Design Automation for DNA Self-Assembled Nanostructures

Constantin Pistol
Department of Computer Science
Duke University
Durham, NC 27708
costi@cs.duke.edu

Alvin R. Lebeck
Department of Computer Science
Duke University
Durham, NC 27708
alvy@cs.duke.edu

Chris Dwyer
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708
dwyer@ece.duke.edu

ABSTRACT

DNA self-assembly is an emerging technology with potential as a future replacement of conventional lithographic fabrication. A key challenge is the specification of appropriate DNA sequences that are optimal according to specified metrics and satisfy various design rules. To meet this challenge we developed a thermodynamics-based design automation tool to evaluate the vast DNA sequence space (2.8k base pairs) and select appropriate sequences. We use this tool to design DNA nanostructures that were previously impossible with existing text distance based tools. We also show that for nanoscale structures our approach produces superior results compared to existing tools.

Categories and Subject Descriptors

J.6 [**Computer Aided Engineering**]: Computer-aided design (CAD). J.3. [**Physical Sciences and Engineering**]: Chemistry, Engineering, Physics.

General Terms: Algorithms, Design, Experimentation, Theory.

Keywords: DNA self-assembly, nanostructure design, optimized self-assembly.

1. INTRODUCTION

DNA self-assembly is an emerging method for the bottom-up fabrication of nanoscale computing systems. The precise binding rules of DNA enable creation of nanostructures with minimum pitch on the order of a few nanometers. These nanostructures can be used to place and interconnect nanoscale components (e.g., crossed carbon nanotube FETs, ring-gated FETs, nanowires).

The challenge in creating DNA nanostructures is to specify the appropriate DNA sequences such that the desired structure (geometry) forms and is thermodynamically stable. To meet this challenge, DNA self-assembly can exploit the common technique of composing a small set of relatively simple motifs to create more sophisticated structures. Many parts of this design process can benefit from design automation. However, in this paper we focus on the key aspect of designing the DNA sequences that control how motifs can bind with each other. Specifically, we seek to find DNA sequences that minimize the strength of unintentional interactions with the other motifs in the set while maximizing the strength of intentional interactions.

This paper presents our approach to evaluate the sequence design space to create a fixed size 60nm X 60nm grid with 20nm pitch. This structure is composed through a hierarchical assembly of motifs. We focus on the design of the final assembly step that combines 16 cruciform motifs (arranged 4x4) to form the final grid structure. For this structure, we must determine the best 96 sequences that satisfy the structural and stability metrics. To accomplish this we implemented an optimization algorithm that is aware of both intentional and unintentional interactions and exploits parallelism to rapidly evaluate the large sequence design space.

We have experimentally verified our method by designing, synthesizing, and assembling the target nanostructure and characterizing it with atomic force microscopy (AFM). We also show that our optimization algorithm produces superior sequences for a 2x2 grid than sequences produced using conventional text-based sequence comparison or random sequence selection.

The remainder of this paper is organized as follows. Sections 2 and 3 provide background on DNA Self-Assembly and DNA motifs, respectively. The metrics for DNA designs are described in Section 4 and our target nanostructures are presented in Section 5. Section 6 describes various design automation approaches and we evaluate these designs in Section 7.

2. DNA Self-Assembly

DNA is an acronym that stands for a class of chemicals known as deoxyribonucleic acids and is widely studied in the context of molecular genetics. We are concerned primarily with DNA as a substrate for fabricating nanostructures, and thus provide a brief review in this context.

DNA's basic building blocks—called a nucleotide—are composed of a phosphodiester covalently bound to a nucleoside or a derivative of a deoxyribose sugar and either a purine or pyrimidine nucleobase. The nucleobases commonly used in DNA self-assembly are the purines: adenine (A) and guanine (G), and the pyrimidines: thymine (T) and cytosine (C). The nucleotides bind to each other to form a linear chain through the phosphodiester bonds. This represents a so-called single-stranded DNA molecule. Geometrically compatible single strands can wrap around each other to form the well-known helical structure, or double stranded molecule (i.e., a double helix).

The double stranded DNA structure is most stable when the pairwise nucleobase interactions are "complementary", i.e., if A pairs with T and G pairs with C. Under these conditions each base

DAC 2006, July 24–28, 2006, San Francisco, California, USA.

pair (bp) is approximately 2 nm wide (diameter of the helix) and on average 0.34 nm long (along the strand, per base). The helical twist of the two strands (in the most common form) is such that a full turn occurs between every 10th and 11th base. Further, the stability of this interaction is only approximately linear per base and depends on neighboring mismatch or complementary interactions [1]. The stability and exact dimensions, orientation, and form of the nucleobase interaction depend on several factors including the pH of the solution and local properties of the DNA.

2.1 Thermodynamics

The central theme in the use of static self-assembly for nanoscale fabrication is the application of an external control over an otherwise spontaneous reaction to direct its outcome [2]. This control directs the assembly of materials into structures that are interesting and relevant to a target design problem. In the context of computer system fabrication the self-assembly is used to direct the formation of switching devices (e.g., transistors and wires) to create logic circuitry, memory, and I/O interfaces.

The temperature of the reaction volume (i.e., the solution) is a simple control in DNA self-assembly. This follows from the experimental evidence that demonstrates the formation of double helices from single strands of complementary DNA as the solution temperature is changed from high to low. The melting temperature (T_m) of a DNA strand is the temperature at which the transition from single strands to double strands has reached 50%. That is, half of the single strands in solution are bound to their complementary strand when the solution temperature is exactly the melting temperature of the strand. The T_m of two strands is dependent on their sequences and the degree to which they are complementary. This simple picture is complicated by the introduction of multiple sequences in solution. Further, the time evolving dynamics of these interactions are still under study [3]. It is the richness of this interaction that is at the root of why DNA can be used to form complex nanostructures.

2.2 Sequence Design

A strand of DNA obeys certain thermodynamic behavior, most importantly that double strands form at temperatures below the T_m of the constituent single strands, and this interaction can be complex if multiple unique (sequences) DNA strands are in solution. Specification of the strand sequences provides external control over the self-assembly process (through temperature control) and determines the formation of structures (through complementarity). Good sequence design leads to a minimization of sequence mismatches, or unintentional interactions between strands of similar but not perfectly complementary sequences, at a given temperature and therefore produces a higher yield of the target structures.

Sequence design is important because it determines many aspects of the target DNA nanostructure (e.g., geometry and stability). Therefore it is critical to have good methods for choosing sequences. One approach is to use abstractions to create increasingly sophisticated structures.

3. DNA Motifs

Complex designs are often created using a relatively small set of common building blocks—called motifs. DNA self-assembly can exploit this same design principle to hierarchically create more sophisticated aperiodic structures. For DNA there are many possible motifs, however we focus on only a few in the context of our target nanostructure (see Section 5). Motifs include junctions that enable three or more double stranded helices of DNA to interact and thus form specific structures (e.g., a triangle, a corner, etc.) Another important motif is a single strand of DNA protruding from a double stranded helix—called a sticky-end.

Two motifs with complementary sequences on their sticky-ends will bind to form a composite motif. These composite motifs may also have embedded sticky-end motifs and thus can also bind with other composite motifs to form another, larger, composite motif. This results in a hierarchical structure for motifs.

The cruciform motif is composed from three smaller motifs: a core, 4 shells, and 4 arms (each arm contains two 5-nt sticky-ends).

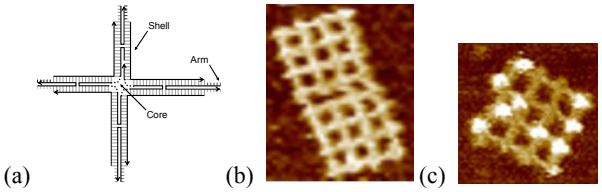

Figure 1 *– (a) Schematic of a cruciform composite motif, (b) AFM image of a hierarchical 8x4 grid, and (c) a protein-patterned 4x4 grid. Each cavity in (a) and (b) is ~20nm wide.*

Figure 1 shows: (a) a schematic of the cruciform motif, (b) an AFM image of a hierarchical 8x4 grid, and (c) a protein-patterned nanostructure each developed in our laboratory using the methods described in this paper [4]. Although motifs provide an easy abstraction for reasoning about DNA nanostructures, there are many potential issues related to sequence dependent physical (structural) properties. For example, the above cruciform motif has a slight curvature in three-space, thus composite motifs formed with this motif must account for this curvature to ensure the desired final geometry is formed. Furthermore, the structural properties of DNA sequences can create strain in the final structure which can prohibit proper formation.

There are many design considerations that must be accounted for in creating a DNA nanostructure. When combined with the vast sequence space, these considerations motivate the need for computer aided design methods.

4. Metrics and Design Rules

The number of possible nanostructures that can be fabricated by motif-based DNA self-assembly is large. However, not all nanostructures can be synthesized efficiently due to the geometric and thermodynamic limitations of DNA hybridization. Such limitations require a set of experimentally verified design rules to act as templates for new designs. In this section we describe the metrics and design rules we have used to design and evaluate new nanostructures.

4.1 Metrics

Metrics enable meaningful comparisons between designs and when coupled with baseline or reference designs can be used to predict fabrication yields. Experimental data indicates that these metrics are correlated with high yields. The two metrics we use to evaluate a design are: (i) the average single-interaction energy measure (SEM) and (ii) the target-interaction likelihood measure (TLM).

SEM— The average single-interaction energy measure is an estimate of the thermal stability of the motif interactions in a design. The SEM can be used as a relative measure of stability in terms of temperature. For example, a design with a large SEM will be stable at higher temperatures than a design with a low SEM. The SEM is

calculated by averaging the melting temperature of each interaction between the motifs in a design. Therefore, a large SEM indicates that the average interaction strength between motifs is also large.

While the thermal stability of a structure is important it is clear that if the structure forms incorrectly, yet with high stability, the fabrication process will produce a large fraction of flawed or defective structures. Therefore, a measure of how likely a structure will form must be coupled with the SEM to obtain a complete measure of the assembly process.

TLM— The target-interaction likelihood measure is an estimate of the potential for a design to assemble a correct structure. The larger this value the less likely it is that a design will form incorrectly. This metric is calculated as the average distance from the diagonal on a non-specific vs. specific melting temperature plot for each motif against all other motifs. Motifs that are close to the diagonal (i.e., motifs with strong non-specific interactions) should be avoided since they will likely contribute to the formation of defective structures. This metric will always be positive.

Thus, the SEM and TLM are metrics that enable a consistent thermodynamic framework in which to compare candidate designs. However, they do not provide insight into the geometric or structural quality of a design. The complexity of this problem motivates the use of design rules in choosing the structure and assembly order for a design. The design rules provide a template for the creation of structures from motifs and can guarantee geometric properties when the design is fabricated.

4.2 Design Rules

We have identified two design rules that enable the fabrication of complex nanostructures. The first is the use of a "corrugation" scheme that alternates the direction of each motif's normal vector across the structure. The second design rule describes a "thermal ordering" of motif interactions based on melting temperatures and the desired structure.

4.2.1 Corrugation

The term "corrugation" was first introduced by Liu et al. [5] to describe a method to combat a curling effect observed in large periodic nanostructures. Each motif has an intrinsic curvature that is sequence dependent and may or may not be approximately zero. Further, the curvature can generate a strained structure if a curved motif is forced (by adjacent motifs) into a planar shape.

A sufficiently large accumulation of strain along a vector that crosses the structure can curl the structure into a tubule. To avoid this strain the corrugation design rule specifies that motifs must be placed into the structure with alternating normal vectors such that their sum over the entire structure is minimal. Regardless of the curvature (or lack thereof) in the motif, the accumulation of strain can be minimized by symmetrically arranging motifs in this way. For example, consider a linear array of identical motifs each with some positive curvature. The structure will curl on itself if the motifs are aligned with each other. Instead, if the motif alignment alternates they will form a straight line.

We have generalized this rule for 2D planar structures and apply it as a template for new designs. The remaining design task (Section 6) is to render the generic template into real nucleotide sequences that can be used to fabricate the target structure. While the corrugation design rule will ensure that the resulting structure is planar, we have found that the order in which motifs assemble into the final structure plays a significant role in the defect rate of the fabrication process. This has led us to apply a second design rule that specifies the motif assembly order.

4.2.2 Thermal Ordering

The DNA of each motif has a specific melting temperature below which an interaction (with other motifs) can take place and will be stable. These melting temperatures (estimated by the SEM) can be ordered from high to low and used as a criterion for picking a given arm-sticky-end sequence and motif. Since we can physically constrain the assembly temperature to be monotonically decreasing we can control the order in which motifs will assemble by choosing sequences with descending SEMs.

Similar to the corrugation design rule, the thermal ordering design rule provides a template (in this case an assembly order) for the nanostructure and must ultimately be rendered into real nucleotide sequences. In the next section we describe our target nanostructure and how the metrics and design rules are applied.

5. Target Nanostructure

As a design example, consider a planar grid of motifs like the ones shown in fig. 1. The cruciform motif described in Section 3 is the basic element for the grid and the design must completely specify the nucleotide sequence for each motif. Using the hierarchical strategy we can reduce the complexity of this problem by using known nucleotide sequences for the cruciform motif and focus on the arm-sticky-end sequences.

A 4x4 grid has 16 cruciform motifs and each cruciform requires four pairs of sequences (one pair for each arm per motif). Since the motifs only bind on the interior of the grid a total of 96 arm-sticky-end sequences are required (96-arm). Prior work has been limited to periodic structures where as many as four or five motifs polymerize into 2D arrays; such systems require fewer arm-sticky-end sequences due to the periodic reuse of motifs [6]. A classic example of this is the "AB" system that requires 20 distinct arm sequences for 2 motifs (20-arm). These sequences must ensure that each motif will bind in only 1 of 16 positions in the grid. We can apply the corrugation and thermal ordering design rules as templates for the grid at the outset and use the SEM and TLM estimates to choose from all candidate arm-sticky-end sequences. To maintain an optimal solution we evaluate all possible 5-nucleotide (5-nt) arm-sticky-end sequences. Section 6 describes the sequence design process in detail and Section 7 evaluates the designs.

6. Design Automation Methods

Given the importance of sequence design for self-assembled systems there are a variety of available tools for this purpose [7-11]. However, these tools use heuristics, simplified interaction models (e.g., sequence text distance) or no hierarchies which make them unable to design large systems (i.e., our systems require non-interacting sequences with >1200 base pairs). Even for small problems theses solutions do not generate a sufficient set stick-ends. For example, the 96-arm motif structure is too large for these methods.

An alternative, but trivial, method is to randomly select sequences. Using a random sequence generator, thermodynamic analysis can be used to evaluate the design. The computational effort is low in this case but there are obviously no guarantees on the optimality of the resulting design.

6.1 Thermodynamic Optimization Tool

To overcome the limitations of text distance and random sequence generation we have implemented a new optimization tool. The tool is a parallel implementation of an exhaustive thermodynamic search that can optimize a target design given a target topology and basic motif design against both the SEM and TLM estimators described in Section 4. The outcome is a set of arm-sticky-end sequences that can be used to generate a set of motifs that are optimized to reduce non-specific interactions during the assembly process. The algorithm evaluates each possible arm-sticky-end sequence against all other candidate sequences and motif sequences and maps their mutual interaction. Self-binding and region mismatching of up to 6 consecutive nucleotides are included in the evaluation.

To calculate the thermodynamic interaction of candidate strands (needed for both the TLM and SLM estimators) we used a modified nearest-neighbor algorithm based on the freely available MELTING4 tool [12]. The MELTING4 code was modified to handle internal and terminal base pair mismatches [1, 13]. Terminal mismatches are treated by "padding" all evaluated sequences with a complementary 3-bp region. This better simulates the motif environment and ensures that the ends are complementary. The padding artificially increases the stability of a configuration (slightly) due to the additional matching base-pairs. This systematic bias means that the calculated values are more reliable as a relative measure of sequence melting temperature.

Given the vast sequence interaction space that must be covered, the execution time on a single processor can quickly become prohibitively large as the size of the target structure increases. To overcome this limitation the algorithm partitions the problem into sub-parts which are then executed in parallel on computing clusters and multi-processor machines. This greatly reduces the time needed to perform an optimization run and allows the application to target larger and more complex DNA nanostructures.

The following pseudo-code generates a TLM-optimized sticky-end solution arm set for the 20-arm or 96-arm systems:

```
 1: FindDNASet(seq_length,set_size,fixed_seq)
 2: {
 3:  arms = generate_all_seq(seq_length);
 4:  arms = remove_verboten(arms);
 5:  arms = add_complements(arms);
 6:  arms = remove_duplicates(arms);
 7:  seq_set = concat_set(arms, fixed_seq)
 8:  results = cross_melt(seq_set);
 9:  results = remove(results, fixed_seq);
10:  results = rank_seq(results);
11:  results = sort_by_TLM(results);
12:  top = head(results, set_size);
13:  return = top;  }
```

The parameters are the target sticky-end length (5 for the cruciform motif), the target sequence set size (20 and 96 respectively) and the fixed sequence set (the A and B cores, shells and arm stub strands).

The first step is to generate all possible sequences of the target sticky-end length. In the next step, "verboten" sequences are removed from the problem space. Verboten sequences are sequences known to have unfavorable properties for self-assembly. The parallel cross-melt algorithm is executed and the resulting interaction data is used to rank the strands based on their specificity (TLM). The top sequences represent the TLM-optimal solution set.

6.2 Alternative Designs

Single Core— The quality of the solution set will improve if the number of fixed motif strands is minimized. This is intuitive, given that each fixed strand adds additional constraints on the solution space. For our target system we apply this by using a single core on all motifs rather than the original dual core motifs (A-type and B-type). The tool can optimize for a single core (e.g., A-only or B-only) system by simply modifying the set of fixed sequences dedicated to the motifs (cores, shells, and arms).

Split Core— Our second approach to improve the design method makes use of the two motif types (A and B) in the context of hierarchical assembly. Both the 20-arm and 96-arm systems are assembled in two separated hierarchical steps. In the first step single strands assemble into motifs. In the second step motifs are mixed and due to their sticky-ends they assemble into the target grid-like structure. Thus, the interaction between the sticky-ends and the fixed strands (cores/shells) is critical in the first level of the hierarchy, when motifs are annealed. In the second step (grid anneal) the motifs are assumed to be thermodynamically stable and the major factor becomes the motif sticky-end interactions.

We use this to generate an optimized set of sticky-ends for each motif type separately. This is equivalent to applying the Single Core method for both A and B motif types. The final set is obtained by combining the results of the two optimization runs with common sequences used only once. The A-type sticky-ends will have sub-optimal interaction with the B core/shells (and vice versa), but the hierarchical assembly process guarantees that they are only simultaneously in assembly solution in the second step.

Since the sticky-ends must be unique across the whole system, the effectiveness of this approach depends on the TLM estimates of the resulting solution sets for the A and B motifs. However, the sequences for the A and B core and shell strands were originally designed to be as different as possible in order to minimize their mutual interaction (i.e., large TLM for each motif type). We expect that this will also translate into significantly different solution sets for each core.

The pseudo-code for the Split A/B design is the following:

```
SplitDesign(set_size)
{
 setA = FindDNASet(5,set_size,A_CoreShells);
 setB = FindDNASet(5,set_size,B_CoreShells);
 setA = sort_by_TLM(setA);
 setB = sort_by_TLM(setB);
 for (i=0;i<(set_size/2);i++)
     seq = head(setA,1);
     ret_setA = concat_set (ret_setA,seq);
     setA = remove(setA, seq);
     setB = remove(setB, seq);
     seq = head(setB,1);
     ret_setB = concat_set (ret_setB,seq);
     setA = remove(setA, seq);
     setB = remove(setB, seq);
 }
return = concat_set(ret_setA, ret_setB);
}
```

The solution sets for each core are separately computed and the final set is assembled from the top sequences of the two sets.

SEM Optimization— The above methods were presented in the context of obtaining TLM-optimal sequence sets for low assembly error rates. However, improved structural stability (SEM) could also be an important design goal. For example, the self-assembled system might need to be stable in certain temperature ranges in order to interface with other systems. Our design tool can optionally trade TLM-optimality for SEM-optimality. This trade-off is controlled with an SEM factor (SF) that expands the candidate set (proportionally) of TLM ranked sequences for subsequent ranking

by their SEM estimates. The pseudo-code below illustrates this process:

```
FindDNASet(seq_length,set_size,fixed_seq,
  sem_factor)
{
  ... (lines 3 to 11 from from original)
  ex_set_size = set_size * (1 + sem_factor);
  top = head (results, ex_set_size);
  top = sort_by_SEM(top);
  final = head (top, set_size);
  return = final;
}
```

7. Evaluation

We evaluate the results of each presented method (AB, A-only, B-only and Split A/B) in terms of specificity and stability (TLM and SEM estimates) as applied to a small 20-arm system and a structurally similar but larger 96-arm system (as described earlier). The results are compared with the expected values for a random sequence design as well as the original 20-arm set from [6] which was generated with the widely used text-distance tool SEQUIN [7]. Table 1 shows the results in terms of SEM, average non-specific T_m (NST_m) and TLM for each method.

20-arm	SEM	NST_m	TLM
AB Core	7.12	-6.87	9.81
AB Core Original*	11.77	4.83	4.24
AB Random	10.04	4.01	3.32
A-Only	7.66	-6.44	9.82
B-Only	9.75	-4.68	10.00
B-Only SF=4	14.08	-0.08	9.65
AB Split	9.75	-4.74	9.99
AB Split SF=7	15.75	2.32	9.31
96-arm			
AB Core	6.66	-6.65	9.25
AB Random	10.04	4.01	3.32
A-Only	7.80	-5.83	9.44
B-Only	8.17	-5.68	9.52
B-Only SF=1	11.19	-1.38	9.11
AB Split	8.28	-5.54	9.57
AB Split SF=7	12.24	-1.29	9.15

Table 1 *– 20-arm and 96-arm results. *No 96-arm original exists.*

To evaluate the upper bounds for the SEM and TLM of each method we remove all fixed sequences (cores/shells) and evaluate the best possible sets for each metric. This simulates a theoretical system in which the core and shells do not interact with the sticky-ends. To obtain the highest SEM design we use a large SF factor. The results are shown in table 2.

20-arm	SEM	NST_m	TLM
No Core – rank by TLM	11.93	-3.29	**10.47**
No Core – rank by SEM	**18.94**	6.95	7.22
96-arm			
No Core – rank by TLM	8.96	-5.16	**9.77**
No Core – rank by SEM	**15.90**	3.39	8.05

Table 2 *– Upper bounds on TLM and SEM with 5-nt sticky-ends*

The random sequence method results show a fairly high SEM, which translates into good expected stability for the target assembly. However the overall specificity (TLM) is very low, suggesting that the system is likely to form defective structures when self-assembling.

The SEQUIN-generated 20-arm original design shows a slight increase in both TLM and SEM when compared to random. However, the TLM is still low and this shows that using simple text-distance metrics and heuristics for optimizing sequence sets can lead to uncertain results if low error rates are desired.

The AB Core set generated with our tool shows a significant improvement in specificity. The target structure is thus much more likely to form correctly when using this method. The stability estimate is lower than the original design which means that the system will disintegrate at slightly lower temperatures. However, the SEM factor (SF) can be used to trade specificity for stability and can be used to balance the design.

Figure 2 shows a scatter plot of three sequence sets for the 20-arm system. The diagonal represents the line of zero specificity (TLM=0) where specific T_m equals non-specific T_m. The distance to this line from any given point (sequence) is the TLM. The AB Core sequences are clustered in a series of points situated at roughly similar TLMs. The random and original designs do not show this pattern and include sequences that are situated on the diagonal itself: these sequences are just as likely to base-pair with the core/shells as they are with their complements! These strands are likely to have a particularly disruptive effect on structure formation in their sets and there is some evidence that this is the case [4].

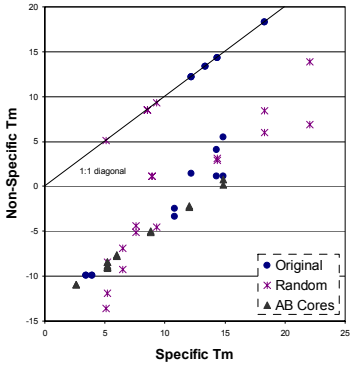

Figure 2 *– AB Core, random and original 20-arm sets*

Figure 3 shows that when the number of fixed sequences (cores/shells) decreases (A-only or B-only vs. AB Cores) the average specificity of the system increases. The specificity of each arm-sticky-end is defined as the minimum of its TLM and the TLM of its complement. (In all 5-nt arm-sticky-end systems there are 600 candidate sequences because 424 of the total possible 1024 sequences are verboten and removed.) This result verifies the intuition that fixed sequences in the motifs restrict the number of high quality (i.e., high specificity) arm-sticky-end sequences.

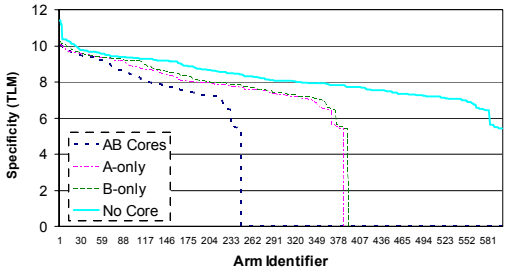

Figure 3 *– Specificity in arm-sticky-end space with different fixed sequence sets*

Figure 3 also shows the systematic bias induced by the thermodynamic optimization tool due to sequence padding. There is an offset of ~5.4 for non-zero TLM values for all the designs due to the always-complementary 3-bp pads used by our tool.

The results of B-only, AB Cores and random methods for the 96-arm design are presented in fig. 4. *Random* has many arm-stick-ends that strongly interact with the core/shells (mapped on the

diagonal). B-only shows the same clustering as AB Cores, but it is on average slightly farther away from the diagonal (higher specificity). The Split AB design is the best performer for TLM-optimized designs, slightly outperforming even the single core B-only method for larger systems. The scatter-plot is very similar to the B-only set in fig. 4 and we omit it for brevity. Figure 5 shows how the SF factor can be used to increase the SEM of a design at the expense of a slightly lower TLM.

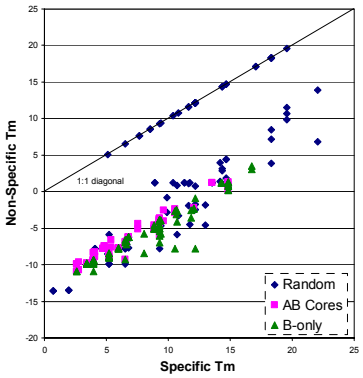

Figure 4 *– B-only, AB Cores and Random for 96-arm sets.*

In fig. 6 two such balanced designs, one based on the B-only method and one on an SEM-targeted Split AB are contrasted with the text-based design of the original AB and with the random method. The theoretical maximum SEM and TLM for 5-nt sticky-end designs are included for comparison. Table 3 lists our best *AB-Cores* arm-sticky-end sequences.

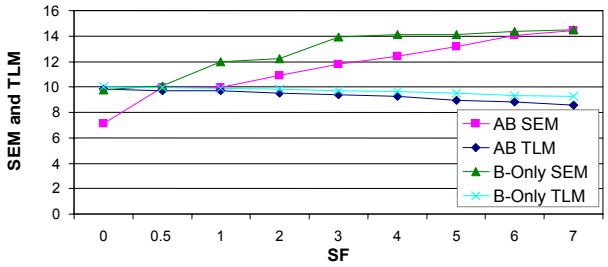

Figure 5 *– Trade-off: specificity can be traded for stability*

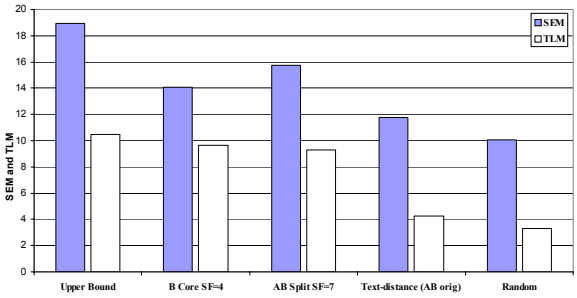

Figure 6 *– 20-arm designs balanced for the TLM and the SEM*

8. Conclusions

The continued scaling of conventional CMOS fabrication techniques faces many challenges that may be overcome by a switch to bottom-up self-assembly. DNA's precise binding rules at small scale (a few nm) make it a potential candidate for future fabrication of nanoscale computing systems. This paper presents a thermodynamics-based computer aided technique for DNA nanostructure design. Compared to existing text-based tools, our approach enables creating structures of previously unattainable size and produces superior designs for small structures.

This work was supported by the NSF (CCR-03-26157), the AFRL (FA8750-05-2-0018), the Duke Provost Common Fund, and equipment donations from Intel and IBM.

Seq.	Comp.	Seq.	Comp.	Seq.	Comp.
CGTGC	GCACG	CCTCG	CGAGG	TATGT	ACATA
CAAGC	GCTTG	ACGAC	GTCGT	TGTAT	ATACA
ACGTC	GACGT	CAGAC	GTCTG	TTAGA	TCTAA
ACAGC	GCTGT	ACTGC	GCAGT	TTACT	AGTAA
TGCAG	CTGCA	TGCTG	CAGCA	TAAGA	TCTTA
CTGTG	CACAG	TGCAC	GTGCA	AATAG	CTATT
AGCTC	GAGCT	AGAGC	GCTCT	AATTC	GAATT
CATGG	CCATG	CTTGG	CCAAG	ATACT	AGTAT
CAATC	GATTG	CATTC	GAATG	TAACT	AGTTA
AATGC	GCATT	ATTGC	GCAAT	TTAGT	ACTAA
AACGT	ACGTT	CTAAC	GTTAG	TACTT	AAGTA
CTTAC	GTAAG	TTACG	CGTAA	TAAGT	ACTTA
TAACG	CGTTA	CATTG	CAATG	TAGAT	ATCTA
ATGAC	GTCAT	ATGCT	AGCAT	TAGAC	GTCTA
TCATG	CATGA	TTGCT	AGCAA	TTCAT	ATGAA
TTGAG	CTCAA	AAGCT	AGCTT	ATTCT	AGAAT
TGCTT	AAGCA	ACTGT	ACAGT	TCAAT	ATTGA
TCACA	TGTGA	AGTAC	GTACT	TTAAC	GTTAA
TACGT	ACGTA	TGTAG	CTACA	AATCT	AGATT
TACTG	CAGTA	AAGTG	CACTT	TGATT	AATCA
TCAAC	GTTGA	TCTGA	TCAGA	TCATT	AATGA
TCTAG	CTAGA	TAGCT	AGCTA	TATGA	TCATA
AATGT	ACATT	ATGTT	AACAT	TTAAG	CTTAA
ATTGT	ACAAT	TTCAA	TTGAA	TTACA	TGTAA
GTTAT	ATAAC	AATAC	GTATT	TTGTA	TACAA
TAATG	CATTA	CAATA	TATTG	TTGAT	ATCAA
TTATG	CATAA	AATTG	CAATT		

Table 3 *- The best 160 x 5-nt arm-sticky-end sequences (and complements) found by our AB-Core method. These arm-sticky-ends are compatible with the tile motif in [4].*

References

[1] N. Peyret, P. A. Seneviratne, H. T. Allawi, and J. SantaLucia, "Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A-A, C-C, G-G, and T-T Mismatches," Biochemistry, vol. 38, pp. 3468-3477, 1999.

[2] G. M. Whitesides and B. A. Grzybowski, "Self-Assembly at All Scales," Science, vol. 295, pp. 2418-2421, 2002.

[3] X. Wang and W. M. Nau, "Kinetics of end-to-end collision in short single-stranded nucleic acids," J. Am. Chem. Soc., vol. 126, pp. 808-813, 2004.

[4] **(a)** C. Dwyer, S. H. Park, T. LaBean, A. Lebeck, "The Design and Fabrication of a Fully Addressable 8-tile DNA Lattice," Proc. of the 2nd Conf. on the Foundations of Nano., pp. 187-191, April 2005. **(b)** S. H. Park, C. Pistol, S. J. Ahn, J. H. Reif, A. R. Lebeck, C. Dwyer, and T. H. LaBean, "Finite-size, Fully-Addressable DNA Tile Lattices Formed by Hierarchical Assembly Procedures," Angewandte Chemie, vol. 45, pp. 735-739, 2006.

[5] D. Liu, S. H. Park, J. H. Reif, and T. H. LaBean, "DNA Nanotubes Self-assembled from TX Tiles as Templates for Conductive Nanowires," in Proc. of the National Academy of Science, vol. 101, pp. 717-722, 2004.

[6] H. Yan, S. H. Park, G. Finkelstein, J. H. Reif, and T. H. LaBean, "DNA Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires," Science, vol. 301, pp. 1882-1884, 2003.

[7] N. C. Seeman, "De Novo Design of Sequences for Nucleic Acid Structural Engineering," Biomolecular Structure & Dynamics, vol. 8, pp. 573-581, 1990.

[8] A. J. Hartemink, D. K. Gifford, and J. Khodor, "Automated Constraint-Based Nucleotide Sequence Selection for DNA Computation." Proc. of the 4th DIMACS on DNA Based Computers, pp. 227-235, 1998.

[9] U. Feldkamp, S. Saghafi, W. Banzhaf, and H. Rauhe, "DNA Sequence Generator: A Program for the Construction of DNA Sequences," in Proc. of DNA7, Lecture Notes in Computer Science, vol. 2340, pp. 23-32, 2001.

[10] P. Yin, et al., "TileSoft: Sequence optimization software for designing DNA secondary structures," TR-CS-2004-09, Duke, 2004.

[11] M. R. Shortreed, S. B. Chang, D. Hong, M. Phillips, B. Campion, D. C. Tulpan, A. Condon, H. H. Hoos, and L. M. Smith, "A thermodynamic approach to designing structure-free combinatorial DNA word sets," Nucleic Acids Research, vol. 22, pp. 4965-4977, 2005.

[12] N. Le Novere, "MELTING, computing the melting temperature of nucleic acid duplex," Bioinformatics, vol. 17, pp. 1226-1227, 2001.

[13] J. SantaLucia Jr. and D. Hicks, "The thermodynamics of DNA structural motifs," Annual Review of Biophysics and Biomolecular Structure, vol. 33, pp. 415-440, 2004.