New Bounds for Motif Finding in Strong Instances

Broňa Brejová, Daniel G. Brown, Ian M. Harrower, and Tomáš Vinař

David R. Cheriton School of Computer Science, University of Waterloo
{bbrejova, browndg, imharrow, tvinar}@cs.uwaterloo.ca

Abstract. Many algorithms for motif finding that are commonly used in bioinformatics start by sampling r potential motif occurrences from n input sequences. The motif is derived from these samples and evaluated on all sequences. This approach works extremely well in practice, and is implemented by several programs. Li, Ma and Wang have shown that a simple algorithm of this sort is a polynomial-time approximation scheme. However, in 2005, we showed specific instances of the motif finding problem for which the approximation ratio of a slight variation of this scheme converges to one very slowly as a function of the sample size r, which seemingly contradicts the high performance of sample-based algorithms. Here, we account for the difference by showing that, for a variety of different definitions of "strong" binary motifs, the approximation ratio of sample-based algorithms converges to one exponentially fast in r. We also describe "very strong" motifs, for which the simple sample-based approach always identifies the correct motif, even for modest values of r.

1 Introduction

Motif finding is a combinatorial abstraction of the very important problem of regulatory sequence detection in bioinformatics. In motif finding, n discrete input sequences, each of length m, are given, as is a parameter L, called the motif length. The most common goal is to find a contiguous substring of length L in each input sequence, minimizing some function of these substrings (called *the motif occurrences*). One objective function is found in the Consensus-Pattern problem:

Definition 1 (Consensus-Pattern)**.** *Given are n sequences, $s_1, \ldots, s_n$, each of length m, over a finite alphabet Σ, and a parameter L. Find a contiguous subsequence x_i of length L from each sequence, and a consensus sequence x of these subsequences, minimizing $\sum_{i=1\ldots n} d_H(x_i, x)$, where $d_H(x, y)$ is the Hamming distance between two strings.*

While this problem is NP-hard, there is a simple sample-based polynomial-time approximation scheme for it. For a given value of r, the algorithm considers all samples of r substrings of length L from the n sequences. A motif derived from each such sample is then evaluated on all sequences, and the best motif is chosen. This algorithm was shown by Li *et al.* [4] to have approximation ratio of $1 + O(1/\sqrt{r})$ for constant-size alphabets. The algorithm also has $O(L(nm)^{r+1})$ runtime, which is polynomial if r is a constant.

M. Lewenstein and G. Valiente (Eds.): CPM 2006, LNCS 4009, pp. 94–105, 2006.

This bound is not especially useful, since the approximation ratio converges to one only very slowly with increasing r. Yet, sample-based algorithms with small values of r are very successful in practice for both motif finding in the abstract and for regulatory sequence detection [2, 5, 6, 9]. One might imagine that the bounds shown by Li *et al.* are weak, and the simple PTAS actually has a much stronger guarantee. However, in 2005 we showed [1] that this is very likely not the case. For a simple variation of the Li *et al.* PTAS (where the only difference is whether the sampling is without replacement or with replacement), we identified a collection of instances of the problem for which the approximation ratio is $1 + \Theta(1/\sqrt{r})$, suggesting that in order to achieve an approximation ratio of $1 + \varepsilon$, one needs a sample size of $r = \Theta(1/\varepsilon^2)$, which is highly impractical.

Still, the instances of Consensus-Pattern for which we proved our previous bounds are very weak motifs. They are binary instances of Consensus-Pattern; in each position of the motif instances, just over half of the entries are the symbol zero, and just under half are the symbol one. Such motifs are likely uninteresting, as they are no stronger than what we might expect to find if we considered random binary noise.

We might prefer to consider motifs bounded away from uniform noise. For such "strong" motifs, we can do much better: we show here that for various definitions of strong motifs, the approximation ratio of the algorithm approaches one exponentially fast as a function of r. In particular, for strong motifs, the approximation ratio is at most $1 + O(f^{-r})$, for a function f that depends only on the strength of the motif, instead of the approximation ratio of $1 + \Theta(1/\sqrt{r})$ shown for the general case.

Here, we show such theorems for a variety of different definitions of "strong" motifs. First, we consider binary motifs where at least a $\frac{1}{2} + \varepsilon$ fraction of all positions in the motif occurrences matches a given consensus of length L. While occasionally, the sampling PTAS can have bad performance on such an instance, we prove that for randomly chosen instances, the expected approximation ratio converges to one exponentially fast as a function of the sample size r. If we instead require consistently strong binary motifs, where each position of the motif has at least $(\frac{1}{2} + \varepsilon)n$ matches in the motif occurrences, we can prove that the PTAS performs well even in the worst case. In fact, for very strong consistent motifs, the PTAS will always find the correct answer, even for small r.

Our results document that while for arbitrary instances of motif finding, the simple sample-based PTAS may have poor convergence properties, for the kinds of motifs that people care about, the approximation ratio converges exponentially fast to the correct answer.

2 Background

The Consensus-Pattern problem, for an alphabet Σ, can be answered by enumerating all $|\Sigma|^L$ possible choices of the consensus pattern, and finding the best matches to each possible pattern, but such an enumeration is not efficient. Instead, one type of efficient heuristic for this problem first enumerates a

polynomial number of candidate consensus patterns, and then finds the best match to each candidate in each of the n sequences, in $O(nmL)$ time per candidate.

One set of candidates is all L-letter substrings of the input strings; there are $(m-L+1)n$ of them, yielding an algorithm with $O(L(nm)^2)$ runtime. Or, we can expand this idea to consider the result of looking at a sample of r substrings of the input. For each such sample, we compute a candidate motif as a consensus of the sample by identifying the most common letter at each position of the motif, breaking ties arbitrarily.

In this paper, we consider two algorithms based on this idea. The first uses samples with replacement, implying that a single substring can occur in the sample multiple times. There are $((m-L+1)n)^r$ such samples; if we try all of them, this yields an algorithm with $O(L(nm)^{r+1})$ runtime. We will refer to this as the PTAS algorithm. Li *et al.* [4] have shown that this simple algorithm is indeed a polynomial-time approximation scheme (a PTAS): the approximation ratio of the algorithm converges to one as the sample size r grows. Unfortunately, the convergence rate they could prove is very slow: they show the approximation ratio is at most $1 + \frac{4|\Sigma|-4}{\sqrt{e}(\sqrt{4r+1}-3)}$.

We will also study a slight modification of the PTAS, in which we consider only samples without replacement. We will refer to it as the SWOR algorithm, for "sampling without replacement". In our previous work [1], we gave specific instances of the problem for which the approximation ratio of the SWOR algorithm is $1 + \Theta(1/\sqrt{r})$ as a function of r. We conjectured that the same lower bound also holds for PTAS, which asymptotically matches the upper bound of Li *et al.*

2.1 Notation and Observations

To simplify our analysis, we will always assume that the input sequences $s_1, \ldots, s_n$ consist solely of the optimal motif occurrences, that is, $m = L$. While Consensus-Pattern is trivial in these cases, since the optimal motif is the consensus string of the input sequences, both PTAS and SWOR are still well-defined and may not always optimize the objective function. In fact, we showed in our earlier work [1] that if one of these algorithms is run on just the motif occurrences themselves, it will do no better than if run on longer sequences. Upper and lower bounds on the approximation ratio that we show for such instances are still applicable to longer sequences.

We will assume that the sequence alphabet Σ is the set $\{0, 1\}$; all of our results here are for binary motifs. If $m = L$, we can always transform the instance of the problem so that the optimal motif is the string 0^L, by relabelling characters in each column that has more ones than zeros. We will use this transformation in some of our results.

Finally, we note that since PTAS always explores more samples than SWOR, its approximation ratio is always at least as good as that of SWOR. Therefore, any upper bound for the approximation ratio of SWOR also applies to PTAS.

2.2 Concentration Bounds

Most of our bounds are obtained by applying the Hoeffding bound [3], which gives concentration bounds on the sum of independent random variables, and an extension of it to certain classes of dependent variables due to Panconesi and Srinivasan [8]. In this section, we summarize the probabilistic bounds we use. We begin with the following variant of the Hoeffding bound from McDiarmid's survey [7, p. 199]; a similar bound can be found in [3, Theorem 1].

Theorem 1 (Hoeffding's bound [7]). *Let $X_1, \ldots, X_n$ be independent random variables, with $0 \le X_k \le 1$ for each k. Let $X = \sum X_k$, let $\mu = E[X]$, let $p = \mu/n$ and let $q = 1 - p$. Then for any $0 \le t < q$,*

$$\Pr[X - \mu \ge nt] \le \left(\left(\frac{p}{p+t} \right)^{p+t} \left(\frac{q}{q-t} \right)^{q-t} \right)^n .$$

Panconesi and Srinivasan [8] have extended the Hoeffding bound to sums of dependent variables that satisfy certain conditions.

Theorem 2. *Let $X_1, \ldots, X_n$ be (not necessarily independent) binary random variables with $\Pr[X_k = 1] = p$ for each k. If for every subset A of $\{1, \ldots, n\}$ and for every $k \notin A$,*

$$\Pr\left[X_k = 1 \,\middle|\, \bigwedge_{j \in A} (X_j = 1) \right] \le \Pr[X_k = 1], \tag{1}$$

then Hoeffding's bound from Theorem 1 also holds for $X = \sum X_k$.

Proof. This is an application of Panconesi and Srinivasan's framework [8] for Chernoff-Hoeffding bounds of sums of dependent variables. Binary variables satisfying equation (1) are 1-correlated in the notation of Panconesi and Srinivasan. For such variables, we can apply the Hoeffding bounds directly, as though the variables were independent.

In particular, let $\hat{X}_1, \ldots, \hat{X}_n$ be independent random variables with $\Pr[X_k = 1] = p$. The variables $X = \sum_k X_k$ and $\hat{X} = \sum_k \hat{X}_k$ have the same expectation, $\mu = np$, and equation (1) implies that $\Pr[\wedge_{j \in A}(X_j = 1)] \le \prod_{j \in A} \Pr(\hat{X}_i = 1)$. Thus, these random variables satisfy the conditions of Theorem 3.2 in [8], and we obtain

$$\Pr[X - \mu \ge \varepsilon\mu] \le \frac{E[e^{h\hat{X}}]}{e^{h(1+\varepsilon)\mu}},$$

where ε and h are positive real numbers. As in the proof of Hoeffding's bound in McDiarmid [7, p. 199], we can prove $E[e^{h\hat{X}}] \le (1 - p + pe^h)^n$. By substituting $\varepsilon = t/p$, we obtain

$$\Pr[X - \mu \ge tn] \le \left(e^{-h(p+t)}(1 - p + pe^h) \right)^n ;$$

setting e^h to $\frac{(p+t)(1-p)}{p(1-p-t)}$, we obtain the desired result. □

Note that independent variables satisfy equation (1) with equality, so Theorem 1 is a special case of Theorem 2.

We will consider dependent binary random variables that are zero with some probability $p > 0.5$ and we will be interested in the probability that fewer than yn of the random variables are zero for some $0.5 \leq y < p$. Theorem 1 can be easily applied to this case, as is shown in the following lemma.

Lemma 1. *Let $X_1, \ldots, X_n$ be binary random variables with $\Pr[X_k = 0] = p$ for each k, where $p \geq 0.5$. If these variables satisfy the condition of Theorem 2, and $1 - p \leq y < p$ then $\Pr[\sum_k X_k \geq (1-y)n] \leq {\beta_y}^n$, where $\beta_y = \left(\frac{1-p}{1-y}\right)^{1-y} \left(\frac{p}{y}\right)^y$.*

Proof. The expectation of the variable $X = \sum_k X_k$ is $\mu = (1-p)n$. By Theorem 2, we easily obtain desired inequality:

$$\begin{aligned}\Pr[X \geq (1-y)n] &= \Pr[X - \mu \geq (p-y)n] \\ &\leq \left(\left(\frac{1-p}{1-p+(p-y)} \right)^{1-p+(p-y)} \left(\frac{p}{p-(p-y)} \right)^{p-(p-y)} \right)^n \\ &= {\beta_y}^n. \end{aligned}$$

□

Note that in the previous lemma, $\beta_y < 1$ for all p and y such that $0 < y < p < 1$. Therefore the probability that fewer than yn out of n variables are zeroes decreases exponentially as a function of n. For $y = 0.5$ we obtain the following special case.

Lemma 2. *Let $X_1, \ldots, X_n$ be binary random variables with $\Pr[X_k = 0] = p$ for each k, where $p \geq 0.5 + \varepsilon$. If these variables satisfy the condition of Theorem 2 then $\Pr[\sum_k X_k \geq n/2] \leq \alpha^n$, where $\alpha = \sqrt{4p(1-p)}$.*

3 Strong Motifs

We begin our analysis by considering motifs for which we know the number of zeros and ones in the motif instance. We do not necessarily fix the optimal motif to be 0^L.

Definition 2 (Strong motifs of fixed content). *A* strong motif of fixed content *p is a binary motif embedded into n sequences, where the total number of zeros in all n occurrences is pnL.*

Theorem 3. *For any value of r and $p > 0.5$, the worst-case approximation ratio of both PTAS and SWOR on strong motifs of fixed content at least p is the same as on arbitrary motifs.*

Proof. Consider the worst-case motif for a particular algorithm and value of r. Let p' be the number of zeros in this motif. If $p > p'$, we pad such an instance with enough columns, filled entirely with zeros, to make an instance of Consensus-Pattern that has at least pnL zeros. We have simply expanded the value of L.

The overall score of both the motif found by the algorithm and of the optimal motif is exactly the same as if we had not padded the instance with the extra columns. □

We have previously shown [1] that for any value of r, we can produce an instance of Consensus-Pattern for which SWOR has approximation ratio at least $1 + \Theta(1/\sqrt{r})$. This bound therefore transfers also to strong motifs of fixed content.

Thus, this definition of strong motifs does not give any better upper bound on the approximation ratio of the PTAS for motif finding than we had previously. The reason is that we allow many columns that are intensely weak and many columns that are very strong. In Section 4, we study motifs with more consistency among columns.

In the remainder of this section, we show that despite this negative result there are few bad instances of strong motifs, and if we choose a random strong motif of fixed content, the expected approximation ratio is much lower than in the worst case.

3.1 Randomly Chosen Strong Motifs

A *random motif of fixed content* p is an instance of the problem chosen uniformly from all $\binom{nL}{pnL}$ instances of the problem with exactly pnL zeros and $(1-p)nL$ ones. In such motifs, the zeros and ones may not be distributed uniformly, so some columns may contain more ones than zeroes. We call such columns *bad columns*; all other columns are *good columns*.

To analyze the expected approximation ratio of PTAS or SWOR on such randomly chosen motif, we divide all instances into *bad instances* and *good instances*. Bad instances have more than $L\alpha^{r/2}$ bad columns, where $\alpha = \sqrt{4p(1-p)}$. Lemma 3 shows that such instances are exponentially rare, and do not influence the expected approximation ratio much. Good instances have at most $L\alpha^{r/2}$ bad columns. In Lemma 4, we will show that for such instances, the approximation ratio is low.

Lemma 3. *The probability that a random binary motif of fixed content p is a bad instance is at most $\alpha^{r/2}$.*

Proof. Let $X_{i,j}$ be a binary random variable representing the symbol in row i and column j of the motif instance. For a given column j, let the number of ones be $X_j = \sum_i X_{i,j}$. Column j is bad if X_j is more than $n/2$. Each one in a column reduces the probability of others, so the variables corresponding to this column satisfy the conditions of Lemma 2, so $\Pr[X_j > n/2] \leq \alpha^n$. Since $n \geq r$, this probability is also at most α^r. By linearity of expectation, the expected number of bad columns is at most $L\alpha^r$.

Since a bad motif contains more than $L\alpha^{r/2}$ bad columns, we are bounding the probability that the number of bad columns is more than $\alpha^{-r/2}$ times its mean. This can be no greater than $1/\alpha^{-r/2}$, by the Markov inequality. □

Lemma 4. *The expected cost of a motif returned by PTAS (or SWOR) on a randomly chosen good instance is less than $nL\left(\frac{1-p+2p\alpha^r}{1-\alpha^{r/2}}\right)$.*

Proof. Let X_j be a random variable representing the number of ones in column j. Consider a random sample without replacement of r rows. Let Y_j be the number of ones in column j of this sample. The consensus of the sample is one when $Y_j \geq r/2$. Finally, let A_j be the score of the consensus character of this random sample in column j.

We want to bound $E[A_j|G]$, where G is the event that the motif instance is good. Since A_j is always non-negative, this conditional expectation is at most $E[A_j]/\Pr[G]$. In Lemma 3 we have shown that $\Pr[G] \geq 1 - \alpha^{r/2}$.

Consider the event Z_j that column j is good and the random sample has consensus zero. As for Lemma 3, Lemma 2 gives that $\Pr[X_j > n/2] \leq \alpha^r$, and $\Pr[Y_j \geq r/2] \leq \alpha^r$. Therefore the probability of Z_j is at least $1 - 2\alpha^r$.

In the case of the event Z_j, we are skewed towards having more zeroes than expected, and therefore

$$E[A_j|Z_j] = E[X_j|Z_j] \leq E[X_j] = n(1-p).$$

If we are not in Z_j, the cost of the column is at most n. Therefore, the expected cost of a single column is

$$\begin{aligned} E[A_j|G] &\leq \frac{\Pr[Z_j] \cdot E[A_j|Z_j] + \Pr[\bar{Z}_j] \cdot E[A_j|\bar{Z}_j]}{\Pr[G]} \\ &\leq \frac{(1-2\alpha^r)n(1-p) + 2\alpha^r n}{1-\alpha^{r/2}} = n \cdot \frac{1-p+2p\alpha^r}{1-\alpha^{r/2}} \end{aligned}$$

By linearity of expectation, the expected cost over all columns is at most $L \cdot E[A_j|G]$, and at least one sample in the SWOR algorithm must give us a motif which has at most this cost. □

With this in mind, we can bound the performance of the PTAS for random motifs of fixed content p.

Theorem 4. *When applied to a random motif of fixed content $p > 0.5 + \varepsilon$, both PTAS and SWOR have expected approximation ratio at most $1 + \alpha^{r/2} \cdot \frac{18-16p+2p^2}{1-p^2}$ for sufficiently large r, where $\alpha = \sqrt{4p(1-p)}$.*

Proof. For sufficiently large r, $\alpha^{r/2}$ is at most $(1-p)/2$. For good instances, the number of ones in good columns is at least $Ln(1-p-\alpha^{r/2})$, which gives a non-negative lower bound on the optimal motif cost. Therefore, according to Lemma 4, the approximation ratio for such instances can be bounded by $\frac{1-p+2p\alpha^r}{(1-\alpha^{r/2})(1-p-\alpha^{r/2})}$.

We have previously shown [1] that any sampling algorithm has approximation ratio no more than 2 on all instances. We will use this upper bound for bad instances. This gives an overall bound of no greater than $\frac{1-p+2p\alpha^r}{(1-\alpha^{r/2})(1-p-\alpha^{r/2})} + 2\alpha^{r/2}$. Rearranging, we obtain the upper bound $1 + \alpha^{r/2} \cdot \frac{4-3p-5\alpha^{r/2}+4\alpha^{r/2}p+2\alpha^r}{(1-\alpha^{r/2})(1-p-\alpha^{r/2})}$. For sufficiently large r, $\alpha^{r/2} \leq (1-p)/2$, and we obtain the desired bound. □

This theorem shows that if either PTAS or SWOR is applied to a random strong motif of fixed content $p > 0.5$, the expected approximation ratio is $1 + O(\alpha^{r/2})$. This bound converges exponentially quickly to one as a function of r.

As a function of p, the bound on the approximation ratio is decreasing, as long as $p > 0.5$ and $r \geq 4$. This property will be important in the next section.

3.2 Strong Motifs of Fixed Expected Content

Perhaps more natural as a model of random motifs is the case where each of the L positions in all n sequences is chosen independently of all others, with probability p.

Definition 3 (Strong motif of fixed expected content). *A* strong motif of fixed expected content p *is a random motif where each position is zero with probability* p*, and one with probability* $1 - p$ *independently of other positions.*

This stochastic model can generate bad instances of the problem again, but it is very rare that such instances occur, and we can again always bound their approximation ratio by 2, so their contribution to the expected approximation ratio is small.

Theorem 5. *For strong binary motifs of expected content* $p > 0.5$*, where* p *is a fixed constant, the expected approximation ratio of the PTAS and SWOR is at most* $1 + O(\gamma^r)$*, for some constant* $\gamma < 1$ *that depends on* p*, but not* r*.*

Proof. Let $q = 1/4 + p/2$. According to Lemma 1, for a strong motif of expected content $p > 1/2$, the probability that the actual motif generated has fewer than qnL zeros is less than ${\beta_q}^{nL}$, where β_q is less than one. This is certainly less than ${\beta_q}^{r}$, since $nL \geq r$. For these weak motifs, we use the upper bound of 2 on the approximation ratio.

The remaining instances are strong motifs of content at least q. We can treat the process as first picking the motif content $\pi \geq q$, then picking a random motif of that fixed content. For a fixed content π, we can apply the bound given in Theorem 4. Since this bound decreases with increased strength of the motif, we can use the upper bound obtained with Theorem 4 for content q for all values of π. Therefore the overall approximation ratio of the algorithm is at most $1 + {\alpha_q}^{r/2} \cdot \frac{18-16q+2q^2}{1-q^2} + 2{\beta_q}^{r/2}$, $\alpha_q = \sqrt{4q(1-q)}$, for sufficiently large r, and by setting $\gamma = \sqrt{\max\{\alpha_q, \beta_q\}}$, we obtain the desired bound. □

3.3 Many Motifs Are Weak

We finish this section by noting that for any value of r, we can pick an instance size n for which in fact most motifs are weak, and for which we conjecture that the PTAS has poor convergence. If we let $p = 0.5$, and sample from the distribution of all binary motifs with expected content p, then all motif instances are equiprobable, so theorizing about the common behaviour of the algorithm also applies to common motif instances. A random motif of this content with

$n = r^2$ sequences is expected to have a constant fraction of columns in which the fraction of zeros in the column is between $1/2 + 1/\sqrt{r}$ and $1/2 + 2/\sqrt{r}$; that is, a significant fraction of the columns will be weak to the point where a random sample without replacement of r motif instances has a constant probability of picking the incorrect symbol for that column. Further, the expected cost of a random motif will be of the order of $(1/2 + \Theta(1/\sqrt{r}))nL$.

A random sample (without replacement) will incorrectly assign the symbols in a constant fraction of the motif's columns, giving an expected cost increase on the order of $\Theta(1/\sqrt{r})nL$ over the optimum, and an overall approximation ratio of $1 + \Omega(1/\sqrt{r})$. We conjecture that this bound applies to all samples, and that the overall performance of the best sample is also $1 + \Omega(1/\sqrt{r})$.

4 Consistently Strong Motifs

In the previous section, we were not able to guarantee a good performance of the PTAS in the worst case. This was because some instances of strong motifs may have contained many columns with approximately the same number of zeros and ones. Here, we study the performance of the PTAS on consistently strong motifs, where each motif column has a large number of zeros in it.

Definition 4 (Consistently strong motif). *A* consistently strong motif of content $p > 0.5$ *is a binary motif embedded into n sequences, where each column of the motif has at least pn zeros.*

We first note the performance of the algorithms PTAS and SWOR on a single column of a consistently strong binary motif.

Lemma 5. *Suppose that we choose a random sample of r rows (with or without replacement) from a motif instance in which a particular column has pn zeros and* $(1-p)n$ *ones, for* $p > 0.5$. *The expected cost of the consensus character of the sample in this column is at most* $n((1-p)(1-\alpha^r)+\alpha^r)$, *where* $\alpha = \sqrt{4p(1-p)}$.

Proof. First, we want to bound the probability that the random sample without replacement has fewer zeroes than ones in this column, in which case the consensus character will be one. This situation satisfies conditions of Lemma 2 and the probability that at least half of the sample will be ones is at most α^r.

Note that for any constant $p > 0.5 + \varepsilon$, for some positive ε, this bound on the probability of erring in a single column converges to zero exponentially fast in r. Therefore, such samples will not have much influence on the expected cost, and we can bound their cost from above by n. The cost of a sample with consensus zero is exactly $n(1-p)$. Therefore, the expected cost is at most $n[(1-p)(1-\alpha^r)+\alpha^r]$. □

Theorem 6. *For sufficiently large r, both PTAS and SWOR, applied to a consistently strong motif of content* $p > 0.5$, *have approximation ratio at most* $1 + \alpha^r \cdot \frac{p}{1-p}$, *where* $\alpha = \sqrt{4p(1-p)}$.

Proof. Let p_i be the content of zeros of the i-th column of the motif. According to Lemma 5, the expected cost $e(p_i)$ of the i-th column is at most $n((1-p_i)(1-{\alpha_{p_i}}^r)+{\alpha_{p_i}}^r)$, where $\alpha_{p_i} = \sqrt{4p_i(1-p_i)}$. The optimal cost of the same column is $o(p_i) = n(1-p_i)$. From the linearity of expectation, the expected approximation ratio of a consensus of a random sample of r rows over all columns of the motif is $R = \frac{\sum_{i=1}^{L} e(p_i)}{\sum_{i=1}^{L} o(p_i)}$.

Note, that for sufficiently large r (in particular, $r > 2/(2p-1)$), the function $e(p')/o(p')$ is decreasing with increasing value of p' for $p' \geq p$. Therefore, $e(p_i) \leq e(p)o(p_i)/o(p)$, and thus

$$R \leq e(p)/o(p) \leq \frac{n((1-p)(1-\alpha^r)+\alpha^r)}{n(1-p)} = 1 + \alpha^r \cdot \frac{p}{1-p}.$$

At least one sample must achieve this bound, by the first moment principle. Since SWOR examines all samples without replacement, the sample found by SWOR achieves the bound. □

If $p > 0.5 + \varepsilon$ for some constant $\varepsilon > 0$, then this ratio converges exponentially quickly to one.

4.1 Very Strong Consistent Motifs

We finish by noting that some motifs are so strong that the PTAS is guaranteed to find them exactly.

We saw in the proof of Lemma 5 that we can bound the probability of making an error for any column, when we sample r motif instances of that column. If the column has frequency p_i of zeros, the error probability was at most ${\alpha_{p_i}}^r$, where $\alpha_{p_i} = \sqrt{4p_i(1-p_i)}$.

If we have a motif whose columns are strong enough so that the sum of the ${\alpha_{p_i}}^r$ is at most one, the standard union bound gives that the probability that at least one sample column has more ones than zeros is less than 1. Thus, there must exist a sample of r motif instances whose consensus is exactly the correct L-letter-long motif. Since the PTAS is exhaustive, we will examine this sample, and it will be found by the algorithm.

In particular, a motif strong enough that the value of ${\alpha_{p_i}}^r$ is always less than $1/L$ will always be found by the PTAS.

Theorem 7. *The PTAS always finds the correct motif when its input is a consistently strong binary motif of length L with probability $p \geq \frac{1}{2} + \frac{\sqrt{1-L^{-2/r}}}{2}$.*

Proof. This is shown by noting that $\frac{1}{2} + \frac{\sqrt{1-L^{-2/r}}}{2}$ is the root in the range $(0.5, 1]$ of $(4p(1-p))^{r/2} = 1/L$, corresponding to the value where α_p goes below $1/L$. □

This value quickly shrinks for values of r that are not especially large: for a length 10 binary motif, if all columns are at least 80% zeroes, examining all samples of size 11 is certain to find the true motif, while samples of size 5 are

all that is needed for motifs of that length where $p = 0.9$. Indeed, for a fixed value of L, if the motif is consistently strong with probability at least $0.5 + f(r)$, where $f(r)$ is a specific function that is only $O(1/\sqrt{r})$, the PTAS will find the optimal motif.

For random motifs, the situation is not as good; obviously, a random motif with probability p might turn out not to be strong. But, for p large enough, the probability of producing a motif that is weak enough that the algorithm has positive probability of failing can easily be estimated, and again converges exponentially rapidly to zero as a function of p or of r, for a fixed motif length L.

5 Conclusion and Open Problems

We have shown a variety of characterizations of "strong" binary instances of Consensus-Pattern for which the simple sampling-based polynomial-time approximation scheme of Li *et al* [4] has an approximation ratio guarantee that converges to one exponentially fast as a function of r, the sample size. This result is in contrast with our previous work, which showed specific instances of Consensus-Pattern for which a variation of the Li *et al.* PTAS can only achieve $1 + \Theta(1/\sqrt{r})$ approximation ratio.

The difference is quite significant; to achieve $1 + \varepsilon$ approximation ratio using the general bound requires samples of size $\Omega(1/\varepsilon^2)$, giving runtimes of $O(L(nm)^{\Omega(1/\varepsilon^2)})$, whereas for strong motifs we show that a sample size of $O(\log(1/\varepsilon))$ is sufficient.

Our bounds apply to random binary motifs of specific strength, or to those for which the probability that any specific position is a zero is fixed to be a constant bounded above 0.5. While it is possible to obtain a difficult-to-solve instance of the problem by chance, such instances are exponentially rare, and as such, do not affect the algorithm's behaviour significantly.

Finally, we show that for strong instances, small samples can guarantee that the motif found is optimal. While the bounds achieved are not practical, this again suggests that motif finding is an easy problem when applied to strong instances, and only hard when applied to irrelevant, weak problem instances.

Open problems. How tight are the bounds for very strong consistent motifs given in Section 4.1? Can we find specific strong instances of Consensus-Pattern for which the sample-based PTAS finds a wrong motif and for which the value of r is close to the one shown in the theorem, or is the bound very loose?

In all our theorems, we have considered only binary alphabet. Our results extend to non-binary alphabets; however, we still require that one of the alphabet symbols has frequency more than 0.5. The problem of regulatory sequence detection is most commonly applied to DNA and protein sequences, and it makes sense to consider instances in which the most common letter in each column is significantly more common than other letters, but still does not achieve frequency more than 0.5. To prove exponential convergence for such instances likely requires a variation on the Chernoff-Hoeffding bounds for multi-outcome variables.

Acknowledgements

All authors are supported by the Natural Science and Engineering Research Council of Canada. We would like to thank Nick Wormald for a helpful conversation and for pointing us to the paper of McDiarmid [7].

References

1. B. Brejova, D.G. Brown, I.M. Harrower, A. Lopez-Ortiz, and T. Vinar. Sharper upper and lower bounds for an approximation scheme for Consensus-Pattern. In A. Apostolico, M. Crochemore, and K. Park, editors, *Combinatorial Pattern Matching, 16th Annual Symposium (CPM 2005)*, pages 1–10, 2005.
2. G.Z. Hertz and G.D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–577, 1999.
3. W.J. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:713–721, 1963.
4. M. Li, B. Ma, and L. Wang. Finding similar regions in many strings. *Journal of Computer and System Sciences*, 65(1):73–96, 2002.
5. C. Liang. COPIA: a new software for finding consensus patterns in unaligned protein sequences. Master's thesis, University of Waterloo, October 2001.
6. J. Liu. A combinatorial approach for motif discovery in unaligned DNA sequences. Master's thesis, University of Waterloo, March 2004.
7. C. McDiarmid. Concentration. In M. Habib, editor, *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.
8. A. Panconesi and A. Srinivasan. Randomized distributed edge coloring via an extension of the Chernoff-Hoeffding bounds. *SIAM Journal on Computing*, 26:350–368, 1997.
9. P.A. Pevzner and S. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, pages 269–278, 2000.