Subject: Bioinformatics

Lesson: NCBI: Similarity searching tool-BLAST

Lesson Developer: Sandipa Das

College/ Department: Department of Botany, University of Delhi

;

Table of Contents

Chapter: NCBI: Similarity searching tool-BLAST

- **Introduction to Sequence Analysis**
- **Sequence analysis tools**
  - **Similarity searching with nucleotide queries (BLASTN)**
  - **Similarity searching with Protein queries (BLASTP)**
- **Summary**
- **Exercise/ Practice**
- **Glossary**
- **References/ Bibliography/ Further Reading**

;

Introduction to Sequence Analysis

Analysis of the sequence data is one of the major challenges of computation biology and is the first step towards understanding molecular basis of development and adaptation. Several types of analysis can be performed that range from

DNA Sequence analysis

- sequence similarity searches
- prediction of genes and other genetic elements
- evolutionary tendencies and trends
- Functional information

RNA analysis

- Expression analysis
- Structure
- Functional information

Protein level

- Domain finding
- Structure prediction
- Evolution
- Function

Genome level

- Comparative genomics
- Genome organization and re-organisation
- Genome annotation

Similarity search with Nucleotide queries

DNA sequence analysis constitutes one of the major applications in bioinformatics. Some of the basic objectives of performing sequence analyses are

- Sequence retrieval
- Finding similar sequence through **similarity searching**
- Phylogenetic or evolutionary analysis
- Finding homology relationships (orthologus and paralogous nature)
- Discovering new genes and genetic elements

;

- Exploring importance of residues (nucleotides and amino acids) that are important for structure and function

Central to the process of searching for similar sequences from database and retrieval are concepts of homology that are derived from evolutionary relationships. DNA data can be used to retrieve similar sequences that have diverged upto 600 million years ago!

Sequences can be retrieved from NCBI database by using the identity of the sequence in the form of **accession number** and/or using the **"Identity"** of the sequence as **"query"** to search against the entire database or by selecting a specific database.

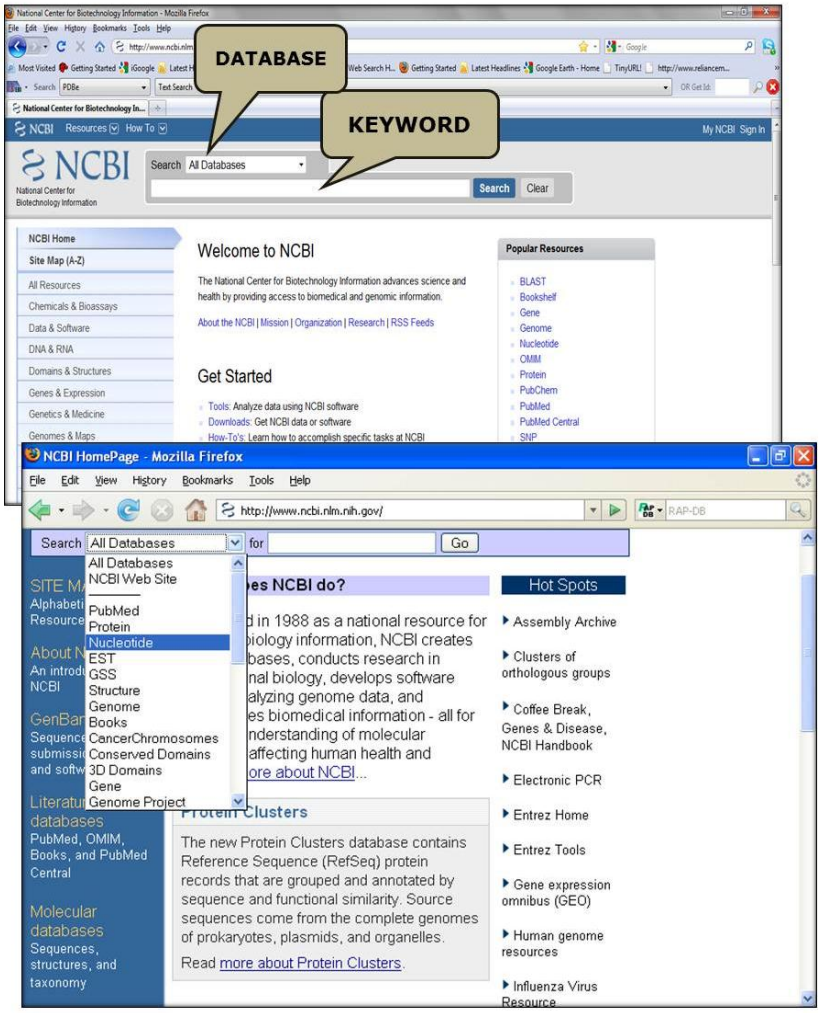


Figure NCBI homepage showing the location of the pull-down menu for accessing databases and search engine

;

Search result for "reverse transcriptase"

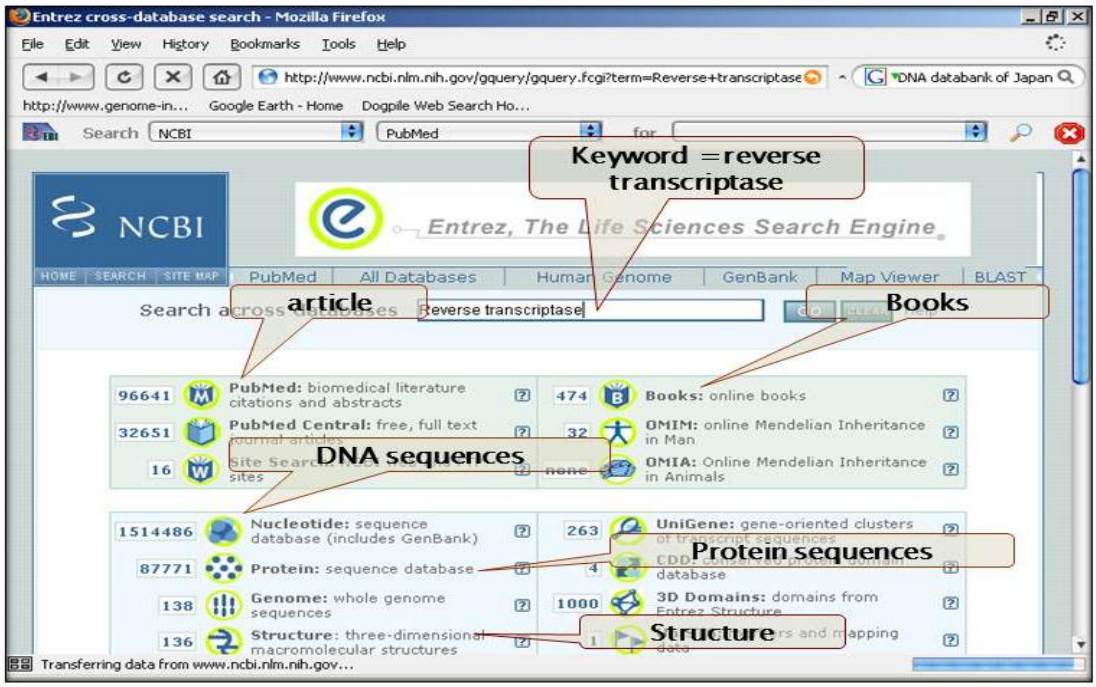
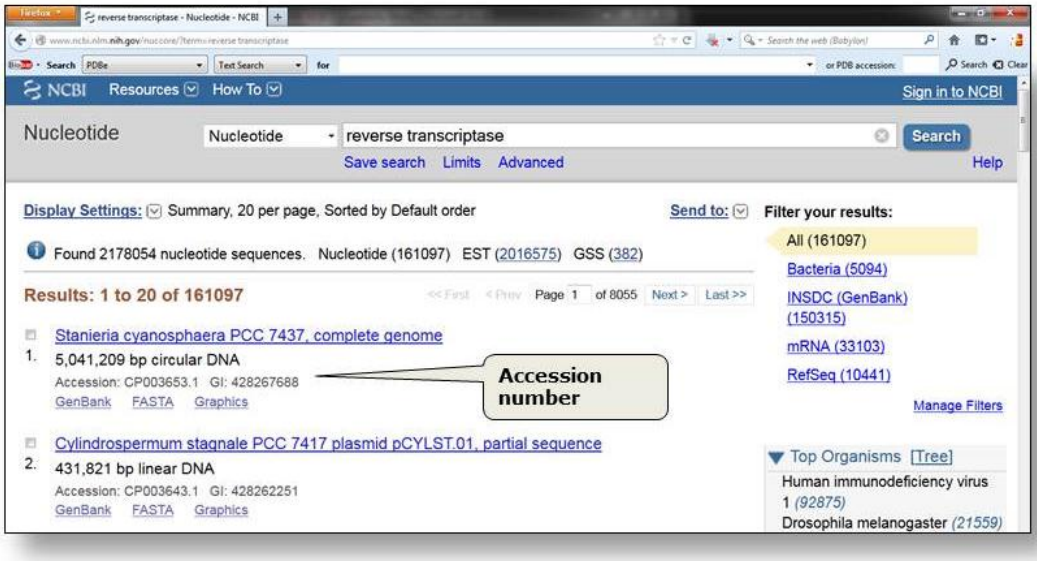


Figure An output window showing results obtained against all databases using "reverse transcriptase" as keyword using the Entrez system.



;

Figure : An output showing the details of search results performed against nucleotide database

Sequence can be retrieved as FASTA formatted sequence or in Genbank format.

FASTA formatted files are simple text files of nucleotide or protein sequence where a single definition beginning with a "greater than (>)" sign is placed at the beginning of the the sequence. This is one of formats that are recognized by almost all sequence analysis softwares. A single file can contain several FASTA formatted sequence that can then be used for analysis such as in multiple sequence alignment.

;

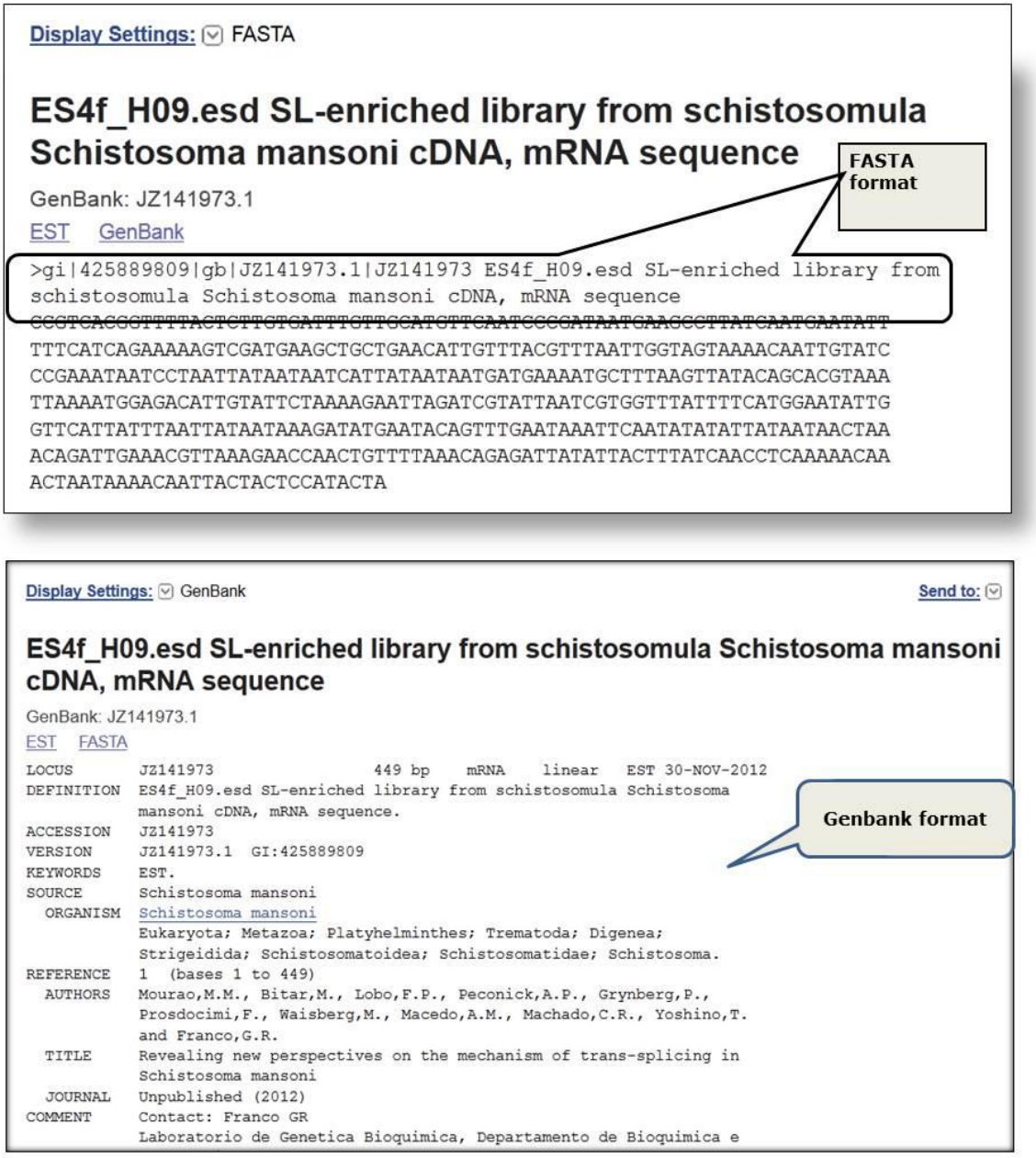


Figure : A FASTA (upper) and Genbank (lower) formatted DNA sequence file

Genbank formatted files contains detailed annotation and the associated sequence Sequence similarity search is performed using a suite of tools called "**BLAST" i.e. Basic Local Alignment Search Tool .** Two distinct types of sequence similarity searches can be

;

performed – **Local** and **Global**. Needleman and Wunsch developed the GLOBAL alignment algorithm (1970) whereas Michael Waterman and Temple Smith co-developed the Smith-Waterman sequence alignment algorithm for LOCAL alignment (1981). Global alignment attempts to find an "optimal or average" similarity via alignment over the entire length between the user provided "query" and "subject" sequences that are part of the database. Local alignment, in contrast attempts to find "local" regions of high similarity between query and subject sequences.

Sequence similarity searches, performed via alignment are a measure of relatedness i.e. sequences that are evolutionary closely related will align over larger distances; in other words similarity is a function of evolutionary relatedness. Similarity searches carried out against subject sequences in the database are based on **pairwise alignment,** i.e. between two sequences at-a-time. One of the two sequences is always the "query" sequence, whereas the subject sequences retrieved from the database changes.

Similarity being a function of evolutionary relationship can also be extended for employing sequence alignments to evaluate molecular phylogeny via **multiple sequence alignment.**

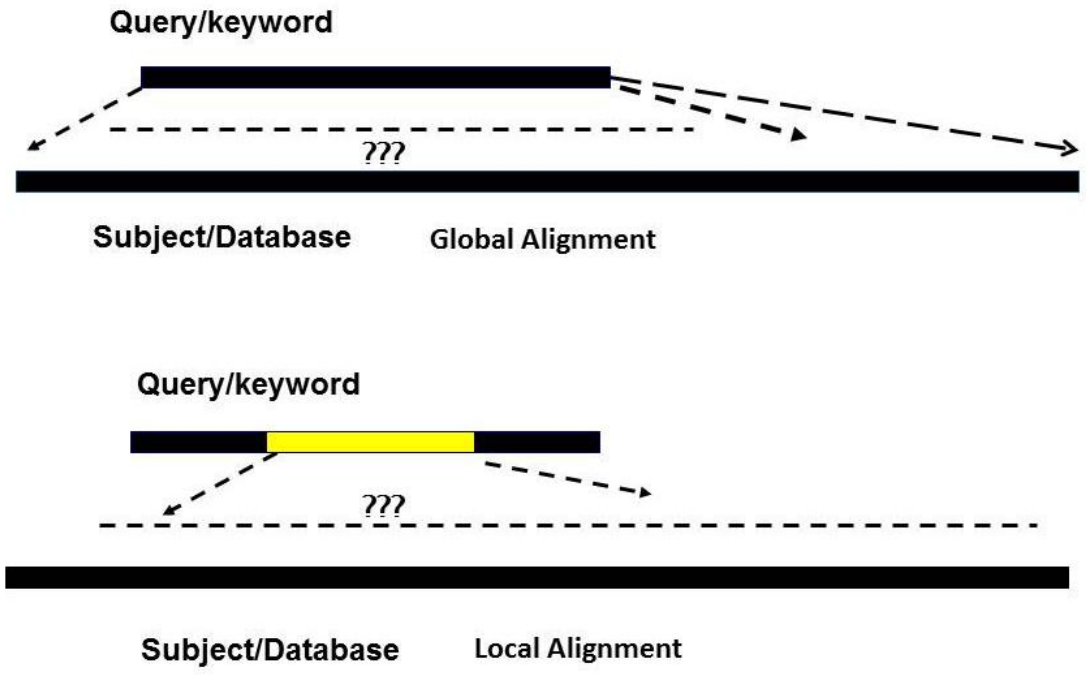


Figure : Pictorial representation of Global and Local alignment

Source: Dr Sandeep Das

;

Selecting an appropriate BLAST program: Local alignment similarity searches can be performed across both DNA and protein sequences independently using BLAST tool/algorithm. Variations in BLAST algorithm also allows researchers to perform similarity search using DNA as query against protein sequence as subject sequences (database) and vice-versa. In addition, translation products (i.e. protein products) of DNA query sequences are also searched against translation products of DNA databases. The following table summarizes the various BLAST algorithms used for sequence similarity searches.

Table: Various BLAST programs

Source: Dr Sandeep Das

Tool / algorithm	Query	Subject (=database)
BLASTN	DNA	DNA
BLASTP	Protein	Protein
BLASTX	Translation products of DNA query	Protein
TBLASTN	Protein	Translation products of DNA subject
TBLASTX	Translation products of DNA query	Translation products of DNA subject

Selecting an appropriate database:

Table : Various databases that need to be selected for BLAST

Source: Sandeep Das

Query	Database
Nucleic Acid (DNA)	nr (non-redundant; all GenBank+EMBL+DDJB but no EST, GSS)
	Chromosome (complete genome and chromosome
	dbSTS (Sequence tagged sites)
	est (EST database)
	gss (Genome survey Sequence)
	pat (Patented sequences)

;

	wgs (whole genome shotgun sequences)
Protein	nr (non-redundant; all GenBank coding + PDB+ Swissprot+ PIR+ PRF)
	PDB (protein database)
	pat (Patented sequences)
	env_nr (protein samples from environmental samples)

Selection of appropriate length of keyword: Performing a search requires the use of keyword/s that are appropriately framed and of suitable length. For example, searching "Google (= datasbe containing "subject") " using either "Delhi" or "University" (as "query") retrieves 457,000,000 and 2,660,000,000 results respectively whereas using the two keyword together i.e. as "University of Delhi" retrieves 46,900,000 results (almost 10 folds or 50 folds less than the earlier results). A further refinement can be obtained by using "University of Delhi North Campus" as the keyword (1,500,000 results; all searches performed on 10.12.12). This example demonstrates the importance of using an appropriate keyword, and the fact that using a "short" keyword also retrieves results that are distantly related to the original query; in contrast using a "long" key word retrieves "closely related" subjects. A similar logic is used to perform similarity searches using DNA sequences.

While performing similarity search using Local alignment tools of BLAST, the query is "divided" into stretches of smaller length, termed as "**word length or K-tuple**" values that is used to initiate the search and extend the search. BLAST attempts to identify an exact matches between subjects to the query words termed as "**word hits**". The "word" match serves as a "seed" that can be extended by BLAST programs in multiple steps to generate the final gapped alignments.

The choice of **"word size or length"** or "**K-tuple**" value is therefore dependent on the user's objective and requirement, and can be suitably selected to regulate specificity, sensitivity and stringency of the search.

DNA based analysis employs "word length" that can vary between 7-256 nucleotides; whereas the "word size" ranges between 2-3 amino acids for proteins.

The ranking of the retrieved subjects are dependent on the similarity between the query and the subjects and several defined matrices are used. The matrix used for scoring similarity in DNA analysis and for ranking is termed as "**scoring matrix or unitary**

;

matrix" and consists of a positive score for similar or identical nucleotides character and a zero (0) or a negative score for a mismatch or dissimilar character.

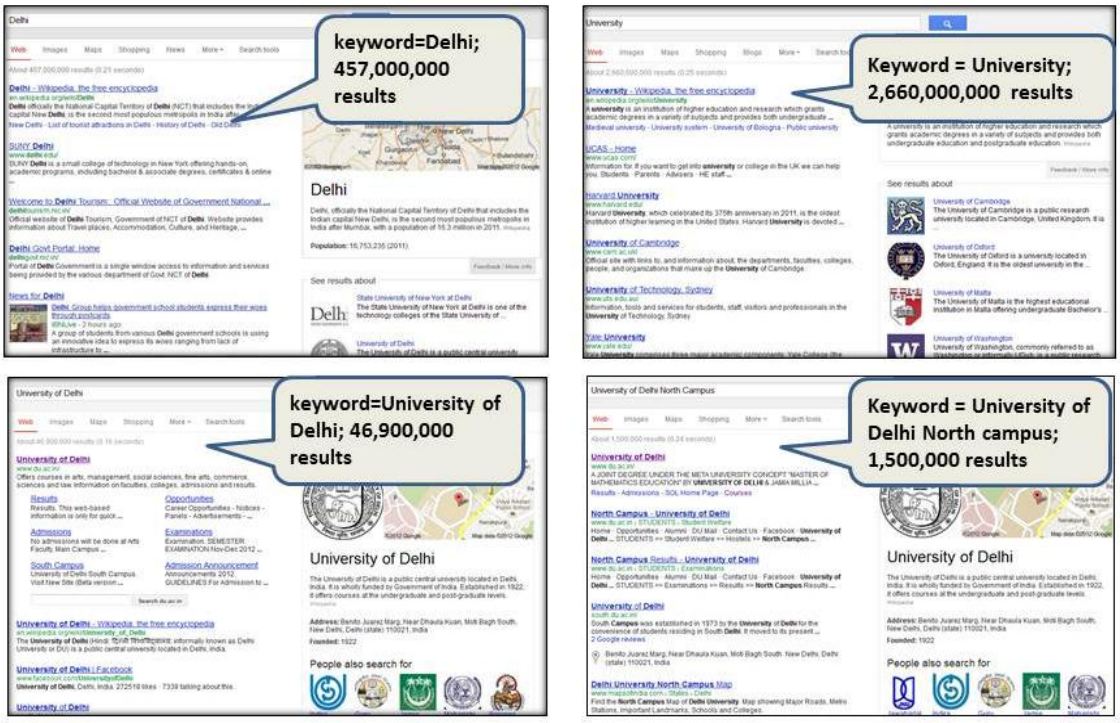


Figure : A hypothetical example showing the relationship between the "query/keyword" length and specificity of the subjects retrieved from the database.

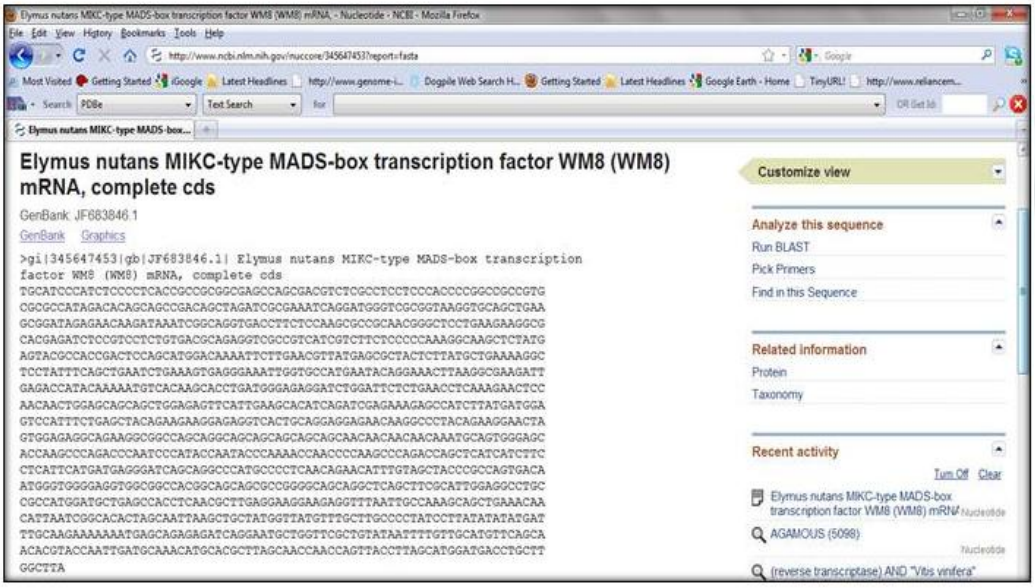


Figure : FASTA formatted query sequence

;

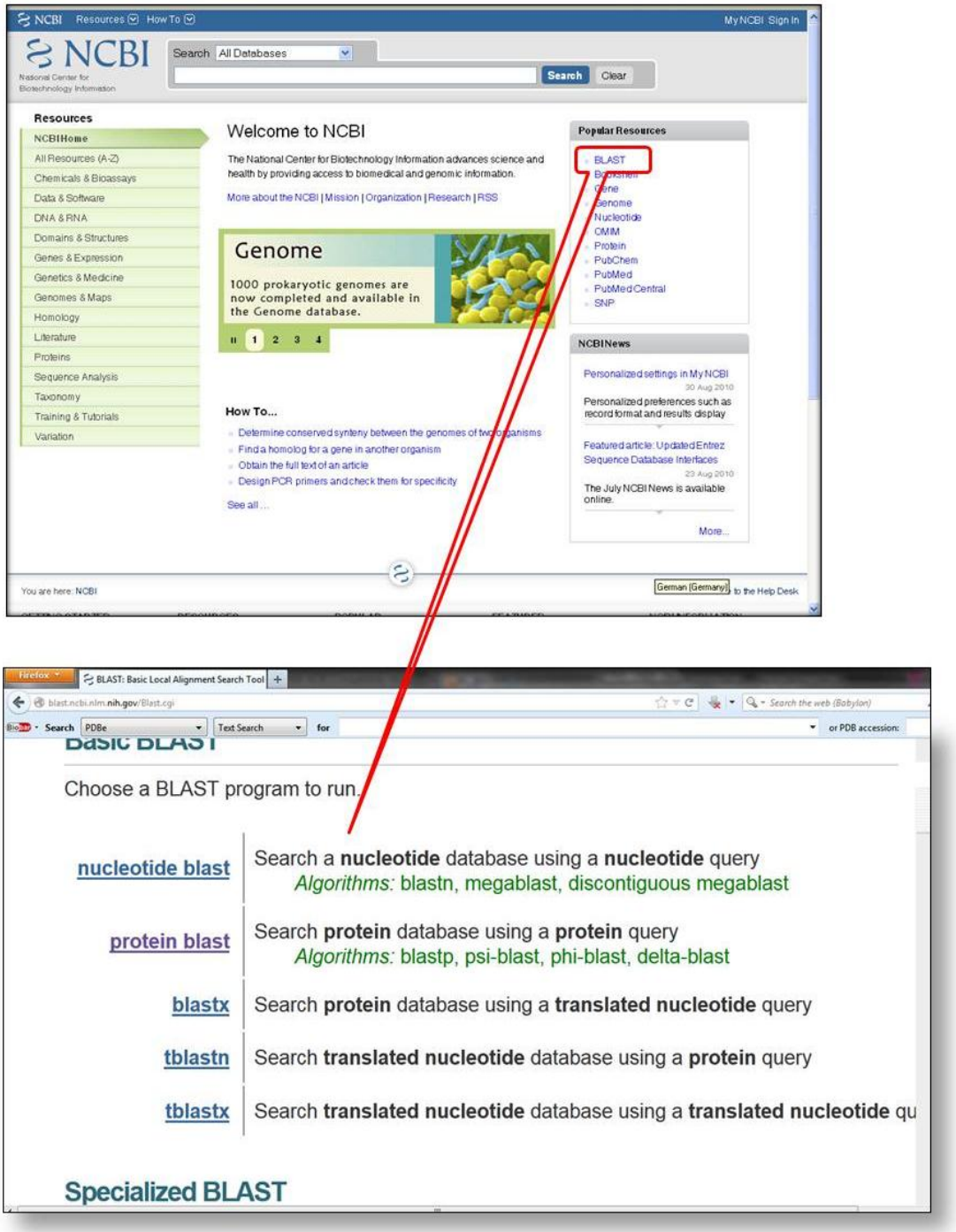


Figure: Accessing BLAST from NCBI

;

Local alignment through BLASTN requires a nucleotide sequence that is used as Query. The query can be 'user provided" or retrieved from NCBI through the use of an appropriate keyword as discussed earlier and BLAST can be accessed from the main homepage of NCBI. Once the BLAST webpage has been accessed, either the query sequence or the accession number (if available) can be entered in the "query box". Subsequently a suitable database is selected followed by BLAST. At present there are **three different variations** of BLASTN available:

a. **BLASTN:** for somewhat related sequences
b. MEGABLAST: for highly similar sequences
c. **DISCONTIGOUS MEGABLAST:** More dissimilar sequences

For beginners it is recommended to use BLASTN, i.e. searching for somewhat similar sequences. These variations of BLASTN differ with respect to the "word size" or "k-tuple" value that is used to initiate and perform BLAST.

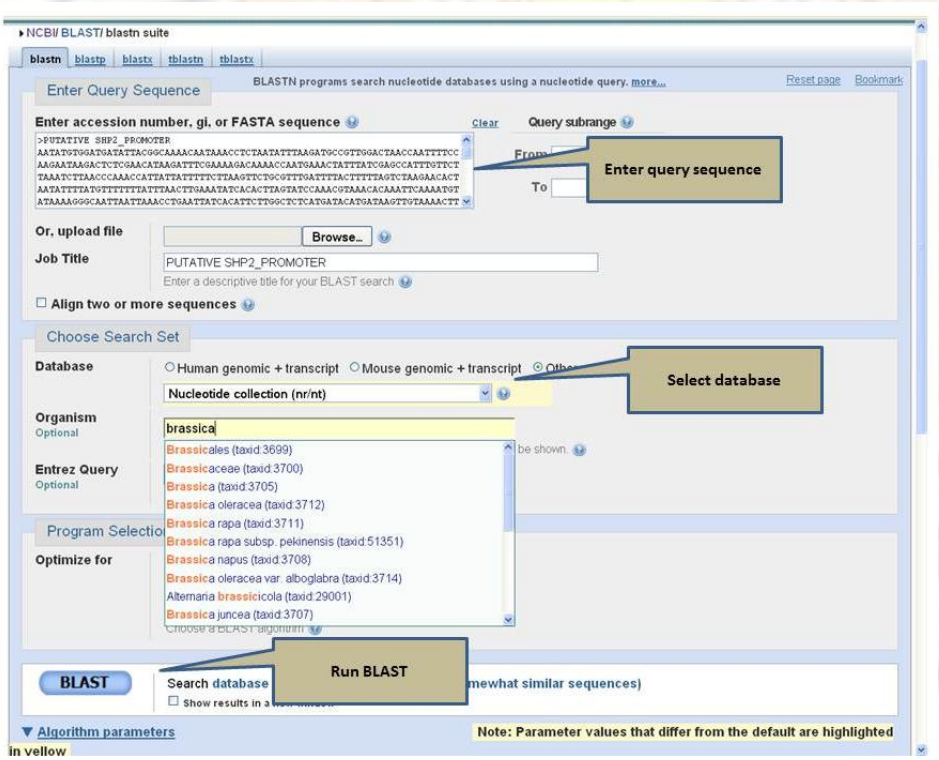


Figure: How to perform basic BLASTN

;

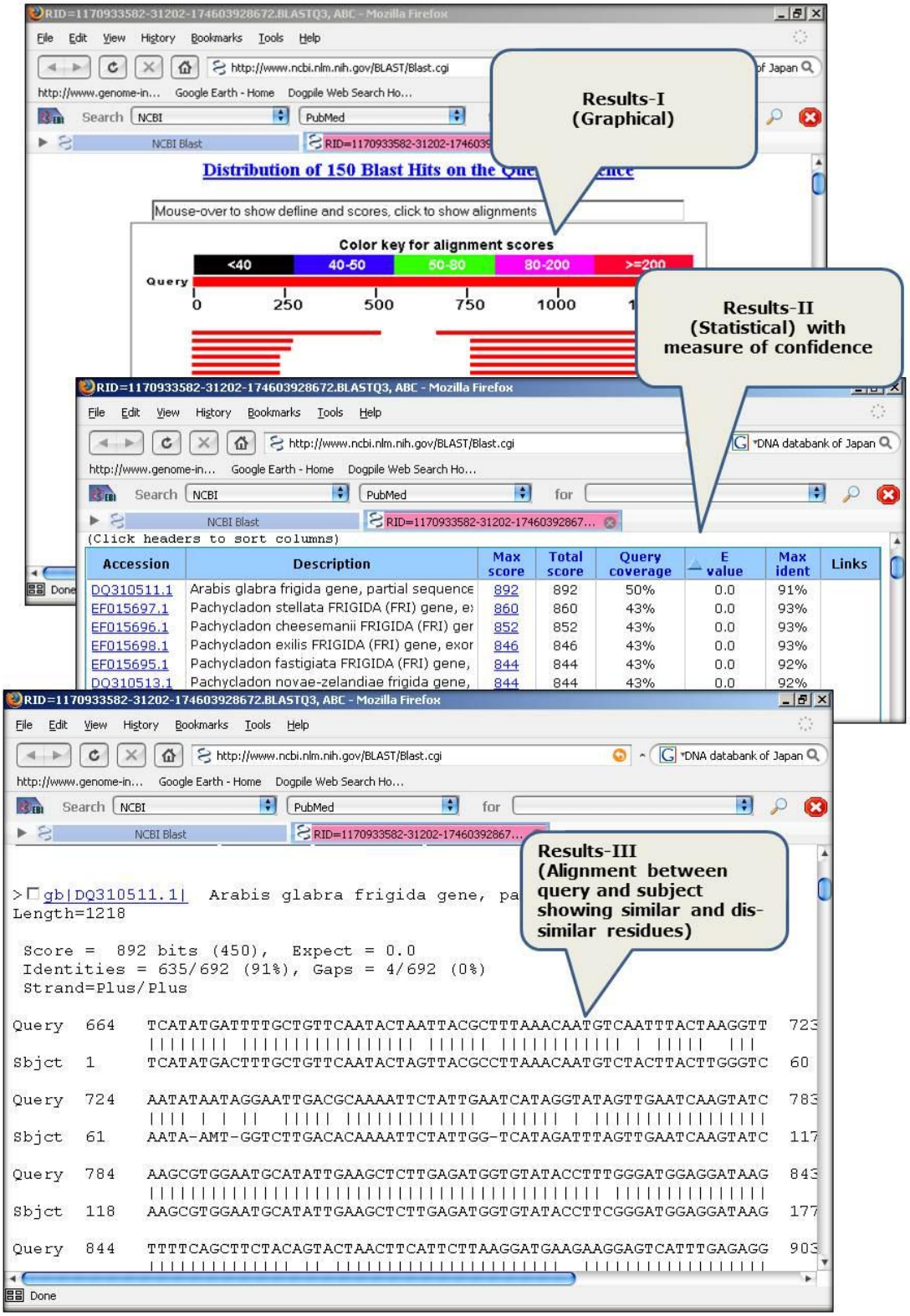


Figure : BLAST results in "graphical", "tabular" and "alignment" format

;

The results window displays the result with the help of a graphical overview and also with the help of detailed pairwise alignment where the subject and the query sequences showing the similar nucleotides are "linked" with the help of a vertical line, whereas differences that arise as a result of mutational events appear as either mis-matched residues or as gaps.

Similarity search with Protein queries

BLASTP or similarity searching using protein query against Protein database is performed using BLASTP algorithm. The webpage is accessed as shown earlier (Figure 1.8) and instead of a nucleotide query, a protein query is used. However, BLASTP differs from BLASTN in the type of matrix that is employed for calculating the similarity and thereby determining the ranking of the subject vis-à-vis the query. In place of unitary matrix, protein similarity searches uses a **mutational probability index** or **substitution matrix.** The commonly used matrices for BLASTP are **PAM (Point Accepted Mutation)** and **BLOSSUM** matrix developed by Margaret Dayhoff (PAM, 1978) and Henikoff and Henikoff (BLOSSUM; Henikoff and Henikoff 1992) respectively. Details about these will be dealt with in the chapter dealing with Multiple sequence Alignment (MSA). There are four types of BLASTP available at NCBI and these are:

a. BLASTP
b. PSI-BLAST (Position specific Iterated BLAST)
c. PHI-BLAST (Pattern Hit Initiated BLAST)
d. DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST

;

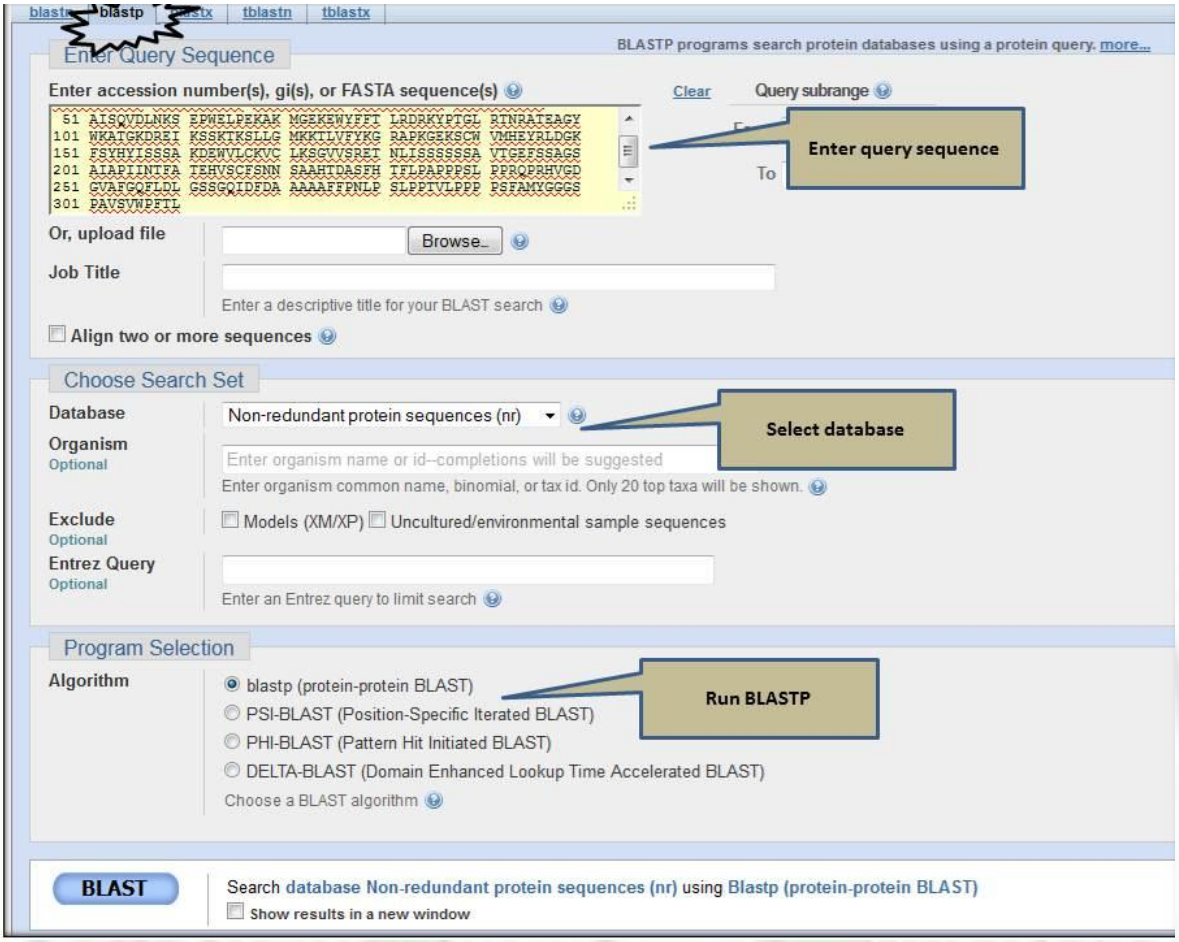
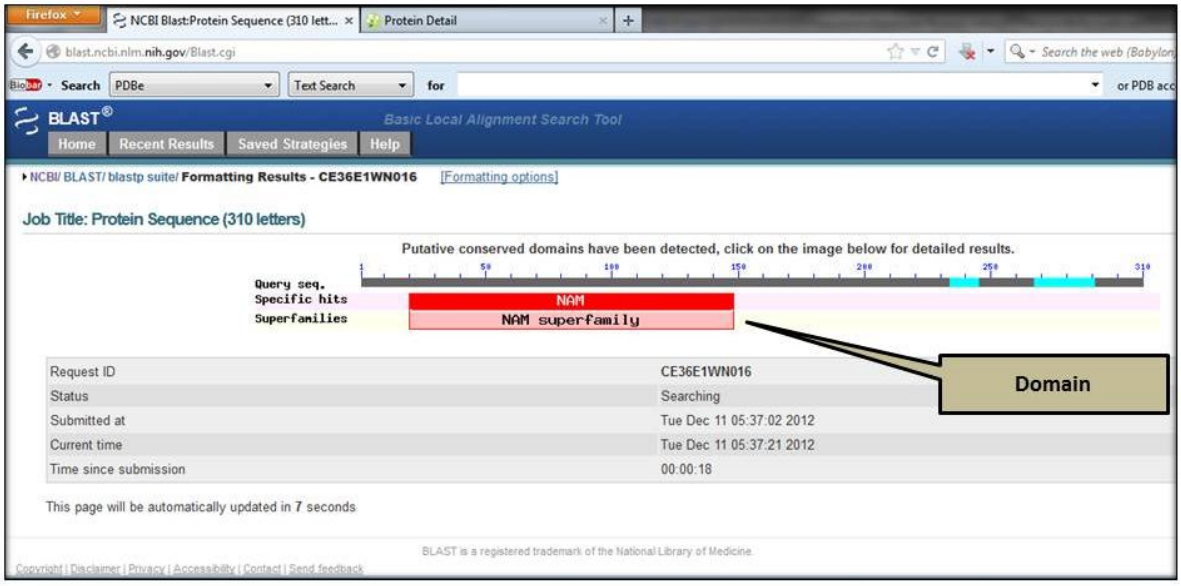
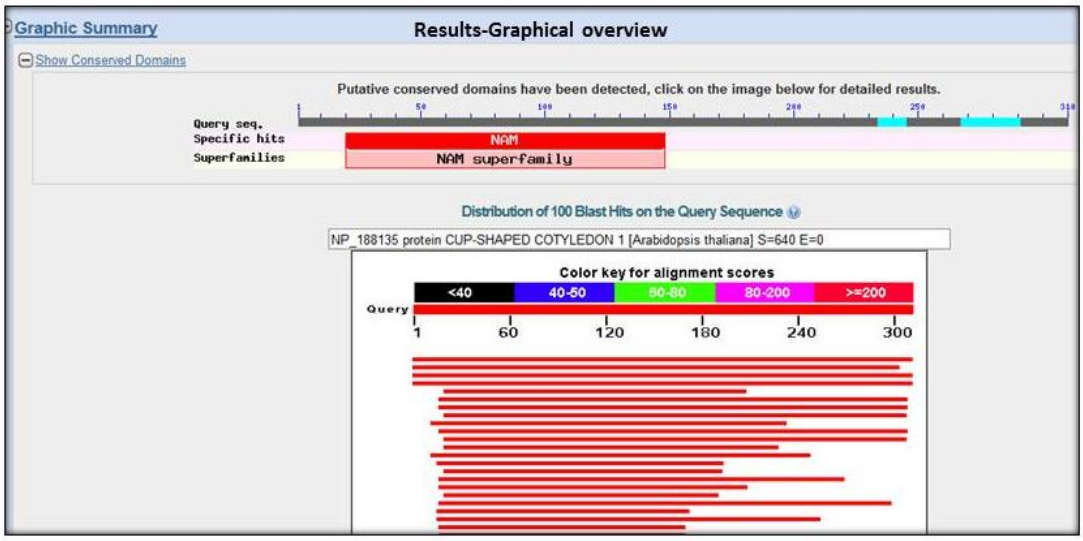


Figure : Accessing and performing BLASTP

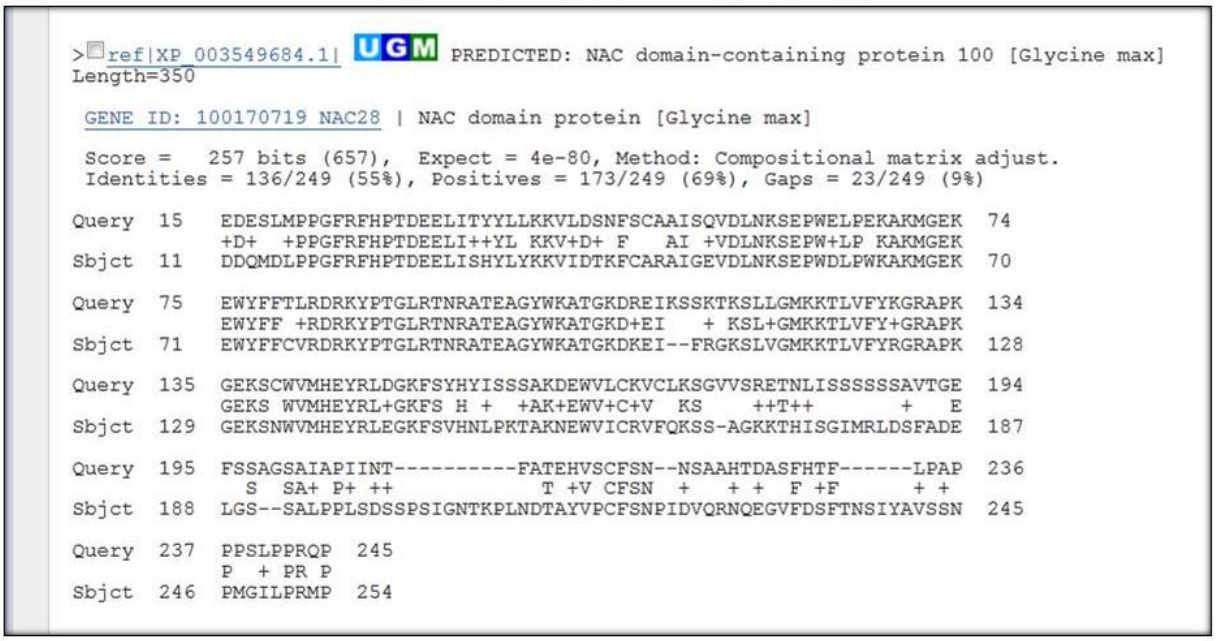
;


NCBI Blast:Protein Sequence (310 lett...
Protein Detail
blast.ncbi.nlm.nih.gov/Blast.cgi
Search
PDBe
Text Search
for
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastp suite/ Formatting Results - CE36E1WN016
[Formatting options]
Job Title: Protein Sequence (310 letters)
Putative conserved domains have been detected, click on the image below for detailed results.
Query seq.
Specific hits
Superfamilies
NAM
NAM superfamily
Domain
Request ID
CE36E1WN016
Status
Searching
Submitted at
Tue Dec 11 05:37:02 2012
Current time
Tue Dec 11 05:37:21 2012
Time since submission
00:00:18
This page will be automatically updated in 7 seconds
BLAST is a registered trademark of the National Library of Medicine
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback


Graphic Summary
Results-Graphical overview
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
Query seq.
Specific hits
Superfamilies
NAM
NAM superfamily
Distribution of 100 Blast Hits on the Query Sequence
NP_188135 protein CUP-SHAPED COTYLEDON 1 [Arabidopsis thaliana] S=640 E=0
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
60
120
180
240
300

;

Results- Pairwise alignment between query and subject



```
>ref|XP_003549684.1| UGM PREDICTED: NAC domain-containing protein 100 [Glycine max]
Length=350

 GENE ID: 100170719 NAC28 | NAC domain protein [Glycine max]

 Score =   257 bits (657),  Expect = 4e-80, Method: Compositional matrix adjust.
 Identities = 136/249 (55%), Positives = 173/249 (69%), Gaps = 23/249 (9%)

Query  15   EDESLMPPGFRFHPTDEELITYYLLKKVLDSNFSCAAISQVDLNKSEPWELPEKAKMGEK  74
            +D+  +PPGFRFHPTDEELI++YL KKV+D+ F   AI +VDLNKSEPW+LP KAKMGEK
Sbjct  11   DDQMDLPPGFRFHPTDEELISHYLYKKVIDTKFCARAIGEVDLNKSEPWDLPWKAKMGEK  70

Query  75   EWYFFTLRDRKYPTGLRTNRATEAGYWKATGKDREIKSSKTKSLLGMKKTLVFYKGRAPK  134
            EWYFF +RDRKYPTGLRTNRATEAGYWKATGKD+EI    + KSL+GMKKTLVFY+GRAPK
Sbjct  71   EWYFFCVRDRKYPTGLRTNRATEAGYWKATGKDKEI--FRGKSLVGMKKTLVFYRGRAPK  128

Query  135  GEKSCWVMHEYRLDGKFSYHYISSSAKDEWVLCKVCLKSGVVSRETNLISSSSSSAVTGE  194
            GEKS WVMHEYRL+GKFS H +  +AK+EWV+C+V  KS    ++T++        +   E
Sbjct  129  GEKSNWVMHEYRLEGKFSVHNLPKTAKNEWVICRVFQKSS-AGKKTHISGIMRLDSFADE  187

Query  195  FSSAGSAIAPIINT----------FATEHVSCFSN--NSAAHTDASFHTF------LPAP  236
              S  SA+ P+ ++           T +V CFSN  +   + +  F +F       + +
Sbjct  188  LGS--SALPPLSDSSPSIGNTKPLNDTAYVPCFSNPIDVQRNQEGVFDSFTNSIYAVSSN  245

Query  237  PPSLPPRQP  245
            P  + PR P
Sbjct  246  PMGILPRMP  254
```

Exercises

1. What is the principle of similarity searching?
2. What are the objectives of analysis of sequence data?
3. Define the following:
    a. Accession number
    b. Query
    c. Subject
4. What are FASTA and Genbank formatted files?
5. Differentiate between Local and Global alignment.
6. Local and Global alignment algorithms were developed by --------& ----------, and ---------- & ----------, respectively.
7. Whether the following statement is true or false:
    a. BLASTN compares DNA sequence against protein sequence
    b. BLASTP compares protein against protein sequence
    c. BLASTX compares DNA against DNA sequence

8. dbSTS and env_nr databases contain ---------- and ---------- sequences respectively.

;

9. A "short" keyword is employed to retrieve closely related sequence. True or false?
10. What is a word size or k-tuple value?
11. With the help of a flowchart, list the steps taken to retrieve a DNA sequence from NCBI with Genbank and FASTA format.
12. What are the three different types of BLASTN?
13. You have been given a nucleotide sequence from *Homo sapiens.* How would you identify and retrieve nucleotide sequences from other closely related organisms
14. Using a query such as *Oryza sativa* retrotransposons" identify and retrieve similar nucleotide sequences from fungi and bacteria
15. Can you identify and retrieve protein sequences from databases using DNA sequence as a query? If yes, trace the steps using a flowchart.
16. What are scoring and substitution matrices?
17. Expand the following:
    a. PAM
    b. BLOSSUM
    c. MSA
18. What are the various sub-types of BLASTP?
19. With the help of a flowchart, enumerate the steps taken to perform BLASTP.

GLossary

a. BLAST: A suite or collection of algorithm for comparison of sequences. Several forms of BLAST such as BLASTN, BLASTP, BLASTX, TBLASTX, and TBLASTN allows user to compare DNA and protein sequences either to similar groups (i.e. DNA to DNA or protein to protein) or between groups (i.e. DNA to protein and vice-versa)

b. E-value: A statistical probability of **expecting** or finding a "hit" or "match" between a query and a subject sequence by chance. Also termed as "expect value"

c. Word size or K-tuple: A minimum number of characters (nucleotide or amino acids) that must have an exact match between query and subject before the alignment can proceed or extend. An appropriate word size can be chosen to increase or decrease sensitivity, accuracy and speed of alignment and similarity searching

;

d. **PAM matrix:** A substitution matrix devised by Margaret Dayhoff (1972, 1978) that is based on rate of substitution of an amino acid by any other amino acid over a given period of evolutionary time or distance. The matrix was derived based on studied performed on small globular proteins

e. **BLOSSUM:** Another substitution matrix devised by Henikoff and Henikoff, but based on substitution rates within domains or conserved regions of proteins found in BLOCKS database (http://blocks.fhcrc.org/help/)

References

1. Works Cited

Altschul S.F, Gish W, Miller W, Myers E W and Lipman D J. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215, 403-410 (1990)

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In "Atlas of Protein Sequence and Structure" 5(3) M.O. Dayhoff (ed.), 345 -352 (1978)

Henikoff, S. and Henikoff, J. Amino acid substitution matrices from protein blocks (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=1438297&query_hl=6&itool=pubmed_docsum) *Proc. Natl. Acad. Sci. USA.* 89(biochemistry): 10915 - 10919 (1992).

2. Suggested Readings

Bioinformatics and Functional Genomics: 2nd Edition, Jonathon Pevsner (2009), Wiley Blackwell

3. Web Links

1.1 http://blast.ncbi.nlm.nih.gov/

1.2 http://blocks.fhcrc.org/help/