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

Anthony A. Philippakis[1,3,4,6], Aaron M. Qureshi[1,5,6], Michael F. Berger[1,4], 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. [5] Department of Mathematics, University of Maryland, College Park, MD 20742. [6] These authors contributed equally to this work.

Correspondence should be addressed to mlbulyk@receptor.med.harvard.edu

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 (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

gaps; 3) methods for selecting de Bruijn sequences that are optimized for determining TF binding site motifs that are wider than the order of the utilized de Bruijn sequence; 4) the utilization of complementary, independently generated de Bruijn sequences for use in replicate PBM experiments. Finally, we note that de Bruijn sequences have previously been utilized in cryptography [13,14], random number generation [13,14] and the design of tags for DNA microarrays [15]. Recently, another group has independently suggested the use of de Bruijn sequences for use in PBM experiments, although that work did not consider the coverage of gapped *k*-mers and did not utilize LFSRs [16]. We hope that this work will be useful to individuals either seeking to design sequences for PBM experiments or analyzing data generated by a PBM utilizing de Bruijn sequences. Additionally, we hope that the mathematical methods developed for this application will lead to other, un-anticipated biological applications.

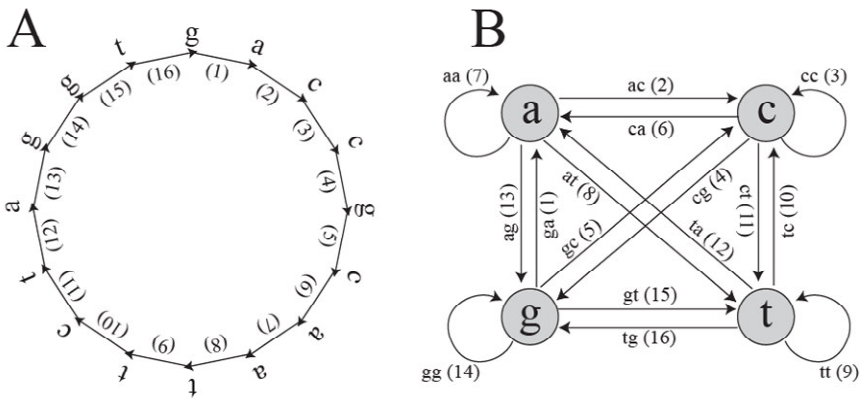

Fig. 1. De Bruijn sequence of order 2 (A) and its associated de Bruijn graph (B).

2 Results

2.1 LFSRs and the Generation of de Bruijn Sequences

For any alphabet Σ of size $|\Sigma|$ and any word length k, there exist sequences $S = \left(s_1 s_2 \ldots s_{|\Sigma|^k + k - 1}\right)$ that are circular (i.e., $s_{|\Sigma|^k+1} \ldots s_{|\Sigma|^k+k-1} = s_1 .. s_{k-1}$) and of length $|\Sigma|^k$ containing all k-mers exactly once when words are considered in a stacked fashion. Such sequences are known as de Bruijn sequences of order k, and their existence can be confirmed by considering the directed graph whose vertices are all k-1-mers and whose edges are all k-mers, where two vertices are connected by an edge if the last k-2 letters of the first vertex are identical to the first k-2 letters of the second. **Fig. 1B** gives an example of such a graph (often referred to as a "de Bruijn

graph" [12]), and we note that graphs of this form have previously been applied to the analysis of repetitive DNA [17] and sequence alignment [18]. Observe that a de Bruijn sequence is equivalent to a walk on this directed graph that traverses every edge (i.e., an Eulerian tour [12]). Since the number of edges going into each vertex is equal to the number of edges that exit it, Euler's theorem guarantees that such paths exist [12]. Indeed, for a given choice of $|\Sigma|$ and k, the number of paths is $\frac{\left(|\Sigma|!\right)^{|\Sigma|^{k-1}}}{|\Sigma|^k}$ [12]; for example, for $|\Sigma|=4$ (i.e., the DNA alphabet) and $k=9$, the number of de Bruijn sequences is greater than $10^{90,000}$.

De Bruijn sequences contain a maximum density of sequence variants, as they contain all distinct k-mers within a sequence of minimum length. Moreover, for any $m>k$ the $|\Sigma|^k$ sequences of length m represented in the de Bruijn sequence will all be distinct; thus, although not all m-mers are represented on an order k de Bruijn sequence, as many distinct m-mers as possible are represented within the given sequence length. Similarly, for all $m'< k$, each m'-mer is represented exactly $|\Sigma|^{k-m'}$ times, insuring that the sampling of m'-mers is uniform.

Clearly, the regularity and variability of de Bruijn sequences makes them a promising tool for designing a universal PBM. An especially facile method of generating such sequences when $|\Sigma| = p^n$, for p a prime and n any integer, is through the use of *linear feedback shift registers* (LFSRs) [13,14]. As background, recall that a Galois field GF(p^n) is a set containing p^n elements that is closed over the multiplication and addition $\{\times, +\}$ operations (one can show that such operations can be suitably defined if and only if the field contains a prime power of elements [19]). For example, **Fig. 2** gives multiplication and addition tables over GF(4) = {0, 1, α, α+1}

+	0	1	α	α+1
0	0	1	α	α+1
1	1	0	α+1	α
α	α	α+1	0	1
α+1	α+1	α	1	0

x	0	1	α	α+1
0	0	0	0	0
1	0	1	α	α+1
α	0	α	α+1	1
α+1	0	α +1	1	α

Fig. 2. Addition and multiplication tables over GF(4)

In order to construct a de Bruijn sequence of order k over the alphabet Σ, take an arbitrary embedding of the alphabet into GF($|\Sigma|$), and consider the following recursive linear equation for generating the i'th element of a sequence $S = \left(s_1 s_2 \ldots s_{\Sigma^k + k - 1}\right)$

from the previous k elements, where arithmetic is performed over GF($|\Sigma|$):

$$s_i = \theta_{k-1} s_{i-1} + \theta_{k-2} s_{i-2} + ... + \theta_0 s_{i-k} \quad . \tag{1}$$

If the coefficients $\theta_i \in$ GF($|\Sigma|$) are chosen so that the polynomial $\sum_{i=0}^{k-1} \theta_i x^i$ is primitive [19], one can demonstrate [14] that the sequence S generated by this recursive equation has periodicity $|\Sigma|^k$-1 and contains every k-mer in GF($|\Sigma|$) except the sequence of k consecutive 0's (this word can be easily included to make S a true de Bruijn sequence by inserting an additional 0 into any of the subsequences of k-1 0*'s* appearing in S). Moreover, S will exhibit the following three properties characteristic of pseudo-random sequences [13,14]:

1) Balance: The number of occurrences in S of each letter differs by no more than 1
2) Low autocorrelation: There is low correlation between pairs of letters separated by a distance j, for any j.
3) Proportional runs: The number of j consecutive occurrences of the same letter ω is n^{k-j} if $\omega \neq 0$ and n^{k-j}-1 if $\omega = 0$.

Thus, de Bruijn sequences generated by LFSRs resemble random sequence, an advantageous property as it guarantees that any trends observed in the data are not an artifact of the method of sequence generation. Moreover, unlike random sequence, LFSRs represent a maximal number of sequence variants (a truly random sequence of equivalent length would represent only $1\text{-}e^{-1} \approx 63\%$ of k-mers on average [1]). Since the DNA alphabet contains a prime power of elements ($4=2^2$), LFSRs are available for use in generating de Bruijn sequences. Indeed, there are (at least) two approaches for using LFSRs to generate de Bruijn sequences over the DNA alphabet. In the first and more natural approach, one takes an arbitrary embedding of {a, c, g, t} into GF(4) = {0, 1, α, α+1} where $\alpha^2=\alpha+1$, and one then picks a primitive polynomial of degree k over GF(4) to use as a LFSR generating a sequence of length 4^k-1. This is schematized in **Fig. 2A**, using the embedding {$a \leftrightarrow 0$, $c \leftrightarrow 1$, $g \leftrightarrow \alpha$, $t \leftrightarrow \alpha+1$} (again, under this embedding the generated sequences do not contain the sequence of k consecutive a's). Alternatively, one can pick a polynomial of degree $2k$ over GF(2)={0,1} and use it to generate a sequence of length 2^{2k}-1 over the 0-1 alphabet. Here, one associates each element of the DNA alphabet with a pair of elements in GF(2), and one must traverse this sequence over GF(2) twice, considering both reading frames. This is schematized in **Fig. 2B**, where we have used the embedding {$a \leftrightarrow 00$, $c \leftrightarrow 10$, $g \leftrightarrow 01$, $t \leftrightarrow 11$}. Henceforth, we shall refer to the embeddings {$a \leftrightarrow 0$, $c \leftrightarrow 1$, $g \leftrightarrow \alpha$, $t \leftrightarrow \alpha+1$} and {$a \leftrightarrow 00$, $c \leftrightarrow 10$, $g \leftrightarrow 01$, $t \leftrightarrow 11$} of the DNA alphabet into GF(4) and GF(2), respectively, as the *standard embeddings* (note that both methods of utilizing LFSRs to generate de Bruijn sequences can be generalized to arbitrary number fields with a prime power of elements). In the next section, we show that de Bruijn sequences generated by primitive polynomials over GF(2) and GF(4) actually behave differently with respect to the coverage of gapped k-mers.

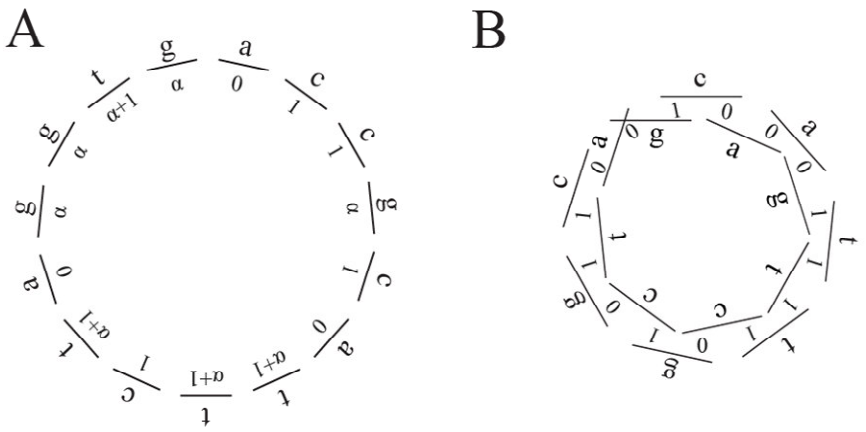

Fig. 3. Generation of de Bruijn sequences over (**A**) 4-letter and (**B**) 2-letter alphabets.

Our basic design, then, is to utilize LFSRs to generate de Bruijn sequences of order *k*, where *k* is as large as possible for a given set of array dimensions and spot lengths. The generated de Bruijn sequence is then computationally segmented into shorter sequences of length *l* corresponding to spots on the array, leaving an overlap of *k*-1 letters between adjacent spots so as not to omit any *k*-mers. For example, consider an array composed of spots of length 30; then all 9-mers could be represented using fewer than 12,000 spots, well within the range of current array dimensions. Such an array would also contain nearly 1/4 of all 10-mers, 1/16 of all 11-mers, etc., and thus could be expected to yield substantial information about TFs having motif widths greater than 9.

2.2 LFSRs and the Coverage of Spaced Seeds

In Berger *et al* [1], we performed a survey of known TFBS motifs in order to determine what value of *k* is required in order to design an array suitable for most TFs. There, we observed that a majority of motifs contained <=9 informative positions. We also observed, however, that for nearly 25% of motifs, the pattern of informative positions was not contiguous and contained one or more gaps (i.e., positions with ≤0.3 bits of information when using the standard position-weight-matrix representation [20]). While de Bruijn sequences of order *k* will, by definition, contain all contiguous *k*-mers, they do not necessarily contain all gapped *k*-mers. Therefore, we inspected the representation of gapped *k*-mers in de Bruijn sequences.

We define a *seed* to be the set of all possible words over the DNA alphabet with a given (possibly gapped) pattern of positions, and we represent the seed with a 0-1 string where 1's are placed at the informative positions. For example, the seed "11" corresponds to the set {*aa, ac, …, tg,tt*} that contains all contiguous 2-mers, and the seed "1001" corresponds to the set {*a(2)a, a(2)c, …, t(2)g, t(2)t}* where the numbers

in parentheses denote the gap size. Words with gaps will be said to be *elements of spaced seeds*, and those without gaps will be said to be *elements of contiguous seeds*. We shall use the variable ζ to represent an arbitrary seed and, for a seed ζ containing k informative positions, we say the *order* of ζ is k. Finally, a given LFSR L is said to *cover* a seed ζ if all its elements except the string composed of all a's (*e.g.*, *aa* and *a(2)a* for the order 2 seeds 11 and 1001, respectively) appear in the sequence generated by L (the reasons for ignoring the elements composed of only the letter a will be explained shortly). Similarly, we shall refer to the *coverage* of ζ by L with the variable $\chi(L, \zeta)$, which takes the value of 1 if the seed ζ is covered by the LFSR and 0 otherwise (again ignoring the element composed of only the letter a).

For a given sequence S over $\{a, c, g, t\}$, let $\mathcal{A}_{k,S}$ denote the set of all (potentially gapped) subsequences of k a's that occur in S; for example, in the sequence shown in **Fig. 2A** $\mathcal{A}_{2,S} = \{a(4)a, a(9)a\}$, and in **Fig. 2B** $\mathcal{A}_{2,S} = \{a(1)a, a(5)a, a(6)a, a(7)a, a(8)a, a(12)a\}$. For ζ a seed of order k and S a sequence generated by a LFSR over GF(q), where q equals either 2 or 4, one can demonstrate the following facts concerning the coverage of spaced seeds by LFSRs.

Proposition 1: a) Under the standard embeddings of the DNA alphabet, ζ is covered by S if and only if $\mathcal{A}_{k,S} \cap \zeta = \varnothing$. **b)** There exists a $j \in \mathbb{N}$ such that every element of ζ not in $\mathcal{A}_{k,S}$ occurs either 0 times or exactly q^j times in S. Also, the element of $\mathcal{A}_{k,S}$ in ζ occurs q^j-1 times.

Proof: Consider the case where q=4. Because our sequence $S = (s_1 s_2 s_3 \ldots)$ is generated by a LFSR, we know that for any i

$$s_i = \theta_{k-1} s_{i-1} + \theta_{k-2} s_{i-2} + \ldots + \theta_0 s_{i-k} \ . \tag{2}$$

Given values of i and m where $m \geq k$, let $\left(s_{i,m}\right)$ denote the subsequence in S of m letters beginning at the letter s_i; also, let this same notation denote the vector of dimension m over GF(q) $\left(s_{i,m}\right) = (s_{i+m-1}, s_{i+m-2}, \ldots, s_i)$. Consider the matrix

$$A = \begin{bmatrix} \theta_{k-1} & \theta_{k-2} & \theta_{k-3} & \cdots & \theta_1 & \theta_0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

It is clear that for any i, $\left(s_{i+1,k}\right) = A\left(s_{i,k}\right)$ and, by induction, for any $j \geq 0$ $\left(s_{i+j,k}\right) = A^j\left(s_{i,k}\right)$. Also, observe that for any $m \geq k$, $\left(s_{i,m}\right)$ can be constructed from $\left(s_{i,k}\right)$ by applying the $m \times k$ matrix

$$\tilde{A}(m)=\begin{bmatrix} \left(A^{m-k}\right)_{1,1} & \left(A^{m-k}\right)_{1,2} & \cdots & \left(A^{m-k}\right)_{1,k-1} & \left(A^{m-k}\right)_{1,k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \left(A^{2}\right)_{1,1} & \left(A^{2}\right)_{1,2} & \cdots & \left(A^{2}\right)_{1,k-1} & \left(A^{2}\right)_{1,k} \\ \theta_{k-1} & \theta_{k-2} & \cdots & \theta_{1} & \theta_{0} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

as $\left(s_{i,m}\right)=\tilde{A}(m)\left(s_{i,k}\right)$ (note that in $\tilde{A}(m)$, the entries $\left(A^{n}\right)_{i,j}$ refer to the matrix element in row i and column j in the n'th power of A). Consider a seed ζ having width m and containing k informative positions. Let $\tilde{A}(m,k)$ be the $k\times k$ sub-matrix of $\tilde{A}(m)$ when restricting to only those rows corresponding to the informative positions of ζ. Consider the set $\left\{\tilde{A}(m,k)\left(s_{i,k}\right)\middle|1\le i<4^{k}-1\right\}$ (i.e., the set of elements of ζ that occur in S). $\tilde{A}(m,k)$ is either invertible or it is not. If $\tilde{A}(m,k)$ is invertible, then its image is all 4^k elements of ζ, and every element of the seed occurs in the LFSR with the exception of the sequence that contains a 0 (equivalently, an *"a"* under the standard embedding) at every informative position, as the kernel of $\tilde{A}(m,k)$ is trivial. Thus, ζ will be covered if the sequence with a's at the informative positions of the spaced seed (which is an element of $\mathcal{A}_{k,S}$) does not appear in S. Similarly, this argument is reversible and so the converse holds; thus, **Prop. 1a** is proven. If $\tilde{A}(m,k)$ is not invertible, then its kernel will contain 4^j elements for $j\in\mathbb{N}$, its image will contain 4^{k-j} elements, and each of these elements will be the image of 4^j vectors $\left(s_{i,k}\right)$. Since every contiguous k-mer $\left(s_{i,k}\right)$ except the sequence of k consecutive a's occurs in S, **Prop. 1b** holds and the proof is completed for the case q=4. The proof for q=2 is nearly identical. Now, however, our matrix A will have dimension $2k\times 2k$. Note that here, the kernel for the matrix analogous to $\tilde{A}(m,k)$ will contain 2^j elements for some j.□

Thus, the spaced seeds that are missed correspond exactly to gapped patterns of a's occurring within the LFSR and, for any spaced seed, the fraction of elements that are represented will be approximately either 2^{-j} when using a polynomial over GF(2), or 4^{-j} when using a polynomial over GF(4). We inspected the coverage of seeds most likely to correspond to known motifs for LFSRs over GF(2) and GF(4), in order to see if some polynomials empirically provided better coverage than others. Here, it is known that the number of primitive polynomials of degree k over a field with q

elements is given by the formula $\frac{\phi(q^k-1)}{k}$ where ϕ is Euler's totient function [21] that returns the number of integers relatively prime to the input value [14,19]; also it is easily seen that the number of spaced seeds of width up to m and order k is given by the formula $\sum_{i=k}^{m}\binom{i-2}{k-2}$.

Because we could not see how to determine $\mathcal{A}_{k,S}$ (and thus the set of covered seeds) for a given LFSR other than by explicit computation, we focused our analysis on the 7776 polynomials over GF(2) of order 18 and the 15,552 polynomials over GF(4) of order 9. For each of the de Bruijn sequences generated by these polynomials, we inspected whether each of the 44 seeds of widths $9 \le m \le 11$ and order $k=9$ was covered. For a given LFSR L, let $C(L)=\frac{1}{44}\sum_{\zeta}\chi(L,\zeta)$ (i.e., the average coverage), where the summation is over all of the spaced seeds ζ with widths between 9 and 11. We were pleased to observe that, by a judicious choice of LFSR, it is possible to completely cover over ~86% (38/44) of these seeds when considering polynomials over GF(4) and ~82% (36/44) of these polynomials over GF(2). Also, the mean coverage of polynomials over GF(4) was ~74±12%, significantly higher than average coverage of ~44±12% for polynomials over GF(2).

2.3 Sampling *k*-mers Larger than the Order of the de Bruijn Sequence

In this section, we demonstrate that the representation of spaced seeds is connected to the uniform sampling of words longer than the order of the shift register. As stated previously, the fraction of m-mers represented in an order k de Bruijn sequence is 4^{k-m} (where $m \ge k$). In this section, we demonstrate that if the sequence covers all spaced seeds of width $\le m$ and order k, then the sampled m-mers are well-spaced and regularly sampled (this will be made precise momentarily), facilitating interpolation to m-mers not represented on the array. Thus, a suitable selection of de Bruijn sequence to cover spaced seeds is valuable to determining TFBSs whose width is greater than the order of the generating de Bruijn sequence.

Let d be the Hamming metric on words of length k over the DNA alphabet [21] (*i.e.*, the metric that counts the number of mismatches between pairs of words). For a de Bruijn sequence of order k, let m be an integer such that $m>k$. We say that the sampling of m-mers is *m,k-spaced* if for each word w of width m occurring in the de Bruijn sequence, there does not exist a distinct word w' in the de Bruijn sequence such that $d(w,w') \le m-k$. Also, we say that the sampling of m-mers is *m,k-equidistant* if 1) for any choice of k-1 positions in w there exists a w' occurring in the de Bruijn sequence that agrees with w at these k-1 positions and such that $d(w,w')=m-k+1$, and 2) the number of words w' appearing in the de Bruijn sequence such that $d(w,w')=m-k+1$ is constant over the choice of w.

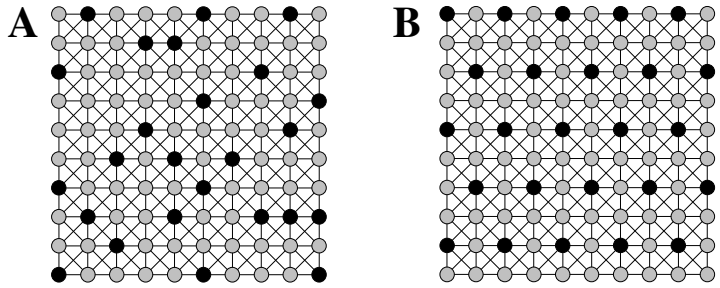

Fig. 4: Cartoon depicting all *m*-mers (grey vertices) and *m*-mers sampled by an order $k<m$ de Bruijn sequence (black vertices). Vertices are connected by an edge if they are 1 mismatch away. (**A**) de Bruijn sequence that samples *m*-mers randomly. (**B**) de Bruijn sequence that samples *m*-mers regularly.

Intuitively, *m*-mers are regularly sampled if they are *m,k*-spaced and *m,k*-equidistant. This is cartooned by the graphs in **Fig. 4**, where nodes represent the 4^m possible *m*-mers, and the black nodes represent the 4^k *m*-mers that are represented within a given de Bruijn sequence of order *k*. In this graph, two *m*-mers are adjacent in the graph if they differ at only one position. A randomly chosen de Bruijn sequence will sample a random collection of *m*-mers (**Fig. 4A**), yet an auspiciously chosen de Bruijn sequence (i.e., *m,k*-spaced and *m,k*-equidistant) will regularly sample *m*-mers (**Fig. 4B**).

One can then prove the following two propositions regarding *m,k*-spacing and *m,k*-equidistance. We note that they apply to general de Bruijn sequences, and not only those de Bruijn sequences generated by an LFSR.

Proposition 2: The sampling of *m*-mers is *m,k*-spaced in an order *k* de Bruijn sequence if and only if all spaced seeds of width $m' \leq m$ and order *k* are covered.

Proof: Assume that the de Bruijn sequence covers all spaced seeds of width *m'* and order *k*. Assume (for contradiction) that there are words *w* and *w'* such that $d(w,w') \leq m-k$; then *w* and *w'* will agree at at least *k* letters. Consider the spaced seed that contains 1's at the positions where *w* and *w'* agree, and let *m'* be the distance between the first and last 1 in this spaced seed (note that *m'* may be less than *m* if *w* and *w'* do not agree at the first or last positions). Then *w* and *w'* will map to the same element of this spaced seed, and so the seed cannot be covered by the pigeonhole principle, giving a contradiction. Conversely, assume that a given de Bruijn sequence is *m,k*-spaced. Let ζ be a spaced seed of width $m' \leq m$ and order *k*. Every element of ζ that appears in the de Bruijn sequence must occur only once. To see this, assume (for contradiction) that there is some element of ζ that occurs more than once. Then there are *m*-mers *w* and *w'* appearing in the de Bruijn sequence that agree at the *k* informative positions of the spaced seed. Then $\tilde{d}(w,w') \leq m-k$, in violation of our assumption that the

sampling of *m*-mers is *m,k*-spaced. Thus, ζ must be covered since the number of its elements that occur in the de Bruijn sequence is 4^k, all of which are distinct. □

Proposition 3: If the sampling of *m*-mers is *m',k*-spaced for all $k \le m' \le m$, it is *m,k*- equidistant.

Proof: For any $m' \le m$, we know that all spaced seeds of width *m'* and order *k* are covered, by **Prop. 2**. Let *w* be an *m*-mer and pick any *k*-1 informative positions in *w*. Since all spaced seeds of width $m' \le m$ and order *k* are covered, there will be exactly three distinct *m*-mers *w'* such that $w' \neq w$ and that occur in the de Bruijn sequence and agree with *w* at these *k*-1 informative positions (call this set *W*). The elements of *W* will all be at a distance of $\tilde{d}(w,w') = m-k+1$ (so condition 1 is satisfied). Also, take a different choice of *k*-1 informative positions in *w*, and consider the set of three words *W'* agreeing with *w* at these *k*-1 informative positions. *W* and *W'* must be disjoint, since if there is a word in common between them, then it would agree with *w* at at least *k* informative positions, and then the de Bruijn sequence could not cover all spaced seeds of width $m' \le m$ and order *k*. This implies that every element has a constant number of *m*-mers at a distance of *m-k*+1, and so condition 2 holds. □

Finally, for the special case of *m=k+1*, one can state the following proposition giving analytic conditions relating *m,k*-spacing and *m,k*-equidistance to the choice of polynomial used for the LFSR:

Proposition 4: a) A de Bruijn sequence of order *k* generated by a LFSR over GF(4) is *k+1,k-spaced* and *k+1,k*-equidistant if and only if none of the coefficients θ_i of the generating polynomial $\sum_{i=0}^{k-1} \theta_i x^i$ are equal to 0. **b)** A de Bruijn sequence of order *k* generated by a LFSR over GF(2) is *k+1,k-spaced* and *k+1,k*-equidistant if and only if it does not contain three consecutive coefficients (θ_i, θ_{i+1}, θ_{i+2}) for even *i* such that $\theta_i \theta_{i+2} + \theta_{i+1}^2 = 0$.

Proof: Assume the case where the de Bruijn sequence is generated by the polynomial $\sum_{i=0}^{k-1} \theta_i x^i$ over GF(4). By **Props. 2** and **3** it is sufficient to prove that all seeds of width *k*+1 and order *k* are covered if and only if all θ_i are non-zero. Coverage of order *k* width *k*+1 seeds is equivalent to asserting that for any *k*+1-mer ($s_1,\ldots s_{k+1}$) and any $1 \le i \le k$, there exists a value $\tilde{s}_i$ such that $\sum_{j=0}^{i-1} \theta_j s_j + \theta_i \tilde{s}_i + \sum_{j=i+1}^{k-1} \theta_j s_j = s_k$. Clearly, this can occur if and only if for all *i* $\theta_i \neq 0$. The proof for polynomials over GF(2) is nearly identical, but involves solving two such equations simultaneously. □

2.4 Complementary de Bruijn Sequences and Replicate Experiments

An additional advantage of our design is that, for any given value of k and desired set of represented gapped k-mers, if one acceptable de Bruijn sequence exists, there will in general be several that could be used (this is easily seen by, for example, permuting the letters or taking the reversal of the de Bruijn sequence). In Berger *et al* [1], we exploited this fact by doing replicate experiments on *distinct* de Bruijn sequences, both of which were *11,10-spaced* and *11,10-equidistant* (i.e., they covered all 10-mers containing a single gapped position). There, we observed that performing replicate experiments on distinct de Bruijn sequences, rather than the same de Bruijn sequence, improved our ability to correctly quantify the binding preferences of the well-characterized TF Zif268. We anticipate that this approach of performing replicate experiments on distinct de Bruijn sequences will be a valuable means for improving PBM experiments. In this section, we inspect some formal aspects of this experimental strategy.

The following proposition implies that all pairs of order k de Bruijn sequences generated by LFSRs will share a constant number of $k+1$-mers.

> **Proposition 5**: Let S and S' be two de Bruijn sequences of order k, both generated by an LFSR over either GF(2) or GF(4). Then exactly 4^{k-1} k+1-mers will be commonly represented on both S and S'.
>
> **Proof:** Assume that S and S' are generated by the polynomials $\sum_{i=0}^{k-1}\theta_i x^i$ and $\sum_{i=0}^{k-1}\theta'_i x^i$, respectively, over GF(4). Then S and S' will share a k+1-mer $(s_{i+k+1},\ldots s_i)$ if and only if $\left\langle \Theta, \tilde{S} \right\rangle = \left\langle \Theta', \tilde{S} \right\rangle \Leftrightarrow \left\langle \Theta - \Theta', \tilde{S} \right\rangle = 0$, where $\Theta=(\theta_{k-1},\ldots,\theta_0)$, $\Theta'=(\theta'_{k-1},\ldots,\theta'_0)$ and $\tilde{S} = (s_{i+k},\ldots,s_i)$. Since the null space of a linear form must always be of dimension k-1, there will be exactly 4^{k-1} values that satisfy this equation. The proof for GF(2) is nearly identical, but involves finding the null space for two linear forms simultaneously. □

Thus, it is not in general possible to optimize the coverage of words longer than the order of the de Bruijn sequences in performing replicate experiments, as the number of k+1-mers represented on at least one of the two de Bruijn sequences will always be $2*4^k-4^{k-1}$. Note that **Prop. 5** also answers a natural question regarding the selection of the optimal order (k) to use for a given set of array dimensions (either on a single array or multiple arrays). It is not immediately clear whether it is better to have 4 different de Bruijn sequences of order 4^{k-1} or one de Bruijn sequence of order 4^k, as each requires an equal number of spots of the same length. **Prop. 5** implies that a de Bruijn sequence of order 4^k is preferable, as de Bruijn sequences of order 4^{k-1} will have overlap with respect to the k-mers that they represent.

Finally, we note that, although it does not seem that complementary order k primitive polynomials can be utilized in order to maximize the coverage of m-mers, $m>k$, we have observed that suitable sets of complementary polynomials can be selected for the coverage of gapped k-mers. Here, we have found by empirical

inspection that if one polynomial misses a given spaced seed, then another polynomial can be identified that covers it. Thus, this parameter can be optimized.

2.5 Open Questions

We see (at least) three broad areas in which further mathematical/algorithmic efforts could lead to improvements in array design. First, assuming the use of LFSRs for generating de Bruijn sequences, there is need for an improved mathematical theory relating the coverage of spaced seeds to the generating polynomial. In this work, we have presented an explicit formula for determining whether a given polynomial represents all *k*-mers with a single gapped position (i.e., *k+1,k*-spaced and *k+1,k*-equidistant de Bruijn sequences), but we have not yet been able to extend this theory to *k*-mers with multiple gaps.

Second, only a small fraction of de Bruijn sequences correspond to sequences generated by an LFSR, and the utility of such non-LFSR-generated de Bruijn sequences remains largely unexplored. In current applications we have utilized LFSRs as they provably satisfy pseudo-randomness properties that are advantageous, since they guarantee that there are no confounding correlations in the experimental data that are an artifact of the methods utilized to generate the de Bruijn sequences. Additionally, LFSRs allow for the complete coverage of certain gapped *k*-mer patterns, which we have observed to be useful for determination of the binding specificities of TFs. However, it may well be the case that there are additional families of de Bruijn sequences that cover even more gapped *k*-mers while still resembling random sequence. Similarly, there may be additional desirable properties of de Bruin sequences that we have not yet considered, and for which LFSRs might not be optimal.

Finally, when considering protein-DNA interactions, it is often a reasonable assumption to identify *k*-mers and their reverse complements, as this symmetry is present in double-stranded DNA. In the case of PBM experiments in particular, this is a reasonable assumption to make if the DNAs are randomly fixed to the slide (although it is debatable whether or not this is appropriate in the case of end-attached DNA, as was the case for the arrays utilized in Berger *et al* [1]). The work presented here could be extended to generate de Bruijn sequences modulo reverse complements (i.e., sequences where only the word or its reverse complement is present, but not both) as was done in a related work [16]. If such sequences could be generated that had the desired pseudo-randomness properties as well as coverage of gapped *k*-mers, then their utilization might be advantageous.

3 Concluding Remarks

We have presented the combinatorial design of a protein binding microarray. Importantly, this design has been optimized in several key aspects: 1) All *k*-mers are represented in a minimum amount of sequence, permitting a maximum number of binding site variants to be represented in a cost-efficient manner. This allows the binding specificities of TFs to be assayed with word-by-word resolution. 2) The

unbiased nature of the construction provides a design that can be used for TFs from any species and/or structural class, making it a universal platform. 3) Our design is flexible, allowing ever greater binding site coverage as array technology improves and feature density increases. For example, all 11-mers can already be represented with Agilent arrays [1], and all 12-mers with NimbleGen technology [22]; moreover, this number is expected to grow quickly. Similarly, our design allows replicate experiments to be performed with distinct de Bruijn sequences, resulting in reduced experimental noise and greater coverage of sequence space. 4) We have utilized de Bruijn sequences generated by LFSRs which provably resemble random sequence. This guarantees that any statistical trends observed in data generated by a PBM experiment are not an artifact of how the sequences were constructed. 5) Our design not only maximizes the coverage of contiguous *k*-mers, but also gapped *k*-mers. This simultaneously allows an interrogation of the binding specificities of TFs with gapped motifs and also improves the ability of the design to approximate the binding specificities of TFs whose width is greater than the order of the de Bruijn sequence.

Indeed, our group has already implemented this design and applied it to determine the binding specificities of five TFs from different organisms and structural classes with an unprecedented level of resolution [1]. There, we demonstrated that this platform could be used to assay the binding preferences of individual binding site variants, allowing us to identify at least one case of positional interdependence in a binding site motif. Similarly, we were able to approximate a binding site motif that was 12 bases in length using a de Bruijn sequence of order 10, attesting to the advantages of a careful and thorough coverage of gapped *k*-mers (point *5* above). Our group is now using this technology to determine the binding specificities of many TFs from a range of organisms, providing a much needed data type for the biological community. Thus, this microarray design provides a powerful, general and robust platform, and its implementation provides an experimental tool that will allow effective determination of TF binding site specificities both now and in the future.

Acknowledgments. We thank Savina Jaeger for critical reading of the manuscript. This work was funded in part by grant R01 HG003985 from NIH/NHGRI to M.L.B. M.F.B. was supported in part by a National Science Foundation Graduate Research Fellowship. A.A.P. was supported in part by a National Defense Science and Engineering Graduate Fellowship, a National Science Foundation Graduate Research Fellowship, and an Athinoula Martinos Fellowship.

References

1. Berger M.F., Philippakis A.A., Qureshi A.M., He F.S., Estep P.W., 3rd, Bulyk M.L.: Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 24 (2006) 1429-1435
2. Bulyk M.L.: Computational prediction of transcription-factor binding site locations. Genome Biol 5 (2003) 201
3. Benos P.V., Lapedes A.S., Stormo G.D.: Is there a code for protein-DNA recognition? Probab(ilistical)ly. Bioessays 24 (2002) 466-475

4. Das P.M., Ramachandran K., vanWert J., Singal R.: Chromatin immunoprecipitation assay. Biotechniques 37 (2004) 961-969
5. Wyrick J.J., Young R.A.: Deciphering gene expression regulatory networks. Curr Opin Genet Dev 12 (2002) 130-136
6. Oliphant A.R., Brandl C.J., Struhl K.: Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol. Cell. Biol. 9 (1989) 2944-2949
7. Bulyk M.L., Huang X., Choo Y., Church G.M.: Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci U S A 98 (2001) 7158-7163
8. Mukherjee S., Berger M.F., Jona G. *et al.*: Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet 36 (2004) 1331-1339
9. Linnell J., Mott R., Field S., Kwiatkowski D.P., Ragoussis J., Udalova I.A.: Quantitative high-throughput analysis of transcription factor binding specificities. Nucleic Acids Res 32 (2004) e44
10. Warren C.L., Kratochvil N.C., Hauschild K.E. *et al.*: Defining the sequence-recognition profile of DNA-binding molecules. Proc Natl Acad Sci U S A 103 (2006) 867-872
11. De Bruijn N.G.: A Combinatorial Problem. Proc. Kon. Ned. Akad. v.Wetensch. 49 (1946) 758-764
12. Gross J.L., Yellen J.: Handbook of Graph Theory. CRC Press, New York (2004)
13. Joyner D., Kreminski R., Turisco J.: Applied Abstract Algebra. The Johns Hopkins University Press, Baltimore, MD (2004)
14. Golomb S.: Shift Register Sequences. Aegean Park Press, Laguna Hills, CA (1967)
15. Ben-Dor A., Karp R., Schwikowski B., Yakhini Z.: Universal DNA tag systems: a combinatorial design scheme. J Comput Biol 7 (2000) 503-519
16. Mintseris J., Eisen M.B.: Design of a combinatorial DNA microarray for protein-DNA interaction studies. BMC Bioinformatics 7 (2006) 429
17. Pevzner P.A., Tang H., Tesler G.: De novo repeat classification and fragment assembly. Genome Res 14 (2004) 1786-1796
18. Zhang Y., Waterman M.S.: An Eulerian path approach to local multiple alignment for DNA sequences. Proc Natl Acad Sci U S A 102 (2005) 1285-1290
19. Stewart I.: Galois Theory. Chapman & Hall, London, UK (1989)
20. Stormo G.D.: DNA binding sites: representation and discovery. Bioinformatics 16 (2000) 16-23
21. Terras A.: Fourier Analysis on Finite Groups and Applications. Cambridge University Press, Cambridge, UK (1999)
22. Singh-Gasson S., Green R.D., Yue Y. *et al.*: Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol 17 (1999) 974-978