

Integrated Design Flow for Universal DNA Tag Arrays

Nisar Hundewale¹, Ion Mandoiu², Claudia Prăjescu², and Alexander Zelikovsky^{1,3}

Keywords: DNA microarray design, design flow, universal tag arrays, tag assignment, probe selection, software tools

1 Introduction

High throughput genomic technologies have revolutionized biomedical sciences, and progress in this area continues at an accelerated pace in response to the increasingly varied needs of biomedical research. Among emerging technologies, one of the most promising is the use of *universal tag arrays (UTA)* [4]. A universal tag array consists of a set of DNA strings called tags, designed such that each tag hybridizes strongly to its antitag (Watson-Crick complement), but does not hybridize to any other antitag. Sample analysis is typically performed by a sequence of hybridization and single-base extension reactions involving reporter probes consisting of application specific primers ligated to antitags. This architecture provides unprecedented assay customization flexibility while maintaining a high degree of multiplexing and low unit cost.

In this poster we describe an integrated flow for designing genomic assays based on universal tag arrays, building upon the DNA microarray design flow presented by Atlas et al. [2] and the integrated probe selection and tag assignment tools presented by Mandoiu et al. [6].

2 Design Flow for UTA-Based Genomic Assays

The steps of the proposed design flow are given in Figure 1. With small modifications, the flow is appropriate for a wide range of genomic analyses, including gene expression, single nucleotide polymorphism (SNP) genotyping, and micro-organism identification via string barcoding [5]. Below we detail the necessary steps and present experimental results for designing an UTA-based assay for studying gene expression in the Herpes B. virus.

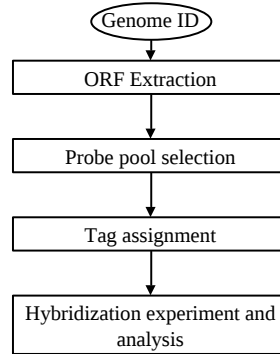


Figure 1: DNA Universal Tag Array Design Flow

Reading Genomic Data and Open Reading Frame (ORF) Extraction: In this step, we use ORF extraction programs to extract the set of ORFs relevant to the application. Because of the differences between prokaryotic and eukaryotic transcription systems there are two types of ORFs. There are two approaches to accomplish this. The first one is ORF-Finder [8]. ORFs can be extracted by means of the genome’s sequence or id using ORF Finder. It uses the prokaryotic approach, there are limitations in gene prediction using ORF finder. A second approach is to

¹Department of Computer Science, Georgia State University, Atlanta, GA 30303. E-mail: nisar@computer.org, alexz@cs.gsu.edu.

²CSE Department, University of Connecticut
371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155. E-mail: {ion.mandoiu,claudia.prajescu}@uconn.edu. Partially supported by a “Large Grant” from the University of Connecticut’s Research Foundation.

³Partially supported by NIH Award 1 P20

Table 1: Tag assignment results.

T_m	# pools	Pool size	Algorithm	1000 tags		2000 tags	
				#arrays	% Util.	#arrays	% Util.
60	1446	1	Pot-Greedy	3	65.35	2	57.05
60	1446	5	Min-deg	3	70.20	2	63.40
67	1560	1	Pot-Greedy	3	69.70	2	61.15
67	1560	5	Min-deg	3	74.90	2	66.15
70	1522	1	Pot-Greedy	3	73.65	2	65.40
70	1522	5	Min-deg	3	76.05	2	69.55

use GenMark [7], which provides identification of protein coding, uses both prokaryotic and eukaryotic; and it has several functions. It uses statistical methods to indicate the true beginning of the ORF and mean coding range of the ORF. GenMark extracts very specific ORFs compared to that of ORF finder.

Probe Pool Selection: The probe pool selection step is responsible for implementing the desired functionality of the DNA array. We use the Promide [9] algorithm to select a large number of possibly overlapping oligonucleotide probes (25-mers in our experiments) from every extracted ORF.² Promide uses a suffix array with additional information to rank all candidate oligos according to their hybridization specificity. It also introduces the concept of master sequences to describe the sequences from which oligos are to be selected. Constraints such as oligo length, melting temperature, and self-complementarity are incorporated in the master sequence at a preprocessing stage and thus kept separate from the main selection problem.

Tag Assignment: In general, it is not possible to assign all tags to primers in an array experiment due to, e.g., unwanted primer-to-tag hybridizations [6]. An assay specific optimization that determines the multiplexing rate (and hence the number of required arrays for a large assay) is the *tag assignment problem*, whereby individual (anti)tags are assigned to each primer. In [6] it has been noted that significant improvements in multiplexing rate can be achieved by combining primer selection with tag assignment. We integrated in our flow both the “potential” greedy deletion algorithm of [3] (which is using only one candidate from each pool) and the pool-aware “min-degree” algorithm in [6].

3 Experimental Results.

We used our flow to design a UTA-based assay for studying the expression of 78 genes in the Herpes B. virus. We varied the prescribed temperature for the hybridization experiment between 60 and 70 degrees Celcius, and selected approximately 20 probes per gene in each case. Table 1 gives the number of arrays and the multiplexing rate (defined as average tag utilization computed over all arrays except the last) for the “potential” greedy deletion algorithm of [3] (which is using only one candidate from each pool) and the pool-aware “min-degree” algorithm in [6].

References

- [1] Affymetrix, Inc. GeneFlex tag array technical note no. 1, available online at http://www.affymetrix.com/support/technical/technotes/genflex_technote.pdf.
- [2] M. Atlas, N. Hundewale, L. Perelygina, and A. Zelikovsky, *Proc. International Conf. of the IEEE Engineering in Medicine and Biology (EMBC'04)*, September 2004, pp. 172–175.
- [3] A. BenDor, T. Hartman, B. Schwikowski, R. Sharan, and Z. Yakhini. Towards optimally multiplexed applications of universal DNA tag systems. In *Proc. 7th Annual International Conference on Research in Computational Molecular Biology*, pages 48–56, 2003.
- [4] S. Brenner. Methods for sorting polynucleotides using oligonucleotide tags. *US Patent 5,604,097*, 1997.
- [5] B. DasGupta, K.M. Konwar, I.I. Mandoiu, and A.A. Shvartsman. Highly scalable algorithms for robust string barcoding, *Proc. 2005 International Workshop on Bioinformatics Research and Applications (IWBR'05)*, 2005.
- [6] I. Mandoiu, C. Prajescu, and D. Trinca, Improved tag set design and multiplexing algorithms for universal arrays, *Proc. 2005 International Workshop on Bioinformatics Research and Applications (IWBR'05)*, 2005.
- [7] M. Bardovsky, *Genemark*, <http://opal.biology.gatech.edu/GeneMark>.
- [8] NIH, *Orf finder*, <http://www.ncbi.nih.gov/gorf/gorf.html>.
- [9] S. Rahmann, Rapid large-scale oligonucleotide selection for microarrays, *Proc. IEEE Computer Society Bioinformatics Conference (CSB'02)*, 2002.

²The probe pool selection step is application dependent, e.g., probe candidates are chosen from the whole genomic sequence in string barcoding applications [5], or immediately preceding the target SNP on the sense or antisense strands in SNP genotyping.