GasCan: A Novel Database for Gastric Cancer Genes and Primers

Sobia Idrees*, Shahid Nadeem, Beenish Ehsan, Muhammad Ibrahim Rajoka

Department of Bioinformatics and Biotechnology
Government College University
Faisalabad, Pakistan
E-mails: sobia_binm@live.com, snadeem63@yahoo.com, merjeena_196@yahoo.com, mibrahimrajoka47@gmail.com

**Corresponding author*

Received: January 11, 2012 **Accepted: July 12, 2012**

Published: July 16, 2012

Abstract: *GasCan is a specialized and unique database of gastric cancer protein encoding genes expressed in human and mouse. The features that make GasCan unique are the availability of gene information, availability of primers for each gene, and built in programmed sequence analysis facility that analyze gene sequences in database itself. Furthermore, DNA sequence analysis tool is provided that can be access freely. GasCan will expand in future to other species, genes and cover more useful information of other species. Flexible database design, expandability and easy access of information to all of the users are the main features of the database. The database is publicly available at http://www.gastric-cancer.site40.net.*

Keywords: *Gastric Cancer, Database, Human, Mouse, Sumatran, Sequence Analysis.*

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].

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 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]. 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 [9].

Computational analysis of biological sequences has become an extremely rich field of modern science and a highly interdisciplinary area, where statistical and algorithmic methods play a

key role. In particular, sequence alignment tools have been at the hearth of this field for nearly 50 years [10].

However, given the burgeoning array of molecular biology databases as well as data retrieval and analysis tools, users are challenged daily to identify the resources that best fit their needs and to use them effectively. This raises questions about the demographics of bioinformatics users, their needs, and libraries' roles in meeting those needs [8].

Due to day by day increase in information present in online resources, data searching is becoming difficult. To tackle with this problem specialized databases are being developed that provide data related to particular problem. Such specialized databases are more quick and easy to use. These databases can be species specific or disease specific or providing information about specific genes, proteins or their respective families. We noticed these new trends in information management and devised a new way to provide gene information of gastric cancer with sequence analysis facility.

The main objectives of our project are to provide:

1. A specialized, minimally redundant, and curated nucleotide sequence database of human and mouse that strives to provide high level annotations, including species based categorization of expressed genes.
2. Automatic sequence analysis facility in database.
3. Designed primer that can help in the amplification of expressed genes.
4. User friendly data retrieval facility so that even novice user can retrieve data without headache.
5. Single platform where researcher can retrieve and perform analysis.
6. Sequence analysis tools.

Materials and methods

Data collection

In the present study to develop the desired database, genes sequences that are expressed in 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

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]. Fig. 1 shows the database structure and working strategy.

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. 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.

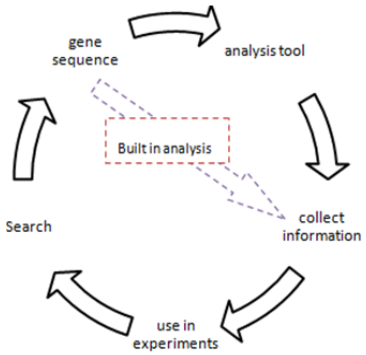

Fig. 1 Database structure and working

3. Article section

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

4. Sequence analysis section

This section is programmed to analyze the sequence. User searches the sequence and before the retrieval, the data is first analyzed by this section and resulting information is sent to user.

Sequence analysis tools

Sequence analysis tools i.e. *in silico* central dogma of molecular biology, complement, reverse complement, nucleotide weight, melting temperature of primers, GC content percentage and nucleotide composition etc. with interactive graphical representation are developed and included in database to facilitate the researchers.

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. 2 shows the proportion of entries of each species. Followings are the salient features of the GasCan.

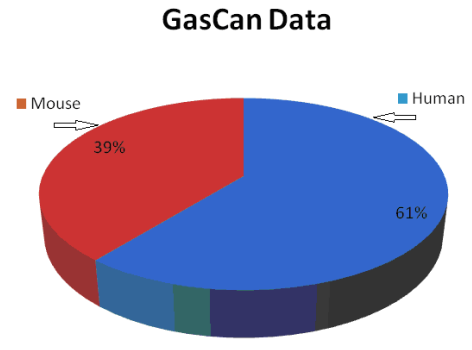

Fig. 2 Proportion of gene data in each species

1. Data searching

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

Data searching by search field
GasCan 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
GasCan provides the facility for the users to search their relevant data by navigating the database. Whenever we click specie, a list of genes related to that specie will appear. From this list we will choose one gene and after clicking it, 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 GasCan that even a new user can search through it with almost no difficulty.

3. Built in primers
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.

3. Built in sequence analysis
Built in sequence analysis facility is the most distinguishing feature of this database. It is a new concept in database designing and can save researcher's time. It accesses information from database, analyze it and then show the results to user along with gene detail.

4. Sequence analysis tools
Sequence analysis tools with interactive graphical representation are developed using multiple web programming languages and are provided in database to facilitate the researchers.

Discussion
As part of the present study, we have developed a specialized database GasCan 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 and mouse.

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]. 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]. 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, GasCan 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 spcialized genomic databases GasCan 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 and sequence analysis facility 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]. GasCan 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 Bioinf, 5(1), 31-38.
4. Buehler L. K., H. H. Rashidi (2005). Bioinformatics Basics: Applications in Biological Science and Medicine, CRC Press, USA.
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. Geer R.C. (2006). Broad Issues to Consider for Library Involvement in Bioinformatics, J Med Libr Assoc, 94(3), 286-298.
9. 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), 119, doi:10.2390/biecoll-jib-2010-119.
10. Giancarlo R., A. Siragusa, E. Siragusa, F. Utro (2007). A Basic Analysis Toolkit for Biological Sequences, Algo Mol Biol, 2(10), 404-406.

Sobia Idrees, M.Phil. in Biotechnology
E-mail: sobia_binm@live.com

Sobia Idrees is a Master student at Government College University, Faisalabad, Pakistan. Based on her BS research project she obtained 1st position in the software competition at national level. Scientific interests: Software Engineering, Web and Database Development, Bioinformatics Sequence Analysis, Drug Designing, Health Biotechnology and Molecular Biology.

Shahid Nadeem, Ph.D.
E-mail: snadeem63@yahoo.com

Dr. Shahid Nadeem got his Ph.D. from the Punjab University, Lahore in 2002. He is working as a senior scientist in NIAB, Faisalabad. In 2009 he joined the Department of Bioinformatics and Biotechnology as a Chairman. He is among the founders of Department of Bioinformatics and Biotechnology in GC University Faisalabad. He is also 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 Biotechnology and its applications, Social, Ethical aspects of Biotechnology also publishing textbooks. Scientific interests: Biotechnology, Plant breeding and Genetics, Microbiology.

Beenish Ehsan, M.Phil. in Biotechnology
E-mail: merjeena_196@yahoo.com

Beenish Ehsan is pursuing Master in Biotechnology at Government College University, Faisalabad. Scientific interests: Phylogenetic Analysis, Molecular Biology, Health Biotechnology.

Muhammad Ibrahim Rajoka, Ph.D. in Biochemistry
E-mail: mibrahimrajoka47@gmail.com

M. I. Rajoka graduated from the University of New South Wales – Australia in 1981 with the distinction of securing second highest percentage in the session. He did Ph.D. in Biochemistry from University of the Punjab in 1990. After serving 41 years in Nuclear Institute for Agriculture and Biology – Faisalabad (19 years) and National Institute for Biotechnology and Genetic Engineering – Faisalabad (22 years), he joined as Professor in the Department of Bioinformatics and Biotechnology. He has supervised 25, 21 and 7 M.Sc., M.Phil. and Ph.D. students respectively and has published 140 research papers and review articles in International and national peer review journals and proceedings of conferences/workshops. He is also an active member of teaching Faculty of the department, teaching bioprocess technology, research methods in biological sciences, technical writing, communication skills and guiding faculty in writing research proposals. Scientific interests: Industrial strain development through conventional breeding and rDNA technology, bioprocess engineering.