Complex models of sequence evolution require accurate estimators as exemplified with the invariable site plus Gamma model

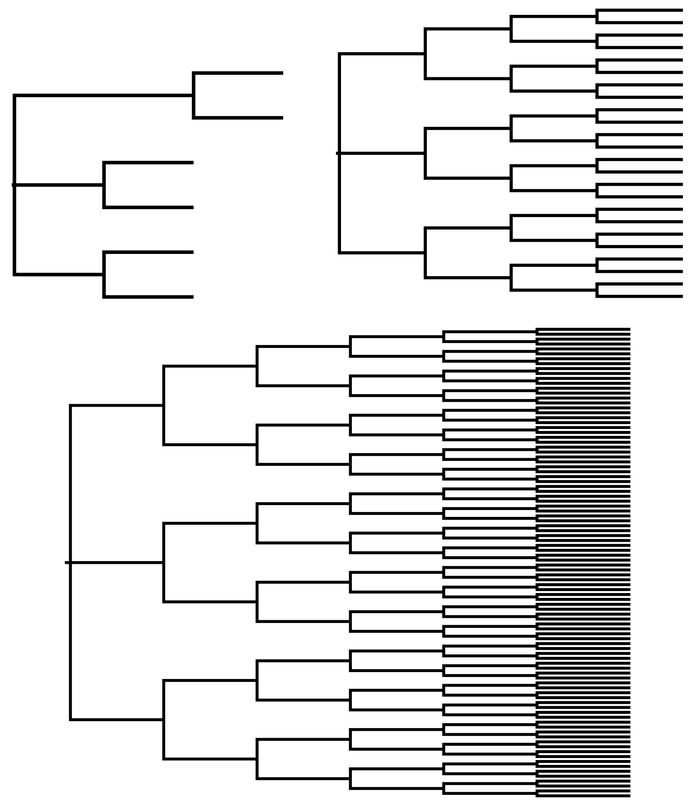
Lam-Tung Nguyen, Arndt von Haeseler and **Bui Quang Minh**

Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories, Austria

Introduction

- The I+Γ (Gu et al. 1995) model (invariable site and Gamma rate heterogeneity across sites) is widely-used in phylogenetics.
- I+Γ model is parameterized by two parameters: p_{inv} (proportion of invariable sites) and α (shape of Γ distribution).
- Its use is controversial due to correlation between p_{inv} and α for strong rate heterogeneity ($\alpha < 1$) (Sullivan & Swofford 1999; Mayrose et al. 2005; Yang 2006):
  - The use of I+Γ is discouraged (Jia et al. 2014; Yang 2006).
  - However, the identifiability of I+*continuous* Γ model was proven (Rogers 2001; Allman & Rhodes 2008; Chai & Housworth 2011).
- We perform simulations to assess accuracy of I+*discrete* Γ estimators in maximum likelihood and Bayesian software.

Simulations under K2P+I+$\Gamma 4$ ($\kappa = 4$)



Current phylogenetic software do not produce accurate estimates for I+Γ model!

PhyML

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
6-taxon tree (tree length = 1.0)											
0.10	0.00	0.10	0.20	0.25	0.39	0.50	0.61	0.89	0.90	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.06	0.09	0.10	0.11	0.56	0.19	0.11	$\hat{\alpha}$
	1.00	1.00	1.00	0.92	0.98	0.99	1.17	1.71	3.68	0.76	$\hat{l}$
0.50	0.19	0.18	0.17	0.17	0.25	0.36	0.43	0.53	0.68	0.69	$\widehat{p_{inv}}$
	0.84	0.62	0.47	0.35	0.33	0.32	0.26	0.23	0.22	0.04	$\hat{\alpha}$
	1.00	1.00	1.00	0.99	0.98	0.98	0.95	0.91	0.91	18.58	$\hat{l}$
1.00	0.11	0.20	0.20	0.19	0.17	0.15	0.17	0.29	0.45	0.77	$\widehat{p_{inv}}$
	1.50	1.47	0.99	0.67	0.45	0.30	0.21	0.17	0.13	0.52	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	0.98	0.94	0.90	0.80	4.54	$\hat{l}$
24-taxon tree (tree length = 4.5)											
0.10	0.41	0.28	0.35	0.30	0.43	0.51	0.69	0.70	0.80	0.93	$\widehat{p_{inv}}$
	0.55	0.30	0.26	0.10	0.13	0.11	0.18	0.10	0.10	0.17	$\hat{\alpha}$
	4.93	4.93	5.14	4.48	4.79	5.93	26.92	4.43	4.53	10.76	$\hat{l}$
0.50	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.51	0.50	0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\hat{\alpha}$
	4.50	4.50	4.50	4.50	4.49	4.51	4.50	4.52	4.49	4.45	$\hat{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.89	$\widehat{p_{inv}}$
	1.01	1.01	1.00	1.00	0.99	0.99	1.00	0.99	1.00	0.87	$\hat{\alpha}$
	4.50	4.50	4.49	4.50	4.50	4.51	4.50	4.52	4.54	136.95	$\hat{l}$
96-taxon tree (tree length = 18.9)											
0.10	0.00	0.10	0.49	0.55	0.61	0.50	0.62	0.70	0.90	0.92	$\widehat{p_{inv}}$
	0.10	0.10	0.28	0.28	0.28	0.10	0.13	0.10	0.45	0.17	$\hat{\alpha}$
	18.83	18.91	13.90	14.00	14.06	18.69	17.95	18.27	35.95	38.45	$\hat{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\hat{\alpha}$
	18.88	18.88	18.88	18.82	18.90	18.88	18.94	18.77	18.89	18.29	$\hat{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.03	$\hat{\alpha}$
	18.90	18.89	18.89	18.87	18.94	18.93	18.99	18.77	18.92	18.49	$\hat{l}$

RAxML

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
6-taxon tree (tree length = 1.0)											
0.10	0.71	0.74	0.77	0.79	0.82	0.84	0.86	0.89	0.90	0.92	$\widehat{p_{inv}}$
	4.63	3.85	3.08	2.36	1.69	1.15	0.67	0.56	0.24	0.22	$\hat{\alpha}$
	1.03	1.04	1.07	1.10	1.17	1.33	1.78	1.74	4.39	1.38	$\hat{l}$
0.50	0.30	0.34	0.40	0.45	0.51	0.57	0.64	0.72	0.81	0.91	$\widehat{p_{inv}}$
	1.31	1.17	1.05	0.95	0.83	0.75	0.68	0.63	0.63	0.73	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	$\hat{l}$
1.00	0.16	0.22	0.29	0.37	0.45	0.53	0.61	0.71	0.80	0.90	$\widehat{p_{inv}}$
	1.89	1.74	1.59	1.45	1.35	1.25	1.17	1.14	1.10	1.05	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	$\hat{l}$
24-taxon tree (tree length = 4.5)											
0.10	0.64	0.67	0.70	0.73	0.76	0.73	0.77	0.82	0.88	0.94	$\widehat{p_{inv}}$
	0.82	0.72	0.64	0.58	0.53	0.20	0.17	0.18	0.24	0.29	$\hat{\alpha}$
	5.18	5.87	6.74	7.69	8.91	16.64	20.97	28.88	8.72	3.38	$\hat{l}$
0.50	0.01	0.11	0.21	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.51	0.51	0.51	0.51	0.51	0.51	0.50	0.50	0.50	0.51	$\hat{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	$\hat{l}$
1.00	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.02	1.02	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00	$\hat{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.50	4.52	$\hat{l}$
96-taxon tree (tree length = 18.9)											
0.10	0.37	0.43	0.49	0.65	0.67	0.68	0.74	0.81	0.87	0.94	$\widehat{p_{inv}}$
	0.27	0.27	0.28	0.62	0.49	0.28	0.26	0.27	0.27	0.28	$\hat{\alpha}$
	13.85	13.88	13.94	13.43	13.79	14.08	14.55	14.32	14.20	11.90	$\hat{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\hat{\alpha}$
	18.89	18.90	18.91	18.89	18.89	18.89	18.89	18.92	18.91	18.88	$\hat{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	$\hat{\alpha}$
	18.90	18.89	18.90	18.89	18.90	18.90	18.90	18.89	18.90	18.92	$\hat{l}$

MrBayes

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
6-taxon tree (tree length = 1.0)											
0.10	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.08	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.02	5.26	$\hat{l}$
0.50	0.10	0.14	0.20	0.28	0.38	0.49	0.60	0.70	0.66	0.74	$\widehat{p_{inv}}$
	0.66	0.57	0.52	0.50	0.48	0.49	0.50	0.50	0.26	0.11	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.51	$\hat{l}$
1.00	0.04	0.10	0.18	0.28	0.23	0.26	0.24	0.32	0.46	0.68	$\widehat{p_{inv}}$
	1.14	1.02	0.96	0.95	0.69	0.57	0.31	0.21	0.13	0.10	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.91	0.80	0.63	$\hat{l}$
24-taxon tree (tree length = 4.5)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\hat{\alpha}$
	4.48	4.50	4.50	4.51	4.51	4.51	4.51	4.52	4.56	4.67	$\hat{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.87	$\widehat{p_{inv}}$
	0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.25	$\hat{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	3.69	$\hat{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.67	$\widehat{p_{inv}}$
	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.10	$\hat{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.51	4.51	7.44	$\hat{l}$
96-taxon tree (tree length = 18.9)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\hat{\alpha}$
	18.85	18.91	18.92	18.92	18.90	18.93	18.99	19.01	18.97	18.96	$\hat{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.77	0.88	$\widehat{p_{inv}}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.39	0.38	$\hat{\alpha}$
	18.88	18.90	18.91	18.90	18.90	18.90	18.90	18.93	13.82	14.36	$\hat{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.89	$\widehat{p_{inv}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.64	$\hat{\alpha}$
	18.90	18.89	18.90	18.89	18.91	18.91	18.91	18.90	17.69	17.69	$\hat{l}$

IQ-TREE

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
6-taxon tree (tree length = 1.0)											
0.10	0.71	0.74	0.77	0.79	0.82	0.84	0.86	0.89	0.90	0.92	$\widehat{p_{inv}}$
	4.80	4.02	3.27	2.51	1.80	1.15	0.64	0.55	0.28	0.17	$\hat{\alpha}$
	1.03	1.04	1.06	1.10	1.17	1.32	1.76	1.73	3.47	1.16	$\hat{l}$
0.50	0.18	0.23	0.30	0.36	0.42	0.51	0.61	0.70	0.80	0.91	$\widehat{p_{inv}}$
	0.82	0.74	0.68	0.62	0.56	0.53	0.52	0.52	0.53	0.68	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.87	$\hat{l}$
1.00	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.02	1.01	1.01	1.00	1.01	1.00	1.02	1.04	1.06	1.03	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	$\hat{l}$
24-taxon tree (tree length = 4.5)											
0.10	0.64	0.67	0.70	0.73	0.76	0.79	0.76	0.81	0.87	0.95	$\widehat{p_{inv}}$
	0.82	0.72	0.64	0.58	0.52	0.29	0.14	0.14	0.14	0.30	$\hat{\alpha}$
	5.17	5.86	6.73	7.68	8.89	62.72	41.99	40.81	41.04	43.68	$\hat{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\hat{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.53	$\hat{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	$\hat{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.51	4.59	$\hat{l}$
96-taxon tree (tree length = 18.9)											
0.10	0.36	0.43	0.59	0.65	0.70	0.75	0.79	0.80	0.87	0.93	$\widehat{p_{inv}}$
	0.27	0.27	0.57	0.62	0.63	0.64	0.63	0.28	0.17	0.16	$\hat{\alpha}$
	13.83	13.87	13.26	13.48	13.69	14.00	14.61	19.03	177.99	74.08	$\hat{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\hat{\alpha}$
	18.89	18.90	18.90	18.88	18.92	18.95	18.99	19.09	19.23	19.57	$\hat{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	$\hat{\alpha}$
	18.90	18.89	18.90	18.90	18.92	18.94	18.96	18.99	19.13	19.56	$\hat{l}$

Cause, Solution and Recommendation

CAUSE:

- Optimization heuristics in different software trapped in suboptimal estimates.

SOLUTION:

- An expectation maximization (EM) algorithm to estimate p_{inv}.
- Restarting parameter estimation from 10 initial p_{inv} evenly spaced between 0 and fraction of constant sites.

RECOMMENDATION:

- Other software developers to employ similar strategy.
- Critically scrutinize heuristics when implementing more complex models.

IQ-TREE-EM

α \ p_{inv}	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	
6-taxon tree (tree length = 1.0)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	0.93	$\hat{l}$
0.50	0.06	0.10	0.19	0.27	0.38	0.48	0.59	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.59	0.52	0.50	0.48	0.49	0.48	0.48	0.52	0.52	0.54	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	$\hat{l}$
1.00	0.01	0.09	0.19	0.29	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.05	1.00	0.98	0.98	1.02	1.03	1.03	1.04	1.00	1.02	$\hat{\alpha}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	$\hat{l}$
24-taxon tree (tree length = 4.5)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\hat{\alpha}$
	4.50	4.49	4.49	4.50	4.51	4.51	4.51	4.52	4.52	4.19	$\hat{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\hat{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	$\hat{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	$\hat{\alpha}$
	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.51	4.50	4.50	$\hat{l}$
96-taxon tree (tree length = 18.9)											
0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	$\hat{\alpha}$
	18.90	18.89	18.90	18.91	18.89	18.92	18.97	18.98	19.02	18.80	$\hat{l}$
0.50	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	$\hat{\alpha}$
	18.90	18.90	18.91	18.89	18.90	18.90	18.90	18.92	18.91	18.86	$\hat{l}$
1.00	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	$\widehat{p_{inv}}$
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	$\hat{\alpha}$
	18.90	18.89	18.90	18.89	18.90	18.90	18.90	18.89	18.90	18.87	$\hat{l}$

References

- Allman ES, Rhodes JA (2008) *Math Biosci* 211:18-33.
- Chai J, Housworth EA (2011) *Syst Biol* 60:713-718.
- Gu X, Fu YX, Li WH (1995) *Mol Biol Evol* 12:546-557.
- Jia FZ, Lo N, Ho SYW (2014) *PLOS One* 9.
- Mayrose I, Friedman N, Pupko T (2005) *Bioinformatics* 21:ii151-ii158.
- Rogers JS (2001) *Syst Biol* 50:713-722.
- Sullivan J, Swofford DL, Naylor GJP (1999) *Mol Biol Evol* 16:1347-1356.
- Yang Z (2006) Oxford University Press.