# Investigation of the gene which causes muscle and connective tissue to turn to bone

Alexia McGregor, s1789372 October 2017

## Question 1

### Is FOP genetic?

The name of the disease I have selected to research is Fibrodysplasia Ossificans Progressiva, more commonly referred to as simply FOP. FOP causes the muscle tissue and connective tissue of a sufferer to be replaced by bone, generally after trauma (even very mild) to a body part [2]. There has been found to be a genetic basis for the disease, with the gene mutating in each individual with the disease, rather than it being hereditary [1].

The human name for the gene that is thought to be involved is activin A receptor type 1 and the official symbol is ACVR1 [3]. This gene is known by other names, namely [2]

- activin A receptor, type I
- activin A receptor, type II-like kinase 2
- activin A type I receptor
- activin A type I receptor precursor
- ActR-IA protein, human
- ACTRI
- ACVR1_HUMAN
- ACVR1A
- ACVRLK2
- ALK2
- hydroxyalkyl-protein kinase
- SKR1

The NCBI website also lists FOP and TSRI as additional names for the ACVR1 gene [3]. Throughout this assignment, I will refer to the gene as both the ACVR1 gene and FOP gene interchangably.

### Are there any homologues for the FOP gene?

'A homologous gene (or homolog) is a gene inherited in two species by a common ancestor. While homologous genes can be similar in sequence, similar sequences are not necessarily homologous.' [7]

To investigate whether there were any homologues present in other species, I took the following steps:

- Searched for ACVR1 on the Homologene database on the NCBI website. Under the 'Genes' section, there was list of genes which have been selected as probable homologues by Homologene.

- Next, to try and test to the best of my ability whether any were true homologues (that they really, or most likely, evolved from the same ancestor, rather than just being coincidentally similar sequences), I downloaded the coding sequence of one of the potential homologues and ran the BLAST search given to me in the lab in lecture 3.

- If the BLAST search returned the original ACVR1 gene in homosapiens, and the e-value was smaller than $10^{-10}$ [8] then I concluded that the two were most likely true homologues.

I first tested the Acvr1/activin A receptor, type 1 gene [3] which is present in a house mouse (M.musculus [15]). The BLAST search returned the original human ACVR1 gene with an e-value of zero, and so I conclude that the Acvr1 and ACVR1 genes are homologues.

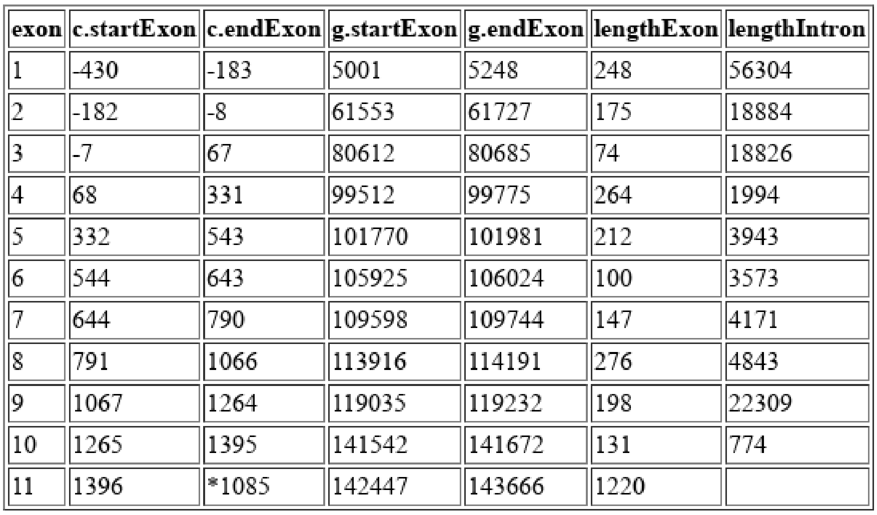
I next tested the sax/saxophone gene [3] in the common fruit fly(D.melanogaster [16]). However, this BLAST search did not return to me any gene for homosapiens at all that had a threshold e-value of less than 0.05, and so I conclude that the sax and the ACVR1 gene are not truly homologous.

# Question 2

## Information about the FOP gene and its transcripts

The location of the ACVR1 gene is 2q24.1 [3]. This means that the gene is located on the second chromosome, at region 2, band 4, sub-band 1 of the longer arm.

There are 16 exons in the genomic content [3] and 15 introns. This NM_001105.4 mRNA transcript, however, has 11 exons and 10 introns [4].



| exon | c.startExon | c.endExon | g.startExon | g.endExon | lengthExon | lengthIntron |
|---|---|---|---|---|---|---|
| 1 | -430 | -183 | 5001 | 5248 | 248 | 56304 |
| 2 | -182 | -8 | 61553 | 61727 | 175 | 18884 |
| 3 | -7 | 67 | 80612 | 80685 | 74 | 18826 |
| 4 | 68 | 331 | 99512 | 99775 | 264 | 1994 |
| 5 | 332 | 543 | 101770 | 101981 | 212 | 3943 |
| 6 | 544 | 643 | 105925 | 106024 | 100 | 3573 |
| 7 | 644 | 790 | 109598 | 109744 | 147 | 4171 |
| 8 | 791 | 1066 | 113916 | 114191 | 276 | 4843 |
| 9 | 1067 | 1264 | 119035 | 119232 | 198 | 22309 |
| 10 | 1265 | 1395 | 141542 | 141672 | 131 | 774 |
| 11 | 1396 | *1085 | 142447 | 143666 | 1220 | |

*Table for a transcript of NM_001105.4 's exons and introns taken from the Leiden open variation database [4].*

The percentage of exons in the transcript is the sum of the length of all the exons dicided by the length of the transcript (both exons and introns), i.e. $\frac{3{,}045}{138{,}666} \times 100 =$ 2.20% and the percentage of introns is calculated in a similar way, i.e. $\frac{135{,}621}{138{,}666} \times 100 =$ 97.8%.

I downloaded the transcript and coding sequences from the NCBI website, by searching for ACVR1 in the gene database, scrolling down the page to the NCBI Reference Sequences (RefSeq) section, clicking on the top entry NM_001105.4 and downloading both sequences as FASTA files. The unique identifier used so that someone else could be sure they were looking at the same sequences was NM_001105.4 [3].

I see from the running the sequences in the Python file given to me in lab 2, that the length of the transcript is 2300 and the length of the coding sequence is 1584, which means that the proportion that is coding is $\frac{1{,}584}{2{,}300} \times 100$, which is around 68.9%.

## Question 3

### Analysing the amino acids in the coding sequence of the FOP gene

By executing the Python code below, I can see that my protein contains all 20 amino acids as 20 letters are given in the output.

Listing 1: Python code to calculate the number of each amino acid in the FOP gene.

```python

from Bio import SeqIO
from Bio.SeqUtils.ProtParam import ProteinAnalysis
from Bio.SeqUtils import ProtParamData

ACVR1codingsequence = next(SeqIO.parse("ACVR1coding.txt", "fasta"))

proteinsequence=str(ACVR1codingsequence.seq.translate())

analysedproteinsequence = ProteinAnalysis(proteinsequence)

analysedproteinsequence.count_amino_acids()
```

The protein contains 61 out of the 64 possible codons, [5]. The three that it does not include are GCG, which encodes alanine, and TAG and TAA, which are both stop codons [5]. Again using my code above, I can see that the most common amino acid is Leucine (L), which is seen 53 times. Six codons for this amino acid exist [5]. The following table illustrates how often each is used:

| Leucine codon | Times used in FOP protein |
|---|---|
| CTT | 10 |
| CTC | 7 |
| CTA | 14 |
| CTG | 2 |
| TTA | 10 |
| TTG | 10 |

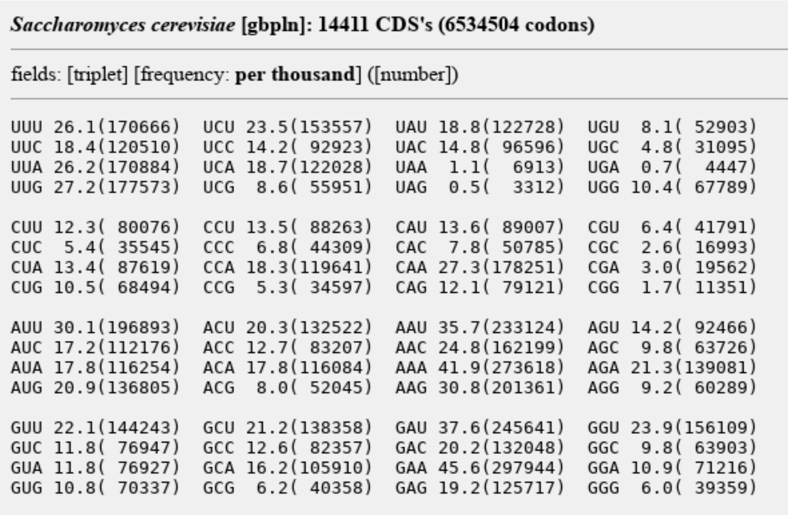
*Information in the table found using [5] .*

## Question 4

### What if I were to express the ACVR1 cDNA sequence in yeast?

"The codon usage database lists the frequency which each codon is used in a species (different species prefer different codons). Sequences which have too many rarer codons result in slowing down transcription and inhibition of protein expression - in extreme cases, rare codons are thought to introduce transcription errors when the rare tRNA is not available. There is no hard threshold, but generally codons with 1% usage or less are considered rare." [assignment sheet]

Codon usage table for yeast



***Saccharomyces cerevisiae* [gbpln]: 14411 CDS's (6534504 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

```
UUU 26.1(170666)  UCU 23.5(153557)  UAU 18.8(122728)  UGU  8.1( 52903)
UUC 18.4(120510)  UCC 14.2( 92923)  UAC 14.8( 96596)  UGC  4.8( 31095)
UUA 26.2(170884)  UCA 18.7(122028)  UAA  1.1(  6913)  UGA  0.7(  4447)
UUG 27.2(177573)  UCG  8.6( 55951)  UAG  0.5(  3312)  UGG 10.4( 67789)

CUU 12.3( 80076)  CCU 13.5( 88263)  CAU 13.6( 89007)  CGU  6.4( 41791)
CUC  5.4( 35545)  CCC  6.8( 44309)  CAC  7.8( 50785)  CGC  2.6( 16993)
CUA 13.4( 87619)  CCA 18.3(119641)  CAA 27.3(178251)  CGA  3.0( 19562)
CUG 10.5( 68494)  CCG  5.3( 34597)  CAG 12.1( 79121)  CGG  1.7( 11351)

AUU 30.1(196893)  ACU 20.3(132522)  AAU 35.7(233124)  AGU 14.2( 92466)
AUC 17.2(112176)  ACC 12.7( 83207)  AAC 24.8(162199)  AGC  9.8( 63726)
AUA 17.8(116254)  ACA 17.8(116084)  AAA 41.9(273618)  AGA 21.3(139081)
AUG 20.9(136805)  ACG  8.0( 52045)  AAG 30.8(201361)  AGG  9.2( 60289)

GUU 22.1(144243)  GCU 21.2(138358)  GAU 37.6(245641)  GGU 23.9(156109)
GUC 11.8( 76947)  GCC 12.6( 82357)  GAC 20.2(132048)  GGC  9.8( 63903)
GUA 11.8( 76927)  GCA 16.2(105910)  GAA 45.6(297944)  GGA 10.9( 71216)
GUG 10.8( 70337)  GCG  6.2( 40358)  GAG 19.2(125717)  GGG  6.0( 39359)
```

*Table found using [6] .*

The codon frequency table is given in frequency per thousand. Therefore, to find codons with 1% usage or less, I need to search for the codons which have a frequency of 10 per thousand are less. These codons are: UCG, UAA, UAG, UGU, UGC, UGA, CUC, CCC, CCG, CAC, CGU, CGC, CGA, CGG, ACG, AGC, AGG, GCG, GGC and GGG. I referred back to the codon usage table for the ACVR1 protein [5] and saw that these codons all occur in my protein except for

UAA, UAG (both stop codons) and GCG. Therefore, all except those three may cause problems if I were to try and express my human cDNA in yeast.

## Question 5

'You are given the following coding sequence fragments. They encode a homologous proteins

in different species. The sequences are aligned to the correct reading frame:

- 1. CTGAAGCGGGAGGCTGAGACGCTGCGGGAGCGGGAGGGC
- 2. CTCAAGCGTGAGGCCGAGACCCTACGGGAGCGGGAAGGC
- 3. GAAGAGCTGAAGAGAGAGGCTGACAATTTAAAGGACAGA
- 4. AACGAGGAGCTCAAGCGAGAAGCTGATACGCTGAAGGAC' [assignment sheet]

### Finding the most likely gene name and its most likely species

In order to go about this problem, I ran the snippets of coding sequences in the BLAST python program that was given to me in lab 3. This would provide me with various genes accompanied by an e-value, with a threshold e-value of being smaller than 0.05, indicating how likely it was that the snippet belonged to that gene. I selected the predictions with the smallest e-value as the correct prediction, and in the case for sequence 2, this was actually several predictions.

#### 1. CTGAAGCGGGAGGCTGAGACGCTGCGGGAGCGGGAGGGC

This coding sequence fragment was predicted to be Mus musculus potassium voltage gated channel, Shab-related subfamily, member 1 (Kcnb1), transcript variant X5, misc_RNA. This is the Kcnb1 gene of a house mouse [15].

#### 2. CTCAAGCGTGAGGCCGAGACCCTACGGGAGCGGGAAGGC

This coding sequence yielded multiple results with the same e-value. All of these were genes belonging to the primate family, with one of these being homosapiens. The predicted genes were:

- Papio anubis potassium voltage-gated channel subfamily B member 1 (KCNB1),
- Gorilla gorilla gorilla potassium voltage-gated channel subfamily B member 1 (KCNB1)
- Callithrix jacchus potassium voltage-gated channel subfamily B member 1-like
- Cebus capucinus imitator potassium voltage-gated channel subfamily B member 1 (KCNB1)
- Homo sapiens potassium voltage-gated channel subfamily B member 1 (KCNB1)
- Pan troglodytes potassium voltage-gated channel subfamily B member 1 (KCNB1)
- Macaca fascicularis potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)
- Pan paniscus potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)

- Nomascus leucogenys potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)
- Cercocebus atys potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)
- Macaca nemestrina potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)
- Mandrillus leucophaeus potassium channel, voltage gated Shab related subfamily B, member 1 (KCNB1)
- Chlorocebus sabaeus potassium voltage-gated channel, Shab-related subfamily,

### 3. GAAGAGCTGAAGAGAGAGGCTGACAATTTAAAGGACAGA

This coding sequence fragment was predicted to be Fundulus heteroclitus potassium voltage-gated channel subfamily B member 1 (kcnb1), transcript variant X2, mRNA, misc_RNA. This is the kcnb1 gene of a mummichog, a type of small killifish [13].

### 4. AACGAGGAGCTCAAGCGAGAAGCTGATACGCTGAAGGAC

This coding sequence fragment was predicted to be Danio rerio potassium voltage-gated channel, Shab-related subfamily, member 1 (kcnb1), mRNA'. This is the kcnb1 gene of a zebrafish [14].

## Comparing the four sequences to each other

Different codons have different frequencies between species, as I looked at a little in question 4. However, it is possible for codons which differ slightly to produce the same amino acid. I can deduce this by reading the codon usage table for ACVR1 [5] and seeing that for each amino acid, there can be various amounts of different codons listed. The resulting proteins from sequences 1 and 2 are the same. However, the sequence of nucleotides in sequences 1 and 2 are slightly different, and hence the same proteins are formed by using slightly different codons as building blocks. For example, both sequences have one alanine amino acid, but it is produced by the GCT codon in sequence 1 and the GCC codon in sequence 2 [5].

### Degeneracy

The differences between the sequences 1 and 2 as I described above are an example of degenerate codons. As there are $4 \times 4 \times 4$ = 64 possible ways of sequencing four nucleotides in groups of three, there are 64 possible codons. However, since there are only 20 amino acids, then some amino acids must clearly have more than one codon associated with it [11], or else some of the nucleotide sequences would have no meaning and would not produce of a protein [12]. This concept is called degeneracy [11].

## Are there any functional implications of degeneracy?

The short answer is ’yes, there are functional implications of degeneracy!’.

### 1. To protect against mutations [12]

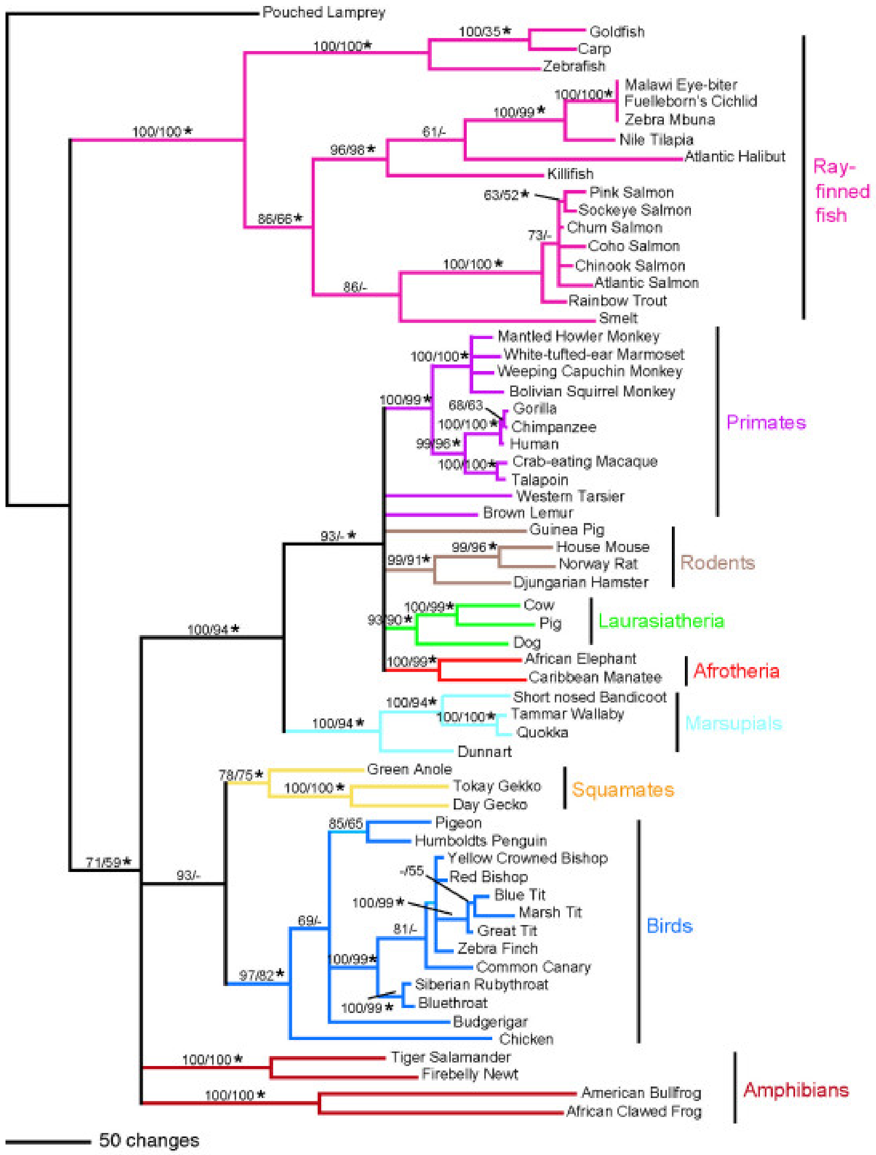
The first and most obvious is the reason I outlined in the previous section on degeneracy - that were not this not the case, then if some codon were to mutate, then this nucleotide sequence of three would hold no meaning, and so would inhibit the production of a protein [12].

### 2. Certain amino acid nucleotide sequences are more ’important’ than others. [12]

Three biologists, one from the university of Harvard and one from the university of Chicago, did some very interesting and revealing experiments. They discovered that if there was an abundance of amino acids, then there was no effect on how many or how quickly proteins were produced when different sequences of nucleotides were used. However, when there were only a few amino acids, then certain sequences would produce many proteins and some would produce hardly any in the same amount of time. This indicates that some amino acid codons are more ’important’ than others, in the sense that they will always try to produce a (probably more important) protein, no matter how much amino acid is available, whereas other amino acid codons do not. The biologists theorize that the more ’important’ amino acids are used in the more ’important’ proteins. [12]

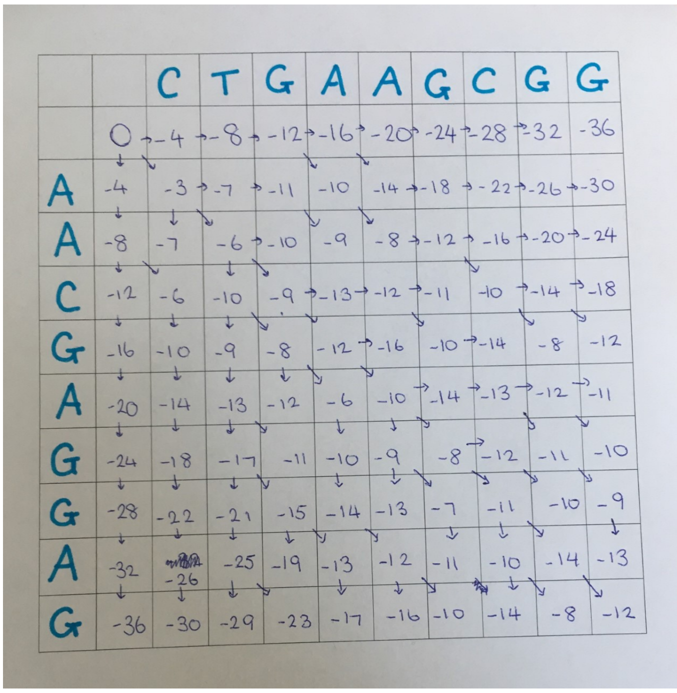
## Using the Needleman Wunsch algorithm to compare the four sequences

I next used the Needleman Wunsch algorithm, using an online calculator [9], to compare the first sequence to the other three. When compared against sequence 1, sequence 2 got a score of 48, sequence 3 -27 and sequence 4 -17 [9]. From these scores, I conclude that the first and second species are the most closely related of the four, and the third and fourth are not very related to the first, but out of those two, the fourth is more related. This is because the higher the Needleman Wunsch score, the more related a species is.



*Phylogenetic tree taken from [10]*

As I can see from the above phylogenetic tree [10], the primates are more closely related to the house mouse. The ray-finned fish are less closely related, but of the two, the zebrafish is more related than the killifish. I can see this as you need to travel down one less branch to find your way from house mouse to zebrafish than from house mouse to killifish. This confirms my findings that the scores from the Needleman Wunsch algorithm make logical sense.



| | | C | T | G | A | A | G | C | G | G |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | -4 | -8 | -12 | -16 | -20 | -24 | -28 | -32 | -36 |
| A | -4 | -3 | -7 | -11 | -10 | -14 | -18 | -22 | -26 | -30 |
| A | -8 | -7 | -6 | -10 | -9 | -8 | -12 | -16 | -20 | -24 |
| C | -12 | -6 | -10 | -9 | -13 | -12 | -11 | -10 | -14 | -18 |
| G | -16 | -10 | -9 | -8 | -12 | -16 | -10 | -14 | -8 | -12 |
| A | -20 | -14 | -13 | -12 | -6 | -10 | -14 | -13 | -12 | -11 |
| G | -24 | -18 | -17 | -11 | -10 | -9 | -8 | -12 | -11 | -10 |
| G | -28 | -22 | -21 | -15 | -14 | -13 | -7 | -11 | -10 | -9 |
| A | -32 | -26 | -25 | -19 | -13 | -12 | -11 | -10 | -14 | -13 |
| G | -36 | -30 | -29 | -23 | -17 | -16 | -10 | -14 | -8 | -12 |

*Performing the Needleman Wunsch algorithm comparing sequences 1 and 4 using only the first three codons*

## References

[1] Connor, J. M., & Evans, D. A. (1982). Genetic aspects of fibrodysplasia ossificans progressiva. *Journal of medical genetics, 19(1), 35-39.*.

[2] Fibrodysplasia ossificans progressiva - Genetics Home Reference. (n.d.). Retrieved October 24, 2017, from https://ghr.nlm.nih.gov/condition/fibrodysplasia-ossificans-progressiva

[3] ACVR1 activin A receptor type 1 [Homo sapiens (human)] - Gene - NCBI. (n.d.). Retrieved October 24, 2017, from https://www.ncbi.nlm.nih.gov/gene/90

[4] Netherlands, L. D. (n.d.). Retrieved October 24, 2017, from https://databases.lovd.nl/shared/genes/ACVR1

[5] (n.d.). Retrieved October 24, 2017, from http://www.bioinformatics.org/sms2/codon_usage.html

[6] (n.d.). Retrieved October 24, 2017, from http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=4932

[7] B. (n.d.). Bioinformatics. Retrieved October 25, 2017, from https://courses.lumenlearning.com/boundless-microbiology/chapter/bioinformatics/

[8] Pearson, W. R. (2013). An introduction to sequence similarity (“homology”) searching. Current protocols in bioinformatics, 3-1.

[9] Needleman-Wunsch and Smith-Waterman Calculator. (n.d.). Retrieved October 25, 2017, from http://bioinformaticnotes.com/Needle-Water/

[10] van Hazel, I., Santini, F., Müller, J., & Chang, B. S. (2006). Short-wavelength sensitive opsin (SWS1) as a new marker for vertebrate phylogenetics. BMC evolutionary biology, 6(1), 97.

[11] (n.d.). Retrieved October 26, 2017, from https://www.nature.com/scitable/topicpage/the-information-in-dna-determines-cellular-function-6523228

[12] Being Degenerate Can Be Very Good! (n.d.). Retrieved October 26, 2017, from http://blog.drwile.com/being-degenerate-can-be-very-good/

[13] Mummichog. (2017, October 18). Retrieved October 26, 2017, from https://en.wikipedia.org/wiki/Mummichog

[14] Zebrafish. (2017, October 25). Retrieved October 26, 2017, from https://en.wikipedia.org/wiki/Zebrafish

[15] House mouse. (2017, October 25). Retrieved October 26, 2017, from https://en.wikipedia.org/wiki/House_mouse

[16] Drosophila melanogaster. (2017, October 22). Retrieved October 26, 2017, from https://en.wikipedia.org/wiki/Drosophila_melanogaster