Research on the Similarity Algorithm of Chromatographic Fingerprint Based on Information Entropy

Hang Wei[1, a], Li Lin[2, b], Pingping Chen[1, c], Dingying Tan [1, d] and Qinqun Chen [1, e]
[1] School of Medical Information Engineering, Guangzhou University of Chinese Medicine, Guangzhou, China

[2] Chinese Material Medical College, Guangzhou University of Chinese Medicine, Guangzhou, China

[a]crwei@gzucm.edu.cn, [b] lwyll @gzucm.edu.cn, [c] chenpingping@gzucm.edu.cn,

[d] tandy @gzucm.edu.cn ,[e] chenqq @gzucm.edu.cn

Keywords: information entropy, chromatographic fingerprint, similarity

Abstract. In order to reduce the error of interangle cosine method or correlation method in the chromatographic fingerprints analysis, an improved similarity algorithm that assigns different weight according to information entropy was proposed in this paper. In this way, 23 samples of Exocarpium Citrus Grandis collected from different places were obviously classified as two different species, which solved the problem that it is difficult to make an accurate identification either by the traditional similarity algorithm or its' optimization by coefficient of variation. Further more, better results were obtained by compared with discrimination model base on principal component analysis and artificial neural network. So, the method based on information entropy was more suitable for quick, effective discrimination of species and origin of traditional Chinese medicine.

Introduction

Traditional Chinese medicine has a long history and still serves widely. However, owning to its complex component, almost not a single effective part but effective group, so quality control is the key to TCM modernization and internationalization. Along with the development of information technique and analysis means of TCM, chromatographic fingerprint technique has become an important method to discriminate varieties, origin and control quality of TCM [1]. The similarity calculation of chromatographic fingerprint of TCM usually adopts interangle cosine method or correlation method. The traditional method could calculate similarity according to the whole data or the data (area or height) of peaks from the fingerprint [2]. However it has certain limitation: the characteristic may be submerged due to its equal weight. Thus this causes the error for the whole similarity results. For example, it is difficult to identify *Citius Grandis* Osbeck. Var. tomentosa Hort.(CGO Var. TH) from *Grandis* (L.) Osbeck (CGO) by traditional method, for the similarity of the two different species is almost more than 0.95 and adjacent. CGO Var. TH and CGO both origin from Exocarpium Citrus Grandis, but the former is much better than the latter in quality [3]. In this paper, a new method of similarity calculation that assigns the different weight according to information entropy is developed. By applying this method in discrimination of two different species of Exocarpium Citrus Grandis by high performance liquid chromatographic (HPLC) data, reasonable results were obtained and the usability of the method was proved out.

Theory

A. The Traditional Similarity Algorithm

The similarity algorithm of chromatographic fingerprint of TCM usually adopts interangle cosine method or correlation method. Both could reflect the shape of chromatographic fingerprint similar to or not, that is, whether TCM both has the same components and the same relative proportion [4].The method of correlation is generally defined as,

$$r_{ij} = (\sum_k (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_k)) / (\sqrt{\sum_k (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_k (x_{jk} - \bar{x}_j)^2}), \quad (k = 1,2,\cdots,m) \tag{1}$$

Therein $\bar{x}_i$ and $\bar{x}_j$ are the means of the two chromatogram data, *m* is the number of peaks.

The method of interangle cosine is defined as,

$$\rho_{ij} = (\sum_k x_{ik} x_{jk}) / (\sqrt{\sum_k x_{ik}^2} \sqrt{\sum_k x_{jk}^2}), (k = 1,2,\cdots,m) \tag{2}$$

B. The Similarity Algorithm Based on Information Entropy

The original similarity algorithms apply the equal weight to calculate similarity, which may submerge the characteristic and thus lead to the misjudgment of the whole similarity. To overcome this limitation, an objective method of assignment weight based on information entropy is proposed. Information entropy measures the uncertainty according to probability theory. It indicates that the more discretely data distributes, the greater the uncertainty is [5].The information entropy of a peak is signed as h_k,

$$h_k = -\sum_l f_{lk} \ln f_{lk}, (l = 1,2,\cdots,n) \tag{3}$$

Where *n* is the number of chromatograms, $f_{lk} = x_{lk} / \sum_{l=1}^{n} x_{lk}$, x_{lk} is the characteristic value of the peak in the chromatogram .While f_{lk}=0， h_k=0.

Equation (3) reflects the balanced degree of each peak: The higher entropy value, the more balanced decision-making information.

To ensure comparability among the various peaks, the balanced degree can be expressed as following,

$$j_k = -\sum_l f_{lk} \ln f_{lk} / \ln n \tag{4}$$

Then the divergence of a peak can be expressed as：

$$d_k = 1 - j_k \tag{5}$$

It is shown that, the more discretely the characteristic value of the peak distributes, the greater the corresponding divergence is, which indicates the more important role it plays. On the contrary, the results contrast. When the characteristic value of the peaks in each fingerprint is equal, that is to say the peak does nothing. At last, the weight of the peak can be assigned by information entropy.

$$w_k = d_k / \sum_{k=1}^{m} d_k = 1 - j_k / m - \sum_{k=1}^{m} j_k \tag{6}$$

Combining entropy weight, the interangle cosine method and correlation method can be optimized as,

$$r_{ij} = (\sum_k w_k^2 (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_k)) / (\sqrt{\sum_k w_k^2 (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_k w_k^2 (x_{jk} - \bar{x}_j)^2}) \tag{7}$$

$$\rho_{ij} = (\sum_k w_k^2 x_{ik} x_{jk}) / (\sqrt{\sum_k w_k^2 x_{ik}^2} \sqrt{\sum_k w_k^2 x_{jk}^2} \tag{8}$$

Results analysis and discussion

Using the original chromatography of 23 samples of Exocarpium Citrus Grandis (12 samples of CGO Var. TH collected from Huazhou, Guangdong and 11 samples of CGO purchased from different places)obtained by the experiment operated in literature [6], a number of peaks can be gotten(Fig.1, Fig.2), which is the key factor to evaluate the integrity of fingerprint . In all the fingerprint of the samples, most peaks appeared in the range of 25.00~68.00 min, which indicated the range can be regarded as the characteristic range of Exocarpium Citrus Grandis. Furthermore, 22 peaks were selected as the “characteristic peaks”, according to the rule that which separates highly and with an area above 0.5%. Then, the data of each HPLC corresponding peak for similarity analysis was standardized as follow. Firstly, the relative retention time are confirmed by naringin (RT=34.70min)

and rhoifolin (RT=41.70min), which are the two most main constituents. And the relative area is normalized.

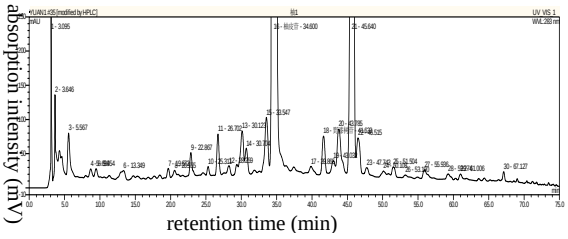

Figure 1. HPLC fingerprint of CGO.Var.TH

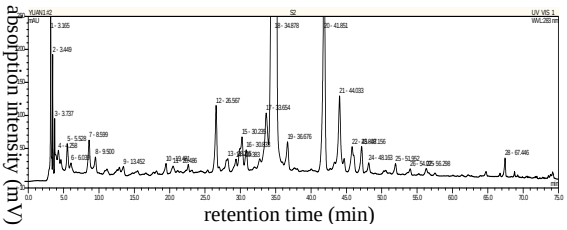

Figure 2. HPLC fingerprint of CGO

Based on information entropy, the divergence and weight of each peak is confirmed, and the results are shown in Tab.1. It indicated that information entropy could reflect the influence degree of each characteristic peak appropriately. The weight of No. 8 Peak(identified as naringin),whose content is much greater than other peaks in both CGO Var. TH and CGO, is reduced greatly as a result of its little divergence. On the other hand, the weight of No.11 Peak (identified as rhoifolin) and No.15 Peak (identified as naringenin) is both increased relatively, which reflect their great divergence .The conclusion above is consistent with the literature [7]. Furthermore, it could be found that the weight of No.10 Peak is also obvious, even higher than those of the content of rhoifolin and naringenin. From the original data, it could be interpreted that this peak almost appears not in GCO but in CGO Var. TH, although its content is inconspicuous and not studied so far. In general, the results revealed that information entropy could offer reasonable measurement of importance of each peak.

TABLE Ⅰ

DIVERGENCE AND WEIGHT OF EACH PEAKS

Peak No.	Divergence	Weight	Peak	Divergence	Weight
1	0.01372	0.00371	12	0.27554	0.07454
2	0.09099	0.02461	13	0.00236	0.00064
3	0.16088	0.04352	14	0.39956	0.10809
4	0.07311	0.01978	15(naringenin)	0.20408	0.05521
5	0.04145	0.01121	16	0.03021	0.00817
6	0.10846	0.02934	17	0.13316	0.03602
7	0.04081	0.01104	18	0.99885	0.27021
8(naringin)	0.00524	0.00142	19	0.16330	0.04418
9	0.20322	0.05497	20	0.04286	0.01160
10	0.38471	0.10407	21	0.08510	0.02302
11(rhoifolin)	0.09186	0.02485	22	0.14714	0.03980

Combined with information entropy, the similarity results of each sample referred to the “common model” by simulative mean all 12 samples of CGO Var. TH are listed in Tab.2.Comparing of similarity results obtained by various methods, we found that the optimum method is information entropy. By entropy, the interangle cosine for each chromatogram of CGO Var. TH to the common model ranged from 0.8957 to 0.9716, in contrast, the values of each chromatogram of CGO concentrated in the vicinity of 0.4. At the same time , improved by entropy, the correlation coefficient among chromatograms of CGO Var. TH were higher than 0.995, however the vales of CGO are below 0.941 excluded CGO 2.On the other side, calculating similarity of all samples of CGO Var. TH and CGO to common model of CGO Var. TH ,the results gotten by traditional method or improvement by coefficient of variation concentrated in the vicinity of 0.99,which means the methods are not suitable to distinguish the two different species of Exocarpium Citrus Grandis.

Among the 23 samples of Exocarpium Citrus Grandis, the content of rhoifolin in CGO Var. TH varied between 0.177% and 1.33%.,conversely those in CGO mostly below 0.17% except for CGO 2.CGO.2 comes from Guangxi, the content of rhoifolin is 0.210%, so that it could be seen as a special case between CGO Var. TH and CGO . Generally the content of rhoifolin in CGO Var. TH is obviously higher than those in CGO, which has proposed as one of the characteristics for identification

of CGO Var. TH and CGO in literature [3]. Therefore, it may reveal the reason why the similarity of CGO 2 is still high even if the information entropy is applied. Thereby, it is suggested that the sample of CGO 2 could be excluded from the similarity analysis.

TABLE II

RESULTS OF SIMILARITY BASED ON TRADITIONAL METHOD AND OPTIMIZED METHODS

Samples	Interangle cosine	Weighted by C.V	Weighted by Entropy	Correlation coefficient	Weighted by C.V	Weighted by Entropy
CGO Var.TH 1	0.9998	1.0000	0.9402	0.9998	0.9984	1.0000
CGO Var.TH 2	0.9999	0.9999	0.9618	0.9998	0.9996	0.9999
CGO Var.TH 3	0.9967	0.9995	0.9129	0.9968	0.9951	0.9995
CGO Var.TH 4	0.9991	0.9998	0.8997	0.999	0.9977	0.9998
CGO Var.TH 5	0.9994	0.9999	0.9467	0.9994	0.9987	0.9999
CGO Var.TH 6	0.9998	1.0000	0.9716	0.9999	0.9992	1.0000
CGO Var.TH 7	0.9998	1.0000	0.941	0.9998	0.9984	1.0000
CGO Var.TH 8	0.9992	0.9999	0.9263	0.9992	0.9981	0.9999
CGO Var.TH 9	0.9997	0.9999	0.9337	0.9997	0.9977	0.9999
CGO Var.TH 10	0.9991	0.9998	0.9397	0.9991	0.9973	0.9998
CGO Var.TH 11	0.9998	1.0000	0.9595	0.9998	0.999	1.0000
CGO Var.TH 12	0.9993	0.9999	0.8957	0.9994	0.9975	0.9999
CGO 1	0.9152	0.9979	0.2923	0.9092	0.1954	0.9969
CGO 2	0.999	0.9997	0.8531	0.9991	0.9981	0.9997
CGO 3	0.9914	0.9988	0.3194	0.9909	0.8824	0.9988
CGO 4	0.0066	0.0269	0.4601	-0.0516	-0.7629	-0.3209
CGO 5	0.9886	0.9987	0.358	0.9891	0.8778	0.9990
CGO 6	0.9957	0.9995	0.3954	0.9955	0.9412	0.9995
CGO 7	0.9918	0.9996	0.3981	0.9917	0.7952	0.9995
CGO 8	0.9861	0.9993	0.4232	0.986	0.7059	0.9992
CGO 9	0.9938	0.9996	0.4951	0.9937	0.8674	0.9996
CGO 10	0.992	0.9987	0.4511	0.9927	0.9216	0.9991
CGO 11	0.9814	0.9988	0.578	0.9816	0.8928	0.9990

In addition, these 23 samples of Exocarpium Citrus Grandis were used to build discrimination model base on principal component analysis and artificial neural network [8] .The result shows that CGO 2 and CGO 23 were mistaken as CGO Var. TH. However, only CGO 2 was misjudged in the process of calculation by correlation coefficient based on information entropy. It proves that CGO 2 could be seen as outlier and the similarity algorithm improved by information entropy is more suitable to identify varieties of herbs by chromatographic fingerprint. The results of comparison are listed in Tab. 3.

TABLE III

ACCURACY OF IDENTIFICATION OBTAINED BY DIFFERENT ANALYSIS METHODS [%]

BP network based on PCA	Correlation method based on information entropy	Interangle cosine method based on information entropy
91.30	95.65	100

Compared with BP network based on principal component analysis (PCA), the similarity algorithm improved by information entropy could offer more efficient discrimination .To a great extent, it owed that this method not only takes all the characteristic peaks in the fingerprint as the basis for discrimination, but also measure the influence degree of each peak in whole. Moreover, this method appears to be more direct, simple and rapid, for it doesn't need network's learning or training.

Summary

Published by Atlantis Press, Paris, France.

In the process of analysis of chromatographic fingerprints of TCM, it is often difficult to make an accurate identification or qualitative analysis by convention methods when confronted with some component which contains particularly high content but changes slightly. How to effectively reflect the degrees of influence of each component is the key to the solution. Thus, information entropy is introduced to the similarity algorithm. In this way, not only can all the peak data be taken into account but also can the weight of each peak be got conveniently and effectively. It gets perfect results in the identification of two different species of Exocarpium Citrus Grandis. This method also can extend to discrimination of different species and analysis of the quality of TCM.

Acknowledgements

The authors are thankful for the financial support provided by National Key Technology Support Program during the Twelfth Five-Year Plan Period by the Ministry of Science and Technology (2011BAI01B02).

Reference

[1] R. Tian, and P. Xie: Traditional Chinese Drug Research & Clinical Pharmacology Vol.17(2006), p. 40.

[2] S. Qiao, Z. Zhu, J. Wang, Y. Chai, and Y. Liu:Academic Journal of Second Military Medical University Vol.25(2004), p.1114.

[3] L. Lin, Z. Chen, X. Yuan, and X. Li:Journal of Guangzhou University of Traditional Chinese Medicine, Vol.2l(2004), p. 308.

[4] P. Xie,in: *Chromatographic Fingerprint of Herbal Medicine*, chapter, 6, People's Medical Publishing House (2005).

[5] R. Qiu,in: *Management decision-making study and application of entropy* , chapter, 1, China Machine Press(2002).

[6] X. Yuan: *Fingerprint and Quantification of Exocarpium Citrus Grandis*, unpublished.

[7] X. Yuan and L. Lin: Chinese Traditional and Herbal Drugs Vol.34(2003), p.764.

[8] H. Wei, L. Lin, Z. Huang, Y. Chen, and X. Yuan: Journal of Guangzhou University of Traditional Chinese Medicine Vol.28(2011), p.272.