TECHNICAL BRIEF

Improved sub-cellular resolution *via* simultaneous analysis of organelle proteomics data across varied experimental conditions

Matthew W. B. Trotter[1], Pawel G. Sadowski[2], Tom P. J. Dunkley[2]*, Arnoud J. Groen[2] and Kathryn S. Lilley[2]*

[1] Anne McLaren Laboratory for Regenerative Medicine and Department of Surgery, University of Cambridge, Cambridge, UK
[2] Cambridge Centre for Proteomics Cambridge Systems Biology Centre, Department of Biochemistry, University of Cambridge, Cambridge, UK

Received: June 10, 2010
Revised: August 5, 2010
Accepted: August 22, 2010

Spatial organisation of proteins according to their function plays an important role in the specificity of their molecular interactions. Emerging proteomics methods seek to assign proteins to sub-cellular locations by partial separation of organelles and computational analysis of protein abundance distributions among partially separated fractions. Such methods permit simultaneous analysis of unpurified organelles and promise proteome-wide localisation in scenarios wherein perturbation may prompt dynamic re-distribution. Resolving organelles that display similar behavior during a protocol designed to provide partial enrichment represents a possible shortcoming. We employ the Localisation of Organelle Proteins by Isotope Tagging (LOPIT) organelle proteomics platform to demonstrate that combining information from distinct separations of the same material can improve organelle resolution and assignment of proteins to sub-cellular locations. Two previously published experiments, whose distinct gradients are alone unable to fully resolve six known protein–organelle groupings, are subjected to a rigorous analysis to assess protein–organelle association *via* a contemporary pattern recognition algorithm. Upon straightforward combination of single-gradient data, we observe significant improvement in protein–organelle association *via* both a non-linear support vector machine algorithm and partial least-squares discriminant analysis. The outcome yields suggestions for further improvements to present organelle proteomics platforms, and a robust analytical methodology *via* which to associate proteins with sub-cellular organelles.

Keywords:
Bioinformatics / Organelle proteomics / Protein localisation / Statistical models / Support vector machines

The spatial organisation of proteins according to their function and location is an important determinant of the specificity of their molecular interactions [1]. Accordingly, the determination of sub-cellular protein location(s) can elucidate a protein's role within the cell and refine knowledge of cellular processes by pinpointing certain activities to specific organelles [2]. Traditional methods to assign proteins to sub-cellular locations are predominantly targeted to a single protein of interest, for example, by creating a GFP-tagged version or raising a specific antibody. Their low-throughput nature has prevented such methods from reaching genome-wide coverage, apart from in a very few

Correspondence: Dr. Matthew W. B. Trotter, Anne McLaren Laboratory for Regenerative Medicine and Department of Surgery, University of Cambridge, West Forvie Building, Robinson Way, Cambridge, UK
E-mail: mwbt2@cam.ac.uk
Fax: +44-1223-763350

Abbreviations: LOPIT, Localisation of Organelle Proteins by Isotope Tagging; **PLSDA,** partial least-squares discriminant analysis; **RBF,** radical basis function

*Current addresses: Dr. Pawel G. Sadowski, Skirball Institute, 5, Lab 18, 540 First Avenue, New York, NY 10016, USA;
Dr. Tom P. J. Dunkley, Astra Zeneca, Alderley Edge, Cheshire, UK

Colour Online: See the article online to view Figs.1 and 4 in colour.

cases where great effort resulted in highly resource-rich studies on very specific systems [3, 4].

Conversely, organelle proteomics involves isolating an organelle of interest and producing a catalogue of the proteins present in that organelle, *via* some form of protein/peptide separation followed by identification using MS. Recent high-throughput methods (summarised in Fig. 1) seek to obviate organelle purification. Quantitative proteomics is employed to characterise the phenotypic distribution of organelles among partially enriched fractions generated by various separation technologies, thereby providing potential to discriminate between genuine organelle residents and contaminants. Some methods seek an enrichment of certain organelle proteins within a small number of highly refined fractions [5–7], but most do not achieve purification of any organelle and may yield noisy data contaminated by false assignments from aberrant purification.

Alternative methods assume that residents of a specific organelle have a characteristic distribution pattern (or profile) along the gradient (first proposed by de Duve [8]), with location determined by matching the gradient profile of a query protein to the profiles of proteins with known sub-cellular location. For example, Protein Correlation Profiling (Fig. 1A) employs label-free quantification methodologies to determine distribution profiles across density gradients [9–11]. Localisation of Organelle Proteins by Isotope Tagging (LOPIT; Fig. 1A) seeks to improve quantification *via* the use of isobaric stable isotope labelling technologies to measure relative protein abundance across a density gradient [12–15]. Protein distributions are approximated by measuring relative abundance among fractions *via* iTRAQ quantification.

Both methods yield a multivariate vector (profile) that approximates protein occupancy along a density gradient designed to separate sub-cellular compartments. The collective isotopic abundance profiles of proteins with known organelle membership within the cells under investigation may be employed to create a phenotypic representation of gradient occupancy for the proteins of a particular organelle. Profile phenotypes may be employed subsequently to assign organelle membership of other proteins present on the gradient. The abstract nature of the experimental output renders its use reliant on the secondary process of mapping between the property measured (gradient occupancy) and the knowledge desired (protein localisation). Thus, methods based on the de Duve principle have an intimate relationship with the field of pattern recognition.

To date, multivariate statistical techniques, including PCA and partial least squares discriminant analysis (PLSDA) [12, 15], have been employed to visualise LOPIT output, cluster proteins according to their gradient profiles and discriminate between the phenotypic profiles of known organelle markers to assign protein–organelle membership. The success of predictive multivariate analysis when applied to the output of organelle proteomics is reliant upon (i) the presence of sufficiently well-described phenotypes for organelles of interest within the data and (ii) the ability to distinguish between, or resolve, these phenotypes across the gradient employed. This work assumes the former and examines the latter.

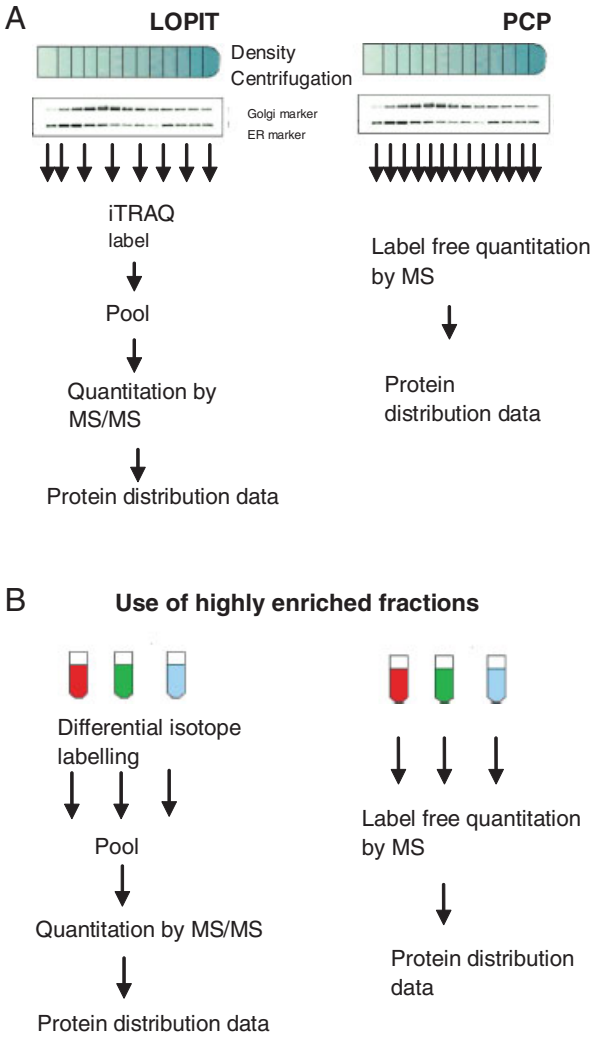

Figure 1. Diagrams of the experimental schema of the common approaches to determining sub-cellular location of organelle proteins by proteomics approaches. The two approaches shown in (A), LOPIT [12] and Protein Correlation Profiling [10], involve partial separation of organelles by density centrifugation and then assessment of the distribution patterns on thousands of proteins among fractions from these gradients using quantitative proteomics technologies. The approaches shown in (B) involve the creation of highly enriched fractions of specific organelles and comparison using proteomics of the enrichment of groups of organelle specific proteins between these fractions [5, 6, 7].

In eukaryotic systems, some organelles, including nucleus, mitochondria and chloroplasts, are relatively easy to obtain in a pure form, whereas many endomembrane organelles are impossible to purify without considerable contamination from other organelles with similar densities.

Similarly, organelle resolution is limited by the gradients applied during partial organelle separation as described above. Present attempts at improving resolution over greater numbers of organelles, towards proteome-wide protein localisation studies, have focused upon an improved approximation to gradient occupancy. For example, original LOPIT methodologies employed four isotopes across four distinct gradient fractions [14], but present methods either employ dual use of four isotopes across eight fractions (duplex method) or single use of a wider range of distinct isotope labels, *e.g.* 6-plex TMT [16] or 8-plex iTRAQ tags [17]. Where the gradient formed results in the identical distribution of organelles with different physical properties; however, ever greater resolution of gradient occupancy will not prevent the profiles of proteins from unresolved organelles having similar phenotypic appearance. Accordingly, even the most sophisticated data analysis will result in superimposition of such profiles and an inability to assign accurately the associated proteins to their organelles.

Here, we present a simple alternative approach to increase organelle resolution from gradient-based separation experiments. Rather than solely seeking to improve the approximation of gradient occupancy, we suggest parallel density centrifugation experiments on identical organelle preparations, wherein the separation gradients employed have distinct density distributions [13]. In a situation wherein no single density condition will achieve optimum resolution of organelles, each of several sufficiently different gradients will resolve certain organelles more optimally than others. Their combination, therefore, should resolve organelle-related profile subsets that are not fully resolved in any individual experiment (*cf.* Fig. 2). We test this assertion *via* a straightforward assessment of computational protein–organelle assignment on data from LOPIT experiments performed on the same biological sample but using different separation gradients. We demonstrate that combination of LOPIT profiles obtained across distinct separation gradients has considerable potential to assist computational protein–organelle assignment. Furthermore, we provide also the first rigorous performance assessment of computational pattern recognition applied to this type of organelle proteomics data and apply a powerful, contemporary technique to obtain protein–organelle associations.

Two Arabidopsis callus LOPIT data sets were collected as described by Dunkley *et al.* (2006) and Sadowski *et al.* (2008) [12, 13]. Each data set results from the application of a different iodixanol equilibrium density gradient (*cf.* Fig. 3) to identically prepared biological samples. The former employs a shallower gradient than the latter, and the two experiments are referred to hereon as shallow and steep, respectively.

The gradient occupancy profiles of both data sets comprise two sets of isotope abundance measurements, each taken from four distinct gradient fractions (duplex experimental protocol), thereby yielding eight values *per* profile. The proteins whose profiles are analysed here (i) appear in both experiments and (ii) are annotated to one of

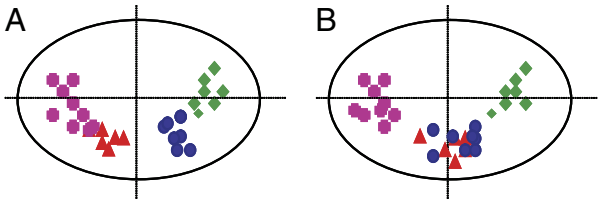

Figure 2. Hypothetical principle component analysis of LOPIT data from two different organelle separation conditions demonstrating potential issues with co-separation of organelles with similar physical properties selected for any given separation procedure. In plot (A), the organelles represented by triangles and pink circles are poorly resolved; in plot (B), the organelles represented by triangles and pink circles are now more fully resolved at the expense of the resolution of the organelles represented by blue circles from that represented by triangles.

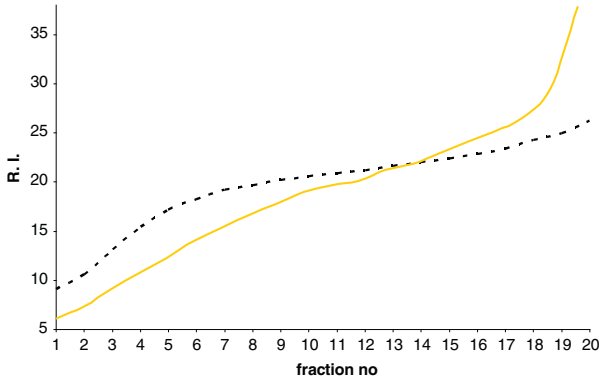

Figure 3. The density profiles of the corresponding gradients used in this study from Dunkley *et al.* (dashed line) and Sadowski *et al.* (solid yellow line). The *y*-axis is refractive index (RI) in brix.

six organelles, as evidenced from annotation within the TAIR8 database (http://www.arabidopsis.org/). The organelles present are endoplasmic reticulum (ER), plasma membrane (PM), Golgi apparatus (GA), mitochondria (MT), vacuole (VA) and plastids (PT). Organelle membership frequencies are displayed in Table 1 and the organelle-annotated profiles of both experiments are visualised by multi-dimensional scaling [18] in Fig. 4, which suggests that shallow and steep gradients resolve certain organelles better than others. For example, profiles from the mitochondria and plastids appear less similar in the steep gradient data projection than they do in the shallow gradient data. Conversely, the endoplasmic reticulum and vacuolar profiles appear less similar in the shallow gradient data projection than they do in that of the steep gradient. Interestingly, as suggested in the introduction, the same visualisation of profiles formed by a concatenation of corresponding shallow and steep profiles appears to obviate these two cases and would appear to suggest improved *general* resolution of all organelles present.

To assess the effect of combining LOPIT data obtained from different gradients, we implemented a straightforward

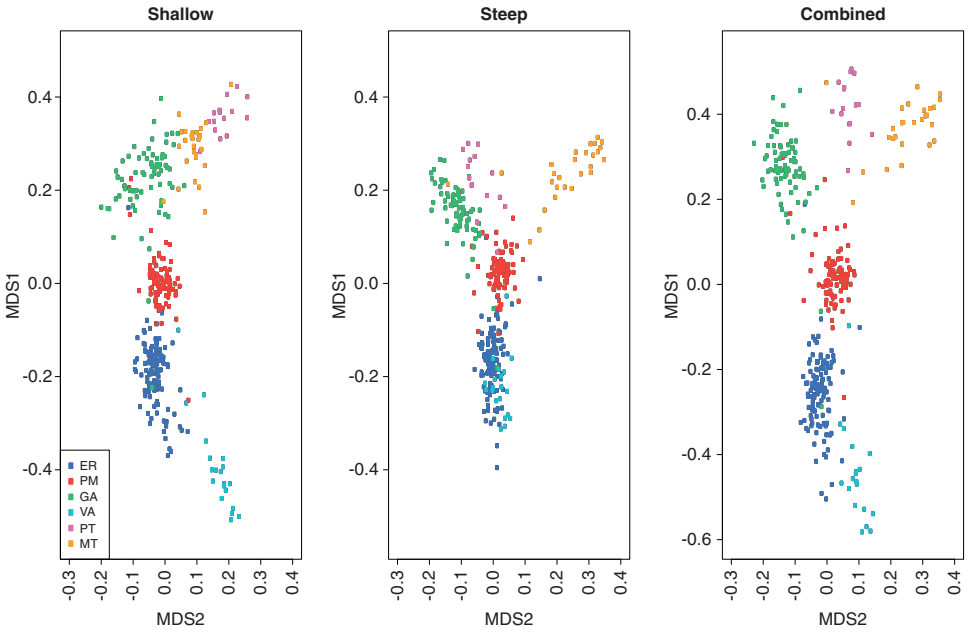

Figure 4. Multi-dimensional scaling plots of annotated LOPIT profiles from shallow (left), steep (centre) and combined (right) gradients. Inter-profile dissimilarity is represented by Euclidean distance. Plots represent projection of multivariate (eight-dimensional) LOPIT profiles onto two dimensions that most account for the original inter-profile dissimilarity.

Table 1. Protein organelle membership frequencies

	ER	PM	GA	MT	VA	PT
Proteins	119	89	76	27	20	16

assessment of computational protein–organelle association when applied to (i) profiles obtained from the shallow gradient, (ii) profiles obtained from the steep gradient and (iii) profiles formed by concatenating corresponding shallow and steep profiles. Successful protein–organelle association was assessed as follows. For each data set, protein profiles were split into stratified training (80%) and test (20%) partitions, respectively, *via* uniform random sampling without replacement from the profiles of each organelle (or data class). Profiles were mapped to known organelle associations (classifier creation) on the training partition and performance of the mapping (estimated generalisation) assessed by using the mapping to predict organelle associations of profiles in the test partition. The partitioning was repeated to create 100 independently selected partitions. The same 100 partitions were employed across shallow, steep and combined data sets.

The generalisation performance of protein–organelle associations made by mappings learned from each training partition was estimated using the category-averaged (or macro) F1 measure [19] to assess predictions made when mappings were applied to predict protein–organelle association on the corresponding test partition. The F1 measure represents the harmonic mean of precision (ratio of correct associations with a particular organelle to number of profiles actually associated to that organelle) and recall (ratio of correct associations with a particular organelle to all associations made with that organelle). The macro-F1 measure, which has interval [0, 1], is the mean F1 measure over all organelles present in the test data. The significance of any observed differences in median estimated generalisation performance across all 100 test partitions was assessed using a paired Wilcoxon rank-sum test of median equality [20].

Protein–organelle associations were obtained using a non-linear support vector machine (SVM) classifier with radial basis function (RBF) kernel function [21]. An SVM seeks to separate the examples of distinct data classes *via* a hyperplane (or hyperplanes) located at maximal distance to the examples it separates, thereby often referred to as being a large-margin classifier. Furthermore, exploitation of the 'kernel trick' – whereby linear solutions are performed in a non-linear transformation of the original data space – places it among a group of learning algorithms known as kernel methods. The SVM algorithm has become familiar across a wide range of pattern recognition applications, owing primarily to its theoretically optimum learning strategy, robustness to noise and high data cardinality, and ease of adaptation to non-linear scenarios [22, 23]. The SVM classifier employed here is that implemented in the kernlab package for the *R* statistical programming environment (http://www.r-project.org). The prediction of multiple classes was handled using the *kernlab* implementation of a native multi-class SVM formulation [24]. On each training partition, the RBF kernel width was set heuristically, as described by [25, 26], and the regularisation parameter *C* selected according to macro-F1 performance over a single round of stratified fivefold cross-validation [27]. Classifier training incorporated weights inversely proportional to relative training class proportions so as to ensure

balanced classification performance, *e.g.* [28]. Additional detail regarding experimental practice is available in Supporting Information, along with tables of the LOPIT profiles analysed.

The boxplot of Fig. 5 displays estimated generalisation performance of protein–organelle associations over the data partitioned as described above. The median macro-F1 performance of the SVM classification framework applied to shallow, steep and combined gradient profiles was 0.917, 0.892 and 0.955, respectively. Over the 100 paired test data partitions of each experiment, median estimated generalisation performance on the combined gradient profiles was significantly higher than that obtained on profiles obtained solely from shallow ($p = 2.14e{-}14$) or steep ($p < 2.20e{-}16$) gradients. Mean F1 measure (and associated standard deviation) for individual organelles are displayed in Table 2 for further detail. It is apparent that, as suggested by the plots of Fig. 4, predictive performance on combined gradient profiles is similar to, or slightly higher than, the best performance among single gradients.

Finally, our choice of base classifier in the above assessment allows us to compare the potential benefit, albeit on a relatively limited number of profiles covering relatively few organelles, of applying a non-linear SVM with performance obtained using PLSDA, as employed in several previous LOPIT analyses. Accordingly, a PLSDA model implemented *via* the caret package for *R*, with Bayesian classification schema and number of components selected using leave-one-out cross-validation on each training partition, was applied to the assessment schema described above. The outcome is described by the boxplot of Fig. 6 and confirms slight but significant performance advantage of the non-linear SVM. The median estimated generalisation performance of PLSDA, in terms of macro-F1 measure across the same 100 test partitions, is 0.896, 0.823 and 0.932 for shallow, steep and combined profiles, respectively. By comparison, the figures reported above for the SVM are higher to a significant extent ($p = 1.93e{-}4$, $p < 2.20e{-}16$ and $p = 1.33e{-}9$, respectively). It is interesting, however, that significant improvement in PLSDA performance is recorded on the combined profiles when compared to shallow ($p = 3.53e{-}7$) and steep ($p < 2.2e{-}16$) profiles, which suggests that the benefit of combining the distinct gradients of our two experiments is not solely available *via* use of the more sophisticated non-linear classifier.

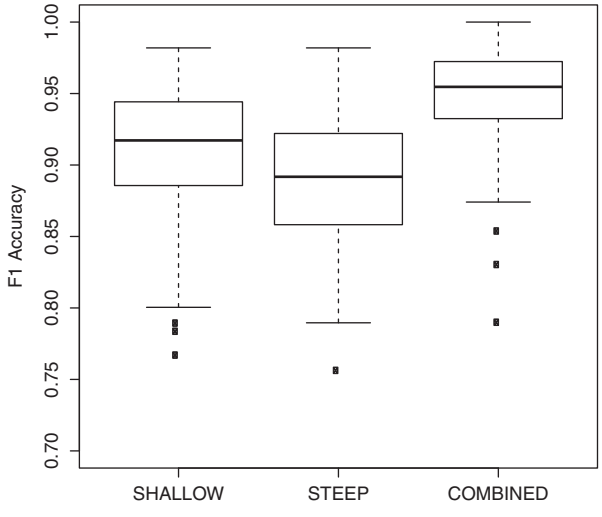

Figure 5. Boxplots, displaying median and inter-quartile range of estimated generalisation performance (macro-F1 measure) over 100 test partitions, for SVM applied to shallow, steep and combined LOPIT profiles.

We have assessed computational protein–organelle association from organelle proteomics data in the circumstance wherein a single experiment is unable to fully resolve the organelles present along a fractionation gradient. Our experimental results suggest that there may be significant benefit to protein–organelle association using our LOPIT platform by combining localisation data obtained across distinct gradients. To demonstrate this approach, we implemented a rigorous analytical platform for the creation and assessment of computational protein–organelle mappings created from fractionation-based platforms. Furthermore, we employed a powerful, contemporary pattern recognition algorithm in order to create the protein–organelle mappings reported.

Our approach to combination (or data fusion) of distinct experimental outputs is simple (concatenation), but appears highly effective in the scenario presented. Going forward, there remains a need to develop the method, (i) in a manner more sophisticated when confronted by an inevitable increase in the number of different gradients, and therefore data dimensionality, required in order to resolve a greater number of organelles towards proteome-wide protein–organelle association studies, and (ii) in a manner that affords organelle association to proteins that do not appear in every gradient employed.

The use of an SVM classifier, widely-reported as being robust to relatively high data cardinality, to map multivariate protein profiles to organelle memberships is already sympathetic to the first challenge. Ultimately, however, increasingly complex gradient representations suggest an eventual need for classifier fusion, rather than data fusion. One can envisage an ensemble of classifiers voting on protein–organelle membership over the output of several distinct gradients (which could also incorporate an approach to overcome the second challenge of missing information from one or more gradients). Further to this, a more integral approach to learning over the output of several gradients could involve data fusion internal to a single classifier. For example, the application of an SVM (or other matrix-based classifier, such as k-nearest neighbour [29]) to a combined kernel (or dissimilarity) matrix. The weighted combination of multiple kernel matrices provided by the Multiple Kernel Learning algorithm [30, 31] would appear well suited to such an approach.

Returning to the need for greater organelle resolution, a more complex aspect of protein–organelle association from

Table 2. Mean F1 measure (and standard deviation) for individual organelles

	ER	GA	MT	PM	PT	VA
Shallow	0.960 (0.025)	0.936 (0.040)	0.870 (0.102)	0.922 (0.037)	0.851 (0.134)	0.911 (0.111)
Steep	0.911 (0.040)	0.947 (0.038)	0.926 (0.079)	0.910 (0.045)	0.897 (0.126)	0.742 (0.135)
Combined	0.968 (0.023)	0.953 (0.034)	0.937 (0.078)	0.940 (0.033)	0.967 (0.062)	0.919 (0.112)

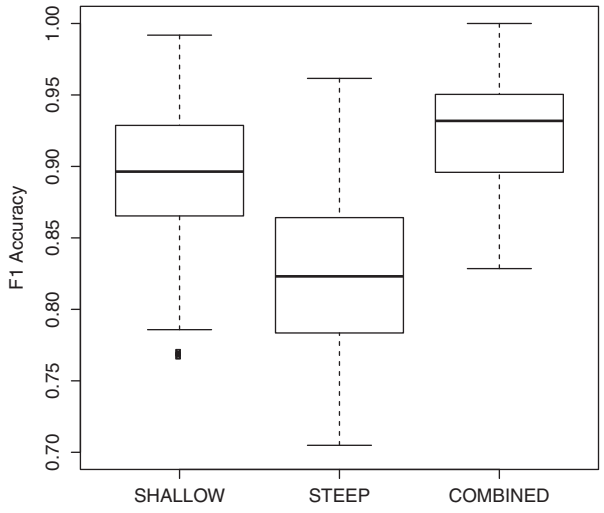

Figure 6. Boxplots, displaying median and inter-quartile range of estimated generalisation performance (macro-F1 measure) over 100 test partitions, for PLSDA applied to shallow, steep and combined LOPIT profiles.

gradient-based studies, which is not tackled here, is that of multiple protein–organelle associations. For example, it is reasonable to assume that proteins located simultaneously in more than one organelle in a particular cellular condition will have gradient abundance profiles that comprise a weighted superimposition of multiple organelle profile phenotypes. Analytically, this scenario may be overcome *via* some form of deconvolution according to the organelle phenotypes present or by probabilistic approaches to protein–organelle mapping. Of potentially greater importance, however, are ensuing practical issues concerning the ability of a single density gradient to not only resolve individual organelles but to do so in a manner that distinguishes the phenotypic profiles of single organelles from the convoluted profiles that arise from multiple organelle residency. Clearly, the task of protein localisation *via* organelle proteomics platforms provides immediate research challenges both practical and analytical.

Regardless of specific future approaches to improve accuracy and resolution in current high-throughput protein localisation platforms, the general theme should be to develop current experimental and analytical protocols for a growing need to investigate protein localisation simultaneously across the entire proteome in a variety of different organisms, cell types and experimental conditions. Application of robust, flexible algorithms, such as the non-linear SVM employed here, to map protein gradient profiles to organelle membership represents a good starting point. The parallel development of present experimental protocols, for example, the use of multiple gradients to resolve physically a larger number of organelles, alongside refined analytical practice promises practical, accurate and high-throughput purification-free approaches to the proteome-wide assignment of proteins to their sub-cellular locations.

Support for this work was provided by grants from EU framework 6 WallNet Consortium awarded to P. G. S. and ERA-PG Consortium (BBSRC grant BB/E024777/1) awarded to A. J. G., and support to M. W. B. T. by an MRC Centre Grant (MRC Centre for Stem Cell Biology and Regenerative Medicine, University of Cambridge). The authors wish to thank Dr. Sean B. Holden, University of Cambridge Computer Laboratory, for reading a revised version of the manuscript.

The authors have declared no conflict of interests.

References

[1] Dreger, M., Subcellular proteomics. *Mass Spectrom. Rev.* 2003, *22*, 27–56.

[2] Lilley, K. S., Dupree, P., Plant organelle proteomics. *Curr. Opin. Plant Biol.* 2007, *10*, 594–599.

[3] Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S. *et al.*, Global analysis of protein localization in budding yeast. *Nature* 2003, *425*, 686–691.

[4] Barbe, L., Lundberg, E., Oksvold, P., Stenius, A. *et al.*, Towards a confocal subcellular atlas of the human proteome. *Mol. Cell. Proteomics* 2008, *7*, 499–508.

[5] Gilchrist, A., Au, C. E., Hiding, J., Bell, A. W. *et al.*, Quantitative proteomics analysis of the secretory pathway. *Cell* 2006, *127*, 1265–1281.

[6] Lam, Y. W., Lamond, A. I., Mann, M., Andersen, J. S., Analysis of nucleolar protein dynamics reveals the nuclear degradation of ribosomal proteins. *Curr. Biol.* 2007, *17*, 749–760.

[7] Andreyev, A. Y., Shen, Z., Guan, Z., Ryan, A. *et al.*, Application of proteomic marker ensembles to subcellular organelle identification. *Mol. Cell. Proteomics* 2010, *9*, 388–402.

[8] de Duve, C., Tissue fractionation. *J. Cell Biol.* 1971, *50*, 20D–55D

[9] Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P. *et al.*, Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 2003, *426*, 570–574.

[10] Foster, L. J., de Hoog, C. L., Zhang, Y. L., Zhang, Y. *et al.*, A mammalian organelle map by protein correlation profiling. *Cell* 2006, *125*, 187–199.

[11] Wiese, S., Gronemeyer, T., Ofman, R., Kunze, M. *et al.*, Proteomics characterization of mouse kidney Peroxisomes by tandem mass spectrometry and protein correlation profiling. *Mol. Cell. Proteomics* 2007, *6*, 2045–2057.

[12] Dunkley, T. P. J., Hester, S., Shadforth, I. P., Runions, J. *et al.*, Mapping the *Arabidopsis* organelle proteome. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 6518–6523.

[13] Sadowski, P. G., Groen, A. J., Dupree, P., Lilley, K. S., Subcellular localization of membrane proteins. *Proteomics* 2008, *8*, 3991–4011.

[14] Tan, D. J. L., Dvinge, H., Christoforou, A., Bertone, P. *et al.*, Mapping Organelle Proteins and Protein Complexes in Drosophila melanogaster. *J. Proteome Res..* 2009, *8*, 2667–2678.

[15] Hall, S. L., Hester, S., Griffin, J. L., Lilley, K. S., Jackson, A. P., The organelle proteome of the DT40 lymphocyte cell line. *Mol. Cell. Proteomics* 2009, *8*, 1295–1305.

[16] Thompson, A., Schafer, J., Kuhn, K., Kienle, S. *et al.*, Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 2003, *75*, 1895–1904.

[17] Ow, S. Y., Cardona, T., Taton, A., Magnuson, A. *et al.*, Quantitative shotgun proteomics of enriched heterocysts from Nostoc sp. PCC 7120 using 8-plex isobaric peptide tags. *J. Proteome Res.* 2008, *7*, 1615–1628.

[18] Kruskhal, J. B., Multi-dimensional scaling by optimising goodness of fit to a non-metric hypothesis. *Psychometrika* 1964, *29*, 1–27.

[19] van Rijsbergen, C. J., *Information Retrieval*, Butterworths, London 1979.

[20] Mann, H. B., Whitney, D. R., On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 1947, *18*, 50–60.

[21] Schölkopf, B., Sung, K., Burges, C., Girosi, F. *et al.*, Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.* 1997, *45*, 2758–2765.

[22] Vapnik, V. N., *The Nature of Statistical Learning Theory*, Srpinger, New York, 1995.

[23] Cristianini, N., Shawe-Taylor, J., *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. Cambridge University Press, Cambridge 2000.

[24] Weston, J., Watkins, C., *Multi-class support vector machines,* in: Verleysen, M. (Ed.) *Proceedings of ESANN99*, Brussels 1999.

[25] Caputo, B., Sim, K., Furesjo, F., Smola, A., Appearance-based object recognition using SVMs: which Kernel should i use? in Proceedings of the NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision, Whistler 2002.

[26] Karatzoglou, A., Smola, A., Hornik, K., Kernlab – An S4 Package for Kernel Methods in R, 2010, (http://cran.r-project.org/web/packages/kernlab/vignettes/kernlab.pdf).

[27] Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceeding of the 14th International Joint Conference on Artificial Intelligence,* Vol. 2, Montreal, Quebec, Canada1995, pp. 1137–1145.

[28] Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Knowledge based analysis of microarray gene expression data by using support vector machines. *Proc Natl. Acad. Sci. USA* 2000, *97*, 262–267.

[29] Cover, T. M., Hart, P. E., Nearest neighbour pattern classification. *IEEE Trans. Inform. Theory* 1967, *13*, 21–27.

[30] Rätsch, G., Sonnenburg, S., Schäfer, C., Learning interpretable SVMs for biological sequence classification. *BMC Bioinformatics* 2006, *7*, 9.

[31] Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y., SimpleMKL. *J. Machine Learn. Res.* 2008, *9*, 2491–2521.

