

# Inferring Complex DNA Substitution Processes on Phylogenies Using Uniformization and Data Augmentation

**Ligia Mateiu** \* and **Bruce Rannala**\*\*

\*Department of Medical Genetics, University of Alberta, Canada

\*\*Genome Center and Section of Evolution & Ecology, University of California, Davis, USA

## SUBSTITUTION RATES

- ★ **Constant rate** → too simple
- ★ **Invariable sites** → still too simple
- ★ **Site specific rates** → more realistic, but overly complex

# GAMMA DISTRIBUTION

- **continuous variation among sites**

Yang, Z. 1993, *Mol.Biol.Evol.* 10:1396-1401

→ **PAML (BASEMLG) < 8 taxa**

- **approximation of the continuous Gamma distribution**

Yang, Z. 1994, *J.Mol.Evol* 39:306-314

→ **PAML (BASEML)**

## OBJECTIVES

- ★ **To allow the substitution rate to vary continuously across sites with the use of Bayesian MCMC method**
- ★ **To increase the efficiency of site-specific rate calculations by using uniformization of a Markov process**

## Transition probabilities calculations:

$$P(t) = e^{-\mathbf{Q}t}$$

### Analytical solutions

JC: 
$$p_{ij}(t) = \frac{3}{4}(1 - e^{-\frac{4}{3}rt})$$

$$p_{ii}(t) = \frac{3}{4}(1 + 3e^{-\frac{4}{3}rt})$$

### Matrix decomposition

GTR: 
$$P(t) = \mathbf{H}e^{Dt}\mathbf{H}^{-1}$$

# Transition probabilities calculations: Uniformization technique

**CTMC**



**Q**

## Transition probabilities calculations: Uniformization technique

**CTMC**  $\longrightarrow$  **DTMC** associated with a **Poisson process**

↓

**Q**

↓

$$\mathbf{P} = \mathbf{Q}/\nu + \mathbf{I}$$

↓

**$\nu$**

$$\nu = 1/\min_i(\pi_i) , i \in \{G,C,A,T\}$$

## Transition probabilities calculations: Uniformization technique

**CTMC** → **DTMC** associated with a **Poisson process**

↓

**Q**

↓

$$\mathbf{P} = \mathbf{Q}/\nu + \mathbf{I}$$

↓

**v**

$$\nu = 1/\min_i(\pi_i), i \in \{G, C, A, T\}$$

$$p_{ij}(w) = \sum_{M=0}^{\infty} \frac{(\nu w)^M w^{-\nu w}}{M!} \times P_{ij}^{(M)}, w = rt$$

## Transition probabilities calculations: Uniformization technique

**CTMC**  $\longrightarrow$  **DTMC** associated with a **Poisson process**

↓

**Q**

↓

$$\mathbf{P} = \mathbf{Q}/\nu + \mathbf{I}$$

↓

**$\nu$**

$$\nu = 1/\min_i(\pi_i), i \in \{G,C,A,T\}$$

$$p_{ij}(w) = \sum_{M=0}^{\infty} \frac{(\nu w)^M w^{-\nu w}}{M!} \times P_{ij}^{(M)}, w = rt$$

↓

**MCMC analysis**

Ross SM Stochastic Processes, Wiley 1983

## Parameters in MCMC

---

---

$\mathbf{x}^-$	nucleotides at the internal nodes	
$\mathbf{M}$	nr of transitions under unif. process	$Uni(0, 50)$
$\mathbf{w}$	branch lengths	$Exp(\lambda)$
$\mathbf{r}$	site specific rates	$\Gamma(\alpha, \alpha^2/\beta)$ Yang, Z. 1993 <i>MolBiolEvol</i> 10:1396
$\theta$	param. GTR subst.model (a,b,c,d,e)	$Dir(1, 1, 1, 1, 1)$ Zwickl D, Holder M 2004 <i>SystBiol</i> 53:877
$\alpha$	param. of Gamma Distr.	$Uni(0, 100)$
$\lambda$	param. of Exp Distr.	$Uni(0, 100)$

---

---

## Simulation studies

- < 8 seqs; 500, 2000, 5000 sites
  - **BYPASSR** ★
  - **PAML** ★ BASEMLG - (with continuous Gamma)
- 10, 20, 50, 250 seqs; 500, 2000, 5000 sites
  - **BYPASSR**
  - **PAML** ★ BASEML - (with discrete Gamma)

★ Bayesian Phylogenetic Analysis of Site Specific Rates

\*Phylogenetic Analysis by Maximum Likelihood (Yang Z. 1997 *CABIOS* 15:555)

# Alpha 2B adrenergic receptor gene

Madsen O et al. *MolBiolEvol* 2002;19:2150

(45 mammalian species, 732 sites)

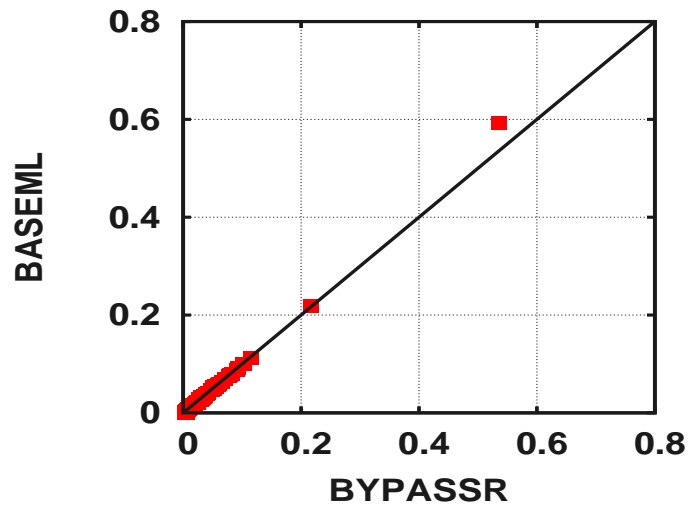
GTR + $\Gamma$				
	BYPASSR <sup>★</sup>	PAML <sup>*</sup>	PAML	PAML
	(3 chains)	5 cat.	20 cat	50 cat.
$\alpha$	0.523	0.516	0.529	0.523
a	0.884	0.856	0.867	0.866
b	0.248	0.241	0.241	0.242
c	0.236	0.227	0.229	0.229
d	0.328	0.324	0.321	0.32
e	0.174	0.171	0.17	0.17
tree length	3.3725	3.501	3.539	3.583

★ Bayesian Phylogenetic Analysis of Site Specific Rates

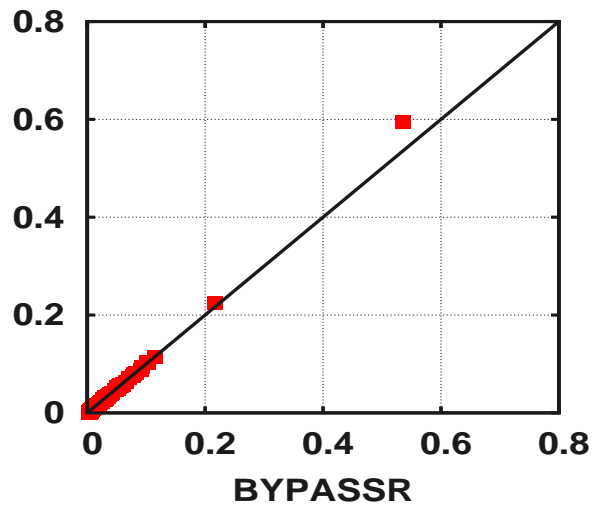
\* PAML(BASEML) (Yang Z. 1997. *CABIOS* 15: 555)

# BRANCH LENGTHS

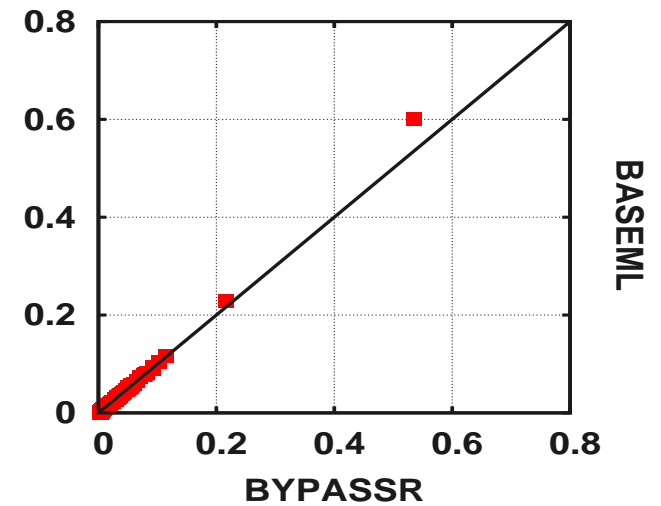
**BYPASSR /  
BASEML (5 cat.)**



**BYPASSR /  
BASEML (20 cat.)**

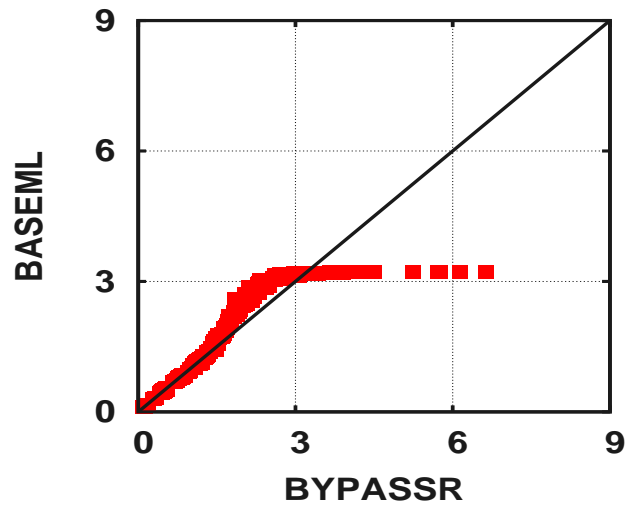


**BYPASSR /  
BASEML (50 cat.)**

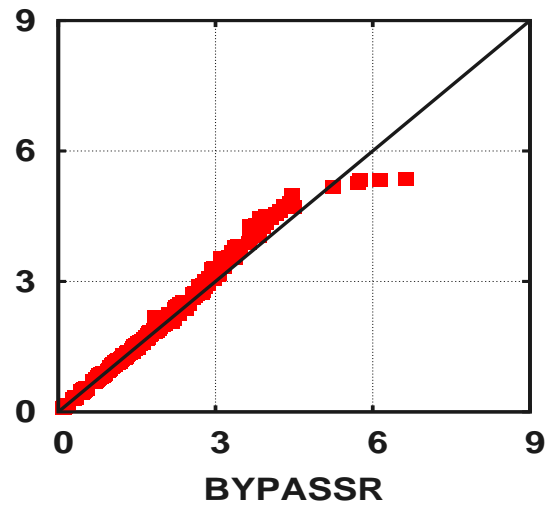


# SITE-SPECIFIC RATES

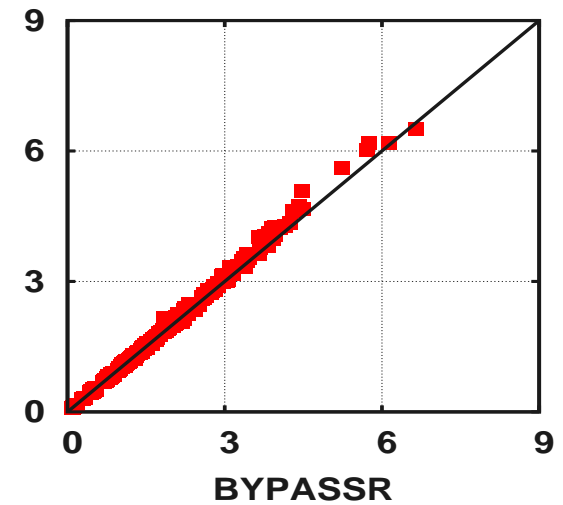
**BYPASSR /  
BASEML (5 cat.)**



**BYPASSR /  
BASEML (20 cat.)**

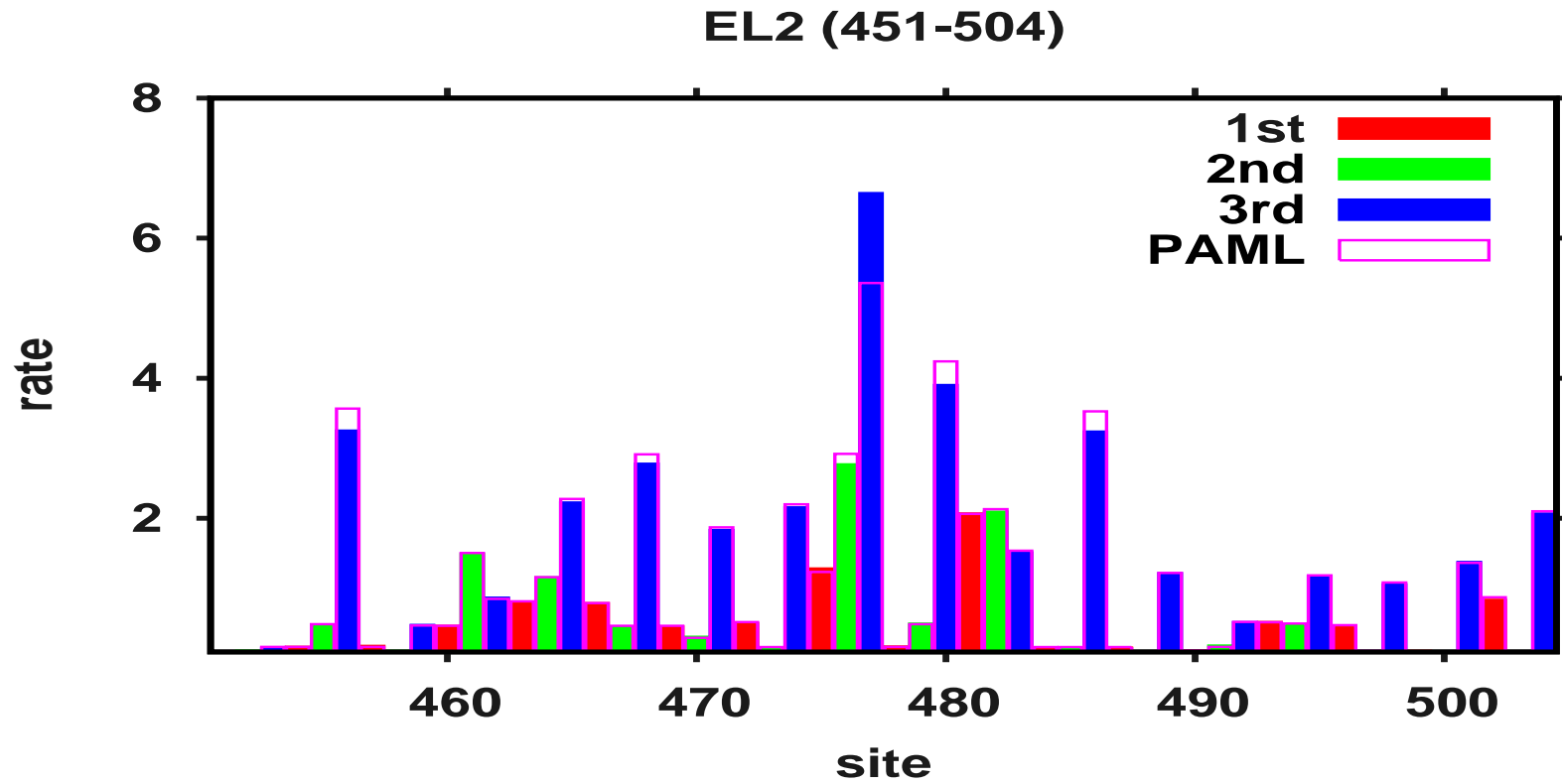


**BYPASSR /  
BASEML (50 cat.)**

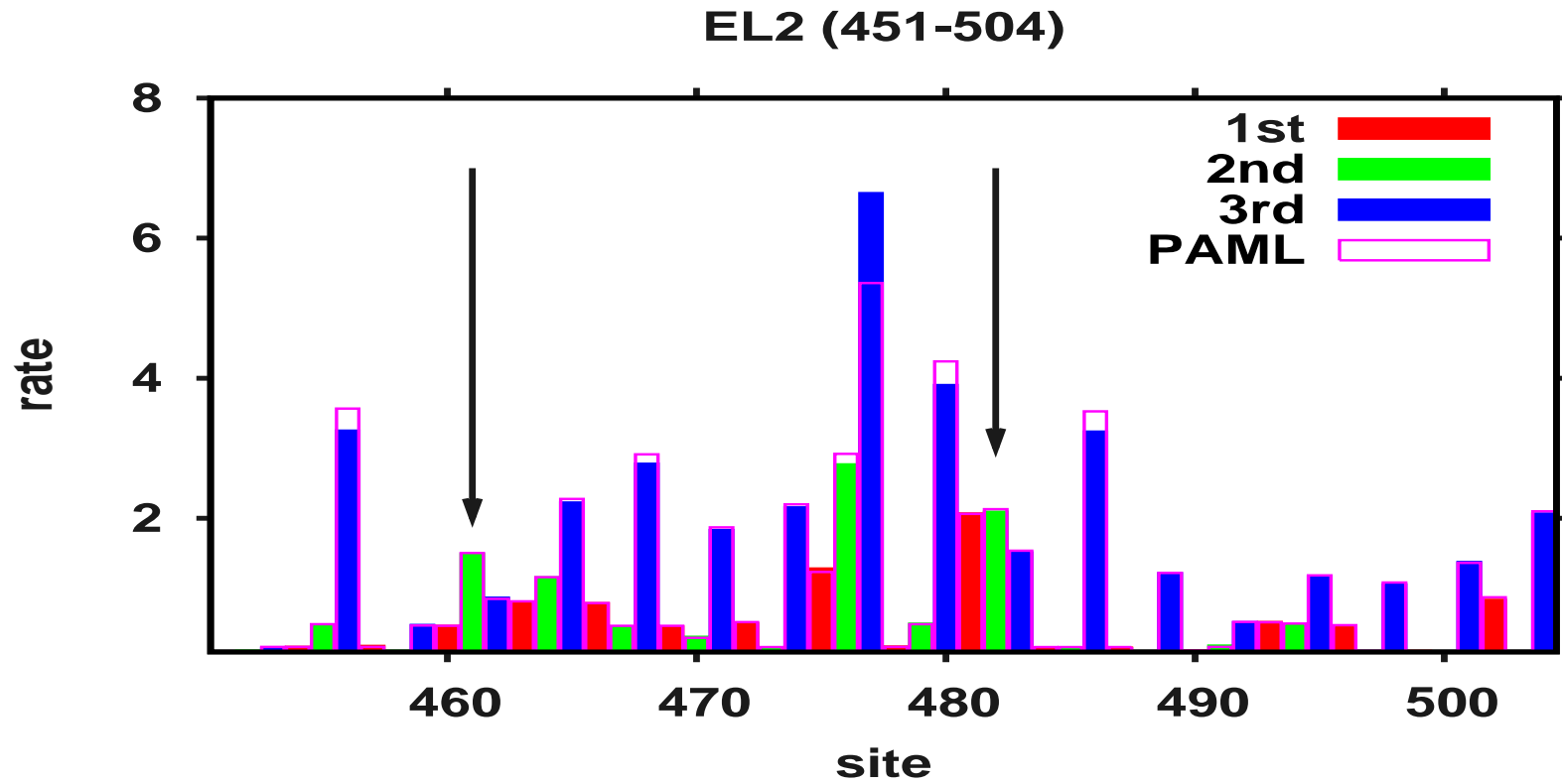


BASEML

# SITE-SPECIFIC RATES AT EL2 DOMAIN



# SITE-SPECIFIC RATES AT EL2 DOMAIN



# CONCLUSIONS

- ★ **Uniformization technique allows the implementation of complex substitution models in Bayesian MCMC framework**
- ★ **Allows each site to have specific rate, even for very large trees**
- ★ **Facilitates searches for selection in a site by site basis**

(Paper submitted to Syst. Biol.)