Design of Compact, Universal DNA Microarrays for Protein Binding Microarray Experiments

Anthony A. Philippakis[1,3,4,*], Aaron M. Qureshi[1,5,*], Michael F. Berger[1,4], and Martha L. Bulyk[1,2,3,4,**]

[1] Division of Genetics, Department of Medicine and
[2] Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115
[3] Harvard/MIT Division of Health Sciences and Technology (HST), Cambridge, MA 02138
[4] Harvard University Graduate Biophysics Program, Cambridge, MA 02138
mlbulyk@receptor.med.harvard.edu
[5] Department of Mathematics, University of Maryland, College Park, MD 20742

Abstract. Our group has recently developed a compact, universal protein binding microarray (PBM) that can be used to determine the binding preferences of transcription factors (TFs) [1]. This design represents all possible sequence variants of a given length *k* (i.e., all *k*-mers) on a single array, allowing a complete characterization of the binding specificities of a given TF. Here, we present the mathematical foundations of this design based on de Bruijn sequences generated by linear feedback shift registers. We show that these sequences represent the maximum number of variants for any given set of array dimensions (i.e., number of spots and spot lengths), while also exhibiting desirable pseudo-randomness properties. Moreover, de Bruijn sequences can be selected that represent gapped sequence patterns, further increasing the coverage of the array. This design yields a powerful experimental platform that allows the binding preferences of TFs to be determined with unprecedented resolution.

Keywords: de Bruijn sequences, linear feedback shift registers, protein binding microarrays, motif, transcription factor.

1 Introduction

Detailed knowledge of the DNA binding specificities of TFs is crucial for both genomic studies attempting to map TFs to their target genes [2], as well as biophysical investigations of protein-DNA interactions [3]. Despite the importance of this data type, the binding preferences of the vast majority of TFs remain unknown, largely due to a historical lack of suitable experimental technologies. While chromatin immunoprecipitation (ChIP) experiments [4] (and, more recently, ChIP-chip experiments [5]) give specific examples of sequences bound by a TF *in vivo*, they do not provide an exhaustive characterization of the sequences that a TF can and

[*] These authors contributed equally to this work.
[**] Corresponding author.

T. Speed and H. Huang (Eds.): RECOMB 2007, LNBI 4453, pp. 430–443, 2007.

(just as importantly) cannot bind. Similarly, approaches such as *in vitro* selection [6] typically identify only a limited number of high-affinity binding sites, making a direct quantification of relative binding preferences difficult.

To address this challenge, our group has developed the protein binding microarray (PBM) technology for high-throughput characterization of the *in vitro* binding specificities of protein-DNA interactions [1,7,8]. Briefly, a DNA-binding protein of interest is expressed with an epitope tag, then purified and applied to a double-stranded DNA microarray. The washed, protein-bound microarray is labeled with a fluorophore-conjugated anti-GST antibody. By scanning the array, quantitative information is generated regarding the preferences of the TF for each of the sequences on the array. Prior work by our group and others has demonstrated that this is an effective technology that allows rapid and high-quality determination of the DNA binding specificities of TFs [1,7-10].

A limitation of previous PBM studies, however, has been the lack of a universal array that can be used for the majority of TFs, regardless of their structural class or genome of origin. Earlier studies have utilized either microarrays containing a limited number of binding site variants chosen for the TF under consideration [7,9], or large genomic fragments obtained from the same genome as the TF (specifically, *S. cerevisiae*) [8]. The former approach has the twofold disadvantage of requiring a new microarray for each additional TF assayed and also requiring some *a priori* knowledge of the DNA binding specificities of the TF; the latter approach suffers from the limitation that longer sequences can contain several binding sites for a given TF, making it difficult to acquire quantitative information on protein-DNA interactions. Thus, a single microarray is desired that represents all possible binding sites of a given width k (i.e., all k-mers), in order to provide a complete survey of all candidate binding sites.

Our group has recently developed such a universal array [1]. The key to our design is two-fold. First, we have selected our double-stranded DNA probes to have a length (L) significantly longer than the motif widths (k) that we intend to inspect, so that each spot contains L-k+1 potential binding sites of width k. For a microarray composed of N spots, this increases the total number of k-mers represented from N (in the naïve construction where there is one k-mer per spot, as has been previously utilized [10]) to $N(L-k+1)$. Second, we have designed these spots to completely cover all k-mer sequence variants, so that a maximal number of distinct k-mers are represented. Consider the circular sequence shown in **Fig. 1A** that contains all 16 2-mer variants exactly once. Such sequences containing all 4^k overlapping k-mers one time are named de Bruijn sequences [11,12] of order k, and the spots of our universal array are obtained by computationally segmenting appropriately chosen de Bruijn sequences, leaving an overlap between adjacent sequences in order to not omit any k-mers. With this design, we are able to represent a maximal number of sequence variants in a minimum amount of sequence.

The implementation of this design, along with generated data for five TFs, has been presented in the work of Berger *et al* [1]. Here, we give an exposition of the underlying combinatorial and algebraic theory utilized in designing the array. Specifically, we provide a mathematical treatment of 1) the motivation for and utilization of linear feedback shift registers (LFSRs) to generate de Bruijn sequences; 2) theoretical developments made by our group in order to design de Bruijn sequences that not only contain contiguous k-mers, but also k-mers with biologically relevant