GEMS Algorithm Description

The details of the preliminary design of the GEMS algorithm can be found in the original publication (1), which is available on the web page (http://genomics10.bu.edu/terrence/gems/). For completeness, we review the algorithm below. Figure 1 provides an outline of the GEMS. A gene expression bicluster is defined as a subset of genes having a consistent (constrained) expression over a subset of conditions. The consistency of gene expression is defined by a range constraint (width = $W$) in the selected subset of conditions. If genes are conserved only in a small number of samples, this module will be over-specific. GEMS searches for a maximal subset of genes and set of samples such that the number of samples included in the bicluster is large enough to meet a size constraint $\alpha$ (fraction of samples). This objective maximization criterion aims to identify statistically significant biclusters.

More formally, we are given a $G \times S$ matrix $M$ where the rows and the columns of the matrix correspond to genes and samples respectively. $M_{i,j}$ is the expression of gene $G_i$ in sample $S_j$. The gene expressions are normalized to be in the (0,1) range.

Geometrically, the data can be seen as a set of $S$ point in a $G$-dimensional unit cube. We are seeking to identify an axis parallel hyper-rectangle (ap-bicluster) in unit cube that includes a subset $S'$ of $S$ that contains at least $K = \alpha S$ samples, such that the number of genes included in the ap-bicluster is maximized. Each side of the axis parallel hyperrectangle is at most $W$ long.

We note that this definition makes our biclusters somewhat unique. A “traditional” notion of a bicluster is defined by a subset of rows and columns in the matrix $M$ such that the subset of samples cluster in the subset of dimensions (columns) included in the bicluster. We impose an additional constraint of width on each dimension (gene) that requires that the set of gene expression for this gene among the selected samples is constrained to be at most $W$.

Let us consider the statistical motivation for this problem. Consider as a NULL model a random process that “sprinkles” points in a $G$-dimensional unit cube using the uniform distribution. In this case the rough estimate of the log-probability of obtaining $K$ points in a small hypercube of volume $W^g$ ($W < 1$) is $\log (W^g)^K = K g \log W$. This probability provides a p-value on the significance of obtaining $K$ points in this volume. Since $K$ and $W$ are fixed, to maximize the significance of the ap-bicluster, we would like to identify the one with the maximum number of genes in it (smallest P-value).
Figure 1. Workflow chart of the GEMS algorithm. $\alpha$: size constraint parameter. $W$: range constraint parameter. $S$: number of samples in microarray dataset.
In the algorithm sketched below we describe an MCMC procedure that samples from a distribution over volumes in a unit cube where the log-probability of a bi-cluster of volume at most $W^g$ is $K g \log W$. Our current approach is based on a Gibbs Sampling paradigm, we plan to extend it to Metropolis style algorithm in the future.

Since microarray data are not selected from a uniform distribution, we need to obtain a better assessment of the probability of having a random hyper-rectangle of volume at most $W^g$ to contain a subset $K$ points in dataset $D$, which we refer to as $P(g | D, W, K)$. In order to estimate this probability, we use Gibbs sampling to sample the space of “statistically significant” biclusters. The process is described below.

**Initiation:**

We initiate the process by choosing a random set of $K$ samples. Once these samples are selected, for each gene $G_i$ we test whether the range of gene expression values for these $K$ samples is smaller than $W$. If yes, the gene $G_i$ is included in the bicluster, otherwise the gene is not. This can be determined in $O(KG)$ time. This is one of key ideas motivating our methodology since it eliminates the possibly exponential search for genes to be included in the ap-bicluster.

**Sampling**

Sampling the space of biclusters is achieved by considering a single sample in the bicluster for replacement with another out-of-cluster sample. To accomplish this, we first estimate the probability of a new subset $S_u$ with $g_u$ genes to be sampled next. This probability is given by

$$P_{\text{sample}}(S_u) = \log P(g_u | D, W, K) / \sum_{u' \in \text{out-samples}} \log P(g_u' | D, W, K) = g_u / \sum_{u' \in \text{out-samples}} g_u'$$  \hspace{1cm} (1)

Where $S_u$ is the subset that results from the replacement with $u$-th out-of-cluster sample, and $g_u$ is the number of genes in this ap-bicluster. During the sampling iterations, we keep track of the changes of the number of genes, and record the maximum gene number $g_{\text{max}}$. We now choose a candidate bicluster with probability $P_{\text{sample}}$ and continue this process until $g_{\text{max}}$ does not change in a predefined number of sampling iterations. We also count the number of biclusters containing $g$ genes obtained during sampling. This provides an empirical estimate of the probability of seeing a bicluster with $g$ genes among the sampled ap-biclusters.
Maximizing

We then select the bicluster containing the sample set $S_{\text{max}}$ with the maximum number of genes ($g_{\text{max}}$). The P-value associated with this bicluster is the empirical frequency of biclusters with $g_{\text{max}}$ genes encountered during the previous Sampling phase.

One variant of the algorithm that will be useful is to update the probability $P_{\text{sample}}(S_u)$ based on the empirical frequency of $g_u$ observed during the Gibbs sampling and use that globally estimated probability instead of the local estimate obtained in equation 1.

Local Search

The bicluster acquired in previous stage is further refined by a local search step. In each step a single sample in the bicluster is considered for replacement with an out-of-cluster sample. Every out-of-cluster sample replacement makes a new bicluster, and the one associated with most genes will be the new bicluster. The local search step iterates until there are no further changes (convergence is easily guaranteed since the number of genes is a monotonic criteria bounded above by $G$. The bicluster becomes optimal in the sense that it cannot be improved by a single replacement.

REFERENCES