期中考试： group presentation

- 小组分工协作， 一起文献汇报。题目中含 **Blast** 的生物信息工具、算法。
- **max 15分钟汇报，讲明why, how, what。**
- 首页写明组员，后面每页右上角 标明主要贡献人**/speaker**。

Sequence Comparison (Part 2)

Multiple Sequence Alignment and patterns/motifs

曹志伟

From Pair-Wise to Multiple Alignment

- **Pari-wise alignment: two sequences**

```
                10        20        30        40            50
1    VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
     |:| :|: | |:||||    | |:||| |: : :|:| :| |     |: |
2    HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                10        20        30        40        50
Initial Score   =   63  Optimized Score  =   98  Significance =  5.51
Residue Identity =  14%  Matches         =   21  Mismatches  =   22
Gaps            =    2  Conservative Substitutions          =   11
```

- **Multiple sequence alignment – MSA:** Simultaneously align **more than two** sequences.

Evolution in a Nutshell

- Amino acids mutate randomly
- Mutations are then selected (accepted) or counter-selected (rejected)
- If a mutation is harmful, it is counter-selected
  - It disappears from the genome
  - You never see it
- Mutations of important positions (such as active sites) are almost always harmful
- You can recognize important positions because they never mutate!
- MSAs reveal these *conserved* positions

Multiple sequence alignment

What are the conserved regions among a set of sequences over the same alphabet?

12345678	**Position Index**
EMQPILLL	**Sequence 1**
DMLR-LL-	**Sequence 2**
NMK-ILLL	**Sequence 3**
DMPPVLIL	**Sequence 4**

What is MSA: A Definition

- 2D table
- Absolute and relative positions

Residues

Sequences

	1	2	3	4	5	6	7	8	9	10
I	Y	D	G	G	A	V	---	E	A	L
II	Y	D	G	G	---	---	---	E	A	L
III	F	E	G	G	I	L	V	E	A	L
IV	F	D	---	G	I	L	V	Q	A	V
V	Y	E	G	G	A	V	V	Q	A	L

Why multiple sequence alignment

1. Determine whether a group of proteins are related

2. Show regions of conservation within a protein family
   → sequence pattern

3. Structure and function prediction

4. Determine evolutionary history of gene families
   → phylogeny tree

多序列比对的广泛应用

- 获得共享序列
  - 多序列比对所得到的与所有序列距离最近的序列，常用于数据库搜索和芯片探针设计
- 序列测序
  - 消除多个机构测序结果的误差
- 突变分析
  - 同一种系不同个体因突变而产生的差异，SNP
- 种系分析
  - 根据基因组的差异判断种系关系，构造种系树

多序列比对的广泛应用

- 保守区段分析
  - 基因组中功能重要的区段不容易接受突变，重要功能的基因高度保守，基因中的外显子尤其保守，基因调节单元如启动子、增强子等高度保守
- 基因和蛋白质的功能分析
  - 测序后同源分析来推断新基因和蛋白质的功能
- RNA和蛋白质的结构分析
  - 从已知序列的结构推断新序列的结构
- 基因组结构分析
  - 揭示基因组的结构特征和进化特征

MSA: How to Align?

```
Seq1   AGAC
Seq2   AC
Seq3   AG
```

?

```
Seq1   AGAC
Seq2   --AC
```

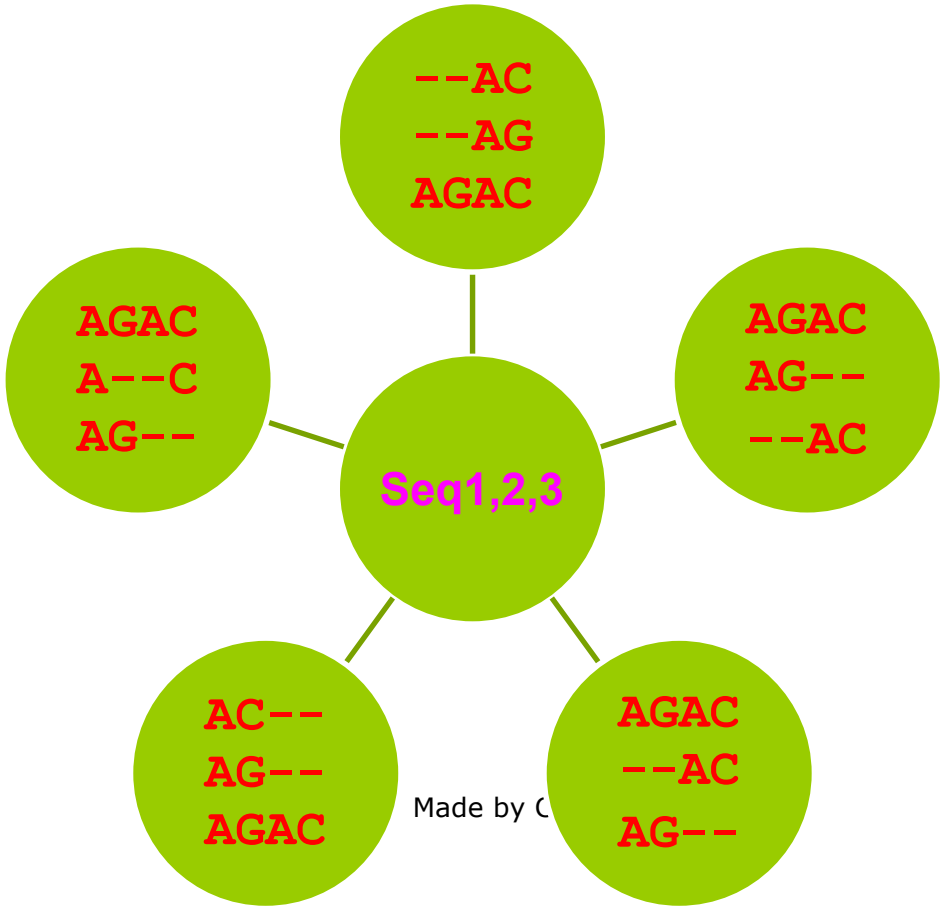
?

```
Seq1  AGAC
Seq3  AG--
```

?

```
Seq2   AC
Seq3   AG
```

MSA: Some Possible Alignments



Made by C

MSA History

- Until 1987 multiple alignments constructed manually from pairwise alignments

- Lipman et al. (1989) pairwise dynamic programming approach applied to multiple sequence alignment - MSA

Commonly Used MSA Methods

1. **Dynamic programming** - extension of pairwise sequence alignment

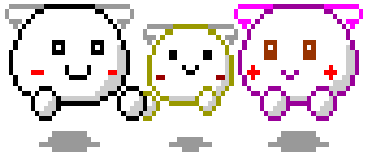
2. **Progressive sequence alignment** - incorporates phylogeny information to guide the alignment process

3. **Iterative sequence alignment** - correct for problems with progressive alignment by repeatedly realigning subgroups of sequence

Progressive Method of MSA

- **Progressive alignment invented in '87 & '88 -** Feng & Doolittle 1987, Higgins and Sharp 1988

- **Based on phylogeny**



How MSA: Progressive method

1 - **Do pairwise alignment of all sequences and calculate distance matrix**

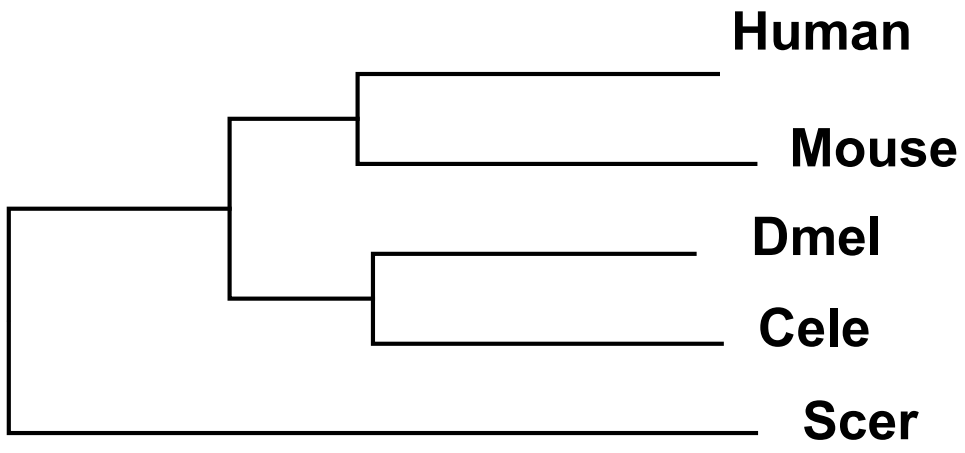
		[1]	[2]	[3]	[4]
Scerevisiae	[1]				
Celegans	[2]	0.640			
Drosophia	[3]	0.634	0.327		
Human	[4]	0.630	0.408	0.420	
Mouse	[5]	0.619	0.405	0.469	0.289

2

1

How MSA: Progressive method

2 - **Create a guide tree based on this pairwise distance matrix**

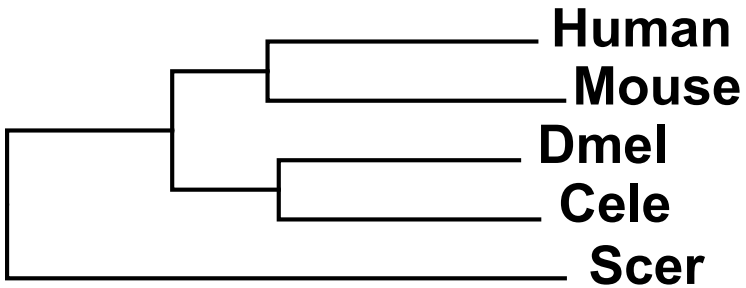


How MSA: Progressive method

3 **- Align progressively following guide tree**

- Start by aligning most closely related pairs of sequences

- Gaps



- At each step align two sequences or one to an existing sub-alignment

Available programs for progressive MSA

- CLUSTAL (Free package):
  - Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73,237-244.
  - http://www.ebi.ac.uk/clustalw/
  - http://clustalw.genome.ad.jp/ (origin 2)

- PILEUP (part of GCG commercial package)
  - http://www.gcg.com

- Others

Example software---ClustalW

Address http://www.clustalw.genome.ad.jp/ Go Links

Clustal W

Multiple Sequence Alignment

CLUSTALW: Multiple Sequence Alignment[help]

General Setting Parameters:

Output Format: CLUSTAL

Pairwise Alignment: **FAST/APPROXIMATE** **SLOW/ACCURATE**

Enter your sequences (with labels) below (copy & paste): **PROTEIN** **DNA**

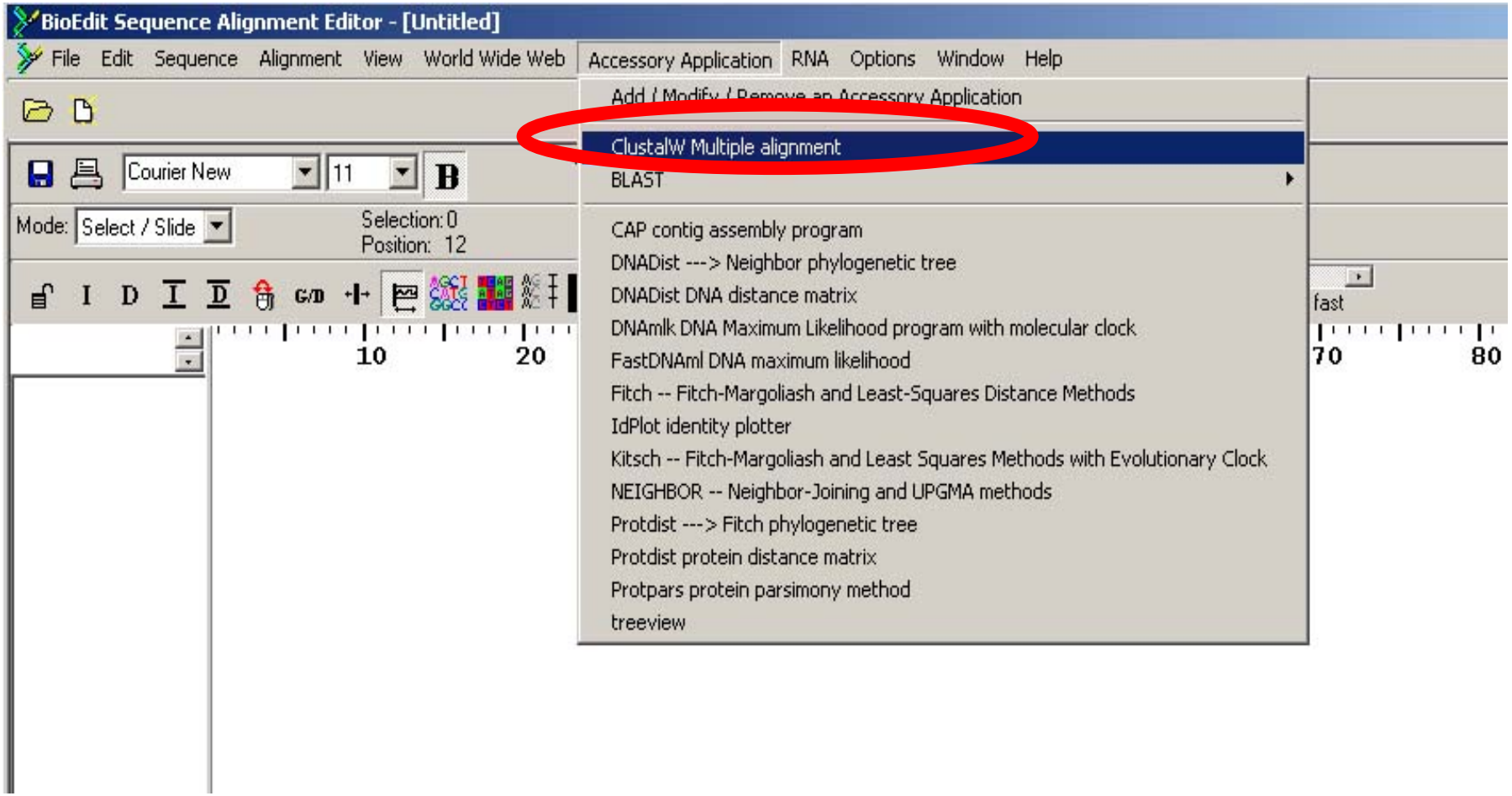
Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

Or give the file name containing your query

Browse...

Execute Multiple Alignment Reset

Example Software---ClustalW (Bioedit)



Steps To Do ClustalW:

Step 1: Prepare the sequences:

- Retrieve sequences

  General considerations:

  - The more the better if family concerned
  - Exclude similar (e.g.>80%) sequences
  - Necessary modification

Steps To Do ClustalW:

Step 2: Input the sequences:

- Put all sequences into one file→ Copy and paste
- Or upload sequences one by one
- Pay attention to sequence format → FASTA format

Steps To Do ClustalW:

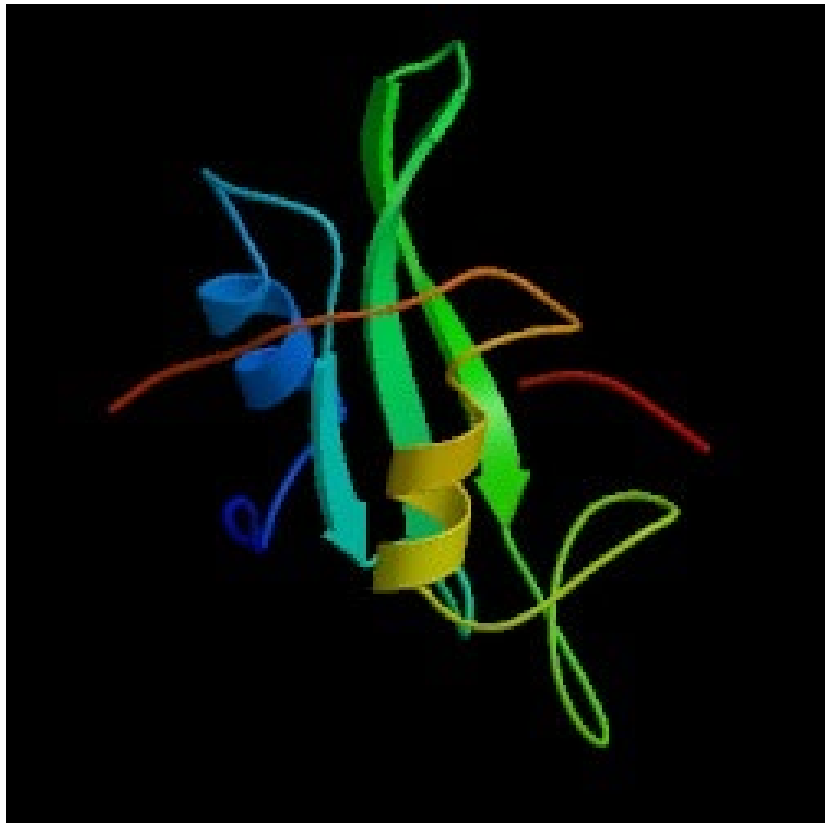
Step 3: Set the parameters:

- Default parameters for protein alignment **General Setting Parameters:**
  - Output Format: CLUSTALW
  - Pairwise Alignment: **FAST/APPROXIMATE**

Example: SH2 domain family

- SH2 domains function as regulatory modules of intracellular signalling cascades

- V-Src Tyrosine Kinase Transforming Protein (Phosphotyrosine Recognition Domain Sh2) Complex With Phosphopeptide A (PDB code 1SHA):



Input Sequences For ClustalW

- >**1SHA-A** V-SRC Tyrosine kinase transforming protein (SH2 domain), from **Rous sarcoma virus**
- >**1A81-A** Chain A, Tandem Sh2 Domain Of The Syk Kinase, from **Homo sapiens**
- >**1JWO-A** Chain A, Sh2 Domain Of The Csk Homologous Kinase Chk, from **Homo sapiens**
- >**1BLJ** Nmr Ensemble Of Blk Sh2 Domain, from **Mus musculus** (house mouse)

Result 1 of ClustalW

CLUSTALW Result

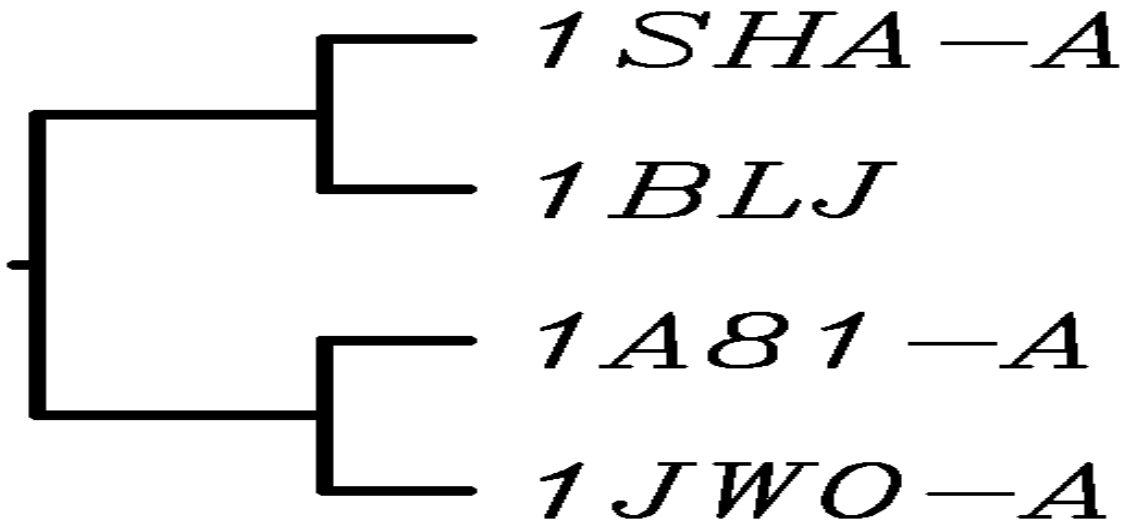
GenomeNet CLUSTALW Server (Kyoto Center) on Thu Oct 31 19:55:40 JST 2002

```
 CLUSTAL W (1.81) Multiple Sequence Alignments


Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: 1SHA-A          104 aa
Sequence 2: 1A81-A          184 aa
Sequence 3: 1JW0-A           97 aa
Sequence 4: 1BLJ            114 aa
Start of Pairwise alignments
Aligning...


Sequences (1:2) Aligned. Score: 29.8077
Sequences (1:3) Aligned. Score: 26.8041
Sequences (1:4) Aligned. Score: 50
Sequences (2:2) Aligned. Score: 100
Sequences (2:3) Aligned. Score: 29.8969
Sequences (2:4) Aligned. Score: 29.8246
Sequences (3:2) Aligned. Score: 29.8969
Sequences (3:3) Aligned. Score: 100
Sequences (3:4) Aligned. Score: 27.8351
Sequences (4:2) Aligned. Score: 29.8246
Sequences (4:3) Aligned. Score: 27.8351
Sequences (4:4) Aligned. Score: 100
```

Result 2 of ClustalW

```
Start of Multiple Alignment
There are 3 groups
Aligning...
Group 1: Sequences:   2      Score:852
Group 2:                     Delayed
Group 3:                     Delayed
Sequence:2     Score:689
Sequence:3     Score:625
Alignment Score 1129
CLUSTAL-Alignment file created  [clustalw.aln]
CLUSTAL W (1.81) multiple sequence alignment


1SHA-A          ------------------------------------------------------------
1BLJ            ------------------------------------------------------------
1A81-A          GRTHASPADLCHYHSQESDGLVCLLKKPFNRPQGVQPKTGPFEDLKENLIREYVKQTWNL
1JWO-A          ------------------------------------------------------------


1SHA-A          -------------------------QAEEWYFGKITRRESERLLLNPENPRGTFLVRESE
1BLJ            ----------------GSVAPVETLEVEKWFFRTISRKDAERQLLAPMNKAGSFLIRESE
1A81-A          QGQALEQAIISQKPQLEKLIATTAHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRARD
1JWO-A          ------------------------LSLMPWFHGKISGQEAVQQLQPPED--GLFLVRESA
                                          .   *:. .*: .:: : :    .  * **:*

1SHA-A          TTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFSSLQQLVAYYSKHADGLC
1BLJ            SNKGAFSLSVKDIT-TQGEVVKHYKIRSLDNGGYYISPRITFPTLQALVQHYSKKGDGLC
1A81-A          NN-GSYALCLLHEG-----KVLHYRIDKDKTGKLSIPEGKKFDTLWQLVEHYSYKADGLL
1JWO-A          RHPGDYVLCVSFGR-----DVIHYRVLHRD-GHLTIDEAVFFCNLMDMVEHYSKDKGAIC
                   * : *.:          * **::   . *   *     * .*  :* :** . ..:

1SHA-A          HRLTNVCPT--
1BLJ            QKLTLPCVNLA
1A81-A          RVLTVPCQKI-
1JWO-A          TKLVRPKRK--
                  *.    .
(
1SHA-A:0.25262,
(
1A81-A:0.33803,
1JWO-A:0.36300)
```

Result 3 of ClustalW: N-J tree



Interpret ClustalW results

- Three characters are used in the results 2:

1. '*' indicates positions which have a single, fully conserved residue
2. ':' indicates that 'strongly' conserved groups
3. '.' indicates that 'weakerly' conserved groups

```
sp|P13860|GUX1_PHACH     SHSSSSTPPTQPTGVTVPQWGQCGGIG---YTGSTTCASPYTCHVLNPYYSQCY--
sp|P38676|GUX1_MEUCR     AKPSSTSTASNPSGTGAAHWAQCGGIG---FSGPTTCPEPYTCAKDHDIYSQCV--
sp|P00725|GUX1_TRIRE     TTRRPATTTGSSPGPTQSHYGQCGGIG---YSGPTVCASGTTCQVLNPYYSQCL--
sp|P45699|GUMK_FUSOX     RGSCPAKTDATAKASVVPAYYQCGGSKSAYPNGNLACATGSKCVKQNEYYSQCVPN
                            .:..  . .  . : ****     .* .*.  .*  :  ****
```

Interpret ClustalW results

- Insertion and deletion, gap

```
sp|P13860|GUX1_PHACH     SHSSSSTPPTQPTGVTVPQWQQCGGLG---YTGSTTCASPYTCHVLNPYYSQCY--
sp|P38676|GUX1_MEUCR     AKPSSTSTASNPSGTGAAHWAQCGGIG---FSGPTTCPEPYTCAKDHDIYSQCV--
sp|P00725|GUX1_TRIRE     TTRRPATTTGSSPGPTQSHYQQCGGIG---YSGPTVCASGTTCQVLNPYYSQCL--
sp|P45699|GUNK_FUSOX     RGSCPAKTDATAKASVVPAYQQCGGKSAYPNGNLACATGSKCVKQNEYYSQCVPN
                            .:..  . .  . : ****     .* .*.  .*  :  ****
```

Consensus

.....QCGG.....G.....C

.....C.....YSQC.....

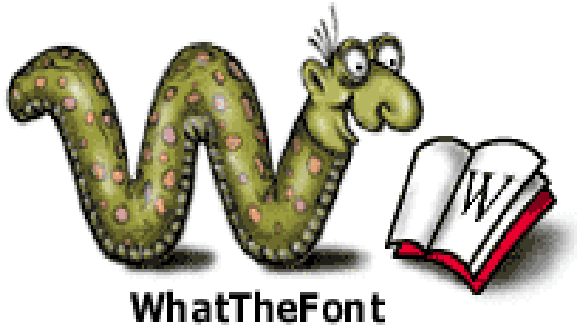
- Consensus sequence →Sequence Pattern

Notes on how to use ClustalW

- Remove signal peptide before alignment, try to compare homologous portion
- Sequence containing a repetitive element (such as a domain)
- **Heuristic algorithm: not guaranteed for perfect alignment**

Notes on how to use ClustalW

- Mobilize your biological knowledge, check the alignment and recheck the alignment
- Manually re-align your sequences if it's bad



MSA software

CLUSTALW或 CLUSTALX(图形化界面)	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/ ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/
T-COFFEE	http://www.ch.embnet.org/software/TCoffee.html
MSA	ftp://fastlink.nih.gov/pub/msa/
DIALIGN	http://www.gsf.de/biodv/dialign.html
DCA	http://bibiserv.techfak.uni-bielefeld.de/dca/
MultAlin	http://protein.toulouse.inra/multalin.html
PILEUP	https://www.sacs.ucsf.edu/secure/cgi-bin/pileup.pl
HMMER	http://hmmer.wustl.edu/
MACAW	ftp://ncbi.nlm.nih.gov/pub/macaw
SAM	http://www.cse.ucsc.edu/research/compbio/sam.html
Vector NTi	http://biocore.unl.edu/coreweb/VectorNTIFrame.htm
Bioedit	http://www.mbio.ncsu.edu/BioEdit/bioedit.html
ProAlign	http://evol-linux1.ulb.ac.be/ueg/ProAlign/
QAlign	http://www.ridom-rdna.de/qalign/

Application of MSA

Example: Drug discovery for SARS

http:/ / www.sciencexpress.org / 13 May 2003 / Page 1/ 10.1126/science.108565

- Coronaviruses are positive-stranded RNA viruses
- Largest viral RNA genomes known to date
- M protein: critical proteinase to produce replicase

Example: Drug discovery for SARS

- Sequence→ structure→ function
  - ✓ Human coronavirus 229E: HCoV;
  - ✓ Porcine transmissible gastroenteritis virus: TGEV;
  - Mouse hepatitis virus: MHV;
  - Bovine coronavirus: BCoV;
  - SARS-associated coronavirus: SARS-CoV;
  - Avian infectious bronchitisvirus: IBV.

Multiple sequence alignment: SARS

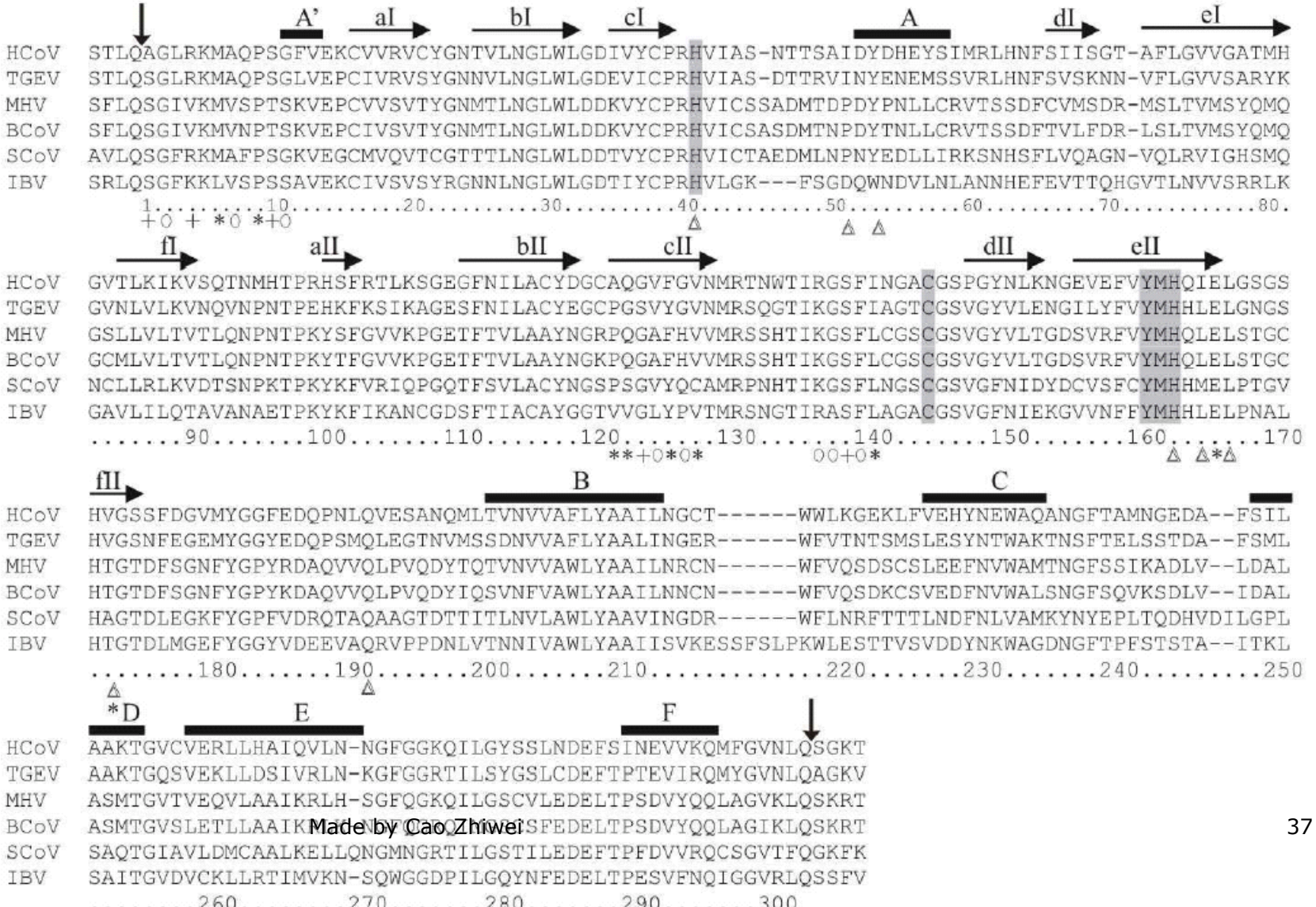
- Which family does SARS virus belong to?
- Sars virus genome (blast result)

  Top hits:

  1. Mouse hepatitis V virus
  2. Avian infectious bronchitis virus
  3. Turkey coronavirus
  4. Porcine epidemic diarrhea virus
  5. …….

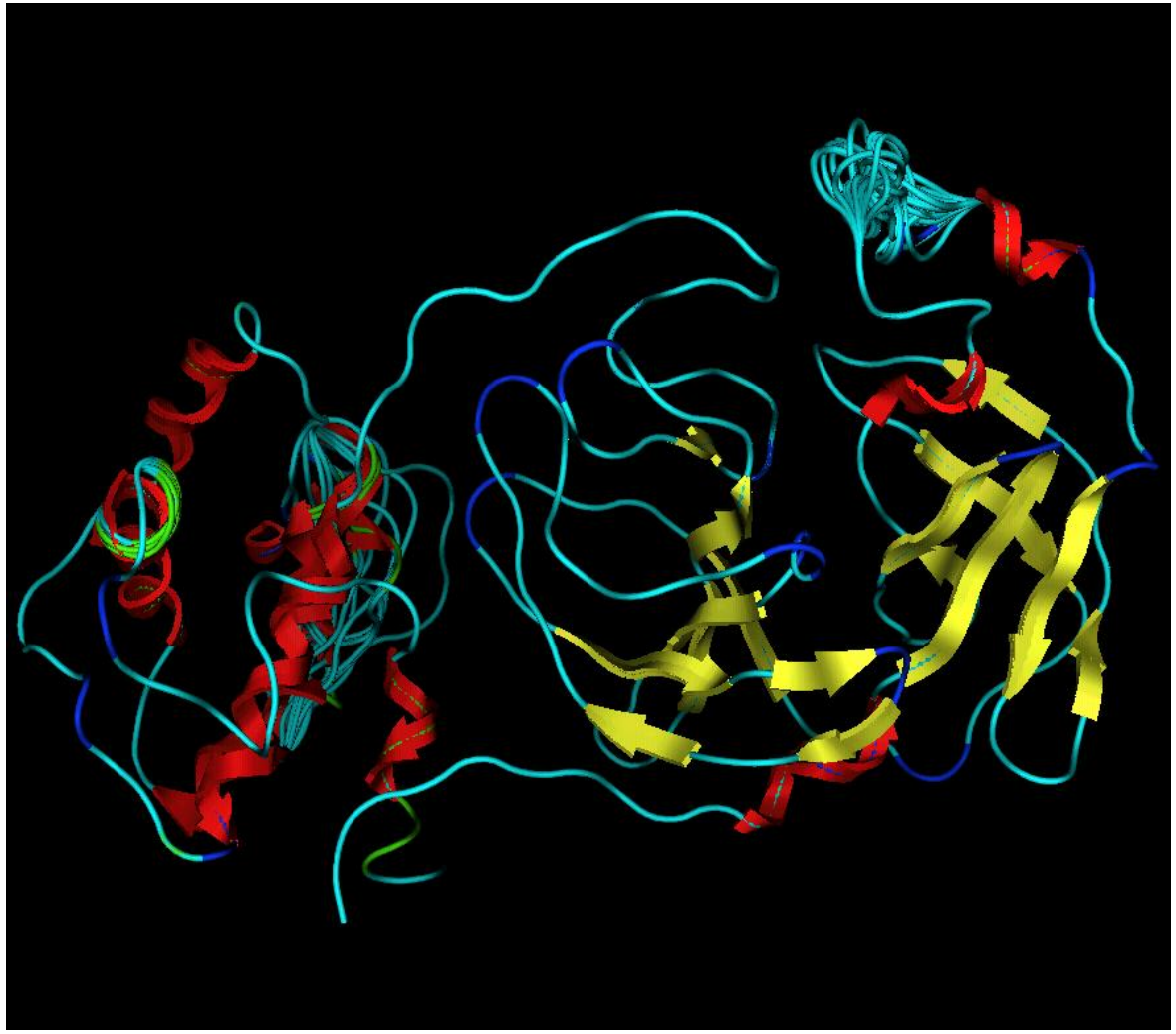
Application of MSA

- Example: Drug discovery for SARS

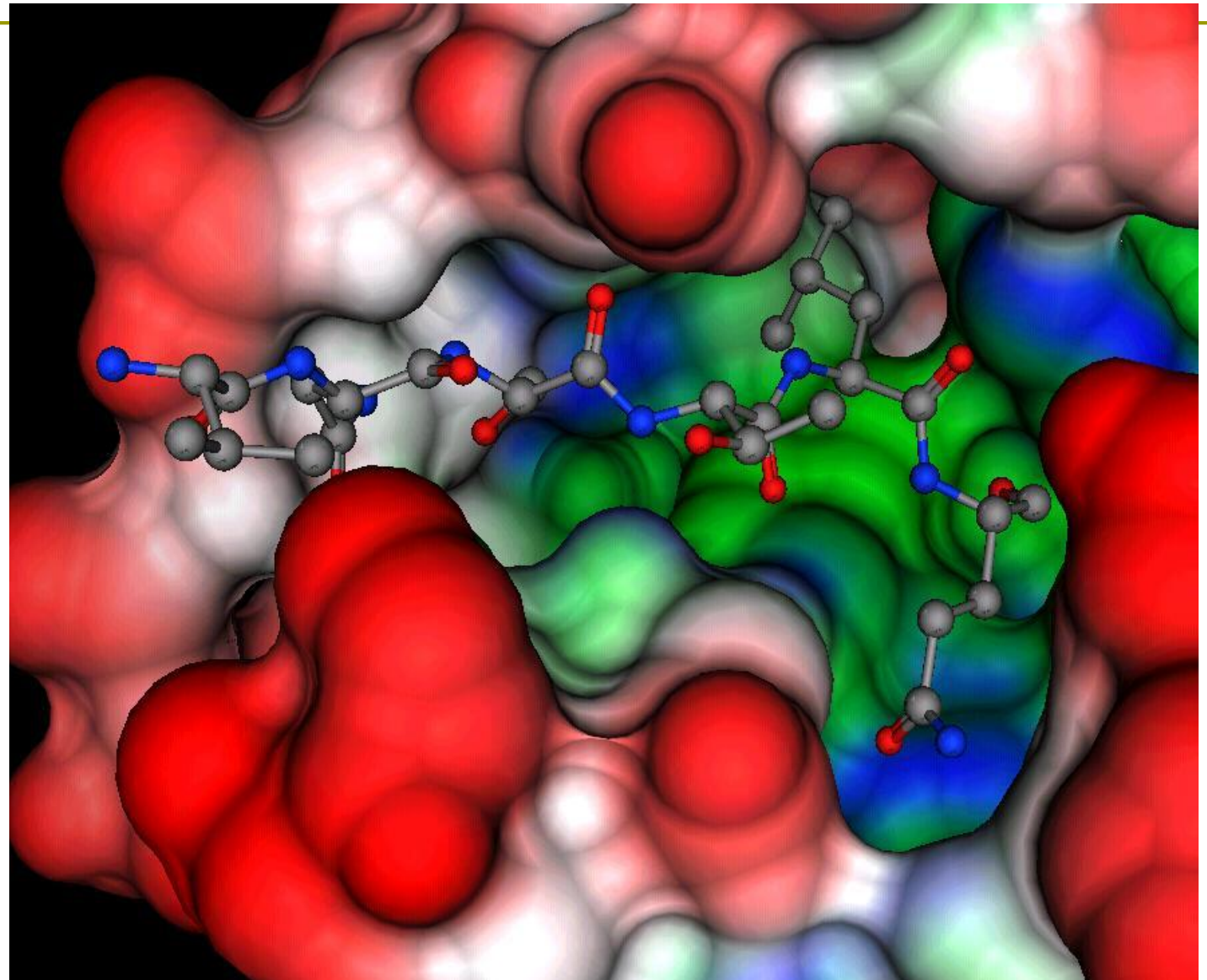


Final model of M protein from SARS virus

http://biocomp.chem.unb.ca:8080/GD/SARS/



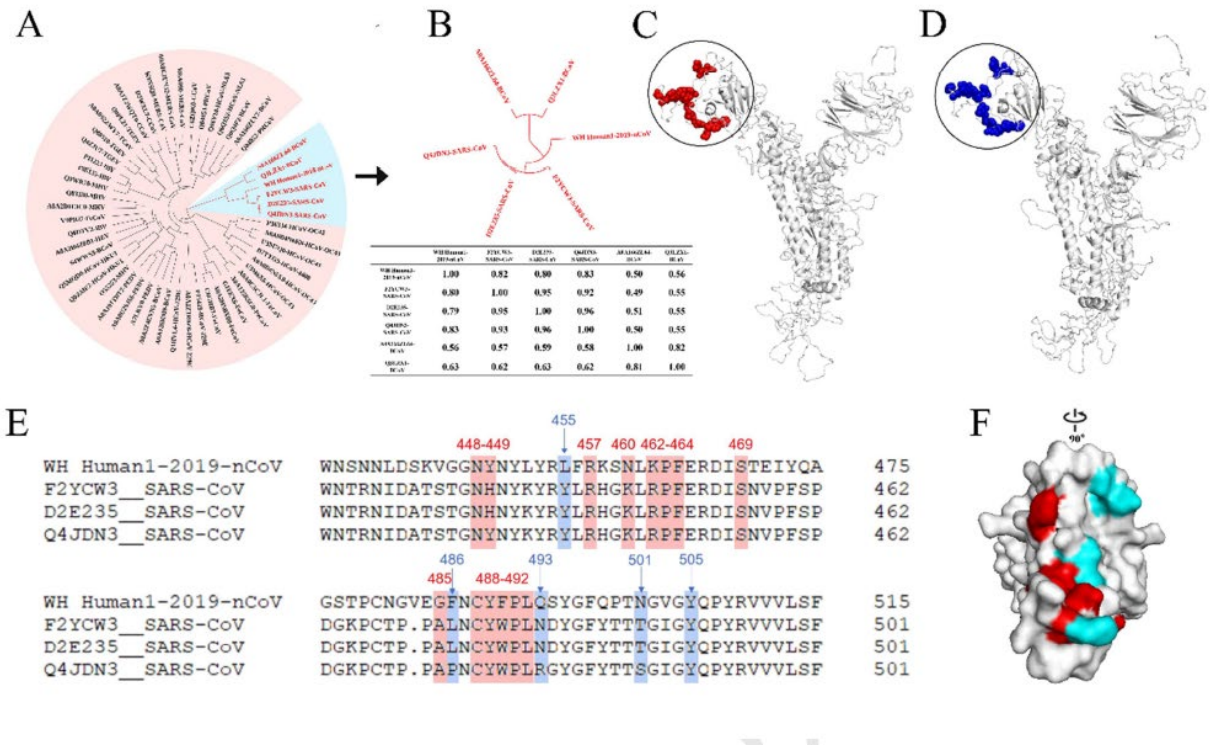
Drug – SARS protein interaction



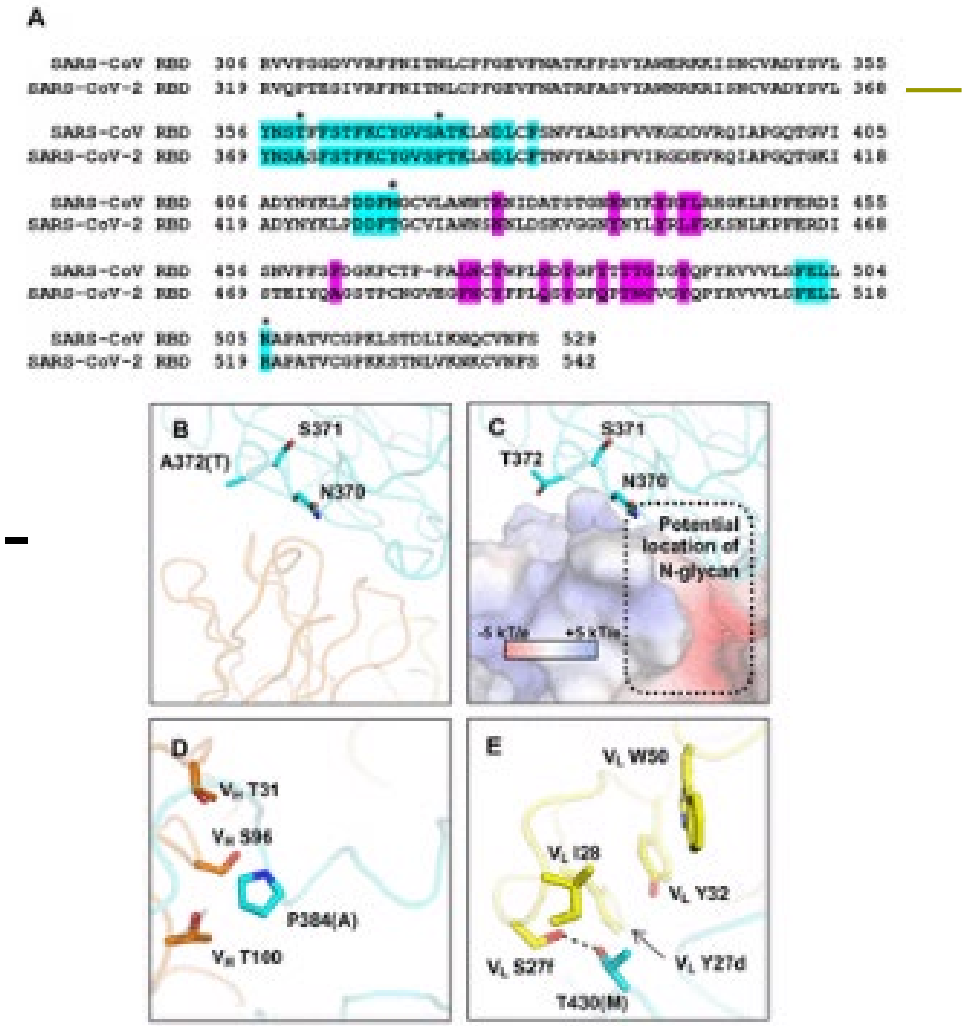
SARS-CoV-2: vaccine

- highly similar epitope was identified between the 2019-nCoV and SARS virus, in the region of the binding site of the S proteins to the human ACE2 receptor.

- Journal of Genetics and Genomics, 2020,47(2):115-11



- CR3022 targets a highly conserved epitope, distal from the receptor-binding site, that enables cross-reactive binding between SARS-CoV-2 and SARS-CoV.
- revealed a conserved, but cryptic epitope shared between SARS-CoV-2 and SARS-CoV
- *Science,* 03 Apr 2020: eabb7269



- try MSA
  - https://www.ebi.ac.uk/Tools/msa/
  - http://www.clustal.org/

- Next: patterns and motif