CAGE: A Database of Cancer Genes of Human, Mouse and Rat

Hassan Tariq[1*], Aurengzeb Muzammil[2], Sana Khalid[1]

[1]*Department of Bioinformatics and Biotechnology*
Government College University, Faisalabad, Pakistan
E-mail: pubht@yahoo.com

[2]*College of Computer Science and Information Sciences*
Government College University, Faisalabad, Pakistan

[*]*Corresponding author*

Received: February 22, 2011 **Accepted: October 10, 2011**

Published: November 30, 2011

Abstract: *CAGE is the database of cancer genes of human, mouse and rat. We have designed PCR oligonucleotide primer sequences for each gene, with their features and conditions given. This feature alone greatly facilitates researchers in PCR amplification of genes sequences, especially in cloning experiments. Currently it encompasses more than 1000 nucleotide entries. Flexible database design, easy expandability, and easy retrieval of information are the main features of this database. The Database is publicly available at cgdb.pakbiz.org.*

Keywords: *Genomic Database, Human, Mouse, Rat.*

Introduction

Historically, databases have been arisen, to satisfy diverse needs, whether it address a biological question of an interest to an individual scientist, to better serve a particular section of biological community, to co-ordinate data from sequencing projects, or to facilitate drug discovery in pharmaceutical companies. According to Nucleic Acids Research annual database issues, in 2010 update, the online Database Collection that accompanies the issue holds 1230 data resources, a growth of 5% over last year [6]. Most available data are computationally derived and include errors and inconsistencies. Effective use of available data in order to derive new knowledge hence requires data integration and quality improvement [8].

The exponential growth of biological data over the past decade has created an enormous challenge to make effective use of the accumulated information. Correctly cataloging, labeling and connecting sequence, structural and functional information of genes and proteins of various trends and laws crucial to our understanding of life on earth as complex systems [4]. Databases are great tools because they offer a unique window on the past. They make it possible to answer today's biological questions by enabling us to analyze sequences that may have been determined as many as 25 years ago, when the whole technology emerged. By doing this, they connect past and present molecular biology and other life sciences [5].

The primary goal of the current project is to develop a specialized, minimally redundant, and curated nucleotide sequence database of human, mouse and rat that strives to provide high-level annotations, including species based categorization of expressed genes, primers designed for the amplification of expressed genes.

Materials and methods

Data collection

In the present study to develop the desired database, genes sequences that are expressed in cancer cells of different species and their relevant annotations were required. To collect the data we searched for protein coding genes in NCBI's 'GenBank' and 'Gene' databases. We used nucleotide sequences in FASTA format and designed primers using Primer3. Then we analyzed the designed primers for primer-dimer formation and secondary structures using NetPrimer.

Database Design

1-Species Section

In this section, we can add new specie; edit the information about currently existing families and delete the currently existing specie.

2-Cancer Type Section

In this section we can also add a new cancer, secondly we can edit the current cancer or other information and thirdly delete the existing cancer.

3-Genes Section

In this section we can add new genes, second, edit the current genes or other information related to it and third, delete existing genes.

Results

In the present study, we included complementary DNA nucleotide sequences for each gene. The primers designed for theses complementary DNA sequences are really useful in their PCR amplification when they are cloned into some sort of vector. The current database also includes protein information of the relevant genes and their function. These features are the result of our flexible database design. Fig. 1 shows the proportion of entries of each species.

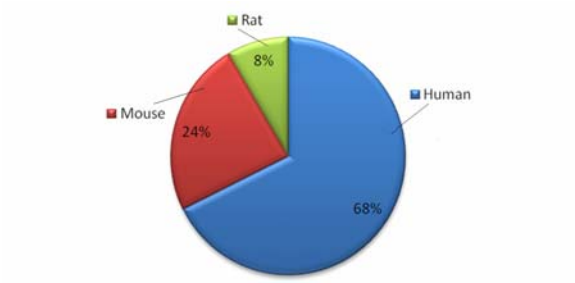

Fig. 1 Percentages of number of records of each specie in the resource

Followings are the salient features of the CAGE:

1. **Data searching**

CAGE provides a very stylish way of searching the data. We can search our required information in two ways:

- **Data searching by search field**

  CAGE facilitates the users to search data by giving keyword related to function, protein, gene. If the record is found in the database then it will show all the results in all possible species.

- **Data searching through navigation**

  CAGE provides the facility for the users to search their relevant data by navigating the database. Whenever we click a specie, a list of types of cancer will appear. From this list we will choose one cancer type and a list of genes present in it will appear, we can view its details by further navigating into it.

2. **Easy and fast access to the information**
   We can get access to data in no time. Data searching is so easy in CAGE that even a new user can search through it with almost no difficulty.
3. **Built in primers**
   Primer designing has been the most distinguishing feature of this database. It is a new concept in database designing. It will help the scientist in PCR amplification of specific gene. Additionally, the conditions and features given pertaining to a particular primer also facilitate scientists to work effectively.

Discussion

As part of the present study, we have also developed a specialized database CAGE to store species based categorized expressed genes nucleotide sequences and their annotations related to genes present in different species. Currently, it is focused on human, mouse and rat.

The goal of biological databases is two fold: information retrieval and knowledge discovery [9]. Similarly, the chief objective in database development was to organize data in such a way in a set of structured records that user community can easily retrieve the desired information. Keeping a good eye on the usage details of the database and the needs of the people using it is the only way to stay grounded [3]. In this project, special efforts are employed to get right details for effective database development, because designing, implementing and running databases are predominantly a series of decisions about intricate details [3]. The sequences in this database may overlap with the primary databases like GenBank [2], but it also has newly submitted data, which was obtained by submitting genes nucleotide sequences in online analysis programs, and then from the outputs of programs different kinds of new data was obtained. Thus, CAGE has its own unique organization and unique related annotations associated with the genes nucleotide sequences.

Although many issues in creating a good database may transcend biology and be valid for all domains, there are special circumstances around biological databases that make them worth treating as a special group (i.e. the free availability that they are on internet, that they keep up with rapidly growing field, and that they maintain high biological relevance) [1]. So like other plant databases CAGE is also free online database. It can be accessed through easy-to-use web interface. All the data in the database is freely available with no restrictions. Its data can be used in wide range of applications and scenarios by users ranging from laboratory scientists to experienced bioinformaticians.

Keeping in view the fact that the manual selection of optimal PCR oligonucleotide sets can be quite tedious and thus lends itself very naturally to computer analysis [7], CAGE also contained PCR oligonucleotide primer sequences for nucleotide sequences of genes. These primers' design is aimed at obtaining a balance between two goals: specificity and efficiency of amplification. In primer designing, this balance is obtained by analyzing the quality of primers with various programs, considering specially avoiding primer-dimer formation and secondary structure in primers.

Acknowledgement

We are thankful to Dr. Shahid Nadeem and all of the researchers of Database and Software Engineering Research Group of the Department of Bioinformatics and Biotechnology, Government College University, Faisalabad, Pakistan.

References

1. Altman R. B. (2004). Building Successful Biological Databases, Brief Bioinf, 5(1), 4-5.

2. Benson D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers (2009). GenBank, Nucleic Acids Res, 38(Database issue), D14-D15.
3. Birney E., M. Clamp (2004). Biological Database Design and Implementation, Brief Bioinformatics, 5(1), 31-38.
4. Buehler L. K., H. H. Rashidi (2005). Bioinformatics Basics: Applications in Biological Science and Medicine, CRC Press, USA, 166 pp.
5. Claverie J. M., C. Notredame (2006). Bioinformatics for Dummies, 2nd Edition, John Wiley & Sons Inc., 69-104.
6. Cochrane G. R., M. Y. Galperin (2010). The 2010 Nucleic Acids Research Database Issue and Online Database Collection: A Community of Data Resources, Nucleic Acids Res, 38(Database issue), D1-D4.
7. Dieffenbach C. W., T. M. Lowe, G. S. Dveksler (1993). General Concepts for PCR Primer Design, PCR Method Appl, 3(3), S30-S37.
8. Ghisalberti G., M. Masseroli, L. Tettamanti (2010). Quality Controls in Integrative Approaches to Detect Errors and Inconsistencies in Biological Databases, J Integr Bioinform, 7(3), 2010-2119.
9. Xiong J. (2006). Essential Bioinformatics, Cambridge University Press, New York, USA.

Hassan Tariq, M.Sc. Bioinformatics
E-mail: pubht@yahoo.com

Hassan Tariq completed his M.Sc. in Bioinformatics from Government College University – Faisalabad in 2011. Since 2009 he is working as a Research Officer in the Department of Bioinformatics and Biotechnology. He is the founder of Database and Software Engineering Research Group in the Department of Bioinformatics and Biotechnology. He is also an active member of teaching Faculty of the department, teaching some of the interesting subjects like Bioinformatics Software Development, also publishing textbooks. Scientific interests: Next Generation Sequencing, Biological Databases, Data Mining, Software Engineering in Bioinformatics, Web Engineering.

Aurengzeb Muzammil, M.Sc. Computer Science
E-mail: thestylish10@yahoo.com

A. Muzammil is an active member of Database and Software Engineering Research Group at Government College University Faisalabad Pakistan. He is currently working as Assistant Professor in the College of Computer Science and Information Sciences. He did his M.Sc. in Computer Sciences from the University of Agriculture, Faisalabad. In research group he is working on ongoing research projects. His research interests are software development, computer networks, wireless networking and bioinformatics.

Sana Khalid, M.Sc. Bioinformatics
E-mail: sanakhalid36@yahoo.com

Sana Khalid is an active member of Database and Software Engineering Research Group at Government College University Faisalabad Pakistan. She graduated from the Government College University Faisalabad in 2010. In research group she is working on ongoing research projects. She has command over many Bioinformatics data analysis and structure prediction tools. Her interests are software development, drug designing, phylogenetic analysis and protein structure prediction.