Advances in Biological Sciences Research (ABSR), volume 4
2nd International Conference on Biomedical and Biological Engineering 2017 (BBE 2017)

Use Text Mining and Complex Networks to Analyze TCM Syndromes

Xing ZHAI[1,a,*], Ping WANG[2] and You-liang HUANG[3]

[1,2,3]Beijing University of Chinese Medicine, 11 Bei San Huan Dong Lu, ChaoYang District, Beijing 100029,Peoples Republic of China

[a]zhaix@bucm.edu.cn

*Corresponding author

Keywords: Text mining, Complex network, Traditional Chinese medicine.

Abstract. Objective: To make intrinsic biological distinctions between qi deficiency pattern and qi stagnation pattern with genes related to NEI. Methods: Establish a data dictionary of NEI-related genes and a keyword lexicon of qi deficiency pattern and qi stagnation pattern. Then retrieve relevant literature on Pubmed database. Obtain compositions of characteristic NEI-related genes of qi deficiency pattern and qi stagnation pattern by using text-mining methodto explore different bioactive materials between the two syndromes. Results: Two kinds of syndromes and their biological networks were constructed. The "qi deficiency" and "qi stagnation" genes based on NEI network were excavated. Conclusion: The genes of qi stagnation pattern relate closely with nerve and endocrine, while those of qi deficiency pattern have a close relationship with the immune system. The intrinsic biological characterizations of TCM patterns can be effectively identified at the level of NEI.

Introduction

In the biomedical field, due to the abrupt growth of the number ofbiological data and biomedical literature, to finding out regulars and new achievements through data mining have become a new hot spot and an important branch of biological study [1]. Text Mining is a specific research area ofthe interdisciplinary subject, data mining. Its main task is to integrate and analyze vast amounts of data to obtain more representative and reliable results [2]. It digs out experimental hypotheses and suggestions from the literatureto enable biologists to verify and achieve new scientific discoveries, which canimprove people's understanding of biomedical phenomena [3].

This study retrieves relevant literature of qi deficiency and qi stagnation patterns on Pubmed database based on the establishments of a data dictionary of NEI-related genes and a keyword lexicon of qi deficiency pattern and qi stagnation pattern. By using literature-mining method, compositionsof characteristicNEI-related genes of the both patterns have been obtained which can be used to explore differentbioactive materials between the two syndromes, including hormones, receptors, cytokines,and neurotransmitters and so on. Meanwhile, taking coronary heart disease as a starting point, clinical data of manifestations of blood stasis and qi stagnation pattern and blood stasis and qi deficiency pattern in patients with coronary disease were collected to verify the results of these articles. This study aims at exploring a systematic, convenient and intuitive biological distinction method and contributing to the establishment of asystematic and objective evaluationsystem of TCM patterns.

Methods

Establishment of Literature Pool of Qi Deficiency Pattern and Qi Stagnation Pattern

Terms related toqi deficiency and qi stagnation pattern that were collected and sorted out according to*Terms of Clinical Practice of Traditional Chinese Medicine in People's Republic of China*, were used as key words to do MeSH (medical subject headings) search in Pubmed database (http://www.ncbi.nlm.nih.gov/pubmed). Abstracts of 41871 articles about qi deficiency pattern and 147,696 articles about qi stagnation pattern were downloaded and saved in format of xml. The process of building literature libraries of related literatures of qi deficiency pattern and qi stagnation pattern was started respectively after those xml documents were sorted out and cleared up.

Table 1. Retrieval keywords of qi deficiency pattern and qi stagnation pattern

Items	Keywords
Qi deficiency	Secret anguish/ vague pain; Dull pain; Dyspnea/ Shortness of breath; Hypodynamia; Spiritlessness; Disinclination to say; Dizziness; Symptoms aggravate after activity; spontaneous perspiration; Light tongue; Weak pulse
Qi stagnation	Distending-pain; Wandering pain/ string pain; Tightness/ oppressive pain/ stuffiness; unstable pain; depression; dysphoria; sigh; Belching; string pulse

Literature Mining Methods

2242 NEI-related genes downloaded from dbNEI database (http://166.111.130.62/portal/root/bi_dbnei/download.jsp) were used as data dictionary to scan literatures in literature libraries of qi deficiency and qi stagnation pattern. Molecular biologicalnetwork was established in the principle of co-occurrence analysis. Co-occurrence analysis was based on a fundamental assumption that if in a large-scale corpus (training corpus), two wordsoften co-occur in the same window unit, it can be considered that these two words are semantically interconnected [4]. The literature mining software (PubMedMiner) which was developed by us at the early stage carried the main parts of literature mining and biological network establishment. The main process of literature mining is shown in Figure 1:

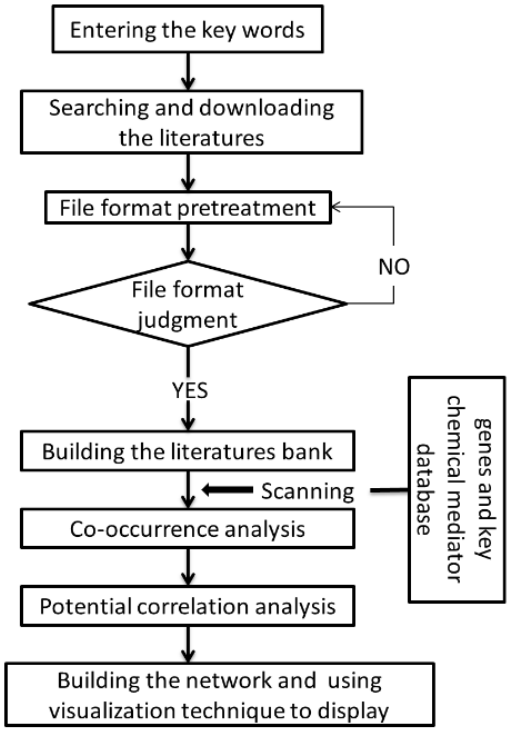

Figure 1. Literature data mining process

Co-Occurrence Analysis

Articles of Related Topics Collected by Co-Occurrence Analysis

The phenomenon of co-occurrence between words suggests the similarity of topics in a way. The greater the co-occurrence of two words isthe more similar the themes of the articles are. For example, if the words shortness of breathfatigue and weary appear simultaneously very frequently in several articles, it can be indicated that these articles may express similar thematic message, which is likely to relate to qi deficiency. According to literature researches [5], the co-occurrence of the words can be calculated by the following formula:

$$v(w_1, w_2) = \frac{1}{2}\left(\frac{P(w_1, w_2)}{P(w_1)} + \frac{P(w_1, w_2)}{P(w_2)}\right) \quad . \tag{1}$$

Wherew_1, $w_2$represent the two key words, for example, if w_1 stands for shortness of breath and w_2 stands for fatigue, then $p(w_1)$ and $p(w_2)$ are the priori probabilities of$w_1$and w_2, showing the probability of occurrence of shortness of breathandweakness and the distribution of them in the document space of an article.$P(w_1,w_2)$ stands for the distribution of (w_1,w_2) in the document space of an article. Generally speaking, the closer the relationship ofthe two words has the stronger thedependence of their occurrenceis which means that they would share a stronger co-occurrence.

The Relationship Between Pattern-Related Key Words and Genes Drawn by Co-Occurrence Analysis

The co-occurrence frequency of terms is widely used to predict the relationship

between terms, which is a currently widely used method with the basic idea that entities appear in the abstract or title of a paper may have a certain relationship. The reliability of this relationship can be strengthened by two approaches. The first approach is to confirm that two closer entities may relate to each other; the second one is to confirm that entities appear simultaneously in different articles may have a certain relationship with each other. [6] [7].

According the research results of reference [8], the following formula as below can be used to calculate the relevancy degree ofterms. Let suppose that $C1 = \{c_{11}, c_{12}, c_{1n}\}$ and $C2 = \{c_{21}, c_{22}, ..., c_{2m}\}$ are two concept sets, where the relevancy degree $RW(c_{1i}, c_{2j})$of c_{1i} and c_{2j} is defined as below:

$$RW(c_{1i}, c_{2i}) = \frac{(F_{ij})^2}{\sum_{p=1}^{m} F_{ip} * \sum_{q=1}^{n} F_{qj}} . \quad (2)$$

In this formula,F_{ij} stands for the co-occurrence frequency of $c_{li}$and c_{2j} in texts. Iftwo concepts $c_{1i}$and c_{2j}were supposed to relate to each other through an intermediate conceptc_{3k}, which is an element of the concept collection $C_3=\{c_{31},c_{32},..., c_{3l}\}$, the relevancy degree of $c_{1i}$and c_{2j} can be calculated by the following formula:

$$RW(c_{1i}, c_{2i}) = p^2 \sum_{k=1}^{i} RW(c_{1i}, c_{3k}) * RW(c_{3k}, c_{2j}) . \quad (3)$$

where P represents the conceptual connection number ofc_{1i} andc_{2j} . Formula 2 is used to calculatethe relationship between syndrome keywords and genes and the direct relevancy degree of gene, while formula 3 is used to calculate the relevancy degree of syndromes keywords and their indirectly associated genes. For example, keyword "Key1"directlyrelates to gene A, and gene A directly relates to the gene B, while Key1 and gene B do not relate directly.

Results

Calculation Results Based on the Combinations of NEI-Based Network-Related Genes of Qi Deficiency Pattern and Qi Stagnation Pattern

It can be indicated from the literature mining resultsand the biological network diagram that there are a total of 317 nodes, 32 of which are the key nodes and 903 edges of qi stagnation pattern network; 32 notes, 8 of which are key notes and 118 edges of of qi deficiency pattern (Table 2). In the network, each node representsa NEI gene excavated from literature library. The connecting line between genes means that two genes appear simultaneously in the abstract of a same article, which is called the degree of this node. If a node degree of is twice greater than the average of all the nodes degrees, then this node degree can be called a key node and its corresponding genes can be called key genes [9].

Table 2. Related genes of qi deficiency pattern and qi stagnation pattern

Qi deficiency Gene(times of repeated occurrence)	MASP1(12),C4(10),C5(8),C7(8),C9(8),C2(7),C3(7),INS(7)
Qi stagnation Gene(times of repeated occurrence)	BDNF(73),INS(58),CRH(38),FST(35),HTR2A(30),CD14(28),LEP(28),PRL(28),AV(27),IL6(26),NPY(26),CCK(25),DRD2(25),MAOA(24),IL10(23),SLC6A4(23),AR(22),CRHR1(21),POMC(19),CD4(18),GCG(17),NRG1(17),EPO(16),CORT(15),NR3C1(15),DRD1(13),DRD4(13),GHRH(13),SLC6A2(13),INSR(12),PTH(12)

Analysis of Pattern Related Genes

NEI network related genes of qi deficiency pattern obtained by literature mining are C4, C5, C7, C9, C2, C3, MASP1 and INS, from which it can be indicated that the qi deficiencyrelated characteristic internal active substancesin NEI network system are closely related to the immune system, especially the complement system.

The NEI -based network-related genes of qi stagnation associated with endocrine system are: AVP, CCK, POMC, GCG, LEP, NPY, GHRH, INS, INSR, PTH, CRH, CRHR1, CORT, PRL, FSH, AR, NR3C1 and EPO. NEI network-related genes of qi stagnationassociated with nervous system are: BDNF, COMT, HTR2A, DRD2, DRD1, DRD4, MAOA, SLC6A4, SLC6A2 and NRG1. The NEI network-related genes of qi stagnation associated with immune system are: CD14, IL6, IL10 and CD4.

The results of intrinsic biological characteristics analysis of qi deficiency based on NEI networkshowed that it is greatly relevant to immune, especially the complement system. The results of intrinsic biological characteristics analysis of qi stagnation based on NEI networkshowed that it is greatly related to neuro-endocrine system, especially the ones related to adrenal gland, thyroid, gonads, andsubstances regulating vascular activity.

Discussion

This study obtained some unique biological active substances of qi deficiency and qi stagnation pattern based on NEI network. It was based on the databases of NEI network- related genes, data mining the literature related to the two common TCM patterns of qi stagnation and qi deficiency with the self-developed PubMedMiner software. The conclusion isthat NEI network related genes closely to neuro-endocrine system, while the ones of qi deficiency associate more closely with immune system was further confirmed by collecting the clinical data of patients with coronary heart disease and selecting patientswithrelativelysimilar age, gender, disease history, family history, and disease duration to have tests with relevant indicators of biological active substances mentioned above, from those who had qi stagnation and blood stasis pattern and qi deficiency pattern and blood stasis pattern.It also further proved that this study could make effectively identification of the intrinsic biological characteristics of TCM patterns.

This study still contains some shortcomings, mainly reflected by the limited depth and range of the study. As for the depth, although we have gotspecific bioactive substances of qi stagnation and qi deficiency pattern, the tests of these clinical

indicators are not commonly made, therefore they cannot be widely used in clinical practice; secondly, the concepts and indicators of the ongoing coronary pharmacological studies still need to be introduced and its mechanism needs to be further explored. As for the study range, this study only concentrates on two patterns: pattern of qi deficiency and blood stasis and pattern of qi stagnation and blood stasis of one single disease, coronary heart disease, and it needs to expand the scope to verify the universality of this method.

Acknowledgement

This research was financially supported by the National Science Foundatio (81603499)

References

[1] Tari L, Anwar S, Liang S, et al. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism[J]. Bioinformatics, 2010, 26(18): i547-i553.

[2] Rodriguez-Esteban R. Biomedical text mining and its applications[J]. PLoS computational biology, 2009, 5(12): e1000597.

[3] Haochang W,Tiejun Z.Biomedical text mining technology research and development [J].Journal of Chinese Information,2008,22(3):89-98.

[4] Bin Q, Ting L. Co-occurrence analysis technology in the application of biomedical information text data mining [J]. The Chinese medicine books intelligence magazine, 2009 (3): 41-43.

[5] Peng C. Based on word co-occurrence research theme text mining model and algorithm [D].Tianjin University, 2010.

[6] AndradeM. A., BorkaP., Automated extraction of information inmoleeular biology. FEBSLetters476(2000)Issue:1-2.12-17.

[7] StePhensM. Palaka1M, MukhoPadhyay S.et al,Deteeting Gene Relations from Medline Abstracts.Pac SymPBiocomPut.2001:483-95.

[8]Xuezhong Z. Text mining application study in traditional Chinese medicine [D]. Zhejiang University, 2004.

[9] Song C, Havlin S, Makse H A. Self-similarity of complex networks [J]. Nature, 2005, 433(7024): 392-395.