STATISTICAL MODELLING OF GENE REGULATORY NETWORKS

[1]Omolola Dorcas Atanda, [2]Angela U Chukwu, Soyinka Ajibola Taiwo[3], [4]Wale-Orojo Oluwaseun Ayobami.

[1].Nile University of Nigeria, Abuja, Nigeria. [2].University of Ibadan, Ibadan, Nigeria.
[3&4]Federal University of Agriculture Abeokuta, Nigeria

Corresponding e-mail: omolola.atanda@nileuniversity.edu.ng, +2348033398222

Biological network analysis is a rapidly growing field which is increasing our understanding of biological process. The study and modeling of biological networks are important in life science today. A wide range of scientists are interested in quantifying the link between nodes in a system, however the linkage is not as straight forward as it might seem. The challenge is, how to extract relevant information and translate this information to knowledge that can yield clinically actionable results. Insights gained from successful computational statistics of networks topology can, in principle, be used to design new experiments that test these insights in a broad context. In this work, a newly derived discretised Power probability density function is proposed for in-degree distribution of gene regulatory networks. A statistical comparison of the newly proposed degree distribution was made with alternative degree distribution in literature.

Keywords: Biological network, network topology, degree distribution, discrete power function

1.0. INTRODUCTION

Gene Regulatory Networks

GRN's is the mathematical and computational representations describing the logic fundamental regulatory occurrences among genes when a specific cell program is operating.
Biological system can be represented by network which are compound sets of binary interactions between different entities. Basically, every single biological unit has interactions with other biological units, from the molecular to the ecosystem level, affording us with the opportunity to model biology using several different types of networks such as molecular interaction, metabolic, neurological and ecological networks.

An understanding of biological networks is crucial to make biological sense of much of the complex data that is now being generated. This growing importance of biological networks is also shown by the increase in publications about network associated topics and the increasing number of research groups dealing with big data like that of biological networks.

One significant property of biological networks that has raised much interest is their heterogeneous topology which are mostly more difficult to analyze; inferring their detailed topology requires wide-ranging statistics. On assumption Power-law distributions habitually provide a good estimate to such network's degree distribution, nonetheless empirical studies have led to some debate concerning their adequacy. Hence, a detailed analyses of network topologies and modeling the **degree distribution** is still one bottleneck in biological networks connectivity.

Topological Measures

To distinguish biological networks, it is important we identify some features that are numerical measures unfolding the pattern of connectivity in the networks. The criteria are referred topological features or measures

- Degree distribution

- The diameter and characteristic path length
- Clustering coefficient
- the network robustness and presence of hubs

1.1 Aim and Objectives of the study

To proposed a Discretized Power Law for modeling the Degree Distribution of Gene Regulatory Networks

i. To derive some of the statistical properties of the newly proposed distribution such as its cumulative distribution, reliability function among others
ii. To make a statistical comparison of the proposed distribution and some other existing assumed degree distribution of biological network using the Akaike Information Criteria (AIC) and other information criterion.

2.0 REVIEW OF LITERATURE

The studied literature such works of Jing-Dong et al. (2005) on ‘, Guelzim et al (2002) , Ravasz et al (2002) and Barabasi and Oltvai (2004) concluded that the degree distribution of biological networks follows a power-law distribution and also that the hubs are at the tail of the distribution.

However,

In the study of Random networks The work of Erdos P and Renyi (1960), the result of their empirical study shows that the distribution of nearest neighbor follows a Poisson distribution. The work on Lethality and Centrality in Protein Networks by Jeong et al (2001) reported that connectivity distribution in some biological networks might be better described by a truncated power law. Moreso, Przytycka and Yu (2004) studied ‘Scale free networks versus evolutionary drift’ their conclusion is that scale free networks contradict Power law distribution. Eric. E et al (2017) also in their study titled ‘Network Enabled Wisdom in Biology, Medicine and Healthcare’ discussed how molecular networks are central into wisdom that can yield clinically actionable results. Eric. E et al, however assumed biological networks are sparse and therefore followed a power law distribution on assumption.

The work of Khanin and Wit (2006) on ‘How Scale Free are Biological Networks’ gave much in -dept on the degree distribution of biological networks in general. The result of their comprehensive study of degree distribution of different 10 published biological networks found out that degree distribution pointedly deviates the assumed power-law distribution also not in the form of scale-free . In the study, Khanin and Wit suggested four other distributions; ‘generalised pareto-law’, stretched exponential’ ‘geometric’ and ‘truncated power-law’ distribution as alternative distributions that may best describe the indegree distribution of biological networks . These suggested alternative distributions was further studied by Vilda and Omolola (2016).

Discrete Analogues of Continuous Probability Distribution

In recent times, many research papers reviewing discrete distributions obtained by discretising a continuous distribution have been seen in many statistical studies. In published literature we found two articles that studied discrete analogues of continuous distributions, that is, the work of Bracquemond and Gaudoin (2003) who extensively studied a discrete life-time distributions derived from continuous and also Lai (2013) presented construction of discrete life-time distributions from continuous one in his paper concerning ‘Issues of construction of discrete life-time distribution’. Our studies also reviewed the work of Subrata Chakraborty, on survey of different ways of obtaining a probability mass functions as analogues of continuous probability distribution.

Discrete Concentration Approach, a method proposed by Roy (2003) and Kemps (2004) is applied in this work. If we denote the continuous random variable to be discretized as x while the discrete analogue by y and the resulting continuous random variable $x,$ the survival function $S_X(x)$, thus the random variable y

$$P(y = k) =$$

$$P(k \leq x < k + 1)$$

$=F_x(k + 1) - F_x(\mathrm{k})$

$= S_x(k) - S_x(k + 1)\ \ where\ \ k = 0,\ 1,\ 2 \dots\dots\dots$ (1)

$$P(X = x) = 0\ and\ F_x(k) = 1 - S_x(k)$$

This method will preserves the survival function such that $S_X(k) = S_Y(k)$

3.0 METHODOLOGY

Discretized Generalised Pareto Distribution (DGPD)

A random variable Y is distributed as Discrete Generalised Pareto (DGP) Distribution with parameters $\mu,\ \sigma$ and ϵ , denoted by dDGP ($\mu,\ \sigma,\ \epsilon$), then the cumulative distribution function, survival function and probability mass function is defined by eq (1) as defined:

$$F(x) = 1 - \left[1{+}\in\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\in}}$$

Let $\theta = \mathrm{e}^{-\frac{\epsilon}{\sigma}} \quad \Rightarrow \theta^{-1} = \mathrm{e}^{\frac{\epsilon}{\sigma}} \Rightarrow ln\,\theta^{-1} = \frac{\epsilon}{\sigma}$

$$\theta^{-1} = \mathrm{e}^{\frac{\epsilon}{\sigma}} \quad \Rightarrow \mathrm{e}^{\frac{\epsilon}{\sigma}(x-\mu)}$$

$$\frac{\epsilon}{\sigma}(x-\mu) = ln\theta^{-(x-\mu)}$$

Let $(x-\mu) = y \qquad 0,\ 1,\ 2,\ \dots\dots\dots\dots\dots$

Hence $F(x) = 1 - \left[1{+}\in\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\in}}$ in discrete form becomes

$F(y) = 1 - [1 + ln\theta^{-y}]^{-\frac{1}{\in}} \qquad ;\ 0 < \theta < 1$ (2)

Also, the survival function in the discrete analogues is expressed as

$$S(y) = 1 - F(y)$$

$$S(y) = \left[1 - \left(1 - [1 + ln\theta^{-y}]^{-\frac{1}{\in}}\right)\right]^{-\frac{1}{\in}}$$

$S(y) = [1 + ln\theta^{-y}]^{-\frac{1}{\in}}\ ;\ y \geq 0$ (3)

Defining the survival at point $y+1$ we have

$S(y+1) = [1 + ln\theta^{-y+1}]^{-\frac{1}{\in}} \quad ;\ \ y > 0,$ (4)

Now we can define the pmf of a Discrete Pareto distribution by the approach in eq (1), Y$\sim DPGD\ (\mu, \sigma, \in)$

$P(Y = y) = S(y) - S(y+1)$; y=0, 1, 2,…………

$f(y; \in, \mu, \sigma) = [1 + ln\theta^{-y}]^{-\frac{1}{\in}} - [1 + ln\theta^{-y+1}]^{-\frac{1}{\in}};$ y = 0, 1,2……. (5)*

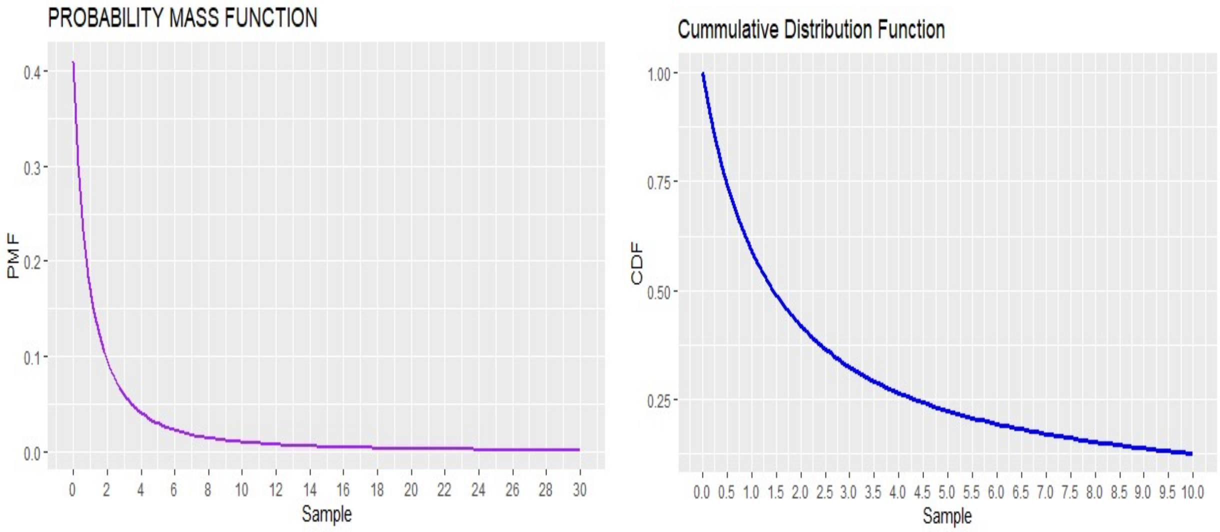

Fig1: The PMF and CDF of Discretized Generalised Pareto Distribution (DGPD)

Statistical Properties of DGPD)
Residual Reliability Function

$$R(i|x) = \frac{\left[1 + ln\theta^{-x+i}\right]^{-\frac{1}{\epsilon}}}{\left[1 + ln\theta^{-x}\right]^{-\frac{1}{\epsilon}}} \qquad (6)$$

Hazard Rate

$$h(x) = 1 - \frac{\left[1 + ln\theta^{-x+1}\right]^{-\frac{1}{\epsilon}}}{\left[1 + ln\theta^{-x}\right]^{-\frac{1}{\epsilon}}} \qquad (7)$$

The Quantile function

$$x = \frac{1}{-ln\theta}\left(\left[1 - F(x)\right]^{-\epsilon} - 1\right) \qquad (8)$$

The likelihood function is given as

$$LogL(x_i) = \sum_{i=1}^{n} log\left[1 - (x_i + 1)ln\theta\right]^{\frac{1}{\epsilon}} - \left[1 - x_i ln\theta\right]^{\frac{1}{\epsilon}} - \frac{1}{\epsilon}\sum_{i=1}^{n} log\left[1 - (x_i + 1)ln\theta\right] - \frac{1}{\epsilon}\sum_{i=1}^{n} log\left[1 - x_i ln\theta\right] \qquad (9)$$

r^th moments about the origin

Using the expression given by ………….in Eq (2.)

$$E(x^r) = \sum_{x=1}^{k-1} [(x+1)^r - x^r] S(x) + \ [S(x=0)]^r \qquad (10)$$

Where $S(x)$ is the survival function

$when\ r = 1$, equation (10) will be

$$E(x^1) = \sum_{x=1}^{k-1} [(x+1)^1 - x^1] S(x) + \ [S(x=0)]^r \qquad (11)$$

Note: $S(x=0) = 0$

Therefore equation (11) becomes

$$E(x) = \sum_{x=1}^{n} S(x)$$

$$E(x) = \frac{1}{(ln\theta^{-1})^{\frac{1}{\in}}} Zeta\left([log\theta^{-1}]^{-1}, \frac{1}{\in}\right) \qquad (12)$$

Corollary

$$\sum_{x=1}^{n} [1 + ln\theta^{-x}]^{-\frac{1}{\in}}$$

$$= \sum_{x=1}^{n} \frac{1}{[1 + ln\theta^{-x}]^{\frac{1}{\in}}}$$

$$= \frac{Zeta\left(\frac{1}{\in}, [ln\theta^{-1}]^{-1}\right)}{(ln\theta^{-1})^{\frac{1}{\in}}};\ 0 < \ \theta < 1, \frac{1}{\in} > 1$$

$$E(x^2) = \sum_{x=1}^{k-1} [(x+1)^2 - x^2] S(x) + \ [S(x=0)]^2$$

$$E(x^2) = E(x) + 2\sum_{x=1}^{k-1} x[1 + ln\theta^{-x}]^{-\frac{1}{\in}}$$

$$E(x^2) = k - (ln\theta^{-1})^{-\frac{1}{\in}} \frac{Zeta\left(\frac{1}{\in}, [ln\theta^{-1}]^{-1}\right)}{(ln\theta^{-1})^{\frac{1}{\in}}})$$
$$- \frac{2}{(ln\theta^{-1})^{1+\frac{1}{\in}}\Gamma_{\frac{1}{\in}}} \left[ln\theta^{-1} Zeta\ \left(\frac{1}{\in} - 2, (ln\theta^{-1})^{-1}\right)\right]\left[ln\theta^{-1} + \frac{1}{\in} - 1\right]$$
$$- \left[Zeta\ \left(\frac{1}{\in} - 1, (ln\theta^{-1})^{-1}\right)\right] \qquad (13)$$

3.1 Data and Results

In the study, six real datasets of human diseases interactions are studied. The data are extracted from OMIM "Online Mendelian Inheritance in Man database" . Cytoscape was used to visualize and generate the degree and other quantifiable measures.

4.0 Summary and Findings

Data 1				
Probability Distribution	AIC	BIC	CAIC	HQIC
GEO	1009.6	1043.21	1277.88	1933.25
SED	994.30	1118.0	2178.80	2051.16
GPD	1113.45	1271.3	1929.92	1925.30
DGPD	16.42	44.57	277.71	334.14
	Data 2			
Probability Distribution	AIC	BIC	CAIC	HQIC
GEO	888.0	977.09	799.37	808.82
SED	112.6	581.33	410.00	465.36
GPD	511.4	618.05	774.65	690.08
DGPD	109.2	199.19	229.31	355.73
	Data 3			
Probability Distribution	AIC	BIC	CAIC	HQIC
GEO	199.6	221.61	187.88	293.25
SED	464.30	409.13	995.80	405.16
GPD	899.45	893.27	929.92	1925.30
DGPD	211.61	324.57	277.71	334.14
	Data 4			
Probability Distribution	AIC	BIC	CAIC	HQIC
GEO	125.62	186.34	399.37	208.82

SED	482.69	652.46	410.00	465.36
GPD	198.42	234.00	274.65	290.08
DGPD	100.02	164.34	129.31	355.73
	Data 5			
Probability Distribution	AIC	BIC	CAIC	HQIC
GEO	221.62	186.34	299.37	258.61
SED	382.69	352.46	417.23	365.43
GPD	298.87	234.00	186.75	179.64
DGPD	211.36	264.34	117.00	115.21
	Data 6			
Probability Distribution	AIC	BIC	CAIC	HQIC
GEO	163.42	129.34	322.11	108.22
SED	211.60	292.43	200.01	337.26
GPD	129.41	223.91	184.65	290.08
DGPD	104.22	101.71	129.31	108.33

Goodness of Fit test result

GEO- Geometric Distribution , SED- Stretched Exponential Distribution GPD- Generalised Pareto Distribution, DGPD – Discretised Generalised Pareto Distribution

The result analysis of the Goodness of fit test (GoF) shows different best fit depending on the test used. However, in our studied cases, Discretised Generalised Pareto Distribution (DGPD) stands out as the best fitted degree distribution of the selected Gene Regulatory networks. It is observed that the Discretised Generalised Pareto Distribution (DGPD) would give the closest estimates of the empirical data. Among the parametric distribution the Discretised Generalised Pareto Distribution (DGPD) is lowest in AIC and has relatively low BIC, CAIC & HQIC.

4.1 Conclusion and Recommendation

The aim of this study is to establish a PDF that can model the degree distribution of biological networks – Gene Regulatory Networks (GRN's) in particular . In the study, we have proposed a Discretised Power law density function for the degree distribution of GRN and also compared with alternative distributions has suggested in literature. Our new distribution provide a more accurate fit to describe the degree distribution of GRN. With more details in the calculations, we arrived at the following conclusion that: the degree distribution of the studied GRN's does not follow Power law nor any of the alternative distribution as suggested by Khanin and Wit but rather a Discretised Generalised Pareto distribution is proposed as the probability distribution to model the degree distribution of GRNS.

Progress can be possible if numerical and analytical studies are blended with proper empirical studies, we recommend an extension on the study of topological measures of different biological networks to divulge more unforeseen window for further review and convincing conclusions. Also, for extension of the work , other types of biological networks such as; Metabolic, cell signaling pathways and Protein Interaction networks should be studied by focusing more on the degree distribution.

REFERENCES

1. Ravasz,E and Barabasi,A.L. (2003) “Hierarchical organization in complex networks”. (Phys.Rev.E67,0261122003)
2. Khanin, R and Wit, E (2006). “How Scale-free are Biological networks” Journal of Computational Biology, Vol 13, pg 810-818.
3. Junker, B. H and Schreiber, N. (2008). “Analysis of Biologıcal Networks”. Wiley & Sons.
4. Ravasz, E., Somera, A.L., Ongru, D.A. and Oltvai, Z.N. (2002). “Hierarchical organization of Modularity in Metabolic Networks”. Vol. 297 no. 5586pp. 1551-1555
5. Guelzim, N., Bottani, S., Bourgine, P and Kepes, F. (2002). “Topological and causal structure of the yeast transcriptional regulatory network” Nature Genetics. Vol. 31, page 60-63
6. Vilda Purutcuoglu and Omolola Odunsi ”Degree Distribution of Real Biological Networks” The Proceedings of The 3rd International Conference on Data Mining, Internet Computing, and Big Data, Konya, Turkey 2016
7. Chakraborty, S. Generating discrete analogues of continuous probability distributions-A survey of methods and constructions. J Stat Distrib App **2,** 6 (2015).
8. Chakraborty Journal of Statistical Distributions and Applications (2015) 2:6 DOI 10.1186/s40488- 015-0028-6
9. C.D. Lai (2013) Issues Concerning Constructions of Discrete Lifetime Models, Quality Technology & Quantitative Management, 10:2, 251-262, DOI: 10.1080/16843703.2013.11673320
10. Hussain T. and Ahmad, M. (2014). “Discrete Inverse Rayleigh Distribution” *Pakistan Journal of Statistics*, 30(2), 203-222.
11. Inusah, S. and Kozubowski, T.J. (2006). “A discrete analogue of the Laplace distribution” *Journal of Statistical Planning and Inference*, 136, 1090-1102.
12. Jazi, M.A., Lai, C.D. and Alamatsaz, M.H. (2010). A discrete Inverse Weibull distribution and estimation of its parameters. Statistical Methodology, 7, 121-132.