

Significance analysis of clustering high-throughput biological data

Hasan H. Otu*, Shakirahmed Kolia, Jon Jones, Osman Osman, and Towia A. Libermann
Beth Israel Deaconess Medical Center Genomics Center and Bioinformatics Core
Harvard Medical School, Boston, Massachusetts 02115
Email: \{hotu, jjones6, tliberma\}@bidmc.harvard.edu, skolia@eecs.tufts.edu, osman@mit.edu

Abstract—In the post-genomic era, the availability of complete genome sequences gave rise to high throughput systems such as gene chips and protein arrays. These techniques revolutionize our understanding of biology by simultaneously probing thousands of biological entities at any given time. Unsupervised classification and clustering have emerged as important methods of analyses, which can be used to group samples with a similar molecular profile and/or molecules with a similar expression profile. However, techniques like hierarchical clustering, k-means, and self organizing maps (SOM) have been extensively used with little attention to the significance of their results. We propose a general method utilizing bootstrap technique to assign confidence levels to clustering results of high throughput biological data. We apply the proposed method to real genomics and proteomics data regarding Renal Cell Cancer (RCC), which is the most common malignancy of the adult kidney. We utilize protein profiles from IL-2 treatment responders and non-responders among metastatic RCC patients using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI TOF-MS). We also use gene expression data using Affymetrix HG-U133A chips for primary RCC tumors, inquiring the Union International Contre le Cancer's (UICC) TNM classification.

I. INTRODUCTION

High throughput techniques such as gene chips and protein arrays measure thousands of data points on the whole network that governs how proteins and genes affect each other. In addition to providing promising applications in life sciences, these innovative methods bring forth a series of computational challenges. These range from problems in experimental design to quantification of the reliability and reproducibility of the results obtained from the experiment [1], [2].

Utilization of high-throughput techniques help researchers understand the molecular dynamics that characterize different conditions in the experiment. When there are two well defined conditions, the main goal is to find molecules whose expression levels have changed significantly between the two conditions. There are numerous methods used for this purpose from simple fold change analysis to more complicated statistical techniques like permutation tests or Bayesian methods [3], [4]. When there are multiple conditions, the problem can be grouped in two subtopics: class prediction and class discovery. Here we refer to the word “class” in the sense of a phenotype that defines a condition in the experiment, e.g. tumor samples in a cancer study.

In class prediction, the goal is to find a set of molecules that can successfully predict samples from a certain condition in the experiment. Class discovery rests on the assumption that there exist molecular profiles, which can differentiate between different biological states. These profiles, which are gene expression patterns in the case of microarray analysis, can help discover new classes that are otherwise unknown. The approaches used to handle class prediction and discovery are supervised and unsupervised learning techniques, respectively.

Prominent supervised learning methods used in microarray analysis are Bayesian classifiers, linear discriminant analysis, nearest neighbor classification trees, and support vector machines among others [5]. In the unsupervised learning counterpart, the prevailing methods have been hierarchical clustering, k-means clustering, self-organizing maps, principal components analysis and variations of these techniques [6]. The application of these algorithms to cluster samples has implications in finding new phenotypic groups. On the other hand, clustering molecules is also important due to the premise that functionally related molecules have similar expression profiles.

The modularity seen in the interaction network formed by biological molecules have implications in finding new pathways. As biological entities work in groups to perform relatively distinct biological functions, clustering of genes or proteins help in understanding the dynamics of cell’s functional organization. For example, when metabolic networks of 43 distinct organisms were analyzed [7], the average clustering coefficient (a measure of the likelihood of a network to form clusters) for these metabolic networks were an order of magnitude larger than that of a random network with similar size. Similar observations have been made for other biological networks [8], [9] showing tight modular clusters. Both class discovery and finding functionally related molecules make unsupervised grouping of these data points in high-throughput biological experiments extremely crucial. Hierarchical clustering [10], SOMs [11], and k-means clustering [12] have been the prevailing methods for such analyses with little effort spent in assigning significance to the results of these methods.

On the other hand bootstrap has emerged as a technique to infer the variability in an unknown distribution through resampling of the observed data set [13]. Variants of this technique has been used for assessing reliability to clusters from gene expression data [14]–[19]; all focusing on the
method of hierarchical clustering. However, these approaches rely on distributional assumptions such as the standard error or differential expression in gene signal values [14], [16], [17], [19] or require a priori knowledge or estimation of the number of clusters [18].

In this paper we propose a method that assigns significance to clustering results of high throughput biological data using bootstrap. For the remainder of the paper we lay out the algorithmic and implementation principles for hierarchical clustering of samples in an experimental setting. The proposed method can easily be extended to clustering of features (e.g. genes in microarray experiments) or clustering using other techniques (e.g. k-means). Our approach is in the spirit of Felsenstein's application of bootstrap to place confidence intervals on phylogenies [20], which has been shown to be unbiased regarding the use of the bootstrap in the tree problem [21].

II. METHODS

Let the expression matrix $M(F, S)$ denote the results of a high-throughput biological experiment where $F = \{f_1, \ldots, f_m\}$ denote the features (e.g. gene or ESTs on a microarray), $S = \{s_1, \ldots, s_n\}$ denote the samples (e.g. patient tumors in a cancer study) and $M_{ij}$ denote the signal value for the $i^{th}$ feature in the $j^{th}$ sample.

A. Hierarchical Clustering

Hierarchical clustering is an iterative clustering method [22] where initially each object is considered as a cluster. At each step two “closest” clusters are joined to form a new cluster until there is only one cluster left. Here we summarize the algorithm using our sample set $S$ where each sample can be considered as a vector of length $m$ representing the signal values of the features.

Initialization: Let $D$ be a $n \times n$ distance matrix where $D_{ij} = d(i, j)$ is the distance between $s_i$ and $s_j$. Popular distance metrics used for this purpose are variants of correlation coefficient or Euclidean distance between the samples. Assign each sample $s_i$ to cluster $C_i$.

Iteration Find $i$ and $j$ for which $D_{ij}$ is minimal. Define a new cluster $C_k = C_i \cup C_j$ and remove clusters $C_i$ and $C_j$. Update the distance matrix $D$ by removing $i^{th}$ and $j^{th}$ columns and rows and adding a new column and row for the new cluster, where $D_{kl} = \min\{d(p, q) : p \in C_k, q \in C_l\}$. Repeat until there is one cluster left.

In particular, the procedure defined above is called “single linkage” clustering due to the definition of $D_{kl}$. If one defines $D_{kl}$ to be the maximum (or average) of the distances between every pair of objects in clusters $C_k$ and $C_l$, the process is called “complete (or average) linkage” clustering. Unless stated otherwise, we present results using the average linkage method; all three methods are implemented in the accompanying software.

Hierarchical clustering procedure can be represented by a binary tree, which we revisit in the next subsection. In this representation, leaves correspond to objects and nodes correspond to new clusters formed by joining two clusters at each iteration.

B. Binary and Consensus Trees

Let $P(S)$ be the power set of our sample set $S$ and $T \subseteq P(S)$. $T$ is called a binary tree if the following conditions are met [23]:

- $S \in T$ and $\{s_i\} \in T$, $i = 1, \ldots, m$.
- If $A, B \in T$ and $A \cap B \neq \emptyset$, then $A \subseteq B$ or $B \subseteq A$.
- If $A \in T$ and $|A| > 1$ then $\exists A_1, A_2 \in T$ such that $A = A_1 \cup A_2$.

A binary tree, therefore, is an ordered tree that has at most two child nodes for each of its nodes.

One of the basic problems in biological classification is combining more than one clustering result in a unified representation. In the case of tree-based clustering methods, the problem becomes that of finding the consensus tree for a collection of trees obtained using a given set of objects. However, there are inherent limitations on the solutions to the consensus tree problem so that no method can prevail as an approach that achieves a list of desired properties [24] (we do not discuss these properties as it is beyond the scope of this paper). Therefore, different methods have emerged as solutions to the consensus tree problem, each capturing some common substructure in the input set of trees [25].

In the proposed method we generate a set of binary trees utilizing the bootstrap method and provide the consensus tree for this input set. We used the so-called greedy consensus tree approach also adapted by PHYLIP and PAUP, two ubiquitous phylogeny analysis programs [25]. This approach refines majority rule, strict, and loose consensus trees providing an interpretation with higher resolution.

C. Bootstrap

Bootstrap involves assessing the accuracy of statistical estimates by resampling from the data. For example, if one is calculating a statistical parameter $p$ using observed data $d_1, \ldots, d_n$ to estimate the true value of the parameter $p_1$, one can assess accuracy of $p$ (in estimating the true value $p_1$) using bootstrap. Let $p = E(d_1, \ldots, d_n)$ where $E$ is some method of statistical calculation. Bootstrap procedure suggests that we resample our data and obtain $d_1^*, \ldots, d_n^*$ where $d_i^*$ is chosen randomly from the original data set. During this resampling some data points might be omitted and some data points might be represented more than once. For example in one of the resampled data sets we might have $d_1^* = d_{15}$, $d_{12}^* = d_{15}$, $d_{27}^* = d_{15}$, $d_2^* = d_{23}$, $d_{11}^* = d_{24}$, and $\exists i$ such that $d_i^* = d_2$, $1 \leq i \leq n$, $61 \leq n$. Hence, in this resampled data set $d_{15}$ and $d_{24}$ from the original data set have been represented two and three times respectively while $d_2$ from the original data set has not been represented at all. For each resampled data set one can compute the parameter $p^* = E(d_1^*, \ldots, d_n^*)$ and repeat this process for a large number of times. The collection of $p^*$’s can be used to assess the accuracy of the estimate $p$ in approximating $p_1$ [13].
In 1985, Felsenstein applied the bootstrap method to place confidence in phylogenies estimated using molecular sequences [20]. In this case, starting data matrix is a multiple alignment of the molecular sequences corresponding to the samples. Resampled data sets are obtained as discussed above by assuming the observed data point $d_i$ is the $i^{th}$ column in the multiple alignment. A phylogenetic tree ($p^*$) is obtained for each resampled data set. The collection of these trees ($p^*$s) are used to place confidence on the estimated tree ($p$) in approximating the true phylogeny ($p_t$). In 1996, Efron et al. showed that this method is not biased and it is a valid use of bootstrap in assessing confidence to the clades in the tree.

Following a similar approach, we resample features from our expression matrix $M(F,S)$ and repeat this process $B$ times performing hierarchical clustering of the samples for each resampled data set. Let $T$ be the hierarchical clustering of samples obtained using $M(F,S)$ and let $T^*$ be the hierarchical clustering of samples obtained using $M(F^*,S)$ where $F^*$ is a resampled data set of features $f_1, \ldots, f_m$ as discussed above. The set of trees ($T^*$s) are used to assign confidence to the clades obtained in hierarchical clustering in two ways. We obtain a consensus tree $T_c$ using $T^*$s and for each clade we indicate the percentage of $T^*$s, which contained the clade. This in turn represents the % confidence in observing the clade in the true tree $T_t$. We repeat this process for the clades that exist in the observed tree $T$ indicating the confidence in the clusters obtained using the original data set.

In high throughput biological experiments number of features far exceed number of samples and a subset of features are represented more than once on the experimental platform yielding redundant information. Therefore, on the average we would assume to see stable clusters showing up on hierarchical clustering trees using resampled data sets. In a more general setting, this would mean class membership for samples encoding the approach outlined here as a general approach, which can be applied to other clustering techniques with slight modifications.

### D. Software

We implemented the proposed algorithm in the software ASSESS (Assigning Significance to Subclusters of Experimental SampleS), which is available through http://www.bidmcgenomics.org/ASSESS. ASSESS is written in JAVA and takes the expression matrix $M(F,S)$ as its input to produce hierarchical clustering for the samples using bootstrap as discussed. The user can view the original tree, any one of the generated “random” trees, and the consensus tree, all showing confidence values for the nodes. SAM-DT provides feature and/or sample standardization, three methods of hierarchical clustering (average, single, complete linkage) and three distance metrics (two variants of correlation and Euclidean). In the results that follow, we have used average linkage clustering with feature standardization utilizing $1 - r$ as the distance metric, where $r$ is the Pearson’s correlation coefficient between the samples. More details about the software can be found in the online manual.

### E. Simulated Data

In the simulated data set, our goal was to mimic a high-throughput biological experiment in a controlled manner to test the validity and properties of the proposed algorithm. We wanted to simulate 4 different biological states, each with four replicates. In order to do this, we started with four samples $A_1, B_1, C_1,$ and $D_1$, which are generated independent from each other, each with 5,000 features sampled independently from a log-normal data distribution.

We then independently added additive white Gaussian noise three times per sample to each sample to obtain our sample set $s_i$, where $s \in \{A, B, C, D\}$ and $i \in \{1, 2, 3, 4\}$. This resulted in a data set with 16 samples composed of four groups where the samples in a given group (e.g., $A_1, A_2, A_3,$ and $A_4$) represent “replicates” of one biological state.

### F. RCC Data

Renal Cell Carcinoma is the most common malignancy of the adult kidney and one of the top ten human malignancies in the developed world. It has a heterogeneous clinical presentation with clear cell RCC (cRCC) being the most common type showing 80% occurrence frequency [26]. Although surgery is highly effective for the treatment of localized RCC, there is no curative treatment available for patients suffering from metastatic disease.

Aggressive treatment with radical nephrectomy and adjuvant immunotherapy has been shown to provide a survival benefit in metastatic RCC patients. Five-year survival rates were improved when IL-2 was used as adjuvant agent instead of IFN-α (19.6% vs. 10%) [27]. At present, there is no reliable bio-marker available for the identification of the subset of patients that will respond to IL-2. The administration of IL-2, however, is associated with considerable toxicity. Therefore it becomes crucial to identify a priori if a patient is likely to respond to IL-2.

An important and general question in cancer treatment is to understand the extent and stage of the cancer for appropriate management of the patient. For this purpose UICC’s TNM classification has emerged as the global standard. However, there has been no study to test whether the transcriptional profiles of tumors follow the T stage suggested by the TNM classification. We analyze two RCC data sets to inquire the IL-2 response and TNM classification problems using two different high-throughput approaches rendering a test bed for the proposed algorithm.

The first data set consists of 34 frozen samples from primary tumors of patients with metastatic cRCC who had undergone radical nephrectomy and adjuvant immunotherapy with IL-2. Twelve patients responded and 22 patients had not responded to IL-2 therapy. We used SELDI-TOF MS to obtain protein profiles for all 34 samples in duplicate using three chip surfaces: hydrophobic, weak cationic exchange chromatography, and immobilized metal affinity capture protein arrays (H50, CM10, and IMAC30 ProteinChip arrays; Ciphergen) [28]. Protein peaks were defined by a signal-to-noise ratio $\geq 3$ and detected using the Ciphergen Biomarker Wizard software.
We interrogated the range of 2000 Da - 40000 Da for peaks and normalized the peak data with the total ion current method after baseline correction.

Our second data set consists of 21 primary tumor cRCC samples with 8 T1, 5 T2, and 8 T3 stage samples. We used Affymetrix HG-U133A chips to obtain the transcriptional profiles of all the samples represented by \( \approx 22,000 \) genes and ESTs [29]. Samples were analyzed using dChip [30], where a smoothing spline normalization method was applied prior to obtaining model-based gene expression values.

III. RESULTS AND DISCUSSION

We have tested the proposed approach in simulated and real (genomics and proteomics) data sets. In all runs we have repeated the bootstrap procedure for one thousand times \((B=1000)\). In the simulated data set the correlation between the “replicates” were between 0.88 and 0.92 showing high similarity. In Figure 1, we show the original and consensus trees with % confidence values.

![Figure 1](image1.png)

(a) Fig. 1. Original (a) and consensus (b) trees obtained from the simulated data set. Numbers at the nodes denote the % confidence for the corresponding clade. Sample names denote the designed group names where we have four distinct groups, A, B, C, and D, each with four replicates.

(b)

The relation between four groups A, B, C, and D are random, which is reflected by the consensus tree. The grouping of As with Bs and Cs with Ds occur 32 and 47% of the time respectively showing low confidence. More importantly, the out-group status of the B group in the original tree is correctly not seen in the consensus tree. Therefore mere observation of the original tree without the aid of the proposed approach would lead one to hypothesize that group A shows similarity to the C-D clade as opposed to the B clade. In reality, the statistical tests do not support evidence for this interpretation as shown by the consensus tree and the confidence values in the original tree.

In our proteomic study, SELDI TOF-MS detected a total of 730 protein peaks (279 on IMAC30, 271 on CM10, and 180 on H50 ProteinChip™ arrays). When we applied the proposed method on this data set, we observed that the distinction between the non-responding (NR) and responding (R) groups was not evident using hierarchical clustering (Figure 2). This suggests the use of supervised analysis to obtain biomarkers distinguishing the two states.

![Figure 2](image2.png)

Fig. 2. Original (a) and consensus (b) trees obtained from the proteomic data regarding IL-2 response in cRCC. Numbers at the nodes denote the % confidence for the corresponding clade. Sample names are patient id numbers followed by the response status: R for responding and NR for non-responding patients.

The two trees, original and consensus, agreed for the majority of cases, except for the placements of patient 129 and the clade containing patients 141 and 176. Similar to the previous case, the very first branching did not prove to be significant implying the lack of existence for two distinguished groups of samples. The improvement in interpreting the results enhanced by the proposed method also become evident for stable sub-clusters. For example, the clade containing samples 275, 132, 244, 70, 234, 171, and 219 prove to be significant and can be used for well supported biological discussions.

In Figure 3, we show the results when the proposed method was applied on the genomic data regarding 21 cRCC samples. These samples are all primary tumors with stages ranging from T1 to T3 based on TNM classification. In 2002, a process to improve the TNM classification has been initiated, which will be tested and evaluated over the next year [31]. One of the points of criticism of the current 1997 TNM staging system has been the cutoff size between stages T1 and T2. While in the 1987 system T2 tumors were defined as larger than 2.5 cm, the 1997 system defines T2 as tumors larger than 7 cm [32]. Recent reports have suggested to reevaluate the cutoff-point between T1 and T2 tumors, as larger T1 tumors were recognized to have an outcome closer to higher T stage tumors [33], [34]. Given that our T1 clear cell RCC tumors displayed a wide range of sizes (range: 3.5 - 7cm; median size: 5.5 cm) we decided to analyze the gene expression pattern of our T1,
The clustering, based on transcriptional profiling of approximately 22,000 genes and ESTs on the Affymetrix HG-U133A chip, did not group the samples according to their T stages. T1 stage tumors that clustered with higher stage tumors (like 12A5 clustering with a T2 and a T3 stage tumor) were not particularly among the large T1 stage tumors (e.g. 12A5 is a 4-cm T1 stage tumor). In the resulting hierarchical clustering, unlike the previous two data sets, first level branching showed medium significance (60 and 61% confidence). However, the 90+ % confidence levels observed in the lower level branchings imply that one should focus on these stable sub-clusters and not over-interpret the two main clusters formed at the top level. These result suggest that TNM classification is not reflected by the complete transcriptional profiling based on the Affymetrix HG-U133A platform. However, this does not negate the possibility of a subset of genes whose expression profiles would correlate with the TNM classification.

Fig. 3. Original (a) and consensus (b) trees obtained from the genomic data regarding TNM classification in cRCC. Numbers at the nodes denote the % confidence for the corresponding clade. Sample names are tumor stages followed by patient id numbers.

The algorithm involves sampling data points with replacement while keeping all of the original experimental conditions. Each resampled data set is used for clustering to infer significance on results from the original data set. In particular we tested the proposed algorithm on hierarchical clustering, which is by far the most popular clustering technique used in literature regarding high throughput biological data.

We showed the simulation of the proposed method by an accompanying platform independent software with test results on a synthetic data set. Our results imply that the use of the proposed method can successfully prevent incorrect interpretation of relations observed in direct application of hierarchical clustering. The confidence assigned by the proposed method guides the user to focus on sub-clusters that occur with high stability.

We also applied the proposed method on real genomics and proteomics data regarding RCC. We used protein profiles from IL-2 treatment responders and non-responders among metastatic RCC patients using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI TOF-MS). We also used gene expression data using Affymetrix HG-U133A chips for primary RCC tumors, inquiring the Union International Contre le Cancer’s (UICC) TNM classification.

Our results suggest that in the proteome data set, we did not identify two distinct groups of samples that correlate with response to IL-2. Similarly, there was not significant association in the clusters formed using different stage cRCC primary tumors that followed TNM staging. In conclusion, the application of the three data sets demonstrated successful use of the proposed method on synthetic and real (genomic and proteomic) data sets. We plan to further the proposed method by applying it to other clustering methods like k-means clustering.

IV. CONCLUSIONS

High throughput systems such as gene chips and protein arrays are being extensively used to understand molecular mechanisms underlying distinct biological states. These methods produce large amounts of data, which is often used by unsupervised classification and clustering techniques to group samples with a similar molecular profile and/or molecules with a similar expression profile. However, there has been little attention paid on the assessment of significance on the results of such analyses.

In this paper we proposed an algorithm that uses the bootstrap technique to assign confidence to clustering results.

REFERENCES


