Introduction to Bioinformatics – Project Report

Aims and Objectives

This paper aims to use standard bioinformatics tools and databases to analyse the promoter, genomic, mRNA, and protein sequences of an unknown gene obtained from human mammary tissue. It also seeks to either prove or disprove the hypothesis that this unknown gene has arisen from exon shuffling by analysing the homology of each exon in the given gene with other human genes.

Promoter Sequence Analysis

The given promoter gene sequence was subjected to a BLAST nucleotide search using the NCBI database and the results showed a 100% homology with the promoter sequence of the *Btn1a1* gene coding for butyrophilin protein in *Mus musculus* (house mouse). ClustalW was used to perform a multiple sequence alignment of the promoter regions of the butyrophilin gene in *Mus musculus* and humans and considerable homology between the two sequences was observed.

Genomic Sequence Analysis

The given unknown gene sequence was subjected to a BLAST nucleotide search using the NCBI database. The sequence was found to have a 99.66% similarity with the genomic sequence of the *Fcamr* gene coding for the Fca/m receptor in *Mus musculus* (house mouse). In order to further narrow down the exact regions of homology, the start and end positions of nucleotides in the homologous regions were identified, which were as follows: 1 – 424, 510 – 1399, 1641 – 2296, 2539 – 2938, 2866 – 2925, 3084 – 3669, and 4448 – 5107. When comparing these regions to the exon and intron regions specified for the unknown gene, it was found that exon 1, part of exon 5, and all the introns were homologous to the *Fcamr* gene sequence in *Mus musculus*.

In order to see if they were also homologous with the *Fcamr* gene sequence in humans, a multiple alignment was done using ClustalW for the unknown gene sequence and the *Fcamr* gene sequences in *Mus musculus* and *Homo sapiens*. Considerable similarity was found between the regions homologous between the unknown gene and *Fcamr* gene in house mouse with the *Fcamr* gene sequence in *Homo sapiens*. A multiple alignment using ClustalW was also performed between the unknown gene sequence and the butyrophilin gene sequence in humans. The results showed homology between a small part of the unknown gene sequence in the centre with the butyrophilin gene sequence.

mRNA Sequence Analysis

In order to identify homologous sequences for other exons of the given gene sequence, the mRNA sequence was used for a BLAST search in NCBI. As expected, nucleotides 1909 – 2566 were found homologous to the *Fcamr* mRNA sequence in *Mus musculus*. The next BLAST search was conducted by removing these nucleotides from the query in order to identify possible

homology with other genes. Interestingly, nucleotides 256 – 416 were found to be homologous to a non-coding RNA sequence found in *Mus musculus*. Two other homologous regions were found between nucleotides 1496 – 1529 to a catalytic subunit of cAMP-dependent protein kinase found in *Aspergillus candidus* and nucleotides 1499 -1542 to an alpha-like domain in ribosomal protein S6 kinase found in *Mesitornis unicolor*.

In order to find homologous regions in the human genome, the search was repeated by setting the filter for human gene sequences. Additionally, in order to acquire more targeted results, the mRNA sequence was divided into exon regions by computing the start and end nucleotides of each exon. Once this was done, each exon sequence was separately subjected to BLAST search in order to obtain targeted homology with human gene sequences. Different parts of the mRNA sequence were found to be homologous to transcript sequences of catalytic subunits of different protein kinases found in humans.

Protein Sequence Analysis

The mRNA sequence was translated using the ExPaSy tool to obtain all possible peptides for all six reading frames of the mRNA sequence. From all the Open Reading Frames (ORFs) obtained, there was a single 493 amino acid-long protein and the rest were all approximately 50 amino acid-long peptides. The long protein was found in frame 1 of the forward strand of the mRNA sequence. The sequence of the 493 amino acid-long protein was subjected to a BLAST search in the Protein Data Bank (PDB) database. The second half of the protein from amino acids 237 – 493 was found to be similar to part of the catalytic subunit alpha-1 in a protein kinase in humans. The same region was also found to be homologous to a serine/threonine protein kinase MARK2 and a calcium/calmodulin dependent protein kinase in humans indicating that this region probably has kinase activity.

In order to acquire possible homology between the first part of the unknown protein sequence and another protein in the PDB database, only the first 240 amino acids were subjected to a BLAST search. The results showed homology between the unknown protein sequence and the heavy chain of the Cod-v fab gene, parts of heavy and light chains of Ig, and H chain of R17 in humans. The protein sequence was also subjected to BLAST search in Uniprot website and it gave similar results as above. Therefore, to sum up the results, as seen in Figure 1, the first part of the unknown protein belongs to the Ig superfamily and the second part of the unknown protein belongs to the protein kinase superfamily.

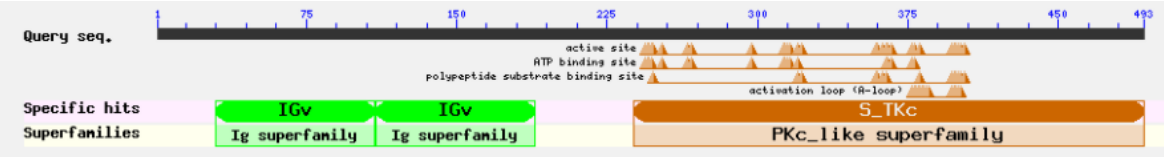

Figure 1: Possible domain structures of the unknown protein. The first half of the unknown protein belongs to the Ig superfamily and the second half of the unknown protein belongs to the protein kinase superfamily.

Once the domain structure of the protein was obtained, the protein sequence was modelled using SWISS MODEL. The acquired figure further proved that the second part of the protein was similar to other protein domains that had protein kinase activity and the first part of the protein was similar to proteins with Ig domains.

Conclusions

Butyrophilin is an important immune modulator and belongs to the human immunoglobulin superfamily of receptor proteins. As they are immune regulators, they come into action when the body is required to mount an immune response against a foreign body. Most of these genes are known to have formed due to events such as gene duplication, diversification, deletion, and formation of pseudogenes. These proteins typically have two Ig domains – IgV and IgC2 – apart from a cytosolic domain (Rhodes et al., 2015).

Fcamr gene codes for a receptor protein that binds to the Fc region of both IgM and IgA molecules. The alpha chains of these receptor proteins are specifically responsible for binding to the immunoglobulin molecules. The receptor proteins bind to the IgM or IgA molecules with high affinity during an immune response and mediate their endocytosis. As it is a transmembrane receptor protein, it uses endocytosis as a means to transport the immunoglobulin molecules from one location to another (Akula et al., 2014).

Protein kinases are regulatory enzymes that function by adding a phosphoryl group from adenosine triphosphate (ATP) to serine and threonine side chains in proteins. Protein kinases represent around 2 to 3% of all genes in humans due to their important regulatory roles in various metabolic pathways. Although these kinases are highly specific in the pathways they regulate, the sequences of the catalytic subunits are found to be highly conserved across all protein kinases (Wang and Cole, 2014).

After a thorough analysis of the sequences of the promoter, unknown gene, unknown mRNA, and unknown protein, several inferences can be made. The unknown promoter sequence is homologous to the promoter sequence of the butyrophilin gene. The butyrophilin protein is an immune modulator and so, the promoter is activated in the presence of an immune response. Upon analysis of the gene, mRNA, and protein sequences, it is seen that there are two distinct functional domains in the protein. The first domain belongs to the Fc receptors superfamily and is capable of binding immunoglobulins IgM and IgA during an immune response. However, unlike the Fcamr protein that also has a transmembrane domain responsible for endocytosis and transport of immunoglobulin molecules, the second part of the unknown protein has protein kinase activity. This indicates that the unknown protein lacks a transmembrane domain and is cytosolic rather than a transmembrane protein. The presence of protein kinase activity indicates that the unknown protein can bind to immunoglobulins and either activate or deactivate them by transferring phosphoryl groups to their serine or threonine side chains.

The possibility that the unknown gene has arisen as a result of exon shuffling is very high. The reason is that neither the butyrophilin gene nor the *Fcamr* gene possesses both the immunoglobulin binding domain and protein kinase catalytic domain. Both these domains bear

similarity to two different proteins and it is very likely that they have come together as a result of recombination between two different gene sequences. Exon shuffling is an evolutionary process that aims to combine exons from different genes in order to create new genes with unique functions (Long et al., 1996). In this case too, the two protein domains seem to belong to exons from two different genes thereby conforming to the rules of the molecular mechanisms of exon shuffling.

References

Akula, S., Mohammadamin, S., & Hellman, L. (2014). Fc Receptors for Immunoglobulins and Their Appearance during Vertebrate Evolution. *PLoS ONE, 9*(5). doi:10.1371/journal.pone.0096903

Long, M., Souza, S. J., Rosenberg, C., & Gilbert, W. (1996). Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proceedings of the National Academy of Sciences, 93*(15), 7727-7731. doi:10.1073/pnas.93.15.7727

Rhodes, D. A., Reith, W., & Trowsdale, J. (2016). Regulation of Immunity by Butyrophilins. *Annual Review of Immunology, 34*(1), 151-172. doi:10.1146/annurev-immunol-041015-055435

Wang, Z., & Cole, P. A. (2014). Catalytic Mechanisms and Regulation of Protein Kinases. *Methods in Enzymology Protein Kinase Inhibitors in Research and Medicine, 548*, 1-21. doi:10.1016/b978-0-12-397918-6.00001-x