DIMACS Technical Report 2003-14
April 2003

Tight Approximability Results for Test Set Problems in Bioinformatics

by

Piotr Berman
Department of Computer Science and Engineering
Pennsylvania State University
University Park, PA 16802
Email: berman@cse.psu.edu

Bhaskar DasGupta[1]
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60607
Email: dasgupta@cs.uic.edu

Ming-Yang Kao[2]
Department of Computer Science
Northwestern University
Evanston, IL 60201
Email: kao@cs.northwestern.edu

[1]Supported in part by NSF grants CCR-0296041 and CCR-0208749.
[2]Supported in part by NSF grant EIA-0112934.

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, NEC Research Institute and Telcordia Technologies (formerly Bellcore).

DIMACS was founded as an NSF Science and Technology Center, and also receives support from the New Jersey Commission on Science and Technology.

ABSTRACT

In this paper, we investigate several versions of a test set problem. In the simplest version, given a family of *tests*, viewed as subsets of the universe, we want to select the smallest subfamily that *separates* every two elements of the universe. We show that this problem is exactly as difficult to approximate as the set cover problem, up to a factor of $1 \pm o(1)$. The same holds for several more elaborate versions of the problem that have applications in bioinformatics; for example, in one version the universe consists of strings over a finite alphabet and each test checks the presence of a specific substring. We also show that if we allow to form tests as unions of the given tests, the problem has essentially the same level of computational intractability as graph coloring.

1 Introduction and Motivation

Test set problems arise naturally in many applications one of which was listed long back as an NP-complete problem by Garey and Johnson [5]. In this paper we investigate the approximability issues of several problems that can be formalized as special cases of the following *test set* problem denoted by $\mathrm{TS}^{\Gamma}(k)$. Let $[i,j]$ be the set of integers $\{i, i+1, \ldots, j\}$, $\mathcal{P}(S)$ be the power set of S *i.e.* $\{A : A \subset S\}$, and let $|X|$ denote the cardinality (resp. length) of X if X is a set (resp. sequence). $\mathrm{TS}^{\Gamma}(k)$ with parameters $\Gamma \subseteq \mathcal{P}([0,2])$ and $k \in \mathbb{N}$ is defined as follows:

Instance: $(n, \mathcal{S})$ where $\mathcal{S} \subset \mathcal{P}([0, n-1])$.

Definition: A k-*test* is a union of at most k sets from $\mathcal{S}$. A test T γ-distinguishes $x, y \in [0, n-1]$ (for a $\gamma \in \Gamma$) if $|\{x,y\} \cap T| \in \gamma$.

Valid solution: A collection $\mathcal{T}$ of k-tests such that $\forall x \in [0, n{-}1]\ \forall y \in [0, n{-}1]{-}\{x\}\ \forall \gamma \in \Gamma\ \exists T \in \mathcal{T}$ such that T γ-distinguishes x and y.

Objective: *minimize* $|\mathcal{T}|$.

This formalism captures problems in several areas in bioinformatics and biological modeling as described below:

Minimum Test Collection Problem: This is precisely $\mathrm{TS}^{\{1\}}(1)$ and has applications in diagnostic testing. Garey and Johnson [5, pp. 71] showed a proof of NP-hardness of this problem via a reduction from the 3-dimensional matching problem.

Condition Cover Problem: Karp *et al.* [8] considered a problem in which the elements are outputs of a Boolean circuit without feedback modeling a biological pathway and there is a set corresponding to each given input perturbation (condition) consisting of those outputs which are *true* for the specific condition in question. If the Boolean circuit is arbitrary, the condition cover problem is identical to $\mathrm{TS}^{\{1\},\{0,2\}}(1)$ since each output can be expressed as a Boolean expression of the inputs.

String Barcoding Problem: For two sequences (strings) $u, v \in \Sigma^*$, v is a substring of x (denoted by $v \prec x$) if $x = uvw$ for some $u, w \in \Sigma^*$. Define a k-*sequence* as a collection of at most k distinct sequences. In this problem, considered by Rash and Gusfield [13] for the case when $k = 1$, we are given a set $\mathcal{S}$ of sequences over some alphabet Σ. For a *fixed* set of m k-sequences $\vec{\mathbf{t}} = (\mathbf{t}_0, \ldots, \mathbf{t}_{m-1})$, the *barcode* $\mathtt{code}(s, \vec{\mathbf{t}})$ for each $s \in S$ is defined to be the Boolean vector (c_0, c_1, c_{m-1}) where c_i is 1 iff there exists a $t \in \mathbf{t}_i$ such that $t \prec s$. We say that $\vec{\mathbf{t}}$ defines a *valid* barcode if for any two distinct strings $s, s' \in \mathcal{S}$, $\mathtt{code}(s, \vec{\mathbf{t}})$ is different from $\mathtt{code}(s', \vec{\mathbf{t}})$. The string barcoding problem over alphabet Σ, denoted by $\mathrm{SB}^{\Sigma}(k)$, has a parameter $k \in \mathbb{N}$ and is defined as follows:

Instance: $(n, \mathcal{S})$ where $\mathcal{S} \subset \Sigma^*$ and $1 \leq k \leq n = |\mathcal{S}|$.

Valid solution: a set of k-sequences $\vec{\mathbf{t}}$ defining a valid barcode.

Objective: *minimize* $|\vec{t}|$.

$\mathrm{SB}^{\Sigma}(k)$ is a special case of $\mathrm{TS}^{\{1\}}(k)$ in which $U = \mathcal{S}$ and for each substring p of each sequence in S there is a test set $\{s \in \mathcal{S} \,:\, p \prec s\}$; valid barcodes can be identified with valid sets of k-tests. A motivation for investigating $\mathrm{SB}^{\Sigma}(k)$ for $k > 1$ comes from its application to DNA microarray designs for efficient virus identification in which the immobilized DNA sequences at an array element are from a set of barcodes. The exact complexity and approximability of the $\mathrm{SB}^{\Sigma}(1)$ problem was left as an open problem by Rash and Gusfield in [13].

Minimum Cost Probe Set Problem with Threshold r: This problem, similar to $\mathrm{SB}^{\Sigma}(1)$, was considered by Borneman *et al.* [2]. Denote by $\mathsf{oc}(x, y)$ the number of occurrences of x in y as a substring, The minimum cost probe set problem with threshold r, denoted by $\mathrm{MCP}^{\Sigma}(r)$, is a variation of $\mathrm{TS}^{\{1\}}(1)$ that was used in [2] for minimizing the number of oligonucleotide probes needed for analyzing populations of ribosomal RNA gene (rDNA) clones by hybridization experiments on DNA microarrays.

For a fixed set of m sequences $\vec{t} = (t_0, t_1, \ldots, t_{m-1})$, an r-barcode $\mathsf{code}(s, \vec{t})$ for any sequence s is defined to be the vector $(c_0, c_1, \ldots, c_{m-1}$ where $c_i = \min\{r, \mathsf{oc}(t_i, s)\}$. Given a set $\mathcal{S}$ of sequences over some alphabet Σ, $\vec{t}$ defines a *valid* r-barcode if for any two distinct strings $s, s' \in \mathcal{S}$, $\mathsf{code}(s, \vec{t})$ is different from $\mathsf{code}(s', \vec{t})$. $\mathrm{MCP}^{\Sigma}(r)$ is now defined as follows:

Instance: $(n, r, \mathcal{S}, \mathcal{P})$ where $\mathcal{S}, \mathcal{P} \subset \Sigma^*$ and $|\mathcal{S}| = n$.

Valid solution: a set of sequences $\vec{t} \in \mathcal{P}^*$ defining a valid r-barcode.

Objective: *minimize* $|\vec{t}|$.

If $\mathcal{P}$ is the set of all substrings of sequences in $\mathcal{S}$, $\mathrm{MCP}^{\Sigma}(1)$ is precisely $\mathrm{SB}^{\Sigma}(1)$. Borneman *et al.* [2] noted that $\mathrm{MCP}^{\Sigma}(r)$ is NP-complete assuming that the lengths of the sequences in the prespecified set P were unrestricted, but no other nontrivial theoretical results are known. All our results on $\mathrm{SB}^{\Sigma}(1)$ apply to $\mathrm{MCP}^{\Sigma}(r)$ with little modification.

2 Summary of Our Results

Our results provide almost matching upper and lower bounds on approximation ratio of polynomial-time algorithms for $\mathrm{TS}^{\{1\}}(1)$, $\mathrm{TS}^{\{1\},\{0,2\}}(1)$, $\mathrm{SB}^{\Sigma}(1)$ and $\mathrm{MCP}^{\Sigma}(r)$ and strong lower bounds on approximation ratios of polynomial-time algorithms for $\mathrm{TS}^{\{1\}}(k)$, $\mathrm{TS}^{\{1\},\{0,2\}}(k)$ and $\mathrm{SB}^{\Sigma}(k)$ for all large k as summarized in Table 1 below:

Moreover, we also prove the following results on the expected size of solutions in random instances of $\mathrm{TS}^{\{1\}}(1)$, $\mathrm{TS}^{\{1\},\{0,2\}}(1)$ and $\mathrm{SB}^{\Sigma}(1)$, assuming all related parameters to be sufficiently large.

Theorem 1
(a) *Let $(n, \mathcal{S})$ be a random instance of $TS^{\alpha}(1)$ such that for any element $i \in [0, n-1]$ and*

<table>
<tr><th rowspan="3">Problem</th><th colspan="4">Approximation Ratio</th><th rowspan="3">Theorem(s)</th></tr>
<tr><th colspan="2">Upper Bound</th><th colspan="2">Lower Bound</th></tr>
<tr><th>Time</th><th>the bound</th><th>the bound</th><th>Assumptions</th></tr>
<tr><td>$\mathrm{TS}^{\{1\}}(1)$</td><td>$O(n^2|\mathcal{S}|)$</td><td>$1+\ln n$</td><td>$(1-\rho)\ln n$</td><td>NP$\neq$DTIME($n^{\log\log n}$)</td><td>2 and 7</td></tr>
<tr><td>$\mathrm{TS}^{\{1\},\{0,2\}}(1)$</td><td>$O(n^2|\mathcal{S}|)$</td><td>$1+\ln 2+\ln n$</td><td>$(1-\rho)\ln n$</td><td>NP$\neq$DTIME($n^{\log\log n}$)</td><td>2 and 7</td></tr>
<tr><td>$\mathrm{SB}^{\Sigma}(1)$</td><td>$O(n^3\ell^2)$</td><td>$1+\ln n$</td><td>$(1-\rho)\ln n$</td><td>NP$\neq$DTIME($n^{\log\log n}$)
$|\Sigma|>1$</td><td>2 and 7</td></tr>
<tr><td>$\mathrm{MCP}^{\Sigma}(r)$</td><td>$O(n^2|\mathcal{P}|+$
$LP)$</td><td>$[1+o(1)]\ln n$</td><td>$(1-\rho)\ln n$</td><td>NP$\neq$DTIME($n^{\log\log n}$)
$|\Sigma|>1$</td><td>2 and 7</td></tr>
<tr><td>$\mathrm{TS}^{\{1\}}(n^{\delta})$</td><td></td><td></td><td>n^{ρ}</td><td>NP$\neq$co-RP
$0<\rho<\delta<1$</td><td>12</td></tr>
<tr><td>$\mathrm{SB}^{\Sigma}(n^{\delta})$</td><td></td><td></td><td>n^{ρ}</td><td>NP$\neq$co-RP
$0<\rho<\delta<\frac{1}{2}$</td><td>12</td></tr>
</table>

Table 1: Summary of our approximability results: $(n,\mathcal{S})$ is an input instance of $\mathrm{TS}^{\Gamma}(k)$ and $\mathrm{SB}^{\Sigma}(k)$, $(n,\mathcal{S},\mathcal{P})$ is an input instance of $\mathrm{MCP}^{\Sigma}(r)$, ℓ is the maximum length of any sequence in $\mathcal{S}$, L is the total length of all sequences in $\mathcal{S}$ and ρ and δ are constants. The column *Assumptions* contains *sufficient* condition(s) for the respective lower bound.

set $S\in\mathcal{S}$ $\Pr[i\in S]=p$ *for some given fixed* $0<p<1$. *Then, an optimal solution* $\mathcal{T}$ *of this instance satisfies* $E[\mathcal{T}]\le c\log_2 n+o(1)$ *where the constant* c *is given by*

$$c\le\begin{cases}-4\log_2(1-2p(1-p)), & \text{if } \alpha \text{ is } \{1\}\\ -4\log_2(\min\{1-2p(1-p),2p(1-p)\}), & \text{if } \alpha \text{ is } \{1\},\{0,2\}\end{cases}$$

(b) *Let* $(n,\mathcal{S})$ *be a randomly generated instance of* $SB^{\Sigma}(1)$ *over a fixed alphabet* Σ *in which each string* $s=s_0s_1\ldots s_{\ell-1}\in\mathcal{S}$ *is of length* ℓ *generated independently randomly with* $\Pr[s_i=a]=1/|\Sigma|$ *for any* i *and any* $a\in\Sigma$. *Also assume that* ℓ *is sufficiently large compared to* n. *Let* $\vec{t}^*$ *be an optimal solution to this instance. Then,* $E(|\vec{t}^*|)=O(\log n)$ *and moreover for each sequence* x *in* $\vec{t}^*$ *also satisfies* $E(|x|)=O(\log\ell)$.

Theorem 1(b) implies that for randomly generated sequences the expected size of the barcode is within a constant factor of its optimal value of $\lceil\log_2 n\rceil$. Moreover, the length of each substring involved in the barcode is only logarithmic in the length of the given sequences, thereby lending some justification in [2] in using an upper bound on the length of the probes.

Techniques Used

(a) Our algorithm to achieve the tight approximation bound in Theorem 2 for $\mathrm{TS}^{\{1\}}(1)$, $\mathrm{TS}^{\{1\},\{0,2\}}(1)$ and $\mathrm{MCP}^{\Sigma}(r)$ is a greedy algorithm that selects tests based on *information content* defined in terms of the change in the partition of the universe when the test is applied. This notion is directly related to the Shannon information complexity [1, 14]. A careful analysis yields an upper bound on the approximation ratio that matches the lower bound in Theorem 7 within a small *additive term*. We believe the analysis will be useful in the context of analyzing other problems involving recursive partitioning of a given universe as well.

(b) The inapproximability results of Theorem 7 are proved by approximation preserving reductions from the set cover problem. To handle the barcode problems for $\Sigma = [0, 1]$ we introduce an artificial intermediate problem in which some tests are called cheap and they help very little in constructing a good set of test. This roughly corresponds to the fact that we cannot avoid tests that do not correspond to sets in the original set cover instance, but we can make them cheap.

(c) The inapproximability results in Theorem 12 are obtained by approximation preserving reductions from the graph coloring problem.

(d) The average case analysis in Theorem 1(b) uses the asymptotic analysis of the generating functions of probabilities of occurrence of a random pattern in a given text [6, 12].

Notational simplification: We will skip (1) in $\mathrm{TS}^{\{1\}}(1)$, $\mathrm{TS}^{\{1\},\{0,2\}}(1)$ and $\mathrm{SB}^{\{0,1\}}(1)$ and we will write "{1}-distinguishes" simply as "distinguishes" or "separates".

The Map. Proofs of most claims in Theorems 2 and 7 appear in Sections 3 and 4, respectively. The remaining proofs appear in the appendix.

3 Approximation Algorithms for Test Set and Minimum Cost Probe Problems

The Set Cover (SC) Problem is defined on an input instance $(U, \mathcal{S})$ such that $\mathcal{S} \subset \mathcal{P}(U)$ with the goal of finding a $\mathcal{C} \subseteq \mathcal{S}$ such that $\bigcup_{A \in \mathcal{C}} A = U$ and $|\mathcal{C}|$ is *minimized*. We can translate the $\mathrm{TS}^{\{1\}}$ problem to SC as follows. Given instance $(n, \mathcal{S})$ of $\mathrm{TS}^{\{1\}}$, we define instance $(U, \tau(\mathcal{S}))$ where $U = \{e \subset [0, n-1] : |e| = 2\}$, $\tau(T) = \{e \in U : |e \cap T| = 1\}$, and $\tau(\mathcal{S}') = \{\tau(T) : T \in \mathcal{S}\}$. The best proven approximation ratio for SC is achieved by a greedy heuristic [7] that, starting from the empty partial set cover, keeps adding new sets to the solution that maximize the number of elements that are not covered as yet. This heuristic for set cover runs in $O(\sum_{T \in \mathcal{S}} |\tau(T)|)$ time and has an approximation ratio of $1 + \ln(\max_{T \in \mathcal{S}} |\tau(T)|)$. Since $\max_{T \in \mathcal{S}} |\tau(T)| = |T|\,(n - |T|) \le \frac{n^2}{4}$, the above translation offers a $O(n^2|\mathcal{S}|)$ time greedy heuristic for $\mathrm{TS}^{\{1\}}$ with an approximation ratio of $(2 \ln n) - \ln 4$. A similar reduction for the $\mathrm{TS}^{\{1\},\{0,2\}}$ (resp. $\mathrm{MCP}^{\Sigma}(r)$) to the SC problem can also be given providing a greedy heuristic with an approximation ratio of $(2 \ln n) - \ln \frac{4}{3}$ (resp. $2 \ln n$). The main result of this section improves upon that simple heuristic as follows.

Theorem 2 *There is an* $O(n^2|\mathcal{S}|)$ *time approximation algorithm for* TS^{Γ} *with approximation ratio* $1 + \ln n$ *for* $\Gamma = \{\{1\}\}$ *and* $1 + \ln 2 + \ln n$ *for* $\Gamma = \{\{1\},]\{0, 2\}\}$. *There is an* $O(n^2|P| + L|P|)$

time approximation algorithm for MCP$^{\Sigma}(r)$ *with approximation ratio* $1+\ln n+\ln\log_2(r'+1)$, *where* $r' = \min\{r, n\}$ *and* L *is the total length of the sequences in* $\mathcal{S}$.

3.1 Proof of Theorem 2 for $\mathrm{TS}^{\{1\}}$

In this section we provide a greedy heuristic for $\mathrm{TS}^{\{1\}}$ running in time $O(n^2|\mathcal{S}|)$ time with an improved approximation ratio of $1+\ln n$. Notice that the upper bound almost matches the lower bound in Theorem 7 for $\mathrm{SB}^{\{0,1\}}$, a special case of $\mathrm{TS}^{\{1\}}$.

First, we consider the problem $\mathrm{TS}^{\{1\}}$. In the definition below and throughout the rest of this section we use $\mathcal{T}+T$ to denote $\mathcal{T}\cup\{T\}$,

Definition 3 *A set of tests* $\mathcal{T}\subset\mathcal{S}$ *defines the following:*

- *an equivalence relation* $\overset{\mathcal{T}}{\equiv}$ *on* $[0, n-1]$ *given by* $i \overset{\mathcal{T}}{\equiv} j$ *if and only if* $\forall T\in\mathcal{T}(i\in T\equiv j\in T)$,
- *a set of permutations* $\Pi_{\mathcal{T}} = \{\pi \in$ (*permutations of* $[0, n-1]$) $: \forall i\in[0, n-1]\ i \overset{\mathcal{T}}{\equiv} \pi(i)\}$,
- *entropy* $H_{\mathcal{T}} = \log_2|\Pi_{\mathcal{T}}|$.
- *information content of a* $T\in\mathcal{S}$ *with respect to* $\mathcal{T}$, $IC(T,\mathcal{T}) = H_{\mathcal{T}} - H_{\mathcal{T}+T} = \log_2\frac{|\Pi_{\mathcal{T}}|}{|\Pi_{\mathcal{T}+T}|}$.

The *information content heuristic* (ICH for short) is the following simple greedy heuristic:

> $\mathcal{T} = \emptyset$
> **while** $H_{\mathcal{T}} \neq 0$ **do**
>   select a $T\in\mathcal{S}-\mathcal{T}$ that *maximizes* $IC(T,\mathcal{T})$
> $\mathcal{T} = \mathcal{T}+T$
> **endwhile**

The correctness of ICH follows from the fact that $H_{\mathcal{T}} = 0$ implies the equivalence classes of $\overset{\mathcal{T}}{\equiv}$ are n singleton sets $\{0\},\{1\},\ldots,\{n-1\}$ and the fact that if $H_{\mathcal{T}}\neq 0$ then there exists a $T\in\mathcal{S}-\mathcal{T}$ with $IC(T,\mathcal{T}) > 0$ (otherwise our problem instance has no feasible solution).

To implement ICH, we iteratively maintain the equivalence classes of $\overset{\mathcal{T}}{\equiv}$ as sorted lists. We also precompute and store $\log_2(i!)$ for each $i\in[1,n]$. Given a specific $T\in\mathcal{S}-\mathcal{T}$, it is easy to compute in $O(n)$ time the equivalence classes of $\overset{\mathcal{T}+T}{\equiv}$ from the equivalence classes of $\overset{\mathcal{T}}{\equiv}$ since an equivalence class E of $\overset{\mathcal{T}}{\equiv}$ is either an equivalence class of $\overset{\mathcal{T}+T}{\equiv}$ or it is partitioned into two equivalence classes $E_1 = E\cap T$ and $E_2 = E - E_1$ of $\overset{\mathcal{T}+T}{\equiv}$; the first case contributes nothing to $IC(T,\mathcal{T})$ while the second case adds $\log_2\binom{|E|}{|E_1|}$ to $IC(T,\mathcal{T})$. Finally, notice that the while loop is executed at most n times.

Now we analyze the approximation ratio of ICH. We will use the convention $x = |X|$ for a set X.

Lemma 4 *If* $\mathcal{T}_0\subset\mathcal{T}_1$ *then* $IC(T,\mathcal{T}_0)\geq IC(T,\mathcal{T}_1)$.

Proof. By induction, it suffices to consider a case when for some test S we have $\mathcal{T}_1 = \mathcal{T}_0 + S$. In this case our claim is

$$H_{\mathcal{T}_0} - H_{\mathcal{T}_0+T} \geq H_{\mathcal{T}_0+S} - H_{\mathcal{T}_0+S+T} \quad \equiv \frac{|\Pi_{\mathcal{T}_0}|}{|\Pi_{\mathcal{T}_0+T}|} \geq \frac{|\Pi_{\mathcal{T}_0+S}|}{|\Pi_{\mathcal{T}_0+S+T}|}.$$

Assume that $E_0, \ldots, E_k$ are the equivalence classes of $\overset{\mathcal{T}_0}{\equiv}$, and let $A_i = E_i \cap S \cap T$, $B_i = E_i \cap S - T$, $C_i = E_i \cap T - S$, $D_i = E_i - S - T$. Observe that

$$\begin{aligned} |\Pi_{\mathcal{T}_0}| &= \prod_{i=0}^{k} (a_i + b_i + c_i + d_i)! & \qquad |\Pi_{\mathcal{T}_0+T}| &= \prod_{i=0}^{k} (a_i + c_i)!(b_i + d_i)! \\ |\Pi_{\mathcal{T}_0+S}| &= \prod_{i=0}^{k} (a_i + b_i)!(c_i + d_i)! & \qquad |\Pi_{\mathcal{T}_0+S+T}| &= \prod_{i=0}^{k} (a_i!)(c_i!)(b_i!)(d_i!) \end{aligned}$$

and thus it suffices to show that

$$\begin{aligned} \prod_{i=0}^{k} \frac{(a_i + b_i + c_i + d_i)!}{(a_i + c_i)!(b_i + d_i)!} &\geq \prod_{i=0}^{k} \frac{(a_i + b_i)!(c_i + d_i)!}{a_i! c_i! b_i! d_i!} \qquad \equiv \\ \prod_{i=0}^{k} \binom{a_i + b_i + c_i + d_i}{a_i + c_i} &\geq \prod_{i=0}^{k} \binom{a_i + b_i}{a_i} \binom{c_i + d_i}{c_i} \end{aligned}$$

It is easy to see that each term $l_i = \binom{a_i+b_i+c_i+d_i}{a_i+c_i}$ on the last left-hand-side is at least as large as the corresponding term $r_i = \binom{a_i+b_i}{a_i}\binom{c_i+d_i}{c_i}$ on the right-hand-side because l_i counts subsets of E_i with $a_i + b_i$ elements, while r_i counts the subsets of E_i that have a_i elements in S and b_i elements in $E_i - S$. ❑

Lemma 5 $IC(\emptyset, T) < n$ *for every test* T.

Proof. $IC(\emptyset, T) = \log_2 \frac{|\Pi_\emptyset|}{\Pi_{\{T\}}} = \log_2 \frac{n!}{t!(n-t)!} = \log_2 \binom{n}{t} < \log_2 2^n = n.$ ❑

Lemma 6 *If* $IC(\mathcal{T}, T) > 0$ *then* $IC(\mathcal{T}, T) \geq 1$.

Proof. The claim is equivalent to the obvious implication: if $|\Pi_{\mathcal{T}}| > 1$ then $|\Pi_{\mathcal{T}}| \geq 2$. ❑

Now we are ready for an amortized analysis of ICH. Suppose that an optimum solution of $(n, \mathcal{S})$ is $\mathcal{T}^* = \{T_1^*, \ldots, T_k^*\}$. During the execution of ICH, for a current partial test set $\mathcal{T}$, let $\mathcal{T}_i = \mathcal{T} + T_1^* + \cdots + T_i^*$ (accordingly, $\mathcal{T}_0 = \mathcal{T}$) and $h_i = IC(\mathcal{T}_{i-1}, T_i^*)$. Notice that $\sum_{i=1}^{k} h_i = \sum_{i=1}^{k} (H_{\mathcal{T}_{i-1}} - H_{\mathcal{T}_{i-1}+T_i^*}) = H_{\mathcal{T}} - H_{\mathcal{T}+\mathcal{T}^*} = H_{\mathcal{T}}$, since $H_{\mathcal{T}+\mathcal{T}^*} = 0$. Let $h_i^* < n$ denote the initial value of h_i *i.e.* the value of h_i with $\mathcal{T} = \emptyset$.

During the j^{th} iteration of the while loop, ICH selects a test T (with, say, $IC(\mathcal{T}, T) = \Delta_j$) and changes $\mathcal{T}$ into $\mathcal{T} + T$. As a result, $H_{\mathcal{T}}$ drops by Δ_j and h_i drops by some $\delta_{i,j}$ with $\sum_{i=1}^{k} \delta_{i,j} = \Delta_j$. This iteration adds 1 to the solution cost. We distribute this cost among the elements of $\mathcal{T}^*$ by charging T_i^* with $\delta_{i,j}/\Delta_j$. Because $h_i = IC(\mathcal{T}_{i-1}, T_i^*) \le IC(\mathcal{T}, T_i^*)$, we know that $\Delta_j \ge h_i$ since otherwise ICH would select T_i^* rather then T. Therefore reducing the current h_i by $\delta_{i,j}$ is associated with a charge that is at most $\delta_{i,j}/h_i$. Let $m(h)$ be the supremum of possible sums of charges that some T_i^* may receive starting from the time when $h_i = h$. By induction on the number of such positive charges we will show that $m(h) \le 1 + \ln h$. If this number is 1, then $h > 0$ and hence $\ln h \ge 0$ (by Lemma 6), while the charge is at most 1. In the inductive step, we consider a situation when T_i^* starts with $h_i = h$, receives a single charge δ/h, h_i is reduced to $h - \delta$ and afterwards, by inductive assumption, T_i^* receives at most $m(h-\delta)$ charges. Because $h - \delta > 0$ we know by Lemma 6 that $h - \delta \ge 1$. Therefore

$$m(h) \le m(h-\delta) + \frac{\delta}{h} \le 1 + \ln(h-\delta) + \frac{\delta}{h} < 1 + \int_1^{h-\delta} \frac{dx}{x} + \int_{h-\delta}^{h} \frac{dx}{x} = 1 + \int_1^h \frac{dx}{x} = 1 + \ln h.$$

By Lemma 5, $h < n$. This proves our claim on the approximation ratio for $\mathrm{TS}^{\{1\}}$.

4 Inapproximability Results for Test Set, String Barcoding and Minimum Cost Probe Set Problems

The NP-hardness of $\mathrm{TS}^{\{1\}}$ follows from the NP-hardness of the minimum test collection problem in [5] from a reduction from the 3-dimensional matching problem and minor modifications of this reduction can be used to prove the NP-hardness of $\mathrm{TS}^{\{1\},\{0,2\}}$ as well. NP-hardness of $\mathrm{MCP}^{\Sigma}(r)$ from the vertex cover problem was mentioned without a proof in [2]. Our goal is to show that it is in fact impossible (under reasonable complexity theoretic assumptions) to approximate these problems any better than mentioned in Theorem 2.

Theorem 7 *For any given constant $0 < \rho < 1$, it is impossible to approximate $SB^{\{0,1\}}$ (a restricted case of $TS^{\{1\}}$), $TS^{\{1\},\{0,2\}}$ or $MCP^{\{0,1\}}(r)$ within a factor of $(1-\rho)\ln n$ in polynomial time unless NP=DTIME($n^{\log\log n}$).*

We can restrict instances of $\mathrm{TS}^{\{1\}\{0,2\}}$ to those such $(n, \mathcal{S})$ that $[0, n-1] \in \mathcal{S}$. If $\mathcal{T}$ is a solution, it remains a solution when we view $(n, \mathcal{S})$ as an instance of $\mathrm{TS}^{\{1\}}$; conversely, if $\mathcal{T}$ is a solution of $(n, \mathcal{S})$ as an instance of $\mathrm{TS}^{\{1\}}$ then $\mathcal{T} + [0, n-1]$ is a solution of $(n, \mathcal{S})$ as an instance of $\mathrm{TS}^{\{1\}\{0,2\}}$. Therefore lower bounds for $\mathrm{TS}^{\{1\}}$ apply to $\mathrm{TS}^{\{1\}\{0,2\}}$.

4.1 Test Set with Order

To make the approximation preserving reduction from Set Cover to $\mathrm{SB}^{\{0,1\}}$ easier to follow, we introduce an intermediate problem called *Test Set with Order* with parameter $k \in \mathbb{N}$ (denoted by TSO^k):

Instance: $(n, k, \mathcal{S})$ where k is a positive integer, $(n, \mathcal{S})$ is an instance of $TS^{\{1\}}$ and $\mathcal{S}$ includes the family of "cheap" sets $\mathcal{S}_0 = \{\{i\} | \; i \in [0, n-1]\} \cup \{[0, i] \; | \; i \in [0, n-1]\}$.

Valid solution: a solution for $(n, \mathcal{S})$.

Objective: minimize $\mathsf{cost}(\mathcal{T}) = |\mathcal{T} - \mathcal{S}_0| + \frac{1}{k}|\mathcal{T} \cap \mathcal{S}_0|$.

Note that TSO^1 is in fact a *special case* of $TS^{\{1\}}$; hence any hardness results proved for TSO^1 would apply to $TS^{\{1\}}$ as well. Our claim follows once the following theorem is proved.

Theorem 8 *For any integer constant $k > 0$ and any constant $0 < \rho < 1$, it is impossible to approximate TSO^k within a factor of $(1 - \rho) \ln n$ in polynomial time unless NP$\subset$DTIME$(n^{\log \log n})$.*

In the rest of this section, we prove the above theorem. We need the following straightforward extension of the hardness result in [3] for a slightly restricted version of SC.

Fact 9 *Assuming NP$\not\subset$DTIME$(n^{\log \log n})$, instances of the SC problem for which the optimal cover requires at least $(\log_2 n)^2$ sets cannot be approximated to within a factor of $(1 - \varepsilon') \ln n$ for any constant $\varepsilon' > 0$ in polynomial time.*

For notational simplicity, assume that kn is an exact power of 2 and $\ell = \log_2(kn)$. The following lemma gives a reduction from SC to TSO^k problem.

Lemma 10 *There exists a polynomial-time computable function τ that maps an instance $(n, \mathcal{S})$ of SC into instance $(2kn, k, \tau(\mathcal{S}))$ of TSO^k such that optimal solutions of $(n, \mathcal{S})$ and $(2kn, k, \tau(\mathcal{S}))$, C^* and $\mathcal{T}^*$ respectively, satisfy the following:*

$$|\mathcal{C}^*| \leq \mathsf{cost}(\mathcal{T}^*) \leq |\mathcal{C}^*| + \ell.$$

Moreover, given any solution X of $(2kn, k, \tau(\mathcal{S}))$, we can in polynomial time construct a solution Y of $(n, \mathcal{S})$ such that $|Y| \leq \mathsf{cost}(X)$.

Proof. We use the notations $a \times A = \{ai | \; i \in A\}$ and $A + B = \{a + b | \; a \in A \,\&\, b \in B\}$. $\tau(\mathcal{S})$ contains the following sets:

$$\begin{array}{rl} \text{cover sets:} & D(A) = 2 \times (k \times S + [0, k-1]) \text{ for } S \in \mathcal{S}; \\ \text{cheap sets:} & \{i\} \text{ and } [0, i] \text{ for each } i \in [0, 2kn - 1]; \\ \text{other sets:} & A_i = \{j \in [0, 2kn-1] | \; j \bmod 2^{i+1} \geq 2^i\} \text{ for } i \in [1, \ell]. \end{array}$$

First, we show that $\mathsf{cost}(\mathcal{T}^*) \leq |\mathcal{C}^*| + \ell$. Given a set cover $\mathcal{C}$ of $(n, \mathcal{S})$ we define the following test set that is a solution of $(2kn, \tau(\mathcal{S}))$: $\mathcal{T} = \{D(A) | \; A \in \mathcal{C}\} \cup \{A_i | \; i \in [1, \ell]\}$. To see that $\mathcal{T}$ is indeed a valid solution, consider $i, j \in [0, 2kn - 1]$. Suppose that i is even and j is not. Then for some $A \in \mathcal{C}$ and $a \in 2 \times [0, k-1]$ we have $(i - 2a)/2k \in A$, and thus $i \in D(A)$ while $j \notin D(A)$. On the other hand, if that i and j have the same parity then they differ on k^{th} bit for some $k \in [1, \ell]$, in which case i and j are distinguished by test A_k. Hence, $\mathsf{cost}(\mathcal{T}^*) = |\mathcal{T}^*| \leq |\mathcal{C}^*| + \ell$.

Next, we show that $|\mathcal{C}^*| \leq \mathtt{cost}(\mathcal{T}^*)$. Given a set of tests $\mathcal{T}$, consider the *partial cover* $\mathcal{C}' = \{A |\ D(A) \in \mathcal{T}\}$, and let $C = \bigcup_{S \in \mathcal{C}'} S$. Consider $i \in [0, n-1] - C$. For $a \in [0, k-1]$ we know that some set of $\mathcal{T}$ distinguished $2ki - 2a$ from $2ki - 2a + 1$. This distinguishing set can only be one of the three sets: $\{2ki - 2a\}$, $\{2ki - 2a + 1\}$ or $[0, 2ki - 2a]$. Note that for each $i \in [0, n-1] - C$ and each $a \in [0, k-1]$ we have a choice of different three sets, so in each such case we use a different element of $\mathcal{T}$. We can conclude that $\mathcal{T}$ contains $k(n - |C|)$ such sets, and thus $\mathtt{cost}(\mathcal{T}) \geq |\mathcal{C}'| + n - |C|$. Since for each $i \in [0, n-1]$ $\mathcal{T}$ must distinguish $2i - 1$ from $2i$, $\mathcal{T}$ must contain one of these three sets: $\{2i-1\}, \{2i\}, [0, 2i-1]$. Note that each $i \in [0, n-1] - C$ has different possibilities, thus for each of them $\mathcal{T}$ contains a different set of choices. We can therefore extend $\mathcal{C}'$ to a cover $\mathcal{C}$ of $(n, \mathcal{S})$ by adding at most $n - |C|$ sets. Hence $|\mathcal{C}| \leq \mathtt{cost}(\mathcal{T})$.

Hence, $\mathtt{cost}(\mathcal{T}^*) \leq |\mathcal{C}^*| + \ell + \frac{1}{k}$. ❑

We can now complete the proof of Theorem 8. Consider an instance of SC as mentioned in Fact 9, transform it to an instance of TSO^k as described in Lemma 10 and let $\mathcal{C}^*$ and $\mathcal{T}^*$ be optimal solutions to the instances of SC and TSO^k, respectively. Suppose that we can approximate TSO^k within a factor of $(1-\rho)\ln n$ and let T' be such an approximate solution. Then, by using Lemma 10 we can find a solution C' to the instance of SC such that

$$\begin{aligned} |C'| &\leq \mathtt{cost}(T') \\ &\leq (1-\rho)\ \ln n\, \mathtt{cost}(\mathcal{T}^*) \\ &\leq (1-\rho)\ \ln n\ (|\mathcal{C}^*| + \ell + 1) \\ &\leq (1-\rho + o(1))\ \ln n\, |\mathcal{C}^*| \qquad \text{since } |\mathcal{C}^*| = \Omega(\ell^2) \text{ and } \ell = \Omega(\log n) \end{aligned}$$

which violates Fact 9 by choosing $\varepsilon' = 1 - \rho + o(1)$.

4.2 Proof of Theorem 7 for $\text{SB}^{\{0,1\}}$

As before, for notational simplicity, assume that kn is an exact power of 2 and $\ell = \log_2(kn)$. First, using the reduction described in the proof of Lemma 10, we provide a reduction of SC to $\text{SB}^{\{0,1\}}$.

Lemma 11 *For any given constant integer $k > 0$, there exists a polynomial-time computable function σ that maps an instance $(n, \mathcal{S})$ of SC into an instance $(2kn, \sigma(\mathcal{S}))$ of $\text{SB}^{\{0,1\}}$, so that if $\mathcal{C}^*$ and $\vec{\mathbf{t}}^*$ are the optimal solutions for $(n, \mathcal{S})$ and $(2kn, \sigma(\mathcal{S}))$, respectively, then*

$$\frac{|\mathcal{C}^*|}{1 + \frac{1}{k}} \leq |\vec{\mathbf{t}}^*| \leq |\mathcal{C}^*| + \ell.$$

Moreover, given any solution $\vec{\mathbf{x}}$ of $(2kn, \sigma(\mathcal{S}))$, we can in polynomial time construct a solution Y of $(n, \mathcal{S})$ such that $\frac{|Y|}{1+\frac{1}{k}} \leq |\vec{\mathbf{x}}|$.

Proof. First, we define a family $\tau(\mathcal{S})$ of subsets of $[0, 2kn - 1]$ using the function τ from Lemma 10. Let $\mathcal{S}_0$ be the family of "special" or "cheap" test sets, and $\mathcal{S}_1 = \tau(\mathcal{S}) - \mathcal{S}_0$. We number the elements of $\mathcal{S}_1$, so $\mathcal{S}_1 = \{B_0, \ldots, B_{m-1}\}$ and let $B_m = [0, 2kn-1] \in \mathcal{S}_0$. For each $i \in [0, 2kn-1]$ we define sequence s_i as a concatenation of alternating groups of 0^{i+1} and a distinct member from the set $\{1^{k+1} \mid i \in B_k\}$, begining and ending with 0^{i+1}. This completes the description of the function σ.

Consider any set cover $\mathcal{C}$ of $(n, \mathcal{S})$. As noted in the proof of Lemma 10, we can map it into a solution for TSO^k without using any cheap tests and with at most $|\mathcal{C}^*| + \ell$ test sets. Then, we replace test B_j with a test sequence $01^{j+1}0$. Thus $|\vec{t}^*| \le |\mathcal{C}^*| + \ell$.

Now consider a solution vector of sequences $\vec{t}$ for $\sigma(\mathcal{S})$. We show how to replace each sequence t of $\vec{t}$ with at most two sets such that the following two statements hold:

(a) if $(t \prec s_p) \neq (t \prec s_q)$ for two sequences s_p and s_q, then the replaced sets $\{1\}$-distinguish p from q;

(b) when we use two sets, one of them is cheap.

By **(a)**, the replacement sets form a solution for the instance $(2kn, k, \tau(\mathcal{S}))$ of TSO^k. By **(b)**, the cost of the this solution for $(2kn, k, \tau(\mathcal{S}))$ is at most $\left(1 + \frac{1}{k}\right) |\vec{t}|$. Finally, by Lemma 10, it is possible to construct from this solution for $(2kn, k, \tau(\mathcal{S}))$ a solution for the set cover instance $(n, \mathcal{S})$ with no more than $\left(1 + \frac{1}{k}\right) |\vec{t}|$ sets. Hence, it only remains to show the replacement. We have the following cases:

Case 1: t contains a substring 10^a1 for some $a > 0$. Then t can be a substring of only s_{a-1}, so we can replace t with a cheap test $\{a-1\}$.

Case 2: Otherwise, t is of the form $0^*1^*0^*$.

Case 2.1: $t = 0^a$ for some $a > 0$. Then t is a substring of all s_i's with $i \ge a-1$, and therefore we can replace it with a cheap test $[0, i-2]$.

Case 2.2: $t = 0^a1^b$ for some $a, b > 0$. If $b > m+1$, t is not a substring of any s_i, so we can discard it. If $b \le m+1$, then this test is equivalent to 0^a because every s_i contains 1^{m+1}.

Case 2.3: $t = 1^a0^b$ for some $a, b > 0$. Similar to Case 2.2.

Case 2.4: $t = 0^a1^b0^c$ where $a, b, c > 0$. Let $d = \max\{a, c\}$; one can see that we can replace t with B_{b-1} and $[0, d-2]$. ❑

We can now complete the proof of our claim in a manner similar to that in the proof of Lemma 10. Consider an instance of SC as mentioned in Fact 9, transform it to an instance of SB$^{\{0,1\}}$ as described in Lemma 11 and let $\mathcal{C}^*$ and $\vec{t}^*$ be optimal solutions to the instances of SC and SB$^{\{0,1\}}$, respectively. Suppose that we can approximate SB$^{\{0,1\}}$ within a factor of $(1-\rho)\ln n$ and let $\vec{t}'$ be such an approximate solution. Then, by using Lemma 11 we can find a solution C' to the instance of SC such that

$$\begin{aligned}
|C'| &\le \left(1 + \tfrac{1}{k}\right) \mathsf{cost}(\vec{t}') \\
&\le \left(1 + \tfrac{1}{k}\right) (1-\rho) \ln n \, \mathsf{cost}(\vec{t}^*) \\
&\le \left(1 + \tfrac{1}{k}\right) (1-\rho) \ln n \, (|\mathcal{C}^*| + \ell + 1) \\
&\le (1 - \rho + o(1)) \ln n \, |\mathcal{C}^*| \qquad \text{since } |\mathcal{C}^*| = \Omega(\ell^2) \text{ and } \ell = \Omega(\log n)
\end{aligned}$$

which violates Fact 9 by choosing $\varepsilon' = 1 - \rho + o(1)$.

5 Stronger Inapproximability Results for $\mathbf{TS}^{\{1\}}(k)$, $\mathbf{TS}^{\{1\},\{0,2\}}(k)$ and $\mathbf{SB}^{\{0,1\}}(k)$

Theorem 12 *For any two given constants $0 < \rho < \delta < 1$, $TS^{\{1\}}(n^{\delta})$ and $TS^{\{1\},\{0,2\}}(n^{\delta})$ cannot be approximated to within a factor of n^{ρ} in polynomial time unless co-RP=NP. The same result holds for $SB^{\Sigma}(n^{\delta})$ also provided $0 < \rho < \delta < \frac{1}{2}$.*

References

[1] Y. S. Abu-Mostafa (editor). *Complexity in Information Theory*, Springer Verlag, 1986.

[2] J. Borneman, M. Chrobak, G. D. Vedova, A. Figueroa and T. Jiang. *Probe Selection Algorithms with Applications in the Analysis of Microbial Communities*, Bioinformatics, Vol. 17 Suppl. 1, 2001, pp. S39-S48.

[3] U. Feige. *A threshold for approximating set cover*, Journal of the ACM, Vol. 45, 1998, pp. 634-652.

[4] U. Feige and J. Kilian. *Zero knowledge and the chromatic number*, Journal of Computer and System Sciences, Vol. 57, No. 2, October 1998, pp. 187-199.

[5] M. R. Garey and D. S. Johnson. *Computers and Intractability - A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co. 1979.

[6] L. J. Guibas and A. M. Odlyzko. *String overlaps, pattern matching, and nontransitive games*, Journal of Combinatorial Theory Series A, Vol. 30, 1981, pp. 183-208.

[7] D. S. Johnson. *Approximation Algorithms for Combinatorial Problems*, Journal of Computer and Systems Sciences, Vol. 9, 1974, pp. 256-278.

[8] R. M. Karp, R. Stoughton and K. Y. Yeung, *Algorithms for Choosing Differential Gene Expression Experiments*, Proc. Third Annual International Conference on Computational Molecular Biology, 1999, pp. 208-217.

[9] D. Gusfield. *Algorithms on Strings, Trees and Sequences*, Cambridge University Press, 1997.

[10] L. Lovasz. *On the Ratio of Optimal Integral and Fractional Covers*, Discrete Mathematics, Vol. 13, 1975, pp. 383-390.

[11] B. Monien and E. Speckenmeyer. *Some further approximation algorithms for the vertex cover problem*, Proceedings of CAAP 83, Lecture Notes in Computer Science Vol. 159, Springer-Verlag, 1983.

[12] A. M. Odlyzko. *Asymptotic enumeration methods*, in Handbook of Combinatorics, Vol. II, R. L. Graham, M. Grötschel and L. Lovász (editors), The MIT Press, 1995, pp. 1063-1230.

[13] S. Rash and D. Gusfield. *String Barcoding: Uncovering Optimal Virus Signatures*, Proc. Sixth Annual International Conference on Computational Molecular Biology, 2002, pp. 254-261.

[14] C. E. Shannon. *Mathematical Theory of Communication*, Bell Systems Technical Journal, Vol. 27, 1948, pp. 379-423, pp. 623-658.

Appendix

A Proof of Theorem 2 for $\mathrm{TS}^{\{1\},\{0,2\}}$

We show how to reduce an instance of $\mathrm{TS}^{\{1\},\{0,2\}}$ to an instance of $\mathrm{TS}^{\{1\}}$ at the expense of increasing the size of the universe by a factor of 2. Such a reduction will obviously prove our claim on the approximation ratio for this problem. Given an instance $(n, \mathcal{S})$ of $\mathrm{TS}^{\{1\},\{0,2\}}$, we create instance $(2n, \sigma(\mathcal{S}))$ of $\mathrm{TS}^{\{1\}}$: $n + T = \{n + i : \ i \in T\}$, $\sigma(T) = T \cup n + ([0, n-1] - T)$, and $\sigma(\mathcal{S}) = \{\sigma(T) : \ T \in \mathcal{S}\}$. Thus, for all $i, j \in [0, n-1]$ with $i \neq j$, the following are true:

$$|\{i,j\} \cap T| = 1 \ \equiv \ |\{i,j\} \cap \sigma(T)| = 1 \ \equiv \ |\{i+n, j+n\} \cap \sigma(T)| = 1$$

$$|\{i,j\} \cap T| \in \{0,2\} \ \equiv \ |\{i, j+n\} \cap \sigma(T)| = 1 \ \equiv \ |\{i+n, j\} \cap \sigma(T)| = 1.$$

Hence a set of tests $T_1, T_2, \ldots, T_k$ is a solution of the instance $(n, \mathcal{S})$ of $\mathrm{TS}^{\{1\},\{0,2\}}$ if and only of the set of tests $\sigma(T_1), \sigma(T_2), \ldots, \sigma(T_k)$ is a solution of the instance $(2n, \sigma(\mathcal{S}))$ of $\mathrm{TS}^{\{1\}}$.

B Proof of Theorem 2 for $\mathrm{MCP}^{\Sigma}(r)$

We use the same greedy algorithm as described in the Section 3. A test set corresponding to a possible choice of a string t in the barcode now partitions the set of sequences $\mathcal{S}$ into at most $r+1$ partitions $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_r$ where $\mathcal{S}_i = \{s \in \mathcal{S} \mid i = \min\{r, \mathrm{oc}(t,s)\}\}$. Using linear time algorithm for finding a pattern in a text (e.g., see [9, page 10]), all the test sets corresponding to the $O(n\ell^2)$ sequences can be constructed in $O(L|\mathcal{P}|)$ time. Hence an overall running time of $O((n^2 + L)|\mathcal{P}|)$ follows.

The analysis of the algorithms is very similar to that in the previous section, so we just point out the differences:

(1) In Lemma 5, $\mathrm{IC}(\emptyset, T) \leq n \log_2(r+1)$ for every test T. This is because $H_{\emptyset} = \log_2 n!$ and $H_{\{T\}} \leq \log_2 \left(\left(\frac{n}{r+1}\right)!\right)^{r+1}$, hence $\mathrm{IC}(\emptyset, T) = \log_2 \frac{n!}{\left(\left(\frac{n}{r+1}\right)!\right)^{r+1}} \leq n \log_2(r+1)$.

(2) A test set that partitions $\mathcal{S}$ into $r+1$ partitions can be thought of as a group r tests each of which partitions $\mathcal{S}$ into two partitions. As a result, the amortized analysis of the previous section applies if we charge each such group of tests in the analysis since Lemma 4 holds for the case when $\mathcal{T}_0$ and $\mathcal{T}_1$ are two such groups of tests.

Note also that even if $r > n$ no test may have information content larger then the initial entropy, *i.e.* larger then $n \log_2 n$. Hence, we have approximation ratio of $1 + \ln n + \ln \log_2 \min\{r+1, n\}$.

C Proof of Theorem 7 for $\mathrm{MCP}^{\{0,1\}}(r)$

We reduce $\mathrm{SB}^{\{0,1\}}$ to this problem. Given an instance $(n, \mathcal{S})$ of $\mathrm{SB}^{\{0,1\}}$, we create an instance $(n, \mathcal{S}')$ of $\mathrm{MCP}^{\{0,1\}}(r)$ where $\mathcal{S}' = \{s^r | s \in \mathcal{S}\}$. Notice that $t \prec s = \frac{1}{r}\min\{r, \mathsf{oc}(t, s^r)\}$ for any two sequences s and t. Hence, there is a 1-1 correspondence between solutions of $\mathrm{SB}^{\{0,1\}}$ and $\mathrm{MCP}^{\{0,1\}}(r)$.

D Proof of Theorem 12

We first prove our claims for $\mathrm{TS}^{\{1\}}(n^\delta)$ and $\mathrm{TS}^{\{1\},\{0,2\}}(n^\delta)$. We give reductions from the *graph coloring* problem whose goal is to produce an assignment of colors to vertices of a given graph $G = (V, E)$ such that no two vertices have the same color and the number of colors is *minimized*. The following strong inapproximability result is a straightforward extension of a hardness result known for graph coloring [4]: for any constant $0 < \rho < \delta < 1$, the graph coloring problem cannot be approximated to within a factor of $|V|^\rho$ even if the maximum size of the independent set in the graph is at most $|V|^\delta$ unless co-RP=NP.

Consider $\mathrm{TS}^\gamma(k)$ (for $\gamma \in \{\{1\}, \{\{1\}, \{0, 2\}\}\}$). Given a graph $G = (V, E)$ of n vertices with $V = \{v_0, v_1, \ldots, v_{n-1}\}$, we create an instance $(2n, \mathcal{S}, n)$ of $\mathrm{TS}^\gamma(n)$ such that, for each v_i, $\mathcal{S}$ contains the test $T_i = \{2i\} \cup \{\ 2j, 2j+1 \mid \{v_i, v_j\} \in E\ \}$. Moreover, $\mathcal{S}$ contains an additional $\log_2 n$ test sets $A_i = \{j \in [0, 2n-1] \mid j \bmod 2^{i+1} \geq 2^i\}$ for $i \in [1, \log_2 n]$ (for simplicity, we assume that n is a power of 2) and one additional test set $[0, 2n-1]$. We now show that G can be colored with χ colors if and only if $\mathrm{TS}^\gamma(n)$ has a n-test cover of size $\chi + 1$:

(a) Suppose that G can be colored with χ colors *i.e.* there exists a function $f : V \to [0, \chi-1]$ such that $f(v_i) \neq f(v_j)$ if $(v_i, v_j) \in E$. For each $i \in [0, \chi-1]$, $\tau_i = \cup_{f(v)=i} T_v$ is a set of n-tests. Moreover, the set of $\chi+1$ n-test sets $\tau_1, \tau_2, \ldots, \tau_{\chi-1}$ and $A = \cup_{i=1}^{\log_2 n} A_i$ $\{1\}$-distinguish every pair of distinct elements in $[0, 2n-1]$ since (1) the pair $\{2i, 2i+1\}$ is $\{1\}$-distinguished by τ_x with $f(v_i) = x$ and (2) for $i \neq j$, both $2i$ and $2i+1$ are $\{1\}$-distinguished from $2j$ and $2j+1$ by some test in A. Finally, inclusion of the test set $[0, 2n-1]$ will ensure that all pairs of distinct elements in $[0, 2n-1]$ are $\{0, 2\}$-distinguished.

(b) Conversely, suppose that $\mathcal{T}$ is a set of n-tests that, for each γ, γ-distinguishes every pair of elements in $[0, 2n-1]$. We first add the n-test set A to $\mathcal{T}$ if it is not already there in some other n-test. Let $\mathcal{T}'$ be the n-tests in $\mathcal{T}$ excluding any 2-tests that contains *only one or both* of A and $[0, 2n-1]$. Now, for each n-test set $T \in \mathcal{T}'$, if $T - (A \cup [0, 2n-1])$ contains both T_i and T_j where $\{i, j\} \in E$, then $T_i \cup T_j$ $\{1\}$-distinguish only those pairs that are $\{1\}$-distinguished by some members of A; hence we can remove every such pair from T. After all the removals for each $T \in \mathcal{T}$ $T - (A \cup [0, 2n-1])$ consists of a set of tests $T_{i_1}, T_{i_2}, \ldots, T_{i_p}$ such that $\{i_x, i_y\} \notin E$ for $x, y \in [1, p]$. Moreover, for each $v_\ell \in V$, T_ℓ occurs in some test in $\mathcal{T}'$ since T_ℓ is the only test that $\{1\}$-distinguishes the pair $\{2\ell, 2\ell+1\}$. Let $\tau_1, \tau_2, \ldots, \tau_{|\mathcal{T}'|-1}$ be the set of n-tests in $\mathcal{T}'$. Our color assignment function f is given by $f(v_i) = j$ if $T_i \in \tau_j$.

We next prove our claim for $\mathrm{SB}^\Sigma(n^\delta)$. Consider an instance $([0, n-1], E)$ of graph coloring, We may assume that no independent set in this graph has more then $n/2$ nodes (because we can find such sets [11]) and that n is even. we create an instance $(n^2, \mathcal{S})$ of $\mathrm{TSO}^n(n/2)$. The *cheap sets* and *other sets* for this instance are defined as in the proof of Lemma 10. Moreover, for a node $u \in [0, n-1]$ we create set

$$S_u = un + 2 \times [0, n/2 - 1] \cup \bigcup_{\{u,v\} \in E} vn + [0, n-1].$$

Suppose that we have a coloring $\mathcal{C}$. Then we can define a set of $(n/2)$-tests consisting of all the *other sets* (there are $2\lceil \log_2 n \rceil$ of them) and for each $C \in \mathcal{C}$ having an $(n/2)$-test $\bigcup_{u \in C} S_u$.

Conversely, given a valid set of N $(n/2)$-tests we will show a coloring with at most $2N$ colors. First, suppose that a $(n/2)$-test T contains $S_u \cup S_v$ for some $\{u, v\} \in E$, one can see that after removing both S_u and S_v from the union defining T, T still separates the same set of pair $\{2i, 2i+1\}$ as before. Therefore we may assume that the set of S_u's used in the definition of T is independent. We use every such set as a color in our coloring. Consider now a node w that is not covered by any of these colors. Then the pairs $un + 2i + \{0, 1\}$, $i = 1, \ldots, n/2$ have to be separated by cheap sets used to define our $(n/2)$-tests. If we have j such nodes, we had to use at least $nj/2$ cheap sets, thus $nj/2 \leq nN/2$, so $j \leq N$. Clearly, we can cover these j nodes with N colors, thus we have a total of $2N$ colors.

Thus if the original graph could be colored with N^* colors, the test set needed for $\mathrm{TSO}^n(n/2)$ needs M tests where $N/2 \leq M < N + 2\log_2 N + 1$. If we can approximate the minimum number of tests with ratio R, we can approximate the number of colors with ratio $2R$. Finally, we translate the resulting instance of $\mathrm{TSO}^{n^2}(n/2)$ into an instance of $\mathrm{SB}^{\{0,1\}}(n)$ in the same way as in the proof of Lemma 11.

E Expected Size of an Optimal Solution of a Random Instance of $\mathbf{TS}^{\{1\}}(1)$, $\mathbf{TS}^{\{1\},\{0,2\}}(1)$ and $\mathbf{SB}^{\Sigma}$

E.1 Proof of Theorem 1(a)

Obviously, an optimal solution $\mathcal{T}$ to an instance $(n, \mathcal{S})$ of a test set problem satisfies $\lceil \log_2 n \rceil \leq |\mathcal{T}| \leq 2n - 2$. Below we show that for randomly generated instances of test set problems as described in Theorem 1, the expected size of the solution is $O(\log n)$.

First, we provide a proof for $\mathrm{TS}^{\{1\}}$. For any $i, j \in [0, n-1]$ with $i \neq j$ and any $S \in \mathcal{S}$, $\Pr[S \text{ does not distinguish } i \text{ and } j] = 1 - 2p(1-p)$. Hence, if we select $c' = c \log_2 n$ distinct sets $S_0, S_1, \ldots, S_{c'-1}$ from $\mathcal{S}$, the probability that none of these sets distinguish i from j is $(1 - 2p(1-p))^{c \log_2 n} \leq n^{-4}$. Hence, $\sum_{1 \leq i < j \leq n} \Pr[\{S_0, S_1, \ldots, S_{c'-1}\} \text{ distinguish } i \text{ from } j] \geq 1 - n^{-2}$ and the result follows since the size of an optimal solution of $(n, \mathcal{S})$ is at most $n - 1$.

For $\mathrm{TS}^{\{\{1\},\{0,2\}\}}$, we can replace each pair (i, j) by two pairs (i_1, j_1) and (i_2, j_2) such that a set $T \in \mathcal{T}$ contains (i_1, j_1) (respectively, (i_2, j_2)) if $|T \cap \{i, j\}| = 1$ (respectively, $|T \cap \{i, j\}| \in \{0, 2\}$).

For any pair (i, j) with $i \neq j$ and any $S \in \mathcal{S}$, $\Pr[(i_2, j_2) \notin S] = 2p(1-p)$. The proof can now be completed in a similar manner.

E.2 Proof of Theorem 1(b)

As noted already, an optimal solution $\vec{t}^*$ of an instance $(n, \mathcal{S})$ of the SB$^\Sigma$ problem satisfies $\lceil \log_2 n \rceil \leq |\vec{t}^*| \leq n-1$. We consider the case when $|\Sigma| > 1$ is a constant and each string $s = s_0 s_1 s_2 \ldots \in \mathcal{S}$ is generated independently randomly with $\Pr[s_i = a] = 1/|\Sigma|$ for any i and any $a \in \Sigma$. Note that $|\Sigma|$ is a constant in the biological applications of this problem, in particular $|\Sigma| = 4$ for DNA sequences and $|\Sigma| = 20$ for protein sequences.

Assume n and ℓ to be sufficiently large for asymptotic results and $\sigma = |\Sigma| > 1$ to be fixed. Using the proof of Theorem 1, it follows that it suffices to show that for a random string $x \in \Sigma^*$ of length $k = O(\log \ell)$, $\Pr[x \prec s] = p$ for some constant $0 < p < 1$ and any $s \in \mathcal{S}$. In [12, Examples 6.4, 6.7, 6.8, 9.3 and 10.11], Odlyzko uses the bounds and generating function described in [6] to give asymptotic bounds on $\Pr[x \prec s]$ when $\sigma = 2$. The result can be generalized to the case of any fixed $\sigma > 2$ in a straightforward manner and below we describe this generalization. For a fixed $x = x_1 x_2 \ldots x_k$, define the correlation polynomial $C_x(z)$ of x as $C_x(z) = \sum_{j=0}^{k-1} c_x(j) z^j$ where $c_x(0) = 1$ and for $1 \leq j < k$, $c_x(j) = \begin{cases} 1 & \text{if } x_1 x_2 \ldots x_{k-j} = x_{j+1} x_{j+2} \ldots x_k \\ 0 & \text{otherwise} \end{cases}$. Let $f_x(\ell)$ be the number of strings in Σ^* of length ℓ that do not contain x as a substring and $F_x(z) = \sum_{\ell=0}^{\infty} f_x(\ell) z^\ell$ be the generating function for this number. Then, $F_x(z) = \frac{C_x(z)}{z^k + (1-\sigma z) C_x(z)}$. ¿From this, the asymptotic bound for $\Pr[x \prec s] = 1 - \sigma^{-\ell} f_x(\ell)$ can be computed as:

$$\Pr[x \prec s] = 1 - \mathbf{e}^{-\frac{\ell}{\sigma^k C_x(1/\sigma)} + O(\ell k \sigma^{-\sigma k})} + O(\mathbf{e}^{-\ell/O(1)})$$

for all sufficiently large n, k and ℓ, where $\mathbf{e}$ is the base of natural logarithm. Note that $1 \leq C_x(\sigma) < 2$ and for a specific x, $C_x(\sigma)$ can be calculated exactly. Hence one can see that choosing $k = \Theta(\log_\sigma \ell)$ gives $\Pr[x \prec s] = p$ for some constant $0 < p < 1$.