STM JOURNALS

International Journal of
Molecular Biotechnological Research

Volume 2, Issue 1, 2024
January–June
DOI (Journal): 10.37591/IJMBR

https://journals.stmjournals.com/ijmbr

Research IJMBR

Gene Annotation of Cancer Vaccine for *Homo sapiens*

Nimilita Chakraborty*

Abstract

Objectives*: Gene annotation helps us to deduce the structural and functional aspects of a gene that encodes for a functional protein in our body. Thus, by determining the coding sequence and gene location we can derive meaningful insights as to what these genes do in our body. In this study, an unknown gene, cancer vaccine for Homo sapiens has been studied and annotated.* ***Methods****: This study was based on a computational approach using various web interface tools to annotate an unknown gene taken from the NCBI Database. The chosen gene was structurally annotated using a GC% content calculator, visually represented using Microsoft Excel, Augustus for gene prediction, and RNAfold to determine the mRNA structure of the same gene. Functional annotation was done using BlastP, gene ontology was confirmed using the UniProt database, a phylogenetic tree was analyzed using HOGENOM database and TMHMM to visualize the transmembrane domain of the protein encoded by the gene, expression of the gene by Bgee, antibody analysis, subcellular localization, and functional analysis was accomplished using Human Protein Atlas, wolF PSORT and InterProScan respectively.* ***Results****: After completing the gene annotation, the cancer vaccine for Homo sapiens query was found to be 99.9% similar to four-jointed box protein 1 precursor [Homo sapiens] which exhibits low cancer tissue specificity and is mostly related to renal and urothelial cancer.* ***Conclusion****: The Cancer vaccine for Homo sapiens entry present in the NCBI Database, which had no annotation previously, was annotated structurally and functionally in this study. Now we can say this entry belongs to the gene coding for a four-box jointed protein-1 precursor protein which is useful for cancer diagnosis in the early stages and is related to poor prognosis of the disease. Often specific peptides are designed for FJX-1 protein which are beneficial in the treatment of cancers showing elevated expression of FJX1 proteins and are often used in the form of vaccines.*

Keywords: Gene prediction, mRNA structure, cancer, vaccination, local sequence alignment, protein, transmembrane domain

INTRODUCTION

Genes are the functional unit of DNA which get transcribed and hence translated to a functional protein. Genes can also be defined as functional units of inheritance that are passed on from parents to offspring generation after generation and contain the information that is required to specify various physical and biological traits. Our entire genome is divided into exonic and intronic regions. Only 1% of the DNA codes for functional proteins contributing to the phenotype. Humans are known to possess approximately 20,000 genes that code for proteins [1]. Errors during DNA replication and cell division lead to changes in the DNA sequence inducing a mutation that makes the gene to be either non-functional, translate truncated protein, or code for a different protein altogether. These mutations which change the sequence of nucleotides and thus open reading frame can disrupt the structure or function of the protein it encodes for, causing a state of disease in the organism. This leads to a genetic disorder. All the relevant information regarding genes and genomes can be mined from the NCBI Database. This database, which is within the National Library of

***Author for Correspondence**
Nimilita Chakraborty
E-mail: nimilita2000@gmail.com

Student Department of Biotechnology, KIIT University, Bhubaneswar, Odisha, India

Received Date: April 19, 2024
Accepted Date: May 01, 2024
Published Date: August 20, 2024

Citation: Nimilita Chakraborty. Gene Annotation of Cancer Vaccine for *Homo sapiens*. International Journal of Molecular Biotechnological and Research. 2024; 2(1): 1–14p.

Medicine at the National Institute of Health, maintained by the government of the United States, is a goldmine of information pertaining to the field of biotechnology and biomedical studies. Also, various bioinformatics tools and services can be availed from this database. The eukaryotic genome is very difficult to identify as it contains varied structures and possesses exons interspersed by the introns. Exons or the expressed sequence contain the protein coding regions, whereas the introns or interspersed sequences are removed after the process of transcription. Vertebrates mostly have 5% of exonic sequences, with most genes having 7 to 8 exons whose length ranges between 145 and 146 base pairs. Whereas introns have an average length of approximately 3365 base pairs. The average length of a coding sequence with introns removed is 1340 base pairs. The longest known coding sequence encodes titin protein and contains 80,780 base pairs exons only. The longest known gene is that which codes for the protein dystrophin which is about 2.4 mb and includes both introns and 97 exons. The complexity of eukaryotic gene identification is increased due to the presence of alternate splicing which translates mRNA from the same gene into different proteins. Thus, to identify genes and annotate genes high-throughput next-generation sequencing and bioinformatics play a crucial role [2, 3]. Thus, all the sequencing data obtained from high-throughput sequencing technologies like NGS need to be annotated and identified for future reference. New sequencing technologies generate 100 times more data as compared to traditional Sanger sequencing techniques which remain unused due to complicated annotating protocols [4]. In this project, various free web interface bioinformatics tools have been used to annotate an unannotated gene, both structurally and functionally, taken from the NCBI Database and using various free bioinformatics analysis tools. For this study, the gene annotation protocol given by Galaxy is an open-source and web-based program that allows scientists and researchers who are not adept in programming languages to use various bioinformatics tools. Galaxy is an online analysis and workflow system tailored for biologists to examine and process their data. Most of the well-known bioinformatics tools are already installed and prepared for use in the provided package. Across the globe, numerous Galaxy servers exist, each uniquely equipped with specialized toolsets and reference data to facilitate the examination of various fields like human genomics, microbial genomics, proteomics, and more.

The Galaxy interface is divided into three sections. The left side is a list of tools, whereas the middle section consists of a viewing pane. By contrast, one can access the analysis performed by the tool and data history [5].

To analyze large datasets, a galaxy provides three advantages to the users: (a) easy and accessible data analysis by scientists without the requirement of any prior knowledge for informatics, (b) reproducible and error-free data analysis, and (c) transparent communication of analysis. With improved features, such as scalability, advancement of tools, easy interactive analysis and visualization, easily comprehensible user interface, improved workflows, and infrastructure enhancement, the Galaxy Project has four complementary components: (a) the main Galaxy public server, which has been operational since 200, consists of a wide range of toolsets for large-scale genomics analysis, terabytes of public data for usage, many shared analysis histories, workflows, and many publication supplements. This server is said to have more than 124,000 registered users, who run approximately 245,000 jobs per month. (b) Open-source software that can be operated by any user on any operating system based on the UNIX operating system. The Galaxy ecosystem provides kits used for the development of Galaxy tool software, API language bindings for numerous programming languages, various software for designing Galaxy interactions, and tools for automatizing the setup and stationing of Galaxy and its plugins, such as tools and visualizations. (c) Galaxy ToolShed is a community-driven tool for sharing Galaxy tools, workflows, and visualizations. It can be found at. Developers and Galaxy administrators can host, distribute, and install Galaxy tools, workflows, and visualizations on this server, which serves as the 'AppStore' for Galaxy servers (d) All facets of the Project benefit greatly from the distinct and complementary subcommunities that make up the Galaxy Community. Users, administrators, developers, resource providers, and educators are just a few stakeholder groups that these subcommunities cater to [6].

METHODOLOGY

First, we needed to identify the portions of the genome that code for functional proteins. Thus, this part of the annotation is defined as structural annotation, where our main motive is to identify and locate open reading frames, gene structures, coding regions, and various regulatory motifs. Galaxy contains protocols and various tools for the structural annotation of a gene, some of which have been used.

Structural Annotation

For preliminary identification of an unannotated gene from the NCBI Database, we mined information from the GenBank file format, which is easily accessible from the NCBI Database. This file format provides information about the gene, such as gene name, function, locus number, and the exact position and sequence from the start and end of the gene, as well as sequence length. Information regarding organisms, strains, data collection, country, and sequencing methodology can be obtained from the GenBank and GFF file formats available in the NCBI Database. This information forms the basis of structural gene annotation.

Sequence Length Count

The total number of base pairs can be counted, and sequence extraction was performed using the FASTA file format available in the NCBI Database.

GC % Count

The greater the percentage of GC, the greater the stability of the gene we counted GC using the GC content calculator tool developed by the Biologics Corp Organization. When feeding the input of our FASTA sequence file, it determines the percentage content of all the nucleotides that make up the gene as well as the GC content.

This tool can be freely accessed at *https://www.biologicscorp.com/tools/gccontent/*.

The nucleotide content of the chosen gene was visually represented using Excel, and the results were as follows:

Thus, a bar chart (Figure 1) and pie chart (Figure 2) were prepared to visualize the nucleotide content of our gene of interest.

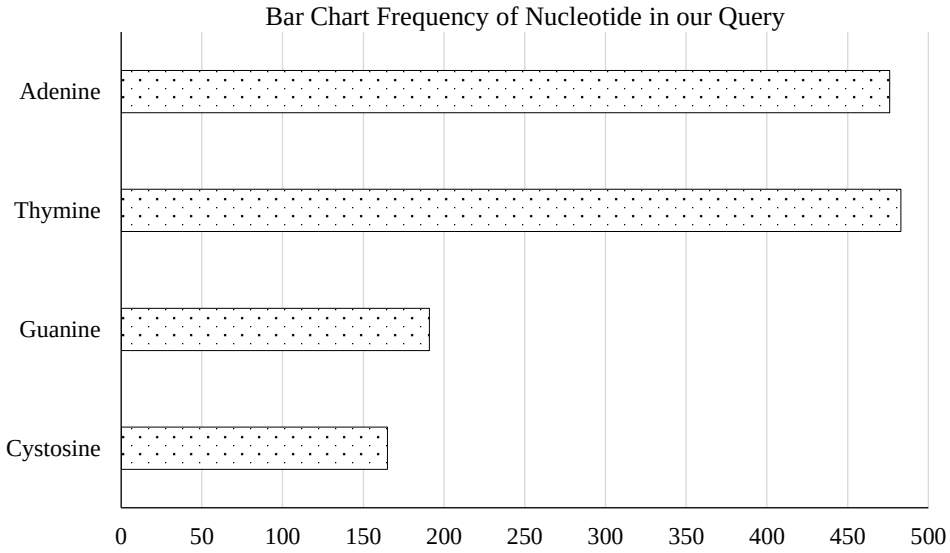

Figure 1. Bar chart of nucleotide frequency in our query using Excel (left).

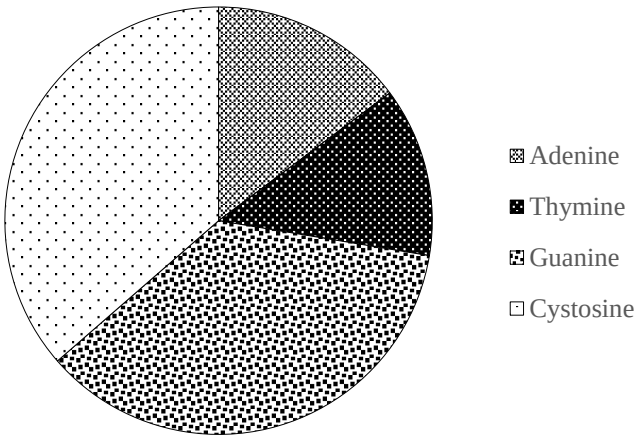

Figure 2. Pie chart representation of nucleotide frequency using Excel (Right).

Gene Prediction

The first step in gene annotation is to identify genes in each genomic sequence. Experimental validation of gene structure is costly and various bioinformatics tools play crucial roles in biological sequence analysis [7]. Based on a generalized hidden Markov model, which provides a probabilistic prediction of a gene and its structure, Augustus provides a prediction of the query genome sequence belonging to a eukaryote and provides an automatic genome annotation for users. The input of the query permits users to specify the position of the splice sites and mentions the start and stop codons. Using this tool, the user can mention the positions of exons and introns if required. This tool is available and can be downloaded at no cost from *https://bioinf.uni-greifswald.de/augustus/submission.php*.

This tool provides outputs in three different files, one of which specifies coding regions. Another method provides information regarding the translated product or protein sequence. The third output specifies all information regarding the start and stop codons of the genes (ORF), their transcript ID, UTR in a detailed manner, and the first and last genes of the given genomic sequence, which can also be marked in a GFF file format [8].

tmRNA tRNA Structure Prediction

Using the tRNA-SE scan tool, users can determine whether a query sequence codes for any tRNA. This free web interface tool software maintained by the University of California, Santa Cruz follows the concept of covariance models that can retrieve the primary and secondary tRNA structure data from the query sequence [9]. This tool can be accessed at *https://trna.ucsc.edu/tRNAscan-SE/*.

Tool RNAfold is a free tool service provided by the Vienna RNA Web Services for mRNA secondary structure prediction. These tools predict the secondary structures of single-stranded DNA and RNA sequences. Currently, the input is limited to 7500 nucleotides for partition function calculations and 10,000 nucleotides for minimum free energy calculations [10]. This tool is available at *http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi*.

Functional Annotation

Blast

The BLAST tool, or basic local alignment search tool, is a program provided by the NCBI Database that compares nucleotide or protein sequences and provides the statistical significance of matches. BLAST also helps the user infer the evolutionary and functional relationships between sequences and helps identify the members of gene families. BLAST is used for various purposes, such as identification of species, locating domains, searching for phylogeny, mapping DNA to a known chromosome, and annotations. The working algorithm of BLAST is based on the concept of a

heuristic model that locates short regions of similarity between the input sequences or between the input sequence and standard sequences in the NCBI Database. This method does not consider the entire sequence space; after the initial match, the tool starts the local alignment from the initial matches. However, BLAST is a local alignment tool that does not ensure optimal alignment; therefore, some sequences may be missed. For a more specific alignment, a Smith-Waterman or global sequence alignment algorithm should be used [11].

The BLAST tool can easily be accessed in the NCBI Database or at any other site.

(*https://blast.ncbi.nlm.nih.gov/Blast.cgi)*

This tool is useful in gene annotation as it helps to determine the sequence or protein that has the highest matching score with our query gene of interest. Greater similarity scores indicate that the gene of interest has a high degree of similarity and its characteristics can be compared with the output protein or gene [12].

PDB Database

The Protein Data Bank is a structural database of proteins, which is the storehouse of the three-dimensional structure of proteins, primarily nucleic acids, and other biological macromolecules. This database, developed in 1971 by Water Hamilton at the Brookhaven National Laboratory, is now an essential database to cater to the research demands of scientists and is a comprehensive resource for accessing and analyzing protein-related data. Their structures, functions, and interactions play important roles in various biological processes. The database is available at *https://www.rcsb.org/search*.

This database was used to mine the structure of the protein that showed the highest degree of similarity to our gene of interest [13].

UniProtKB Database

The UniProt Knowledge Database is an extremely useful database that contains integrated protein information with cross-references to many sources. It is also a primary database for protein sequences and functional annotation of proteins, based on experimental evidence. It combines a network of databases, centralizing all levels of protein sequence annotation [14]. This can be accessed from https://www.uniprot.org/.

In our experiment, we retrieved the following information from the UniProt database: (1) interaction of our query gene with other genes; (2) compartmental localization of the protein; (3) subcellular localization of the protein according to UniProt annotation; (4) Gene ontology annotation with ancestral charts; (5) Taxonomy and organism lineage; (7) PTM of the protein; (8) protein expression; and (9) family and domain analysis.

InterProScan

InterProScan, which is maintained by the European Bioinformatics Institute, is a functional annotation tool. It is a software package that allows the user to scan a novel protein or nucleotide against InterPro membrane database signatures. Thus, users can functionally characterize their novel proteins or nucleotides using InterProScan, which runs on a scanning algorithm against the InterPro database in an integrated manner. The InterPro member databases currently include Prosite, Prints, Pfam, and ProDom.

InterProScan functionally characterizes proteins by classifying them into families and predicting their domains and important sites [15]. InterProScan can be freely accessed at *https://www.ebi.ac.uk/interpro/search/sequence/*.

In our gene annotation, InterProScan was run with the FASTA file of the query protein obtained from the NCBI Database, which provided output in the tsv file format of our functionally annotated gene.

WoLF PSORT

This tool is a protein subcellular localization prediction tool. Based on the PSORT principle, WoLF PSORT is an extended version of the PSORT ii program, which retrieves information on the subcellular localization of the query protein. This tool uses the sequence of sorting signals, amino acid makeup, and functional motifs present, and modifies the amino acid sequence of the protein into numerical localization features. Following conversion, prediction is performed using a straightforward k-nearest neighbor classifier [16].

It can be accessed at *https://wolfpsort.hgc.jp/*.

In our experiment, we inputted the FASTA file of our protein sequence to reconfirm and compare the results with the subcellular localization results obtained from the UniProt database.

TMHMM

TMHMM is a prominent bioinformatics tool maintained by the Department of Health Technology (DTU) and is useful for projecting and visualizing the transmembrane helices of integral membrane proteins**.** This tool, in conformity with the hidden Markov model, helps users visualize portions of the query protein on the outside, inside, or within the folds of the membrane [17]. TMHMM2. Zero was used to analyze the protein of interest and can be used and freely accessed at *https://services.healthtech.dtu.dk/services/TMHMM-2.0/*.

Bgee

Bgee is a database that searches and compares the patterns of gene expression in various animal species.

It offers a logical response to the query, "Where is a gene expressed?" It has advanced evolutionary biology, cancer research, and agricultural research.

Upon entering the accession ID of the protein, its expression pattern can be easily analyzed [18]. The tool is available at https://www.bgee.org/.

HOGENOM Database

A large collection of homologous gene families, corresponding phylogenetic trees, and sequence alignments are available for a variety of organisms through the phylogenomic database HOGENOM [19]. This information is available at http://hogenom.univ-lyon1.fr/.

This was used in our gene annotation to crosscheck the phylogenetic results obtained from the UniProt database.

The Human Protein Atlas

Initiated in 2003, the Human Protein Atlas is a Swedish program whose goal is to map every human protein in cells, tissues, and organs by combining multiple omics technologies, such as transcriptomics, mass spectrometry-based proteomics, antibody-based imaging, and systems biology. The Human Protein Atlas tissue section is dedicated to mRNA and protein expression profiles of genes in human tissues. Protein expression data from 44 normal human tissue types were obtained via immunohistochemistry-based protein profiling using antibodies [20]. This information is available at https://www.proteinatlas.org/.

In our gene annotation, this database was used to derive the following information (1) General protein information, (2) Protein expression and localization, (3) Immune cell-type expression, (4) Immunohistochemistry data reliability, (5) Phylogenetic summary, and (6) disease relationships.

RESULTS

The sequence, a cancer vaccine for *Homo sapiens* obtained from the NCBI Database, showed 100 percent sequence similarity with the four-box jointed protein 1 precursor *Homo sapiens* (FJX 1). After running the BLAST, a 100% percent identity was shown with fjx1 protein. This query sequence was structurally noted to have a length of 1314 nucleotides, that is, 657 base pairs). Also, the GC content was found to be 73%. From tool AUGUSTUS (a web server for gene finding in eukaryotes), deduce the start and stop codon of the gene which are 'ATG' and 'TAA' respectively. Also, the translated amino acid sequence was determined which is 'MRGAAATAGLWLLALGSLLALWGGLLPPRTELPASRPPEDRLPRRPARSGGPAPAPRFPLPP PLAWDARGGSLKTFRALLTLAAGADGPPRQSRSEPRWHVSARQPRPEESAAVHGGVFWSRG LEEQVPPGFSEAQAAAWLEAARGARMVALERGGCGRSSNRLARFADGTRACVRYGINPEQI QGEALSYYLARLLGLQRHVPPLALARVEARGAQWAQVQEELRAAHWTEGSVVSLTRWLPN LTDVVVPAPWRSEDGRLRPLRDAGGELANLSQAELVDLVQWTDLILFDYLTANFDRLVSNL FSLQWDPRVMQRATSNLHRGPGGALVFLDNEAGLVHGYRVAGMWDKYNEPLLQPPRSAAT APPRLPRQAHFAL.'

From the tRNA-SE scan, we found that our query sequence contained no tRNA genes.

Using the RNAfold Tool, the minimum free energy of the primary structure of the mRNA was found to be -715.10 kcal/mol.

The centroid secondary structure of the same tool was determined to be -575.60 kcal/mol.

In addition, the primary and secondary RNA structures are represented visually, as shown in Figures 3 and 4.

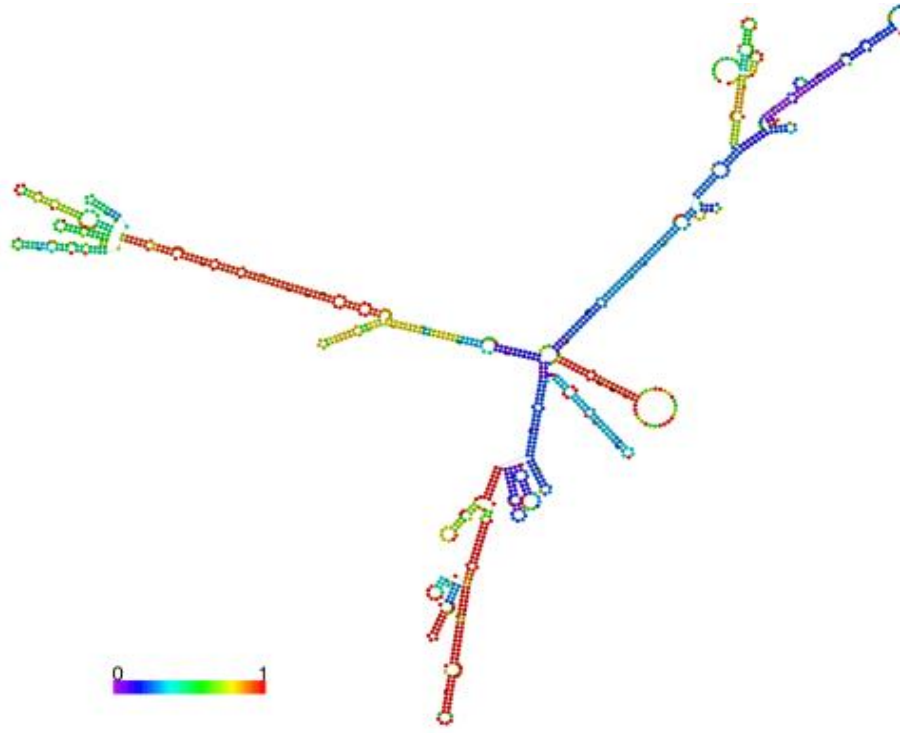

Figure 3. MFE secondary structure represented encoding base pair probabilities of the input nucleotide sequence as obtained from RNAfold Tool (top).

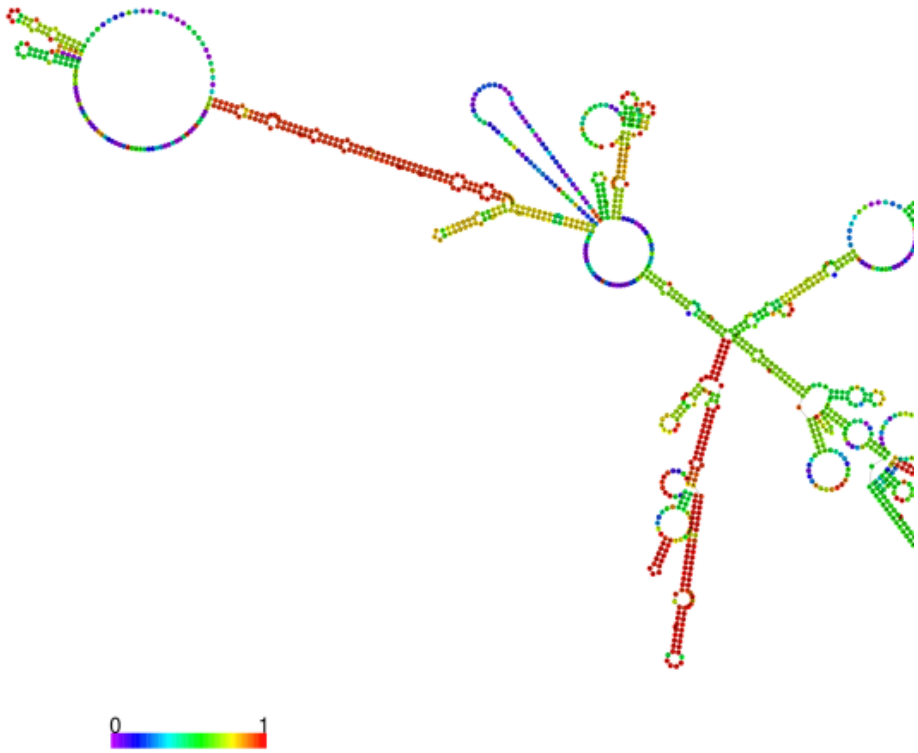

Figure 4. Centroid structure represented encoding base pair probabilities (down).

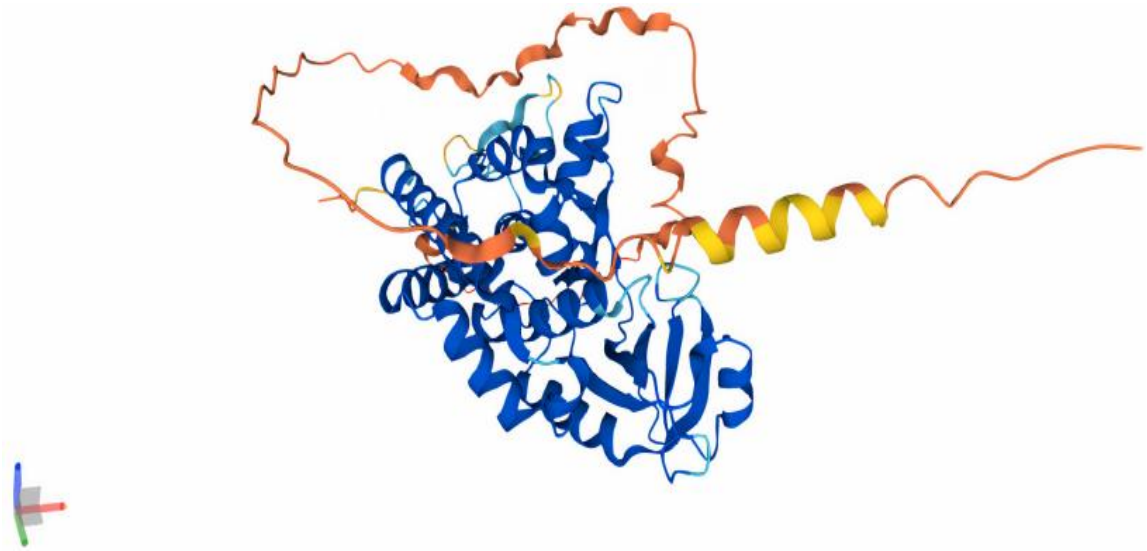

Figure 5. Alpha-fold structure of FJX1 protein(left) obtained from NCBI Database.

The gene was functionally annotated using BLAST, as previously described, and showed 100 percent identity with the fjx1 protein. An alpha-fold structure was obtained from the blast itself as well as its graphical representation. The FJX 1 protein sequence was obtained from the NCBI Database. The alpha-fold structure of the FJX1 protein(left) obtained from the NCBI Database is shown in Figure 5.

The query sequence also showed a 99.72% percent identity with the protein name a putative secreted ligand (*Homo sapiens*). However, no information is available regarding the alpha secondary structure of this protein in the blasts. Hence, the protein was searched in the NCBI Database to obtain GenBank ID Cab53246.1. This ID was used to search for structures in the PDB database, but no

information regarding the structure of this protein could be retrieved. When searched in the UniProt database, this same protein was found to be an isoform of the four-box joined protein 1 *Homo sapiens*.

Using the InterProScan functional prediction tool, we derived a graphical representation of the functional domains of the FJX 1 protein. Other information like FJX 1 is a 437 amino acid protein belonging to the species *Homo sapiens*, proteome ID: Up500005640, and acts as an inhibitor of dendrite extension and branching.

The Woftsport tool was used to obtain the subcellular localization of the fjx1 protein, which gave the results for 32 nearest neighbors, given in tabular format.

When the FJX 1 protein *Homo sapiens* was searched in the UniProt database, the following information was obtained:

The function of the protein is found to inhibit dendritic extension and branching.

As depicted in Figure 6, fjx1 interacts with several other genes, such as DKKL1 and PIGN, where it interferes with the expression of these two genes, whereas interactions with other genes, such as NOG, LEF1, and INS, are found to interact with each other.

The compartmental localization of the protein was found to be in the extracellular space and it was secreted.

The post-translational modification was observed to have a signal sequence from 1–24th nucleotide position with the sequence of the signal peptide being MGRRMRGAAATAGLWLLALGSLLA.

Glycosylation was observed at the 248th n position and was reported to undergo proteolytic cleavage according to the UniProtKB database.

As reported by the UniProtKB database in reference to the Bgee or expression analysis database, this protein is expressed in the ventricular region and 141 other tissues and has low tissue specificity according to the organism-specific database.

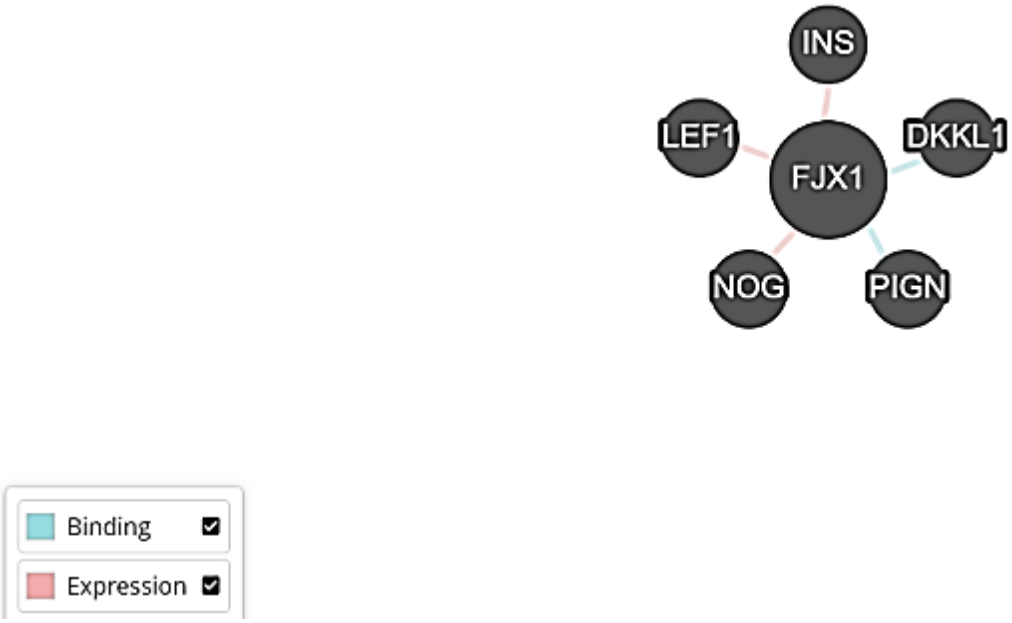

Figure 6. Representing the interaction of FJX1 gene interaction with other genes.

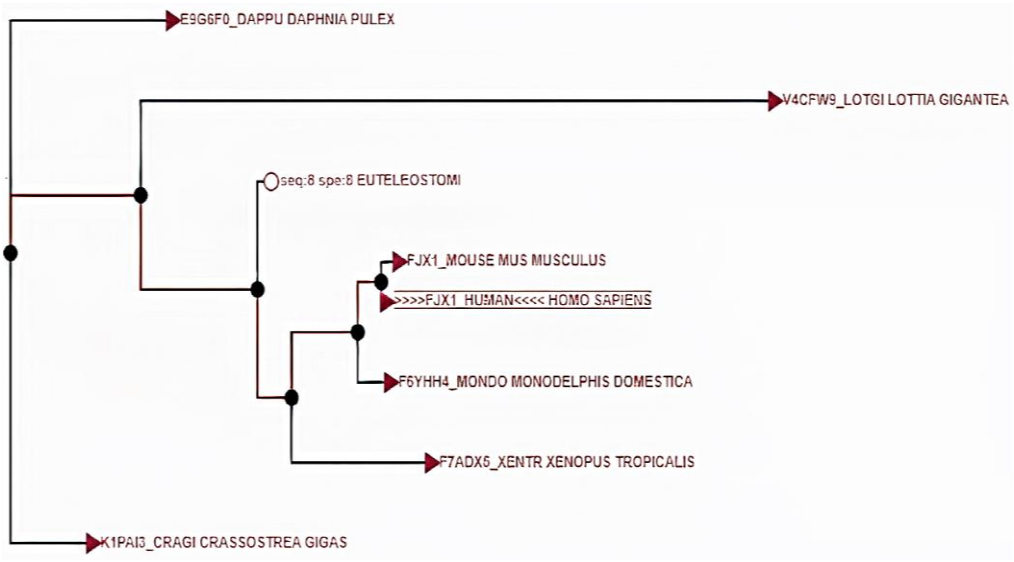

Figure 7. Phylogenetic tree of FJX1 *Homo sapiens*.

Table 1. Information for gene annotation as obtained from Ensembl Database.

Attributes	**Information obtained**
Gene synonyms	FLJ22416, FLJ25593
Location	Chromosome 11: 35,618,460-35,620,865 forward strands.
About this transcript	1 exon, annotated with 18 domains and features, associated 1109 variant alleles and maps to 290 oligo probes.
Gene	This transcript is a product of gene ENSG 00000179431.7
Statistics	Exons: 1, Coding exons: 1, Transcript length: 2,406 bps, Translation length: 437 residues
Type	Protein coding

The alpha-fold structure was confirmed to be the same as that depicted in the BLAST, as shown in the UniProt database.

Phylogenetic analysis and tree construction were performed as reported by UniProtKB in reference to the HOGENOM database, as reported by UniProtKB. from the tree in Figure 7, FJX1 *Homo sapiens* shares the closest homology to the FJX1 protein expressed in Mus musculus.

UniProtKB, in reference to the sequence Database Ensemble, provides the following information about the protein FJX1 *Homo sapiens*.

Gene ontology refers to the knowledge of the biological domain with respect to three aspects: molecular function, biological process, and cellular component. GO terms have been used by many groups, such as UniProtKB curators, to annotate gene products in a computationally tractable manner.

The GO annotations for FJX1 *Homo sapiens* according to the UniProtKB database are presented in Table 1.

Using the Human Protein Atlas Database, we deduced that this protein has low tissue specificity, has only a single transcript, has no protein interaction, subcellular localization is localized to the vesicles' only, predicted location is intracellular, mainly shows Neuronal Transcription, and has low Human Brain specificity. In the brain, this protein is primarily expressed in the astrocytes. This protein, as previously reported, was not detected in the immune cells. It is a prognostic marker for renal cancer (unfavorable) and urothelial cancer (unfavorable). Cancer specificity was enhanced in glioma, and cancer cell line specificity was also found to be low, as shown in Figure 8.

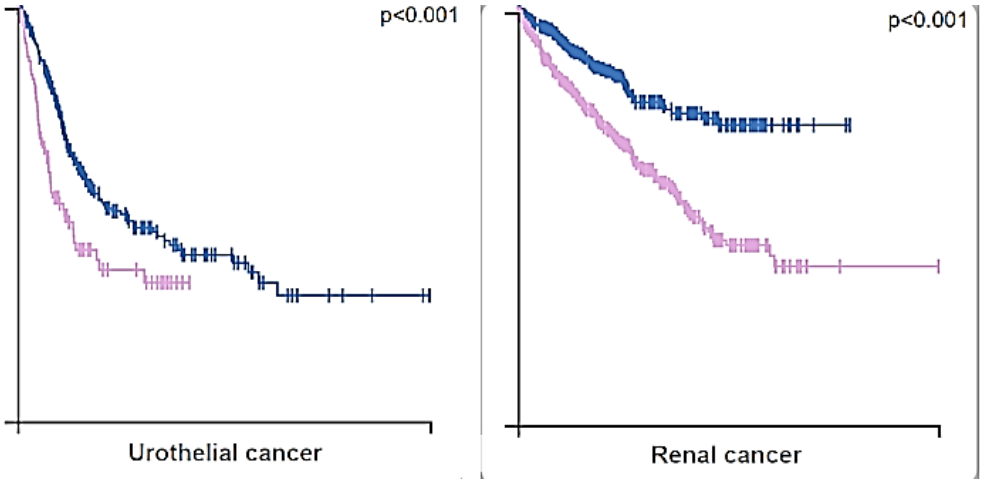

Figure 8. Representing expression of FJX1 marker in renal (left) and urothelial cancer (right).

Table 2. Representing GO annotation of FJX1 *Homo sapiens* is reported by the UniProtKB database.

Gene ontology	Characteristics
Cellular component	Extracellular space
Biological process	Cell–Cell signaling
Biological process	Retinal layer formation

Table 3. The information is derived from the TMHMM tool.

No. of predicted TMH	1
Expected no. of AA in TMH	20.99596
Total probability of N-in	0.99725

Using the TMHMM tool, we inferred that FJX1 *Homo sapiens* has a single transmembrane domain, and there are approximately 20 amino acids in the transmembrane domain, as shown in Table 2, Table 3.

DISCUSSION

From the above interpretation, we can conclude that the human vaccine for *Homo sapiens* entry into the NCBI Database is a four-box jointed protein belonging to the *Homo sapiens* species, the features of which were annotated using various bioinformatics tools. The results were cross-checked and compared among various databases that were found to match. Furthermore, as indicated in the structural annotation, this gene has high GC content, which makes it highly stable, and the marker expression in renal, urothelial, and gliomas makes it an ideal candidate for cancer vaccines, although this protein vaccine is reported to have low tissue specificity. This mRNA coding sequence, when translated into a four-box jointed protein, functions as an inhibitor of dendrite extension and branching through FJX1. *Homo sapiens* has not been characterized properly. Various studies have shown that nasopharyngeal cancer, a highly metastatic cancer, is widely prevalent in southeast China. Our current understanding of the molecular pathophysiology of nasopharyngeal carcinoma (NPC) remains insufficient to manage the disease better. The expression of the four-jointed box 1 (*fjx1*) gene was often found to be enhanced in malignant tissues of patients with nasopharyngeal carcinoma compared to healthy cells using experiments based on microarray analysis. The human gene FJX1, also known as the four-joint (*fj*) gene found in Drosophila and fjx1 in mice, has been linked to pathways involved in cancer progression. However, the precise role of *FJX1* in humans is not fully understood. Primary NPC tissue samples were used to confirm overexpression of FJX1 mRNA. Contrary to this, a subset of NPC tissues was experimentally determined to be around (42%) and exhibited significantly elevated levels of FJX1 protein than the normal epithelial cells and tissues, which showed no FJX1 expression. Moreover, fjx1 was discovered to be overexpressed in TCGA and

microarray datasets of various cancers, such as ovarian, colorectal, and head and neck cancers. siRNA knockdown and several studies on NPC cell lines of NPC revealed that FJX1 is related to increased cellular invasion, growth of cells requiring anchorage, and cellular proliferation. As the levels of FJX1 expression increased, cyclin d1 and e1 mRNA levels also increased in these cell lines, demonstrating that FJX1 is responsible for orchestrating proteins involved in the cell cycle. Since FJX1 overexpression contributes to the aggressive phenotype of NPC cells, further research is necessary to determine whether FJX1 can be used as a therapeutic target for NPC [21]. FJX1 is currently being investigated as a potential target for immunotherapy in NPC and other cancers.

This gene was structurally and functionally annotated according to the Galaxy protocol for genome annotation, using similar bioinformatics tools.

CONCLUSION

In conclusion, most tools available for genome or gene annotation are command-line tools that require a good grasp of the Linux operating system and commands to run them. Gene annotation plays an important role in various applications. First, when scientists discover a novel or putative genome or gene, they need to annotate it using various bioinformatics tools, including functional and gene finding annotation. Over the past three decades, computational tools for sequence annotation have been developed; however, for easy and quick annotation of entire genomes, better tools are needed. Annotation is key to improving comprehension and retention of information. Thus, each annotation detail of a gene has been added. Thousands of new genomes have been sequenced and assembled as a result of the genome sequencing revolution. However, genome annotation still makes use of technology that has essentially not changed in the last 20 years. The sheer volume of genomes makes fully automated annotation procedures necessary, but annotation errors are common, if not more so, than in the past. Thus, new technologies or tools are urgently needed, where scientists can easily annotate a gene or genome irrespective of its size in a hassle-free and integrative manner [22].

Acknowledgment

We acknowledge the Department of Bioinformatics, BioNome, Bengaluru, India for providing computational facilities and support for scientific research services. We thank Ms. Samiksha Bhor Ma'am for her assistance during the project.

Conflict of Interest

The authors declare no conflict of interest

Funding

Not applicable/ This research received no external funding.

Abbreviations

AA	Amino Acid
BLAST	Basic Local Alignment Search Tool
GC content	Guanine Cytosine Content
GFF	General Feature Format
HPA	Human Protein Atlas
NCBI	National Centre for Biotechnology Information
ORF	Open Reading Frame
PDB	Protein Data Bank
PTM	Post-Translational Modification
TSV	Tab Separated Value
UTR	Untranslated Region
NPC	Nasopharyngeal Carcinoma

REFERENCES

1. Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. Ann Med. 2002;34(2):88–95. doi: 10.1080/07853890252953473. PMID: 12108579.
2. Mercer TR, Mattick JS. Understanding the regulatory and transcriptional complexity of the genome through structure. Genome Res. 2013;23(7):1081–8. doi: 10.1101/gr.156612.113. PMID: 23817049.
3. Puente XS, Sánchez LM, Gutiérrez-Fernández A, Velasco G, López-Otín C. A genomic view of the complexity of mammalian proteolytic systems. Biochem Soc Trans. 2005;33(Pt 2):331–4. doi: 10.1042/BST0330331. PMID: 15787599.
4. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52(4):413–35. doi: 10.1007/s13353-011-0057-x. PMID: 21698376.
5. Ramsey J, Rasche H, Maughmer C, Criscione A, Mijalis E, Liu M, et al. Galaxy and Apollo as a biologist-friendly interface for high-quality cooperative phage genome annotation. PLoS Comput Biol. 2020;16(11):e1008214. doi: 10.1371/journal.pcbi.1008214. PMID: 33137082.
6. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46(W1):W537–44. doi: 10.1093/nar/gky379. PMID: 29790989.
7. Wang Z, Chen Y, Li Y. A brief review of computational gene prediction methods. Genomics Proteomics Bioinformatics. 2004;2(4):216–21. doi: 10.1016/s1672-0229(04)02028-5. PMID: 15901250.
8. Hoff KJ, Stanke M. Predicting genes in single genomes with AUGUSTUS. Curr Protoc Bioinformatics. 2019;65(1). doi: 10.1002/cpbi.57. PMID: 30466165.
9. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA genes in genomic sequences. Methods Mol Biol. 2019;1962:1–14. doi: 10.1007/978-1-4939-9173-0_1.
10. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA Websuite. Nucleic Acids Res. 2008;36:W70–4. doi: 10.1093/nar/gkn188. PMID: 18424795.
11. Oehmen C, Nieplocha J. ScalaBLAST: A scalable implementation of BLAST for high-performance data-intensive bioinformatics analysis. IEEE Trans Parallel Distrib Syst. 2006;17(8):740–9. doi: 10.1109/TPDS.2006.112.
12. Neumann RS, Kumar S, Shalchian-Tabrizi K. BLAST output visualization in the new sequencing era. Brief Bioinform. 2014;15(4):484–503. doi: 10.1093/bib/bbt009. PMID: 23603091.
13. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, et al. Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr. 1998;54(Pt 6 Pt 1):1078–84. doi: 10.1107/s0907444998009378. PMID: 10089483.
14. UniProt Consortium. UniProt: A hub for protein information. Nucleic Acids Res. 2015;43(D1):D204–12. doi: 10.1093/nar/gku989. PMID: 25348405.
15. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40. doi: 10.1093/bioinformatics/btu031. PMID: 24451626.
16. Magwanga RO, Lu P, Kirungu JN, Cai X, Zhou Z, Wang X, et al. Whole genome analysis of cyclin dependent kinase (CDK) gene family in cotton and functional evaluation of the role of CDKF4 gene in drought and salt stress tolerance in plants. Int J Mol Sci. 2018;19(9):2625. doi: 10.3390/ijms19092625. PMID: 30189594.
17. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol. 2001;305(3):567–80. doi: 10.1006/jmbi.2000.4315.
18. Bastian FB, Roux J, Niknejad A, Comte A, Fonseca Costa SS, Mendes de Farias T, et al. The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. Nucleic Acids Res. 2021 Jan 8;49(D1): D831–47. doi: 10.1093/nar/gkaa793.
19. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinform. 2019;20(4):1125–36. doi: 10.1093/bib/bbx120. PMID: 29028872.

20. Pontén F, Jirström K, Uhlen M. The Human Protein Atlas—A tool for pathology. J Pathol. 2008;216(4):387–93. doi: 10.1002/path.2440. PMID: 18853439.
21. Chai SJ, Ahmad Zabidi MM, Gan SP, Rajadurai P, Lim PVH, Ng CC, et al. An oncogenic role for four-jointed Box 1 (FJX1) in nasopharyngeal carcinoma. Dis Markers. 2019;2019:3857853. doi: 10.1155/2019/3857853. PMID: 31236144.
22. Salzberg SL. Next-generation genome annotation: We still struggle to get it right. Genome Biol. 2019;20(1):92. doi: 10.1186/s13059-019-1715-2. PMID: 31097009.

