Data mining

To retrieve the data, we had four primary sources:

* Uniprot: a database of proteins, containing a variety of functional and structural information
* European Bioinformatics Institute which has also contains a variety of information on many proteins in an easily parsable XML format
* Ensembl which provides genomic information on a number of proteins and their transcripts
* PhosphoSitePlus: a database of post-translational modifications in human, mouse and rat proteins

The datamining was completed via python scripts and the various APIs available on the respective sources. The data was incorporated into a number of csv files, which could then be used to populate our database.

Kinase data

For our database we required a variety of information on human kinases. We retrieved a list of human kinases from the uniport database (<https://www.uniprot.org/docs/pkinfam>) and used this list to decide which proteins we needed to obtain information for. We then use the uniport API (<https://www.uniprot.org/help/api>) as well as the European Bioinformatics Institute protein data bank (<https://www.ebi.ac.uk/pdbe/node/1>). With EBI, we were able to get protein information in an XML format which could then be parse using the BeautifulSoup package in python to retrieve the relevant information. This information included the gene and protein names for each kinase, the domains and positions, the phosphosites on these kinases as well as the protein sequence for each kinase.

Substrate and phosphosite data