Sharper Upper and Lower Bounds for an Approximation Scheme for Consensus-Pattern

Ian Harrower
School of Computer Science
University of Waterloo
imharrow@cs.uwaterloo.ca

Joint work with Broňa Brejová, Daniel G. Brown, Alejandro López-Ortiz and Tomáš Vinař.

Motif Discovery

- Given a collection of strings, we seek a motif:
  an approximate substring of all input strings
- Useful objective functions NP-Hard to optimize
- Many effective heuristic sample-based algorithms
- Approximation guarantees exist for a simple PTAS

Best Known Guarantees

- Let r be the size of the sample. Consider all samples of that size
- [Li, Ma, Wang. STOC '99] Simple-sample based PTAS, if $r \geq 3$.
- Many common algorithms use essentially this approach with $r \leq 3$
- But known bounds for PTAS are hopeless
  - DNA ($A = 4$): guarantee around 13
  - Proteins ($A = 20$): guarantee around 77

Approx. guarantee: $1 + \frac{4A-4}{\sqrt{e}\left(\sqrt{4r+1}-3\right)}$ (A = alphabet size)

- Increasing the sample size $\Rightarrow$ hopelessly long runtimes
- Why are sample driven algorithms successful?

Our Results

Stronger bounds for small samples:

- Tight approximation guarantee of 2, for $r = 1$
- Approximation guarantee between 1.50 and 1.53, for $r = 3$

New lower bounds:

- Lower bound of $1 + \Theta(1/r^2)$, for general r
- Lower bound of $1 + \Theta(1/\sqrt{r})$, for related algorithm

Conjecture that approximation ratio is independent of alphabet size (A)

Background: Problem Definition

- Given:
  - n string $s_1, \ldots, s_n$
    - Each strings of length m
    - Strings over an alphabet of size A: $\{0, \ldots, A-1\}$
  - Motif Length L

Background: Problem Definition

- Given:
  - n string $s_1, \ldots, s_n$
    - Each strings of length m
    - Strings over an alphabet of size A: $\{0, \ldots, A-1\}$
  - Motif Length L

- Find:
  - Substring t_i in each sequence s_i
    - Length of t_i is L
  - Motif string s
  - Where we optimize some objective function of $t_1, \ldots, t_n$ and s

Background: Problem Definition

- Given:
  - n string $s_1, \ldots, s_n$
    - Each strings of length m
    - Strings over an alphabet of size A: $\{0, \ldots, A-1\}$
  - Motif Length L

- Find:
  - Substring t_i in each sequence s_i
    - Length of t_i is L
  - Motif string s
  - Where we optimize some objective function of $t_1, \ldots, t_n$ and s

Consensus-Pattern : Minimize total hamming distance to center: $\sum_i d_H(s, t_i)$

A Simple PTAS

- Simple PTAS [Li, Ma, Wang. STOC '99] :
  - For all choices of r substrings of length L
    - Find consensus string M_C
    - Choose one substring from each sequence that minimizes distance to M_C
  - Return minimum score among these solutions

- Runtime: $\Theta(L(nm)^{r+1})$ (There are $\Theta((nm)^r)$ samples)

- Note: Sampling done $with$ replacement, same substring can occur multiple times

- We will call this algorithm LMW

First Result: A Tight Bound for $r = 1$

For $r = 1$, approximation ratio at most 2, and this bound is tight

- Observation 1: Can restrict attention to $m = L$
  - Consider a CONSENSUS-PATTERN instance with optimal solution $t_1^*, \ldots, t_n^*, s^*$
  - Running LMW on sequences $t_1^*, \ldots, t_n^*$ will examine a subset of the samples on the original instance

  $\Rightarrow$ Approx. ratio for entire instance as good as ratio on only the optimal solution

- Note: When $m = L$ optimal solution is trivial to find, but LMW may not find it

- If $m = L$ can assume WLOG the optimal motif is 0^L

$r = 1$: Upper Bound of 2

Approx. ratio of LMW is at most 2 for all values of r and all alphabet sizes.

- Let c be the cost of optimal motif 0^L
- Let a_i be the number of non-zero sites in s_i. So $c = \sum_i a_i$
- Motif s_i (considered by LMW for all r) has cost at most $c + a_i n$
- Sum of costs for each possible s_i is $nc + n\sum_i a_i = 2nc$
- Mean cost is $2c$
  - First moment principle implies some s_i gives cost at most $2c$
  - So LMW is a 2-approximation, if $r = 1$

$r = 1$: Lower Bound of 2

LMW with $r = 1$ has instances for which the approx. bound is arbitrarily close to 2.

- Consider the identity matrix I_n (n strings of length n)
- Cost of optimal solution 0^n is n
- LMW selects a single row as motif. WLOG suppose it selects 10^{n-1}

  $\Rightarrow$ Cost is $n - 1 + (n - 1)(1) = 2n - 2$

$\Rightarrow$ Approx. ratio is $2 - 2/n$, which converges to 2 as $n \to \infty$

- Note: same holds for $r = 2$

$r = 3$: Worst-case Ratio Near 1.5

- Lower bound 1.5
  - For any k, consider, $n = 2k$, $L = 2$
  - All $2k$ strings of the form $0i$ or $i0$, $i = 1 \ldots k$
  - Optimal cost $2k$, LMW cost is $3k - 1$

  $\Rightarrow$ Approx. ratio at least 1.5

- Upper bound $(64 + 7\sqrt{7})/54 \approx 1.528$
  - Again proved using first moment principle

$$\begin{array}{cc} 0 & 1 \\ 0 & 2 \\ \vdots & \\ 0 & k \\ 1 & 0 \\ 2 & 0 \\ \vdots & \\ k & 0 \end{array}$$

A Strong Lower Bound on Approx. Guarantee

Consider modification of LMW, where we allow only a $single$ sample from each input sequence. This modified algorithm has approx. guarantee at least $1 + \Theta(1/\sqrt{r})$.

$\Rightarrow$ To obtain a bound of $1 + \varepsilon$ requires that r is $\Omega(1/\varepsilon^2)$.

- Assume $r = (2k+1)^2$, set $n = 2r$
- Let $L = \binom{2r}{r+\sqrt{r}}$: include all possible columns with $r - \sqrt{r}$ 1s and $r + \sqrt{r}$ zeros.

A Sample Instance

$$
\begin{array}{cccc c cc}
1 & 1 & 1 & 1 & & 0 & 0 \\
1 & 1 & 1 & 1 & & 0 & 0 \\
1 & 1 & 1 & 1 & & 0 & 0 \\
1 & 1 & 1 & 1 & & 0 & 0 \\
1 & 1 & 1 & 1 & & 0 & 0 \\
1 & 0 & 0 & 0 & & 0 & 0 \\
0 & 1 & 0 & 0 & & 0 & 0 \\
0 & 0 & 1 & 0 & & 0 & 0 \\
0 & 0 & 0 & 1 & \ldots & 0 & 0 \\
0 & 0 & 0 & 0 & & 0 & 0 \\
0 & 0 & 0 & 0 & & 0 & 0 \\
0 & 0 & 0 & 0 & & 1 & 0 \\
0 & 0 & 0 & 0 & & 0 & 1 \\
0 & 0 & 0 & 0 & & 1 & 1 \\
0 & 0 & 0 & 0 & & 1 & 1 \\
0 & 0 & 0 & 0 & & 1 & 1 \\
0 & 0 & 0 & 0 & & 1 & 1 \\
0 & 0 & 0 & 0 & & 1 & 1
\end{array}
$$

- $r = 9$
- $n = 18$
- 6 1s and 12 0s per column
- $L = \binom{18}{12} = 18564$
- Optimal solution $0^{\binom{18}{12}}$ with cost $6 \times \binom{18}{12}$

- Optimal solution is 0^L with cost $L \cdot (r - \sqrt{r})$
- Note: any combination of r rows distinct gives rise to an equivalent solution

A Probabilistic Approach

- Consider, p_r: probability a random sample of size r has a 1 in a particular column. (Fraction of errors in chosen motif)

- By linearity of expectation, the chosen motif is expected to have Lp_r 1s

  – Cost is

$$\underbrace{L \cdot p_r \cdot (r + \sqrt{r})}_{\text{columns with a 1 in motif}} + \underbrace{L \cdot (1 - p_r)(r - \sqrt{r})}_{\text{columns with a 0 in motif}}$$

- All solutions from the algorithm have this same cost

- Approximation ratio $> 1 + 2p_r/\sqrt{r}$

A Probabilistic Approach - Finishing the Proof

- Have shown: approximation ratio $> 1 + 2p_r/\sqrt{r}$
- We wish to bound p_r below by some constant
  - p_r is probability $\geq r/2$ rows are 1 in the sample (for a given column)
  - p_r is based on sampling r times from a population of $r + \sqrt{r}$ 0s and $r - \sqrt{r}$ 1s
  - As $r \to \infty$
    - Number of 1s in the sample approaches $\frac{1}{2}(r - \sqrt{r})$
    - Number of 1s in the sample approaches a normal distribution (Central Limit Theorem for Finite Populations)
    - Bound probability we have $\geq r/2$ 1s, $p_r \geq 0.023$

$\Rightarrow$ Approx. ratio for modified algorithm at least $1 + \Theta(1/\sqrt{r})$

What Can We Show about LMW?

- Previous lower bound only proved when sampling without replacement
- No known upper bounds for modified algorithm
- Can show $1 + \Theta(1/r^2)$ lower bound for LMW

Conjectures:

- $1 + \Theta(1/\sqrt{r})$ bounds holds for LMW and that the instance generated is actually the worse case example.
- The modified algorithm for sampling without replacement is a PTAS