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 run against the Mammalian taxon with otherwise default settings. FASTA searches were run against the Mouse and Human databases as running against the full Mammalian database produced results that aligned well but did not tell us much about the sequence.

For the gene and mRNA sequences, BLAST returned a high similarity to *Mus musculus* Fcamr gene for Fca/m receptor (ACC: AB071978.1) at 99.66% identity to the gene (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.

The promoter sequence provided a few 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, these results indicate that bp region 510 – 605 seems to have high repetition in the mouse genome aligning well with multiple chromosomes (average E-value is approximately 1e-26 for this portion of the promoter), which indicates a shared signaling motif within this promoter. 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 naïve approach to identifying these sequences, the previously identified 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 Sequence 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, respectively) 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 good likeness to the Fcamr gene in both BLAST and FASTA.

In full, the results from analyzing these exons look to be good evidence for exon shuffling, as the results are consistently from one of two different gene families (Fcamr/butyrophilin).

*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 Butyrophilin (available data suggests both range from about 450 to 530 amino acids in length) and represents the largest contiguous area of 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 either end of the sequence.

This sequence was then run 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). BLASTp identified three conserved domains. The first two of these domains were for immunoglobulin-variable heavy chain family, which corresponds with the 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 search 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 this 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 TMMHMM, the leading immunoglobulin portion of this protein was identified 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, the leading portion of the protein was found 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 built using SWISS-MODEL and I-TASSER of both the protein as a whole and the individual domains (Ig-V and catalytic).6 The models were in keeping with what was expected from the structural predictions made above.

The I-TASSER server’s CO-FACTOR and COACH tools were used to determine potential ligand binding. The results for the Ig region of the protein feature some interesting potentials in that four out of the five potentials were for antibodies (PDBs: 1A4KB, 2v17H, 2igfH, 1rivH). Immunoglobulin-immunoglobulin (antibody-antibody) interactions have been linked to the autoimmune system and implicate this protein to behave as a secondary antibody.10

To explore further, the *Homo sapien* representatives for Fcamr and Butyrophilin were put into a FASTA file with the translated protein and aligned using several MSA techniques (Clustal Omega, MAFFT, MUSCLE, Expresso, and M-Coffee).9 These MSAs indicated high levels of homology in the leading regions of the respective proteins. Plugging the Fcamr and Butyrophilin representatives into InterPro it was found that both have Ig like regions in the same region as the translated protein. The lagging region is undefined in Fcamr, but Butyrophilin’s is named SPRY and may be involved in innate immunity.7

*Conclusions*

Evidence indicates this novel gene is likely the product of exon shuffling between an Fcamr-like and a Butyrophilin-like (BTN1A1-like) gene. The corresponding protein is likely a membrane-bound receptor for a signal transduction event related to the binding of some antigen (or more likely a primary antibody) to immunoglobulin-like extracellular region of the protein and may be involved in the innate immune system in antigen presentation, as Butyrophilin has been described to do.7,8 Physiologically, if this protein is involved with the innate immunity, it is likely that this protein causes an immune response in order to keep harmful antigens from harming the infant.

These conclusions are derived from the analyses of the conserved immunoglobulin-like structure at the beginning of this protein that aligns well with representative sequences for both Fcamr and Butyrophilin. In keeping with this line of reasoning, the catalytic domain of this protein likely activates a downstream enzyme via phosphorylation.

[Supplementary Materials, i.e. results and graphics, available at <https://github.com/WillMc93/AS.410.633-Final/>]

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>
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>
9. <https://github.com/WillMc93/AS.410.633-Final/tree/master/MSAs>
10. Kawa, S., Kitahara, K., Hamano, H., Ozaki, Y., Arakura, N., Yoshizawa, K., … Bahram, S. (2008). A novel immunoglobulin-immunoglobulin interaction in autoimmunity. *PloS one*, *3*(2), e1637. doi:10.1371/journal.pone.0001637. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2249926/>