1. Introduction

Google’s™ PageRank algorithm assesses the “importance” of a web-page based on the “importance” of web-pages which link to it. We modify this algorithm to:

- provide a more accurate interpretation of microarray experiments which is more robust to noise.
- highlight potentially active genes.
- provide a ranked list of genes based on their importance in an experiment.

Data from different sources is combined to construct a network.
- Gene Ontology annotations.
- Gene expression data.

Assumption: high connectivity to genes with a significant change of expression may indicate that a genes relative ranking should be increased. See Figure 1.

2. Network Construction

- Google™ views the world-wide web as a directed graph.
- We construct an undirected graph with weighted nodes.
- Nodes represent each gene and are weighted with the absolute value of the expression level.
- Genes are connected by an edge in the graph if they share a GO annotation.

3. The Algorithm

Original PageRank

Hyperlinks into a page are viewed as a “vote of confidence.” Votes from a high-ranked page have a greater influence on the rank of a page.

Original PageRank Definition

The algorithm is defined iteratively as shown in (1):

$$r_j^t = (1-d) + d \sum_{i=1}^{N} w_{ij} r_i^{t-1}$$  \hspace{1cm} (1)

- $r_j^t$: rank of node $j$ at the $t$th iteration.
- $w_{ij}$ = 1 if there is a link between nodes $i$ and $j$, $w_{ij}$ = 0 otherwise.
- $d$: number of links from gene $i$.
- $d$: a free parameter in the range (0,1).

Modified Algorithm

Connections to other changed genes are an indication of importance. Connections to a gene with a high differential expression carry more weight.

Instead of giving each gene a rank of (1-d) ‘for free’, as in the original algorithm, we modify this to give each gene (1-d)*$w_{ij}$, where $w_{ij}$ is the absolute value of its expression level (2).

This is an instance of ‘personalised PageRanking’, where the personalisation vector is $x$ (3).

$$r_j^t = (1-d)x + d \sum_{i=1}^{N} w_{ij} r_i^{t-1}$$  \hspace{1cm} (2)

This iteration is shown to converge to the solution of the linear system given in (3):

$$x = \left( I - dA + dW \right)^{-1} e$$  \hspace{1cm} (3)

- $e$: vector of expression levels.
- $W$: the adjacency matrix of the network.
- $D$: a diagonal matrix where $D_{ii}$ of degree gene $i$.
- $r_j^t$: vector of rankings

True generalisation of page rank; for $x = f$ we revert to the original algorithm.

Random Walk Interpretation: equivalent to a random walk where we teleport to a random node with probability proportional to expression level.

Combining expression and connectivity. For $d=0$ we return the ranking based on pure expression level. For $d=1$ we return ranking based on the degree of each node. Setting $d$ in the range (0,d) we interpolate between expression and connectivity extremes.

4. Results

Synthetic Networks

- Artificially constructed networks of 1000 genes.
- Set A: 100 genes with random expression levels from a $N(0,1)$ distribution.
- Set B: 900 genes with random expression levels from a $N(0,1)$ distribution.

We hope to identify the set A genes as the top 100 ranked genes. The median rank of set A is measured after PageRanking (min50).

Figure 2 demonstrates the importance of a high relative connectivity between the genes in set A compared to the connections in the network as a whole.

Real Networks

- The GO Cellular Component network has the required connectivity. See Figure 3.
- We work only with the down regulated genes (due to the high connectivity amongst the 200 most down-regulated genes.)

Experiment: take 200 genes and set each of their expression levels to zero in turn. We measure the difference in their ranked position before and after PageRanking.

Results:

1. Top 200 genes: We are able to boost the ranked position of each gene. In some cases the gene returned to its original ranked position. See Figure 4.

2. Random 200 genes: The ranked position is not boosted, except where the random gene belonged to the highly connected region. See Figure 5.

5. Conclusions

- We have derived an algorithm which combines expression and connectivity data.
- The algorithm is robust to noise and may also highlight potentially active genes.
- The algorithm is an instance of Google’s™ personalised PageRank algorithm (3).
- Further testing is being carried out to quantify the level of preferential connectivity required for good performance of the algorithm.

6. References


Acknowledgement

This work has been funded by a Glasgow-Strathclyde University alliance, Synergy, providing a PhD studentship.
For further information see: http://www.strath.gla.ac.uk/synergy/