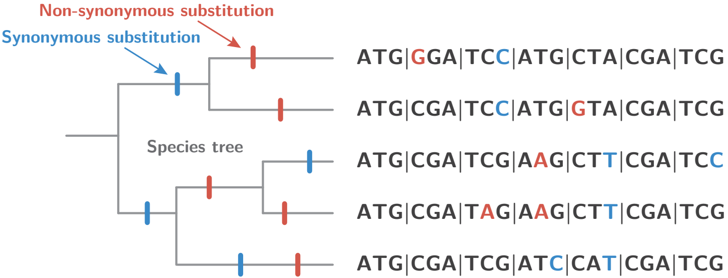
Bayesian Model selection: marginal likelihoods, cross-validation and information criteria

Nicolas Lartillot

June 10, 2022

Sequences as documents of evolutionary history



ATG|GGA|TCC|ATG|CTA|CGA|TCG

ATG|CGA|TCC|ATG|GTA|CGA|TCG

ATG|CGA|TCG|AAG|CTT|CGA|TCC

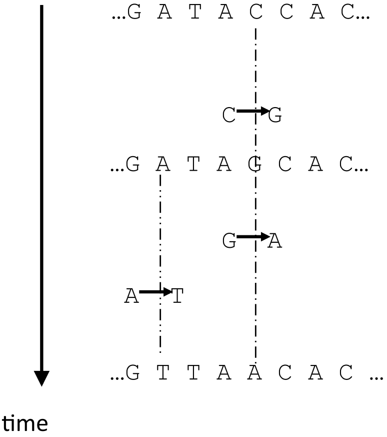
ATG|CGA|TAG|AAG|CTT|CGA|TCG

ATG|CGA|TCG|ATC|CAT|CGA|TCG

- reconstructing the phylogeny
- inferring, and testing hypotheses about, evolutionary processes

→ model-based approach

Probabilistic model of nucleotide substitution



...G A T A C C A C...

C → G

...G A T A G C A C...

G → A

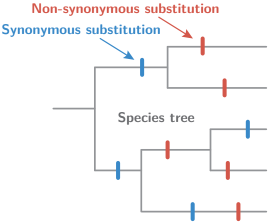
A → T

...G T T A A C A C ...

time

- all sites assumed to evolve independently
- under a continuous-time Markov model of nucleotide substitutions

Coding sequences: from nucleotides to amino-acids



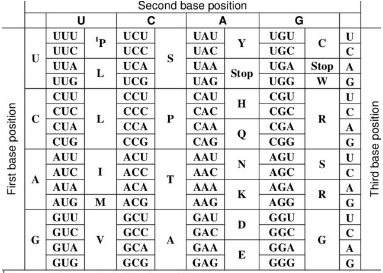
ATG|GGA|TCC|ATG|CTA|CGA|TCG

ATG|CGA|TCC|ATG|GTA|CGA|TCG

ATG|CGA|TCG|AAG|CTT|CGA|TCC

ATG|CGA|TAG|AAG|CTT|CGA|TCG

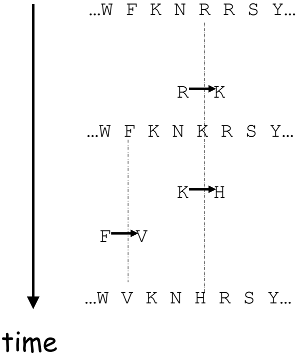
ATG|CGA|TCG|ATC|CAT|CGA|TCG



	Second base position								
First base position	U		C		A		G		Third base position
U	UUU	[1]P	UCU	S	UAU	Y	UGU	C	U
	UUC		UCC		UAC		UGC		C
	UUA	L	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG		UGG	W	G
C	CUU	L	CCU	P	CAU	H	CGU	R	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Q	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	I	ACU	T	AAU	N	AGU	S	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	K	AGA	R	A
	AUG	M	ACG		AAG		AGG		G
G	GUU	V	GCU	A	GAU	D	GGU	G	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	E	GGA		A
	GUG		GCG		GAG		GGG		G

[1] The one letter symbol of amino acids.

Probabilistic model of amino-acid replacement



...W F K N R R S Y...

R → K

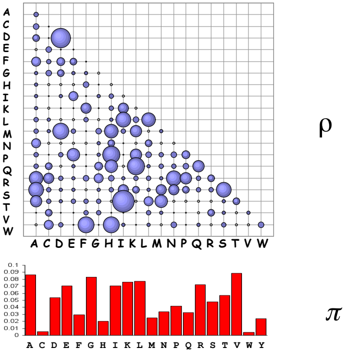
...W F K N K R S Y...

K → H

F → V

...W V K N H R S Y...

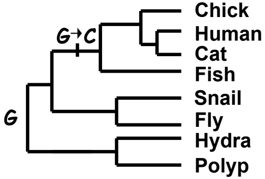
time



Instant rate matrix (Q) $Q_{lm} = \rho_{lm}\,\pi_m$

Bayesian phylogenetics

phylogenetic tree (T)



Observed sequence alignment (D)

```
A C A C A T T A
A G A C A T T A
A G A C A T T A
A C A C A T T A
T C A G A T C A
T A G G A T C A
A C A G G T C A
A C A G G T C A
```

$$p(\theta \mid D) \quad = \quad \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$$

θ: tree and model parameters

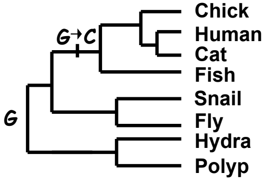
$p(\theta)$: prior (over tree and model parameters)

$p(D \mid \theta)$: likelihood (probability of data given tree and parameters)

$p(\theta \mid D)$: posterior (over tree and model parameters)

Bayesian phylogenetics

phylogenetic tree (*T*)



Observed sequence alignment (*D*)

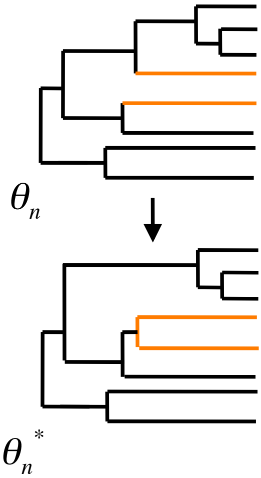
```
A C A C A T T A
A G A C A T T A
A G A C A T T A
A C A C A T T A
T C A G A T C A
T A G G A T C A
A C A G G T C A
A C A G G T C A
```

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$$

Commonly used priors

- uniform over tree topologies
- alternatively: birth-death process over tree
- generally: vague priors over continuous model parameters

Sampling from posterior by Markov Chain Monte Carlo



1. Propose a move $\theta_n \rightarrow \theta_n^*$

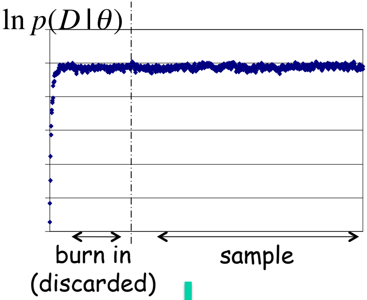
   According to kernel $q(\theta,\theta^*)$

2. Accept with probability

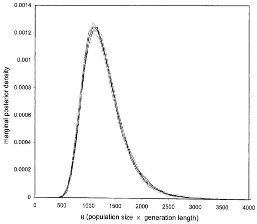
$$p = Min\left\{1, \frac{p(\theta_n^* \mid D)q(\theta_n^*,\theta_n)}{p(\theta_n \mid D)q(\theta_n,\theta_n^*)}\right\}$$

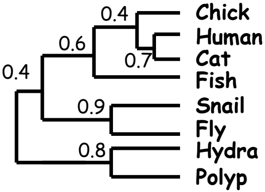
3. Iterate

Inference by marginalization of the posterior

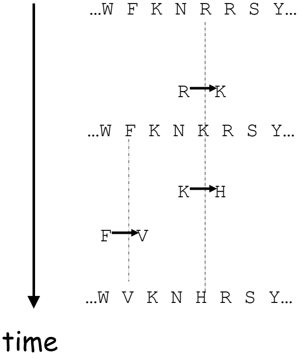


$$(\theta_k)_{k=1..K} \sim p(\theta \mid D)$$





Selecting among models of sequence evolution



…W F K N R R S Y…

R→K

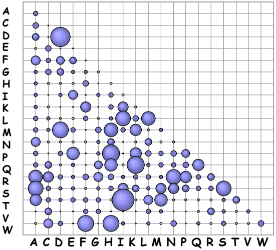
…W F K N K R S Y…

K→H

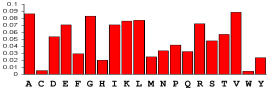
F→V

…W V K N H R S Y…

time



ρ



π

Instant rate matrix (Q) $Q_{lm} = \rho_{lm} \pi_m$

Amino-acid replacement matrices

- universal matrices pre-estimated on large datasets (JTT, LG)
- general time-reversible (GTR) model re-estimated on current data

→ should one use a universal matrix or re-estimate it on current data?

The different aims and meanings of model selection

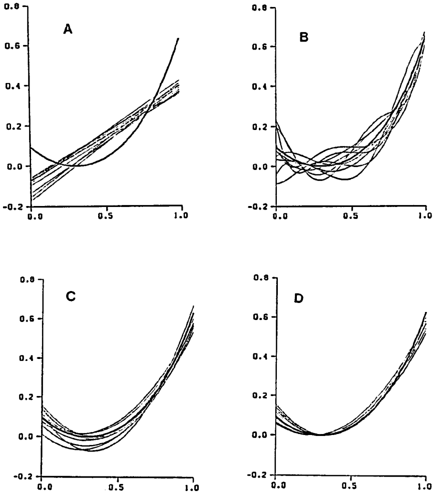
Hypothesis testing

- choosing between alternative hypotheses about processes
- frequentist: likelihood ratio tests
- Bayes: marginal likelihoods and model posterior probabilities
- 0/1 loss (false negatives / false positives)

Approximation

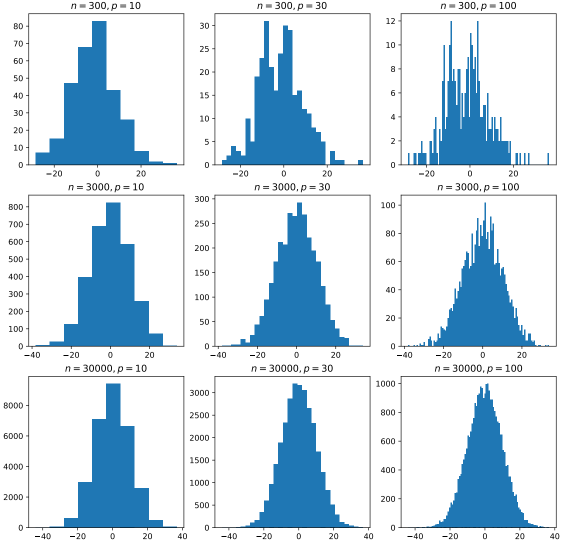
- aim is not true model identification, but accurate estimation
- minimizing quadratic error or information loss
- leave-one-out cross-validation
- information criteria of the Akaike family

Polynomial regression



Burnham and Anderson, 2002

Making a histogram: how many bins?



The information loss and risk

Information loss: Kullback-Leibler divergence

given two distributions F and G with densities f and g:

$$D(f,g) \quad = \quad E_{Y \sim F} \left[\ln f(Y) - \ln g(Y) \right] \geq 0$$

Information risk: expected information loss

- (unknown) true distribution F, of density f;
- based on data $X = (X_i)_{i=1..n} \sim F$, estimate histogram $\hat{f}_{p,X}$;
- define information risk (of using p when F is true) as:

$$R(p,F) \quad = \quad E_{X \sim F, Y \sim F} \left[\ln f(Y) - \ln \hat{f}_{p,X}(Y) \right]$$

where $Y \sim F$ would be a new data point from the same source

Estimating the information risk

$$R(p,F) = E_{X\sim F,Y\sim F}\left[\ln f(Y) - \ln \hat{f}_{p,X}(Y)\right]$$

- minimizing $R(p,f)$ w.r.t. p is equivalent to maximizing:

$$L(p,F) = E_{X\sim F,Y\sim F}\left[\ln \hat{f}_{p,X}(Y)\right]$$

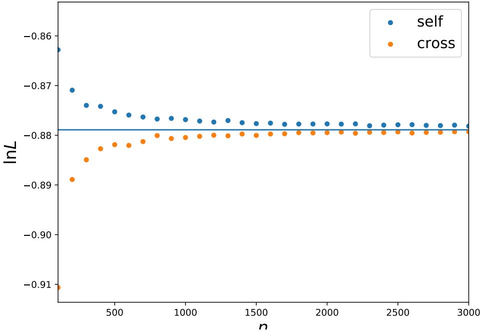
- self-consistent estimate

$$L_{self}(p,F) = \frac{1}{n}\sum_i \ln \hat{f}_{p,X}(X_i)$$

- leave-one-out cross-validation ($X_{(i)}$: data set with X_i removed):

$$L_{cross}(p,F) = \frac{1}{n}\sum_i \ln \hat{f}_{p,X_{(i)}}(X_i)$$

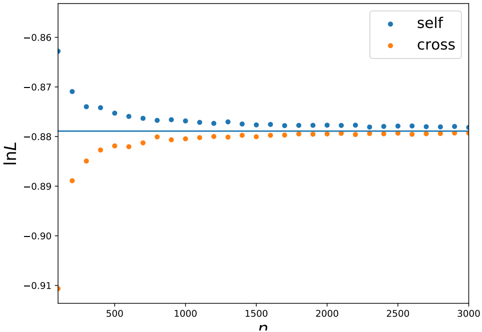
Self versus cross log likelihood



$$L_{cross}(p, F) \;=\; L_{self}(p, F) - \frac{p}{n}$$

p/n: optimism, or generalization gap

Self versus cross log likelihood



$$L_{cross}(p, F) = L_{self}(p, F) - \frac{p}{n}$$

$$AIC = -2L_{max} + 2p$$

p/n: optimism, or generalization gap

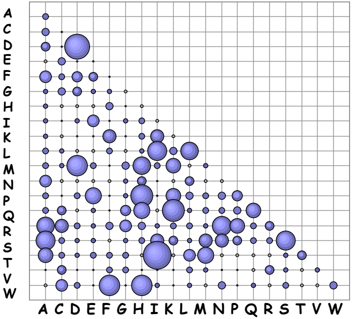
Variants on Akaike criterion

- AIC (Akaike): makes good model assumption
- TIC (Takeuchi): more general (accounts for model violation)
- RIC (Shibata) and GIC (Konishi): TIC for penalized likelihood
- DIC (Spiegelhatler et al): heuristic derivation in a Bayesian context
- wAIC (Watanabe): formal derivation in a Bayesian context
- BIC (Schwartz): not based on information-loss

Information criteria and cross-validation

- operationally, AIC-type criteria: asymptotic estimates of LOO-CV

Selecting amino-acid replacement models



Experiment

- M1: using existing ’universal’ empirical matrix (LG)
- M2: re-estimating the 190 exchange rates on current data (GTR)
- uniform prior over the 190 relative exchange rates (standard)
- comparing M1 and M2 increasingly large empirical datasets

The Bayesian leave-one-out (LOO-CV) score

Definition

- data $X = (X_i)_{i=1..n}$
- $X_{(i)}$: data X with entry X_i removed

$$CV = \frac{1}{n}\sum_{i=1}^{n} \ln p(X_i \mid X_{(i)})$$

Harmonic mean estimator based on posterior sample

$$\begin{aligned}\frac{1}{p(X_i \mid X_{(i)})} &= \int \frac{1}{p(X_i \mid \theta)}\, p(\theta \mid X) d\theta \\ &\simeq \frac{1}{K}\sum_{k=1}^{K} \frac{1}{p(X_i \mid \theta_k)}\end{aligned}$$

with $\theta_k \sim p(\theta \mid X)$ (Gelfand, 1992)

Marginal likelihood estimation: sequential Monte Carlo

Principle

- data $X = (X_i)_{i=1..n}$
- $X_{1:i}$: first i observations

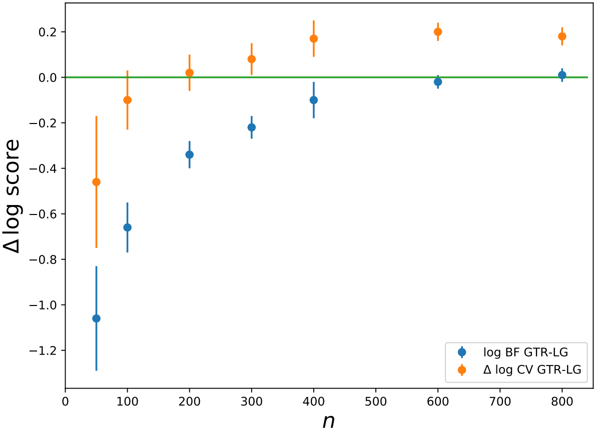
$$\begin{aligned} p(X) &= \prod_{i=1}^{n} p(X_i \mid X_{1:i-1}) \\ p(X_i \mid X_{1:i-1}) &\simeq \frac{1}{K} \sum_{k=1}^{K} \frac{1}{p(X_i \mid \theta_{ik})} \end{aligned}$$

with $\theta_{ik} \sim p(\theta \mid X_{1:i-1})$

Algorithm

- do a sequential MCMC, adding observations one by one
- at each step, run for a few cycles and estimate $p(X_i \mid X_{1:i-1})$

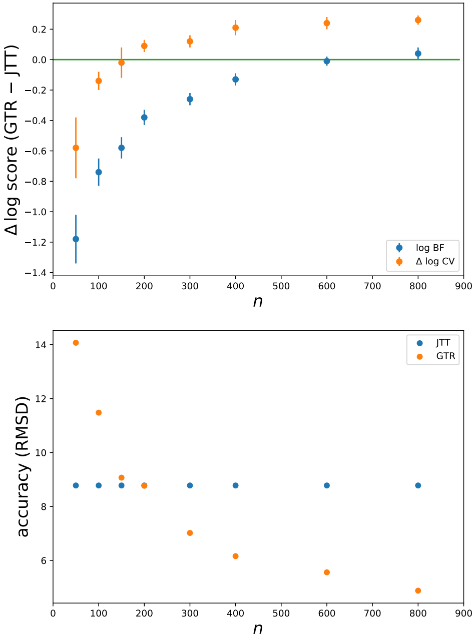
BF versus LOO-CV: empirical data



Simulation experiment

Experiment

- posterior predictive simulations under the LG model
- M1: using an empirical matrix different from LG (JTT)
- re-estimating the 190 exchange rates on current data (GTR)
- doing this on increasingly large simulated datasets


Δ log score (GTR − JTT)
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−1.2
−1.4
0
100
200
300
400
500
600
700
800
900
n
log BF
Δ log CV
accuracy (RMSD)
14
12
10
8
6
0
100
200
300
400
500
600
700
800
900
n
JTT
GTR

Watanabe's information criterion (wAIC)

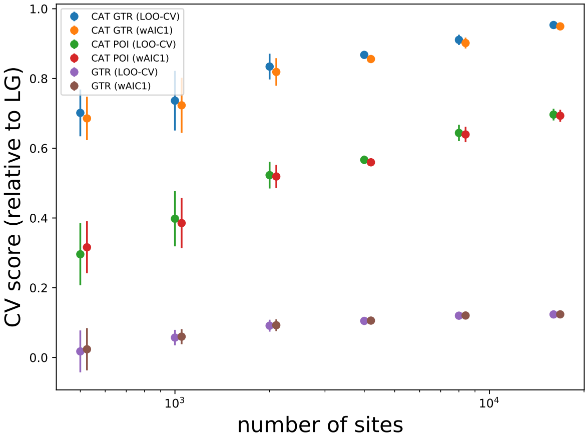
Principle

data $X = (X_i)_{i=1..n}$

$$\texttt{wAIC} \quad = \quad \frac{1}{n}\sum_i \ln E_{post}[p(X_i \mid \theta)] - \frac{1}{n}\sum_i V_{post}[\ln p(X_i \mid \theta)]$$

posterior expectation and variance estimated by MCMC

wAIC is a good approximation to LOO-CV



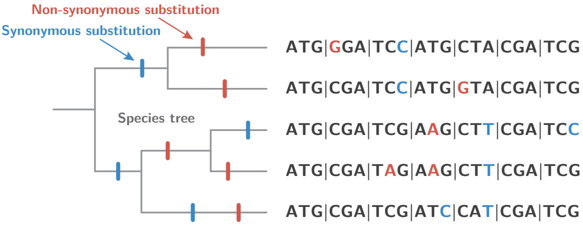
Summary 1

- LOO-CV better than BF for selecting best-approximating model
- BF is generally conservative, in particular under vague priors

not shown:

- wAIC (but not DIC) gives a good approximation to LOO-CV
- still better approach: prior centered on LG with tunable variance

Characterizing the selective regime: codon models



Non-synonymous substitution

Synonymous substitution

Species tree

ATG|GGA|TCC|ATG|CTA|CGA|TCG

ATG|CGA|TCC|ATG|GTA|CGA|TCG

ATG|CGA|TCG|AAG|CTT|CGA|TCC

ATG|CGA|TAG|AAG|CTT|CGA|TCG

ATG|CGA|TCG|ATC|CAT|CGA|TCG

R: 61 × 61 codon substitution matrix

$$\begin{aligned} R_{\mathrm{ACA}\to\mathrm{ACC}} &= \mu \\ R_{\mathrm{ACA}\to\mathrm{ATA}} &= \mu\,.\,\omega \\ R_{\mathrm{ACA}\to\mathrm{AGC}} &= 0\ \ldots \end{aligned}$$

- μ: mutation rate
- $\omega = dN/dS$: net effect of selection on non-synonymous changes
- if $\omega > 1$: positive selection (non-syn mutations are advantageous)

→ determine whether a given gene is under positive selection?

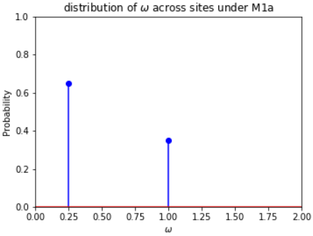
Gene-level dN/dS estimates

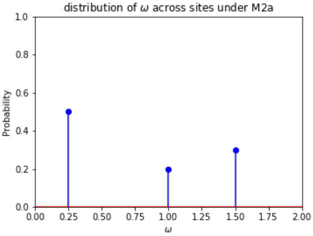
gene	dN/dS
ATP synthase	0.02
Albumine	0.23
SAMHD1	0.43
APOBEC	0.52
BRCA1	0.85
Interleukine 6	1.91

- genes like SAMHD1: implicated in defense against retroviruses
- likely under positive selection at least in part of its sequence
- method based on gene-level dN/dS insufficiently sensitive

→ modulating dN/dS over the sequence

Random-effect site models

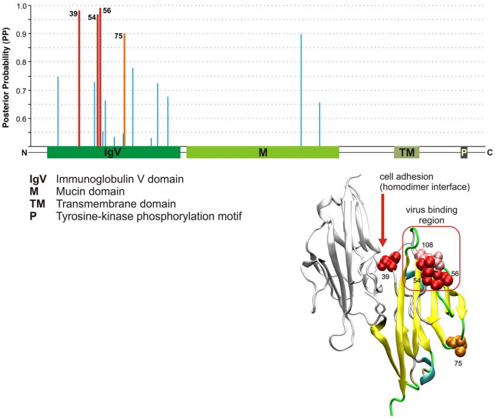




Model structure

- M1: site are iid from a 3-component distribution
- $\omega_- < 1$, $\omega_0 = 1$, $\omega_+ > 1$, with proportions π_-, π_0, π_+
- M0: $\pi_+ = 0$ (no site under positive selection)

The maximum likelihood / empirical Bayes approach



HAVCR1 gene, Kosiol et al 2008, PLoS Genet

- parameters (lengths, nucrates, ω's and π's) estimated by ML
- gene-level inference: likelihood ratio test between M0 / M1
- identification of positively selected sites by empirical Bayes

The Bayesian approach

A spike-and-slab prior on w_+

- with probability $1 - \pi_+$, $w_+ = 0$
- with probability π_+, $w_+ > 0$
- gene-level inference: posterior prob. that $w_+ > 0$
- identification of positively selected sites by hierarchical Bayes

Alternative priors

key parameters for effect size under M1: w_+ and $\Delta\omega_+ = \omega_+ - 1$

- $\pi_+ = 0.02$, 0.1 or 0.5
- informative: beta(1,9) on w_+ and expo(1) on $\Delta\omega_+$
- uninformative: beta(1,1) on w_+ and expo(10) on $\Delta\omega_+$
- hierarchical: π_+ and hyper-parameters shared across genes

Simulation experiment

- maximum likelihood implementation fitted on 1000 genes
- gene sequences re-simulated under M0 (90%) and M1 (10 %)
- → maximum likelihood and Bayesian analysis on these dta
- → accuracy and calibration (nominal versus true FDR)

The false discovery rate

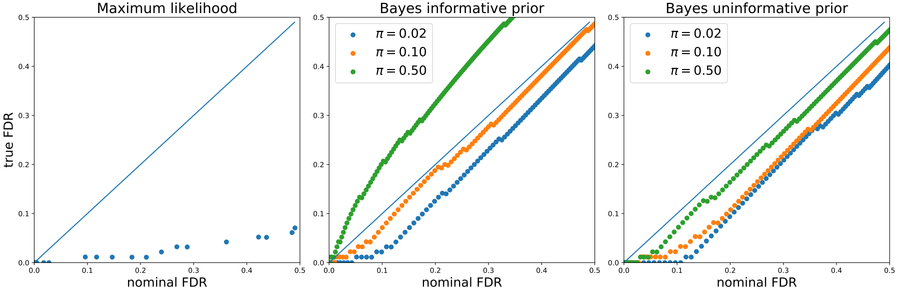
Based on p-values (Benjamini and Hochberg)

- rejecting null when $p < 0.01$
- null rejected for 40 out of 1000 genes
- at $\alpha = 0.01$, 10 false expected
- $\rightarrow$ nominal $FDR = 10/40 = 0.25$

Bayesian FDR estimate

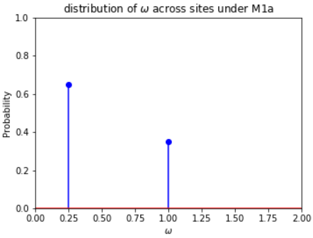
- rejecting null when post prob for alternative (pp) is > 0.80
- compute mean pp over selected genes $\overline{pp} = 0.92$
- $\rightarrow$ nominal $FDR = 1 - \overline{pp} = 0.08$

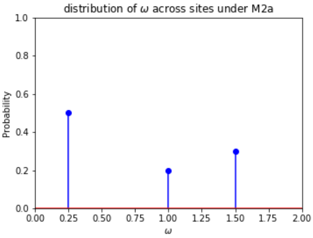
FDR calibration on simulated data



- maximum likelihood conservative; p-values under M0 are not χ^2_2

Random-effect site models



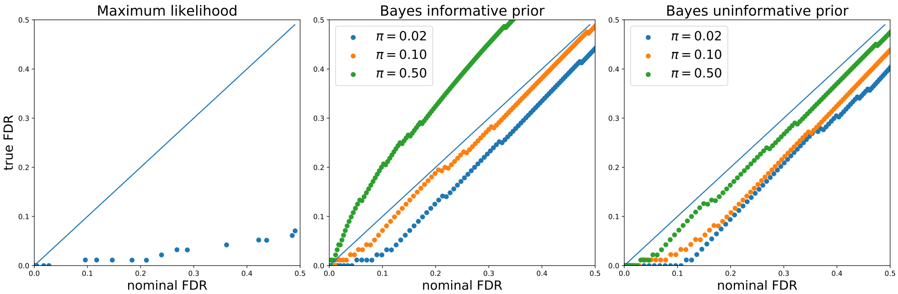


Model structure

- M1: site are iid from a 3-component distribution
- $\omega_- < 1$, $\omega_0 = 1$, $\omega_+ > 1$, with proportions π_-, π_0, π_+
- M0: $\pi_+ = 0$ (no site under positive selection)

M0 obtained by setting $w_+ = 0$ or $\omega_+ = 1 \rightarrow$ log-likelihood ratio not χ^2

FDR calibration on simulated data

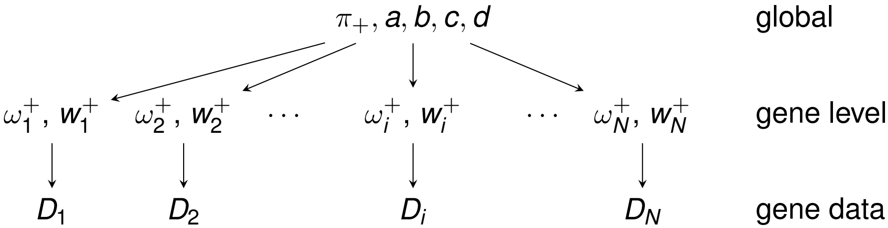


- Bayes: sensitive to π but also to the prior on effect size under M1
- best is prior with $\pi_+ = 0.1$, beta(1,1) on w_+ and expo(10) on $\Delta\omega_+$
- roughly corresponds to true prevalence and effect size distribution

Hierarchical models for exome-wide analyses

Gene-level prior

- with probability $1 - \pi_+$, $w_+ = 0$
- with probability π_+, $w_+ \sim beta(a, b)$
- $\omega_+ \sim 1 + gamma(c, d)$



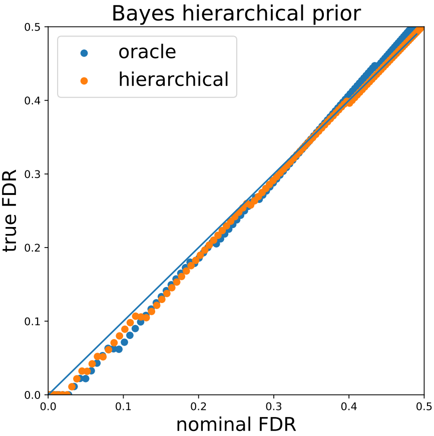
π_+, a, b, c, d global

ω_1^+, w_1^+ ω_2^+, w_2^+ $\cdots$ ω_i^+, w_i^+ $\cdots$ ω_N^+, w_N^+ gene level

D_1 D_2 D_i D_N gene data

- calibrating prior by sharing information across genes
- MPI parallelism, code with component design
- $\sim$ 10000 genes, $\sim$ 100 species: 24 / 48 hours on $\sim$ 1000 cores

FDR calibration on simulated data



Summary 2

Maximum likelihood and LRT

- in practice, null distribution of LRT can be complicated
- → standard frequentist FDR can difficult to calibrate

Bayes

- model posterior probabilities can have good frequentist properties
- but requires care on the calibration of the hyper prior
- possibility to calibrate on small subset of genes

Global summary

The two problems behind model selection . . .

- testing hypotheses (true model identification)
- approximation (best-approximating model identification)

. . . and their respective solutions

- LOO-CV / wAIC suitable for selecting best-approximating model
- model posterior probabilities (with calibrated priors): adequate for testing hypotheses
- marginal likelihoods or Bayes factors not adequate in either case

Frequentist properties of Bayes

- hierarchical setting: FDR-type frequentist properties
- uninformative setting: type-I error frequentist properties

Some references

- AIC - Akaike, H. 1974 A new look at the statistical model identification. IEEE Trans. Automat. Contr., 19(6), 716–723.
- TIC - Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. Mathematical Sciences, No. 153, pp. 12-8 (in Japanese).
- GIC - Konishi, S. and Kitagawa, G. 1996 Generalised information criteria in model selection. Biometrika, 83(4), 875–890.
- RIC - Shibata, R. 1989 Statistical aspects of model selection. In From data to model (ed. J. C. Willems), pp. 215–240. Springer New York.
- BIC - Schwarz, G. 2006 Estimating the Dimension of a Model. Ann. Statist., 6(2), 461–464.
- wAIC - Watanabe, S. 2010 Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. The Journal of Machine Learning Research, 11, 3571–3594.
- AIC and LOOCV - Stone, M. 1977 An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J. R. Statist. Soc. B, pp. 44–47.
- Cross-val not consistent - Shao, J. 1993 Linear Model Selection by Cross-Validation. Journal of the American Statistical Association, 88(422), 486–494.
- Burnham and Anderson. 2003. Model Selection and Multimodel Inference.
- Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. Ecology, 95(3), 631–636. http://doi.org/10.1890/13-1452.1
- Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), Psychol Methods. 1–27. http://doi.org/10.1037/a0027127