Software Requirement Specification

Author: Dorzhey Matkhanov

01/30/2024

Version: 1.0

**Introduction**

* **Purpose of this Document –** to describe and specify design and requirements of software Scgog (System in future use).
* **Scope of this document –** Main objective is to provide understanding and clarifications of expectation of System and development project. The estimated cost is $0 and estimated development time is 3 months.
* **Overview –** System will centralize most common Data Analysis techniques on Single Cell data. Most of techniques are focused on but are not limited to Machine Learning application.

**General description**

In this, general functions of product which includes objective of user, a user characteristic, features, benefits, about why its importance is mentioned. It also describes features of user community.

Single Cell biology is a rapidly growing field with amount data increasing even more. In purpose to help users to analyze this data without deep expertise in programming and Machine Learning, saving time and efforts. Users can be anyone from researchers and developers to amateurs interested in exploring a new field.

**Functional Requirements**

System is implemented in Python programming language in the form of a package, i.e. independent distributable element. From now on mentioned as *System*.

System is required to implement following Single Cell data analysis steps:

1. System pre-processes raw Single Cell data and validates it
2. System implements Quality Control
3. System conducts Normalization
4. System conducts Integration of datasets
5. System conducts Dimension reduction and Clustering of the cell-level
6. System conducts Cell type annotation
7. System implements Machine Learning analysis
   1. Cross-validation on reduced dataset
   2. Benchmarking of 5 different models by choice
   3. Hyperparameter optimization of the chosen model

**Data Requirements**

Input data should be

1. scRNA-seq data in form of Cell-Gene matrix and meta data
2. scATAC-seq data in form of Cell-Cell matrix, peak annotation data, and meta data

Both in Comma-Separated Values (.csv) format

**Interface Requirements**

To ensure high level of data analysis implementation System is required to be able to communicate with the User using Command-line Interface (CLI), in form of commands or option choice.

**Performance Requirements**

System is required to implement most reliable and popular data analysis methods. System is required to minimize inherited inaccuracies of high-level implementation of data analysis

**Design Constraints**

User’s machine must possess all technical prerequisites listed in design documents. User must be familiar with using CLI, and possess basic knowledge of required input parameters of each step in Single Cell data analysis.