Analysis of *Drosophila melanogaster* cDNA Sequences to Model Human Disease

Matthew Ryan

Partner: Aaron Allender

BIOL 230W, Section 014

TAs: Mary McGoldrick & Ayesha Samad

10/21/15

Introduction

Diseases in humans are often caused by mutations in DNA that lead to protein misfolding and thus malfunctioning of the protein. Monitoring diseases that are caused by protein misfolding can be difficult in humans as the observations can sometimes be risky for human health. Proper protein folding is vital to ensure the proper function of the protein because the three-dimensional shape of the protein is necessary for its correct function. The protein may be folded into a globular shape for an enzyme or a fibrous shape based on how the protein folds. Protein misfolding can be catastrophic for the function of a protein. Human life spans are also too long to determine a specific gene as the cause of of the disease. Thus, scientists use the homology between proteins in different organisms to to understand the function of the protein.

Homology exists between proteins of different species which allows for comparisons to be drawn between the mechanisms that control the proteins themselves. This homology between proteins can be seen in protein domains which are regions of a protein that have been conserved throughout the evolution of an organism. These protein domains have specific specific functions that can function and evolve independently. Protein families, which are groups of related proteins that come from common ancestors, have have similar structures, function and sequence.[1] The organisms that scientists use that contain these homologous proteins are called model organisms.

Drosophila melanogaster is a very common model organism used in laboratory breeding because it is relatively inexpensive, easy to breed, well studied, has a short life cycle, and high levels of homology with *Homo sapiens*. These qualities of *Drosophila* make experimentation faster and allow scientists to observe results over many generations. *Drosophila* homologues

exist for around 70% of human cancer genes.[2] There are several homologous protein domains that have been previously discovered in other experiments already such as mutations in the protein domain LRRK2 of the fruit fly, which has homology to the DJ-1 protein domain in humans which has been linked to Parkinson's Disease by Mel Feany at Harvard Medical School.[2] Due to the extreme homology between *Drosophila* and *Homo sapiens,* there is great potential to use *Drosophila* to study human disease as many of the genes are present in both species. With the presence of the same genes in both organisms, it is reasonable to expect that similar effects would be seen from alterations in these genes and their proteins in each organism.

The goal of this experiment was to isolate and identify cDNA sequences in *Drosophila melanogaster* which codes for homologous proteins which will act as models for human disease. cDNA is is created from an mRNA template that does not contain introns. Once complete, the sequence can be amplified and submitted to the National Center for Biotechnology Information to identify homologues and their properties. The cDNA sequence taken from *Drosophila* will most likely have a human homology directly linked to a genetic disorder or other notable human proteins. The main research question being investigated is "What potential models for understanding human disease emerge when we search the human genome database with library sequences from *Drosophila melanogaster*?[3] The other research questions being investigated are:

1. What *Drosophila* protein is identified via the sequence of the cDNA and what is the role of this protein in *Drosophila?* What is the known function of any identified protein domains?
2. What human proteins and/or protein domains show homology with cDNA sequences prepared from a *Drosophila* cDNA library?

3. What is the function of each protein in humans (if known)? In particular, does it play a role in human disease? What is the known function of any homologous protein domains?
4. How might studying the control and function of this protein in *Drosophila* contribute to our understanding of mechanisms controlling human disease involving this protein?[3]

Materials and Methods

All procedures for the experiment were carried out as directed in the lab manual.[3] The experiment was carried out over several weeks and was separated into colony picking, plasmid isolation, gel electrophoresis, sequencing, and bioinformatics search.

Initially, two colonies of non-pathogenic *E. Coli* with plasmids for ampicillin resistance, which were grown on agar plates treated with ampicillin, were selected. The ampicillin was added to ensure that only the *E. Coli* was being grown on the plates and to prevent bacterial contamination. The *E. Coli* cells also contained the genes for *Drosophila* cDNA sequences.

After the colonies were incubated for a week, the cDNA was then isolated from the plasmid by rinsing the cell with buffer and using a centrifuge on the solution to facilitate separation of the materials. A negative control tube was also created in order to confirm that there was no DNA contaminating the sample. This step was vitally important to ensure isolation from any human genetic information that could contaminate the *Drosophila* sequences. Any contamination could result in inaccurate readings of the motifs present during the bioinformatics search.

The cDNA sequence was then amplified via Polymerase Chain Reaction (PCR). PCR puts the desired DNA sample through thermal cycling, which breaks the hydrogen bonds and separates the two strands so that primers can bind to and help replicate the DNA which is then cooled and the resulting strands are then copied as well. The PCR mix contained buffer, dNTPs, primers, and Taq polymerase. Taq polymerase is responsible for adding the necessary nucleotides to the separated strands of DNA. PCR was necessary to multiply the desired sequences of DNA so that they could be further analyzed.

The cDNA sequences were then analyzed through gel electrophoresis. The masses of the amplified cDNA sequences were found by running them along a DNA ladder of known masses to yield estimated mass values for the samples according to where they line up on the ladder. The larger DNA fragments tend to travel a shorter distance while the small DNA fragments travel a longer distance. The negatively charged phosphate groups allow the fragments to run from the negative (black) end to the positive (red) end of the gel. The brightness of each lane's material would later be analyzed to determine which lane's DNA would be searched in the bioinformatics database.

During the last week of the investigation, the cDNA sequence was then searched on a bioinformatics database to find homologous human proteins to the corresponding *Drosophila* sequence. The DNA sequence was trimmed before sequences of unidentifiable pairs were seen in the software program MEGA. This was to ensure that any extraneous information was cut out and to prevent it from altering the search results. This sequence was then exported and opened in the NCBI BLAST. The sequence was then compared to other homologous sequences to see which proteins were present in humans as well.

Results

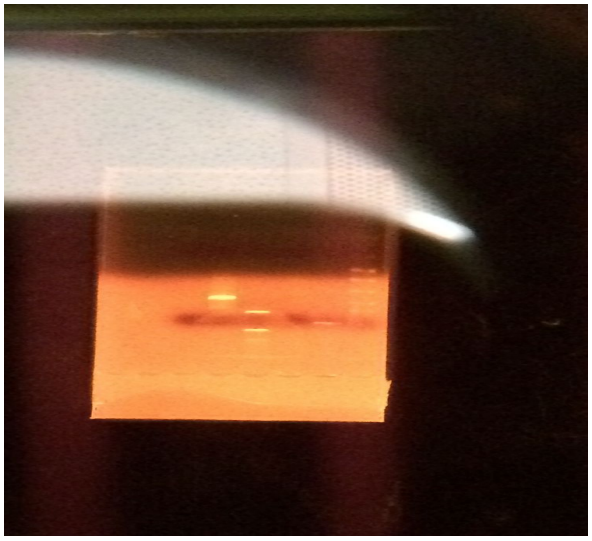

Figure 1. (The picture presented is an image of the agarose gel electrophoresis. This picture was taken from a different group as there was not a clear picture for the original gel, however the rest of the information is relevant to the original material.[4])

Well 1 contained the PCR DNA ladder, followed by plasmid A DNA, plasmid A PCR, plasmid B DNA, plasmid B PCR, and the PCR negative control in the following wells. The gel electrophoresis pulled the negatively charged DNA fragments toward the positive end of the gel. The smaller DNA fragments tended to move further as can be observed in the photograph as they had less resistance to the gel in respect to the larger fragments. After looking at lanes 3 and 5, the PCR clearly worked for both samples A and B as both bands were at the same level. Also, despite what is the seen in the picture above, the band for sample A's PCR products was brighter in the gel, which was a slightly better result than in sample B. Thus, the DNA plasmids from sample A were chosen for sequencing at the Nucleic Acid Facility instead of the plasmids from sample B.

After submitting the sequence to MEGA, the chromatogram was very readable and the messy sections were cut out using the software. The cDNA sequence was then submitted to the NCBI website with the reference ID IAZ255UP01R to determine the identity of the mRNA. The NCBI reference sequence for the mRNA was NM_001274092.1 and was named *Drosophila melanogaster clueless (clue)*, transcript variant B. The identification of this search was 100% accurate.

A protein blast was then performed from the mRNA to find the protein that was coded for for this mRNA. It had protein ID of NP_001261021.1 named *clueless, isoform B* [*Drosophila melanogaster*]. This protein contained the domain CLU_N with a function that is not yet known. It also contained the domain CLU which is needed in the correct function of mitochondria and mitochondrial transport, however its exact function is also unknown. TPR1 and CLU-central were also present but had unknown functions.[5]

A BLAST was then performed to find homologues in humans with the request ID 1AZPZD9E01R. The closest found human protein in terms of percent of identical amino acids to *clueless isoform B* was KIAA0664, isoform CRA, which had a 61% homology rate. Another protein found was called Clustered mitochondria which possessed a 55% homology to the original protein. This information could then be used to determine which protein was most prevalent in humans and *Drosophila,* and that protein's function would then be analyzed to determine any potential links between its function and human diseases.

Discussion

As can be seen in the investigation, there are four protein domains that exist in this specific *Drosophila* sequence and it also contains numerous homologous proteins that exist in

humans as well. There are obvious homologous similarities between *Drosophila* DNA sequences and human sequences, which indicates that there are certainly homologous proteins that code for a human disease. The *Drosophila* DNA can certainly be used to study certain diseases in humans.

The results of this investigation did present a certain level of obscurity regarding the function of the protein domains in the given sequence. Moreover, the name of the protein, *clueless isoform B,* in itself shows that scientists are not yet sure of its exact function. However, each homologous protein plays a role in its respective organisms, such as *clueless isoform B.* The CLU domain is vital in mitochondrial function and mitochondrial transport in both *Drosophila*, which may indicate that it could be used to study mitochondrial diseases such as Pearson Syndrome in humans. It is also believed that *clueless isoform B* may play a role in humans. This disease results in bone marrow and pancreas dysfunction which is caused by a single mitochondrial DNA deletion; people with this disease often do not survive past childbirth.[6] Misregulation of the gene that codes for this protein may lead to the development of Pearson Syndrome, however this cannot be concluded definitively.

It is difficult to determine with absolute certainty if *clueless isoform B* is related to the development of Pearson Syndrome in humans. While it is believed that *clueless isoform B* plays a role in human development of Pearson Syndrome, it is difficult to say for sure. The research and information gathered during this investigation point to this protein potentially playing a role, but there is no indisputable evidence from this investigation alone. After running the protein through the CDART program to determine human protein homology, it was deduced that the closest human protein to *clueless isoform B* was *KIAA0664, isoform CRP [Homo Sapiens],* as

stated prior. However, the homology rate was only 61% and while this is a relatively similar relationship, there is too much of a disparity to conclude that the *Drosophila* gene could be used to show the disease develops in humans. This does however give a lead for this disease to be further investigated due to the potential similarities in both *Drosophila* and humans.

A potential source of error in this experiment may have occurred during the trimming of the trace file in MEGA because if the sequence was trimmed too long or too short than it could have resulted in different matches during the BLAST search, which could have led to incorrect human homologs. As always, there is always the potential for bacterial contamination in the DNA as it is difficult to be sure of if the previous groups kept their pipet tips as sterile as possible. If that group did not, the integrity of this investigation may have been compromised.

The DNA extracted for this investigation coded for a protein in *Drosophila melanogaster* which had homologous protein domains in humans. The homologous domain is involved in mitochondrial function and transport and mutations in this domain may result in mitochondrial diseases such as Pearson Syndrome. If scientists are able to further understand the function of *clueless isoform B*, then it may help scientists understand the function of the human homolog involved in mitochondrial diseases. The discovery of this homologous protein in fruit flies also indicates that there are additional proteins in fruit flies that have human homologs which could potentially play a role in other human diseases. More research must be conducted on *clueless isoform B* in particular to give a clearer idea of the protein's exact function. If this investigation could be replicated and with more intrusive research on this particular domain, perhaps a clearer link could be established. Further sequencing of the fruit fly genome may lead to understanding

additional human diseases and functions due to the apparent protein homology present between the two species.

References

1. Flower. "The Lipocalin Protein Family: Structure and Function."*Biochemical Journal*. U.S. National Library of Medicine, n.d. Web. 14 Oct. 2015.
2. Lee, Soojin. "Evaluation of Traditional Medicines for Neurodegenerative Diseases Using Drosophila Models." *Evaluation of Traditional Medicines for Neurodegenerative Diseases Using Drosophila Models*. Hindawi Publishing Corporation, 2014. Web. 20 Oct. 2015.
3. Penn State Biology Department. “Biology 230 Laboratory Manual.” Lab handbook. The Pennsylvania State University, PA. 2015.
4. Brustle, Kent. “DNA Analysis Lab- Week 1” BIOL 230W. 2015. The Pennsylvania State University, PA.
5. "Basic Local Alignment Search Tool." *National Center for Biotechnology Information*. N.p., n.d. Web. 21 Oct. 2015.
6. Cormier, V. "Pearson's Marrow-pancreas Syndrome. A Multisystem Mitochondrial Disorder in Infancy." *JCI - Pearson's Marrow-pancreas Syndrome. A Multisystem Mitochondrial Disorder in Infancy.* The Journal of Clinical Investigation, n.d. Web. 14 Oct. 2015.