行政院國家科學委員會補助專題研究計畫成果報告

※ ※※※※※※※※※※※※※※※※※※※※※※※※
※
※ ※
※ **與微陣列數據資料相關之生物資訊研究服務計劃** ※※
※
※ ※
※ ※※※※※※※※※※※※※※※※※※※※※※※※
※

計畫類別： ¦ **個別型計畫** æ**整合型計畫**

計畫編號：NSC －89－2316 －B － 002 －035

執行期間： 89**年** 09**月**01**日至** 90**年**07**月**31**日**

計畫主持人：張啟仁博士

執行單位：國立台灣大學醫學院臨床醫學研究所

中　華　民　國　90 年　10　月　30　日

行政院國家科學委員會專題研究計畫成果報告

計畫編號：NSC 89-2316-B-002-035
執行期限：89年09月01日至90年07月31日
主持人：張啟仁　執行機構及單位名稱 國立台灣大學醫學院臨床醫學研究所
計畫參與人員：劉孝斌　執行機構及單位名稱 國立台灣大學醫學院臨床醫學研究所

一、中文摘要

由於現代化實驗室工具的發展，提供了生物醫學科學家許多便捷的工具，使得他們能很輕易地接觸、取得及分析因網際網路的發達所取得的資料。在台灣或台大醫院的研究群中，因大型生物資訊資料庫能很容易地被取得、納入、分享並迅速地分析，那麼生物醫學上的新發現即能很快地散佈於所有研究者之中。因此，許多費時、費錢的實驗，也得以免除重覆執行的命運。為了建立一個良好的「研究服務」團隊，我們提出了此包括臨床醫師、實驗室生物科學家及生物統計學家為一合作群體的生物資訊科學研究計劃。

台灣大學醫學院附屬醫院臨床醫學研究部已經成立了新的「微陣列」方法的研究服務，透過基因微陣列方法的幫助，基因資料可以在很短的時間之內大量製造出來。因此，資料的整理、除錯、建檔及分析的需求亦相對地增加，臨床研究者對此資料庫的依賴與需求增加甚多。依過去的經驗，資料取得之後的建檔、資料處理及統計分析都是由計劃主持人及其助理擔任，這不僅浪費了許多寶貴的人力與時間，並且所處理之後的結果往往也比不上專業的資料處理人員及統計學家所做的準確。有鑑於此，冀望透過此一研究計劃，結合對基因微陣列方法與疾病研究有興趣之臨床醫師，配合基因微陣列方法的專家，再加上資料庫處理人員與生物統計學家的整合，成立台大醫院生物資訊科學研究群，期能使台灣的生物資訊學能有大幅的進步以邁入世界領導地位。

關鍵詞： 研究服務， 微陣列，生物統計，生物資訊科學。

Abstract

Modern laboratory tools, computer technology advances and ubiquity of the internet offer unprecedented opportunity for scientists to gain access to, share, and analyze critical data and information stored in databases over the cyberspace. Scientific discovery can be expedited and many wastefully and costly experiments can be avoided if the vast information could be stored, shared, analyzed and opened to the research scientists in any clinical research institute. One of the successful collaboration examples among physicians, laboratory scientists, and biostatisticians has been established within National Taiwan University Hospital Research Group from interested physicians led by Dr. PC Yang, Chairman and Professor of Internal Medicine, National Taiwan University Hospital (NTUH) has successfully implemented their own research topics in collaboration with, Dr. Jeremy Chen, a Microarray specialist and further collaborated with this PI, Dr. CJ Chang. The three roles has been worked as a research group within NTUH.

Recent development of Microarray technology has enabled research scientists approaching their own area into a new era. With the collaboration from scientists in generating Microarray data, investigators can look into the research problems from a vast point of view, i.e. from a large data generated by array machine. However, due to the fast growing technique and astonishing data output, data analysis of this new Bioinformatic became an important issue in biomedical research.

Understanding what and how the

Bioinformatics can provide has given clinical physicians in medical center a starting point to reconsider an infrastructure of Bioinformatics facility as a research service resource. A Microarray facility center has been established under the direction of Dr. Jeremy Chen in which research services and education in using Microarray machine and technique were rapidly provided within NTUH. In the meantime, statistical data analysis support for Microarray data in Bioinformatic area has also been provided in NTUH research campus. A NSC supported grant 斉 ioinformatics Research Services Facility Using Microarray Data (NSC89-2316-B-002-035)" has provided us a good starting point to setup the connection among Microarray data analysts, Microarray data generators, and research investigators. A paper utilized this research collaborating pathway has been finished and submitted for publication.

Huge data sets can easily be generated from the fast speed machine, and thus the demands of data collection, clearing, selection, and data management in the database is therefore strongly needed. However, the ability to tackle such problems can only be made and solved by limited attached programs from the array management software such as Genecluster from MIT. No formal statistical consideration of the huge data set has been considered from statistical point of view. They were all performed automatically from the software by individual PIs and their assistants. To our experience, it is considered as a waste of manpower and valuable time. Thus, we had put all the specialties together to form a research team including investigators who provide the hypothesis and samples from their patients for Microarray analysis; the Microarray experts who provide the genetic array analysis; and biostatistician who provides the database management and data analysis since last year in NTUH.

However, even with a good starting point of the statistical support to the data generated from Microarray research. There still exist some potential problems in handling the data from the early data management stage to the later statistical modeling and analysis stage. Therefore, this proposal is hereby proposed due to the strong research demands in Microarray data analyses and the fact of lacking good statistical methods to the array data. This 3 years grant proposal will include four major research topics in studying gene expression generated from Microarray machine concurrently; they are 1)data selection; 2)data filtering; 3)cluster analysis; and 4)discriminant classification analysis. Our goals are to study, develop practical and advanced statistical data mining methods for Microarray data, especially when the bridge connecting the array data generated from laboratory and data analyst using advanced computer technology have been well established. This article provides guidance for report writing under the Grant of National Science Council beginning from fiscal year 1998.

Keywords: Bioinformatics, Microarray, Cluster Analysis, Discriminant Analysis National Science Council.

2. Objective and Goal

Modern laboratory tools, computer technology advances and ubiquity of the Internet offer unprecedented opportunity for scientists to gain access to, share, and analyze critical data and information stored in databases over the cyberspace. Scientific discovery can be expedited and many wastefully and costly experiments can be avoided if the vast information could be stored, shared, analyzed and opened to the research scientists in any clinical research institute such as National Taiwan University Hospital. In order to establish such research services facility and make a good collaborative example among clinical physicians, laboratory scientists, and biostatistical data analyst, We had established a collaborating research group within National Taiwan University Hospital that include the Microarray laboratory, and Bioinformatic supporting groups.

Some US government research institutions such as General Clinical Research Centers (GCRCs) established by the National Center for Research Resources (NCRR) of NIH in USA, are leading contributors to biomedical clinical research[1]. Many infrastructures of Bioinformatics workshops, discussions and seminars had been held during last couple years in USA under the recommendations from National Science and Technology Council 旧 Subcommittee on Biotechnology. In the meantime, a national Bioinformatics conference organized by GCRC institutions was firstly held at San Antonio, Texas in Nov. 1999; and then at New Orleans, Louisiana in October 2000. During both meetings, conclusions and executive summary of definition in Bioinformatics and how it is promoted and utilized by clinical research institution such as medical center is made.

Specific definition of the Informatics and Bioinformatics were discussed, although the conference was not meant to make a clear distinguish in between, suggestions in defining Informatics and Bioinformatics were indicated as following:

Informatics is the application of information science and technology to the conduct of clinical research.

Bioinformatics is the science of data collection and analysis using mathematics, statistics, computer science and knowledge of the biology of the system under study to reach significant and relevant conclusions.

Research Services

Understanding the definition of the Bioinformatics has given clinical physicians in medical center a starting point to reconsider an infrastructure of Bioinformatics core as a 遲esearch service" resource. As a matter of fact, due to the strong interest demands and lack of supports from specialties in the area of Biostatistics and data coordinating facilities to the clinical research physicians in NTUH. A formal proposal of 遲esearch services" named 斉 iostatistical and Data Coordinating Center" has been proposed and presented within NTUH research environment. However, the strong impact of data generated from genetic researches have also enabled the research scientists searching for a core facility in which provides data generating, data managing, database construction and statistical analysis.

Common questions raised by the physicians doing genomic related research are how they utilize the laboratory and biostatistical facilities to collaborating with their own interests.

The fact of the new biotech, chip-based Microarray technology allowing high speed, high capacity analysis of gene expression has given them a new direction in conducting clinical researches[2-5]. Huge data derived from Microarray devices have to be stored in an efficient manner through the high speed and well organized computer database system. Complicated statistical data analysis utilizing the data generated from Microarray and clinical outcome has to be reconsidered and reevaluated in a more efficient manner. Thus, to promote the newly developed Bioinformatics concepts in a clinical setting such as in medical center has to be through the exhausted collaborating research services.

In fact, A Microarray facility center has been established under the direction of Dr. Jeremy Chen. In the meantime, statistical data analysis support for Microarray data in Bioinformatic area has also been provided in NTUH research campus. A NSC supported grant 斉 ioinformatics Research Services Facility Using Microarray Data (NSC89-2316-B-002-035)" has provided us a good starting point to setup the connection among Microarray data analysts, Microarray data generators, and research investigators.

Concepts

In last few years, Bioinformatics has been developed, as the discipline required

implementing the integration of the component parts into a cohesive study environment. It is recognized as the 昼cientific processing of biomedical data." Thus bioinformatics has coupled a variety of methods, techniques, and computer approaches not previously used in experimental environments.

Nowadays, in the fields ranging from structural biology to genomic to biomedical imaging, large data will be quickly generated and available for research use. These all due to the fast biomedical chip-based technologies such as Microarray. These powerful techniques, developed less than six years ago, allow high-speed, high-capacity analysis of gene expression.

To accommodate such huge data generated from the fast speed chip, and to transform these data into information and to analyze for future valuable knowledge. Data management and Internet computer technology plus the long existing mathematical and biostatistical tools have to be merged as a unit as a research service facility. The advanced statistical methods to handle large Microarray data become very important issues.

For a standard approach of Microarray data analysis, huge data sets can easily be generated from this fast speed Microarray machine, and thus the demands of data collection, data clearing and data management in the database is therefore strongly needed. However, the ability to tackle such problem can only be made and solved by limited attached programs from the array management software. No formal statistical consideration of the huge data set has been considered from statistical experts. They were all performed automatically from the software by individual PIs and their assistants. To our experience, it is considered as a waste of manpower and valuable time. Thus, we had put all the experts together to form a research team including investigators who provide the hypothesis and samples from their patients for Microarray analysis; the Microarray experts who provide the genetic array analysis; and biostatistician who provides the database management and data analysis since last year in NTUH.

However, even with a good starting point of the statistical support to the data generated from Microarray research. There exist some potential problems in handling the data from the early data management stage to the later statistical modeling stage. Therefore, this proposal is hereby proposed due to the strong research demands in Microarray data analyses and the fact of lacking good statistical methods to the array data. To mention some approaches, few background controlling methods has been proposed elsewhere; some cluster analysis, discriminant analysis, and classification model have also been developed in identifying underlying patterns in complex data[6-9]. The techniques are essentially different, such as cluster analysis is to cluster points in multidimensional space if the outcome is not clear defined; discriminat, classification analysis are to identify the pattern form different time points or different type of treatments if outcome is defined. We will focus in these theoretical and practical statistical issues within this proposal.

Goals

Methods and systems supporting Microarray data analysis can be adapted to the challenge from each investigator within NTUH and possibly by their basic science collaborators. Statistical collaboration, data coordinating, and data manipulating can be utilized as good research tools in the molecular biological related disease researches. Without a more thoroughly discover of the statistical methods, data management approaches and systematical analysis plan of the array data, one can not provide and propose good results in biomedical research.

We like to use this proposed grant to review the existing data management approaches, available existing statistical

methods, to further discuss some potential revisions, and to develop new statistical approaches. The goals are:

1. To set up a mechanism bridging the data collecting from individual PIs' data derived in Dr. Jeremy Chen旧 Microarray laboratory and develops a standard operating procedure storing the valuable data into database.
2. To review the existing data management approaches and statistical methods appended to Microarray data analysis software. This will include GeneCluster, Cluster and GeneSpring.
3. To research practical issues in the data clearing, selection and data filtering procedures.
4. To study the current available statistical methods applicable to the data generated from Microarray laboratory. This will include cluster analysis and discriminant classification analysis. Four major research aims in this three year grant proposal are i): data selection; ii): data filtering; iii): cluster analysis; and iv): discriminant analysis.

As of now, the statistical analysis approach to the Microarray data is a very new and under strong demand applied science. It is also considered as one major part in newly established Bioinformatics topic in Taiwan. We feel obligated in participating this research because of some practical reasons. Within NTUH, there are strong needs from cancer research scientist utilizing Microarray technique. The Microarray research service facility was just established within the campus. All the research scientists and Microarray expert have the abilities and credits in performing this advanced technology. The PI, Dr. CJ Chang, has more than 10 years experiences performing statistical and Bioinformatics related consultation previously in USA and recently in Taiwan. Dr. Jeremy Chen, is the most experienced person in performing Microarray machine in Taiwan. With his USA paten of using Microarray technique, the ability will certainly help the research scientists in NTUH and in Taiwan. Owing to the collaboration in between and with the support from this grant, we expect to develop practical statistical methods to Microarray data

3. Results

Improvement of Microarray technique makes scientists could explore thousands genomes in a time and this technique shorten much time in biomedical research. Investigators could get exciting result from Microarray data analysis, however, once we can not control and reduce influences of backgroud noise, the decision we make might seriously not be able to reproduce again. The stability of data greatly affects the conclusion of cluster analysis. We focus concentration on controlling of background noise. The project is not only doing research for Microarray data analysis but offering Microarray consultation to investigators who need data management and statistical analysis. Since the project started, we have been involved in several PIs' Microarray studies. For instance, Drs. **楊洋池，許世明，楊志新**, and **陳建尉** within NTUH. All the studies are related to cancer research and all are in the process of either laboratory work or report writing. In all cases, the clients have successfully explored some potential genes in their particular topics.

In the data management and data filtering before further Microarray statistical analysis. We have also made some progress in collaborating with Drs. **范盛娟**, **張久媛** in IBMS, Academia Sinica, and Dr. **連怡斌** in **彰化師範大學**.

When controlling background noise, we firstly evaluated the size and distribution of background noise. Secondly, we evaluate the consistency of repeated Microarray data by using the formula A=B + K**脫**oise. A is the first experiment baseline data and B is second experiment baseline data. Noise follows normal distribution, and we make different value of K. We tried to correlate two experiment data into a regression model.

After our careful evaluation, we found that even under the same assumptions, we could not find a fixed distribution for background noise. It varies with different model considered and different conditions of array experiments. What we need is to figure out the distribution of K. Only we know it□ accuracy then we can control variation of Microarray data and conclude an effective result.

The methodologies of data filtering became a wide discussion topic by many scientists including our research team. We have successfully compared and implemented many statistical approaches to the data filtering procedures utilizing SAM, Cook□ distance, and ACE procedures to the normal vs. normal; normal vs. tumor cells in the array analysis. A manuscript in comparing data filtering methods for cDNA microarray data is under preparation.

4. Discussion

How to evaluate accuracy of a microarray experiment is important. This will subject to the data derived from the experiments. Whether the data collected from this experiment will correctly be used needs lots of attention in the quality itself. We tried and evaluated many approaches under different situations when the data is subject to some errors. It appears to us that no matter how good the statistical methods are, we can 急 derive good results from data without a quality experiment.

5. Assessment

This grant proposal has given us a good start in 1): collaborating with research scientists in cancer research; 2): collaborating with laboratory scientists in microarray experiments; 3): the most importantly, collaborating with biostatistician in proposing better and effective methods in handling the data quality issues in microarray data. In conjunction with the current on going grant, one manuscript is under preparation and more findings and discussion is in the process.

6. References

[1] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., Cluster analysis and display of genome-wide expression pattern. Proceedings of the National Academy of Sciences of the United States of America. 95, 14863-8.

[2] Lee, M.L., Kuo, F.C., Whitemore G.A., and Sklar, J., Importance of replication in microarray gene expression studies: statistical and evidence from repetitive cDNA hybridizations. Proceedings of the National Academy of Sciences of the United States of America. 97,9834-9.

[3] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentitation. Proc. National Academy Science of the United States America . 99, 2907-2912.