

Heritable Clustering Algorithms for Recapturing Epigenetic Progression in Breast Cancer

Zailong Wang¹, Pearly Yan^{2,4}, Dustin Potter^{1,2,4}, Charis Eng^{2,3,4}, Tim H. Huang^{2,4} and Shili Lin^{1,5*}

¹Mathematical Biosciences Institute, The Ohio State University, 231 W. 18th Avenue; ²Department of Molecular Virology, Immunology, and Medical Genetics, ³Division of Human Genetics, Department of Internal Medicine, ⁴Human Cancer Genetics Program, Comprehensive Cancer Center, The Ohio State University, 420 W. 12th Avenue; ⁵Department of Statistics, The Ohio State University, 1598 Neil Avenue, Columbus, OH 43210, USA

*Author and address for correspondence:

Shili Lin, PhD
Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210-1247
USA
Tel: (614) 292-7404
Fax: (614) 292-2096
Email: shili@stat.ohio-state.edu

Running Head: Recapitulating tumor progression pathways

Abstract

Motivation: In order to recapitulate tumor progression pathways using epigenetic data, we developed novel clustering and pathway reconstruction algorithms, collectively referred to as heritable clustering. This approach generates a progression model of altered DNA methylation from tumor tissues diagnosed at different developmental stages. The samples act as surrogates for natural progression in breast cancer and ideally allow the algorithm to uncover distinct epigenotypes that describe the molecular events underlying this process. Furthermore, our likelihood-based clustering algorithm has great flexibility, allowing for incomplete epigenotype or clinical phenotype data and also permitting dependencies among variables.

Results: Using this heritable clustering approach, we analyzed methylation data obtained from 86 primary breast cancers to recapitulate pathways of breast tumor progression. Detailed annotation and interpretation are provided to the optimal pathway recapitulated. Our results indicate that the proposed heritable clustering algorithms are a useful tool for methylation profiling and for stratifying clinical variables of breast cancer.

Availability: The heritable clustering program (Matlab codes) and supplementary information can be accessed at <http://www.stat.ohio-state.edu/~statgen/Pathway.html>.

Contact: shili@stat.ohio-state.edu

1 Introduction

Recapitulating pathways of tumor progression by tracing specific molecular lesions is necessary for understanding the disease and for developing novel drug targets and therapies. The idea of utilizing DNA methylation profiles to recapitulate tumor progression is even more enticing in that these epigenetic marks are stable and heritable in tumor genomes [2]. Specifically, this event occurs by the addition of a methyl group to a cytosine residue of a CpG dinucleotide [19]. It is recognized that in the normal genome, DNA methylation plays a role in mammalian development, imprinting, and X chromosome inactivation [20]. Recent advances further highlight a critical role of epigenetically mediated gene silencing in tumorigenesis [2]. Unmethylated CpG islands, located in the promoter regions of tumor suppressor/gatekeeper genes, become densely methylated during tumorigenesis.[16, 7, 13]. Once the de novo methylation takes place, this new mark is maintained in subsequent cycles of cell replication by DNA methyltransferases and other associated proteins, like polycomb repressors [21, 8]. The consequence of these molecular events is a gradual accumulation of DNA methylation in an affected promoter CpG island. In addition, methylation-associated silencing of tumor suppressor genes can result in cells with a growth advantage. The number of hypermethylated genes tends to increase in more malignant cells, and clonal expansion of proliferating cells may generate specific tumor types marked by their unique epigenetic signatures [16, 13]. This epigenetic event is inherently stable, and the silencing information is stored in the DNA methylation code of a tumor. Therefore, DNA methylation analysis can be retrospectively performed on clinical samples, allowing for studies of tumor progression history and for clinicopathological correlation.

With the implementation of the state-of-the-art microarray technologies, it is now possible to obtain methylation signatures of multiple genes simultaneously and to classify tumors based on their global methylation patterns [5, 18, 17, 24]. The idea of conducting a human epigenome project has recently been conceptualized [14] and is expected to facilitate our fundamental understanding of aberrant epigenetic mechanisms in cancer. This type of large-scale research provides an unprecedented challenge to develop novel statistical tools for analyzing a gargantuan amount of data. In this study, we developed novel clustering

and pathway reconstruction algorithms, collectively called heritable clustering, to evaluate a set of methylation microarray data previously generated in our laboratory [4]. Progressive accumulation of hypermethylated CpG islands was used to characterize breast tumor progression pathways. Utilizing these novel algorithms, we correlated specific methylation profiles with patients clinical phenotypes and reconstructed the epigenetic history germane to tumorigenesis.

2 Tumor progression pathways and Recapitulation

Tumor progression pathways are constructed based on the following characteristics: (1) most CpG islands are unmethylated in normal cells, (2) CpG island hypermethylation is heritable in tumor cells, and (3) multiple methylated loci are progressively accumulated during tumorigenesis. Based on these properties, we hypothesized that tumor cells have unique epigenetic signatures that are associated with specific cancer subtypes (phenotypic information). Specifically, we seek to construct patterns and relationships among hypermethylated genes that are progressively accumulated during tumorigenesis. As it is not possible for us to obtain tissues from the same patients at different stages of tumor progression over time, methylation data derived from tumors of different patients are used as surrogates for reconstructing tumor progression history.

To accommodate the heritable nature of de novo methylation, a progression tree ought to adhere to the notion that the hypermethylated loci acquired at each node are passed on to its progeny node(s). As such, hypermethylated loci are progressively accumulated; and the parental methylated loci are subsets of their progeny's. Furthermore, the phenotypes of progeny nodes are hypothesized to be more aggressive than the parents. Although existing clustering algorithms (e.g., hierarchical clustering or K-means) are available for clustering samples, no suitable method can be applied to give temporal directions of progression among different epigenetic clusters. Such a challenge impedes us from adopting published clustering algorithms without major modifications (for an indepth review and comparison of the clustering methods most widely used for analyzing microarray data, see [6]). Therefore, we developed the heritable clustering algorithms to identify and organize clusters into a tree

and to recreate tumor progression pathways.

3 Heritable Clustering

Three stages of heritable clustering are laid out here. First, we determined the number of clusters, and second we assigned the tumor samples into clusters. Finally, the clusters were organized into a tree structure to capture tumor progression pathways. Other well-known clustering methods were also considered as alternatives. Except for the likelihood approach, which is based on probabilistic modeling (see 3.3), all the other methods considered here make use of a distance metric. Thus, we describe our chosen distance, or the equivalence - similarity, measure next before we detail the three stages of heritable clustering.

3.1 Similarity measures for epigenotypes and phenotypes

For our purposes, the data generated by methylation microarray are interpreted as categorical in nature (henceforth described as epigenotype) – 1: hypermethylated; 0: un-methylated. Methylation progression patterns among tumor samples are integral to the inheritance property of our model; hence, the capacity to capture such patterns is a requisite of any clustering method employed. Under these constraints, we choose to design our algorithm based on the concept of ε -similarity [25], which defines distance and similarity measures suitable for our analysis. Specifically, the Hamming distance [1] defines the distance between two binary vectors of equal length as the number of elements that have different bits. This distance measure is adopted for describing the distance between the epigenotypes of two tumor samples. For each tumor t , $t = 1, \dots, T$, let $\mathbf{X}_t = \{X_{tg}, g = 1, \dots, G\}$ be the epigenotype vector at G loci. The epigenotype distance between two tumor samples t_i and t_j is then defined as

$$d_g(i, j) = \frac{1}{G} \sum_{g=1}^G |X_{t_i g} - X_{t_j g}|.$$

Since most phenotype data (e.g., clinical stage, histological grade, or hormone receptor status) are categorical in nature, we assume that each tumor phenotype is a discrete ordinal,

or can be ordered sensibly beginning from 0 to $K_p - 1$, where K_p is the number of the categories for phenotype p . Similar to the notation for epigenotypes, we use $\mathbf{Y}_t = \{Y_{tp}, p = 1, \dots, P\}$ to denote the vector of phenotypes for tumor t . Then the phenotype distance between two tumors t_i and t_j is

$$d_p(i, j) = \frac{1}{P} \sum_{p=1}^P \frac{|Y_{t_i p} - Y_{t_j p}|}{K_p - 1}.$$

Finally, the similarity measure between two tumors t_i and t_j is defined as

$$S(t_i, t_j) = 1 - (w \cdot d_p(i, j) + (1 - w) \cdot d_g(i, j)),$$

where $0 \leq w \leq 1$ is a weight parameter to balance the contributions from epigenotype and phenotype similarities. Two tumors, t_i and t_j , are said to be ε -similar if and only if $S(t_i, t_j) \geq \varepsilon$, where $0 \leq \varepsilon \leq 1$ represents the level of similarity. In the proposed heritable clustering method, if two tumors are sufficiently similar, they will be clustered into the same group. The selection of an appropriate ε depends on the desired degree of similarity within a cluster. The lower the ε value, the less similarity (i.e. more variation) within each cluster is allowed. To balance the contributions from epigenotypes and phenotypes, and to guarantee a reasonable level of similarities among tumor samples within each cluster, we suggest considering the parameters w and ε in the following ranges: $0.2 \leq w \leq 0.8$ and $0.5 \leq \varepsilon \leq 1$.

3.2 Determination of number of clusters

For each combination of weight and similarity (w, ε) (e.g., $0.2 \leq w \leq 0.8, 0.5 \leq \varepsilon \leq 1$), we use the following steps to determine the number of clusters.

1. Begin with two tumors $\{t_i, t_j\}$ that minimize the similarity measure $S(t_i, t_j)$. If the minimal value of S is greater than ε , assign all tumors into one cluster and stop. Otherwise let $C_1 = \{t_i\}$ and $C_2 = \{t_j\}$ and go to the next step.
2. Suppose there exist K clusters C_1, \dots, C_K . Let n_k be the number of tumors in C_k and t_{ki} be the i -th tumor to be added to $C_k, k = 1, \dots, K$, and $i = 1, \dots, n_k$. Let t be

a tumor sample that has not yet been assigned to any of the clusters. The similarity score between t and each of the existing cluster is defined as:

$$S(t, C_k) = \sum_{i=1}^{n_k} S(t, t_{ki})/n_k.$$

Let $k^* = \arg \max\{S(t, C_k), k = 1, \dots, K\}$. If $\min_{1 \leq i \leq n_{k^*}} S(t, t_{k^*i}) \geq \varepsilon$, then $C_{k^*} = C_{k^*} \cup \{t\}$; otherwise create a new cluster $C_{K+1} = \{t\}$.

3. Repeat step 2 until all tumor samples are assigned to clusters. Then calculate the total similarity score

$$T_S(w, \varepsilon) = \sum_{k=1}^{K(w, \varepsilon)} AS_k,$$

where $AS_k = \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} S(t_{ki}, t_{kj})/n_k^2$ is the average similarity in cluster k and $K(w, \varepsilon)$ is the corresponding number of clusters with parameters w and ε .

What remains is to find the optimal cluster number K and its corresponding parameter pair (w, ε) . In general, if the number of clusters K is large, then T_S is also large, and consequently $-\log(T_S)$ is small. This leads us to propose a model selection criterion following the formulation of Akaike's Information Criterion (AIC) [9]. The main idea is to maximize total similarity subject to a penalty term for over stratification. Specifically, we seek (K, w, ε) that satisfies

$$(K, w, \varepsilon) = \arg \max\{f(w, \varepsilon) = \log(T_S(w, \varepsilon)) - \frac{K(w, \varepsilon)}{P + G}; \quad 0.2 \leq w \leq 0.8, 0.5 \leq \varepsilon \leq 1\},$$

where P and G denote the number of phenotypes and epigenotypes, respectively. Note that the second term in the objective function f is to penalize over estimation of the number of clusters. It is designed to balance the number of clusters and total similarity, as in AIC.

3.3 Clustering samples

We then turn to clustering algorithms to group samples into K clusters, where K is the optimal number of clusters determined from the previous stage. Here we describe two novel algorithms, one based on similarity and the other based on likelihood. Both methods are iterative procedures like that of k-means, and therefore it is worth noting that the first stage

of heritable clustering also produces clusters as a by-product, which can be conveniently used as initial clusters here. These two algorithms will be evaluated, in the next section, through performance comparisons with several popular clustering algorithms (e.g., k-means and hierarchical clustering) using a breast tumor dataset.

Distance-based similarity approach

The idea of this similarity (SIM) algorithm is to group samples such that each tumor is more similar to those in the same cluster than to those in a different cluster. Let $a(t)$ denote the average similarity between tumor t and all the other tumors in cluster C_{k^*} to which t belongs. For any of the other clusters $C_k, k \neq k^*$, let $d(t, k)$ be the average similarity of t to all samples in C_k . We denote by $b(t) = \max_{k \neq k^*} d(t, k)$ the similarity between t and its nearest “neighbor” cluster. If $ab(t) = (a(t) - b(t)) \geq 0$, we say that t is correctly assigned to its current cluster, C_{k^*} , otherwise, it is a candidate for switching its cluster membership.

Algorithm: SIM

1. Calculate $ab(t)$ for each t and let $abmin = \min_t ab(t)$.
2. If $abmin < 0$, move the corresponding t to its “neighbor” cluster.
3. Repeat 1 - 2 until $abmin \geq 0$.

Likelihood-based method

Unlike the SIM algorithm or most other existing algorithms in the literature, the likelihood-based method proposed here does not depend on any measure of distance. This leads to greater flexibility, in that the approach can deal with both discrete and continuous data, missing observations for some of the variables, and dependencies among the variables. We briefly describe the method and provide the algorithm below; technical details of the method are deferred to the appendix for readability. The idea is similar, in spirit, to that reported in Cai et al. [3] for SAGE data. We assume that samples in each cluster (C_k) follow the same parametric distribution, with the maximum likelihood estimates of the parameter vector ($\hat{\theta}^k$) representing the profile of the cluster. For a tumor sample t , if its “likelihood”

$P(\mathbf{X}_t, \mathbf{Y}_t \mid \hat{\boldsymbol{\theta}}^k)$, where P is a probability measure in the currently assigned cluster C_{k^*} , is smaller than that in any of the other clusters, then t is a candidate for membership switches (see Appendix for details).

Algorithm: LH

1. For each tumor t , calculate $L_t(k^*) = -\log P(\mathbf{X}_t, \mathbf{Y}_t \mid \hat{\boldsymbol{\theta}}^{k^*})$ and $L_t(k \neq k^*) = \min_{k \neq k^*} \{-\log P(\mathbf{X}_t, \mathbf{Y}_t \mid \hat{\boldsymbol{\theta}}^k)\}$.
2. If $\max_t (L_t(k^*)/L_t(k \neq k^*)) > 1$, move the corresponding t to cluster $C_{\hat{k}}$, where $\hat{k} = \arg \min_{k \neq k^*} \{-\log P(\mathbf{X}_t, \mathbf{Y}_t \mid \hat{\boldsymbol{\theta}}^k)\}$.
3. Repeat 1 - 2 until $\max_t \leq 1$.

3.4 Building progression tree

In the final stage of heritable clustering, clusters generated from the previous stage will be assembled into a tree structure to represent the pathway of tumorigenesis. We first describe the concepts of cluster centers and scores, which are essential for our pathway discovery (PD) algorithm.

The clusters as previously described become the nodes of the tumor progression model. In order to derive pathways between nodes, a vector representative of the epigenotype and phenotype signatures of the tumors within a given node needs be defined. Node centers and scores (both epigenotypic and phenotypic) derived from each cluster are used to define such a vector and is referred to as the node label.

The epigenotype center of a cluster is determined by the epigenotype status common to the majority of tumors in the cluster. Let V_{kg} denote the set of epigenotype statuses at locus g over all tumors in cluster C_k , and $P(V_{kg})$ be the number of 1s in V_{kg} . Then the *epigenetic node center (ENC)* for locus g in cluster C_k is defined as:

$$GC_{kg} = \begin{cases} 1, & \text{if } P(V_{kg}) \geq \text{card}\{V_{kg}\}/2; \\ 0, & \text{otherwise;} \end{cases}$$

where $\text{card}\{V_{kg}\}$ is the cardinality of the set V_{kg} . The epigenotype score, or degree, of the node g is then defined based on the calculated node center as follows:

$$GS_k = \frac{1}{G} \sum_{g=1}^G GC_{kg},$$

i.e., the proportion of 1s in the set of ENCs for the node. The epigenotype score of a node is interpreted as measuring the extent of methylation of the tumors within the cluster.

With the definition of epigenotype centers and scores, it is now possible to define heritability of a progeny C_j from a parental node C_i :

$$H(C_i, C_j) = \frac{\text{card}\{g \mid GC_{ig} = GC_{jg} = 1, g = 1, \dots, G\}/G}{GS_i}.$$

The value of $H(C_i, C_j)$, which is between 0 and 1 and meaningful only if $GS_i \leq GS_j$, is the degree of heritability. Strict inheritance is defined when $H = 1$. Under this condition, all hypermethylated loci in a parental node are inherited by its progeny nodes. Note that the heritability is defined on the loci methylation signature of the ENC and not the methylation signature of the tumors that comprise the node. Such a definition of heritability is faithful to the recapitulation nature of the method.

In an analogous manner, phenotype centers and scores are used to capture the clinical progression in tumorigenesis. The center of a phenotype in a cluster is taken to be the weighted average of the phenotype values of the samples in the cluster rounded to the nearest integer. Let n_p be the number of categories for phenotype p and $c_{ki} = \text{card}\{Y_{tp} = i \mid t \in k\}$ be the count of category i in cluster $C_k, i = 1, \dots, n_p$. Then the phenotypic node center of phenotype p in cluster C_k is

$$PC_{kp} = \left\lfloor \frac{\sum_{i=1}^{n_p} i c_{ki}}{\sum_{i=1}^{n_p} c_{ki}} + 0.5 \right\rfloor,$$

where $\lfloor \cdot \rfloor$ is the floor of the value being bracketed.

The phenotype score for cluster C_k is then calculated as

$$PS_k = \frac{1}{P} \sum_{p=1}^P \frac{PC_{kp}}{K_P - 1}.$$

This score can be interpreted as measuring the average phenotypic value of the tumors in the cluster, with a larger score being indicative of more advanced tumors. Analogous to the concept of epigenotype heritability, we assume that phenotypic scores follow a temporal

order. That is, a node with a small score represents a tumor that occurred temporally before a tumor represented by a node with a larger score. Our PD algorithm is built to capture this chronological characteristic.

Algorithm: PD

1. Sort nodes in ascending order according to their epigenotypic scores with ties determined by their phenotypic scores. Assume, with possible relabeling, that the ordered nodes are C_1, \dots, C_K . Set node C_1 as the initial root node.
2. Suppose C_1, \dots, C_{k-1} have been used to build the tree. Consider the next node C_k . Let C_i denote a current terminal node (a node without any progeny) that satisfies (a) $H(C_i, C_k) \geq h$ (preset level of heritability, say, $h = 1$ for strict heritability) and (b) for each phenotype p , $PC_{ip} \leq PC_{kp}$. If such a node can be uniquely identified, then C_k is added as its progeny. If there are $m (> 1)$ candidates satisfying both the conditions (a) and (b), then C_k is added as a progeny of the node C_{i_m} with maximal epigenotype score GS_{i_m} (which should also have maximal phenotype score GP_{i_m} if there are still ties).
3. If no current terminal nodes can be a parent of C_k as none of them satisfy (a) and (b) in step 2, then consider the previous generation of nodes successively until a parental node is found. If a parent has not been identified up to the root node, then add C_k as its sibling node, and create a new (pseudo) root node with epigenotype and phenotype centers and scores set to be the minimal values.
4. Go back to step 2 until all the nodes are connected to the tree.

4 Application to Breast Tumor Progression Pathway

4.1 Data

Methylation analyses were initially performed on 93 breast carcinomas from unrelated patients, and their sample amplicons were deposited on the array [4]. The studied gene probes were hybridized to the array sequentially to generate composite methylation signatures [4]. A total of 10 genes were studied for their methylation status (0, unmethylated; 1, hypemethylated) in these tumor samples. These genes were chosen for analysis because of their known involvement in tumor suppression [23]. For a description of the methods used for generating the methylation data as well as assigning the discrete methylation values see [4]. Since gene BRCA2 is not methylated in any tumors, it is excluded from the final data analysis and model building. The remaining 9 genes used for pathway recapitulation are GPC3, RASSF1A, WT1, uPA, HOXA5, p16, 3OST3B, BRCA1, and DAPK1.

There are also five clinical phenotypes, and the categorical values of each phenotype are considered ordinal, with the lowest level to be adjusted to 0 for the heritable clustering analysis:

- Y1=age (1: age > 55; 2: age \leq 55),
- Y2=ER/PR (1: +/+; 2: +/- or -/+; 3: -/-),
- Y3=histology (1: well- differentiated (WD); 2: moderately-differentiated (MD); 3: poorly-differentiated (PD)),
- Y4=clinical stage (1, 2, 3, or 4), and
- Y5=metastasis status (0: no; 1: yes).

Of the 93 samples for which epigenotyps are available, seven have missing data on some of the phenotypic measurements. Therefore only the 86 samples that have complete data on both epigenotyps and phenotypes were used in the analyses presented in this section. However, the likelihood method is amenable to the full set of 93 samples, as will be elaborated in the Discussion section.

Table 1: The top five clustering outcomes (ranked by the values of the objective function f) and the corresponding w and ε values from the *NodeDiscovery* procedure.

Rank	w	ε	#Cluster(K)	Total Similarity (T_S)	$f(w, \varepsilon)$
1	0.7	0.7	15	13.42	1.53
2	0.8	0.7	12	10.76	1.52
3	0.2	0.7	17	15.18	1.51
4	0.6	0.7	16	13.97	1.49
5	0.3	0.7	18	15.97	1.48

4.2 Number of clusters

For the 86 samples with complete data, our model selection method was employed to find the optimal number of clusters and its associated parameter values for weight and similarity. Specifically, to apply our model selection procedure, we considered parameters w and ε in the range of 0.2 and 0.8, and 0.5 and 1, respectively, in increments of 0.1. We arranged the resulting values of the objective function $f(w, \varepsilon)$ in ascending order. The result that optimizes our objective function has 15 clusters, with both the w and ε values being 0.7, as shown in Table 1. Also shown in the table are the next four best results according to the criterion. Note that the numbers of clusters among this group are all quite similar. The full table is available as supplementary material.

4.3 Clustering analysis

We applied the two clustering algorithms developed, SIM and LH, to group the 86 tumor samples into 15 clusters. For the LH approach, the nine epigenotypes were treated as independent binomial variables, as was the age phenotype. However, since ER/PR status and histology ($\rho = 0.46; p < 0.0001$) were significantly positively correlated, these two variables were modeled jointly as follows: $p_0 = P(|Y_2 - Y_3| = 0)$, $p_1 = P(|Y_2 - Y_3| = 1)$, and $p_2 = P(|Y_2 - Y_3| = 2) = 1 - p_0 - p_1$. Similarly, the high positive correlation between clinical stage and metastasis ($\rho = 0.62; p < 0.0001$) led us to the joint modeling of

Table 2: Goodness-of-fit test for different cluster methods with different criteria.

Method	LH	SIM	K-means	PAM	H-clust
-log(LH)	75.16	80.50	91.18	126.27	131.13
Silhouette	0.11	0.32	-0.08	0.28	0.23
Entropy	77.19	70.44	80.39	116.55	123.88

these two variables, which in essence was a binomial probability distribution with parameter $p = P((Y4 \leq 2 \ \& \ Y5 = 0) \text{ or } (Y4 \geq 2 \ \& \ Y5 = 1))$.

In addition to these two novel clustering methods, we also analyzed the same set of data using three popular clustering methods in the literature, namely, K-means, PAM, and hierarchical clustering (H-clust), setting the number of clusters to 15. For these three popular algorithms and SIM, the distance measure was as described before with w and ε set to correspond to the choice of the optimal number of clusters (table 1). For all algorithms, the starting clustering assignments all use the by-product from the model selection step.

Three objective criteria, likelihood, silhouette, and entropy, were used to evaluate the outcomes of the various clustering algorithms. These criteria all try to measure the tightness of the samples within each cluster and/or the separation between clusters, albeit from different perspectives. Our results in Table 2 show that LH outperformed the others under the likelihood criterion, with SIM being a close second. On the other hand, SIM came out a winner as far as silhouette and entropy are concerned. Both PAM and H-clust were not too far off from the optimal achieved by SIM in terms of silhouette, but they are not competitive under the other two criteria. The performance of K-means is the opposite. While it did a descent job evaluated under the likelihood and entropy criteria, there were virtually no separations between the clusters, reflected in the negative silhouette value.

4.4 Pathway recapitulation

We applied our PD algorithm to each of the five clustering outcomes presented in Table 1. Shown in Figure 1 is the recapitulated pathway built from the nodes derived from SIM,

which performed the best for two of the three criteria evaluated and was also the second best under the third criterion. The red spots in these figures correspond to hypermethylated loci. The gene name of each locus is given in the corresponding cell of the 3×3 matrix arranged in the top-left corner of the figure. The number in the brackets below each node plot is the number of tumors in the cluster. The data above each node plot are the phenotype centers (arranged in the same order as that described in the Data subsection) and score for that node. Finally, each tree built adheres to strict heritability. Pathway trees built based on clustering results from the other clustering algorithms can be found in the supplementary material.

4.5 Interpretation

The progression tree presented in Figure 1 depicts the optimal outcome from heritable clustering using the methylation profiles of 9 promoter CpG islands analyzed in 86 primary breast tumors. The analysis selects node centers that not only preserve the heritability of promoter hypermethylation, but also uncover pathways that are associated with specific clinical phenotypes. This progression tree portrays a linear relationship between methylation heritability and tumor progression. With the exception of branches A2-B7-C2, all the other branches show progressive accumulation of more hypermethylated loci during the development of breast cancer. In this regard, tumors at terminal nodes C1, C2, C3, and C4 tend to have higher scores (ranges: 0.47-0.93) of advanced phenotypes than other shorter branches. This proof-of-principle study, therefore, confirms the previous observation that aggressive tumors tend to exhibit higher levels of promoter hypermethylation [22, 10, 12]. It will be interesting to find out in a future study if these progression pathways are predictive in tumors exhibiting similar methylation profiles.

This progression tree also displays a complex relationship between methylation heritability and three clinical phenotypes (i.e., age at diagnosis, hormone receptor status, and metastasis). First, a large proportion of post-menopausal patients (age cutoff > 55 years old, an indicator of less aggressive tumors) tend to have less or no methylation in this 9 gene set (the majority of patients in node A1) whereas only a handful of pre-menopausal patients (age cutoff ≤ 55 years old, and indicator of more aggressive tumors) had similar

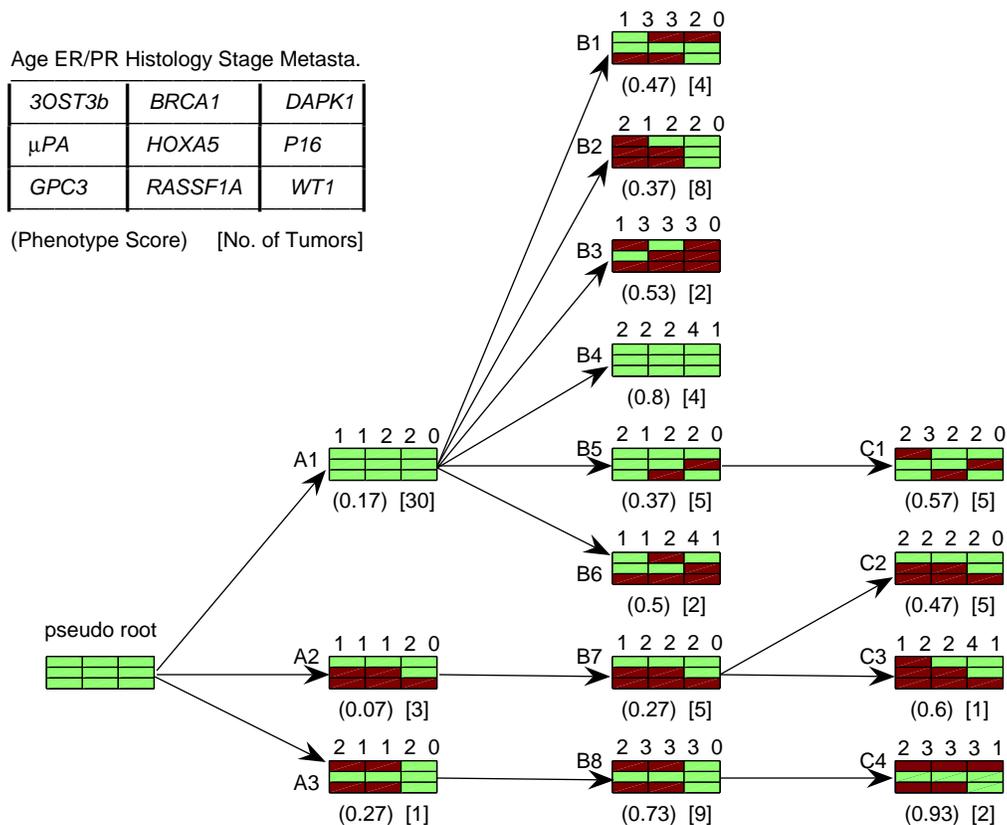


Figure 1: Progression pathway with $w = 0.7$, $\varepsilon = 0.7$ and likelihood updating method. The methylation data analyzed here are from 86 primary breast cancers. A set of 9 gene promoter CpG islands is investigated. The gene list is shown in the upper-left corner in a 3×3 format corresponding to the 3×3 blocks in each node of the progression tree. Red boxes indicate methylation in that specific gene promoter whereas green boxes indicate no detectable methylation. There are five phenotype measurements for each tumor. They are: age (1, age > 55 ; 2, age ≤ 55), ER/PR (1, +/+; 2, +/- or -/+; 3, -/-), histology (1, well-differentiated, WD; 2, moderately-differentiated, MD; 3, poorly-differentiated, PD), clinical stage (1, 2, 3, or 4), and metastasis status (0, M0; 1, M1). The phenotype center for each tumor phenotype is listed above each node in the order described above. The phenotype score of each node is presented within the parentheses and the number of tumors within each cluster is within the brackets below each node. The tree presented here conforms to strict heritability and ideal tumor phenotype progression.

methylation genotypes (among those in B4). Based on this feature, the majority of these tumors are assigned to an early node, A1, from which progressive accumulation of different loci likely disseminates to different branches. This observation seems to indicate that age-related methylation [11] may play a more prominent role in the breast cancer etiology of older patients (i.e., post-menopausal women). Second, hormone receptor (ER/PR)-positive tumors can be clustered to early nodes (e.g., A1, A2, and A3) in this progression tree while hormone receptor-negative tumors are grouped into terminal nodes (e.g., B1, B3, C1, and C4) or non-terminal, second-generation nodes (e.g., B8). These results corroborate a common notion that aggressive tumors usually exhibit a hormone receptor-negative phenotype [15, 24]. Third, in our progression model metastatic tumors are clustered into a node after tumors with no metastasis. Except node B4, metastatic nodes B6, C3, and C4 usually have higher methylation events (≥ 5 loci) than non-metastatic nodes. Similarly, we observe that tumors with advanced histology grades and clinical stages usually develop from a low grade or stage during methylation progression.

Taken together, the outcome of this statistical modeling is in sync with a stepwise accumulation of molecular alterations during breast tumorigenesis. Because of its heritable nature, promoter hypermethylation can be viewed as a molecular relic from which the history of breast cancer can be reconstructed. Future studies will be expanded to conduct the analysis of more hypermethylated loci in a large size of tumor samples, allowing for further testing of our heritable clustering model. Such an analysis will provide insight into the molecular mechanisms of promoter hypermethylation and is useful for the prognostication of breast cancer in a clinical setting.

5 Discussion

In this paper, we have developed novel methods for each of the three stages of the heritable clustering procedure. Although existing clustering methods are applicable to the second stage of the procedure, our proposed clustering algorithms, SIM and LH, outperformed their counterparts based on three objective evaluation criteria, for the data examined. Furthermore, our heritable clustering procedure seems to be able to capture the biological essence

of tumor progression, as discussed below in general, and as elaborated in 4.5 for the breast tumor example in particular. It is especially worth noting that the LH algorithm can incorporate dependencies among the variables and include samples with some missing values. To demonstrate the applicability of the LH algorithm for incomplete samples, we reanalyzed the breast tumor data using all 93 samples. The resulting pathway tree is available as supplementary material. Despite these initial encouraging results, further experimentations are warranted to fully evaluate the proposed methods. In particular, we are especially interested in applying the heritable clustering method to larger scale tumor progression pathway analyses using stroma and tumor data.

The preliminary application of the heritable clustering algorithms to the breast tumor data demonstrates its effectiveness in identifying pathways with unambiguous epigenetic and phenotypic progression. The constructed pathway summarizes the epigenetic and phenotypic data in a way that corresponds to the current understanding of tumor progression. Further, the potential of methylation profiles to be used for characterizing tumor progression has been demonstrated. The resulting trees from our tumor progression pathway recapitulation procedure depend on a number of factors including: 1) distance between tumors (epigenotype and phenotype); 2) balance between epigenotype and phenotype data; 3) similarities within clusters; and 4) heritability between nodes. The best results are those that reflect the underlying biological processes that lead to the formation of the primary tumors. Our heritable clustering method is designed based on the assumption that epigenetic changes are stably passed from progenitor to progeny cells [13]. Depending on what stage each tumor is diagnosed, some might have accumulated more epigenetic alternations than others as they have progressed more. In this paper, we capitalize on these epigenetic hallmarks to recapitulate breast tumor progression pathways utilizing CpG island hypermethylation data

In building the tumor progression pathway, the assumption is based on the heritable nature of CpG island hypermethylation passing from the parent node to its progeny nodes as tumor progresses. Therefore, the progeny nodes of tumor cells accumulate more hypermethylated gene promoters as they are further along in the progression pathway. The progeny tumor cells are likely to be more aggressive and have more proliferative advantages than the parental cells. Hence, we built a tree model by linking the nodes or clusters based on strict

heritability and their phenotype scores.

In practice, it is unlikely to recreate a linear temporal clini-copathological history of a cancer developing over time in a single patient as it is unethical to remove part of the tumor and allows a portion to grow for research purposes. To overcome this challenge common to all human genetic and epigenetic studies, we propose to view CpG island hyper-methylation as "molecular relics" whereby one can trace how much each tumor has progressed by examining the overall methylation profile as such information is stably transmitted from parent cells to their progeny. The heritable clustering method developed in this paper is designed to uncover the different paths breast tumors can progress. Our results from the breast tumor application indicate that this approach will select meaningful progression models and will assist in the interpretation of pathways having biological and clinical significance.

In this present application of the heritable clustering method, the epigenetic and clinical phenotypic values took on discrete values. However, the method can be extended to analyze other data types where the numerical values of the data are continuous. For instance, the method is well suited for modeling methylation data expressed as intensity ratios from two-color microarray experiments or transcript factor binding enrichment on gene promoters from ChIP-on-chip experiments.

Appendix: Likelihood Clustering Approach

We assume that the epigenotype vectors (\mathbf{X}_t) for tumor t follows a common parametric family of distributions with its own parameter vector $\boldsymbol{\theta}_{tG}$. Analogously, we use $\boldsymbol{\theta}_{tP}$ to denote the parameter vector for the distribution of the clinical phenotype vector \mathbf{Y}_t . That is,

$$\begin{aligned}\mathbf{X}_t &= \{X_{t1}, X_{t2}, \dots, X_{tG}\} \sim P(\cdot | \boldsymbol{\theta}_{tG}), \\ \mathbf{Y}_t &= \{Y_{t1}, Y_{t2}, \dots, Y_{tP}\} \sim P(\cdot | \boldsymbol{\theta}_{tP}).\end{aligned}$$

Thus, \mathbf{X}_t and \mathbf{Y}_t are jointly distributed as

$$(\mathbf{X}_t, \mathbf{Y}_t) \sim P(\cdot | \boldsymbol{\theta}_t = (\boldsymbol{\theta}_{tG}, \boldsymbol{\theta}_{tP})).$$

If \mathbf{X}_t and \mathbf{Y}_t are assumed to be independent, then

$$P(\mathbf{X}_t, \mathbf{Y}_t \mid \boldsymbol{\theta}_{tG}, \boldsymbol{\theta}_{tP}) = P(\mathbf{X}_t \mid \boldsymbol{\theta}_{tG})P(\mathbf{Y}_t \mid \boldsymbol{\theta}_{tP}).$$

The goal is to group tumors with similar epigenotypes and phenotypes according to their parameter vectors. That is, we assume that tumors within a cluster (C_k) share the common distributional parameter vector $\boldsymbol{\theta}^k = \{\boldsymbol{\theta}_G^k, \boldsymbol{\theta}_P^k\}$, which represents the cluster profile. Let $I_k(t) = 1$ if tumor t is in cluster C_k , otherwise it is 0. Thus, the joint likelihood is

$$L(\boldsymbol{\theta}_G^k, \boldsymbol{\theta}_P^k, k = 1, \dots, K \mid \mathbf{X}_t, \mathbf{Y}_t, t = 1, \dots, T) = \prod_{k=1}^K \left\{ \prod_{t=1}^T [P(\mathbf{X}_t, \mathbf{Y}_t \mid \boldsymbol{\theta}_G^k, \boldsymbol{\theta}_P^k)]^{I_k(t)} \right\},$$

where K is the number of clusters, and T is the number of tumor samples. Suppose that $\hat{\boldsymbol{\theta}}_G^k$ and $\hat{\boldsymbol{\theta}}_P^k$ are the maximum likelihood estimate of $\boldsymbol{\theta}_G^k$ and $\boldsymbol{\theta}_P^k$, respectively, $k = 1, \dots, K$, then it is natural to evaluate how well a particular tumor sample fits into the assigned cluster by computing

$$k^* = \operatorname{argmin}_k \{-\log P(\mathbf{X}_t, \mathbf{Y}_t \mid \hat{\boldsymbol{\theta}}_G^k, \hat{\boldsymbol{\theta}}_P^k); k = 1, \dots, K\}.$$

If C_{k^*} is not the same as its currently assigned cluster, then tumor t is a candidate for switching cluster membership. This basic idea may lead to various clustering algorithms, including the one used for our primary breast tumor data.

The above formulation of the likelihood clustering approach provides a general setting in which dependencies among epigenotypes (e.g., hypermethylated promoter regions binded to the same transcription factor) and phenotypes (e.g, tumor grade and metastasis status) can be easily incorporated. In the breast tumor example, we have discrete epigenotypes (hypermethylated or not) and phenotypes (ordinal), therefore binomial and multinomial are the natural choice of parametric families for the distributions of the variables. However, the framework can be flexibly adapted to any other type of data, such as continuous measurements of methylation intensities. Finally, the approach can make use of tumor samples that have missing data on some of the variables; the contribution to the corresponding likelihood from such a sample will be set to unity by convention.

Acknowledgments

The authors wish to thank Professor Chi-Ren Shyu at the University of Missouri for suggesting the initial idea of Heritable Clustering. ZW is supported by NSF Postdoctoral Fellowship under Agreement No. 0112050 through MBI at OSU. This work is also supported by the National Cancer Institute grants P50CA113001, P30CA16058 and R01CA069065, and OSU James Cancer Center fund. CE is a recipient of the Doris Duke Distinguished Clinical Scientist Award. SL is supported in part by NSF grant DMS-0306800.

References

- [1] R Baeza-Yates and B Ribeiro-Neto, *Modern information retrieval*, The ACM Press, NY, 1999.
- [2] SB Baylin, *Dna methylation and gene silencing in cancer*, Nat Clin Pract Oncol (2005), no. 2, Suppl 1(S1):S4–S11.
- [3] L Cai, H Huang, S Blackshaw, JS Liu, C Cepko, and WH Wong, *Clustering analysis of sage data using a poisson approach*, Genome Biology (2004), no. 5, R51.
- [4] CM Chen, HL Chen, TH-C Hsiau, AH-A Hsiau, H Shi, G Brock, SH Wei, CW Caldwell, PS Yan, and TH-M Huang, *Methylation target array for rapid analysis of cpg island hypermethylation in multiple tissue genomes*, Am J Pathol (2003), no. 163, 37–45.
- [5] JF Costello, *Comparative epigenomics of leukemia*, Nat Genet **37** (2005), no. 3, 211–212.
- [6] S Datta and S Datta, *Comparisons and validation of statistical clustering techniques for microarray gene expression data*, Bioinformatics (2003), no. 19, 459–466.
- [7] AP Feinberg, *The epigenetics of cancer etiology*, Semin Cancer Biol **6** (2004), no. 14, 427–432.
- [8] DP Genereux, BE Miner, CT Bergstrom, and CD Laird, *A population-epigenetic model to infer site-specific methylation rates from double-stranded dna methylation patterns*, Proc. Natl. Acad. Sci. USA (2005), in press.

- [9] T Hastie, R Tibshirani, and J Friedman, *The elements of statistical learning: Data mining, inference and prediction*, Springer-Verlag, New York., 2001.
- [10] R Henrique, C Jeronimo, MO Hoque, S Nomoto, AL Carvalho, VL Costa, J Oliveira, Teixeira, C Lopes, and D Sidransky, *Mt1g hypermethylation is associated with higher tumor stage in prostate cancer*, *Cancer Epidemiol Biomarkers Prev* **5** (2005), no. 14, 1274–1278.
- [11] JP Issa, *Age-related epigenetic changes and the immune system*, *Clin Immunol* **1** (2003), no. 109, 103–108, Review.
- [12] ———, *Cpg island methylator phenotype in cancer*, *Nat Rev Cancer* **12** (2004), no. 4, 988–993.
- [13] PA Jones and SB Baylne, *The fundamental role of epigenetic events in cancer*, *Nature Review Cancer* (2002), no. 3, 415–428.
- [14] PA Jones and R Martienssen, *A blueprint for a human epigenome project: the aacr human epigenome workshop*, *Cancer Res* **65(24)** (2005), no. 15, 11241–11246.
- [15] JC Keen, E Garrett-Mayer, C Pettit, KM Mack, Manning, JG Herman, and NE Davidson, *Epigenetic regulation of protein phosphatase 2a (pp2a), lymphotactin (xcl1) and estrogen receptor alpha (er) expression in human breast cancer cells*, *Cancer Biol Ther* **12** (2004), no. 3, 1304–1312.
- [16] PW Laird, *Cancer epigenetics*, *Hum Mol Genet* **14 Spec No 1** (2005), no. 15, R65–76.
- [17] YW Leu, PS Yan, VX Jin, JC Liu, EM Curran, WV Welshons, SH Wei, RV Davuluri, C Plass, Nephew, and TH-M Huang, *Loss of estrogen receptor signaling triggers epigenetic silencing of downstream targets in breast cancer*, *Cancer Res* (2004), no. 64, 8184–8192.
- [18] M Weber M, JJ Davies, D Wittig, EJ Oakeley, M Haase, WL Lam, and D Schubeler, *Chromosome-wide and promoter-specific analyses identify sites of differential dna*

- methylation in normal and transformed human cells*, Nat Genet **8** (2005), no. 37, 853–862.
- [19] R Jaenisch R and A Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals*, Nat Genet. (2003), no. 33, Suppl:245–54.
- [20] KD Robertson, *Dna methylation and human disease*, Nat Rev Genet. (2005), no. 6, 597–610.
- [21] E Vire, C Brenner, R Deplus, L Blanchon, M Fraga, C Didelot, L Morey, A Van Eynde, D Bernard, JM Vanderwinden, M Bollen, M Esteller, L Di Croce, Y de Launoit, and F Fuks, *The polycomb group protein ezh2 directly controls dna methylation*, Nature (2005), no. 14.
- [22] SH Wei, C-M Chen, G Strathdee, J Harnsomburana, CR Shyu, F Rahmatpanah, H Shi, SW Ng, PS Yan, KP Nephew, R Brown, and TH-M Huang, *Methylation microarray analysis of late-stage ovarian carcinomas distinguishes progression-free survival in patients and identifies candidate epigenetic markers*, Clin Cancer Res (2002), no. 8, 2246–2252.
- [23] M Widschwendter and PA Jones, *Dna methylation and breast carcinogenesis*, Oncogene (2002), no. 21, 5462–5482.
- [24] PS Yan, C-M Chen, H Shi, F Rahmatpanah, SH Wei, and CW Caldwell TH-M Huang, *Dissecting complex epigenetic alterations in breast cancer using cpg island microarrays*, Cancer Res (2001), no. 61, 8375–8380.
- [25] JP Yoon, V Raghavan, and V Chakilam, *Bitcube: a three dimensional bitmap indexing for xml documents*, J. Intellingent Systems (2001), no. 17, 241–254.