

Phylogeny of three IBV proteins

GENE 8940 Coronavirus Genome Analysis Project

Eric Talevich, Wenyuan Xiao, Yupeng Wang, Celine Hong and
Xinyu Liu

Institute of Bioinformatics, University of Georgia

Apr. 28, 2009

Choosing sequences to analyze

$$\text{Sequences} = \begin{bmatrix} \text{CAL557} \\ \text{CAL95} \\ \text{CONN66} \\ \text{CONN72} \\ \text{CONN83} \\ \text{CONN91} \\ \text{MASS65} \\ \text{MASS72} \\ \text{MASS79} \\ \text{MASS06} \end{bmatrix} \times \begin{bmatrix} \text{Spike} \\ \text{Nucleocapsid} \\ \text{Polyprotein} \end{bmatrix} \times \begin{bmatrix} \text{Nucleotide} \\ \text{Peptide} \end{bmatrix}$$

Extracting sequences

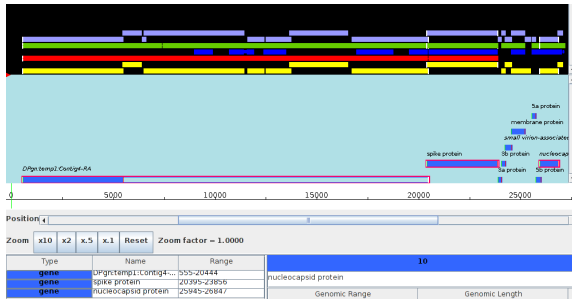


Figure: Selected genes in an example IBV genome (Mass65)

Spike protein (Xinyu)

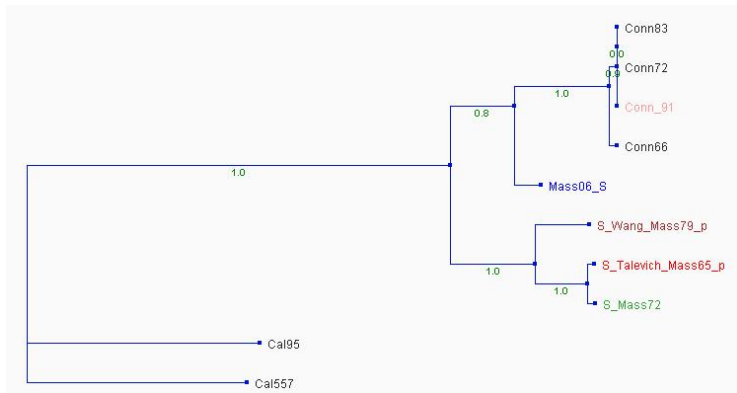
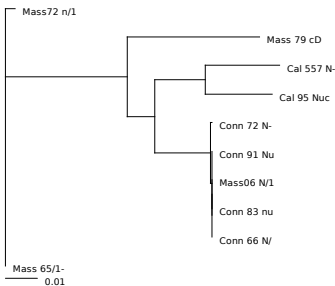
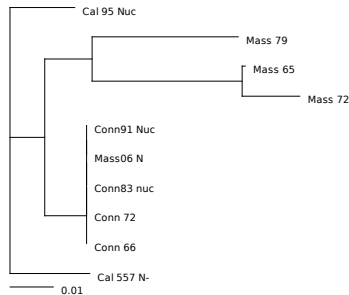


Figure: Phylogram of S protein by PhyML maximum likelihood method

Nucleocapsid protein (Celine)



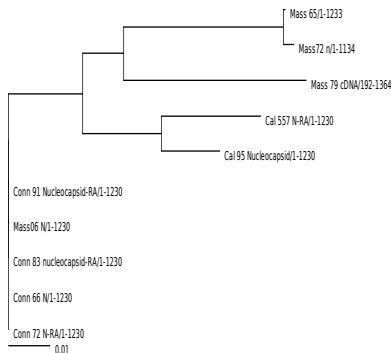
(a) DNA



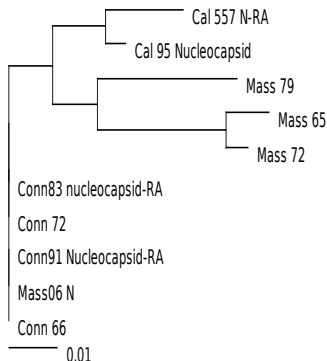
(b) Protein

Figure: Phylograms generated in Phylip by Neighbor Joining method

Nucleocapsid protein (Celine)



(a) DNA



(b) Protein

Figure: Alternative phylograms generated by ClustalX's NJ method, with automatic gap and bootstrap filtering

Polyprotein 1ab (Wenyuan)



Figure: Phylogram of DNA sequences via Phylip NJ method

dN/dS statistic (Yupeng)

Codon usage in genomes, total:

Phe	TTT	84	Ser	TCT	79	Tyr	TAT	46	Cys	TGT	7
	TTC	36		TCC	1		TAC	25		TGC	13
Leu	TTA	5		TCA	46	***	TAA	0	***	TGA	0
	TTG	10		TCG	23		TAG	0	Trp	TGG	60
Leu	CTT	44	Pro	CCT	86	His	CAT	28	Arg	CGT	79
	CTC	10		CCC	22		CAC	11		CGC	43
	CTA	40		CCA	118	Gln	CAA	68		CGA	8
	CTG	21		CCG	14		CAG	54		CGG	11
Ile	ATT	57	Thr	ACT	49	Asn	AAT	93	Ser	AGT	42
	ATC	20		ACC	1		AAC	23		AGC	56
	ATA	31		ACA	54	Lys	AAA	97	Arg	AGA	72
Met	ATG	40		ACG	14		AAG	180		AGG	37
Val	GTT	58	Ala	GCT	108	Asp	GAT	155	Gly	GGT	183
	GTC	34		GCC	34		GAC	72		GGC	20
	GTA	4		GCA	113	Glu	GAA	51		GGA	127
	GTG	23		GCG	11		GAG	22		GGG	27

dN/dS statistic

Yang & Nielsen method ¹ (yn report):

seq.	seq.	S	N	dN	dS	dN/dS
10	1	218.2	681.8	0.0346	0.3460	0.100
10	2	214.3	685.7	0.0290	0.3011	0.096
10	3	217.8	682.2	0.0269	0.2137	0.126
10	4	217.8	682.2	0.0269	0.2137	0.126
10	5	217.8	682.2	0.0269	0.2137	0.126
10	6	217.8	682.2	0.0269	0.2137	0.126
10	7	217.8	682.2	0.0269	0.2137	0.126
10	8	221.0	679.0	0.0302	0.3186	0.095
10	9	206.2	693.8	0.0072	0.0049	1.469

¹Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17:32-43

Conclusions

- 1 Phylogeny \approx geography
- 2 Analysis is **highly** sensitive to alignment and curation techniques
- 3 Unrooted tree is probably more appropriate
- 4 Strong negative selection against nucleocapsid mutations (dN/dS)

Questions?