dipartimentoinformatica
Università degli Studi dell'Aquila

Entropy-based Improved Approximation for Test Cover

Pasquale Caianiello, Alessandro Sabetta

Technical Report TRCS 001/2013

Dipartimento di Informatica
Università degli Studi dell'Aquila
Via Vetoio Loc. Coppito
I-67010 L'Aquila, Italy

http://www.di.univaq.it

Technical Report TRCS Series

Entropy-based Improved Approximation for Test Cover

Pasquale Caianiello [*]*Alessandro Sabetta* [†]

Dipartimento di Informatica Università dell'Aquila Via Vetoio, Coppito 67100 AQ, Italy

11th November 2004

Abstract

This paper studies the polynomial time heuristic algorithm MIC for the Test Cover problem that selects, with a greedy strategy, a test that maximizes the classical Shannon entropy function. The Test Cover Problem is known not to admit a polynomial time algorithm with performance bound $(1-\epsilon)(\log m)$, for any $\epsilon > 0$ unless $NP \subset DTIME(m^{\lg_2 \lg_2 m})$. We prove that MIC approximation ratio is $\leq \frac{1}{\ln 2} \ln m + O(\ln \ln m) \approx 1.45 \ln m$, improving previous best approximation algorithms with ratio of $\approx 2 \ln m$.

1 Introduction

Given a ground set of items $\{1, ..., m\}$ and a collection of tests $T_1, ..., T_n$, $T_i \subseteq \{1, ..., m\}$ the Test Cover Problem (TCP) aims at constructing a minimum cardinality set of tests that distinguishes all items. It naturally arises in problems of medical diagnosis and pattern recognition, and has been recently addressed in a number of works approaching the Single Nucleotide Polymorphism (SNPs) selection problem [1, 11, 14] for haploid identification and reconstruction in computational biology.

This paper provides an estimate of the approximation ratio of the Mutual Information Clustering (MIC) Algorithm for such problems, as proposed in [2]. In the rest of this introduction we report on existing literature on the subject; in Section 2 we recall basic concepts from Information Theory by

[*] E-mail: caianiel@di.univaq.it

[†] E-mail: vanalesab@libero.it

applying them to partitions of a set, and we provide a slight generalization of partition entropy extremal properties as reported by Shastri and Govil [13]. Finally in Section 3 we present the algorithm, prove its correctness, and provide an upper bound for its approximation ratio.

1.1 Test Cover and Set Cover

The input of the Test Cover Problem (TCP) is a set of *items* $\{1, ..., m\}$, and a collection of *tests* $T_1, ..., T_n$, $T_i \subseteq \{1, ..., m\}$. A test T_j *covers* or *differentiates* the item pair $\{h, i\}$ if either $h \in T_j$ or $i \in T_j$, i.e., if $|T_j \cap \{h, i\}| = 1$. A subcollection $\Gamma \subseteq \{T_1, ..., T_n\}$ of tests is a *test cover* if each of the $m(m-1)/2$ item pairs is covered by at least one test in Γ. The goal is to find a test cover of minimum cardinality, if one exists.

The TCP has a close relationship with the Set Cover Problem (SCP), which inputs a set of elements $\{1, ..., M\}$ and a collection of *sets* $S_1, ..., S_N$, $S_i \subseteq \{1, ..., M\}$. Its goal is to find a *set cover* of minimum cardinality, if one exists, that is a subcollection $\Upsilon \subseteq \{S_1, ..., S_N\}$ of sets such that each of the M elements is covered by at least one set in Υ, i.e. $\bigcup_{S \in \Upsilon} = \{1, ..., M\}$.

In facts, a TCP instance T can be transformed into an equivalent SCP instance S_T with $M = \frac{(m)(m-1)}{2}$ elements and $N = n$ subsets, by constructing an element $o_{i,j}$ in S_T for each pair $\{i, j\}$ of different items in T. Subsets in S_T are such that $o_{i,j} \in S_l \Leftrightarrow |T_l \cap \{i, j\}| = 1$.

Thus, an algorithm for the SCP also works for the TCP. All known TCP approximation algorithms are based on this transformation [4].

1.2 The Greedy TCP Algorithm

In particular, the greedy algorithm for the SCP, which selects a subset covering the largest number of yet uncovered elements, directly gives a greedy algorithm for the TCP that chooses a test that covers the largest number of yet uncovered pairs. Hence, given a TCP instance T, the Greedy TCP builds the equivalent SCP instance S_T and applies the greedy algorithm for the SCP to find a set cover σ. The Greedy TCP returns the tests collection $\tau = \{T_i | S_i \in \sigma\}$.

Theorem 1.1. *[6, 8, 9] The Greedy TCP approximation ratio is* $1 + 2 \ln m$.

1.3 Inapproximability results

Moret and Shapiro [10] showed how to reduce the SCP to the TCP. Thus they argue that the TCP is NP-Hard and extend other inapproximability results for the SCP to the TCP:

Theorem 1.2. *[10] The TCP has no polynomial-time algorithm with performance bound $o(\lg m)$, unless $P = NP$, and no polynomial-time algorithm with performance bound $(1 - \epsilon) \ln m$, for any $\epsilon > 0$, unless $NP \subset DTIME(m^{\lg_2 \lg_2 m})$.*

2 Entropy of partitions

Entropy is a measure of uncertainty or *information* of a random variable introduced by Shannon [12]. As Shannon himself stated, it is the only function satisfying all the *reasonable properties* required for such a measure.

In the following we recall basic definitions and properties of entropy by applying them to our scenario of sets of tests in connection with the theory of set partitions. We refer to [3] for a thorough treatment of Information Theory.

Definition 2.1 ($\mathcal{R}_\Gamma$). *A tests collection $\Gamma \subseteq \{T_1, ..., T_n\}$ determines a binary relation $\mathcal{R}_\Gamma$ on the set $\{1, ..., m\}$. We say that two elements i, j (possibly equal) are related, and write $\mathcal{R}_\Gamma(i, j)$ if no test in Γ can distinguish them. More formally:*

$$\mathcal{R}_\Gamma(i, j) = \forall T_l \in \Gamma,\ (i, j \in T_l) \vee (i, j \notin T_l)$$

Proposition 2.2. *$\mathcal{R}_\Gamma$ is an equivalence relation.*

Definition 2.3 (Δ_Γ). *Let Γ be a tests collection over $\{1, ..., m\}$, we define Δ_Γ as the partition induced on the set $\{1, ..., m\}$ by $\mathcal{R}_\Gamma$.*

A partition Δ over the set $\{1, ..., m\}$ has a natural probability mass function defined over its elements:

$$p(x) = Pr\{x\} = \frac{|x|}{m},\ x \in \Delta$$

which allows the following

Definition 2.4 (Entropy). *Let $\Gamma \subseteq \{T_1, ..., T_n\}$ be a tests collection. Let $\Delta_\Gamma = \{A_1, ..., A_k\}$ be the partition induced by Γ on the set $\{1, ..., m\}$. The entropy $H(\Gamma)$ of Γ is defined by:*

$$H(\Gamma) = H(\Delta_\Gamma) = -\sum_{x \in \Delta_\Gamma} p(x) \lg_2 p(x)$$

2.1 Joint and conditional partition entropy

Joint and conditional entropy are analogously defined relying over the natural joint and conditional probabilities:

Definition 2.5 (Joint Entropy of Γ, Υ). *Let $\Gamma, \Upsilon \subseteq \{T_1, ..., T_n\}$ be tests collections over $\{1, ..., m\}$. The joint entropy $H(\Gamma, \Upsilon)$ is:*

$$H(\Gamma, \Upsilon) = H(\Delta_\Gamma, \Delta_\Upsilon) = -\sum_{x \in \Delta_\Gamma} \sum_{y \in \Delta_\Upsilon} p(x, y) \lg_2 p(x, y)$$

where:

$$p(x, y) = Pr\{x \cap y\} = \frac{|x \cap y|}{m},\ x \in \Delta_\Gamma,\ y \in \Delta_\Upsilon$$

Proposition 2.6. *Let $\Gamma, \Upsilon \subseteq \{T_1, ..., T_n\}$ be tests collections over $\{1, ..., m\}$.*

$$H(\Gamma, \Upsilon) = H(\Gamma \cup \Upsilon)$$

Definition 2.7 (Conditional Entropy of Γ known Υ). *Let $\Gamma, \Upsilon \subseteq \{T_1, ..., T_n\}$ be tests collections over $\{1, ..., m\}$. The entropy $H(\Gamma|\Upsilon)$ is defined by:*

$$H(\Gamma|\Upsilon) = H(\Delta_\Gamma|\Delta_\Upsilon) = -\sum_{x \in \Delta_\Gamma} \sum_{y \in \Delta_\Upsilon} p(x, y) \lg_2 p(x|y)$$

where:

$$p(x, y) = Pr\{x \cap y\} = \frac{|x \cap y|}{m},\ x \in \Delta_\Gamma,\ y \in \Delta_\Upsilon$$

$$p(x|y) = Pr\{x|y\} = \frac{|x \cap y|}{|y|}\ x \in \Delta_\Gamma,\ y \in \Delta_\Upsilon$$

The tight coherence of partition entropy with Shannon's allows to deduce the analogous chain rule:

Theorem 2.8 (Chain Rule).

$$H(\Gamma_1, \Gamma_2, ..., \Gamma_n) = \sum_{i=1}^{n} H(\Gamma_i|\Gamma_{i-1}, ..., \Gamma_1)$$

2.2 Extremal partition entropy

Extremal properties of partition entropy have been studied by Shastri and Govil [13]. We report their results and a slight generalization.

Theorem 2.9 (Maximum Entropy). *[13] Let $\Delta = \{A_1, ..., A_k\}$ be a partition on the set $\{1, ..., m\}$ such that $\forall A_i, A_j \in \Delta,\ ||A_i| - |A_j|| \leq 1$. Let $m = qk + r$ with $0 \leq r < k$. Then*

$$H(\Delta) = H_{max(k)} = -\frac{1}{m}\left\{r(q+1)\lg_2\frac{q+1}{m} + (k-r)q\lg_2\frac{q}{m}\right\}$$ [1]

is the maximum entropy-value for partitions with size k.

The result can be improved with the following:

Corollary 2.10 (Maximum Entropy). *Let $\Delta = \{A_1, ..., A_k\}$ be a partition on the set $\{1, ..., m\}$ such that $\forall A_i, A_j \in \Delta,\ ||A_i| - |A_j|| \leq 1$. Then $H(\Delta) = H_{max(k)}$ is the maximum entropy-value for the partitions with size at most k.*

Proof. It is sufficient to show that $H_{max(k)} \leq H_{max(k+1)}$. We obtain:

$$H_{max(k)} = -\frac{1}{m}\left\{r_1(q_1+1)\lg_2\frac{q_1+1}{m} + (k-r_1)q_1\lg_2\frac{q_1}{m}\right\}$$

$$H_{max(k+1)} = -\frac{1}{m}\left\{r_2(q_2+1)\lg_2\frac{q_2+1}{m} + (k+1-r_2)q_2\lg_2\frac{q_2}{m}\right\}$$

where $m = q_1 k + r_1$, $0 \leq r_1 < k$, $m = q_2(k+1) + r_2$ and $0 \leq r_2 < k+1$. Let's consider this two possible situations according to the results of the divisions $\frac{m}{k}$ and $\frac{m}{k+1}$.

- Let's suppose that $q_1 = q_2 = q$ and $r_2 = r_1 - q$.

$$H_{max(k)} - H_{max(k+1)} = ... = \left[\frac{q(q+1)}{m}\right]\left[\lg_2\frac{q}{m} - \lg_2\frac{q+1}{m}\right] < 0$$

- Now we suppose that $q_2 < q_1$.

$$H_{max(k)} - H_{max(k+1)} \leq (\lg_2(q_2+1) - \lg_2 q_1) \leq \lg_2 q_1 - \lg_2 q_1 = 0$$

□

Theorem 2.11 (Minimum Entropy). *[13] Let $\Delta = \{A_1, ..., A_k\}$ be a partition on the set $\{1, ..., m\}$ such that for all $A_i \in \Delta$ but exactly one, $|A_i| = 1$. Then*

$$H(\Delta) = H_{min(k)} = -\frac{1}{m}\left\{(k-1)\lg_2\frac{1}{m} + (m-k+1)\lg_2(1-\frac{k-1}{m})\right\}$$

is the minimum entropy-value for partitions with size k.

[1] This expression is slightly different from the one given in [13] which is incorrect.

Again, the result can be improved with the following:

Corollary 2.12 (Minimum Entropy). *Let $\Delta = \{A_1, ..., A_k\}$ be a partition on the set $\{1, ..., m\}$ such that for all $A_i \in \Delta$ but exactly one, $|A_i| = 1$. Then $H(\Delta) = H_{min(k)}$ is the minimum entropy-value for partitions with size at least k.*

Proof. It is sufficient to show that $H_{min(k)} \leq H_{min(k+1)}$. We obtain:

$$H_{min(k)} = -\frac{k-1}{m} \lg_2 \frac{1}{m} - \frac{m-k+1}{m} \lg_2 \frac{m-k+1}{m}$$

$$H_{min(k+1)} = -\frac{k}{m} \lg_2 \frac{1}{m} - \frac{m-k}{m} \lg_2 \frac{m-k}{m}$$

Then:

$$H_{min(k)} - H_{min(k+1)} = ... = -\frac{1}{m} \lg_2(m-k+1) < 0$$

□

3 The MIC algorithm

The MIC (*Mutual Information Clustering*) Algorithm chooses step by step, with greedy strategy, a most informative test.

```
input $(\{1, ..., m\}, \{T_1, ..., T_n\})$
$\Gamma = \emptyset$
while $H(\{T_1, ..., T_n\}|\Gamma) \neq 0$ do
begin
    $T \in \arg\max_{T_i \in (\{T_1,...,T_n\} \setminus \Gamma)} H(\{T_i\}|\Gamma)$
    $\Gamma = \Gamma \cup \{T\}$
end
return $\Gamma$
```

Table 1: The MIC Algorithm

Worst case running time of the MIC Algorithm is $O(mn^2)$.

3.1 Correctness

In this section we assume that input allows a cover, i.e. there are no identical elements.

Theorem 3.1. *Let* Γ *be a tests collection over*$\{1, ..., m\}$. *The following are equivalent:*

1. $H(\{T_1, ..., T_n\}|\Gamma) = 0$
2. $H(\Gamma) = \lg_2 m$
3. *each of the* $m(m-1)/2$ *items pair is covered (or distinguished) by at least one test in* Γ

The proof is simple. The intuitive argument consists in noting that the information necessary for distinguishing/identifying all elements is $\lg_2 m$ and, once $H(\Gamma)$ is $\lg_2 m$, if Γ is known no other test can provide extra information.

This theorem ensures that Γ is a test cover when MIC exits the while-cycle. It can be observed that building a possibly suboptimal test cover is conceptually equivalent to finding a tests collection with complete information about all/remaining tests.

3.2 Approximation ratio

In this section we prove our main result: the approximation ratio of the MIC Algorithm is $\frac{1}{\ln 2} \ln m + O(\ln \ln m)$.

Definition 3.2. *Let*

G_i *be the test selected by MIC at step* i.

Γ_i *be the test collection constructed by MIC after step* i: $\Gamma_i = \{G_1, ..., G_i\}$ *with* $\Gamma_0 = \emptyset$.

H_i *be the information of* G_i, *given previously chosen tests:* $H_i = H(G_i|\Gamma_{i-1})$

Observaton 3.3 (Information $H(\Gamma_k)$). *From the chain rule follows*

$$H(\Gamma_k) = \sum_{i=1}^{k} H(G_i|G_{i-1}, ..., G_1) = \sum_{i=1}^{k} H(G_i|\Gamma_{i-1}) = \sum_{i=1}^{k} H_i$$

Lemma 3.4 (H_i Monotonicity). H_i *is non-increasing, i.e.* $\forall i,\ H_i \geq H_{i+1}$.

Proof. Since G_i is selected before G_{i+1} we have $H(G_i|\Gamma_{i-1}) \geq H(G_{i+1}|\Gamma_{i-1})$. Thus

$$H_i = H(G_i|\Gamma_{i-1}) \geq H(G_{i+1}|\Gamma_{i-1}) \geq H(G_{i+1}|\Gamma_i) = H_{i+1}$$

because conditioning reduces entropy. □

Lemma 3.5 (Upper Gap). *Let Δ be any tests collection over $\{1, ..., m\}$. If $H(\Delta) > \lg_2 m - \frac{2}{m}$ then $H(\Delta) = \lg_2 m$*

Proof. On the one end, the entropy of the partition $\Delta_m = \{\{1\}, ..., \{m\}\}$ of exactly m classes is $H(\Delta_m) = \lg_2 m$. On the other, the maximal-entropy Corollary 2.10 guarantees that, among partitions with strictly less than m classes, the maximal entropy value is obtained with a partition Δ_{max} consisting of $m-2$ classes of cardinality 1, and one of cardinality 2. Thus:

$$H(\Delta_{max}) = \frac{m-2}{m}\lg_2 m + \frac{2}{m}\lg_2 \frac{m}{2} = \lg_2 m - \frac{2}{m}$$

□

The following lemma guarantees that after k iterations MIC constructs a collection Γ_k with at least half the information of any other collection with k tests. This evaluates the error made by MIC after k steps, and it is the key result towards the proof of its approximation ratio.

Lemma 3.6. *Let $\tau = \{t_1, ..., t_{|\tau|}\}$ be any tests collection and $\Gamma_{|\tau|}$ the tests collection constructed after $|\tau|$ iterations of MIC . Then $H(\Gamma_{|\tau|}) \geq \frac{H(\tau)}{2}$.*

Proof. Suppose by contradiction that $H(\Gamma_{|\tau|}) < \frac{H(\tau)}{2}$. Hence

$$H(\tau|\Gamma_{|\tau|}) \geq H(\tau) - H(\Gamma_{|\tau|}) > H(\tau) - \frac{H(\tau)}{2} = \frac{H(\tau)}{2}$$

Now, if we rewrite $H(\tau|\Gamma_{|\tau|})$ by applying the chain rule we get

$$H(\tau|\Gamma_{|\tau|}) = \sum_{i=1}^{|\tau|} H(t_i|t_{i-1}, ..., t_1, \Gamma_{|\tau|}) > \frac{H(\tau)}{2}$$

and we can note that since the $|\tau|$ terms on the left side of the inequality are all non-negative, at least one of them (say the k-th) is such that:

$$H(t_k|t_{k-1}, ..., t_1, \Gamma_{|\tau|}) > \frac{H(\tau)}{2|\tau|}$$

Then:

$$H(t_k|\Gamma_{|\tau|}) \geq H(t_k|t_{k-1}, ..., t_1, \Gamma_{|\tau|}) > \frac{H(\tau)}{2|\tau|}$$

This proves that there exists t_k such that $H(t_k|\Gamma_{|\tau|}) > \frac{H(\tau)}{2|\tau|}$, which implies that $t_k \notin \Gamma_{|\tau|}$.

By the monotonicity of H_i (Lemma 3.4), we have that for all $i \leq |\tau|$ $H_i \geq \frac{H(\tau)}{2|\tau|}$. Thus

$$H(\Gamma_{|\tau|}) = \sum_{i=1}^{|\tau|} H_i \geq |\tau| \frac{H(\tau)}{2|\tau|} = \frac{H(\tau)}{2}$$

leading to a contradiction. □

It should be noted that the previous lemma needs no hypothesis on the initial status of the computation. The lemma is still valid if MIC had already chosen some tests Υ. The following generalizes lemma 3.6 to the general case.

Lemma 3.7. *Let's suppose that during its computation, MIC has constructed the set* $\Upsilon = \Gamma_{|\Upsilon|}$. *Let* $\tau = \{t_1, ..., t_{|\tau|}\}$ *be any tests collection and let* $\widetilde{\Gamma} = \Gamma_{|\Upsilon|+|\tau|} \backslash \Gamma_{|\Upsilon|}$ *be the tests collection constructed in the subsequent* $|\tau|$ *iterations. Then* $H(\widetilde{\Gamma}|\Upsilon) \geq \frac{H(\tau|\Upsilon)}{2}$.

Proof. The proof is identical to the previous lemma 3.6 and can be obtained by substituting for all entropy expressions $H(x)$ with $H(x|\Upsilon)$ and $H(x|y)$ with $H(x|(y \cup \Upsilon))$. □

Theorem 3.8 (MIC Progression). *Let* τ^* *be an optimal test cover. Let* Π_i *be the test collection built by MIC after* $i|\tau^*|$ *steps:* $\Pi_i = \{G_1, ..., G_{i|\tau^*|}\}$ *Then, for any* $i \geq 1$*:*

$$H(\tau^*|\Pi_i) \leq \frac{\lg_2 m}{2^i}$$

Proof. By induction on i:

i=1. By Lemma 3.6, after $|\tau^*|$ iterations, MIC finds a partial cover Π_1 such that $H(\Pi_1) \geq \frac{\lg_2 m}{2}$. Thus

$$H(\tau^*|\Pi_1) = H(\tau^* \cup \Pi_1) - H(\Pi_1) \leq \lg_2 m - \frac{\lg_2 m}{2} = \frac{\lg_2 m}{2}$$

Inductive step. Assume that MIC has gone through $(i-1)|\tau^*|$ iterations and has constructed a partial cover Π_{i-1}, and that

$$H(\tau^*|\Pi_{i-1}) \leq \frac{\lg_2 m}{2^{i-1}}$$

By Lemma 3.7, in the subsequent $|\tau^*|$ iterations, MIC selects the set $\Pi = \Pi_i \setminus \Pi_{i-1}$ of new elements such that $H(\Pi|\Pi_{i-1}) \geq \frac{H(\tau^*|\Pi_{i-1})}{2}$. Thus

$$H(\tau^*|\Pi_i) = H(\tau^*|\Pi_{i-1}) - H(\Pi|\Pi_{i-1}) \leq \frac{H(\tau^*|\Pi_{i-1})}{2} \leq \frac{\lg_2 m}{2^i}$$

□

Theorem 3.9 (MIC Approximation Ratio). *Let τ^* be an optimal test cover. Then MIC builds a test cover with cardinality at most $|\tau^*|\lceil \lg_2(m \lg_2 m)\rceil$.*

Proof. By the Upper Gap Lemma 3.5, MIC iteration ends when the remaining entropy reaches its theoretical bound:

$$H(\tau^*|\Pi_i) \leq \frac{\lg_2 m}{2^i} < \frac{2}{m}$$

The second inequality holds when $i > \lg_2(m \lg_2 m)$, so MIC goes through at most $|\tau^*|\lceil \lg_2(m \lg_2 m)\rceil$ iterations. □

Corollary 3.10. *MIC approximation ratio is $\frac{1}{\ln 2} \ln m + \mathcal{O}(\ln \ln m)$*

Proof.

$$\lceil \lg_2(m \lg_2 m)\rceil \leq \frac{1}{\ln 2} \ln m + \mathcal{O}(\ln \ln m)$$

□

4 Conclusion

This paper shows that Shannon's information proves to be a better measure to guide greedy selection strategies than simpler counting measures for the TCP. In particular, the use of information allows a relevant improvement of the approximation ratio of a greedy algorithm that uses set cardinality. We are confident that similar improvements may be obtained for other NP-hard problems whose approximating algorithms are based on simple greedy strategies and, in future work, we plan to investigate its application on further problems.

We should also mention that [2] reports on a broad experimentation of MIC both in simulated and natural genomic scenarios. MIC achieves nearly optimal results for feasible tests and mostly better results than other strategies proposed for genomic data, a fact that recommends for a broader use of Shannon information as a measure of dependency in genomics.

5 Acknoledgement

We wish to thank Guido Proietti for useful recommendations over previous drafts. This work was partially supported by MIUR 40 ”Ragionamento su aggregati e numeri a supporto della programmazione”.

References

[1] V. Bafna et al (2003). Haplotypes and Informative SNP Selection Algorithms: Don’t Block out Information. *RECOMB’03*, Berlin.

[2] P. Caianiello, W. Casey, B. Mishra, A. Sabetta (2004). Mutual Information Clustering Algorithms for Haplotyping, Tech. Rep., Dept. of Computer Science, New York University.

[3] T.M. Cover and J.A. Thomas (1989). *Elements of Information Theory*, Wiley, New York.

[4] K.M.J. De Bontridder et al (2003). Approximation Algorithms for the Minimum Test Cover Problem. *Mathematical Programming B 98*, 477-491.

[5] U. Feige (1998). A threshold of ln n for approximating set cover. *Journal of the ACM 45*, 634-652.

[6] M.R. Garey and D.S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco.

[7] B.V. Halldòrsson, M.M. Halldòrsson, R. Ravi (2001). On the Approximability of the Test Collection Problem. *Proceedings of the 9th Annual European Symposium on Algorithms, 2001*, 158-169.

[8] D.S. Johnson (1972). Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences 9*, 256-278.

[9] L. Lovász (1975). On the ratio of optimal integral and fractional covers. *Discrete Mathematics 13*, 383-390.

[10] B.M.E. Moret and H.D. Shapiro (1985). On minimizing a set of tests. *SIAM Journal on Scientific and Statistical Computing 6*, 983-1003.

[11] N. Patil et al (2001). Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. *Science 294*.

[12] C. E. Shannon (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 27*, 379-423, 623-656

[13] A. Shastri and R. Govil (2001). Optimal discrete entropy. *Applied Mathematics E-notes 1*, 73-76.

[14] K. Zhang et al (2002). A Dynamic Programming Algorithm for Haplotype Block Partitioning. *PNAS 99*.