

International Journal of Scientific Research and Reviews

Computational Model Using Anderson Darling Statistics and Pearson Correlation Coefficients of Different Distribution Functions for Stress Activation Protein Kinase

Jain Shruti^{1*} and Chauhan D S²

Department of Electronics and Communication Engineering,
¹Jaypee University of Information Technology, Waknagaht, Solan, HP. 173234, India.
²GLA Mathura, Uttar Pradesh, 281406. India

ABSTRACT:

Multicellular organisms have three subfamilies of mitogen-activated protein kinases (MAPKs) that control a vast array of physiological processes. This paper stresses on stress activation protein kinase (SAPK/JNK) which are activated when cells are exposed to heat shock or UV radiation. SAPK protein is activated by different cycles of TNF, EGF and insulin proteins. The computational analysis using adjustment Anderson darling statistics (AD) and Pearson correlation coefficient (PCC) for different distribution functions. For good results AD value should be lower which was obtained for 100 ng/ml of TNF, 0 ng/ml of EGF and 0 ng/ml of insulin concentration for every distribution function. The PCC value should be higher which was obtained for 5 ng/ml of TNF, 1 ng/ml of EGF and 0 ng/ml of insulin concentration for herd Johnson method. The best value of PCC is 0.995.

KEYWORDS: SAPK, parametric, non-parametric, Anderson darling, Pearson correlation coefficient.

***Corresponding author**

Shruti Jain

Department of Electronics and Communication Engineering,
Jaypee University of Information Technology, Waknagaht, Solan, HP. 173234, India
E-mail: jain.shruti15@gmail.com

1. INTRODUCTION

The mitogen-activated protein kinases (MAPKs)^{1, 2, 3} regulate diverse cellular programs by relaying extracellular signals to intracellular responses. In mammals more than a dozen MAPK enzymes that regulate differentiation, cell proliferation, survival and motility^{4, 5}. These are classified into main three groups^{6,7}. (a) mediated by mitogenic and differentiation i.e. extracellular signal regulated kinase (ERK); (b) mediated to stress and inflammatory cytokines i.e. stress activation protein kinase (SAPK) / Jun N terminal kinase (JNK) pathway; (c) for high osmolarity glycerol i.e. p38/HOG. This paper stresses on SAPK/JNK which are activated when cells are exposed to heat shock or UV radiation. SAPK can be occur due to three main input proteins *tumor necrosis factor- α* (TNF)^{8, 9, 10}, *epidermal growth factor (EGF)*^{11, 12, 13, 14, 15, 16} and *insulin*^{17, 18, 19, 20}. The epidermal growth factor (EGF)/ insulin and their receptors are the first receptor pairs. EGFR is a member of receptor tyrosine kinase (RTK) family or known as human epidermal growth factor receptor (HER). Both receptors plays important role in cell survival^{21, 22, 23, 24, 25}.

This paper aims to do analysis by calculating adjusted Anderson Darling statics (AD) and Pearson Correlation Coefficients (PCC) using Normal, Kalpan Meier and Herd Johnson methods for Maximum Likelihood (ML) and Least Square (LS) approach for various distribution functions. For calculating AD^{26, 27} we can use both methods i.e. ML and LS while for calculating PCC, only LS approach can be used. Specifically, we have used ten different concentrations levels (ng/ml) of EGF, insulin and TNF- α for cell death/ cell survival for HT-29 human colon carcinoma cells. Later in this paper, section 2 explains the Adjusted Anderson Darling stat (AD) with its results; Section 3 explains Pearson Correlation Coefficient (PCC) and concluded in the end.

2. COMPUTATIONAL MODEL USING ANDERSON DARLING STATISTICS OF DIFFERENT DISTRIBUTION FUNCTIONS FOR SAPK

An adjusted Anderson darling statistics (AD) test is a method of determining whether, x_1, x_2, \dots, x_n can be assumed as a sample of n observations from any given distribution function^{26, 27}. In general, any distribution function is defined as

$$F(x) = \int_{-\infty}^{\infty} f(x) dx \quad -\infty < x < \infty \quad (1)$$

where $f(x)$ is a density of $F(x)$. When X is a random variable with distribution function $F(x) = \Pr \{ X \leq x \}$, then distribution function equation is represented as

$$\Pr \{X \leq x\} = \Pr \{F(X) \leq x\} = x, \quad 0 \leq x \leq 1. \tag{2}$$

A test of the hypothesis that x_1, x_2, \dots, x_n is a sample from a specified distribution $K(x)$, is equivalent to a test that $x_1 = K(x_1), x_2 = K(x_2), \dots, x_n = K(x_n)$ is a sample from $X(0, 1)$.

An AD test is a comparison of $K_n(x)$ with $K(x)$ where $K_n(x)$ is an empirical distribution function (edf) or cumulative distribution function (cdf). The hypothesis

$$H_0 : F(x) = K(x), \quad -\infty < x < \infty, \text{ is rejected}$$

The statistics is expressed as :

$$W_n^2 = n \int_{-\infty}^{\infty} [K_n(x) - K(x)]^2 \phi[K(x)] dK(x) \tag{3}$$

$$W_n^2 = n \int_{-\infty}^{\infty} [K_n(x) - K(x)]^2 \phi[K(x)] k(x) dx \tag{4}$$

where $k(x)$ is the density of $K(x)$. $\phi(z)$ is a weight function whose value is ≥ 0 , which can be represented as

$$\phi(z) = \frac{1}{z(1-z)} \tag{5}$$

Replacing the value of $\phi(z)$ from equation 5 to equation 3 we get equation 6 which is AD statistic equation :

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[K_n(x) - K(x)]^2}{K(x)[1 - K(x)]} dK(x) \tag{6}$$

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[K_n(x) - K(x)]^2}{K(x)} dK(x) + n \int_{-\infty}^{\infty} \frac{[K_n(x) - K(x)]^2}{[1 - K(x)]} dK(x) \tag{7}$$

From the above equation AD can be written as

$$A_n^2 = -n - \frac{1}{n} \sum_{j=1}^n (2j-1) \left[\log v_{(j)} - \log(1 - v_{(n-j+1)}) \right] \quad (8)$$

where n is the sample size and v is normal cdf. Value of v_j is different for different distribution functions.

This paper introduces the AD test for comparing distribution. AD test is more powerful when comparing two distributions which vary in scale or shift or symmetry or that which have the similar mean and standard deviation but differs on the tail ends only. In addition, the AD test has a type I error rate corresponding to alpha. It also requires less data to reach sufficient statistical power.

The most common distributions are Normal distribution (ND), Logistic distribution (LD), Exponential distribution (ED), Weibull distribution (WD), Lognormal distribution (LND e or LND 10). The AD statics is used to measure the area between the nonparametric step function and fitted curve for the distribution. We can also say that AD is a squared distance that is weighted more likely to occur in the peaks of the distribution. AD is used for the comparison of the fit for the particular distribution. Lower the value of AD better the fitting distribution. AD explains how far the plot points fall from the fitted line in a probability plot or how the data follow a particular distribution. In general, kalpan meier method is used to get AD values. p - value for this statistics is not calculated.

A normal distribution is expressed as

$$f(x, \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (9)$$

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

Fig 1 shows the AD for the normal, kalpan meier and herd johnson method values for the ML and LS approach for 09 different values of three input proteins.

A Weibull distribution function : The probability density function (pdf) of a WD is expressed as

$$f(x; \lambda, k) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (10)$$

where $k > 0$ is the *shape parameter* and $\lambda > 0$ is the *scale parameter* of the distribution. If $k = 2$ and $\lambda = \sqrt{2} \lambda$, than weibull function equals to Rayleigh distribution. Fig 2 shows the AD for the normal, kalpan meier and herd johnson method values for the ML and LS approach for 09 different values of three input proteins

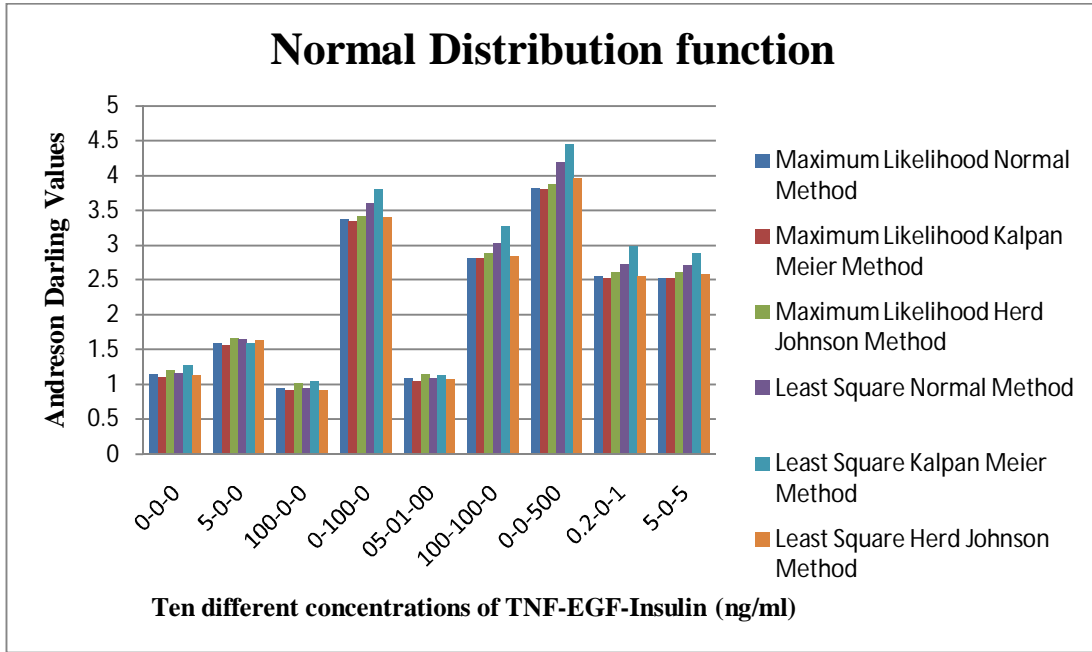


Fig 1 : The normal, kalpan meier and herd johnson method values for the ML and LS approach for ND

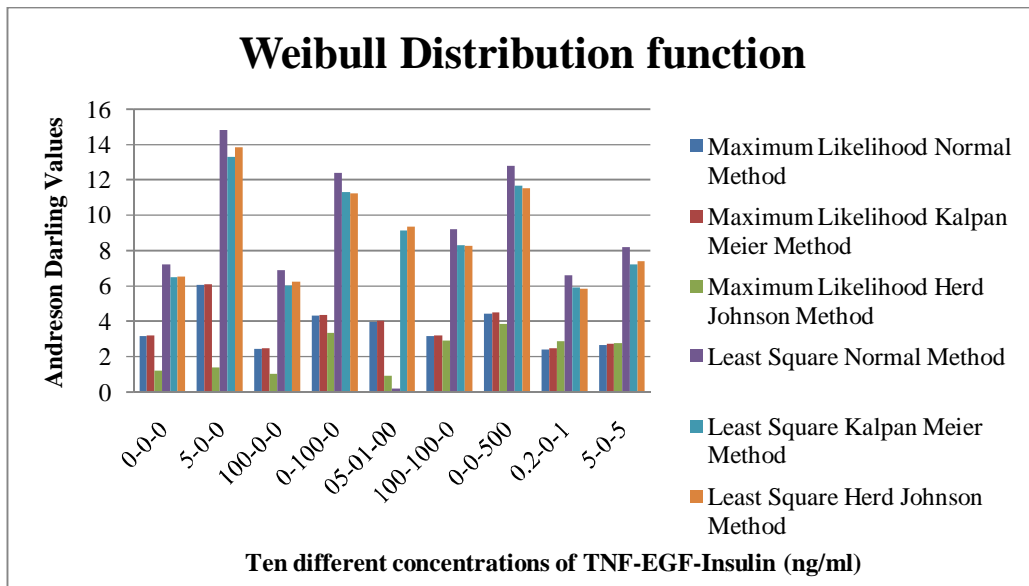


Fig 2 : Three different methods using the ML and LS approach for WD

An exponential distribution function : In Equation 11 if $k = 1$ than WD equals to ED. The pdf of an ED is expressed as

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

(11)

If $\lambda > 0$ is than the distribution, is called the *rate parameter*. Fig 3 shows the AD for the normal, kalpan meier and herd johnson method values for the ML and LS approach for 10 different values of three input proteins

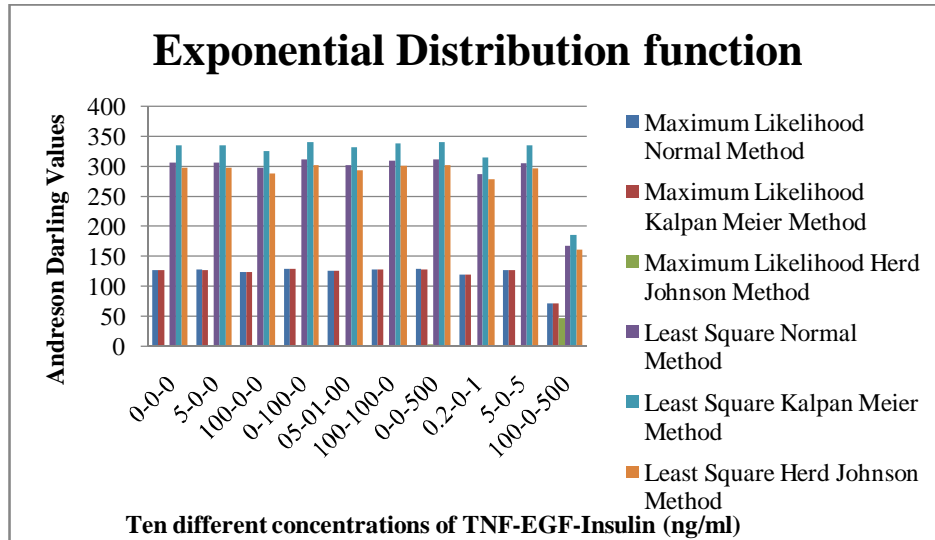


Fig 3: Three different methods using the ML and LS approach for ED

A lognormal distribution function: If the random variable X is log normally distributed then $Y = \log(X)$ is ND function. Similarly if Y is a normal distribution than $X = \exp(Y)$ has a log ND. The log normal function only takes real values. Fig 4 shows the AD for the normal, kalpan meier and herd johnson method values for the ML and LS approach for 09 different values of three input proteins.

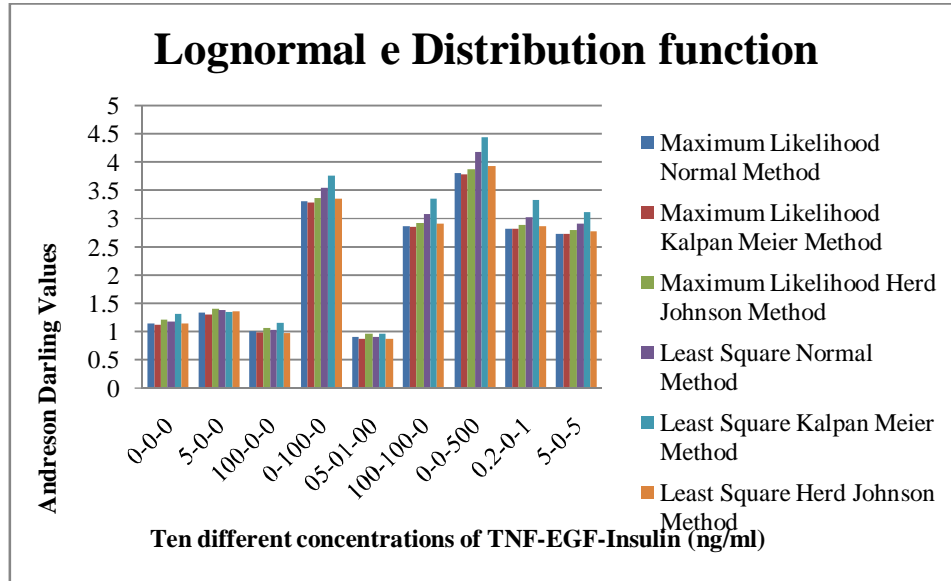


Fig 4: Three different methods using the ML and LS approach for LND e

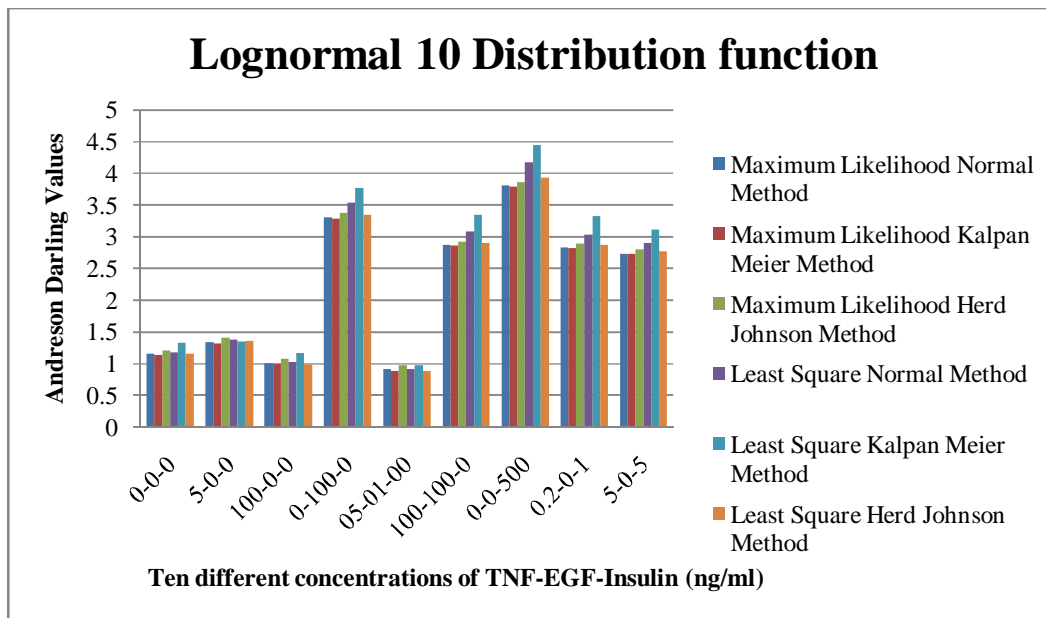


Fig 5: Three different methods using the ML and LS approach for LND

A logistic distribution function/ sech- square distribution: The pdf of the LD is given by:

$$f(x; \mu, s) = \frac{e^{-\frac{x-\mu}{s}}}{s \left(1 + e^{-\frac{x-\mu}{s}} \right)} = \frac{1}{4s} \operatorname{sech}^2 \left(\frac{x-\mu}{2s} \right) \tag{12}$$

Fig 6 shows the AD for the normal, kalpan meier and herd johnson method values for the ML and LS approach for 09 different values of three input proteins.

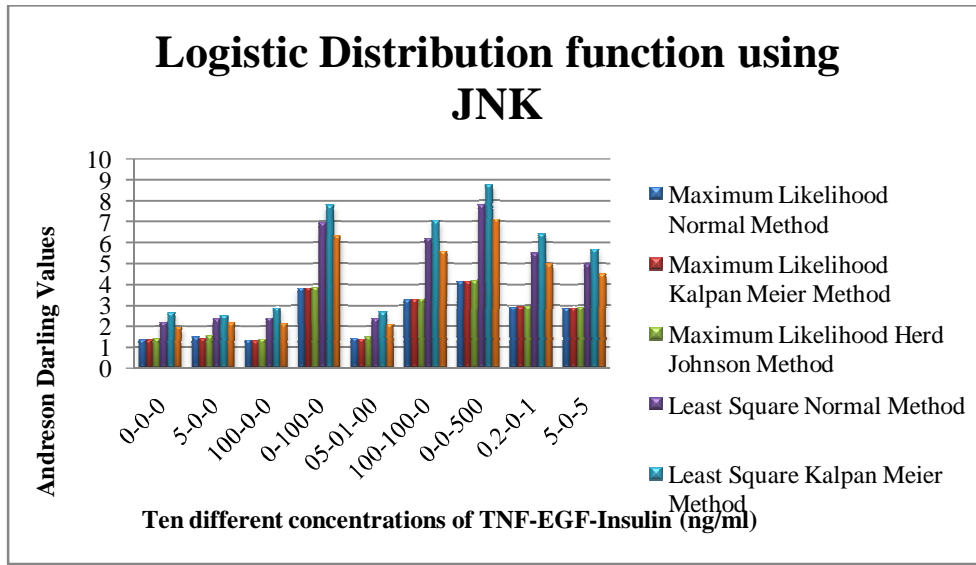


Fig 6: Three different methods using the ML and LS approach for LD

For ND, WD, LD and LND the best AD statistics is for 100 ng/ml of TNF, 0 ng/ml of EGF and 0 ng/ml of insulin concentration for ML and LS method. The tenth value i.e. 100 ng/ml of TNF, 0 ng/ml of EGF and 500 ng/ml of insulin concentration is too high that's why we have discarded that value.

3. COMPUTATIONAL MODEL USING PEARSON CORRELATION COEFFICIENT (PCC) OF DIFFERENT DISTRIBUTION FUNCTIONS FOR SAPK

There are four different types of Correlations : Pearson's r or Pearson Product-Moment Correlation: Spearman's r , Point-Biserial r and Phi (ϕ) Correlation.

Pearson's r is the calculation of the measuring the relation between two continuous variables. It is defined as the ratio of the covariance (CC) of two variables x and y representing a set of numerical data to the square root of the covariance of the multiplication of single variables x and y respectively. Pearson's r is defined by the equation 13.

$$r = \frac{CC_{xy}}{\sqrt{CC_{xx} C_{yy}}} = \frac{CC_{xy}}{\sigma_x \sigma_y} = \pm 1 \quad (13)$$

$r = \pm 1$ shows that the slope is negative or positive i.e. anti correlation or correlation.

$$\text{where } CC_{xy} = \frac{1}{N-1} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \quad (14)$$

likewise we can write the formula of CC_{xx} , CC_{yy} .

Table 1 to Table 5 shows the PCC for ND, WD, LND e , LND 10 and LD respectively. Higher the value of PCC better is the distribution curve. The PCC is only calculated for LSE approach. PCC values lies in the range of 0 and 1.

Table 1 : PCC for Normal Distribution using LSE method

	Normal Method	Kalpan Meier Method	Herd Johnson Method
0-0-0	0.991	0.988	0.992
5-0-0	0.988	0.987	0.989
100-0-0	0.992	0.991	0.994
0-100-0	0.983	0.981	0.984
5-1-0	0.993	0.991	0.994
100-100-0	0.984	0.981	0.985
0-0-500	0.979	0.977	0.981
0.2-0-1	0.984	0.982	0.986
5-0-5	0.984	0.983	0.985
100-0-500	0.839	0.834	0.841

Table 2 : PCC for Weibull Distribution using LSE method

	Normal Method	Kalpan Meier Method	Herd Johnson Method
0-0-0	0.974	0.978	0.977
5-0-0	0.939	0.945	0.943
100-0-0	0.964	0.968	0.968
0-100-0	0.948	0.952	0.952
5-1-0	0.963	0.967	0.966
100-100-0	0.954	0.959	0.958
0-0-500	0.943	0.949	0.948
0.2-0-1	0.962	0.967	0.966
5-0-5	0.946	0.951	0.950
100-0-500	0.840	0.843	0.844

Table 3 : PCC for Lognormal e Distribution using LSE method

	Normal Method	Kalpan Meier Method	Herd Johnson Method
0-0-0	0.990	0.987	0.991
5-0-0	0.989	0.989	0.990
100-0-0	0.992	0.990	0.993
0-100-0	0.983	0.981	0.984
5-1-0	0.994	0.992	0.995
100-100-0	0.983	0.981	0.985
0-0-500	0.979	0.976	0.981
0.2-0-1	0.983	0.980	0.985
5-0-5	0.983	0.982	0.985
100-0-500	0.826	0.821	0.828

Table 4 : PCC for Lognormal 10 Distribution using LSE method

	Normal Method	Kalpan Meier Method	Herd Johnson Method
0-0-0	0.990	0.987	0.991
5-0-0	0.989	0.989	0.990
100-0-0	0.992	0.990	0.993
0-100-0	0.983	0.981	0.984
5-1-0	0.994	0.992	0.995
100-100-0	0.983	0.981	0.985
0-0-500	0.979	0.976	0.981
0.2-0-1	0.983	0.980	0.985
5-0-5	0.983	0.982	0.985
100-0-500	0.826	0.821	0.828

Table 5 : PCC for Logistic Distribution using LSE method

	Normal Method	Kalpan Meier Method	Herd Johnson Method
0-0-0	0.980	0.973	0.983
5-0-0	0.977	0.974	0.980
100-0-0	0.980	0.975	0.984
0-100-0	0.966	0.960	0.970
5-1-0	0.983	0.977	0.986
100-100-0	0.967	0.960	0.971
0-0-500	0.960	0.954	0.964
0.2-0-1	0.968	0.961	0.972
5-0-5	0.969	0.964	0.972
100-0-500	0.815	0.806	0.819

For 5 ng/ml of TNF, 1 ng/ml of EGF and 0 ng/ml of insulin concentration give the best result of PCC i.e. 0.995 with Herd Johnson method.

CONCLUSION:

In this paper we have used the computational techniques to make a best linear model using ten concentrations combination of different pro-survival and pro- death protein for SAPK. We have calculated the Anderson darling statistics for normal, kalpan meier and herd Johnson method for least square and maximum likelihood approaches for different distribution functions. Pearson correlation coefficients were also calculated for normal, kalpan meier and herd Johnson method for least square approach. The value of AD should be lower while PCC value should be higher so as to make best fit model. For 100 ng/ml of TNF, 0 ng/ml of EGF and 0 ng/ml of insulin concentration gives the best result of AD and for 5 ng/ml of TNF, 1 ng/ml of EGF and 0 ng/ml of insulin concentration give the best result of PCC i.e. 0.995 with Herd Johnson method.

REFERENCES

1. Jain S., Communication of signals and responses leading to cell survival / cell death using Engineered Regulatory Networks. PhD Dissertation, Jaypee University of Information Technology, Solan, Himachal Pradesh, India. 2012.
2. Jain S, Bhooshan SV, Naik PK, Model of Mitogen Activated Protein Kinases for Cell Survival/Death and its Equivalent Bio-Circuit, Current Research J of Biological Sciences. 2010; 2(1) : 59-71.
3. Weiss R. Cellular computation and communications using engineered genetic regulatory networks. PhD Dissertation, MIT. 2001.
4. Gaudet S, Kevin JA, John AG, Emily PA, Douglas LA, Peter SK. A compendium of signals and responses triggered by prodeath and prosurvival cytokines. Manuscript M500158-MCP200, 2005.
5. Jain S, Regression analysis on different mitogenic pathways, Network Biology. 2016; 6(2): 40-46 .
6. Jain S, Regression modeling of different proteins using linear and multiple analysis, Network Biology. 2017;7(4): 80-93.
7. Jain S, Mathematical Analysis using Frequency and Cumulative Distribution functions for Mitogenic Pathway, Research Journal of Pharmaceutical, Biological and Chemical Sciences. May - Jun 2016; 7(3): 262-72.

8. Kevin JA, John AG, Suzanne G, Peter SK, Douglas LA, Michael YB. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science*. 2005; 310: 1646-53.
9. Jain S, Implementation of Marker Proteins Using Standardised Effect, *J of Global Pharma Technology*. 2017; 9(5): 22-27.
10. Thoma B, Grell M, Pfizenmaier K, Scheurich P. Identification of a 60-kD tumor necrosis factor (TNF) receptor as the major signal transducing component in TNF responses. *J Exp Med*. 1990; 172: 1019-23.
11. Jain S, Naik PK, Bhooshan SV. Mathematical modeling deciphering balance between cell survival and cell death using Tumor Necrosis Factor α . *Research J of Pharmaceutical, Biological and Chemical Sciences*. 2011; 2(3): 574-83.
12. Normanno N, Luca AD, Bianco C, Strizzi L, Mancino M, Maiello MR *et al* ,Epidermal growth factor receptor (EGFR) signaling. *Cancer Gene*. 2006; 366: 2–16.
13. Jain S, Compedium model using frequency / cumulative distribution function for receptors of survival proteins: Epidermal growth factor and insulin, *Network Biology*. 2016; 6(4): 101-110.
14. Jain S , Chauhan DS. Mathematical Analysis of Receptors For Survival Proteins. *International J of Pharma and Bio Sciences*. 2015; 6(3): 164-176.
15. Jain S, Naik PK, Bhooshan SV. Mathematical modeling deciphering balance between cell survival and cell death using insulin. *Network Biology*. 2011; 1(1):46-58.
16. Jain S, Naik PK, Bhooshan SV. A System Model for Cell Death/ Survival using SPICE and Ladder Logic. *Digest Journal of Nanomaterials and Biostructures*. 2010; 5(1): 57-66.
17. Jain S, Naik PK. System Modeling of cell survival and cell death : A deterministic model using Fuzzy System, *International Journal of Pharma and BioSciences*. 2012; 3(4): 358-73.
18. Jain S, Naik PK, Communication of signals and responses leading to cell death using Engineered Regulatory Networks, *Research Journal of Pharmaceutical, Biological and Chemical Sciences* . July – Sep 2012; 3(3): 492-508.
19. JM Lizcano and DR Alessi. The insulin signalling pathway. *Curr Biol*.2002; 12: 236-38.
20. White MF. Insulin Signaling in Health and Disease. *Science*. 2003; 302: 1710–11.
21. Jain S, Naik PK. System Modeling of cell survival and cell death: A deterministic model using Fuzzy System. *International Journal of Pharma and BioSciences*. 2012; 3(4): 358-73.
22. Jain S, Naik PK, Bhooshan SV. Nonlinear Modeling of cell survival/ death using artificial neural network. In *The Proceedings of International Conference on Computational Intelligence and Communication Networks*, Gwalior, India. 2011; 565-68.

23. Jain S, Naik PK, Sharma R, A Computational Modeling of cell survival/ death using VHDL and MATLAB Simulator, Digest Journal of Nanomaterials and Biostructures. 2009; 4(4): 863- 79.
 24. Jain S, Naik PK, Bhooshan SV. Petri net Implementation of Cell Signaling for Cell Death. International Journal of Pharma and Bio Sciences. 2010; 1(2):1-18.
 25. Jain S, Chauhan DS. Linear and Non Linear Modeling of Protein Kinase B/ Akt. In: Proceeding of the International Conference on Information and Communication Technology for Sustainable Development, Ahmedabad, India.2015; 81-88.
 26. Jain S, Parametric and Non Parametric Distribution Analysis of Akt for Cell Survival/Death, International Journal of Artificial Intelligence and Soft Computing. 2017; 6(1): 43- 55.
 27. Anderson TW, Darling DA. Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes, Ann. Math. Stat., 1952; 23: 193-212,
 28. Marsaglia G, Tsang WW, Wang J. Evaluating Kolmogorov's distribution, Journal Statistical Software, 2003; 8(18):45-52.
-