Design of Geometric Molecular Bonds

David Doty
Department of Computer Science
University of California, Davis
Davis, CA 95616
doty@ucdavis.edu

Andrew Winslow
Départment d'Informatique
Université Libre de Bruxelles
1050 Brussels, Belgium
awinslow@ulb.ac.be

***Abstract*—An example of a *nonspecific* molecular bond is the affinity of any positive charge for any negative charge (like-unlike), or of nonpolar material for itself when in aqueous solution (like-like). This contrasts *specific* bonds such as the affinity of the DNA base A for T, but not for C, G, or another A. Recent experimental breakthroughs in DNA nanotechnology [4], [11] demonstrate that a particular nonspecific like-like bond ("blunt-end DNA stacking" that occurs between the ends of any pair of DNA double-helices) can be used to create specific "macrobonds" by careful geometric arrangement of many nonspecific blunt ends, motivating the need for sets of macrobonds that are *orthogonal*: two macrobonds not intended to bind should have relatively low binding strength, even when misaligned.**

To address this need, we introduce *geometric orthogonal codes* that abstractly model the engineered DNA macrobonds as two-dimensional binary codewords. While motivated by completely different applications, geometric orthogonal codes share similar features to the *optical orthogonal codes* studied by Chung, Salehi, and Wei [3]. The main technical difference is the importance of 2D geometry in defining codeword orthogonality.

I. Introduction

A. Experimental DNA nanotechnology

DNA nanotechnology began in the 1980s when Seeman [9] showed that artificially synthesized DNA strands could be designed to automatically self-assemble nanoscale structures, rationally designed through the choice of DNA sequences. In the past 20 years, the field has witnessed a dramatic surge in the development of basic science, *in vitro* applications such as chemical oscillators and molecular walkers, and *in vivo* applications such as drug delivery, cellular RNA sensing, and genetically encoded structures [2].

A technological pillar of the field is *DNA origami*, developed by Rothemund [7], a simple, fast, inexpensive, and reliable method for creating artificial 2D and 3D DNA structures, with a control resolution of a few nanometers. DNA origami requires a single long *scaffold* strand of DNA; the most commonly used is the 7249-nucleotide single-stranded genome of the bacteriophage virus M13mp18, widely and cheaply available from many biotech companies. The scaffold is mixed with a few hundred shorter ($\approx$ 32nt) synthesized strands called *staples*, each of which is designed to bind (through Watson-Crick complementarity) to 2-3 regions of the scaffold. Via thermal annealing, the staples fold the scaffold strand into a shape dictated by the choice of staple sequences, hence the term *origami*. The process is illustrated in Figure 1(a), with the results shown in Figure 1(b).

Although the Watson-Crick pairing of bases between two single strands of DNA is very specific, DNA is known to undergo other, less specific interactions. One well-studied interaction is called a *stacking bond*, formed when *any* pair of terminated double helices — known as *blunt ends* — face each other, as shown in Figure 1(c). Since two edges of a standard DNA origami rectangle consist entirely of blunt ends, DNA origami rectangles are known to bind along their edges to form long polymers of many origamis, despite the fact that no hybridization between single strands occurs between them. One way to avoid stacking between origamis is to leave out staple strands along the edge, so that rather than blunt ends, there are single-stranded loops of the scaffold strand [7].

Woo and Rothemund [11] turned the lemon of unintended origami stacking into lemonade with the following idea: leave out *some* of the staples along the edge, but keep others; see Figure 1(d). Although each individual blunt end binds nonspecifically to any other, the only way for *all* blunt ends along an edge to bind is for the other origami to have them in the same relative positions. Thus, geometric placement of blunt ends makes the entire side of an origami into a specific "macrobond". Figure 1(d) shows how this enforces that the origami's left side binds only to a right side.

The idea extends from 2D origami rectangles with 1D edges, to 3D origami boxes with 2D rectangular faces as demonstrated by Gerling, Wagenbauer, Neuner, and Dietz [4].[1] They rationally design polymers of many origamis with prescribed sizes and shapes such as the 4-mer ABCD shown in Figure 1(e).

The preceding is an idealized description: nonidealities cause unintended pairs of macrobonds to bind spuriously. Figure 1(f) shows two macrobonds aligning enough of their blunt ends to attach stably, through two mechanisms that can be seen in the image: *flexibility* of DNA helices and *misalignment* of macrobonds.

This paper is an attempt to attack the latter problem with coding theory.

[1]The motif is slightly different. Rather than helices orthogonal to the origami face, they are parallel. Also, two whole helices protrude (see Figure 1(e), and the complementary face should have a two-helix "gap" where the other helices fit, forming four total stacking bonds.)

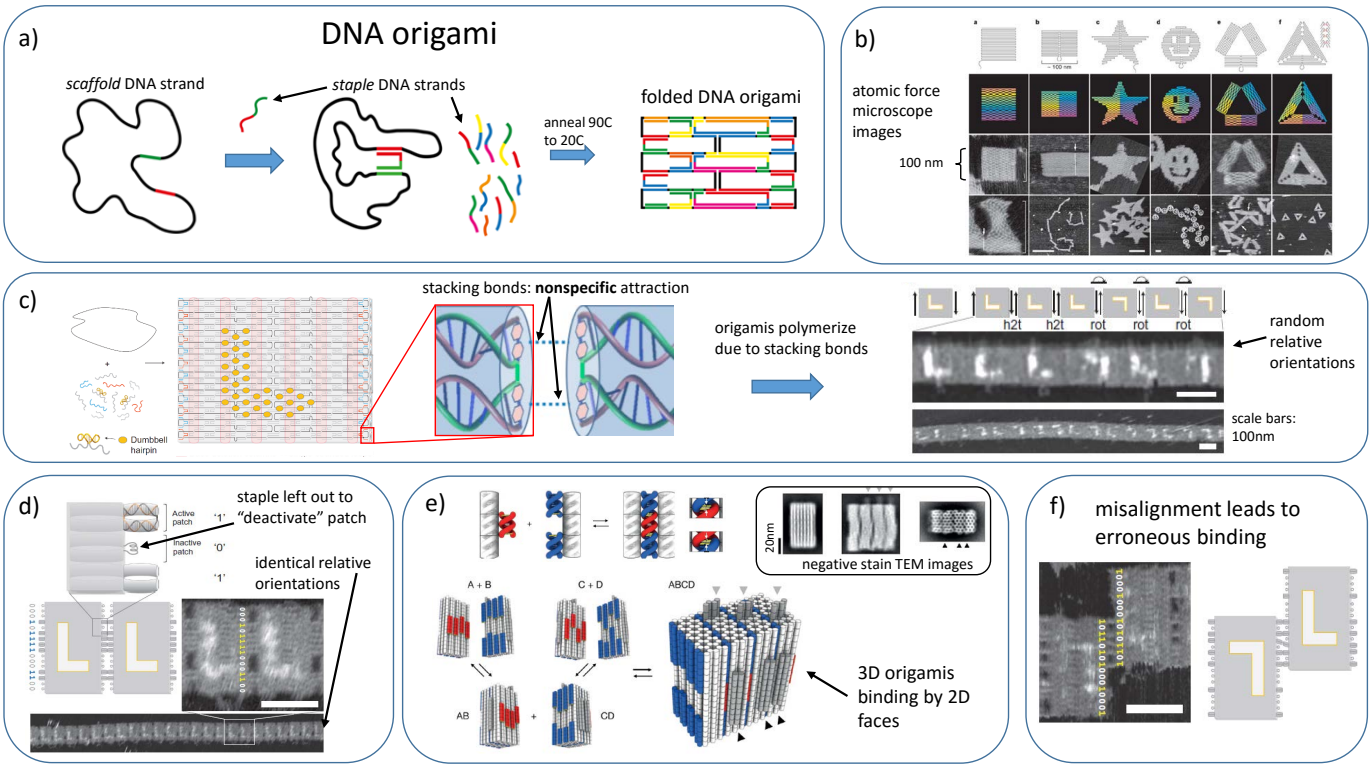

Fig. 1. Illustration of DNA origami and geometrically programmable stacking bonds. a) DNA origami illustration. (source: `http://openwetware.org/wiki/Biomod/2014/Design`) b) Atomic force microscope images of nanoscale shapes assembled by DNA origami technique. (source: [7]) c) Stacking bonds are nonspecific attraction occurring between the *ends* of two DNA helices, such as those that appear on the edges of a DNA origami. They cause origamis to polymerize (form long chains) in solution. Markers on the origami surface (an asymmetrical L shape) reveal that the orientation of origamis in a polymer is random; i.e., some are "upside down" relative to others. (source: `http://openwetware.org/wiki/Biomod/2014/Kansai/Experiment` and [11]) d) By removing certain stacking bonds at specific locations to create a binary pattern on the edge of a DNA origami, the whole edge becomes a specific "macrobond" that binds most strongly to another edge with the same pattern; in this case, the left side of an origami binds favorably to the right side, and less favorably to another left side. (source: [11]) e) The technique also works to bind 3D origami using 2D patterns of stacking bonds on their faces, using a slightly different motif than in part (d). The placement of the nonspecific bonds gives the entire face a higher affinity for another face with a complementary pattern. (source: [4]) f) One source of error is the matching of many stacking bonds between two faces because of misalignment. (source: [11])

B. Statement of model and main result

Although inspired by work in DNA nanotechnology, the design of specific macrobonds formed by geometric arrangements of nonspecific bonds is fundamental and likely to be part of the future of nanotechnology, even if based on substrates other than DNA. We abstract away several details of DNA origami in mathematically formulating the problem.

We imagine each 2D face of a monomer (e.g., a DNA origami) is a discrete $n \times n$ square, denoted $S_n = \{0, 1, \ldots, n-1\}^2$, with n representing the placement resolution of nonspecific bonds, called *patches*. A *macrobond* is a subset $M \subseteq S_n$. Given a vector $\vec{v} \in \mathbb{Z}^2$, $M + \vec{v} = \{ \vec{m} + \vec{v} \mid \vec{m} \in M \}$ denotes M translated by $\vec{v}$. A parameter $w \in \{2, \ldots, n^2\}$ denotes the *target strength*. A parameter $\lambda \in \{1, \ldots, w-1\}$ denotes the *orthogonality strength threshold*. An (n, w, λ) *geometric orthogonal code* is a set of macrobonds $\mathcal{M} = \{M_1, \ldots, M_\ell\}$, where each $M_i \subseteq S_n$ and $|M_i| = w$, so that for all $1 \leq i < j \leq \ell$, two conditions hold:

low cross-correlation: $\forall \vec{v} \in \mathbb{Z}^2$, $|M_i \cap (M_j + \vec{v})| \leq \lambda$.
low autocorrelation: $\forall \vec{v} \in \mathbb{Z}^2 \setminus \{\vec{0}\}$, $|M_i \cap (M_i + \vec{v})| \leq \lambda$.

Informally, an (n, w, λ) geometric orthogonal code is a set of macrobonds with n^2 available space for potential binding sites, total binding strength w, and spurious binding strength limited to at most λ. As with any code, the goal is to maximize the number of codewords $|\mathcal{M}|$. The main result of this paper is that for all $n, \lambda \in \mathbb{Z}^+$ with n a prime power and $2 \leq \lambda \leq n$, there exists an efficiently computable (n, n, λ) geometric orthogonal code $\mathcal{M}$ with $|\mathcal{M}| = 2n^{\lambda-1} - 2n^{\lambda-2}$.

Examine the physical implementation of patches shown in Figure 1(e), and observe that a "bump" patch on one macrobond cannot insert into a "hole" patch on another macrobond if they are rotated relative to each other, *unless* the rotation is by 180°, i.e., flipped along each axis.[2]

[2]There are physical reasons to dismiss this possibility, since the energy of a stacking bond appears to weaken if the relative rotation angle of the two DNA phosphate backbones are rotated 180° relative to each other: see the image on the right of Figure 1(c), where only two of four possible orientations appear. The other two would put the L pattern *underneath* the origami; presumably they are absent due to weaker stacking energy of rotated helices. Nonetheless, if one imagines a mixture of different angles being used in different patches, rather than only one angle as in [11], then it may be reasonable to assume a worst-case scenario in which a 180° rotation could make any of the patch backbone angles match between two macrobonds.

(Otherwise the blunt ends will not be flush.) To model this, define an (n, w, λ) *geometric flipping orthogonal code* to be an (n, w, λ) geometric orthogonal code that, defining $\mathsf{flip}(M_i) = \{ (n-1-x, n-1-y) \mid (x,y) \in M_i \}$, also obeys $|\mathsf{flip}(M_i) \cap (M_j + \vec{v})| \leq \lambda$ for all $1 \leq i \leq j \leq \ell$ and all $\vec{v} \in \mathbb{Z}^2$. We demonstrate codes that obey this extra constraint as well.

C. Related work

The most directly related theoretical work are the binary *optical orthogonal codes* studied by Chung, Salehi, and Wei [3], which minimize the number of overlapping 1's (analogous to our nonspecific patches) between codewords; overlapping 0's (analogous to neutral non-binding sites) are not penalized. Also, these codes consider all possible translations of codewords; a codeword requires orthogonality not only to translations of other codewords (cross-correlation) but also to nonzero translations of itself (autocorrelation).

The major difference between optical orthogonal codes and our work is the geometric nature of our codes.[3] Each codeword represents a 2D face of a 3D molecular structure, so translations in both x and y coordinates must be considered.

One could imagine applying the optical orthogonal codes of [3] directly to our problem setting. Indeed, every (n^2, n, λ) optical orthogonal code is in fact an (n, n, λ) geometric orthogonal code by interpreting each 1D codeword as the concatenation of the n rows of a 2D codeword. However, applied to this scenario, the upper bounds on possible code size proven in [3] are lower than our lower bounds (though within a factor 2) in Theorems II.1 and II.2.[4] In other words, an optical orthogonal code is more constrained, and these constraints imply that the codes we construct are provably larger than any possible optical orthogonal code with the same parameters.

Two-dimensional optical orthogonal codes have been studied [1], [6], [10]. The 2D nature of these codes reflects the fact that two different variables (e.g., time and wavelength) determine where 1's and 0's appear in the codeword. However, these techniques do not apply directly to our problem, since they consider only translation in one dimension (time) while we must consider simultaneous translations along both dimensions.

Huntley, Murugan, and Brenner [5] have also studied specific engineered molecular bonds from an information theory perspective. They study a different model in which translation is disallowed. They study "color" coding: extending the patches to allow some specificity, so that only equal-color patches can bind; see Section III for a discussion of this issue. They compare with "shape" coding: allowing the shape of a 1D edge to be nonflat, thus providing steric hindrance as an additional mechanism to prevent unintended binds; also discussed as an open question in Section III. They run simulations to show that randomly selected shape codes perform better (i.e., for a given orthogonality have greater size) than randomly selected color codes.

[3] Another difference with our setting is that optical orthogonal codes are more stringent in defining orthogonality under translation. In [3], translations are assumed to be modulo the codeword size, whereas in our setting such "wrapping" does not make sense: a molecular structure α moving off the end of another structure β does not appear on the opposite side of β, hence could not contribute to the binding strength. They also use different parameters to bound autocorrelation and cross-correlation, but for the setting we are modeling, these both correspond to spurious molecular bonds, so it makes sense to use the same threshold for each.

[4] Applying the Johnson-like bound of Theorem 1 of [3] to our setting gives (assuming λ is constant with respect to n) the upper bound $n^{\lambda-1} + O(n^{\lambda-2})$, compared to our lower bound $2n^{\lambda-1} - 2n^{\lambda-2}$ (Theorem II.1).

II. Results

A. Lower bounds

The following theorem shows how to construct geometric orthogonal codes *without* the flipping constraint.

Theorem II.1. *For each prime power $n \in \mathbb{Z}^+$ and $\lambda \in \{2, \ldots, n-1\}$, there exists an (n, n, λ) geometric orthogonal code of size $2n^{\lambda-1} - 2n^{\lambda-2}$.*

Proof. Translation by a vector $\vec{v} = (\delta_x, \delta_y)$ with $|\delta_x|$ or $|\delta_y| \geq n$ implies correlation is 0. So assume $|\delta_x|, |\delta_y| < n$.

Let $\mathbb{F}_n$ denote the finite field of order n, where n is a prime power, and let $0_{\mathbb{F}_n}, 1_{\mathbb{F}_n} \in \mathbb{F}_n$ respectively represent the additive and multiplicative identity elements of $\mathbb{F}_n$. Each macrobond is defined by a degree-λ polynomial $p(x) = a_\lambda x^\lambda + a_{\lambda-1}x^{\lambda-1} + \cdots + a_1 x + a_0$ over $\mathbb{F}_n$, where each coefficient $a_i \in \mathbb{F}_n$, $a_\lambda \neq 0_{\mathbb{F}_n}$, $a_{\lambda-1} = 0_{\mathbb{F}_n}$, and $a_0 \in \{0_{\mathbb{F}_n}, -1_{\mathbb{F}_n}\}$. Put the elements of $\mathbb{F}_n$ in 1-1 correspondence with the integers $\{0, 1, \ldots, n-1\}$ by $f : \mathbb{F}_n \to \{0, 1, \ldots, n-1\}$ defined by the recurrence $f(0_{\mathbb{F}_n}) = 0$ and $f(x + 1_{\mathbb{F}_n}) = f(x) + 1$. For $x \in \mathbb{F}_n$, associate x with an index $f(x)$ specifying a row or column of the macrobond. For ease of notation, $f(x)$ is denoted by x and context distinguishes whether $x \in \{0, \ldots, n-1\}$ or $x \in \mathbb{F}_n$.

For a given polynomial p, the corresponding macrobond $M_p = \{ (x, p(x)) \mid x \in \mathbb{F}_n \}$, i.e., a patch in column x on row $p(x)$. There are $2(n-1)n^{\lambda-2}$ such polynomials, one for each sequence of $n-1$ choices for a_λ, $n^{\lambda-2}$ choices for $a_{\lambda-2}, a_{\lambda-3}, \ldots, a_1$, and 2 choices for a_0.

All that remains is to prove that no pair of macrobonds have correlation more than λ under translation by some vector (δ_x, δ_y). Alternatively, suppose there exist macrobonds M_p and M_q with $p(x) = \sum_{i=0}^{\lambda} a_i x^i$ and $q(x) = \sum_{i=0}^{\lambda} b_i x^i$ and correlation $> \lambda$. By our choice of f, translating the macrobond M_p to $M_p + (\delta_x, \delta_y)$ results in a macrobond represented by the polynomial $p(x + f^{-1}(\delta_x)) + f^{-1}(\delta_y)$. As mentioned, for ease of notation, interpret (δ_x, δ_y) as an element of $\mathbb{F}_n^2$.

By the fundamental theorem of algebra, $p(x + \delta_x) + \delta_y = q(x)$ for all $x \in \mathbb{F}_n$. Expanding the two leading terms of $p(x + \delta_x) + \delta_y$ and $q(x)$ implies $\delta_x \lambda a_\lambda = a_{\lambda-1} + \delta_x \lambda a_\lambda = b_{\lambda-1} = 0_{\mathbb{F}_n}$. Then since $\lambda, a_\lambda \neq 0_{\mathbb{F}_n}$, it must be that $\delta_x = 0_{\mathbb{F}_n}$.

Expanding all terms of $p(x) + \delta_y$ and $q(x)$ implies $\delta_y = \delta_y + a_0 + \sum_{i=1}^{\lambda} a_i \delta_x^i = b_0 \in \{0_{\mathbb{F}_n}, -1_{\mathbb{F}_n}\}$. So $\delta_y \in \{0, n-1, -(n-1)\}$. Again by the fundamental theorem of algebra, for any $c \in \mathbb{F}_n$, $p(x) = c$ for at most λ values of x. So $p(x) = 0$ and $p(x) = n-1$ for at most λ values of x each. So if $\delta_y \in \{n-1, -(n-1)\}$, then correlation is $\leq \lambda$. So $\delta_y = 0$ and $(\delta_x, \delta_y) = (0, 0)$.

Thus every macrobond has low auto-correlation. Since $\delta_x = \delta_y = 0_{\mathbb{F}_n}$, it must be that $p = q$ and thus $M_p = M_q$, so any pair of *unequal* macrobonds have low cross-correlation. □

We now show how to obtain a geometric *flipping* orthogonal code, using similar techniques to the proof of Theorem II.1.

Theorem II.2. *For each prime power $n \in \mathbb{Z}^+$ and $\lambda \in \{2, \ldots, n-1\}$, there exists an (n, n, λ) geometric flipping orthogonal code of size $((n-1)n^{\lambda-2} - n^{\lceil \lambda/2 \rceil})/2$ if n is odd, and $((n-1)n^{\lambda-2} - 2^{\lfloor \lambda/2 \rfloor + 1} n^{\lceil \lambda/2 \rceil})/2$ if n is even.*

Proof. The construction is a modification of Theorem II.1 obtained by taking a subset of the code that avoids high correlation in the new orientation. For the sake of contradiction, assume that there exist polynomials p and q and a translation vector (δ_x, δ_y) such that M_q has correlation $> \lambda$ with $\mathsf{flip}(M_p) + (\delta_x, \delta_y)$. By definition, $\mathsf{flip}(M_p) = \{ (n-1-x, n-1-p(x)) \mid x \in \mathbb{F}_n \} = \{ (x, n-1-p(n-1-x)) \mid x \in \mathbb{F}_n \}$, so by the fundamental theorem of algebra, $n-1-p(n-1-x+\delta_x)+\delta_y = q(x)$ for all $x \in \mathbb{F}_n$.

The $\lambda - 1$ term of q is 0 by construction. Expanding the two leading terms of $n-1-p(n-1-x+\delta_x)+\delta_y$ with this constraint implies that $a_\lambda \lambda (n-1+\delta_x)(-1)^\lambda = 0$. Then since $\lambda, a_\lambda, (-1)^\lambda \neq 0$, it must be that $n-1+\delta_x = 0$. So $\delta_x = 1$ and $n-1-p(n-1-x+\delta_x)+\delta_y = n-1-p(-x)+\delta_y$.

Restrict the code to only macrobonds with polynomials $\sum_{i=0}^{\lambda} a_i x^\lambda$ such that $a_0 = 0_{\mathbb{F}_n}$, i.e., a_0 may no longer be $-1_{\mathbb{F}_n}$. Expanding all terms of $n-1-p(-x)+\delta_y$ and $q(x)$ leads to $n-1+\delta_y = n-1+a_0+\delta_y = b_0 = 0$. So $\delta_y = 1$ and thus $n-1-p(n-1-x+\delta_x)+\delta_y = n-p(n-x) = -p(-x) = q(x)$. Expanding these polynomials implies that for all i, $a_i(-1)^{i+1} = b_i$.

Define a *complement* of a polynomial $p(x)$ to be $\sum_{i=1}^{\lambda} a_i(-1)^{i+1}x^i$, i.e. the polynomial $q(x)$ such that $-p(-x) = q(x)$. As the above shows, the current code allows two macrobonds to have correlation $> \lambda$ points if one of them is flipped, but only if the two corresponding polynomials are complements. Observe that every polynomial has a unique complement, and some polynomials are complements of themselves. Self-complement polynomials have auto-correlation more than λ and complementary pairs have cross-correlation more than λ. A flipping code can be obtained by taking any subset of the code that contains no polynomial and its complement.

A self-complementary polynomial is one in which for all even i, $a_i + a_i = 0$. If n is odd, this occurs only if $a_i = 0$, and if n is even, then it also occurs if $a_i = n/2$ (i.e., the field element $\sum_{i=1}^{n/2} 1_{\mathbb{F}_n}$). Thus, the number of self-complementary polynomials is at most $n^{\lceil \lambda/2 \rceil}$ when n is odd and, since there are two choices for each even i coefficient, the number is $2^{\lfloor \lambda/2 \rfloor + 1} n^{\lceil \lambda/2 \rceil}$ when n is even.

So first remove all such self-complementary polynomials from the code. Of all the remaining, each has a unique complementary polynomial; remove one of them arbitrarily, cutting the number of remaining in half. So the flipping code has size $((n-1)n^{\lambda-2} - n^{\lceil \lambda/2 \rceil})/2$ for odd n, and $((n-1)n^{\lambda-2} - 2^{\lfloor \lambda/2 \rfloor + 1} n^{\lceil \lambda/2 \rceil})/2$ for even n. □

B. Upper bounds

The following theorem shows an upper bound on the size of a geometric orthogonal code.

Theorem II.3. *For constant λ, any (n, w, λ) geometric orthogonal code has size at most $16\lambda^{\lambda-1} n^{2\lambda}/w^{\lambda+1} + O(n^{2\lambda-1}/w^{\lambda+1})$.*

Proof. The approach is as follows. First, obtain an upper bound on the number of $(\lambda+1)$-patch arrangements in an $r \times c$ region that are pairwise distinct under translations. Second, observe that every macrobond in a (n, w, λ) code induces $\binom{w}{\lambda+1}$ such $(\lambda+1)$-patch arrangements. Dividing the former by the latter yields an upper bound on the number of macrobonds in any (n, w, λ) code.

Consider the number of $(\lambda+1)$-patch patterns with a $r \times c$ bounding box. Every such pattern has a patch incident to each side of the bounding box in one of 6 configurations, e.g., two in opposite corners, two in adjacent corners and one along the opposite edge, etc. The greatest number of patch placements is possible when just two patches are incident to the bounding box, placed in opposite corners. The 6 configurations have 16 total symmetries, thus there are at most $16(rc)^{\lambda-1}$ $(\lambda+1)$-patch patterns with a $r \times c$ bounding box.

Then the number of $(\lambda+1)$-patch patterns in a $n \times n$ region that are pairwise distinct up to translation is at most $\sum_{r=1}^{\lambda} \sum_{c=1}^{\lambda} 16(rc)^{\lambda-1} = 16n^{2\lambda}/\lambda^2 + O(n^{2\lambda-1})$, with λ constant. Now observe that $\binom{w}{\lambda+1} \geq (w-\lambda-1)^{\lambda+1}/(\lambda+1)^{\lambda+1} = w^{\lambda+1}/\lambda^{\lambda+1} + O(w^\lambda)$, again where λ is a constant. Dividing the former by the latter yields the desired bound. □

Corollary II.4. *For constant λ, any (n, n, λ) geometric orthogonal code has size at most $16\lambda^{\lambda-1} n^{\lambda-1} + O(n^{\lambda-2})$.*

Note that the upper bound of Corollary II.4 asymptotically matches the lower bound of Theorem II.1.

Although simple and efficiently computable, it is worth asking if the technique of Theorem II.1 is overkill, compared to the most obvious way of attempting to generate codes: picking codewords at random. The next theorem shows that this approaches yields much smaller codes.

Theorem II.5. *With probability at least $1/2$, a randomly selected set of $(\lambda+1)(\sqrt{2n^{(\lambda+1)}\ln(2)}+1)/n$ macrobonds with one patch per column is not a (n, n, λ) geometric orthogonal code.*

Proof. Consider placements of one patch per column in nonoverlapping *blocks* of $\lambda+1$ consecutive columns as selecting letters in an alphabet of size $n^{\lambda+1}$. Because the blocks are nonoverlapping, each letter selection is independent. Then each macrobond is partially specified by $\lfloor n/(\lambda+1) \rfloor$ letters defining the patch placements in the first $(\lambda+1)\lfloor n/(\lambda+1) \rfloor$ columns, where any pair of bonds containing the same letter, including a bond and itself, have correlation $> \lambda$.

The chance that a sequence of k randomly selected letters from an alphabet of size $n^{\lambda+1}$ does not contain a repeated letter is $\prod_{i=0}^{k-1}(1 - i/n^{\lambda+1}) \leq \prod_{i=0}^{k-1} e^{-i/n^{\lambda+1}} \leq e^{-\sum_{i=0}^{k-1} i/n^{\lambda+1}} \leq e^{-(k-1)^2/(2n^{\lambda+1})}$. Thus the probability that it *does* contain a repeated letter is at least $1-e^{-(k-1)^2/(2n^{\lambda+1})}$. By algebra, the inequality $1/2 \leq 1 - e^{-(k-1)^2/(2n^{\lambda+1})}$ holds provided $\sqrt{2n^{\lambda+1}\ln(2)} + 1 \leq k$.

Since each macrobond induces $\lfloor n/(\lambda+1) \rfloor$ letters, a set of $(\lambda+1)(\sqrt{2n^{(\lambda+1)}\ln(2)}+1)/n$ macrobonds induces $(\lambda+1)(\sqrt{2n^{(\lambda+1)}\ln(2)}+1)/n \cdot \lfloor n/(\lambda+1) \rfloor \geq \sqrt{2n^{(\lambda+1)}\ln(2)}+1$ letters and thus with probability $\geq 1/2$ contains two with correlation $> \lambda$. □

III. Open questions

A number of directions for future work suggest themselves.

1) We chose to define a macrobond as a subset of an $n \times n$ square for convenience and because it worked well with our proof technique using polynomials over finite fields. An obvious generalization is to find geometric orthogonal codes that work over $n \times m$ rectangles for $n \neq m$.
2) Our lower bound technique works for $w = n$, where w is the desired number of patches per macrobond. Can we generalize to arbitrary w, possibly matching the upper bound of Theorem II.3?
3) In defining orthogonality of two macrobonds, we allow them to translate relative to each other, but not to rotate, except by 180°. Imagining a 2D macrobond based on generalizing the scheme of Figure 1(d) in the most obvious way — in which the blunt ends face orthogonal to the origami face rather than parallel to it as in Figure 1(e) — would imply that the shape of a patch would no longer automatically disallow non-180° rotations.[5] Thus, it would be interesting to consider adding a rotational constraint to the definition of geometric orthogonal code.
4) We model patches as completely non-specific bonds. DNA blunt ends are *relatively* nonspecific, but even so, a GC/CG stack for instance is significantly stronger than an AT/TA stack. The macrobonds employed in [11] actually use only GC/CG stacks to enforce uniformity (other stack types are allowed in [4]). One can imagine ways to add some specificity to patches by choice of terminating base pair, or possibly by using DNA sticky ends in place of stacking bonds. The problem would then be better modeled by defining a macrobond to be a function $M : S_n \to C \cup \{\mathsf{null}\}$, where C is a finite set of "colors", and null represents the absence of a patch. Then, two aligned patches with colors $c_1, c_2 \in C$ would have strength $\mathsf{str}(c_1, c_2)$ (for C being the set of possible terminating base pairs, $\mathsf{str}(c_1, c_2)$ is in Table 1 of [8]).
5) Our formalization of the concept resembles Figure 1(e) more than 1(d) in the sense that there are two types of faces ("bump" type faces and "dent" type faces), and a macrobond is always formed between opposite-type faces. In contrast, macrobonds formed in Figure 1(d) are between faces of the same "type." In this case, one could imagine a macrobond coming into contact with another copy of itself through flipping along one axis only.
6) Woo and Rothemund [11] study a related technique for creating specific macrobonds. It uses shape complementarity rather than patch placement, in which the shape of an edge can sterically prevent patches from bonding. It would be interesting to study codes based on this technique (see also [5]).

[5] As mentioned in Section I-B, there are physical reasons to conjecture that such rotations have weaker stacking bonds than the "standard" rotation, but it would be interesting to model a worst-case scenario in which rotated macrobonds could align many patches.

Acknowledgements. The authors are grateful to Matt Patitz for organizing the 2015 University of Arkansas Self-assembly Workshop, where this project began, and to the participants of that workshop and support from NSF grant CCF-1422152. We are also indebted to the participants of the 2015 Workshop on Coding Techniques for Synthetic Biology at the Univ. of Illinois, Urbana-Champaign, especially Han Mao Kiah, Farzad Farnoud, Urbashi Mitra, and Olgica Milenkovic, for bringing optical orthogonal codes to our attention and giving valuable feedback. We thank Paul Rothemund and Sungwook Woo for explaining known physical properties of stacking bonds. D.D. was supported by NSF grants CCF-1219274, CCF-1442454, and the Molecular Programming Project under NSF grant 1317694. We thank anonymous reviewers for comments and pointing out mistakes in an early draft.

References

[1] Ram Chandra Singh Chauhan, Yatindra Nath Singh, and Rachna Asthana. Design of two dimensional unipolar (optical) orthogonal codes through one dimensional unipolar (optical) orthogonal codes. Technical Report 1309.2254, arXiv, 2013.

[2] Yuan-Jyue Chen, Benjamin Groves, Richard A. Muscat, and Georg Seelig. DNA nanotechnology from the test tube to the cell. *Nature Nanotechnology*, 10:748–760, 2015.

[3] Fan Chung, Jawad A. Salehi, and Victor K. Wei. Optical orthogonal codes: Design, analysis and applications. *IEEE Transactions on Information Theory*, 35(3):595–604, May 1989.

[4] Thomas Gerling, Klaus F. Wagenbauer, Andrea M. Neuner, and Hendrik Dietz. Dynamic DNA devices and assemblies formed by shape-complementary, nonbase pairing 3D components. *Science*, 347(6229):1446–1452, 2015.

[5] Miriam H. Huntley, Arvind Murugan, and Michael P. Brenner. The information capacity of specific interactions. Technical Report 1602.05649, arXiv, 2016.

[6] Yu-Chei Lin, Guu-Chang Yang, Cheng-Yuan Chang, and Wing C. Kwong. Construction of optimal 2D optical codes using $(n, w, 2, 2)$ optical orthogonal codes. *IEEE Transactions On Communications*, 59:194–200, 2011.

[7] Paul W. K. Rothemund. Folding DNA to create nanoscale shapes and patterns. *Nature*, 440(7082):297–302, 2006.

[8] John SantaLucia Jr and Donald Hicks. The thermodynamics of DNA structural motifs. *Ann. Rev. Biophys. Biomol. Struct.*, 33:415–440, 2004.

[9] Nadrian C. Seeman. Nucleic-acid junctions and lattices. *Journal of Theoretical Biology*, 99:237–247, 1982.

[10] E.S. Shivaleela, A. Selvarajan, and Talabattula Srinivas. Two-dimensional optical orthogonal codes for fiber-optic CDMA networks. *Journal of Lightwave Technology*, 23(2):647–654, Feb 2005.

[11] Sungwook Woo and Paul W. K. Rothemund. Programmable molecular recognition based on the geometry of DNA nanostructures. *Nature Chemistry*, 3:620–627, 2011.