Daniel Alfonsetti

UROP Faculty Supervisor: Troy Littleton, M.D., Ph.D.

UROP Direct Supervisor: Karen Leopold Cunningham

Term: Fall 2018

9 Sep. 2018

Analysis of polyglutamine tracts in synaptic proteins

Project Overview

Chemical synapses are junctions between neurons that facilitate neuronal communication in the brain. At this junction, the chemical signal from the presynaptic neuron is carried across a specialized region in the cell membrane called the active zone. The active zone is composed of a protein scaffold that facilitates the efficient fusion of the signal-carrying vesicles and the subsequent release of their chemical contents into the intracellular space between the neurons (the synaptic cleft). Understanding synaptic function and regulation is crucial to understanding how we can think, learn, and adapt to changing environments. Moreover, pathological alterations in synapse function underlie many important neurological diseases.

A certain set of neurological diseases are caused by protein alterations in which a certain segment of the protein's amino acid sequence, called a polyglutamine tract (polyQ tract), gets elongated. A polyQ tract is a section of a protein that contains several consecutive glutamine monomers. Long polyQ tracts (more than 20 amino acids) are frequent among proteins of healthy cells in the *Drosophilidae* family of flies. In humans, polyQ tracts are the most common homo-amino-acid tract encoded by a single repeated codon (CAG) (Atanesyan, 2012). However, mistakes during DNA replication (such as strand slippage, misalignment, and stalling) can increase the number of glutamine repeats to abnormal levels, resulting in the so called polyglutamine diseases. Such diseases include Huntington's disease and spinocerebellar ataxia. Longer polyQ tracts are generally associated with more severe disease symptoms and earlier disease onset in life (Rinaldi, 2015). The pathogenic polyQ proteins are prone to aggregation, and biophysical studies have shown that inter-protein hydrogen bonding between the glutamines drives this aggregation (Natalello, 2011). There is debate as to whether these aggregations contribute to the pathogenesis or if smaller oligomers are driving the disease (Todd and Lim, 2013*)*.

Although expanded polyQ tracts are well known for their involvement in neurodegenerative disease, shorter polyQ tracts are present in many genomes and play biological roles in healthy cells. For example, prior studies in healthy eukaryotes have shown polyQ tracts to be in proteins involving regulation of gene expression. These regulatory proteins contain glutamine-rich activation domains involved in protein-protein interactions (Atanesyan, 2012).

Despite the established importance of polyQ tracts in both disease contexts and in normal cellular processes, little research has been performed on the function of polyQ tracts outside of neurodegenerative diseases and transcriptional regulation. However, recent preliminary research by Ph.D. student Karen Leopold Cunningham (my supervisor) from the Littleton Lab at MIT has identified polyQ tracts in several core synaptic proteins, including Rab3-interacting molecule (RIM), Rim Binding Protein (RBP) and bruchpilot (BRP). Further exploration of glutamine enrichment in synaptic proteins revealed that polyQ-enriched regions are present in a large number of proteins involved in synapse structure, function, and plasticity. The goal of this project, therefore, is to rigorously determine if synaptic proteins are significantly enriched for polyQ tracts in flies and humans on a genome-wide scale. Because the polyQ tracts promote protein-protein interactions in both proteins associated with polyglutamine diseases and proteins involved in gene regulation, we hypothesize that polyQ tracts serve a similar function in synaptic proteins. More specifically, we hypothesize that the polyQ tracts help to tether proteins together and confer stability to the active zone scaffold.

Personal Role & Responsibilities

While complimentary experimental tests are being performed in parallel by my supervisor, my responsibilities are completely computational in nature. I will be responsible for creating an open source, reusable, scalable pipeline that can perform, but is not limited to, the following analyses using open source data from online databases (such as FlyBase) and potentially also from the Littleton Lab:

1. For each protein in the input genome, run a hidden markov model (HMM) to annotate polyQ and non-polyQ regions along the amino acid sequence. Tune HMM parameters with the Baum-Welch algorithm.
2. Annotate proteins based on which cell types they are expressed in. Determine if polyQ proteins are enriched among expressed proteins for certain cell types, and compare the amount of enrichment between the types (i.e neurons vs cardiomyocytes). Repeat the process after removing nuclear proteins (since many are known to have polyQ enriched regions). Use the nuclear proteins as positive controls.
3. Within the neuronal proteome, annotate whether each protein is associated with the synapse or not. Determine if more synaptic proteins are enriched for polyQ tracts than synaptic proteins that are not enriched. Do the same for non-synaptic, neuronally expressed proteins.
4. For proteins with multiple isoforms, determine whether all isoforms are enriched or whether alternative splicing generates some isoforms with polyQ tracts and some without. When possible, use existing experimental data to determine whether the presence of a polyQ tract correlates with the neuronal/synaptic expression or localization of that isoform.

5. Determine if/how polyQ enriched protein expression levels change with synapse developmental stages.
6. Perform gene set enrichment analysis to identify categories of genes that are enriched for polyQ tracts.

This pipeline analysis will be performed on both the fly and human genomes. Time permitting, the code may be generalized to analyze other sequence patterns (instead of just polyQ regions).

Goals

The overall aim of the project is to determine if glutamine tracts are enriched in synaptic proteins and, if so, to use publicly available data to further elucidate their potential biological roles at synapses. My personal goal for this project is to create an easily scalable and generalizable computational pipeline for protein sequence analysis that others both inside and outside the lab can use and build upon after my UROP.

Personal Statement

My long-term career goal is to integrate computer science, neuroscience, and cognitive science in order develop more sophisticated artificial intelligence. Conversely, I'm interested in applying artificial intelligence and machine learning to computational neuroscience and genomics research. I think this UROP will primarily serve the second goal and will give me more programming and modelling practice in computational biology while simultaneously giving me an opportunity to start learning about neuroscience and the brain.

References

1. Atanesyan L, Günther V, Dichtl B, Georgiev O, Schaffner W. Polyglutamine tracts as modulators of transcriptional activation from yeast to mammals. *Biol Chem*. 2012 Jan;393(1-2):63-70.

2. Rinaldi, C. & Fischbeck, K. H. (2015) Pathological Mechanisms of Polyglutamine Diseases. *Nature Education* 8(4):5.

3. Natalello A, Frana AM, Relini A, Apicella A, Invernizzi G, Casari C, Gliozzi A, Doglia SM, Tortora P, Regonesi ME. A major role for side-chain polyglutamine hydrogen bonding in irreversible ataxin-3 aggregation. *PLoS One*. 2011 Apr 13;6(4):e18789.

4. Todd TW, Lim J. Aggregation Formation in the Polyglutamine Diseases: Protection at a Cost? *Molecules and Cells*. 2013;36(3):185-194.