A statistical test based on the Bernstein inequality to discover multi-level structures in bio-molecular data.

Alberto Bertoni (1) and *Giorgio Valentini* (1)

DSI - Dipartimento di Scienze dell' Informazione dell' Università degli Studi di Milano.

Motivation

The unsupervised exploration and identification of structures (i.e. clusterings) underlying complex bio-molecular data is a central issue in several branches of bioinformatics, ranging from transcriptomics to proteomics and functional genomics.
Several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters in clustered bio-molecular data.
A major problem with stability-based methods is the detection of multi-level structures underlying the data (e.g. hierarchical subclasses of diseases, or hierarchical functional classes of genes), and the assessment of their statistical significance.
Recently, we proposed a chi-squared-based statistical test of hypothesis to assess the significance of the "optimal" number of clusters and to discover multiple structures simultaneously present in bio-molecular data: however, some assumptions about the distribution of the data are needed to estimate the reliability of the obtained clusterings.

Methods

In this contribution we propose a new distribution-free approach that does not assume any "a priori" distribution of the similarity measures.
In particular we consider the problem of the assessment of the reliability of a clustering procedure using a stability-based approach: multiple clusterings are obtained by introducing perturbations into the original data, and a clustering is considered reliable if it is approximately maintained across multiple perturbations. Different procedures may be applied to randomly perturb the data, ranging from bootstrapping techniques, to noise injection into the data or random projections into lower dimensional subspaces.
To assess the statistical significance and to discover multi-level structures underlying bio-molecular data, we propose a method based on Bernstein inequality, that makes no assumption about the distribution of the data, thus assuring a reliable application of the method to a large range of bioinformatics problems.

Results

Figure 1 (see http://homes.dsi.unimi.it/~valenti/papers/BITS07/Bertoni_Valentini_BITS_2007.html) summarizes the results obtained with the classical Golub's *leukemia* data set, by applying two variants of the Bernstein-based and the chi-squared-based statistical tests. K-clusterings above the dashed horizontal line are significant, that is they are significantly more reliable than those below the horizontal line (at 0.001 significance level): all the tests correctly detect 2 and 3-clustering as the most reliable, according to the biological meaning of the data (two classes: ALL and AML, with AML splittable in two other B-cell

and T-cell subclusters).
The Bernstein test, due to its more general assumptions (no particular distributions and no independence are assumed for the random variables that represent the similarity between clusterings), is more sensitive with respect to the chi-squared-based test to multiple structures simultaneously present in the data. Nevertheless it is less selective, that is subject to more false positives. Assuming independence between the random variables, we obtain a more selective Bernstein inequality-based test with intermediate characteristics between the chi-squared and the Bernstein test that assumes independence between variables (red curve in Figure 1).
Results with other DNA microarray data sets show the effectiveness of the proposed test based on the Bernstein inequality in the assessment of the reliability of multiple hierarchical structures discovered in bio-molecular data.

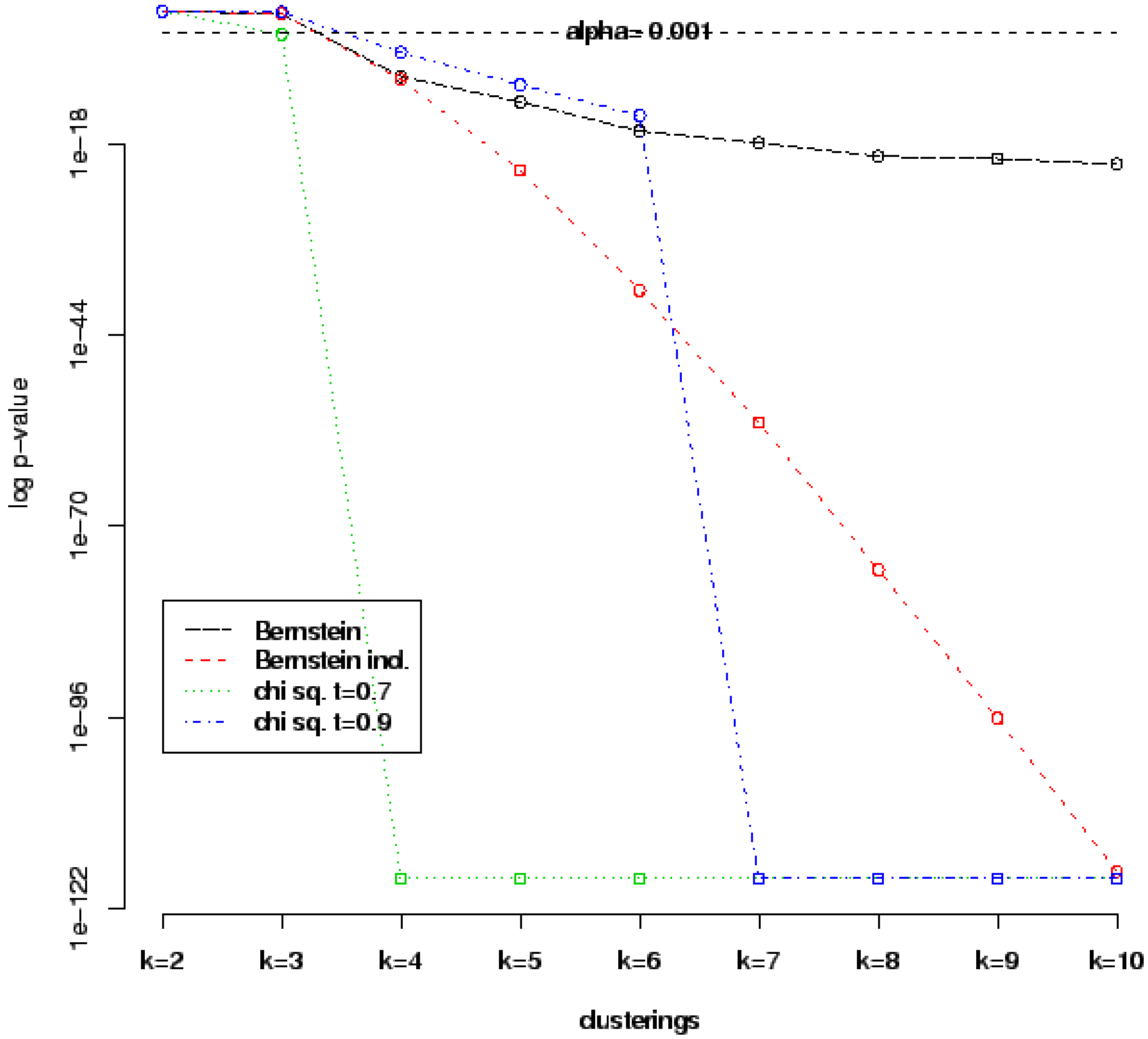

Figure 1. Plot of the p-values computed according to different statistical tests. In abscissa is represented the number of clusters sorted from the most reliable to the least reliable (with respect to the stability indices computed by random projections and using the k-means clustering algorithm). The the p-values are represented in log scale. The straight horizontal dashed line, represents the 0.001 significance level: K-clusterings above the dashed line are significant, that is their reliability significantly differ from the k-clusterings below the dashed horizontal line.