BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database

Volume 6(3)

A database of six eukaryotic hypothetical genes and proteins

Katika Prabhakara Surya Adinarayana[1]*, Tanuka Sai Sravani[2], Chamarthi Hareesh[2]

[1]Department of Anatomy, Andhra Medical College, Visakhapatnam – 530001, India; [2]Bio-Lab, Research Gateway for Biosciences (RGBio), 47-3-30, Dwaraka Nagar, 5th Lane, Visakhapatnam – 530016, India; Katika Prabhakra Surya Adinarayana - Email: kpsanarayana@rediffmail.com; Phone: +91-9848443943; *Corresponding author

Received February 18, 2011; Accepted March 23, 2011; Published April 22, 2011

Abstract:
Assigning functions to proteins of unknown function is of considerable interest to the proteomic researchers as the genes encoding them are conserved over various species. Here, we describe HypoDB, a database of hypothetical genes and proteins in six eukaryotes. The database was collected and organized based on the number of entries in each chromosome with few annotations. Hypothetical protein database contains information related to gene and protein sequences, chromosome number and location, secondary and tertiary structure related data.

Availability: The database can be accessed at http://www.trimslabs.com/database/hypodb/index.html

Background:
Data pertaining to hypothetical proteins expressed in many eukaryotes would help researchers to search for potential proteins of interest with unknown functions **[1]**. However, many such hypothetical protein encoding genes are conserved over various species, which can be revealed from comparative genome analysis **[2-4]**. To predict a function for each of the protein coding regions, a comparative sequence analysis against all functionally elucidated sequences in protein sequence databases would reveal the necessary information for sequence retrieval, functional prediction and homologous sequences **[5, 6]**, further which, multiple sequence alignments would reveal possible functional insights on cellular process or biological function **[9, 10]**. A hypothetical protein showing one or more significant structural homolog is predicted to have similar molecular properties **[7, 8]**. On the other hand, conserved hypothetical proteins are found in both prokaryotes and eukaryotes, the function of which can be predicted by domain homology searches, secondary, tertiary structure predictions, and gene annotations. Hence, data on hypothetical proteins from NCBI database was collected and organized in the form of a database using html and javascript. The database contains information regarding gene/protein sequences, chromosome number and location, secondary and tertiary structure information, ProFunc server data, primary analysis tools (mol.wt, ionization constant etc.), expression levels of the sequences and related data.

Methodology:
Construction of database:
HypoDB is constructed using html and JavaScript and can be accessed at http://www.trimslabs.com/database/hypodb/index.html. Data were collected from NCBI GenBank and SWISS-PROT databases. HypoDB includes hypothetical proteins of 8 organisms. The complete list of organisms with their scientific and general names was given in **Table 1 (see Supplementary material)**. They are provided as records and organized to simplify the task of finding relevant data for proteins in the related organism. In order to make the database available online, HTML pages are constructed using Javascript. Hypothetical protein database contains information on hypothetical gene and protein sequences in the form of records. The data were categorized based on the number of hypothetical genes and proteins in each chromosome of six eukaryotes. Each record when accessed returns the nucleotide and protein sequence and annotation such as accession numbers, source organism and chromosome number. An example of an entry in human chromosome 1, LOC100131311 is given in **Table 2 (see Supplementary material)**.

Utility:
The database is of much utility to researchers working in the fields of functional proteomics and genomics. Such data on hypothetical genes and proteins represents a prominent research area to annotate the genes of interest and predict functional regions. However, given the insight into the technological advances in bioinformatics, function prediction and assigning functionally important sites within the protein sequence is advantageous to identify the mutations that might have resulted to unknown function of the particular gene. Therefore, this database of hypothetical genes and proteins would be a useful source to study or predict the functional regions of a protein. Data was segregated based on the number of entries in each chromosome of six eukaryotes, provided with an easy way of access.

References:
[1] Doolittle WF. *Trends Genet.* 1998 **14**: 307 [PMID: 9724962]
[2] Lorbach E *et al. Biol Chem.* 1998 **379**: 1355 [PMID: 9865609]
[3] Wilson CA *et al. J Mol Biol*. 2000 **297**: 233 [PMID: 10704319]
[4] Zarembinski TI *et al. Proc Natl Acad Sci U S A.* 1998 **95**: 15189 [PMID: 9860944]
[5] Eisenstein E *et al. Curr Opin Biotechnol.* 2000 **11**: 25 [PMID: 10679350]

[6] Uchiyama I. *Nucleic Acids Res.* 2003 **31**: 58 [PMID: 12519947]
[7] Kinoshita K & Nakamura H. *Protein Sci.* 2003 **12**: 1589 [PMID: 12876308]
[8] Liu Z *et al*. *BMC Genomics*. 2008 **9**: 509 [PMID: 18973670]
[9] Discala C *et al. Nucleic Acids Res.* 2000 **28**: 8 [PMID: 10592168]
[10] Ebihara A *et al. Protein Sci*. 2006 **15**: 1494 [PMID: 16672237]

Edited by P Kangueane
Citation: Adinarayana *et al.* Bioinformation 6(3): 128-130 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: List of organisms selected in the study

Scientific Symbol	Scientific Name	General name
Ha	*Homo sapiens*	human
Ms	*Mus musculus*	laboratory mouse
Oa	*Ornithorhynchus anatinus*	Duck billed Platypus
Md	*Monodelphis domestica*	Gray Short-tailed Opossum
Tg	*Taeniopygia guttata*	Zebra
Ec	*Equus caballus*	Horse
Pn	*Pan troglodytes*	chimpanzee
Ss	*Sus scrofa*	Pig

Table 2: An example of human chromosome 1, LOC100131311 entry in database

Hypothetical Protein	**LOC100131311**
Gene Description	hypothetical LOC100131311
Gene Type	protein coding
Chromosome	1
Locus	1q21.3
Nucleotide Seq	1344bp NC_000001.10 GI:224589800 >ref|NC_000001.10|:150550796-150552139 Homo sapiens chromosome 1, GRCh37 primary reference assembly ATGCAGCTTTCTTGGTTTATGGTCTTCAAGTGTTTAGCCACAAAGGCACCAAAAGAAATGAGAGTCACAA TCCTGCCCCAGTTTGTTACGCCGTCGCTGAAAACATGGATCATCACTCGAGACAACGATTTCACATCGTC TTCGTTTTTGATGTCCAGTTTCCGAAGCATGCCTGAGAAAGAAAAGCATGCAGGTCCTCACGGCCTCCTT TGTCTAAACCGGCGCAAGATTCGCCTGCCCACCCGCGGCGGGAAAATCGCTACTGGGATTTACAGAACTC AGGTTGACCCCACTTGAAATTGACATCCCACCCTTTCCGGTCTTTGAACAAGAGCTGCCATTTCCAAAAG AATCAAGATGGGCGGAAACAATGACTCATGGCCAGAATATTCTGGCTTCAGGAATAGGATGAGACACGTT TCAACACTGACTCGTTTCGGTTTCCAACCCACCCTTGGCGGGTGAGTCCGGGGAGAGATGGAAAAAAGGG AGTGAGGCCTTGGCGATTAATGAACCCCCTTACCTTGGAAGGCCGTCTCGTGGTTGCGCTGCACGCCATC CCCAACCCGTCGTAAGGTCTCCAGCGCCTTCCTGCTGGTGGCCCCAGACCTGCCCATTGGCTTTGTGTCC TTGGCGCCGGTGGCCTGCTCCCGAAGGTACCGAGAGATAATCTCCAGCGACTGCCGGTACAACTCGTCCT CCTCCTCCTCTGCTGGCGGCGGCGTCGAGGGTAGTGACCCGTCCGTACTGGTGTTATTACCAGATTCCCC GACCAACTCCAGCAGCGGCAGGACAGCCGGCCGCTTCCCGAGAGGCTCCGGCTCGTACCCGTCCAGCTCC TCTTCGGGCGACATGATGGCGTCAGCGGCCGGGGCTTCCATCTCCTCAAGCGGCGCCGCGCGGCGGGTGG GCGCGAAGAAAAGCAGCCTCGCGGGGGTCGCGGTGACGTCGGGGACCTCGGCGCCAATGGGCGGCGGCCG CGCGACCCTCCGGGAGTCTGGCGTGAGGGTGGACGGGGGGCTTGCGCCGGCGCTTCCGCCAATCACCGCG CCGGCCTCCCCTCCCCCTATCTCTCGCCGGGCCGAGGCCTCCTTCTCCGTAGCCAAAAGTCGCCCTCCCG GGCGGGTGGCGCCGCCGCTGCCGGCCCCCAAGCCGGCCCCCCCACAGTAGAGGTTGAGTCCGATTACCGC GTTTCTTTTGAGGCCAAACATTGCCAGTCGCCGCCGCCGCCTGGCTGAGAAAACTGGGGAAGACCCCGAC TCCTTACTGGAAGGAAGCGGAAGTGAGAAGTGGCGAGCAGCTCCTTTATCACGGTTTTAGGGCGGCCAGT CCTACGGGGTGGCG
Coding Region	join(1..282,532..906,1008..1238) "GeneID:4170"
Protein Seq	295 aa XP_001718239.1 GI:169162655 >gi|169162655|ref|XP_001718239.1| PREDICTED: hypothetical protein [Homo sapiens] MQLSWFMVFKCLATKAPKEMRVTILPQFVTPSLKTWIITRDNDFTSSSFLMSSFRSMPEKEKHAGPHGLL CLNRRKIRLPTRGGKIATGIYRTQAVSWLRCTPSPTRRKVSSAFLLVAPDLPIGFVSLAPVACSRRYREI ISSDCRYNSSSSSSAGGGVEGSDPSVLVLLPDSPTNSSSGRTAGRFPRGSGSYPSSSSSGDMMASAAGAS ISSSGAARRGGRGACAGASANHRAGLPSPYLSPGRGLLLRSQKSPSRAGGAAAAGPQAGPPTVEVESDYR VSFEAKHCQSPPPPG