Cluster Analysis for Large Scale Gene Expression Studies

High throughput gene expression analysis is becoming more and more important in many areas of biomedical research. cDNA microarray technology is one very promising approach for high throughput analysis and provides the opportunity to study gene expression patterns on a genomic scale. Thousands or even tens of thousands of genes can be spotted on a microscope slide and relative expression levels of each gene can be determined by measuring the fluorescence intensity of labeled mRNA hybridized to the arrays. Beyond simple discrimination of differentially expressed genes, functional annotation (guilt-by-association) or diagnostic classification requires the clustering of genes from multiple experiments into groups with similar expression patterns. A platform independent Java package of tools has been developed to simultaneously visualize and analyze a whole set of gene expression experiments. After reading the data from flat files several graphical representations of hybridizations can be generated, showing a matrix of experiments and genes, where multiple experiments and genes can be easily compared with each other. Fluorescence ratios can be normalized in several ways to gain a best possible representation of the data for further statistical analysis. Hierarchical and non hierarchical algorithms have been implemented to identify similar expressed genes and expression patterns, including: 1) hierarchical clustering, 2) k-means, 3) self organizing maps, 4) principal component analysis, and 5) support vector machines. More than 10 different kinds of similarity distance measurements have been facilitated, ranging from simple Pearson correlation to more sophisticated approaches like mutual information. Moreover, it is possible to map gene expression data onto chromosomal sequences. The flexibility, variety of analysis tools and data visualizations makes this software suite a valuable tool in future functional genomic studies.

Keywords: microarray, cluster analysis, genomics, bioinformatics, java

Clusteranalyse von umfangreichen Genexpressionsdaten

Die Hochdurchsatz-Genexpressionsanalyse wird immer wichtiger in vielen Bereichen der biomedizinischen Forschung. Die cDNA Micorarray-Technologie ist ein sehr vielversprechender Ansatz für Hochdurchsatzanalyse und bietet die Möglichkeit, Genexpressionsmuster auf Genomebene zu studieren. Tausende oder sogar zehntausende von Genen können auf ein Mikroskopplättchen gedruckt werden, und die relative Expression von jedem Gen kann durch die Messung der Fluoreszenzintensität der zu den Arrays hybridisierten und markierten RNA gemessen werden. Geht man über die simple Unterscheidung von unterschiedlich exprimierten Genen hinaus, benötigen die funktionelle Annotation oder diagnostische Klassifikation das Clustern von Genen von multiplen Experimenten in Gruppen von ähnlich exprimierten Genen. Es wurde ein plattform-unabhängiges Java-Paket von Werkzeugen für die simultane Visualisierung und Analyse von ganzen Genexpressionsexperimentensets entwickelt. Nachdem die Daten eingelesen worden sind, können verschiedene graphische Repräsentationen der Hybridisierungen erstellt werden, die eine Matrix von Experimenten und Genen zeigt, mit der mehrere Experimente und Gene leicht miteinander verglichen werden können. Die Fluoreszenzverhältnisse können auf mehrere Arten normalisiert werden, um eine bestmögliche Repräsentation der Daten für die weitere statistische Auswertung erlangen zu können. Es wurden hierarchische und nicht hierarchische Clusterverfahren für die Identifikation von ähnlich exprimierten Genen und Expressionsmustern implementiert, einschließlich: 1.) Hierarchisches Clustern, 2) k-means, 3.) Self Organizing Maps, 4) Principal Component Analysis und 5) Support Vector Machines. Mehr als 10 verschiedene Arten von Ähnlichkeitsdistanzmessungen wurden implementiert, von der simplen Pearson Korrelation zu aufwendigeren Verfahren wie Mutual Information. Weiters ist es möglich, Genexpressionsdaten auf chromosomale Sequenzen zu projizieren. Die Flexibilität, Auswahl an Analysewerkzeugen und Datenvisualisierungen machen diese Software zu einem wertvollen Werkzeug für zukünftige Studien von Genomen.

Schlüsselworte: Mikroarray, Clusteranalyse, Genomische Forschung, Bioinformatik, Java