BIOINFORMATICS
MBChB/Bpharm/BDS Level II

Lecture 1

Dr Victor Mobegi

Introduction to Bioinformatics

- Bioinformatics is the storage, retrieval and analysis of biological data
- Another definition: computational techniques for solving biological problems
- Computers and internet resources can maximize the biological information available to a researcher.
- This can not only make the work of a researcher easier and more productive but also enable one to answer biological questions that would be impossible without electronic help e.g. Sequence analyses
- This course will introduce you to the more commonly used bioinformatics tools and resources.

Databases

- A Database is a structured collection of information.
- Databases provide the means for automated storage, retrieval and sharing of large volumes of data.
- This includes literature and molecular databases.
- Consists of basic units called records or entries.
- Each record consists of fields, which hold pre-defined data related to the record.
- For example, a protein database would have protein entries as records and protein properties as fields (e.g., name of protein, length, amino-acid sequence)

Qualities of an ideal Database

- Comprehensive, but easy to search.
- Should be annotated.
- A simple, easy to understand structure.
- Cross-referenced.
- Minimum redundancy.
- Easy retrieval of data.

Why use Databases

- Huge amount of data is being generated in experiments including high-throughput genomics, proteomics and metabolomics
- Need for storing and sharing large datasets has grown tremendously
- Archiving, curation, analysis and interpretation of all of these datasets are a challenge
- Convenient methods for proper storing, searching & retrieving are necessary
- Databases provide the means for automated storage, retrieval and sharing of these large volumes of data

Types of databases

- Literature databases: PubMed, Google Scholar, OMIM
- Molecular databases providing genomic, proteomic and metabolomic data

  -Nucleotide sequence databases: GenBank, EMBL, DDBJ

  -Protein databases: Swiss-Prot, Genpept, PROSITE, PDB
- Molecular databases providing metabolic pathways data e.g. KEGG, MetaCyc, Reactome
- Specialized databases: EuPathDB, TTD

How to find a database

- **Database Journals**

Database: The Journal of Biological Databases and Curation

http://database.oxfordjournals.org/

- **Nucleic Acids Research** offers **Database Issue** every year

http://nar.oxfordjournals.org/

- **Database portals**

DBD (Database of Biological Databases)

http://www.biodbs.info/

Types of sequence databases

- Can be broadly divided into 2 classes: Primary databases and secondary databases
- Primary Databases

– Original submissions by experimentalists

– Content controlled by the submitter

Examples: GenBank, Trace Archive, SRA

- Secondary Databases

– Derived from primary data

– Contain results from the analysis of the sequences in the primary databases

– Content controlled by third party

Examples: Refseq, Pfam, PROSITE, RefSNP

Nucleotide sequence databases

- Databases are the core resource for bioinformatics and many programs access databases of information.
- Frequently used classes are the biological sequence (nucleotide & protein) databases.
- Nucleotide sequence databases include:

-GenBank

http://www.ncbi.nlm.nih.gov/Genbank/

-European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena) previously EMBL (European Molecular Biol Lab)

http://www.ebi.ac.uk/embl/

-DDBJ (DNA Data Bank of Japan)

http://www.ddbj.nig.ac.jp/

- Entries in the GenBank, EMBL and DDBJ databases are synchronized on a daily basis and so should be identical in content.

- Accession numbers are managed in a consistent manner
- Each of these databases consists of entries, each consisting of a single sequence preceded by annotation that puts the sequence in its biological, functional and historical context.
- However, the format of the records in these databases is different
- The growth of the DNA sequence data has been phenomenal.
- When using databases it is important to note which release of the sequence databases were used.
- Because of enormous size of the databases, to ease management they are now broken up into divisions.
- Most of these divisions are organised on the basis of taxonomy (Prokaryotes, plants, fungi, mammals etc.)
- These divisions are useful as they make it easier to search only in the relevant part of the database.

Accession numbers

- Accession numbers are used as unique and unchanging numbers
- GenBank/EMBL accession numbers were originally a letter followed by 5 digits (e.g. X32152, M22239).
- When the number of sequences in databases increased, accession numbers format changed to 2 letters followed by 6 digits (e.g. AL234556, BF345788)
- RefSeq (NCBI's reference sequence database) accession numbers format: 2 letters, and underscore, and 6 digits e.g NM_000492
- Sequences in database are updated, corrected and merged giving different versions of the sequence
- Version numbers are appended to the accession number after a dot e.g. NM_000492.2

Sample GenBank record

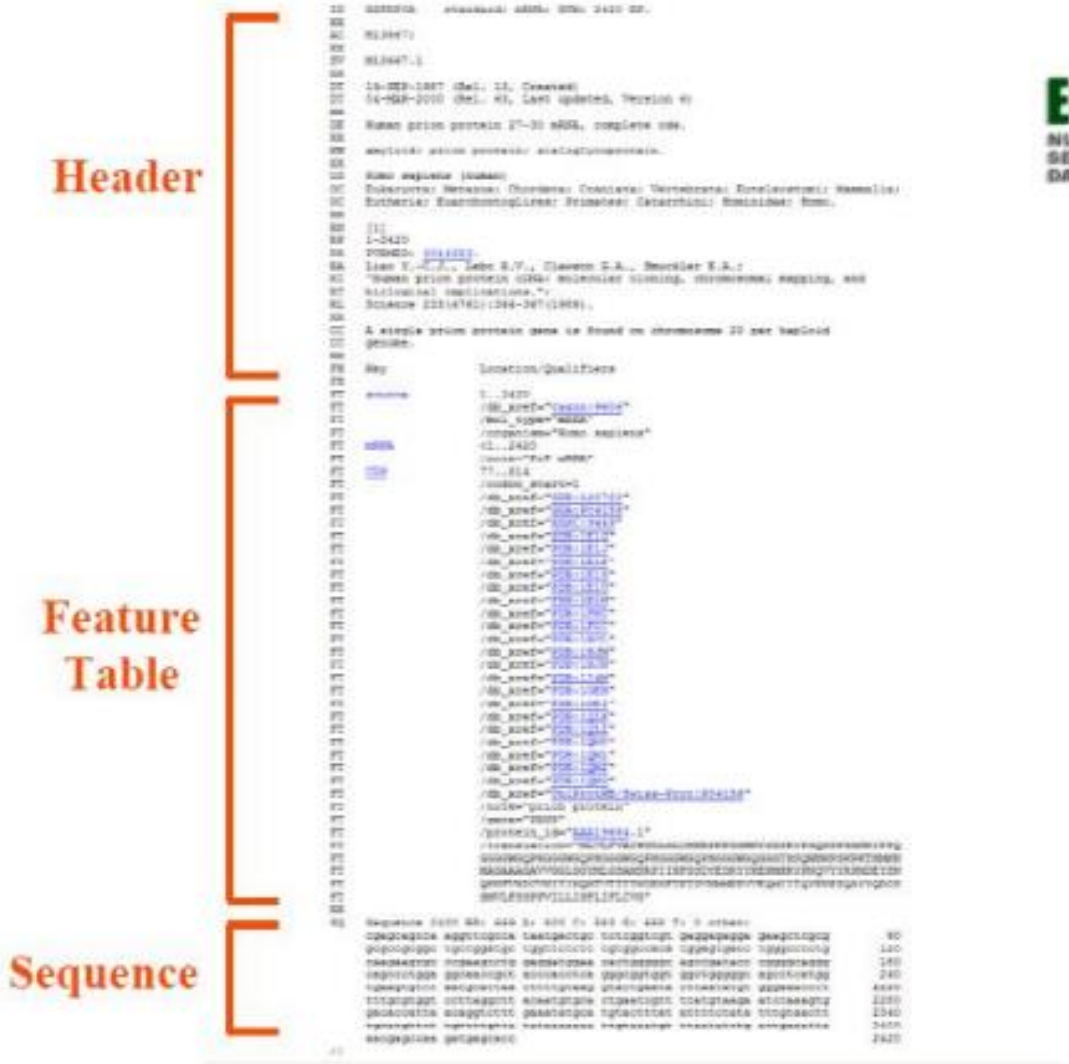
Header

Feature Table

Sequence

```
LOCUS       HUMPRPOA                2420 bp    mRNA    linear   PRI 19-JUL-1994
DEFINITION  Human prion protein 27-30 mRNA, complete cds.
ACCESSION   M13667
VERSION     M13667.1  GI:190466
KEYWORDS    amyloid; prion protein; sialoglycoprotein.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini;
            Hominidae; Homo.
REFERENCE   1  (bases 1 to 2420)
  AUTHORS   Liao,Y.C., Lebo,R.V., Clawson,G.A. and Smuckler,E.A.
  TITLE     Human prion protein cDNA: molecular cloning, chromosomal mapping,
            and biological implications
  JOURNAL   Science 233 (4761), 364-367 (1986)
   PUBMED   3014653
COMMENT     Original source text: Human, cDNA to mRNA, clones lambda [3,6,7].
            A single prion protein gene is found on chromosome 20 per haploid
            genome.
FEATURES             Location/Qualifiers
     source          1..2420
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
     gene            1..2420
                     /gene="PRNP"
     mRNA            <1..2420
                     /gene="PRNP"
                     /product="PrP mRNA"
     CDS             77..814
                     /gene="PRNP"
                     /note="prion protein"
                     /codon_start=1
                     /protein_id="AAA19664.1"
                     /db_xref="GI:190470"
                     /translation="MLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPP
                     QGGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNM
                     KHMAGAAAAGAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDE
                     YSNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYY
                     QRGSSMVLFSSPPVILLISFLIFLIVG"
ORIGIN      171 bp upstream of SmaI site; chromosome 20.
        1 cgagcagcca aggttcgcca taatgactgc tctcggtcgt gaggagaggg gaagctcgcg
       61 gcgccgcggc tgctggatgc tggttctctt tgtggccacc tggagtgacc tgggcctctg
      121 caagaagcgc cccgagcctg gaggatggaa cactgggggc agccgatacc cggggcaggg
      181 cagccctgga ggcaacccct acccacctca gggcggtggt ggctgggggc agcctcatgg
      241 tggtggctgg gggcagcctc atggtggtgg ctgggggcag cccccatggtg gtggctgggg
      301 acagcctcat ggtggtggct ggggtcaagg aggtggcacc cacagtcagt ggaacaagcc
     2161 tgaagtgtct aatgcattaa cttttgtaag gtactgaata cttaatatgt gggaaacccct
     2221 tttgcgtggt ccttaggctt acaatgtgca ctgaatcgtt tcatgtaaga atccaaagtg
     2281 gacaccatta acaggtcttt gaaatatgca tgtactttat attttctata tttgtaactt
     2341 tgcatgttct tgttttgtta tataaaaaaa ttgtaaatgt ttaatatctg actgaaatta
     2401 aacgagccaa gatgagcacc
//
```

Sample EMBL record



EMBL
NUCLEOTIDE SEQUENCE DATABASE

a) GenBank (or DDBJ) flat file format

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
```

Header

b) EMBL flat file format

```
ID   U49845; SV 1; linear; genomic DNA; STD; FUN; 5028 BP.
XX
AC   U49845;
XX
DT   07-MAY-1996 (Rel. 47, Created)
DT   25-MAR-2010 (Rel. 104, Last updated, Version 5)
XX
DE   Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and
DE   Rev7p (REV7) genes, complete cds.
XX
KW   .
XX
OS   Saccharomyces cerevisiae (baker's yeast)
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
XX
RN   [1]
RP   1-5028
RX   PUBMED; 8846915.
RA   Roemer T., Madden K., Chang J., Snyder M.;
RT   "Selection of axial growth sites in yeast requires Axl2p, a novel plasma
RT   membrane glycoprotein";
RL   Genes Dev. 10(7):777-793(1996).
XX
RN   [2]
RP   1-5028
RA   Roemer T.;
RT   ;
RL   Submitted (22-FEB-1996) to the INSDC.
RL   Biology, Yale University, New Haven, CT 06520, USA
```

Header

DNA sequence formats

- There are a number of ways you can write, store and transmit simple one-dimensional sequence files. Some commonly used sequence file formats are shown below:

- 1) Plain sequence format

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

Sequence formats

- 2) FASTA format

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

- Start of sequence marked by line starting with '>'

- 3) EMBL format

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                                368
//
```

- Start of sequence marked by line starting with 'SQ'
- End of sequence marked by two slashes (//).

Sequence formats

- 4) GenBank format

```
LOCUS       AB000263                 368 bp    mRNA    linear   PRI 05-FEB-1999
DEFINITION  Homo sapiens mRNA for prepro cortistatin like peptide, complete
            cds.
ACCESSION   AB000263
ORIGIN
        1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
       61 ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
      121 caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
      181 aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
      241 gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
      361 gacctgaa
//
```

- Start of the sequence marked by word ORIGIN
- End of sequence marked by two slashes (//)

Protein Databases

- With increasing number of completed genome sequences from both eukaryotic and prokaryotic organisms, it is important to focus on gene products
- Variety of protein sequence databases have grown up to reflect the huge amount of data generated from the large scale analysis of these gene products.
- Protein sequence databases can be divided into universal databases storing proteins from all species and specialised protein databases storing information about specific families or groups of proteins or about proteins of specific organism.
- DNA sequences deposited in the DNA sequence database (GenBank, EMBL and DDBJ) are automatically translated to produce Sequence repositories such as TrEMBL and GenPept.

- In these databases, the protein data is stored with little or no manual intervention
- This provides a more nearly comprehensive coverage of protein sequences, but at the expense of the quality of annotation.

UniProt

- The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data
- UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR)
- New protein sequence database that is the result of a merge from SWISS-PROT and PIR

- The UniProt databases are:-

  -UniProt Knowledgebase (UniProtKB)

  -UniProt Reference Clusters (UniRef)

  - UniProt Archive (UniParc)

UniProt Knowledgebase (UniProtKB)

- Consists of two sections:-

1. UniProtKB/Swiss-Prot (manually annotated)
2. UniProtKB/TrEMBL" (automatically annotated)

- TrEMBL (Translated EMBL Nucleotide Sequence Data Library)
- TrEMBL is a computer-annotated protein sequence database supplementing the SWISS-PROT Protein Sequence Data Bank.
- TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database not yet integrated in SWISS-PROT.

Source of UniProtKB protein sequences

- Data in UniProtKB is derived from the translation of the coding sequences (CDS) which have been submitted to the public nucleic acid databases, the EMBL/GenBank/DDBJ databases
- A protein identifier ("protein_id") is assigned to the translated CDS
- Protein sequences are automatically integrated into UniProtKB/TrEMBL

Eukaryotic Pathogen Database resources (EuPathDB)

- EuPathDB Bioinformatics Resource Center (http://eupathdb.org/eupathdb/) is a portal for accessing genomic-scale datasets associated with the eukaryotic pathogens in the following websites:-
- AmoebaDB: http://amoebadb.org/amoeba/
- CryptoDB: http://cryptodb.org/cryptodb/
- FungiDB: http://fungidb.org/fungidb/
- GiardiaDB: http://giardiadb.org/giardiadb/
- MicrosporidiaDB: http://microsporidiadb.org/micro/
- PlasmoDB: http://plasmodb.org/plasmo/
- ToxoDB: http://toxodb.org/toxo/
- TritrypDB: http://tritrypdb.org/tritrypdb/

National Cancer Institute (NCI) clinical trials Database

Website: http://www.nci.nih.gov/clinicaltrials/

- This is an important resource in cancer research.
- Using this resource, pharmacogenomic correlations between specific variants in genes like cellular tumor antigen p53 (TP53), BRAF, ERBBs and ATAD5 and anti-cancer agents nutlin, vemurafenib, erlotinib and bleomycin can be done.
- This data can be used to validate and generate novel hypothesis

Therapeutic Target Database (TTD)

This database can be used to:
-Search for drugs
-search for targets
-Search for biomarkers
Etc

Example:
If you search for gene target for the anti-cancer drug nutlin, it will be identified as cellular tumor antigen p53 (TP53)

Website: http://bidd.nus.edu.sg/group/cjttd/

Homology Searching

- The most widely used bioinformatics protocol is to search a database for sequences similar to a candidate sequence
- If sequences are similar at some statistically significant level they share a common ancestor
- If two sequences are similar, they are likely to have a similar structure and function
- There are several algorithms for doing homology searches against databases but the standard for homology searching is the BLAST family of programs
- Another example of homology search algorithm is FastA

BLAST

- Basic Local Alignment Search Tool (BLAST) is a family of programs carrying out different classes of search:
- **Blastn**-searches a DNA sequence against a DNA database such as EMBL or GenBank
- **Blastp**- searches a protein sequence against a protein database such as Swissprot or TrEMBL
- **Blastx**-searches a DNA sequence (translated in all 6 reading frames) against a protein database
- **tBlastn**- Searches a protein sequence against a translated nucleotide

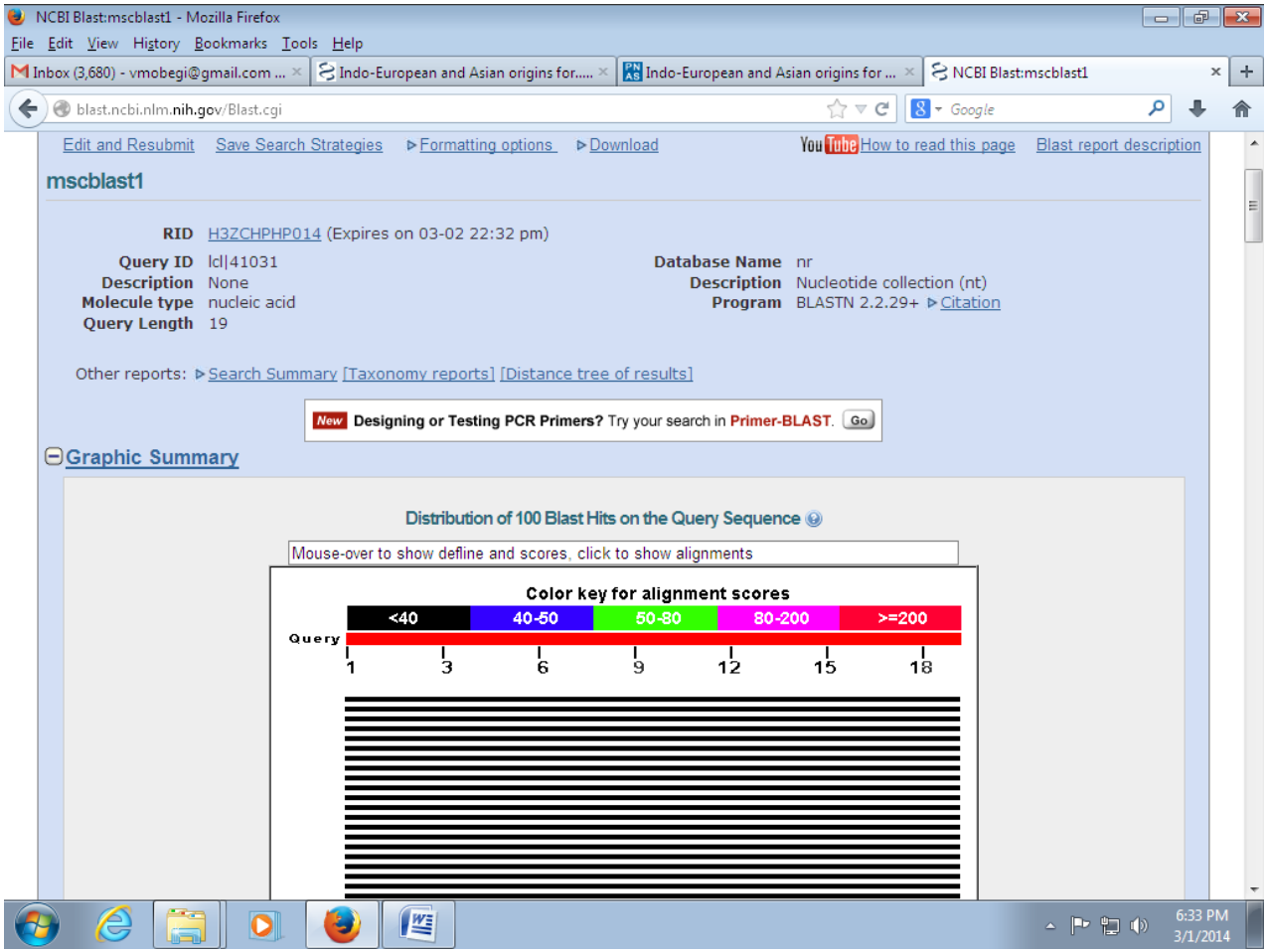
Task 1

- Go to NCBI and click on BLAST
- Select Standard Nucleotide BLAST by selecting “blastn” option
- To Enter Query Sequence copy and paste the sequence TCTATATTCCACATTTCTC
- In the Job Title type mscblast1
- Click on BLAST button
- The following results will be returned.

Part 1

-Will include information about the length of nucleotide sequence used for the search

-Graphic summary of color key for alignment scores for all hits (100 hits in this case)

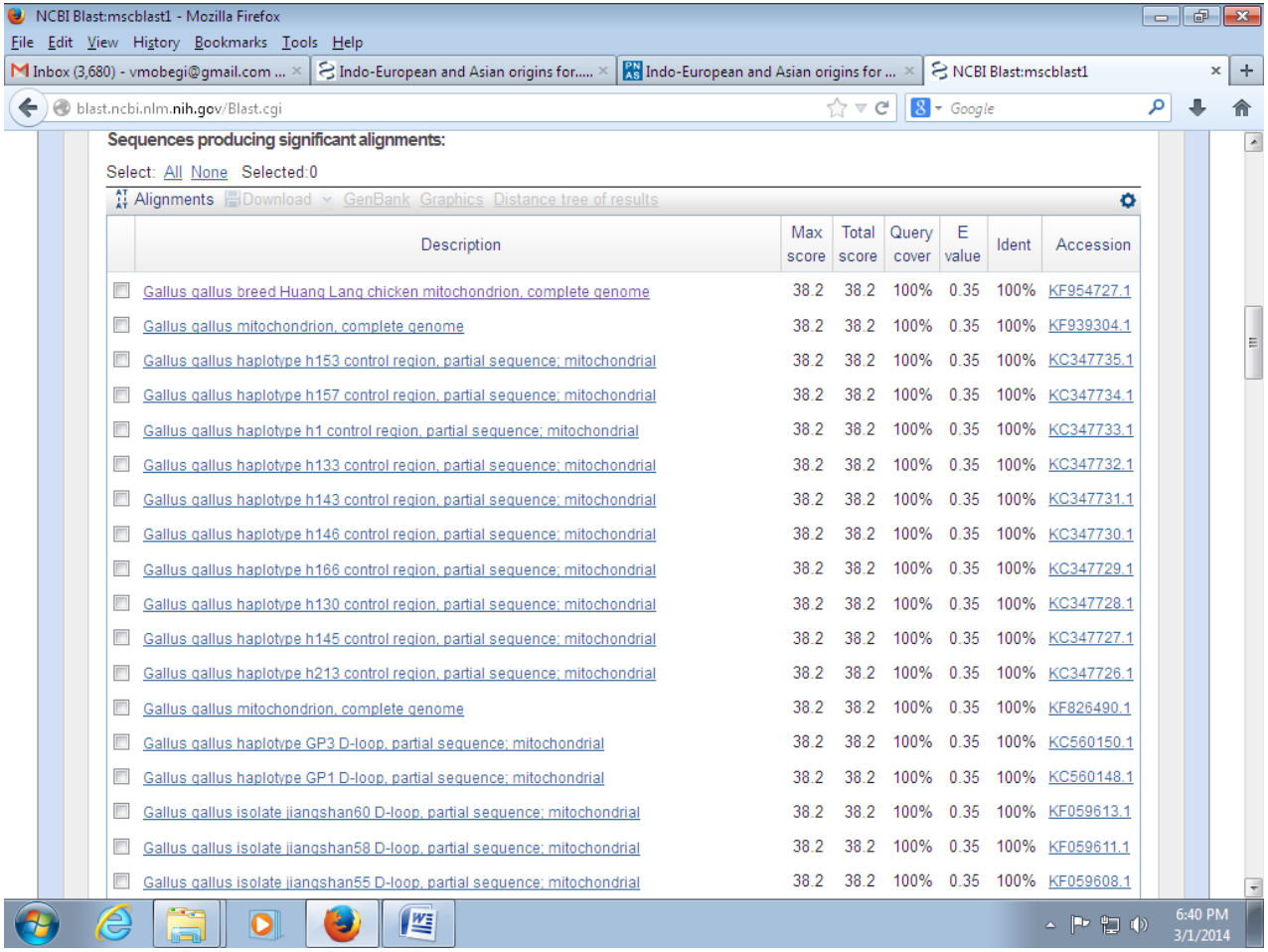


Part 2:

-Gives description of what organism and gene/genome matches the query sequence

-Gives maximum score for each alignment and Expectation (E) value

-Information of accession number of the sequence



Sequences producing significant alignments:

Select: All None Selected:0

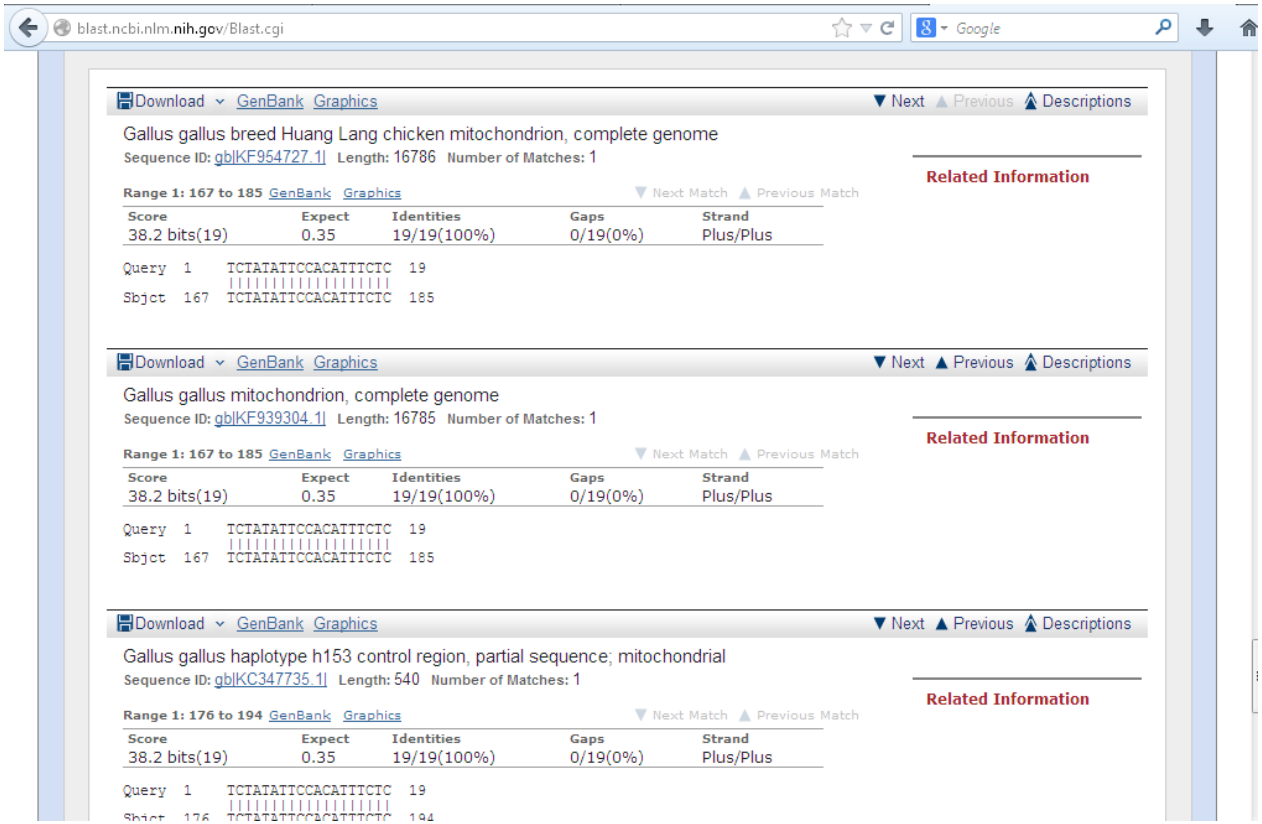
Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Gallus gallus breed Huang Lang chicken mitochondrion, complete genome	38.2	38.2	100%	0.35	100%	KF954727.1
Gallus gallus mitochondrion, complete genome	38.2	38.2	100%	0.35	100%	KF939304.1
Gallus gallus haplotype h153 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347735.1
Gallus gallus haplotype h157 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347734.1
Gallus gallus haplotype h1 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347733.1
Gallus gallus haplotype h133 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347732.1
Gallus gallus haplotype h143 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347731.1
Gallus gallus haplotype h146 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347730.1
Gallus gallus haplotype h166 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347729.1
Gallus gallus haplotype h130 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347728.1
Gallus gallus haplotype h145 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347727.1
Gallus gallus haplotype h213 control region, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC347726.1
Gallus gallus mitochondrion, complete genome	38.2	38.2	100%	0.35	100%	KF826490.1
Gallus gallus haplotype GP3 D-loop, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC560150.1
Gallus gallus haplotype GP1 D-loop, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KC560148.1
Gallus gallus isolate jiangshan60 D-loop, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KF059613.1
Gallus gallus isolate jiangshan58 D-loop, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KF059611.1
Gallus gallus isolate jiangshan55 D-loop, partial sequence; mitochondrial	38.2	38.2	100%	0.35	100%	KF059608.1

Part 3:

-Shows the actual alignment giving percentage of identities

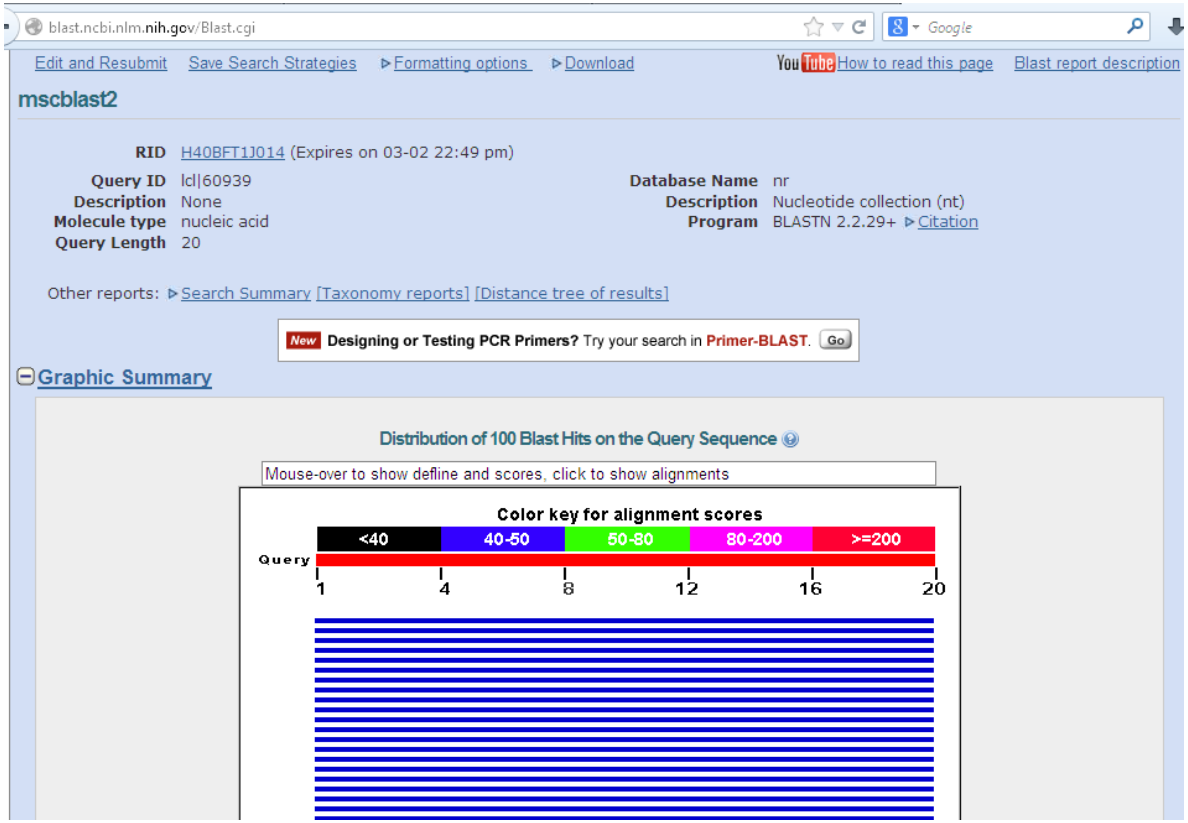
-Gives the length of the subject sequence that the query sequence matches



Task 2

- To Enter Query Sequence copy and paste the sequence AGGACTACGGCTTGAAAAGC
- In the Job Title type mscblast2
- Click on BLAST button
- The following results will be returned.
- Note the graphic color here is blue (maximum score of 40-50) whereas in task 1 it was black (<40).

Graphic for alignment scores



Task 3

- To Enter Query Sequence copy and paste the sequence

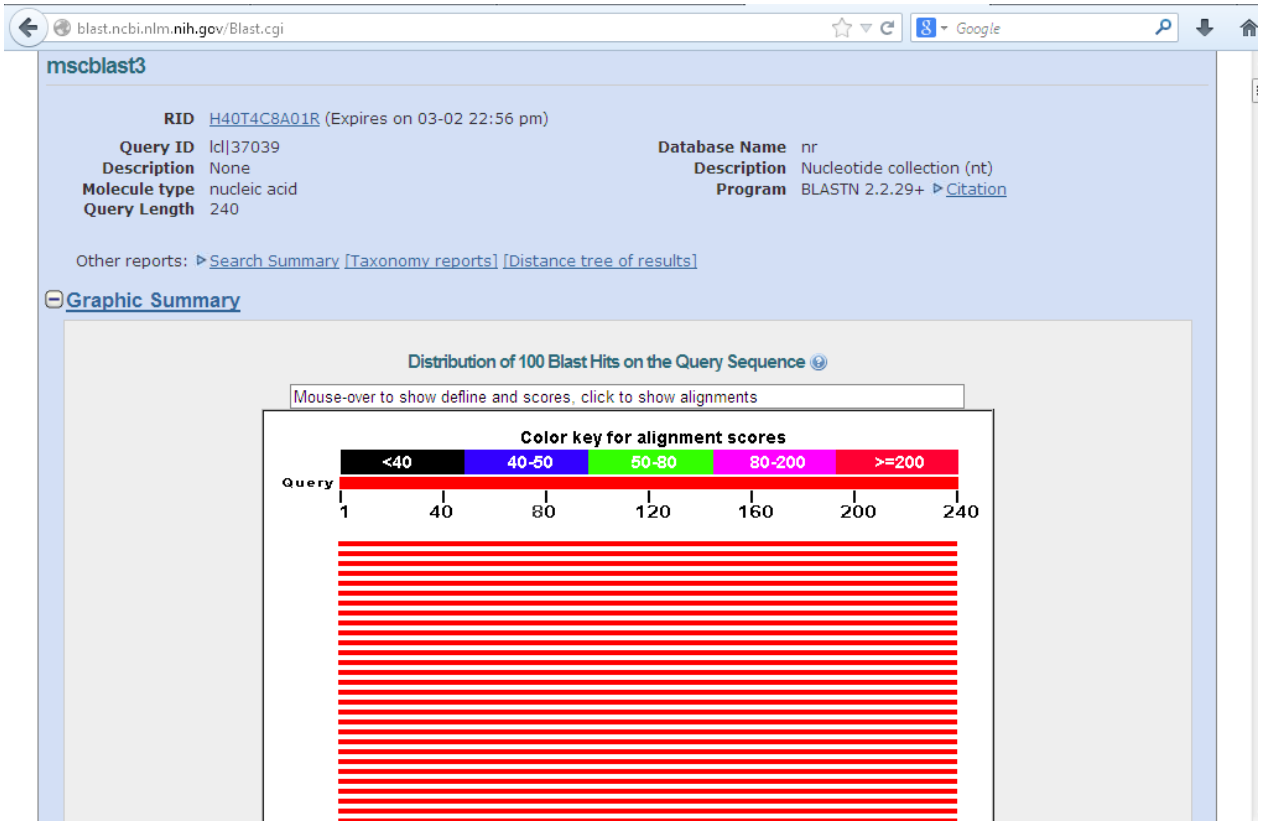
aattttattt tttaacctaa ctcccctact aagtgtaccc ccccttttccc ccccaggggg ggtatactat gcataatcgt gcatacattt atataccaca tatattatgg taccggtaat atatactata tatgtactaa acccattata tgtatacggg cattaatcta tattccacat ttctcccaat gtccattcta tgcatgatcc aggacacact cattcaccct ccccatagac

- In the Job Title type mscblast3
- Click on BLAST button
- The following results will be returned.

Part 1:

-Query length i.e 240 is given

-Note that the graphic color is now red for all the hits (i.e maximum score of >=200)

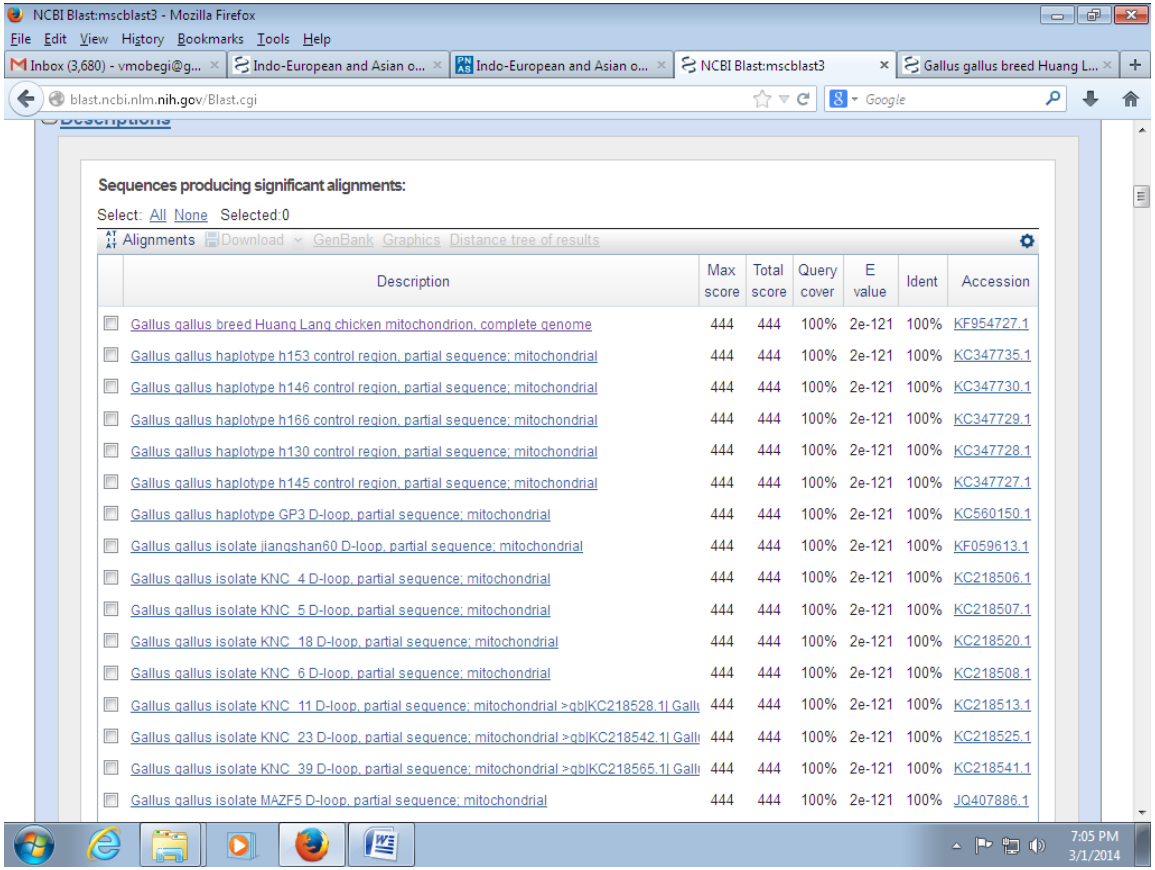


Part 2:

-Now maximum score for each top hit alignment is 444 and Expectation (E) value is 2e-121

Why do you think maximum score is higher and E lower compared to the previous (i.e Task 1 and 2)?

The longer the query sequence the more the confidence associated with its match. In other words it is unlikely that the sequence matched by chance.



Description	Max score	Total score	Query cover	E value	Ident	Accession
Gallus gallus breed Huang Lang chicken mitochondrion, complete genome	444	444	100%	2e-121	100%	KF954727.1
Gallus gallus haplotype h153 control region, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC347735.1
Gallus gallus haplotype h146 control region, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC347730.1
Gallus gallus haplotype h166 control region, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC347729.1
Gallus gallus haplotype h130 control region, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC347728.1
Gallus gallus haplotype h145 control region, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC347727.1
Gallus gallus haplotype GP3 D-loop, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC560150.1
Gallus gallus isolate jiangshan60 D-loop, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KF059613.1
Gallus gallus isolate KNC_4 D-loop, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC218506.1
Gallus gallus isolate KNC_5 D-loop, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC218507.1
Gallus gallus isolate KNC_18 D-loop, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC218520.1
Gallus gallus isolate KNC_6 D-loop, partial sequence; mitochondrial	444	444	100%	2e-121	100%	KC218508.1
Gallus gallus isolate KNC_11 D-loop, partial sequence; mitochondrial >gb|KC218528.1| Gall	444	444	100%	2e-121	100%	KC218513.1
Gallus gallus isolate KNC_23 D-loop, partial sequence; mitochondrial >gb|KC218542.1| Gall	444	444	100%	2e-121	100%	KC218525.1
Gallus gallus isolate KNC_39 D-loop, partial sequence; mitochondrial >gb|KC218565.1| Gall	444	444	100%	2e-121	100%	KC218541.1
Gallus gallus isolate MAZF5 D-loop, partial sequence; mitochondrial	444	444	100%	2e-121	100%	JQ407886.1

Task 4

- Go to NCBI BLAST. Perform homology search for the sequence
- AATAAAGATT GTAAAATGAT AATTTGGTTT ATTCAACCAA CGATTTTTTA CATAATTTTT

1. What is the maximum score and E for the top hit?
2. Which organism and what gene does this sequence belong to?
3. What is the length of the gene?
4. On what chromosome is the gene located?
5. From which isolate has the sequence been generated?

BLASTp

- Used to search a protein sequence against a protein database such as Uniprot.

-If your DNA sequence is coding, you can translate it and use blastp to search protein database. Otherwise you can retrieve protein sequence from database.

-You can access BLAST in many different ways and many different sites. The default parameters may be significantly different , the databases may not be updated on the same schedule as such may be significantly different in size or level of redundancy.

Lecture 2

Protein structure and function prediction

Expert Protein Analysis System (ExPASy)

- ExPASy is the Swiss Institute of Bioinformatics (SIB) Bioinformatics Resource Portal which provides access to scientific databases and software tools (i.e., resources) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc
- It was initially a protein and proteomics server dedicated to the analysis of protein sequences and structures
- In this course we will consider just a few tools but you will note that there are several tools available here that may be useful in other applications

Signal peptides

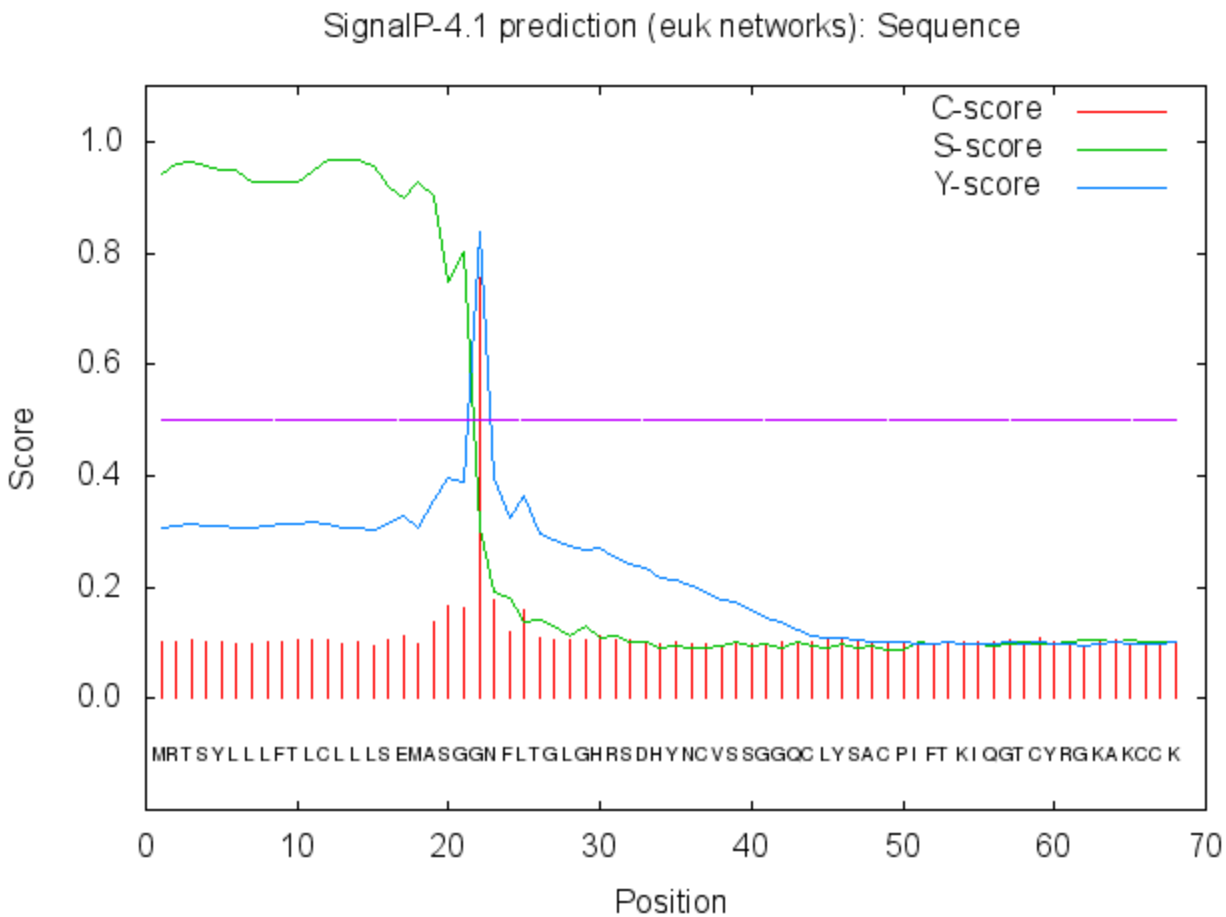
- **SignalP** predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks.

Exercise 1

- Go to Uniprot http://www.uniprot.org/
- In Search in box select Protein Knowledgebase (UniProtKB) and in Query box type "human beta-defensin 1"
- Select entry P60022 and entry name DEFB1_HUMAN
- Click on the entry P60022 to get information about this protein

- Scroll down to where you have protein sequence and copy the sequence
- Go to proteomics category and click on "protein modifications"
- **Under tools click on SignalP** http://www.cbs.dtu.dk/services/SignalP/
- This takes you to **SignalP server**
- Now paste the sequence you copied earlier to the box provided under SUBMISSION
- Use Default parameters and press Submit

Output from SignalP



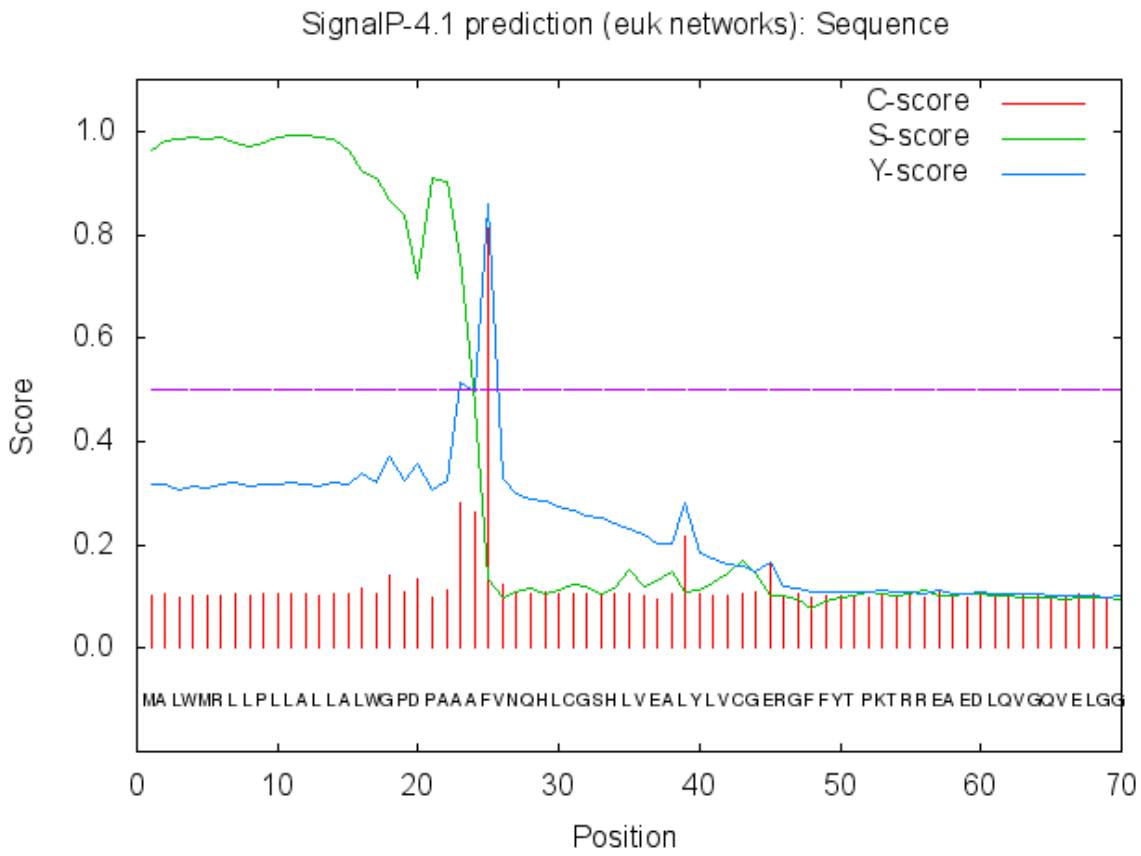
Measure Position Value Cutoff signal peptide? max. C 22 0.756 max. Y 22 0.837 max.
S 14 0.968 mean S 1-21 0.925 D 1-21 0.885 0.450 YES

Name=Sequence SP='YES' Cleavage site between position 21 and 22

Exercise 2

- Go to Uniprot http://www.uniprot.org/
- In Search in box select Protein Knowledgebase (UniProtKB) and in Query box type “Insulin"
- Select entry P01308 and entry name INS_HUMAN
- Click on the entry P01308 to get information about this protein
- Follow the rest of the steps as in Exercise 1 above.

Output from SignalP



Measure Position Value Cutoff signal peptide? max. C 25 0.812 max. Y 25 0.861 max. S 11 0.993 mean S 1-24 0.917 D 1-24 0.891 0.450 YES Name=Sequence SP='YES' Cleavage site between pos. 24 and 25

>Sequence ; MatureChain: 25-110
FVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEG
SLQKRGIVEQCCTSICSLYQLENYCN