Gene expression

Gene-Expression Omnibus integration and clustering Tools in SeqExpress

John Boyle
Holbeck, George Street, Cambridge CB4 1AJ, UK

Received on September 9, 2004; revised on February 21, 2005; accepted on February 23, 2005
Advance Access publication March 3, 2005

ABSTRACT
Summary: SeqExpress, a gene-expression analysis suite, has been extended to offer a number of cluster generation, refinement and visualization techniques. The cluster generation methods have been specialized to deal with aspects of the sparseness and extreme values that occur within microarray data. The results of such cluster analysis can then be refined using either: a functional enrichment based procedure, which examines each cluster to see if it possesses an unusually high or low concentration of ontology terms; or by using Expectation–Maximization to find a mixture of model based distributions within the datasets. Visualizations are provided both to explore and compare the results of the cluster generation algorithms. In addition, a tool has been developed which integrates SeqExpress with the Gene-Expression Omnibus repository. The tool provides seamless access to the large number of experimental results in the repository, so that they can be visualized and analysed locally using SeqExpress.

Availability: SeqExpress is available as a 6 MB download from http://www.seqexpress.com and runs under Windows. A server-based version is available and is required for the GEO integration. SeqExpress is not affiliated with any academic institution, funding body or commercial organization and is free to use by all.

Contact: john@seqexpress.com

INTRODUCTION
The SeqExpress gene-expression application suite has been extended to provide integration with the Gene-Expression Omnibus (GEO) (Edgar et al., 2002). In addition, a number of cluster generation, refinement and visualization techniques have been implemented. This functionality is incorporated into SeqExpress along with a number of new data transformation, projection, visualization and analysis options. These extensions expand the functionality of the previous implementation of SeqExpress which was originally designed as a visualization tool (Boyle, 2004).

CLUSTER GENERATION
Four clustering generation techniques have been implemented within SeqExpress.

Distance measure based generation. Clusters are derived on the basis of the numerical experimental results, i.e. based solely on the geometry of the gene-expression vectors. A variety of distance measures can be used to compare the gene vectors, including Euclidian and Pearson distance-based techniques. Depending on the data distribution, initial centroids can be generated or chosen randomly from the data points within the dataset. By using the data points themselves as the initial centroids, rather than the points with the farthest distance or randomly generated points, a large number of clusters can initially be generated for areas of high density within the expression levels. To help alleviate problems due to areas of sparsity or little local variation during an iteration cycle it is possible to remove clusters that have less than a user-defined minimum threshold of members. To minimize the effect of outliers, an anchoring procedure can be selected, so that the resulting centroids within a sample are adjusted at each iteration to match those of a ‘real value’.

Graph based generation. In SeqExpress graphs are calculated and then partitions are generated using Metis (Karypis and Kumar, 1998). Graphs can be generated either based on predefined relationships that are specified in an ontology or by calculating the minimal spanning tree. If an ontology is used to generate the graph then the edges between the genes are defined by the ontology, and the length of the edges corresponds to a distance function between the expression profiles. In this manner, the ontology represents the knowledge of how different factors could cause the specific expression profiles by describing the groupings and relationships of gene products. By default, ontologies based on the Gene Ontology (GO) (The Gene Ontology Consortium, 2000) are used.

Heuristics based generation. A specialized workflow has been designed to robustly deal with SAGE experiment data. First the initial components of the system are derived, then a modified version of normal based Expectation–Maximization is used to find subdistributions within the data. After a level of convergence is reached, the models are adjusted, either by merging similar distributions or dividing ones with high standard deviations, and the iterative procedure is then repeated.

Hierarchy based generation. The hierarchical clustering techniques within SeqExpress build a graph structure which represents the differences between the different genetic profiles, and the results are then visualized. SeqExpress has two different hierarchical clustering algorithms.

(1) Semi-discrete decomposition (SDD) builds the hierarchy by establishing how the genes contribute towards significant areas of local density (peaks/troughs). This technique uses an SDD (O’Leary and Peleg, 1983) to compare local subspaces within the expression data. SDD is used to identify a predefined number of significant ‘bumps’ (areas of high density).
in a matrix, and generate approximations which describe how data items are affected by them.

(2) Hierarchical clustering builds the hierarchy by establishing which two genes are the closest to each other, then combining these into a single node and repeating until the tree is complete. The definition of closeness depends on a distance model; SeqExpress provides a number of such distance models.

**CLUSTER REFINEMENT**

SeqExpress can be used to apply two types of refinement operation to alter the shape, size and properties of a set of clusters:

*Category refinement.* Clusters are examined to see if they exhibit functional enrichment. The clusters are analysed for unusually high or low concentration of a particular category. The refinement mechanism uses a ‘user-defined’ ontology graph which has associated gene instances. The terms in the graph, and potentially associated generalizations, are examined to see if the specific cluster shows any functional enrichment. The functional enrichment is modelled as a hyper-geometric distribution.

*Mixture model refinement.* Expectation–Maximization is used to find a mixture of model based distributions within a dataset. The clusters are used to define the initial models within the system, the refinement procedure then iterates until the solution fails to improve. B-Spline, Pearson, Euclidian and Normal based models are supported. To enable better fitting a free-energy parameter can be either manually defined or automatically discovered.

**CLUSTER VISUALIZATION**

Further comparison and exploration of the cluster analysis is possible using an intracluster visualization tool. This tool uses both hierarchical and parallel plots to visualize similarities between the generated clusters. The validity of refined clusters can be also be explored: if category refinement has been used then the probability (and precision) scores are shown for the terms for which the cluster exhibits enrichment; and if mixture-modelling was used then the contribution that each gene made towards the cluster model can be viewed.

In addition, an intercluster visualization tool is provided to compare the results of two different cluster analyses or the results of a cluster analysis against a biological relevant categorization.

**GEO INTEGRATION**

The GEO integration tool provides a convenient means for the local analysis of remote, publicly available experiments. As the tool is designed for convenience, as much of the data retrieving, file parsing and database loading process is automated (Fig. 1).

The tool synchronizes with GEO by scanning the repository contents so that new items can be flagged and salient information retrieved (title, keywords, species and other experimental details). The experimental datasets and series can be browsed locally, and then marked for retrieval. Inbuilt customizable rules are used to map between the different annotation types, allowing for the automatic mappings of identifiers. So, if a platform file has Unigene identifiers, then the GO terms for each entry will be resolved at run time (by matching the platform row id to the Unigene cluster, then to the corresponding set of LocusLink ids and finally to the GO terms).

Where possible, a ‘one-click’ access approach is adopted, so that associated data files (e.g. platform or annotation files) are retrieved automatically and database configuration rules are used for parsing. In situations where choices have to be made (e.g. experiment selection from data files, choices on which annotations are required), simple wizards are provided.

**ACKNOWLEDGEMENT**

The author would like to thank Tanya Barrett from the NCBI for providing advice on the structure and format of GEO, and George Karypis from the University of Minnesota for the permission to use Metis.

**REFERENCES**


