SUBTLE MOTIF DISCOVERY FOR DETECTION OF DNA REGULATORY SITES

MATTEO COMIN*

Dept. Information Engineering, University of Padova, Italy
E-mail: ciompin@dei.unipd.it

LAXMI PARIDA

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA.
E-mail: parida@us.ibm.com

We address the problem of detecting consensus motifs, that occur with subtle variations, across multiple sequences. These are usually functional domains in DNA sequences such as transcriptional binding factors or other regulatory sites. The problem in its generality has been considered difficult and various benchmark data serve as the litmus test for different computational methods. We present a method centered around unsupervised combinatorial pattern discovery. The parameters are chosen using a careful statistical analysis of consensus motifs. This method works well on the benchmark data and is general enough to be extended to a scenario where the variation in the consensus motif includes indels (along with mutations). We also present some results on detection of transcription binding factors in human DNA sequences.
Availability: The system will be made available at www.research.ibm.com/computationalgenomics.
Keywords: pattern discovery, subtle motifs, consensus motifs, transcription factors, binding sites.

1. Introduction

The problem of detecting common motifs across DNA sequences for locating regulatory sites, transcription binding factors or even drug target binding sites is of prime importance. The main difficulty is that these motifs have subtle variations at each occurrence. This problem has been of interest to both biologists and computer scientists. A satisfactory practical solution has been elusive although the problem is defined very precisely:

Problem 1. ***(The Consensus Motif Problem):*** Given t sequence s_i on an alphabet Σ, a length $l > 0$ and a distance $d \geq 0$, the task is to find all patterns p, of length l that occur in each s_i such that each occurrence p_i' on s_i has at most d mismatches with p.

The problem in this form made its first appearance in 1984 [16]. In this discussion, the alphabet Σ is $\{A, C, G, T\}$ and the problem is made difficult by the fact that each occurrence of the pattern p may differ in some d positions and the occurrence of the consensus pattern p may not have $d = 0$ in any of the sequences. In the seminal paper [16], Waterman and coauthors provide exact solutions to this problem by enumerating *neighborhood* patterns, i.e., patterns that are at most d Hamming distance from a candidate pattern. Sagot gives a good summary of the (computational) efforts in [14] and offers a solution that improves the time complexity of the earlier algorithms by the use of generalized

*Work done during internship at IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA.

suffix trees. These clever enumeration schemes, though exact, have a drawback that they run in time exponential in the pattern length.

This problem of detecting common subtle patterns across sequences is nevertheless of great interest and various statistical and machine learning approaches, which are inexact but more efficient, have been proposed [9,10,3,6]. One of the questions that can be asked to compare and test the efficacy of such methods of consensus motif detection systems is: *Given a set of sequences that harbor (with mutations) k motifs , what percentage of the k motifs does the system recover?* When k is large, all of the above approaches give good average-case performance under this criterion.

Yet another question to ask is : *Given a set of sequences that harbor (with mutations) ONE motif p , does the system recover p?* This is a rather difficult criterion to meet since these algorithms use some form of local search based on Gibbs sampling or expectation maximization or even clever heuristics. Hence it is not surprising that they may miss p. However, a question of this form is a biological reality. Consider the following, somewhat contrived, variation of Problem 1 which is an attempt at simplifying the computational problem.

Problem 2. ***(The Planted** (l, d)**-motif problem):*** Given t sequence s_i' on Σ, a pattern p of length l is embedded in s_i', with exactly d errors (mutations), to obtain the sequence s_i of length n, for each $1 \leq i \leq t$. The task is to recover p, given s_i, $1 \leq i \leq t$ and the two numbers l and d.

Pevzner and Sze tantalized the community with the *challenge problem*, which was Problem 2 with parameters $n = 600$, $t = 20$, $l = 15$ and $d = 4$ [11]. A thrust of this paper also was the need for the deployment of combinatorial approaches to tackle this thorny problem. One of the algorithms they presented was an exact algorithm where the challenge problem was reduced to finding a t-sized clique in a t-partite graph with at most $n - l + 1$ vertices in each partition. Even the best known heuristics for clique finding problem failed to detect the clique corresponding to the signal. The second algorithm was based on enumerating possible patterns and checking their candidacy for being the subtle pattern using clever heuristics and an exhaustive search in a reduced space. A similar algorithm, with different heuristics was presented in [12,7].

One of the most effective algorithms, we found, was the one discussed by Buhler and Tompa [4]. The probabilistic algorithm uses a random projection h and hashes each input l-mer x into bucket $h(x)$. Any hash bucket with sufficiently many entries is explored as a potential embedded motif. This approach solved the challenge problem and some more. There has been a flurry of activity around this problem of subtle motifs [5,7,8]. See also [13] for some practical implementations of exact approaches.

Overview of our approach. We first clarify the different "motifs" used in this paper: Our central goal is to detect the *consensus* or the *embedded* or the *planted* motif in the given data sets which is also sometimes referred to as the *signal* in the data or the *subtle signal*. When a motif is not qualified with these terms, it refers to a substring that appears in multiple sequences, with possible wild cards.

We propose an approach that uses unsupervised motif discovery to solve Problem 2. We show that this method works well for the more general Problem 1 as well. Recall that the signal ("subtle motifs") is embedded in t random sequences. The problem is compounded by the fact that although the consensus motif is solid (i.e., an l-mer without wild cards or dont-care characters), it is not necessarily contained in any of the t sequences. However, if we can obtain a correct alignment of the m sequences, then it is relatively easy to extract the consensus motif satisfying the (l, d) constraint. In other words, one of the difficulties of the problem is that the sequences are unaligned. The extent of similarity across the sequences is so little that any global alignment scheme cannot be employed. So we tackle this problem in two steps: First, we identify *potential signal (PS)* segments of interest in the input sequences. This is done by using the imprints of the discovered motifs on the input. Second, amongst these segments, we carry out an exhaustive comparison and alignment to extract the consensus motif. This delineation into two steps helps us also address a more realistic version of the problem that includes insertion and deletion in the consensus motif:

Problem 3. ***(The Indel Consensus Motif Problem):*** Given t sequence s_i on an alphabet Σ, a length $l > 0$ and a distance $d \geq 0$, the task is to find all patterns p, of length l that occur in each s_i such that each occurrence p_i' on s_i is at an edit distance (mutation, insertion, deletion) at most d from p.

The main focus of our method is in obtaining good quality PS segments and restricting the number of such segments to keep the problem tractable. The Type I error or false negative errors, in detecting PS segments, are reduced by using appropriate parameters for the discovery process based on a careful statistical analysis of consensus motifs which is discussed in Section 2. The Type II error or false positive errors are reduced by using irredundant motifs [2] and their statistical significance measures [1] discussed in Section 3.1. Loosely speaking, irredundancy helps to control the extent of over-counting of patterns and the pattern-statistics helps filter the true signal from the signal-like-background. In the scenario where indels (insertions and/or deletes) are permitted along with mutations, the unsupervised discovery process detects *extensible* motifs (instead of *rigid* motifs that have a fixed imprint length in all the occurrences). Also, the second step uses *gapped* alignments.

2. Statistics of consensus motifs

Here we make some calculations, under simplifying assumptions, to justify our unsupervised motif discovery approach to the problem. We consider the most general version of the problem which is formally stated as Problem 3 in the last section. Recall that this setting permits insertion and deletion as well as mutation in the embedded motif.

Given t sequences of length l each, a pattern satisfies *quorum* K if it occurs in $K' \geq K$ of the given t sequences. Further it is of *maximal* size h, if in each of the K' occurrences, the size cannot be increased without decreasing the number of occurrences K' (see [1] for a more rigorous definition).

For simplicity, the sequences are the same length l as the consensus motif and all the t sequences are aligned and we will further assume that a pattern occurs at most once in each sequence. Let q be the probability of any position in the input data to be contained in a pattern and let $P_{maximal}(K, H, q)$ be the probability that a pattern with maximal H solid characters and quorum K occurs in the input data. Then [a]

$$P_{maximal}(K, H, q) = \sum_{k=K}^{t} \binom{t}{k} \left(1 - q^H\right)^{t-k} q^{H\,k} (1 - q^k)^{l-H} \tag{1}$$

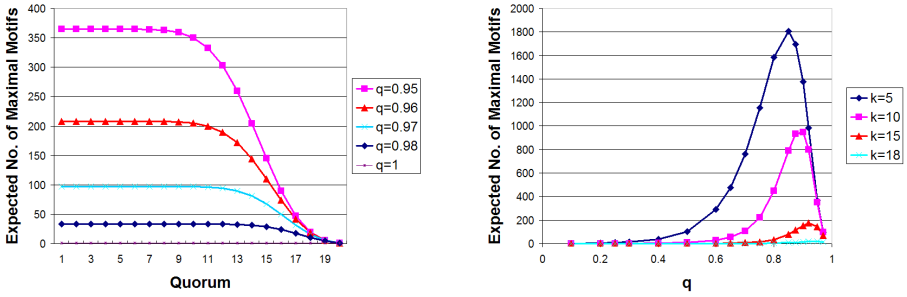
Let $Z_{K,q}$ be a random variable denoting the number of maximal motifs with quorum K and q as defined above, and, $E(Z_{K,q})$ denotes the expectation of $Z_{K,q}$. Note that for maximal motifs, it is the case that the occurrences of two distinct motifs are independent events. Further, using linearity of expectations, we obtain (for a fixed t and l),

$$E(Z_{K,q}) = \sum_{h=1}^{l} \binom{l}{h} P_{maximal}(K, h, q) \tag{2}$$

Computing q. Consider the case where the embedded motif is constructed with some d edit operations. Let the edit operations be (1) mutation, (2) deletion and (3) insertion. Let q_M be the probability of mutation, q_X the probability of deletion and q_I the probability of insertion with $q_M + q_X + q_I = 1$.

Note that for simplicity we have assumed that the t sequences are aligned. For example, the table on the left below shows exactly one edit applied to the signal motif and the table on the right shows

[a] Note that if the motifs are not maximal then $P(K, H, q) = \sum_{k=K}^{t} \binom{t}{k} \left(1 - q^H\right)^{t-k} q^{H\,k}$. Also if motif m_1 is a maximal version of motif m_2 then the occurrences of m_1 and m_2 are not independent.



(a) (b)

Figure 1. For $t = 20$, $l = 20$, the expected number of maximal motifs $E(Z_{K,q})$, is plotted against (a) quorum K shown along the X-axis (for different values of q), and, (b) against q shown along the X-axis (for different values of quorum K).

the alignment of the embedded motifs.

Edits	signal = ACGTAC						
M	A	C	G	T	C	C	
X	A	G	T	A	C		
I	A	C	G	A	T	A	C
M	A	C	C	T	A	C	
M	G	C	G	T	A	C	

Alignment						
A	C	G	-	T	c	C
A	-	G	-	T	A	C
A	C	G	a	T	A	C
A	C	c	-	T	A	C
g	C	G	-	T	A	C

Assume that d out of the l positions are picked at random on the embedded motif for exactly one of the edit operations, insertion, deletion or mutation. Then it is easy to see that probability q of a position to be contained in a motif is:

$$q = 1 - \frac{d}{l}(q_M + q_X) \tag{3}$$

Now, it is straightforward to compute the value of q, to estimate $E(Z_{K,q})$ of Equation (2), given different scenarios. For example, consider the following cases.

(1) Exactly d mutations $q_M = 1$ $q = 1 - d/l$

(2) Exactly d edits $q_M = q_X = q_I = 1/3$ $q = 1 - 2d/3l$

Also note that when *no more than* d' *edit operations* are carried out on the embedded motif, it is usually interpreted as each collection of $0, 1, 2, \ldots, d'$ positions being picked with equal probability, and thus $d = d'/2$ for Equation (3).

2.1. *Rationale for using unsupervised motif discovery*

A motif of length l that occurs across $t' \leq t$ sequences provides a local alignment of length l for the t' sequences which, in a sense, justifies the simplified scenario of the last section. The best case scenario, for our problem, is when the embedded motif m is identical in all t sequences and the discovery process detects this *single* maximal motif with quorum t. So the scenarios closer to the best case should have fewer (but important) maximal motifs. Figure 1(a) shows the expected number of motifs with different values of q and quorum K. Notice that the expected number of motifs saturates for small values of K and falls dramatically as K increases. The saturation at lower values occurs since we are seeking *maximal* motifs. Thus as q increases the saturation occurs at a higher value of K. Figure 1(b) shows the variation of the expected number of maximal motifs with q which is unimodal, for different values of K. The value of q is determined by the given problem scenario and thus a large value of K is a good handle on controlling the number and "quality" of maximal motifs.

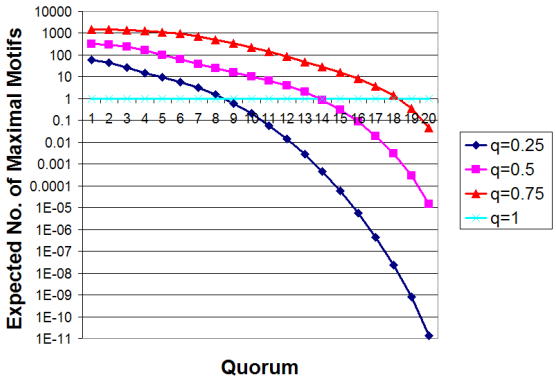


Figure 2. For $t = 20, l = 20$, the expected number of maximal motifs $E(Z_{K,q})$, is plotted against quorum K shown along the X-axis, for different values of q, in a logarithmic scale. Notice that when $q = 1$, the curve is a horizontal line at $y = 1$. Note that for DNA sequences, $q = 0.25$ corresponds to the random input case.

The *signal* is embedded in the *background* and it is important to exploit the characteristics that distinguishes one from the other. In our case, we assume that the background is random, in other words it is assumed to be randomly generated using an i.i.d process. Under this condition, it is easy to see that $q = 1/4$. Thus we need to compare $E(Z_{K,q})$ with $E(Z_{K,1/4})$, the expectation for the random case. To compare these expectation curves, particulary around small values (close to 1 in the Y-axis), we study the plots of $\log(E(Z_{K,q}))$ against quorum K in Figure 2.

For example, consider the case when $q = 0.75$; this is the value of q for the *challenge problem* of Section 1. In Figure 2, this is shown by the red curve and for large K, say $K \geq 16$, the expected number of motifs is small. Also, the corresponding expected numbers for the random case is extremely low, thus providing a strong contrast in the number of expected motifs. Hence the reasonable choice for the quorum parameter K is 16 or more, in the unsupervised discovery process.

Before we conclude this section, we must point out that in the case where the embedded motif is changed with insertions and/or deletions (indels), the q value is computed appropriately using Equation (3) and the corresponding expectation curve in Figure 2 is studied. However, the burden is heavier for the unsupervised discovery process and we use the extensible (or, variable-sized gaps) motif discovery capability in Varun [1].

3. Subtle Motif Finder: Our Approach

Here we present our approach, *SubtleMotif*, which detects the consensus motifs in two steps. We first locate regions in the sequence called *potential signal* (PS) segments. The statistical analysis of the previous section suggests that the detection of PS segments via unsupervised motif discovery is indeed possible. The two important parameters in the combinatorial discovery process are K and D: K is the quorum or the minimum number of sequences where the pattern must occur and D is the size of the gap between any two solid characters in the pattern. In the second step we carry out a local alignment of these short segments and extract the consensus motif.

3.1. *(Step 1) Detecting PS segments*

As seen in the last section, we expect to see more maximal patterns in the signal region than the background in an appropriate range of quorum K. We extract all common motifs across sequences using an unsupervised combinatorial motif discovery process. We use the system Varun [1] for this

purpose. This allows us to discover motifs with "dont-cares" or wild-cards. The number of such characters is controlled by the parameter D in Varun, which is a bound on the number of "dont-cares" between any two solid characters in a pattern. Next, we simply count the number of motifs that cover a position i on the input. The first prediction of the PS segments are the positions (i's) with high counts.

This elementary rule works well for simple cases like Problem 2 with $n = 600$, $t = 20$, $l \leq 10$ and $d = 2$. Here the PS segments are predicted accurately. However, for $d > 2$ we found it is difficult to distinguish the true from the false PS segments using this simple approach. To weed out these wrong PS segments, we explored other means of pruning the motifs using some combinatorial and statistical approaches. Firstly, we use the idea of *irredundant* or *basis* motifs [2], to avoid overcounting of patterns that cover the same region multiple times on the sequence. Secondly, we consider only those motifs that have a significant z-score and also, biased the motif count at a position i on the input with the probability of the occurrence of that motif. Due to space constraints, we omit the discussion on irredundant motifs and their statistical significance evaluations and instead direct the reader to [2,1].

We use Varun to discover irredundant motifs in the input data. In the right, the motif discovery parameters are K and D and $l = 15$, $d = 4$, $t = 20$, $n = 600$ and the value of q is $11/15 \approx 0.73$, using Equation (3). Column I shows the number of correct PS segments predicted using *all* motifs and column II shows the same using only *irredundant motifs*. In all the cases, there is an increase in the number of correctly detected positions for the latter.

K	D	I	II
20	2	2	3
19	2	0	1
20	3	3	4
19	3	1	2
20	4	2	5

We compute the z-score of each irredundant motif using our previous result (Equation (5) in [1]) and filter these motifs based on a cut-off threshold z-score. We further use a weighted count for each input position in the imprint of the motif m, where the weight is $(1/p_m)$ and p_m is computed as in Equation (4) in [1]. Figure 3 shows the results for a variety of settings comparing the use of statistical methods (both z-score and weighted counting), called Method II, with the one that does not use them, called Method I.

Notice that using Method II, we can restore all 10 positions of the $n = 200$, $t = 20$, $l = 10$ and $d = 2$ of Problem 2. In the experiments for $l = 15$ and $d = 4$, we can recover 4 positions correctly out of 20. We find that only in two cases Method I recovers more PS segment positions than Method II. However, in all the remaining 22 cases, Method II outperforms Method I.

Since it is very difficult to detect 100% of the PS segments correctly in this step alone, we use these partial PS segments in the next step to reconstruct the true signal.

3.2. *(Step 2) Processing PS Segments*

In the previous step we identified the potential signal (PS) segments in the input. Next, we merge the information form each sequence by combining different PS segments. Assuming that the PS segment is predicted correctly, the planted motif is embedded in this segment. If the length of the consensus motif is known, say l, then the PS segment is constrained to be substring of length $2 \times l$. Thus given a candidate position i in sequence s, the signal is contained in the interval $s[i - l, i + l]$.

We next pick one PS segment from each sequence to "locally align" the segments across some C sequences. We enumerate all the $\binom{t}{C}$ configurations here. Let the C PS segments, each from a distinct sequence, be given as $(s_{i_1}[b_{i_1}, e_{i_1}], s_{i_2}[b_{i_2}, e_{i_2}], ..., s_{i_C}[b_{i_C}, e_{i_C}])$. We make the assumption that the starting position x_{i_j} of the consensus motif in sequence s_{i_j} lies in the substring $s_{i_j}[b_{i_j}, e_{i_j}]$, i.e., $b_{i_j} \leq x_{i_j} \leq e_{i_j}$. We are seeking all possible alignments of length l using these PS segments. We use the following measure to evaluate an alignment. The *majority* string s_m, of length l, is simply the string obtained by using the majority base at each aligned position (column). The score f is the sum total of the aligned positions in all the C segments that agree with s_m. For example, consider the aligned segments here below where $l = 8$, $d = 3$, $C = 5$, and $f = 27$.

$$
\begin{array}{rl}
(1) & \text{—-}A\ C\ T\ G\ C\ T\ C\ C\text{—-} \\
(2) & \text{—-}A\ G\ G\ G\ T\ T\ G\ A\text{—-} \\
(3) & \text{—-}C\ C\ G\ G\ T\ T\ G\ A\text{—-} \\
(4) & \text{—-}C\ C\ T\ C\ T\ A\ C\ A\text{—-} \\
(5) & \text{—-}A\ C\ G\ G\ T\ -\ C\ A\text{—-} \\
s_m = & \mathbf{A\ C\ G\ G\ T\ T\ C\ A}
\end{array}
$$

Since our first step is very tightly controlled, we found in practice that there are only a few candidate PS segments. Also, in the model that uses insertion and deletion (i.e., the length of the imprint of the occurrence of the consensus motif in each sequence is not necessarily l), we use the same score by keeping track of the alignment columns: deletions and insertions result in gaps in some sequences in the alignment (see sequence 5 in the above example). We consider all those alignments, whose score f exceeds a fixed threshold T_C. In all our experiments we have used $C = 3$ and the values of T_C are reported in the experiments.

Extracting the consensus motif across t sequences. At the previous step, we have multiple alignments, where each alignment is across some $C(\leq t)$ sequences. From these we need to extract the consensus motif across all the t sequences. For each alignment, we designate the majority substring s_m (see last section) as the putative consensus motif. Then we scan all the t input strings for the occurrence of s_m with at most d errors which can be done in linear time. For each sequence, we pick the best occurrence, i.e., the one with the minimum edit distance from s_m. In practice, this step very quickly discards the erroneous consensus motifs and quickly converges to the one(s) satisfying the distance constraint of d.

$l = 10, d = 2$

Motif params			Methods	
K	D	M	I	II
10	2	95	8	7
9	2	236	8	10
8	2	434	7	8

(a) $n = 100, t = 10$

Motif params			Methods	
K	D	M	I	II
20	2	281	10	10
19	2	459	12	13
18	2	588	18	18

(b) $n = 200, t = 20$

$t = 20, l = 15, d = 4$

Motif params			Methods	
K	D	M	I	II
20	2	539	2	4
19	2	647	6	7
18	2	837	12	12
20	3	952	5	6
19	3	1164	11	10
18	3	1582	13	13
20	4	1454	8	9
19	4	1832	9	10
18	4	2577	11	11

(c) $n = 300$

Motif params			Methods	
K	D	M	I	II
20	2	1588	2	2
19	2	3526	1	1
18	2	5456	1	1
16	2	7316	1	2
20	3	3348	4	4
19	3	7885	2	2
18	3	12444	1	1
17	3	15318	2	3
16	3	17017	0	1

(d) $n = 400$

Figure 3. Number of PS segment positions predicted correctly using Methods I and II for different parameters. The motif discovery parameters are K and D and M is the total number of irredundant motifs discovered in the input. The values of q, obtained using Equation 3, are as follows: (a) & (b) $q = 0.8$, (c) & (d) $q = 0.73$.

4. Results

Let P be the set of all positions covered by the prediction and S be the same set for the embedded motif. The score of the prediction P, with respect to the embedded motif, can be given as (see [15]): $score = \frac{|P \cap S|}{|P \cup S|}$. The score is 1 if the prediction is 100% correct. However, even for values much smaller than 1, the embedded motif may be computed correctly. This measure is rather stringent and so we use yet another measure, the *solution coverage* (SC) score. This is defined as the number of sequences that contains at least one occurrence of the predicted motif whose distance from the prediction is within the problem constraint i.e., bounded by d. Again if the coverage is equal to the total number of sequences t, then the prediction can be considered 100% correct.

Results on benchmark synthetic data. We report our results in terms of these two measures in Figure 4 averaged over eight random experiments. Each experiment is defined by the four parameters n, t, l and d. In the unsupervised motif discovery process of the first step we use parameters $K = t = 20$ and $0 < D < 4$. The high K value was suggested by the statistical analysis in Section 2 and confirmed by our experiments in Section 3.1. In the second step we use $C = 3$ based on our experiments reported in Figure 3. In Figure 4(a),(b) and (c), we show the performance measures for various instances of Problem 2. We compare our results with what we found as the best performing algorithm, PROJECTION [4]. In all cases our best results are similar, or slightly better, than PROJECTION as shown in Figure 4. We observe that as we increase the number of gaps D, the $score$ improves. In particular if $D = 0$ (i.e., solid motifs), the chances of success drops dramatically. We observe a similar tendency in Problem 3 as shown in Figure 4(d) and (e). Although this version of the problem, with indels, should be harder, we find that the method gives surprisingly good results.

K	D	N	Score	SC
20	1	2	0.066	10
20	2	2	0.415	12
20	3	4	0.95	20
20	4	3	0.94	20

(a) $l = 15, d = 4$
$Score_{PRJ} = 0.93$

K	D	N	Score	SC
20	0	0	0.02	10
20	1	1	0.49	11
20	2	1	0.8	20
20	3	1	0.93	20
20	4	2	0.91	20

(b) $l = 17, d = 5$
$Score_{PRJ} = 0.93$

K	D	N	Score	SC
20	1	2	0.75	11
20	2	2	0.95	20
20	3	4	0.95	20

(c) $l = 19, d = 6$
$Score_{PRJ} = 0.96$

K	D	N	Score	SC
20	0	1	0.05	5
20	1	3	0.75	20
20	2	3	0.81	20

(d) $l = 15$, 3 mutations & 1 indel

K	D	N	Score	SC
20	0	1	0.09	8
20	1	3	0.68	20
20	2	4	0.78	20

(e) $l = 15$, 2 mutations & 2 indels

Figure 4. In all cases, $t = 20$, $n = 600$. The motif discovery parameters are K and D and we use $C = 3$ and the values of T_C are as follows: (a) 32 (b) 36 (c) 40 (d) & (e) 30. The results are are averaged over 8 random problem instances. N is the total number of PS segments predicted correctly. See text for definitions of $Score$ and SC. $Score_{PRJ}$ is the score for the PROJECTION algorithm by Tompa et al.

Results on Human `hm01r` data. We have tested the system on various real data sets and we give details of one such case- that of detecting transcription binding factors on human DNA sequences on the data set suggested by Tompa [15]. The details are as follows:

No	pos	$Predictions$	M	I
0	-101	$T\ G\ A\ C\ G\ T\ C\ A$	$-$	1
1	-299	$T\ G\ C\ -\ G\ T\ C\ A$	1	$-$
2	-71	$T\ G\ A\ C\ A\ T\ C\ A$	1	1
3	-69	$A\ T\ G\ A\ -\ G\ T\ C\ A\ G$	$-$	2
4	-527	$T\ G\ C\ G\ A\ T\ G\ A$	2	1
6	-173	$T\ G\ A\ -\ C\ T\ A\ A$	2	$-$
7	-1595	$T\ G\ A\ -\ A\ T\ G\ A$	2	$-$
8	-221	$T\ G\ G\ -\ G\ T\ C\ T$	2	$-$
9	-69	$T\ G\ A\ -\ C\ T\ G\ C$	3	$-$
10	-105	$T\ G\ A\ -\ A\ T\ C\ A$	1	$-$
12	-780	$T\ G\ C\ -\ G\ T\ C\ A$	1	$-$
14	-1654	$A\ T\ G\ A\ -\ A\ T\ C\ A$	1	1
15	-69	$A\ T\ G\ A\ -\ G\ T\ C\ A\ A$	$-$	2
16	-97	$T\ G\ A\ -\ G\ T\ A\ A$	1	$-$
17	-1936	$A\ T\ G\ A\ -\ A\ T\ C\ A$	1	1
$signal$		$\mathbf{T\ G\ A\ \ G\ T\ C\ A}$		

The parameters for this data set are $n = 2000$, $t = 18$. Note that we had to estimate l and d through a series of trials. l was estimated to be 7 and d to be 3. We use parameters $K = 18$ and $D = 1$ in the motif discovery process in Step 1 and use $C = 3$ and $T_C = 12$ in Step 2. We identify the signal in 15 of the 18 sequences at positions given in the pos column. We miss the signal in only one sequence (sequence no 5) and the signal is absent in two other sequences (no 11 and 13). We reconstruct the consensus sequence as $TGAGTCA$ which is at most 3 edit distance away from the "embedded" signals. In the table M denotes number of mutations and I the number of insertions; no deletions were found.

5. Concluding Remarks

The problem of detecting subtle consensus motifs is tricky and a purely combinatorial or a purely statistical approach has been unsatisfactory (see Section 1). It appears it requires a delicate combination of the two methods. We have presented a method that uses unsupervised combinatorial pattern discovery, followed by a careful statistical refinement and processing. Since we use tried-and-tested tools such as pattern discovery, in the first step, and local alignment, in the second step, we have focussed more on choosing and combining appropriate parameters. Also, extension of the method to handling a more general scenario such as inclusion of indels (insertion and/or deletion) in the embedded motif has been relatively straightforward. We achieved this by using extensible motifs in the pattern discovery process of the first step and gapped alignment in the second step. The results on benchmark data and some real DNA sequences have been very encouraging. We are looking at the yet harder instance of the problem which is the task of finding subtle motifs within the same sequence.

References

1. A. Apostolico, M. Comin, and L. Parida. Conservative extraction of over-represented extensible motifs. *ISMB (Supplement of Bioinformatics)*, 21:9–18, 2005.
2. A. Apostolico and L. Parida. Incremental paradigms for motif discovery. *Journal of Computational Biology*, 11(4):15–25, 2004.
3. T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. pages 28–36, 1994.
4. Buhler and Tompa. Finding motifs using random projections. In *Proceedings of the Annual Conference on Computational Molecular Biology (RECOMB01)*, pages 69–75. ACM Press, 2001.

5. Eleazar Eskin and Pavel Pevzner. Finding composite regulatory patterns in DNA sequences. In *Bioinformatics*, volume 18, pages 354–363, 2002.
6. G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.
7. Keich and Pevzner. Finding motifs in the twilight zone. In *Annual International Conference on Computational Molecular Biology*, pages 195–204, Apr, 2002.
8. Uri Keich and Pavel Pevzner. Subtle motifs: defining the limits of motif finding algorithms. In *Bioinformatics*, volume 18, pages 1382–1390, 2002.
9. C. E. Lawrence and Reilly A.A. An expectaion maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function and Genetics*, 7:41–51, 1990.
10. C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, Oct, 1993.
11. P. A. Pevzner and S.-H. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 269–278. AAAI Press, 2000.
12. Alkes Price, Sriram Ramabhadran, and Pavel Pevzner. Finding subtle motifs by branching from sample strings. In *Bioinformatics*, number 1, pages 149–155, 2003.
13. S. Rajasekaran, S. Balla, and C.-H Huang. Exact algorithms for planted motif problems. *Journal of Computational Biology*, 12(8):1117–1128, 2005.
14. M. F. Sagot. Spelling approximate repeated or common motifs using a suffix tree. *Latin 98: Theoretical Informatics, Lecture Notes in Computer Science*, 1380:111–127, 1998.
15. Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble1, Giulio Pavesi, Graziano Pesole, Mireille Rgnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23:137–144, 2005.
16. M.S. Waterman, R. Aratia, and D.J. Galas. Pattern recognition in several sequences: Consensus and alignment. *Bulletin of Mathematical Biology*, 46(4):515–527, 1984.