curATime

curAHack Challenges 2025

Challenge 1: Interactive Visualization of Paired scRNA-seq and scTCR-seq Data

In interdisciplinary research environments, making complex datasets accessible and easy to explore is crucial for collaboration. While many tools exist for visualizing single-cell RNA sequencing (scRNA-seq) data, most focus solely on transcriptomic information. In the context of the curATarget project, paired single-cell transcriptomic and T cell receptor (TCR) clonotype analysis are combined to investigate the immune environment of atherosclerotic plaques. As multiome approaches, which capture multiple modalities per cell, become more common, there is an increasing need for tools that enable interactive exploration of such data.

The goal of this challenge is to create an interactive interface that allows users to explore paired scRNA-seq and scTCR-seq data, specifically related to atherosclerotic plaques. This may involve the development of a dashboard, or perhaps more feasibly, the implementation of a plugin for an existing platform like the ShinyCell app to visualize immune receptor diversity analyses. Ultimately, the aim is to provide a user-friendly, interactive way to explore complex multi-omics data, facilitating deeper insights into immune responses in disease contexts.

Challenge 2: Multiclass Classification of Anti-CRISPR Proteins

Anti-CRISPR (Acr) proteins inhibit the CRISPR-Cas systems and are being explored to enhance the precision and safety of gene editing, for example in therapeutic applications in cardiovascular diseases (CVDs). These proteins can prevent the Cas-gRNA complex from binding to DNA or deactivate the Cas effector, thereby reducing off-target effects. However, classifying Acr proteins remains a challenge, especially for novel Acrs that lack sequence homology to already known proteins. Traditional methods, such as sequence alignment and manual curation, are time-consuming and may not capture functional similarities across diverse Acr families.

Machine learning (ML) presents a promising solution to classify both known and novel Acr proteins into functional classes. However, developing accurate models is complicated by high sequence diversity, data imbalance, and limited labeled data. Some Acr families are underrepresented, leading to class imbalance, while many Acr proteins lack clear functional annotations. The goal is to develop a robust, ML-based multiclass classification model capable of handling equence variability and class imbalance, while also predicting functional classes for novel Acrs without requiring extensive sequence homology. This could be implemented in the curATime project curATarget.

The challenge is to develop a data-driven classification solution for identifying new anti- CRISPR proteins. This solution could involve using traditional sequence encoders combined with machine learning (ML) or deep learning classifiers, as well as hierarchical or ensemble models to capture relationships among different Acr families. Such a model would streamline Acr protein identification, reduce the need for manual curation, and support both academic research and therapeutic applications of CRISPR technology. Additionally, it would provide valuable insights into CRISPR resistance mechanisms and guide the design of Acr-based CRISPR modulators.

Challenge 3: AlphaFold and Adversarial Examples

AlphaFold is an AI-driven software tool designed to predict the tertiary and quaternary structures of proteins and protein complexes. The output can producing models that surpass the quality of those generated by traditional methods. However, like other AI models, AlphaFold can exhibit unpredictable behavior when processing certain input data, known as adversarial examples. These are instances where small, intentional perturbations in the input features cause the machine learning model to make incorrect predictions.

In our internal use of AlphaFold, we have encountered such examples, particularly with TCR-p- MHC complexes (referring to the interaction between a T cell receptor (TCR) and a peptide-major histocompatibility complex (p-MHC)), where minor changes in input sequences led to erroneous predictions of the relative orientations of protein complex members. Understanding and identifying these adversarial examples is crucial for improving the reliability of AI models used for protein structure prediction.

The aim of this challenge is to enhance the robustness of predictive systems for TCR-p-MHC complexes by exploring adversarial examples and improve the accuracy of antigen predictions, particularly those related to TCRs in atherosclerotic plaques. This, in turn, will contribute to advancing our understanding of immune responses in disease contexts.

Challenge 4: Intercellular Communication in Cardiovascular Disease

Single-cell sequencing is a key tool for assessing disease progression at the cellular level. In atherosclerosis, single-cell transcriptomic analysis has provided insights into plaque-specific macrophage differentiation. Cell Chat, an open-source R package, infers intercellular communication networks from single-cell transcriptomic data on ligand-receptor or pathway level.

However, the ligand-receptor communication network is limited by its reliance on curated peer-reviewed datasets, which often lack comprehensive disease-specific interactions. Understanding communication differences between diseased and healthy groups can be challenging due to insufficient literature annotations. Furthermore, analyzing the vast number of potential intercellular communications is both time-consuming and difficult to contextualize within known cardiovascular disease mechanisms.

High-resolution single-cell RNA sequencing data is crucial for disease characterization. Within the curATime cluster, this data can identify cell-type-specific targets for diseases like atherothrombosis. Efficient analysis of CellChat outputs can simplify interpretation and aid in identifying potential drug targets. Automated literature integration can reduce manual effort in analyzing large communication networks.

In the context of the curATarget and the curAIknow project, the challenge is to develop a framework that explains differences in intercellular communication between cardiovascular disease and healthy groups. This framework should leverage existing literature and knowledge bases, enrich datasets with relevant interactions, and discover new mechanisms. It should also account for interaction probabilities. The goal is to create a knowledge graph that integrates annotations, CellChat outputs, and novel findings to provide actionable insights.

Challenge 5: Leveraging scRNA+scATAC-seq for Automatic Cell Type Annotation

Cell type annotation is crucial in single-cell analysis, used to decode cellular diversity. Traditional methods using single-cell RNA-sequencing (scRNA-seq) map gene expression profiles to known marker genes, often failing to capture rare or transitional states. Single-cell multiome sequencing, which combines gene expression and chromatin accessibility (scATAC-seq), offers enhanced resolution. However, integrating these modalities for annotation remains a challenge. While scRNA-seq offers high transcriptional resolution, it does not capture transient cell states. Multiome sequencing provides additional insights into regulatory elements. Marker databases are useful but often limited. Integrating scATAC-seq data can identify novel regulatory elements and refine annotations.

The limitations include marker gene databases not resolving closely related cell types, underutilization of chromatin accessibility data, complexity in integrating scRNA-seq and scATACseq, and the tedious and manual process of current annotation methods. Automating the integration of gene expression and chromatin accessibility enables the reliable identification of unknown cellular subpopulations, benefiting the curATime community by studying transitions in immune and cardiovascular cells. By

automating the integration of gene expression and chromatin accessibility enables the reliable identification of unknown cellular subpopulations. This approach benefits the curATime community by studying transitions in immune and cardiovascular cells.

In the context of the curAIntervent project, the goal is to develop a (semi-)automated cell-type annotation pipeline that utilizes both gene expression and chromatin accessibility. This pipeline should pre-process and integrate both data types, assign cell types using public or customized databases, highlight regulatory regions using scATAC-seq data, and provide a scalable, reproducible solution.