International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

In Silico Studies on Identification of Novel Therapeutic Targets for Treatment of Diabetes: A Review

Shivani Singh

Goel Institute of Technology and Management, Lucknow, U.P., India

Abstract: ***The expressed sequence tags (ESTs) are major entities for gene discovery, molecular transcripts, and single nucleotide polymorphism (SNPs) analysis as well as functional annotation of putative gene products. In our quest for identification of novel diabetic genes as virtual targets for type II diabetes, databases and Methods used in silico evaluation are illustrated. The functional and structural annotations of these proteins revealed some important features which may lead to the discovery of novel therapeutic targets for the treatment of diabetes.***

Keywords: Expressed sequence tags, Single nucleotide polymorphism, molecular transcripts

1. Introduction

It is essential to get a clear picture of the genes and its products involved in understanding the behaviour and functioning of different biochemical functions, which is evident by the functional associations of DNA, RNA, and proteins. In the light of genomics and proteomics, rapid development in technologies such as microarray, sequencing, and spectrometry has contributed vast data to analysis and prediction. Expressed sequence tags (ESTs) are small sequence of nucleotide extensions originating from cDNA repositories (200-800 bases). These are capable of recognizing the complementary full-length gene and are often utilized for identifying an activated gene.

The method of EST processing includes the sequencing of individual fragments of spontaneous clones from an organism's cDNA library, either 5' end or 3' end. Many ESTs at a time can be generated by a automation of DNA isolation and single sequencing reaction , sequencing, and analysis. ESTs are exponentially increasing in different public databases after their original identification and involvement as primary tools in human gene discovery [1], which will continue until sufficient funding for decoding activities is available.

A significant number of ESTs are also isolated from model organisms such as Caenorhabditiselegans, Drosophila, rice, and Arabidopsis, even if the initial ESTs were of human origin. For study and review, public databases such as dbEST[2], TIGR Gene Indices[3], and UniGene [4-6] now contain ESTs from a variety of species. Furthermore, several privately funded, in-house collections of ESTs accessible for study are held by many industrial establishments. ESTs are currently commonly utilized for genetic analysis, annotation, complementary genome, mapping, gene prediction, polymorphism analysis recognition of gene structure, and expression studies in the genomics and molecular biology communities to determine the viability of alternative transcripts and promote proteome analysis.Hyperglycemia, glucosuria, negative nitrogen balance, and occasionally ketonemia are a metabolic condition characterised by diabetes.

Retinopathy, neuropathy, and peripheral artery insufficiency are the clinical signs associated with it. Obese people are much more susceptible to diabetes and have a sedentary life. A new study shows that 150 million individuals are affected and almost 300 million more will be diabetic by 2025[7]. The non-insulin-dependent (type II diabetes or NIDDM) accounts for 95 percent of the reported cases of the disease out of the three main forms of diabetes.

There is no standard approach to the treatment of this disease and combined therapy from multiple approaches is generally implemented. The global Type II diabetes epidemic has contributed to the emergence of new methods for its treatment. The discovery of peroxisome proliferator activated receptors (PPARs) for nuclear receptors heralded a new age in understanding insulin receptor patho-physiology and its associated complications [8].

The receptor for the fibrate class of hypolipidemic agents is considered to be PPARs, whereas PPAR agonists minimise hyperglycemia without enhancing insulin secretion. There is no standard approach to the treatment of this disease and combined therapy from multiple approaches is generally implemented. The global Type II diabetes epidemic has contributed to the emergence of new methods for its treatment. The discovery of peroxisome proliferator activated receptors (PPARs) for nuclear receptors heralded a new age in understanding insulin receptor patho-physiology and its associated complications [8].

The receptor for the fibrate class of hypolipidemic agents is considered to be PPARs, whereas PPAR agonists minimise hyperglycemia without enhancing insulin secretion. Protein tyrosine phosphatase-1B (PTP1B) and glycogen synthase kinase-3 (GSK-3) are the only other validated targets. PTP-1B is a single catalytic-domain cytosolic phosphatise [9]. It is a nonspecific PTP in vitro and phosphorylates a large range of substrates. In vivo, insulin signalling by dephosphorylation of unique phosphotyrosine residues on the insulin receptor is involved in down-regulation. GSK-3

is a type of protein kinase that mediates the phosphorylation, in particular of cell substrates, of certain serine and threonine residues.

This phosphorylation mainly inhibits the target proteins, as glycogen synthase[10-12] is inhibited in the case of glycogenesis. Although a lot of research focuses on validated targets such as PTP1B, PPARs, and GSKs, the purpose of this paper is to recognise new diabetic genes as virtual targets. The method is solely in silico and accessible in public databases through review of ESTs.

Databases

For several species, databases such as dbEST, TIGR Gene Indices, and UniGene are the most valuable tools that contain raw and EST clusters. DbEST is the largest repository of NCBI-maintained EST data. DFCI's TIGR Gene Indices list the ESTs of several species alphabetically. NCBI's UniGene comprises transcript sequence gene-oriented clusters resulting from alignments between transcript sequences and genomic sequences originating from the same gene.

Table 1: The current information content of these three databases is represented

Date	Database name	Information content
01 / 04 / 2012	dbEST	ESTs 8315296
28 / 04 / 2006	TIGR Gene Indices	ESTs 7233257 HTs 234976
3 / 12 / 2011	UniGene	mRNAs 209412 Models 212 HTC 20115 3’ ESTs 1693253 5’ ESTs 4027153

The UniGene database was searched for human diabetes gene clusters to initiate an in silico study, which identified seven gene entries whose mRNA and ESTs data are listed by pratibha et.al.2013.[10].

Table 2: Details on human diabetes under UniGene (mRNA and ESTs)

mRNA	Source	ESTs	Gene Name
06	*Homo sapiens*	141	Ankyrin repeat domain 23 (ANKRD23)
12	*Homo sapiens*	46	Glucokinase (GCK)
14	*Homo sapiens*	10	Arginine vasopressin receptor 2 (AVPR2)
06	*Homo sapiens*	160	Ras-related associated with diabetes (RRAD)
07	*Homo sapiens*	61	Aquaporin 2 (AQP2)
10	*Homo sapiens*	217	Islet cell autoantigen 1 (ICA1)
09	*Homo sapiens*	181	SRY (Sex determining region Y) box 13 (SOX13)

2. Methods used Insilico evaluation

Pre-processing of EST

The EST sequences often are low-quality since they are produced without validation automatically and therefore require large failure rates. During the processing, the ESTs are often contaminated by vector sequences, since a portion of the vector is also sequenced along with the EST sequences. To minimize overall variability and increase effectiveness in further research, these sequences should be excluded from the EST. For example, a comparison of ESTs with different non-redundant vector databases identifies the contamination that is removed before analysing. The EMVEC [13, 14] database eliminates contamination of vector using NCBI BLAST2 [15, 16] from the EST sequences. In study, of pratibha et.al.2011 [9] the use of UniGene clusters is evident as each cluster generates combined data from dbEST, the mRNA database of GenBank, and electronically spliced genomic DNA. They are further clustered and washed (either by bacterial vector sequences or by linker sequences) from contamination.

Clustering & Assembly for EST:

In order to minimise duplication, the goal behind EST clustering is to accumulate overlapping ESTs from the very same transcript of a single gene into a specific cluster. This is essential because all the data expressed from a specific gene is clustered into an index class that reflects that particular gene's information. The clustering or aggregation is mainly performed by searching for similarities between sequences in pairwise sequences and consists of three main phases.In the first stage, weak regions with readings of both 5' and 3' are detected and removed.

The overlap areas between the sequences are then determined and after its detection, the incorrect overlaps are eliminated. In the second step, in descending order of overlap ratings, readings are joined to form contigs. Then all forward-reverse constraints are required to fix the resulting contiguities. A multiple sequence alignment of reads is built in the third step and a consensus sequence is determined for each contig along with a quality value for each basis. In the measurement of overlaps and the creation of several sequence alignments, base quality values are used. The CAP3 Server was exposed to cluster analysis of tissue-based ESTs from six identified genes [17].

Similarity of Database Searches:

Consensus sequences or continues (putative genes) collected from clustering are very valuable if its usability is known and repository similarity analysis utilizing specific freely accessible tools such as BLASTN and BLASTX is only possible. The ESTs are subsequently matched to the genome sequence of the organism for transcriptome examination using advanced programmed such as BLAT (BLAST as alignment tool)[18] to aid genome mapping and gene discovery.

ESTs Logical Translation:

Data or EST sequences is informative when its ontology functions and structure are apparent , for this the ESTs are associated by most detailed and reliable polypeptide translation to protein-centric annotations. The fact that the proteins serve as better models for the recognition of motifs and domains for the analysis of protein localization and gene ontology assignment is the fact that governs this process. EST translations are triggered by the concept of protein-coding regions or ORFs (open reading frames) from consensus or contiguous sequences.

The Functional Annotation

The functioning of a presumed polypeptide is anticipated by aligning sequences of protein, motifs, and family with non-redundant databases; this is because proteins serve as better models for operational annotation by imposing multiple alignment, HMM generatio, profile, analysis of phylogenetic, domains, and analysis of motive.

References

[1] Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., ...& Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, *252*(5013), 1651-1656.

[2] Boguski, M. S., Lowe, T. M., &Tolstoshev, C. M. (1993). dbEST—database for "expressed sequence tags". *Nature genetics*, *4*(4), 332-333.

[3] Lee, Y., Tsai, J., Sunkara, S., Karamycheva, S., Pertea, G., Sultana, R., ...&Quackenbush, J. (2005). The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic acids research*, *33*(suppl_1), D71-D74.

[4] Pontius, J. U., Wagner, L., & Schuler, G. D. (2003). 21. UniGene: A unified view of the transcriptome. *The NCBI Handbook. Bethesda, MD: National Library of Medicine (US), NCBI*.

[5] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., ...& Geer, L. Y. (2006). Database resources of the national center for biotechnology information. *Nucleic acids research*, *34*(suppl_1), D173-D180.

[6] Schuler, G. D. (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *Journal of Molecular Medicine*, *75*(10), 694-698.

[7] Dunstan, D. W., Zimmet, P. Z., Welborn, T. A., De Courten, M. P., Cameron, A. J., Sicree, R. A., ... & Atkins, R. (2002). The rising prevalence of diabetes and impaired glucose tolerance: the Australian Diabetes, Obesity and Lifestyle Study. *Diabetes care*, *25*(5), 829-834.

[8] Behera, P. M., Behera, D. K., Panda, A., Dixit, A., &Padhi, P. (2013). In Silico Expressed Sequence Tag Analysis in Identification of Probable Diabetic Genes as Virtual Therapeutic Targets. *BioMed Research International*, *2013*.

[9] Zhang, Z. Y. (2001). Protein tyrosine phosphatases: prospects for therapeutics. *Current opinion in chemical biology*, *5*(4), 416-423.

[10] Woodgett, J. R. (1994, August). Regulation and functions of the glycogen synthase kinase-3 subfamily.In *Seminars in cancer biology* (Vol. 5, No. 4, pp. 269-275).

[11] Woodgett, J. R. (2001). Judging a protein by more than its name: GSK-3. *Science's STKE*, *2001*(100), re12-re12.

[12] Ali, A., Hoeflich, K. P., &Woodgett, J. R. (2001). Glycogen synthase kinase-3: properties, functions, and regulation. *Chemical reviews*, *101*(8), 2527-2540.

[13] Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., ... & Lopez, R. (2002). The EMBL nucleotide sequence database. *Nucleic acids research*, *30*(1), 21-26.

[14] Etzold, T., Ulyanov, A., & Argos, P. (1996). [8] SRS: information retrieval system for molecular biology data banks. *Methods in enzymology*, *266*, 114-128.

[15] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., &Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.

[16] Shpaer, E. G., Robinson, M., Yee, D., Candlin, J. D., Mines, R., &Hunkapiller, T. (1996). Sensitivity and selectivity in protein similarity searches: a comparison of Smith–Waterman in hardware to BLAST and FASTA. *Genomics*, *38*(2), 179-191.

[17] Huang, X., &Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome research*, *9*(9), 868-877.

[18] Lottaz, C., Iseli, C., Jongeneel, C. V., & Bucher, P. (2003). Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics*, *19*(suppl_2), ii103-ii112.

[19] Iseli, C., Jongeneel, C. V., & Bucher, P. (1999, August). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. In *ISMB* (Vol. 99, pp. 138-148).

[20] Zdobnov, E. M., &Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, *17*(9), 847-848.