PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

New clustering method for expression array data

ArticleInfo		
ArticleID	:	3667
ArticleDOI	:	10.1186/gb-2000-1-6-reports0078
ArticleCitationID	:	reports0078
ArticleSequenceNumber	:	17
ArticleCategory	:	Paper report
ArticleFirstPage	:	1
ArticleLastPage	:	4
ArticleHistory	:	RegistrationDate : 2000–10–27 Received : 2000–10–27 OnlineDate : 2000–11–26
ArticleCopyright	:	BioMed Central Ltd2000
ArticleGrants	:	

ArticleContext	:	130591166

Rachel Brem

Abstract

An algorithm has been described that clusters expression data from microarray experiments with respect to subsets of genes and other variables, rather than to all data at once.

Significance and context

Expression arrays measure the RNA levels of thousands of genes at once, from cells grown under many different conditions. Genomicists have examined microarrays in search of data that 'cluster', or behave similarly across many experiments. A usual strategy is to find genes that express RNA in similar global patterns across all conditions - for example, in all body tissue types. The same analysis can be used to cluster tissue types with respect to the genes they express. Though useful on a global level, this approach cannot always pick out patterns among small subsets of data. To see why, imagine a group of genes called Q genes whose RNA levels are identical in gut, brain and muscle; in skin, on the other hand, each Q gene's RNA level is different. The similarity of RNA levels does not extend to all tissues, so a global clustering method may not detect the Q group. Getz *et al*. have attempted to address this failing with a new procedure called coupled two-way clustering (CTWC). CTWC looks for clusters in small chunks of data, rather than in the entire array matrix as a whole. Thus in the Q group example, CTWC will focus on the chunk of the array containing gut, brain and muscle alone, and will find the relation among Q genes.

Key results

Getz *et al*. applied CTWC to two array data sets of human genes, each containing about 125,000 data points. The first contained RNA levels from bone marrow mononuclear cells of leukemia patients, the second from colon tumors and normal colon samples. CTWC run on these array data found 30 to 100 gene and patient clusters. Many of these could not be rationalized against known gene families or similarities among patients. However, when Getz *et al*. manipulated clusters by hand (for example, running CTWC on a subset of the data chosen by the user) some known features of the data emerged. Clusters of patients fell into subdivisions based on cancer type, mode of treatment or the experimental protocol used in extracting RNA. Clusters of genes agreed with known cell-type differences.

Methodological innovations

CTWC begins with a global clustering search in an array, over all genes and all body tissues (or other condition variables such as phase in the cell cycle). Getz *et al*. use standard quantitative definitions of 'similarity' during the clustering. Imagine that the global clustering procedure results in, among others, a gene cluster containing the genes g1, g2, g3 and g4, and a tissue cluster containing the tissues t1 and t2. Now the method focuses on the RNA levels of genes g1-g4 measured in tissues t1 and t2: CTWC looks for more clusters within just this small chunk of the array. This type of search then goes on within all other chunks formed in a similar way. New clusters are then recombined and examined for still more clusters. The cycle of clustering and subdividing goes on until no more groups of data meet the clustering criteria.

Links

The Computational physics group website of Getz *et al.* lists all clusters found in the experimental data.

Conclusions

Getz *et al*. conclude that CTWC finds clusters of genes and tissues which agree with known classifications in their test data; other clusters may lead to new hypotheses about regulatory cascades in cancer.

Reporter's comments

The CTWC method is a technical advance designed to reduce the signal-to-noise ratio in array data clustering. But to find biologically relevant clusters in this paper, Getz *et al*. had to analyze parts of CWTC's output by hand. It is not yet clear whether any given cluster from the automated form of CTWC will be useful as a hypothesis generator in biology.

Table of links

Proceedings of the National Academy of Sciences of the United States of America

Computational physics group

References

1. Getz G, Levine E, Domany E: Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A. 2000, 97: 12079-12084. 0027-8424