ERCIM "ALAIN BENSOUSSAN"
FELLOWSHIP PROGRAMME

Scientific Report

First name / Family name	SUMANTA/RAY
Nationality	
Name of the *Host Organisation*	CWI, NETHERLANDS
First Name / family name of the *Scientific Coordinator*	LEEN/STOUGIE
Period of the fellowship	PROFESSOR
	01/07/2019 to 31/08/2020

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

Single cell RNA sequencing (scRNA-seq) technologies provide unprecedented opportunities to capture high-quality gene expression snapshots in individual cells. Technological advances have recently enabled to process several thousands of cells per scRNA-seq experiment. A fundamental step in single-cell analysis is to type the individual cells one analyzes. The most immediate approach is to cluster the population of cells under analysis into different groups. The groups to which individual cells belong then further determine the identity of the individual cells. This way of typing and annotating cells (reflecting unsupervised learning approaches) has been prevalent in identifying biologically coherent populations of cells in scRNA-seq data so far.

However, the process of assigning biological meaning to cell clusters is both complicated and time consuming, because it requires to inspect the identified cell clusters manually. Arguably, the procedure may even cancel the very advantages of single cell typing, because manual annotation requires to rely on prior knowledge, which had typically been gained by analyzing bulks of cells, and not single cells. Clearly, single cell experiments themselves constitute optimal resources for revealing and defining novel cell types. With the revelation and availability of ever more cell sub-/types, cell states, possibly even at the level of cells transiting between types and states, manual inspection will be too tedious,

inaccurate, or just impossible in terms of manpower resources. New methodology is required that allows to determine cell labels (sub-/types, states) in an automated fashion. Supervised learning-based approaches address these points by being able to determine the identity of single cells although the characteristic molecular mechanisms have not yet been fully understood. This explains their recent gain in popularity.
Tapping additional resources, such as protein expression data (e.g. sc-CITE-seq data), methylation or chromatic accessibility data (the latter provided by e.g. sc-ATAC-seq) and combining them with basic sc-RNA-seq data offers further advantages. When integrating additional data appropriately, one can expect both enhanced basic classification and the identification of cell (sub-)types that remain invisible on the RNA level alone. The driving methodical challenges are the heterogeneity of data sets, and the lack of general models for coherent integration.

In this project, we address the two challenges of establishing advanced supervised learning-based methodology and the smooth integration of additional data in single cell typing. Beyond addressing these challenges in themselves, we also focus on their combination.

As for supervised learning, we employ capsule networks, as a most recent and advanced deep learning approach that has proven its superiority in other, prominent areas of application so far. Here, we demonstrate how to use capsule networks to equally great advantage also in single cell typing. We leverage the trade marks of capsule networks for typing cells at utmost accuracy, economic use of training data (Capsule Networks achieve competitive accuracy already on only half or even only a third of the data in comparison with state-of-the-art approaches), and biologically meaningful interpretation of results.
46 Note that the reduced demand of training data helps to identify also types where cells suffer from a relative lack of coverage. Biological interpretation addresses the notorious complaint that high-performance deep learning is little explainable: here we can deliver explanations along with outstanding performance. We also present novel technology that enables a coherent integration of additional data. We create data representations based on non-negative matrix factorization and variational autoencoders that support the coherent integration of data from additional experimental resources, such as sc-ATAC-seq, sc-CITE52 seq, or sc-bisulfite sequencing, and which is generally applicable. Experiments point out that these representations enhance the performance in classification even further. That is, we demonstrate that mastering the methodical challenges of the integration of additional data indeed means an important step up in classifying single cells, as was anticipated in earlier studies

In this project, we provide the following contributions:
(1) We provide the first capsule network based approach (MarkerCapsule) that successfully deals with sequencing data in general, and single cell sequencing data in particular. Note that earlier Capsule Network based biological applications have only addressed protein structure prediction, the prediction of secretory proteins in saliva based on non-sequencing based proteomics data raised in 2007/8 and in network and disease biology

(where their possible advantages for processing heterogeneous, multi-omics data were pointed out).
(2) Although studying the integration of additional data has gained considerable momentum recently, our approach is the first supervised learning approach that is explicitly designed to integrate data from multiple single cell analysis techniques. While the state of the art in supervised learning based single cell typing is able to process integrated data when suitable data representations are provided, our approach is the first one to explicitly provide such representations
(3) Our capsule network based approach outperforms all supervised learning methods that are state of the art in single cell type classification on integrated data. This confirms that the theoretical achievement yields relevant practical advantages as well.
(4) We demonstrate that MarkerCapsule requires (substantially) less training data than prior approaches for achieving optimal performance. This enables to identify sparsely covered cell types, without incurring losses in prediction accuracy.
(5) We demonstrate that the primary capsules of the MarkerCapsule network have a clear and intuitive interpretation: we show that primary capsules reflect marker genes for cell types where marker genes are known. If marker genes are not known for a cell type, genes captured by primary capsules that are important for predicting such types are likely to suggest plausible marker genes.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

1. S Ray and A Schönhuth, Automated discovery of cell types in multi-omics scRNA-seq data using Capsule Networks (Manuscript in preparation)
2. S Ray, S Lall, A Mukhopadhyay, S Bandyopadhyay and A Schönhuth, ``Predicting potential drug targets and repurposable drugs for COVID-19 via a deep generative model for graphs", arXiv preprint arXiv:2007.02338.
3. F. Nielsen, G. Marti, S. Ray and S. Pyne, ``Clustering patterns connecting COVID-19 dynamics and Human mobility using optimal transport" arXiv preprint, arXiv:2007.10677
4. S. Ray, S.Lall and S.Bandyopadhyay, “CODC: A copula based model to identify differential coexpression”, npj System Biology and Applications,6, 20, (2020).
5. S. Pyne, S. Ray, R. Gurewitsch , M. Aruru, “Transition from Social Vulnerability to Resiliency `vis-`a-vis COVID-19”, Statistics and Applications, Vol.18, pp-197-208, (2020).

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

“Capsule Network: A new way to analyze single cell RNA-seq data”. at Institute of Informatics, University of Warsaw, Poland, 9 March, 2020.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

During my fellowship I visited the Institute of Informatics, University of Warsaw (MIMUW), Warsaw. In MIMUW, my host was DR. Ewa Szczurek. We discussed several ideas and one project, particularly in the domain of single cell clustering analysis. We are mainly interested in the clustering of immune cells collected from human cell atlas data. We discussed several ways to address the challenges of single cell downstream analysis. I was there for 14 days and we have several fruitful discussions with her research team on different possibilities and ideas for the analysis of immune cell atlas. We are now continuing the proposed project and hope to have some good work in the future.

Validated on 7/9/2020 by
Leen Stougie