms solve almost all practical problems, they are not guaranteed to solve a given problem.

Some of the other approaches include suffix tree -based approaches [Sag98, PavMP01]. The SPELLER algorithm presented in [Sag98] first builds a suffix tree for the input sequence. It then examines all possible patterns traversing through the suffix tree. If the paths to k different leaves of length l mismatch with the pattern in at most d positions, then the pattern is reported. Starting with zero characters at the root, the pattern is extended one character at a time. At any time if there are less than k different paths in the suffix tree that mismatch in at most d positions with the current pattern, the search is stopped and the (alphabetical) next pattern of the same length, or the next pattern of a shorter length is searched. The complexity of the algorithm is given as $O(nt^2l^d4^d)$.

In the sequence driven approach, each l-gram is searched separately. The Mitra-Count algorithm [EskP02] is based on the idea that if all the l-grams are searched concurrently, then only the information about those l-grams that meet the current search criteria need to be stored. This will reduce the memory requirements drastically. The MITRA algorithm searches the pattern search space in a depth first manner, abandoning the search whenever the search criterion is no longer met. For this it uses the mismatch tree data structure. The path from the root to a node at depth m in the mismatch tree represents a prefix of the pattern of length m. The list of l-grams from the sample whose m-length prefixes mismatch in at most d positions with the path label of the current node are stored at the node. The tree is built in a depth-first fashion. Whenever the size of the list of l-grams at a node falls below k, the node is discarded, and the sub tree of the node is never searched. Whenever the search reaches a depth l, the pattern corresponding to the path label is reported. The algorithm is memory efficient, since only the nodes that lie in the current path need to be stored at any time. An improved algorithm, Mitra-Graph, also presented in [EskP02] applies WINNOWER -like pair wise similarity information in order to maintain a graph at each node of the mismatch tree. If two l-grams L_1 & L_2 mismatch in d_1 & d_2 positions respectively with the node label, and if their suffixes beyond the current depth mismatch in q positions, then the two l-grams are connected by an edge if $d_1 + d_2 + q \geq 2d$. The nodes can be discarded if there is no possibility for a k-clique in the graph. Even though there is an extra overhead of maintaining the graph and extending the graph at each node, much smaller pattern sub-space needs to be searched in Mitra-Graph. The theoretical complexity of Mitra is claimed to be the same as that of the SPELLER algorithm.

3 The PRUNER Algorithm

3.1 Our Contributions

Our approach is based on the WINNOWER algorithm [PevS00, Lia03]. As in WINNOWER, we build a graph based on pair-wise similarity information, and prune the graph eliminating vertices that can not be part of a solution. However, after this point, we employ a different approach. The algorithms try to successively remove edges from the graph, after checking all the patterns that mismatch in at most d positions from both the l-grams that are connected by the edge. We categorize the edges into two groups. Group1 consists of edges that connect l-grams that differ in more than d positions, and Group2 consists of edges that connect l-grams that differ in less than or equal to d positions. In the following sections, we will show that there will be relatively fewer patterns that mismatch in at most d positions from both the l-grams that are connected by a Group1-edge. Precisely, we will show that there will be at most $O(l^{\frac{d}{2}}\sigma^{\frac{d}{2}})$ such patterns for every Group-1 edge. Each Group-2 edge, on the other hand, can have $O(l^d\sigma^d)$ such patterns. We present a technique which enumerates all the patterns corresponding to each Group1-edge, checks each one of them to see if they satisfy the search criteria, and removes the Group1-edge. We show that at least k monad patterns can be reported without enumerating the patterns corresponding to the Group-2 edges, there by avoiding the $O(l^d\sigma^d)$ complexity. Unlike WINNOWER and cWINNOWER[Lia03], our algorithm is guaranteed to find a solution in $O(n^2t^2l^{\frac{d}{2}}\sigma^{\frac{d}{2}})$ time using $O(ntl^{\frac{d}{2}}\sigma^{\frac{d}{2}})$ space.

3.2 Problem Statement

In the discussion that follows, for convenience in illustration, we treat the input sample as a single sequence of size n. The time and space complexities are not affected by this simplification. In section3.5, we explain the enhancements to handle t different sequences, instead of a single sequence. Therefore, the problem can be stated as follows: given a string S of length n over the alphabet $\sum = \{A, C, G, T\}$, the problem is to find a pattern P of length l that occurs at least k times in S with at most d mismatches in each occurrence.

3.3 Terms and Definitions

We denote a length-l substring(an l-gram) of S starting at position i in S by L_i. A score $h = D(L_i, L_j)$ indicates the number of positions in which the two l-grams L_i, L_j mismatch. We denote the set of patterns that mismatch with both L_i and L_j in at most d positions by $\rho(L_i, L_j)$. We refer to the set $\rho(L_i, L_j)$ also as the set of patterns that are consistent with L_i and L_j. We now describe, briefly, how to compute the size of the set $\rho(L_i, L_j)$. Let P be any pattern such that $P \in \rho(L_i, L_j)$. Now, it is important to note that $\rho(L_i, L_j) = \{\Phi\}$ if $h > 2d$. We have to enumerate all the different possibilities for P. Also, let us divide each l-gram into two regions: M-region, consisting of positions in which L_i and

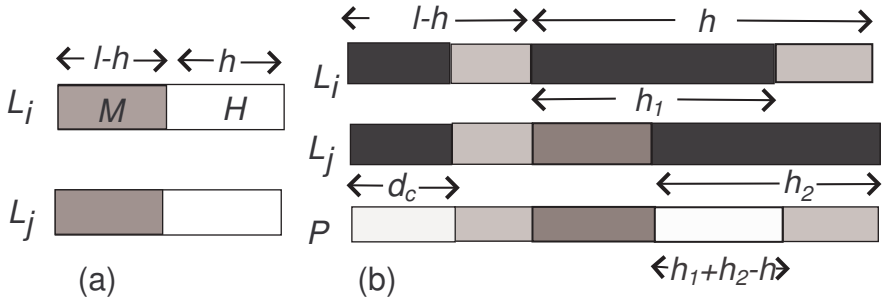

Fig. 3. (a) The matching(M) and mismatching(H) regions of L_i, L_j. (b) Different regions of the pattern P. The regions in black are the regions in which L_i, L_j mismatch with P.

L_j match with each other, and the H-region, consisting positions in which L_i and L_j mismatch with each other, as shown in Figure 3-(a). Both the regions are shown to be contiguous for simplicity in illustration. In reality, these regions need not be contiguous. Now, let us assume patterns L_i and L_j mismatch with P in d_c positions within the M-region. Additionally, let L_i mismatch with P in h_1 positions, and let L_j mismatch with P in h_2 positions, as shown in Figure 3-(b). Again, none of these regions needs to be contiguous.

Now, d_c mismatch positions can be chosen from $l-h$ positions in $\binom{l-h}{d_c}$ ways. At each one of these positions, we have $\sigma - 1 = 3$ symbols to choose from. Similarly, h_1 positions in which L_i can mismatch with P can be selected from h positions in $\binom{h}{h_1}$ ways. The remaining $h-h_1$ positions in L_i have to match with P, and hence they mismatch with P in L_j (since we know that L_i mismatches with L_j in these positions). The remaining $(h_1 + h_2 - h)$ positions in which L_j mismatches with P can be selected from h_1 positions in $\binom{h_1}{h_1+h_2-h}$ ways. We have $\sigma - 2 = 2$ options at each one of these $(h_1 + h_2 - h)$ positions, since P mismatches with both L_i and L_j. Therefore, the total number of patterns in (L_i, L_j) is given by:

$$|\rho(L_i, L_j)| = \sum_{d_c=0}^{\lfloor \frac{2d-h}{2} \rfloor} \left[\binom{l-h}{d_c} 3^{d_c} \sum_{h_1=h-d+d_c}^{d-d_c} \sum_{h_2=h-h_1}^{d-d_c} \binom{h}{h_1} \binom{h_1}{h_1+h_2-h} 2^{h_1+h_2-h} \right], \text{ if } h \leq 2d$$
$$= 0 \text{ otherwise} \quad (1)$$

In the above expression, $|\rho(L_i, L_j)|$ increases when h decreases. When $d < h \leq 2d$, the maximum value of $|\rho(L_i, L_j)|$ occurs when $h = d + 1$. When $h = d + 1$, the maximum value that d_c can take is given by $d_c = \frac{d-1}{2}$ which is equal to $\frac{d}{2}$ when d is odd, and $\frac{d}{2} - 1$ when d is even. Now, $\binom{l-h}{d_c} 3^{d_c}$is in $O(l^{d_c} 3^{d_c})$. Therefore, on the whole, $|\rho(L_i, L_j)|$ is in $O(l^{\frac{d}{2}} 4^{\frac{d}{2}})$.

3.4 The PRUNER-I and PRUNER-II Algorithms

In both the algorithms, we construct a graph $G(L, E)$ where each vertex is an l-gram in the input sample, and there is an edge $(L_i, L_j, D(L_i, L_j))$ connecting two l-grams L_i and L_j if $D(L_i, L_j)$ is less than or equal to $2d$. We then successively remove vertices representing l-grams from the graph $G(L, E)$ that have a degree less than $k-1$, and remove the edges that are incident on these vertices. Until this point, our algorithms are no different from WINNOWER. However, they differ from WINNOWER in the following steps.

Both the PRUNER-I and the PRUNER-II algorithms process each vertex successively. The PRUNER-I algorithm enumerates the consistent patterns for every group1-edge (i.e., edges between l-grams which mismatch in more than d positions). It then computes how many times each pattern repeats. It does this by adding all the consistent patterns for each edge to a list, sorting and scanning the list. Each time a pattern appears, it means that the pattern is within d mismatches from another l-gram. Hence, if a pattern repeats $k-1$ times, it means that the pattern is within d mismatches from $k-1$ other l-grams. However, since we have not yet processed the Group2-edges(i.e., edges connecting l-grams that mismatch in d or fewer positions), we can not yet discard the patterns that repeat less than $k-1$ times. We do not want to evaluate all the consistent patterns for the Group2-edges, as there are too many ($O(l^d 4^d)$) such patterns. Therefore, we will have to take each pattern in the list, and compare it with each l-gram that is connected to the current vertex through a Group2-edge. Only then will we know how many times each one of those patterns has repeated. An efficient way of doing all this is presented below.

At each node L_i, we enumerate the consistent patterns $\rho(L_i, L_j)$ for all the Group1-edges, i.e., edges $(L_i, L_j, D(L_i, L_j))$, such that $d < D(L_i, L_j) \leq 2d$. We add these patterns to a list $\eta(i)$, and remove the edge $(L_i, L_j, D(L_i, L_j))$. Lemma 1 states that we can safely remove the edge $(L_i, L_j, D(L_i, L_j))$ after enumerating and adding $\rho(L_i, L_j)$ to $\eta(i)$.

Lemma 1. *After a vertex L_i in $(L_i, L_j, D(L_i, L_j))$ is processed, there can be no new patterns in $\rho(L_i, L_j)$ that were not reported while processing L_i, but will be reported while processing the vertex L_j.*

Proof. Let us assume that there is a pattern $P \in \rho(L_i, L_j)$ that was not reported while processing node L_i, but will be reported while processing node L_j. This means that there are a set of l-grams $\psi(P)$ other than L_i, such that for each $L_q \in \psi(P)$, there is an edge $(L_q, L_j, D(L_q, L_j))$ connecting L_q and L_j, and $D(L_q, P) \leq d$. Additionally, since P will be reported while processing L_j, $|\psi(P)| \geq k-2$. Now, since for each $L_q \in \psi(P)$, $D(L_q, P) \leq d$ and $D(L_i, P) \leq d$ (as $P \in \rho(L_i, L_j)$ by definition), it implies that $D(L_i, L_q) \leq 2d$. Therefore, for each $L_q \in \psi(P)$ there is an edge $(L_i, L_q, D(L_i, L_q))$ connecting L_i and L_q. Since $|\psi(P)| \geq k-2$, and $P \in \rho(L_i, L_j)$, there are at least $k-1$ edges incident in L_i which contain P as one of their consistent patterns. Therefore, pattern P must have been reported while processing node L_i. Hence there can be no pattern $P \in \rho(L_i, L_j)$ that is not reported while processing L_i that can be reported while processing L_j. □

Now, we need to find out how many times each pattern is repeated in $\eta(i)$. An easy way of doing this will be to sort $\eta(i)$, and scan $\eta(i)$. As each pattern in $\eta(i)$ is a length-l string of a fixed alphabet, $\eta(i)$ can be sorted in linear time using radix sort. Let a pattern P repeat m times in $\eta(i)$. Let R be the degree of node L_i after processing and removing all Group1-edges. As explained in section1, R is expected to be very small. We do the following:

- If $m + R < k - 1$, we discard P. The number of times P repeats can increase by at most R, by comparing P with each one of the Group2-edges. If $m < k - 1 - R$, there is no way that P can repeat $k - 1$ times. So we can discard P.
- If $m \geq k - 1$, report P, since it is clear that P has already occurred at least $k - 1$ times.
- If $k - 1 \leq m + R < k - 1 - R$, we compare P with all l-grams that are still connected to L_i. For each such l-gram that mismatches P in at most d locations, we increment the repeat count of P. If the repeat count reaches $k - 1$, we report P. Other wise, we discard P.

Before we leave L_i and proceed to process the next vertex, we can do one more thing - we can remove the vertex L_i from the graph if $R < k - 1$, without ever enumerating the consistent patterns for these edges. Lemma 2 proves this.

Lemma 2. *If the residual degree R of vertex L_i is less than $k-1$ after processing and removing all Group1-edges of L_i, there can be no new patterns that will be reported by processing the Group2-edges.*

Proof. Let us assume that there is a pattern P that was not reported while processing the Group1-edges, but will be reported while processing the Group2-edges. Since we will be reporting P, and since $R < k - 1$, there should have been at least one Group1-edge $(L_i, L_q, D(L_i, L_q)$ such that $P \in \rho(L_i, L_q)$. Therefore, P was checked and reported while processing vertex L_q. Hence there can be no new patterns that will be reported by processing the Group2-edges. □

We are now left with a graph in which the score of each edge is at most d, and degree of each remaining vertex is at least $k - 1$. Therefore, *if the graph has any vertices left, there will be at least* k *vertices left in each connected component of the graph.* In practice, we do not expect any vertices to remain at this stage, as our assumption is that there are not too many patterns that meet the search criteria. All the l-grams that do remain until this stage are themselves valid solutions, since they mismatch in at most d positions with at least $k - 1$ other l-grams. Hence, we report all the remaining l-grams. Beyond this, there might be other patterns in the graph that meet the search criteria, but in a general case, we assume that there are fewer than k distinct monad patterns in the given sample. In the almost impractical scenario that there are more than k distinct monad patterns, the algorithms we present report at least k of them. The PRUNER-I algorithm is presented in detail in figures 4 and 5.

The PRUNER-II algorithm is very similar to the PRUNER-I algorithm in concept. However, the PRUNER-II algorithm attempts to eliminate the potentially huge memory requirements of the PRUNER-I algorithm. While processing

```
ProcessLGram()
Inputs: G(L,E), i, l, d, k
Output: Reports patterns in the d-mismatch neighborhood of L_i that satisfy the search criteria
 1. PatternList ← {φ}
 2.for every j such that (L_i, L_j, D(L_i, L_j)) ∈ E do
 3.          if D(L_i, L_j) > d /* checking if (L_i, L_j, D(L_i, L_j)) is a Group-1 edge*/
 4.             PatternList ← PatternList ∪ ρ(L_i, L_j)
/* the set ρ(L_i, L_j) of consistent patterns is enumerated by a subroutine at this point*/
 5.          E ← E − (L_i, L_j, D(L_i, L_j)) /* The Group-1 edge is immediately removed */
 6.          end if
 7. end for
 8. RadixSort(PatternList)
 9. Cnt ← 0 /* Cnt is the number of times the current pattern has repeated */
10. for j ← 1 to |PatternList| − 1
11.     if PatternList_j = PatternList_{j−1}
12.        Cnt ← Cnt + 1
13.     else if Cnt ≥ k − 1 − Degree(L_i) /*Degree(L_i) is the residual degree (the degree of
          Group-2 edges of L_i) , since all the Group-1 edges have been removed in step 6.*/
14.          for every r such that (L_i, L_r, D(L_i, L_r)) ∈ E do /* for each Group-2 edge */
15.                  if D(PatternList_j, L_r) ≤ d
        /*check if the pattern PatternList_j is in the d-mismatch neighborhood of L_r*/
16.                     Cnt ← Cnt + 1
17.                  end if
18.          end for
19.          if Cnt > k − 1
20.             Report(PatternList_j) /*PatternList_j is an (l, d) − k pattern */
21.          end if
22.          Cnt ← 0
23.     end if
24. end for
```

Fig. 4. The routine that checks the d-mismatch neighborhood of each l-gram

each node L_i, the PRUNER-I algorithm maintains a list $\eta(i)$ that contains all the patterns that are consistent with each one of the Group1-edges. When the number of such edges is huge, the amount of memory required for $\eta(i)$ may be too big. Especially, this might be the case when d is large and the d/l ratio is large, in which case the graph $G(L, E)$ will be highly connected.

At each vertex L_i, the PRUNER-II algorithm processes edges one by one. For each edge $(L_i, L_j, D(L_i, L_j))$, it enumerates the set of consistent patterns $\rho(L_i, L_j)$. For each consistent pattern $P \in \rho(Li, Lj)$, if we compare P with all the l-grams that are directly connected with vertex L_i, we can determine if P mismatches in at most d positions with at least $k-1$ of them. However, a deeper analysis reveals that it not necessary to compare P with all the l-grams that share an edge with L_i. For any l-gram L_q, if $D(L_q, P) \leq d$, then $D(L_q, L_j)$ will be less than or equal to $2d$. This means that the l-grams L_q and L_j will also be connected. Therefore, we only need to compare P with all vertices L_q such that the edge $(L_q, L_j, D(L_q, L_j)) \in E$. If at least $k-2$ of them mismatch with P in fewer than d positions, it reports P. Otherwise, P is discarded. As in the PRUNER-I algorithm, it removes the edge $(L_i, L_j, D(L_i, L_j))$ after checking all the patterns in $\rho(L_i, L_j)$.

3.5 Extending the Algorithm to Handle Multiple Sequences

When the input sample is made of t sequences of length n each, and the problem is to find an (l, d) motif that occurs in at least k of them, the graph $G(L, E)$ will be a t-partite graph. At each vertex in the graph, we need to maintain and update another variable, which we call t-degree. The variable t-degree stores the number of distinct sequences in t that the current vertex is connected to. In the

```
SearchForPatterns()
Inputs: S, l, d, k, n
 1. Buildgraph(S,l,d,k,n) /*The routine that builds the graph G(L,E)*/
 2. PruneGraph(G(L,E),l,d,k,n) /* The pruning routine which removes
                                  all the vertices with degree < k − 1*/
 3. for i ← 0 to n − l + 1 do
 4.     ProcessLGram(G(L,E),i,l,d,k,n)
 5.     ifDegree(L_i) < k − 1
 6.       RemoveLGram(G(L, E), i) /* remove the vertex L_i
        (and the edges incident on L_i) from the graph */
 7. end for
 8. PruneGraph(G(L,E), l, d, k, n) // remove l-grams with degree < k − 1
 9. for i ← 0 to n − l + 1 do /*check if any l-grams are still remaining*/
10.     if Degree(L_i) > k − 1
11.       Report(L_i) /* report all remaining l-grams*/
12.     end if
13. end for
```

Fig. 5. The PRUNER-I algorithm

algorithms that we discussed above, whenever we are referring to the degree of a vertex, we will be using t-degree instead of the actual degree of the vertex. Whenever we are checking for a pattern P, it is no longer sufficient to check if the pattern is within d mismatches from $k-1$ other l-grams. We need to make sure that the l-grams are derived from $k-1$ distinct sequences in the sample. The implementation typically involves maintaining a bit-vector of length t for the pattern that is being considered. Whenever the pattern is within d-mismatches from an l-gram, the bit corresponding to the sequence from which the l-gram is derived is set to 1. P satisfies the search criteria if at least $k-1$ (or whatever is necessary at that point in the algorithm) bits are set to 1.

3.6 Complexity analysis

Building the graph involves calculating the mismatch count for each l-gram pair (L_i, L_j) such that L_i and L_j are derived from different input sequences. There are $(n-l+1)$ l-grams for each input sequence, and $n(t-1)$ other l-grams for each l-gram in the input sequence. Therefore, building the graph takes $O(n^2t^2)$. Pruning the graph involves removing all the edges incident on each vertex whose degree is less than $k-1$. In the worst case, we might have to delete all the nodes, so the maximum number of edges that need to be removed is $((k-1)nt-1)$, which is $O(ntk)$. This time is common for both PRUNER-I and PRUNER-II. In the PRUNER-I algorithm, each l-gram can have up to $n(t-1)$ $2d$-mismatch neighbors. Therefore, at each l-gram, we might have to enumerate the consistent patterns with $n(t-1)$ other l-grams. The maximum number of these consistent patterns as discussed in section 3.3, is $O(l^{\frac{d}{2}}4^{\frac{d}{2}})$. Hence the worst-case time complexity at each node is given by $O(ntl^{\frac{d}{2}}4^{\frac{d}{2}})$. We need to store all these patterns in a list, so we need $O(ntl^{\frac{d}{2}}4^{\frac{d}{2}})$ space. In the worst case, we will have to process $(t-k+1)n$ l-grams, since no new patterns can be discovered after removing all the vertices corresponding to $(t-k+1)$ sequences in the sample. Therefore, the overall complexity is given by $O(n^2t(t-k+1)l^{\frac{d}{2}}4^{\frac{d}{2}})$. If k is small w.r.t. t, this will be $O(n^2t^2l^{\frac{d}{2}}4^{\frac{d}{2}})$. When $k=t$, the complexity of the PRUNER-I algorithm is $O(n^2tl^{\frac{d}{2}}4^{\frac{d}{2}})$. In case of the PRUNER-II algorithm, each edge is processed separately. All the patterns consistent with each edge (L_i, L_j) have to be compared with all the l-grams that are connected to both L_i and L_j.

Table 1. Performance of the algorithms

Test case (l,d)	d/l ratio	PRUNER-I Time	PRUNER-I Memory (MB)	PRUNER-II Time	PRUNER-II Memory (MB)	Test case (l,d)	d/l ratio	PRUNER-I Time	PRUNER-I Memory (MB)	PRUNER-II Time	PRUNER-II Memory (MB)
13,3	0.23	0.17	43	0.14	43	13,4	0.31	12.26	166	29.58	278
14,4	0.28	5.35	198	7.09	178	15,4	0.27	2.28	122	1.34	91
16,4	0.25	0.56	51	0.56	43	16,5	0.31	69.00	540	284.59	247
17,5	0.29	29.54	315	36.58	161	18,5	0.28	13.19	174	13.19	92
19,5	0.263	6.02	101	0.54	48	20,5	0.25	2.16	65	0.13	21
21,5	0.238	0.13	10	0.13	11	22,5	0.227	0.17	56	0.13	7
22,6	0.273		649	3.16	83	23,6	0.261	18.57	525	0.13	64
24,6	0.25	1.12	720	0.16	11	25,6	0.24	1.17	592	0.18	60
26,6	0.231	1.08	613	0.16	62	27,7	0.259	out of memory		2.27	614
28,7	0.25	out of memory		0.36	640	29,7	0.241	0.38	749	1.41	640

In the worst case, there can be $n(t-2)$ vertices that are connected to both L_i and L_j. The total number of the edges could be $n^2t(t-1)$ in the worst case. The edge can have $O(l^{\frac{d}{2}}4^{\frac{d}{2}})$ patterns that are consistent with it, so the total time taken will be $O(n^3t^3l^{\frac{d}{2}}4^{\frac{d}{2}})$. Each pattern could be compared separately; therefore the space needed is approximately the same as that necessary for the graph.

4 Results

The algorithms were tested on generated samples containing 20 sequences of 600 nucleotides each. The sequences are implanted with randomly mutated patterns at randomly chosen positions. Each occurrence of the pattern is allowed to have up to d mismatches. The tests were carried out on a Pentium-4 3.2 GHz PC with 2GB of memory, running Redhat Linux 9.0. The time/memory results are presented in Table 1. The PRUNER-I algorithm ran out of memory for the (27,7) and the (28,7) cases. The implanted pattern was detected in all the remaining test cases.

5 Conclusion

We have presented two new algorithms for finding the monad patterns. Both the algorithms perform extremely well on the challenge problem of (15,4) on 20 input sequences of 600 nucleotides. As d increases in comparison to l, i.e., when the d/l ratio increases, the PRUNER-I algorithm takes a longer time and a larger memory. The PRUNER-I algorithm runs out of memory for large values of l and d. The PRUNER-II algorithm, on the other hand, can handle large values of l and d, but reacts very sharply to the d/l ratio. As long as the d/l ratio is around 0.25, the PRUNER-II algorithm performs very well, independent of the actual values of l and d. Unlike Winnower and cWinnower, the algorithms we presented here are not sensitive to k. Our algorithms will be able to detect patterns even for very small values of k. The only concern when dealing with very small values of k is that there could be random signals in the input sample that meet the search criteria. An interesting observation from the test cases is that the graph

itself starts consuming more and more space as the d/l ratio gets bigger. This is because there are more and more edges in the graph, as there are a larger number of l-gram pairs that mismatch in less than $2d$ positions. In the future, we plan to investigate compact representations for the graph. Another approach may involve using a two-pass algorithm. WINNOWER or cWINNOWER can be used initially in order to remove some spurious edges. Our algorithms can be applied in the second pass. As the graph has much fewer edges now, PRUNER-I or PRUNER-II may have very good performance. For the first pass, we can use a low sensitivity version of WINNOWER or cWINNOWER in order to maximize the speed.

References

[BuhT2001]Buhler J. and Tompa B.: Finding motifs using random projections *Proc. of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB01)* (2001) 69–76

[EskKGP03]Eskin E., Keich U., Gelfand M.S., Pevzner P.A.: Genome-wide analysis of bacterial promoter regions. *Proc. of the Pacific Symposium on Biocomputing PSB − 2003* (2003) Kau'i, Hawaii, January 3-7, 2003

[EskP02]Eskin E. and Pevzner P.A.: Finding composite regulatory patterns in DNA sequences. *Proc. of the Tenth International Conference on Intelligent Systems for Molecular Biology (ISMB-2002)* (2002) Edmonton, Canada, August 3-7

[GuhS01]Guha Thakurtha D. and Stormo G.D.: Identifying target sites for cooperatively binding factors. *Bioinformatics* **15** (2001) 563–577

[HerS99]Hertz G.Z. and Stormo G.D.: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **10**, (1999) 1205–1214

[Lia03]Liang S.: cWINNOWER Algorithm for finding fuzzy DNA motifs. *Proc. of the 2003 IEEE Computational Systems Bioinformatics conference (CSB 2003)* (2003) 260–265

[MarS00]Marsan L. and Sagot M.: Algorithms for extracting structured motifs using suffix tree with applications to promoter and regulatory site consensus identification. *Journal of Computational Biology* **7**, (2000) 345–360

[PavMP01]Pavesi G., Mauri G. and Pesole G.: An algorithm for finding signals of unknown length in DNA sequences. *Proc. of the Ninth International Conference on Intelligent Systems for Molecular Biology* (2001)

[PevS00]Pevzner P.A. and Sze S.: Combinatorial approaches to finding subtle motifs in DNA sequences. *Proc. of the Eighth International Conference on Intelligent Systems for Molecular Biology* (2000) 269–278

[PriRP03]Price A., Ramabhadran S. and Pevzner A.: Finding subtle motifs by branching from sample strings. *Bioinformatics* **19**, (2003) 149–155

[Sag98]Sagot M.: Spelling approximate or repeated motifs using a suffix tree. *Lecture notes in computer science* **1380**, (1998) 111-127

[vanRC00]van Helden J., Rios A.F., and Collado-Vides J.: Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research* **28**, (2000) 1808–1818