

EXPANDER – a 'one stop shop' for microarray data analysis

Ron Shamir¹, Adi Maron-Katz¹, Amos Tanay¹, Chaim Linhart¹, Israel Steinfeld¹,
Roded Sharan¹, Yosef Shiloh², and Ran Elkon²

Keywords: gene expression analysis, clustering, biclustering, promoter elements, GO analysis

1 Introduction.

A major challenge in the analysis of microarray gene expression data is to mine meaningful biological knowledge out of the huge volume of raw data. Here we present *EXPANDER 2.0* (EXpression ANalyzer and DisplayER) - an integrative package for the analysis of gene expression data, designed as a 'one-stop shop' tool that implements various data analysis algorithms ranging from the initial steps of normalization and filtering, through clustering and biclustering, to high-level functional enrichment analysis that points to biological processes that are active in the examined conditions, and promoter cis-regulatory elements analysis revealing transcription factors that control the observed transcriptional response. EXPANDER is available with pre-compiled functional Gene Ontology (GO) and promoter sequence-derived data files for yeast, worm, fly, rat, mouse and human, supporting high-level analysis applied to data obtained from these six organisms. The integrated analysis capabilities provided by EXPANDER and its built-in support of multiple organisms make it unique among the many tools available for microarray data analysis. The package is freely available for academic users at <http://www.cs.tau.ac.il/~rshamir/expand>

2 Main features.

Normalization and Filtering - *EXPANDER* supports non-linear regression and quantiles equalization as well as the simpler row/column standardization procedures. Genes can be filtered based on fold-change factors and minimal variation criteria. The system can be used to analyze both cDNA microarrays and Affymetrix datasets.

Cluster analysis - *EXPANDER* implements several of the most popular clustering algorithms - SOM, K-means, and hierarchical clustering, as well as CLICK [1], a graph theoretic based algorithm developed in our lab.

Bicluster analysis – The SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) biclustering algorithm [2] is the preferable partition-analysis approach when large heterogeneous datasets that encompass dozens of conditions are involved. The SAMBA version 2.0 that is integrated in *EXPANDER* has an improved performance and can handle datasets with thousands of conditions profiled over entire genomes.

Functional enrichment analysis - *EXPANDER* uses the TANGO (Tool for ANalysis of GO enrichment) algorithm for performing functional enrichment tests that fully account for extensive multiple testing effects [3]. TANGO repeatedly shuffles genes to compute an empirical distribution of maximum p-values for functional enrichment obtained across a random sample of clusters that maintain the exact size characteristics of the analyzed clusters.

Six readily available model organisms - *EXPANDER* is provided with pre-compiled functional annotation files for six organisms: yeast (*S. cerevisiae*), worm (*C. elegans*), fly (*D. melanogaster*), rat (*R. norvegicus*), mouse (*M. musculus*) and human.

Cis-regulatory element analysis – *EXPANDER* integrates clustering and promoter analysis of gene clusters and biclusters by incorporating the PRIMA (**PR**omoter **I**ntegration in **M**icroarray **A**nalysis) tool [4]. Given target sets and a background set of promoters, PRIMA performs statistical tests aimed at identifying transcription factors whose binding site signatures are significantly more prevalent in any of the target sets than in the background set. Expander uses preprocessed fingerprint files to speed up PRIMA performance and allows the user to interactively explore possible cis-regulatory motifs in clusters.

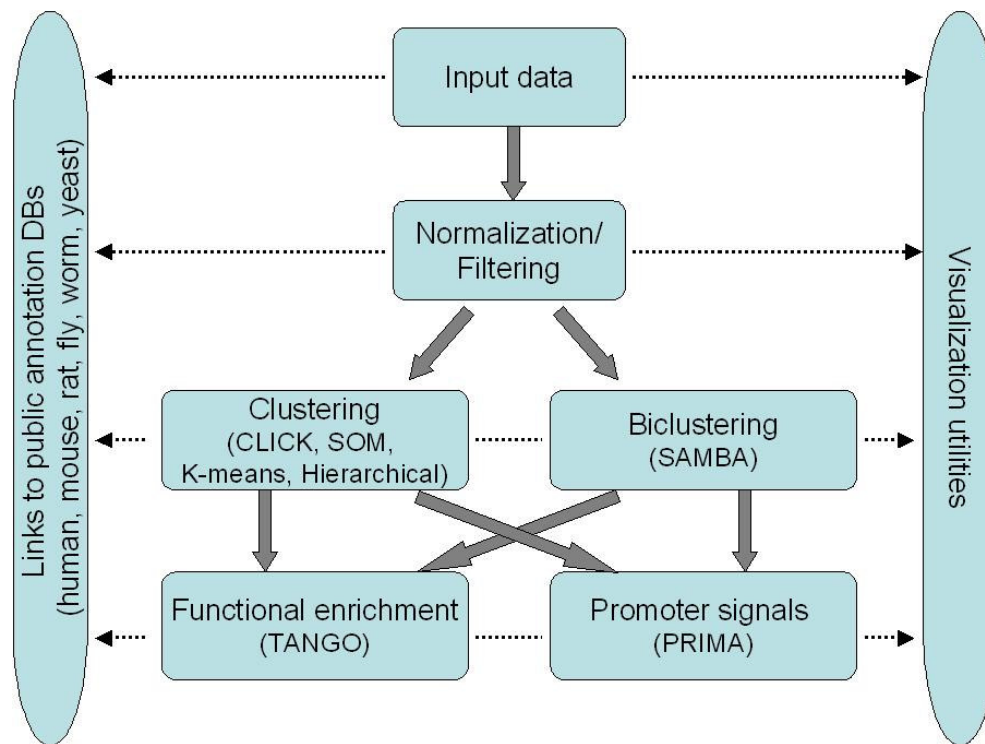


Figure 1: The EXPANDER system high-level design.

References

- [4] Elkon, R., C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh. 2003. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res* **13**: 773-780
- [1] Sharan, R. and R. Shamir. 2000. CLICK: a clustering algorithm with applications to gene expression analysis. *Proc Int Conf Intell Syst Mol Biol* **8**: 307-316.
- [2] Tanay, A., R. Sharan, M. Kupiec, and R. Shamir. 2004. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A* **101**: 2981-2986.
- [3] Tanay, A., Blienberg, O., Kritz, D., Warsaw, R. TANGO: functional enrichment with rapid correction for multiple testing. Manuscript in preparation.