A strategy for increasing the confidence of target gene/protein predictions

Chi-Ying Lee[1], Wen-Chang Chang[1], Kuan-Hsien Lin[1], Han Cheng[2], An-Sheng Cheng[1,*]

[1] Department of Medicinal Plant Development, Yupintang Traditional Chinese Medicine Foundation, Taiwan, R.O.C.

[2] School of Health Sciences, Purdue University, USA

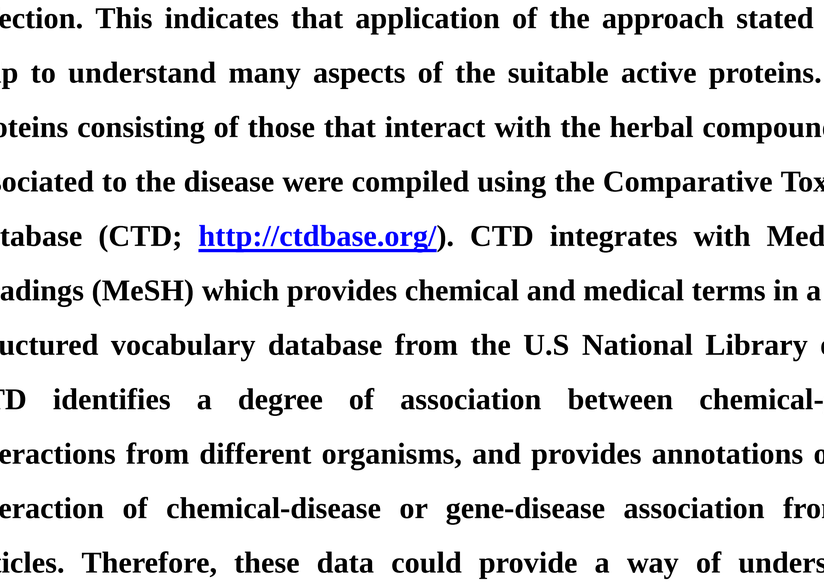
Abstract

In the post-genome era, the access of information is no longer a difficult task but rather the selection of the relevant information from vast amounts of data. Bioinformatics involves bringing biology, information science, statistics, biophysics and biochemistry to address the biological and medical complications. This is achieved through mass data collection, database establishment, integration, annotation and corresponded biological significance, in other words, using known data to solve unknown problems. Quality data collection, integration of analysis and problem-solving of strategies are important factors in the field of bioinformatics. The integration and collection of data relies on the database and facilitation of software, but difficulties arises in determining the analysis strategy. The present study aims to create a new analysis strategy for Traditional Chinese Medicine (TCM) to understand the role of gene/protein in disease mechanisms, through bibliometric analysis, interpretation of complex disease phenotype and major compounds of the herbs. In the Comparative Toxicogenomics Database (CTD), disease-protein relationships provide an Inference Score as a reference, but do not provide the screening criteria. We thus use the known disease target proteins to establish the selection criteria at 95% confidence interval, from 21724 liver cirrhosis related proteins we screened out 380 proteins. This strategy can help us identify the disease target protein of the herb in order to improve the effectiveness of Chinese herbal medicine, which to date has remained poorly researched and documented. Through the contributions of clinical foundations of TCM and various connecting applications of scientific research we aim to develop personalized care of TCM into modern medical applications.

Introduction

Studying the influence on the affected physiological processes would help to better understand the mechanism of actions of each protein in the body; while on the other hand, understanding the signal transduction pathways within the system would help to enhance the efficacy of the compound as well as reducing its side effects. The cause of a disease is commonly related to the abnormal intracellular regulation of the molecular mechanism, which in turn affects the cell function. In the case of a complex disease, it often involves the abnormal regulation of multiple genes and functional proteins. The strategy in developing new drugs is to target multiple disease-specific genes and proteins in order to interfere with the pathology network of a disease. Moreover, studies have also revealed that drugs with multiple targets demonstrate higher efficacy than single target drugs in treating tumor, DM, mental health illness and infection. This indicates that application of the approach stated above could help to understand many aspects of the suitable active proteins. The lists of proteins consisting of those that interact with the herbal compounds and those associated to the disease were compiled using the Comparative Toxicogenomics Database (CTD; http://ctdbase.org/). CTD integrates with Medical Subject Headings (MeSH) which provides chemical and medical terms in a hierarchical structured vocabulary database from the U.S National Library of Medicine. CTD identifies a degree of association between chemical-gene/protein interactions from different organisms, and provides annotations on molecular interaction of chemical-disease or gene-disease association from reference articles. Therefore, these data could provide a way of understanding the complex chemical-gene or chemical-protein interaction networks. However, CTD offers a mean of gathering information on the active gene/protein of disease medicines without confident strategy.

Method



Results

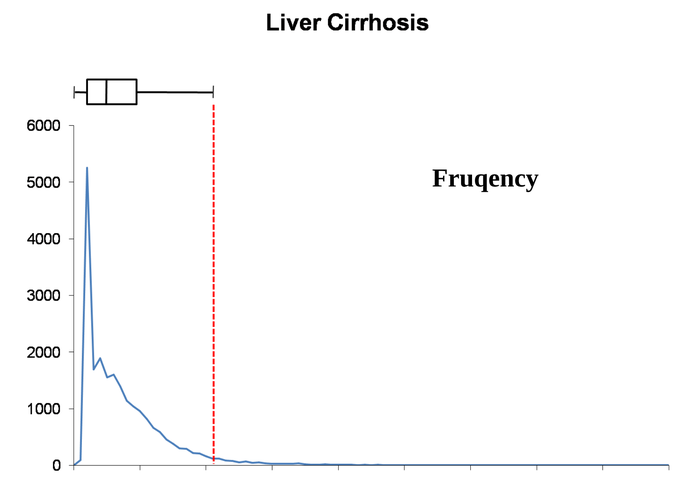


Figure 1. Interquartile range (IQR) of the liver cirrhosis-protein inference score distribution. The median is significantly different to mode, and the upper fence score is 41.885 (99.7% confidence interval), number of proteins is 898. The distribution is not an ideal model for inference score description.

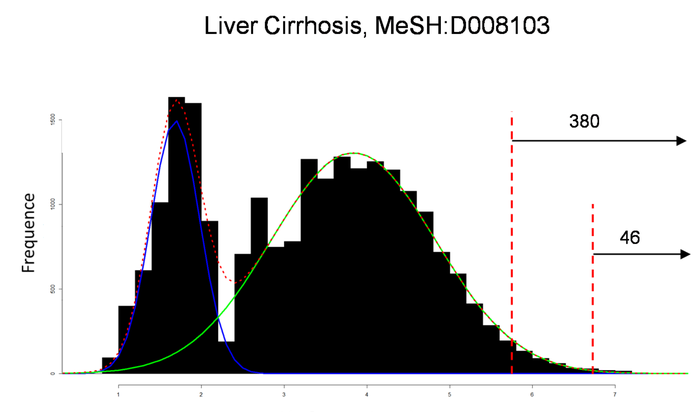


Figure 2. In different confidence interval, the significant activated proteins. Blue line is noise distribution, green line is real distribution and red line is mixing two normal distributions. In CI 95%, there are 380 proteins; in CI 99.7%, there are 46 proteins.

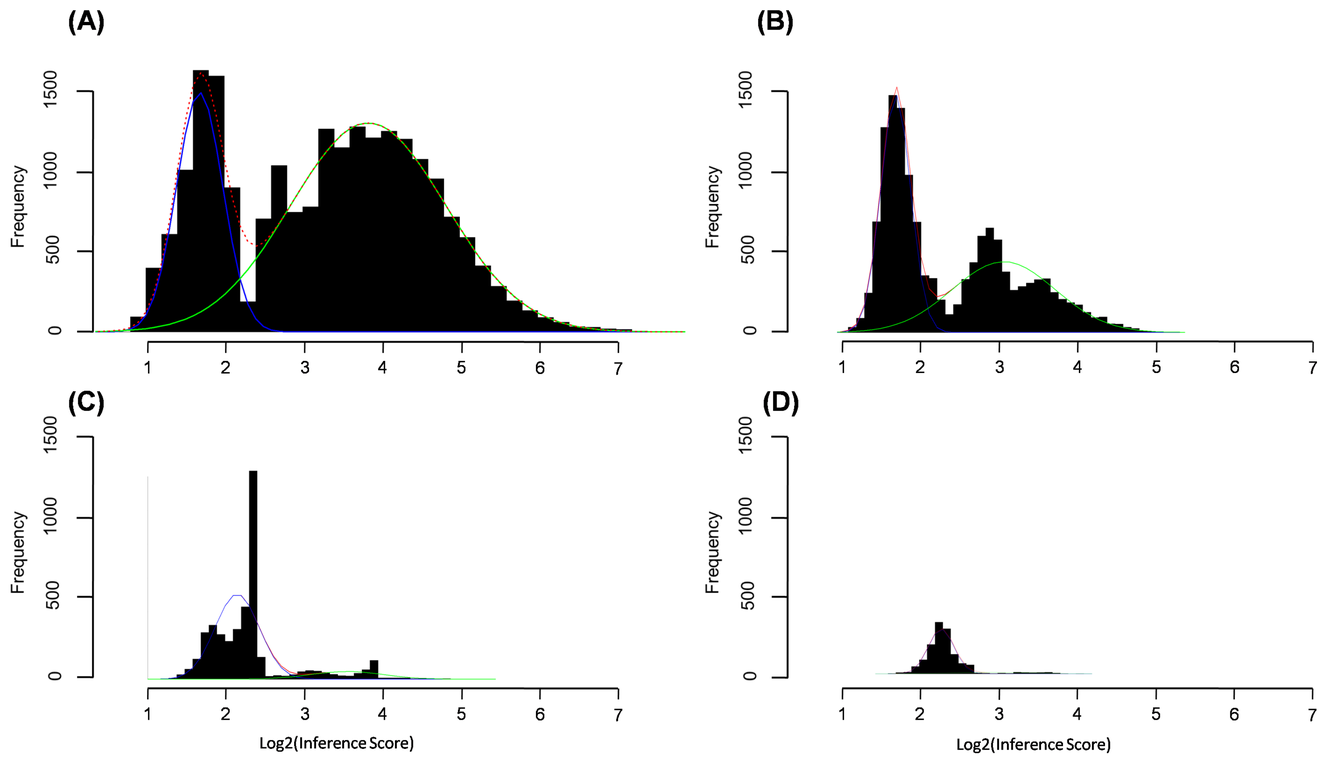


Figure 3. Disease-protein log2(inference score) distribution of different scales. (A) number of scale is about 20 k; (B) about 10 k; (C) about 7 k; (D) about 3k

Conclusion

We proposed a strategy to screen for the predicted disease associated proteins from the CTD that eliminates the noise in the process in order to avoid unnecessary data analysis. Moreover, by selecting the proteins that are significantly active or studied in considerable detail from the right distribution will assist us later in the analysis.