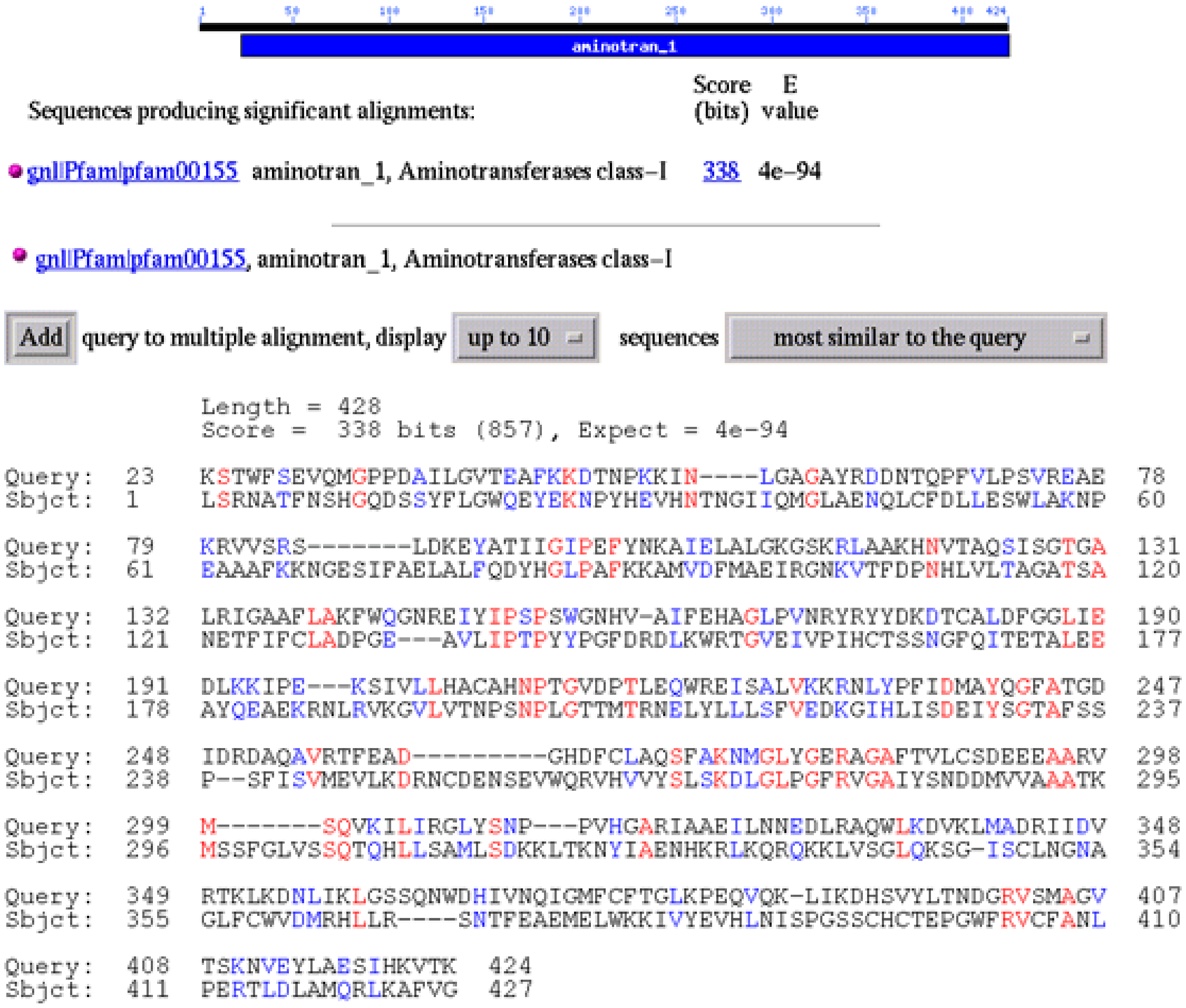
# COVID modeling

James Morton

# Sequence search

- Sequence similarity forms the basis of
  - Protein function prediction
  - Evolutionary analysis
  - Protein interaction analysis
  - Many more applications

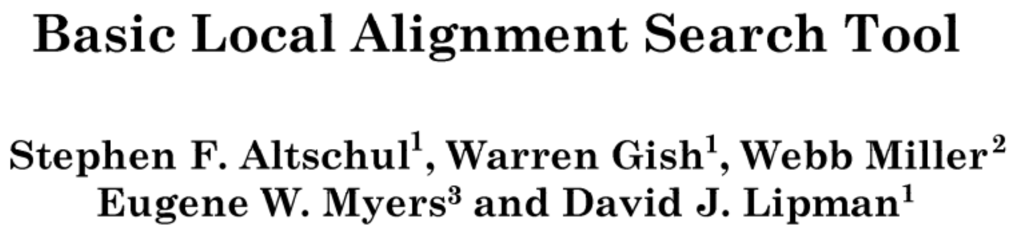
# BLAST



```
Sequences producing significant alignments:                  Score    E
                                                             (bits)  value

● gnl|Pfam|pfam00155  aminotran_1, Aminotransferases class-I    338   4e-94

● gnl|Pfam|pfam00155, aminotran_1, Aminotransferases class-I
```

Add query to multiple alignment, display up to 10 sequences most similar to the query

```
          Length = 428
          Score =  338 bits (857), Expect = 4e-94

Query:  23   KSTWFSEVQMGPPDAILGVTEAFKKDTNPKKIN----LGAGAYRDDNTQPFVLPSVREAE  78
Sbjct:  1    LSRNATFNSHGQDSSYFLGWQEYEKNPYHEVHNTNGIIQMGLAENQLCFDLLESWLAKNP  60

Query:  79   KRVVSRS-------LDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHNVTAQSISGTGA  131
Sbjct:  61   EAAAFKKNGESIFAELALFQDYHGLPAFKKAMVDFMAEIRGNKVTFDPNHLVLTAGATSA  120

Query:  132  LRIGAAFLAKFWQGNREIYIPSPSWGNHV-AIFEHAGLPVNRYRYYDKDTCALDFGGLIE  190
Sbjct:  121  NETFIFCLADPGE---AVLIPTPYYPGFDRDLKWRTGVEIVPIHCTSSNGFQITETALEE  177

Query:  191  DLKKIPE---KSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAYQGFATGD  247
Sbjct:  178  AYQEAEKRNLRVKGVLVTNPSNPLGTTMTRNELYLLLSFVEDKGIHLISDEIYSGTAFSS  237

Query:  248  IDRDAQAVRTFEAD---------GHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARV  298
Sbjct:  238  P--SFISVMEVLKDRNCDENSEVWQRVHVVYSLSKDLGLPGFRVGAIYSNDDMVVAAATK  295

Query:  299  M-------SQVKILIRGLYSNP---PVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDV  348
Sbjct:  296  MSSFGLVSSQTQHLLSAMLSDKKLTKNYIAENHKRLKQRQKKLVSGLQKSG-ISCLNGNA  354

Query:  349  RTKLKDNLIKLGSSQNWDHIVNQIGMFCFTGLKPEQVQK-LIKDHSVYLTNDGRVSMAGV  407
Sbjct:  355  GLFCWVDMRHLLR----SNTFEAEMELWKKIVYEVHLNISPGSSCHCTEPGWFRVCFANL  410

Query:  408  TSKNVEYLAESIHKVTK  424
Sbjct:  411  PERTLDLAMQRLKAFVG  427
```



## Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

150k citations

1,688,561 papers

# Recent work

Bepler et al 2019 : LSTMs

- Contact map prediction
- Sequence similarity

Alley et al 2019 : LSTMs

- Protein engineering

Rives et al 2019: Attention model

- Contact map prediction

Rao et al 2019: LSTMs + Attention

- Contact map prediction
- Protein engineering

Madani et al 2019: Attention

- Protein engineering

**Recommended strategy**: Learn residue coordinates (per protein) in an unsupervised manner. Then use the coordinates for downstream tasks (i.e. classification / regression)

**Benefits**: requires *much* less labeled data

# Our benchmark

Pfam 10k triplets (sequence based)
SCoP: 6k triplets (structure based)

PFAM: 10M sequences
Uniref90: 100M sequences

| | **PFam** | **SCoP** |
|---|---|---|
| ELMo PFam | 0.95 | **0.81** |
| RoBERTa base PFam | 0.52 | 0.50 |
| RoBERTa base Uniref90 | 0.98 | **0.77** |
| BLAST | 0.99 | 0.04 |

# Existing limitations

- Viral proteins are known to be under-represented in databases
  - Metagenomics samples are better representative
- The more proteins, the better search will become

- BFD: 2.5B proteins from Uniref + metagenomics
  - Note: add link
- Compute resources
  - Lack of openly available models

# What do we need

- Current compute estimate: 7 days 1024 V100 GPUs (170k GPU hours)
  - Previous study: 128 V100s, 4 days, 250M proteins
  - Requirements: 32GB RAM per GPU, 100GB CPU RAM per node
    - 2TB of hard disk storage per node at a minimum
    - Homogenous compute (each node has the same GPU hardware and the same number of GPUs)
    - MPI is preferred

# Next immediate steps

- Make model publicly immediately available upon completion
  - Followed by benchmarks within 1 week

- Task 1: Perform protein search against existing drug databases

- Task 2: Build drug-protein interaction model (finetuning)

# Model Size

- ~700M parameters
    - 36 layers
    - 16 attention heads
    - 1536 embedding dimensions
    - 4096 FFN