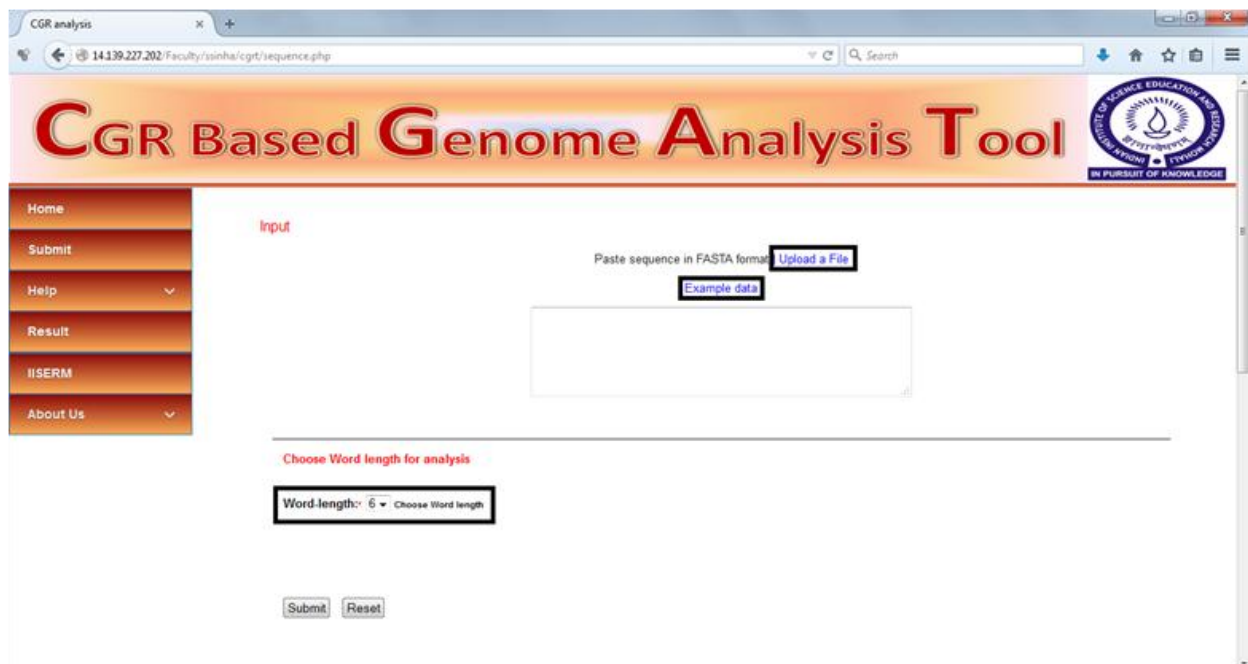


Usage

Input

The DNA sequences to be analyzed can be either pasted in the text box on "Submit" page or uploaded in the form text file. All the sequences must be in FASTA format. The allowed DNA characters in file are A, T, G, C, R, Y and N.

Whole genome sequences of HIV-1 subtypes A, B, C, D, F, G, SIV and HTLV (Dataset used in Pandit et al Mol. Phylo. Evol., 2012) have been provided as an example. Click on the "Example data" link on Submit page to upload this data for analysis.



The screenshot shows the web interface of the "CGR Based Genome Analysis Tool". The browser address bar displays "14.139.227.202/Faculty/sinha/cgrt/sequence.php". The page has a header with the tool's name in large red letters and a logo of the Institute of Science, Education and Research (ISER) on the right. A left sidebar contains navigation links: Home, Submit, Help, Result, IISERM, and About Us. The main content area is titled "Input" and contains a text box for "Paste sequence in FASTA format". Above this box are links for "Upload a File" and "Example data". Below the text box is a section titled "Choose Word length for analysis" with a dropdown menu currently set to "Word-length: 6" and a "Choose Word length" link. At the bottom of the input section are "Submit" and "Reset" buttons.

Selection of "Word Length"

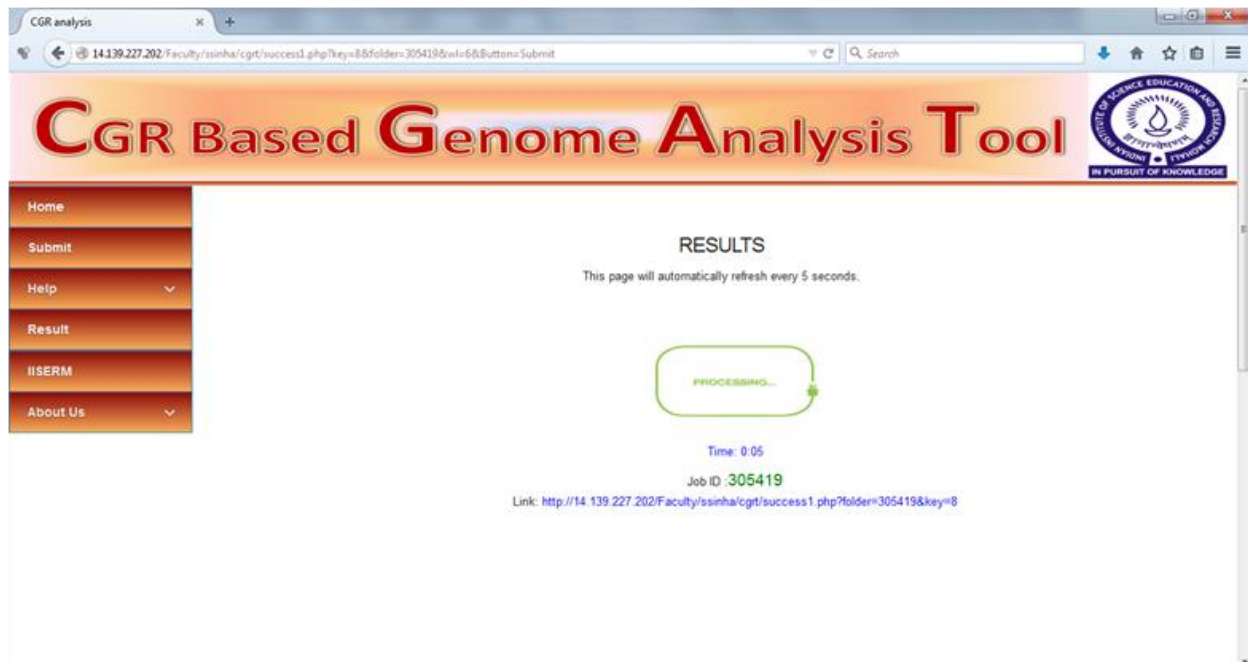
The clustering of the sequences is based on the distances calculated from the frequencies of DNA words. The word length to be used for the calculation can be specified by the user at the submission page. This default word length used is 6.

After selecting the word length click on "Submit"

Selection of Out group

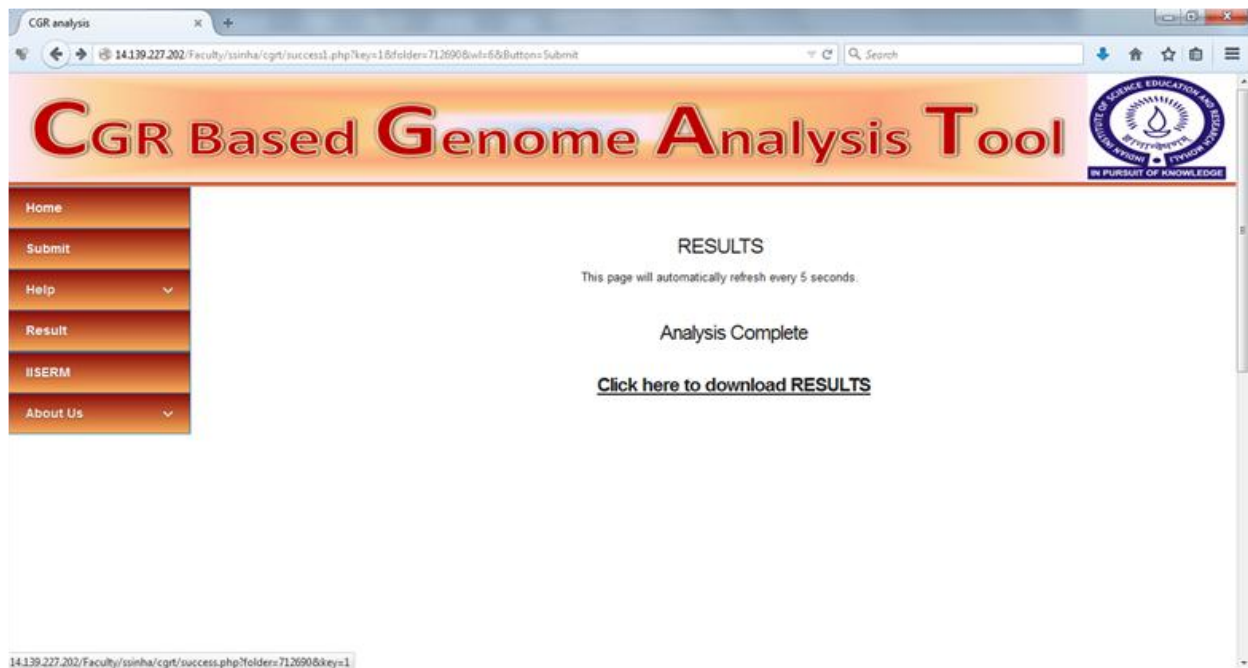
An out group needs to be specified for construction of Neighbor Joining Tree by Phylip. Select the out group from the dropdown menu and confirm the submission.

The calculation is started immediately and each submission is provided a unique Job ID.



Results

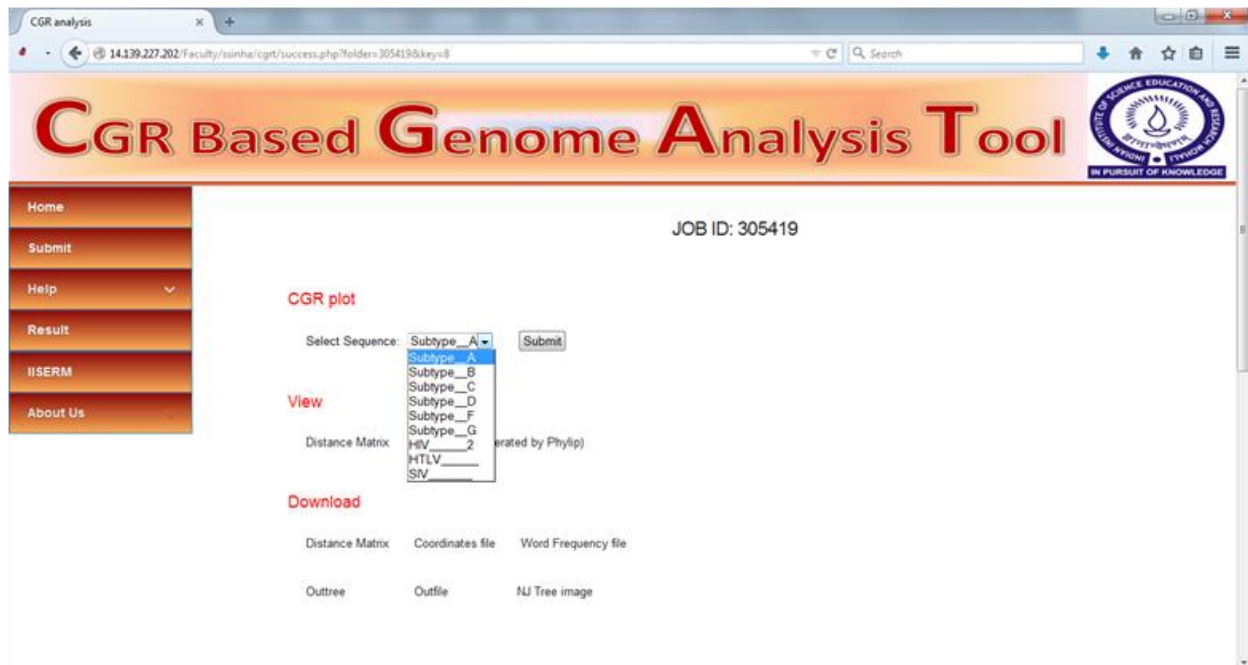
Once the Analysis is complete the results page shows a link to all the results. Click on the link to access results.



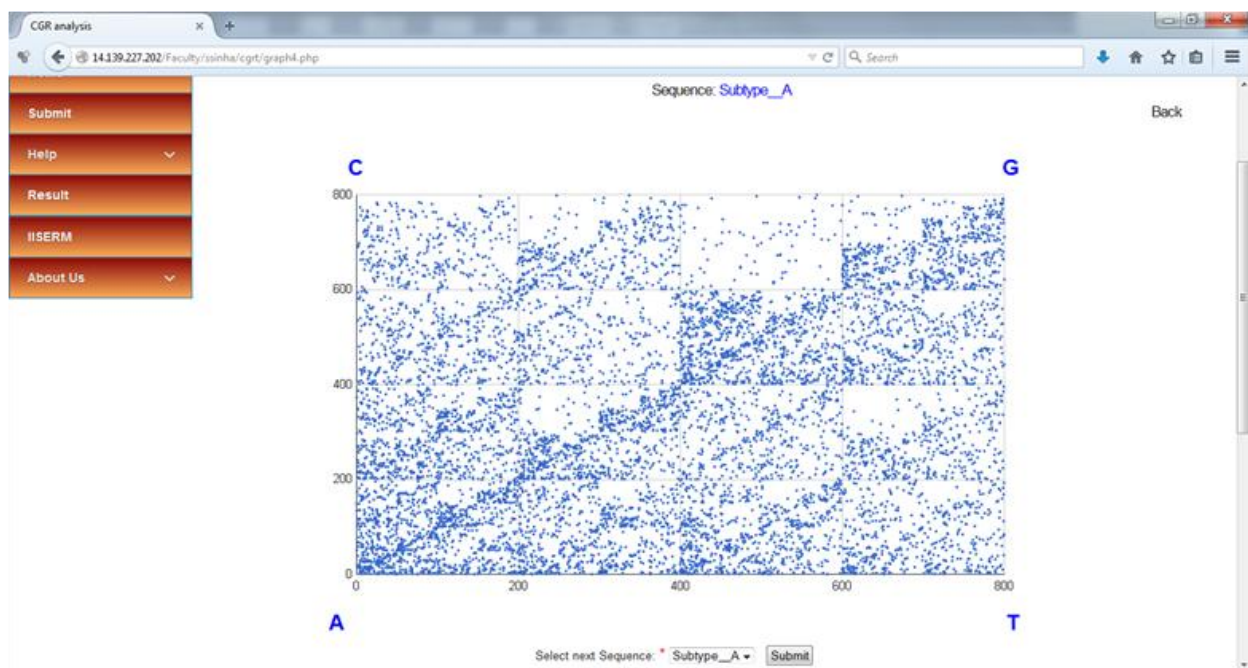
The results comprise of following

Visualization of CGR plot

CGRs for each sequence can be visualized by selecting the sequence from dropdown menu.



CGR plot for the Whole genome sequence of HIV-1 Subtype A as visualized on the browser, has been shown in the following screenshot



This plot can be saved by right click and selecting "Save image as" option.

View Distance Matrix

The screenshot shows the 'CGR Based Genome Analysis Tool' web interface. The browser address bar displays '141.39.227.202/Faculty/sinha/cgr/success.php?folder=305419&keys='. The page title is 'CGR Based Genome Analysis Tool'. A sidebar on the left contains links: Home, Submit, Help, Result, IISERM, and About Us. The main content area shows 'JOB ID: 305419'. Under the 'CGR plot' section, there is a 'Select Sequence' dropdown menu set to 'Subtype_A' and a 'Submit' button. Below this, the 'View' button is highlighted with a red box. Under the 'Download' section, there are links for 'Distance Matrix', 'Coordinates file', 'Word Frequency file', 'Outtree', 'Outfile', and 'NJ Tree image'. The IISERMA logo is visible in the top right corner.

Distance matrix generated by the CGR method can be used directly into phylip can be viewed on the browser. It can also be downloaded directly from this page.

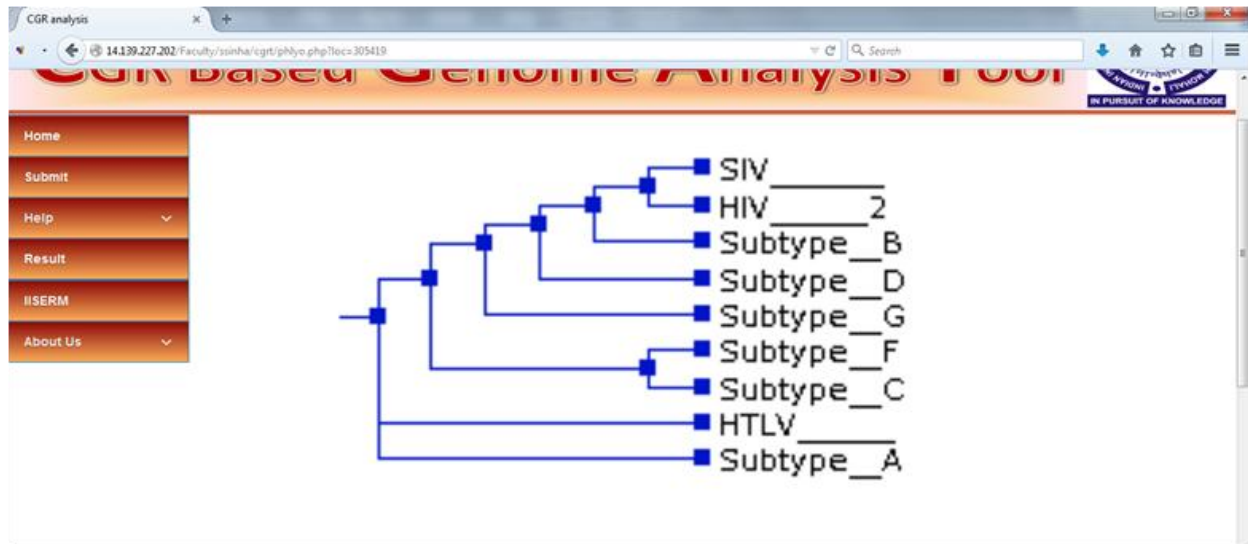
The screenshot shows the 'Distance matrix' output page of the 'CGR Based Genome Analysis Tool'. The browser address bar displays '141.39.227.202/Faculty/sinha/cgr/distance.php?loc=305419'. The page title is 'CGR Based Genome Analysis Tool'. The sidebar on the left is the same as in the previous screenshot. The main content area shows the 'Distance matrix' section with a text box containing the following data:

```
9
Subtype_A 0 2999.97 3438.13 3337.13 3484.09 3407.32 2452.14 3434.23 2957.84
Subtype_B 2999.97 0 2968.83 2851.82 3024.55 2936.58 1734.59 2971.58 2399.86
Subtype_C 3438.13 2968.83 0 3312.33 3458.88 3383.75 2416.44 3414.23 2927.66
Subtype_D 3337.13 2851.82 3312.33 0 3362.16 3282.62 2272.84 3312.29 2812.6
Subtype_F 3484.09 3024.55 3458.88 3362.16 0 3430.36 2482.86 3459.66 2882.71
Subtype_G 3407.32 2936.58 3383.75 3282.62 3430.36 0 2376.54 3382.09 2895.55
HIV_2 2452.14 1734.59 2416.44 2272.84 2482.86 2376.54 0 2417.9 1664.76
HTLV 3434.23 2971.58 3414.23 3312.29 3459.66 3382.09 2417.9 0 2925.99
SIV 2957.84 2399.86 2927.66 2812.6 2895.55 1664.76 2925.99 0
```

Below the text box, there are 'Download' and 'Back' buttons. The IISERMA logo is visible in the top right corner.

View NJ Tree

NJ tree, created by Phylip using the distance matrix shown above, can be visualized on the browser. This visualization is created using the Notung (version 2.6) package. This image can also be downloaded directly from this page in PNG format.



Download Results

All the results can be downloaded from the same page. The file that can be downloaded are

1) Distance Matrix file- It is (.txt) text file that contains pairwise distance matrix generated using input sequences. We used Euclidean distance method to calculate pairwise distances between multiple whole genome sequences. It contains output in standard PHYLIP distance matrix format.

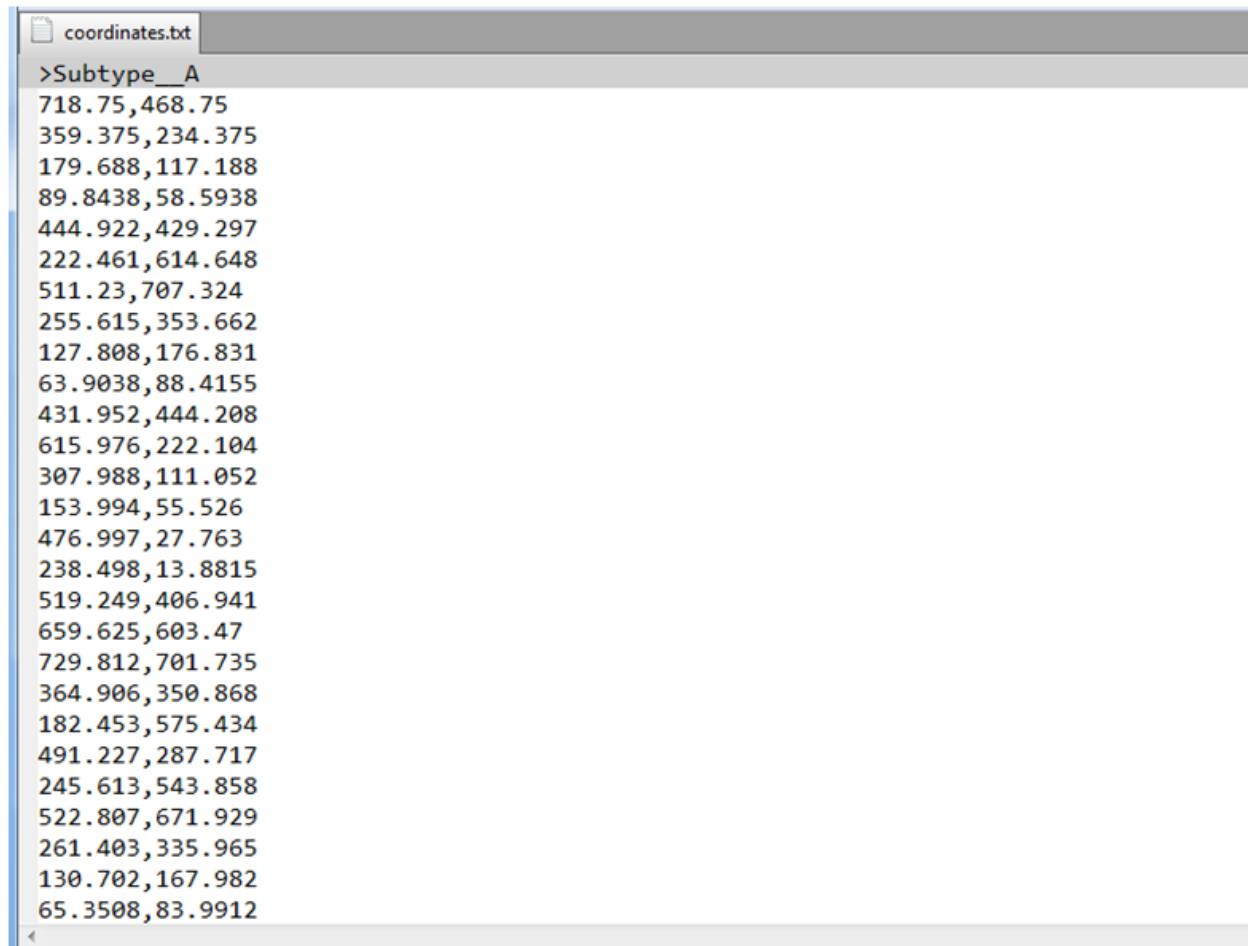
Following is the screenshot of the file for example data.

distance-1.txt									
9									
Subtype__A	0	2999.97	3438.13	3337.13	3484.09	3407.32	2452.14	3434.23	2957.84
Subtype__B	2999.97	0	2968.83	2851.82	3024.55	2936.38	1734.59	2971.58	2399.86
Subtype__C	3438.13	2968.83	0	3312.33	3458.88	3383.75	2416.44	3414.23	2927.66
Subtype__D	3337.13	2851.82	3312.33	0	3362.16	3282.62	2272.84	3312.29	2812.6
Subtype__F	3484.09	3024.55	3458.88	3362.16	0	3430.36	2482.86	3459.66	2982.71
Subtype__G	3407.32	2936.38	3383.75	3282.62	3430.36	0	2376.54	3382.09	2895.55
HIV_____2	2452.14	1734.59	2416.44	2272.84	2482.86	2376.54	0	2417.9	1664.76
HTLV_____	3434.23	2971.58	3414.23	3312.29	3459.66	3382.09	2417.9	0	2925.99
SIV_____	2957.84	2399.86	2927.66	2812.6	2982.71	2895.55	1664.76	2925.99	0

2) Coordinates file- Coordinate file is a (.txt) text file that contains x,y coordinates for each point in CGR of each sequence. It contains co-ordinates in following format.

>First Sequence name
 X-coordinates, Y-coordinates
 ...
 ..
 >Second Sequence name
 X-coordinates, Y-coordinates
 ...

Following is the screenshot of the file for example data.



3) Word Frequency file- It is a (.txt) text file that contains frequencies of all different k-letter words corresponding to CGR map.

For example – At k=3 in example data set

CCC	GCC	CGC	GGC	CCG	GCG	CGG	GGG
ACC	TCC	AGC	TGC	ACG	TCG	AGG	TGG
CAC	GAC	CTC	GTC	CAG	GAG	CTG	GTG
AAC	TAC	ATC	TTC	AAG	TAG	ATG	TTG

CCA	GCA	CGA	GGA	CCT	GCT	CGT	GGT
ACA	TCA	AGA	TGA	ACT	TCT	AGT	TGT
CAA	GAA	CTA	GTA	CAT	GAT	CTT	GTT
AAA	TAA	ATA	TTA	AAT	TAT	ATT	TTT

(2-d matrix of all possible 3 letter words using nucleotides A,T,G,C)

For the example data at word length 3

Subtype__A

```

78 73 15 117 15 26 26 186
110 74 192 94 20 19 238 199
98 126 83 55 274 2610 126 95
146 125 122 66 283 204 187 117
144 207 28 255 98 112 11 91
249 138 321 119 116 95 174 113
224 274 130 157 142 159 82 82
402 204 231 157 273 142 176 137

```

Each sequence name is followed by 2-d matrix of frequency values for 3 letter words shown above as calculated from CGR map.

Thus, one can find frequency of occurrence of any k-letter word in sequences using this tool. As shown in above example CCC has frequency of 78 in given input sequence denoted by Subtype__A

Following is the screenshot of the file for example data.



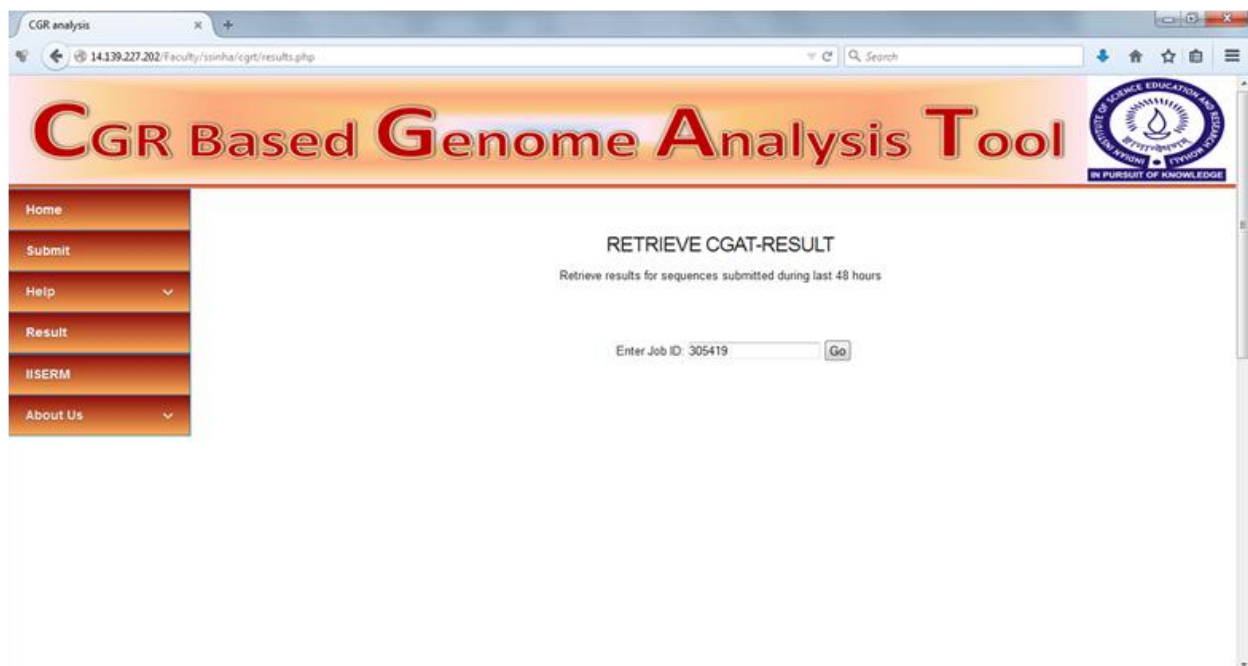
4) Outtree and Outfile – CGAT also lets users download outfile and outtree files generated by Phylip. These files contains information about trees generated in standard NEWICK format.

These files are compatible with standard bioinformatics tools like TreeView , MEGA etc. to create, view , edit and customize trees.

5) NJ Tree- The tree generated by the Notung program using the Phylip Outtree input can be downloaded in PNG format.

Retrieval of old results

The results of each analysis will be saved on the server for 48 hours after the analysis has been performed and these can be retrieved using the unique job id assigned at the time of analysis. For retrieving the results using job id, go the "Results" tab and submit the job id in the text box. Following is the screenshot of retrieval of the analysis performed on example data.



Contact us –

For further queries, suggestions and any technical issues regarding CGAT, send an email at [thind.amarinder\[at\]gmail.com](mailto:thind.amarinder[at]gmail.com)