

Universal DNA Tag Systems: A Combinatorial Design Scheme

Amir Ben-Dor* Richard Karp† Benno Schwikowski‡ Zohar Yakhini§

Abstract

Custom-designed DNA arrays offer the possibility of simultaneously monitoring thousands of hybridization reactions. These arrays show great potential for many medical and scientific applications such as polymorphism analysis and genotyping. Relatively high costs are associated with the need to specifically design and synthesize problem specific arrays. Recently, an alternative approach was suggested that utilizes fixed, universal arrays. This approach presents an interesting design problem—the arrays should contain as many probes as possible, while minimizing experimental errors caused by cross-hybridization. We use a simple thermodynamic model to cast this design problem in a formal mathematical framework. Employing new combinatorial ideas, we derive an efficient construction for the design problem, and prove that our construction is near-optimal.

1 Introduction

Oligonucleotides are short single-stranded pieces of DNA (typically 15-50 nucleotides) made by chemical synthesis. In solution, oligonucleotides tend to specifically *hybridize* with their Watson-Crick complements

([21]), and form a stable DNA duplex. This specificity is exploited in molecular hybridization assays, in which oligonucleotides are used as probes to identify any complementary (or near complementary) DNA from a complex mixture of target DNA.

Array-based hybridization assays, introduced in the late 1980s [6, 13, 15, 3, 5, 8], offer the possibility of simultaneously monitoring a multitude (currently up to tens of thousands) of hybridization reactions. In such an assay, a target-specific set of oligonucleotides is synthesized on a solid support surface (e.g., silicon or glass). A fluorescently labeled target sample mixture of DNA or RNA fragments is then brought in contact with the treated surface, and allowed to hybridize with the synthesized oligonucleotides. Scanning the fluorescent labels of the fragments attached to the array reveals information about the content of the sample mixture. Theoretically, the assay conditions are such that hybridization only occurs in sites on the surface that are Watson-Crick complements to some substring in the target. In practice, cross-hybridization is a main source of cross-signal contamination in any array-based hybridization assay.

Array-based hybridization assays show great potential for many different applications such as SNP genotyping [12], gene expression profiling [4], and resequencing DNA [14, 12]. Recently, S. Brenner and others [1, 2] suggested an alternative approach based on *universal arrays* containing oligonucleotides called *antitags*. The Watson-Crick complement of each antitag is called a *tag*. The tag-antitag pairs are designed so that each tag hybridizes strongly to its complementary antitag, but not to any other antitag. In this approach, the analysis of a DNA sample consists of two steps: solution-phase hybridization followed by solid-phase hybridization. In the first step, hybridization takes place between the target DNA in solution and a set of oligonucleotide precursors called reporter molecules. Each reporter molecule consists of a target-specific part ligated to a unique tag.

*Department of Computer Science & Engineering, University of Washington. Supported by the Program in Mathematics and Molecular Biology (amirbd@cs.washington.edu).

†International Computer Science Institute and Mathematical Sciences Research Institute, University of California at Berkeley. Supported by NSF grant at the University of Washington (karp@icsi.berkeley.edu).

‡Department of Computer Science & Engineering, University of Washington. Supported by the German Academic Exchange Service (DAAD) (benno@cs.washington.edu).

§Chemical and Biological Systems Department, Agilent Laboratories, a Hewlett-Packard Subsidiary (zohary@hp1.hp.com).

Reporter/target hybridization events are registered (e.g. by an enzymatic reaction). In the second step the modified precursors are introduced to the array. Tags form duplexes with the corresponding antitags. Thus, the reporter molecules are sorted into different locations on the array and hybridization events can be determined. This approach has several advantages:

- Complicated array manufacturing processes are required only for the fixed, universal component of the assay. These universal components can therefore be mass-produced, significantly reducing manufacturing costs.
- The assay components that need to be designed for a specific target are involved in solution phase processes. The relevant nucleic acid chemistry and thermodynamics is better understood than that of surface-based processes. Therefore a more efficient and effective design process is facilitated.

As an example, we describe a multiplexed SNP genotyping assay. SNPs (single nucleotide polymorphisms) are differences, across the population, in a single base, within an otherwise conserved genomic sequence [9]. Genotyping is a process that determines the variants present in a given sample, over a set of SNPs. This assay uses off-the-shelf universal components: a universal set of oligonucleotide tags and a universal array of antitags. The antitags, immobilized on the array, are Watson-Crick complements of the tags in the mixture. The whole system will be called a DNA Tag/AntiTag system and in short a DNA TAT system. Consider a set of SNPs to be genotyped. The assay is performed as follows (see Fig. 1):

1. A set of *reporter molecules* (one for each SNP) is synthesized in solution. Each reporter molecule consists of two parts that are ligated (in string language: concatenated) together. The first part is the Watson-Crick complement of the upstream sequence that immediately precedes the polymorphic site of the SNP. The second part of each reporter molecule is a unique tag from the universal set of tags.
2. When an individual is to be genotyped, a sample is prepared that contains the sequences flanking each of the SNP loci. The sample is mixed with the reporter molecules. Solution-phase hybridization then takes place. Assuming that specificity is perfect, this results in the flanking sequences of the SNPs paired only with the appropriate reporter molecule.
3. Single nucleotides, **A, C, T, G**, fluorescently labeled with four distinct colors, are added to the mixture.

These labeled nucleotides hybridize to the polymorphic site of each SNP and are ligated to the corresponding reporter molecule. That is, each reporter molecule is extended by exactly one labeled nucleotide.

4. The extended reporter molecules are separated from the sample fragments, and brought into contact with the universal array. Assuming that specificity is perfect, the tag part of each reporter molecule will only hybridize to its complementary antitag on the array. Thus the extended reporter molecules sort into the array sites where the corresponding antitag is present.
5. For each site of the array, the fluorescent colors present at that site are detected. The colors indicate which bases were used for the extension at the corresponding SNP site, and thus reveal the SNP variations present in the individual.

The design problem for a DNA TAT system presents a tradeoff. Clearly, it is desirable to have as many tags as possible, in order to maximize the number of SNPs that can be genotyped in parallel. On the other hand, if too many tags are used, similar tags will necessarily entail cross-hybridization events (where tags hybridize to foreign antitags), reducing the accuracy of the assay.

This design problem was identified in previous work and several formulations and solutions were proposed [10, 1, 2, 18, 11]. These papers differ both in the way hybridization is modeled, and in the algorithmic approach employed to find a good DNA TAT system. In [10] a TAT system is described as a part of a strategy for surface-based DNA computing. The authors take a coding theory approach and choose to model cross-hybridization constraints as general Hamming distance conditions. A set of 108 8-mers, with a 50% G/C content, which differ in at least 4 bases from each other, is constructed, and experimentally tested for cross-hybridization.

In [1] the method of using a DNA TAT system to sort target DNA is presented, together with several examples of applications. The model assumption is that two oligonucleotides of length n need to have perfectly complementary substrings of length more than λ in order to form a reasonably stable duplex. A set of n -mers is said to be a λ -free code if no two elements of the set have a common substring of length more than λ . Given n , the design problem implied in [1] is to construct the largest possible λ -free code.

A *De Bruijn sequence of order λ* is a cyclic sequence in which each possible λ -mer occurs exactly once [7]. In [2] the authors observe that by parsing a De Bruijn sequence of order λ , an optimal λ -free code, of size

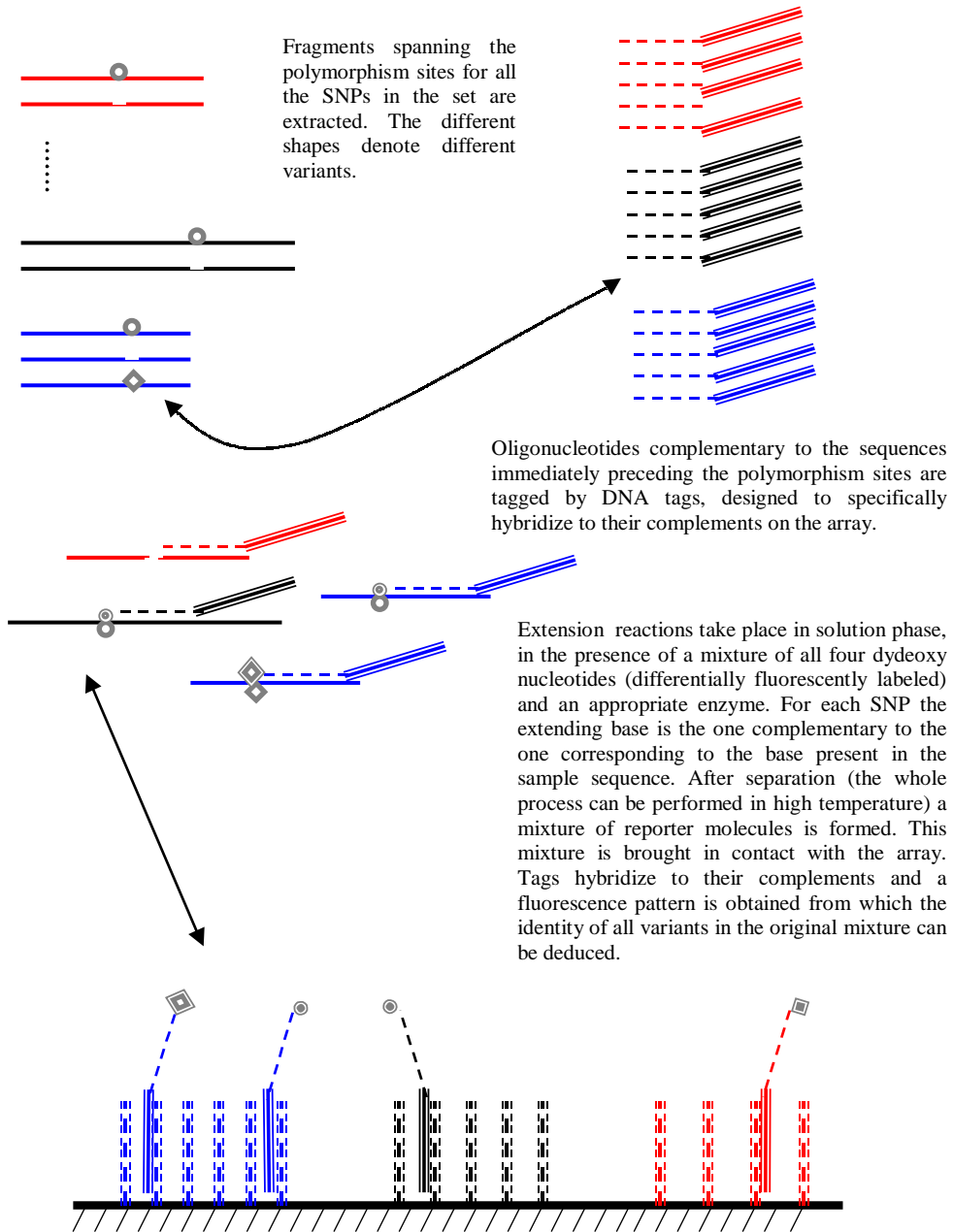


Figure 1: Schematic for SNP genotyping using a DNA TAT system

$4^\lambda/(n - \lambda + 1)$ can be obtained. However, the authors also recognize the shortcoming of their highly simplified mathematical model. To capture cross-hybridization, a model has to embody thermodynamic properties of DNA duplexes. The work we present here improves on this aspect by using De Bruijn sequences in a different way.

Additionally, the authors of [2] suggest a greedy approach to designing DNA TAT systems—starting with an empty set and iteratively adding a new tag to it, provided it does not hybridize with any of the complements of the tags already included. This approach clearly allows the use of arbitrarily complex thermodynamical models. However, there is no analysis for the performance of such greedy heuristics.

Recently, a small prototype universal array was used to detect K-ras mutations in tumor and cell line DNA [11]. The results reported suggest that universal arrays may be used to rapidly detect low-abundance mutations in any gene of interest.

In this paper we use a simple thermodynamic model of hybridization to give a precise formulation of the TAT system design problem. We employ new combinatorial ideas to provide an efficient construction and prove that our solution is near-optimal.

1.1 Thermodynamic model

In this section we describe a simple thermodynamic model of DNA duplex formation, and use this model to derive a formal TAT design problem. DNA duplexes are held together by weak hydrogen bonds formed between Watson-Crick complementary nucleotides. Denaturation (or melting) of a DNA duplex is generally achieved by heating the solution to a temperature which disrupts the hydrogen bonds. The energy required to separate complementary DNA strands is dependent on a number of factors, notably: strand length—longer duplexes contain a large number of hydrogen bonds and require more energy to separate them—and base composition, because C-G pairs have one more hydrogen bond than A-T pairs, strands with higher C-G content are more difficult to separate than those with low C-G content [20].

As the assay temperature increases, the probability that a duplex will be separated increases. A useful measure of the hybridization behavior of two oligonucleotides U and V is their *melting temperature* $t_M(U, V)$ (in degrees Celsius). This is the temperature at which, under stated experimental conditions, half of the U and V oligonucleotides will be in a single-stranded form, and half will occur in duplexes.

Let f denote the fraction of duplexes formed at a specific array site between an immobilized oligonucleotide and a fluorescently labeled tag. We assume that we are

given two design parameters $0 \leq \alpha \leq \beta \leq 1$ such that:

- If $f > \beta$, the resulting fluorescent level is interpreted as a positive result.
- If $f < \alpha$, the resulting fluorescent level is reported as a negative result.

Moreover, we assume that the experiment is performed at temperature t . The parameters α and β can be translated into two temperature parameters, C and H ($C < t < H$) with the following property: If a duplex has a melting temperature of at most C , in the experiment (done at temperature t), at most a fraction ϵ of the duplexes will form. Similarly, if a duplex has a melting temperature of at least H , then at temperature t , at least a fraction δ of the duplexes will form.

The design goal is to construct a large TAT system such that

1. for each tag U , $t_M(U, \bar{U}) \geq H$, and
2. for any two distinct tags U and V , $t_M(U, \bar{V}) < C$.

Such a system, if used with temperature t , would allow each tag-antitag hybridization to be detected, without any cross-hybridization errors. Note that, in a practical application, C can be lowered and H can be increased to provide a buffer against parameter inaccuracies and limitations in the melting temperature model used, imperfect temperature control, and further experimental biases.

For estimating the melting temperature of a duplex between a tag and its antitag we employ the *2-4 rule* that is commonly used for short oligonucleotides [20]: The melting temperature of a sequence and its complement is approximately twice the number of A-T base pairs plus four times the number of C-G base pairs. This model also enables us to state a condition that is sufficient to exclude cross-hybridization errors. The sufficient condition is derived from a characterization of the duplex formation process by Southern et al. [19]

“...the process begins by the formation of a transient nucleation complex from the interaction of very few base pairs. Duplex formation proceeds, one base pair at a time, through a zippering process. At any point the reaction may go in one of two directions, pairing or separation: if bases are complementary and freely available for pairing, duplex formation is more likely to proceed; if bases are non-complementary or a stable structure inhibits base pair formation, the block to the zippering process may drive the nucleation complex to fall apart. Duplex formation, and hence duplex yield, will be determined by the

stability of the nucleation complex and of intermediates up to the point in the zippering process where the likelihood of strand separation is negligible.”

Based on the above characterization and the 2–4 rule, we make the following assumptions.

1. Stable hybridization is always initiated by the formation of a nucleation complex.
2. A nucleation complex is a region of perfect base pairs between the tag and the foreign antitag.
3. The melting temperature of the nucleation complex can be computed according to the 2–4 rule.

Following these assumptions, we aim at avoiding the situation where a tag contains a substring x and a non-complementary antitag contains \bar{x} , where $t_M(x, \bar{x}) \geq C$. Since tags and antitags are complementary to each other, the above requirement can be stated in terms of solely the tag sequences.

2 TAT system design problem

We can now formally define the TAT system design problem. Our goal is to construct a TAT system with a maximum number of tag–antitag pairs such that the following properties are satisfied:

- (1) For each tag–antitag pair (U, \bar{U}) the melting temperature (using the 2–4 rule) satisfies $t_M(U, \bar{U}) \geq H$
- (2) For any two distinct tags U and V , and for each oligonucleotide x that occurs as a substring in both U and V , $t_M(x, \bar{x}) < C$.

We now present a conservative formalization of the above TAT design problem. Specifically, we not only forbid that a string x with $t_M(x, \bar{x}) \geq C$ occurs in any two distinct tags, but we also forbid that it occurs twice in the same tag. This mild restriction has also been imposed in previous work [2] and allows a rigorous analysis and a near-optimal constructive solution.

As usual, we model oligonucleotides as strings over the alphabet $\Sigma = \{\mathbf{A}, \mathbf{C}, \mathbf{T}, \mathbf{G}\}$. We assume that the parameters C and H are fixed. To keep our exposition simple, we assign to each string the number that corresponds to half of its melting temperature in degrees Celsius.

Definition 1. The weight $w(s)$ of a string $s = a_1 a_2 \cdots a_k$ is $\sum_{i=1}^k w(a_i)$, where $w(\mathbf{A}) = w(\mathbf{T}) = 1$ and $w(\mathbf{C}) = w(\mathbf{G}) = 2$. Given two parameters c and h , we call a set T of strings or “tags” a valid $c-h$ code if the following two conditions are satisfied:

Condition 1 Each tag has a weight of h or more.

Condition 2 Any substring of weight c or more occurs at most once.

Note that a valid $c-h$ code corresponds to a solution of the TAT design problem in which the lower melting temperature C is $2c$ and the upper melting temperature H is $2h$. We call the problem of finding a maximum valid $c-h$ code the **Combinatorial Tag Design Problem**. See Figure 2 for an example of a valid code.

GACCAAT	CAGCTAT	GTCGATA	CTGGTTA
CATTATCA	GAAATTCT	CTTAATGA	GTATTTGT
ATATAGTG	TAAAACCTC	AATAAGAG	TTTTACAC

Figure 2: A valid 4–10 code with 12 tags

In the next section we derive an upper bound, in particular, implies that the code in Fig. 2 is optimal, that is: there exists no valid 4–10 code with more than 12 tags.

3 Upper bound

In this section we derive a tight upper bound for the number of tags in a valid $c-h$ code. The idea is to associate a numerical resource with each tag. We show that any tag has to use a certain minimum amount of resource, and the global resource usage over all tags cannot exceed a certain maximum. The upper bound on the number of possible tags then follows from dividing the global upper bound by the minimum amount of resource used by each tag.

Roughly speaking, the limited resource we consider consists of those substrings in a tag with a weight of c or more that can occur only once in a valid $c-h$ code. The following definition captures the minimal *suffixes* that can occur only once in a valid $c-h$ code.

Definition 2. We call a string t a c -token if $w(t) \geq c$, but t does not properly contain a suffix of weight $\geq c$.

It is straightforward to see that a substring of weight $\geq c$ occurs twice if and only if some c -token occurs twice. We can therefore replace Condition 2 in our problem by the following equivalent condition.

Condition 2’ Any c -token occurs at most once.

Hereafter, we refer to a c -token simply as a *token*. With each token, we associate its *tail weight*, the weight of its terminal character. The tail weight of a tag T is the sum of the tail weights of all the tokens it contains as substrings. Figure 3 gives an example for $T = \text{GACCAAT}$ and $c = 4$.

Tag T :	GACCAAT	Token's tail weight
	GAC	2
	CC	2
Tokens:	CCA	1
	CAA	1
	CAAT	1
		Tail weight of T : 7

Figure 3: Tokens and tail weight

Notice that all characters of T except the first two terminate a token and thus contribute their weight to the tail weight of T . The first two characters do not terminate a token because they do not terminate a suffix of weight $\geq c$. In the general case of a tag T in a valid c - h code, the maximal prefix that does not contain a suffix of weight $\geq c$ has a total weight of at most $c - 1$. Since the weight of a tag in a valid code is at least h (Condition 1), we have the following lower bound.

Lemma 1. *Any tag in a valid c - h code has a tail weight of at least $h - c + 1$.*

Based on Conditions 1 and 2', we now derive the upper bound on the total tail weight of a valid c - h code. We use $\langle n \rangle$ to denote the set of strings with weight $n \in \mathcal{N}$, and G_n to denote the number of such strings. It is straightforward to derive the recurrence $G_1 = 2$, $G_2 = 6$, and $G_n = 2 \cdot G_{n-2} + 2 \cdot G_{n-1}$ for $n \geq 3$, and for the sake of simplicity we define $G_0 := 1$. Using standard techniques for solving recurrences it can be shown that, for all $n \in \mathcal{N}$, G_n is the nearest integer to

$$\left(\frac{3 + \sqrt{3}}{6} \right) \cdot (1 + \sqrt{3})^n.$$

To compute the maximal total tail weight of the tokens in a valid code, we partition tokens into four classes. In our symbolic representation of these classes, we denote any character with a weight of 1 (**A** or **T**) by **W** ("weak") and a character with a weight of 2 (**C** or **G**) by **S** ("strong"). To see how the set of tokens is partitioned, observe that any token is terminated by either a strong or a weak character, and it has a weight of either c or $c + 1$. Tokens of weight $c + 1$ begin with a strong character, since a string of weight $c + 1$ that begins with a weak character properly contains a suffix of weight c . Figure 1 lists the corresponding four classes of tokens, their maximal cardinalities, and the maximal total tail weight they can contribute in a valid code.

The total tail weight of each class is computed by multiplying its size by the weight of the terminal character of its members. Only the last row requires additional explanation: Observe that any token in the class $\mathbf{S}\langle c - 2 \rangle\mathbf{W}$ contains a token of the form $\mathbf{S}\langle c - 2 \rangle$ as a substring, and thus only $2 \cdot G_{c-2}$ tokens of this form can

Token class	Max. occurrences in valid code	Max. tail weight
$\langle c - 2 \rangle\mathbf{S}$	$2 \cdot G_{c-2}$	$4 \cdot G_{c-2}$
$\mathbf{S}\langle c - 3 \rangle\mathbf{S}$	$4 \cdot G_{c-3}$	$8 \cdot G_{c-3}$
$\langle c - 1 \rangle\mathbf{W}$	$2 \cdot G_{c-1}$	$2 \cdot G_{c-1}$
$\mathbf{S}\langle c - 2 \rangle\mathbf{W}$	$2 \cdot G_{c-2}$	$2 \cdot G_{c-2}$

Table 1: Bounds on the number of tokens and their tail weight in a valid c - h code

exist in a valid code. Therefore the number of tokens of the form $\mathbf{S}\langle c - 2 \rangle\mathbf{W}$ cannot exceed $2 \cdot G_{c-2}$. Summing up the rightmost column proves the following lemma.

Lemma 2. *The total tail weight of all tags contained in a valid c - h code is at most*

$$2 \cdot G_{c-1} + 6 \cdot G_{c-2} + 8 \cdot G_{c-3}.$$

Combining this with Lemma 1 yields the following upper bound.

Theorem 1. *Any valid c - h code contains at most*

$$\frac{2 \cdot G_{c-1} + 6 \cdot G_{c-2} + 8 \cdot G_{c-3}}{h - c + 1} \text{ tags.}$$

For $h = 10$ and $c = 4$, the upper bound is $\frac{2 \cdot 16 + 6 \cdot 6 + 8 \cdot 2}{10 - 4 + 1} = 12$, which proves:

Corollary 1. *The 4-10 code in Fig. 2 is optimal.*

4 Our construction using circular strings

In this section we describe a method of constructing a nearly optimal c - h code for arbitrary values of c and h . Specifically, our code comprises at least

$$\frac{2 \cdot G_{c-1} + 6 \cdot G_{c-2} + 4 \cdot G_{c-3}}{h - c + 3} - 1 \text{ tags.}$$

Comparing our code with the upper bound, and using the recurrence for G_n , one finds that our method achieves at least $0.89 \cdot (h - c + 1) / (h - c + 3)$ of the upper bound. For the values $c = 12$ and $h = 30$ that can be seen as relevant in practice, our construction yields 12119 tags, which corresponds to 87.6% of the upper bound of 13840 one gets from Theorem 1.

Throughout our exposition we will assume that the parameters c and $h \geq c$ are fixed. In addition we will assume that c is even. The case of an odd value of c requires only small modifications.

4.1 Construction overview

Our construction proceeds in two stages. In the first stage we construct a set of *circular strings* in which each token occurs at most once. The characters of a circular string are arranged in a cyclic order, and when convenient, we will assume that a specific character is designated as its origin.

In the second stage of our construction, the tags of our design are extracted as substrings from the circular strings, as illustrated in Fig. 4. To satisfy Condition 1, each of the extracted substrings has a weight of h or more. To satisfy Condition 2', the overlap between two tags has a weight of at most $c - 1$.

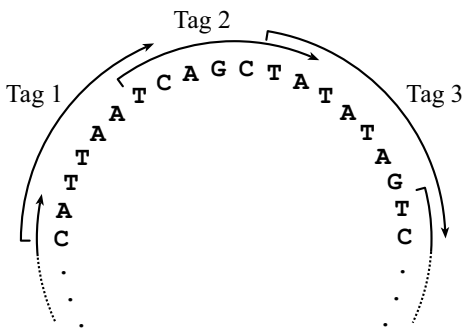


Figure 4: Second stage—Extracting tags from circular strings

Tags can be extracted from a circular string by a straightforward greedy algorithm that iterates the following operation. Starting at some position, the algorithm collects characters until their cumulative weight reaches or exceeds h , forming one tag, and then tracks back over as many characters as possible without collecting a weight of c or more. This operation is repeated until some overlap of weight $\geq c$ with the first extracted tag occurs, and the last retrieved tag is discarded. Given the best start position, this algorithm produces the largest number of tags that are substrings of a given circular string and can be included in a valid code. Observe that, since each character has a weight of 1 or 2, each tag extracted in this manner has a weight of at most $h + 1$, and the overlap between two tags is at least $c - 2$. Therefore, each circular string C leads to at least $\frac{w(C)}{h-c+3} - 1$ tags. This lower bound for the number of tags extracted from each cycle motivates us to consider the following formal problem.

Definition 3 (Circular String Problem). *Given the parameters $c > 0$ and $h > c$, construct a set \mathcal{C} of circular strings that contain any substring of weight*

$$\begin{aligned} \mathbf{A} &= (\mathbf{W}, \mathbf{0}) \\ \mathbf{T} &= (\mathbf{W}, \mathbf{1}) \\ \mathbf{C} &= (\mathbf{S}, \mathbf{0}) \\ \mathbf{G} &= (\mathbf{S}, \mathbf{1}) \end{aligned}$$

Figure 5: Encoding of each character into one meta-character and one bit

$\geq c$ at most once, and maximize

$$\sum_{C \in \mathcal{C}} \left(\frac{w(C)}{h-c+3} - 1 \right) \quad (1)$$

The construction we will describe optimally solves this problem. Specifically, our construction will yield a single cycle with the maximal possible weight among all set of circular strings that contain each substring of weight $\geq c$ at most once.

4.2 Meta-Strings and De Bruijn sequences

Our construction is based on the encoding of the nucleotides as given in Fig. 5. Each character $a \in \Sigma$ is identified with a pair (μ, β) , where $\mu \in \{\mathbf{W}, \mathbf{S}\}$ and $\beta \in \{\mathbf{0}, \mathbf{1}\}$. Extending this to strings over $\{\mathbf{A}, \mathbf{C}, \mathbf{T}, \mathbf{G}\}$, we identify each string s with its pair of *meta-string* $\mu(s)$ and *bit string* $\beta(s)$. In this context, we will also call the string s an *instance* of the meta-string μ . Figure 6 gives an example.

CCTGCAGGACGT	<i>String $s=(\mu(s), \beta(s))$</i>
SSWSSWSSWSSW	<i>Meta-string $\mu(s)$</i>
001100110011	<i>Bit string $\beta(s)$</i>

Figure 6: A string corresponds to its meta-string, bit string pair

Each circular string (or “cycle”) in our construction will be an instance of a long circular meta-string that arises from repeating a shorter meta-string. If s is a string, we will denote k repetitions of s by s^k .

De Bruijn sequences To avoid generating several identical tokens from repetitions of a meta-string μ , our construction will ensure that each instance of the μ is paired with a different pattern in the bit-string.

For $k \in \mathcal{N}$, a binary De Bruijn sequence of order k is a cyclic binary sequence of length 2^k in which each possible substring of length k occurs exactly once [7]. Such sequences exist for all $k \in \mathcal{N}$ and can be constructed in linear time. We assume that a fixed De Bruijn sequence of order k is given for each $k \in \mathcal{N}$. Reading this

sequence once, starting from a specific offset i relative to a fixed origin position, we obtain a linear string, a *linearization* that we denote by \mathcal{D}_k^i .

4.3 Cycle construction

Each cycle in our construction is based on a meta-string μ of weight c . Before describing the case of general meta-strings μ , we illustrate the construction principle for a special case.

Simple case We consider meta-strings μ with the following two properties.

- (P1) $\gcd(|\mu| + 1, 2^{|\mu|}) = 1$, i. e., the greatest common divisor of $|\mu| + 1$ and $2^{|\mu|}$ is 1.
- (P2) $\mu\mathbb{W}$ cannot be represented as a concatenation of two or more identical substrings.

For meta-strings μ that satisfy the two conditions above, the cycle is

$$C_0(\mu) = \left((\mu\mathbb{W})^{2^{|\mu|}}, (\mathcal{D}_{|\mu|}^0)^{|\mu|+1} \right).$$

Figure 7 shows $C_0(\mu)$ for $c = 4$ and $\mu = \text{SS}$. Notice

$$\begin{array}{l} \mu = \text{SS} \quad |\mu| = 2 \quad \mathcal{D}_{|\mu|}^0 = \text{0011} \\ \hline \text{SSWSWSWSWS} \quad (\mu\mathbb{W})^{2^{|\mu|}} \\ \text{001100110011} \quad (\mathcal{D}_{|\mu|}^0)^{|\mu|+1} \\ \hline \text{CCTGCAGGACGT} \quad C_0(\mu) \end{array}$$

Figure 7: Construction of the cycle $C_0(\text{SS})$

that the meta-string $(\mu\mathbb{W})^{2^{|\mu|}}$ not only contains the meta-string $\mu = \text{SS}$ four times as a substring, but also contains the meta-strings SSW and SWS four times.

General case The construction for general meta-strings μ is a generalization of the above construction. Let α be the shortest *period* of $\mu\mathbb{W}$, i. e., the shortest substring such that $\mu\mathbb{W}$ can be written as $\mu\mathbb{W} = (\alpha)^p$, and set $k = k(\mu) = \gcd(|\alpha|, 2^{|\mu|})$. For any meta-strings μ with $w(\mu) = c$, our code contains the k cycles

$$C_i(\mu) = \left((\alpha)^{2^{|\mu|}/k}, (\mathcal{D}_{|\mu|}^i)^{|\alpha|/k} \right), \quad i = 0, \dots, k-1.$$

Using the above notation, the set of cycles we construct is

$$\mathcal{C} := \bigcup_{w(\mu)=c} \bigcup_{i=0}^{k(\mu)-1} C_i(\mu).$$

4.4 Validity of our construction

In this section we outline how one can prove the validity of our construction, and in particular, that any token occurs at most once. Consider a token U of weight c , and let (μ, β) denote the corresponding pair of meta-string and bit string. Clearly, U occurs if and only if both μ and β occur together. The idea behind the proof is to first show that in our construction all of the instances of μ occur in the same cycle. We then show that, in this cycle, μ and β occur together at most once.

The formal proof for the simple case, i. e., meta-strings μ with the above properties (P1) and (P2), goes along the following lines.

Lemma 3. *The meta-string μ occurs only in the cycle $C_0(\mu) = \left((\mu\mathbb{W})^{2^{|\mu|}}, (\mathcal{D}_{|\mu|}^0)^{|\mu|+1} \right)$.*

Lemma 4. *The meta-string μ occurs exactly $2^{|\mu|}$ times in the meta-cycle of $C_0(\mu)$. Moreover, the distance between any two occurrences of μ is a multiple of $|\mu| + 1$.*

Lemma 5. *The bit string β occurs exactly $|\mu| + 1$ times in the bit cycle of $C_0(\mu)$ (once in each De Bruijn sequence $\mathcal{D}_{|\mu|}^0$). Moreover, the distance between any two occurrences of β is a multiple of $2^{|\mu|}$.*

It follows directly from the above three lemmas that U occurs at most once in our construction. The above proof can easily be extended to tokens U with a weight of $c + 1$, and general meta-tokens μ .

4.5 Token content of our cycles

We now examine how many tokens of weight c and $c + 1$ the set of cycles \mathcal{C} contains.

Lemma 6. *\mathcal{C} contains each token of weight c exactly once.*

Proof. Let μ be an arbitrary meta-string of weight c . Observe that, in each cycle $C_i(\mu)$, $i = 0, \dots, k-1$, μ occurs $2^{|\mu|}/k$ times, each time paired with a different μ -bit substring of $\mathcal{D}_{|\mu|}^i$. Therefore \mathcal{C} contains all $k \cdot 2^{|\mu|}/k = 2^{|\mu|}$ distinct instances of each meta-string of weight c , which means that each token of weight c occurs exactly once. \square

Lemma 7. *\mathcal{C} contains exactly half of the tokens of weight $c + 1$, and each of these occurs exactly once.*

Proof. Let μ be an arbitrary meta-string of weight $c + 1$. Since we have assumed that c is even, μ contains at least one weak character. Thus, μ can be represented as a rotation of some meta-string $\mu'\mathbb{W}$, where μ' is a meta-string of weight c . Therefore, all instances of μ occur in the cycles of the type $C_i(\mu')$, alternating with

instances of μ' . Due to the alternation, the numbers of instances of μ' and μ contained in any $C_i(\mu')$ are identical. Instances of μ in these cycles are also distinct, because each time μ occurs, it occurs with a distinct bit string. Since we have seen in Lemma 6 that all instances of μ' occur exactly once, and the maximally possible number of instances for μ is twice as high, we can conclude that exactly half of the instances of μ occur in \mathcal{C} . \square

Together with Table 1, Lemmas 6 and 7 yield:

Corollary 2. *The total weight of \mathcal{C} is $2 \cdot G_{c-1} + 6 \cdot G_{c-2} + 4 \cdot G_{c-3}$.*

4.6 Pasting the cycles together

Before extracting tags from the cycles in \mathcal{C} , we combine all cycles into a single cycle, without modifying the set of tokens that occur. This avoids the “end effect” that occurs when we extract tags from the cycles: Recall that every cycle can possibly leave all characters of a nearly complete tag unused. By pasting all cycles into a single cycle before the extraction, this situation occurs only once.

With the basic operation we present here, one can paste any two cycles A and B together if they share a common substring s with a weight of $c - 1$. If this is the case, A can be written as A_0s , where A_0 is some string, and, analogously, B can be written as B_0s . Then $\text{paste}(A, B) := A_0sB_0s$ defines a new cyclic string. The following lemma guarantees that the tail weight and the validity of the set of tokens contained in the circular code is preserved.

Lemma 8. *For any two cycles A and B that share a common substring of weight $c - 1$, $\text{paste}(A, B)$ contains exactly the union of the tokens contained in A and the tokens contained in B .*

Proof idea. Observe that any token contained in $\text{paste}(A, B) = A_0sB_0s$ is a substring of $A = A_0s$ or $B = B_0s$. On the other hand, each of these tokens occurs exactly once in $\text{paste}(A, B)$.

Lemma 9. *There exists a sequence of paste operations that merges all cycles of \mathcal{C} into a single cycle.*

Proof. The central observation is that each cycle $C_i(\mu)$, where μ is a meta-string containing $k \geq 1$ strong characters, can be pasted with a cycle of the form $C_j(\nu)$, where ν contains only $k - 1$ strong characters. To see how this works, observe that the circular meta-string of $C_i(\mu)$ can be expressed as a repetition of $\mu'S$ for some meta-string μ' of weight $c - 1$. Consider an instance s of μ' in $C_i(\mu)$. The string sT is a token of weight c and,

according to Lemma 6, does occur in some cycle $C_j(\nu)$ with $\nu = \mu'W$. Observe that the meta-string ν contains only $k - 1$ strong characters. Since both cycles $C_i(\mu)$ and $C_j(\nu)$ contain s as a substring, they can be pasted together.

Iterating the above paste operation leads from any given cycle $C_i(\mu)$ to the one cycle $C_i(W^c)$ that contains no strong characters. Since substrings of weight $c - 1$ are preserved by the paste operation, all cycles of \mathcal{C} can indeed be pasted into a single cycle. \square

Together with Corollary 2, Lemmas 8 and 9 prove our earlier claim.

Theorem 2. *The above construction yields at least*

$$\frac{2 \cdot G_{c-1} + 6 \cdot G_{c-2} + 4 \cdot G_{c-3}}{h - c + 3} - 1 \quad \text{tags.}$$

As mentioned before, this means that, asymptotically, our code achieves 89.9% of the upper bound from Theorem 1. For the values of $c = 12$ and $h = 30$, our code achieves 87.6% of the upper bound.

5 Optimality of our construction

In this section we show that our construction solves the Circular String Problem optimally. To this end, we prove that any set of cycles that is valid (i. e., no token is repeated) has a weight of at most $2 \cdot G_{c-1} + 6 \cdot G_{c-2} + 4 \cdot G_{c-3}$. Note that, using the recursion for G_n , this can be written as $G_c + 2 \cdot G_{c-1}$.

Let \mathcal{C}' denote a set of cycles. We continue to discuss even values of c , and use r to denote $c/2$. To bound the total weight of \mathcal{C}' , we will bound the number of weak characters and the number of strong characters in \mathcal{C}' . We need the following technical lemma.

Lemma 10. *For $k \in \{0, \dots, r - 1\}$, the number of instances of the meta-string S^kW in \mathcal{C}' is at most $2^k G_{2(r-k)}$.*

Proof. To bound the number of instances of S^kW , we bound the number of tokens that have an instance of S^kW as a suffix. There are two cases to consider:

- Tokens of weight $2r$, which have the form $\langle 2r - 2k - 1 \rangle S^kW$. There are $2^{k+1} G_{2(r-k)-1}$ tokens of this type.
- Tokens of weight $2r + 1$, which have the form $S \langle 2r - 2k - 2 \rangle S^kW$. Since the prefix $S \langle 2r - 2k - 2 \rangle S^k$ has a weight of $2r$, these tokens cannot occur more than $2^{k+1} G_{2(r-k)-2}$ times in \mathcal{C}' .

Therefore, the total number of tokens that have an instance of $\mathbf{S}^k\mathbf{W}$ as a suffix is

$$2^{k+1}G_{2(r-k)-1} + 2^{k+1}G_{2(r-k)-2} = 2^k G_{2(r-k)}.$$

As each instance of $\mathbf{S}^k\mathbf{W}$ in \mathcal{C}' is the suffix of exactly one token, our claim follows. \square

The following lemma is straightforward to prove.

Lemma 11. *\mathcal{C}' contains at most 2^r instances of the meta-string \mathbf{S}^r .*

The case of $k = 0$ in Lemma 10 implies that \mathcal{C}' contains at most G_{2r} weak characters. To bound the number of strong characters in \mathcal{C}' , we need the following lemma.

Lemma 12. *The number of strong characters in \mathcal{C}' equals the number of instances of the meta-string $\mathbf{S}^k\mathbf{W}$ (over $k = 1, \dots, r-1$), plus the number of instances of \mathbf{S}^r .*

Proof. We prove this lemma by mapping each character s in \mathcal{C}' to instances of $\mathbf{S}^k\mathbf{W}$ or \mathbf{S}^r in \mathcal{C}' . If the $r-1$ characters $\sigma_1, \dots, \sigma_{r-1}$ that follow s in \mathcal{C}' are all strong, we map s to the string $s, \sigma_1, \dots, \sigma_{r-1}$, an instance of \mathbf{S}^r in \mathcal{C}' . Otherwise, let i be the minimal index such that σ_i is a weak character. In this case, s is mapped to $s, \sigma_1, \dots, \sigma_i$, an instance of $\mathbf{S}^k\mathbf{W}$ in \mathcal{C} . It is easy to verify that this mapping is one-to-one and onto. \square

To establish a bound on the number of strong characters in \mathcal{C}' , we can now sum up the above bounds on the total number of instances of $\mathbf{S}^k\mathbf{W}$ ($k = 1, \dots, r-1$) and the number of instances of \mathbf{S}^r in \mathcal{C} . In the following lemma we show that this sum (and thus the upper bound on the number of strong characters in \mathcal{C}') equals G_{2r-1} .

Lemma 13.

$$\sum_{k=1}^r 2^k G_{2(r-k)} = G_{2r-1}$$

Proof. Using $i = r - k$, the above equality is equivalent to the following

$$\sum_{i=0}^{r-1} 2^{r-i} \cdot G_{2i} = G_{2r-1}$$

which can be proven by induction on r , using the above recursion for G_n . \square

Lemma 13 yields an upper bound of G_{c-1} for the number of strong characters in \mathcal{C}' . Together with the upper bound of G_c on the number of weak characters, we get the following.

Theorem 3. *The total weight of any valid set of cycles is at most $G_c + 2 \cdot G_{c-1}$.*

This proves that our construction is indeed optimal.

6 Discussion

This paper formulates the design of a universal set of tags as a combinatorial problem and achieves a provably near-optimal solution. Our formulation rests on two assumptions:

1. If a sequence has weight greater than or equal to h (corresponding to melting temperature greater than or equal to $2h$ in the 2-4 model) then the sequence will hybridize to its complement.
2. Sequence x will fail to hybridize to sequence y provided that there is no string z of weight greater than or equal to c such that z is a substring of x and z^C is a substring of y .

In practice the choice of the parameters h and c will depend on the concentrations of the reagents involved in the hybridization process and on other hybridization conditions.

Since our model of hybridization is only an approximate rule of thumb, it is inevitable that our design will include tag-antitag pairs that violate the two assumptions. Secondary structure in a tag may cause it to fail to hybridize to its antitag, even though its weight is greater than or equal to h . A tag and a foreign antitag may hybridize together because of long substrings that are nearly complementary but not exactly complementary.

Such violations depend on very specific properties of sequences, such as unusual dinucleotide composition, high-weight near-perfect matches or specific structural motifs. For example, a principal type of secondary structure within a DNA sequence is a hairpin, which can occur when the sequence contains two high-weight complementary substrings separated by at least three nucleotides. Such features occur infrequently in random sequences, and our method of tag construction does not appear to be strongly biased toward their occurrence. The 4-10 code in Figure 2 contains no complementary substrings of weight 4 in a single tag. The 12-30 code our method yields contains no tag containing complementary substring of weight higher than 12. Only approximately $\sim 0.1\%$ (14 tags out of a total of 12119) contain a pair of complementary substrings of weight 12.

Moreover, we can guard against possible bias by randomizing the choice of De Bruijn sequences in the cycle formation stage of our construction. Thus, we believe that violations will not occur frequently in our set of tags and antitags if the parameters h and c are chosen conservatively.

In order to make our design useful it will be necessary to verify our belief that violations are infrequent, and then get rid of the violations that do exist by deleting

some tags. We can perform these tasks by a combination of computational and experimental approaches.

The computational approach depends on the availability of refined models of DNA secondary structure and of duplex formation between (not necessarily complementary) DNA sequences. For special types of duplexes, such refined models are already available [17, 16]. Given such models, we can computationally screen our tags for secondary structure interfering with hybridization, and screen the tag-antitag pairs for undesired hybridization. Because of the huge number of tag-antitag pairs, an exhaustive approach to the latter task may be infeasible. Instead, it may be possible to use mathematical properties of our design and of the model of duplex formation to limit the search. For example, it may be possible to show that a tag x and a foreign antitag y are likely to form a duplex only if they have been constructed from highly similar meta-tokens.

Another problem lies in the experimental validation of tag-antitag systems. Once a set of tags has been screened computationally, one can perform further screening by building universal arrays and exposing them to sets of antitags. Choosing these sets of antitags for this screening procedure poses another new design problem.

Acknowledgements

We thank Deborah Nickerson for introducing us to the problem of designing universal arrays, and we gratefully acknowledge the helpful remarks of the anonymous referees.

References

- [1] S. Brenner, *Methods for sorting polynucleotides using oligonucleotide tags*, US Patent 5,604,097, 1997.
- [2] M. S. Morris, D. D. Shoemaker, R. W. Davis and M. P. Mittmann, *Methods and compositions for selecting tag nucleic acids and probe arrays*, European Patent Application 97302313, 1997.
- [3] D. Solas A. C. Pease, E. J. Sullivan, M. T. Cronin, C. P. Holmes, and S. P. A. Fodor. Oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. of Sci. USA*, (91):5022–5026, 1994.
- [4] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues by oligonucleotide arrays. *PNAS*, 96:6745–6750, June 1999.
- [5] A. P. Blanchard and L. Hood. Sequence to array: probing the genome's secrets. *Nature Biotechnology*, 14:1649, 1996.
- [6] R. Drmanac G. Lennon S. Drmanac I. Labat R. Crkvenjakov and H. Lehrach. Partial sequencing by oligohybridization: Concept and applications in genome analysis. In *Proceedings of the first international conference on electrophoresis supercomputing and the human genome* Edited by C. Cantor and H. Lim, pages 60–75, Singapore, 1991. World Scientific.
- [7] N.G. de Bruijn. A combinatorial problem. *Proc. Kon. Ned. Akad. v. Wetensch.*, 49:758–764, 1946.
- [8] J. DeRisi, V. Iyer, and P. Brown. Exploring the metabolic genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.
- [9] D. G. Wang et al. Large-scale identification, mapping and genotyping of single nucleotide polymorphisms in the human genome. *Science*, (280):1077–1082, 1998.
- [10] A. G. Frutos, Q. Liu, A. J. Thiel, A. M. W. Sanner, A. E. Condon, L. M. Smith, and R. M. Corn. Demonstration of a word design strategy for DNA computing on surfaces. *Nucleic Acids Research*, 25(23):4748 – 4757, 1997.
- [11] N. Garry, N. Wotiwski, J. Day, R. Hammer, G. Barany, and F. Barany. Universal dna microarray method for multiplex detection of low abundance point mutations. *J. Mol. Bio.*, (292):251–262, 1999.
- [12] Joseph G Hacia. Resequencing and mutational analysis using oligonucleotide micro arrays. *Nature Genetics*, 21(1):42–47, January 1999.
- [13] K. R. Khrapko, A. Khorlin, I. B. Ivanov, B. K. Chernov, Y. P. Lysov, S. Vasilenko, V. Floreny'ev, and Mirzabekov. Hybridization of DNA with oligonucleotides immobilized in gel: A convenient method for detecting single base substitutions. *Molecular Biology*, 25(3):581–591, 1991.
- [14] M. Kozal, N. Shah, N. Shen, R. Fucini, R. Yang, T. Merigan, D. D. Richman, M. S. Morris, E. Hubbell, M. Chee, and T. R. Gingeras. Extensive polymorphisms observed in hiv-1 clade b protease gene using high density oligonucleotide arrays: implications for therapy. *Nature Medicine*, (7):753–759, 1996.
- [15] C. Y. Lin, K. H. Hahnenberger, M. T. Cronin, D. Lee, N. M. Sampas, and R. Kanemoto. A method for genotyping cyp2d6 and cyp2c19 using genechip probe array hybridization. In *ISSX Meeting*, 1996.
- [16] N. Peyret, P.A. Seneviratne, H.T. Allawi, and J. Santalucia Jr. Nearest-neighbor thermodynamics and nmr of dna sequences with internal a.a, c.c, g.g, and t.t mismatches. *Biochemistry*, 38(12):3468–3477, March 1999.
- [17] John SantaLucia. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci.*, 95(4):1460–1465, February 1998.
- [18] D.D. Shoemaker, D.A. Lashkari, D. Morris, M. Mittmann, and R.W Davis. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genetics*, 4(14):450–456, 1996.
- [19] Edwin Southern, Kalim Mir, and Mikhail Shchepinov. Molecular interactions on microarrays. *Nature Genetics*, 21(1):5–9, January 1999.
- [20] Tom Strachen and Andrew P. Read. *Human Molecular Genetics*. Bios scientific publishers, 1996.
- [21] J. D. Watson, M. Gilman, J. Witkowski, and M. Zoller. *Recombinant DNA*. Scientific American Books, New York, 1996.