AS.410.712.SU22
Adv. Practical Computer Concepts for Bioinformatics
Chaojie (James) Yuan
Final Project Documentation

Documentation for the Design and Development of a CFTR Gene Variants Search Tool

Tool Background and Motivation

Cystic fibrosis (CF) is a rare autosomal recessive genetic disease resulting from mutations in the CFTR gene. According to my experience, it is difficult to directly find variants and single nucleotide polymorphisms (SNPs) of a single gene. Rather, it is more common to locate a gene in a whole genome sequencing data and find variants of the gene of interest that way. Therefore, I created a search tool webpage interface dedicated to variants of the CFTR gene that could take different types of search inputs and provide matching CFTR variants as outputs accordingly.

CFTR2 is a website maintained by Johns Hopkins University[1]. It maintains information about all the CFTR mutations that have been reported. It is mainly used to determine whether the variant or variant combination is CF-causing. Though the data stored in CFTR2 are comprehensive, the website could be a bit difficult to utilize if you do not have some variants in mind. It requires the user to put in at least one variant to perform a search. Therefore, I decided to redesign the search options on my web interface and move the CFTR variants data from CFTR2 to a table on the MySQL database, so it would be easier for the user to find their variant(s) of interest.

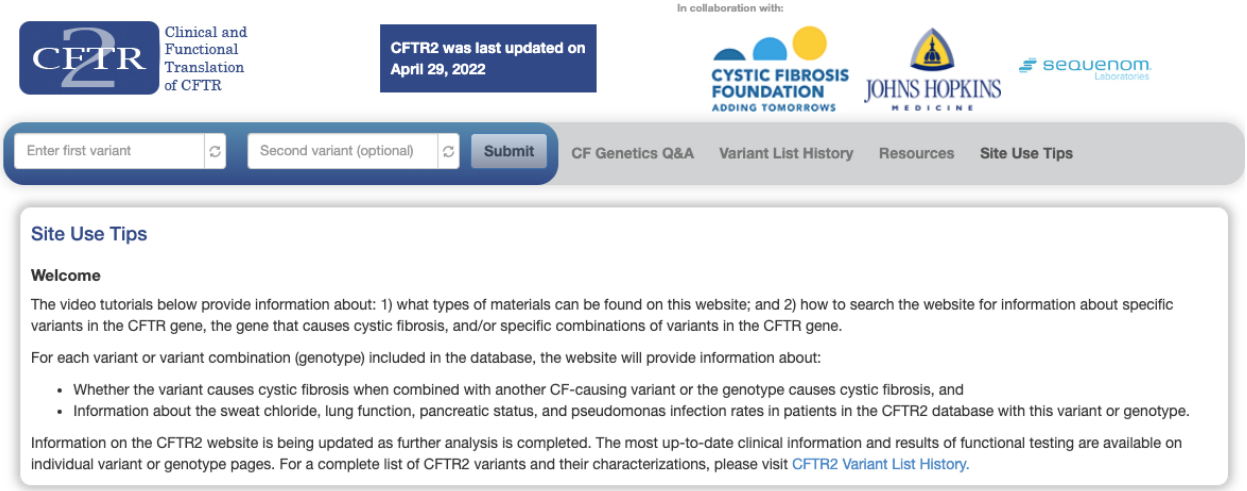

Figure 1. CFTR2 website interface.

Tool Functionality

The following is a screenshot of the webpage:
(http://bfx3.aap.jhu.edu/cyuan23/final_project/search.html)

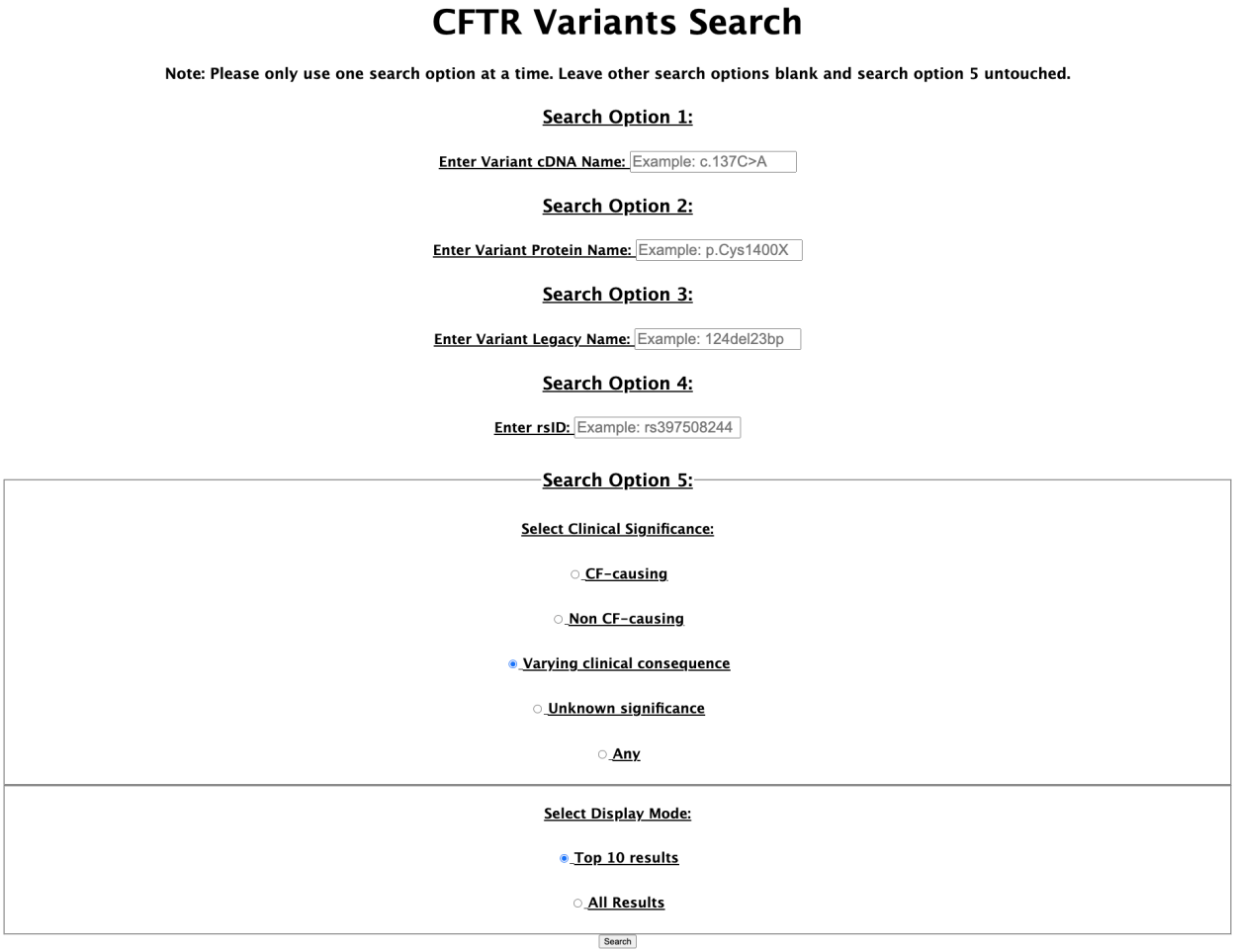

Figure 2. Screenshot of the search webpage.

As shown above, the search tools could take 5 types of search inputs:

1. Variant cDNA name
2. Variant protein name
3. Variant legacy name
4. rsID of the variant
5. Clinical significance of the variant

Note that it only takes one type of search input at a time. The user needs to clear out other unwanted search inputs to allow the proper functioning of the search tool. Also, the autocomplete functionality is incorporated in search options 1 through 4.

Example output:

The following screenshot shows the results if search options 1 through 4 were left blank, and "Varying clinical consequence" was selected in search options 5 and top 10 results was selected as the preferred display mode:

CFTR Variants Search Results

49 matches found.

Entries #	Variant cDNA Name	Variant Protein Name	Variant Legacy Name	rsID	Alleles in CFTR2	Allele Frequency in CFTR2	% Pancreatic Insufficient	Variant Clinical Consequence
1	c.350G>A	p.Arg117His	R117H	rs78655421	1854	0.01305	23	Varying clinical consequence
2	c.3454G>C	p.Asp1152His	D1152H	rs75541969	571	0.00402	24	Varying clinical consequence
3	c.1210?12T[5]	No protein name	5T	rs1805177	516	0.00363	28	Varying clinical consequence
4	c.1210-33_1210-6GT[12]T[4]	No protein name	5T;TG12	not found	182	0.00128	14	Varying clinical consequence
5	c.[350G>A;1210?12T[7]]	p.[Arg117His;No protein name]	R117H;7T	not found	125	0.00088	15	Varying clinical consequence
6	c.14C>T	p.Pro5Leu	P5L	rs193922501	60	0.00042	10	Varying clinical consequence
7	c.3808G>A	p.Asp1270Asn	D1270N	rs11971167	55	0.00039	17	Varying clinical consequence
8	c.1736A>G	p.Asp579Gly	D579G	rs397508288	53	0.00037	35	Varying clinical consequence
9	c.1210-33_1210-6GT[13]T[4]	No protein name	5T;TG13	not found	43	0.00030	11	Varying clinical consequence
10	c.3208C>T	p.Arg1070Trp	R1070W	rs202179988	36	0.00025	34	Varying clinical consequence

Figure 3. Screenshot of the result webpage.

As shown above, the table has 9 columns:

- Entries #
- Variant cDNA name
- Variant protein name
- Variant legacy name
- rsID
- Alleles in CFTR2
- Allele frequency in CFTR2:
- % pancreatic insufficient
- Variant clinical consequence

Note that on the top the page shows there are 49 matches found. This indicates that there are 49 CFTR variants in total that have "varying clinical consequence". However, the page only displays 10 entries due to the user selected to only display top 10 results, which are ranked by descending alleles in CFTR2. CFTR2 is a database run by Johns Hopkins University and it keeps records of 485 *CFTR* variants, their clinical consequences. Also, the webpage returns another parameter called % pancreatic insufficient, which indicates the percentage of patients with variant in trans with ACMG-PI variant, with variant in homozygosity, or with another variant expected to lead to no CFTR protein production. In short, the webpage will rank the results from most common among the database to the least common as long as there are more than one matches.

Tool Description

The CFTR variants search tool utilizes the following three technologies to carry out its functionality:

1. SQL relational database: Data of 485 CFTR variants are stored in a table on the MySQL database.

2. Python-based Computer Gateway Interface (CGI): A CGI script is used to accept the search inputs from the search HTML. Then it connects to the MySQL database on the JHU BFX server via mysql.connector package, and sends appropriate queries to the database and access entries in the table. Then it stores all the results and sends them to the result HTML.
3. CSS/HTML5/JavaScript(JQuery) Graphic User Interface (GUI): A HTML5 page is used to collect user search inputs. A JavaScript file incorporates autocomplete functionality into the HTML search boxes. Another HTML5 page is used as a template file for the result table. 2 CSS files (one for the search page and another for the results page) are used for styling the web interface.

Database Population

Create table in mysql database:

```
MariaDB [cyuan23]> CREATE TABLE cftr_variant(
    -> variant_cDNA_name VARCHAR(255) PRIMARY KEY,
    -> variant_protein_name VARCHAR(255),
    -> variant_legacy_name VARCHAR(255),
    -> rs_ID VARCHAR(255),
    -> alleles_in_CFTR2 INT NOT NULL,
    -> allele_frequency_in_CFTR2 DECIMAL(5,5),
    -> percent_pancreatic_insufficient INT NOT NULL,
    -> variant_clinical_consequence VARCHAR(255));
Query OK, 0 rows affected (0.01 sec)
```

Populate table (full command not shown):

```
MariaDB [cyuan23]> INSERT INTO cftr_variant (variant_cDNA_name, variant_protein_name, variant_legacy_name, rs_ID, alleles_in_CFTR2, allele_frequency_in_CFTR2, percent_pancreatic_insufficient, variant_clinical_consequence)
    -> VALUES("c.(?_-1270)_(53+1_54-1)del", "No protein name", "CFTRdelePr-1", "not found", 3, 0.00002, 67, "CF-causing"),
    -> ("c.-9_14del23", "No protein name", "124del23bp", "rs397508136", 6, 0.00004, 100, "CF-causing"),
    -> ("c.-86>C", "No protein name", "125G/C", "rs1800501", 8, 0.00006, 0, "Non CF-causing"),
    -> ("c.(?_1)_(53+1_54-1)del", "p.Glu2GlyfsX17", "CFTRdele1", "not found", 6, 0.00004, 100, "CF-causing"),
    -> ("c.1A>G", "p.Met1Val", "M1V", "rs397508328", 26, 0.00018, 84, "CF-causing"),
    -> ("c.4C>T", "p.Gln2X", "Q2X", "rs397508740", 5, 0.00004, 100, "CF-causing"),
    -> ("c.[4C>T;7A>T]", "p.[Gln2X;Arg3Trp]", "Q2X;R3W", "not found", 4, 0.00003, 100, "CF-causing"),
    -> ("c.11C>A", "p.Ser4X", "S4X", "rs397508173", 14, 0.0001, 100, "CF-causing"),
    -> ("c.14C>T", "p.Pro5Leu", "P5L", "rs193922501", 60, 0.00042, 10, "Varying clinical consequence"),
    -> ("c.38C>T", "p.Ser13Phe", "S13F", "rs397508635", 3, 0.00002, 33, "CF-causing"),
    -> ("c.44T>C", "p.Leu15Pro", "L15P", "rs1562876459", 4, 0.00003, 0, "CF-causing"),
    -> ("c.50delT", "p.Phe17SerfsX8", "182delT", "rs397508742", 9, 0.00006, 86, "CF-causing"),
    -> ("c.50dupT", "p.Ser18GlnfsX27", "176insT", "rs397508714", 7, 0.00005, 83, "CF-causing"),
    -> ("c.(53+1_54-1)_(164+1_165-1)del", "No protein name", "CFTRdele2", "not found", 46, 0.00032, 100, "CF-causing"),
    -> ("c.(53+1_54-1)_(489+1_490-1)del", "No protein name", "CFTRdele2-4", "not found", 4, 0.00003, 100, "CF-causing"),
    -> ("c.53+1G>T", "No protein name", "185+1G->T", "rs397508746", 8, 0.00006, 100, "CF-causing"),
    -> ("c.54-5940_273+10250del21kb", "p.Ser18ArgfsX16", "CFTRdele2,3", "not found", 417, 0.00294, 100, "CF-causing"),
    -> ("c.54-5842_489+401del", "No protein name", "IVS1-5842_IVS4+401del", "not found", 4, 0.00003, 100, "CF-causing"),
    -> ("c.57G>A", "p.Trp19X", "W19X", "rs397508762", 4, 0.00003, 100, "CF-causing"),
    -> ("c.79G>A", "p.Gly27Arg", "G27R", "rs397508796", 3, 0.00002, 67, "CF-causing"),
    -> ("c.79G>T", "p.Gly27X", "G27X", "rs397508796", 10, 0.00007, 100, "CF-causing"),
    -> ("c.88C>T", "p.Gln30X", "Q30X", "rs397508815", 3, 0.00002, 100, "CF-causing"),
    -> ("c.91C>T", "p.Arg31Cys", "R31C", "rs1800073", 23, 0.00016, 30, "Non CF-causing"),
    -> ("c.92G>T", "p.Arg31Leu", "R31L", "rs149353983", 7, 0.00005, 0, "Unknown significance"),
    -> ("c.115C>T", "p.Gln39X", "Q39X", "rs397508168", 47, 0.00033, 100, "CF-causing"),
    -> ("c.137C>A", "p.Ala46Asp", "A46D", "rs151020603", 4, 0.00003, 100, "CF-causing"),
    -> ("c.164+1G>A", "No protein name", "296+1G->A", "rs397508243", 3, 0.00002, 100, "CF-causing"),
    -> ("c.164+1G>T", "No protein name", "296+1G->T", "rs397508243", 4, 0.00003, 100, "CF-causing"),
    -> ("c.164+1_165-1_(1584_+1_1585-1)del(2619+1_2628-1)_(2988+1_2989-1)del", "No protein name", "CFTRdele3-10,14b-16", "not found", 4, 0.00003, 100, "CF-causing"),
    -> ("c.164+28A>G", "No protein name", "296+28A->G", "rs34010645", 9, 0.00006, 67, "Unknown significance"),
    -> ("c.164+2T>C", "No protein name", "296+2T->C", "rs121908800", 3, 0.00002, 100, "CF-causing"),
    -> ("c.164+4dupT", "No protein name", "296+3insT", "rs397508244", 4, 0.00003, 50, "CF-causing"),
    -> ("c.165-3C>T", "No protein name", "297-3C->T", "rs200337193", 3, 0.00002, 100, "Unknown significance"),
    -> ("c.165-1G>A", "No protein name", "297-1G->A", "rs397508249", 5, 0.00004, 100, "CF-causing"),
    -> ("c.166G>A", "p.Glu56Lys", "E56K", "rs397508256", 5, 0.00004, 0, "CF-causing"),
    -> ("c.168delA", "p.Glu56AspfsX35", "300delA", "rs397508269", 8, 0.00006, 100, "CF-causing"),
    -> ("c.169T>G", "p.Trp57Gly", "W57G", "rs397508272", 10, 0.00007, 100, "CF-causing"),
    -> ("c.170G>A or c.171G>A", "p.Trp57X", "W57X", "rs397508279 or rs121909925", 14, 0.0001, 100, "CF-causing"),
    -> ("c.175dupA", "p.Arg59LysfsX10", "306insA", "rs397508294", 21, 0.00015, 94, "CF-causing"),
    -> ("c.174_177delTAGA", "p.Asp58GlufsX32", "306delTAGA", "rs397508295", 6, 0.00004, 100, "CF-causing"),
    -> ("c.178G>A", "p.Glu60Lys", "E60K", "rs77284892", 7, 0.00005, 40, "CF-causing"),
    -> ("c.178G>T", "p.Glu60X", "E60X", "rs77284892", 296, 0.00208, 99, "CF-causing"),
    -> ("c.200C>T", "p.Pro67Leu", "P67L", "rs368505753", 239, 0.00168, 34, "CF-causing"),
    -> ("c.202A>T", "p.Lys68X", "K68X", "not found", 3, 0.00002, 100, "CF-causing"),
    -> ("c.220C>T", "p.Arg74Trp", "R74W", "rs115545701", 36, 0.00025, 15, "Varying clinical consequence"),
```

```
Query OK, 485 rows affected (0.02 sec)
Records: 485  Duplicates: 0  Warnings: 0
```

The data are from an Excel spreadsheet named **CFTR2_29April2022.xlsx** downloaded from the CFTR2 website. It is the most recent version (April 2022). It is available both on Canvas and on Github.

Tool Limitations

- One of the biggest limitations of this search tool website is that it requires the user to do further analysis on their own. It does provide useful information such as the rsID, variant cDNA name and variant protein name, etc so the user could tell whether the variant is a coding variant or a noncoding variant, whether the variant is an insertion or an deletion.

- The second limitation is that I could not find any information about the patient demographics or any patient information at all. This could be due to the fact that the patient information is confidential and is protected by Johns Hopkins. However, I do believe that knowing more information about the patients would help immensely with the association study of variants and cystic fibrosis. Imagine a variant has very high allele frequency among the CFTR2 patient population, and it is CF-causing. That very variant may not have a high frequency or be CF-causing at all for a patient population that has nothing in common with the CFTR patient population.

- Though the website is perfectly functional, it could be further improved aesthetically. More CSS styling could be applied, but due to time constraints of the summer semester, I only had time to make sure that it functions as promised.

References:

1. The Clinical and Functional TRanslation of CFTR (CFTR2); available at http://cftr2.org