Prediction of Prokaryotic Transcription Units from Microarray Data Revisited

Ulrich Bodenhofer, Wilhelm Lichtberger, Frank Klawonn

In prokaryotic genomes, a transcription unit is a set of one or more co-located genes that are transcribed into one single mRNA molecule that codes for more than one protein. Transcription units are mostly disjoint, but may also overlap. Minimal non-overlapping unions of transcription units are commonly called operons.

The identification of transcription units is the first step towards the construction of regulatory networks for such species. Since co-transcribed genes, in principle, result in correlated expression levels, microarray expression measurements are promising data sources for inferring transcription units. Previous works are based on correlation-based non-overlapping clustering of genes, and have only been partly successful. The problems arise from two main sources: (1) if a transcription unit is never expressed in the given data set, no correlations occur at all; (2) overlapping transcription units cannot be detected sufficiently well with classical correlation measures.

We propose a new framework for inferring hypotheses about transcription units from microarray data. First, we use FARMS, a summarization method based on a probabilistic factor analysis model, which allows for a better distinction between non-correlation and noise. Second, we use support vector machines for determining possible interrelations between genes on the basis of their intergenic distance and a specifically designed correlation measure that accounts for the two challenges highlighted above. This approach has been tested on E.coli data, for which it is able to identify transcription units with high accuracy. The method, therefore, cannot only help to complete the knowledge about E.coli. For less investigated prokaryotes, it allows to infer hypotheses about transcription units in a quick and reliable manner.