Int J Pharm Bio Sci 2015 July ; 6(3): (B) 164 - 176

Research Article **Bioinfomatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

MATHEMATICAL ANALYSIS OF RECEPTORS FOR SURVIVAL PROTEINS

SHRUTI JAIN*[1] AND DURG SINGH CHAUHAN[2]

[1]Department of Electronics and Communication Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh.
[2] GLA University, Mathura, Uttar Pradesh

ABSTRACT

In this paper, we have used the mathematical analysis to make a best linear model of the receptors of the survival proteins i.e. Epidermal growth factor (EGF) and Insulin using ten concentrations combinations. The model was made using different types of regression analysis in which regression coefficient (r^2), adjusted regression coefficient (r^2_{adj}),PRESS, regression coefficient, cross validation (q^2_{cv}), Durban Watson statistics and other parameters was analyzed. We have plotted the probability density function (pdf) for normal distribution functions for the receptors i.e. *epidermal growth factor receptor (EGFR) and insulin receptor (IRS)* which is coming OK. We have also plotted the survival function and hazard function for normal distribution function using different methods(kalpan meier and herd johnson) for maximum likelihood and least square methods. Non parametric functions were also plotted.

KEYWORDS: EGF, Insulin, Receptors, Regression Analysis, probability distribution function, Anderson darling adjustment values.

SHRUTI JAIN
Department of Electronics and Communication Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh.

***Corresponding author**

INTRODUCTION

Biological signaling networks process extracellular cues to control important cell divisions such as survival - apoptosis, growth-quiescence, and proliferation-differentiation[1]. Communication between the response of cells to extracellular signals such as cytokines, growth factors, and hormones is mediated by receptors that transduce cellular cues into changes in intracellular physiology. Downstream of receptors, signal communication networks [2, 3] are controlled by large sets of proteins acting in concert. In case of programmed cell death, TNF-α [4, 5] functions as apoptosis cues, whereas growth factors such as EGF [6-9] and insulin [10-13] exert survival effects. Increasingly, systematic methods are being applied to the interpretation and computational analysis of cell signaling[14-17]. These methods are useful for codifying existing prior knowledge in pursuit of in-silico predictions. The signal processing pertaining to cell survival/ apoptosis is ripe for applying various computational techniques to tease out the key biochemical changes associated with critical cell decisions because of the availability of experimental data. However, the experimental data that have been available are from the measurement of a wide range of parameters including protein abundance, localization, enzymatic activity, and posttranslational modification. In addition, protein measurements involve a variety of techniques, including western blots, kinase assays, protein microarrays, and imaging. The heterogeneous nature of the data for cell signaling studies present a challenging problem for data integration into a single coherent model. The decision between cell survival/ apoptosis is well regulated by three input signals : TNF, EGF and insulin[18]. These factors in single or in combination activate various key players in the network pertaining to cell survival/ apoptosis. Many proteins involved in this process that interact systematically regulating a specific pathway or cross talk with other proteins of different pathways. As a result, many pathways activated simultaneously leading to many biochemical and physiological changes inside the cell. The final outcome of whether a cell dies or survives depends on the concentrations of key players among the pathways. In this work our purpose is to determine the pdf, survival function of normal distribution functions using different methods (like Kalpan Meier and Herd Johnson) for maximum likelihood and least square methods. This paper focuses on the mathematical modeling using regression analysis of a signals and responses triggered by the receptors of EGF and insulin factors leading to cell survival/ apoptosis.

COMMUNICATION OF SIGNAL TRIGGERED BY EGF/ INSULIN LEADING TO CELL SURVIVAL/ APOPTOSIS

The epidermal growth factor (EGF) binds with EGF receptor (EGFR) at the outer side of cell membrane and phosphorylates the tyrosine residues of the receptor sub-units. These phosphorylated tyrosine sites allow other proteins to bind through their Src homology 2 (SH2) domains leading to the activation of downstream signaling cascades: the RAS/extracellular signal regulated kinase (ERK) pathway, the phosphatidylinositol 3 kinase(PI3K) pathway [19, 20] and the activator of transcription (JAK/ STAT) pathway [21]. The EGF signal is terminated primarily through endocytosis of the receptor-ligand complex. A number of signal transduction pathways branch out from the receptor signaling complex are shown in Figure 1.

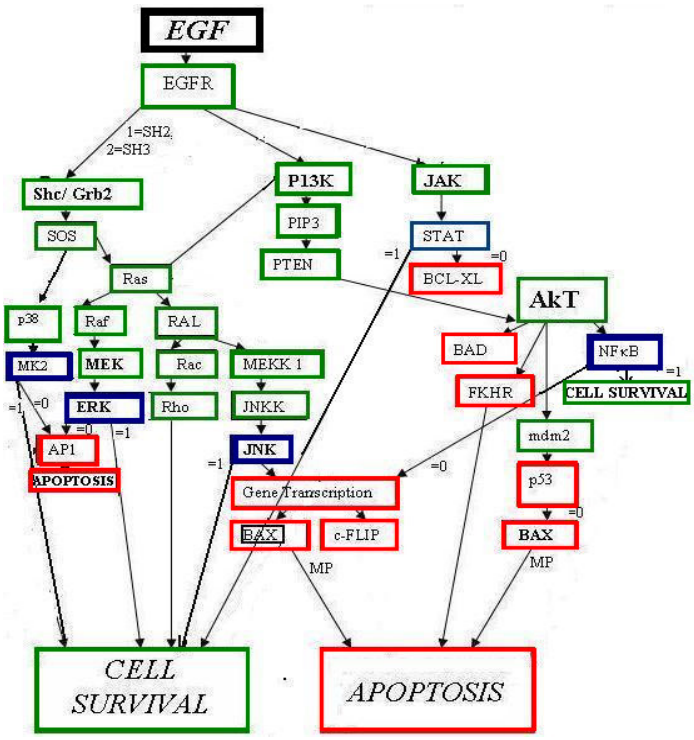

Figure 1

Illustration of signal communication network triggered by EGF. The box marked in red color are the proteins involved in cell apoptosis pathway, the box marked in green color are the proteins involved in cell survival pathway and the box marked in blue color are the proteins involved in both cell survival/ apoptosis.

Insulin is a hormone that binds to its receptor (the insulin receptor) on cell membranes and initiates signal transduction leading to cell survival/ apoptosis. Binding of insulin to its receptor induces phosphorylation of tyrosine residue on the inner part of the receptor. The phosphorylated tyrosine residues allow other intracellular proteins to bind to the intracellular domain of the receptor, and become phosphorylated. The signaling pathways activated by insulin are shown in Figure 2.

MATHEMATICAL MODELLING

For mathematical modeling we have used different types of regression analysis like linear, partial least square, k nearest neighbours, random forest, mean and SVM regression. We have designed a model using regression analysis for cell survival/ cell death. Different parameters are calculated with the different concentrations of the TNF, EGF and Insulin [1, 16] shown in table 1.

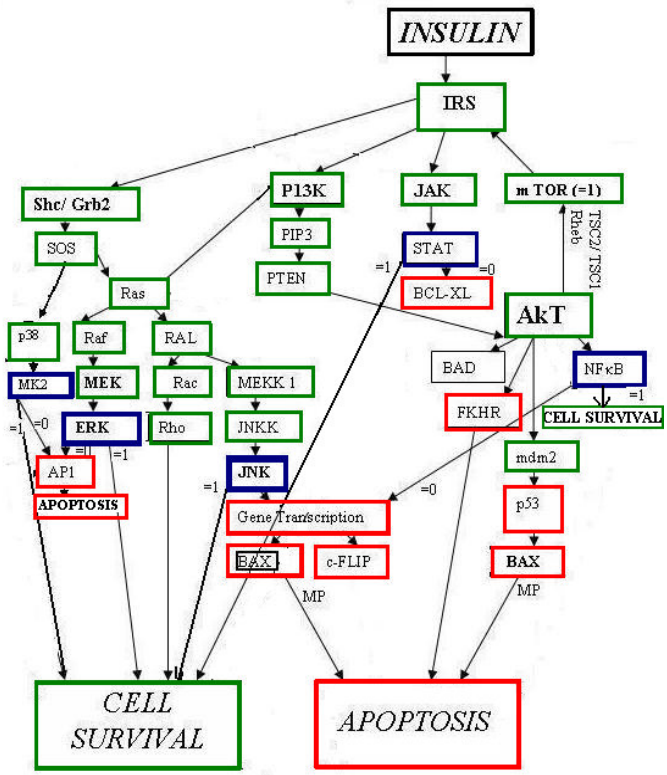

Figure 2

Illustration of signal communication network triggered by Insulin. The box marked in red color are the proteins involved in cell apoptosis pathway, the box marked in green color are the proteins involved in cell survival pathway and the box marked in blue color are the proteins involved in both cell survival/ apoptosis.

Table 1

Ten cytokine combinations of TNF, EGF and Insulin.

	a	b	c	d	e	f	g	h	i	j
TNF(ng/ml)	0	5	100	0	5	100	0	0.2	5	100
EGF(ng/ml)	-	-	-	100	1	100	-	-	-	-
Insulin(ng/ml)	0	0	0	-	-	-	500	1	5	500

a) For r^2 , r^2_{pred}, r^2_{adj}

The best equation was taken based on the statistical parameters such as regression coefficient (r^2), adjusted regression coefficient (r^2_{adj}) and predicted regression coefficient(r^2_{pred}). A data set has values y_i each of which has an associated modeled value f_i. The values y_i are called the observed values and the modeled values f_i are sometimes called the predicted values. $\bar{y}$ and $\bar{f}$ are the means of the observed data and modeled (predicted) values, respectively.

$$r^2 = 1 - \frac{\sum (y_i - f_i)^2}{\sum (y_i - \bar{y})^2} \quad or \quad r^2 = \frac{\sum (f_i - \bar{f})^2}{\sum (y_i - \bar{y})^2} \qquad ..1$$

We clubbed all the concentrations of TNF, EGF and Insulin and only normalized output (EGFR and IRS) was taken and we get the regression equation as

Final Output for EGFR = 0.531 -0.000396 *a* +0.000011 *b* +0.000029 *c* -0.000095 *d* +0.000046 *e* -0.000102 *f* +0.000206 *g* +0.000389 *h* -0.000275 *i* -0.000393 *j.* while

Final Output for IRS = 0.607 +0.000041 *a* -0.000010 *b* -0.000005 *c* -0.000156 *d* -0.000040 *e* -0.000027 *f* -0.000022 *g* -0.000024 *h* -0.000095 *i* -0.000202 *j.* where a, b, c ….j defines the combinations of TNF, EGF and Insulin shown in table 1. The parameters which we have calculated for *EGFR* are : S = 0.006142, r^2 = 91.7% , r^2 (adj) = 91.4% , = r^2 (pred) = 91.00% while the parameters which we have calculated for *IRS* are S = 0.006020, r^2 = 92.0% , r^2 (adj) = 91.7%, r^2 (pred) = 91.36%. We have also calculated : Mean square error (MSE) , Root mean square error (RMSE), Mean absolute error (MAE), Relative square error (RSE) , Root relative square error (RRSE) and Relative absolute error (RAE) for both the receptor proteins shown in table 2 and table 3 respectively.

Table 2
Various parameters using different regression methods for EGFR

	MSE	RMSE	MAE	RSE	RRSE	RAE
PLS Regression	0.0391	0.1977	0.0932	0.1671	0.4088	0.1992
Linear Regression	0.0391	0.1977	0.0932	0.1671	0.4088	0.1992
SVM Regression	0.0389	0.1972	0.0923	0.1663	0.4078	0.1973
K nearest neighbours regression	0.0611	0.2471	0.0774	0.2610	0.5109	0.1654
Mean	0.2376	0.4875	0.4711	1.0157	1.0078	1.0067
Random Forest regression	0.0481	0.2193	0.1498	0.2055	0.4533	0.3202

Table 3
Various parameters using different regression methods for IRS

	MSE	RMSE	MAE	RSE	RRSE	RAE
PLS Regression	0.0000	0.0062	0.0050	0.0896	0.2994	0.2638
Linear Regression	0.0000	0.0062	0.0050	0.0896	0.2994	0.2638
SVM Regression	0.0005	0.0216	0.0207	1.0755	1.0371	1.0969
K nearest neighbours regression	0.0001	0.0071	0.0056	0.1164	0.3412	0.2941
Mean	0.0004	0.0210	0.0190	1.0125	1.0062	1.0042
Random Forest regression	0.0000	0.0066	0.0053	0.0999	0.3161	0.2795

b) PRESS

The prediction sum of squares (PRESS), similar to the sum of squares of the residual error (SSE), is the sum of squares of the prediction error. Usually, the smaller the PRESS value, the better the model's predictive ability. In our case the PRESS value for *EGFR* is coming out to be 0.011756, while for *IRS* is 0.011288.

c) The regression coefficient cross validation (q^2_{cv}),

The predictive capability of the equation is determined using the leave-one-out cross validation method. The cross validation regression coefficient (q^2_{cv}) was calculated by the following equation.

$$q^2_{cv} = 1 - \frac{PRESS}{TOTAL} = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \quad ..2$$

For a perfect model value of q^2_{cv} should be close to one and its value is less than r^2. In our case the q^2_{cv} value for *EGFR* is coming out to be 0.91004 while for *IRS* is 0.91362.

d) Watson Statistics Durbin

If e_t is the residual associated with the observation at time *t*, then the test statistic is

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \quad ..3$$

where T is the number of observations. Note that if one has a lengthy sample, then Statistical Ideas site shows this can be linearly mapped to the Pearson correlation of the time-series data with its lags. Since *d* is approximately equal to 2(1 – *r*), where *r* is the sample autocorrelation of the residuals. If *d* > 2 indicated negatively correlation, *d* = 2 indicates no autocorrelation, *d* < 2 indicates positive serial correlation and if *d* < 1 causes alarm. The value of *d* always lies between 0 and 4. The Durbin-Watson statistic for EGFR is 2.00 while for IRS is 2.05 and the analysis of the variance was shown in Table 4 and Table 5, which shows the sum of squares and mean squares of the regression and residual error for EGFR and IRS respectively.

Table 4
Analysis of Variance for all combinations of EGFR

Source	dF	SS	MS	F
Regression	10	0.119780	0.011978	317.51
Residual Error	289	0.010903	0.000038	
Total	299	0.130683		

Table 5
Analysis of Variance for all combinations of IRS

Source	dF	SS	MS	F
Regression	10	0.120209	0.012021	331.70
Residual Error	289	0.010473	0.000036	
Total	299	0.130683		

e) Standard Error Coefficients

We use the standard error of the coefficient to measure the precision of the estimate of the coefficient. The smaller the standard error, the more precise the estimate. Dividing the coefficient by its standard error calculates a *t*-value. If the *p*-value associated with this *t*-statistic is less than alpha level, we conclude that the coefficient is significantly different from zero. Table 6 and Table 7, shows the regression analysis in terms of standard error coefficients, *t*-value, *p*-value and VIF for EGFR and IRS respectively.

Table 6
Regression analysis in terms of standard error coefficients, t-value, p value and VIF for EGFR

Predictor	Coefficient	Standard Error Coefficient	t-Value	p- Value	VIF
Constant	0.53102	0.03120	17.02	0.000	
0-0-0	-0.0003958	0.0004144	-0.96	0.340	2.6
5-0-0	0.0000112	0.0001812	0.06	0.951	23.5
100-0-0	0.0000292	0.0001428	0.20	0.838	57.0
0-100-0	-0.00009452	0.00005177	-1.83	0.069	216.9
5-1-0	0.00004638	0.00009832	0.47	0.637	10.3
100-100-0	-0.0001020	0.0001947	-0.52	0.601	135.2
0-0-500	0.0002055	0.0001215	1.69	0.092	54.8
0.2-0-1	0.0003894	0.0001104	3.53	0.000	147.6
5-0-5	-0.0002755	0.0001553	-1.77	0.077	36.3
100-0-500	-0.0003932	0.0001602	-2.45	0.015	168.3

Table 7
Regression analysis in terms of standard error coefficients, t-value, p value and VIF for IRS

Predictor	Coefficient	Standard Error Coefficient	t-Value	p- Value	VIF
Constant	0.60747	0.02273	26.73	0.000	
0-0-0	0.00004100	0.00005796	0.71	0.480	3.0
5-0-0	-0.00000956	0.00005695	-0.17	0.867	5.5
100-0-0	-0.00000530	0.00004477	-0.12	0.906	34.2
0-100-0	-0.00015624	0.00005956	-2.62	0.009	75.8
5-1-0	-0.00004002	0.00005312	-0.75	0.452	8.4
100-100-0	-0.00002721	0.00005600	-0.49	0.627	15.3
0-0-500	-0.00002179	0.00004705	-0.46	0.644	39.7
0.2-0-1	-0.00002416	0.00005232	-0.46	0.645	1.6
5-0-5	-0.00009460	0.00003498	-2.70	0.007	105.9
100-0-500	-0.00020225	0.00004208	-4.81	0.000	38.9

DISTRIBUTION FUNCTIONS

There are different distribution functions like Normal, Weibull, Lognormal base e, Lognormal base 10, Exponential, and Logistic. In this paper we have discussed only normal distribution function but with different methods like Kalpan Meier method, normal method and Herd Johnson method using Maximum likelihood and least square techniques. A normal distribution is expressed as

$$f(x,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \text{... (4)}$$

where μ defines the *mean* or *expectation* (median and mode) of the distance; σ is *standard deviation*. If μ = 0 and σ = 1 the distance is called standard/unit normal

distance. The probabilty density function (*pdf*), probability function, survival function and hazard function : normal distribution using *maximum likelihood for EGFR* is shown in fig 3, normal distribution using *least square for EGFR* is shown in fig 4, normal distribution using *maximum likelihood for IRS* is shown in fig 5 and normal distribution using *least square for IRS* is shown in fig 6 .

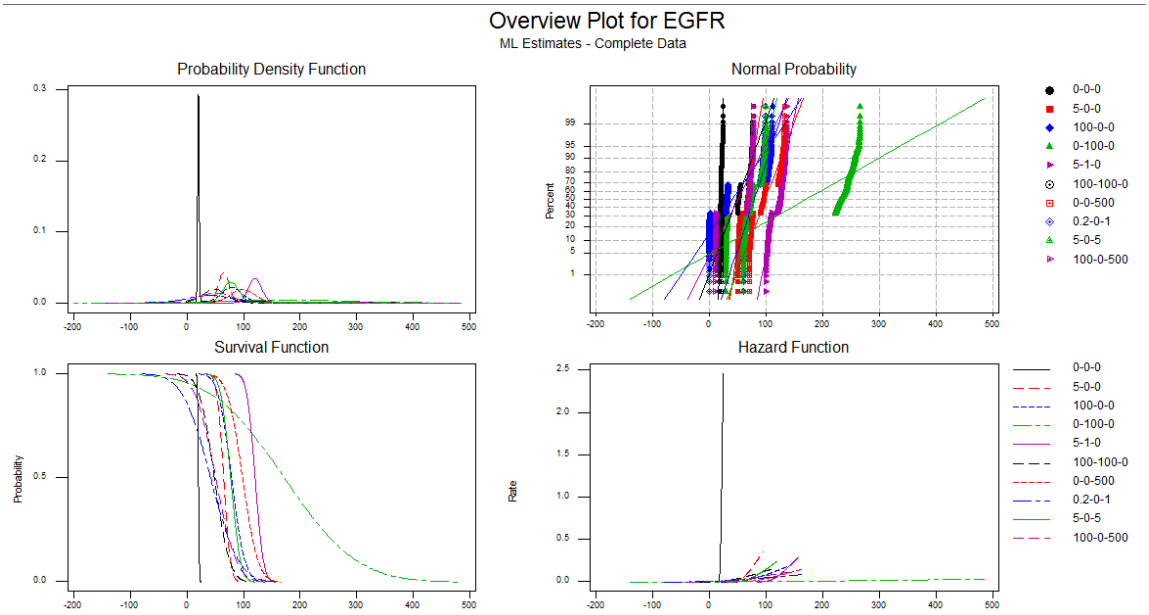

Figure 3
pdf, probability, survival function and hazard function for Normal distribution using Maximum likelihood for EGFR.

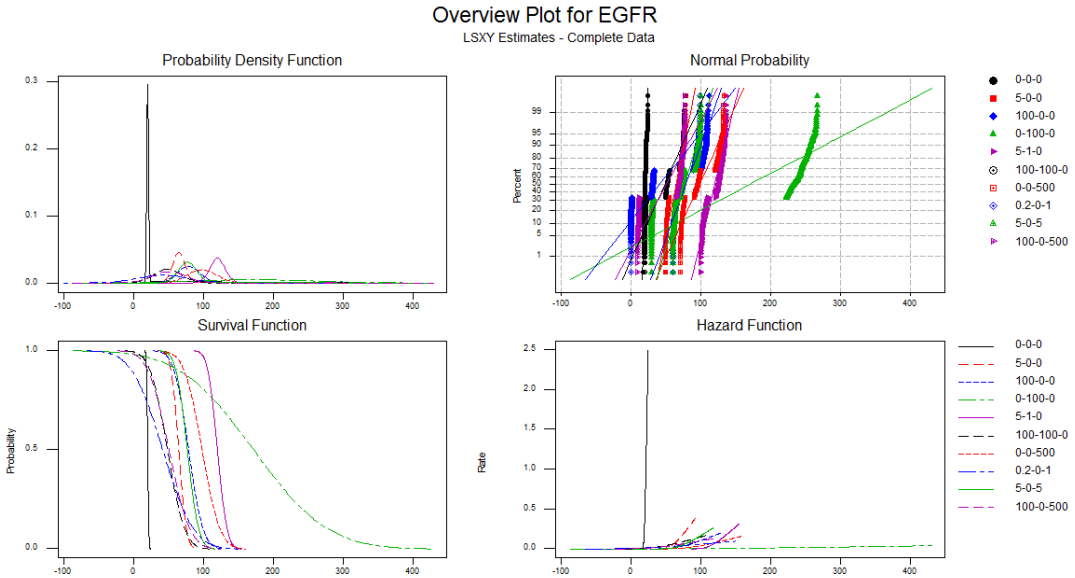

Figure 4
pdf, probability, survival function and hazard function for Normal distribution using Least Square for EGFR.

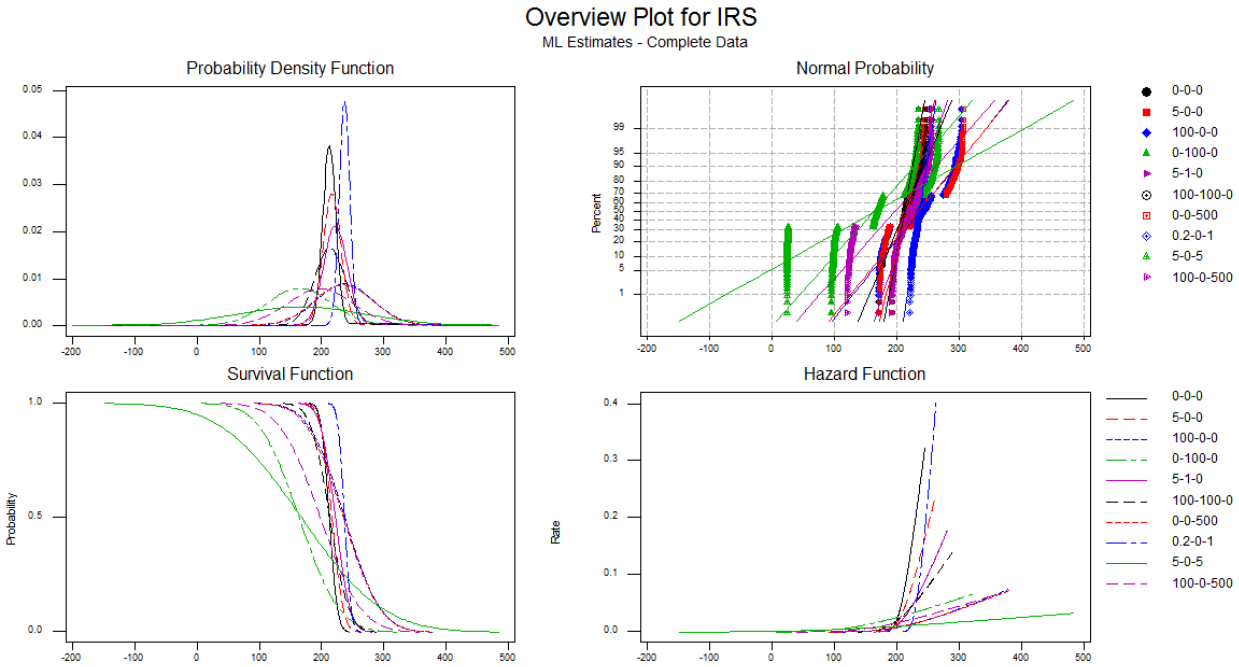

Figure 5
pdf, probability, survival function and hazard function for Normal distribution using Maximum likelihood for IRS.

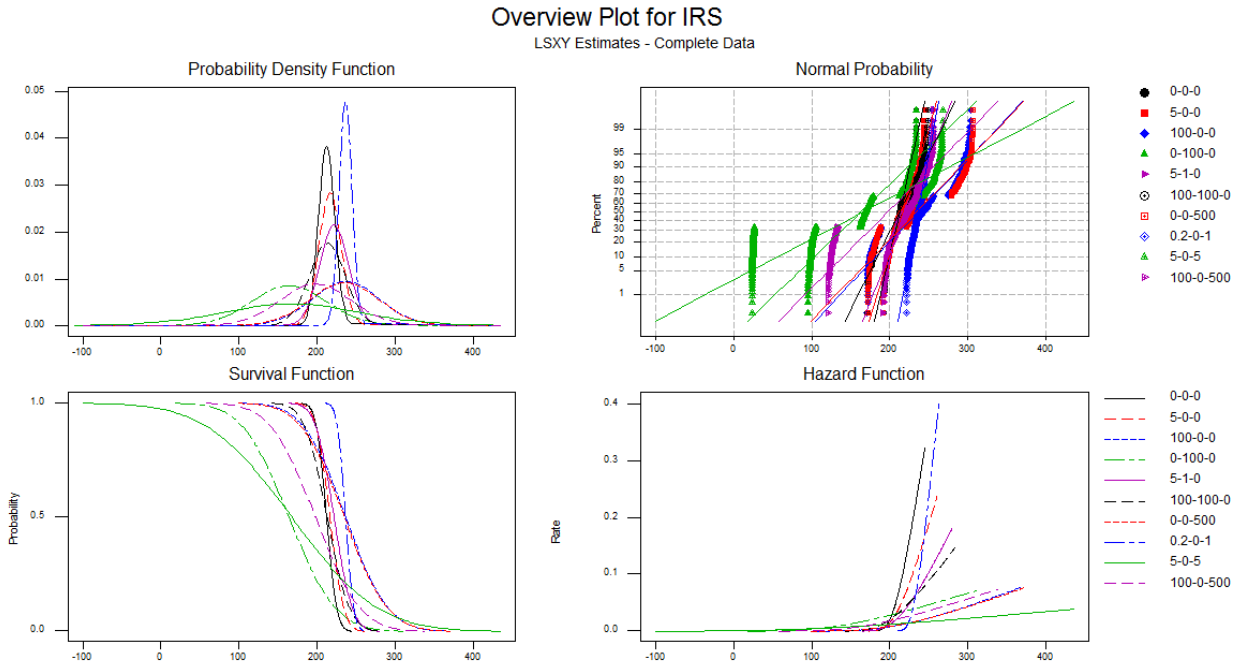

Figure 6
pdf, probability, survival function and hazard function for Normal distribution using Least Square for IRS.

We have calculated the Anderson darling adjustment values using two other methods known as Kalpan Meier and Herd Johnson method for normal distribution and results are shown in table 8 and table 9 for EGFR and IRS respectively.

Table 8

Anderson darling adjustment values for different methods using normal distribution function for maximum likelihood and least square techniques for EGFR.

	Maximum Likelihood			Least Square		
	Normal Method	Kalpan Meier Method	Herd Johnson Method	Normal Method	Kalpan Meier Method	Herd Johnson Method
0-0-0	2.13	2.10	2.21	2.18 0.982	2.04 0.984	2.21 0.983
5-0-0	20.90	20.96	20.89	25.49 0.918	26.58 0.914	24.67 0.920
100-0-0	33.83	33.92	33.85	44.70 0.871	45.20 0.871	43.45 0.874
0-100-0	46.77	46.88	46.66	67.84 0.825	70.36 0.820	65.84 0.827
5-1-0	17.00	17.03	17.00	20.10 0.932	21.06 0.928	19.45 0.933
100-100-0	19.04	19.10	19.08	23.82 0.916	24.74 0.913	22.99 0.918
0-0-500	14.23	14.26	14.29	17.17 0.935	17.63 0.933	16.56 0.937
0.2-0-1	26.29	26.36	26.33	34.27 0.890	34.96 0.888	33.20 0.892
5-0-5	15.06	15.09	15.12	18.10 0.933	18.47 0.932	17.48 0.936
100-0-500	46.80	46.91	46.69	67.88 0.825	70.41 0.820	65.88 0.827

Table 9

Anderson darling adjustment values for different methods using normal distribution function for maximum likelihood and least square techniques for IRS.

	Maximum Likelihood			Least Square		
	Normal Method	Kalpan Meier Method	Herd Johnson Method	Normal Method	Kalpan Meier Method	Herd Johnson Method
0-0-0	0.96	0.93	1.03	0.99 0.991	1.02 0.990	0.95 0.993
5-0-0	2.95	2.93	3.02	3.16 0.983	3.29 0.981	3.00 0.985
100-0-0	12.47	12.51	12.52	14.90 0.941	15.57 0.938	14.33 0.944
0-100-0	16.75	16.80	16.80	20.69 0.924	21.48 0.921	19.95 0.926
5-1-0	4.12	4.11	4.19	4.58 0.976	4.79 0.974	4.34 0.977
100-100-0	16.38	16.42	16.39	19.35 0.935	20.25 0.931	18.70 0.937
0-0-500	12.00	12.02	12.06	14.32 0.942	14.81 0.940	13.78 0.945
0.2-0-1	0.96	0.94	1.04	0.99 0.991	1.02 0.990	0.96 0.993
5-0-5	40.74	40.85	40.66	57.24 0.844	59.41 0.839	55.51 0.846
100-0-500	32.35	32.43	32.30	42.72 0.875	44.40 0.871	41.41 0.878

Above are parametric hazard functions. Fig 7 and Fig 8 shows the non-parametric hazard function and Kaplan Meier Survival function for EGFR and IRS respectively.

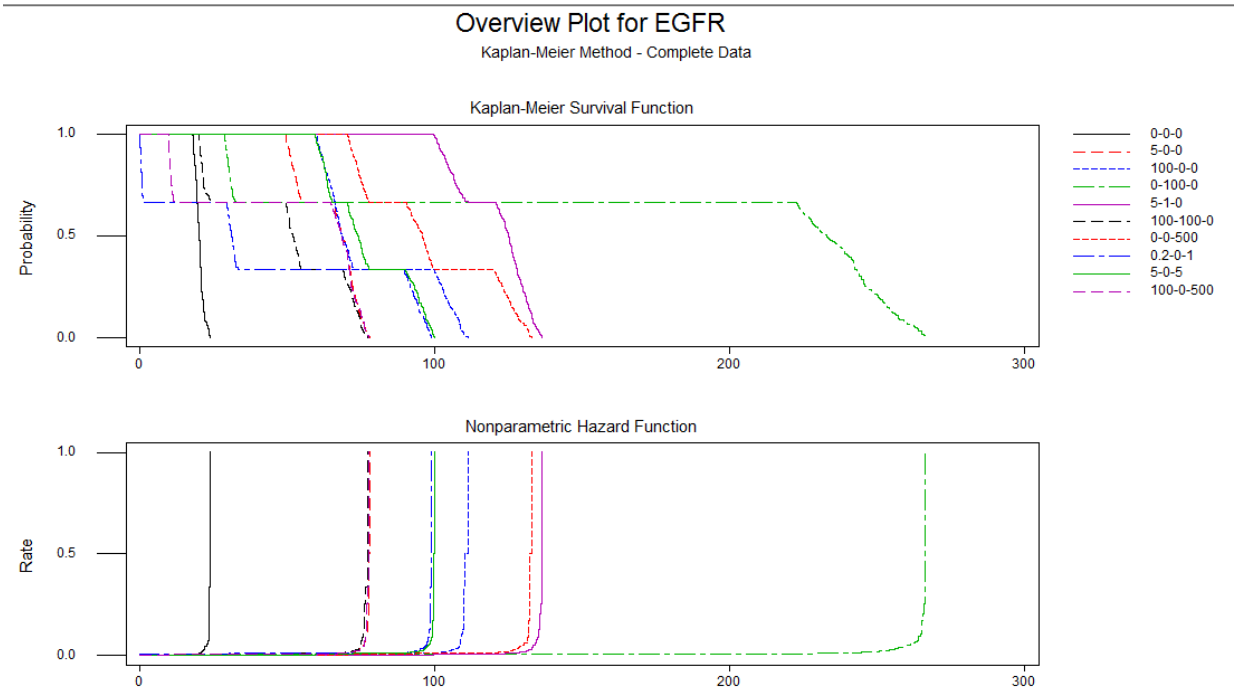

Figure 7
The non-parametric Hazard function and Kaplan Meier Survival function for EGFR.

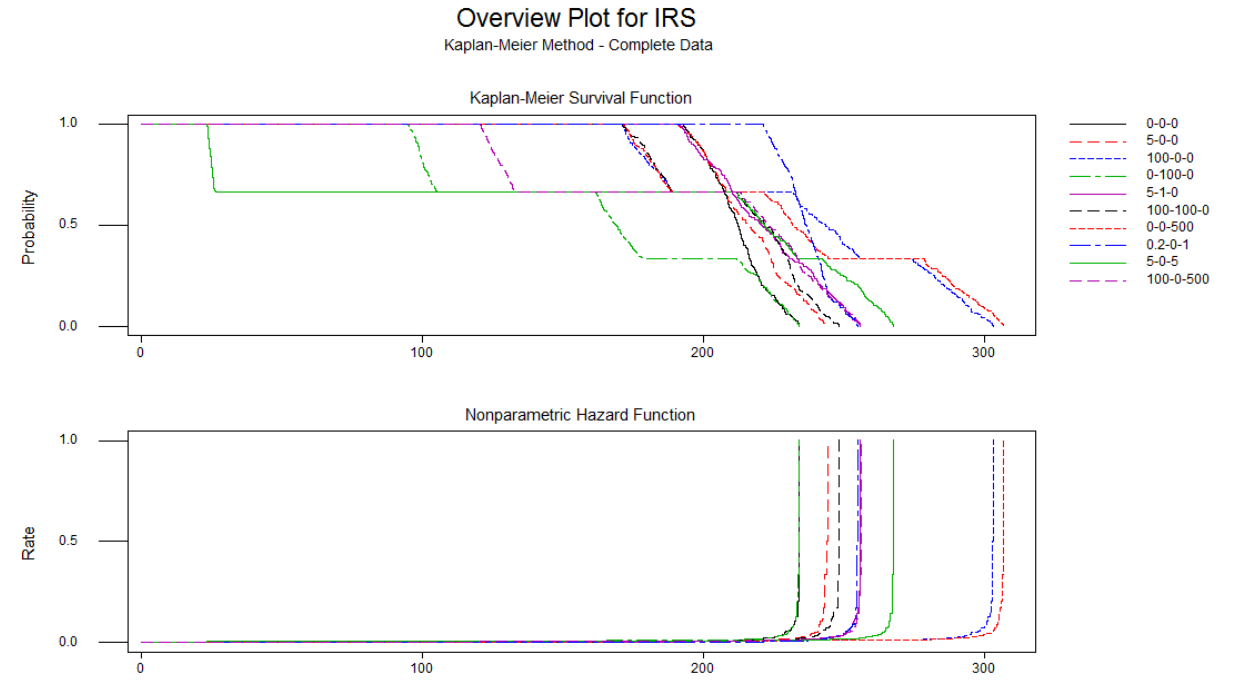

Figure 8
The non-parametric Hazard function and Kaplan Meier Survival function for IRS.

CONCLUSION

It has been revealed that survival and apoptosis signals induced by the receptors of EGF and insulin are temporarily separated and this is reflected in our model by the differences between the values of the parameters used. In this we have find all the results for regression coefficient (r^2), adjusted regression coefficient (r^2_{adj}), PRESS, regression coefficient cross validation (q^2_{cv}), Durbin Watson statistics, and *t*-values for our 10 data sets which comes out to be correct. Later we have used normal distribution functions to find the pdf. We have plotted the probability function, survival function and hazard function using different distributions for EGFR and IRS. More generally, these models are flexible, able to incorporate qualitative and noisy data, and powerful enough to produce quantitative predictions and new biological insights about the operation of signalling networks.

REFERENCES

1. Weiss, R.,Cellular computation and communications using engineered genetic regulatory networks . PhD Thesis, MIT, 2001.
2. Gaudet S, Janes K A., Albeck J G., Pace E A., Lauffenburger Do A, and Sorger P K., A compendium of signals and responses trigerred by prodeath and prosurvival cytokines Manuscript M500158-MCP200(2005).
3. Janes K A, Albeck J G, Gaudet S, Sorger P K, Lauffenburger D A, Yaffe B., A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis; Science 310**,** 1646-1653 (2005).
4. Brockhaus M, Schoenfeld HJ, Schlaeger EJ, Hunziker W, Lesslauer W, and Loetscher H Identification of two types of tumor necrosis factor receptors on human cell lines by monoclonal antibodies. Proc Natl Acad Sci USA 87, 3127-3131(1990).
5. Thoma B, Grell M, Pfizenmaier K, and Scheurich P, Identification of a 60-kD tumor necrosis factor (TNF) receptor as the major signal transducing component in TNF responses. J Exp Med 172, 1019-23(1990).
6. Libermann T A , Razon T A., Bartal A D, Yarden Y., Schlessinger J and Soreq H, Expression of epidermal growth factor receptors in human brain tumors Cancer Res. 44,753-760 (1984).
7. Normanno N, De Luca A, Bianco C, Strizzi L, Mancino M, Maiello MR, Carotenuto A, De Feo G, Caponiqro F, Salomon DS. , Epidermal growth factor receptor (EGFR) signaling in cancer Gene 366, 2–16 (2006).
8. Ullrich A., Schlessinger J., Signal transduction by receptors with tyrosine kinase activity : Cell, vol 61, 203-211, (1990).
9. Arteaga C., Targeting HER1/EGFR: a molecular approach to cancer therapy : Semin Oncol, vol 30, 314, (2003).
10. Lizcano J. M. , Alessi D. R. , The insulin signalling pathway. Curr Biol. 12, 236-238 (2002).
11. White M F., The insulin signaling system and the IRS proteins Diabetologia 40, S2–S17(1997)
12. White M F. , Insulin Signaling in Health and Disease Science 302, 1710–1711 (2003).
13. Jain S, Naik P.K., Bhooshan S.V., Mathematical modeling deciphering balance between cell survival and cell death using insulin, Network Biology, 1(1):46-58, (2011).
14. Jain S, Naik P.K., Bhooshan S.V., Petri net Implementation of Cell Signaling for Cell Death", International Journal of Pharma and Bio Sciences , 1-18, 1 (2), (2010).
15. Jain S, Naik P.K., System Modeling of cell survival and cell death : A deterministic model using Fuzzy System, International Journal of Pharma and Bio Sciences,358-373, 3(4): (Oct- Dec 2012)

16. Jain S, Naik P.K., Bhooshan S.V., A System Model for Cell Death/ Survival using SPICE and Ladder Logic, Digest Journal of Nanomaterials and Biostructures (DJNB), 5(1) : 57-66, (2010).
17. Jain S, Naik P.K. , Sharma R, A Computational Modeling of cell survival/ death using VHDL and MATLAB Simulator, Digest Journal of Nanomaterials and Biostructures (DJNB), 4 (4): 863- 879, (2009).
18. Jain S., Communication of signals and responses leading to cell survival / cell death using Engineered Regulatory Networks. PhD Thesis, Jaypee University of Information Technology, Solan, Himachal Pradesh, India, (2012).
19. Jorrissen R. N., Walker F., Pouliot N., Garrett T. J., Ward C. W., Burgess A.W., Epidermal growth factor receptor: mechanisms of activation and signaling, Exp. Cell Res., vol 284, pp 31-53, (2003).
20. Kim D., Chung J., Akt: Versatile Mediator of Cell Survival and Beyond, Journal of Biochemistry and Molecular Biology, vol 35(1), pp 106-115, (January 2002).
21. Kisseleva T., Bhattacharya S., Braunstein J., Schindler C. W., Signaling through the JAK/STAT pathway, recent advances and future challenges, Gene., vol 285, pp 1-24, (2002).