FACHHOCHSCHULE KIEL
University of Applied Sciences

Design and Implementation of a software-based Realtime-Simulation-System for high speed Nanopore Sequencing Data Analysis

Nadine Kraft

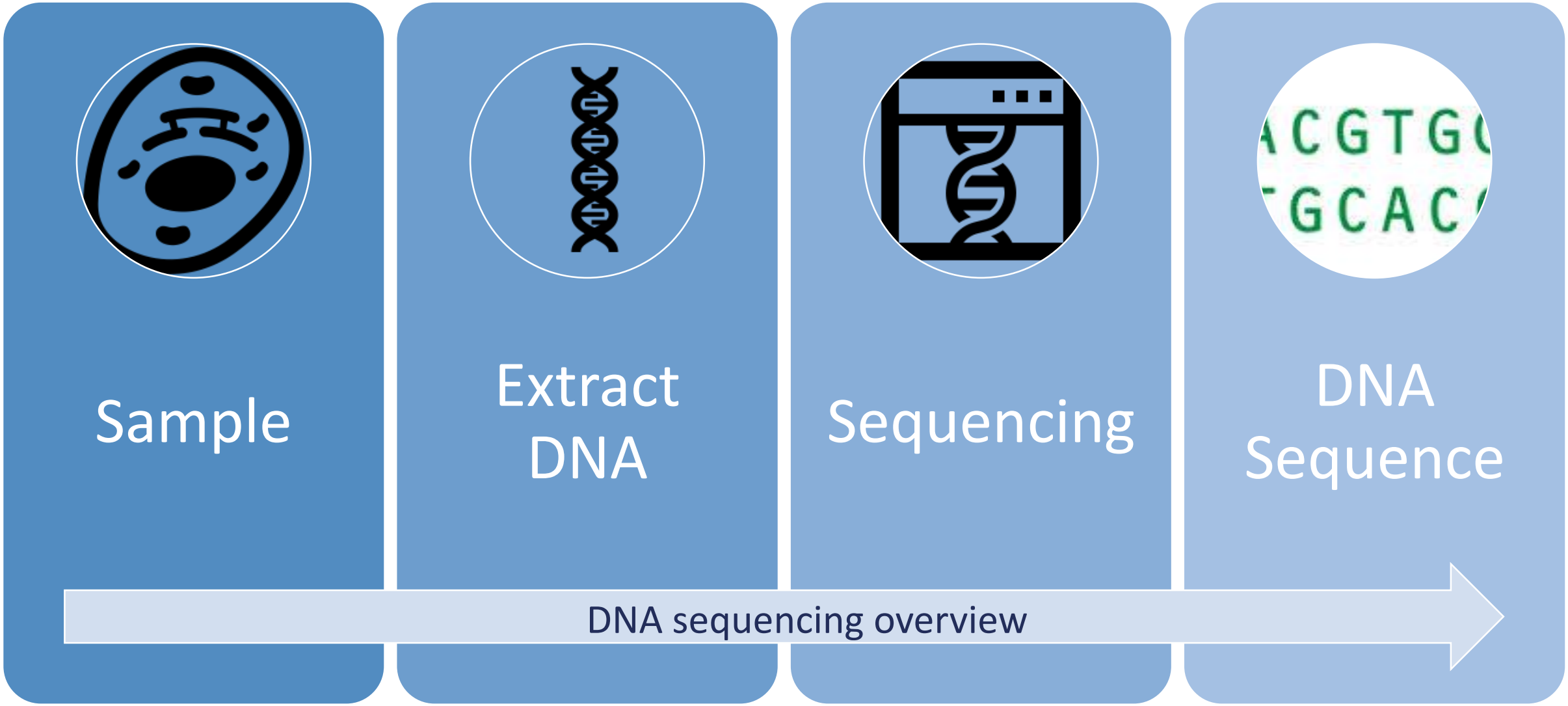
Kiel University of Applied Sciences

Erasmus+

Table of Contents

1. Introduction
2. Fundamentals
3. Development of the Simulator
4. Verification
5. Conclusion

Introduction



Introduction

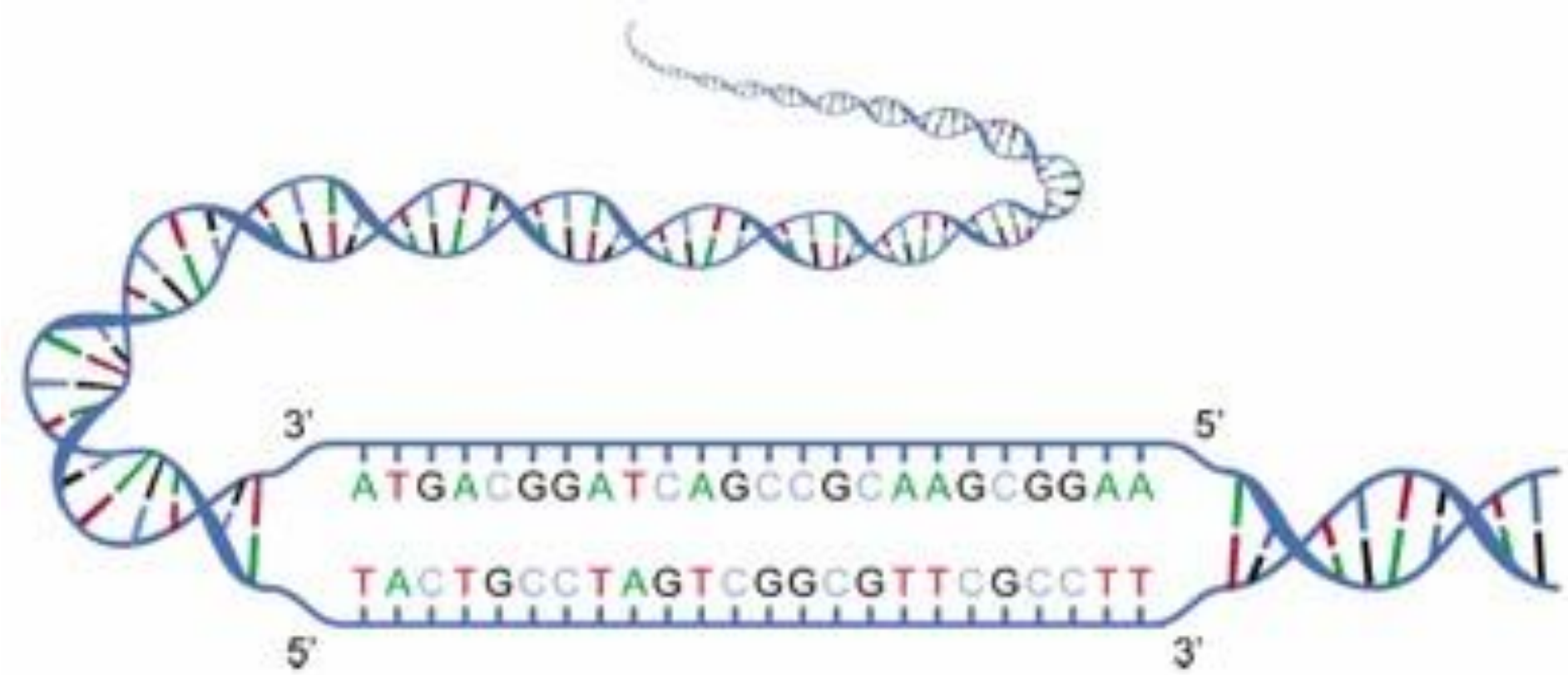
Oxford Nanopore Technologies MinION Sequencer

- Small and portable high-throughput sequencing platform
- Used for sequencing DNA or RNA molecules

Source: https://nanoporetech.com/products
Accessed on 15.09.2017

Basic Definitions

DNA



Source: G. Wolfgang (2010-2017) ; http://www.biologie-schule.de/desoxyribonukleinsaeure-dna.php; Accessed on 19.07.2017

Basic Definitions

K-mer (K-tuple)

- **Definition:** Element of the set of all possible substrings of length k that are contained in a string

A C G T T A C C C T T

Sequence divided into 5-mers

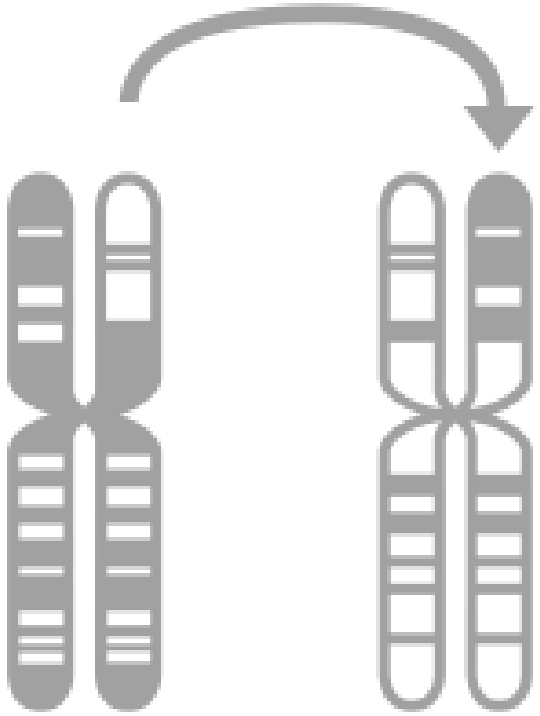
ACGTT
CGTTA
GTTAC
TTACC
TACCC
ACCCT
CCCTT

Basic Definitions

Mutation

Definition: Any alteration in the sequence of nucleotides is known as a gene mutation.

- **Point mutation:** Substitution of one nucleotide for another
- **Frame-shift mutation:** Deletion or insertion of nucleotides.



Basic Definitions

Sequence

Definition: Order of nucleotides within the DNA

- Termed as primary structure

DNA-Sequencing

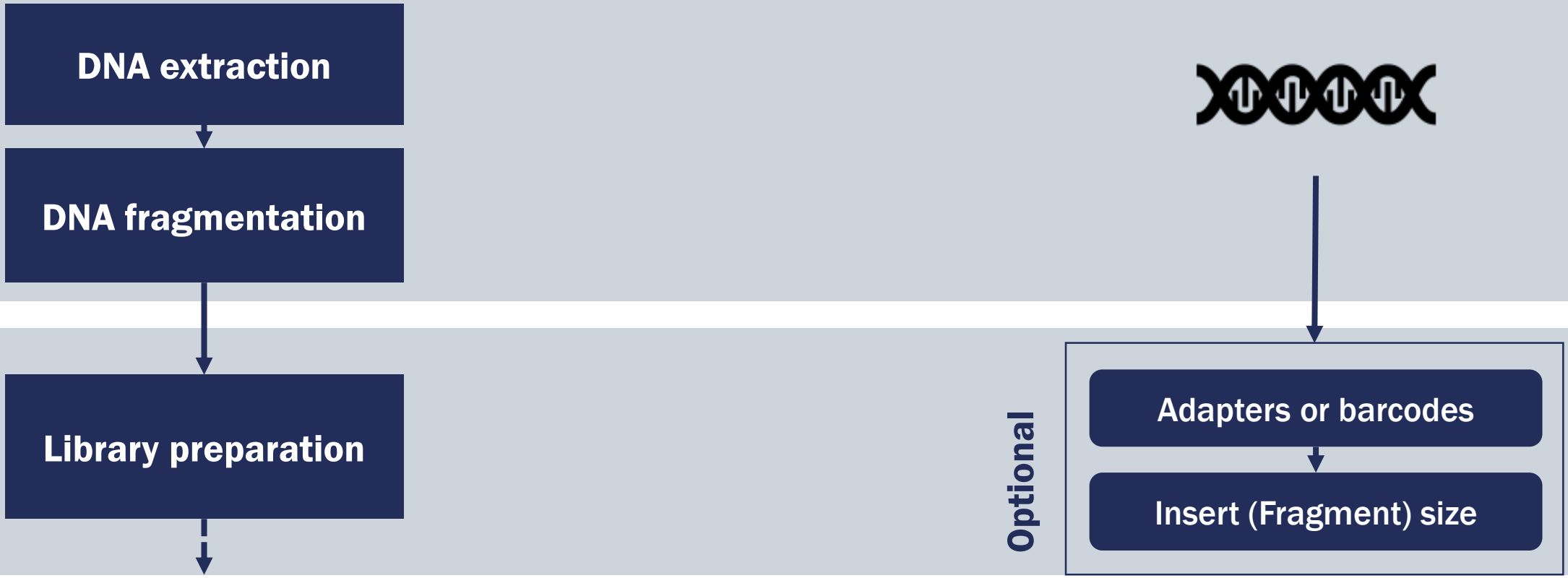
Definition: Method to identify the bases within a DNA sequence

- Approaches:
  - Next-generation-sequencing method
  - Third-generation-sequencing method

NEXT-GENERATION-SEQUENCING

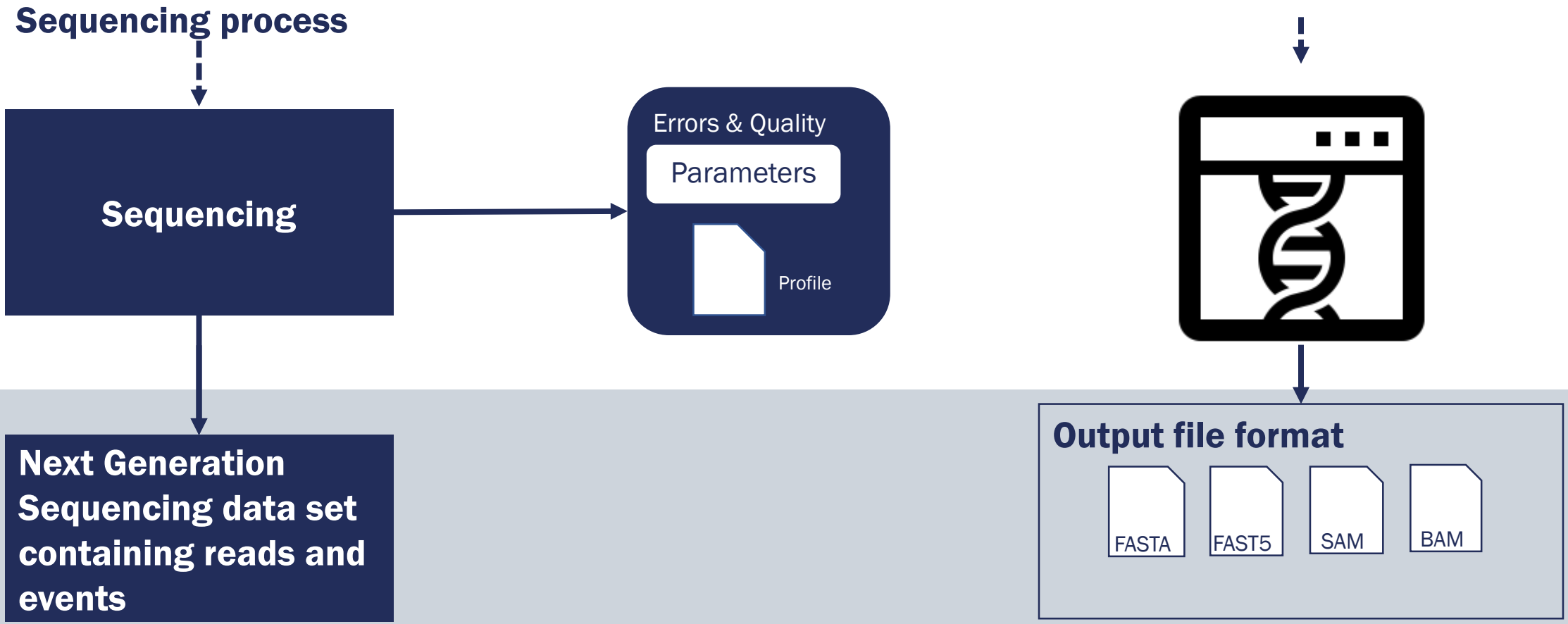
Flowchart Next Generation Sequencing Method

Sequencing process



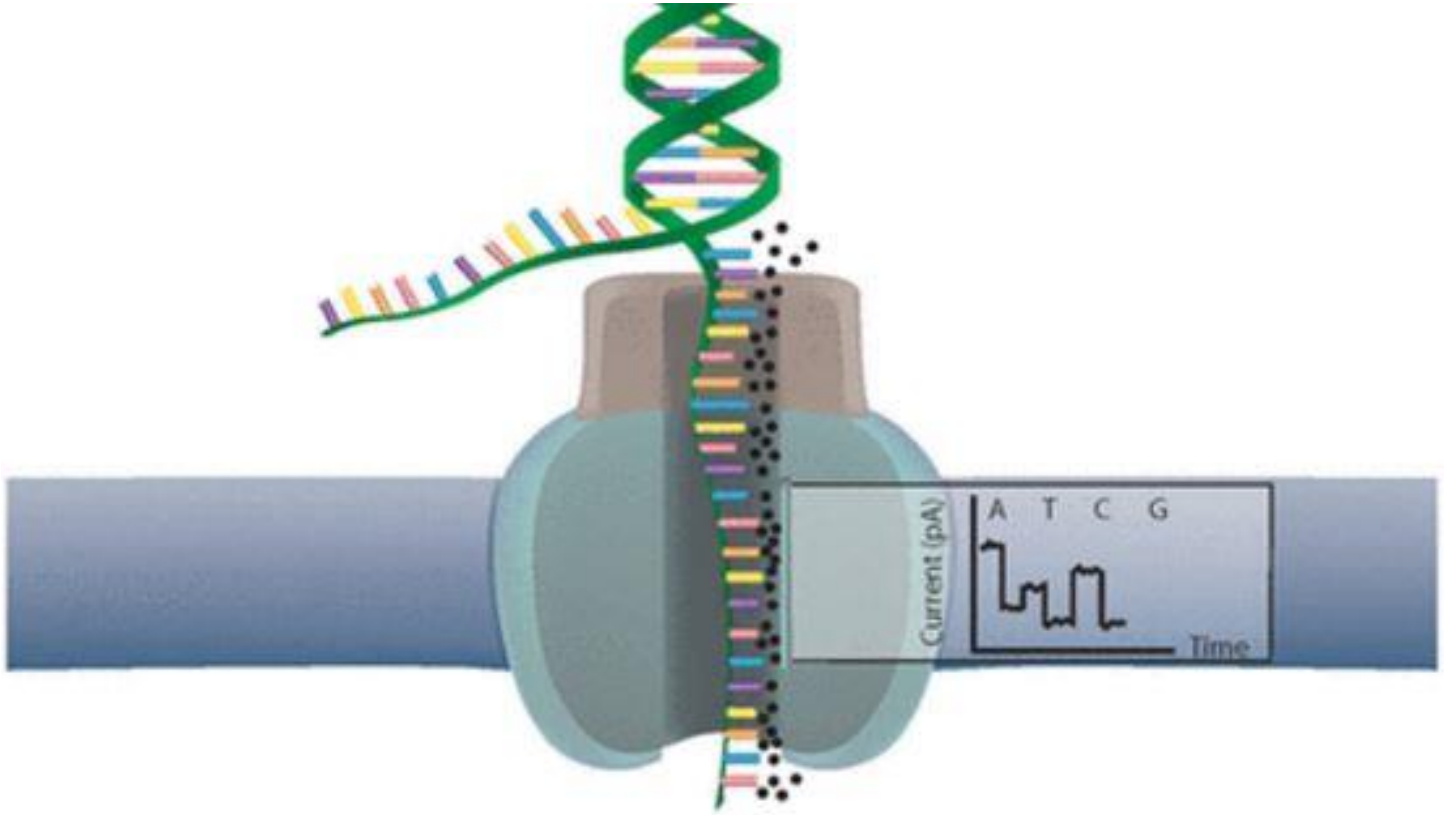
Next-Generation-Sequencing

Flowchart Next Generation Sequencing Method



THIRD-GENERATION-SEQUENCING

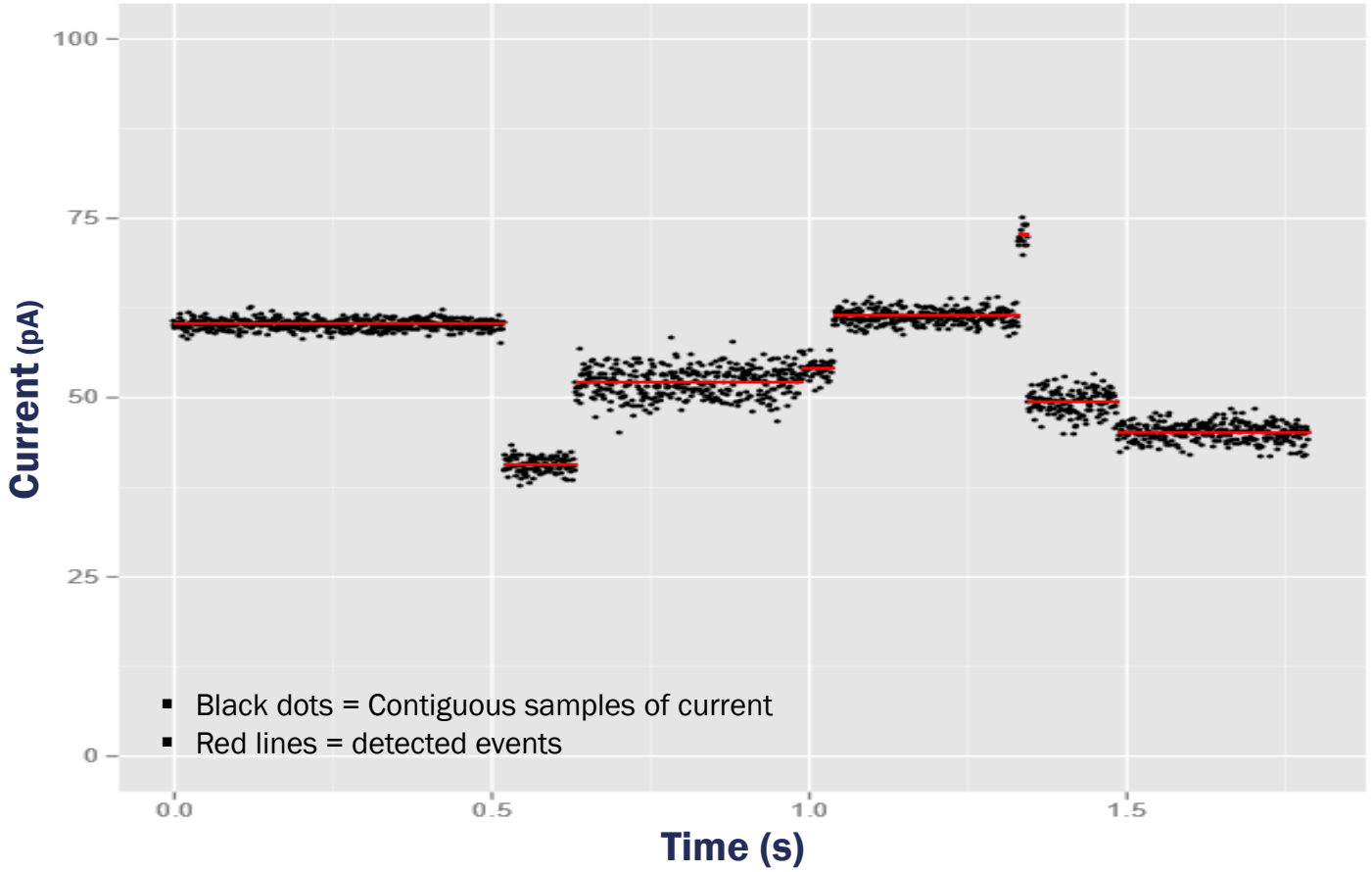
MinION - Workflow



Source: Churko et al. (2013) Overview of High Throughput Sequencing Technologies to Elucidate Molecular Pathways in Cardiovascular Diseases. Circulation Research (doi: 10.1161/CIRCRESAHA.113.300939)

THIRD-GENERATION-SEQUENCING

Contiguous samples of current make up detected events



Third-Generation-Sequencing

MinION events

- **Base-calling**: Translation of detected events into DNA sequence
- Events consist of:
  - Mean current value
  - Corresponding variance and duration
  - Start point
  - Corresponding model state
  - Move command indicating event or pseudo event

Contiguous samples of current make up detected events

	mean	start	stdv	length	model_st...	move
0	65.91390...	314.824	1.526084...	0.0015	GCGTA	0
1	62.71160...	314.8255...	1.238966...	0.002	GCGTA	0
2	80.68004...	314.8275	2.026480...	0.002	CGTAC	1
3	81.08922...	314.8295	2.027599...	0.00375	CGTAC	0
4	71.36225...	314.83325	0.691531...	0.002	GTACT	1
5	68.84934...	314.8352...	4.013957...	0.00425	TACTT	1
6	76.89461...	314.8395	1.146159...	0.002250...	ACTTC	1
7	77.78018...	314.84175	1.206087...	0.0025	ACTTC	0
8	78.66971...	314.84425	0.867185...	0.00125	ACTTC	0

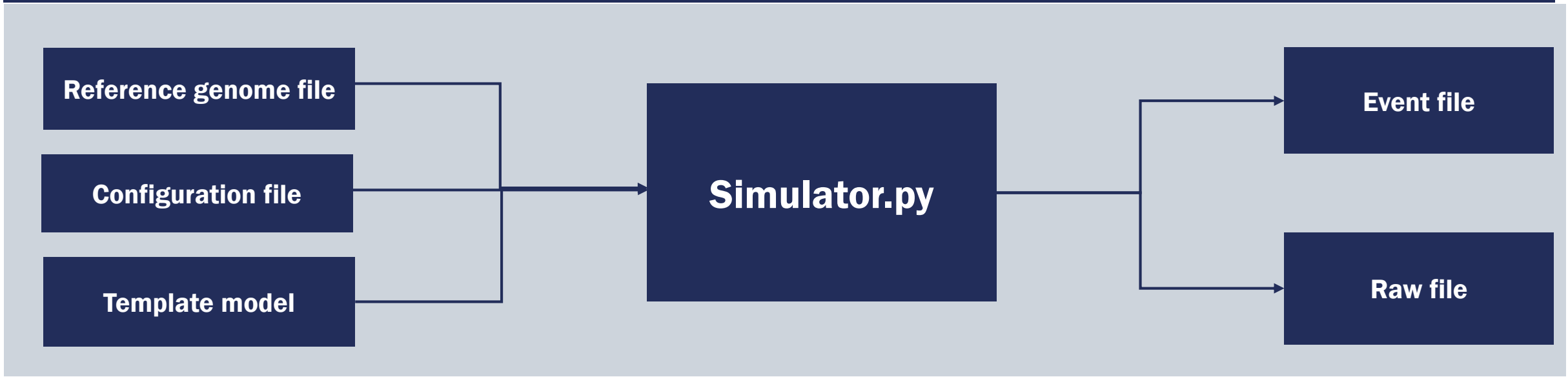
Development of the Simulator

Requirements

- The simulator must perform a complete simulation of an ONT MinION sequencer
- The simulator must produce realistic data
- The system must take configurations of real conducted experiments

Development of the Simulator

System Architecture



Input files

- Sequence
- Configuration file
- Template model

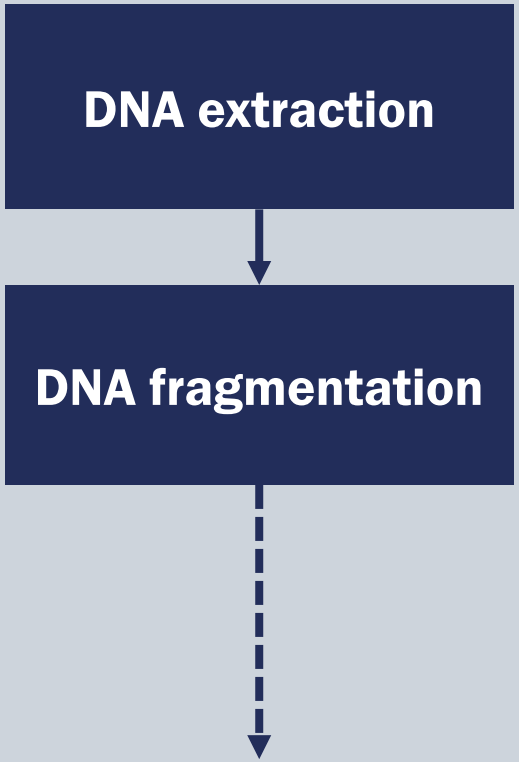
Output of the simulator

- Event file
- Raw file for the simulated reads

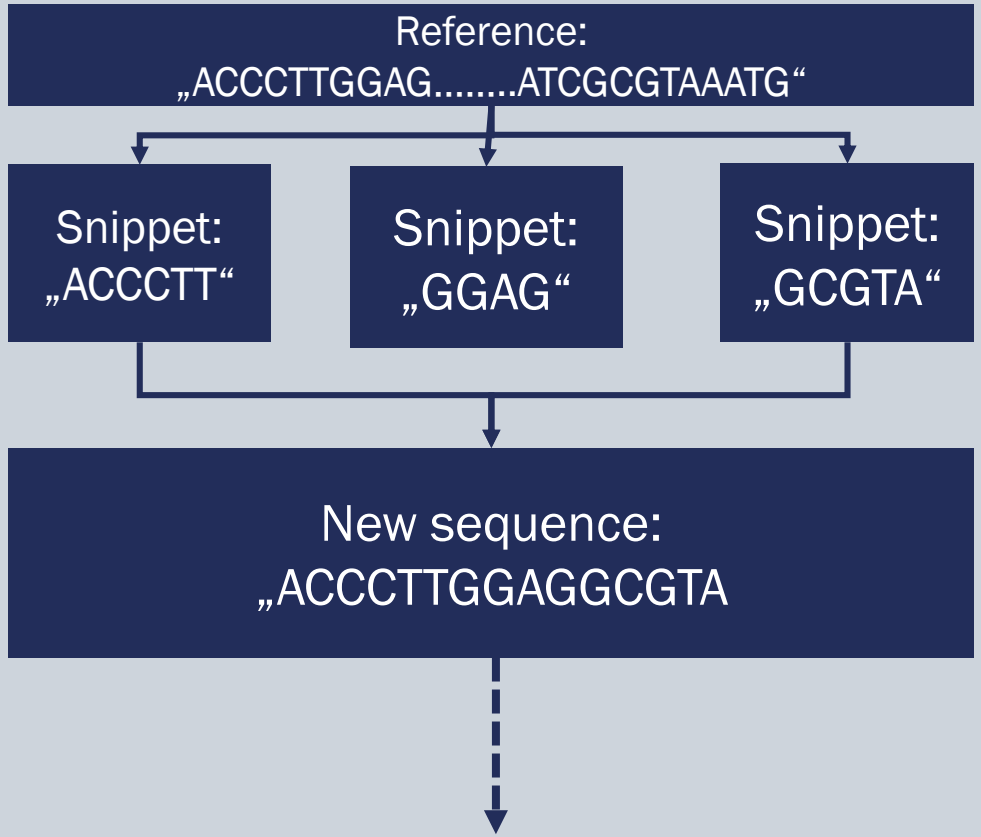
Implementation

Comparison sequencing and simulation Process

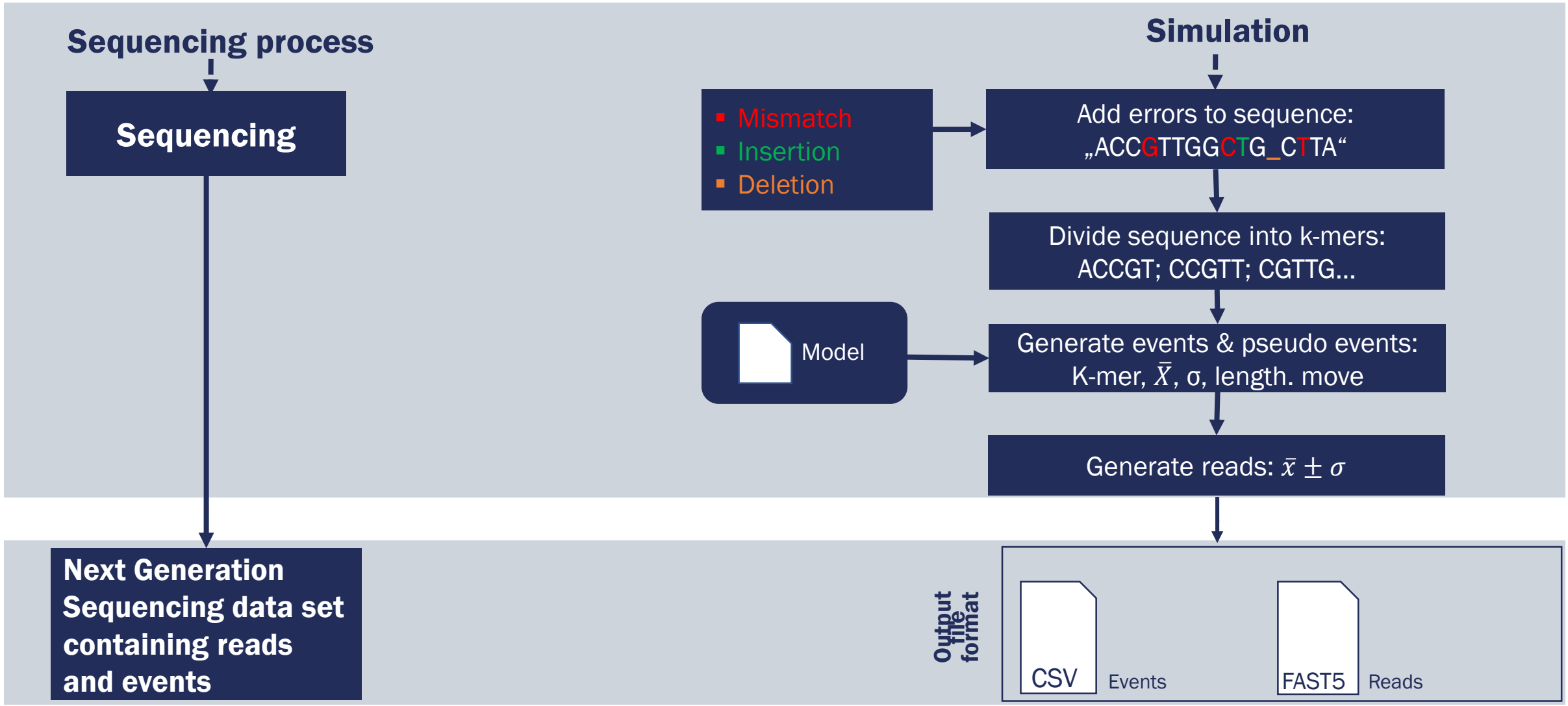
Sequencing process



Simulation



Implementation



Configuration

Configuring the simulation program

- Model template file
- Reference FASTA file
- Defining the error-rate
- Configuration file

System architecture

class SimSignalGenerator

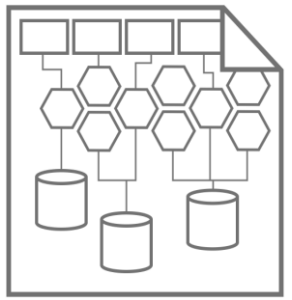
- self
- ref_file
- config_file
- error_rate
- debug

- def init (self, ref_file, error_rate, config_file, debug)
- def load_model(self, model_file)
- def load_reference(self, ref_file)
- def load_config(self, config_file)
- def generate(self, snip_count, generate_events)

class SIConfigFile

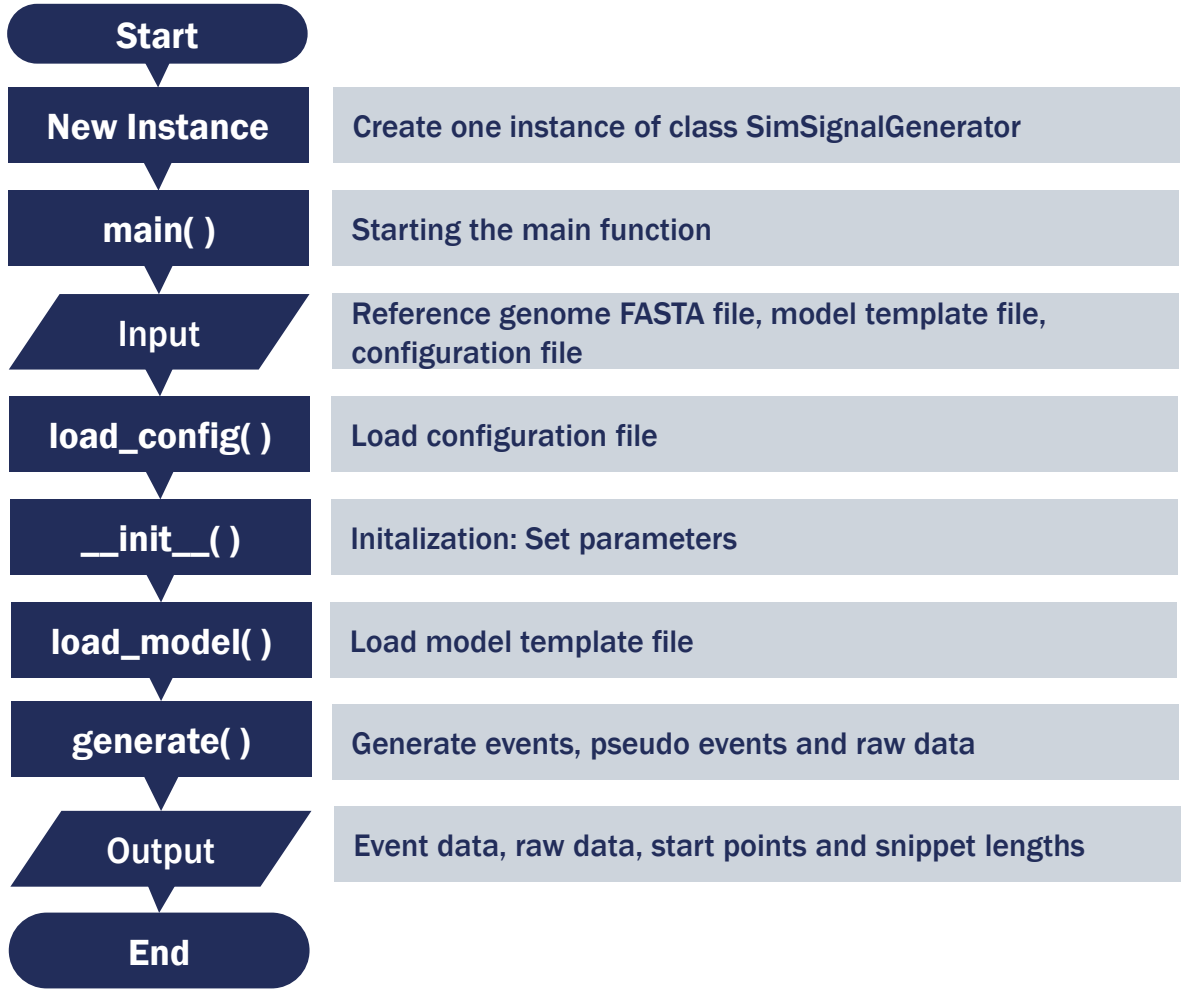
- read_length_distribution
- bases_per_second
- pores_number
- max_active_pores
- read_until
- wear_out

- load_file(config_file)



Software Design

Object Oriented Design



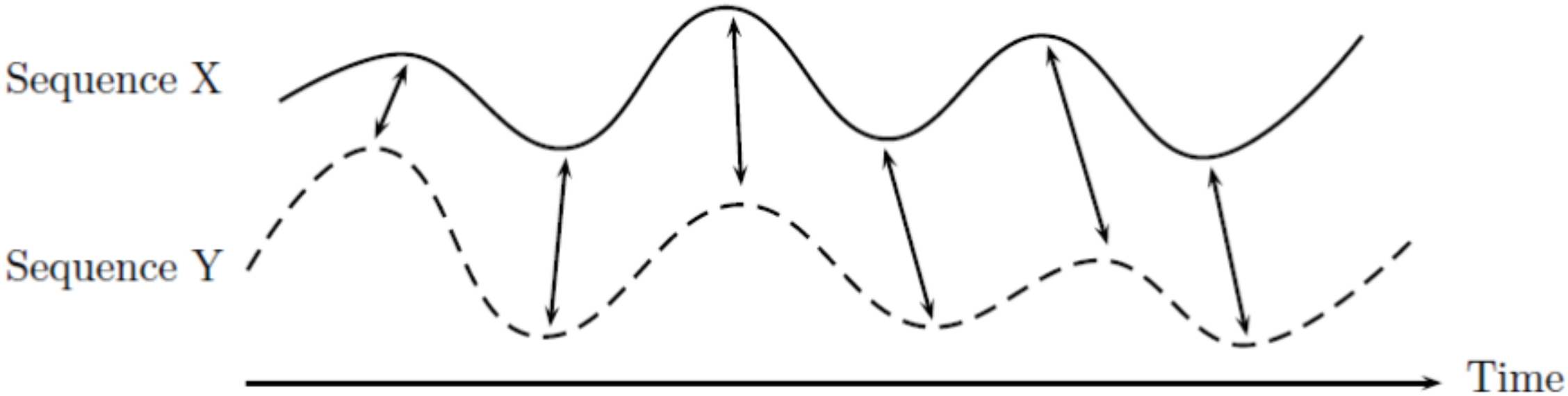
The simulator in action

Live Demonstration

Verification Methology – Dynamic time warping

Dynamic time wrapping

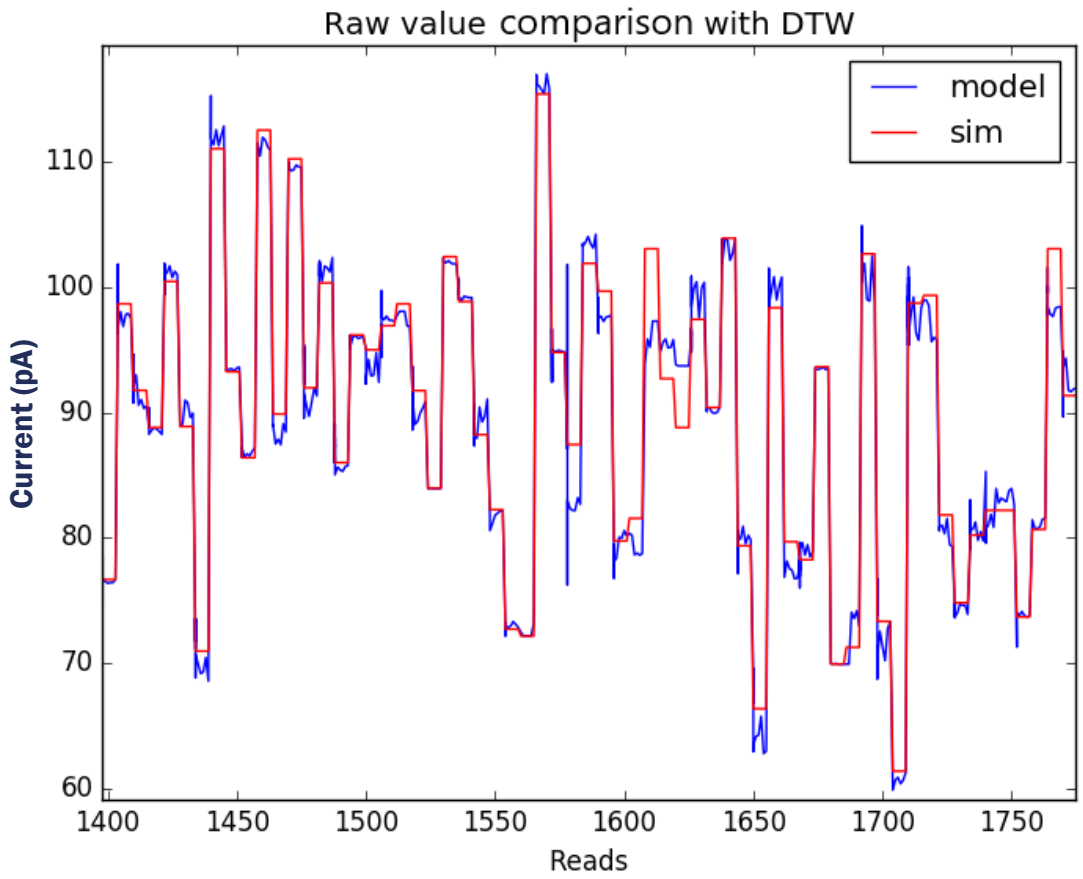
Time alignment of two time-dependent sequences



Source: M. Müller: Information Retrieval for Music and Motion, page 70, 2007 Springer

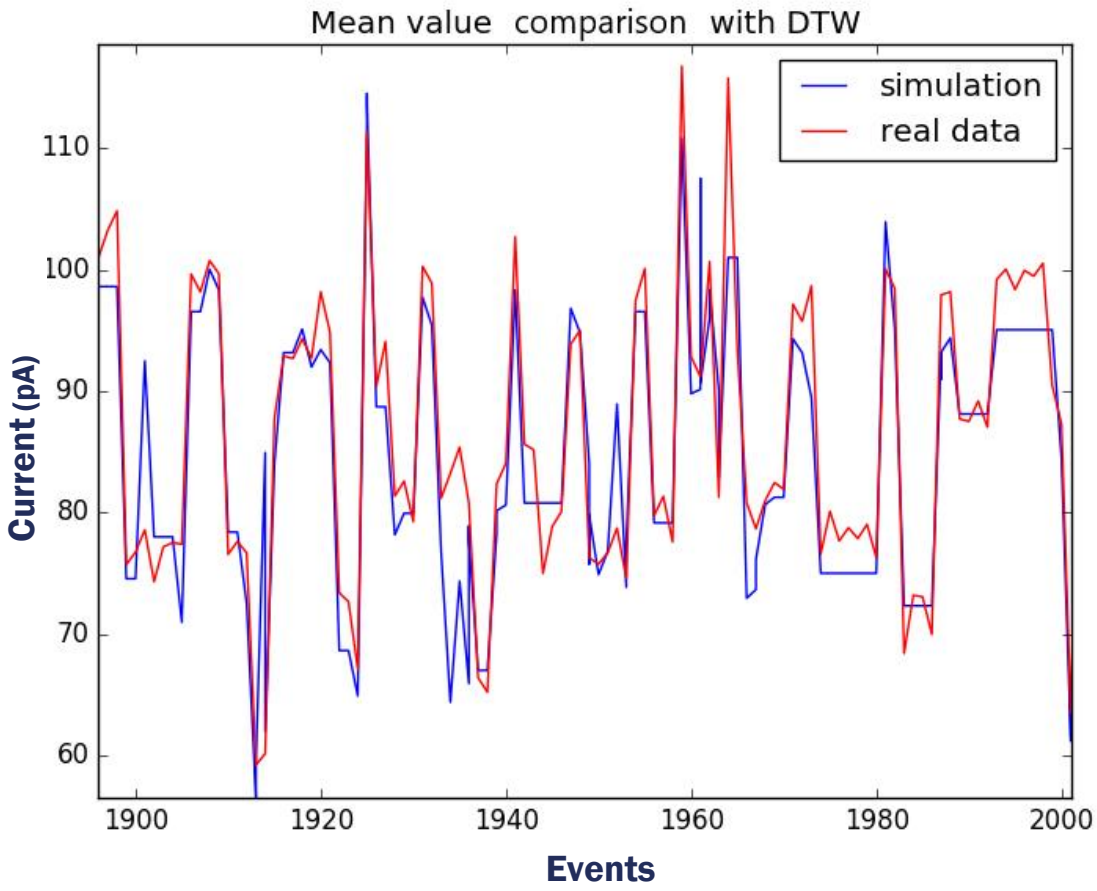
Comparing simulated and model raw data

Plot alignment of Simulated raw data and model raw data



Comparing simulated and experimental event data

Plot alignment of Simulated event data with event data of real conducted experiments



Conclusion

Conclusion

- Implementation of a tool for the simulation of the Oxford Nanopore Technologies MinION sequencer
- The simulator successfully integrates the simulatION part 1 program manager for the configuration of the simulated data
- Development of realistic simulated datasets containing event and raw signal data
- Simulator fulfills all the demanded requirements and delivers adequate results

Questions