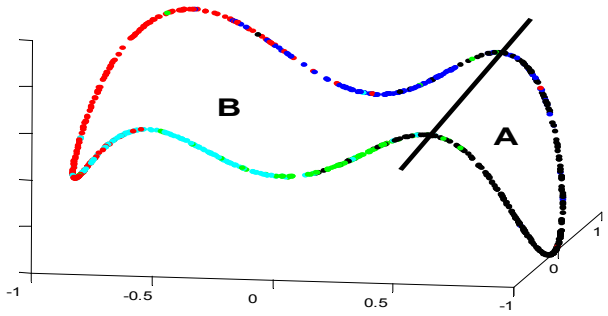
Recursive Top-Down Quantum Clustering of Biological Data

Roy Varshavsky[1,*], David Horn[2] and Michal Linial[3]

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel, [2]School of Physics and Astronomy, Tel Aviv University, Israel, [3]Deptartment of Biological Chemistry, The Hebrew University of Jerusalem, Israel

Motivation: Hierarchical tree is a natural way to present different granularity in gene expression, proteins, function annotations and more. In most of the cases where such representation is desired, the number of clusters may be large and a priori unknown, hence global clustering is insufficient (e.g., K-Means, QC). Many of the available known hierarchical clustering algorithms are bottom-up (agglomerative). The main drawback is that arbitrary considerations (e.g., dependency on the similarity and metric representations, determining the number of clusters) are applied. We present a new procedure, Top-Down Quantum Clustering (TDQC), taking advantage of the potential value assigned to each data point by QC [1]. It overcomes the tendency of QC to generate a small number of clusters and miss some internal structures.

Method: Top-down, recursive clustering



The Algorithm:

1. [Optionally] To the original dataset apply preprocessing [2]:
2. Run QC
3. Divide the data to:
   a. The cluster with the global minimum (A in Fig 1)
   b. The remaining clusters (B in Fig 1)
4. Recursively go to 1

Stop dividing when a set includes <= 2 elements

Fig 1: Potential values of the Spellman dataset (compressed to 2 normalized dimensions, see [2]). Also shown is a partitioning of the dataset into two groups. The color code represents Spellman's expert view of the 5 cell cycle phase (G1-black, S-green, S/G2- cyan, G2/M-red, M/G1-blue)

Results: TDQC was applied to gene expression data and tested on several benchmarks. Here we illustrate the results for the Spellman experiment that analyzes gene expression stages in the yeast genome [3]. According to their expert view there are 5 major partitions of the 800 genes that correspond to the phases of the cell cycle. Fig. 2 shows a graphic result of the TDQC algorithm. We were able to automatically partition the data to the main 5 groups. Interestingly, our results suggest some refinement of the expert view. A small subset of the assigned genes in S/G2 (cyan) is more likely to be associated with G2/M (red). The biological relevance of our results and the power of the method and its application to other protein sets and to gene expression data will be discussed.

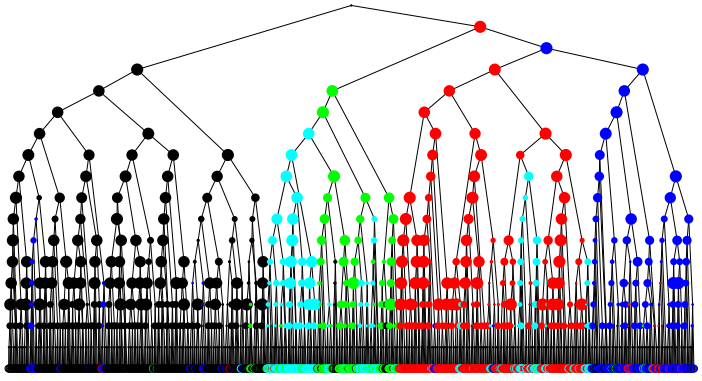


Fig 2: Hierarchical tree of the 800 Spellman genes according to TDQC. Color codes are the same as in Fig. 1. Dot sizes indicate statistical enrichment levels.

[1] Horn, D. and Gottlieb, A. (2002) Algorithm for data clustering in pattern recognition problems based on quantum mechanics, Physical Review Letters, 88.

[2] Varshavsky, R., Linial, M. and Horn, D. (2005) COMPACT: A Comparative Package for Clustering Assessment. Lecture Notes in Computer Science (3759) 159-167 Springer-Verlag.

[3] Spellman P.T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D., Futcher B. P. T.: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998, 9(12): 3273-97.