

MESA: automated assessment of synthetic DNA fragments and simulation of DNA synthesis, storage, sequencing and PCR errors

Bioinformatics, 2020, 1–5

doi: 10.1093/bioinformatics/btaa140

Advance Access Publication Date: 4 March 2020

Original Paper

OXFORD

Sequence analysis

MESA: automated assessment of synthetic DNA fragments and simulation of DNA synthesis, storage, sequencing and PCR errors

Michael Schwarz^{1,†}, Marius Welzel ^{1,†}, Tolganay Kabdullayeva², Anke Becker²,
Bernd Freisleben¹ and Dominik Heider ^{1,*}

MESA

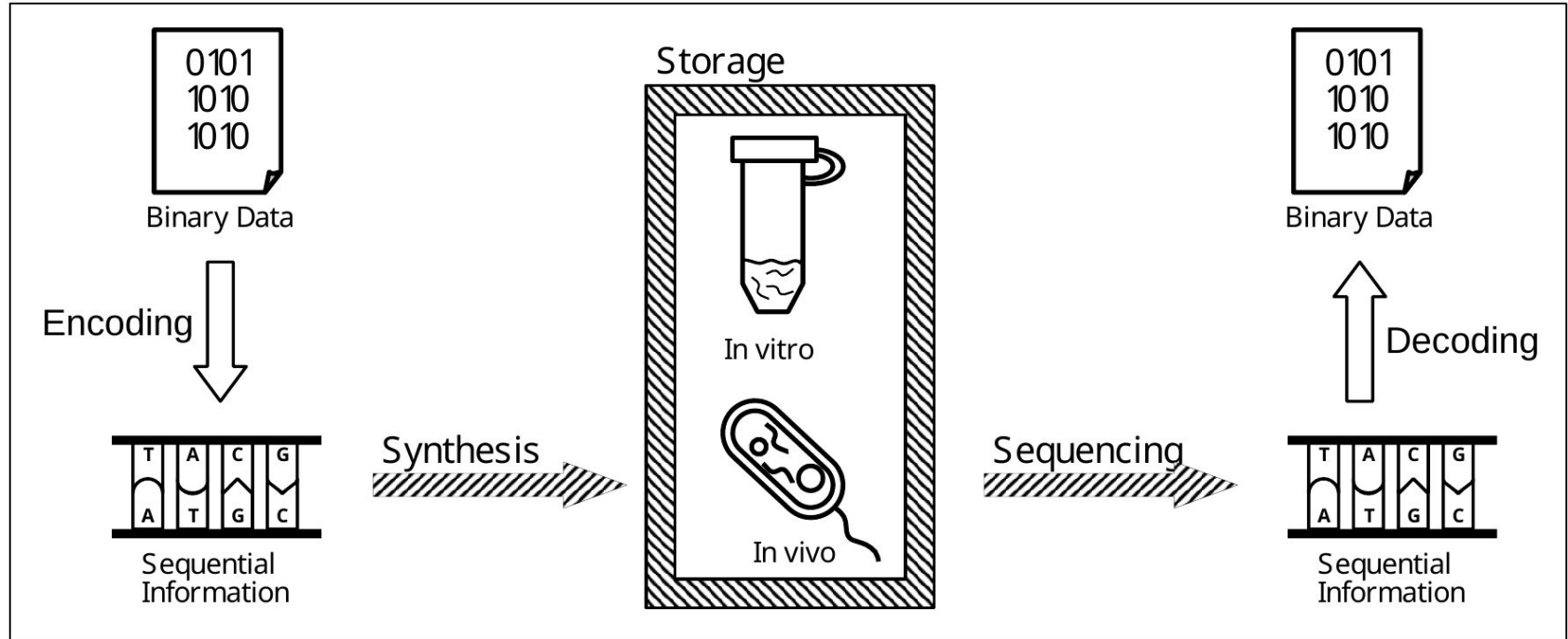
<https://mesa.mosla.de/>

Part 1: overview and general usage

Marius Welzel, M. Sc.

21.04.22

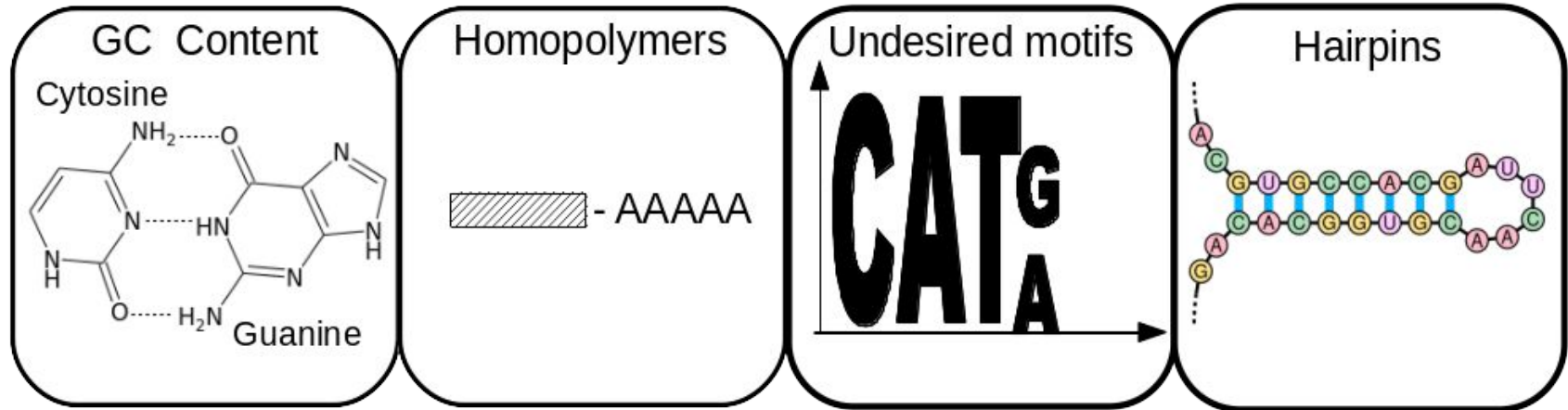
Data storage in DNA



Hannah F. Löchel, Fractals in Bioinformatics: Visualization, Analysis, and construction of Molecular Codes, 2021

Motivation: Sequence Evaluation

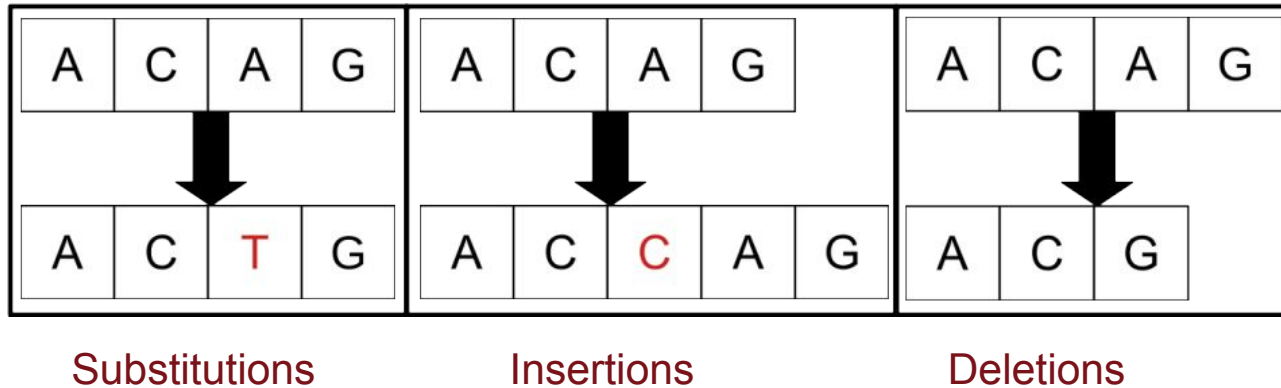
- Rapidly changing requirements warranted a customizable tool to evaluate sequences regarding constraints



Hannah F. Löchel, Marius Welzel, Georges Hattab, Anne-Christin Hauschild, Dominik Heider,
Fractal construction of constrained code words for DNA storage systems, Nucleic Acids Research, 2021.

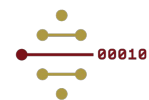
Motivation: Error Simulation

- Classical channel models are less applicable to DNA data storage
- DNA data storage channel simulation can help with the developments of codecs



Motivation: Customizability

- Rapid technological progress in the area of DNA sequencing and synthesis
- Different experimental conditions lead to different constraints that encoded data has to adhere to
- → Tools to evaluate sequences and simulate errors have to reflect this by being highly customizable



Sequence to analyse

AAACCACT

Error Detection

Manage Undesired Motifs

GC Window

50

Default Graph



Change GC Probability

Kmer Window

10

Default Graph



Change Kmer Probability

Calculated Error Prob.



Default Graph



Change Homopolymer Prob.

Error Simulation

Advanced Error-Simulation Settings



Synthesis Method

ErrASE



PCR Cycles

30

PCR Polymerase

Taq



Months of storage to be simulated

24

Storage Host

E coli



Sequencing Method

Paired End



Secondary Structure prediction

Max. Expect



Temperature (*K)

310.15



Configuration

Seed

Random

Download current config

Upload config/fasta file

Send E-Mail



Submit

Motif selection



Manage Undesired Motifs

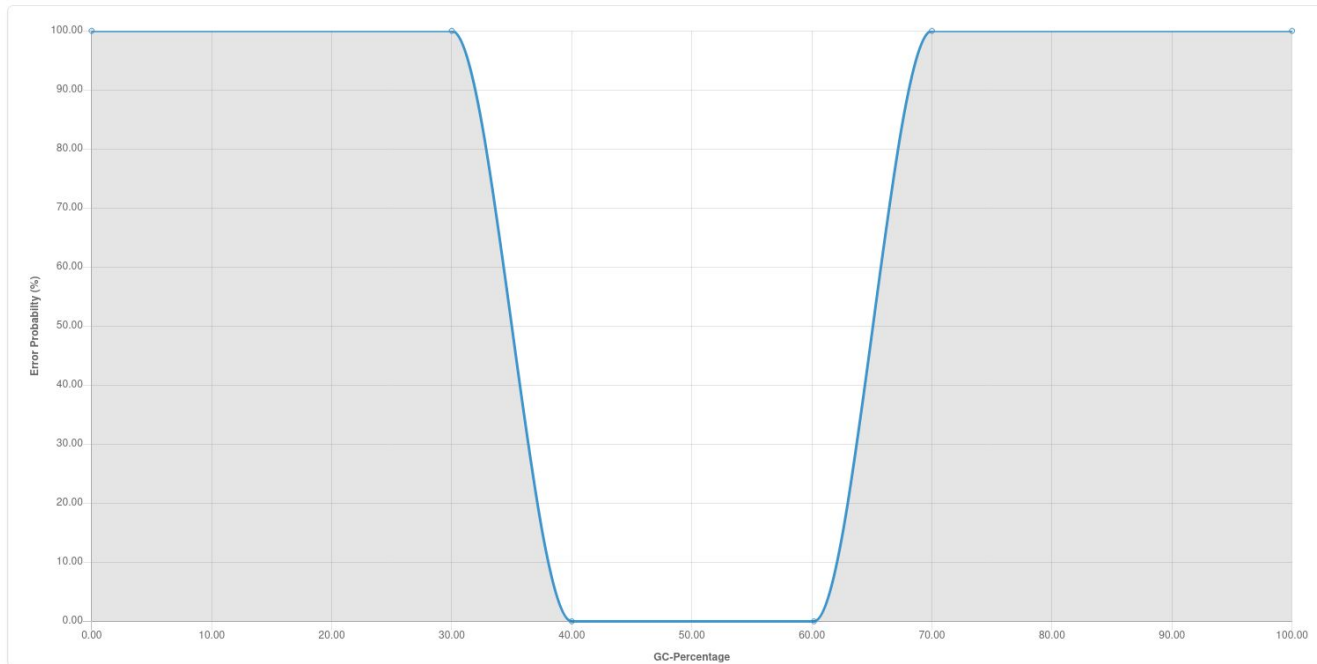
☐ All

Changes here wont be saved in your profile!

Enabled:

<input type="checkbox"/>	Sequence TATAAA	Error Probability 100.0	Description Eukaryotic promotor recogniti	Validated: <input checked="" type="checkbox"/>
<input type="checkbox"/>	Sequence TTGACA	Error Probability 100.0	Description Prokaryotic promoter recognit	Validated: <input checked="" type="checkbox"/>
<input type="checkbox"/>	Sequence TGTATAATG	Error Probability 100.0	Description Prokaryotic promoter recognit	Validated: <input checked="" type="checkbox"/>
<input type="checkbox"/>	Sequence GCCACCATGG	Error Probability 100.0	Description Eukaryotic ribosomal binding s	Validated: <input checked="" type="checkbox"/>
<input type="checkbox"/>	Sequence ACCACCATGG	Error Probability 100.0	Description Eukaryotic ribosomal binding s	Validated: <input checked="" type="checkbox"/>
<input type="checkbox"/>	Sequence AATAAA	Error Probability 100.0	Description Eukaryotic polyadenylation sig	Validated: <input checked="" type="checkbox"/>
<input type="checkbox"/>	Sequence TTGTGTGTTG	Error Probability 100.0	Description Eukaryotic polyadenylation sig	Validated: <input checked="" type="checkbox"/>
<input type="checkbox"/>	Sequence ATAACTTCGTATAGCATACATTATACGAAGTTAT	Error Probability 100.0	Description loxP https://doi.org/10.1016/j.jl	Validated: <input checked="" type="checkbox"/>
<input type="checkbox"/>	Sequence ATAACTTCGTATAGCATACATTATACGAACGGTA	Error Probability 100.0	Description loxR https://doi.org/10.1016/j.jl	Validated: <input checked="" type="checkbox"/>

GC content, homopolymer length and kmers



X-value: 15 Y-value: 15

Add / Modify point Remove point Set Y-Values < X Set Y-Values > X

Reset Use Interpolation Allow drag along X-Axis Publish

Name: Default Graph Delete Save copy of Template

Sequence evaluation



Your Results:

Overall:

GTGGAGTACACTGTAAAGATGGCAACAGTCCCCCCCCCGGAGCCTCGCGTATATCGGCCTAGAGTGATAGTCGCGCTAGGGACGTGGATTAGCAATCGAGTGACGTGGGCAGCTTACATAAAATG

Error Probability: 6.03%

[Download original sequence as FastQ](#)

Sequence evaluation



Expand Results

Subsequences:

GTGGAGTACACTGTAAAGATGGCAACAGTCCCCCCCCCCCCGAGCCTCGCGTATATCGGCCTAGAGTGATAGTCGCGCTAGGGACGTGGATTAGCAATCGCAGTGACGTCGGGCAGCTTACTATAAATG

Error Probability: 100.0, Description: Eukaryotic promotor recognition motif
[https://doi.org/10.1016/0022-2836\(90\)90223-9](https://doi.org/10.1016/0022-2836(90)90223-9)

GC-Content:

GTGGAGTACACTGTAAAGATGGCAACAGTCCCCCCCCCCCCGAGCCTCGCGTATATCGGCCTAGAGTGATAGTCGCGCTAGGGACGTGGATTAGCAATCGCAGTGACGTCGGGCAGCTTACTATAAATG

Kmer:

GTGGAGTACACTGTAAAGATGGCAACAGTCCCCCCCCCCCCGAGCCTCGCGTATATCGGCCTAGAGTGATAGTCGCGCTAGGGACGTGGATTAGCAATCGCAGTGACGTCGGGCAGCTTACTATAAATG

Homopolymers:

GTGGAGTACACTGTAAAGATGGCAACAGTCCCCCCCCCCCCGAGCCTCGCGTATATCGGCCTAGAGTGATAGTCGCGCTAGGGACGTGGATTAGCAATCGCAGTGACGTCGGGCAGCTTACTATAAATG

Error simulation



Error Simulation

Advanced Error-Simulation Settings

Synthesis Method

ErrASE

▼

PCR Cycles

30

▼

Months of storage to be simulated

24

▼

Sequencing Method

Paired End

▼

PCR Polymerase

Taq

▼

Storage Host

Depurination at pH 8 and 293.15K

▼

- The default execution order for the error simulation is
Synthesis → Storage → PCR → Sequencing

Literature error rates

Synthesis	PCR	Storage	Sequencing
Column synthesized oligos	Taq	<i>in-vitro</i> (Depurination)	Illumina
Microarray based oligo pools	Pfu	<i>in-vivo</i> (<i>E. coli</i> , <i>S. cerevisiae</i> , <i>D. melanogaster</i> , <i>M. musculus</i> , <i>H. sapiens</i>)	Nanopore
	Phusion	Evolutionary models (Jukes-Cantor, Kimura)	PacBio
	Pwo	Channel models (Erasure, WGN)	

Custom simulation order



Advanced Error-Simulation Settings



Synthesis Methods

Column Synthesized Oligos
ErrASE
MutS
Consensus Shuffle
Microarray based Oligo Pools
Oligo Hybridization based error correction
High-temperature ligation/hybridization based error correction

Sequencing Methods

Illumina
Single End
Paired End
Nanopore
1D
2D
None

Storage Methods

H sapiens
M musculus
D melanogaster
S cerevisiae
In-vitro
Erasure Channel with an error probability of 0.5 percent
White Gaussian Noise with an error probability of 0.5 percent

PCR Methods

None
None
Polymerases
Phusion
Taq
Pwo
Pfu

Month (Storage)

120

Cycles (PCR)

30

Execution Order

Synthesis
MutS
Storage/PCR
Taq (30 cycle(s))
M musculus (12 month(s))
H sapiens (120 month(s))
Sequencing
Paired End

Philipps



Universität
Marburg

JUSTUS-LIEBIG-



UNIVERSITÄT
GIESSEN



LOEWE

Exzellente Forschung für
Hessens Zukunft

Error simulation



Advanced Error-Simulation Settings



Synthesis Method

ErrASE

PCR Cycles

30

Months of storage to be simulated

240

Sequencing Method

2D

PCR Polymerase

Taq

Storage Host

Depurination at pH 7 and 293.15K

Fully modified Sequence:

AACTTGCCCTGGTTGAGGATTATGGAACCCACCCAGCCTCGGTCATCAGCCACTAATGTCAGCAATAAGAGTTGCGGAGTCCTATGGCAGCCTATCGACCATGGTTATAAACCGATGGC

Was T, is now AT, Error Source: sequencing, Error Type: insertion

GC-Content: 50.83% Tm: 78.94°C Start-Pos: 0 End-Pos: 130

Download fully modified sequence as FastQ

Philipps



Universität
Marburg

JUSTUS-LIEBIG-



UNIVERSITÄT
GIESSEN



LOEWE

Exzellente Forschung für
Hessens Zukunft

Error simulation

Fully modified Sequence:













AACTTGCCCTGGTTGAGGATATGGAACCCCTCCAGCCTCGGTCATCAGCCACTAATGTCAGCAATAAGAGTTGCGGAGTCCTATGGCAGCCTATCGACCATGGTTATAAACCGATGGC

Was T, is now AT, Error Source: sequencing, Error Type: insertion

GC-Content: 50.83% Tm: 78.94°C Start-Pos: 0 End-Pos: 130

[Download fully modified sequence as FastQ](#)

- Errors are color coded by type and source

Color	Error source	Error type
	Synthesis	Insertion
	Synthesis	Deletion
	Synthesis	Mismatch/Substitution
	Storage	Insertion
	Storage	Deletion
	Storage	Mismatch/Substitution
	Sequencing	Insertion
	Sequencing	Deletion
	Sequencing	Mismatch/Substitution
	PCR	Insertion
	PCR	Deletion
	PCR	Mismatch/Substitution

Error simulation



Advanced Error-Simulation Settings



Synthesis Method	Oligo Hybridization based error correction
PCR Cycles	30
Months of storage to be simulated	240
Sequencing Method	2D
PCR Polymerase	Taq
Storage Host	White Gaussian Noise with an error probability of 0.5 percent

Fully modified Sequence:

A AATGAGTGGTCAATTGTATGGCCTCGACCTGCTGGACACAGTGCCTCGTTGCCCTGCTCTCTAAAGTGGGGTCTCTCGGCTCACCCTTGCCGAGGCTCCGATTGGAGCAACCAATTGTGTGGTCC

GC-Content: 0% Tm: Selected

Was TAC, is now TGC, Error Source: sequencing, Error Type: pattern_mismatch
Was A, is now T, Error Source: storage, Error Type: pattern_mismatch

Download fully modified sequence as FastQ

Secondary structure prediction



Secondary Structure prediction

Max. Expect



Temperature (°K)

310.15



Max-Expect dot-Sequence:

(((((.....)))))..(((.....))))......(((.....)))

Download as:

SVG-Image

PS-Image

PDF-Image

CT-File

DOT-File

PFS-Distribution-File

Philipps



Universität
Marburg

JUSTUS-LIEBIG-



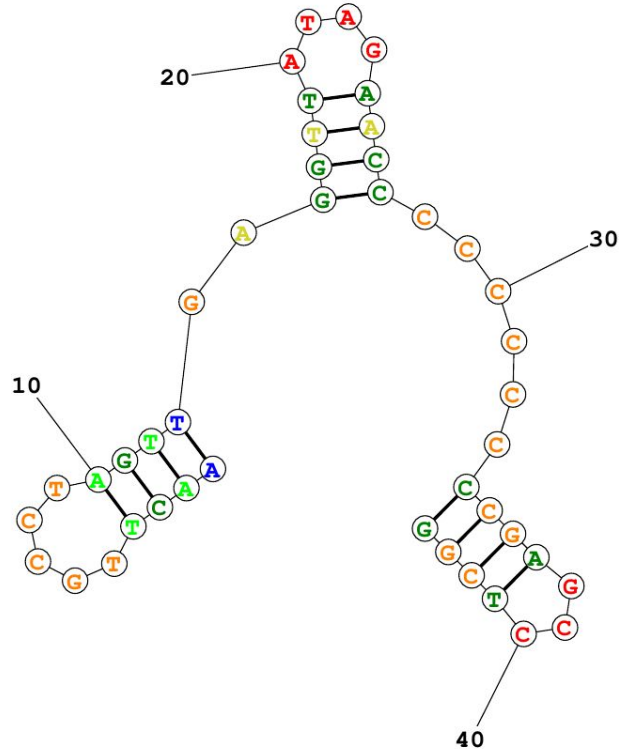
UNIVERSITÄT
GIESSEN



LOEWE

Exzellente Forschung für
Hessens Zukunft

Secondary structure prediction



Probability $\geq 99\%$

99% > Probability >= 95%

95% > Probability >= 90%

90% > Probability >= 80%

80% > Probability >= 70%

70% > Probability >= 60%

60% > Probability >= 50%

50% > Probability

ENERGY = 3.9

- Supports FASTA files as input (using the upload FASTA button or by drag and drop)
- The results of the evaluation can be returned as pseudo-FASTQ, using the error probabilities as quality scores
- The results of the error simulation can also be downloaded in FASTQ format
- The pseudo-random elements of the simulation are seeded

MESA

Part 2: customization, management and interfacing

Michael Schwarz, M. Sc.

21.04.22

Real-time GC-content and T_m calculation

Quick analysis of the GC-content or T_m of a subsequence:

- Selected subsequence will be analyzed on the fly
- Melting point calculation adopted from *Primer3* and

[Rychlik W, Spencer WJ and Rhoads RE (1990) "Optimization of the annealing temperature for DNA amplification in vitro", Nucleic Acids Res 18:6409-12]
[Breslauer KJ, Frank R, Blöcker H and Marky LA (1986) "Predicting DNA duplex stability from the base sequence" Proc Natl Acad Sci 83:4746-50]

- Default settings from *Primer3* 2.3.6:
 - strand concentration: 250e^{-9}
 - Na⁺/K⁺ concentration: 50e^{-3}
 - Mg: 0
 - dNTP: 0

Fully modified Sequence:

AGATGAGTGACCGCGCCCTAGTAGTATTATATAATTAGCCGCGCGATGCC

GC-Content: 23.53% T_m: 35.81°C Start-Pos: 23 End-Pos: 40

[Download fully modified sequence as FastQ](#)

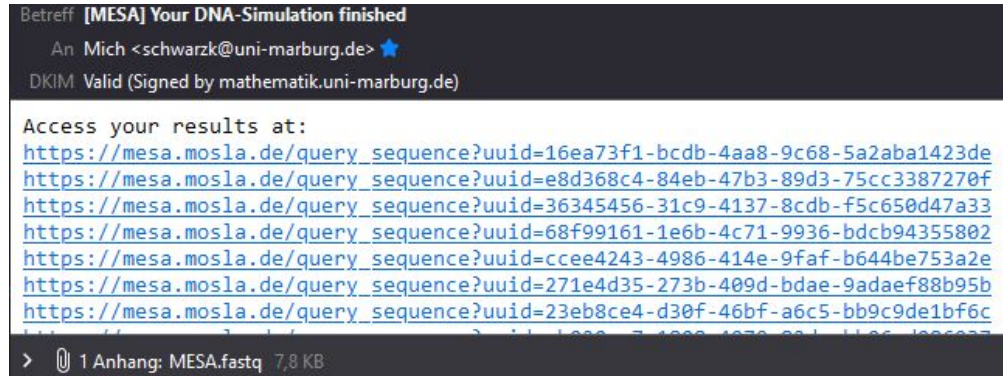
E-Mail notifications

Send E-Mail ☒

Submit

Email

- used for long-running tasks
- automatically enabled for fasta files
 - fastq file containing all modified sequences



FASTA processing



Batch-processing of sequence possible

- Input via .fasta-file
- Uploading via:
 - file picker
 - drag & drop

FASTA file loaded. Max. Expect is unchecked now and the results will be sent to your E-Mail

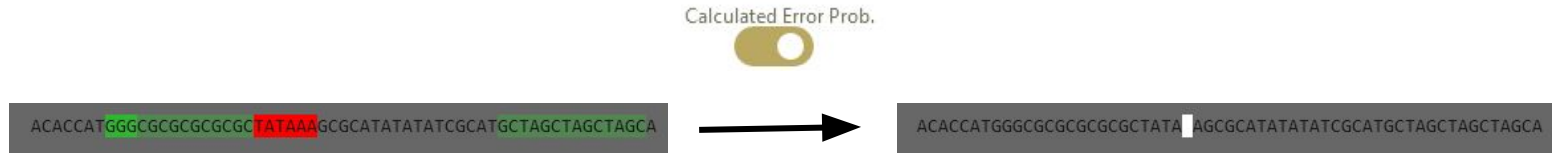
Your Query:

Sequence to analyse

FASTA FILE LOADED. YOUR RESULTS WILL BE SENT TO YOUR E-MAIL

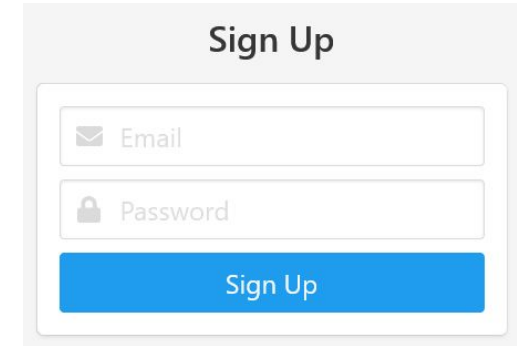
Error detection based Simulation

- By default the two steps are separated
 - Error detection used to ensure sequence adheres to minimal requirements (e.g. to be able to synthesize)
 - Hard to simulate DNA storage cycle for sequences that can not be synthesized
- “Calculated Error Prob.” allows the simulation to use the computed error probabilities for error simulation
 - Limits configuration to GC, k-mer and homopolymer graphs!



Customization of settings and rules

- Existing rules might not be sufficient for desired experiment
- New technology with different requirements emerge
- Creating an account allows:
 - persistent customization of rules
 - access to experiment history
 - E-mail notifications and fastq results via mail*
 - API usage



Sign Up

Email

Password

Sign Up

Changing existing rules

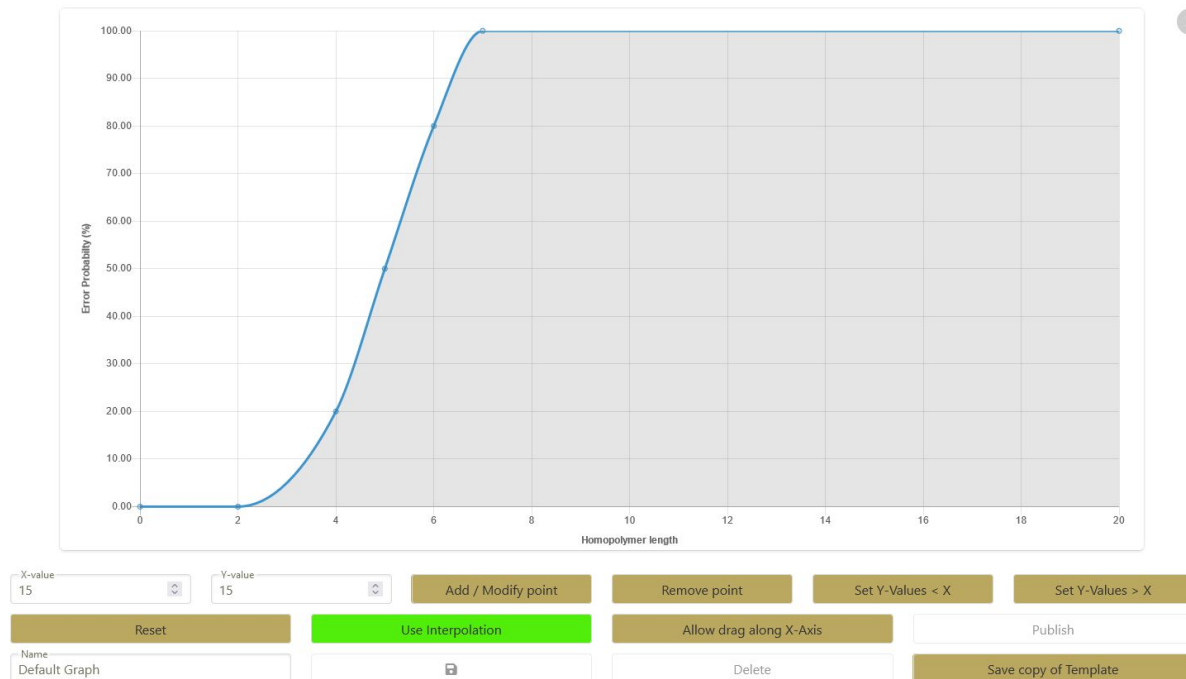
Graph based configuration:

- GC
- K-mer
- Homopolymer

Can be modified at the
“Simulate”-page

Registered users:

- copy existing config
- save and publish changes



Changing existing rules

Sequence TATTGAAGCATATATTATGGTATATGCTTGCCAT	Error Probability 100.0	Description reversed - lox1R https;	Create copy
Sequence AAAAAAAACGGTGTGGGGTGTGTTT	Error Probability 55.0000000000	Description test	Publish  Delete
Sequence Undesired Subsequence	Error Probability 50,0	Description Description	Add

Changing existing rules - Simulation Settings

Add new Rule

Description: Raw Error Rate:

Distribution

Deletion: Insertion: Mismatch:

Deletion

A: C: G: T:

Homopolymer: Random:

Insertion

A: C: G: T:

Homopolymer: Random:

Mismatch

Original DNA-Sequence: # possible Mismatches:

☐ Start Position: End Position:

Mismatch 0: Mismatch 1: Mismatch 2:

Mismatch 0: Mismatch 1: Mismatch 2:

DNA-Sequence: # possible Mismatches:

Changing existing rules - Simulation Settings

“Storage” setting allows calculation of the depurination rate based on pH and temperature

- Can be used as an estimation for the error rate
- Error rate should be treated as “per base per month”

The following calculation is based on “Non-Enzymatic Depurination of NucleicAcids: Factors and Mechanisms” (<https://doi.org/10.1371/journal.pone.0115950>)

With T = absolute Temperature (°K) and k = the Depurination Rate per Base and per Second:

$$\text{pH} < 2.5, \lg(k) = 14.6 - 0.707 \times \text{pH} - 5.63 \times 10^3 / T$$

$$\text{pH} \geq 2.5, \lg(k) = 16.5 - 0.982 \times \text{pH} - 5.85 \times 10^3 / T$$

Calculation

pH-Value

7.0



Temperature (°K)

293



Depurination Rate

0.0120238485140381



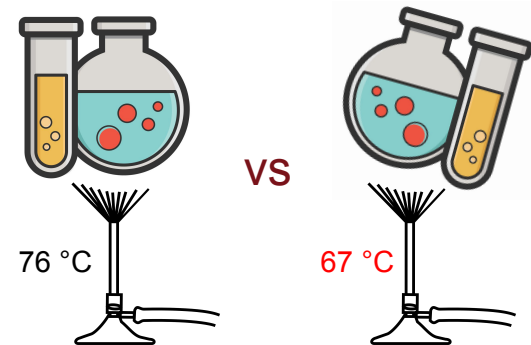
/ Base / Month

Set Error Rate

Sharing configurations

- Goal: Comparable scenarios for different experiments
 - Usual approach:
 - No information about fine-grain settings
 - Sharing test setup manually - prone to errors!
 - Better: Share config file to ensure correct settings
- Solution:
 - Pre-defined rules
 - link-sharable results
 - downloadable config
 - seed-based simulation
 - publishing of configurations

Download current config



Management - Requesting validation

- Users can request validation of created settings:

Description: Example
 Raw Error Rate: 40.0000000000000000
 Publish

ErrASE: 0.1250000000000000
 Create copy View

Description: Nuclease-based error correction
 Raw Error Rate: 0.0330000000000000
 Create copy View

Publish your custom configuration to make it available for everyone to use.
 If you just want to share with a coworker consider using "Download current config" or sharing your result using the Share-Link

Description for the Validation (this Text should help us to verify your config)
 Text with sources for validation and comments

Request validation

Description: Example
 Raw Error Rate: 40.0000000000000000
 Publish Modify Delete

- | Description | Raw Error Rate | Activity | Library | Modify | Delete |
|----------------------------------|--------------------|------------------------------|---------|--------|--------|
| Depurination at pH 8 and 193.15K | 0.0000000000000000 | Column Synthesized Oligos | | | |
| Depurination at pH 8 and 193.15K | 0.0000000000000000 | Eukaryotes | Publish | | |
| Depurination at pH 8 and 193.15K | 0.0000000000000000 | Illumina | | Modify | Delete |
| Depurination at pH 7 and 193.15K | 0.000000000000005 | In-vitro | | | |
| Depurination at pH 7 and 193.15K | 0.000000000000005 | Microarray based Oligo Pools | Publish | | Delete |
| Depurination at pH 7 and 193.15K | 0.000000000000000 | Nanopore | | Modify | Delete |
| Depurination at pH 7 and 253.15K | 0.000009000000000 | None | Publish | | Delete |
| Depurination at pH 7 and 253.15K | 0.000009000000000 | PacBio | | Modify | Delete |
| Depurination at pH 7 and 293.15K | 0.012310000000000 | Polymerases | | | |
| Depurination at pH 7 and 293.15K | 0.012310000000000 | Prokaryotes | Publish | | Delete |
| Depurination at pH 7 and 293.15K | 0.012310000000000 | User defined | | Modify | Delete |
| Example | 40 | User defined | Publish | | Delete |

Administration overview

Administration Settings

Undesired motifs

Synthesis

Sequencing

PCR

Storage

Graphs

Users

Previous Results

ID

0

E-Mail

nouser@mosla.de

Validated: ☐

Admin: ☒



Delete

ID

2

E-Mail

test@localhost

Validated: ☒

Admin: ☒



Delete

User Profile



History	Account-Management	API-Key
Previous Results:		
38b2c95e-e9cb-4310-8625-f8a041822db1	Valid until Sat Apr 8 10:39:29 2023	Set expiration Days 365 <input type="button" value="Change"/> <input type="button" value="Delete"/>
237675b4-b2e8-4e5f-8c0b-1e4b579cba9f	Valid until Sat Apr 8 10:39:12 2023	Set expiration Days 365 <input type="button" value="Change"/> <input type="button" value="Delete"/>

Profile		
History	Account-Management	API-Key
aAnv21uZzTgCMLxeZVqAVaK7fotXLFxIDz7vSfAiqas		

API-Integration



MESA exposes a REST-API:

- API description inside “swagger.json”



API includes:

- GC, Homopolymer, k-mer, subsequences
- secondary structure prediction + image retrieval
- full simulation (single sequence)
- full simulation (fasta-file as input)

→ All results as JSON

API-Integration



```
In [1]: import json, requests
```

```
In [6]: MESA_URL = 'http://127.0.0.1:5000/api/all'
API_KEY = "aAnv21uZzTgCMLxeZVqAVaK7fotXLFxIDz7vSfAiqas"
header = {'content-type': 'application/json;charset=UTF-8'}
with open ("mesa.json", "r") as f:
    config = json.load(f)
```

Download current config

```
In [10]: def requestMESA(config, seq, api_key):
    config['sequence'] = seq
    config['asHTML'] = False
    config["key"] = API_KEY
    res = requests.post(MESA_URL, data=json.dumps(config), headers=header)
    return res.json()
```

```
In [12]: requestMESA(config, "AAACGTGACTGACTAGTCGAGCGGTACGATCG", API_KEY)
```

```
Out[12]: {'AAACGTGACTGACTAGTCGAGCGGTACGATCG': {'res': {'all': [{'base': 'A',
    'endpos': 2,
    'errorprob': 0.060294117368628816,
    'identifier': 'homopolymer_0',
    'startpos': 0}],
    'dot_seq': '.....((((.....)).....',
    'fastqMod': '---IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII',
    'fastqOr': '---IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII',
    'gccontent': [],
    'homopolymer': [{'base': 'A',
    'endpos': 2,
    'errorprob': 0.060294117368628816,
    'identifier': 'homopolymer_0',
    'startpos': 0}],
    'kmer': {},
    'maxexpectid': '0039b145685c4b66bad3e76c363f7b86',
    'modified_sequence': 'AAACGTGACTGACTAGTCGAGCGGTACGATCG',
    'modify': [],
    'seed': '1362981422',
    'sequence': 'AAACGTGACTGACTAGTCGAGCGGTACGATCG',
    'uuid': 'f4322eed-323f-4ffe-bb20-227f2c2789b3']}}}
```

Setup / Installation

MESA was designed as a web application:

- Available at: <https://mesa.mosla.de>
- For small or non-critical experiments: No need for a local instance
- Large, time-consuming or sensitive experiments can be performed on a local machine:



using a docker-compose file and the existing containers,
MESA can be installed in minutes



Pull & change the source code to your needs

Initial setup

3 Steps:

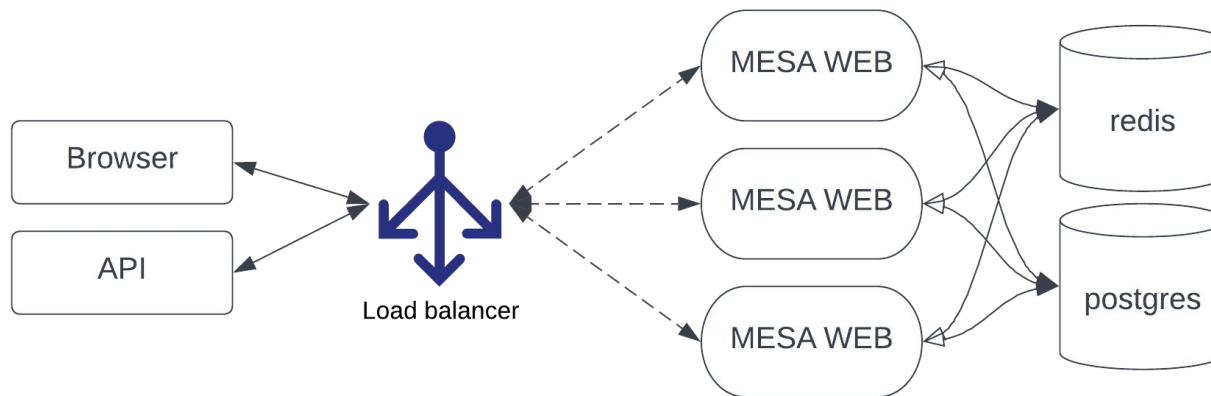
- Get MESA from GitHub
- change “docker-compose.yml” according to your needs
- run: `docker-compose up (-d)`

The docker-compose file includes:

- Mail-setup
- secret keys for Cookie and session generation
- exception logging
- self-signed and lets-encrypt based SSL
- database passwords

Load balancing

- High demand can be handled by using an additional load-balancer



Conclusion

- MESA is a system for error-detection and -simulation for DNA storage
- Key features:
 - fast, flexible and configurable
 - sharing and publishing of results and configurations
 - comparable tests
 - easy integration in existing tool-chains

Contact: {marius.welzel, peter.schwarz}@uni-marburg.de

URL: <https://mesa.mosla.de>

Thank you!

Questions?