Inferring Potts models for evolutionary correlated data

Edwin Rodriguez, Pierre Barrat-Charlaix, Martin Weigt

Statistical modeling of protein sequences

Protein family

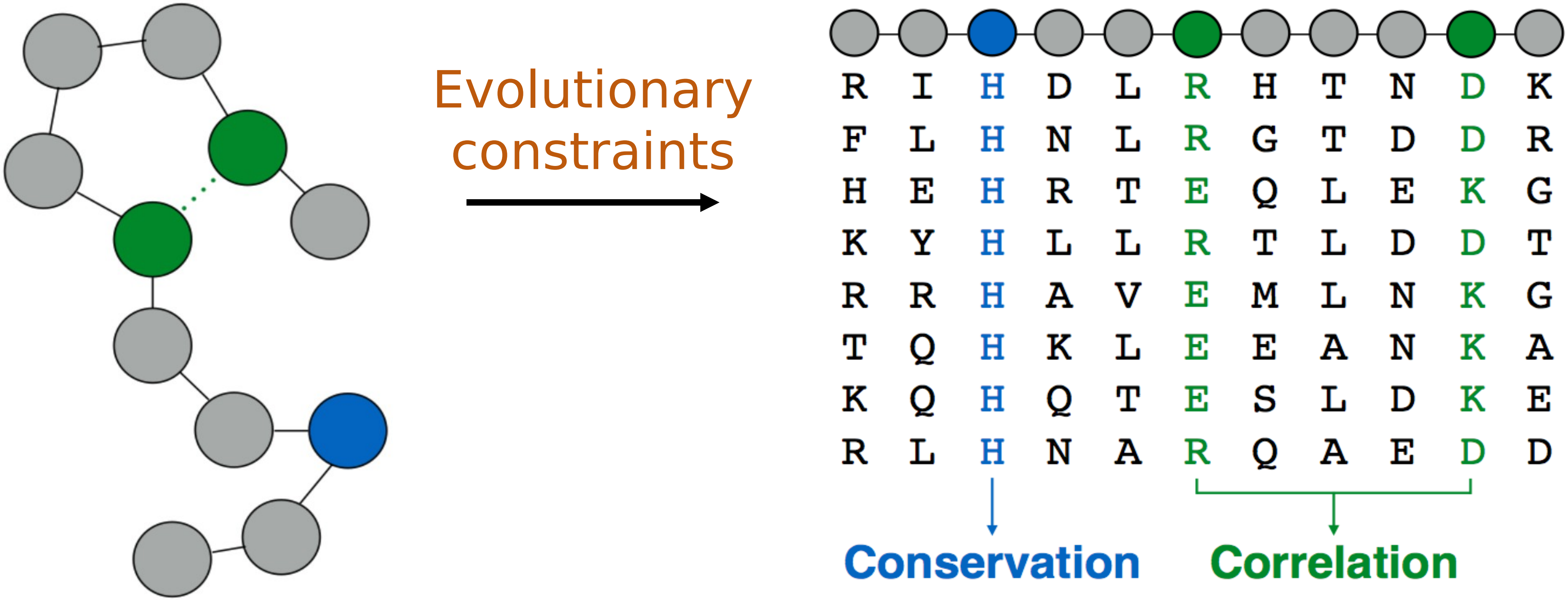
Evolutionary constraints

Multiple Sequence Alignment

```
...
YHCDKCSMSFAAPSRLNKHMRTH
HKCSYCSKAFIKKTLLKAHERTH
-QCEECGKQFAYSHSLKTHMMTH
YVCNVCGNLFRQHSTLTIHMRTH
-TCEFCGKNFERNGNYVEHRRTH
FVCGVCNKGFNSRTYLLEHMNKH
YVCHFCGKAVTNRESLKTHVRLH
YSCNVCDKSFTQRSSLVVHQRTH
FECQICGKSFKRSVQLKYHMEIH
YKCATCQKSFKRSQELKSHGKLH
HACGICGKTFPNNSSLEKHKHIH
YVCDKCGRSFSQRSSLTIHQRYH
YTCNVCGKTVTTKKSYTNHVKIH
FKCGVCGKFYKNESSLKTHSKIH
-QCEECGEIFNHKSSLNKHLLKH
YACEYCDKRFGDKQYLTQHRRVH
FKCDECGQCFSQRSSLNRHKRYH
YECDICGICFNQRSTMTSHRRSH
```

Information?

Global statistical model



Evolutionary constraints

```
R  I  H  D  L  R  H  T  N  D  K
F  L  H  N  L  R  G  T  D  D  R
H  E  H  R  T  E  Q  L  E  K  G
K  Y  H  L  L  R  T  L  D  D  T
R  R  H  A  V  E  M  L  N  K  G
T  Q  H  K  L  E  E  A  N  K  A
K  Q  H  Q  T  E  S  L  D  K  E
R  L  H  N  A  R  Q  A  E  D  D
```

Conservation

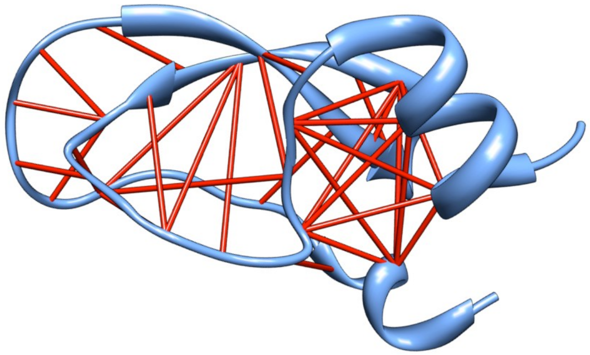
Correlation

$$P(a_1, \ldots, a_N) = \frac{1}{Z} \exp\left(\sum_{i,j=1}^{L} \overset{\text{Couplings}}{J_{ij}(a_i, a_j)} + \sum_{i=1}^{L} \overset{\text{Fields}}{h_i(a_i)} \right)$$

Direct Coupling Analysis (DCA)

Only information used is $f_i(a)$ and $f_{ij}(a, b)$

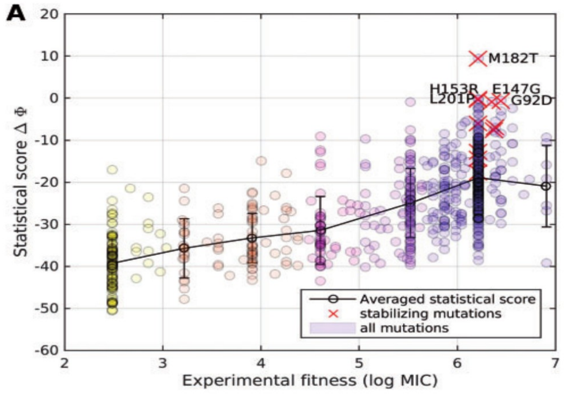
DCA: Successful model



- **Predicting 3D structure**

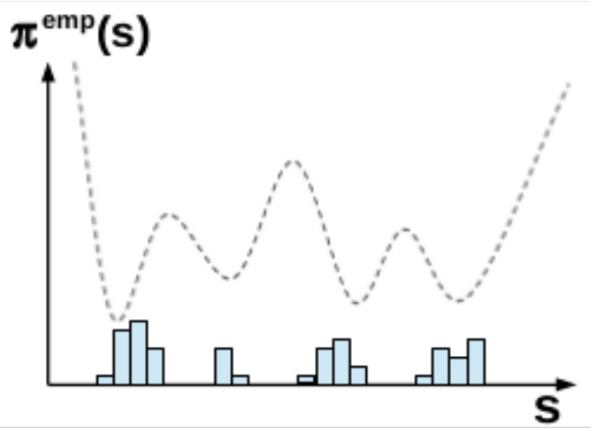
Morcos *et al.*, PNAS, 2011

Ovchinnikov *et al.*, Science, 2017



- **Predicting effect of mutations**

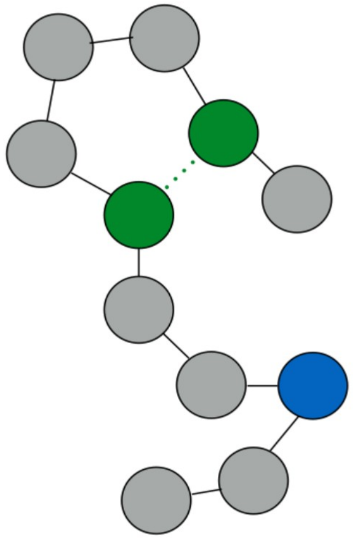
Figliuzzi *et al.*, MBE, 2015

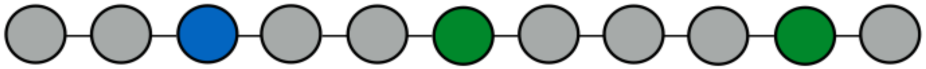


- **Designing new sequences**

Martin’s talk, this morning

Phylogenetic biases





R	I	H	D	L	R	H	T	N	D	K
F	L	H	N	L	R	G	T	D	D	R
H	E	H	R	T	E	Q	L	E	K	G
K	Y	H	L	L	R	T	L	D	D	T
R	R	H	A	V	E	M	L	N	K	G
T	Q	H	K	L	E	E	A	N	K	A
K	Q	H	Q	T	E	S	L	D	K	E
R	L	H	N	A	R	Q	A	E	D	D

Conservation **Correlation**

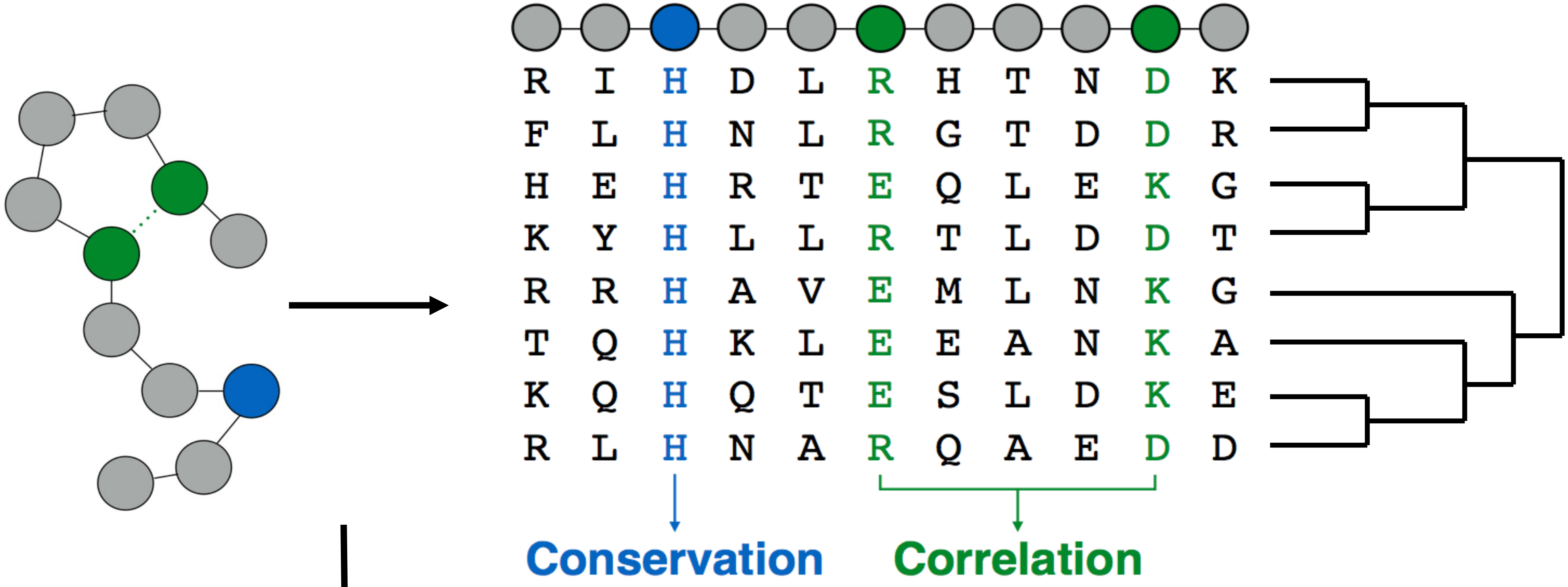
i.i.d. **sample**

$$f_i(a) \quad f_{ij}(a,b)$$

Couplings Fields

$$P(a_1,\dots,a_N) = \frac{1}{Z}\exp\left(\sum_{i,j=1}^{L} J_{ij}(a_i,a_j) + \sum_{i=1}^{L} h_i(a_i)\right)$$

Phylogenetic biases

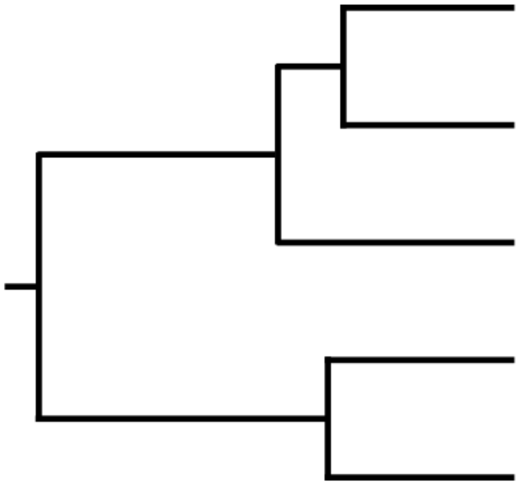


R	I	H	D	L	R	H	T	N	D	K
F	L	H	N	L	R	G	T	D	D	R
H	E	H	R	T	E	Q	L	E	K	G
K	Y	H	L	L	R	T	L	D	D	T
R	R	H	A	V	E	M	L	N	K	G
T	Q	H	K	L	E	E	A	N	K	A
K	Q	H	Q	T	E	S	L	D	K	E
R	L	H	N	A	R	Q	A	E	D	D

Biased statistics $f_i(a)$ $f_{ij}(a,b)$

Biased parameters $$P(a_1,\ldots,a_N)=\frac{1}{Z}\exp\left(\sum_{i,j=1}^{L}\overset{\text{Couplings}}{J_{ij}(a_i,a_j)}+\sum_{i=1}^{L}\overset{\text{Fields}}{h_i(a_i)}\right)$$

Phylogenetic biases



... R I **H** D L **R** H **T** N ...

... F L **H** N L **R** G **T** D ...

... H E **H** R T **E** Q **T** E ...

... K Y **D** L L **R** T **L** D ...

... R R **D** A V **E** M **L** N ...

→ Biased statistics $f_i(a)$ $f_{ij}(a, b)$

Phylogenetic tree → **Changes spectre of the correlation matrix**

Power law tails in phylogenetic systems
Qin & Colwell, 2017

Correcting for biases

Reweighting sequences

Sequence σ_i

Weight $w_i = 1/(\#\ seqs\ with\ >\ 80\%\ similarity\ to\ \sigma_i)$

Uncontrolled method...

Correcting for biases

Reweighting sequences

Sequence σ_i

Weight $w_i = 1/(\# \; seqs \; with \; > \; 80\% \; similarity \; to \; \sigma_i)$

Uncontrolled method...

Can we do better?

Given the phylogenetic tree…

- Principled way to correct statistics for phylogeny
- Translating this into a DCA model
- Assessing the quality of the method on artificial/protein data

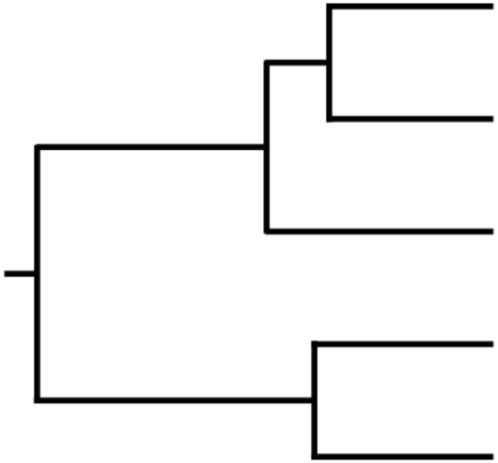
Maximum likelihood

…	R	I	**H**	D	L	**R**	H	**T**	N	…
…	F	L	**H**	N	L	**R**	G	**T**	D	…
…	H	E	**H**	R	T	**E**	Q	**T**	E	…
…	K	Y	**D**	L	L	**R**	T	**L**	D	…
…	R	R	**D**	A	V	**E**	M	**L**	N	…

Likelihood: *i.i.d.* sequences

$$\mathcal{L}(Data|J,h) = \prod_n P(\sigma_n|J,h)$$

Maximum likelihood



```
... R I H D L R H T N ...
... F L H N L R G T D ...
... H E H R T E Q T E ...
... K Y D L L R T L D ...
... R R D A V E M L N ...
```

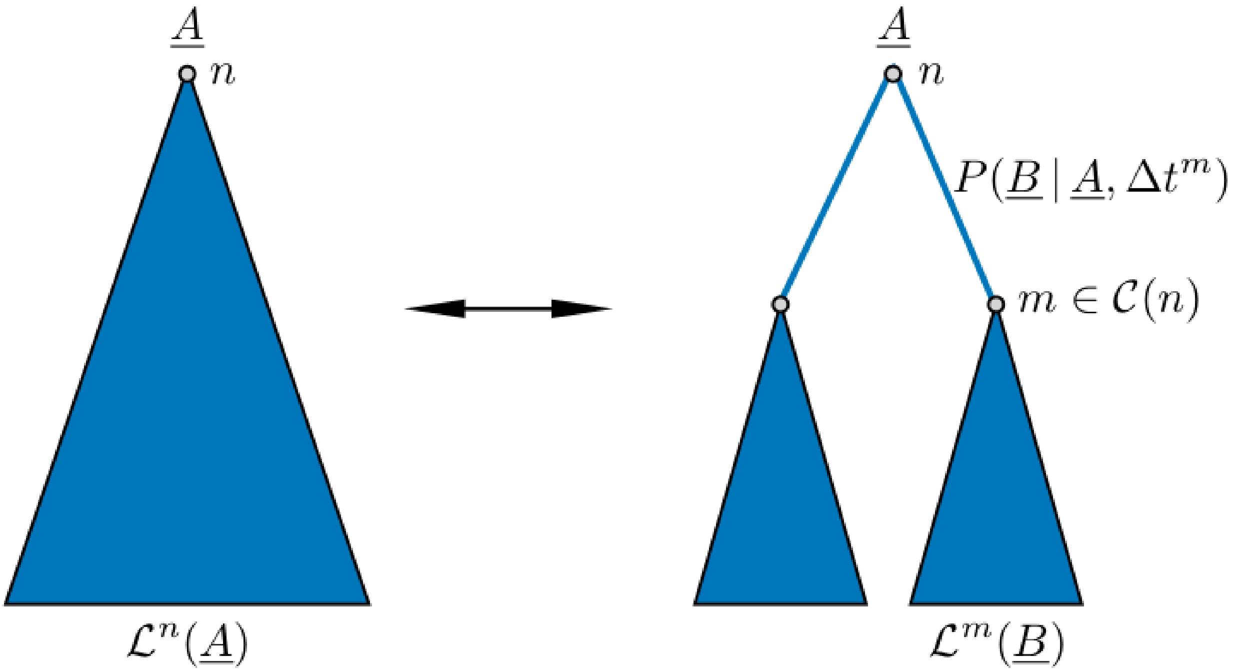
Likelihood

$$\mathcal{L}(Data|J,h) \neq \prod_n P(\sigma_n|J,h)$$

Correcting the likelihood

Evolutionary model (*i.e.* propagator) → $P(B|A, \Delta t)$

Felsenstein's pruning algorithm


$\underline{A}$
n
$\underline{A}$
n
$P(\underline{B}\,|\,\underline{A}, \Delta t^m)$
$m \in \mathcal{C}(n)$
$\mathcal{L}^n(\underline{A})$
$\mathcal{L}^m(\underline{B})$

$$\mathcal{L}^n(A) = \prod_{B \in \mathcal{C}(A)} \sum_{\{B\}} P(B|A, \Delta t)\mathcal{L}^m(B)$$

Evolutionary model

$$P(B|A,\Delta t) \quad \textbf{?}$$

Based on the Potts model?

$$P(a_1,\ldots,a_N) = \frac{1}{Z}\exp\left(\sum_{i,j=1}^{L} \overset{\text{Couplings}}{J_{ij}(a_i,a_j)} + \sum_{i=1}^{L} \overset{\text{Fields}}{h_i(a_i)}\right)$$

$$\downarrow$$

$$P(B|A,\Delta t,J,h)$$

Evolutionary model

$$P(B|A, \Delta t) \quad \textbf{?}$$

Based on the Potts model?

$$P(B|A, \Delta t, J, h)$$

→ **Two major problems**

- Sum of all possible trajectories from **A** to **B** → **Intractable**
- Sum over all configurations of internal nodes → **Intractable**

$$\mathcal{L}^n(A) = \prod_{B \in \mathcal{C}(A)} \boxed{\sum_{\{B\}}} P(B|A, \Delta t) \mathcal{L}^m(B)$$

→ **Need of an approximation**

Evolutionary model

Independent sites approximation:

“Real” frequency $\omega_i(A_i)$

Mutation rate μ

Position *i* of the alignment

$$P(B_i|A_i,\Delta t) = \underset{\text{No mutation}}{\underline{e^{-\mu\Delta t}\delta_{A_i,B_i}}} + \underset{\text{>1 mutation}}{\underline{(1-e^{-\mu\Delta t})\omega_i(B_i)}}$$

Evolutionary model

Independent sites approximation:

“Real” frequency $\omega_i(A_i)$

Mutation rate μ

Position *i* of the alignment

$$P(B_i|A_i,\Delta t) = \underbrace{e^{-\mu\Delta t}\delta_{A_i,B_i}}_{\text{No mutation}} + \underbrace{(1-e^{-\mu\Delta t})\omega_i(B_i)}_{\text{>1 mutation}}$$

Likelihood

$$\mathcal{L}_i^n(A_i|\omega_i) = \prod_{B\in\mathcal{C}(A)} \sum_{\{B_i\}} P(B_i|A_i,\Delta t)\mathcal{L}_i^m(B_i|\omega_i)$$

→ Cannot account for correlations!

Evolutionary model

Independent pairs approximation: “Real” frequency $\omega_{ij}(A_i, A_j)$

Pairs (*i*,*j*) evolve independently of each other

$$\begin{aligned}
P(B_i, B_j | A_i, A_j, \Delta t) = {} & \overset{\text{No mutation}}{\overline{e^{-2\mu\Delta t}\delta_{A_i,B_i}\delta_{A_j,B_j}}} \\
& + \underset{\text{One mutation}}{\underline{e^{-\mu\Delta t}(1 - e^{-\mu\Delta t})\left(\omega_{ij}(B_i|A_i)\delta_{A_j,B_j} + \omega_{ij}(B_j|A_j)\delta_{A_i,B_i}\right)}} \\
& + \underset{\text{>2 mutations}}{\underline{(1 - e^{-\mu\Delta t})^2\omega_{ij}(B_i, B_j)}}
\end{aligned}$$

With constraints $\forall j, \ \sum_b \omega_{ij}(a, b) = \omega_i(a)$

and $\forall i, \ \sum_a \omega_{ij}(a, b) = \omega_j(b)$

Evolutionary model

Independent sites approximation: “Real” frequency $\omega_i(A_i)$

$$P(B_i|A_i,\Delta t) = \underbrace{e^{-\mu\Delta t}\delta_{A_i,B_i}}_{\text{No mutation}} + \underbrace{(1-e^{-\mu\Delta t})\omega_i(B_i)}_{\text{>1 mutation}}$$

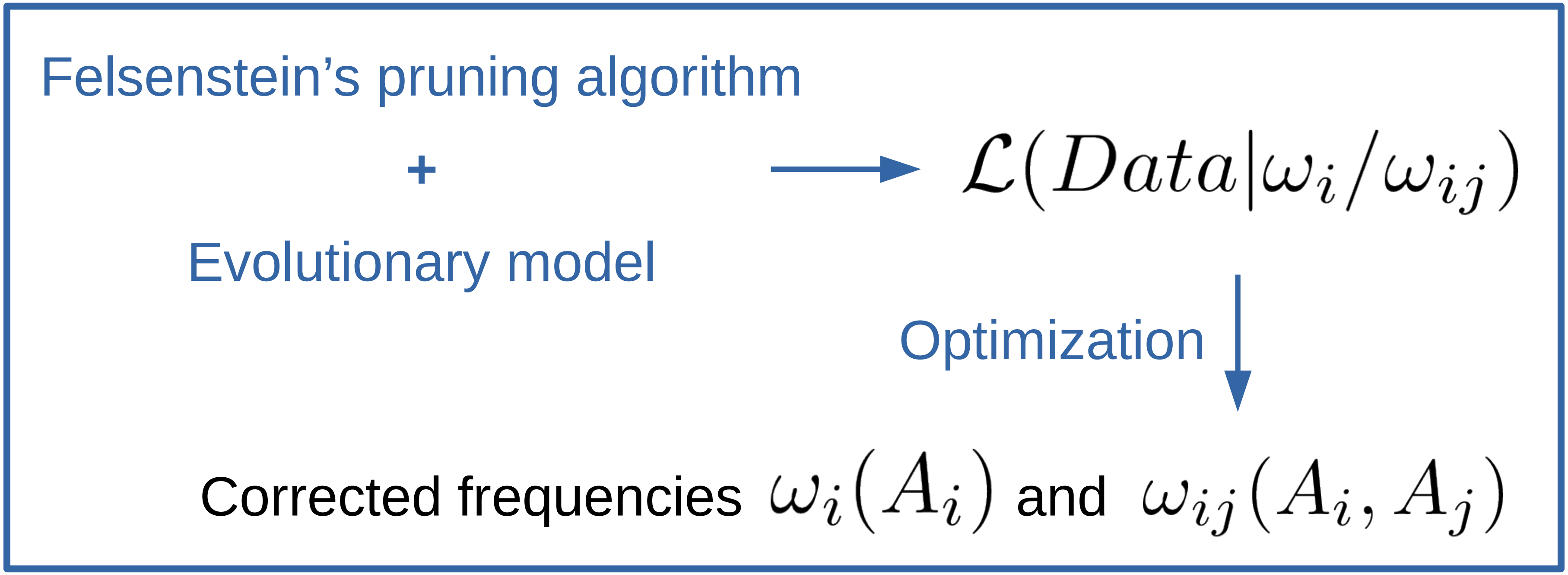
→ Cannot account for correlations!

Independent pairs approximation: “Real” frequency $\omega_{ij}(A_i, A_j)$

$$\begin{aligned}P(B_i,B_j|A_i,A_j,\Delta t) = {} & \overbrace{e^{-2\mu\Delta t}\delta_{A_i,B_i}\delta_{A_j,B_j}}^{\text{No mutation}} \\ & \underbrace{+\, e^{-\mu\Delta t}(1-e^{-\mu\Delta t})\left(\omega_{ij}(B_i|A_i)\delta_{A_j,B_j} + \omega_{ij}(B_j|A_j)\delta_{A_i,B_i}\right)}_{\text{One mutation}} \\ & \underbrace{+\,(1-e^{-\mu\Delta t})^2\omega_{ij}(B_i,B_j)}_{\text{>2 mutations}}\end{aligned}$$

Correcting for phylogenetic effects

- Principled way to correct statistics for phylogeny



Felsenstein's pruning algorithm

+

Evolutionary model

$$\longrightarrow \mathcal{L}(Data|\omega_i/\omega_{ij})$$

Optimization

Corrected frequencies $\omega_i(A_i)$ and $\omega_{ij}(A_i, A_j)$

- Translating this into a DCA/Potts model
- Assessing the quality of the method on artificial data

Testing the method: artificial data

Potts model

$$P^0(\sigma) \propto e^{-\mathcal{H}^0(\sigma)}$$

Sparse couplings

$$\mathcal{H}^0(\sigma) = -\sum_{i<j} \boxed{J^0_{ij}(\sigma_i, \sigma_j)} - \sum_{i=1}^{L} h^0_i(\sigma_i)$$

Tree

Propagator

Testing the method: artificial data

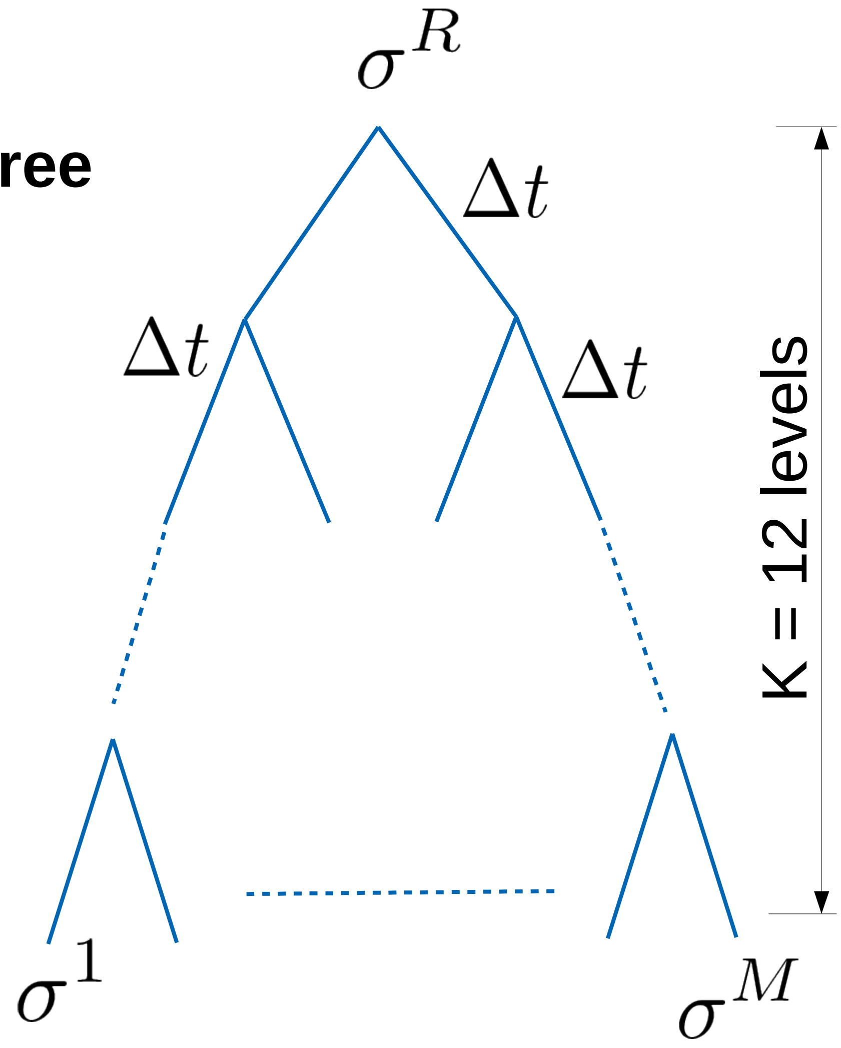
Potts model

$$P^0(\sigma) \propto e^{-\mathcal{H}^0(\sigma)}$$

Sparse couplings

$$\mathcal{H}^0(\sigma) = -\sum_{i<j} \boxed{J^0_{ij}(\sigma_i, \sigma_j)} - \sum_{i=1}^{L} h^0_i(\sigma_i)$$

Tree



Propagator

- # Mutations per branch

$$\mu L \Delta t = 3$$

- New state after mutation

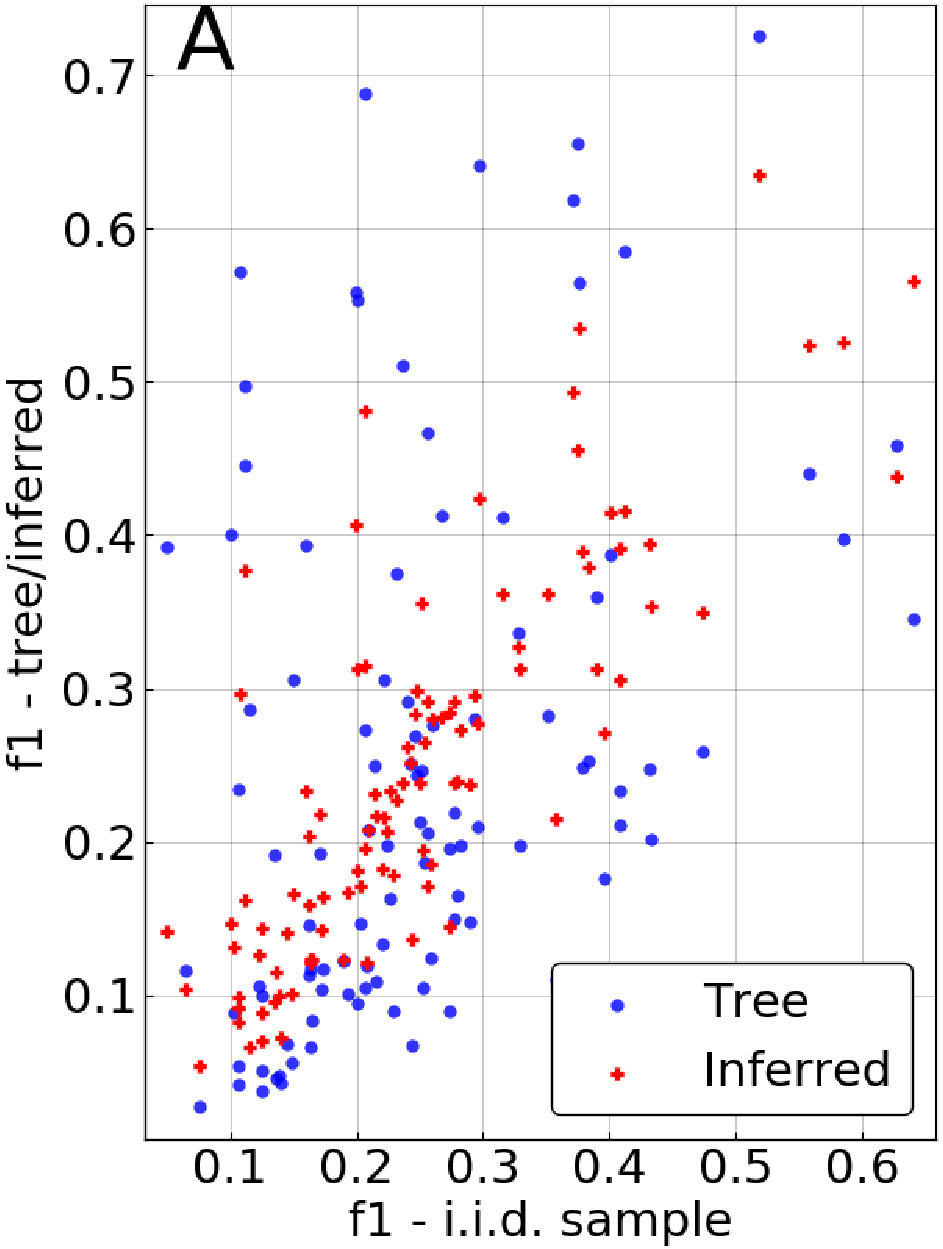
$$P^0(\sigma_i | \sigma_{\setminus i})$$

~Gibbs sampling

- 30 repetitions, different σ^R

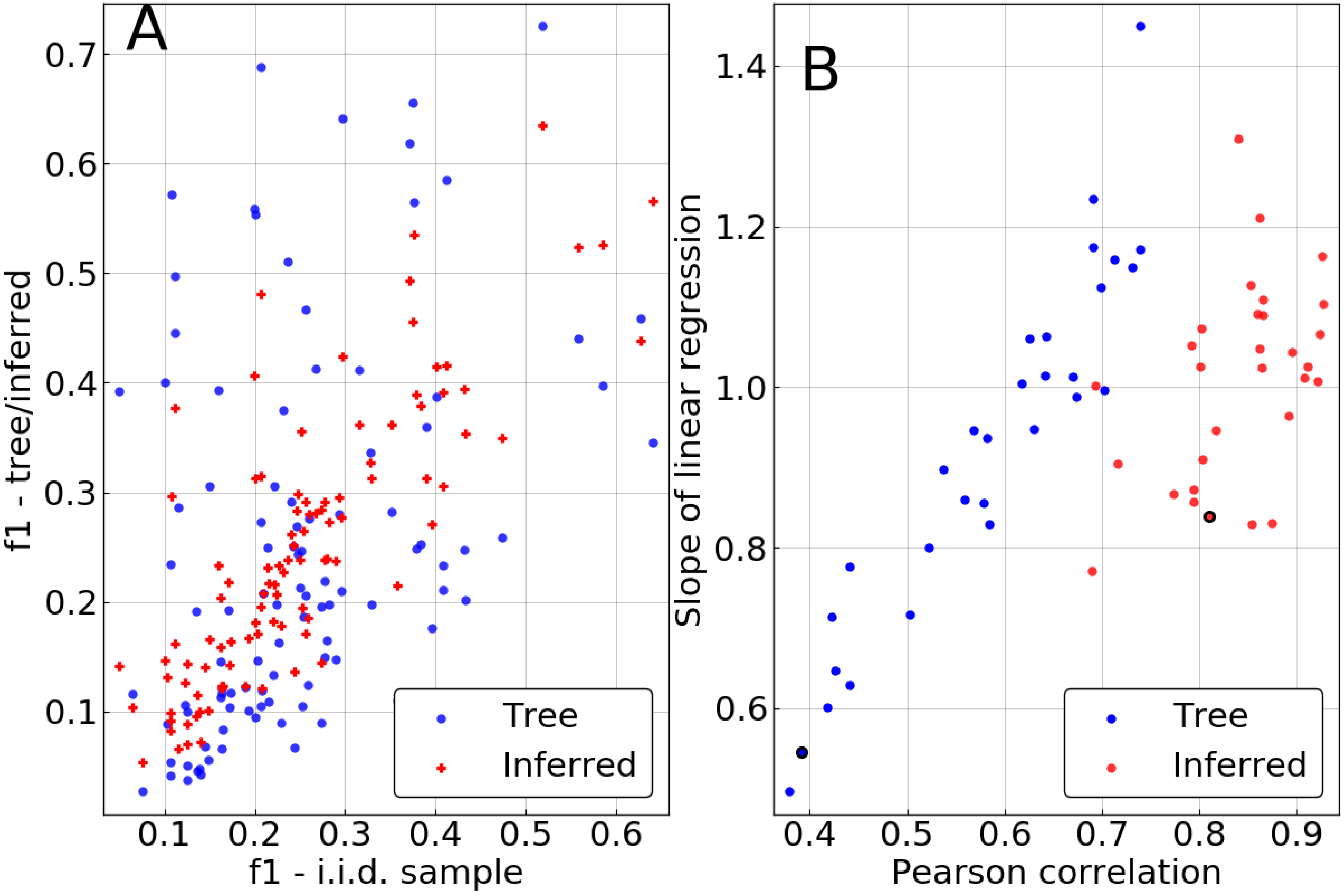
Phylogenetic inference corrects statistics

Single site frequencies ω_i : inferred vs true



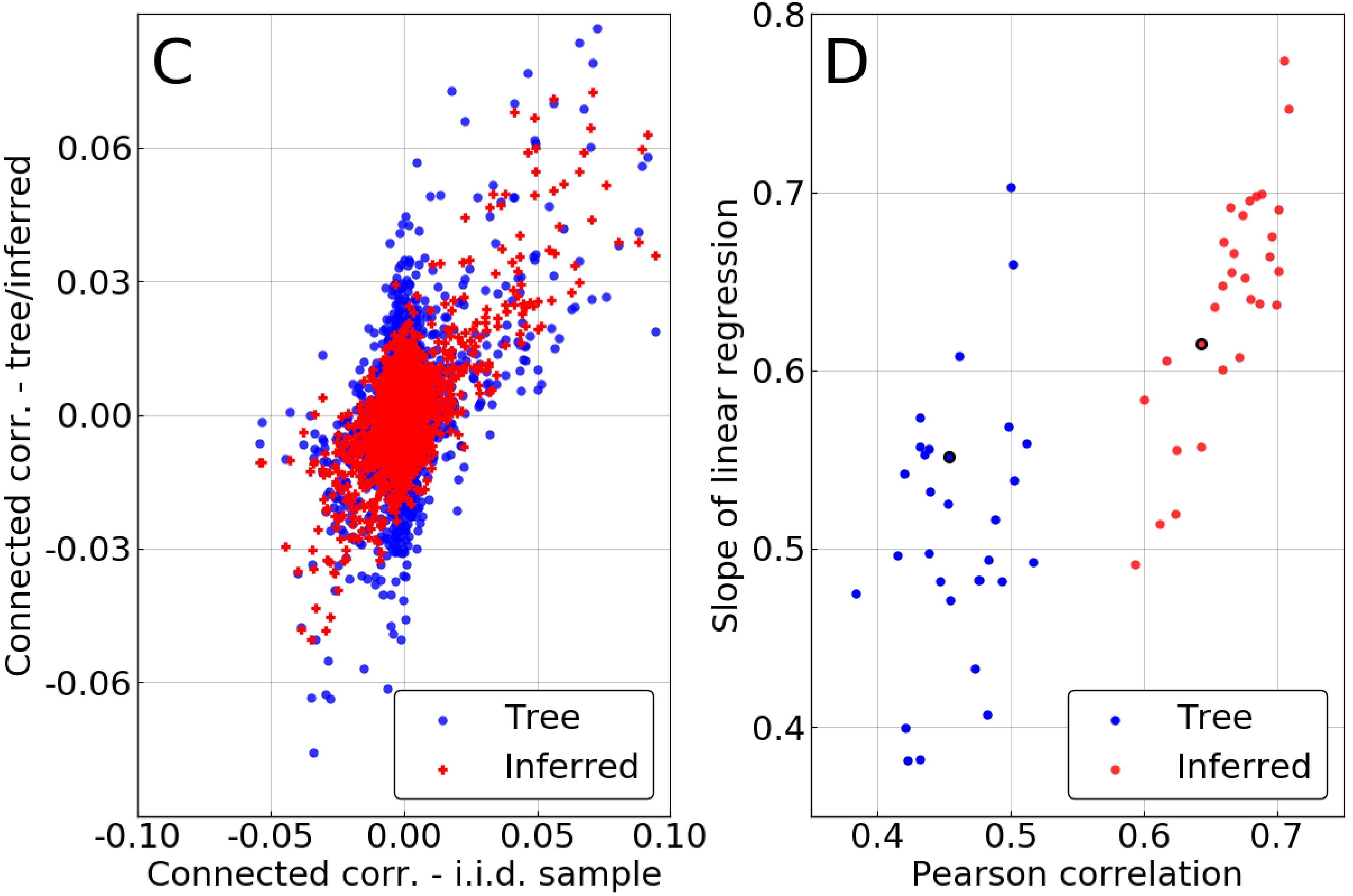
Phylogenetic inference corrects statistics

Single site frequencies ω_i : inferred vs true

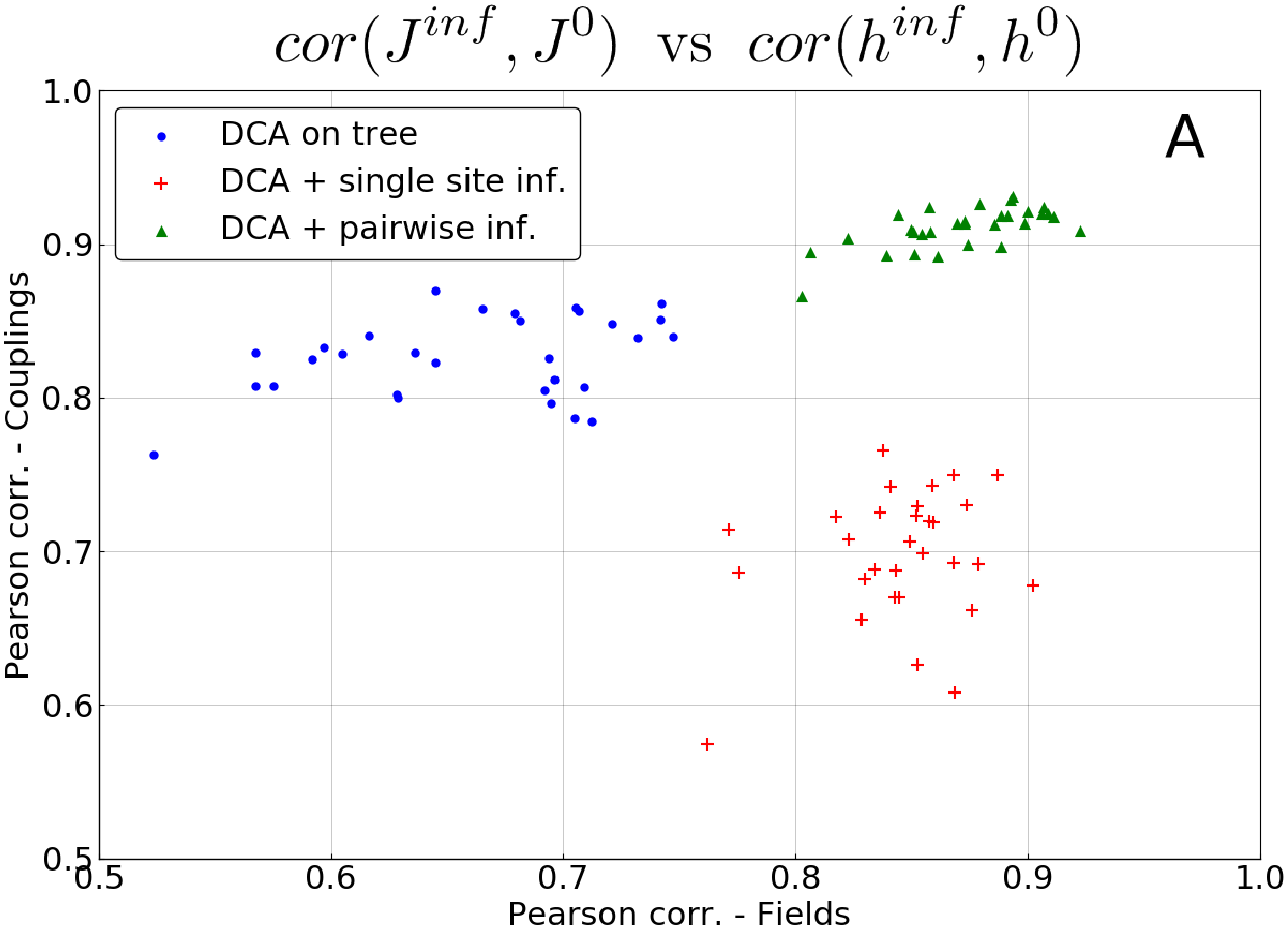


Phylogenetic inference corrects statistics

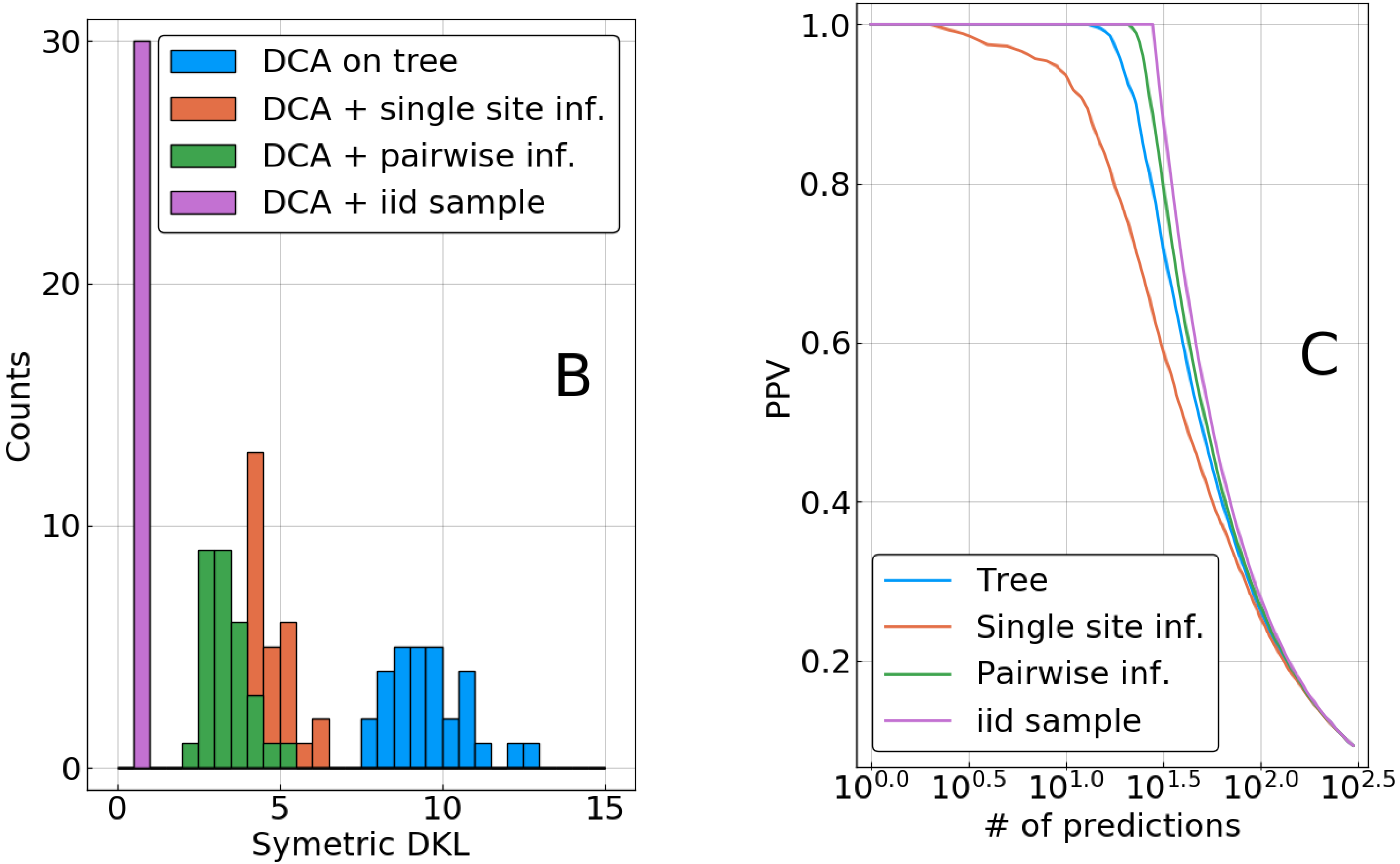
Connected correlations $\omega_{ij} - \omega_i \omega_j$: inferred vs true



Improved DCA parameters



Improved DCA parameters



Prediction of mutational effects

Single mutations from “wild-type” sequence

σ^1 ↓↑↑↑↑↑↑↑ E^1

σ^2 ↑↓↑↑↑↑↑↑ E^2

…

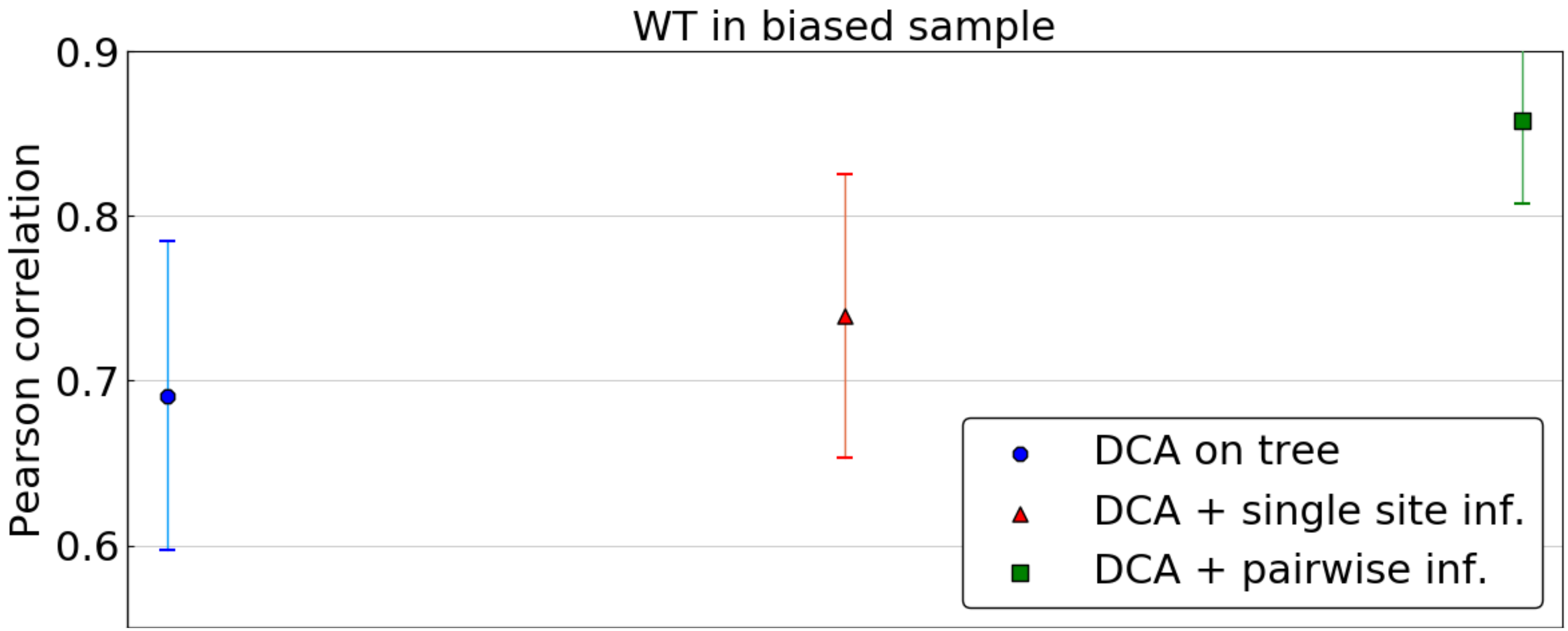
σ^K ↑↑↑↑↑↑↑↓ E^K

→

Quality of prediction

$$cor(E^i, \mathcal{H}^{inf}(\sigma^i))$$

Prediction of mutational effects



Single mutations from "wild-type" sequence

σ^1 ↓↑↑↑↑↑↑↑ E^1

σ^2 ↑↓↑↑↑↑↑↑ E^2

...

σ^K ↑↑↑↑↑↑↑↓ E^K

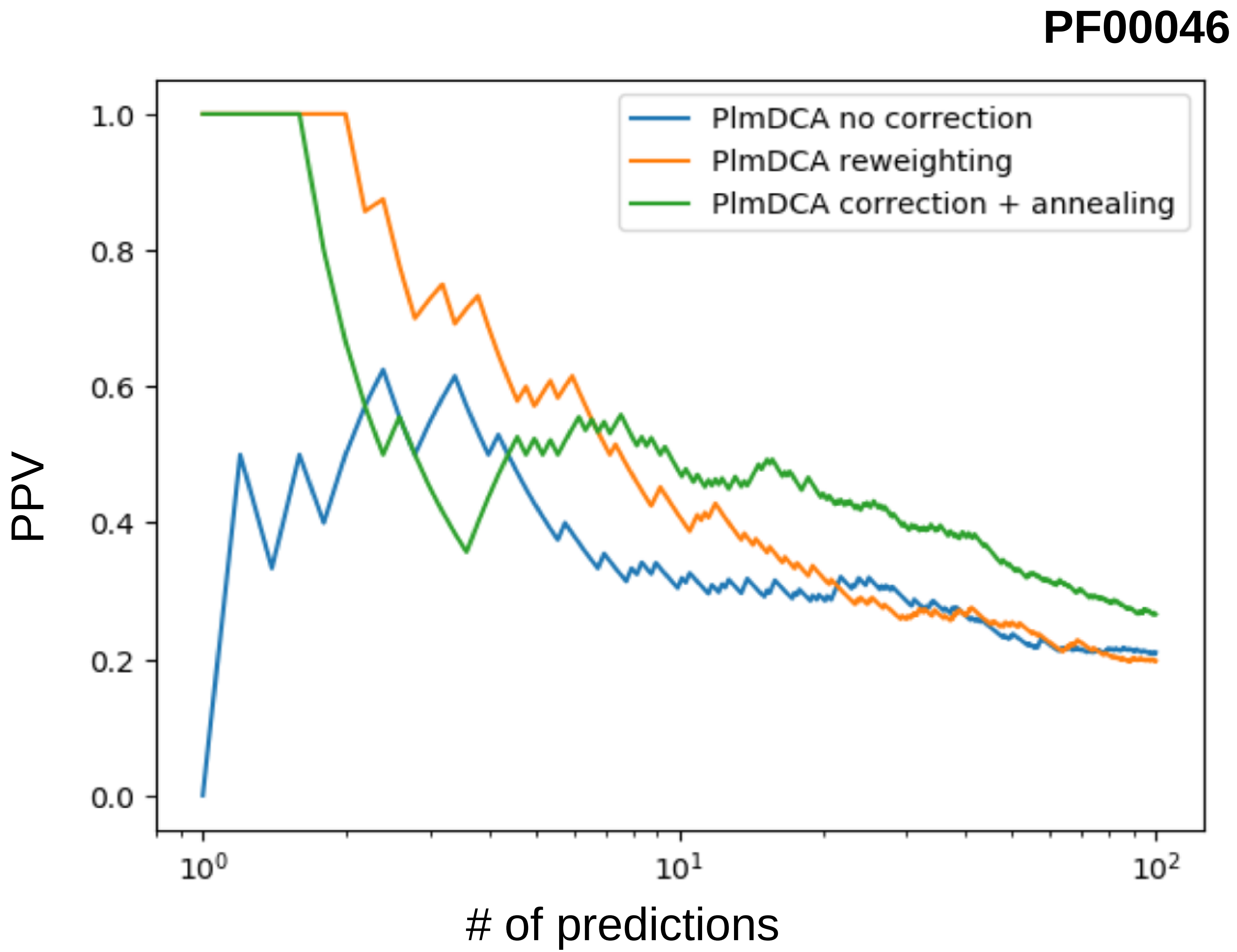
→

Quality of prediction

$$cor(E^i, \mathcal{H}^{inf}(\sigma^i))$$

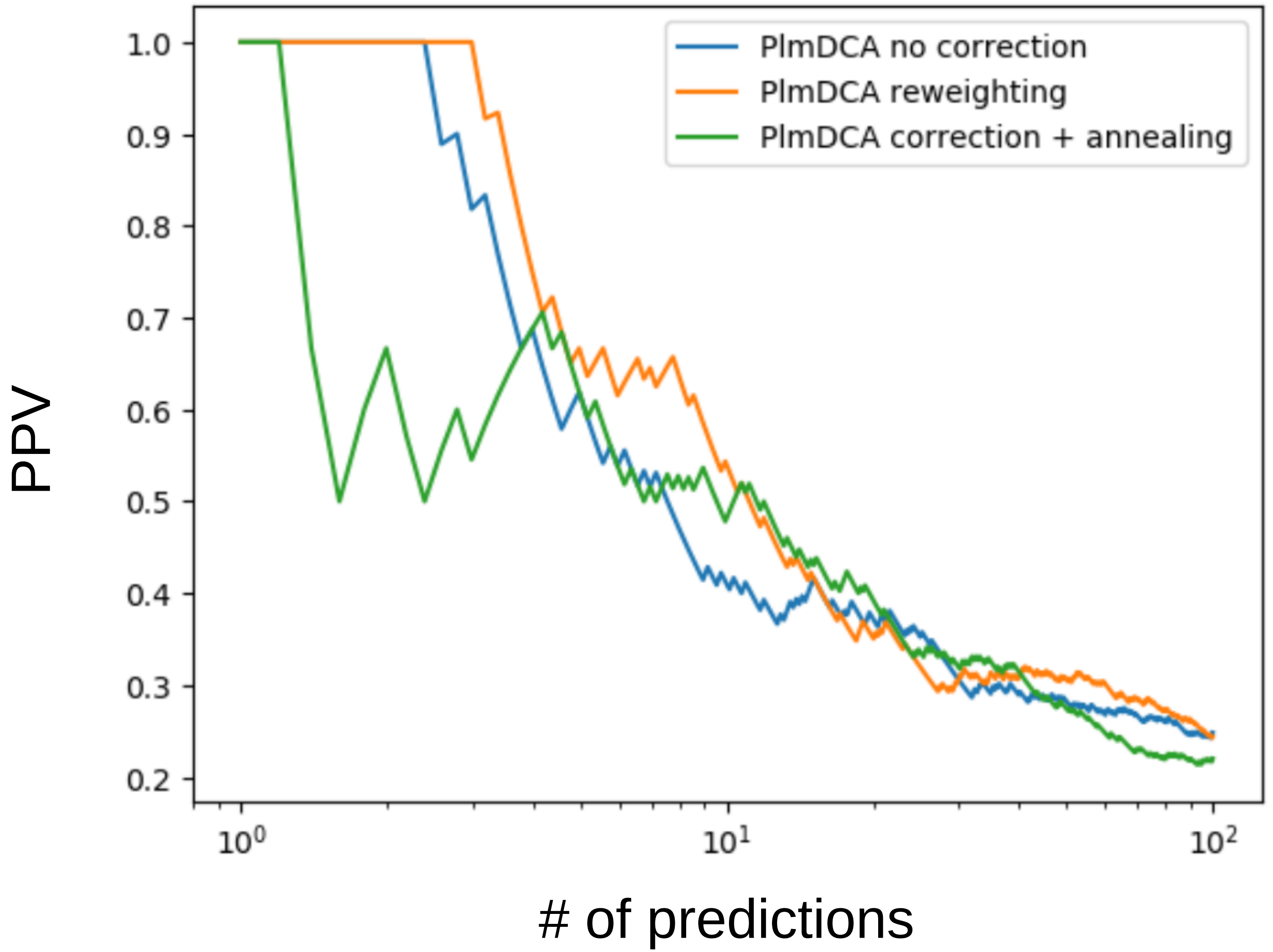
What about protein families?

Contact prediction in protein families



Contact prediction in protein families

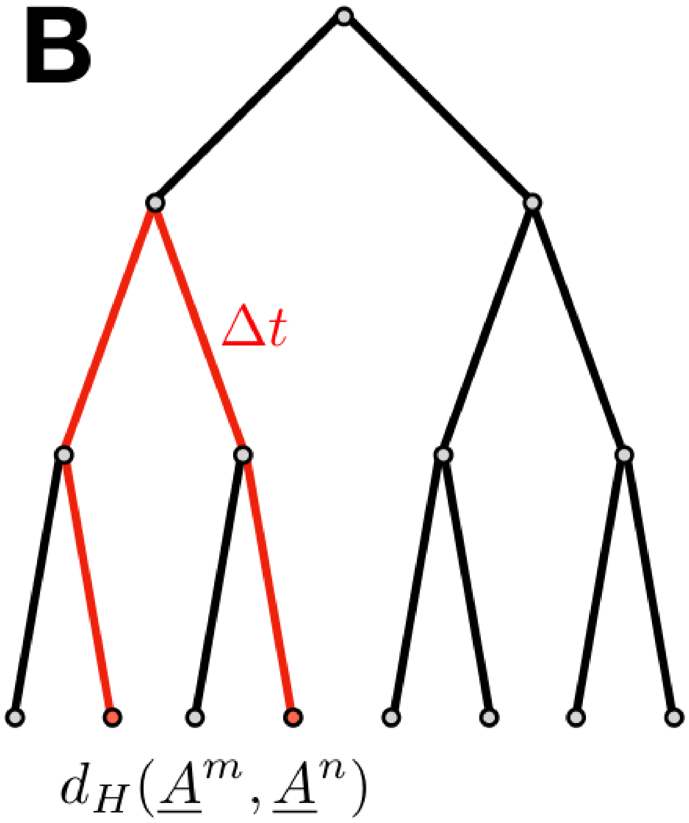
PF00084

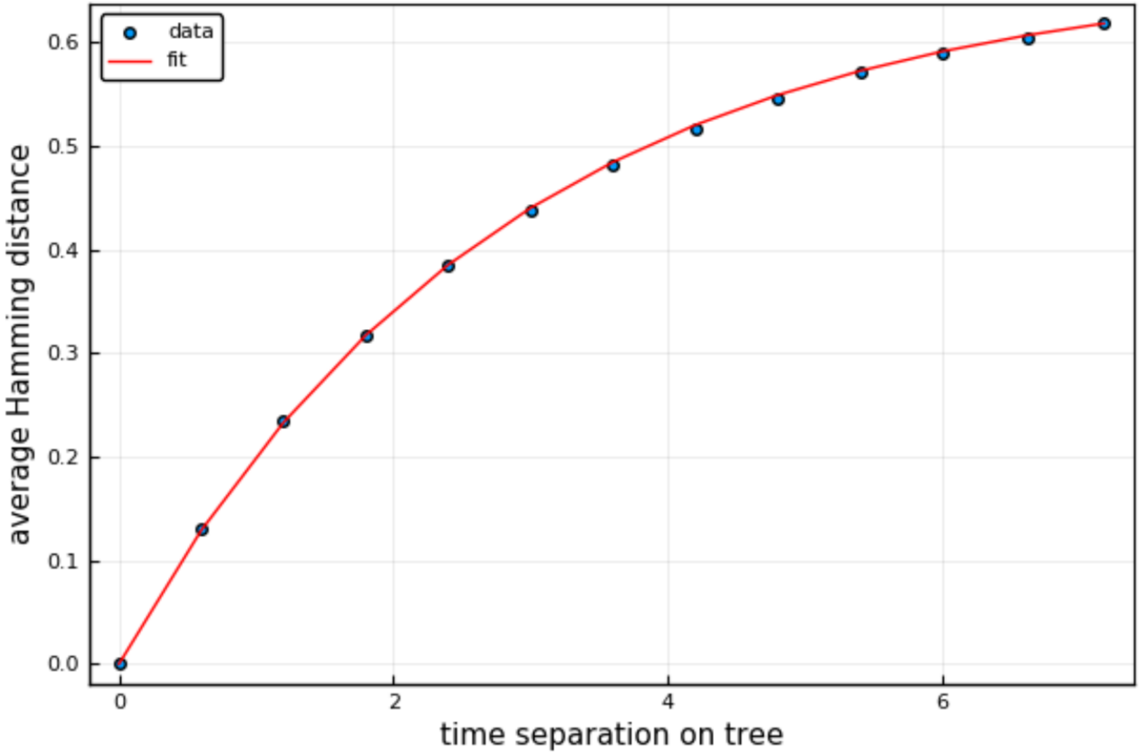


→ **Mixed results...**

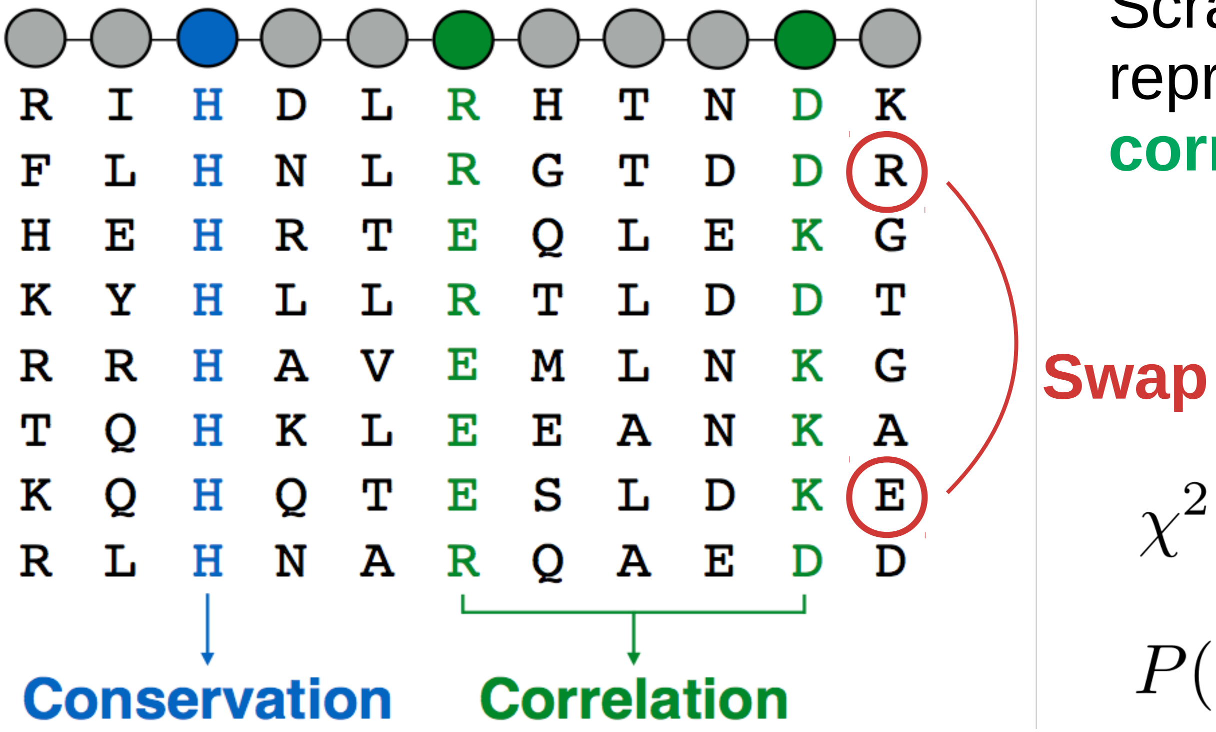
Thank you!

Fitting mu


B
Δt
$d_H(\underline{A}^m, \underline{A}^n)$


data
fit
average Hamming distance
time separation on tree
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6

Alignment from frequencies



R	I	H	D	L	R	H	T	N	D	K
F	L	H	N	L	R	G	T	D	D	R
H	E	H	R	T	E	Q	L	E	K	G
K	Y	H	L	L	R	T	L	D	D	T
R	R	H	A	V	E	M	L	N	K	G
T	Q	H	K	L	E	E	A	N	K	A
K	Q	H	Q	T	E	S	L	D	K	E
R	L	H	N	A	R	Q	A	E	D	D

Scrambling the alignment to reproduce **conservation** and **correlation**

$$\chi^2 = ||C - C^{target}||$$

$$P(\vec{a}) \propto e^{-\beta\chi^2} \quad \text{and} \quad \beta \to 0$$

Bialek & Ranganathan, arXiv, 2007