Consensus Clustering

We have developed an algorithm for generating consensus clusters. Firstly, an upper triangular $n \times n$ agreement matrix is generated with each cell containing the number of agreements amongst methods for clustering together the two variables, represented by the indexing row and column indices (see Figure 1). This matrix is then used to cluster variables based upon their cluster agreement (as found in the matrix).

$$\begin{array}{cc}
 & \text{To Gene} \\
\text{From Gene} & \begin{bmatrix}
0 & A_{12} & A_{13} & A_{14} & \cdots & \cdots & \cdots & \cdots & \cdots & A_{1n} \\
\vdots & 0 & A_{23} & A_{24} & \cdots & \cdots & \cdots & \cdots & \cdots & A_{2n} \\
\vdots & \vdots & 0 & A_{34} & \cdots & \cdots & \cdots & \cdots & \cdots & A_{3n} \\
\vdots & \vdots & \vdots & \ddots & \ddots & & & & & \vdots \\
\vdots & \vdots & \vdots & & \ddots & \ddots & & & & \vdots \\
\vdots & \vdots & \vdots & & & \ddots & A_{ij} & & & \vdots \\
\vdots & \vdots & \vdots & & & & \ddots & \ddots & & \vdots \\
\vdots & \vdots & \vdots & & & & & \ddots & \ddots & \vdots \\
\vdots & \vdots & \vdots & & & & & & \ddots & A_{(n-1)n} \\
0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0
\end{bmatrix}
\end{array}$$

Figure 1. The Agreement Matrix, *A*

Consensus clustering attempts to maximise a metric that rewards variables in the same cluster if they have high agreement and penalises variables in the same cluster if they have low agreement. In particular, the algorithm tries to maximise agreement using the function in Equation (1) to score each cluster of size s_i, where A is the agreement matrix, G_{ij} is the jth element of i, G_i, and α is a user-defined parameter (the agreement threshold), which determines whether the score for the cluster is incremented or decremented.

$$f(G_i) = \sum_{j=1}^{s_i - 1} \sum_{k=j+1}^{s_i} (A_{G_{ij}G_{ik}} - \alpha) \tag{1}$$

If α is equal to *Min*, the minimum value in *A*, then clearly Equation (1) is maximised when all variables are placed into the same cluster. Alternatively, when α is equal to *Max*, the maximum value in *A*, Equation (1) is maximised when each variable is placed into its own cluster. Therefore, a sensible value for α should be chosen that lies between the minimum and the maximum agreement. The α parameter will have a large effect on the content of the final clusters and so this must be chosen carefully. Based upon the distribution of the agreement matrix, *A*, in a number of datasets, we have found that a good value for α is: (*Max* + *Min*)/2, where *Max* is the maximum value in *A* and *Min* is the minimum.

The search algorithm is a form of Simulated Annealing which is an iterative improvement search technique that starts with a random solution to a given problem, and then tries to increase its worth by a series of small changes. If such a small change is better than the previous solution then further changes are made from this new point. But when the new solution is worse than the old one, it is not discarded, but accepted with a probability according to Equation (2),

$$\Pr(\text{accept new}) = e^{-x}, x = \frac{E_{new}(G') - E_{old}(G)}{\theta_t} \tag{2}$$

where $E_{new}(G')$ is the new value for the validity of the new cluster arrangement, G', and $E_{old}(G)$ is the value before the change was made. The value θt is referred to as the temperature at iteration t, where θ_t is updated to $c\theta_{t-1}$ after each iteration where $0<c<1$ is a constant. The required number of clusters is represented as lists of variables and a change is a random move of a variable from one cluster to another. The algorithm is shown below.

```
Input: Agreement Matrix (n×n), $A$; Number of Clusters sought, $m$; Number of Iterations, $Iter$; Agreement Threshold, $\alpha$; Initial Temperature, $\theta_0$; Cooling Rate, $c$
Generate $m$ empty clusters
Randomly distribute the variables (genes) 1..$n$ between the $m$ clusters
Score each cluster according to Equation (1)
For $i$ = 1 to $Iter$ do
    Move a variable (gene) from one random cluster to another
    Set $Change$ to difference in score according to Equation (1)
    If Change < 0 Then
        Calculate probability, $p$, according to Equation (2)
        If p > random(0,1) Then Undo Move
    End If
    $\theta_i = c\theta_{i-1}$
End For
Output: Set of Consensus Clusters
```