A direct proof of entropy-based directed random walk

X.H. Tay[1,], Tole Sutikno[2], Shahreen Kasim[1*], Mohd Farhan Md. Fudzee[1], Rohayanti Hassan[3], C.S. Seah[4]

[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400, Batu Pahat, Johor, Malaysia
[2]Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia
[4]Faculty of Accounting & Management, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Bandar Sungai Long, 43000, Kajang, Malaysia

*shahreen@uthm.edu.my

ABSTRACT: Random walks and Shannon entropy have been studied by probability theorists for decades, and they have applications in many fields of mathematics, physics, chemistry, biology and computer science. This paper is particularly aimed (1) To apply Shannon entropy in Directed Random Walk algorithm in order to calculate the distribution value along a biological pathway, (2) To discover the interrelationship between entropy and connectivity of nodes via entropy-based Directed Random Walk (e-DRW), and (3) To show the role of entropy in revealing the biological insights for a gene and a pathway in the biological network. An equation is introduced to discover the connectivity of nodes in directed graph via probability values calculated from Shannon entropy formula. A direct proof of calculation is presented using entropy-based Directed Random Walk with gene expression data.

Keywords: Directed *Random Walk; Shannon Entropy*

1. INTRODUCTION

Nowadays, random walk is applied in various fields including mathematics, physics and biology. With the development of this graph-theoretical algorithm, random walk on graphs can be efficiently employed to solve many biological problems. Biological networks can be represented as graphs and random walk serves as a useful tool in studying the relationship between nodes in graph. In this paper, we focus on directed random walk model in biological network and present a direct proof of how this graph-theoretical algorithm can help in identifying the interrelationships between entropy and connectivity of nodes in biological networks.

2. METHODOLOGY

Entropy-based Directed Random Walk (e-DRW) is aimed to discover relationship between entropy and connectivity of nodes in directed graph via probability values calculated from Shannon entropy formula. Shannon entropy is proved to be useful in calculating the amount of information or uncertainty of a sequence in genomic data [3]. Entropy of gene is used as a weight parameter to estimate the variability in expression for a single gene.

2.1 *Entropy-based directed random walk*

Given an edge weighted graph G = (V, E, w), where V , E and w denote the vertex set, the edge set and the edge weight respectively, node entropy [4] can be defined as

$$H(vi) = -\sum_{j=1}^{di} P(vi)\ \log_2 P(vi) \qquad (1)$$

where P(vi) represents the probability of gene weight in average between two connected nodes, N1 and N2 with the formula

$$P(vi) = \frac{z(vi)}{\sum_{j=1}^{di} z(vi)} \qquad (2)$$

The entropy-based Directed Random Walk (e-DRW) starts random walker from a single node and the walker transits from its current node to another randomly selected neighbor (forward) node based on edge weights or goes back to previous node with probability r. e-DRW can be defined as

$$\mathrm{H}(\mathrm{v}_{t+1}) = (1 - \mathrm{r})\ \mathrm{E}^{\mathrm{T}}\ \mathrm{H}(\mathrm{v}_t) + \mathrm{r}\ \mathrm{H}(\mathrm{v}_0) \qquad (3)$$

where $\mathrm{H}(\mathrm{v}_t)$ represents transition probability of i^th node which is transmitted from i-1 node. $\mathrm{H}(\mathrm{v}_0)$ is the initial entropy probability vector. r is set to 0.7 as defined by Liu [5] and E^{T} is an entropy edge-weighted adjacency matrix developed from the original directed graph (with edges). $\mathrm{H}(\mathrm{v}_{t+1})$ denotes the final entropy probability vector.

3. RESULTS AND DISCUSSION

This section presents a direct proof of entropy-based Directed Random Walk (e-DRW) by implementing entropy as weight variable with gene expression data (GSE19188) as demonstrated in previous study [6]. Table 1 shows the weight of nodes after data pre-processing using Gene Chip Robust Multi-Array Averaging (GCRMA). The average gene weights between two connected nodes are calculated. Assume a directed graph G, where V represents vertex, V = {1, 2, 3, 4, 5}.

Table 1 Weight of each node that implement in graph G

Nodes	Weight	Average Weight
1	2.338914	5.405962
2	8.47301	7.308555
3	6.1441	4.6235445

4	3.102989	7.2433195
5	11.38365	8.2665215
6	5.149393	5.149393

Based on Table 2, the edge weight of last node (node 6) is remained as the same value because there is no further adjacent node connected to the last node along the pathway. Before calculating the vector of e-DRW, we need to first compute the entropy value for each node. Table 2 shows the calculation of node entropy based on probability of edge weight.

Table 2 Calculation of node entropy based on probability of edge weight

Nodes	Weight	Probability	Node Entropy
1	2.338914	0.142272283	0.400211932079
2	8.47301	0.192344084	0.4574327005688
3	6.1441	0.121680883	0.369788203437
4	3.102989	0.190627238	0.455789726058
5	11.38365	0.217555523	0.4787309283615
6	5.149393	0.135519987	0.3907583305158

The calculation of e-DRW whereby r (restart probability) is sets to 0.7, M (adjacency matrix) is sets to 1 and the initial probability vector is sets to 0 are shown as Table 3 below:

Table 3 Calculation of e-DRW

Nodes	Distribution Vector
H(v1)	(1-0.7) (1) (0.400211932079) + (0.7) (0) = 0.120663579
H(v2)	(1-0.7) (1) (0.120663579) + (0.7) (0.4574327005688) = 0.356221964
H(v3)	(1-0.7) (1) (0.356221964) + (0.7) (0.369788203437) = 0.365718331
H(v4)	(1-0.7) (1) (0.365718331) + (0.7) (0.455789726058) = 0.428768307
H(v5)	(1-0.7) (1) (0.428768307) + (0.7) (0.4787309283615) = 0.463742142
H(v6)	(1-0.7) (1) (0.463742142) + (0.7) (0.3907583305158) = 0.412653474

From the results shown in Table 3, the low value of vector in e-DRW is due to the calculation of entropy. Entropy plays an important role in identifying the level of biological information density in a gene and along the biological pathway. Based on Shannon's information theory in evolutionary biology [7], low entropy value is associated with low information content while high entropy value indicates high information content.

Besides, the increasing vector in entropy-based Directed Random Walk is proved to be related to cancer cell formation as stated by scientific researchers [8]. Thus, by implementing entropy value in entropy-based Directed Random Walk, it can reveal biological insights for a gene and pathway. It also demonstrates a systemic link between gene expression changes at the nodes with entropy changes in biological network.

4. CONCLUSION

Entropy as weight variable plays an important role in revealing the biological insights for a gene and a pathway. According to Shannon's information theory in evolutionary biology, low entropy indicates more predictability and high entropy reflects more uncertainty. The entropy metric is useful to calculate the amount of information or uncertainty of a biological sequence. With the proven results after implementation, it is shown that entropy vector not only applicable in calculating the distribution values along a biological pathway using entropy-based Directed Random Walk, it also disclosed the essential biological information behind the concepts of information theory.

ACKNOWLEDGEMENT

This study was supported by Universiti Tun Hussein Onn Malaysia under REGG FASA 1/2021 (VOT NO. H888).

REFERENCES

[1] Codling, E. A., Plank, M. J., & Benhamou, S. (2008). Random walk models in biology. Journal of the Royal society interface, 5(25), 813-834.

[2] Pearson, K. (1905). The problem of the random walk. Nature, 72(1867), 342 342.

[3] Akhter, S., Bailey, B. A., Salamon, P., Aziz, R. K., & Edwards, R. A. (2013). Applying Shannon's information theory to bacterial and phage genomes and metagenomes. Scientific reports, 3(1), 1-7.

[4] Chen, Z., Dehmer, M., Emmert-Streib, F., & Shi, Y. (2015). Entropy of Weighted Graphs with Randi c Weights. Entropy, 17(6), 3710-3723.

[5] Liu, W., Li, C., Xu, Y., Yang, H., Yao, Q., Han, J., ... & Li, X. (2013). Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. Bioinformatics, 29(17), 2169-2177.

[6] Seah, C. S., Kasim, S., Fudzee, M. F. M., & Mohamad, M. S. (2017, September). A direct proof of significant directed random walk. In IOP Conference Series: Materials Science and Engineering (Vol. 235, No. 1, p. 012004). IOP Publishing.

[7] Adami, C. (2011). The use of information theory in evolutionary biology. arXiv preprint arXiv:1112.3867.

[8] West, J., Bianconi, G., Severini, S., & Teschendorff, A. E. (2012). Differential network entropy reveals cancer system hallmarks. Scientific reports, 2(1), 1-8.