A Method of Integrating Spatial Proteomics and Protein-Protein Interaction Network Data

Steven Squires, Rob Ewing, Adam Prügel-Bennett, and Mahesan Niranjan

University of Southampton, UK
{ses2g14,rob.ewing,apb,mn}@soton.ac.uk

Abstract. The increase in quantity of spatial proteomics data requires a range of analytical techniques to effectively analyse the data. We provide a method of integrating spatial proteomics data together with protein-protein interaction (PPI) networks to enable the extraction of more information. A strong relationship between spatial proteomics and PPI network data was demonstrated. Then a method of converting the PPI network into vectors using spatial proteomics data was explained which allows the integration of the two datasets. The resulting vectors were tested using machine learning techniques and reasonable predictive accuracy was found.

Keywords: Bioinformatics · Spatial proteomics · Machine learning

1 Introduction

Proteins can only perform their function in direct physical contact with other proteins or parts of the cell, therefore knowing the location of a protein (known as spatial proteomics) can aid in understanding its function. It has also been shown that there is a direct connection between diseases and subcellular protein localisation [1], consequently understanding certain diseases and cellular function depends on a reliable and accurate knowledge of protein localisation.

Protein-protein interactions (PPIs) have been studied for many years due, in part, to their importance in understanding cellular function. Interactions between pairs or groups of proteins, or proteins and other parts of the cell, have significant consequences for cell functionality including links to disease [2]. PPI networks chart these known or predicted interactions.

There is considerable interest in combining multiple sources of high throughput biological measurements. Examples include the integration of spatial and temporal patterns of gene expression [3], combining sequence and secondary structure of proteins [4], and the integrated analysis of the transcriptome and proteome [5].

Spatial proteomics and PPI networks should have significant similarities. A pair of proteins can only physically interact if they are in the same spatial location at the same time, hence we would expect that there would be a link between proteins that interact and those that share a spatial location. In principle, accurate PPI networks might be able to predict which proteins co-localise.

Conversely, spatial proteomics cannot on its own specify whether an interaction exists, but if two proteins are in the same compartment it may be more likely due to the increased likelihood that they share a function. In addition, proteins that never exist in the same spatial location cannot directly interact.

There has been work conducted which uses the PPI networks to make predictions on protein localisation. In particular, a recently published paper used PPI networks together with sequence predictors to classify proteins into spatial locations [6]. In contrast, we use the spatial proteomics profiles themselves and integrate them together with the PPI network data. In doing so we propose to aid the development of analytical tools for the analysis of proteomics data.

Our contributions are, first, to demonstrate the strong relationship between spatial proteomics and PPI network data. We then provide a mechanism to integrate the two datasets along with some useful visualisation techniques. We demonstrate that prediction of spatial localization from fractionation profiles can potentially be enhanced by the inclusion of information taken from PPI interactions and that interactions themselves are somewhat predictable from spatial profiles.

This paper is structured as follows: in Section 2 we discuss the spatial proteomics and PPI datasets along with the methods we use; in Section 3 we demonstrate the strong correlation between PPI and spatial proteomics data, the visualisation benefits of our technique, and the predictive power of the datasets. Finally, in Section 4 we provide a brief discussion of our results.

2 Methods

2.1 Spatial Proteomics and PPI Network Datasets

Spatial proteomics data is obtained from experiments which separate the contents of the cell into fractions and measure the relative abundance of each protein within each fraction. Proteins with a similar profile of fractional abundances are believed to occupy the same spatial location [7]. These proteins are then mapped to organelles by using marker proteins with similar profiles whose location is known [8, 9]. The marker proteins tend to be extracted from literature and need to be highly reliable as they set the mapping from profiles to locations. In this study we used two sets of data obtained from Arabidopsis thaliana [10] and Drosophila melanogaster [11]. Spatial proteomics data is generally of a fairly low dimensionality (usually under 10 fractions) and the datasets contain 689 and 888 proteins for Arabidopsis and Drosophila respectively. We consider two marker sets for each organism, the same as the authors use [10, 11]. The first, which we call the original marker set, are those proteins with known location extracted from literature. There are 27 for Arabidopsis and 55 for Drosophila. The authors then use the spatial proteomics datasets to assign previously unknown proteins to an organelle. We use the same assigned proteins which are named as extended marker sets in this paper.

Two PPI datasets were used: STRING [12] and BIOGRID [13]. These datasets have different methodologies to extract PPIs and present the results differently but results we have gained from both, where comparable, are consistent.

2.2 Combining Spatial Proteomics and PPI Network Data

Spatial proteomics data is produced in a format suitable for applying standard machine learning classification techniques as there is a matrix with m dimensions, n datapoints and each datapoint is within a class. In contrast the PPI data is presented as pairs of known interactions and needs to be converted into a form suitable for applying machine learning methods.

We create a simple fixed dimensional representation to capture information held in interaction networks and apply standard machine learning techniques. We do not consider more sophisticated techniques such as graph kernels [14] in this work because the amount of experimental data relating to subcellular measurements is small. Consider three organelles α, β and γ and a protein of interest, A. The PPI data for A is transformed into a three-dimensional vector by calculating the number of interactions between protein A and the marker proteins within α, β and γ respectively. Protein A is then represented as a vector with each dimension associated with an organelle. We normalise the vector by dividing by the number of proteins within each organelle and then dividing each protein individually by the sum of the vectors across each organelle. We then scale up each protein components so the sum across the PPI vector equals one.

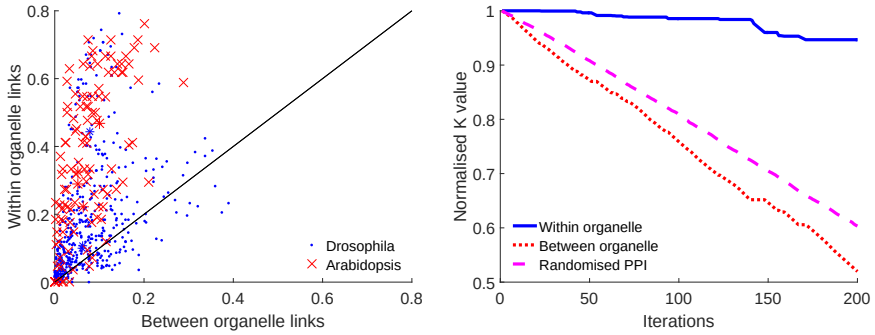
3 Results

3.1 Correlation between Spatial Proteomics and PPI Network Data

For the integration of spatial proteomics and PPI network data to be a valuable analytical technique it should first be demonstrated that they are structurally similar. A key structural similarity is the relationship between the spatial location and the chance of an interaction occurring. In the STRING database approximately 10% and 7% of proteins have links for Drosophila and Arabidopsis respectively; proteins located in the same organelle should, in general, have a higher likelihood of interacting. The ratio of interactions to potential interactions was measured for each protein in the extended marker set. Figure 1(a) shows the fraction of interactions occurring within the same organelle against the fraction to proteins in other organelles. The dots are the proteins for Drosophila and crosses for Arabidopsis, the stars are the averages for each organelle and the black line is the average expected if there was no correlation between the datasets. Any protein above the line has a higher fraction of links to proteins within its organelle than to proteins in other organelles. A majority of proteins have significantly more links within the organelles than between them.

The probability of these results occurring by chance from an uncorrelated PPI network (P-values) was calculated using the hypergeometric distribution.

Within an individual organelle, the probability of the number of interactions observed occurring by chance ranges from $P = 10^{-12}$ to $P = 10^{-404}$. We can reasonably conclude that the marker proteins in the two datasets are correlated.



(a) (b)

Fig. 1. a) The fraction of interactions for each extended marker protein within the organelle is plotted against the fraction of interactions to proteins in different organelles. The black line is the expectation if there is no correlation and the stars show the averages for the organelles. The vast majority of proteins lie above the line. b) The clustering algorithm iteratively removes links, preferentially removing PPI links which connect clusters together rather than those links that are within a cluster. The fall in number of links between proteins within the same organelle (solid line) is slower than those links between proteins in different organelles (dotted line) or a randomised PPI network (dashed line).

If similar clusters of proteins are formed in the PPI networks as in the organelle groupings in spatial proteomics data it would provide further evidence of the correlation between the data types. A clustering algorithm was developed based upon previous work [15, 6] which removes links sequentially based on the structure of the PPI network. The first links to be removed are those considered to be joining separate clusters together. If the structures of the datasets are similar we would expect the number of links between proteins in different organelles to fall off much faster than between proteins within the same organelle. The normalised K value [15] gives a measure of how well connected a group of proteins are, if it falls fast then these proteins are unlikely to be in a cluster together. The average number of links between proteins within the same organelle, K_{in}, and the equivalent number of links between proteins in different organelles, K_{out}, were calculated and normalised. In Figure 1(b) we show that the fall in K_{out} (dotted line) is far faster than for K_{in} (solid line) or for a randomised PPI network (dashed line) demonstrating that the clusters formed in the PPI data are similar to those in spatial proteomics data.

3.2 Visualisation of the Datasets

Visualisation of the spatial proteomics and PPI network data is important for both gaining a qualitative understanding of the quality of the data and for observation of potential patterns.

For spatial proteomics data the ability to classify the proteins depends on differences in the profiles for proteins in different organelles [8]. Part of our contribution is to add the PPI vectors to create vectors with additional dimensions which add extra information to the spatial proteomics data. The average vectors for the original marker proteins for each organelle together with the additional PPI dimensions are in Figure 2(a). The average profile for the PPI dimensions shows peaks at the dimension associated with each organelle (noted by their first letter on the plots). Any protein not showing a significant peak in a PPI dimension while being well classified by the spatial proteomics profile may be of particular biological interest. The organelles shown are the endoplasmic reticulum (ER), mitochondrion (Mito), plasma membrane (PM), Golgi apparatus (Gol) and the vacuole (Vac).

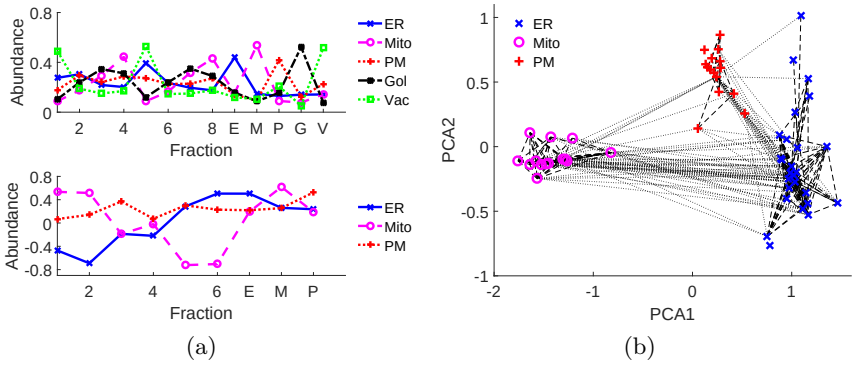
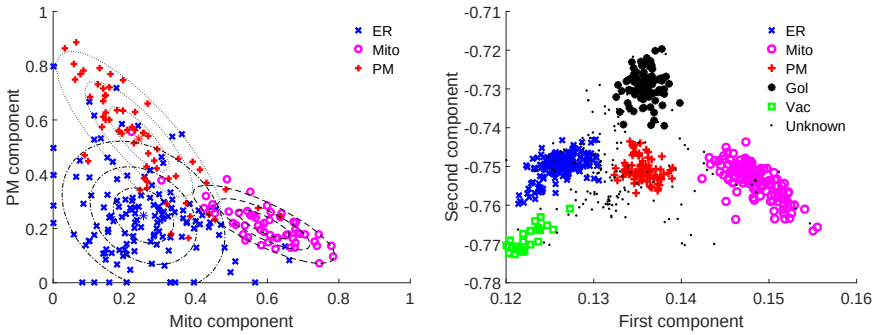


Fig. 2. a) The average profiles for the original marker proteins for Arabidopsis (top) and Drosophila (bottom) are shown. The first eight (for Arabidopsis) and six (Drosophila) fractions are from the spatial proteomics data with the remaining created from the PPI data with labels associated with the first letter of the relevant organelle. b) The original marker protein profiles for Drosophila were projected onto their principal components and PPI network links added. Proteins that have strong connections outside of their organelle can be investigated.

Visualisation of the PPI network data is difficult as the number of links are large. Here, an example of what the network looks like for the small number of original marker proteins for Drosophila is shown in Figure 2(b). The spatial proteomics profiles were projected onto their principal components and known interactions from the STRING database are shown. It may be useful to inspect the links between individual proteins and the markers in this manner to gain insight into how each protein is linked. While there are many links between

proteins in different organelles there are significantly more to proteins within the same organelle. It should also be noted that with only 55 proteins the network plots are already difficult to inspect in this format.

While the PPI networks may be difficult to visualise, the PPI vectors are somewhat easier when two components of the vector are compared. In Figure 3(a) the PM component was plotted against the Mito component for all the Drosophila expanded marker proteins. The PPI vectors were also used to create a Gaussian probabilistic model, with means (shown as stars) and covariances extracted from the vectors. Contours of equal probability are then plotted which aids with understanding the separation and structure of the data. For example, the Mito vectors are much tighter bound than the PM vectors which visually represents that the Mito proteins are more closely connected to other Mito proteins than to proteins in other organelles than the PM proteins are. The structure of the plot is exactly as would be expected with the PM and Mito proteins tending to reside high up the PM and Mito axes respectively and the ER proteins following a fairly isotropic distribution near the origin.



(a) (b)

Fig. 3. a) The PPI vectors for the PM and Mito components for Drosophila marker proteins are plotted. The PM proteins tend to cluster high up the PM axis and the Mito proteins along the Mito axis. The ER proteins are based near the origin. The contours of equal probability are centred on the average of each organelle. b) The application of a Fisher Linear Discriminant to the combined spatial proteomics and PPI vectors for Arabidopsis allows for effective partition of the extended marker proteins into their respective organelles.

We also show the effect of projection using Fisher discriminant directions for the combined vector for Arabidopsis in Figure 3(b). Most of the extended marker proteins are well separated into their organelles, the combined vector is able to effectively partition the proteins into different compartments.

3.3 Predictive Power

We have demonstrated the similarity of the two datasets and some visualisation techniques. Now we will show that there is considerable predictive power in the method of combining datasets

First we show the PPI networks can predict spatial location. As we have converted the PPI network into a vector we can apply standard machine learning techniques. The protein extended marker data was partitioned randomly into two and a support vector machine [16] (SVM) was trained on half of the proteins with the PPI vector as input and the organelles as output classes. The trained SVM was then tested on the remainder of the randomised data. The process, with different random partitions, was then repeated two hundred times. The classification accuracies, sensitivities (true positive rate) and specificities (true negative rates) are shown as boxplots in Figure 4(a). Generally, the SVM trained on the PPI data was able to predict the location of the proteins approximately 70% of the time. The most notable exception was the vacuole where very poor predictions are made. There are only small numbers of vacuole proteins in the extended marker set which is likely to be the reason for the poor predictive ability.

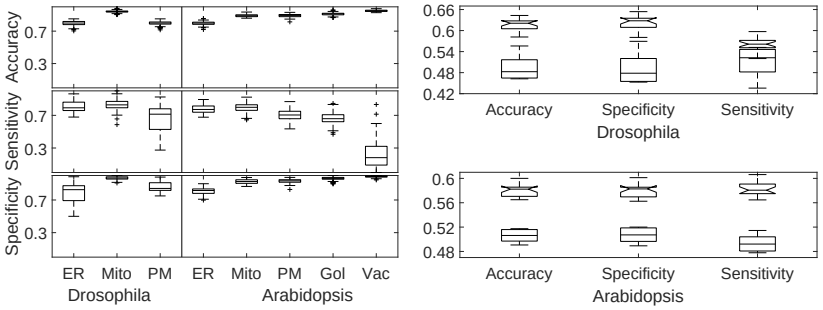


Fig. 4. a) The fraction of extended marker proteins predicted correctly from two hundred runs of an SVM classifier based on the PPI vectors as input and the organelle markers as output classes. The classifier is able to correctly predict the organelle around 70% of the time. b) The spatial proteomics data can make weak predictions on the existence of PPI links. The notched plots show the accuracy, specificity and sensitivity for Drosophila (top) and Arabidopsis (bottom) while the square plots are the results for a randomised PPI network. There is a clear (but faint) signal from the spatial proteomics data.

The inverse problem is more challenging. To attempt to use the spatial proteomics data to estimate whether a PPI exists between two proteins, first each pair of proteins were merged together to create a combined vector by multiplying each component of each vector by all the components of the other. The

six dimensional Drosophila vector, for example, was transformed to a 36 dimensional combined vector. A two-class SVM for interactions and non-interactions was then trained. The new dataset is the combination of all the protein pairs so contains 393,828 protein pairs for Drosophila and 237,016 for Arabidopsis. The data was split into training sets of protein pairs and, as the data is highly skewed towards non-interactions, the sampling was biased to force the training set to contain 50% interactions. The process was repeated twenty times. The accuracy, sensitivity and specificity of the predictions are shown in the notched plots of Figure 4(b). The equivalent SVM was applied to a randomised PPI network and shows what would be expected if there was no signal available (the non-notched plots). While the signal from the real data is small, it is consistently larger than the results from the randomised PPI network and can make some predictions about the PPI network.

4 Discussion

In this paper, we show how sub-cellular proteomics measurements can be combined with information contained in protein-protein interaction networks. Our work shows that there is significant correlation between spatial protein expression in cells and protein interaction information. Using a simple representation of interaction data in a fixed dimensional space, we show that predictions can be made in both directions between spatial proteomics and PPI networks.

There are many potential benefits from using the combined datasets. Differences in classification between the datasets may be of particular interest as it may imply interesting cases such as proteins that exist in multiple compartments or false data that should be re-evaluated. Confidence in conclusions can also be increased if the same conclusion is drawn using two separate datasets. Inspection of the data using some of the visualisation techniques discussed may also be useful for increasing understanding of data quality and building intuition.

References

1. Park, S., Yang, J.S., Shin, Y.E., Park, J., Jang, S.K., Kim, S.: Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. Molecular systems biology 7(1), 494 (2011)
2. Ideker, T., Sharan, R.: Protein networks in disease. Genome research 18(4), 644–652 (2008)
3. Samsonova, A.A., Niranjan, M., Russell, S., Brazma, A.: Prediction of gene expression in embryonic structures of drosophila melanogaster. PLoS computational biology 3(7), e144 (2007)
4. Wieser, D., Niranjan, M.: Remote homology detection using a kernel method that combines sequence and secondary-structure similarity scores. In silico biology 9(3), 89–103 (2009)
5. Gunawardana, Y., Fujiwara, S., Takeda, A., Woo, J., Woelk, C., Niranjan, M.: Outlier detection at the transcriptome-proteome interface. Bioinformatics 31(15), 2530–2536 (2015)

6. Du, P., Wang, L.: Predicting human protein subcellular locations by the ensemble of multiple predictors via protein-protein interaction network with edge clustering coefficients. PloS one 9(1), e86879 (2014)
7. De Duve, C., Beaufay, H.: A short history of tissue fractionation. The Journal of cell biology 91(3), 293 (1981)
8. Gatto, L., Breckels, L.M., Burger, T., Nightingale, D.J., Groen, A.J., Campbell, C., Mulvey, C.M., Christoforou, A., Ferro, M., Lilley, K.S.: A foundation for reliable spatial proteomics data analysis. Molecular & Cellular Proteomics pp. mcp–M113 (2014)
9. Itzhak, D.N., Tyanova, S., Cox, J., Borner, G.H.: Global, quantitative and dynamic mapping of protein subcellular localization. Elife 5, e16950 (2016)
10. Dunkley, T.P., Hester, S., Shadforth, I.P., Runions, J., Weimar, T., Hanton, S.L., Griffin, J.L., Bessant, C., Brandizzi, F., Hawes, C., et al.: Mapping the arabidopsis organelle proteome. Proceedings of the National Academy of Sciences 103(17), 6518–6523 (2006)
11. Tan, D.J., Dvinge, H., Christoforou, A., Bertone, P., Martinez Arias, A., Lilley, K.S.: Mapping organelle proteins and protein complexes in drosophila melanogaster. Journal of proteome research 8(6), 2667–2678 (2009)
12. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., et al.: String 8a global view on proteins and their functional interactions in 630 organisms. Nucleic acids research 37(suppl_1), D412–D416 (2008)
13. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: Biogrid: a general repository for interaction datasets. Nucleic acids research 34(suppl_1), D535–D539 (2006)
14. Kondor, R.I., Lafferty, J.: Diffusion kernels on graphs and other discrete input spaces. In: ICML. vol. 2, pp. 315–322 (2002)
15. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of the United States of America 101(9), 2658–2663 (2004)
16. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013)