# BSC: a clustering program for DNA array expression data

G. Iazzetti[1], V. CalabrÚ[1], S. Saviozzi[2], A. Weisz[3], L. Lania[1], R. Calogero[2]

1 Dipartimento di Scienze Genetica, Biologia Generale e Molecolare, Università di Napoli Federico II, Via Mezzocannone 8, 80134 Napoli
(Italy)
2 Dipartimento di Scienze Cliniche e Biologiche, Università di Torino, Ospedale S. Luigi, Regione Gonzole 10, 10043 Orbassano - TO (Italy)
3 Istituto di Patologia Generale e Oncologia, Seconda Università di Napoli, Larghetto S. Aniello a Caponapoli 2, 80138 Napoli (Italy)

In any living cell that undergoes a biological process, different subset of its genes are expressed in different stages of the process. The particular genes expressed at a given stage and their relative abundance is crucial to the cell's proper function. Measuring genes expression levels in different developmental stages, different bodies tissues, different clinical conditions and different organisms is instrumental in understanding biological processes. Therefore, large-scale gene-expression analysis can accelerate the discovery of key biological processes. Independent of whether the data set originate from drug responses, disease models or molecular-anatomy studies, the first step in any analysis begins with the categorization of expression patterns according to their similarity and existing functional annotations (clustering). In cluster analysis the aim is the partition of entities into group based on given features of each entity so that the groups are homogeneous and well separated.

Commonly used clustering methods include hierarchical (Eisen, et al. 1998, Proc. Natl. Acad Sci. 95, 14863-14868; DeRisi et al. 1996, Genetics 14, 457-460; Spellman et al. 1998, Mol. Biol. 9, 3273-3297), k-mean clustering (White et al. 1999, Science, 286, 2179-2184), self-organizing maps (SOM) (Tamayo et al. 1999, Proc. Natl. Acad. Sci. 96, 2907-2912) and graph-theoretic approaches (Ben-Dor et al. 1999, J. Comp. Biol., 6, 281-297). Hierarchical clustering divides genes into a strict divisive hierarchy of nested subsets. In particular, a hierarchical agglomerative processing consists of repeated cycles where the two closest remaining items (those with the smallest distance) are joined by a node/branch of a tree, with length of the branch set to the distance between the joined items. The two joined items are removed from the list of items and they are replaced by an item that represent the new branch; the process is repeated until only one-item remains. Instead, SOMs allow the imposition of partial structure by specifying an initial number of N clusters that converge on the optimal clusters. K-mean clustering requires the specification of the number of clusters, N, as the SOMs, but it does not allow the specification of their optimal position in the k-dimensional space.

Moreover, the k-mean clustering is faster then SOMs but the N clusters may not converge on the optimal clusters as well as in the case of SOMs. In this report we describe the BSC (BestScoreClustering) clustering program that build clusters as unrelated entities, which are subsequently ordered and portrayed on the bases of their similarities. The program runs on Win98, and WinNT operating systems and it can be download free of charge together with an help file and a sample data set at http://dscb041.sluigi.unito.it/BSC/BSC_welcome.htm. The clustering approach implemented in BSC not only keep in consideration the similarity existing between the series of data but uses also other parameters, which can be important to outline similarity between genes (Chromosome location, metabolic pathway, biological function, etc). The similarities between the data series are defined using a Score Similarity Value (SSV). The SSV varies between 1 and 0, with 1 meaning that the two series are identical and 0 meaning they are completely independent. SSV embeds various parameters including Pearson correlation coefficient (PCC), chromosomal localization, gene function, metabolic pathway and any other peculiar gene characteristic defined by the user. BSC clustering in contrast with SOM and k- mean clustering does not assume a given number of clusters and an initial spatial structure of them, but finds out cluster number and structure based on the data. Moreover, k-mean clustering and SOM are characterized by the random selection of the genes that will be used in the N clusters as centroids, this approach induces some fluctuations in the clusters composition from run to run performed on the same set of genes. This problem is not present in BSC because the seeding couple of the first cluster is that showing the highest SSV within the analysed data. BSC generates clusters characterized by items having a SSV higher of a Score Similarity Cut off Value (SSCV), which can be set by the user (single clustering) or defined by the program via a recursive analysis (loop clustering). The loop clustering generates various set of clusters characterized by a score value describing the quality of the generated cluster. The use of a low SSCV does not influence much the size of the clusters but frequently allows the creation on new clusters containing the items rejected using a more stringent SSCV. To outline the clusters characterized by very similar shapes BSC ordinates them by their similarity essentially as described for the gene clustering.