The Epic Battle between Markov and Phylogenetic Invariants
Graphs

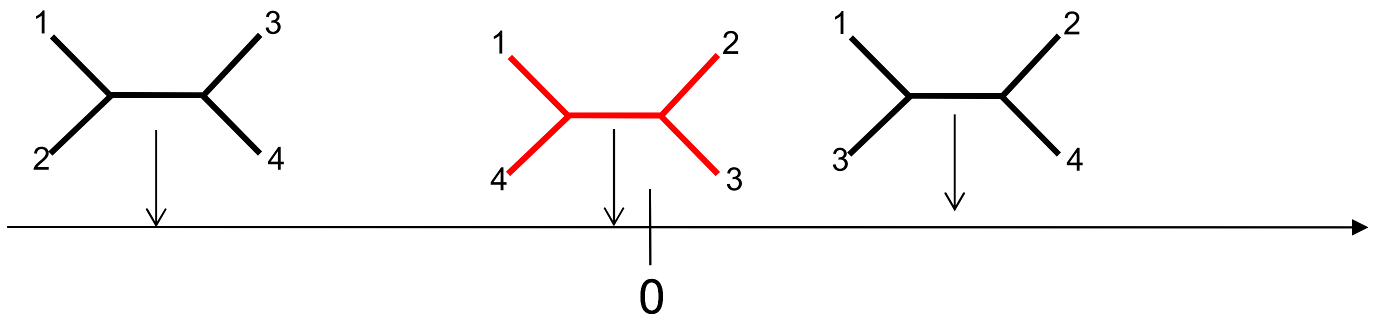
Jeremy Sumner

Peter Jarvis

Amelia Taylor

Barbara Holland

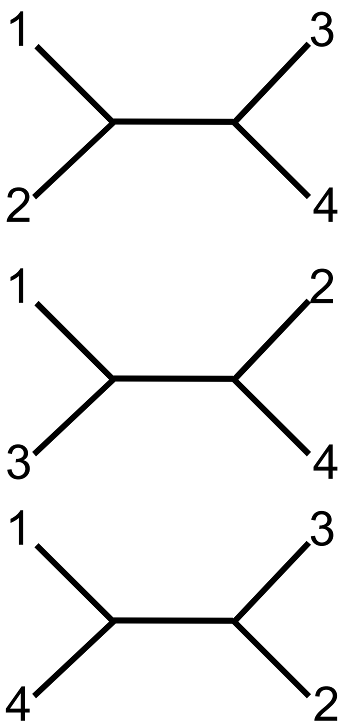
Unsigned Sum of Squares



- Get one number for each tree
- Pick the one that is closest to zero

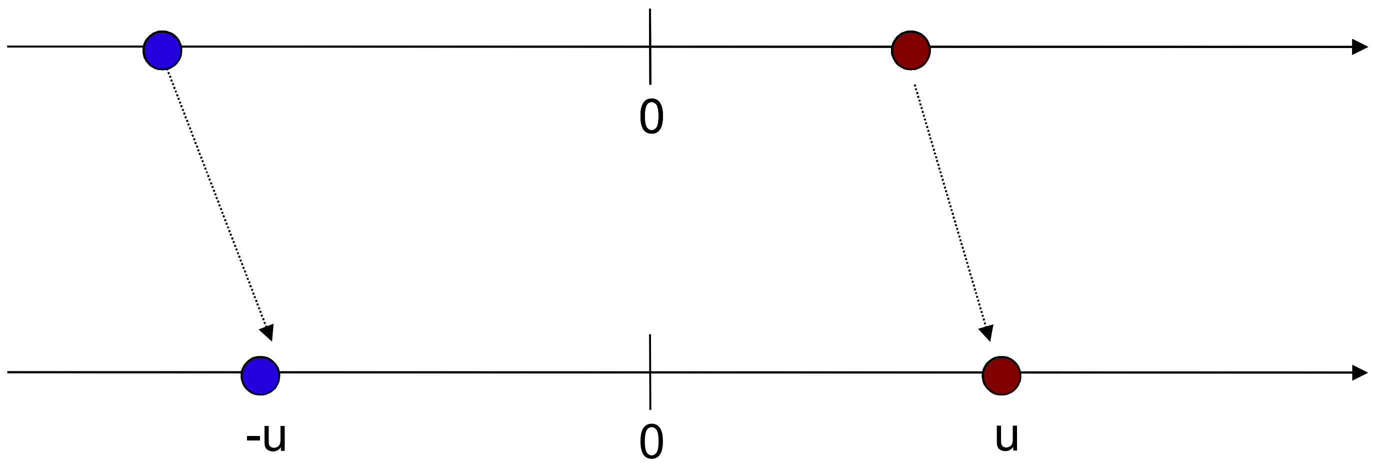
Signed Sum of Squares (SSS)

$q_1 + q_2 + q_3 = 0$

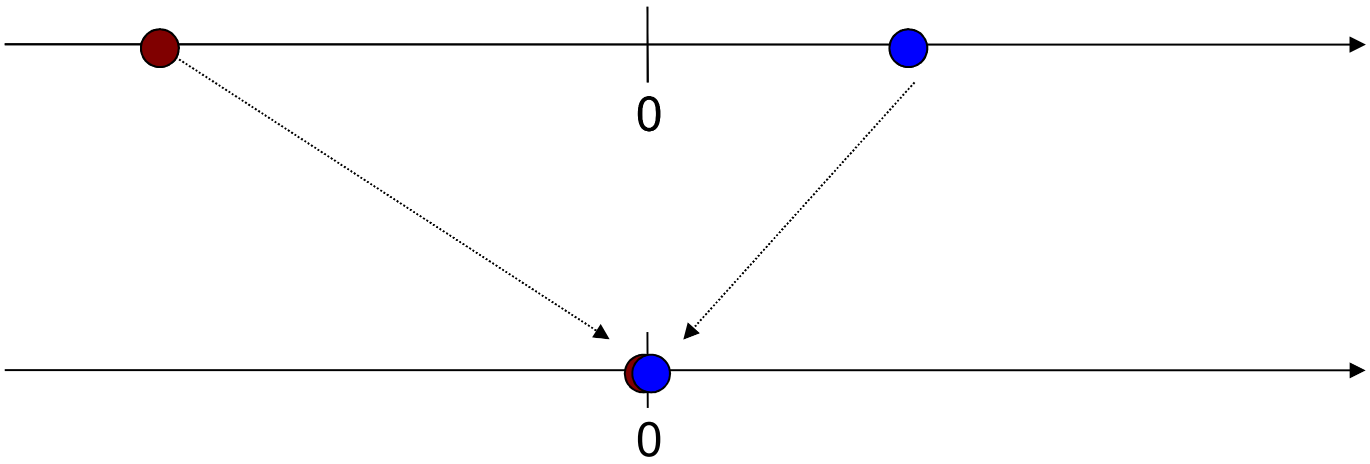


q_1	q_2	q_3
0	-u	u
v	0	-v
-w	w	0

q_1	q_2	q_3
*	**-u**	**u**
v	0	-v
-w	w	0



q_1	q_2	q_3
0	**-u**	**u**
v	0	-v
-w	w	0



u=0

Signed Sum of Squares (SSS)

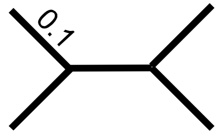
- Get three (really 2) numbers
- Consult the table of signs
- Pick the quartet tree that minimises the residual sum of squares (SSS)

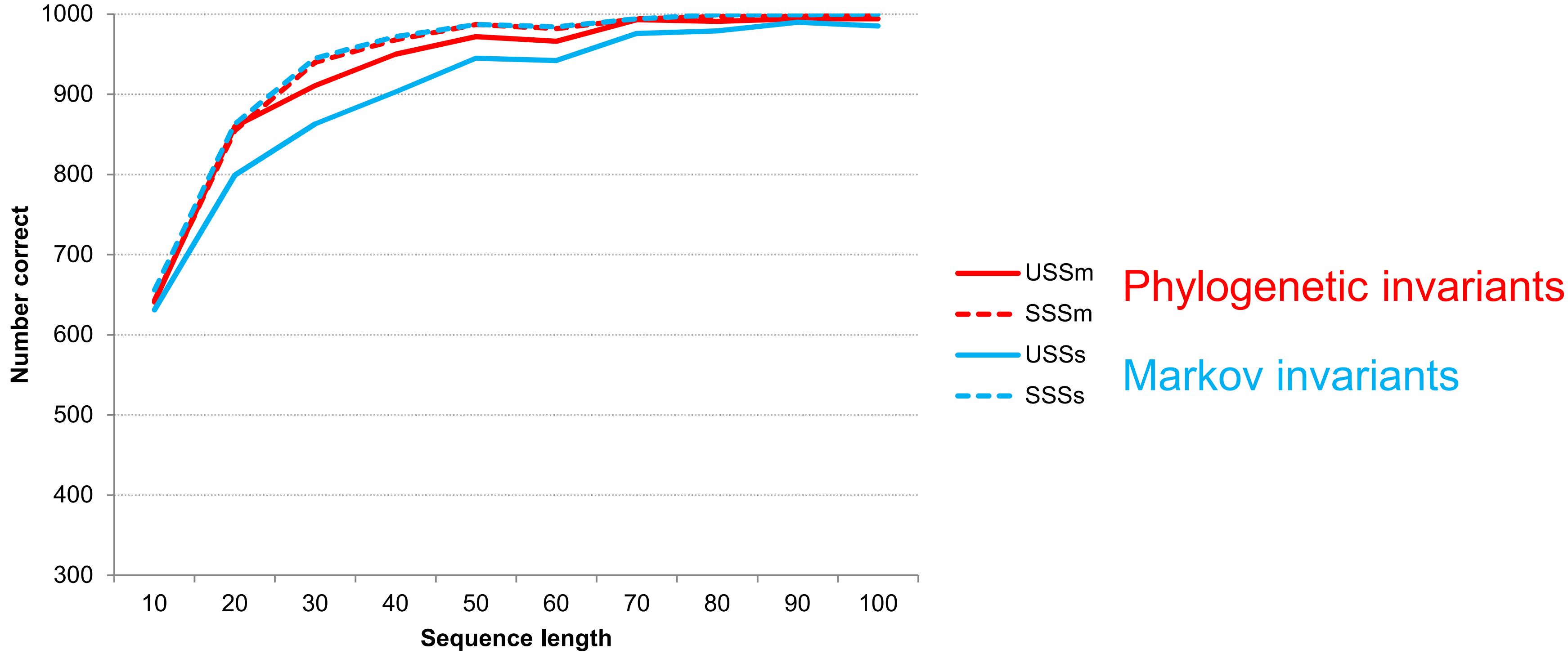
The contestants

- Phylogenetic Invariants (minors) with
  - Unsigned Sum of Squares (USSm)
  - Signed Sum of Squares (SSSm)

- Markov Invariants (squangles) with
  - Unsigned Sum of Squares (USSs)
  - Signed Sum of Squares (SSSs)

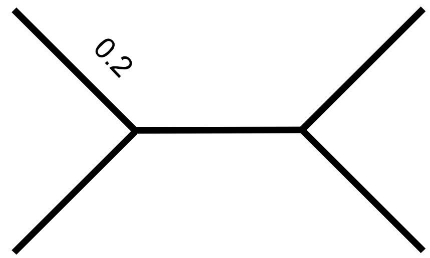
Simulation - 1

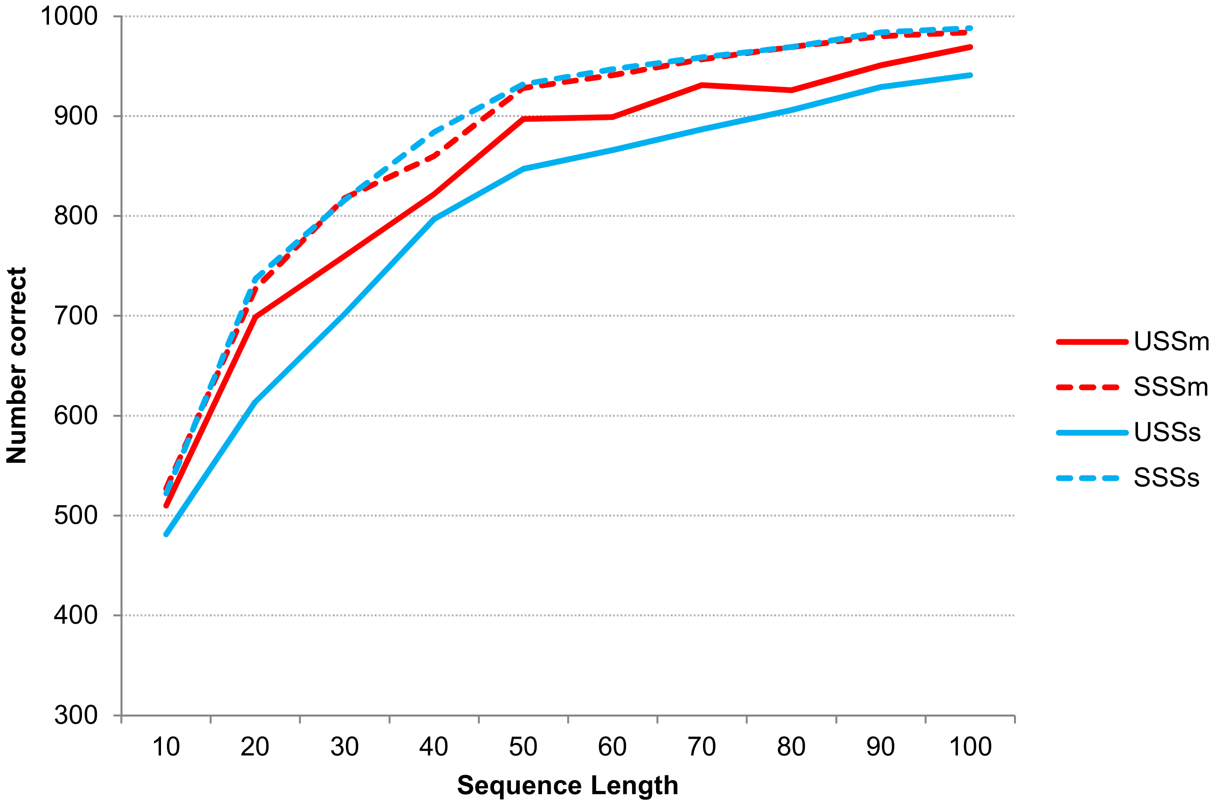




Tie between signed variants

Simulation - 2



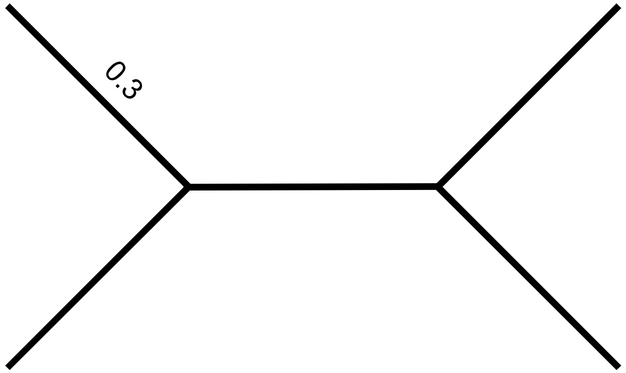


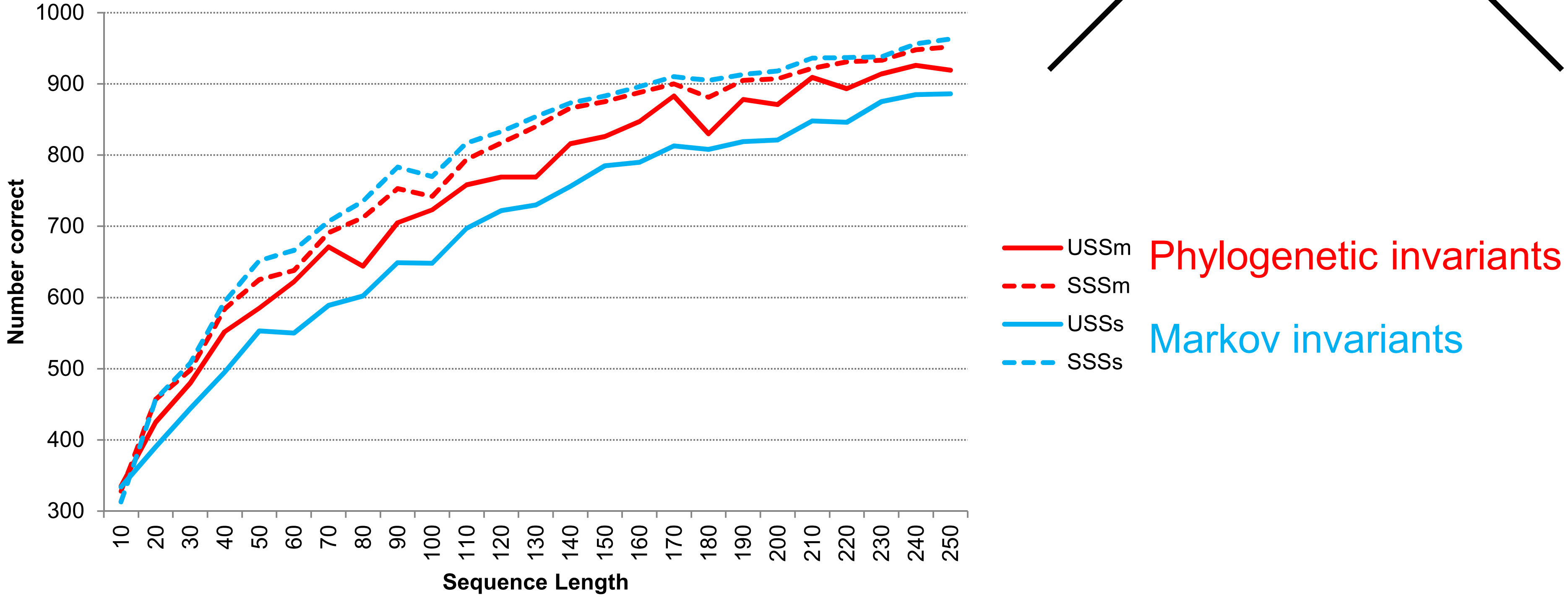
Phylogenetic invariants

Markov invariants

Narrow victory for signed squangles

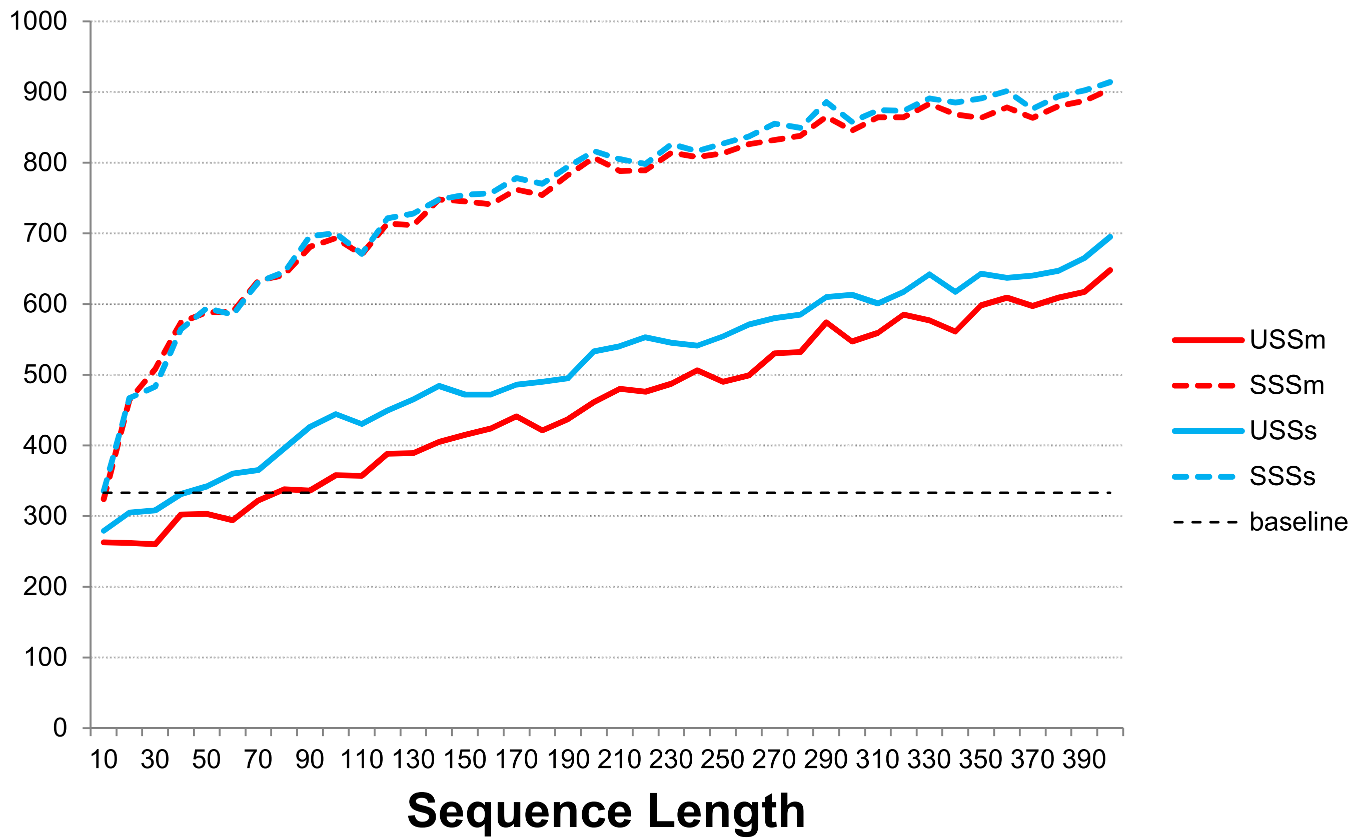
Simulation - 3

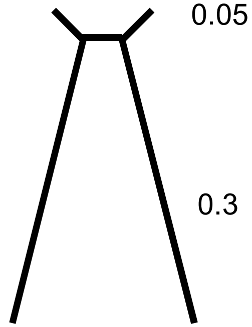




Narrow victory for signed squangles

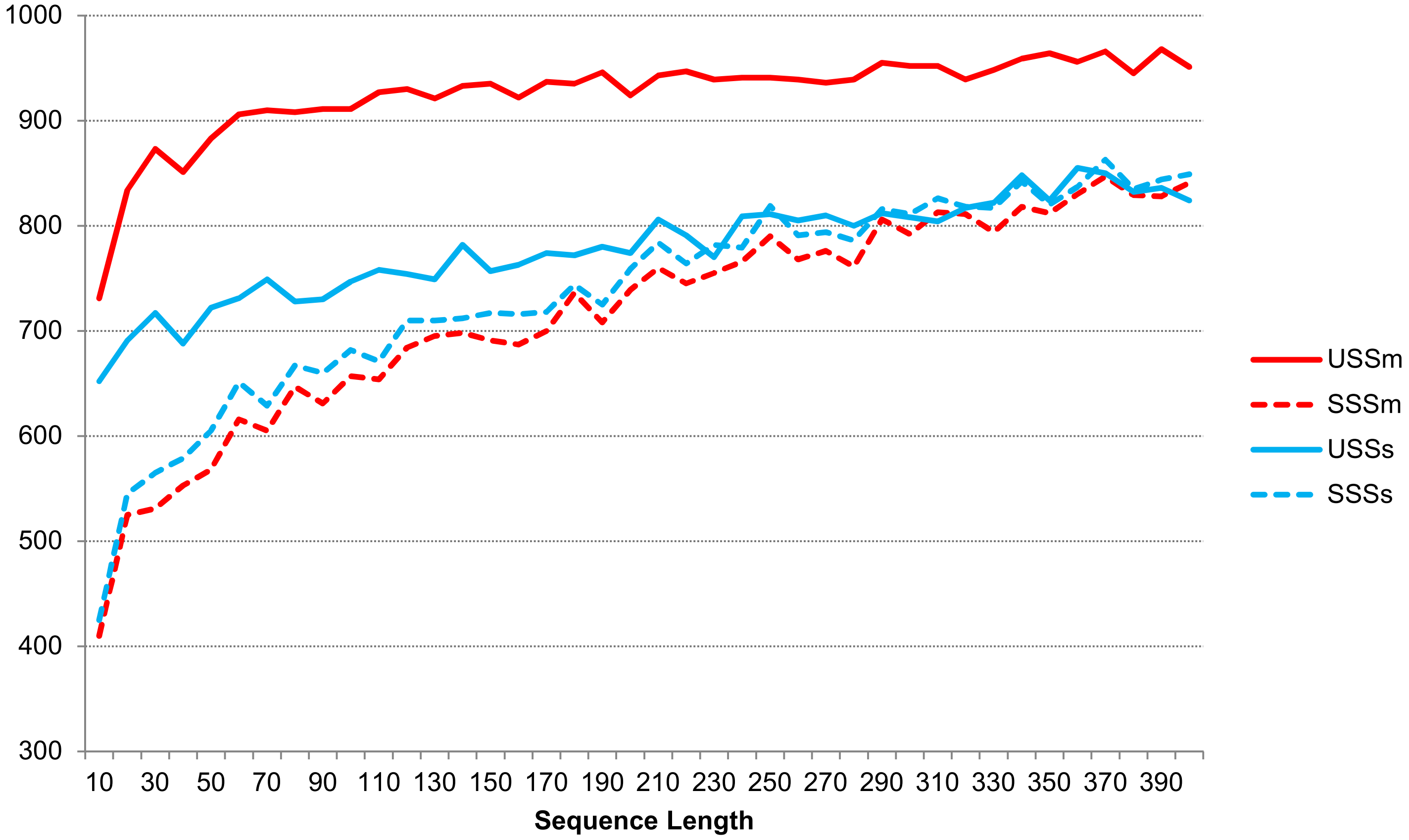
Simulation – Felsenstein Zone

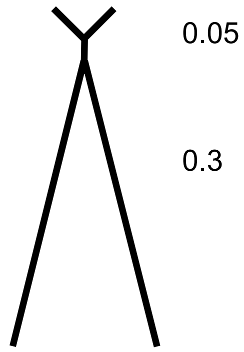




Narrow victory for signed squangles

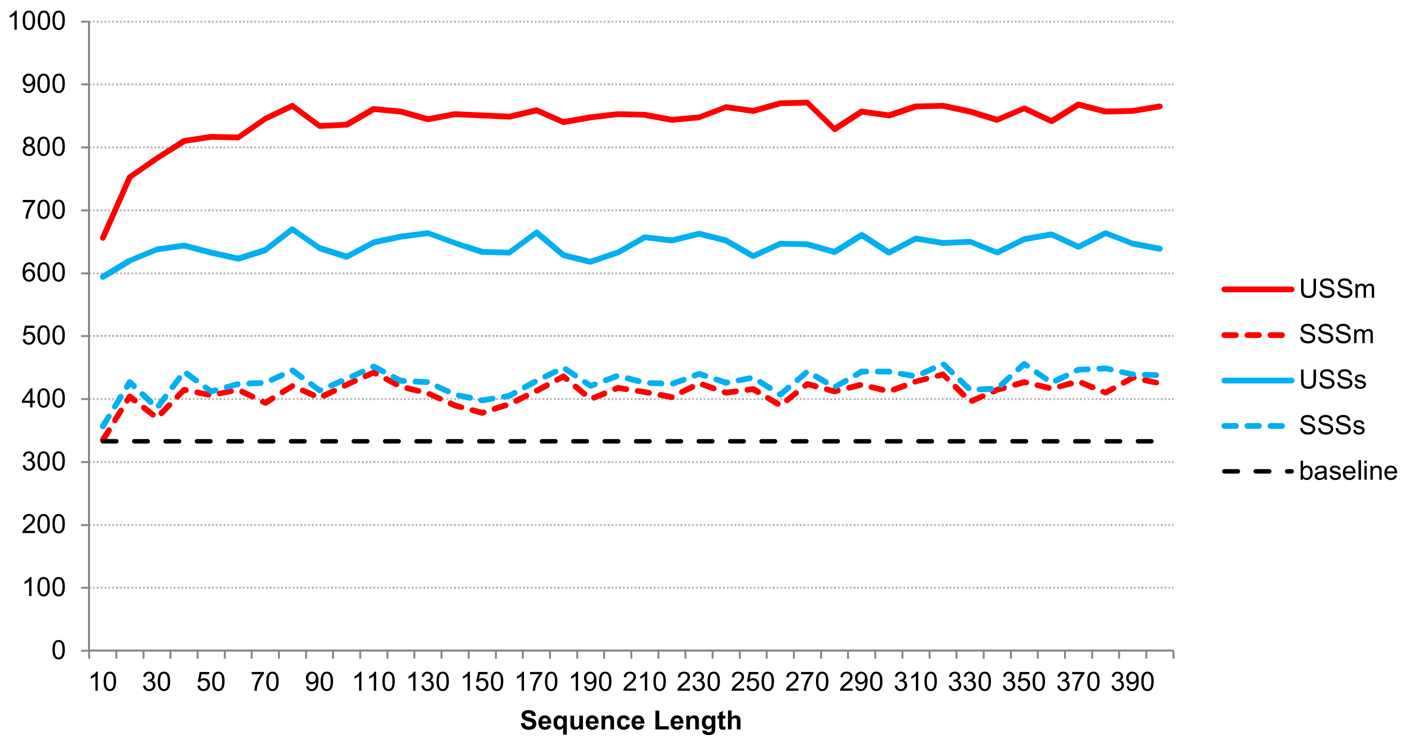
Simulation – Farris Zone

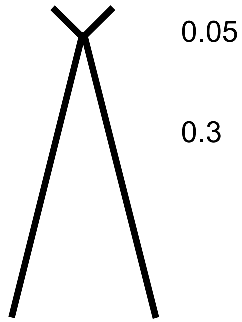




Glorious victory for unsigned minors?

Bias


1000
900
800
700
600
500
400
300
200
100
0
10 30 50 70 90 110 130 150 170 190 210 230 250 270 290 310 330 350 370 390
Sequence Length
USSm
SSSm
USSs
SSSs
baseline


0.05
0.3

More from minors?

- We get 16 minors (for each flattening). So far we’ve summed the Residual Sum of Squares over all 16.
- What if some minors are more **discriminating** than others?
- Are there sensible subsets of the minors to look at? How do minors “move around” under leaf permutations?

Flattenings

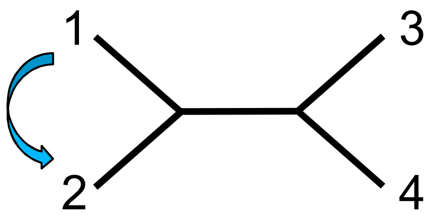
```
1234
0000 a
0001 b
0010 c
0011 d
0100 e
0101 f
0110 g
0111 h
1000 i
1001 j
1010 k
1011 l
1100 m
1101 n
1110 o
1111 p
```

$\xrightarrow[\text{12|34}]{\text{Flatten}}$

	00	01	10	11
00	a	b	c	D
01	e	f	g	h
10	i	j	k	l
11	m	n	o	p

Flattenings and Permutations

1234			2134	
0000	a		0000	a
0001	b		0001	b
0010	c		0010	c
0011	d		0011	d
0100	e		1000	e
0101	f		1001	f
0110	g		1010	g
0111	h	$\longrightarrow$	1011	h
1000	i		0100	i
1001	j		0101	j
1010	k		0110	k
1011	l		0111	l
1100	m		1100	m
1101	n		1101	n
1110	o		1110	o
1111	p		1111	p

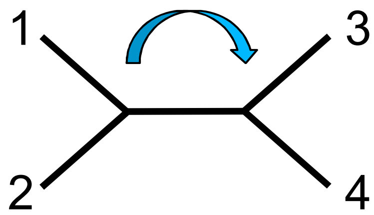


$\xrightarrow[12|34]{\text{Flatten}}$

	00	01	10	11
00	a	b	c	d
01	i	j	k	l
10	e	f	g	h
11	m	n	o	p

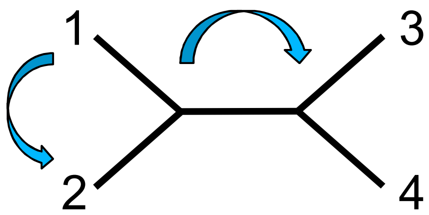
Flattenings and Permutations

1234		3412	
0000	a	0000	a
0001	b	0100	b
0010	c	1000	c
0011	d	1100	d
0100	e	0001	e
0101	f	0101	f
0110	g	1001	g
0111	h	1101	h
1000	i	0001	i
1001	j	0101	j
1010	k	1001	k
1011	l	1101	l
1100	m	0011	m
1101	n	0111	n
1110	o	1011	o
1111	p	1111	p



$\xrightarrow[12|34]{\text{Flatten}}$

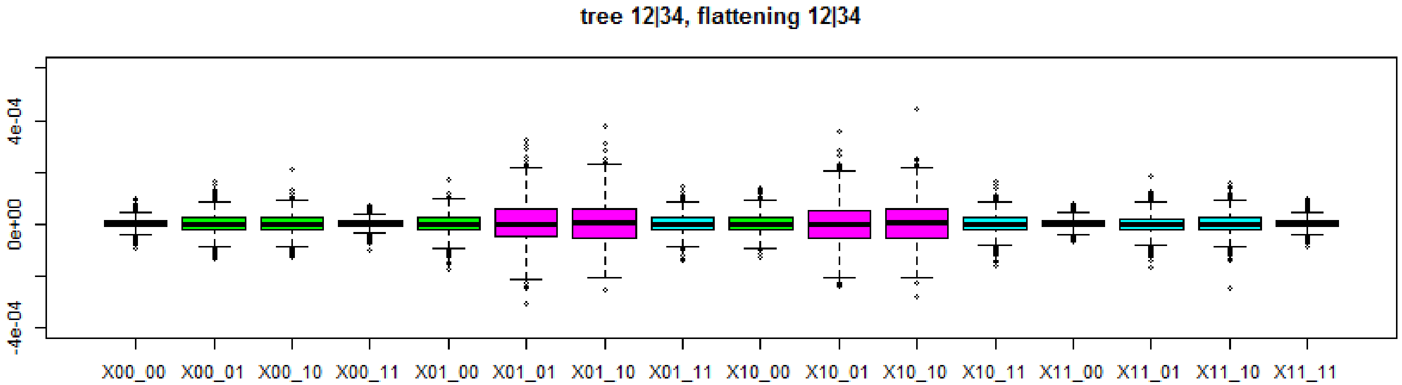
	00	01	10	11
00	a	e	i	m
01	b	f	j	n
10	c	g	k	o
11	d	h	l	p

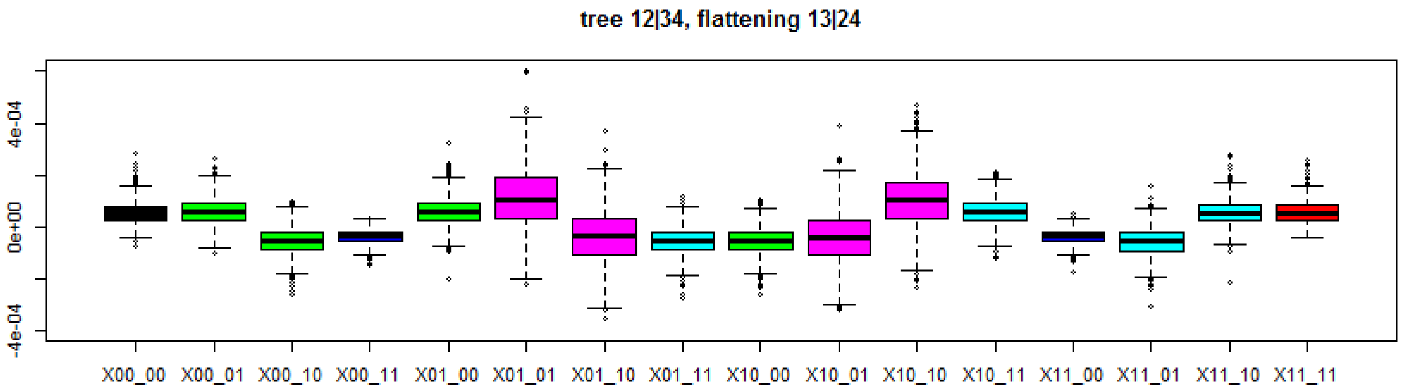


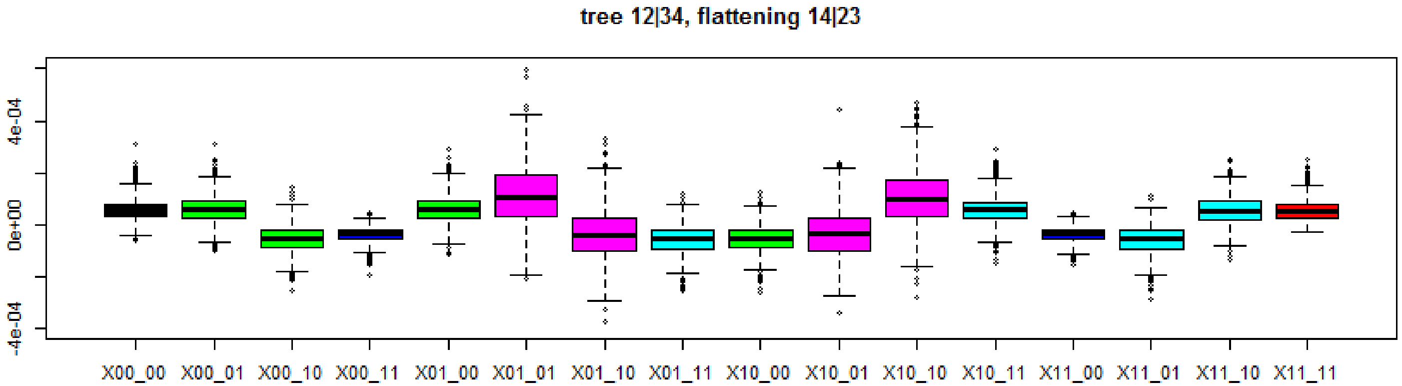
	00	01	10	11
00	a	e	i	m
01	b	f	j	n
10	c	g	k	o
11	d	h	l	p

	00	01	10	11
00	a	e	i	m
01	b	f	j	n
10	c	g	k	o
11	d	h	l	p

Accuracy	
a	0.819
p	0.822
dm	0.803
bcei	0.840
hlno	0.849
fgjk	0.829
sum	0.867



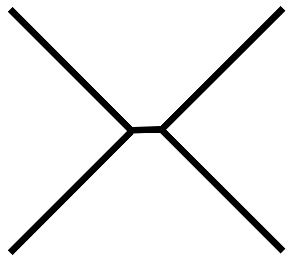


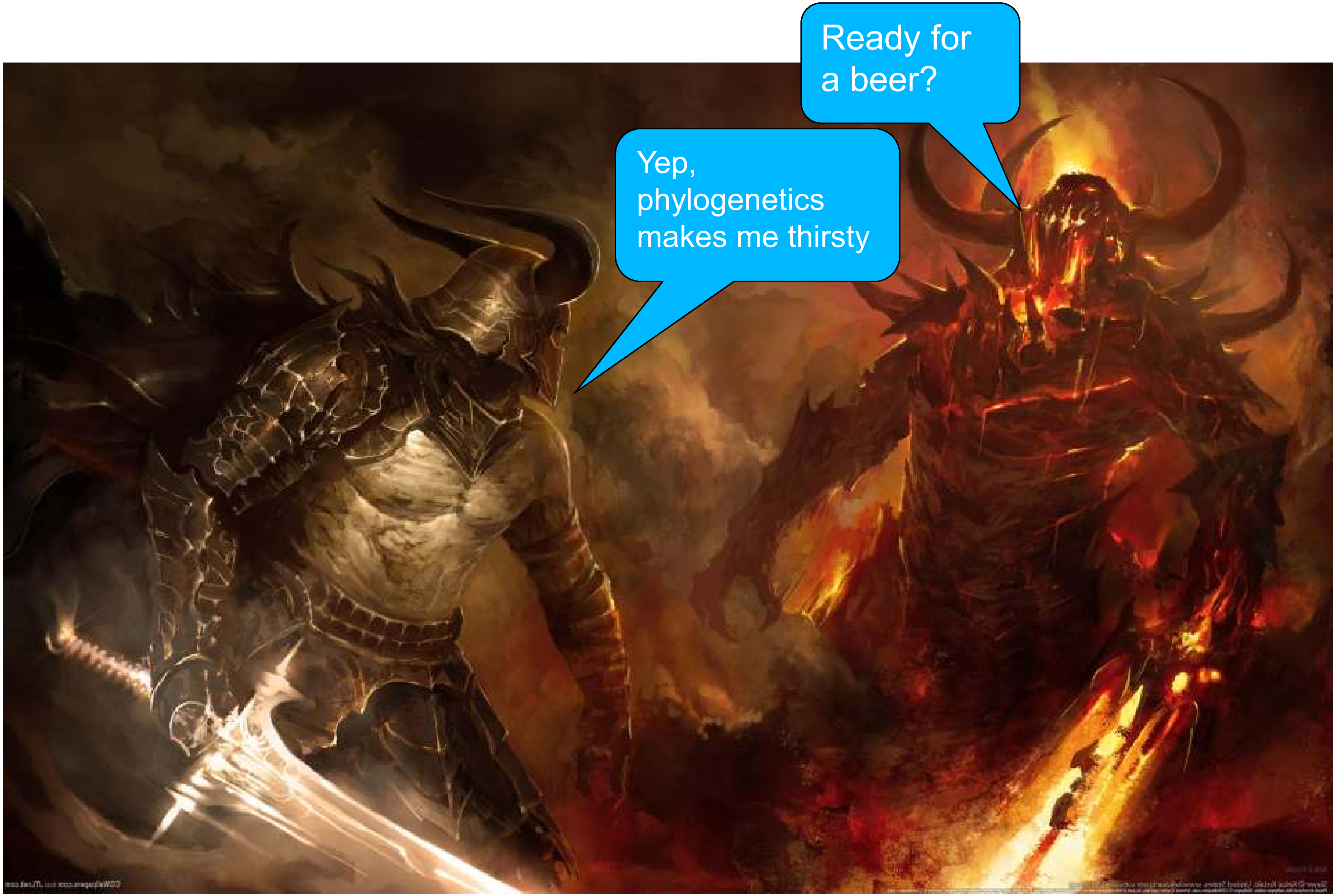


2 states to 4 states

- There are squangles for 4 state data
  - 5^{th} order polynomials with 66,744 terms
  - 0.81 secs / quartet
- Squangles for 2 state data are
  - 3^{rd} order polynomials with 6 terms
  - faster
- How effective is recoding followed by the simpler 2-state squangle?

Short edge	Long edge	4 state squangle accuracy	2 state squangle accuracy
0.01	0.2	0.80	0.61
0.01	0.4	0.42	0.38
0.02	0.2	0.96	0.84
0.02	0.4	0.48	0.40
0.04	0.2	1.00	0.95
0.04	0.4	0.65	0.69




Ready for a beer?
Yep, phylogenetics makes me thirsty