
www.ruthe.de
I know that look! Steffen wants to talk about Misfits again
Ruthe

Phylogenetic inference with real confidence

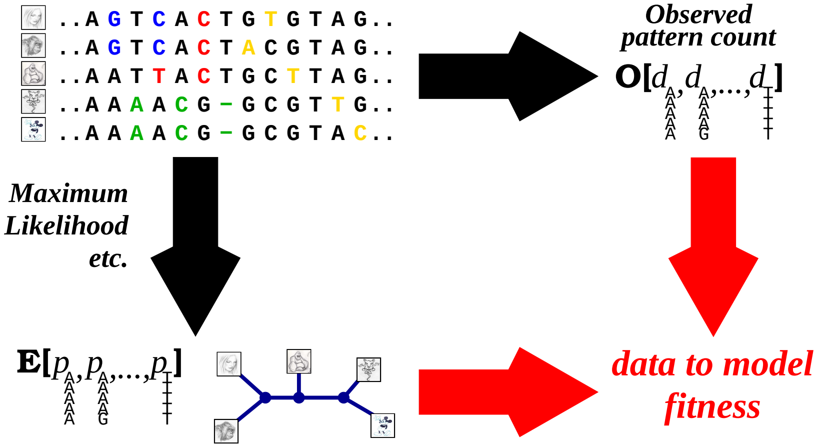
Steffen Klaere

University of Auckland

Acknowledgement

- David Fletcher
- Barbara Holland
- Michael Charleston
- Stephane Guindon
- Bradley Liu
- Daisy Shepherd
- Mung-Yuen Crystal Ng
- Jessica Leigh

Motivation



..A G T C A C T G T G T A G..
..A G T C A C T A C G T A G..
..A A T T A C T G C T T A G..
..A A A A C G – G C G T T G..
..A A A A C G – G C G T A C..
Observed pattern count
O[d, d, ..., d]
Maximum Likelihood etc.
E[p, p, ..., p]
data to model fitness

A tenth crucial question regarding model use in phylogenetics

John Gatesy

Department of Biology, University of California – Riverside, Spieth Hall 2314, Riverside, CA 92521, USA

> *'Unfortunately, in most phylogenetic analyses, even if a log-likelihood is calculated, it is never compared with the best log-likelihood value to see if the models being considered are adequate. In fact, the fit is almost invariably awful, which may explain why such comparisons are not often made.'*
>
> J. Reeves, 1992 [1]

Kelchner and Thomas addressed nine key questions regarding the use of stochastic models in phylogenetics [2]. A tenth crucial question was not explored in detail in their TREE review: 'In modern systematic studies, how often is the fit between model and data absolutely poor?' At their best, models should provide adequate explanations of complex biological patterns. Yet until now, systematists have been preoccupied primarily with the relative fit of competing models to DNA datasets [3]. Simple methods for detecting an absolutely poor fit between DNA sequence data and a particular model have existed for some time [1,4–6], but, unfortunately, these tests have been implemented in relatively few cases [7]. Perhaps either buoyed by studies that asserted the robustness of model-based methods [8] or daunted by computational demands, systematists have hidden their heads in the sand for ~15 years.

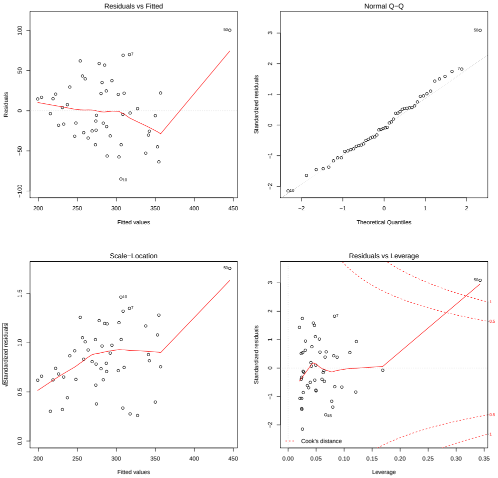
In most published studies, statistical criteria are applied to determine which model, among a set of competing models, fits the empirical data best [3]. From the initial set, a particular model can be chosen as 'optimal,' but might simply represent the best of several extremely poor choices, none of which fit the empirical data well. Given the simplicity of most models, it is possible that model selection in modern systematics is analogous to an overweight man shopping in the petites department of a women's clothing store. A particular garment might fit the portly man best, but this does not imply a good overall fit. Likewise, to assume that any of the simple molecular models commonly utilized by systematists [3] provide a good fit to the data is a leap of faith, especially considering that the most parameter-rich model (i.e. the largest dress in the store) often is chosen as the best for published data matrices [2].

Adequate models of molecular evolution are a prerequisite for successful interpretation of data in the model-based approach to systematics [1–10]. For example, statistical consistency (touted as a hallmark of this framework [9]) and accuracy of branch support values are not guaranteed given a mismatch of model to data [8,10]. The fact that the goodness of fit between DNA data and current models is unknown is a disturbing aspect of phylogenetic analysis in the 21st century. Are molecular models poor fits to the highly complex datasets compiled by modern systematists? Unfortunately, and a bit embarrassingly, we still do not know the answer to this tenth crucial question for the great majority of published datasets [7].

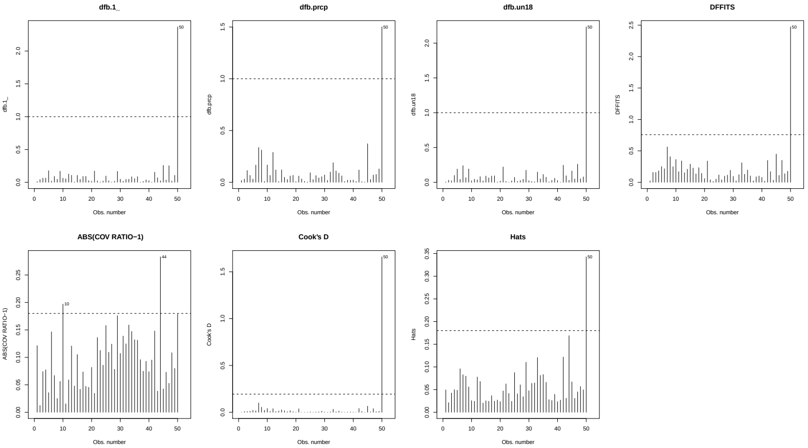
Linear Regression: Omnibus test

```
Call:
lm(formula = educ ~ urban + percap + under18, data = educ.df)
---
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -555.92562  123.46634  -4.503 4.56e-05 ***
urban         -0.00476    0.05174  -0.092    0.927
percap         0.07236    0.01165   6.211 1.40e-07 ***
under18        1.55134    0.31545   4.918 1.16e-05 ***
---
Residual standard error: 40.53 on 46 degrees of freedom
Multiple R-squared:  0.5902, Adjusted R-squared:  0.5634
F-statistic: 22.08 on 3 and 46 DF,  p-value: 5.271e-09
```

Linear Regression: Outliers vs. Fitted values



Linear Regression: Influential observations



Model fitness in ML: Omnibus test

- Deviance statistic

$$G = 2 \sum_{j=1}^{M} N_i \left(\ln p_i - \ln \left(N_i / N \right) \right)$$

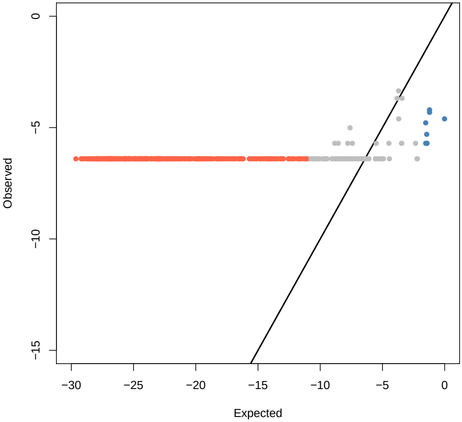
  Assess significance using parametric bootstrap (Goldman, 1993a)

- Pearson X^2 statistic

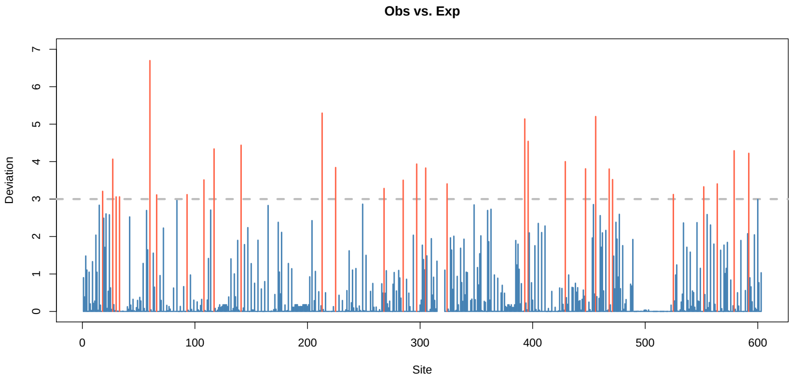
$$X^2 = \sum_{j=1}^{4^n} \frac{\left(N_i - Np_i \right)^2}{Np_i} = \sum_{j=1}^{M} \frac{Np_i \left(N_i - Np_i \right)}{Np_i}$$

  Assessment similar to above.

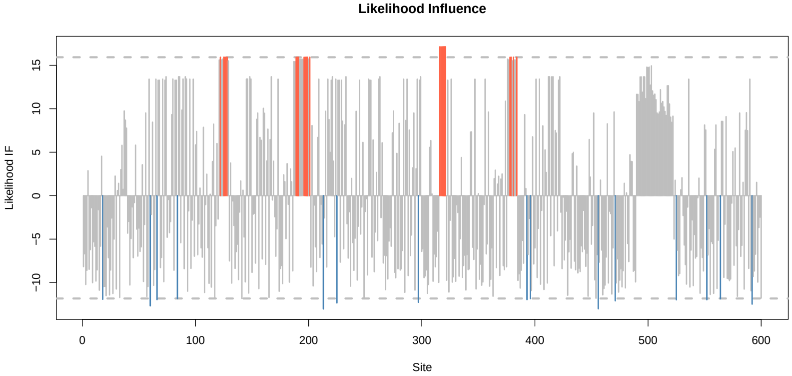
Model fitness in ML: Outliers vs. Fitted values



Model fitness in ML: Outliers vs. Fitted values



Model fitness in ML: Outliers vs. Fitted values



Model fitness in ML: Influence measures

Topology: Robinson-Foulds, SPR, NNI etc.

Branches: Weighted RF, branch score distance, geodesic distance

Parameters: Leave-one-out

$$IF_h(\theta_j) = \frac{\left|\widehat{\theta}_j - \widehat{\theta}_j[-h]\right|}{\text{se}\left(\widehat{\theta}_j\right)}$$

Problem: What is the standard error of the shape parameter?

Model fitness in ML: Simulation settings

General settings: 19 taxa, 600 sites...

Simulation 1: All good...

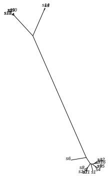
Simulation 2: Concatenated alignment...

Simulation 3: Total chaos...

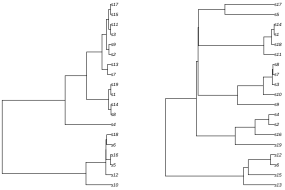
Simulation 4: Actual data...

Trees

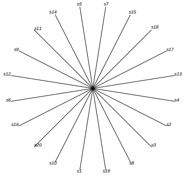
Simulation 1



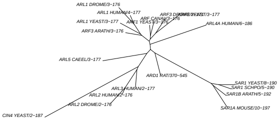
Simulation 2



Simulation 3

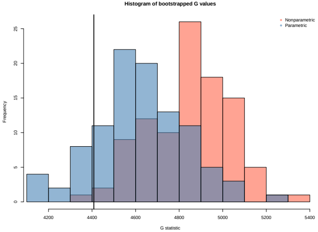


Actual data

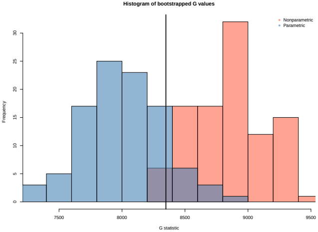


G Statistics

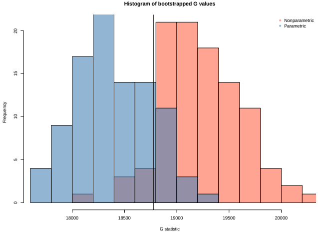
Simulation 1



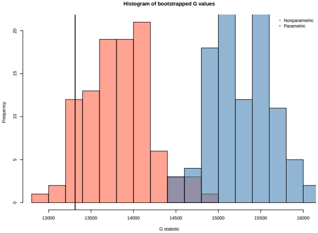
Simulation 2



Simulation 3

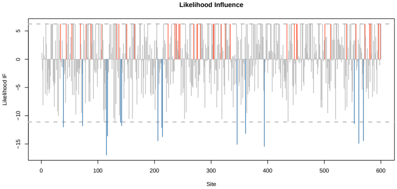


Actual data

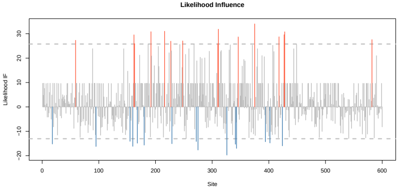


Likelihood IF

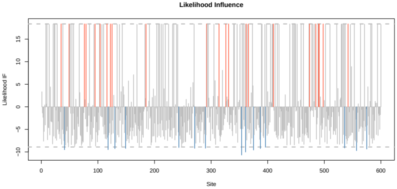
Simulation 1



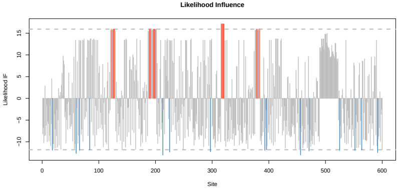
Simulation 2



Simulation 3

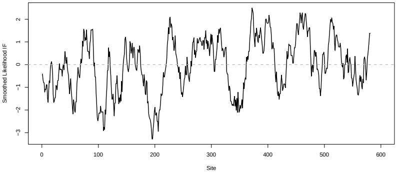


Actual data

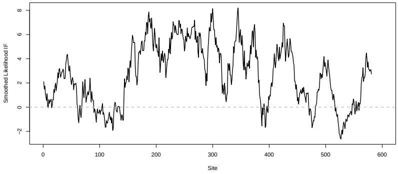


Smoothing the Likelihood IF

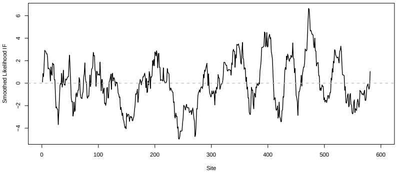
Simulation 1



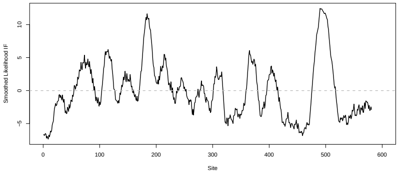
Simulation 2



Simulation 3

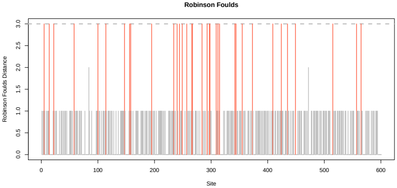


Actual data

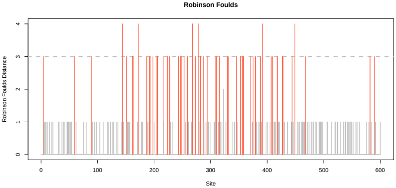


Robinson Foulds distance

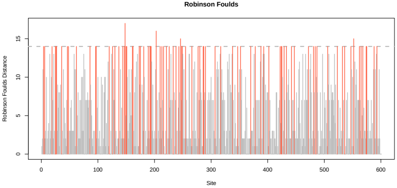
Simulation 1



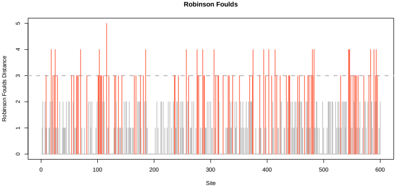
Simulation 2



Simulation 3

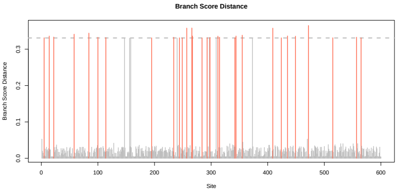


Actual data

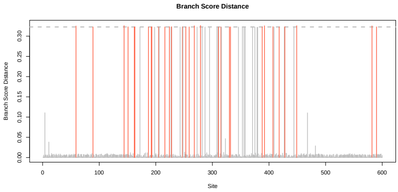


Branch score distance

Simulation 1



Simulation 2

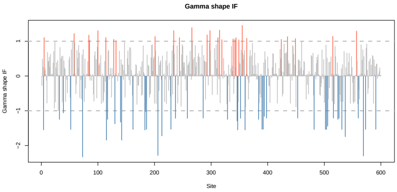


Simulation 3

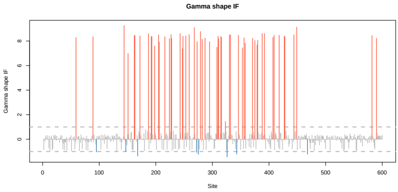
Actual data

Gamma Shape Parameter

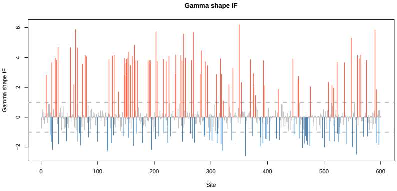
Simulation 1



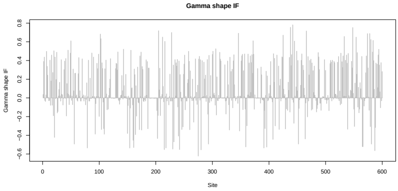
Simulation 2



Simulation 3

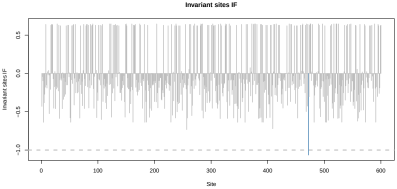


Actual data

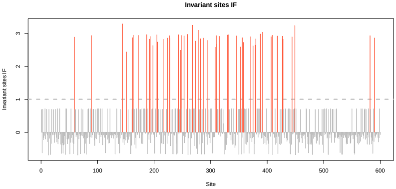


Invariant Sites Parameter IF

Simulation 1

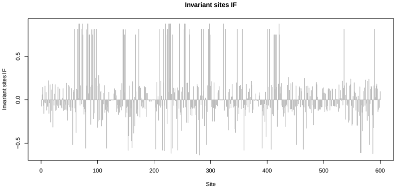


Simulation 2



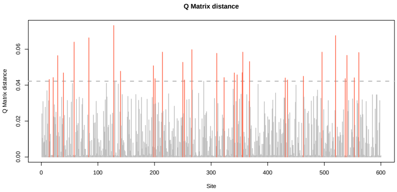
Simulation 3

Actual data

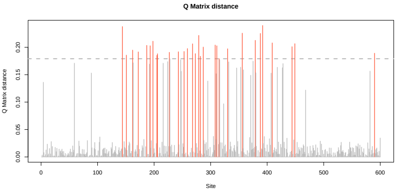


Q Matrix

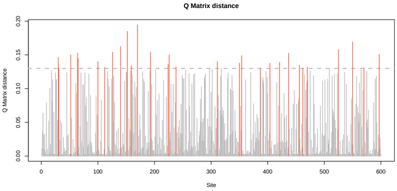
Simulation 1



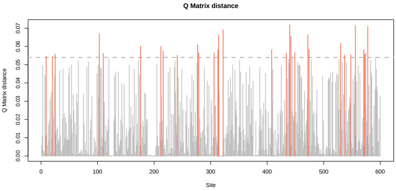
Simulation 2



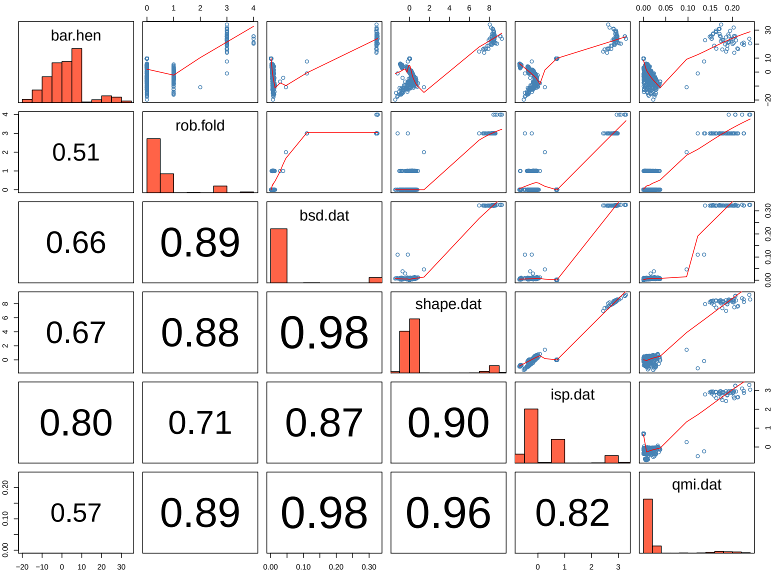
Simulation 3



Actual data

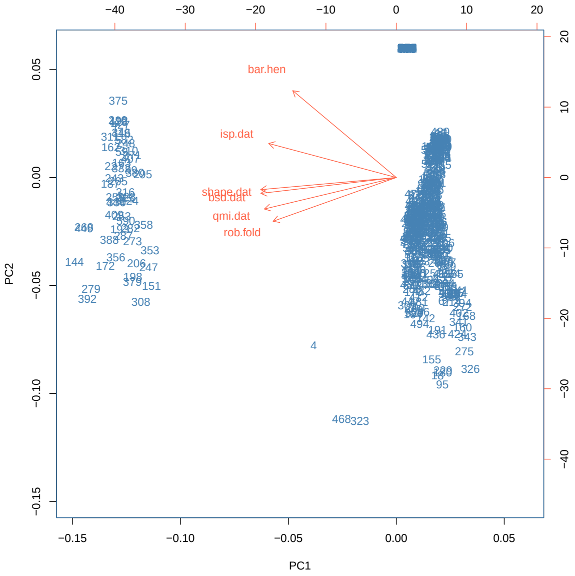


Other ideas

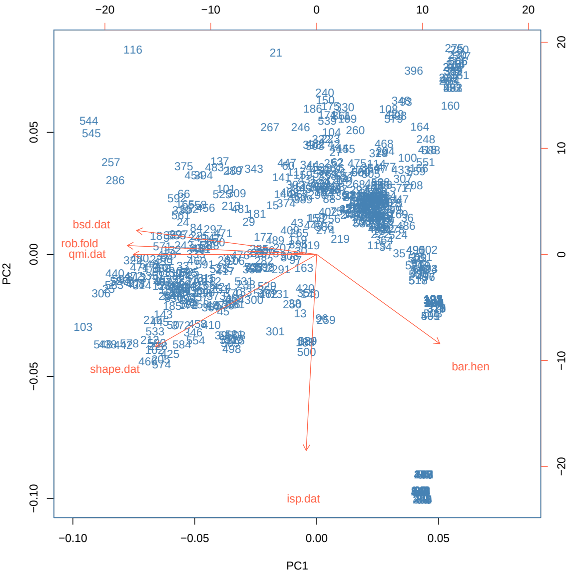


bar.hen
rob.fold
bsd.dat
shape.dat
isp.dat
qmi.dat
0.51
0.66
0.89
0.67
0.88
0.98
0.80
0.71
0.87
0.90
0.57
0.89
0.98
0.96
0.82

Other ideas



Other ideas



−20
−10
0
10
20
bsd.dat
rob.fold
qmi.dat
shape.dat
bar.hen
isp.dat
−0.10
−0.05
0.00
0.05
PC1
PC2

Resulting questions

- Which influence measures are essential?
- Are the single sites informations available from inferences as informative as leave-one-out measures?
- What is the unit of information? Site or taxon?
- Do site-wise comparisons make sense in the age of genomics?
- Should we use blocks of sites instead of single sites?
- What is an appropriate confidence interval for topologies?

Questions?