Hands-on 3: Local and Distributed Execution

Daniel Crawl

Shweta Purawat

Goals

- Create an actor to run BLAST in different environments
  - Local execution (blastall)
  - (Part 2) Distributed execution (mpiblast) on a cluster
- Configure ExecutionChoice to run BLAST
  - Inputs, Outputs, Parameters
  - Local and MPI choices

Running BLAST Locally

- Example command line:

```
blastall -i query -d ref -p blastn -m 8 -e 1E-5 > align
```

Running BLAST Locally

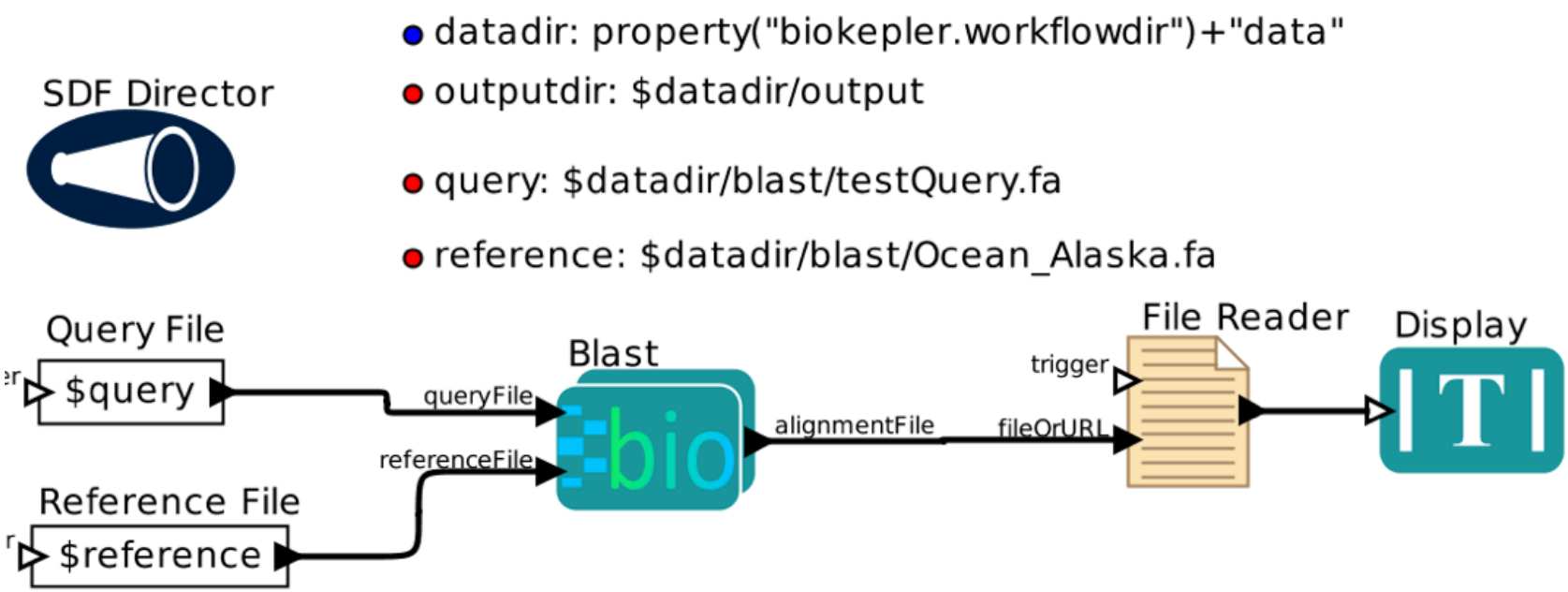
- Example command line:

```
blastall -i query -d ref -p blastn -m 8 -e 1E-5 > align
```

Inputs Parameters Output

Step 1: Open Blast workflow

- Workflow is /home/biokepler/Blast.kar



Step 2: Set Parameter Values

program: blastall

Input File Parameters

queryFile (-i):

referenceFile (-d):

Output File Parameters

checkOutputTimestamp: ☑

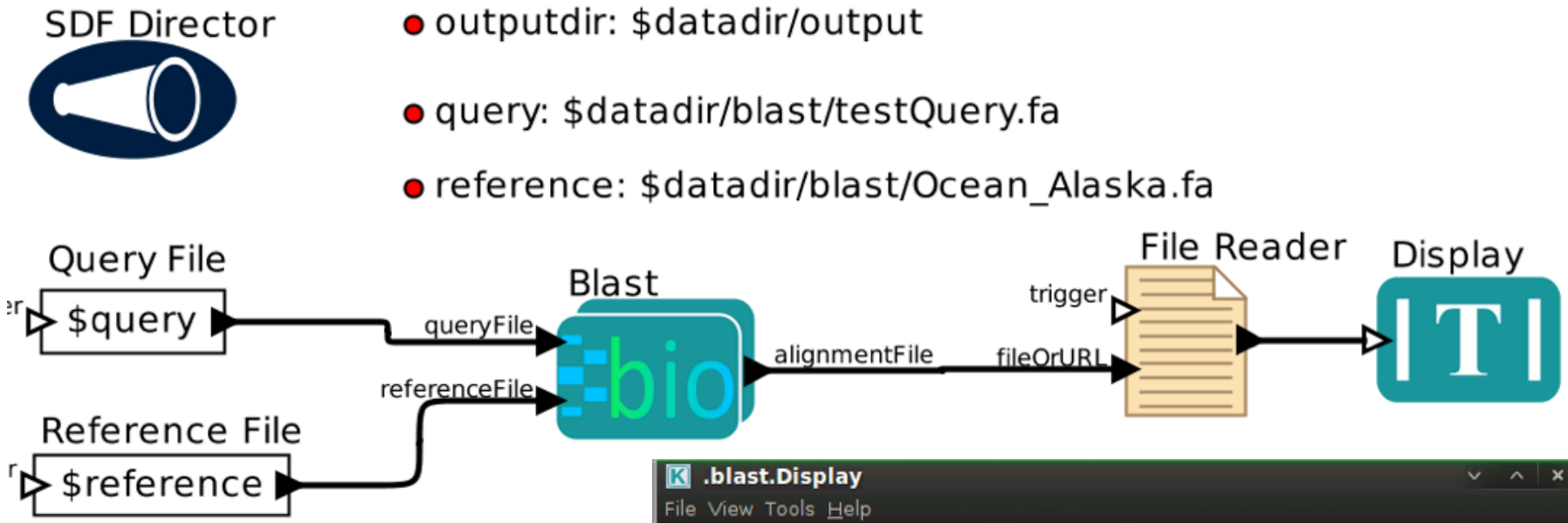
alignmentFile (>): $HOME/alignment.txt

Parameters

additionalOptions: -e 1E-5 -p bastn -m 8

Step 3: Run Finished Workflow

- datadir: property("biokepler.workflowdir")+"data"
- outputdir: $datadir/output
- query: $datadir/blast/testQuery.fa
- reference: $datadir/blast/Ocean_Alaska.fa



.blast.Display

File View Tools Help

56.5102262	17551515	83.12	77	13	0
56.5105189	17540078	100.00	39	0	0
56.5105189	17522675	100.00	39	0	0
56.5105189	17549965	100.00	38	0	0
56.5105189	17531612	100.00	38	0	0
56.5105189	17529086	100.00	38	0	0
56.5105189	17523468	100.00	38	0	0
56.5105189	17525933	94.87	39	2	0
56.5105189	17523067	94.12	34	2	0
56.5105189	17538416	90.24	41	4	0
56.5105189	17535335	93.94	33	2	0
56.5114036	17539194	91.67	36	3	0
62.6454064	17551781	100.00	26	0	0
69.7151151	17532466	96.88	32	0	1
69.7151151	17549965	93.55	31	2	0
69.7151151	17531612	93.55	31	2	0
69.7153657	17529886	96.55	29	1	0

mpiBLAST

- mpiBLAST is open source MPI based implementation of database segmentation for parallel BLAST search
- Super-linear performance gain with database segmenting technique
- Ideal database fragment is largest fragment that can sit in the memory
- Making fragments smaller than available memory adds to the overhead

mpiBLAST Algorithm

mpiBLAST algorithm consists of three steps:

- Segmenting and Distributing the database
- Running mpiBLAST queries on each node
- Merging the results from each node in a single output file

mpiBlast commands

- **Formatting and Segmenting the database**

  mpiformatdb –n **NP** -i ref.fa

- **Querying the database and merging the results**

  mpirun –n **NP** mpiblast -p blastp –d ref.fa -i query.fas –o blast_results.txt

Goal: To run mpiBlast on HPC Cluster remotely through Kepler workflow: SDSC Gordon Supercomputer

mpirun –n **NP** mpiblast -p blastp -i query -d ref -m 8 -e 1E-5 > align

Inputs

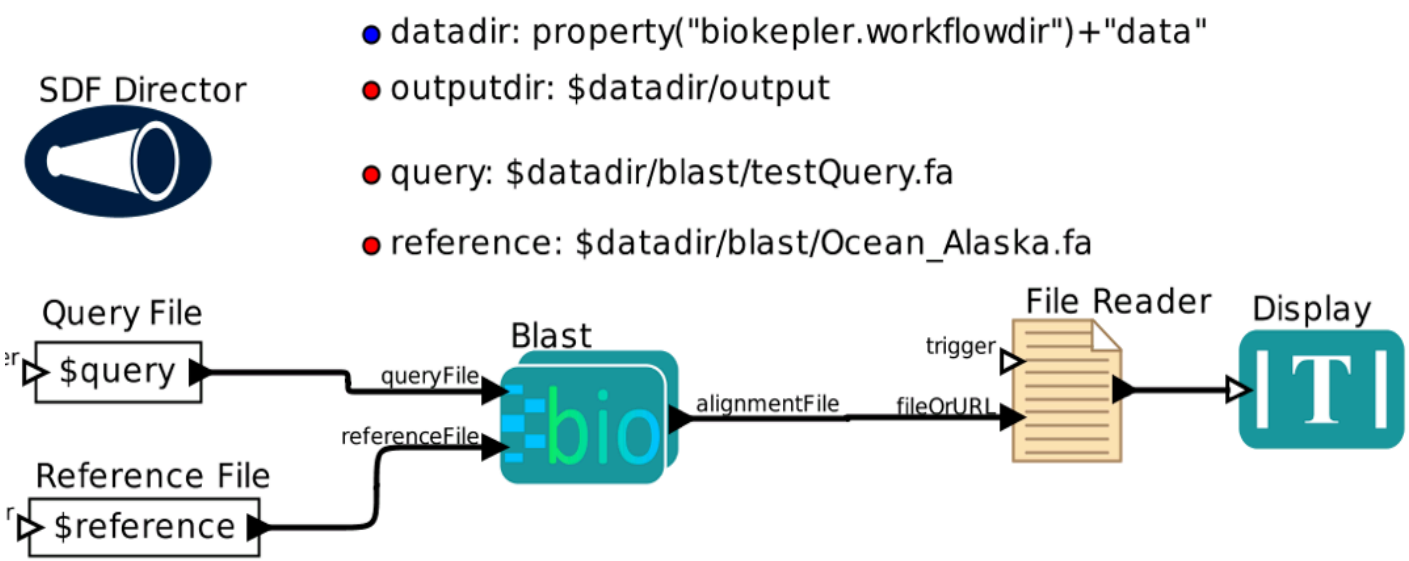
Parameters

Output

NP – multiple parallel processes

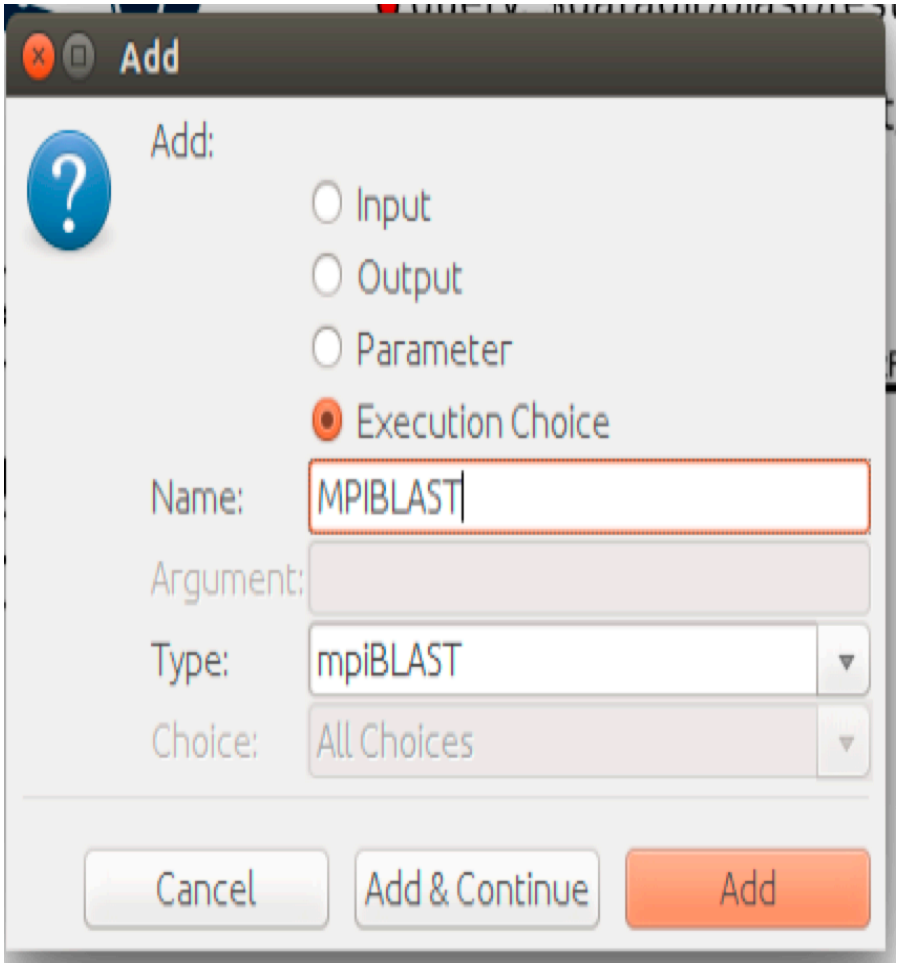
Step 1: Open blast workflow

– Go to Step 2, if Blast.kar is already open
– Else the workflow: /home/biokepler/Blast.kar



Step 2: Add new execution choice for MPIBLAST

- Double-click on the Blast actor and click on Add.
- Change radio button to Execution Choice.
- Select Type: mpiBLAST.
- Set Name to MPIBLAST.
- Click Add to close the dialog



Step 3:Configure the New Execution Choice

- Double-click on Blast actor
- Change to MPIBLAST tab
- Configure MPIBLAST tab:

Please use Gordon username assigned to you in place of **LOGIN**.

First parameter: TargetHost

- Edit "TargetHost" parameter and set its value to be LOGIN@gordon.sdsc.edu

Second parameter: cmdFile

- If already set move next parameter
- Else set its value to be $HOME/mpiBLAST/blastp_gordon.sh

Third parameter: commandLine

- Keep as it is. Move to next parameter.

Fourth parameter: outputFile

- If already set move next parameter
- Else set its value to be $alignmentFile

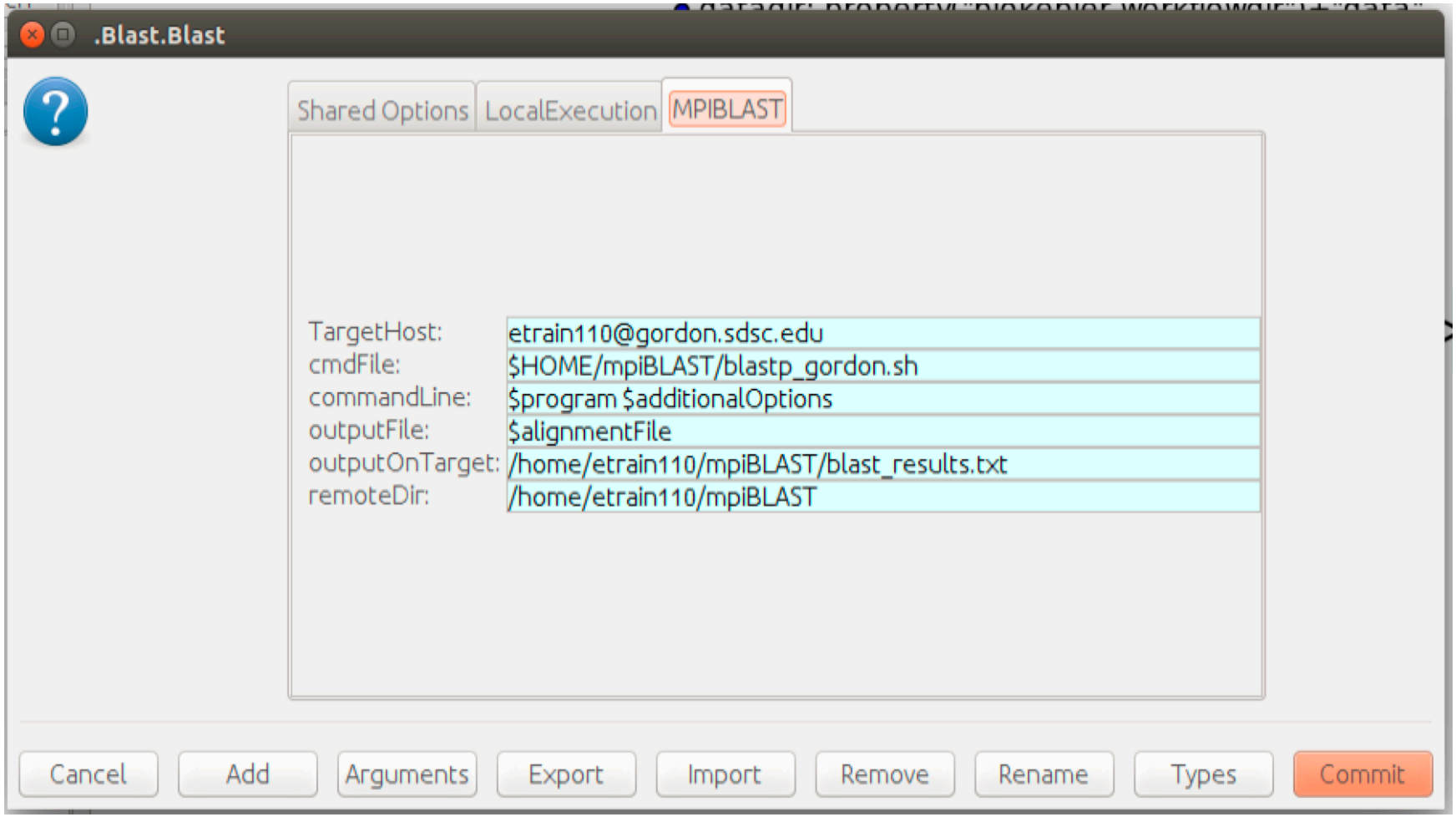
Fifth parameter: outputOnTarget

- Set its value to be /home/LOGIN/mpiBLAST/blast_results.txt

Sixth parameter: remoteDir

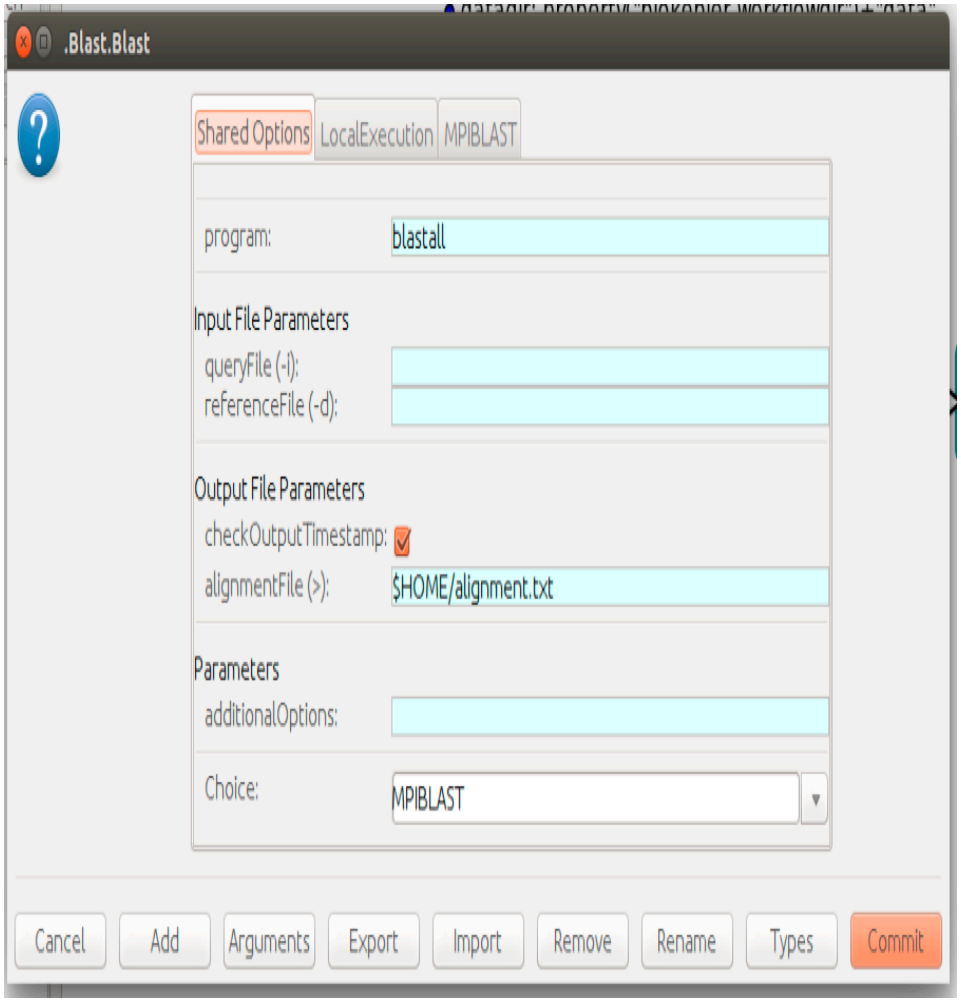
- Set its value to be /home/LOGIN/mpiBLAST

Step 3:Configure the New Execution Choice



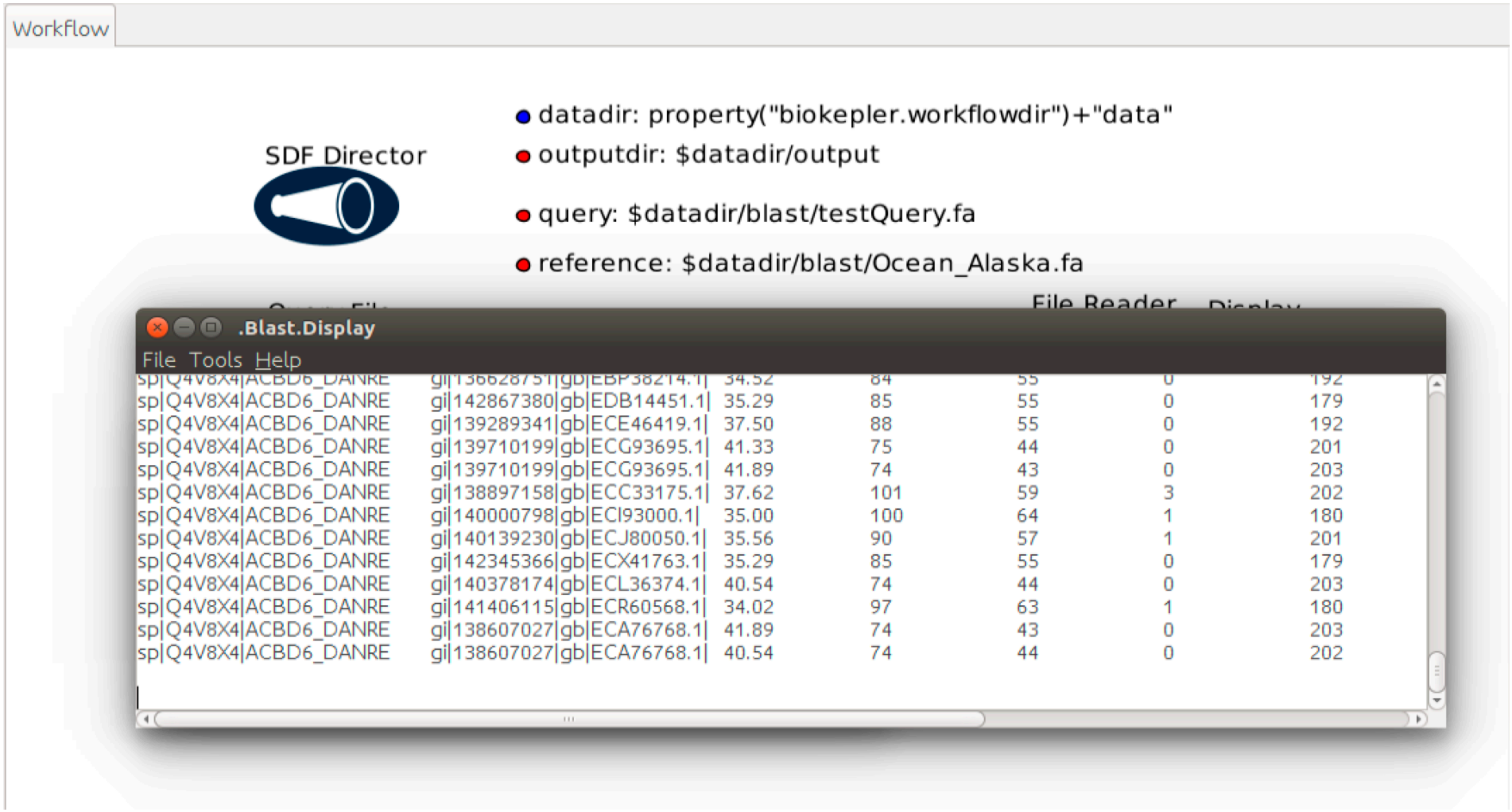
Step 4: Set the Execution Choice to MPIBLAST

- Double-click on the Blast actor and click on Add.
- Go to “Shared Options” tab
- Set Choice to MPIBLAST.
- Click on commit to save the changes and close the configuration box.



Step 5: Run Finished Workflow

- Save and Run the workflow.



mpiBLAST Performance

#Nodes	Run Time (sec)	Speed-Up
1	80775	1.00
4	8752	9.23
8	4548	17.76
16	2437	33.15
32	1350	59.83
64	851	94.92
128	474	170.41

Parallel Computing: Software Technology, Algorithms, Architectures & Applications
Gerhard Joubert, Wolfgang Nagel, Frans Peters, Wolfgang Walter
Elsevier, Sep 23, 2004