INT. J. BIOAUTOMATION, 2014, 18(2), 67-74

CGPD: Cancer Genetics and Proteomics Database – A Dataset for Computational Analysis and Online Cancer Diagnostic Centre

Muhammad Rizwan Riaz[#*], Attia Iram[#]

Department of Bioinformatics and Biotechnology
Government College University
Faisalabad, Pakistan
E-mails: rizi.leo20@gmail.com,
irramattia@yahoo.com

[*]*Corresponding author*

[#]*Both authors contributed equally*

Received: December 17, 2013 **Accepted: March 23, 2014**

Published: June 30, 2014

Abstract: *Cancer Genetics and Proteomics Database (CGPD) is a repository for genetics and proteomics data of those Homo sapiens genes which are involved in Cancer. These genes are categorized in the database on the basis of cancer type. 72 genes of 13 types of cancers are considered in this database yet. Primers, promoters and peptides of these genes are also made available. Primers provided for each gene, with their features and conditions given to facilitate the researchers, are useful in PCR amplification, especially in cloning experiments. CGPD also contains Online Cancer Diagnostic Center (OCDC). It also contains transcription and translation tools to assist research work in progressive manner. The database is publicly available at http://www.cgpd.comyr.com.*

Keywords: *Cancer, Database, CGPD, OCDC, Bioinformatics, Genetics.*

Introduction

Cancer belongs to a group of diseases which can be lethal at any stage of human life. Cancer exists because of somatic selection; mutations in somatic cells result in some dividing faster than others, in some cases generating neoplasms. Neoplasms grow, or do not, in complex cellular ecosystems. Cancer is relatively rare because of natural selection; our genomes were derived disproportionally from individuals with effective mechanisms for suppressing cancer. Cancer occurs nonetheless for the same six evolutionary reasons that explain why we remain vulnerable to other diseases. These four principles – cancers evolve by somatic selection, neoplasms grow in complex ecosystems, natural selection has shaped powerful cancer defenses, and the limitations of those defenses have evolutionary explanations – provide a foundation for understanding, preventing, and treating cancer [1].

CGPD is a biological database that was developed to acquire all the relevant information about cancer genomics and proteomics realities in *Homo sapiens*. There are many biological repositories having specific information about different biological aspect for example DDBJ [7, 9], GenBank [10], UniProtKB [14], PDBe [3], CAGE [12], etc. CGPD contains all the information like nucleotide sequence of all the genes that could be involved in developing specific type of cancer. The cause of mutation depends on the environment in which a human live, but the effect of a bad environment can be thousand times increased by targeting the genes which are specific for cancer development. So in this situation, one can hope to

eliminate cancer from human's life by focusing on these genes and control cancer through these genes.

There are many types of cancers based on their multiple classification schemas. A view of cancers merely growing is being replaced by recognition that they evolve according to well-understood principles of somatic selection, along trajectories that can be described by established methods for tracing phylogenies. This has practical applications for understanding the significance of heterogeneity within tumors, and implications for diagnosis and treatment [1]. Furthermore, a single gene can have its more than one variant, exhibiting different behavior or structure. The overwhelming majority of polymorphisms studied are single nucleotide polymorphisms (SNPs) that occur with a frequency of > 1% in the normal population (in contrast to "mutations" that occur with a frequency of < 1%). It is estimated that up to 10 million SNPs are probably present in the human genome though not all have thus far been identified. Naturally, most of these SNPs do not occur in coding sequences and even those that do, are not associated with any alteration in the amino acid sequence and are therefore of no functional consequence [4].

The CGPD was developed as a relational database using MYSQL software. The database is searched by PHP script. The open source program Apache HTTP Server was used to build HTTP Server.

A researcher cancer can obtain all the required information from CGPD not only the sequences but the website also has a unique and definite feature called "online cancer diagnostic center", where you can find the diagnostic tool for cancer. One can find the type of cancer by selecting the appropriate symptoms of the patient. Cancer is somehow holding a major role in the history of mankind. As it does not only cause death or a miserable human life but also in other sense it also can be considered as a major source to study the life, either it can be of a single cell or a whole multi cellular organism. Previous research is not only for the purpose to eradicate cancer completely but also to focus that how life take its form from just a single cell to whole population or a clump of undifferentiated cells called the tumor. In recent years, biological databases have greatly developed and increased very well, and they became a normal part of the biologist's everyday toolbox. Biological databases can be broadly classified into sequence and structure databases. In general, databases can be furthermore classified into primary and secondary databases. Primary database consisting of data derived experimentally and secondary database contains organized data which has been derived from primary database. CGPD is a secondary database which contains data derived from other primary databases and it has organized data architecture.

Materials and methods

Selection of programming languages

The first step was to determine the programming language to create the "Front-end" i.e. interface, HTML, CSS and AJAX were used to design GUI and PHP used to connect backend database with the GUI. PHP is a widely deployed dynamic web language specifically created for developing web pages with flexible and powerful built-in functions that allow for quick access to a comprehensive online database. Along with PHP, other fundamental web technologies were employed to develop the database, such as Cascading Style Sheet (CSS) and HTML.

Data collection

The data for CGPD is collected from these sources:

- GenBank [10]: nucleotide sequence of the corresponding genes. That can be involved in any type of the gene.
- UniProt/KB [14]: amino acid sequences of the proteins involved in cancer.
- Primer sequences designed by Primer3 [5] and Primer-Blast tool [8].
- Promoter sequences used in CGPD are taken from BDGP 1999 NNPP version 2.2 [11].

CGPD is a relational secondary database providing all information about the types of cancers based on their locality in the specific body part. It has the information which is very specific and definite for a cancer type.

Designation of the database

CGPD is a relational database developed by using:

- MYSQL 5.5.15;
- PHP 5.3.7;
- Apache server 2.2.19.

Online cancer diagnostic center

OCDC is a tool is to predict type of cancer on the basis of symptoms. User can select specific symptoms from a list and this tool will predict cancer type on the basis of probabilities of selected symptoms.

Cancer can be of any type and according to its type it may have multiple symptoms. Some symptoms might not lead to cancer while others could. On the basis of results it also suggests the further possible diagnostic tests.

The design of OCDC is very user friendly. Algorithm defines a particular type of cancer on selection of one or more different symptoms. So for this purpose some special formulae were required to calculate the probability of a particular type of cancer. In scripting of this design, logical operators are used. Like for overlapping symptoms we use “OR” and for different symptoms we use the “AND” operator.

According to the symptoms that a user had selected, the result appears as a type of cancer for those specific symptoms. In results the tool also recommends further steps in diagnostics like which test should be performed.

Results and discussion

CGPD contain information like nucleotide sequences, amino acid sequences, primers and promoter sequences of genes or proteins related to fourteen types of cancers. All genes involved in each type of cancer are given in Table 1. Information should not overlap or misunderstood in any of its format in a relational database.

The most important thing for a researcher is to find accurate sequence regarding his specific type of tumor. The information available in soft form like previously approved DNA or RNA sequence of a gene involved in cancer can be easily retrieved from CGPD. In addition to that, this database also provides primer and promoter sequences for those genes. This database is

updated with new genes and their relevant information which are reported to be linked to any type of cancer in the literature.

Table 1. Genes involved in different types of cancers

Cancer types	Genes
Colon cancer	APC MSH2 MSH6 MLH1 PMS1
Breast cancer	BRCA1 BRCA2 CDH1 PTEN STK11 AR TP53 BARD1 BRIP1 DIRAS3
Ovarian cancer	CASC4
Lung cancer	FRA3B EGFR KRAS NRG1 TP53 PA2G4 TP63 FHIT
Cervical cancer	CCNB1 AKT1 BLC2 MMP3
Prostate cancer	HPC1 CSAD EPHB2 CBX4 MSR1 ELAC2
Testicular cancer	KITLG
Brain cancer	OSM ABHD2 COL3A1 FIGF CSPG4 PPFIBP1 RAD51B PDGFB AIM2 ABHD2 DDR1 FOSL1 CSPG4 PTPRF RAD51B SULF2 PLEKHB1 PREX1 PRKG2 NGFR
Kidney cancer	PBRM1 ARID1A
Liver cancer	IQGAP1 CMAS
Skin cancer	MAP2K1 DHRS3
Thyroid cancer	TGFB3
Uterine cancer	SLC2A1 IL8 IDO1

Data searching

CGPD provides its users with a very simple and configured way of searching the data. This is actually through selection of user's required fields; in this case these may be DNA/RNA/primer/promoter sequence, through a drop down menu or list. First user will have to specify the type of cancer and then he has to select which sequence is required (Fig. 1) and then the results will appear in the next windows with the entire gene's information that a user needs.

DNA mismatch repair (MMR) proteins are ubiquitous players in a diverse array of important cellular functions. In its role in post-replication repair, MMR safeguards the genome correcting base mispairs arising as a result of replication errors. Loss of MMR results in greatly increased rates of spontaneous mutation in organisms ranging from bacteria to humans. Mutations in MMR genes cause hereditary nonpolyposis colorectal cancer, and loss of MMR is associated with a significant fraction of sporadic cancers. The rapidly increasing information about these mutations need to be collected and appropriately stored to facilitate further studies on the biological and clinical significance of the findings [2] and we have same

thing for other types of cancer. The management of this huge amount of information is actually important. The Human Genome Project has increased the rate of DNA sequence accumulation to the point where information management has become a formidable task [1]. Not only to make huge databases containing large amount of data, data must be in an organized form for specific purposes like searching the genes involved in cancer. So here CGPD serves the same purpose. It also has some very useful tools like OCDC and other tools having some desktop applications of bioinformatics for the manipulation and management of data. It is a relational database with a very friendly interface providing its users a feasible approach to study cancer genes and proteins. Recent years have seen an explosion in the amount of available biological data. More and more renomes are being sequenced and annotated, and protein and gene interaction data are accumulating. Biological databases have been invaluable for managing these data and for making them accessible. Depending on the data that they contain, the databases fulfill different functions. Although they are architecturally similar, so far their integration has proved problematic [6, 13]. And for this purpose databases like CGPD should be built with their appropriate and definite purposes.

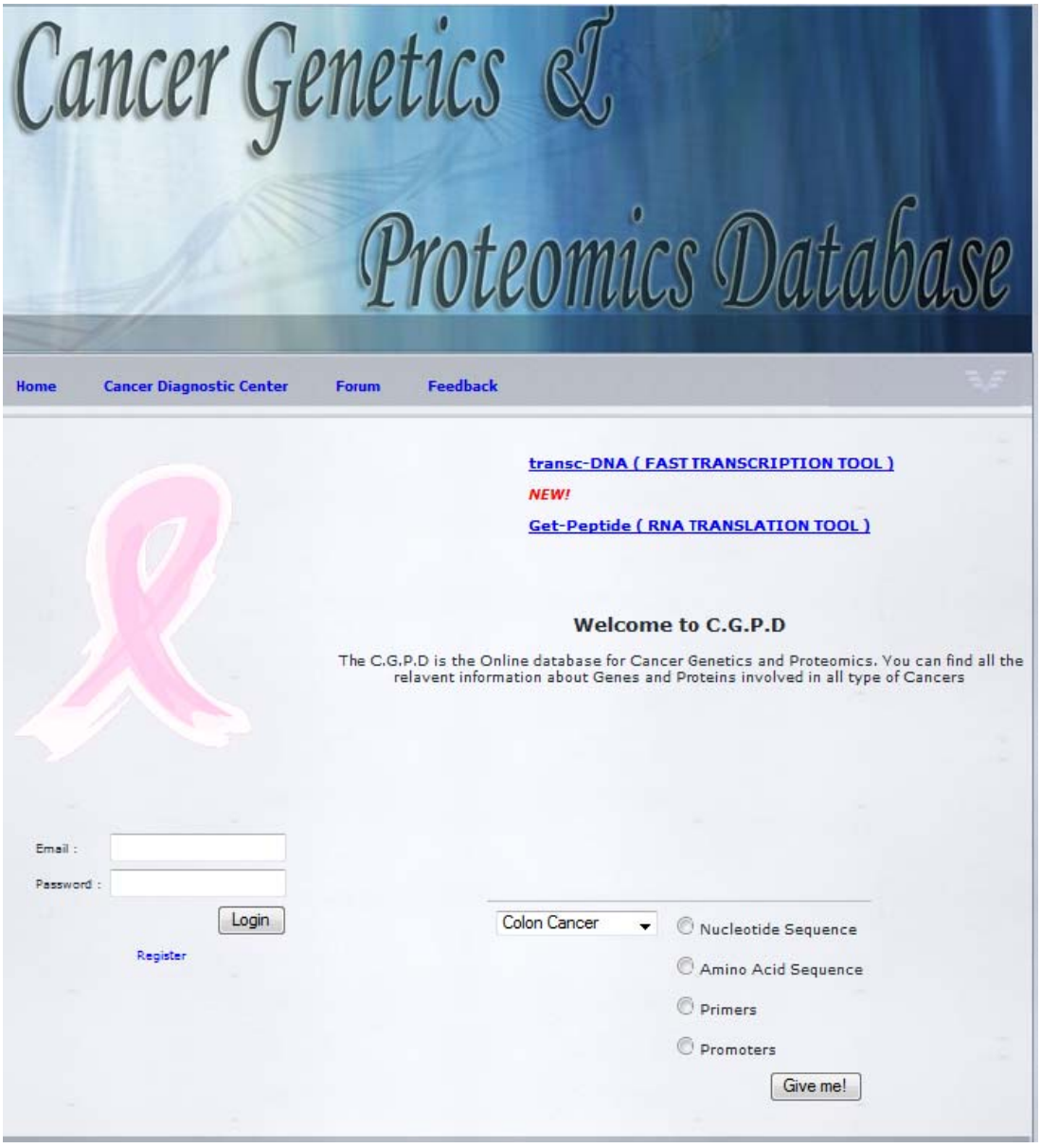

Fig. 1 Home page of CGPD

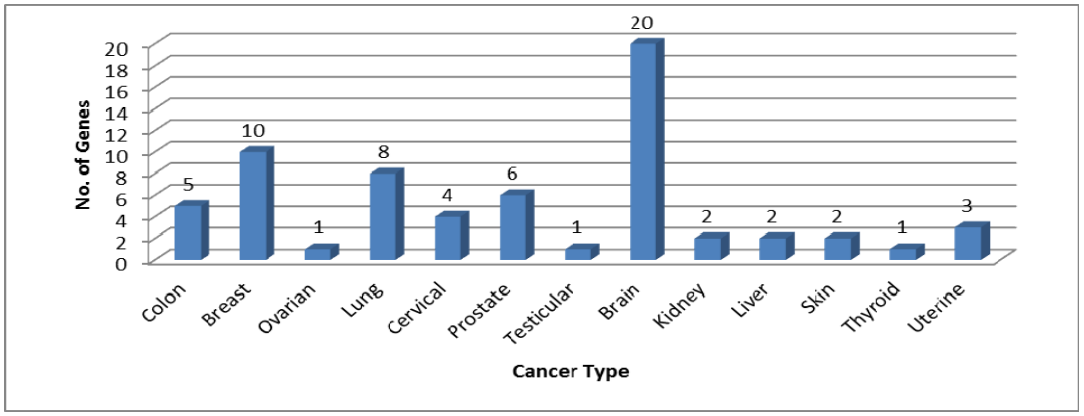

Fig. 2 Chart showing database statistics

Conclusion

CGPD provides appropriate dataset for cancer research. 72 genes were selected initially and extracted their sequence features like primers and promoters. All this information is available to user in very user-friendly manner with references. Moreover, protein sequences and its features are also made available. OCDC is very useful application, which can predict types of cancer on the basis of known symptoms. Decision making ability of this algorithm is quite efficient. Lot of more functionality will be embedded soon.

References

1. Aktipis C. A., R. M. Nesse (2013). Evolutionary Foundations for Cancer Biology, Evolutionary Applications, 6(1), 144-159, doi: 10.1111/eva.12034.
2. Benson D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers (2013). GenBank, Nucleic Acids Research, 41(D1), D36-D42, doi: 10.1093/nar/gks1195.
3. Demaria S., E. Pikarsky, M. Karin, L. M. Coussens, Y.-C. Chen, E. M. El-Omar, M. T. Lotze (2010). Cancer and Inflammation: Promise for Biological Therapy, Journal of Immunotherapy (Hagerstown, Md.: 1997), 33(4), 335-351, doi: 10.1097/CJI.0b013e3181d32e74.
4. Hofmann K., W. Stoffel (1993). TMBASE – A Database of Membrane Spanning Protein Segments, Biol Chem Hoppe-Seyler, 374: 166.
5. Hsieh P., K. Yamane (2008). DNA Mismatch Repair: Molecular Mechanism, Cancer, and Ageing, Mechanisms of Ageing and Development, 129(7-8), 391-407, doi: 10.1016/j.mad.2008.02.012
6. Magrane M., U. Consortium (2011). UniProt Knowledgebase: A Hub of Integrated Protein Data, Database: The Journal of Biological Databases and Curation, doi:10.1093/database/bar009.
7. Ogasawara O., J. Mashima, Y. Kodama, E. Kaminuma, Y. Nakamura, K. Okubo, T. Takagi (2013). DDBJ New System and Service Refactoring, Nucleic Acids Research, 41(D1), D25-D29, doi: 10.1093/nar/gks1152.
8. Peltomäki P., H. F. Vasen (1997). Mutations Predisposing to Hereditary Nonpolyposis Colorectal Cancer: Database and Results of a Collaborative Study, The International Collaborative Group on Hereditary Nonpolyposis Colorectal Cancer, Gastroenterology, 113(4), 1146-1158.

9. Reese M. G. (2001). Application of a Time-delay Neural Network to Promoter Annotation in the *Drosophila melanogaster* Genome, Computers & Chemistry, 26(1), 51-56.
10. Rozen S., H. J. Skaletsky (1998). Primer3, Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.
11. Tariq H., A. Muzammil, S. Khalid (2011). CAGE: A Database of Cancer Genes of Human, Mouse and Rat, Int J Bioautomation, 15(3), 179-182.
12. Tariq H., T. Niaz (2011). Orgene: An Organ Based Categorized Human Genome Database, Journal of Biochemical Technology, 3(2), 266-269.
13. Velankar S., Y. Alhroub, A. Alili, C. Best, H. C. Boutselakis, S., Caboche, G. J. Kleywegt (2011). PDBe: Protein Data Bank in Europe, Nucleic Acids Research, 39 (Database issue), D402-D410, doi:10.1093/nar/gkq985.
14. Ye J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, T. L. Madden (2012). Primer-BLAST: A Tool to Design Target-specific Primers for Polymerase Chain Reaction, BMC Bioinformatics, 13, 134, doi: 10.1186/1471-2105-13-134.

Muhammad Rizwan Riaz
E-mail: rizi.leo20@gmail.com

Muhammad Rizwan Riaz is a bioinformatician, graduated from Government College University, Faisalabad, in 2012. Now he is doing MS Bioinformatics at COMSATS Institute of Information Technology, Islamabad.

Attia Iram
E-mail: irramattia@yahoo.com

Attia Iram graduated from Government College University, Faisalabad, in 2012. Now she is doing M.Phil. in Biotechnology at National Institute of Biotechnology and Genetic Engineering, Faisalabad.