John Barker – Project Proposal - Binding Database Query Tool

For this assignment, my hope is to design a web tool that will query the UniProt database for a given gene product and list a summary of its known interactions documented in the IntAct database. The user can then select one of these known interacting targets, at which point the system will query UniProt for proteins that are similar to the originally provided protein, then iterate through these proteins to look at progressively more and more dissimilar UniProt entries and check IntAct to see whether they too have a recorded interaction with the given target. The tool will then summarize its results for the user.

The user will begin with a simple search box that allows them to provide a UniProt identifier to get started. This simple HTML form will be the starting point for the search and there will be links back to it for the user to start over at the various steps of the process.

On the back end of the tool, once given a UniProt identifier for a given protein, the tool will look this protein up. If no entry is found, it will return an error, otherwise it will output a summary of the protein – its name, the organism it comes from, and its accession number if available. Given that the protein of interest is now known, the program should attempt to collect UniProt’s list of similar proteins and load their identifying details – name, organism, accession – into memory for later. If there are more than 100 similar proteins in the database I will cap the list at this length, for the sake of memory economy. Having done this, it will query the IntAct database and output interaction targets. All of this information will then be displayed in a formatted page with the summary at the top, a column of similar proteins on the bottom left, and a table of interaction targets on the bottom right, each one having been given a selection button or hyperlink that will allow the user to communicate their selection to the tool.

Upon providing this selection, the program will then review the IntAct database for the given interaction target and check the similar UniProt proteins against it to see whether there is a documented interaction. The tool will iterate through the similar proteins in order of increasing dissimilarity, and if there are three consecutive proteins that have no documented interaction it will stop accessing the database and assume that the rest are too dissimilar and also do not interact – but these assumptions will be displayed differently than the “confirmed no interaction” results. The displayed summary page will be updated so that the selected interaction target is highlighted and the list of similar proteins is color-coded for interaction, where the column rows of interacting similar proteins are colored green, the column rows of proteins confirmed to have no documented interaction are in yellow, and the column rows of proteins assumed to have no documented interaction based on being more dissimilar than a sequence of three non-interacting similar proteins are marked in red.

The intent behind this tool would be to provide a quick visualization of the boundaries of a given protein group where it ceases to have known interaction with a given target. This would provide researchers with two possible insights – either a clear understanding of where an interacting subgroup of a protein family can be defined, or where there are possible yet uncharacterized interactions that could be experimentally confirmed.