CS426: Introduction to Computational Biology

Section Week 3

Bioinformatics Databases and Tools

- Why would we use such databases
  - When obtaining a new DNA sequence, one needs to know whether it has already been deposited in the databanks fully or partially, or whether they contain any *homologous sequences*(sequences which are descended from a common ancestor).
  - Some of the databases contain annotation which has already been added to a specific sequence. Finding annotation for the searched sequence or its *homologous sequences* can facilitate its research.
  - Find similar non-coding DNA stretches in the database:for instance repeat elementsor regulatory sequences.
  - Other uses for specific purpose, like locating false priming sites for a set of PCR oligonucleotides.
  - Search for *homologous proteins* - proteins similar in their sequence and therefore also in their presumed folding or structure or function.

Primary sequence databases

DNA (nucleotide)		Protein	
EMBL	UK	PIR	US
GenBank	US	MIPS	Germany
DDBJ	Japan	Swiss-Prot	Swiss
Celera	Celera	TrEMBL	Swiss
		NRL_3D	US
		GenPept	US

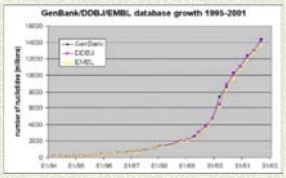
List of primary sequence databases and their locations.

Primary sequence databases

- There are several problems with databases today:
  - Databases are regulated by users rather than by a central body (except for Swiss-Prot).
  - Only the owner of the data can change it.
  - Sequences are not up to date.
  - Large degree of redundancy in databases and between databases.
  - Lack of standard for .elds or annotation.

DNA Databases (Nucleotide Sequences)

- Growing faster then the protein databases.



- Largest databases: Genbank (US), EMBL (Europe - UK), DDBJ (Japan).

DNA Databases

- EMBL : URL http://www.ebi.ac.uk/embl/
  - EMBL is a DNA sequence database from European Bioinformatics Institute (EBI).
  - EMBL includes sequences from direct submissions, from genome sequencing projects, scientific literature and patent applications.
  - Its growth is exponential,
  - supports several retrieval tools:
  - SRS for text based retrieval and Blast and FastA for sequence based retrieval.

DNA Databases

- GeneBank: http://www.ncbi.nlm.nih.gov/
  - sequence database from National Center Biotechnology Information (NCBI).
  - It incorporates sequences from publicly available sources

Protein Databases

- PIR – Protein Sequence Database
  - http://pir.georgetown.edu/
  - was developed in the early 1960's.
  - four sections:
  - **PIR1** - fully classified and annotated entries.
  - **PIR2** - preliminary entries, not thoroughly reviewed.
  - **PIR3** - unverified entries, not reviewed.
  - **PIR4** - conceptual translations.

Protein Databases

- Swiss-Prot: http://us.expasy.org/sprot/
  - Established in 1986
  - Provides high-level annotations, including description of protein function, structure of protein domains, post-translational modi.cations, variants, etc. It aims to be minimally redundant.
- TrEMBL - Translated EMBL
  - was created in 1996
  - It contains translations of all coding sequences in the EMBL nucleotide sequence database.
  - SP-TrEMBL contains entries that will be incorporated into Swiss-Prot
  - REM-TrEMBL contains entries that are not destined to be included in Swiss-Prot,

Protein Databases

- **GenPept** http://www.ncbi.nlm.nih.gov/
  - GenPept is a supplement to the GenBank nucleotide sequence database.
  - translations of coding regions in GenBank entries.
- **NRL_3D**
  http://pir.georgetown.edu/pirwww/dbinfo/nrl3d.html
  - produced and maintained by PIR.
  - contains sequences extracted from the Protein DataBank (PDB)

Protein Databases

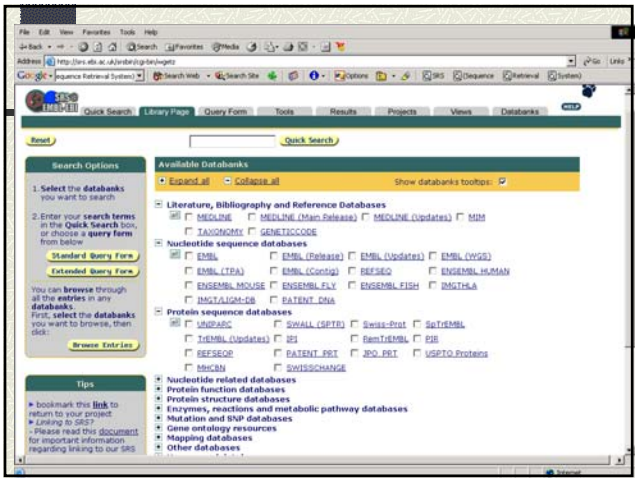
- **Summary of protein sequence databases**
  - PIR(1-4) - comprehensive, poor quality of annotation (even in PIR1).
  - Swiss-Prot - poor sequence coverage, highly structured, excellent annotation.
  - GenPept most comprehensive, poor quality of annotation.
  - NRL 3D - least comprehensive but is directly relating to structural information.

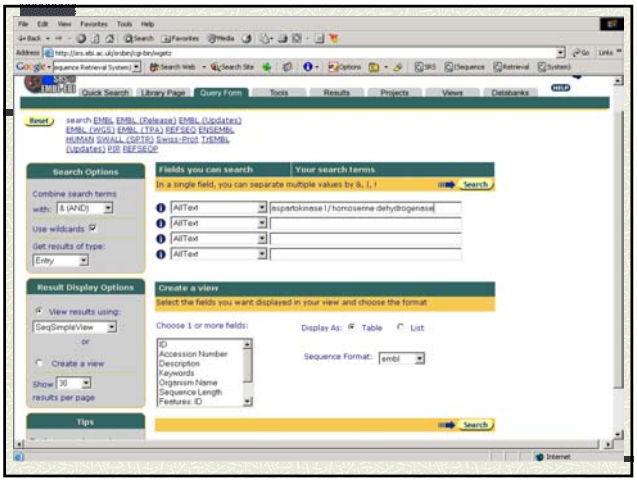
Database searching

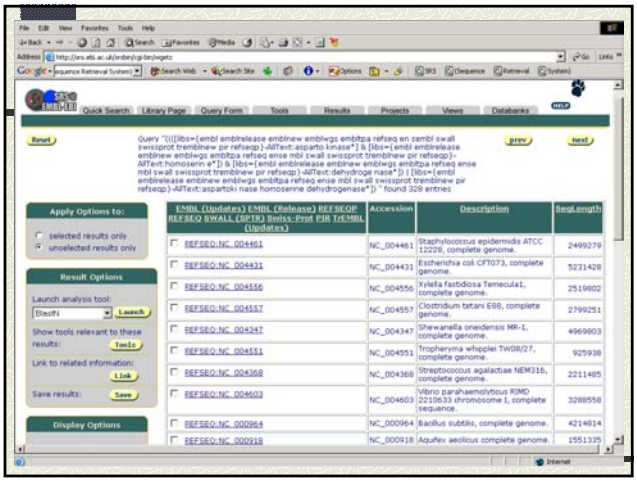
- **Text based search** - Searching the annotations. Examples:SRS, GCG's Lookup, Entrez.
- **Sequence based search** - Searching the sequence itself. Examples:Blast, FastA, SW.

Text based retrieval tools

- **SRS (Sequence Retrieval System)**
  - http://srs.ebi.ac.uk/
  - Developed by EBI
  - provides a homogeneous interface to over 80 biological databases
  - includes databases of sequences, metabolic pathways, transcription factors, application results (like BLAST, SSEARCH, FASTA), protein 3-D structures, genomes, mappings, mutations, and locus specific mutations.
  - Before entering a query, one selects one or more of the databases to search.
  - It is possible to send the query results as a batch query to a sequence search tool.

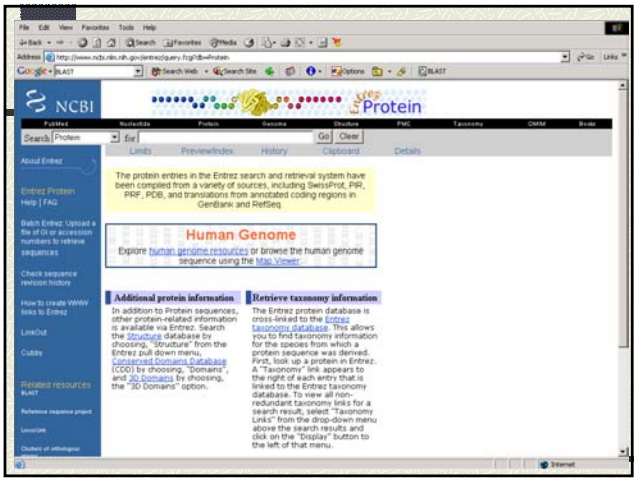


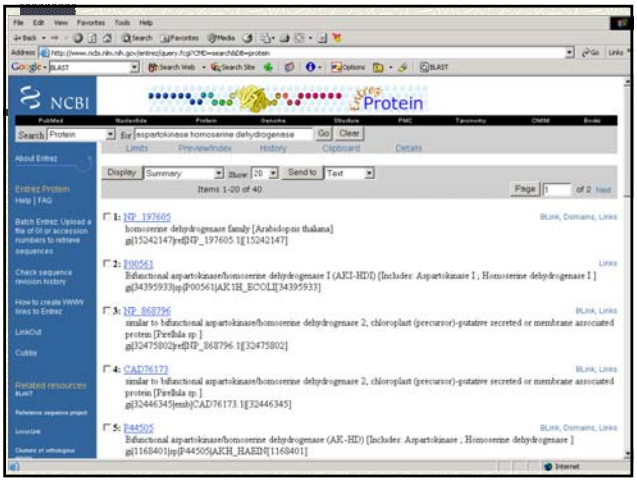




Text based retrieval tools

- **Entrez**
  - http://www.ncbi.nlm.nih.gov/Entrez/
  - Developed at NCBI
  - entry point for exploring the NCBI's integrated databases.
  - easy to use, but unlike SRS, the search is limited.





Sequence Based Searching

- **DNA search versus Protein search**
  - A DNA sequence is a string of length *n* over an alphabet of size 4. Its protein translation is a string of length *n*/3 over an alphabet of size 20. Statistically, the expected number of random matches in some arbitrary database is larger for a DNA sequence.
  - DNA databases are much larger than protein databases, and they grow faster. This also means more random hits.
  - Translation of a DNA sequence to a protein sequence causes loss of information.
  - Protein sequences are more biologically preserved than DNA sequences.

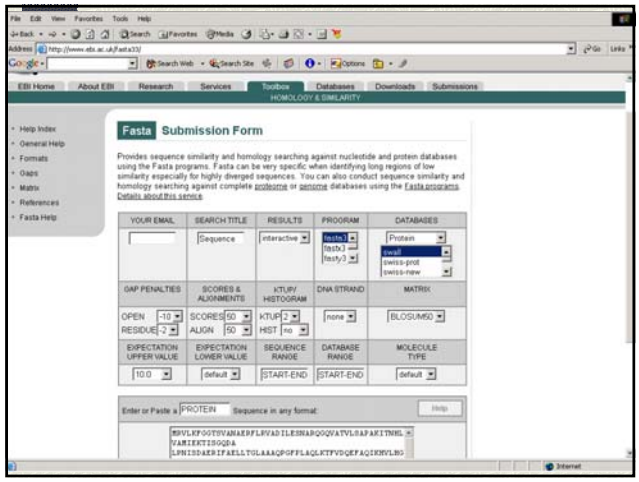
FastA

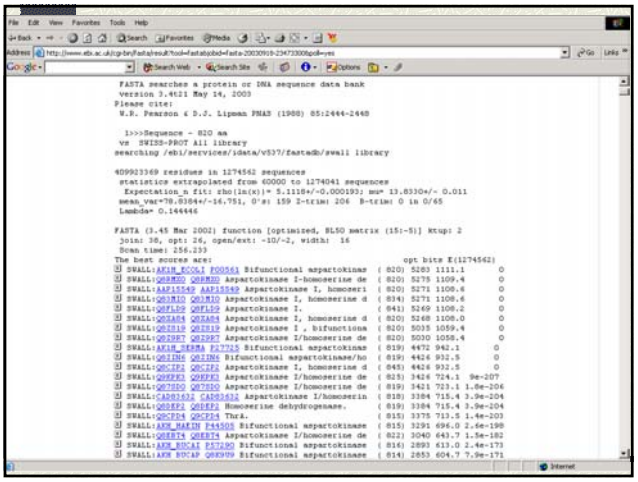
PROGRAM	FUNCTION
fasta3	scan a protein or DNA sequence library for similar sequences
fastx/y3	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
tfastx/y3	compares a protein to a translated DNA data bank
fasts3	compares linked peptides to a protein databank
fastf3	compares mixed peptides to a protein databank

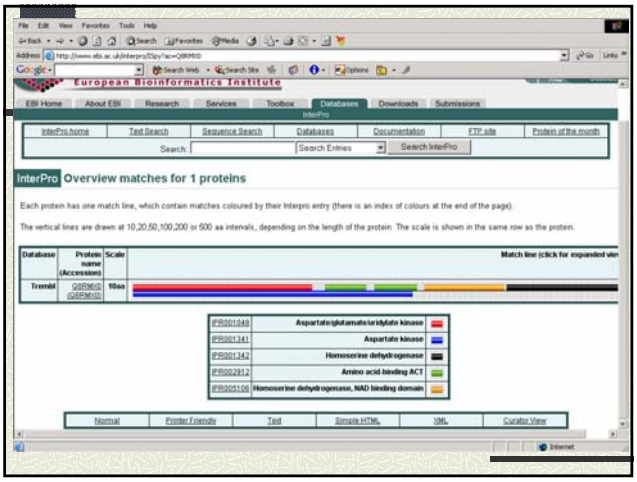
- http://www.ebi.ac.uk/fasta33/
- Under different circumstances it is favorable to use different programs:
  - To identify an unknown protein sequence use either FastA3 or tFastX3.
  - To identify structural DNA sequence:(rep eated DNA, structural RNA) use FastA3, first with *ktup* = 6 and then with *ktup* = 3.
  - To identify an EST use FastX3 (check whether the EST codes for a protein homologous to a known protein).
  - Use *ktup* = 1 for oligonucleotides (length < 20).

FastA

- Output
  - The standard FastA output contains a list of the best alignment scores and a visual representation of the alignments.
  - Sequences with E-score less than 0.01 are almost always found to be homologous.
  - Sequences with E-score between 1 and 10 frequently turnout to be related as well.

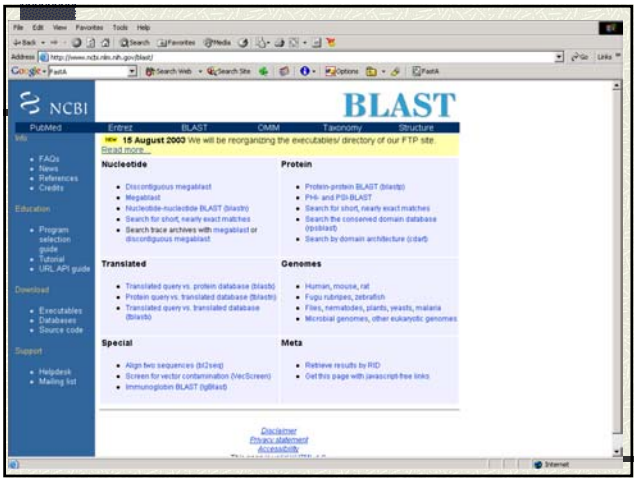


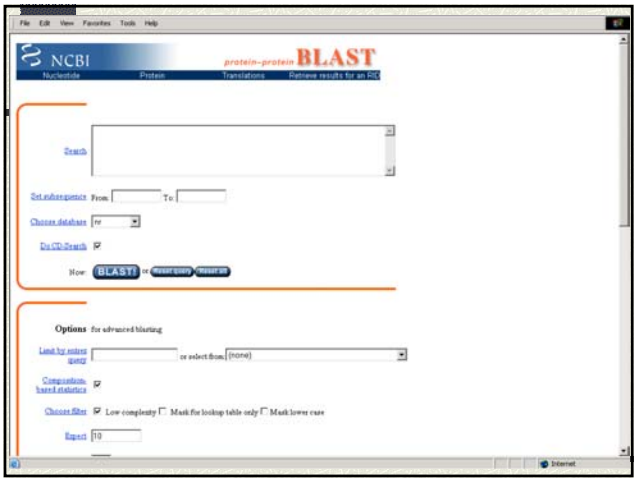


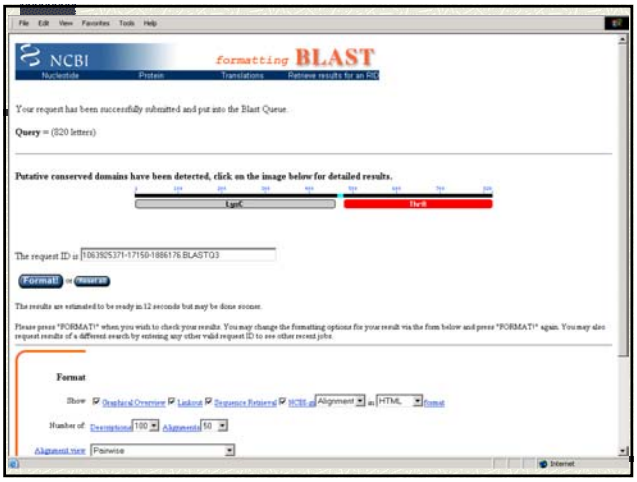


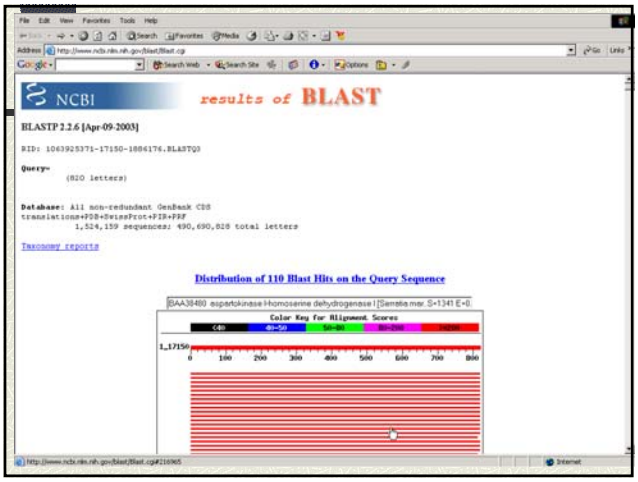
BLAST - Basic Local Alignment Search Tool

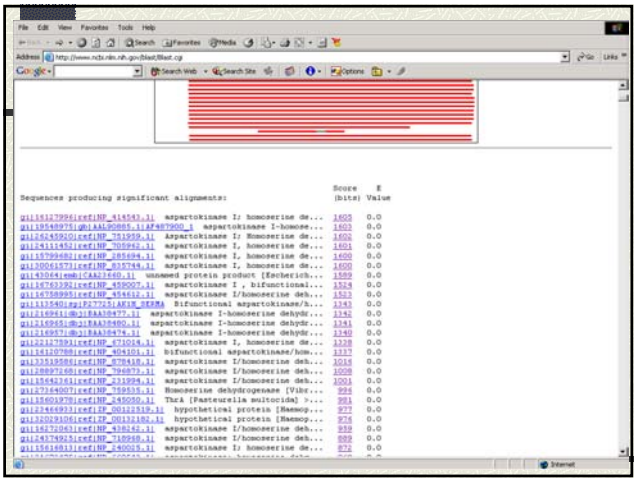
- BLAST programs were designed for fast database searching,with minimal sacrifice of sensitivity for distantly related sequences.
- http://www.ncbi.nlm.nih.gov/BLAST/

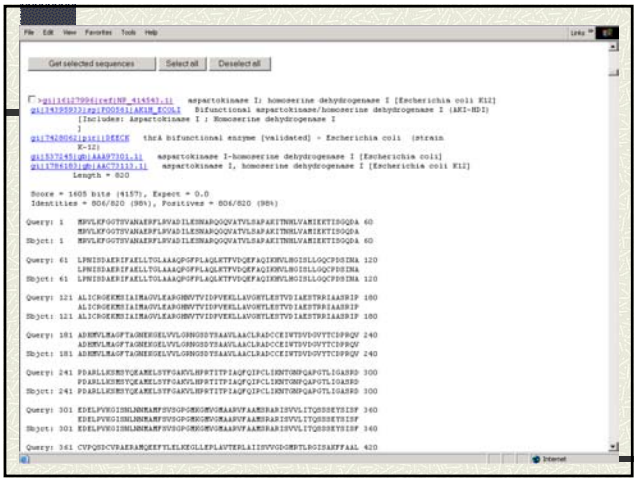





NCBI
formatting BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Your request has been successfully submitted and put into the Blast Queue.
Query = (820 letters)
Putative conserved domains have been detected, click on the image below for detailed results.
LysC
ThrB
The request ID is 1063925371-17150-1886176.BLASTQ3
Format!
reset all
The results are estimated to be ready in 12 seconds but may be done sooner.
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Alignment
as
HTML
format
Number of Descriptions
100
Alignments
50
Alignment view
Pairwise


NCBI
results of BLAST
BLASTP 2.2.6 [Apr-09-2003]
RID: 1063925371-17150-1886176.BLASTQ3
Query=
(820 letters)
Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF
1,524,159 sequences; 490,690,828 total letters
Taxonomy reports
Distribution of 110 Blast Hits on the Query Sequence
BAA38480 aspartokinase-Homoserine dehydrogenase I [Serratia mar. S=1341 E=0
Color Key For Alignment Scores
1..1715
0
100
200
300
400
500
600
700
800


Sequences producing significant alignments:
Score (bits)
E Value


Get selected sequences
Select all
Deselect all
Query: 1 MPVLKFGGTSVANAERFLRVADILESNARQQQVATVLSAPAKITNHLVAMIEKTISGQDA 60
Sbjct: 1 MPVLKFGGTSVANAERFLRVADILESNARQQQVATVLSAPAKITNHLVAMIEKTISGQDA 60
Query: 61 LPNISDAERIFAELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINA 120
Sbjct: 61 LPNISDAERIFAELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINA 120
Query: 121 ALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIP 180
Sbjct: 121 ALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIP 180
Query: 181 ADHMVLAAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240
Sbjct: 181 ADHMVLAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV 240
Query: 241 PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300
Sbjct: 241 PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300
Query: 301 EDELPVKGISNLSNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEYSISF 360
Sbjct: 301 EDELPVKGISNLSNMAMFSVSGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEYSISF 360
Query: 361 CVPQSDCVRAERAMQEEFYLELKEGLLEPLAVTERLAIISVVGDGMRTLRGISAKFFAAL 420

Comparison of the Programs

- Concept:
  - BLAST produce local alignments, while FastA is a global alignment tool. BLAST can report more than one HSP per database entry, while FastA reports only one segment(match).
- Speed:
  - BLAST > FastA
  - BLAST (package) is a highly e.cient search tool.

Comparison of the Programs

- Sensitivity:
  - FastA > BLAST (old version!)
  - FastA is more sensitive, missing less homologous sequences on the average (but the opposite can also happen - if there are no identical residues conserved, but this is infrequent). It also gives better separation between true hits and random hits.
- Statistics:
  - BLAST calculates probabilities, and it sometimes fails entirely if some of the assumptions used are invalid.
  - FastA calculates significance 'on the fly' from the given dataset which is more relevant but can be problematic if the dataset is small.