William McElhenney

*Initial Sequence Evaluation*

The first step performed was BLAST and FASTA searches with the three given sequences (gene, mRNA, and promoter). BLAST searches were ran against the Mammalian taxon with otherwise default settings. FASTA searches were ran against the Mouse and Human databases as running against the full Mammalian database produced results that, while they aligned well, did not tell us much about the sequence.

For the gene sequence, BLAST returned a high similarity to *Mus musculus* Fcamr gene for Fca/m receptor (ACC: AB071978.1) at 99.66% identity, an E-value of 0.0, and query coverage of 70%. The other results from this search corresponded to the same gene. The default FASTA search returned similar results with Fcamr being the first in the list with lowest E-value.

For the mRNA sequence, BLAST and FASTA searches returned high similarity to the same *Mus musculus* Fcamr gene as the gene sequence did.

The promoter sequence provided a few new and relevant details. The BLAST and FASTA searches for this sequence indicated high similarity to the *Mus musculus* butyrophilin (Btn1a1) gene, *promoter region,* and complete cds (ACC: U67065.1). Intriguingly, it can be seen in these results that the bp region 510 – 605 seems to have high repetition in the mouse genome aligning well multiple chromosomes (average E-value is approximately 1e-26 for this portion of the promoter), which indicates a shared signaling motif within this promoter (we did not try to identify it). Butyrophilin is a nice result, as Entrez states that “Butyrophilin is the major protein associated with fat droplets in the *milk* . . . and may have arisen relatively recently in evolution by the *shuffling of exons* between 2 ancestral gene families.”1

*Exon Sequence Evaluation*

Having completed the naive approach to identifying these sequences, the exons from the gene were then run against BLAST and FASTA. For the BLAST settings, the word size was reduced to 16, and the filter for low-complexity regions was turned off. FASTA settings were left as before in *Initial Evaluation*.

Exon 1 (bp 1 – 510) aligned well with the already identified Fcamr gene in *Mus musculus* in both BLAST and FASTA.

Exons 2 and 3 (bp 1401 – 1640 and bp 2299 – 2538) are the same sequence. This sequence aligned with a new gene in BLAST, *Homo sapiens* gene for immunoglobulin heavy chain variable region (ACC: AB203310.1), with an E-value of 0.042, 96% identity, and 11% query coverage. This is not a particularly high quality alignment, but it is interesting in relation to Butyrophilin, as Butyrophilin is a member of the Immunoglobulin superfamily.1 These results were mirrored in FASTA but with much better statistics (E-values were below 1e-10 and query coverage was greater than 50%).

Exon 4 (bp 2941 – 3081) produced 1 very poor result in BLAST, but FASTA produced several results indicating a likeness to Butyrophilin. These results further implicate this gene as having a similar structure/function to Butyrophilin.

Exon 5 (bp 3672 – 5121) returned very good likeness to the Fcamr gene in both BLAST and FASTA.

*Protein Structure and Function Evaluation*

The mRNA was translated using the ExPASy translate tool. In 5’3’ Frame 1 a 493 residue sequence was found. The length of this sequence is comparable to human and mouse Butryophilin (available data suggests both range from about 450 to 530 amino acids in length), and represents the largest contiguous area of the coding sequence available. This sequence is taken directly from the middle of the mRNA and largely cuts out the FCAMR corresponding portions of the mRNA at the ends of the sequence.

This sequence was then ran through BLASTp and FASTA to determine similarity to existing proteins. BLASTp with the BLOSSUM62 matrix found this protein to be most similar to a variety of mammalian calcium/calmodulin-dependent protein kinase type 1s (nearly all the results from restricting the BLASTp search to mammals resulted in finding similarity to these kinases).identified three conserved domains. The first two of these domains were for immunoglobulin-variable heavy chain family, which corresponds with our results from aligning exons 2 and 3. The third domain is for the catalytic domain of serine/threonine protein kinase family, and takes up approximately the last-half of the protein. These domains were confirmed by analysis with PROSITE, Pfam, and InterPro.5 The FASTA results with BLOSSUM80 mirror the BLASTp results with all the results being for different forms of serine/threonine protein kinase. The Ig-V Heavy portions were identified using the same BLASTp seach strategy, but the query range was limited to the relevant parts of the protein (residues 1 – 225) to prevent overtake by kinase results. This search indicated that the front-half of the protein was significantly similar to several variants of Ig-V and Ig-V Heavy chains. Mirroring this change in procedure in FASTA allowed us to mirror the results there as well.

Using PHOBIUS and TMMHMMwe identified the leading immunoglobulin portion of this protein to be extracellular, while the lagging kinase portion appears to be cytoplasmic5. These two domains are separated by a strong transmembrane signal.5 However, using a SignalP to analyze this protein, we found the leading portion of the protein to most likely be a signal peptide (98.86% likely).5 Structural predictions with PHD and PSIPRED on these domains by suggest that the immunoglobulin portion of this protein begins with an alpha helix and is followed by beta stranding into the catalytic domain and the catalytic domain itself is composed of a series of helices interspersed with beta stranding.5

Models were made using SWISS-MODEL and I-TASSER of both the protein as a whole, and the individual domains themselves (Ig-V Heavy and kinase).5,6 Note that while both programs were able to find templates that fit well with the separated regions of the protein they would only find kinases for the full protein. The templates found corresponded well with the structural predictions for the protein.

In order to try and determine function of the protein, the protein structure was compared to both Fcamr and Butryophilin using data from InterPro available on representative sequences for the two genes. Both Fcamr and Butryophilin have immunoglobulin like areas in what would be the leading region (Ig-V Heavy) of our protein. The lagging region is undefined in Fcamr, but is named SPRY in Butryophilin which maybe involved in innate immunity.7 The structural predictions for

References:

1. <https://www.ncbi.nlm.nih.gov/gene?cmd=Retrieve&dopt=full_report&list_uids=696>

2. [https://www.uniprot.org/uniprot/Q13410#function](https://www.uniprot.org/uniprot/Q13410" \l "function)

3. https://www.omim.org/entry/601610

4. <https://github.com/WillMc93/AS.410.633-Final/>

5. https://github.com/WillMc93/AS.410.633-Final/tree/master/Report/Supplementary

6. <https://github.com/WillMc93/AS.410.633-Final/tree/master/Model>

7. D'Cruz, A. A., Babon, J. J., Norton, R. S., Nicola, N. A., & Nicholson, S. E. (2012). Structure and function of the SPRY/B30.2 domain proteins involved in innate immunity. *Protein science : a publication of the Protein Society*, *22*(1), 1–10. doi:10.1002/pro.2185

8. https://www.nature.com/articles/nri3818