A Selective Voting Convex-Hull Ensemble Procedure for Personalized Medicine

Radhakrishnan Nagarajan, Ph.D.¹, Ralph L. Kodell, Ph.D.²
¹Division of Biomedical Informatics, University of Arkansas for Medical Sciences
²Department of Biostatistics, University of Arkansas for Medical Sciences

Abstract

Genes work in concert as a system as opposed to independent entities and mediate disease states. There has been considerable interest in understanding variations in molecular signatures between normal and disease states. However, a majority of techniques implicitly assume homogeneity between samples within a given group and use a fixed set of genes in discerning the groups. The proposed study overcomes these caveats by using a selective-voting convex-hull ensemble procedure that accommodates molecular heterogeneity within and between groups. The significance of the study is its potential to selectively retrieve sample-specific ensemble sets and investigate variations in the networks corresponding to the ensemble set across these samples. These characteristics fit well within the scope of personalized medicine and comparative effectiveness research that emphasize on patient-tailored interventions. While the results are demonstrated on colon cancer gene expression profiles the approach as such is generic and can be readily extended to other settings.

Introduction

Understanding the differences between normal and disease states from high-throughput molecular assays using suitable computational techniques is an area of active research. Several techniques have been proposed in literature to identify genes that change markedly between normal and disease groups from their high-throughput molecular signatures. A majority of these techniques implicitly assume homogeneity between the samples within a given group. Such homogeneity assumptions need not necessarily be true in light of inherent differences between the members within that group. More importantly, homogeneity assumptions can be violated across multiple scales. For instance, lack of homogeneity or heterogeneity at the molecular level can be an outcome of inherent stochasticity underlying transcriptional mechanism, microenvironment effects, plasticity, de-differentiation, crosstalk between pathways, co-expression of multiple differentiation programs. On the other hand, heterogeneity between patient populations can be attributed to genetic factors, co-morbidities, variations in disease progression and other clinical parameters. Therefore, accommodating heterogeneity between patients within a given group may be a critical step prior to developing patient-tailored interventions and falls well within the objectives of comparative effectiveness research and personalized medicine.

Popular techniques that discern molecular signatures between a given pair of groups include clustering and classification techniques. However, these techniques traditionally use the same fixed set of genes to distinguish between normal and disease groups. Many methods have also been proposed for combining the outputs of classifiers into an ensemble including those for combining class labels and continuous outputs. Alternative approaches include selecting ensemble members from among many models of the same type as well as different types. A comprehensive review of ensemble-based methods can be found elsewhere. Majority voting among ensemble members is a common approach for combining class labels to predict the class of an unknown sample. Two popular methods that employ majority voting are bagging and boosting. The Random Forest method is a popular ensemble classifier that uses bagging, while LogitBoost uses boosting. Recently, Classification by Ensembles from Random Partitions (CERP) designed specifically for tree-based classifiers has also been shown to be competitive and a suitable alternative to bagging and boosting. Other notable classification based methods on high-dimensional feature include the random subspace method, attribute bagging and genetic algorithm-based approaches. While all of these methods show high accuracy when classifying cases not included in the training set, they tend to follow the standard “one-size-fits-all” approach to predicting class membership for unknown samples, i.e. every member of the trained ensemble of classifiers votes on every test sample. While classification accuracy is used as a benchmark, developing suitable techniques that accommodate population heterogeneity may be critical for a more realistic representation, generalization and possible incorporation in clinical workflows. More importantly, such techniques are likely to yield patient specific ensemble sets as opposed to class specific ensemble sets.
Recently, Kodell et al.\textsuperscript{29} introduced a model-free convex-hull-based approach\textsuperscript{30} in conjunction with a majority voting strategy for ensemble classification and compared its performance to other popular classification techniques using established performance measures. The convex-hull procedure\textsuperscript{29} by very definition captures second-order interaction and the geometry of the predictor variables in the two-dimensional space without imposing any model assumptions. This essentially renders the approach model-free. Since the ensemble set consists of gene pairs, the procedure is robust to data sets with missing values. Missing values are common across high-throughput molecular assays such as microarrays and can be an outcome of non-specific binding, poor hybridization and other measurement artifacts. Missing value of a gene across a single sample usually results in dropping that particular gene from the entire analysis. While imputation techniques can prove to be suitable alternative, they predict the missing value under certain implicit statistical assumptions and can lead to bias when these assumptions are violated. For the convex-hull selective voting procedure proposed in this study, voting is done by pairs of genes. Given a set of $n$ genes, the number of possible pairs or voters is $n(n-1)/2$. Thus a missing value in a gene across a given patient would essentially imply dropping the pairs corresponding to the gene for that particular patient in the analysis and not removal of the gene from the entire analysis. The selective-voting procedure is especially appealing since it lends itself to adapt the ensemble set across the samples retrieving personalized molecular signatures.

Unlike Kodell et al.\textsuperscript{29}, the present study investigates the choice of a novel selective-voting strategy and convex-hull pruning procedure that relies on pairs whose interaction contributes significantly in discerning a given pair of groups while accommodating possible molecular heterogeneity within a group. A logistic regression pre-processing filter is used to first identify gene pairs with significant interaction term. Such an approach was inspired by traditional belief that associated genes are likely to have biologically relevant relationship (i.e. functional relationship)\textsuperscript{31,32}. Capturing genes with significant interaction can also be useful in constructing network abstractions similar to those of relevance networks\textsuperscript{32}. These pairs are subsequently mapped onto two-dimensional space for selective voting using convex hulls. A novel pruning procedure is proposed to minimize the overlap between the hulls corresponding to the control and the cancer samples. This procedure has to be contrasted to Kodell et al.\textsuperscript{29}, that used a distance metric and majority voting to assign the samples in the overlap region to the hulls. The pruning and selective voting process used in the present study results in sample-specific ensemble sets. Subsequently, network abstraction corresponding to the ensemble sets across each sample is generated. Since no explicit directionality information is available, these networks are in the form of undirected graphs. Several studies have successfully demonstrated the usefulness of network abstractions for obtaining preliminary system-level insights into biological pathways and signaling mechanisms\textsuperscript{9,31}. It is our belief that the network abstractions of the ensemble sets are likely to provide insights into the intricate wiring of the genes and variation in the topological properties across the ensemble sets. Such an understanding may be an important step prior to devising meaningful patient-tailored interventions and falls within the scope of personalized medicine and comparative effectiveness research.

**Methods**

**Gene Expression Data**

The microarray gene expression data set\textsuperscript{1} (http://microarray.princeton.edu/oncology/affydata/index.html) used in the present study is publicly available and consists of 62 colon tissue samples (40 samples being from cancerous colon tissue of patients with colon adenocarcinoma and 22 samples being from normal colon tissue (controls) taken from the 40 cancer patients). Alon et al.\textsuperscript{1} identified 2000 of the 6500 genes generated using Affymetrix oligonucleotide array to have high intensity levels across the 62 tissue samples. The present study focuses on this set of 2000 genes. The expression level of each of the 2000 genes was subsequently log\textsubscript{2}-transformed and normalized across the 62 samples by subtracting the mean and dividing by the standard deviation.

**Pre-processing Filter**

Microarray gene expression data sets are high-dimensional by their very design and provide an unbiased simultaneous screening of the transcriptional activity. However, several studies have reiterated the fact that only a fraction of the genes on these arrays are biologically meaningful. This aspect may possibly explain the widely observed positively-skewed distribution of the microarray gene expression profiles with a majority of the genes exhibiting low intensities. We use a logistic regression pre-processing filter in order to minimize noise, alleviate computational burden on the convex-hull selective voting process and select pairs whose interaction contribute significantly to the class prediction. As noted earlier, identifying gene pairs with significant interaction terms is expected to be useful in capturing possible functional relationships. The logistic regression model was chosen to be of the form ($Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \epsilon$) where $Y$ represents the class-prediction labels (22 control, 40 cancer) with $X_1$, $X_2$ and $X_1X_2$ representing the main and interaction effects of the given pair of genes. In the present
study, pairs with significant ($\alpha = 0.01$) interaction term (i.e. $\beta_2$) were deemed interesting and used for subsequent analysis. It is customary to incorporate multiple-testing correction to control for false-discovery rate (FDR) in the logistic regression procedure. However, in the present study imposing FDR (Benjamini-Hochberg) drastically reduced the number of pairs to zero. Therefore, we excluded the FDR correction from the pre-processing filter. The reduced set after pre-processing consisted of 3016 pairs of genes. All subsequent analysis and discussion were restricted to these 3016 pairs.

Convex-Hull Selective Voting Algorithm

**Step 1:** Store the expression profile of $m$ pairs of genes (i.e. $m$ potential voting members of the ensemble) across $n$ samples identified using the pre-processing filter in $X_{mxn}$. The $n$ samples are comprised of control ($n_{\text{con}}$) and cancer ($n_{\text{can}}$) samples such that $n = n_{\text{con}} + n_{\text{can}}$. Initialize the control and cancer voting matrices $v^{\text{con}}(r, s) \leftarrow 0$; $v^{\text{can}}(r, s) \leftarrow 0$, where $r = 1 \ldots 20$, $s = 1 \ldots 62$.

**Step 2:** Set $r \leftarrow 1$.
Choose a window length $k1$ and divide $X_{mxn}$ into test $U_{mx(k-1)}$ and training $V_{mk2}$ datasets such that $k2 = n \setminus k1$ and $k2 \gg k1$. The class labels of the $n$ samples in $X_{mxn}$ are stored in the vector $L_{ixn}$ where $L(k) = 1$ implies the $k^{th}$ sample is control.

Step 3: Consider the gene pair $(i, j) \in m, i \neq j$. Their training and test data sets are $(V_{i(k2)j2})$ and $(U_{i(k1)j1})$ respectively. Project the training data $(V_{i(k2)j2})$ onto a two-dimensional space and generate the convex-hulls $\Psi^{\text{con}}, \Psi^{\text{can}}$ corresponding to the control (con) and cancer (can) samples. If $\Psi^{\text{con}} \cap \Psi^{\text{can}} \neq \emptyset$ then the hulls overlap. In such a case, we iteratively prune the hulls till they’re separated using the procedure described below.

**Pruning:**

(i) Drop the vertices of the hulls in the overlap region and also their immediate neighbors.

(ii) Generate the pruned hulls ($\Psi^{\text{con}}_*, \Psi^{\text{can}}_*$).

(iii) Set \( \Psi^{\text{con}}_* = \Psi^{\text{con}}, \Psi^{\text{can}}_* = \Psi^{\text{can}} \).

If $\Psi^{\text{con}}_* \cap \Psi^{\text{can}}_* = \emptyset$, then go to Step 4.
Else go to (i).

**Step 4:** Vote the samples in the test data $(U_{i(k)j})$, $k = 1 \ldots k1$ as follows:

If $(U_{i(k)j}) \in \Psi^{\text{con}}$, then the control vote is incremented vote $v^{\text{con}}(r, k) \leftarrow v^{\text{con}}(r, k) + 1$.
If $(U_{i(k)j}) \in \Psi^{\text{can}}$, then the cancer vote is incremented vote $v^{\text{can}}(r, k) \leftarrow v^{\text{can}}(r, k) + 1$.

**Step 5:** Repeat Steps 3 and 4 for each pair of genes $(i, j) \in m, i \neq j$. The number of times each sample is voted as control and cancer is stored in $v^{\text{con}}$ and $v^{\text{can}}$ respectively.

**Step 6:** Repeat Steps 2-5 using the next window of $k1$ samples as test sample $U_{mx1}$ and the remaining ($k2 = n \setminus k1$) as training samples, $V_{mk2}$.

**Step 7:** Repeat Step 6 till each sample is voted on as a test case once.

**Step 8:** Repeat Steps 1-7, by randomly permuting the columns of $X_{mxn}$ 20 times (i.e. $r \leftarrow r + 1$, $r \leq 20$).

**Predicted labels ($L^*$):**

For the $k^{th}$, $k = 1 \ldots n$ sample, parametric $t$test ($\alpha = 0.01$) was used to determine the significant separation of $v^{\text{con}}[k]$ and $v^{\text{can}}[k]$.

$L^*(k) = 1$, (i.e. control sample) provided the separation between the means corresponding to $v^{\text{con}}[k]$ and $v^{\text{can}}[k]$ are statistically significant ($t$test, $\alpha = 0.01$) and $E(v^{\text{con}}[k]) > E(v^{\text{can}}[k])$ where $E$ represents the means. The ensemble set ($G^{\text{con}}$) of interest will contain those pairs of genes that contributed to $v^{\text{con}}$.

$L^*(k) = 2$, (i.e. cancer sample) provided the separation between the means corresponding to $v^{\text{con}}[k]$ and $v^{\text{can}}[k]$ are statistically significant ($t$test, $\alpha = 0.01$) and $E(v^{\text{con}}[k]) > E(v^{\text{can}}[k])$ where $E$ represents the means. The ensemble set ($G^{\text{can}}$) of interest will contain those pairs of genes that contributed to $v^{\text{can}}$.

$L^*(k) = 0$, provided the separation between the means of the corresponding to $v^{\text{con}}[k]$ and $v^{\text{can}}[k]$ is not statistically significant($t$test, $\alpha = 0.01$).

The $k^{th}$ sample is said to be correctly classified if $L^*(k) = L(k)$.

**Step 9:** Estimate the following statistics, Accuracy (ACC), Sensitivity (SEN), Specificity (SPC), Positive Predicted Value (PPV), Negative Predicted Value (NPV).
Given the small sample size of the dataset in the present study, we opted for a leave-one-out cross-validation (LOOCV) procedure, i.e. \( |k| = 1 \). Thus no repeats were necessary in Step 8 resulting in \( r = 1 \), Step 1). Voting was done by a direct comparison of \( v^{\text{con}} \) and \( v^{\text{can}} \) across the 62 samples (22 control, 40 cancer) as opposed to using a ttest since only a single realization was used \( r = 1 \). The control and cancer votes across the 62 samples were first determined (i.e. \( v^{\text{con}}(k), v^{\text{can}}(k), k = 1 \ldots 62 \)). If \( v^{\text{con}}(k) > v^{\text{can}}(k) \), then the \( k^{\text{th}} \) sample was labeled as a control sample. On the other hand, if \( v^{\text{con}}(k) < v^{\text{can}}(k) \), then the \( k^{\text{th}} \) sample was labeled as a cancer sample. No labels were assigned when \( v^{\text{con}}(k) = v^{\text{can}}(k) \), i.e. a tie. An example of the iterative pruning of the convex-hulls generated using a training set of 19 control and 36 cancer samples for a given pair of genes \((X, Y)\) is shown in Fig. 1 for clarity. Of interest is to note the significant overlap between the cancer and control training samples, Fig. 1a. However, the iterative pruning process minimizes the overlap till the cancer and control samples are completely separated in Fig. 1c. The overlay of the test samples (4 cancer, 3 control) on the pruned hulls, Fig. 1d, shows that ~50% of the cancer test samples and ~66% of the control test samples are correctly classified while the remaining are unclassified. Thus not every test sample is voted on by the gene pair \((X, Y)\).

**Figure 1.** Iterative pruning process for separating the convex-hulls corresponding to 19 control and the 36 cancer training samples using the given pairs of genes \((X, Y)\) is shown in (a-d). The vertices in the overlap region that are dropped in the pruning process are shown by the up-arrow (cancer) and down-arrow (control) in (a) and (b). The cancer and control training samples are shown by squares (white) and circles (white) respectively. Their corresponding hulls are shown by colored and white polygons in (a-d). The overlap between the hulls (a) are minimized by the iterative pruning process till the hulls are completely separated in (c). The four cancer and the three control test samples are shown by squares (dark) and circles (dark) respectively in (d).

**Network Abstraction of the Ensemble Sets**

As noted earlier, genes work in concert as a system and mediate phenotypic outcomes/disease states. Networks and graphs have proven to be convenient and useful abstractions of the underlying signaling mechanisms and pathways. Understanding their variation across the patient samples is one of the objectives of the proposed study and is likely to be useful in identifying candidate genes for possible intervention. The logistic regression pre-processing filter was used to identify pairs of genes whose interaction contribute significantly in discerning the control and cancer samples. Subsequently, we use the convex hull selective voting algorithm once \( r = 1 \) with the leave-one out procedure to identify pairs that vote across each of the samples. For the network abstraction (undirected graphs), we consider only pairs in the ensemble set (Step 8). For instance, if the given sample is a control sample, we consider only pairs corresponding to \( v^{\text{con}} \). Each gene is represented by a vertex in the undirected graph and the pairs of genes by an edge. Since we consider pairs with no explicit directional information, the resulting graph by definition is undirected. Such networks often exhibit interesting topological and statistical properties.
Figure 2. The number of gene pairs in the control \((\nu^{con})\) and cancer \((\nu^{can})\) ensemble sets for the 62 patient tissue samples (22 control and 40 cancer) obtained using the convex-hull ensemble selective voting technique (LOOCV) is shown in (a) and (b) respectively. The labeling of the samples is such that the control samples (1…22, solid line) precede the cancer samples (23…62, dotted line). The misclassified samples are shown by circles in (a-c). Six of the misclassified samples (control samples: 8, 12, 18; cancer samples: 55, 58, 59) were also reported as outliers in the original study on hierarchical clustering of the 62 samples using 20 most differentially expressed genes. The original labels of these six tissue samples were \((N8, N12, N34)\) and \((T33, T36, T37)\) (see the dendrogram in Fig. 4, Alon et al., 1999). Heatmap representing the overlap between pairs of 22 control ensemble sets \(G^{con}(k)\) \(k = 1 \ldots 22\) is shown in (c) (top triangle). A similar analysis representing the overlap between the 40 cancer ensemble sets \(G^{can}(k)\) \(k = 23 \ldots 62\) is also shown in (c) (bottom triangle). The magnitude of overlap is proportional to the grayscale intensities with brighter color representing larger overlap of ensemble sets between those particular pairs of samples. Dark streaks are observed across the control and well as cancer misclassified samples (circles) indicating that the ensemble sets for these samples may not share sufficient number of pairs in common with the others.

Results and Discussion

Convex-Hull Selective Voting

Prior to investigating the network of genes and their variation across the patient samples we chose to determine the performance of the proposed convex-hull selective voting approach in classifying the control and cancer sample using popular performance metrics for completeness. The Accuracy (ACC = 82.3), Sensitivity (SEN = 92.5), Specificity (SPC = 63.6), Positive Predictive Value (PPV = 82.2) and Negative Predictive Value (NPV = 82.3) estimated using the proposed approach were comparable to those we had estimated earlier using more traditional algorithms in an earlier study. While the proposed method is model-free, some of the established algorithms demand judicious choice of various parameters for improved performance. These parameter choices are often non-trivial and may be data dependent failing to generalize across new samples. These inherent discrepancies in the
nature of the approaches prevent us from conducting a direct comparison of the performance metrics. The control \((\nu^{\text{con}})\) and the cancer \((\nu^{\text{canc}})\) votes estimated across the 62 samples are shown in Fig. 2a and 2b respectively. Of interest is to note that the number of control samples misclassified were more than that of the cancer samples. On a related note, the number of control samples is considerably lower than that of the cancer samples. The control samples (22) were also obtained from the 40 cancer subjects as opposed to normal subjects. We believe these may contribute to the discrepancies in the misclassification between the controls and the cancer samples. As expected, for the correctly classified control samples we have \(\nu^{\text{con}}(k) > \nu^{\text{canc}}(k), k = 1 \ldots 22\), Figs. 2a-2b. A similar behavior was observed for correctly classified cancer samples i.e. \(\nu^{\text{canc}}(k) > \nu^{\text{con}}(k), k = 23 \ldots 62\) Fig. 2a-2b. It should be noted that Figs. 2a-2b also reflect variation in the number of pairs of genes across the control and the cancer ensemble sets. This in turn justifies the need for selective-voting strategies and reflects inherent molecular heterogeneity between the samples. In order to obtain further insight, we investigated the overlap in voting members of the 22 control ensemble sets \(G^{\text{con}}(k), k = 1 \ldots 22\) and the 40 cancer ensemble sets \(G^{\text{canc}}(k), k = 23 \ldots 62\). A grayscale heatmap was used to visualize the extent of overlap between the controls and between the cancer ensemble sets, Fig. 2c with lighter color representing larger overlap. Of interest is to note the characteristic dark streaks in heatmap corresponding to misclassified control and the cancer samples. The dark streaks reflect minimal overlap between the ensemble sets corresponding to the misclassified samples and rest. The dark streaks were especially pronounced across the control samples since the size of the control ensemble sets (Figs. 2a-2b) were relatively smaller than those of the cancer ensemble sets. There were also additional samples that were correctly classified but shared minimal overlap with the other samples within each group. These samples are accompanied by intermittent dark streaks in Fig. 2c.

**Figure 3.** Nodes in the dominant components across the 22 control (left) and 40 cancer (right) ensemble sets are shown in (a) and (b) respectively. The corresponding edge densities of these graphs are shown in (c) and (d) respectively. The misclassified control and cancer samples are represented by circles in the respective subplots.

**Network Abstraction of the Ensemble Sets**

Network abstractions were generated from the adjacency matrix corresponding to the pairs of genes for the 22 control \((G^{\text{con}}(k), k = 1 \ldots 22)\) and 40 cancer \((G^{\text{canc}}(k) = 1 \ldots 40)\) ensemble sets. For the control ensemble sets, nine genes had non-zero degree centrality across the 22 sets. These genes were (Hsa. 4689; Hsa. 5444; Hsa. 832; Hsa. 1130; Hsa. 996; Hsa. 821; Hsa. 831; Hsa.692; Hsa.3305). Out of these nines genes (Hsa. 5444; 40S Ribosomal Protein S24 (human)) was also identified as one of the co-regulated proteins in the ribosomal cluster in the original study (see Table 1, Alon et al., 1999). These genes were shown to be accompanied by relatively low intensities in
the control tissue as opposed to cancer tissue. The gene (Hsa. 3305; tropomyosin alpha chain smooth muscle (human)) was also reported in the original study. This gene was chosen as one of the 17 genes used to estimate the muscle index and distinguish the control and cancer specimens (see Fig. 4, Alon et al., 19991). The cancer ensemble sets failed to exhibit any overlap across the 40 sets. The absence of overlap in the case of cancer may possibly be due to increased heterogeneity in molecular signatures across diseased states as opposed to normal. However, based on the results from Fig. 2c, we did note that 29 of the 40 samples had a considerably large number of pairs of genes in their ensemble sets. Three genes that had non-zero degree centrality across these 29 ensemble sets were (Hsa. 10755; Hsa.692; Hsa. 122). Out of these three genes (Hsa. 1221; actin aortic smooth muscle (human)) was identified as one of the 17 smooth muscle genes used to compute the muscle index and subsequently distinguish the control and cancer tissues (see Fig. 4, Alon et al., 19991).

The connectivity of the graphs was investigated and the dominant components of each of the graphs corresponding to the ensemble sets were retrieved. The number of nodes in the dominant component across each of the graphs for the 22 control and 40 cancer ensemble sets is shown in Figs. 3a - 3b respectively. It should be noted that the number of nodes corresponding to the dominant component of the misclassified samples were considerably lower than others indicating sparse connectivity across the misclassified samples. This was true for the graphs corresponding to misclassified samples across control as well as cancer ensemble sets. Subsequently, the edge density of these dominant components was estimated. The graph density δ of an undirected graph G with V vertices (nodes) and E edges is given by the expression δ(G) = 2E/V(V − 1). In contrast to the number of nodes, the graph densities of the dominant component corresponding to the misclassified samples was considerably higher than correctly classified ones, Figs. 3c-3d. These results may be an outcome of the relatively small number of classifiers (Figs. 2a-2c) in the ensemble sets for the misclassified samples as opposed to others within the control and the cancer groups. Other established metrics did not provide any marked insights into the sample variations in the present study. However, these need to be investigated in greater detail.

Discussion

Traditional approaches for discerning normal and disease states from their molecular signatures implicitly assume homogeneity within groups. Such an assumption may be a serious limitation of these approaches as there can considerable differences between samples within a given group attributed to heterogeneity across multiple scales. There has been recent interest and emphasis on developing methodologies that can accommodate heterogeneity and individual variations that can lead to patient-tailored interventions. Such an approach falls within personalized medicine and comparative effectiveness research. The present study investigated the choice of logistic regression pre-filtering approach in conjunction with convex-hull ensemble selective voting strategy and network abstractions in understanding molecular heterogeneity in colon cancer gene expression profiles. While the pre-filtering is useful in identifying pairs of genes whose interaction contribute significantly to the prediction of the classes, the convex-hull selective voting is useful in identifying sample-specific marker genes. The convex-hull approach is model free and robust to missing values common in high-throughput molecular assays. The ability of this approach to capture second-order dependencies and geometry in the two-dimensional space makes it appealing. In contrast to more traditional classification approaches, the approach presented in this study lends itself for generating network abstractions of the ensemble sets and investigate their variation across the samples. Such networks have conventionally been useful in obtaining system-level insights and proven to be convenient abstractions of underlying signaling mechanisms. Their understanding may especially be necessary in developing patient-tailored interventions. The results presented on publicly available colon cancer gene expression profiles also validated some of the molecular markers identified in the original study. While the proposed approach is demonstrated on gene expression profiles, its generic nature lends itself to be adapted readily across other settings.

Acknowledgements

This research was supported by National Cancer Institute Grant 1R01CA152667.

References

1. Alon U et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences. 1999; 96:6745-6750.
2. Perou CM et al. Molecular portraits of human breast tumours. Nature. 2000; 406(6797):747-52.
3. Golub TR et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286:531-537.
4. Tibshirani R et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings National Academy of Science. 2002; 99:6567-6572.
5. Khan J et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine. 2001; 7:673-679.
6. Aubin et al. Heterogeniety of the osteoblast phenotype, Endocrinologist 9 (1999), pp. 25–31.
7. Candeliere et al (2001) G.A. Candeliere, F. Liu and J.E. Aubin, Individual osteoblasts in the developing calvaria express different gene repertoires, Bone. 2001; 28: 351–361.
8. Ross S E et al. Inhibition of adipogenesis by Wnt signaling. Science. 2000; 289:950–953.
9. Sachs K et al. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005; 308: 523–529.
10. Allinen M. et al. Molecular characterization of the tumor microenvironment in breast cancer. Cancer Cell. 2004; 6:17–32.
11. Elowitz MB et al. Stochastic Gene Expression in a Single Cell. Science. 2002; 297(5584):1183-6.
12. West M et al. Embracing the complexity of genomic data for personalized medicine. Genome Research. 2006; 16(5):559-66.
13. Common, J.E.A et al. Clinical and Genetic Heterogeneity of Erythrokeratoderma Variabilis. Journal of Investigative Dermatology 2005; 125: 920–927.
14. Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine. 2006; 6:21-45.
15. Zhou Z-H, Wu J and Tang W. Ensembling neural networks: Many could be better than all. Artificial Intelligence. 2002; 137:239-263.
16. Caruana R et al. Ensemble selection from libraries of models. Proceedings of the 21st International Conference on Machine Learning. New York: ACM (2004) 8 pp.
17. Rokach L. Collective-agreement-based pruning of ensembles. Computational Statistics and Data Analysis. 2009; 53:1015-1026.
18. Rokach L. Ensemble-based classifiers. Artificial Intelligence Review. 2010; 33:1-39.
19. Lam L and Suen CY. Application of majority voting to pattern recognition: an analysis of its behavior and performance. Systems, Man & Cybernetics. 1997; 27:553-568.
20. Breiman L. Bagging predictors. Machine Learning. 1996; 24:123-140.
21. Freund Y, Schapire RE. Experiments with a new boosting algorithm. Proceedings of the 13th International Conference on Machine Learning. 1996; 148-156.
22. Breiman L. Random forest. Machine Learning. 2001; 45:5-32.
23. Friedman J, Hastie T and Tibshirani R. Adaptive logistic regression: a statistical view of boosting. Annals of Statistics. 2000; 28:337-374.
24. Moon H et al. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. Artificial Intelligence in Medicine. 2007; 41:197-207.
25. Ahn H et al. Classification by ensembles from random partitions for high-dimensional data. Computational Statistics and Data Analysis. 2007; 51:6166-6179.
26. Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 832-844.
27. Bryll R, Gutierrez-Osuna R and Quek F. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognition 36 (2003) 1291-1302.
28. Rokach L. Genetic algorithm-based feature set partitioning for classification problems. Pattern Recognition 41 (2008) 1676-1700.
29. Kodell RL. et al. A model-free ensemble method for class prediction with application to biomedical decision making. Artificial Intelligence in Medicine. 2009; 46:267-276.
30. Barber CB. Dobkin DP. and Huhdanpaa HT. The Quickhull Algorithm for Convex Hulls. ACM Trans. Mathematical Software. 1996; 22(4): 469-483.
31. Gardner TS. et al. Inferring genetic networks and identifying compound mode of action via expression profiling. Science. 2003; 301:102-105.
32. Butte AJ. Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pacific Symposium on Biocomputing. 2000; 418-429.