Minimum Multicolored Subgraph Problem in Multiplex PCR Primer Set Selection and Population Haplotyping

M.T. Hajiaghayi[1], K. Jain[2], L.C. Lau[3], I.I. Măndoiu[4], A. Russell[4], and V.V. Vazirani[5]

[1] Laboratory for Computer Science, MIT, hajiagha@mit.edu
[2] Microsoft Research, kamalj@microsoft.com
[3] Department of Computer Science, University of Toronto, chi@cs.toronto.edu
[4] CSE Department, University of Connecticut, {ion,acr}@cse.uconn.edu
[5] College of Computing, Georgia Institute of Technology, vazirani@cc.gatech.edu

Abstract. In this paper we consider the minimum weight multicolored subgraph problem (MWMCSP), which is a common generalization of minimum cost multiplex PCR primer set selection and maximum likelihood population haplotyping. In this problem one is given an undirected graph G with non-negative vertex weights and a color function that assigns to each edge one or more of n given colors, and the goal is to find a minimum weight set of vertices inducing edges of all n colors. We obtain improved approximation algorithms and hardness results for MWMCSP and its variant in which the goal is to find a minimum number of vertices inducing edges of at least k colors for a given integer $k \leq n$.

1 Introduction

In this paper we consider the following *minimum weight multicolored subgraph problem* (MWMCSP): given an undirected graph G with non-negative vertex weights and a color function that assigns to each edge one or more of n given colors, find a minimum weight set of vertices of G inducing edges of all n colors. We also consider the generalization of MWMCSP in which one seeks a minimum weight set of vertices inducing edges of at least k colors for a given integer $k \leq n$, referred to as the *minimum weight k-colored subgraph problem* (MWkCSP), and the unweighted versions of MWMCSP and MWkCSP, denoted MMCSP and MkCSP, respectively. As detailed below, MWMCSP and its variants model two important bioinformatics problems: minimum cost multiplex PCR primer set selection and maximum likelihood population haplotyping.

1.1 Primer set selection for DNA amplification by PCR

A critical step in many high-throughput genomic assays is the cost-effective amplification of DNA sequences containing loci of interest via biochemical reactions such as the *Polymerase Chain Reaction* (PCR). In its basic form, PCR requires a pair of short single-stranded DNA sequences, referred to as *PCR primers*, flanking the amplification locus on the two strands of the template. In *multiplex*

PCR, multiple genomic loci are amplified simultaneously (and a primer may simultaneously participate in multiple amplifications). In addition to constraints on individual primer properties that affect reaction efficiency, such as primer melting temperature and lack of secondary structure, multiplex PCR primer set selection must ensure various *pairwise* compatibility constraints between selected primers. Since the efficiency of PCR amplification falls off exponentially as the length of the amplification product increases, an important practical constraint is that the two primer sites defining a product must be within a certain maximum distance L of each other. In applications such as spotted microarray synthesis [1] a further pairwise compatibility constraint is the requirement of unique amplification: for every desired amplification locus there should be a pair of primers that amplifies a DNA fragment surrounding it but no other fragment.

Subject to these constraints, one would like to minimize the total cost of the primer set required to amplify the n given loci. As noted by Fernandes and Skiena [1], primer selection problem subject to pairwise compatibility constraints can be easily reduced to M(W)MCSP: each candidate primer becomes a graph vertex and each pair of primers that feasibly amplifies a desired locus becomes an edge colored by the respective locus number. More generally, the problem of selecting the minimum size/cost set of primers required to amplify at least k of the n loci reduces to M(W)kCSP; this problem arises when several multiplex reactions are required to amplify the given loci.

1.2 Maximum likelihood population haplotyping

The most common form of genomic variation between individuals, is the presence of different DNA nucleotides, or *alleles*, at certain chromosomal locations, commonly referred to as *single nucleotide polymorphisms* (SNPs). For diploid organisms such as humans, the combinations of SNP alleles in the maternal and paternal chromosomes of an individual are referred to as the individual's *haplotypes*. Finding the haplotypes in human populations is an important step in determining the genetic basis of complex diseases [2].

With current technologies, it is prohibitively expensive to directly determine the haplotypes of an individual, but it is possible to obtain rather easily the conflated SNP information in the so called *genotype*. The *population haplotyping problem* (PHP) seeks to infer the set of haplotypes explaining the genotypes observed in a large population. Formally, a haplotype is represented as a 0/1 vector – e.g., by representing the most frequent SNP allele as a 0 and the alternate allele as a 1 – while a genotype is a 0/1/2 vector, where 0 (1) means that both chromosomes contain the respective SNP allele and 2 means that the two chromosomes contain different SNP alleles. We say that a set $\mathcal{H}$ of haplotypes *explains* a given set $\mathcal{G}$ of genotypes if, for every $g \in \mathcal{G}$, there exist $h, h' \in \mathcal{H}$ with $h + h' = g$, where $h + h'$ is the vector whose i-th component is equal to 2 when $h_i \neq h'_i$, and to the common value of h_i and h'_i when $h_i = h'_i$.

Several optimization objectives have been considered for PHP and the related *genotype phasing* problem, which seeks a pair of haplotypes explaining each of the given genotypes – see, e.g., [3–5] for recent surveys. In the *maximum likelihood* approach to PHP, one assumes an a priori probability p_h for every possible haplotype h (inferred, e.g., from genotype frequencies [6]), and seeks the *most*

likely set $\mathcal{H}$ of haplotypes explaining the observed genotypes, where the likelihood of a set $\mathcal{H}$ is given by $L(\mathcal{H}) = \prod_{h \in \mathcal{H}} p_h$. In the special case when all a priori haplotype probabilities are equal, likelihood maximization recovers the *maximum parsimony* approach to PHP [7, 8], in which one seeks the *smallest* set $\mathcal{H}$ of haplotypes explaining $\mathcal{G}$.

The maximum likelihood PHP can be reduced to MWMCSP by associating a vertex of weight $-\log p_h$ to each candidate haplotype h, and adding an edge (h, h') colored by $h+h'$ whenever $h+h'$ is one of the given genotypes. Maximum parsimony PHP reduces to MMCSP in a similar way. Notice that in resulting M(W)MCSP instances each edge is assigned at most one color (in fact, color classes form a matching in the underlying graph). This property is no longer true for the more general versions of PHP in which the input contains missing data [9], i.e., when the input consists of *partial genotypes* which are vectors over the alphabet $\{0, 1, 2, *\}$, and the goal is to resolve each "$*$" symbols into a 0, 1, or a 2, and find a most likely/smallest set of haplotypes that explain the resolved genotypes. We also remark that the reductions of PHP to M(W)MCSP are not polynomial, as the number of haplotypes compatible with the given genotypes may be exponential. Nevertheless, in practice the reductions yield instances of manageable size [7].

1.3 Previous work

Gusfield [7] proposed an (exponential size) integer program formulation for the maximum parsimony PHP. He reports that the commercial integer programming solver CPLEX finds optimal solutions in practical running time for instances with up to 150 genotypes and up to 100 SNPs. For the same problem, Wang and Xu [10] proposed a greedy heuristic and an exact branch and bound algorithm. Lancia et al. [8] proved that maximum parsimony PHP is APX-hard, and gave two straightforward algorithms with approximation factors of $\sqrt{n}$ and q, where n is the number of genotypes and q is the maximum number of haplotype pairs compatible with a genotype. These results immediately imply APX-hardness of M(W)MCSP and M(W)kCSP (even when only one color can be assigned to each edge), and can be shown to yield approximation factors of $\sqrt{n}$ and $\mathfrak{m}$ for MMCSP with one color per edge and MWMCSP, respectively, where $\mathfrak{m}$ is the maximum size of a color class (i.e., the maximum number of edges sharing the same color).

Brown and Harrower [11] and Lancia et al. [12] independently proposed polynomial size integer programs for maximum parsimony PHP. Although these formulations are more compact than the one proposed by Gusfield [7], experimental results in [11] indicate that they often take longer to solve for instances of practical interest, even when augmented with sophisticated sets of valid constraints. This may be explained by the fact that there is no known integrality gap for the formulations in [12] and [11], whereas the results in Section 4, imply an integrality gap of $O(\sqrt{q} \log n)$ for Gusfield's formulation. The formulations in [11] and [12] do not seem to extend to the maximum likelihood PHP problem.

Fernandes and Skiena [1] studied MMCSP with at most one color per edge in the context of multi-use primer selection for synthesis of spotted microarrays.

They gave practical greedy and densest-subgraph based heuristics for the problem and proved, by a direct reduction from set cover, that even this special case of MkCSP cannot be approximated within a factor better than $(1-o(1))\ln n - o(1)$, where n is the number of colors. Konwar et al. [13] introduced a string-pair covering formulation for multiplex PCR primer set selection when there are only amplification length constraints, and proved that in this special case a modification of the classical greedy algorithm for set cover gives an approximation factor of $1 + \ln(nL)$, where L is the upperbound on the amplification length. The algorithm in [13] cannot enforce arbitrary pairwise compatibility constraints, such as ensuring amplification uniqueness.

Very recently, Hassin and Segev [14] showed that a suitable adaptation of the greedy set cover algorithm yields an approximation factor of $O(\sqrt{n \log n})$ for the MMCSP, and Huang et al. [15] gave a factor $O(\log n)$ approximation algorithm for maximum parsimony PHP based on semidefinite programming.

1.4 Our results and techniques

In this paper we give several approximation algorithms and hardness results for MWMCSP and its variants. Unlike approximation factors in [8, 14, 15], our results hold for the weighted version of the problem and do not require the assumption that edges belong to a single color class. Our contributions are as follows:

- First, in Section 2, we present a $\sqrt{k(1 + \ln \Delta)}$ approximation algorithm for MkCSP using an algorithm of Slavik [16] for the partial set cover problem. Here Δ is the maximum number of colors assigned to an edge.
- Then, in Section 3, we present evidence of potential polynomial inapproximability for MkCSP problem by showing a novel reduction from the densest k-subgraph maximization problem to our minimization problem. We believe that our approach can serve as a general technique to reduce hardness from other budgeted graph-theoretic maximization problems to the corresponding minimization problems.
- Finally, in Section 4, we give an $O(\sqrt{\mathfrak{m} \log n})$ approximation algorithm for MWMCSP, where $\mathfrak{m}$ is the maximum size of a color class and n is the number of colors. For PCR primer set selection with arbitrary pairwise compatibility constraints in addition to amplification length constraints, $\mathfrak{m} = O(L^2)$. Thus, reduction to MWMCSP gives an approximation factor of $O(L \log n)$. For maximum likelihood PHP, $\mathfrak{m} = O(2^t)$, where t is the maximum number of 2's in a genotype. Thus, our algorithm yields an approximation factor of $O(2^{t/2}\sqrt{\log n})$ in this case. Our approximation algorithm for MWMCSP is based on LP-rounding, and we show that the approximation factor is almost tight by showing a matching (up to the logarithmic factor) integrality gap for the underlying linear program.

2 Approximation algorithm for MkCSP

Notice that arbitrarily picking a set of k color classes and an arbitrary edge from each color class yields a factor $O(\sqrt{k\Delta})$ approximation for MkCSP, where

Δ denotes the maximum number of colors that can be assigned to an edge. The following theorem gives an improved approximation algorithm.

Theorem 1. *There exists an approximation algorithm with factor* $\sqrt{2kH(\Delta)} = O(\sqrt{k(1+\ln \Delta)})$ *for MkCSP, where* $H(\Delta) = 1 + \frac{1}{2} + \ldots + \frac{1}{\Delta}$.

Proof. The algorithm is as follows. Let X be the set of selected vertices; initially empty. While the number of colors covered is less than k, we choose an edge with maximum number of uncovered colors and add both of its endpoints to X (if they are not already in X). Let i be the number of edges that we choose in this process, clearly $i \leq k$. We know that $|X| \leq 2i$. On the other hand, by a result of Slavik [16], we know that the above greedy algorithm for the partial set cover problem, i.e., finding the minimum number of sets to cover at least k elements, is an $H(\Delta)$ approximation algorithm. This means that the minimum number of edges needed to cover at least k colors is at least $i/H(\Delta)$. It is easy to see that, in order to induce at least $i/H(\Delta)$ edges, the optimum MkCSP solution should pick at least $\sqrt{2i/H(\Delta)}$ vertices. The approximation factor follows immediately by using this lower bound.

Remark. For the case when $k = n$ and $\Delta = 1$, i.e., for MMCSP with one color per edge, the above algorithm corresponds to the $\sqrt{k}$-approximation algorithm of [8]. It is also worth mentioning that using the approximation algorithm of Gandhi, Khuller and Srinivasan [17] for partial set cover in the proof of Theorem 1, we can obtain an $\sqrt{2k\mathfrak{m}}$ approximation algorithm for MkCSP, where $\mathfrak{m}$ is the maximum number of edges sharing the same color. The reduction established in next section suggests that the approximation factor for MkCSP cannot be easily improved.

3 Hardness result for MkCSP

In this section, we show an interesting relation between the approximability of MkCSP and that of the densest k-subgraph problem. Formally, we show that if there is a polynomial time f-approximation algorithm $\mathcal{A}$ for MkCSP, then there is a polynomial time $2f^2$-approximation algorithm for the densest k-subgraph problem. Given a graph G and a parameter k, the densest k-subgraph problem is to find a set of k vertices with maximum number of induced edges. The densest k-subgraph problem is well-studied in the literature [18, 19]. The best known approximation factor for the densest k-subgraph problem is $O(\min\{n^\delta, n/k\}))$ for some $\delta < 1/3$ and improvement seems to be hard [20, 19]. The connection between MkCSP and the densest k-subgraph problem suggests that significant improvements in the approximation ratio for MkCSP would require substantially new ideas.

Theorem 2. *If there is a polynomial time f-approximation algorithm $\mathcal{A}$ for MkCSP, then there is a polynomial time $2f^2$-approximation algorithm for the densest k-subgraph problem.*

Proof. Given a graph G with m edges, we would like to find a set of k vertices with maximum number of edges in the subgraph induced by this set. We assign to each edge of G a different color and use $\mathcal{A}$ to find the approximate solutions for MkCSP on the resulting graph. Suppose l is the maximum color coverage requirement for which $\mathcal{A}$ outputs a solution Y with at most k vertices. That is, there are l colors assigned to the subgraph induced by Y, and the approximate solution returned by $\mathcal{A}$ when $l+1$ colors are required to be covered contains at least $k+1$ vertices. Let the optimal solution to the densest k-subgraph problem contain *opt* edges. We shall prove that $opt \leq 2f^2 l$ and thus Y is a solution to the densest k-subgraph problem which is within a factor of $\frac{1}{2f^2}$ to the optimal solution.

By our choice of l and the fact that $\mathcal{A}$ is an f-approximation algorithm, any $\lfloor \frac{k}{f} \rfloor$ vertices of G can induce at most l colors. Consider a subset X with k vertices. The total number of colors induced by all possible subsets of $\lfloor \frac{k}{f} \rfloor$ elements of X is at most $\binom{k}{\lfloor \frac{k}{f} \rfloor} l$. Notice that each edge is counted exactly $\binom{k-2}{\lfloor \frac{k}{f} \rfloor - 2}$ times. So, the total number of edges in X is at most

$$\frac{\binom{k}{\lfloor \frac{k}{f} \rfloor} l}{\binom{k-2}{\lfloor \frac{k}{f} - 2 \rfloor}} = \frac{k(k-1)}{\lfloor \frac{k}{f} \rfloor \lfloor (\frac{k}{f} - 1) \rfloor} l \leq f^2 l \frac{k(k-1)}{(k-f)(k-2f)} < 2f^2 l.$$

The last inequality holds since we can assume without loss of generality that $k > 4f^2$ (otherwise, any connected subgraph on k vertices is a $2f^2$-approximation), and also that k is a constant such that $\frac{k(k-1)}{(k-f)(k-2f)} < 2$. Since X is an arbitrary set with k vertices, $opt \leq 2f^2 l$ and this completes the proof.

4 LP-rounding based approximation

In this section we give an $O(\sqrt{\mathfrak{m} \log n})$ approximation algorithm for MWMCSP, where $\mathfrak{m}$ is the maximum size of a color class. Our algorithm uses LP-rounding, and we show that the approximation guarantee is matched (up to a logarithmic factor) by the integrality gap of the underlying linear program. Proofs are omitted due to space constraints.

Let $G = (V, E)$ be the input graph and $\mathcal{X} = (\chi_1, \ldots, \chi_n)$ be the family of nonempty "color classes" of edges (without loss of generality we assume that $\bigcup_i \chi_i = E$). We use the following integer program formulation of MWMCSP:

$$\min \sum_v w_v x_v, \text{ subject to } \quad \begin{aligned} &\forall \chi \in \mathcal{X}, \sum_{e \in \chi} y_e \geq 1 \ , \\ &\forall v \in V, \forall \chi \in \mathcal{X}, \sum_{v \in e \in \chi} y_e \leq x_v \ , \\ &\forall e \in E, y_e \geq 0, \forall v \in V, x_v \geq 0 \ . \end{aligned}$$

Here the x_v and y_e are variables associated with the vertices and edges of the graph, and the w_v denote the positive weights given in the problem instance.

Our formulation is related to that introduced by Gusfield [7] for maximum parsimony PHP. Gusfield's formulation lacks weights, and replaces our second set of constraints by the simpler requirement that $y_e \leq x_v$ for every edge e incident to a vertex v. The two sets of constraints are identical for MMCSP instances obtained by reduction from maximum parsimony PHP, since color classes are independent sets of edges in this case. However, using the stronger set of constraints is essential in establishing our approximation guarantee for arbitrary M(W)MCSP instances; a simple example shows that in this case the integrality gap with Gusfield's constraints is $\Omega(\mathfrak{m})$.

For any given weight function $w : V \rightarrow \mathbb{R}^+$ and color classes $\mathcal{X}$, we let $\mathcal{I}(G, w, \mathcal{X})$ denote the optimum value of the MWMCSP integer program, and $\mathcal{I}_\ell(G, w, \mathcal{X})$ denote the optimum value of the linear program obtained by allowing the variables x_v and y_e to take values in $[0, 1]$.

Theorem 3. *MWMCSP can be approximated in polynomial time within an approximation factor of* $O(\sqrt{\mathfrak{m} \log |\mathcal{X}|})$*, where* $\mathfrak{m} = \max_{\chi \in \mathcal{X}} | \chi |$.

Theorem 4. *For every* $s \geq 0$ *there is a pair* $(G, \mathcal{X})$ *for which* $\mathfrak{m} = s$ *and* $\mathcal{I}(G, \mathcal{X}) \geq \Omega(\sqrt{\mathfrak{m}})\mathcal{I}_\ell(G, \mathcal{X})$.

Theorem 4 suggests that the above linear program may have limited value in achieving approximation results beyond the $\sqrt{\mathfrak{m}}$ threshold. It is worth mentioning that the integrality gap in Theorem 4 holds for Gusfield's maximum parsimony PHP formulation [7] as well. As mentioned in Subsection 1.2, in this case, the graph is more restricted, that is, each vertex is a 0/1 vector and each edge between vertices h and h' has a unique color $h + h'$ (which is a 0/1/2 vector). Still we can construct such a restricted graph which shows the integrality gap is the same as that of Theorem 4.

5 Conclusions

In this paper we have proposed the first non-trivial approximation and inapproximability results for the MWMCSP problem and several of its variants capturing important applications in computational biology. Interesting open problems include closing the gap between approximation guarantees and inapproximability results for MWMCSP, and obtaining non-trivial approximations for MWkCSP (an approximation factor of k is obtained, e.g., by picking the lightest edge of each color.) An important constraint on the primers for multiplex PCR not modeled by MWMCSP is that they shouldn't cross-hybridize. This motivates studying the variant of MWMCSP in which certain edges are marked as "forbidden", and the goal is to find a minimum multicolored induced subgraph with no forbidden edges.

Acknowledgments

IIM's work was supported in part by NSF CAREER Award IIS-0546457 and a Faculty Large Research Grant from the University of Connecticut Research Foundation.

References

1. Fernandes, R., Skiena, S.: Microarray synthesis through multiple-use PCR primer design. Bioinformatics **18** (2002) S128–S135
2. Clark, A.: The role of haplotypes in candidate gene studies. Genet. Epid. **27** (2004) 321–333
3. Bonizzoni, P., Vedova, G.D., Dondi, R., Li, J.: The haplotyping problem: An overview of computational models and solutions. Journal of Computer Science and Technology **18** (2003) 675–688
4. Halldorsson, B., Bafna, V., Edwards, N., Lippert, R., Yooseph, S., Istrail, S.: A survey of computational methods for determining haplotypes. In: Proc. of the DIMACS/RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotype Inference (2004) 26–47
5. Niu, T.: Algorithms for inferring haplotypes. Genet. Epid. **27** (2004) 334–347
6. Halperin, E., Hazan, E.: HAPLOFREQ - estimating haplotype frequencies efficiently. In: Proc. 9th Annual International Conference on Research in Computational Molecular Biology (2005) 553–568
7. Gusfield, D.: Haplotyping by pure parsimony. In: Proc. 14th Annual Symp. on Combinatorial Pattern Matching (2003) 144–155
8. Lancia, G., Pinotti, C., Rizzi, R.: Haplotyping populations: complexity and approximations. Technical Report DIT-02-0080, University of Trento (2002)
9. Lin, S., Chakravarti, A., Cutler, D.: Haplotype and Missing Data Inference in Nuclear Families. Genome Res. **14**(8) (2004) 1624–1632
10. Wang, L., Xu, Y.: Haplotype inference by maximum parsimony. Bioinformatics **19** (2003) 1773–1780
11. Brown, D., Harrower, I.: A New Integer Programming Formulation for the Pure Parsimony Problem in Haplotype Analysis. In: Proc. 4th International Workshop on Algorithms in Bioinformatics (2004) 254–265
12. Lancia, G., Pinotti, M., Rizzi, R.: Haplotyping populations by pure parsimony: Complexity of exact and approximation algorithms. INFORMS Journal on Computing **16** (2004) 348–359
13. Konwar, K., Măndoiu, I., Russell, A., Shvartsman, A.: Improved algorithms for multiplex PCR primer set selection with amplification length constraints. In: Proc. 3rd Asia-Pacific Bioinformatics Conference (2005) 41–50
14. Hassin, R., Segev, D.: The set cover with pairs problem. In: Proc. 25th Annual Conference on Foundations of Software Technology and Theoretical Computer Science (2005) 164–176
15. Huang, Y.T., Chao, K.M., Chen, T.: An approximation algorithm for haplotype inference by maximum parsimony. Journal of Computational Biology **12** (2005) 1261–1274
16. Slavik, P.: Improved performance of the greedy algorithm for partial cover. Information Processing Letters **64** (1997) 251–254
17. Gandhi, R., Khuller, S., Srinivasan, A.: Approximation algorithms for partial covering problems. Journal of Algorithms **53** (2004) 55-84
18. Feige, U., Kortsarz, G., Peleg, D.: The dense k-subgraph problem. Algorithmica **29**(3) (2001) 410–421
19. Khot, S.: Ruling out PTAS for graph min-bisection, densest subgraph and bipartite clique. In: Proc. 45th Annual IEEE Symposium on Foundations of Computer Science (2004) 136-145
20. Feige, U.: Relations between average case complexity and approximation complexity. In: Proc. 34th Annual ACM Symposium on Theory of Computing (2002) 534–543