行政院國家科學委員會專題研究計畫　成果報告

生物資訊分析工具之測試平臺

計畫類別：個別型計畫
計畫編號：NSC92-2213-E-009-098-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：國立交通大學資訊科學學系

計畫主持人：胡毓志

計畫參與人員：林婉嫻　王美華　賴昀君　吳秉蔚

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 93 年 8 月 26 日

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　　　　　生物資訊分析工具之測試平台　　　　　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：個別型計畫
計畫編號：NSC 92－2213－E－009 －098－
執行期間：92 年 8 月 1 日至　93 年 7 月 31 日

計畫主持人：國立交通大學資訊科學系 胡毓志

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
■出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

執行單位：交通大學資訊科學系

中　華　民　國　93　年　8　月　23　日

行政院國家科學委員會專題研究計畫成果報告
國科會專題研究計畫成果報告撰寫格式說明
Preparation of NSC Project Reports

計畫編號：NSC 92-2213-E-009 -098-
執行期限：92 年 8 月 1 日至 93 年 7 月 31 日
主持人：胡毓志　交通大學資訊科學系
計畫參與人員：林婉嫻 王美華 賴昀君 吳秉蔚 交通大學資訊科學系

一、中文摘要

在眾多不同的基因組定序計劃執行之後，生物序列等相關資料業已被大量產生，若再加上目前生物檢測技術的日新月異，各式各樣的生物資料量早已超越了過去生物學的傳統資料分析方法，雖然我們已能夠迅速有效地產生製造有關的原始生物資料，然而，我們對資料的瞭解卻仍然相當欠缺，我們需要好的分析工具幫助我們儘速破解生物資料之謎。

縱使今日不管是學術界或產業界都竭盡心力發展新的分析方法，但是，現今生物資訊分析工具的評估與檢測仍是過於草率且欠缺標準，藉以實驗檢測工具效能的資料大多由工具研發者自行蒐集與挑選，加上沒有完整詳實的說明，其他人根本無從重覆實驗以確定其真正的實用價值，例如，僅僅簡述待測基因的名字，公共資料庫的網站等等，這些都無法提供有用的訊息以供他人再利用，此外，原始實驗資料的前、後置處理，例如，正規化(normalization)，離散化(discretization)，以及資料擷取(extraction)等，都會影響實驗的結果。沒有確切的資料，詳實的資料說明，任何錯誤的假設，造成資料準備的不確實，都將會造成不一致的實驗結論。而建構一個新的生物資訊實驗與研究測試平臺之目的，即是在提供一個標準的測試環境，所有的使用者將可以使用相同的資料，完備的準備程序，以維持實驗的客觀性與一致性。

我們建構一個以網際網路為基礎的生物資訊測試平臺，它將涵蓋主要的生物資料種類與分析內容，我們預期它能成為一個客觀評估生物資訊分析方法的公共環境，藉以凝聚來自不同研究領域的心血，共同為生物資訊的發展做最大的努力。

關鍵詞：基因調控、基因家族、調控訊號

Abstract

Multiple various genome projects have produced an explosive amount of biosequence data. In addition, significant improvements and new developments of biology-related instrumentation have also generated a wide variety of biological data for further study. It requires tremendous efforts from various research fields to finally achieve this. Though the data could now be generated effectively and efficiently, it seems that our biological knowledge has not been able to increase in the same pace of the growth of biological data. This imbalance has consequently stimulated the development of many different analysis tools to bridge the gap.

Despite that academia and industry have scaled up their data generating activities, and significant efforts have been made on the development of computational tools, so far bioinformatic data analysis research has been relatively limited and rather *ad hoc* in terms of the data in use. No matter what the data source is, a clear specification of the format and content of the data is crucial to any following study. Unfortunately, this type of descriptions is missed in most research papers. Note that information such as the names of the genes/ORFs used in experiments, references about who previously used the same data, vague descriptions of how the data is prepared, and so on, is not adequate for a precise reproduction of the experiments. Using the data in any other fashion than the way it was originally applied can lead to discrepancy in experiments.

We have constructed an online bioinformatics test bed. Based on a clean definition of the data sets and their availability, this test bed can serve as a bridge to help attract more efforts from other research fields, and thus stimulate further progress in bioinformatics. In addition, the test bed can allow researchers to replicate experiments and studies as well as a common environment for exploratory research in bioinformatics by gathering more challenging problems in different application domains and motivating more researchers of various interests. It will accelerate the development of new techniques to extract and understand hidden information in the data.

Keywords: 基因組，生物資訊，測試平臺

Introduction

Multiple various genome projects have produced an explosive amount of biosequence data. In addition, significant improvements and new developments of biology-related instrumentation have also generated a wide variety of biological data for further study, e.g., protein structures, protein-protein interactions, gene expression levels, etc. (Lo Conte et al. 1999; Marcotte et al., 1999; DeRisi et al., 1997; Wodicka et al. 1997) It requires tremendous efforts from various research fields to finally

achieve this. Though the data could now be generated effectively and efficiently, it seems that our biological knowledge has not been able to increase in the same pace of the growth of biological data. This imbalance has consequently stimulated the development of many different analysis tools to bridge the gap. We can foresee that once the Human Genome Project is complete, the analysis tasks will be even more demanding than ever. After all, the real success of this grand project is determined by the degree to which we solve the most difficult puzzle ever, i.e., to understand every bit of the information hidden in this god-sent book, the human genome. The only way to get us closer to the goal is to develop better analysis algorithms. Despite that academia and industry have scaled up their data generating activities, and significant efforts have been made on the development of computational tools, so far bioinformatics data analysis research has been relatively limited and rather ad hoc in terms of the data in use. The data to be analyzed are typically derived in two ways. First, the data may be directly generated from some biological laboratories. They produce and analyze the data of their own interest. Second, the data may have been studied and published in literatures, and have been stored in some public databases for free access. However, no matter what the data source is, a clear specification of the format and content of the data is crucial to any following study. Unfortunately, this type of descriptions is missed in most research papers. Note that information such as the names of the genes/ORFs used in experiments, references about who previously used the same data, vague descriptions of how the data is prepared, and so on, is not adequate for a precise reproduction of the experiments. Using the data in any other fashion than the way it was originally applied can lead to discrepancy in experiments, e.g., applying different physicochemical properties in predicting protein functions may get different classifications (King et al., 2001). Publishing conclusions without providing access to the data that support those conclusions is no science (States, 2001).

We have constructed an online data archive of biological data sets, and expect it to play the following roles. First, based on a clean definition of the data sets and their availability, this archive can serve as a bridge to help attract more efforts from other research fields, and thus stimulate further progress in bioinformatics. Second, the archive can serve as a test bed that allows researchers to replicate experiments and studies. It also makes possible the quantitative comparisons between computational methods. Researchers not only have access to the exact data previously used, but also have full knowledge of how to correctly reuse them in experiments in order to avoid the undesirable biases and keep the required consistency. Finally, this archive will also serve as a common environment for exploratory research in bioinformatics by gathering more challenging problems in different application domains and motivating more researchers of various interests. It will accelerate the development of new techniques to extract and understand hidden information in the data.

Background

One can view science as a search through a space of theories which requires two basic components, a generator and a test (Kibler and Langley, 1990). A generator produces new theories, and a test yields information regarding the quality of the generated theories. As science puts its emphasis on observations, the purpose of the repeated generate-and-test procedure is to identify an ideal theory. A good theory is one that must not only accurately describe a large class of observations, but also make definite predictions about the results of future observations (Hawking, 1988). Normally we can consider any bioinformatics analysis tool as a theory generator, thus its result (i.e. its output) becomes a theory that is expected to accurately describe a significant number of observations, i.e., the data provided. To evaluate a theory, there is always a temptation to emphasize formal approaches. However, given the fact that many computational tools are too complex for formal analysis, and that biological behaviors are full of variables, empirical studies of these computational tools/algorithms must retain a central role. Through experimentation we can test the performance of the tools and also better understand under what conditions they perform the best. Just like the literature of the machine learning community in the early days prior to the creation of its data archives, most current bioinformatic research papers lack a clear description of the data used when introducing new analysis algorithms or tools. For example, some authors only cited the database where they retrieved the data, but they did not explain whether or how the data was actually processed and represented for later experiments. Note that even if the data is gatherer from the same source, it can be later described by different representations. As the expressiveness of

representations may not be the same, the data in use can carry different information, and consequently influence the final experimental results (Hu, 1998). Under such circumstances, it is very difficult for readers of interest to replicate the experiments, or further improve the algorithms. This will definitely hamper the progress of bioinformatics because researchers have no access to the correct data. The primary purpose of building a data archive is thus to provide the exact data used before, so researchers in all fields can use the same data as a test bed to evaluate existing or new algorithms.

NCTU BioInfo Archive

The creation of the NCTU BioInfo Archive is greatly motivated by the Machine Learning Data Repository and the KDD Archive both maintained by the Information and Computer Science Department of the University of California at Irvine. They are widely used by industrial and academic researchers, and have thus become the most frequently cited benchmark for empirical evaluation of new and existing learning and mining algorithms (Bay et al., 2000). Like these data archives, we envision for this new bioinformatics data archive to serve as a common playground for researchers to test algorithms of interest, and help gain insight into a wider class of problems. Adopting the design philosophy of the KDD Archive at UCI, the preliminary goal of the NCTU BioInfo Archive is to store the data sets ranging over a wide variety of data types and problem tasks related to bioinformatics. All the data sets are characterized and stored according to its data type and the associated analysis task with the aim of ensuring the precise reference and easy access for the users. In the following subsections, we will explain the data types, present the analysis tasks, and finally introduce the documentation. Unlike others, we further divide the data type into two classes, syntactic and biological. We use the syntactic data type to address the way the data is presented, and use the biological data type to indicate the biological meanings the data is related to.

We define the syntactic data type as the underlying representation of data, including the attributes used to describe the data and the structure of the attributes. This data type is only concerned about what the data look like rather than the implied biological meanings. For those researchers from other communities than biological sciences, without the burden of biology jargons, the syntactic data type is sufficient for conducting analysis of these data. Take a DNA sequence data set represented as a FASTA file for example. The syntactic data type shows that each instance in the data set is described as a series of symbolic attributes with four possible values (i.e. A, G, C and T). This information is adequate for a computer scientist or a statistician to apply an appropriate method to DNA motif prediction. Any domain knowledge about the nucleotides or the FASTA format is not needed. The key point of separating syntactic from biological is to attract more efforts from other research fields which may otherwise be discouraged by those unnecessary biological terms.

The amount of new types of biological data has been increasing as the advent of new technologies. This not only provides more data to work on, but also opens new doors for research. For instance, prior to the existence of microarrays and biochips, there was little study of gene expressions on a genomic scale. In order to accommodate a wide variety of data, we also organize the data sets based on their basic biological meanings. Though there are many different criteria by which we can partition the data, we currently characterize the data into the following three simple types. We notice that tremendous efforts have been put into the development of bioinformatics ontology. An ontological description of data certainly conveys more information than our current biological types. Nevertheless, the ontologies used within the community to provide knowledge are very different and specific to their intended use (Ashburner et al., 2000; Pouliot et al., 2001; Baker et al., 1999; Chen et al., 1997). For example, TaO (Baker et al., 1998) is an ontology of bioinformatics tasks and thus includes concepts such as ProteinId and AccessionNumber, which are not really part of molecular biology. It cannot be substituted for EcoCyc's (Karp and Paley, 1996) ontology, which was specifically designed to cover E. coli. genes, metabolism, regulation and signal transduction. Since at the present there is no ontology available that is capable of covering the whole of molecular biology and bioinformatics tasks, we do not incorporated any ontological terms in the data types. Simple biological data types like the following are enough for distinguishing between different biological data without compromising the aim of our data archive as a general data resource for research and experimentation.

Sequence Data: It is the most basic data type, including DNA sequences, RNA sequences

and protein sequences. For example, a sequence data set may consist of the regulatory regions of a gene family (van Helden et al., 1998).

Gene Expression Data: It is probably the most widely discussed data type these days. According to how the data is generated and collected, it may convey different biological meanings. For a specific gene, it could be the expression level change over time as represented by a single time series if recorded continually at different time points in a single experiment (DeRisi et al., 1997). On the other hand, the expression data may be gathered from various experiments for each gene as a concatenation of different and independent mRNA expression measurements (Brown et al., 1999).

Physicochemical Data: This may include all kinds of physicochemical properties that can be used to present biological meanings or functions of biological activities. For example, researchers used sequential hydrophilicity profiles (Hopp and Woods, 1981), surface tension and charge (Lo Conte et al., 1999) to characterize proteinprotein interactions.

The complexity of biological process in life has generated a wide variety of problem tasks of interest. Thus besides the data type, we also organize the data based on its analysis task. We roughly divide the problem tasks into four categories, and for each category, we present the related studies and potential research.

Classification: Classification is a task that involves an accurate prediction of the value of a categorical variable typically named “class”. Many problems in bioinformatics belong to this task such as protein functional class prediction, protein structure prediction, protein interaction prediction, and so on. Given a previously unseen data item, the goal of classification is to assign the most appropriate class to it.

Regression: Similar to classification, regression is also a prediction of the value of a target variable. However unlike classification the target variable in regression is continuous instead of categorical. Some possible applications of regression include predicting the protein structure in terms of its 3D coordinates, the gene expression level at a particular time point, and so on.

Clustering: The goal of clustering is to partition the data into meaningful groups based on some predefined measure, and it usually provides a basis for bootstrapping further analysis process of the data. For example, based on the assumption that genes falling into the same expression cluster are likely to share functions, clusters of coexpressed genes render the basis for further study of transcription coregulation.

Pattern Discovery: The rationale behind pattern discovery is to identify the implied regularity in the data. With this regularity as potential data characteristics, it is easier to carry out further analysis of the existing and the new data. Besides, the patterns found can be used as new attributes to transform the original data representation (Hu et al., 2000). The typical research related to pattern discovery is common motif detection in biosequences. These motifs are believed to reveal specific biological meanings such as protein structures, functions, regulation of transcription and translation, etc.

The success of an archive depends on not only the availability of data, but also on the good documentation. In the archive, each data set is associated with a documentation file that provides other information than the data itself to increase the usefulness of the data. The documentation file contains the following information when available.

Data Description: This is the major part of the documentation file. It includes the information of syntactic and biological data types of the data set, to what application area the data belongs, from what organism the data set was derived, etc.

Goals: This describes the goals of the experiments in which the data set is used. To be more precise, the goal specifies the problem task, e.g., to find some target motifs in the sequence data set.

Previous Experimental Results: It is necessary to have the correct previous experimental results in order to carry out a fair comparative study. If available, we keep the related previous results for researchers to perform further experiments and studies.

Literature References: In addition to the experimental results, researchers may be interested in other useful information, such as the biological background of the data, the reasons for the data generation and its representation as well as the validation and suggested future work.

Data Preprocess and Postprocess: Data preprocess means how the data is prepared for and used by the tools/algorithms. The data may be different from the raw data originally generated from a biological laboratory after noise filtering and normalization. On the other hand, data postprocess is done on the output of the tool/algorithm. The original data output from the algorithm may require modification before it can

be used to draw the final conclusion. As these processes add biases (Mitchell, 1997), they can significantly affect the results. We include the information of data processes for later experiments to keep the consistency.

Donors: The contributors are recorded for reference.

The NCTU BioInfo Archive currently contains more than 40 data sets. To ensure the validity of data, we only collect those data sets that have been published or maintained in wellknown public databases.

五、參考文獻

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000) “Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium”, Nature Genetics, Vol 25, p25-29.
2. Baker, P.G., Goble, C., Bechhofer, S., Paton, N., Stevens, R. and Brass, A. (1999) “An ontology for bioinformatics applications”, Bioinformatics, Vol 15, p510-520.
3. Baker, P.G., Brass, A., Bechhofer, S., Goble, C., Paton, N. and Stevens, R. (1998) “TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview”, Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology, p25-34.
4. Bay, S., Kibler, D., Pazzani, M. and Smyth, P. (2000) “The UCI KDD Archive of Large Data Sets for Data Mining Research and Experimentation”, ACM SIGKDD, 2, p14-18.
5. Brown, M.,Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M. and Haussler, D. (1999) ”Knowledgebased analysis of microarray gene expression data using support vector machines”, Proceedings of the National Academy of Science, 97(1), p262-267 .
6. Chen, R.O., Felciano, R. and Altman, R.B. (1997) “RIBOWEB: Linking structural computations to a knowledge base of published experimental data”, Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology, p84-87.
7. DeRisi, J., Iyer, V. and Brown, P. (1997) “Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale”, Science, 278, p680-696.
8. Hawking, S. (1988) “A Brief History of Time”, New York: Bantam Books. Hopp, T. P. and Woods, K. R. (1981) “Prediction of Protein Antigenic Determinants from Amino Acid Sequences”, Proc. Natl. Acad. Sci. USA, 78, p3824-3828.
9. Hu, Y. (1998) “Constructive Induction: Covering Attribute Spectrum”, in Feature Extraction, Construction and Selection: A Data Mining Perspective, H. Liu and H. Motoda, eds. (Kluwer Academic Publisher), p257-272.
10. Hu, Y., Sandmeyer, S., McLaughlin, C. and Kibler, D. (2000) “Combinatorial Motif Analysis and Hypothesis Generation on a Genomic Scale”, Bioinformatics, Vol 16, 3, p222-232.
11. Karp, P. and Paley, S. (1996) “Integrated Access to Metabolic and Genomic Data”, Journal of Computational Biology, Vol 3, 1, p191-212.
12. Kibler, D. and Langley, P. (1990) “Machine Learning as an Experimental Science”, in Readings in Machine Learning, Tom Dierteric and Jude Shavlik, eds. Morgan Kaufmann, p38-43, 1990.
13. King, R. D., Karwath, A. Clare, A. and Dehaspe, L. (2001) “The Utility of Di_erent Representations of Protein Sequence for Predicting Functional Class”, Bioinformatics, 17, p445-454.
14. Lo Conte, L., Chothia, C. and Janin, J. (1999) “The Atomic Structure of Protein-protein Recognition Sites”, J. Mol. Biol., 285, p2177-2198.
15. Marcotte, E., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O. and Eisenberg, D. (1999) “Detecting Protein Function and Protein-protein Interactions from Genome Sequences”, Science, 285, p751-753.
16. Mitchell, T. (1997) Machine Learning, McGraw-Hill, New York. Pouliot, Y., Gao, J., Su, Q.J., Liu, G.G and Ling, X.B. (2001) “DIAN: A Novel Algorithm for Genome Ontological Classification”, Genome Research, p1766-1779.
17. States, D. Editorial (2001) “Time to Defend What We have Won”, Bioinformatics, 17, p299.
18. van Helden, J. V., Andre, B, and Collado-Vides, J. (1998) “Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies”, Journal of Molecular Biology, 281, p827-842.
19. Wodicka, L., Dong, H., Mittmann, M., Ho, M. and Lockhart, D. (1997) “Genomewide Expression Monitoring in Saccharomyces

cerevisiae”, Nature Biotechnology, 15, p1359-1367.