A Bioconductor workflow for the Bayesian Analysis of Spatial proteomics

true
true

Introduction

Quantifying uncertainty in the spatial distribution of proteins allows for novel insight into protein function. Many proteins live in a single location within the cell, however there are those that reside in mutiple locations and those that dynamically relocalise. Functional comparmentalisation of proteins allows the cell to control biomolecular pathways and biochemical process within the cell. Therefore, proteins with multiple localisation may have mutiple functional roles. Machine learning algorithms that fail to quantify uncertainty are unable to draw deeper insight into understanding cell biology.

We present a worflow for the Bayesian analysis of spatial proteomics using the t-augmented Gaussian mixture (TAGM) model proposed in:

> A Bayesian Mixture Modelling Approach For Spatial Proteomics Oliver M Crook, Claire M Mulvey, Paul D. W. Kirk, Kathryn S Lilley, Laurent Gatto bioRxiv 282269; doi: https://doi.org/10.1101/282269

The above manuscript provides a detailed description of the model, rigorous comparisons and testing on many spatial proteomics datasets and a case study on mouse pluripotent stem cells. Revisiting these details is not the purpose of this computational protocol, rather we present how to correctly use the software and provide step by step guidance. In breif, the TAGM model posits that each annotated sub-cellular niche can be described by a Gaussian distribution. Thus the full complement of proteins within the cell is captured as a mixture of Gaussians. The highly dynamic nature of the cell means that many proteins are not well captured by any of these multivariate Gaussian distributions, and thus the model also include a outlier component mathematically desribed as multivariate student's t distribution. The heavier tails of the t distribution allow it better capture dispersed proteins.

To perform inference in the TAGM model there are two approaches. The first allows you to produce *maximum a posteriori* estimates of posterior localisation probabilities; that is, the posterior probability that a protein localises to that class. Whilst this is a true and interpretable summary of the TAGM model it is only point estimate. For a richer analysis, we present a Markov-chain Monte-Carlo method to beform fully Bayesian inference in our model, allowing us to obtain full posterior localisation distributions.