The Gene Network: Development of a Database System Showing the Relationships Among the Genes of All Species

Nobuyuki Miyajima
miyajima@kazusa.or.jp
Shinobu Nakayama
nakayama@kazusa.or.jp
Mitsuyo Kohara
kohara@kazusa.or.jp
Satoko Hayashi
hayashi@kazusa.or.jp

Department of Genome Informatics, Kazusa DNA Research Institute
1532-3 Uchino Yana Kisarazu, Chiba 292 Japan

Abstract

We have developed a sophisticated method called "The Gene Network" for elucidating the relationships existing among all genes. This was accomplished by employing 24222 gene symbols contained in the MEDLINE database of Entrez rel. 12.0, then determining their inter-relationships by examining the frequency of appearance among one another. This new method enables construction of gene maps which graphically show their relationships, there by enhancing the understanding of them. We expect The Gene Network will have the future capability to allow navigation through the "world of genes."

1 Introduction

Numerous recent advances in molecular biology techniques, especially those employed for DNA sequencing and the polymerase chain reaction (PCR) method, have elucidated many unknown genes at an amazing rate. The resultant multitude of world-wide published information, however, cannot be fully utilized since it is almost impossible to absorb and digest all such data. To maximize the utilization of published data associated with the genes of all species, we focused our attention on clarifying the relationships among them so as to gain a sophisticated understanding of unknown relationships on unrecognized facts. Consequently, our efforts led to the development of an integrated database of genes, which we named "The Gene Network."

宮嶋伸行、中山しのぶ、小原光代、林聡子：かずさ DNA 研究所，〒292 千葉県木更津市矢那内野 1532-3

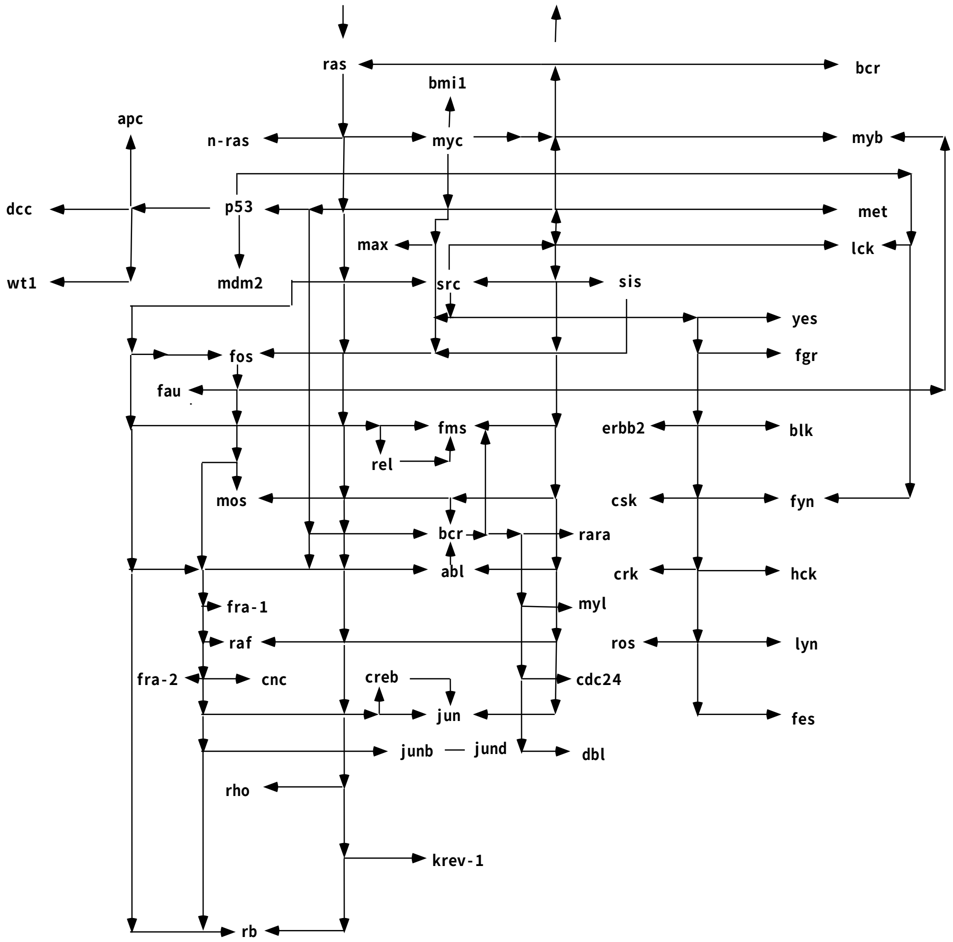

Figure 1: The most congested area of Gene World.

2 Methods and Results

We initially obtained Entrez release 12.0 from the National Center for Biotechnology Information (NCBI). This database contains 213993 MEDLINE records, which comprise MEDLINE's molecular sequence data subset. By using several perl scripts, we investigated the frequency of Gene Symbols occurring in this data subset.

Genes are normally discussed in only one article of MEDLINE data because some correlation should exist among them. Therefore, based on this concept, we calculated the frequency of different genes present in all the articles on one gene. Regarding the ras gene, for example, all the other genes co-discussed in articles on it can be determined. Following the examination of all Gene Symbol items contained in MEDLINE data, we obtained a putative correlation table for 24222 genes. It should be noted that the resultant correlating genes are sequentially ranked according to the total numbers of correlations found.

To make the relationships among genes easier to grasp, we decided to graphically show their relationships. Using the correlation table of a particular gene, we employed modified version of tgif software to draw arrows between genes(Fig. 1), i.e., if two genes have more than two co-discussed citations, they can be connected by an arrow.