Tight Approximability Results for Test Set Problems in Bioinformatics

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

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

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

† Supported by NSF grant CCR-0208821.

‡ Supported by NSF Grants CCR-0296041, CCR-0206795, CCR-0208749 and a CAREER grant IIS-0346973.

¶ Supported by NSF grant EIA-0112934.

A General Framework

Problem $\mathbf{TS}^{\Gamma}(k)$: $\Gamma \subseteq 2^{\{0,1,2\}}$, k a positive integer k

Instance: $(n, \mathcal{S})$ where $\mathcal{S} \subseteq 2^{\{0,1,2,\cdots,n-1\}}$

Terminologies:

- A k**-test** is a union of at most k sets from $\mathcal{S}$
- For a $\gamma \in \Gamma$ and two distinct elements $x, y \in \{0, 1, 2, \cdots, n-1\}$, **a k-test T γ-distinguishes x and y** if $|\{x, y\} \cap T| \in \gamma$.

Valid solutions: A collection $\mathcal{T}$ of k-tests such that

$$(\forall \mathbf{x}, \mathbf{y} \in \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \cdots, \mathbf{n}-\mathbf{1}\} \;\; \forall \gamma \in \Gamma) \;\; \mathbf{x} \neq \mathbf{y}$$

$$\Longrightarrow$$

$$\exists \mathbf{T} \in \mathcal{T} \text{ such that } T \ \gamma\text{-distinguishes } x \text{ and } y$$

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

Some Problems Captured by General Framework

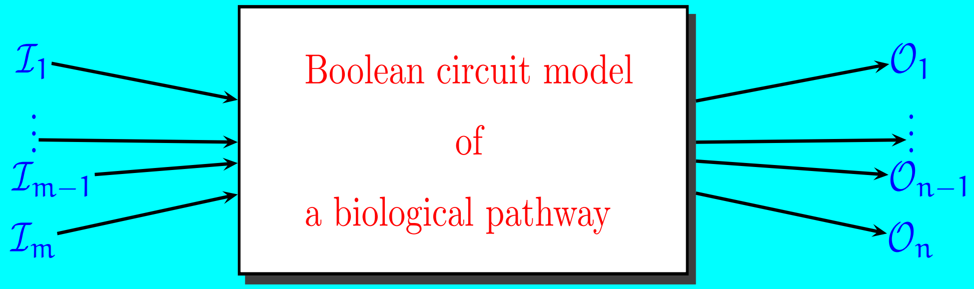
Minimum Test Collection Problems

- Equivalent to $\mathrm{TS}^{\{1\}}(1)$
  - Objects $\mathcal{O}_1, \mathcal{O}_2, \cdots, \mathcal{O}_n$
  - $\mathcal{O}_i, \mathcal{O}_j$ is distinguished by test T if $T \cap \{\mathcal{O}_i, \mathcal{O}_j\} = 1$.
- Applications: diagnostic testing
- Reference: Garey and Johnson's book on NP-completeness, page 71.

More Problems Captured by General Framework

Condition Cover Problems

- Captured by $\mathrm{TS}^{\{1\},\{0,2\}}(1)$



  - test : assignment σ of $I_1, I_2, \ldots, I_m$ to 0/1 values.
    For this assignment, m^{th} output is $\mathcal{O}_m(\sigma)$
  - Set of tests given as part of input
  - $\mathcal{O}_i, \mathcal{O}_j$ are distinguished if
    - $\exists \sigma : \mathcal{O}_i(\sigma) = \mathcal{O}_j(\sigma)$ & $\exists \rho : \mathcal{O}_i(\rho) \neq \mathcal{O}_j(\rho)$
- Applications: verifying a multi-output feedforward Boolean circuit as a model of specific biological pathways

More Problems Captured by General Framework

Simplest String Barcoding Problems ($\mathbf{SB}^{\Sigma}(1)$)

- very special case of $\mathrm{TS}^{\{1\}}(k)$
  - Given: set $\mathcal{S}$ of sequences over alphabet Σ
  - Definition of barcode: for a sequence s and a set of sequences $\mathbf{t}$, barcode$(s, \vec{\mathbf{t}})$ is the Boolean vector (c_0, c_1, c_{m-1}) where c_i is 1 if t_i is a substring of s.

$\vec{\mathbf{t}}$

	t_1 =A	t_2 =CC	t_3 =TTT	t_4 =GT
s =ACC	1	1	0	0
	substring			not substring

barcode$(s, \vec{\mathbf{t}}) = 1100$

Simplest String Barcoding Problems (continued)

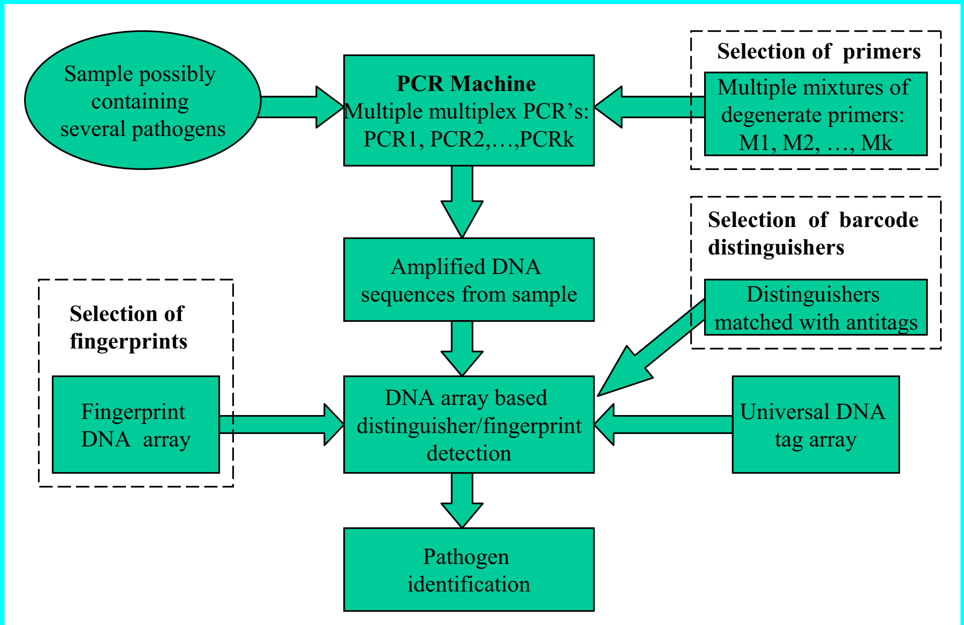
– Valid solutions: a set of sequences $\vec{\mathbf{t}}$ such that
$\forall s, s' \in \mathcal{S} \; : \; s \neq s' \equiv \text{barcode}(s, \vec{\mathbf{t}}) \neq \text{barcode}(s', \vec{\mathbf{t}})$

$\mathcal{S}$	$\vec{\mathbf{t}}$			
	A	CC	TTT	GT
s_1 =AAC	1	0	0	0
s_2 =ACC	1	1	0	0
s_3 =GGGG	0	0	0	0
s_4 =GTGTGG	0	0	0	1
s_5 =TTTT	0	0	1	0

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

Simplest String Barcoding Problems (continued)

- Applications:
    - database compression/fast database search for DNA sequences
    - DNA microarray designs for unknown pathogen identification



More Problems Captured by General Framework

Minimum Cost Probe Set with Threshold r (MCP$^{\Sigma}(r)$):

- variation of TS$^{\{1\}}(1)$

  Given : sets $\mathcal{S}$ and $\mathcal{P}$ of sequences over alphabet Σ and an integer $r > 0$

  Definition of r-barcode : for a sequence s and a set of sequences $\mathbf{t}$, r-barcode$(s, \vec{t})$ is the integer vector (c_0, c_1, c_{m-1}) where

$$c_i \;=\; \min\{\, r, \text{number of occurrences of } t_i \text{ in } s \,\}$$

Example of 2-barcode ($r = 2$)

$\vec{\mathbf{t}}$

	t_1 =A	t_2 =CC	t_3 =AC	t_4 =G
s =ACCCCA	2	2	1	0
	$\min\{2, 2\}$ r	$\min\{2, 3\}$ r		

$$2\text{-barcode}(s, \vec{\mathbf{t}}) = 2210$$

Valid solutions: set of sequences $\vec{\mathbf{t}} \subseteq \mathcal{P}$ such that

$$\forall s, s' \in \mathcal{S} \;:\; s \neq s' \equiv r\text{-barcode}(s, \vec{\mathbf{t}}) \neq r\text{-barcode}(s', \vec{\mathbf{t}})$$

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

- A special case of $\mathrm{MCP}^{\Sigma}(r)$ is $\mathrm{SB}^{\Sigma}(1)$

  If $\mathcal{P}$ is the set of all substrings of all sequences in $\mathcal{S}$ then $\mathrm{MCP}^{\Sigma}(1)$ is precisely $\mathrm{SB}^{\Sigma}(1)$

- Applications: in minimization the number of oligonucleotide probes needed for analyzing populations of ribosomal RNA gene (rDNA) clones by hybridization experiments on DNA microarrays

Relationships between problems

most general

$$
\begin{array}{ccccc}
 & & & & \text{condition cover} \\
 & & & & \equiv \\
TS^{\Gamma}(k) & \supseteq & TS^{\Gamma}(1) & \supseteq & TS^{\{1\},\{0,2\}}(1) \\
 & & & & \supseteq \\
 & & \text{Minimum Test Collection} & \equiv & TS^{\{1\}}(1) \\
 & & & & \supseteq \\
 & & & & SB^{\Sigma}(1) \\
 & & & & \supseteq \\
MCP^{\Sigma}(r) & \supseteq & MCP^{\Sigma}(1) & \supseteq & SB^{\{0,1\}}(1)
\end{array}
$$

most general (above $MCP^{\Sigma}(r)$)

most restrictive

- Upper bounds are proved for the more general cases ($\mathrm{TS}^{\{1\},\{0,2\}}(1)$, $\mathrm{TS}^{\{1\}}(1)$ and $\mathrm{MCP}^{\Sigma}(r)$)
- Lower bounds are proved for the most restrictive case ($\mathrm{SB}^{\{0,1\}}(1)$ and $\mathrm{TS}^{\{1\},\{0,2\}}(1)$)
- Lower bounds are also proved for $\mathrm{TS}^{\{1\},\{0,2\}}(k)$, $\mathrm{TS}^{\{1\}}(k)$ and $\mathrm{SB}^{\{0,1\}}(k)$ when k is large

condition cover

$\equiv$

$\mathrm{TS}^{\Gamma}(k) \supseteq \mathrm{TS}^{\Gamma}(1) \supseteq \mathrm{TS}^{\{1\},\{0,2\}}(1)$

$\mathrm{TS}^{\Gamma}(1) \supseteq \mathrm{TS}^{\{1\}}(1)$

Tight upper & lower bounds

Minimum Test Collection $\equiv \mathrm{TS}^{\{1\}}(1)$

$\mathrm{SB}^{\Sigma}(1) \supseteq \mathrm{TS}^{\{1\}}(1)$

$\mathrm{SB}^{\{0,1\}}(1) \supseteq \mathrm{SB}^{\Sigma}(1)$

$\mathrm{MCP}^{\Sigma}(r) \supseteq \mathrm{MCP}^{\Sigma}(1) \supseteq \mathrm{SB}^{\{0,1\}}(1)$

$$\begin{array}{ccccc} \mathrm{TS}^{\Gamma}(k) & \supseteq & \mathrm{TS}^{\{1\}}(k) & \supseteq & \mathrm{SB}^{\{0,1\}}(k) \\ \rotatebox{90}{$\supseteq$} & & & & \\ \mathrm{TS}^{\{0\},\{1,2\}}(k) & & & & \end{array}$$

`lower bounds`

`for large k`

Summary of our results

(matching upper/lower bounds)

Problem	Approximation Ratio		
	Upper Bound (algorithm)	Lower Bound	
		the bound	Assumptions
$\mathrm{TS}^{\{1\}}(1)$	$1+\ln n$	$(1-\varepsilon)\ln n$	NP$\not\subset$DTIME$(n^{\log\log n})$
$\mathrm{TS}^{\{1\},\{0,2\}}(1)$	$1+\ln 2+\ln n$	$(1-\varepsilon)\ln n$	NP$\not\subset$DTIME$(n^{\log\log n})$
$\mathrm{SB}^{\Sigma}(1)$	$1+\ln n$	$(1-\varepsilon)\ln n$	NP$\not\subset$DTIME$(n^{\log\log n})$ $\lvert\Sigma\rvert > 1$
$\mathrm{MCP}^{\Sigma}(r)$	$[1+o(1)]\ln n$	$(1-\varepsilon)\ln n$	NP$\not\subset$DTIME$(n^{\log\log n})$ $\lvert\Sigma\rvert > 1$
$\mathrm{TS}^{\{1\}}(n^{\delta})$ $\mathrm{TS}^{\{1\},\{0,2\}}(n^{\delta})$		n^{ε}	NP$\neq$co-RP $0<\varepsilon<\delta<1$
$\mathrm{SB}^{\{0,1\}}(n^{\delta})$		n^{ε}	NP$\neq$co-RP $0<\varepsilon<\delta<\frac{1}{2}$

ε and δ are *arbitrary* constants

Comparison of our results with those in [1,2]

[1] B. V. Halldórsson, M. M. Halldórsson and R. Ravi. *On the Approximability of the Minimum Test Collection Problem*, Proc. Ninth Annual European Symposium on Algorithms, Lecture Notes in Computer Science 2161, pp. 158-169, 2001.

[2] K. M. J. De Bontridder, B. V. Halldórsson, M. M. Halldórsson, C. A. J. Hurkens, J. K. Lenstra, R. Ravi and L. Stougie. *Approximation algorithms for the test cover problem*, Mathematical Programming-B, Vol. 98, No. 1-3, 2003, pp. 477-491.

Comparison of our results with those in [1,2]

- The authors in [1,2] proved a $(1-\varepsilon)\ln n$ lower bound for approximation for $\mathrm{TS}^{\{1\}}(1)$

  We prove a lower bound of $(1-\varepsilon)\ln n$ for the very special case $\mathrm{SB}^{\{0,1\}}$

$$\boxed{\mathrm{TS}^{\{1\}}(1)} \supseteq \boxed{\mathrm{SB}^{\{0,1\}}(1)}$$

$(1-\varepsilon)\ln n$ of [1,2] | our $(1-\varepsilon)\ln n$

Comparison of our results with those in [1,2] (continued)

- The proof in [1,2] from set-cover to $\mathrm{TS}^{1}(1)$ does not seem to be easily transformable to provide a lower bound for $\mathrm{SB}^{\{0,1\}}$ with a similar quality of non-approximability because of the special nature of $\mathrm{SB}^{\{0,1\}}$

- We therefore needed to introduce an artificial intermediate problem (the "test set with order with positive integer parameter m" problem, denoted by TSO^{m}) which we could then translate to $\mathrm{SB}^{\{0,1\}}$ in a non-trivial manner

set cover	$\leftarrow$	TSO^{m}	$\leftarrow$	$\mathrm{SB}^{\{0,1\}}$
$(1-\varepsilon_1)\ln n$		$(1-\varepsilon_2)\ln n$		$(1-\varepsilon_3)\ln n$
non-approximability		non-approximability		non-approximability

Comparison of our results with those in [1,2] (continued)

- In general, TSO^m is neither equivalent to or nor a special case of $\mathrm{TS}^{\{1\}}(1)$.

Summary of Other Techniques Used

- Algorithm 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 the change of information content of the partition of the universe
  - A set of tests $\mathcal{T}$ defines an entropy $\mathsf{H}_{\mathcal{T}}$
  - notion of information content IC

$$\begin{aligned} \mathcal{T} &= \text{already selected tests} \\ \mathsf{T} &= \text{new test} \\ \mathrm{IC}(\mathsf{T},\mathcal{T}) &= \mathsf{H}_{\mathcal{T}} - \mathsf{H}_{\mathcal{T}\cup\mathsf{T}} \end{aligned}$$

  - Greedy heuristics

$\mathcal{T} = \emptyset$

while $\mathsf{H}_{\mathcal{T}} \neq 0$ **do**

select a $\mathsf{T} \in \mathcal{S} - \mathcal{T}$ that *maximizes* $\mathrm{IC}(\mathsf{T},\mathcal{T})$

$\mathcal{T} = \mathcal{T} \cup \mathsf{T}$

endwhile

- The inapproximability results for $\mathrm{TS}^{\{1\}}(n^{\delta})$, $\mathrm{TS}^{\{0\}.\{1,2\}}(n^{\delta})$ and $\mathrm{SB}^{\{0,1\}}(n^{\delta})$ are obtained by approximation preserving reductions from the graph coloring problem.

Thank You for your attention!!