Academic Sciences International Journal of Pharmacy and Pharmaceutical Sciences

ISSN- 0975-1491 Vol 4, Suppl 3, 2012

Research Article

DEVELOPMENT OF CANCER DATABASE BY AMALGAMATION OF VARIOUS EXISTING DATABASE

SORABH SINGHAL[1], K. RAMANATHAN[2*], V. SHANTHI[3], R. RAJASEKARAN[2*]

[1]Program Analyst, Cognizant Technology Services, [2]Bioinformatics Division School of Bioscience and Technology, [3]Industrial Biotechnology Division, School of Bioscience and Technology, VIT University. Email: rrajasekaran@vit.ac.in and kramanthan@vit.ac.in

Received: 28 Jan 2012, Revised and Accepted: 07 Apr 2012

ABSTRACT

Cancer (medical term: malignant neoplasm) is a class of disease in which a group of cells display uncontrolled growth through division beyond normal limits, invasion that intrudes upon and destroys adjacent tissues, and sometimes metastasis, in which cancer cells spread to other locations in the body via lymph or blood. These three malignant properties of cancers differentiate them from benign tumors, which are self-limited, and do not invade or metastasize. Cell reproduction is an extremely complex process, which is normally tightly regulated by several classes of genes including oncogenes and tumor suppressor genes. Hereditary or acquired abnormalities in these regulatory genes can lead to uncontrolled cell growth, and the development of cancer. There are different of Cancer like bladder cancer, breast cancer, colon cancer, endometrial cancer, kidney cancer (renal cell), leukemia, lung cancer, melanoma, non-Hodgkin lymphoma, pancreatic cancer, prostate cancer, thyroid cancer. To find different types of genes for different types of cancer with other related genetic information is of great importance for clinicians and cancer immunologists to detect. In the present work, a database containing various information about the different type of cancer causing gene, NCBI ID, PDB IB, SWISS PROT ID, KEGG pathway and other information has been developed with Codon bias Concept. The main emphasis in this database is introduction of Codon Usage Bias which refers to differences in the frequency of occurrence of synonymous codons in coding DNA. The advantages of this database are that it is freely available and can be used even by a layman. It is easily up gradable and quite cost effective. The most important advantage is that it would be helpful in carrying out further research in developing countries like India where cancer has a high occurrence even in poorer sections of the society.

Keywords: Cancer, Database, Codon Bias Concept, Gene, KEGG pathway.

INTRODUCTION

Cancer is the uncontrolled growth of abnormal cells anywhere in a body. The abnormal cells are termed cancer cells, malignant cells, or tumor cells. Many cancers and the abnormal cells that compose the cancer tissue are further identified by the name of the tissue that the abnormal cells originated from (for example, breast cancer, lung cancer, colon cancer). Cancer is not confined to humans; animals and other living organisms can also be affected by cancer. Frequently, cancer cells can break away from the original mass of cells, travel through the blood and lymph systems, and lodge in other organs where they can again repeat the uncontrolled growth cycle. This process of cancer cells leaving an area and growing in another body area is termed metastatic spread or metastatic disease [1, 2].

A database is an organized collection of data for one or more purposes. A database related to cancer or any disease may be helpful either in the diagnosis or in the treatment of a particular type of cancer. As such there are many types of databases available in the field of cancer which is based on different type of methodology or the work performed. The databases currently available normally give us an idea about any particular type of gene causing cancer or about mRNA expression in the gene causing cancer. However, most of the these databases do not provide details of all the cancer types at one place, thus a need was felt to develop a database which can fulfill this gap. The database developed by the authors with Codon Bias Concept provides information about different types of cancers and their causing gene, NCBI ID, PDB IB, SWISS PROT ID, KEGG pathway and other information which can be retrieved easily.

MATERIALS AND METHODS

Implementation of Asp.Net and SQL

The database developed by the authors is a combination of ASP.Net and SQL. Data related to different cancer types was collected from various sources like NCBI, OMIM, PubMed, PDB etc. Typical data collected related to gene material obtained from various sources is listed at Table 1.

Table 1: Typical data collected for the study and its sources

Description	Sources
Genes Responsible	Data collected from various sources especially from cancer genetics web of cancer index [3]
Nucleotide Sequence	Data obtained from Nucleotide, National Center of Biotechnology Information(NCBI) [4]
Protein Sequence	Source was Protein, NCBI [5]
Swiss Prot ID	Data obtained from Uniprot protein knowledgebase [6]
Misses Variation	Mutdb was the source maintained by Mooney lab, Buck Institute [7]
PDB ID	Protein structures were available on RCSB Protein Data Bank [8]
Microarray Database	Data obtained from European Bioinformatics Institute (EBI) [9]
Codon Bias Concept	DAMBE software was used to calculate the concept
Micro RNA Database	Was obtained by miRBase maintained by University of Manchester [10]
KEGG Pathway	Different pathway images was collected from Kyoto Encyclopedia of Genes and Genomes (KEGG) [11]
OMIM	Source was Online Mendelian Inheritance in Man(OMIM), NCBI [12]
PubMed	Different research papers were referred from PubMed, NCBI [13]
Drugs and Medication	Various sources were there for finding various drugs and medication especially medlineplus [14, 15]

Development of Database

The authors preferred to use asp.net instead of PHP because it is easily available, upgradable and code generation is simpler than any other language. Generating the database through ASP.NET will create it in the form of website which can be accessed easily through internet. **ASP.NET** is a web application framework developed and marketed by Microsoft to allow programmers to build dynamic web sites, web applications and web services.

It contains all the HTML codes with the programming in both C/C++. Fig 1 & 2 depicts the code generation process of the database.

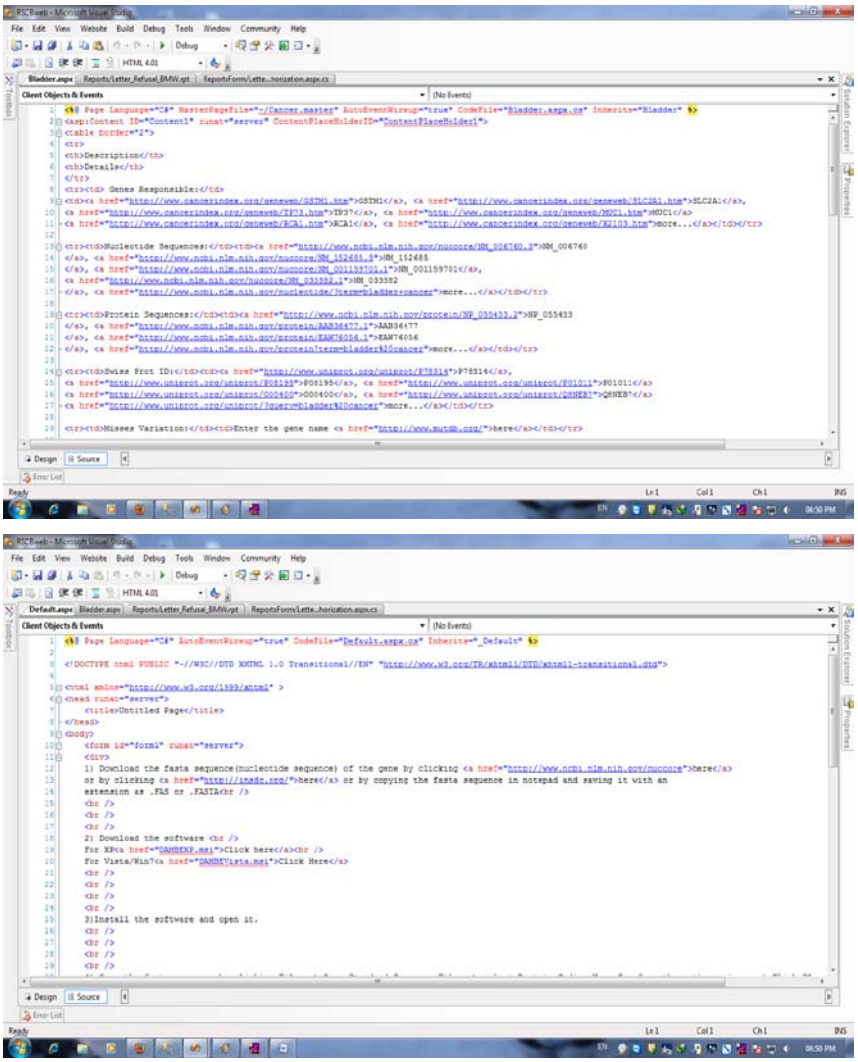

Fig. 1& 2: Code generation process of the database

HTML is written in the form of HTML elements consisting of *tags*, enclosed in angle brackets (like <html>), within the web page content. HTML tags normally come in pairs like <h1> and </h1>. The first tag in a pair is the *start tag*, the second tag is the *end tag* (they are also called *opening tags* and *closing tags*). In between these tags web designers can add text, tables, images, etc.

HTML tags which are used in this database are as follows:

- <html> - The HTML html tag is the container that contains all other HTML elements. All other HTML elements are nested between the opening <html> and </html> tags.
- <head> - The head can contain other HTML tags that provide information about the document such as title, description, keywords etc. Most of this information is not displayed in the browser (although the title usually appears in the browser's title bar) but can be useful for search engines etc. The head tag is placed between the <html> and the <body> tags.
- <body> - The HTML body tag is used for indicating the main content section of the HTML document. The body tag is placed between the </head> and the </html> tags.
- <table> - The HTML <table> tag is used for defining a table. The table tag contains other tags that define the structure of the table.
- <tr> - The HTML <tr> tag is used for specifying a table row within a table. A table row contains one or more td and/or the tags which determine individual cells.
- <td> - The HTML <td> tag is used for specifying a cell (or table data) within a table.
- <a> - The HTML a tag is used for creating a hyperlink to either another document, or somewhere within the current document.
- <form> - The HTML form tag is used for declaring a form. The <form> tag is used in conjunction with form-associated elements. To create a form, you can nest form-associated elements inside the opening/closing <form> tags.

SQL is used at backhand which contains the code for retrieving the information of the particular cancer type. There are several SQL commands used like:

The **SQL SELECT** clause selects data from one or more database tables and/or views. In its basic form the SQL SELECT syntax looks like this: Syntax: FROM Table1 SELECT ColumnName1, ColumnName2,..., The **SQL ORDER BY** clause defines in what order to return a data set retrieved with a SQL SELECT statement. Here is an example of using SQL ORDER BY to order the rows in our Weather table by city: Syntax: SELECT * FROM <table name> ORDER BY <column name>

RESULT AND DISCUSSION

The main page of the database contains the list of options for different types of cancer in the form of a drop down list. Choose the type of cancer for which the information has to be retrieved. After clicking, the page will directly come to specific cancer page. From

there, the information can be seen just by clicking on the part for which information is needed. For the codon bias concept or codon usage bias, a tutorial has been incorporated in which various steps are detailed to calculate the codon bias concept. Database on different types of Cancer has been developed which can be seen and can be operated as shown in Fig 3, 4, 5.

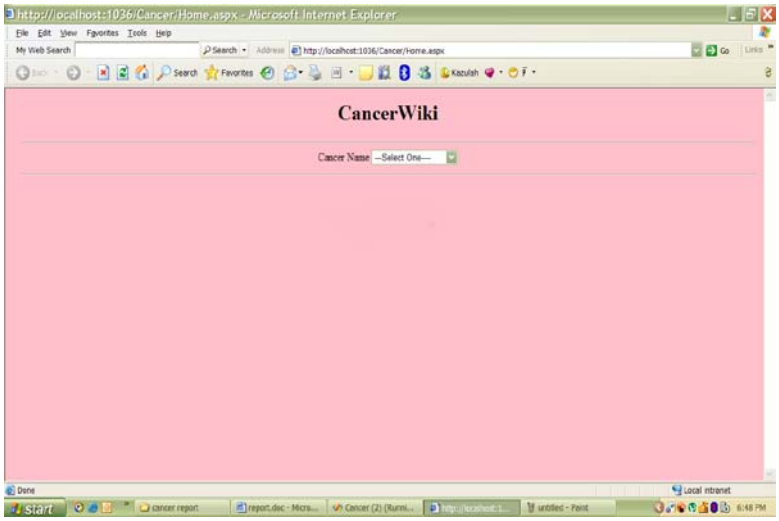

Fig. 3: Main Page of the database

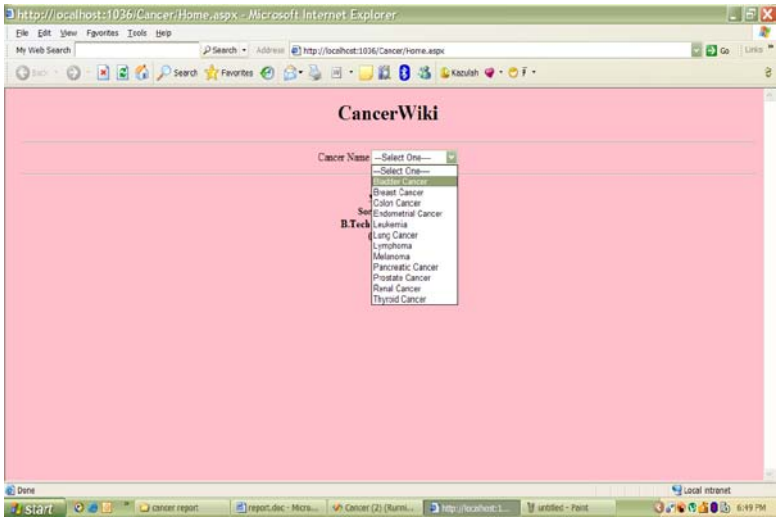

Fig. 4: Selection of the type of Cancer

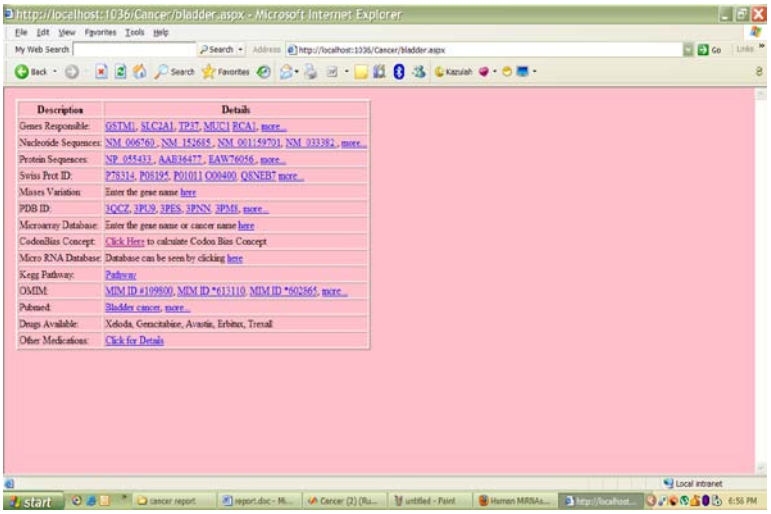

Fig. 5: Page showing the details of the particular type of Cancer

There are many different databases in cancer having unique features like different gene material, mRNA, microarray expression etc. These data bases contain gathered information from various sources written in a computer code. CaSNP is a database for interrogating copy number alterations of cancer genome from SNP array data have presented the CaSNP database for identifying and visualizing CNAs in cancers at any specific region within the human genome. CaSNP stores pre-computed from other software or methods. Besides the tabular display, the heat map view displays SNP copy numbers in colors, enabling users to intuitively and comprehensively visualize the results and facilitating finding novel CNA regions in subset of samples. Besides, we provided a scenario of using CaSNP to explore cancer biomarkers or genes through a meta-analysis, and proved CaSNP's ability in suggesting novel oncogenes/tumor suppressors, whether a protein coding gene or an ncRNA [16].

Another database is dbDEMC: a database of differentially expressed miRNAs in human cancers includes miRNAs that have the expression information in different cancers determined by statistical analysis of microarray data. Analysis results from different data sources indicated that the expression levels of a specific miRNA are often different and even contradict with each other among many experiments. The different cancer subtypes, cell lines and the experiment platforms used may be the explanation for these differences. In this sense, the integration of the data from different source provides more rich and reliable miRNA expression information in cancers [17].

Database for cancer immunology is a database which aims to develop a tumoral microenvironment (TME) database for storing and maintaining all the data which are arising from different immunological experiments. This information comprises pathological information, patient related information, experiment related information, and data from clinical treatments. This database was especially developed for tumor microenvironment related data, but the flexible design suggests that it can be used for other cancer related information as well [18].

Another database of mRNA gene expression profiles of multiple human organs constructed a gene expression database capturing the mRNA transcriptional levels for 19 different organs from 158 normal human tissues from 30 donors. The database builds on the published studies by an increased number of samples, an increased number of biological repeats for the same tissue, an increase in the number of detected unique clones, and the different array technology used (cDNA). Biological repeats were hybridized individually, allowing estimating intra organ variation of transcript levels. In the analysis, 18,927 unique genes passed quality filtering [19].

Besides the above, There are also many other databases which provide different types of information on cancer like The EUROCARE-4 which is a database on cancer survival in Europe [20]; National Virtual Specimen Database for early cancer detection [21]; Cancer Resource Database, a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge [22]; The Capsure Database, a methodology for clinical practice and research in prostate cancer [23]; Database of Cancer Induction by low-dose radiation in mammals [24]; Adverse Event Reporting compared with sponsor database for cancer clinical trials [25]; Gene Expression Database for the molecular pharmacology of cancer [26]; ONCOMINE: a cancer microarray database and integrated data-mining platform [27]; effect of anti-inflammatory drugs on overall risk of common cancer: case control study in general practice research database [28]; RTCGD: retroviral tagged cancer gene database [29]; A database of protein expression in lung cancer [30] and many more.

Codon Bias Concept or Codon usage bias refers to differences in the frequency of occurrence of synonymous codons in coding DNA. A codon is a series of three nucleotides (triplets) that encodes a specific amino acid residue in a polypeptide chain or for the termination of translation (stop codons). There are 64 different codons (61 codons encoding for amino acids plus 3 stop codons) but only 20 different translated amino acids. The overabundance in the number of codons allows many amino acids to be encoded by more than one codon. Because of such redundancy it is said that the genetic code is degenerate. Different organisms often show particular *preferences* for one of the several codons that encode the same amino acid- that is; a greater frequency of one will be found than expected by chance.

Different factors have been proposed to be related to codon usage bias, including gene expression level (reflecting selection for optimizing translation process by tRNA abundance), %G+C composition (reflecting horizontal gene transfer or mutational bias), GC skew (reflecting strand-specific mutational bias), amino acid conservation, protein hydropathy, transcriptional selection, RNA stability, optimal growth temperature and hyper saline adaptation.

Codon Usage bias is calculated by the concept of Codon Adaptation Index (CAI). The index uses a reference set of highly expressed genes from a species to assess the relative merits of each codon, and a score for a gene is calculated from the frequency of use of all codons in that gene. The index assesses the extent to which selection has been effective in molding the pattern of codon usage. In that respect it is useful for predicting the level of expression of a gene, for assessing the adaptation of viral genes to their hosts, and for making comparisons of codon usage in different organisms. The index may also give an approximate indication of the likely success of heterologous gene expression. CAI in the present study is calculated by the use of DAMBE software which gives a significant value for a gene sequence. CAI values range from 0 to 10, with higher values indicating a higher proportion of the most abundant codons [31, 32].

In the present database, more emphasis has been given on diversity of different types of cancer by giving main information about them at one place like genes responsible or their protein id or swiss prot id etc. This will enable the researchers to find all the information needed at one place without making much effort. Main highlight of the database is the concept of codon usage bias which is new for any database based on cancer and can be useful in research of cancer. Application of codon usage bias will help in predicting the level of expression of a gene, assessing the adaptation of viral genes to their hosts and for making comparisons of codon usage in different organisms. The index may also give an approximate indication of the likely success of heterologous gene expression. In future this database can be further integrated by microarray data which can further help in other genetic information of different types of cancer.

CONCLUSION

To find different types of genes for different types of cancer with other related genetic information is of great importance for clinicians and cancer immunologists to detect. To achieve this goal, we developed a database that allows for an effective retrieval of genetic information in different cancers. This database brings together different data in order to investigate different types of cancer and their genetic information and other relevant information. The advantages of this database are that it will be freely available and can be used even by a layman. It is easily upgradable and quite cost effective. The most important advantage is that it would helpful for the developing countries like INDIA where cancer id a very prominent disease. Because of the many unique features of this database, it will greatly facilitate the identification of cancer-related information and the discrimination and determination of different cancer types.

ACKNOWLEDGEMENT

The authors express deep sense of gratitude to the management of Vellore Institute of Technology for all the support, assistance, and constant encouragement to carry out this work.

REFERENCES

1. Chapter in a book: Moscow JA and Cowan KH. In: Goldman L, Ausiello D, editors. Biology of cancer. 23rd ed. Philadelphia: Saunders Elsevier; 2007. chap 187.
2. Thun MJ. In: Goldman L, Ausiello D, editors. Epidemiology of cancer. 23rd ed. Philadelphia: Saunders Elsevier; 2007. chap 185.
3. www.cancerindex.org

4. www.ncbi.nlm.nih.gov/nuccore
5. http://www.ncbi.nlm.nih.gov/protein
6. Magrane M and Consortium U UniProt Knowledgebase: a hub of integrated protein data. Database; doi:10.1093/database/bar009
7. Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, Heiland R, Mooney SD MutDB: update on development of tools for the biochemical analysis of genetic variation. Nucleic Acids Research 2008; 36 Suppl 1: D815-19.
8. Bermen HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE The Protein Data Bank. Nucleic Acids Research 2000; 28(1): 235-42
9. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S and Brazma A ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acid Research 2003; 31(1):68-71.
10. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ miRBase: tools for microRNA genomics. Nucleic Acid Research 2008; 36(Database Issue):D154-58
11. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H and Kanehisa M KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 1999; 27(1): 29-34.
12. http://www.ncbi.nlm.nih.gov/omim
13. http://www.ncbi.nlm.nih.gov/pubmed
14. http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002267/
15. www.nlm.nih.gov/medlineplus
16. Cao Q, Zhou M, Wang X, Meyer CA, Zhang Y, Chen Z, Li C and Liu XS CaSNP: A database for interrogating copy number alterations of cancer genome from SNP array data. Nucleic Acids Research 2011; 39(Database issue): D968–74.
17. Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, Yao L, Zhang Y, Miao R, Cao Y, Zhao Y, Zhong Y, Zhao H dbDEMC: a database of differentially expressed miRNAs in human cancers. BMC Genomics 2010; 11 Suppl 4: S5.
18. Institute for Biomedical Engineering, Graz University of Technology, Graz, Australia, Institut National de la Sant´e et de la Recherche M´edical Unit´e 255, Centre de Recherches Biom´edicales des Cordeliers, Paris, France, Database for cancer immunology
19. Son CG, Bilke S, Davis S, Greer BT, Wei JS, Whiteford CC, Chen QR, Cenacchi N and Khan J Database of mRNA gene expression profiles of multiple human organs, Genome Research 2009; 15(3): 443-50.
20. Angelis RD , Francisci S, Baili P, Marchesi F, Roazzi P, Belot A, Crocetti E, Pury P, Knijn A, Coleman M, Capocaccia R The EUROCARE-4 database on cancer survival in Europe: Data standardisation, quality control and methods of statistical analysis. European Journel on Cancer 2009; 45(6):909-30.
21. Crichton D, Kincaid H, Kelly S, Thornquist M, Johnsey D, Winget M, Srivastava S A National Virtual Specimen Database for Early Cancer Detection. IEEE conference on Computer-based medical systems 2003; pp 117-23
22. Volume with Supplement: Ahmed J, Meinel T, Dunkel M, Murgueitio MS, Adams R, Blasse C, Eckert A, Preissner S and Preissner R Cancer Resource: A comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge, Nucleic Acids Research 2010; 39 Suppl 1: D960-D967.
23. Lubeck DP, Litwin MS, Henning JM, Stier DM, Mazonson P, Fisk R, Carroll PR The capsure database: a methodology for clinical practice and research in prostate cancer. Urology 1996; 48(5):773-77.
24. Duport P A database of cancer induction by low-dose radiation in mammals: overview and initial observations. International Journal of Low Radiation 2003; 1(1): 120-31.
25. Scharf O and Colevas AD Adverse Event Reporting in Publications Compared With Sponsor Database for Cancer Clinical Trials. Journel of Clinical Oncology 2006; 24(24):3933-38.
26. Scherf U, Ross DT, Waltham M, Smith LH, Lee1 JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO and Weinstein JN A gene expression database for the molecular pharmacology of cancer. Nature Genetics 2000; 24(3): 236-44.
27. Rhodes DR, Yu J, Shanker K, Deshpandez N, Varambally R, Ghosh D, Barrette T, Pandeyb A and Chinnaiyan AM ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform. Neoplasia 2004; 6(1):1-6.
28. Langman MJS, Cheng KK, Gilman EA, and Lancashire RJ Effect of anti-inflammatory drugs on overall risk of common cancer: case-control study in general practice research database, British Medical Journel 2000; 320(7250):1642-46.
29. Akagi K, Suzuki T, Stephens RM, Jenkins NA and Copeland NG RTCGD: Retroviral tagged cancer gene database. Nucleic Acids Res. 2004; 32(Database issue): D523-27.
30. Oh JMC, Brichory F, Puravs E, Kuick R, Wood C, Jean Rouillard JM, Tra J, Kardia S, Beer D, Hanash S A database of protein expression in lung cancer. Proteomics 2001; 1(10):1303-19.
31. Palidwor1 GA, Perkins TJ, Xia X A General Model of Codon Bias Due to GC Mutational Bias. PLoS One 2010; 5(10):1-11.
32. Supek F and Vlahovic K INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics 2004; 20(14):2329-30.