CSSI Element: EAGER: Bridging the Last Mile

Rajiv Ramnath; Bryan Carstens; Jian Chen; Drew Duckett; Swathi Vallabhajosyula
{ Carsten.12; chen.8028; ramnath.6; duckett.17; vallabhajosyula.2 }@buckeyemail.osu.edu
Institutions: Ohio State University

THE OHIO STATE UNIVERSITY

Motivation

With the onset of any research/project, most of the time is spent on data acquisition (published and verified data sets suitable for the study), pre-processing (noise reduction, visualizing and manipulating the data to fit our research) and tool exploration (state-of-art techniques). Some preliminary analysis along with small-scale computations is needed to compare the tools' results while adjusting relevant software parameters and modal parameters. During this entire process, one has to explore various resources (how-to guides, research papers/journals, textbooks, internet) and/or seek ad-hoc advice from colleagues, collaborators, advisors. Most of the suggestions/recommendations go undocumented unless they were implemented and thus recorded. Also, different researchers need guidance during different stages of their research progress and in various forms. This project proposes the use of artificial intelligence to build a cyberinfrastructure tool that assists by utilizing past experiences and other resources to carter to individual researcher's needs.

Process for Computationally Driven Science

a. General Flow

b. Genome Sequencing
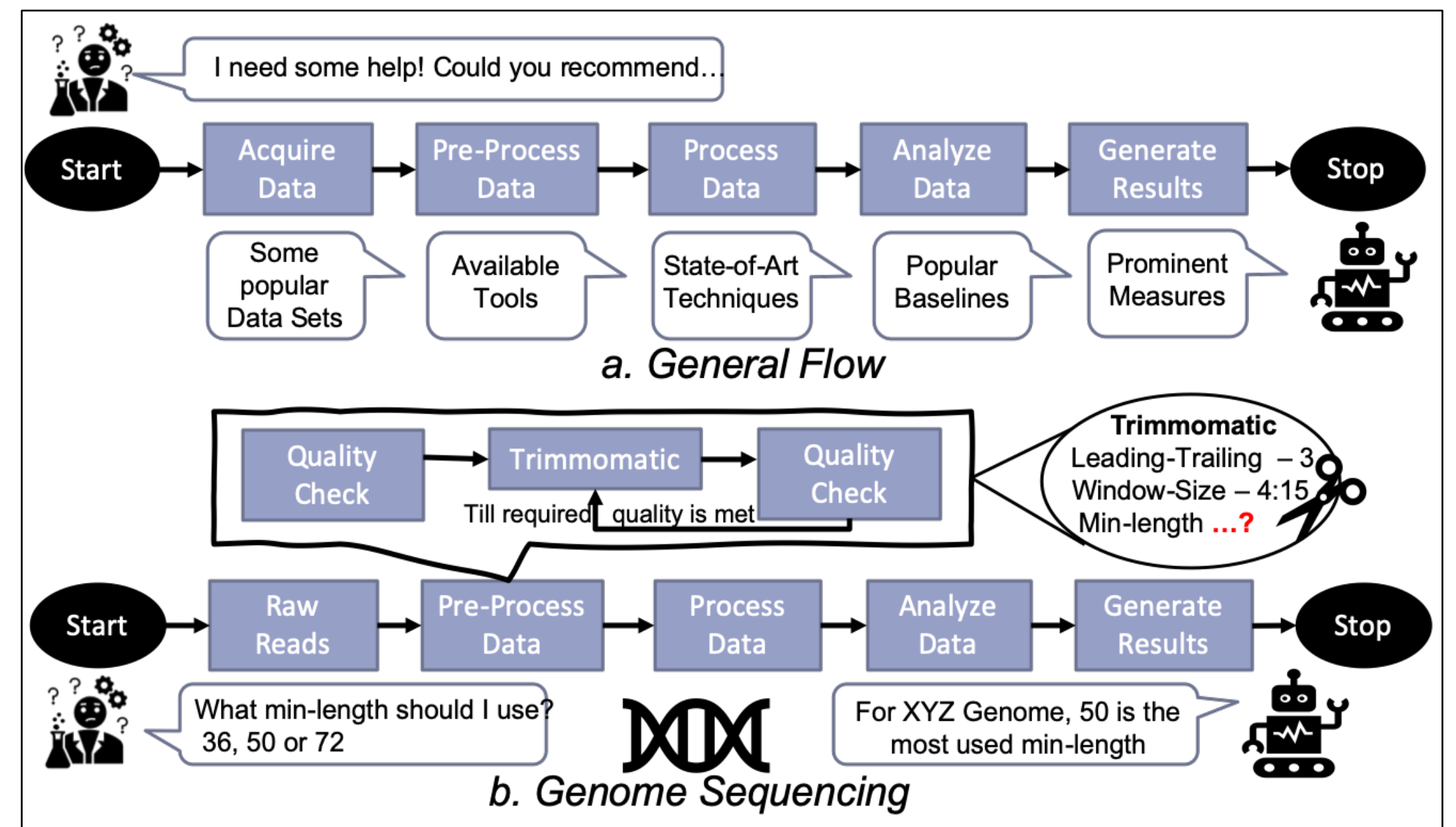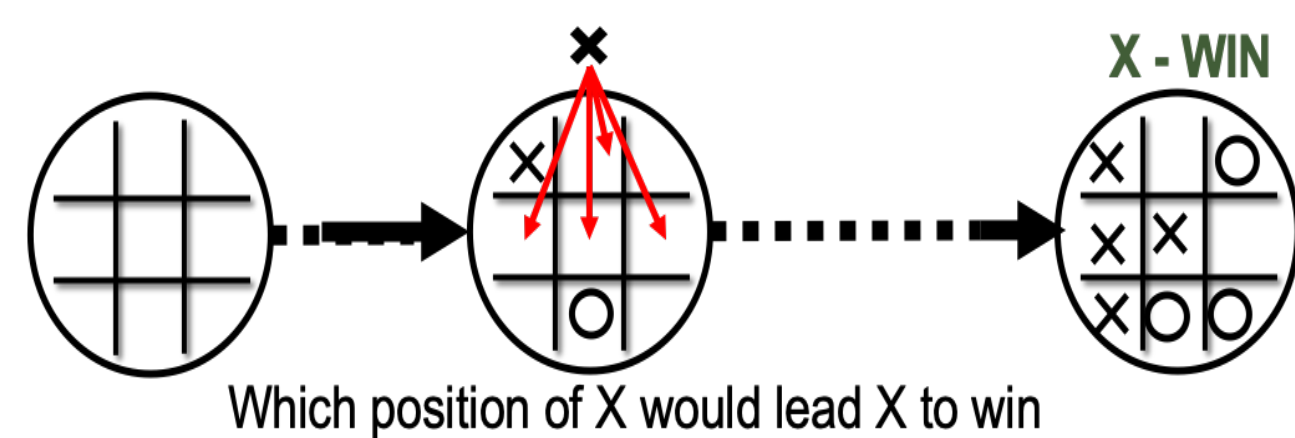
Domain

We chose Genome Sequencing as the domain of our interest. Sequencing is used for many important tasks such as analyzing how biological diversity is generated; to identify evolutionary contributions and understand the role of the ecology and environment in the formation of our target population's genetic structure

Complexity

DNA sequencing is an NP-Hard problem. A raw read file for a genome (water-vole) is around 15 Gb in size. Each step through the genome sequence from around few hours to two days. Some recommended parameter setting could reduce days of redo work.

Which position of X would lead X to win

Models:

Linear Regression
Reinforcement learning

References

1. https://carstenslab.osu.edu/index.html
2. https://www.ceti.cse.ohio-state.edu/front-page#page-content
3. Blazewicz, J. and M. Kasprzak, "Complexity of DNA Sequencing by Hy- bridization," Theoretical Computer Science, 290, 2003, pp. 1459-1473.