J Biochem Tech (2011) 3(2): 266-269
ISSN: 0974-2328

Orgene: An organ based categorized human genome database

Hassan Tariq, Tariq Niaz

Received: 17 August 2010 / Received in revised form: 24 August 2010, Accepted: 28 June 2011, Published online: 01 July 2011,

Abstract

OrGene is an organ based categorized human genome database developed for the categorization of human genome on the basis of organs. The function of the proteins encoded by these genes is also made available. OrGene will expand in future to other organs, genes and cover more useful information of other species. Primers provided for each gene, with their features and conditions given to facilitate the researchers, are useful in PCR amplification, especially in cloning experiments. 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.orgene.pakbiz.org.

Key words: Organ, Database, Human genome

Introduction

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 are crucial to our understanding of life on earth as complex systems. Information stored must be correct, complete and internal relationships among elements easy to navigate. Computational tools and databases are essential to the management and identification of patterns among database elements that reflect biological systems (Buehler and Rashidi, 2005).

Hassan Tariq

Department of Bioinformatics and Biotechnology, Government College University, Faisalabad, Pakistan

Tel.: 0092 3447797859
Email: hassantariq9@yahoo.com

Tariq Niaz

Ayub Agriculture Research Institute, Faisalabad, Pakistan

Historically, databases have arisen to satisfy diverse needs, whether it address a biological question of interest to an individual scientist, to better serve a particular section of biological community, to coordinate data from sequencing projects, or to facilitate drug discovery in pharmaceutical companies. The workhorses of modern biology, databanks now number in hundreds and house information of all kinds. Indeed, a special isssue of the Journal Nucleic Acid Research, with an online molecular biology database catalogue (Baxevanis, 2003), is devoted every year to the documentaton of ongoing and new databases initiatives, and there is even a database of databases, DBCat (Discala et al., 2000).

The databases in common use today each have different objective. Primary DNA sequence repositories, such as EMBL, GenBank and DDBJ, are those that attempt to keep a comprehensive record of all sequenced DNA as it become available. As primary databases are all-inclusive, they are inevitably fairly shallow in terms of the information they contatin. By contrast, secondary databases aim to combine information from several different primary database entries and to provide added information not present in their primary sources.

Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high throughput experiment technology, and computational analyses. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures (Altman, 2004).

Relational database concepts of computer science and Information retrieval concepts of digital libraries are important for understanding biological databases. Biological database design, development, and long-term management are a core area of the discipline of Bioinformatics. Data contents include gene sequences, textual descriptions, attributes and ontology classifications, citations, and tabular data. These are often described as semi-structured data, and can be represented as tables, key delimited records, and XML structures. Cross-references among databases are common, using database accession numbers (Bourne, 2005).

The most important product of the sequencing of a genome is a complete, accurate catalogue of genes and their products, primarily messenger RNA transcripts and their cognate proteins (Shoemaker *et al.*, 2001). At present, it is estimated that the human genome encodes for <30,000 genes. However, it has been suggested that only a fraction, perhaps 10,000 genes, are actively transcribed in normal cell processes (Lander *et al*., 2001 and Venter *et al*., 2001). The human genome resources at NCBI allow complete access to varied information regarding the human genome. These resources include information on genes and human health, the nucleotide genome and clone registry, navigatable maps based and different marker types (physical and linkage maps), transcribed sequences allowing a functional view of the genome, cytogenetics and comparative genomes (Buehler and Rashidi, 2005).

As the annotation of the information present in these online resources increasing exponently, data searching or data retrieval is a major problem for most of the users. Especially those who are not so much familiar with these databases are really suffering. So day by day the number of secondary databases is increasing. These databases provide quick access to the required information. 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 categorize the present human genome information which we find at public databases like NCBI, UniProt etc.

Keeping in view the sequence retrieval difficulty, we planned to develop a user friendly categorized human genome database with following objectives

- Organ based categorization of gene expression
- Primer designing for the amplification of expressed genes

Material and Methods

Data Collection

In the present study we required genes and their relevant proteins which are expressed in each organ. So to collect our desired data we designed a methodology.

- Searched for human proteins and their tissue specificity from UniProt Knowledgebase
- Searched for the protein coding genes in NCBI's Genbank
- Searched for the nucleotide sequences of the relevant genes
- Designed primers for these nucleotides by using Primer3

To analyze the designed primers we used NetPrimer . There are certain things that we considered while analyzing primers.

- Primer-Dimer formation
- Secondary Structures in Primers

Sections

There are two sections, that is, organ and gene section. Both of them include three basic functionalities. First, we can add a new organ/gene, second, we can edit the current information and third, we can delete as existing record.

Results

Currently, the database includes almost all of the major organs like brain, heart, liver, kidney, stomach etc. In the present study, we included complementary DNA nucleotide sequences for each gene.

Genes

Sr. No.	Accession #	Gene Name	Organ	Defination	View Gene Details
1	HSAG_024529	CDC73	Adrenal Glands	Homo sapiens cell division cycle 73, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae) (CDC73), mRNA.	view
2	HSAG_000198	HSD3B2	Adrenal Glands	Homo sapiens hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2 (HSD3B2), mRNA.	view
3	HSAG_003167	SULT2A1	Adrenal Glands	Homo sapiens sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone (DHEA)-preferring, member 1 (SULT2A1), mRNA.	view
4	HSAG_004689	MTA1	Adrenal Glands	Homo sapiens metastasis associated 1 (MTA1), mRNA.	view
5	HSAG_006196	PCBP1	Adrenal Glands	Homo sapiens poly(rC) binding protein 1 (PCBP1), mRNA.	view
6	HSAG_014790	JAKMIP2	Adrenal Glands	Homo sapiens janus kinase and microtubule interacting protein 2 (JAKMIP2), mRNA.	view
7	HSAG_017680	ASPN	Adrenal Glands	Homo sapiens asporin (ASPN), mRNA.	view
8	HSAG_001007253	ERV3	Adrenal Glands	Homo sapiens endogenous retroviral sequence 3 (includes zinc finger protein H-plk/HPF9) (ERV3), mRNA.	view
9	HSAG_144707	PROM2	Adrenal Glands	Homo sapiens prominin 2 (PROM2), mRNA.	view

Figure 1. Showing different genes in a particular tissue or organ

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 in a particular tissue / organ. These features are the result of our flexible database design. Followings are the salient features of the OrGene:

Data searching

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

- Data searching by Search field

  OrGene, 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 organs.

- Data searching through Navigation

  OrGene, provides the facility for the users to search their relevant data by navigating the database. Whenever we click specie, a list of organs will appear. From this list of organs we can choose one organ and a list of genes present in it will appear, we can view its details by further navigating into it as shown in the figure 1.

- Easy and fast access to the information

  We can get access to data in no time. Data searching is so easy in OrGene that even a new user can search through it with almost no difficulty.

- 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 as shown in figure 2.

Discussion

The human genome project has had an impact on both biological research and its political organization. The project has generated both anticipated and novel information; in the later category are the description of the unusual distribution of genes, the prevalence of non-protein-coding genes, and the extraordinary evolutionary conservation of some regions of the genome (Little *et al*., 2005). This approach is new to the most of people. We made use of the richness of the annotation of human genes and proteins in many different databases and literature. These resources also use software for Customized Annotation of Genome Regions (Huntley *et al*., 2003). Beside all the information related to genes and proteins, we also designed primers for the nucleotide sequences of these genes.
At present there are many databases having huge amount of data related to a large number of species such as GenBank (Benson *et al*., 2008). But the purpose of current project is to sort out the information relevant to humans from the current databases and arrange them in such a way that anyone can access it very easily. Keeping in view the fact that it is a rather challenging task to categorize the information related to human genome, a database was developed on the basis of organs. **T**he Human Genome Project has increased the rate of DNA sequence accumulation to the point where information management has become a formidable task. The

Symbols	Descriptions
S	The position of the 5' base of the primer. For a Left Primer or Hyb Oligo this position is the position of the leftmost base. For a Right primer it is the position of the ***rightmost*** base.
L	The length of the primer or oligo.
T	The melting temperature of the primer or oligo.
G	The percent of G or C bases in the primer or oligo.
A	The self-complementarity score of the oligo or primer (taken as a measure of its tendency to anneal to itself or form secondary structure).
3'	The 3' self-complementarity of the primer or oligo (taken as a measure of its tendency to form a primer-dimer with itself).

Figure 2. Showing primers designed for the specific sequence

central repositories for this avalanche of data, GenBank, EMBL (European Molecular Biology Laboratory), and DDBJ (DNA Data Bank of Japan), continue to accumulate DNA sequences at an unprecedented rate. For example, the total number of nucleotides stored in the GenBank database more than doubles every 18 months (Benson *et al.,* 2008). Then the rich amount of annotation relevant to gene, its structure and its expression has added up this difficulty. So there was an intense need of a specialized human database.
It has been the highlight of the current project that it facilitates the data searching not only for advance users but also for beginners. For advance users we built a direct search method and for the beginners we built a navigation searching option. The main purpose of both of these options is to make the information retrieval fast and easy which is difficult and sometimes too much slow in case of public databases such as NCBI's Gene and Nucleotide databases.

The usefulness of organ based categorization of human genome is that it can help researchers who are working on the molecular basis of a particular disease related to a particular organ. The functional annotation of human genome by artificial transcription factor-based random genome perturbation method would provide genomic tool for annotation and classification of genes in the human genome and those of many other organisms (Lee et al., 2003). Thus, it will be helpful in designing better medicines or drugs too. We can design drugs which can be sight specific or targeted by understanding the molecular basis of the site of action. This will not only make molecular data more accessible for the researchers but also more understanding of the molecular basis of a disease.

References

Altman, RB (2004) Building successful biological databases". Brief. Bioinformatics, 5(1):4–5

Baxevanis AD (2003) The Molecular Biology Database Collection: 2003 update. Nucleic Acid Res 31:1-12

Benson DA, Karsch-Mizrachi et al (2008) Genbank, Nucleic Acids Res 36: D25-30

Bourne, P (2005). Will a biological database be different from a biological journal. PLoS Comput Biol 179–81

Discala C, Benigni X, Barilbt E, Vaysseix G (2000) *DBCAT:* A catalogue of 500 biological databases. Nucleic Acid Res 28(1):8-9

Huntley D, Hummerich H, Smedley et al (2003) GANESH: Software for Customized Annotation of Genome Regions. Genome Res 13:2195-2202

Lander ES, Linton LM et al (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Lee DK, Park JW et al (2003) Toward a functional annotation of the human genome using artificial transcription factors. Genome Res 13:2708-2716

Little PFR (2005) Structure and function of the human genome. Genome Res 15:1759-1766

Shoemaker DD, Schadt EE et al (2001) Experimental annotation of the human genome using microarray technology. Nature 409:922-927

Venter J, Adams MD et al (2001) The sequence of the human genome. Science 291(5507):1304–1351