JULIE GARCIA

INTRO TO BIONIFORMATICS – FINAL PROJECT – April 29, 2017

Promoter Analysis

To analyze the promoter sequence, I entered it into Promoter 2.0 Prediction Server and found three possible promoter regions, two marginal predictions at positions 400 and 1000, and one highly likely prediction at position 1700. I then searched FPROM, which found 2 promoter regions, one at position 315 (with a TATA Box upstream at 284), and the second at 1554 (TATA Box at 1525). CONSITE also found TATA boxes at 283 and 1524. The tools TSSG and TSSW did not find any promoters. CpGFinder found zero CpG Islands in the sequence, even when I lowered the length parameter to 100 bps.

Using PROMO, I found 71 possible binding sites for transcription factors. TBP binding sites, which indicate TATA box, were found at 615 and 750, which was inconsistent with my previous findings. Binding sites for TFIID, which are generally associated with the core promoter region and help RNA Polymerase to bind, were found in clusters between positions 200-300, 500-650, 740-860, 1040-1170, one at 1500 and between 1860-2020. C/EBPbeta binding sites, also known as CAAT boxes were found clustered between 200-300, 500-660, 730-860, 1080-1230 and 1960-2040.

I ran BLASTn on the promoter sequence and found several homologous sequences, including an identical promoter region in Mus Musculus for the gene butyrophilin (*Btn1a1*) in the mammary gland with accession number U67065.1. This was the complete CDS and promoter for this gene. Other homologous sequences were mutant alleles of this gene, several clones, and a 99% similarity with the *Btn1a1* gene of Rattus norvegicus. Since the promoter region of *Btn1a1* in Mus Musculus was an exact match, I searched for it in the Eukaryotic Promoter Database (EPD) and found that it has two promoters.

I then ran BLASTn searching only human nucleotide sequences and found several highly homologous sequences in the 300-600 region of the promotor. These were partial sequences from clones, ATP/GTP binding protein genes in addition to several non-coding RNA matches. This suggests that this gene could be silenced by a non-coding RNA. Interestingly the promoter region of the novel gene did not match the promoter region of the human Btn1a1 gene. Combining this data with the promoter prediction data, I would estimate that the core promoter region is between positions 250-350 with a possible alternative promoter around position 1525.

Gene Analysis

To analyze the gene sequence, first I ran BLASTn to find homologous sequences. The results showed the best match was from the Mus Musculus *Fcamr* gene for the Fca/m receptor. Several other close matches were found including a mutant allele from *Fcamr*, variants of the *Fcamr* gene and several Mus Musculus clones. The only other species where there was a highly homologous match was in an mRNA ribosomal sequence in the Mesitornis unicolor (bird) species. Running BLASTn with "more dissimilar sequences" checked, gave me results from several different species. Most of these new hits were from mRNA whose products were protein kinases. Included in these results I also found the mRNA for mouse and rat butyrophilin in the mammary gland, for which the promoter region also was also a match.

To find the sequence on the human genome, I ran BLASTn, and filtered by human. Here the results gave several DNA, mRNAs and clones for many different types of protein kinases in addition to matches with genes that contained immunoglobulin domains. The immunoglobulin domains matched up with exon 2 of the DNA (pos. 1401-1640) and exon 3 (pos. 229-2538), while the protein kinase homologies were in exon 5 (pos. 3672-5121). I also found a position on the genome with a 74% homology with accession number NC_000007.14. I ran a search at this position through GEO Profiles, to find any expression data associated with this locus. I found expression data in the mammary epithelial cell line of a cyclin-dependent kinase. Most datasets found were in relation to studies of breast cancer.

Using ORFFinder I found 40 possible open reading frames (ORFs). The longest ORF was 546 nucleotides, located at position 3900, which is inside exon 5. Running SMARTBLAST on this ORF came up with five hits, from five different species. These were all in PKc_like domains, which are protein kinase domains. This matches up with what I found in the BLASTn results for the gene sequence. ORF4, at position 2935-3306 is 372 nt long. Running SMARTBLAST on this sequence came up with the Mus Musculus butyrophilin protein again. Other matching ORFs found significant homologies in hypothetical proteins of various species at several positions along the sequence.

mRNA Analysis

To analyze the mRNA sequence, I first ran a BLASTn search on the non-redundant nucleotide database. As expected, exact matches came up for the complete CDS and several variants of the Mus Musculus Fcamr gene, as well as the Mesitornis unicolor ribosomal mRNA. Expanding the search, I ran BLASTn again with “more dissimilar sequences” checked and again found the Mus Musculus mammary gland butyrophilin mRNA. I also ran BLASTx, to search only homologous proteins from similar translated sequences, and found numerous matches with proteins that have protein kinase functions. These proteins lined up with the area of the sequence that would have been spliced from exon 5 of the original sequence. A tBLASTx and BLASTx search only on human nucleotide sequences and proteins brought up hundreds of matches, mostly serine/threonine kinase proteins and mRNA, in addition to several proteins with immunoglobulin domains. Again, the immunoglobulin domains matched up with exon 2 and 3 of the original gene sequence, and the protein kinase domains matched up with exon 5.

ORFFinder found a very long ORF from position 424-1905 that when searched in SMARTBLAST, matched up with the protein kinase regions in various species. Specifically, the catalytic domain of CAMK family Serine/Threonine Kinases. Running BLASTp on this ORF again found the two conserved Ig-like domains and protein kinase domain in the expected locations. One interesting find was a protein kinase from the species Dictyostelium dicoideum, that when analyzing the taxonomy tree appears to have a common ancestor to humans.

Protein Analysis

I translated the mRNA in ExPASy Translate, and found the protein sequence that matched up with the longest ORF in the mRNA Analysis. The sequence is 493 aa long and is attached on the final page.

First, I wanted to find any known homologous proteins. I ran BLASTp, against the non-redundant protein database. The results showed two putative conserved domains in the Ig superfamily, and one in the protein kinase superfamily. The portion of the sequence matching the protein kinase superfamily, from ~240-493 found homologies in proteins with kinase domains from several different species. PROSITE confirmed the three domains found with BLASTp. Two Ig-like domains (30-118 aa, 127-190 aa) and a protein kinase domain (238-493 aa). All of this was consistent with the domains found in the mRNA sequence.

Since I didn't find the mammary gland protein as a match, I ran a pairwise alignment in BLASTp, with the novel protein sequence and one of the Mus Musculus butyrophilin accession numbers, XP_006516598.1. This resulted in a 37% match between amino acid 90-200. This is roughly in the second Ig-like domain of the novel protein.

To find the secondary structure of this novel protein I ran PHD, and found that the protein is 32.45% Alpha Helix, 32.25% Extended Strand, and 35.29% Random Coil. It showed that the Ig-like domains are made up of completely of extended strand and random coil. The protein kinase domain contains 9 alpha helices. TMHMM showed two transmembrane areas at position 1-30 and position 190-210 of the protein. These transmembrane areas surround the the two Ig-like domains and suggests that these domains are protruding from the membrane. SignalP found a peptide signal at position 1-30. Phobius showed similar results to TMHMM, with the Ig-like domains in the non-cytoplasmic region and the kinase domain in the cytoplasmic region.

Combining this data, I believe that the protein is an integral membrane bound protein. It starts with a signal peptide in the intracellular region, has an alpha helix embedded in the membrane, the two Ig-like domains are extracellular, followed by another alpha helix embedded in the membrane, ending with the protein kinase domain in the intracellular region. ProtFun showed that this protein is an enzyme in the cell envelope, that has immune response functionality.

Conclusion

The protein product of this novel gene, in the splice variant given, is an integral membrane bound protein that has an intracellular protein kinase domain, most likely serine/threonine specific. Protein kinases transfer phosphoryl groups from ATP to other proteins, switching them on or off, as part of a cell signaling pathway. Serine/threonine specific kinases generally play a role in cell proliferation, cell differentiation, and embryonic development.

This novel protein also has two extracellular immunoglobulin (Ig) domains. Ig-like domains sit on the outside of the cell and bind foreign particles, like viruses, bacteria, or other large foreign molecules and flag them for degradation. We know this protein originates in the mammary tissue of humans and it has high homology to the protein butyrophilin in mice (in both the promoter region and the gene sequence) and humans. Butyrophilin in humans is associated with the production of milk droplets in lactating women and has two Ig-like domains, in addition to a B30.2 SPRY domain, which is involved in calcium release channels in mammals. This domain was originally discovered in a Dictyostelium dicoideum kinase, which showed high homology with the protein kinase domain in our novel protein. UniGene EST Profile Viewer also showed butyrophilin expressed in the lymph nodes of the mammary tissue in humans. I would hypothesize that this novel gene is involved in the mammary epithelial cell signaling pathway, possibly during embryonic development or in breast cancer.

For exon shuffling to have occurred, this gene would have been derived by mixing the exons from another gene. I believe that there is evidence for exon shuffling in this protein.

Immunoglobulins have often been derived by exon duplication. They are made up of repeated units, as we can see in this novel protein with two repeated Ig units. There is also evidence that this gene has reused portions of other genes, such as the protein kinase region that has very high similarity with several other protein kinase domains in other proteins and other species, for example Dictyostelium dicoideum which shares a common ancestor with humans. Each of these three domains match up with exons located in the gene sequence.

REFERENCES

Nelson, David L., Michael M. Cox, and Albert L. Lehninger. *Lehninger principles of biochemistry*. New York: W.H. Freeman, 2017. Print.

Ogg, S. L., A. K. Weldon, L. Dobbie, A. J. H. Smith, and I. H. Mather. "Expression of butyrophilin (Btn1a1) in lactating mammary gland is essential for the regulated secretion of milk-lipid droplets." *Proceedings of the National Academy of Sciences*101.27 (2004): 10084-0089. Web.

Watson, James D. *Molecular biology of the gene*. 7th ed. Boston: Pearson, 2014. Print. p. 497-500

https://www.ncbi.nlm.nih.gov/protein/166197658
http://www.cbs.dtu.dk/cgi-bin/webface2.fcgi?jobid=5903CAE900004C55881A42D6&wait=20
http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3
http://linux1.softberry.com/cgi-bin/programs/promoter/fprom.pl
http://linux1.softberry.com/cgi-bin/programs/promoter/tssg.pl
https://blast.ncbi.nlm.nih.gov/Blast.cgi
http://linux1.softberry.com/cgi-bin/programs/promoter/cpgfinder.pl
http://web.expasy.org/cgi-bin/translate/dna_aa
https://www.ncbi.nlm.nih.gov/orffinder/
https://www.ncbi.nlm.nih.gov/nucleotide/1246078?report=genbank&log$=nuclalign&blast_rank=52&RID=G80MNFU3014
http://rfam.xfam.org/
https://npsa-prabi.ibcp.fr/cgi-bin/secpred_phd.pl
http://prosite.expasy.org/cgi-bin/prosite/ScanView.cgi?scanfile=96036875075.scan.gz
http://www.cbs.dtu.dk/cgi-bin/webface2.fcgi?jobid=5904254700006425A4A95BCE&wait=20
http://www.cbs.dtu.dk/cgi-bin/webface2.fcgi?jobid=5904280600006425BAC026F3&wait=20
http://phobius.sbc.su.se/cgi-bin/predict.pl
https://www.ebi.ac.uk/interpro/entry/IPR001870
https://www.ncbi.nlm.nih.gov/UniGene/ESTProfileViewer.cgi?uglist=Hs.153058

Supporting Materials:

>novel protein, homologous with CAMK kinases

MACLWSFSWPSCFLSLLLLLLLQLSCSYASVTLSCKASGFTFSSYYVSWVRQPPGKGLE
WLGYIGSDVSYSEASYKGRVTISKDNSKNDVSLTISNLRVEDTGTYYCAVSVTLSCKAS
GFTFSSYYVSWVRQPPGKGLEWLGYIGSDVSYSEASYKGRVTISKDNSKNDVSLTISNLR
VEDTGTYYCAVTPWIVAVAIILLALGFLTIGSIFFTWKLYKERSSLRKKEFGSKERLLEYE
LGEKLGSGAFGKVYKGKHKDTGEIVAIKILKKRSLSEKKKRFLREIQILRRLSHPNIVRLL
GVFEEDDHLYLVMEYMEGGDLFDYLRRNGLLLSEKEAKKIALQILRGLEYLHSRGIVHR
DLKPENILLDENGTVKIADFGLARKLESSSYEKLTTFVGTPEYMAPEVLEGRGYSSKVDV
WSLGVILYELLTGKLPFPGIDPLEELFRIKERPRLRLPLPPNCSEELKDLIKKCLNKDPEKR
PTAKEILNHPWF