How many Clusters: A Validation Index for arbitrary shaped clusters.
A. Baya & P. M. Granitto

Additional file 2: Full Figures

Figure 2 compares the clustering solution of 100 sub-samples, where each run considers 95% of the patterns of the dataset from Table II. The (x) axis shows number of cluster and the (y) axis shows the mean value of the cRand (corrected Rand index) or ARI (adjusted Rand index) for the three selected clustering algorithms (EAC-av, PKNNG-av and Spectral Clustering). This figure presents the performance of the three algorithms based on the golden rule i.e. biological classes. The peaks of cRand indexes coincide with the number of biological classes (Table II shows the number of biological classes for each case). In all examples bellow, Figure 2 show that EAC-av, PKNNG-av and Spectral Clustering find clusters related with the biological classes. This set of solutions are used by the automatic validation methods to look for the most likely number of clusters. The graphical methods used to compare to the IC-av index have a different sub-sampling scheme and they do not use this results.

An analysis of THY data with EAC-av and PKNNG-av showed that solutions beyond 3 clusters tend to have outliers. To control this behaviour the solutions with groups formed by one point were not considered as clusters. Consistent with the previous behaviour, Panel (a) from Figure 2 show that cRand is almost null for EAC-av and PKNNG-av solutions with more than 3 clusters.

Fig. 2. Clustering performance. Panels (a) to (f) show the performance of 3 clustering algorithms, EAC-av (green), PKNNG-av (blue) and Spectral Clustering (black) on six real gene expression data.