Journal of Computers Vol. 31 No. 6, 2020, pp. 246-257
doi:10.3966/199115992020123106019

Applying Metagenomics Sequencing Data to Assistantly Analysize Acne Disease Based on FP-Growth Method

Xue-Yi Gao, Yu Wang*, Meng-Ru Sun

Beijing Key Laboratory of Big Data Technology for Food Safety, School of Artificial Intelligence, Beijing Technology and Business University, Beijing, China
xueyi_g@163.com, wangyu@btbu.edu.cn, cecilia_smr@163.com

Received 18 September 2019; Revised 19 September 2020; Accepted 21 October 2020

Abstract. Acne, as a high incidence of chronic inflammatory skin disease, has a complex etiology and pathogenesis, and microbial colonization is currently considered as one of the important causes. Therefore, in this paper metagenomics data is analyzed using Frequent-Pattern Growth (FP-Growth) method for aided diagnosis of acne disease. The main ideas include that firstly, the data sets are transformed into binary form of 0 or 1. The original data of lipids whose component content is 0 is set 0, and the original data of lipids whose component content is not 0 is set 1. Then the data sets are scanned to build a frequent pattern tree (FP-tree) based on the frequency of each element and support. Finally, FP-tree is used to determine frequent itemsets. The element items in frequent itemsets correspond to the lipids which are highly correlated with the corresponding data. Experimental results on dataset including normal control (NC), acne healthy skin (HS) and acne diseased skin (DS) show that the proposed method can determine the frequent itemsets of different sample sets. By comparing the difference of frequent itemsets, lipids that can distinguish different skin states are also determined, which can provide guiding help for the auxiliary analysis and treatment of skin acne.

Keywords: acne, Frequent-Pattern Growth, lipid, metagenomics

1 Introduction

Acne is a chronic inflammatory skin disease involving the hair follicles and sebaceous glands. The main areas of infection are the sebaceous area such as the face, chest, shoulders and back. The clinical manifestations are papules, acne, pustules, cysts, nodules and scars of different degrees [1]. Acne is a skin disease with a high incidence and a lifetime prevalence is up to 85%, mainly affecting teenagers and young adults [2]. The harm of acne not only affects the appearance of patients, but also causes anxiety, inferiority, depression and other psychological problems [3]. Therefore, the diagnosis and treatment of acne has very important research value and practical significance.

Although the pathogenesis of acne is very complicated, according to the research by related scholars, it can be attributed to four big reasons including the increase of androgen and cortex secretion, excessive keratinization of the hair follicle sebaceous gland, colonization of microbes in the hair follicle sebaceous gland, and secondary inflammation. The reasons are partly related to genetic, immune, dietary and environmental factors [4]. The colonization of microbes in the hair follicles is a hot research topic in the study of the pathogenesis of acne. At present, some studies have found that some microorganisms are closely related to the incidence of acne, such as propionibacterium, staphylococcus and malassezia [5]. At the same time, Jia [6] et al. demonstrated that the skin micro-ecological environment has a great impact on the skin state. But the work is all conducted on a single microorganism, and acne pathogenesis is unknown due to the joint action of multiple microbes. Therefore, in this paper metagenomics are used to analyze the pathogenic factors of acne.

* Corresponding Author

Metagenomics analyzes the growth of microbial communities by measuring the nucleic acid sequence of all microorganisms in the sample [7], overcoming the shortcomings of traditional microbial isolation and culture. Data shows that only 1% of microorganisms in nature can be isolated and cultured, so the development of metagenomics provides new possibilities for analyzing the remaining 99% of unknown microorganisms [8]. At the same time, some studies have found that the metagenomics sample data is closely related to human health status [9]. For example, bifidobacteria can affect the immune function of the baby and thymus development, leading to the growth of related diseases [10]. The ratio between the number of firmicutes and bacteroides in the intestinal tract can be used as a new indicator to measure obesity [11]. Li et al. [12] expounded the role of sebum in the mechanism of acne pathogenesis by a lot of researches. The amount of metagenomic data is relatively large. The machine learning method is suitable for large dataset processing, and has been successfully applied to the research of metagenomics. For example, Koren et al. [13] use clustering method, distance measurement and other methods to detect intestinal type. Qin et al. [14] use genome-wide association analysis to analyze the relationship between type 2 diabetes and changes in intestinal microbial composition.

M.S et al. [15] used mobile devices to capture facial skin data of acne patients and identify acne lesions. Their work divided acne lesions into two categories: papules and pustules. First of all, mobile devices were used to capture the face data of acne patients, carry out spatial calibration, and conduct normalized processing on the images. The regions of interest (ROI) were extracted for acne assessment, and the classification accuracy of papules and pustules was 98%. E.B et al. [16] will be acne as a model disease, using macro genome sequencing, determined the propionic acid bacillus in healthy skin and acne propionic acid bacillus phage abundance is higher. And the authors used the metagenomic data to construct a quantitative prediction model. The model in the study cohort (85%), and independent sample set (86%) in the clinical status of the skin has carried on the classification precision. Sun et al. [17] use multiple sets of canonical correlation analysis methods to analyze lipids that have different effects on different sample sets, and to distinguish different skin states by these lipids. However this method will weaken the effect of the lipids on acne with less content to a certain extent. Therefore, in this paper correlation rules mining algorithm are used to analyze acne metagenomic data. The advantage is that this method will first change the numerical data into binarization, overcoming the defect that the effect of some data are weakened due to the small order of original data.

Common algorithms for mining association rules are Apriori, frequent-pattern Growth (FP-Growth), Eclat, etc. The core idea of Apriori algorithm [18] is the recursive algorithm of two-stage frequency set idea. Frequent itemsets are searched by layer-by-layer search strategy, and association rules are determined by frequent itemsets. Due to too many candidate frequent itemsets generated by this method and too many times of scanning database, the algorithm efficiency is low. In view of the above two problems, Han et al. [19] propose FP-Growth algorithm, whose core idea is to compress data to a Frequent Pattern Tree (FP-Tree), and frequent itemsets can be mined in FP-Tree. FP-Tree only scans data sets twice, overcoming the problem that Apriori algorithm scans data set multiple times, which effectively reduces the time complexity. Different from Apriori, FP-Growth algorithm, Eclat [20] algorithm is based on the depth-first algorithm. The original data is inverted for finding the intersection of frequent k itemset to generate candidate k+1 itemset, cutting the k+1 itemset to generate frequent k+1 itemset, and finding the intersection to generate candidate k+2 itemset, and so on, until itemset is unified. Based on the advantages and disadvantages of the above algorithms, FP-Growth algorithm with moderate time complexity and space complexity is selected to analyze the acne metagenome data and to determine lipids that can distinguish different sample sets.

In summary, acne has a great impact on people's quality of life, and the etiology of acne disease is not yet clear. The purpose of this study is to determine the types of lipids related to acne diseases from the perspective of metagenomics, try to determine the pathogenesis of acne diseases. During the experiment, many algorithms were compared, FP-Growth was determined and successfully determined several types of microorganisms related to acne diseases. Six lipids including No. 604, No. 1149, No. 1213, No. 1265, No. 2277 and No. 2506 can judge whether tested individuals suffer from acne. When patients with acne receive treatment, if the content of No. 1318, No. 1320, No. 1328 and No. 1331 is continuously reduced, it indicates that their skin condition is improved.

2 Method

2.1 The Data Set

The data of acne metagenomics used in this study are provided by Key Laboratory of Cosmetic of China National Light Industry, Beijing technology and business university. During the data collection, chromatographic equipment (Waters ACQUITY UPLC i-class (Waters Corporation, Milford, Massachusetts, USA) was used for the test, with a flow rate of 0.3mL/min and injection volume of 2.0μL.Ultra Performance Liquid Chromatography (UPLC) system eluate was introduced into a high-resolution quality measurement equipment (Waters Xevo g2-xs qtof-ms (Waters Corporation, Milford, Massachusetts, USA) under Chromatography flow velocity. The equipment had an interface of Electrospray ionization (ESI) which was operated in a positive ion mode. Nitrogen was used as the atomized and dissolved gas. The number of high-resolution quality measurement obtained by the Masslynx 4.1 (Waters Corporation, Milford, Massachusetts, USA) was the raw data. The raw data includes 35 healthy volunteers' facial skin data as Normal Control group (NC), and 35 patients' facial skin data with acne facial skin disease (DS) data and healthy skin (HS), and each of the data is collected 2520 sequences.

2.2 FP-Growth Algorithm

The core idea of the FP-Growth algorithm is to compress the data set into an FP-tree while preserving the relationship between the attributes in the data set. Then the frequent itemsets are solved in the FP-tree. The FP-Growth algorithm can be summarized into two stages including the construction of frequent pattern tree and the mining of frequent itemsets.

For the acne metagenomic data set, let each person be a transaction and 2520 sequences as element items. Since the data used in this experiment are numeric data, and the FP-Growth algorithm requires the data to be Boolean, the data in this experiment are processed to 0-1 before the experiment, i.e. the non-existent lipids on the component content are zeroized, and the existing lipids are set 1.This method can overcome the problem that the effect of lipid data is weakened due to its small order of magnitude to some extent.

A simple example is illustrated the FP-Growth algorithm steps. Assume that the data set is as shown in Table 1. Mining its frequent itemsets is divided into the following two steps.

Table 1. The data set

ID	element items
001	r, z, h, j, p
002	z, y, x, w, v, u, t, s
003	z
004	r, x, n, o, s
005	y, r, x, z, q, t, p
006	y, z, x, e, q, s, t, m

The construction of frequent pattern tree. The specific steps to build a frequent pattern tree are as follows:

(1) The frequency of each element item is counted in the original data set. According to the data set as shown in Table 1, a total of 17 element items can be calculated. Then occurrence times of each element item in the data set can be calculated separately. The main function is to generate tree nodes of the tree of frequent patterns.

(2) Support filtering is made. Not all the tree nodes calculated by step (1) are required. The minimum support is set to eliminate items with less than frequency of occurrence. For element items that do not meet the minimum support, it can be thought that this element item is infrequent. According to the prior property, if an element item is infrequent, all supersets of this element item are infrequent. In this example, the minimum support is set to 3. The element items obtained after support filtering are shown in Table 2. And the minimum support is set 33 in the final experiment

Table 2. The element items obtained after support filtering

element items	the frequency of element item
r	3
z	5
y	3
x	4
t	3
s	3

(3) Frequent element item is ordered. The element items filtered by the support degree (i.e. frequent element items) are sorted according to the frequency of occurrence, and the data set is also reordered according to the sorted element items, as shown in Table 3.

Table 3. The reordered data set according to the sorted element items

ID	the reordered data set
001	z, r
002	z, x, y, t, s
003	z
004	x, r, s
005	z, x, y, t, r
006	z, x, y, t, s

(4) Frequent pattern trees are built. Null is started, and frequent element items are added to the tree. The value of the existing element item is increased if it already exists. A branch to the tree is added if it does not. The frequent pattern tree is shown as Fig. 1.

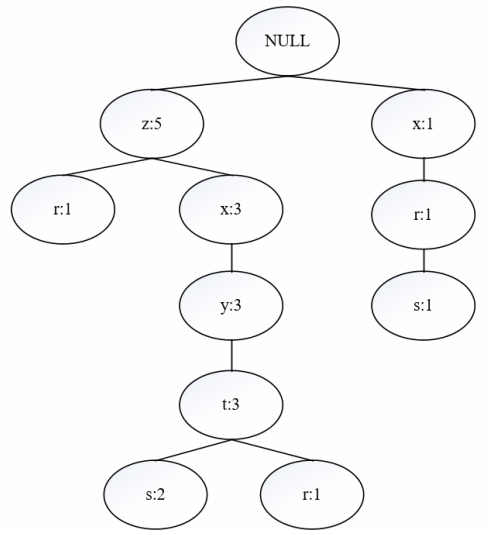


Fig. 1. The built frequent pattern tree equation

The mining of frequent itemsets. After the frequent pattern tree is built, the frequent itemsets can be obtained by mining the frequent pattern tree. The three basic steps for extracting frequent itemsets from the frequent pattern tree are as follows.

(1) The conditional pattern base is obtained from the frequent pattern tree which is composed of a certain frequent element item in a frequent pattern tree together with its prefix paths, that is, The found element item is considered as a set of paths to the end, and each path is actually a prefix path as shown Table 4.

Table 4. The obtained conditional pattern base from the frequent pattern tree

the frequent itemset	the conditional pattern base
z	{}: 5
x	{z}: 3, {}: 1
r	{z, x, y, t}: 1, {z}: 1, {x}: 1
y	{z, x}: 3
t	{z, x, y}: 3
s	{z, x, y, t}: 2, {x, r}: 1

(2) The conditional frequent pattern tree is constructed by using conditional pattern base. To find more frequent itemsets, a conditional frequent pattern tree is created for each frequent element item. For a frequent element item, the frequent element item whose occurrence frequency is greater than the minimum support is determined in its conditional pattern base, and these element items are frequent items.

(3) Steps (1) and (2) are repeated until the frequent pattern tree is empty. All frequent items are recorded to form a frequent itemset. Taking element item {t} as an example, Fig. 2 shows the conditional pattern base using t, and finally obtains the frequent itemset of element item {t}, in which the minimum support is 3.

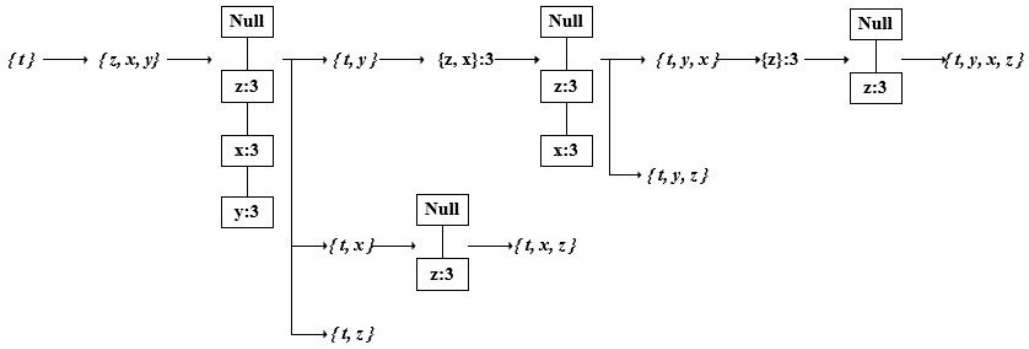


Fig. 2. The frequent itemset with element item {t}

3 Result

In order to verify the effectiveness of the proposed algorithm, a series of experiments are carefully designed in this paper. FP-Growth algorithm is used to analyze the sample sets of DS, HS and NC respectively, and the frequent itemsets corresponding to each sample set is obtained. The sample sets are the lipid content of each subject.

After data mining analysis by FP-Growth algorithm, frequent itemsets obtained from DS, HS and NC sample sets not only contain the same element, but also contain different elements. By comparing the similarities and differences of frequent itemsets in different sample sets, an auxiliary analysis can be conducted for skin acne.

3.1 Lipids that Co-exist in Frequent Itemsets of Three Sample Sets

Table 5 shows lipids and their specific descriptions in frequent itemsets of DS, HS and NC samples.

Table 5. The same elements in the frequent itemset of three sample sets

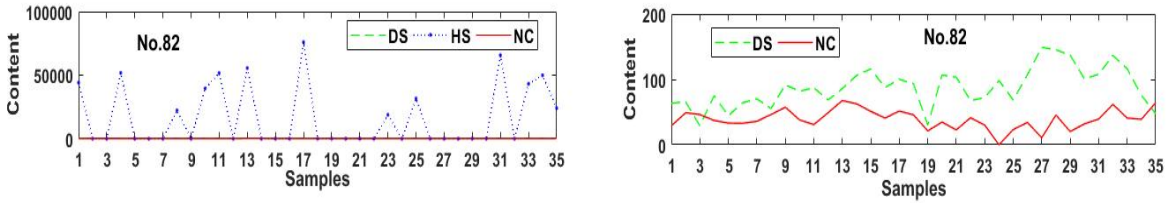
No	specific descriptions
82	2E,4E,8E,10E-Dodecatetraenedioic acid
172	9, 10-dihydroxy-Octadecanedioic acid
2053	PS(P-16:0/18:1(9Z))

No. 82, No. 172 and No. 2053 are lipids in frequent itemsets of DS, HS and NC sample sets, and their contents in different sample sets have obvious change trend as shown in Fig. 1. In the figure, the abscissa

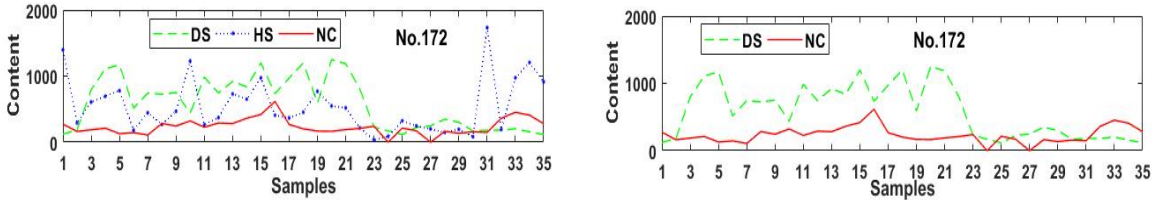
represents the sample number and the ordinate represents the corresponding lipid content. In Fig. 3(a), the left figure shows the content of the lipid represented by No. 82 in DS, HS and NC samples. It can be seen that the content of No. 82 in healthy skin samples of acne patients increased significantly. In Fig. 3(a), the right figure shows the lipid content of No. 82 in the affected skin of acne patients and the normal control group. Although the content of No. 82 in the two samples is not high, it can be clearly seen that the content of No. 82 in the DS sample set is generally higher than that in the NC sample set.

In Fig. 3(b), the lipid content represented by No. 172 in DS, HS and NC samples is shown on the left. It can be seen that the lipid content of No. 172 in the normal control group is the lowest, followed by the healthy skin sample set of acne patients, and the content of the diseased skin sample set of acne patients is the highest. According to Fig. 3(b), lipids represented by No. 172 can distinguish three sample sets in terms of lipid content.

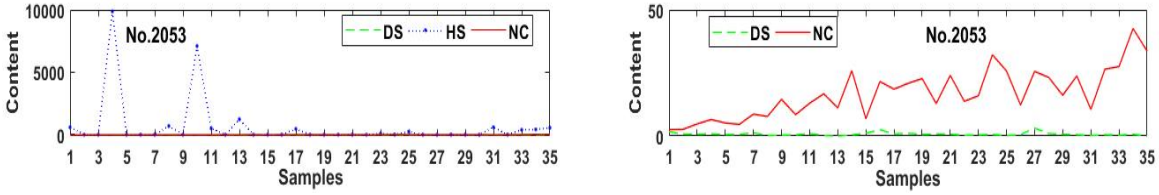
In Fig. 3(c), the content of lipid No. 2053 in DS, HS and NC sample sets is shown. According to Fig. 3(c), the lipid content of No. 2053 in the healthy skin sample set of acne patients is relatively high, and the content of No. 2053 in the diseased skin sample set of acne patients and the normal control group is shown in the right figure of Fig. 3(c). The content of No. 2053 in the normal control group is significantly higher than that in the diseased skin sample set of acne patients. The content of No. 2053 in the diseased skin samples of acne patients is the least, while the content of No. 2053 in the other two samples is relatively higher. It can be considered that when an acne patient is receiving treatment, if the lipid content of No. 2503 increases, it indicates that the treatment is effective and the skin condition of the patient is improving continuously.



(a) No. 82



(b) No. 172



(c) No. 2053

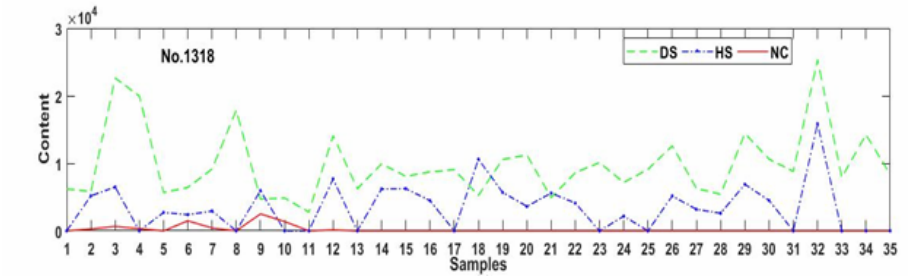
Fig. 3. The content of same lipid in frequent itemsets of DS, HS and NC sample sets

3.2 Lipids that Exist only in the Frequent Itemset of the DS Sample Set

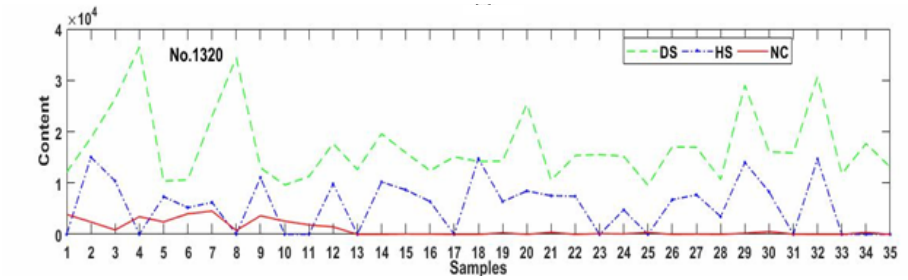
In addition, FP-Growth algorithm is used to analyze the data of the acne metagenomic sample sets, obtaining the lipids corresponding to the frequent itemset that only exist in one of the sample sets and do not exist in the other two sample sets. There are 16 lipids in total, whose specific description is similar to Table 1, which will not be shown here because of limited space.

Fig. 4 shows lipids that only appear in the frequent itemset of DS sample sets, and the contents of these lipids are significantly different in DS, HS and NC sample sets. In Fig. 4, the abscissa represents the

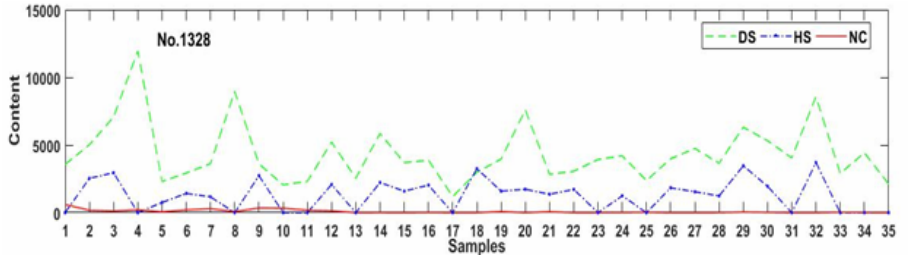
sample number and the ordinate represents the lipid content. It can be seen from Fig. 4 that the lipid contents represented by No. 1318, No. 1320, No. 1328 and No. 1331 have the same trend with respect to DS, HS and NC sample sets, that is, the content of the diseased skin sample set of acne patients is the highest, followed by the healthy skin of acne patients, and finally the normal control group. Therefore, these lipids can be used as indicators to distinguish DS, HS and NC skin states.



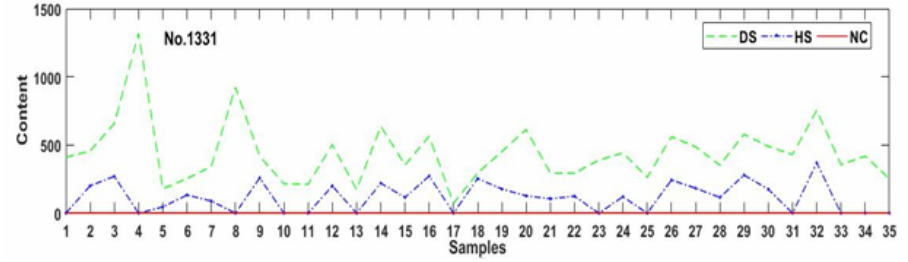
(a) No. 1320



(b) No. 1328



(c) No. 1331



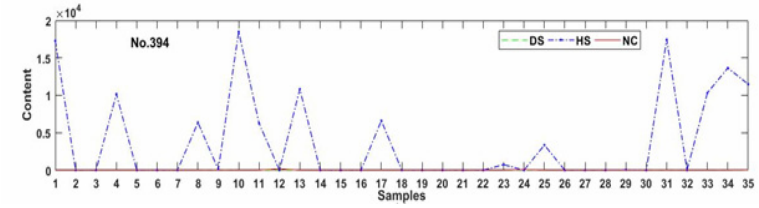
(d)

Fig. 4. Lipids only in the frequent itemset of DS sample set

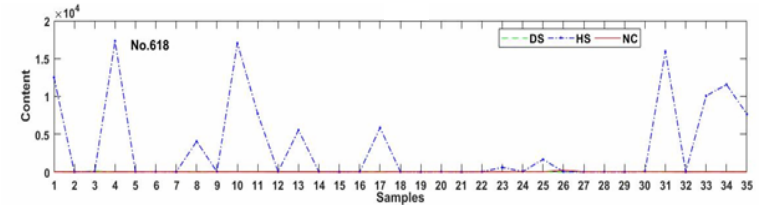
3.3 Lipids that Exist only in the Frequent Itemset of the HS Sample Set

Meanwhile, FP-Growth algorithm is used to analyze acne metagenomics data of healthy skin sample set of acne patients, and the lipids that only appeared in the frequent items of HS sample set can be obtained, a total of 28 kinds. Among them, 6 lipids show significant differences in the samples of DS, HS and NC groups, including No. 394, No. 618, No. 776, No. 1358, No. 2065 and No. 2334. In the figure, the abscissa represents the sample number in the three sample sets, and the ordinate represents the corresponding lipid content.

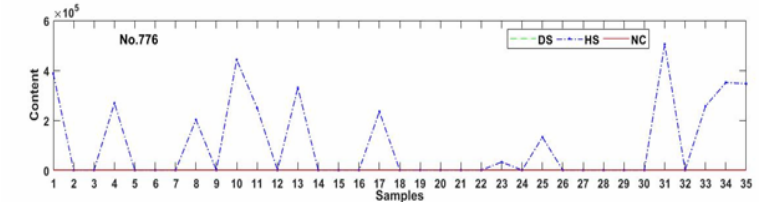
As shown in Fig. 5, the contents of No. 394, No. 618, No. 776, No. 1358, No. 2065 and No. 2334 show the same change trend in the samples of DS, HS and NC, that is, their contents are significantly high in the samples of the NC group, while the contents of the samples of the diseased skin of acne patients and the healthy skin of acne patients are 0. Although not all samples of 6 lipids in the NC group show high values, the lipid contents of No. 394, No. 618, No. 776, No. 1358, No. 2065 and No. 2334 show a significant increase compared with the other two samples.



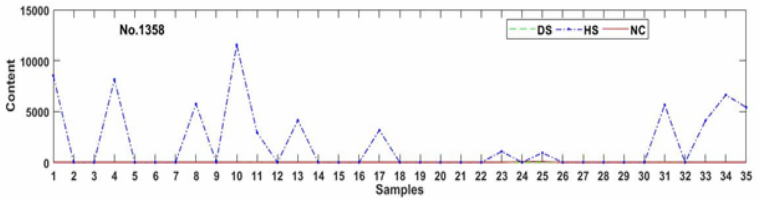
(a) No. 394



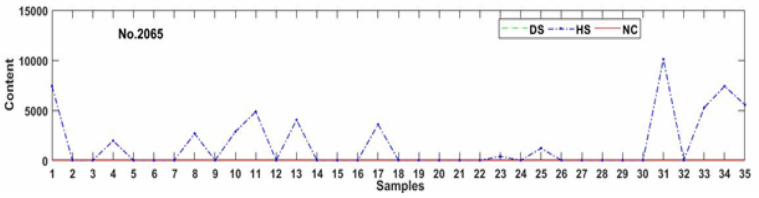
(b) No. 618



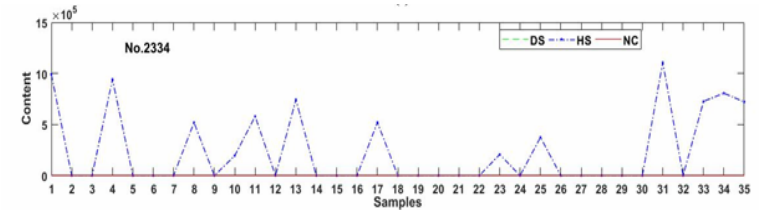
(c) No. 776



(d) 1358



(e) No. 2065



(f) No. 2334

Fig. 5. Lipids only in the frequent itemset of HS sample set

3.4 Lipids that Exist Only in the Frequent Itemset of NC Sample Set

Finally, in this paper, FP-Growth algorithm is used to analyze the sample set data of the NC group, and lipids that only exist in the frequent itemset of NC sample set can be obtained, with a total of 16 lipids. Fig. 6 shows that there are significant differences on lipid contents of DS, HS and NC groups. In Fig. 4, the abscissa represents the sample number and the ordinate represents the corresponding lipid content. As can be clearly seen from Fig. 6, the contents of 6 kinds of lipids including No. 604, No. 1149, No. 1213, No. 1265, No. 2277, No. 2506 in NC group are obviously higher than those of patients with acne skin disease and the patients' healthy skin. And the contents of these 6 kinds of lipids in the diseased skin sample set and healthy skin samples of acne patients are near the zero, and the contents of sample set in the NC group will have a more dramatic increase. As can be seen from the above, No. 604, No. 1149, No. 1213, No. 1265, No. 2277, No. 2506 can be used as indicators to determine whether the treatment of acne patients is effective or not.

3.5 Lipids that Can DistingUish Different Sample Sets

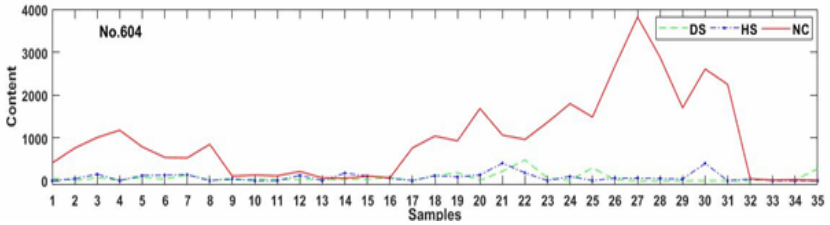
In summary, after analyzing the sample sets of DS, HS and NC according to the FP-Growth algorithm, the frequent itemsets corresponding to each sample set can be obtained. Different sample sets can be distinguished according to the similarities and differences of lipids in the frequent itemsets and the contents of lipids in different sample sets. Among them, lipids represented by No. 604, No. 1149, No. 1213, No. 1265, No. 1318, No. 1320, No. 1328, No. 1331, No. 2277 and No. 2506 show a certain change trend in DS, HS and NC sample sets, as shown in Fig. 7. In the figure, the abscissa represents the sample sets of DS, HS and NC, and the ordinate represents the average content of lipids in the corresponding sample set.

In Fig. 7(a), the average lipid contents of No. 1318, No. 1320, No. 1328 and No. 1331 show a decreasing trend in DS, HS and NC sample sets, that is, the lipid content is the highest in the diseased skin of acne patients, followed by the healthy skin of acne patients, and finally the NC group. As shown in Fig. 7(a), the 4 lipids including No. 1318, No. 1320, No. 1328 and No. 1331 can be used as the indicators to judge the skin state. When the acne patients receive treatment, if the contents of the lipids in these four lipids are continuously reduced, it indicates that the skin state of the patients is continuously improved and the treatment is effective.

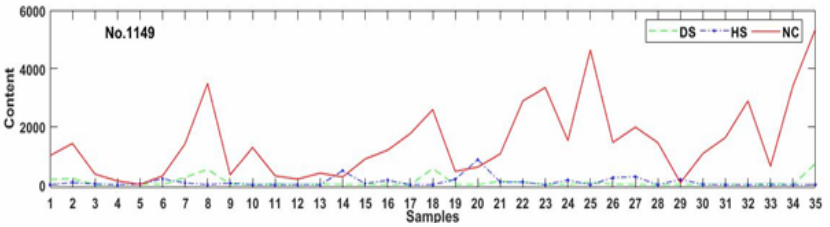
Fig. 7(b) shows the changes in the average content of 6 lipids in the sample sets of DS, HS and NC. The 6 lipids include No. 604, No. 1149, No. 1213, No. 1265, No. 2277 and No. 2506 respectively. As can be clearly seen from Fig. 7(b), the average content of these 6 lipids in the diseased skin of acne patients and the healthy skin of acne patients is roughly unchanged, while there is a significant increase in the NC group. Therefore, these 6 lipids can be used as indicators to determine whether the skin is healthy or not.

4 Conclusion

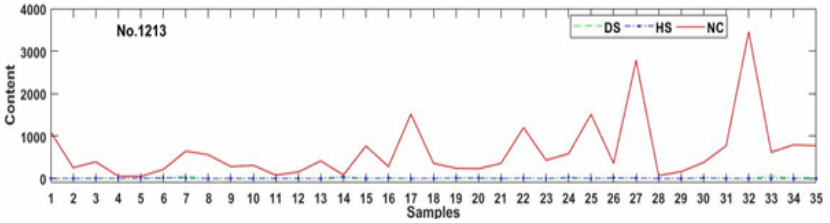
Acne is a skin disease with a very high incidence, which has a great impact on the life of patients with acne. Therefore, the treatment and auxiliary analysis of acne have a good application prospect. In this paper, FP-Growth algorithm is used to analyze acne metagenomic data, and the data is transformed into a Boolean type. While frequent itemsets of the DS, HS and NC sample set can be obtained, the influence of value size in the analysis process is eliminated. Lipids whose content has obvious changes in three groups of samples can be determined by comparing the similarities and differences between lipids in frequent itemsets. The experimental results show that FP-Growth algorithm can effectively determine lipids that can distinguish different sample sets. Six lipids including No. 604, No. 1149, No. 1213, No. 1265, No. 2277 and No. 2506 can judge whether tested individuals suffer from acne. When patients with acne receive treatment, if the content of No. 1318, No. 1320, No. 1328 and No. 1331 is continuously reduced, it indicates that their skin condition is improved. From what has been discussed above, lipids that distinguish different sample sets can be effectively obtained by analyzing acne metagenomic data with FP-Growth method, which provids theoretical support for the prevention, diagnosis and treatment of acne.



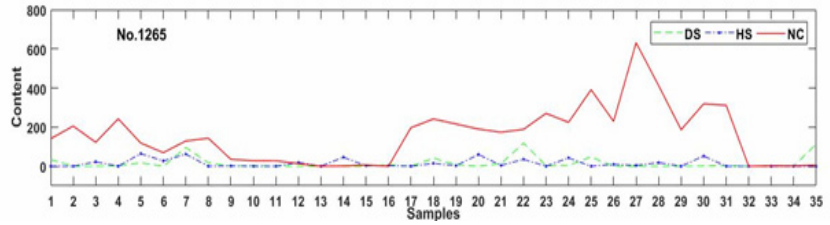
(a) No. 604



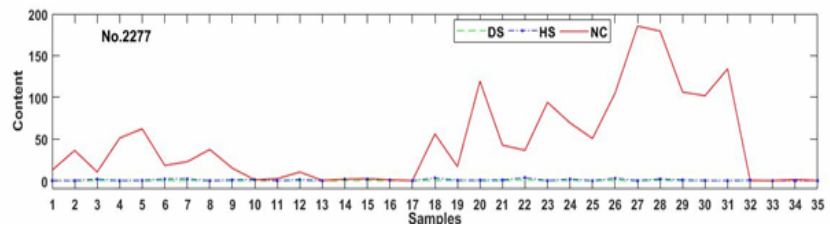
(b) No. 1149



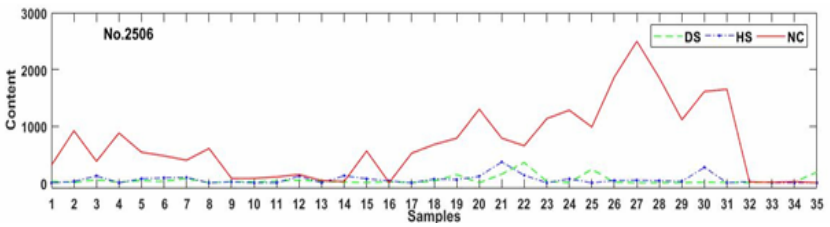
(c) No. 1213



(d) 1265

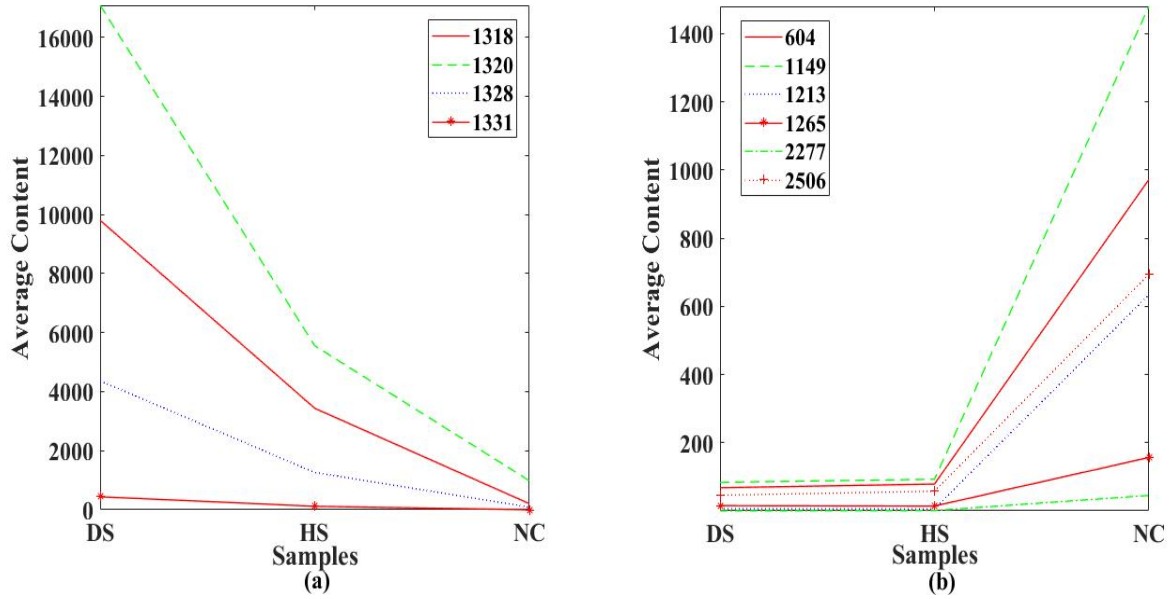


(e) No. 2277



(f) No. 2506

Fig. 6. Lipids only in the frequent itemset of NC sample set



(a) No. 1318, No. 1320, No. 1328, No. 1331

(b) No. 604, No. 1149, No. 1213, No. 1265, No. 2277, No. 2506

Fig. 7. Contents change trend of lipids

This method has been successful in identifying the types of lipids associated with acne, but there is still a lot of work to be done, such as classification issues. According to the lipids determined in this paper, an effective classifier can be determined to quickly and accurately determine the skin status of new individuals when they are tested for acne diseases, providing a fast and accurate method for the medical community to detect acne diseases. This is a forward-looking topic.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61671028), and Joint Project of Beijing Natural Science Foundation and Beijing Municipal Education Commission (No. 21JD0016).

References

[1] G. Maroni, M. Ermidoro, F Previdi, G. Bigini, Automated detection, extraction and counting of acne lesions for automatic evaluation and tracking of acne severity, in: Proc. 2017 IEEE Symposium Series on Computational Intelligence (SSCI), 2017.

[2] A.U. Tan, B.J. Schlosser, A.S. Paller, A review of diagnosis and treatment of acne in adult female patients, International Journal of Women's Dermatology 4(2)(2018) 56-71.

[3] T.B. Zhang, Y.P. Bai, Progress in the treatment of acne vulgaris, Chinese Journal of Dermatovenereology of Integrated Traditional and Western Medicine 18(2)(2019) 180-182.

[4] T.X. Cong, D. Hao, X. Wen, X.-H. Li, G. He, X. Jiang, From pathogenesis of acne vulgaris to anti-acne agents, Archives for Dermatological Research 311(5)2019 337-349. DOI: 10.1007/s00403-019-01908-x.

[5] Y. Wu, J. Ji, L.L. Zhang et al., Roles of microorganisms in the pathogenesis of acne, The Chinese Journal of Dermatovenereology 30(3)(2016) 311-314.

[6] Y. Jia, Y. Gan, C.F. He, Z. Chen, C. Zhou, The mechanism of skin lipids influencing skin status, Journal of Dermatological Science 89(2)(2018) 112-119.

[7] J. Handelsman, Metagenomics: application of genomics to uncultured microorganisms, Microbiology and Molecular Biology Reviews 69(1)(2005) 195.

[8] J. Ma, Research on classification of metagenomics sequencing fragment, Harbin Institute of Technology, 2017.

[9] X. Zhu, S.B. Zhu, T.J. Zhang, Association between metagenome and human health: a research progress, Chinese Journal of Public Health 35(1)(2019) 122-124.

[10] M.N. Huda, Z. Lewis, K.M. Kalanetra, M. Rashid, S.M. Ahmad, R. Raqib, F. Qadri, M.A. Underwood, D.A. Mills, C.B. Stephensen, Stool Microbiota and vaccine responses of infants, Pediatrics 134(2)(2014) 362-372.

[11] T. Korem, D. Zeevi, J. Suez, A Weinberger, T Avnit-Sagi, M Pompan-Lotan, E Matot, G. Jona, A. Harmelin, N. Cohen, A. Sirota-Madi, C.A. Thaiss, M. Pevsner-Fischer, R. Sorek, R.J. Xavier, E. Elinav, E. Segal, Growth dynamics of gut microbiota in health and disease inferred from single metagenomics samples, Science 349(6252)(2015) 1101-1106.

[12] X.C. Li, C.F. He, Z. Chen, C. Zhou, Y. Gan, Y. Jia, A review of the role of sebum in the mechanism of acne pathogenesis, Journal of cosmetic dermatology 16(2)(2017) 168-173.

[13] O. Koren, D. Knights, A. Gonzalez, L. Waldron, N. Segata, R. Knight, C. Huttenhower, R.E. Ley, A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets, PLoS Computational Biology 9(1)(2013) e1002863.

[14] J.J. Qin, Y. R. Li, Z.M. Cai et al., A metagenome-wide association study of gut microbiota in type 2 diabetes, Nature 490(2012) 55-60.

[15] F. Vasefi, W. Kemp, N. Mackinnon, M. Valdebran, K. Huang, H. Zhang, Automated facial acne assessment from smartphone images, in: Proc. Imaging, Manipulation, & Analysis of Biomolecules, Cells, &Tissues XVI. 2018.

[16] E. Barnard, B. Shi, D. Kang, N. Craft, H. Li, The balance of metagenomics elements shapes the skin micrebiome in acne and health. Scientific Reports 6(1)(2016) 39491.

[17] M.R. Sun, Y. Wang, C.F. He et al., Assisted analysis of acne metagenomic sequencing data using multi-set canonical correlation analysis methods, CAAI Transactions on Intelligent Systems. DOI: 10.11992/tis.201810005.

[18] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proc. 20thVLDB Conference Santiago, 1994.

[19] J.J. Han, J. Pei, Y.W. Yin, Mining frequent patterns without candidate generation, in: Proc. 2000 ACM SIGMOD International Conference on Management of Data, 2000.

[20]M.J. Zaki, Scalable algorithms for association mining, IEEE Transactions on Knowledge and Data Engineering 12(3)(2000) 372-390.