UNIVERSITÄT DES SAARLANDES

ZBI ZENTRUM FÜR BIOINFORMATIK

Overview and Setup

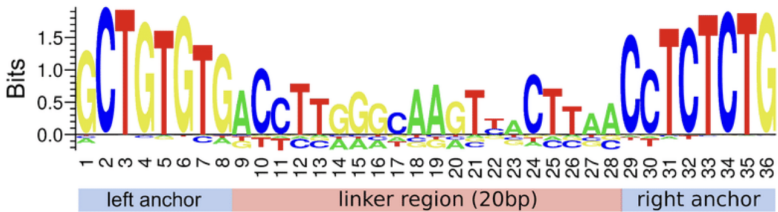
Just-in-time Compiled Python for Bioinformatics Research

Johanna Elena Schmitz, Jens Zentgraf and Sven Rahmann

ISMB 2024, Montréal, Canada (July 12, 2024)

Welcome!

We would like to write a (fast!) Python tool
that finds all occurrences of a given DNA motif in a given genome, e.g.
the ZNF768 binding pattern `RCTGTGYRN(17,23)CYTCTCTG`
[Rohrmoser et al.: "MIR sequences recruit zinc finger protein ZNF768 to expressed genes", Nucl. Acid Res. 47(2): p. 707, 2019]



- The linker region in fact has variable length (17-23 bp)
- Some of the positions allow more than one nucleotide (`R` = `A` or `G` at position 1)

Overview

We will discuss today

- the difference between interpretation, lazy and eager/early compilation,
- the possibilities and limitations of the `numba` just-in-time compiler,
- when `numba` can accelerate your code (and when it cannot),
- the pre-requisites for compiling a function,
- the differences between compilable and non-compilable Python code,
- how to speed up an initial Python implementation to handle larger data faster,
- how to parallelize Python in spite of the Global Interpreter Lock (GIL) with compiled functions,
- using a DNA motif search application as an example

Schedule

9:00	Setup Introduction to the numba just-in-time compiler for Python: small examples, how the compilation works, possibilities, limitations
10:45	Coffee break
11:00	Introduction to DNA motif search. Automaton-based pattern search and a bit-parallel algorithm.
13:00	Lunch break
14:00	Transforming a Python implementation to a numba-compiled implementation: separation of high-level and low-level code parts; managing memory allocations; introduction of type annotations
16:00	Coffee break
16:15	Parallelization: Using threads to parallelize the application; replacing the command-line interface by a simple GUI using streamlit
18:00	End of the tutorial day

Setup

What You Need

- a laptop with an up-to-date Python installation (3.12)
- a package manager, ideally `conda` or `mamba`, to install `numba`;
  we recommend to download and install **`miniforge`**.
- an internet connection to download our materials and the t2t reference genome

Detailed setup instructions will follow in the hands-on sessions.