

# Discovering multi-level structures in bio-molecular data through the Bernstein inequality

Alberto Bertoni<sup>1</sup>, Giorgio Valentini<sup>\*1</sup>

<sup>1</sup>DSI, Dipartimento di Scienze dell' Informazione, Università degli Studi di Milano, Via Comelico 39, Milano, Italy

Email: Alberto Bertoni - bertoni@dsi.unimi.it; Giorgio Valentini\* - valentini@dsi.unimi.it;

\*Corresponding author

## Abstract

---

**Background:** The unsupervised discovery of structures (i.e. clusterings) underlying data is a central issue in several branches of bioinformatics. Methods based on the concept of stability have been recently proposed to assess the reliability of a clustering procedure and to estimate the "optimal" number of clusters in bio-molecular data. A major problem with stability-based methods is the detection of multi-level structures (e.g. hierarchical functional classes of genes), and the assessment of their statistical significance. In this context, a chi-square based statistical test of hypothesis has been proposed; however, to assure the correctness of this technique some assumptions about the distribution of the data are needed.

**Results:** To assess the statistical significance and to discover multi-level structures in bio-molecular data, a new method based on Bernstein's inequality is proposed. This approach makes no assumptions about the distribution of the data, thus assuring a reliable application to a large range of bioinformatics problems. Results with synthetic and DNA microarray data show the effectiveness of the proposed method.

**Conclusions:** The Bernstein test, due to its loose assumptions, is more sensitive than the chi-square test to the detection of multiple structures simultaneously present in the data. Nevertheless it is less selective, that is subject to more false positives, but adding independence assumptions, a more selective variant of the Bernstein inequality-based test is also presented. The proposed methods can be applied to discover multiple structures and to assess their significance in different types of bio-molecular data.

---

## Background

Unsupervised cluster analysis of bio-molecular data is one of the main and well-established research lines in bioinformatics [1]. Classes of co-expressed genes, classes of functionally related proteins, or subgroups of patients with malignancies differentiated at bio-molecular level can be discovered through clustering algorithms, and several other tasks related to the analysis of bio-molecular data require the development and application of unsupervised clustering techniques [2–4]. Anyway, in most bioinformatics problems, we need to assess the reliability of the discovered clusters, as well as the proper selection of the "natural" number of clusters underlying the data [5].

Recently, several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters [6, 7]: 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 [8], to noise injection into the data [9] or random projections into lower dimensional subspaces [10, 11].

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). For instance, it is possible that data exhibit a hierarchical structure, with subclusters inside other clusters, and we need to detect these multi-level structures, possibly estimating their reliability and statistical significance. In [7], it is proposed a  $\chi^2$ -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, by this approach, on one hand some assumptions about the distribution of the similarity measures are needed to estimate the reliability of the obtained clusterings, and on the other hand test results depend on the choice of user-defined parameters.

In this contribution we propose a distribution-free approach that does not assume any "a priori" distribution of the similarity measures, and that does not require any user-defined additional parameter. The proposed approach is based on the classical Bernstein inequality [12], and for its loose assumptions about the distribution of the data may in principle be applied to any unsupervised model order selection problem. More precisely the proposed stability-based method may be applied to several tasks related to the unsupervised analysis of complex bio-molecular data: (a) the assessment of the reliability of a given clustering solution; (b) the clustering model order selection, that is the discovery of the "natural" number of clusters in the data; (c) the assessment of the statistical significance of a given clustering solution; (d)

the discovery of multiple structures underlying the data, i.e. the detection of multiple reliable clustering solutions at a given significance level.

## Methods

In the following sections we summarize the characteristics of the stability-based procedures for the assessment of the reliability of clusterings, and we introduce our proposed method based on the Bernstein inequality.

### Model order selection through stability based procedures

Let be  $\mathcal{C}$  a clustering algorithm,  $\rho(D)$  a given random perturbation procedure applied to a data set  $D$  and  $sim$  a suitable similarity measure between two clusterings (e.g. the Jaccard similarity [13]). Among the random perturbations we recall random projections from a high dimensional to a low dimensional subspace [14], or bootstrap procedures to sample a random subset of data from the original data set  $D$  [8]. Fixing an integer  $k$  (the number of clusters), we define  $S_k$  ( $0 \leq S_k \leq 1$ ) as the random variable given by the similarity between two  $k$ -clusterings obtained by applying a clustering algorithm  $\mathcal{C}$  to data pairs  $D_1$  and  $D_2$  obtained by randomly and independently perturbing the original data  $D$ .

If  $S_k$  is concentrated close to 1, then the corresponding clustering is stable with respect to a given controlled perturbation and hence it is reliable. This idea, mutated by a qualitative method proposed in [15], can be formalized using the integral  $g(k)$  of the cumulative distribution  $F_k$  of  $S_k$  [7]:

$$g(k) = \int_0^1 F_k(s) ds \quad (1)$$

If  $g(k)$  is close to 0 then the values of the random variable  $S_k$  are close to 1 and hence the  $k$ -clustering is stable, while for larger values of  $g(k)$  the  $k$ -clustering is less reliable. This observation comes from the following fact:

**Fact :**  $E[S_k] = 1 - g(k)$ ,  $Var[S_k] \leq g(k)(1 - g(k))$ .

*Proof :*

Let be  $f_k(s)$  the probability density function of  $S_k$ ; then

$$E[S_k] = \int_0^1 s f_k(s) ds = \int_0^1 s F_k'(s) ds = 1 - \int_0^1 F_k(s) ds = 1 - g(k)$$

Moreover:

$$Var[S_k] = E[S_k^2] - E[S_k]^2 \leq E[S_k] - E[S_k]^2 = g(k)(1 - g(k))$$

□.

Hence,  $g(k) \simeq 0$  implies  $Var[S_k] \simeq 0$ . As a consequence,  $g(k)$  or equivalently  $E[S_k]$  can be used as a good index of the reliability of the  $k$ -clusterings (clusterings with  $k$  clusters).  $E[S_k]$  may be estimated by the empirical mean  $\xi_k$  of  $n$  replicated similarity measures between pairs of perturbed clusterings:

$$\xi_k = \sum_{j=1}^n \frac{S_{kj}}{n} \quad (2)$$

where  $S_{kj}$  represents the similarity between two  $k$ -clusterings obtained through the application of the algorithm  $\mathcal{C}$  to a pair of perturbed data.

We may perform a sorting of the  $\xi_k$ :

$$(\xi_2, \xi_3, \dots, \xi_{H+1}) \xrightarrow{sort} (\xi_{p(1)}, \xi_{p(2)}, \dots, \xi_{p(H)}) \quad (3)$$

where  $p$  is an index permutation such that  $\xi_{p(1)} \geq \xi_{p(2)} \geq \dots \geq \xi_{p(H)}$ . In this way we obtain an ordering of the clusterings, from the most to the least reliable one.

Exploiting this ordering, we proposed a  $\chi^2$ -based statistical test to detect and to estimate the statistical significance of multiple-structures discovered by clustering algorithms [7]. The main drawbacks of this approach consists in an implicit normality assumption for the distribution of the  $S_k$  (random variables that measure the similarity between two perturbed  $k$ -clusterings, see above), and in a user defined threshold parameter that determines when two  $k$ -clusterings can be considered similar and "stable". Indeed, in general we have no guarantee that the  $S_k$  random variables are normally distributed; moreover the "optimal" choice of the threshold parameter seems to be application dependent and may affect the overall test results.

In this contribution, to address these problems we propose a new statistical method that, adopting a stability-based approach, makes no assumptions about the distribution of the random variables and does not require any user-defined threshold parameter.

### Hypothesis testing based on Bernstein inequality

We briefly recall the Bernstein inequality, because this inequality is used to build-up our proposed hypothesis testing procedure.

**Bernstein inequality.** If  $Y_1, Y_2, \dots, Y_n$  are independent random variables s.t.  $0 \leq Y_i \leq 1$ , with  $\mu = E[Y_i]$ ,  $\sigma^2 = Var[Y_i]$ ,  $\bar{Y} = \sum Y_i/n$  then

$$Prob\{\bar{Y} - \mu \geq \Delta\} \leq e^{\frac{-n\Delta^2}{2\sigma^2 + 2/3\Delta}} \quad (4)$$

Using the Bernstein inequality, we would estimate if for a given  $r$ ,  $2 \leq r \leq H$ , there exists a statistically significant difference between the reliability of the best  $p(1)$  clustering and the  $p(r)$  clustering (eq. 3). In other words we may state the null hypothesis  $H_0$  and the alternative hypothesis in the following way:

$H_0$ :  $p(1)$  clustering is not more reliable than  $p(r)$  clustering, that is  $E[S_{p(1)}] \leq E[S_{p(r)}]$

$H_a$ :  $p(1)$  clustering is more reliable than  $p(r)$  clustering, that is  $E[S_{p(1)}] > E[S_{p(r)}]$

To this end, consider the following random variables:

$$P_i = S_{p(1)} - S_{p(i)} \quad \text{and} \quad X_i = \xi_{p(1)} - \xi_{p(i)} \quad (5)$$

We start considering the first and last ranked clustering  $p(1)$  and  $p(H)$ . In this case the above null hypothesis  $H_0$  becomes:  $E[S_{p(1)}] \leq E[S_{p(H)}]$ , or equivalently  $E[S_{p(1)}] - E[S_{p(H)}] = E[P_H] \leq 0$ . The distribution of the random variable  $X_H$  (eq. 5) is in general unknown; anyway note that in the Bernstein inequality no assumption is made about the distribution of the random variables  $Y_i$  (eq. 4). Hence, fixing a parameter  $\Delta \geq 0$ , considering true the null hypothesis  $E[P_H] \leq 0$ , and using Bernstein inequality, we have:

$$Prob\{X_H \geq \Delta\} \leq Prob\{X_H - E[P_H] \geq \Delta\} \leq e^{\frac{-n\Delta^2}{2\sigma^2 + 2/3\Delta}} \quad (6)$$

Considering an instance (a measured value)  $\hat{X}_H$  of the random variable  $X_H$ , if we let  $\Delta = \hat{X}_H$  we obtain the following probability of type I error:

$$P_{err}\{X_H \geq \hat{X}_H\} \leq e^{\frac{-n\hat{X}_H^2}{2\sigma_H^2 + 2/3\hat{X}_H}}$$

with  $\sigma_H^2 = \sigma_{p(1)}^2 + \sigma_{p(H)}^2$ .

If  $P_{err}\{X_H \geq \hat{X}_H\} < \alpha$ , we reject the null hypothesis: a significant difference between the two clusterings is detected at  $\alpha$  significance level and we continue by testing the  $p(H-1)$  clustering. More in general if the null hypothesis has been rejected for the  $p(H-r+1)$  clustering,  $1 \leq r \leq H-2$  then we consider the  $p(H-r)$  clustering, and using the Boole inequality, we can estimate the type I error:

$$P_{err}(H-r) = Prob\left\{ \bigvee_{H-r \leq i \leq H} X_i \geq \hat{X}_i \right\} \leq \sum_{i=H-r}^H Prob\{X_i \geq \hat{X}_i\} \leq \sum_{i=H-r}^H e^{\frac{-n\hat{X}_i^2}{2\sigma_i^2 + 2/3\hat{X}_i}} \quad (7)$$

As in the previous case, if  $P_{err}(H-r) < \alpha$  we reject the null hypothesis: a significant difference is detected between the reliability of the  $p(1)$  and  $p(H-r)$  clustering and we iteratively continue the procedure estimating  $P_{err}(H-r-1)$ .

This procedure stops if either of these cases succeeds:

I) The null hypothesis is rejected till  $r = H - 2$ , that is  $\forall r, 1 \leq r \leq H - 2, P_{err}(H - r) < \alpha$ : all the possible null hypotheses have been rejected and the only reliable clustering at  $\alpha$ -significance level is the top ranked one, that is the  $p(1)$  clustering.

II) The null hypothesis cannot be rejected for  $r \leq H - 2$ , that is,  $\exists r, 1 \leq r \leq H - 2, P_{err}(H - r) \geq \alpha$ : in this case the clusterings that are significantly less reliable than the top ranked  $p(1)$  clustering are the  $p(r + 1), p(r + 2), \dots, p(H)$  clusterings.

Note that in this second case we cannot state that there is no significant difference between the first  $r$  top-ranked clusterings, since the upper bound provided by the Bernstein inequality is not guaranteed to be tight. To answer to this question, we may apply the  $\chi^2$ -based hypothesis testing proposed in [7] to the remaining top ranked clusterings to establish which of them are significant at  $\alpha$  level, but in this case we need to assume that the similarity measures between pairs of clusterings are distributed according to a normal distribution.

If we assume that the  $X_i$  random variables (eq. 5) are (at least approximately) independent, we can obtain a variant of the previous Bernstein inequality-based approach, that we name *Bernstein ind.* for brevity. By this approach we should in principle obtain lower  $p$ -values, thus assuring lower false positive rates than the *Bernstein* test without independence assumptions.

With these independence assumptions the null hypothesis  $H_0$  and the alternative hypothesis for the *Bernstein ind.* test can be formulated as follows:

$H_0$ :  $\exists i, 2 \leq i \leq r \leq H$  such that  $E[S_{p(1)}] \leq E[S_{p(r)}]$ : it does exist at least one  $p(i)$ -clustering equally or more reliable than the first one in the group of the first  $r$  ordered clusterings.

$H_a$ :  $\forall i, 2 \leq i \leq r \leq H E[S_{p(1)}] > E[S_{p(r)}]$ : all the clusterings in the group of the first  $r$  ordered clusterings are less reliable than the first one.

If we assume that the null hypothesis is true, using the independence among the  $X_i$  random variables, we may obtain the type I error:

$$P_{err}(r) = Prob\left\{ \bigwedge_{2 \leq i \leq r} X_i \geq \hat{X}_i \right\} = \prod_{i=2}^r Prob\{X_i \geq \hat{X}_i\} \leq \prod_{i=2}^r e^{\frac{-n\hat{X}_i^2}{2\sigma_i^2 + 2/3\hat{X}_i}} \quad (8)$$

Starting from  $r = H$ , if  $P_{err}(r) < \alpha$  we reject the null hypothesis: a significant difference is detected between the reliability of the  $p(1)$  and the other first  $r$ -clustering and we iteratively continue the procedure estimating  $P_{err}(r - 1)$ . As in the *Bernstein* test, the procedure is iterated until we remain with a single clustering (and this will be the only significant one), or until  $P_{err}(r) \geq \alpha$  and in this case we cannot reject

the null hypothesis and the first  $r$  clusterings can be considered equally reliable. Note that, strictly speaking, in this case we can only say that at least one of the first  $r$  clusterings is equally or more reliable than the first one.

## Results and Discussion

In this section we apply the *Bernstein* test to synthetic and DNA microarray data analysis, and compare it to the previously proposed  $\chi^2$ -based test [7]. For the experiments we used the *mosclust* R package [16], and all the data used in the experiments are available from the authors.

### Analysis of hierarchical structures in synthetic data

In order to show the effectiveness of the proposed approach we consider synthetic data with a priori known multi-level hierarchical structure. To this end we generated a two-dimensional synthetic data set with a three-level hierarchical structure (Fig. 1) using the *clusterv* R package [17]: at a first level three large clusters are present in the data (black ovals); at a second level we have a 6-clustering (red ovals) and finally a third-level 12-clustering may be detected (blue ovals).

We performed analysis of the stability of the clusters (see Section *Model order selection through stability based procedures*), by applying random subsample techniques to perturb the data (100 subsample pairs each composed by 80 % of the data randomly drawn without replacement) and Ward’s hierarchical clustering algorithm [18] with dendrogram cuts from  $k = 2$  to  $k = 15$  clusters. Then we computed the similarity indices for each  $k$  from 2 to 15: their empirical cumulative distribution is shown in Fig. 2. From Fig. 2 we may observe that 3 and 6-clusterings similarities are closely concentrated near 1, thus showing that these clusterings are clearly detectable by the hierarchical clustering algorithm. Indeed both  $\chi^2$ -based and *Bernstein*-based test of hypothesis select these clusterings at  $10^{-5}$  significance level. Nevertheless, the *Bernstein* test selects also the 7-clustering (false positive) and the 12-clustering (true positive) (Table 1). As a second experiment we considered a 1000-dimensional synthetic multivariate gaussian distributed data set. These data are characterized by a two-level hierarchical structure: at a first level we have two main clusters with inside each one three other clusters, thus resulting in a second level 6-clustering. Each of the six second-level clusters is distributed according to a hyperspherical gaussian distribution and each cluster contains only 20 examples, thus resulting in a sparse data set (low number of examples in a high dimensional space). We applied the Prediction Around Medoids clustering algorithm [19], and we perturbed the data through *Bernoulli* random projections [7], from a 1000 to a 479-dimensional subspace,

considering the reliability of clusterings composed from 2 to 10 clusters. In this case both the  $\chi^2$ -based and the *Bernstein* based iterative procedures correctly detect 2 and 6-clusterings at  $10^{-4}$  significance level. With these high dimensional data the Bernstein test is not subject to false positives, but also the *chi*<sup>2</sup> test correctly detects all the structures underlying the data.

### Discovery of multi-level structures in DNA microarray data

As an example of the application of the Bernstein test to the discovery of multiple structures in bio-molecular data, we consider two classical DNA microarray data sets: *Leukemia* [20] and *Lymphoma* [21]. The *Leukemia* data set is composed by a group of 25 acute myeloid leukemia (AML) samples and another group of 47 acute lymphoblastic leukemia (ALL) samples, that can be subdivided into 38 B-Cell and 9 T-Cell subgroups, resulting in a two-level hierarchical structure.

We applied both resampling and random projections to lower dimensional subspaces to perturb the original data using the *R* package *mosclust* [16] that implements the Bernstein-based test and the stability measures described in Sect. *Model order selection through stability based procedures*.

Fig. 3 shows the empirical cumulative distributions of the similarity values and Table 2 the p-values of the clusterings sorted according to their  $\xi$  values (eq. 2), using *Bernoulli* random projections [7]. Our proposed procedure detects the 2 – *clustering* as the most reliable at 0.01 significance level, according to the fact that two biologically meaningful groups (ALL, acute lymphoblastic leukemia and AML, acute myeloid leukemia) are present in the data. Choosing a significance level  $\alpha = 10^{-5}$  we cannot reject the null hypothesis that a 2-clustering is less or equally reliable than a 3-clustering: in this case 2 structures (2 and 3-clusterings) are detected as simultaneously present in the data, reflecting the biological fact that ALL can be subdivided into two subclasses (B-cell and T-cell ALL).

The results obtained with two variants of the  $\chi^2$  [7] and *Bernstein* based statistical tests are compared in Fig. 4 (k-means algorithm) and Fig. 5 (PAM, Prediction Around Medoids algorithm [19]) : log p-values are represented in ordinate, while in abscissa the number of clusters are sorted according to the empirical mean of the corresponding pairwise similarities (eq. 2). In both figures a straight horizontal dashed line represents a significance level  $\alpha = 0.001$ : k-clusterings above the dashed line are significant, that is their reliability significantly differ from the k-clusterings below the dashed horizontal line. Note that the k-means (Fig. 4) and PAM (Fig. 5) clustering algorithms provide a slightly different ranking of the similarity indices, but in most cases 2 and 3 clusterings are considered significantly more reliable than the others, according to the biological characteristics of the data. 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 less selective (in the sense that it may consider reliable a larger number of  $k$ -clusterings) than the  $\chi^2$ -based test that make assumptions about the distribution of the random variables. This is confirmed by the fact that *Bernstein* p-values decrease more slowly with respect to the  $\chi^2$  test (Fig. 4 and 5), thus resulting in a better sensitivity to multiple structures present in the data. The main drawback of this behaviour is the larger probability of false positives. Note that the *Bernstein ind.* test shows an intermediate trend between the *Bernstein* and  $\chi^2$  test (red lines in Fig. 4 and 5): the assumption of independence between the random variables yields a more selective Bernstein inequality-based test less subject to false positives, but potentially less sensitive to multiple structures underlying the data.

The *Lymphoma* gene expression data set [21] comprises three different lymphoid malignancies: Diffuse Large B-Cell Lymphoma (DLBCL), Follicular Lymphoma (FL) and Chronic Lymphocytic Leukemia (CLL). The data provides expression levels for 4026 genes [22], with 62 samples subdivided in 42 DLBCL, 11 CLL and 9 FL. We performed pre-processing of the data according to [21], replacing missing values with 0 and then normalizing the data to zero mean and unit variance across genes. We considered both resampling techniques and random projections to perturb the data. In particular, data have been resampled by randomly drawing 80% of the available data without replacement, and data have been projected using *Bernoulli* projections with  $\epsilon = 0.2$  corresponding to 413-dimensional subspaces.

Fig. 6 and 7 show the empirical cumulative distribution of the similarity measures for different numbers of clusters, using the hierarchical Ward’s clustering algorithm and respectively *Bernoulli* random projections (Fig. 6) and subsampling perturbation techniques (Fig. 7). Considering random projections both the *Bernstein* and  $\chi^2$ -based tests correctly select 2 and 3-clusterings at 0.001 significance level. On the contrary, using subsampling techniques only the *Bernstein* inequality based test select as significant both 2 and 3-clusterings, while the  $\chi^2$  tests select only the 2-clustering (Table 3). These results confirm that the *Bernstein* test is more sensitive to multiple structures underlying the data.

Considering the *Leukemia* and *Lymphoma* data sets, the proposed *Bernstein* test achieves results competitive with state-of-the-art stability methods proposed in the literature. Indeed the *Model Explorer* algorithm, based on subsampling techniques, correctly detect only the 2-clustering structure both in *Leukemia* and *Lymphoma* [15]. Another subsampling-based method (*Figure of Merit* [23]) detects 2, 8 and 19-clusterings in *Leukemia* and 2 and 9-clusterings in *Lymphoma*. Stability methods that apply supervised algorithms to assess the quality of the discovered clusterings correctly detect only a 3-clustering in

*Leukemia* and a 2-clustering in *Lymphoma* [6,24]. Our previously proposed  $\chi^2$ -based test correctly detects both 2 and 3-clusterings in both data sets, if random projections are used as perturbation method, but it fails to detect the 3-clustering in *Lymphoma* when subsampling techniques are applied. On the contrary, the *Bernstein* test discovers both the two-level structures in *Leukemia* and *Lymphoma*, independently of the applied perturbation method.

The experimental results with both synthetic and gene expression data support the hypothesis that the *Bernstein* test is more sensitive to multiple structures underlying the data. Indeed in the first experiment with synthetic data it correctly predicts also the third level of structure, that is the 12-clustering; on the other hand it is subject to false positives, as shown by the wrong discovery of a 7-clustering (Table 1).

These results are confirmed by the fact that *Bernstein* p-values decrease more slowly with respect to the  $\chi^2$  test (Fig. 4 and 5): in this way for a given significance level it is likely that the *Bernstein* test selects larger sets of structures underlying the data. The risk of an increased rate of false positives may be balanced by the assumption of independence between the random variables, yielding to the proposed *Bernstein ind.* test (eq. 8), less subject to false positives, but potentially less sensitive to multiple structures underlying the data.

In real applications to complex bio-molecular data, we suggest to apply both *Bernstein*-based and  $\chi^2$ -based procedures: structures discovered by both tests are likely to be significant, and *Bernstein*-based tests can discover potential structures not detectable with the more selective  $\chi^2$ -based test. Moreover the computational burden due to the application of the  $\chi^2$  and *Bernstein*-based iterative procedures is irrelevant with respect to the execution of clustering algorithms.

## Conclusions

We proposed a test of hypothesis based on Bernstein inequality to estimate if there is a significant difference between the reliability of different clusterings performed on the same data. Our proposed method can be applied to discover multiple or hierarchical structures, using different clustering algorithms and different perturbation methods. Even if in our experiments we applied the *Bernstein* test to the analysis of gene expression data, this approach may be in principle applied to discover multiple structures in any type of complex bio-molecular data. Indeed no user-defined parameters are required, and very loose assumptions are made about the distribution of the data and the distribution of the similarity values used to estimate the stability of the discovered clusterings, thus assuring a reliable application of the method to a large range of bioinformatics problems.

Our experiments with synthetic and gene expression data show that *Bernstein*-based tests are more sensitive than  $\chi^2$ -based tests to multiple structures embedded in the data: in this way not self-evident structures may be detected too, as well as subtle relationships between the data. A drawback of the *Bernstein* test is its larger expected rate of false positives, but assuming independence between the empirical means of the similarity values a new test (*Bernstein ind.*), less subject to false positives, has been proposed.

Developments of this work could consist in the adaptation and application of the proposed methods to large scale bioinformatics problems, to discover multiple structures underlying the data when a very large number of clusters is potentially involved.

## Authors contributions

The authors equally contributed to this paper.

## Competing interests

The authors declare that they have no competing interests.

## Acknowledgments

We would like to thank the anonymous reviewers for their comments and suggestions. This work has been developed in the context of *CIMAINA* Center of Excellence, and it has been funded by the Italian COFIN project *Linguaggi formali ed automi: metodi, modelli ed applicazioni*.

## References

1. Dopazo J: **Functional Interpretation of Microarray Experiments**. *OMICS* 2006, **3**(10).
2. Gasch P, Eisen M: **Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering**. *Genome Biology* 2002, **3**(11).
3. Dyrskjot L, Thykjaer T, Krühøffer M, Jensen J, Marcussen N, Hamilton-Dutoit S, Wolf H, Ørntoft T: **Identifying distinct classes of bladder carcinoma using microarrays**. *Nature Genetics* 2003, **33**(jan.):90–96.
4. Kaplan N, Friedlich M, Fromer M, Linial M: **A functional hierarchical organization of the protein sequence space**. *BMC Bioinformatics* 2004, **5**.
5. Handl J, Knowles J, Kell D: **Computational cluster validation in post-genomic data analysis**. *Bioinformatics* 2005, **21**(15):3201–3215.
6. Lange T, Roth V, Braun M, Buhmann J: **Stability-based Validation of Clustering Solutions**. *Neural Computation* 2004, **16**:1299–1323.
7. Bertoni A, Valentini G: **Model order selection for bio-molecular data clustering**. *BMC Bioinformatics* 2007, **8**(Suppl.3).

8. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data.** *Machine Learning* 2003, **52**:91–118.
9. McShane L, Radmacher D, Freidlin B, Yu R, Li M, Simon R: **Method for assessing reproducibility of clustering patterns observed in analyses of microarray data.** *Bioinformatics* 2002, **18**(11):1462–1469.
10. Smolkin M, Gosh D: **Cluster stability scores for microarray data in cancer studies.** *BMC Bioinformatics* 2003, **36**(4).
11. Bertoni A, Valentini G: **Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses.** *Artificial Intelligence in Medicine* 2006, **37**(2):85–109.
12. Hoeffding W: **Probability inequalities for sums of independent random variables.** *J. Amer. Statist. Assoc.* 1963, **58**:13–30.
13. Jain A, Murty M, Flynn P: **Data Clustering: a Review.** *ACM Computing Surveys* 1999, **31**(3):264–323.
14. Achlioptas D: **Database-friendly random projections.** In *Proc. ACM Symp. on the Principles of Database Systems*, Contemporary Mathematics. Edited by Buneman P, New York, NY, USA: ACM Press 2001:274–281.
15. Ben-Hur A, Elisseeff A, Guyon I: **A stability based method for discovering structure in clustered data.** In *Pacific Symposium on Biocomputing, Volume 7*. Edited by Altman R, Dunker A, Hunter L, Klein T, Lauderdale K, Lihue, Hawaii, USA: World Scientific 2002:6–17.
16. Valentini G: **Mosclust: a software library for discovering significant structures in bio-molecular data.** *Bioinformatics* 2007, **23**(3):387–389.
17. Valentini G: **Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data.** *Bioinformatics* 2006, **22**(3):369–370.
18. Ward J: **Hierarchical grouping to optimize an objective function.** *J. Am. Stat. Assoc.* 1963, **58**:236–244.
19. Kaufman L, Rousseeuw P: *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley 1990.
20. Golub T, et al.: **Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.** *Science* 1999, **286**:531–537.
21. Alizadeh A, Eisen M, Davis R, Ma C, Lossos I, Rosenwald A, Boldrick J, Sabet H, Tran T, Yu X, Powell J, Yang L, Marti G, Moore T, Hudson J, Lu L, Lewis D, Tibshirani R, Sherlock G, Chan W, Greiner T, Weisenburger D, Armitage J, Warnke R, Levy R, Wilson W, Grever M, Byrd J, Botstein D, Brown P, Staudt L: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403**:503–511.
22. Alizadeh A, et al.: **The Lymphochip: a specialized cDNA microarray for genomic-scale analysis of gene expression in normal and malignant lymphocytes.** In *Cold Spring Harbor Symp. Quant. Biol.* 2001.
23. Levine E, Domany E: **Resampling method for unsupervised estimation of cluster validity.** *Neural Computation* 2001, **13**(11):2573–2593.
24. Dudoit S, Fridlyand J: **A prediction-based resampling method for estimating the number of clusters in a dataset.** *Genome Biology* 2002, **3**(7):1–21.

## Captions of the figures

Figure 1. Synthetic data set: a three-level hierarchical structure with 3, 6 and 12 clusters.

Figure 2. Synthetic data set: empirical cumulative distribution functions of the similarity measures for different number of clusters (Hierarchical clustering).

Figure 3. Leukemia data set. Empirical cumulative distributions of the similarity measures for different numbers of clusters  $k$ .

Figure 4. K-means clustering: log p-value computed for  $\chi^2$ -based and *Bernstein*-based statistical tests.

Ordinate: log p-value; abscissa: number of clusters sorted according the computed similarity means.

Figure 5. PAM clustering: log p-value computed for  $\chi^2$ -based and *Bernstein*-based statistical tests.

Ordinate: log p-value; abscissa: number of clusters sorted according the computed similarity means.

Figure 6. Lymphoma data set. Empirical cumulative distribution functions of the similarity measures for different number of clusters  $k$ . Perturbation technique: random projections

Figure 7. Lymphoma data set. Empirical cumulative distribution functions of the similarity measures for different number of clusters  $k$ . Perturbation technique: resampling

## Tables

**Table 1 - Synthetic data: comparison of the  $\chi^2$  and Bernstein inequality-based tests.**

*Bernstein ind.* stands for the *Bernstein* test with assumption of independence between the random variables representing the empirical mean of the similarity measures.

Test	Structures discovered ( $10^{-5}$ significance level)
$\chi^2$	3-clustering 6-clustering
<i>Bernstein ind.</i>	3-clustering 6-clustering 7-clustering
<i>Bernstein</i>	3-clustering 6-clustering 7-clustering 12-clustering

**Table 2 - Leukemia data set: empirical means ( $\xi$ ) and p-values computed according to the Bernstein inequality.**

Num.clusters	p-values	$\xi$
2	—	0.8664
3	1.0561e-04	0.7521
4	1.2165e-08	0.6850
5	1.0554e-12	0.6196
6	3.9321e-14	0.5922
7	1.7630e-14	0.5878
8	2.3732e-15	0.5822
9	2.7570e-16	0.5690
10	1.6297e-17	0.5491

**Table 3 - Lymphoma data: comparison of the  $\chi^2$  and Bernstein inequality-based tests.**

$t$  represents the threshold level for the  $\chi^2$ -based test.

<b>Test</b>	<b>Structures discovered</b> (0.001 significance level)
$\chi^2$ (t=0.9)	2-clustering
$\chi^2$ (t=0.7)	2-clustering
<i>Bernstein ind.</i>	2-clustering 3-clustering
<i>Bernstein</i>	2-clustering 3-clustering