UNIVERSITY OF CAMBRIDGE
Department of Biochemistry

Dr Laurent Gatto
Principal Investigator
Computational Proteomics Unit

20th May 2018

Dear Editors,

We would like you to consider our manuscript, 'A Bayesian Mixture Modelling Approach For Spatial Proteomics' for publication in PLoS Computational Biology. The manuscript is currently available on the bioRχiv preprint server[1].

Our work falls squarely within the PLoS cross-journal call for *Machine Learning in Health and Biomedicine* papers, where we propose a new machine learning method for analysing proteomics data and machine learning-driven research providing insights on biological mechanisms of protein sub-cellular localisation.

In our manuscript, we describe a novel modelling approach to analyse high-throughput mass spectrometry (MS)-based spatial proteomics data, which enables to pinpoint the sub-cellular localisation of proteins. Traditionally, prediction of protein localisation from experimental data relied on discriminative supervised machine learning methods to assign proteins of unknown location based on a set of marker proteins of know localisation. Such machine learning algorithms fail to quantify uncertainty in their allocations, which substantially limits the interpretability of the data and the underlying biological phenomena.

Here, we present a probabilistic generative model for MS-based spatial proteomics data using a mixture of multivariate Gaussian distributions. To account for proteins residing in multiple location or in un-annotated sub-cellular niches, we augment our model with an outlier component described by a heavy-tailed multivariate student-t distribution, leading us to a T Augmented Gaussian Mixture model (TAGM). Our new methods compare competitively against contemporary methods and provide superior interpretability of the spatial proteome. Bayesian inference is the natural paradigm for which to perform uncertainty quantification and is an immediate product of our approach. As a result of proteome-wide uncertainty quantification, we provide biological interpretation to an additional 20% of proteins compared to a state-of-the-art descriptive classifier.

We have validated our algorithms through evaluation of eight experimental datasets, from six different cell lines from four different species, classifying thousands of proteins to tens of sub-cellular compartments with high generalisation accuracy. We further applied the method to determine protein localisation within pluripotent mouse embryonic stem cells, providing protein sub-cellular localisation of proteins that previously could not be reliably assigned a sub-cellular niche.

The TAGM methods are implemented in the R language for statistical computing and are distributed as part of the well documented `pRoloc` Bioconductor package for spatial proteomics data analysis. The software has been available and maintained since November 2012 and, with its associated Bioconductor packages, enables to import, processing, analyse and visualise spatial proteomics data. The source of our manuscript and calculations underlying our results are available at https://github.com/lgatto/2018-TAGM-paper.

Thanking you very much in advance for considering our manuscript for submission.

On behalf of the authors,

Yours sincerely,
Dr Laurent Gatto

[1] https://doi.org/10.1101/282269