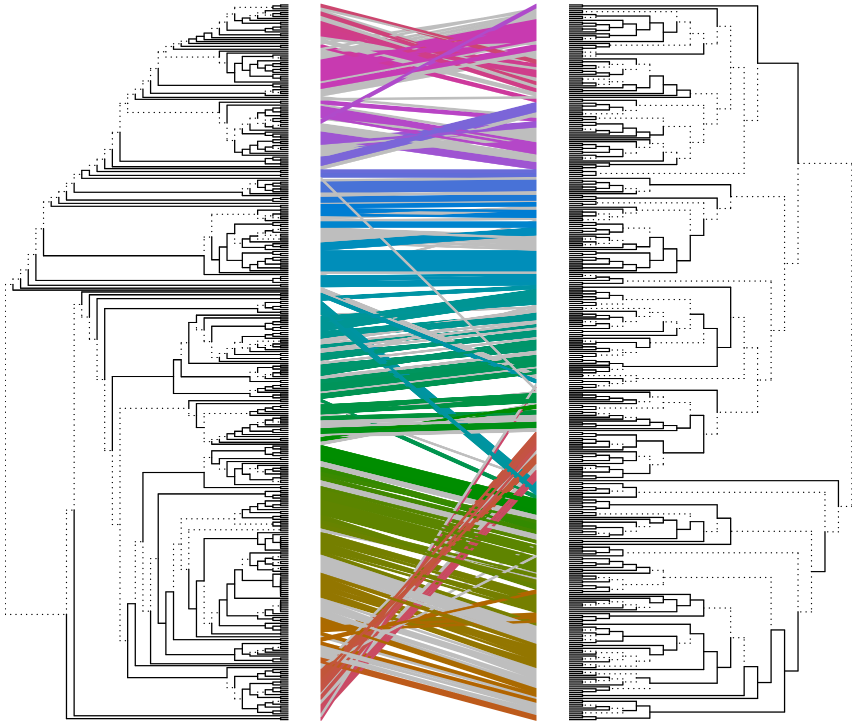
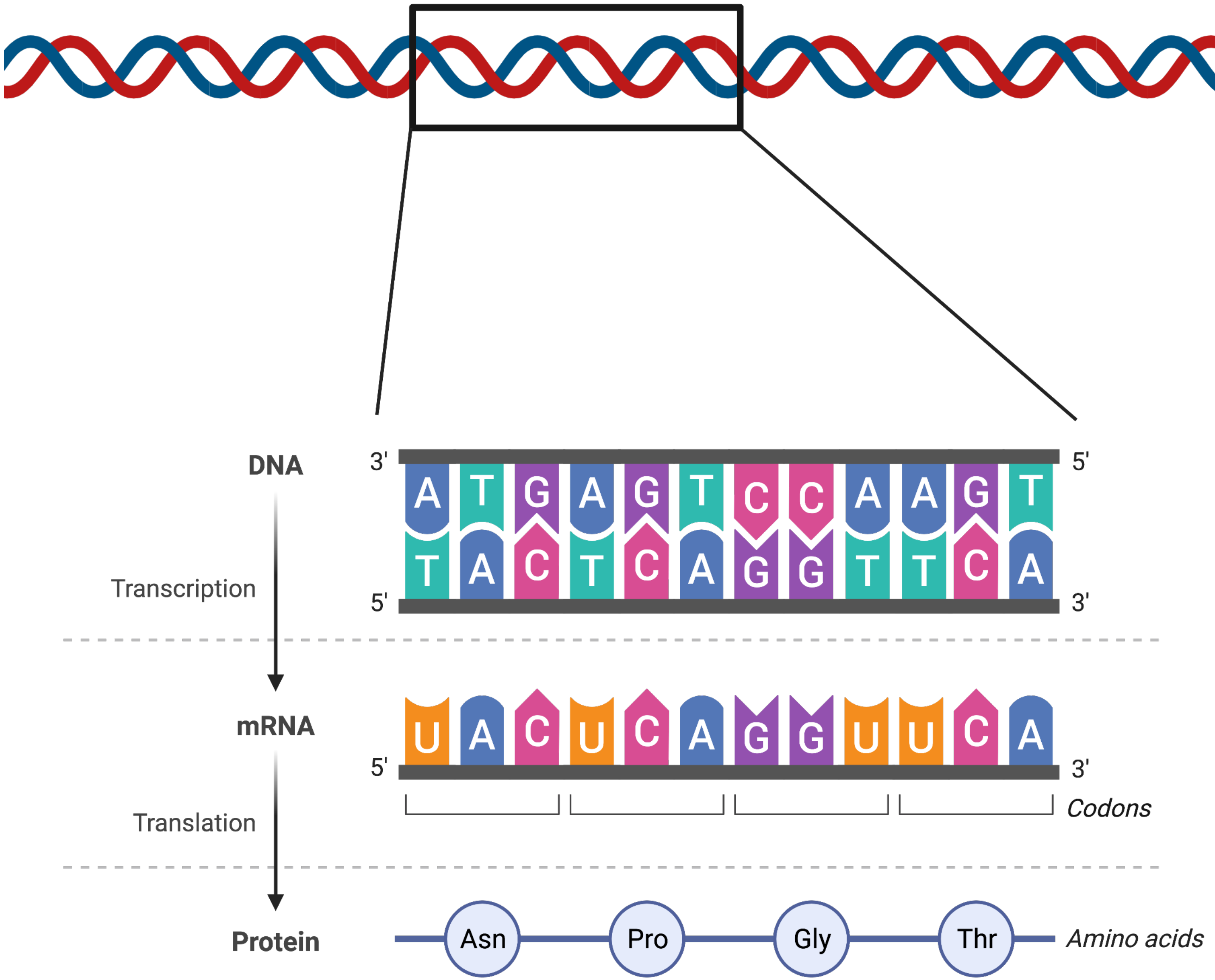
Computational Biology on the Open Science Grid



Nicholas Cooley
Wright Lab
University of Pittsburgh
Department of Biomedical Informatics

It is relatively straightforward to predict where genes are in genomes, even newly sequenced, novel isolates.

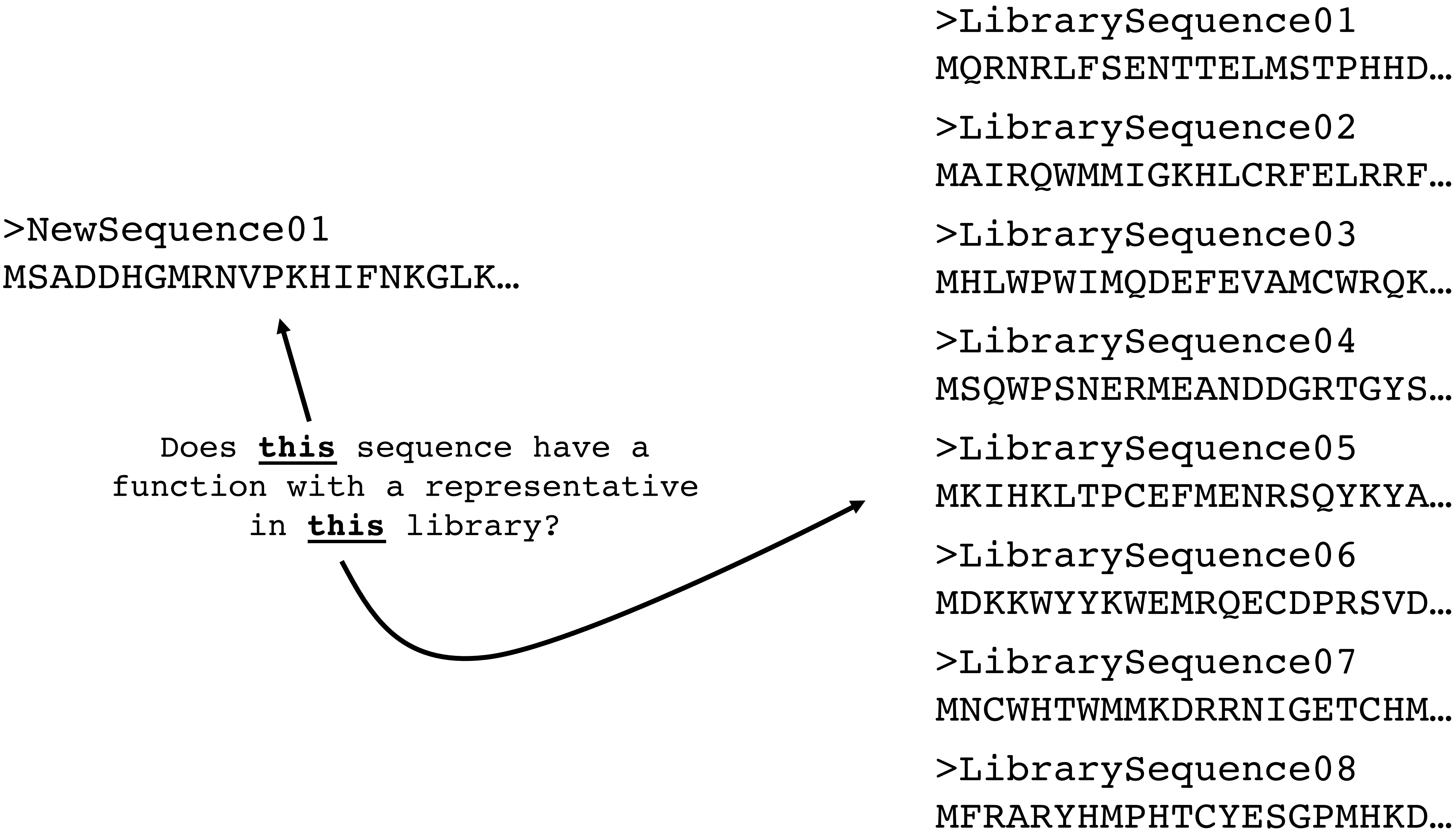
Figuring out what new genes do however, is not trivial.



Created with BioRender.com

Classification / Annotation:
Does this sequence have the same function or job as a sequence in some training data whose function or job is known?

Give a novel sequence a descriptive and succinct label that represents that sequence’s function.



```
>NewSequence01
MSADDHGMRNVPKHIFNKGLK…
```

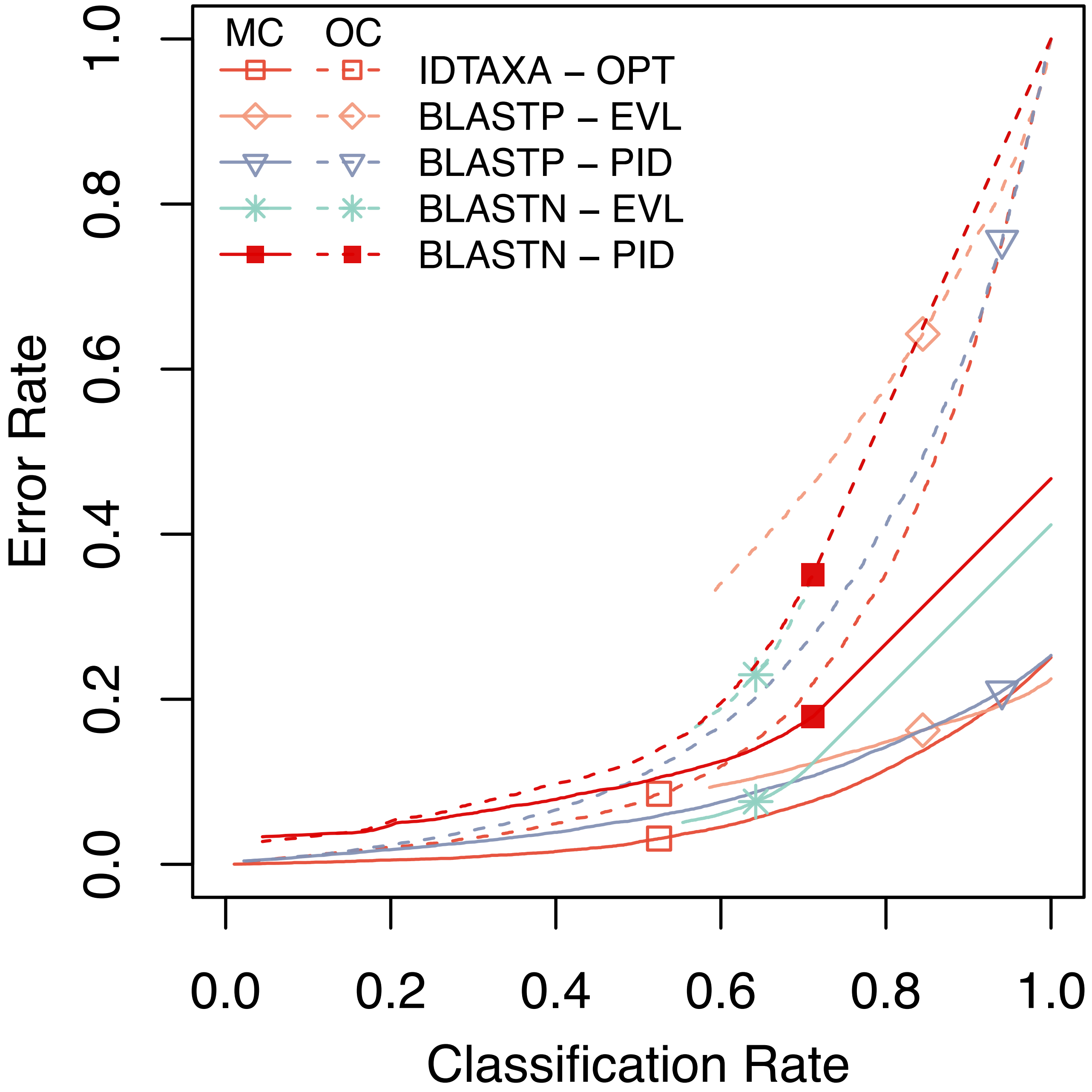
Does **this** sequence have a function with a representative in **this** library?

```
>LibrarySequence01
MQRNRLFSENTTELMSTPHHD…
>LibrarySequence02
MAIRQWMMIGKHLCRFELRRF…
>LibrarySequence03
MHLWPWIMQDEFEVAMCWRQK…
>LibrarySequence04
MSQWPSNERMEANDDGRTGYS…
>LibrarySequence05
MKIHKLTPCEFMENRSQYKYA…
>LibrarySequence06
MDKKWYYKWEMRQECDPRSVD…
>LibrarySequence07
MNCWHTWMMKDRRNIGETCHM…
>LibrarySequence08
MFRARYHMPHTCYESGPMHKD…
```

TL;DR we built a classifier:

- Accurate functional classification is difficult
- Emphasis on conservative classification to avoid overclassification of truly novel sequences

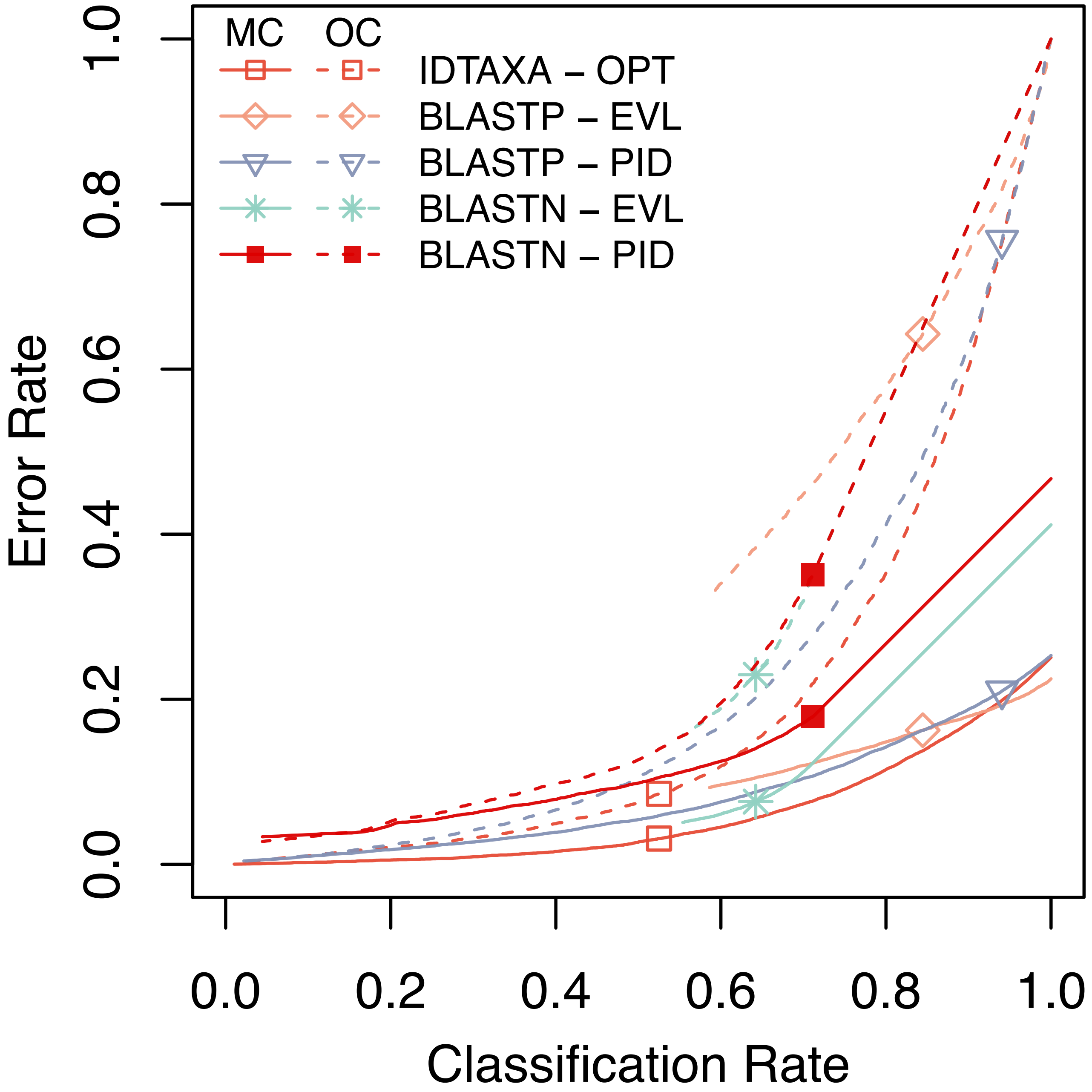
Publication coming soon!



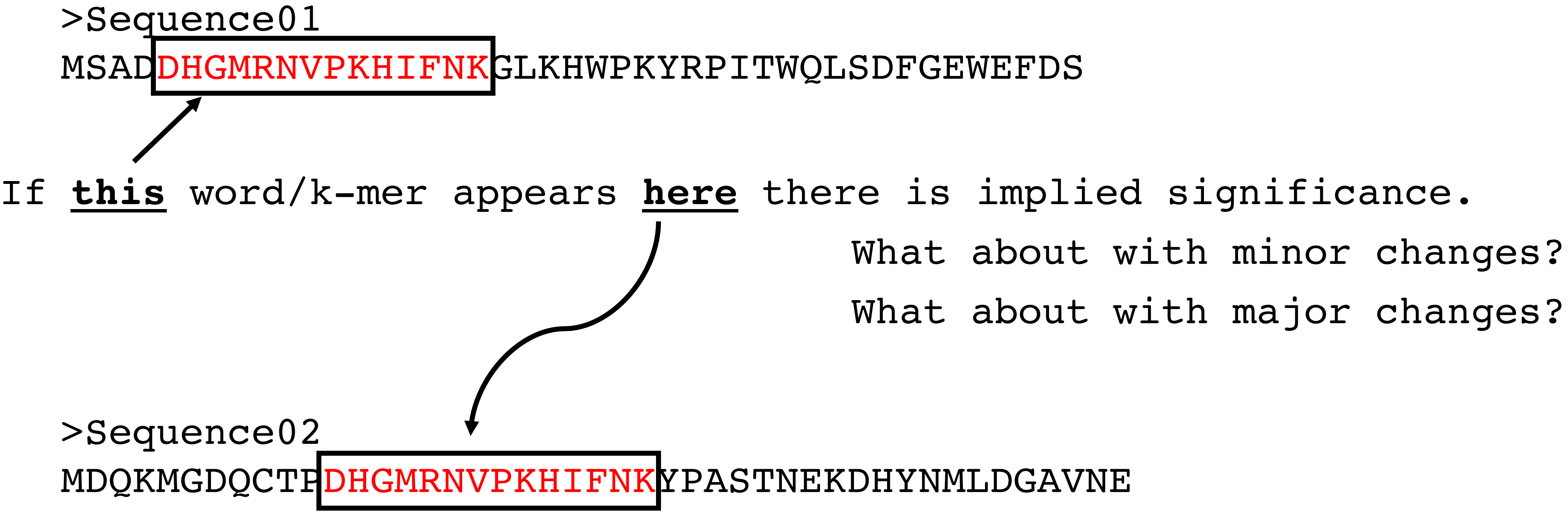
This work would not have been possible without the open science grid:

- Data preparation
- Parameter tuning
  - k-mer characteristics
- Cross validation
- Testing, testing, testing …

Publication coming soon!



One last bit of biology to introduce:



>Sequence01
MSADDHGMRNVPKHIFNKGLKHWPKYRPITWQLSDFGEWEFDS

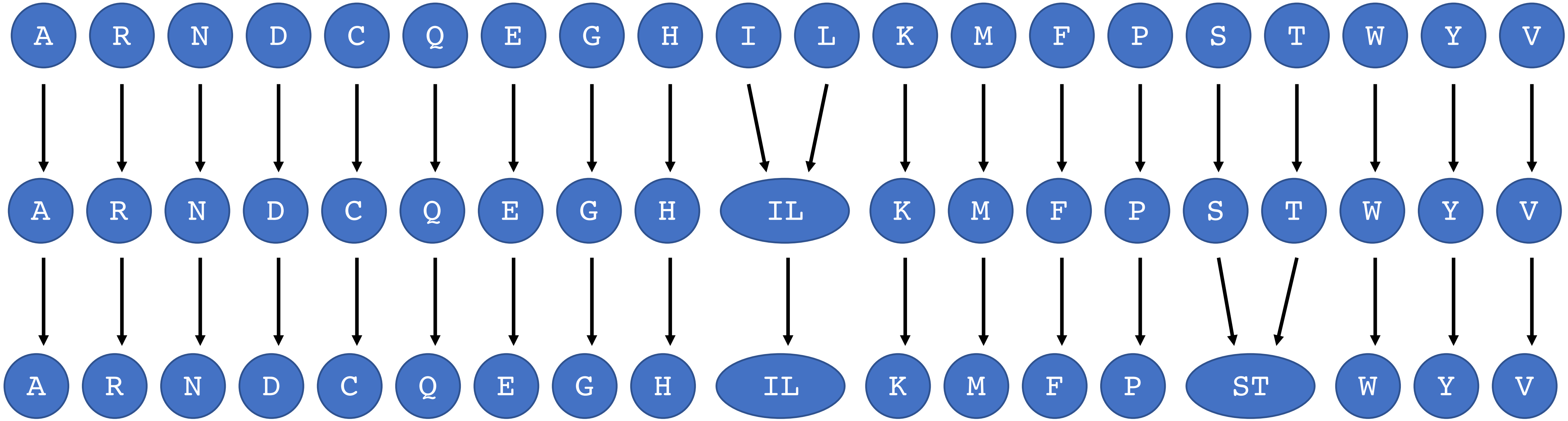
If **this** word/k-mer appears **here** there is implied significance.

What about with minor changes?

What about with major changes?

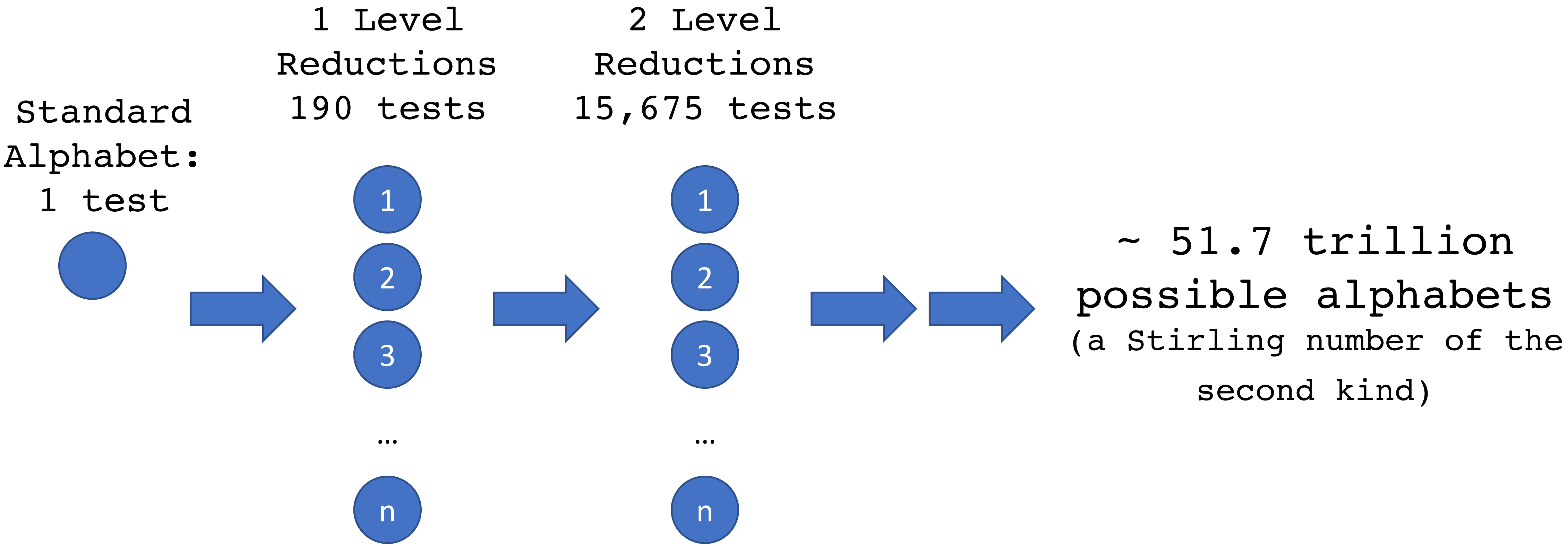
>Sequence02
MDQKMGDQCTPDHGMRNVPKHIFNKYPASTNEKDHYNMLDGAVNE

Performance of our classifier improves when the standard amino acid alphabet is substituted with a reduced alphabet:


A R N D C Q E G H I L K M F P S T W Y V
A R N D C Q E G H IL K M F P S T W Y V
A R N D C Q E G H IL K M F P ST W Y V

Test alphabet performance, perform a reduction, test again, repeat ad nauseum.

If reduced alphabets provide improved performance, how do we select the *best* reduced alphabet?



Standard Alphabet: 1 test

1 Level Reductions 190 tests

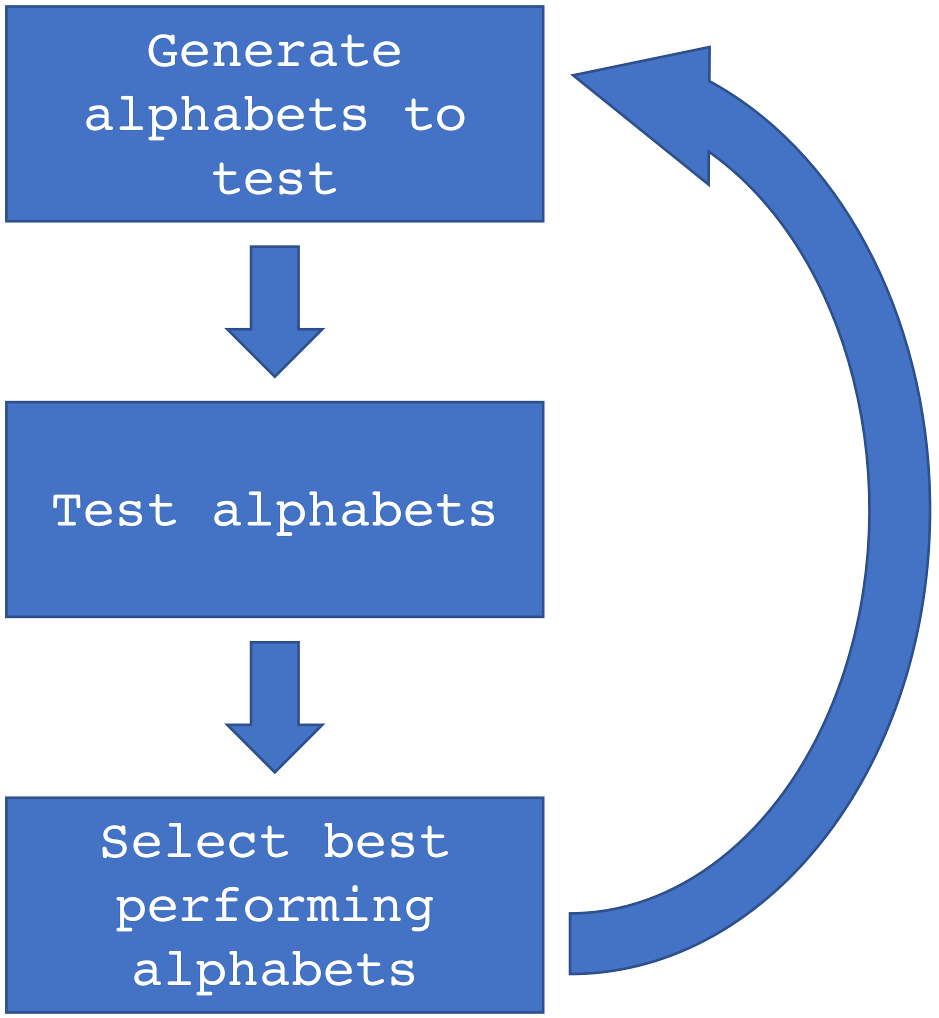
2 Level Reductions 15,675 tests

~ 51.7 trillion possible alphabets

(a Stirling number of the second kind)

51.7 trillion tests is probably too many tests.

- Iterate down through alphabet sizes
- Only test reductions of highest performing alphabets from previous level
- Avoid brute force testing of every possible alphabet



Generate alphabets to test

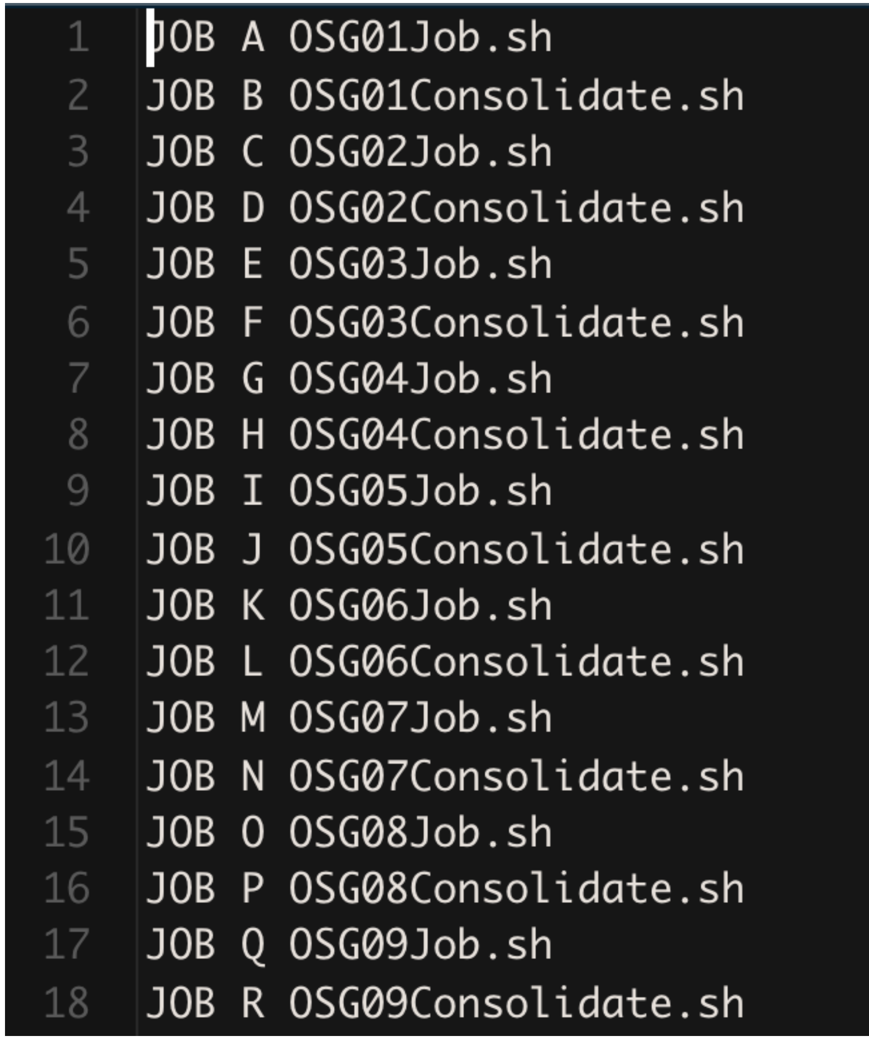
Test alphabets

Select best performing alphabets

We've got a DAG for that …

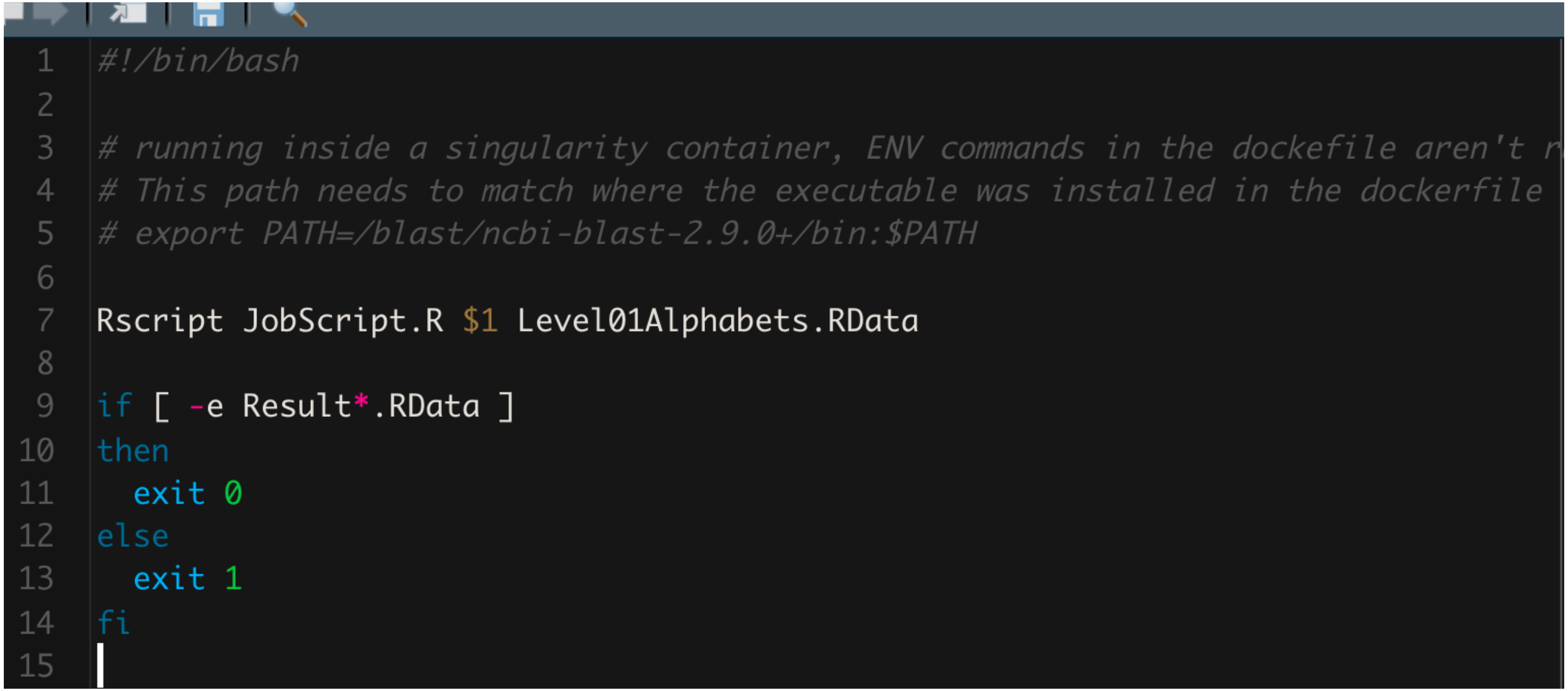
Test, Consolidate, Repeat

- Testing alphabets has modest requirements:
  - 1 CPU
  - 1 GB disk
  - > 4 GB memory
- Consolidating results at each level has trivial requirements:
  - 1 CPU
  - 1 GB disk
  - 2 GB memory
    - During consolidation, parameters for next level are set

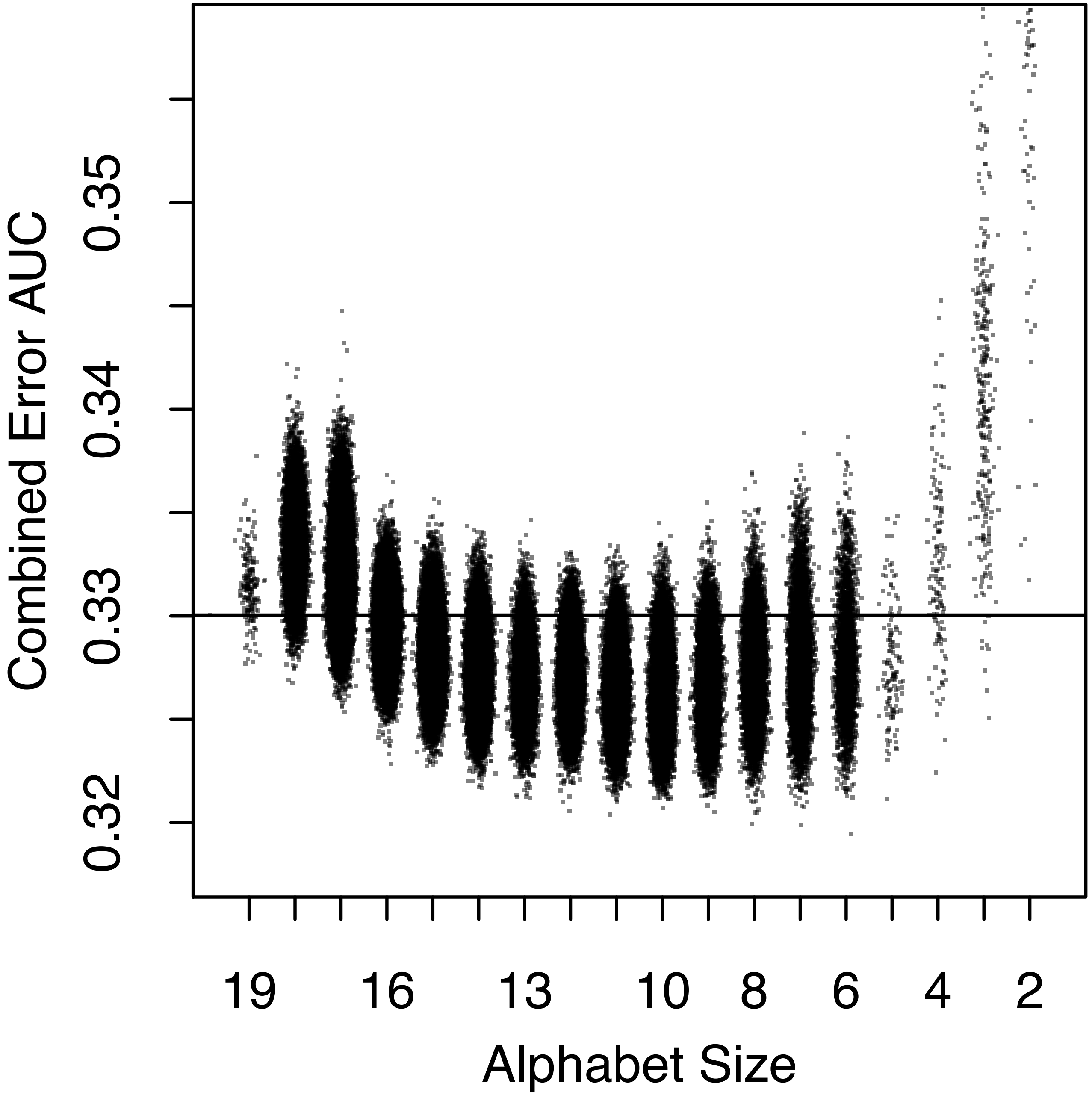


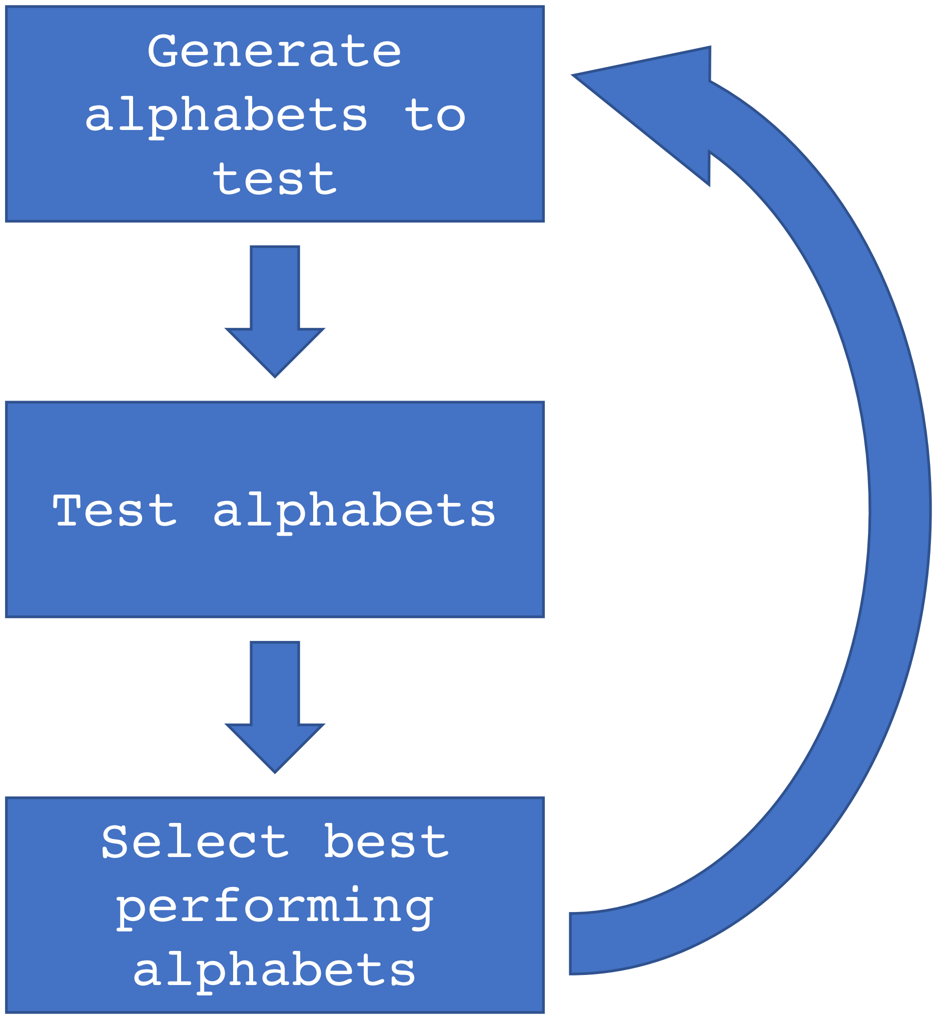
```
JOB A OSG01Job.sh
JOB B OSG01Consolidate.sh
JOB C OSG02Job.sh
JOB D OSG02Consolidate.sh
JOB E OSG03Job.sh
JOB F OSG03Consolidate.sh
JOB G OSG04Job.sh
JOB H OSG04Consolidate.sh
JOB I OSG05Job.sh
JOB J OSG05Consolidate.sh
JOB K OSG06Job.sh
JOB L OSG06Consolidate.sh
JOB M OSG07Job.sh
JOB N OSG07Consolidate.sh
JOB O OSG08Job.sh
JOB P OSG08Consolidate.sh
JOB Q OSG09Job.sh
JOB R OSG09Consolidate.sh
```

We've got a DAG for that …
And we try to keep it simple:



```
#!/bin/bash

# running inside a singularity container, ENV commands in the dockefile aren't r
# This path needs to match where the executable was installed in the dockerfile
# export PATH=/blast/ncbi-blast-2.9.0+/bin:$PATH

Rscript JobScript.R $1 Level01Alphabets.RData

if [ -e Result*.RData ]
then
  exit 0
else
  exit 1
fi

```



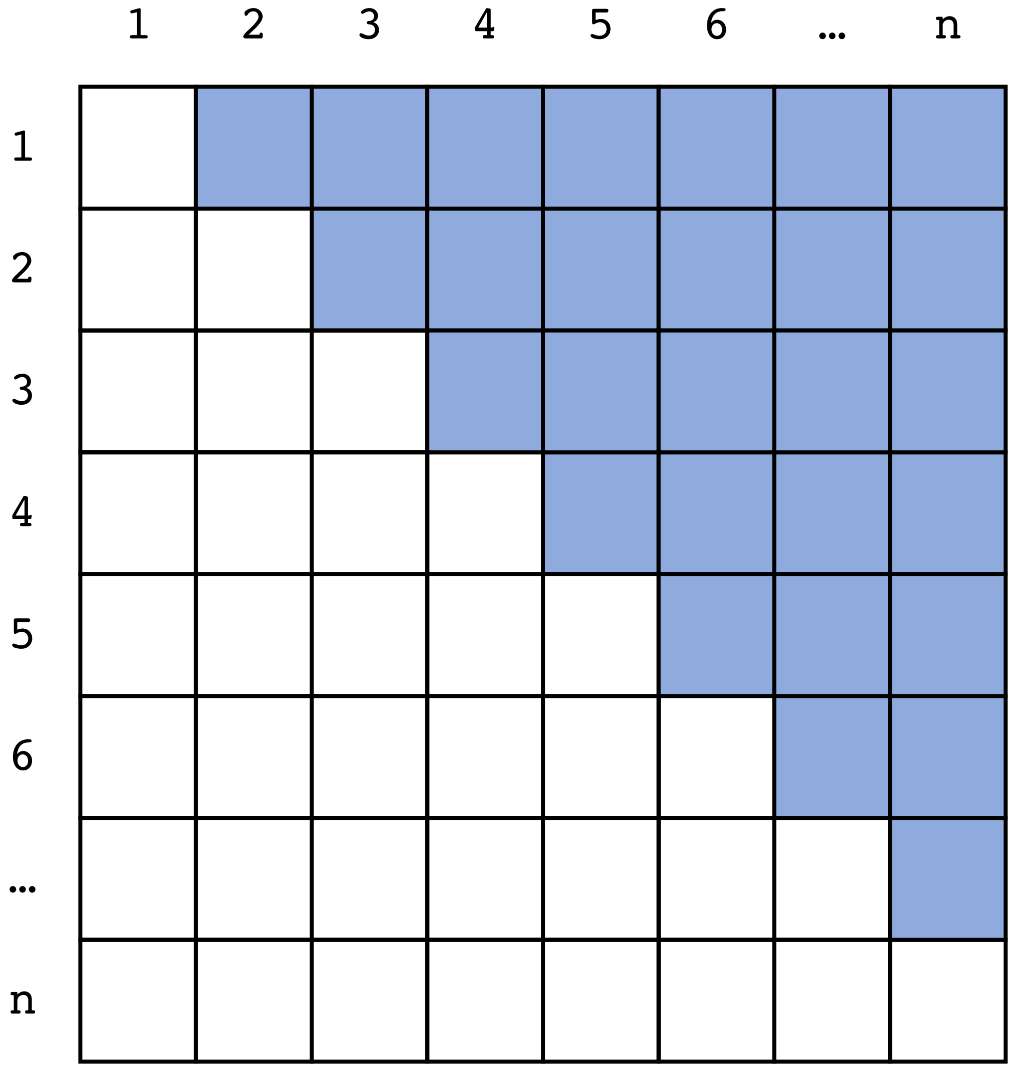


Generate alphabets to test

Test alphabets

Select best performing alphabets

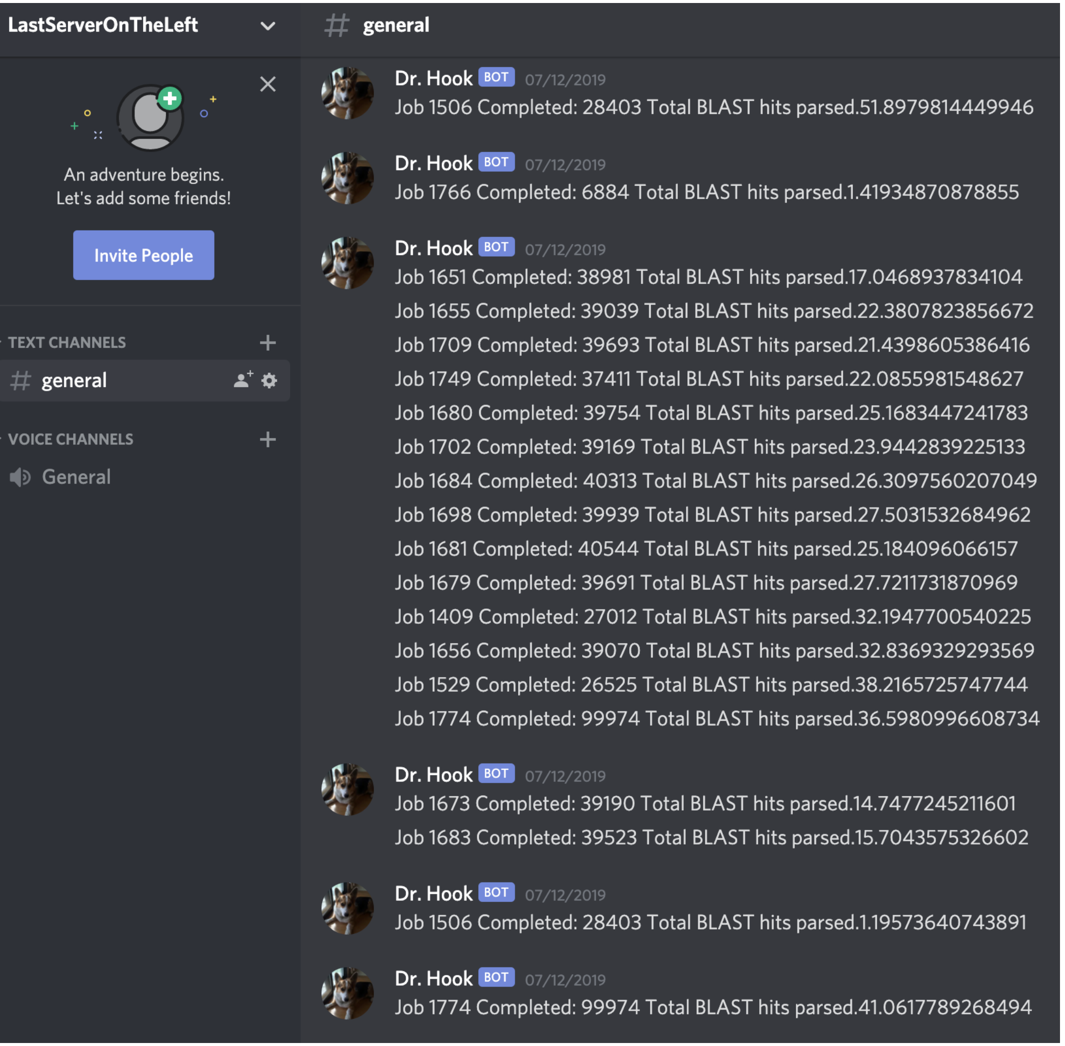
Other cool tidbits:



A typical job set up for us: Perform all pairwise comparisons in a set of genomes – give 1 comparison to each node.

With relatively trivial requirements for nodes (1 GB disk, 2 GB memory, 1 CPU) we can complete ~ 70,000 jobs at 10 minutes per job in a weekend.

Other cool tidbits:



```
Dr. Hook BOT 07/12/2019
Job 1506 Completed: 28403 Total BLAST hits parsed.51.8979814449946

Dr. Hook BOT 07/12/2019
Job 1766 Completed: 6884 Total BLAST hits parsed.1.41934870878855

Dr. Hook BOT 07/12/2019
Job 1651 Completed: 38981 Total BLAST hits parsed.17.0468937834104
Job 1655 Completed: 39039 Total BLAST hits parsed.22.3807823856672
Job 1709 Completed: 39693 Total BLAST hits parsed.21.4398605386416
Job 1749 Completed: 37411 Total BLAST hits parsed.22.0855981548627
Job 1680 Completed: 39754 Total BLAST hits parsed.25.1683447241783
Job 1702 Completed: 39169 Total BLAST hits parsed.23.9442839225133
Job 1684 Completed: 40313 Total BLAST hits parsed.26.3097560207049
Job 1698 Completed: 39939 Total BLAST hits parsed.27.5031532684962
Job 1681 Completed: 40544 Total BLAST hits parsed.25.184096066157
Job 1679 Completed: 39691 Total BLAST hits parsed.27.7211731870969
Job 1409 Completed: 27012 Total BLAST hits parsed.32.1947700540225
Job 1656 Completed: 39070 Total BLAST hits parsed.32.8369329293569
Job 1529 Completed: 26525 Total BLAST hits parsed.38.2165725747744
Job 1774 Completed: 99974 Total BLAST hits parsed.36.5980996608734

Dr. Hook BOT 07/12/2019
Job 1673 Completed: 39190 Total BLAST hits parsed.14.7477245211601
Job 1683 Completed: 39523 Total BLAST hits parsed.15.7043575326602

Dr. Hook BOT 07/12/2019
Job 1506 Completed: 28403 Total BLAST hits parsed.1.19573640743891

Dr. Hook BOT 07/12/2019
Job 1774 Completed: 99974 Total BLAST hits parsed.41.0617789268494
```

Monitoring jobs in real time is complicated.

```
npcooley$ watch –n 5 condor_q
```

But we’re not always at a work computer and ssh’d into our login node

Discord can collect results for us but …

Acknowledgements

Wright Lab:

- Erik Wright
- Maria Bond
- Andrew Beckley
- Allison Petrick
- Sam Blechman
- Shania Khatri
- Nishant Panicker

Funding:

- T15LM007059
- 1DP2AI145058-01

Open Science Grid:

- OSG User School
- Christina Koch
- Lauren Michael
- Mats Rynge
- Carrie Brown

1) Pordes, R. et al. (2007). "The Open Science Grid", J. Phys. Conf. Ser. 78, 012057.doi:10.1088/1742-6596/78/1/012057.

2) Sfiligoi, I., Bradley, D. C., Holzman, B., Mhashilkar, P., Padhi, S. and Wurthwein, F. (2009). "The Pilot Way to Grid Resources Using glideinWMS", 2009 WRI World Congress on Computer Science and Information Engineering, Vol. 2, pp. 428–432. doi:10.1109/CSIE.2009.950.