Survival-Expression Map and Essential Forms of Survival-Expression Relations for Genes

Yajing Chen[1], Shikui Tu[1], and Lei Xu[1,2](✉)

[1] Center for Cognitive Machines and Computational Health, and Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
{cyj907,tushikui,leixu}@sjtu.edu.cn

[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

Abstract. The relation between survival and gene expression has been investigated in many studies. Some used a univariate Cox model to detect genes with expression significantly related to survival. Some built a multivariate Cox model to analyze the influence of multiple genes on death risk. The original Cox model assumes a linear relation between survival and expression. But some evidence implied the existence of non-linear relation. Whether the survival-expression relations for different genes share some particular forms remain unknown. Here, we clustered the survival-expression (S-E) relations by k-means. We also developed a survival-expression (S-E) map to display the S-E relations for each cluster and summarized four essential forms of relations. We believe that the four essential S-E forms might assist the discovery of therapeutic targets and enhance the understanding of mechanisms in cancers.

Keywords: Cox regression · Spline · Survival-expression map · Essential survival-expression relation

1 Introduction

Many studies performed survival analysis to investigate the relation between survival and gene expression. Some built a multivariate Cox model to predict survival from the expression of several genes [2,9–11]. Some aimed to find prognostic biomarkers and used a univariate Cox model to discover the significant survival-related genes [2,3,6]. As the Cox model is linear, these studies implicitly assumed a linear survival-expression (S-E) relation. Others applied non-linear models and found evidence of non-linear relations [5,7,8]. However, whether the S-E relations for different genes share some particular forms and what these forms look like remain unknown.

Here, we aimed to find out the essential forms of S-E relations and clarify their patterns. After computing the survival rates for each gene using Cox regression with natural splines, we applied K-means on the resulting S-E relations, namely, the survival rates arranged by increasing expression. We proposed

© Springer International Publishing AG 2017
Y. Sun et al. (Eds.): IScIDE 2017, LNCS 10559, pp. 641–649, 2017.
DOI: 10.1007/978-3-319-67777-4_58

a survival-expression (S-E) map to display S-E relations for multiple genes simultaneously. It is a heat map whose rows are genes and columns are samples. Each row display the S-E relation for a gene. We analyzed the gene expression data for breast cancer, lung adenocarcinoma, lung squamous cell carcinoma. For all of them, the S-E relations can be clustered into four groups. The S-E maps for the four clusters revealed the differences in changing rates and tendencies of S-E relations. We claimed that each cluster represents an essential form of S-E relations, which might play a distinct role in biological systems. Careful investigation in the S-E maps uncovered the variation in the proportion and detailed configurations of the essential S-E forms for different cancers, indicating the heterogeneity of cancers. Surprisingly, even though lung adenocarcinoma and lung squamous cell carcinoma are both non-small-cell lung carcinoma, their S-E maps display significant difference.

In conclusion, we discovered four essential S-E forms by S-E maps for different cancers. We believe that the discovery might assist the finding of therapeutic targets to improve prognosis.

2 Methods

2.1 Cox Regression with Natural Splines

To analyze the relation between survival and gene expression, we applied Cox regression with natural splines (degree of freedom is 2). Cox regression model is a proportional hazard model which defines hazard rate λ as:

$$\lambda(t|x) = \lambda_0(t) exp(\beta^T x) \tag{1}$$

where $\lambda(t|x)$ represents the hazard rate at time t with covariate x, $\lambda_0(t)$ is the baseline hazard function which is irrelevant to x. Cox regression does not have to explicitly specify the form of baseline hazard function $\lambda_0(t)$, as it is cancelled out during the computation of β. The survival rate $S(t|x)$ is defined as:

$$S(t|x) = exp(-\mathit{\Lambda}(t|x)) \tag{2}$$

where $\mathit{\Lambda}(t|x)$ is the cumulative hazard function:

$$\mathit{\Lambda}(t|x) = \int_0^t \lambda(T|x)\, dT \tag{3}$$

However, the computation of survival rate requires us to specify the baseline hazard $\lambda_0(t)$. Thus, we define cumulative baseline hazard $\mathit{\Lambda}_0(t)$ as the Nelson-Aalen estimator [1]:

$$\mathit{\Lambda}_0(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{4}$$

The $\beta^T x$ in the hazard function is replaced by the natural splines $s(x)$. The hazard function becomes:

$$\mathit{\Lambda}(t|x) = exp(s(x))\mathit{\Lambda}_0(t) \tag{5}$$

The algorithm is implemented in the R language. The Cox model is built by the *coxph* function in the *survival* package, and natural splines is given in the *ns* function in the *splines* package. Finally, we obtained the survival rates for each gene and patient.

2.2 Survival-Expression (S-E) Map

The S-E map is a heat map that combines multiple S-E relations as its rows. Each S-E relation is the survivals sorted by increasing expression. Thus, the survivals in the same column but different rows of the S-E map might be computed for different patients. The colors indicate the magnitude of survival rates, where green represents low survival and red represents high survival. Rows of the S-E map are rearranged according to the results of hierarchical clustering, so as to display similar patterns in the S-E relations.

3 Experiments and Results

3.1 Data Preprocessing

We analyzed three cancers in this study, i.e., breast cancer (BRCA), lung adenocarcinoma (LUAD) and lung squamous cell cancer (LUSC). They were all RNA-sequencing expression profiles downloaded from the UCSC Xena browser (https://xenabrowser.net/datapages/, [4]). The RNA-sequencing expression data revealed the gene activity inside specific tissues. They were generated by the Cancer Genome Atlas (TCGA) research network (http://cancergenome.nih.gov/). The survival data include information about the overall and disease-free survival for each patient, which can also be obtained from TCGA. We selected breast cancer (BRCA) because it has the largest sample size in the TCGA hub. We selected lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) because they have large sample size and both of them are subtypes of non-small-cell lung cancer. The statistics of these data sets are listed in Table 1. The number of genes in each data set is 20530.

Table 1. Statistics of the three data sets.

Cancer	Abbreviation	Sample size
Lung squamous cell carcinoma	LUSC	494
Lung adenocarcinoma	LUAD	503
Breast cancer	BRCA	1080

3.2 Computation of the Overall Survival

We applied Cox regression with natural splines ($df = 2$) to analyze the influence of gene expression on survival. As the prognosis of BRCA is better than LUSC and LUAD, we computed the 5-year overall survival for LUAD and LUSC, and the 10-year overall survival for BRCA. Then, the survival rates were sorted according to the increasing expression for each gene, resulting in a survival-expression (S-E) relation. We performed likelihood ratio tests to evaluate the reliability of the fitting Cox models. Genes with small p-values were thought to display reliable S-E relations, which were selected for further investigation. The resulting p-value distributions for the three cancers are different. We chose the genes with $p < 0.001$ for LUAD, $p < 0.02$ for LUSC and $p < 0.01$ for BRCA, so that the number of selected genes ranged between 500 and 1000, i.e., 651, 445 and 793 respectively for LUAD, LUSC and BRCA.

3.3 Discovery of Four Essential Forms of Survival-Expression Relations

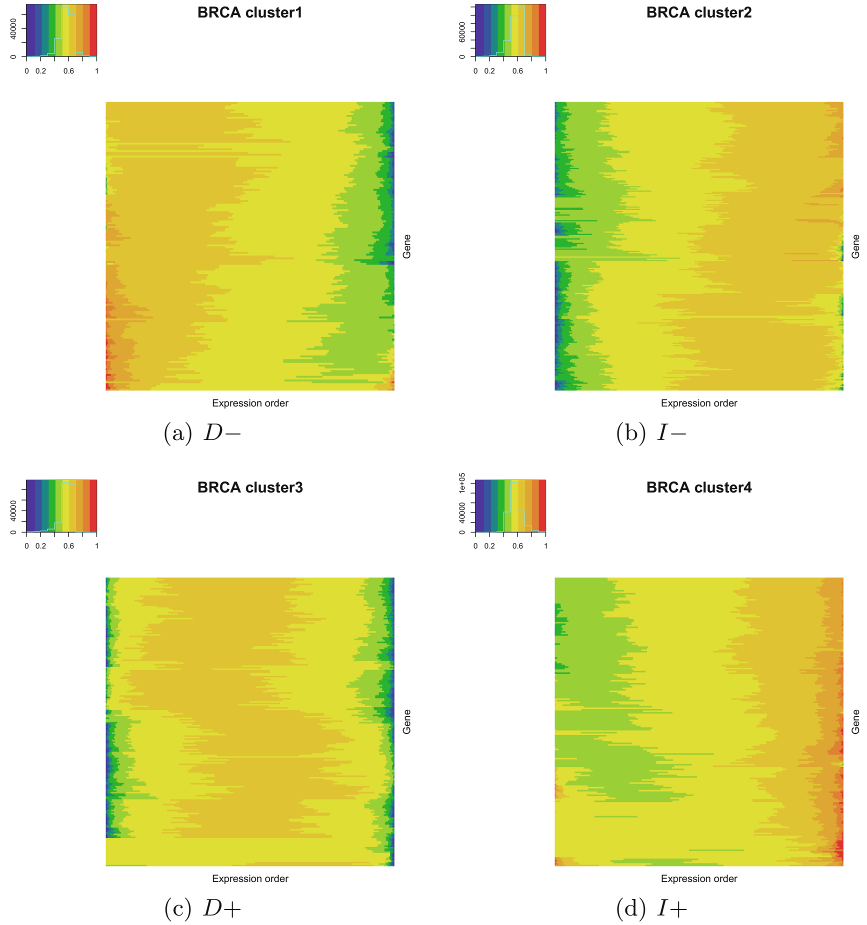
After applying k-means with $k = 4$ on the S-E relations, we made the S-E maps for the four clusters. Figs. 1, 2 and 3 display the results for BRCA, LUAD and LUSC respectively. The four clusters for each cancer mainly show two patterns. One is increasing survival with increasing expression (e.g. Fig. 1(b), (d)), while the other is decreasing survival with increasing expression (e.g. Fig. 1(a), (c)), each of which covers two clusters. The two clusters with the same changing tendency display different changing rates at different expression. One has a fast changing rate at low expression (e.g. Fig. 1(a), (b)), while the other has a fast changing rate at high expression (e.g. Fig. 1(c), (d)).

We claim that these four clusters represent four essential forms of S-E relations in biological systems. We named them as $I+$, $I-$, $D+$ and $D-$ according to the changing rates and tendencies. Details are shown in Table 2.

Table 2. Characteristics of four essential S-E forms.

S-E form	Changing tendency	Changing rate
$I+$	Increasing	Slow at low expression, fast at high expression
$I-$	Increasing	Fast at low expression, low at high expression
$D+$	Decreasing	Slow at low expression, fast at high expression
$D-$	Decreasing	Fast at low expression, low at high expression

After careful investigation in the four S-E forms, we found some bell-shape S-E relations. For example, at the bottom of Fig. 1(a), the colors of some S-E relations display patterns as orange-yellow-green-yellow, indicating the changing tendencies as high-low-high. Similar patterns can also be found in Fig. 1(b), (c) and (d). The opposite tendency in the tail of a bell-shape S-E relation seems to locate in the region where survivals change slowly. For example, at the bottom of Fig. 1(a) for $D-$, the bell-shape S-E relations show a slight increase at the high expression region, where survivals change slowly. At the bottom of Fig. 1(d), though the major tendency is increasing, we can see a slight decreasing at the low expression region, where survivals also change slowly. Similar results are shown in Figs. 2 and 3.



(a) $D-$

(b) $I-$

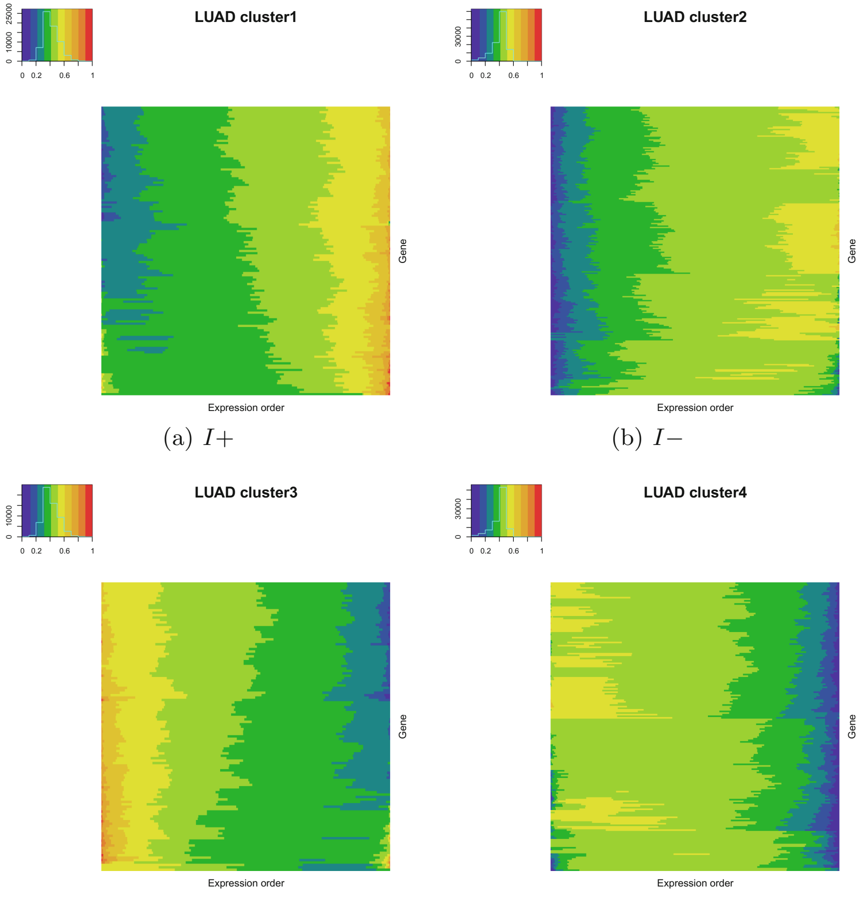
(c) $D+$

(d) $I+$

Fig. 1. The four essential S-E forms for breast cancer (Color figure online)

3.4 Analysis of the S-E Maps Between Cancers

The above section showed the similarities between the same S-E forms for three cancers. In this section, we focused on difference in the S-E forms between different diseases. We mainly compared the S-E maps for LUAD and LUSC, which are both non-small-cell lung cancer. First of all, the color configurations of the S-E maps for LUAD and LUSC are different. For LUAD, the major colors are green and blue, while for LUSC, the major colors become yellow and green. The



(a) $I+$

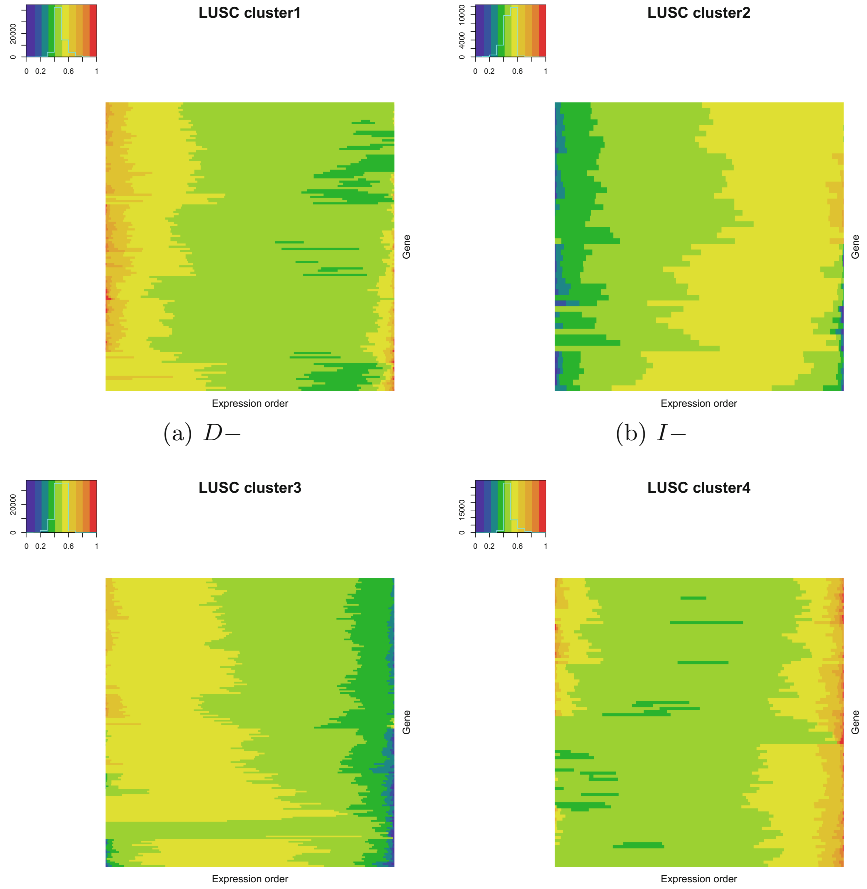
(b) $I-$

(c) $D-$

(d) $D+$

Fig. 2. The four essential S-E forms for lung adenocarcinoma (Color figure online)

difference in colors indicates that the patients with LUAD have a lower 5-year overall survival rate than LUSC. Second, the proportion of bell-shape S-E relations in LUSC are larger than LUAD. In LUSC, they cover almost half of the $D-$, $I-$ and $I+$ forms (Fig. 3(a), (b) and (d)). But we cannot see such a large proportion in LUAD. Third, the ratios of each S-E form between LUAD and LUSC are different, as shown in Table 3. However, the difference is not significant by chi-square test ($p = 0.2133$). These results indicate that the mechanisms and characteristics of LUSC and LUAD might be disease-specific, though they both belong to non-small-cell lung cancer.



(a) $D-$ (b) $I-$

(c) $D+$ (d) $I+$

Fig. 3. The four essential S-E forms for lung squamous cell cancer (Color figure online)

Table 3. The number of genes in each S-E form for LUSC and LUAD

Cancer	I+	I−	D+	D−
LUSC	94 (21%)	51 (11%)	167 (38%)	133 (30%)
LUAD	131 (20%)	221 (34%)	180 (28%)	119 (18%)

4 Conclusion

The relation between survival and gene expression is implicitly assumed to be linear in many studies. Univariate Cox regression is often applied to evaluate the significance of the association between survival and gene expression. Some researchers performed Cox regression with splines to analyze the non-linear survival-expression relation, but no one has ever bothered to reveal the essential forms of survival-expression relations in the biological systems. In this study, after obtaining the survival-expression (S-E) relations by Cox regression with natural splines, we clustered the S-E relations into 4 groups and drew the S-E maps for each group. The S-E maps showed that the four clusters had their special S-E configurations, which might demonstrate the essential forms of S-E relations. Different S-E forms have various changing rates and tendencies in survival versus expression. Bell-shape S-E relations exist in each form and the opposite tendencies tend to appear in tail of the expression region where survivals change slowly. Comparisons between S-E maps for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) showed the difference in their 5-year overall survival rates. The coverage of bell-shape S-E relations is larger in LUSC. The proportion of the four essential S-E forms are not significantly different between these two types of lung cancers. The configurations of the four essential S-E forms seem to be disease-specific, which reflects the complexity of the biological system and the heterogeneity of cancers.

We believe that the essential S-E forms will provide more information for analyzing the prognostic roles of genes and understanding the mechanisms of cancers.

Acknowledgments. This work was supported by the Zhi-Yuan chair professorship start-up grant (WF220103010) from Shanghai Jiao Tong University.

References

1. Aalen, O.: Nonparametric inference for a family of counting processes. Ann. Stat. **6**, 701–726 (1978)
2. Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.A., Chinnaiyan, A.M.: Delineation of prognostic biomarkers in prostate cancer. Nature **412**(6849), 822–826 (2001)
3. Diamandis, E.P., Scorilas, A., Fracchioli, S., Van Gramberen, M., De Bruijn, H., Henrik, A., Soosaipillai, A., Grass, L., Yousef, G.M., Stenman, U.H., et al.: Human kallikrein 6 (hK6): a new potential serum biomarker for diagnosis and prognosis of ovarian carcinoma. J. Clin. Oncol. **21**(6), 1035–1043 (2003)
4. Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., Haussler, D., Zhu, J.: The UCSC cancer genomics browser: update 2015. Nucleic Acids Res. **43**, D812–D817 (2014)
5. Li, H., Luan, Y.: Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. Bioinformatics **21**(10), 2403–2409 (2005)
6. Luo, L.Y., Katsaros, D., Scorilas, A., Fracchioli, S., Bellino, R., van Gramberen, M., de Bruijn, H., Henrik, A., Stenman, U.H., Massobrio, M., et al.: The serum concentration of human kallikrein 10 represents a novel biomarker for ovarian cancer diagnosis and prognosis. Cancer Res. **63**(4), 807–811 (2003)
7. Rini, B., Goddard, A., Knezevic, D., Maddala, T., Zhou, M., Aydin, H., Campbell, S., Elson, P., Koscielny, S., Lopatin, M., et al.: A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies. Lancet Oncol. **16**(6), 676–685 (2015)

8. Rockova, V., Abbas, S., Wouters, B.J., Erpelinck, C.A., Beverloo, H.B., Delwel, R., van Putten, W.L., Löwenberg, B., Valk, P.J.: Risk stratification of intermediate-risk acute myeloid leukemia: integrative analysis of a multitude of gene mutation and gene expression markers. Blood **118**(4), 1069–1076 (2011)
9. Sotiriou, C., Neo, S.Y., McShane, L.M., Korn, E.L., Long, P.M., Jazaeri, A., Martiat, P., Fox, S.B., Harris, A.L., Liu, E.T.: Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc. Nat. Acad. Sci. **100**(18), 10393–10398 (2003)
10. Van't Veer, L.J., Dai, H., Van De Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**(6871), 530–536 (2002)
11. Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., et al.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet **365**(9460), 671–679 (2005)