Georgia Advanced Computing Resource Center
UNIVERSITY OF GEORGIA

How to Run NCBI BLAST on zcluster at GACRC

BLAST: Basic Local Alignment Search Tool

Georgia Advanced Computing Resource Center
University of Georgia
Suchitra Pakala
pakala@uga.edu

OVERVIEW

- What is BLAST? Why use it? How does it work?
- Different BLAST programs and Running BLAST
- Running on zcluster and other considerations

Part 1:

What is BLAST?

Why use it?

How does it work?

Introduction to BLAST

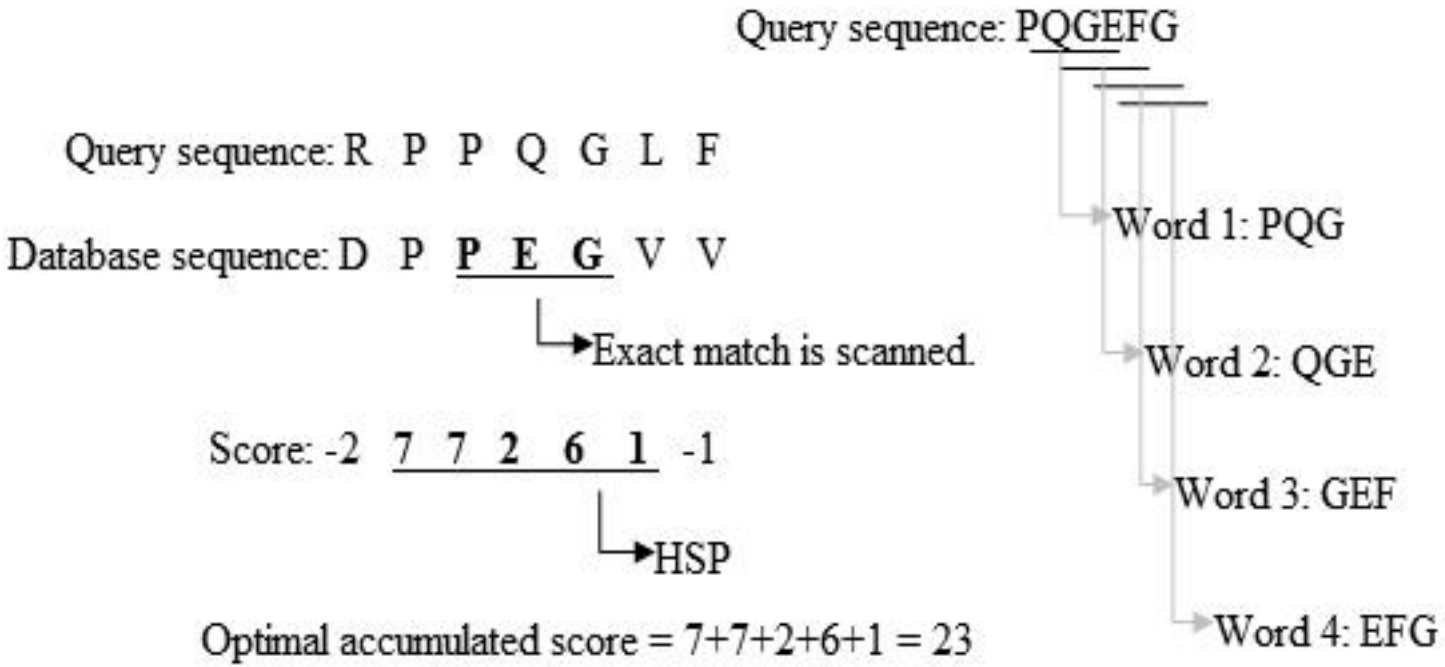
- **BLAST: B**asic **L**ocal **A**lignment **S**earch **T**ool
- Rapid sequence comparison of a "query" sequence against a large set of sequences – the "reference database"
- "Local" means it searches and aligns sequence segments, rather than aligning the entire sequence
- *Defacto* standard for search and alignment
- Fast, accurate, open source, available on several platforms

BLAST - Uses

- Identify a species
- Identify unknown sequences
- Identify homologous sequences in a database
- Find domains and hence determine protein function
- Phylogenetic analysis
- Placing sequences on chromosome
- Genome annotation
- Find variants of genes

BLAST – How does it work?

- Initial short matches – “words”
- Build local alignments in regions with common words – “seeding”
  - Seed extensions
- Scoring matrix
- Alignment score; thresholds



Part 2:

BLAST Programs

Reference Databases

Running BLAST

BLAST Programs

- blastn (nucleotide BLAST)
- blastp (protein BLAST)
- blastx (translated BLAST)
- tblastn (translated BLAST)
- tblastx (translated BLAST)

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, delta-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

blastn AGTCCGATGA AGCGCAAGTCAGTCCGATGAGCTAGCAGCT

blastp FLPQLLGNVL MFIPLIFLPQLLGNVLVCVLAHHFGKEFTP

blastx TGGCGAACCTTGGCTGCTGGATG MANLGCWMLVLFVATWSDLGLCKKRP

TGG-CGA-ACC-TTG-GCT-GCT-GGA-TG

T-GGC-GAA-CCT-TGG-CTG-CTG-GAT-G

TG-GCG-AAC-CTT-GGC-TGC-TGG-ATG

WRTLAAG

GEPWLLD

ANLGCWM

3 more frames from reverse
NOT shown here

Number of Searches Performed

Program	Input		Database
• blastn	DNA	1 →	DNA
• blastp	protein	1 →	protein
• blastx	DNA	6 →	protein
• tblastn	protein	6 →	DNA
• tblastx	DNA	36 →	DNA

DNA => 6-frame translation

BLAST Databases

- Publicly available collections of reference sequences and databases
  - nr/nt - non-redundant nucleotide collection
  - Refseq proteins
  - Species specific
    - Human
    - Mouse

- Custom databases
  - Build your own!
  - Example: Fungal ITS

Creating your own BLAST reference database

makeblastdb -in my_genomes.fasta -dbtype nucl -out my_genomes_db -title my_genomes_db

- Index files are created:

  my_genomes_db.nhr
  my_genomes_db.nin
  my_genomes_db.nog
  my_genomes_db.nsd
  my_genomes_db.nsi
  my_genomes_db.nsq

Run a BLAST Search – Where?

- NCBI BLAST server – web based searches
- Local infrastructure – zcluster
- Personal computer – yes, you can do it on your laptop
  - Integrated bioinformatics workbenches (CLC Genomics Workbench, etc)
  - Command line
    - http://www.ncbi.nlm.nih.gov/guide/howto/run-blast-local/
    - Download the executables that are readily available for your machine (windows / mac)
    - Installation instructions in manual
    - Obtain readily available reference databases AND/OR Build your own

Run a BLAST search – How?

- 4 “simple” steps:
  - Provide “Query” sequence
  - Select the BLAST program
  - Choose the database to search
  - Choose optional parameters
- “BLAST” away

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

NP_000509

Querysubrange

From

To

Or, upload file Browse...

Job Title NP_000509:beta globin [Homo sapiens]

Enter a descriptive title for your BLAST search

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism
Optional

Any Human *A.thaliana* Mouse Custom...

Search only sequences from selected organism

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST Search database **nr** using **Blastp (protein-protein BLAST)**

Show results in a new window

Algorithm parameters

STEP1: Choose the Sequence

- Sequence can be input in FASTA format or an accession number

Example of the FASTA format for a BLAST query

STEP 2: Choose the BLAST Program

Georgia Advanced Computing Resource Center
UNIVERSITY OF GEORGIA

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

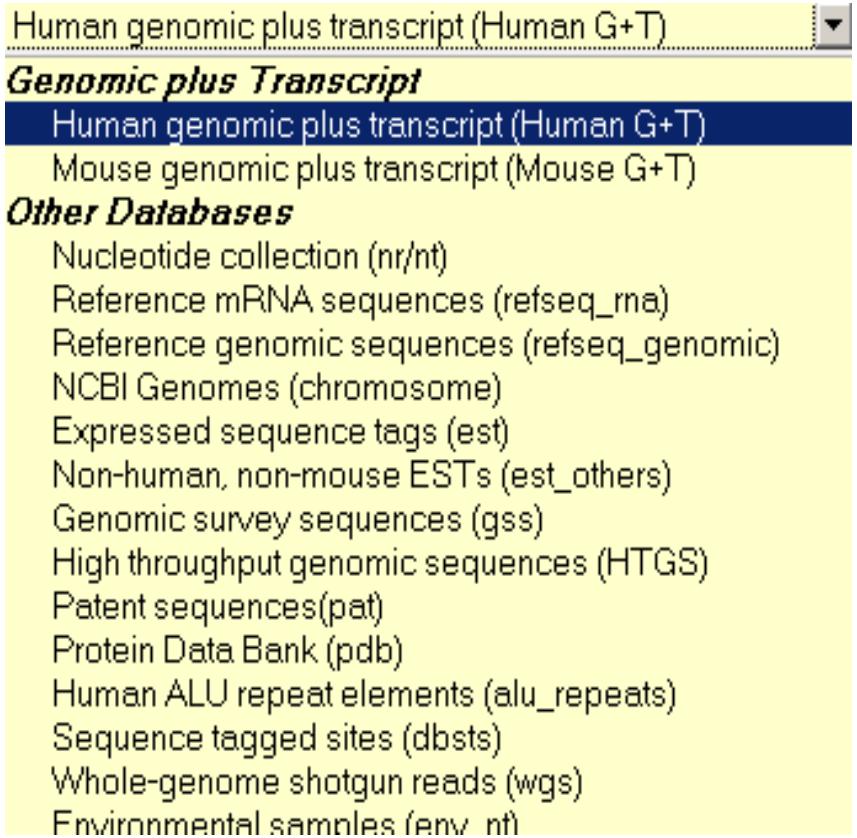
Specialized BLAST

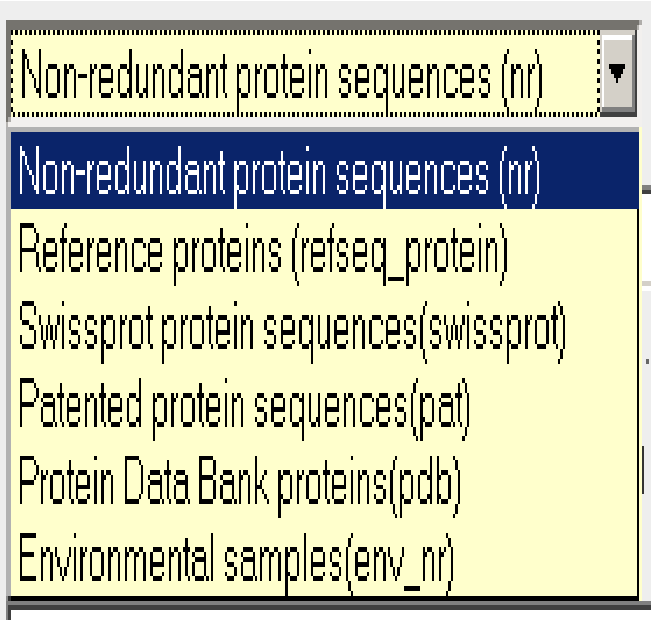
Choose a type of specialized search (or database name in parentheses.)

- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search for SNPs (snp)
- Screen sequence for vector contamination (vecscreen)
- Align two sequences using BLAST (bl2seq)

Step 3: Choose the database

nr = non-redundant (most general database)
dbest = database of expressed sequence tags
dbsts = database of sequence tag sites
gss = genomic survey sequences





Step 4: Choose Optional Parameters

You can….

- Choose the organism to search
- Turn filtering on/off
- Change the substitution matrix
- Change the expect (e) value
- Change the word size
- Change the output format

Step 4: Select optional search parameters

Organism

Algorithm

BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI BLAST / blastp suite

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

NP_000509.1

Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Non-redundant protein sequences (nr)

Organism Optional Enter organism name or id--completions will be suggested Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST

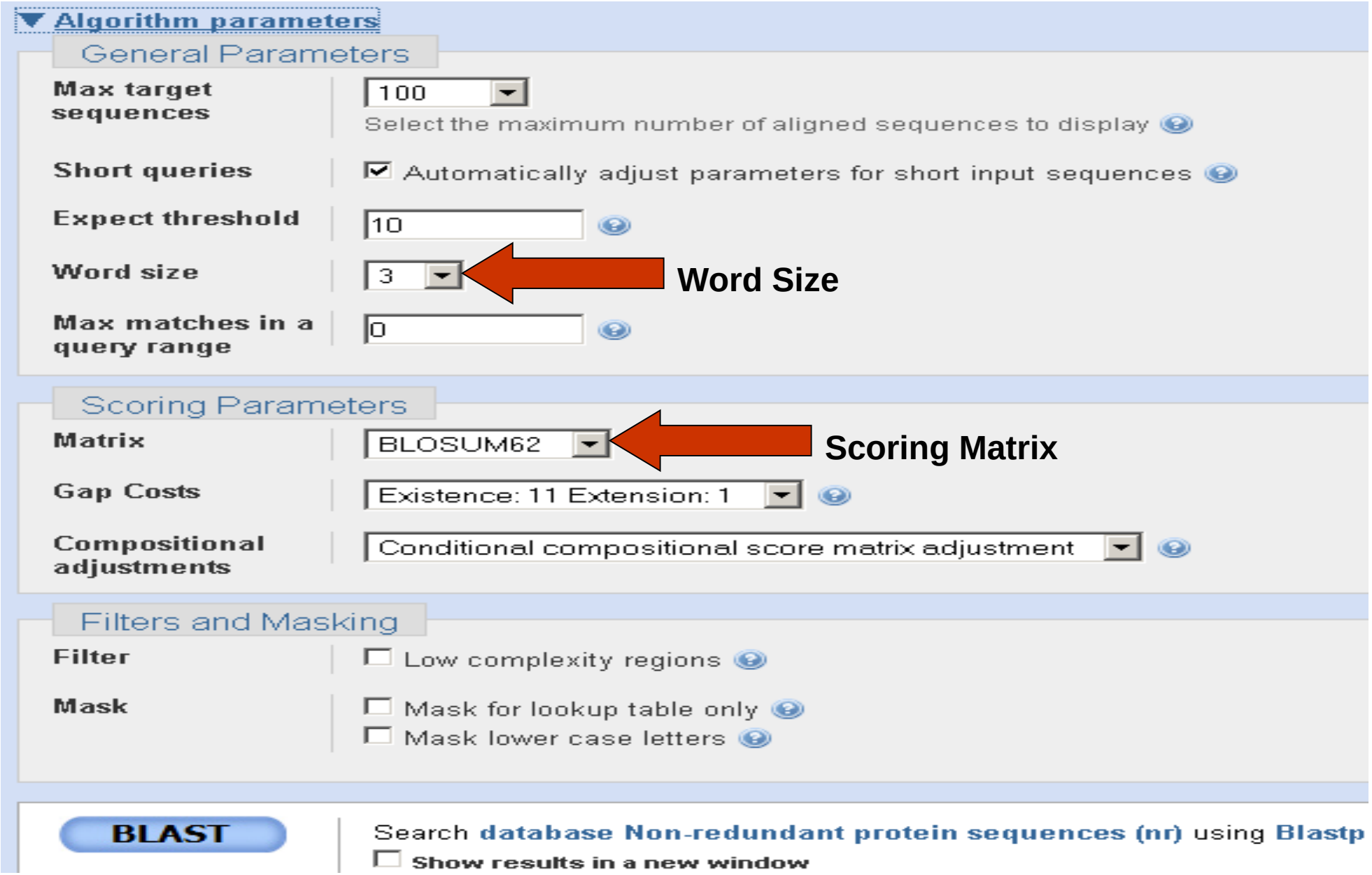
Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

Step 4: optional blastp search parameters

Georgia Advanced Computing Resource Center
UNIVERSITY OF GEORGIA



BLAST Output

Web interface – Summary of Results

Georgia Advanced Computing Resource Center
UNIVERSITY OF GEORGIA

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics

	Description	Max score	Total score	Query cover	E value	Ident	Accession
	unnamed protein product [Homo sapiens]	105	105	49%	2e-25	100%	BAG58727.1
	PREDICTED: major prion protein isoform X7 [Mandrillus leucophaeus]	105	105	49%	2e-25	100%	XP_011830622.1
	PREDICTED: major prion protein isoform X8 [Colobus angolensis palliatus]	105	105	49%	2e-25	100%	XP_011796437.1
	prion protein precursor [Microcebus murinus]	105	105	49%	3e-25	96%	NP_001296849.1
	PREDICTED: major prion protein isoform X5 [Mandrillus leucophaeus]	105	105	49%	3e-25	100%	XP_011830615.1
	PREDICTED: major prion protein isoform X6 [Colobus angolensis palliatus]	105	105	49%	3e-25	100%	XP_011796435.1
	PREDICTED: major prion protein isoform X4 [Colobus angolensis palliatus]	105	105	49%	4e-25	100%	XP_011796433.1
	PREDICTED: major prion protein isoform X3 [Mandrillus leucophaeus]	105	105	49%	4e-25	100%	XP_011830601.1
	prion protein [Galeopterus variegatus]	105	105	49%	5e-25	94%	AAN16488.1
	prion protein (p27-30) (Creutzfeld-Jakob disease, Gerstmann-Strausler-Scheinker syndror	105	105	49%	5e-25	100%	AAV38283.1
	prion protein [Homo sapiens]	105	105	49%	5e-25	100%	CAA58442.1

BLAST Output

Web interface – Alignments

Georgia Advanced Computing Resource Center
UNIVERSITY OF GEORGIA

Alignments

Download GenPept Graphics

unnamed protein product [Homo sapiens]

Sequence ID: dbj|BAG58727.1| **Length:** 229 **Number of Matches:** 1

Range 1: 1 to 49 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
105 bits(263)	2e-25	Compositional matrix adjust.	49/49(100%)	49/49(100%)	0/49(0%)	+1

```
Query  1    MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRY  147
            MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRY
Sbjct  1    MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRY  49
```

Download GenPept Graphics

PREDICTED: major prion protein isoform X7 [Mandrillus leucophaeus]

Sequence ID: ref|XP_011830622.1| **Length:** 229 **Number of Matches:** 1

Range 1: 1 to 49 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
105 bits(263)	2e-25	Compositional matrix adjust.	49/49(100%)	49/49(100%)	0/49(0%)	+1

```
Query  1    MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRY  147
            MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRY
Sbjct  1    MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRY  49
```

In Summary

- Key results
  - Score
  - Query Coverage
  - Expected value (evalue)
  - % Identities

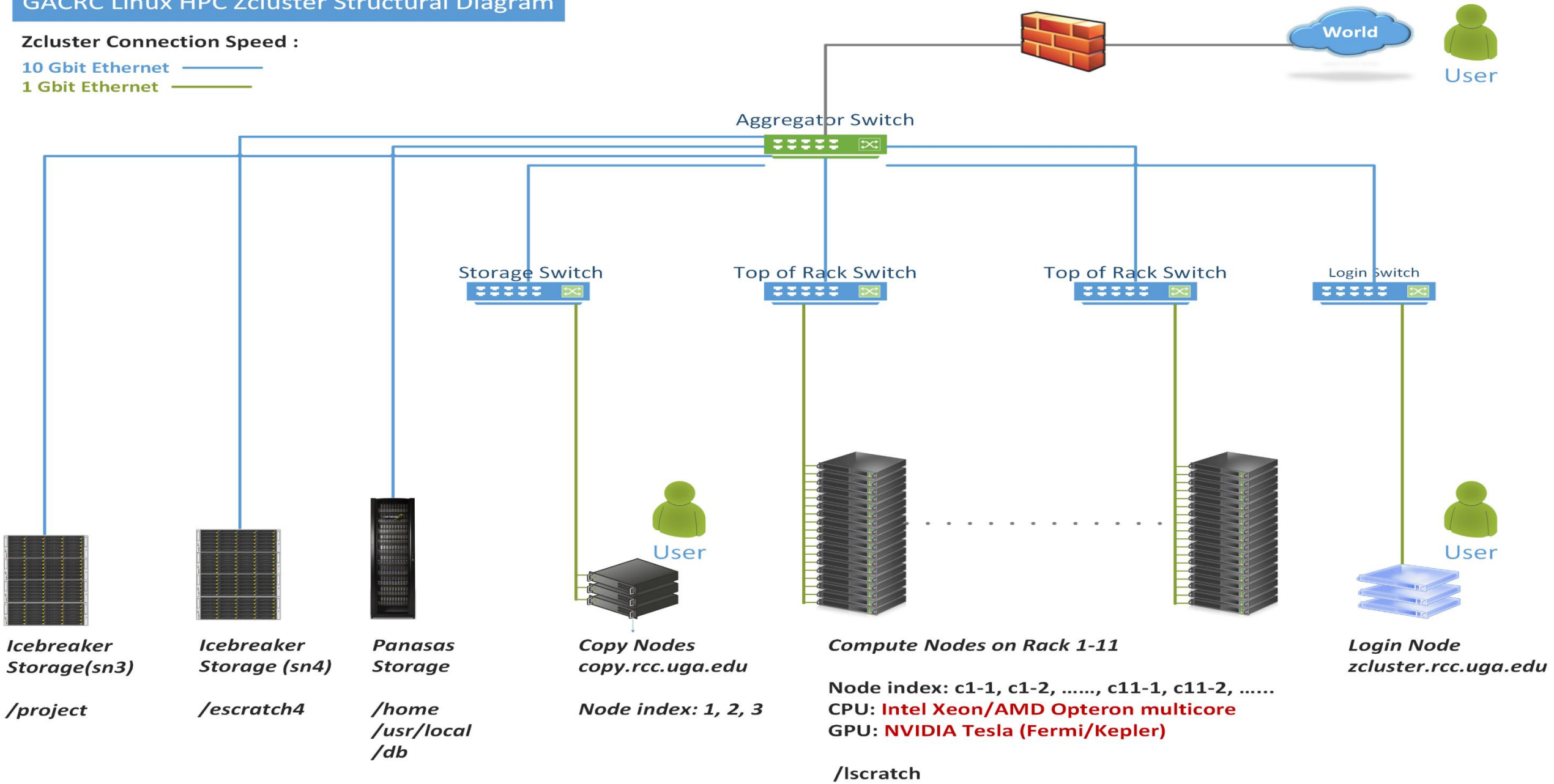
- HSPs – High Scoring Segment Pairs

Part 3:

Running on zcluster

And other considerations

GACRC Linux HPC Zcluster Structural Diagram



zcluster – General Information

- Batch-queueing System:
  - Jobs can be started (submitted), monitored, and controlled
  - Determine which compute node is the best place to run a job
  - Determine appropriate execution priority for a job to run
- On zcluster: Sun Grid Engine (SGE)

zcluster – Computing Resources

Queue Type	Queue Name	Nodes	Processor	Cores/Node	RAM(GB)/Node	Cores	NVIDIA GPU
Regular	rcc-30d	45	Intel Xeon	12	48	540	N/A
		150		8	16	1200	
High Memory	rcc-m128-30d	1	Intel Xeon	8	128	8	N/A
		4		8	192	32	
		10		12	256	120	
	rcc-m512-30d	2		32	512	64	
Multi Core	rcc-mc-30d	6	AMD Opteron	32	64	192	N/A
Interactive	interq	2	AMD Opteron	48	132	96	N/A
GPU	rcc-sgpu-30d	2	Intel Xeon	8	48	16	4 Tesla S1070 cards
	rcc-mgpu-30d	2		12	48	24	9 Tesla (Fermi) M2070 cards
	rcc-kgpu-30d	4		12	96	24	32 Tesla (Kepler) K20Xm cards

***Total peak performance:* 23 Tflops**

So, Where to run the BLAST job?

- Directly on login node of zcluster – **NO. Please don't.**
- After login to zcluster, Submit to the queue. **rcc-30d** is recommended

How does the command look like?

```
blastx -query my_input.fasta
       -db nt
       -out my_input_blastx_nr.out
       -num_threads 4
       -outfmt 6
       -evalue 1e-5
       -max_target_seqs=2
```

- Write a shell script [blast_AF293.sh]

```
#!/bin/bash
cd /home/gacrc-instruction/pakala/Blast/AF293
time /usr/local/ncbiblast+/latest/bin/blastn -num_threads 4 \
-query AFU293.fna -db /db/ncbiblast/latest/nt -out blast_afu.out \
-outfmt 6 -evalue 1e-5 -max_target_seqs=2
```

- Submit to queue

```
qsub -q rcc-30d -pe thread 4 ./blast_AF293.sh
```

- Monitor job

```
qstat, qacct, qsj, qdel
```

Check and Cancel Jobs @ zcluster

- To check the status of all queued and running jobs: qstat

qstat	→ shows your job in the pool
qstat –u "*"	→ shows all the jobs in the pool
qstat –j 12345	→ shows detailed information, e.g., maxvmem, about the job with JOBID 12345
qstat –g t	→ list all nodes used by your jobs

- To cancel a queued or running job: qdel

qdel –u pakala	→ deleted all your jobs
qdel 12345	→ deletes your job with JOBID 12345

- To list detailed information about a job: qsj, qacct

qsj 12345	→ shows information, e.g., maxvmem, about the RUNNING job with JOBID 12345
qacct –j 12345	→ shows information, e.g., maxvmem, about the ENDED job with JOBID 12345

The best part: The Results!!!

Georgia Advanced Computing Resource Center
UNIVERSITY OF GEORGIA

Qseqid	Sseqid	Pident	Length	Mismatch	Gapopen	Qstart	Qend	Sstart	Send	Evalue	Bitscore
ch334	gi|56|ref|NP_757.1|	17.42	597	364	26	33168	33694	741	1278	0.002	60.1
ch423	gi|60|ref|WP_500.1|	50.85	118	25	2	11856	11973	580	664	2.00E-04	60.1
ch425	gi|60|ref|WP_500.1|	55.05	109	28	3	11863	11971	809	896	0.01	54.7

Qseqid	Query Seq-id
Sseqid	Subject Seq-id
Pident	Percentage of identical matches
Length	Alignment length
Mismatch	Number of mismatches
Gapopen	Number of gap openings
Qstart	Start of alignment in query
Qend	End of alignment in query
Sstart	Start of alignment in subject
Send	End of alignment in subject
Evalue	Expect value
Bitscore	Bit score

Alignment view options	
-outfmt <String>	
0	pairwise
1	query-anchored showing identities
2	query-anchored no identities
3	flat query-anchored, show identities
4	flat query-anchored, no identities
5	XML Blast output
6	tabular
7	tabular with comment lines
8	Text ASN.1
9	Binary ASN.1
10	Comma-separated values
11	BLAST archive format (ASN.1)

Other parameters

- **There are a LOT of parameters that can be adjusted**
- **Beyond the scope of this introduction**
- **Example:**

```
*** Restrict search or results
-gilist <String>
  Restrict search of database to list of GI's
   * Incompatible with:  negative_gilist, seqidlist, remote, subject,
  subject_loc
-seqidlist <String>
  Restrict search of database to list of SeqId's
   * Incompatible with:  gilist, negative_gilist, remote, subject,
  subject_loc
-negative_gilist <String>
  Restrict search of database to everything except the listed GIs
   * Incompatible with:  gilist, seqidlist, remote, subject, subject_loc
-max_target_seqs <Integer, >=1>
  Maximum number of aligned sequences to keep
  Not applicable for outfmt <= 4
```

What about "LARGE" jobs?

RCCBatchBlastPlus

- rccbatchblastplus
  - Splits big query file into smaller units
  - Submits each unit to queue to BLAST

qlogin
cd working_directory
rccbatchblastplus blastn –query inputfile –out outputfile -db targetdatabase –size size-of-split-unit –queue queue-name

- rccbatchblastplus-check
  - Run after rccbatchblastplus finishes
  - Checks the results and merges them into one final output file

qlogin
rccbatchblastplus-check

- https://wiki.gacrc.uga.edu/wiki/RCCBatchBlastPlus

Another option: Mpiblast

- Open-source, parallel implementation of NCBI BLAST
- Takes advantage of distributed computational resources, i.e., a cluster, through explicit MPI communication
- Utilizes all available resources at cluster

mpiformatdb
mpirun
mpiblast

- https://wiki.gacrc.uga.edu/wiki/Mpiblast

Quick links

- NCBI Blast+
https://wiki.gacrc.uga.edu/wiki/NCBI_Blast_%2B

- Legacy Blast
https://wiki.gacrc.uga.edu/wiki/NCBI_Blast

- Batch Blast
https://wiki.gacrc.uga.edu/wiki/RCCBatchBlastPlus

- MPI Blast
https://wiki.gacrc.uga.edu/wiki/Mpiblast

- Common Blast databases
https://wiki.gacrc.uga.edu/wiki/BLAST_Databases

THANK YOU