Assessment of Data Reliability
of Wireless Sensor Network for Bioinformatics

Ting Dong

School of Information Engineering
Yulin University
Shanxi Yulin, China
E-mail: 2367094502@qq.com

Abstract: As a focal point of biotechnology, bioinformatics integrates knowledge from biology, mathematics, physics, chemistry, computer science and information science. It generally deals with genome informatics, protein structure and drug design. However, the data or information thus acquired from the main areas of bioinformatics may not be effective. Some researchers combined bioinformatics with wireless sensor network (WSN) into biosensor and other tools, and applied them to such areas as fermentation, environmental monitoring, food engineering, clinical medicine and military. In the combination, the WSN is used to collect data and information. The reliability of the WSN in bioinformatics is the prerequisite to effective utilization of information. It is greatly influenced by factors like quality, benefits, service, timeliness and stability, some of them are qualitative and some are quantitative. Hence, it is necessary to develop a method that can handle both qualitative and quantitative assessment of information. A viable option is the fuzzy linguistic method, especially 2-tuple linguistic model, which has been extensively used to cope with such issues. As a result, this paper introduces 2-tuple linguistic representation to assist experts in giving their opinions on different WSNs in bioinformatics that involve multiple factors. Moreover, the author proposes a novel way to determine attribute weights and uses the method to weigh the relative importance of different influencing factors which can be considered as attributes in the assessment of the WSN in bioinformatics. Finally, an illustrative example is given to provide a reasonable solution for the assessment.

Keywords: Bioinformatics, Assessment, Reliability, Wireless sensor network, 2-tuple linguistic term set.

Introduction

Owing to the rapid development of science and technology, bioinformatics, a focal point of biotechnology, integrates much knowledge from biology, mathematics, physics, chemistry, computer science and information science [1, 11]. As the fruit of human genome project, the branch of information science deals with the acquisition, processing, storing, allocation, analysis and interpretation of biological information. It is mainly aimed at disclosing the complexity of the genome information structure and the basic rule of genetic language [5, 12].

Up to now, bioinformatics has developed through three stages: the pre-genomic era, the genomic era and the post-genomic era. The first stage focused on the establishment of biological database, the development and application of retrieval tool, and the comparison and analysis of DNA and protein sequence. The second stage emphasized on the identification of new genes by measuring and analyzing nucleotide sequences. The third stage centered on the mapping and annotation of human genomes. For example, Nailwai and Chauhan [4] employed bioinformatics tools to analyze the consequences of non-synonymous SNPs of USP9Y gene in human. Senturk et al. applied a resource allocation for bioinformatics in multi-cloud environments [8].

The wireless sensor network (WSN) is often used in synergy with bioinformatics because of its function of collection and aggregation of bioinformatics information [3, 6, 8-10]. Consisting of multiple tiny nodes, the WSN can persistently detect the environment around the sensors and send sensor data to base station via multi-hop routing. The acquisition of surrounding information relies on a mass of sensors, the number of which adds up to thousands. Nevertheless, a WSN on such a large scale may have lower accuracy and poorer data quality, which may impact any further analysis based on its results. The situation would deteriorate due to the dynamic nature of the WSN.

Therefore, the WSN has to be more reliable before it is integrated with bioinformatics to make the data utilization in bioinformatics more efficient. The influencing factors on the WSN's reliability will be analyzed in details in the next section.

Targeted at the above problem, this paper introduces the 2-tuple linguistic model, an immensely popular method [2, 7], to help with the expression of the preference information related to the WSN in bioinformatics. First, the factors that may impact the assessment of WSN's reliability are obtained by analyzing the relevant concepts of WSN in bioinformatics; second, some definitions and the operation rule of 2-tuple linguistic term sets are introduced, and an attribute weighting method is developed to weigh the relative importance of the attributes of the assessment; third, two operators are imported to integrate the assessment information provided by experts [13]; finally, the reliability of WSN in bioinformatics is assessed with the proposed linguistic method to obtain the ranking order of the attributes. Besides, the proposed assessment framework is verified through the application to several WSNs.

The following sections of this paper mainly discuss these issues. Section 2 determines the attributes of data reliability of the WSN in bioinformatics. Section 3 develops the proposed 2-tuple fuzzy linguistic model and introduces the relevant concepts. Section 4 applies the established 2-tuple linguistic method to assessing the WSN's reliability based on the index system. The conclusion and relative references are provided in Section 5.

Attributes of WSN data reliability assessment

For the better application to the military, the WSN was extensively studied in the 1970s. In recent decades, however, this technology has enjoyed tremendous popularity in academic and practical fields, especially bioinformatics.

Since the founding of the National Center for Biotechnology Information in the US, numerous related organizations have been established, such as the European Bioinformatics Institute in Europe, and the Nippon Biotechnology Information Center in Japan. The emergence of these institutes has attracted much attention across the world. It is also worth mentioning that the genetic sequencing technology has been updated constantly in the same period, from the second generation to the third generation. Against this backdrop, bioinformatics came into being, as can be seen in Fig. 1.

As mentioned in the Introduction, the future of bioinformatics hinges on the data reliability of the WSN in bioinformatics. According to previous research, the data reliability of the WSN in bioinformatics should be assessed from the aspects of quality, benefits, service, timeliness and stability, as can be seen in Fig. 2.
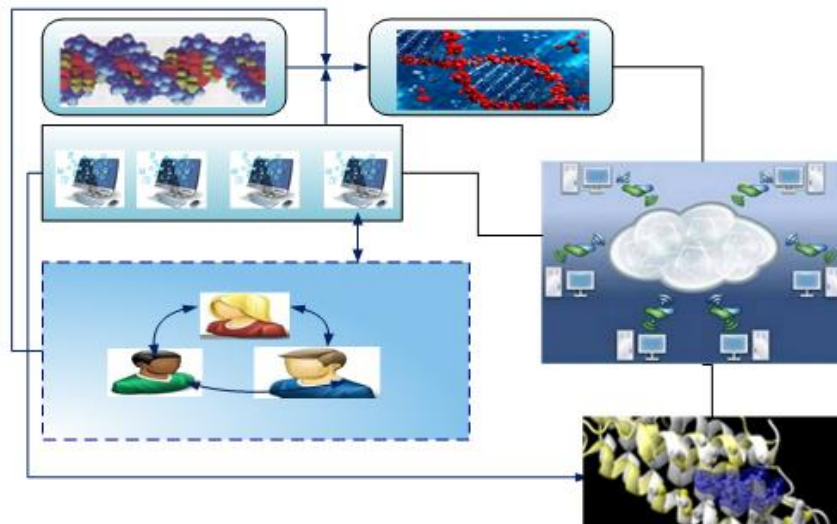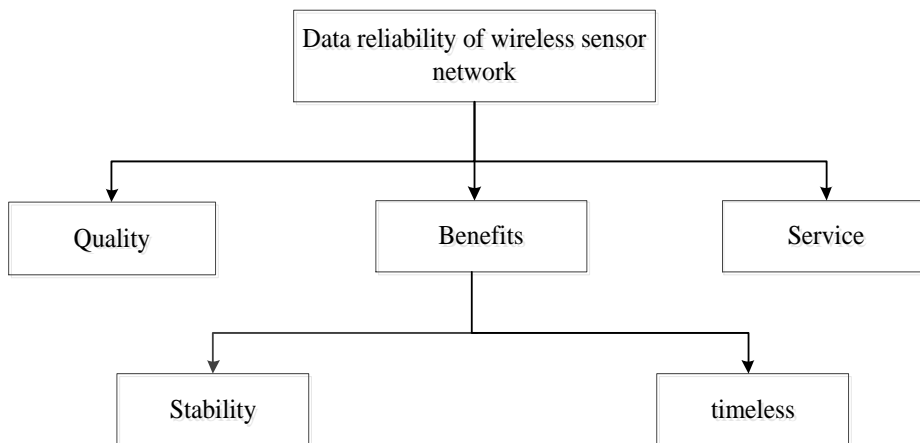
Fig. 1 The emergence of bioinformatics

Fig. 2 The attributes of WSN data reliability assessment

2-tuple linguistic term sets

This section introduces the 2-tupe fuzzy linguistic term sets, and, on this basis, constructs a representation model for decision-making with the aim to obtain the decision information of experts.

Relative concepts of 2-tuple linguistic representation model

There is a common defect in classical linguistic computing models and linguistic term sets: the lack of experts' decision information. If such models are used directly by inexperienced researchers, the results are very likely to be false. To overcome the defect, the author provides the following definitions of the 2-tuple linguistic term sets:

Definition 1. Let there be a linguistic set denoted by $L = \{l_\chi | \chi = 0, 1,\ldots, \varphi\}$ with odd cardinality, where $\gamma \in [0, \varphi]$ is a value that supports the results of linguistic aggregation operation, and $\chi \in [-0.5, 0.5]$. Then, the 2-tuple linguistic term set can be generated by combining parameters γ and χ:

$\Delta: [0, \varphi] \rightarrow \overline{L}$, $\overline{L} \times [-0.5, 0.5)$,

$\Delta(\gamma) = (l_m, \chi)$ with $n = \text{round}(\gamma)$ and $\chi = \gamma - n$,

where round($\cdot$) stands for the rounding operation that assigns the linguistic term l_n to the closest term γ; χ denotes the value of the symbolic translation.

In addition,

$$\Delta^{-1}: \overline{L} \times [-0.5, 0.5) \to [0, \varphi],$$

$$\Delta^{-1}(l_n, \chi) = \chi + n = \gamma.$$

Function Δ^{-1} is the mapping of a linguistic term to a value, and $\Delta^{-1}(l_n, \chi)$ returns the linguistic term to its equivalent value $\gamma \in [0, \varphi]$, where (l_n, χ) is denoted by l_n.

It is obvious that if χ equals to 0, the 2-tuple linguistic term set could be reduced to the classical linguistic term set:

$$l_n \in L \Rightarrow (l_n, 0) \in \overline{L}.$$

Moreover, a 2-tuple negation operator can be obtained by $Neg(l_n, \chi) = \Delta(\varphi - \Delta^{-1}(l_n, \chi))$.

In light of the above definition of 2-tuple linguistic term set and in reference to [2], the operational rules are defined as follows:

Given that the two linguistic terms l_n, $l_m \in \overline{L}$ and two parameters a_1, $a_2 \in [0, 1]$, the following can be obtained:

(1) $l_n \oplus l_m = l_{n+m}$;

(2) $l_n \oplus l_m = l_m \oplus l_n$;

(3) $al_n = l_{an}$;

(4) $(a_1 + a_2)l_n = a_1 l_n \oplus a_2 l_n$;

(5) $a(l_n \oplus l_m) = al_n \oplus al_m$.

If the two linguistic terms are denoted by (l_n, χ_1) and (l_m, χ_2), the following can be obtained:

(1) If $n < m$, then $(l_n, \chi_1) < (l_m, \chi_2)$;

(2) If $n > m$, then:

when $\chi_1 = \chi_2$, (l_n, χ_1) and (l_m, χ_2) contain the same linguistic information;

when $\chi_1 < \chi_2$, $(l_n, \chi_1) < (l_m, \chi_2)$;

when $\chi_1 > \chi_2$, $(l_n, \chi_1) > (l_m, \chi_2)$.

An example is given to make a clear illustration of the meaning of 2-tuple linguistic term set:

Suppose a student has to select an appropriate laptop from the following four brands: Lenovo, Dell, Apple and Acer, and define the fuzzy linguistic term set as $L = \{l_0$, Extremely Bad (*EB*); l_1, Very Bad (*VB*); l_2, Bad (*B*); l_3, Middle (*M*); l_4, Good (*G*); l_5, Very Good (*VG*); l_6, Extremely Good (*EG*)\}. Based on the linguistic term set and his/her own preferences, the student's assessment of the four laptops would be:

$L_{Lenovo} = \{l_4, 0.3\}$, $l_{Dell} = \{l_4, 0.1\}$, $l_{Apple} = \{l_5, -0.2\}$, and $l_{acer} = \{l_3, 0.4\}$.

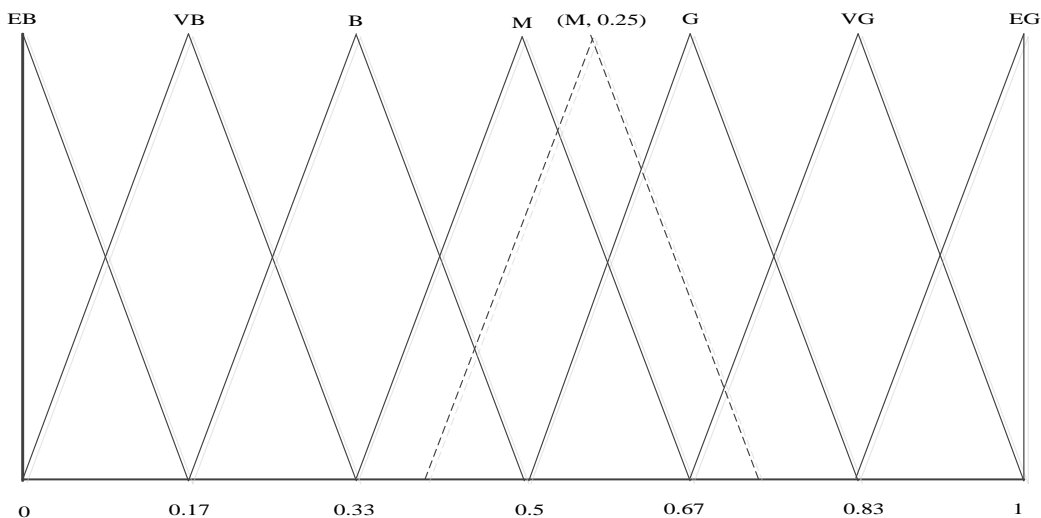The 2-tuple term sets are illustrated in Fig. 3.

Fig. 3 2-tuple linguistic term set

In light of the assessment, the student decides to choose Apple. If the student wants to make a more detailed assessment of these laptops, he/she could import some attributes, namely price, quality, brand, etc.

The weights of attributes and experts involved in the assessment
For the purpose of assessing the performance of bioinformatics, the attributes and experts are weighed in the following manner.

The weight of an attribute or an expert reveals the relative importance of the attribute or expert against other attributes or experts. Different weights may result in vastly different decisions. Therefore, it is crucial to identify the weight of each attribute or expert involved in the reliability assessment of WSN in bioinformatics.

Based on previous studies on attribute or expert weighting, there are two types of weight determination methods: subjective and objective. The subjective method relies heavily on subjective judgments, work experience or knowledge of experts or decision makers. Typical examples are eigenvector algorithm, linear/non-linear programming model and so on. In other words, subjective weights are directly provided by experts or decision makers. In contrast, the objective method rests on the assessment information of alternatives for each attribute given by experts or decision makers. Typical examples include deviation method, distance measure, entropy measure, correlation coefficient and the alike. The methods should be selected in accordance of the specific circumstances of the assessment. This paper chooses

the objective method to determine the weights of attributes, and the subjective method to identify the weights of experts.

Definition 2. Let there be a set of attributes denoted by $C = \{c_1, c_2, \ldots, c_t\}$ and 2-tuple linguistic terms of each attribute. The deviation measure of these attributes can be defined as:

$$Dev(c_1, c_2, \ldots, c_t) = d(L_n, L_m) = \Delta \, | \, \Delta^{-1}(l_n, \chi) - \Delta^{-1}(l_m, \chi)|. \tag{1}$$

The weights of attributes are obtained by the above Eq. (1)

$$Max \sum_{j=1}^{t} \frac{wet_i d\left(L_i, L_j\right)}{t-1}$$

s.t. $0 \leq wet_i \leq 1$

$$\sum_{i=1}^{n} wet_i = 1$$

The attribute weights are determined in accordance with the above optimization model.

Aggregation of linguistic information

For the purpose of making an overall linguistic assessment of all alternatives, the author proposes the operators of 2-tuple linguistic term sets of each attribute.

Definition 3. Let $L = \{l_\chi | \chi = 0, 1, \ldots, \varphi\}$ be a 2-tuple linguistic term set, and $wet = \{wet_1, wet_2, \ldots, wet_t\}$ be a collective weight vector satisfying $0 \leq wet_i \leq 1$ and $\sum_{i=1}^{n} wet_i = 1$. The weight average operator of 2-tuple linguistic terms is developed in the following manner:

$$LWAO(c_1, c_2, \ldots, c_t) = \sum_{i=1}^{n} wet_i L_{ij} = \Delta(\sum_{i=1}^{t} wet_i \Delta^{-1} \Delta(\gamma_{ij})), \tag{2}$$

where $\Delta(\gamma_{ij}) = (L_{ij}, \varphi_{ij})$. If $wet_i = 1/n$, the weight average operator will be equal to the arithmetic average operator.

Definition 4. Let $L = \{l_\chi | \chi = 0, 1, \ldots, \varphi\}$ be a 2-tuple linguistic term set. The arithmetic average operator of 2-tuple linguistic terms is developed in the following manner:

$$LWAO(c_1, c_2, \ldots, c_t) = \sum_{i=1}^{n} \frac{1}{n} L_{ij} = \Delta(\sum_{i=1}^{t} \frac{1}{n} \Delta^{-1} \Delta(\gamma_{ij})), \tag{3}$$

where $\Delta(\gamma_{ij}) = (L_{ij}, \varphi_{ij})$.

The assessment process

In this section, the relability assessment of the WSN in bioinformatics is explained step by step.

Step (1): The decision makers or experts identify a linguistic term set, construct t attributes, and provide q alternatives.

Suppose the relability assessment of the WSN in bioinformatics has q alternatives Q_j ($j = 1, 2, …, q$) and t attributes Att_i ($i = 1, 2, …, t$), and denote the relative weights of the t attributes by $wet_i = wet_1, wet_2, …, wet_t$. There is:

$$0 \le wet_i \le 1 \text{ and } \sum_{i=1}^{n} wet_i = 1. \tag{4}$$

Then, construct an assessment matrix denoted by $Q = [Q_{ij}]_{q \times t}$, where the assessment alternatives provided by a decision maker are denoted by A_{ij} ($j = 1, 2, …, q$; $I = 1, 2, …, t$). In this case, a normal decision making matrix $Q = [Q_{ij}]_{q \times t}$, can be generated as follows:

$$Q_{q \times t} = \begin{pmatrix} Q_{11} & Q_{12} & \cdots & Q_{1t} \\ Q_{21} & Q_{22} & \cdots & Q_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{q1} & Q_{q2} & \cdots & Q_{qt} \end{pmatrix}. \tag{5}$$

Step (2): Collect the assessment information given by experts or decision makers to construct an asessment matrix.

Step (3): Calculate the weight of each attribute and expert by the above-mentioned optimization model.

Step (4): Combine the assessment alternatives of all attributes with the weights of experts and attributes by the information aggregation operators to generate an overall assessment of all alternatives.

Step (5): Compare the alternatives with each other by the operational rules.

Step (6): Obtain the ranking order of q alternatives to get the optimal alternative for this problem.

Step (7): Exit the assessment process.

The whole process is illustrated in details in Fig. 4.

The application of the assessment

In order to evaluate the reliability of the WSN in the bioinformatics, three experts are invited as the decision makers from the Department of Science and Technology. They are denoted by E_1, E_2 and E_3, and assigned the weights of (0.4, 0.3, 0.3), respectively.

Four types of WSNs in bioinformatics are selected as the four alternatives, which are denoted as Q_1, Q_2, Q_3 and Q_4. The attributes mentioned in Section 2 are expressed by Att_1, Att_2, Att_3, Att_4, and Att_5. Besides, the weights for these attributes are determined as $wet = \{0.28, 0.22, 0.13, 0.21, 0.17\}$ by the proposed method in Section 3. In total, the experts provide a set of eleven linguistic terms $L = \{l_0 = EB, l_1 = VB, l_2 = B, l_3 = LB, l_4 = M, l_5 = LG, l_6 = G, l_7 = VG, l_8 = EG\}$. Their opinions and preferences on the four alternatives on the five attributes are expressed by the 2-tuple fuzzy linguistic term set mentioned in the previous sections.
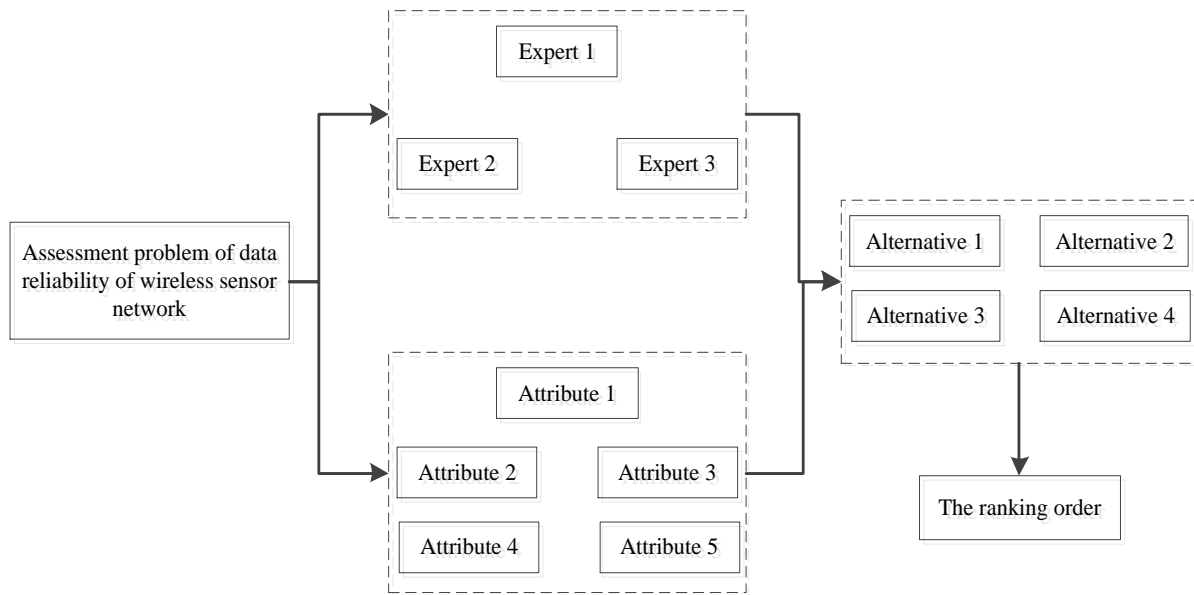
Fig. 4 The assessment process

Based on their experience and preferences, the experts evaluate each assessment alternative Q_j (j = 1, 2, 3, 4) on each attribute Att_i (i = 1, 2, …, 5) by the 2-tuple linguistic representation method. Next, the assessment matrix is constructed, (Tables 1-3) and the 2-tuple linguistic operators in Eq. (3) are introduced to aggregate the assessments on all attributes. The resulting ranking orders are $Q_2 > Q_4 > Q_3 > Q_1$.

Finally, Q_2 is proved to be the optimal solution. However, the assessment should not stop here. The reasons making it the optimal solution should be further analyzed based on specific attributes.

Table 1. The decision matrix of expert E_1

*0.01	Q_1	Q_2	Q_3	Q_4
Att_1	$(l_1, 20)$	$(l_5, 25)$	$(l_3, 40)$	$(l_5, -15)$
Att_2	$(l_7, 15)$	$(l_8, -30)$	$(l_4, 20)$	$(l_4, -10)$
Att_3	$(l_2, -50)$	$(l_6, -20)$	$(l_5, -30)$	$(l_5, -10)$
Att_4	$(l_2, 25)$	$(l_5, 20)$	$(l_2, 20)$	$(l_6, 30)$
Att_5	$(l_3, -15)$	$(l_4, 30)$	$(l_3, 10)$	$(l_3, 20)$

Table 2. The decision matrix of expert E_2

*0.01	Q_1	Q_2	Q_3	Q_4
Att_1	$(l_5, -15)$	$(l_4, 35)$	$(l_6, 30)$	$(l_4, 40)$
Att_2	$(l_4, 30)$	$(l_5, 20)$	$(l_2, 40)$	$(l_4, -10)$
Att_3	$(l_1, 40)$	$(l_7, -10)$	$(l_3, 20)$	$(l_5, -20)$
Att_4	$(l_3, 20)$	$(l_6, 0.3)$	$(l_3, 10)$	$(l_5, 30)$
Att_5	$(l_3, 30)$	$(l_6, 0.2)$	$(l_5, 20)$	$(l_6, 20)$

Table 3. The decision matrix of expert E_3

*0.01	Q_1	Q_2	Q_3	Q_4
Att_1	$(l_2, 15)$	$(l_3, 25)$	$(l_4, 30)$	$(l_5, 40)$
Att_2	$(l_3, 40)$	$(l_6, 20)$	$(l_5, 30)$	$(l_7, -10)$
Att_3	$(l_1, 20)$	$(l_7, 20)$	$(l_3, 30)$	$(l_4, 40)$
Att_4	$(l_3, 30)$	$(l_6, 30)$	$(l_2, 20)$	$(l_3, -20)$
Att_5	$(l_1, 20)$	$(l_5, 20)$	$(l_2, 10)$	$(l_1, 45)$

Conclusion

Bioinformatics is a fast-developing interdisciplinary science that combines biology, mathematics, physics, chemistry, computer science with information science. It is widely adopted in such fields as fermentation technology, environmental monitoring, food engineering, clinical medicine and military. Nevertheless, the data or information gathered from these fields may be not effective. To solve the problem, it is necessary to construct a WSN-based bioinformatics system because the WSN is a useful tool to collect such information. This gives rise to the importance of information reliability in the bioinformatics system. There are five influencing factors on the reliability, including quality, benefits, service, timeliness and stability. These factors should be evaluated with a method that can handle both qualitative and quantitative assessment information. A viable option for the evaluation is the fuzzy linguistic method, especially 2-tuple linguistic model. Therefore, this paper introduces the 2-tuple linguistic representation is introduced to help experts provide their opinions related to different WSN-based bioinformatics systems that involve multiple factors. In addition, the author comes up with a novel way to identify attribute weights and takes the approach to weigh the relative importance of different factors that can be considered as attributes in this assessment problem. In the end, the assessment process is applied to an illustrative example to verify its effect. The bioinformatics system proposed in this paper will be improved further in future research.

References

1. Hao L. J., Q. F. Liang (2000). The Development and Application of Bioinformatics, Journal of Shangdong Institute of Light Industry, 14(2), 37-41.
2. Herrera F, L. Martinez (2000). A 2-tuple Fuzzy Linguistic Representation Model for Computing with Words, IEEE Transactions on Fuzzy Systems, 8(6), 746-752.
3. Martinez L., F. Herrera (2012). An Overview on the 2-tuple Linguistic Model for Computing with Words in Decision Making: Extensions, Application and Challenges, Information Sciences, 207, 1-18.
4. Nailwai M., J. B. Chauhan (2017). Analysis of Consequences of Non-synonymous SNPs of USP9Y Gene in Human Using Bioinformatics Tools, Meta Gene, 12, 13-17.
5. Ouchitachen H., A. Hait, N. Idrissi (2017). Improved Multi-Objective Clustering Algorithm in Wireless Sensor Network, Egyptian Informatics Journal, 18(1), 45-54.
6. Qin J., J. Wang (2010). A Human Cut Microbial Gene Catalog Established by Deep Metagenomic Sequencing, Nature, 464, 59-65.
7. Rodriguez R. M., A. Labella, L. Martinez (2016). An Overview on Fuzzy Modelling of Complex Linguistic Preferences in Decision Making, International Journal of Computational Intelligence Systems, 26, 81-94.
8. Senturk I. F., P. Balakrishnan, A. Abu-Doleh, K. Kaya, Q. Malluhi, U. V. Catalyurek (2017). Are Source Provisioning Framework for Bioinformatics Applications in Multi-cloud Environments, Future Generation Computer Systems, in press.
9. Teng L. P. (2012). Overview of Security in Wireless Sensor Networks, Computer Security, 11, 24-29.

10. Wang C. D. (2015). A Modified Machine Learning Method Used in Protein Prediction in Bioinformatics, International Journal Bioautomation, 19(1), 25-36

11. Zhao P. N., W. B. Zhang, S. J. Chen (2010). Predicting Secondary Structural Folding Kinetics for Nucleic Acids, Biophysical Journal, 98, 1617-1625.

12. Zhao Y., R. S. Gu, S. M. Du (2012). The Research Status and Development Tendency of Bioinformatics, Journal of Medical Informatics, 33(5), 2-6.

13. Zhou Z. H. (2016). Total Variation-based Denoising Model for Bioinformatics Images, International Journal Bioautomation, 20(4), 457-470.

Ting Dong, M.Sc.
E-mail: 2367094502@qq.com

Ting Dong received her B.Sc. degree in Xi'an Institute of Posts and Telecommunications in 2004, and M.Sc. degree in Sichuan University in 2011. Her research interest is software engineering.