CellMontage - a fast database search system for gene expression profiles

Wataru Fujibuchi
Research Scientist
Sequence Analysis Team

With the rapid increase of gene expression profile data since the inventions of convenient experimental systems such as microarrays, methods for mining information from the database of gene expression profiles are in great demand. We are developing a system to detect similar cells, analogous to finding similar sequences with FASTA or BLAST. Our **CellMontage** system can quickly calculate gene expression correlation from large databases of gene expression profiles.

This system has three highly-useful advantages; (1)robust - as it uses the rank correlation coefficients which work for cross-platform comparison, (2) marker genes are not necessarily needed - as CellMontage calculates the whole trend of similarity from all the genes, and (3)fast in searching - as we devised an algorithm for quick search. Previous approaches have avoided using gene expression data across different platforms because the microarray systems are totally different among companies and it has been believed that meaningful cross-platform comparisons are too difficult. However we found that our system can correctly rank similar cell types to the top of the match list, mostly independent of the platforms. Furthermore, unlike the typical marker methods, the system can find the similar cell types without needing a thousands of predefined marker gene sets for all the cell types.

Currently, by linking to Gene Expression Omnibus database at NCBI in NIH, roughly 30,000 expression profiles are stored in CellMontage system, and more than 10,000 data are searchable for human profiles. Using this system as shown in Figure 1, one can quickly find the known cells most similar to the query profile. A user can assess if a new tissue, which is naturally unknown or made by tissue engineering, is truly what he/she thinks it should be. Thus one immediate application is for detecting contamination by surrounding tissue in cell samples. Also the search is extendable to cross-species profile comparison such as the similarity measurement between human and mouse cells - when their ortholog genes are used. With improved linkage to clinical records, the system may be applied to the disease diagnosis field in future. Above all, the speed is advantageous for microarray analysis. In fact, as shown in Figure 2 where 10,000 genes are used, the CellMontage calculates the similarity for 1,000 profiles in as fast as one second so that it can give a list of cell types similar to your query in a moment.

In addition, it is noteworthy that the data has the information of cell types (i.e., annotations) with their gene expression patterns. From this point of view, we started an taxonomic project for annotating human normal cell types based on multiple views. For instances, we challenge problems of defining cell types, such as "what is the morphological and structural differences between liver and muscle cells?" and "Can we predict those differences by their gene expression profiles?". A motto in the field of protein function prediction has been "from sequences to structures, and functions". We expect a similar phrase such as "from gene expression profiles to cell structures, and cell functions" will characterize the gene expression analysis field.

Reference

JP-AN 2004-280257, "Gene expression profile retrieval method, program and system", Wataru Fujibuchi and Paul B. Horton (27th Sep. 2004)

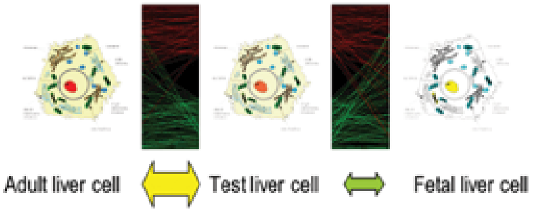
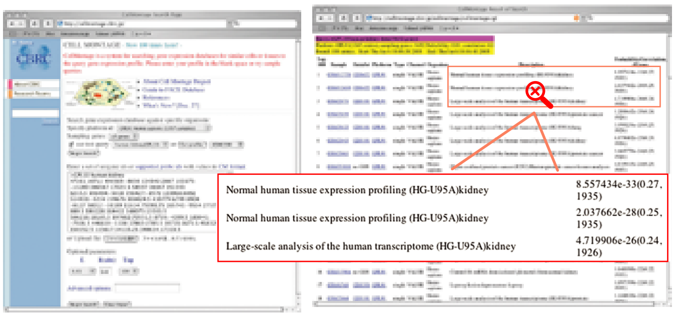


Fig.1

The principle of profile matching by rank correlation. When genes are ordered by their expression levels measured from an adult human liver and then compared with liver profiles from an adult and a fetus, it is observed that the gene expression orders are more conserved between adults than between adult and fetus. CellMontage system can quickly detect such differences and output the cells in the order of gene expression similarity. In the old marker methods that target only specific genes, the relationships of gene orders are not directly investigated. However, the matching methods based on the rank correlation can directly supply information of relative status of genes and their networks. To analyze how the gene expression status is related with the cell structure or function, we are developing a database of reference normal cells.



Normal human tissue expression profiling (HG-U95A)kidney	8.557434e-33(0.27, 1935)
Normal human tissue expression profiling (HG-U95A)kidney	2.037662e-28(0.25, 1935)
Large-scale analysis of the human transcriptome (HG-U95A)kidney	4.719906e-26(0.24, 1926)

Fig.2

CellMontage retrieval page (http://cellmontage.cbrc.jp/) (left) and the example results (right). When users input the UniGene numbers and their expression values as a query profile (left), the system will show the results by quickly search the database (right). Here, 1,345 oligonucleotide array profiles are searched with 9,432 UniGenes using a cDNA array profile from human kidney as a query, the system gave a result within 2 seconds and the top three are kidney or kidney-related data.