Journal of Engineering and Applied Sciences 1 (1) : 66-68, 2006
© Medwell Online, 2006

Region Query System for Two-dimensional Electrophoresis Images Based on R-tree Index Structure

[1]Kevin I-J Ho, [2]Tung-Shou Chen, [2]Chun-Wei Tsai and [2]Ming-Bin Chune
[1]Department of Information Management Science, Chung-Shan Medical University Taiwan, R.O.C.
[2]Department of Information Management, National Taichung Institute of Technolog, Taiwan, R.O.C.

Abstract: In Post-Genome Era, Functional Genomic becomes a major research field. Two-Dimensional Electrophoresis (abbreviated as 2-DE) technique is a central technique for analyzing proteins. The final process and goal of this technique is to compare different experimental data and construct a database for querying. In this article, we present a query system, based on the R-Tree index technique, to support the regional searching function. With the regional searching ability, researchers may retrieve all the 2-DE images consisting of protein molecules in a given region. The process of the proposed system consists of two phases, one for extracting the regional properties of proteins, the other for constructing the R-Tree.

Key words: Region query system, electrophoresis images, r-tree index structure

INTRODUCTION

The construction of DNA model, done by Watson and Crick[7,8] in 1953, starts the study of life. After several decades, researchers had completed the sequencing of human genome. Since then, *Post-Genome Era* has become. In this era, proteomics or proteome analysis plays an important role. Proteome analysis concerns the separation, identification and quantitation of proteins in biological samples, such as tissues or cells. The purpose of the analysis is to reveal how the living cell works and the function of geonome. Applications of the proteome analysis are wide, including the development of medicine and the monitoring of medical treatment.

Two-dimensional electrophoresis technique, abbreviated as 2-DE, is one of the important tools in proteome analysis. 2-DE converts the protein expression patterns into two-dimensional digital images and facilitate the proteome analysis. Mixed proteins are separated due to the differences in electronic charges, which make the proteins be separated along the X-axle and their molecular weights along the Y-axle. Through the proper staining process, the composition and distribution of the proteins of a sample can be captured in a digital image. Fig. 1 shows a sample 2-DE image.

The protein pattern differences shown on 2-DE images may be very subtle and tedious to observe and analysis by human beings. So, the digital image process becomes a natural part of the proteome analysis. Most of the researches related to digital image analysis of 2-DE images focus on the segmentation of the images and matching of corresponding protein spots[6,9,11]. Therefore, the protein analysis is restricted on the spot information. To analyze the information related a specific region can be only done by manually. It costs tremendous resource. Hence, in this article, we propose and implement a query system, based on the R-Tree index structure, to provide the ability of display all the information of the protein spots located within a given rectangle region.

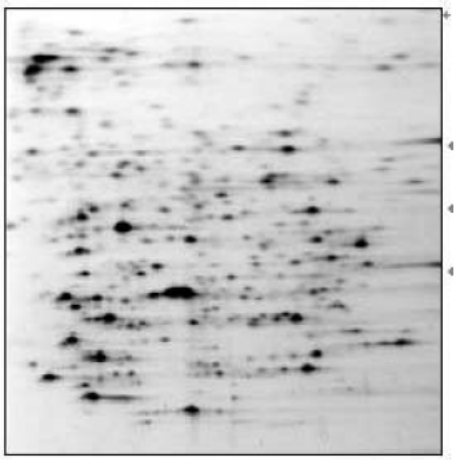

Fig. 1: A sample 2-DE image

This article is organized as follows. In this section, we first introduced the relationship between the analysis of 2-DE images and proteomics. We also state the motivation of our study. In Section 2, we first briefly introduce the R-Tree index structure and then delineate the process of the proposed query system in detail.

Corresponding Author: Chun-Wei Tsai, Department of Information Management, National Taichung Institute of Technolog, Taiwan, R.O.C., Taiwan

Then, experiment results will be shown in the third section. Finally, we draw some concluding remarks and future refinement of the proposed system in the last section.

PROPOSED SYSTEM

The information of protein spots on 2-DE images can be treated as spatial data. Each protein spot is equivalent to an object in spatial data. Searching for all the objects in an area is a common operation for spatial data. To speed up the searching, an index based on the spatial locations of objects are required. A. Guttman[4] proposed a dynamic index structure for searching *n*-dimensional spatial data based on the spatial locations of objects. The proposed index structure facilitates the regional searching. In this article, we employs the R-Tree index structure to construct a query system for searching the protein spot information in a specific region on a 2-DE image.

Basically, an R-Tree is a height-balanced tree, which is similar to a B-Tree[2,3]. The spatial objects contained in a spatial database are represented as a tuple and each tuple has an unique identifier used to retrieve the object. Each leaf node is of the form (I, tuple- identifier), where I = (I0; I2, ..., In-1) is a *n*-dimensional rectangle as a bounding box of the spatial object indexed by the leaf node. Ii, 0 <= i <= *n* - 1, is the closed bounded interval along the i^{th} dimension. Each non-leaf node has the form (I, child - pointer), where *I* represents the bounding box of all the areas covered by its children and child-pointer is a list of addresses of its children. Fig. 2 gives a R-tree example.

In the proposed system, we record the attributes of a protein spot in a leaf node, instead of the tuple-identifier. The attributes include the coordinate of the bounding box of the protein spot, the size of the protein spot, representing by the number of pixels and the coordinate of the center point. Non-leaf nodes are the same as in the R-tree.

For each 2-DE image, we first identify all the protein spots. In this implementation, we simply use the threshold of gray-level value to locate the region of a protein spot. Then, we decide the bounding box for each protein spot based on the method proposed in[5]. After that, we create a R-tree for each image by using the algorithms proposed in[4].

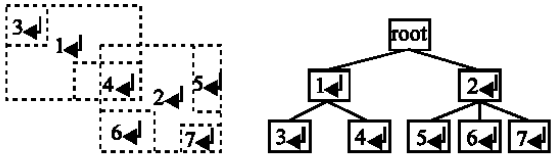

Fig. 2: R-tree sample

EXPERIMENT RESULTS

We implemented the region query system for 2-DE images in Java under Microsoft Windows XP on a Intel based personal computer. To test the correctness and performance of the proposed system, we randomly collect 124 2-DE images from Internet as the testing images. Then, we built the R-tree for each testing image, where we set *M* and *m* as 3 and 2, respectively. Fig. 3 shows the protein spots, the areas circling by white lines, of one of the testing image given in Fig. 1. The rectangles correspond to the non-leaf and leaf nodes of the R-Tree are shown in Fig. 4.

In the experiment, all the testing images are 512*512 pixels. In average, it takes 4.3 seconds to build up a R-tree for an image. And, searching the regional information of one 2-DE image takes around only 1 second. So, we can conclude that the proposed system is helpful for extracting the protein spot information within a specific area. Moreover, for more than 50 times queries, the system correctly shows all the protein spots contained in the query regions.

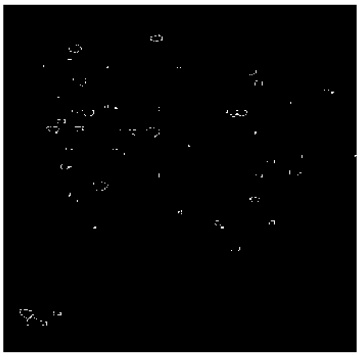

Fig. 3: Detected protein spots

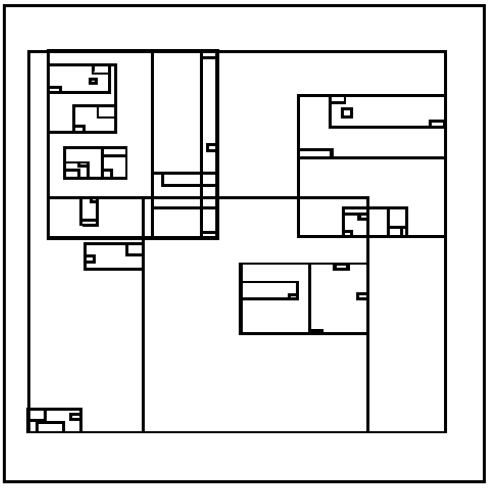

Fig. 4: Rectangles in R-tree

CONCLUSION

In this study, we propose and implement a query system to assist the analysis of regional information contained in 2-DE images. In the proposed system, 2-DE images are preprocessed. We store the attributes of protein spots and their spatial locations by using R-tree index structure. So, when a region information of 2-DE images is searched, we perform the text-comparison. In the past, several methods have been proposed to analyze the information contained in 2-DE images, but all these methods only provide spot information. Therefore, whenever biologists need to get the information of protein spots located in a region, instead of at a point, they need to do it manually. With the help of the proposed system, it can be done efficiently.

Currently, we are working on the refinement of the proposed system in two different directions. One is to compare the performance and accuracy of using different ways to bound protein spots, such as the *convex hulls* and *minimum bounding n - corner convexes*. The other is to investigate if there are other attributes which can be stored in leaf nodes or aggregated in non-leaf nodes, so as to provide more information for proteome analysis.

ACKNOWLEDGEMENT

This work was supported by the National Science Council of Taiwan, ROC, under the contract of NSC–94–2213–E–025–001,NSC–93–2213–E–025–002and NSC–92–2213–E–025–003.

REFERENCES

1. Appel, R.D., J.R. Vargas, P.M. Palagi, D.Walther and D.F. Hochstrasser, 1997. Melanie ii - a third generation software package for analysis of two-dimensional electrophoresis images: Ii. algorithms. Electrophoresis, 18: 2735-2748.
2. Bayer, B. and E. McCreight, 1970. Organization and maintenance of large ordered indices. Procedings of 1970 ACM-SIGFIDET Workshop on Data Description and Access, Houston, Texas, 107-141.
3. Corner, D., 1979. The ubiquitous b-tree. Computing Surveys, 11: 121-138.
4. Guttman, A., 1984. R-tree: a dynamic index structure for spatial searching. Procedings of ACM SIGMOD, 47-57.
5. Papadias, D. and Y. Theodoridis, 1997. Spatial relations, minimum bounding rectangles and spatial data structures. Intl. J. Geograph. Inform. Sci., 11 :111-138.
6. Lars, P., 2002. Analysis of Two-Dimensional Electrophoresis Gel Images. PhD thesis, Technical University of Denmark.
7. Watson, J.D. and F.H.C. Crick, 1953. Genetical implications of the structure of deoxyri- bose nucleic acid. Nature, 171: 964.
8. Watson, J.D. and F.H.C. Crick, 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature, 171: 737-738.
9. Ye, X., C. Y. Suen, M. Cheriet and E. Wang, 1999. A recent development in image analysis of electrophoresis gels. Procedings of Vision Interface 99, Trois-Rivi _ eres, Canada, pp: 432-438.