利用資訊擷取的方法來尋找生物醫學文獻中基因與疾病的關係

Discovering the gene and disease relationship from biomedical literature database by applying the information retrieval technology

中文摘要

目前生物科技發展日新月異,從過去對基因序列研究到目前基因功能性的研究,我們對人類的了解也越益進步. 此外,目前生物晶片的發明,與技術的發展與價格的下降,讓我們有機會一窺人類身上基因活動的全貌,也因此各種基因相關的資訊也隨之接踵而來.

此外,基因科技的發展對臨床醫師也帶來極大的衝擊,從過去對單一基因的研究,到目前數百到數千甚至數萬個基因研究,在質與量方面都帶來相當大的衝擊. 而 MEDLINE 是臨床醫師與生物醫學研究者最重要的文獻資料庫,總數超過一千萬篇的文章,帶來許多寶貴的資訊. 但是,動輒數以萬計的文章,與數以萬計的基因,對我們人類的認知能力也是極大的一個挑戰.因此,我們急欲利用一些資訊擷取的技術,來幫助我們消化數量龐大的生物文獻,而對臨床工作人員來說,基因與疾病的關係,是我們最迫切想要知道的訊息,所以本研究致力於應用資訊擷取技術來尋找文獻中基因與疾病的關係.

本研究主要針對如何建立基因與疾病的機率模型,我們建立了兩個機率模型,並比較其優缺點.

英文摘要

With the development of biomolecular technology, there is getting more and more information derived from genome research. Besides, the microarry was introduced to allow people study genome wide pattern of gene expression profile, the scientists have the opportunity to study the function of genes.

At the same time, the functional genomic research also bring a great impact to clinicians which usually study single gene or study disease at biochemistry level. In traditional, the MEDLINE always is the major resource for clinicians research. Recently, the explosion amount of the genomic related research bring for clinicians is too complicated to understand it. For examples, when talking about one disease, there are approximate over ten thousand of articles and hundred genes in it. It is almost impossible for clinicians to digest the knowledge. So it is urgent that there must be some computational tools developed to help clinicians observing the gene and disease relationship

In this research, we focus on constructing the probabilistic model of gene and disease relationship. By using two models to represent the knowledge from biomedical literature database, we can compare the two models in system performance and

precision.