

# PoSeiDon

A Web Server for the Detection of Evolutionary  
Recombination Events and Positive Selection

---

Martin Hölzer

February 15, 2017

Friedrich Schiller University Jena

RNA Bioinformatics and High Throughput Analysis

32nd TBI Winterseminar in Bled

# Evolutionary Biology

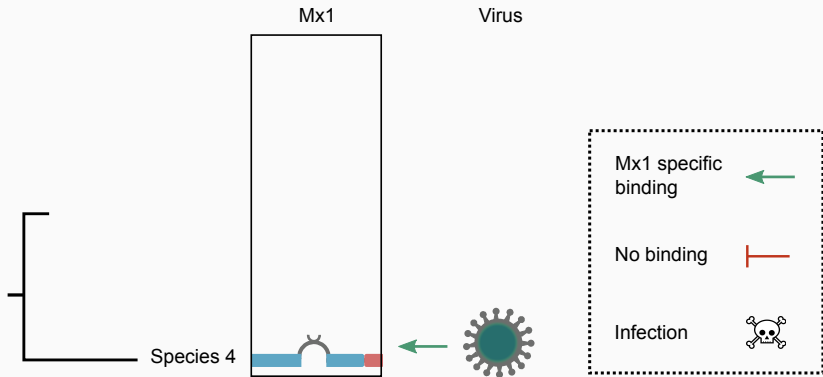
How do protein-coding genes evolve through time and in the various organisms that exist today?



# Positive Selection

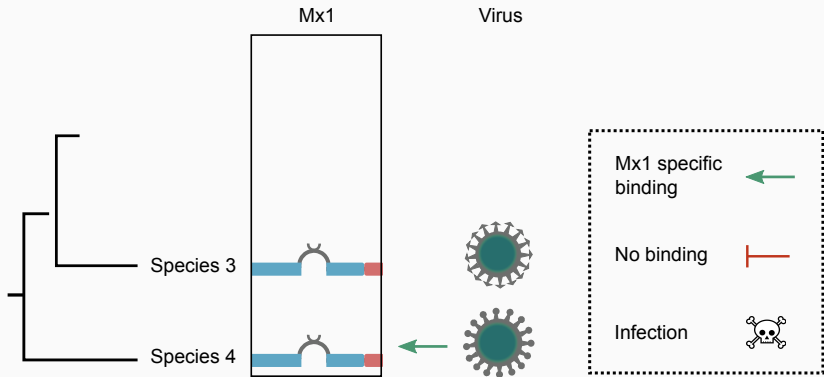
---

# 'Arms Race'

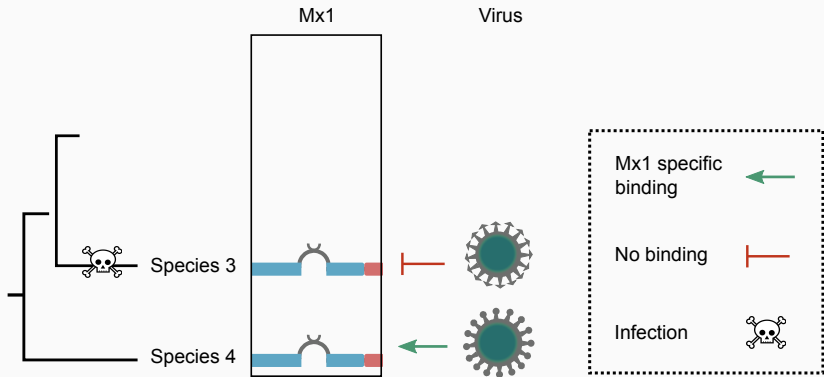




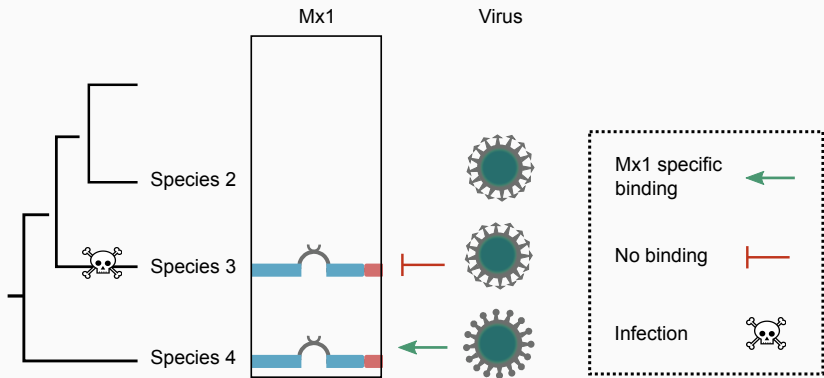
# 'Arms Race'



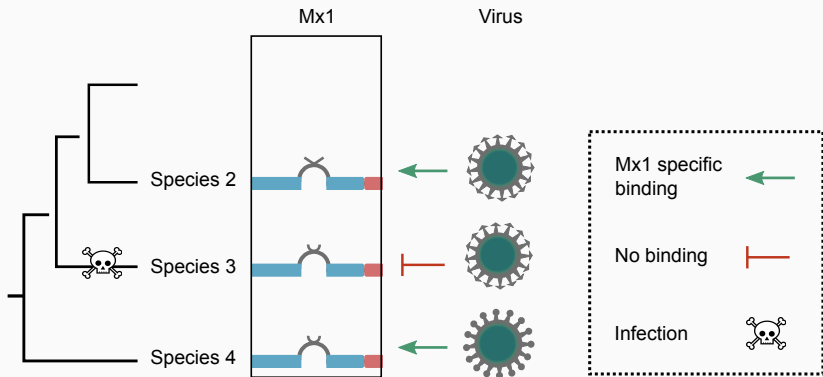
# 'Arms Race'



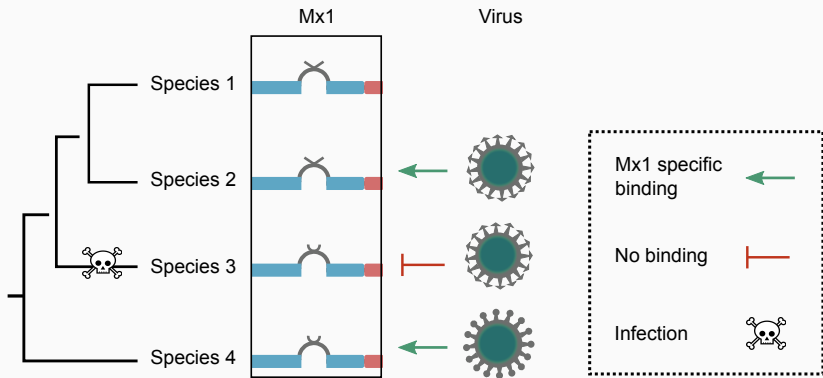
# 'Arms Race'



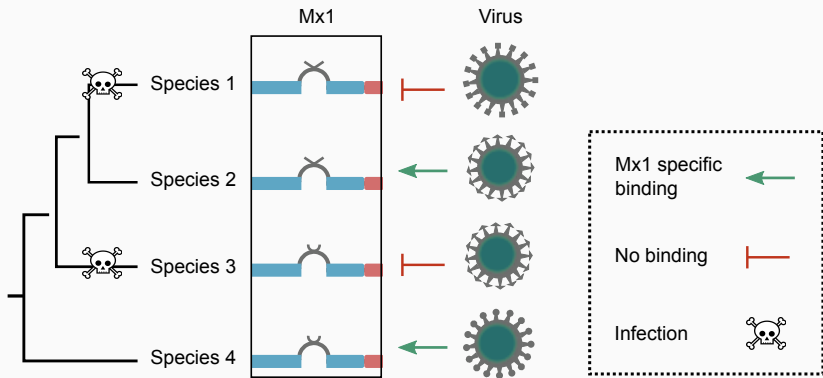
# 'Arms Race'



# 'Arms Race'



# 'Arms Race'



# Co-Evolution and detecting natural selection

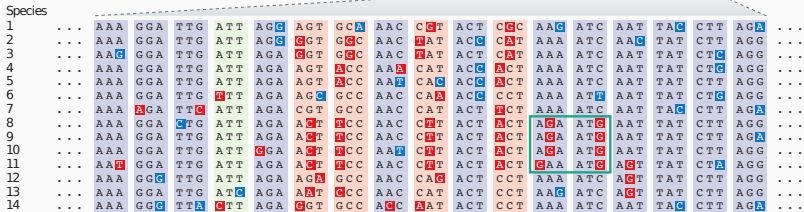
Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	...				
1	...	AAA	GGA	TTG	ATT	AGG	AGT	GCA	AAC	CST	ACT	CSC	AAG	ATC	AAT	TAC	CTT	AGA	...
2	...	AAA	GGA	TTG	ATT	AGG	GGT	GGC	AAC	TAT	ACC	CAT	AAA	ATC	AAAC	TAT	CTT	AGG	...
3	...	AA	GGA	TTG	ATT	AGA	GGT	GGC	AAC	TAT	ACT	CAT	AAA	ATC	AAT	TAT	CTC	AGG	...
4	...	AAA	GGA	TTG	ATT	AGA	AGT	ACC	AA	CAT	ACC	ACT	AAA	ATC	AAT	TAT	CTG	AGG	...
5	...	AAA	GGA	TTG	ATT	AGA	AGT	ACC	AA	CA	ACC	ACT	AAA	ATC	AAT	TAT	CTT	AGG	...
6	...	AAA	GGA	TTG	ATT	AGA	AGT	GCC	AAC	CA	ACC	CCT	AAA	ATC	AAT	TAT	CTG	AGG	...
7	...	AAA	AGA	TTT	ATT	AGA	CGT	GCC	AAC	CAT	ACT	CT	AAA	ATC	AAT	TAT	CTT	AGA	...
8	...	AAA	GGA	TTG	ATT	AGA	AGT	TCC	AAC	CT	ACT	ACT	AGA	ATC	AAT	TAT	CTT	AGG	...
9	...	AAA	GGA	TTG	ATT	AGA	AGT	TCC	AAC	CT	ACT	ACT	AGA	ATC	AAT	TAT	CTT	AGG	...
10	...	AAA	GGA	TTG	ATT	GGA	AGT	TCC	AA	CT	ACT	ACT	AGA	ATC	AAT	TAT	CTT	AGG	...
11	...	AA	GGA	TTG	ATT	AGA	AGT	TCC	AAC	CT	ACT	ACT	GAA	ATC	AGT	TAT	CTA	AGG	...
12	...	AAA	GGG	TTG	ATT	AGA	AGA	GCC	AAC	CA	ACT	CCT	AAA	ATC	AGT	TAT	CTT	AGG	...
13	...	AAA	GGA	TTG	ATT	AGA	AGT	GCC	AAC	CAT	ACT	CCT	AAA	ATC	AGT	TAT	CTT	AGG	...
14	...	AAA	GGG	TTA	CTT	AGA	GGT	GCC	AAC	CA	ACT	CCT	AAA	ATC	AAT	TAT	CTT	AGG	...

<b>Synonymous substitution</b> (no amino acid replacement)
<b>Non-synonymous substitution</b> (amino acid replacement)
$dN/dS < 1$ (negative selection)
$dN/dS \sim 1$ (neutral evolution)
$dN/dS > 1$ (positive selection)
<b>Lineage-specific selection</b> (episodic selection)

Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	...					
1	...	K	G	L	I	R	S	A	N	R	T	R	K	I	N	Y	L	R	...	
2	...	K	G	L	I	R	G	G	N	N	Y	T	H	K	I	N	Y	L	R	...
3	...	K	G	L	I	R	S	T	N	Y	T	H	K	I	N	Y	L	R	...	
4	...	K	G	L	I	R	S	T	K	H	T	T	K	I	N	Y	L	R	...	
5	...	K	G	L	I	R	S	T	N	H	T	T	K	I	N	Y	L	R	...	
6	...	K	G	L	F	R	S	A	N	Q	T	P	K	I	N	Y	L	R	...	
7	...	K	R	F	I	R	R	A	N	H	T	S	K	I	N	Y	L	R	...	
8	...	K	G	L	I	R	T	S	N	L	T	T	R	M	N	Y	L	R	...	
9	...	K	G	L	I	R	T	S	N	L	T	T	R	M	N	Y	L	R	...	
10	...	K	G	L	I	G	T	S	N	L	T	T	R	M	N	Y	L	R	...	
11	...	N	G	L	I	R	T	S	N	L	T	T	E	M	S	Y	L	R	...	
12	...	K	G	L	I	R	R	A	N	Q	T	P	K	I	S	Y	L	R	...	
13	...	K	G	L	I	R	N	P	N	H	T	P	K	I	S	Y	L	R	...	
14	...	K	G	L	L	R	G	A	T	N	T	P	K	I	S	Y	L	R	...	

Sironi, Manuela, et al. "Evolutionary insights into host-pathogen interactions from mammalian sequence data." *Nature Reviews Genetics* 16.4 (2015): 224-236.

# Co-Evolution and detecting natural selection



**Synonymous substitution**  
(no amino acid replacement)

**Non-synonymous substitution**  
(amino acid replacement)

$dN/dS < 1$   
(negative selection)

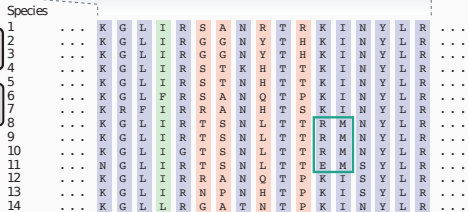
$dN/dS \sim 1$   
(neutral evolution)

$dN/dS > 1$   
(positive selection)

**Lineage-specific selection**  
(episodic selection)

$$dS = \frac{\text{\# syn substitutions}}{\text{\# syn sites}}$$

$$dN = \frac{\text{\# non-syn substitutions}}{\text{\# non-syn sites}}$$



Sironi, Manuela, et al. "Evolutionary insights into host-pathogen interactions from mammalian sequence data." *Nature Reviews Genetics* 16.4 (2015): 224-236.



# Co-Evolution and detecting natural selection

Species

1	...	AAA	GGA	TTG	ATT	AGG	AGT	GCA	AAC	CST	ACT	CSC	AAG	ATC	AAT	TAC	CTT	AGA	...
2	...	AAA	GGA	TTG	ATT	AGG	GGT	GGC	AAC	BAT	ACC	CAT	AAA	ATC	AAAC	TAT	CTT	AGG	...
3	...	AA	GGA	TTG	ATT	AGA	GGT	GGC	AAC	BAT	ACT	CAT	AAA	ATC	AAT	TAT	CTC	AGG	...
4	...	AAA	GGA	TTG	ATT	AGA	AGT	ACC	AA	CAT	ACC	ACT	AAA	ATC	AAT	TAT	CTG	AGG	...
5	...	AAA	GGA	TTG	ATT	AGA	AGT	ACC	AA	CA	ACC	ACT	AAA	ATC	AAT	TAT	CTT	AGG	...
6	...	AAA	GGA	TTG	ATT	AGA	AGT	GCC	AAC	CA	ACC	CCT	AAA	ATC	AAT	TAT	CTG	AGG	...
7	...	AAA	AGA	TTT	ATT	AGA	CGT	GCC	AAC	CAT	ACT	HCT	AAA	ATC	AAT	TAT	CTT	AGA	...
8	...	AAA	GGA	CTG	ATT	AGA	ACT	TCC	AAC	CT	ACT	ACT	AGA	ATC	AAT	TAT	CTT	AGG	...
9	...	AAA	GGA	TTG	ATT	AGA	ACT	TCC	AAC	CT	ACT	AGA	ATC	AAT	TAT	CTT	AGG	...	
10	...	AAA	GGA	TTG	ATT	GGA	ACT	TCC	AA	CT	ACT	AGA	ATC	AAT	TAT	CTT	AGG	...	
11	...	AA	GGA	TTG	ATT	AGA	ACT	TCC	AAC	CT	ACT	GAA	ATC	AGT	TAT	CTA	AGG	...	
12	...	AAA	GG	TTG	ATT	AGA	AGA	GCC	AAC	CA	ACT	CCT	AAA	ATC	AGT	TAT	CTT	AGG	...
13	...	AAA	GGA	TTG	ATT	AGA	AA	GCC	AAC	CAT	ACT	CCT	AAA	ATC	AGT	TAT	CTT	AGG	...
14	...	AAA	GG	TT	ATT	AGA	GGT	GCC	A	CA	ACT	CCT	AAA	ATC	AAT	TAT	CTT	AGG	...

**Synonymous substitution**  
(no amino acid replacement)

**Non-synonymous substitution**  
(amino acid replacement)

$dN/dS < 1$   
(negative selection)

$dN/dS \sim 1$   
(neutral evolution)

$dN/dS > 1$   
(positive selection)

**Lineage-specific selection**  
(episodic selection)

$$dS = \frac{\# \text{ syn substitutions}}{\# \text{ syn sites}}$$

$$dN = \frac{\# \text{ non-syn substitutions}}{\# \text{ non-syn sites}}$$

$$\omega = \frac{dN}{dS}$$

Species

1	...	K	G	L	I	R	S	A	N	R	T	R	K	I	N	Y	L	R	...		
2	...	K	G	L	I	R	G	G	N	Y	T	H	K	I	N	Y	L	R	...		
3	...	K	G	L	I	R	S	T	K	H	T	T	K	I	N	Y	L	R	...		
4	...	K	G	L	I	R	S	T	N	H	T	T	K	I	N	Y	L	R	...		
5	...	K	G	L	I	R	S	T	S	A	N	Q	T	P	K	I	N	Y	L	R	...
6	...	K	R	F	I	R	R	A	N	H	T	S	K	I	N	Y	L	R	...		
7	...	K	G	L	I	R	T	S	N	L	T	T	R	M	N	Y	L	R	...		
8	...	K	G	L	I	R	T	S	N	L	T	T	R	M	N	Y	L	R	...		
9	...	K	G	L	I	R	T	S	N	L	T	T	R	M	N	Y	L	R	...		
10	...	K	G	L	I	G	T	S	N	L	T	T	R	M	N	Y	L	R	...		
11	...	N	G	L	I	R	T	S	N	L	T	T	E	M	S	Y	L	R	...		
12	...	K	G	L	I	R	R	A	N	Q	T	P	K	I	S	Y	L	R	...		
13	...	K	G	L	I	R	N	P	N	H	T	P	K	I	S	Y	L	R	...		
14	...	K	G	L	L	R	G	A	T	N	T	P	K	I	S	Y	L	R	...		

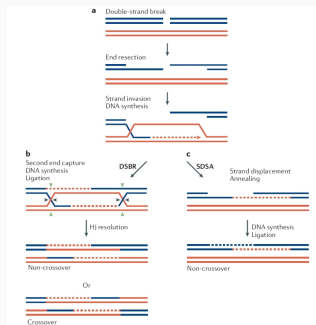
Sironi, Manuela, et al. "Evolutionary insights into host-pathogen interactions from mammalian sequence data." *Nature Reviews Genetics* 16.4 (2015): 224-236.

# Recombination

---

# Genetic Recombination

- Rearrangement of genetic information within and among DNA molecules
- If recombination, then possibly no unique tree topology can describe the evolutionary history of the whole sequence

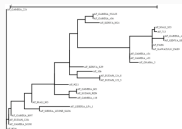
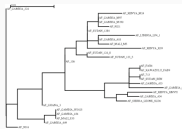
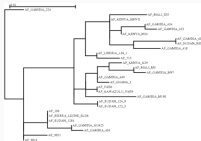
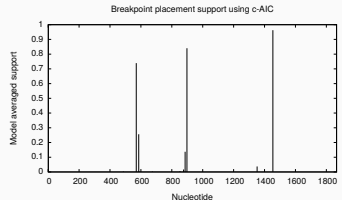


Clancy, S. (2008) "Genetic recombination." *Nature Education* 1(1):40

# Genetic Recombination

## GARD found evidence of 3 breakpoints

BP#	c-AIC	$\Delta$ c-AIC	Segments
0	7902.02		1-1866
1	7642.2	259.812	898   1454
2	7518.5	123.708	724   1454
3	7424.64	93.8553	570   898   1454
4	7424.64	0	570   898   1454

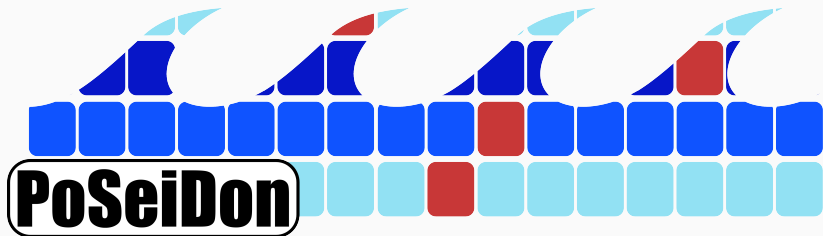


Pond, Sergei L. Kosakovsky, *et al.* "GARD: a genetic algorithm for recombination detection." *Bioinformatics* 22.24 (2006): 3096-3098.

## Previous Work

---





---

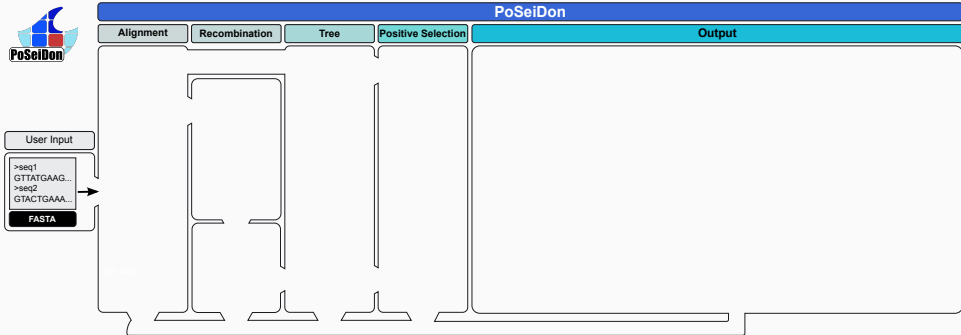
Hölzer, Martin and Marz, Manja "PoSeiDon: A Web Server for the Detection of Evolutionary Recombination Events and Positive Selection." Submitted as Applications Note.

# The Pipeline

---

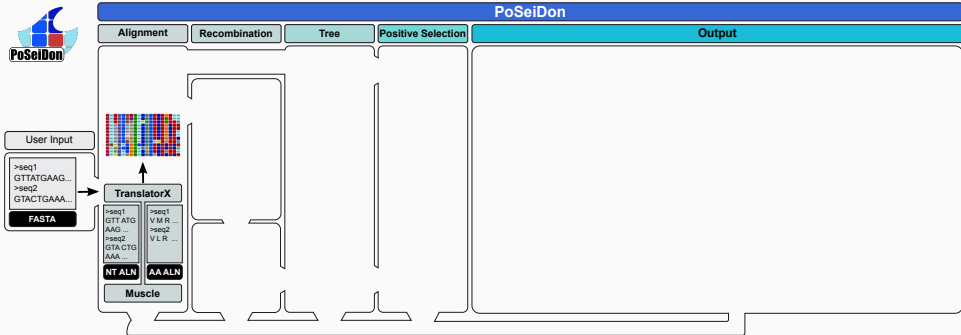


# The Pipeline



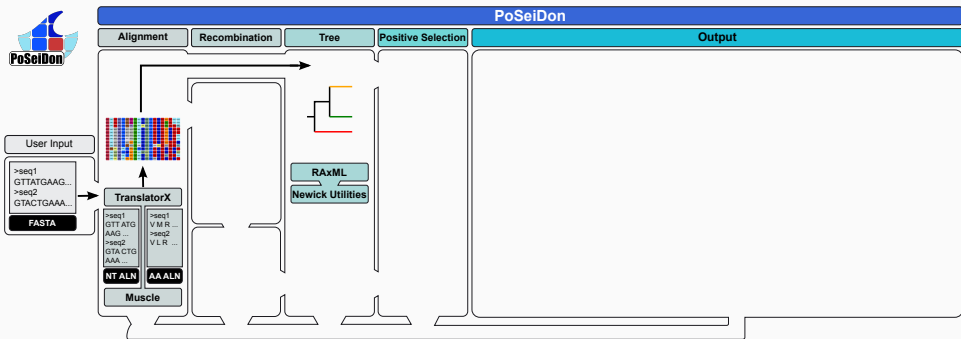
Hölzer, Martin and Marz, Manja "PoSeiDon: A Web Server for the Detection of Evolutionary Recombination Events and Positive Selection." Submitted as Applications Note.

# The Pipeline



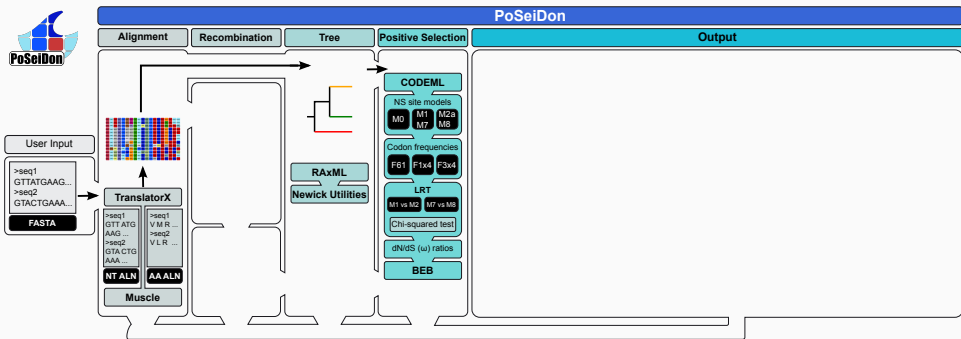
Abascal, Federico, Rafael Zardoya, and Maximilian J. Telford. "TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations." *Nucleic acids research* (2010): gkq291.

# The Pipeline



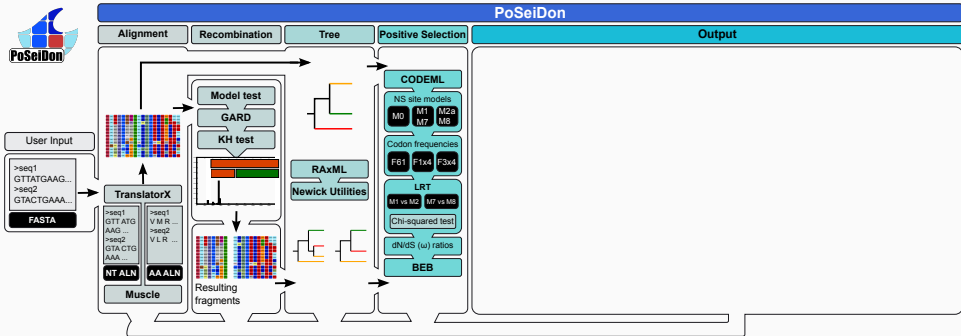
Stamatakis, Alexandros. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* 30.9 (2014): 1312-1313.

# The Pipeline



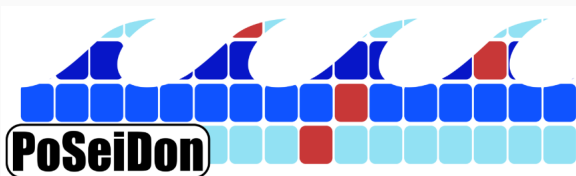
Yang, Ziheng. "PAML 4: phylogenetic analysis by maximum likelihood." *Molecular biology and evolution* 24.8 (2007): 1586-1591.

# The Pipeline



Pond, Sergei L. Kosakovsky, et al. "GARD: a genetic algorithm for recombination detection." *Bioinformatics* 22.24 (2006): 3096-3098.





## Positive Selection Detection and Recombination Analysis

Your E-Mail\*

Your name

Project title

Reference

Outgroup

Use also insignificant breakpoints:

Select File\*\*



Here we present **PoSeiDon**, a [pipeline](#) to detect significant positively selected sites and possible recombination events in an alignment of multiple coding sequences. Sites that undergo positive selection can give you insights in the evolutionary history of your sequences, for example showing you important mutation hot spots, accumulated as results of virus-host arms races during evolution.

**PoSeiDon** is easy to use: just provide your nucleotide coding sequences as one multiple FASTA file and enter your E-Mail address. After all calculations finished, **PoSeiDon** will send you a link to access all data. You can also provide optional information like one or multiple species as *Outgroup* (for tree drawing) and a name of a *Reference* species (amino acids of positive selected sites will be shown in relation to this species).





# Outlook

---

- improve interface
- support branch-site models
- upload your own tree/alignment
- distribute source code to run on local machines



- improve interface
- support branch-site models
- upload your own tree/alignment
- distribute source code to run on local machines



- [www.rna.uni-jena.de/poseidon](http://www.rna.uni-jena.de/poseidon)

- improve interface
  - support branch-site models
  - upload your own tree/alignment
  - distribute source code to run on local machines
  - check and include suggestions of the reviewers 😊
- 
- [www.rna.uni-jena.de/poseidon](http://www.rna.uni-jena.de/poseidon)



Thanks!



**Questions?**





*“Now, here, you see, it takes all the running you can do, to keep in the same place.”*

---

– Red Queen to Alice in Lewis Carroll's *“Through the Looking-Glass”*



# Run for your Lives!



*“Now, here, you see, it takes all the running you can do, to keep in the same place.”*

---

Iron Maiden et al. “The Number of the Beast (album).” EMI, *Battery Studios*, 39:11, London, England (1982)

## Calculating selective pressure, $K_a/K_s$

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes

## Calculating selective pressure, $K_a/K_s$

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - $K_a$ : number of non-synonymous substitutions per non-synonymous site

## Calculating selective pressure, $K_a/K_s$

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - $K_a$ : number of non-synonymous substitutions per non-synonymous site
  - $K_s$ : number of synonymous substitutions per synonymous site

## Calculating selective pressure, $K_a/K_s$

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - $K_a$ : number of non-synonymous substitutions per non-synonymous site
  - $K_s$ : number of synonymous substitutions per synonymous site
- homologous genes with a  $K_a/K_s$  ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**

## Calculating selective pressure, $K_a/K_s$

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - $K_a$ : number of non-synonymous substitutions per non-synonymous site
  - $K_s$ : number of synonymous substitutions per synonymous site
- homologous genes with a  $K_a/K_s$  ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**
- if all the mutations are neutral or disadvantageous, the ratio will be in the range 0 to 1

## Calculating selective pressure, $K_a/K_s$

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - $K_a$ : number of non-synonymous substitutions per non-synonymous site
  - $K_s$ : number of synonymous substitutions per synonymous site
- homologous genes with a  $K_a/K_s$  ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**
- if all the mutations are neutral or disadvantageous, the ratio will be in the range 0 to 1
- however, if some of the mutations are advantageous and some disadvantageous, the ratio could be less than 1

## Calculating selective pressure, $K_a/K_s$

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - $K_a$ : number of non-synonymous substitutions per non-synonymous site
  - $K_s$ : number of synonymous substitutions per synonymous site
- homologous genes with a  $K_a/K_s$  ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**
- if all the mutations are neutral or disadvantageous, the ratio will be in the range 0 to 1
- however, if some of the mutations are advantageous and some disadvantageous, the ratio could be less than 1
- $K_a/K_s = dN/dS = \omega$



## Calculating selective pressure, $K_a/K_s$

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - $K_a$ : number of non-synonymous substitutions per non-synonymous site
  - $K_s$ : number of synonymous substitutions per synonymous site
- homologous genes with a  $K_a/K_s$  ratio above 1 are evolving under positive selection
  - at least some of the mutations concerned must be **advantageous**
- if all the mutations are neutral or disadvantageous, the ratio will be in the range 0 to 1
- however, if some of the mutations are advantageous and some disadvantageous, the ratio could be less than 1
- $K_a/K_s = dN/dS = \omega$   $\kappa = Ts/Tv$

## Example: calculating $K_a/K_s$

	Val			Met			Arg			Thr		
G	T	T	A	T	G	A	A	G	A	C	C	Total

# Example: calculating $K_a/K_s$

		Val		Met		Arg		Thr			
	G	T	T	A	T	G	A	A	C	C	Total
degeneracy			(4)			(2)			(4)		

# Example: calculating $K_a/K_s$

	Val			Met			Arg			Thr			Total
	G	T	T	A	T	G	A	A	G	A	C	C	
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$

# Example: calculating $K_a/K_s$

	Val			Met			Arg			Thr			Total
	G	T	T	A	T	G	A	A	G	A	C	C	
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$\frac{9}{3}$
# syn sites	0	0	1	0	0	0	0	0	$\frac{1}{3}$	0	0	1	$\frac{2}{3}$

## Example: calculating $K_a/K_s$

	Val			Met			Arg			Thr			Total
	G	T	T	A	T	G	A	A	G	A	C	C	
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$
# syn sites	0	0	1	0	0	0	0	0	$\frac{1}{3}$	0	0	1	$2\frac{1}{3}$

- nucleotide sites give  $9\frac{2}{3}$  non-synonymous sites and  $2\frac{1}{3}$  synonymous sites in this peptide







# Example: calculating $K_a/K_s$

	Val			Met			Arg			Thr			Total
	G	T	T	A	T	G	A	A	G	A	C	C	
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$
# syn sites	0	0	1	0	0	0	0	0	$\frac{1}{3}$	0	0	1	$2\frac{1}{3}$

- nucleotide sites give  $9\frac{2}{3}$  non-synonymous sites and  $2\frac{1}{3}$  synonymous sites in this peptide

	Val			Leu			Arg			Thr			Total
	G	T	<b>A</b>	<b>C</b>	T	G	A	<b>A</b>	<b>A</b>	A	C	C	
# substitutions													
# non-syn	0	0	0	1	0	0	0	0	0	0	0	0	1
# syn	0	0	1	0	0	0	0	0	1	0	0	0	2

## Example: calculating $K_a/K_s$

Now,

$$K_a = \frac{\#nonsyn\_substitutions}{\#nonsyn\_sites} = \frac{1}{9\frac{2}{3}} = 0.103$$

and

$$K_s = \frac{\#syn\_substitutions}{\#syn\_sites} = \frac{2}{2\frac{1}{3}} = 0.857$$

Thus,

$$\frac{K_a}{K_s} = \frac{dN}{dS} = \omega = \frac{0.103}{0.857} = 0.12$$

- *CodonFreq=F3x4* (estimation of codon frequency distribution)
- *NSsites=M2a* (positive selection)

Model	NSsites	$p$	Parameters
M0 (one ratio)	0	1	$\omega$
M1a (neutral)	1	2	$p_0$ ( $p_1 = 1 - p_0$ ), $\omega_0 < 1, \omega_1 = 1$
M2a (selection)	2	4	$p_0, p_1$ ( $p_2 = 1 - p_0 - p_1$ ), $\omega_0 < 1, \omega_1 = 1, \omega_2 > 1$
M3 (discrete)	3	5	$p_0, p_1$ ( $p_2 = 1 - p_0 - p_1$ ) $\omega_0, \omega_1, \omega_2$
M7 (beta)	7	2	$p, q$
M8 (beta& $\omega$ )	8	4	$p_0$ ( $p_1 = 1 - p_0$ ), $p, q, \omega_s > 1$

NOTE.—The site models are implemented using the control variable NSsites in CODEML, and  $p$  is the number of free parameters in the  $\omega$  distribution.

⇒ [codeml.variable.mlc]