BITS :: Call for Abstracts 2024 - Oral communication

Type	Oral communication
Session	Bioinformatics AI, Models and Tools
Title	Machine learning for expression-based multi-label patient subtyping and advanced recognition of primary and secondary assignments
All Authors	Cascianelli S(1), Masseroli M(1)
Affiliation	

(1) Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Giuseppe Ponzio 34, 20133, Milano, Italy.

Motivation

Transcriptional subtyping is pivotal both for biomedical research and clinical domains, especially for oncology. Patient stratification into molecular subtypes of a given cancer is widely studied and adopted to identify differences and peculiarities among cases and contribute to patient clinical handling. While single-sample methods are increasingly investigated to support clinical practice, multi-label strategies able to associate each patient with more than one class have barely been addressed to date for subtyping purposes, despite their wide application to other fields. Current stratification methods tend to oversimplify patient molecular characterization by assigning only the most prominent subtype, although distinct molecular and functional traits of cell populations can be reflected and appreciated at the bulk sample level using multiple subtype assignments.
To address this gap, we developed an innovative computational workflow, named MULTI-STAR, which guarantees comprehensive and reliable single-sample multi-label expression-based subtyping of patients.

Methods

MULTI-STAR (MULTI-label SubTyping and Advanced Recognition) includes: 1) a multi-label characterization step, which can extend any state-of-the-art similarity-based transcriptional subtyping technique to a multi-label perspective; 2) a multi-label classification step, which uses previous characterizations to obtain promising machine learning solutions that predict the subtype(s) portraying any new single patient. Assigning one or multiple classes requires finding how many and which subtypes should be associated with a sample to adequately describe it: learning this task in the absence of data already associated with multiple labels is challenging. Accordingly, the MULTI-STAR characterization step focuses on generating reliable multi-label references using similarity measures from an existing state-of-the-art subtyping method together with M-cut and further filtering strategies. Then, the MULTI-STAR classification step rigorously explores predictive models implementing multi-label strategies of algorithm adaptation and problem transformation to identify the multi-label solution that better depicts inner heterogeneity for the subtyping task at hand. The weighted average F1-score is chosen as the optimization metric during cross-validation and hyper-parameter tuning to balance the contributions of all the subtype assignments rather than focusing on multi-label accuracy, which could be biased by major classes. All the multi-label classifiers under exam are compared in cross-validation and testing using a plethora of label-based and example-based multi-label and class-specific performance measures. Furthermore, the most promising approaches are carefully evaluated in terms of the biological and clinical relevance of their patient predictions to identify the best predictive solution for the subtyping task at hand.

Results

MULTI-STAR can dissect molecular heterogeneity of disease at the bulk level, providing computationally and biologically valid multi-label models and predictions, given as inputs the patient expression profiles and an existing state-of-the-art subtyping standard. Subtype assignments are further categorized into primary and secondary assignments to distinguish the most prominent subtype (primary) from the additional significant secondary assignments, improving the comprehensive multi-label perspective.
We validated MULTI-STAR independently on breast (BC) and colorectal (CRC) cancer subtyping, using RNA-seq expression profiles from The Cancer Genome Atlas and considering, respectively, the PAM50 Intrinsic Subtypes and the ColoRectal Cancer Intrinsic Subtypes, as target classes of interest. A large part of BC and CRC cases emerged as significantly assigned to more than one subtype from the multi-label characterization step, and the advanced recognition of multi-label assignments in the classification step demonstrated to enhance the prognostic capabilities of the stratification by considering not only the patient primary subtype but also the so-far overlooked secondary subtypes. Indeed, enrichment analyses focused on prognosis-related subtypes showcased the importance of considering secondary assignments to provide more significantly valuable indications concerning clinical events, like relapse in BC or survival in CRC, overcoming the prognostic relevance of more straightforward single-label subtyping methods. Thus, MULTI-STAR both provides clinically promising multi-label classification solutions for BC and CRC and represents a significant general methodological result. In fact, it can be applied to reliably dissect the inner heterogeneity of any given disease based on transcriptional subtyping. Furthermore, it can be potentially extended to deal with other state-of-the-art subtyping approaches, not only based on transcriptional dataset-level similarity techniques.

Info

This work was supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGeneration EU program, within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line

on Artificial Intelligence)	
filename	-
Figure	
-	
Availability	-

Dissemination Material

Social	
-	
Summary	
-	

Corresponding Author

Name, Surname	Silvia, Cascianelli
Email	silvia.cascianelli@polimi.it
Submitted on	02.05.2024