Characterization of lung tumor subtypes through gene expression cluster validity assessment [*]

Giorgio Valentini, Francesca Ruffino
DSI - Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
Via Comelico 39, Milano, Italia
e-mail: {valenti,ruffino}@dsi.unimi.it

Abstract

The problem of assessing the reliability of clusters patients identified by clustering algorithms is crucial to estimate the significance of subclasses of diseases detectable at bio-molecular level, and more in general to support bio-medical discovery of patterns in gene expression data. In this paper we present an experimental analysis of the reliability of clusters discovered in lung tumor patients using DNA microarray data. In particular we investigate if subclasses of lung adenocarcinoma can be detected with high reliability at bio-molecular level. To this end we apply cluster validity measures based on random projections recently proposed by Bertoni and coworkers. The results show that at least 2 subclasses of lung adenocarcinoma can be detected with relatively high reliability, confirming and extending previous findings reported in the literature.

Keywords: Cluster validity; clustering algorithms; bio-molecular taxonomy of tumors; DNA microarray data analysis.

1 Introduction

An open problem in microarray data analysis is the assessment of the reliability of clustering results, since clustering algorithms may find clusters even if no structure is present. Indeed a quantitative data-driven estimate of the reliability of the discovered clusters can support bio-medical researchers in the validation

[*] The authors would like to emphasize that any reference to lung tumors or any other disease should be interpreted in apotropaic sense.

of novel subgroups identified at bio-molecular level. In particular the definition of a more refined bio-molecular taxonomy of tumoral diseases could improve the prediction of patient outcome, the selection of therapies targeted to the bio-molecular characteristics of patients and the search for molecular targets for chemotherapy [1]. In this context the assessment of the validity of clusters discovered in DNA microarray data plays a fundamental role [14, 3].

Different measures and indices have been proposed in the literature to estimate the reliability of clusters discovered by unsupervised learning methods. Classical indices are usually based on the ratio between intercluster and intracluster distances [16, 13, 7]; however, they focus on the validity of the number of the discovered clusters, without providing an estimate of the validity of each individual cluster. An exception is represented by the ”silhouette index” that provides an estimate of the reliability of single clusters as well as an estimate of the membership of each example to a specific cluster [27].

Some recent approaches to estimate cluster reliability are based on the concept of stability with respect to perturbations [23, 25, 28, 26]. In the context of gene expression data, that are usually characterized by relatively high level of noise [12], stability may be considered an important property. Indeed we may study the impact of ”small” perturbations of the original data on the characteristics and composition of the discovered clusters to get insights into their stability: a cluster is considered reliable if ”stable” with respect to data perturbations. The perturbations may be introduced by adding noise [25], or using subsamples of the original data [23, 26] or random subsets of the original feature space [28].

Recently Bertoni and coworkers proposed reliability indices for individual clusters and clusterings based on random projections of the original data [5, 6, 4]. Their method is related to the Smolkin and Gosh [28] approach based on an unsupervised version of the random subspace method [19]. Extending the unsupervised random subspace approach to more general random projections, they proposed cluster stability measures based on similarity between randomly projected data [5]. The proposed reliability measures are well suited to very high dimensional data, as gene expression data usually are [6].

In this paper we apply the stability measures based on random projections to the analysis of the reliability of subclasses discovered in lung tumor patients using high-dimensional gene expression data. The traditional lung tumor classification is based on clinicopathological features, but it has been shown that lung pathologists agreed on lung adenocarcinoma in less than 50% of cases [29]. Moreover there is clinical evidence of different prognostic classes that do not correspond to known histopathological subclassification of lung cancer [10]. For these reasons we try to investigate if a bio-molecular unsupervised analysis of lung cancer and in particular of lung adenocarcinoma may reveal subclasses not detectable with a traditional histopathological approach.

In the next section we summarize the main characteristics of the measures based on random projections for cluster validity assessment. Then in Sect. 3 we provide an extensive experimental analysis of the validity of clusters discovered in lung tumor patients, by applying hierarchical, c-mean and PAM clustering algorithms. In Sect. 4 we discuss the experimental results, showing the effectiveness of random projection-based validity measures for supporting the discovery of novel subclasses of lung adenocarcinoma.

2 Measures based on random projections for cluster validity assessment

2.1 Random projections and cluster reliability

The measures based on random projections estimate the reliability of individual clusters exploiting the redundancy inherent to microarray gene chips. Indeed the number of genes in a chip is usually much larger than the number of samples, and we may reasonably expect that using subsets of genes to perform clustering of tissues, we may obtain meaningful clusters of data. The main idea behind this approach consists in evaluating the stability of the clusters discovered in the original high dimensional space comparing them with the clusters discovered in randomly projected lower dimensional subspaces. In this context the concept of reliability is tied to the concept of stability: a cluster is considered reliable if it is stable, that is if that cluster is maintained in the projected space without too large changes. To properly evaluate the reliability of the clusters, the random projections should not induce too large modifications of the distances between the examples in the projected space. To this end, the concept of random projections with bounded metric distortions, according to the Johnson-Lindenstrauss (JL) theory [21], is used. It has been shown that random projections that obey the Johnson Lindenstrauss lemma do not induce too large distortions [6]. In particular the JL lemma shows that the bounds of the distortion induced by JL lemma-compliant projections depend only on the logarithm of the cardinality of the available data and on the dimensionality of the projected subspace, while, quite surprisingly, does not depend on the dimensionality of the original space [21]. For more details on random projections and the JL lemma see the original Johnson and Lindenstrauss paper [21], the Bertoni's paper [6] or the web-site of *clusterv*, an R package that implements the reliability measures based on random projections (`http://homes.dsi.unimi.it/~valenti/SW/clusterv`).

2.2 Stability measures

The procedure to measure cluster reliability can be divided into several steps [6]:

1. Multiple random projections of the data are generated, choosing a subspace dimension in concordance with the *JL lemma*.

2. Each instance of the projected data is given as input to a suitable clustering algorithm.

3. The resulting clusters are compared with that obtained in the original high dimensional space.

4. The stability measure of an individual cluster is computed by counting how many pairs of elements of the cluster in the original space are preserved in the clusters obtained in the projected space.

Fig. 1 summarizes the procedures needed to compute the stability measures: given a data set S with n examples (Original data), and a set $< O_1, O_2, \dots, O_k >$ of clusters $O_i \in S, \ 1 \leq i \leq n$ generated through a clustering algorithm C, the random projections $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ from the original d-dimensional data set generate m d'-dimensional projected data sets $P_1, P_2, \dots, P_m$. Then a clustering algorithm C is applied to the projected data sets $P_r, \ 1 \leq r \leq m$, giving rise to m clusterings $< A_{r1}, A_{r2}, \dots, A_{rk} >, \ 1 \leq r \leq m$, where $A_{rs}, \ 1 \leq r \leq m, \ 1 \leq s \leq k$ represents the s^{th} cluster of the r^{th} clustering. The m clusterings are finally compared with the original clustering $< O_1, O_2, \dots, O_k >$ obtained by applying the clustering algorithm C to the original high dimensional data. The comparison and the estimate of the reliability measures (boxes inside the dotted line of Fig. 1) are implemented through a *pairwise similarity matrix*. More precisely, the elements $M_{ij}^{(r)}$ of the $n \times n$ symmetric similarity matrix $M^{(r)}$, computed at the r^{th} iteration of the clustering on the projected subspace, store the memberships of examples pairs i, j to the same cluster [15]:

$$M_{ij}^{(r)} = \sum_{s=1}^{k} \chi_{A_{rs}}[i] \cdot \chi_{A_{rs}}[j] \tag{1}$$

where $i, j \in \{1, 2, \dots, n\}$, $A_{rs} \subseteq P_r$ is a cluster returned by a clustering algorithm, k the number of clusters, and $\chi_{A_{rs}} \in \{0, 1\}^n$ is the characteristic vector

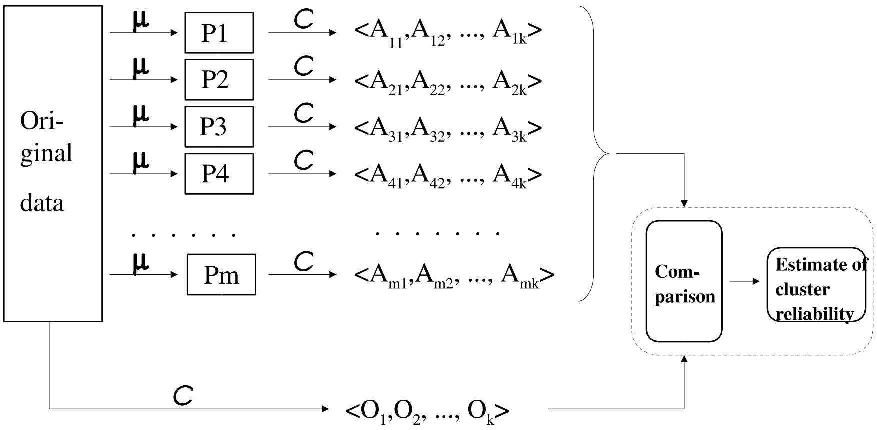

Figure 1: Schematic representation of the procedures needed to compute the stability measures based on random projections.

of A_{rs}, i.e. $\chi_{A_{rs}}[i] = 1$ if $i \in A_{rs}$, otherwise $\chi_{A_{rs}}[i] = 0$. In other words $M_{ij}^{(r)}$ denotes if elements i and j belong to the same cluster. Using multiple random projections of the data we generate multiple instances of projected data that are used by a clustering algorithm to provide multiple sets of clusters (clusterings). Then multiple similarity matrices (one for each clustering) are built, and averaging between them, a similarity matrix $\overline{M}$ that stores the memberships of examples pairs i, j to the same cluster across multiple clusterings is finally obtained.

Using the previously computed similarity matrix, the *stability index* s for an individual cluster O is:

$$s(O) = \frac{1}{|O|(|O|-1)} \sum_{(i,j) \in O \times O, i \neq j} \overline{M}_{ij} \tag{2}$$

The index $s(O)$ estimates the stability of a cluster O by measuring how much the projections of the pairs $(i, j) \in O \times O$ occur together in the same cluster in the projected subspaces. The stability index has values between 0 and 1: low values indicate no reliable clusters, high values denote stable clusters. An overall measure of the stability of the clustering may be obtained averaging between the stability indices:

$$S(k) = \frac{1}{k} \sum_{r=1}^{k} s(O_r) \tag{3}$$

In this case also we have that $0 \leq S(k) \leq 1$, where k is the number of clusters. Finally, the *Assignment-Confidence (AC)* index estimates the confidence of the assignment of an example i to a cluster O, by measuring the frequency by which i appears with the other elements of the cluster O:

$$AC(i,O) = \frac{1}{|O|-1} \sum_{j \in O, j \neq i} \overline{M}_{ij} \tag{4}$$

3 Experimental analysis of the validity of clusters discovered in lung tumor patients

In this section we apply the measures based on random projections to evaluate the reliability of the clusters discovered in gene expression data obtained from lung tumor patients. In particular we try to understand if and at which extent we can characterize subtypes of lung tumors at bio-molecular level. To this end we analyze the results obtained with different clustering algorithms, largely applied for unsupervised analysis of gene expression data within the community of bioinformaticians and bio-medical researchers. At first we provide a validity analysis of clusters discovered with agglomerative hierarchical clustering [24, 31], then we propose the same analysis with PAM (Prediction Around Medoids) [22] and c-mean clustering algorithms [18].

3.1 Experimental environment

The *lung tumor* data set [8] collects 203 histologically defined specimens: 186 lung tumors, subdivided in 139 lung adenocarcinoma (AD), 21 squamous cell lung adenocarcinoma (SQ), 20 pulmonary carcinoids (COID), 6 small-cell lung adenocarcinoma (SMCL) and 17 normal lung (NL) specimens [8]. Each U95A Affymetrix oligonucleotide array provides the gene expression levels of 12600 genes.

From the 12600 original genes of the U95A Affymetrix oligonucleotide array 3312 passed the filter (genes with standard deviation units less than 50 have been excluded), according to the procedures described in [8] and then the gene expression levels have been normalized with respect to the mean and standard deviation. We implemented the pre-processing procedures with R scripts.

Then we evaluated the reliability of the clusters discovered with hierarchical, c-mean and PAM clustering algorithms by using 50 *Plus-Minus-One (PMO)* random projections [6] with a maximum predicted distortion equal to 1.1. ($\epsilon = 0.1$), according to the *JL lemma*. This choice of the ϵ value assures that the euclidean distances between the samples are very likely to be preserved within the bounds of a 10% distortion. For each clustering algorithm we analyzed the reliability of clusterings and individual clusters for a number of clusters ranging from 2 to 20. In particular we computed the *overall stability index* (eq. 3), the *stability indices* for each cluster (eq. 2) and the *Assignment-Confidence index* (eq. 4) for each sample, using the *clusterv* R package [30] to write the R software applications needed for the cluster validity analyses. A summary of the reliability analysis of clusters discovered with hierarchical, PAM and c-means algorithms is summarized in the following sections. Full experimental results, such as the values of the stability indices for different ϵ-level distortions induced by PMO projections, and detailed information about the composition of the obtained clusters, as well as the R source code used for the experiments, are available from the Supplementary Information web-page: `http://homes.dsi.unimi.it/∼valenti/SW/web-lung-validity`.

3.2 Validity analysis of clusters discovered with hierarchical clustering

We applied our stability measures using *PMO* projections and the Ward's hierarchical clustering [31] to analyze the reliability of the discovered subclasses. The results summarized in Tab. 1 and Fig. 2 partially confirmed that the clusters defined by established histological classes [8] are quite reliable. At first, the overall stability indices suggest that pulmonary carcinoid tumors (COID) constitute a well-defined and separated cluster among the different subclasses of lung tumors. Indeed the highest overall stability index is obtained with $K = 2$ clusters, and the first cluster, that collects all the COID patients, shows an individual stability index very close to 1. Moreover, also with $K = 3$ clusters the first COID cluster is highly supported by the s index (Tab. 1). Note that the third cluster, that groups together both normal (NL), lung adenocarcinoma (AD) and squamous cell lung adenocarcinomas (SQ), shows a stability index largely lower than that of the other two clusters (0.692 against 0.998 and 0.836); this fact witnesses the low reliability of the third heterogeneous cluster. Anyway, also partitions characterized by larger number of clusters show relatively high values of the overall stability index, supporting the Bhattacharjee et al. thesis of distinct subclasses of lung adenocarcinoma [8]. For instance, with $K = 4$ clusters, the COID and normal lung (NL) subclasses are classified as reliable by the s index, the big second cluster characterized by several adenocarcinomas (AD) with Small-Cell-Lung-adenocarcinoma (SMCL) and some normal examples is scored as relatively quite reliable ($s = 0.8168$), while the fourth cluster

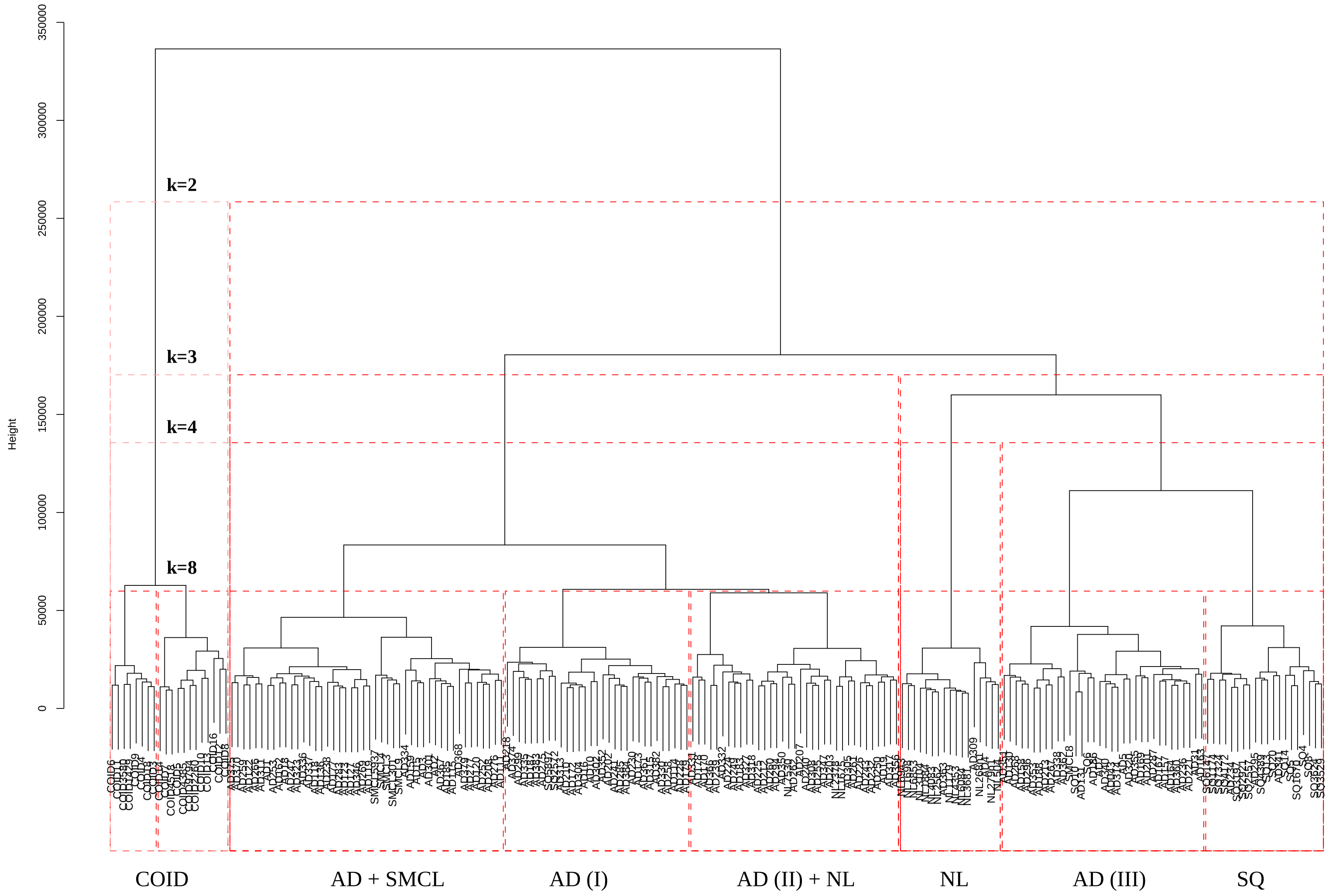

Figure 2: Hierarchical clustering of *Lung tumor* examples (Ward method). Gray dotted lines cut the dendrogram such that exactly k clusters are produced, for $k = 2, 3, 4, 8$. Considering 8 clusters, the first two refers two pulmonary carcinoids patients (COID), the third to a group of lung adenocarcinoma together with small-cell lung adenocarcinoma patients (SMCL), the fourth to a first group of lung adenocarcinoma patients (AD I), the fifth to a second group of adenocarcinoma patients with 3 normal patients (AD II + NL), the sixth to normal (NL) patients, the seventh to a third group of adenocarcinoma patients (AD III) and the last to squamous cell lung adenocarcinomas (SQ). See Table 1 for the corresponding stability indices.

Table 1: Estimate of cluster stability achieved with hierarchical clustering

Overall Stability indices S of the clusterings										
Number of clusters:	2	3	4	5	6	7	8	9	10	20
S value:	0.999	0.842	0.879	0.833	0.851	0.820	0.859	0.846	0.844	0.831
Stability indices s of individual clusters										
2 clusters:	1	2								
s value:	0.998	1.000								
3 clusters:	1	2	3							
s value:	0.998	0.836	0.692							
4 clusters:	1	2	3	4						
s value:	0.998	0.816	0.931	0.772						
5 clusters:	1	2	3	4	5					
s value:	0.998	0.642	0.910	0.950	0.664					
6 clusters:	1	2	3	4	5	6				
s value:	0.998	0.833	0.883	0.908	0.875	0.607				
7 clusters:	1	2	3	4	5	6	7			
s value:	0.806	0.833	0.864	0.908	0.839	0.772	0.723			
8 clusters:	1	2	3	4	5	6	7	8		
s value:	1.000	0.996	0.822	0.863	0.908	0.830	0.769	0.684		
9 clusters:	1	2	3	4	5	6	7	8	9	
s value:	1.000	0.996	0.821	0.859	0.908	0.740	0.749	0.723	0.820	
10 clusters:	1	2	3	4	5	6	7	8	9	10
s value:	1.000	0.996	0.811	0.859	0.908	0.774	0.825	0.738	0.715	0.815
20 clusters:	1	2	3	4	5	6	7	8	9	10
s value:	1.000	1.000	1.000	0.986	0.805	0.840	0.931	0.873	1.000	0.738
	11	12	13	14	15	16	17	18	19	20
s value:	0.715	0.656	0.976	0.698	0.990	0.520	0.909	0.570	0.727	0.681

that groups together adenocarcinoma and squamous cell lung adenocarcinomas (SQ) is scored as less reliable ($s = 0.7157$) (Tab. 1 and Fig. 2). With $K = 8$ the first two subclasses of COID patients are highly reliable ($s \simeq 1$), and quite reliable also the other ones, up to the sixth cluster (normal lung). Interestingly enough, the cluster 3,4,5 could be interpreted as three distinct subclasses inside adenocarcinoma patients, with a relatively high individual cluster stability (Tab. 1, $K = 8$). A clinical follow-up study could confirm if these distinct clusters may correspond to different prognostic subtypes. Cluster 7 also represents another cluster of adenocarcinomas with also SQ and SMCL specimens inside, even if its individual stability index is quite smaller ($s = 0.7692$). In the Bhattacharjee et al. paper, it has been shown that several adenocarcinomas express high levels of squamous-associated genes such as keratin SQ-markers, displaying also histological evidence of squamous features. In our experiments the cluster 8 that groups together most of the squamous cell lung adenocarcinoma (SQ) patients is not strongly supported by the s index, even if we consider $K = 5$ or $K = 6$ clusters, supporting the hypothesis that this group shares common characteristics with other lung adenocarcinoma patients. Note that some adenocarcinoma patients belong to this cluster and other SQ specimens belong to AD(III) adenocarcinoma cluster (Fig. 2); these results support the hypothesis that squamous cell lung adenocarcinoma and a subclass of adenocarcinoma may be part of the same disease at bio-molecular level.

3.3 Validity analysis of clusters discovered with PAM clustering

The validity analysis of the clusters discovered with *PAM (Prediction Around Medoids)* [22] clustering algorithm are summarized in Tab. 2. In this case $2, 3$ and 4 clusters are considered highly reliable ($S > 0.95$), approximately at the same degree.

Moreover the overall stability indices show that all the clusterings, at least to 10 clusters, are quite reliable. When two clusters are considered, we have that the second one collects the pulmonary carcinoid samples (COID) and a single small-cell lung adenocarcinoma sample (SMCL), and the first one all the other samples. Both are highly reliable, and the first bigger one shows the highest

Table 2: Estimate of cluster stability achieved with PAM clustering

Overall Stability indices S of the clusterings										
Number of clusters:	2	3	4	5	6	7	8	9	10	20
S value:	0.971	0.975	0.962	0.940	0.905	0.879	0.888	0.812	0.827	0.749
Stability indices s of individual clusters										
2 clusters:	1	2								
s value:	1.000	0.943								
3 clusters:	1	2	3							
s value:	0.991	0.980	0.953							
4 clusters:	1	2	3	4						
s value:	0.950	0.949	0.949	1.000						
5 clusters:	1	2	3	4	5					
s value:	0.913	0.904	0.941	0.977	0.966					
6 clusters:	1	2	3	4	5	6				
s value:	0.880	0.836	0.921	0.829	0.970	0.993				
7 clusters:	1	2	3	4	5	6	7			
s value:	0.842	0.664	0.904	0.864	0.937	0.970	0.975			
8 clusters:	1	2	3	4	5	6	7	8		
s value:	0.840	0.769	0.895	0.885	0.940	0.825	0.970	0.981		
9 clusters:	1	2	3	4	5	6	7	8	9	
s value:	0.771	0.638	0.932	0.836	0.940	0.634	0.826	0.963	0.766	
10 clusters:	1	2	3	4	5	6	7	8	9	10
s value:	0.720	0.535	0.908	0.806	0.944	0.693	0.830	0.948	1.000	0.883
20 clusters:	1	2	3	4	5	6	7	8	9	10
s value:	0.814	0.576	0.450	0.377	0.485	0.430	0.732	0.302	0.623	0.888
	11	12	13	14	15	16	17	18	19	20
s value:	0.977	0.892	0.758	1.000	1.000	1.000	0.981	0.847	0.866	0.981

stability ($s = 1$). With 3 clusters a third highly reliable cluster ($s = 0.980$) is added: it includes all the normal samples plus some adenocarcinoma samples (AD). Two distinct subclasses of AD samples are detected with 4 clusters: the first contains 118 samples, with some SMCL and SQ samples, the second one 43 samples with some added SQ samples. Analyzing the Assignment-Confidence index of the samples, we see that its value for several SQ samples is lower than 0.5, showing that the membership of some SQ samples to AD subclasses is low. The cluster with normal samples (the third) is better defined and the fourth cluster that comprises now only COID samples has the highest reliability ($s = 1$). With 5 clusters we register the same previous clusters, but the cluster of COID samples is split into two reliable clusters. With 6 clusters two AD subclusters are also detected with relatively high reliability; moreover a new cluster characterized by squamous cell lung adenocarcinomas (SQ) is found as quite reliable ($s = 0.829$); note that this cluster includes also 3 AD samples, but the AC index is low for at least one of them ($AC = 0.5$). Three and four subclasses of adenocarcinoma patients are detected respectively with 7 and 8 clusters. Note that in both cases the overall stability index (close to 0.9), supports the structures found in the data. Anyway, the second subcluster of AD samples shows a relatively low reliability ($s = 0.664$ and $s = 0.769$ respectively with 7 and 8 clusters), while the other are quite stable ($s \sim 0.85$). When 10 clusters are considered, we find four subclasses of adenocarcinoma (cluster n.1,2,4 and 6), but they are less supported by the stability index. A fifth AD subclass that includes also SMCL samples is quite stable ($s = 0.830$). The last three clusters refer to 3 subclasses of COID, with a high stability index (respectively $0.948, 1.000$ and 0.883). With 20 clusters the overall stability index is quite low ($S = 0.749$) and most of the cluster are poorly reliable, apart from COID clusters and some small cluster of AD and SCLC patients.

Summarizing, the results support the hypothesis of different subtypes inside lung adenocarcinoma, but the number and the limits between the subclasses are not clearly defined.

3.4 Validity analysis of clusters discovered with c-means clustering

The values of the overall stability computed with the c-means clustering algorithm are not so high as those computed with PAM. This fact reflects the larger instability of c-means with respect to PAM algorithm. However with c-means we have relatively large values of the overall stability index for a large range of clusters, from 2 to at least 8 (Tab. 3). With 2 clusters we have the largest value of the overall stability index ($S = 0.969$); the second cluster that includes all the COID samples and a single SMCL sample achieves the highest stability ($s = 1$), but also the first cluster with all the remaining samples is highly reliable ($s = 0.939$). With 3 clusters, two subclasses of adenocarcinoma samples are detected, even if the second one is less reliable than the first ($s = 0.811$ vs. $s = 0.981$), and these subclasses are maintained also with 4 clusters, where a reliable cluster of normal samples ($s = 0.944$) is also present. With 5 clusters a third cluster of mixed AD and SQ samples is added ($s = 0.842$), but with 6 clusters c-means finds 4 AD subclasses (cluster 1,2,3 and 4), but only the second is reliable ($s = 0.874$). Considering 7 clusters, the number of the AD subclasses is reduced but their stability is reduced, as well as the value of the overall stability index ($S = 0.7$). With 8 clusters we come back to more AD subclasses, but at least 3 of them show a very low reliability. When more than 8 clusters are generated, the overall stability decreases, as well as the reliability of the adenocarcinoma subclasses: only the SQ, COID and normal clusters show an high stability index.

Table 3: Estimate of cluster stability achieved with c-mean clustering

Overall Stability indices S of the clusterings										
Number of clusters:	2	3	4	5	6	7	8	9	10	20
S value:	0.969	0.929	0.887	0.874	0.792	0.702	0.743	0.668	0.672	0.706
Stability indices s of individual clusters										
2 clusters:	1	2								
s value:	0.939	1.000								
3 clusters:	1	2	3							
s value:	0.981	0.811	0.996							
4 clusters:	1	2	3	4						
s value:	0.941	0.736	0.944	0.929						
5 clusters:	1	2	3	4	5					
s value:	0.922	0.685	0.957	0.842	0.963					
6 clusters:	1	2	3	4	5	6				
s value:	0.586	0.874	0.747	0.636	0.996	0.910				
7 clusters:	1	2	3	4	5	6	7			
s value:	0.502	0.429	0.421	0.740	0.822	1.000	1.000			
8 clusters:	1	2	3	4	5	6	7	8		
s value:	0.592	0.474	0.576	0.874	0.813	0.955	0.729	0.931		
9 clusters:	1	2	3	4	5	6	7	8	9	
s value:	0.505	0.457	0.662	0.447	0.771	0.480	0.984	0.796	0.914	
10 clusters:	1	2	3	4	5	6	7	8	9	10
s value:	0.421	0.495	0.489	0.449	0.823	0.698	0.719	0.913	0.843	0.872
20 clusters:	1	2	3	4	5	6	7	8	9	10
s value:	0.598	0.485	0.477	0.494	0.735	0.685	0.477	0.617	0.504	0.683
	11	12	13	14	15	16	17	18	19	20
s value:	0.535	0.833	0.932	0.783	0.691	0.867	1.000	0.985	0.913	0.830

4 Discussion

The analysis of the reliability of clusters based on random projections of the original data depends on the choice of the clustering algorithm. This is common to all the methods based on perturbations [28, 23, 25, 6]. In this paper we presented the stability results obtained with different clustering algorithms, largely used between the community of bioinformaticians. Indeed, on one hand it is well-known that different clustering algorithms can identify different features and characteristics of the data [20], and on the other hand patterns consistently identified by different clustering algorithms and supported by high values of validity indices can be considered robust and reliable [9].

Comparing the results obtained with different clustering algorithms and using the validity indices based on random projections, we may get insights into the most significant clusters of patients and we can perform a quantitative evaluation of the reliability of the discovered clusters. For instance, with 2 clusters all the clustering algorithms find the same two clusters: a clusters of COID examples and a larger one collecting all the other examples. In all cases the overall stability index and the individual stability index of each discovered cluster is very high (equal to 1 or very close to 1). This is not surprising, because it is known from bio-molecular cancer research that pulmonary carcinoids cases are divergent from malignant lung tumors [2]. With 3 clusters the PAM algorithm finds a first highly reliable cluster, similar to the third found by the hierarchical clustering algorithm, that shows a significantly lower stability index, because hierarchical clustering includes in this clustering normal samples too (even if with a relatively low AC index), while PAM separates normal samples in the second highly reliable cluster.

Considering different number of clusters, the overall results confirm the hypothesis of distinct subclasses among lung adenocarcinoma. Indeed at least two subclasses are clearly detected by all the three clustering algorithms with high reliability (see Sect. 3 and Supplementary Information). Moreover other two subclasses are detected by both PAM and c-means clustering algorithms, even if their reliability is lower (especially with c-means). Hierarchical clustering finds only three different AD subclasses (Fig. 2), quite consistent and reliable. However, the bounds between AD subclasses are not clearly defined: we have different sample composition of the subclasses if we consider different clustering algorithms. Anyway, analyzing the AC indices (eq. 4) of the examples that belong to the different subclasses we may understand which are the examples more responsible for the uncertainty of clusters bounds.

From these results, we could hypothesize a hierarchical structure of the adenocarcinoma subclasses, considering two quite well defined subclasses (clearly detected e.g. with PAM when 4 clusters are considered), and two other ones (less reliable) derived from the previous two when SQ samples segregated in a highly reliable separated cluster (see the 8-clusters PAM clustering in Supplementary Information and Tab. 2). Note that a reliable separated cluster with SQ samples is found by all the three clustering algorithms, as well as a reliable cluster with normal samples. In all cases (except in part for clusters discovered with hierarchical clustering) if we try to find finer structures using more subdivisions of the available data (e.g. more that 10 clusters), the reliability of the overall clustering decreases, and only some small clusters can be considered reliable according to their stability indices: in other words we cannot find significant structures in the data (see Supplementary Information and Tab. 1, 2 and 3). Summarizing, the stability analysis across different clustering algorithms confirm the hypothesis of distinct subclasses among lung adenocarcinoma [8]: from 2 to 4 distinct subclasses are detected at different degree of reliability (according to the stability index of each individual cluster). These classes, defined without using any a priori information about the examples, need to be clinically validated, and follow-up studies could be considered in order to evaluate if they can be considered relevant for prognosis and outcome prediction purposes.

5 Conclusions

We evaluated the reliability of clusters discovered by hierarchical, c-mean and PAM clustering algorithms in lung adenocarcinoma patients, showing that the reliability measures based on random projections can support bio-medical researchers in the identification of stable clusters of patients and in the discovery of new subtypes of diseases characterized at bio-molecular level. In particular we detected from 2 to 4 lung adenocarcinoma subclasses, with different degrees of reliability and we found also other reliable clusters, such as squamous cell lung adenocarcinomas mixed with subgroups of lung adenocarcinomas, confirming at bio-molecular level the histological evidence of squamous features in subsets of AD samples. Note that in previous studies [8], using only DNA microarray data of lung adenocarcinoma patients, 4 subclasses of adenocarcinomas have been detected by using hierarchical clustering and a probabilistic model-base clustering [11]. Subclasses of AD patients have been hypothesized also in another study of lung adenocarcinoma [17]. Our analysis confirms and extends these results, providing also a quantitative estimate of the reliability of the discovered subclasses using stability measures based on random projections. A future work could consists in clinical follow-up studies to understand if the discovered subclasses corresponds to relevant clinically distinct diseases detectable at bio-molecular level.

6 Acknowledgments

We would like to thank the reviewers for their comments and suggestions. This work has been developed in the context of the *CIMAINA* Center of Excellence.

References

[1] A. Alizadeh, D.T. Ross, C.M. Perou, and M. van de Rijn. Towards a novel classification of human malignancies based on gene expression. *J. Pathol.*, 195(1):41–52, 2001.

[2] R Anbazhagan et al. Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles. *Cancer Research*, 59:5119–5122, 1999.

[3] F. Azuaje. A cluster validity framework for genome expression data. *Bioinformatics*, 18(2):319–320, 2002.

[4] A. Bertoni, R. Folgieri, F. Ruffino, and G. Valentini. Assessment of clusters reliability for high dimensional genomic data. In *BITS 2005, Bioinformatics Italian Society Meeting*, Milano Italy, 2005.

[5] A. Bertoni and G. Valentini. Random projections for assessing gene expression cluster stability. In *IJCNN 2005, The IEEE-INNS International Joint Conference on Neural Networks*, Montreal, 2005.

[6] A. Bertoni and G. Valentini. Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses. (submitted).

[7] J.C. Bezdek and N.R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man and Cybernetics Part B*, 28:301–315, 1998.

[8] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, and M. Meyerson. Classification of human lung carcinoma by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, 98(24):13790–13795, 2001.

[9] N. Bolshakova, F. Azuaje, and P. Cunningham. An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, 21:451–455, 2005.

[10] O.S. Breathnach et al. Clinical features of patients with stage iiib and iv bronchioloalveolar carcinoma of the lung. *Cancer*, 86(7):1165–1173, 1999.

[11] P. Cheeseman and J. Stutz. Bayesian classification (autoclass): Theory and results. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurasamy, editors, *Advances in Knowledge Discovery and Data Mining*, volume 2, pages 153–180. MIT Press, Cambridge, MA, 1996.

[12] J.J. Chen, R. Delongchamp, C. Tsai, H. Hsueh, F. Sisatare, K. Thompson, V. Deasi, and J. Fuscoe. Analysis of variance components in gene expression data. *Bioinformatics*, 20(9):1436–1446, 2004.

[13] D.L. Davies and D.W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 1(2):224–227, 1979.

[14] S. Dudoit and J. Fridlyand. A prediction-based method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7):1–21, 2002.

[15] S. Dudoit and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9):1090–1099, 2003.

[16] J. Dunn. Well separated clusters and optimal fuzzy partitions. *J. Cybernetics*, 4:95–104, 1974.

[17] M.E. Garber et al. Diversity of gene expression in adenocarcinoma of the lung. *PNAS*, 98(24):13784–13789, 2001.

[18] J.A. Hartigan and M.A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[19] T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[20] A.K. Jain, M.N. Murty, and P.J. Flynn. Data Clustering: a Review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[21] W.B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. In *Conference in modern analysis and probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. Amer. Math. Soc., 1984.

[22] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[23] M.K. Kerr and G.A. Curchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *PNAS*, 98:8961–8965, 2001.

[24] B. King. Step-wise clustering procedures. *J. Am. Stat. Assoc.*, 69:86–101, 1967.

[25] L.M. McShane, D. Radmacher, B. Freidlin, R. Yu, M.C. Li, and R. Simon. Method for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469, 2002.

[26] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52:91–118, 2003.

[27] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comp. App. Math.*, 20:53–65, 1987.

[28] M. Smolkin and D. Gosh. Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, 36(4), 2003.

[29] J.B. Sorensen, F.R. Hirsch, A. Gazdar, and J.E. Olsen. Interobserver variability in histopahologic subtyping and grading of pulmonary adenocarcinoma. *Cancer*, 71:2971–2976, 1993.

[30] G. Valentini. Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data. *Bioinformatics*, 22(3):369–370, 2006.

[31] J.H. Ward. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58:236–244, 1963.