Structured sequences emerge from randomness when replicated by templated ligation

Patrick W. Kudella[1], Alexei V. Tkachenko[2], Sergei Maslov[3,4], Annalena Salditt[1], Dieter Braun*[1]

[1] Systems Biophysics and Center for NanoScience, Ludwigs-Maximilian-Universität München, 80799 Munich, Germany
[2] Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, New York 11973, USA
[3] Department of Bioengineering, University of Illinois at Urbana-Champaign, 1270 Digital Computer Laboratory, MC-278, Urbana, Illinois 61801, USA
[4] Carl R. Woese Institute for Genomic Biology, University of Illinois, Urbana-Champaign, Illinois 61801, USA

Supplementary Information

1. Table of Contents

1. Table of Contents 1
2. Polyacrylamide gel electrophoresis 2
3. Concentration Measurement in PAGE Gels by Image Analysis 2
3.1. Baseline of Gels 3
4. NUPACK simulation comparing 12mer and 10mer AT complexes 4
5. Random sequence pool parameter space 4
5.1. Random sequence pool ligation dynamics as function of temperature 4
5.2. Full random ATGC 12mer sequence pool 5
5.3. Comparison with Toyabe/ Braun 6
6. GC only random sequence pools of different starting lengths 7
7. Ligation time variation 8
8. LabVIEW program for sequence analysis 9
9. Randomness of initial pools 11
9.1. 12mer AT random sequence pool 11
9.2. GC random sequence pool 12
9.3. Error estimation in random sequence pool sequencing 13
10. Entropy Reduction 15
11. Pearson Correlation Coefficient 15
12. Domain Wall 17
13. Correlation of Ligation Sites, Inter-Ligation Sites, Strand Start and Strand End 18
14. Z-score landscape 18
15. First strands ligation 19
16. Ligation site shift 20
17. x8 experiment, designed and selected pools of eight sequences 21
17.1. “Replicator” design 21
17.2. x8 experiment after 50 temperature cycles 21
18. A:T-composition and kinetic simulations 23
18.1. Composition and hairpin formation 23
18.2. Kinetic model 24
19. Network-families resemblance 24
20. x8-pools de Bruijn graphs 26
21. Random Templated Ligation Simulation based on 12mer pool motif bias 27
22. List of DNA used 31
23. Supplementary Information Sources 32

2. Polyacrylamide gel electrophoresis

For analyzing the dynamics and product yield of the random sequence ligation we use polyacrylamide gel electrophoresis (PAGE) with SYBR gold post-staining. The gels are 15 % acrylamide and are run in a solution of 50 % urea and 1x TBE buffer at about 50 °C posing denaturing conditions.

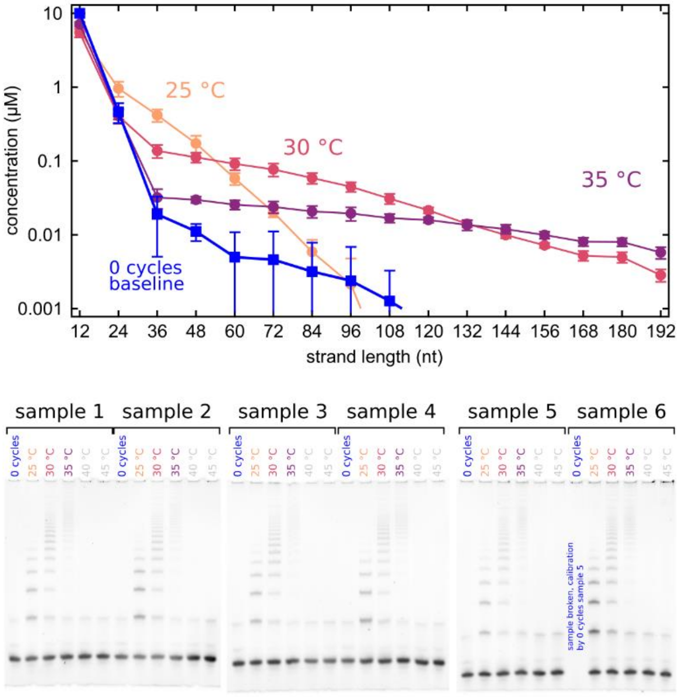
The gel is mixed from the *Roth* Rotiphorese DNA sequencing system. One 0.75 mm thick gel with a 15 tooth comb needs about 5 ml gel mixture which contains 3 ml gel concentrate, 1.5 ml gel diluent, 0.5 ml buffer concentrate, 25 µl APS and 2.5 µl TEMED. After 30 min of pre-run at 400 V, the gel pockets are loaded with a total of 4 µl of sample made from 0.89 µl of 10 µM sample and 3.11 µl of 2x loading dye (for about 10 ml add 9.5 ml formamide, 0.5 ml glycerol, 1 µl EDTA (0.5 M), and 100 µl Orange G dye (e.g. from *New England Biolabs*)). The sample is drawn into the gel in a first step of the run with 50 V for 5 min, then the gel electrophoresis is run for about 30 min at 300 V.

After the run, gels are submersed into 50 ml of 1x TBE buffer with 5 µl of 10.000x *SYBR Gold Nucleic Acid Gel Stain* from *Thermo Scientific* for 5 min. The stained gel is washed in 1x TBE buffer two times and imaged in a *bio-rad ChemiDoc MP* System.

Analysis of the gel images are done in self-written *LabVIEW* code, annotations are made in *GIMP* and *inkscape*.

3. Concentration Measurement in PAGE Gels by Image Analysis

We developed a LabView program for detailed analysis of PAGE images. The main problem with quantification in PAGE is the inhomogeneous fluorescence for ssDNA and dsDNA, as well as base-order. In our experiments all possible products are made from the same length monomers that are only have A and T as. The necessary prerequisite for the tool is a baseline: this lane is loaded with the monomers and buffer only. This baseline sample did not experience temperature cycling but was kept at constant 4 °C in the fridge. All other samples are compared to this sample.



SI-Fig. 1, ***reproducibility experiment for concentration estimation tool,***
top: *Calculating the concentration from each experiment enables the calculation of an average concentration estimation and a standard deviation error.*
bottom: *Three gels with a total of six experiments show very similar structure.*

In the first step, gel images are loaded in the tool and the outermost lanes are marked with cursors. The program automatically selects the lanes and space in between the lanes for background correction. For every lane the intensity over the gel position is calculated and corrected with the average intensity taken from the inter-lane areas left and right of the sample lane.

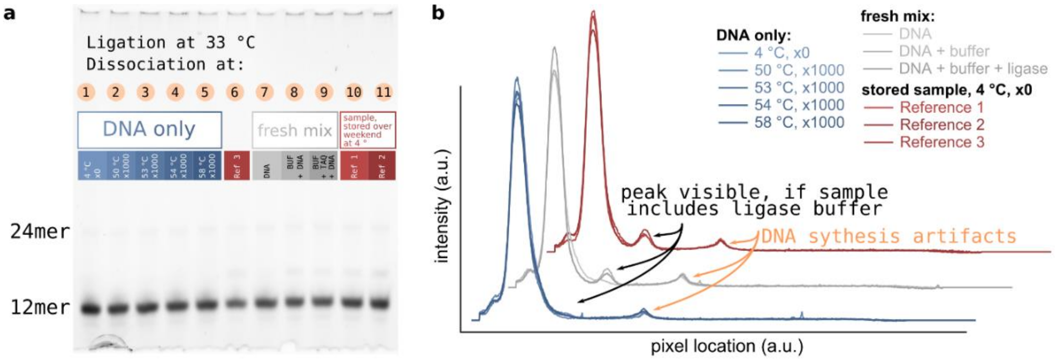
The intensity data is then normalized: the total intensity in every lane must be the same as the total concentration of the monomer pool in the baseline lane. The peak areas are marked with cursers and integrated with the simple trapezoidal method. For the final concentration estimation, the total intensity of the *x*th peak is divided by *x* – as a *x*-mers' intensity is *x* times greater than that of a monomer.

Performing the same experiment several times and calculating the concentration from the different gels each time enables us to calculate a standard deviation, as shown in SI-Fig. 1. The deviation of the six samples is small and the traces for the different ligation temperatures can be easily distinguished.

3.1. Baseline of Gels

All gels show two different distinct artefacts. First, there is a small peak directly after the 12mer peak. This peak is only visible if the sample contains the ligase buffer. We suspect the SYBR gold post-staining also dyes a component of the buffer. The gels in Fig. 1d of the main manuscript, in SI-Fig. 2a and the intensity analysis in SI-Fig. 2b all show this peak. In comparison a sample with just DNA and MilliQ water does not show this peak.

The second peak unfortunately runs at a length where one would expect the 24mer products of the ligation reaction. This peak is always visible at the same location with the same intensity. This is probably an artefact from DNA synthesis. As the DNA is synthesized starting from the 3'end, it can occur, that a second DNA backbone is synthesized to the first base. Those structures would run comparable to a 24mer, but probably don't take part in the ligation reaction. It can, however, not be analyzed by NGS, as the library preparation chemistry is sensitive to DNA backbone errors.

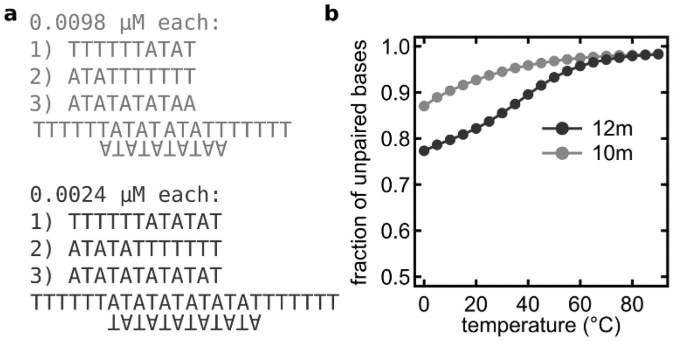


*SI-Fig. 2, **PAGE of AT-only random sequence pool with and without ligase for baseline comparison:***
***a** PAGE of AT-only random 12mer DNA with different buffer conditions. 1) DNA only in MilliQ water at 4 °C in the fridge for ~60 h. 2) DNA only in MilliQ water, 1000 temperature cycles between 33 °C and 50 °C. 3) DNA only in MilliQ water, 1000 temperature cycles between 33 °C and 53 °C. 4) DNA only in MilliQ water, 1000 temperature cycles between 33 °C and 54 °C. 5) DNA only in MilliQ water, 1000 temperature cycles between 33 °C and 58 °C. 7) fresh solution of DNA in MilliQ water, immediately mixed with loading dye and quenched. 8) fresh solution of DNA in MilliQ water and ligase buffer, immediately mixed with loading dye and quenched. 9) fresh solution of DNA in MilliQ water, ligase buffer and ligase, immediately mixed with loading dye and quenched. 10), 11) and 6) AT-only random DNA in ligase buffer and ligase, stored at 4 °C in the fridge for ~60 h.*
***b** Baseline-corrected intensity plots of the gels: all lanes show the same artifact at a position where 24mer DNA would run. This might be an artifacts of the synthesis at the 3' end of DNA. The small peak close to the 12mer-peak is only visible if the ligase-buffer is in the sample. All of those artifact-peaks are very similar across different experimental conditions.*

4. NUPACK simulation comparing 12mer and 10mer AT complexes

As mentioned in the main manuscript double stranded complexes in a conformation that can be ligated by the TAQ DNA ligase are needed in order for the experiment to work. Taking later results as the input for the most probable sequences for a complex of three strands we use NUPACK to calculate an approximation to the melting curve. SI-Fig. 3a shows the three sequences with lengths 10 nt and 12 nt. SI-Fig. 3b shows the fraction of unpaired bases as a function of temperature. The concentration of strands is set to 0.0098 µM and 0.0024 µM which is equivalent to their concentration as if they had the same abundance as in a 10mer (respectively) 12mer AT-random sequence pool.

In the temperature range where ligation reactions are possible because the ligase is active, there is a significantly lower number of paired bases marking a lower number of double stranded DNA for the 10mers. With an addition of only 2 bases per stand, the total sequence space only grows by a factor of 4. And NUPACK suggest a higher number of double strands at temperatures of 25 °C to 33 °C – conditions similar to the experimental ligation conditions in the main manuscript.



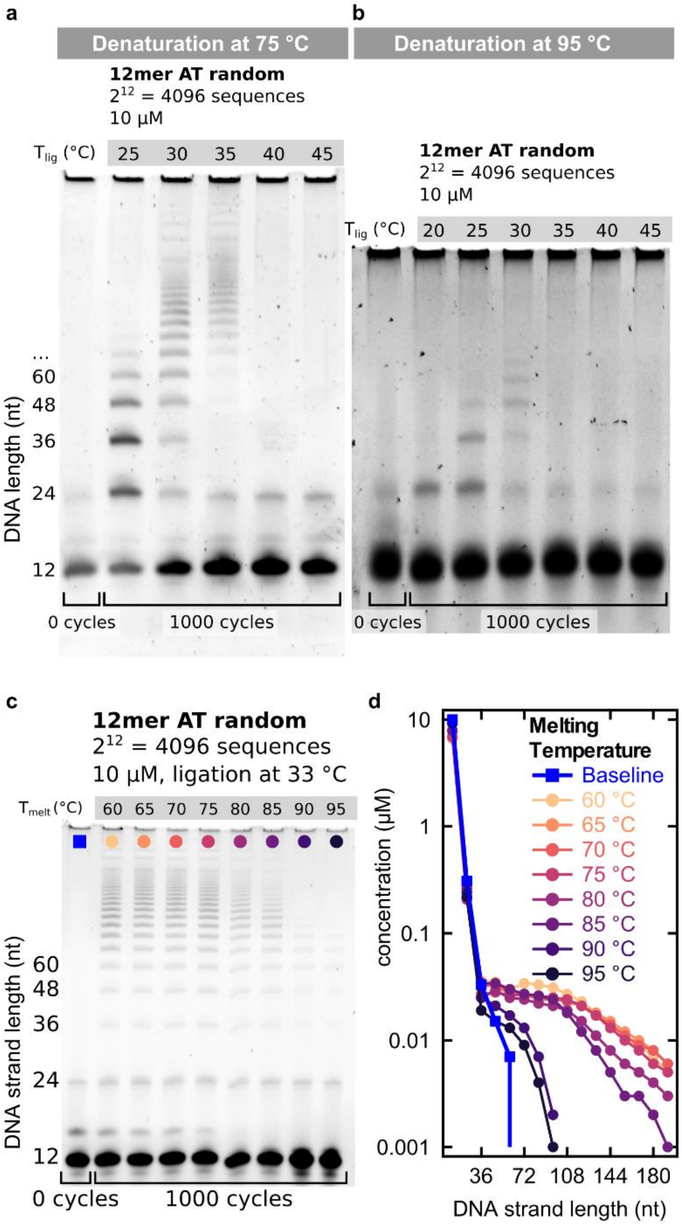
*SI-Fig. 3, **NUPACK simulation for 12mer and 10mer AT sequences duplex formation:***
a *10mer and 12mer long strands that form triplet structures suitable for TAQ DNA ligation.*
b *The NUPACK tool gives a fraction of unpaired bases over temperature for both systems. This is comparable to a melting curve: the lowest possible value on the y-axis is about 0.33 because in ideal conformation there are 24 of the 36 bases paired, leaving 1/3 unpaired. At the ligation temperature of 33 °C there is a significantly larger amount of paired 12mers than 10mers.*

5. Random sequence pool parameter space

The pool made from 12mer AT-random sequence DNA strands produces oligomers of different lengths by the templated ligation reaction under temperature cycling. The dynamics of the system, best visualized by PAGE, depends on several parameters discussed below.

5.1. Random sequence pool ligation dynamics as function of temperature

The dynamics of formation of longer strands from the monomers in the pool is highly dependent on the temperature of ligation and dissociation. As a rule of thumb, higher temperatures of ligation produce more long strands with non-exponentially decreasing character. Higher dissociation temperature reduces the amount of long sequences dramatically. In SI-Fig. 4 four gels show the described behavior. SI-Fig. 4a shows 12mer AT-only random sequence pool products for varying ligation temperatures and a dissociation of 75 °C. The sample at 25 °C shows a strong exponential decay of longer sequences. For 30 °C the exponential decay is weaker and the amount of long sequences is higher. For 35 °C only long sequences emerge and short 24mer and 36mer sequences are absent in the gel. For high temperatures of 40 °C and above no product emerges. In SI-Fig. 4b the same sample is run with a dissociation temperature of 95 °C. The overall behavior is similar, but the amount of long strands severely reduced.

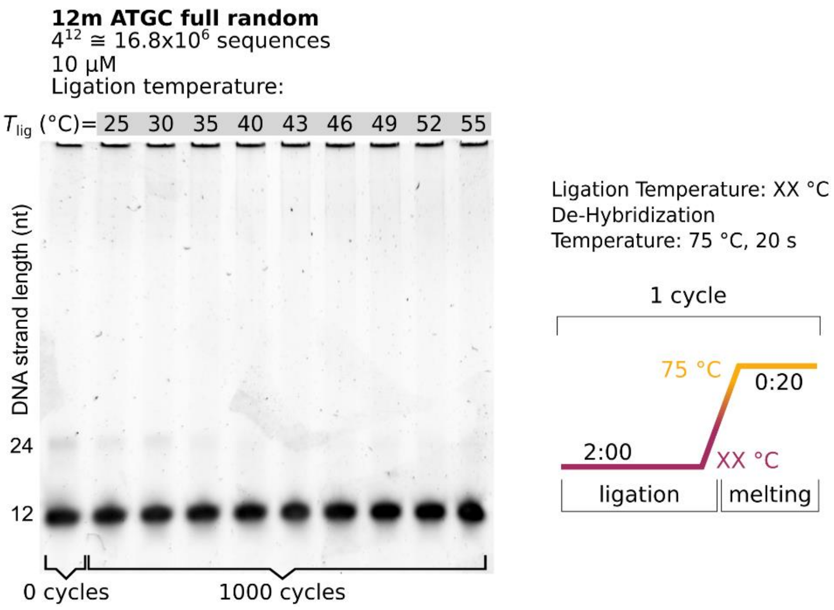


SI-Fig. 4, ***PAGE of 12mer AT only random sequence pool, cycled for 1000 times, different temperature cycle conditions:***
a *The dissociation temperature is 75 °C the left-most line is the reference sample with zero temperature cycles. The temperature given above each lane is the temperature of the ligation reaction.*
b *The dissociation temperature is 95 °C the left-most line is the reference sample with zero temperature cycles. The temperature given above each lane is the temperature of the ligation reaction.*
c & d *With a fixed ligation temperature of 33 °C the dissociation temperature is varied in between lanes from 60 °C to 95 °C. The shoulder of short and medium sized oligomers is retained up until 85 °C but the concentration of long oligomers is lower, the higher the dissociation temperature.*

5.2. Full random ATGC 12mer sequence pool

Despite the comparably easy analysis of binary sequence data in AT-only or GC-only experiments, the emergence of potential motives from full random, four bases DNA might be interesting. In preliminary results under similar experimental conditions as described above, we could not detect oligomer products emerging from 12mer ATGC-random pools (see SI-Fig. 5). Comparing the sequence space reveals the vast difference: 4^{12}≈16.8*10^6, for four bases and 2^{12}=4096 for two bases. Assuming a linear correlation (lower boundary) of temperature cycles to sequence space for the emergence of a well observed oligomer product distribution would mean an increase of experimental time from about 60 hours for 1000 temperature cycles to 204.800 hours (>23 years). With changes to the experiment it might be possible to decrease the experimental time further, but the system will have either reduced complexity, or less freedom to form complexes.

Anyhow, the underlying results found in this work might very well still be valid: suppression of self-folding, emergence of specific sequences at ligation sites due to better binding geometry or higher binding energy, suppression of poly-G sequences due to their stickiness might all be found in a potential full random templated ligation experiment.



*SI-Fig. 5, **T_{lig} parameter sweep for ATGC full random sample:***
The full random sample including all four bases A, T, G, and C does not yield any oligomer product after 1000 temperature cycles. The ligation temperatures T_{lig} were selected in a range where oligomer products emerged for both AT-random and GC-random.

5.3. Comparison with Toyabe/ Braun

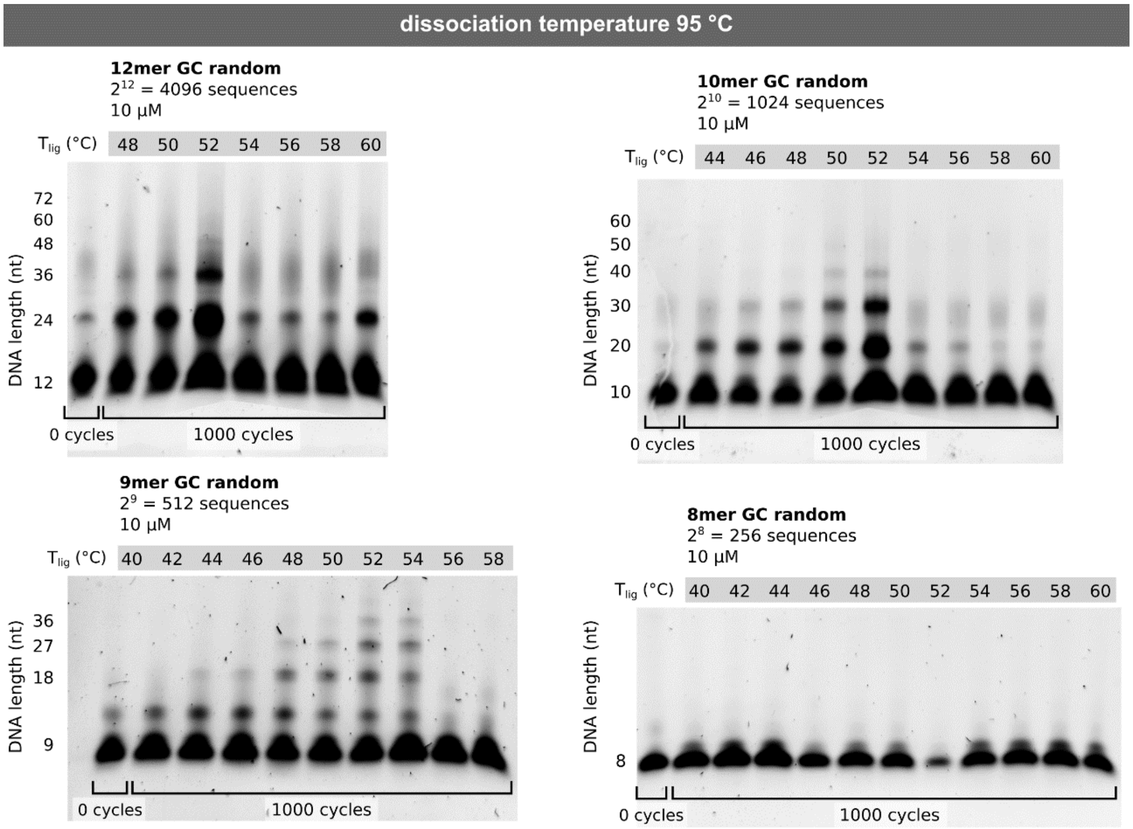
The 2019 study of Toyabe and Braun (Ref. (21) of the main manuscript) explored possible cooperation of selected DNA motives of length 20 nt. The strands were designed to build networks, that could withstand a simulated decay (serial dilution) when fed with the motive-monomers. The ligation was also done with the TAQ DNA ligase from *new england biolabs*. For the dissociation temperature they chose 95 °C, the highest possible temperature where the ligase can survive for short time and DNA does not denaturate, while double strands are dissociated. They kept the dissociation temperature time as short as the PCR temperature cycler allows (1 s). The ligation temperature was selected as 67 °C, the melting temperature of the motive-monomers. They used three different submotives (a, b, c) and their respective reverse complements ($\bar{a}$, $\bar{b}$, $\bar{c}$), and different oligomers to start the reaction, like ab + $\overline{ba}$ noted as AB together. In a typical reaction six monomers plus 2 for each oligomer (two different 40 nt dimers and one 60 nt trimer) were spanning a sequence space of twelve different sequences.

Although the total sequence space for 12mer "monomers" in the experiment in the main manuscript is easy to calculate (2^{12}=4096) and as shown below also a good estimate for the actually present DNA strands in the random sequence mix, the comparison is not straight forward. In both experiments, the hybridization is the actually interesting mechanism. In Toyabe and Brauns work, the strands are designed to only hybridize to their reverse complement and without any overhangs. In contrast, strands from the AT-only random sequence pool can hybridize in multiple different configurations and a majority of those will likely inhibit ligation. The sequence space of complex formation is therefore significantly larger than the increase in sequence species might suggest. This slows the overall reaction rate substantially and explains, why the random sequence templated ligation reaction needs longer cycle times.

The melting temperature of a dsDNA strand depends on several parameters like the amount of paired bases, the GC-content and stacking interactions. In Toyabe and Brauns work all "monomers" were designed to have the same melting temperature. In the random sequence pool, dsDNA complexes presumably have very different dissociation temperatures, mainly due to the different amount of paired bases, dangeling ends and the actual sequence of bases.

6. GC only random sequence pools of different starting lengths

The main reason for using AT-only samples is the lower melting point for dsDNA. Using a GC-only random sequence pool, complexes have a higher melting temperature and the experiments then also need a higher dissociation temperature in which the ligase degrades faster. Nevertheless, the basic templated ligation elongation experiment is possible for GC-only pools. With the higher melting temperature of the 12mers, we also tried lower lengths for the random sequence pool. As explained in the methods section of the main manuscript, the dissociation temperature is a limiting factor for the shortest possible monomer length of AT-only strands. SI-Fig. 6 shows that the 8mer GC-only random sequence pool does not produce oligomers in 1000 temperature cycles. This is not surprising, as the manufacturer states 4mer overhangs are not ligated by the TAQ DNA ligase. From all other pools of lengths 9mer (remarkable, as a complex made from three 9mers must have a 5mer and a 4mer overhang – which is apparently ligated), 10mer and 12mer oligomers emerge, although way less long oligomers and less oligomers over all. The bands, especially for the monomer-length, tend to smudge into/ close to the next bands. There is also a more pronounced smear of each product band.

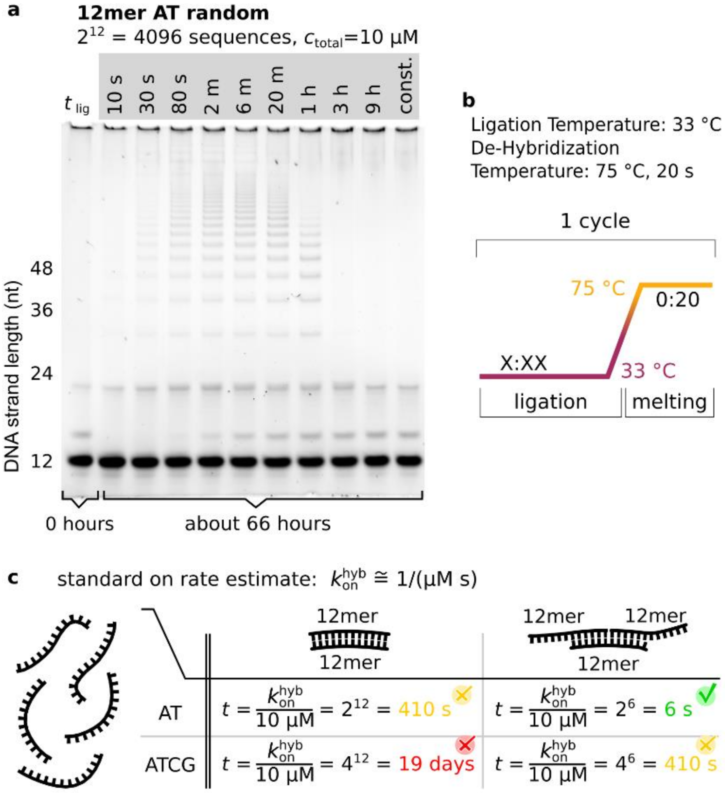


SI-Fig. 6, ***PAGE of several lengths of GC only random sequence pools, cycled for 1000 temperature cycles with different conditions:***
top left: *12mer sample with sequence space 4096. Although using a denaturizing PAGE conditions, bands tend to smear out for longer strands.*
top right: *10mer sample with sequence space 1024.*
top right: *9mer sample with sequence space 512. For this length, the length distribution for reaction products looks the most like the AT only random sequence 12mer shown in the main manuscript.*
top right: *8mer sample with sequence space 256. Here, the strands are too short and the ligase doesn't ligate the three-part complexes anymore.*

The best working conditions for the GC-random pools seems to be around 52-54 °C, which is significantly higher than for the AT-only random sequence pools (33 °C).

7. Ligation time variation

The sequence space of a pool of DNA dictates the ligation time necessary for the emergence of products. For the 12mer AT-only random sequence pool temperature cycling is necessary, as seen in SI-Fig. 7a. For this sample, the increase of the ligation time in each respective temperature cycle yields more product in total until about t_{lig} = 6 min. For very short (such as 10 s) and very long ligation times per cycle, there is no product at all. This due to two different reasons: for templated ligation, there first need to be double stranded complexes. The effective on-rate for hybridization of two oligomer strands can be assumed to be in the range of about 1/(µM s) (reasonable value in the range of literature values (1–4)). For the experimental system it is now possible to estimate the time strands need to form a complex, see SI-Fig. 7c. Only short enough strands find their reverse complement sequence due to the vast sequence space. For long t_{lig} there are simply not enough temperature cycles to measure the growth of emerging oligomer product strands.



SI-Fig. 7, ***PAGE of AT-only random sequence pool (orange) and NN (blue) sample:***
a *For longer ligation times the gel shows more product for the AT-only random sequence pool sample. The last column marks the sample without temperature cycling. For ligation time steps of 10 s the gel doesn't show any product at 24mer length or longer. For longer times, oligomers start to emerge. For even longer cycle times, the total amount of temperature cycles decreases and the product strand concentration decreases.*
b *The sketch shows the temperature-cycle scheme used here.*
c *Using a standard on-rate for hybridization (see Ref. (1–4)) of about 1 per µM and second it is possible to estimate the time it takes for strands to form complexes. The time scale is dominated by the sequence space of the possible complex formations. For six on six strands with bases AT only the complex takes approximately 6 s to form. For four bases, this already takes about 410 s (almost seven minutes). The same holds true for the formation of fully hybridized 12mers, which explains, why the first ligation reactions that are assumed to consist of three 12mer complexes with dangling ends are possible. For four bases and 12mer on 12mer a single cycle would need to be about 19 days long.*

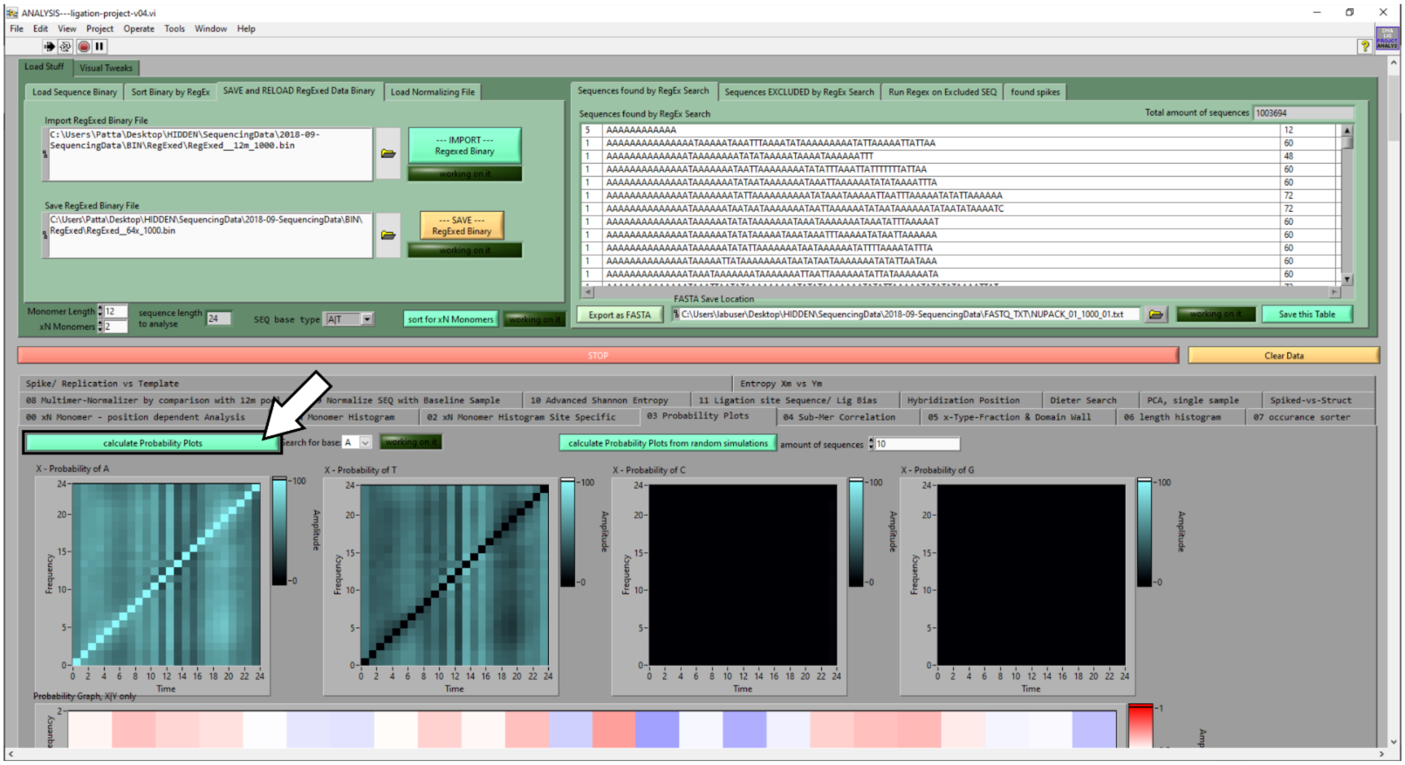
We chose our temperature cycling conditions optimized for the emergence of oligomer products with 20 s dissociation temperature and 120 s ligation temperature. With the heating and cooling in between those steps the entire experiment lasts about 60 h for the 1000 temperature cycles.

Additionally, as seen in section 5.1, our experimental system is strongly temperature dependent. We expect this to have a stronger influence on the overall length and product concentration distribution than the activity of the ligase.

8. LabVIEW program for sequence analysis

Sequence analysis is predominantly performed with self-written LabVIEW programs. LabVIEW is a graphical programming language suited for fast programming of high level data structures. A LabVIEW program is called a “Virtual Instrument” (VI) and consists of the front panel including graphs, controls and tables and the block diagram including the logic of the program. Overall, LabVIEW is very similar in performance compared to other high level scripting languages but has several advantages like the by default included user interface (UI) or the possibility to store large datasets in the VI.

When an *illumina* sequencing run is completed the data is stored on a server of the Gen Center Munich (https://www.genzentrum.uni-muenchen.de/index.html) running an instance of galaxy (Ref. (52) of the main manuscript). galaxy-demultiplexing scripts are used for the demultiplexing step of the data, as described in the methods section. The demultiplexed FASTA file is then downloaded and further analysis is done with the custom-VIs.


ANALYSIS---ligation-project-v04.vi
File Edit View Project Operate Tools Window Help
Load Stuff
Visual Tweaks
Load Sequence Binary
Sort Binary by RegEx
SAVE and RELOAD RegExed Data Binary
Load Normalizing File
--- IMPORT --- Regexed Binary
working on it
--- SAVE --- RegExed Binary
working on it
sort for xN Monomers
working on it
Sequences found by RegEx Search
Sequences EXCLUDED by RegEx Search
Run Regex on Excluded SEQ
found spikes
Export as FASTA
working on it
Save this Table
STOP
Clear Data
Spike/ Replication vs Template
Entropy Xm vs Ym
08 Multimer-Normalizer by comparison with 12m pool
09 Normalize SEQ with Baseline Sample
10 Advanced Shannon Entropy
11 Ligation site Sequence/ Lig Bias
Hybridization Position
Dieter Search
PCA, single sample
Spiked-vs-Struct
00 xN Monomer - position dependent Analysis
01 Monomer Histogram
02 xN Monomer Histogram Site Specific
03 Probability Plots
04 Sub-Mer Correlation
05 x-Type-Fraction & Domain Wall
06 length histogram
07 occurance sorter
calculate Probability Plots
working on it
calculate Probability Plots from random simulations
X - Probability of A
X - Probability of T
X - Probability of C
X - Probability of G
Frequency
Time
Amplitude
100
0
24
20
15
10
5
0
0 2 4 6 8 10 12 14 16 18 20 22 24
Probability Graph, X|Y only
2

Import RegExed Binary File: C:\Users\Patta\Desktop\HIDDEN\SequencingData\2018-09-SequencingData\BIN\RegExed\RegExed__12m_1000.bin

Save RegExed Binary File: C:\Users\Patta\Desktop\HIDDEN\SequencingData\2018-09-SequencingData\BIN\RegExed\RegExed__64x_1000.bin

Monomer Length: 12
xN Monomers: 2
sequence length to analyse: 24
SEQ base type: A|T

Sequences found by RegEx Search

Total amount of sequences: 1003694

5	AAAAAAAAAAAA	12
1	AAAAAAAAAAAAAAATAAAAATAAATTTAAAATATAAAAAAAAAATATTAAAAATTATTAA	60
1	AAAAAAAAAAAAAATAAAAAAAATATATAAAAATAAAATAAAAAATTT	48
1	AAAAAAAAAAAAAATAAAAAAATAATTAAAAAAAATATATTTAAATTATTTTTTTATTAA	60
1	AAAAAAAAAAAAAATAAAAAAATATAATAAAAAAATAAATTAAAAAATATATAAAATTTA	60
1	AAAAAAAAAAAAAATAAAAAAATATTAAAAAAAAAATATAAATAAAAATTAATTTAAAAATATATTAAAAAA	72
1	AAAAAAAAAAAAAATAAAAAATAATAATAAAAAAATAATTAAAAAATATAATAAAAAATATAATATAAAATC	72
1	AAAAAAAAAAAAAATAAAAAATATATAAAAAAATAAATAAAAAAATAAATATTTAAAAAT	60
1	AAAAAAAAAAAAAATAAAAAATATATAAAAATAAATAAATTTAAAAATATAATTAAAAAA	60
1	AAAAAAAAAAAAAATAAAAAATATATTAAAAAAATAATAAAAAATATTTTAAAATATTTA	60
1	AAAAAAAAAAAAAATAAAAATTATAAAAAAAATAATATAATAAAAAAATATATTAATAAA	60
1	AAAAAAAAAAAAAATAAATAAAAAAATAAAAAAATTAATTAAAAAATATTATAAAAAATA	60

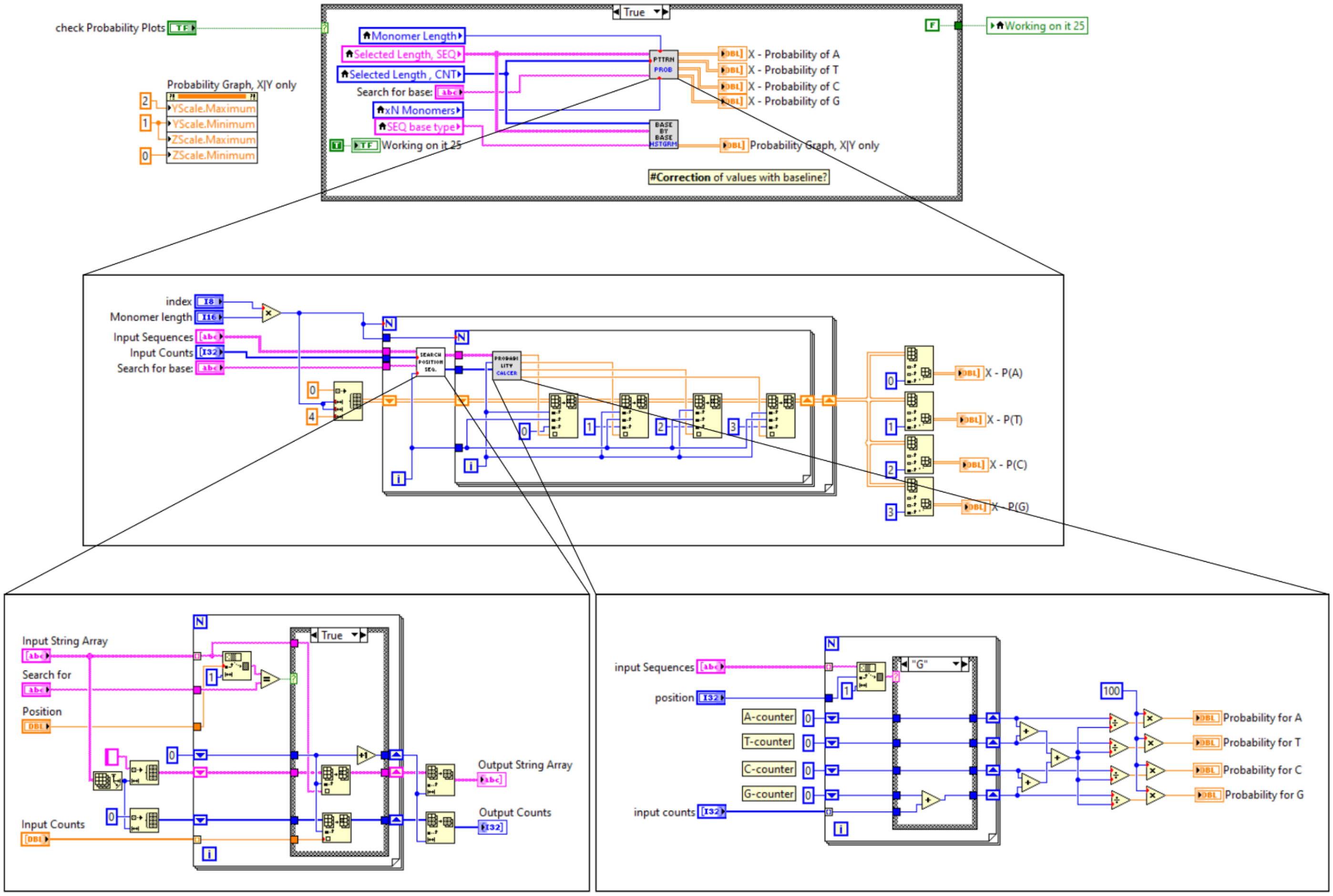
FASTA Save Location: C:\Users\labuser\Desktop\HIDDEN\SequencingData\2018-09-SequencingData\FASTQ_TXT\NUPACK_01_1000_01.txt

Search for base: A

amount of sequences: 10

*SI-Fig. 8, **screenshot of the main LabVIEW sequence analysis program:***
In comparison to most other scripting languages, the front panel of a VI is designed to be a responsive UI. This makes analysis with different parameters and exporting to plotting tools easy.

SI-Fig. 8 and SI-Fig. 9 show screenshots from parts of the VI front panel and block diagram (and the so called Sub-VIs, which behave like function-calls). The function shown here is run when the button on the front panel is pressed. The function calculates an image, where the probability of all other bases is calculated given a certain selected base (here A). The probability is plotted for all possible positions of the fixed base resulting in a 2D plot of the size oligomer-length squared.



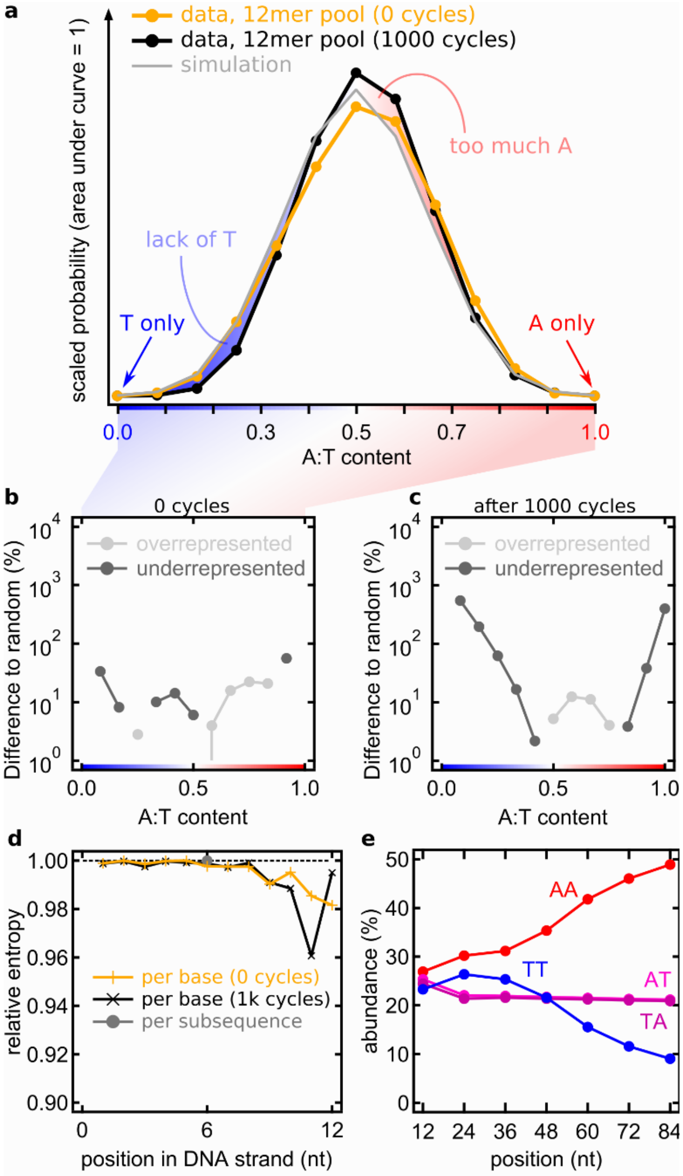
*SI-Fig. 9, **screenshot of the block diagram and Sub-Vis for one example function:***
The VI-architecture shown here is only one of multiple ways/styles of program in LabVIEW. The top level function is executed when the button in the Front Panel marked in SI-Fig. 8 is pressed. The function then calls several sub-Vis that work like calling a function. In the end the data is printed to graph indicators.

9. Randomness of initial pools

An essential part of the selection experiment is the initial random pool. The randomness of the pool might influence the downstream reaction significantly. Therefore, the randomness of the pool must be considered.

9.1. 12mer AT random sequence pool

For a completely random sequence distribution we would expect a binomial-shaped distribution for the A:T content, centered around the 50:50 mark. SI-Fig. 10a shows the 12m AT-only random sequence pool has too many strands with about 60-70 % A and too little strands with 60-80 % T.



*SI-Fig. 10, **12mer random sequence pool A:T fraction vs simulated random:***

a *Analyzing the A-to-T fraction of the 12mer strands of the remaining pool and comparing them to the corresponding curve of a simulated completely random pool reveals a bias towards A-type sequences. Specifically, ratios of 6:6, 7:5 and 8:4 A:T are overrepresented while 4:8, 3:9 and 2:10 A:T are underrepresented.*

b

c *The difference of the underrepresented and overrepresented of AT fraction in a) in percent. Poly-A and especially poly-T strands are underrepresented.*

d *The entropy of the ensemble of all strands is not reduced (black circle) due to the large amount of random sequence strands with about 50:50 % A:T-ratio but for the second to last base there is a drop in entropy. This is comparable to Figure 3a in the main manuscript.*

e *Plotting the abundance of all four possible 2 nt long submotives AA, AT, TA and TT as a function of product length shows, that for the initial pool AA, AT and TA submotives are about equal in abundance, while TT is underrepresented. For longer strands the TT submotives become less abundant. The total amount of A-type sequences is growing, as the increase of the AA submotive to almost 50 % of all 2 nt submotives indicates.*

The lack of poly-A and poly-T sequences is obvious when comparing the distribution to a perfectly random one (simulated, light grey). The simulated curves could be obtained by a simple binomial function, but we opted to calculating a great amount of AT-only random sequence 12mers with the analysis tool and analyzing their distribution with the same A:T-fraction algorithm as the sequence data.

The more bases of one kind are found in one strand, the less frequently they are sequenced. Still, in the experiment, there are enough total strands of A-type or T-type for the experiment to work, but the lack of T-type sequences in the monomers might help explain the bias in A-type to T-type in oligomer products.

In the entropy reduction (SI-Fig. 10d), there is only a minimal reduction compared to a perfectly random distribution, if calculated for the entire strand. The analysis for each position shows a lower entropy for the second to last base position. Also, in Figure 1b in the main manuscript, there is a distinct band in the base probability graph for the 12mers for base A.

Overall, we sequenced 4067 of the 4096 possible submotives (99.29 % coverage). The bias towards more A in the 12mer random sequence pool is probably no artefact, as long oligomers of A-type are significantly more pronounced than T-type sequences. As mentioned in the manuscript, an initial imbalance in A-T content is enhanced in every further elongation of a oligomer strand due to the templated ligation reaction.

Analyzing the abundance of all possible 12mer submotives is a difficult task, and it is even more difficult to visualize the results. In the main manuscript we show that reducing the length of analyzed submotives to a length of 6 nt reduces the amount of motives to 2^6=64 while retaining almost all local sequence-motive properties. This can also be done for very short submotives of length 2 nt, to analyze an initial or gradual bias in the pool or resulting oligomers. Therefore, we compare all possible motives AA, AT, TA and TT in SI-Fig. 10e: For the initial 12mers there is a lack of TT motives, while the other three motives AA, AT and TA have about the same abundance. There might have been a small difference in the base attachment probability of a DNA strand during DNA synthesis and is likely the reason for the base-composition bias in the initial 12mers (Figure 2c). For 24mers, the probability of AT and TA decreases while the probability of AA and TT increase. For longer oligomers the TT- motives get less abundant, while AA gets more abundant. AT and TA motives, mainly responsible for the ligation site A-T alternating sequence patterns stay at about 21 % abundance.

9.2. GC random sequence pool

DNA made from bases A and T only has a lower melting temperature compared to ATGC or GC-only DNA because of the weaker Watson-Crick basepairing of A and T, as described in section 7.

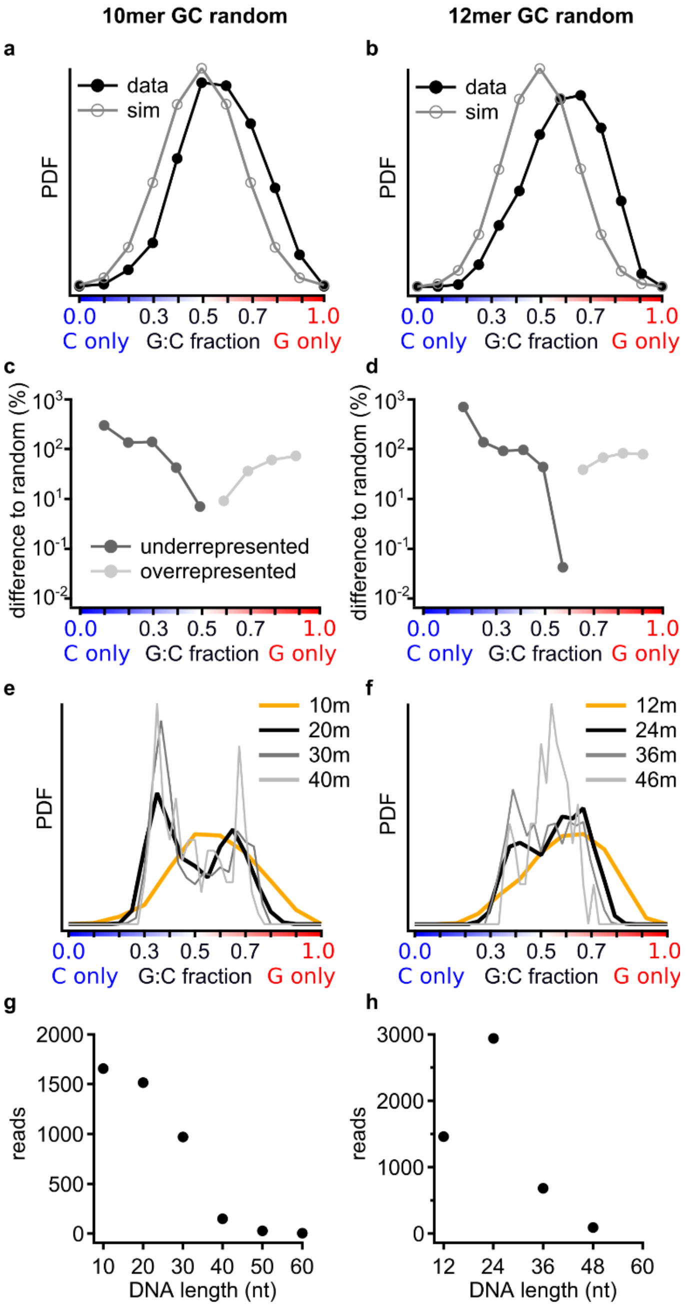
Nevertheless, the experiment works for GC-only samples as well. The sequences space is 2^{12}=4096 different sequences (and 2^{10}=1024 for 10mer monomers).

GC-only samples were subjected to similar experimental conditions, library preparation and illumina sequencing as the AT-only samples. GC-only experiments and subsequent sequencing typically yield significantly less reads and especially less reads for long oligomers, as seen in SI-Fig. 11g and h.

As shown in SI-Fig. 11, the GC-fraction of 10mer and 12mer samples are comparable to AT-only samples: There is a shift towards too much G, but it is significantly more pronounced than the shift towards A in AT-only. The difference to a completely random distribution is in both cases larger. Additionally, the symmetry of underrepresented poly-A and poly-T sequences (as in SI-Fig. 10b) is not seen for GC. C-type strands are underrepresented, while G-type strands are overrepresented in 10mer and 12mer GC-only “monomers”.

The PDF of GC-fractions of different oligomer-lengths shows a bimodal distribution for the 10mer sample. For the 12mer sample, only the 24mer seems to show a bimodal distribution. For longer oligomers the analysis is inconclusive as the amount of analyzed sequences is substantially smaller than for AT-only samples due to a very low readout success-rate in GC-only illumina sequencing.

Interestingly, for the 10mer samples, the bimodal distribution seems to favor C-type sequences, although the initial 12mers are predominantly G-type sequences. DNA with a lot of bases G are often referred to as “sticky”. Guanine domains tend to form duplexes with other strands or fold on themselves(5). With this additional binding mechanism absent in AT-only DNA the G:C-fraction might favor longer C-type sequences, because G-type sequences stick together or to themselves and are thus not taking part in templated ligation reactions.



*SI-Fig. 11, **GC-only random 10mer and 12mer samples:***
***a, b** For 10mer and 12mer monomers there is a shift towards too much G.*
***c, d** G-type sequences are strongly over and C-type sequences strongly underrepresented in both, 10mer and 12mer GC random sequence pools.*
***e** 10mer GC-only samples start again from an about Gaussian shaped GC-fraction distribution with the before mentioned shift towards G. 20mers and 30mers show the distinct emergence of a bimodal distribution – a G-type and C-type.*
***f** 12mer GC-only samples only show a bimodal distribution for 24mers. For longer oligomers the amount of analyzed sequences is low and there seems to be no clear shape.*
***g, h** Length histograms show only a small amount of reads, although the experimental conditions and library preparation are similar to AT-only samples.*

9.3. Error estimation in random sequence pool sequencing

In contrast to sequencing data of well-defined strand elongation experiments, where it is possible to already include the primer, barcode and possibly the binding section to adhere the strand in the flow cell of the illumina sequencer, this is not possible here. The short 12mer DNA strands need be able to be ligated on both 5' and 3' end without a long tail of additional sequence for later characterization. Therefore, we utilize, as described in the methods section of the main manuscript, a library preparation kit to enzymatically attach a primer and barcodes to our sample strands of different lengths. This attachment might be biased towards certain sequence motives. As there can't be a "real" reference sample because of the randomness of the pool, we can only estimate and search for systematic biases in our results. There are several results that point towards different errors:

- The comparison of 2 nt motives in strands of all lengths (SI-Fig. 10d) shows, that motives AA, AT, TA are comparable in abundance, while TT is underrepresented. This might hint towards a bias in strand synthesis. In longer strands the abundance of AT and TA is about constant while the oligomer products are segregated into A-type and T-type strands, with T-types being less abundant due to the initial asymmetry.
- Although there might be bias in attaching the random sequence CT-tail to 12mers and their oligomer products during library preparation, this enzymatic elongation (possibly done by a terminal transferase, unclear because the manufacturer doesn't provide further information) is likely only biased by the few bases close to the 3'-end of the strand. SI-Fig. 10c suggests a very small bias in the entropy for bases in positions 9, 10, 11 and 12 that might stem from either DNA synthesis or sequencing. Despite that, most analysis of sequencing data performed here is done on the entire strand or the only the center subsequences, that are predominantly far enough away from the 3'-end to be bias from the selection. In graphs like Figure 2, 3 and 4 of the main manuscript, results show great reproducibility over different lengths, and the results fit the theoretical models well (that don't have a bias for sequencing or synthesis artifacts).

Both points support the assumption, that the library preparation kit is well suited to prepare random sequence stands for NGS without introducing strong biases. The asymmetry of the pool and underrepresented sequence motives might account for greater uncertainties in comparison to an error due to library preparation or sequencing.

10. Entropy Reduction

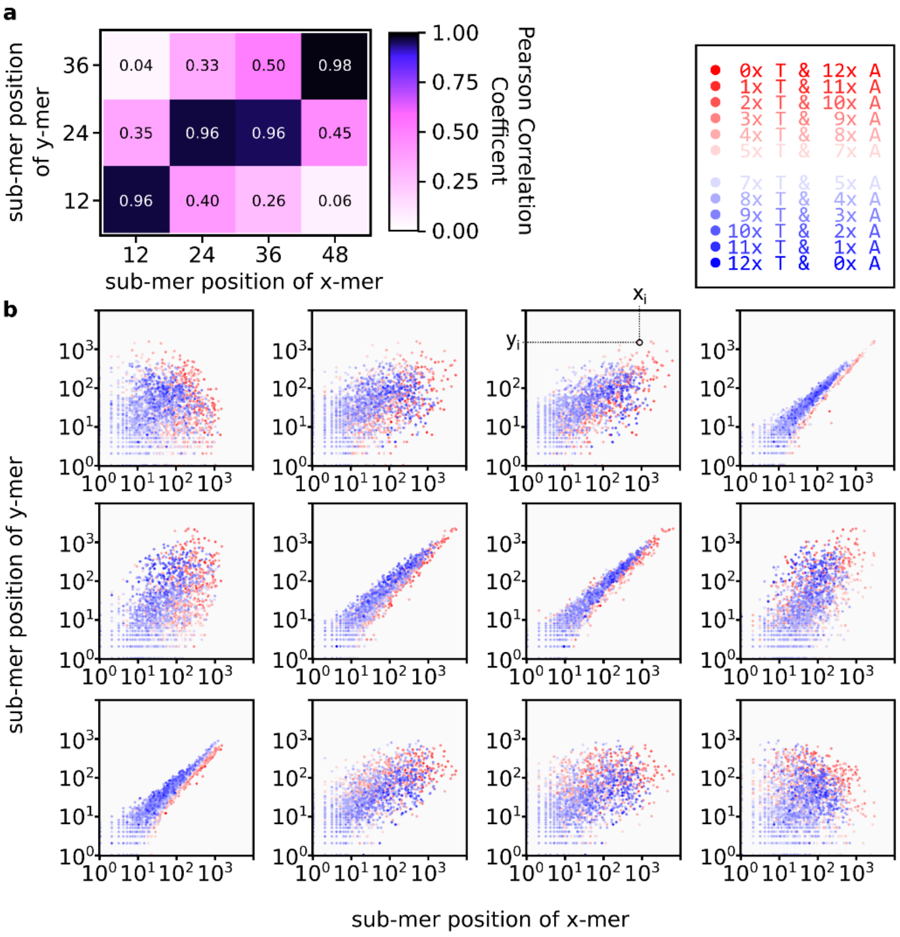
The entropy calculation used here is done as described by Shannon(6) and adapted as by Derr *et al.*(Ref. (43) of the main manuscript):

1) $H_k(s) = -\sum_i p_i \log_2(p_i)$

with the entropy $H_k(s)$ of a sequence s of length L with i as the index of a unique substring of length k. The frequency of the ith k-mer in s is called p_i. For the entropy of single base positions, the length of the substring is 1, for the subsequences it is 12. For the plot in Figure 2a we calculated the entropy of a large set (10^7) of randomly generated AT-sequences with the same algorithm and divided it by the value calculated for the sequencing data. A value of 1 then corresponds to no reduction in sequence entropy in comparison to an ensemble of random sequences. A value close to 0 represent the case of a very low sequence diversity with only a handful of sequences in all strands. For the entropy reduction in single positions the y-axis is scaled by a factor of 12 for easy comparison with the 12mer subsequences.

11. Pearson Correlation Coefficient

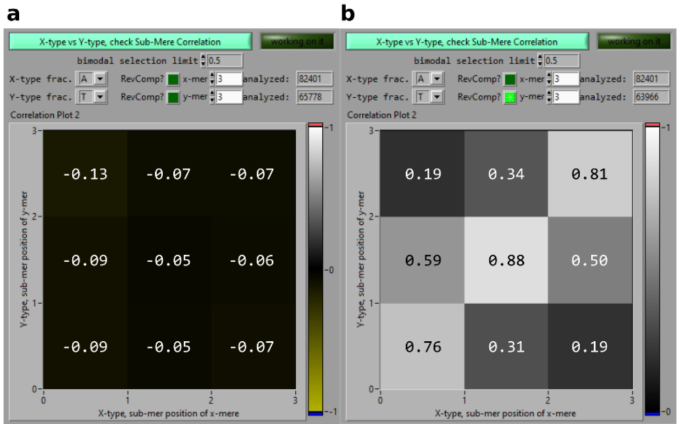
Comparing two populations with the same elements but different distributions can be done with the sample Pearson Correlation Coefficient (sPCC). On the x-axis the elements of one distribution are plotted, sorted from smallest probability to highest probability. The same is done for the second population but on the y-axis. If the populations correlate, elements that are common in one, are also common in the other, the same is true for uncommon elements. The resulting plot can be described by a linear function. If the populations don't correlate, the resulting distribution has no particular shape. SI-Fig. 12b shows all possible sPCC plots for the comparison of all 36mers to all 48mers and all subsequence positions.



*SI-Fig. 12, **sample Pearson Correlation Coefficient for 48mers vs 36mers:***
a *Color-coded sPCC from the sub-plots shown in b.*
b *sPCC plots for all possible combinations of 12mer subsequences for the comparison of 36mer and 48mers. The color codes the A:T fraction for each sequence. In the second subfigure from the top right, points x_i and y_i from equation (3) of the main manuscript are shown.*

The first, the last and the two center subsequences have high correlations, while the first compared to last have very low correlations. SI-Fig. 12b also plots the A:T-ratio: the highly correlating start sequences have two distinct populations for the A-types (blue) and the T-types (red) with different slopes. There is a clear trend towards more A-type sequences in longer oligomers, indicated by the flatter in slope. The plot in SI-Fig. 12a is the color representation of the sPCC values calculated from the linearity of the correlation plots.

Comparing the 36mer A-type sequences to the 36mer T-type sequence reveals a slight negative correlation, as shown in SI-Fig. 13a. Strands with a lot of A are uncommon in T-types and *vice versa*. But comparing the 36mer A-types with the reverse complement of the 36mer T-types reveals an already known pattern. T-type sequences are just the reverse complement of A-type sequences. This is no surprise, as both groups can function as the template for one another.


a
X-type vs Y-type, check Sub-Mere Correlation
working on it
Y-type, sub-mer position of y-mer
X-type, sub-mer position of x-mere
0
1
2
3
-1
-0
--1
b
X-type vs Y-type, check Sub-Mere Correlation
working on it
Y-type, sub-mer position of y-mer
X-type, sub-mer position of x-mere
0
1
2
3
-1
-0

a

bimodal selection limit 0.5

X-type frac. A RevComp? ■ x-mer 3 analyzed: 82401

Y-type frac. T RevComp? ■ y-mer 3 analyzed: 65778

Correlation Plot 2

-0.13	-0.07	-0.07
-0.09	-0.05	-0.06
-0.09	-0.05	-0.07

b

bimodal selection limit 0.5

X-type frac. A RevComp? ■ x-mer 3 analyzed: 82401

Y-type frac. T RevComp? ■ y-mer 3 analyzed: 63966

Correlation Plot 2

0.19	0.34	0.81
0.59	0.88	0.50
0.76	0.31	0.19

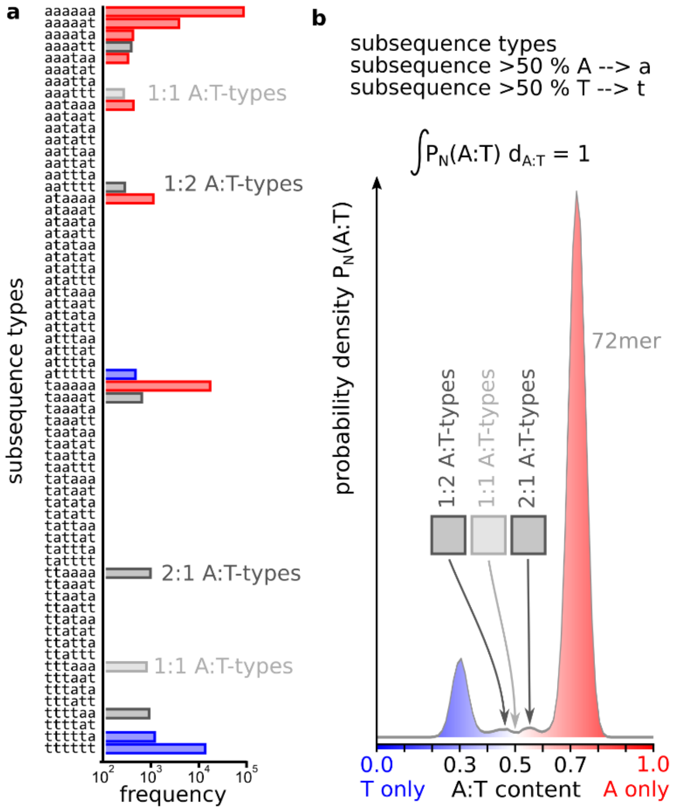
SI-Fig. 13, ***Pearson Correlation Graph for 36mer A-type vs 36mer reverse complement T-type strands:***
Screenshot of the LabVIEW analysis tool.
a *Comparing 36mer A-types and 36mer T-types reveals no correlation.*
b *Comparing the reverse complement of the T-type sequences instead gives a very similar pattern as already known for the comparison of different lengths.*

12. Domain Wall

Starting with the 24mer products, oligomers can be categorized as either A-type or T-type sequences with A:T ratios of about 70:30 or 30:70. But for long oligomers like 72mers the originally bimodal distribution starts to has additional small peaks around A:T fractions of 0.45 and 0.55. SI-Fig. 14 shows the most common subsequence-sequences. Here, every 12mer subsequence is analyzed for its A:T content, if a strand has more than 50 % A it's called "a", for more than 50 % T it's called "t". So a 36mer strand like "AATTAATAAATA-TAATTAAATTTT-AAAAAAATAATA" is noted as "ata". This results in a histogram of 6-sequences long representation of a strand.

As expected, the majority of strands is either completely A-type ("aaaaaa") or completely T-type ("tttttt"). The second most common group is strands with a single subsequence not matching the others, like in an A-type strand a subsequence-sequence like "aaaata".

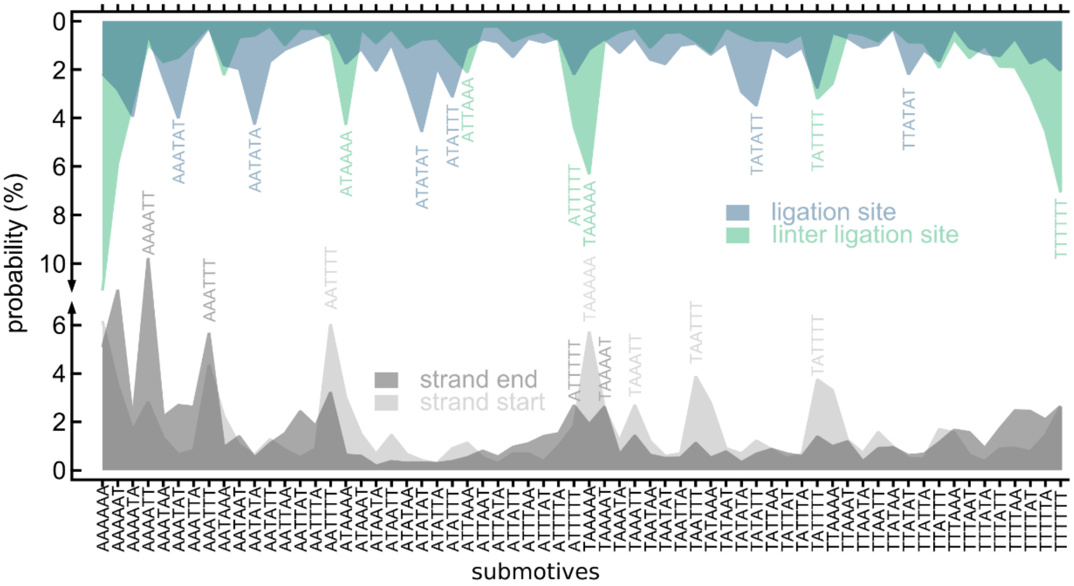
The small additional peaks mark strands with significant read counts that do not match the first two groups and are either 1:1 A- and T-type ("aaattt" and "tttaaa") marked in light grey, or they consist of 2 non-matching subsequences, like "ttaaaa" marked in dark grey. Interestingly, the non-matching subsequences are all consecutive: "aaaatt", "aatttt", "ttaaaa" and "ttttaa", even "taaaat". This is in agreement with the possible self-folding mechanism and the templated ligation mechanism in the experiment discussed in the main manuscript. This particular effect shown here is best described as a domain wall: if a sequence has a certain appearance, like "aaaa" there is a higher chance to attach "aa" than "at", which in turn has a higher probability than "tt".



*SI-Fig. 14, **Most common subsequence types and comparison with AT-composition graph:***
***a** Noting A-type and T-type subsequences in a 72mer as "a" and "t" shows, that not only the entire strand is made from 70 % of A or T, but so is each subsequence. Rare subsequences include aaattt and tttaaa that might fold on themselves and sequences with two non-matching subsequences, like aatttt. In those strands, the non-matching subsequences are predominantly consecutive.*
***b** The groups of subsequence-sequence motives give rise to additional peaks in the originally bimodal A:T-fraction graph shown in the main manuscript.*

13. Correlation of Ligation Sites, Inter-Ligation Sites, Strand Start and Strand End

The ligation site and the inter-ligation sites show distinctly different sequence patterns, as discussed in the main manuscript Figure 4b. But the Analysis of the sPCC shows, there should also be a clear difference between the start and the end of the oligomers. SI-Fig. 15 shows the distribution of all 64 possible 6 nt long submotives in 36mer oligomers. On the bottom, the start and end strands are plotted. They do have certain similarities like a peak for the sequence AAATTT, but are overall clearly different. The same holds for the ligation site and the inter ligation sites and both are again different, compared to the start and end sequences. As suggested by the sPCC the combination of different submotives describes where they can be found in a oligomer.



SI-Fig. 15, ***6 nt long submotives on 36mer reaction products, analyzed in groups of strand position:***
Plotting the probability to find a submotive in the region of the ligation site (blue) and in between ligation sites (green) shows, they have different distributions. This is known from Figure 4b of the main manuscript. The graph suggests, that there is no clear correlation between start- (light grey) and end-of-strand (dark grey) submotives. Interestingly, there is almost no correlation between any of those four groups of submotives.

14. Z-score landscape

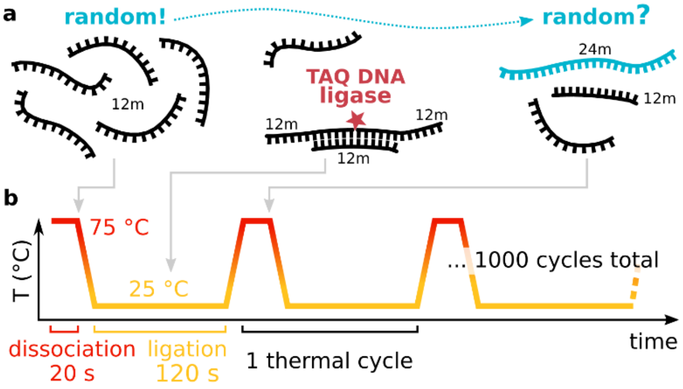
The z-score is used when comparing differently scaled and distributed normal distribution datasets. The z-score shifts and normalizes the dataset by the mean value of the distribution. Data points above the mean of the dataset are given in positive values, data points smaller than the dataset mean are represented by negative values. Equation 1) gives the mathematical representation of the z-score of a sample with index *i*, *j*:

$$2)\quad Z_{ij} = \frac{x - \mu}{\sigma} = \frac{N_{ij}^{\text{observed}} - N_{ij}^{\text{expected}}}{\sqrt{N_{ij}^{\text{expected}}}} \quad \text{with} \quad N_{ij}^{\text{expected}} = \frac{N_i N_j}{N_{\text{total}}}.$$

For the z-score landscape in the main manuscript transitions with either significantly higher or significantly lower than mean probabilities are plotted.

15. First strands ligation

At some point in the experiment during temperature cycling the first ligation events occurs. There is no possibility to find the exact conformation for those events; the results might also be included in longer strands. They might also just be precursor needed for the first reactions that is then not rebuild and simply be so rare, that they are not sequences. Therefore, some estimations must be made to get a simple model for such a reaction. SI-Fig. 16 shows a schematic sketch of a likely conformation. Two strands from the 12mer pool are brought together by a third strand, that templates the other two by Watson-Crick basepairing. Such a complex can be ligated by the ligase enzyme. The products are then the connected 12mers, now as a single 24mer ssDNA strand, and the templating 12mer.



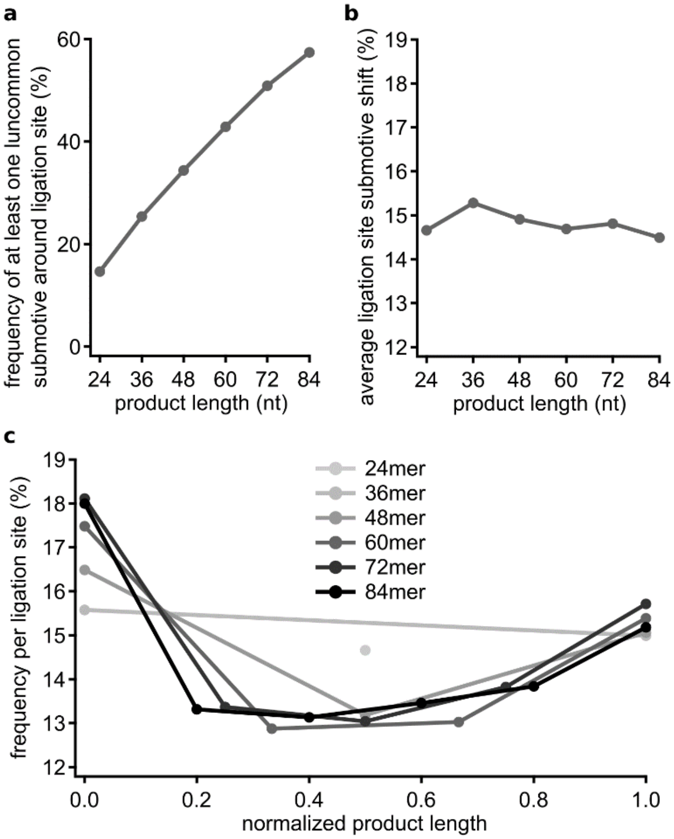
*SI-Fig. 16, **schematic drawing of the hypothesized conformation for the first ligation events:***
***a** The first ligation events must have occurred solely between 12mer monomer strands. As a simple model, those were in a substrate+ substrate +template conformation as shown. Two 12mer strands are brought 3'-end-to-5'start by a templating third strand. The ligase can connect such structures and the result is a 24mer and the templating 12mer.*
***b** Shows the different pool and structure conformations in an exemplarily drawn temperature cycle. At dissociation temperatures, there aren't any dsDNA strands. At the ligation temperature, strands might be in conformations, that can be ligated by the ligase. In subsequent dissociation, those complexes are melted to ssDNA again.*

In this simple model, there is no dynamic aspect included: one might expect to get 24mer sequences as a results for a small amount of temperature cycles already. But as shown in Figure 1 of the main manuscript, the first substantial bands are 36mer and 48mer, while 24mer sequences seem very scarce. Those bands can hardly be seen by eye and even for higher cycle counts, the resulting band structure suggest a non-trivial growth mode. Therefore, it's not possible to conclude, that 24mers that were sequenced have a close correlation to the first emerging sequences.

16. Ligation site shift

In the graph showing the probability for submotives on 72mer junctions (Figure 4a, b in the main manuscript) there is a cluster of junctions with a higher probability of including poly-A on the ligation junction. This is not expected, as all graphs up until here show a distinct AT-alternating pattern for ligation junctions. This effect might best be described as a hybridization or ligation site shift. With a long single stranded unbound oligomer, there is no reason shorter strands like a monomer should hybridize exactly onto a 12mer subsequence. It might rather hybridize with its center on a ligation site. If this monomer is extended by another strand, the AT-alternating and poly-A, poly-T pattern the product inherits from the template is shifted relative to the 5'-end of the strand. The abundance and length dependence of this effect is further described in the following.

Detecting those shifts is done by analyzing the submotives on ligation junctions. As the analyzed strands are all a multiple of 12 nt in length, the junctions have to be after a multiple of 12 read bases. Filtering oligomer-junction submotives for poly-A and poly-T of lengths of minimum 4 nt and maximum 6 nt in a region of plus and minus three bases from the junction produces the data shown in SI-Fig. 17.



SI-Fig. 17, ***Ligation site shift:***
a *Frequency of finding at least one uncommon submotive at a ligation site per strand. Longer strands have more junctions and an about linear increase for the probability of having at least one uncommon submotive around at least one ligation site.*
b *The average probability of an uncommon submotive per ligation site as a function of the product length is about 15 % for all lengths.*
c *The ligation site submotive shift shown in b is not constant per length. The probability to find an uncommon submotive on a ligation site is larger on the outer ligation sites.*

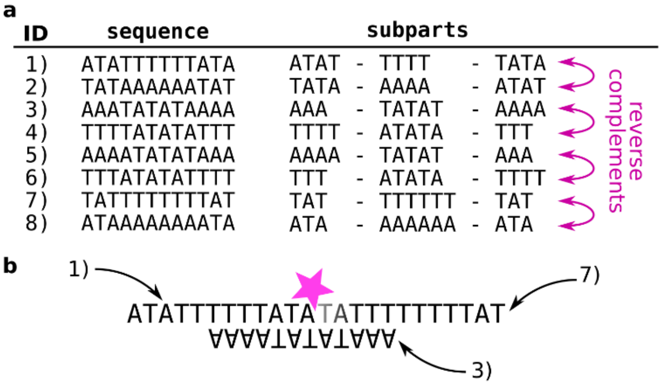
SI-Fig. 17a shows the frequency of finding at least one of those uncommon poly-A or poly-T submotives on at least one ligation junction. With a linear increase, longer strands like 84mers have a chance of about 60 % to include at least one ligation site shift. In SI-Fig. 17b the abundance per junction is analyzed, and as expected from a the average ligation site shift probability is about constant in all oligomers and about 15 %. But SI-Fig. 17c shows, this frequency is not the same for all junctions. The junctions in the oligomer center have a lower probability of including a poly-A or poly-T submotive compared to the outer strands. This can be seen for all oligomer lengths.

After all, Figure 1b and Figure 4a, b of the main manuscript suggest, that overall the ligation site is most common after a multiple of 12 bases.

17. x8 experiment, designed and selected pools of eight sequences

17.1. “Replicator” design

For the comparison of the most common selected sequences of the AT-random sequence 12mer pool we build a set of eight sequences designed to elongate. As described in section 15, we chose strands that are able to form three strand complexes, with one templating strand and two substrates. The sequences are then made from regions of alternating bases and poly-bases: All strands have either two poly-base parts and one alternating bases part, or *vice versa*. For every strand there is also its reverse complement strand, as shown in SI-Fig. 18.

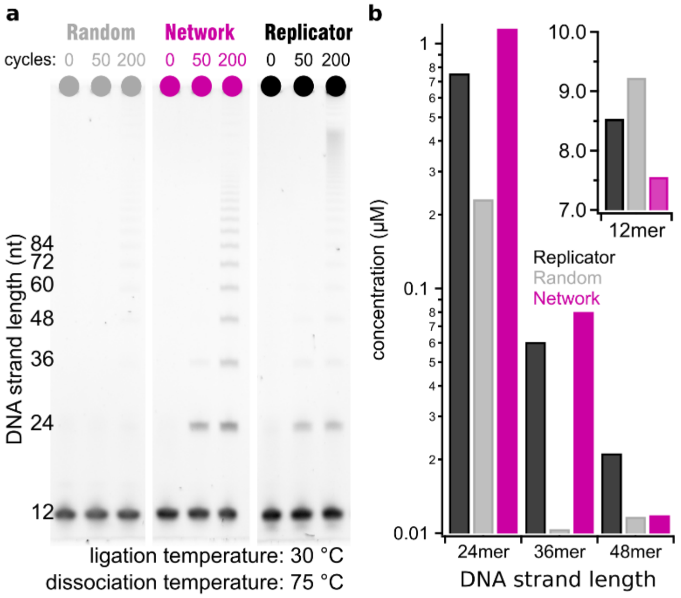


ID	sequence	subparts
1)	ATATTTTTTATA	ATAT - TTTT - TATA
2)	TATAAAAAATAT	TATA - AAAA - ATAT
3)	AAATATATAAAA	AAA - TATAT - AAAA
4)	TTTTATATATTT	TTTT - ATATA - TTT
5)	AAAATATATAAA	AAAA - TATAT - AAA
6)	TTTATATATTTT	TTT - ATATA - TTTT
7)	TATTTTTTTTAT	TAT - TTTTTT - TAT
8)	ATAAAAAAAATA	ATA - AAAAAA - ATA

SI-Fig. 18, ***x8 “Replicator” sequence design:***
a *Sequences and subsequence parts of the Replicator pool. There is either alternating or poly base sections, always in a 1-2-1 or 2-1-2 conformation. The stretches differ in length in order to form properly ligatable dsDNA conformations.*
b *An example for a presumed ligation construct in early temperature cycles. Strands with IDs 1) and 7) act as the substrate while being templated by strand 3).*

17.2. x8 experiment after 50 temperature cycles

The experiment comparing three pools of eight sequences each described in the main manuscript was done for 50 and for 200 temperature cycles. Figure 5c shows the PAGE gels and Figure 5d the concentration quantification for 200 temperature cycles. SI-Fig. 19b shows the concentration estimation for 50 temperature cycles. The inset shows the concentrations for the 12mer band. The Random sample depletes the pool the least, by not even 10 %, while the Network depletes the pool the furthest. The oligomer product concentrations are also the lowest for the Random sample, while the Replicator sample and the Network sample have both about exponentially decaying product concentrations over length.



*SI-Fig. 19, **concentration quantification for x8 pool comparison and 50 temperature cycles:***
a *Graph similar to Figure 4c in the main manuscript showing a PAGE gel of the x8 pools comparison.*
b *Concentration quantification of the PAGE gel in a, drawn as a bar plot showing the concentrations after 50 temperature cycles.*

18. A:T-composition and kinetic simulations

18.1. Composition and hairpin formation

One striking evidence of the selection in the system is the formation of two subpopulations of oligomers: A-type and T-type. A likely origin of this behavior is that the templating ability of a given sequence would be substantially reduced by the formation of internal hairpins and other secondary structures. The same is true for the substrate chains. An easy way to suppress the formation of hairpins is by introducing a composition bias in a sequence. Consider a sequence of length N, which is random but whose composition p (i.e. the fraction of 'A'-bases among all nucleotides) is different from 0.5. How likely is it to find an internal hairpin of length l_0, i.e. two non-overlapping mutually complementary regions of length l_0, within that sequence? The probability of any two given bases in the sequence to be complementary is $2p(1-p)$, which yields the probability $[2p(1-p)]^{l_0}$ for two given segments of length l_0 to be complementary. The number of ways in which two such non-overlapping segments can be chosen on a sequence of length N is $(N-2l_0)^2/2$ (assuming large N). By requiring the expected number of hairpins of length l_0 to be 1, we obtain an equation that relates the sequence length N to the length of a typical maximum hairpin within it:

$$3)\quad N = 2l_0 + \sqrt{2}[2p(1-p)]^{-l_0/2}$$

If the sequences generated by our autocatalytic reaction were completely random, subject only to the constraint of the fixed mean composition for both T-type and A-type chains, we would obtain an ensemble of sequences that maximizes entropy under that constraint:

$$4)\quad S = \int_0^1 \left[f(p)\ln\left(\frac{f(p)}{f_N(p)}\right) - \lambda\, p f(p)\right] dp$$

Here $f(p)$ is the probability density function (PDF) of sequences with composition p at the ensemble, and $f_N(p)$ is the PDF of all sequences of length N with that composition. The latter follows a regular unbiased binomial distribution (sequence of length N is statistically equivalent to N sequential tosses of a coin), and can be well approximated by a Gaussian curve $f_N(p) \sim \mathrm{e}^{-2N(p-1/2)^2}$. Maximization of the above entropy leads to a regular Gibbs-Boltzmann distribution, with abundances of individual sequences proportional to $\mathrm{e}^{\lambda p}$, and the overall PDF

$$5)\quad f(p) \sim f_N(p)\mathrm{e}^{\lambda p} \sim \mathrm{e}^{\pm Npx_0 - 2N(p-1/2)^2}$$

Here $= \pm Nx_0$, corresponds to shift of the mean compositions for A-type and T-type subpopulations, to $p = 1/2 \pm x_0$, respectively. Note that the pre-factors may be different for the two subpopulations, reflecting the compositional bias in the initial ensemble.

As was demonstrated in the prior theoretical works(Ref. (18) and Ref. (19) of the main manuscript, Tkachenko and Maslov, 2015, 2018), the templated ligation leads to exponential distribution of chain length, much like regular step-growth polymerization process(7). The preferential ligation of fragments that belong to the same sub-population (A-type or T-type), leads to distribution that is characterized by two different characteristic chain length (which depends of abundance of corresponding 12-mers, and their ligation probabilities). The overall compositional PDF for the two sub-populations has the following form:

$$6)\quad P(x) = \frac{\beta^{\left(\frac{N}{12}-2\right)} f(p) + f(1-p)}{\beta^{\left(\frac{N}{12}-2\right)} + 1} = \sqrt{\frac{2N}{\pi}}\left(\frac{\beta^{\left(\frac{N}{12}-2\right)}\mathrm{e}^{-Nx_0x} + \mathrm{e}^{Nx_0x}}{\beta^{\left(\frac{N}{12}-2\right)} + 1}\right)\mathrm{e}^{-2N\left(x^2 + x_0^2\right)}$$

Here $x = p - 1/2$ and the parameter β accounts for the compositional bias observed for chains longer than N=24. Note that due to the exponential length distribution with unequal means, the contrast between the two subpopulations is getting exponentially enhanced for longer sequences.

Note that the assumption of maximum entropy is in fact excessive. Our analysis shows that the entropy of the generated oligomer pool is reduced well beyond a simple compositional bias. However, if all other types of selection were not correlated with the value of p, the maximum entropy ensemble would give a correct PDF, $P(x)$. That is indeed the case, as demonstrated in Figure 2g in the main text.

18.2. Kinetic model

In the previous section we demonstrated how the experimentally observed A- and T-type subpopulations can be quantitatively described by postulating that the sequence entropy is maximized, subject to the constraint of a fixed average sequence composition within each of two subpopulations. The bias of their compositions with respect to $p = 1/2$ was, in turn, linked to a suppression of hairpin formation on both template and two substrate chains. Below, we supplement this analysis by a kinetic model that directly illustrates the mechanism of such sequence selection for the case of templated autocatalytic formation of 24mers from pairs of 12mers.

Specifically, we adapt the model developed in Ref. (19) of the main manuscript (Tkachenko and Maslov, 2018) to include effects of hairpin formation which reduces the activity of the reaction of both template and substrate strains. This leads to the following system of kinetic equations:

7) $$\dot{d}_{ij} = \lambda \left(\alpha_{j^*i^*} d_{j^*i^*}\right)(\alpha_i l_i)\left(\alpha_j r_j\right)$$

Here d_{ij}, l_i and r_j are the concentrations of a specific 24mer and its constituent "left" and "right" 12mers. $d_{j^*i^*}$ is the concentration of its complementary 24mer which, within this model, acts as a sole template for d_{ij}. λ is the ligation-rate which in the current version of the model is assumed to be sequence independent.

Coefficients α are activities of respective DNA fragments (12mers i and j, and 24-mer i^*j^*) that depend on the length of the longest internal hairpin l_s for a sequence s. They are calculated assuming that each of the fragments is described by a 2-state model, i.e. with and without hairpin. The activity α is given by the probability to find a given fragment in the hairpin-free state:

8) $$\alpha_s = \frac{1}{1+e^{-(G_0+\Delta G l_S)/\mathrm{kT}}}$$

Here ΔG is the hybridization free energy per single base ($\Delta G \approx 1.5\ kT$ for random AT-based sequences(8), and G_0 is a threshold free energy that accounts for termination of the hybridized region (about 1.5 kT at each side), and for the loop formation. Based on the analysis of Ref. (9), we assume $G_0 \approx 6\ kT$. For each of the 2^{12} 12mers and 2^{24} 24mers, we have determined the longest hairpin length l_s, and the corresponding activity factor α_s.

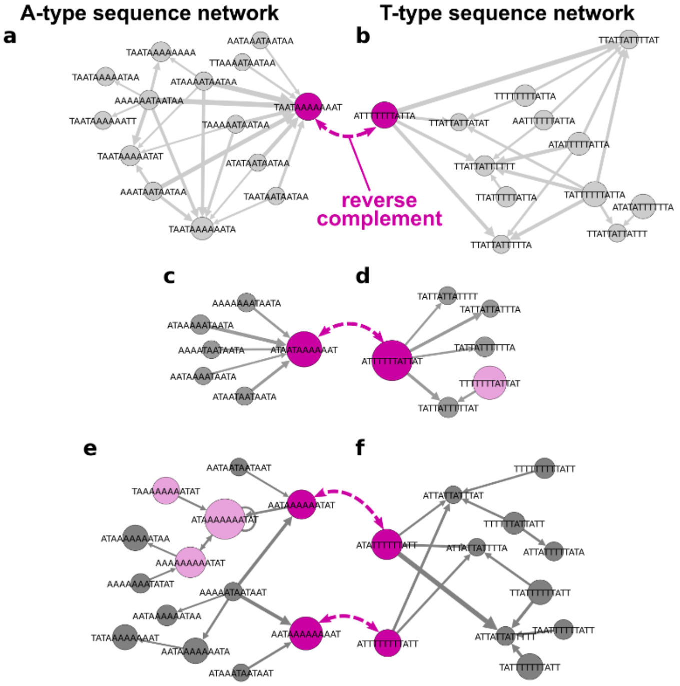
We then solved the set of equations eq(8) numerically, and observed the development of the bimodal composition profile indicating the emergence of A-type and T-type subpopulations within 24mers, as shown in Figure 2d in the main text.

This kinetic model only accounts for templating of 24mers by other 24mers. It is therefore only applicable at relatively early stages of the experiment. Once longer oligomers appear, one needs to account for sequence-dependent suppression of activity due to formation of hairpins within those longer chains. As discussed above, the average composition can be used as a reasonable proxy to find the length of the longest internal hairpin within a sequence. As a result, the overall composition PDF is well described by the minimalistic, maximum entropy model presented in 18.1.

19. Network-families resemblance

Figure 5a and b in the main manuscript show 12mer subsequences common in oligomers longer than 48 nt as a de Bruijn network graph, that depicts each sequence as a node and sequences that follow each other are connected by an edge. Nodes and edges are scaled with the abundance of the sequence and the connection.

As stated in Figure 4 and SI-Fig. 13 A-type and T-type sequences are mostly the reverse complement of each other. In the eight most common A-type and T-type sequences, there are four sequences of which the reverse complement can be found in the other respective group (reverse complement dark pink, other four most common sequences light pink color). The network of the most common sequences is not entirely connected, but consists of several "families" of sequences. Those families can themselves be a very intricate network, or just contain one sequences that tends to follow itself.



SI-Fig. 20, ***Network families show similar network structures:***
a & b *Both families are intricate networks, that have multiple internal connections as well as several start and end sequences.*
c & d *Both networks are very directed, with the A-type sequence ending in one specific sequence, and the T-type sequence, starting with a specific sequence.*
e & f *The two families both have two sequences that are found in the most common and reverse complement sequences of the network. The sequences are connected in the network by one (A-type) and two (T-type) intermediate sequences.*

SI-Fig. 20 shows the six network families with the four most common and reverse complement sequences. The families of A-type and T-type with the respective common reverse complement sequence each have similar forms, that are distinctly different for the three groups. SI-Fig. 20 a, b are a network with multiple internal connections and several start- and end-sequences. In contrast, for SI-Fig. 20 c, d the networks are very directional. The A-type sequence is the end-sequence for all connected sequences, while the T-type is the start for all connected sequences. And in both cases, the attached sequences have no relevant edges connecting them among themselves. In SI-Fig. 20 e, f the families each contain two sequences that have a reverse complement in the other respective family. In the A-type there is one, in the T-type there are two intermediate sequences connecting them in the network graph. In both cases, there are no edges that would suggest commonly finding both sequences in the same oligomer strand.

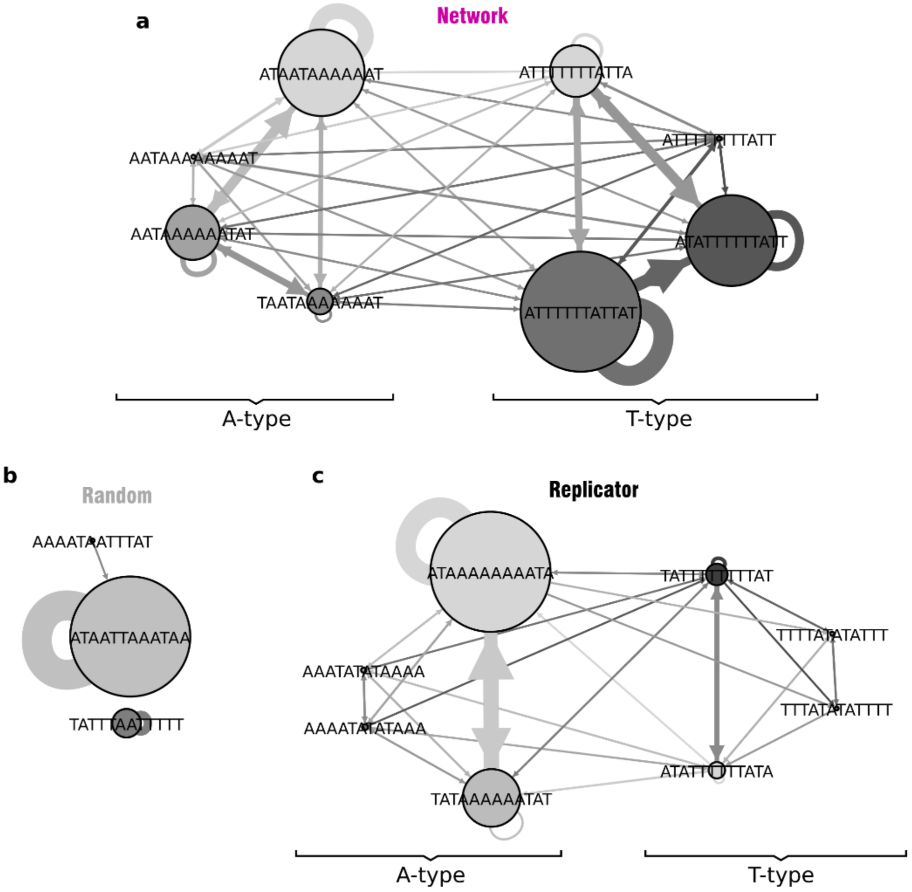
Despite the differences in the form of the families, the new initial pool made from the most common four A-type and four T-type sequences has a high oligomer production rate and compared to the total amount of oligomer products a remarkably similar sPCC matrix. This is despite the networks explicitly not taking the first and last subsequence in oligomer products into account – as those are distinctly different to the center subsequences (discussed above).

20. x8-pools de Bruijn graphs

The de Bruijn sequence network in section 19 shows a subset of common sequences in the resulting oligomer product strands. For the pools with only eight different strands discussed before, the sequence network can be plotted entirely. In SI-Fig. 21 the de Bruijn networks for the x8 samples are shown. The network sample forms an intricate network and includes all possible 64 edges. As for the AT-random sample, the x8 network sample forms A-type and T-type sequence groups. Those are better connected internally, as to the reverse complement group. Together, the subset from the AT-random sample selected by the templated ligation behaves very similar to the original system.

The sPCC matrix for the random x8 sample suggests a high level of similarity for almost all positions. The de Bruijn network reveals that oligomers are almost entirely made from a single sequence (ATAATTAAATAA). From the eight sequences in the pool, five are not sequenced in oligomers at all.

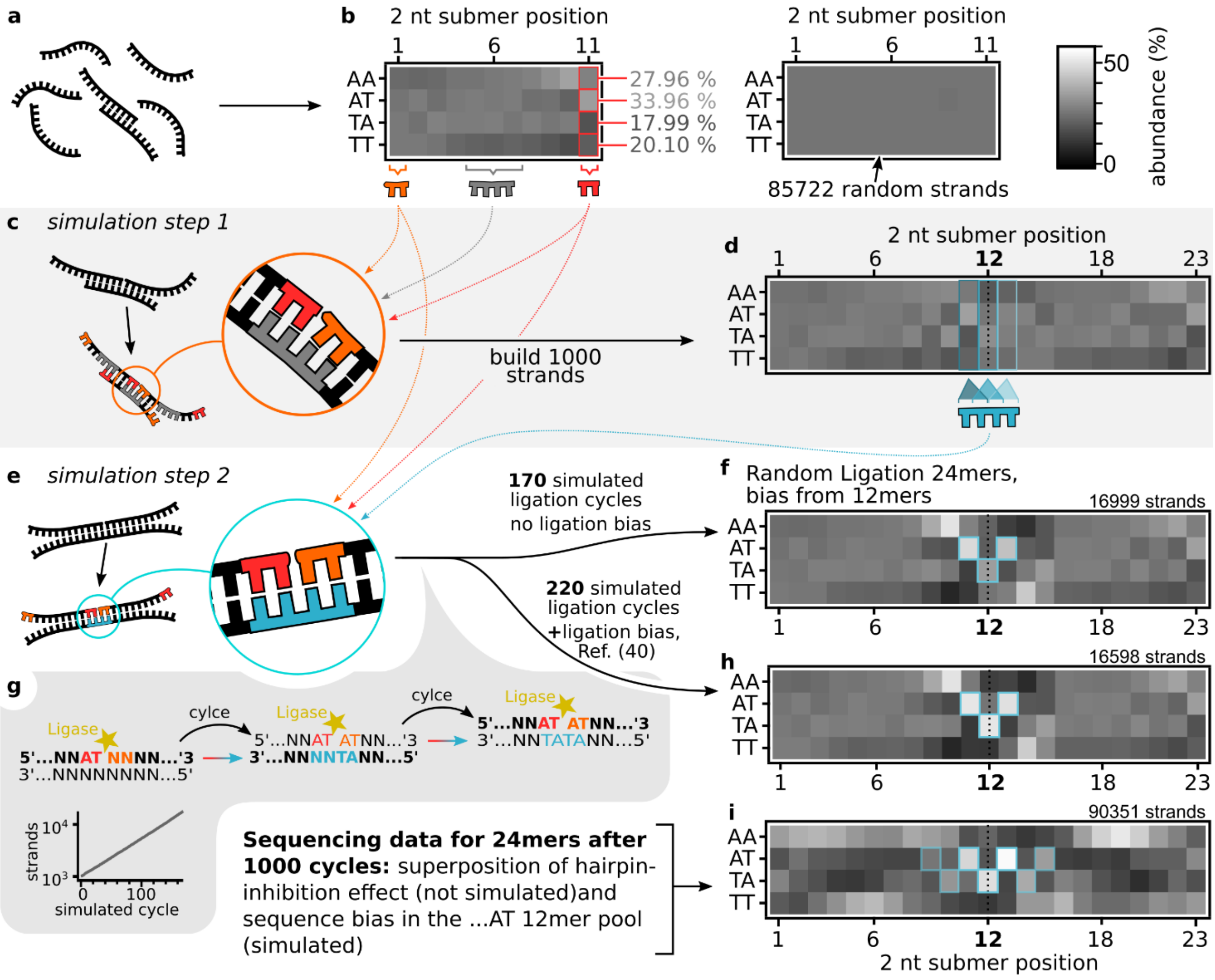
The replicator sample predominantly forms oligomers made from the four sequences with poly-A and poly-T at the center and alternating sequence motives on the start and the end of the 12mer “monomers”. Because the alternating section starts (and ends) either with ATA…. or with TAT… alternating motives are common, leading to an alternating sequence motive …ATATA… on the ligation site, known from the AT random sample. And again, A-type strands and T-type strands frequently connect to themselves, while connections from one type to the other are rare forming the A-type and T-type oligomer groups that inhibit hairpin formation.



SI-Fig. 21, ***de Bruijn network graphs for the three x8 pools:***
a *The network sample shows an intricate network, where all eight sequences are interconnected. The sequence ATTTTTTATTAT is most common. There are frequent connections to other T-type sequences. As for the AT-random oligomers, the network sample oligomers tend to stay purely A-type or T-type.*
b *For the random sample, as expected from the sPCC-matrix, long strands are made predominantly from two sequences, that both tend to follow themselves. From the other six sequences in the pool, five are not found in oligomer products at all.*
c *The replicator sample network is also interconnected, but has significantly fewer edges. The most common sequences ATAAAAAAAATA and TATAAAAAATAT are often found in an alternating motif, and so are their reverse complements. Oligomers with poly-A or poly-T at the ligation site are rare in comparison.*

21. Random Templated Ligation Simulation based on 12mer pool motif bias

In multimer product strands, the ligation site can often be identified by simply searching for the sequence pattern “ATAT”. The ligation-site sequence landscape in Fig. 4a of the main manuscript shows that the ligation site is actually diverse, but still, some sequence-patterns (like “ATAT”) are very common. This might indicate a preference of the TAQ DNA ligase for certain sequence motifs, which is equivalent to a sequence-dependent ligation rate. We selected the ligase for its sequence specificity and low error rate (see Ref. 39 and Ref. 40 of the main manuscript), but a sequence-dependent ligation rate might be responsible for the analyzed sequence entropy reductions and selections. To quantify a possible ligase bias, we simulated a simple particle-based random templated ligation.



*SI-Fig. 22, **random templated ligation of 12mers based on the sequence motif bias in “monomer” strands:***

a *random sequence 12mer strands with random A, T sequences.*

b *plot of the abundance of 2 nt long subsequences for each position in the 12mer “monomer” strands. At the last position (11) the bias towards the sequence motif “AT” is more dominant than expected for random. On the right: the same amount but truly random sequence strands.*

c & d *Simulation step 1: building a pool of 1000 random 24mer strands, based on the sequence motif frequencies in 12mer strands known from NGS sequencing. For the initial 24mer template strands we assume complexes made from three strands. The rate (or more accurately for this simulation: the probability) of ligation is based on the abundance of the template strand, times the abundance of the 3’-end sequence motifs of 12mers times the abundance of the 5’-start sequence motifs of 12mers. Because there is the before mentioned bias on the 3’ end, the motif “AT” is also visible for 24mers at positions 11 and 23.*

e & f *Simulation step 2: as known from Ref. (10) (see main manuscript) the predominant growth mode for a system with already existing longer templates is blunt-end “primer extension”. For “dimers” in this experimental setting, this corresponds to one 24mer templating the growth of two 12mer “monomers”. In the simulation here, the template motif is taken from the center position in the 24mers build in step 1, panel c & d. Again, the ligation rate is dependent on the abundance of the template sequence motifs (light blue) and the abundance of the 12mer binding motifs (red and orange). Newly ligated 24mers are included as possible template strands in the next cycle. This leads to an exponential growth of strands in the 170 cycles of this simulation.*

g *Hypothetical model for the emergence of the ligation site …ATAT…motifs. The bias towards “…AT”-3’ is amplified, due to its higher abundance in 12mer strands. Over time, even from an about random sequence motif distribution of template sequences (light blue, panel d), this “…AT”-3’-motif*

bias induces an increased abundance of the corresponding "AT" reverse complement sequence motif at position 13 of the "dimer" strands. The resulting 2 nt submotif abundances per position clearly show the "…ATAT…" pattern at the ligation site (panel f).
h *In Ref. (40) of the main manuscript Lohman et al. characterize the TAQ DNA ligase specificity for three complementary strands, but single possible mismatches at the last position of the first substrate and the first position of the last substrate. At standard buffer conditions, all possible four perfect reverse complements scored a ligation efficiency of ">80%" and only the template "AA" with substrates "…T-3'" and "5'-T…" had an efficiency of "50-80%". Implementing this bias in the simulation changes the outcome only marginally. The "AT" pattern induced at position 13 is clearly visible.*
i *2 nt sequence motifs for the sequenced 24mer strands after 1000 temperature cycles. In this sequencing data, the strands already experienced a selection due to the inhibition of hairpin sequences (see Fig. 2c, d). The ligation site pattern "…ATAT…" is still clearly visible.*

The 12mer AT-only random sequence pool the experiment is based on, is synthesized by *biomers.net*. and as discussed in SI-section 9.1, the pool is about random with a bias towards A-type strands. But additionally, there is a specific sequence bias at the 3'-end of the 12mer strands, see inset in Fig. 1c of the main manuscript: the last two bases have a higher than expected frequency to be "AT" (expected: 1/4 = 25 %, sequenced: 34 %). This might cause the emergence of the ligation site "ATAT" motif bias in multimers.

This question can be addressed with a simple simulation: if the ligation reaction was truly random and unbiased by the ligase, the emergence of multimers would only be governed by the abundance of template and substrate motifs. The simulation consists of two major steps:

1. The 12mer "monomer" strands from the NGS data are analyzed for their position dependent subsequence motifs. The abundance of all subsequence motifs of length ≥ 2 is calculated (2^{length} individual motifs). For the first templated ligation reactions in the experimental system we assume a complex made from three individual 12mer strands, as described in SI-section 15. One 12mer acts as the template for two 12mer substrate strands. The probability of such an initial 24mer formation in this simulation here is calculated as follows: one strand of the sequenced 12mer pool is randomly chosen. This choice is weighted by the abundance of all sequenced 12mers (e.g. strand "AAATAATTATAT" is sequence 12 times, "AAATATAAAATA" is sequenced three times, therefore its four times more likely to randomly choose the first of those two). Then the reverse complement of the templating center section of the chosen 12mer is bisected. The probability to find a certain "dimer" strand is the product of the normalized abundance of motifs for the 3'-end of a substrate strand and the 5'-start of another substrate strand (e.g.
    - probability of selecting strand "AAAT**AATA**TTAT": 14/85722 → template motif "AATA" with reverse complement "TATT"
    - substrate 1 with 3'-end "TA", probability: 18 %
      → randomly choose one strand of all possible 12mers that end in "TA"
    - substrate 2 with 5'-start "TT", probability: 26 %
      → randomly choose one strand of all possible 12mers that start with "TT"
    - probability of ligation: 0.18*0.26 = 0.0468).

   Here, the tool calculates 1000 "initial" 12mer stands by randomly choosing a weighted template and then randomly choosing one of the 12mer strands with the corresponding sequence end or sequence start (also weighted). From this set of strands an analysis with 2 nt subsequence motifs is performed, as shown in SI-Fig. 22d.
2. The experiment runs for 1000 temperature cycles and we expect to predominantly sequence strands, that are ligated comparably late in the experimental timeframe, as the strand concentration of multimer products increases non-linearly over time. From the detailed Gillespie simulation in Ref. (10) we know, that the growth mode in a "developed" system is mainly what they call "primer extension", which corresponds to a blunt-ended final complex in our experiments: in contrast to the initial ligations in the first step, the 24mers are used as template strands and ligate two 12mers. Successfully ligated new 24mers are added to the pool of 24mer template strands. This second simulation step is repeated for several cycles, similar to the experiment, resulting in an exponentially growing amount of strands, see SI-Fig. 22g.

For a more detailed resolution of the ligation site, the simulation can analyze subsequence motifs of lengths ≤ (monomer length)/2. For the simulation done here in shown in SI-Fig. 22 motifs of length 3 nt are analyzed (therefore, the ligation site sequence has a length of 6 nt). But for visualization purposes 2 nt motifs are plotted, as the binary code is easier to read. Longer motifs provide a more detailed graph with more “sequence-sensitivity” but might still have the same result. An increase in motif length also significantly increases the calculation time. For 2 nt motifs there are only 2^2=4 possible motifs with a mean probability of 25 % per substrate in the ligation step, but for 4 nt motifs there are 2^4=16 possible motifs with a mean probability of 6.25 % per substrate in the ligation step.

In SI-Fig. 22 this simulation is performed for the AT-only random sample after 1000 temperature cycles. The selection of this sample in comparison to the one which was not cycled at all is important since we want to analyze emerging strands in an already “developed” system, as mentioned above. After 170 cycles in the second simulation step the system size increased to 16999 strands. In the 2 nt motif analysis of the emerging strands three different regions can be seen. Until position seven or eight the motifs are about random, but “TT” motifs are less abundant. The same holds true for the region from position 16 to 23. Here, the last position does also show an increase in “AT” motifs compared to random, as expected. The center region from position 9 to 15 has more pronounced frequencies for certain motifs: “…A**ATAT**T…”. SI-Fig. 22g depicts the presumed amplification mode of this motif bias. The template sequence in the first simulation step has only small deviations from random. Therefore, a biased 3’-end motif will be found more frequently in “dimers” as well. In the second simulation step those strands act as template and the original position the 3’-end is now the template for the 5’start of the second substrate strand. And because the reverse complement of “AT” is also “AT”, the center motif becomes “…ATAT…”.

When comparing this graph to the one made for the NGS data of 24mers after 1000 temperature cycles, this pattern in the center is very similar. In contrast though, the sequenced 24mers also show the results of the hairpin-inhibition selection shown in Fig. 2c, d of the main manuscript. Here, “AA” and “TT” motifs are more abundant in the first and last region, as this reduced the probability of self-folding due to a lack of reverse complement bases. The ligation region in the simulation is restricted to the center of the 24mer strands. In the experiment, 12mers might hybridize with an offset of 2 nt towards either side, but with a reduced probability (see Ref. (10)). This will induce a similar “AT” pattern at the positions 9 and 15.

In Ref. (40) of the main manuscript Lohman *et al.* characterize the TAQ DNA ligase specificity for three complementary strands, but single possible mismatches at the last position of the first substrate and the first position of the last substrate. At standard buffer conditions, all possible four perfect reverse complements scored a ligation yield of “>80%” and only the template “AA” with substrates “…T-3’” and “5’-T…” had a yield of “50-80%”. Mismatched ligation sites in otherwise perfectly hybridized complexes were ligated with “2-10%” yield. From Ref. (39) it is known, that the overall specificity of the TAQ ligase is high and rarely ligates incorrect bases, but the overall yield is only about 80 %. Lohman *et al.* find an above 80 % yield for the template-“AA” case for increased buffer pH-conditions of 8.0 (standard: 7.5). Unfortunately, Ref. (40) only shows the coarsely binned values and not more detailed ligation yield data. Ref. (39) and the pH-experiments in Ref. (40) indicate that, the ligation rates could be about similar, around 80 %. Anyhow, implementing this sequence bias as a worst-case-scenario for comparison with an unbiased simulation produces panel h in SI-Fig. 22. The exact ligation rate per ligation event is random but linearly scaled in the yield ranges given by Ref. (40), e.g. for the “AA”-template case all ligation rates between 0.5 and 0.8 are similarly likely and different for each simulated ligation.

The 2 nt submotif graphs for all three panels f (unbiased), h (biased) and i (NGS data) are remarkably similar. Effects that (presumably) occur in the NGS data include the above mentioned ligation site shift (see SI-section 16 as well) or the inhibition of hairpin sequences. These effects apparently change the start- and end-regions of the 24mer, as the ligated center sections seem to be dominated by the “AT”-pattern induction due to the “AT” bias at the last position of 12mer strands following the mechanism described in panel g.

Again, just as the shift towards predominantly A-type ligation product strands seen in Figure 2e, f stems from a small bias in the original 12mer pool, the emerging “ATAT” pattern at the ligation site stems from a sequence motif bias at the 3’-end of monomer strands. This mechanism will also work for a bias towards “TA” and also at the 5’-start. This simulation shows an emerging ligation site pattern without a sequence-dependent ligation rate and produces results comparable to the NGS data set. With the mechanism being understood, it is still possible, that the TAQ DNA ligase has a ligation rate bias. But this bias must be less significant than the biased initial pool and probably only affects region close to the ligation site. Our findings of the hairpin inhibition effect and the systems' selection of a minimal set of most suitable 12mer motifs is most certainly unaffected by the ligase.

22. List of DNA used

List of DNA used			
Name	**Length (nt)**	**Sequence (5' to 3')**	**Modification**
AT-random_12m	12	WWWWWWWWWWWW	5'-POH
GC-random_12m	12	SSSSSSSSSSSS	5'-POH
GC-random_10m	10	SSSSSSSSSS	5'-POH
AT_x8_Replicator_01	12	ATATTTTTTATA	5'-POH
AT_x8_Replicator_02	12	TATAAAAAATAT	5'-POH
AT_x8_Replicator_03	12	AAATATATAAAA	5'-POH
AT_x8_Replicator_04	12	TTTTATATATTT	5'-POH
AT_x8_Replicator_05	12	AAAATATATAAA	5'-POH
AT_x8_Replicator_06	12	TTTATATATTTT	5'-POH
AT_x8_Replicator_07	12	TATTTTTTTTAT	5'-POH
AT_x8_Replicator_08	12	ATAAAAAAAATA	5'-POH
AT_x8_Random_01	12	AAAATAAAATAT	5'-POH
AT_x8_Random_02	12	ATAATTAAATAA	5'-POH
AT_x8_Random_03	12	TAAAAATTATTT	5'-POH
AT_x8_Random_04	12	TTAAATTTTATA	5'-POH
AT_x8_Random_05	12	TATTTAATTTTT	5'-POH
AT_x8_Random_06	12	TAAAAATTAATA	5'-POH
AT_x8_Random_07	12	AAAATAATTTAT	5'-POH
AT_x8_Random_08	12	TTATATAAAATA	5'-POH
AT_x8_Network_A-type_01	12	ATAATAAAAAAT	5'-POH
AT_x8_Network_A-type_02	12	AATAAAAAAAAT	5'-POH
AT_x8_Network_A-type_03	12	AATAAAAAATAT	5'-POH
AT_x8_Network_A-type_04	12	TAATAAAAAAAT	5'-POH
AT_x8_Network_T-type_01	12	ATTTTTTATTAT	5'-POH
AT_x8_Network_T-type_02	12	ATATTTTTTATT	5'-POH
AT_x8_Network_T-type_03	12	ATTTTTTTTATT	5'-POH
AT_x8_Network_T-type_04	12	ATTTTTTTATTA	5'-POH

23. Supplementary Information Sources

1. M. Kinjo, R. Rigler, Ultrasensitive hybridization analysis using fluorescence correlation spectroscopy. *Nucleic Acids Res.* (1995) https:/doi.org/10.1093/nar/23.10.1795.

2. J. G. Wetmur, N. Davidson, Kinetics of renaturation of DNA. *J. Mol. Biol.* (1968) https:/doi.org/10.1016/0022-2836(68)90414-2.

3. I. I. Cisse, H. Kim, T. Ha, A rule of seven in Watson-Crick base-pairing of mismatched sequences. *Nat. Struct. Mol. Biol.* (2012) https:/doi.org/10.1038/nsmb.2294.

4. I. Schoen, H. Krammer, D. Braun, Hybridization kinetics is different inside cells. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 21649–21654 (2009).

5. E. A. Venczel, D. Sen, Synapsable DNA. *J. Mol. Biol.* (1996) https:/doi.org/10.1006/jmbi.1996.0157.

6. C. E. Shannon, A Mathematical Theory of Communication. *Bell Syst. Tech. J.* (1948) https:/doi.org/10.1002/j.1538-7305.1948.tb01338.x.

7. P. J. Flory, *Principles of polymer chemistry* (1953).

8. J. SantaLucia, D. Hicks, The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* (2004) https:/doi.org/10.1146/annurev.biophys.32.110601.141800.

9. D. P. Wilson, A. V. Tkachenko, J. C. Meiners, A generalized theory of DNA looping and cyclization. *EPL* (2010) https:/doi.org/10.1209/0295-5075/89/58005.

10. J. Rosenberger, *et al.*, Self-assembly of informational polymers by templated ligation, *submitted* (2020).