A Web based Database for Hypothetical Genes in the Human Genome

Sivashankari Selvarajan[1] and Piramanayagam Shanmughavel[2]

[1] Assistant Professor and Head, Department of Bioinformatics, Kongunadu Arts & Science College, Coimbatore, Tamilnadu, India

[2] Assistant Professor, Computational Biology and Bioinformatics Lab, Department of Bioinformatics, Bharathiar University, Coimbatore, Tamilnadu, India

ABSTRACT
Due to accumulation of genomic data, the function of a vast amount of genes and the proteins encoded by them are unknown. Unless, the function of proteome encoded by the entire genome is not known, the biochemical processes and their importance cannot be understood. Also, the computational annotation returns a gene without any homolog in the protein database it encodes it as 'hypothetical'. Due to advancements in annotation projects, many genes whose evidence for expression invivo is not known and due to lack of similar protein could not be assigned function. This pose a challenge to functional genomics and automatic annotation of hypothetical genes are done at a faster rate using developed annotation tools to know the function of the hypothetical genes. Moreover, when the hypothetical genes are present in human, it is really a lacuna and hence functional annotation of the hypothetical genes in the human genome is the need of the hour. Hence, this work attempts to annotate the hypothetical genes in Human and makes the results publicly accessible using a web based database using PHP and MySQL.

Keywords: Hypothetical database, human hypothetical genes.

1. INTRODUCTION
The genome sequencing was started in 1995 with the sequencing of first cellular life form *haemophilus influenzae* and human genome sequencing started in 1990 with the probable time of completion of 15 years. Due to advancements in sequencing efforts, the human genome project was completed in 2003 well ahead of the estimated year 2005. The human genome project revealed the three billion base pairs encrypted within the twenty three pairs of chromosomes in the human genome. Also, the Human Genome contains 30,000 genes, constituting just 1% of the ~3 billion base pairs of the total human DNA. Among these, there are genes (called Hypothetical ORFs) which code for the so-called "hypothetical proteins" whose existence is either validated experimentally or predicted computationally but its function is not yet reported. Hence, after the completion of the genome sequences, the challenge ahead for all biologists is to use the data to interpret the function of the protein, the cell, and the organism. This can be achieved by a process called annotation which involves identification of genes within the chromosome, its fine structure, determination of protein products encodes by the gene and understanding the function *(Venter et al., 2001).* A group of these genes may be involved in many pathological disorders and hence are of pharmaceutical significance. Thus, annotation is an essential process of understanding the entire mechanism behind the cellular processes and molecular functions of a genome. However, there were inconsistencies in the accuracy of genome annotation in the initial stages which are now gone due to advancements in computational algorithms and potentiality of bioinformatics. After annotation of the Human Genome a number of genes (59%) reported by the project were hypothetical and annotated genes with unknown function *(Venter et al 2001).*

In biochemistry, a **hypothetical protein** encoded by a **hypothetical gene** is a protein whose existence has been predicted, for which there is no experimental evidence for expression *in vivo (*Zarembinski *et al* 1998). As a result, the function of such genes is not known. This is due to the fact that they are predicted using computational methods, which rely on signals in DNA sequences to predict it as a gene or based on similarity to genes in other organisms. In this case, the function of these homologous genes is also not known. Not only in Human Genome, in all genomes sequenced to date, a large portion of these organisms' protein coding regions encodes polypeptides of unknown biochemical, biophysical, and/or cellular functions. The usual scenario involving a hypothetical protein is in gene identification during genome analysis. When the bioinformatics tool used for the gene identification finds a large open reading frame without an analog in the protein database, it returns "hypothetical protein" as an annotation remark.

Despite several efforts, only 50-60 % of genes have been annotated in most completely sequenced genomes and their functions are known. The rest 40% of the genes in any genome is totally unknown in terms of its functions. The experimental characterization of such a huge number of hypothetical genes will take many decades before the biological function encoded by such hypothetical genes is known. Thus, automated genome sequence analysis and annotation methods may provide ways to understand genomes.

1.1. FUNCTIONAL ANNOTATION

The high-throughput genome projects have resulted in a rapid accumulation of genome sequences for a large number of organisms and large number of genes with unknown function (Hypothetical). To fully realize the value of the data, scientists need to identify proteins encoded by these genes and understand how these proteins function in making up a living cell. With experimentally verified information on protein function lagging far behind, computational methods are needed for reliable and large-scale functional annotation of proteins. **Functional annotation** is the process of identifying for a given gene its biological function, interaction with other elements, involvement in metabolic pathways, and any other piece of information that helps in understanding when and how a gene influences the overall system.

The first step in assessing a new protein sequence is always to see whether the string of amino acids is similar to a known sequence: that is, to scan sequence databases [GenBank, European Molecular Biology Laboratory (EMBL)] for homologous proteins. PSI-BLAST (Altschul *et al.,* 1997), which iteratively scans a sequence database to automatically build protein-specific profiles, is able to detect distant relations and reliably provide statistical significance. However, data derived from protein structures show that even PSI-BLAST can only identify about 2/3 of all evolutionary relationships (Salmov *et al.,* 1999).

An alternative approach to finding a sequence homolog is to scan the sequence against a library of protein domain families. As more sequence and structure data have been gathered, it seems increasingly likely that there are a limited number of ancestral protein domains (Chothia, 1993), which have duplicated and evolved into large families with great structural and functional diversity (Todd *et al.,* 2001). Further diversity is introduced by mixing and matching these domains during evolution. These protein families can be thought of as the "elements of the periodic table of biology," from which biological complexity is created. In these libraries of protein domains [such as Pfam (Bateman *et al.,* 2000), SMART (Schultz *et al.,* 1998), and COGS (Tatusov *et al.,* 2000)], a sequence alignment for each domain is constructed, which allows a novel sequence to be matched rapidly to domains already in the library. In the computer, a family is encoded as a profile or a Hidden Markov model (Krogh *et al.,* 1994) (HMM) that can sometimes detect more distant relatives than PSI-BLAST.

Many protein functional sites are well conserved and exhibit specific sequence motifs, which can provide surprisingly sensitive and specific search tools. Motif libraries, combined in the InterPro (Apweiler *et al.,* 2001) resource, are sometimes useful in recognizing distant relatives or annotating a sequence. A new sequence can be rapidly scanned against these libraries to help identify functional sites within it.

If no sequence homolog is found, or, as is more likely, no structural data are found for the family of interest, then a library of protein domain structures can be scanned to attempt fold recognition (i.e., compare sequence to structure rather than sequence to sequence) (Jones *et al.,* 1992). The sequence is optimally aligned with each fold in turn. The match is scored using information derived from sequence similarity measures, secondary structure prediction, and empirical energy functions (from observed residue separations in proteins of known structure). The prediction is the matched protein structure that gives the best score. In blind tests (Moult *et al.,* 1999), these methods are often able to identify more distant relatives than sequence comparisons alone. Theoretically, these methods aim to recognize all proteins with similar folds, but in practice only those with a common ancestor are found. The order of genes on a genome or pathway analysis can also be helpful for some proteins (Huynen *et al.,* 2000). These searching methods will generate a sequence alignment of the query and the matched sequence from which a structure may be built. 3D structure can aid the assignment of function, motivating the challenge of structural genomics projects to make structural information available for novel uncharacterized proteins. Structure-based identification of homologues often succeeds where sequence-alone-based methods fail, because in many cases evolution retains the folding pattern long after sequence similarity becomes undetectable. Finally, structural features can be used, after modeling the structure of a protein from its sequence or solving its structure. Protein fold class can be strongly indicative of function, while other structural features, such as secondary structure content, cleft size and 3D structural motifs are also useful. (Dobson et al., 2004)

Nevertheless, prediction of protein function from sequence and structure is a difficult problem, because homologous proteins often have different functions. Many methods of function prediction rely on identifying similarity in sequence and/or structure between a protein of unknown function and one or more well-understood proteins. Alternative methods include inferring conservation patterns in members of a functionally uncharacterized family for which many sequences and structures are known. However, these inferences are tenuous. Such methods provide reasonable guesses at function, but are far from foolproof. It is therefore fortunate that the development of whole-organism approaches and comparative genomics permits other approaches to function prediction when the data are available. These include the use of protein-protein interaction patterns, and correlations between occurrences of related proteins in different organisms, as indicators of functional properties. Even if it is possible to ascribe a particular function to a gene product, the protein may have multiple functions. A fundamental problem is that function is in many cases an ill-defined concept. (Whisstock, 2003).

2. METHODOLOGY

The initial step in function annotation is to classify the hypothetical genes into ‘characterised’ and ‘uncharacterised’ based on availability of functional information in NCBI (Pruitt et al, 2007). The characterised hypothetical genes were assigned functional categories from COG/SCOP(Tatusov et al., 2003; Andreeva et al., 2007). The uncharacterised hypothetical genes were

analysed both at protein and gene level. The protein level annotations included domain prediction using pfam and Superfamily prediction using Superfamily database (Gough et al., 2001). The gene level annotations included querying against BLAST2GO which is a data integration tool for functional annotation. Finally, using the protein and gene level annotations functions were predicted to the uncharacterised hypothetical genes and assigned functional category using COG/SCOP.

The results of the annotation were created as a web based database using MySQL as backend and PHP as front end.

2.1 CREATION OF DATABASE

MySQL is a relational database management system (RDBMS) that is developed by Michael Widenius runs as a server providing multi-user access to a number of databases. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL is owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Sun Microsystems, a subsidiary of Oracle Corporation. MySql has many advantages in comparison to Oracle since it's an open source and easy to learn package. PHP is a general-purpose scripting language that is especially suited to server-side web development where PHP generally runs on a web server. It is an Open Source, readily available and dual-licensed. Any PHP code in a requested file is executed by the PHP runtime, usually to create dynamic web page content. It can also be used for command-line scripting and client-side GUI applications. PHP can be deployed on most web servers, many operating systems and platforms, and can be used with many relational database management systems. It is available free of charge, and the PHP Group provides the complete source code for users to build, customize and extend for their own use. PHP is available free of cost and is flexible.

2.2 THE DATABASE WEBSITE

The Database is publicly accessible and is available at www.humanhypo.com. The site includes the following main features: complete list of all hypothetical genes in human genome hyperlinked to individual NCBI pages; Hypothetical genes organized by functional category; a database Help page. A sample webpage is available in fig.1.

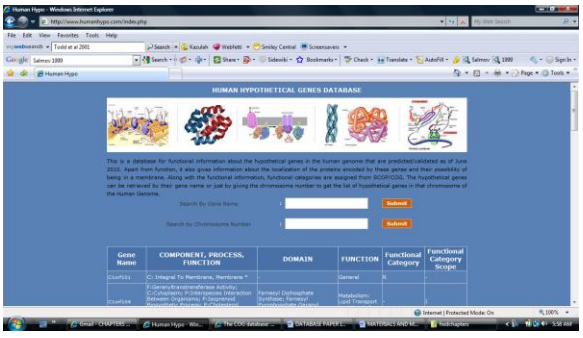

Fig.1 The Human Hypothetical Database –Home Page

2.3 FEATURES OF DATABASE

The Database for hypothetical genes in the Human Genome contains approximately 700 hypothetical genes. The database has information about gene name, Protein name, the component, process and function in which they might be involved, domains in it, Superfamily to which the protein belong to and finally the functional category of SCOP/COG. The Database can be queried using the hypothetical gene name directly or using the chromosome number of the human genome. The main feature of the database is its ability to connect to NCBI webpage for each gene.

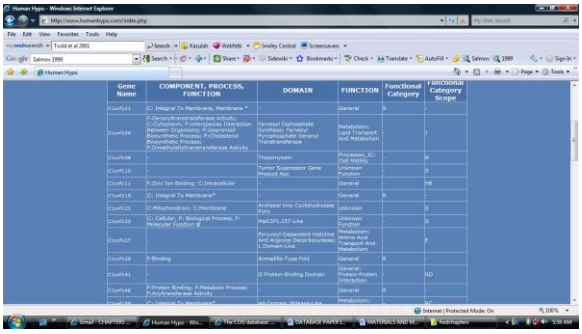

Fig.2 The Human Hypothetical Database –Result Page

3. CONCLUSION

Hypothetical genes pose a general threat to the research community. That too, when its in the Human Genome it does not complete information about the Human Behavior. Also, the proof for expression of hypothetical genes can be made only knowing its computational annotation. Hence, this database is a rich source of information for those who are in Hypothetical gene research of Human genome.

REFERENCES

[1] J. C. Venter *et al*., *Science* 291, 1304 (2001).

[2] Zarembinski, T. I., Hung, L.-W., Mueller-Dieckmann, H.-J., Kim, K.-K., Yokota, H ., Kim, R. & Kim, S.-H. (1998). Proc. Natl Acad. Sci. USA, 95, 15189-15193

[3] Todd AE, Orengo CA, Thornton JM, Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol. 2001 Apr 6;307(4):1113-43.

[4] Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001). "Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure." *J. Mol. Biol.*, 313(4), 903-919.

[5] Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, B Sridhar Rao, Sergei Smirnov, Alexander V Sverdlov, Sona Vasudevan, Yuri I Wolf, Jodie J Yin, and Darren A

Natale , The COG database: an updated version includes eukaryotes, BMC Bioinformatics. 2003;4: 41.

[6] Krogh A, Larsson B, von Heijne G, Sonnhammer EL, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J Mol Biol. 2001 Jan 19;305(3):567-80.

[7] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 2006;34:D187–D191.

[8] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997; 25:3389–3402.

[9] Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001). "Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure." *J. Mol. Biol.*, 313(4), 903-919.

[10] Antonina Andreeva1,Dave Howorth1, John-Marc Chandonia, Steven E. Brenner, Tim J. P. Hubbard, Cyrus Chothia5 and Alexey G. Murzin, Data growth and its impact on the SCOP database: new developments Nucleic Acids Research, 2008, Vol. 36, Database issue D419–D425.

[11] Pruitt KD, Tatusova, T, Maglott DR, NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res 2007 Jan 1;35(Database issue):D61-5

[12] Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Tim D. Williams, María José Nueda, Montserrat Robles, Manuel Talón, Joaquín Dopazo and Ana Conesa, High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008 June; 36(10): 3420–3435.