



# A Composite Model for Subgroup Identification and Prediction via Bicluster Analysis

Hung-Chia Chen<sup>1,2</sup>, Wen Zou<sup>1</sup>, Tzu-Pin Lu<sup>1,3</sup>, James J. Chen<sup>1,2\*</sup>

**1** Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas, United States of America,

**2** Graduate Institute of Biostatistics and Biostatistics Center, China Medical University, Taichung, Taiwan, **3** Department of Public Health, Graduate Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan

## Abstract

**Background:** A major challenge in the analysis of large and complex biomedical data is to develop an approach for 1) identifying distinct subgroups in the sampled populations, 2) characterizing their relationships among subgroups, and 3) developing a prediction model to classify subgroup memberships of new samples by finding a set of predictors. Each subgroup can represent different pathogen serotypes of microorganisms, different tumor subtypes in cancer patients, or different genetic makeups of patients related to treatment response.

**Methods:** This paper proposes a composite model for subgroup identification and prediction using biclusters. A biclustering technique is first used to identify a set of biclusters from the sampled data. For each bicluster, a subgroup-specific binary classifier is built to determine if a particular sample is either inside or outside the bicluster. A composite model, which consists of all binary classifiers, is constructed to classify samples into several disjoint subgroups. The proposed composite model neither depends on any specific biclustering algorithm or patterns of biclusters, nor on any classification algorithms.

**Results:** The composite model was shown to have an overall accuracy of 97.4% for a synthetic dataset consisting of four subgroups. The model was applied to two datasets where the sample's subgroup memberships were known. The procedure showed 83.7% accuracy in discriminating lung cancer adenocarcinoma and squamous carcinoma subtypes, and was able to identify 5 serotypes and several subtypes with about 94% accuracy in a pathogen dataset.

**Conclusion:** The composite model presents a novel approach to developing a biclustering-based classification model from unlabeled sampled data. The proposed approach combines unsupervised biclustering and supervised classification techniques to classify samples into disjoint subgroups based on their associated attributes, such as genotypic factors, phenotypic outcomes, efficacy/safety measures, or responses to treatments. The procedure is useful for identification of unknown species or new biomarkers for targeted therapy.

**Citation:** Chen H-C, Zou W, Lu T-P, Chen JJ (2014) A Composite Model for Subgroup Identification and Prediction via Bicluster Analysis. PLoS ONE 9(10): e111318. doi:10.1371/journal.pone.0111318

**Editor:** Peter Csermely, Semmelweis University, Hungary

**Received** January 17, 2014; **Accepted** September 30, 2014; **Published** October 27, 2014

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

**Funding:** The research by Hung-Chia and Tzu-Pin was supported by the Post-doctoral Fellowship Program at the NCTR administered by the Oak Ridge Institute for Science and Education through an interagency agreement between USDOE and USFDA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* Email: jamesj.chen@fda.hhs.gov

## Introduction

Recent advances in biotechnology have generated great interest in the development of statistical methods and data mining techniques to analyze massive amounts of biological and medical data for understanding biological processes, discovering new species, or identifying new biomarkers for safety assessment, disease diagnostics and prognostics, and prediction of treatment response, etc. For example, metagenomics utilizes DNA sequence data to detect and identify representative species in environmental and clinically relevant samples and to discover genes or organisms with novel or useful functional properties [1–4].

In clinical treatment, patients are heterogeneous due to differences in genetic pre-dispositions, lifestyle, and disease characteristics. Personalized medicine utilizes genomic predictors of target patient population for assignment of more effective

therapies to ensure safety and avoid adverse events or unnecessary treatment [5,6]. A main goal is to develop a procedure that can classify patients into subgroups representing different disease characteristics or different responses to a specific treatment. For example, acute lymphoblastic leukemia (ALL) is a heterogeneous disease, including several subtypes (T-ALL, E2A-PBX1, BCR-ABL, TEL-AML1, MLL) differing in their response to chemotherapy [7–9]. Identifying important leukemia subtypes to accurately assign patients to specific risk/treatment groups is a difficult and expensive process, requiring the combined expertise of hematologist/oncologist, pathologist, and cytogeneticist [9].

In food safety surveillance, serotyping of pathogen strains is usually the first important step for identification and characterization of *Salmonella* isolates in outbreak investigations. However, standard methods for serotype identification of strains are tedious and time-consuming [10,11]. Considering there are over 2,500

outbreak strains of unknown or new serotypes, development of a procedure for early and fast screening and source tracking is essential. PFGE (pulsed-field gel electrophoresis) genotyping method has been used to investigate the relatedness of individual cases, and to confirm an outbreak of a disease and determine its possible source [10–13]. Previous works [10,11,14–16] reported that serotypes of *Salmonella* isolates could be deduced and predicted based on PFGE fingerprints. Thus, PFGE fingerprint profiling using data mining algorithms can potentially provide a possible alternative method for fast screening and identifying *Salmonella* serotypes.

In the aforementioned applications, the primary goal is to develop a class prediction model that can accurately identify population subgroups (cancer or strain subtypes) for new samples. There are three main aims: 1) classifying samples into distinct subgroups from large and complex unlabeled multivariate data, 2) characterizing the relationships among the subgroups identified, and 3) developing a prediction model to classify subgroup memberships of new samples by finding a set of predictor variables.

Classification is the standard approach to developing a model for class prediction of new samples. Classification is a supervised analysis, in which each sample has a predefined class label. A classification model builds a mathematical function for predicting class memberships of new unlabeled samples by learning the relationships between the class memberships of samples and their attributes from the sampled data [17–21]. The objective of this learning is to search for a prediction function and a least number of predictor variable that maximizes the probability of classification accuracy. In other words, a classification model utilizes class label information to optimize predictive accuracy. Without class labels, classification analysis is not viable for sample classification and prediction. Furthermore, standard classification algorithms are only applicable to the samples from the classes that are present in the sampled data. The algorithms are incapable of classifying the samples from classes other than those presented within the dataset, such as classifying new cancer subtypes in clinical medicine or new serotypes in pathogen identification.

Cluster analysis is the standard data mining technique for identification of structures and patterns in the data by partitioning samples into disjoint subgroups and finding their relationships. There are hierarchical and non-hierarchical clustering algorithms. The hierarchical algorithm clusters the objects into a tree-like dendrogram [22]. The hierarchical clustering method can provide the relationship among the samples or the clusters; however, it is inefficient for determining subgroups when the number of samples is large. The non-hierarchical clustering algorithms divide objects into a pre-specified number of groups; k-means [23] and self-organizing maps (SOM) [24] are two commonly known algorithms. Specification of the number of subgroups is a challenge when the number of subgroups is large.

Clustering techniques provide a global analysis of samples by partitioning samples with similar attributes in the same cluster. Each sample is assigned to one and only one cluster, based on all attributes. In many applications, such as gene expression experiments, functionally related genes may exhibit a similar pattern only in a subset of patients with certain medical conditions, not in all patients; also, some genes may involve more than one function or no function at all, and associate with more than one condition or no condition. A primary goal in these applications is to identify those subsets of co-expressed/co-regulated genes with associated subsets of samples with similar conditions. Cluster analysis cannot effectively identify the substructures between a subset of genes and a subset of samples. Biocluster analysis

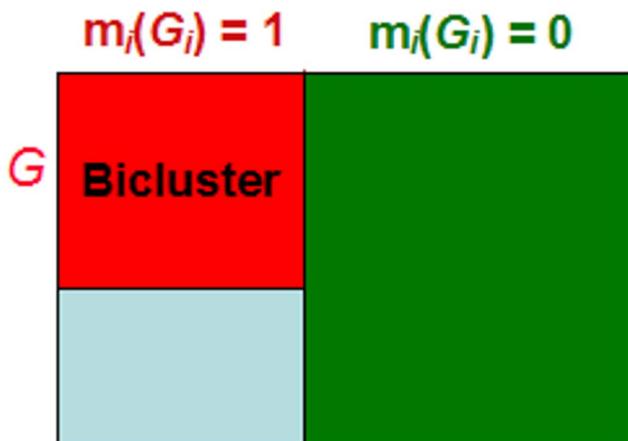
provides an approach to identify substructures in the sampled data. Bioclustering techniques identify biclusters by simultaneously clustering both samples and attributes [25–39]. Each bicluster is defined as a subset of attributes associated with a subset of samples. For an overview of bioclustering methods see the reviews of Madeira and Oliveira [28] and Kriegel et al. [32]. Alternatively, Baker et al. [40,41] developed GeneWeaver system aiming to integrate multiple data sources to identify associations between phenotypes and gene sets. The system was capable of demonstrating the clustered genes and phenotypes as hierarchical associations. Recently, Zhang et al. [42] further developed an approach to finding maximum bicliques in bipartite graphs, which was incorporated into the GeneWeaver system. Biocluster analysis can be viewed as an application of GeneWeaver to identify substructures in single study.

Both cluster and biocluster analyses are unsupervised analyses, in which samples do not have a predefined class label. These two methods are effective techniques for subgroup identification and characterization, but, are inefficient for subgroup prediction. Several supervised bioclustering procedures have been proposed for classification of labelled sample datasets [43–46]; these methods incorporate label information into the process of building biclusters. More discussion in the use of cluster/bicluster analysis for prediction and supervised bioclustering procedures are given in the Discussion section.

In this paper, we propose a composite modeling approach for subgroup identification and prediction via a biocluster analysis. The proposed approach combines an unsupervised bioclustering technique to identify potential sample subgroups in the first step, and a supervised classification technique to predict sample subgroup memberships in the second step. The proposed composite model neither depends on any specific bioclustering algorithm or patterns of biclusters, nor on any classification algorithms. Any bioclustering methods can be used in the first step of biocluster identification. This paper uses a SVD-based bioclustering algorithm to identify constant biclusters [39]; this method has been shown to perform well in extensive comparisons with various bioclustering methods, and found to be generally superior in terms of sensitivity and specificity. The primary focus of this paper is subgroup classification and prediction. Three well-known classification algorithms are considered in the second step of subgroup classification and prediction: support vector machine [17,18], random forests [19], and diagonal linear discriminant analysis [21]. The proposed composite model for subgroup identification and prediction is applied to a synthetic dataset and three real datasets for illustration.

## Methods

Consider a two-way data matrix with rows representing the measured attributes and columns representing samples. Many singular value decomposition (SVD) approaches for biocluster analysis of microarray data have been proposed and demonstrated to be effective [34–39]. In this paper, a SVD-based bioclustering method [39] was used to identify substructures between subsets of attributes and subsets of samples. An advantage of SVD-based bioclustering methods is that the bioclustering results do not depend on the random starting seeds. In the proposed approach, first a set of biclusters was identified using the SVD-based bioclustering method [39], followed by generating a set of binary classifiers, each built from one of the biclusters identified. A composite model is then developed to classify samples into disjoint subgroups described below.

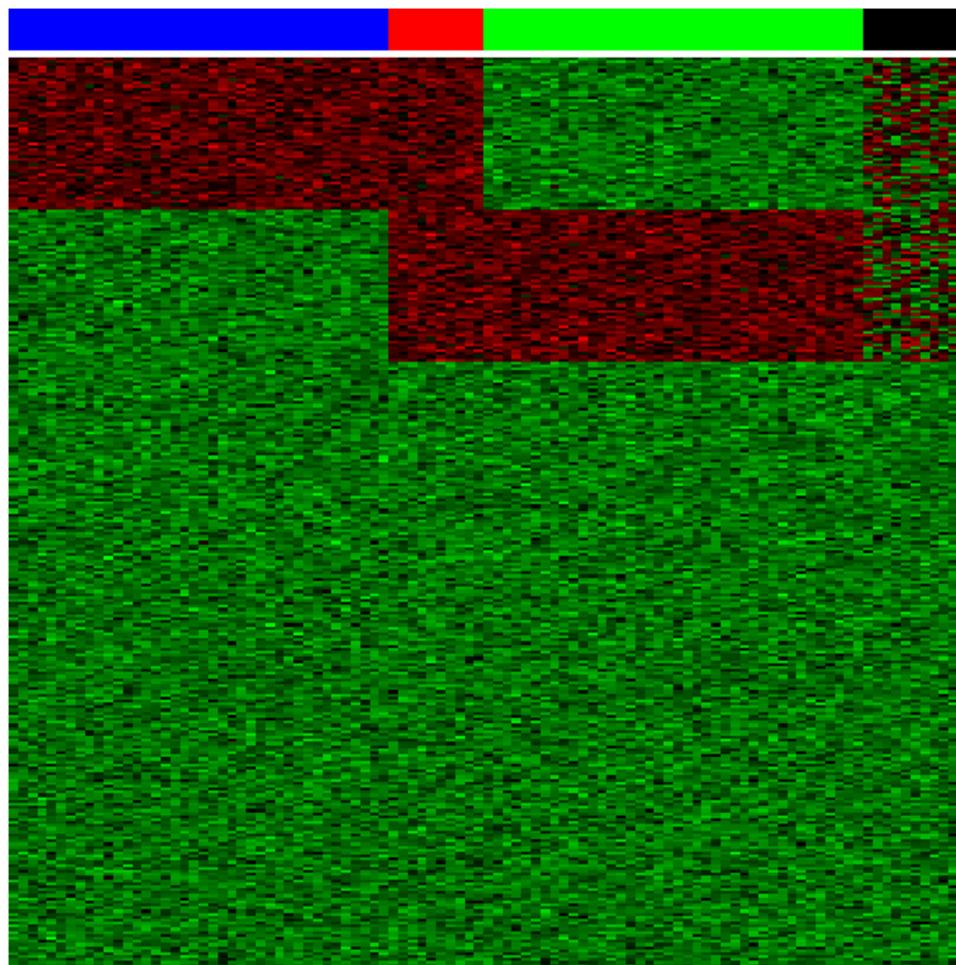


**Figure 1. Subgroup-specific binary classifier.** For each bicluster  $C_i = \{G_i, S_i\}$ , a subpopulation-specific genomic binary classifier  $m_i(G_i) = I\{s \in S_i\}$ , is built to predict where or not a sample is in the subpopulation  $S_i$ , where  $I$  is an indicator function.

doi:10.1371/journal.pone.0111318.g001

Denote the collection of biclusters identified as  $C = \{C_1, C_2, \dots, C_k\}$ . Each bicluster  $C_i$  consists of a subset of samples  $S_i$  that have similar attributes  $G_i$  ( $i = 1, \dots, k$ ). Thus, each  $S_i$  represents a subgroup in the sampled population. A subgroup-specific binary classifier  $m_i$  can be built to determine whether or not a sample  $s$  with the attribute  $\mathbf{g}$  is in the associated subgroup  $S_i$ , that is,  $m_i(\mathbf{g} | G_i) = I\{s \in S_i\}$ , where  $I$  is an indicator function (Figure 1). A composite classification model  $M$ , which consists of the collection of the binary classifiers  $M = \{m_1, \dots, m_k\}$ , is developed to partition samples into several disjoint subgroups described below.

For a given sample  $s$  with the attribute  $\mathbf{g}$ , each component binary classifier predicts whether or not the sample  $s$  belongs to its corresponding subgroup, where there are  $k$  predictive outcomes. Denote yes as “1” and no as “0”. Suppose the composite classification model consists of five binary classifiers ( $m_1, \dots, m_5$ ) with the corresponding subgroups ( $S_1, \dots, S_5$ ). For example, the outcome  $(1, 0, 0, 0, 0)$  of the composite model implies that the sample is in  $S_1$ ,  $(0, 0, 1, 1, 0)$  implies that the sample is in  $S_3$  and  $S_4$ , and  $(0, 0, 0, 0, 0)$  implies that the sample is not in any of the five subgroups. For  $k$  binary classifiers, there are  $2^k$  possible patterns of predictive outcomes. Each pattern represents a subgroup. However, when  $k$  is modest or large, many patterns would contain very few samples or no samples at all. When the number of patterns is



**Figure 2. A synthetic  $300 \times 100$  data matrix consists of two main bicluster regions with the size of  $50 \times 50$  having 10 overlapping columns.** The columns represent 100 samples consisting of 4 subgroups: S1 (columns 1–40, blue), S2 (columns 41–50, red), S3 (columns 51–90, green), and S4 (columns 91–100, black); the first 100 rows represent attributes: G1 (rows 1–50), G2 (rows 1–100), G3 (rows 51–100), and G4 (rows 1–100).

doi:10.1371/journal.pone.0111318.g002

**Table 1.** Upper panel, frequency distributions of classification patterns identified by the SVM composite model (m1, m2, m3, m4) for the synthetic training dataset consisting of 4 subgroups, S1, S2, S3, and S4; Lower panel, performance of the SVM composite prediction model for the test dataset of 1,000 simulated samples.

Subgroup Pattern	S1 (n = 40)	S2 (n = 10)	S3 (n = 40)	S4 (n = 10)	Total (n = 100)
<b>Training</b>					
0010	<b>40</b>	0	0	4	44
0100	0	0	<b>40</b>	2	42
1111	0	<b>10</b>	0	1	11
0110	0	0	0	1	1
0111	0	0	0	1	1
1011	0	0	0	1	1
Sensitivity	1	1	1	0	0.90
Specificity	0.93	0.99	0.97	1	0.98
<b>Test</b>					
Sensitivity	1.000	0.994	1.000	0.654	0.964
Specificity	0.968	0.997	0.963	0.999	0.990

Table values are the averages over 1,000 repetitions.

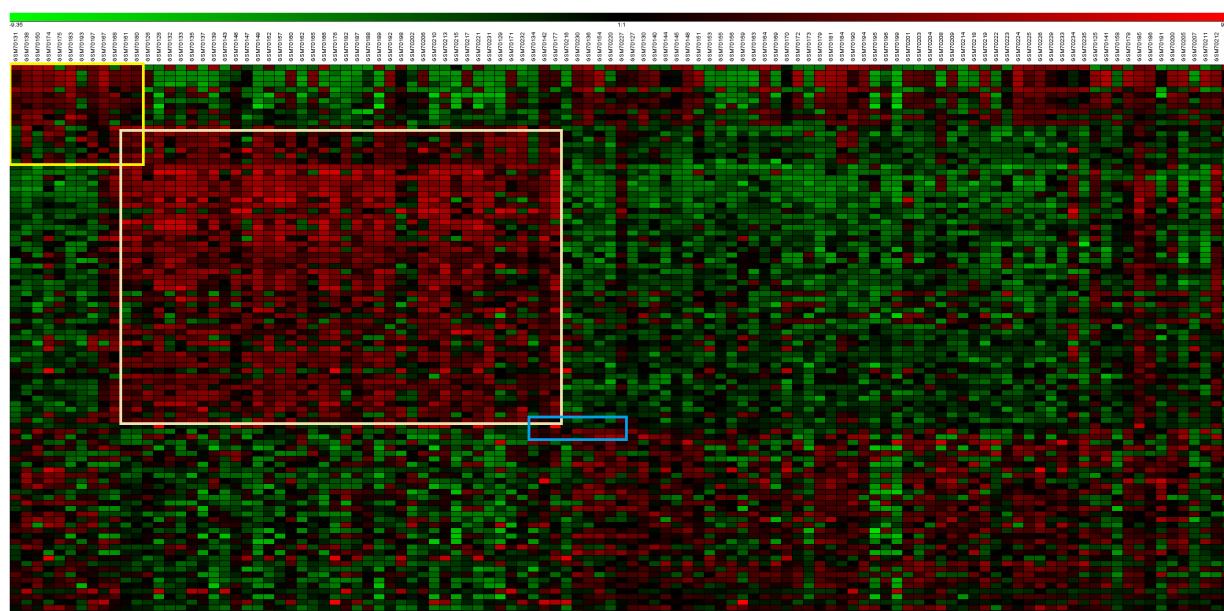
doi:10.1371/journal.pone.0111318.t001

large, a minimum of  $n^* = 5-10$  samples may be set as the criterion to form a (major) subgroup for further analysis. The patterns that contain less than  $n^*$  samples are referred to as minority subgroups.

Binary classifiers can be developed using any classification algorithms. This paper uses the three well-known algorithms: support vector machine (SVM) [17,18], random forests (RF) [19], and diagonal linear discriminant analysis (DLDA) [21]. These three algorithms were shown to perform well and have been the most popular classification algorithms for class prediction of high dimensional data [47].

In the development of a classification model, the most important consideration is to unbiasedly evaluate its “performance”. The common measures of performance are sensitivity (the proportion

of correct positive classifications out of the number of true positives), specificity (the proportion of correct negative classifications out of the number of true negatives), and accuracy (the total number of correct classifications out of the total number of samples). Procedures with both high sensitivity and high specificity will have high accuracy. To obtain unbiased estimates, the current sampled data are divided into a training set and a separate test set [48]; the training set is used for model development, and the test set is used for performance assessment. The split-sample and cross-validation methods are commonly used to evaluate performance of a classifier. The split-sample method randomly splits the data into two subsets from either the entire data or a designated test dataset. Split-sample validation is useful when the sample size is large.



**Figure 3. Lung Cancer data: three biclusters are identified, 55x40, 18x22 and 4x10, using top 100 genes.**  
doi:10.1371/journal.pone.0111318.g003

**Table 2.** Subgroup classification for the 111 lung cancer patients of the GSE3141 dataset using the composite model with the SVM, RF and LDA algorithms, and K-means (2-means, 3-means and 4-means) cluster analysis.

Methods	Subgroup pattern	Adenocarcinoma	Squamous cell carcinoma
SVM	000	39	6
	010	2	0
	100	12	52
RF	000	38	5
	010	2	1
	100	13	50
	110	0	2
DLDA	000	35	4
	010	7	2
	100	11	37
2-Means	110	0	15
	0	42	6
	1	11	52
3-Means	0	32	3
	1	14	3
	2	7	52
4-Means	0	33	4
	1	9	2
	2	6	22
	3	5	30
Total		53	58

doi:10.1371/journal.pone.0111318.t002

Cross validation involves repeatedly splitting the sampled data into a training set and test set to generate different training and test sample partitions to repeatedly estimate “accuracy” measures. Leave-one-out is a cross validation in which one sample is left out as a test set while all the other samples constitute the train set. The “accuracy” measures are estimated after all samples are tested. This paper uses both leave-one-out and split-sample for performance evaluation.

## Results

### Simulation Experiment

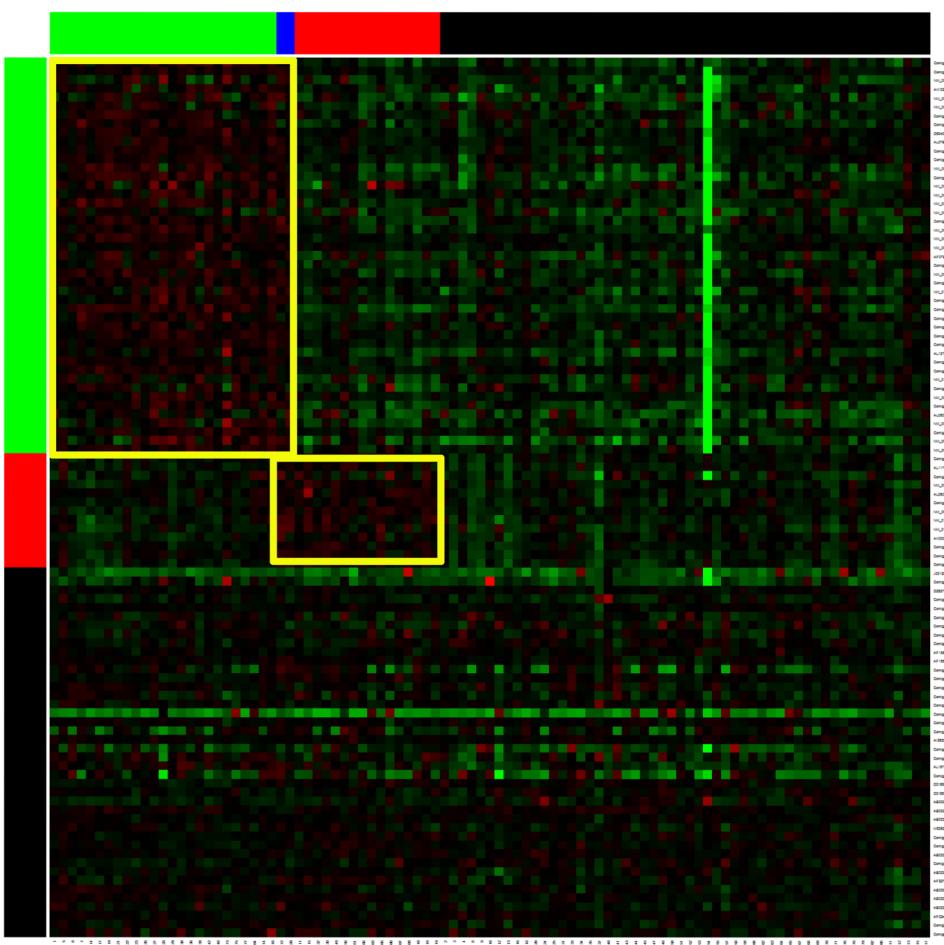
A simulation experiment was conducted to illustrate the proposed approach using asynthetic dataset of size 300 (rows)  $\times$  100

(columns). The dataset consisted of two main biocluster regions with the size of 50  $\times$  50 having 10 overlapping columns. The first main biocluster consisted of rows 1–50 and columns 1–50, and the second biocluster consisted of rows 51–100 and columns 41–90. The remaining columns 91–100 were in neither bioclusters. The biocluster (signal) data were generated from the normal distribution  $N(11, 1.22)$  and background data were generated from the normal random variable  $N(6, 1)$ . For masking purpose, random signals were also generated in the first 100 attributes for the last 10 samples. This dataset can be summarized as four bioclusters as follows. The columns represent 100 samples consisting of 4 subgroups: S1 (columns 1–40, blue), S2 (columns 41–50, red), S3 (columns 51–90, green), and S4 (columns 91–100, black); the first

**Table 3.** Subgroup classification for the 97 breast cancer patients (46 from patients who developed distant metastases within 5 years and 51 from patients who continued to be disease-free after a period of at least 5 years) using the SVM, RF, and DLDA composite models.

Subgroup pattern	SVM	Random Forest	DLDA
(0,0)	57	64	47
(0,1)	6	4	15
(1,0)	31	27	28
(1,1)	3	2	7
Logrank test for (0,1) vs (1,0) subgroups	0.284	0.519	0.599

doi:10.1371/journal.pone.0111318.t003



**Table 4.** Frequency distributions of classification patterns identified by the SVM composite model ( $m_1$ – $m_{10}$ ) for the *Salmonella* PFGE training dataset consisting of five serotypes.

13 Subgroups (n≥5)	4,5,12:i- n = 1113	Hadar n = 982	Oranienburg n = 997	Thompson n = 990	Typhi n = 972	Total n = 5054
0000000000	27	73	126	49	142	417
1000000000	653	0	0	0	0	653
1000000100	211	0	0	0	0	211
1000001000	212	0	0	0	0	212
0000001000	6	0	0	0	0	6
0100000000	0	0	0	0	829	829
0010000000	0	1	0	940	0	941
0001000000	0	0	34	0	0	34
0000100000	0	0	215	0	0	215
0001100000	0	0	593	0	0	593
0001110000	0	0	10	0	0	10
0000110000	0	0	9	0	0	9
0000010000	0	905	10	1	1	917
0000000100	2	0	0	0	0	2
0000011000	2	0	0	0	0	2
0010010000	0	3	0	0	0	3
Correct identification	1082	905	861	940	971	4759
Sensitivity	0.967	0.922	0.864	0.950	0.999	0.942
Specificity	1	0.997	1	1.000	0.932	0.982

Sixteen classification patterns are identified; 13 of the 16 have frequencies of at least 5 (last column). The last two rows show the sensitivity and specificity of the model performance.

doi:10.1371/journal.pone.0111318.t004

samples sample size was 1,000, the number of  $S_4$  samples was about 100 in each evaluation. Unlike the analysis of training samples, sufficient number of data from  $S_4$  was generated to form a subgroup and identified.

Tables S1 and S2 are the results from the RF and DLDA algorithms, respectively. The performances of the three algorithms are similar, in general. All three algorithms show high sensitivity and specificity in identifying (test) the  $S_1$  and  $S_3$  samples.  $S_2$  has the attributes across two subgroups  $S_1$  and  $S_3$ .  $S_4$  was designed to have indefinite attributes and difficult to be identified. The pattern corresponding to  $S_2$  is (1,1,1,1) using SVM and DLDA, and the pattern is (1,1,0,0) using RF. SVM appears to perform slightly better than RF and DLDA. For  $S_4$ , as expected, the sensitivity is low in all three algorithms.

#### Analysis of a lung cancer dataset

A public lung cancer microarray dataset was used to evaluate the performance of the proposed procedure and compare with k-means cluster analysis. The dataset was from a study (GSE3141) of using gene expression signatures to identify patterns of oncogenic pathway deregulation in lung cancer subtypes [49]. The original GSE3141 dataset was retrieved from the Gene Expression Omnibus [50]. The dataset consisted of 111 lung cancer samples with 53 adenocarcinoma (AD) and 58 squamous cell carcinoma (SQ) subtypes. This analysis was performed to distinguish these two lung cancer subtypes assuming no information on the sample subtypes. In the analysis, a quantile normalization algorithm was performed to remove the systematic biases. For each probe, standard error was calculated across all samples and ranked

decreasingly. The top 100 probes with the largest standard errors were selected as attribute variables.

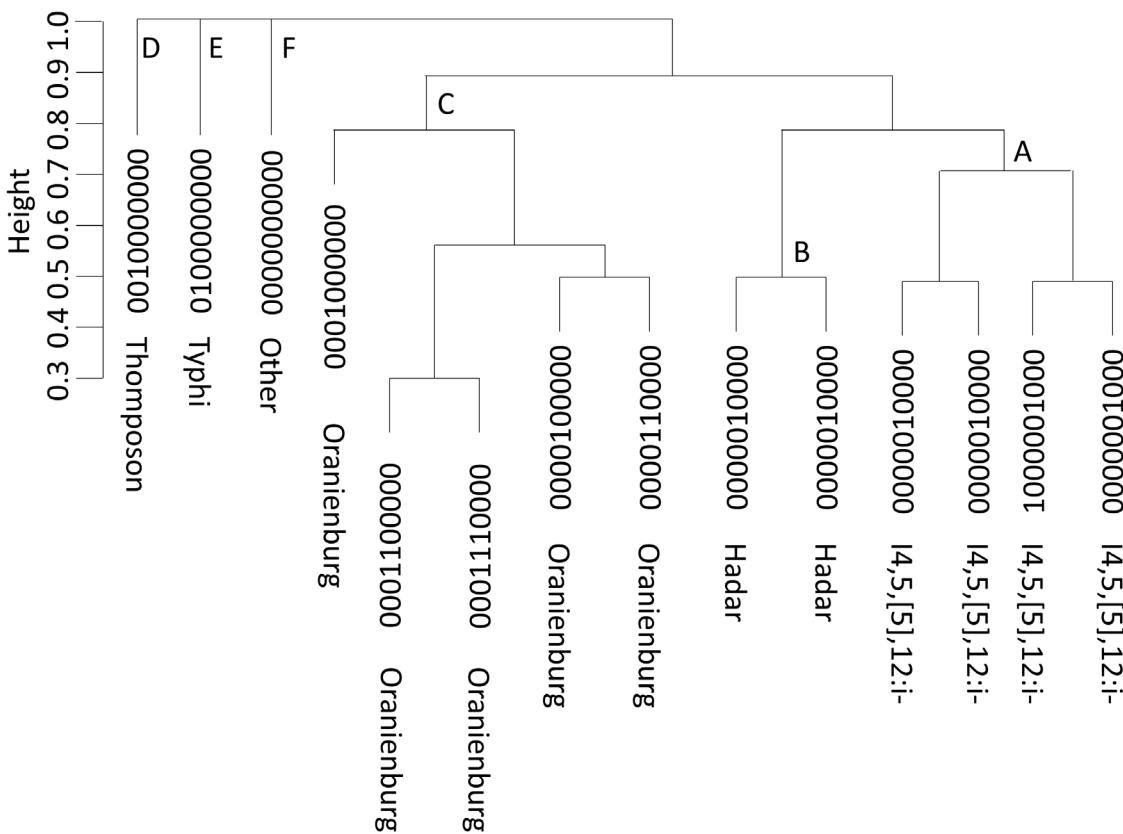
The proposed approach was performed on the data matrix of 100 genes by 111 samples. The biocluster analysis identified 32 bioclusters. A cutoff of at least 10 samples was used to eliminate small bioclusters, such as sizes of  $2 \times 2$  or  $2 \times 3$ , resulting in 3 clusters (Figure 3). The sizes of the three bioclusters were  $55 \times 40$ ,  $18 \times 22$ , and  $4 \times 10$ . A composite model  $M = \{m_1, m_2, m_3\}$  was built based on the three bioclusters. The leave-one-out cross (LOU) validation was used to classify each sample into one of the possible 8 subgroups.

Table 2 shows the results from the composite models and k-means methods for  $k = 2, 3, 4$ . Note that unlike the composite model using LOU, all 111 samples were used in the k-means analysis. The SVM algorithm identified three patterns (0,0,0), (0,1,0), and (1,0,0), while RF and DLDA identified four patterns (0,0,0), (0,1,0), (1,0,0), and (1,1,0). The classifier  $m_1$  generated from the biocluster  $C_1$  appears to be associated with the SQ subtype. Note that the classifier  $m_3$  by itself or in combination with  $m_1$  and  $m_2$  assigned none samples in a subgroup. That is, all samples, including 10 samples from  $C_3$  were not in  $C_3$ , as predicted by  $m_3$ . Based on the majority rule, SVM, RF, and DLDA correctly identified 41, 40, and 42 out of the 52 AD subtypes, respectively. All three algorithms identified 52 out of the 58 SQ subtypes. The performance between the composite models and 2-mean are generally similar. The 52 SQ subtypes identified by the 2-means and by the three composite models are identical. The 42 AD subtypes identified by 2-means contained those 41, 40, and all 42 ADs identified by the SVM, RF, and DLDA composite models,

**Table 5.** Frequency distributions of subgroup patterns identified by the SVM composite model for the *Salmonella* PFGE test dataset, which consisted of 5,055 isolates from five training serotypes and 1,000 additional “Decoy” isolates.

14 subgroups (n≥5)	A n=1156	B n=992	C n=930	D n=1047	E n=930	F n=1000	Total n=6055
000000000	38	73	122	56	133	747	1169
100000000	699	0	0	0	1	176	876
100000100	204	0	0	0	0	3	207
100001000	204	0	0	0	0	11	215
000001000	5	0	0	0	0	2	7
010000000	0	0	0	0	795	3	798
001000000	0	0	0	987	0	42	1029
000100000	0	0	24	0	0	0	24
000010000	0	0	216	0	0	1	217
000110000	0	0	544	0	0	0	544
000111000	0	0	8	0	0	0	8
000011000	0	0	6	0	0	0	6
0000010000	0	911	9	1	0	10	931
1000010000	0	8	0	0	0	1	9
Minority (10)	6	0	1	3	1	4	15
Correct	1112	919	798	987	795	747	5358
Sensitivity	.962	.926	.858	.943	.855	.747	.885
Specificity	.961	.996	1.00	.992	.999	.917	.977

The serotypes [4,15],12:i-, Hadar, Oranienburg, Thompson, Typhimurium, and Decoy were labeled as A, B, C, D, E, and F, respectively. n is the number of isolates in the serotypes. Fourteen of 24 identified classification patterns had frequencies at least 5. The last two rows show the sensitivity and specificity of the model performance.  
doi:10.1371/journal.pone.0111318.t005



**Figure 6. Hierarchical cluster analysis of the 14 subgroups identified from the test dataset using the average linkage distance.** The 14 subgroups consist of 5 major subgroups: 1. Thompson (001000000); 2. Typhimurium (010000000); 3. Decoy (000000000); 4. Oranienburg (000100000, 000010000, 000110000, 0001110000, 0000110000); 5. Hadar (0000010000, 1000010000) and 14,[5],12-i- (1000000000, 1000000100, 1000001000, 0000001000).

doi:10.1371/journal.pone.0111318.g006

respectively. Using  $n^* = 5$ , SVM identified two subgroups, including 39 AD and 52 SQ subtypes; RF identified two subgroups of 38 AD and 50 SQ subtypes; DLDA identified four subgroups with 42 AD and 52 SQ subtypes.

In the 4-means analysis, Groups 0 and 1 were from the split of Group 0 in the 2-means analysis, and Groups 2 and 3 were from the split of Group 1. However, the results of the 3-means analysis were peculiar. For example, there were 32, 14, and 7 adenocarcinomas for Groups 0, 1, and 2, respectively. Comparing with the 4-means analysis, the 32 consisted 21, 9, and 2 from Groups 0, 1, and 2, respectively; similarly, the 14 consisted of 12 and 3 from Groups 0 and 2, respectively.

## Analysis of the breast cancer dataset

The dataset of van't Veer et al. [51] contained 97 breast cancers (46 from patients who developed distant metastases within 5 years and 51 from patients who continued to be disease-free after a period of at least 5 years). The outcome was cancer-related survival time with 6391 genes as predictor variables.

Two biclusters with dimensions of  $45 \times 27$  and  $13 \times 18$  were identified from the 6391 genes and 97 patients (Figure 4). Two patients belonged to both biclusters; two binary classifiers,  $m_1$  and  $m_2$ , were developed. The leave-one-out cross validation analysis divided the 97 patient into 4 subgroups. Table 3 shows the results from the composite models. The  $m_1$  classifier identified low risk group patients and  $m_2$  identified high risk group patients. Figure 5 shows the plots of the survival time for four subgroups from the SVM model. Figures S1 and S2 are the plots from the RF and

DLDA composite models, respectively. The logrank tests for the differences between the two major subgroups (0,1) versus (1,0) were 0.284, 0.510, and 0.599 for SVM, RF, and DLDA, respectively.

## Analysis of the *Salmonella* isolate dataset

The *Salmonella* isolate dataset consisted of 45,924 PFGE isolates covering 32 mostly encountered serotypes published by Zou et al. [16]. The sample isolates were genotyped by the Pulsed-Field Gel Electrophoresis (PFGE) with DNA bands representing the presence and absence of a feature in a location as a fingerprint of isolates. Each isolate has 60 or 61 bands. Five serotypes, I4,[5],12:i, Hadar, Oranienburg, Thompson, and Typhimurium, were randomly selected for data analysis. Each serotype consisted of about 2,000 isolates. The analysis was to illustrate the use of the proposed composite model to identify the five serotypes and their subtypes, if any, and evaluate its performance as compared with the k-means clustering and SVM and RF classifications when the test set contained isolates from the serotypes that are not observed in the training set. The DLDA algorithm was not considered in this example since the PFGE fingerprints were binary features.

The data were first randomly divided into a training and a test dataset for each serotype. The bicluster analysis identified 10 biclusters and built 10 binary classifiers ( $m_1-m_{10}$ ) from the training dataset. The SVM and RF composite models then were applied to each training sample; 16 patterns were identified. The SVM model identified 16 patterns. Based on  $n^*=5$  as a cutoff, 13 subgroups were identified (Table 4). Note that the classifier  $m_9$

**Table 6.** Sensitivity (SN), specificity (SP), and accuracy of the composite model, k-means, and support vector machine (SVM) procedures for the PFGE test dataset.

		<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>ACC</b>
SVM Composite	SN	.962	.926	.858	.943	.855	.747	.885
	SP	.961	.96	1	.992	.99	.917	
RF	SN	.960	.927	.847	.945	.873	.750	.894
	SP	.961	.96	1	.992	.991	.921	
Composite	SN	.997	.976	.981	.973	.975	0	.819
	SP	.919	.955	.960	.974	.974	1	
5-means	SN	.960	.947	.932	.966	.917	.816	.924
	SP	.966	.96	.999	.999	.999	.947	
6-means	SN	.891	.941	.919	.976	.895	.941	.927
	SP	.989	.969	.999	.999	1	.924	
10-means	SN	.900	.942	.916	.954	.952	.944	.934
	SP	.990	.99	.999	.999	.999	.933	
15-means	SN	.999	.985	.998	.995	1	0	.831
	SP	.945	.975	.970	.966	.942	1	
SVM	SN	.999	.994	.998	.996	.999	0	.832
	SP	.941	.967	.954	.963	.974	1	

Serotypes A–F are defined in Table 4.  
doi:10.1371/journal.pone.0111318.t006

and  $m_{10}$  by itself or in combination with other classifiers assigned no samples into a subgroup. The interpretation (and presentation) of the performance was based on the known serotypes; in the analysis, the serotype was determined by majority rule. The sensitivities were from 86% to 99%, specificities were high at least 99%, except the Typhimurium with 93.2%. The overall accuracy was 94.2% and specificity was 98.2%.

The RF model identified 15 patterns, one less than the SVM model (Table S3). The difference is in the serotypes 4,5,12:i-identification. In the SVM classification, there were two patterns (1000000000) and (1000000100) with 653 and 211 for a total of 864 isolates, respectively. In the RF classification, there was no (1000000100) pattern, instead, the pattern (1000000000) consisted of 898 isolates. In addition to  $m_9$  and  $m_{10}$ ,  $m_8$  assigned no samples into a subgroup. Based on  $n^* = 5$  as a cutoff, 12 subgroups were identified. The sensitivities were from 86% to 99%, specificities were 93% to 100%. The overall sensitivity (accuracy) was 94.2% and specificity was 98.2%.

The SVM and RF composite models were applied to the test dataset, which included 1,000 additional samples (named “Decoy”) from the serotypes other than the five training serotypes. The analysis of the test dataset classification described below is for the SVM composite model, the results for the RF composite model are given in Table S4.

The SVM model identified 24 classification patterns from the 10 binary classifiers  $m_1$ – $m_{10}$ . Based on the  $n^* = 5$  as a cutoff, 14 subgroups were identified (Table 5), where 13 of the 14 were identical to the 13 subgroups that were identified in the training data. The additional subgroup consisted of 8 Hadar isolates. The serotypes and their associated binary classifiers were: 4,5,12:i-:  $m_1$ ,  $m_7$ , ( $m_1$ ,  $m_7$ ), ( $m_1$ ,  $m_8$ ); Hadar:  $m_6$ , ( $m_1$ ,  $m_6$ ); Oranienburg:  $m_4$ ,  $m_5$ , ( $m_4$ ,  $m_5$ ), ( $m_5$ ,  $m_6$ ), ( $m_4$ ,  $m_5$ ,  $m_6$ ); Thompson:  $m_3$ ; Typhimurium:  $m_2$ . The sensitivities between the training and test datasets were similar for the data of the five training serotypes. The overall specificity was lower since there were 1,000 additional “Decoy” isolates (Table 4 and Table 5). For the “Decoy” serotype, the sensitivity and specificity were 74.7% and 91.7%, respectively. The accuracies were 95.9% and 96.1% by excluding and including the “Decoy” isolates, in the calculation, respectively. The relationships among the 14 subgroups were further analyzed using the hierarchical cluster using the Euclidean distance function and the average agglomeration method (Figure 6). The 14 subgroups identified all 5 major serotypes and their subtypes, and the “Decoy” serotype: 1. Thompson (0010000000); 2. Typhimurium (0100000000); 3. Decoy (0000000000); 4. Oranienburg contained 5 subtypes (0001000000, 0000100000, 0001100000, 0001110000, 0000110000); 5. Hadar contained 2 subtypes (0000010000, 1000010000); 6. 14,[5],12:i- contained 4 subtypes (1000000000, 1000000100, 10000001000, 00000001000).

The PFGE test data were further analyzed using the k-means clustering to identify serotypes and their subtypes, and the SVM and RF algorithms to predict serotypes (including 1,000 Decoy isolates). Table 6 shows the sensitivity, specificity and accuracy of the three procedures. The k-means analysis was performed for  $k = 5$  to 15; only the results for  $k = 5$ , 6, 10, and 15 are presented. The k-means analysis was also based on majority rule to determine the serotype. The k-means’ performances were similar except for  $k = 5$ , in which the number of clusters were mis-specified. It appears that k-means has generally better performance than the composite model, except when a smaller  $k$  is specified. The SVM has much better performance than either the composite model or k-mean methods for the test dataset without Decoy data, the accuracy is more than 99%. The SVM is unable to predict the Decoy data since their serotypes are not in the training classes.

Currently, PFGE is routinely used molecular subtyping method by CDC (Centers for Disease Control and Prevention) and state health labs in the US for *Salmonella* surveillance and outbreak investigation [52], the ability to rapidly identify the serotype or a subtype of a *Salmonella* isolate is essential. The same serotype may have different subtypes, such as *Salmonella* Newport, and Dublin etc. These subtypes are closely related with their gene composition and variations. Current routine serotyping methods cannot provide sufficient information for subtype classification. The serotype subtype classification is important for the studies of genetic diversity and evolution. The composite model not only contributes to the PFGE-based characterization and surveillance of *Salmonella* isolates in outbreak investigations, also provides a better understanding of *Salmonella* genetic diversity and epidemiology.

## Discussion

Cluster analysis has been the primary data mining technique for dividing samples into disjoint subgroups where the samples in a cluster contain all attributes that characterize the cluster. Bicluster analysis techniques are being developed to identify which subsets of attributes are associated with which subsets of samples [34–39]. A bicluster analysis divides the samples into disjoint subgroups, where each sample in the subgroup corresponds to one or more subsets of attributes; and where there may be one additional subgroup formed by the samples not in any biclusters which are not associated with any subset of attributes. Both cluster analysis and bicluster analysis are powerful techniques for classifying samples into subgroups, but they are inefficient for prediction purpose. Either method can predict new samples by pooling the current samples with new samples then performing the same analysis. However, the subgroup membership of a current sample before and after the pooling may be different. Alternatively, either method may also assign the new sample using a classification algorithm such as k-NN (k-Nearest Neighbors) to develop a prediction model; note that k-NN requires specification of  $k$  and a distance measure between the new samples and the subgroups.

In the analysis of the lung cancer and PFGE datasets, Tables 4 and 5 show that k-means can outperform the proposed procedure when the number of clusters are correctly specified; however, it is often difficult to determine  $k$  when the sample size or the number of subgroups is large such as the PFGE data. Clustering analysis does not perform well if there is a subgroup of samples that are made of diverse subtypes, e.g., Decoy subgroup. The major advantages of the proposed procedure over k-means are: 1) it does not require pre-specifying the number of clusters, and 2) it uses a subset of attributes for each bicluster, instead of entire set of attributes, to develop a binary classifier. The composite model further identifies the relationships among subgroups based on their patterns of partition. Figure 6 clearly shows six distinct classes representing five serotypes and their sub-serotypes, and an unknown serotypes group. Finally, the hierarchical clustering tree can provide relationships among the clusters by a cutoff however, there seems to have no standard criterion or algorithm for choosing a cutoff; the cutoff is often made by visual inspection. When the number of samples and/or the number of clusters is large, such as the PFGE data, the visual inspection becomes infeasible.

Biclustering algorithms have been extended to supervised biclustering classification for labelled sampled data [43–46]. There are the CCC-biclustering algorithm to classify good versus poor responders [43], the co-clustering algorithm to discriminate between two sample classes (Class A versus Class B) [44], the

subspace co-expression analysis to discover differential co-expression patterns to classify normal versus cancer samples [45], and the LAS (large average submatrix) to classify five breast cancer subtypes [46]. These methods are supervised bioclustering-based classifiers (or class-discriminant bioclusters) [43], classification algorithms which were developed while optimizing the class discriminative ability from the label information. A two-class supervised bioclustering algorithm can be extended to a multiclass classification algorithm. However, classification algorithms are unable to characterize the subgroup relationships without further analysis. The composite model considers unlabeled data; the objectives are not only to classify samples into subgroups and predict new samples but also to characterize the relationships among subgroups. Recently, Geraci et al. [53] proposed “Butterfly”, a discrete dynamic system, for visualization, clustering, and classification of unlabeled data. Butterfly provided a 2D representation of the relationship between samples according to a set of variables. The system first generated a set of 2D cluster models, after performing a feature reduction step, and evaluated by binary classifiers, and finally showed the visual representation of the top classification models. On the other hand, the composite model is a general procedure applicable for two-class or multiclass prediction using bioclusters with or without feature reduction.

In the proposed approach, a binary classifier is developed to predict whether or not a sample is in the associated biocluster. For the samples that are assigned into two or more bioclusters, the composite model will separate those samples into a new subgroup. Some classifiers, either by itself or in combination with other classifiers, may assign only a small number samples, or none, into a subgroup. The PFGE analysis appeared to support some comments of Odibat and Reddy [44] that the bioclustering approach itself is inadequate for subgroup discrimination. The Oranienburg serotype consisted of at least 5 subtypes (Table 5). It would need three bioclusters,  $C_4$ ,  $C_5$ , and  $C_6$ , to identify (discriminate between) these subtypes. For example, the two bioclusters  $C_5$  and  $C_6$  in combination identified seven Oranienburg isolates. In addition, the biocluster  $C_9$  and  $C_{10}$  were not shown in any of the 14 patterns.

The composite model uses  $k$  bioclusters as a basis to generate up to  $2^k$  disjoint subgroups. Those small bioclusters are too small to be considered as representative subgroups for further partition. The composite model assigns each sample to one and only one subgroup, including those samples in the small bioclusters. In the lung cancer example, the composite model was composed of three binary classifiers from three “large” bioclusters of at least ten samples, out of the 32 bioclusters identified. These three binary classifiers could generate up to 8 subgroups. However, only three subgroup patterns were identified. The smallest subgroup (0,1,0) contained only two samples (Table 2). Similarly, in the breast cancer example, two “large” bioclusters were used. There were two small subgroups containing three and six samples. In the PFGE example, the composite model identified 16 subgroups based on 10 bioclusters in the training dataset (Table 4). The numbers of the samples in the three smallest subgroups were 2, 2, and 3. The model identified 24 patterns in the training dataset (Table 5). The total number of samples for 10 smallest subgroups combined was 15, less than 2 on the average. The composite model is capable of identifying small subgroups.

Specification of the threshold  $n^*$  can be based on the sample size and study objectives. For example, in personal medicine applications, patients are typically classified as high-risk versus low-risk or responders versus non-responders. The subgroups are identified for treatment recommendation. Different cancer subtypes or risk groups are subjected to different treatments. In the lung cancer

example, the treatments for the two subtypes are different. In the breast cancer example, patients in the high risk group would be recommended to more aggressive treatment. In both examples,  $n^*$  was set at ten. In the PFGE example, the primary objective was to develop a model to identify/predict serotypes/subtypes of unknown isolates. Knowing that there were many subtypes,  $n^*$  was set at five. Ten bioclusters were used to develop ten classifiers. The three small “subgroups” with sizes 2, 2, and 3 can be further investigated, if necessary.

In this paper, a minimum of five samples is recommended,  $n^* = 5$ . In the lung cancer example, three bioclusters with sample sizes of 40, 22, and 10 were used to generate subgroups. Cluster  $C_3$  consisted of 10 samples. As discussed, the prediction results by  $m_3$  were that all 10 samples were outside the  $C_3$  biocluster. These 10 samples were assigned primarily based on the classifiers  $m_1$  and  $m_2$ . In other words, bioclusters  $C_1$  and  $C_2$  were sufficient to develop the composite model in the sample assignment. In general, the samples from small bioclusters are likely to be assigned to some larger bioclusters. An explanation is that there are much more samples outside the biocluster region than the samples inside; a binary classifier tends to favor the majority class prediction in order to maximize total accuracy. Smaller bioclusters ( $n < 5$ ) can be used to develop a composite model. However, classifier developed by a small biocluster is likely to predict that the samples are outside the biocluster. This problem is known as class-imbalanced classification [54]. Furthermore, for large binary data matrix, there may be hundreds of  $2 \times 2$ ,  $2 \times 3$ ,  $3 \times 2$ , and  $3 \times 3$  bioclusters.

The notion of the composite modeling approach via bioclusters for class prediction is intuitive and straightforward. For a given biocluster, a sample is either inside or outside the biocluster. There are  $k$  predicted outcomes for each sample. Each predicted pattern represents a subgroup. In the simulation experiment, four bioclusters  $C_1$ – $C_4$  were identified. The sizes of  $C_1$ – $C_4$  were  $100 \times 16$ ,  $50 \times 51$ ,  $50 \times 58$ , and  $100 \times 15$ , respectively. Samples 1–50 and samples 41–90 were in bioclusters  $C_3$  and  $C_2$ , respectively; and samples 41–50 appeared in all four bioclusters  $C_1$ – $C_4$ . Table 1 shows that the composite model performed well in classification of the sample 1–90 since the four binary classifiers were developed based on the four bioclusters. Samples 91–100 were not in any of the four bioclusters, these samples are not associated with any subsets of attributes. In the PFGE data, there were 10 bioclusters with the sizes:  $8 \times 1097$ ,  $13 \times 813$ ,  $9 \times 938$ ,  $10 \times 596$ ,  $5 \times 787$ ,  $10 \times 938$ ,  $5 \times 175$ ,  $3 \times 178$ ,  $3 \times 468$ , and  $2 \times 109$ . There were many overlapping bioclusters. These bioclusters represented relative large numbers of samples with small numbers of attributes. On the other hand, in the lung data, the three bioclusters with the sizes of  $55 \times 40$ ,  $18 \times 22$ , and  $4 \times 10$  were smaller bioclusters relatively. In the simulation, lung cancer and PFGE examples, where the subgroups were known, the SVD-based bioclustering algorithm was able to capture the critical subgroup structures. The composite model appeared to perform reasonable well. In the proposed approach, any types of biocluster patterns and any bioclustering methods can be used to develop a composite model. However, the performance of a composite model highly depends on the bioclusters used to generate binary classifiers. A good bioclustering method is essential for the next step of subgroup classification and prediction.

The three classification algorithms, SVM, RF, and DLDA, are considered for the development of a composite prediction model. The SVM and RF have been the most popular and successful classification algorithms and applied to numerous areas of applications. These two algorithms can be applied to high dimensional data without feature selection. DLDA is a variant of the Fisher’s linear discriminant analysis. DLDA has been shown to be robust against imbalanced class size data [54], where the

numbers of samples in the biocluster and outside differs substantially. When the class sizes are imbalanced, the standard classification algorithms, such as SVM and RF, will favor majority class prediction resulting in poor performance. Among the three algorithms, SVM appears to perform consistently well.

Personalized medicine is the goal of much current research. A general aim is to identify a set of molecular biomarkers that can match disease of an individual patient with an optimal therapy. Several procedures have been proposed utilizing the classification and regression tree [19] for subgroup identification. These procedures partitioned the entire covariate space into subsets of patients that are homogeneous with respect to the set of covariates [55–58]. This paper proposes a composite prediction model as an alternative procedure to classify samples into subgroups according their associated attributes. Unlike the supervised classification tree approach, the proposed procedure is an unsupervised approach. The procedure provides an approach to classifying patients into subgroups of having different outcomes of interest, such as genotypic factors, phenotypic outcomes, efficacy/safety measures, or responses to treatments; the relationships among the subgroups identified can be further examined [59,60]. However, the approach presented does not consider outcome measures that are associated with specific drug treatment. In other words, the applications focus on the prognostic model, not predictive model, in the context of personalized medicine [48].

## Supporting Information

**Figure S1** The prediction model divided the 97 patients into four subgroups using RF. The logrank test for differences among the four subgroups (0,0), (0,1), (1,0), and (1,1) was 0.717. (TIF)

**Figure S2** The prediction model divided the 97 patients into four subgroups using DLDA. The logrank test for differences among the four subgroups (0,0), (0,1), (1,0), and (1,1) was 0.186. (TIF)

**Table S1** Upper panel. Frequency distributions of classification patterns identified by the RF composite model (m1, m2, m3, m4) for the synthetic training dataset consisting of 4 subgroups, S1, S2, S3, and S4. Lower panel. Performance of the RF composite

## References

- Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* 14: 303–310.
- Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 38: 525–552.
- Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, et al. (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol* 6: R78.
- Li J, Nelson K, McWhorter AC, Whittam TS, Selander RK (1994) Recombinational basis of serovar diversity in *Salmonella* enterica. *Proc Natl Acad Sci U S A* 91: 2552–2556.
- Langreth R, Waldholz M (1999) New era of personalized medicine: targeting drugs for each unique genetic profile. *Oncologist* 4: 426–427.
- Rothenberg ML, Carbone DP, Johnson DH (2003) Improving the evaluation of new cancer treatments: challenges and opportunities. *Nat Rev Cancer* 3: 303–309.
- Pui CH, Evans WE (1999) Acute lymphoblastic leukemia in infants. *J Clin Oncol* 17: 438–440.
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537.
- Ycoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 1: 133–143.
- Liebana E, Guns D, Garcia-Migura L, Woodward MJ, Clifton-Hadley FA, et al. (2001) Molecular typing of *Salmonella* serotypes prevalent in animals in England: assessment of methodology. *J Clin Microbiol* 39: 3609–3616.
- Gaul SB, Wedel S, Erdman MM, Harris DL, Harris IT, et al. (2007) Use of pulsed-field gel electrophoresis of conserved XbaI fragments for identification of swine *Salmonella* serotypes. *J Clin Microbiol* 45: 472–476.
- Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, et al. (1995) Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* 33: 2233–2239.
- Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV (2001) PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg Infect Dis* 7: 382–389.
- Zou W, Lin WJ, Hisz KB, Chen HC, Keys C, et al. (2012) Prediction system for rapid identification of *Salmonella* serotypes based on pulsed-field gel electrophoresis fingerprints. *J Clin Microbiol* 50: 1524–1532.
- Zou W, Lin WJ, Foley SL, Chen CH, Nayak R, et al. (2010) Evaluation of pulsed-field gel electrophoresis profiles for identification of *Salmonella* serotypes. *J Clin Microbiol* 48: 3122–3126.
- Zou W, Chen HC, Hisz KB, Tang H, Foley SL, et al. (2013) Meta-analysis of pulsed-field gel electrophoresis fingerprints based on a constructed *Salmonella* database. *PLoS One* 8: e59224.
- Vapnik V (1998) Statistical learning theory: Wiley, New York.
- Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Machine Learning* 46: 389–422.
- Brieman L, Friedman J, Olshen R, Stone C, Steinberg D, et al. (1995) CART: Classification and Regression Trees: Stanford, CA.
- Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Springer.

prediction model for the test dataset of 1,000 simulated samples. Table values are the averages over 1,000 repetitions. (DOC)

**Table S2** Upper panel. Frequency distributions of classification patterns identified by the DLDA composite model (m1, m2, m3, m4) for the synthetic training dataset consisting of 4 subgroups, S1, S2, S3, and S4. Lower panel. Performance of the DLDA composite prediction model for the test dataset of 1,000 simulated samples. Table values are the averages over 1,000 repetitions. (DOC)

**Table S3** Frequency distributions of classification patterns identified by the RF composite model (m<sub>1</sub>–m<sub>10</sub>) for the *Salmonella* PFGE training dataset consisting of five serotypes. Sixteen classification patterns are identified; 13 of the 16 have frequencies of at least 5 (last column). The last two rows show the sensitivity and specificity of the model performance. (DOC)

**Table S4** Frequency distributions of subgroup patterns identified by the RF composite model for the *Salmonella* PFGE test dataset, which consisted of 5,055 isolates from five training serotypes and 1,000 additional “Decoy” isolates. The serotypes I4,[5],12:i, Hadar, Oranienburg, Thompson, Typhimurium, and Decoy were labeled as A, B, C, D, E, and F, respectively. n is the number of isolates in the serotypes. Fourteen of 24 identified classification patterns had frequencies at least 5. The last two rows show the sensitivity and specificity of the model performance. (DOC)

## Acknowledgments

The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration. The authors are grateful to reviewers for much helpful comments and suggestions for revising and improving this paper.

## Author Contributions

Conceived and designed the experiments: JJC HCC. Performed the experiments: HCC. Analyzed the data: HCC TPL. Contributed reagents/materials/analysis tools: WZ. Wrote the paper: JJC HCC TPL.

21. Dudoit S, Fridlyand J, Speed T (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 97: 77–87.
22. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95: 14863–14868.
23. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* 22: 281–285.
24. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 96: 2907–2912.
25. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterrecker A, et al. (2010) FABIA: factor analysis for bicluster acquisition. *Bioinformatics* 26: 1520–1527.
26. Cheng Y, Church GM (2000) Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* 8: 93–103.
27. Tanay A, Sharan R, Shamir R (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 18 Suppl 1: S136–144.
28. Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform* 1: 24–45.
29. Jiang D, Tang C, Zhang A (2004) Cluster analysis for gene expression data: A Survey. *IEEE Transaction on Knowledge and Data Engineering* 16: 1370–1386.
30. Barkow S, Bleuler S, Prelic A, Zimmermann P, Zitzler E (2006) BiCAT: a biclustering analysis toolbox. *Bioinformatics* 22: 1282–1283.
31. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhmann P, et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22: 1122–1129.
32. Kriegel H-P, Kroger P, Zimek A (2009) Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transaction on Knowledge Discovery from Data* 9: 1–58.
33. Tagkopoulos I, Slavov N, Kung S (2005) Multi-class biclustering and classification based on modeling of gene regulatory networks. In Fifth IEEE Symposium on Bioinformatics and Bioengineering: 89–97.
34. Kluger Y, Basri R, Chang JT, Gerstein M (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res* 13: 703–716.
35. Yang WH, Dai DQ, Yan H (2007) Biclustering of microarray data based on singular value decomposition. *Proc 2007 Int Conf Emerging technologies in knowledge discovery and data mining*: 194–205.
36. Busygin S, Prokopyev O, Pardalos PM (2008) Biclustering in data mining. *Comput Oper Res* 35: 2964–2987.
37. Lee M, Shen H, Huang JZ, Marron JS (2010) Biclustering via sparse singular value decomposition. *Biometrics* 66: 1087–1095.
38. Sill M, Kaiser S, Benner A, Kopp-Schneider A (2011) Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics* 27: 2089–2097.
39. Chen HC, Zou W, Tien YJ, Chen JJ (2013) Identification of bicluster regions in a binary matrix and its applications. *PLoS One* 8: e71680.
40. Baker EJ, Jay JJ, Bubier JA, Langston MA, Chesler EJ (2012) GeneWeaver: a web-based system for integrative functional genomics. *Nucleic Acids Res* 40: D1067–1076.
41. Baker EJ, Jay JJ, Philip VM, Zhang Y, Li Z, et al. (2009) Ontological Discovery Environment: a system for integrating gene-phenotype associations. *Genomics* 94: 377–387.
42. Zhang Y, Phillips CA, Rogers GL, Baker EJ, Chesler EJ, et al. (2014) On finding biclusters in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. *BMC Bioinformatics* 15: 110.
43. Carreiro AV, Anunciacao O, Carrico JA, Madeira SC (2011) Prognostic prediction through biclustering-based classification of clinical gene expression time series. *J Integr Bioinform* 8: 175.
44. Odibat M, Reddy CK (2013) Efficient mining of discriminative co-clusters from gene expression data. *Knowl Inf Syst* DOI:101007/s10115-013-0684-0.
45. Fang G, Kuang R, Pandey G, Steinbach M, Myers CL, et al. (2010) Subspace differential coexpression analysis: problem definition and a general approach. *Pac Symp Biocomput*: 145–156.
46. Shabalin AA, Weigman VJ, Perou CM, Nobel AB (2009) Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics* 3: 985–1012.
47. Baek S, Tsai CA, Chen JJ (2009) Development of biomarker classifiers from high-dimensional data. *Brief Bioinform* 10: 537–546.
48. Chen JJ, Lin WJ, Chen HC (2013) Pharmacogenomic biomarkers for personalized medicine. *Pharmacogenomics* 14: 969–980.
49. Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439: 353–357.
50. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30: 207–210.
51. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415: 530–536.
52. Barrett TJ, Gerner-Smidt P, Swaminathan B (2006) Interpretation of pulsed-field gel electrophoresis patterns in foodborne disease investigations and surveillance. *Foodborne Pathog Dis* 3: 20–31.
53. Geraci J, Dharsee M, Nuin P, Haslehurst A, Koti M, et al. (2014) Exploring high dimensional data with Butterfly: a novel classification algorithm based on discrete dynamical systems. *Bioinformatics* 30: 712–718.
54. Lin WJ, Chen JJ (2013) Class-imbalanced classifiers for high-dimensional data. *Brief Bioinform* 14: 13–26.
55. Foster JC, Taylor JM, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. *Stat Med* 30: 2867–2880.
56. Su X, Tsai C-L, Wang H, Nickerson DM, Bogong L (2009) Subgroup Analysis via Recursive Partitioning. *The Journal of Machine Learning Research* 10: 141–158.
57. Ruberg SJ, Chen L, Wang Y (2010) The mean does not mean as much anymore: finding subgroups for tailored therapeutics. *Clinical Trials* 7: 574–583.
58. Lipkovich I, Dmitrienko A, Denne J, Enas G (2011) Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med* 30: 2601–2621.
59. Wang SJ, O'Neill RT, Hung HJ (2010) Statistical considerations in evaluating pharmacogenomics-based clinical effect for confirmatory trials. *Clin Trials* 7: 525–536.
60. Chen DT, Hsu YL, Fulp WJ, Coppola D, Haura EB, et al. (2011) Prognostic and predictive value of a malignancy-risk gene signature in early-stage non-small cell lung cancer. *J Natl Cancer Inst* 103: 1859–1870.