Ancestral protein reconstruction using autoregressive generative models

Matteo De Leonardis, Andrea Pagnani, **Pierre Barrat-Charlaix**

DISAT, Politecnico di Torino

Politecnico di Torino

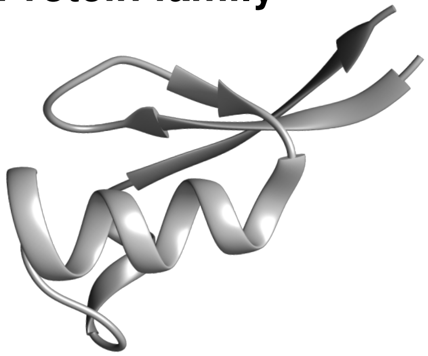
1859

Collaborators
Andrea Pagnani (PoliTo)
Matteo De Leonardis (PoliTo)

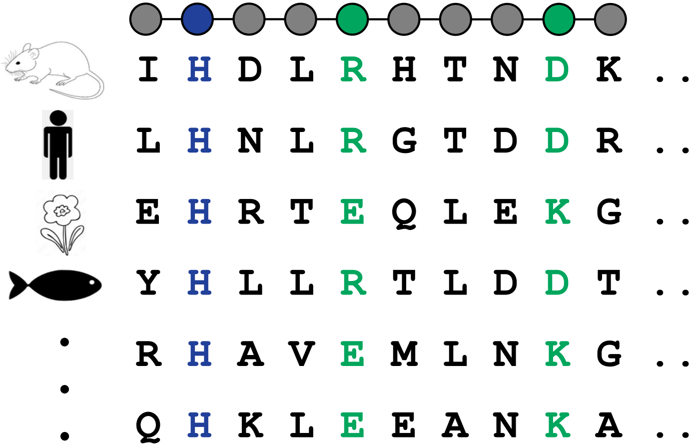
Statistical modeling of protein sequences

Multiple Sequence Alignment (MSA)

Protein family



Evolutionary constraints



In different species:

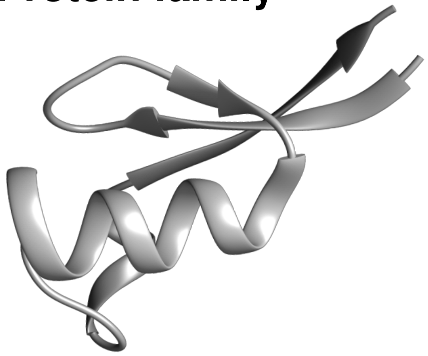
~ same structure

~ same function

- Regulation (DNA-binding, protein inhibitor…)
- Signaling (two-component signaling)
- Fundamental (ribosome…)
- Antibiotic resistance

Statistical modeling of protein sequences

Protein family

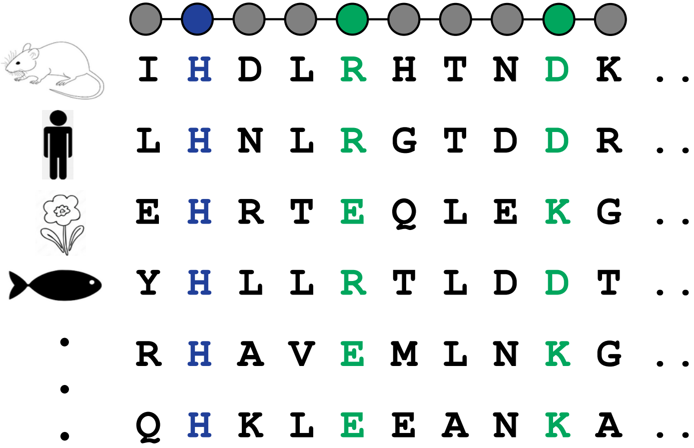


In different species:
~ same structure
~ same function

- Regulation (DNA-binding, protein inhibitor…)
- Signaling (two-component signaling)
- Fundamental (ribosome…)
- Antibiotic resistance

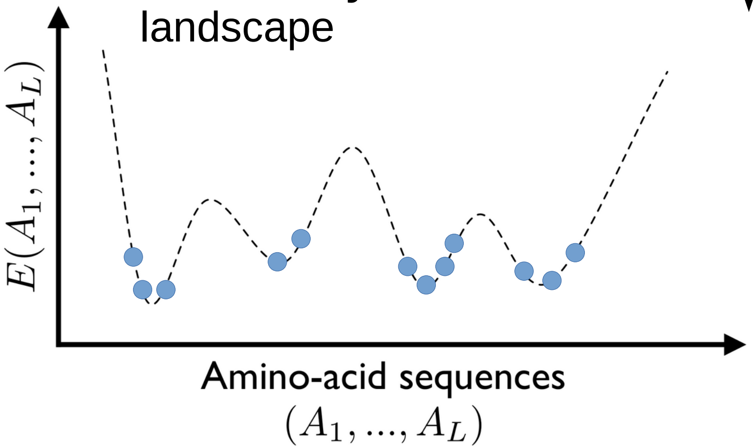
Evolutionary constraints →

Multiple Sequence Alignment (MSA)



Learning

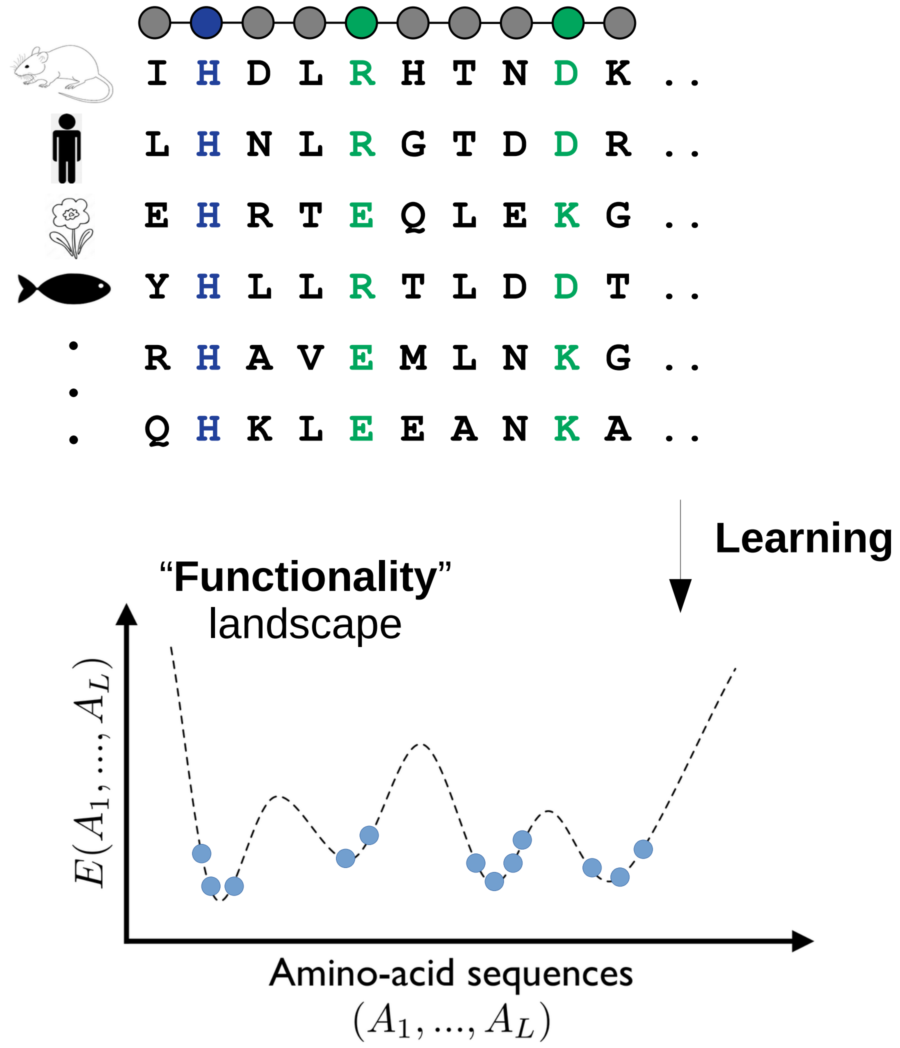
"Functionality" landscape



Bad sequences

Good sequences

Statistical modeling of protein sequences

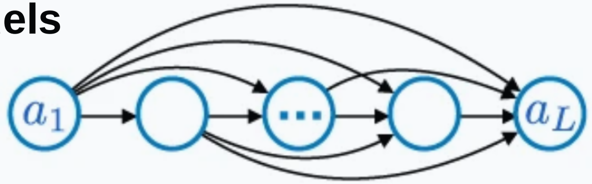


Potts model

Couplings Fields

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

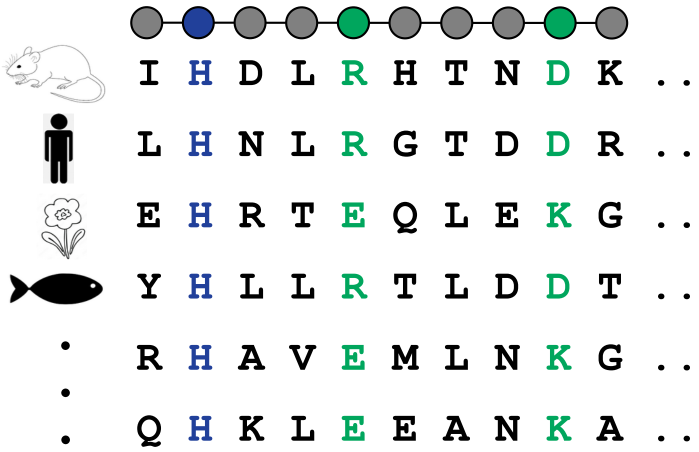
Autoregressive models



Deep learning

Variational autoencoders (VAE)
Transformers (MSATransformer, ESM, ...)

Statistical modeling of protein sequences

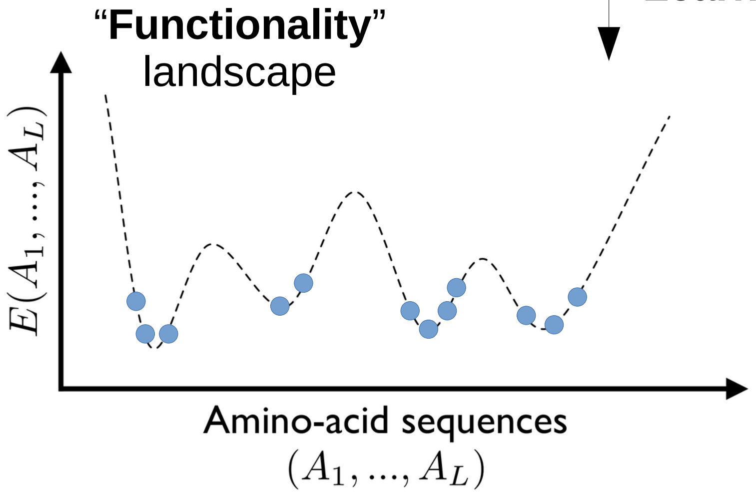


Used for

- Contacts in 3D structure
- Effect of mutations
- **Generative models** → **Design functional synthetic proteins!**

[Russ et. al. Science 2020]

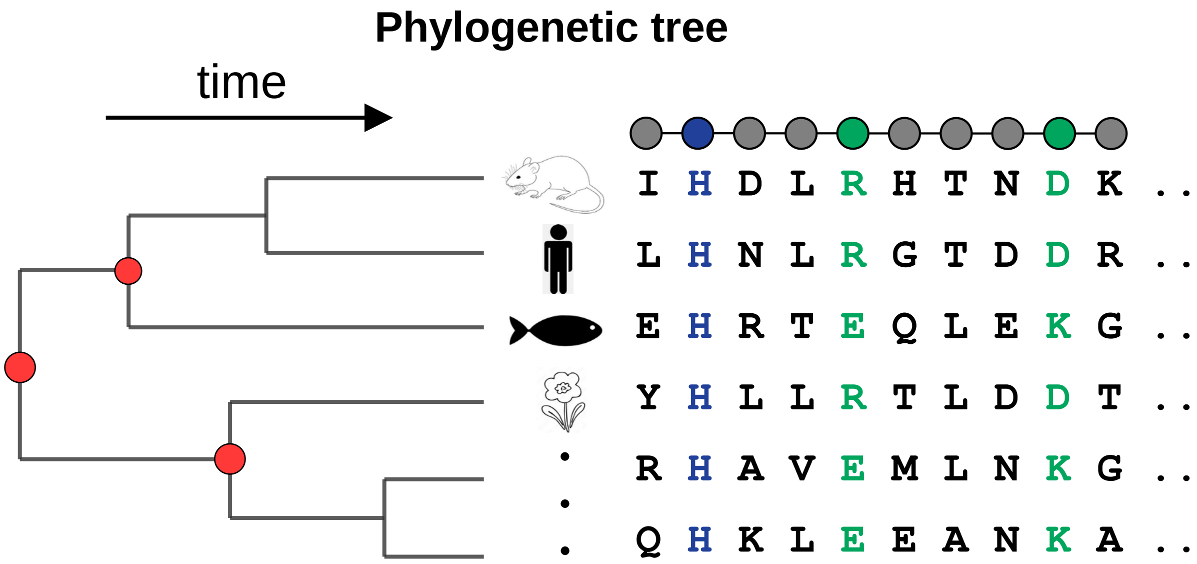
Learning



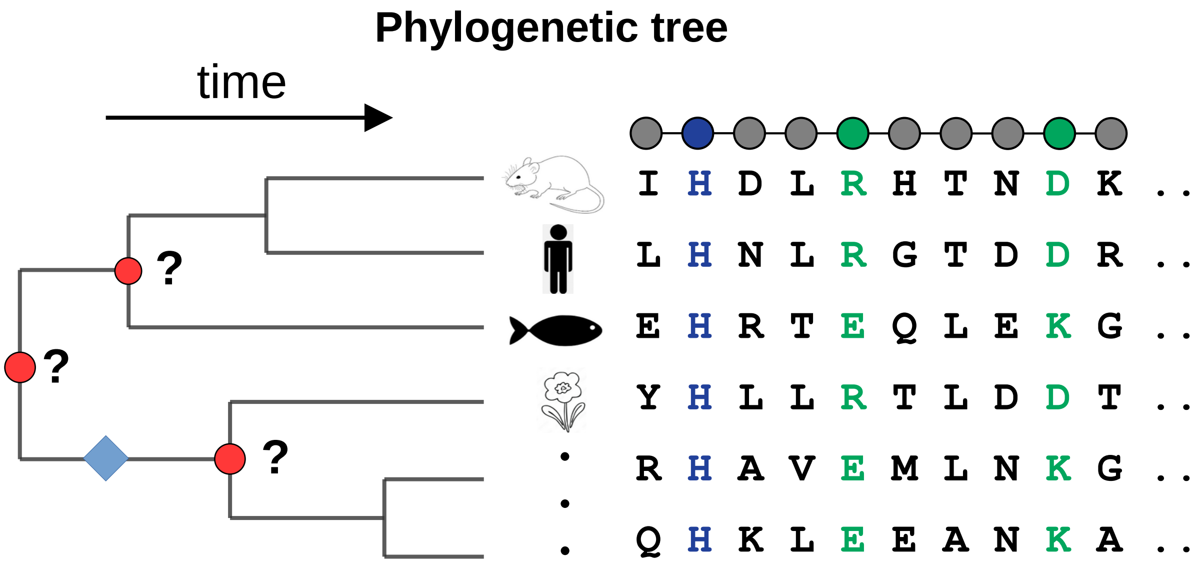
Key ingredient: **Epistasis**

→ Columns of the MSA are **not** independent

Ancestral sequence reconstruction



Ancestral sequence reconstruction



Why?

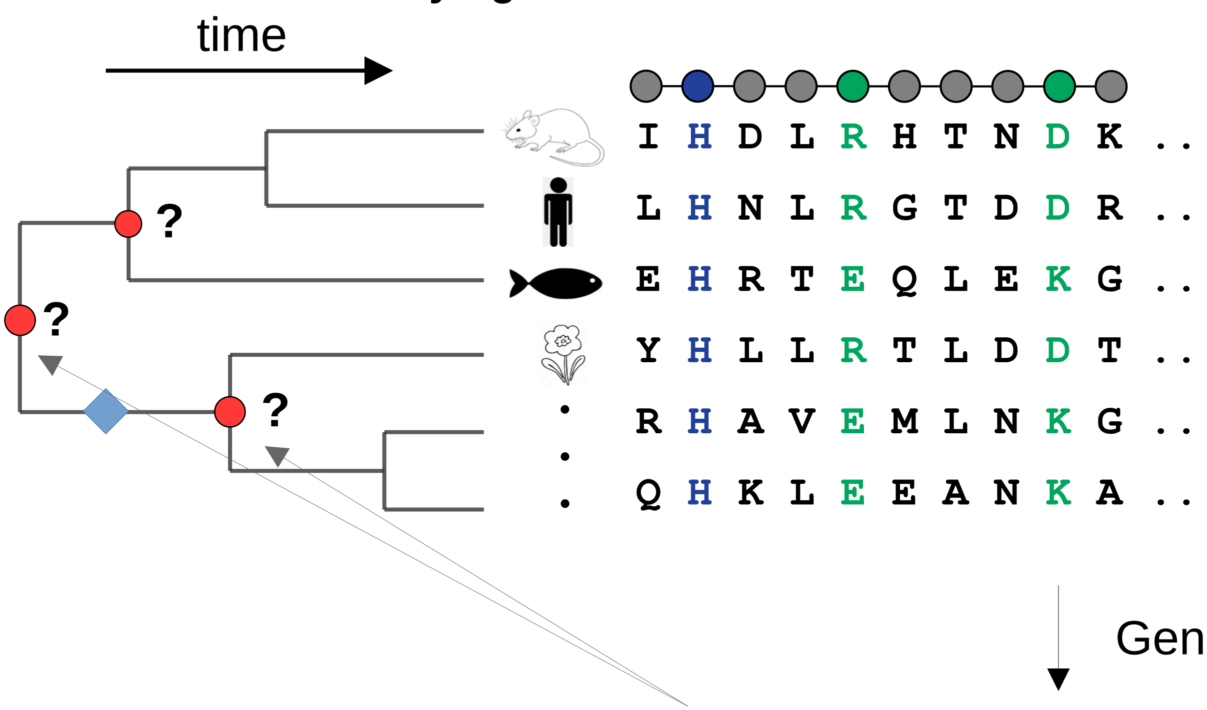
- What were ancient proteins like?
- Sequence – Function relationship

Applications to protein design

- Thermostable proteins
- Proteins with given specificity

Ancestral sequence reconstruction

Phylogenetic tree



Why?

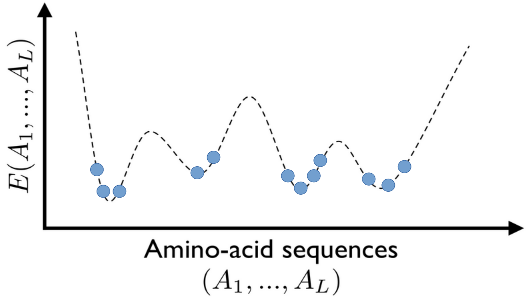
- What were ancient proteins like?
- Sequence – Function relationship

Applications to protein design

- Thermostable proteins
- Proteins with given specificity

Generative model

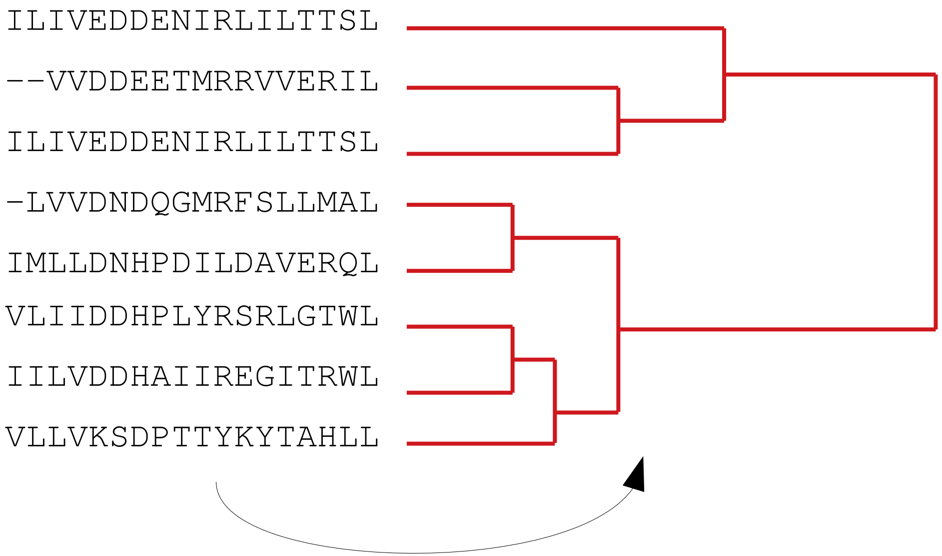
$$P(a_1, \ldots, a_N) = \frac{1}{Z} \exp\left(\sum_{i,j=1}^{L} J_{ij}(a_i, a_j) + \sum_{i=1}^{L} h_i(a_i) \right)$$



Ancestral sequence reconstruction

Phylogenetic tree

topology + branch length



Phylogenetic inference

- IQ-TREE – [Minh *et. al.*, MBE 2020]
- Fasttree – [Price *et. al.*, *PLOS* 2010]
- BEAST – [Suchard *et. al.*, *Virus Evol.* 2018]

Sequence evolution model

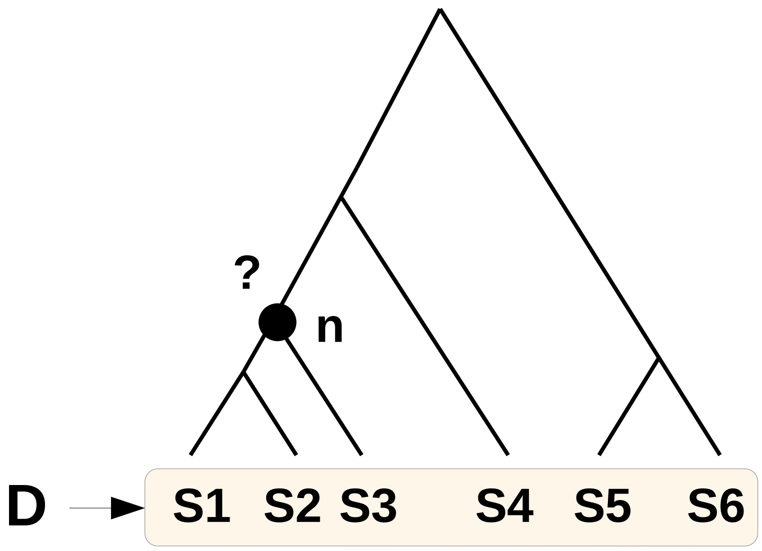
$\mathbf{x}^a$ ACTAGCTGGTC

Δt ↓ Evolution

$\mathbf{x}^c$ A**T**TAGCTG**C**TC

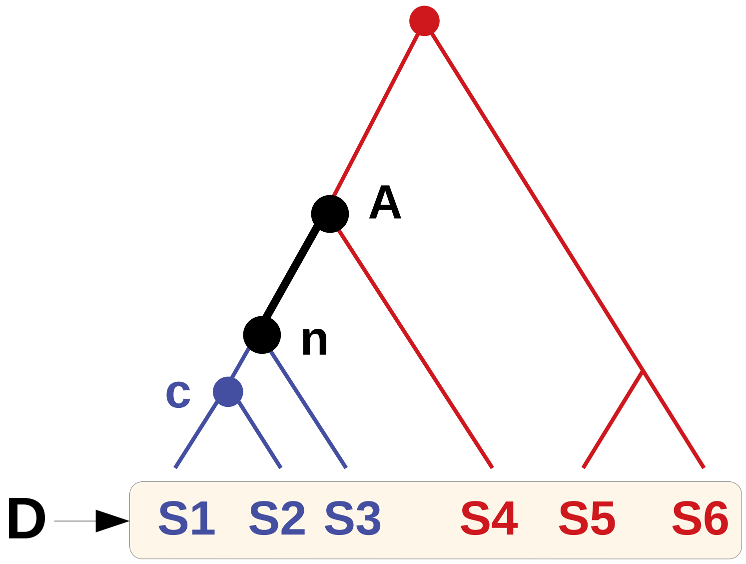
$$P(\mathbf{x}^c|\mathbf{x}^a, \Delta t)$$

Inference of internal states: Felsenstein's algorithm



$$\mathcal{L}_n(\mathbf{x}) = P(\mathcal{D}|n = \mathbf{x})$$

Inference algorithm



$$\mathcal{L}_n(\mathbf{x}) = P(\mathcal{D} | n = \mathbf{x})$$

"Down" likelihood

$\mathcal{L}_n^d(\mathbf{x})$ → Probability of data below ***n***, if ***n*** is in state ***x***

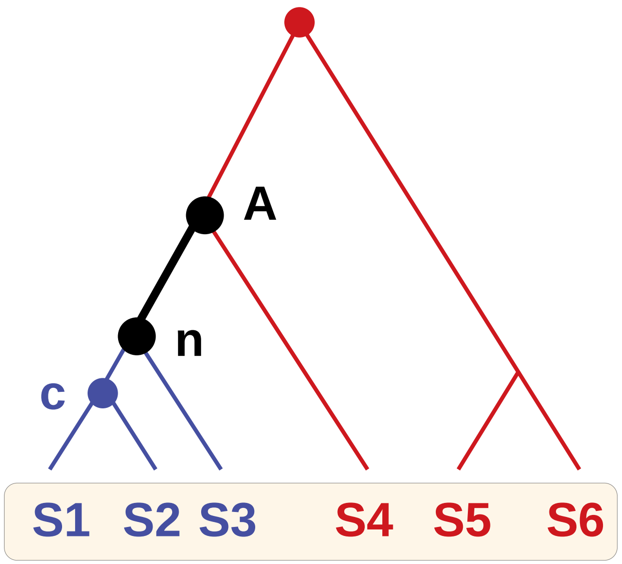
If ***n*** is leaf → $\mathcal{L}_{\text{leaf}}^d(\mathbf{x}) = \delta(\mathbf{x}, \mathbf{S})$

Otherwise → $\mathcal{L}_n^d(\mathbf{x}) = \prod_{c \in \mathcal{C}(n)} \sum_{\{\mathbf{z}\}} P(\mathbf{z}|\mathbf{x}, t_c)\mathcal{L}_c^d(\mathbf{z})$

Here $P(\mathbf{S}_3|\mathbf{x}, t_3) \cdot \sum_{\{\mathbf{z}\}} P(\mathbf{z}|\mathbf{x}, t_c)\mathcal{L}_c^d(\mathbf{z})$

→ One pass up from the leaves to compute $\mathcal{L}_n^d(\mathbf{x})$

Inference algorithm



$$\mathcal{L}_n(\mathbf{x}) = P(\mathcal{D} | n = \mathbf{x})$$

"Down" likelihood

$\mathcal{L}_n^d(\mathbf{x})$ → Probability of data below ***n***, if ***n*** is in state ***x***

$$\mathcal{L}_n^d(\mathbf{x}) = \prod_{c \in \mathcal{C}(n)} \sum_{\{\mathbf{z}\}} P(\mathbf{z}|\mathbf{x}, t_c) \mathcal{L}_c^d(\mathbf{z})$$

"Up" likelihood

$\mathcal{L}_n^u(\mathbf{y})$ → Probability of data above ***n***, if ***A*** is in state ***y*** where ***A ~ ancestor(n)***

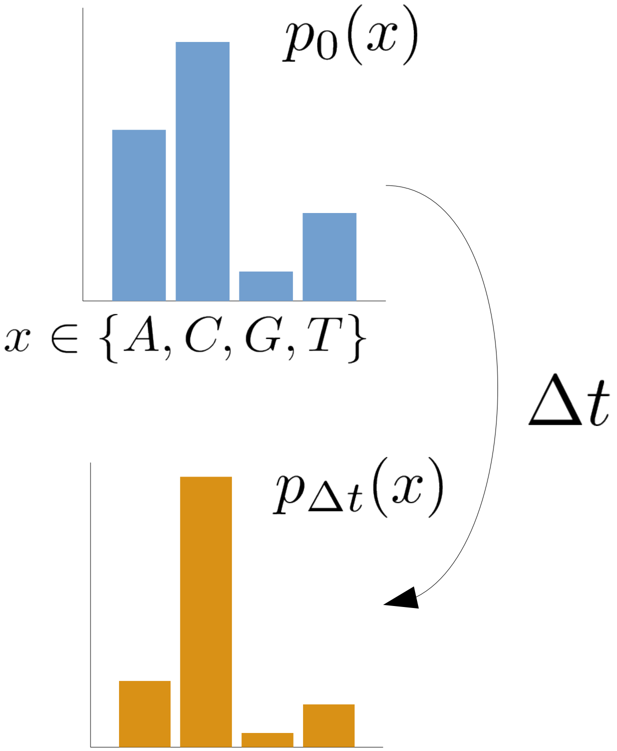
$$\mathcal{L}_n^u(\mathbf{y}) = \sum_{\{\mathbf{z}\}} P(\mathbf{y}|\mathbf{z}, t_A) \mathcal{L}_A^u(\mathbf{z}) \cdot \prod_{c \in \mathcal{C}(A)} \sum_{\{\mathbf{z}\}} P(\mathbf{z}|\mathbf{y}, t_c) \mathcal{L}_c^d(\mathbf{y})$$

$$\mathcal{L}_n(\mathbf{x}) = \sum_{\{\mathbf{y}\}} \mathcal{L}_n^u(\mathbf{y}) P(\mathbf{x}|\mathbf{y}, t_n) \mathcal{L}_n^d(\mathbf{x})$$

- **linear** in number of nodes
- depends only on **evolution model**

Sequence evolution model

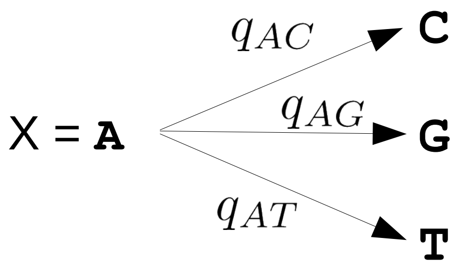
Focus on one position: $x \in \{A, C, G, T\}$



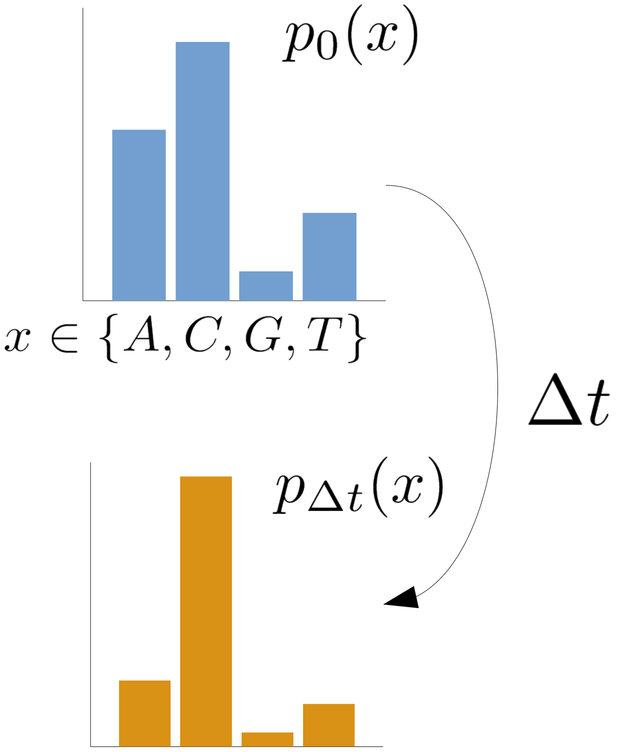
Sequence evolution model

Continuous time Markov chain

Focus on one position: $x \in \{A, C, G, T\}$



$$Q = \begin{pmatrix} -q_A & q_{CA} & q_{GA} & q_{TA} \\ q_{AC} & -q_C & q_{GC} & q_{TC} \\ q_{AG} & q_{CG} & -q_G & q_{TG} \\ q_{AT} & q_{CT} & q_{GT} & -q_T \end{pmatrix}$$



$$\dot{p} = p \cdot \mu Q$$

$$p_{\Delta t} = p_0 \cdot e^{\mu \Delta t \cdot Q}$$

$$q_x = \sum_{y \neq x} q_{xy}$$

Transition rate matrix Q

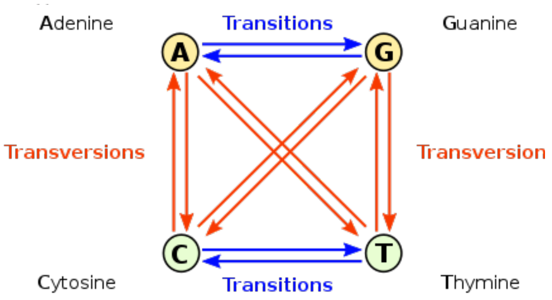
Reversibility: $\pi_x P(y|x,t) = \pi_y P(x|y,t)$

If reversibility $Q = \mathbf{H} \cdot \begin{pmatrix} \pi_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \pi_q \end{pmatrix}$ →

Q encodes
- Possibilities of mutations
- Equilibrium distribution

H → **Symmetric**

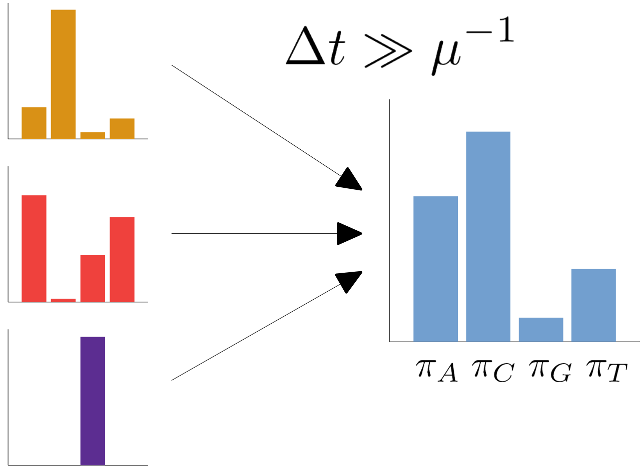
Nucleotides



Amino acids → **H** ~ Genetic code

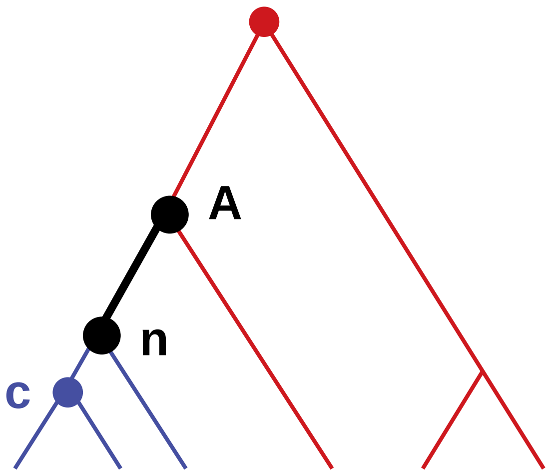
$(\pi_1 \ldots \pi_q)$ → **Equilibrium state**

$\Delta t \gg \mu^{-1}$



Independent-site reconstruction

$$p_{\Delta t} = p_0 \cdot e^{\mu \Delta t \cdot Q}$$



One Q_i per sequence position *i*

Reconstruct for one position *i*

"Down" likelihood

$$\mathcal{L}^d_{x,i}(x_i) = \prod_{c \in \mathcal{C}(n)} \sum_{z_i=1}^{q} \left(e^{Q_i t_c}\right)_{x_i z_i} \mathcal{L}^d_c(z_i)$$

$$\mathcal{L}_n(x_i) = P(\mathcal{D} | n_i = x_i)$$

How do we find the right Q matrices?

Independent-site reconstruction

[Jones et al., 1992]
[Le and Gascuel, 2008]

$$p_{\Delta t} = p_0 \cdot e^{\mu \Delta t \cdot Q}$$

$$Q = \mathbf{H} \cdot \begin{pmatrix} \pi_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \pi_q \end{pmatrix}$$

State of the art

Q
- Same for all positions
- Fixed matrix, pre-learned (JTT, LG, ...)

μ
Can change across positions. Typically
- Four rates $\{\mu_1, \ldots, \mu_4\}$
- Proportion of invariable sites $\mu = 0$

Drawbacks
- Consider positions independently
- **Ignores** functional constraints

Independent-site reconstruction

[Jones et al., 1992]
[Le and Gascuel, 2008]

$$p_{\Delta t} = p_0 \cdot e^{\mu \Delta t \cdot Q}$$

$$Q = \mathbf{H} \cdot \begin{pmatrix} \pi_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \pi_q \end{pmatrix}$$

State of the art

Q
- Same for all positions
- Fixed matrix, pre-learned (JTT, LG, ...)

μ Can change across positions. Typically
- Four rates $\{\mu_1, \dots, \mu_4\}$
- Proportion of invariable sites $\mu = 0$

Drawbacks
- Consider positions independently
- **Ignores** functional constraints

Generative sequence model → Propagator?

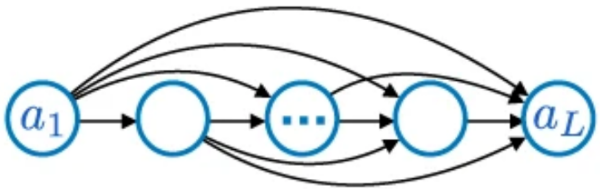
$$P(a_1, \dots, a_N) = \frac{1}{Z} \exp\left(\sum_{i,j=1}^{L} J_{ij}(a_i, a_j) + \sum_{i=1}^{L} h_i(a_i) \right)$$

P **is not factorized**

Evolution with autoregressive models

Autoregressive model

$$P(a_1 \dots a_L) = \prod_{i=1}^{L} p_i(a_i | a_1 \dots a_{i-1})$$



→ Need to know $p(a_i | a< i)$

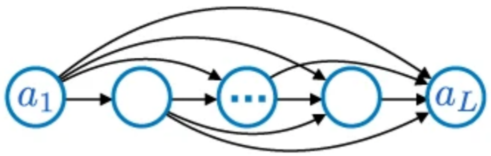
ArDCA → $p(a_i | a< i) \propto \exp\left(\sum_{j<i} J_{ij}(a_i, a_j) + h_i(a_i)\right)$

- Easy to infer (from alignment)
- Interpretable (Jij ~ contacts)
- Good generative properties

[Trinquier *et. al., Nature Comm* 2021]

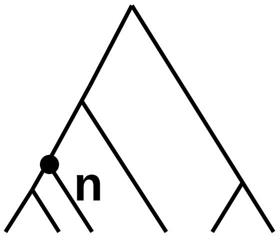
Evolution with autoregressive models

Autoregressive model



$$P(\mathbf{a}) = \prod_{i=1}^{L} p_i(a_i | a_{<i})$$

Given the context $a_{<i} = a_1 \dots a_{i-1}$
we know the equilibrium frequencies for a_i



Suppose we reconstructed *i-1* positions

$\rightarrow a^n_{<i} = a^n_1 \dots a^n_{i-1}$

$\rightarrow$ Equilibrium frequencies $p_i(a_i | a^n_{<i})$

Evolution model

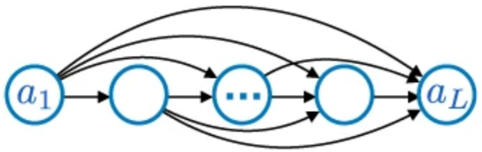
- One position at a time: **almost factorized**
- Use knowledge of **functional constraints**
- **Cost**: need data to infer model

Evolution *towards* n for position i

$$Q_i^{\rightarrow n} = \mathbf{H} \cdot \begin{pmatrix} p_i(1|a^n_{<i}) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & p_i(q|a^n_{<i}) \end{pmatrix}$$

Testing on simulations

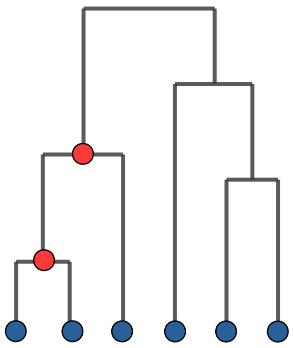
Reference generative model



PF00072 : Response regulator receiver domain

+

Random simulated tree



T: tree height

M: # leaves

Simulated evolution

Sequences of leaves

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

Reconstruction

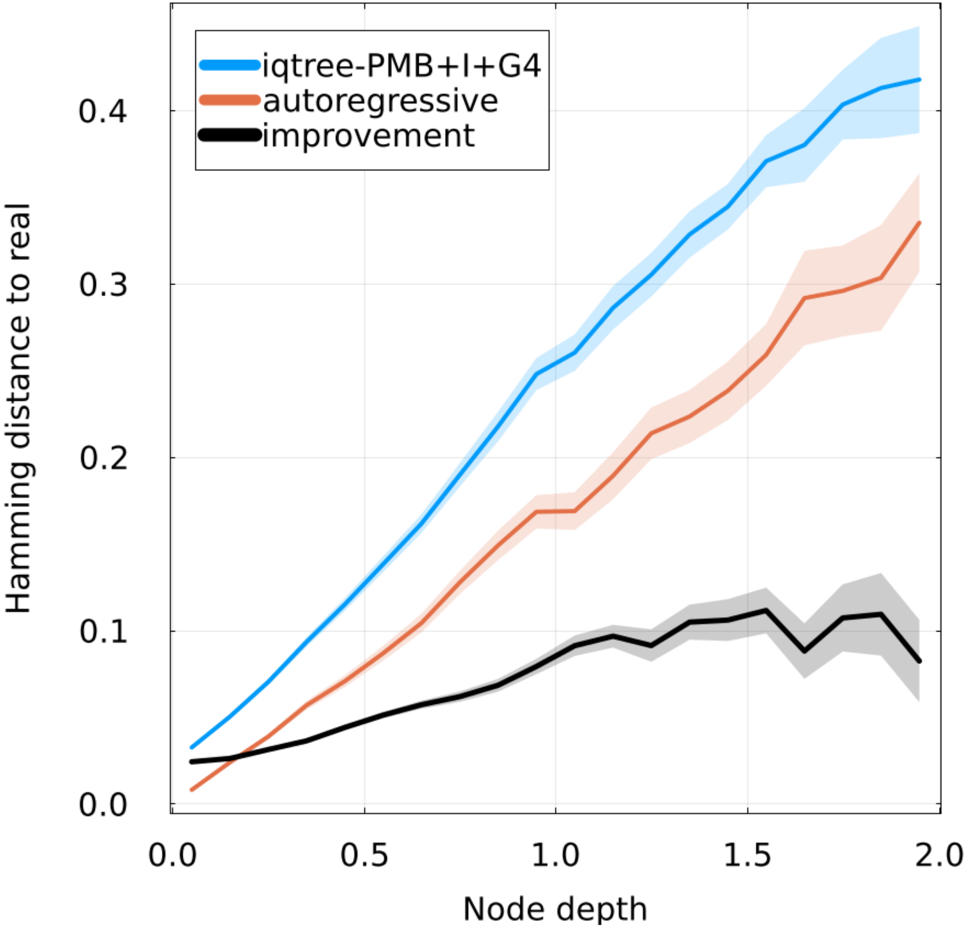
- **Autoregressive**
- **IQ-TREE**: representative state of the art method

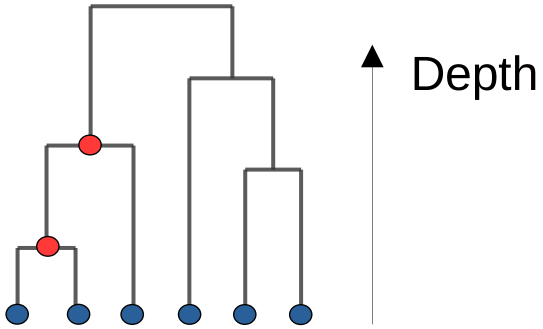
[Minh *et. al.*, MBE 2020]

Compare to real internal sequences!

Results: Maximum Likelihood reconstruction

$$\mathcal{L}_n(\mathbf{x}) = P(\mathcal{D}|n = \mathbf{x})$$



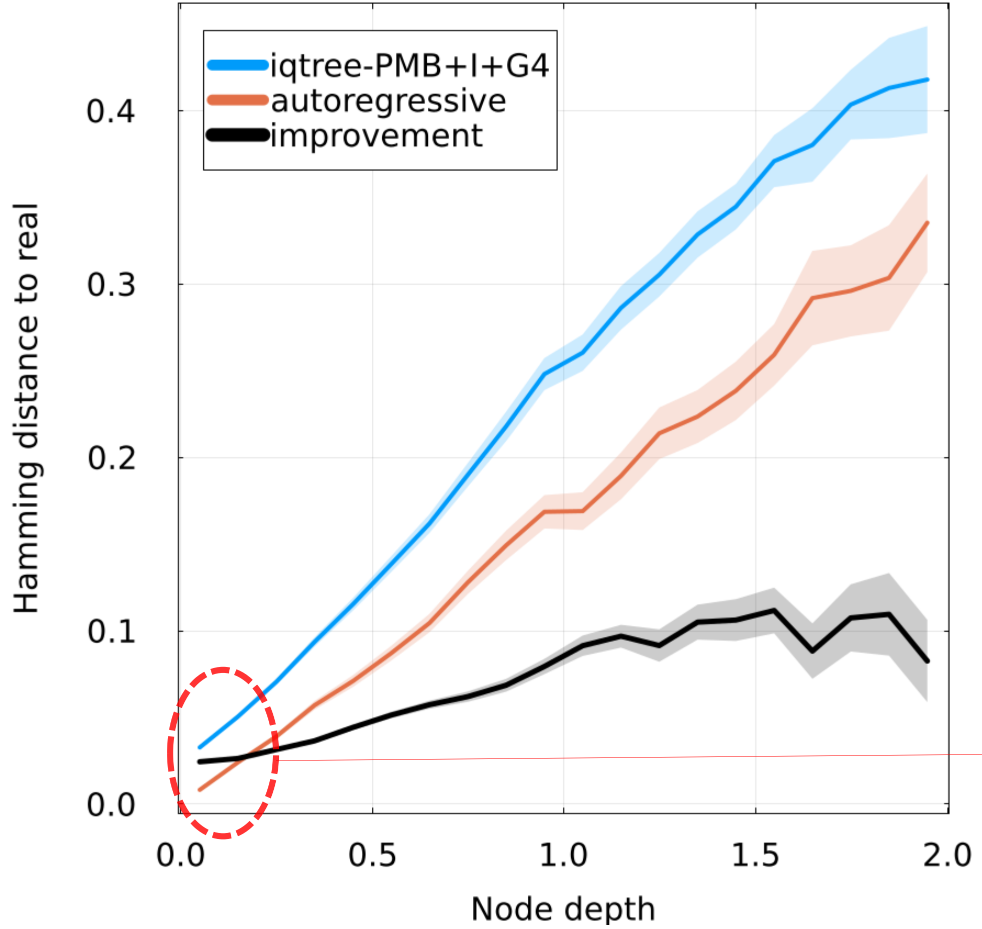


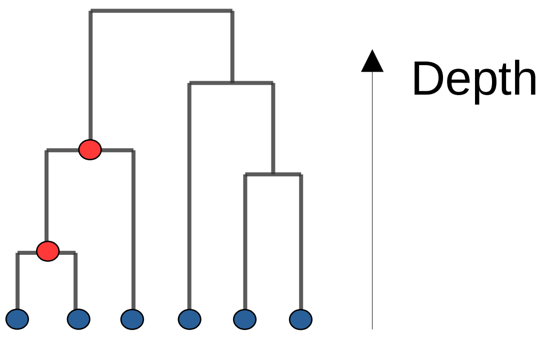
ML reconstruction $\mathbf{x} = \text{argmax}\, \mathcal{L}_n$

Most commonly used in litterature

Results: Maximum Likelihood reconstruction

$$\mathcal{L}_n(\mathbf{x}) = P(\mathcal{D}|n = \mathbf{x})$$





ML reconstruction $\mathbf{x} = \operatorname{argmax} \mathcal{L}_n$

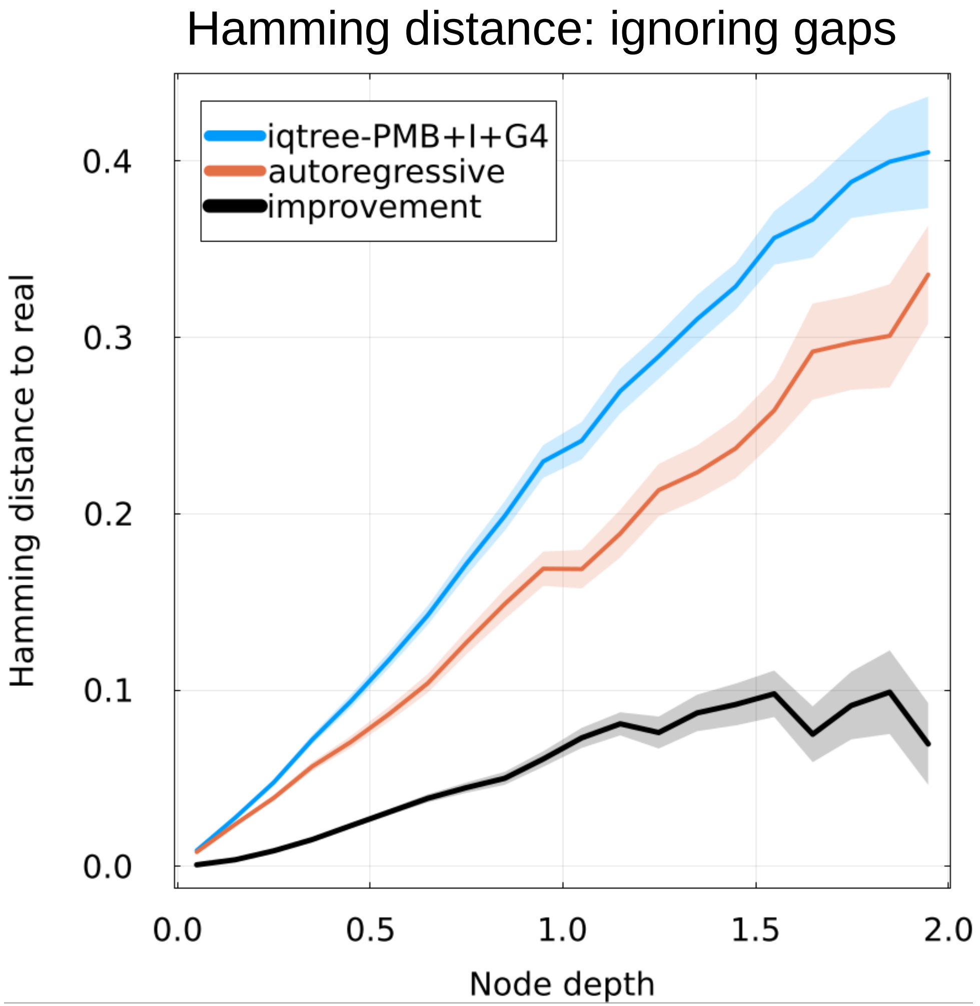
Most commonly used in litterature

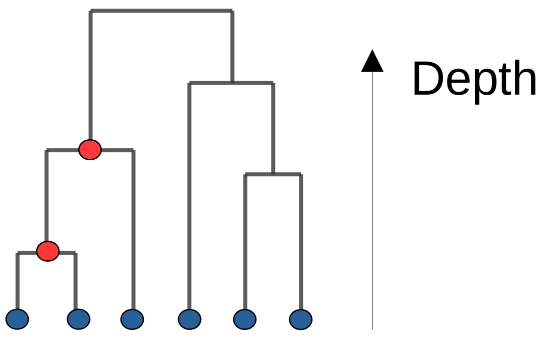
Gaps!

- Represented in ArDCA...
- … but not in IQ-TREE

Results: Maximum Likelihood reconstruction

$$\mathcal{L}_n(\mathbf{x}) = P(\mathcal{D}|n = \mathbf{x})$$

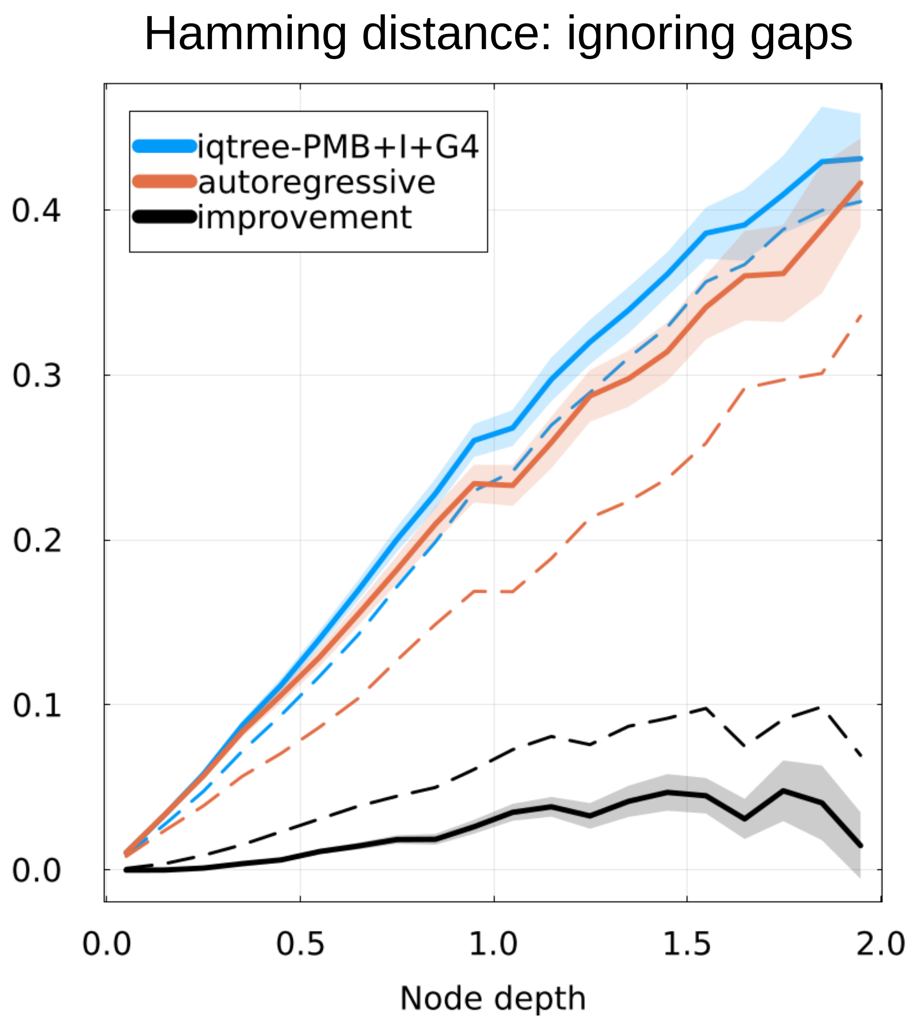




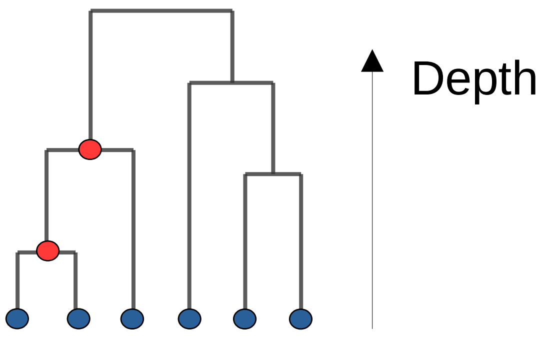
ML reconstruction $\mathbf{x} = \text{argmax}\,\mathcal{L}_n$

Most commonly used in litterature

Results: ML + Bayesian reconstruction



$$\mathcal{L}_n(\mathbf{x}) = P(\mathcal{D}|n = \mathbf{x})$$



ML reconstruction $\mathbf{x} = \operatorname{argmax} \mathcal{L}_n$

Most commonly used in literature

Bayesian $P(\mathbf{x}) \propto \mathcal{L}_n(\mathbf{x})$

Rarely used in practice

Testing on simulations: with a different evolver?

Reference generative model

Simulated evolution

Sequences of leaves

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

Another model that has epistasis

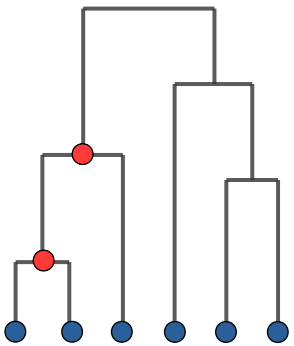
Potts model

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

+

Random simulated tree

Reconstruction



T: tree height

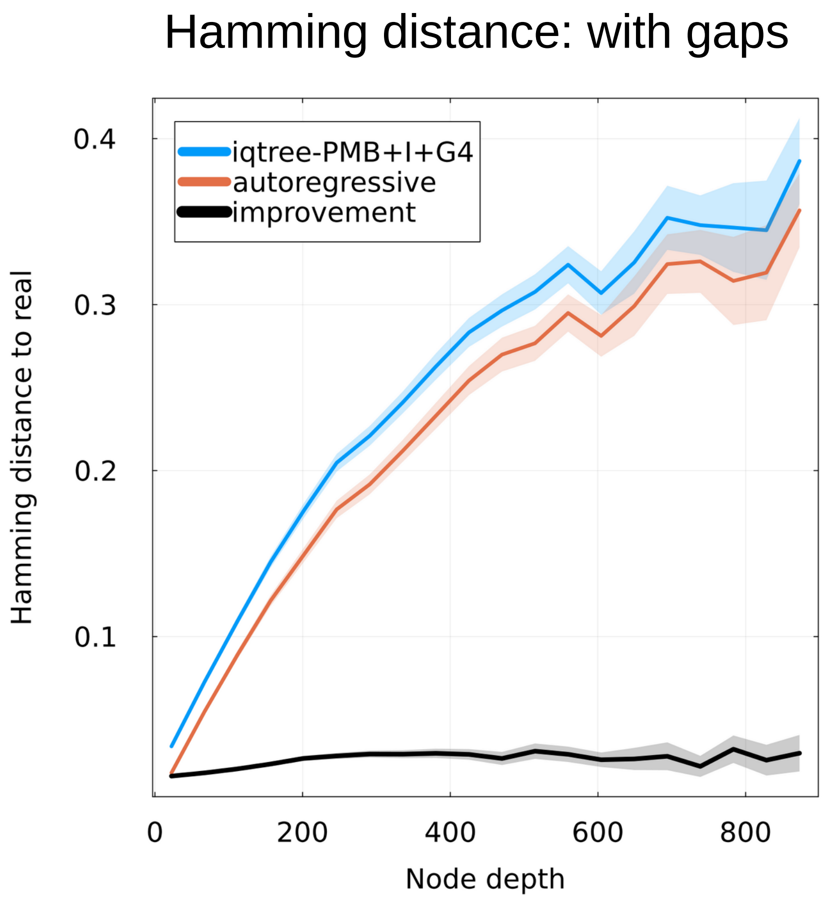
M: # leaves

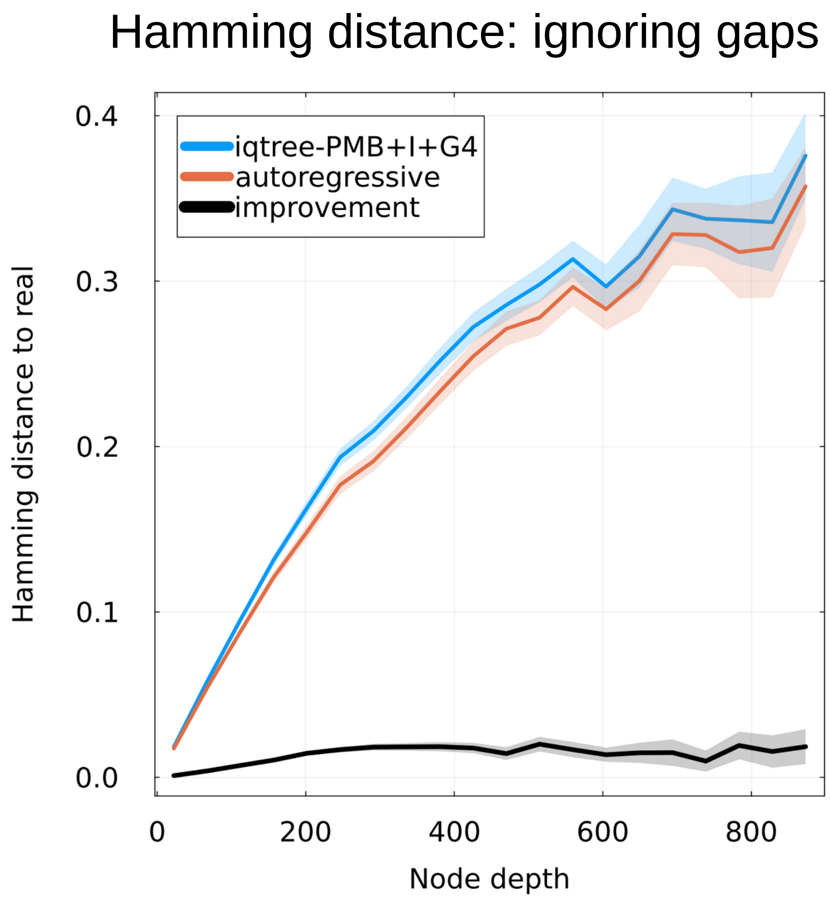
- **Autoregressive**
- **IQ-TREE**: representative state of the art method

[Minh *et. al.*, MBE 2020]

Compare to real internal sequences!

Testing on simulations: with a different evolver?



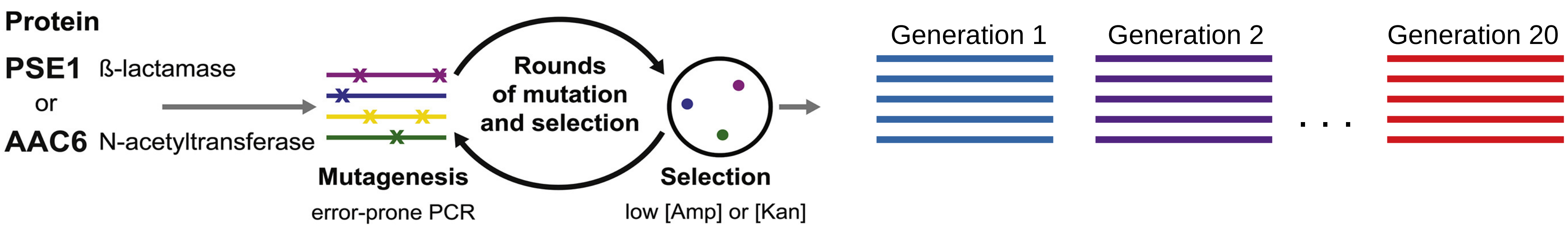


→ **Systematic improvement, but small**

Experimental data

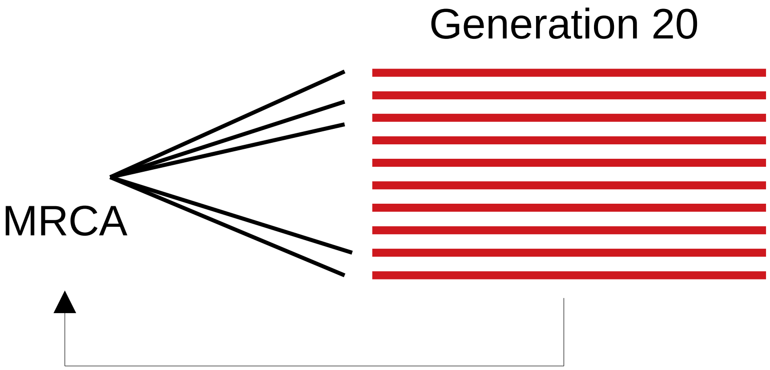
[Stiffler et. al. Cell 2020]

The sequence of MRCA is known!



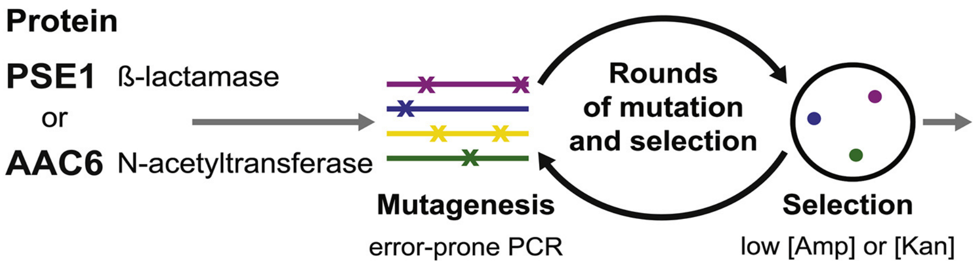
but not the tree...

→ Simplification: star tree

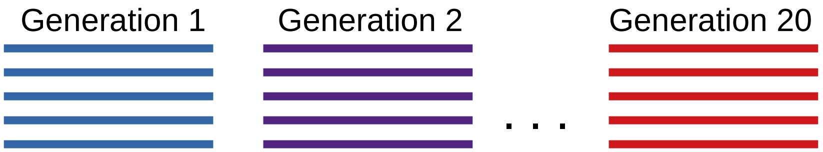


Experimental data

[Stiffler et. al. Cell 2020]

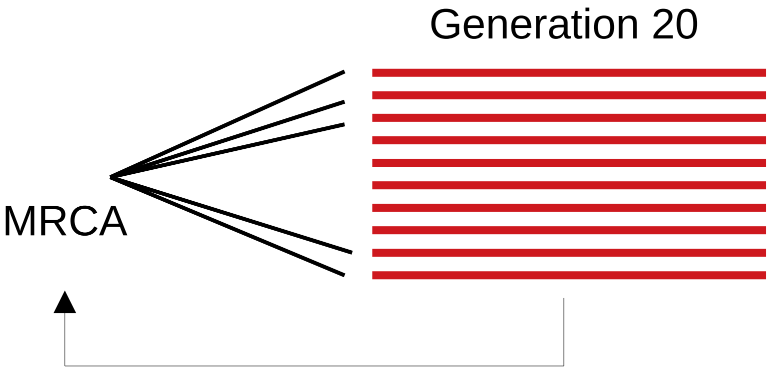


The sequence of MRCA is known!

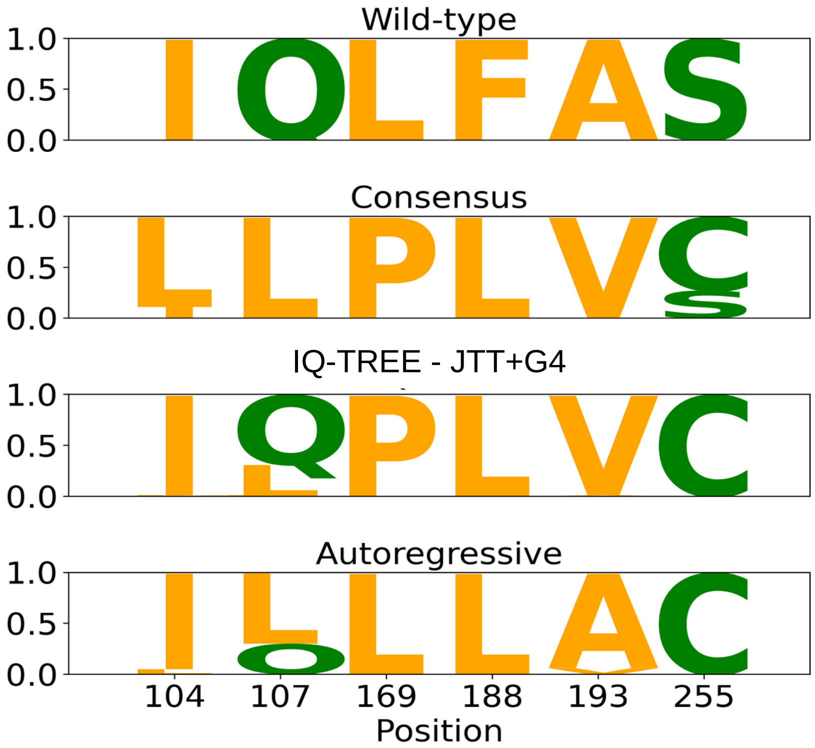


but not the tree...

→ Simplification: star tree



6 “interesting” positions



Biases of the ML reconstruction

ML reconstruction $\mathbf{x} = \mathrm{argmax}\, \mathcal{L}_n$

- Most commonly used in literature
- Experimentally: functional & highly thermostable
- **Problem**: Is the best sequence representative?

→ **Biases**

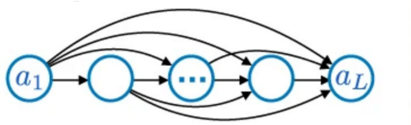
[Williams et. al., PLOS CB 2006]

Bayesian $P(\mathbf{x}) \propto \mathcal{L}_n(\mathbf{x})$

- Set of sequences at each internal node
- Rarely used in practice
- Sometimes non-functional
- More representative / Less subject to biases?

Biases of the ML reconstruction

Reference generative model



$$P(a_1 \dots a_L) = \prod_i P(a_i | a_1 \dots a_{i-1})$$

→ **Probability** of reconstructed sequences?

~proxy for function

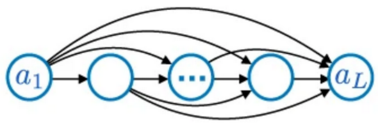
Biases of the ML reconstruction

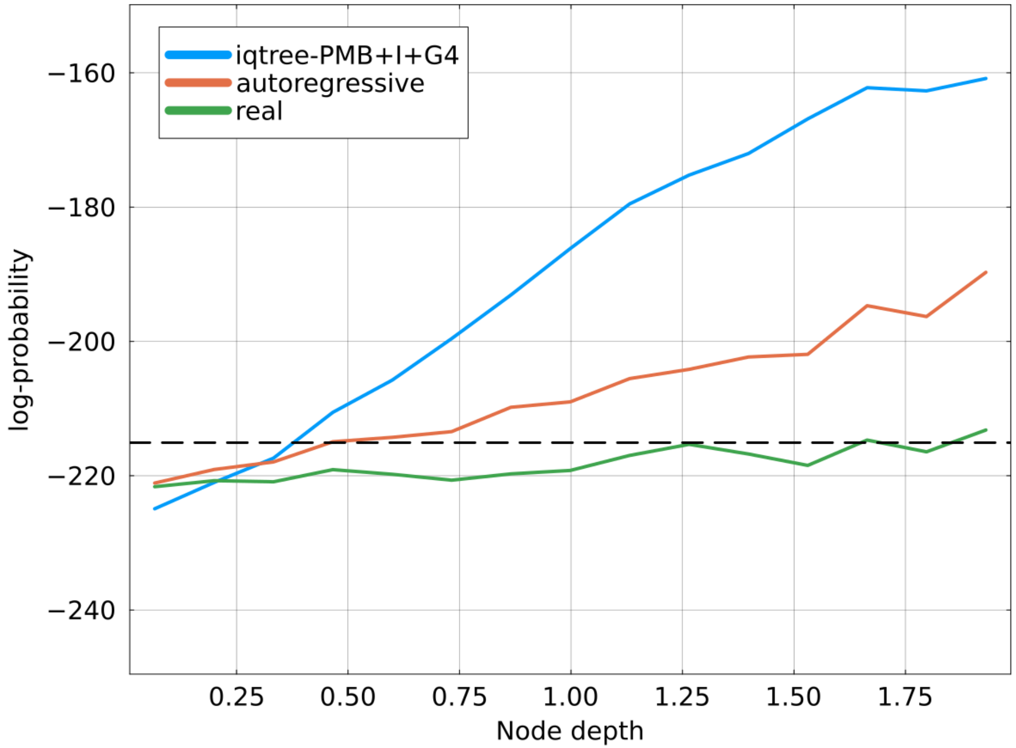
Reference generative model

$$P(a_1 \dots a_L) = \prod_i P(a_i | a_1 \dots a_{i-1})$$

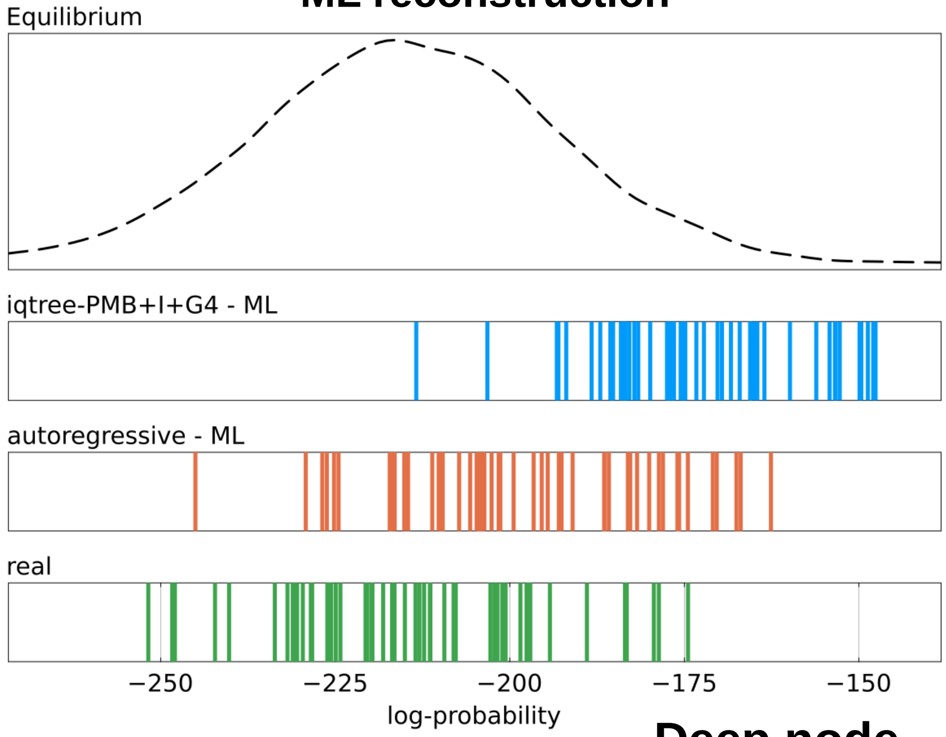
→ **Probability** of reconstructed sequences?

~proxy for function





ML reconstruction



Deep node

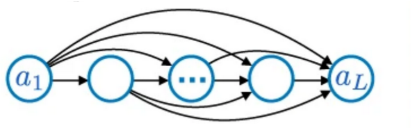
Is Bayesian less biased?

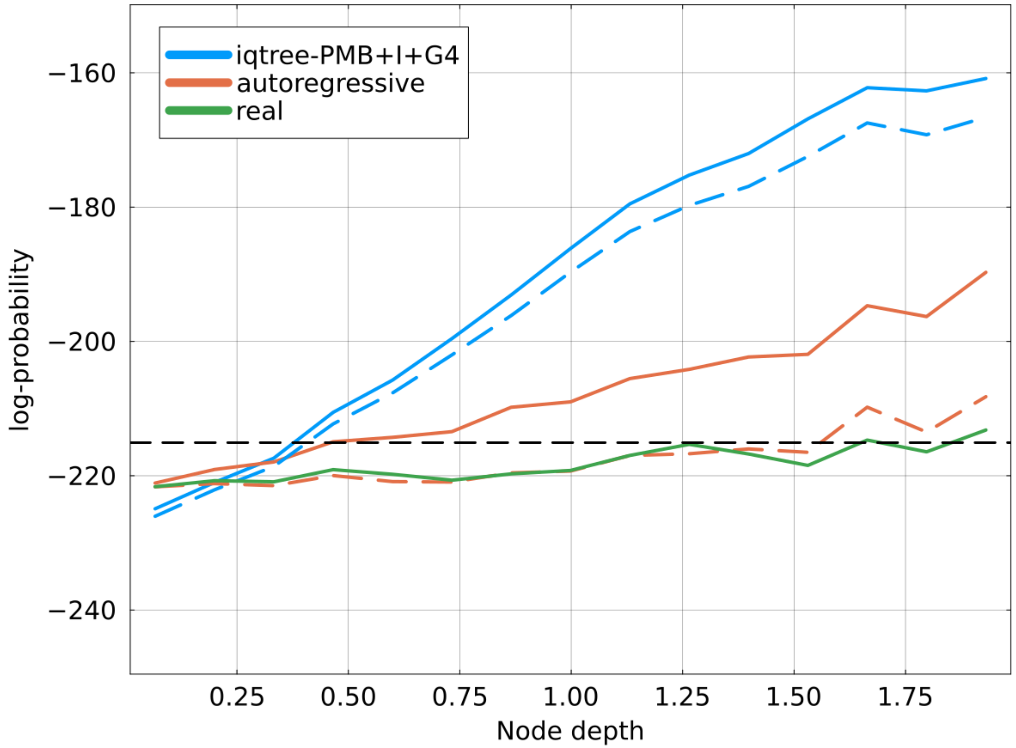
Reference generative model

$$P(a_1 \ldots a_L) = \prod_i P(a_i | a_1 \ldots a_{i-1})$$

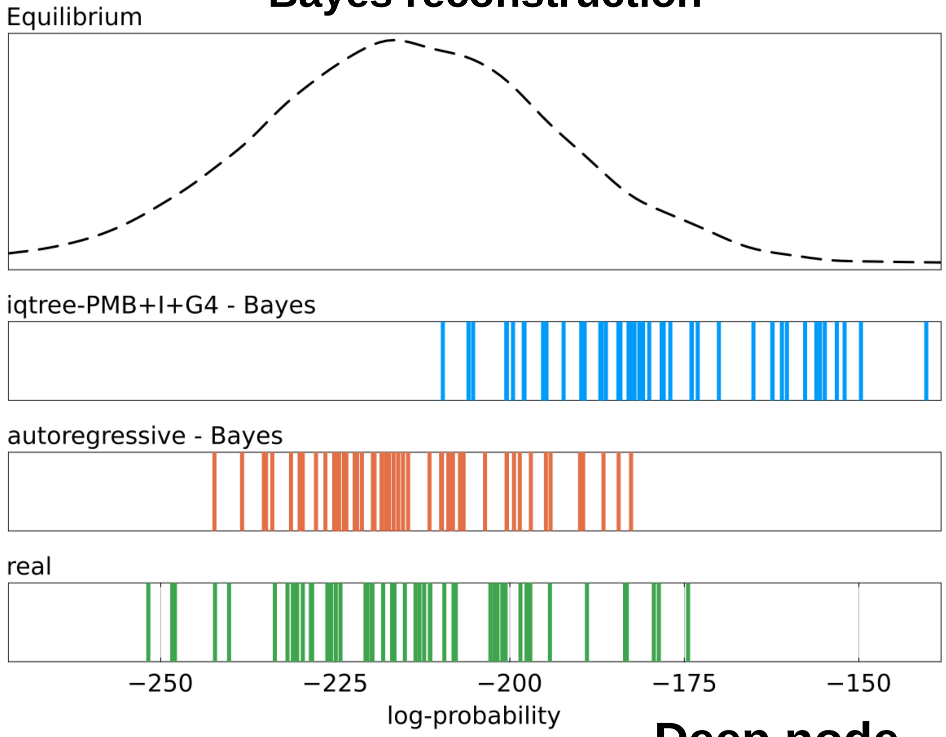
→ **Probability** of reconstructed sequences?

~proxy for function





Bayes reconstruction



Deep node

Biases of the ML reconstruction

ML reconstruction $\mathbf{x} = \mathrm{argmax}\, \mathcal{L}_n$

- Most commonly used in literature
- Experimentally: functional & highly thermostable
- **Problem**: Is the best sequence representative?

→ **Biases**

[Williams et. al., PLOS CB 2006]

Bayesian $P(\mathbf{x}) \propto \mathcal{L}_n(\mathbf{x})$

- Set of sequences at each internal node
- Rarely used in practice
- Sometimes non-functional
- More representative / Less subject to biases?

For ML or site-independent model

→ **Bias in log-probability**

Biases of the ML reconstruction

ML reconstruction $\mathbf{x} = \text{argmax}\, \mathcal{L}_n$

- Most commonly used in literature
- Experimentally: functional & highly thermostable
- **Problem**: Is the best sequence representative?

→ **Biases**

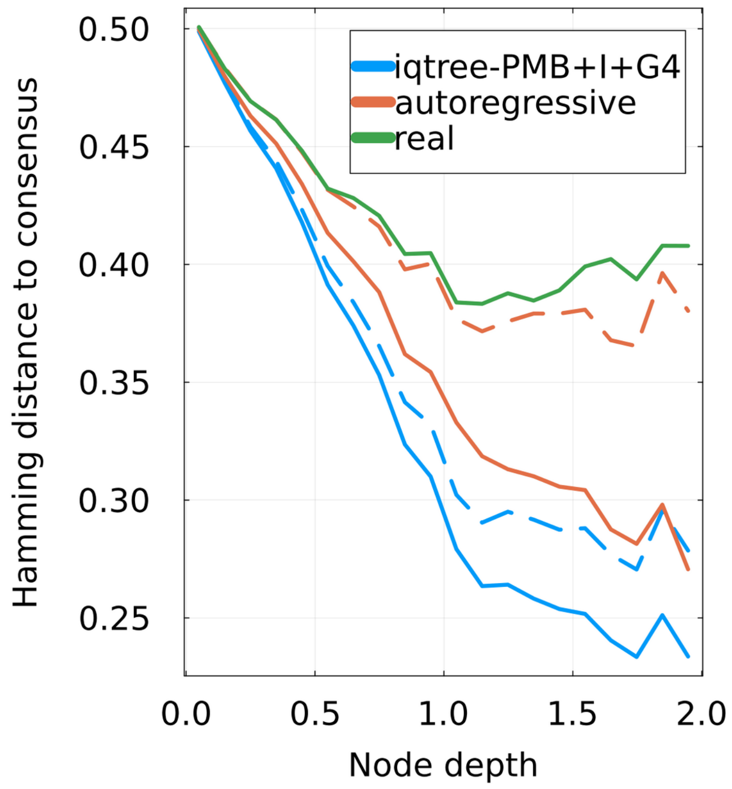
[Williams et. al., PLOS CB 2006]

Bayesian $P(\mathbf{x}) \propto \mathcal{L}_n(\mathbf{x})$

- Set of sequences at each internal node
- Rarely used in practice
- Sometimes non-functional
- More representative / Less subject to biases?

For ML or site-independent model

→ **Bias in log-probability**

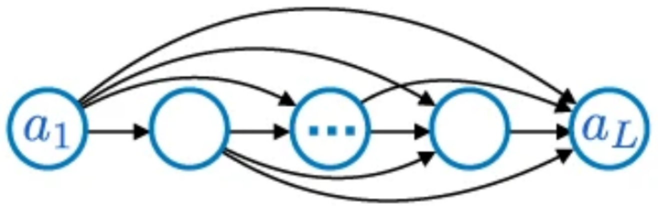


→ **Bias towards consensus!**

Autoregressive dynamics are not Markovian

Autoregressive model

$$P(a_1 \dots a_L) = \prod_i P(a_i | a_1 \dots a_{i-1})$$



Autoregressive dynamics

$$P_i(a|b, a_{<i}, t) = e^{-t}\delta_{ab} + (1 - e^{-t})p_i(a|a_{<i})$$

(if **H** uniform)

No mut.   >1 mut.

$$P(\mathbf{a}|\mathbf{b}, t) = \prod_{i=1}^{L} P_i(a|b, a_{<i}, t) \xrightarrow[t \to \infty]{} P(\mathbf{a})$$

Autoregressive dynamics are not Markovian

Toy model

Binary sequences, L=2

Global balance: $$P(\mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{y}) P(\mathbf{x}|\mathbf{y})$$

x = (- +) → P(x) << 1

y = (+ +) → P(y) ~ 1/2

Frequent
+ +
- -
Rare
- +
+ -

$$p_1(-) = 1/2$$

$$p_2(+|-) \ll 1$$

Autoregressive dynamics are not Markovian

Toy model

Binary sequences, L=2

Frequent
+ +
- -
Rare
- +
+ -

$$p_1(-) = 1/2$$

$$p_2(+|-) \ll 1$$

Global balance: $P(\mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{y}) P(\mathbf{x}|\mathbf{y})$

x = (- +) → P(x) << 1

y = (+ +) → P(y) ~ 1/2

$$P(\mathbf{x}|\mathbf{y},t) = (1-e^{-t})p_1(-) \cdot \{e^{-t} + (1-e^{-t})p_2(+|-)\} \rightarrow \mathbf{O(1)}$$

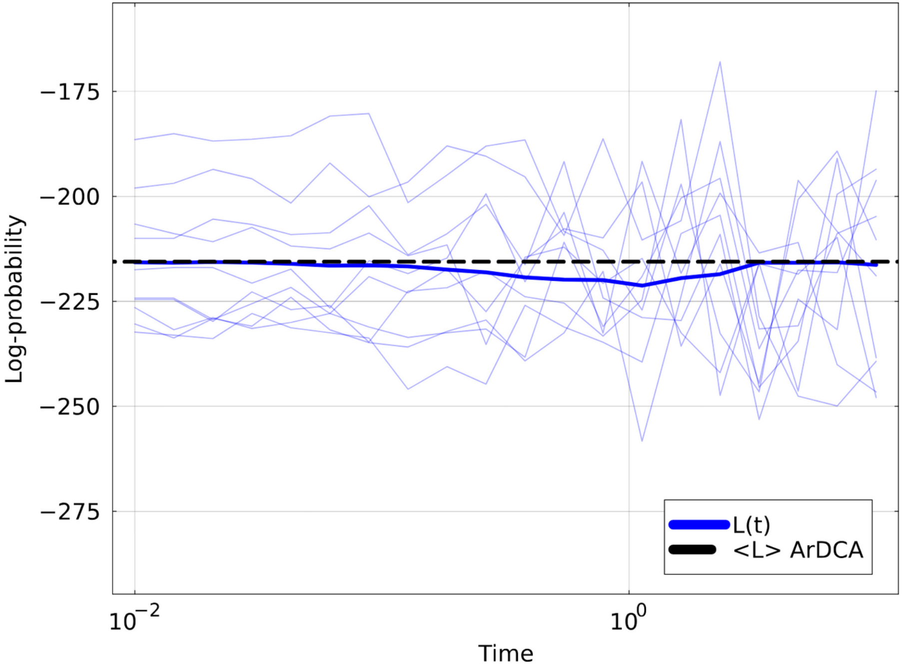
Mut. at pos 1 — No mut at 2 — $\ll 1$

→ **Global balance cannot hold!**

Autoregressive dynamics are not Markovian

Consequences

- Irreversibility
- Dynamics go out of equilibrium
- Very likely not realistic



But...

- Correct dynamics for t<<1 and t>>1
- Quantitatively small effect

Conclusion

Evolution based on a generative model

- Autoregressive architecture: “almost factorized”
- Converges to generative distribution at long times
- Caveats: not Markov, irreversible

Ancestral Sequence Reconstruction

- Better handling of gaps
- Systematic improvement over state of the art models (simulations)
- Improvement on directed evolution data

Biases of maximum likelihood

- ML reconstruction is “too good” → Bias to consensus!
- Bayesian reconstruction with good model: more representative!

Conclusion

Authors
Matteo De Leonardis (PoliTo)
Andrea Pagnani (PoliTo)
P.B.C

Evolution based on a generative model

- Autoregressive architecture: “almost factorized”
- Converges to generative distribution at long times
- Caveats: not Markov, irreversible

Ancestral Sequence Reconstruction

- Better handling of gaps
- Systematic improvement over state of the art models (simulations)
- Improvement on directed evolution data

Biases of maximum likelihood

- ML reconstruction is “too good” → Bias to consensus!
- Bayesian reconstruction with good model: more representative!

Thank you for listening

Politecnico di Torino