2017-2018 FALL

BIN 503

BIOLOGICAL DATABASES AND DATA ANALYSIS TOOLS

FINAL PROJECT

LEUKEMIA

PROJECT MEMBERS:

BAŞAK BAHÇIVANCI

ELİF DOĞAN DAR

BENGİ RUKEN YAVUZ

INTRODUCTION

Leukemia is cancer that starts in the tissue that forms blood. In a person with leukemia, the bone marrow makes abnormal white blood cells. The abnormal cells are leukemia cells. Unlike normal blood cells, leukemia cells don't die when they should. They may crowd out normal white blood cells, red blood cells, and platelets. This makes it hard for normal blood cells to do their work. There are four main types of leukemia which are:

- Acute lymphoblastic leukemia (ALL)
- Acute myelogenous leukemia (AML)
- Chronic lymphocytic leukemia (CLL)
- Chronic myelogenous leukemia (CML)

ALL is the most common type of leukemia in young children. On the other hand, AML is the most common type of acute leukemia in American adults and the average age of a patient with AML is 67. (Acute Myeloid Leukemia, 2014). In PubMed, there is also relatively more publications about AML compared to other types of leukemia, with the number of 23,736 publications. In further analysis in this project, we also use expression data with the type of acute myeloid leukemia.

In addition, there is no standard staging system for leukemia. Generally, cancers are staged with respect to the size and spread of tumors. However, since leukemia already occurs in the developing blood cells within the bone marrow, the stages of leukemia are often characterized by blood cell counts and the accumulation of leukemia cells in other organs, like the liver or spleen. (Leukemia stages, n.d.). Moreover, estimated new cases in 2014 in U.S. is 3.1% of all new cases in the same year. Therefore, compared to other cancers, leukemia is relatively rare. (SEER Stat Fact Sheets: Leukemia, n.d.).

According to findings from OMIM database, by using information from phenotype description case number, #601626, it is obtained that acute myeloid leukemia can be caused by heterozygous mutation in the CEBPA gene. Somatic mutations in several genes have been found in cases of AML, e.g., in the CEBPA, ETV6, JAK2, KRAS2, NRAS, HIPK2, FLT3, TET2, ASXL1, IDH1, CBL, DNMT3A, NPM1, and SF3B1 genes. Findings from OMIM is also supported by findings from COSMIC database, which states following genes for acute myeloid leukemia as the top mutated genes: NPM1, FLT3, DNMT3A, NRAS, TET2, RET, ATM, CEBPA, IDH2, RUNX1, respectively. Moreover, susceptibility to the development of AML may be caused by germline mutations in certain genes, including GATA2, TERC, and TERT. AML may also be part of the phenotypic spectrum of inherited disorders, including platelet disorder with associated myeloid malignancy, caused by mutation in the RUNX1 gene, and telomere-related pulmonary fibrosis and/or bone marrow failure, caused by mutation in the TERT or the TERC gene. Furthermore, for acute lymphoblastic leukemia, susceptibility locus for acute lymphoblastic leukemia (ALL1) has been mapped to chromosome 10q21 and locus ALL2, which has been mapped to chromosome 7p12.2; and ALL3, which is caused by mutation in the PAX5 gene on chromosome 9p. Moreover, for chronic lymphocytic leukemia , by the light of findings with gene description number on OMIM, *616989, gene CLLU1 is exclusively upregulated in chronic lymphocytic leukemia. CLLU1OS is located on the opposite strand from CLLU1, a gene that is exclusively upregulated in chronic lymphocytic leukemia. Finally, for CML, it is most frequently caused

by a translocation between chromosomes 22 and 9, creating a BCR/ABL fusion gene encoding a tyrosine kinase.

According to TCGA database, TCGA researchers have observed relatively few mutations per AML patient. They found that overall, AML genomes have relatively few mutations, and such tumors are among the least mutated adult cancers. An average of 13 mutated genes per tumor were found, in contrast to breast, lung or pancreatic cancer, which often have hundreds of mutated genes. (The Cancer Genome Atlas Network, 2013). Other results were also somewhat surprising. Researchers were aware that mutations in genes which help control cell growth and development, and specifically known as signaling genes, were very common in AML, and thought that all AML samples may have at least one signaling gene mutation. But according to reasearchers, TCGA findings showed that these genes are mutated in only 60 percent of cases. These include mutations in the gene FLT3, which occur in about a third of cases, making it one of the most commonly mutated genes in AML. FLT3 is important for normal blood cell development. The researchers also found that many AML patients have concurrent mutations in three commonly mutated genes: FLT3, NPM1 and DNMT3A. Patients with this combination of gene mutations appear to have a unique subtype of AML. (The Cancer Genome Atlas Network, 2013). Following table is obtained from COSMIC database supports the findings that we obtained from TCGA database, and shows mutated genes and in parenthesis, the frequency of mutated genes among samples.

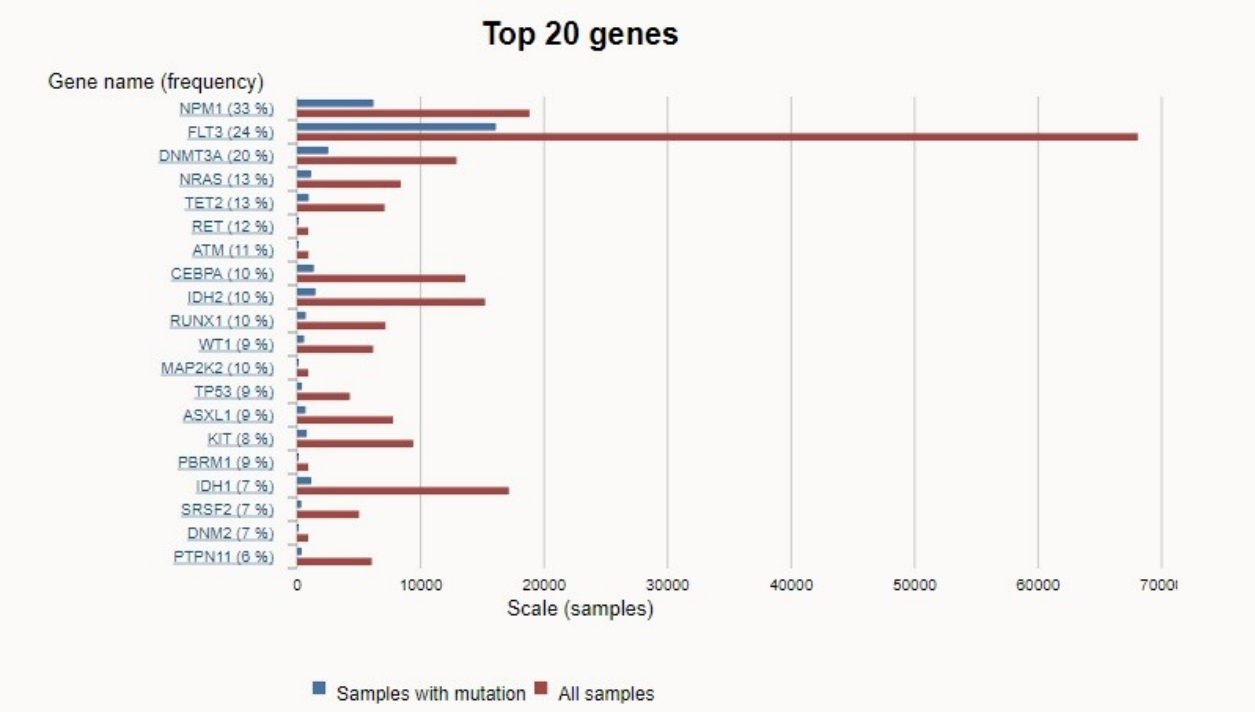
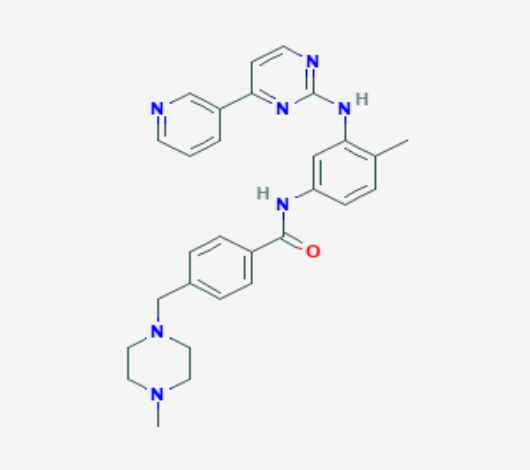


FIGURE 1: COSMIC database table of the top 20 mutated genes by tissue

To be more clear, for NPM1 gene in the Figure 1, there is 16082 samples with NPM1 gene mutated out of 68053 samples which is shown as percentage, in the parenthesis as 33 %.

Drugs

From OMIM database, phenotype description case number #608232 includes information about an approved drug for the treatment of Leukemia called STI-571, which is also called as Imatinib. According to DrugBank database, the Imatinib is used for the treatment of Philadelphia chromosome positive chronic myeloid leukemia (Ph+ CML), Ph+ acute lymphoblastic leukaemia, myelodysplastic/myeloproliferative diseases, aggressive systemic mastocytosis, hypereosinophilic syndrome and/or chronic eosinophilic leukemia (CEL),



dermatofibrosarcoma protuberans, and malignant gastrointestinal stromal tumors (GIST).

FIGURE 2: PubChem 2D Structure of STI-571 (Imatinib)

According to Druker et al. (2001), since tyrosine kinase activity is essential to the transforming function of BCR-ABL, reasearchers reasoned that an inhibitor of the kinase may be an effective treatment for CML. They found that indeed a tyrosine kinase inhibitor STI-571 was well tolerated and had significant antileukemic activity in patients with CML in whom treatment with standard chemotherapy had failed.

Another approved drug is that Arsenic Trioxide. According to information on OMIM database, Arsenic is an ancient drug used in traditional Chinese medicine which has attracted worldwide interest because it shows substantial anticancer activity in patients with acute promyelocytic leukemia (APL).

FIGURE 3: DrugBank Chemical Structure of Arsenic Trioxide.

Zhang et al. (2010) concluded that the identification of PML as a direct target of arsenic trioxide provides insights into the drug's mechanism of action and its specificity for APL.

Likewise, in The ChEMBL database it is stated that Arsenic trioxide (ATO) is an effective cancer therapeutic drug for acute promyelocytic leukemia and has potential anticancer activity against a wide range of solid tumors. However, according to description of the drug on DrugBank, Arsenic trioxide is a chemotheraputic agent of idiopathic function used to treat leukemia that is unresponsive to first line agents. It is suspected that arsenic trisulfide induces cancer cells to undergo apoptosis. DrugBank points out that due to the toxic nature of arsenic, this drug carries significant health risks. Therefore, one can conclude that Arsenic Trioxide might be a debatable drug in terms of pros and cons.

Other popular drugs are found for leukemia by literature review, i.e., Prednisone for use in patients with chronic lymphocytic leukemia and Sprycel (Dasatinib) kinase inhibitor approved for use in patients with chronic myelogenous leukemia. (National Cancer Institute, 2017)

ER DIAGRAM

In the ER diagram, we have shown the entities as rectangles, namely GENES, VARIATIONS, PROTEINS, FUNCTION, BIOACTIVITY AND EXPRESSION. Attributes were tied to the corresponding entities as ellipses. Key attributes were shown as underlined and bold, e.g. uniprot_id in BIOACTIVITY. Also, there are some entities who doesn't have a unique attribute, in that cases, we chose combined primary keys. For example, for PROTEINS we have chosen uniprot_id and pdb_id as combined primary key. There are also multivalued attributes which are shown as double borders around the ellipses, for example, in the FUNCTION table for each uniprot_id there are multiple inter_pro entries which are seperated in a cell. Relations were shown as diamonds, for example, GENES "HAS" "VARIATIONS". Double line between the entity and the relation means it totally participated, e.g., for each VARIATIONS entry there is an entry in the GENES table but not every GENES entry has VARIATIONS entry. Also, there is "N" and "1" written on the sides of the diamond which tells that there is N to 1 relationship between them, for every gene there are many variations entries. However, between PROTEINS and FUNCTION there is a 1 to 1 relationship, for every proteins entry there is only one corresponding entry in the function table. Foreign keys discussion can be found in data retrieval phpmyadmin part.

DATA RETRIEVAL

We gathered data related to "leukemia" from ENSEMBL and GEO databases. First by using Ensembl/Biomart tool we constructed GENES table. We chose Ensembl Genes 91 and Human genes(GRCh38.p10) dataset. As filter, we only used "leukemia", we chose all phenotypes which includes leukemia in them and we choose Gene stable ID, HGNC symbol and gene start(bp) as attributes. There are 12 rows in that table. Similarly we exported PROTEINS table, attributes are Gene stable ID, UniProtKB/Swiss-Prot ID and PDB ID, it resulted in 164 rows. Lastly, we constructed VARIATIONS table by using additional Human Somatic Structural Variants (GRCh38.p10) dataset, also we used an additional filter dbSNP as variant source. Attributes are Gene stable ID, variant name and variant start in translation (aa), this table has 20 entries. We exported all these tables as txt files.

We used UniProt Retrieve/ID maping to generate FUNCTION and BIOACTIVITY tables. We retrieved uniport_id list in PROTEINS table by using query "SELECT DISTINCT uniprot_id FROM proteins". For the BIOACTIVITY table, we provided these Uniprot ID list as identifiers and "from UniProtKB AC/ID to UniProt KB" as options. In the resulting page, we

chose Uniprot ID, Chembl and Drugbank as columns. Then we downloaded this table as text file. It has 8 rows. Similarly, we constructed the FUNCTION table with 12 rows. Attributes for this table are Uniprot ID, InterPro and Gene ontology IDs.

Finally, we retrieved EXPRESSION table by using GEO database and GEO2R. We made an advanced search by using the keywords “leukemia” and ”healthy” in GEO datasets. There were 11 datasets in the resulting page, among which we chose “Acute myeloid leukemia” data set with GSE9476 reference series. In this dataset we have 26 acute myeloid leukemia (AML) patients with normal hematopoietic cells at a variety of different stages of maturation from 38 healthy donors. Then we entered reference number to GE02R tool, and defined healthy and leukemia groups. The we saved all results in EXPRESSION table as text file. There are 22283 entries in this table.

DATABASE CONSTRUCTION WITH PHPMYADMIN

In the next step, we constructed our database by using phpmyadmin. We created a database and 6 tables in it. GENES table has columns ensemble_gene_id, hgnc_sym and chr_pos_start with data types varchar(15), varchar(10) and varchar(4). Ensembl_gene_id is the primary key for this table. PROTEINS table has columns ensemble_gene_id, uniport_id and pdb_id with corresponding data types varchar(15), varchar(10) and varchar(4). Uniprot_id and pdb_id are the combined primary keys for this table. VARIATIONS table has columns ensemble_gene_id, variant_name and var_start with data types varchar(15), varchar(11) and int(3). Variation_name and var_start are the combined primary keys for VARIATIONS table. In the FUNCTION table we have uniprot_id, go_id and inter_pro columns with data types varchar(10), varchar(200) and varchar(150), here primary key is uniport_id. BIOACTIVITY table has attributes uniprot_id, chembl and drugbank with data types, varchar(10), varchar(200) and varchar(150), here primary key is uniport_id. The reason behind that long data type is that there are multiple chembl and drugbank entries for each protein. Finally, we have EXPRESSION table with attributes and data types gene_symbol(varchar(25)), gene_title(varchar(80)), id(varchar(15)), p_value(decimal(6,5)), adj_p_value(decimal(6,5)), t(decimal(6,5)), b(decimal(6,5)) and log_fc(decimal(6,5)). In this table id, adj_p_value and gene_symbol are combined primary keys. Id and adj_p_value combination was also unique but we needed gene_symbol to be a primary key for foreign key constraints. Then we tied these tables to each other as follows: ensemble_gene_id in PROTEINS table is a foreign key for ensemble_gene_id in GENES table. Ensembl_gene_id in the VARIATIONS table is a foreign key for ensemble_gene_id in the GENES table. Hgnc_sym in the GENES table is a foreign key for gene_symbol in the EXPRESSION table. Uniprot_id in the FUNCTION table is a foreign key for uniprot_id in the PROTEINS table and uniprot_id in the BIOACTIVITY table is a foreign key for uniprot_id in the PROTEINS table.

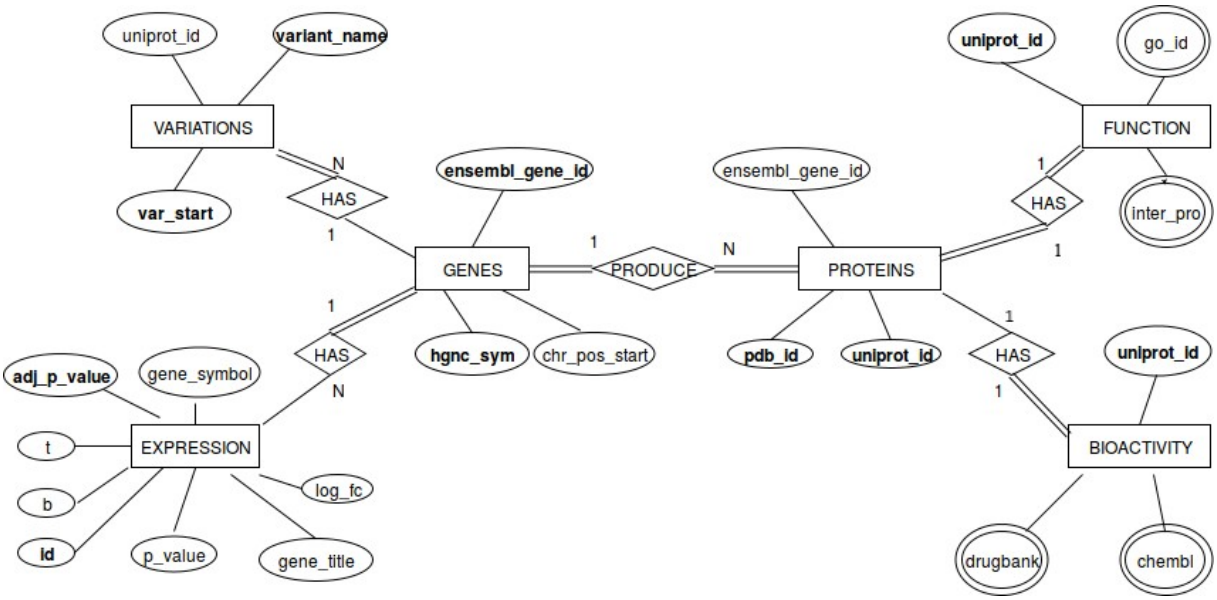


FIGURE 4: ER Diagram

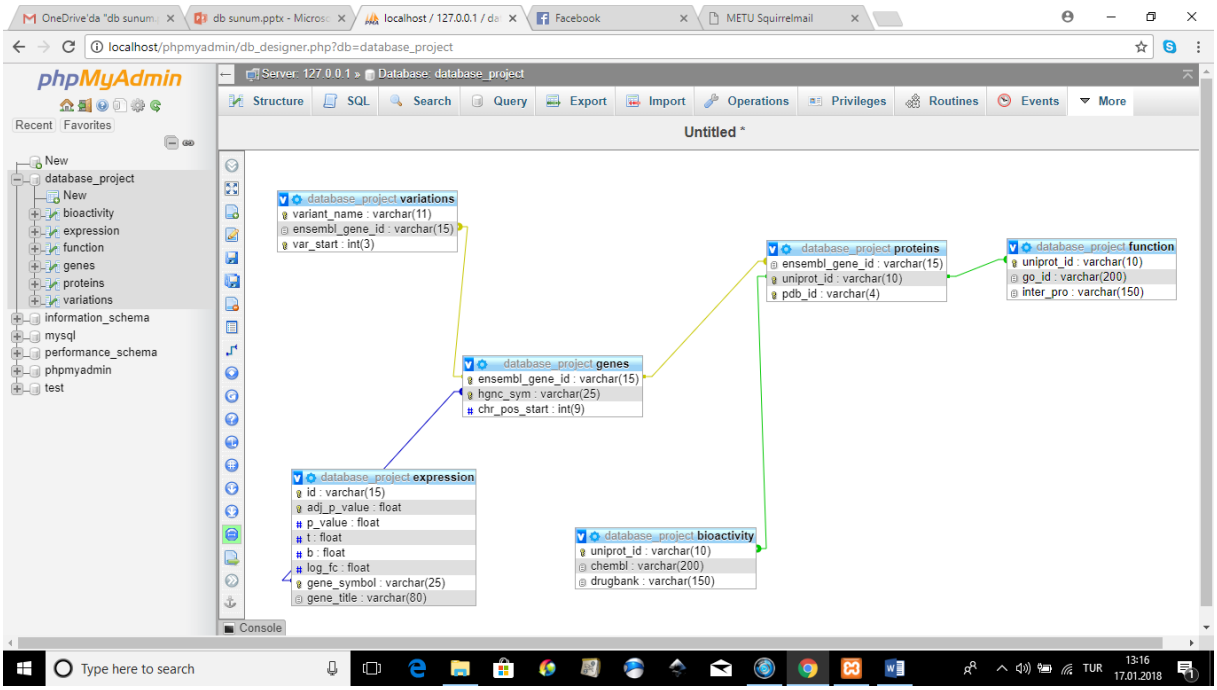


FIGURE 5: Diagram in the phpmyaadmin for the database

WEB PAGE

We prepared a web page which is connected to the database named “proje” created via phpmyadmin. We prepared the interface by using HTML. The interface requires “Gene Symbol” from the user and makes a search in “proje” according to this input. After preparing interface, we connected this interface to the database “poje” by using PHP. We inserted the following SQL query into the PHP code which makes a search using the given gene symbol

starting from the Genes table and then connects the other tables in the database by using this information.

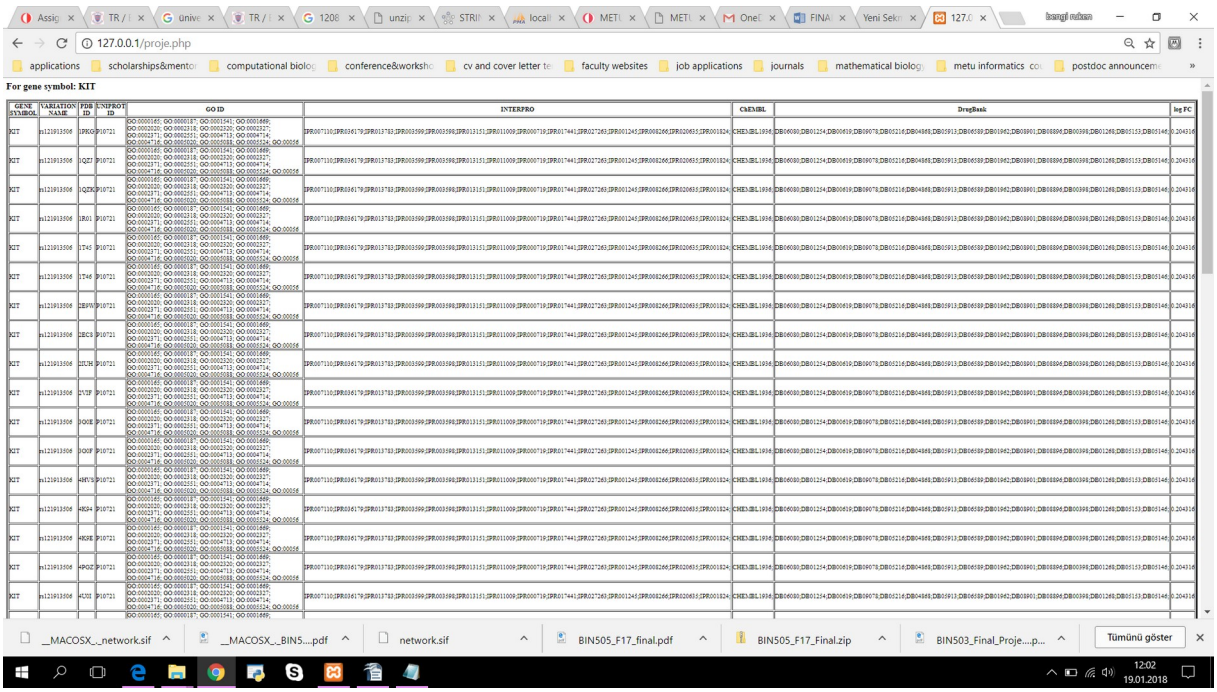
SELECT DISTINCT G.hgnc_sym, V.variant_name, P.pdb_id, P.uniprot_id, F.go_id, F.inter_pro, B.chembl, B.drugbank, E.log_fc

FROM genes AS G, variations AS V, proteins AS P, function AS F, bioactivity AS B, expression AS E

WHERE G.hgnc_sym= ? AND G.ensembl_gene_id=P.ensembl_gene_id AND G.ensembl_gene_id=V.ensembl_gene_id AND P.uniprot_id=F.uniprot_id AND B.uniprot_id=P.uniprot_id AND G.hgnc_sym=E.gene_symbol

PHP and HTML codes can be found in the attached files.

For “Gene Symbol=KIT”, “Gene Symbol=KRAS”, “Gene Symbol=FLT3”, we obtain the results from the database. For the rest of the genes, there is no variation information. So that when we enter a gene symbol different than “KIT”,”KRAS” or “FLT3”, our query works properly but we see empty columns on the interface. You can see the screenshots below.



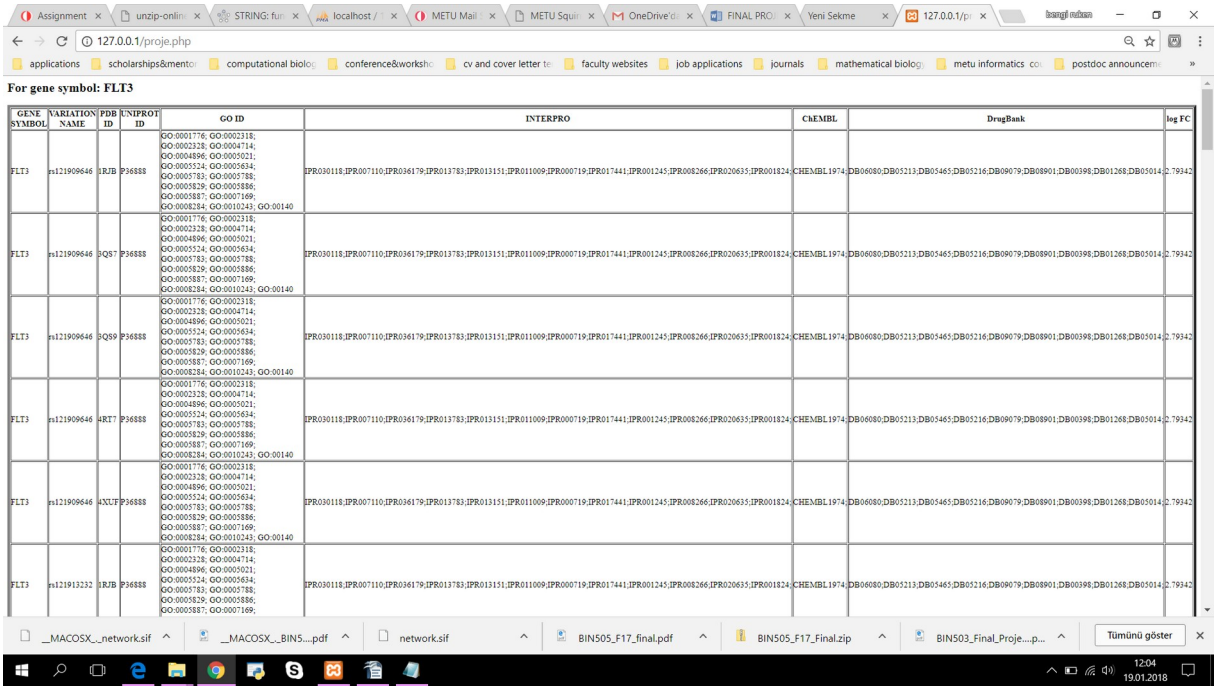
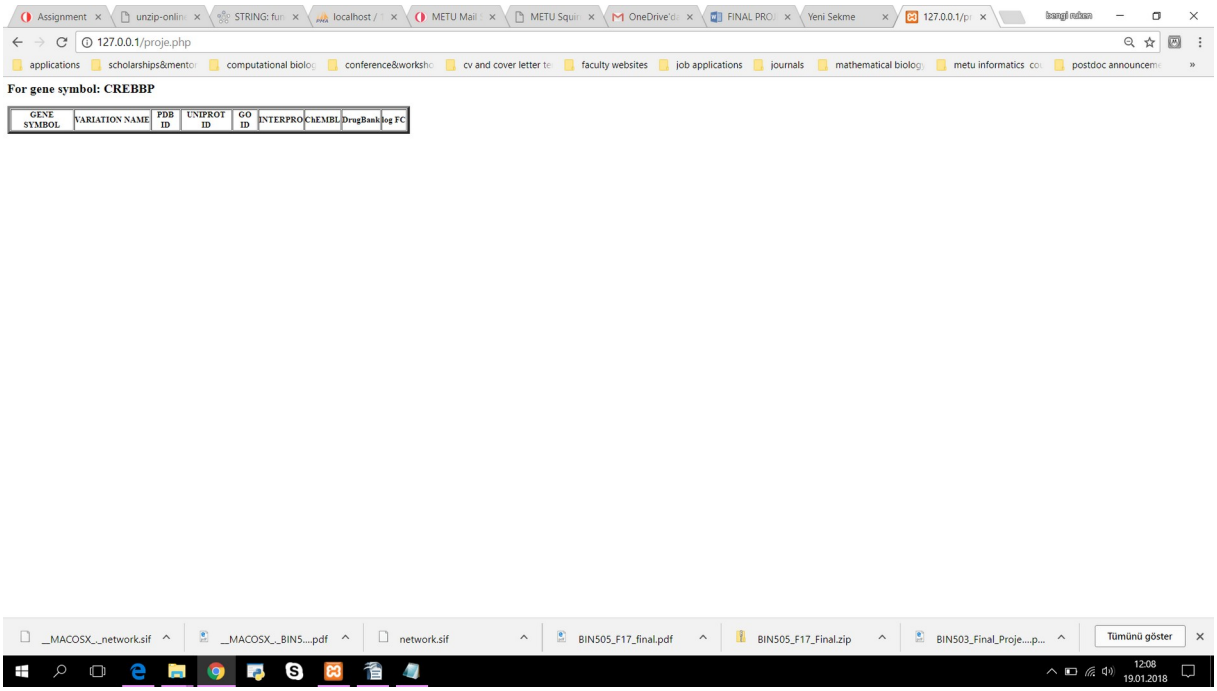


FIGURE 7: Results for FLT3 gene in the database



DISCUSSION

For getting the list of genes that we are going to use in DAVID, we used the following SQL query(Figure 6):

```
SELECT DISTINCT G.hgnc_sym

FROM genes AS G, proteins AS P, bioactivity as B

WHERE G.ensembl_gene_id= P.ensembl_gene_id AND

        B.uniprot_id=P.uniprot_id AND
```

P.pdb_id IS NOT NULL AND

(B.chembl IS NOT NULL OR B.drugbank IS NOT NULL) ;

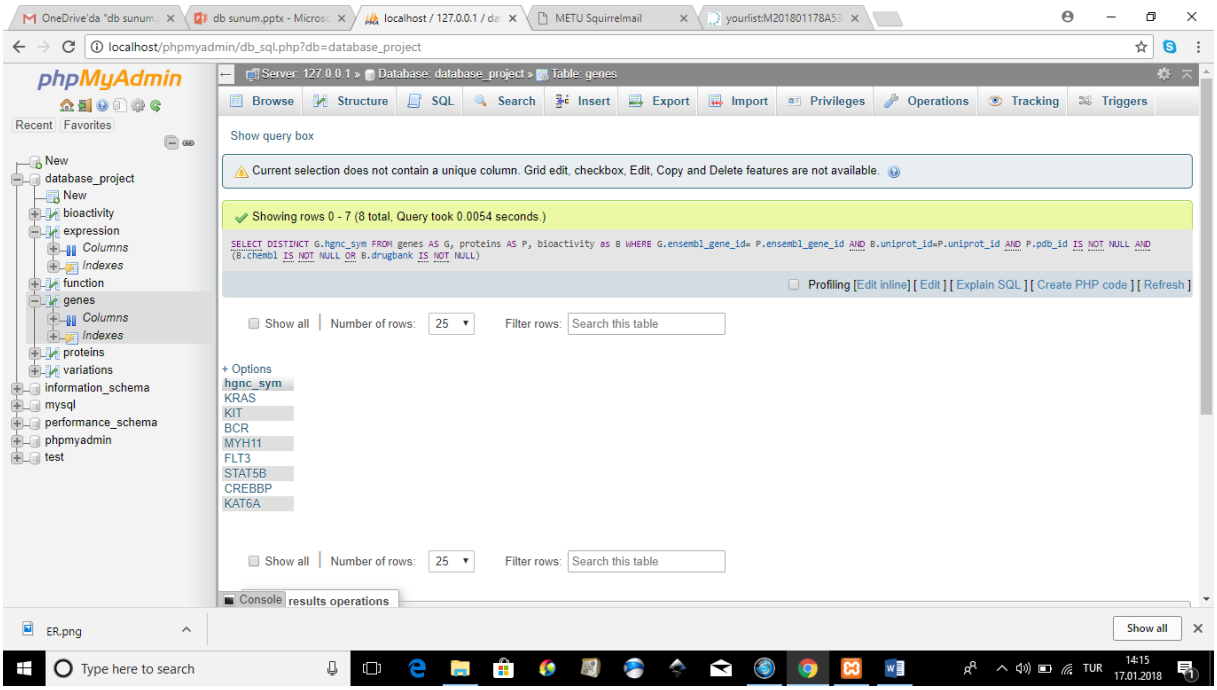


FIGURE 6: Result of the query in the database

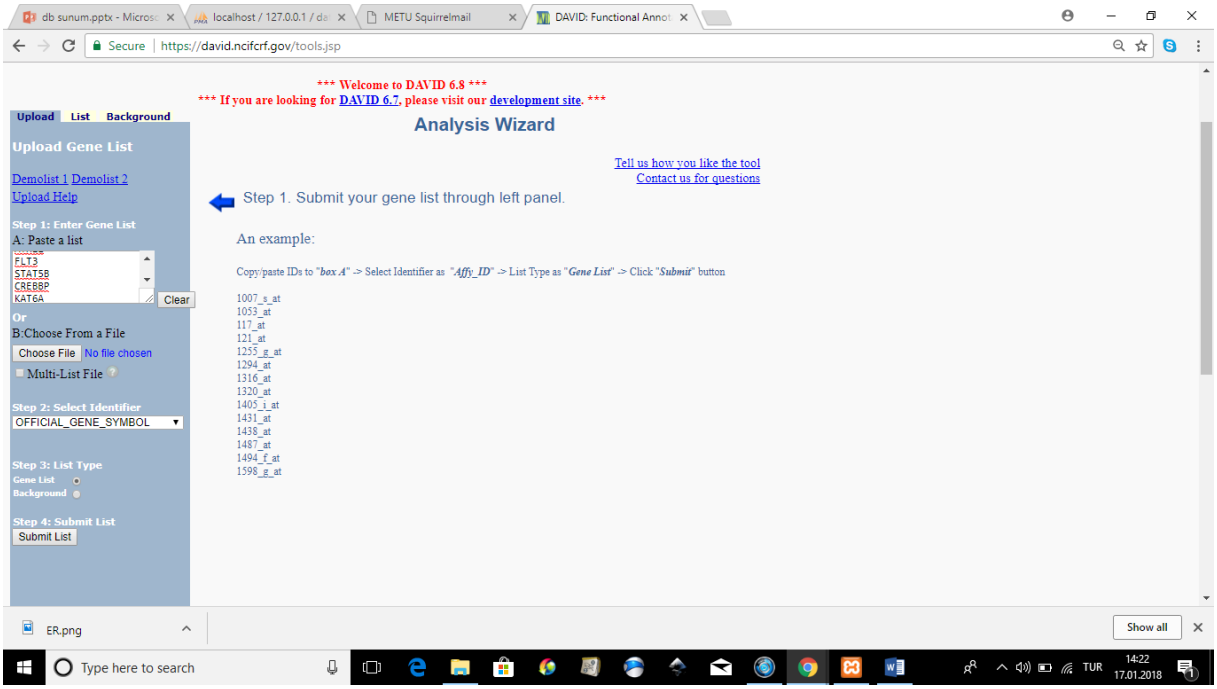


FIGURE 7: Uploading page of DAVID

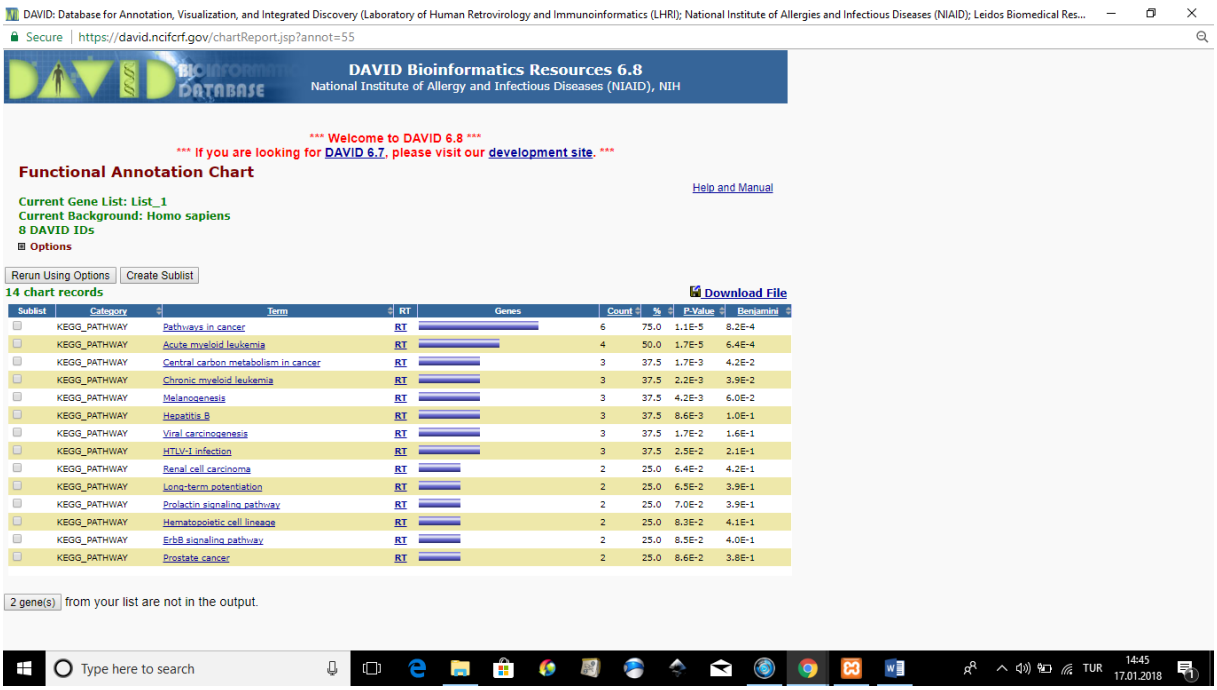


FIGURE 8: Functional Annotation chart of the enrichment analysis in DAVID

There are 8 genes in the resulting list. We go to DAVID's web page. Upload this list by choosing "official_gene_symbol" as identifier. Then we choose "Homo sapiens" in the list and background tabs (Figure 7). Then we choose "functional annotation clustering" (Figure 8). We have our enrichment analysis in the resulting page. We are specifically interested in KEGG pathways and GO functions. For pathways information, we click on pathways, and KEGG pathways chart. It opens a list of pathways with the number of genes participated in and p values in a new window. Uppermost pathway in the list is with the highest significance, and significance decreases as we go down. Pathway names in the list have direct link to the corresponding KEGG page. When we click on the pathway link we see the whole pathway with genes from our list having blinking red star next to them. The pathway with the highest significance is "Pathways in cancer" (Figure 9) with 6 genes participated from our list. CREBBP gene is on the path which results in sustained angiogenesis, which is the sustained development of new blood vessels. BCR and STAT5B participates in evading apoptosis, evading the death of cells. KIT, FLT3 and KRAS participates in proliferation which is the rapid increase in cell number. Second pathway is related to many leukemia subtypes: "Acute myeloid leukemia" (Figure 10) with 4 genes from our list. KIT and KRAS gene participates in the path which activates the proliferative genes. FLT3 and STAT5B helps signaling for antiapoptosis. We can also see the tumor suppressors RUNX1 and CEBPA in the pathway. Third pathway is the "Central carbon metabolism in cancer" (Figure 11) with 3 genes. This pathway shows how the carbon metabolism works in cancer and 3 of our genes have a role in this pathway. Fourth pathway is the "Chronic myeloid leukemia" with 3 genes. STAT5B and BCR are on the path to survival and BCR with KRAS are on the path to proliferation again. Not surprisingly, all these pathways are cancer related pathways.

To get information about function we click on "gene ontology" tab. Here we are interested in biological processes(GOTERM_BP) to see what processes they participate in. The term with the highest significance is the "immune system process" (GO:0002376)(Figure 12), when we click on it, it gives a direct link to the corresponding Quick GO page. 7 genes from our list

have this go term which says they are part of the development or functioning the immune system. When there is a problem with these genes it causes problems in the bodys defence system. Second go term with the highest significance is “developmental process” (GO:0032502) which is a very generic term, doesn’t give much specific information. Other terms in GOTERM_BP_1 list are also similar generic terms. Then I also checked GOTERM_BP_2 list. These lists are ordered with a increasing specificity. Therefore, most significant terms in each list are important. The most significant go term in the second list is “leukocyte migration” (GO:0050900), which says 4 genes in our list plays a role in the change of the location of the white blood cells, which are responsible from fighting with foreign substances and disease in blood, which is obviously related to leukemia.

We can retrieve drugs and chembl entries related to the genes in our database with the following SQL query (Figure 13):

```
SELECT DISTINCT G.hgnc_sym, P.uniprot_id, B.drugbank, B.chembl

FROM genes AS G, proteins AS P, bioactivity as B

WHERE G.ensembl_gene_id= P.ensembl_gene_id AND

    B.uniprot_id=P.uniprot_id AND

    P.pdb_id IS NOT NULL AND

    (B.chembl IS NOT NULL OR B.drugbank IS NOT NULL)
```

For example, there is a drug related to KRAS gene with drugbank entry DB07780. This is an experimental drug, still not available on the market. For KIT gene there are lots of drug entries, some of them are experimental, some are investigational and some of them are approved. For example, Imatinib with drugbank entry DB00619 (Figure 14) is an approved drug related to this gene. It works as an inhibitor of some specific enzyme which results in less increase in cancer cells. Although there are many drugs related to these genes, there are also genes in our list which doesn’t have any drug in drugbank database related to them. CREBBP gene is one of them, but it has a chembl entry, CHEMBL5747 (Figure 15), which says that there are compounds that can be drug candidates for the future investigation.

Sometimes it can be difficult to find compounds with the targets we specifically want . There can be compounds who interact with our target which has many side effects or there are other compounds whom it can interact with. In those cases it might be a good idea to use the upstream genes as targets because if we can inhibit them, we will be breaking the path coming to the gene responsible from the disease. For example, lets take KRAS gene in that pathway, it has only one drug in the drugbank which is not approved yet. If we target PTPN11 gene which comes right before KRAS in the pathway, we can investigate for future drugs. There is no drug entry for this gene yet, but there is chembl entry which is CHEMBL3864.

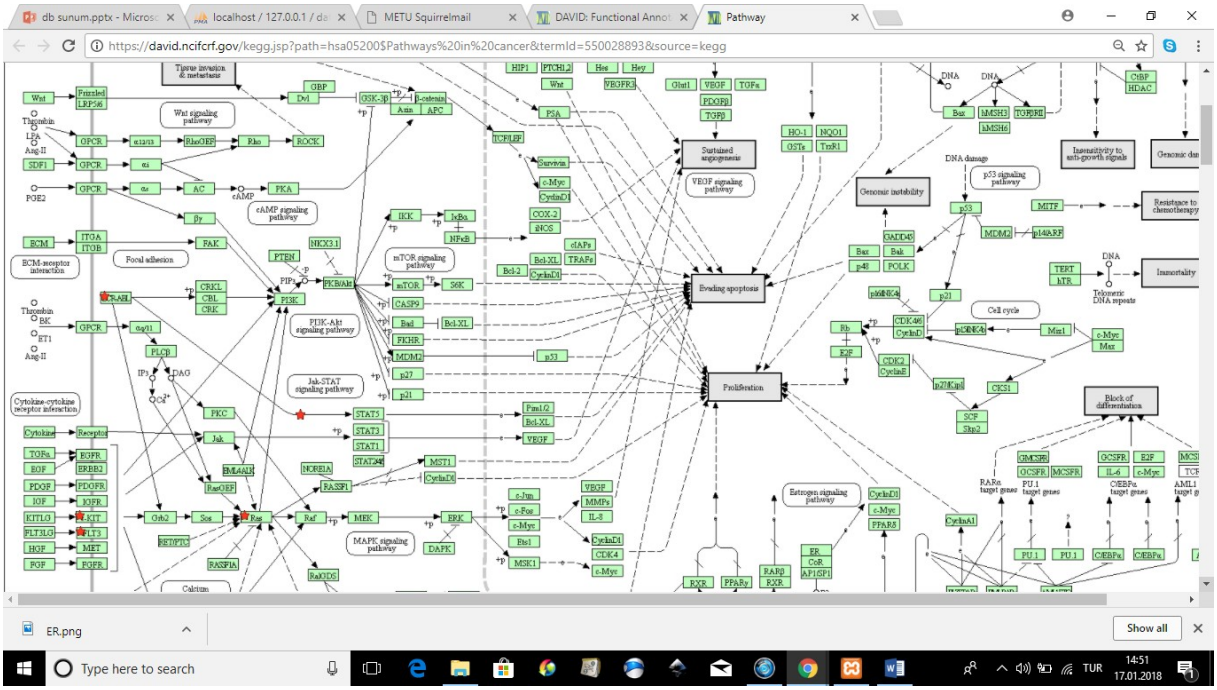


FIGURE 9: KEGG Pathway: Pathways in cancer

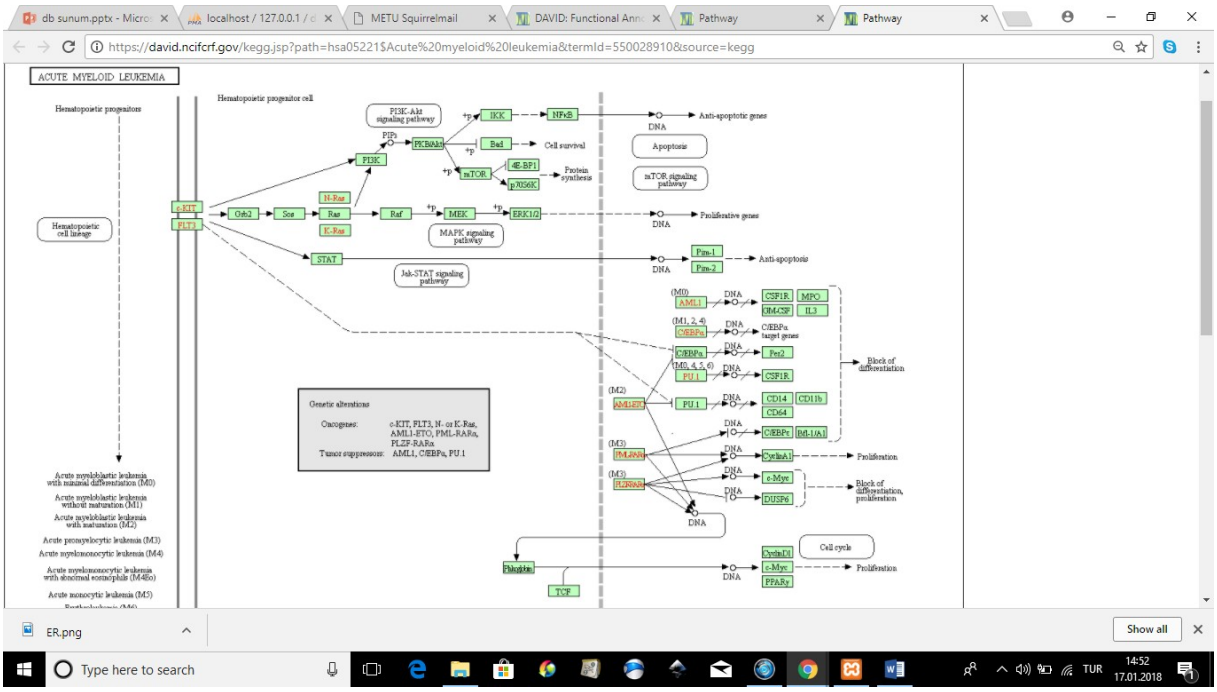


FIGURE 10: KEGG Pathway: Acute myloid leukemia

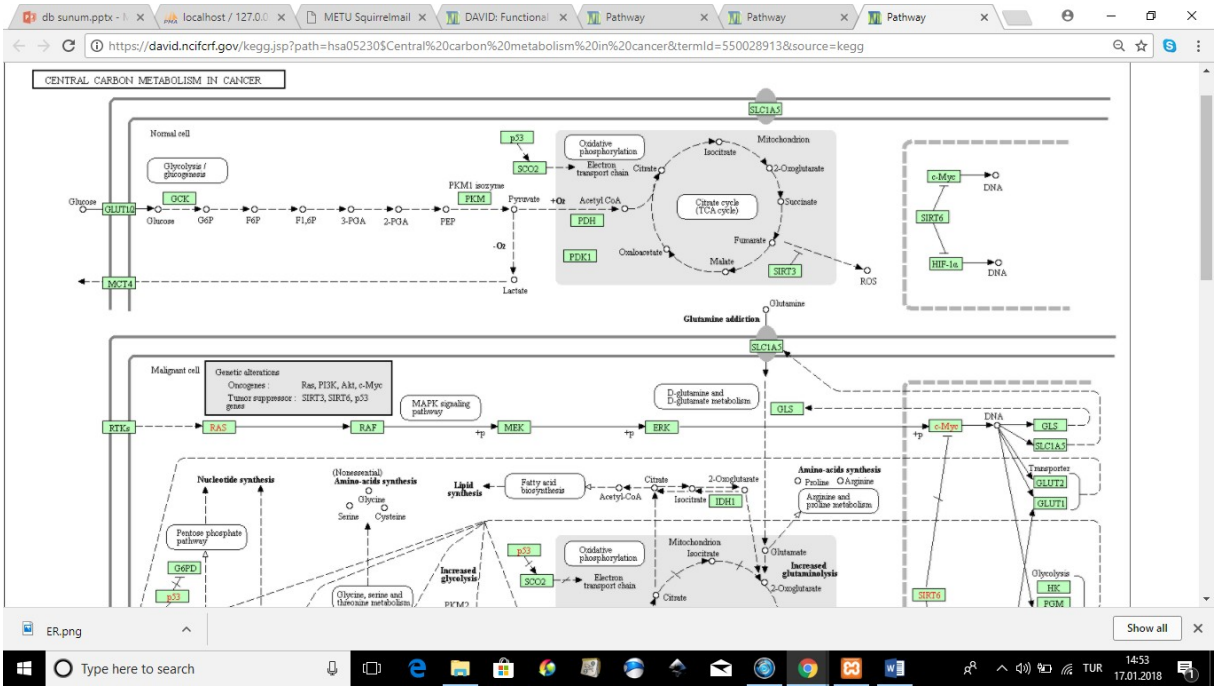


FIGURE 11: KEGG Pathway: Central carbon methabolism in cancer

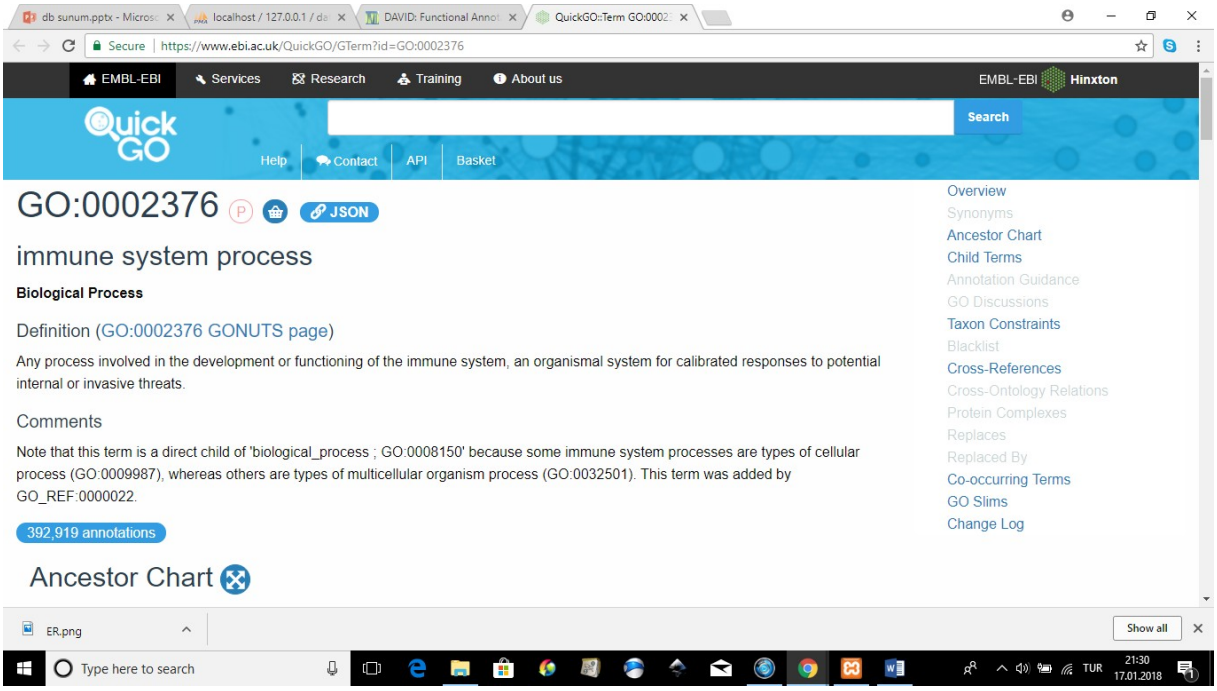


FIGURE 12: Quick GO entry for “immune ststem process”

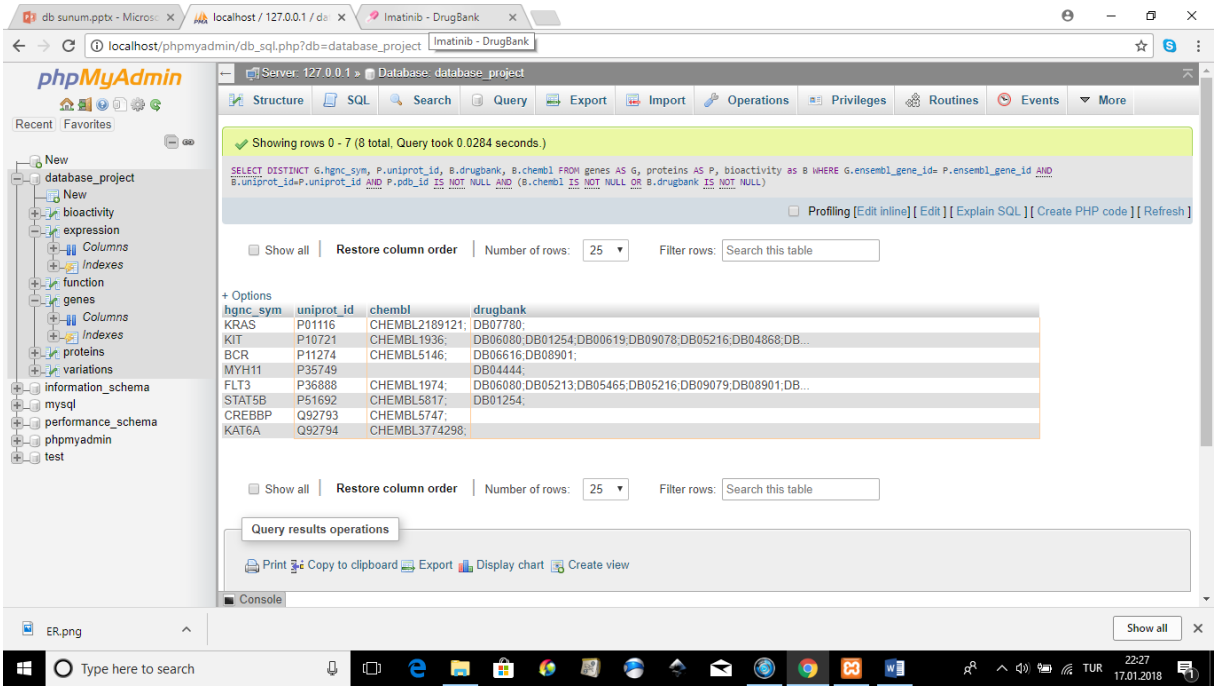


FIGURE 13: Drugbank and Chembl entries fort he genes in our list

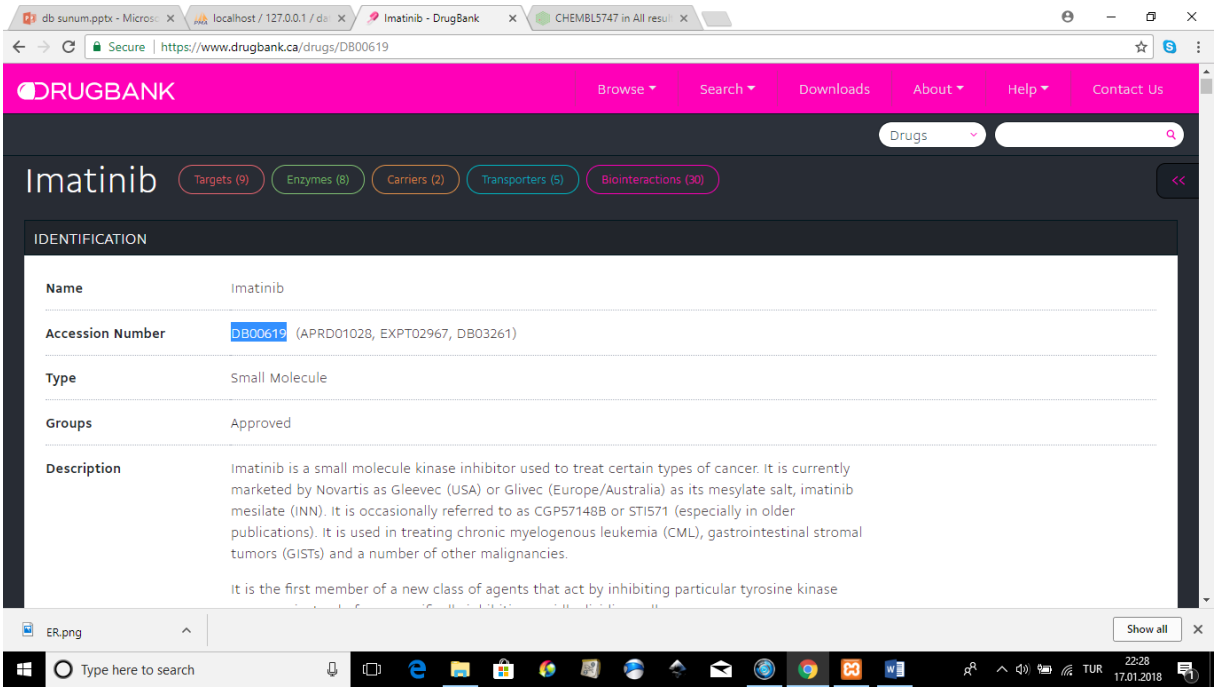


FIGURE 14: Drugbank entry for Imatinib

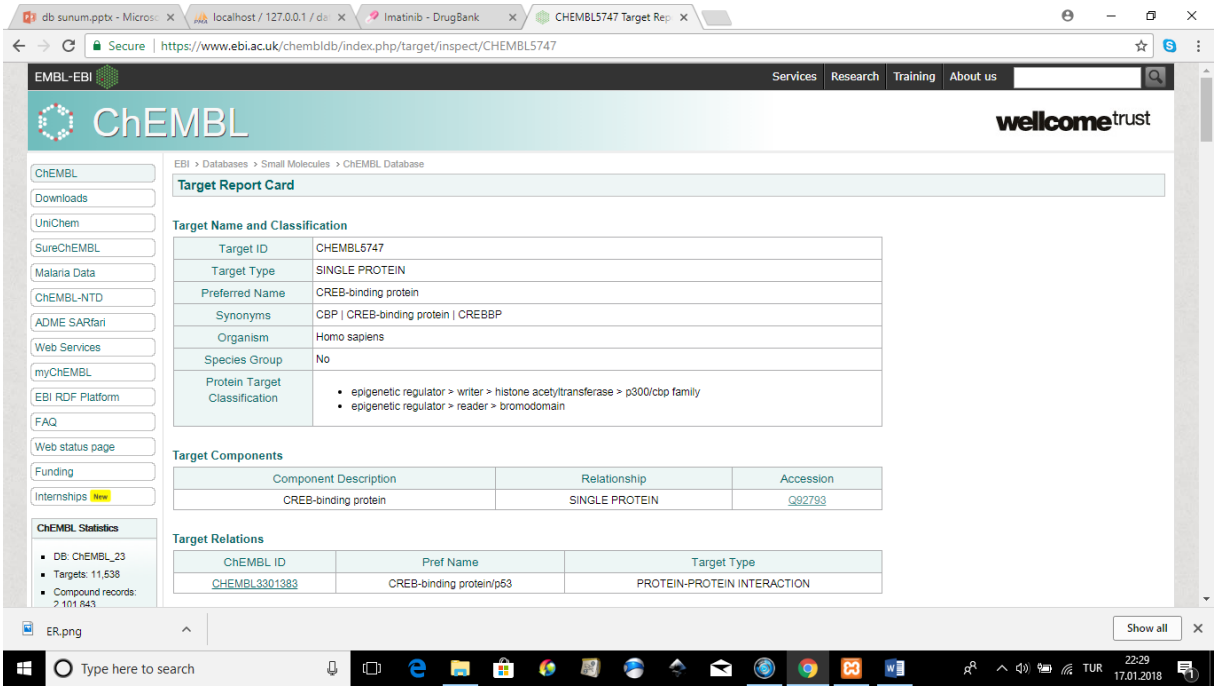


FIGURE 15: Chembl entry

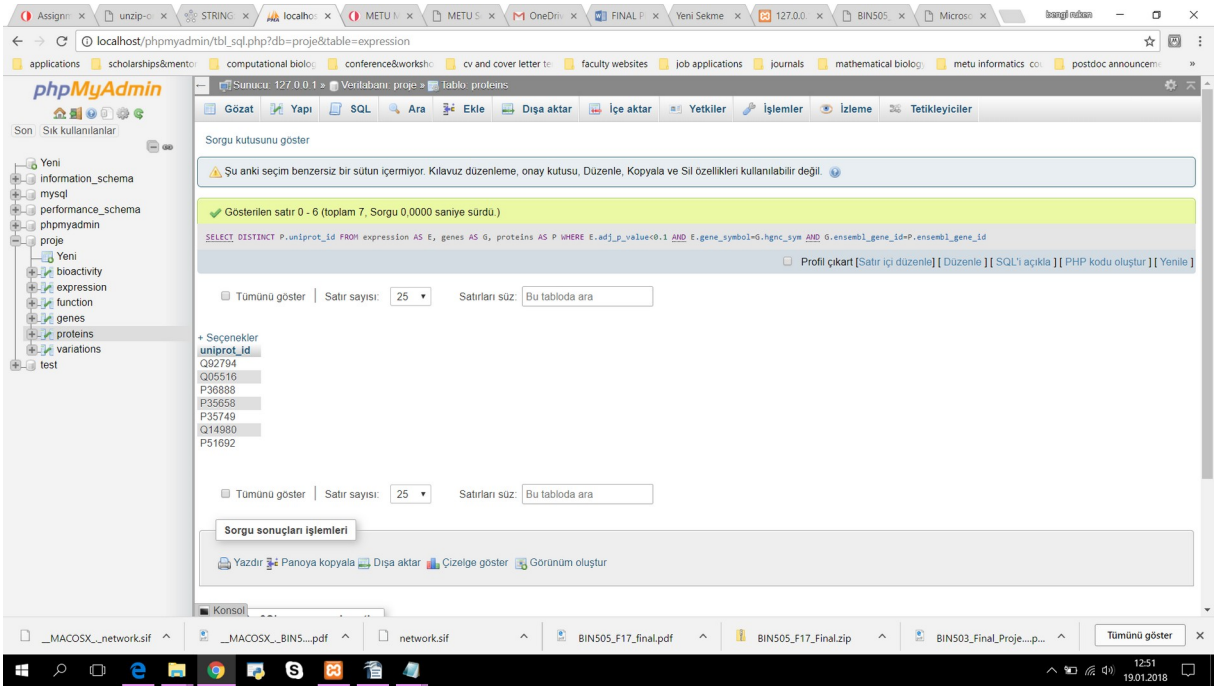
The following part contains information about the genes in STRING database satisfying the following properties:

a)Write down a query to count a list of genes whose adjusted p-value in GEO2R is smaller than 0.3.

```
SELECT COUNT(DISTINCT E.gene_symbol)

FROM expression AS E

WHERE E.adj_p_value<0.3
```

b)Write down a query to retrieve a list of Uniprot IDs of genes whose adjusted pvalue in GEO2R is smaller than 0.1 and those having a UniProt ID.

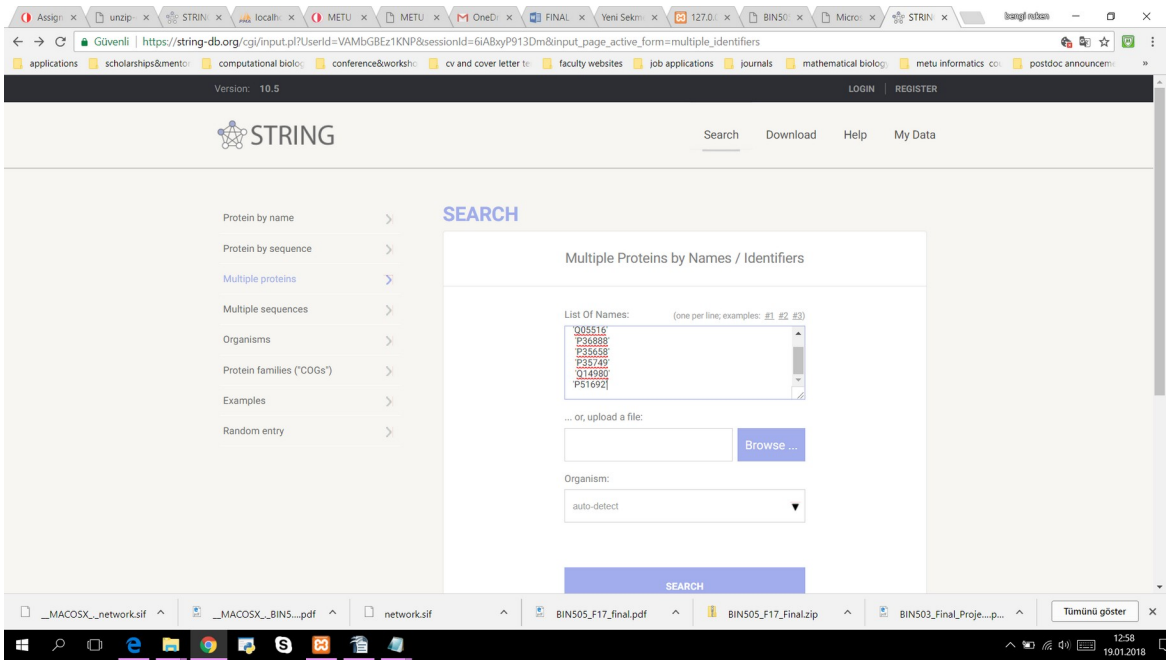
```
SELECT DISTINCT P.uniprot_id

FROM expression AS E, genes AS G, proteins AS P

WHERE E.adj_p_value<0.1 AND E.gene_symbol=G.hgnc_sym AND
G.ensembl_gene_id=P.ensembl_gene_id
```

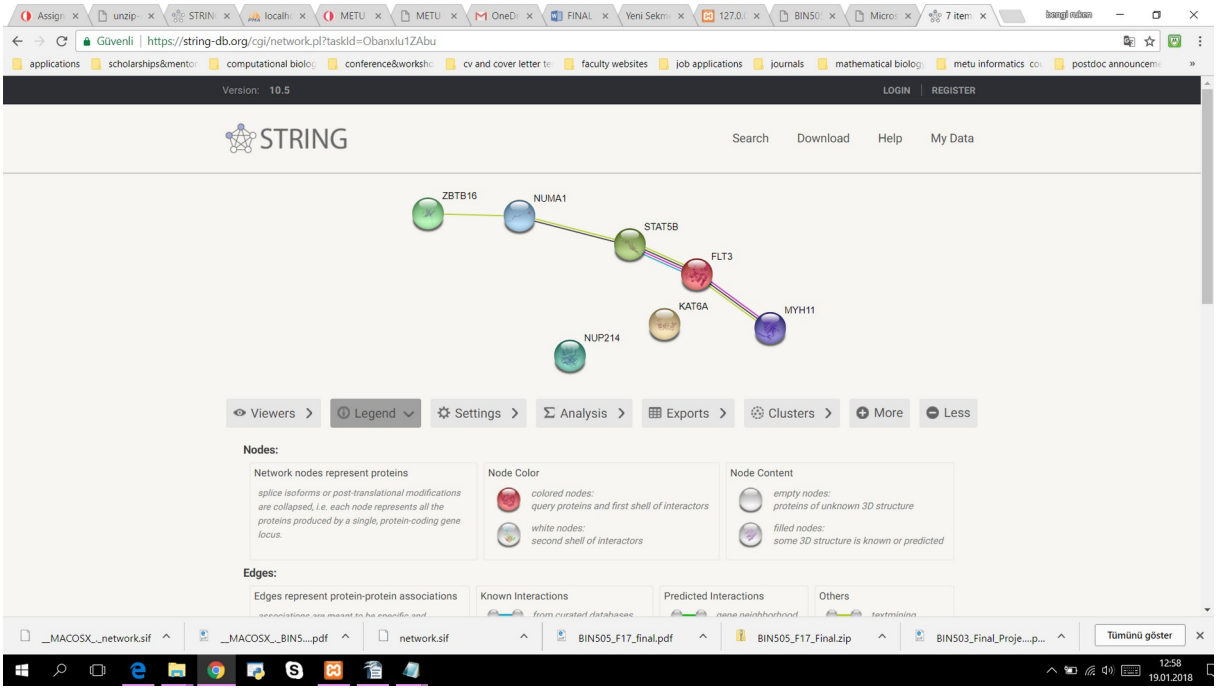


c) In part b the query returned the following Uniprot Ids:

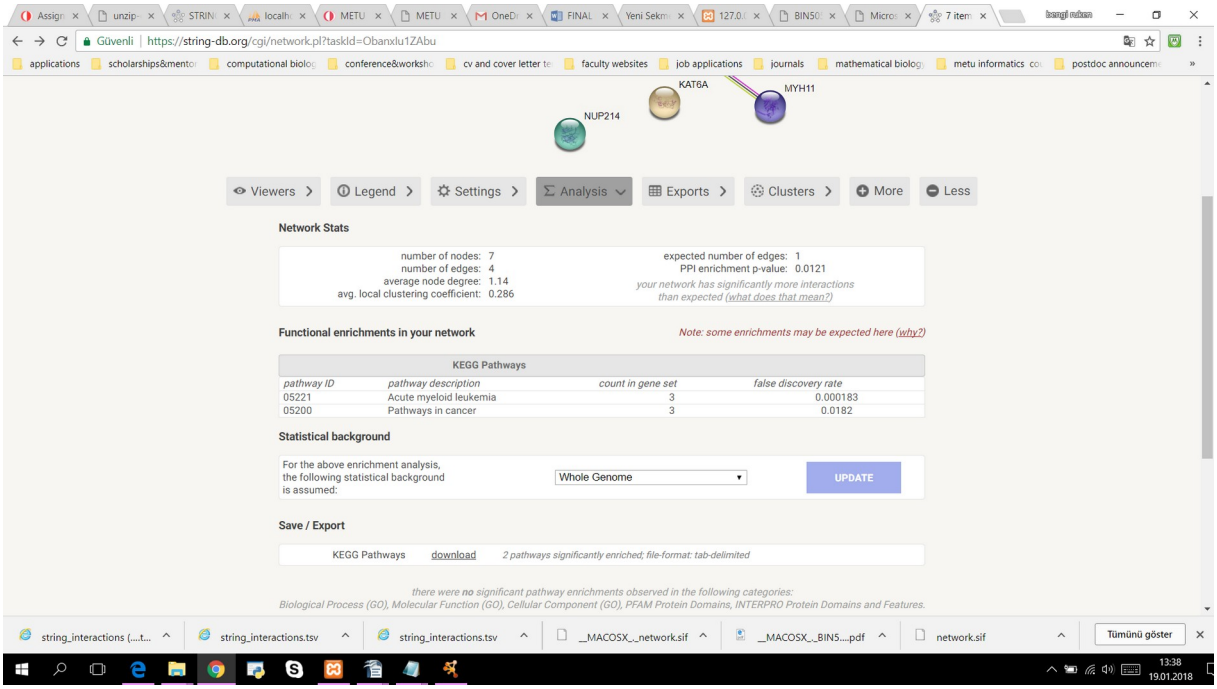
'Q92794' 'Q05516' 'P36888' 'P35658' 'P35749' 'Q14980' 'P51692'

We wrote these Ids to the “Multiple Proteins” segment of the STRING database. You can see the results in the following figures.

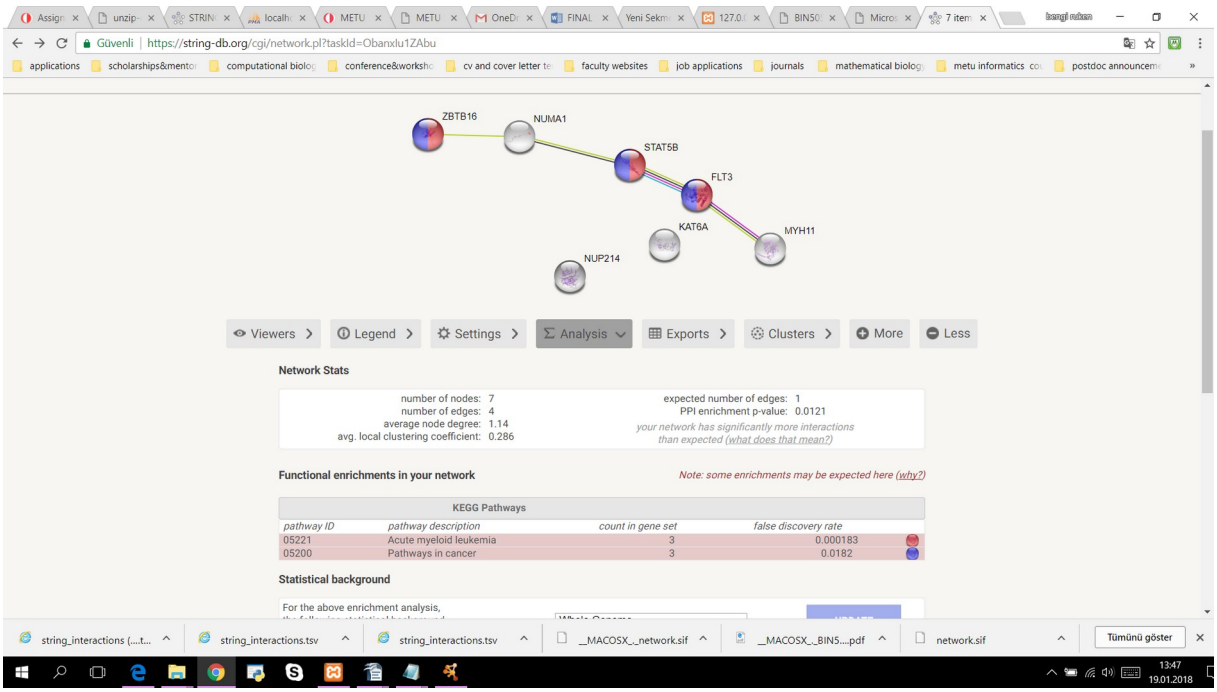




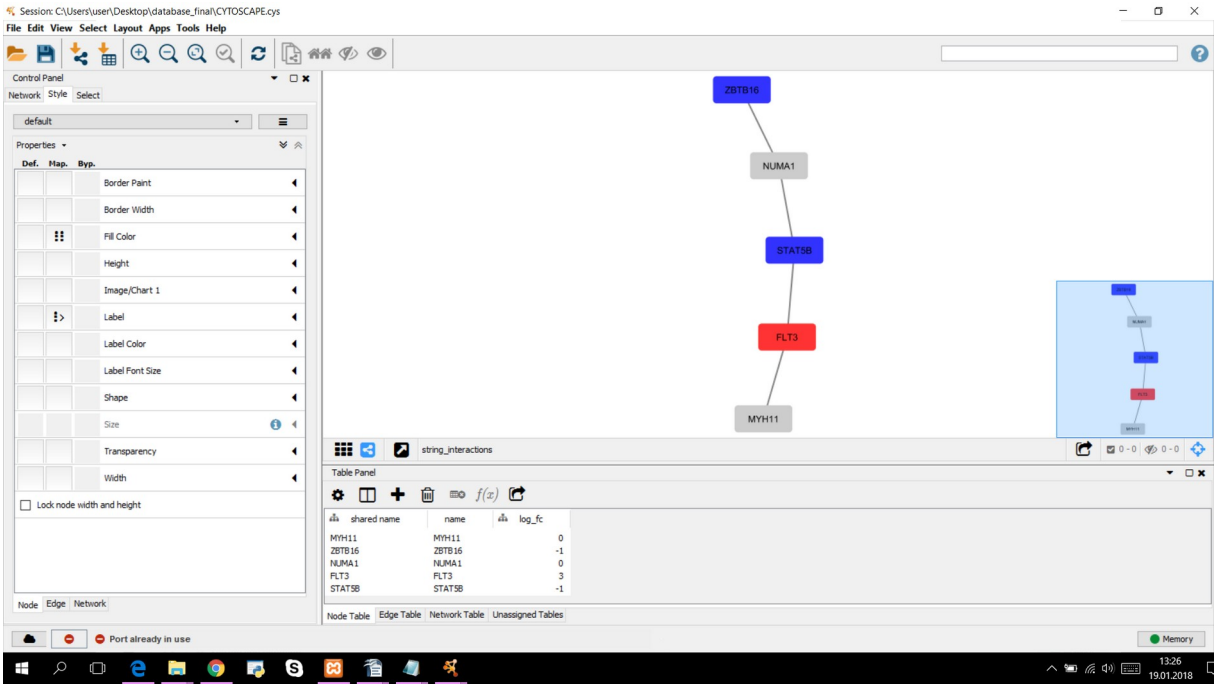
5 of the proteins in the given collection have more interactions among themselves. 2 proteins in the collection does not have any interactions with others, they are independent. As you see in the following figure 3 proteins in the given collection in the are involved in the pathways related to “acute myeloid leukemia” and cancer.



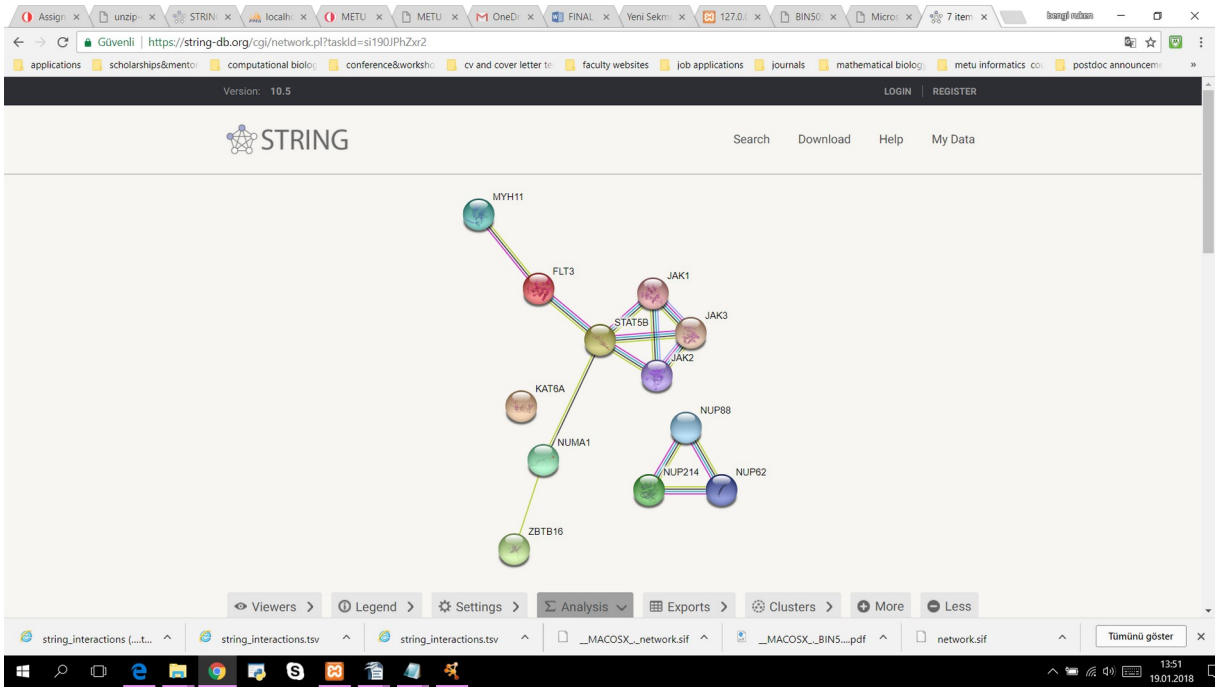
The proteins represented in red nodes are related to “acute myeloid leukemia” and the ones represented in blue nodes are related to cancer pathways. These three proteins are “ZBTB16,STAT5B,FLT3”. You can see the related screenshot below.



We uploaded the the information about the proteins with uniprot ids 'Q92794' 'Q05516' 'P36888' 'P35658' 'P35749' 'Q14980' 'P51692' to Cytoscape. By using the “Fill colour” option in Cytoscape we gave colour to the nodes according their logFC values. The proteins with red colour have positive logFC values and are up regulated, the proteins with blue colour have negativ logFC values and are down regulated. The proteins with grey color have 0 logFC values. Below you can see the protein intercations in force-directed layout.



If we click to the more button in STRING, we see the second shell interactions of the proteins. According to STRING, our set of proteins is not very well connected. The set of proteins may not have been studied well by STRING database and their interactions may not be known by this database.



REFERENCES

Acute Myeloid Leukemia. (2014, April 29). Retrieved from https://cancergenome.nih.gov/cancersselected/acutemyeloidleukemia

Druker, B. J., Sawyers, C. L., Kantarjian, H., Resta, D. J., Reese, S. F., Ford, J. M., Capdeville, R., Talpaz, M. Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. New Eng. J. Med. 344: 1038-1042, 2001. Note: Erratum: New Eng. J. Med. 345: 232 only, 2001. [PubMed: 11287973]

Leukemia stages. (n.d). Retrieved from https://www.cancercenter.com/leukemia/stages/

National Cancer Institute. (2017, May 19). Drugs Approved for Leukemia. Retrieved from https://www.cancer.gov/about-cancer/treatment/drugs/leukemia#1

SEER Stat Fact Sheets: Leukemia. (n.d.). Retrieved from https://web.archive.org/web/20140706145706/http://www.seer.cancer.gov/statfacts/html/leuks.html

The Cancer Genome Atlas Network. Genomic and epigenomic landscape of adult de novo acute myeloid leukemia. New England Journal of Medicine. Online May 1, 2013. In print May 30, 2013. DOI: 10.1056/NEJMoa1301689

https://www.ensembl.org/biomart/martview/27306e3a45b4ce142e6ad783e763c5de

https://www.ebi.ac.uk/QuickGO/

https://david.ncifcrf.gov/tools.jsp

http://www.cytoscape.org/download.php

https://www.ebi.ac.uk/interpro/

http://www.uniprot.org/uniprot/?query=reviewed:yes

phpMyAdmin

http://www.genome.jp/kegg/

https://string-db.org/cgi/input.pl?UserId=input_page_show_search=on

https://www.ncbi.nlm.nih.gov/geo/geo2r/