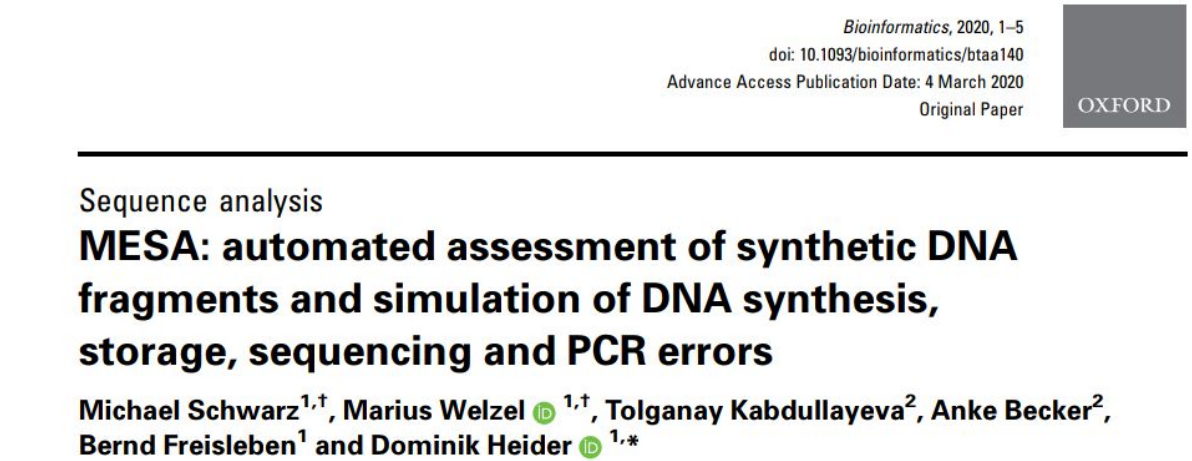
MESA: automated assessment of synthetic DNA fragments and simulation of DNA synthesis, storage, sequencing and PCR errors

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung



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 Kabdullayeva[2], Anke Becker[2], Bernd Freisleben[1] and Dominik Heider [1,*]

Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

MESA

https://mesa.mosla.de/

Part 1: overview and general usage

Marius Welzel, M. Sc.

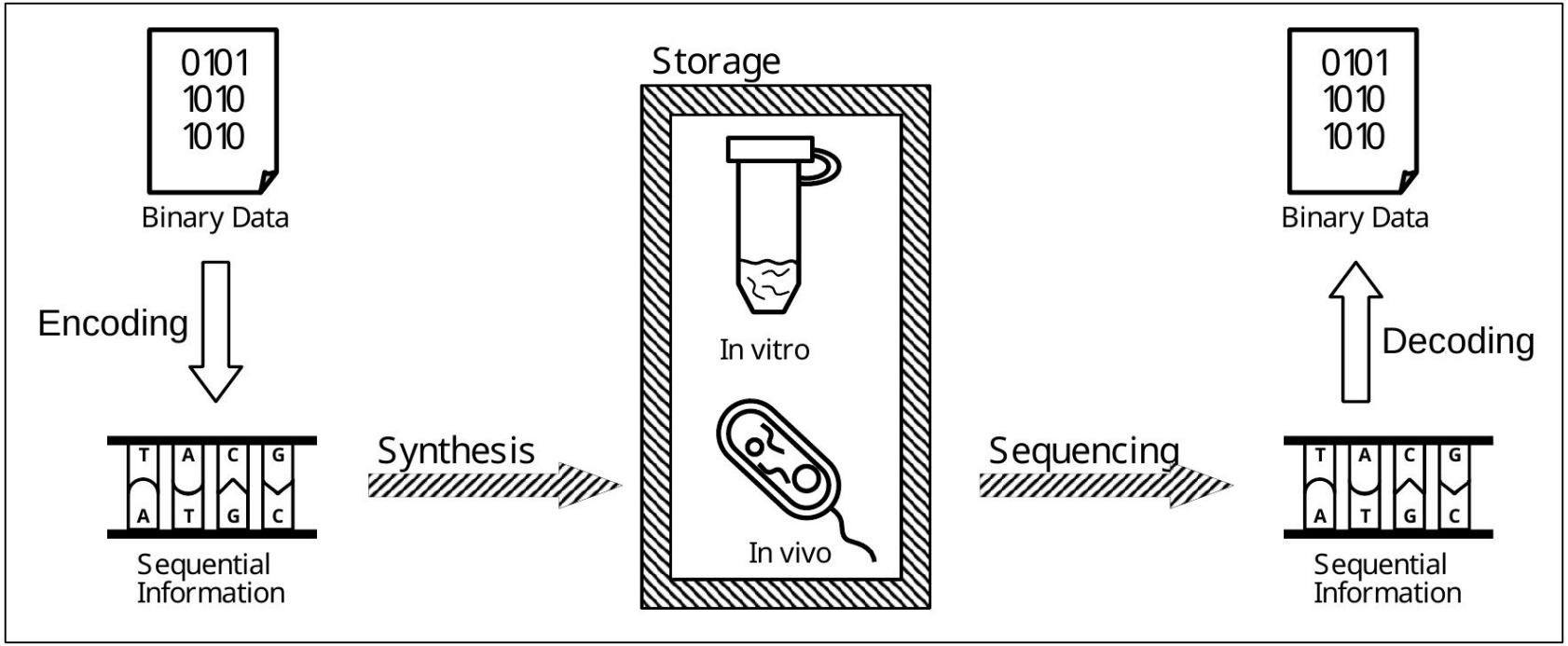
21.04.22

Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

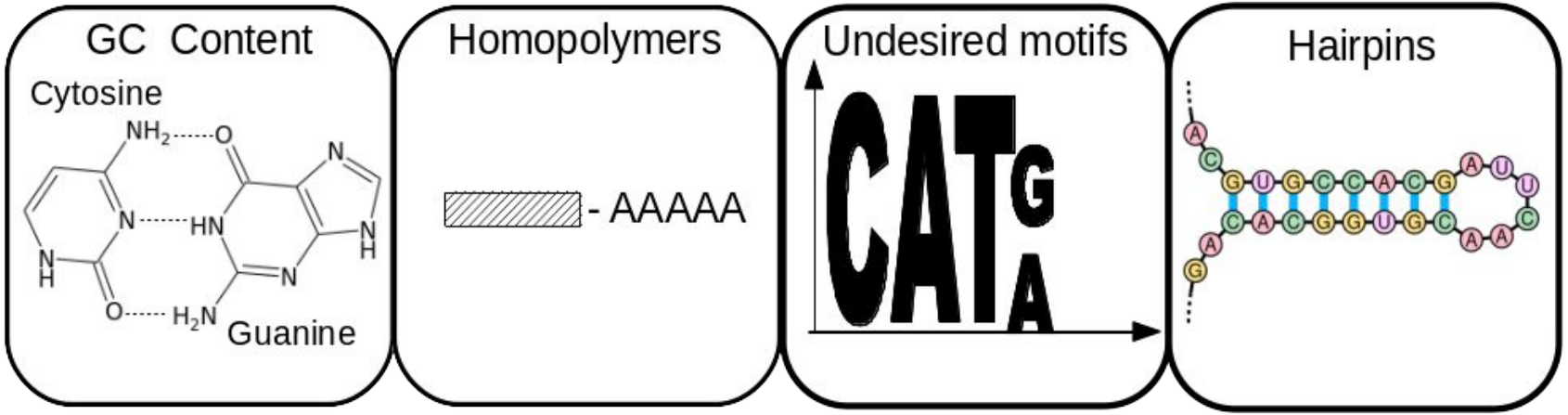
Data storage in DNA



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

Motivation: Sequence Evaluation

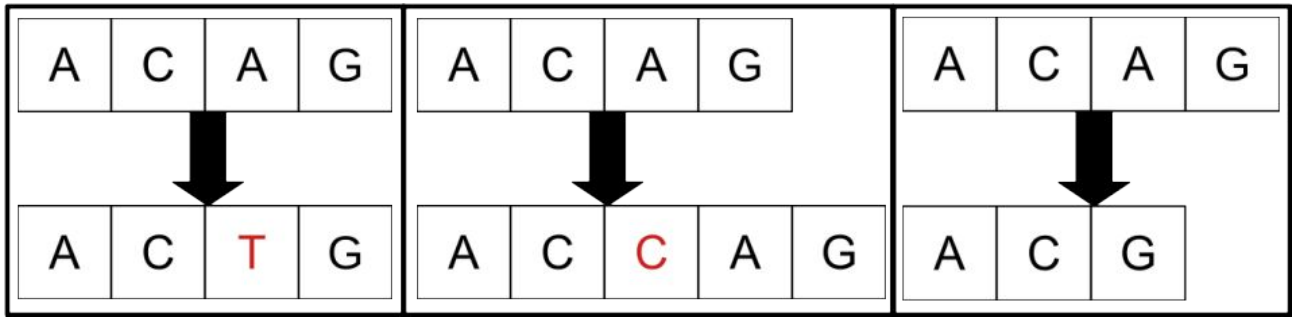
- Rapidly changing requirements warranted a customizable tool to evaluate sequences regarding constraints



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



Substitutions Insertions Deletions

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

AAACCAGT

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

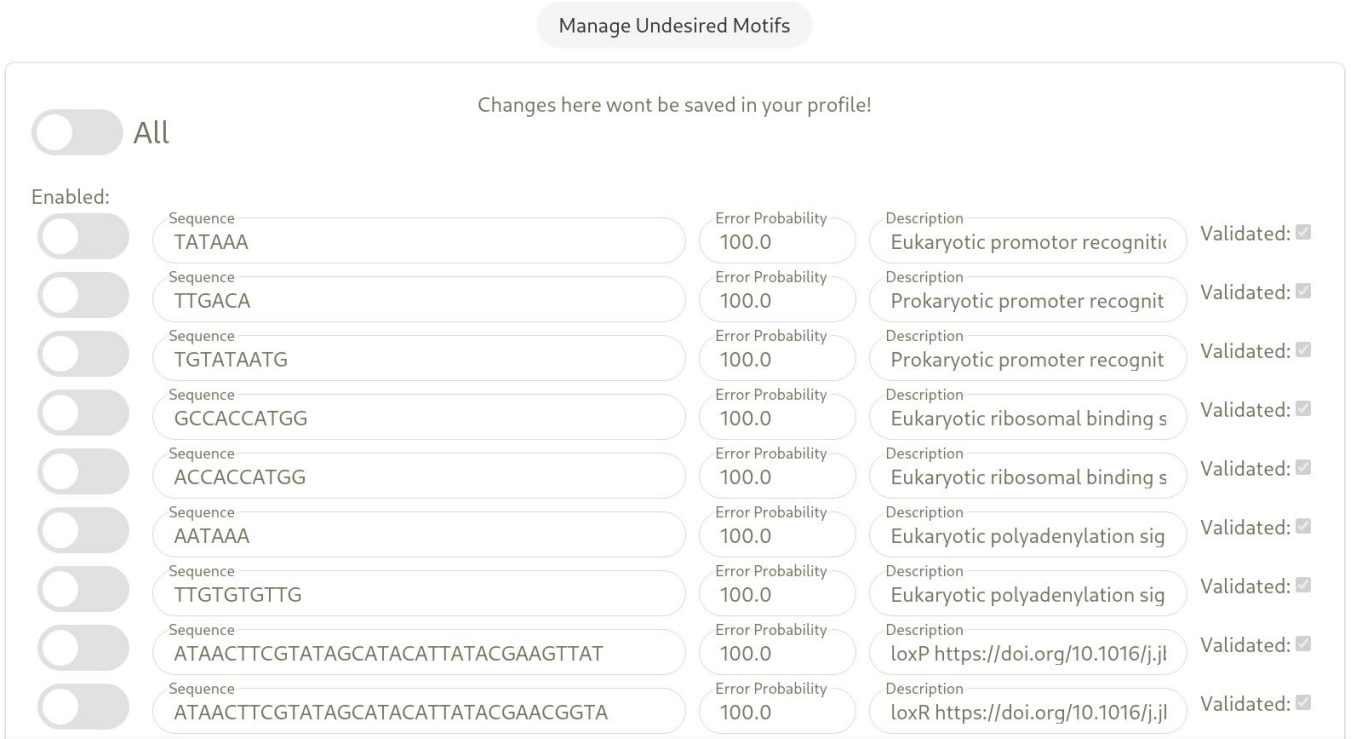
Download current config

Upload config/fasta file

Send E-Mail

Submit

Motif selection

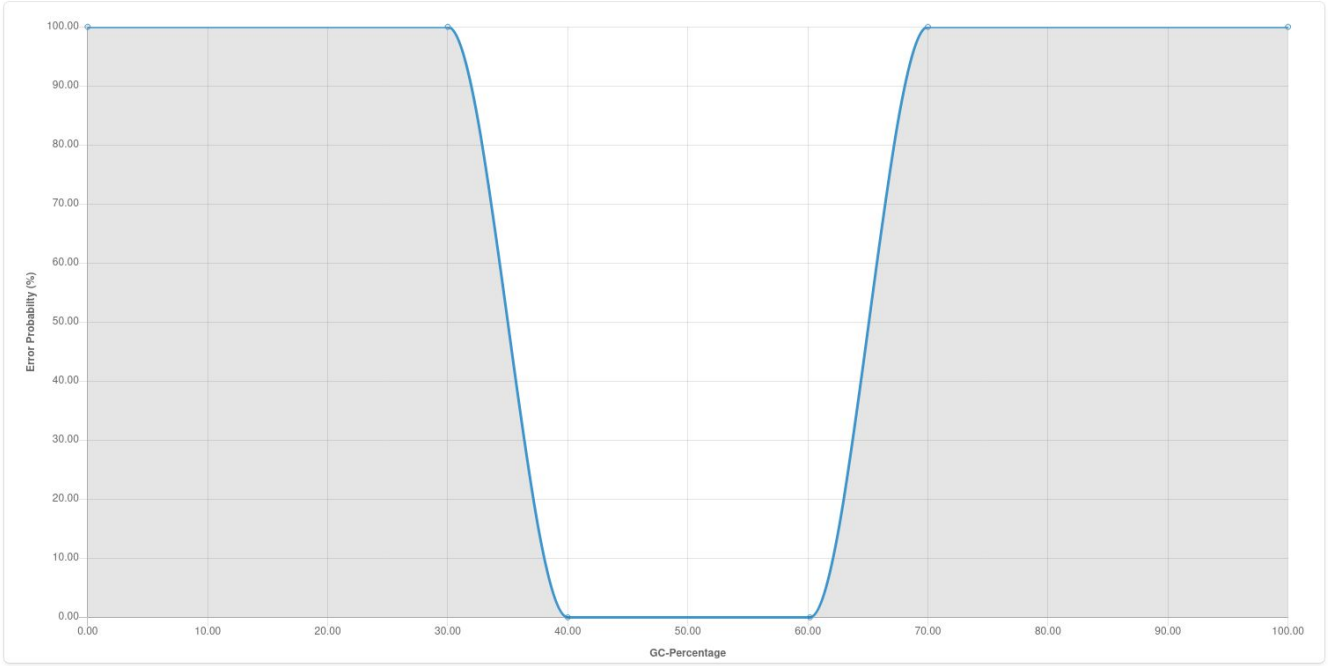


Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

GC content, homopolymer length and kmers

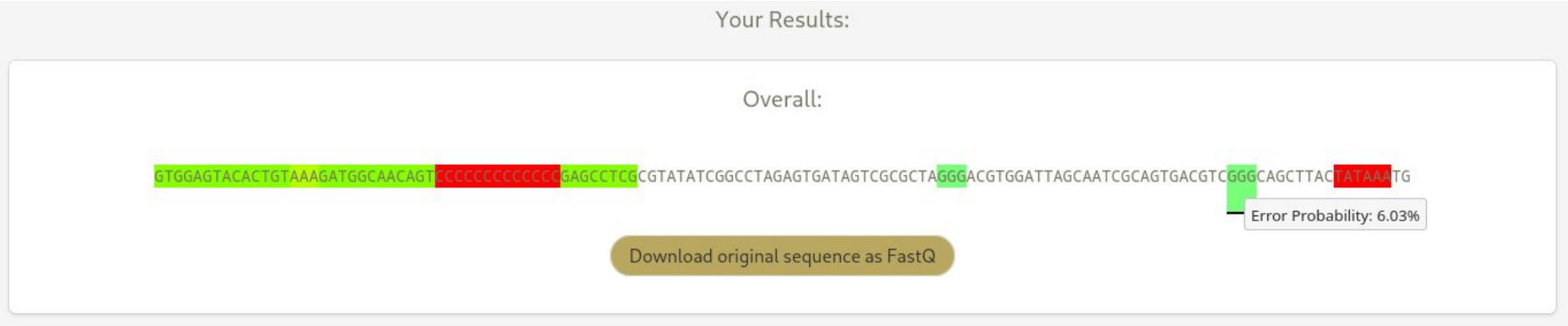


Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

Sequence evaluation



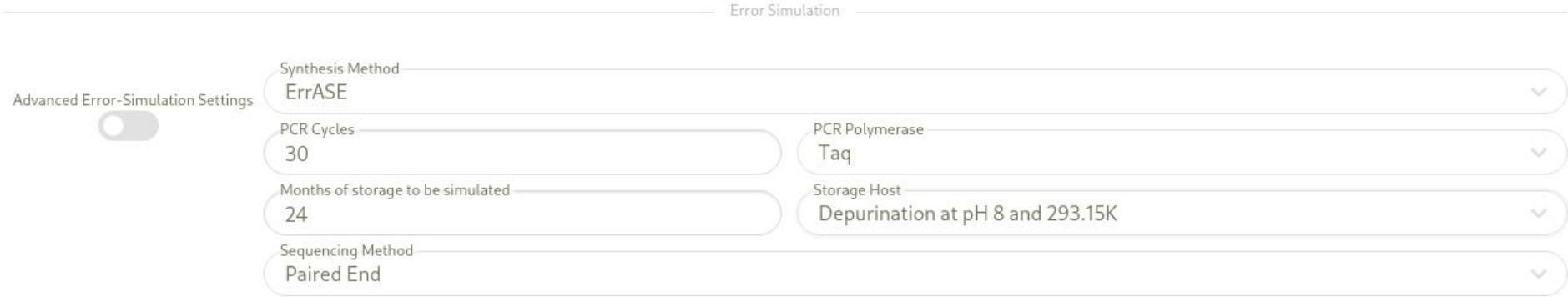
Sequence evaluation

Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

Error simulation

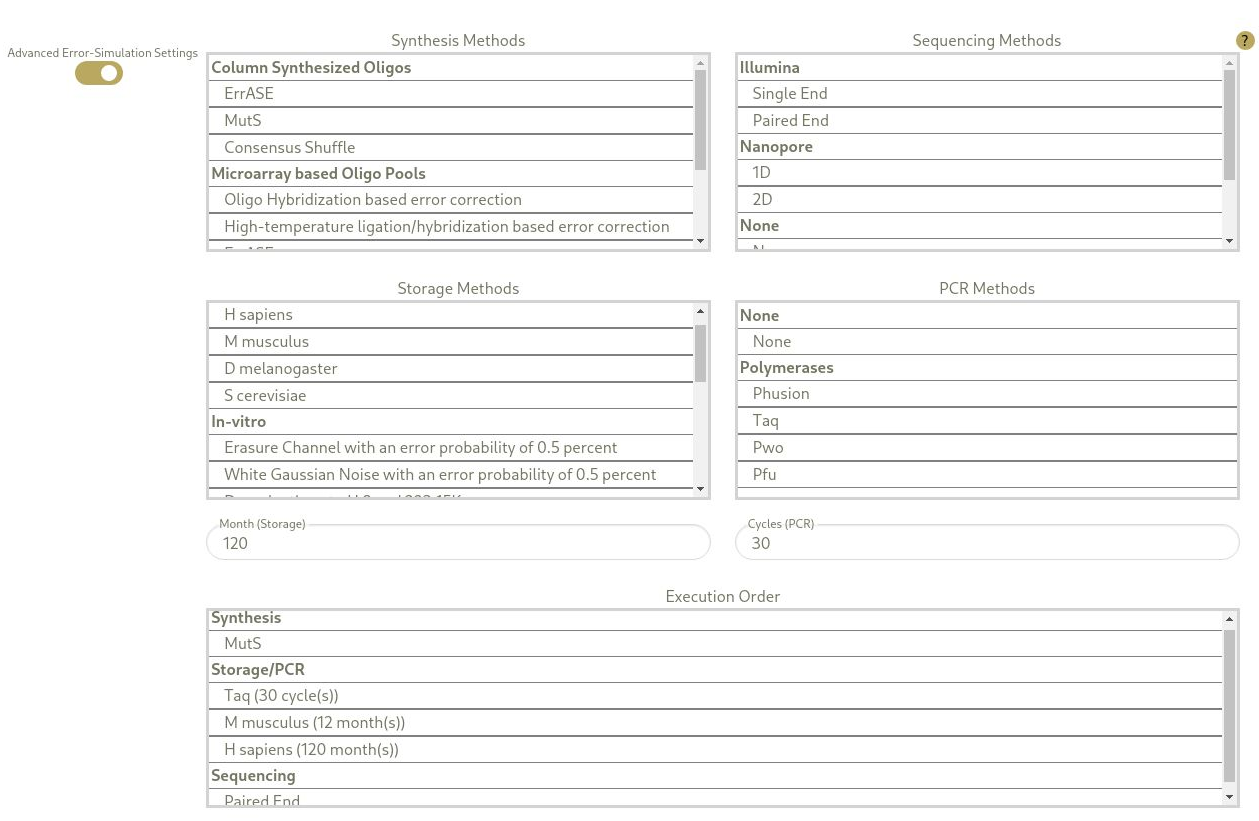


- The default execution order for the error simulation is
  Synthesis → Storage → PCR → Sequencing

Literature error rates

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

Custom simulation order

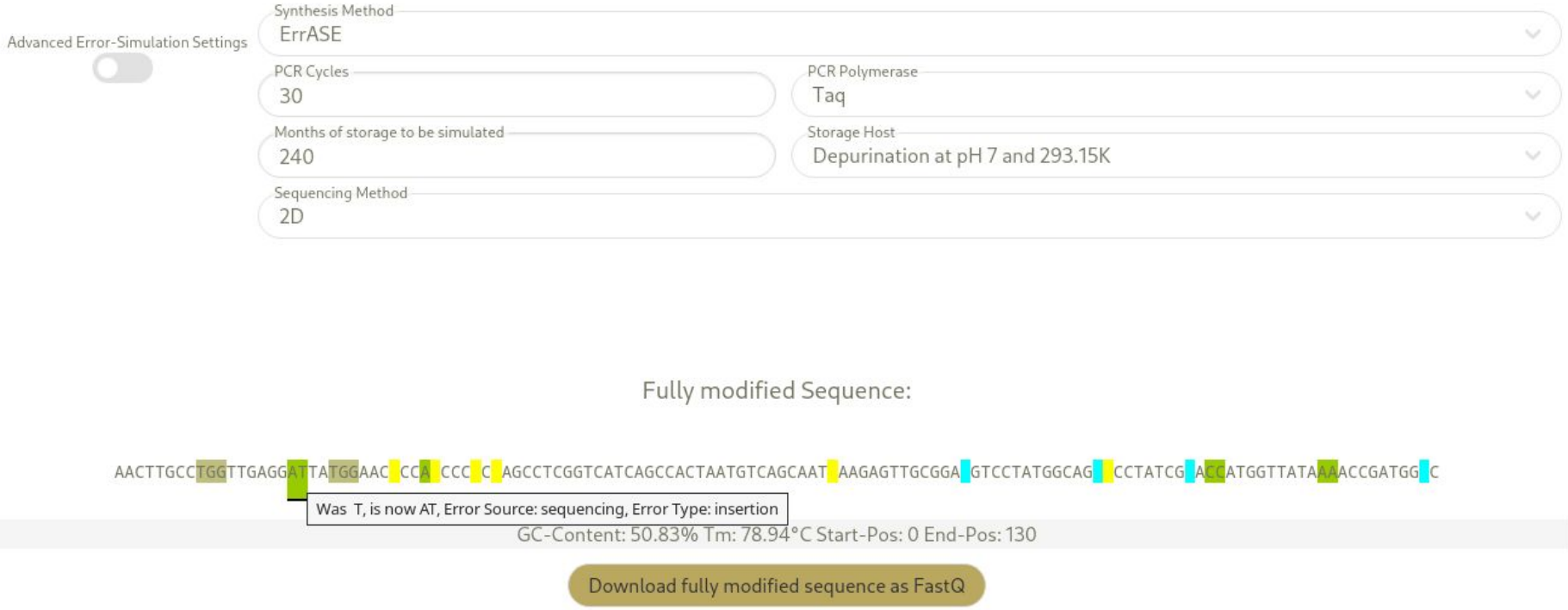


Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

Error simulation

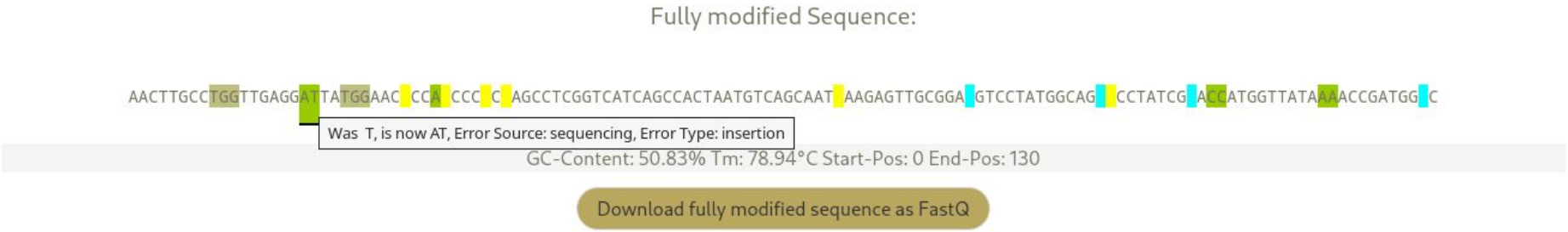


Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

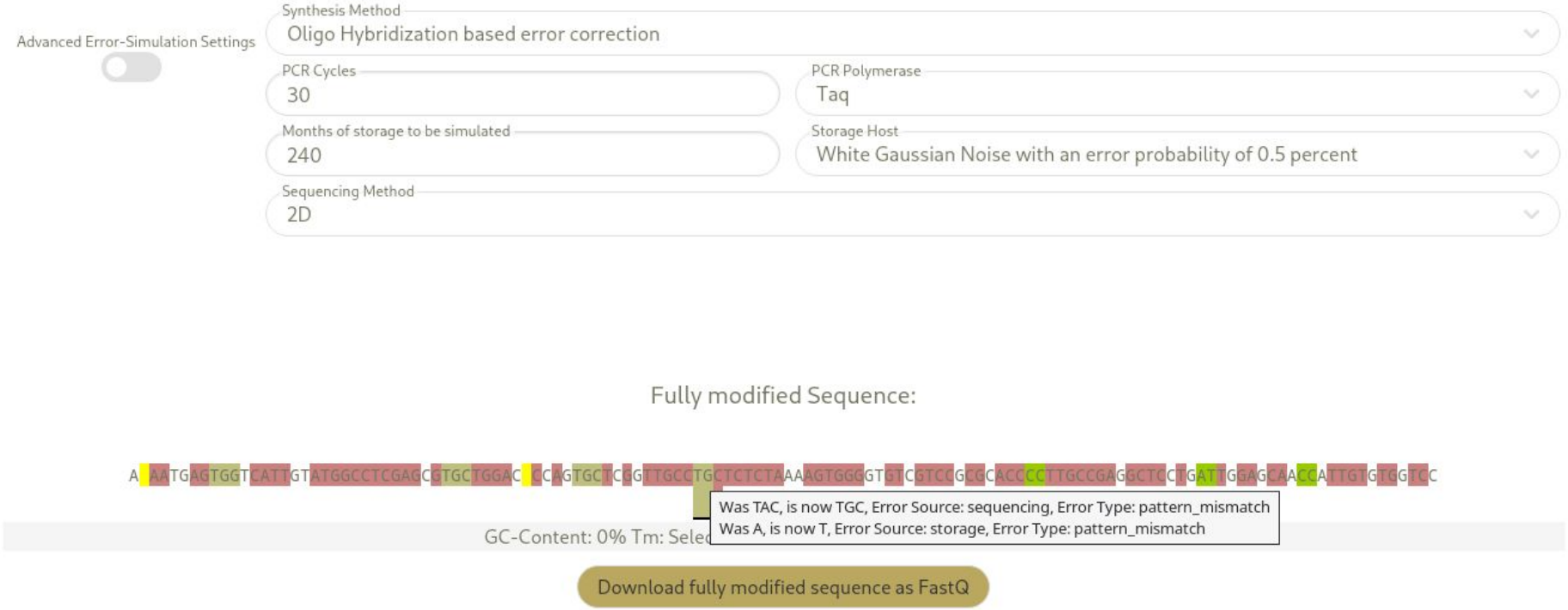
Error simulation



- 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

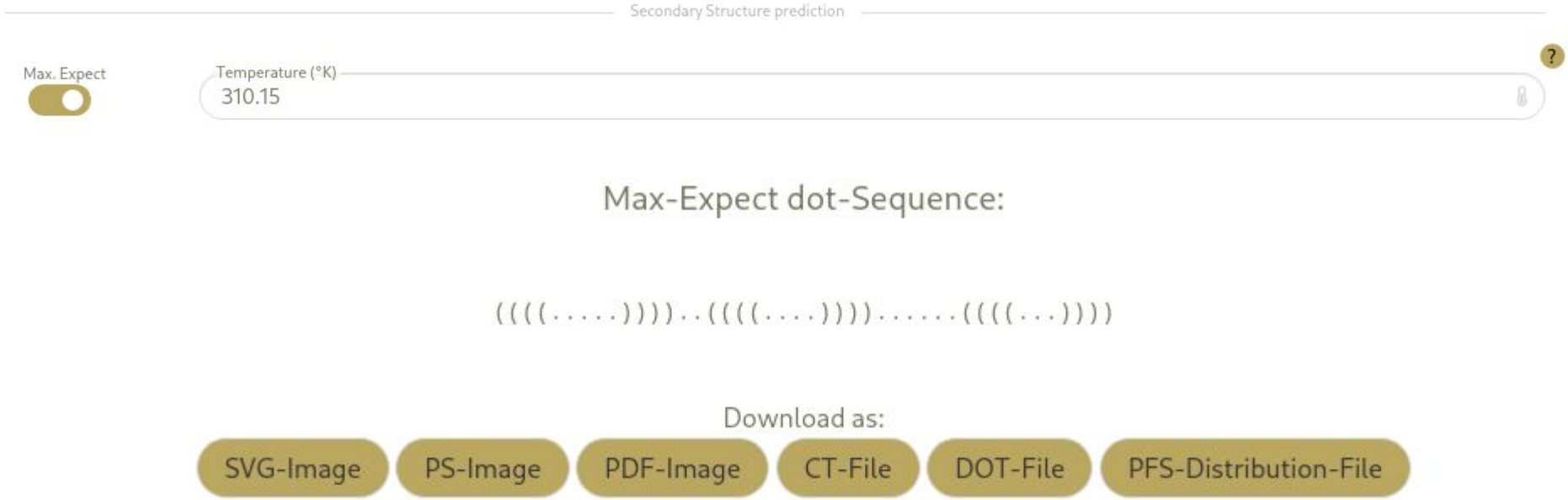


Philipps Universität Marburg

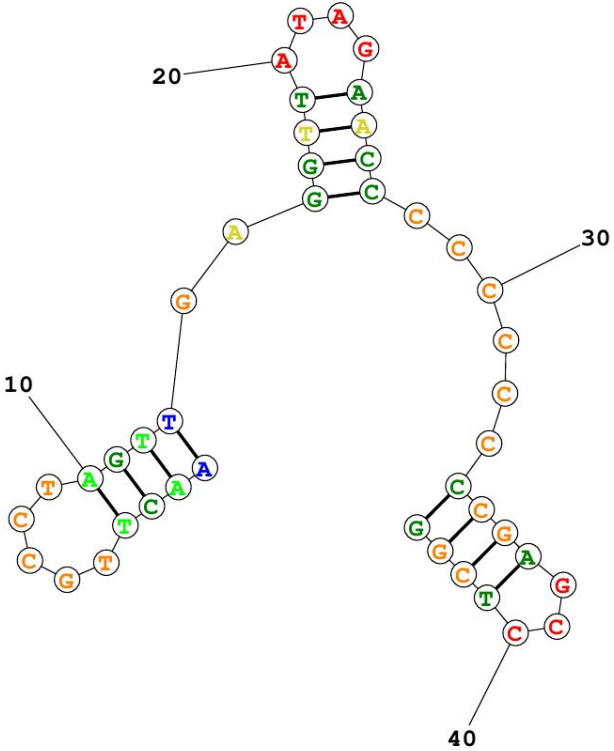
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

Secondary structure prediction



Secondary structure prediction



Probability >= 99%
99% > Probability >= 95%
95% > Probability >= 90%
90% > Probability >= 80%
80% > Probability >= 70%
70% > Probability >= 60%
60% > Probability >= 50%
50% > Probability

ENERGY = 3.9

IO

- 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

Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

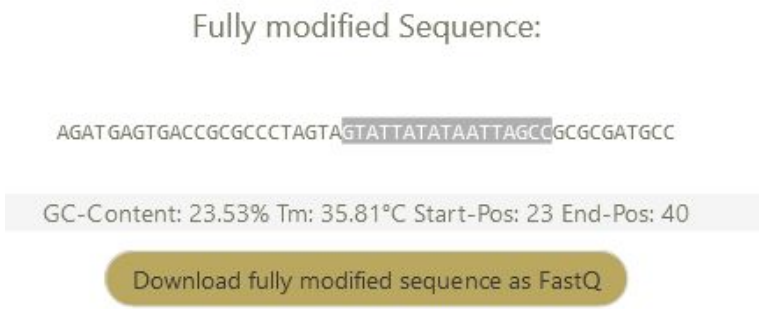
Real-time GC-content and Tm calculation

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

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

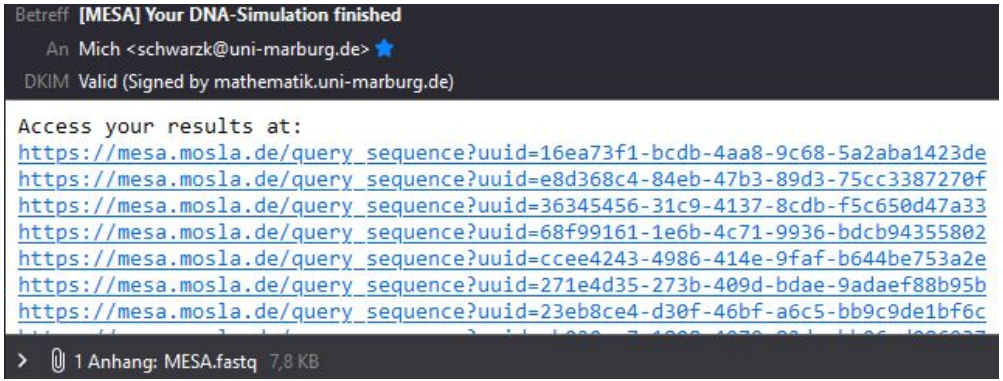
[Rychlik W, Spencer WJ and Rhoads RE (1990) "Optimization of the annealing temperature for DNA amplification in vitro", Nucleic Acids Res 18:6409-12]
[Breslauer KJ, Frank R, Blöcker H and Marky LA (1986) "Predicting DNA duplex stability from the base sequence" Proc Natl Acad Sci 83:4746-50]

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



E-Mail notifications

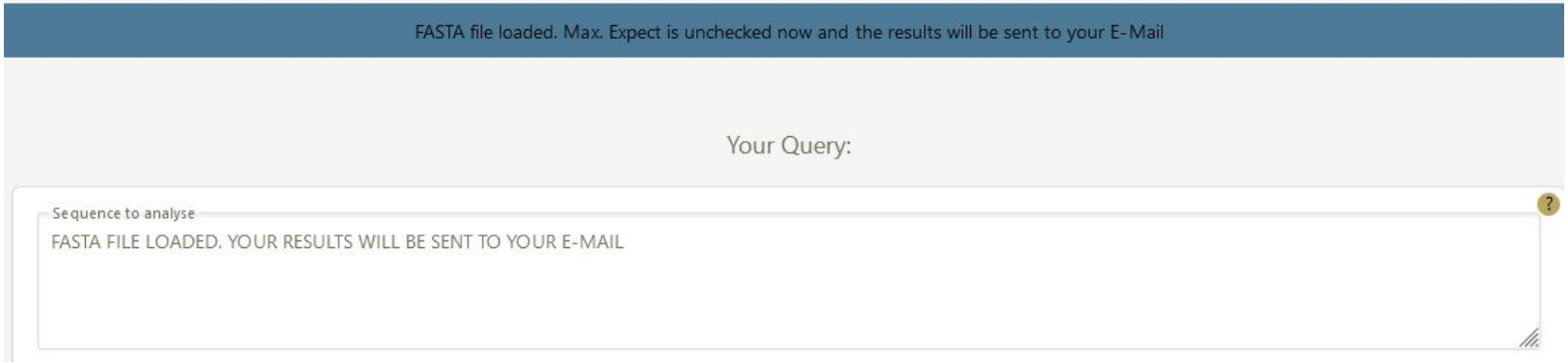
- used for long-running tasks
- automatically enabled for fasta files
  - fastq file containing all modified sequences



FASTA processing

Batch-processing of sequence possible

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



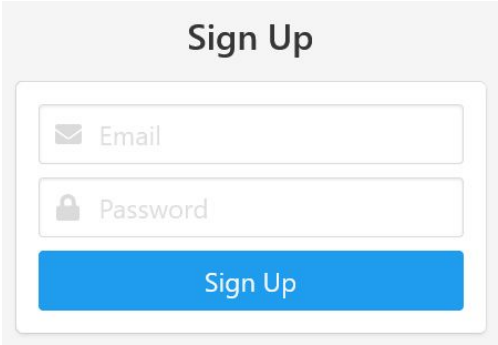
Error detection based Simulation

- 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



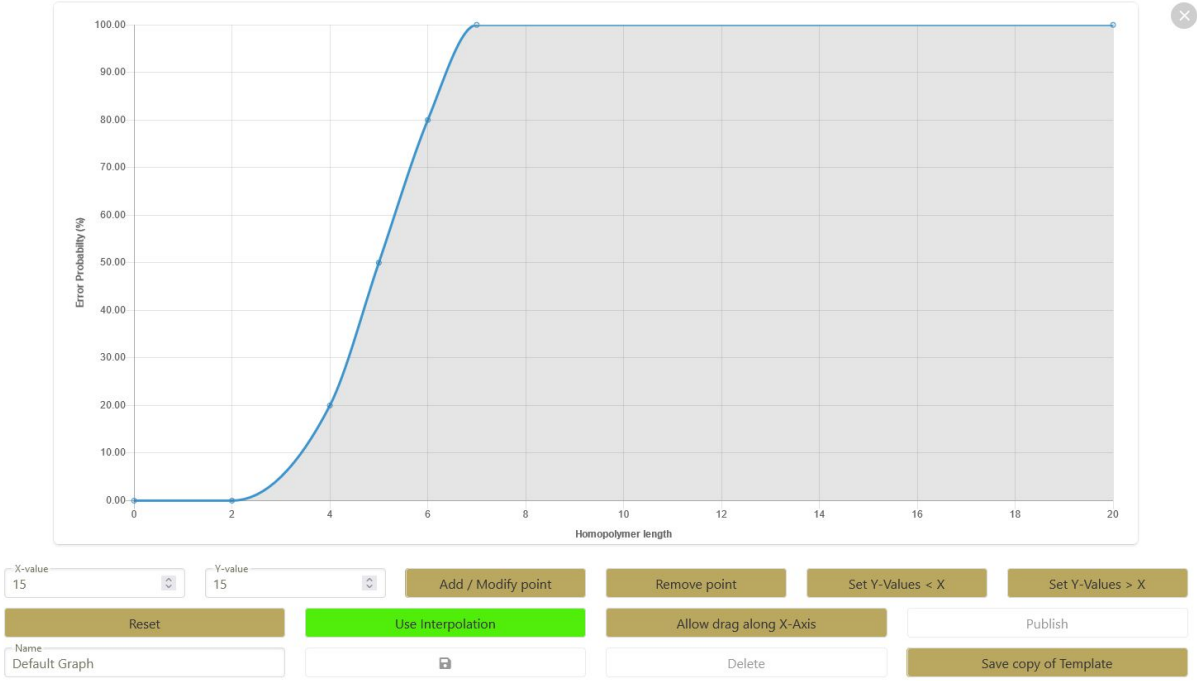
Changing existing rules

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung

Graph based configuration:
- GC
- K-mer
- Homopolymer

Can be modified at the “Simulate”-page

Registered users:
- copy existing config
- save and publish changes



Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

Changing existing rules

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung

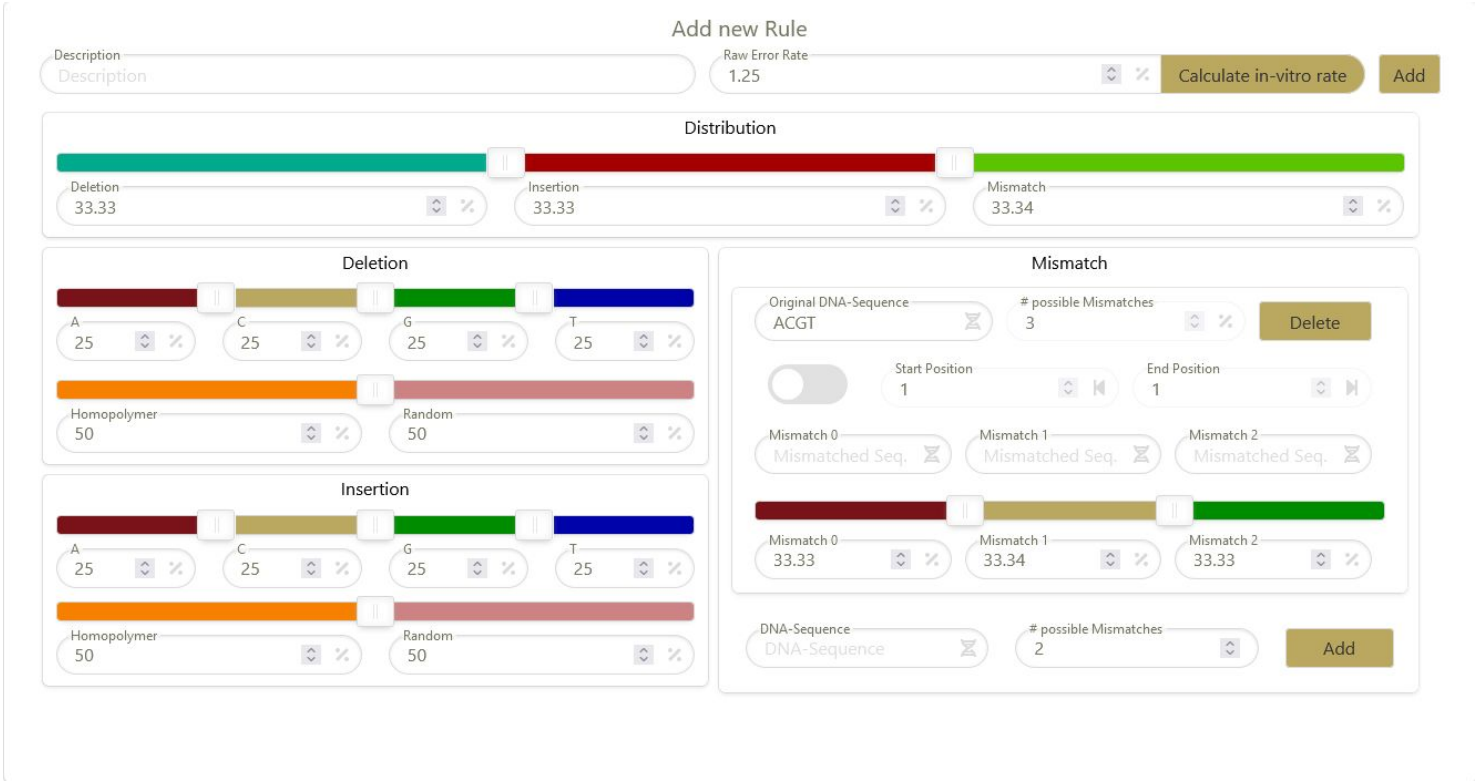
Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

Changing existing rules - Simulation Settings

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung



Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

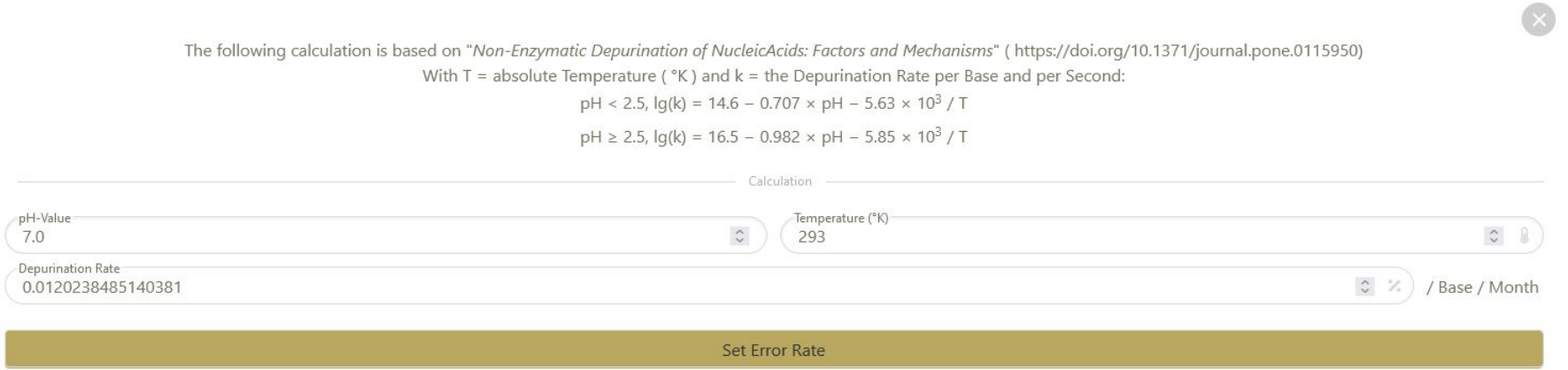
LOEWE
Exzellente Forschung für Hessens Zukunft

Changing existing rules - Simulation Settings

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung

"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"



Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

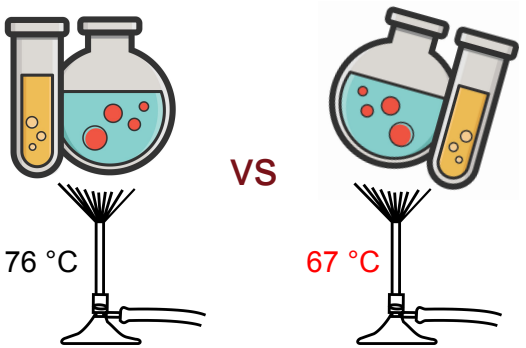
LOEWE
Exzellente Forschung für Hessens Zukunft

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

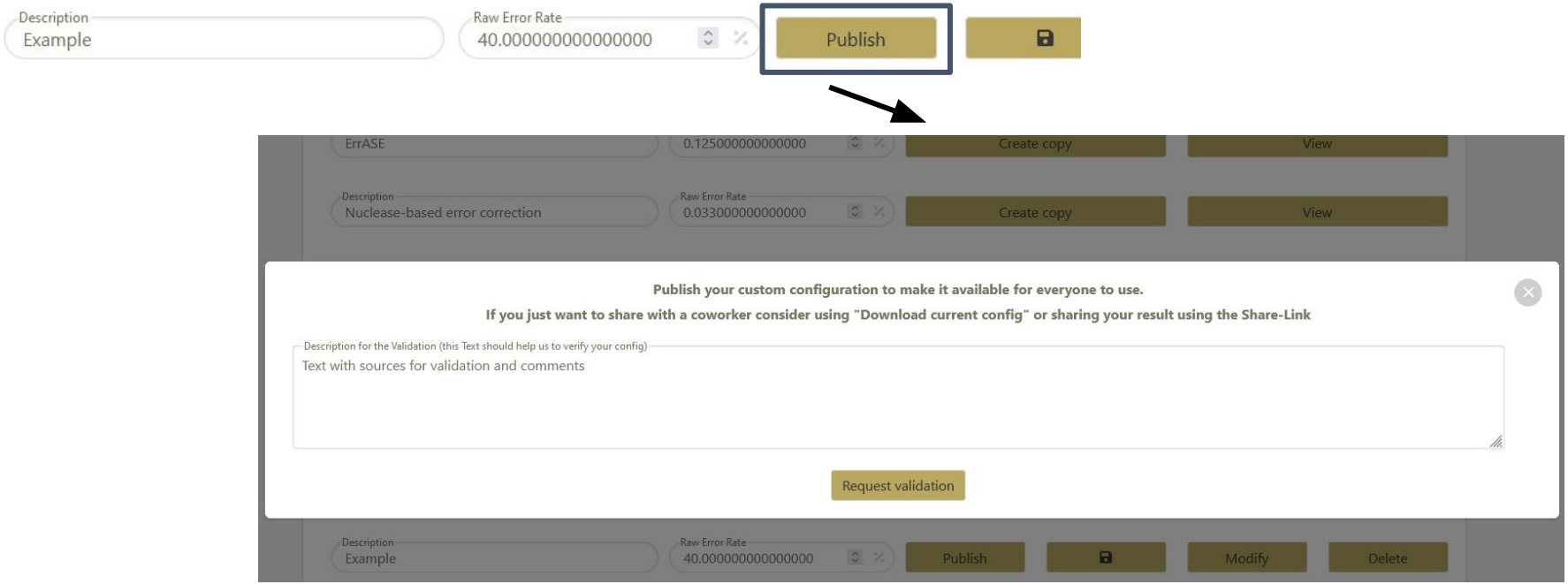
- Solution:
  - Pre-defined rules
  - link-sharable results
  - downloadable config
  - seed-based simulation
  - publishing of configurations

Download current config



Management - Requesting validation

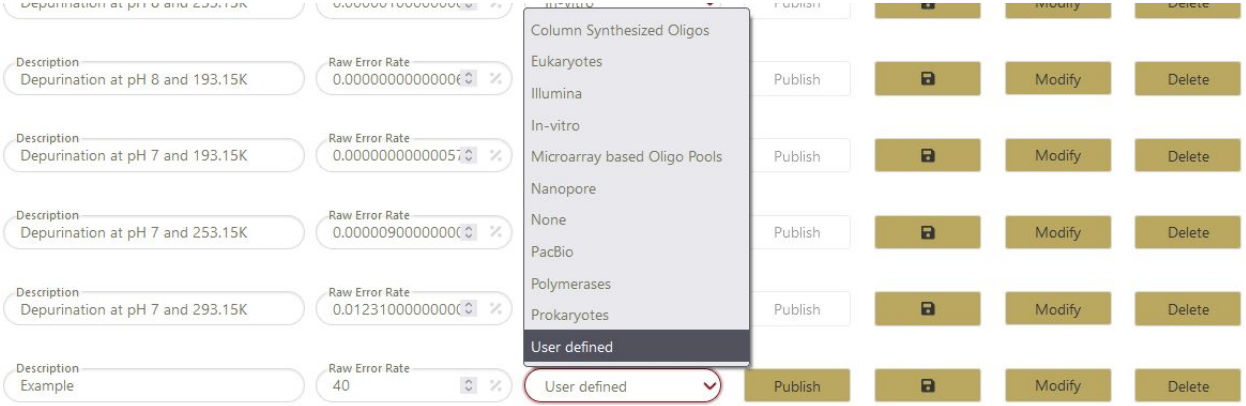
- Users can request validation of created settings:



Management - Administration

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung

- (if set up) admins get e-mail for each validation request
- Admin can review, categorize and modify setting
- “Publish” will show validation comment and allow publication
- Unpublishing performed automatically when modified



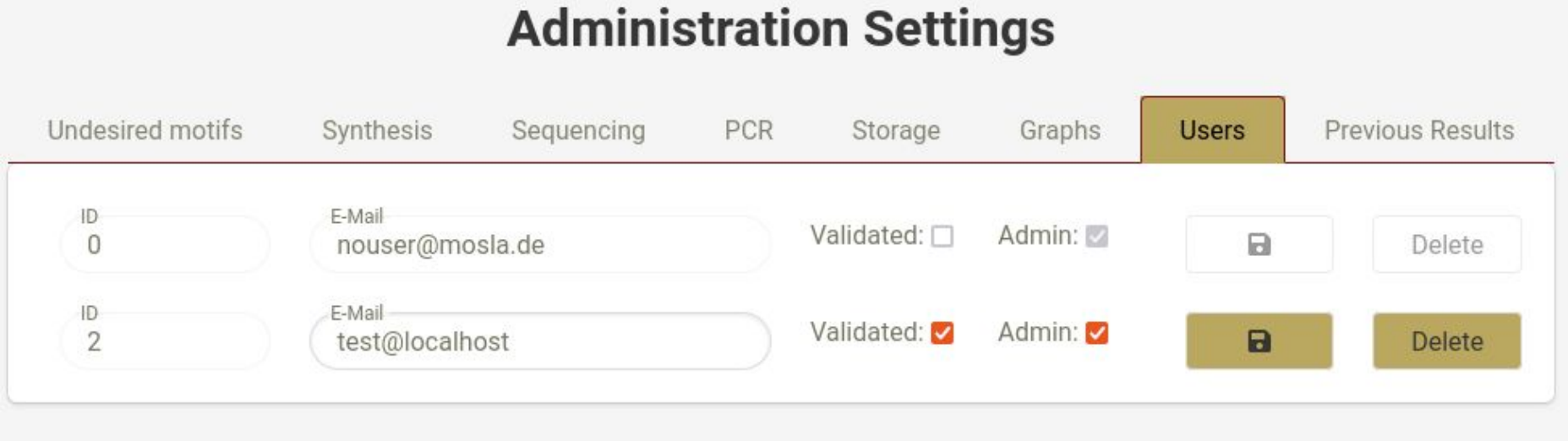
Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

Administration overview

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung

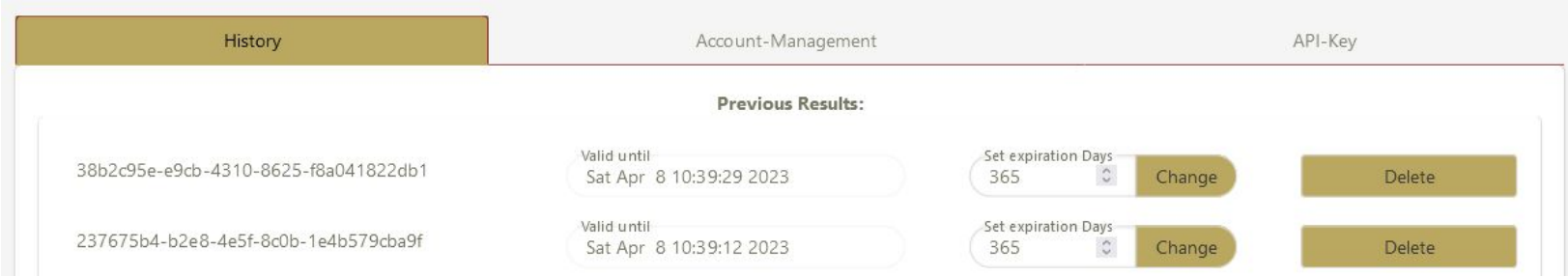


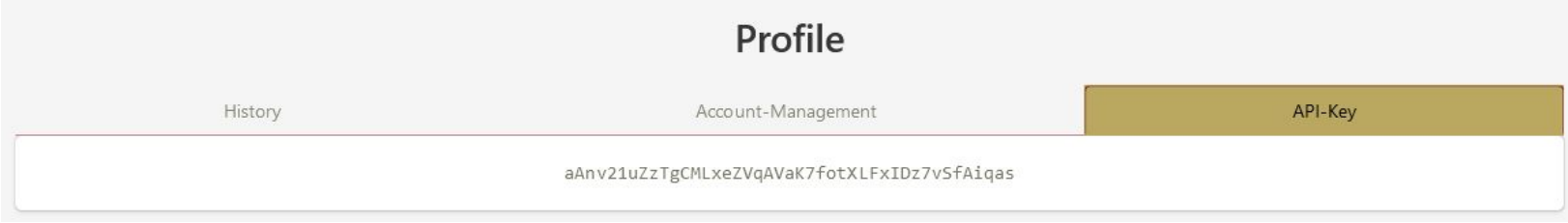
Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

User Profile





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)

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, "AAACGTGACTGACTGACTAGTCGAGCCGTACGATCG", API_KEY)

Out[12]: {'AAACGTGACTGACTGACTAGTCGAGCCGTACGATCG': {'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': 'AAACGTGACTGACTGACTAGTCGAGCCGTACGATCG',
              'modify': [],
              'seed': '1362981422',
              'sequence': 'AAACGTGACTGACTGACTAGTCGAGCCGTACGATCG',
              'uuid': 'f4322eed-323f-4ffe-bb20-227f2c2789b3'}}}
```

Download current config

Setup / Installation

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung

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:

docker

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

GitHub

Pull & change the source code to your needs

Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

Initial setup

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung

3 Steps:

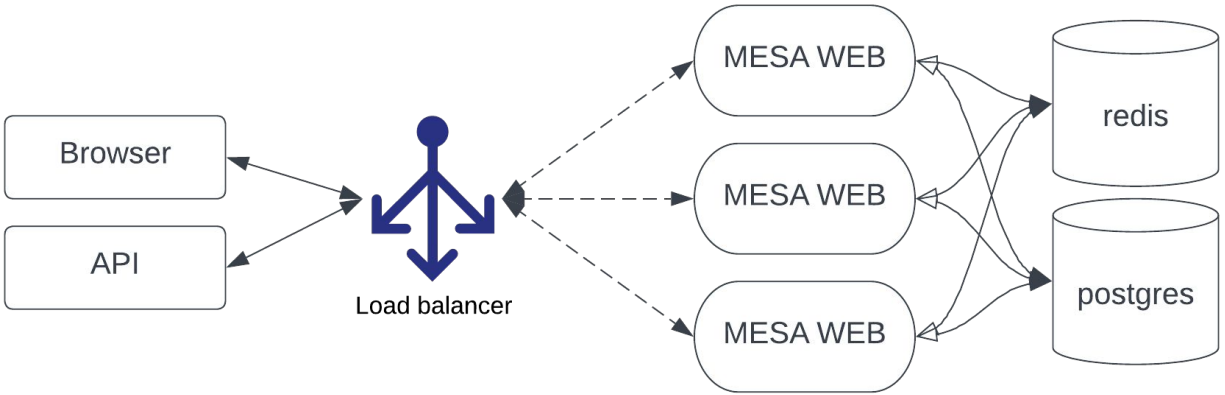
- Get MESA from GitHub
- change “docker-compose.yml” according to your needs
- run: `docker-compose up (-d)`

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

MOSLA
MOLEKULARE SPEICHER
zur Langzeit-Archivierung

- 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!

Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft

MOSLA
00010
MOLEKULARE SPEICHER
zur Langzeit-Archivierung

Questions?

Philipps Universität Marburg

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LOEWE
Exzellente Forschung für Hessens Zukunft