frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 13 January 2023
DOI 10.3389/fgene.2022.1086379

OPEN ACCESS

EDITED BY
Deepak Kumar Jain,
Chongqing University of Posts and Telecommunications, China

REVIEWED BY
Lei Shi,
Luliang University, China
Tiefeng Wu,
Qingdao University of Technology, China
Fenghui Dong,
Nanjing Forestry University, China

*CORRESPONDENCE
Xiaopei Cao,
✉ kana0720@zjnu.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 01 November 2022
ACCEPTED 09 December 2022
PUBLISHED 13 January 2023

CITATION
Zhang C and Cao X (2023), Biological gene extraction path based on knowledge graph and natural language processing.
Front. Genet. 13:1086379.
doi: 10.3389/fgene.2022.1086379

Biological gene extraction path based on knowledge graph and natural language processing

Canlin Zhang[1] and Xiaopei Cao[2]*

[1]Sorenson Communications, Salt Lake City, UT, United States, [2]College of Creative Culture and Communication, Zhejiang Normal University, Jinhua, Zhejiang, China

The continuous progress of society and the vigorous development of science and technology have brought people the dawn of maintaining health and preventing and controlling diseases. At the same time, with the update and iteration of bioinformatics technology, the current biological gene research has also undergone revolutionary changes. However, a long-standing problem in genetic research has always plagued researchers, that is, how to find the most needed sample genes from a large number of sample genes, so as to reduce unnecessary research and reduce research costs. By studying the extraction path of biological genes, it can help researchers to extract the most valuable research genes and avoid wasting time and energy. In order to solve the above problems, this paper used the Bhattacharyya distance index and the Gini index to screen the sample genes when extracting the characteristic genes of breast cancer. In the selected 49 public genes, 6 principal components were extracted by principal component analysis (PCA), and finally the experimental results were tested. It was found that when the optimal number of characteristic genes was selected as 5, the recognition rate of genes reached the highest 90.31%, which met the experimental requirements. In addition, the experiment also proved that the characteristic gene extraction method designed in this paper had a removal rate of 99.75% of redundant genes, which can greatly reduce the time and money cost of research.

KEYWORDS

biological gene extraction, knowledge graph, natural language processing, path research, biological gene

1 Introduction

In a rapidly developing world, various diseases are increasingly threatening people's healthy life. Many of these diseases can be resisted through physical defense, but many diseases come from human genes. How to study human-related diseases and find solutions for corresponding diseases is a difficult problem that cannot be ignored in current life sciences. Each genetic disease has its own characteristics, and the ever-changing gene expression is the key to preventing medical staff from curing such diseases. Coincidentally, how to extract the most needed genes when studying disease genes is also the key for researchers to solve the problem of genetic diseases. From the above description, the importance of biological gene extraction pathway research can be found. In order to solve this problem, this paper selects the theory-assisted research of knowledge graph and natural language processing.

This paper adopts the knowledge graph and natural language processing theory for the research on the extraction path of biological genes, the purpose of which is to apply the technology of knowledge extraction of knowledge graph to the extraction of biological genes. In this way, the extraction of genes in clinical medicine and biological research can be better achieved, thereby further promoting the development of life sciences. The Bhattacharyya distance index and Gini index used in this paper have a good screening effect on massive research samples, and the PCA method can also further realize the extraction of characteristic genes. The innovation of this paper includes the following aspects: 1) It is not limited to the activity research of specific genes, but seeks to reduce the research cost and optimize the research process of similar research from the methodological level. 2) The combination of knowledge graph and gene extraction path is realized, which provides a more mature theoretical basis for optimizing the gene extraction path.

2 Literature review

It is not a novel thing to study biological genes to help people better understand and use biological genes. Xu L B discovered in his research on biological genes that biological rhythms are an important mechanism for organisms to receive external signals and regulate their own behavior. To this end, he summarized some core biological rhythms including biological clocks to assist biological breeding and human disease prevention and control (Xu et al., 2020). Ebigwai J K found that most of the existing protein research tools are only useful for the study of protein interaction PPI, so he proposed a new comprehensive PPI information extraction tool for support vector machine classifier, which is very important for protein gene research (Ebigwai et al., 2020). In order to better identify and predict the encoding of the genome, Abbas B proposed a new tool to rank, compare and identify the recurring properties in the hidden Markov model (HMM) (Abbas et al., 2019). Hasan M was opposed to synthetic gene editing and transgenic technologies that are widely used in food research and development. He hoped that the relevant departments can formulate more perfect biosafety regulations to reduce the risks related to food safety (Hasan et al., 2017). Do H pointed out that the mining technology of association rules can promote the study of different gene expression. To this end, she designed CPMiner, a data mining method for processing biological data, and dedicated to a unified framework for extracting gene expression using association rules (Do et al., 2020). Although these studies have carried out research on biological genes to a greater or lesser extent, they are all limited to a specific type or

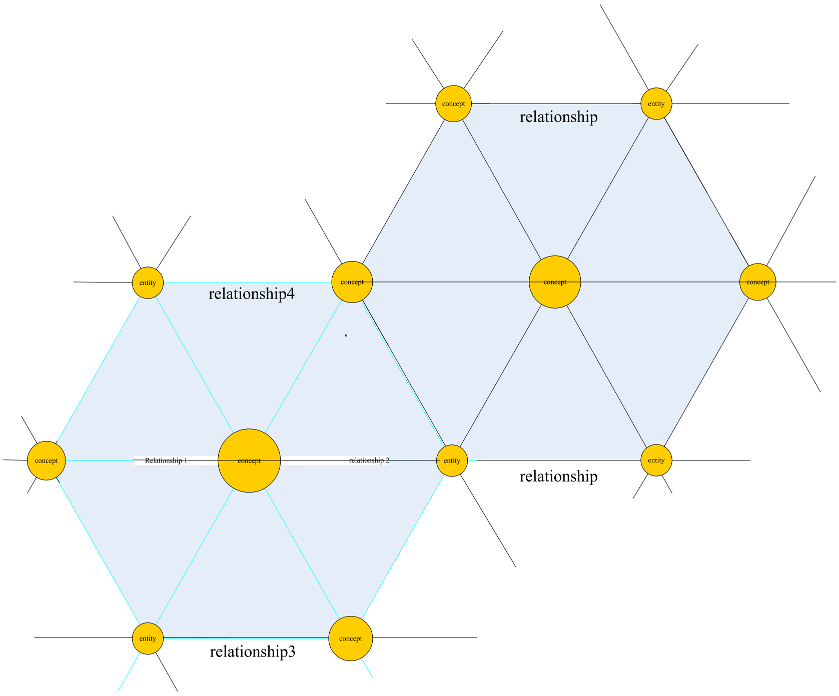

FIGURE 1
Knowledge graph network.

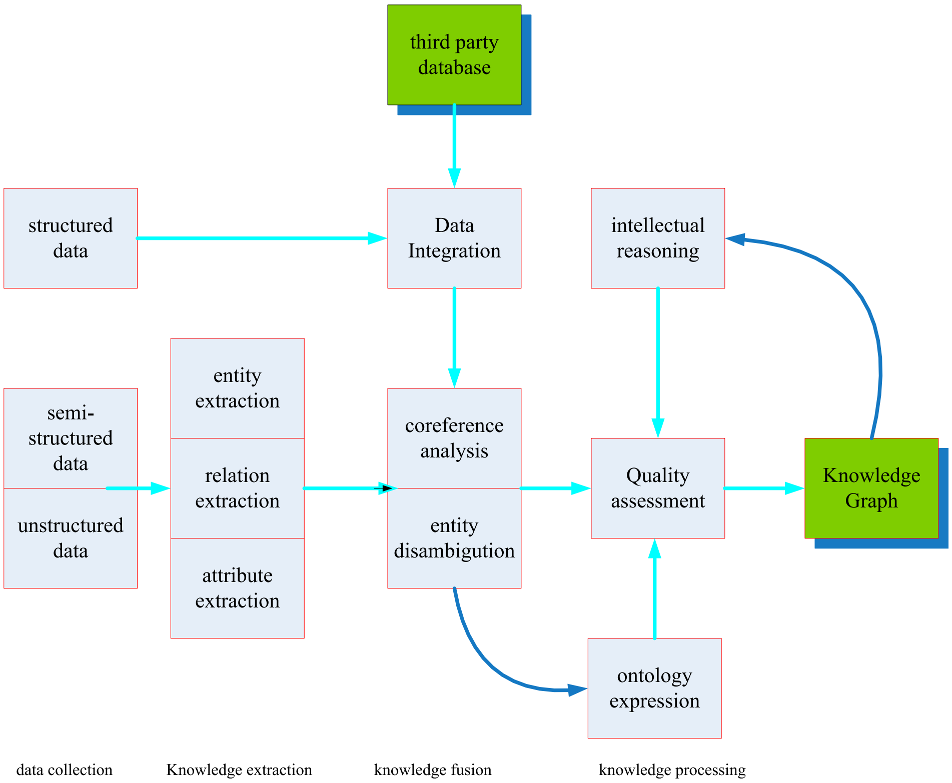

FIGURE 2
Construction process of the knowledge graph.

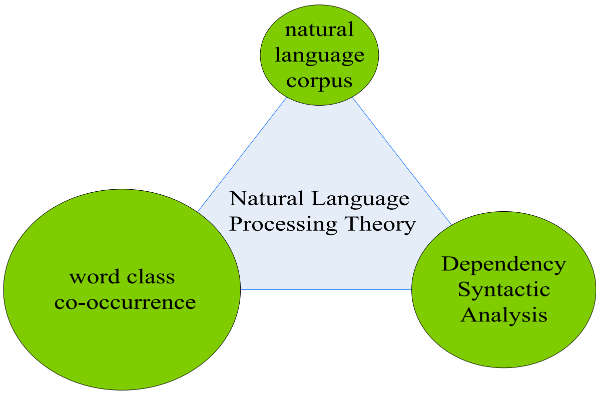
natural language corpus
Natural Language Processing Theory
word class co-occurrence
Dependency Syntactic Analysis

FIGURE 3
Natural language processing theory.

even a specific gene. However, this paper takes a unique approach to study the extraction path of biological genes, and provides a new research direction for the current gene research.

In recent years, with the spread of the concept of knowledge graph, more and more people have begun to conduct in-depth research on knowledge graph. Am A pointed out that the types of knowledge graphs are very broad, including public and private types. Knowledge graphs can also be obtained through the Internet, which requires statistical and linguistic methods (AlMarshad et al., 2021). Cai J L pointed out that a new multi-layer convolutional network model can be used in the model construction and connection prediction of large-scale knowledge graphs, which can build highly complex knowledge graphs based on its high parameter efficiency (Cai et al., 2019). Nuaima R H applied a new learning framework in the study of reasoning large-scale knowledge graphs (KG), which can construct knowledge graphs with continuous states in the KG vector space more accurately, diversely and efficiently (Nuaima et al., 2018). Shi J found that existing self-symptom detectors are far from meeting the needs of clinical decision support systems. Therefore, she studied an automated knowledge graph that can automatically learn diseases and conditions in electronic medical records to identify patients' diseases (Shi et al., 2017). Lin Y Z found that the time limit is an invisible challenge to the mining and exploration of large-scale knowledge graphs. To this end, she proposed an online mining algorithm, which can greatly improve the speed and accuracy of mining knowledge graphs within a certain period of time (Lin et al., 2017a). Such a comprehensive

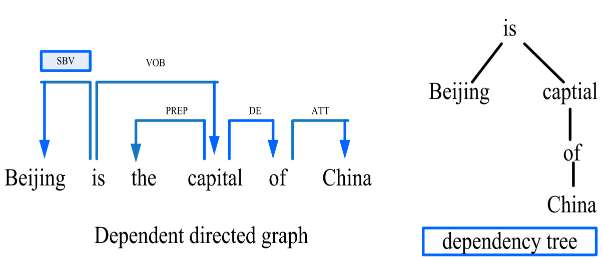

FIGURE 4
Process of dependency parsing.

knowledge map research provides a pioneering idea for the combination of knowledge map and biological gene extraction path in this paper. This paper looks forward to solving the problem of biological gene extraction path that has puzzled researchers for a long time by using the related theory of exponential graph and natural language processing.

3 Methods of biological gene extraction research based on knowledge graph and natural language processing

3.1 Knowledge graph and natural language processing

3.1.1 Knowledge graph

Knowledge graph is a semantic network proposed by Google in 2012 to describe the relationship between concepts in the real world in order to improve the search ability of its own engine (Paulheim and Cimiano, 2017). In fact, the essence of the knowledge graph is a conceptual network woven by related relationships between physical machines in the real world. As shown in Figure 1, on this network, nodes represent real objects or concepts, and the lines connecting each node are network edges and represent the correlation between them.

By combining the knowledge network with the Internet, it is possible to establish the relationship between things on the Internet with the help of the knowledge graph, and on this basis, to combine the all-encompassing information on the Internet to create new knowledge, which is a major breakthrough in artificial intelligence. The knowledge graph has been inseparable from artificial intelligence since it was proposed, and is widely used in search engines, intelligent question answering, and personalized recommendation (Lin et al., 2017b). The combination of the two can reduce the tediousness of manual search, and use Internet tools to visualize the searched knowledge.

The construction process of the knowledge graph is based on the original data and use technical means to extract objective knowledge facts from the original data, then extracting knowledge elements from the knowledge facts and storing knowledge elements in a database. The construction of the knowledge graph hides the update of the knowledge graph (Zhang et al., 2017). Because the real world is constantly moving and changing, and new concepts are produced every moment, along with the demise of concepts, the construction of knowledge on this basis also means the renewal of knowledge. As shown in Figure 2, the construction of knowledge graph includes three stages, namely knowledge extraction, knowledge fusion and knowledge processing. Each knowledge construction is an iterative process of updating.

Among them, information extraction is the starting link of knowledge graph construction (Natthawut and Ryutaro, 2018). This link is mainly the process of extracting entities and the interrelationships of entity attributes from real entities. Through this link, a large number of discrete entities and their interrelationships can be achieved. Then, through the knowledge fusion link, the data obtained in the information extraction link is cleaned and classified. The redundant and repeated data are eliminated, and errors are eliminated to improve the hierarchy and logic of knowledge. Actually this coincides with the research on the biological gene extraction pathway studied in this paper. The research topic of this paper is to extract the core genes from the complex genome. The third link is the knowledge processing link. In this link, knowledge needs to be processed to obtain systematic and structured knowledge. In addition, a dynamic knowledge network needs to be finally obtained by constructing ontology, inferring knowledge and evaluating quality.

3.1.2 Natural language processing

Natural language processing (NLP) is an important research tool for computers, especially in artificial intelligence. It mainly studies the means of communication between humans and computers, that is, natural language (Jia et al., 2017). In the information age, people can access information on the Internet more quickly because of natural language processing, and can also respond more quickly to changes in online life. The most typical example is the interception of e-mail spam. With the help of natural language processing, people can easily set up blocking words, so that the mailbox can block spam by locking key words to avoid harassment. The processing of natural language is inseparable from the natural language corpus. Natural language corpus is an indispensable foundation for processing natural language. Meanwhile, with the increasing development of natural language processing related work, the theory of natural language processing has gradually matured. As shown in Figure 3, the processing theory of natural language includes three parts: natural language corpus, dependency syntax analysis and part-of-speech co-occurrence (Zhu et al., 2017).

TABLE 1 Sample gene datasets.

Datasets	Category	Sample amount	Gene amount
Leukemia data set	ALLI.ALLB.AWL	46	6844
Polio data set	4	22	3629
Breast cancer data set	3	50	2000
Colon cancer data set	ALL.AML	82	96548

TABLE 2 Basic description of sample data.

Breast cancer sample data division	Lesion sample	Normal sample
Sample size	28	22
Training set	20	10
Test set	8	12

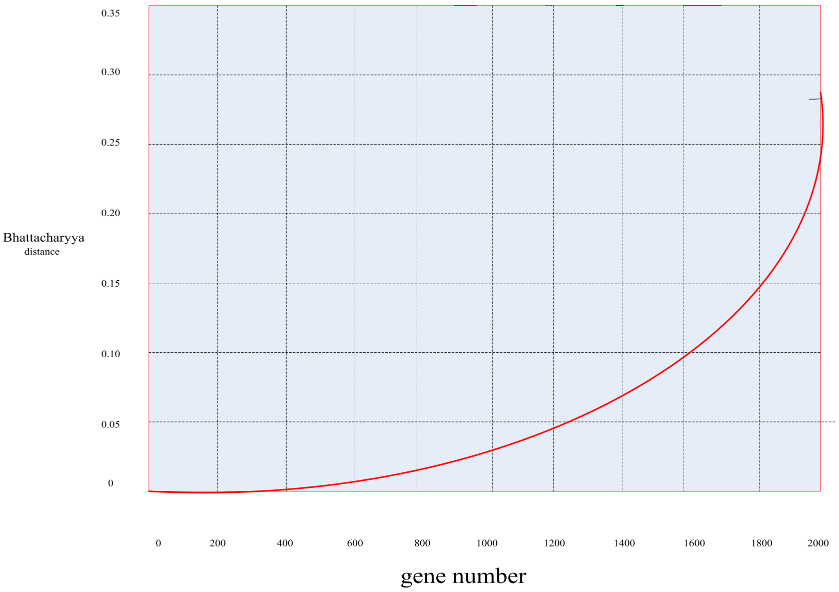

FIGURE 5
Bhattacharyya distance distribution curve of breast cancer genes.

The study of natural language corpus comes from the interdisciplinary subject of corpus linguistics. Corpus linguistics is a hybrid science that combines linguistics, computer science and applied linguistics (Wang et al., 2018). Through the study of a large number of real language materials, this discipline summarizes some abstract language laws that are detached from specific languages and words, and applies these language laws to natural language processing, which plays an important role in improving the language learning ability of the machine. Corpus linguistics has been around for a hundred years since its inception, and so far corpora are still an integral part of natural language processing.

Dependency parsing is an important part of natural language processing. Its working principle is mainly to reveal its intrinsic syntactic structure by analyzing the dependencies between various language components, and to assist the research by constructing a corresponding dependency tree. The workflow of dependency parsing is based on the existence of dependencies

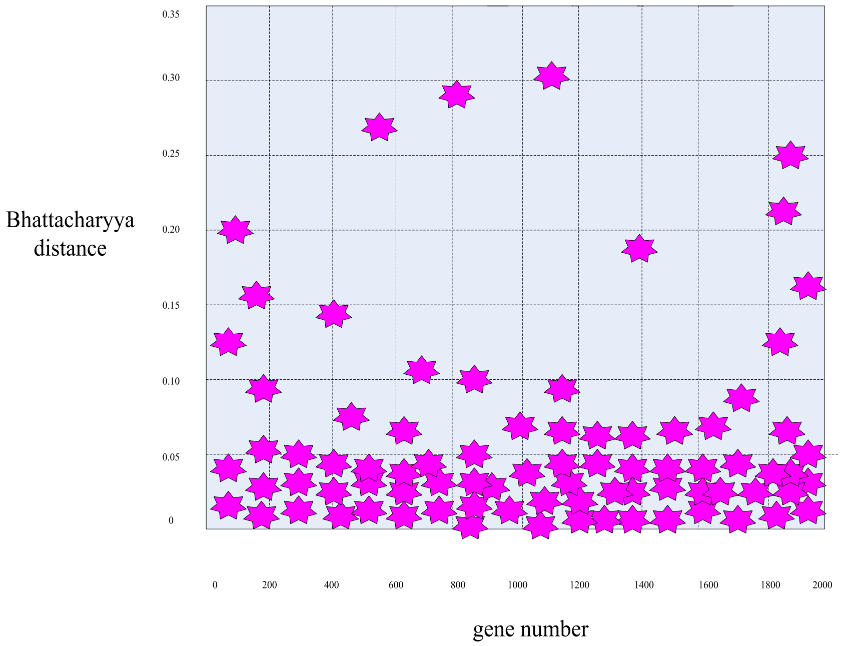

FIGURE 6
Scatter plot of the distribution of Bhattacharyya distance for breast cancer genes.

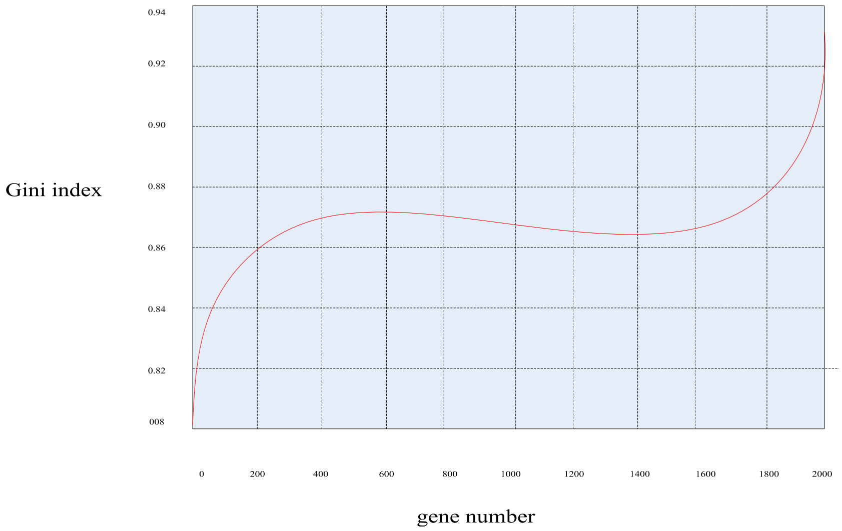

FIGURE 7
Gini index distribution curve of breast cancer genes.

between different natural language units. Generally speaking, a dependency relationship consists of two parts, the core word and the modifier. Researchers can analyze the structure of sentences through dependency syntax analysis technology to determine the main components of sentences, so as to help machines better understand natural language and achieve more efficient human-computer interaction. As shown in Figure 4, the analysis process of dependency parsing is as follows:

Co-occurrence theory, literally, is a theory that studies the relationship of things that occur together (Tom et al., 2018).

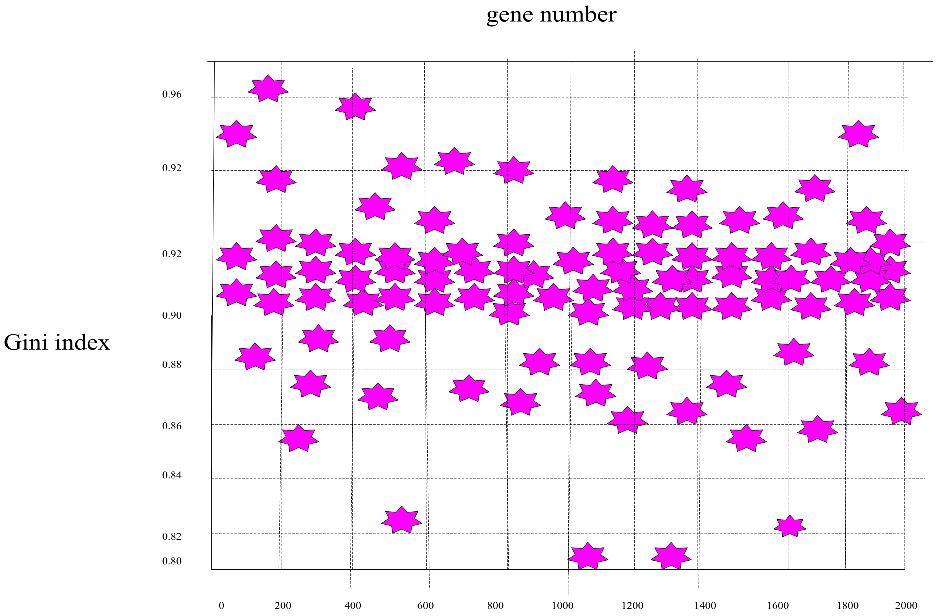

FIGURE 8
Scatter plot of the distribution of Gini index distribution of breast cancer genes.

Generally speaking, the co-occurrence relationship is the most basic relationship in the entity relationship. In fact, the co-occurrence relationship refers to inferring the affinity between different entities by counting the number of co-occurrences between them in the same text, and this relationship is also called co-occurrence relationship in the field of natural language. Co-occurrence relationship can be divided into broad and narrow sense. Different from the limited scope of the co-occurrence relationship in the narrow sense, the generalized contribution relationship expands the scope of this relationship to time and space, and when analyzing this generalized co-occurrence relationship, it can further increase the prediction of the real world. For example, Walmart found that sales of egg tarts and flashlights can increase during the hurricane season, and the increase in sales and the arrival of a hurricane are co-occurrences in a broad sense. Based on this broad co-occurrence, Walmart would put together egg tarts and anti-hurricane supplies before the hurricane to increase sales.

3.2 Feature gene extraction based on knowledge graph and natural language processing

3.2 1 Feature gene extraction

Feature gene extraction refers to finding out the genes that are not relevant for classification, extracting those representative genes. Then these genes are grouped into a subset for classification (Wong et al., 2018). There are two feature selection methods, namely feature selection and feature extraction. In a broad sense, feature extraction refers to a kind of mapping. The main methods are principal component method and partial least squares method. Feature selection only removes redundant genes and noise genes. Feature selection is difficult to achieve the optimal classification effect because the misclassification rate is too large, which is not considered in this paper.

3.2 2 Scoring guidelines for gene expression data

In view of the high-dimensional and high-noise characteristics of gene expression, this paper believes that dimensionality reduction and denoising of gene expression are very necessary. Scoring guidelines are integral to this process. This paper mainly uses feature score criterion (FSC) and Fisher discriminant method. The scoring formulas are shown in Formulas 1, 2:

$$FSC(g_i) = \left|\frac{\eta_i^+ - \eta_i^-}{\omega_i^+ - \omega_i^-}\right| \tag{1}$$

$$FDR(g_i) = \frac{(\eta_i^+ - \eta_i^-)^2}{(\omega_i^+)^2 - (\omega_i^-)^2} \tag{2}$$

In Formula 1 and Formula 2, η_i^+ and η_i^- represent the mean, while ω_i^+ and ω_i^- represent the standard deviation.

TABLE 3 List of principal component characteristic data of public genes.

Main ingredient	Eigenvalues	Contribution rate (%)	Feature contribution rate (%)
A	28.174	39.658	39.658
B	10.742	18.649	49.165
C	9.645	12.359	57.761
D	8.742	8.542	61.952
E	4.568	3.406	71.962
F	2.691	2.864	77.294
G	1.669	2.173	79.429
H	1.485	1.749	81.263
I	1.293	1.599	83.648

3.2.3 Feature gene extraction method

The purpose of eigengene extraction is to extract the most eigengenes from a group of genes, and the number of a group of eigengenes is extremely large in the experiment. Therefore, the selection of the extraction method of eigengenes is very important. The commonly used eigengene selection methods are as follows:

Filtering method: The filtering method refers to the selection according to the data of the information itself. This paper mainly uses the following two filtering methods, namely the signal-to-noise ratio and the t-statistic method. The signal-to-noise ratio is the simplest filtering method. The criterion for selecting feature vectors by using the signal-to-noise ratio method is the degree of difference in the vector attribute space in different categories of samples. Assuming that there is a gene g), the calculation formula of its signal-to-noise ratio $d(g)$ is as Formula 3:

$$d(g) = \left| \frac{\eta_g^+ - \eta_g^-}{\omega_g^+ - \omega_g^-} \right| \tag{3}$$

In Formula 3, η_g^+ and η_g^- represent the mean, while ω_g^+ and ω_g^- represent the standard deviation.

The t-statistic method refers to the use of the t-part theory to infer the probability of the occurrence of differences, and its calculation formula is shown in Formula 4:

$$t_i = \frac{\bar{x}_i - \overline{y_i}}{\sqrt{\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}}} \tag{4}$$

In Formula 4, $\bar{x}_i$ and $\overline{y_i}$ represent the mean, and s_{1i}^2 and s_{2i}^2 represent the variance. It can be seen that the larger the calculated t value, the greater the difference, that is, the greater the influence of gene i on the sample classification.

3.3 Path of feature gene extraction based on PCA

Gene expression data are characterized by small samples, high dimensionality, and high noise, and there are very few genes associated with diseases. The presence of a large number of redundant genes makes the expression difference between the diseased and normal samples small, and the contribution to the classification is also small. Therefore, it is necessary to use a certain gene microarray technology to propose genes from a large number of genes, and extract biological genes that truly meet the characteristics (Balsmeieri et al., 2018).

3.3.1 Construction of gene expression data

Gene data covers the basic characteristics of genes and is a display of the physiological state of biological cells. The gene expression profile can be obtained by using gene chip technology. The expression profile data can be expressed as Formula 5:

$$X = \left(x_{ij}\right) = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} \\ x_{21} & x_{22} & \dots & x_{2j} \\ \vdots & \vdots & \dots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} \end{bmatrix} \tag{5}$$

In Formula 5, x_{ij} represents the j-th sample for the i-th gene.

3.3.2 Elimination of irrelevant genes

There are tens of thousands of genes in a biological sample, and at least a few are needed for the experiment. Other unrelated genes can cause gene redundancy and interfere with the experiment. Therefore, this paper chooses a hybrid gene extraction method that combines the Bhattacharyya distance and the Gini index. The formula for calculating the Bhattacharyya distance of a gene is shown in Formula 6:

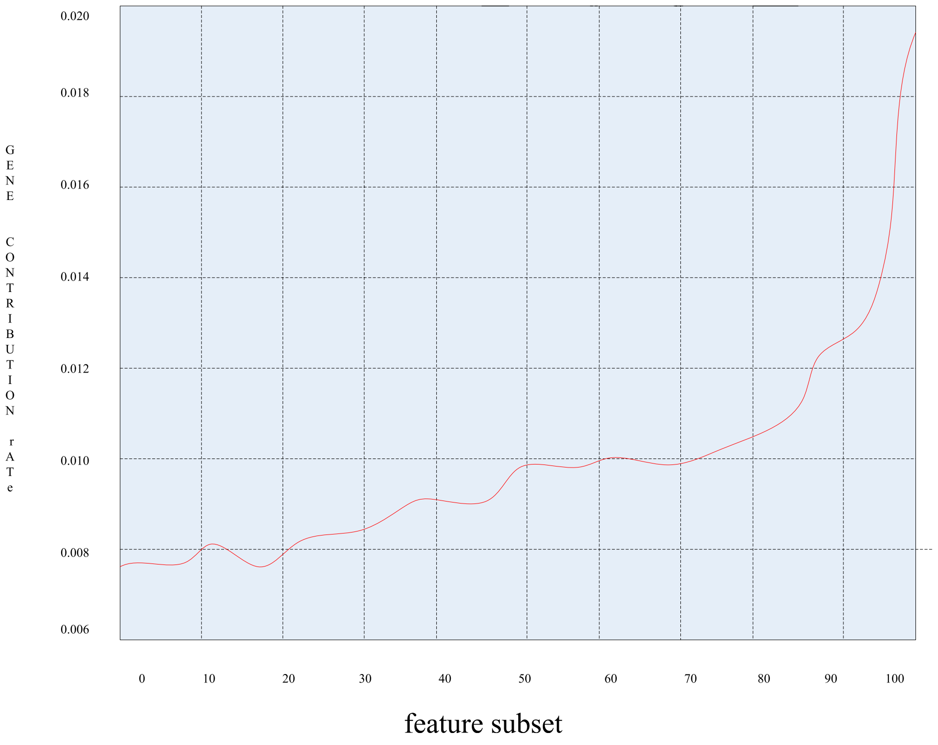

FIGURE 9
Graph of the contribution rate of each principal component.

$$B_d = \frac{1}{8}\frac{(\eta_1 - \eta_2)^2}{(\omega_1^2 + \omega_2^2)} + \frac{1}{2}\ln\left(\frac{\omega_1^2 + \omega_2^2}{2\omega_1\omega_2}\right) \quad (6)$$

Among them, η_1 and η_2 represent the mean of genes in the first and second samples respectively, while ω_1 and ω_2 represent the sample variance.

Gini index is an index commonly used in data mining to evaluate the goodness of classification nodes (Wi et al., 2017). Before calculating the Gini index, the original data should be discretized. The discrete formula is shown in Formula 7:

$$S_{ij} = Int\left[20 \times \frac{n_{ij} - \min(i)}{\max(i) - \min(i)} + 0.5\right] \quad (7)$$

Among them, $Int[\,]$ represents rounding, and $\max(i)$ and $\min(i)$ represent the maximum and minimum values of gene g_i.

The discretized data is substituted into Formula 8:

$$Gini(k) = 1 - \sum_{j=0}^{20}\left[p_{ij}\right]^2 \quad (8)$$

In Formula 8, $Gini(k)$ refers to the Gini index of k-type genes, and p_{ij} refers to the relative rate of level j in category k.

The formula for calculating the Gini index of gene g_i is as Formula 9:

$$Gini(g_i) = \sum_{k=1}^{2}\frac{n_k}{n}Gini(k), i \in [1, 2000] \quad (9)$$

In Formula 9, n is the total number of samples and n_k is the number of samples in the k-th class. Therefore, the smaller the Gini value of gene g_i, the more taxonomic information this gene contains, and *vice versa*.

The Bhattacharyya distance and the Gini index are combined together, and these two indicators are integrated as a standard to measure the gene classification information, which avoids the defect of “signal-to-noise ratio” and other methods that do not generate variance in the differential expression of genes (Diamantopoulos et al., 2017).

3.3.3 Selection of principal component feature

PCA refers to using the idea of dimensionality reduction to replace many original variables with new variables on the basis of keeping the principal components unchanged. These new variables can reflect most of the information of the original variables (Johnny et al., 2017; UzmaAl-Obeidat et al., 2022). The purpose of principal component analysis is to extract as few comprehensive variables with strong classification ability as possible as the representative of all variables, so as to retain the original variable information and reduce the number of

TABLE 4 Detection results of signature genes.

Number of genes	Classification accuracy (%)	Number of misjudgments
6	78.82	7
10	84.62	5
5	90.31	4
16	67.74	5
20	88.52	4

variables. Assuming that there are n samples, and each sample contains m variables, then there is Formula 10:

$$X_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix} \tag{10}$$

On this basis, the data is standardized to obtain Formula 11, Formula 12, Formula 13:

$$x_{ij} = \frac{x_{ij} - \overline{x_j}}{h_j} \tag{11}$$

$$\overline{x_j} = \frac{1}{n}\sum_{n=1}^{1} x_{ij} \tag{12}$$

$$h_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2} \tag{13}$$

$\overline{x_j}$ means the mean of the j-th attribute, and h_j means that the variance is normalized to have a mean of 0 and a standard deviation of 1 for each data.

Then the correlation coefficient matrix is obtained as Formula 14:

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1t} \\ P_{21} & P_{22} & \ldots & P_{2t} \\ \vdots & \vdots & \vdots & \vdots \\ P_{m1} & P_{m2} & \ldots & P_{mt} \end{bmatrix} \tag{14}$$

The correlation coefficient p_{ij} represents the correlation between x_i and x_j, and the calculation formula is as Formula 15:

$$p_{ij}\frac{\sum_{k=1}^{n}\left(x_{ki} - \overline{x_i}\right)\left(x_{kj} - \overline{x_j}\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ki} - \overline{x_i}\right)^2 \sum_{k=1}^{n}\left(x_{kj} - \overline{x_j}\right)^2}}, \quad i, j = 1, 2, ..., m \tag{15}$$

Solving the characteristic formula according to the correlation coefficient matrix P, there is Formula 16:

$$\left|\gamma I - P\right| = 0 \tag{16}$$

The characteristic formula is solved to get different eigenvalues. Sorting these eigenvalues yields Formula 17:

$$\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_m \geq 0 \tag{17}$$

Then the standard orthonormal vector is calculated by the existing eigenvalues to obtain Formula 18:

$$e_i = \left(e_{1i}, e_{2i}, .., e_{mi}\right)^T \tag{18}$$

On this basis, the analysis of different principal component contribution rates obtains Formula 19:

$$\lambda_i = \frac{\gamma_i}{\sum_{k=1}^{m}\gamma_k}, i \tag{19}$$

After calculating the contribution rate of each principal component, the cumulative sum is obtained to obtain the cumulative contribution rate as shown in Formula 20:

$$\lambda_{(i)} = \frac{\sum_{k=1}^{i}\gamma_k}{\sum_{k=1}^{m}\gamma_k}, \quad i = 1, 2, ..., m \tag{20}$$

Finally, the gene contribution rate Formula 21 can be found:

$$g_j = \frac{\sum_{i=1}^{t}\left|d_{ij}\right|}{\sum_{i=1}^{t}\sum_{k=1}^{m}\left|d_{ik}\right|}, \quad j = 1, 2, .., m \tag{21}$$

4 Experiment and results of feature gene extraction based on PCA

4.1 Experimental data acquisition and processing

With the fast-paced work and life, the living pressure of modern women has greatly increased. Staying up late and insomnia are quite common. In such a stressful life rhythm, breast cancer, the killer disease of women's health, has quietly appeared in more and more women. In order to alleviate this situation, this paper took breast cancer as the research object, and analyzed the characteristic gene extraction pathway of breast cancer, hoping to enrich the breast cancer gene extraction pathway and contribute to the control and prevention of breast cancer.

In order to verify the feature gene extraction path proposed in this paper, several common disease gene datasets were collected in this paper, as shown in Table 1:

As shown in Table 1, the total number of samples in the selected four diseases was 200 cases. It can be seen that the number of samples of breast cancer genes was large, but the total number of genes was relatively small. The former satisfied the quantitative requirements of the experiment, and the latter represented the less complex number of genes, which was convenient for screening. Even though all four of these diseases were common, the choice of breast cancer genes was more beneficial than other options.

Through the investigation, it can be seen that among the 50 samples in this breast cancer data set, 28 were diseased genes, and the remaining 22 were normal samples. In this regard, 30 samples were used as the training set, and the remaining 20 samples were used as the test set. The description of the samples is shown in Table 2:

4.2 Experimental process and results

4.2.1 Elimination of irrelevant genes

Based on the characteristics of sample genes with small number of samples and many redundant genes, although there were 2000 genes in 20 samples, there were still a large number of genes that were not related to the content of the experiment. Therefore, this paper combined the two methods of Bhattacharyya distance and Gini index to measure to promote the standardization of gene classification. This method avoided the defect that methods such as "signal-to-noise ratio" did not express differential gene expression due to variance. On this basis, this paper focused on the former and the latter as a supplement, so that the priority and secondary could better extract public genes.

First, according to the distance calculation formulas mentioned above, the Bhattacharyya distance value of these 2000 sample genes was calculated to obtain the distribution curve diagram of the Bhattacharyya distance of these genes in Figure 5 and the distribution scatter diagram in Figure 6.

First, it can be seen from Figures 5, 6 that in general, only a small number of genes have high information content, while most of the genes have a Bhattacharyya distance index between 0 and 0.05. This indicated that the genes within this range had no significant difference in mean and variance between normal samples and cancer samples, so it can be easily deleted as an irrelevant gene. Considering the validity and accuracy of the experiment, in this experiment, the Bhattacharyya distance index of these 2000 genes was sorted from small to large, and the corresponding genes ranked in the bottom 200 were selected as further experimental objects, eliminating the other 1800 irrelevant genes.

Similarly, according to the calculation formula of the Gini index mentioned above, the Gini index values of these 2000 genes were calculated, and the Gini index distribution curve diagram of the sample genes can be obtained in Figure 7 and the distribution scatter diagram in Figure 8.

It can be seen from Figures 7, 8 that the Gini index values of these 2000 sample genes were mostly between 0.88 and 0.92. In fact, the smaller the Gini index of a gene, the more equal the data expression, and the greater the amount of classification information of the gene. Since the experimental design of this paper was mainly based on the Bhattacharyya distance ranking, supplemented by the Gini index ranking, the same sorting method was used to eliminate irrelevant genes, and 200 spare genes were extracted.

After selecting 200 genes for each of the Bhattacharyya distance index and the Gini index as experimental subjects, among these 400 genes, the common genes were found whose duplicates were regarded as both. It could be determined that the amount of taxonomic information in these public genes far exceeded that of the excluded irrelevant genes. In the end, 49 common genes were obtained.

4.2.2 Extraction of characteristic genes

The dimensionality reduction idea of PCA was used to select the features of the 49 common genes finally selected, and the feature genes could be selected by the contribution rate of each gene in the finally selected principal components.

The eigenvalue, contribution rate and feature contribution rate of each principal component obtained by PCA of the 49 public genes are shown in Table 3:

It can be seen from Table 3 that the eigenvalues of the first six principal components were all greater than 2, and the eigenvalues corresponding to the seventh principal component were all less than 2. The cumulative contribution rate of the first six principal components has reached 85.478%. Therefore, only the first six principal components could be further analyzed.

When further analyzing the first six principal components, the contribution rate of each principal component has been calculated, and Figure 9 was obtained by arranging the contribution rate of each principal component in ascending order:

It can be seen from Figure 9 that the contribution rate of these 49 genes varied greatly, and the gene contribution rate varied greatly between 0.010 and 0.018. That is to say, those genes with a large gene contribution rate contain more classification information of the sample and contribute more to the classification. It can be seen that these genes contain a great amount of information, and the corresponding sample genes can be extracted as characteristic genes.

4.2.3 Classification inspection

On the basis of extracting the eigengenes above, this paper used the support vector machine (SVM) radial basis kernel function to detect the extracted eigengenes. The results are shown in Table 4:

It can be seen from Table 4 that when the number of selected eigengenes was 5, the recognition rate of genes reached a maximum of 90.31%. Therefore, it can be determined that only 5 genes in the 2000 sample genes finally contained rich information and meet the experimental requirements. The 5 most useful genes were extracted from 2000 genes, and the rate of removing redundant genes reached 99.75%, which just proved that the characteristic gene extraction path designed in this paper has a good classification effect on breast cancer genes.

5 Conclusion

This paper studied the extraction path of biological genes related to knowledge graph and natural language processing. The main significance of this research was to provide more abundant channels for gene extraction in modern medicine. In the past biological gene extraction research, most people were limited to the activity of the extracted genes, but ignored how to accurately extract the characteristic genes from the source to better meet the experimental needs, which is also the starting point of this study. In this regard, the sample genes of the research object were standardized by two screening criteria, the Bhattacharyya distance index and the Gini index. Then the PCA method was used to further extract the eigengenes, and finally the eigengenes were verified. It was proved that the redundant gene removal rate of the characteristic gene extraction method in this paper reached 99.75%, which provides a certain reference value for modern biological research.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CZ: write paper XC: revise paper.

Conflict of interest

CZ was empolyed by Sorenson Communications.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbas, B., Tjolli, I., Dailami, M., and Munarti (2019). Phylogenetic of sago palm (Metroxylon sagu) and others monocotyledon based on mitochondrial nad2 gene markers. *Biodiversitas J. Biol. Divers.* 20 (8), 2249–2256. doi:10.13057/biodiv/d200820

AlMarshad, L. K., AlJobair, A. M., Al-Anazi, M. R., Bohol, M. F. F., Wyne, A. H., and Al-Qahtani, A. A. (2021). Association of polymorphisms in genes involved in enamel formation, taste preference and immune response with early childhood caries in Saudi pre-school children. *Saudi J. Biol. Sci.* 28 (4), 2388–2395. doi:10.1016/j.sjbs.2021.01.036

Balsmeieri, B., Assaf, M., Chesebro, T., Fierro, G., Johnson, K., Johnson, S., et al. (2018). Machine learning and natural language processing on the patent corpus: Data, tools, and new measures. *J. Econ. Manag. Strategy* 27 (3), 535–553. doi:10.1111/jems.12259

Cai, J. L., Yan, Y. F., Feng, G. W., and Jing, S. (2019). Dynamic change in the gene expression profile of rat benign prostate hyperplasia tissue after complete denervation. *Zhonghua nan ke xue = Natl. J. Androl.* 25 (11), 971–977.

Diamantopoulos, T., Roth, M., Symeonidis, A., and Klein, E. (2017). Software requirements as an application domain for natural language processing. *Lang. Resour. Eval.* 51 (2), 495–524. doi:10.1007/s10579-017-9381-z

Do, H., Thu, T., and Tran, N. (2020). Indigenous Lien Minh chicken of Vietnam: Phenotypic characteristics and single nucleotide polymorphisms of GH, IGFBP and PIT candidate genes related to growth traits. *Biodiversitas J. Biol. Divers.* 21 (11), 5344–5352.

Ebigwai, J. K., Ferdinand, A., and Ubi, G. M. (2020). Resolving taxonomic ambiguity between two morphological similar plant taxa using maturase K gene analysis. *J. Biol. Sci.* 20 (1), 13–21. doi:10.3923/jbs.2020.13.21

Hasan, M., Siddique, M. A., Hossain, M. A., and Rahman, M. S. (2017). 16S rRNA gene sequence based identification of Vibrio spp. in shrimp and tilapia hatcheries of Bangladesh. *Dhaka Univ. J. Biol. Sci.* 26 (1), 45–58. doi:10.3329/dujbs.v26i1.46349

Jia, Y., Wang, Y., Jin, X., Lin, H., and Cheng, X. (2017). Knowledge graph embedding: A locally and temporally adaptive translation-based approach. *ACM Trans. Web* 12 (2), 1–33. doi:10.1145/3132733

Johnny, D., Velupillai, S., George, G., Holden, R., Kikoler, M., Dean, H., et al. (2017). Detection of suicidality in adolescents with autism spectrum disorders: Developing a Natural Language Processing approach for use in electronic health records. *AMIA Symp.* 2017, 641–649.

Lin, Y. Z., Ou, D. L., Chang, H. Y., Lin, W. Y., Hsu, C., and Chang, P. L. (2017). Simultaneous visualization of the subfemtomolar expression of microRNA and microRNA target gene using HILO microscopy. *Chem. Sci.* 8 (9), 6670–6678. doi:10.1039/c7sc02701j

Lin, Z. Q., Xie, B., Zou, Y. Z., Zhao, J. F., Li, X. D., Wei, J., et al. (2017). Intelligent development environment and software knowledge graph. *J. Comput. Sci. Technol.* 32 (002), 242–249. doi:10.1007/s11390-017-1718-y

Natthawut, K., and Ryutaro, I. (2018). An automatic knowledge graph creation framework from Natural Language text. *Ieice Trans. Inf. Syst.* 101 (1), 90–98. doi:10.1587/transinf.2017swp0006

Nuaima, R. H., Roeb, J., Hallmann, J., Daub, M., Otte, S., and Heuer, H. (2018). Effector gene vap1 based DGGE fingerprinting to assess variation within and among *Heterodera schachtii* populations. *J. nematology* 50 (4), 517–528. doi:10.21307/jofnem-2018-055

Paulheim, H., and Cimiano, P. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web* 8 (3), 489–508. doi:10.3233/sw-160218

Shi, J., Li, W., Gao, Y., Wang, B., Li, Y., and Song, Z. (2017). Enhanced rutin accumulation in tobacco leaves by overexpressing the NtFLS2 gene. *Bioence Biotechnol. Biochem.* 81 (9), 1721–1725. doi:10.1080/09168451.2017.1353401

Tom, Y., Devamanyu, H., Soujanya, P., and Cambria, E. (2018). Recent trends in deep learning based Natural Language Processing. *IEEE Comput. Intell. Mag.* 13 (3), 55–75. doi:10.1109/mci.2018.2840738

UzmaAl-Obeidat, F., Tubaishat, A., Shah, B., and Halim, Z. (2022). Gene encoder: A feature selection technique through unsupervised deep learning-based clustering for large gene expression data. *Neural Comput. Applic* 34, 8309–8331. doi:10.1007/s00521-020-05101-4

Wang, C., Ma, X., and Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature. *Comput. Geosciences* 112, 112–120. doi:10.1016/j.cageo.2017.12.007

Wi, C. L., Sohn, S., Rolfes, M. C., Seabright, A., Ryu, E., Voge, G., et al. (2017). Application of a Natural Language Processing algorithm to asthma ascertainment: An automated chart review. *Am. J. Respir. Crit. Care Med.* 196 (4), 430–437. doi:10.1164/rccm.201610-2006OC

Wong, A., Plasek, J. M., Montecalvo, S. P., and Zhou, L. (2018). Natural Language processing and its implications for the future of medication safety: A narrative review of recent advances and challenges. *Pharmacother. J. Hum. Pharmacol. Drug Ther.* 38 (8), 822–841. doi:10.1002/phar.2151

Xu, L. B., Zhao, Z. G., Xu, S. F., Zhang, X. X., Liu, T., Jing, C. Y., et al. (2020). The landscape of gene mutations and clinical significance of tumor mutation burden in patients with soft tissue sarcoma who underwent surgical resection and received conventional adjuvant therapy. *Int. J. Biol. Markers* 35 (3), 14–22. doi:10.1177/1724600820925095

Zhang, C., Miao, Z., Xiao, H., Hu, Z., and Ji, Y. (2017). Knowledge graph embedding for hyper-relational data. *Tsinghua Sci. Technol.* 22 (2), 185–197. doi:10.23919/tst.2017.7889640

Zhu, Y., Zhou, W., Xu, Y., Liu, J., and Tan, Y. (2017). Intelligent learning for knowledge graph towards geological data. *Sci. Program.* 2017 (1), 1–13. doi:10.1155/2017/5072427