Poster L-37

Fuzzy Biclustering of Microarray Data - borrowing ideas from sequence profiles

Authors:
Stefan Tomiuk (*Miltenyi Biotec GmbH*)
Michael Brosda (*Miltenyi Biotec GmbH*)
Corinna Scholz (*Miltenyi Biotec GmbH*)
Kay Hofmann (*Miltenyi Biotec GmbH*)

Short Abstract: We present a novel approach to biclustering that is seeded by a small group of genes known to share a common regulator. The experiments are weighted by their support for the intra-seed similarity, and this weighting gets refined while new members of the group are being added in an iterative process.

Long Abstract:
Hierarchical clustering of microarray data is a widely used method for identifying groups of co-regulated genes (over multiple experiments), or alternatively, groups of similar experiments (over multiple genes). Often, however, conventional clustering methods fail to identify biologically meaningful similarities. Problematic are e.g. expression similarities between genes that are limited to a small fraction of experiments that activate a common regulatory mechanism. Different regulators, which are active in the other experiments, may have divergent effects on the members of the co-expression group, making their detection very difficult. This problem can be adressed by so-called biclustering methods, which aim to define subsets of genes that are co-regulated in a subset of experiments.
We present a novel approach to biclustering, which makes use of the analogy between the comparison of transcription profiles and that of biological sequences. Sequence analysis is a relative mature area of bioinformatics and has produced a number of algorithms for sensitive sequence comparisons and database searches. One of these methods, the ‘sequence profile’ technique, exploits the fact that in biological sequence families some residues are highly conserved, while other residues are more variable. The incorporation of prior knowledge on residue conservation greatly facilitates the search for new members of an existing sequence family.
This situation is perfectly analogous to a biclustering setting, where a small number of genes is known to share a common regulatory mechanism, and the task is to increase this initial cluster by searching the expression data collection for genes that are likely to share that mechanism. In the first step, the initial ‘query set’ of genes is analysed with the aim to identify those experiments supporting an existing relationship between the query genes. Experiments, in which the query genes are consistently regulated, are being assigned a high relative weight, while experiments without such regularities get a low or even zero weight. In the second phase, the derived weights are applied to a searching step, with the aim to identify additional genes that are expressed similarly to the query genes - at least in the ‘critical’ experiments. If those genes pass a given significance criterion, they are included into the query group. The third phase is analogous to the ‘iterative profile refinement’ step in sequence comparison: The augmentation of the initial query family by newly found members generally leads to an improved knowledge on the relative importance of the experiments.

Thus, a re-evaluation of the experiment weights and a repeated searching step can be expected to find additional similarities passing the significance criterion.
This whole process can be repeated until no new significant members are found. The final result consists of a list of genes that are assumed to share a common regulation. A second result is a list of weights showing the relative contribution of the experiments to the reported gene cluster. In a typical biclustering application, both genes and experiments are subjected to a binary member/non-member classification. By contrast, our method yields a binary classification for one dimension (typically genes) while cluster-membership of the second dimension (typically experiments) is expressed as a quantitative measure. Thus, we propose the term ‘fuzzy biclustering’ for this type of application.