BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 9 2003, pages 1153–1154
DOI: 10.1093/bioinformatics/btg143

Coupled two-way clustering server

Gad Getz* and Eytan Domany

Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel

Received on December 25, 2002; revised on November 4, 2002; accepted on November 6, 2002

ABSTRACT

Summary: The CTWC server provides access to the software, CTWC1.00, that implements **C**oupled **T**wo **W**ay **Cl**ustering (Getz *et al.*, 2000), a method designed to mine gene expression data.

Availability: Free, at http://ctwc.weizmann.ac.il.

Contact: ctwc_support@weizmann.ac.il

Supplementary information: The site has a link to an *example* which provides figures and detailed explanations.

A DNA chip experiment provides expression levels, E_{gs}, of thousands of genes g for up to 100 samples s, summarized in an *expression table* of $\approx 10^6$ entries. Analysis of such data has several aims: (1) identify genes whose expression levels reflect biological processes of interest (such as development of cancer); (2) group the samples (e.g. tumors) into classes, possibly in a clinically relevant way, and (3) provide clues for the function of genes (proteins) of yet unknown role.

First one filters the genes (Alon *et al.*, 1999), leaving a set $G1$ to work with. Next, cluster all genes of $G1$ on the basis of their expression levels over the set of all samples, $S1$ [an operation denoted by $G1(S1)$], and cluster $S1$ using all the genes of $G1$ [$S1(G1)$]. In general, however, only a small subset of N_r genes are relevant for one particular biological process of interest. Since usually $N_r \ll |G1|$, the 'signal' of these genes may be masked by the 'noise' generated by the (much more numerous) other genes. Furthermore, to assign samples into two clinically meaningful classes (e.g. adenoma and carcinoma), we may have to remove first a previously identified group of samples (e.g. healthy tissue), and cluster only the remaining $N_s' < N_s$ tumors. Thus one should analyze, one at a time, special *submatrices* of E_{gs}. CTWC (coupled two way clustering) is a heuristic, iterative method to search for informative $N_r \times N_s'$ submatrices among the exponentially many possible ones. In the first two steps, $G1(S1)$ and $S1(G1)$, we identify and *register* stable, statistically significant clusters of genes, GI with $I = 2, 3, \ldots$ and of samples, SJ, $J = 2, 3, \ldots$. Next, we cluster every one of the stable sample groups SJ (including $S1$), using the expression levels of every stable gene group GI, one at a time. Such a clustering operation, denoted by $SJ(GI)$, may generate new stable sample subgroups. Similarly, one reclusters every gene group GI on the basis of every sample group SJ. New stable gene and sample clusters that emerge are added to the respective registers and used in the next iterative step, until the emergent new clusters are smaller than some preset threshold. A typical positive finding of the method is such a statement (Getz *et al.*, 2000): 'A particular group of samples SJ (e.g. patients suffering from ALL leukemia) breaks into two clear subgroups (e.g. T-ALL and B-ALL) on the basis of the expression levels of a group of genes GK'.

CTWC uses as its 'clustering engine' an algorithm called superparamagnetic clustering (SPC) (Blatt *et al.*, 1996). SPC places in the same cluster objects that are 'close' to one another, producing a dendrogram, as a parameter T, that controls resolution, is varied. SPC is stable against addition of noise to the data and can identify irregular shaped clouds of points as clusters. Most importantly, SPC provides for each cluster a 'stability' index, whose value is indicative of the extent to which the cluster is 'real', and not due to noise in the data. The index is based on the physical intuition that underlies SPC; a stable cluster behaves as an independent 'ordered magnetic grain' for a wide range of values of T (Blatt *et al.*, 1997). Using this index we exhaustively scan (and cluster, one at a time) those submatrices, whose genes and samples constitute stable clusters. CTWC has been used successfully to study data from experiments on colon cancer, leukemia (Getz *et al.*, 2000), breast cancer (Kela, 2002; Getz *et al.*, 2003), glioblastoma (Godard *et al.*, 2003), skin cancer (Dazard *et al.*, 2003) and antigen chips (Quintana *et al.*, 2003).

THE SERVER is frequently updated. Here we present a detailed, step by step 'roadmap' of the server, from data entry to viewing the results. We recommend that the instructions be read while viewing the *example (ES)* found at the CTWC site.

Data preparation and Entry: Filter the genes down to $|G1| < 3000$ (in our example we kept 2000 genes). The resulting matrix E_{gs} is uploaded in the format used in `Cluster` (Eisen *et al.*, 1998), of an ASCII table separated by tabs (see *ES* links 1,2). Three optional preprocessing

*To whom correspondence should be addressed.

operations, can be performed after uploading the data matrix; *Scaling*, *Thresholding* and *Log* (see *ES* link 3, where only the first two were performed). Optionally the user may upload also a $P \times N_s$ table of P 'predefined sets', whose entry L_{is} can be 1/0/blank, indicating that sample s belongs to set i/does-not-belong to set i/has unknown assignment (the example includes $P = 4$ categories; tumor,normal, protocols A and B—see *ES*, links 4,5). One can upload a similar table for genes.

Creating Projects and Analyses: Each user creates projects in his account. A *Project* is related to a dataset E_{gs} and to two tables of predefined sets. Every project may contain several *Analyses*; each uses a particular set of running parameters. Within an analysis there are *processes*; each is a CTWC run defined by its initial gene and sample sets and the desired iteration depth.

An analysis is specified by its clustering parameters for SPC (try first the default values!), which are explained in the site. Here we mention only *Min T*, *Max T* and ΔT, that govern the range and step size that specify the parameter T, which controls the resolution. At '*Min T*' there should be a single cluster, and at '*Max T*' many small clusters.

Another set of parameters, used by CTWC, define a stable cluster: (a) a '*minimal cluster size*' must be exceeded; (b) the number of cluster members lost, when T increases by ΔT, must be less than '*ignore dropout size*'; (c) conditions (a,b) must hold for at least '*stable delta T*' temperature steps. Clusters that qualify as stable are used in subsequent CTWC iterations.

Execution of Analysis: $G1/S1$ are the default for the initial gene/sample clusters used. In subsequent runs one can apply CTWC to a sub-matrix, defined either by a stable cluster that was found in a previous *Process*, or by one of the predefined sets. Specify the iteration depth of CTWC: try first '*depth*'=1 for samples and genes, performing $G1(S1)$ and $S1(G1)$. If the parameters gave suitable results, proceed to deeper levels (see *ES* link 8). Starting the analysis invokes a run; upon completion it generates output files and notifies the user by e-mail.

Results: Each execution generates *results* pages. The main one lists all stable gene (GI) and sample (SJ) clusters. Our example uses depth 1 for genes and 2 for samples (see *ES* link 9), showing for each stable cluster its stability index, the clustering operation in which it was found, and a table of all the clustering operations that were applied to it, and the clusters found by them.

Additional tables (for genes and for samples) relate stable clusters to the predefined labels. Each stable cluster is represented by a row and each predefined label by a column. The table element of cluster Cx and set Py contains the purity ($|Cx \cap Py|/|Cx|$) and efficiency ($|Cx \cap Py|/|Py|$) indices that measure the extent to which Cx captures Py, and a score that measures the likelihood to obtain such overlap by chance. Significant overlaps are linked to the clustering operation that found them, allowing an easy search for clusters that capture known sets in the data. In the example $S7$ overlaps with normal samples and $S5$ with protocol B. The links show that $S7$ was found in $S1(G5)$ whereas $S5$ was identified in $S1(G4)$. This example demonstrates how different sets of genes (e.g. $G4$, $G5$) can yield very different separations of the samples ($S1$).

Links from the main *results* page point to two kinds of pages. (a) A cluster page contains a list of its members and whether they belong to predefined sets (see *ES* links 10,11). (b) A page describing a clustering operation, containing tables and figures (see *ES* links 12–14), such as a dendrogram, depicting hierarchical partitioning of the data, and the distance matrix, which shows the distances between the clustered objects (genes or samples), after reordering them according to the dendrogram.

ACKNOWLEDGEMENTS

Funding for the server was provided by Yeda and the Levine Fund.

REFERENCES

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Blatt,M., Wiseman,S. and Domany,E. (1996) Superparamagnetic clustering of data. *Phys. Rev. Lett.*, **76**, 3251–3254.

Blatt,M., Wiseman,S. and Domany,E. (1997) Data clustering using a model granular magnet. *Neural Comp.*, **9**, 1805–1842.

Dazard,J.-E., Gal,H., Amariglio,N., Rechavi,G., Domany,E. and Givol,D. (2003) Genome-wide comparison of human keratinocyte and squamous cell carcinoma responses to UVB irradiation: implications for skin and epithelial cancer. *Oncogene*, in press.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Getz,G., Levine,E. and Domany,E. (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.

Getz,G., Gal,H., Kela,I. and Domany,E. (2003) Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. *Bioinformatics*, **19**, 1079–1089.

Godard,S., Getz,G., Delorenzi,M., Kobayashi,H., Farmer,P., Nozaki,M., Diserens,A.-C., Hamou,M.-F., Dietrich,P.-Y., Regli,L. *et al.* Taxonomy and Classification of Human Astrocytic Gliomas on the basisof gene expression, submitted.

Kela,I. (2001) Clustering of gene expression data, M.Sc. Thesis, Weizmann Institute.

Quintana,F., Getz,G., Hed,G., Domany,E. and Cohen,I.R. (2003) Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: A bio-informatic approach to immune complexity, submitted.