Reports

GENESUS: A two-step sequence design program for DNA nanostructure self-assembly

Takanobu Tsutsumi¹, Takeshi Asakawa^{1,†}, Akemi Kanegami², Takao Okada², Tomoko Tahira¹, and Kenshi Hayashi¹

¹*Division of Genome Analysis, Research Center for Genetic Information, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan* and ²*Research Institute of Biomolecule Metrology Co., Ltd., Tsukuba, Ibaraki, Japan*

[†]Present address: Business Design and Innovation Laboratory, Corporate R&D, Sony Corporation, Tokyo, Japan

BioTechniques 56:180-185 (April 2014) doi 10.2144/000114157

Keywords: sequence design; DNA nanostructures; self-assembly; dynamic programming algorithm; octahedron; free energy

Supplementary material for this article is available at www.BioTechniques.com/article/114157.

DNA has been recognized as an ideal material for bottom-up construction of nanometer scale structures by self-assembly. The generation of sequences optimized for unique self-assembly (GENESUS) program reported here is a straightforward method for generating sets of strand sequences optimized for self-assembly of arbitrarily designed DNA nanostructures by a generate-candidates-and-choose-the-best strategy. A scalable procedure to prepare single-stranded DNA having arbitrary sequences is also presented. Strands for the assembly of various structures were designed and successfully constructed, validating both the program and the procedure.

Building arbitrary nanostructures using self-assembling DNA strands is a promising bottom-up approach in nano-engineering that is being exploited for various applications (e.g., to create vehicles for drug delivery, molecular machinery, or DNA-computing modules) (1–3). This is because the DNA strands are linear heteropolymers carrying information as nucleotide sequences, and the inter- or intramolecular local contacts of the strands to form rigid helices can be specified using this information. In addition, strands with arbitrary sequences tens of nucleotides long can be chemically synthesized at low cost.

Successful assembly of a target structure by annealing a mixture of DNA strands depends on specific and exclusive segmental pairings at intended positions on the involved strands. In the pioneering work of Seeman, the SEQUIN program was developed to fulfill these requirements (4). A core algorithm of the program composes a set of strand sequences following the n-tuple uniqueness principle, which is realized by non-redundantly picking overlapping oligonucleotides from an exhaustive list of n-mer oligonucleotide pools (called critons or

vocabularies). Similar approaches have been adopted in subsequent sequence selection programs (5–8). These programs have their own advantages, but they require human intervention to run effectively, and the final selection of the sequence set is left ambiguous. Alternatively, random sequences are initially assigned to the strands and then allowed to evolve following the genetic algorithm to maximize the probability of assembling a designed structure that is then evaluated using available thermodynamic parameters (9,10). These methods are highly sophisticated but computationally demanding. They are also not suitable for designing structures having pseudo-knots.

An alternative approach to those described above is the use of naturally occurring or randomly generated sequences as the general-purpose quasi-unique sequences (11–16). The rationale of adopting these panacea sequences is that the overall stability of the designed segmental pairings usually far exceeds that of incidental aberrant pairings. The successful construction of various structures using this strategy demonstrate its usefulness, but the assembly often requires excess staple strands

and extensive annealing time (11–14), or the strategy is likely to be limited to the assembly of single-strand tiles where every assembly step is a short coaxial extension of two or more pre-existing helices so that highly cooperative and accurate assembly is expected (15,16). Also, the use of these panacea sequences always carries the risk of forming unexpected unfavorable pairings since the redundancy or complementarity have not been formally removed in these sequences.

We describe below a new, straightforward program, GENESUS, for the selection of true unique sequences optimized for the self-assembly of arbitrary DNA nanostructures. The program first generates candidate strand sets that are chosen following the n-tuple uniqueness principle; then, it selects the set that has the least chance of forming unwanted pairings using a modified dynamic programming algorithm. We validate the program by constructing self-assembled DNA nanostructures whose sequences are chosen using the program. We also present scalable procedures to prepare single-stranded DNAs with arbitrary sequences that are suitable for building various DNA nanostructures.

Method summary:

To facilitate DNA nanostructure self-assembly, the generation of sequences optimized for unique self-assembly (GENESUS) program has been developed. In addition, a streamlined scalable procedure has been established for preparing single strands from cloned segments carrying arbitrary sequences that serve as scaffolds of the nanostructure.

Materials and methods

Development and implementation of GENESUS

The GENESUS program outlined in Figure 1 was developed using the C programming language. Executables for Windows or LINUX, as well as the source code, are available at the GENESUS site (<https://sites.google.com/site/qgenesus/>). Running the program on a personal computer equipped with an Intel Core 2 Quad CPU at 2.66 GHz and 3.25 GB RAM, it took approximately 1 min to generate sequences for the regular octahedron with helices of 21 bp shown in this work. The program can handle structures with helix species of different lengths and melting temperatures, although all of the structures described in this report had helix species of only one length and T_m value.

Determination of nucleic acid concentrations

The concentrations of single-stranded DNAs were determined by their A_{260} values, using the extinction coefficients of their component nucleotides (17). The concentrations of RNA and double-stranded DNA were determined by measuring A_{260} , assuming that $A_{260} = 1$ for solutions containing 40 mg/mL RNA or 50 mg/mL double-stranded DNA.

Strand preparations

All staples and PCR primers were purchased as cartridge-purification grade from Invitrogen (Carlsbad, CA) and were then dissolved and diluted in 1× TE buffer (10 mM Tris-HCl, 1 mM EDTA) to 10 μ M stocks. The scaffold strands were prepared as depicted in Figure 2. The fragments containing scaffold sequences were synthesized and blunt-end cloned into the *EcoRV*

site of pUC57 (GenScript, Piscataway, NJ). The plasmid was linearized by *Pst*I (New England Biolabs, Ipswich, MA), and the regions containing scaffold sequences were amplified by PCR using primer pairs where one primer harbored the T7 promoter sequence (CCAAGCCTTCTAATAC-GACTCACTATAGGGAGA) at its 5' end. The PCR products were purified using the QIAquick PCR Purification Kit (QIAGEN, Venlo, Netherlands), and 85 ng aliquots were subjected to in vitro transcription using the T7 RiboMAX Express Large Scale RNA Production System (Promega, Madison, WI), followed by purification using the GeneChip IVT cRNA Cleanup Kit (QIAGEN) to obtain RNAs with sequences complementary to the scaffolds. An approximately 500-fold increase of the strand copy number was achieved by this in vitro transcription reaction. An aliquot of the product RNA was subjected to reverse transcription with the primer used in the PCR (but without the T7 promoter) and SuperScript III Reverse Transcriptase (Invitrogen), followed by treatment with RNase H (Invitrogen) to cleave RNA and DNA purification (QIAquick PCR Purification Kit, QIAGEN). Starting with 8 mg of RNA template, approximately 4.5 mg of purified single-stranded DNA was obtained, which then served as a scaffold.

Assembly

Mixtures of strands (5 nM each, unless specified) in TAM buffer (40 mM Tris, 20 mM acetic acid, 12.5 mM MgCl₂) were annealed by adopting the following thermal profile using a T-Gradient thermal cycler (Biometra, Goettingen, Germany): an initial incubation at 96°C for 5 min, followed by rapid cooling to 60°C, slow cooling to 55°C over the course of 24 h (approximately 30 min per 0.1°C), and a final rapid cooling to 20°C.

Agarose gel electrophoresis

Electrophoresis through 2% agarose gels was performed in TAM buffer at 50 V at room temperature with buffer circulation. The gel was then stained with the SYBR Gold Nucleic Acid Gel Stain (Molecular Probes, Eugene, OR) following the conditions recommended by the manufacturer. The images of stained gels were obtained using a LAS-1000 plus fluorescent image analyzer (FUJIFILM, Tokyo, Japan).

Quantification of bands by densitometry

The intensity of the bands in the gel image was measured using a Science Lab 2001 Image Gauge Ver. 4.0 (FUJIFILM). The overlapping bands were resolved, and the ratios of areas of component bands were

Figure 2. Strand preparation. The scaffold sequences are shown in red (DNA) or yellow (RNA). The T7 promoter tag sequence is shown in blue. The primers for PCR or reverse transcription are shown by thin arrows.

determined using GelBandFitter (18). A PCR ladder marker (New England Biolabs) was used as an internal reference to determine the double-stranded DNA mass.

Results and discussion

GENESUS selects the sequences of a set of strands that are optimized for the self-assembly of arbitrarily designed DNA nanostructures by a two-step generate-candidates-and-choose-the-best strategy. The program is based on the principle that virtually any DNA nanostructure can be viewed as an aggregate of double helices that are connected by single-stranded junctions. GENESUS consists of two modules: Generation of Unique Segment Pairs (GUSP) and Compilation and Evaluation of Strand Sets (CESS) (Figure 1).

GUSP generates sets of unique segment pairs (USPs) that are candidate sequences for helices in the designed structure. Here, unique means that the set consists of non-redundant n -tuples and is constructed by tiling unique n -mer sequence pairs (seed pairs) with r -mer overlaps, where $r = n - 1$. This is achieved by the extension of growing segment pairs by picking n -mer seed pairs that overlap the growing ends by r -mer from a randomly ordered n -mer seed pair table until the segment pairs of desired lengths are collected (Supplementary Material, GUSP algorithm; Supplementary Figure S1) (19). Each seed pair in the table can only be picked once during the collection of a set of segment pair sequences. Tables of seed pairs for r from 4 to 7 have been prepared using the Seed Maker module and are supplied with the program. The module also has the function of removing simple repeats to avoid slippage

Figure 1. Flowchart of GENESUS. This program consists of GUSP and CESS modules. See text for details.

in the annealing process and optionally of excluding user-specified sequence motifs (see Supplementary Material, Seed Pair tables).

The seed pair extension method guarantees that both redundancy and sequence symmetry of predetermined length (i.e., seed pair length) are simultaneously excluded from the set of USPs. Each time the seed pair extension reaches the predetermined helix length l , the candidate USP is evaluated by its melting temperature (T_m) and saved as the approved USP if its value is within a preset range. Setting a T_m range for each helix group allows the design of structures that can be assembled by one pod stepwise annealing, if necessary.

The number of USPs obtainable by GUSP is limited because the repertoire size of seed pairs is finite (less than $4^{n/2}$). Furthermore, premature termination of segment pair extension can be caused by the incidental exhaustion of the required seed pairs. We empirically determined the maximum numbers of USPs obtainable with GUSP by performing 1000 runs of the program at various l and n values and found that, on average, 60%–70% of the upper limit (USP counts if all seed pairs are used) were collected, with very narrow distributions (Supplementary Material, Maximum number of USPs obtainable by GUSP; Supplementary Table S1).

We also determined the length of unique sequence pairs that were collectable by repeated runs of modified GUSP (without length limit and T_m range setting) and found that extension terminated when 50%–60% of the available seed pairs were used, again with narrow distributions (Supplementary Material, Longest unique sequence pairs obtainable by modified GUSP; Supplementary Figure S2, Supplementary Table S2). These sequences, or any of their parts, should be suitable for use as scaffolds in the conventional origami strategy of DNA nanostructure self-assembly. The longest unique sequence pair at $r = 7$ after 1000 trials was 17,999 bp. When the collections were carried out by further modifying GUSP with a GC content restriction of 40%–60% at sliding windows of 10 bp, the longest collected sequence was 13,402 bp (Supplementary Material, Strand sequences). These figures were approximately twice the length of M13mp18, which has been widely used as a scaffold DNA nanostructure assembly using the origami strategy (11–14).

The salient feature of GENESUS is in the second part of the program, the CESS module. At the start of CESS, the sequences of the candidate strand set are composed by linking GUSP-generated USPs according to the design (Supplementary Material, CESS algorithm; Supplementary Figure S3; Supplementary Material, Design file;

Figure 3. Design of octahedron multimers. Six structures (a to f) were designed and named as shown. Scaffold strands are red and staple strands are black. In the multimers, connector strand sets are shown in various colors. All segment-junctions carried T5 as hinges (not shown in this figure).

Supplementary Figure S4). At this step, junction sequences between segments can be inserted. Then, the most stable aberrant pairing in the set (i.e., the pairing with the lowest free energy other than the intended pairs) is exhaustively searched by a modified dynamic programming algorithm, DPAL_AB (Supplementary Material, DPAL_AB), in which legitimate pairings are masked at the free energy calculation. DPAL_AB uses the nearest neighbor parameters of Santa-Lucia et al. (20), and the free energy is calculated allowing for single-base mismatches and gaps. The DPAL_AB test is repeated for all combinations of strands within the set to obtain the free energy of the most stable aberrant pairing for the set.

In the CESS evaluation steps, the possible formation of aberrant stem-loop or bulge formations is not considered, contrary to

other, similar programs (8). This is because such structures, if present, are likely to be superseded by competing proper pairings of full-length USPs, which always have longer base-pair stretches than seed pairs. Neglecting the possible formation of unfavorable stem-loops may not eliminate the possibility of kinetic trapping and delayed formation of intended structures, but the extent of this effect is unclear. The benefit of neglecting these possible unwanted structures is the resulting simplification of strand set evaluation and reduction of the time complexity for evaluation from $O(n^3)$ to $O(n^2)$ (10).

We did not include steps to avoid branch migrations at helix junctions (21). This is because the migration requires the presence of two-fold symmetrical nucleotides at the branch site and can be easily prevented by inserting non-symmetrical spacer nucleotides.

tides. Even without such spacers, branch migration inevitably causes a rotation of the involved helices that is unlikely to be allowed in the actual assembly of most nanostructures.

The GUSP-CES procedures are repeated for predetermined rounds, and the best strand set is selected as the one whose worst aberrant pairing is the least harmful (that is, the set whose most stable aberrant pairing has the highest free energy). The results are presented in the output files (Supplementary Material, Output files). The number of GUSP-CES cycle iterations is preset without any particular reason, and users are left to decide whether to accept the results by considering the lowest free energy value of aberrant pairing for the chosen strand set. The next version of GENESUS will be improved, for example, by including an algorithm that stops the cycle when the free energy value begins to plateau.

Another important feature of the GENESUS program is that by collecting the helix sequence set as sequences consisting of unique tuples, the sequence space to be tested at the evaluation step of CES is drastically reduced, in contrast to some published methods that start by collecting strand sets of random sequences (22). The sets of strand

sequences, compiled by linking helix sequences with junctions, are evaluated by the free energy of unintended pairings that are estimated using DPAL_AB. These selection steps also eliminate possible stable aberrant pairings at helix-junctions, whose sequences are not derived from tiles of unique seed sequences.

The rational selection of unique sequences inevitably limits the number of collectable helices. This limitation can be overcome by adopting a hierarchical assembly process strategy, where segment pairings are designed to proceed at two or more distinct temperature ranges. In this strategy, the whole structure is to be divided into domains, and the segmental pairings that connect the domains are designed to assemble at temperatures lower than the temperature of intra-domain assembly. Then, sequences can be shared among domains. GENESUS has the option of individually specifying the melting temperature of each USP to enable this strategy. The ability to individually set the melting temperatures of USPs also allows ordered assembly of the structure that otherwise cannot be assembled because of topological restraints caused by random initiation of pairing.

The validity of GENESUS was tested by constructing a regular octahedron (T5-OCT) with an edge length of 21 bases using a scaffold

and 4 staple strands whose sequences were chosen by the program. In this design, its helix junctions were uniformly T₅ stretches that are the minimum length required to bridge two contacting helices in any configuration (Supplementary Material, Design of a regular octahedron; Supplementary Figures S4 and S5).

The direct chemical synthesis of single-stranded DNA in large amounts (such as micrograms) is costly, especially when it is long. We established a low-cost and scalable enzymatic method to prepare single-stranded DNA of arbitrary sequence starting from cloned sequences. The method involves two steps of target sequence amplification (i.e., PCR and in vitro transcription), as depicted in Figure 2 and detailed in the Materials and methods section. A scaffold designed as described above was synthesized by this method and used in the assembly reaction. A small aliquot was used at each amplification step to obtain strands at quantities sufficient for in vitro self-assembly of DNA nanostructures. An equimolar mixture of scaffold and staples was annealed under various conditions, and the formation of T5-OCT was confirmed by gel electrophoresis of the assembled products (Supplementary Material, Assembly of T5-hinged octahedron). As shown in Supplementary Figure S6, assembly of presumptive complete octahedrons was observed only when all strands were present in the annealing mixture.

Next, we estimated the efficiency of the assembly of T5-OCT by densitometric analysis of the electrophoretic images of assembled products (Supplementary Material, Efficiency of assembly; Supplementary Figures S7 and S8). It turned out that approximately 75% of the scaffold was integrated into the octahedron in the equimolar assembly. The percentages reached a plateau of more than 90% when the concentrations of the staple strands were twice that of the scaffold or higher, demonstrating efficient assembly without requiring excess staples (Supplementary Table S3).

We further constructed multimers of octahedra as examples of more complex structures (Supplementary Material, Design and assembly of octahedral multimers). To do this, we designed three additional T5-hinged octahedra that have structures similar to T5-OCT (but consisting of strands with different sequences). To connect the octahedra, the sequences of full-length segments in the two staples at the contacting edges of the two octahedra were swapped, creating connector strand sets that were to be integrated into the assembly in a double crossover configuration (Figure 3). Trimers and tetramers with I- and L-shapes were designed in a similar manner.

Figure 4. Construction of octahedral multimers. Octahedral multimers were assembled by annealing the mixtures of scaffolds and staples (5 nM each). The annealed products were subjected to electrophoresis in a 2% agarose gel. Lane M: 100 bp double-stranded DNA ladder marker; lane 1: mixture of staple strands only; lane 2: scaffold strand only; lane 3: T5 Supplementary Figure S9-OCT; lane 4: (T5-OCT)₂; lane 5: (T5-OCT)₃-I; lane 6: (T5-OCT)₃-L; lane 7: (T5-OCT)₄-I; lane 8: (T5-OCT)₄-L; lane 9: (T5-OCT)₄-I without connector strand set C; lane 10: (T5-OCT)₄-L without connector strand set D; lane 11: (T5-OCT)₄-I without connector strand set B; and lane 12: (T5-OCT)₄-L without connector strand set B. Expected assembly products are drawn schematically for each lane below the photograph. The presence of faint bands in some multimer lanes suggests the incomplete participation of some connector strands in the assembly. See Supplementary Material, Design and assembly of octahedral multimers, and Figure 3.

A total of 2481 base oligonucleotides were designed using GENESUS (sequences in Supplementary Material, Strand sequences). The strands were prepared as described above, and mixtures of various combinations of the strands were annealed. Electrophoresis of the annealing products showed that the use of connector strand sets resulted in the formation of the expected multimers with size-dependent retardation (Figure 4). Omitting connector strands caused dissociation of the multimers to the expected substructures (lanes 9 to 15), confirming that the strands played the expected roles in multimerization. Thus, we conclude that the regular octahedron and its multimers were successfully assembled.

By expanding this strategy, it should be possible to configure an addressable space consisting of a body-centered cubic lattice in which the regular octahedron is the structural unit. Addressing space with the lattice described here has the advantage of leaving more space in the lattice than previously reported systems, such as tessellation or using blocks, which essentially fill the space with DNA strands.

In summary, we developed a straightforward program, GENESUS, to select a set of unique strand sequences in a two-step generate-candidates-and-choose-the-best strategy to achieve self-assembly of target structures. The selected set has sequences having the lowest chance of aberrant pairing among candidate sets. Future improvement of the program includes evaluation of the chosen sequence set by an energy minimization algorithm to assess the probability of target structure self-assembly. We also presented an enzymatic method to prepare single-stranded DNA with arbitrary sequences. The procedure includes two amplification steps followed by taking aliquots for the next steps. The quantity of the final product can be easily expanded to, for example, 1000 × the quantity described here (micrograms of single strand produced starting from nanograms of cloned plasmid DNA) by simply using larger aliquots. We designed strand sequences to build regular octahedrons and their multimers by GENESUS and confirmed their validity by actually assembling the structures at a high efficiency.

Author contributions

T.Ts. wrote the program, performed experiments, and drafted the paper. T.A. contributed to the early phase of program development. A.K. and T.O. discussed the results. T.Ta. discussed the results, and wrote the paper. K.H. conceived and supervised the project, and wrote the paper.

Acknowledgments

This work was supported by the New Energy and Industrial Technology Development Organization (NEDO).

Competing interests

The authors declare no competing interests.

References

1. Lu, C.H., B. Willner, and I. Willner. 2013. DNA nanotechnology: from sensing and DNA machines to drug-delivery systems. *ACS Nano* 7:8320-8332.
2. Douglas, S.M., I. Bachelet, and G.M. Church. 2012. A logic-gated nanorobot for targeted transport of molecular payloads. *Science* 335:831-834.
3. Qian, L., E. Winfree, and J. Bruck. 2011. Neural network computation with DNA strand displacement cascades. *Nature* 475:368-372.
4. Seeman, N.C. 1990. De novo design of sequences for nucleic acid structural engineering. *J. Biomol. Struct. Dyn.* 8:573-581.
5. Zhu, J., B. Wei, Y. Yuan, and Y. Mi. 2009. UNIQIMER 3D, a software system for structural DNA nanotechnology design, analysis and evaluation. *Nucleic Acids Res.* 37:2164-2175.
6. Feldkamp, U., H. Rauhe, and W. Banzhaf. 2003. Software tools for DNA sequence design. *Genet Program Evolvable Mach* 4:153-171.
7. Feldkamp, U. 2010. CANADA: Designing nucleic acid sequences for nanobiotechnology applications. *J. Comput. Chem.* 31:660-663.
8. Kick, A., M. Bonsch, and M. Mertig. 2012. EGNAS: an exhaustive DNA sequence design. *BMC Bioinformatics* 13:138-154.
9. Goodman, R.P. 2005. NANEV: a program employing evolutionary methods for the design of nucleic acid nanostructures. *Biotechniques* 38:548-550.
10. Dirks, R.M., J.S. Bois, J.M. Schaeffer, E. Winfree, and N.A. Pierce. 2007. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Review* 49:65-88.
11. Rothmund, P.W.K. 2006. Folding DNA to create nanoscale shapes and patterns. *Nature* 440:297-302.
12. Douglas, S.M., H. Dietz, L. Tim, B. Hogberg, F. Graf, and W.M. Shih. 2009. Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* 459:414-418.
13. Han, D., S. Pal, J. Nangreave, Z. Deng, Y. Liu, and H. Yan. 2011. DNA origami with complex curvatures in three dimensional space. *Science* 332:342-346.
14. Ke, Y., N.V. Voigt, K.V. Gothelf, and W.M. Shih. 2012. Multilayer DNA Origami Packed on Hexagonal and Hybrid Lattices. *J. Am. Chem. Soc.* 134:1770-1774.
15. Wei, B., M. Dai, and P. Yin. 2012. Complex shapes self-assembled from single-stranded DNA tiles. *Nature* 485:623-626.
16. Ke, Y., L. Ong, W.M. Shih, and P. Yin. 2012. Three-dimensional structures self-assembled from DNA bricks. *Science* 338:1177-1183.

17. Sambrook, J. and D.W. Russell. 2001. *Molecular Cloning. A laboratory manual* 3rd edn., Vol. 3. p. A8.20-A8.21. Cold Spring Harbor Laboratory Press, Woodbury, NY.
18. Mitov, M.I., M.L. Greaser, and K.S. Campbell. 2009. GelBandFitter – A computer program for analysis of closely spaced electrophoretic and immunoblotted bands. *Electrophoresis* 30:848-851.
19. Asakawa, T., K. Nishi, R. Mizuno, K. Yoneda, T. Okada, and K. Hayashi. 2006. Build-to-order nanostructures using DNA self-assembly. *Thin Solid Films* 509:85-93.
20. SantaLucia, J., Jr. and D. Hicks. 2004. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* 33:415-440.
21. Seeman, N.C. 1982. Nucleic acid junction and lattices. *J. Theor. Biol.* 99:237-247.
22. Tanaka, F., A. Kameda, M. Yamamoto, and A. Ohuchi. 2005. Design of nucleic acid sequences for DNA computing based on a thermodynamic approach. *Nucleic Acids Res.* 33:903-911.

Received 11 December 2013; accepted 04 February 2014.

Address correspondence to: Kenshi Hayashi, Division of Genome Analysis, Research Center for Genetic Information, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan. E-mail: hayashi.kenshi@gmail.com

To purchase reprints of this article, contact: biotechniques@fosterprinting.com