#AprendeBioinformáticaEnCasa

Multiple Sequence Alignments and Jalview

@ Your , 25/05/2020

Rafael Torres-Perez
rafael.torres@cnb.csic.es

BioinfoGP
Bioinformatics for Genomics and Proteomics

Similarity

	Bilabial		Labio-dental		Inter-dental		Dental		Alveolar		Palatal		Velar	
	SOR	SON	SOR	SON	SOR	SON	SOR	SON	SOR	SON	SOR	SON	SOR	SON
Oclusiva	p	b					t	d					k	g
Fricativa		β	f		θ	ð			s	s̬		ʝ	x	ɣ
Africada											ʧ	ʤ		
Nasal		m		ɱ		n̪		n̪		n		ɲ ɲ		ŋ
Lateral						l̪		l̪		l		ɭ ʎ		
Vibrante simple										ɾ				
Vibrante múltiple										r				

Alignments

- Refers to:
    . the Process (algorithm) and,
    . the Representation of its result

- Match, mismatch, gap [open, extension, penalties], scoring

- Conservation, consensus, occupancy

- Similarity, Identity

- Families, domains

- Function

Alignments

- Utility:

- Relations between sequences: functionality, philogeny, evolutionary history

- Variation: polimorphism, pathogenicity.

- MSA is based on Pairwise alignment (PA)

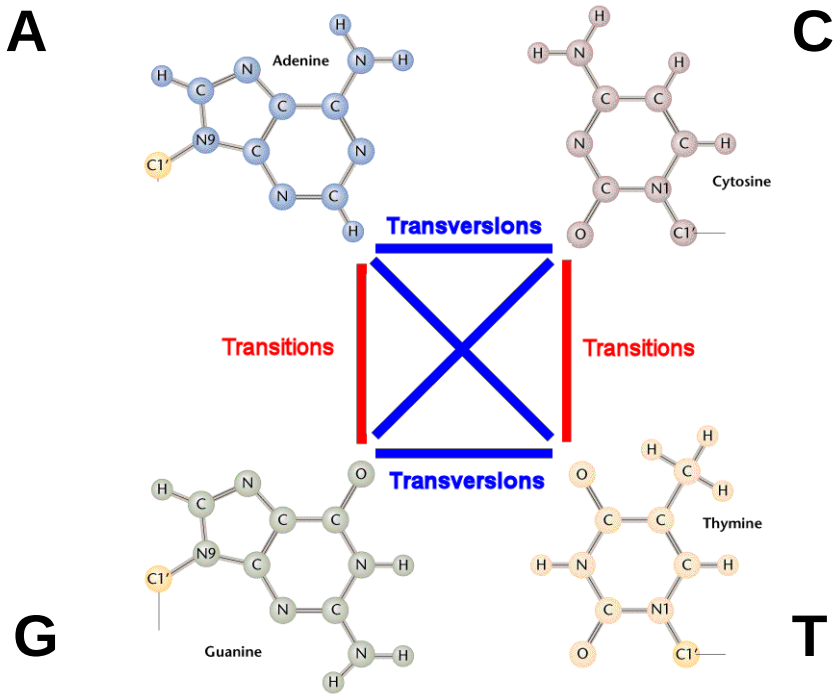
. but...“*PAs whispers… M(S)As shouts out loud*” (Hubbard et al., 1996)

- Manual MSA is tedious

- Sequences must be of similar length (except if looking for domain sharing) and composition.

- There is no unique MSA result. Scoring algorithms provide methods to compare alignments.

Scoring matrices for DNA sequences



GAATC

CATAC

-5 + 10 + -5 + -5 + 10 = 5

A hypothetical substitution matrix:

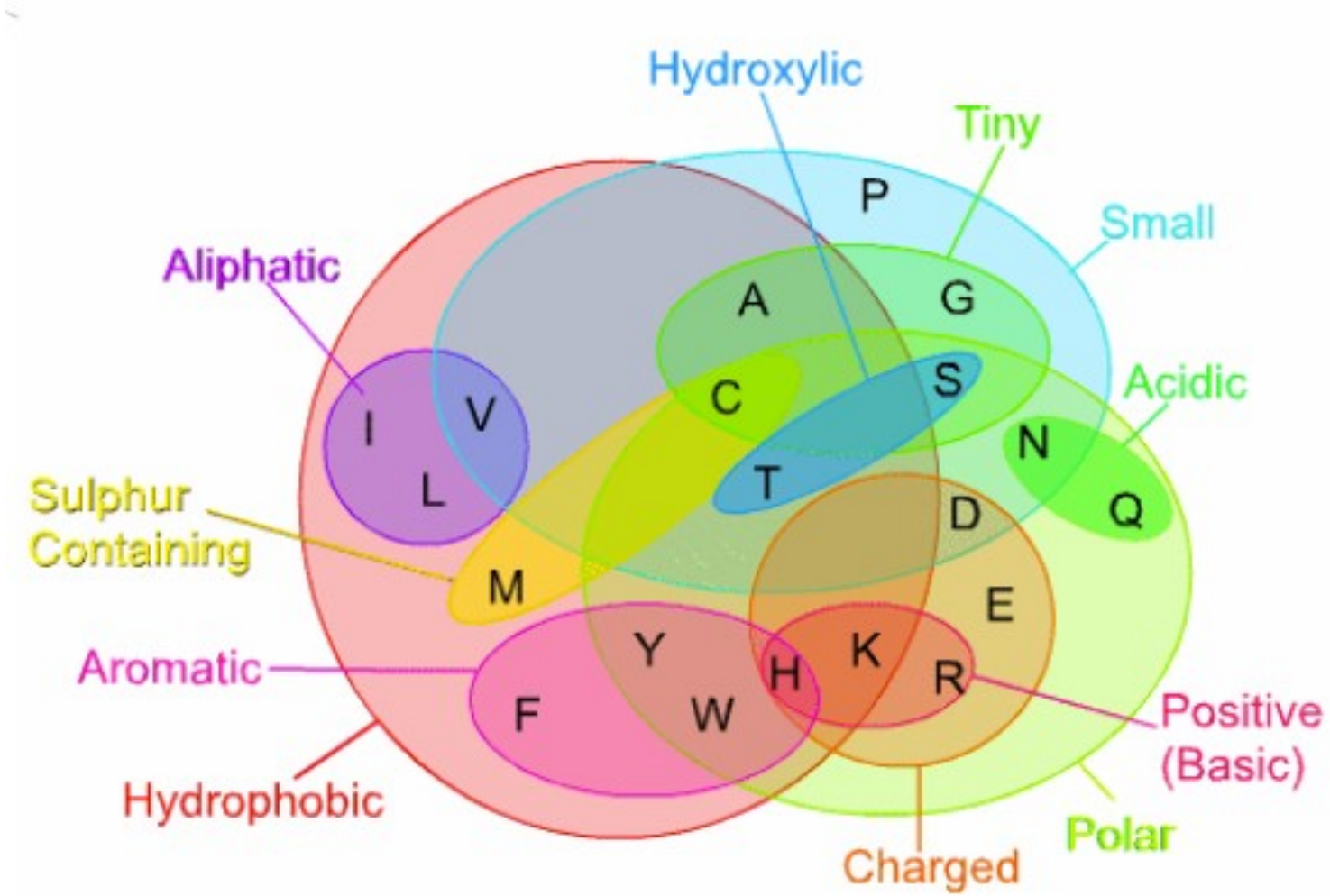
	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

DNA MSA for evolutionary history, Protein MSA for functional relationship

Second base

First base	U	C	A	G	Third base
U	UUU, UUC — Phenylalanine F UUA, UUG — Leucine L	UCU, UCC, UCA, UCG — Serine S	UAU, UAC — Tyrosine Y UAA — Stop codon UAG — Stop codon	UGU, UGC — Cysteine C UGA — Stop codon UGG — Tryptophan W	U C A G
C	CUU, CUC, CUA, CUG — Leucine L	CCU, CCC, CCA, CCG — Proline P	CAU, CAC — Histidine H CAA, CAG — Glutamine Q	CGU, CGC, CGA, CGG — Arginine R	U C A G
A	AUU, AUC, AUA — Isoleucine I AUG — Methionine start codon M	ACU, ACC, ACA, ACG — Threonine T	AAU, AAC — Asparagine N AAA, AAG — Lysine K	AGU, AGC — Serine S AGA, AGG — Arginine R	U C A G
G	GUU, GUC, GUA, GUG — Valine V	GCU, GCC, GCA, GCG — Alanine A	GAU, GAC — Aspartic acid D GAA, GAG — Glutamic acid E	GGU, GGC, GGA, GGG — Glycine G	U C A G

Simililarity (: / .) in aminocids based on physicochemical properties...



…but, **in fact, the used substitution matrices** are made by analyzing the **observed frequencies** of substitutions in families of **similar** proteins.

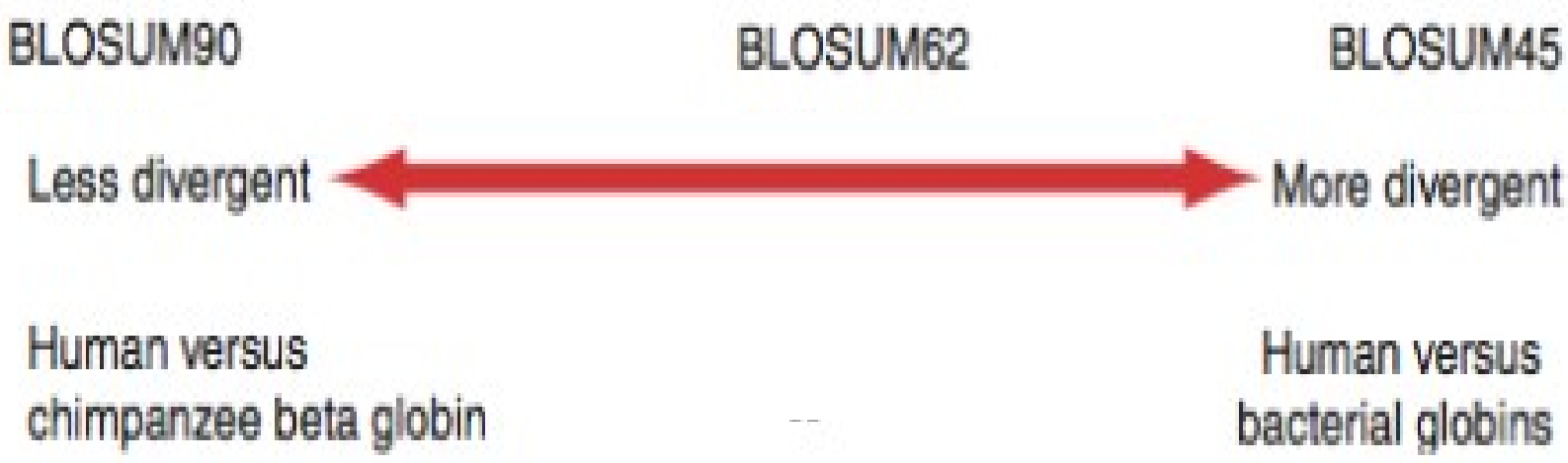
BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

KKLA
RRIA
Score: 10

Which matrix to choose?

It depends on how distant are your sequences each other...



Global and Local Alignment

Global Alignment

- Attemps to match as much of the sequence (heat-to-tail) as possible
- Recommended for suspected similar sequences in composition and length
- Main algorithm: Needleman-Wunsch (https://www.ebi.ac.uk/Tools/psa/emboss_needle/)

Local Alignment

- Try to find the regions with highest density of matches (best matching subsequences)
- Suitable for aligning more divergent or distant related sequences (often different lengths)
- Main algorithm: Smith-Waterman (https://www.ebi.ac.uk/Tools/psa/emboss_water/)

Both of them, Needleman-Wunsch (NW) and Smith-Waterman (SW) use a technique named Dynamic Programming, that assures to get highest-ranked alignments

```
Global FTFTALILLAVAV
       F--TAL-LLA-AV

Local  FTFTALILL-AVAV
       --FTAL-LLAAV--
```

Computational Complexity (~time of processing)

Pairwise alignment: reaching an optimal solution:

ATGGCCCTGTGGATGCGCCT
CTGGTGCTGAGGTTGCGCTT

L=20

Exhaustive (brute-force) search: 3^L^2 = 3^20^2 = **7·10^190**

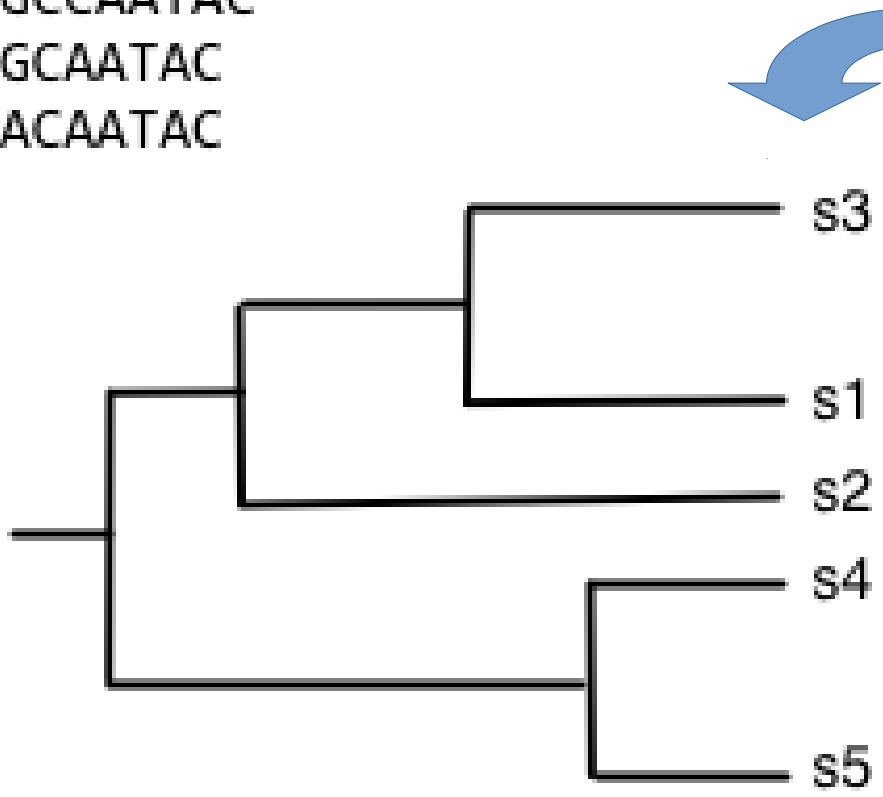
Dynamic programming (DP): 3·L^2 = 3·20^2 = **1200**

MSA scales one dimension (3D scoring matrix) and even the DP is not enough to reach an optimal alignment in a reasonable time

MSA uses **heuristics** (sometimes combined with DP) to reach an <u>approximate-to-optimal</u> (good) solution.

Heuristics (Progressive method)

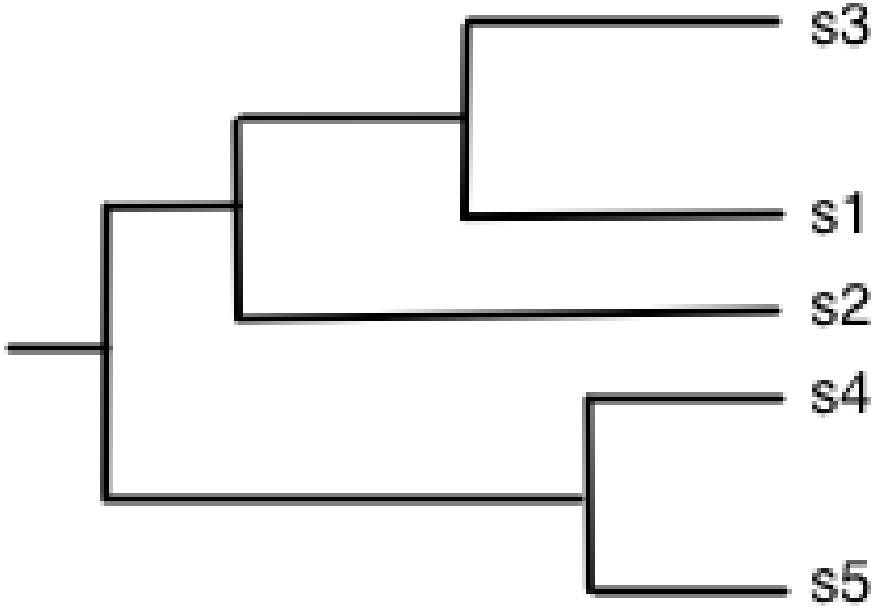
```
s1: ACCGTGAAGCCAATAC
s2: ACGTGCAACCATTAC
s3: AGCGTGCAGCCAATAC
s4: AGGGTGCCGCAATAC
s5: AGGGTGCCACAATAC
```



- best scored pairwise alignments

Heuristics (Progressive method)

```
s1: ACCGTGAAGCCAATAC
s2: ACGTGCAACCATTAC
s3: AGCGTGCAGCCAATAC
s4: AGGGTGCCGCAATAC
s5: AGGGTGCCACAATAC
```



- k-mers (“words”) shared

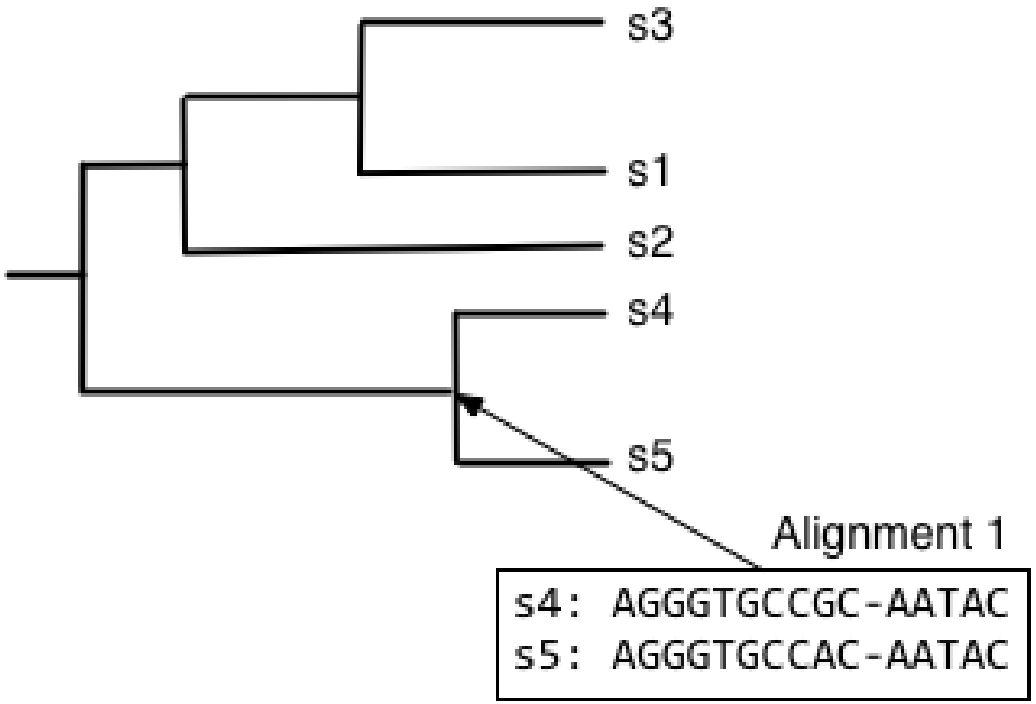
K-mers s1

```
ACC
CCG
CGG
GGT
GTG
TGA
GAC
ACC
CCA
CAG
AGT
GTT
TTG
TGA
GAC
ACC
CCA
CAG
AGT
GTA
```

```
ACC
GGT
GAC
CAG
TTG
ACC
AGT
```

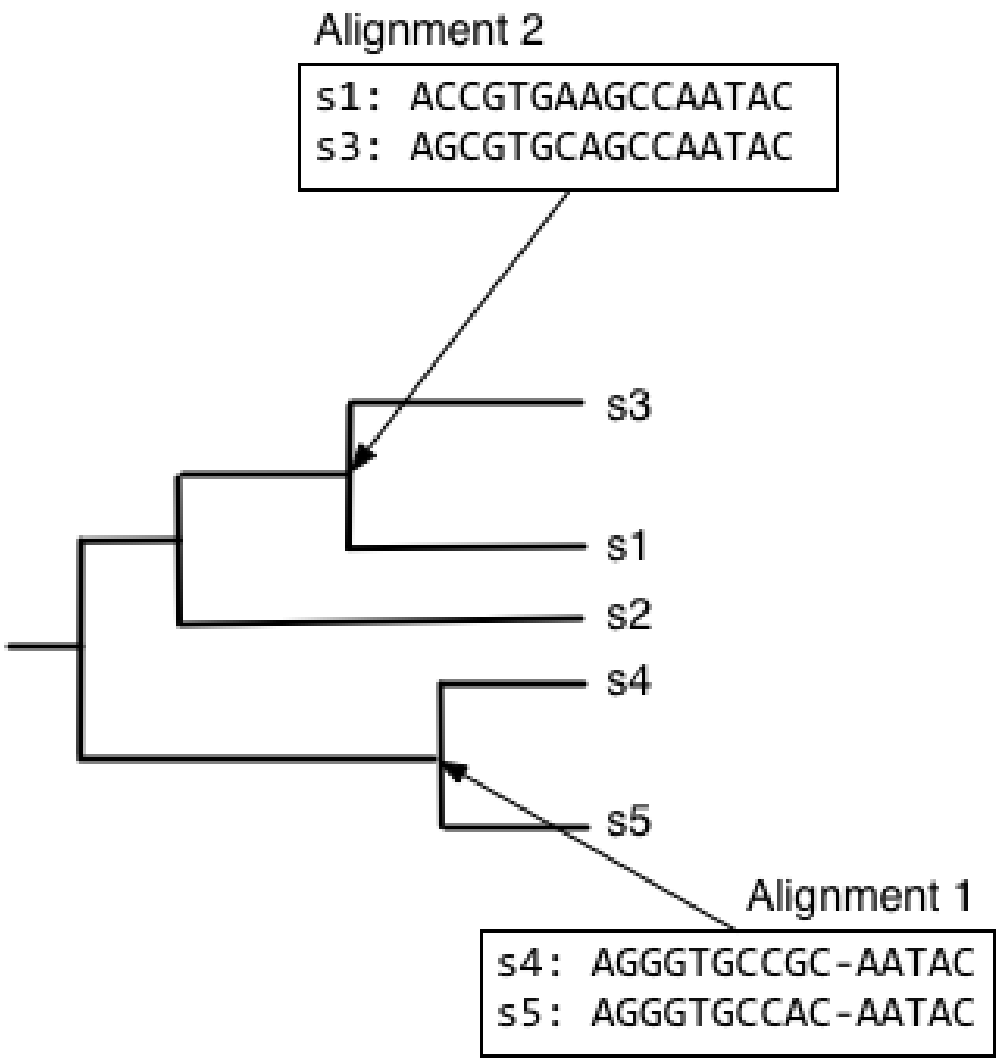
Heuristics (Progressive method)

```
s1: ACCGTGAAGCCAATAC
s2: ACGTGCAACCATTAC
s3: AGCGTGCAGCCAATAC
s4: AGGGTGCCGCAATAC
s5: AGGGTGCCACAATAC
```



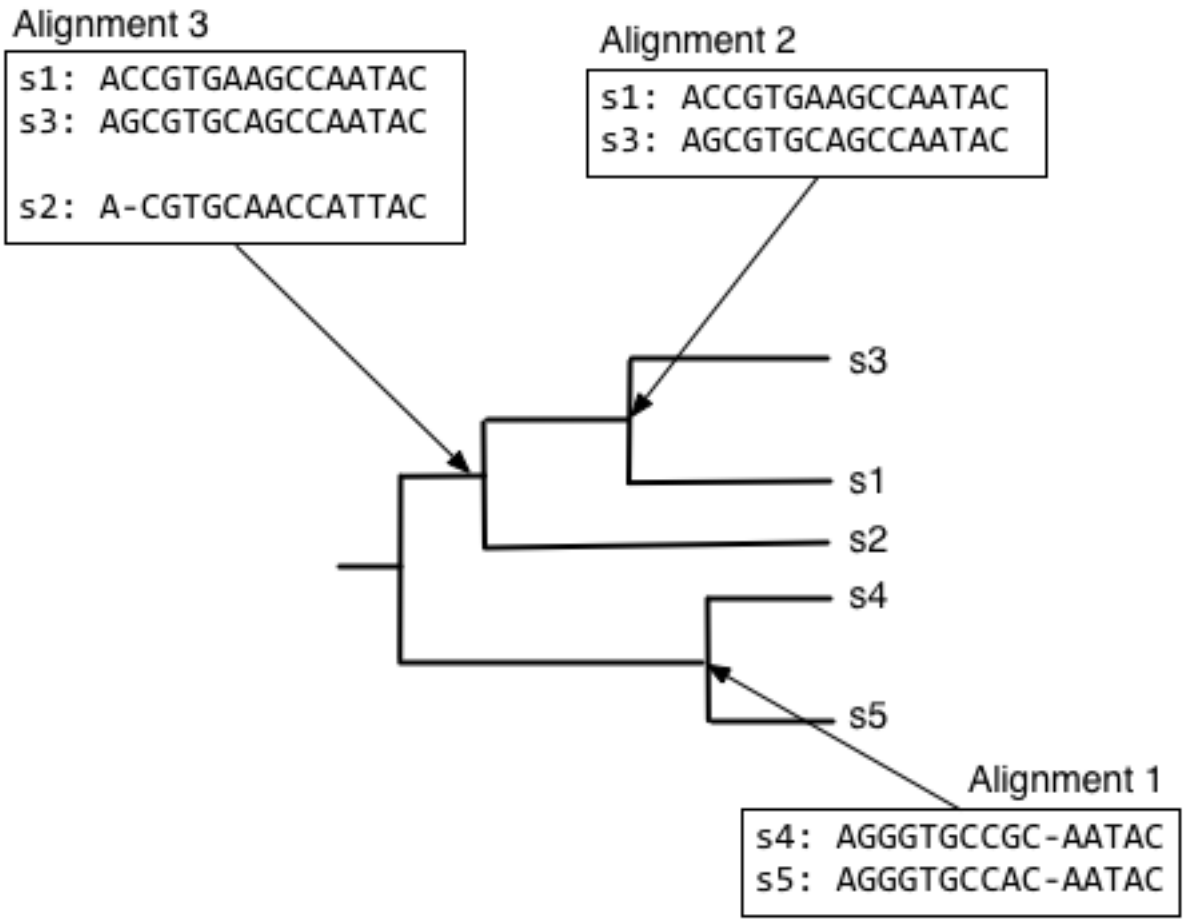
Heuristics (Progressive method)

```
s1: ACCGTGAAGCCAATAC
s2: ACGTGCAACCATTAC
s3: AGCGTGCAGCCAATAC
s4: AGGGTGCCGCAATAC
s5: AGGGTGCCACAATAC
```



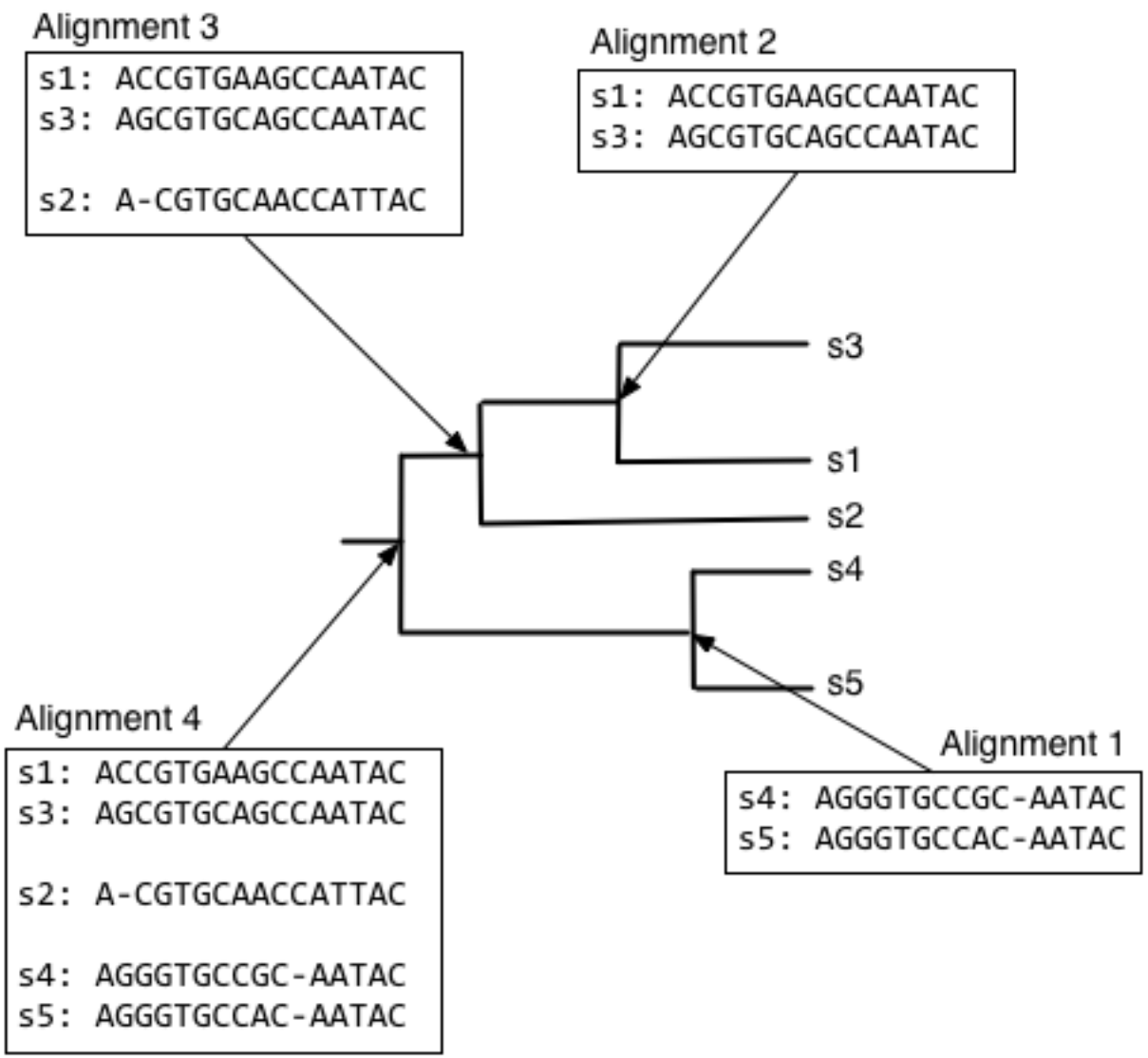
Heuristics (Progressive method)

```
s1: ACCGTGAAGCCAATAC
s2: ACGTGCAACCATTAC
s3: AGCGTGCAGCCAATAC
s4: AGGGTGCCGCAATAC
s5: AGGGTGCCACAATAC
```



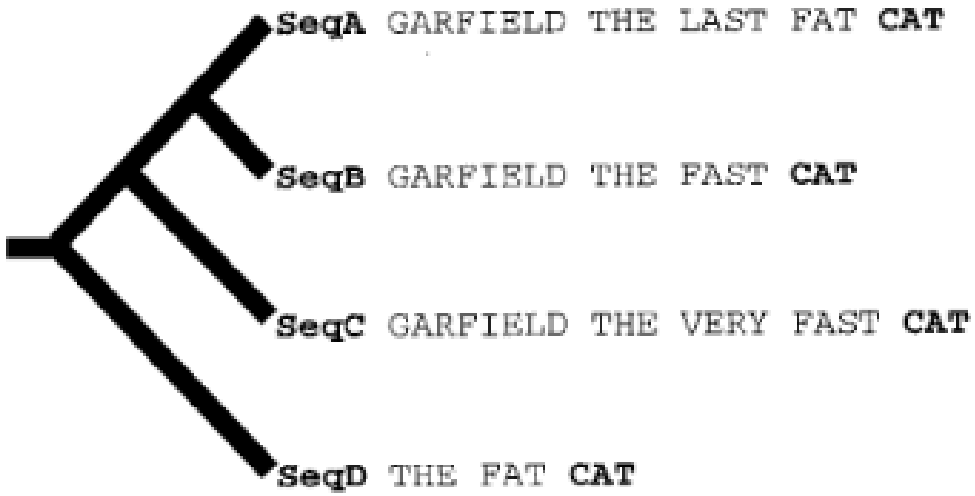
Heuristics (Progressive method)

```
s1: ACCGTGAAGCCAATAC
s2: ACGTGCAACCATTAC
s3: AGCGTGCAGCCAATAC
s4: AGGGTGCCGCAATAC
s5: AGGGTGCCACAATAC
```



Guide tree is crucial...

a)Regular Progressive Alignment Strategy




```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD -------  THE ---- FA-T CAT
```

Notredame, C et al (2000)

More algorithms/methods have been developed: **Iterative**, **HMM**, **Consistency**, etc… to make improvements

MSA tools we use

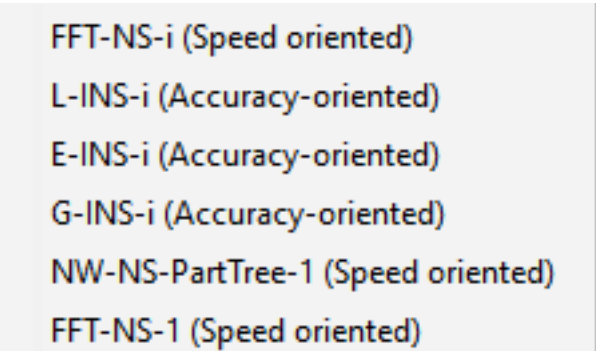
https://www.ebi.ac.uk/Tools/msa/

Tool	Comment	Suitable for	Max # sequences	Max file size
Clustal Omega https://en.wikipedia.org/wiki/Clustal http://www.clustal.org/omega/	Uses seeded trees (popular progressive alignment) and HMM profile-profile. Fast	Medium-large MSAs (proteins, DNA or RNA)	4000	4 Mb
MAFFT https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html	Fast Fourier Transforms. *Fast.*	Medium to large MSAs (protein, DNA)	500	1 Mb
MUSCLE	Log-Expectation method. Accurate and especially good with proteins	Medium MSAs	500	1 Mb
T-Coffee	Consistency-based MSA (combines several aligners). Perhaps a bit slow	Small MSAs	500	1 Mb

Which is faster?

Well, it is not an easy question. It depends on the tool, tool algorithm(1) and version, the parameters, how many/how long are the sequences...

(1) e.g. In MAFFT there are very different algorithms available:



*"One can see that for 100 sequences, default MAFFT is **faster** than default Clustal Omega and default MUSCLE. MUSCLE has a higher-speed option, which employs a smaller number of refinements than the default (two as opposed to 16). For 100 sequences this option is **faster** than Clustal Omega, but still not as fast as default MAFFT. However, as the number of sequences is increased to around 2000, Clustal Omega **overtakes** the high speed MUSCLE version, and for around 10,000–20,000 sequences, **overtakes** default MAFFT. "*

https://onlinelibrary.wiley.com/doi/full/10.1002/pro.3290

EL PAÍS

ccu cggcgggca

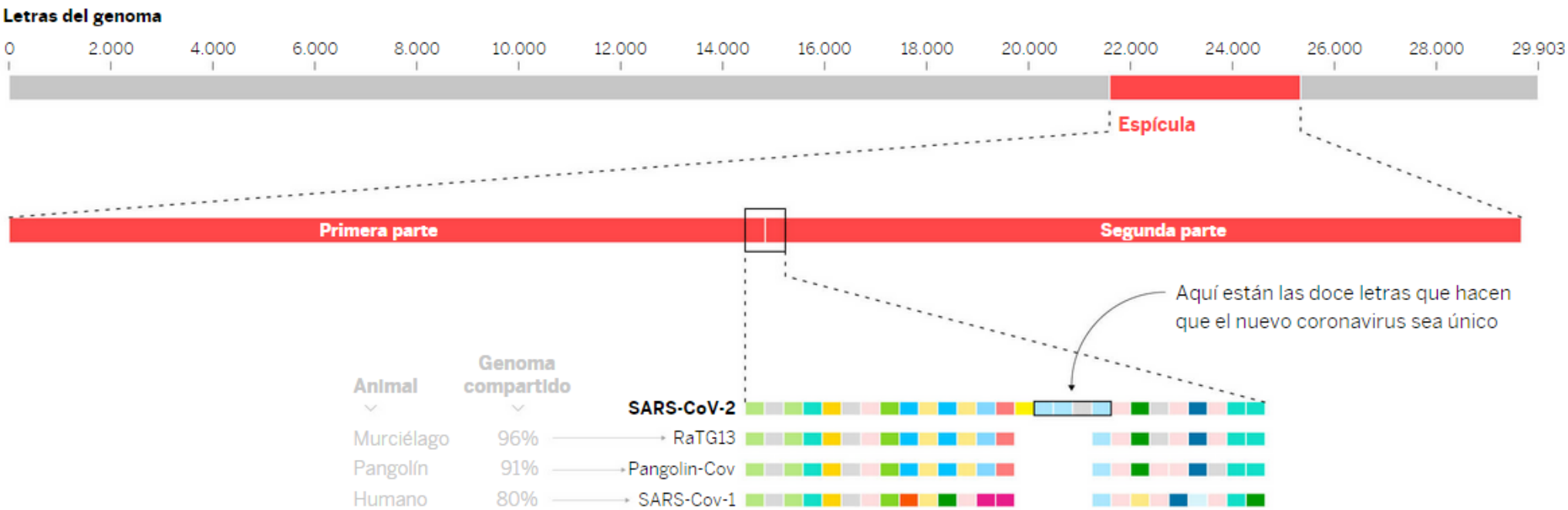
Las doce letras que cambiaron el mundo

MANUEL ANSEDE | ARTUR GALOCHA | MARIANO ZAFRA

Let's Find (Ctrl+F) in JalView this motif in the alignments of coronaviruses genomes (exercise 4) and (PRRA) in the S Proteins (exercise 5):

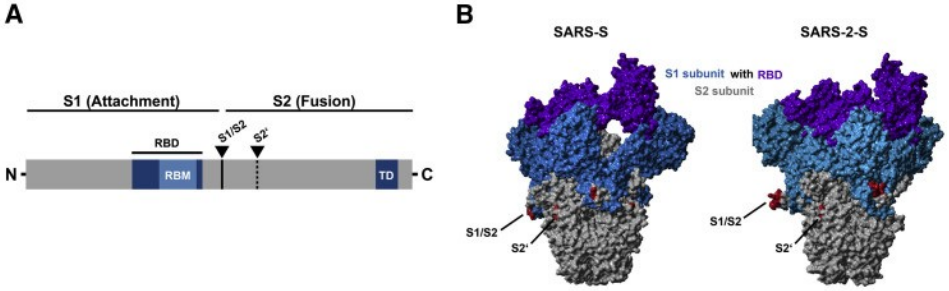
- Which genomes or S proteins carry the motif (DNA: CCTCGGCGGGCA, S proteins: PRRA)?

- Based on the MAFFT alignments: which look to have some "similar" motif (which ones show a gap in this zone)?



capacidad de contagio y de su virulencia

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7194065/ (Hoffmann et al, 2020)



C

	S1/S2	S2'
Human SARS-CoV BJ01	655 - GICASYHTVSLL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Human SARS-CoV CUHK-W1	655 - GICASYHTVSLL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Human SARS-CoV Tor2	655 - GICASYHTVSLL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Human SARS-CoV Frankfurt-1	655 - GICASYHTVSLL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Human SARS-CoV Urbani	655 - GICASYHTVSLL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Civet SARS-CoV civet020	655 - GICASYHTVSSL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Civet SARS-CoV SZ16	655 - GICASYHTVSSL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Raccoon dog SARS-CoV A030	655 - GICASYHTVSSL----RSTS - 670	790 - DPLKPTKRSFIED - 802
SARS-CoV-2	669 - GICASYQTQTNSPRRARSVA - 688	808 - DPSKPSKRSFIED - 820
Pangolin CoV MP789	n/a - GICASYQTQTNS----RSVS - n/a	n/a - DPSKPSKRSFIED - n/a
Bat SARSr-CoV RaTG13	669 - GICASYQTQTNS----RSVA - 684	804 - DPSKPSKRSFIED - 816
Bat SARSr-CoV LYRa11	659 - GICASYHTASLL----RNTD - 674	794 - DPSKPTKRSFIED - 806
Bat SARSr-CoV LYRa3	659 - GICASYHTASLL----RNTG - 674	794 - DPSKPTKRSFIED - 806
Bat SARSr-CoV RsSHC014	656 - GICASYHTVSSL----RSTS - 671	791 - DPLKPTKRSFIED - 803
Bat SARSr-CoV Rs4084	656 - GICASYHTVSSL----RSTS - 671	791 - DPLKPTKRSFIED - 803
Bat SARSr-CoV WIV1	656 - GICASYHTVSSL----RSTS - 671	791 - DPLKPTKRSFIED - 803
Bat SARSr-CoV Rs3367	656 - GICASYHTVSSL----RSTS - 671	791 - DPLKPTKRSFIED - 803
Bat SARSr-CoV Rs7327	656 - GICASYHTVSSL----RSTS - 671	791 - DPLKPTKRSFIED - 803
Bat SARSr-CoV Rs9401	656 - GICASYHTVSSL----RSTS - 671	791 - DPLKPTKRSFIED - 803
Bat SARSr-CoV Rs4231	655 - GICASYHTVSSL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Bat SARSr-CoV WIV16	655 - GICASYHTVSSL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Bat SARSr-CoV Rs4874	655 - GICASYHTVSSL----RSTS - 670	790 - DPLKPTKRSFIED - 802
Bat SARSr-CoV ZC45	646 - GICASYHTASIL----RSTS - 661	781 - DPSKPSKRSFIED - 793
Bat SARSr-CoV ZXC21	645 - GICASYHTASIL----RSTG - 660	780 - DPSKPSKRSFIED - 792
Bat SARSr-CoV Rf4092	634 - GICASYHTASTL----RGVG - 649	769 - DPSKPTKRSFIED - 781
Bat SARSr-CoV Rf/JL2012	636 - GICASYHTASLL----RSTG - 651	771 - DPLKPTKRSFIED - 783
Bat SARSr-CoV JTMC15	636 - GICASYHTASLL----RSTG - 651	771 - DPLKPTKRSFIED - 783
Bat SARSr-CoV 16BO133	636 - GICASYHTASLL----RSTG - 651	771 - DPLKPTKRSFIED - 783
Bat SARSr-CoV B15-21	636 - GICASYHTASLL----RSTG - 651	771 - DPLKPTKRSFIED - 783
Bat SARSr-CoV YN2013	633 - GICASYHTASTL----RSIG - 648	768 - DPSKPTKRSFIED - 780
Bat SARSr-CoV Anlong-103	633 - GICASYHTASTL----RSVG - 648	768 - DPSKPTKRSFIED - 780
Bat SARSr-CoV Rp/Shaanxi2011	640 - GICASYHTASVL----RSTG - 655	775 - DPSKPTKRSFIED - 787
Bat SARSr-CoV Rs/HuB2013	641 - GICASYHTASVL----RSTG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV YNLF/34C	641 - GICASYHTASVL----RSTG - 656	776 - DPLKPTKRSFIED - 788
Bat SARSr-CoV YNLF/31C	641 - GICASYHTASVL----RSTG - 656	776 - DPLKPTKRSFIED - 788
Bat SARSr-CoV Rf1	641 - GICASYHTASHL----RSTG - 656	776 - DPLKPTKRSFIED - 788
Bat SARSr-CoV 273	641 - GICASYHTASHL----RSTG - 656	776 - DPLKPTKRSFIED - 788
Bat SARSr-CoV Rf/SX2013	639 - GICASYHTASLL----RSTG - 654	774 - DPLKPTKRSFIED - 786
Bat SARSr-CoV Rf/HeB2013	641 - GICASYHTASLL----RSTG - 656	776 - DPLKPTKRSFIED - 788
Bat SARSr-CoV Cp/Yunnan2011	641 - GICASYHTASLL----RNTG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV Rs672	641 - GICASYHTASTL----RSVG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV Rs4255	641 - GICASYHTASTL----RSVG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV Rs4081	641 - GICASYHTASTL----RSVG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV Rm1	641 - GICASYHTASVL----RSTG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV 279	641 - GICASYHTASVL----RSTG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV Rs/GX2013	642 - GICASYHTASVL----RSTG - 657	777 - DPSKPTKRSFIED - 789
Bat SARSr-CoV Rs806	641 - GICASYHTASLL----RSTG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV HKU3-1	642 - GICASYHTASVL----RSTG - 657	777 - DPSKPTKRSFIED - 789
Bat SARSr-CoV Longquan-140	642 - GICASYHTASVL----RSTG - 657	777 - DPSKPTKRSFIED - 789
Bat SARSr-CoV Rp3	641 - GICASYHTASTL----RSVG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV Rs4247	642 - GICASYHTASTL----RSVG - 657	777 - DPSKPTKRSFIED - 789
Bat SARSr-CoV Rs4237	641 - GICASYHTASTL----RSVG - 656	776 - DPSKPTKRSFIED - 788
Bat SARSr-CoV As6526	641 - GICASYHTASTL----RSVG - 656	777 - DPSKPTKRSFIED - 789
Bat SARSr-CoV BtKY72/KEN	660 - GICAKFGS---D---KIRMG - 673	793 - DPKKLSYRSFIED - 805
Bat SARSr-CoV BM48-31	658 - GICAKYTNVSST---LVRSG - 674	794 - DPAKPSSRSFIED - 806
	****.:	** * : ******

D

Alpha HCoV-NL63	735 - GICADGSLI----PVRPRNSS - 751	860 - RNIRSSRIAGRSALED - 875
Alpha HCoV-229E	554 - GVCADGSII----AVQPRNVS - 570	679 - LPTSGSRVAGRSAIED - 694
Beta 2a HCoV-OC43	753 - GYCVDYSK-----NRRSRGAI - 768	901 - LGSECSKASSRSAIED - 916
Beta 2a HCoV-HKU1	742 - GFCVDYNSPSSSSSRRKRRSI - 762	895 - LGPHCGS-SSRSFFED - 909
Beta 2b SARS-CoV	655 - GICASYHTVS-L----LRSTS - 670	790 - DP---LKPTKRSFIED - 802
Beta 2b SARS-CoV-2	669 - GICASYQTQT-NSPRRARSVA - 688	808 - DP---SKPSKRSFIED - 820
Beta 2c MERS-CoV	734 - SLCALPDTPSTLTPRSVRSVP - 754	877 - VSISTGSRSARSAIED - 892
	. *. *	: ** :**

Betacovs not infecting human
(plus SARS_Cov_2)

Betacovs infecting Human