Selected Gene Groups

We selected 26 genes in 4 well-defined functional groups (genes encoding 10 glutamate receptor subunits, 7 enzymes in catecholamine metabolism, 5 cytoskeletal proteins and 4 enzymes in tyrosine and phenylalanine (Y/F) synthesis).

Precision-Recall

The number of true positives (tp; words in the algorithm-derived list also found in the investigator-derived list), false positives (fp; words in the investigator-derived list not found in the algorithm-derived list) and false negatives (fn; words in the investigator-derived list not found in the algorithm-derived list) were counted and then the Precision (P; the fraction of retrieved keywords that were retrieved) were calculated for the results from the algorithm under various parameters (see above) as follows:

\[ P = \frac{tp}{tp + fn} \]

\[ R = \frac{tp}{tp + fn} \]

Error-minimization

The optimum combination of these parameters plus the z-score threshold was determined by minimizing the function 

\[ (P \times R) - \alpha \times \text{error} \]

including low-quality keywords is weighted more heavily than the cost of exciting high-quality keywords (Huisden et al. 2001). This function is plotted on the left for one combination of parameters using \( r^2 = 4 \) and \( r^2 = 4 \). The right panel shows the minimum of each plot for the 32 individual combinations of parameters. Overall the best performance was achieved with the random background set, Porter strong stemming, the PD+ stop list and a z-score acceptance threshold of \( \pm 6-10 \).

3D-projection of Clusters

To visualize these clusters, the 26x26 matrix was collapsed to a 3x26 matrix by single value decomposition, a commonly used method for dimension-reduction. These three reduced dimensions were plotted against each other to visualize the 26 genes in three-dimensional space. Four functionally relevant clusters were produced. Once again, tyrosine transaminase was placed with the dopamine biosynthesis enzymes instead of the Y/F synthesis enzymes.

Functional Clustering of Genes

A 26x26 gene x gene matrix was created with the cells containing the sum of products of z-scores for shared keywords; larger values reflect stronger and more extensive keyword associations between genes-gene pairs. For each gene pair (A,B) and every word i they share, the cell values are \( \sum_{i=1}^{26} \frac{z\text{-score} \text{ (gene } A \text{) } \times z\text{-score} \text{ (gene } B \text{) }}{1000} \). A modified bond energy algorithm (Navathe et al., 1984) was used to group genes into 4 clusters based on shared keyword associations, and the resulting gene clusters are boxed with the appropriate color. The clustering program first sorts the symmetrical matrix by sequentially placing each column adjacent to the column it gives the maximum sum of products between adjacent cells (or their “bond-energy”). The algorithm then partitions the matrix by sequentially finding the maximum sum of quotients between adjacent cells. The bond energy algorithm correctly assigned 25 of 26 genes to the appropriate clusters. The single “mistake”, tyrosine transaminase, was assigned to the catecholamine metabolism cluster rather than the Y/F synthesis cluster.

Clustering using “CLASSIT”

Using the same 4 groups of 26 genes, we clustered them using a different algorithm, CLASSIT (a derivative of the unsupervised learning algorithm, COBWEB). Testing various parameters such as z-score threshold, similarity measures, term-frequency weighting, dimension reduction, the best results (as seen in the table below) were obtained using no z-score threshold, a cosine similarity criterion, no weighting, and 18 dimensions. All genes were sorted into their original groups except for tyrosine transaminase (as before), belonging to the “Y/F Synthesis” group but sorted with “Catecholamine Transaminase” as key. Thus, keyword associated with genes provide sufficient information (i.e. distinctive features) to cluster those genes into functionally relevant groups by two different clustering approaches.

Introduction

To facilitate the interpretation of large data sets generated by DNA microarray studies, we are developing a text mining system to extract keywords from MEDLINE abstracts associated with individual gene names and 2) investigating several clustering algorithms to determine relative distances between genes based on shared keywords. The basic mechanisms of our keyword extraction algorithm was described previously (Soc Neurosci Abstr 2001, 557.4). Recent progress in evaluating the performance of this algorithm through Precision-Recall calculations and in using extracted keywords to accurately cluster predefined groups of genes are reported here.

KEYWORDS ASSOCIATED WITH GENE CLUSTERS ARE INFORMATIVE

We provided the four lists of shared keywords to approximately 20 students, postdoctoral fellows and faculty, asking them to guess a major function of the genes in each cluster. Even though this was an informal survey, the finding that a large majority of guesses were accurate adds credence to the notion that our clustering and keyword lists can be useful in allowing rapid sorting and evaluation of large lists of genes identified in microarray experiments.

Conclusions

1. Precision-Recall methods indicate that the best performance of our keyword selection algorithms is obtained with the Random background set, Porter stemming, and an extensive, customized stop-list.

2. Clustering methods can reconstruct known functional gene groups by shared keyword associations.

3. Extracted keywords are informative of the common underlying function of clustered genes, and can therefore aid scientists in determining the functional basis of a gene group.