Unsupervised stability-based ensembles to discover reliable structures in complex bio-molecular data

Alberto Bertoni and Giorgio Valentini

DSI, Dipartimento di Scienze dell' Informazione,
Università degli Studi di Milano,
Via Comelico 39, 20135 Milano, Italia.
{bertoni,valentini}@dsi.unimi.it

Abstract. The assessment of the reliability of clusters discovered in bio-molecular data is a central issue in several bioinformatics problems. Several methods based on the concept of stability have been proposed to estimate the reliability of each individual cluster as well as the "optimal" number of clusters. In this conceptual framework a clustering ensemble is obtained through bootstrapping techniques, noise injection into the data or random projections into lower dimensional subspaces. A measure of the reliability of a given clustering is obtained through specific stability/reliability scores based on the similarity of the clusterings composing the ensemble. Classical stability-based methods do not provide an assessment of the statistical significance of the clustering solutions and are not able to directly detect multiple structures (e.g. hierarchical structures) simultaneously present in the data. Statistical approaches based on the chi-square distribution and on the Bernstein inequality, show that stability-based methods can be successfully applied to the statistical assessment of the reliability of clusters, and to discover multiple structures underlying complex bio-molecular data. In this paper we provide an overview of stability based methods, focusing on stability indices and statistical tests that we recently proposed in the context of the analysis of gene expression data.

1 Introduction

Clustering of complex of bio-molecular data represents one of the main problems 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, 3, 4]. From a general standpoint the discovered clusters depend on the clustering algorithm, the initial condition, the parameters of the algorithm, the distance or correlation measure applied to the data and other clustering and data-dependent factors [5].

Moreover the bioinformatics domain raises specific and challenging problems that characterize clustering applications in bio-molecular biology and medicine. In particular the integration of multiple data sources [6], the very high dimensionality [7, 8], and the visualization of the data [9, 10], as well as interactive data analysis in clustering genomic data [11, 12] represent relevant topics in the unsupervised analysis of bio-molecular data.

Another relevant problem is the assessment of the reliability of the discovered clusters, as well as the proper selection of the "natural" number of clusters underlying bio-molecular data [13, 14]. Indeed in many cases we have no sufficient biological knowledge to "a priori" evaluate both the number of clusters (e.g. the number of biologically distinct tumor classes), as well as the validity of the discovered clusters (e.g. the reliability of new discovered tumor classes). Note that this is an intrinsically "ill-posed" problem, since in unsupervised learning we lack an external objective criterion, that is we have not an equivalent of a priori known class label as in supervised learning, and hence the evaluation of the validity/reliability of the discovered classes becomes elusive and difficult.

Most of the works focused on the estimate of the number of clusters in gene expression data [15, 16, 17, 18, 19], while the problem of stability of each individual cluster has been less investigated. Nevertheless, the stability and reliability of the obtained clusters is crucial to assess the confidence and the significance of a bio-medical discovery [20, 21].

Considerings the complexity and the characteristics of the data used in bioinformatics applications (e.g. the low cardinality and very high dimensionality of DNA microarray data), classical parametric methods in many cases may fail to discover structures in the data. This is the main reason why non parametric methods, based on the concept of the stability, have been recently introduced in the context of significant bioinformatics problems.

In particular, several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters in complex bio-molecular data [22, 23, 24, 17, 25, 26]. In this conceptual framework 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. Several perturbation techniques have been proposed, ranging from bootstrap techniques [19, 16, 23], to random projections to lower dimensional subspaces [21, 27] to noise injection procedures [20].

Another major problem related to stability-based methods is to estimate the statistical significance of the structures discovered by clustering algorithms. To face this problem we proposed a χ^2-based statistical test [26] and a test based on the classical *Bernstein inequality* [28, 29]. These statistical tests may be applied to any stability method based on the distribution of similarity measures between pairs of clusterings. We experimentally showed that by this approach we may discover multiple structures simultaneously present in the data (e.g. hierarchical structures), associating a *p-value* to the clusterings selected by a given stability-based method for model order selection [30, 31].

In this paper we introduce the main concepts behind stability based methods, focusing on the work developed in [27, 26, 29]. More precisely, in the next section an overview of the main characteristics of stability-based methods is given. Then in Sect. 3 a stability index, proposed in [26] to assess the reliability of a clustering solution, is described. Sect. 4 introduces two statistical tests to assess the significance of overall clustering solutions, while Sect. 5 provides an introduction to stability indices proposed in [27] to estimate the reliability of each individual cluster inside a given clustering. Then the main drawbacks and limitations of the proposed approaches, as well as new research lines are discussed and the conclusions end the paper. In the appendix 7, we briefly describe the main characteristics of two *R* packages implementing the stability indices and statistical tests described in the previous sections.

2 An overview of stability based methods

A major requirement for clustering algorithms is the reproducibility of their solutions on other data sets drawn from the same source. In this context, several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters in clustered data [23, 24]: 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.

A general stability-based algorithmic scheme for assessing the reliability of clustering solutions may be summarized in the following way:

1. For a fixed number k of clusters, randomly perturb the data many times according to a given perturbation procedure.
2. Apply a given clustering algorithm to the perturbed data
3. Apply a given clustering similarity measure to multiple pairs of k-clusterings obtained according to steps 1 and 2.
4. Use appropriate similarity indices (stability scores) to assess the stability of a given clustering.
5. Repeat steps 1 to 4 for multiple values of k and select the most stable clustering(s) as the most reliable.

Several approaches have been proposed to implement the first step: a random "perturbation" of the data may be obtained through bootstrap samples drawn from the available data [19, 23], or random noise injection into the data [20] or random subspace [21] or random projections into lower dimensional subspaces [27].

The application of a given algorithm (step 2) represents a choice based on "a priori" knowledge or assumptions about the characteristics of the data. To estimate the similarity between clusterings (step 3), classical measures, such as the Rand Index [32], or the Jaccard or the Fowlkes and Mallows coefficients [5] or their equivalent dot-product representations [16] may be applied. More precisely,

for a given clustering algorithm $\mathcal{C}$ applied to a data set X, we may obtain the following clustering:

$$\mathcal{C}(X,k) =< A_1, A_2, \ldots, A_k >, \quad \cup_{i=1}^{k} A_i = X \tag{1}$$

For each clustering $C = \mathcal{C}(X,k)$ we may obtain a *pairwise similarity matrix* M with $n \times n$ elements, where n is the cardinality of X:

$$M_{i,j} = \begin{cases} 1, & \text{if } \exists r \in \{1,\ldots,k\},\ x_i \in A_r \text{ and } x_j \in A_r,\ i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Given two clusterings $C^{(1)}$ and $C^{(2)}$ obtained from the same data set X, we may compute the corresponding similarity matrices $M^{(1)}$ and $M^{(2)}$. Then we count the number of entries $M_{i,j}$ for which $M^{(1)}$ and $M^{(2)}$ have corresponding values equal to 1 (that is the number of entries N_{11} for which the clusterings agree about the membership of a pair of examples to the same cluster). Equivalently we may compute N_{10}, that is the number of entries for which a given pair of examples belongs to the same cluster in $C^{(1)}$, but does not belong to the same cluster in $C^{(2)}$. N_{01} and N_{00} can be computed in the same way. From this quantities we may compute the classical similarity measures between clusterings:
the *Matching* coefficient:

$$M(C^{(1)}, C^{(2)}) = \frac{N_{00} + N_{11}}{N_{00} + N_{11} + N_{10} + N_{01}} \tag{3}$$

the *Jaccard* coefficient:

$$M(C^{(1)}, C^{(2)}) = \frac{N_{11}}{N_{11} + N_{10} + N_{01}} \tag{4}$$

and the *Fowlkes and Mallows* coefficient:

$$M(C^{(1)}, C^{(2)}) = \frac{N_{11}}{\sqrt{(N_{01} + N_{11})(N_{10} + +N_{11})}} \tag{5}$$

Several stability indices for model order selection have been proposed in the literature (see, e.g. [20, 21, 23, 24]): very schematically they can be divided into indices that use statistics of the similarity measures [21, 27] or their overall empirical distribution [16, 26].

The last step, that is the selection of the most stable/reliable clustering, given a set of similarity measures and the related stability indices, has been usually approached by choosing the best scored clustering (according to the chosen stability index). A major problem in this last step is represented by the estimate of the statistical significance of the discovered solutions.

3 A stability index based on the distribution of the similarity measures

In [26] we extended the approach proposed by *Ben-Hur, Ellisseeff and Guyon* [16], by providing a quantitative estimate of a *stability score* based on the overall distribution of the similarities between pairs of clusterings.

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 Fowlkes and Mallows similarity [33]). For instance ρ may be a random projection from a high dimensional to a low dimensional subspace [34], or a bootstrap procedure to sample a random subset of data from the original data set D [16].

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 pairs D_1 and D_2 of random independently perturbed data. The intuitive idea is that if S_k is concentrated close to 1, the corresponding clustering is stable with respect to a given controlled perturbation and hence it is reliable.

As an example, consider a a 1000-dimensional synthetic multivariate gaussian data set with relatively low cardinality (60 examples), characterized by a two-level hierarchical structure, highlighted by the projection of the data into the two main principal components (Fig. 1): indeed a two-level structure, with respectively 2 and 6 clusters is self-evident in the data.

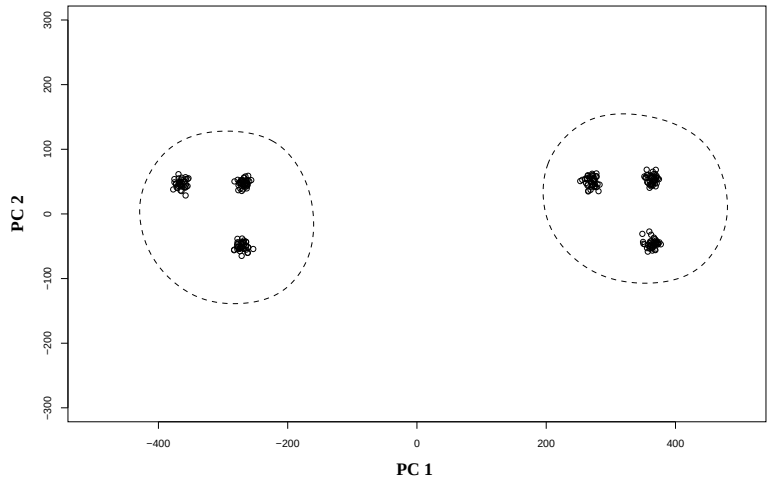

Fig. 1: A two-level hierarchical structure with 2 and 6 clusters is revealed by principal components analysis (data projected into the two components with highest variance).

We can estimate S_k, for the number of clusters k varying e.g. from 2 to 9. This can be performed by using 100 pairs of *Bernoulli* projections [26], with a distortion bounded to 20 % with respect to the original data, yielding to random projections from 1000 to 479-dimensional subspaces, and using PAM (Partitioning Around Medoids) as clustering algorithm [35]. The distribution of the similarity values is depicted in Fig. 2: the histograms of the similarity measures for $k = 2$ and $k = 6$ clusters are tightly concentrated near 1, showing that these clusterings are very stable, while for other values of k the similarity measures are spread across multiple values.

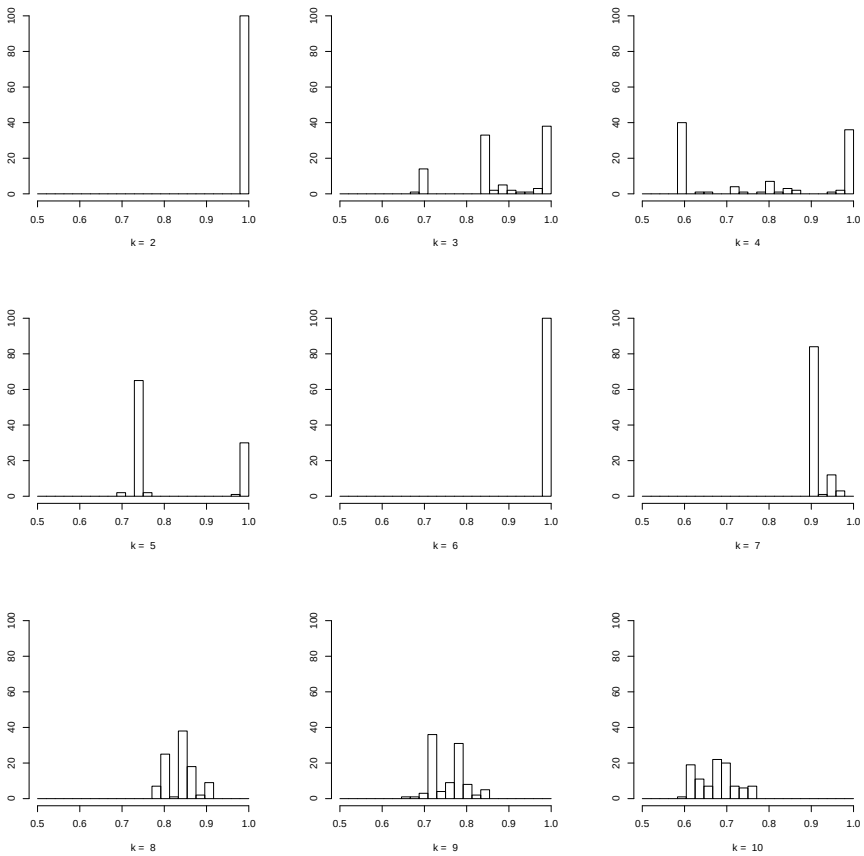

Fig. 2: Histograms of the similarity measure distributions for different numbers of clusters.

These results suggest that we could try to exploit the *cumulative distribution* of the similarities between pairs of clusterings to compute a reliable *stability score* for a given k-clustering. Indeed if the similarities are spread across multiple values, the clustering is unstable, while if they are cumulated close to 1 the clustering is stable. In the following we derive a more formal derivation of a stability score based on the cumulative distribution of the similarity measure between clusterings.

Let be $f_k(s)$ the density function of the random variable S_k, and

$$F_k(\bar{s}) = \int_{-\infty}^{\bar{s}} f_k(s)ds \tag{6}$$

its cumulative distribution function.

We define $g(k)$ as the integral of the cumulative distribution function:

$$g(k) = \int_0^1 F_k(s)ds \tag{7}$$

Intuitively $g(k)$ represents the "concentration" of the similarity values close to 1; that is, if $g(k) \simeq 0$ then the distribution of the values of S_k is concentrated near 1, or, in other words, the k-clustering is stable. On the other hand, if $g(k) \simeq 1$ then the clusterings are totally unstable, while if the distribution is close to the uniform distribution, we have $g(k) \simeq 1/2$.

We may directly estimate eq. 7 by numerical integration, or we may more easily obtain $g(k)$ from the estimate of the expectation $E[S_k]$:

$$\begin{aligned} E[S_k] &= \int_0^1 s f_k(s)ds = \int_0^1 s F_k'(s)ds \\ &= sF_k(s)\,|_0^1 - \int_0^1 F_k(s)ds = 1 - \int_0^1 F_k(s)ds \end{aligned} \tag{8}$$

Hence from eq. 8 we may easily compute $g(k)$:

$$g(k) = \int_0^1 F_k(s)ds = 1 - E[S_k] \tag{9}$$

Eq. 9 shows that we have a very stable and reliable clustering ($E[S_k]$ close to 1), if and only if $g(k)$ is close to 0.

In practice we can compute the empirical means ξ_k of the similarity values, while varying the number of clusters k from 2 to H and then we can perform a sorting of the obtained values:

$$(\xi_2, \xi_3, \ldots, \xi_H) \stackrel{sort}{\rightarrow} (\xi_{p(1)}, \xi_{p(2)}, \ldots, \xi_{p(H-1)}) \tag{10}$$

where p is the permutation index such that $\xi_{p(1)} \geq \xi_{p(2)} \geq \ldots \geq \xi_{p(H-1)}$. Roughly speaking, this ordering represents the "most reliable" $p(1)$-clustering down to the least reliable $p(H-1)$-clustering and ξ_k provides a *stability score* of the obtained k-clustering.

4 Statistical tests to assess the significance of overall clustering solutions

In [26] we proposed a χ^2 test to assess the significance of clustering solutions and to discover multiple structures underlying gene expression data. Moreover, in [28, 36], we proposed 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, using the classical Bernstein inequality [37].

4.1 A χ^2-based test to discover multiple structures in bio-molecular data

Consider a set of k-clusterings $k \in \mathcal{K}$, where $\mathcal{K}$ is a set of numbers of clusters. By estimating the expectations $E[S_k]$ or equivalently by computing eq. 7 through numerical integration, we obtain a set of values $\mathcal{G} = \{g_k | k \in \mathcal{K}\}$. We can sort G obtaining $\hat{\mathcal{G}}$ with values $\hat{g}_i$ in ascending order. For each k-clustering we consider two groups of pairwise clustering similarities values separated by a threshold t^o (a reasonable threshold could be $t^o = 0.9$.). Thus we may obtain: $P(S_k > t^o) = 1 - F_k(s = t^o)$, where $F_k(s = t^o)$ is computed according to eq. 6. If n represents the number of trials for estimating the value of S_k then $x_k = P(S_k > t^o)n$ is the number of times for which the similarity values are larger than t^o. The x_k may be interpreted as the successes from $|\mathcal{K}|$ binomial populations with parameters θ_k. If the number of trials n is sufficiently large, and setting X_k as a random variable that counts how many times $S_k > t^o$, we have that the following random variable, for sufficiently large values of n is distributed according to a normal distribution:

$$\frac{X_k - n\theta_k}{\sqrt{n\theta_k(1-\theta_k)}} \sim N(0,1) \tag{11}$$

A sum of i.i.d. squared normal variables is distributed according to a χ^2 distribution:

$$\sum_{k \in \mathcal{K}} \frac{(X_k - n\theta_k)^2}{n\theta_k(1-\theta_k)} \sim \chi^2 \tag{12}$$

Considering the null hypothesis H_0: all the θ_k are equal to θ, where the unknown θ is estimated through its pooled estimate $\hat{\theta} = \frac{\sum_{k \in K} x_k}{|\mathcal{K}| \cdot n}$, then the null hypothesis may be evaluated against the alternative hypothesis that the θ_k are not all equal using the statistic

$$Y = \sum_{k \in \mathcal{K}} \frac{(x_k - n\hat{\theta})^2}{n\hat{\theta}(1-\hat{\theta})} \sim \chi^2_{|\mathcal{K}|-1} \tag{13}$$

If $Y \geq \chi^2_{\alpha,|\mathcal{K}|-1}$ we may reject the null hypothesis at α significance level, that is we may conclude that with probability $1 - \alpha$ the considered proportions are different, and hence that at least one k-clustering significantly differ from the others. Using the above test we start considering all the k-clustering. If a significant difference is registered according to the statistical test we exclude the last clustering (according to the sorting of $\mathcal{G}$). This is repeated until no significant difference is detected (or until only 1 clustering is left out): the set of the remaining (top sorted) k-clusterings represents the set of the estimate stable number of clusters discovered (at α significance level).

It is worth noting that the above χ^2-based procedure can be also applied to automatically find the optimal number of clusters independently of the applied perturbation method.

Anyway, note that with the previous χ^2-based statistical test we implicitly assume that some probability distributions are normal. Moreover test results depend on the choice of user-defined parameters (the threshold t^o). Using the

classical Bernstein inequality [37] we may apply a partially "distribution independent" approach to assess the significance of the discovered clustering.

4.2 A *Bernstein inequality*-based test to discover multiple structures in bio-molecular data

We briefly recall the Bernstein inequality, because this inequality is used to build-up our proposed hypothesis testing procedure, without introducing any user defined parameter.

Bernstein inequality. If $Y_1, Y_2, \ldots, Y_n$ 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

$$Prob\{\bar{Y} - \mu \geq \Delta\} \leq e^{\frac{-n\Delta^2}{2\sigma^2 + 2/3\Delta}} \tag{14}$$

Consider the following random variables:

$$P_i = S_{p(1)} - S_{p(i)} \quad \text{and} \quad X_i = \xi_{p(1)} - \xi_{p(i)} \tag{15}$$

We start considering the first and last ranked clustering $p(1)$ and $p(H)$. In this case the null hypothesis becomes: $E[S_{p(1)}] \leq E[S_{p(H)}]$, that is: $E[S_{p(1)}] - E[S_{p(H)}] = E[P_H] \leq 0$. The distribution of the random variable X_H (eq. 15) 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. 14). Hence, fixing a parameter $\Delta \geq 0$, considering true the null hypothesis $E[P_H] \leq 0$, and using Bernstein inequality, we have:

$$Prob\{X_H \geq \Delta\} \leq Prob\{X_H - E[P_H] \geq \Delta\} \leq e^{\frac{-n\Delta^2}{2\sigma^2 + 2/3\Delta}} \tag{16}$$

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 α 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 by union bound we can estimate the type I error:

$$P_{err}(H-r) = Prob\{\bigvee_{H-r \leq i \leq H} X_i \geq \hat{X}_i\} \leq \sum_{i=H-r}^{H} Prob\{X_i \geq \hat{X}_i\} \leq \sum_{i=H-r}^{H} e^{\frac{-n\hat{X}_i^2}{2\sigma_i^2 + 2/3\hat{X}_i}} \tag{17}$$

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 to $r = H - 2$, that is $\forall r, 1 \leq r \leq H - 2, \quad P_{err}(H - r) < \alpha$: all the possible hypotheses have been rejected and the only reliable clustering at α-significance level is the top ranked one, that is the $p(1)$ clustering.
II) The null hypothesis cannot be rejected for $r < H - 2$, that is, $\exists r, 1 \leq r \leq H - 2, \quad P_{err}(H - r) \geq \alpha$: in this case the clusterings that are significantly less reliable than the top ranked $p(1)$ clustering are the $p(r + 1), p(r + 2), \ldots, p(H)$ 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 χ^2-based hypothesis testing proposed in [26] to the remaining top ranked clusterings to establish which of them are significant at α level, but in this case we need to assume that the similarity measures between pairs of clusterings are distributed according to a normal distribution.

For applications of the χ^2-based and the *Bernstein inequality*-based to the analysis of bio-molecular data see e.g. [26, 28, 29]. The experimental results show that Bernstein test is more sensitive to multiple structures underlying the data, but at the same time more susceptible to false positives with respect to the χ^2 test.

5 Stability indices for the assessment of the reliability of individual clusters

In this section we provide an overview of the approach proposed in [27] to assess the validity of each individual cluster, using random projections to lower dimensional subspaces as perturbation methods.

5.1 Perturbations through randomized embedding

Dimensionality reduction may be obtained by mapping points from a high to a low-dimensional space, approximately preserving some characteristics, i.e. the distances between points. In this context randomized embeddings with low distortion represent a key concept. Randomized embeddings have been successfully applied both to combinatorial optimization and data compression [38].

A *randomized embedding* between L_2 normed metric spaces with distortion $1 + \epsilon$, with $\epsilon > 0$ and failure probability P is a distribution probability over mappings $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$, such that for every pair $p, q \in \mathbb{R}^d$, the following property holds with probability $1 - P$:

$$\frac{1}{1+\epsilon} \leq \frac{||\mu(p) - \mu(q)||_2}{||p - q||_2} \leq 1 + \epsilon \tag{18}$$

The main result on randomized embedding is due to Johnson and Lindenstrauss [39], who proved the existence of a randomized embedding $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$

with distortion $1+\epsilon$ and failure probability $e^{\Omega(-d'\epsilon^2)}$, for every $0 < \epsilon < 1/2$. As a consequence, for a fixed data set $S \subset \mathbb{R}^d$, with $|S| = n$, by union bound, for all $p, q \in S$, it holds:

$$Prob\left(\frac{1}{1+\epsilon} \leq \frac{||\mu(p) - \mu(q)||_2}{||p-q||_2} \leq 1+\epsilon\right) \geq 1 - n^2 e^{\Omega(-d'\epsilon^2)} \tag{19}$$

Hence, by choosing d' such that $n^2 e^{\Omega(-d'\epsilon^2)} < 1/2$, it is proved the following: *Johnson-Lindenstrauss (JL) lemma*: Given a set S with $|S| = n$ there exists a $1+\epsilon$-distortion embedding into $\mathbb{R}^{d'}$ with $d' = c\ \log n/\epsilon^2$, where c is a suitable constant.

The embedding exhibited in [39] consists in random projections from $\mathbb{R}^d$ into $\mathbb{R}^{d'}$, represented by matrices $d' \times d$ with random orthonormal vectors. Similar results may be obtained by using simpler embeddings, represented through random $d' \times d$ matrices $P = 1/\sqrt{d'}(r_{ij})$, where r_{ij} are random variables such that:

$$E[r_{ij}] = 0, \qquad Var[r_{ij}] = 1$$

For sake of simplicity, we call random projections even this kind of embeddings. In particular in [34] matrices are proposed such that their entries are uniformly chosen in $\{-1, 1\}$, or in $\{-\sqrt{3}, 0, \sqrt{3}\}$, by choosing 0 with probability $2/3$ and $-\sqrt{3}$ or $\sqrt{3}$ with probability $1/6$. In this case the *JL lemma* holds with $c \simeq 4$.

Consider now a data set represented by a $d \times n$ matrix X whose columns represent n d-dimensional observations. Suppose that $d' = 4\ \log n/\epsilon^2 << d$; the *JL lemma* guarantees the existence of a $d' \times d$ matrix P such that the columns of the "compressed" data set $X^P = PX$ have approximately the same distance (up to a distortion $1+\epsilon$) of the corresponding columns in X. Moreover there is a randomized algorithm that, having in input X, outputs X^P in time $\mathcal{O}(dd'n)$ with high confidence.

5.2 Stability measures for individual clusters

The *JL* lemma shows that we may generate relatively low-distorted random projected data. Our aim is to exploit random projections to estimate stability of clusters, because random projections do not induce relevant distortions (as long as we provide a projection into a sufficiently high-dimensional subspace).

Given a finite set $X \subset \mathbb{R}^d$, we denote (with abuse of notation) with X the metric space $< X, f >$, where $f(x,y) = ||x-y||_2,\ x, y \in \mathbb{R}^d$. In the following of this section we consider a fixed random projection $\mu : \mathbb{R}^d \to \mathbb{R}^{d'}$ that verifies the *JL* lemma, and we propose a stability index for clustering by using a pairwise similarity matrix between the projected examples.

Let $\mathcal{C}$ be a clustering algorithm, that, having in input X, outputs a set of k clusters:

$$\mathcal{C}(X) =< A_1, A_2, \ldots, A_k >,\ A_j \subset X, 1 \leq j \leq k \tag{20}$$

Then we compute a "similarity" matrix M, with indices in X, using the following algorithm:

1. Generate t independent projections $\mu_i : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$, $1 \leq i \leq t$, such that $d' = 4\,\frac{\log|X|+\log t}{\epsilon^2}$
2. Apply $\mathcal{C}$ to the new projected data $\mu_i(X)$, obtaining a set of clusterings, for $1 \leq i \leq t$:

$$\mathcal{C}(\mu_i(X)) =< B_1^i, \ldots, B_k^i >,\ B_j^i \subset X_i, 1 \leq j \leq k \tag{21}$$

   where B_j^i is the j^{th} cluster of the i^{th} clustering.
3. Set the elements M_{xy} of the similarity matrix:

$$M_{xy} = \frac{1}{t}\sum_{j=1}^{k}\sum_{i=1}^{t}\chi_{B_j^i}(\mu_i(x)) \cdot \chi_{B_j^i}(\mu_i(y)) \tag{22}$$

   where $\chi_{B_j^i}$ is the characteristic function for the cluster B_j^i.

Since the elements M_{xy} measure the occurrences of the examples $\mu_i(x), \mu_i(y) \in \mu_i(X)$ in the same clusters B_j^i for $1 \leq i \leq t$, then M represents the "tendency" of the projections to belong to the same cluster. It is easy to see that $0 \leq M_{xy} \leq 1$, for each $x, y \in X$.

Using the similarity matrix M (eq. 22) we propose the following *stability index* s for a cluster A_i [27]:

$$s(A_i) = \frac{1}{|A_i|(|A_i|-1)}\sum_{(x,y)\in A_i \times A_i, x \neq y} M_{xy} \tag{23}$$

The index $s(A_i)$ estimates the stability of a cluster A_i in the original non projected space, by measuring how much the projections of the pairs $(x, y) \in A_i \times A_i$ occur together in the same cluster in the projected subspaces. The stability index has values between 0 and 1: values near 1 denote stable clusters, while lower values indicate less reliable clusters. The above stability index is very similar to that proposed by [23]. The main difference of our approach consists in the way the similarity matrix is computed: we applied randomized projections into lower dimensional subspaces, while [23] applied bootstrap techniques.

An overall measure of the stability of the clustering in the original space may be obtained averaging between the stability indices:

$$S(k) = \frac{1}{k}\sum_{i=1}^{k} s(A_i) \tag{24}$$

In this case also we have that $0 \leq S(k) \leq 1$, where k is the number of clusters.

Experimental applications of the stability indices (eq. 23 and 24) to the discovery of bio-molecular subclasses of malignancies are described in [40, 41, 27].

6 Drawbacks of stability-based methods and new research lines

Despite their successful application to several real-world problems, and in particular in bioinformatics, the theory underlying stability-based methods is not

well-understood and several problems remain open from a theoretical standpoint [42]. Moreover, using clustering stability in a high sample setting can be problematic. In particular it has been shown the bounding the difference between the finite sample stability and the "true stability" can exist only if one makes strong assumptions on the underlying distribution [43].

Moreover stability-based method may converge to a suboptimal solution owing to the shape of the data manifold and not to the real structure of the data [13], especially if the distribution of the data obey to a some rule of symmetry.

A problem that cannot be directly addressed by stability-based methods is the detection of "no structure" in the data. However, we may obtain an indirect evidence of "no structure" if the stability scores are always very low and comparable for a large set of numbers of clusters, or if the statistical tests consider equally reliable all or a large part of the possible clusterings.

Another problem relies on the characteristics of the perturbations that may induce bias into the stability indices used to estimate the reliability of the discovered clusters. In particular if the intensity of the perturbation is too high, significant distortions can be introduced, and the structure of the data cannot be preserved. For instance we showed that random subspace perturbations can induce significant distortions into real gene expression data. We showed also that random projections obeying the Johnson-Lindenstrauss lemma may induce bounded distortions, thus providing a theoretically-founded way to perturb the data approximately preserving their underlying structure [27, 26]. Unfortunately similar results are not available when we introduce perturbations through resampling techniques or noise injection into the data.

Apart from these theoretical problems, that need to be considered in future research work, we would like to cite at least two other problems that to our opinion are relevant in the bioinformatics context.

The first one is related to problems characterized by a very high number of possible clusters and clusterings. These problems naturally come from genomics and proteomics: consider, e.g. the unsupervised search for functional classes of genes or proteins. In this context the number of possible clusters is too high for classical stability based methods (consider e.g. Gene Ontology taxonomy that includes thousands of possible functional classes [44]), and the the iterative approach is too computationally expensive. For relatively moderate sized problems parallel computation could be a solution, but from a more general standpoint the problem is too complex and requires the development of new specific algorithms. A possible solution could be the reduction of possible candidate clusters, making some assumption about the characteristics of the data. To this end approaches based on hierarchical clustering ensembles and non parametric tests have been recently proposed [45, 46].

A related problem of paramount importance in medicine is the detection of stable clusters considering at the same time both patients and the genes involved in subclass separation. To this end a new approach based on stability indices for biclustering algorithms has been recently proposed [47].

A second problem is related to data integration. Indeed different sources of biomolecular data are available for unsupervised analysis and for the analysis of the reliability of clustering results. Even if this topic has been investigated in supervised analysis of bio-molecular data [48, 49], largely less efforts have been devoted to the unsupervised analysis and in particular to the integration of multiple sources of data in the context of stability based methods. However, the integration of multiple sources of data to assess the validity of clustering solutions should in principle significantly improve the reliability of stability-based methods.

7 Conclusions

We presented an overview of stability based methods to estimate the reliability of clusterings discovered in bio-molecular data. These methods, if jointly used with statistical tests specifically designed to discover multiple structures, can be successfully applied to assess the statistical significance and to discover multiple structures in complex bio-molecular data. Summarizing, stability based methods can be applied for:

1. Assessment of the reliability of a given clustering solution
2. Assessment of the reliability of a each cluster inside a clustering
3. Assessment of the reliability of each example to a given cluster
4. Clustering model order selection: selection of the "natural" number of clusters.
5. Assessment of the statistical significance of a given clustering solution
6. Discovery of multiple structures underlying the data

In this introduction we focused on methods, without discussing in detail applications to real bio-molecular data. Anyway, bioinformatics applications of stability based methods can be found in most of the papers cited in this paper (see e.g. [21, 16, 20, 40]). Several problems not discussed in the paper remain open, ranging from the applicability of stability-based methods to problems characterized by very high number of examples and clusters (e.g.: discovery of functional classes of proteins), to their theoretical foundations [42].

Appendix: R software packages implementing stability based methods

Two main *R* packages, implementing stability based methods, are freely available on the web:

1. *Mosclust*: **M**odel **o**rder **s**election for **clust**ering problems. It implements stability based methods to discover the number of clusters and multiple structures underlying bio-molecular data [30]
2. *Clusterv*: **Cluster v**alidation. It implements a set of functions to assess the reliability of individual clusters discovered by clustering algorithms [25]

Overview of the *clusterv* R package

The *clusterv* R package implements a set of functions to assess the reliability of clusters discovered by clustering algorithms [25] This library is tailored to the analysis of high dimensional data and in particular it is conceived for the analysis of the reliability of clusters discovered using DNA microarray data.

Indeed cluster analysis has been used for investigating structure in microarray data, such as the search of new tumor taxonomies [50],[3],[51]. It provides a way for validating groups of patients according to prior biological knowledge or to discover new "natural groups" inside the data. Anyway, clustering algorithms always find structure in the data, even when no structure is present instead. Hence we need methods for assessing the validity of the discovered clusters to test the existence of biologically meaningful clusters.

To assess the reliability of the discovered classes, *clusterv* provides a set of measures that estimate the stability of the clusters obtained by perturbing the original data set. This perturbation is achieved through random projections of the original high dimensional data to lower dimensional subspaces, approximately preserving the distances between examples, in order to avoid too large distortions of the data. These random projections are repeated many times and each time a new clustering is performed. The obtained multiple clusterings are then compared with the clustering for which we need to evaluate its reliability. Intuitively a cluster will be reliable if it will be maintained across multiple clusterings performed in the lower dimensional subspaces. The measures provided by *clusterv* are based on the evaluation of the stability of the clusters across multiple random projections. By these measures we can assess:

1. the reliability of single individual clusters inside a clustering
2. the reliability of the overall clustering (that is, an estimate of the "optimal" number of clusters)
3. the confidence by which example may be assigned to each cluster

The *clusterv* R source package is downloadable from the *clusterv* web-site: `http://homes.dsi.unimi.it/~valenti/SW/clusterv/`

Overview of the *mosclust* R package

The *mosclust* R package (that stands for **m**odel **o**rder **s**election for **clust**ering problems) implements a set of functions to discover significant structures in bio-molecular data [30]. One of the main problems in unsupervised clustering analysis is the assessment of the "natural" number of clusters. Several methods and software tools have been proposed to tackle this problem (see [13] for a recent review).

Recently, several methods based on the concept of stability have been proposed to estimate the "optimal" number of clusters in complex bio-molecular data [22, 23, 24, 17, 25]. In this conceptual framework 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.

Several perturbation techniques have been proposed, ranging form bootstrap techniques [19, 16, 23], to random projections to lower dimensional subspaces [21, 27] to noise injection procedures [20]. All these perturbation techniques are implemented in *mosclust*.

The library implements indices of stability/reliability of the clusterings based on the distribution of similarity measures between multiple instances of clusterings performed on multiple instances of data obtained through a given random perturbation of the original data.

These indices provides a "score" that can be used to compare the reliability of different clusterings. Moreover statistical tests based on χ^2 and on the classical Bernstein inequality [37] are implemented in order to assess the statistical significance of the discovered clustering solutions. By this approach we could also find multiple structures simultaneously present in the data. For instance, it is possible that data exhibit a hierarchical structure, with subclusters inside other clusters, and using the indices and the statistical tests implemented in *mosclust* we may detect them at a given significance level.

The *mosclust* R source package is downloadable from the *mosclust* web-site: `http://homes.dsi.unimi.it/~valenti/SW/mosclust/`

References

[1] Dopazo, J.: Functional interpretation of microarray experiments. OMICS **3** (2006)

[2] Gasch, P., Eisen, M.: Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering. Genome Biology **3** (2002)

[3] Dyrskjøt, L., Thykjaer, T., Kruhøffer, M., Jensen, J., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., Ørntoft, T.: Identifying distinct classes of bladder carcinoma using microarrays. Nature Genetics **33** (2003) 90–96

[4] Kaplan, N., Friedlich, M., Fromer, M., Linial, M.: A functional hierarchical organization of the protein sequence space. BMC Bioinformatics **5** (2004)

[5] Jain, A., Murty, M., Flynn, P.: Data Clustering: a Review. ACM Computing Surveys **31** (1999) 264–323

[6] Kasturi, J., Acharya, R.: Clustering of diverse genomic data using information fusions. Bioinformatics **21** (2005) 423–429

[7] Avogadri, R., Valentini, G.: Fuzzy ensemble clustering based on random projections for dna microarray data analysis. Artificial Intelligence in Medicine (2008) available on line: doi:10.1016/j.artmed.2008.07.014.

[8] Swift, S., Tucker, A., Liu, X.: An analysis of scalable methods for clustering high-dimensional gene expression data. Annals of Mathematics, Computing and Teleinformatics **1** (2004) 80–89

[9] Napolitano, F., Raiconi, G., Tagliaferri, R., Ciaramella, A., Staiano, A., Miele, G.: Clustering and visualization approaches for human cell cycle gene expression data analysis. Int. J. Approx. Reasoning **47** (2008) 70–84

[10] Azuaje, F., Dopazo, J.: Data Analysis and Visualization in Genomics and Proteomics. Wiley (2005)

[11] Giardine, B., Riemer, C., Hardison, R., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W., Nekrutenko, A.: Galaxy: a platform for interactive large-scale genome analysis. Genome Res. **15** (2005) 1451–1455
[12] Ciaramella, A., Cocozza, S., Iorio, F., Miele, G., Napolitano, F., Pinelli, M., Raiconi, G., Tagliaferri: Interactive data analysis and clustering of genomic data. Neural Networks **21** (2008) 368–378
[13] Handl, J., Knowles, J., Kell, D.: Computational cluster validation in post-genomic data analysis. Bioinformatics **21** (2005) 3201–3215
[14] Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics **19** (2003) 1090–1099
[15] Ben-Dor, A., Shamir, R., Yakhini, Z.: Clustering gene expression patterns. Journal of Computational Biology **6** (1999) 281–297
[16] Ben-Hur, A., Ellisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In Altman, R., Dunker, A., Hunter, L., Klein, T., Lauderdale, K., eds.: Pacific Symposium on Biocomputing. Volume 7., Lihue, Hawaii, USA, World Scientific (2002) 6–17
[17] Dudoit, S., Fridlyand, J.: A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology **3** (2002) 1–21
[18] Yeung, K., Haynor, D., Ruzzo, W.: Validating clustering for gene expression data. Bioinformatics **17** (2001) 309–318
[19] Kerr, M., Curchill, G.: Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. PNAS **98** (2001) 8961–8965
[20] 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 **18** (2002) 1462–1469
[21] Smolkin, M., Gosh, D.: Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics **36** (2003)
[22] Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V.: Molecular classification of malignant melanoma by gene expression profiling. Nature **406** (2000) 536–540
[23] 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 **52** (2003) 91–118
[24] Lange, T., Roth, V., Braun, M., Buhmann, J.: Stability-based validation of clustering solutions. Neural Computation **16** (2004) 1299–1323
[25] Valentini, G.: Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data. Bioinformatics **22** (2006) 369–370
[26] Bertoni, A., Valentini, G.: Model order selection for bio-molecular data clustering. BMC Bioinformatics **8** (2007)
[27] Bertoni, A., Valentini, G.: Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses. Artificial Intelligence in Medicine **37** (2006) 85–109
[28] Bertoni, A., Valentini, G.: Discovering Significant Structures in Clustered Biomolecular Data Through the Bernstein Inequality. In: Knowledge-Based Intelligent Information and Engineering Systems, 11th International Conference, KES 2007. Volume 4964 of Lecture Notes in Computer Science., Berlin (2007) 886–891

[29] Bertoni, A., Valentini, G.: Discovering multi-level structures in bio-molecular data through the Bernstein inequality. BMC Bioinformatics **9** (2008)
[30] Valentini, G.: Mosclust: a software library for discovering significant structures in bio-molecular data. Bioinformatics **23** (2007) 387–389
[31] Bertoni, A., Valentini, G.: Randomized embedding cluster ensembles for gene expression data analysis. In: SETIT 2007 - IEEE International Conf. on Sciences of Electronic, Technologies of Information and Telecommunications, Hammamet, Tunisia (2007)
[32] Rand, W.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66** (1971) 846–850
[33] Jain, A., Dubes, R.: Algorithms for clustering data. Prentice Hall, Englewood Cliffs, NJ (1988)
[34] Achlioptas, D.: Database-friendly random projections. In Buneman, P., ed.: Proc. ACM Symp. on the Principles of Database Systems. Contemporary Mathematics, New York, NY, USA, ACM Press (2001) 274–281
[35] Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
[36] Bertoni, A., Valentini, G.: Assessment of clusters reliability for high dimensional genomic data. In: BITS 2007, Bioinformatics Italian Society Meeting, Napoli Italy (2007)
[37] Hoeffding, W.: Probability inequalities for sums of independent random variables. J. Amer. Statist. Assoc. **58** (1963) 13–30
[38] Indyk, P.: Algorithmic Applications of Low-Distortion Geometric Embeddings. In: Proceedings of the 42nd IEEE symposium on Foundations of Computer Science, Washington DC, USA, IEEE Computer Society (2001) 10–33
[39] Johnson, W., Lindenstrauss, J.: Extensions of Lipshitz mapping into Hilbert space. In: Conference in modern analysis and probability. Volume 26 of Contemporary Mathematics., Amer. Math. Soc. (1984) 189–206
[40] Valentini, G., Ruffino, F.: Characterization of lung tumor subtypes through gene expression cluster validity assessment. RAIRO - Theoretical Informatics and Applications **40** (2006) 163–176
[41] Bertoni, A., Valentini, G.: Random projections for assessing gene expression cluster stability. In: IJCNN 2005, The IEEE-INNS International Joint Conference on Neural Networks, Montreal (2005)
[42] Ben-David, S., von Luxburg, U., Pal, D.: A sober look at clustering stability. In: 19th Annual Conference on Learning Theory, COLT 2006. Volume 4005 of Lecture Notes in Computer Science., Springer (2006) 5–19
[43] Ben-David, S., von Luxburg: Relating clustering stability to properties of cluster boundaries. In: 21st Annual Conference on Learning Theory (COLT 2008). Lecture Notes in Computer Science, Springer (2008) 379–390
[44] Harris, M., et al.: The Gene Ontology (GO) database and informatics resource. Nucleic Acid Res. **32** (2004) D258–D261
[45] Brehelin, L., Gascuel, O., Martin, O.: Using repeated measurements to validate hierarchical gene clusters. Bioinformatics **24** (2008) 682–688
[46] Avogadri, R., Brioschi, M., Ruffino, F., Ferrazzi, F., Beghini, A., Valentini, G.: An algorithm to assess the reliability of hierarchical clusters in gene expression data. In: Knowledge-Based Intelligent Information and Engineering Systems, 12th International Conference, KES 2008. Volume 5179 of Lecture Notes in Computer Science., Springer (2008) 764–770

[47] Filippone, M., Masulli, F., Rovetta, S., Zini, L.: Stability indexes and performances of biclustering algorithms. In: CIBB 2008. This issue of Lecture Notes in Computer Science. Springer-Verlag, Berlin (2008)
[48] Troyanskaya, O., et al.: A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *saccharomices cerevisiae*). Proc. Natl Acad. Sci. USA **100** (2003) 8348–8353
[49] Guan, Y., Myers, C., Hess, D., Barutcuoglu, Z., Caudy, A., Troyanskaya, O.: Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biology **9** (2008)
[50] 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 **403** (2000) 503–511
[51] Lapointe, J., Li, C., Higgins, J., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A., Tibshirani, R., Botstein, D., Brown, P., Brooks, J., Pollack, J.: Gene expression profiling identifies clinically relevant subtypes of prostate cancer. PNAS **101** (2004) 811–816