Probabilistic modelling of protein sub-cellular localisation

Laurent Gatto[1,*], Oliver M. Crook[2]

[1] de Duve Institute - UCLouvain, Avenue Hippocrate 75, 1200 Brussels, Belgium
[2] Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, UK

*laurent.gatto@uclouvain.be http://lgatto.github.io/about

Introduction

In biology, **localisation is function** - understanding the sub-cellular localisation of proteins is paramount to comprehend the context of their functions. The cellular sub-division allows cells to establish a range of distinct micro-environments (Figure 1), each favouring different biochemical reactions and interactions and, therefore, allowing each compartment to fulfil a particular functional role.

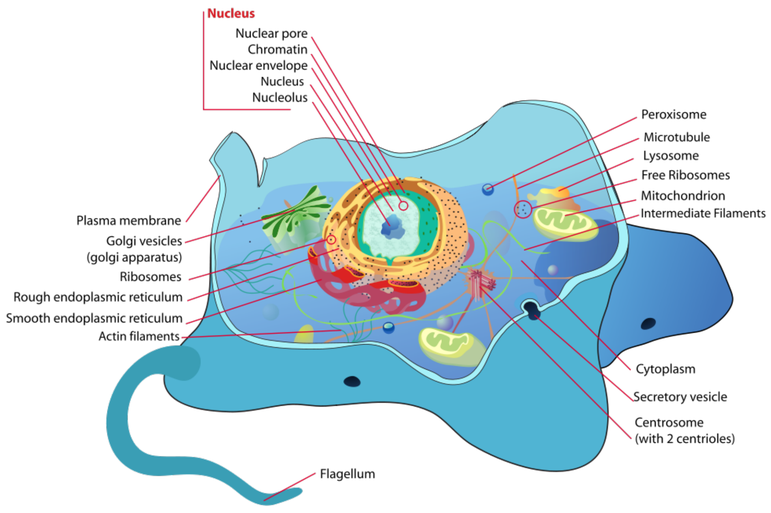

Figure 1 : Structure of an animal cell (credit: Wikipedia).

Mass spectrometry (MS)-based spatial proteomics (Figure 3) and contemporary machine learning enable to build **proteome-wide spatial maps** (Figure 2), informing us on the location of thousands of proteins.

Nevertheless, while some proteins can be found in a single location within a cell, up to half of proteins may reside in multiple locations, can dynamically re-localise, or reside within an unknown functional compartment, leading to considerable uncertainty in associating a protein to their sub-cellular location.

Recent advances enable to probabilistically model protein localisation as well as quantify the uncertainty in the location assignments, thus leading to better and more trustworthy biological interpretation of the data.

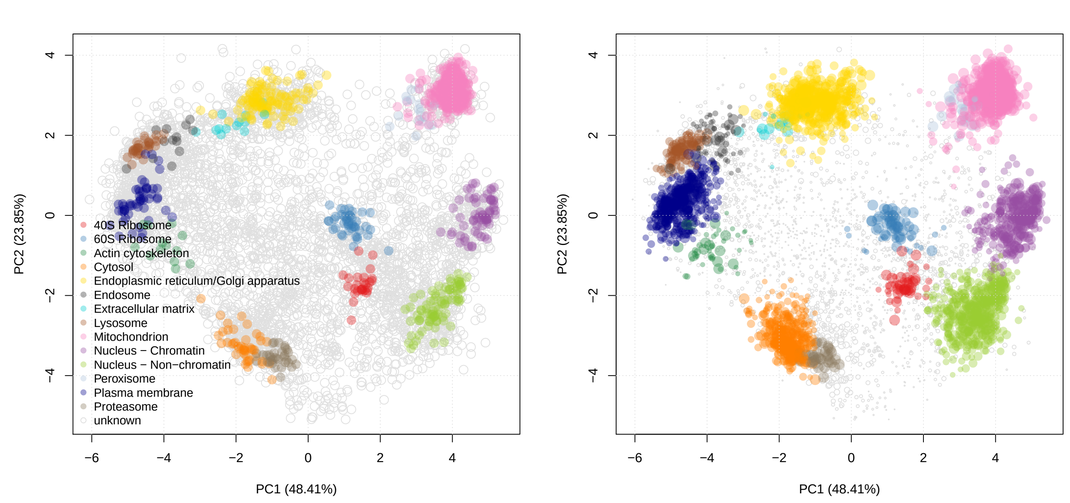

Figure 2 : Principle component analysis of the pluripotent mouse embryonic stem cell spatial map: each dot represents one protein samples along 20 gradient fractions (Figure 3). Left: among the 5032 proteins, 926 marker proteins (well known and curated proteins that can condidently assigned to a unique location) depicting 14 distinct and well resolved sub-cellular niches. Right: assignment of proteins of *unknown* location to one of the annotated classes. The size of the dots represent the assignment probability.

Mass spectrometry-based spatial proteomics

The *hyper*LOPIT protocol (Figure 3) uses density gradients to separate organelles and macro-molecular complexes. A set of discrete fractions are then collected and proteins are extracted, identified and quantified by mass spectrometry. The quantitaive proteins profiles display location specific patterns that are used for clustering and localisation analyses (classification) (Figure 2).

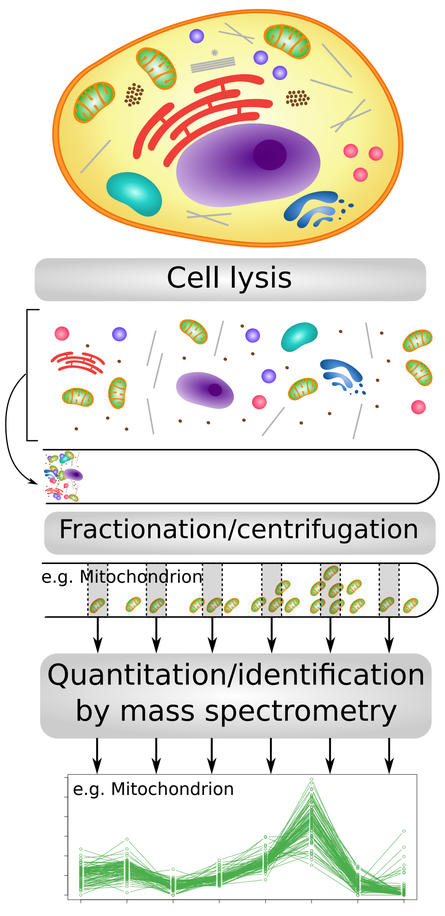

Figure 3 : LOPIT and *hyper*LOPIT separation of the cell content along a density gradient.

Results and conclusions

By implementing a probabilistic model for mass spectrometry-based spatial proteomics, we are in a position to deconvolute biologically important localisation patterns and confidently assign more proteins to their most likely sub-cellular localisations .

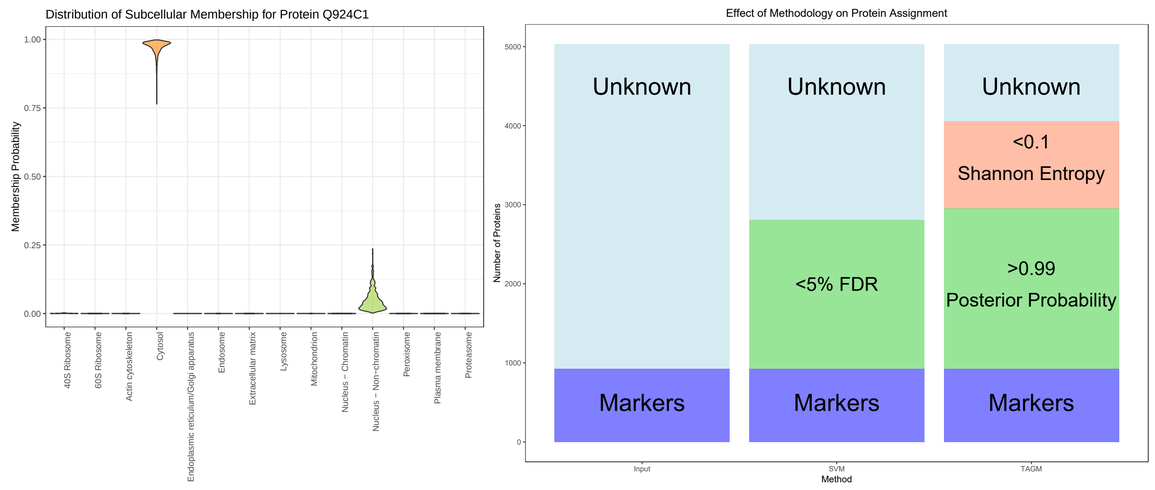

Figure 4 : More results and biologically more relevant protein localisation results.

Exportin 5 (Q924C1) forms part of the micro-RNA export machinery of the nucleus, transporting miRNA from the nucleus to the cytoplasm for further processing. It then translocates back through the nuclear pore complex to return to the nucleus. Exportin 5 can then continue to mediate further transport between nucleus and cytoplasm. The support vector machine (SVM) was unable to assign a localisation of Exportin 5, with its assignment falling below a 5% FDR to wrongly assign this protein to the proteasome. This incorrect assertion by the SVM was confounded by the similarity between the cytosol and proteasome profiles. Figure 4 (left) demonstrates, according to our model, that Exportin 5 most likely localises to the cytosol but there is some uncertainty with this assignment. This uncertainty is reflected in possible assignment of Exportin 5 to the nucleus non-chromatin and this uncertainty is a manifestation of the the fact that the function of this protein is to shuttle between the cytosol and nucleus.

Figure 4 (right) shows the number of proteins initially curated (markers), those that are confidently assigned a unique localisation using a SVM classifier with a manually-assigned 5% false discovery threshold (center) and those that are assigned with at least 99% posterior probability and low uncertainty (right). This demonstrates the effect of applying different methodologies on protein assignment when applied the mouse pluripotent embryonic stem cell data. Roughly 2000 proteins are classified using either SVM and TAGM-MCMC; however, TAGM-MCMC can draw additional conclusions about an extra 1000 proteins by quantifying uncertainty.

Probabilistic model

We present a probabilistic generative model for MS-based spatial proteomics data. Our model posits that each annotated sub-cellular niche can be modelled by a multivariate Gaussian distribution. Thus, the full complement of annotated proteins is captured by a mixture of multivariate Gaussian distributions. With the prior knowledge that many proteins are not captured by known sub-cellular niches, we augment our model with an outlier component. Outliers are often dispersed and thus this additional component is described by a heavy-tailed distribution: the multivariate Student's t-distribution, leading us to a T Augmented Gaussian Mixture model (TAGM).

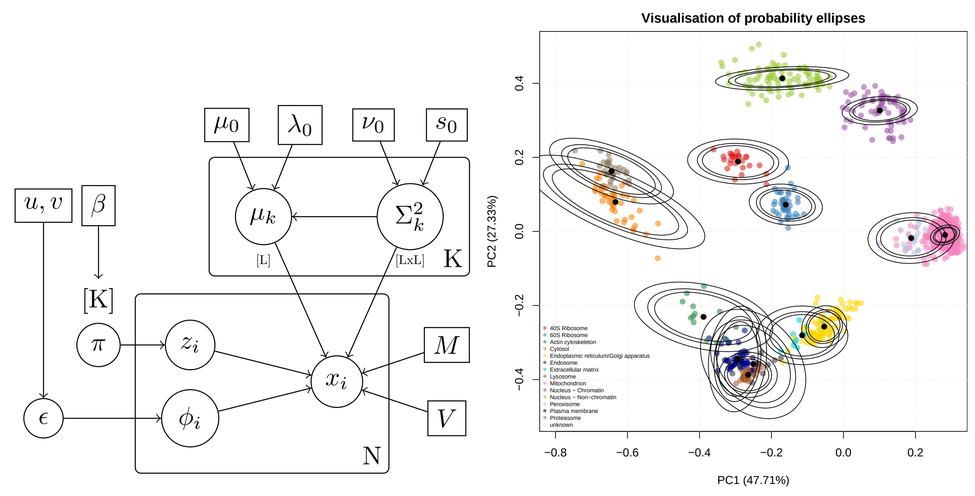

Figure 5 : Left: Plate diagram for TAGM model. This diagram specifies the conditional independencies and parameters in our model. Right: Illustration of how the model describes the data. Each ellipse contains a proportion of total probability of a particular multivariate Gaussian density. The outer ellipse contains $\mathbf{99\%}$ of the total probability whilst the middle and inner ellipses contain $\mathbf{95\%}$ and $\mathbf{90\%}$ of the probability respectively.

Funding This work was supported by a Wellcome Trust and the Biotechnology and Biological Sciences Research Council (BBSRC).

References

A draft map of the mouse pluripotent stem cell spatial proteome Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, Naake T, Gatto L, Viner R, Martinez Arias A, Lilley KS. Nat Commun. 2016 Jan 12;7:8992. doi: 10.1038/ncomms9992.

A Bayesian Mixture Modelling Approach For Spatial Proteomics Crook OM, Mulvey CM, Kirk PDW , Lilley KS, Gatto L bioRxiv 282269; doi: 10.1101/282269