Web supplement

Reference Expression Data Set

Gene expression can be measured with multiple tools, each resulting in slightly different results. Even within the same technique results are known to vary wildly due to different experimental conditions, cross hybridization, different calibration procedures for relating fluorescence intensity to a cellular concentration or different protocols for harvesting and reverse-transcribing cellular mRNA. In contrast to expression levels determined by SAGE, gene chips are sometimes suffering from a saturation of the signal for highly expressed mRNA transcripts. On the other hand, SAGE values are less reliable for lowly expressed genes.

To examine yeast mRNA expression levels in a comprehensive and robust manner we attempted to assemble a reference expression data set. This reference data, through iterative scaling and merging, integrates overlapping data from several gene chip experiments, and in addition includes the SAGE expression data for highly expressed genes to correct for saturation on the gene chips. The “averaged” data reduces the inherent noise and errors in the individual expression sets. By combining several data sources, the reference expression set is also more complete than the individual data sets since it encompasses a larger set of ORFs.

The following is an outline of the methods used to construct the set. We start with the values of one gene chip data set $U_i$ where $i$ is used throughout as a subscript to denote gene number. We then transform the values of the next gene chip data set $X_i$ to $Y_i$ with the following non-linear regression:

$$\min \sum_i (Y_i - U_i)^2 \quad \text{with} \quad Y_i = AX_i^B$$

where $A$ and $B$ are the parameters of the regression. Note that the regression is carried out on the intersection of the two datasets but that the following transformation applies to all data points $X_i$. We merge and average the data to create a new reference expression level $V_i$ as follows:

If $U_i$ and $Y_i$ are both defined for gene $i$ and $|Y_i - U_i| < \alpha$

Then $V_i = \frac{1}{2} (Y_i + U_i)$

Else, if only $Y_i$ exists, $V_i = Y_i$

Else $V_i = U_i$

When an expression level for a particular ORF exists in only one of the datasets, we just incorporate this value. When both data sets have values for an ORF, we average the values if they are within 15% of each other; otherwise, we just stay with the original
expression value $U_i$. We used $\alpha = 15\%$ in order to prevent outliers from skewing the result. Other data sets are subsequently included in the same procedure, continuing the iteration from the new expression values $V_i$. The initial iteration starts with the Young expression set as $U_i$ since we have the highest confidence in its accuracy.

For merging the Church set onto the Young set the regression parameters are:

\begin{align*}
A &= 7.00 \\
B &= 0.72
\end{align*}

For merging the Samson set onto the combined Church and Young data set, the regression parameters are:

\begin{align*}
A &= 1.63 \\
B &= 0.76
\end{align*}

The SAGE data was not included in the above procedure since it is of a fundamentally different nature. An advantage of the SAGE technology over gene chips is that there is no possible signal saturation for high expression levels, as is possible for chips (Futcher et al. 1999). Conversely, SAGE values are less reliable for lowly expressed genes since there is a chance that one might not sequence a SAGE tag corresponding to such a gene altogether. Therefore, if, after the last iteration, the average Gene Chip expression level $V_i$ was both above a certain threshold $\beta$ and below the SAGE expression level $S_i$ for the same gene, it was replaced with the SAGE value; otherwise the average Gene Chip value was kept. This gave us our final expression set $w_{mRNA}$. Our treatment of the SAGE data is modeled after that in Futcher et al. (1999), and like them, we used $\beta = 16$.

The following figures show scatter plots of different data sets against one another. Supplemental figure 1 shows a scatter plot of the Young set on the ordinate and the Church set on the abscissa. The black line represents a linear fit, the red line a non-linear fit according to the equation above. It can be seen from the plot that the relationship between the two datasets can be better described with a non-linear formula. Supplemental figure 2 shows a scatter plot of the combined Young and Church data against the Samson data. Again, it can be seen that the relationship between the two datasets is of non-linear nature. Finally, supplemental figure 3 shows a scatter plot of the reference expression data set against the original Young data. The reference expression data follows the Young data relatively closely. Only for the most highly expressed genes the reference set values tend to be higher than the Young set values because of the integration of the SAGE data.
Supplemental figure 2

The figure shows a scatter plot with mRNA expression values. The x-axis represents mRNA expression for Samson, while the y-axis represents mRNA expression for Young/Church combined. The data points are scattered across the graph, and two lines are drawn to illustrate the correlation between the two sets of expression values.