Identifying Cancer Subtypes based on Somatic Mutation Profile

Sungchul Kim
POSTECH
subright@postech.ac.kr

Lee Sael
SUNY Korea & Stony brook University
sael@sunykorea.ac.kr

Hwanjo Yu
POSTECH
hwanjoyu@postech.ac.kr

ABSTRACT

Tumor stratification is one of the basic tasks in cancer genomics for a better understanding of the tumor heterogeneity and better targeted treatments. There are various biological data that can be used to stratify tumors including gene expression and sequencing data. In this work, we use the somatic mutation data. Two types of somatic mutation profiles are generated and clustered using k-means clustering with appropriate distance measures to obtain cancer subtypes for each cancer type: binary somatic mutation profile and weighted somatic mutation profile. According to the predictive power of clinical features and survival time of the identified subtypes, the binary somatic mutation profile with Jaccard distance (B-Jac) performed the best and the weighted somatic mutation profile with Euclidean distance (W-Euc) performed comparably. Both approaches performed significantly better than the typical usage of somatic mutation, i.e. the binary somatic mutation profile with Euclidean distance (B-Euc).

Categories and Subject Descriptors

J.3 [**Computer Applications**]: LIFE AND MEDICAL SCIENCES—*Medical information systems; Biology and genetics* ; H.3.1 [**Information Systems**]: Information Search and Retrieval—*Information Search and Retrieval*

General Terms

Algorithms, Theory

Keywords

Tumor startification; Somatic mutation; Clustering analysis

1. INTRODUCTION

Cancer is a complex and heterogeneous disease, mutating combinations of genes that vary greatly among patients. Recently, due to the advancements in the sequencing technology, large-scale projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have undertaken sequencing of thousands of tumor samples [7, 15]. Accordingly, with increased numbers of tumor genomics data, we are now able to stratify cancer by genomic contents which are closer to fundamental causalities and effects than the classical stratifications that are based mostly on clinical observations. However, due to its massiveness, there is an urgent need for the ability to analyze molecular defects in cancers efficiently and systematically. There already are many methods that attempt to systematically characterize cancer subtypes to understand the tumor heterogeneity and to provide more effective treatments for patients based on genomic contents [10, 14]. However, most prior attempts identify tumor subtypes based on molecular profiles of mRNA expression data which are known to suffer from systematic errors called the ”batch effect” [12, 11]. A few works provide tumor subtypes which have clinical correlation including patient survival and response to chemotherapy [1, 2].

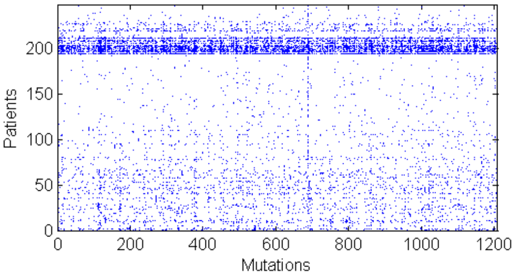


Figure 1: Plot matrix of somatic mutation profile (Chromosome 17 in UCEC data)

In this work, to identify tumor subtypes, we focus on the somatic mutation data. Somatic mutations are identified by comparing a tumor sample with a matched normal sample. Generally, somatic mutations are sparse in that proportion of mutation is small compared to the whole genome size and they are heterogeneous in that there are several types of mutation possible for a given loci in a genome. It has been observed that in several cases patients with same clinical profiles do not share even a single mutation [5, 6]. Fig 1 shows the plot matrix of somatic mutations in Chromosome 17, involved in allelic losses and inactivation of tumor suppressor genes for the development of endometrial carcinoma of the

DTMBIO'14, November 7, 2014, Shanghai, China.
Copyright 2014 ACM 978-1-4503-1275-2/14/11 ...$15.00.
http://dx.doi.org/10.1145/2665970.2665980.

uterus [9]. Only a few patients and mutations of loci have enough number of instances. The sparse and heterogeneous character of the somatic mutation makes it difficult to determine the similarity and the difference among cancer genome. To overcome this difficulty, we introduce two somatic mutation profile and distance pairs. The first approach use the Jaccard distance to correctly compute the distance among patient somatic mutation profiles represented as binary vectors. Then, we introduce a new type of somatic mutation profile weighted on the entire mutation frequency that uses Euclidean distance. For identifying tumor subtypes, we exploit k-means clustering which have been popularly used in bioinformatics due to its simplicity and interpretability. For evaluation of the proposed techniques, we compute statistical significance between identified subtypes and the clinical features, and estimate its predictive accuracy on the survival time.

2. MATERIALS AND METHODS

This work identifies tumor subtypes using k-means clustering based on two different representations of somatic mutation data with proper distance measurement. We downloaded three types of somatic mutation data: Ovarian serous cystadenocarcinoma (OV), Uterine Corpus Endometrial Carcinoma (UCEC), Lung adenocarcinoma (LUAD) from the TCGA data portal[1] [8]. The paients having less than ten mutations were discarded. Finally, there are 247 patients with 20446 genes for the UCEC cancer, 442 patients with 12432 genes for OV cancer, 517 patients with 18068 genes for LUAD cancer.

Binary somatic mutation profile and Jaccard distance: For each patient, the somatic mutation profile is represented as a binary vector where each entry indicates binary state on gene; 1 if any somatic mutation (i.e., a single-nucleotide base change and the deletion/insertion of bases) is present in the gene, and 0 otherwise. The Jaccard coefficient is a common measurement for computing the similarity between sample sets or between binary data. Given two objects, X and Y, each with n binary attributes, the Jaccard coefficient first count the number of attributes that are consistent between X and Y. Each attribute is a binary values (0 or 1). Then, the Jaccard coefficient is computed as $J_s = \frac{A}{A+B+C}$ where A is the total number of attributes where X and Y both have a value of 1, B is the total number of attributes where the attribute of X is 0 and that of Y is 1, C is the total number of attributes where the attribute of X is 1 and that of Y is 0, and D is the total number of attributes where X and Y both have a value of 0. The Jaccard distance is defined as one minus the Jaccard coefficient and is used to measure the dissimilarity between sample sets. It can also be computed as $J_d = 1 - J_s = \frac{B+C}{A+B+C}$ The Jaccard distance is not dependent of the entire dimension of the profile. Thus it reduces the sparsity problem in comparing the binary somatic mutation profiles.

Weighted somatic mutation profile: Weighted somatic mutation profile is introduced to address the heterogeneity in frequency of somatic mutations among patients and to enrich the frequently occurring mutations. Each binary values in the binary profile is converted as $x_i = \frac{f(x_i)}{x_{max}}$ where $f(x_i)$ is the mutation count in i-th gene, x_i, and x_{max} is the maximum mutation count over all genes. By this conversion, the frequently mutated genes in the profiles will have higher impact on computing distances. With binary values, it is difficult to distinguish a set of patient profile pairs in similarity/dissimilarity measure when they have same number of overlapped mutated genes regardless of which genes they are. However, alternate profile with real values that gives weight on genes make it easier to distinguish a set of patient profile pairs which have same distances in the Jaccard distance.

Clustering analysis via K-means: To identify tumor subtypes, we exploit k-means clustering technique due to its simplicity and efficiency in high dimensional data. The goal of k-means clustering is to find the positions of cluster centers, y_i, that minimize the distance from the data points to the cluster, which is formalized as $\arg\min_C \sum_{i=1}^{k} \sum_{x \in C_i} d(x, y_i)$ where C_i is the set of points that belong to cluster $i, C = \{C_1, C_2, \ldots, C_k\}$, and $d(x, y_i)$ is one of distance measures according to the data representation.

2.1 Evaluation measures

To verify the association of the tumor subtypes with clinical data, we compute their statistical significance using χ^2 statitics. We also run survival analysis on each subtype using Cox proportional hazards regression model [3] and compute log-rank statistics. The χ^2 statistic has been widely used to determine the statistical dependency of the categorical values. More precisely, it is a statistical test on whether the two pair of categorical variable is independent (null hypothesis) or that they are dependent (alternative hypothesis). The log-rank test is used to compare survival distributions (or hazard functions) of two samples that are right censored. In this test, the null hypothesis is that there is no difference between survival distributions.

3. EXPERIMENTS

In this section, we provide experimental results to verify the effectiveness of two approaches for somatic mutation profile with appropriate distance measures, 1) binary profile with the Jaccard distance (B-Jac) and 2) weighted profile with the Euclidean distance (W-Euc). We compare the outputs of the two methods with the that of a baseline method: binary profile with the Euclidean distance (B-Euc). Then, k-means clustering is used to stratify the profile distance pairs for each tumor types. In addition, clusters of small size (less than ten patients) are removed.

3.1 Association of subtypes and clinical feature

To verify the biological importance of the identified subtypes, we first conducted experiments to investigate whether the subtypes are predictive of the observed clinical data. Six clinical features are generated and used for the UCEC data. The six features are created by combining the histologic grade and the residual surgical resection (the two histologic grade times the three residual surgical resection) [4]. To measure the association between the extracted subtypes and the clinical features, χ^2 statistic is used. The result shows that the B-Euc is not appropriate for tumor stratification in the prediction of the clinical data. When k is small, the tumor subtypes based on the B-Jac are more closely associated with the clinical features. However, as k increases, the tumor subtypes based on the W-Euc are more

[1] http://cancergenome.nih.gov/

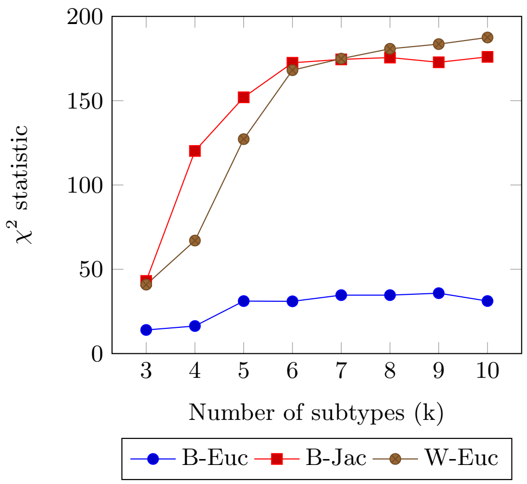


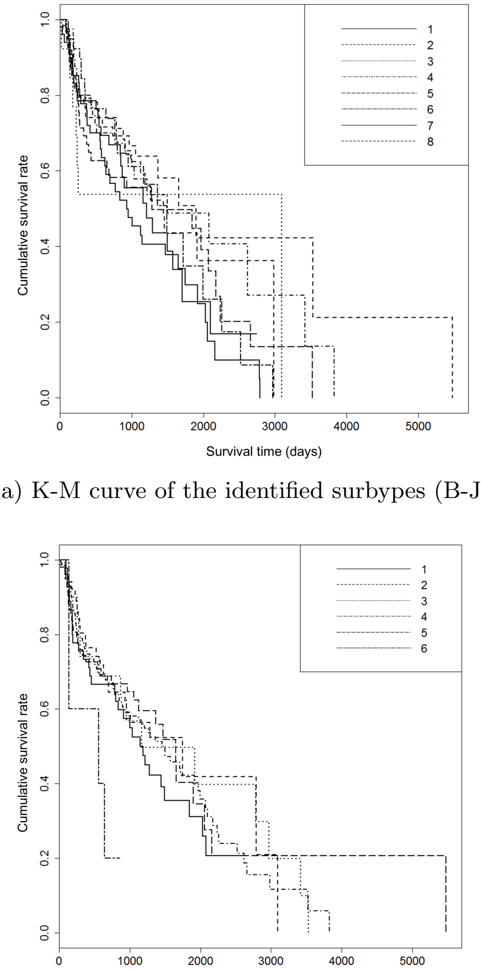
Figure 2: The association of the identified subtypes with clinical features of histologic grade and residual surgical resection (data:UCEC)

closely associated with the clinical features than the other approaches (Fig 2).

3.2 Survival analysis

To verify the relationship between the extracted subtypes and patient survival time, we fit a Cox proportional hazards regression model using R survival package [13], and conduct a log likelihood-ratio test which is used to compare the fit of two models. The full model includes the subtypes and the clinical features, and the baseline model includes the clinical features include age, clinical stage, neoplasm histologic grade, and residual surgical resection. In addition, smoking history is used for the LUAD data. In UCEC, a survival analysis is not possible due to the low mortality rates of patients.

We provide results of the best performance cases after generation of various number of subtypes for each data type. Clusters of small size (less than ten) considered as outliers and are removed from the result. According to the result, B-Euc failed to identify meaningful tumor subtypes that have clinical correlations. In contrast, the identified subtypes by other two approaches (B-Jac and W-Euc) are good indicators for patient survival time. First, in the case of ovarian cancer, the tumor subtypes identified based on the B-Jac have meaningful predictive power of the survival time (the log-rank P-value is 1.09×10^{-3} where the number of subtypes is eight). The mean survival of the most aggressive tumor subtype is approximately 27 months and that of the least aggressive tumor subtype is more than 36 months (Fig 3-(a)). The tumor subtypes identified based on the B-Jac also have meaningful predictive power of the survival time (the log-rank P-value is 2.23×10^{-3} where the number of subtypes is six). The mean survival of the most aggressive tumor subtype is approximately 15 months and the mean survival of the least aggressive tumor subtype is more than 41 months (Fig 3-(b)). Similarly, in the case of lung cancer, the tumor subtypes identified based on the B-Jac have significant predictive power of the survival time (the log-rank P-value is 1.08×10^{-7} where the number of subtypes is seven). The mean survival of the most aggressive tumor subtype is approximately 13 months and that of the least aggressive tumor subtype is more than 22 months (Fig 4-(a)). The tumor subtypes identified based on the B-Jac also have significant predictive power of the survival time (the log-rank P-value is 7.09×10^{-8} where the number of subtypes is six). The mean survival of the most aggressive tumor subtype is approximately ten months. The mean survival of the least aggressive tumor subtype is more than 20 months (Fig 4-(b)).



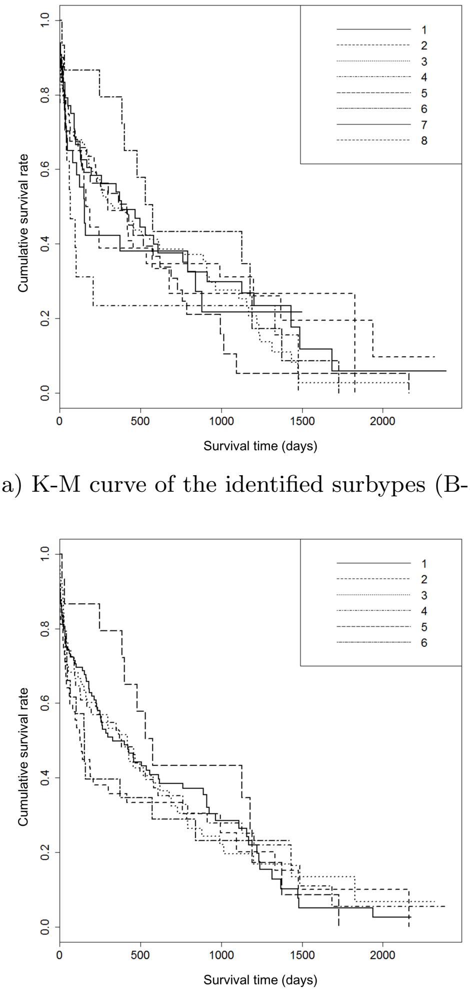
(a) K-M curve of the identified surbypes (B-Jac)

(b) K-M curve of the identified surbypes (W-Euc)

Figure 3: Kaplan-Meier survival plots of the identified subtypes with patient survival time (data:OV, Legend: subtype IDs)

4. CONCLUSION

In this work, we identify tumor subtypes based on somatic mutation profiles. To resolve the sparsity problem of the somatic mutation data, we use appropriate distance measure for the binary profile and suggest a new profile called weighted somatic mutation profile. More precisely, Jaccard distance is used for binary mutation profile, and Euclidean distance is used for somatic mutation profile weighted ac-



(a) K-M curve of the identified surbypes (B-Jac)

(b) K-M curve of the identified surbypes (W-Euc)

Figure 4: Kaplan-Meier survival plots of the identified subtypes with patient survival time (data:LUAD, Legend: subtype IDs)

cording to the mutation frequency. For a tumor stratification, k-means clustering is used due to its simplicity and interpretability. According to the results, for the typical binary somatic mutation profile, the Jaccard distance is more appropriate than the Euclidean distance. Also, the weighted somatic mutation profile with the Euclidean distance performs comparably to the binary somatic mutation profile with the Jaccard distance according to the predictive power of the clinical features and the survival time analysis.

5. ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the NRF of Korea (2013005259), MSIP under "IT Consilience Creative Program" supervised by the NIPA (NIPA-2013-H0203-13-100), the ICT R&D program of MSIP/IITP [14-824-09-014, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center)], Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education Science and Technology (No. 2012M3C4A7033344), and the Industrial Core Technology Development Program (10049079, Development of Mining core technology exploiting personal big data) funded by the Ministry of Trade Industry and Energy (MOTIE, Korea).

6. REFERENCES

[1] Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011.

[2] T. cancer genome atlas network. Comprehensive molecular characterization of human colon and rectal cancer, 2012.

[3] J. Fan and R. Li. Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*, 30(1):74–99, 2002.

[4] P. Hermanek and C. Wittekind. Residual tumor (r) classification and prognosis. *Semin Surg Oncol*, 10(1):12–20, 1994.

[5] D. C. Koboldt et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, Sept. 2012.

[6] M. S. Lawrence et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, pages 1–5, June 2013.

[7] E. R. Mardis. Genome sequencing and cancer. *Current Opinion in Genetics & Development*, 22:245–250, June 2012.

[8] T. C. G. A. R. Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, Sept. 2013.

[9] A. Okamoto, Y. Sameshima, Y. Yamada, S. Teshima, Y. Terashima, M. Terada, and J. Yokota.

[10] A. Prat and C. M. Perou. Deconstructing the molecular portraits of breast cancer. *Molecular Oncology*, 5(1):5–23, Feb. 2011.

[11] J. S. Reis-Filho and L. Pusztai. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*, 378(9805):1812 – 1823, 2011.

[12] A. Sommer, F. Hilpert, and N. Arnold. Gene Expression Profiling of Epithelial Ovarian Cancer. *Current Genomics*, 7(2):115–135, Apr. 2006.

[13] T. M. Therneau. *A Package for Survival Analysis in S*, 2014. R package version 2.37-7.

[14] K. Wang, J. Kan, S. T. Yuen, S. T. Shi, K. M. Chu, S. Law, T. L. Chan, Z. Kan, A. S. Y. Chan, and W. Y. Tsui. Exome sequencing identifies frequent mutation of arid1a in molecular subtypes of gastric cancer. *Nature Genetics*, (12):1219–1223, 2011.

[15] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin. Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics*, 14(10):703–718, Sept. 2013.