Comparison of Similarity Coefficients on Morphological Rodent Tuber

[1,2]Iwan Binanto
[1]Computer Science Department,
BINUS Graduate Program – Doctor of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480

[2]Informatics Department,
Sanata Dharma University,
Yogyakarta, Indonesia 55002
iwan@usd.ac.id

[1]Harco Leslie Hendric Spits Warnars, [2]Bahtiar Saleh Abbas, [3]Yaya Heryadi
Computer Science Department,
BINUS Graduate Program,
Doctor of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
[1]spits.hendric@binus.ac.id,
[2]bahtiars@binus.edu,
[3]yayaheryadi@binus.edu

[1,2]Nesti Fronika Sianipar
[1]Food Technology Department,
Faculty of Engineering,
Bina Nusantara University,
Jakarta, Indonesia 11480

[2]Research Interest Group
Biotechnology,
Bina Nusantara University,
Jakarta, Indonesia 11480
nsianipar@binus.edu

Lukas
Cognitive Engineering Research Group (CERG),
Faculty of Engineering,
Universitas Katolik Indonesia Atma Jaya,
Jakarta, Indonesia 12930
lukas@atmajaya.ac.id

Horacio Emilio Perez Sanchez
Bioinformatics and High Performance Computing Research Group (BIO-HPC), Universidad Católica de Murcia (UCAM), Guadalupe, Spain 30107
hperez@ucam.edu

***Abstract*— Many comparisons of similarity coefficient done by researchers, especially in the field of biology. This comparison aims to find the most appropriate similarity coefficient for some cases. Many results found that Sorensen-dice coefficient and Jaccard coefficient is close or even identical. But Jaccard coefficient can not handle properly for sets with real-value or weighted sets or any pair of vectors. So, Jaccard coefficient redefined as Generalized Jaccard Coefficient. This paper shows the correlation between Sorensen-dice coefficient with Generalized Jaccard Coefficient using Spearman's correlation as predecessors research did and using ANOVA to ensure the results. This research find that the comparison between them is less similar from predecessors research.**

Keywords*— *Generalized Jaccard Similarity, Sorensen-Dice Similarity, similarity coefficient, comparison, rodent tuber

I. INTRODUCTION

The similarity is necessary to examine the objects of investigation; in this case, the mutant of Rodent Tuber (Typhonium flagelliforme Lodd.) derived from breeding with its parent, called control plant. The research of Rodent Tuber was performed by Sianipar, et al. in [1]–[5] utilizing NTSys, which is proprietary software. One of their research objectives is to find similarity. By the discovery of similarity, it will be easier to find its dissimilarity, because the real purpose of the breeding is to find the diversity of produced mutants [6] [7].

One of Sianipar's investigations is the morphological observation of Rodent Tuber, which has been given gamma irradiation. According to this investigations, gamma irradiation at 6 Gy's dose was able to increase the number of shoots and leaves, and also the height of the plant of the Rodent clones which are compared to the control plants [4]. This paper using the data from [4] as in Table I.

Sianipar et al. measure the similarity between the mutants of Rodent Tuber and the control plant using Sorensen-Dice coefficient [1]–[5]. The formula of Sorensen-Dice coefficient is:

$$SDC(A,B) = \frac{2\,|A \cap B|}{|A| + |B|} \quad (1)$$

Beside of Sorensen-Dice coefficient, there are many coefficient similarities, one of them is Jaccard coefficient which had approximately identical results in [8], [9] or has close result in [10] or a very close result in [11] to Sorensen-Dice coefficient. The Jaccard coefficient created for analyses in phytology [12] and works well with binary data as well as Sorensen-Dice coefficient. Many research using Jaccard coefficient for measuring similarities in a various field [8]–[17]. The formula of Jaccard coefficient is:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

Jaccard coefficient is simple and effective in many applications [13], [18] but it can not handle properly for sets with real-value or weighted sets [18] or any pair of vectors [19], therefore it redefine and explained well as the Generalized Jaccard Coefficient in [19], for short we call it GJS, and also introduced and used in [18]–[22] as:

$$GJS(A,B) = \frac{\sum_i min(A_i, B_i)}{\sum_i max(A_i, B_i)} \quad (3)$$

978-1-5386-9422-0/18/$31.00 ©2018 IEEE

This paper discusses Generalized Jaccard Coefficient compared to Sorensen-Dice Coefficient (result from proprietary software namely NTSys) using Spearman's correlation as [8]–[11] did.

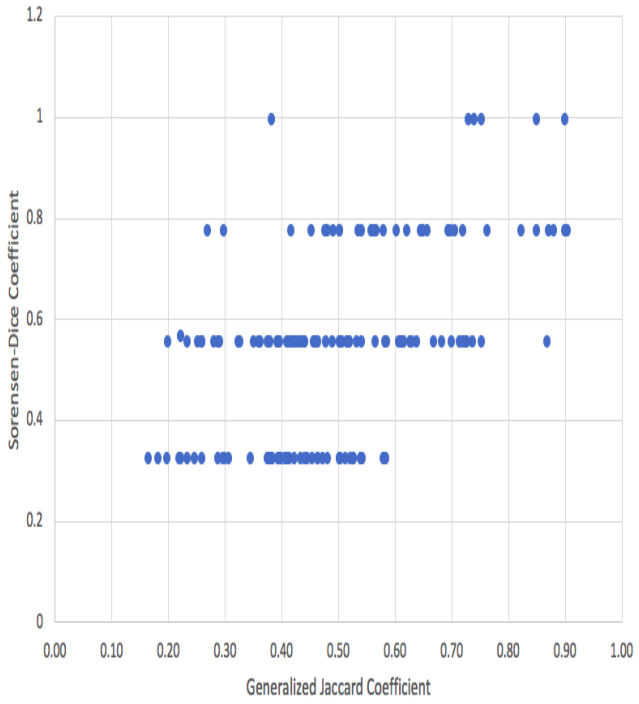

Fig. 1. Plot Sorensen-Dice and Generalized Jaccard coefficient

II. Literature Review

Rodent Tuber is a plant native to Indonesia that has been used as traditional medicine for many years. This plant contains detoxification and anti-cancer compounds. These anticancer compounds exist in all parts of the plant, including roots, tubers, stems, and leaves. Unfortunately, this plant does not have much genetic diversity, so it becomes an obstacle regarding obtaining plants that have higher anticancer compounds. Sianipar et al. began to develop mutants using gamma radiation [23]. To test the genetic diversity of the mutant plants produced, Sianipar et al. did a similarity test using the NTSys software with Sorensen-Dice coefficient [1]–[5].

Duarte et al. in [8] compared eight similarity coefficients using the Spearman's correlation and dendrogram to test similarity in common beans based on the RAPD marker. One of the result is Sorensen-Dice, and the Jaccard coefficient has an identical result. Murguia et al. in [9] compared nine similarity coefficients to estimate the effect of biogeographic classification; the result is Sorensen-Dice and Jaccard coefficient had identical results. Silva et al. in [10] compared eight similarity coefficient using Spearman's correlation, and the result is Sorensen-Dice and Jaccard coefficient had a close result. Dalirsefat et al. in [11] compared three similarity coefficient one of the comparison tools is the Spearman's correlation and of the result of a correlation value between Sorensen-Dice and Jaccard coefficient is one which means exactly same.

Shrivastava (2016) in [19] said that GJS (A, B) is often used to compare web documents, histograms (especially images), gene sequences, etc. Those are weighted sets or pair of vectors. Weighted sets or any pair of vectors are more commonly found than binary sets. If *A* and *B* are binary or sets, then the similarity measure is called Jaccard coefficient as mentioned in [8]–[17]. According to [18], [19], Jaccard coefficient cannot handle properly for sets with real-value called weighted sets or any pair of vectors.

III. Method

This paper uses raw data and Sorensen-Dice similarity table from [4] as in Table I and Table IV respectively. Generalized Jaccard coefficient calculated with formula (3) and have a result as in Table V. It was done using Microsoft Excel.

To calculate the correlation, each similarity table converted to be 1 column, so we have two columns which are Generalized Jaccard column and Sorensen-Dice column. From here, we can plot the data as in Fig. 1.

Then Spearman's correlation calculated to find the value of correlation between Table IV and Table V. It done using MATLAB with the script as below:

```
a = xlsread('Book2.xlsx','A:A')
b = xlsread('Book2.xlsx','B:B')
[RHO] = corr(a,b,'Type','Spearman');
```

The script generates RHO value 0.5052, which is the value of Spearman's correlation.

To ensure the correlation between Generalized Jaccard coefficient and Sorensen-Dice coefficient, we construct a hypothesis which are:

H_0: No correlation between Generalized Jaccard coefficient and Sorensen-Dice coefficient
H_a: There is a correlation between Generalized Jaccard coefficient and Sorensen-Dice coefficient

This hypothesis evaluated with ANOVA using Microsoft Excel and the result provided as in Table II.

TABLE I. Raw Data From [4]

Clone	Shoot	Leaf	Plant Height (cm)
control	0	1	3.5
6-3-3-6	1	6	4
6-9-3	2.5	3.5	4
6-9-4	0.4	4	12.5
6-2-5-3	0.5	7	12
6-3-2-5	1.5	8	13.5
6-1-1-2	3.5	2	6
6-9-1	2.5	11	4.5
6-2-4-1	0	2	3
6-6-3-7	0.5	6	7.5
6-6-3-6	1	6	12.5
6-2-7	0	5.5	12
6-2-6-3	0	5	5.5
6-1-2	4.5	15	8.3
6-1-1-6	1	2	5
6-2-8-2	2.5	11.5	6.5
6-9-5	0	12.5	10.3
6-3-3-10	0	1.5	7.5

978-1-5386-9422-0/18/$31.00 ©2018 IEEE

IV. Results And Discussions

Duarte et al. in [8] concluded that the result is Sorensen-Dice and the Jaccard coefficient has an identical result. Murguia et al. in [9] had a result that Sorensen-Dice and Jaccard coefficient had identical results. Silva et al. in [10] concluded that Sorensen-Dice and Jaccard coefficient had a close result. Dalirsefat et al. in [11] had the result that correlation value between Sorensen-Dice and Jaccard coefficient is one which means exactly same. They made a comparison between Sorensen-dice coefficient and Jaccard coefficient where both are used binary data. This paper uses Generalized Jaccard coefficient for real-value data. According to [19], Jaccard coefficient similar to Generalized Jaccard coefficient. But in this research, the result of Spearman's correlation is 0.5052 as above, which means there is a moderate positive correlation, as in Table III [24]. It is not close, very close, nor even identical.

TABLE II. ANOVA Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
GJS	153	75.8055923	0.49546139	0.02869388
DICE	153	84.09	0.54960784	0.03544327

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.22428566	1	0.22428566	6.99393895	0.0086034	3.87222952
Within Groups	9.74884679	304	0.03206857			
Total	9.97313245	305				

TABLE III. Interpreting Correlation Coefficient [24]

Correlation Value	Interpretation
0.90 to 1.00 (-0.90 to -1.00)	Very High Positive/Negative Correlation
0.70 to 0.90 (-0.70 to -0.90)	High Positive/Negative Correlation
0.50 to 0.70 (-0.50 to -0.70)	Moderate Positive/Negative Correlation
0.30 to 0.50 (-0.30 to -0.50)	Low Positive/Negative Correlation
0.00 to 0.30 (0.00 to -0.30)	Negligible Correlation

TABLE IV. Result Of Sorensen-Dice Coefficient

	control	6-3-3-6	6-9-3	6-9-4	6-2-5-3	6-3-2-5	6-1-1-2	6-9-1	6-2-4-1	6-6-3-7	6-6-3-6	6-2-7	6-2-6-3	6-1-2	6-1-1-6	6-2-8-2	6-9-5	6-3-3-10
control	1																	
6-3-3-6	0.56	1																
6-9-3	0.78	0.56	1															
6-9-4	0.78	0.33	0.56	1														
6-2-5-3	0.56	0.33	0.33	0.78	1													
6-3-2-5	0.33	0.78	0.33	0.56	0.78	1												
6-1-1-2	0.56	0.33	0.78	0.56	0.33	1	1											
6-9-1	0.56	0.56	0.78	0.33	0.33	0.33	0.56	1										
6-2-4-1	1	0.56	0.78	0.78	0.56	0.33	0.56	0.56	1									
6-6-3-7	0.56	0.56	0.33	0.56	0.78	0.56	0.56	0.33	0.56	1								
6-6-3-6	0.33	0.78	0.33	0.56	0.78	1	0.33	0.33	0.33	0.56	1							
6-2-7	0.56	0.56	0.33	0.78	1	0.78	0.33	0.33	0.56	0.78	0.78	1						
6-2-6-3	0.56	0.56	0.33	0.56	0.78	0.56	0.56	0.33	0.56	1	0.56	0.78	1					
6-1-2	0.33	0.33	0.56	0.33	0.33	0.33	0.78	0.78	0.33	0.56	0.33	0.33	0.56	1				
6-1-1-6	0.56	0.56	0.56	0.56	0.33	0.56	0.78	0.33	0.56	0.56	0.56	0.33	0.56	0.56	1			
6-2-8-2	0.33	0.33	0.56	0.33	0.33	0.33	0.78	0.78	0.33	0.56	0.33	0.33	0.56	1	0.56	1		
6-9-5	0.56	0.33	0.33	0.78	0.78	0.56	0.33	0.56	0.57	0.56	0.56	0.78	0.56	0.56	0.33	0.56	1	
6-3-3-10	0.78	0.33	0.56	0.78	0.56	0.33	0.78	0.33	0.78	0.78	0.33	0.56	0.78	0.56	0.78	0.56	0.56	1

TABLE V. Result Of Generalized Jaccard Coefficient

	control	6-3-3-6	6-9-3	6-9-4	6-2-5-3	6-3-2-5	6-1-1-2	6-9-1	6-2-4-1	6-6-3-7	6-6-3-6	6-2-7	6-2-6-3	6-1-2	6-1-1-6	6-2-8-2	6-9-5	6-3-3-10
control	1																	
6-3-3-6	0.75	1																
6-9-3	0.45	0.68	1															
6-9-4	0.27	0.43	0.42	1														
6-2-5-3	0.23	0.53	0.37	0.82	1													
6-3-2-5	0.20	0.48	0.38	0.73	0.85	1												
6-1-1-2	0.39	0.45	0.65	0.42	0.38	0.38	1											
6-9-1	0.25	0.61	0.56	0.34	0.47	0.52	0.44	1										
6-2-4-1	0.73	0.45	0.50	0.30	0.26	0.22	0.43	0.3	1									
6-6-3-7	0.32	0.72	0.50	0.63	0.72	0.61	0.50	0.5	0.36	1								
6-6-3-6	0.23	0.56	0.40	0.87	0.90	0.85	0.41	0.4	0.26	0.72	1							
6-2-7	0.26	0.50	0.38	0.87	0.90	0.76	0.38	0.4	0.29	0.70	0.90	1						
6-2-6-3	0.43	0.72	0.58	0.53	0.54	0.46	0.52	0.5	0.48	0.75	0.54	0.60	1					
6-1-2	0.16	0.40	0.36	0.40	0.50	0.54	0.41	0.6	0.18	0.50	0.48	0.44	0.38	1				
6-1-1-6	0.56	0.58	0.64	0.42	0.38	0.35	0.70	0.4	0.63	0.52	0.41	0.38	0.61	0.29	1			
6-2-8-2	0.22	0.54	0.49	0.41	0.54	0.58	0.49	0.9	0.24	0.60	0.51	0.46	0.51	0.74	0.39	1		
6-9-5	0.20	0.42	0.30	0.56	0.69	0.67	0.30	0.6	0.22	0.58	0.63	0.64	0.46	0.70	0.29	0.71	1	
6-3-3-10	0.50	0.38	0.41	0.53	0.46	0.39	0.58	0.3	0.47	0.64	0.46	0.51	0.56	0.32	0.62	0.37	0.39	1

978-1-5386-9422-0/18/$31.00 ©2018 IEEE

V. CONCLUSIONS

In previous research on the comparison between Jaccard coefficient and Sorensen-Dice coefficient [8]–[11], showing the results that both have close correlations up to identical. But Jaccard coefficient can not handle properly for sets with real-value or weighted sets [18] or any pair of vectors [19], so the Generalized Jaccard coefficient is used. In this study, Sorensen-Dice coefficient compared with Generalized Jaccard coefficient and the result is there is a moderate correlation with the Spearman's correlation value is 0.5052. This result less similar than the previous research in [8]–[11]. Because of this, we are not recommending to use Generalized Jaccard coefficient if already use Sorensen-Dice coefficient to avoid confusion.

REFERENCES

[1] N. F. Sianipar, Ariandana, and W. Maarisit, "Detection of Gamma-Irradiated Mutant of Rodent Tuber (Typhonium flagelliforme Lodd) In Vitro Culture by RAPD Molecular Marker," vol. 14, pp. 285–294, 2015.

[2] D. Laurent, N. F. Sianipar, Chelen, Listiarini, and A. Wantho, "Analysis of Genetic Diversity of Indonesia Rodent Tuber (Typhonium flagelliforme Lodd.) Cultivars Based on RAPD Marker)," in *The 3rd International Conference on Biological Science 2013 (The 3rd ICBS-2013)*, 2015, vol. 2, pp. 139–145.

[3] N. F. Sianipar, D. Laurent, R. Purnamaningsih, and I. Darwati, "SHORT COMMUNICATION Genetic Variation of the First Generation of Rodent Tuber (Typhonium flagelliforme Lodd .) Mutants Based on RAPD Molecular Markers," vol. 22, no. 2, pp. 98–104, 2015.

[4] N. F. Sianipar, R. Purnamaningsih, D. L. Gumanti, Rosaria, and M. Vidianti, "Analysis of Gamma Irradiated-Third Generation Mutants of Rodent Tuber (Typhonium flagelliforme Lodd .) Based on Morphology , RAPD , and GC-MS Markers," *Pertanika J. Trop. Agric. Sci.*, vol. 40, no. 1, pp. 185–202, 2017.

[5] N. F. Sianipar, R. Purnamaningsih, D. L. Gumanti, Rosaria, and M. Vidianti, "Analysis Of Gamma Irradiated Fourth Generation Mutant Of Rodent Tuber (Typhonium Flagelliforme Lodd.) Based On Morphology And RAPD Markers," *J. Teknol.*, vol. 78, no. 5–6, pp. 41–49, 2016.

[6] R. Hesananda *et al.*, "Supervised Classification Karakter Morfologi Tanaman Keladi Tikus (Typhonium Flagelliforme) Menggunakan Database," *J. Sist. Komput.*, vol. 7, no. 2, pp. 50–58, 2017.

[7] T. Siswanto *et al.*, "The Genomic Plant Warehouse Framework: A Systematic Literature Review," *Proc. 2017 Int. Conf. Inf. Manag. Technol.*, no. November, pp. 244–248, 2017.

[8] J. M. Duarte, J. B. Dos Santos, and L. C. Melo, "Comparison of similarity coefficients based on RAPD markers in the common bean," *Genet. Mol. Biol.*, vol. 22, no. 3, pp. 427–432, 1999.

[9] M. Murguia and J. L. Villasenor, "Estimating the effect of the similarity coefficient and the cluster algorithm on biogeographic classifications," *Ann. Bot. Fenn.*, vol. 40, no. December, pp. 415–421, 2003.

[10] A. da Silva Meyer, A. A. F. Garcia, A. Pereira de Souza, and C. Lopes de Souza, "Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L)," *Genet. Mol. Biol.*, vol. 27, no. 1, pp. 83–91, 2004.

[11] S. B. Dalirsefat, A. da S. Meyer, and S. Z. Mirhoseini, "Comparison of Similarity Coefficients used for Cluster Analysis with Amplified Fragment Length Polymorphism Markers in the Silkworm , Bombyx mori," *J. Insect Sci.*, vol. 9, no. 71, pp. 1–8, 2009.

[12] P. Jaccard, "The distribution of the flora in the alphine zone," *New Phytol.*, vol. XI, no. 2, pp. 37–50, 1912.

[13] S. Pal, F. Yu, T. J. Moore, R. Ramanathan, A. Bar-Noy, and A. Swami, "An efficient alternative to Ollivier-Ricci curvature based on the Jaccard metric," pp. 1–22, 2017.

[14] V. Thada and V. Jaglan, "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm," *Int. J. Innov. Eng. Technol.*, vol. 2, no. 4, pp. 202–205, 2013.

[15] S. Kosub, "A note on the triangle inequality for the Jaccard distance," *arXiv1612.02696v1 [cs.DM] 8 Dec 2016 A*, no. 1, pp. 1–5, 2016.

[16] D. Fogaras and B. Rácz, "Scaling link-based similarity search," in *Proceedings of the 14th international conference on World Wide Web - WWW '05*, 2005, p. 641.

[17] C. S. Loh, I. H. Li, and Y. Sheng, "Comparison of similarity measures to differentiate players' actions and decision-making profiles in serious games analytics," *Comput. Human Behav.*, vol. 64, pp. 562–574, 2016.

[18] W. Wu, B. Li, L. Chen, and C. Zhang, "Consistent Weighted Sampling Made More Practical.," in *2017 International World Wide Web Conference Committee (IW3C2)*, 2017, pp. 1035–1043.

[19] A. Shrivastava, "Exact Weighted Minwise Hashing in Constant Time," *arXiv Prepr. arXiv1602.08393*, no. 2, 2016.

[20] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," *Proc. thiry-fourth Annu. ACM Symp. Theory Comput. - STOC '02*, p. 380, 2002.

[21] V. Kashyap, D. B. Brown, B. Liblit, D. Melski, and T. Reps, "Source Forager: A Search Engine for Similar Source Code," 2017.

[22] Z. Shirzadi *et al.*, "Enhancement of automated blood flow estimates (ENABLE) from arterial spin-labeled MRI," *J. Magn. Reson. Imaging*, vol. 47, no. 3, pp. 647–655, 2017.

[23] N. F. Sianipar, A. Wantho, Rustikawati, and W. Maarisit, "The Effects of Gamma Irradiation on Growth Response of Rodent Tuber (Typhonium flagelliforme Lodd .) Mutant in In Vitro Culture," *HAYATI J. Biosci.*, vol. 20, no. 2, pp. 51–56, 2013.

[24] M. M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, 2012.