

Bioinformatics Assignment

Anand.M.P

B120361CS

S8, B.Tech Computer Science

Question 1:

Create an implementation for the Global Sequence Alignment algorithm.

Input:

It takes three inputs; two DNA sequences, and one scoring criteria, which would determine how the matches, mismatches, insertions, and deletions should be scored.

Output:

One of the plausible alignments and the corresponding score of the alignment

Solution:

The code for the above problem is attached with the mail.

Question 2:

Create an implementation for the Exon Chaining Problem.

Input:

A set of exons is provided as the input. Each input is given in the following format,

L R W

Where 'L', represents the left index of the interval, 'R' the right index of the interval, and 'W' the weight of the interval.

Output:

The maximum score that can be obtained by using non-overlapping exon sequences.

Solution:

The code for the above problem is attached with the mail.

Question 3:

Study of Clustal Omega :

Use the Clustal Omega tool to carry out Multiple Sequence Alignment, and generate a phylogram tree of 7 hemoglobin sequences, downloaded from UniPort, of 7 different organisms. Perform two different test cases and analyze the results obtained. Create a step by step documentation of the same.

Answer:

The two different test sets used for performing the activity have been listed below.

Test case 1:

1. Humans - *Homo sapiens*, Hemoglobin subunit beta
2. Rat - *Rattus norvegicus*, Hemoglobin subunit beta-1
3. Mouse - *Mus musculus*, Hemoglobin subunit beta-1
4. Rabbit - *Oryctolagus cuniculus*, Hemoglobin subunit beta-1/2
5. Sperm Whale - *Equus caballus*, Hemoglobin subunit alpha
6. Horse - *Physeter catodon*, Hemoglobin subunit beta-1/2
7. Chimpanzee - *Pan troglodytes*, Hemoglobin subunit alpha

Test case 2:

1. Cat - *Felis catus*, Hemoglobin subunit beta-A/B
2. Olive baboon - *Papio Anubis*, Hemoglobin subunit beta
3. Sheep - *Ovis aries*, Hemoglobin subunit beta
4. Atlantic cod - *Gadus morhua*, Hemoglobin subunit beta
5. Chicken - *Gallus gallus*, Hemoglobin subunit alpha-2
6. Golden hamster - *Mesocricetus auratus*, Hemoglobin subunit beta
7. Australian ghost bat- *Macroderma gigas*, Hemoglobin subunit alpha-1/2

Procedure:

1. Hemoglobin sequence for different organisms were collected from UniProt(www.uniprot.org).

UniProtKB results

Filter by: Reviewed (1,351), Unreviewed (14,159), Popular organisms (Human, Mouse, Zebrafish, Rat, Bovine), Search terms (Filter "hemoglobin" as: disease, gene name, gene ontology, keyword).

Entry	Entry name	Protein names	Gene names	Organism	Length
<input checked="" type="checkbox"/> P09905	HBB_PHYCD	Hemoglobin subunit beta-1/2	HBB	Physeter catodon (Sperm whale) (Physeter macrocephalus)	146
<input type="checkbox"/> P73925	TRHBN_SYNY3	Group 1 truncated hemoglobin GlnN	glnN slr2097	Synechocystis sp. (strain PCC 6803 / Kazusa)	124
<input type="checkbox"/> Q67XG0	GLB3_ARATH	Two-on-two hemoglobin-3	GLB3 At4g32690,F4D11.110	Arabidopsis thaliana (Mouse-ear cress)	175
<input checked="" type="checkbox"/> P01958	HBA_HORSE	Hemoglobin subunit alpha	HBA	Equus caballus (Horse)	142
<input type="checkbox"/> P15163	HBA_LEPWE	Hemoglobin subunit alpha-1/2		Leptonychotes weddellii (Weddell seal)	141
<input type="checkbox"/> P08258	HBA_MANSP	Hemoglobin subunit alpha-1/2		Mandrillus sphinx (Mandrill) (Papio sphinx)	141
<input type="checkbox"/> Q9NZD4	AHSP_HUMAN	Alpha-hemoglobin-stabilizing protein	AHSP EDRF,ERAF	Homo sapiens (Human)	102
<input type="checkbox"/> P02074	HBB_ODOVI	Hemoglobin subunit beta-3	HBB	Odocoileus virginianus virginianus (Virginia white-tailed deer)	145
<input type="checkbox"/> P02075	HBB_SHEEP	Hemoglobin subunit beta	HBB	Ovis aries (Sheep)	145
<input checked="" type="checkbox"/> P69907	HBA_PANTR	Hemoglobin subunit alpha	HBA1	Pan troglodytes (Chimpanzee)	142

Figure 1 Hemoglobin sequences at UniProt website

2. Select the 7 hemoglobin sequences for the first test case and click download. This will download all the 7 sequences in FASTA format, which will be then used as input for Clustal Omega tool.

UniProtKB results

Filter by: Reviewed (1,351), Unreviewed (14,159), Popular organisms (Human, Mouse, Zebrafish, Rat, Bovine), Search terms (Filter "hemoglobin" as: disease, gene name, gene ontology, keyword).

Entry	Entry name	Protein names	Gene names	Organism	Length
<input checked="" type="checkbox"/> P09905	HBB_PHYCD	Hemoglobin subunit beta-1/2	HBB	Physeter catodon (Sperm whale) (Physeter macrocephalus)	146
<input type="checkbox"/> P73925	TRHBN_SYNY3	Group 1 truncated hemoglobin GlnN	glnN slr2097	Synechocystis sp. (strain PCC 6803 / Kazusa)	124
<input type="checkbox"/> Q67XG0	GLB3_ARATH	Two-on-two hemoglobin-3	GLB3 At4g32690,F4D11.110	Arabidopsis thaliana (Mouse-ear cress)	175
<input checked="" type="checkbox"/> P01958	HBA_HORSE	Hemoglobin subunit alpha	HBA	Equus caballus (Horse)	142
<input type="checkbox"/> P15163	HBA_LEPWE	Hemoglobin subunit alpha-1/2		Leptonychotes weddellii (Weddell seal)	141
<input type="checkbox"/> P08258	HBA_MANSP	Hemoglobin subunit alpha-1/2		Mandrillus sphinx (Mandrill) (Papio sphinx)	141
<input type="checkbox"/> Q9NZD4	AHSP_HUMAN	Alpha-hemoglobin-stabilizing protein	AHSP EDRF,ERAF	Homo sapiens (Human)	102
<input type="checkbox"/> P02074	HBB_ODOVI	Hemoglobin subunit beta-3	HBB	Odocoileus virginianus virginianus (Virginia white-tailed deer)	145
<input type="checkbox"/> P02075	HBB_SHEEP	Hemoglobin subunit beta	HBB	Ovis aries (Sheep)	145
<input checked="" type="checkbox"/> P69907	HBA_PANTR	Hemoglobin subunit alpha	HBA1	Pan troglodytes (Chimpanzee)	142

Download selected (3) / Download all (15510)

Format: FASTA (canonical) / Compressed / Uncompressed

Preview first 10

Figure 2 Download Hemoglobin Sequence

- At the Clustal Omega website (<http://www.ebi.ac.uk/Tools/msa/clustalo/>), the sequence obtained from UniProt is added as the input. The default settings of the tool were left unchanged. Click the submit button to start processing the input.

The screenshot shows the Clustal Omega web interface. The browser address bar displays www.ebi.ac.uk/Tools/msa/clustalo/. The page title is "Multiple Sequence Alignment". Below the title, a description states: "Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#)."

STEP 1 - Enter your input sequences

Enter or paste a set of **PROTEIN** sequences in any supported format:

```
VHLTGEKSGLTALWAKVNVVEIGGEALGRLLVYPWTORFFEHFGDLSTADAVMKNPKV
KKHGQKVLASFGEGLKHLNKGTFATLSELHCDKLHVDPENFRLLGNVLVVVLARHFGK
EFTPELOTAYQKVVAGVANALAHKYH
>sp|P6907|HBA_PANTR Hemoglobin subunit alpha OS=Pan troglodytes GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAHWGVGAHAGEYGAELRMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAAVHDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFT
AVHASLDKFLASVSTVLTSKYR
```

Or, upload a file: No file chosen

STEP 2 - Set your parameters

OUTPUT FORMAT: **Clustal w/o numbers**

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

If you plan to use these services during a course please [contact us](#).

Please read the [FAQ](#) before seeking help from our support staff.

Figure 3 Clustal Omega welcome screen

- Wait for some time while the clustal omega tool processes the input.

The screenshot shows the Clustal Omega job result page. The browser address bar displays www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20160417-085900-0577-54064452-pg. The page title is "Clustal Omega". The navigation bar includes "Input form", "Web services", and "Help & Documentation". The main content area displays:

Your job is currently running... please be patient

The result of your job will appear in this browser window.

Job ID: [clustalo-I20160417-085900-0577-54064452-pg](#)

Please note the following

- You may press Shift+Refresh or Reload on your browser at any time to check if results are ready.
- You may bookmark this page to view your results later if you wish.
- Results are stored for 7 days.

5. The Multiple Sequence Alignment (MSA) for the input sequences is generated as the output along with the phylogenetic tree.

```

CLUSTAL O(1.2.1) multiple sequence alignment

sp|P01958|HBA_HORSE      -MVLSAADKTNVKAASKVGGHAGEYGAELERMF LGFP TTKTYFPHF DLS-----HGS
sp|P69907|HBA_PANTR      -MVLSPADKTNVKAAGKVGAHAGEYGAELERMF LSFPTTKTYFPHF DLS-----HGS
sp|P02091|HBB1_RAT       MVHLTDAEKA AVNGLWGK VNP--DDVGGEALGRLLVVYPW TQRYFDSFGD LSSASAIMGN
sp|P02088|HBB1_MOUSE     MVHLTDAEKA AVSC LWGKVNS--DEVGGEALGRLLVVYPW TQRYFDSFGD LSSASAIMGN
sp|P09905|HBB_PHYCD      -VHLTGEESGLTALWAKVNV--EEIGGEALGRLLVVYPW TQRF EFHFGD LSTADAVMKN
sp|P68871|HBB_HUMAN      MVHLTPEEKSA VTALWGK VNV--DEVGGEALGRLLVVYPW TQRF EFSGD LSTPAVMGN
sp|P02057|HBB_RABIT      MVHLSSEKSA VTALWGK VNV--EEVGGEALGRLLVVYPW TQRF EFSGD LSSANAVMMN
      : *:  *:  :.  **.      : *.*** *:  : *:  *:  *

sp|P01958|HBA_HORSE      A QVKAGHKGVGDALTAVGH LDDLPGALSNLSDLHAHKLRVDPVNF KLLSHCLLSTLAVH
sp|P69907|HBA_PANTR      A QVKG HKGVADALTNAVGH LDDMPNALSALD LHAHKLRVDPVNF KLLSHCLLVTLAAH
sp|P02091|HBB1_RAT       P KVKAGHKGVINAFNDGLKHL DNLKGTFAHLS ELHCDKLHVDPENFRLLGNMIVIVLGHH
sp|P02088|HBB1_MOUSE     A KVKAGHKGVITAFNDGLNHL DLSLKGTFASL ELHCDKLHVDPENFRLLGNMIVIVLGHH
sp|P09905|HBB_PHYCD      P KVKKHGQKVLASFGEGLKHL DNLKGT FATLS ELHCDKLHVDPENFRLLGNLVVVLAH
sp|P68871|HBB_HUMAN      P KVKAGHKQVLGAFSDGLAHL DNLKGT FATLS ELHCDKLHVDPENFRLLGNVLVCLAH
sp|P02057|HBB_RABIT      P KVKAGHKQVLAASFGLSHLD NNLKGTFAKLS ELHCDKLHVDPENFRLLGNLVIVLSH
      : ** *: **:  :.  : *:  :.  : : : *.**..*:*:* *:*.. :.  : .

sp|P01958|HBA_HORSE      LPNDFTPAVHASLDKFLSSVSTVLTSKYR
sp|P69907|HBA_PANTR      LPAEFTPAVHASLDKFLASVSTVLTSKYR
sp|P02091|HBB1_RAT       LGKEFTPCAQA AFQKV VAGVASALAHKYH
sp|P02088|HBB1_MOUSE     LGKDFTPAQAQAFQKV VAGVATALAHKYH
sp|P09905|HBB_PHYCD      FGKEFTP ELQTAYQKV VAGVANALAHKYH
sp|P68871|HBB_HUMAN      FGKEFTP PVQAAYQKV VAGVANALAHKYH
sp|P02057|HBB_RABIT      FGKEFTP QVQAAYQKV VAGVANALAHKYH
      : :*** : : :*:  :.  :*:  :*:  :*

```

Figure 4 Multiple Sequence Alignment for test case 1

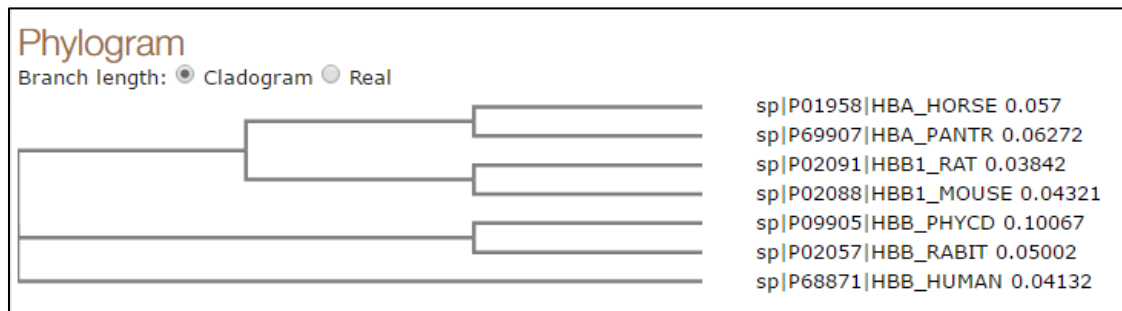
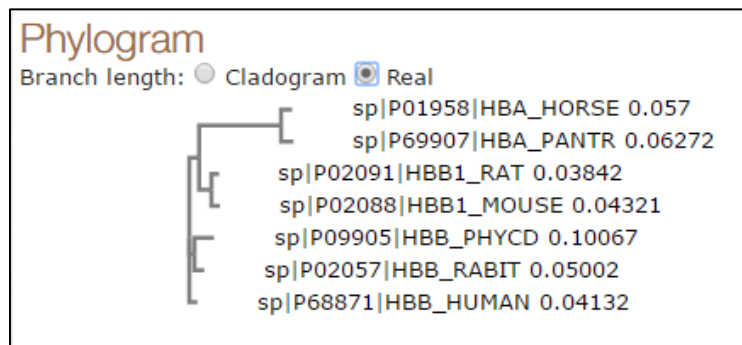


Figure 5 Phylogram for test case 1



6. Similar steps were performed for test case 2 and the following results were obtained.



Figure 6 Multiple Sequence Alignment for test case 2

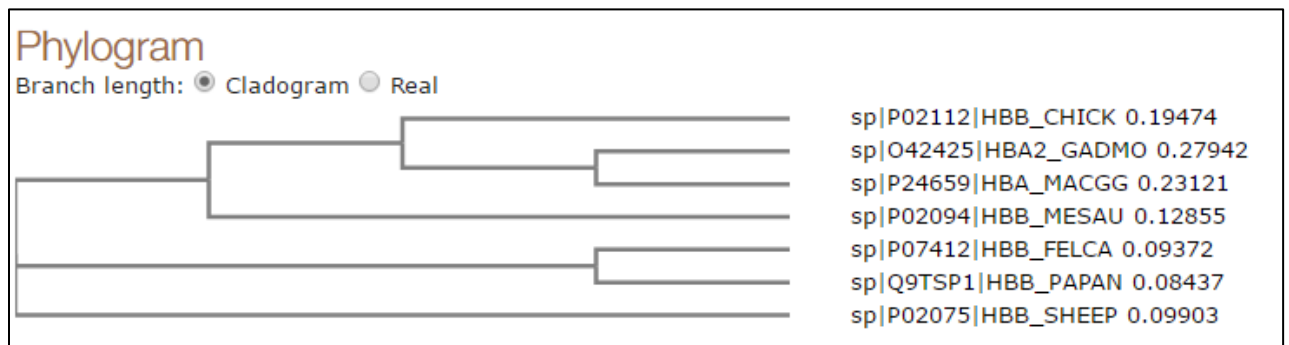
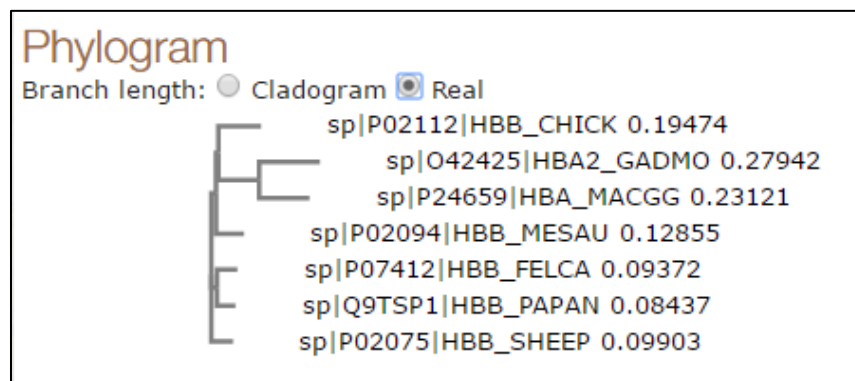


Figure 7 Phylogram for test case 2



7. Some of the Input Parameters used by the Clustal Omega for generating the output are as follows:

Input Parameters	
program	clustalo
version	1.2.1
Output guide tree	false
Output distance matrix	false
Dealign input sequences	false
mBed-like clustering guide tree	true
mBed-like clustering iteration	true
Number of iterations	0
Maximum guide tree iterations	-1
Maximum HMM iterations	-1
Output alignment format	clustal
Output order	

Figure 8 Input Parameters used by Clustal Omega

Observations:

Test Case 1:

It can be observed from the phylogram obtained from the test set 1 that humans are least similar to the rest of the organisms in that particular set. Whereas, rat and mouse are the most similar as it would have expected. This clearly indicates that they are closely related to each other.

Test Case 2:

It can be observed from the phylogram obtained from the test set 2 that sheep is least similar to the rest of the organisms in that particular set. Whereas, cat and olive baboon were found have great amount of similarities, indicating that they are closely related.

Question 4:

Collect the chromosome 4 sequence of humans from NCBI website and use this sequence to perform a Genscan and analyze the results obtained.

Procedure:

1. The chromosome 4 sequence was collected from the NCBI website. Taking into consideration the enormous size of the sequence, a small portion of it was later used as the input for Genscan.

The screenshot displays the NCBI website interface. The browser's address bar shows the URL www.ncbi.nlm.nih.gov/nuccore/NC_018915.2. The page title is "Homo sapiens chromosome 4, alternate assembly CHM1_1.1, whole genome shotgun sequence". The NCBI Reference Sequence is NC_018915.2. The page includes a "FASTA" link and a "Graphics" link. The "Go to" section lists various links: Locus, Definition, Accession, Version, Dblink, Keywords, Source, Organism, and Comment. The "Related information" section lists links: Assembly, BioProject, BioSample, Components (Core), Full text in PMC, Gene, Genome, HomoloGene, Identical GenBank Sequence, Map Viewer, and Protein. The "Change region shown" and "Customize view" sections are also visible.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Advanced Help

NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! [Read more...](#)

GenBank Send Change region shown Customize view Analyze this sequence Run BLAST Pick Primers Related information Assembly BioProject BioSample Components (Core) Full text in PMC Gene Genome HomoloGene Identical GenBank Sequence Map Viewer Protein

Due to the size of this record, annotated features are not shown by default. [Display features.](#)

Homo sapiens chromosome 4, alternate assembly CHM1_1.1, whole genome shotgun sequence

NCBI Reference Sequence: NC_018915.2

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS NC_018915 191040880 bp DNA linear CON 12-MAR-2015

DEFINITION Homo sapiens chromosome 4, alternate assembly CHM1_1.1, whole genome shotgun sequence.

ACCESSION NC_018915 GPC_000001163

VERSION NC_018915.2 GI:528476651

DBLINK BioProject: [PRJNA178030](#)

BioSample: [SAMN02205338](#)

Assembly: [GCF_000306695.2](#)

KEYWORDS WGS; RefSeq.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

COMMENT [REFSEQ INFORMATION](#): The reference sequence is identical to [CH001612.2](#).

On Aug 13, 2013 this sequence version replaced gi:409253460.

Assembly name: CHM1_1.1

Figure 9 NCBI website

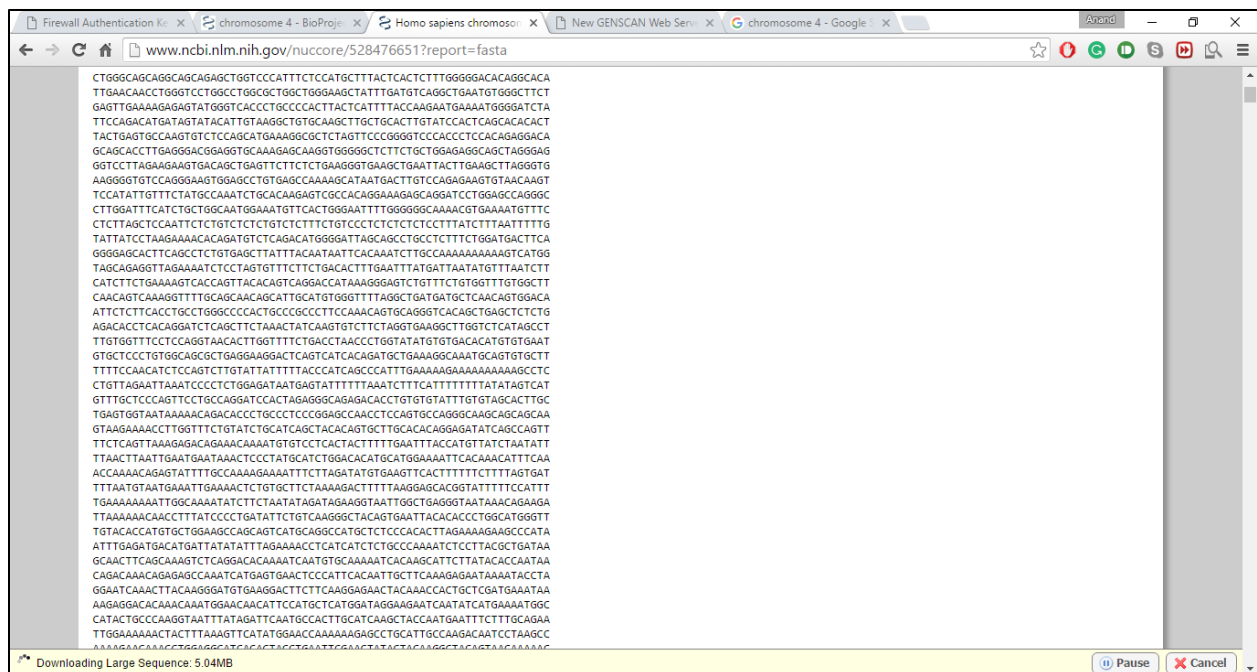


Figure 10 Downloaded chromosome 4 sequence

2. A small portion of the downloaded sequence was used to create a text file, which was later uploaded to Genscan website for processing. The default settings of the website were left unchanged. File was given the name chromosome_4_seq_1.txt. Run GENSCAN was clicked to start processing the input file.

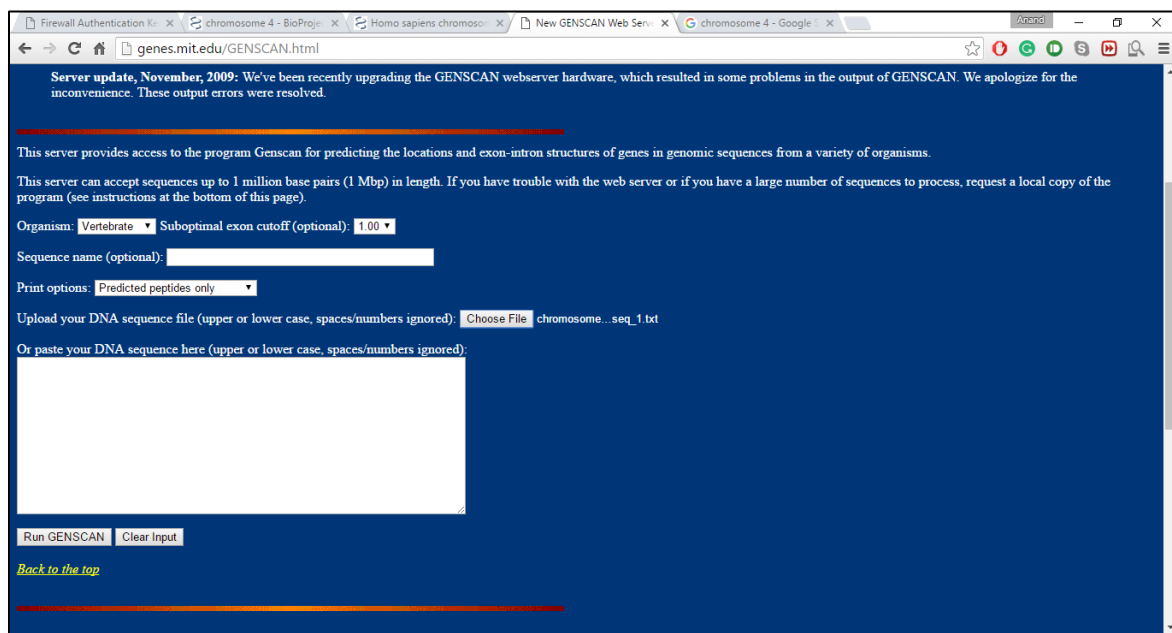


Figure 11 Genscan website

3. Following results were generated by by genscan after the processing was over. Predictions for a new set of genes or exons, and peptide sequences were provided as the output.

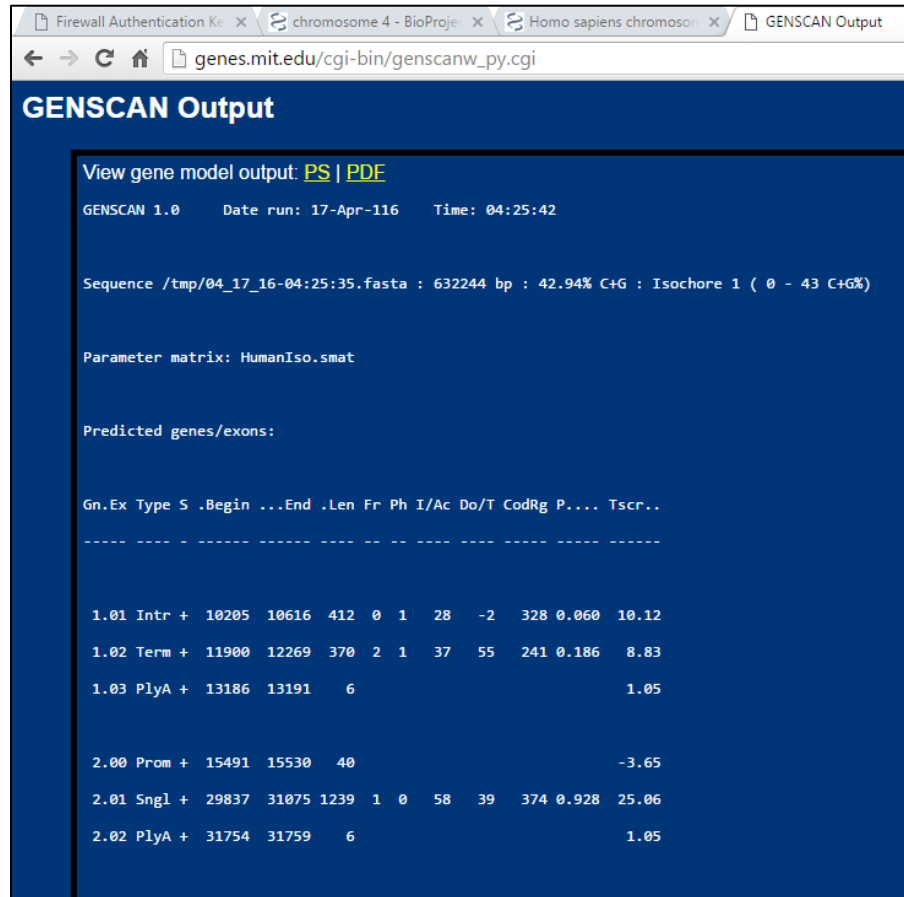


Figure 12 Predicted Genes or Exons

```
Firewall Authentication K x chromosome 4 - BioProj x Homo sapiens chromosc x GE
genes.mit.edu/cgi-bin/genscanw_py.cgi

36.05 Intr - 627928 627883 46 0 1 99 50 64 0.000 1.19
36.04 Intr - 629328 629179 150 2 0 64 110 86 0.000 6.76
36.03 Intr - 629644 629374 271 2 1 21 55 221 0.000 7.78
36.02 Intr - 629855 629725 131 1 2 99 99 34 0.000 5.02
36.01 Init - 631810 631167 644 1 2 61 58 430 0.022 32.14

Suboptimal exons with probability > 1.000

Exnum Type S .Begin ...End .Len Fr Ph B/Ac Do/T CodRg P.... Tscr..
-----
NO EXONS FOUND AT GIVEN PROBABILITY CUTOFF

Predicted peptide sequence(s):

>/tmp/04_17_16-04:25:35.fasta|GENSCAN_predicted_peptide_1|260_aa
XCSSVIAFPKSLQRKTELLLRDALQVCAEENAAPPSQRHSAIAGAENKVGAAQAQRKTT
ARPWGARRRPREACHRGAGAWGAAQTQRRTPAQRRDGSAAQAQTRTAANRLAGGATQAQT
HTAARRHDGTPRRRTDGCVVAVGSPAGSWGHCRAALLAVYNGHAPFWPLGALQDALAHSV
```

Figure 13 Predicted Peptide Sequence(s)