Predicting function of plant polysaccharides synthesis enzymes

P. Dupree, K. Evans, S.P. Preston, M. Segura and S. White

July 31, 2009

1 Background

Polysaccharides are the main constituents of the cell wall of plants. They are highly diverse, which suggests that many enzymes participate in their synthesis, and estimations based on plant genomes and homology of sequences reveal that a large number of protein families are indeed involved.

Determining gene function experimentally is laborious and time consuming, so when searching for genes that perform a specific function it is useful to prioritise candidates according to some measure of "closeness" to genes whose function is known. A promising way to prioritise is according to their localisation in cell compartments, since cellular functions tend to be performed by proteins grouped within the same organelle.

In the laboratory a technique called LOPIT has been developed to localise in parallel all the proteins of a plant sample. LOPIT translates the localisation problem into a classification of "traces" of proteins as they migrate in a separation medium. The localisation of unknown proteins to their possible organelles can be predicted through comparison with proteins of known localisation (Gold standards).

Data output from LOPIT consists of a collection of traces where each trace corresponds to the measurement of abundance of a protein across 7 samples. Then, a typical data matrix contains around 1000 proteins (rows) and 7 samples (columns).

Another way to prioritise candidates comes from the fact that proteins working together are in many cases associated in larger functional aggregates or complexes. This kind of association can be inferred from a second experimental technique called ProCoDeS. During ProCoDes proteins are also separated as in LOPIT, but organelle membranes are previously eliminated so the co-migration of proteins is given mainly by the physical attachment that exists between proteins. In ProCoDeS the similarity of traces suggests a physical link between the proteins and hence it can reveal which proteins are in the same complex.

So far, the identification of protein complexes from the data has been done by fitting peaks to the traces and then comparing the relative migration of these peaks. The structure of ProCoDeS data is similar to that of LOPIT data.

2 Prior Work

2.1 Localisation Correlation Analysis

The assumption that co-localisation within the cell implies similar protein function is the basis of the analysis. The LOPIT and ProCoDeS profiles provide information about possible localisation in the same organelle and protein complex respectively. In order predict functionality candidate unknown proteins need to be compared with known bait proteins.

The correlation of bait protein's profile to the rest of detected proteins can be calculated and then an ordered list of all pairwise comparisons produced. It is hoped that the proteins of similar function to the bait occur at high correlation level, near the top of the list. Each of the thousand proteins will generate such a list for each experiment, questions on how to combine lists across the same and different experiments must be answered.

The correlation lists from two known bait proteins having the same function are not exactly the same, so (Mutual) Ranks were used in order to unify multiple correlations lists, though there are issues with this approach. Fundamentally, the correlation analysis does not separate satisfactorily bait proteins of different function. The analysis is limited since it fails to use any additional information about the profile covariance structure of proteins within a known organelle or complex. The profiles are also points on a simplex, which may add some bias to the correlation

2.2 Combining Experiments

Considering the pairwise correlation of protein profiles has been used to form (ranked) lists of likely candidate proteins. One aim is to combine multiple experiments to form better candidate lists. The LOPIT and ProCoDeS profiles distinguish organelles and complexes respectively. A combination of the two experiments is to plot the correlation of all proteins against a bait for both experiments.

The most pressing issues are: the number of proteins with known organelle and complex is small, not all proteins are present in both experiments. The scales of both experiments are expressed in arbitrary units reflecting the abundance of each protein across the whole trace. However the scales represent different kind of information and hence they are not directly comparable. The question arises about what region in the "double correlation plot" contains the good candidates. It is difficult to quantify the combined information and such plots do not extend to comparing many experiments together.

2.3 Principal Component Analysis and Clustering

Considering a single experiment initially, the data set consists of approximately one thousand profiles (vectors of length four to seven). It is difficult to visualise high dimensional data, so initial exploratory analysis to reduce the dimension has been done. Principal Component Analysis (PCA) is a dimension reduction technique, it performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance. From first to last the axes of the transformed data account for decreasing observed variability. It is common to plot the data projected onto the first two principal components.

The PCA plot gives evidence to the co-localisation assumption, known proteins are seen to cluster. Using the principal components to form candidate lists can be based on the distance of a protein from the centroids of known protein sets (organelle or complex). There are several distance metrics that can be considered. PCA is a graphical technique, which is useful to present the profile vectors in a more understandable form. However, since the dimension of the full profile is small (seven at most) there is no need to work with a reduced set of dimensions for any statistical analysis (reduction is usually performed when there are hundreds of predictors).

Cluster Analysis can be used to group the proteins, and dendrograms used to illustrate the clustering based on various metrics. However, given the large number of unknown proteins there are incorrectly grouped known proteins on the dendrogram. Additionally, the number of clusters is unknown and predicted to be much larger (in particular for the ProCoDeS) than the number of known groups; also the number of proteins within each group is unknown. This is a significant issue for techniques such as k-means clustering, where the number of clusters must be initially set.

3 Meeting discussion

3.1 Distance metrics in real space

LOPIT and ProCoDeS assays lead to a "profile" for each protein,

$$\mathbf{x}_i \in \left\{ \mathbf{w} : \sum_{j=1}^{m} w_j = 1,\, w_j \geq 0 \right\}. \tag{1}$$

Inference of protein function from such data is based on two premises: i) that two proteins with similar profiles are likely to be co-localised in the cell/organelle; and ii) that co-localised proteins are likely to share a common function.

There are many ways to measure the similarity (or difference) between two profiles (denoted below as **a** and **b**); possible candidate metrics are:

- Pearson (product-moment) correlation coefficient

$$\frac{1}{n-1}\sum\left(\frac{a_i-\bar{a}}{s_a}\right)\left(\frac{b_i-\bar{b}}{s_b}\right)\in[-1,1], \tag{2}$$

  where s_a and s_b are the standard deviations of the elements of $\mathbf{a}$ and $\mathbf{b}$;

- Euclidean distance (L2-norm): $m^{-1}\sum_{i=1}^{m}(a_i-b_i)^2\in\mathbb{R}^+$;

- Mean absolute differences (L1-norm): $m^{-1}\sum_{i=1}^{m}|a_i-b_i|\in\mathbb{R}^+$.

A procedure to find, amongst n candidates proteins, those that might be functionally linked to some bait protein, indexed say by b, is to calculate the pairwise similarities, denoted $(S_{i,b})_{i=1,..n}$, between the bait and each of the other proteins, then order these and assign a rank to each candidate protein according to its position in the ordered list. The rank is thus:

$$R(a,b)=1+\sum_{i=1}^{n}\mathbb{I}\{S_{a,b}<S_{i,b}\}, \tag{3}$$

where $\mathbb{I}$ is the indicator function, which equals 1 if its argument is true and 0 otherwise.

Rank $R(a,b)$ has the advantage of being intuitive, plus one can use it with any choice of similarity metric, S. However, is has two disadvantages: i) it is asymmetric in its arguments, i.e. in general $R(a,b)\neq R(b,a)$; and ii) it is calculated pairwise. Thus there is no clear way to generalise the method to determine likely candidate proteins that belong to some “module”.

To overcome the asymmetry problem, the *Arabidopsis thaliana* trans-factor and cis-element prediction database (`http://atted.jp`) uses “mutual rank”,

$$M(a,b)=\sqrt{R(a,b)\cdot R(b,a)}, \tag{4}$$

which is the geometric mean of the ranks $R(a,b)$ and $R(b,a)$, and thus symmetric in its arguments. Calculating $(M(a,b))_{a=1\ldots n,b=a\ldots n}$ and ordering the values according to size is a way to calculate the most probable links amongst all the $\frac{1}{2}n(n-1)$ pair of proteins.

One could potentially extend the idea of geometric averaging to obtain a single measure of similarity between a protein and a bait module (i.e. a set of proteins $B=\{b_1,\ldots b_p\}$):

$$N(a;B)=\left(R(a,b_1)\cdot R(a,b_2)\cdots R(a,b_p)\right)^{1/p}. \tag{5}$$

We discussed this idea at length in the study group, but eventually decided that this was unlikely to be an effective way to identify candidate proteins that belong to a module. This decision was partly because the method of taking geometric averages of ranks has no theoretical foundation and it is thus difficult to trust that the method will produce robust predictions. Furthermore because (5) is based on amalgamating pairwise quantities, it neglects an important source of information, namely the structure of the dispersion of the bait-module observations.

Instead of calculating and combining pairwise quantities, it is preferable to work directly with quantities that measure distance from a single protein to a group of proteins. The simplest choice is Euclidean distance

$$d_E(\mathbf{a};B)=\sqrt{(\mathbf{a}-\bar{\mathbf{b}})^T(\mathbf{a}-\bar{\mathbf{b}})}, \tag{6}$$

where $\bar{\mathbf{b}}=p^{-1}\sum_{i=1}^{p}\mathbf{b}_i$ is the centroid of the set $B=\{\mathbf{b}_1,\ldots,\mathbf{b}_p\}$. An alternative choice—which has the advantage that it incorporates the structure of the dispersion of the observations—is the “Mahalanobis distance”, defined as

$$d_M(\mathbf{a};B)=\sqrt{(\mathbf{a}-\bar{\mathbf{b}})^T\Sigma^{-1}(\mathbf{a}-\bar{\mathbf{b}})}, \tag{7}$$

where $\bar{\mathbf{b}}$ is again the centroid, Σ is the covariance matrix of the set B, and Σ^{-1} denotes the matrix inverse of Σ (or the generalised inverse if Σ is rank-deficient). The Euclidean and Mahalanobis distances are equal only when Σ is the identity matrix, i.e. when the observations are isotropically dispersed. When the data have non-isotropic dispersion, the Mahalanobis distance penalises an observation $\mathbf{a}$ if it is positioned relative to the centroid in a direction in which the variance of $\mathbf{b}_1,\ldots,\mathbf{b}_p$ is small; see Figure 1.

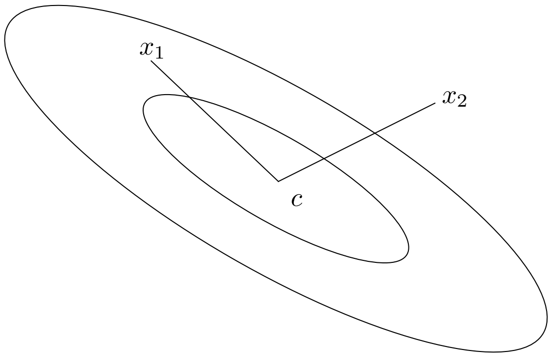

Figure 1: Schematic illustrating Mahalanobis distance. The ellipses represent contours of density. The points marked x_1 and x_2 have equal Euclidean distance to the centroid c, but x_1 has a smaller Mahalanobis distance to c than x_2 does.

3.2 The unit-sum constraint and the possible need for other metrics

The techniques discussed in the above section are suitable for analysing data that lie in Euclidean space. LOPIT and ProCoDeS traces, however, have the constraint (eqn (1)) that their elements are non-negative and sum to one. A trace is thus a point on a subspace of Euclidean space called the m-dimensional simplex. Data of this kind, which are common in some other fields (notably geology) are called "compositional data", and an individual observation, i.e. a trace in the present context, is called a "composition".

The unit-sum constraint affects the covariance structure of the elements of a composition. This can lead to spurious results if compositional data are analysed using standard techniques of multivariate analysis (see Aitchison [4, 5, 6, 7]). In particular, for a vector of uncorrelated random variables, ($(W_1, \ldots, W_m)$ such that $\mathrm{corr}(W_i, W_j) = 0$ for $i \neq j$), elements of the corresponding composition $(X_1, \ldots, X_m) = (W_1 + \cdots + W_m)^{-1}(W_1, \ldots, W_m)$ are in general correlated on account of the scaling factor that appears in each element. The problem is more acute when the composition contains fewer elements and, as Aitchison [6] pointed out, most apparent for $m = 2$ since in this case

$$\mathrm{cov}(X_1, X_2) = \mathrm{cov}(X_1, 1 - X_1) = -\mathrm{var}(X_1) = -\mathrm{var}(X_2), \tag{8}$$

and the correlation is

$$\mathrm{corr}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{var}(X_1)\,\mathrm{var}(X_2)}} = -1. \tag{9}$$

Hence we may need to be cautious when interpreting what Aitchison calls "crude" analyses, i.e. those that are based on the covariances of compositional components calculated in the same way as if the observations were lying in unconstrained real space.

Progress in this area has been made by recognising that a composition provides information about only the relative sizes of its components, and that analysis should be based on studying ratios of components. Aitchison [5] introduced the log-ratio transformation (a bijective mapping from the simplex onto a real space) and suggested that standard multivariate techniques could be legitimately used on the transformed observations; see [4] for a discussion of principal component analysis of compositional data.

We return to the question which is significant in defining both centrality and variability of data: how can we best measure differences between compositions? See [1, 2, 3, 9, 10, 11] for a debate on the properties that a difference measure should have. Commonly used is Aitchison's simplicial distance [4],

$$d_S(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{m} \left\{ \ln \frac{a_i}{g(\mathbf{a})} - \ln \frac{b_i}{g(\mathbf{b})} \right\}^2}, \tag{10}$$

where $g(\cdot)$ denotes the geometric mean of the elements of its argument. A natural centroid based on this metric is the point to which the sum across observations of simplicial distances is minimised. This point is

$$\mathrm{cen}\,(\mathbf{x}_1, \ldots, \mathbf{x}_n) = (h_1 + \cdots + h_n)^{-1} (h_1, \ldots, h_n), \tag{11}$$

where h_j is the geometric mean across observations of the jth component, i.e. $h_j = n^{-1}\left((\mathbf{x}_1)_j \times \cdots \times (\mathbf{x}_n)_j\right)$. Usefully, with the transformation

$$\mathbf{y} = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_m}{g(\mathbf{x})}\right), \tag{12}$$

then $d_S(\mathbf{x}_i, \mathbf{x}_j) = d_E(\mathbf{y}_i, \mathbf{y}_j)$ and $\text{cen}\,(\mathbf{x}_1, \ldots, \mathbf{x}_n) = n^{-1} \sum_{i=1}^{n} \mathbf{y}_i$ (see Martín-Fernández et al. [8]); hence after an initial data transformation (12) we can apply standard multivariate techniques to the transformed observations.

4 Numerical tests

We tested several methods to see how well they performed in identifying co-localised proteins based on LOPIT data. We tried the "crude" approach of working with plain compositional data, and the "log-ratio" approach of first transforming the data using (12). For each of the crude and log-ratio approaches we implemented both the Euclidean and Mahalanobis distances, meaning that we tried a total of four methods.

We used a "leave-one-out" test algorithm, summarised by the following pseudo-code (in which we assume we have n proteins, of which j belong to group G).

```
- for protein $i$ from 1 to $n$,
    - if $i$ belongs to $G$,
        - calculate the centroid $c$ (and, in the Mahalanobis case, the covariance matrix $\Sigma$) for the other
          $(j-1)$ members of $G$,
    - else
        - calculate $c$ (and, in the Mahalanobis case, $\Sigma$) for the $j$ members of $G$,
    - end
    - calculate the distance $d$ $(= d_E$ or $d_M)$ of protein $i$ from $c$,
- end
- rank proteins according to ascending size of $d$, and define first $m$ to be the most likely candidate
  members of $G$.
- plot receiver operator characteristic (ROC) curve (a plot of true positives versus false positives for all
  values of $1 \leq m \leq n$) .
```

We repeated this for each of the groups of known proteins; each group corresponds to a line on the ROC plots (Figs 2–5). At the stage of plotting the ROC curves, we have two choices of how to deal with the unclassified observations (of which there is a large number): we could treat them as false positives, or else omit them from consideration altogether. The ROC curves in Figures 2–5 show results both ways.

4.1 Results

Figures 2 and 3 show ROC curves for protein membership of organelles, and Figures 4 and 5 are for protein membership of complexes. In an ROC curve, if the number of true positives rises very quickly it shows that the method can successfully predict which proteins are members of a particular group. In this regard, each of the four methods shows good discriminatory power, even the "crude" methods that take no special account of the unit-sum constraint.

To decide from ROC curves which method is best is subjective; one needs first to decide upon an appropriate criterion based on the relative costs of missing true positives and of finding false positives. For example, if missing true positives is more costly than finding false positives then an appropriate (scalar) measure of performance might be the proportion of false positives at which we achieve 0.3 true positives (criterion A, say). If vice versa, we might set this level at 0.7 (criterion B). There other ROC curve summaries too, such as the area bounded by the ROC curve and the lines $x = 0$ and $y = 1$.

A broad summary of the performance of the methods based on Figures 2–5 is as follows. Considering first criterion A and the organelle data (Figs 2 and 3), the best performance comes from using metric d_M on

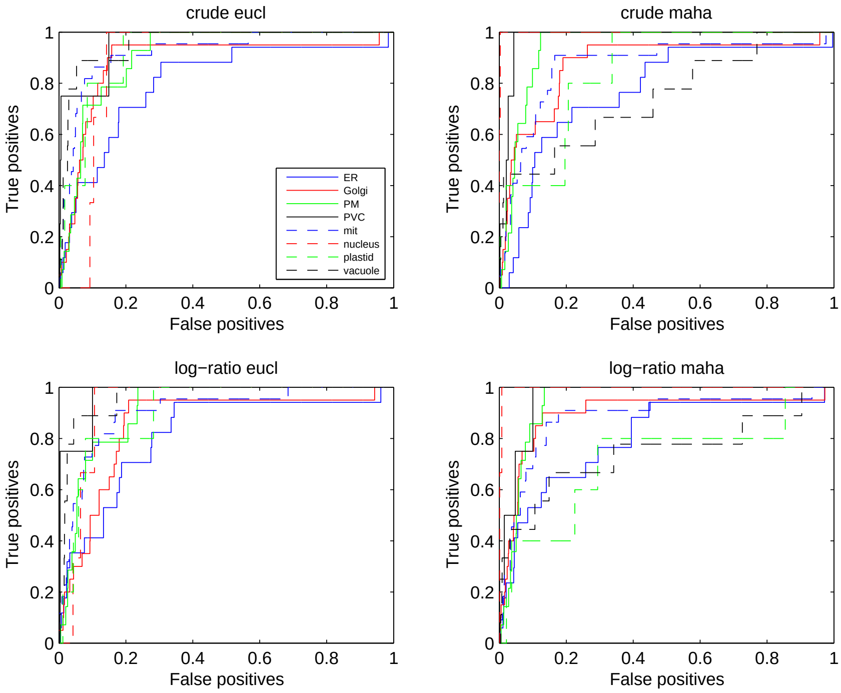

Figure 2: ROC curves for organelle membership with all unknown observations classified as false positives, using (upper-left) Euclidean distance, crude data; (upper-right) Mahalanobis distance, crude data; (lower-left) Euclidean distance, log-ratio transformed data; (lower-right) Mahalanobis distance, log-ratio transformed data.

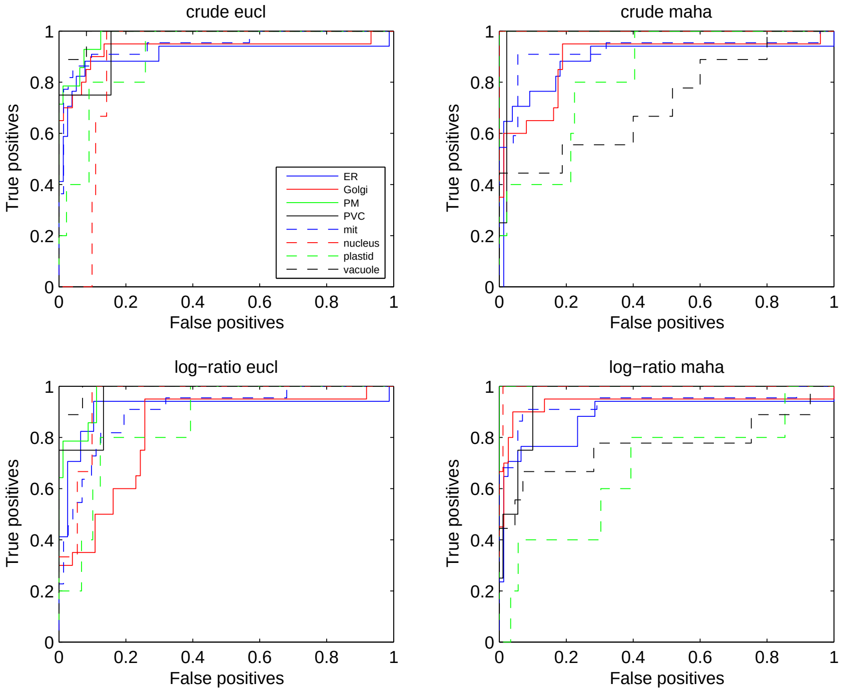

Figure 3: ROC curves for organelle membership based only on data for proteins whose group is known, using (upper-left) Euclidean distance, crude data; (upper-right) Mahalanobis distance, crude data; (lower-left) Euclidean distance, log-ratio transformed data; (lower-right) Mahalanobis distance, log-ratio transformed data..

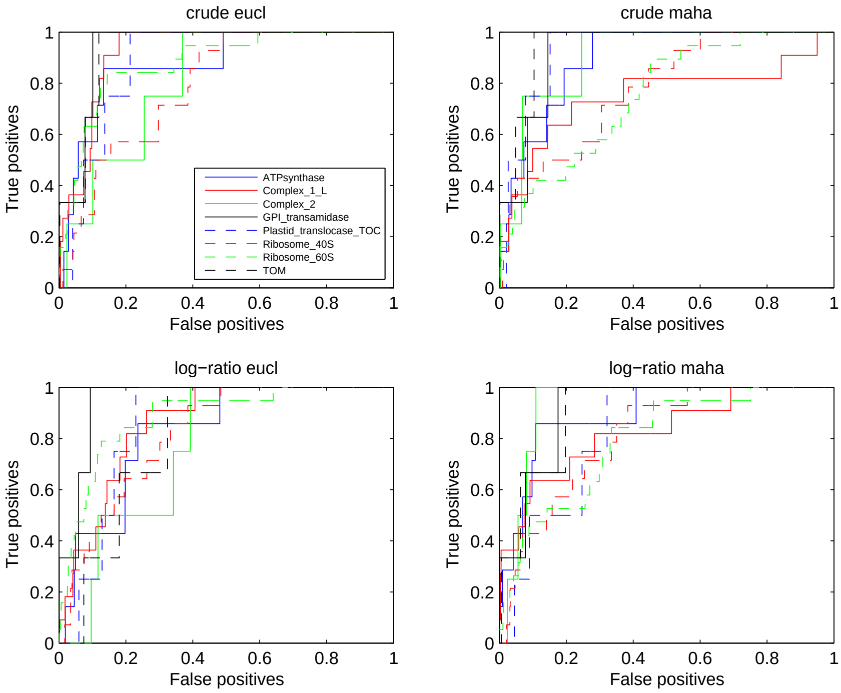

Figure 4: ROC curves for complex membership with all unknown observations classified as false positives, using (upper-left) Euclidean distance, crude data; (upper-right) Mahalanobis distance, crude data; (lower-left) Euclidean distance, log-ratio transformed data; (lower-right) Mahalanobis distance, log-ratio transformed data.

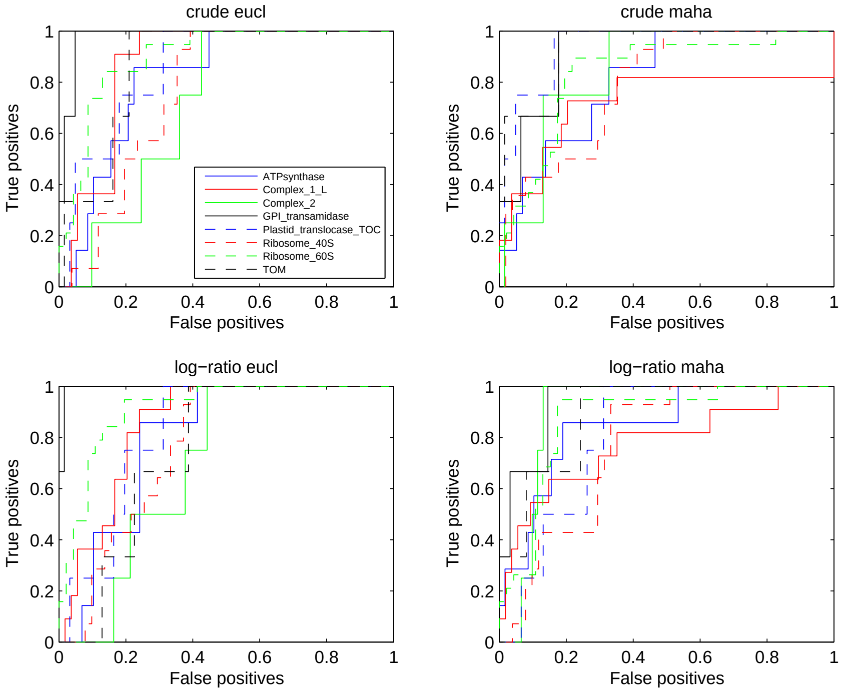

Figure 5: ROC curves for complex membership based only on data for proteins whose group is known, using (upper-left) Euclidean distance, crude data; (upper-right) Mahalanobis distance, crude data; (lower-left) Euclidean distance, log-ratio transformed data; (lower-right) Mahalanobis distance, log-ratio transformed data.

either the crude or the log-ratio transformed data, the latter being slightly superior. For criterion A and the complex data, metric d_M is again the best, but with the crude method this time better than using the log-ratio transformation. However, if we instead use criterion B, these conclusions are reversed: the crude method using simple Euclidean distance d_E appears best for both organelle and complex data.

5 Bayesian Analysis for LOPIT and ProCoDeS data

In a Bayesian framework, probabilities are random variables on $[0, 1]$ that represent degrees of belief. An advantage of this approach is that it allows us to specify prior beliefs then update them in light of experimental data. This offers a natural way to combine data from experiments of different types (e.g. LOPIT and ProCoDeS) and allows the user to weight the importance of each experiment.

In this section we construct an informative prior on the probability that a protein belongs to a particular group and use Bayesian analysis to calculate the posterior distribution. Throughout, we denote the squared Mahalanobis distance of $\mathbf{x}$ from the set Y by $\mathrm{d}_M^2(\mathbf{x}, Y)$. This distance is defined

$$\mathrm{d}_M^2(\mathbf{x}, Y) = (\mathbf{x} - \bar{\mathbf{y}})^T \Sigma_Y^{-1} (\mathbf{x} - \bar{\mathbf{y}}),$$

where $\bar{\mathbf{y}}$ is the centroid of the set Y and Σ_Y is the covariance matrix of the elements of the set Y.

Let $\theta_{\mathbf{x},Y}$ be the probability that $\mathbf{x}$ is in set Y. Suppose we have N proteins in the experiment, of which N_Y are known to be in the set Y. We construct a prior on $[0, 1]$ for $\theta_{\mathbf{x},Y}$, such that $p(0 \leq \theta_{\mathbf{x},Y} \leq 0.5) = (N - N_Y)/N$ and $p(0.5 < \theta_{\mathbf{x},Y} \leq 1) = N_Y/N$ as shown in Figure 6

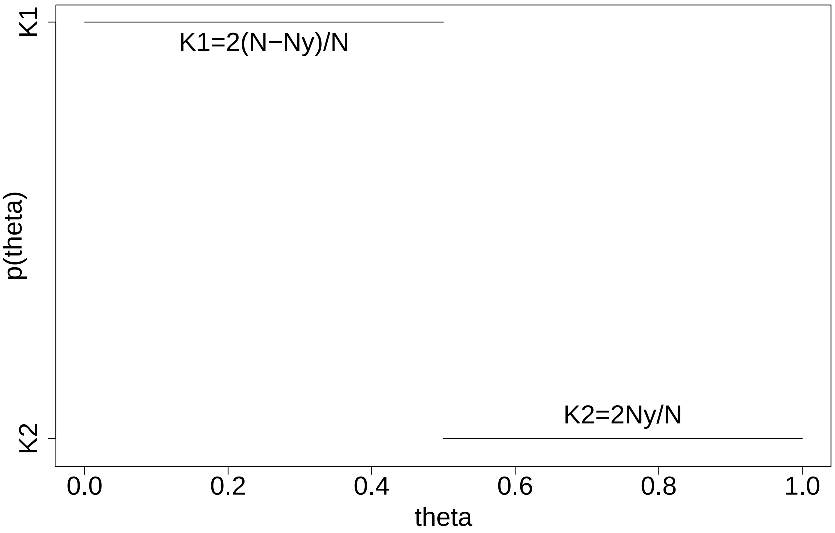

Figure 6: The prior density for $\theta_{\mathbf{x},Y}$

We are interested in the posterior distribution of θ, $\pi(\theta_{\mathbf{x},Y}|D)$, where the data, D, consists of the Mahalanobis distance $d_M(\mathbf{x}, Y)$.

The relevant equation for the posterior distribution of $\theta_{\mathbf{x},Y}$ is

$$\pi(\theta_{\mathbf{x},Y}|\mathrm{d}_M^2(\mathbf{x}, Y) = k) = \frac{\pi(\mathrm{d}_M^2(\mathbf{x}, Y) = k|\theta_{\mathbf{x},Y})\pi(\theta_{\mathbf{x},Y})}{\int_0^1 \pi(\mathrm{d}_M^2(\mathbf{x}, Y) = k|\theta_{\mathbf{x},Y})\pi(\theta_{\mathbf{x},Y})\mathrm{d}\theta_{\mathbf{x},Y}}. \quad (13)$$

For the likelihood, $\pi(\mathrm{d}_M^2(\mathbf{x}, Y) = k|\theta_{\mathbf{x},Y})$, we use a mixture model. If $\mathbf{x} \in Y$, we assume that $\pi(\mathrm{d}_M^2(\mathbf{x}, Y) = k|\theta_{\mathbf{x},Y})$ is an exponential distribution, and if $\mathbf{x} \notin Y$, we assume that $\pi(\mathrm{d}_M^2(\mathbf{x}, Y) = k|\theta_{\mathbf{x},Y})$ is a uniform distribution on the interval $[0, M]$, where $M = 2 \max_{\mathbf{x}} \mathrm{d}_M^2(\mathrm{x}, Y)$ and the maximum is taken over all of the proteins in the experiment. Thus $\pi(\mathrm{d}_M^2(\mathbf{x}, Y) = k|\theta_{\mathbf{x},Y}) = \theta f_1(k) + (1 - \theta)f_2(k)$, where $f_1(k) = A_Y e^{-A_y k}$ and $f_2(k) = 1/M)$.

Although we do not know the value of A_Y, we can use the maximum likelihood estimate, $\hat{A}_Y$, which maximises $\prod_{\mathbf{x}\in Y} A_Y e^{-A_Y \mathrm{d}^2_M(\mathbf{x},Y)}$. Taking logs, we wish to maximise

$$g(A_Y) = N_Y \log A_Y - A_Y \sum_{\mathbf{x}\in Y} \mathrm{d}^2_M(\mathbf{x}, Y).$$

Differentiating and setting the result equal to zero yields

$$\hat{A}_Y = \frac{N_Y}{\sum_{\mathbf{x}\in Y} \mathrm{d}^2_M(\mathbf{x}, Y)}.$$

The normalising constant in equation (13) is given by

$$\begin{aligned}
\int_0^1 \left[\theta_{\mathbf{x},Y} \hat{A}_Y e^{-\hat{A}_Y k} + \frac{1-\theta_{\mathbf{x},Y}}{M}\right] \pi(\theta_{\mathbf{x},Y}) \mathrm{d}\theta_{\mathbf{x},Y} &= \hat{A}_Y e^{-\hat{A}_Y k} \mathbb{E}(\theta_{\mathbf{x},Y}) + \frac{1}{M} - \frac{1}{M}\mathbb{E}(\theta_{\mathbf{x},Y}) \\
&= \mathbb{E}(\theta_{\mathbf{x},Y}) \left[\hat{A}_Y e^{-\hat{A}_Y k} - \frac{1}{M}\right] + \frac{1}{M},
\end{aligned}$$

where $\mathbb{E}(\theta_{\mathbf{x},Y})$ is the prior expectation of $\theta_{\mathbf{x},Y}$.

In fact, we can calculate $\mathbb{E}(\theta_{\mathbf{x},Y})$ exactly.

$$\begin{aligned}
\mathbb{E}(\theta_{\mathbf{x},Y}) &= \int_0^1 \theta_{\mathbf{x},Y} \pi(\theta_{\mathbf{x},Y}) \mathrm{d}\theta_{\mathbf{x},Y} \\
&= \int_0^{1/2} \theta_{\mathbf{x},Y} 2\frac{N-N_Y}{N} \mathrm{d}\theta_{\mathbf{x},Y} + \int_{1/2}^1 \theta_{\mathbf{x},Y} 2\frac{N_Y}{N} \mathrm{d}\theta_{\mathbf{x},Y} \\
&= 2\frac{N-N_Y}{N}\left[\frac{\theta^2_{\mathbf{x},Y}}{2}\right]_0^{1/2} + 2\frac{N_Y}{N}\left[\frac{\theta^2_{\mathbf{x},Y}}{2}\right]_{1/2}^1 \\
&= 2\frac{N-N_Y}{N}\frac{1}{8} + 2\frac{N_Y}{N}\frac{3}{8} \\
&= \frac{1}{4} + \frac{N_Y}{2N}.
\end{aligned}$$

Putting it all together we have

$$\pi(\theta_{\mathbf{x},Y} | \mathrm{d}^2_M(\mathbf{x}, Y) = k) = \frac{(\theta_{\mathbf{x},Y} \hat{A}_Y e^{-\hat{A}_Y k} + (1-\theta_{\mathbf{x},Y})/M)\pi(\theta_{\mathbf{x},Y})}{\left(\frac{1}{4} + \frac{N_Y}{2N}\right)\left(\hat{A}_Y e^{-\hat{A}_Y k} - \frac{1}{M}\right) + \frac{1}{M}}.$$

We wish to calculate the posterior mean of θ conditional on the value $\mathrm{d}^2_M(\mathbf{x}, Y)$, i.e. $\mathbb{E}(\theta_{\mathbf{x},Y}|\mathrm{d}^2_M(\mathbf{x}, Y) = k)$. After a little algebra we obtain

$$\mathbb{E}(\theta_{\mathbf{x},Y}|\mathrm{d}^2_M(\mathbf{x}, Y) = k) = \frac{1}{C_k}(C_1 \hat{A}_Y e^{-\hat{A}_Y k} + C_2),$$

where

$$\begin{aligned}
C_k &= \left(\frac{1}{4} + \frac{N_Y}{2N}\right)\left(\hat{A}_Y e^{-\hat{A}_Y k} - \frac{1}{M}\right) + \frac{1}{M} \\
C_1 &= \frac{1}{2}\frac{N_Y}{N} + \frac{1}{12} \\
C_2 &= \frac{1}{6M}
\end{aligned}$$

If we write $D_1 = 1/4 + N_Y/(2N)$ and $D_2 = 1/M(1 - D_1)$, then we have

$$\mathbb{E}(\theta_{\mathbf{x},Y}|\mathrm{d}^2_M(\mathbf{x}, Y) = k) = \frac{C_1 \hat{A}_Y e^{-\hat{A}_Y k} + C_2}{D_1 \hat{A}_Y e^{-\hat{A}_Y k} + D_2}.$$

We can show that with the constants defined as above, the derivative of this posterior mean function is always negative, i.e. the posterior mean is a monotonically decreasing function of the Mahalanobis distance from a particular protein to the set in question. Hence in the case of a single experiment, we obtain identical results to those given in the non-Bayesian setting described above. This justifies choosing the prior as in figure (6). As we stated earlier, one could go on to apply the Bayesian approach to integrate data from multiple experiments.

6 Summary and directions for future work

We have discussed approaches for inferring function of proteins based on LOPIT data, and preliminary numerical results show promising performance. The approach is based on the assumptions that proteins with similar LOPIT profiles are co-localised, and that co-localised proteins are likely to have the same function.

In order to assess the likelihood of a protein belonging to a particular group, e.g. being expressed in a particular organelle, we used two metrics (Euclidean and Mahalanobis) on both raw and log-ratio transformed LOPIT data. We used ROC curves to show that all four methods had good discriminatory power.

We also formulated the problem in a Bayesian framework. We showed that by specifying an informative prior on the basis of known group membership and using the posterior mean as a summary statistic, the Bayesian analysis gives the same results as the non-Bayesian method. The approach, however, provides a natural way to combine results from multiple experiments.

Areas for future work include the following:

- Implement the Bayesian method.
- Thoroughly explore other possible metrics (especially simplicial metrics) and systematically compare their performance. Two metrics worthy of inclusion are:
    - the Pearson correlation coefficient, the metric upon which the *Arabidopsis thaliana* trans-factor and cis-element prediction database (`http://atted.jp`) is heavily based;
    - variants of the Mahalanobis distance that employ a regularised covariance matrix. This could potentially improve performance for groups where there is only a small number of known members, especially where the number is smaller than the dimension m of LOPIT data (a case for which here we have employed the generalised inverse).

References

[1] J. Aitchison. Comment on measures of variability for geological data by d. f. watson and g. m. philip. *Mathematical Geology*, 22(2):223–226, Feb. 1990. doi: 10.1007/BF00891826. URL `http://dx.doi.org/10.1007/BF00891826`.

[2] J. Aitchison. On criteria for measures of compositional difference. *Mathematical Geology*, 24(4):365–379, May 1992. doi: 10.1007/BF00891269. URL `http://dx.doi.org/10.1007/BF00891269`.

[3] J. Aitchison. Delusions of uniqueness and ineluctability. *Mathematical Geology*, 23(2):275–277, Feb. 1991. doi: 10.1007/BF02066299. URL `http://dx.doi.org/10.1007/BF02066299`.

[4] J. Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, Apr. 1983. doi: 10.1093/biomet/70.1.57. URL `http://biomet.oxfordjournals.org/cgi/content/abstract/70/1/57`.

[5] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982. ISSN 00359246. doi: 10.2307/2345821. URL `http://www.jstor.org/stable/2345821`. ArticleType: primary_article / Full publication date: 1982 / Copyright 1982 Royal Statistical Society.

[6] J. Aitchison. *The statistical analysis of compositional data.* Chapman and Hall, New York, 1986.

[7] J. Aitchison and J. J. Egozcue. Compositional data analysis: Where are we and where should we be heading? *Mathematical Geosciences*, 37(7):829–850, Oct. 2005. doi: 10.1007/s11004-005-7383-7. URL `http://dx.doi.org/10.1007/s11004-005-7383-7`.

[8] J. A. Martín-Fernández, C. Barcelo-Vidal, and V. Pawlowsky-Glahn. A critical approach to non-parametric classification of compositional data. In *Proceedings of the 5 th Conference of the International Federation of Classification Societies, Universit La Sapienza, Rome*, 1998.

[9] D. F. Watson. Reply to comment on measures of variability for geological data by d. f. watson and g. m. philip. *Mathematical Geology*, 22(2):227–231, Feb. 1990. doi: 10.1007/BF00891827. URL `http://dx.doi.org/10.1007/BF00891827`.

[10] D. F. Watson. Reply to delusions of uniqueness and ineluctability by j. aitchison. *Mathematical Geology*, 23(2):279, Feb. 1991. doi: 10.1007/BF02066300. URL `http://dx.doi.org/10.1007/BF02066300`.

[11] D. F. Watson and G. M. Philip. Measures of variability for geological data. *Mathematical Geology*, 21 (2):233–254, Feb. 1989. doi: 10.1007/BF00893217. URL `http://dx.doi.org/10.1007/BF00893217`.