# 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'



#### Co-Evolution and detecting natural selection

Species					peries																						
1		AAA	GGA	TTG	ATT	AGC	AGT	GCA	AAC	COT	AC	т.	C	EC.	AA	а 🤉	ATC	A	ΑТ	т	A	C '	тт	AG	A		
2		AAA	GGA	TTG	ATT	AGG	GGT	GIEC	AAC	TAT	AC	C	C	т	AA	A	ATC	A	A	т	ΑТ	C	тт	AG	G		
3			GGA	TTG	ΔΨΨ	AGA	GT	GIEC	AAC	TAT	AC	TT T	C	TT.	2 2	a 1	A T C	Δ	ΔΨ	T	ΔΨ	<u> </u>	T	A (	G		
4			GGA	TTG	ΔΨΨ	AGA	AGT	ACC		CAT	AC		N	- m	2 2	a 1	A T C	Δ	ΔΨ	T	ΔΨ	<u> </u>	T	A (	G		
5		AAA	GGA	TTG	ATT	AGA	AGT	ACC	AAD	CAG	AC	C	A	C T	AA	A	ATC	A	АТ	T	АТ	C	TΤ	AC	G		
6		AAA	GGA	TTG	TT	AGA	AG	GCC	AAC	CAN	AC	C	CO	C T	AA	A	АТП	A	АТ	т	АТ	C	те	AG	G		
7		AAA	AGA	тт	ATT	AGA	CGT	GCC	AAC	CAT	AC	т	m	C T	AA	A	ATC	A	АТ	т	AC	C	тт	AG	A		
8		AAA	GGA	T G	ATT	AGA	A	TCC	AAC	CIT	AC	. т	A	C T	AG	A J	ATC	A	АТ	т	ΑТ	C	тт	AG	G		
9		AAA	GGA	TTG	ATT	AGA	AOT	TC C	AAC	CIT	AC	т	A	СТ	AG	A	ATC	A	AT	т	AT	C	тт	AG	A		
10		AAA	GGA	TTG	ATT	GGA	AOT	n c c	AA	CIT	AC	т	A	СТ	AG	A	ATC	Α.	АТ	т.	АТ	C	тт	AG	G		
11		AA	GGA	TTG	ATT	AGA	AOT	n c c	AAC	CIT	AC	т	A	СТ	GA	A	ATC	A	GТ	т.	АТ	C	ТА	AG	G		
12		AAA	GGG	TTG	ATT	AGA	AGA	GCC	AAC	CAC	AC	т	CO	СТ	AA	A 2	ATC	А	Gт	т.	АТ	C	тт	AG	G		
13		AAA	GGA	TTG	ATC	AGA	AAT	CCC	AAC	CAT	AC	т	CO	СТ	AA	G 1	ATC	А	Gт	т.	АТ	C	тт	AG	G		
14		AAA	G G G	ттА	СТТ	AGA	GGT	GCC	ACC	AAT	ΑC	т	CO	СТ	ΑA	A A	АТС	A.	ΑТ	т.	AC	C !	тт	AG	А		
			_	٦				Spec	ies	1111																	
Synony	ymous si	ubstituti	ion					1		v	C	т	т	D	c		J D	m	D	v	т	M	v	т	D		
(no am	ino acid	replace	ement)					2		· · ·	c	T	T	D	c .			T T	LI LI	v	Ť	N	v	T	D		
Non-sy	nonvm	nus subs	titution					â		K K	G	T.	T	R	G		J V	Ť	н	K	Ť	N	v	T.	R		
(amino	acid ret	placeme	ent)					4		ĸ	G	т.	T	P	s	r F	х н	Ť	T	ĸ	Ť	N	v	T.	R	• •	
			,					5		ĸ	G	т.	T	P	s	T D	л н	Ť	Ť	ĸ	Ť	N	v	T.	R	• •	
dN/dS	<1							6		ĸ	G	т.	F	P	s			Ť	P	ĸ	Ť	N	v	T.	R	• •	
inegati	ve selec	tion)						7		. K	R	F	T	R	R	A	J H	Ť	ŝ	ĸ	Ť	N	Ŷ	T.	R		. :
dN/dS	~1							8		ĸ	G	т.	т	R	T	5 1	а т.	т	T	R	M	N	Y	T.	R		
íneutra	al evolut	ion)						9		ĸ	G	т.	т	R	T	5 1	а т.	т	T	R	м	N	Y	T.	R		
18.11.107								10		ĸ	G	т.	т	G	T	5 1	а т.	т	T	R	м	N	Y	T.	R		
aiv/as	>1							11		N	G	т.	т	R	T	5 1	а т.	т	T	E	м	s	Y	T.	R		
(positiv	/e select	lion)						12		K	G	L	I	R	R	AP		т	P	K	I	s	Y	L	R		
Lineag	e-specif	ic selec	tion					13		. к	G	L	I	R	N	P	N H	т	Р	K	I	S	Y	L	R		
(episod	dic selec	tion)						14		. к	G	L	L	R	G.	A	r N	т	Р	K	I	N	Y	L	R		
Lineag (episod	e-specif dic selec	ic selec tion)	tion					13 14		. K	G G	L L	I L	R R	N G	P N A 1	H C N	T T	P P	K K	I I	s N	Y Y	L L	R 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

CTT AG CTT AG CTC AG	G A G G G G
CTT AG CTC AG	G G G G G G G G G G G G G G G G G G G
CTC AG	GG
CTG AG	
	C C
CTT AG	66
CTC AC	66
CTT AG	GN
CTT AG	66
CTT AG	GE
CTT AG	GG
CTA AG	G G G G G G G G G G G G G G G G G G G
CTT AG	G G
CTT AG	G G
CTT AG	G 🗛 🛛
NVT	P
N V L	R
NYL	R
SYL	R
SYL	R
S Y L	R
N Y L	R
	C T T A C C C T T A C C T

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	iperies																										
1		AAA	GGA	TTG	ATT	AGG	AGT	GCA	AAC	CGT	AC	ст	C	C	AA	A	тс	Α.	ΑТ	Т	AC	C	гт	AG	A		
2		AAA	GGA	TTG	ATT	AGG	GGT	GC	AAC	AT	AC		C	т	AAA	A	тC	Α.	AC	т	АТ	C 1	гт	AG	G		
3		AAG	GGA	TTG	ATT	AGA	GGT	GCC	AAC	ТАТ	AC	ст	C	т	AAA	A	тC	Α.	ΑТ	т	АТ	C S		AG	G		
4		AAA	GGA	TTG	ATT	AGA	AGT	ACC	AAA	CAT	AC	C	A	ст	AAA	A	тC	Α.	ΑТ	т	АТ	C S	r G	AG	G		
5		AAA	GGA	TTG	АТТ	AGA	AGT	ACC	AAT	CAC	AC	C	A	СТ	AAA	A	тС	Α.	ΑТ	т	АТ	C S	гт	AG	G		
6		AAA	GGA	ТТG	ттт	AGA	A G 🖸	GCC	AAC	CAA	AO	С	CO	ст	AAA	A	тт	Α.	ΑТ	т	ΑТ	C 1	P G	AG	G		
7		AAA	AGA	т т С	ΑΤΤ	AGA	CGT	GCC	AAC	CAT	AO	ст	т	СТ	AAA	A	ΤС	Α.	ΑТ	т	AC	C 1	гт	AG	А		
8		AAA	GGA	CTG	ΑΤΤ	AGA	ACT	TCC	AAC	СТТ	AO	ст	A (	СТ	AGA	A	ΤG	Α.	ΑТ	т	ΑТ	C 1	гт	AG	G		
9		AAA	GGA	ТТG	ΑΤΤ	AGA	ACT	TCC	AAC	СТТ	AO	ст	A (	СТ	AGA	A	ТG	Α.	ΑТ	т	ΑТ	C 1	гт	AG	А		
10		AAA	GGA	ТТG	ΑΤΤ	GGA	ACT	TCC	A A 🖪	СТТ	AO	ст	A (	СТ	AGA	A	ТG	Α.	ΑТ	т	ΑТ	C 1	гт	AG	G		
11		AA 🔳	GGA	ТТG	ΑΤΤ	AGA	ACT	TCC	AAC	СТТ	AO	ст	A (	СТ	GAA	A	ТG	A	GT	т	ΑТ	C 1	ΡA	ΑG	G		
12		AAA	G G G	ΤΤG	ΑΤΤ	AGA	AGA	GCC	AAC	CAG	ΑC	ст	CO	СТ	AAA	A	ΤС	А	GT	т	ΑТ	C 1	гт	AG	G		
13		AAA	GGA	ΤΤG	A T C	AGA	AAT	CCC	AAC	CAT	AC	ст	CO	СТ	AA	A	ΤС	А	GΤ	т	ΑТ	C 1	гт	ΑG	G		
14		AAA	G G 🖸	ттА	СТТ	AGA	GGT	GCC	ACC	AAT	A C	ст	CO	СТ	AAA	A	ΤС	Α.	ΑΤ	т	AC	C :	гт	AG	А		
C		de antes as		7				Spec	ies																		
Synon	ymous su	roplace	(on (mont)					1		. K	G	L	Ι	R	SI	N	R	т	R	K	I	N	Y	L	R		
(IIU al I	IIIO aciu	replace	erneriu)	1	# syr	n substi	utions	2		. к	G	L	I	R	G	N	Y	т	н	к	I	N	Y	L	R		
Non-sy	/nonymc	ous subs	titution	as	=	≠ syn sit	es	3		. к	G	L	I	R	G	N	Y	т	Н	к	I	N	Y	L	R		
(amino	acid rep	placeme	ent)					<b>J</b> 4		. к	G	L	I	R	S 1	K	Н	т	т	к	I	N	Y	L	R		
dN/dS	<1							5		. к	G	L	I	R	S 1	N	н	т	т	ĸ	I	N	Y	L	R		
inenati	ive selec	tion)			# non-s	yn subs	titutions	6		. K	G	L	F	R	SZ	N	Q	т	Ρ	к	I	N	¥	L	R		
megue	10 00100			dN=	# n	on-syn s	ites	7		. K	R	F	Ι	R	R I	N	Η	т	s	K	I	Ν	¥	L	R		
dN/dS	~1							<b>J</b> 8		. K	G	L	Ι	R	Т	N	L	т	т	R	М	N	¥	L	R		
(neutra	al evoluti	ion)			_		_	9		. K	G	L	Ι	R	Т	N	L	т	т	R	М	N	¥	L	R		
dN/dS	>1				1	d١	1	10		. K	G	L	I	G	тя	N	L	т	т	R	М	N	¥	L	R		
loositiv	ve select	ion)			$\omega$	=	-	11		. N	G	L	I	R	T S	N	L	т	т	Е	М	s	¥	L	R		•
						as		12		. к	G	L	I	R	R A	N	Q	т	Ρ	K	I	s	Y	L	R	• •	•
Lineag	le-specif	IC Selec	uon					13		. к	G	L	I	R	NI	N	H	т	Ρ	K	I	s	Y	L	R	• •	•
episod	aic selec	uon)						14		. K	G	L	L	R	G I	T	N	т	Ρ	K	I	N	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**



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

**Previous Work** 

#### Positive Selection in bat MX1



Fuchs, Jonas et al., "Evolution and antiviral specificity of interferon-induced Mx proteins of bats." Submitted.

#### **Positive Selection Detection and Recombination Analysis**



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



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



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



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



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



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



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

#### The Interface



#### Positive Selection Detection and Recombination Analysis

Your E-Mail*	
Your name	
Project title	
Reference	species_1
Outgroup	species_1,species_2
Use also insignifica	ant breakpoints:
Select File**	
MOORLAND	WOOD MOOR
Geben Sie den an	ezeigte

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).

#### The Output



# 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



- 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 <sup>(2)</sup>



• www.rna.uni-jena.de/poseidon



# Thanks!



# Questions?

#### **Backup slides**



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

<sup>-</sup> 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)

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

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

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - *K*<sub>a</sub>: number of non-synonymous substitutions per non-synonymous site
  - K<sub>s</sub>: number of synonymous substitutions per synonymous site

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - *K*<sub>a</sub>: 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

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - *K*<sub>a</sub>: 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

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - *K*<sub>a</sub>: 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

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - *K*<sub>a</sub>: 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
- $K_a/K_s dN/dS \omega$

- $K_a/K_s$  ratio is an indicator of selective pressure acting on protein-coding genes
  - *K*<sub>a</sub>: 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

• 
$$K_a/K_s - dN/dS - \omega$$
  $\kappa - Ts/Tv$ 

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



		Val			Met			Arg	5		Thr		
	G	Т	Т	А	Т	G	А	Α	G	А	С	С	Total
degeneracy			(4)						(2)			(4)	
# non-syn sites	1	1	0	1	1	1	1	1	$\frac{2}{3}$	1	1	0	$9\frac{2}{3}$

		Val			Met			Arg			Thr		
	G	Т	Т	Α	Т	G	А	А	G	А	С	С	Total
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	$\frac{1}{2}$	0	0	1	$2\frac{1}{2}$

		Val			Met			Arg	5		Thr		
	G	Т	Т	А	Т	G	А	А	G	А	С	С	Total
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			Met			Arg			Thr		
	G	Т	Т	А	Т	G	А	А	G	А	С	С	Total
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	$\frac{1}{2}$	0	0	1	$2\frac{1}{2}$

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

 Val
 Leu
 Arg
 Thr

 G
 T
 A
 C
 T
 G
 A
 A
 C
 T otal

	Vai				iviet				Arg	5			Inr		
	G	Т	Т	А	Т	G		A	А	G		А	С	С	Total
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	13		0	0	1	$2\frac{1}{3}$
•	nucleo peptio	otide de	sites gi	ve 9 <u>2</u> n	ion-sy	nony	mous	sit	es an	$1d \ 2\frac{1}{3}$	synon	ymo	us si	tes in	this
		Val		1	Leu			A	Arg			Th	r		
	G	Т	Α	С	Т	G	А		A	Α	А	С	С	То	tal
# substitutions															
# non-syn	0	0	0	1	0	0	0		0	0	0	0	0		1

		Val			Met			Arg			Thr		
	G	Т	Т	А	Т	G	А	А	G	А	С	С	Total
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	1/2	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		
	G	Т	Α	С	Т	G	А	Α	Α	А	С	С	Total
# 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

#### Now,

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

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

