DOWNLOADING THE PDC GENES FROM NCBI

- Setting Email for Entrez:
    - The `Entrez.email` variable is set to provide the email address to NCBI's Entrez utilities. This is a requirement by NCBI to identify users and prevent abuse of their services.
- Function `download_gene_sequence`:
    - This function downloads the gene sequence from NCBI using specified parameters.
    - Parameters:
        - `accession_number`: Accession number of the gene sequence in the NCBI database.
        - `start`: Start position of the gene sequence.
        - `end`: End position of the gene sequence.
        - `pdc_gene_number`: Identifier for the PDC (Pyruvate decarboxylase) gene.
        - `reverse_complement` (optional): Boolean flag indicating whether to retrieve the reverse complement of the sequence.
    - Steps:
        - Print Start Message: Displays a message indicating the start of the download process for the specified gene sequence.
        - Fetch Sequence: Uses NCBI's Entrez utility (`efetch`) to fetch the gene sequence from the nucleotide database.
        - Read Sequence: Reads the fetched sequence using Biopython's `SeqIO` module.
        - Close Handle: Closes the handle to the sequence data retrieved from NCBI.
        - Reverse Complement: Optionally applies reverse complement to the sequence if specified.
        - Write to File: Writes the downloaded gene sequence to a FASTA file named according to the PDC gene number.
        - Print Success Message: Displays a message indicating successful download of the gene sequence.
    - Returns: The downloaded sequence as a `SeqRecord` object.
- Downloading PDC1 Gene Sequence:
    - Downloads the PDC1 gene sequence from NCBI.
    - Parameters:
        - `accession_number_pdc1`: Accession number of the PDC1 gene sequence.
        - `start_pdc1`: Start position of the PDC1 gene sequence.
        - `end_pdc1`: End position of the PDC1 gene sequence.
    - Flags reverse complement to be applied.
- Downloading PDC5 Gene Sequence:
    - Downloads the PDC5 gene sequence from NCBI.
    - Parameters:

        - `accession_number_pdc5`: Accession number of the PDC5 gene sequence.
        - `start_pdc5`: Start position of the PDC5 gene sequence.
        - `end_pdc5`: End position of the PDC5 gene sequence.
- Downloading PDC6 Gene Sequence:
    - Downloads the PDC6 gene sequence from NCBI.
    - Parameters:
        - `accession_number_pdc6`: Accession number of the PDC6 gene sequence.
        - `start_pdc6`: Start position of the PDC6 gene sequence.
        - `end_pdc6`: End position of the PDC6 gene sequence.
    - Flags reverse complement to be applied.

Applying Blast

Blast through code.

- I have provided the code to blast the PDC's with Saccharomyces cerevisiae genome
- Importing Necessary Modules:
    - The code imports required modules from Biopython, namely `NCBIWWW` and `NCBIXML`, for performing BLAST searches and parsing BLAST results, respectively.
- Defining `perform_blast()` Function:
    - This function performs a BLAST search against the NCBI BLAST database using the provided query sequence.
    - It returns a handle to the BLAST search results.
- Defining `parse_blast_results()` Function:
    - This function parses the BLAST search results to identify SNPs (Single Nucleotide Polymorphisms).
    - It iterates through each BLAST record, alignment, and High Scoring Pair (HSP) to identify SNPs.
    - Identified SNPs are stored in a set.
- Iterating Over Files and Performing BLAST Searches:
    - The code iterates over a list of file indices representing different genes.(ie PDC1, DC2, PDC3)

    - For each gene, it prepares a query sequence by extracting the sequence from the corresponding FASTA file.
    - It then performs a BLAST search using the `perform_blast()` function and parses the results using the `parse_blast_results()` function.
- Additional Notes:
    - Biopython's functionality is utilised for BLAST searches and result parsing.
    - The code specifically searches for SNPs in the Saccharomyces cerevisiae genome using the provided query sequences.
    - SNPs are identified by comparing the query and subject sequences from BLAST alignments.
- Summary:
    - This script automates the process of searching for SNPs in the Saccharomyces cerevisiae genome using BLAST.
    - It defines functions for performing BLAST searches and parsing results, iterates over a list of genes, and extracts sequences from FASTA files to perform BLAST searches.
    - Finally, it prints the identified SNPs
- OUTPUT ->

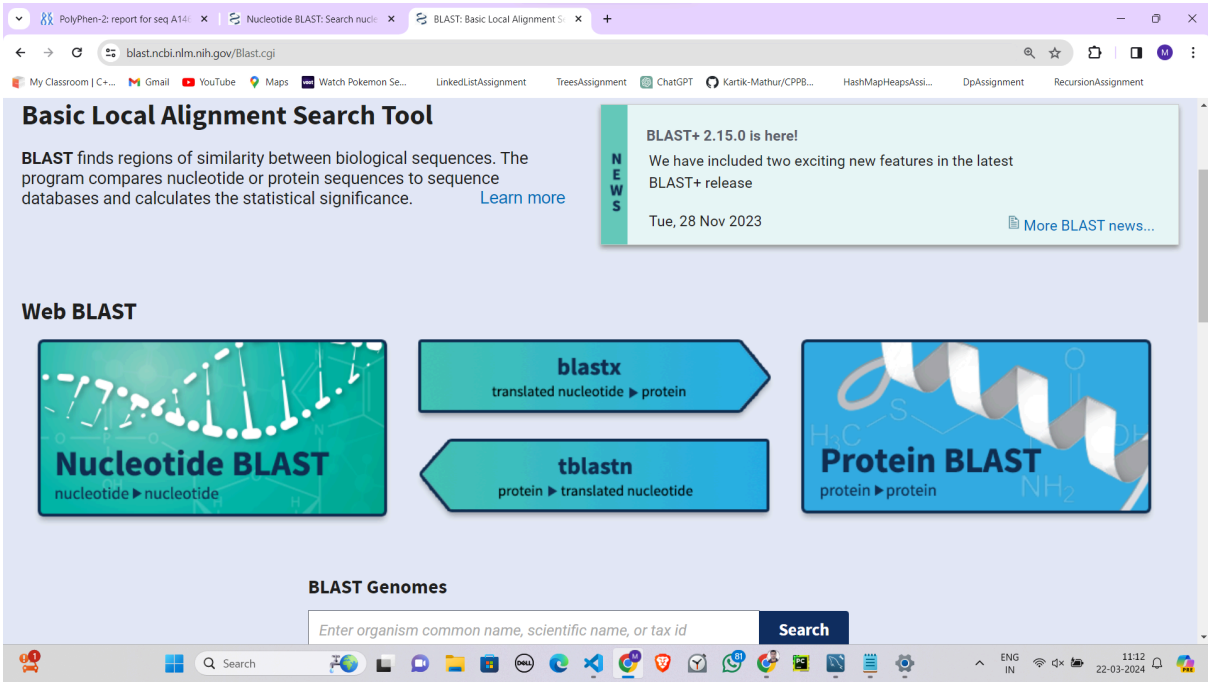
```
------------------------------------------------

For PDC  1
****Alignment****
Sequence: gi|1586061137|gb|CP036478.1| Saccharomyces cerevisiae strain ySR128 chromosome XII, complete sequence
Length: 1076801
E-value: 0.0
ATGTCTGAAATTACTTTGGGTAAATATTTGTTCGAAAGATTAAAGCAAGTCAACGTTAACACCGTTTTCGGTTTG...
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||...
ATGTCTGAAATTACTTTGGGTAAATATTTGTTCGAAAGATTAAAGCAAGTCAACGTTAACACCGTTTTCGGTTTG...

****Alignment****
Sequence: gi|1586061137|gb|CP036478.1| Saccharomyces cerevisiae strain ySR128 chromosome XII, complete sequence
Length: 1076801
E-value: 0.0
ATGTCTGAAATTACTTTGGGTAAATATTTGTTCGAAAGATTAAAGCAAGTCAACGTTAACACCGTTTTCGGTTTG...
|||||||||||  ||  ||  ||||||||||  ||  ||||||||  |  |||||||||  |||||||| |||||||||...
ATGTCTGAAATAACCTTAGGTAAATATTTATTTGAAAGATTGAGCCAAGTCAACTGTAACACCGTCTTCGGTTTG...

****Alignment****
Sequence: gi|1389182097|gb|CP029160.1| Saccharomyces cerevisiae strain SY14 chromosome I, complete sequence
Length: 11848804
E-value: 0.0
ATGTCTGAAATTACTTTGGGTAAATATTTGTTCGAAAGATTAAAGCAAGTCAACGTTAACACCGTTTTCGGTTTG...
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||...
ATGTCTGAAATTACTTTGGGTAAATATTTGTTCGAAAGATTAAAGCAAGTCAACGTTAACACCGTTTTCGGTTTG...
```

```
Number of snps -> 783
Some of the snps ............ -> [12, 15, 16, 18, 21, 27, 30, 33, 42, 44, 45, 51, 54, 55, 56]
```

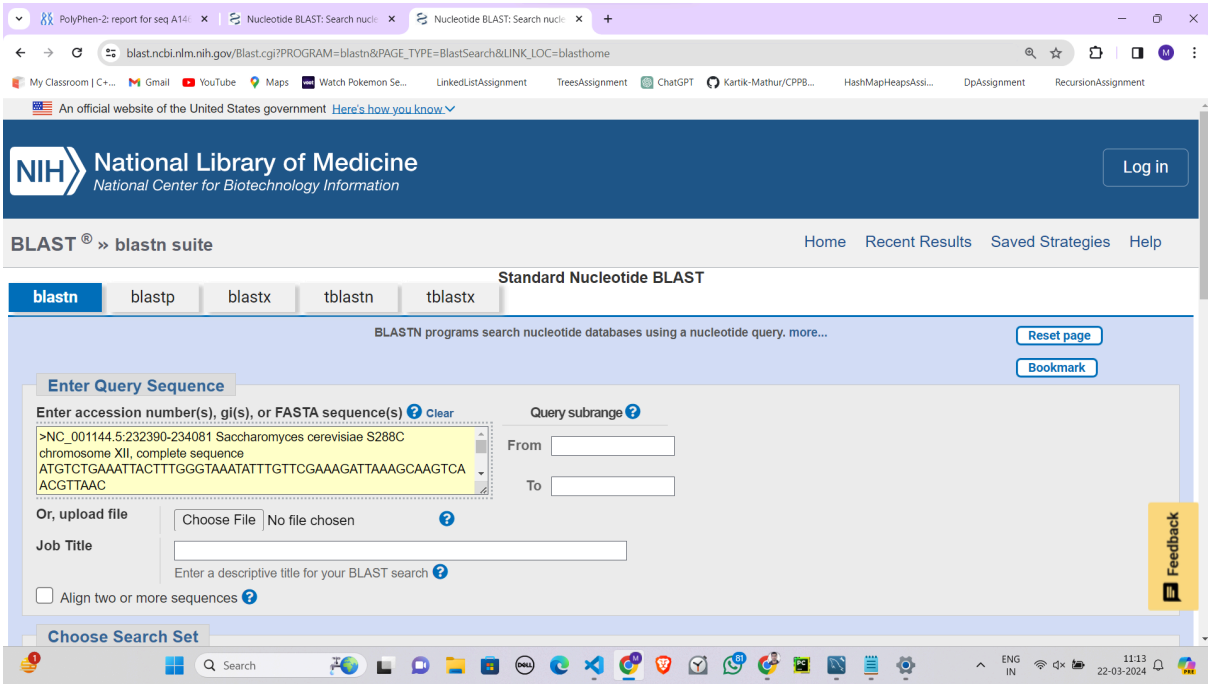
- NOTE- *The above ss is only for pdc 1. Also the number of snp indicates number of positions and does not include number of combinations of mutations on a single position. Therefore my number of snps might be less.*
- NOTE: The number of snp is indicated for separate PDC strains
    - First printing the PDC gene number
    - Then printing the subject sequence information
    - Then the alignment
    - At the end of each pdc I am also printing the total number of SNP found and printing first 10 or 15 positions.

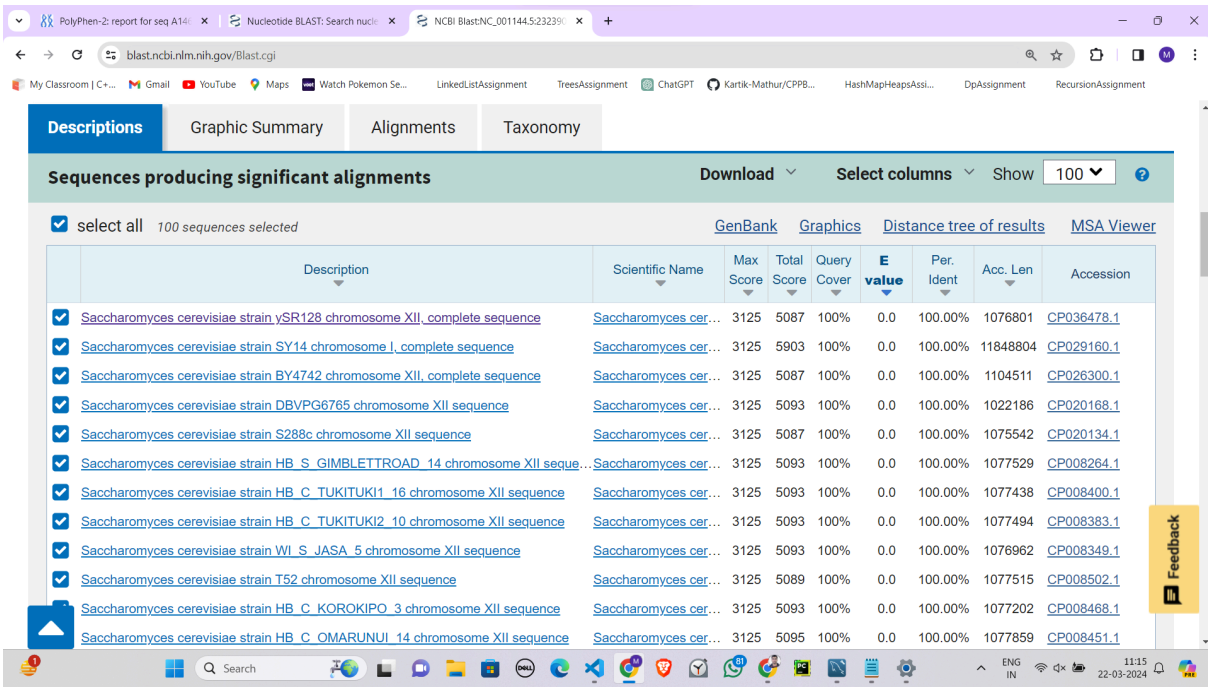
BLAST through UI

- First visit the official blast website (link for blast website)



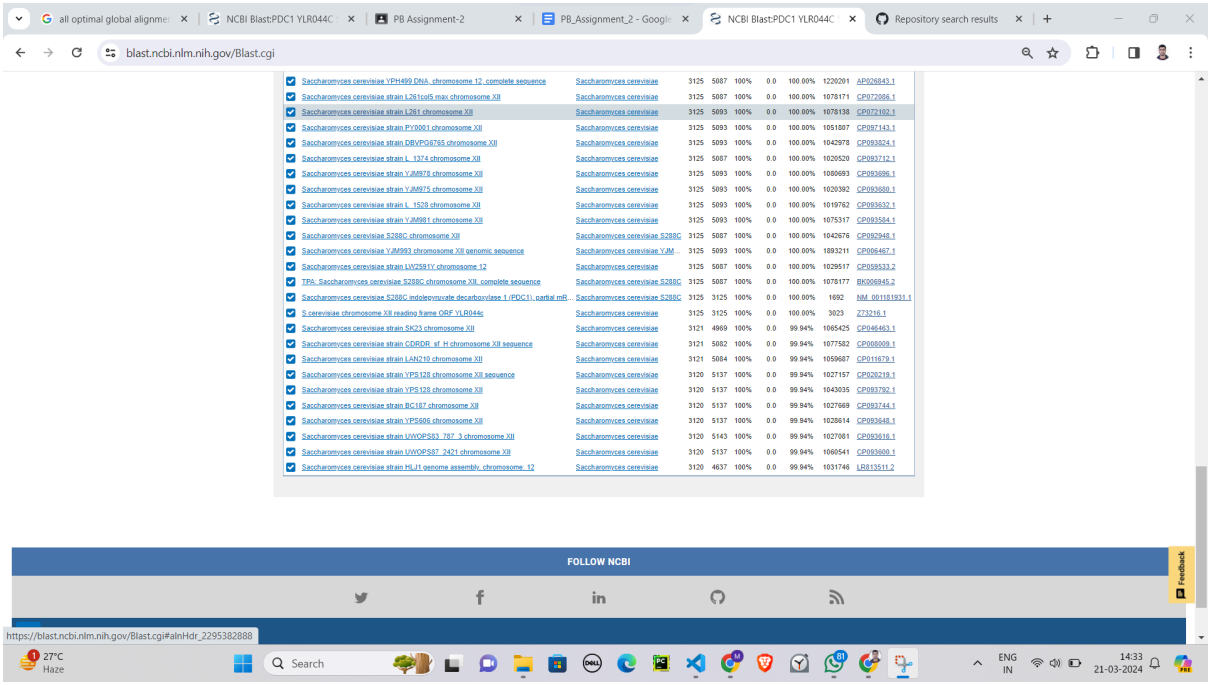
- Then Click on the Nucleotide Blast (nucleotide -> nucleotide) button
- After this step the following page will open

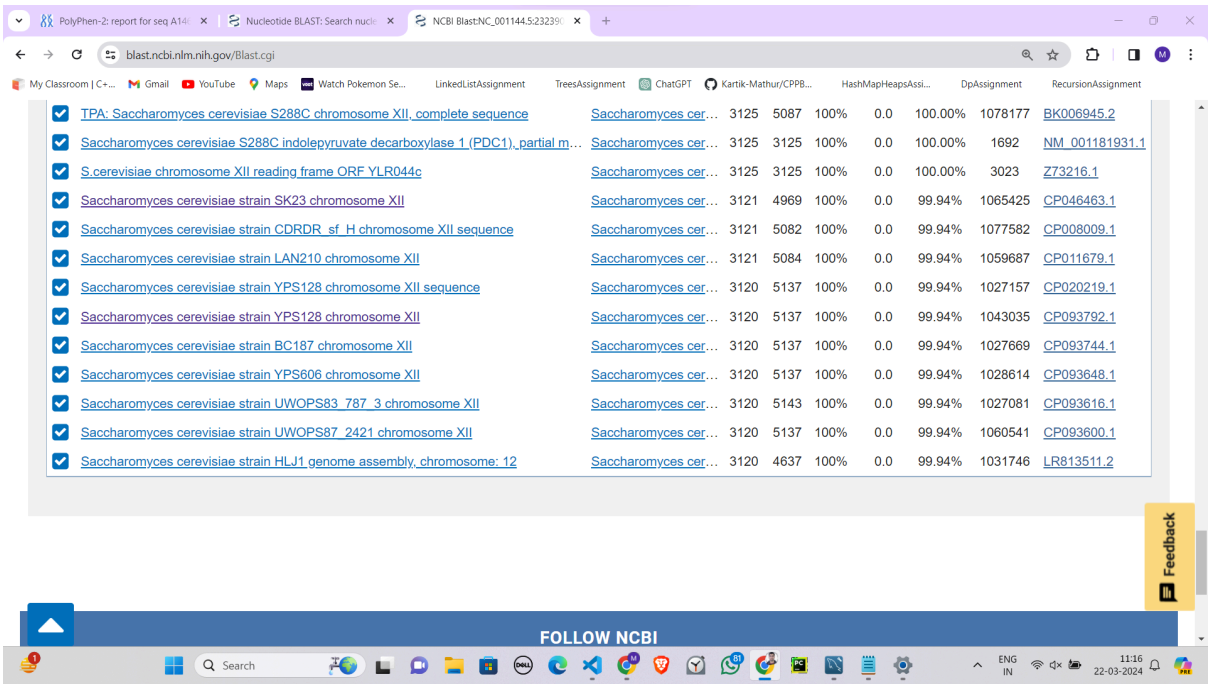
- 

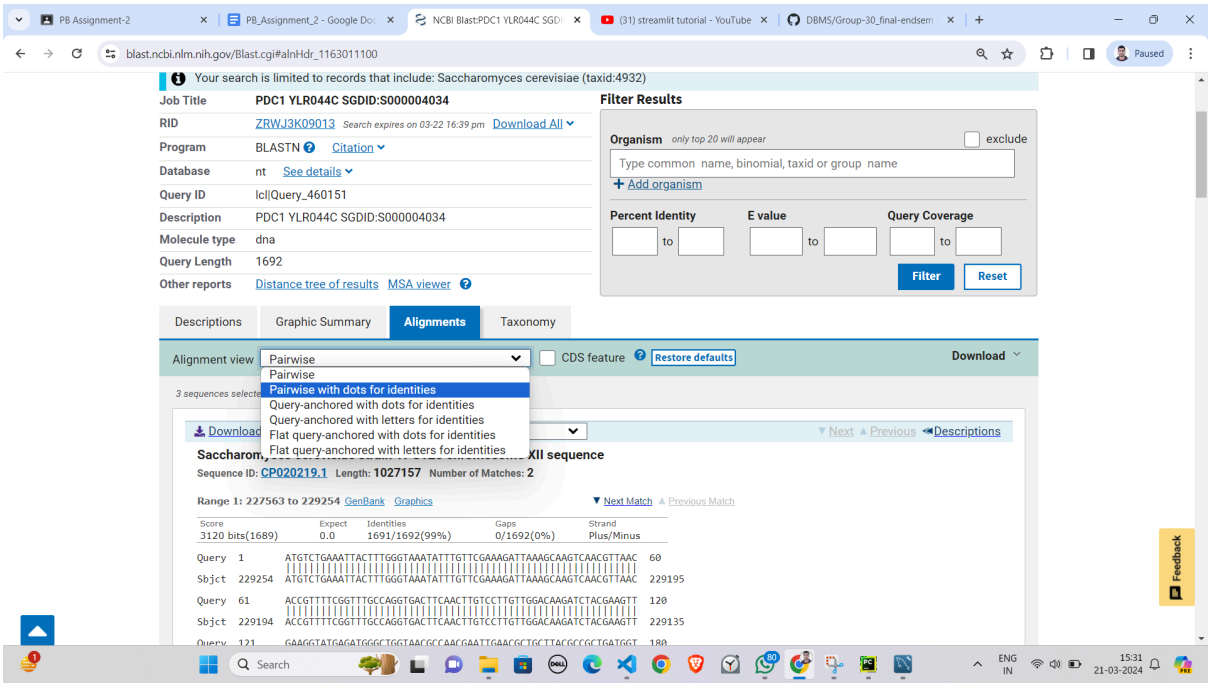
- Copy paste the fasta file of the pdc gene 1 in the QUERY Sequence Box shown in the above web page. Also enter the organism name ie Saccharomyces cerevisiae across which you need to perform blast of the query. After filling this information simply click on the blast button.
- 

- The above page will appear on the screen as a result.
- Note that there is a column of percent identity as highlighted in the above image. As we are finding SNP (Single Nucleotide Polymorphism) which means that only one nucleotide should change in the alignment. Implies that Percent Identity = {len(query)-1/len(query)}*100

- So as the len of pdc gene 1 is 1692 so Percent Identity = (1691/1692)*100 ie approx 99.94%
- In the more general sense, it is not necessary for only one nucleotide to change in a sequence. Check the references section for more details.
-





- Click on alignment section set alignment view as dot plot.(See below)
- 

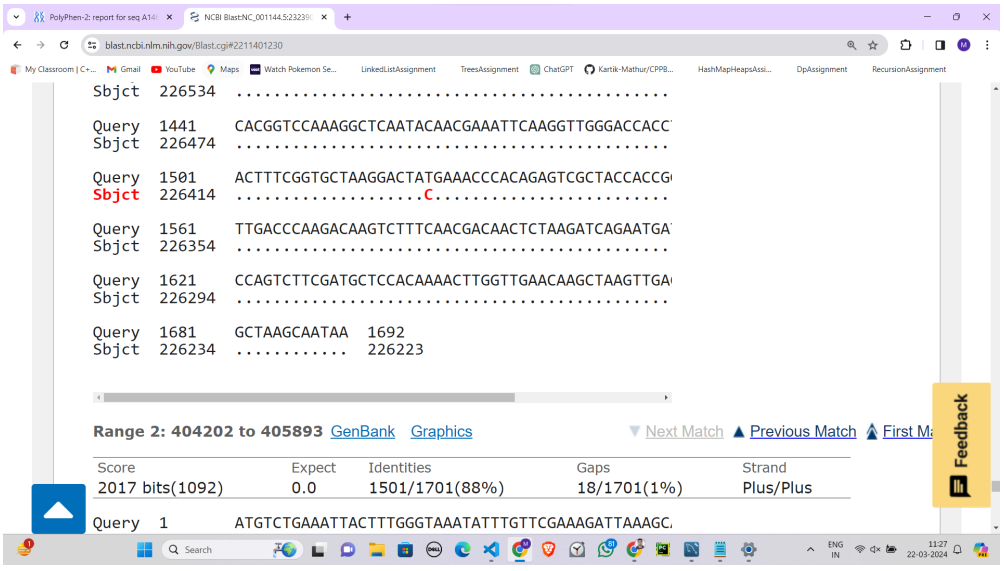
- On choosing **Saccharomyces cerevisiae strain YPS128 chromosome XII sequence** I get an alignment overview with the SNP position 1521
- 

Range 2: 393754 to 395445 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

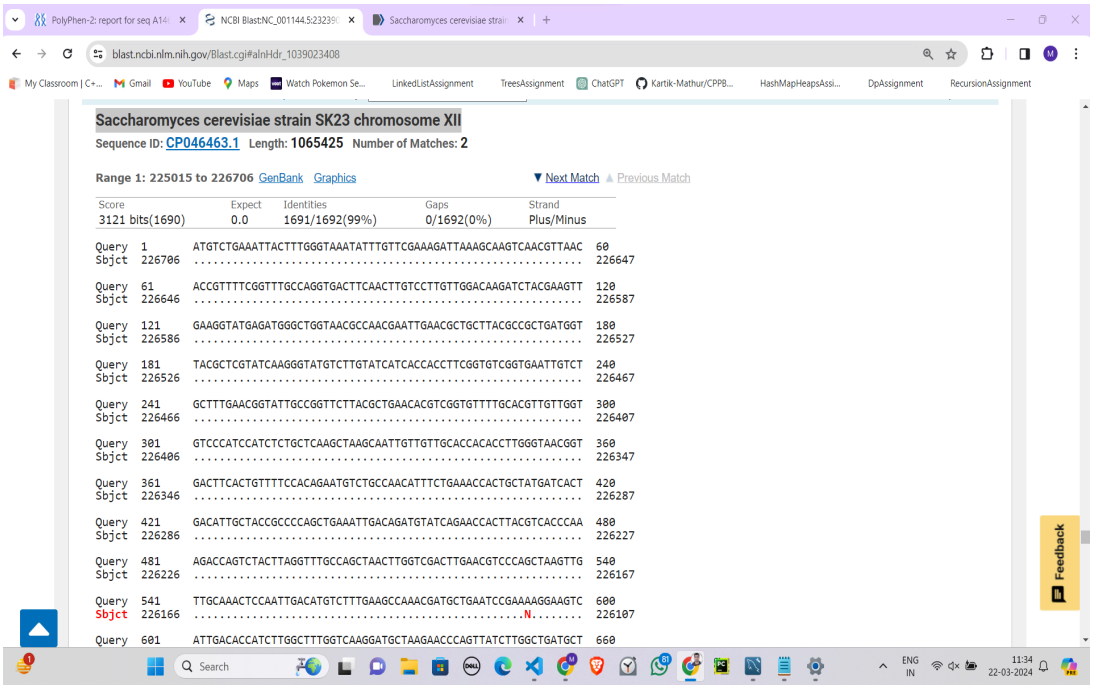
Score	Expect	Identities	Gaps	Strand
1967 bits(1065)	0.0	1492/1701(88%)	18/1701(1%)	Plus/Plus

```
Query  1       ATGTCTGAAATTACTTTGGGTAAATATTTGTTCGAAAGATTAAAGCAAGTCAACGTTAAC  60
Sbjct  393754  ..........A..C..A..........A..T........G.GC.........TG....  393813

Query  61      ACCGTTTTCGGTTTGCCAGGTGACTTCAACTTGTC-CTTGTTGGACAAGATCTACGAAGT  119
Sbjct  393814  .....C...................T........T...-.....T...C.T..T.....  393872

Query  120     TGAAGGTATGAGATGGGCTGGTAACGCCAACGAATTGAACGCTGCTTACGCCGCTGATGG  179
Sbjct  393873  CA.......................T................C..T..T........  393932

Query  180     TTACGCTCGTATCAAGGGTATGTCTTGTATCATCACCACCTTCGGTGTCGGTGAATTGTC  239
Sbjct  393933  ......................C.....T..T.............T...........  393992

Query  240     TGCTTTGAACGGTATTGCCGGTTCTTACGCTGAACACGTCGGTGTTTTGCACGTTGTTGG  299
Sbjct  393993  .........T.........................T.......................  394052

Query  300     TGTCCCATCCATCTCTGCTCAAGCTAAGCAATTGTTGTTGCACCACACCTTGGGTAACGG  359
Sbjct  394053  ...T............T.........................T..T.............  394112

Query  360     TGACTTCACTGTTTTCCACAGAATGTCTGCCAACATTTCTGAAACCACTGCTATGATCAC  419
Sbjct  394113  ..................................................C........  394172

Query  420     TGACATTGCTACCGCCCCAGCTGAAATTGACAGATGTATCAGAACCACTTACGTCACCCA  479
Sbjct  394173  ...T.......A...T..............................C...ACT.....  394232

Query  480     AAGACCAGTCTACTTAGGTTTGCCAGCTAACTTGGTCGACTTGAACGTCCCAGCTAAGTT  539
Sbjct  394233  ..............G.....................T.................C.....  394292

Query  540     GTTGCAAACTCCAATTGACATGTCTTTGAAGCCAAACGATGCTGAATCCGAAAAGGAAGT  599
Sbjct  394293  A...G.............T...................C......G.T...GCT.....  394352

Query  600     CATT-GACACCATCTTGGCTTTGGTCAAGGATGCTAAGAACCCAGTTATCTTGGCTGATG  658
Sbjct  394353  TG..A..-..TG.TG.T.AA...A.....................................  394411

Query  659     CTTGTTGTTCCAGACACGACGTCAAGGCTGAAACTAAGAAGTTGATTGACTTGACTCAAT  718
Sbjct  394412  .....GC...T.....T..T.........................G.............  394471

Query  719     TCCCAGCTTTCGTCACCCCAATGGGTAAGGGTTCCATTGACGAACAACACCCAAGATACG  778
Sbjct  394472  ......T..A.....................G.T.........................  394531

Query  779     GTGGTGTTTACGTCGGTACCTTGTCCAAGCCAGAAGTTAAGGAAGCCGTTGAATCTGCTG  838
Sbjct  394532  ............T...........T.GA............A.G..T..A..........  394591
```

- For strain **Saccharomyces cerevisiae strain UWOPS87_2421 chromosome XII** SNP is 1521



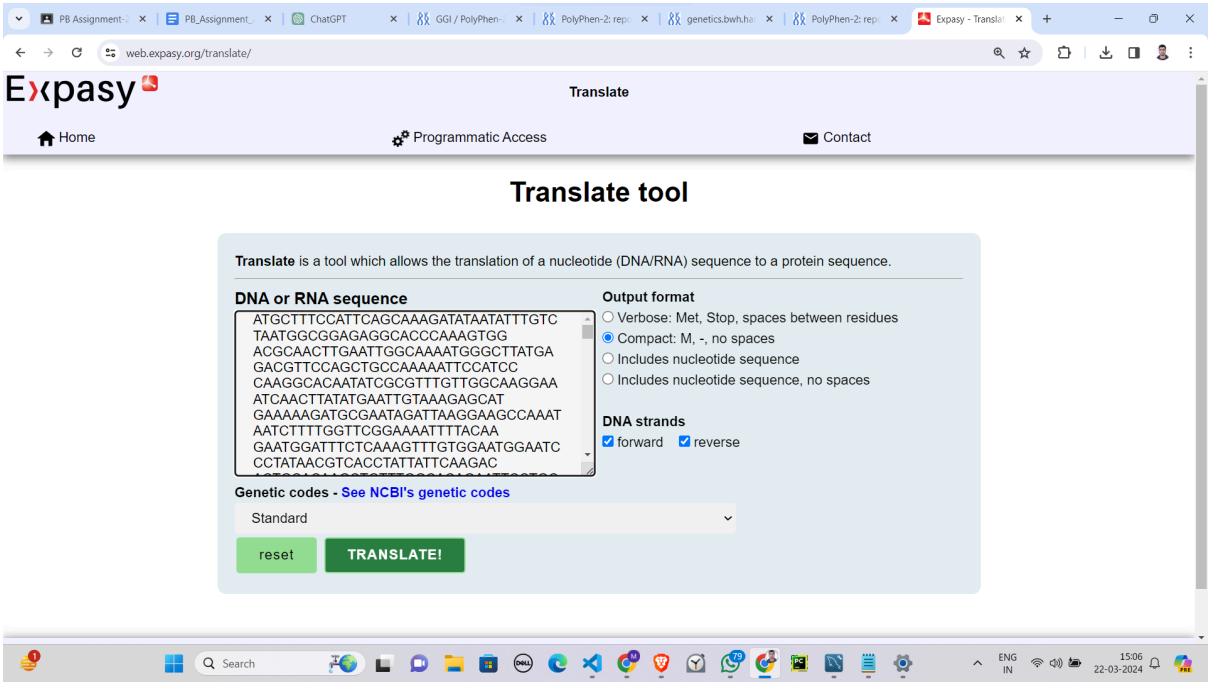
- For strain **Saccharomyces cerevisiae strain SK23 chromosome XII** SNP is 592

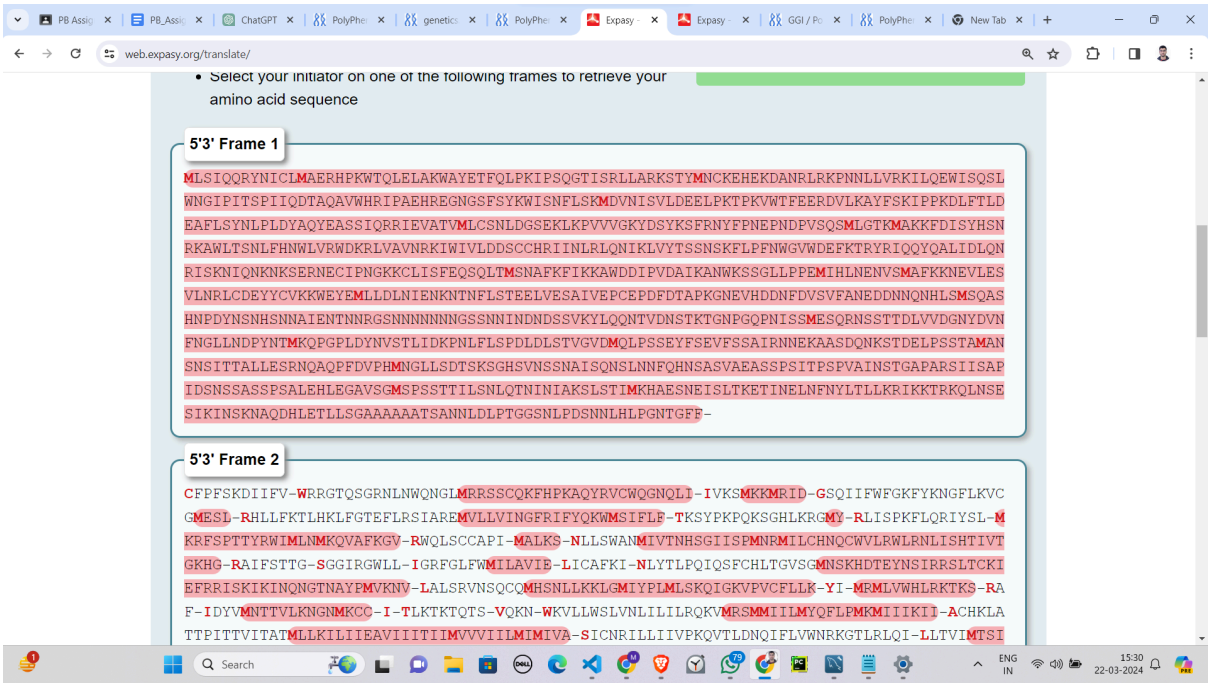


- Similarly following the same steps for PDC 5 we get SNP at
  - 879 for **Saccharomyces cerevisiae strain WI_S_OAKURA_4 chromosome XII sequence**
  - 1214 for **Saccharomyces cerevisiae strain WA_C_CODDINGTON_2 chromosome XII, partial sequence**
  - And many more
- Similarly following the same steps for PDC 6 we get SNP at
  - 1142 for **Saccharomyces cerevisiae strain L_1528 chromosome VII**
  - 1437 for **Saccharomyces cerevisiae YJM1356 chromosome VII genomic sequence**
  - And many more

EXPASY TOOL FOR TRANSLATION

- Use this link to open the page (expassy link)

- 

- Copy paste the nucleotide sequence in the box and press on translate button.
- The output will be the following where there would be multiple frames shown in red and we would have to chose
- 

BIOPYTHON CODE FOR TRANSLATION

Reading the Fasta file and extracting the sequence from it

```
def extractSeq(file_name):
    for seq_record in list(SeqIO.parse(file_name, "fasta"))[:5]:


        pdc_gene_seq = str(seq_record.seq)

        return  pdc_gene_seq
```

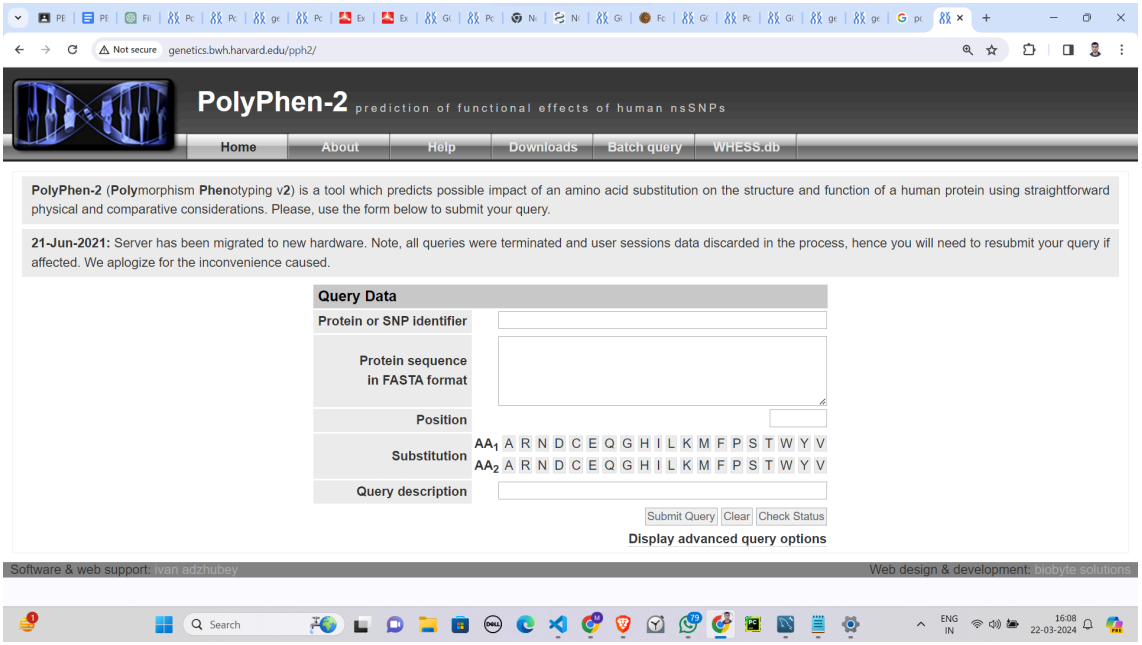
```
def translate_sequence(sequence):
    seq = Seq(sequence)
    protein_sequence = seq.translate()
    return str(protein_sequence)

sequ = extractSeq("PDC5.fasta")
sequStrain = extractSeq("strain.fasta")
print(translate_sequence(sequ))
print(translate_sequence(sequStrain))
```

- Function `extractSeq(file_name)`:
  - Purpose: This function extracts DNA sequences from a FASTA file.
  - Parameters:
    - `file_name` (string): The name of the FASTA file containing DNA sequences.
  - Returns:
    - A string representing the DNA sequence extracted from the FASTA file.
  - Description:
    - Iterates over the first five sequence records in the FASTA file specified by `file_name`.
    - Converts the DNA sequence of each record to a string.
    - Returns the DNA sequence of the first record found.
- Function `translate_sequence(sequence)`:
  - Purpose: This function translates a DNA sequence into a protein sequence.

    - Parameters:
        - `sequence` (string): The DNA sequence to be translated.
    - Returns:
        - A string representing the translated protein sequence.
    - Description:
        - Converts the input DNA sequence into a `Seq` object.
        - Utilizes Biopython's `translate()` function to translate the DNA sequence into a protein sequence.
        - Converts the translated protein sequence into a string and returns it.
- Main Code:
    - Calls `extractSeq("PDC5.fasta")` to extract a DNA sequence from the "PDC5.fasta" file.
    - Calls `translate_sequence()` to translate the extracted DNA sequence into a protein sequence.
    - Prints the translated protein sequence to the console.
- Overall Description:
    - The code snippet first extracts a DNA sequence from a FASTA file using the `extractSeq()` function.
    - It then translates the extracted DNA sequence into a protein sequence using the `translate_sequence()` function.
    - Finally, it prints the translated protein sequence to the console.
    - This code can be used to quickly extract and translate DNA sequences for further analysis in protein-related studies.

POLYPHEN PREDICTION TOOL

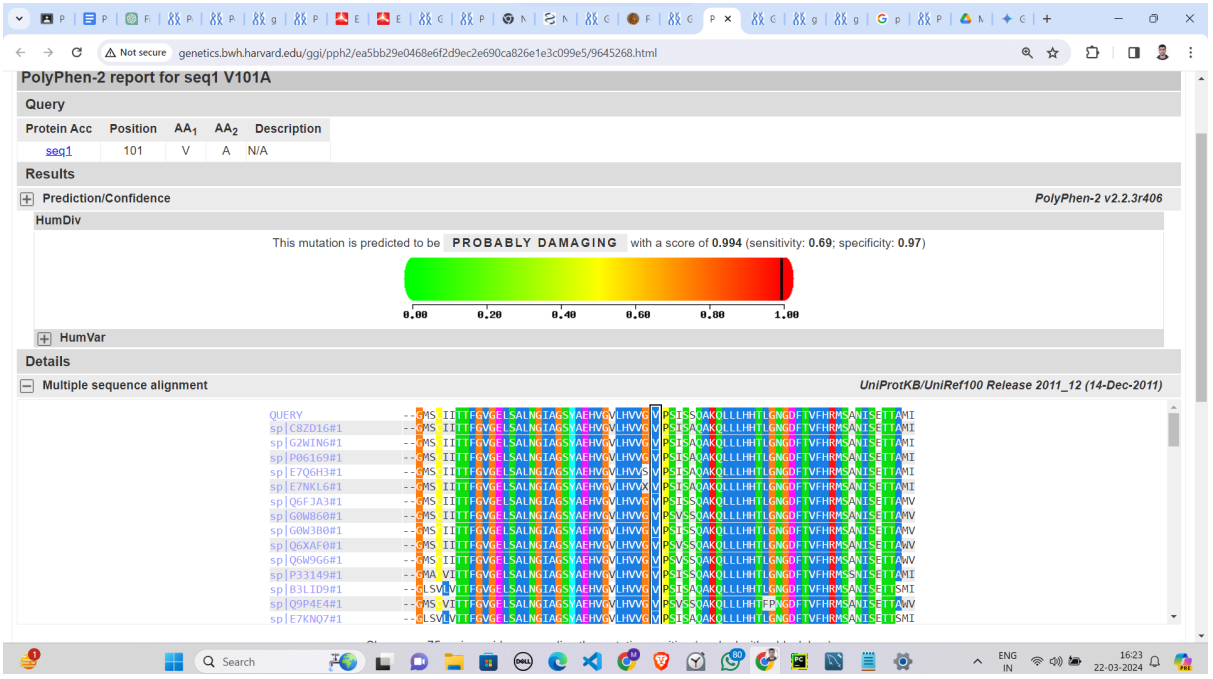
- Go to the PolyPhen-2 web server: PolyPhen-2.
- Paste or upload the amino acid sequence of the protein you want to analyse.
- Ensure that your input sequence is in FASTA format and contains only the amino acid sequence (without headers or additional information).
- Choose the type of mutation you want to predict and its position. (Ex G -> A at 143 position)


Not secure
genetics.bwh.harvard.edu/pph2/
PolyPhen-2
prediction of functional effects of human nsSNPs
Home
About
Help
Downloads
Batch query
WHESS.db
PolyPhen-2 (Polymorphism Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. Please, use the form below to submit your query.
21-Jun-2021: Server has been migrated to new hardware. Note, all queries were terminated and user sessions data discarded in the process, hence you will need to resubmit your query if affected. We aplogize for the inconvenience caused.
Query Data
Protein or SNP identifier
Protein sequence in FASTA format
Position
Substitution
AA1 A R N D C E Q G H I L K M F P S T W Y V
AA2 A R N D C E Q G H I L K M F P S T W Y V
Query description
Submit Query
Clear
Check Status
Display advanced query options
Software & web support: Ivan Adzhubey
Web design & development: biobyte solutions
Search
ENG
IN
16:08
22-03-2024

PREDICTIONS

1. Result 1

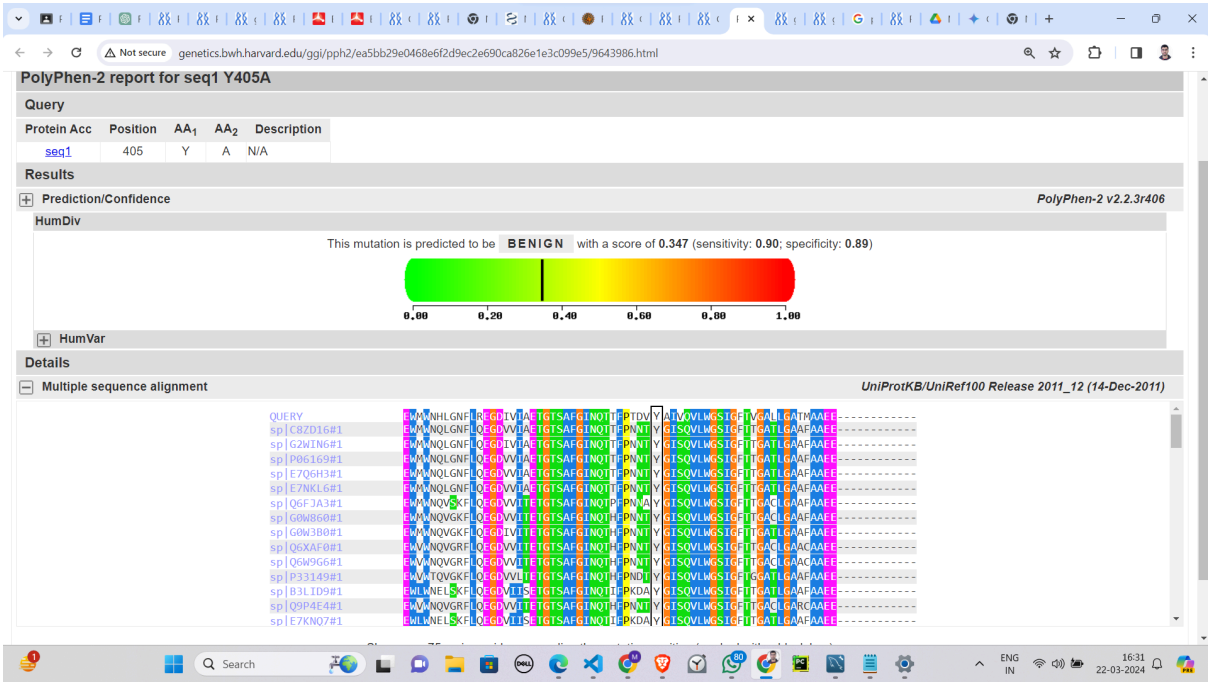
a. **STRAIN USED:** Pyruvate decarboxylase 6 (PDC6) - Saccharomyces cerevisiae
b. **POSITION:** 101
c. **MUTATION:** Valine -> Alanine (V -> A)
d. **PREDICTION:** From the report, the mutation seq1 V101A is labelled as “PROBABLY DAMAGING” with a score of 0.994 (sensitivity: 0.96, specificity: 0.97). This score is quite high, suggesting a high confidence in the prediction. The sensitivity and specificity values also indicate a high true positive rate and a low false positive rate, respectively.



2. Result 2

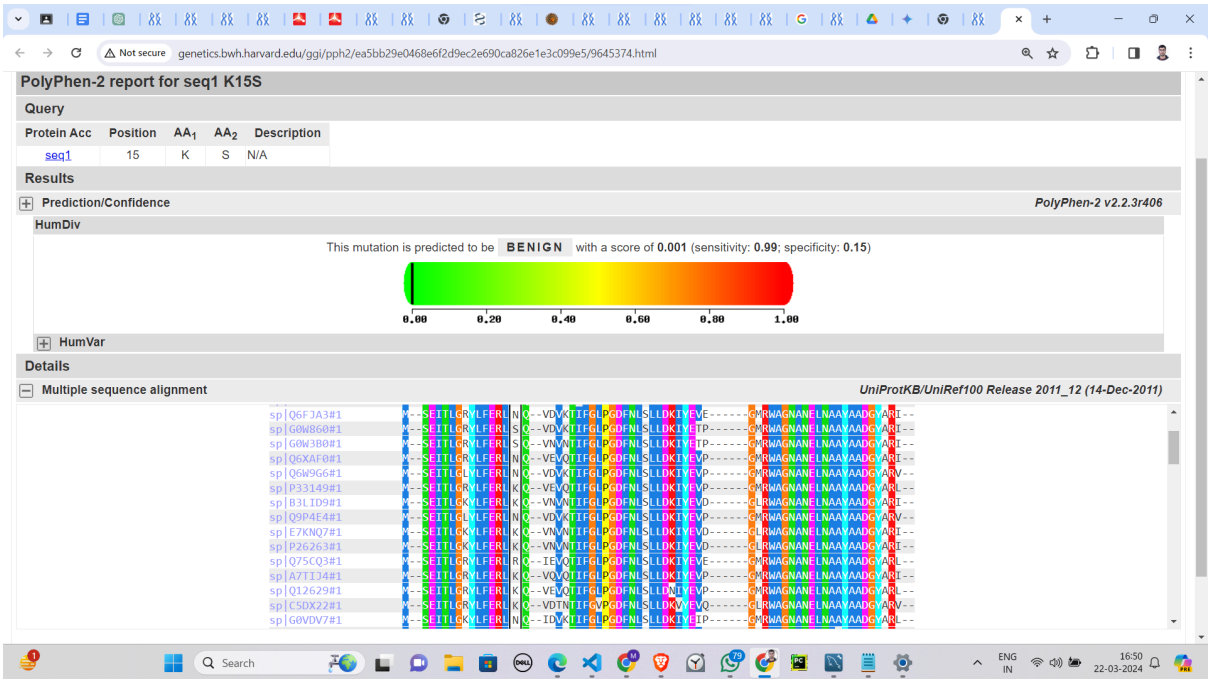
a. **STRAIN USED:** Pyruvate decarboxylase 5 (PDC5) - Saccharomyces cerevisiae
b. **POSITION:** 406
c. **MUTATION:** Tyrosine -> Alanine (Y -> A)
d. **PREDICTION:** This report indicates that a specific mutation (seq1 Y405A) is predicted to be “BENIGN” with a score of 0.347. The sensitivity is 0.90 and specificity is 0.89, suggesting high accuracy in the prediction model. The confidence in this prediction is visualised on a gradient scale from green (benign) to red (probably damaging), with this particular mutation falling into the green zone.

   Based on this information, one can predict that this particular protein mutation (Y405A) would not have damaging effects or contribute to disease development due to its benign nature as indicated by the low score of 0.347 on PolyPhen-2’s scale.



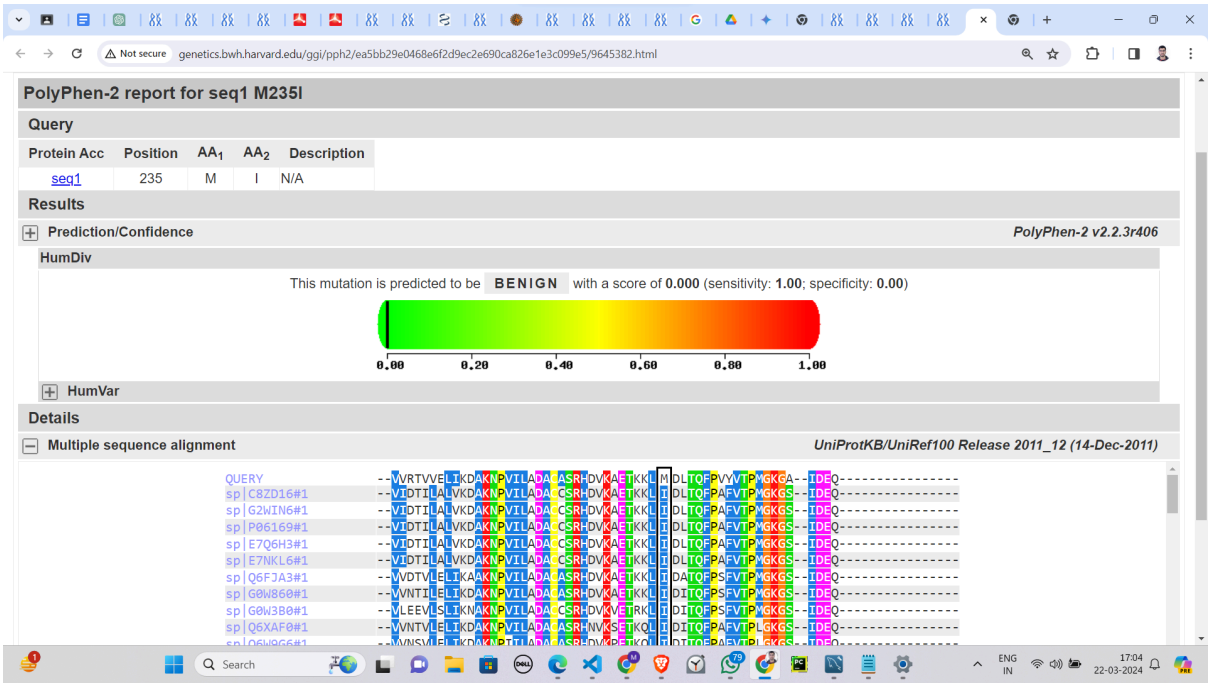
3. Result 3

a. **STRAIN USED:** Pyruvate decarboxylase 1 (PDC1) - Saccharomyces cerevisiae
b. **POSITION:** 15
c. **MUTATION:** Lysine -> Serine (K -> S)
d. **PREDICTION:** The report you've shared indicates that a specific mutation (seq1 K15S) is predicted to be "BENIGN" with a low score of 0.001 and sensitivity and specificity as 0.99 and 0.15 respectively. In the context of PolyPhen-2, a "benign" prediction suggests that the mutation is not expected to have a significant impact on the protein's function. The score, which ranges from 0 (benign) to 1 (probably damaging), provides a measure of confidence in this prediction. A score of 0.001 is very close to 0, indicating a high confidence that the mutation is benign



4. Result 4

a. **STRAIN USED:** Pyruvate decarboxylase 5 (PDC5) - Saccharomyces cerevisiae
b. **POSITION:** 235
c. **MUTATION:** Methionine-> Isoleucine (M -> I)
d. **PREDICTION:** This report indicates that a specific mutation (seq1 M235I) is predicted to be “BENIGN” with a score of 0.000.The sensitivity is 1 and specificity is 0. As it is labelled as “benign” and the score is 0.000, it suggests that this mutation might not have significant implications on the protein’s function.



CITATIONS

1. **Effects of pyruvate decarboxylase (pdc 1, pdc 5) gene knockout on the production of metabolites in two haploid Saccharomyces cerevisiae strains**

   Zhang W;Kang J;Wang C;Ping W;Ge J;
   https://pubmed.ncbi.nlm.nih.gov/33881948/

2. **Development of Single Nucleotide Polymorphism (SNP)-Based Triplex PCR Marker for Serotype-specific Escherichia coli Detection**

   Rahman et al.
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8874422/

3. **PolyPhen-2 Tutorials**

   The Jackson Laboratory is an independent, nonprofit organization focusing on mammalian genetics research to advance human health. Our mission is to discover precise genomic solutions for disease and empower the global biomedical community in the shared quest to improve human health.
   https://www.jax.org/-/media/jaxweb/files/education-and-learning/tutorials/polyphen2_tutorials_clickable.pdf

4. **ExPASy: The proteomics server for in-depth protein knowledge and analysis**

   Gasteiger et al.
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC168970/