CENTER FOR
GENOME RESEARCH &
BIOCOMPUTING

"Introduction to Unix/Linux"
INX_U18, Day 7, 2018-08-08

Basic Local Alignment Search Tools (BLAST) and "blasting"

Learning Outcome(s):

Describe the differences between the different BLAST alignment programs, i.e., algorithms, identify several important options for these programs, and run them to produce tab-delimited output.

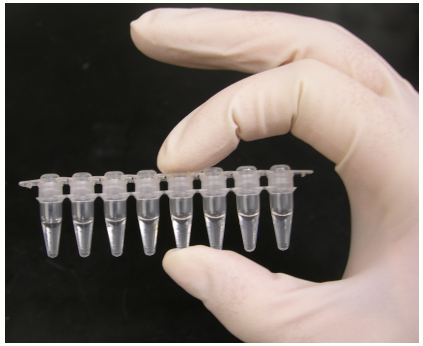
Matthew Peterson, OSU CGRB, matthew@cgrb.oregonstate.edu

Basic Local Alignment Search Tool

- **BLAST** was developed in the 90s and its paper has been cited tens of thousands of times!
- It is an algorithm to compare **query** sequences against a **subject** set, e.g., composed of:
  - Nucleotides of DNA: **A**, **C**, **T**, **G** (and **N**)
  - Protein format (Amino-acids):
    - Glycine **G** = GGT, GGC, GGA, GGG
    - Lysine **K** = AAA, AAG
    - etc..

Oregon State UNIVERSITY

BLASTing Example



Do my sample sequences **match:** **?**

CC 0, Credit: Madeleine Price Ball),https://commons.wikimedia.org/wiki/File:PCR_tubes.png
CC BY 4.0 attribution
Figure 1: Living things may be single-celled or complex, multicellular organisms. They may be plants, animals, fungi, bacteria, or archaea. This diversity results from evolution. (credit "wolf": modification of work by Gary Kramer; credit "coral": modification of work by William Harrigan, NOAA; credit "river": modification of work by Vojtěch Dostál; credit "fish" modification of work by Christian Mehlführer; credit "mushroom": modification of work by Cory Zanker; credit "tree": modification of work by Joseph Kranak; credit "bee": modification of work by Cory Zanker)
Download for free at http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@10.117."
http://philschatz.com/biology-book/contents/m44575.html

Oregon State UNIVERSITY

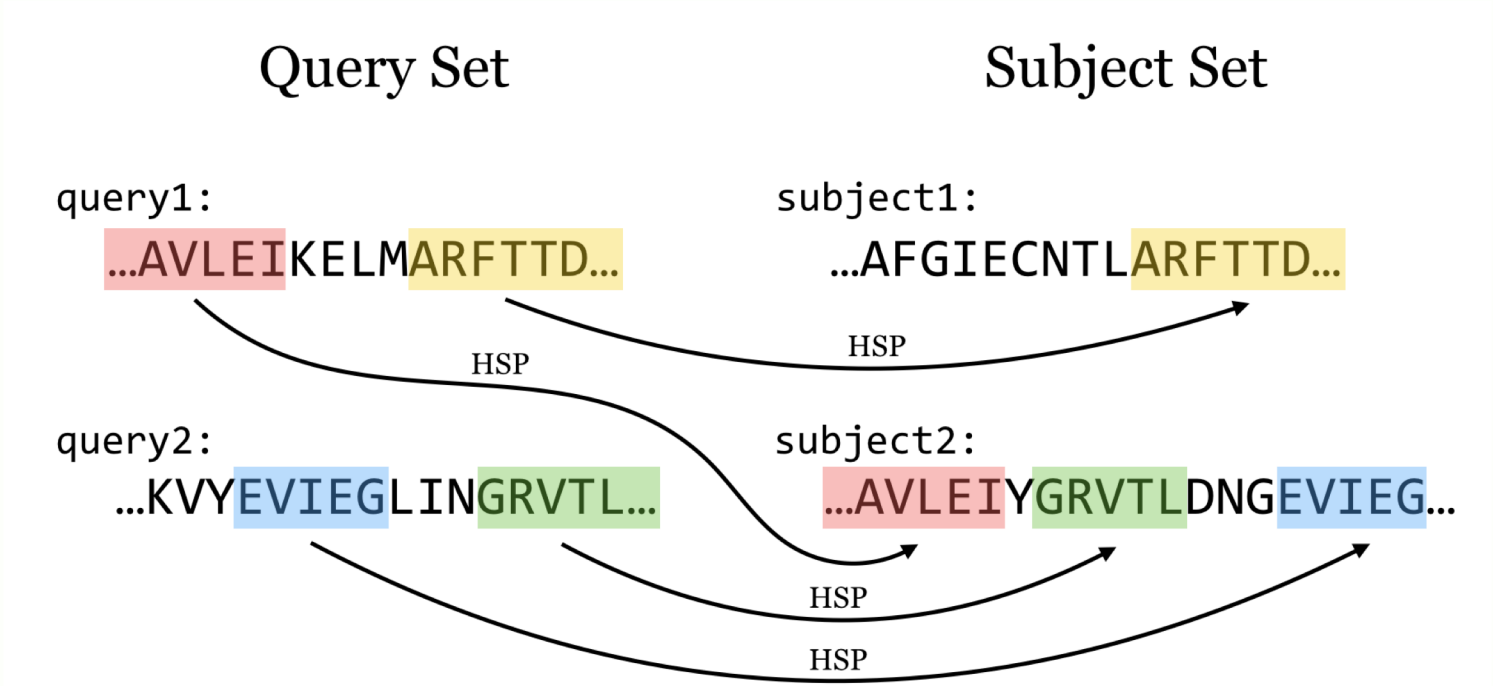
Other common uses for BLAST

- Identifying species
  - "identify a species or find homologous species … using DNA from an unknown species" (good to determine contamination in your samples)
- Locating domains
  - Find known (protein) domains (within your query sequence)
- Establishing phylogeny
  - Create an (approximated) phylogenetic tree using the BLAST web-page
- DNA mapping
  - Find similar genes in your query sequence based upon location in other organisms' database(s)
- Comparison
  - "Locate common genes in two related species"

Oregon State UNIVERSITY

https://en.wikipedia.org/wiki/BLAST#Uses_of_BLAST

Query vs. Subject sets

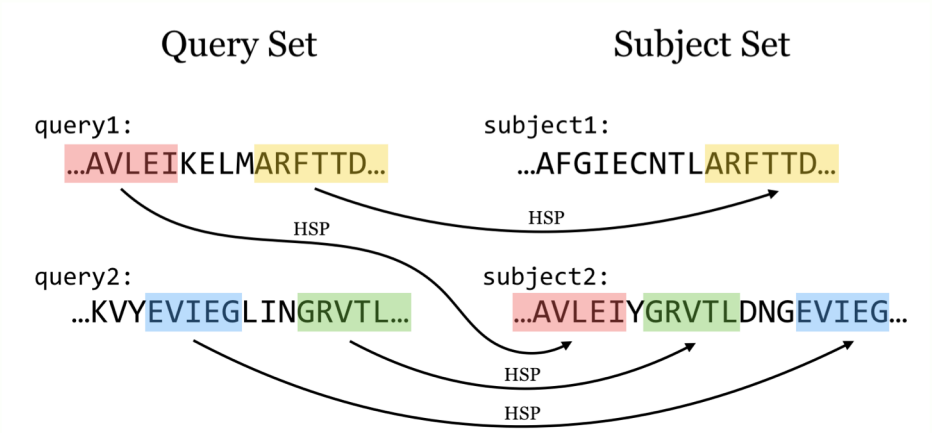


HSP = High Scoring Pair

http://library.open.oregonstate.edu/computationalbiology/chapter/command-line-blast/

HSPs, scoring, and e-values

- Each *close* match of a High Scoring Pair (HSP) is considered a "**hit**" (there can be more than 1 hit)
- Each HSP is assigned a **score** as to how close the **subject** matched the **query** sequence
- Each HSP is associated with an **e-value**, i.e., the chance a hit was found by chance



e-value (Expect value)

- As the **score** of an alignment increases, the **e-value** (number of **hits** expected by chance) will decrease exponentially. “Essentially, the E value describes the random background noise”
- As the **e-value** approaches 0 (less chance) the match becomes more "significant"
- What is the "best" **e-value**? It's *just an estimate* (influenced by length of match, database size, etc.) Adjust your e-value, re-run, audit. Also read the methods sections of papers using BLAST.

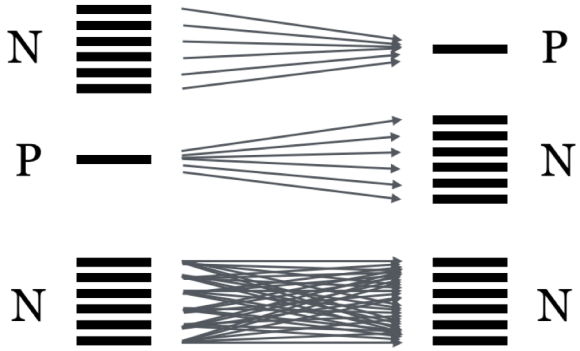
Oregon State UNIVERSITY

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect

BLAST+ suite of tools

- **BLAST**+ from NCBI is a set of tools for **N**ucleotide and **P**rotein comparisons

Program	Query Type	Subject Type	Computation
blastn	N	N	~ 1X
blastp	P	P	~ 1X
blastx	N	P	~ 6X
tblastn	P	N	~ 6X
tblastx	N	N	~36X

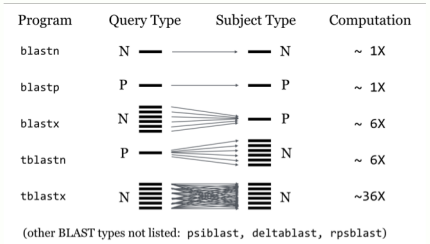


(other BLAST types not listed: `psiblast`, `deltablast`, `rpsblast`)

Oregon State UNIVERSITY

Comparisons

- **blastn** = **N**ucleotide to **N**ucleotide
- **blastp** = **P**rotein to **P**rotein
- **blastx**, **tblastn** compare **N** to **P** and **P** to **N**



Program	Query Type	Subject Type	Computation
blastn	N	N	~ 1X
blastp	P	P	~ 1X
blastx	N	P	~ 6X
tblastn	P	N	~ 6X
tblastx	N	N	~36X

(other BLAST types not listed: psiblast, deltablast, rpsblast)

by converting **N**ucleotide sequences into **P**rotein sequences in all 6 reading frames (three on the forward DNA strand and three on the reverse), ~6x the time of work (CPU cycles)

```
1.  ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2.  A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3.  AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A
```

(**Start** codon (**M**ethionine) vs. **Sto**p codons, 5' to 3')

https://en.wikipedia.org/wiki/Open_reading_frame

Oregon State UNIVERSITY

BLOSUM / PAM Scoring matrices

- Scoring matrices account for mutations (evolution) of (protein) sequences over time
- **BLO**cks **SU**bstitution **M**atrix (**BLOSUM**)
- **P**oint **A**ccepted **M**utation (**PAM**) matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

https://en.wikipedia.org/wiki/BLOSUM
https://en.wikipedia.org/wiki/Point_accepted_mutation

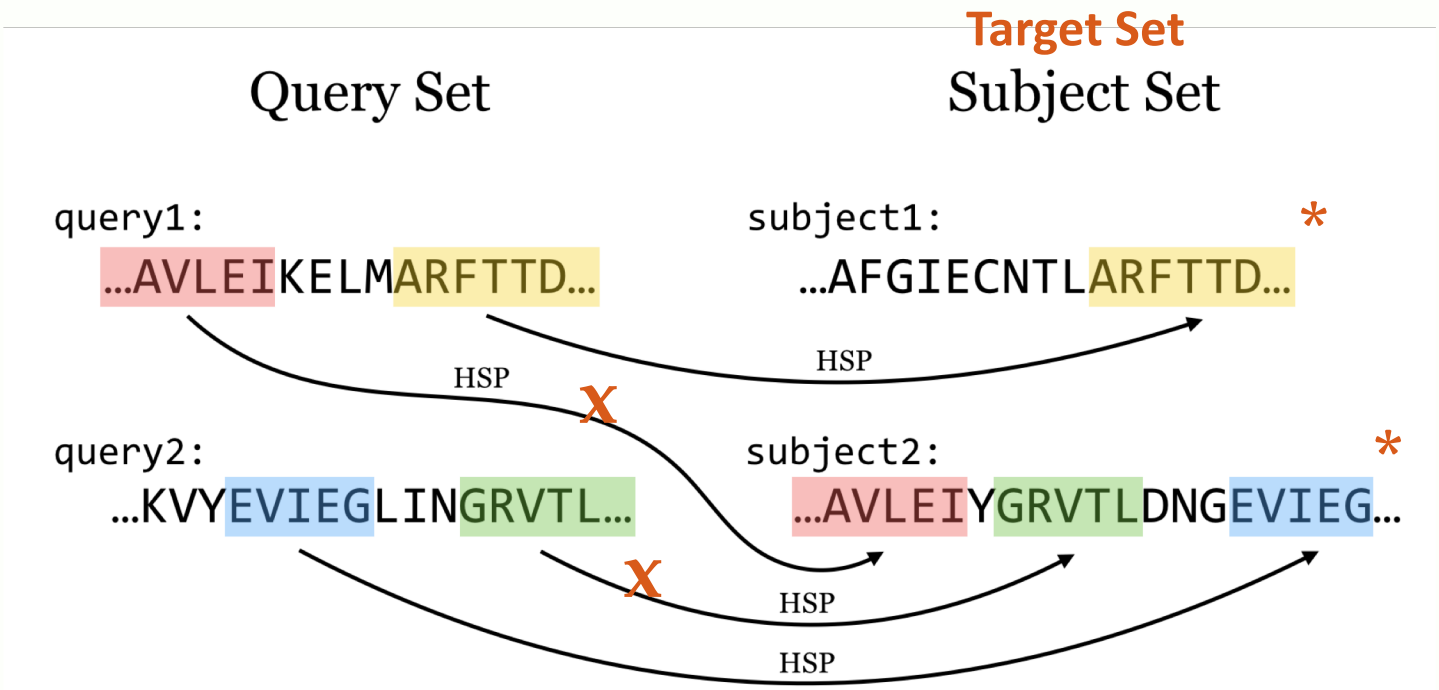
Oregon State UNIVERSITY

`blastn -help` options

-query <fasta file>
-subject <fasta file>
-evalue <real number> (e.g., 0.001 or 1e-6)
-outfmt <integer> (default 0)
-max_target_seqs <integer>
For `outfmt` format 6, 7, or 10, only report HSPs for the best <integer> different subject sequence
-max_hsps <integer>
For each query/target pair, only report the best <integer> HSPs.
-out <output file>

Oregon State UNIVERSITY

-max_target_seqs=1 –max_hsps=1



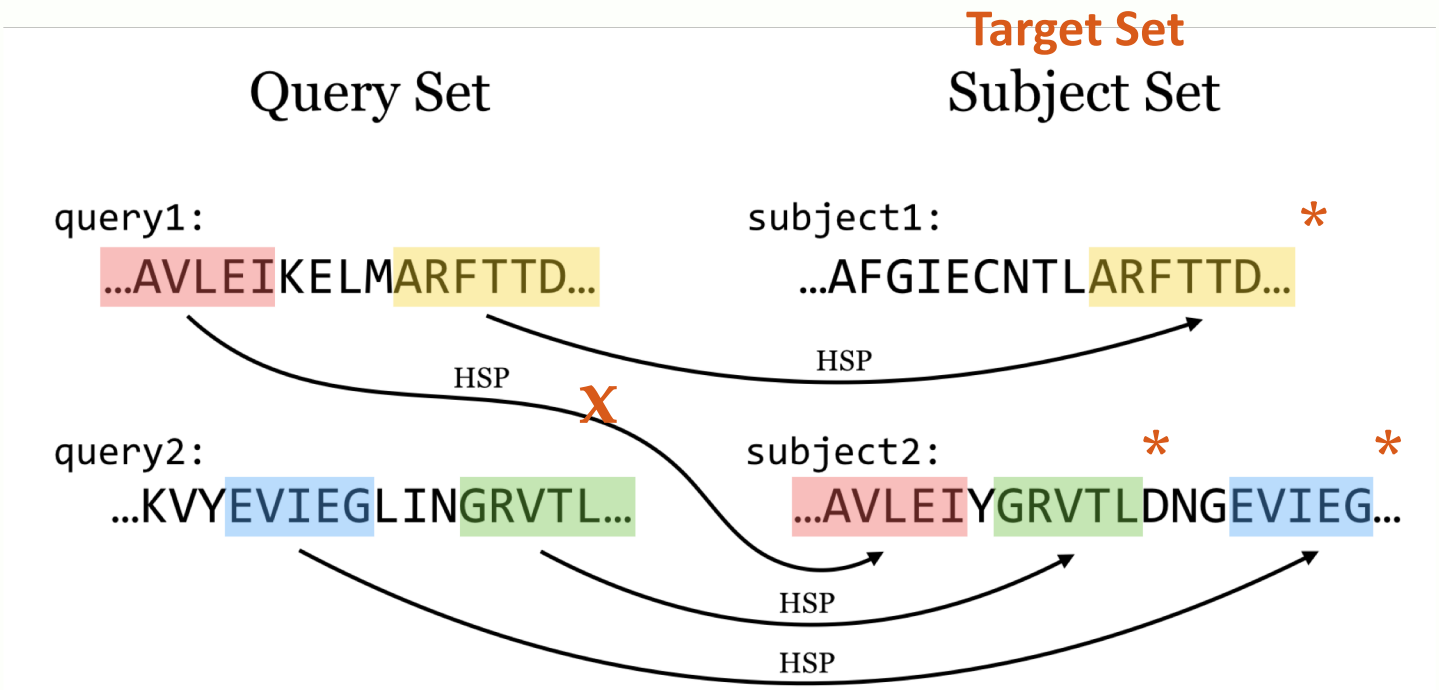
query1:subject1 = ARFTTD… (subject2 not included)

query2:subject2 = EVIEG (assume blue is better)

Only 1 sequence with 1 **HSP** per "hit"

Oregon State University

-max_target_seqs=1 –max_hsps=2



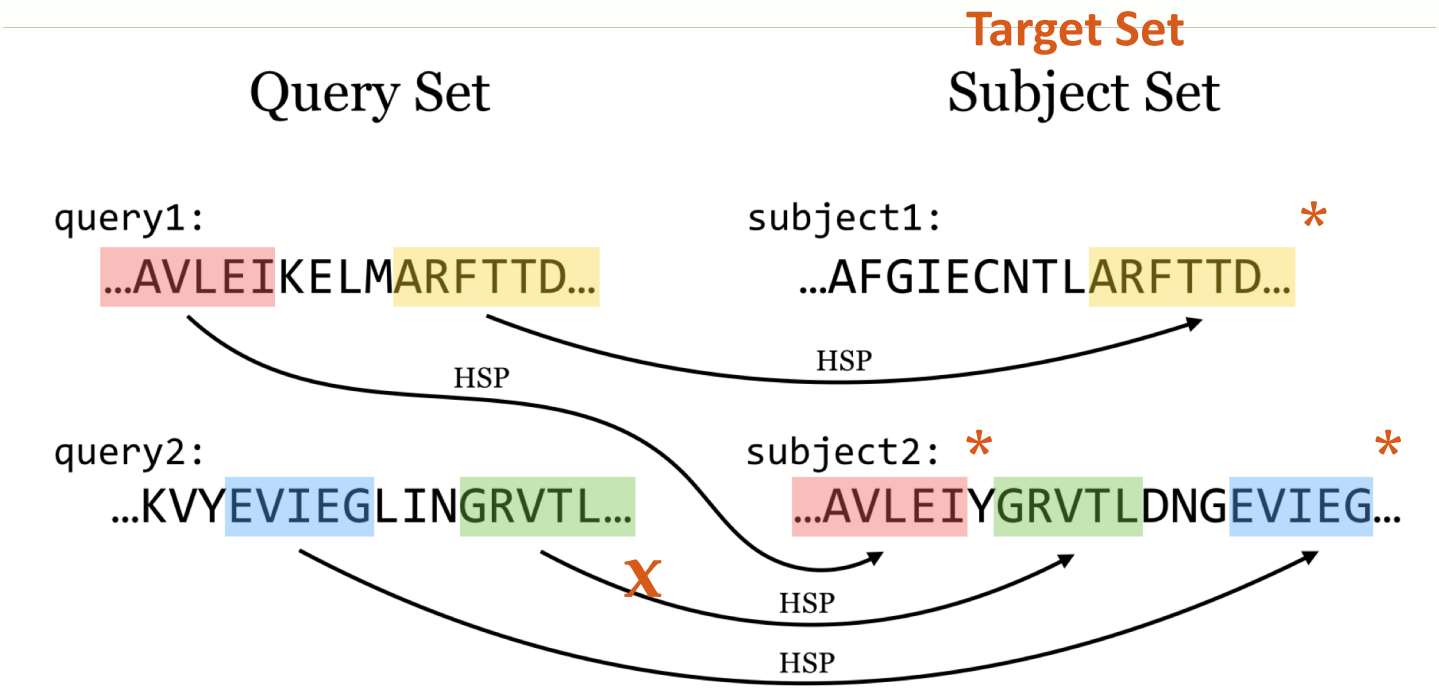
query1:subject1 = ARFTTD… (subject2 not included)

query2:subject2 = EVIEG (assume blue is best)

query2:subject2 = GRVTL (2nd best **HSP**)

Oregon State University

-max_target_seqs=2 –max_hsps=1



query1:subject1 = …ARFTTD…

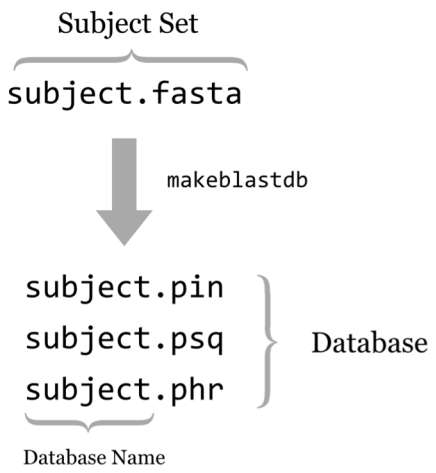
query1:subject2 = …AVLEI

query2:subject2 = EVIEG (blue is ~ best)

Oregon State University

BLAST databases

- BLAST is not efficient in searching **subject** FASTAs
- BLAST+ provides **makeblastdb** to convert a FASTA file into an indexed, easily searchable binary file (1's and 0's).



-in <fasta file>
-out <database name>
-dbtype <type> (**p**rot or **n**ucl)
-title <title> (in HTML reports)
-parse_seqids (include FASTA sequence IDs)

(ID used in outputs and other BLAST+ tools)

Oregon State UNIVERSITY

BLAST databases options

- Once the BLAST database (db) is created we can supply **`blastn`** with additional parameters:

`-db` `<database name>`

Search against the db we just made instead of a FASTA file specified by `-subject`

`-num_threads` `<integer>`

The number of CPU cores to use (in parallel) during your "blasting," which will speed things up.

Oregon State UNIVERSITY

BLAST databases locations

- When using the **-db** option BLAST will look in three locations for databases:

1) **$PWD**
2) **$HOME**
3) **$BLASTDB**

BLAST databases list

You can list BLAST databases in the specified directory, e.g.,

```
blastdbcmd -list $BLASTDB/
```

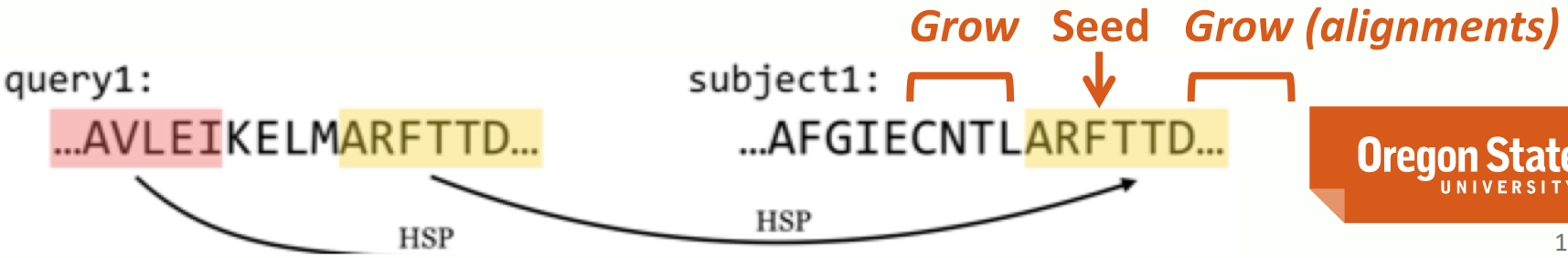
This will give you a set of "human readable" names for each database in that directory.

The blastdbcmd can also be used to extract information out of the databases, e.g., IDs reported in the BLAST output files.

Oregon State UNIVERSITY

"Basic" BLAST Algorithm

1) Index database for "words" of specific k-mer size
   e.g., 3 (bases) for protein, 11 for nucleotide
2) Search for High Scoring Pair (**HSP**) matches in db,
   aka "**seeding**" (finding "seeds" in your sequence)
3) Extend HSP "hits" with a **local alignment** to
   "***grow***" the **seeds** in size, increasing their **score**
4) Report matches above **HSP** and scores



Yeast exome database

Example: Look for proteins that are similar in sequence to other proteins in the Yeast exome.

Originally from yeastgenome.org

```
wget
http://library.open.oregonstate.edu/doc/computationalbiology/orf_trans.fasta
```

Create Yeast genome database (db) from FASTA

```
makeblastdb -in orf_trans.fasta -out orf_trans
-dbtype prot -title "Yeast Open Reading Frames"
-parse_seqids
```

Oregon State UNIVERSITY

`blastp` cutoff parameters

BLAST the Yeast exome against *itself*

`-max_target_seqs 2`

Since we are blasting the Yeast exome against itself each query will perfectly match, so we want > 1 to see what else may match!

`-max_hsps 1`

We want only the best **H**igh **S**coring **P**air per hit

`-evalue 1e-6`

Commonly-used cutoff (0.000001); read the literature to guide your research methods!

Oregon State UNIVERSITY

`blastp` output parameters

`-outfmt 7 'qseqid sseqid …'`

Tab-separated output with comment lines as:
1) query sequence ID, **2)** subject sequence ID, **3)** HSP alignment length, **4)** percent identity of the alignment, **5)** subject sequence length, **6)** query sequence length, **7)** start and **8)** end positions in the query and subject, and the **9)** E-value.

`-out yeast_blastp_yeast_top2.txt`

Send output to a file instead of **stdout** (terminal)

blastp Yeast exome against itself

```
blastp -query orf_trans.fasta
       -db orf_trans
       -max_target_seqs 2
       -max_hsps 1
       -evalue 1e-6
       -outfmt '7 qseqid sseqid
       length qlen slen qstart
       qend sstart send evalue'
-out yeast_blastp_yeast_top2.txt
-num_threads 4
```

Yeast exome BLAST output

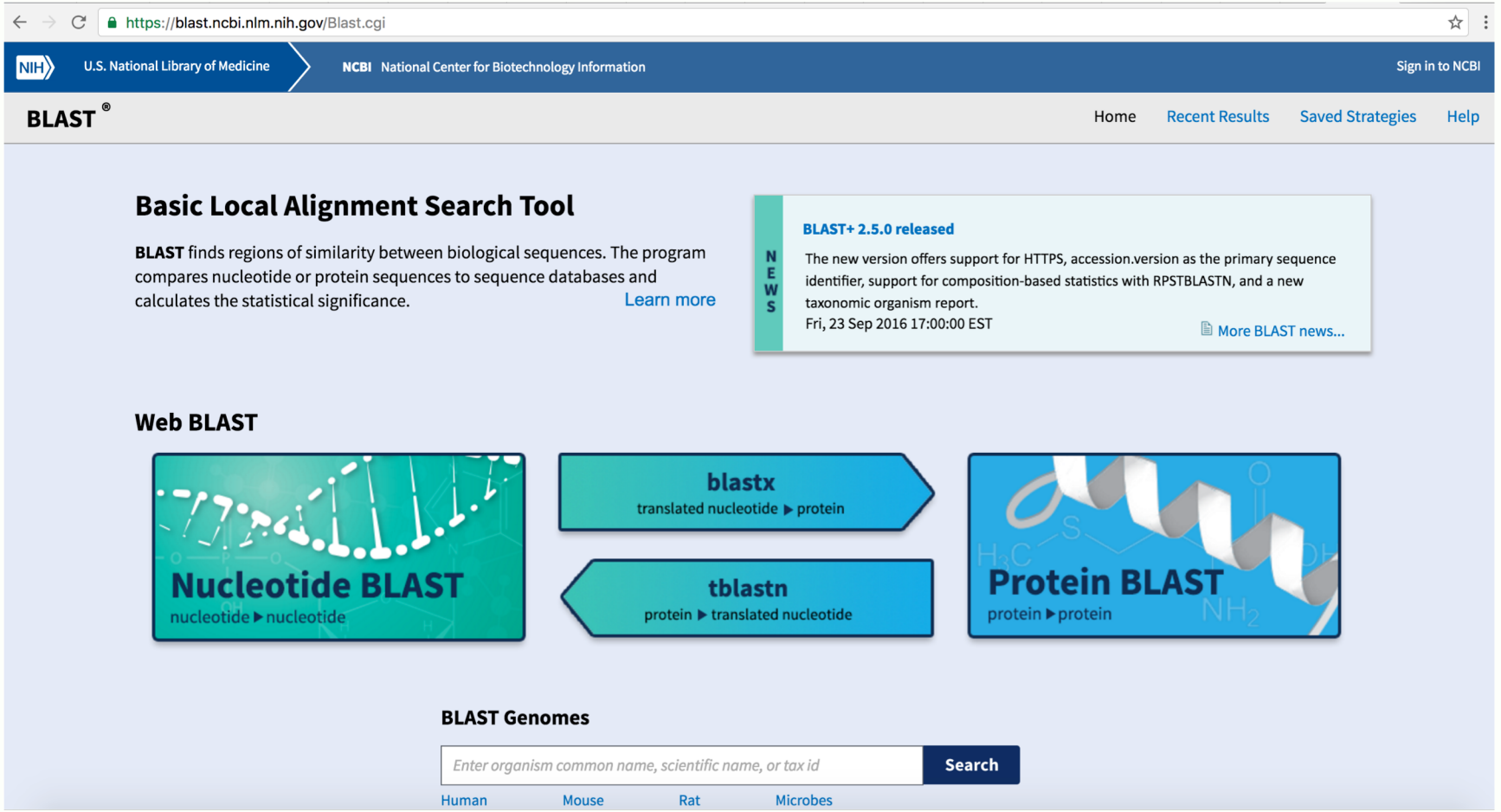
```
...
# BLASTP 2.2.30+
# Query: YAL003W EFB1 SGDID:S000000003, Chr I from 142174-142253,142620-143160,
# Database: orf_trans
# Fields: query id, subject id, alignment length, query length, subject length,
# 1 hits found
YAL003W YAL003W 207     207     207     1       207     1       207     2e-148
# BLASTP 2.2.30+
# Query: YAL005C SSA1 SGDID:S000000004, Chr I from 141431-139503, Genome Release
# Database: orf_trans
# Fields: query id, subject id, alignment length, query length, subject length,
# 2 hits found
YAL005C YAL005C 643     643     643     1       643     1       643     0.0
YAL005C YLL024C 643     643     640     1       643     1       640     0.0
...
```

`less -S` `yeast_blastp_yeast_top2.txt`

Expected self-match, e.g, **YAL005C**

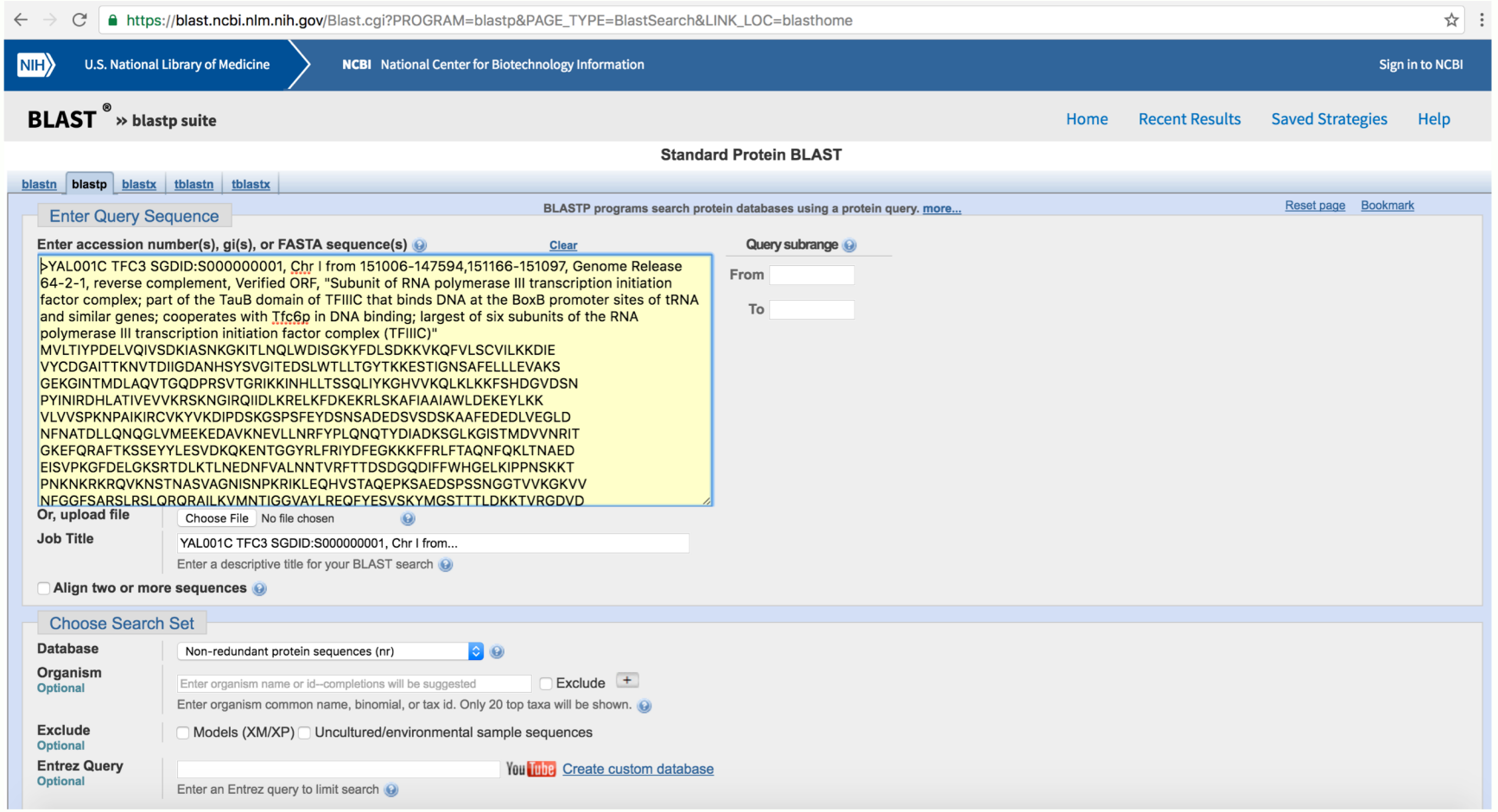
Oregon State UNIVERSITY

https://blast.ncbi.nlm.nih.gov/



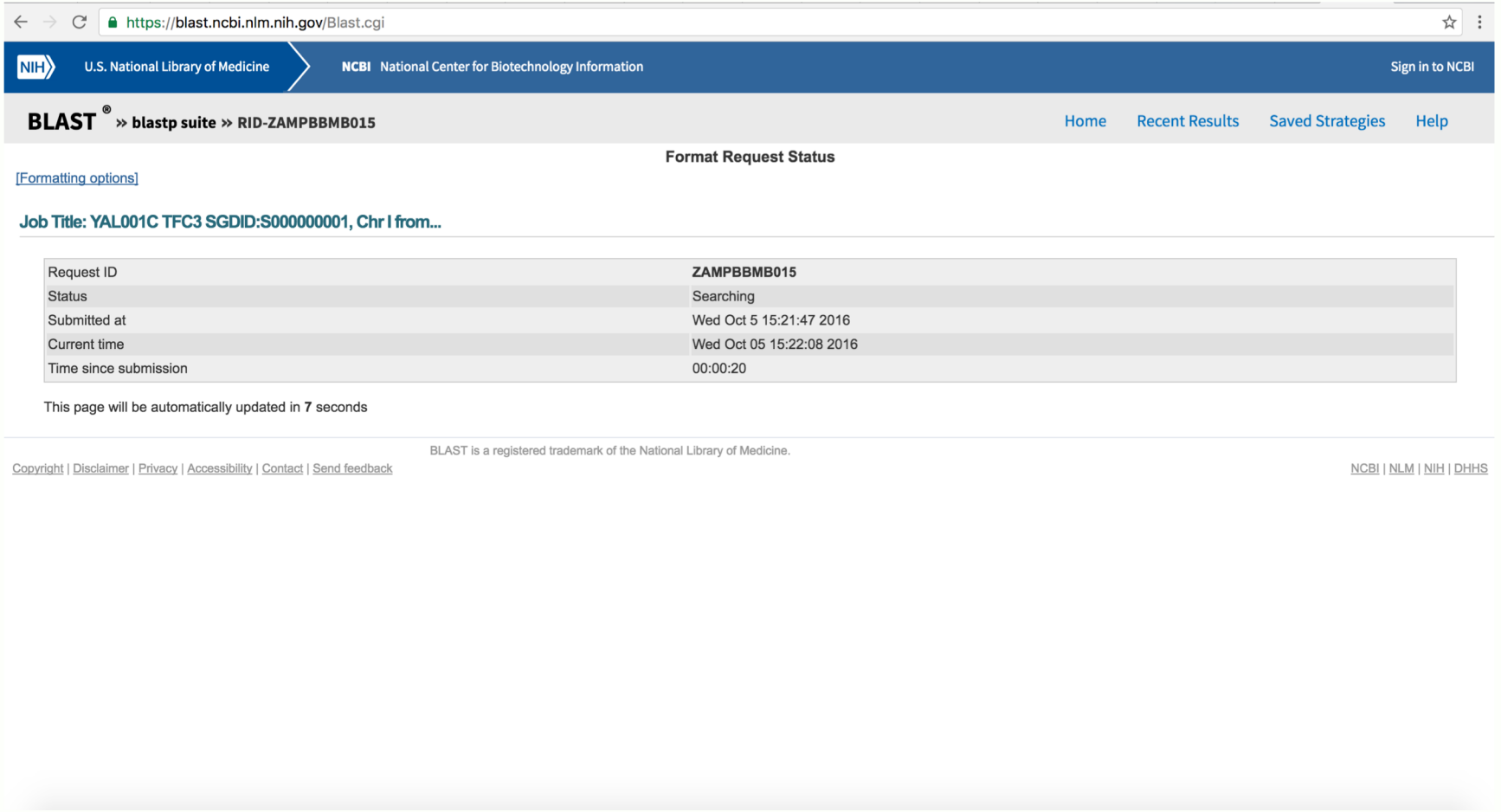
Oregon State UNIVERSITY

blastp suite Yeast exome sequence



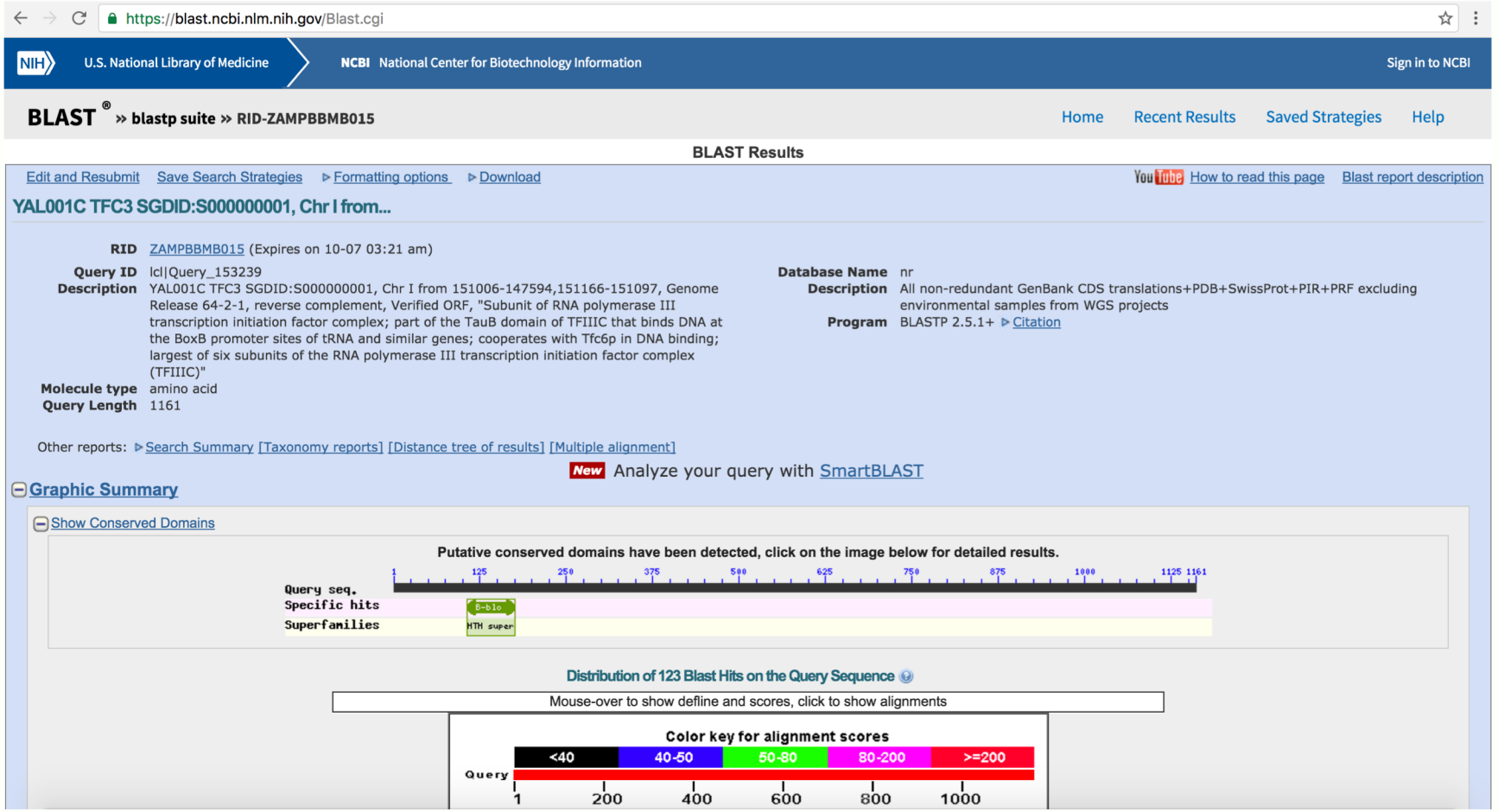
Oregon State UNIVERSITY

blastp processing..



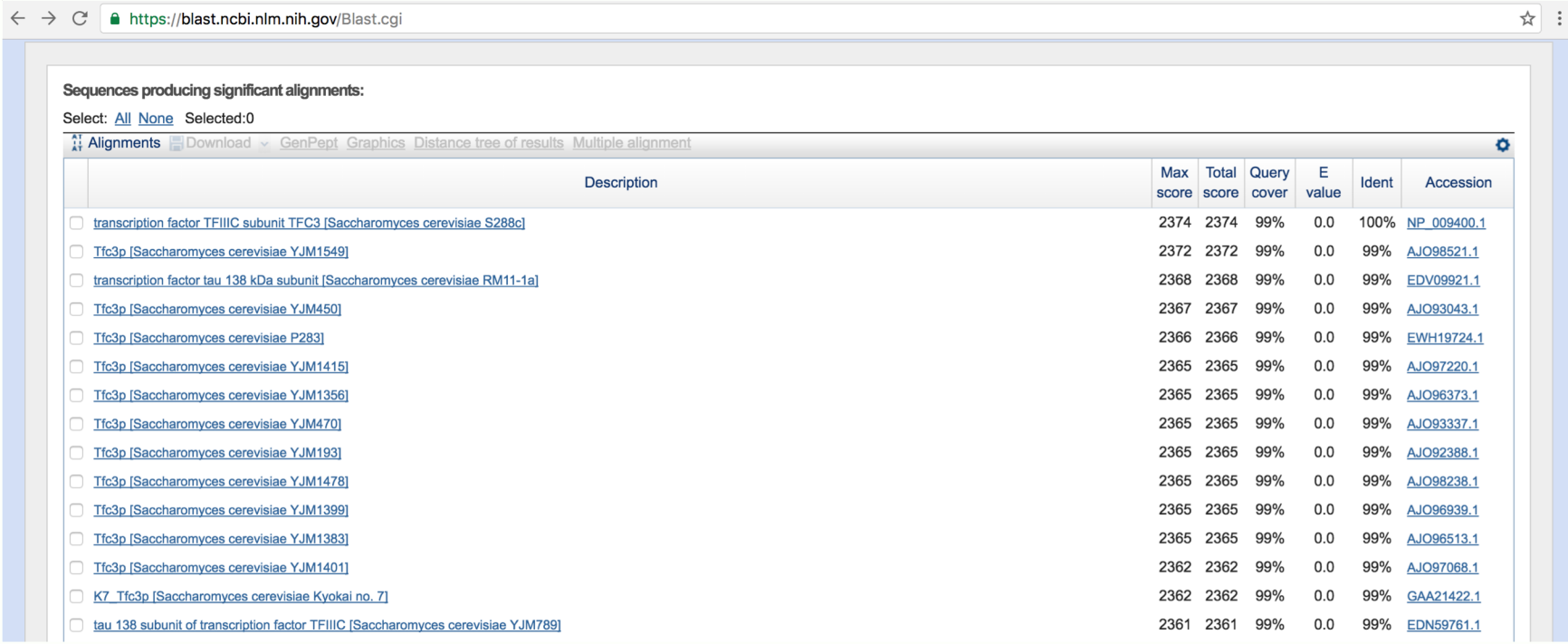
Oregon State UNIVERSITY

blastp results (top)



Oregon State UNIVERSITY

blastp results (bottom)



https://blast.ncbi.nlm.nih.gov/Blast.cgi

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
transcription factor TFIIIC subunit TFC3 [Saccharomyces cerevisiae S288c]	2374	2374	99%	0.0	100%	NP_009400.1
Tfc3p [Saccharomyces cerevisiae YJM1549]	2372	2372	99%	0.0	99%	AJO98521.1
transcription factor tau 138 kDa subunit [Saccharomyces cerevisiae RM11-1a]	2368	2368	99%	0.0	99%	EDV09921.1
Tfc3p [Saccharomyces cerevisiae YJM450]	2367	2367	99%	0.0	99%	AJO93043.1
Tfc3p [Saccharomyces cerevisiae P283]	2366	2366	99%	0.0	99%	EWH19724.1
Tfc3p [Saccharomyces cerevisiae YJM1415]	2365	2365	99%	0.0	99%	AJO97220.1
Tfc3p [Saccharomyces cerevisiae YJM1356]	2365	2365	99%	0.0	99%	AJO96373.1
Tfc3p [Saccharomyces cerevisiae YJM470]	2365	2365	99%	0.0	99%	AJO93337.1
Tfc3p [Saccharomyces cerevisiae YJM193]	2365	2365	99%	0.0	99%	AJO92388.1
Tfc3p [Saccharomyces cerevisiae YJM1478]	2365	2365	99%	0.0	99%	AJO98238.1
Tfc3p [Saccharomyces cerevisiae YJM1399]	2365	2365	99%	0.0	99%	AJO96939.1
Tfc3p [Saccharomyces cerevisiae YJM1383]	2365	2365	99%	0.0	99%	AJO96513.1
Tfc3p [Saccharomyces cerevisiae YJM1401]	2362	2362	99%	0.0	99%	AJO97068.1
K7_Tfc3p [Saccharomyces cerevisiae Kyokai no. 7]	2362	2362	99%	0.0	99%	GAA21422.1
tau 138 subunit of transcription factor TFIIIC [Saccharomyces cerevisiae YJM789]	2361	2361	99%	0.0	99%	EDN59761.1

Numerous hits to ***Saccharomyces cerevisiae***

Expected? Unusual?

Oregon State UNIVERSITY