Programming Project #1

Aim:

To assess the phosphorylation of proteins involved in glycolysis in response to insulin. This is to compare the results of Poh’s screen to Sean’s phosphor data.

Need to extract all of the genes from the KEGG pathways for

- glycolysis/gluconeogenesis

- fructose/mannose metabolism

- insulin signalling pathway

Need to search Sean’s phosphor data for each gene and return genes for which there is data. There may be several phosphorylation sites for each gene.

For each gene need to know;

* Are they phosphorylated?
* If so how many times?
* Is phosphorylation insulin regulated?

Step 1: Sean’s data

Downloaded phosphorylation data - supplementary table S2 ‘**All Identified Phosphorylation Sites’** from paper (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3690479/?report=classic>)

Prepared the data by putting it in excel with columns as follows:

1. Gene name (abbreviated)
2. IPI for gene
3. sequence window
4. 15 sec /starved (log2) median
5. 30 sec /starved (log2) median
6. 1 min /starved (log2) median
7. 2 min /starved (log2) median
8. 5 min /starved (log2) median
9. 10 min /starved (log2) median
10. 20 min /starved (log2) median
11. 60 min /starved (log2) median
12. LY exps insulin/starved (log2) median
13. MK exps insulin/starved (log2) median
14. TC exp1 15 sec / starved (log2
15. TC exp2 15 sec / starved (log2)
16. TC exp3 15 sec / starved (log2)
17. TC exp1 30 sec / starved (log2
18. TC exp2 30 sec / starved (log2)
19. TC exp3 30 sec / starved (log2)
20. TC exp1 1 min / starved (log2
21. TC exp2 1 min / starved (log2)
22. TC exp3 1 min / starved (log2)
23. TC exp1 2 min / starved (log2
24. TC exp2 2 min / starved (log2)
25. TC exp3 2 min / starved (log2)
26. TC exp1 5 min / starved (log2
27. TC exp2 5 min / starved (log2)
28. TC exp3 5 min / starved (log2)
29. TC exp1 10 min / starved (log2
30. TC exp2 10 min / starved (log2)
31. TC exp3 10 min / starved (log2)
32. TC exp1 20 min / starved (log2
33. TC exp2 20 min / starved (log2)
34. TC exp3 20 min / starved (log2)
35. TC exp1 60 min / starved (log2
36. TC exp2 60 min / starved (log2)
37. TC exp3 60 min / starved (log2)
38. LY exp1 insulin/starved (log2)
39. LY exp2 insulin/starved (log2)
40. LY exp3 insulin/starved (log2)
41. MK exp1 insulin/starved (log2)
42. MK exp2 insulin/starved (log2)
43. MK exp3 insulin/starved (log2)

- Search for all the IPIs in the KEGG list present in Sean’s list. >>print(IPI)

- Of these, report how many phosphor sites (sequence window) are present for each IPI. >>print(sequence windows)

- For LY expts how many entries are there besides NaN for each phosphor site? >>print(LY values not NaN col 37-39), average these, un log them determine std dev. If more than 1.

- For MK expts how many entries are there besides NaN for each phosphor site? >>print(MK values not NaN col 37-39), average these, un log them determine std dev. If more than 1.

- Are there entries besides NaN in the timecourse? If so find the highest median value in the timecourse (column D-K) and tell me which column it is so I know what time point, then find the raw experiment data.

Step 2: KEGG

Extracted gene names for each pathway.

Searched how to do this came up with posts on Biostars (https://www.biostars.org/p/69336/) , which led me to the central ‘API?’ page (<http://www.kegg.jp/kegg/docs/keggapi.html>) where it explains the URLs for certain data structure. From this and the biostar example I entered the ID for glycolysis and gluconeogenesis (00010) in the URL for human gene list (<http://rest.kegg.jp/get/hsa00010>). This list also contains other info not just genes. ID for fructose mannose (00051), insulin signalling pathway (04910).

Copy and pasted lists into spreadsheet. But realise I want to separate the different names/accessions into different cells.

Converting identifiers to UniprotKB\_accession:

Using DAVID to convert Seans IPIs to UniprotKB accession:

* found out that IPI (international protein index) was retired in 2011 and they advise the use of UniprotKB accession numbers instead. This is probably why KEGG has different numbers. Will have to convert IPIs in Sean’s list to uniprotKD accession. There are applications to do this.
* I used the DAVID converter to get a conversion list for uniprot accession numbers <http://david.abcc.ncifcrf.gov/conversion.jsp>. This was saved as ipi\_uniprot\_list.txt
* I replaced the IPIs in Seans list with uniprot accession: program saved as sophie1.

Exporting KEGG pathway as uniprot accession:

* realise I need the mouse ids so that is **mmu00010** for glycolysis pathway rather than hsa for human.
* Found gene list at <http://www.genome.jp/dbget-bin/get_linkdb?-t+genes+path:mmu00010>
* But its KEGG ids. Must convert.
* First of all, needed to clean up the list so there were no odd “ symbols. See sophie\_1
* Using the API converter I can convert 1 gene at a time exx. <http://rest.kegg.jp/conv/uniprot/mmu:100042025>
* The result for this example is 2 uniprot ids. Maybe this is why it doesn’t work for the whole pathway?
* Fed in cleaned up kegg list one at a time and stored output in a list.