Visualization of Gene Database on WEB Using Self Organizing Maps

Yasuhiko Miyazaki
miyazaki@dna.ec.saga-u.ac.jp

Hiroshi Dozono
hiro@dna.ec.saga-u.ac.jp

Hisao Tokushima
tokusima@dna.ec.saga-u.ac.jp

Yoshio Noguchi
nogu@dna.ec.saga-u.ac.jp

Faculty of Science and Engineering, Saga University, 1 Honjiyo Saga 840-8502 Japan

Keywords: genome comparison, bioinformatics, self-organizing map

1 Introduction

Recently, entire genome sequences for some species have been determined as the progress of genome analysis, and many database about the sequences, proteins and structures are introduced to the public on the WEB. The amount of the information in each database is huge and each database is comprised of the multivariate data. As the datamining method which can extract the meaningful information from genome database, Self Organizing Maps (SOM) are considered to be applicable. SOM is a kind of neural network which can categorize the input patterns according to the similarities among themselves autonomously by unsupervised competitive learning [1]. After the learning, the genome data can be visualized on the 2-dimensional plains by mapping the multi-dimensional genome data which are used in the learning phase. The genome data which have similar features are mapped closely on the plain, thus each position on the plain indicates the features of the data which are mapped to the position. Some applications of SOMs for genome analysis were reported [2, 3].

In this report, we describe a system which can automatically organize the map of genome data from WEB and can visualize the relation among them according to the keywords or sequences which are comprised in the genome data.

2 Method

To get the genome data from WEB, we used the search services which are equipped to the database. As a genome database, we used the gene database of KEGG because KEGG returns the result of the search as HTML (Hyper Text Markup Language) stream which can be easily interpreted in consecutive steps. As a genome database, the NCBI (National Center for Biotechnology Information) database is also open for public, but it returns the results of the search as JavaScript which needs interpreter to extract the information from them. Our system query the database in GenomeNet using arbitrary keyword inputted from the form tailored to the system and the get the results of searches as HTML streams. The HTML streams are interpreted and the information of the URL (Uniform Resource Locater) to the objective genome data are extracted. The genome data are transferred from database using HTTP (Hyper Text Transfer Protocol).

Using the data of genes as genome data, the gene data are comprised of the some fields, which contains the Entry name, gene name, related pathways and sequences of aminoacid and nucleotides, e.g. Analyzing the gene data using the keywords as feature vectors, the frequencies of the keywords in the gene data compose the feature vectors which are used for training SOM. Analyzing the gene data using the sequence data of genes, the sequence of the amino-acid or nucleotides are extracted. These data are used for training SOM directly or after preprocessing. To train the map directly using sequences, the SOMS are trained using all sequences of n-tuples in the sequence. To train the map after preprocessing, the SOMS are trained by the feature vectors that are comprised of the frequency of each n-tuple which are proposed in [3]. The size of the feature vector becomes large using large amount of keyword or n-tuple of long length and the amount of the computations for learning becomes huge. We used the SOM combined with Principal Component Analysis.
(PCA) to reduce the size of feature vectors without losing important information.

After learning, the learned data are mapped on the organized map again. Using the keywords as feature vectors, some keywords are selected from GUI (Graphical User Interface) and the map is colored by the frequency of the keyword. From this analysis, we can visualize the relations among the keywords on the map depending on the genome data used in the learning phase. Using the sequence data as feature vectors, some set of genes are selected from GUI and each position on the map is colored by the most similar sequence. From this analysis, we can visualize the relations of genes depending on the pathways, species and evolitional differences.

Figure 1: Overview of the system based on SOM

3 Result and Discussion

In this report, we described the system based on SOM, which automatically gets the genome data from WEB and visualize the relation of them according to the keywords or sequence data. Now, we have poor GUIs, we will remake the GUI until GIW 2005 and show it at poster presentation.

As the future work, we should improve the usages of keywords using the dictionary of bio-informatics to avoid the useless keyword and improve the system as to visualize the genome data other than the gene data and make our system compatible with NCBI database. And we must make GUI which is friendly for biologist or medical doctors.

References

