Shortcomings of artificial data analysis

Although artificial data are helpful to gain understanding in properties of algorithms, creating realistic scenarios and choosing good performance measures remains extremely tricky. As mentioned before, we decided to use artificial data from a recent comparative study (Prelic *et al.*, 2006) in order to avoid any bias in creating our own data sets. However, one has to keep in mind this data suffer some artifacts:

- The modules are either noisy or overlapping. In reality, transcriptional modules are both noisy and overlapping.
- It is unclear whether the interaction in the overlap is best simulated by additive, averaging, multiplicative, or Boolean logical models.
- Whenever modules overlap, the intersections are valid modules too, and (if statistically significant) their unions could be considered coherent transcription units as well. Indeed, depending on the user parameter (that specifies the resolution) our method finds the intersection, the 'module' or the union of two or more 'modules', explaining the somewhat lower bicluster relevance scores in those scenarios (see supplementary material). Query-driven biclustering allows obtaining a detailed local multi-resolution view rather than a global single-resolution view. Whenever interesting changes in bicluster composition occur, critical resolution values can be identified.
- In reality, the number of genes in an interesting module is much smaller than the number of genes in the background. The artificial data we use reflects this assumption poorly, with modules that contain 10% of the total gene content. However, for computational reasons it is often infeasible to perform extensive simulation studies on large data sets.
- The artificial data sets contain modules of equal size and (nearly) equal strength. In real data sets, we expect a few strong modules to dominate (for example ribosome biogenesis in yeast expression arrays). In the latter situation, the advantage of query based approaches will be more evident.

REFERENCES

Prelic,A. et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics.*, **22**, 1122-1129.