

# Weakest Link Models for Detecting Small Groups of Genes to Predict Lung Cancer Survival

Thomas J. Richards  
Simmons Center for Interstitial Lung  
Diseases  
NW 628 MUH, 3459 Fifth Ave  
Pittsburgh PA 15213  
1-412-692-2145  
richardstj2@upmc.edu

Roger S. Day  
University of Pittsburgh Cancer  
Institute  
201 N. Craig Street, Suite 325  
Pittsburgh PA 15213  
1-412-383-1537  
day@upci.pitt.edu

Naftali Kaminski  
Simmons Center for Interstitial Lung  
Diseases  
NW 628 MUH, 3459 Fifth Ave  
Pittsburgh PA 15213  
1-412-647-3156  
kaminskin@upmc.edu

## ABSTRACT

In this paper, we describe a meta-analysis of four lung cancer survival data sets, with the goal of detecting gene interactions affecting survival. The analysis uses the weakest link family of statistical models, a novel class of models for survival analysis that incorporate a useful type of biological interaction among quantitative gene expression levels. A weakest link type of biological interaction posits a curve in gene expression space, the “Curve of Optimal Use (COU),” such that all genes necessary for impacting patient survival work together efficiently only if the expression levels are related as specified by the COU. Otherwise, one gene, the “weakest link” gene, will be least optimally set, relative to the other necessary genes. In applying weakest link models to lung cancer data sets, we specified substantively meaningful classes of genes included in all four lung cancer data sets and searched for dyads of genes that interact biologically to impact the hazard function for death. Testing for weakest link effects of gene expression levels on survival is also possible for triads of genes, and for an arbitrary number of genes. We detected biological interactions among genes implicated in the cell cycle and genes that encode matrix metalloproteinases (MMPs). These interactions are not detectable as conventional statistical interactions. Performance of the method will be assessed using cross-validation on the most predictive dyads of genes.

## Categories and Subject Descriptors

J.3 [Life and Medical Sciences]: *Biology and genetics.*

## General Terms

Algorithms, Theory.

## Keywords

data analysis; biological interaction; statistical interaction; joint effects; statistical model; microarray; nonparametric.methods; lung cancer; meta-analysis.

## 1. INTRODUCTION

Lung cancer accounts for the majority of cancer deaths in the industrialized world. The ability to detect a small number of prognostic markers for lung cancer would have tremendous public health impact, especially if such markers could be used to design more effective treatments. DNA microarrays facilitate the detection of markers for lung cancer prognosis, but only when accompanied by appropriate techniques of data analysis.

Data analysis aimed at relating microarray data to prognosis requires statistical models that incorporate biologically meaningful assumptions about the *joint effects* of gene expression levels. One-gene-at-a-time techniques that aim to detect effects via thousands of single-gene hypothesis tests with multiplicity adjustment may lack statistical power to detect cases in which gene expression levels *jointly*, but not individually, impact prognosis. Such ‘interactions without main effects’ are *biological* interactions and may not manifest as conventional *statistical* interactions.

In this work we describe the application to microarray analysis of the weakest link (WL) family of statistical models that incorporate biological interactions in survival analysis. Conventional statistical methods assume that differential expression of any gene *always* impacts the hazard function. More plausibly, there exists in high-dimensional expression space a curve along which genes that jointly impact survival are *simultaneously* over- or under-expressed *to sufficient degrees*. We call this unknowable curve in gene expression space the ‘curve of optimal use,’ or COU. All genes that are necessary to affect patient survival are efficiently utilized whenever gene expression profiles lie on the COU. For a gene expression profile not on the COU, a WL model posits that one gene, the ‘weakest link gene’, will be least optimally expressed, relative to the others. Only by identifying and manipulating the weakest link gene can Nature or man increase or decrease the hazard of death, the instantaneous probability of lung cancer death, conditional upon being at risk. The identity of the weakest link gene changes according to the location in gene expression space.

To facilitate rapid screening we propose a nonparametric estimation of the COU by applying quantile matching to the empirical distributions of gene expression profiles. Profile likelihood is used for maximum likelihood estimation.

## 2. METHODS

### 2.1 Data Pre-Processing

Genes included in all four of the contest data sets were matched to their LocusLink identifiers (IDs). There were 2000 unique LocusLink IDs present in all 4 data sets. Data sets were assembled by concatenation. Because data from spotted arrays (Ontario and Stanford data sets) were expressed in terms of ratios (disease to normal on the same arrays), data from Affymetrix chips were also converted to ratios. The denominator used for a given gene was the geometric mean of expression levels for the gene from all normal samples in a data set. All data were log2-transformed for analysis.

The 2000 genes present in all 4 data sets were placed into substantive classes using NIAID's DAVID (Database for Annotation, Visualization, and Integrated Discovery; <http://david.niaid.nih.gov/david/>). The substantive classes of genes used, with their abbreviations, are as follows:

- Cell Cycle (CELL; 24 LocusLink IDs)
- Apoptosis (AP; 12 LocusLink IDs)
- Extracellular Matrix Proteins (ECM; 18 LocusLink IDs)
- Matrix Metalloproteinases (MMP; 10 LocusLink IDs)
- WNT (WNT; 11 LocusLink IDs)

### 2.2 Statistical Methods

#### 2.2.1 Weakest Link Statistical Models

We seek a statistical model into which prognostic factors can be easily incorporated, without changing the identity of the weakest link gene for any patient. The problem can be simplified by transforming from high-dimensional gene expression space to the unit square, by using the empirical CDF's of two gene expression measures at a time. A simple parameterization for the COU can be devised that facilitates both maximum likelihood estimation and model selection. Consider a data set with two quantitative gene expression measures  $\{(X_{i1}, X_{i2}) : i = 1, \dots, n\}$ . Denote the respective CDF's of  $x_1$  and  $x_2$  by  $F_1(x_1)$  and  $F_2(x_2)$  and their empirical counterparts as  $\hat{F}_1(x_1)$  and  $\hat{F}_2(x_2)$ . For ease of reference, and to emphasize the fact that the range of the  $\hat{F}_j(x_j)$ 's is the unit interval, the symbols  $p_1$  and  $p_2$  will be used in place of  $\hat{F}_1(x_1)$  and  $\hat{F}_2(x_2)$ , respectively. A joint

effect of  $x_1$  and  $x_2$  on the hazard rate can be posited in terms of either quantiles in  $x_1$ - $x_2$  space or percentiles in  $p_1$ - $p_2$  space; the two formulations are equivalent. The weakest link survival model, which generalizes the Cox proportional hazards model to effects that are not compensatory, can be stated for a dyad of two genes, as follows:

$$I_i(t) = I_0(t) \exp[\mathbf{b} \mathbf{r}_i].$$

The gene expression levels enter the model via the quantity  $?_i$ , defined as one of the following:

$$?_i = \begin{cases} p_1 & \text{if } p_2 > \mathbf{f}_\gamma(p_1) \\ \mathbf{f}_\gamma^{-1}(p_2) & \text{if } p_2 < \mathbf{f}_\gamma(p_1) \end{cases};$$

$$?_i = \begin{cases} \mathbf{f}_\gamma^{-1}(p_2) & \text{if } p_2 > \mathbf{f}_\gamma(p_1) \\ p_1 & \text{if } p_2 < \mathbf{f}_\gamma(p_1) \end{cases};$$

$$?_i = \begin{cases} p_1 & \text{if } p_2 > \mathbf{f}_\gamma(1-p_1) \\ 1-\mathbf{f}_\gamma^{-1}(p_2) & \text{if } p_2 < \mathbf{f}_\gamma(1-p_1) \end{cases}; \text{ or}$$

$$?_i = \begin{cases} 1-\mathbf{f}_\gamma^{-1}(p_2) & \text{if } p_2 > \mathbf{f}_\gamma(1-p_1) \\ p_1 & \text{if } p_2 < \mathbf{f}_\gamma(1-p_1) \end{cases},$$

where the mapping  $\mathbf{f}_\gamma : [0,1] \rightarrow [0,1]$  is defined as

$$\mathbf{f}_\gamma(p) = 1 - F[F^{-1}(1-p) - \Delta],$$

and  $F$  is a prespecified symmetric function on the unit interval. In these models the curve of optimal use (COU) is specified as  $p_2 = \mathbf{f}_\gamma(p_1)$  or  $p_2 = \mathbf{f}_\gamma(1-p_1)$ . The functions  $p_2 = \mathbf{f}_\gamma(p_1)$  and  $p_2 = \mathbf{f}_\gamma(1-p_1)$  are graphed in Figures 1 and 2, respectively, for varying  $\Delta$ .

Each choice among the four definitions for the covariate  $p_i$  specifies a partition of gene expression space into regions, according to which gene is the weakest link. The four choices of  $p_i$  can be succinctly expressed, as  $\min\{p_1, p_2\}$ ,  $\max\{p_1, p_2\}$ ,  $\max\{p_1, 1 - p_2\}$ , and  $\min\{p_1, 1 - p_2\}$ , respectively. Using standard notation for probabilities,  $q_j = 1 - p_j$ , the four cases are assigned shorthand names,  $\min p_1 p_2$ ,  $\max p_1 p_2$ ,  $\max p_1 q_2$ , and  $\min p_1 q_2$ , respectively. The putative COU's and contours of constant

hazard for these four models are depicted in Figure 3. In the minp1p2 and maxp1q2 models (maxp1p2 and minp1q2 models),  $x_1$  is the weakest link above (below) the COU line and  $x_2$  is the weakest link below (above) the COU line. The direction of increasing hazard along the COU line thus depends upon only the sign of the coefficient  $\beta$ .

With a single  $\beta$  coefficient, the four models incorporate the essential weakest link properties. There is a COU: projection is done onto a specified line in  $p_1$ - $p_2$  space. In a given region of gene expression space, the hazard rate depends upon a single gene, the weakest link gene. In the minp1p2 and maxp1p2 models the genes impact the hazard rate in the same direction; in the maxp1q2 and minp1q2 models the genes impact the hazard in opposite directions.

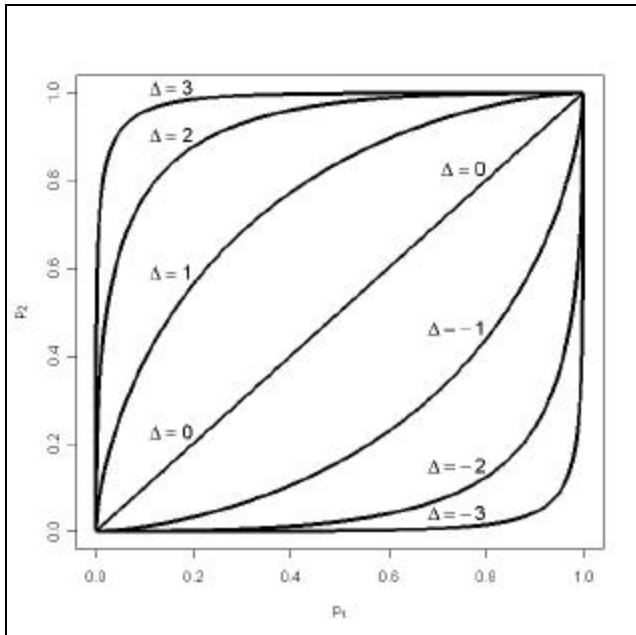


Figure 1: The function  $p_2 = f_{\beta}^{\gamma}(p_1)$  for varying D

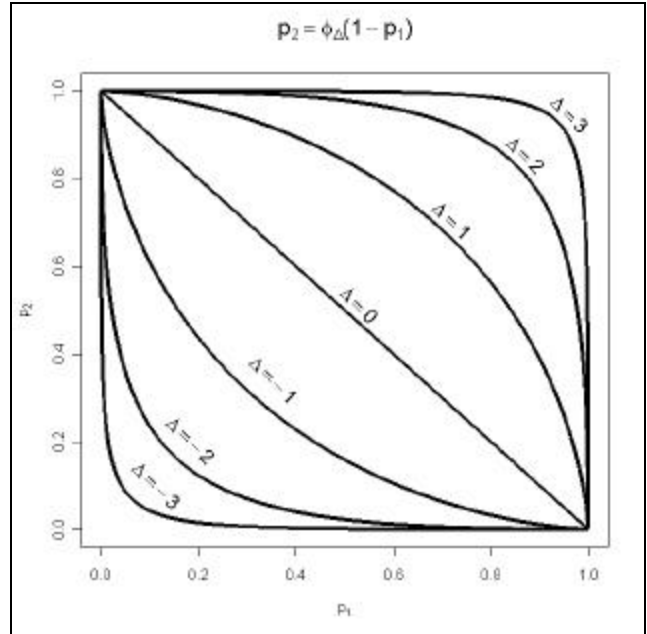


Figure 2: The function  $p_2 = \phi_{\Delta}(1 - p_1)$  for varying D

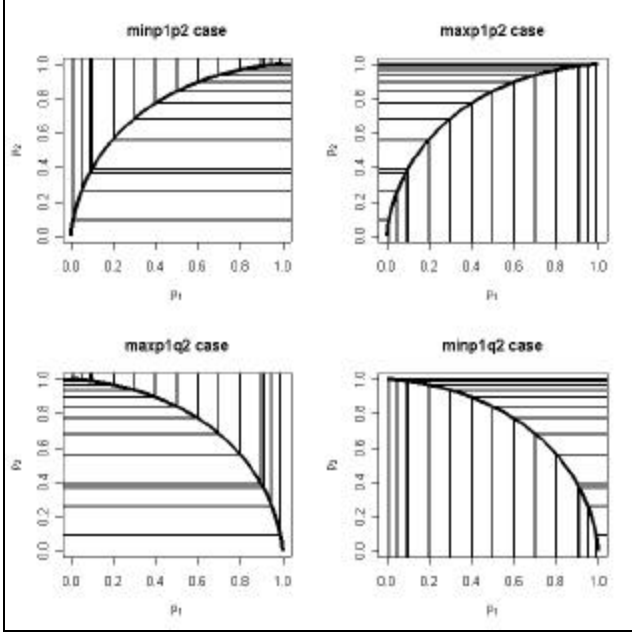
For fixed  $p$ ,  $\phi(p)$  is an increasing function of  $\Delta$ . The four expressions for the covariate  $\rho_i$  can be simply expressed as

$$\min\{p_1, \mathbf{f}_{\gamma}^{-1}(p_2)\},$$

$$\max\{p_1, \mathbf{f}_{\gamma}^{-1}(p_2)\},$$

$$\max\{p_1, 1 - \mathbf{f}_{\gamma}^{-1}(p_2)\},$$

$$\text{and } \min\{p_1, 1 - \mathbf{f}_{\gamma}^{-1}(p_2)\}.$$



**Figure 3: WL model COU and hazard contours**

A further generalization in the development of the WL model involves invariance to the order in which genes are entered into the model. To ensure that the estimated response probabilities depend only upon which genes are involved and not upon the order in which those variables enter into the model, the following schema will be used to define the covariate  $\rho_i$  for the two gene expression levels:

- I. Compute the empirical CDFs  $p_1^*$  and  $p_2^*$  using the expression of gene 1 to compute  $p_1^*$ ;
- II. Compute the empirical CDFs  $p_1^{**}$  and  $p_2^{**}$  using the expression of gene 2 to compute  $p_1^{**}$ ;
- III. In place of  $p_1$  and  $p_2$  in the definition of  $\rho_i$  above, substitute the averages  $\frac{p_1^* + p_1^{**}}{2}$  and  $\frac{p_2^* + p_2^{**}}{2}$ , respectively.
- IV. Perform a Bonferroni-by-4 correction for multiple testing;

As the parameter  $\Delta$  tends to infinity, the log-likelihood of each of the four WL models defined above tends to the maximized log-partial likelihood of a univariate Cox PH model based on a reduced data set. The observations in the reduced data set are determined by the behavior of the COU as  $\Delta$  tends to infinity. Figures 1 and 3 together show that as  $\Delta$  tends to  $+\infty$ , the COU for both the minp1p2 and maxp1p2 models tends to the Northwest corner of the unit square; as  $\Delta$  tends to  $-\infty$ , the COU for both the minp1p2 and maxp1p2 models tends to the Southeast

corner of the unit square. Figures 2 and 3 together show that as  $\Delta$  tends to  $+\infty$ , the COU for both the maxp1q2 and minp1q2 models tends to the Northeast corner of the unit square; as  $\Delta$  tends to  $-\infty$ , the COU for both the minp1p2 and maxp1p2 models tends to the Southwest corner of the unit square. The minp1p2 model thus tends to a univariate Cox PH model with covariate the average of the empirical CDF's of the two gene expression levels. This follows from the behavior of  $\min\{p_1, p_2\}$  in Figure 3, as  $\Delta$  tends to  $-\infty$ . Similar statements apply to all four of the WL models as  $\Delta$  tends to positive or negative infinity.

### 2.2.2 Fitting Weakest Link Models

The WL model can be fitted to survival data by computing the covariate  $\rho_i$  and applying the Cox proportional hazards model (Cox, 1972), as implemented in standard software for survival analysis (Therneau and Grambsch, 2000). A profile likelihood estimation technique can be used for the survival model, but with the partial likelihood, rather than the full likelihood. It can be that the log-partial likelihood for a  $PWL_1$  model is constructed by stitching together pieces of ordinary Cox PH regression log-partial likelihoods: parameter vectors on the same local PH slice of the WL model log-partial likelihood assign the same weakest link for each patient in the data set. As a maximum likelihood algorithm bounces around parameter space, landing at iteration  $k$  on parameter vector  $q^{(k)}$ , it can be determined whether the weakest link log-partial likelihood assumes a (local) maximum value within the PH log-partial likelihood slice on which the current parameter vector  $q^{(k)}$  lies. This can be done by fitting a specified standard Cox PH model that can be easily derived from the data set on hand.

## 3. RESULTS

We fitted the pairwise weakest link models to all pairs of LocusLink IDs, as specified in Table 1:

**Table 1: Classes of Models fitted in Meta Analysis**

Class 1	Class 2	# Models	# signif
Cell Cycle	Apoptosis	288	34
Cell Cycle	ECM	432	57
Cell Cycle	MMP	240	25
Cell Cycle	WNT	264	34
Apoptosis	ECM	216	25
Apoptosis	MMP	120	15
Apoptosis	WNT	132	16
ECM	MMP	180	25
ECM	WNT	198	25

MMP	WNT	110	10
-----	-----	-----	----

There are a total of 266 statistically significant pairs, out of 2180 substantively interesting models using marker pairs. Of these 266 pairs, 65 were associated with increasing risk ( $\beta > 0$ ), and 201 were associated with decreasing risk ( $\beta < 0$ ). These statistically significant pairs of markers were distributed among the 4 model types as follows: 67 minp1p2, 89 maxp1p2, 41 maxp1q2, and 69 minp1q2. The joint results for these two descriptions is as follows:

**Table 2: Statistically Significant Model Types and Directions**

Model Type	$\beta > 0$	$\beta < 0$
Minp1p2	1	66
Maxp1p2	0	89
Maxp1q2	4	37
Minp1q2	60	9

There were 155 (= 66 + 89) LocusLink ID pairs for which the effects were protective and in the same direction, but only one pair (LLid = 1280, 1000) for which the effect was to increase the hazard. The symbols for this pair for increased risk are COL2A1 (in ECM) for LLid = 1280 and CDH2 (in WNT) for LLid = 1000. The 15 most significant pairs found to be associated with decreased risk, in the same direction, are shown in table below.

**Table 3: 15 Most significant Dyads, same direction**

Class 1	Symbol 1	Class 2	Symbol 2	model
CELL	EP300	ECM	COL5A1	minp1p2
AP	TNFRSF6	ECM	COL4A2	maxp1p2
AP	TNFRSF6	MMP	MMP11	maxp1p2
ECM	COL1A2	WNT	CDH13	maxp1p2
ECM	COL1A2	WNT	CTNND2	maxp1p2
CELL	MADH4	MMP	TIMP3	minp1p2
ECM	COL1A2	WNT	CDH5	minp1p2
ECM	COL1A2	WNT	CDH8	maxp1p2
ECM	COL1A2	MMP	TIMP3	Minp1p2
ECM	COL1A2	MMP	MMP17	Minp1p2
CELL	MADH3	AP	BCL2L2	maxp1p2
CELL	MADH3	ECM	COL15A1	maxp1p2
CELL	EP300	WNT	WNT5A	minp1p2
ECM	COL1A2	MMP	MMP19	maxp1p2

AP	TNFRSF6	ECM	COL6A3	maxp1p2
----	---------	-----	--------	---------

Table 3 lists the 15 most significant dyads of genes that predict improved hazard for survival, where both genes must be expressed to a high degree in order to promote survival. Table 4 presents the same criterion information for the 15 most predictive genes that impact the hazard in opposite directions, as shown in Figure 3 above.

**Table 4: 15 Most significant Dyads, opposite direction**

Class 1	Symbol 1	Class 2	Symbol 2	$\beta$	Model
CELL	MADH4	MMP	MMP2	+	minp1q2
CELL	MADH4	ECM	COL6A1	+	minp1q2
CELL	MADH4	ECM	COL5A1	+	minp1q2
ECM	COL1A2	WNT	CDH2	-	minp1q2
CELL	MADH4	ECM	FN1	+	minp1q2
ECM	COL2A1	WNT	CTNND1	-	maxp1q2
AP	TNFRSF6	ECM	TNC	+	minp1q2
CELL	MADH3	WNT	CDH11	+	minp1q2
AP	TNFRSF1B	WNT	CTNND2	+	minp1q2
ECM	COL2A1	WNT	WNT5A	-	maxp1q2
CELL	MADH4	AP	BCL2L2	+	minp1q2
AP	TNFRSF6	ECM	COL11A1	+	minp1q2
CELL	MADH4	WNT	CAV1	+	minp1q2
AP	TNFRSF6	ECM	COL18A1	+	minp1q2
CELL	MADH3	AP	BCL2	-	maxp1q2

## 4. Conclusion

In this paper we applied the weakest link family of statistical models to survival data in lung cancer. We focused on identifying dyads of genes that jointly impact patient survival, neither of which may individually predict survival. We showed proof of concept that the weakest link model family is easily interpretable in the context of gene expression analysis, that the weakest link concept is easily stated, both conceptually and mathematically.

We detected many more dyads than expected by chance. The dyads of genes we identified made biological sense; indeed, the tests performed were selected based on substantive considerations. Biological verification is required, however. In the very near future we plan to extend this analysis to incorporate triads of genes, requiring the addition of another function to specify the COU. We also intend to validate results statistically using cross-validation.