Tests of Positive Selection based on the Comparison of Polymorphism and Divergence

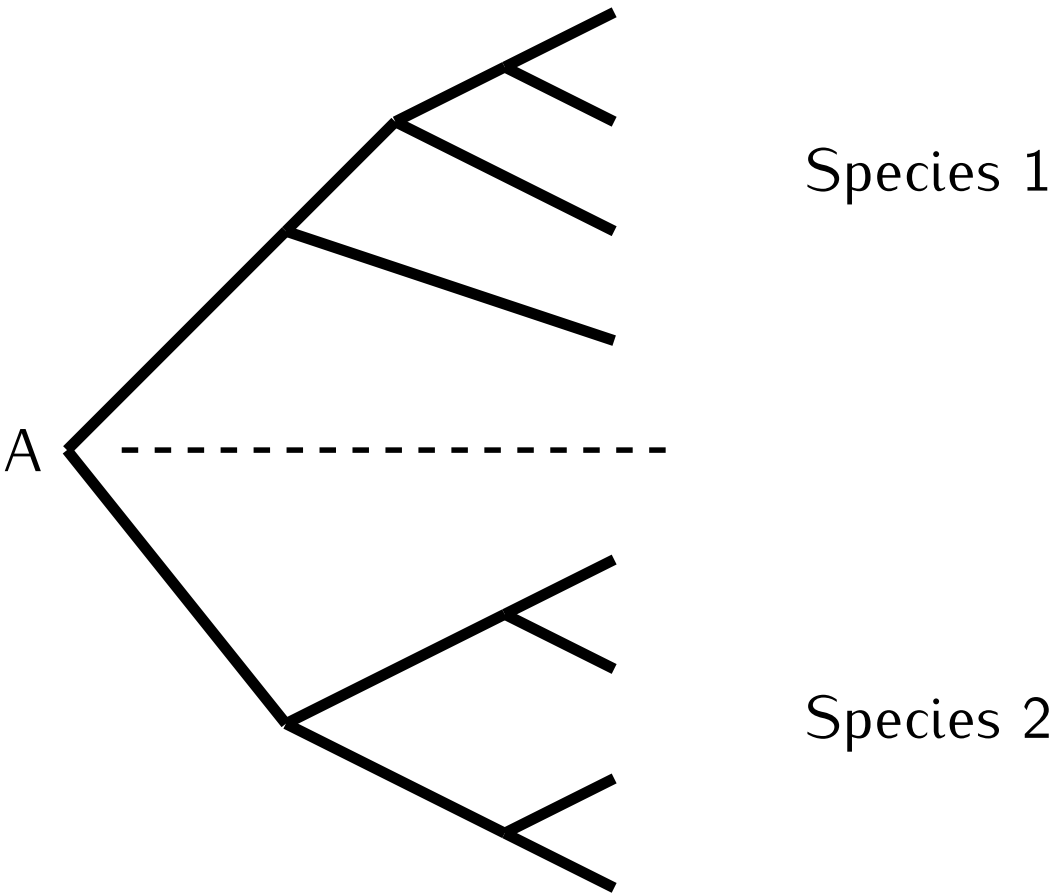
Julien Dutheil

dutheil@evolbio.mpg.de

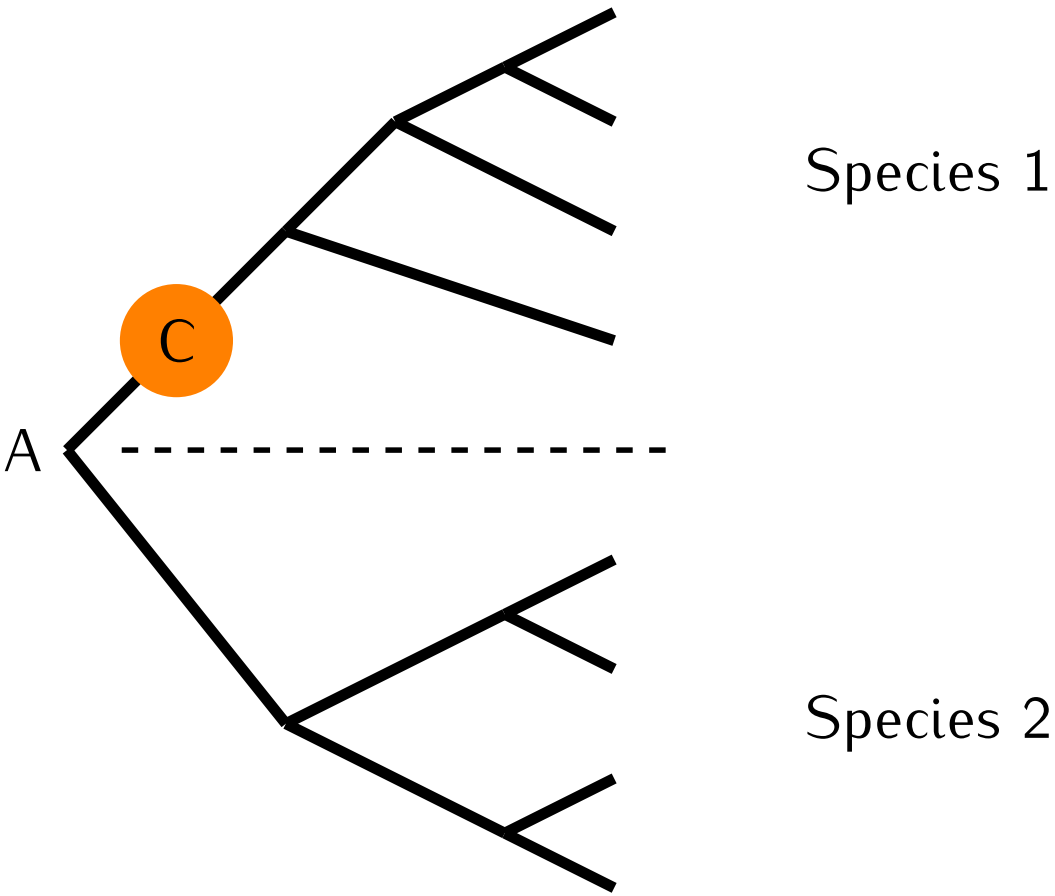
Max Planck Institute for Evolutionary Biology

June 22nd 2015

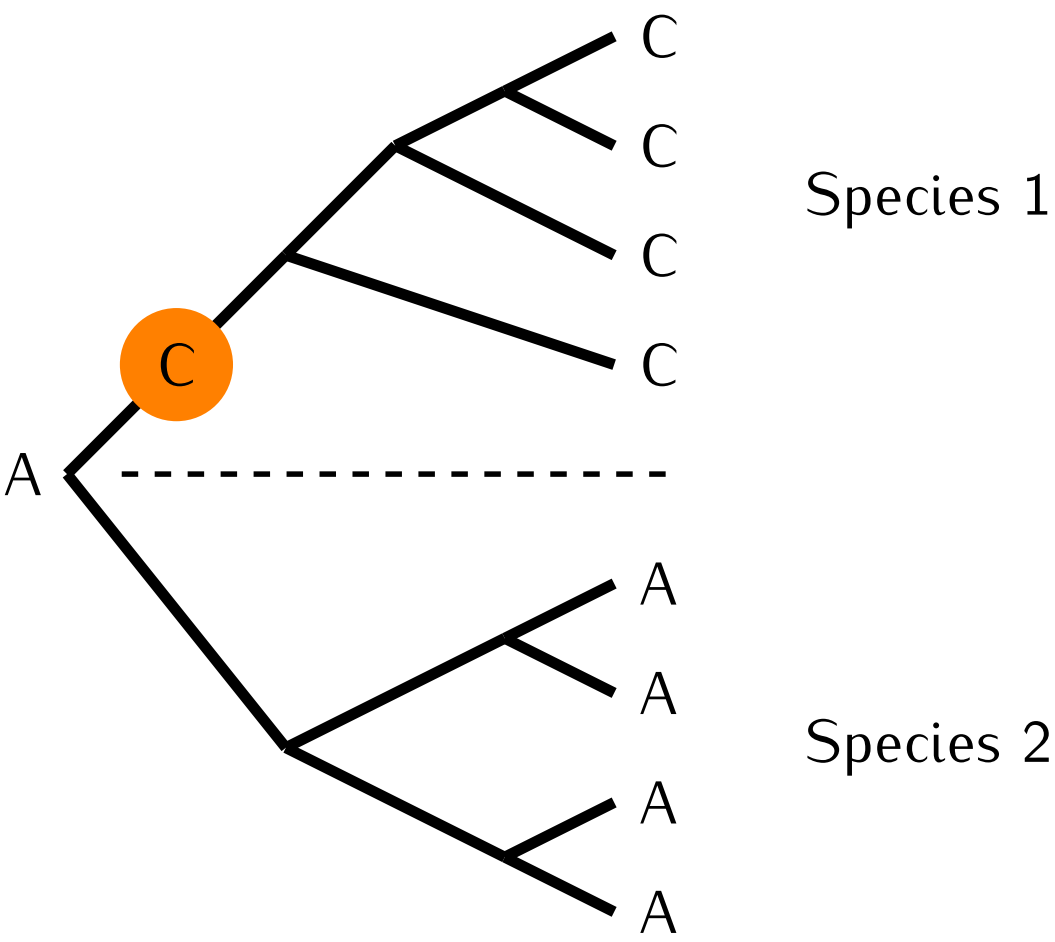
Within vs. between species



Within vs. between species

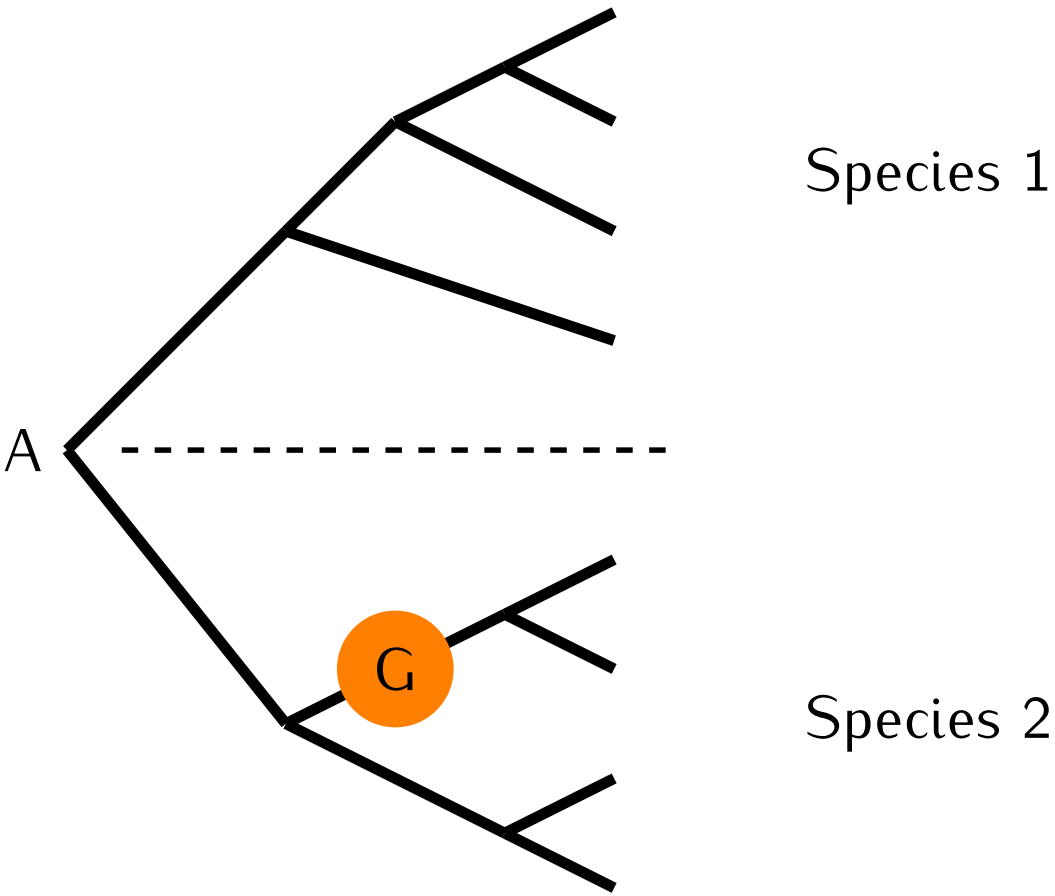


Within vs. between species

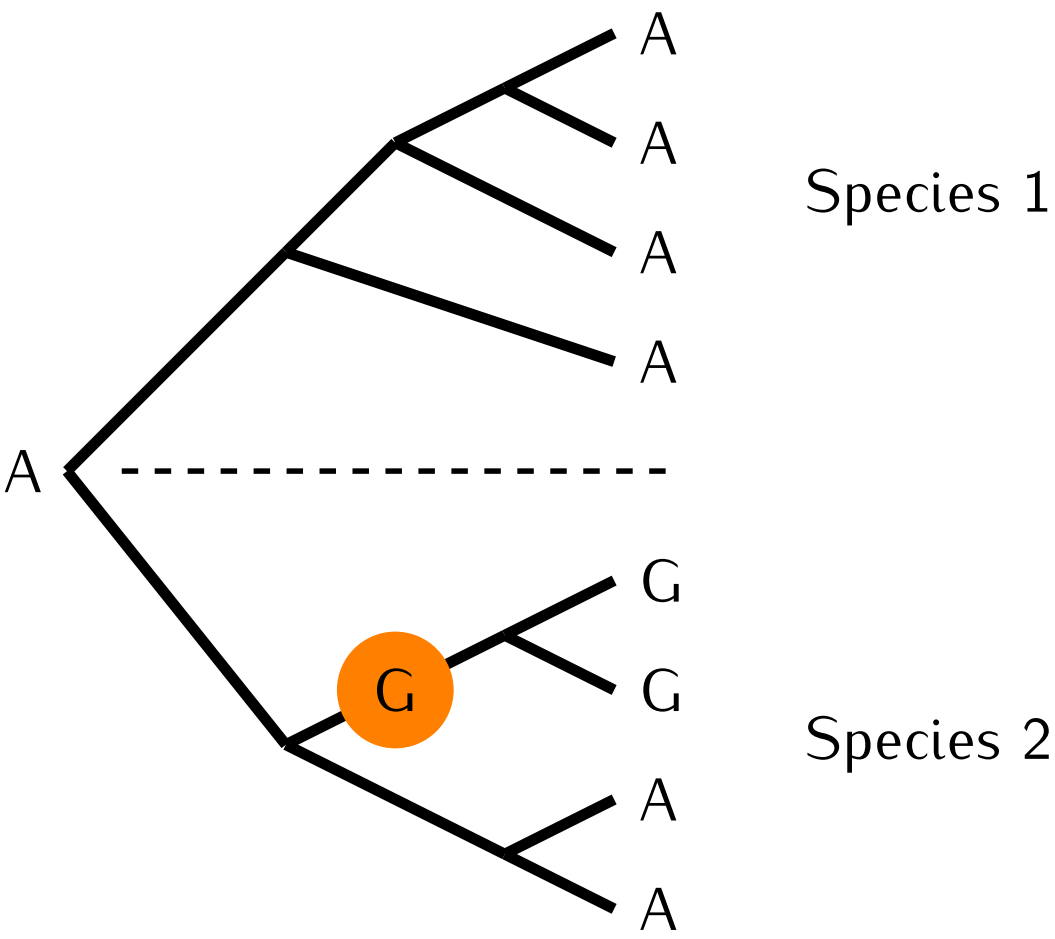


- Mutations on *interspecies branches* lead to fixed differences between species

Within vs. between species



Within vs. between species



- Mutations on *interspecies branches* lead to fixed differences between species
- Mutations on *intraspecies branches* lead to polymorphism in one species

If all mutations are neutral...

☞ The ratio of polymorphic sites vs. fixed differences sites is constant along the genome!

If all mutations are neutral...

☞ The ratio of polymorphic sites vs. fixed differences sites is constant along the genome!

If mutation rate varies between sites but is constant over time in the two species, two predictions:

1. the ratio of polymorphism vs. divergence is constant between genes

If all mutations are neutral...

☞ The ratio of polymorphic sites vs. fixed differences sites is constant along the genome!

If mutation rate varies between sites but is constant over time in the two species, two predictions:

1. the ratio of polymorphism vs. divergence is constant between genes
2. the ratio of non-synonymous to synonymous polymorphism equals the ratio of non-synonymous to synonymous divergence

Polymorphism and divergence are two facets of the same process

The HKA test

Hudson, Kreitman and Aguadé (1987)

- Compare at least 2 loci in 2 species, with polymorphism data in at least 1 species

The HKA test

Hudson, Kreitman and Aguadé (1987)

- Compare at least 2 loci in 2 species, with polymorphism data in at least 1 species
- If mutation rate is constant in time:
  - Regions with high mutation rate display high levels of polymorphism and divergence
  - Regions with low mutation rate display low levels of polymorphism and divergence

The HKA test

Hudson, Kreitman and Aguadé (1987)

- Compare at least 2 loci in 2 species, with polymorphism data in at least 1 species
- If mutation rate is constant in time:
  - Regions with high mutation rate display high levels of polymorphism and divergence
  - Regions with low mutation rate display low levels of polymorphism and divergence
- 'Goodness-of-fit' test to assess how consistent distinct regions are with a constant mutation rate

The HKA test

Hudson, Kreitman and Aguadé (1987)

- Compare at least 2 loci in 2 species, with polymorphism data in at least 1 species
- If mutation rate is constant in time:
  - Regions with high mutation rate display high levels of polymorphism and divergence
  - Regions with low mutation rate display low levels of polymorphism and divergence
- 'Goodness-of-fit' test to assess how consistent distinct regions are with a constant mutation rate
- Assumes free recombination between regions and no recombination within regions

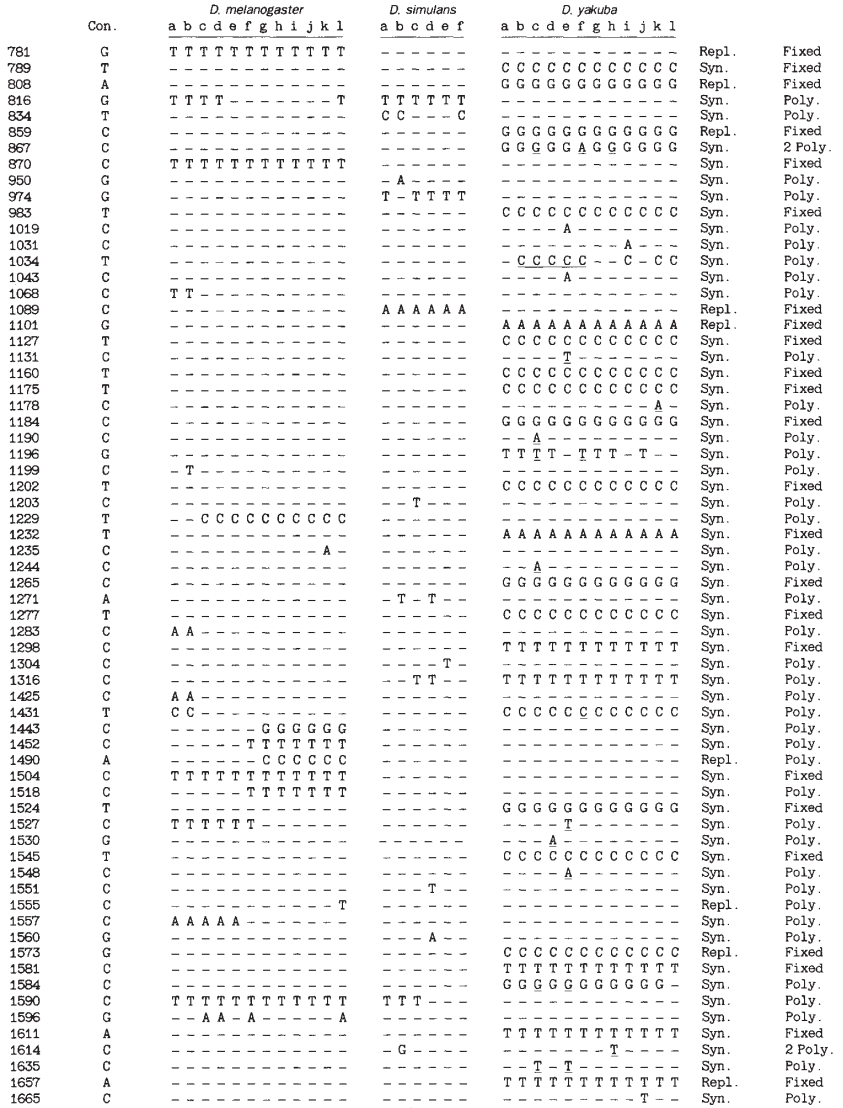
The MK test
McDonald and Kreitman (1991)

- One coding gene in at least 2 species, with polymorphism data for at least 1 species

The MK test

McDonald and Kreitman (1991)

- One coding gene in at least 2 species, with polymorphism data for at least 1 species
- Count synonymous and non-synonymous polymorphisms and fixed differences



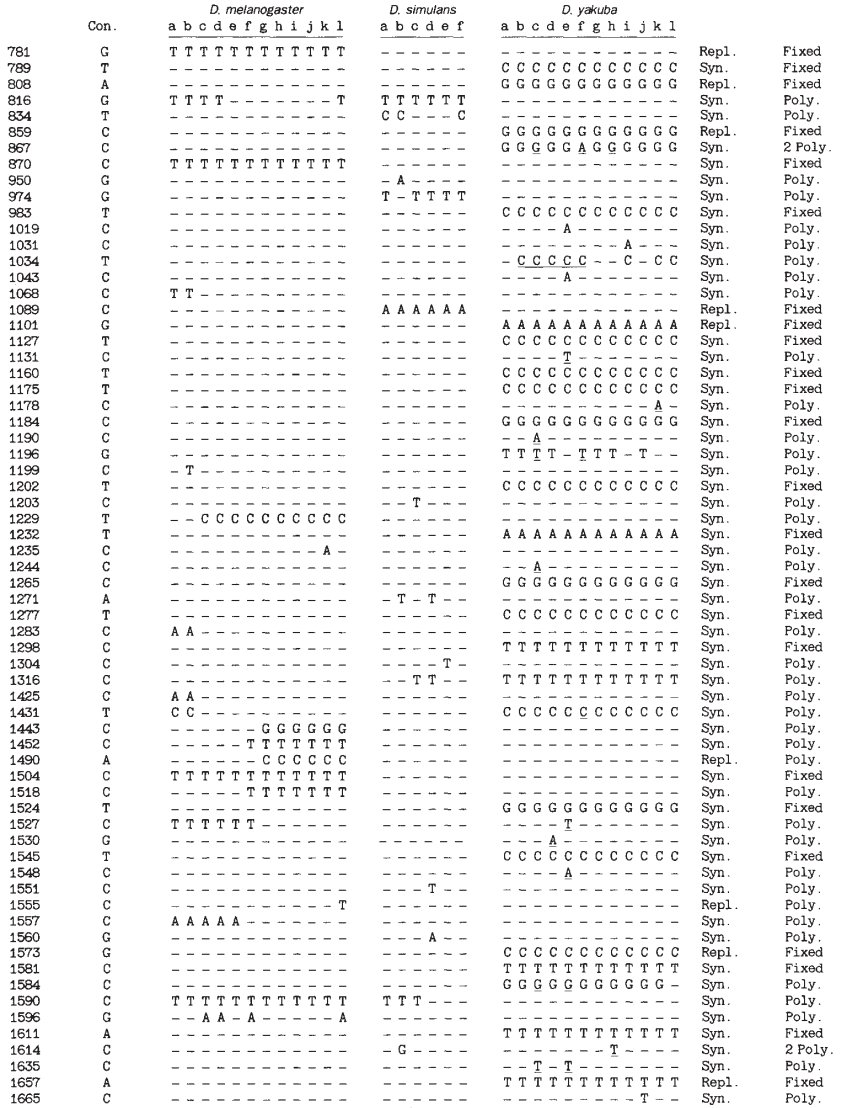
		D. melanogaster	*D. simulans*	*D. yakuba*		
	Con.	a b c d e f g h i j k l	a b c d e f	a b c d e f g h i j k l		
781	G	T T T T T T T T T T T T			Repl.	Fixed
789	T			C C C C C C C C C C C C	Syn.	Fixed
808	A			G G G G G G G G G G G G	Repl.	Fixed
816	G	T T T T T	T T T T T T		Syn.	Poly.
834	T		C C C		Syn.	Poly.
859	C			G G G G G G G G G G G G	Repl.	Fixed
867	C			G G G G A G G G G G G	Syn.	2 Poly.
870	C	T T T T T T T T T T T T			Syn.	Fixed
950	G		A		Syn.	Poly.
974	G		T T T T T		Syn.	Poly.
983	T			C C C C C C C C C C C C	Syn.	Fixed
1019	C			A	Syn.	Poly.
1031	C			A	Syn.	Poly.
1034	T			C C C C C C C C	Syn.	Poly.
1043	C			A	Syn.	Poly.
1068	C	T T			Syn.	Poly.
1089	C		A A A A A A		Repl.	Fixed
1101	G			A A A A A A A A A A A A	Repl.	Fixed
1127	T			C C C C C C C C C C C C	Syn.	Fixed
1131	C			T	Syn.	Poly.
1160	T			C C C C C C C C C C C C	Syn.	Fixed
1175	T			C C C C C C C C C C C C	Syn.	Fixed
1178	C			A	Syn.	Poly.
1184	C			G G G G G G G G G G G G	Syn.	Fixed
1190	C			A	Syn.	Poly.
1196	G			T T T T T T T T T	Syn.	Poly.
1199	C	T			Syn.	Poly.
1202	T			C C C C C C C C C C C C	Syn.	Fixed
1203	C		T		Syn.	Poly.
1229	T	C C C C C C C C C C C			Syn.	Poly.
1232	T			A A A A A A A A A A A A	Syn.	Fixed
1235	C	A			Syn.	Poly.
1244	C			A	Syn.	Poly.
1265	C			G G G G G G G G G G G G	Syn.	Fixed
1271	A		T T		Syn.	Poly.
1277	T			C C C C C C C C C C C C	Syn.	Fixed
1283	C	A A			Syn.	Poly.
1298	C			T T T T T T T T T T T T	Syn.	Fixed
1304	C		T		Syn.	Poly.
1316	C		T T	T T T T T T T T T T T	Syn.	Poly.
1425	C	A A			Syn.	Poly.
1431	T	C C		C C C C C C C C C C C C	Syn.	Poly.
1443	C	G G G G G G			Syn.	Poly.
1452	C	T T T T T T T			Syn.	Poly.
1490	A	C C C C C C			Repl.	Poly.
1504	C	T T T T T T T T T T T T			Syn.	Fixed
1518	C	T T T T T T T			Syn.	Poly.
1524	T			G G G G G G G G G G G G	Syn.	Fixed
1527	C	T T T T T T		T	Syn.	Poly.
1530	G			A	Syn.	Poly.
1545	T			C C C C C C C C C C C C	Syn.	Fixed
1548	C			A	Syn.	Poly.
1551	C		T		Syn.	Poly.
1555	C	T			Repl.	Poly.
1557	C	A A A A A			Syn.	Poly.
1560	G		A		Syn.	Poly.
1573	G			C C C C C C C C C C C C	Repl.	Fixed
1581	C			T T T T T T T T T T T	Syn.	Fixed
1584	C			G G G G G G G G G G G	Syn.	Poly.
1590	C	T T T T T T T T T T T T	T T T		Syn.	Poly.
1596	G	A A A			Syn.	Poly.
1611	A			T T T T T T T T T T T	Syn.	Fixed
1614	C		G	T	Syn.	2 Poly.
1635	C			T T	Syn.	Poly.
1657	A			T T T T T T T T T T T	Repl.	Fixed
1665	C			T	Syn.	Poly.

The MK test

McDonald and Kreitman (1991)

- One coding gene in at least 2 species, with polymorphism data for at least 1 species
- Count synonymous and non-synonymous polymorphisms and fixed differences
- Build a contingency table and perform a G-test

	Fixed	Polym.
Non-syn.	7	2
Synon.	17	42



Inter-specific codon models

Yang (1998)

- Consider at least 2 species, with one sequence per species

Inter-specific codon models

Yang (1998)

- Consider at least 2 species, with one sequence per species
- Assumes a known phylogeny (at least topology)

Inter-specific codon models

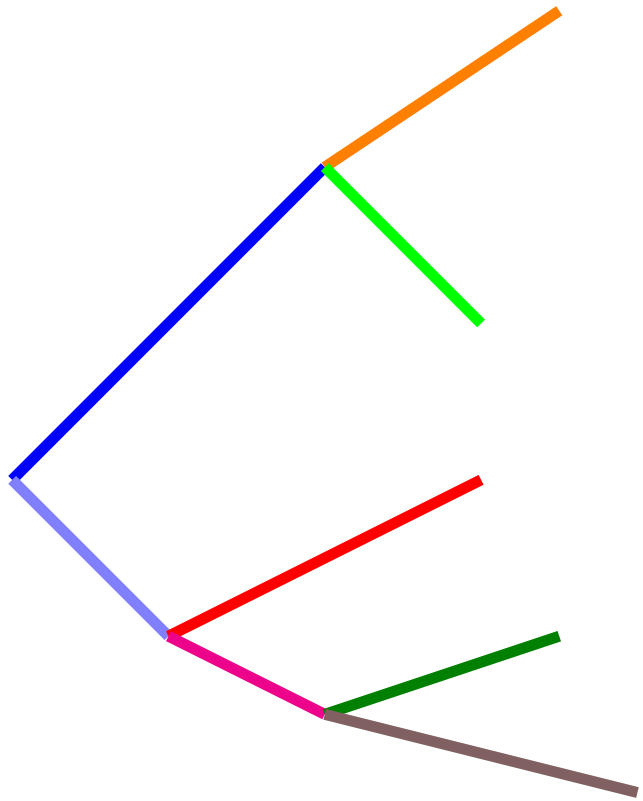
Yang (1998)

- Consider at least 2 species, with one sequence per species
- Assumes a known phylogeny (at least topology)
- Non-homogeneous model: distinct branches in the tree are allowed to have evolved with distinct $\omega = dN/dS$:

Inter-specific codon models

Yang (1998)

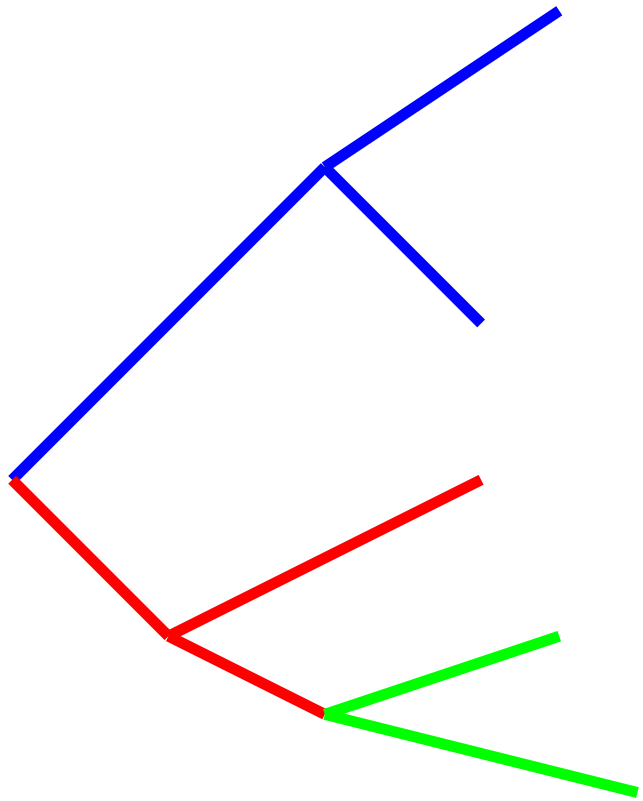
- Consider at least 2 species, with one sequence per species
- Assumes a known phylogeny (at least topology)
- Non-homogeneous model: distinct branches in the tree are allowed to have evolved with distinct $\omega = dN/dS$:
  - One per branch $\Rightarrow$ *branch* model



Inter-specific codon models

Yang (1998)

- Consider at least 2 species, with one sequence per species
- Assumes a known phylogeny (at least topology)
- Non-homogeneous model: distinct branches in the tree are allowed to have evolved with distinct $\omega = dN/dS$:
  - One per branch $\Rightarrow$ *branch* model
  - Several clades $\Rightarrow$ *clade* model



Inter-specific codon models

Yang (1998)

- Consider at least 2 species, with one sequence per species
- Assumes a known phylogeny (at least topology)
- Non-homogeneous model: distinct branches in the tree are allowed to have evolved with distinct $\omega = dN/dS$:
  - One per branch $\Rightarrow$ *branch* model
  - Several clades $\Rightarrow$ *clade* model
- Other parameters are constant throughout the tree

Finding the best model

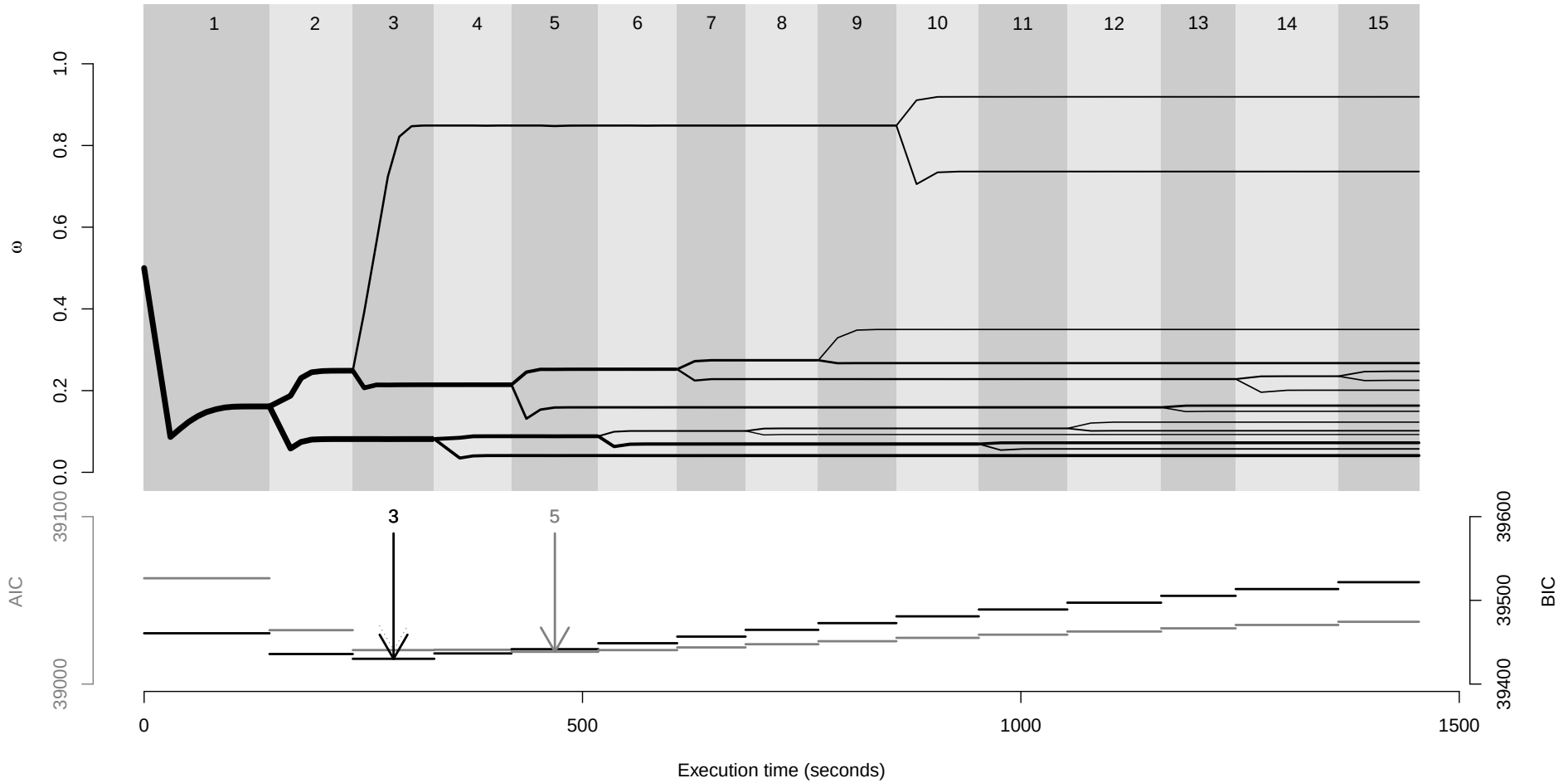
Dutheil et al. (2012)

- The branch model suffers from overparametrization issues
- The clade model needs an *a priori* knowledge

Finding the best model

Dutheil et al. (2012)

- The branch model suffers from overparametrization issues
- The clade model needs an *a priori* knowledge



Combining site and branch heterogeneity

Yang and Nielsen (2002), Zhang, Nielsen and Yang (2005)

- Consider a dataset with several species, with one species per branch and known phylogeny

Combining site and branch heterogeneity

Yang and Nielsen (2002), Zhang, Nielsen and Yang (2005)

- Consider a dataset with several species, with one species per branch and known phylogeny
- Consider two models, with and without selection. Branches where positive selection might have occurred are known *a priori*

Combining site and branch heterogeneity

Yang and Nielsen (2002), Zhang, Nielsen and Yang (2005)

- Consider a dataset with several species, with one species per branch and known phylogeny
- Consider two models, with and without selection. Branches where positive selection might have occurred are known *a priori*
- Branches evolving under positive selection are called *foreground* branches, others *background* branches

Combining site and branch heterogeneity

Yang and Nielsen (2002), Zhang, Nielsen and Yang (2005)

- Consider a dataset with several species, with one species per branch and known phylogeny
- Consider two models, with and without selection. Branches where positive selection might have occurred are known *a priori*
- Branches evolving under positive selection are called *foreground* branches, others *background* branches
- Background branches evolve under the M1a model, foreground branches under the M2a model

Combining site and branch heterogeneity

Yang and Nielsen (2002), Zhang, Nielsen and Yang (2005)

- Consider a dataset with several species, with one species per branch and known phylogeny
- Consider two models, with and without selection. Branches where positive selection might have occurred are known *a priori*
- Branches evolving under positive selection are called *foreground* branches, others *background* branches
- Background branches evolve under the M1a model, foreground branches under the M2a model
- Likelihood ratio test to compare with a homogeneous M1a model.