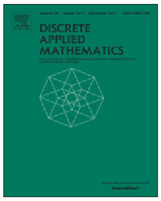
Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam



Hardness and approximation of the asynchronous border minimization problem[☆]

Cindy Y. Li [a], Alexandru Popa [b,c,*], Prudence W.H. Wong [d], Fencol C.C. Yung [d]

[a] NHS Blood and Transplantation, National Health Service (NHS), Bristol, UK
[b] Department of Computer Science, University of Bucharest, Bucharest, Romania
[c] National Institute for Research and Development in Informatics, Bucharest, Romania
[d] Department of Computer Science, University of Liverpool, Liverpool, UK

ARTICLE INFO

Article history:
Received 24 January 2015
Received in revised form 3 May 2017
Accepted 23 August 2017
Available online xxxx

Keywords:
NP-hardness
Approximation algorithms
Border length minimization

ABSTRACT

We study a combinatorial problem arising from the microarray synthesis. The objective of the Border Minimization Problem (BMP) is to place a set of sequences in the array and to find an embedding of these sequences into a common supersequence such that the sum of the "border length" is minimized. A variant of the problem, called P-BMP, is that the placement is given and the concern is simply to find the embedding.

An exponential time algorithm has been proposed for the problem but it is unknown whether the problem is NP-hard or not. In this paper, we give a comprehensive study of different variations of BMP by presenting NP-hardness proofs and approximation algorithms. We show that BMP, P-BMP, and 1D-BMP are all NP-hard and 1D-BMP is polynomial time solvable. The interesting implications include (i) the BMP is NP-hard regardless of the dimension (1D or 2D) of the array; (ii) the array dimension differentiates the complexity of the P-BMP; and (iii) for 1D array, whether placement is given differentiates the complexity of the BMP. Another contribution of the paper is devising approximation algorithms, and in particular, we present a randomized approximation algorithm for BMP with approximation ratio $O(n^{1/4}\log^2 n)$, where n is the total number of sequences.

1. Introduction

In this paper, we study an optimization problem called (asynchronous) border minimization problem (BMP), arising from a biological problem of microarray synthesis. We first describe the BMP (formal definition is given in Section 2) and then explain its relation with the biological problem. The input is a set of sequences $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$. We want to find a common supersequence $\mathcal{D}$ of $\mathcal{S}$ and an embedding ε_i for each sequence s_i into $\mathcal{D}$, where ε_i is obtained by inserting spaces into s_i up to length $|\mathcal{D}|$ with the constraint that the jth position of ε_i is either the character at the jth position of $\mathcal{D}$ or a space. The border length of s_i with respect to s_j is the number of non-space positions of ε_i that are different from ε_j. We then have to "place" the sequences into a $\sqrt{n} \times \sqrt{n}$ array such that the total border length is minimized (the total border length is the

[☆] Preliminary versions of the paper appeared in "Approximating Border Length for DNA Microarray Synthesis" in Proceedings of the 5th Annual Conference on Theory and Applications of Models of Computation, 2008, pp. 410–422 and "Hardness and Approximation of The Asynchronous Border Minimization Problem" in Proceedings of the 9th Annual Conference on Theory and Applications of Models of Computation, 2012, pp. 164–176.

[*] Corresponding author at: Department of Computer Science, University of Bucharest, Bucharest, Romania.

E-mail addresses: Ying.Li@nhsbt.nhs.uk (C.Y. Li), alexandru.popa@fmi.unibuc.ro (A. Popa), pwong@liverpool.ac.uk (P.W.H. Wong), ccyung@graduate.hku.hk (F.C.C. Yung).

http://dx.doi.org/10.1016/j.dam.2017.08.025
0166-218X/© 2017 Elsevier B.V. All rights reserved.

Please cite this article in press as: C.Y. Li, et al., Hardness and approximation of the asynchronous border minimization problem, Discrete Applied Mathematics (2017), http://dx.doi.org/10.1016/j.dam.2017.08.025.

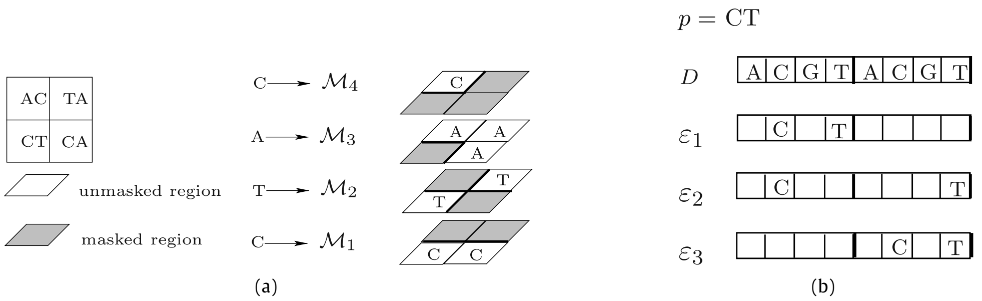


Fig. 1. (a) Asynchronous synthesis of a 2 × 2 microarray with four input sequences AC, TA, CT, CA in the four respective cells (left). The deposition sequence $\mathcal{D}$ = CTAC corresponds to the sequence of four masks $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$ (right). The corresponding embeddings are −−AC, −TA−, CT−−, and C−A−. The masked regions are shaded. The borders between the masked and unmasked regions are represented by bold lines. (b) Different embeddings of the sequence s = CT into deposition sequence $\mathcal{D} = (\text{ACGT})^2$.

sum of the border length between every two sequences that are neighbors in the array). We study the complexity of BMP and give approximation algorithms.

Motivation. DNA and peptide microarrays [8,13] are important research tools used in gene discovery, multi-virus discovery, disease and cancer diagnosis. Apart from measuring the amount of gene expression [29], microarrays are an efficient tool for making a qualitative statement about the presence or absence of biological target sequences in a sample, e.g., peptide microarrays are used for detecting tumor biomarkers [6,25,31]. Microarray design raises a number of challenging combinatorial problems, such as probe selection [17,23,30], deposition sequence design [20,26] and probe placement and synthesis [3–5,15,18,19].

A microarray is a plastic or glass slide consisting of thousands of sequences called *probes*. The synthesis process [12] consists of two components: *probe placement* and *probe embedding*. In the probe placement the goal is to place each probe to a unique array cell. In the probe embedding we want to find a common supersequence of all sequences, called the *deposition sequence*, and a sequence of 2D arrays, called *masks*. The cells of a mask can be either opaque or transparent allowing the deposition of the character associated with the mask. For any cell, concatenating the characters for which the cell is transparent has to be the same as the probe in that cell of the microarray. See Fig. 1(a) for an example. The embedding of a probe placed in a cell c is a sequence in which the ith character is "−" if cell c is opaque in the ith mask, or the ith character of the deposition sequence if transparent (see Fig. 1(b)).

Due to diffraction, the cells on the *border* between the masked and the unmasked regions are often subject to unintended illumination [12], and can compromise experimental results. As the microarray chip is expensive to synthesize, unintended illumination should be minimized. The magnitude of unintended illumination can be measured by the *border length* of the masks used, which is the number of borders shared between masked and unmasked regions, e.g., in Fig. 1(a), the border length of $\mathcal{M}_1$, $\mathcal{M}_3$, $\mathcal{M}_4$ is 2 and $\mathcal{M}_2$ is 4. Note that the sum of the border length of all the masks is the same as the sum of border length as defined by the corresponding embedding (cf. the first paragraph).

In this paper we study the asynchronous synthesis where a mask may deposit a character to different positions of different probes. For example, in Fig. 1(a), we want to synthesize the microarray with the four sequences AC, TA, CT, CA in the respective cells as shown in the left hand side. The right is four masks $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$, where $\mathcal{M}_1$ deposits the character C and there are two transparent cells at the bottom row and two opaque cells at the top of $\mathcal{M}_1$, and so on. This sequence of masks shows an asynchronous synthesis because $\mathcal{M}_2$ deposits the character T to the second position of the sequence CT and the first position of TA (different positions of different probes). On the other hand, in synchronous synthesis, each deposition character can only be deposited to the ith position of the probes for a particular i. The synchronous variant of the problem was first studied [15]. For this problem, if the placement is fixed, the border length is unique and is proportional to the Hamming distance of neighboring probes. Thus the only problem is the placement of the probes. The synchronous version is NP-hard [21], $O(\sqrt{n})$-approximable [22] and there are also some experimental results [4,18,19]. Notice that the NP-hardness of the synchronous BMP [21] does not imply that asynchronous BMP – the problem that we study – is NP-hard.

Previous work on asynchronous BMP. The Asynchronous Border Minimization Problem (BMP) was introduced by Kahng et al. [18]. The problem appears to be difficult as they studied a special case in which the deposition sequence is given and the embeddings of all but one probes are known. A polynomial time dynamic programming algorithm was proposed to compute the optimal embedding of this single probe. This algorithm is used as the basis for several heuristics [3–5,18,19] that are shown experimentally to reduce unintended illumination. The dynamic programming [18] computes the optimal embedding of a single probe in time $O(\ell|\mathcal{D}|)$, where ℓ is the length of a probe and $\mathcal{D}$ is the deposition sequence. The algorithm can be extended to an exponential time algorithm to find the optimal embedding of all n probes in $O(2^n\ell^n|\mathcal{D}|)$ time. It is however unknown whether the general problem is NP-hard or not. This naturally raises a number of questions. Let us denote by P-BMP the problem with placement already given.

Please cite this article in press as: C.Y. Li, et al., Hardness and approximation of the asynchronous border minimization problem, Discrete Applied Mathematics (2017), http://dx.doi.org/10.1016/j.dam.2017.08.025.