CancerDR Manual

Developed at Dr. G.P.S. Raghava Group

CONTENTS

Introduction

CancerDR is the compilation of mutation data and pharmacological drug profiles from Catalogue Of Somatic Mutations (COSMIC) and Cancer Cell Line Encyclopedia (CCLE). In COSMIC, 138 anticancer drugs targeting a wide range of therapeutic targets were screened on 275-507 cell lines and in case of CCLE, 24 anticancer drugs were screened on more than 500 cell lines. We have compiled both of these data on a single platform in user-friendly format to make out some useful conclusions on cancer drug resistance. Methods of drug screening on cancer cell lines are available on COSMIC and CCLE websites respectively. We obtained the mutational status of 116 unique drug targets in cell lines form the parallel sequencing data available at CCLE website. Then we tried to correlate the mutation status of targets with IC50 values of the drugs. Along with this, we have been predicted the structures of these therapeutic targets. To correlate the effect of mutation on target structure, we predicted the tertiary structures of all mutants reported in CCLE for 116 drug targets. In addition to this, structural alignments of wild type drug targets with their mutants are generated and covered in the CancerDR. For sequence alignment purpose, all the given mutations were mapped on the target sequence. Then alignment of wild type and their mutants were generated. These sequences can also be used for the alignment of short reads to find out the mutations in a given target gene. After sequence alignment, we generated the phylogenetic trees of mutants to show the distance between them. We also tried to cluster the cell lines on the basis of IC50 values of anticancer drugs assayed on them. This will help to identify the most sensitive or most resistant cell lines for a particular drug. Similarly, we cluster the drugs to identify which drug is most effective or less effective on a particular cell line.

This manual will help users to understand the tools integrated in the CancerDR for their efficient use.

Note: Most of the CancerDR tool requires Java. So, make sure that Java is installed on your system.

1. Search Tool

Search tool is divided into four parts:

1.1 Drug Targets: CancerDR contains the information about 116 anticancer drug targets. For the efficient retrieval of this information, we have integrated this tool. If user wants the information about some drug target e.g. AKT1, just type AKT1 in the input box and click search button at the bottom of the page. This will return only that information, which is checked by the user in their corresponding check box. User can check, CHECK ALL button to display all the information (Figure 1).

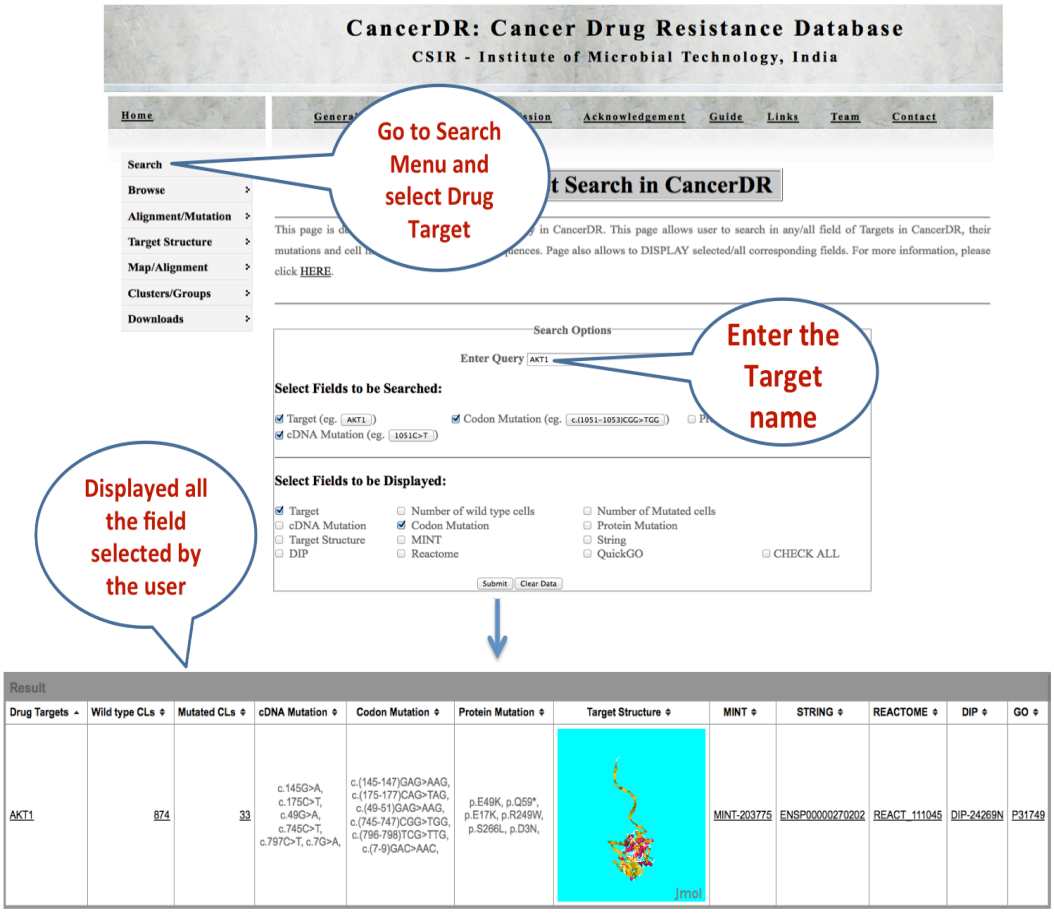


Figure 1. Drug target Search tool showing the result of searching AKT1 as drug target.

1.2 Cell Lines: User can use this tool to find out the information about any of the cell line available in CancerDR. Type the cell line name in input box (e.g. A498) and click submit button. It will return the information like tissue type and number of drugs tried against

that cell line with their respective IC50 (μM) values (Figure 2).

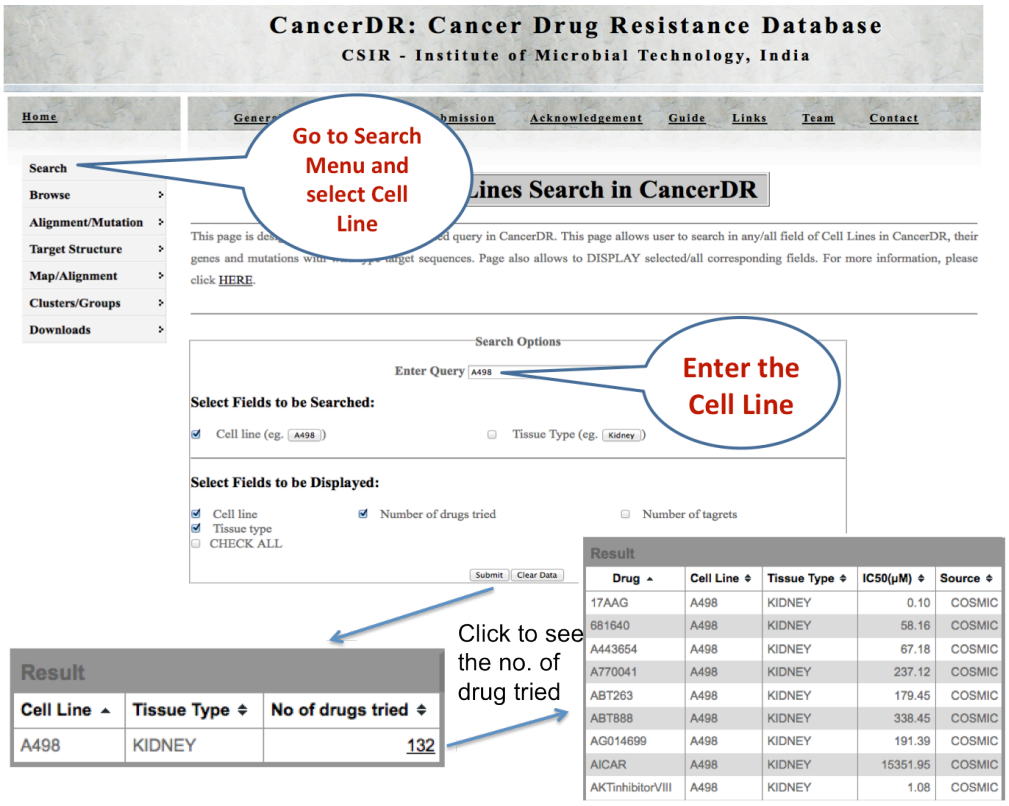


Figure 2. Cell line search tool showing the search of A498 cell line.

1.3 Drugs: 148 drugs available in the CancerDR can be explored by the drugs search tool. Type the drug name user wish to search e.g. ABT888 and click submit button after checking the box user wants to display. This tool can obtain information like drug target(s), H-bond donor, H-bond acceptor, molecular weight and structure of the drug (Figure 3).

1.4 Structure: If user wishes to search some drug on the basis of structure or wants to find some drug similar to a given molecule, this tool can help in that case. User either has to draw the structure by using JME editor or can upload a MOL/SDF/SMILE file of test molecule and click submit button. It will return all the molecules from CancerDR, which has same structure like test molecule (Figure 4).

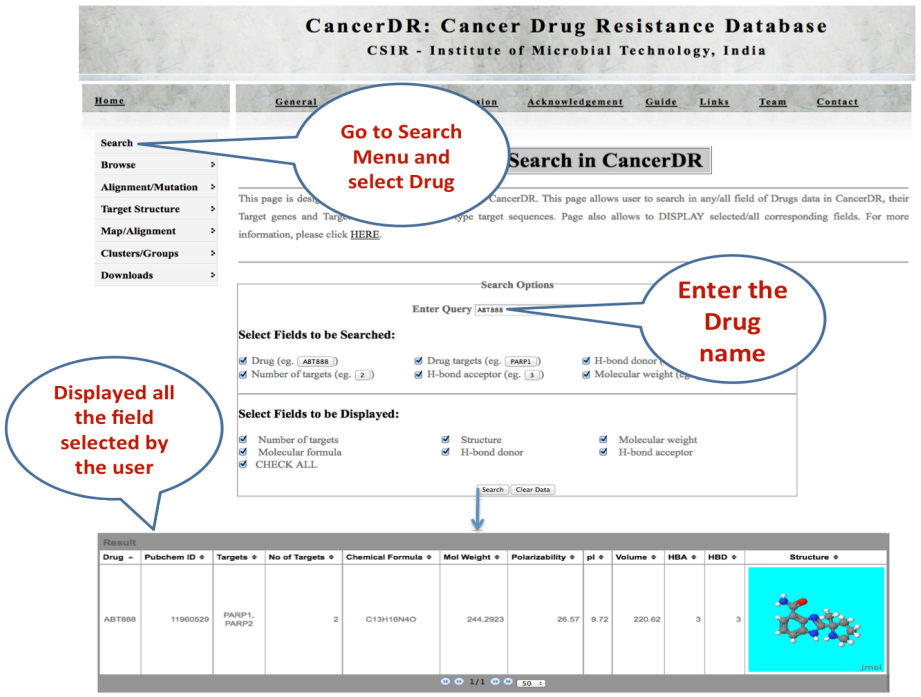


Figure 3. Drug search tool showing the search of ABT888 anticancer drug.

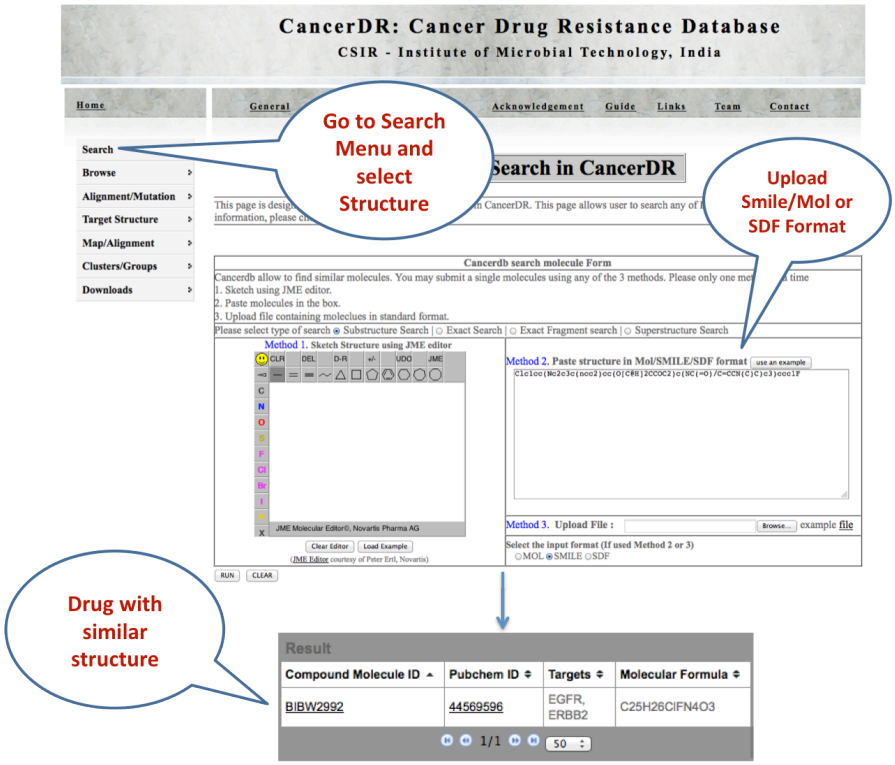


Figure 4. Structure search tool for similarity search of drugs.

2. Browse Tool

Browse tool divided into four parts

2.1 Major Fields: This tool is integrated to explore the database on the basis of fields like tissue type, target class and mutated targets. User can know how many cell lines are belonging to which particular tissue type and how many drugs are tried on each of these cell lines with their respective IC50 (μM) values (Figure 5a).

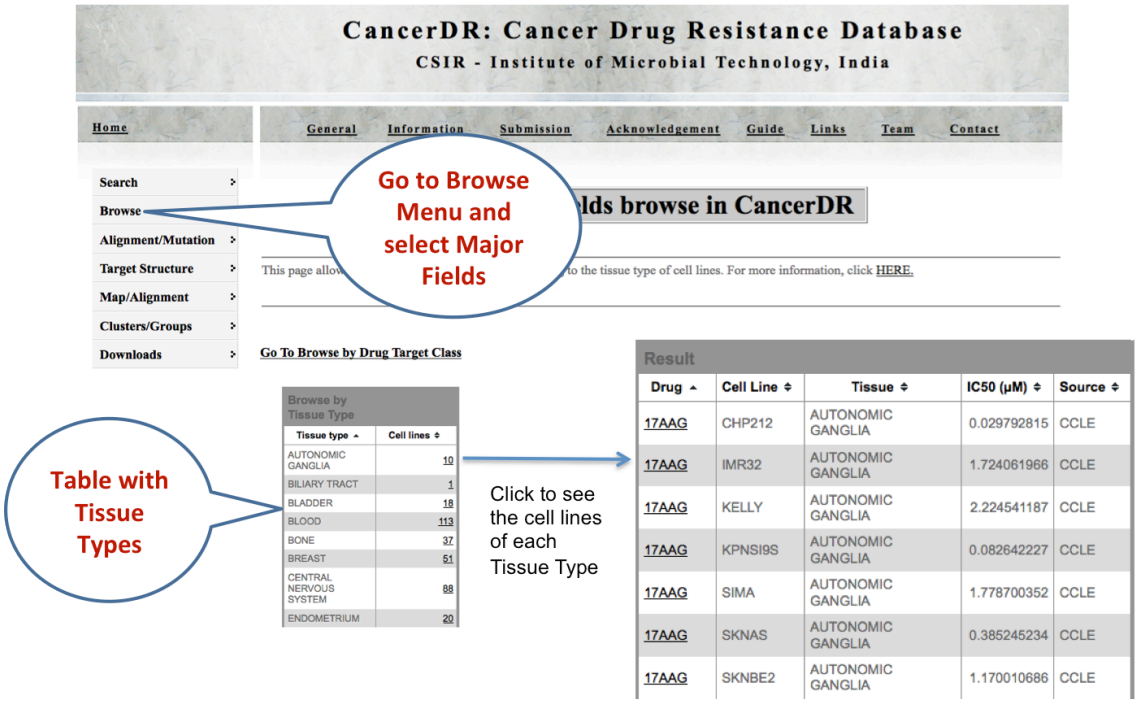


Figure 5a. Tissue types browse option in CancerDR to browse data according to the tissue types.

User can know how many drugs hitting which target or pathway and on how many cell lines that drug was tried with their respective IC50 (μM) values (Figure 5b). Along with this, user can know how many mutants are present for a particular target and which cell line has which mutants. Further they can also get the drug sensitivity information of a particular drug on particular cell line, which can help in relating the effect of mutations on drug sensitivity (Figure 5c).

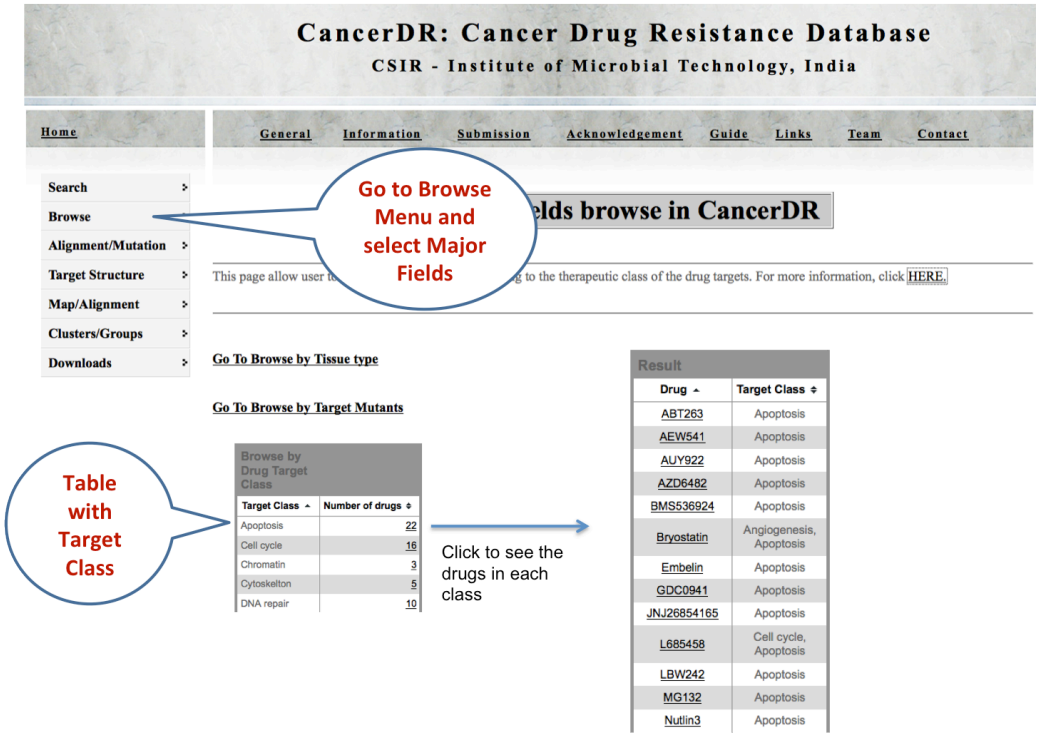


Figure 5b. Browse according to the drug target class.

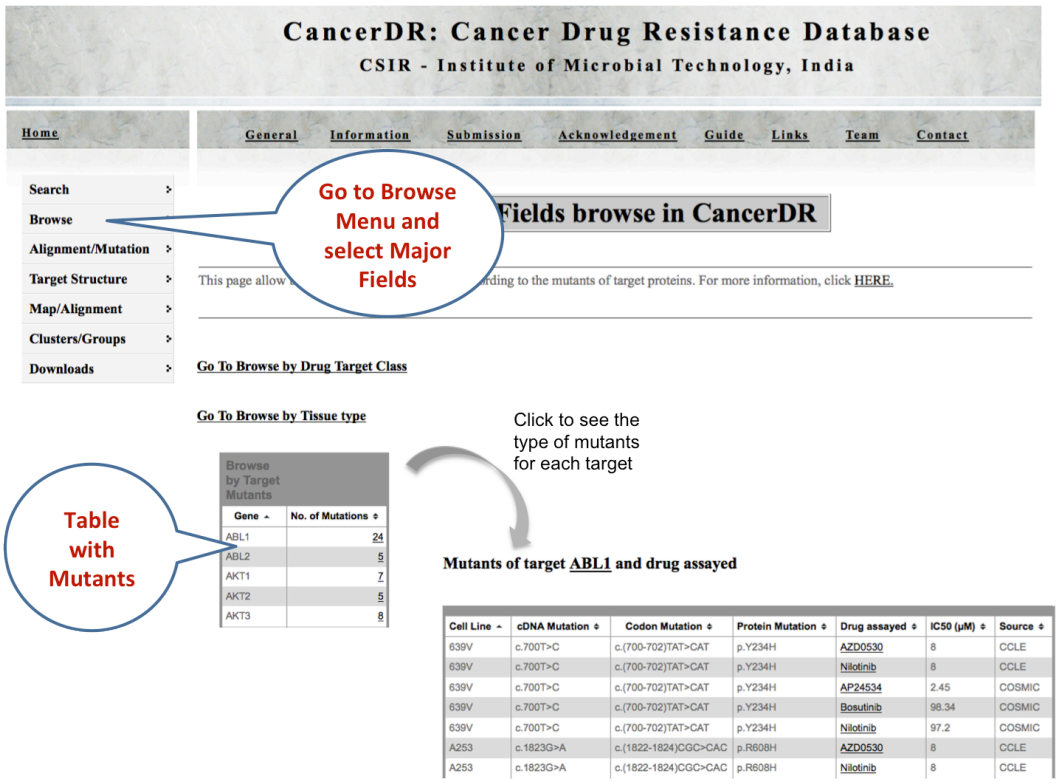


Figure 5c. Browsing on the basis of types of mutations present in drug targets.

2.2 Drug Targets: Drug target browsing facilitates the user to get the comprehensive information of drug targets such as links for mutations, protein-protein interactions, pathway interactions, gene ontologies, genome browser, phylogenetic tree, chromosomal position and number of cell lines with mutated drug targets. All the databases are linked to a single window for gathering fast information about a drug target (Figure 6).

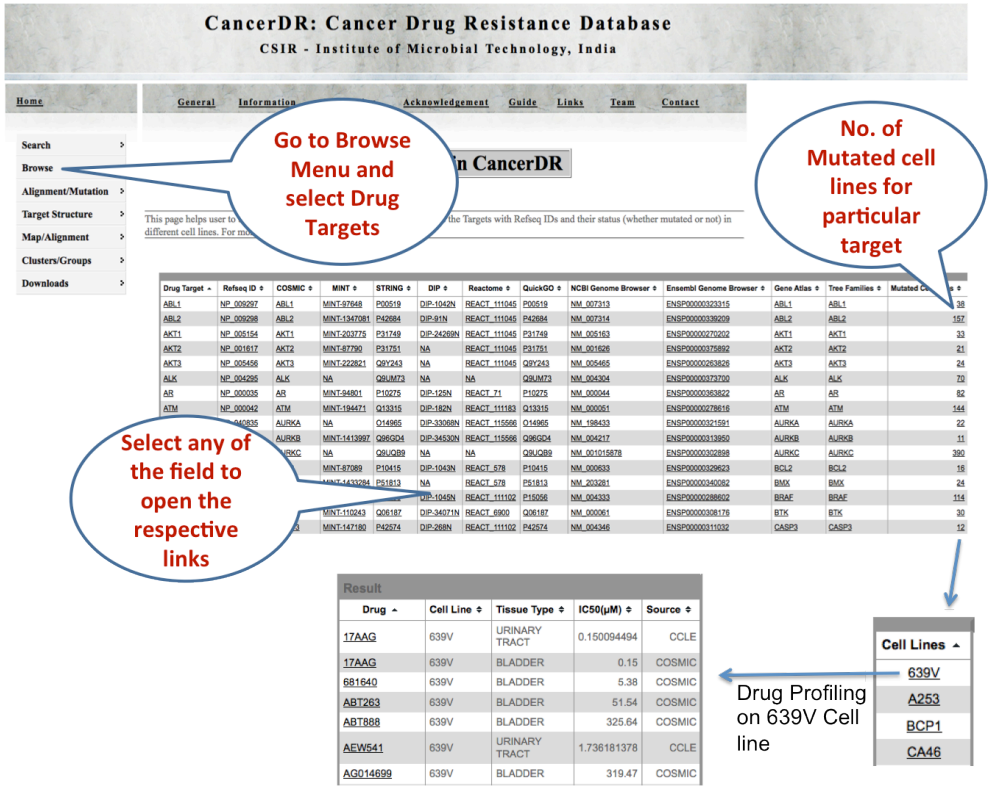


Figure 6. Screen shots showing the usage of drug targets tool in browse menu.

2.3 Cell Line: Here user can browse according to cell lines. This browsing option enables the user to look at the number of drugs and their targets along with the information regarding the histology of the cell lines (Figure 7).

2.4 Drugs: This tool enables the user to get as much as information about the drugs on a single window with their structures. With this option, the user can look at different

descriptors of drugs like pI, Volume, Hydrogen Bond Acceptor, Hydrogen Bond Donor, Structure link, molecular target, along with other chemical properties and structures of drugs (Figure 8).

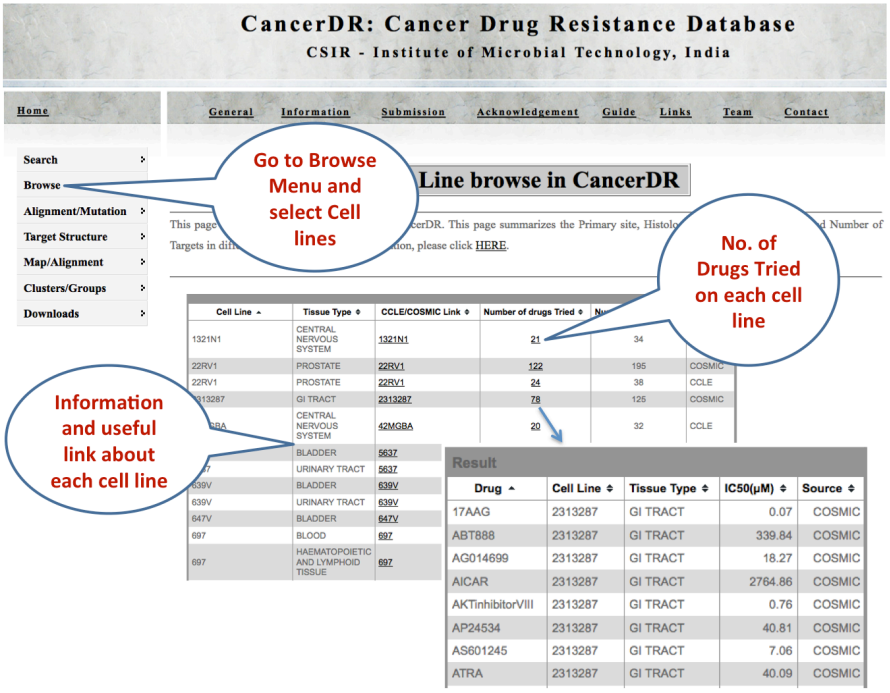


Figure 7. Browse cell lines option.

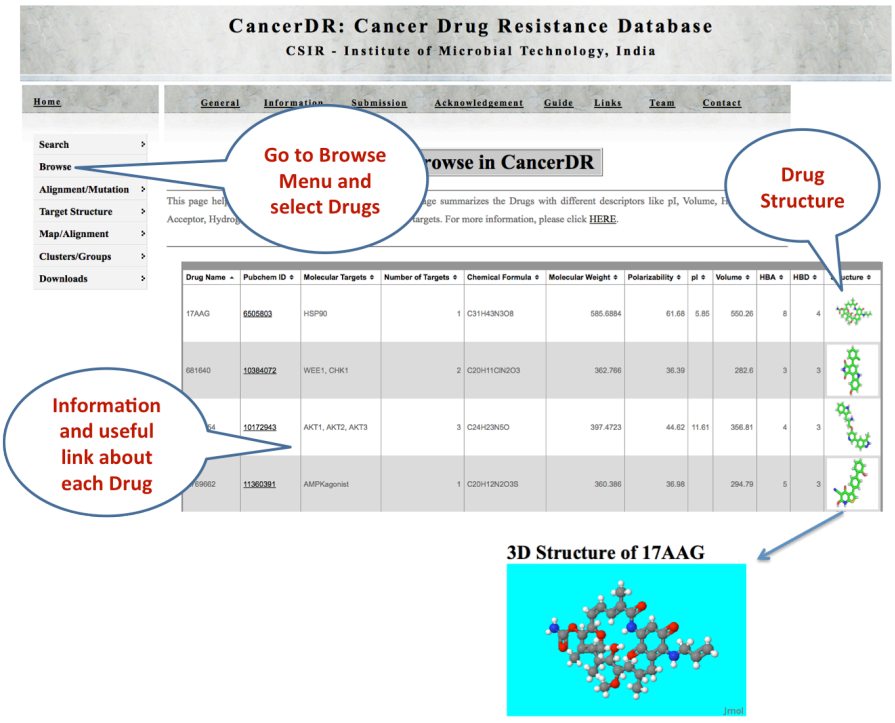


Figure 8. Drug Browsing to explore the information about drugs.

3. Alignment/Mutation

Multiple sequence alignment of variants of drug targets allows users to identify acceptable mutations. Similarly, multiple sequence alignment in mutants of drug targets facilitates in identification of mutations that affect sensitivity of drug.

3.1 Total align: This page allows users to view multiple sequence alignment of mutants and variants of targets. It helps user to distinguish the variable regions in the sequences with variations among mutants/natural variants and respective targets in CancerDR. In order to view multiple sequence alignment of target, please click on RED button for mutants and BLUE button for variants (Figure 9).

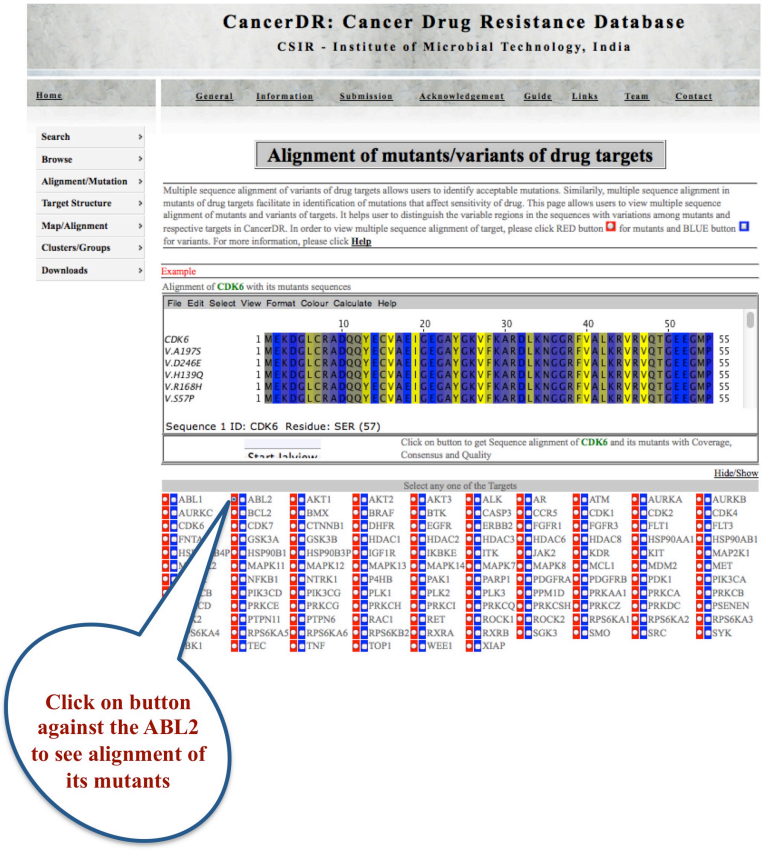


Figure 9. Multiple sequence alignment.

By clicking on any target (say ABL2; red button for its mutants), a Jalview window will

be opened below the list of the mutants of ABL2. Upper part of the main window shows the multiple sequence alignment, which can be presented by number of ways using menu bar options. Base of this window contains a button, upon clicking which opens a popup window displaying more intuitive sequence alignment with Coverage, Consensus and Quality information of alignment (Figure 10).

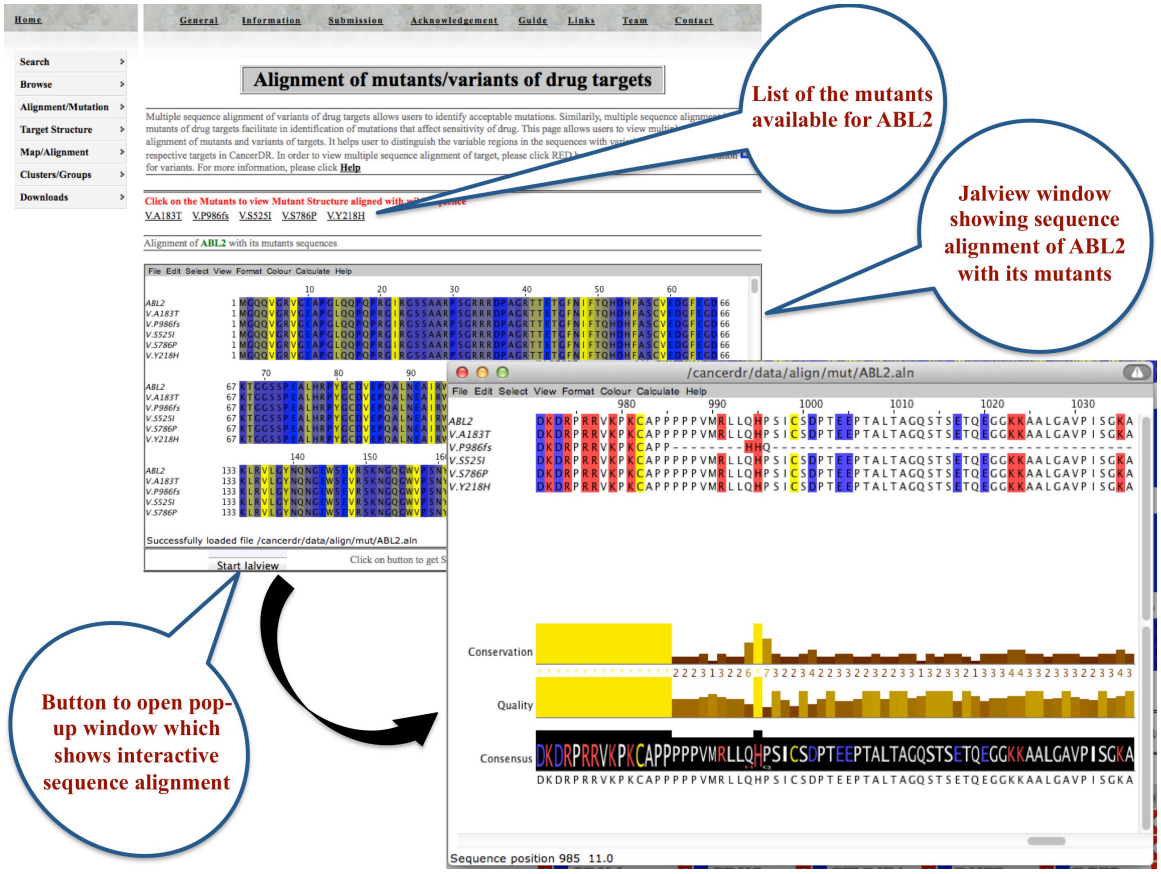


Figure 10. Multiple sequence alignment display.

Each mutant in that uppermost list is clickable which opens into next page having link of Jalview button. Upon clicking this button, two popup windows open: one showing the sequence alignment of ABL2 with its mutants (including the selected one) and second window shows the predicted structure of selected mutant for comparison with the multiple sequence alignment in the first window. Moving the cursor over the structure shows the corresponding sequence alignment portion in the other window. Jalview gives

the users different options to highlight the desired regions in the alignment through the menu bar (Figure 11).

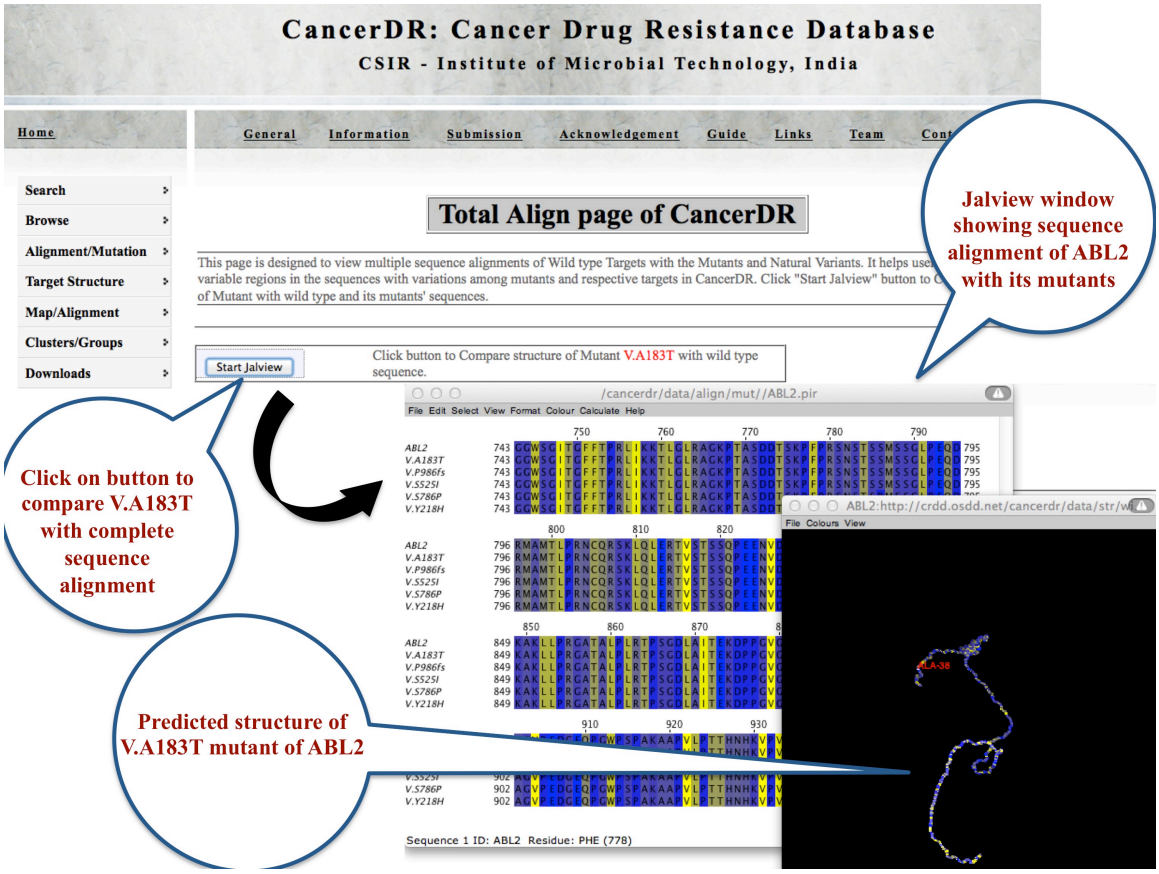


Figure 11. Multiple sequence alignment with structure of mutant.

3.2 Custom align: This module allows users to perform multiple sequence alignment of selected sequences of mutants of a drug target. This option also allows users to align their own sequence against any selected mutant's sequence of a drug target. Further, mutants are sorted based upon the increasing IC50 values for a given drug at a given time and then can be selected for the alignment (Figure 12).

For instance, by clicking on AKT2, we get three drugs, which have been tested, on these targets (A443654, AKTinhibit and MK2206). By selecting A443654 and submitting, we get three hits (*i.e.* for target AKT2 using drug A443654 three cell lines have been reported with three different kind of mutations and IC50 values). Now, we can select any two mutants or one mutant and user defined/provided sequence. After selecting

any two hits and clicking on Align button, Jalview window opens, which shows the alignment of selected mutants in the interactive way (as described in the previous section). This module is useful to select mutants based upon IC50 values and compares their sequences interactively (Figure 13, 14).

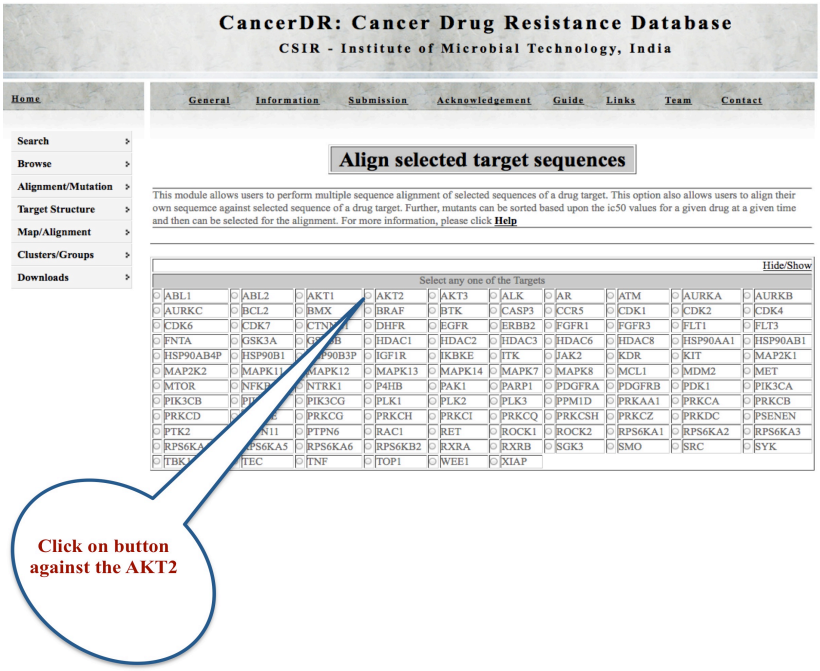


Figure 12. Customized sequence alignment.

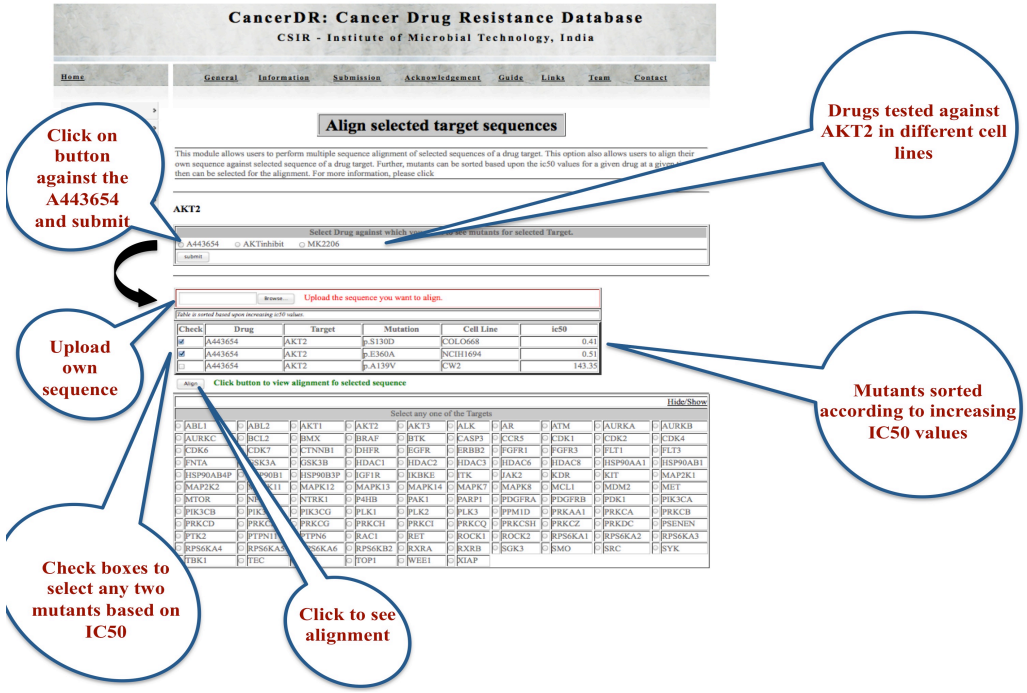


Figure 13. Customized sequence alignment.

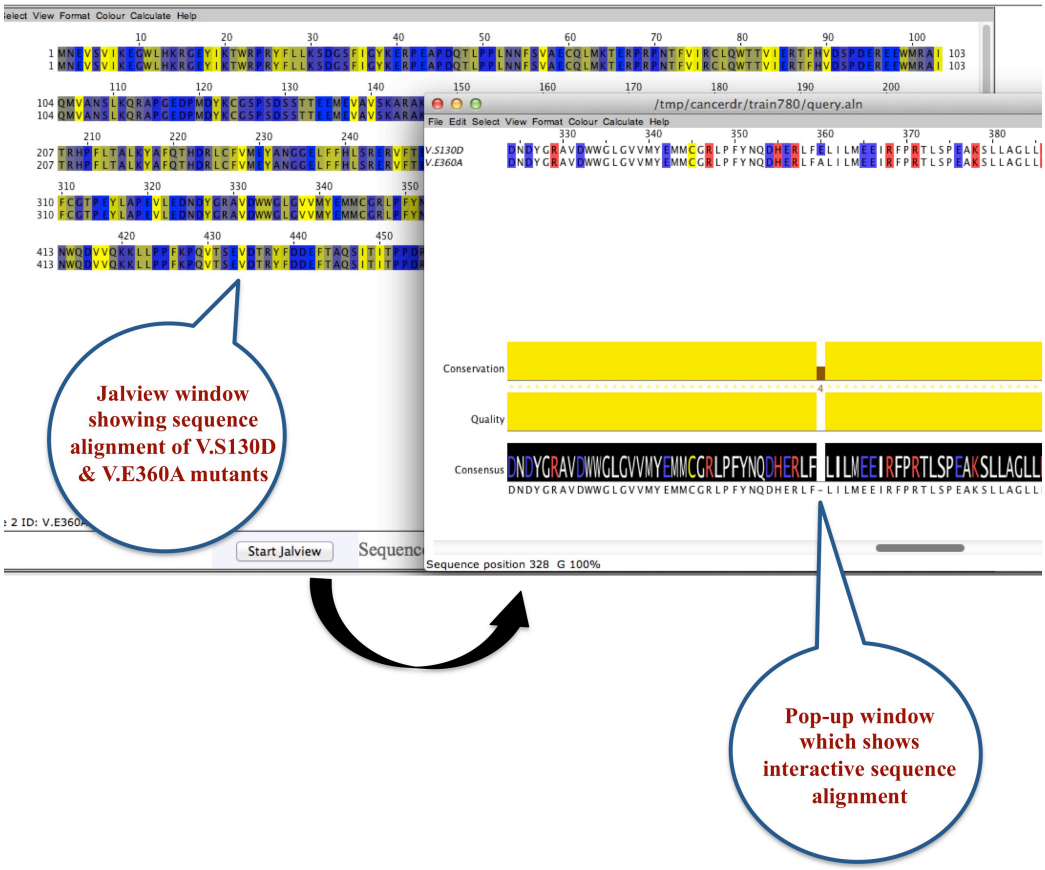


Figure 14. Customized sequence alignment.

3.3 Mutants: This page allows the user to display different types of mutations reported for a drug target at amino acid level, codon level and cDNA level (Figure 15).

3.4 Structure alignment: This module allows users to visualize structure alignment of mutants/variants of a target. User can select mutants by clicking on RED button and natural variants by clicking on BLUE button of a target. Structure alignment (using MUSTANG software) is visible in jmol applet, which is accompanied by the details of the sequences aligned to achieve that alignment. Also download option is given to download the structurally aligned file above the Jmol applet. In few cases, where extremely variable length of sequences is present, a few sequences have been removed to achieve the multiple structure alignment (Figure 16).

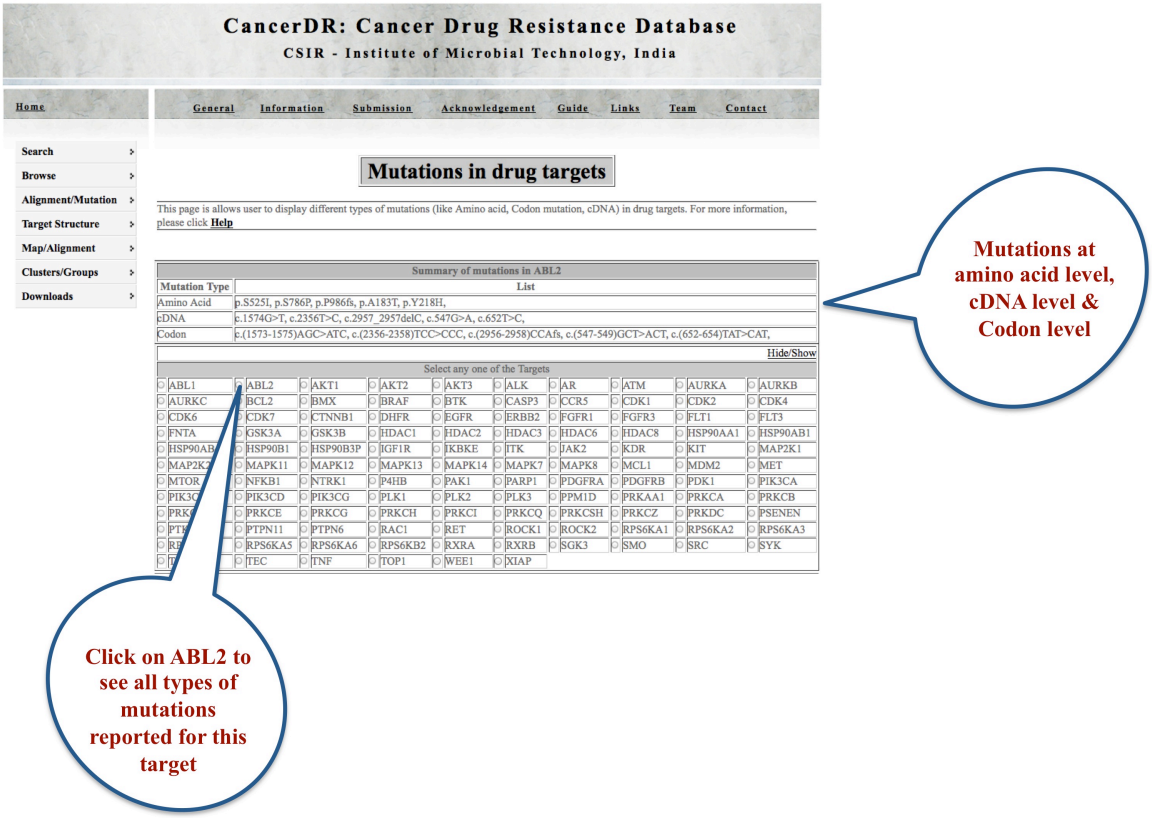


Figure 15. Mutation information for each drug target.

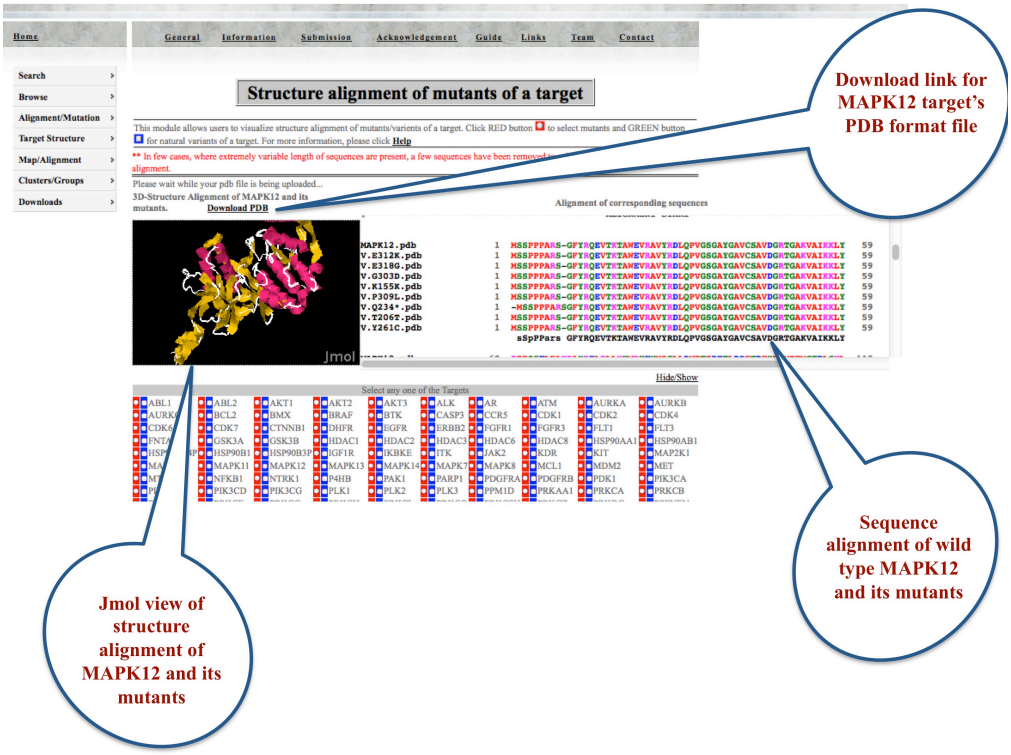


Figure 16. Structure alignment.

4. Target structure

4.1 Tertiary: Predicted tertiary structure (using HHsuite) of each drug target and their mutants are available through this module. This page is designed to view structures of wild type targets and their mutants. It helps the user to see 3D view in Jmol against selected targets in CancerDR. Also, user can download the selected structure for analysis (Figure 17a,b). This page also shows the reliability score of modelled structure, which include percent coverage, identity with the template and percentage of allowed region in ramachandran plot.

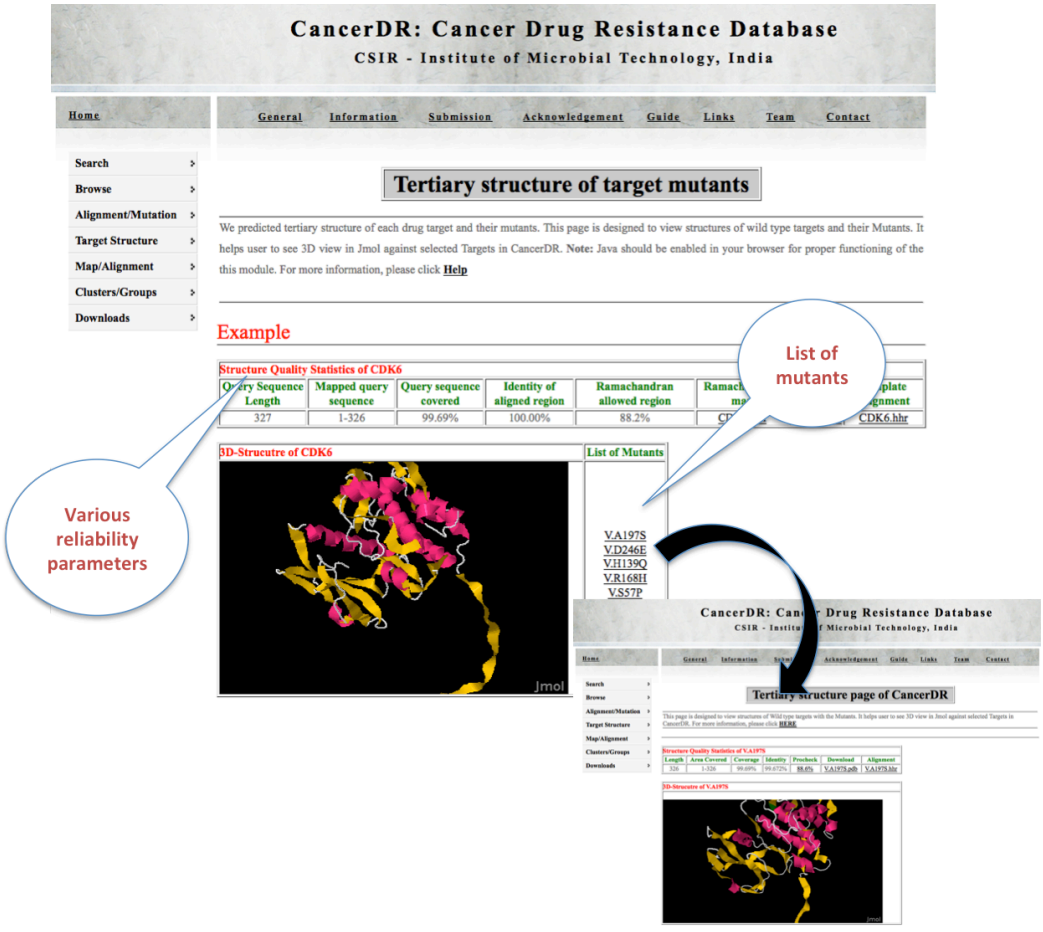


Figure 17 (a). Tertiary structure information.

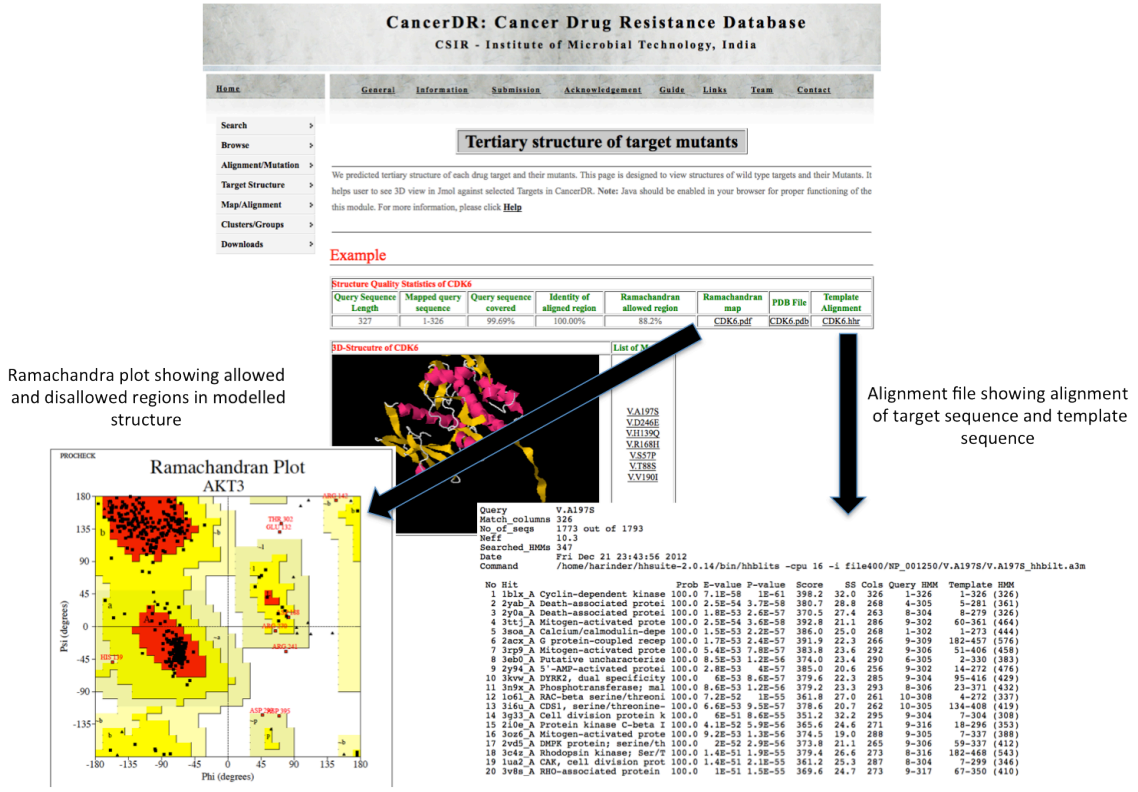


Figure 17 (b). Aignment file and ramachandran plot.

Reliability Parameters:

Mapped query sequence: This score indicates, which part of the query sequence is covered by the templates.

Query sequence covered: It is the percentage of the length of query sequence covered by the templates.

Identity of aligned region: This indicates, what is the identity of query and template in the aligned regions.

Ramachandran allowed region: This parameter tells about how much

percentage of the modeled structure falls in the allowed region of ramachandran plot.

HHR file format:

No: the index of the database match.

Hit: the first 30 characters of the name line.

Prob: the Probability of template to be a true positive. For the probability of being a true positive, the secondary structure score in column SS is taken into account, together with the raw score in column Score. True positives are de_ned to be either globally homologous or they are at least homologous in parts, and thereby locally similar in structure. More precisely, the latter criterion demands that the MAXSUB score between query and hit is at least 0.1. In almost all cases the structural similarity will we be due to a global OR LOCAL homology between query and template.

E-value: The E-value gives the average number of false positives ('wrong hits') with a score better than the one for the template when scanning the database. It is a measure of reliability: Evalues near to 0 signify a very reliable hit, an E-value of 10 means about 10 wrong hits are expected to be found in the database with a score at least this good. Note that E-value and P-value are calculated without taking the secondary structure into account!

P-value: The P-value is the E-value divided by the number of sequences in the database. It is the probability that in a pairwise comparison a wrong hit will score at least this good.

Score: the raw score is what comes out of the (Viterbi) HMM-HMM alignment excluding the secondary structure score. Informally speaking, it is the sum over the similarities of aligned pro_le colmuns minus the gap penalties.

SS: the secondary structure score. This score tells you how well the PSIPRED-predicted (3-state) or actual DSSP-determined (8-state) secondary structure sequences agree with each other. PSIPRED con_dence values are used in the scoring, low con_dences getting less statistical weight.

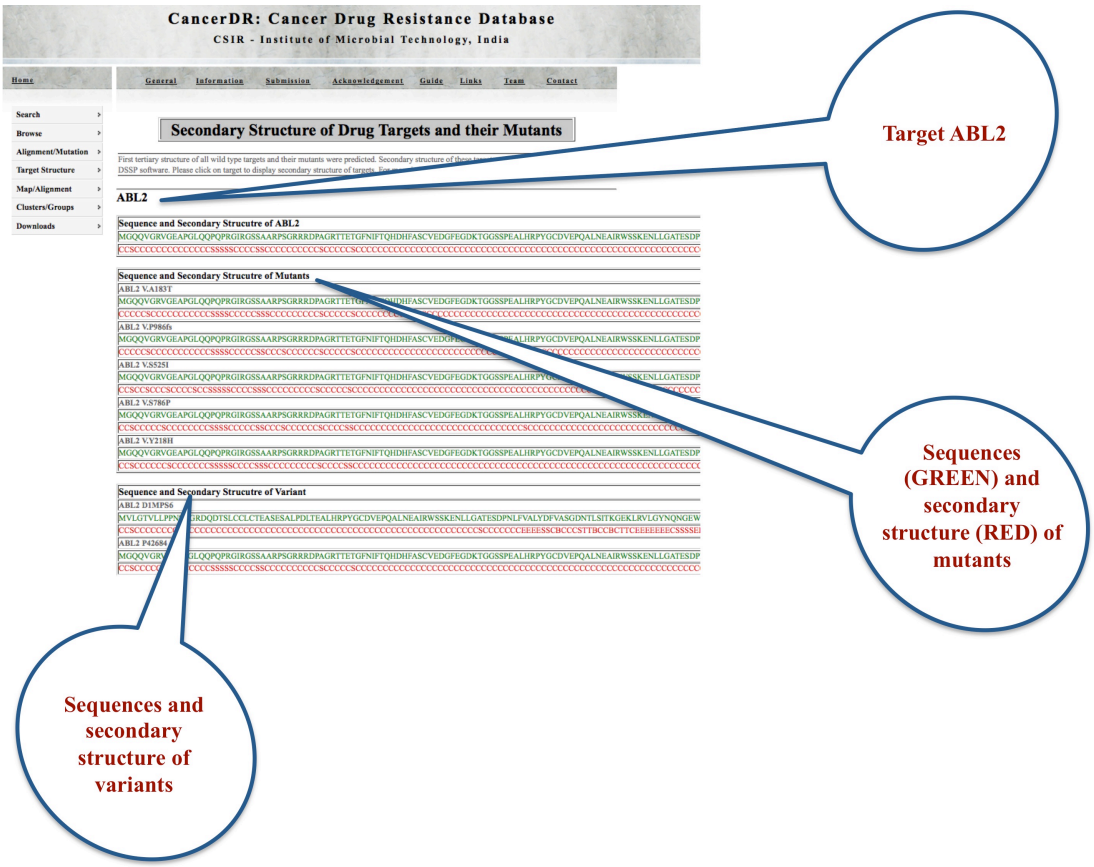
Cols: the number of aligned Match columns in the HMM-HMM alignment.

Query HMM: the range of aligned match states from query HMM.

Template HMM: the range of aligned match states from the database/template HMM and, in parenthesis, the number of match states in the database HMM.

For further information, please visit http://toolkit.tuebingen.mpg.de/hhblits.

4.2 Secondary: First tertiary structure of all wild type targets and their mutants were predicted. Then secondary structure of these targets and mutants were assigned using



DSSP software. By clicking on the target, sequence and DSSP state of wild type target is followed by its mutants and variants (Figure 18).

Figure 18. Secondary structure information.

4.3 Compare: This page is designed to compare the structures of two mutants of a target. User can also compare his/her own structure against any mutant of a selected target in CancerDR. For target ABL2, there are 5 mutants. So for comparing structures of any two mutants, user can select those and provide his/her valid email. Results are forwarded to the user when the job is completed (Figure 19).

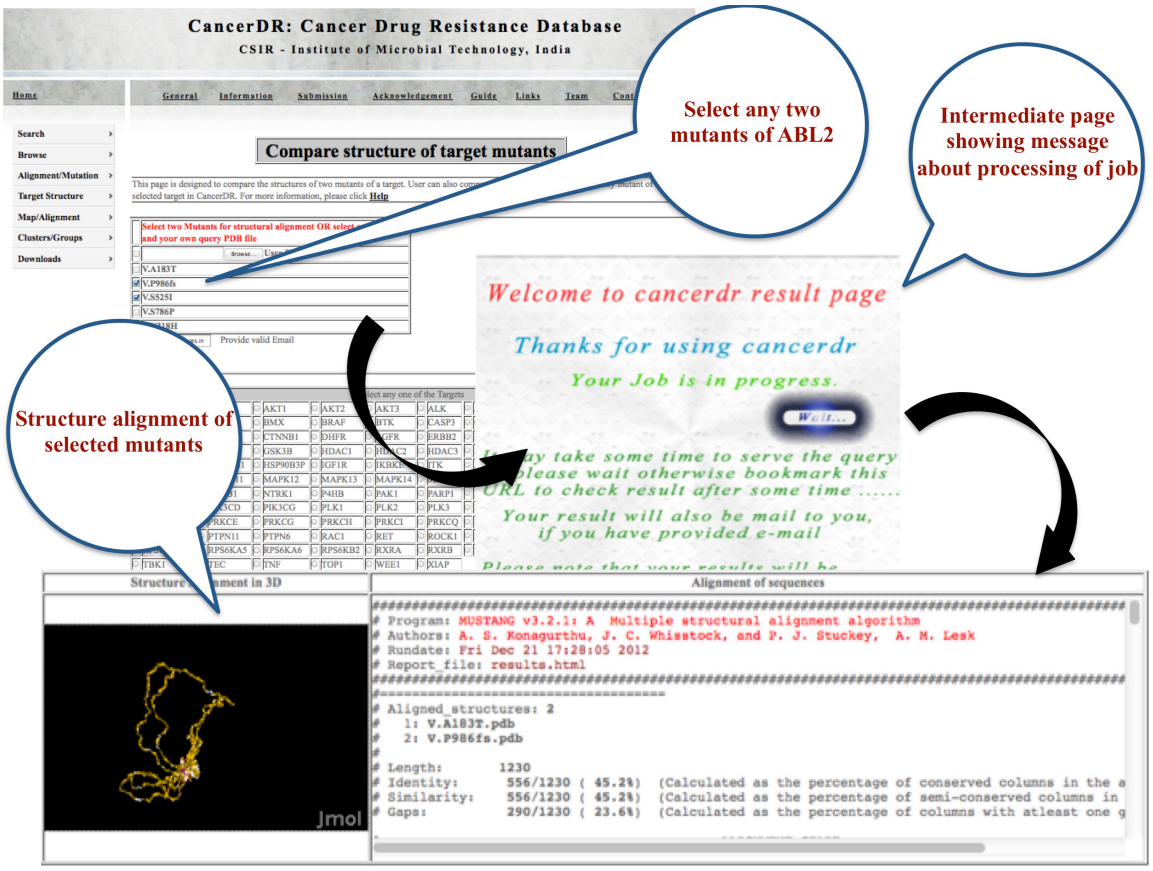


Figure 19. Structure alignment of selected mutants selected by the user.

4.4 User sequence: This module allows users to predict the structure of their sequence using homology modeling (using MODELLER). User has to either paste the query sequence in FASTA format or upload the corresponding file containing query sequence and select any one of the targets against which structure has to be built. The selected target serves as a template for the structure prediction (Figure20).

4.5 Structures in PDB: We have also collected available PDB structures available for all the targets. There are multiple structures available for every target. These can be accessed via the PDB ID hyperlink for that target (Figure 21).

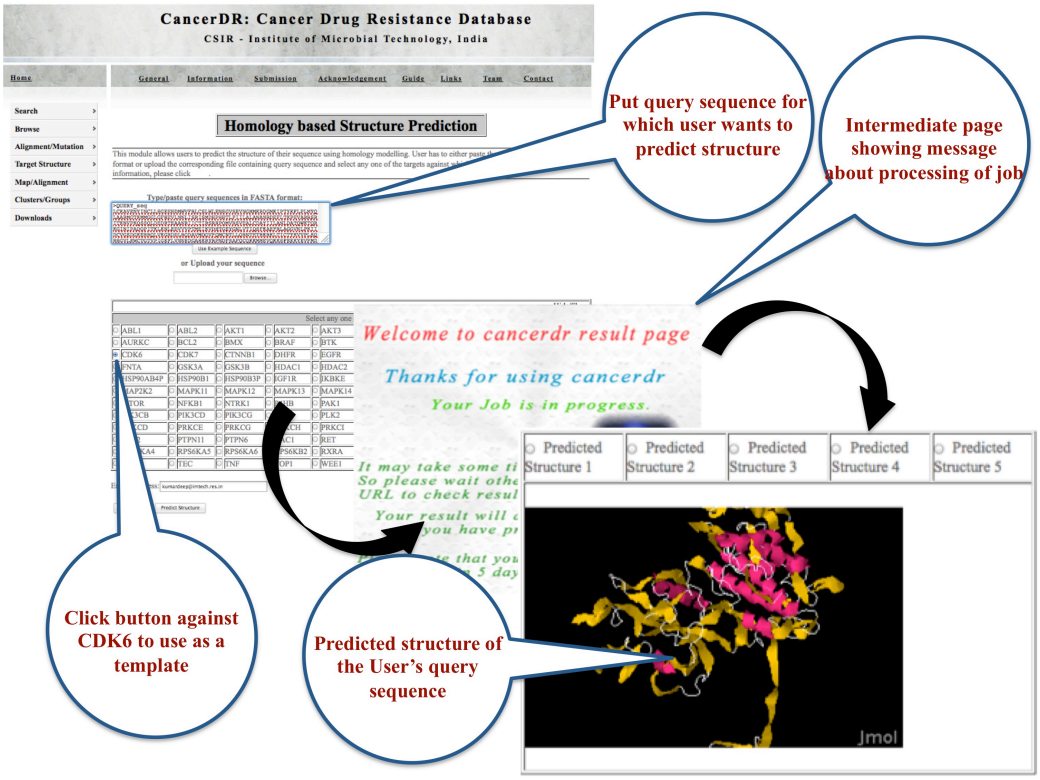


Figure 20. Tertiary structure prediction of user sequence.

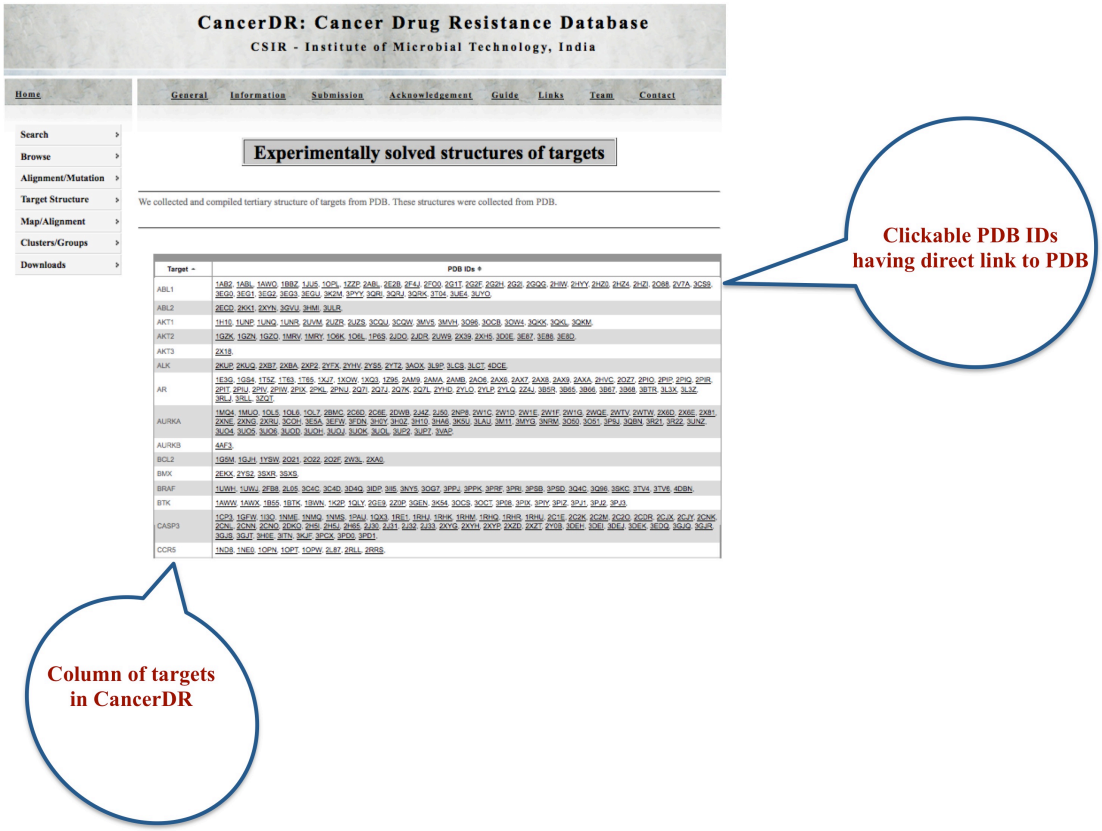


Figure 21. Experimentally validated structure information (PDB structures).

5. Map/Alignment

Due to advancement in Next Generation Sequencing (NGS) technologies whole genome, transcriptome and exome sequencing have been frequently used to find out the mutations in the cancer samples. The data produced by the NGS technologies could be in form of short reads; contigs (if short reads were assembled) or genes, predicted from assembled contigs (or genomic fragments). Map/Alignment sections of our CancerDR allow the user to align Short reads, contigs and genes to cancer targets mentioned in this database. We have three separate modules for all three purposes.

5.1 Short Reads: This module allows the user to align and visualize short reads (i.e. Illumina reads) to cancer targets. Any mutation, at a particular position in the cancer sequencing sample can be easily visualized. User can provide single end reads (i.e. .fastq files) or paired end reads (i.e. .fastq file) for the alignment. Short reads file should be in .fastq format. For the paired end reads alignment two separate files of forward read and reversed reads are necessary. Single end reads file should be provided in third box (Figure 22).

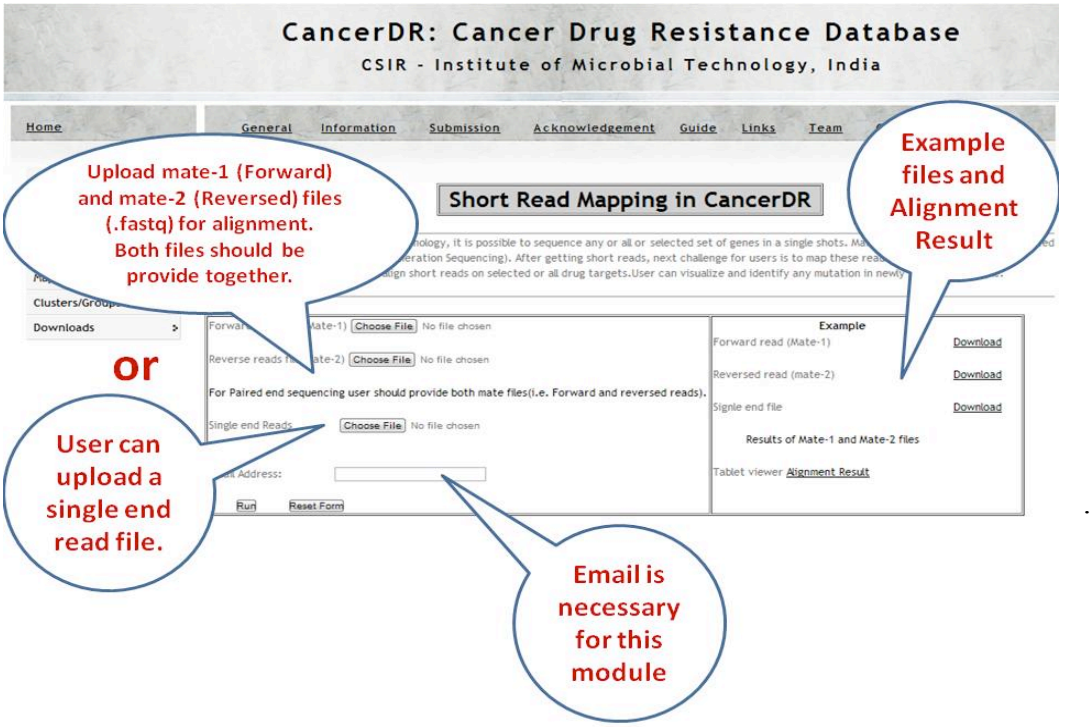


Figure 22. Submission page for short reads alignment to cancer targets.

User can check the example file results by clicking **Alignment Results** link in the example section of the page. User must provide a valid Email address in this section so that results link can be send though email in case short reads alignment taking time (Figure 23).

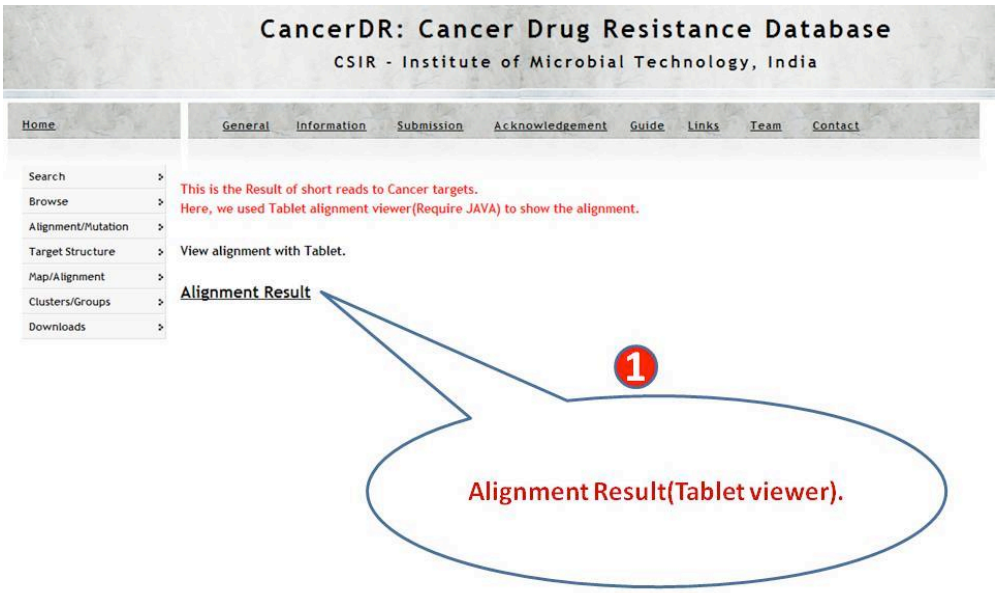


Figure 23. Alignment result link, appear after submission of short reads.

After submitting the short reads files, Alignment results will be available after some time (depends upon the amount of data and load on our server) (Figure 24). User should have JAVA to view the alignment results.

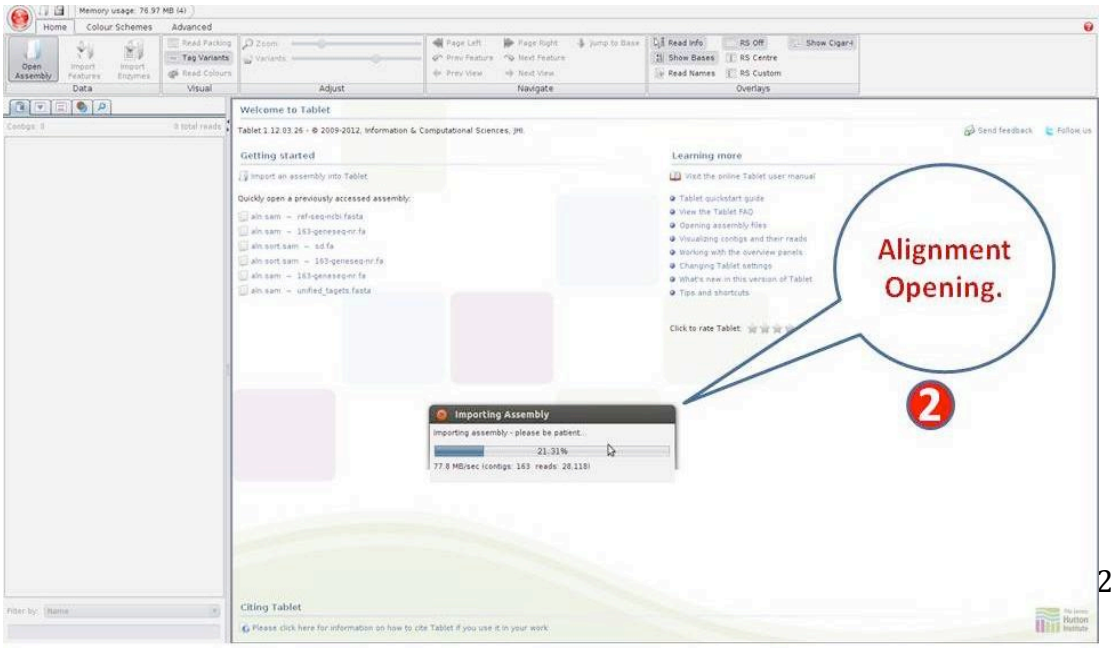


Figure 24. Processing of alignment results in Tablet viewer.

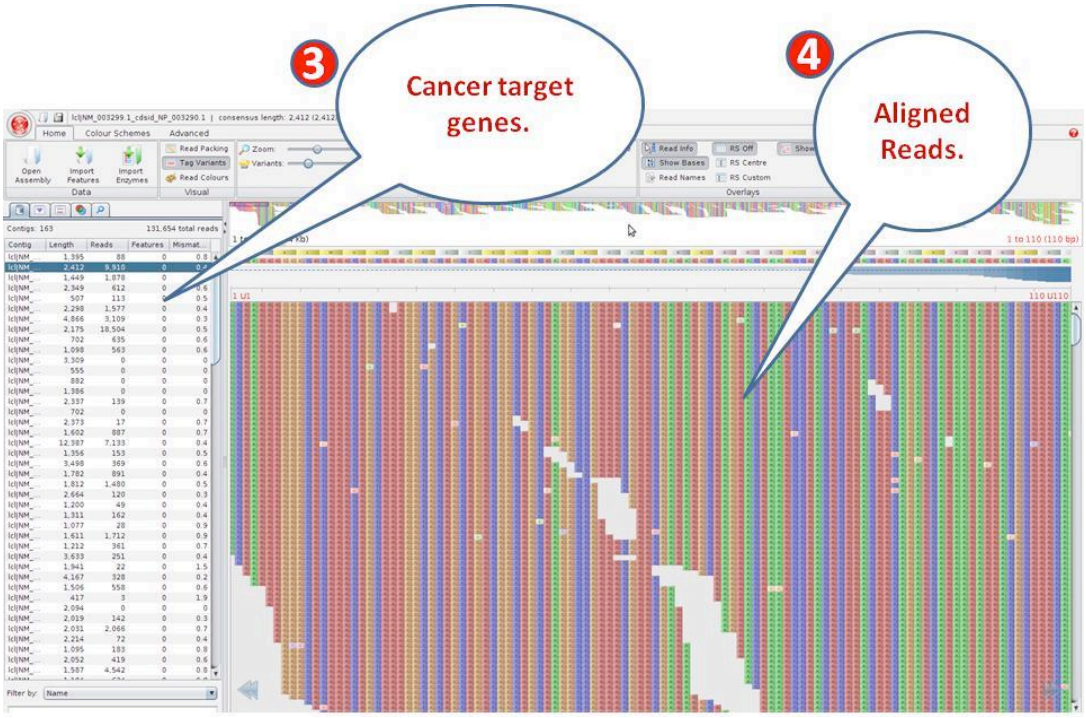


Figure 25. Alignment of short reads to cancer targets.

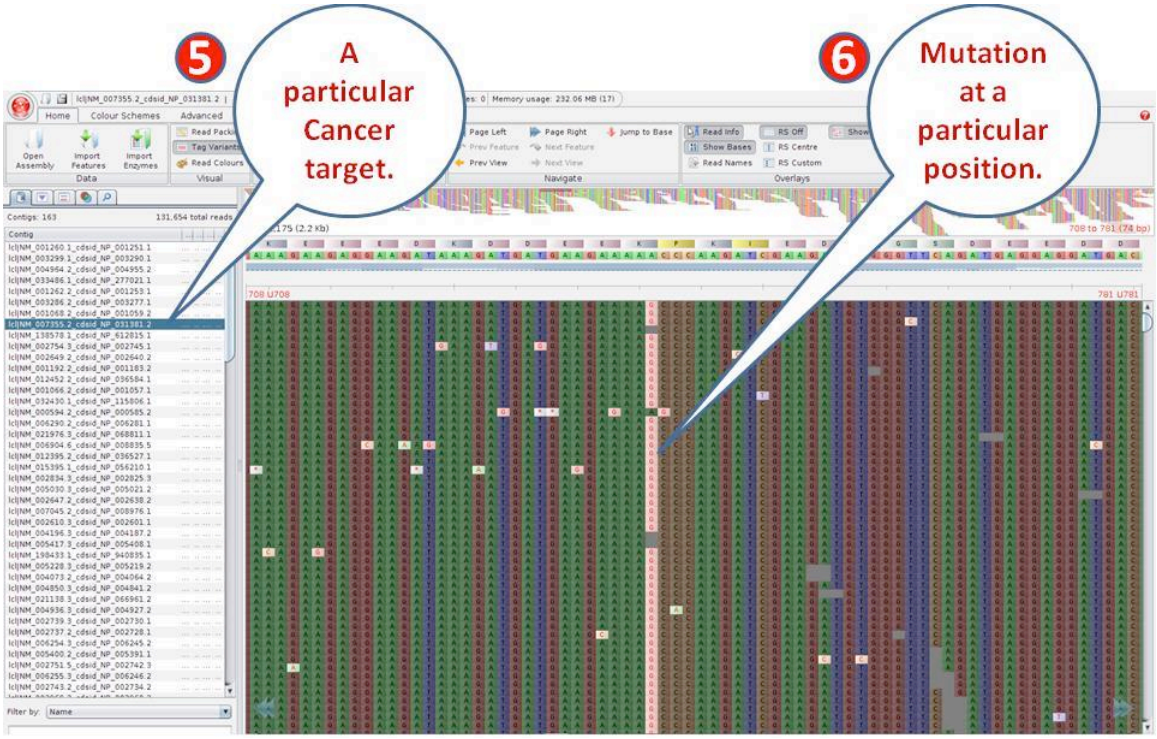


Figure 26. Visualization of a mutation in cancer sequence sample for a particular gene.

Tablet viewer window will start opening the alignment files in the next step (Figure 25). Alignment results will provide the information about number of reads align to each cancer target (i.e. gene). User can select a particular gene and check for mutation present at any position in the sequenced cancer sample (Figure 26).

Please contact Dr. G P S Raghava (raghava@imtech.res.in) in case you want to align bulk amount of data to Cancer targets, we will provide ftp service for the purpose.

5.2 Contigs: This module is prepared for the alignment of genes (both nucleotide and protein sequences), predicted from user provided contigs. At this module, user can provide contigs (or genomic fragments) of cancer sequenced samples and find out any change in the sequence by aligning to our cancer targets (i.e. wild type genes) by BLAST (Figure 27).

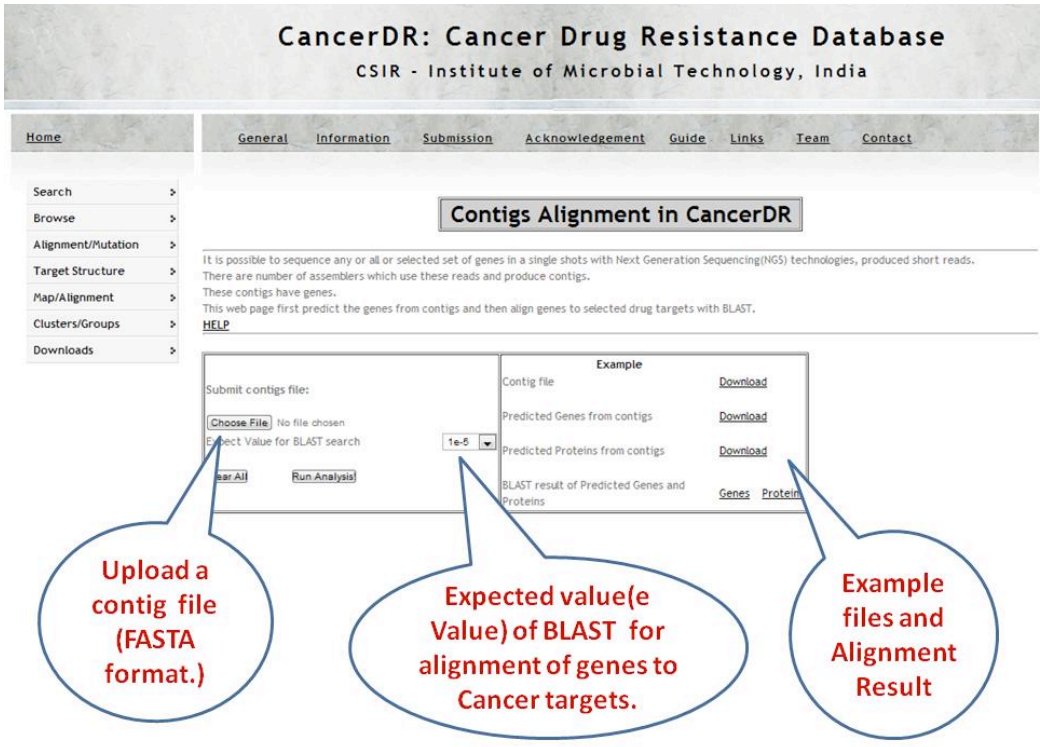


Figure 27. Contigs (or genomic fragments) submission page.

Nucleotide sequenced (contigs of genomic fragments) should be in fasta file format. User can select variable E value cutoff for the alignment. This module works in two steps. First, genes have been predicted from contigs with the help of Augustus software. Second, predicted genes have been aligned to cancer targets with BLAST.

BLAST alignment results can be downloaded and checked for any variation present in the genes predicted from cancer genome fragments at nucleotide and protein level both.

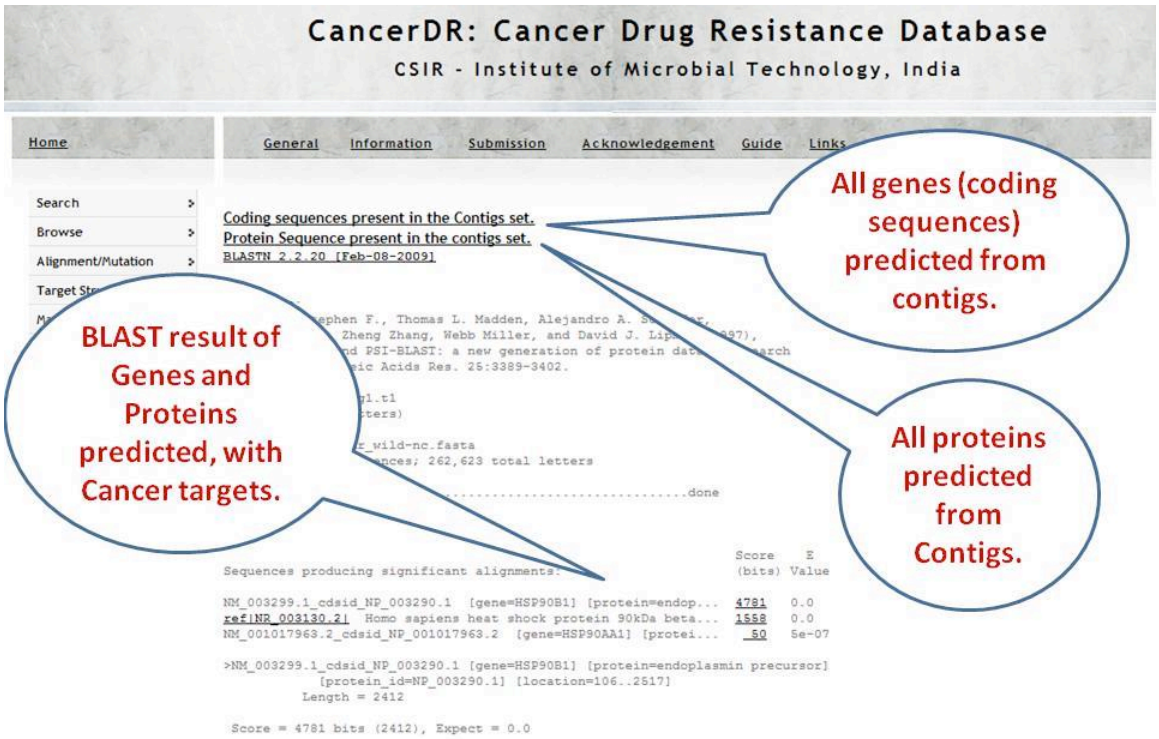


Figure 28. Alignment (i.e. BLAST) results page of genes predicted from contigs.

5.3 Genes: This module has been prepared for the alignment of gene (both nucleotide and amino acid) sequences with cancer target at this database. This strategy provides the idea about any variation in the sequenced cancer genes with respect to wild type sequences of same gene present at our database. These web pages have the option for both single and multiple fasta sequences with multiple E value options for BLAST alignment. Nucleotide and protein sequences both can be provided at this module to align with cancer targets. Example section provides the idea about input and output files used at this module.

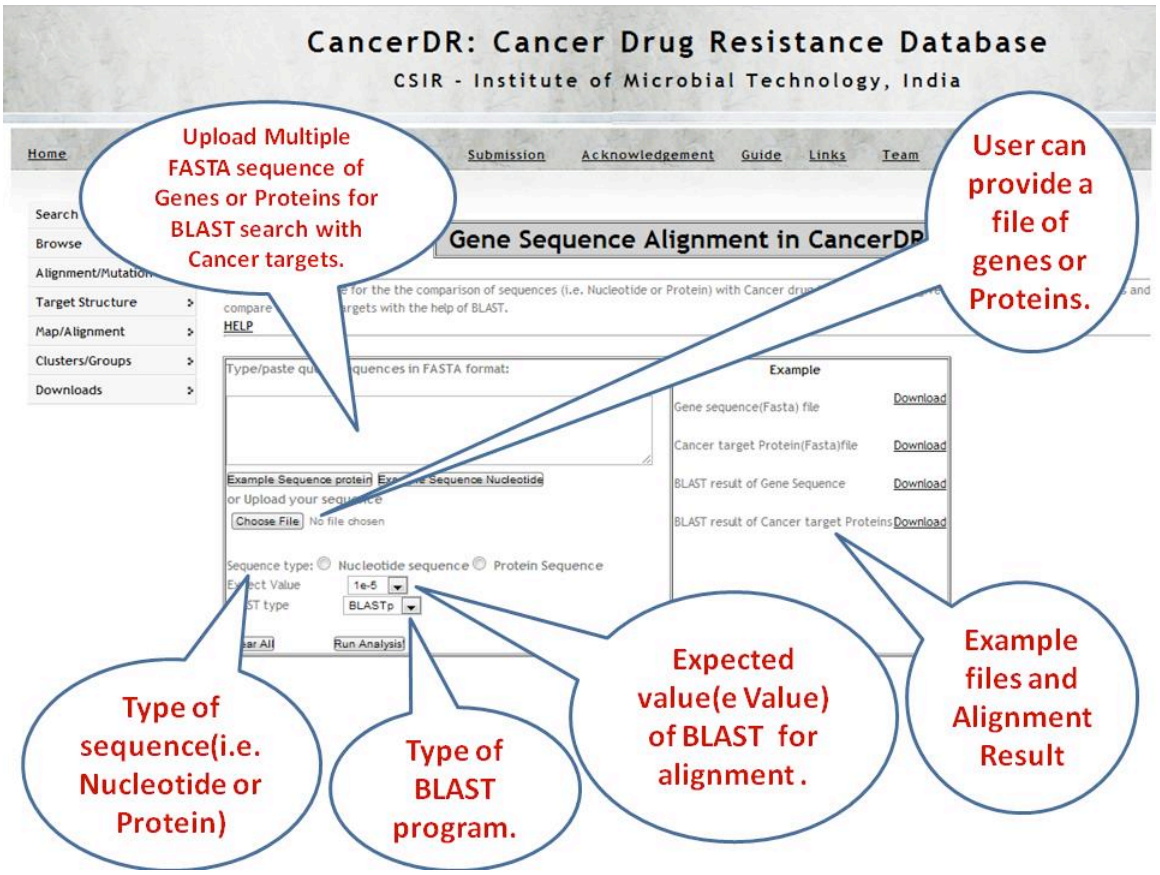


Figure 29. Genes (i.e. Nucleotide or proteiins) sequence submission page.

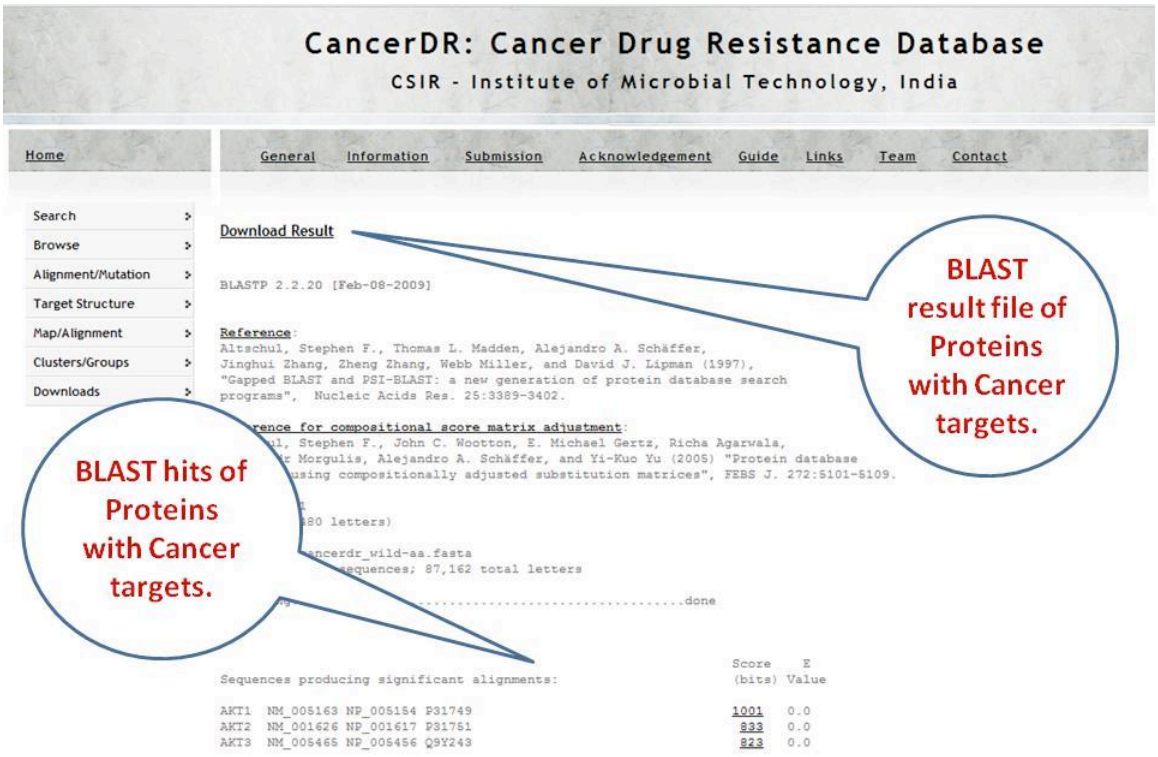


Figure 30. Alignment (i.e. BLAST) results page of genes.

6. Clustering Tool

Clustering tool is subdivided into three parts:

6.1 Clustering of sequences in a drug target: We have provided this tool to look at the distance of different mutated sequences of a selected target as an alignment tree. Here user can select a drug target from the given list. The Alignment-tree diagram will be displayed creating the tree of all possible mutated protein sequences for that selected drug target. There are two different selection options (Figure 31a) for drug targets, i) Red button- for mutants of given drug target, ii) Blue button for natural variants of the query drug target. Clicking red or blue buttons displays name of query with 'Start Jalview' button in middle table, which on submission/starting, displays windows for Clustal-W alignment and alignment tree (Figure 31b). Finally the user can look at the distance of different mutant sequences in the alignment against selected target.

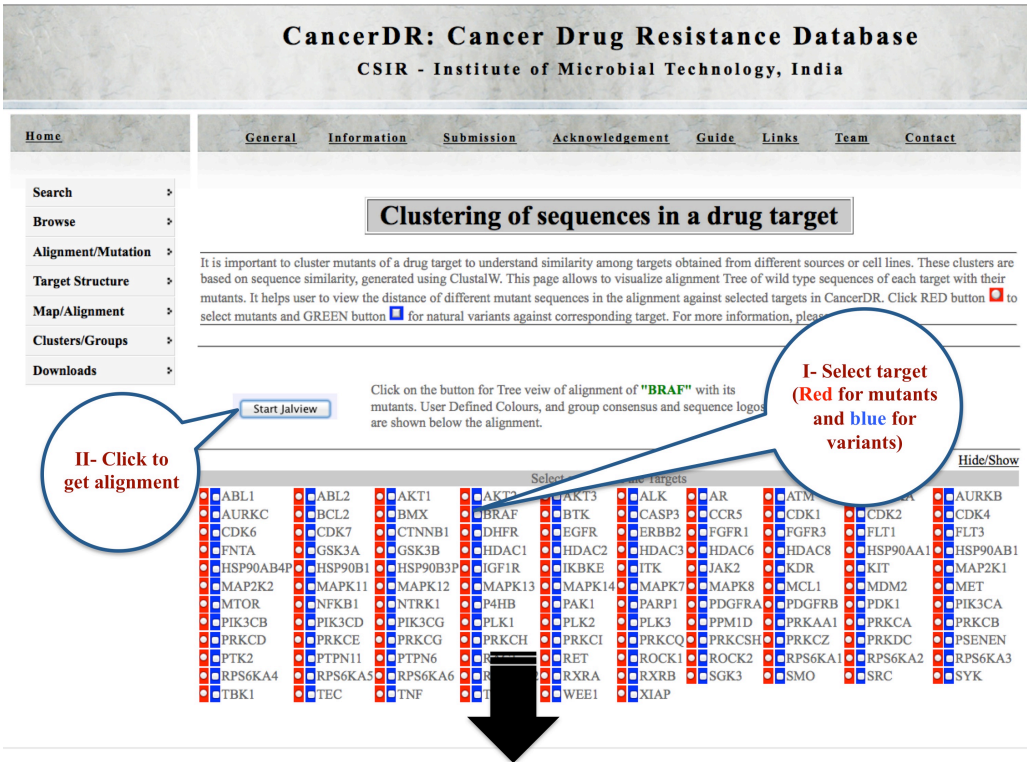


Figure 31a. Tool for the drug target clustering on the basis of their sequences.

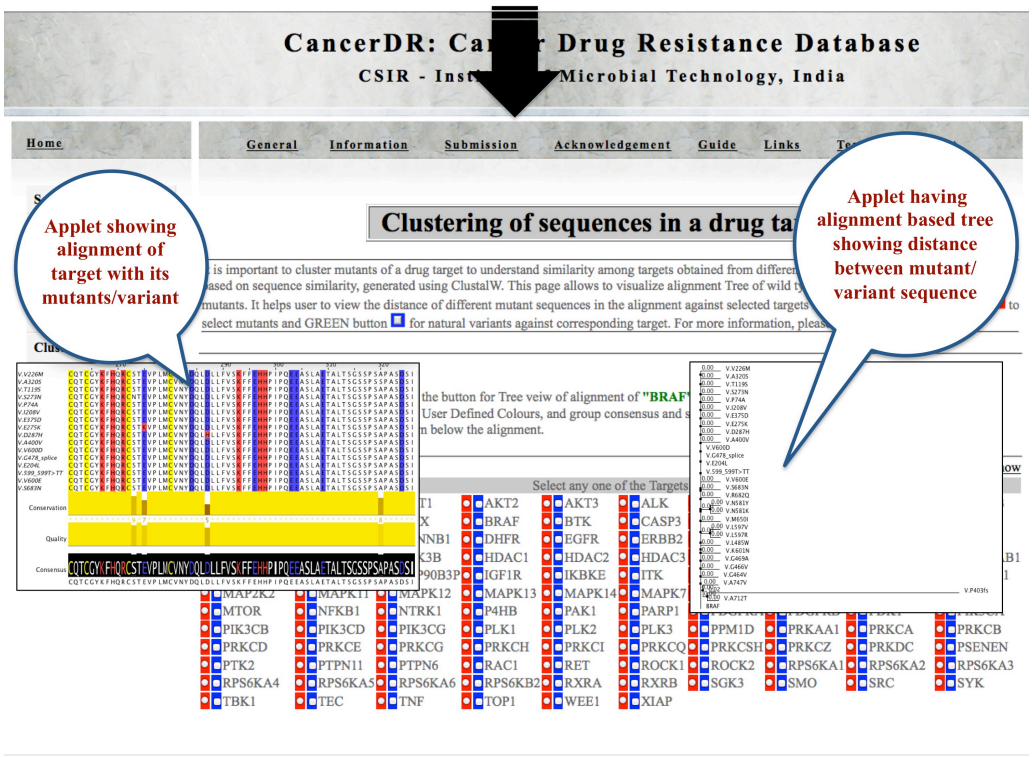


Figure 31b. Jalview showing the clustered drug target sequences.

6.2 Clustering of Cell lines: This page allows user to cluster cell lines based on their sensitivity towards a drug. We group cell lines for each drug based on their relative sensitivity to reference (highest sensitivity) or inhibition IC50 value in a given range. User can either choose tissue type or all cell lines (Figure 32a) or can select the cell lines with mutated drug targets (Figure 33a). The second column gives the number of cell lines screened against the drug as it is equal to or greater than reference IC50 (lowest of all). Subsequent columns, up to eight, give grouping based on IC50 value equal to or greater than 3, 5, 25, 100 and 250 times of reference IC50. The columns from nine to sixteen show the group of cell lines having absolute value of IC50 lying in different ranges (R1, R2, R3, R4, R5, R6, R7 and R8) (Figure 32b & c).

In addition, 'MUTANTS' button in the tissue list leads to a table on the next page (Figure 33a), which has similar ranges of IC50 along with links to bar plots at the last column. The bar plots show the distribution of IC50 values along different mutants of the drug-target coming from various cell line (Figure 33b).

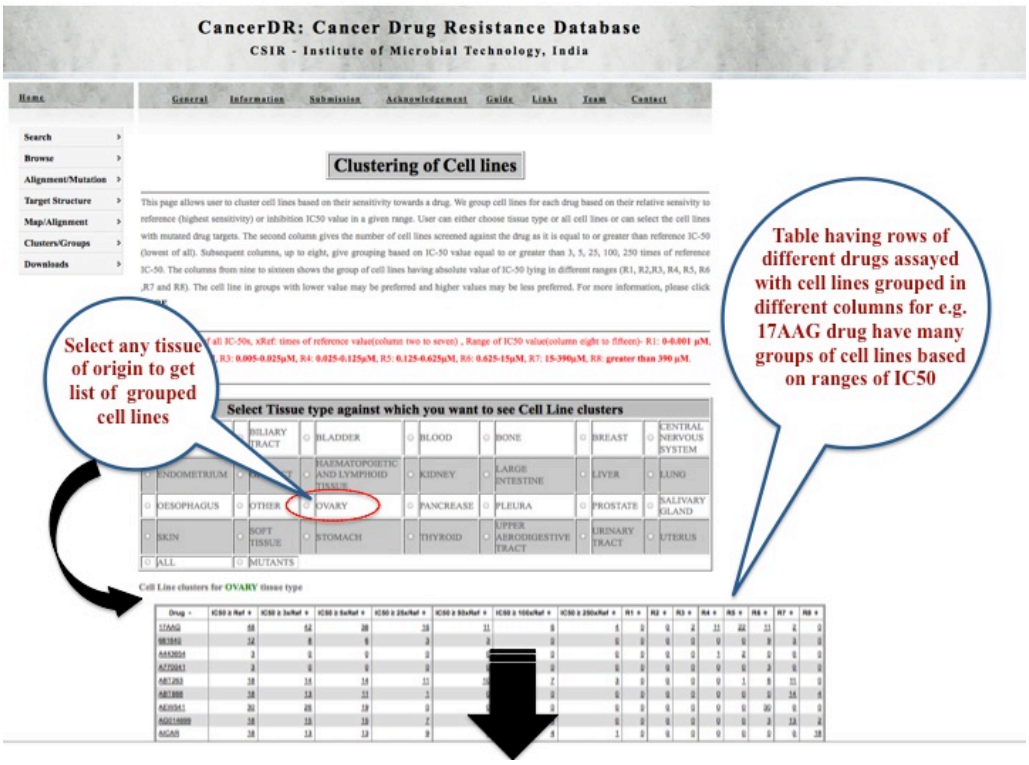


Figure 32a. Screen shot showing the cell lines clustering function.

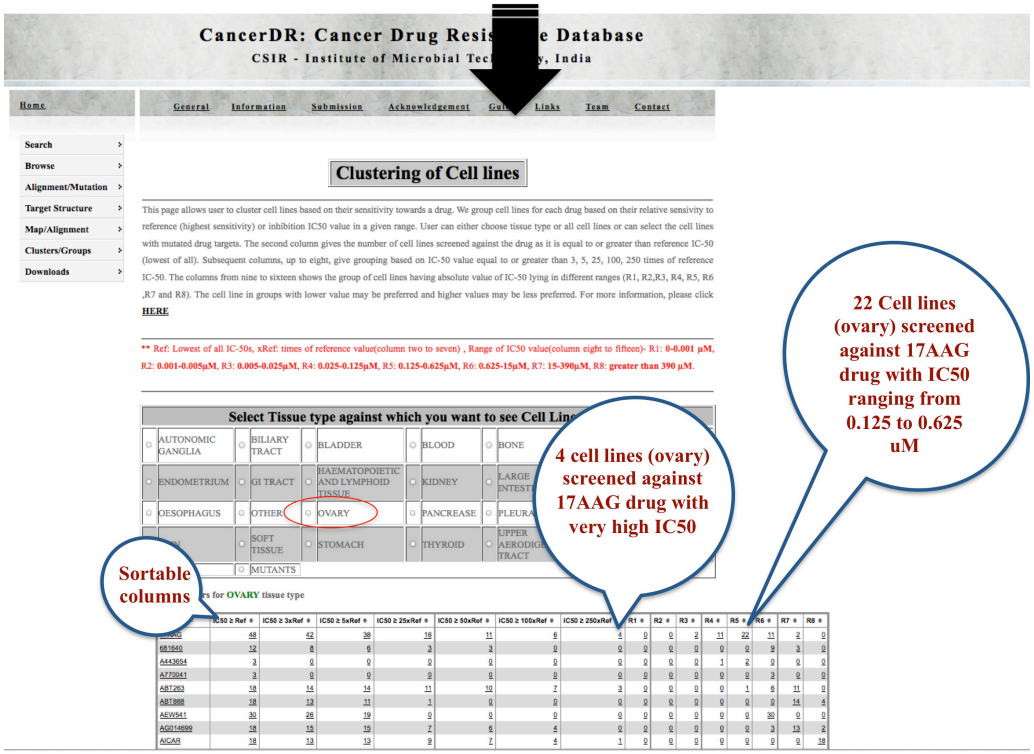


Figure 32b. Result of clustering of ovary cell lines.

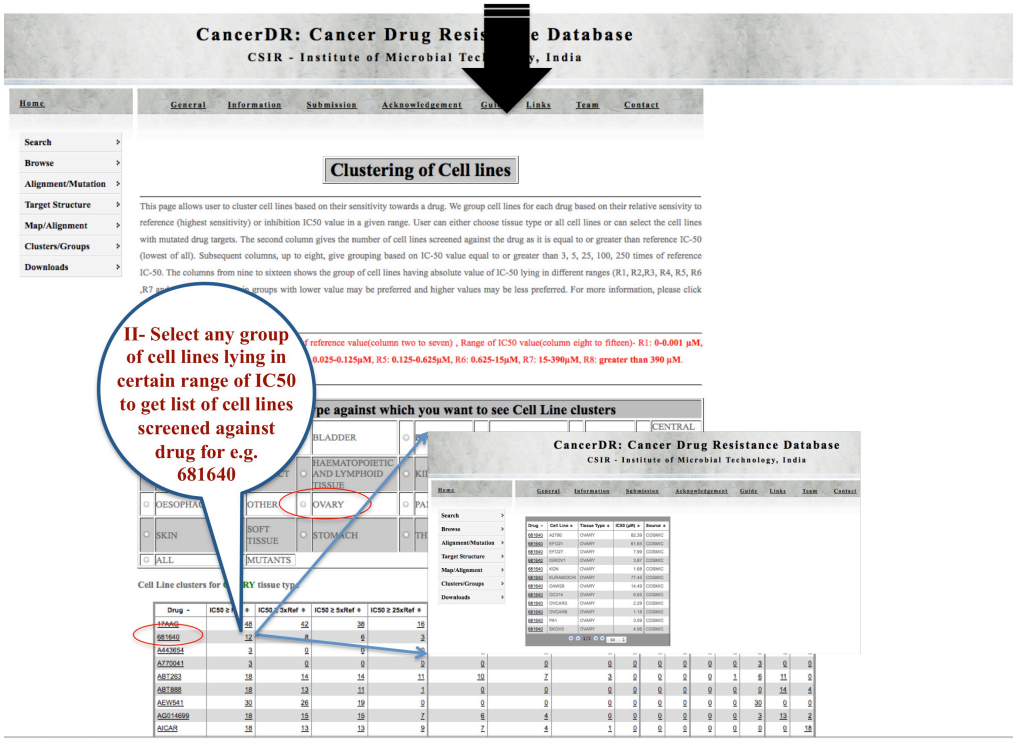


Figure 32c. Further exploration of the clustering tool to obtain IC50 values.

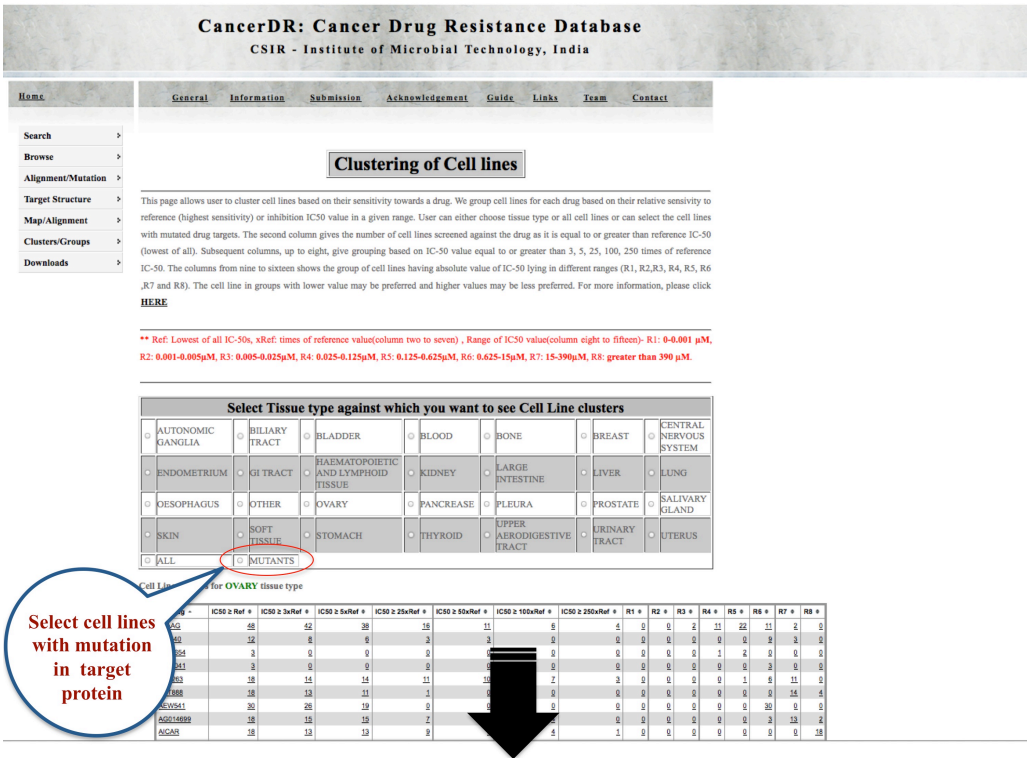


Figure 33a. Clustering of cell lines which are mutated in some drug target.

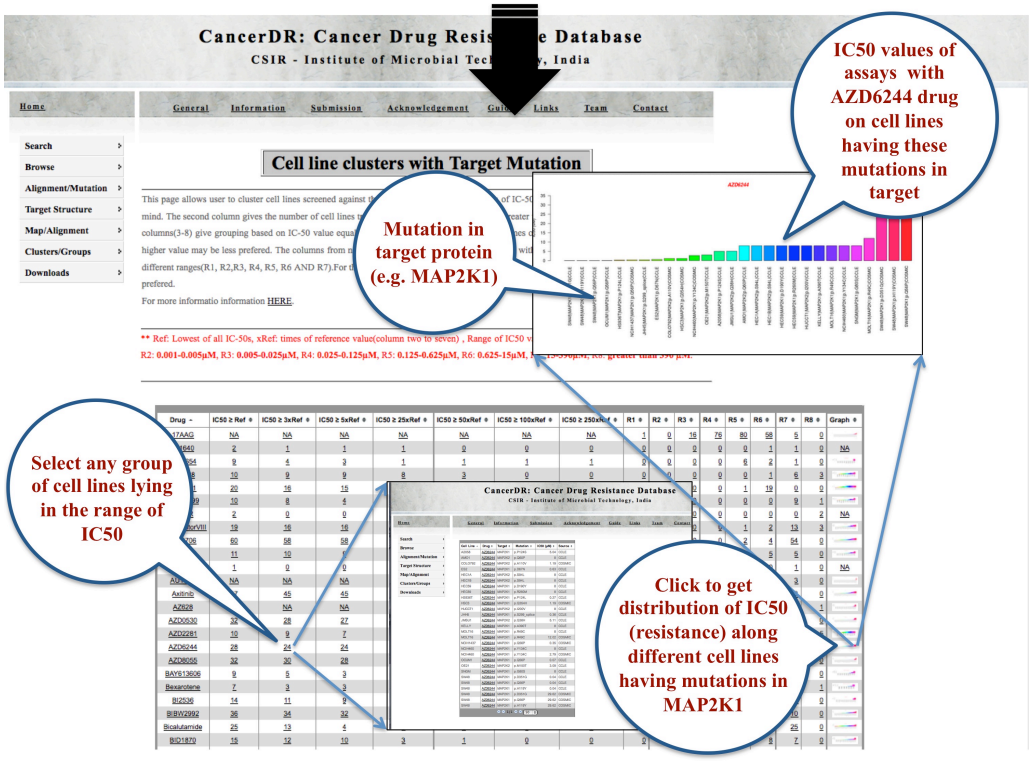


Figure 33b. Screenshot showing the Pharmacological profiling of AZD6244 drug.

6.3 Clustering of drugs: It is important to know which drugs have similar level of effectiveness for a given cell line or tissue type. This page allows the user to cluster drugs screened against selected cell line and group on the basis of IC50 or relative drug sensitivity. User can either choose the tissue type, all cell lines or can select the cell lines with mutated drug targets (Figure 34a). The second column gives the number of drugs screened against the particular cell line as it is equal to or greater than reference IC50 (lowest of all IC50s). Subsequent columns, up to eighth, give grouping based on IC50 value equal or greater than 3, 5, 25, 100, and 250 times of reference IC50. The columns from nine to sixteen show the group of drugs with absolute value of IC50 lying in different ranges (R1, R2, R3, R4, R5, R6, R7 and R8). The drugs in groups with higher value are important for resistance-mechanism related studies (Figure 34b).

The 'MUTANTS' button in the tissue list leads to a table on the next page, which has similar ranges of IC50 along with links to bar plots at the last column. The bar plots show the distribution of IC50 values along different cell lines having mutation in the targets of corresponding drugs.

Clicking on any number in displayed table leads to a new table having list of drugs grouped in that range of IC50 (Figure 34c).

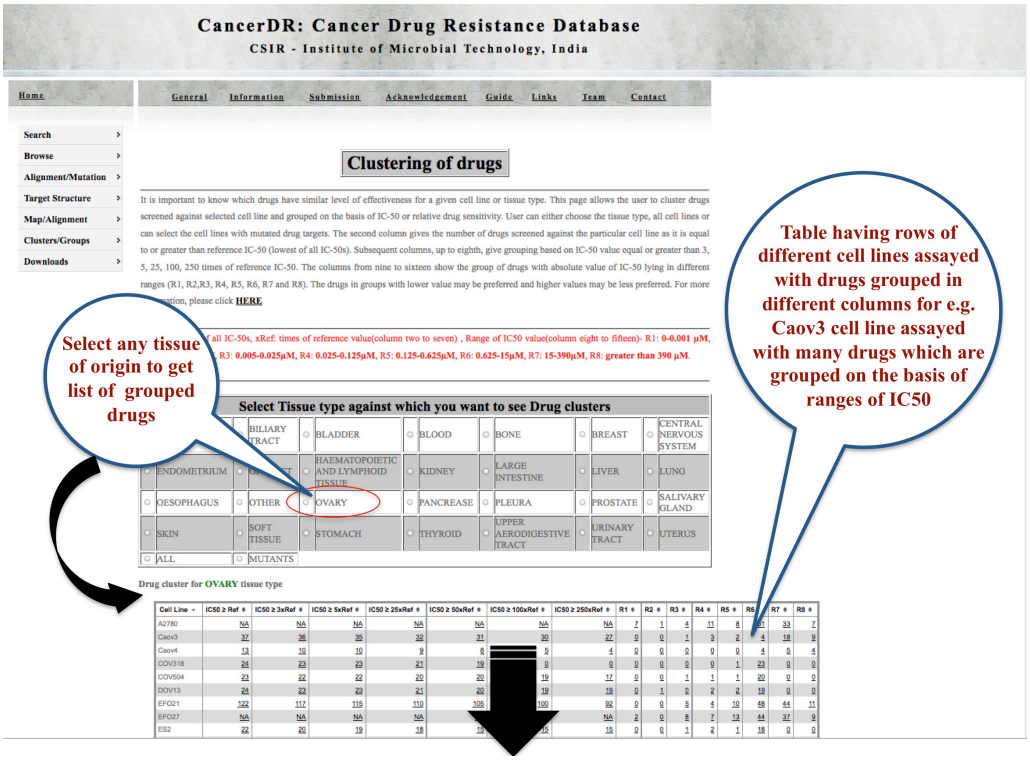


Figure 34a. Clustering of drugs assayed on ovary cell lines.

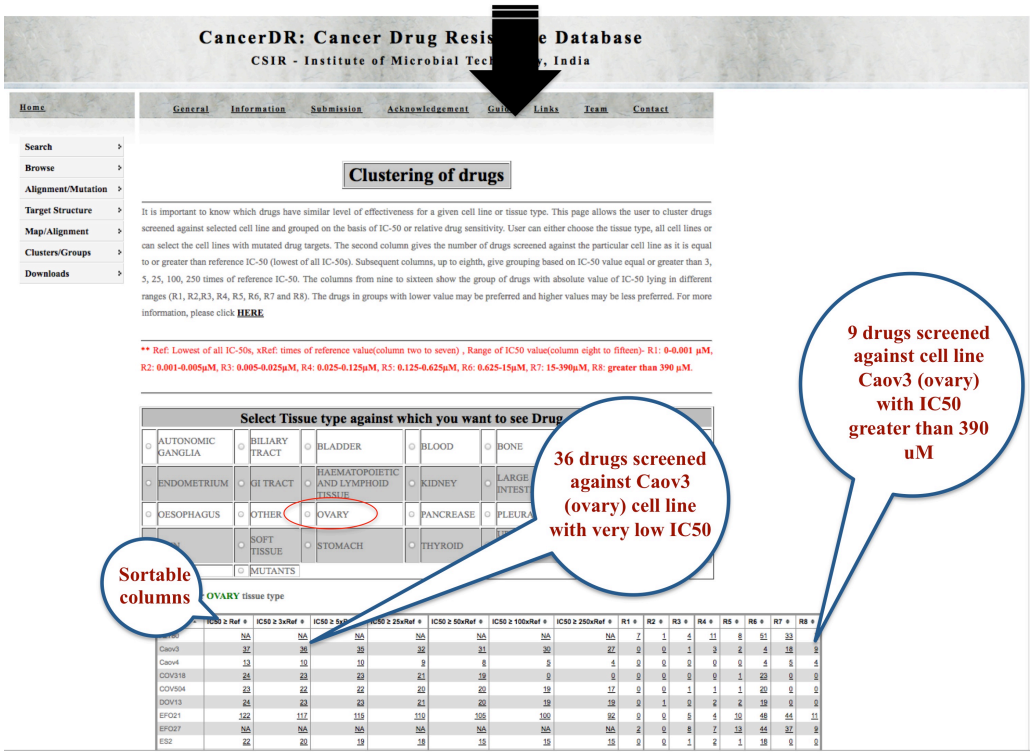


Figure 34b. Result showing clustered drugs.

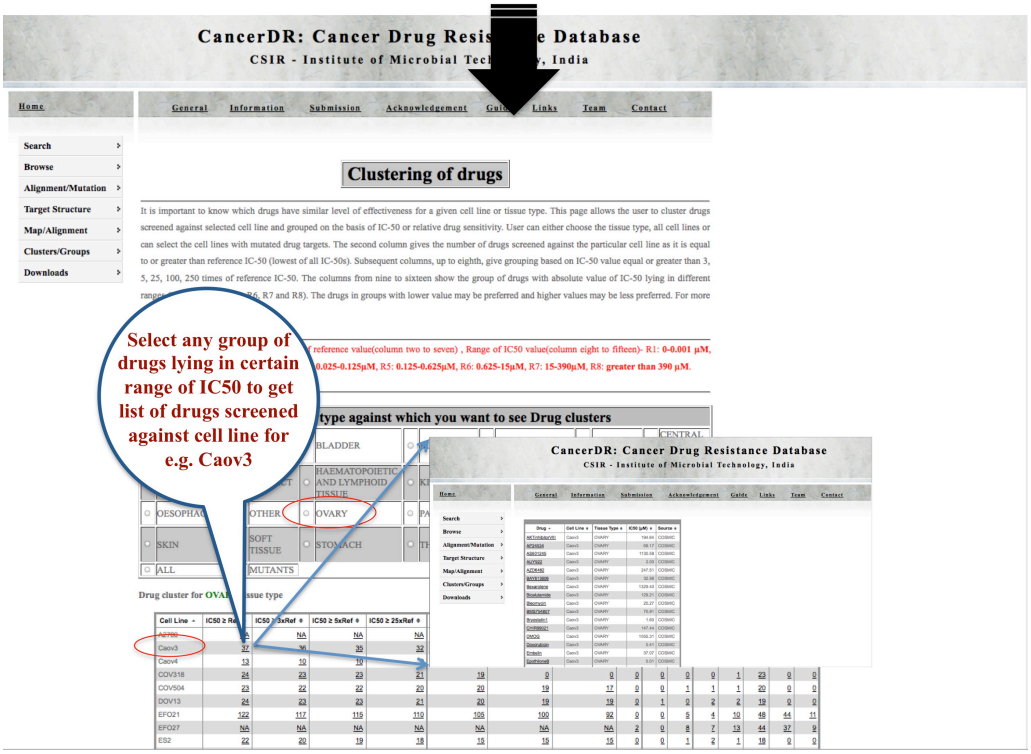


Figure 34c. IC50 values of assayed drugs on the respective cell lines.

7. Downloads

This section is divided into 4 modules: for downloading all target sequences and their mutants/variants, for downloading multiple sequence alignments of targets and their mutants/variants, for downloading predicted structures and for downloading experimental PDB structures (whichever are available).

8.1 Sequences: This page allows the users to download the sequences of drug targets (wild type) and their mutants/natural variants. User can download either a specific or all target sequences. User has to click on RED button to select mutants and BLUE button for natural variants against corresponding targets. User can either download a specific target and its mutants/natural variants or can browse whole directory containing all targets and their mutants/variants. Further, facility has been provided to synchronize whole sequence data at user's local disc using rsync command (Figure 35).

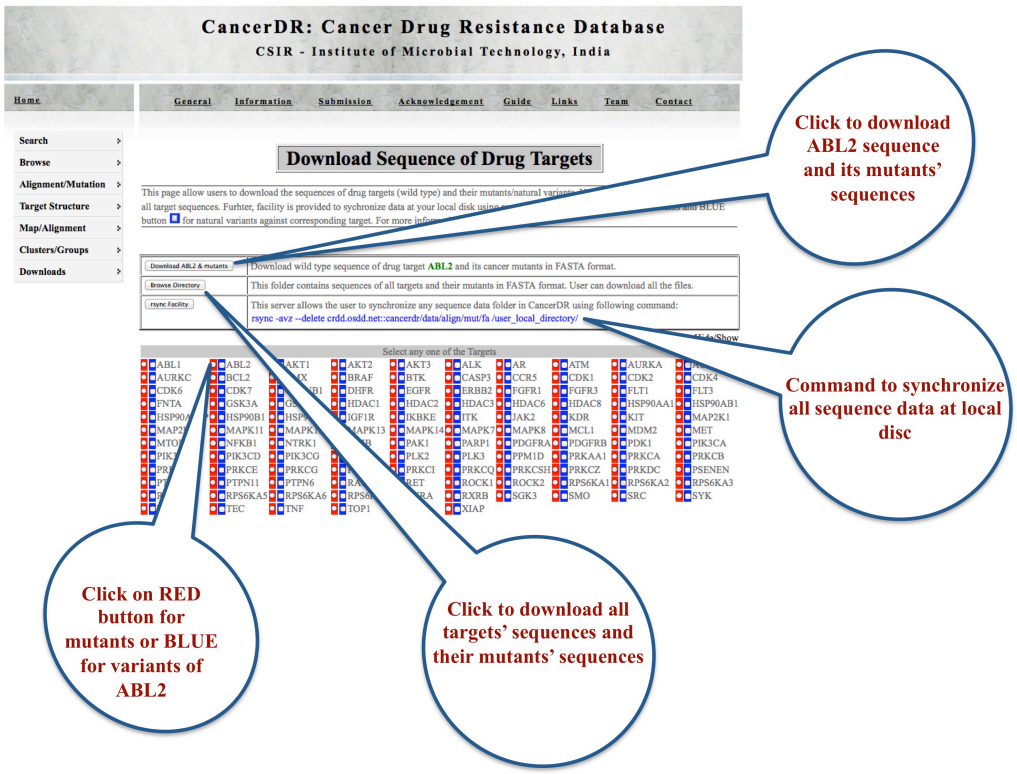


Figure 35. Sequence download.

7.2 Alignments: This module is to download the sequence alignment files in the CLUSTALW format of drug targets (wild type) with their mutants/natural variants. User can download either a specific or all sequence alignments. User has to click on RED button to select mutants' alignment and BLUE button for natural variants' alignment for the corresponding selected targets. User can either download alignment file containing a specific target and its mutants/natural variants or can browse whole directory containing all alignment files. Further, facility has been provided to synchronize all alignment data at user's local disc using rsync command (Figure 36).

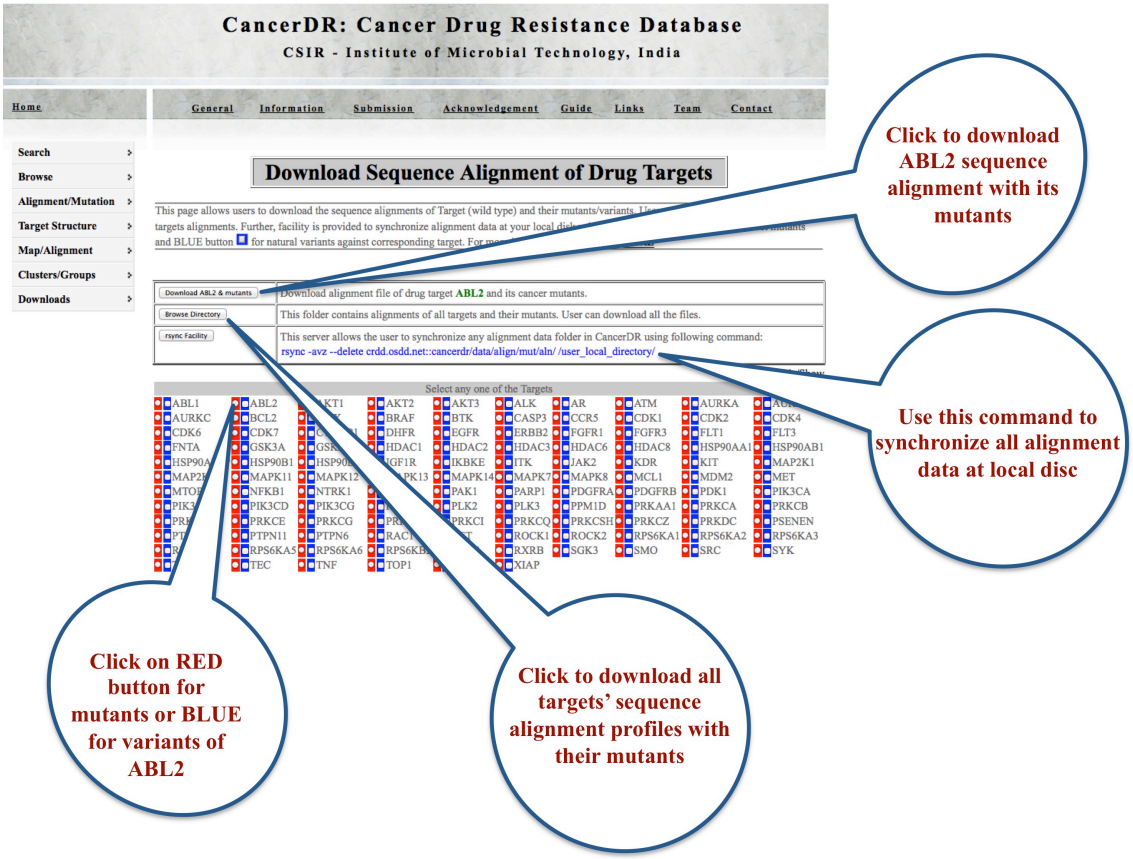


Figure 36. Sequence alignment download.

7.3 Structures: This section assists the users to download the predicted structures of drug targets (wild type) along with their mutants/natural variants. User can download either a specific target structure and structure of its mutants/variants or can browse directory containing all targets' structures and their mutants/variants. User has to click on RED button to select mutants and BLUE button for natural variants against corresponding targets. Facility has been provided to synchronize data at user's local disc using rsync command (Figure 37).

7.4 PDB structures: This section allows the users to download the experimentally etermined structures of drug targets (wild type) available from the Protein Data Bank. User can download either a specific target PDBs or can browse directory containing all targets'. User can also avail rsync facility to synchronize the structural data from the server to the local disc (Figure 38).

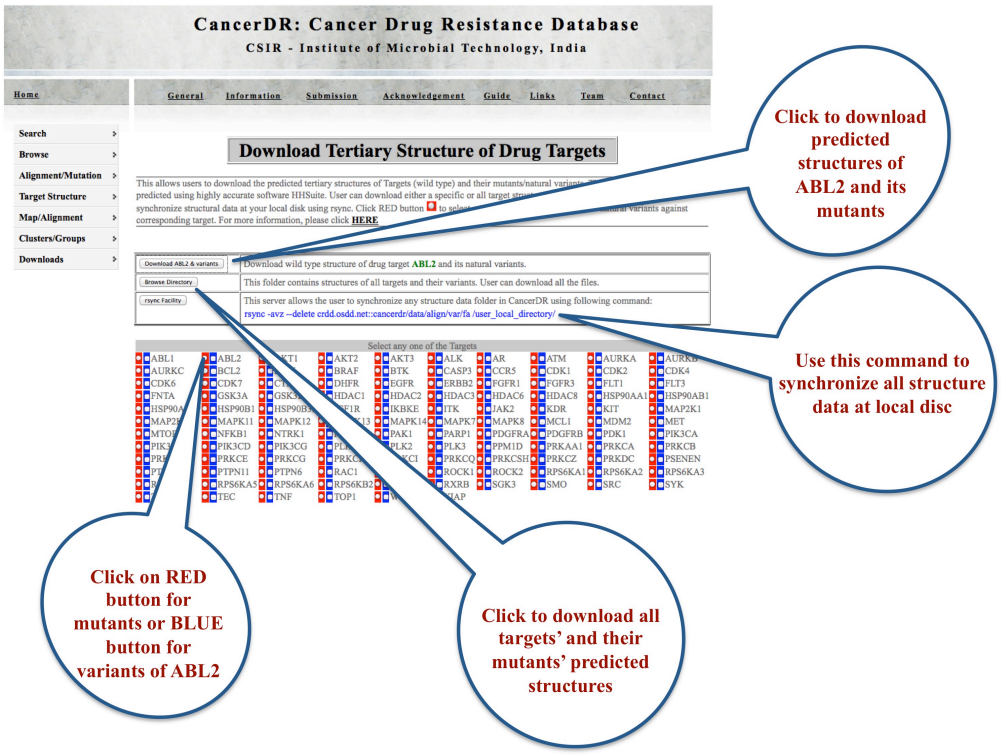


Figure 37. Predicted structure download.

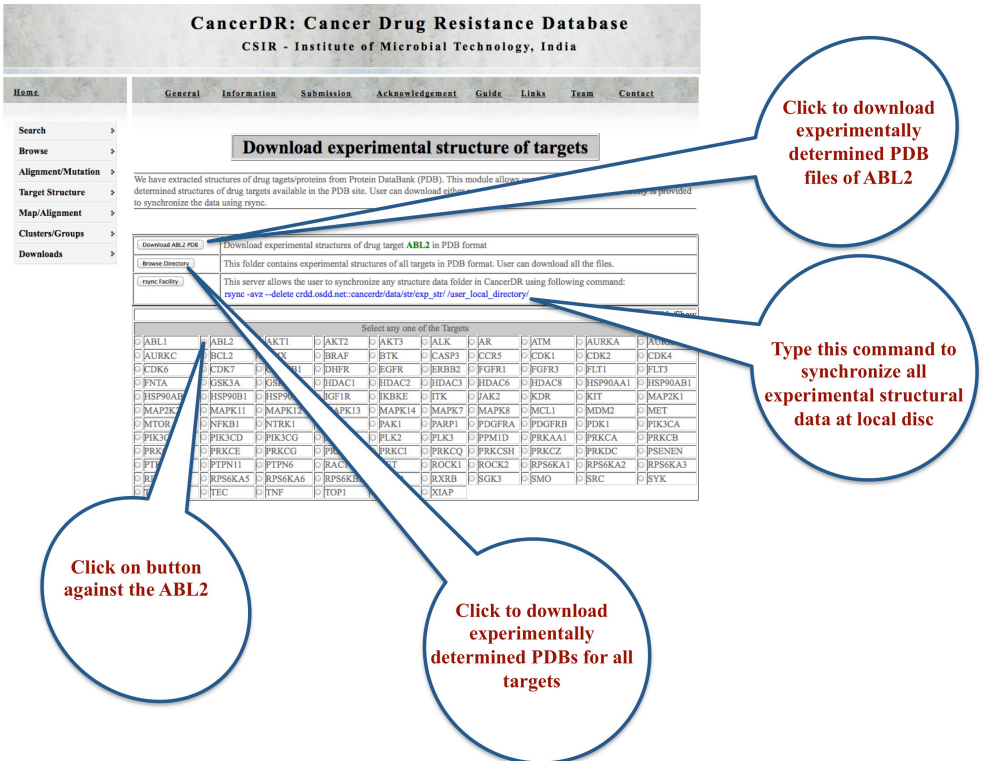


Figure 38. PDB structure (experimentally validated) download.