Application of Structural Alignment in Immunology

Daron Standley & John Rozewicki
Research Institute for Microbial Diseases
Department of Genome Informatics
Osaka University

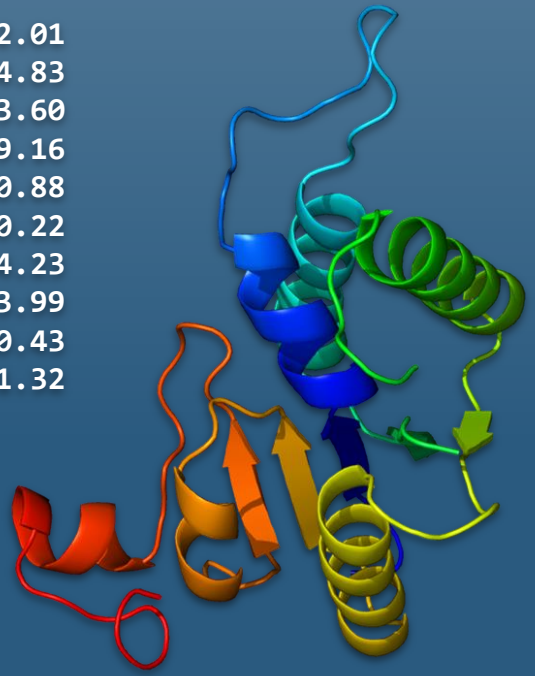
Sequence and Structure

PDB/mmCIF

```
ATOM     65  N   GLY A  31     -51.397  -9.757  12.360  1.00 22.01
ATOM     66  CA  GLY A  31     -51.023  -9.667  13.758  1.00 24.83
ATOM     67  C   GLY A  31     -51.645 -10.743  14.620  1.00 23.60
ATOM     68  O   GLY A  31     -50.969 -11.318  15.465  1.00 29.16
ATOM     69  N   SER A  32     -52.926 -11.040  14.407  1.00 20.88
ATOM     70  CA  SER A  32     -53.596 -12.031  15.253  1.00 20.22
ATOM     71  C   SER A  32     -52.979 -13.389  15.010  1.00 24.23
ATOM     72  O   SER A  32     -52.770 -14.157  15.953  1.00 23.99
ATOM     73  CB  SER A  32     -55.098 -12.110  14.975  1.00 20.43
ATOM     74  OG  SER A  32     -55.727 -10.876  15.253  1.00 31.32
```

FASTA

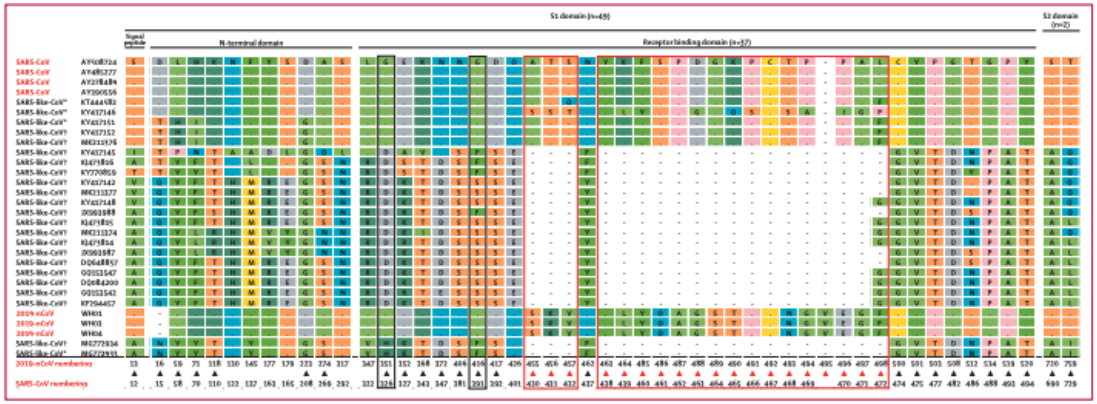
```
>3V33_A
GGGTPKAPNLEPPLPEEEKEGSDLRPVVIDGSNVAMSHGNKEVFSCRGILLAVNWFL
ERGHTDITVFVPSWRKEQPRPDVPITDQHILRELEKKKILVFTPSRRVGGKRVVCYD
DRFIVKLAYESDGIVVSNDTYRDLQGERQEWKRFIEERLLMYSFVNDKFMPPDDPLG
RHGPSLDNFLRKKPLTLEHRKQPCPYGRKCTYGIKCRFFHPERPSCPQRSVA
```



Example: Regnase-1

Multiple Sequence Alignment (MSA)

Sequences were aligned using MAFFT



THE LANCET

Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding

Roujian Lu, Xiang Zhao*, Juan Li*, Peihua Niu*, Bo Yang*, Honglong Wu*, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, Yuhai Bi, Xuejun Ma, Faxian Zhan, Liang Wang, Tao Hu, Hong Zhou, Zhenhong Hu, Weimin Zhou, Li Zhao, Jing Chen, Yao Meng, Ji Wang, Yang Lin, Jianying Yuan, Zhihao Xie, Jinmin Ma, William J Liu, Dayan Wang, Wenbo Xu, Edward C Holmes, George F Gao, Guizhen Wu¶, Weijun Chen¶, Weifeng Shi¶, Wenjie Tan¶*

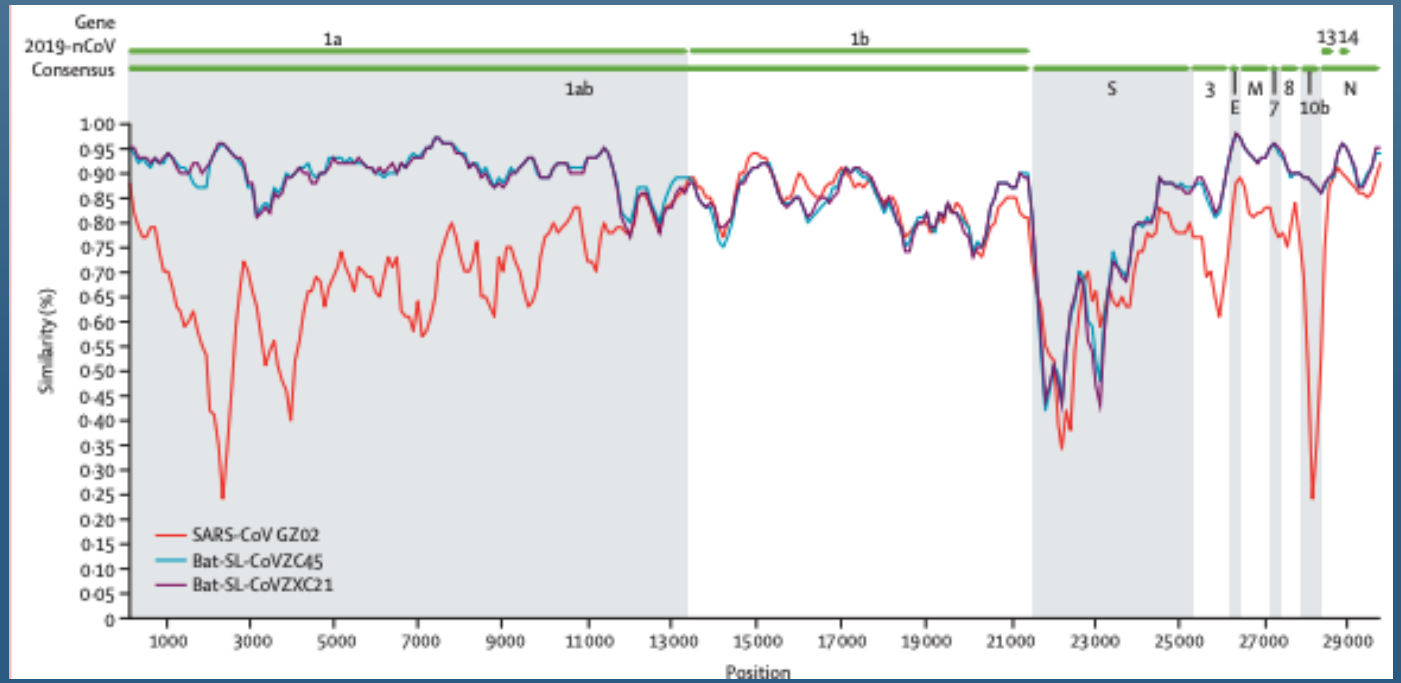
The Novel Coronavirus Epidemic in China: How to Help Researchers Using Sequence Alignment on 2019-nCoV with MAFFT

by Shen Huang

2020-01-27

Importance of MSAs

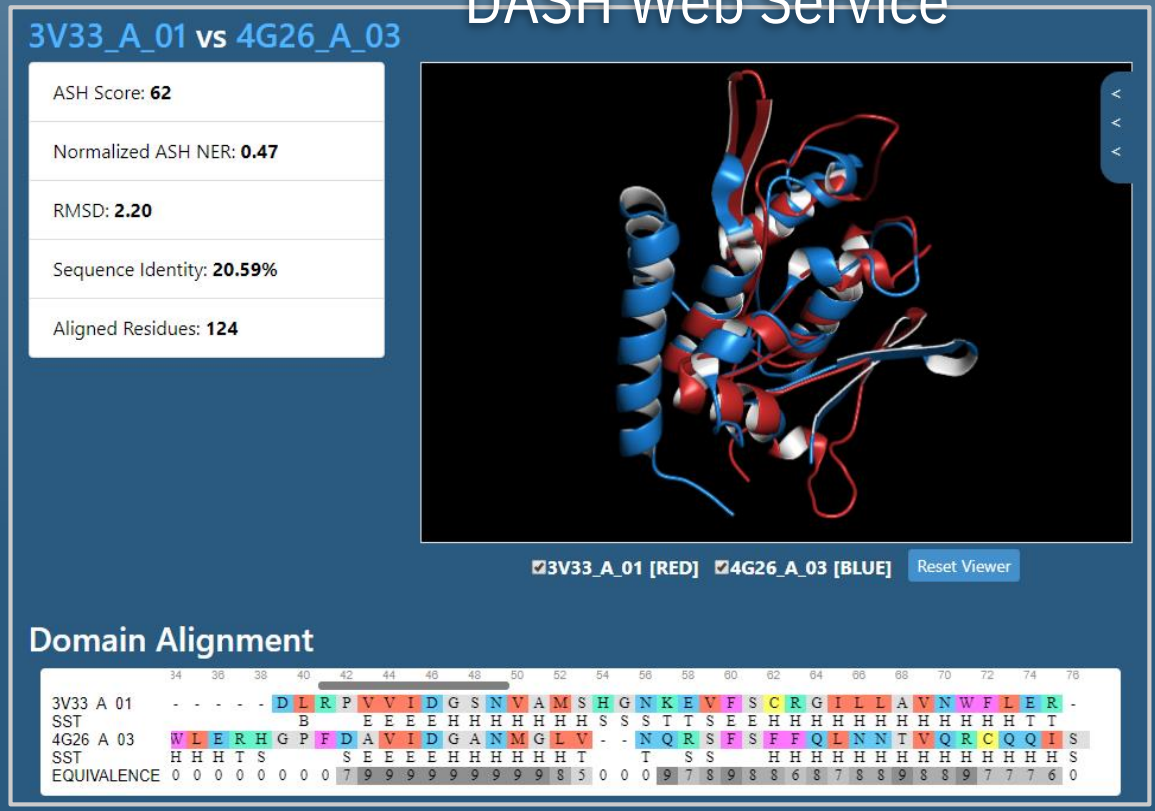
Sequence identity between the consensus of 2019-nCoV and representative betacoronavirus genomes



Structural Alignment

- 3D structures can also be aligned in 3D space
- Corresponding residues can be scored/calculated

DASH Web Service



DASH:
Database of Aligned Structural Homologs

- ASH structure alignment database
- Tracking:
  - ~60 million chain alignments
  - ~100 million domain alignments
- Integrated with MAFFT through public REST interface

- Conceived in 2003 at IPR
- Development started in 2008
- Released to public in 2019

Rozewicki, et. al.
Nucleic Acids Research
2019

Why structural alignments?

Most common use: Protein Function Prediction

Zc3h12a: Unknown Function

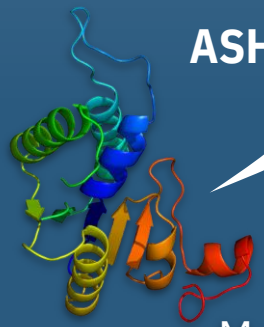
MSGPCGEKPVLEASPTMSLWEFEDSHSRQGTPRPGQELAAEEASALELQMKVDFFRKLGY
SSTEIHSVLQKLGVQADTNTVLGELVKHGTATERERQTSPDPCPQLPLVPRGGGTPKAPN
LEPPLPEEEKEGSDLRPVVIDGSNVAMSHGNKEVFSCRGILLAVNWFLERGHTDITVFVP
SWRKEQPRPDVPITDQHILRELEKKKILVFTPSRRVGGKRVVCYDDRFIVKLAYESDGIV
VSNDTYRDLQGERQEWKRFIEERLLMYSFVNDKFMPPDDPLGRHGPSLDNFLRKKPLTLE
HRKQPCPYGRKCTYGIKCRFFHPERPSCPQRSVADELRANALLSPPRAPSKDKNGRRPSP
SSQSSSLLTESEQCSLDGKKLGAQASPGSRQEGLTQTYAPSGRSLAPSGGSGSSFGPTDW
LPQTLDSLPYVSQDCLDSGIGSLESQMSELWGVRGGGPGEPGPPRAPYTGYSPYGSELPA
TAAFSAFGRAMGAGHFSVPADYPPAPPAFPPREYWSEPYPLPPPTSVLQEPPVQSPGAGR
SPWGRAGSLAKEQASVYTKLCGVFPPHLVEAVMGRFPQLLDPQQLAAEILSYKSQHPSE

Zc3h12a Model

Template (2QIP)

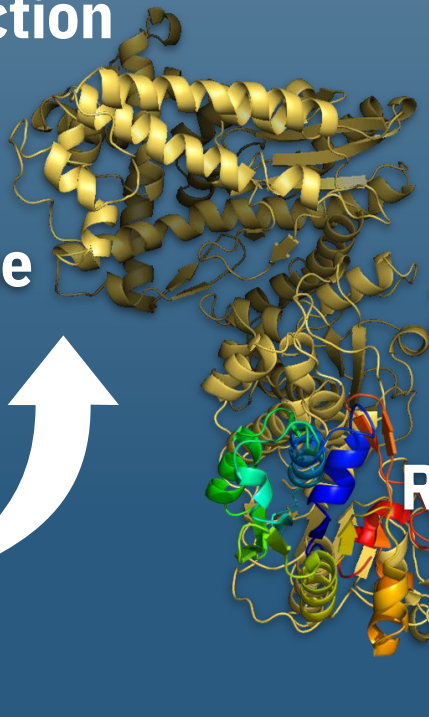
2QIP Function Unknown

Protein 3000 Target



ASH

Taq Polymerase



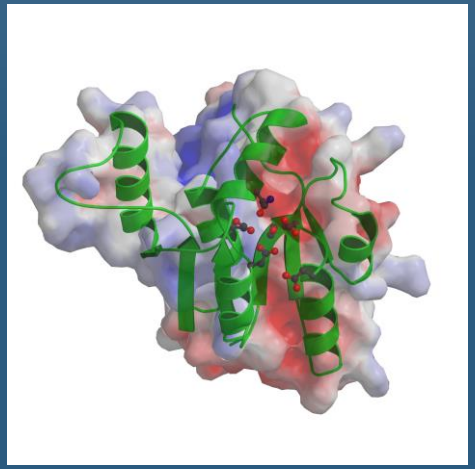
Zc3h12a -> Regnase-1

Matsushita, et. al.; Nature 2009

Other refs

Why multiple alignments?

MAFFT and MAFFT-DASH MSAs of Regnase, Taq polymerase and 2QIP



MAFFT Alignment

```
Q5D1E8/112-297 GGGTPKAPNLEPPLPEEEKEGSDLRPVVIDGSNVAMS--HGNKEVFSCRG-ILLAVNWFL
Q79YT8/  1-156 MQSD-------------HKEK---IAILVDVQNVYYTCREAY------RSNFDYNQFWYV
P19821/  1-144 MRGM---------LPLFEPKG---RVLLVDGHHLAYRTFHALKGLTTSRGEPVQAVYGFA

Q5D1E8/112-297 --------ERGHTDITVF---VPSWRKE---------QPRPDVPITDQHILRELEKKKIL
Q79YT8/  1-156 ATQEKEVVSAKAYAIASNDPKQRQFHH----------------------ILRGVGFEVML
P19821/  1-144 KSLLKALKEDGDAVIVVFDAKAPSFRHEAYGGYKAGRAPTPEDFPRQLALIKEL------

Q5D1E8/112-297 VFTPSRRVGGKR-----VVCYDDRFIVKLA------YESDGIVVS----------NDTYL
Q79YT8/  1-156 KPYIQRRDGSAKGDWDVGITLD---AIEIA------PDVDRVIL---------------V
P19821/  1-144 -------------------VD---LLGLARLEVPGYEADDVLASLAKKAEKEGYEVRIL

Q5D1E8/112-297 RKKPL
Q79YT8/  1-156 SGDGD
P19821/  1-144 TADKD
```

Conserved Aspartic Acid

MAFFT-DASH Alignment

```
Q5D1E8/112-297 GGGTPKAPNLEPPLPEEEKEGSDLR-PVVIDGSNVAMSHGNK----------EVFSCRGI
Q79YT8/  1-156 ----------------MQSDH-KEKIAILVDVQNVYYTCREAYR----------SNFDY
P19821/  1-144 MRGM---------LPLFEPKG---R-VLLVDGHHLAYRTFHALKGLTTSRGEPVQAVYGF

Q5D1E8/112-297 LLAVNWFLERGHTDITVFVP----SWRKE---------QPRP-DVPITDQHILRELEKKK
Q79YT8/  1-156 NQFWYVATQEKEVVSAKAYA----IASND----------PKQ-------RQFHHILRGVG
P19821/  1-144 AKSLLKALKEDGDAVIVVFDAKAPSFRHEAYGGYKAGRAPTPEDFPRQLALIKELVDLLG

Q5D1E8/112-297 ILVFTPS---RRVGGKRVVCYDDRFIVKLAYE-----SD-----GIVVSNDTYLRKKPL
Q79YT8/  1-156 FEVMLKPYIQRRDGSAK--GDWDVGITLDAIE-----IAPDVDRVILVSGDGD------
P19821/  1-144 LARLEVP----------GYEADDVLASLAKKAEKEGYE-----VRILTADKD------
```

PDB & PDBj in 2003

- Only 24,000 entries in the PDB
- Structural Genomics was starting to take off
- National goal for Japanese labs to solve 3000 protein structures in 3 years
- Good structural alignment tools were becoming more and more necessary

Structure Resources

Annotation DB's

- Domain parsing
- Biological hierarchies
- Structure neighbors

Examples: CATH, SCOP

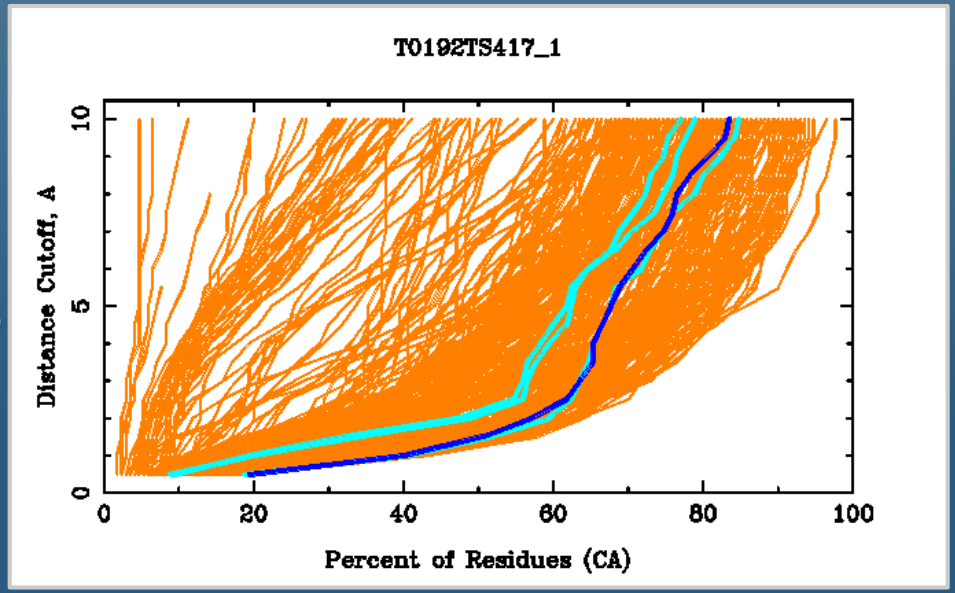
Structure Aligners

- Structural comparison
- Superimposing
- Scoring

Examples: Dali, GDT

GDT: Global Distance Test

- Used for CASP
- Local alignments generated with LGA
- Weighted sum of the number of aligned residues within 20 different distance cutoffs (0.5, 1, 1.5…10 Å)

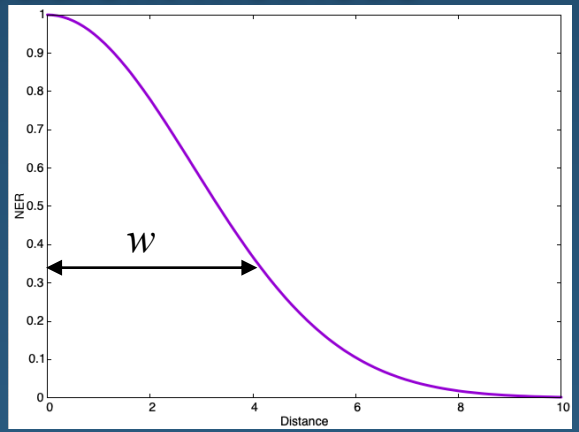


ASH:
Alignment of Structural Homologs

- Double Dynamic Programming combines local sampling with global scoring
- NER score replaces GDT discrete distance cutoffs with smooth Gaussian
- Allows direct optimization of superposition gradient methods

Standley, Toh, Nakamura
Proteins 57, 2004

$$\mathrm{NER} = \Sigma\ e^{-(d/w)^2}$$



ASH:
Alignment of Structural Homologs

- Genetic ASH (GASH)
  - Standley, Toh, Nakamura; BMC Bioinformatics 2005
- Rapid ASH (RASH)
  - Standley, Toh, Nakamura; BMC Bioinformatics 2007
- SeSAW
  - Standley, Yamashita, Kinjo, Toh, Nakamura; Bioinformatics 2010

Can we make a database?

Workflow

1) Choose representative structures from the PDB
2) Slice representatives into domains
3) Align all domains against all domains
4) Build composite chain-level alignments

Seems simple!

Database Workflow

1) ***Choose representative structures*** from the ***PDB***
2) ***Slice*** representatives into ***domains***
3) ***Align*** all domains against all domains
4) Build ***chain-level alignments***
5) ***Add, remove, and modify*** as the PDB is updated

Oh, all of these steps are somewhat hard...

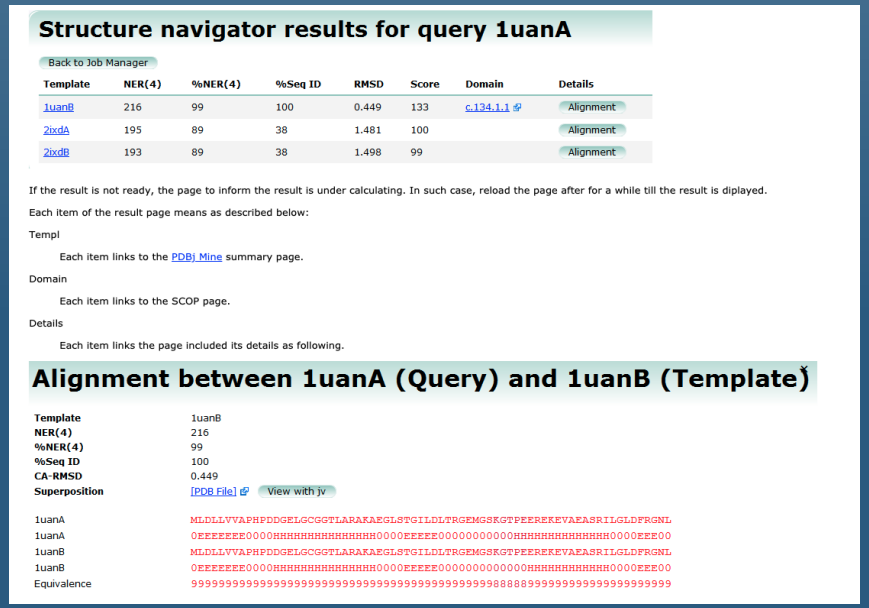
PDB Issues

- Entries added every week...
  - ... and also removed
- Chain ID's sometimes change
- Parsing data is non-trivial
  - Alternate Locations
  - Insertion Codes
- Sequence/structure mismatch

Very difficult to build something on top of the PDB.

Structure Navigator

Structure Navigator was an early prototype of DASH at PDBj

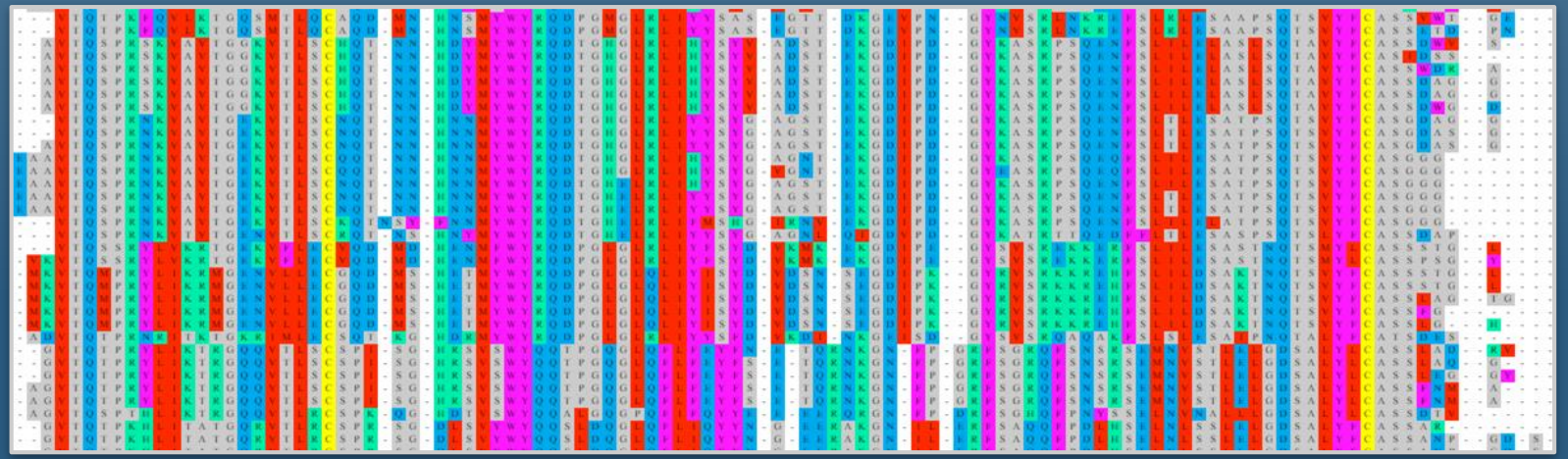


Issues with Structure Navigator

- Complicated front-end and back-end code
- Slow to update
- Difficult to maintain
- Not widely used
- **No clear sense of quality (completeness or accuracy) of the results**

How to assess quality?

Multiple sequence alignment might work



MAFFT + ASH

- ASH residue-wise similarity used as restraints
- MSA benchmark sets can be used to evaluate the quality of the alignments
- PDP-ASH step to map domain alignments back to chain level

- MAFFTash web service released in 2009
- Tested, but not published

MAFFTash Issues

- Required user to manually select PDB templates
- Slow performance if data wasn't cached
- Cache problems:
  - Missing data could have multiple meanings
  - Cached output sometimes contained errors
  - Difficult to remove things from cache
- Many points of failure:
  - Shell scripts (BASH/TCSH), Perl scripts, C code...

2017

- Hired at RIMD to do computer infrastructure
- DASH update failing

Can we fix DASH?

Discoveries

- DASH is split between 2 different databases
  - Cache for MAFFTash
  - Generic domain alignment DB
- Weekly update crashes
- Intertwined with many other databases
- Empty REST interfaces (cannot be called by other software)

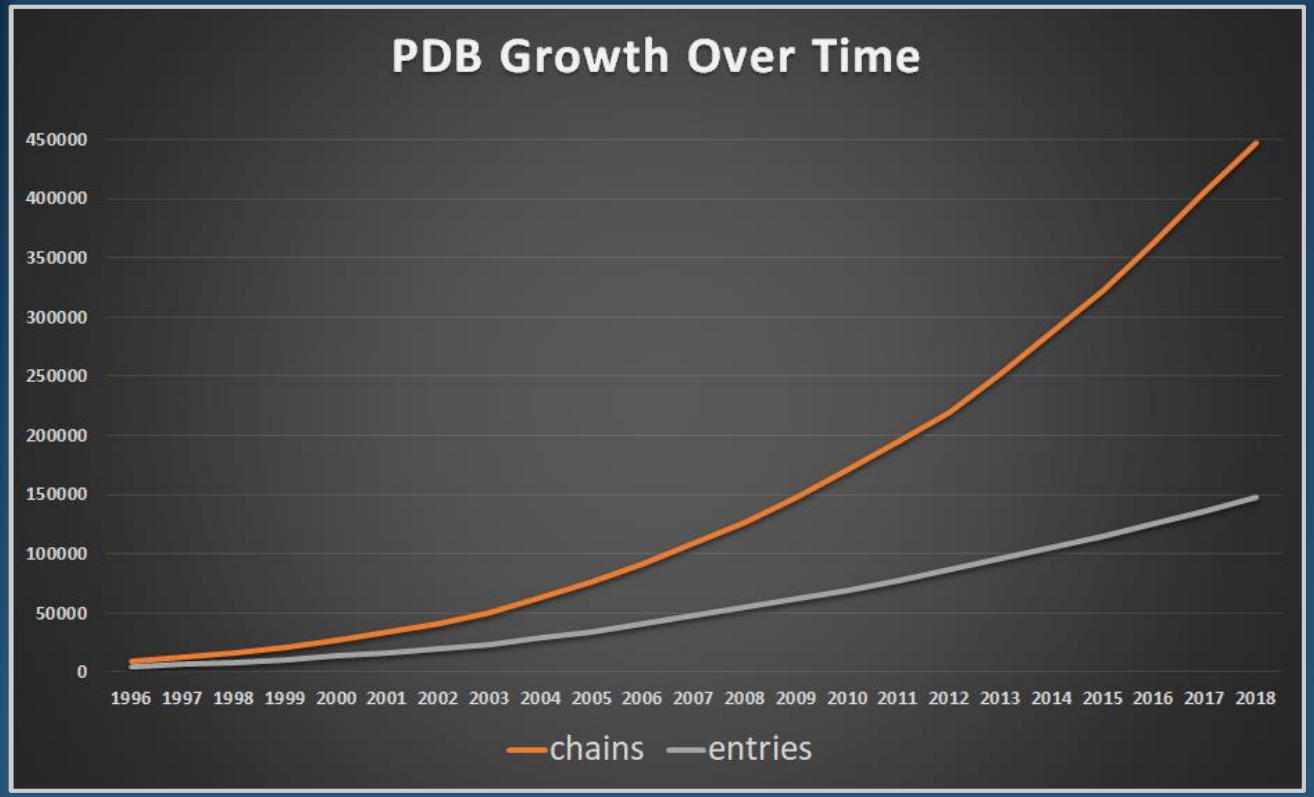
Let's start over!

Database Workflow

1) ***Choose representative structures*** from the ***PDB***
2) ***Slice*** representatives into ***domains***
3) ***Align*** all domains against all domains
4) Build ***chain-level alignments***
5) ***Add, remove, and modify*** as the PDB is updated

Oh, all of these steps are somewhat hard...

PDB Growth



Entry size is also increasing!

2018 Rebuild Plan

- Use Google Cloud for scaling
- Keep ASH as the core
- Update software
- Rewrite old workflow code in Go language
- Version control everything

What is “Go?”

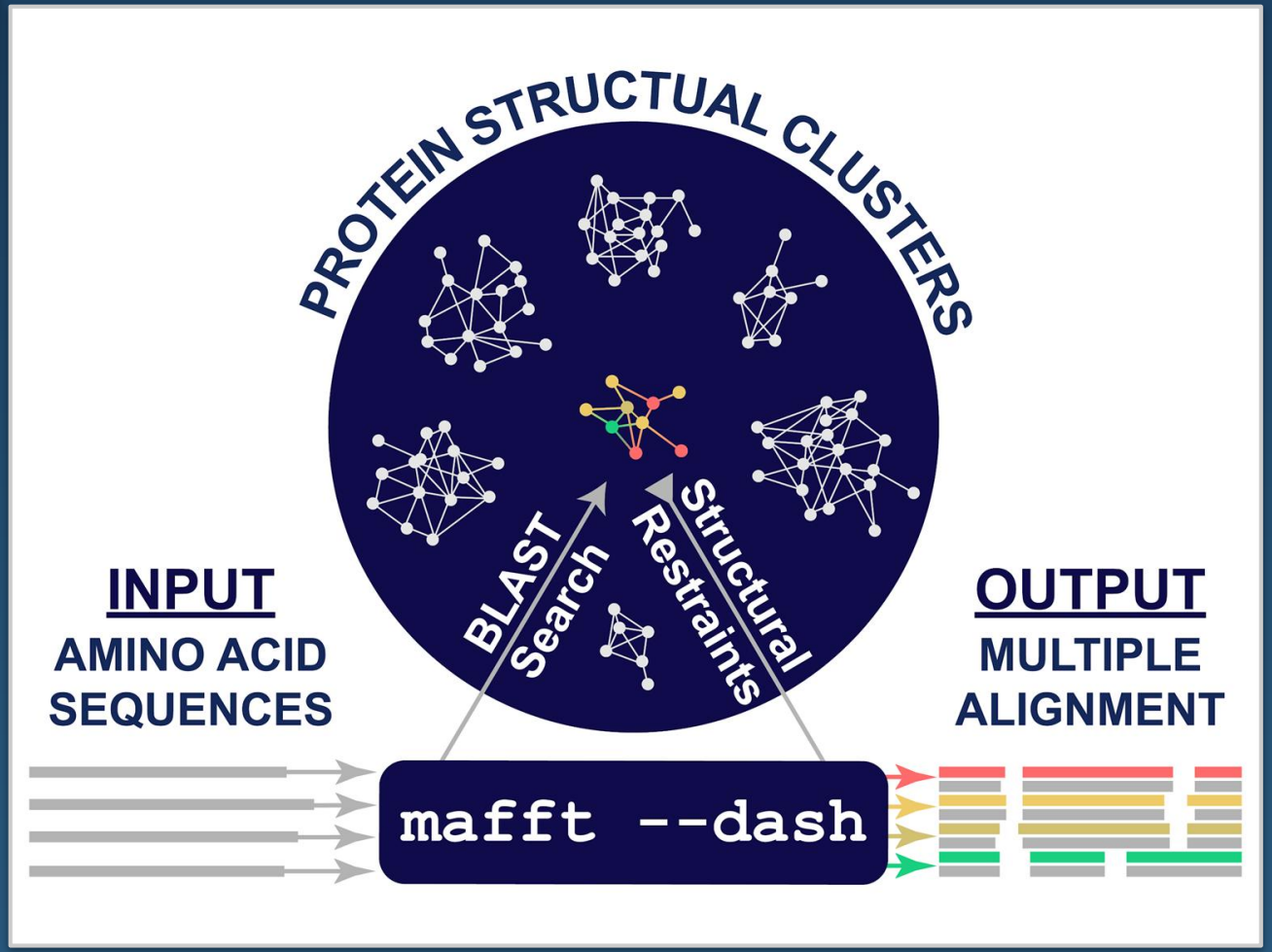
- Static Binaries
- Informative crashes
- Large standard library
  - SQL, HTTP requests, Zip, etc.
- Built-in support for multi-threading
- Faster than Python/Perl
- Safer than C

2018

- January
  - Google Cloud begins
- March
  - 5.5 B domain comparisons calculated on ~2500 CPU cores
- June
  - New PDP-ASH is written
- July
  - First MSA benchmark
- August
  - Alpha version opened
- October
  - First comparison with Promals3D
- December
  - Public beta started

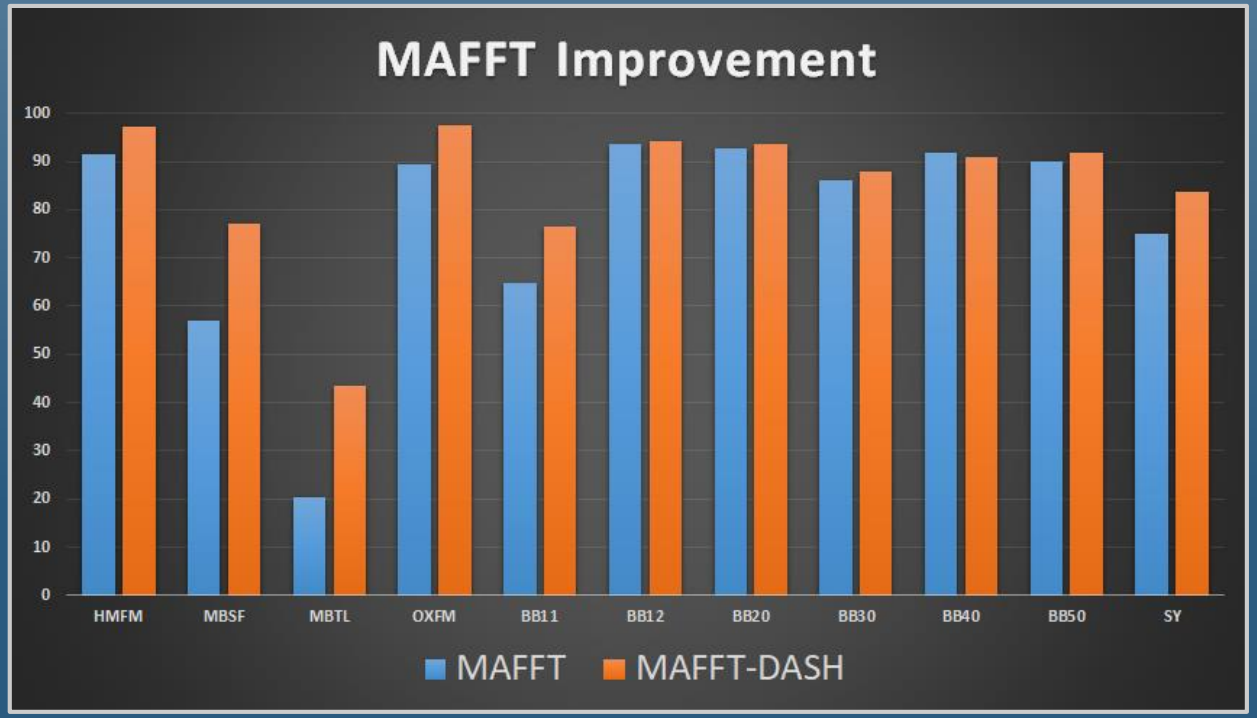
MAFFT-DASH

MAFFT and DASH can talk to each other



MAFFT-DASH Benchmarks

- 10% overall improvement
- >20% improvement on hardest cases
- Much faster than other tested methods



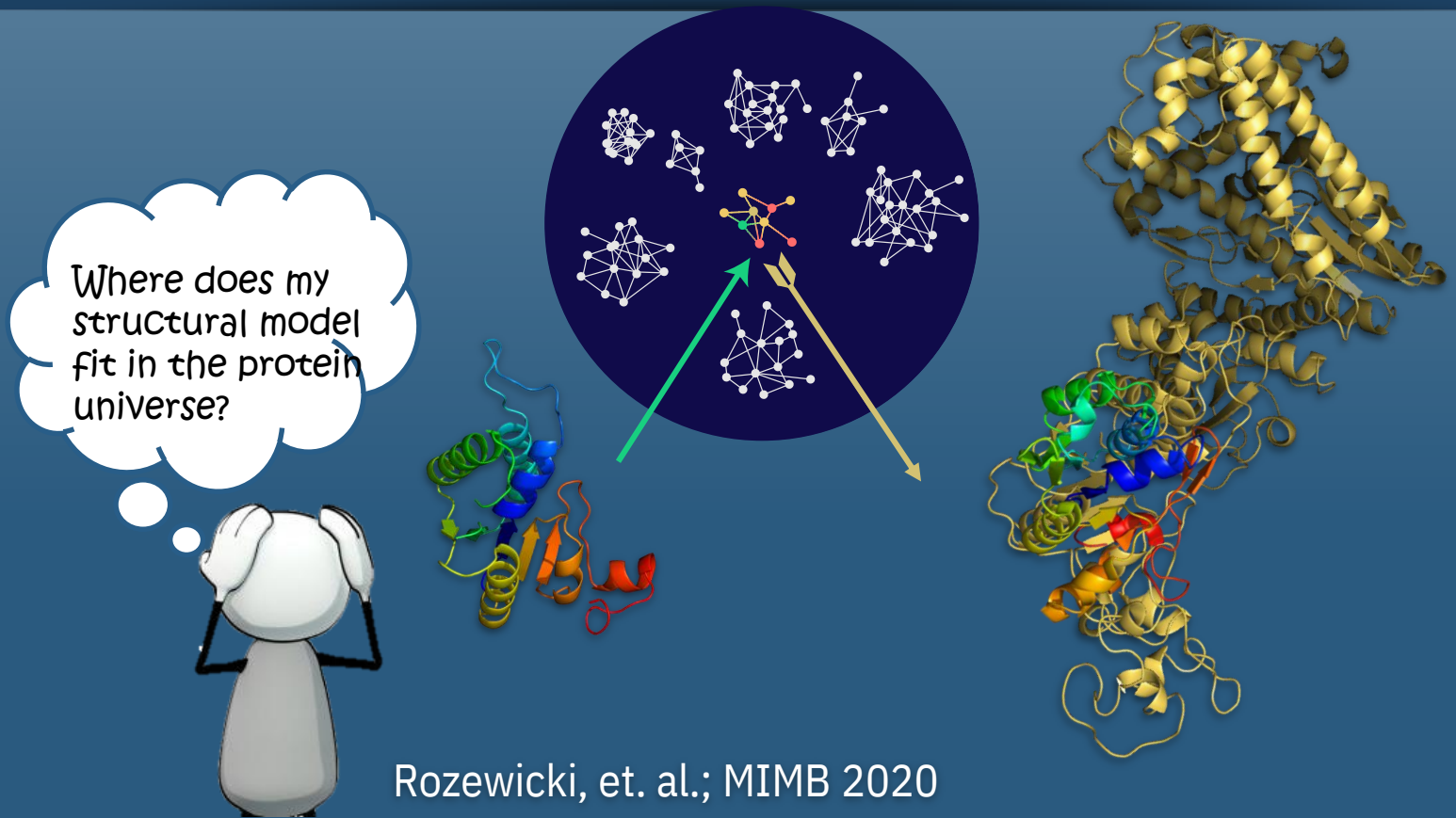
Rozewicki, et. al.; NAR 2019

2019 Improvements

- ASH Rewritten in Go
  - Many bug fixes
  - Multi-core support
  - JSON output
- Rotation Matrices & Translation Vectors
- Search by Structure

Rozewicki, et. al.; MIMB 2020

Search by structure



Rozewicki, et. al.; MIMB 2020