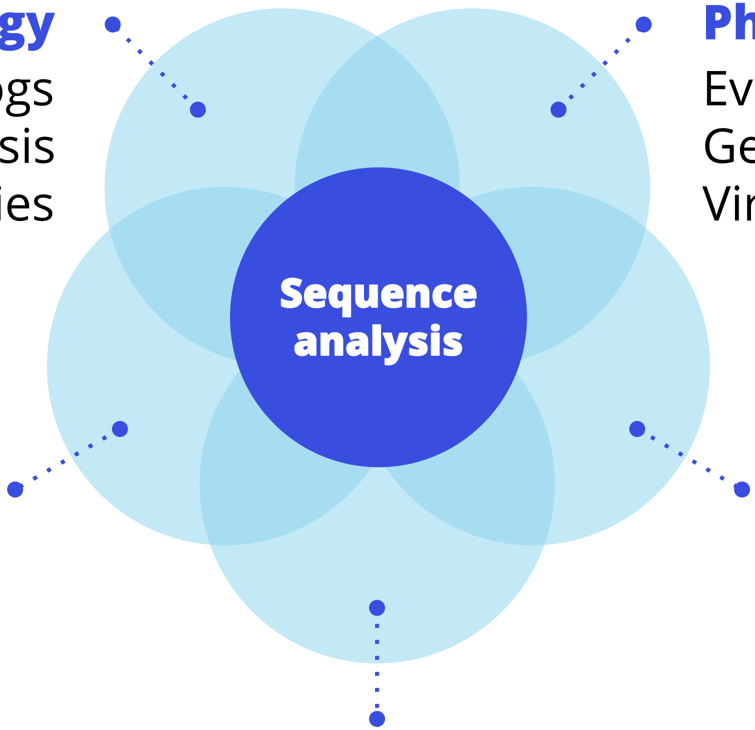
Introduction to MSA

|Multiple Sequence Alignment|

Dr. Urmila Kulkarni-Kale
S. P. Pune University| University of Southeastern Norway | Citadel Precision Medicine
urmila.Kulkarni.kale@gmail.com

Sequence analysis: core of virus bioinformatics



Sequence analysis

Homology
Detection of homologs
Pattern, Motif, Domain analysis
Gene & protein families

Phyloinformatics
Evolutionary studies
Gene based vs Genome based
Viral Typing & diagnostic

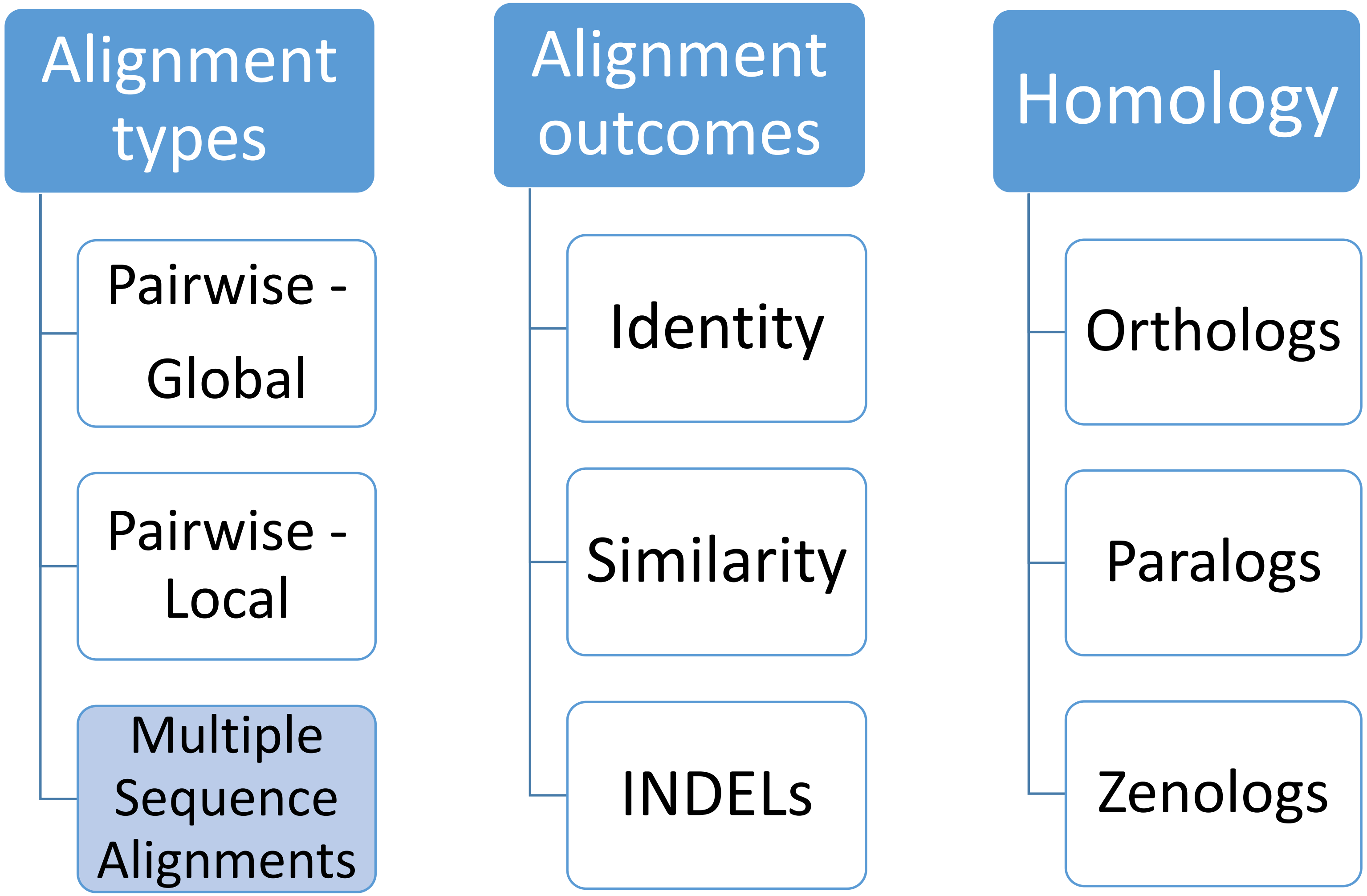
Functional characterisation
Sequence-Structure-Function
Conservation of folds
Drug designing & docking

Immunoinformatics
B- & T- cell epitope predictions
Antigenic variations
Vaccine design

Genomics
Assembly & Annotation
Variation analysis
Population scale analysis

* Bioinformatics, Statistics, Computer Science & Engineering

Sequence alignments: Concepts



Alignment types

- Pairwise - Global
- Pairwise - Local
- Multiple Sequence Alignments

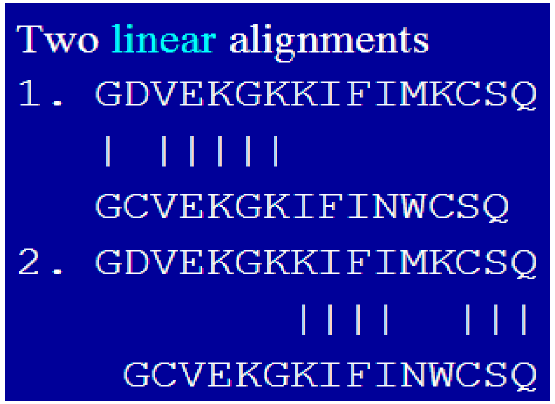
Alignment outcomes

- Identity
- Similarity
- INDELs

Homology

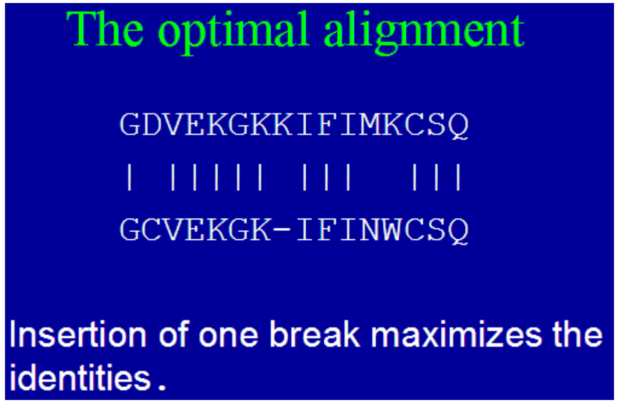
- Orthologs
- Paralogs
- Zenologs

Pairwise Sequence Alignments: Concepts



Two linear alignments

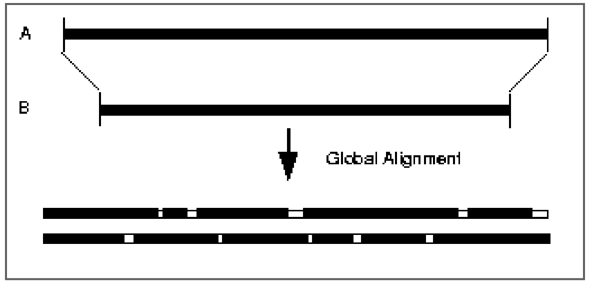
```
1. GDVEKGKKIFIMKCSQ
    |  |||||
   GCVEKGKIFINWCSQ
2. GDVEKGKKIFIMKCSQ
             ||||  |||
    GCVEKGKIFINWCSQ
```



The optimal alignment

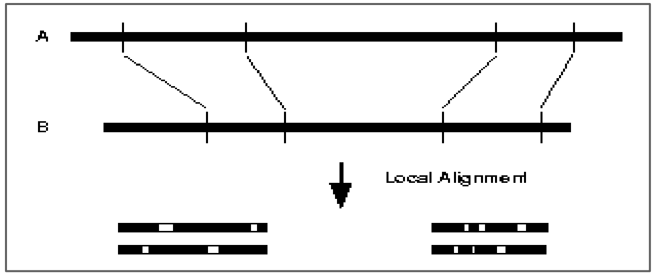
```
GDVEKGKKIFIMKCSQ
| ||||| |||  |||
GCVEKGK-IFINWCSQ
```

Insertion of one break maximizes the identities.



Global

Local

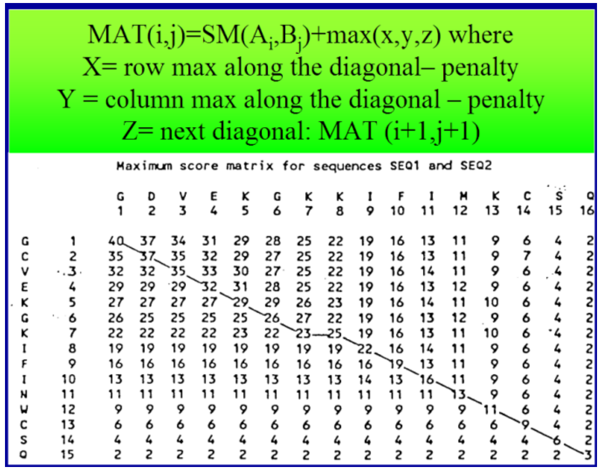


Dynamic programming:

Random walk in 2Ds

Trace the optimal path in matrix

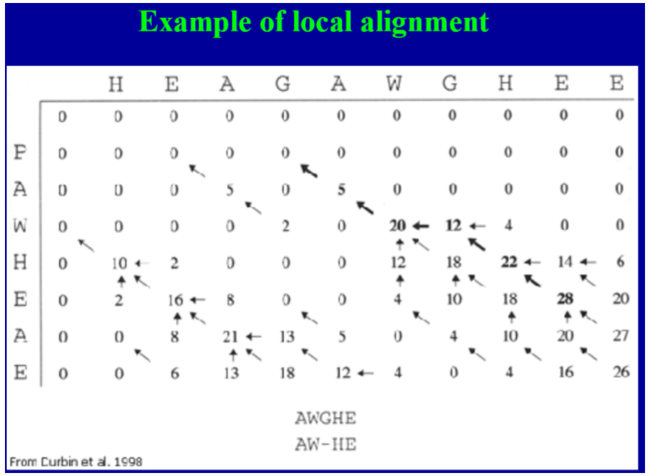
Optimization function: minimize breaks & maximize similarity



MAT(i,j)=SM(A_i,B_j)+max(x,y,z) where
X= row max along the diagonal– penalty
Y = column max along the diagonal – penalty
Z= next diagonal: MAT (i+1,j+1)

Maximum score matrix for sequences SEQ1 and SEQ2

		G	D	V	E	K	G	K	K	I	F	I	M	K	C	S	Q
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
G	1	40	37	34	31	29	28	25	22	19	16	13	11	9	6	4	2
C	2	35	37	35	32	29	27	25	22	19	16	13	11	9	7	4	2
V	3	32	32	35	33	30	27	25	22	19	16	14	11	9	6	4	2
E	4	29	29	29	32	31	28	25	22	19	16	13	12	9	6	4	2
K	5	27	27	27	27	29	29	26	23	19	16	14	11	10	6	4	2
G	6	26	25	25	25	25	26	27	22	19	16	13	12	9	6	4	2
K	7	22	22	22	22	23	22	23	25	19	16	13	11	10	6	4	2
I	8	19	19	19	19	19	19	19	19	22	16	14	11	9	6	4	2
F	9	16	16	16	16	16	16	16	16	16	19	13	11	9	6	4	2
I	10	13	13	13	13	13	13	13	13	14	13	16	11	9	6	4	2
N	11	11	11	11	11	11	11	11	11	11	11	11	13	9	6	4	2
W	12	9	9	9	9	9	9	9	9	9	9	9	9	11	6	4	2
C	13	6	6	6	6	6	6	6	6	6	6	6	6	6	9	4	2
S	14	4	4	4	4	4	4	4	4	4	4	4	4	4	4	6	2
Q	15	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3



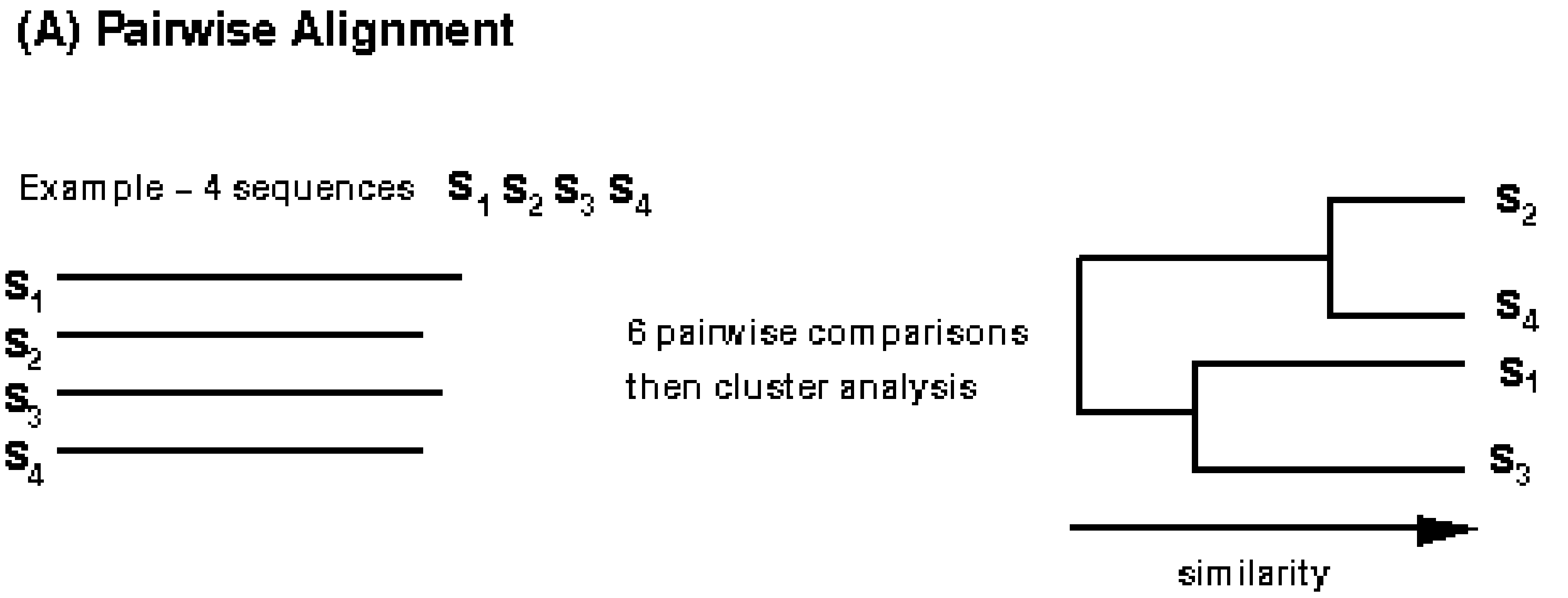
Example of local alignment

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0	0
W	0	0	0	0	2	0	20 ←	12 ←	4	0	0
H	0	10 ←	2	0	0	0	12	18	22 ←	14 ←	6
E	0	2	16 ←	8	0	0	4	10	18	28	20
A	0	0	8	21 ←	13	5	0	4	10	20	27
E	0	0	6	13	18	12 ←	4	0	4	16	26

```
AWGHE
AW-HE
```

From Curbin et al. 1998

Multiple Sequence Alignment (MSA)
Progressive Alignment Model implemented in ClustalW



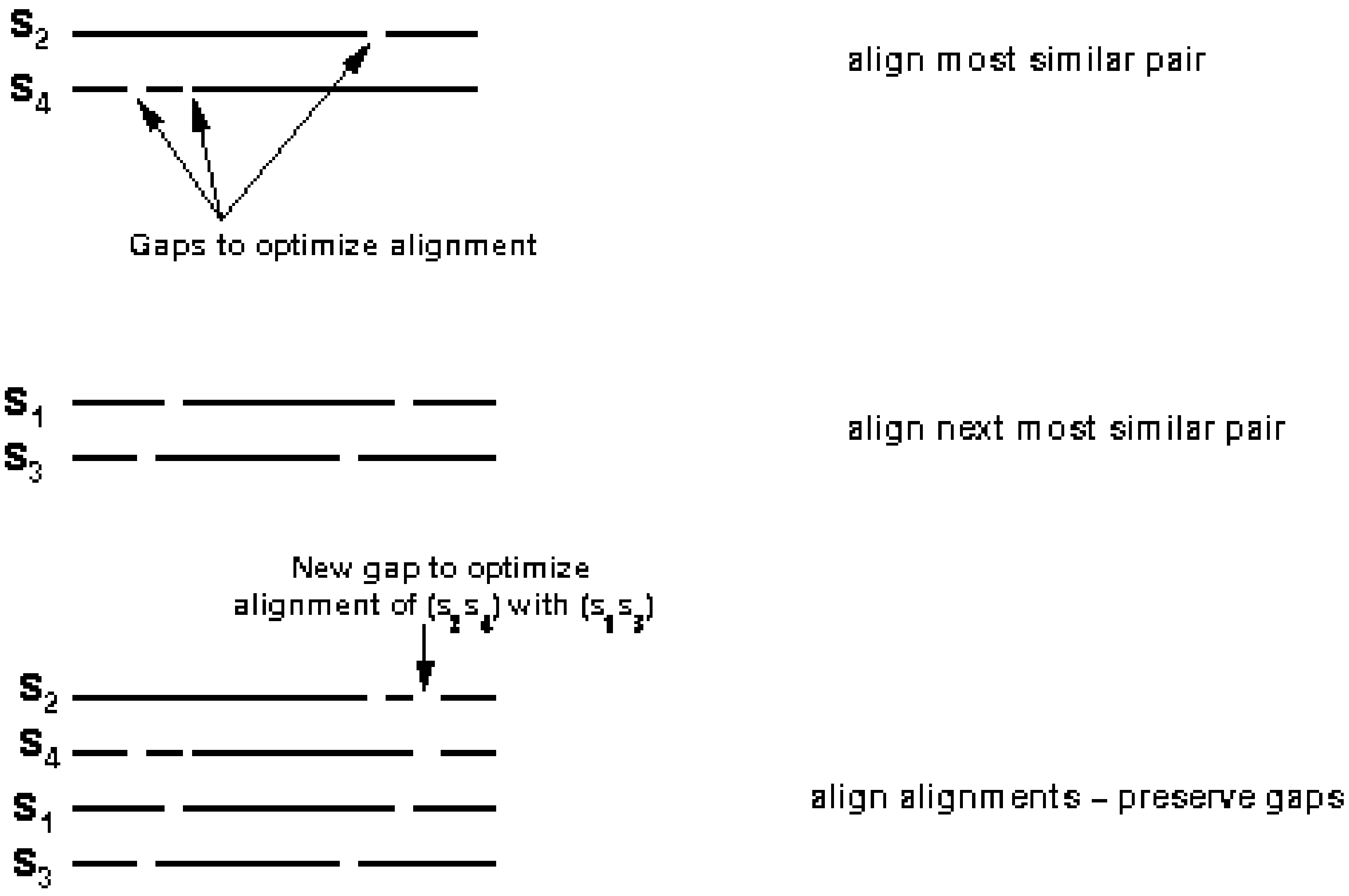
(A) Pairwise Alignment

Example – 4 sequences S_1 S_2 S_3 S_4

6 pairwise comparisons then cluster analysis

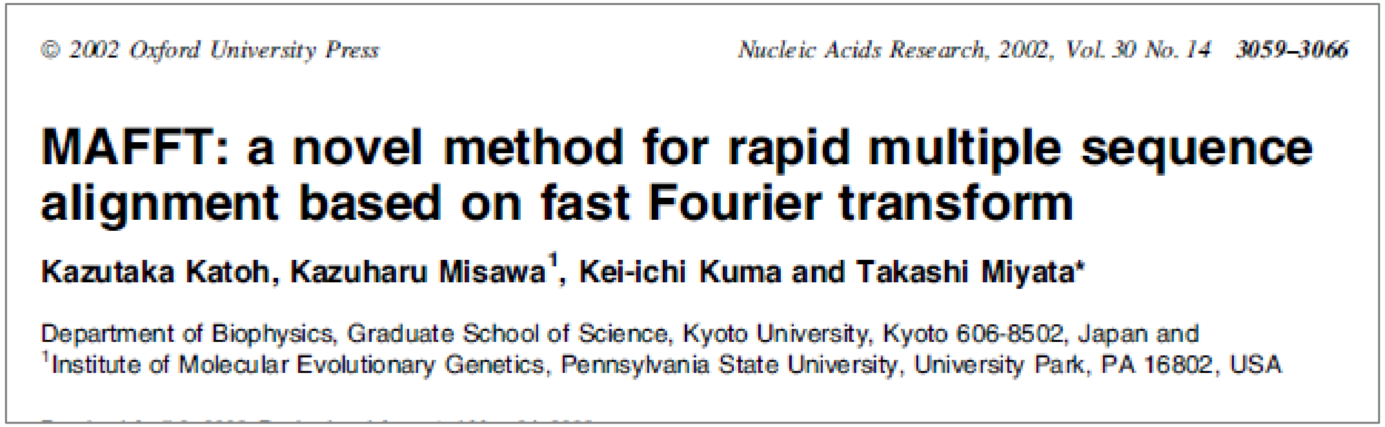
No of pairwise alignments: N*(N-1)/2

(B) Multiple alignment following the tree from A


S2
S4
align most similar pair
Gaps to optimize alignment
S1
S3
align next most similar pair
New gap to optimize alignment of (s2 s4) with (s1 s3)
S2
S4
S1
S3
align alignments – preserve gaps

MAFFT Algorithm

Multiple Alignment using Fast Fourier Transform



© 2002 Oxford University Press

Nucleic Acids Research, 2002, Vol. 30 No. 14 **3059–3066**

MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform

Kazutaka Katoh, Kazuharu Misawa[1], Kei-ichi Kuma and Takashi Miyata*

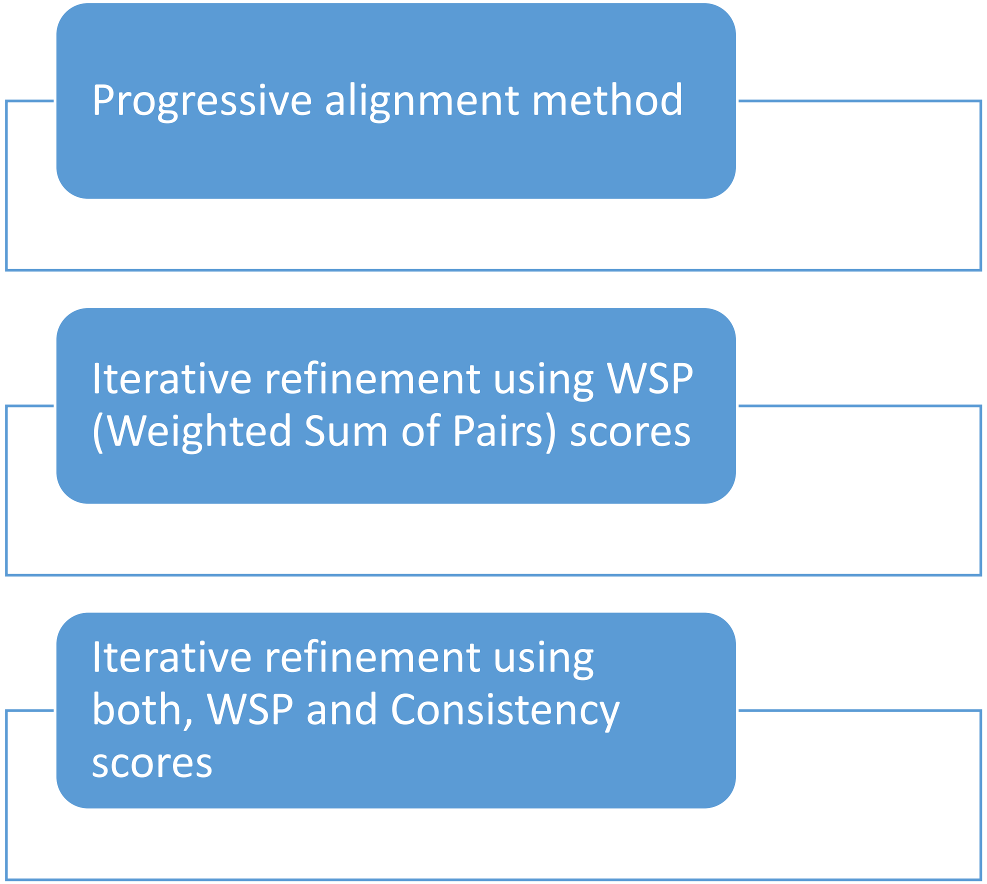
Department of Biophysics, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan and
[1]Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA 16802, USA

MSA Salient Features

- Fast Fourier transform (FFT): rapid detection of homologous segments

- Homologous regions are rapidly identified by FFT, in which
  - nucleotide sequence is converted to four-dimensional vectors whose components are the frequencies of A, T, G and C at each column
  - amino acid sequence is converted to a sequence composed of volume and polarity values of each amino acid residue
- Homologous regions in sequences are marked using sliding window analysis using threshold criteria of correlation score
  - Window size of 30 sites

- Contiguous homologous segments of length 150 sites are identified and combined



Progressive alignment method

Iterative refinement using WSP (Weighted Sum of Pairs) scores

Iterative refinement using both, WSP and Consistency scores

Weighted Sum of Pairs

- WSP is calculated as proposed by Gotoh, 1995 as follows:
- For a MSA '*A*' of alignment length *l* and composed of *N* nucleotides or protein sequences, the sum of pair scores of *A* i.e., SP (*A*) is defined as

$$\mathrm{SP(A)} = \sum_{j=2}^{N}\sum_{k=1}^{j-1} S_{j,k}$$

where $S_{j,k}$ is the score of pairwise alignment between j^{th} and k^{th} sequences as defined by a scoring matrix within *A*

- WSP is defined as:

$$\mathrm{WSP(A)} = \sum_{j=2}^{N}\sum_{k=1}^{j-1} w_{j\ k} S_{j,k}$$

where $w_{j,k}$ is the weight of individual pairs of sequences in *A*

Consistency Scores

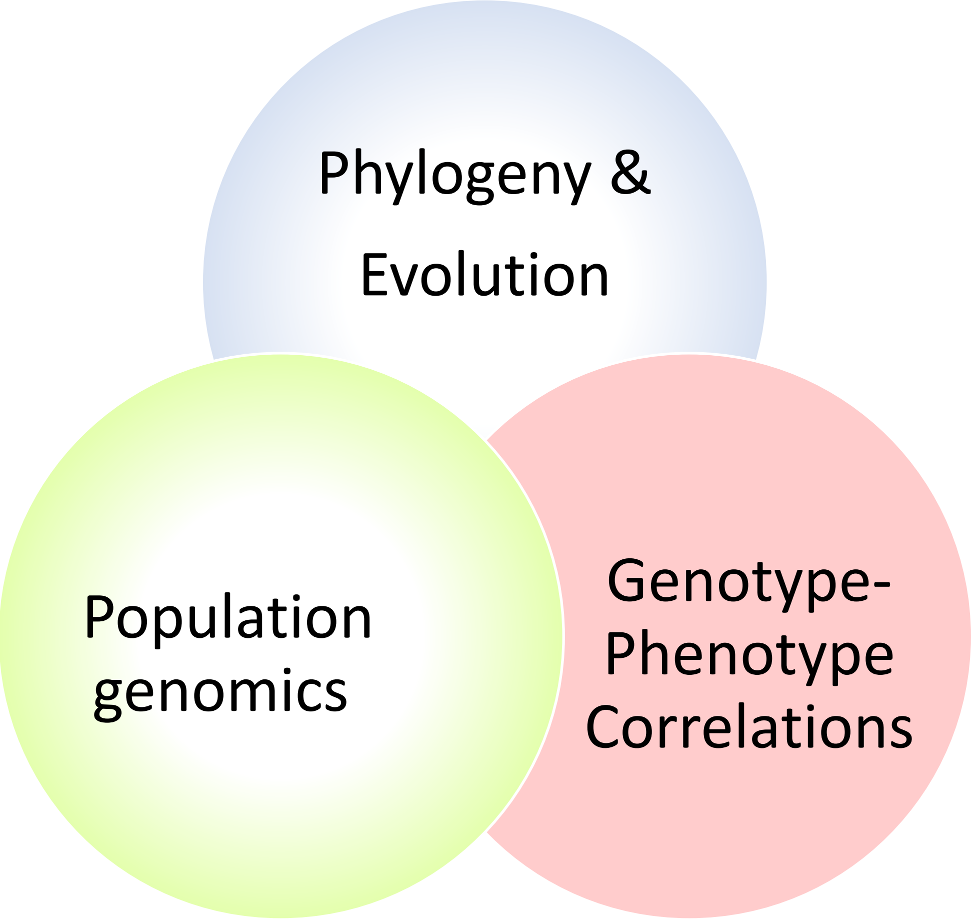
- Consistency scores adopted from TCOFFEE algorithm
- Two alignment scores (local and global) are derived based on a library of local and global alignments
- In the library, each alignment is represented as a list of pairwise residue matches
- The pairwise residue score for a similar pair between local and global alignment libraries is the sum of both scores
- Weights are assigned based on **consistency** of given residue pair relative to other residue pairs in the library

MSA: Interpretations & Applications

Tracking Evolution of SARS-CoV-2 at the onset of pandemic

Virus bioinformatics @ SPPU
Data led discovery framework: DBT CoE| DeiTY CoE|

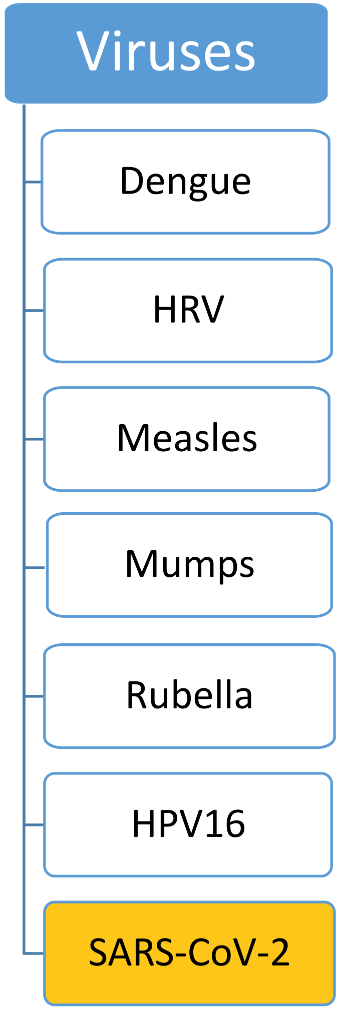
Approaches | Case studies



Phylogeny & Evolution

Population genomics

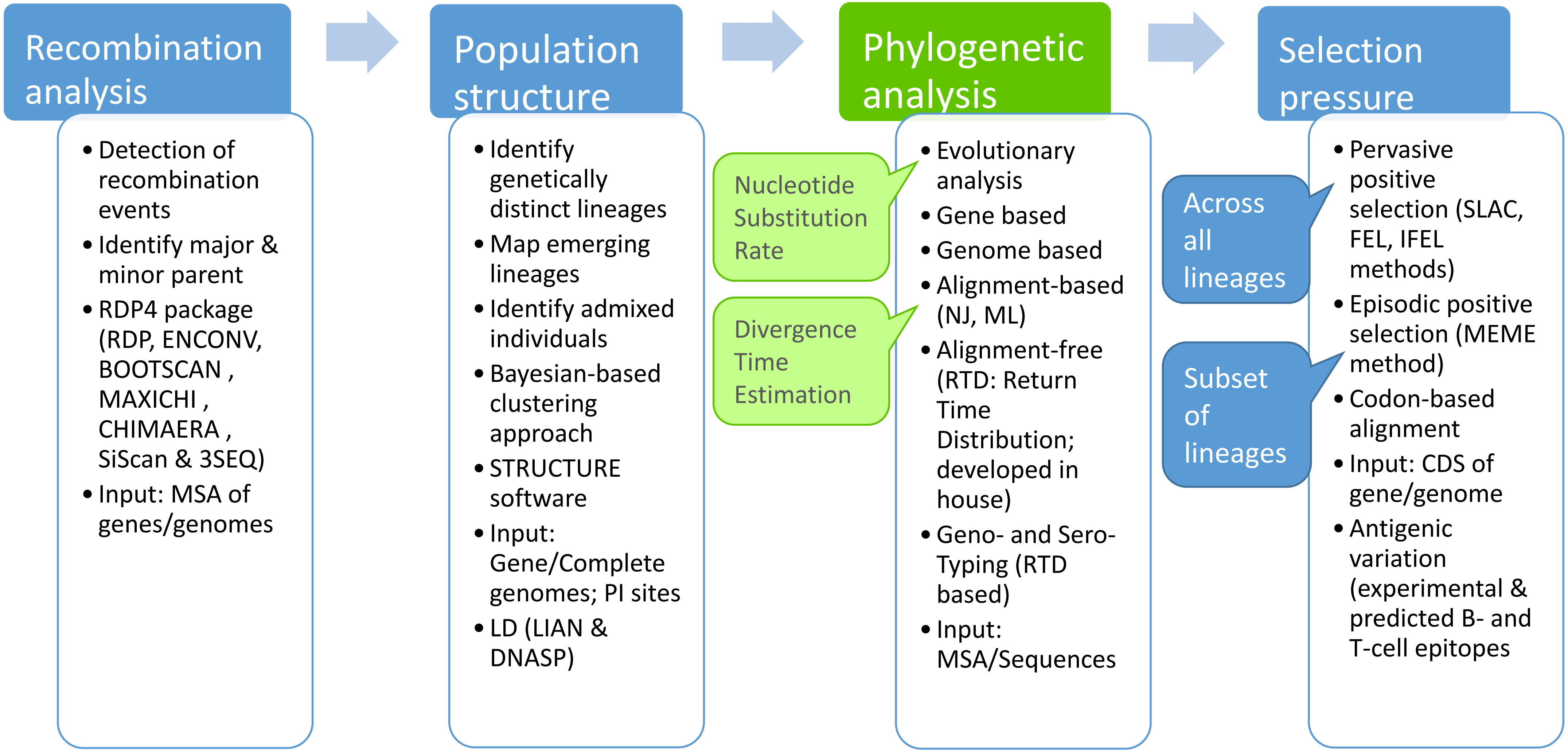
Genotype-Phenotype Correlations



Viruses

- Dengue
- HRV
- Measles
- Mumps
- Rubella
- HPV16
- SARS-CoV-2

Exploring the unknown of (un)known Data → Discovery



Recombination analysis

- Detection of recombination events
- Identify major & minor parent
- RDP4 package (RDP, ENCONV, BOOTSCAN , MAXICHI , CHIMAERA , SiScan & 3SEQ)
- Input: MSA of genes/genomes

Population structure

- Identify genetically distinct lineages
- Map emerging lineages
- Identify admixed individuals
- Bayesian-based clustering approach
- STRUCTURE software
- Input: Gene/Complete genomes; PI sites
- LD (LIAN & DNASP)

Nucleotide Substitution Rate

Divergence Time Estimation

Phylogenetic analysis

- Evolutionary analysis
- Gene based
- Genome based
- Alignment-based (NJ, ML)
- Alignment-free (RTD: Return Time Distribution; developed in house)
- Geno- and Sero-Typing (RTD based)
- Input: MSA/Sequences

Across all lineages

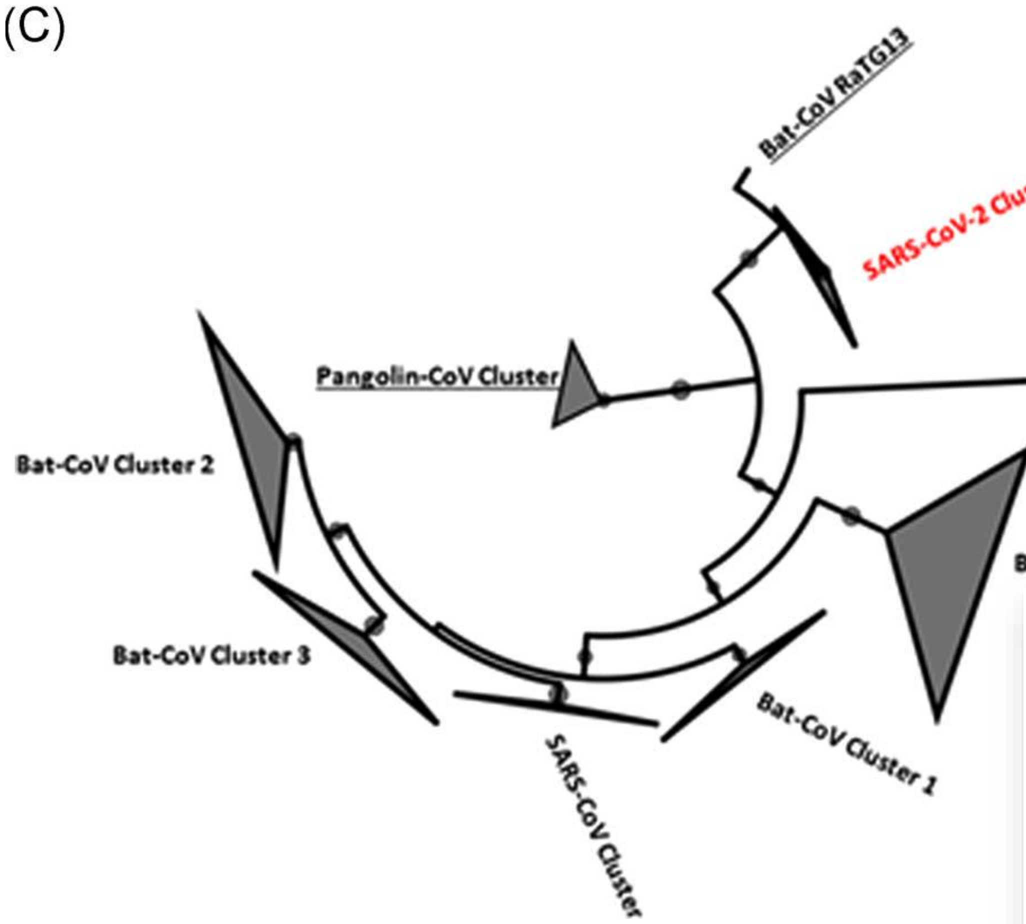
Subset of lineages

Selection pressure

- Pervasive positive selection (SLAC, FEL, IFEL methods)
- Episodic positive selection (MEME method)
- Codon-based alignment
- Input: CDS of gene/genome
- Antigenic variation (experimental & predicted B- and T-cell epitopes

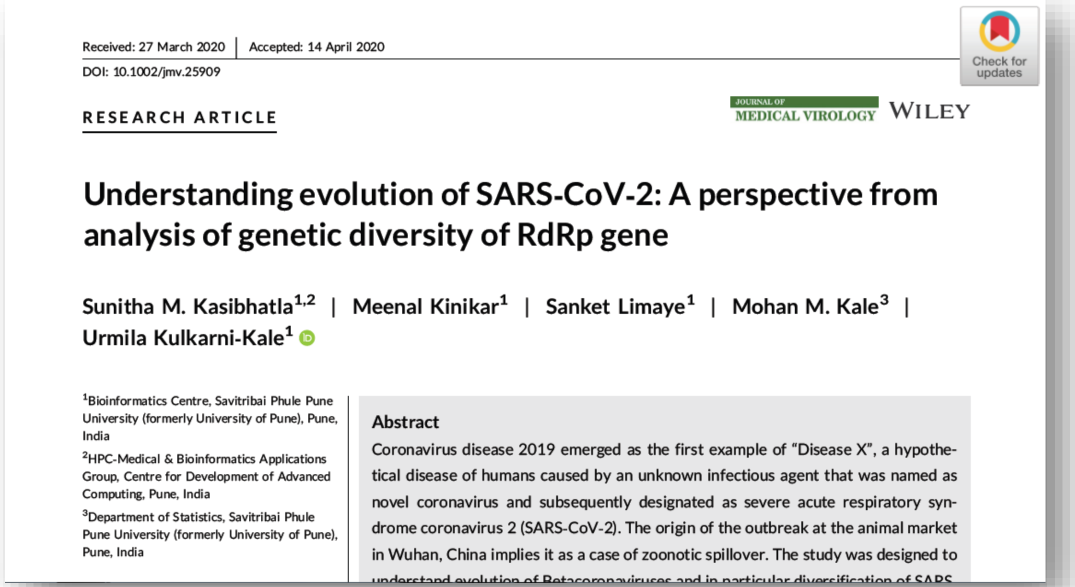
SARS-CoV-2|Phylogenetic analysis of RdRp gene

(C)



Method: Maximum Likelihood (ML)

● Bootstrap support >70%


Received: 27 March 2020 | Accepted: 14 April 2020
DOI: 10.1002/jmv.25909

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Understanding evolution of SARS-CoV-2: A perspective from analysis of genetic diversity of RdRp gene

Sunitha M. Kasibhatla[1,2] | Meenal Kinikar[1] | Sanket Limaye[1] | Mohan M. Kale[3] | Urmila Kulkarni-Kale[1]

[1]Bioinformatics Centre, Savitribai Phule Pune University (formerly University of Pune), Pune, India

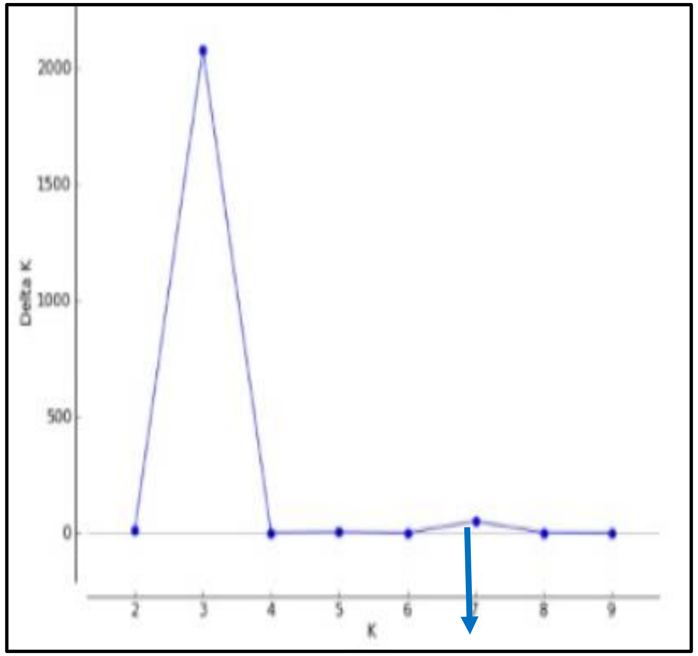
[2]HPC-Medical & Bioinformatics Applications Group, Centre for Development of Advanced Computing, Pune, India

[3]Department of Statistics, Savitribai Phule Pune University (formerly University of Pune), Pune, India

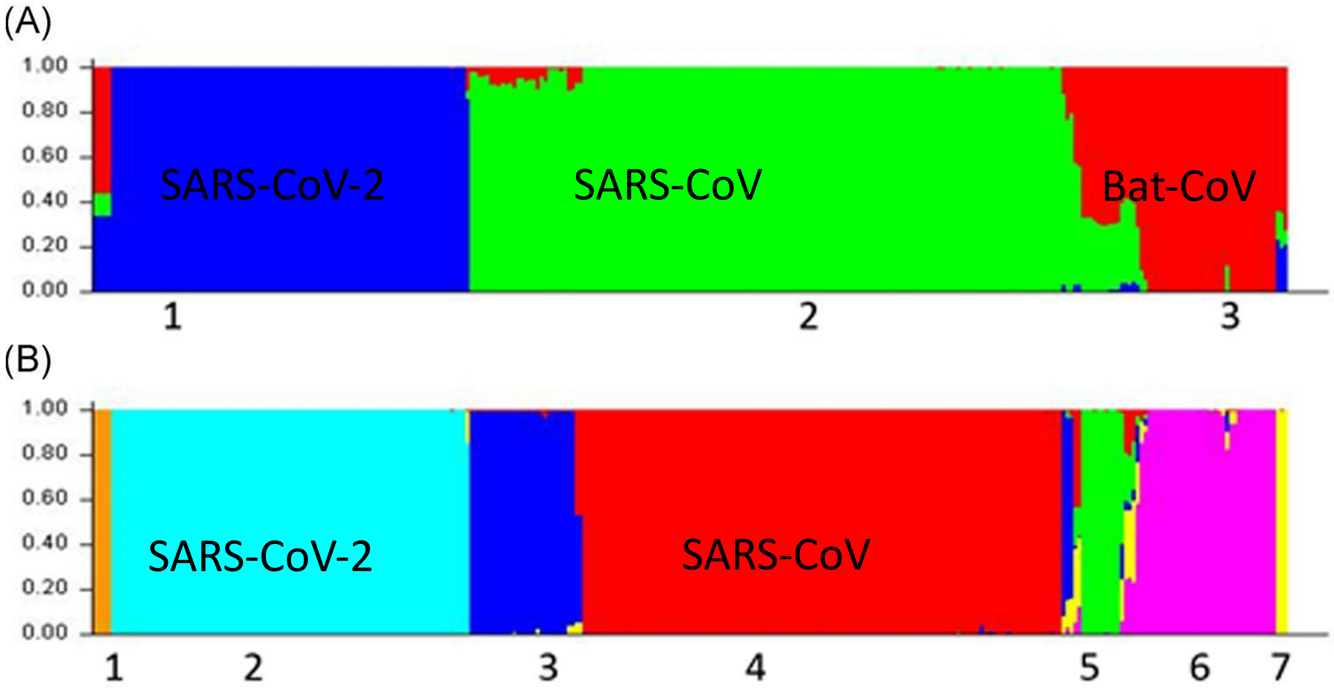
Abstract

Coronavirus disease 2019 emerged as the first example of "Disease X", a hypothetical disease of humans caused by an unknown infectious agent that was named as novel coronavirus and subsequently designated as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The origin of the outbreak at the animal market in Wuhan, China implies it as a case of zoonotic spillover. The study was designed to understand evolution of Betacoronaviruses and in particular diversification of SARS

SARS-CoV-2: population structure

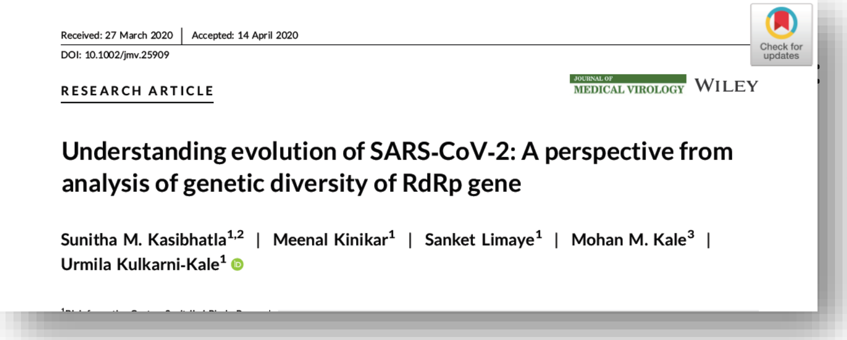


Plot of *K vs* delta *K* which depicts the major peak at *K* =3 and minor peak at *K* =7.



A: Population stratification at optimal peak k=3 wherein the labels 1, 2, and 3 represent SARS-CoV-2, SARS-CoV, and Bat-CoV.

: Population stratification at minor peak k=7 wherein the labels **1 (Pangolin-CoV), 2(SARS-CoV-2), 3(Bat-CoV-Cluster_1), 4(SARS-CoV), 5(Bat-CoV-Cluster_2), 6(Bat-CoV-Cluster_3), 7(Bat-CoV-Cluster_4)**



Received: 27 March 2020 | Accepted: 14 April 2020

DOI: 10.1002/jmv.25909

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

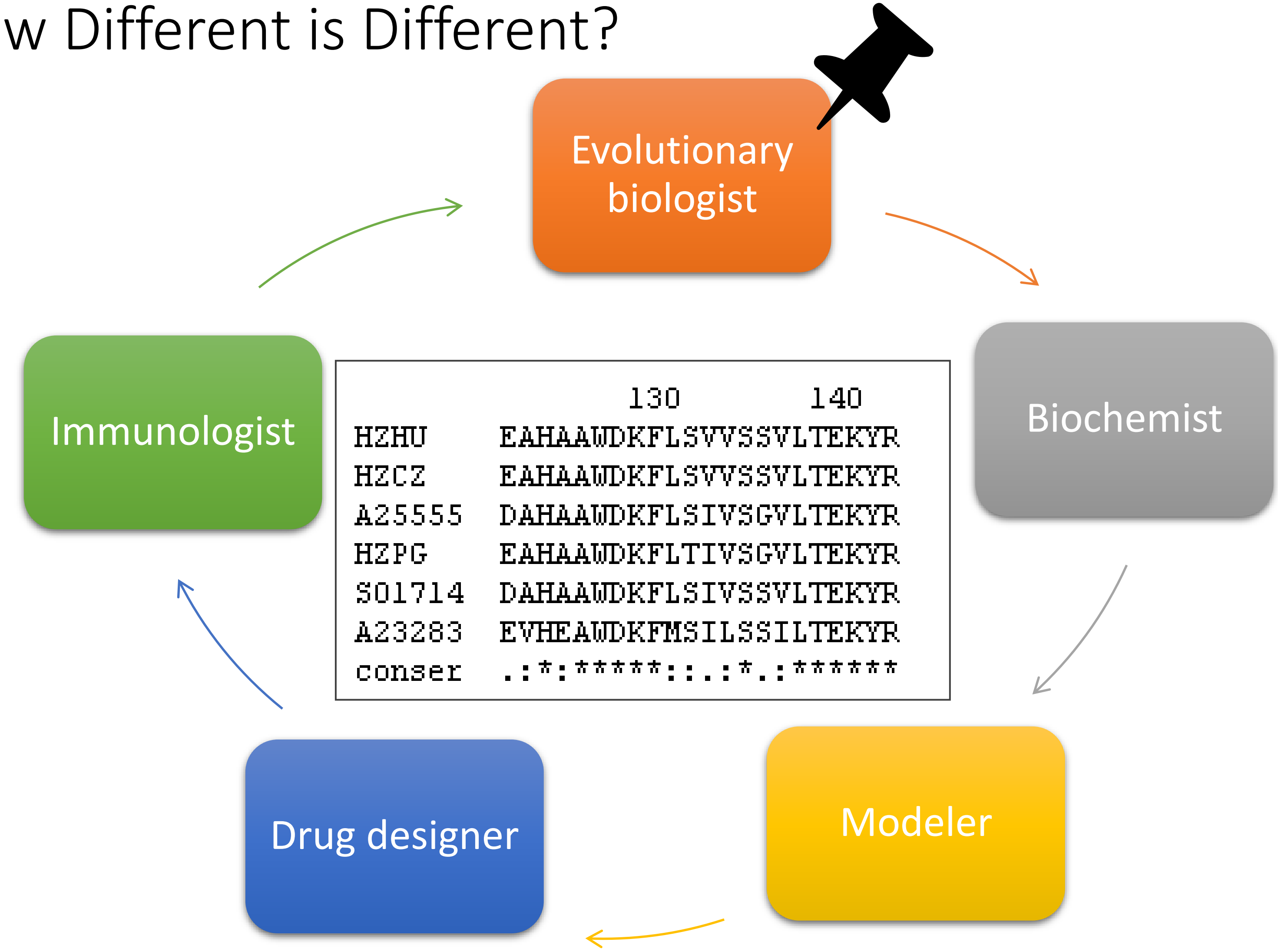
Check for updates

Understanding evolution of SARS-CoV-2: A perspective from analysis of genetic diversity of RdRp gene

Sunitha M. Kasibhatla[1,2] | Meenal Kinikar[1] | Sanket Limaye[1] | Mohan M. Kale[3] | Urmila Kulkarni-Kale[1]

MSA
Genome to Function via Sequence & Structure

How Similar is Similar?
How Different is Different?



```
                130       140
HZHU    EAHAAWDKFLSVVSSVLTEKYR
HZCZ    EAHAAWDKFLSVVSSVLTEKYR
A25555  DAHAAWDKFLSIVSGVLTEKYR
HZPG    EAHAAWDKFLTIVSGVLTEKYR
S01714  DAHAAWDKFLSIVSSVLTEKYR
A23283  EVHEAWDKFMSILSSILTEKYR
conser  .:*:*****::.:*.:******
```

Depends on the question you want to ask and answer using alignment output

Data to Information
Perspectives from similarity (JEV story)

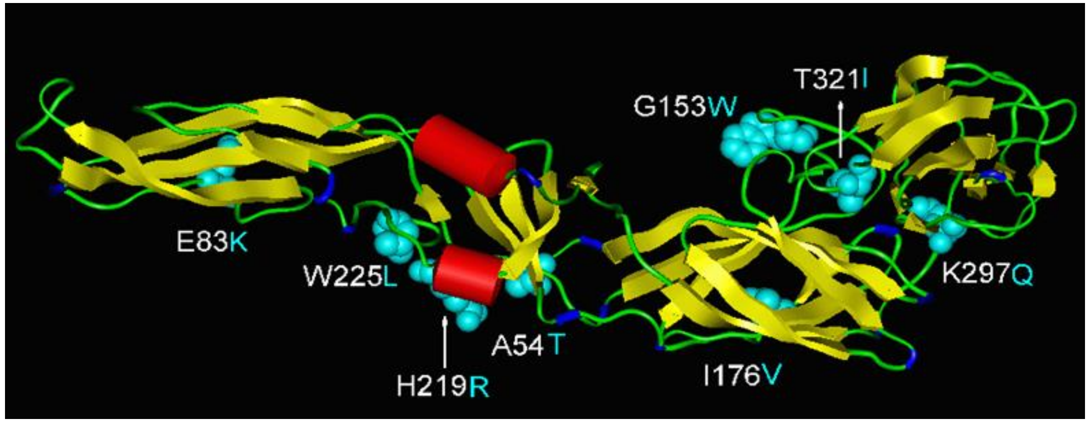
Species & Strain specific variations

```
Loop1 in TBEV: TLAEEHQGG
Loop1 in JEVN:  HNEKRADSS
Loop1 in JEVS:  HNKKRADSS
```

Biodiversity & Biocomplexity:

Isolates, strains, genotypes, serotypes, species & clades

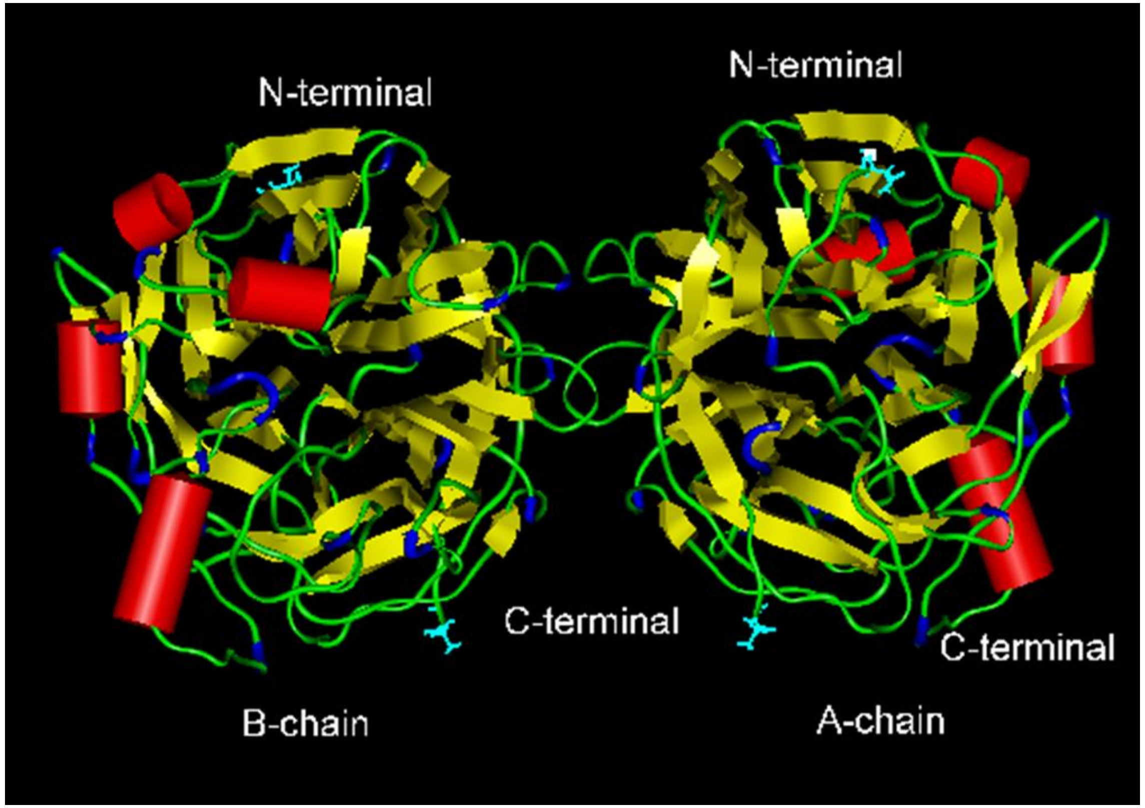
Genus, subfamilies, families

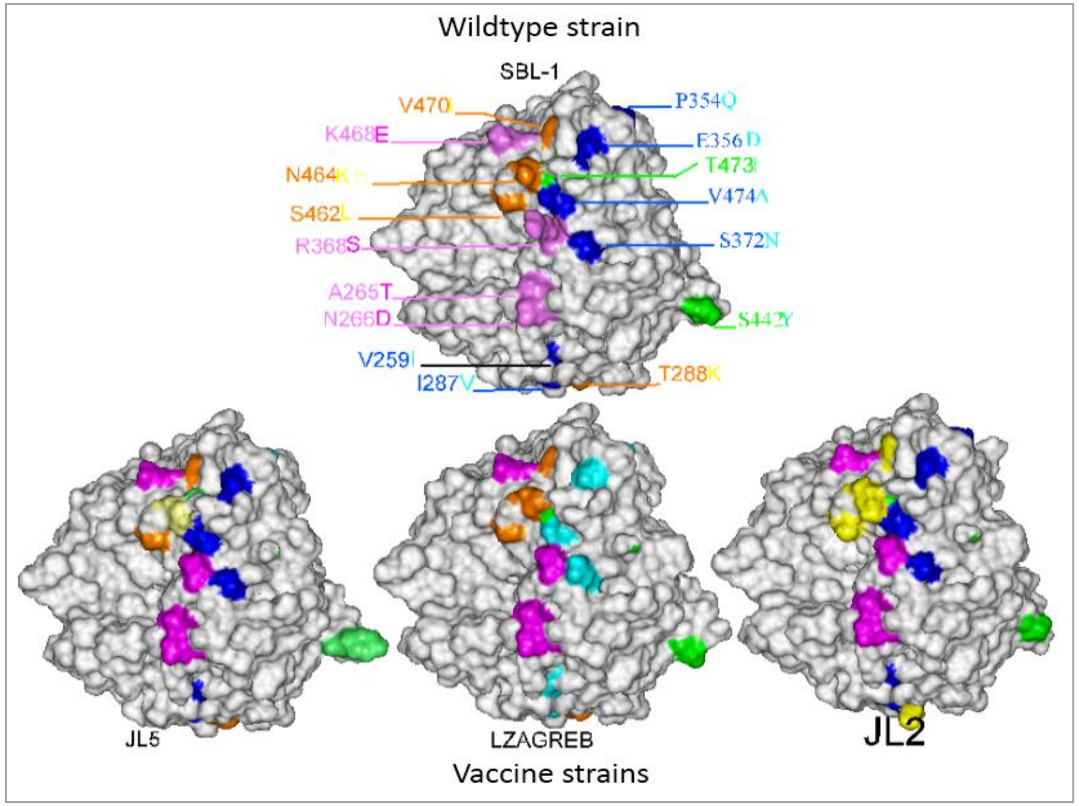


Sequence-based analyses

Structure-based analyses

MSA → Explaining immune escape
Case study: Mumps virus story

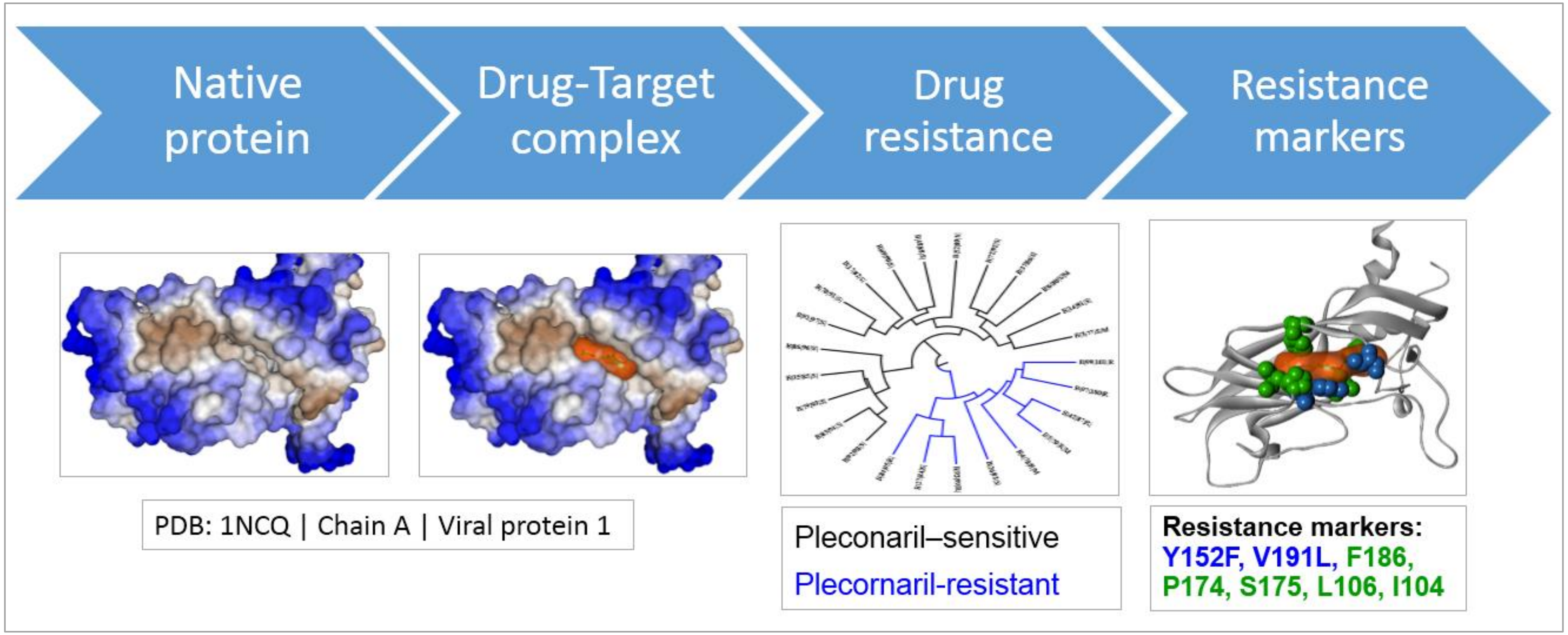




Kulkarni-Kale et al., 2007.

Funded by: Serum Institute of India

MSA →Tracing emergence of drug resistance
Case study: Human Rhinoviruses



Native protein → Drug-Target complex → Drug resistance → Resistance markers

PDB: 1NCQ | Chain A | Viral protein 1

Pleconaril–sensitive

Plecornaril-resistant

Resistance markers:
Y152F, V191L, F186,
P174, S175, L106, I104

Waman et al., Unpublished

MAFFT References

- Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul 15;30(14):3059-66. doi: 10.1093/nar/gkf436. PMID: 12136088; PMCID: PMC135756.
- Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005 Jan 20;33(2):511-8. doi: 10.1093/nar/gki198. PMID: 15661851; PMCID: PMC548345.
- Katoh K, Toh H. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. BMC Bioinformatics. 2008 Apr 25;9:212. doi: 10.1186/1471-2105-9-212. PMID: 18439255; PMCID: PMC2387179.
- Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 2008 Jul;9(4):286-98. doi: 10.1093/bib/bbn013. Epub 2008 Mar 27. PMID: 18372315.
- Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol. 2009;537:39-64. doi: 10.1007/978-1-59745-251-9_3. PMID: 19378139.
- Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics. 2010 Aug 1;26(15):1899-900. doi: 10.1093/bioinformatics/btq224. Epub 2010 Apr 28. PMID: 20427515; PMCID: PMC2905546.
- Katoh K, Frith MC. Adding unaligned sequences into an existing alignment using MAFFT and LAST. Bioinformatics. 2012 Dec 1;28(23):3144-6. doi: 10.1093/bioinformatics/bts578. Epub 2012 Sep 27. PMID: 23023983; PMCID: PMC3516148.
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772-80. doi: 10.1093/molbev/mst010. Epub 2013 Jan 16. PMID: 23329690; PMCID: PMC3603318.
- Katoh K, Standley DM. MAFFT: iterative refinement and additional methods. Methods Mol Biol. 2014;1079:131-46. doi: 10.1007/978-1-62703-646-7_8. PMID: 24170399.