International Journal of Data Mining Techniques and Applications
Volume: 01 Issue: 01 June 2015 Page No.17-20
ISSN: 2278-2419

GASCAN: A Novel Database for Gastric Cancer Genes and Primers

V.Khanaa[1], Krishna Mohanta[2]
[1]Dean - Information. Bharath University, Chennai,
[2] Sri Sai Ram Engg. College Chennai
Email: drvkannan62@yahoo.com, krishnamohanta@gmail.com

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 availability of gene information, availability of primers for each gene, with their features and conditions given that are useful in PCR amplification, especially in cloning experiments and to make it more unique built in programmed sequence analysis facility is provided that analyze gene sequences in database itself, resulting sequence analysis information can be valuable for researchers in different experiments. 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.

Key words: Gastric Cancer, Database, Human, Mouse, Sumatran, Sequence Analysis

I. 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;

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

II. MATERIALS AND METHODS

A. 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.

B. Database Design

In this project, special efforts are employed to get right details for effective database development, because designing,

International Journal of Data Mining Techniques and Applications
Volume: 01 Issue: 01 June 2015 Page No.17-20
ISSN: 2278-2419

implementing and running databases are predominantly a series of decisions about intricate details [3]. Fig.1 shows the database structure and working strategy
.

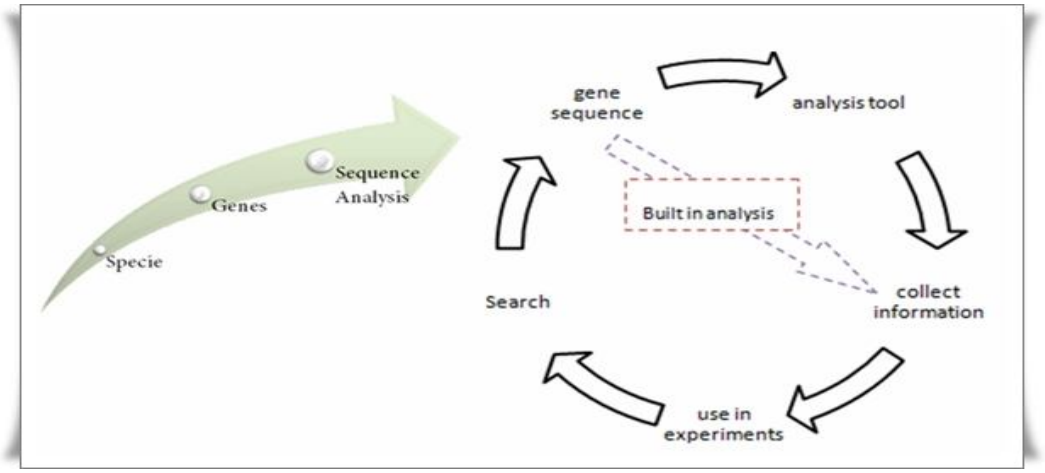

Fig. 1: Database structure and working.

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

b) 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.

c) 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.

d) 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.

III. 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.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:

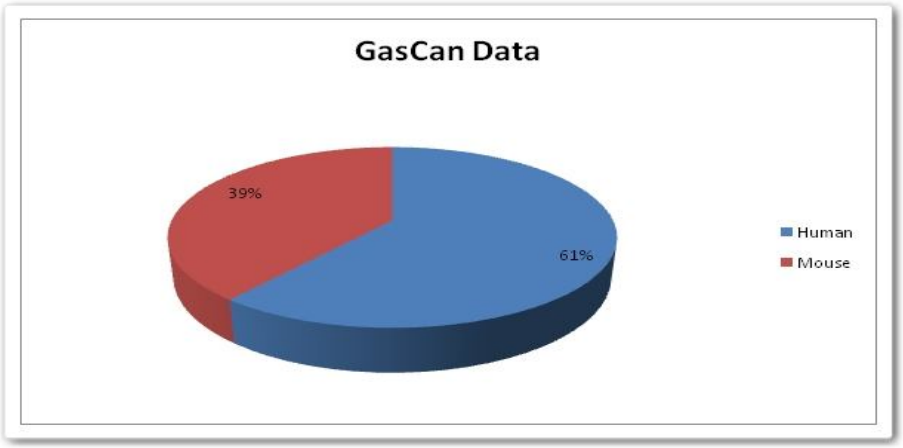

Fig. 2: Proportion of gene data in each specie.

International Journal of Data Mining Techniques and Applications
Volume: 01 Issue: 01 June 2015 Page No.17-20
ISSN: 2278-2419

AY039238	PPHLN1	Homo Sapiens	Periphilin-1	Homo sapiens gastric cancer antigen Ga50 (HSPC232) mRNA, complete cds.	View Detail
AF049910	TACC1	Homo Sapiens	Transforming acidic coiled-coil-containing protein 1	Homo sapiens TACC1 (TACC1) mRNA, complete cds.	View Detail
X76732	NUCB2	Homo Sapiens	Nucleobindin-2	Homo sapiens cell line GM10618 DNA binding protein NEFA precursor,mRNA, complete cds	View Detail
BC031838	RRP7A	Homo Sapiens	Ribosomal RNA-processing protein 7 homolog A	Homo sapiens ribosomal RNA processing 7 homolog A (S. cerevisiae), mRNA (cDNA clone IMAGE:5094598), partial cds.	View Detail
AF494509	GKN2	Homo Sapiens	Gastrokine-2	Homo sapiens GDDR (GDDR) mRNA, complete cds.	View Detail
AB034695	EMCN	Homo Sapiens	Endomucin	Homo sapiens mRNA for endomucin-2, complete cds.	View Detail
AY039242	NAA15	Homo Sapiens	N-alpha-acetyltransferase 15, NatA auxiliary subunit	Homo sapiens gastric cancer antigen Ga19 mRNA, complete cds.	View Detail
AY074490	CAPRIN2	Homo Sapiens	Caprin-2	Homo sapiens EEG1L (EEG1) mRNA, complete cds; alternatively spliced.	View Detail
AF438406	GCRG224	Homo Sapiens	Putative gastric cancer-related gene 224 protein	Homo sapiens GCRG-P224 mRNA, complete cds.	View Detail
AB096683	FAM72D	Homo Sapiens	Protein FAM72D	Homo sapiens GCUD2 mRNA for Gastric cancer up-regulated-2, complete cds	View Detail
AB083199	KCMF1	Homo Sapiens	E3 ubiquitin-protein ligase KCMF1	Homo sapiens mRNA for FIGC1, complete cds.	View Detail
D43772	GRB7	Homo Sapiens	Growth factor receptor-bound protein 7	Homo sapiens mRNA for GRB-7 SH2 domain protein, complete cds.	View Detail

Fig. 3: Human Gastric Cancer gene data.

We can search our required information in two ways:

A. Data searching

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.

B. 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.

C. **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.

D. 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.

E. Sequence analysis tools

Sequence analysis tools with interactive graphical representation are provided in database to facilitate the researchers.

IV. CONCLUSION

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 (Birney *et al.,* 2004). 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 (Birney *et al.,* 2004).The sequences in this database may overlap with the primary databases like GenBank (Benson *et al.,* 2009) 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) (Altman *et al.,* 2004)

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 (Dieffenbach *et al.,* 1993).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.

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.

International Journal of Data Mining Techniques and Applications
Volume: 01 Issue: 01 June 2015 Page No.17-20
ISSN: 2278-2419

REFERENCES

[1] Altman R. B. (2004) Building Successful Biological Databases, Brief Bioinf. 5(1), 4-5.182
[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] Geer, R.C. (2006). Broad issues to consider for library involvement in bioinformatics, JMed LibrAssoc., 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), 2010-2119.
[10] Giancarlo, R., A. Siragusa, E. Siragusa, F. Utro (2007). A basic analysis toolkit for biological sequences, Algo Mol Biol., 2(10), 404-406.
[11] Xiong J. (2006). Essential Bioinformatics, Cambridge University Press, New York, USA.