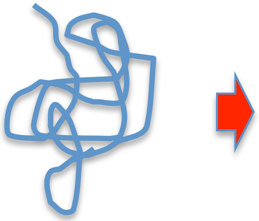
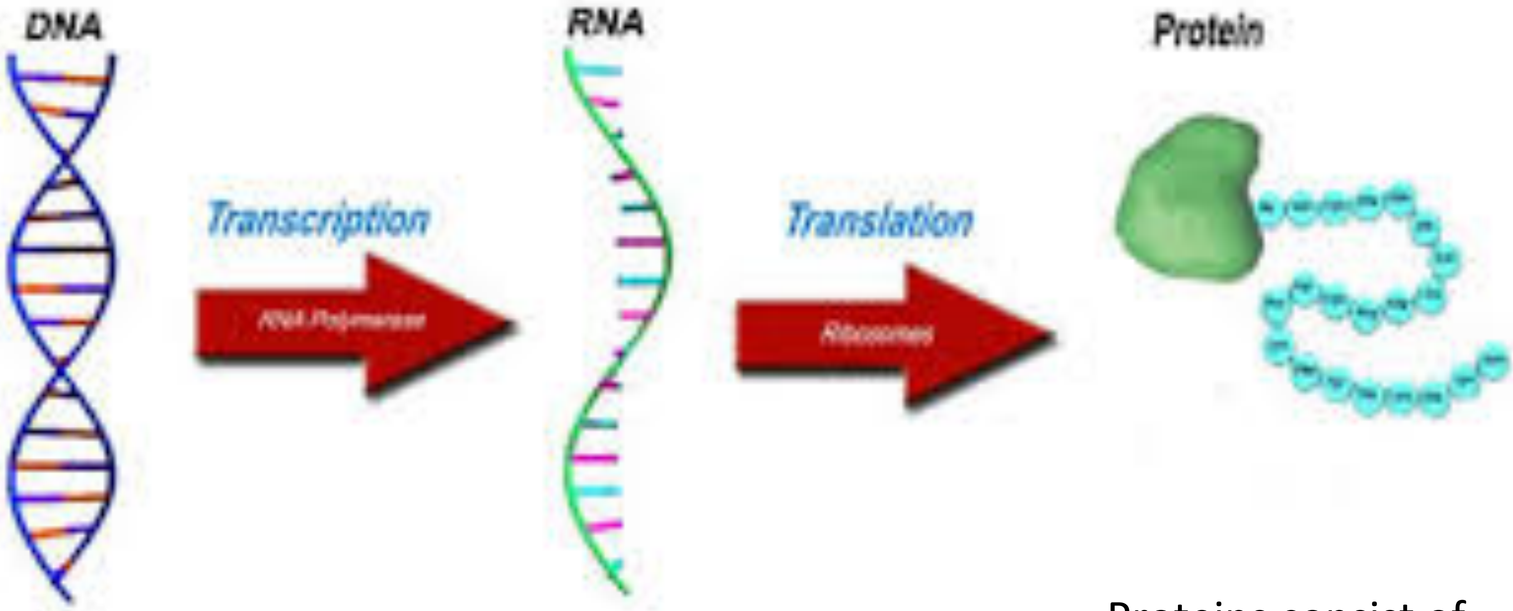
Lecture 3

Sequence analysis and databases search

Recap some ideas





Genome consists of entire DNA sequence of a species

DNA is the *blueprint.* Contains individual *genes.* Genes are composed of four nucleotides (A,C,G,T)

RNA is the transition of DNA to proteins.
It is composed of four nucleotides (A,C,G,U)

Proteins consist of 20 amino acids RAYFT ….
Proteins perform the actual *functions*

More ideas

- One of the major goals of bioinformatics is to assign the function of a gene or protein by sequence comparison

Seq1-AGTGTGACGTGTGC   Nucleotides
Seq2-ACTGTGACCTGTGC

Seq1-MAPKPEPKKEA   aminoacids
Seq2-MPPKPEPKKET

Finding Function By Sequence Similarity

Finding Function By Sequence Similarity

- Similarity indicates conserved function
- Comparing sequences helps us understand function
- Locate similar gene in another species to understand your new gene
- Similarity could tell (sometimes) evolutionary relation.
- IMPORTANT: similarity is not Transitive!!!

similarity not transitive!

- If SeqA is “similar” to Seq2, and Seq3 is “similar” To Seq2. Seq1 is not always similar to Seq3

- YYYYYY is similar to XXXXXXXYYYYYY
- XXXXXX is similar to XXXXXXXYYYYYY
- Is YYYYYY similar to XXXXXX ? … NO!!!!

- Unless the regions of similarity (alignments) are overlapping!

Identifying Similarity

- Algorithms to match sequences
  - Needleman-Wunsch
  - Smith Waterman
  - BLAST

Needleman-Wunsch

- Global alignment algorithm
- Example
- align UNIVGEORGIA and FLORIDA
- UNIVGEORGIA
- -------FLOR-IDA

- SCORE FOR MATCH +1, AND -1 FOR GAPS OR MISMATCHES

Needleman-Wunsch
GLOBAL

Smith-Waterman

- Modification of Needleman-Wunsch
- It looks for local alignments (partial)

BLAST

- Basic Local Alignment Search Tool
- Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, and David J. Lipman (NIH)

- Set of programs that search sequence databases for statistically significant similarities
- Five traditional BLAST programs:
  - BLASTN – nucleotides
  - BLASTSP, BLASTX, TBLASTN, TBLASTX - proteins

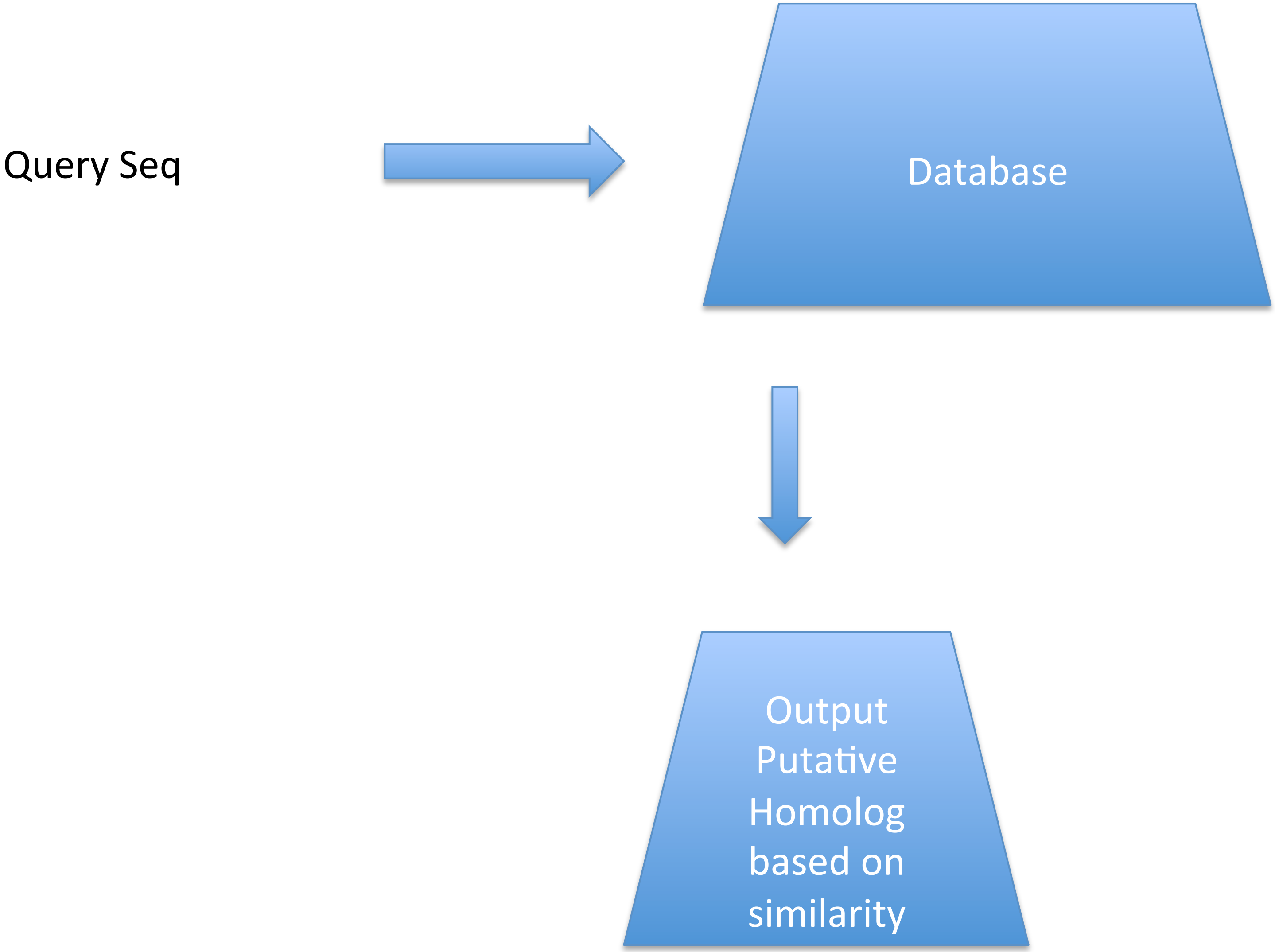
BLAST programs

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database.You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Other BLAST programs

Program	Notes
Megablast	Nearly identical sequences (Nucleotide only)
PSI-BLAST	Automatically generates a position specific score matrix (PSSM) (Protein only)

The BLAST Search Algorithm



Query Seq

Database

Output Putative Homolog based on similarity

The BLAST Search Algorithm

Query Word (W=3)

GSVEDTTTGSQQSLAAALLNKCKTPQGQLRVNQWIKPQMDKNRIEERLNLVAAFVEDAEL

Neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PHG	13
PQA	12
PQN	12
etc…	

Neighborhood score threshold (T=13)

```
Query    325  SLAALLNKCKTPQGQRLVNQWIKQLMDKNRI
              +LA++L+   TP G R+       LM    +
sbject   290  TLASVLDCTVTPNGSRILHMVRDLMKNTSSL
```

High-scoring Segment Pair (HSP)

Matrix

BLOSUM 45	**BLOSUM 62**	**BLOSUM 90**
PAM 250	**PAM 160**	**PAM 100**
More Divergent		***Less Divergent***

BLOSUM 62 is the default matrix in BLAST

The BLAST Search Algorithm

Query Word (W=3)

GSVEDTTTGSQQSLAAALLNKCKTPQGQLRVNQWIKPQMDKNRIEERLNLVAAFVEDAEL

Neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PHG	13
PQA	12
PQN	12
etc…	

Neighborhood score threshold (T=13)

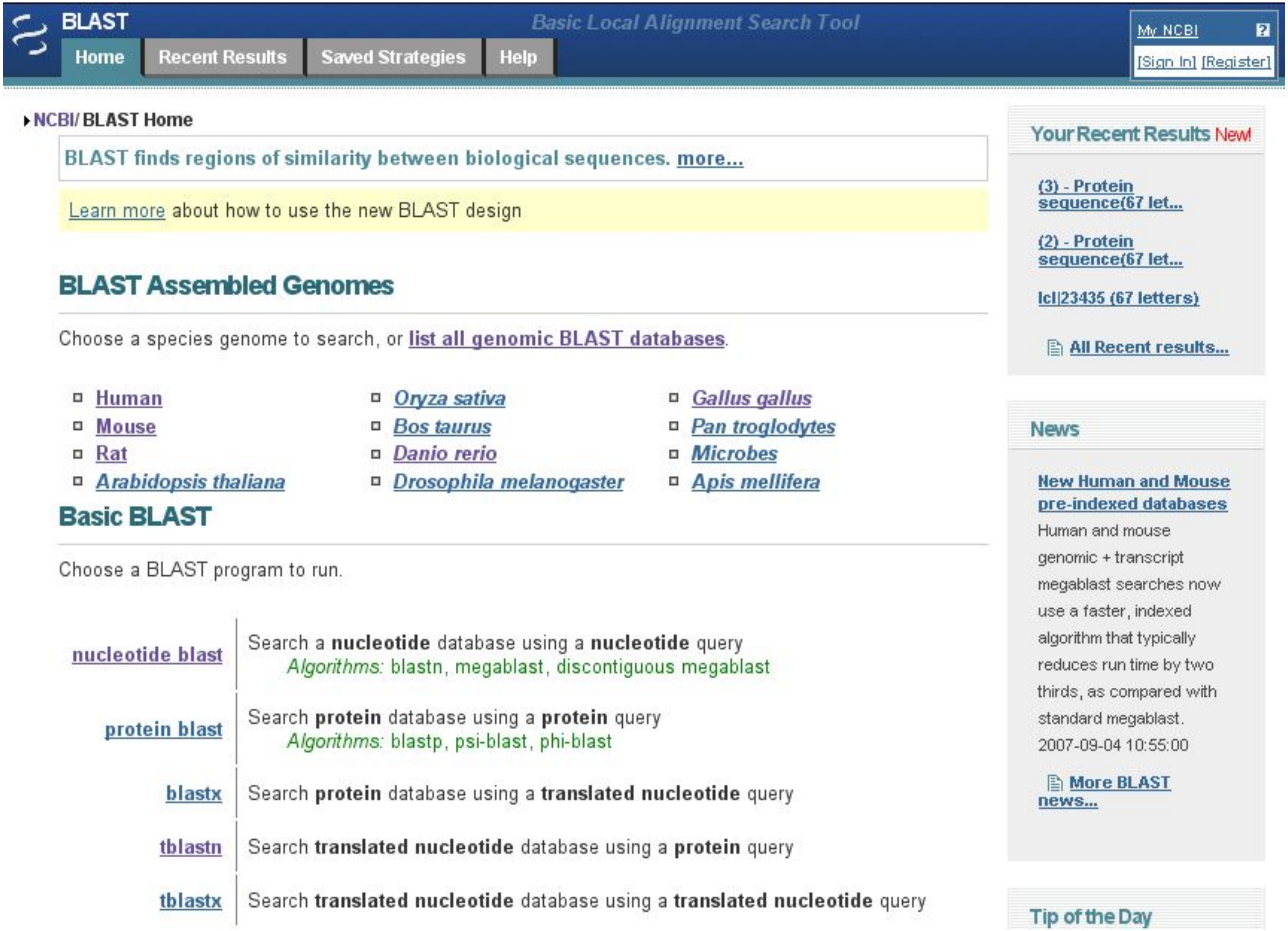
```
Query   325  SLAALLNKCKTPQGQRLVNQWIKQLMDKNRI
             +LA++L+   TP G R+       LM    +
sbject  290  TLASVLDCTVTPNGSRILHMVRDLMKNTSSSL
```

High-scoring Segment Pair (HSP)

Score and the e-value

- The quality of the alignment is represented by the Score (S).
- The score of an alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (PAM, BLOSUM) whereas gap scores are assigned empirically .
- The significance of each alignment is computed as an E value (E).
- Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

Blast online in NCBI



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Your Recent Results New!

(3) - Protein sequence(67 let...

(2) - Protein sequence(67 let...

lcl|23435 (67 letters)

All Recent results...

News

New Human and Mouse pre-indexed databases

Human and mouse genomic + transcript megablast searches now use a faster, indexed algorithm that typically reduces run time by two thirds, as compared with standard megablast.
2007-09-04 10:55:00

More BLAST news...

Tip of the Day

BLAST in your computer

-BLAST command Line for batch searches

-We will use the Blast program installed in the hipergator

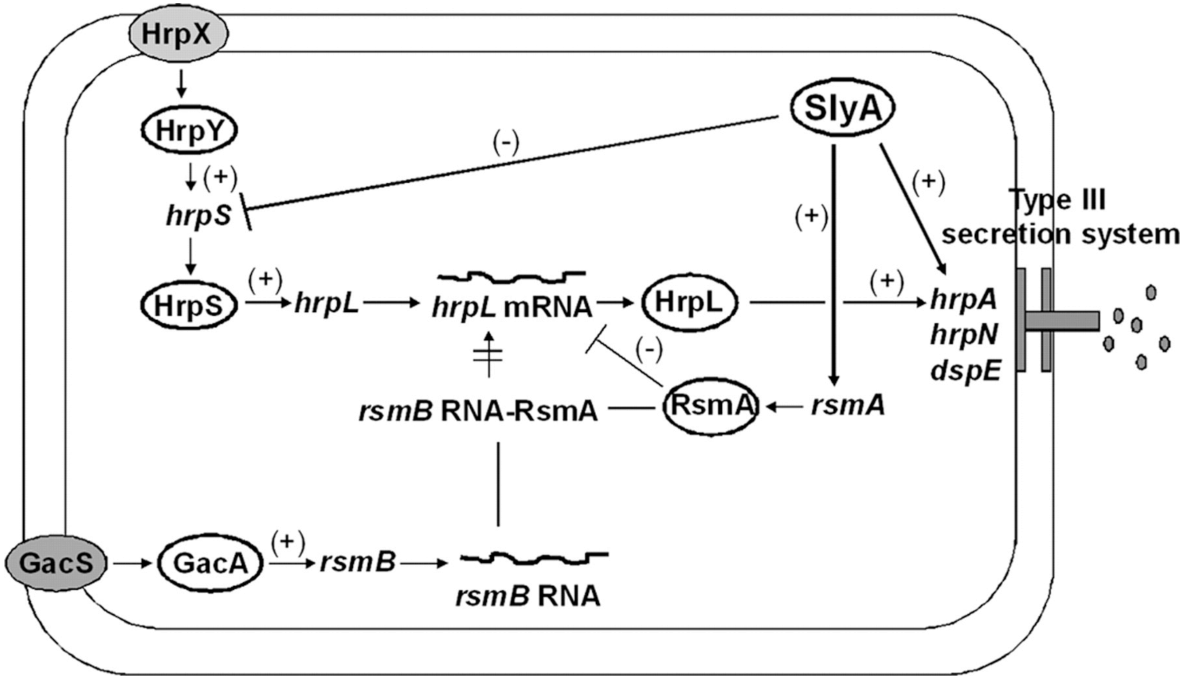
```
module load ufrc
srundev -t 60 -m=4G
module load ncbi_blast
```

Basic commands

- makeblastdb –in your_sequences_db –dbtype nucl or prot
- blastn -query your_sequences –db your_database –out your_results –outfmt 1 (full format), 6 (tab format)
- *if you want to blast proteins change blastn for blastp or use the other programs blastx, tblastn , tblastx, depending of your query and/or db

Exercise 1
Local blast with your own databases

- HrpA is an essential component of the type III secretion system (TTSS) which pathogens use to inject virulence factors directly into their host cells, and to cause disease. The TTSS has an Hrp pilus appendage for channelling effector proteins through the plant cell wall and this pilus elongates by the addition of HrpA pilin subunits at the distal end.



Exercise 1

- You have two files that contain the aminoacid (PseudomonasX.faa) and nucleotide (PseudomonasX.fna) sequences of the genome of Pseudomonas X
- You have a protein sequence of hopY1 (hopY1.faa) an important component of the TSS, obtained from the Pseudomonas database

- Using Blast , check is Pseudomonas X has an homologue of hopY1.

Lets start a session with the hipergator (interactive)
module load ufrc
srundev –t 120 –m=2g

- module load ncbi_blast

- FIRST:
- Build the blastdb of Pseudomonas.faa and Pseudomonas.ffn
- SECOND
- Lets use as a query hopY1.faa
- Search Pseudomonas.faa using HopY1.faa to output file named hrpBlastp
- Search Pseudomonas.ffn using HopY1.faa to output file named hrptBlastn
- Try different formats for the output (outfmt 1,2,3,4,5,6,7) and compare

Exercise 2
Nucleotide identity using blast
and parsing results with unix commands

- Samples of soil were submitted to your laboratory. You need to discard the presence of plant pathogens in this sample. The lab technician used universal primers to amplify rDNA from bacteria. After processing all the amplicons, the sequences were saved in a multi-fasta file ampli_queries.fna. In addition you have a Database with rDNA of plant pathogens (hopY1.faa)
- Using blast tools lets investigate if your amplicons contain plant pathogens.
- "a prokaryotic species is considered to be a group of strains that are characterized by a certain degree of phenotypic consistency, showing over 97% of 16S ribosomal RNA (rDNA) gene-sequence identity"

Exercise 2

- 1. Format your database first
- 2. check the content of your fasta files
- 3. Obtain the best hit among all the amplicons (based on identity third column)
- 4. Is there other amplicon whit Hits (more than 97% 3th column) ?
- UNIX COMMANDS in ACTION!!
- 5. obtain the best hit for amplicon AMP3
- 6. Generate a report of hits. The report should display 3 columns with the 10th best hits, first column should list the amplicon ID, The second column the species name and the third the identity percentage
- Generate a ordered report with the best hits for every amplicon (hint type blast –h to display more options)

Where do you find databases for nucleotides and proteins

- ftp://ftp.ncbi.nih.gov/genbank

- Databases for plnat pathogens:
- http://cpgr.plantbiology.msu.edu/
- http://www.pathoplant.de/

- Formats:
- GBK or GBS format: genbank format
- Fna format :fasta format for nucleotides
- Fnn format: fasta format for nucleotides
- GFF PPT formats: tabular formats
- Asn format