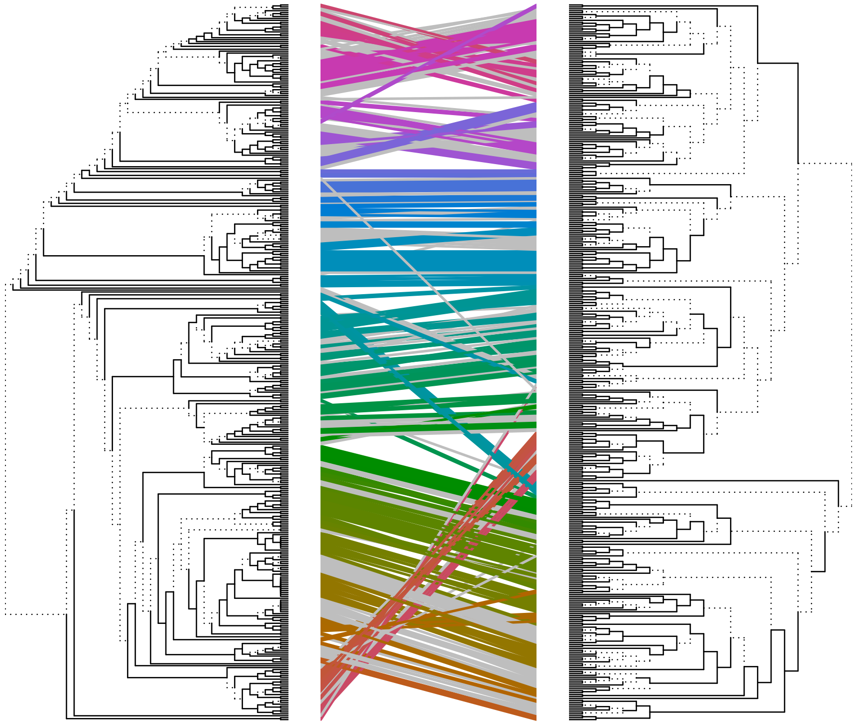
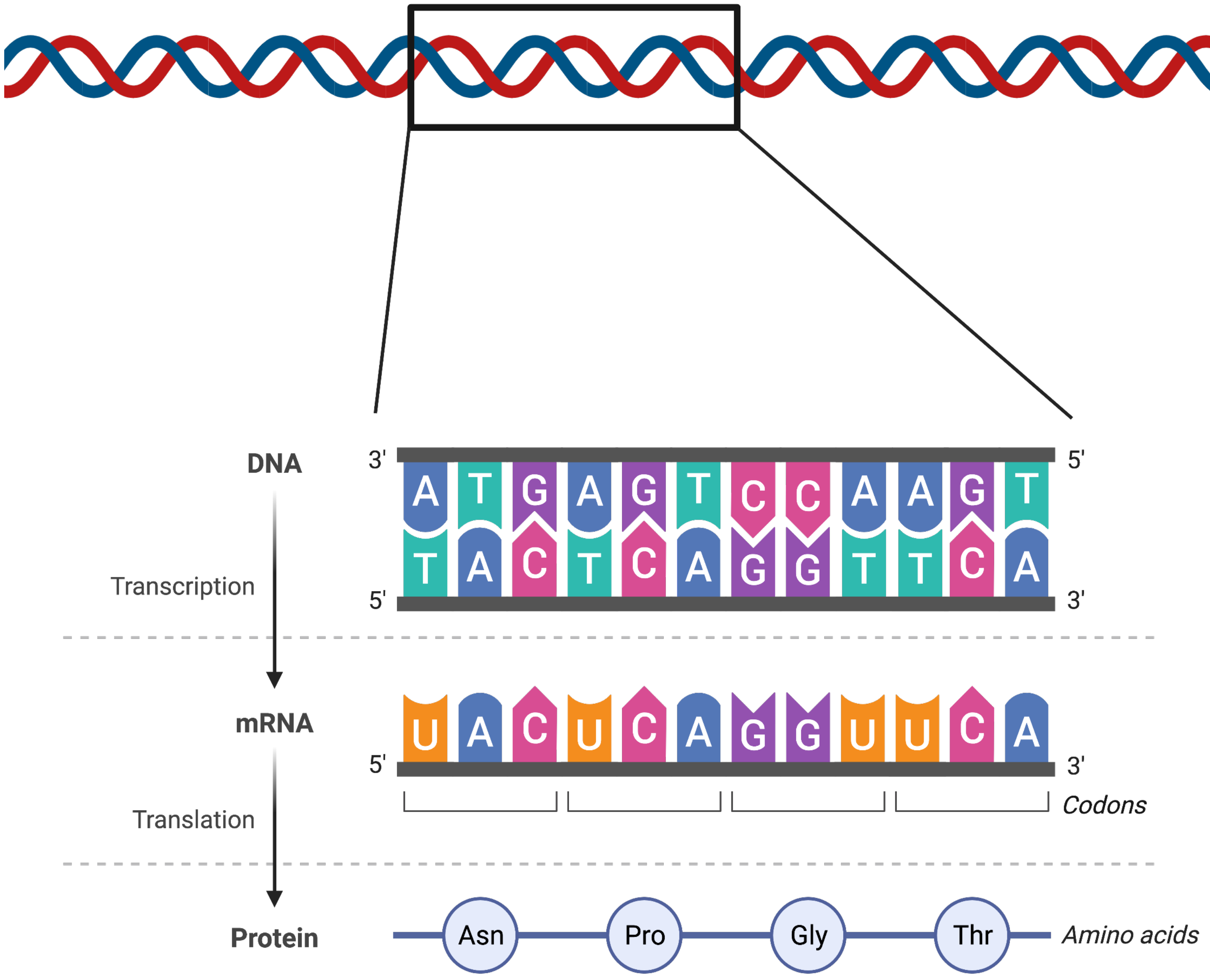
Building tools with the help of the OSG



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…
```

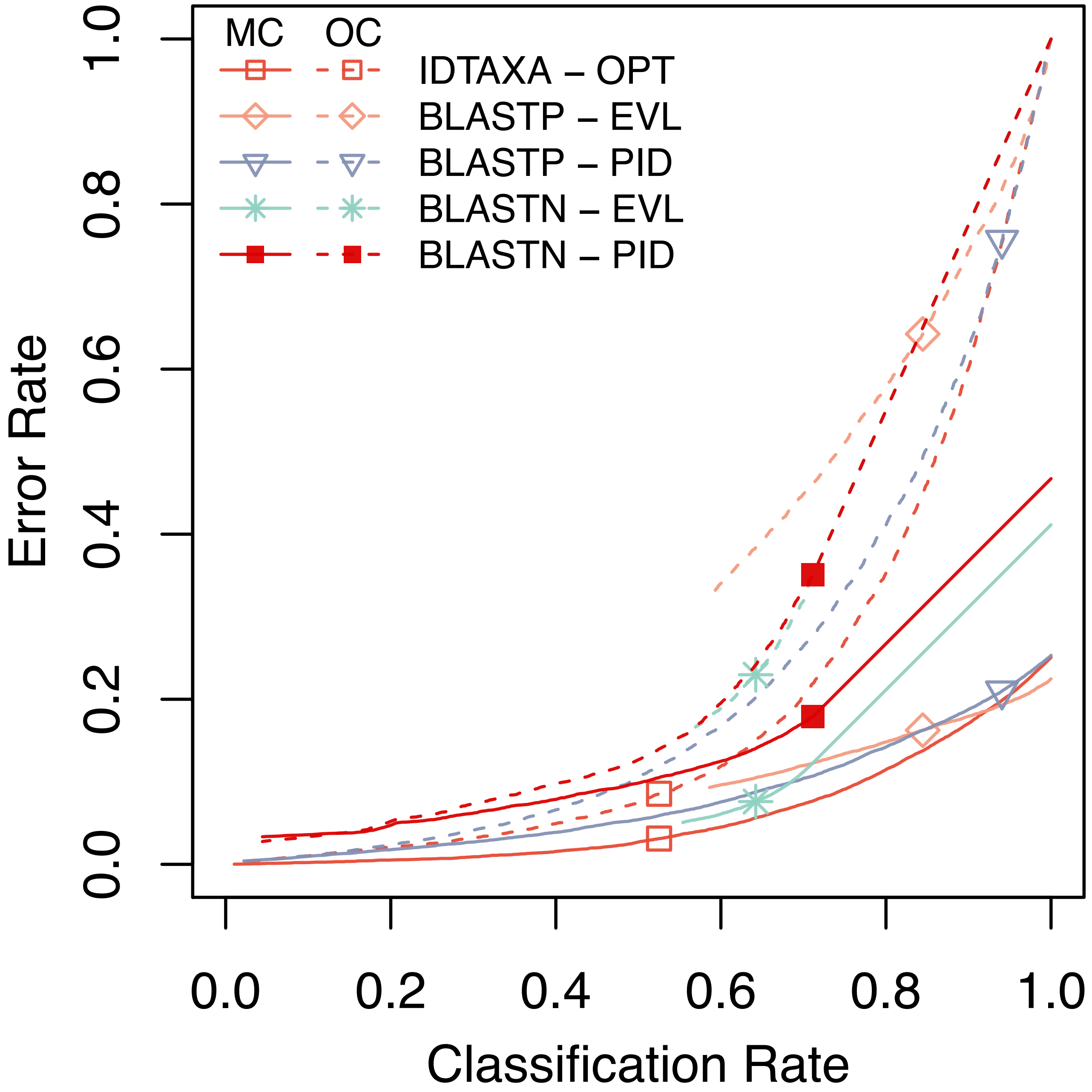
Does **this** sequence have a function with a representative in **this** library?

```
>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

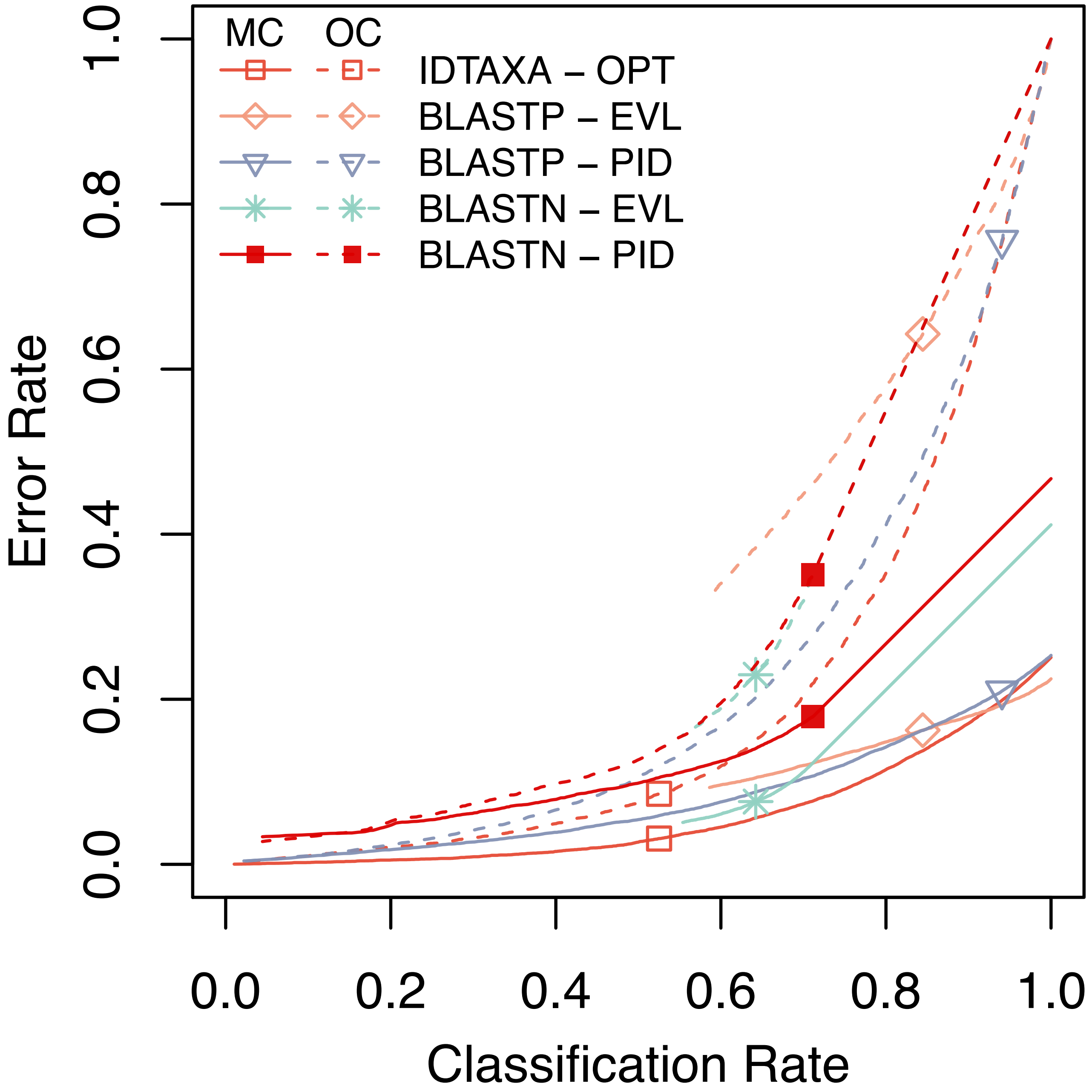
We're responding to reviewer comments on this paper now!



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

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

We're responding to reviewer comments on this paper now!



One last bit of biology to introduce:

>Sequence01
MSADDHGMRNVPKHIFNKGLKHWPKYRPITWQLSDFGEWEFDS

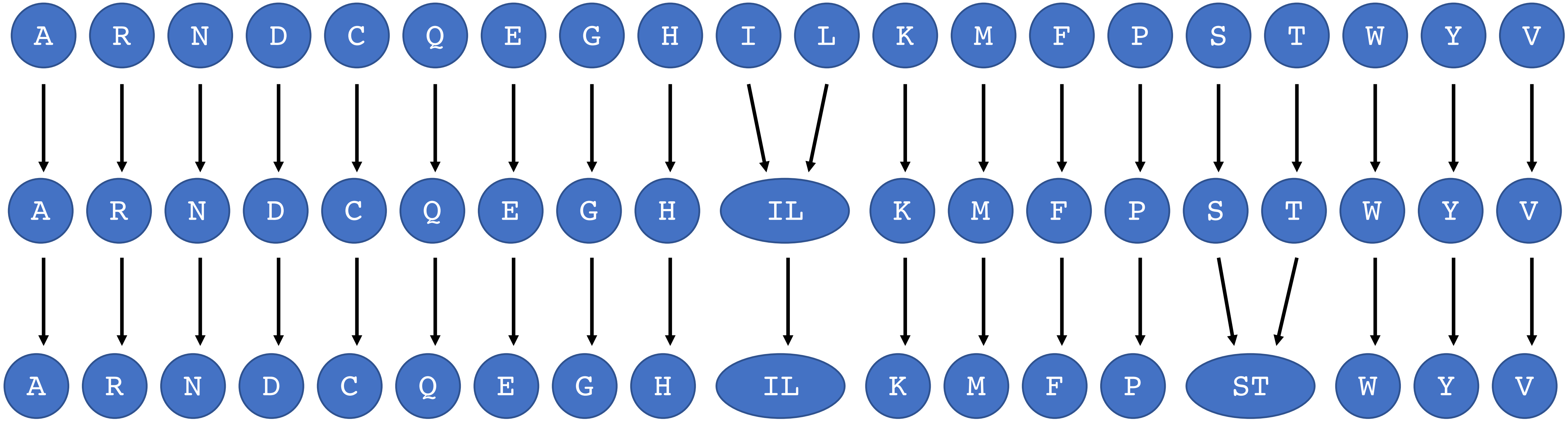
If **this** word/k-mer appears **here** there is implied significance.

What about with minor changes?

What about with major changes?

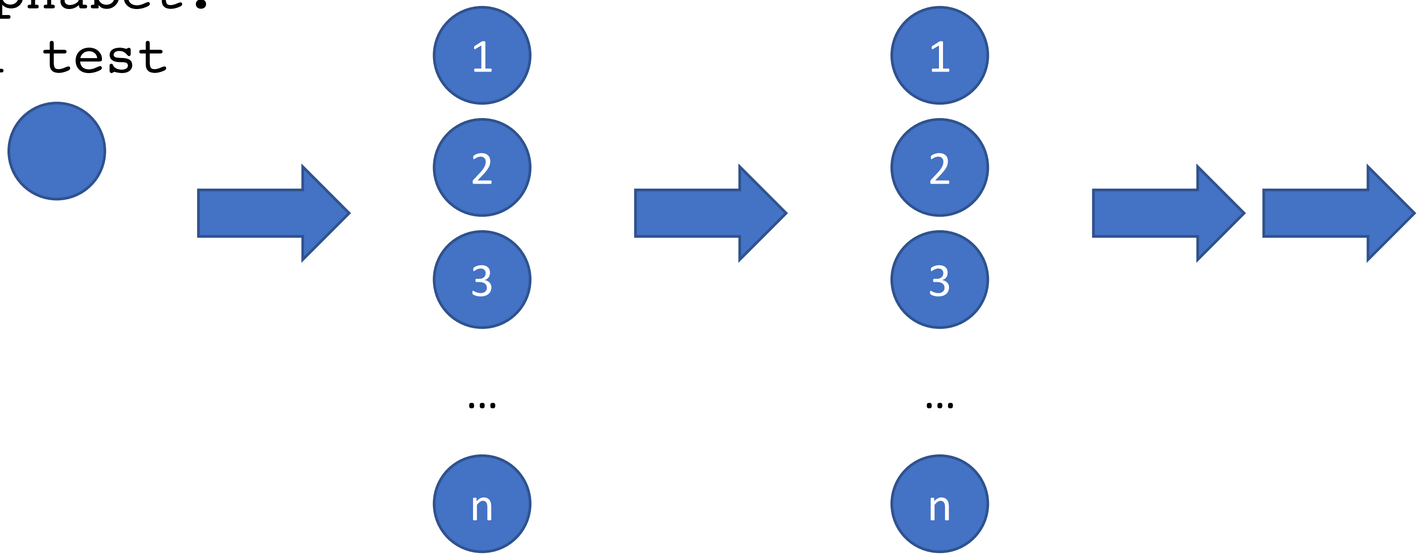
>Sequence02
MDQKMGDQCTPDHGMRNVPKHIFNKYPASTNEKDHYNMLDGAVNE

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



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

Size 19
Alphabets:
190 tests

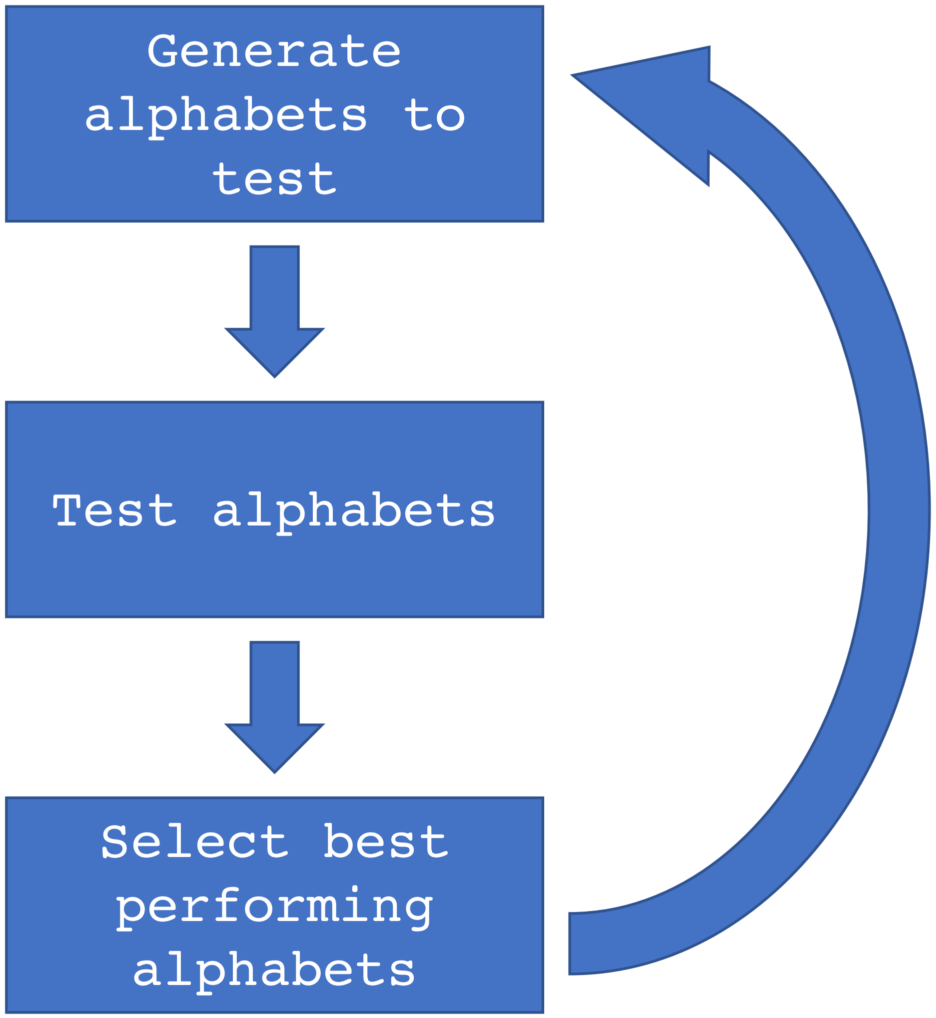
Size 18
Alphabets:
15,675 tests

~ 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

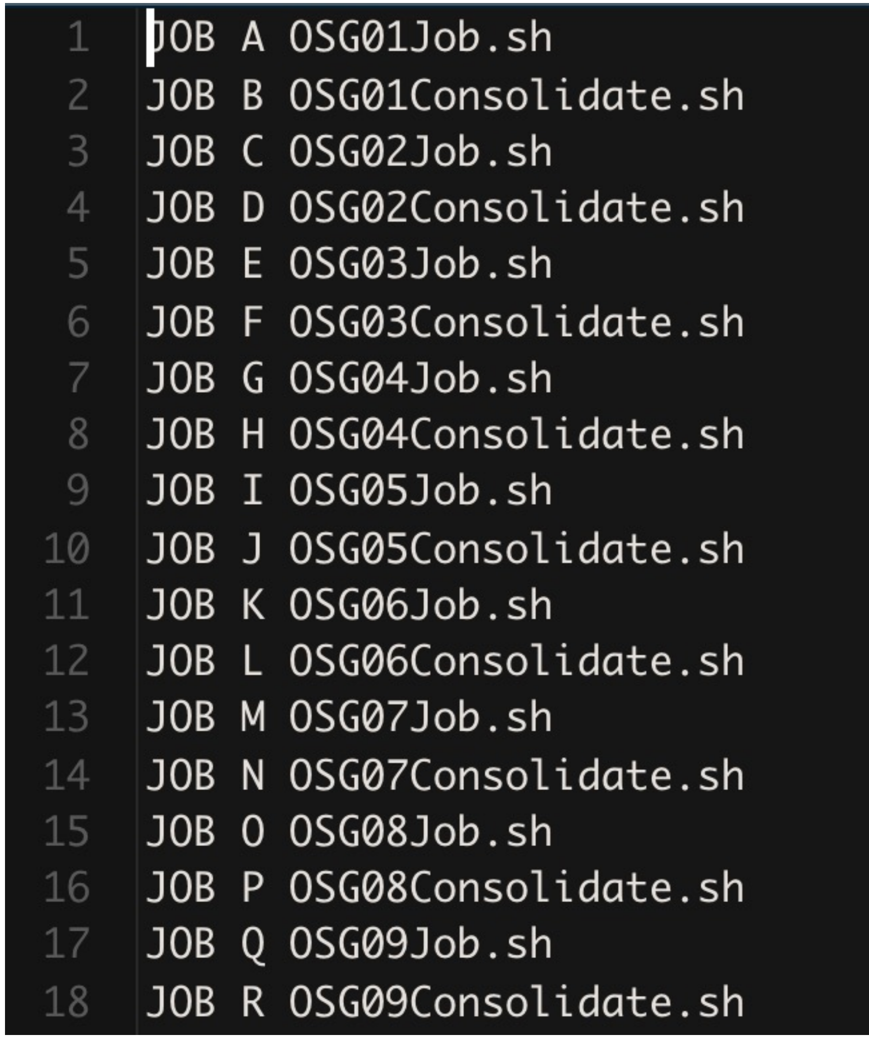
Test alphabets

Select best performing alphabets

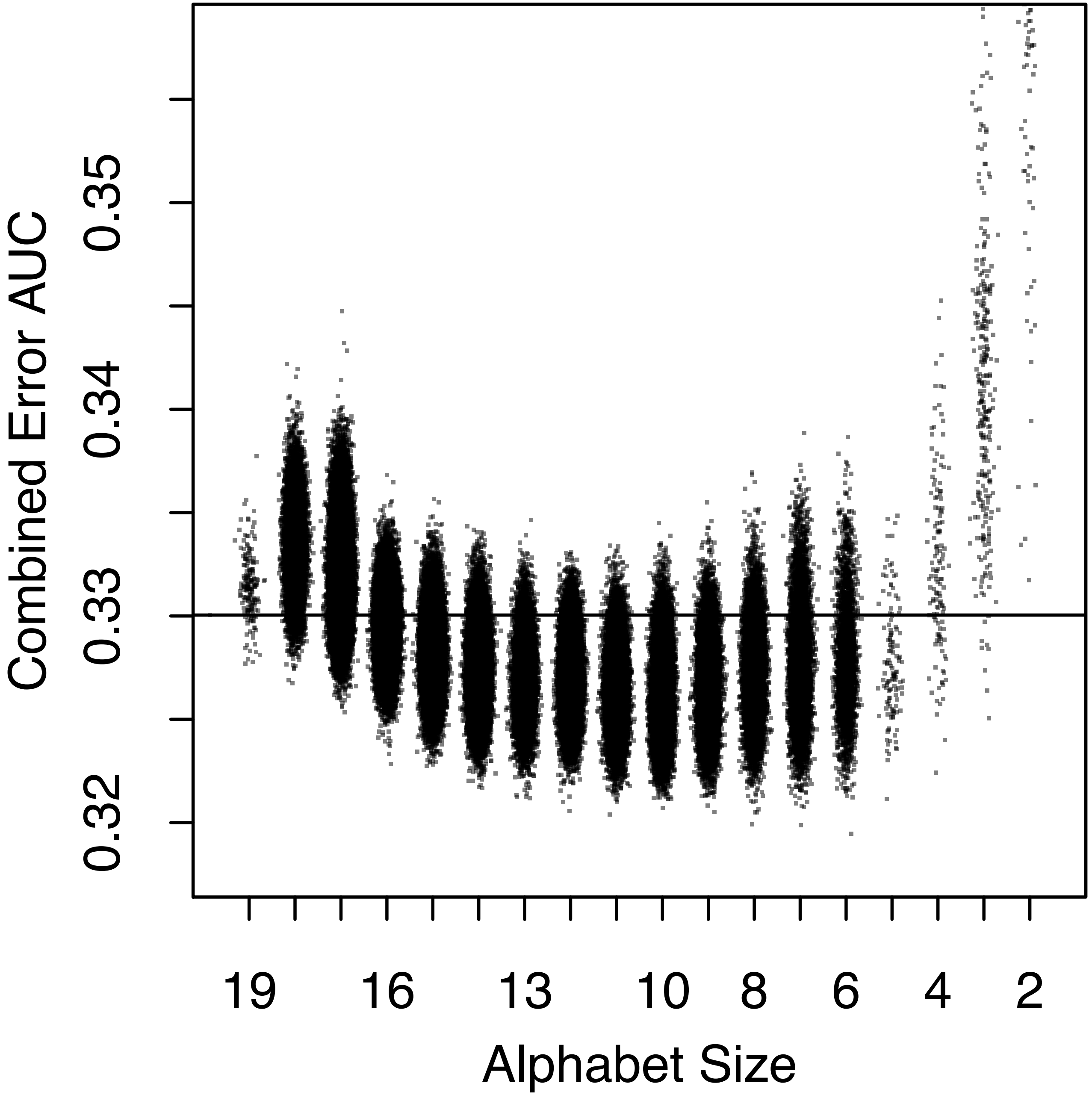
We’ve got a DAG for that …

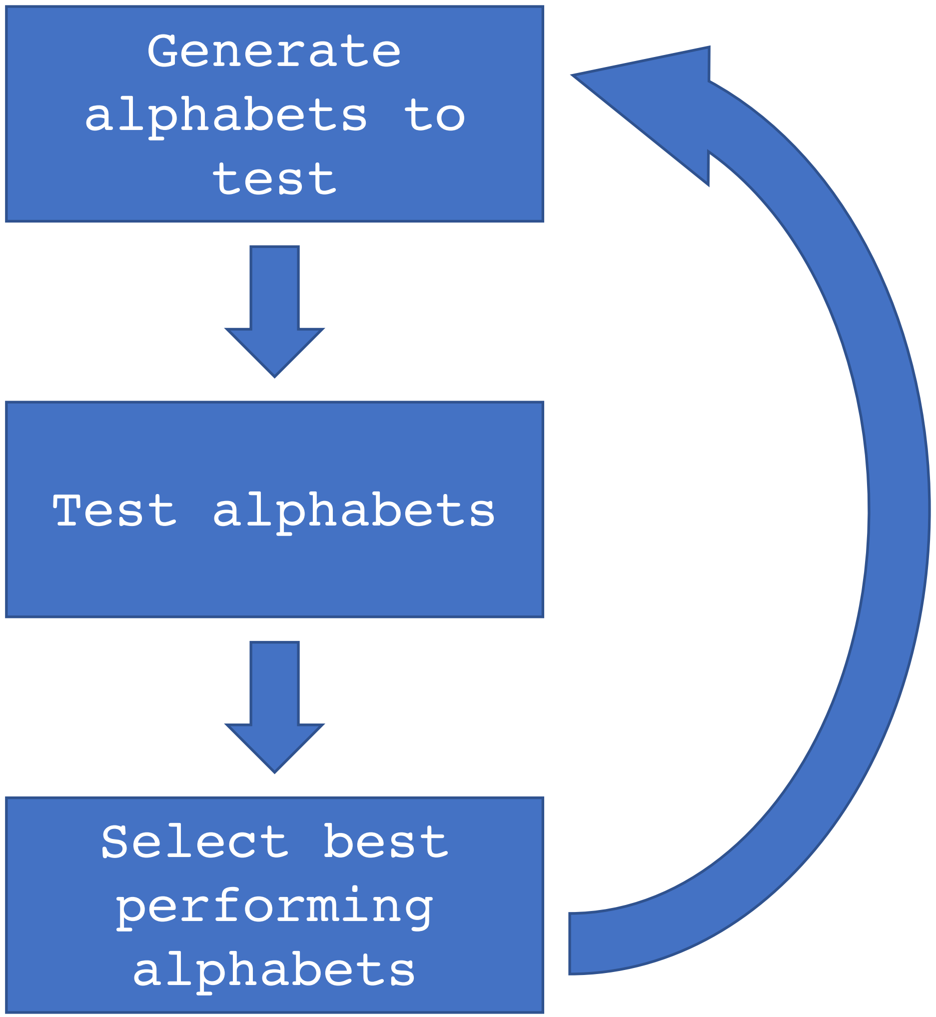
Test, Consolidate, Repeat

- 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
```




Generate alphabets to test

Test alphabets

Select best performing alphabets

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.