Clustering biological data by unraveling hidden transitive substructures

Tobias Wittkop[*]

Buck Institute for Age Research
Novato, CA 94945, USA
Email: twittkop@buckinstitute.org

Jan Baumbach

Max-Planck Institute for Informatics
Saarbrücken, Germany
Email: jbaumbac@mpi-inf.mpg.de

Clustering is a computational technique for the assignment of objects into groups of similar elements. Generally, it is widely used for business data interpretation, natural language analyses, and image processing. Typical bioinformatics applications are the detection of homologous proteins and the identification of co-expressed genes.

Here, we introduce Transitivity Clustering and its accompanying software framework TransClust, a homogeneous data partitioning method based on Weighted Transitive Graph Projection. It aims for unraveling hidden transitive substructures in a given similarity graph deduced from a pairwise similarity measure. Transitivity Clustering is an efficient technique that is capable of processing hundreds of thousands of data points while still being robust against outliers and noise. A single, intuitive density parameter determines the number and the size of the clusters; with provable attributes. In addition, the model has been extended in order to allow for the integration of existing knowledge and the computation of an hierarchical or overlapping clustering.

The software implementation of Transitivity Clustering, TransClust, is available online at http://transclust.cebitec.uni-bielefeld.de as web application, as standalone tool, and as plugin for the network analysis tool Cytoscape. It provides results of similar or superior accuracy to those of alternative approaches. It is unique in that it features an easy-to-use clustering environment that contributes to all the important steps in a cluster analysis: (1) the choice and evaluation of a meaningful similarity function, (2) the detection of an appropriate density parameter, (3) the efficient computation of a clustering, and (4) the interpretation and evaluation of the clustering results.

[*] Corresponding author.