READ ME-Data Mining- Genomic location of phosphosites

This folder contains all the code and relevant files to create a function within the software allowing the user to explore and search for phosphosites by genomic location.

Genomic locations of phosphosites.ipynb takes the file containing all information on kinase, substrates and phosphosites (new\_clean\_human\_kinase\_substrates.csv). From this, a new data frame is generated containing only substrate accession numbers (hk\_sub\_Acc.csv). This file was used to obtain the chromosomal locations of the substrate genes, from the Ensembl Biomart search tool, results of this are in Biomart\_sub\_gene\_locations.csv.

Biomart dataframe are merged with substrate accession numbers. From this data frame, substrate accession numbers were iterated over and input into an EBI API to extract all phosphosite coordinates for each substrate and format the information into a data frame. (PS\_genomic\_location.csv)

Genomic locations\_merge.ipynb

Combines relevant data into a single data frame and creates genome browser url for each substrate.

From new\_clean\_human\_kinase\_substrate.csv create new data frame containing relevant information e.g. kinases, substrates, phosphosites and neighbouring sequences.

Substrate gene locations from the Biomart data frame are joined by substrate gene name.

Phosphosite genomic locations are merged by substrate accession number and phosphosite.

To create URLs for each substrate gene, whole gene start and stop genomic coordinates are obtained from Ensembl Biomart, using gene accession numbers as before. The output of this search is saved as sub\_gene\_start\_end.csv

Read in sub\_gene\_start\_end.csv.

Format appropriate UCSC genome browser URL to include specific chromosome and gene start/end genomic coordinates.

Add this as a new column to the Final\_Phosphosite\_genomic\_locations.csv

<https://www.ebi.ac.uk/proteins/api/doc/#!/coordinates/search>

<https://www.ensembl.org/biomart/martview/6ab3e0b6b1bb8e02b80167e14d154936>