DNA methylation as a molecular relic to recapitulate tumor progression in breast cancer

Zailong Wang[1], Pearlly Yan[2,4], Charis Eng[2,3,4], Tim H. Huang,[2,4] and Shili Lin[5*]

[1]Mathematical Biosciences Institute, The Ohio State University, 231 W. 18th Avenue; [2]Department of Molecular Virology, Immunology, and Medical Genetics, [3]Division of Human Genetics, Department of Internal Medicine, [4]Human Cancer Genetics Program, Comprehensive Cancer Center, The Ohio State University, 420 W. 12th Avenue; [5]Department 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 CpG island hypermethylation data, a novel clustering algorithm called heritable clustering was developed. This new approach should be capable of creating a tumor progression model by utilizing data from tumor tissues diagnosed at different stages. These samples act as surrogates for natural tumor progression in breast cancer and ideally should allow the algorithm to uncover distinct epigenotype and phenotype that describe the molecular events underlying this process.

Results: Using this heritable clustering algorithm to analyze CpG island methylation data obtained from 50 primary breast cancers, we built several tumor progression trees in an attempt to recapitulate the pathways of breast tumor progression. One of these pathways was selected for detailed annotation and interpretation. Our results indicate that the proposed heritable clustering algorithm can provide an effective tool using methylation profiles and clinical variables to stratify tumor subtypes and stages.

Supplementary Information: The program implementing the method (Matlab codes) can be accessed at http://www.stat.ohio-state.edu/∼statgen/Pathway.html

Contact: shili@stat.ohio-state.edu.

1 Introduction

Recapitulating pathways of breast tumor progression by tracing its molecular lesions is necessary for understanding the disease and for the establishment of novel drug targets and therapies. The implementation of this concept by utilizing DNA methylation profiles are even more enticing in that DNA methylation is heritable and stable. A methylation enzyme (DNMT1) present in our cells ensures that any changes in methylation status is maintained and pass on to the next generation. This implies that methylation events responsible for silencing critical tumor suppressor genes that leads to tumorigenesis are captured in the DNA epigenetic code of a tumor. As DNA is more stable than RNA, researchers can retrospectively perform methylation analysis on samples collected at earlier times which tend to have more complete survival and patient clinicopathological information.

DNA methylation occurs at the 5' carbon of cytosine. It is noted that normally unmethylated CpG island (a stretch of DNA with higher than predicted amount of CG dinucleotides) located in the promoter region of cancer cells undergoes dense hypermethylation during tumor progression. The number of hypermethylated genes tends to increase with the malignant potential of the tumor. Methylation associated silencing of tumor suppressor genes can result in cells with a growth advantage, and clonal expansion of these proliferating cells bear specific epigenetic signatures reflecting different types or stages of various tumors. With state-of-the-art microarray technologies, it is now possible to obtain the methylation signature of multiple genes simultaneously and to classify tumors based on their global patterns of DNA methylation. This type of research provides an unprecedented opportunity to improve our understanding of the overall molecular mechanisms leading to the development of cancer (Alizadeh *et al.* 2000; Welsh et al. 2001; Yan *et al.* 2001). The microarray technique used to generate the data analyzed in this paper was described by Yan *et al.* (2001) and Chen *et al.* (2003). In this study, we set out to test the hypothesis that solid tumor development and progression are characterized by the progressive accumulation of epigenetic events. We have, therefore, utilized multiple hypermethylated loci to cluster and reconstruct the epigenetic history germane to tumorigenesis.

To recapitulate epigenetic progression using hypermethylation data, several challenging issues arise. First it is difficult to collect tumor tissues from the same patient at different stages. Instead, tumor tissues from different patients at different stages of progression with distinct phenotypes are studied. The second issue is the inapplicability of existing clustering algorithms to meet the need of our progression model. To deal with these, we develop a novel method called heritable clustering approach that has the ability to identify clusters and organize them into a tree to recapitulate tumor progression pathways.

The paper is presented in five sections. In section 2 we describe the concept of tumor progression pathways and their recapitulation. Some related clustering approaches are also briefly discussed. Section 3 shows in detail the heritable clustering approach and algorithm, including the *NodeDiscovery* and *PathwayDiscovery* procedures. The application of the algorithm to a breast cancer dataset is shown in section 4 followed by discussion on other possible applications in section 5.

2 Concepts

2.1 Tumor progression pathways

Tumor progression pathways are constructed based on the following properties:

- Most CpG islands are unmethylated in normal cells.
- CpG island hypermethylation is heritable in tumor cells.
- Multiple methylated loci are progressively accumulated during tumorigenesis.

Based on these properties, tumor cells could have unique epigenetic signatures (with consequent unique gene expression signatures) that are associated with specific cancer subtypes (phenotypic information). To represent possible tumor progression pathways, it is suitable to use directed acyclic graphs (DAGs) or tree diagrams (Cormen *et al.* 2001) due to the hierarchical nature of pathways.

2.2 Tumor progression recapitulation

The concept of tumor progression recapitulation is similar to a well-known biological evolution methodology–phylogenetics (e.g., see Eldridge and Cracraft (1980) and Brooks and McLennan (1991) for a phylogenetic process). They all consider the following matters:

- Construct the biological history based on assumptions, hypotheses, and available events.
- Utilize DAGs in the explanation of hypotheses.
- Include the temporal factor in terms of chronological order in the hypotheses.

More specifically, tumor progression recapitulation seeks to construct patterns and relationships among hypermethylated genes that are progressively accumulated during tumorigenesis. Similar to phylogenetic approaches, the data collected for the discovery of tumor progression do not have temporal information from the same individuals. That is we do not have tissues from the same patients at different stages of tumor progression over time. An acceptable approach of this type of progression model can be represented by using the available phenotypes from different tumors at different stages obtained from different patients.

Although the phylogenetic approach can be suitably adapted to tumor progression models, there are more requirements for tumor progression recapitulation. These include

- the characteristics of each branch of the tree; hence the phenotypes of progeny nodes are hypothesized to be more aggressive than the parents; and
- that the hypermethylated loci acquired at each node are inherited by the progeny cell(s) such that the hypermethylated loci progressively accumulated, and the parents' hypermethylated loci are subsets of their progeny's.

Therefore there exist two challenges in clustering tumors in the nodes of our tumor progression trees as follows:

1. Number of clusters is unknown.

2. Each cluster should make "biological sense".

These challenges impede us from adopting any published clustering algorithms without major modifications. A sampling of these algorithms is described in the following.

2.3 Related clustering approaches

Datta and Datta (2003) have compared and validated six statistical clustering algorithms applied to microarray gene expression data. The techniques they tested include hierarchical clustering with correlation and with partial least squares, K-Means clustering, divisive clustering (Theodoridis and Koutroumbas, 1998), fuzzy logic based clustering (Klir and Yuan, 1995), and model-based clustering (Yueng *et al.* 2001). Datta and Datta concluded that there exist some dependencies between the clustering results and the methods used, i.e., the clustering results are not concordant among the tested methods. In addition to this discrepancy in outcomes, clustering methods such as hierarchical clustering, K-Means clustering, and self-organizing maps (Kohonen, 1998) also require the user to predefine the number of clusters best describing the data set. This requirement also applies to clustering methods based on cost function optimization where the number of clusters is fixed.

A recent approach called adaptive quality-based clustering (Smet *et al.* 2001) was proposed to cluster gene expression data without knowing the predefined parameters, such as the number of clusters and data distribution. Another gene expression clustering approach called diametrical clustering (Dhillion *et al.* 2003), which can give "biologically meaningful" results, presented an interesting aspect of anti-correlated gene clustering. It clusters samples that share similar functions but are expressed differently.

3 Methods

Our proposed heritable clustering algorithm consists of two procedures: *NodeDiscovery* and *PathwayDiscovery*. *NodeDiscovery* is used to cluster tumor samples into a set of nodes. *PathwayDiscovery* is subsequently used to build a pathway tree from the set of nodes un-

covered by *NodeDiscovery*. These two procedures use three criteria: similarity, node centers and scores, and heritability, which will be described in turn in the following subsection.

3.1 Criteria

3.1.1 Similarity measures for epigenotypes and phenotypes

Data generated by methylation array are binary in nature (henceforth described as epigenotype; 1: hypermethylated; 0: unmethylated) and progression patterns among the clusters are integral to the inheritance property of our model. Under these constraints, we choose to design our algorithm based on the concept of ε-similarity (Yoon *et al.* 2001), which defines distance and similarity measures suitable for our analysis. Specifically, the Hamming distance (Baeza-Yates and Ribeiro-Neto, 1999) 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 (vectors). Let X_{tg} be the epigenotype status for tumor t at locus g, $t = 1, \cdots, T$, $g = 1, \cdots, G$. The epigenotype distance between two tumor samples t_i and t_j is defined as

$$d_g(i,j) = \sum_{g=1}^{G} (X_{t_i g} - X_{t_j g})^2.$$

Denote by $mindg$ and $maxdg$ the minimum and maximum of $d_g(i,j)$ over all pairs of tumor samples (t_i, t_j). A rescaled (to be between 0 and 1) epigenotype distance is

$$sd_g(i,j) = \frac{d_g(i,j) - mindg}{maxdg - mindg}.$$

For clinical data, we assume that each tumor phenotype is a discrete ordinal, or can be ordered sensibly. Let X_{tp} be the value for tumor t at phenotype p, $t = 1, \cdots, T$ and $p = 1, \cdots, P$. Analogous to the definition of epigenotype distance, the phenotype distance between two tumors t_i and t_j is defined as

$$d_p(i,j) = \sum_{p=1}^{P} (X_{t_i p} - X_{t_j p})^2.$$

Using *mindp* and *maxdp* to denote the minimum and maximum of $d_p(i,j)$ over all (t_i, t_j), we define the rescaled phenotype distance as

$$sd_p(i,j) = \frac{d_p(i,j) - mindp}{maxdp - mindp}.$$

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

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

where $0 \leqslant w \leqslant 1$ is a weight parameter to balance the contributions from epigenotype and phenotype similarities. Two vectors, t_i and t_j, are said to be ε-similar if and only if $S(t_i, t_j) \geqslant \varepsilon$, where $0 \leqslant \varepsilon \leqslant 1$ represents the level of similarity. If two vectors 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: $0.2 \leqslant w \leqslant 0.8$ and $0.5 \leqslant \varepsilon \leqslant 1$.

3.1.2 Epigenotype and phenotype node centers and scores

The *PathwayDiscovery* procedure steps through a series of processes to grow pathway trees. It makes use of the concepts of node centers and scores, both epigenotypic and phenotypic, defined on each cluster. The epigenotype center of a cluster can be calculated by taking the majority of each marker among tumor vectors. Let V_{kg} denote the g-th marker vector over tumors in cluster k, and $P(V_{kg})$ be the number of 1's in V_{kg}, then

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

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

$$GS_k = \text{number of 1's in the node center vector } \{GC_{kg}, g = 1, \cdots, G\}.$$

From the above definition, we can interpret the epigenotype score of a node as measuring the extend of methylation of the tumors within the cluster.

In building a pathway tree, we also use the concepts of phenotype centers and scores 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} be the count of category s_i in cluster $k, i = 1, \cdots, n_p$. Then the center of phenotype p in cluster k is

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

where $\lfloor\ \rfloor$ is the floor of the value being bracketed. Its rescaled (0 - 1) value is

$$SPC_{kp} = \frac{PC_{kp} - min_t(X_{tp})}{max_t(X_{tp}) - min_t(X_{tp})}.$$

The phenotype score for cluster k is then calculated as

$$PS_k = \frac{1}{P} \sum_{p=1}^{P} SPC_{kp}.$$

This score can be interpreted as measuring the phenotypic (clinical) degree of the tumors in the cluster, with a larger score being indicative of more advanced tumors.

3.1.3 Heritability

One more concept is needed for building a heritable tree, which is heritability of a child node C_j from its parent node C_i in terms of the epigenotypic node centers. This is defined in terms of the epigenotype scores and the number of commonly methylated loci. Specifically, using $com(C_i, C_j)$ to denote the number of loci with common 1's in their epigenotypic node center vectors $\{GC_{ig}, g = 1, \cdots, G\}$ and $\{GC_{jg}, g = 1, \cdots, G\}$. If $GS_i \leqslant GS_j$, we define heritability of C_j from C_i as:

$$H(C_i, C_j) = \frac{com(C_i, C_j)}{GS_i}.$$

The value of $H(C_i, C_j)$, which is between 0 and 1, is the degree of heritability. Strict inheritance is defined when $H = 1$. Under this condition, all hypermethylated loci in a parent node is inherited by its progeny nodes.

3.2 Heritable clustering

Our proposed heritable clustering algorithm maintains a level of heritability from parental clusters to progeny clusters. The process of building the tumor progression pathways is to group tumor vectors into nodes (clusters) using the *NodeDiscovery* procedure first. Then the *PathwayDiscovery* procedure is employed to build links among these nodes.

3.2.1 NodeDiscovery–clustering samples

The *NodeDiscovery* procedure for clustering tumor vectors consists of the following steps:

1. Begin with two vectors $\{t_i, t_j\}$ that are the least similar. If this similarity is greater than ε, assign all tumor vectors 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 $C_1, \cdots, C_K$. Let n_k be the number of tumors in C_k and t_{ki} be the i-th tumor vector in $C_k, k = 1, \cdots, K, i = 1, \cdots, n_k$. Let t be a tumor sample that has not been assigned to any of the clusters yet. Compute the similarity score between t and each of the existing cluster: $S(t, C_k) = \sum_{i=1}^{n_k} S(t, t_{ki})/n_k$. Let $k^* = \arg\max\{S(t, C_k), k = 1, \cdots, K\}$. If $\min_{1 \leqslant i \leqslant n_{k^*}} S(t, t_{k^*i}) \geqslant \varepsilon$, 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 node centers and scores for both epigenotypes and phenotypes in each cluster.

This procedure returns the clusters $C_1, \cdots, C_K$ and their node centers and scores, which will be used in building the tumor progression pathway in *PathwayDiscovery*.

It remains to find the set of appropriate weights, w, and levels of similarity, ε. For each combination of (w, ε) (e.g., $0.2 \leq w \leq 0.8, 0.5 \leq \varepsilon \leq 1$) investigated, define a total similarity score for the clustering outcome:

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

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. Note that this definition is applicable to clustering outcomes that contain clusters with a single tumor sample. In general, if the number of clusters K is large, then T_S is usually also large, and consequently $-2ln(T_S)$ is small. This observation leads us to the proposal of a model selection criterion following the formulation of Akaike's Information Criterion (AIC) (Hastie et al. 2001). Specifically, we seek (w, ε) that satisfies

$$\begin{aligned}(w, \varepsilon) &= \arg\min f(w, \varepsilon) \\ &= \arg\min\{-2ln(T_S) + \frac{2K(w, \varepsilon)}{P + G};\ 0.2 \leq w \leq 0.8,\ 0.5 \leq \epsilon \leq 1\}\end{aligned}$$

as well as its neighbors with near optimal values. This scheme offers multiple candidate clustering sets for building progression trees, which increases the chances of finding a biologically plausible tumorigenesis pathway. Note that the second term in the above model selection formula is used to penalize over estimation of the number of clusters. It is designed to balance the number of clusters and total similarity, as in AIC.

3.2.2 PathwayDiscovery–Building progression tree

The *PathwayDiscovery* procedure for building tumor progression tree consists of the following steps:

1. Sort nodes first by their phenotypic scores followed by their epigenotypic scores for ties, both in ascending order. Assume that the ordered nodes are $C_{(1)}, \cdots, C_{(K)}$, then set $C_{(1)}$ as the current root node.

2. Suppose $C_{(1)}, \cdots, C_{(k-1)}$ 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, $PC(C_{ip}) \leqslant PC(C_{(k)p})$. 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), we sort them according to their epigenotype scores such that $GS(C_{i1}) \leqslant GS(C_{i2}) \leqslant \cdots \leqslant GS(C_{im})$. Then $C_{(k)}$ is added as a progeny of C_{im}.

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 current root node, then add $C_{(k)}$ as its sibling node, and create a new (pseudo) root node.

4. Go back to step 2 until all the nodes are connected to the tree.

4 Results

Table 1 and Figure 1 present the clustering results from breast cancer methylation data using our heritable clustering algorithm. The resulting pathways are represented pictorially by progression trees, which explain accumulation of methylated loci. The phenotypic characteristics of each node in the pathway are shown in the corresponding plots.

4.1 Data preparation

Methylation analyses were performed on 50 breast carcinomas from unrelated patients. For each tumor, we studied 10 genes for their methylation status (0, unmethylated; 1, hypemethylated). 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 phenotype measurements for each tumor. They are: age (1, age $\geqslant$ 65; 2, age $<$ 65), 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 data is organized into a two-dimensional matrix with each row representing one tumor (or patient) and each column representing either phenotype ordinal values (five columns) or gene methylation status (nine columns).

4.2 Pathway recapitulation

For our dataset, we first used the *NodeDiscovery* procedure with $w = 0.2, 0.3, \cdots, 0.8$ and $\varepsilon = 0.5, 0.6, \cdots, 1$ to find the appropriate w and ε values. We arranged the resulting values of the objective function $f(w, \varepsilon)$ in ascending order. Table 1 shows the top five w and ε values of the clustering results from *NodeDiscovery* and the corresponding numbers of clusters and the values of the objective function.

In Figure 1, we present the tree built using *PathwayDiscovery* from the clustering outcome that optimized the selection criterion (first row of Table 1). The red spots in the figure 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-right corner of the figure. The number 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 preparation subsection) and score for that node. Finally, the tree built adheres to strict heritability.

4.3 Interpretation

The progression tree presented in Figure 1 depicts the optimal outcome from *NodeDiscovery* using the methylation profiles of 9 promoter CpG islands frequently hypermethylated in primary breast tumors. In this particular progression tree, the proposed clustering method selects for node centers which not only preserve strict heritability of promoter methylation but also uncover pathways with perfect progression in the 5 selected breast tumor phenotypes. This is an important criterion in that promoter hypermethylation in this set of tumor associated and/or tumor suppressor genes may lead to tumorigenesis or tumor progression whereas hypermethylation of bystander CpG islands may have little consequences. As such, we propose that tumors with more aggressive phenotypes should exhibit higher level of

methylation in this gene panel than the less aggressive tumors. The first phenotype studied is the age of diagnosis. It is known that a young age of tumor onset generally correlates with a more aggressive disease. Often DNA methylation plays less of a role in tumorigenesis in this subset. As there are 26 patients older than 65 and 24 patients younger than 65, a cutoff of 65 will definitely segregate pre-menopausal patients (thus patients with the more aggressive cancer) from the post-menopausal patients. Hormone receptor status or ER/PR status is another phenotype that distinguishes early, less aggressive tumors from late, more aggressive tumors. Therefore, tumors expressing a measurable level of ER and PR (an assigned value of 1 or 2) should be clustered to the early nodes of the tree while tumors without ER and PR expression should appear closer to the terminal nodes. Another tumor phenotype that should follow stringent progression is tumor metastasis. A tumor that has shed a portion of its cells to distant sites such as lymph nodes represents a late stage, aggressive tumor. As such, tumors with a metastasis value of 1 should not appear in a node before tumors with no metastasis (metastasis value = 0). Tumor phenotypes relating to histology and tumor staging should progress similarly from a low grade or stage to a high grade or stage in the progression tree. This particular computation outcome clusters nodal relationships that are completely in sync with our proposed model of strict heritability and perfect progression.

Our previous analysis on this data set showed that a large number of tumors have concurrent hypermethylation in the promoter of GPC3 and RASSF1A. The progression tree presented in Figure 1 shows that promoter methylation of these two genes are early event in tumorigenesis with more tumors showing hypermethylation in RASSF1A than GPC3. In the later nodes, all of the tumors exhibit hypermethylation in these two gene promoters. Also, this algorithm singles out a subset of patients with a young age of disease onset and very aggressive tumors (no hormone receptor expression, high-grade and late stage tumors with metastatic disease) to a separate branch with no subsequent nodal development. This cluster (circled) has three methylated loci (an intermediate level of methylation) yet with advanced disease. The formation of this distinct branch corroborates our earlier discussions on breast tumors from young patients whereby DNA methylation might play a less important

role in their disease progression. The preliminary application of this clustering algorithm has proven to be effective in identifying pathways with unambiguous epigenetic and phenotypic progression.

5 Discussion

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 (Jones and Baylin, 2002). 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.

For this purpose, the heritable clustering method has been established as described in section 3. We have proposed a novel criterion to balance epigenotype and phenotype distances between tumors by combining the two using a weighting scheme. This combination is also used to choose appropriate similarities for grouping different tumors into one cluster. 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. We hence 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 clinicopathological 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 hypermethylation 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. From the results presented in section 4, this approach will select meaningful progression models and will assist in the interpretation of pathways having biological and clinical significance. To our knowledge, this method utilizes a novel and distinctive idea in clustering that is not found in the current literature.

This method can be extended to analyze other data types whereby the numerical value of the data can be continuous rather than discrete (i.e., 0 and 1). For instance, the method is well suited for modeling methylation data expressed as intensity ratios from two-color microarray experiment or transcript factor binding enrichment on gene promoters from ChIP-on-chip experiment. We do not anticipate the need to change the distance formula in NodeDiscovery procedure when applying to microarray data that are continuous in nature. However, user needs to determine the representative node centers used to cluster their data set in the process of generating the pathway trees. The extension of this method to cluster and to recapitulate progression pathway of other biological events is currently under investigation in our group.

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. SL is supported in part by NSF grant DMS-0306800 and NIH grant 1R01HG002657-01A1. CE is a recipient of the Doris Duke Distinguished Clinical

Scientist Award.

REFERENCES

Alizadeh, A. & *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **4051**, 503-511.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999) Modern information retrieval. *The ACM Press, NY.*

Bonner, R. F., Emmert-Buck, M., Cole, K., Pohida, T., Chuaqui, R., Goldstein, S., and Liotta, L. A. (1997) Laser capture microdissection: molecular analysis of tissue, *Science (Washington DC)*, **278**, 1481-1482.

Brenner, A.J. and Aldaz, C.M (1997) The genetics of sporadic breast cancer. *Prog. Clin. Biol. Res.*, **396**, 63-82.

Chen, C.M., Chen, H.L., Hsiau, TH-C, Hsiau, AH-A, Shi, H., Brock, G., Wei, S.H., Caldwell, C.W., Yan, P.S., and Huang, TH-M. (2003) Methylation target array for rapid analy-sis of CpG island hypermethylation in multiple tissue genomes. *Am J Pathol*, **163**, 37-45.

Brooks, D.R. and McLennan, D.A. (1991) Phylogeny, ecology, and behavior: a re-search program in comparative biology. *Univ. of Chicago Press.*

Cormen, T.H., Leiserson, C.E., Rivest, R.L., and Stein, C. (2001) Introduction to algorithms, second edition. *The MIT Press, Cambridge, MA.*

Datta, S. and Datta, S. (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459-466.

Dhillon,I.S., Marcotte, E.M., and Roshan, U. (2003) Diametrical clustering for identifying anticorrelated gene clusters. *Bioinformatics*, **19**, 1612-1619.

Eldridge, N., and Cracraft, J. (1980) Phylogenetic Patterns and the evolutionary process. *Columbia University Press*, New York, USA.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction. *Springer-Verlag, New York.*

Jones, P.A. and Bayline, S.B. (2002) The fundamental role of epigenetic events in cancer.

Nature Review Cancer, **3**, 415-428.

Klir, G.J. and Yuan, B. (1995) Fuzzy sets and Fuzzy logic: theory and applications. *Prentice-Hall Inc., NJ.*

Kohonen, T. (1998) Self-organization of very large document collections: state of the art. *Proc. ICANN98*, **1**, 65-74.

McPherson, K., Steel, C.M. and Dixon, J.M. (1994) ABC of breast diseases: Breast cancer epidemiology, risk factors and genetics. *Br. MeJ.*, **309**, 1003-1006.

Mitchell, T.M. (1997) Machine Learning. *The McGraw-Hill Companies Inc., Boston.*

Simin, K., Wu, H., Lu, L., Pinkel, D., Albertson, D., & *et al.* (2004) pRb inactivation in mammary cells reveals common mechanisms for tumor initiation and progression in divergent epithelia. DOI: 10.1371/ journal.pbio.0020022

Smet, F.D., Mathys, J., Marchal, K., Thijs, G., Moor, B.D., and Moreau, Y. (2002) Adaptive quality-based clustering of gene expression profile. *Bioinformatics*, **18**, 735-746.

Theodoridis, S. and Koutroumbas, K. (1998) Pattern recognition. *Academic CA.*

Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Monk, B.J., Lockhart, D.J., Burger, R.A., and Hampton, G.M. (2001) Analysis of gene ex-pression profiles in normal and neoplastic ovarian tissue samples identifies candi-date molecular markers of epithelial ovarian cancer. *Proc. Natl. Acad. Sci. USA*, **98**, 1176-1181.

Yan, P.S., Chen, C-M, Shi, H., Rahmatpanah, F., Wei, S.H., Caldwell, C.W., and Huang, T-M. (2001) Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Res*, **61**, 8375-8380.

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., and Ruzzo, W.L. (2001) Model-based clustering and data transformations for gene expression data. *Technique report UW-CSE-2001-04-02. University of Washington, WA, USA.*

Yoon, J.P., Raghavan, V., and Chakilam, V. (2001) BitCube: a three dimensional bitmap indexing for XML documents. *J. Intellingent Systems*, **17**, 241-254.

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)	$f(w, \epsilon)$
1	0.8	0.9	14	13.53	-3.21
2	0.5	0.8	14	13.32	-3.18
3	0.4	0.8	14	13.28	-3.17
4	0.3	0.8	14	13.08	-3.14
5	0.2	0.7	13	12.06	-3.12

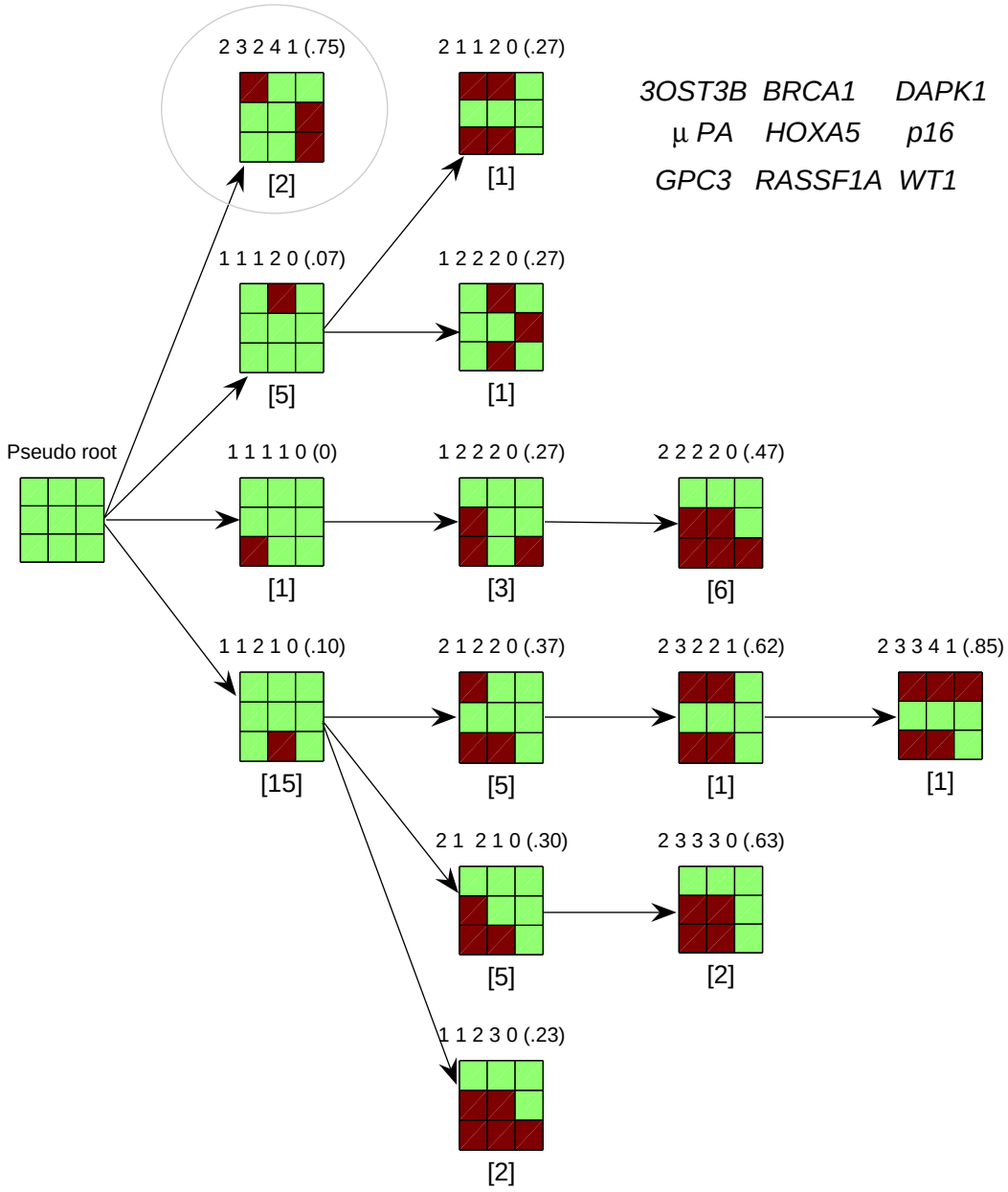
2 3 2 4 1 (.75)
[2]
2 1 1 2 0 (.27)
[1]
3OST3B BRCA1 DAPK1
μ PA HOXA5 p16
GPC3 RASSF1A WT1
1 1 1 2 0 (.07)
[5]
1 2 2 2 0 (.27)
[1]
Pseudo root
1 1 1 1 0 (0)
[1]
1 2 2 2 0 (.27)
[3]
2 2 2 2 0 (.47)
[6]
1 1 2 1 0 (.10)
[15]
2 1 2 2 0 (.37)
[5]
2 3 2 2 1 (.62)
[1]
2 3 3 4 1 (.85)
[1]
2 1 2 1 0 (.30)
[5]
2 3 3 3 0 (.63)
[2]
1 1 2 3 0 (.23)
[2]

FIGURE LEGENDS

Figure 1: Progression pathway with $w = 0.8$ and $\varepsilon = 0.9$. The methylation data analyzed here are from 50 primary breast cancers. A set of 9 gene promoter CpG islands is investigated. The gene list is shown in the upper-right 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 $\geqslant 65; 2, age < 65$), 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 bracket. The number below each node is the number of tumors within the cluster. The tree presented here conforms to strict heritability and ideal tumor phenotype progression.