



The Abdus Salam
**International Centre
for Theoretical Physics**
50th Anniversary 1964–2014



2584–5

Spring College on the Physics of Complex Systems

26 May – 20 June, 2014

Including Phylogeny

Martin Weigt
Université Pierre et Marie Curie
Paris
France

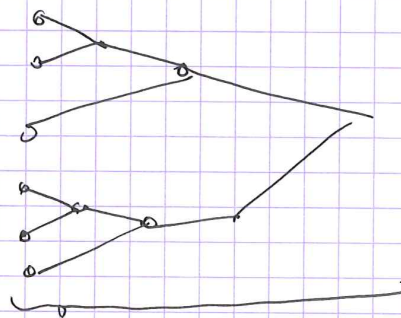
INCLUDING PHYLOGENY

(2.1)

- consider regulation of orthologous gene
 - search motifs for evolutionary conserved regulatory sites
 - independence assumption of seqs. not OK, include phylogeny of species (assume to be known)

Data:

$$D = \begin{pmatrix} a_1^1 & \dots & a_L^1 \\ \vdots & & \vdots \\ a_1^n & \dots & a_L^n \end{pmatrix}$$



species tree T

→ topology

→ branch length (time between speciation events)

How to bring these two together?

→ data generated by evolution on phylogenetic tree!

Assume: independent evolution of sites $i=1, \dots, L$
(final modeling: PWM)

→ may concentrate on one site at a time

→ skip index i for simplicity of notation

one site $\rightarrow$ PWM $a \in \{A, C, G, T\}$

$\hookrightarrow$ 3 independent parameters

$$(w(A), w(C), w(G))$$

$$\rightarrow w(T) = 1 - w(A) - w(C) - w(G)$$

Need to characterize evolution of this site:

crucial quantity:

$$P(a | b, t) = \text{transition probability} \\ = \text{prob (find nt. } a \text{ at time } t \\ \text{given nt } b \text{ at time } 0)$$

Some properties of $P(a | b, t)$

$$P(a | b, t=0) = \delta_{a,b} = \begin{cases} 1 & \text{if } a=b \\ 0 & \text{if } a \neq b \end{cases}$$

$\rightarrow$ no change in zero time

$$\lim_{t \rightarrow \infty} P(a | b, t) = w(a) \quad \text{PWM}$$

$\rightarrow$ long-term evolution samples according to PWM

$\rightarrow$ PWM = stationary distribution of evolution.

$$\begin{aligned} &\text{Prob of no mutation in time } t \\ &= e^{-\mu t} \end{aligned}$$

with mutation rate μ

$$P(a | b, t) \rightarrow 12 \text{ parameters}$$

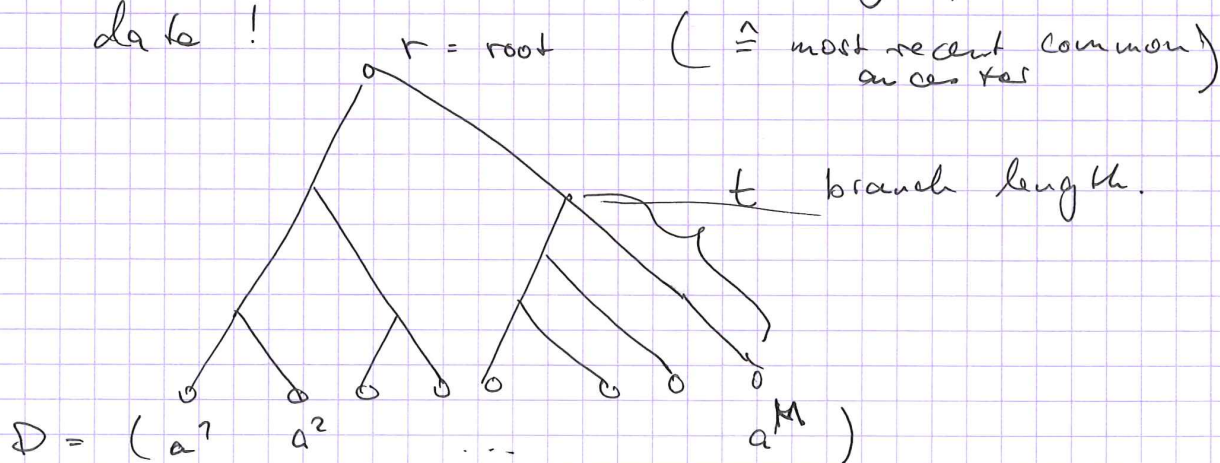
(4×4 nt combinations - 4 normalization constraints)

A simple "time-reversible" model:

$$P(a|b, t) = e^{-\mu t} \delta_{a,b} + (1 - e^{-\mu t}) w(a)$$

- prob $e^{-\mu t}$ of no substitution
- prob $(1 - e^{-\mu t})$ of (at least one) substitution, fixed ~~according to~~ according to stationary distribution.
- 4 parameters $\mu, w(a)$
(stationary dist. + rate of convergence)

⇒ Can calculate the probability of our data!



$$\Rightarrow P(D \mid w, \mu, T)$$

Problem: know only states of leaves, but not of internal nodes (= ancestral species ...)

(2.4)

Assume for a moment, we know all ancestral nucleotides

- S_i = nt on node i
 $\rightarrow S_i = a^i$ if i leaf of T
- $p(i)$ = parent node of i
- $c(i)$ = children nodes of i
- t_i = branch length $p(i) \rightarrow i$

$$\begin{aligned} \Rightarrow P(\{S_i\}_{i \in T} \mid \omega, \mu, T) \\ &= \prod_{i \in T, i \neq r} P(S_i \mid S_{p(i)}, t_i) \\ &\quad \times \underbrace{P(S_r)}_{\text{root node, no ancestor}} \\ &\quad \rightarrow P(S_r) = \omega(S_r) \end{aligned}$$

But we do not know the nt of the internal nodes of T

$\rightarrow$ need to average over them

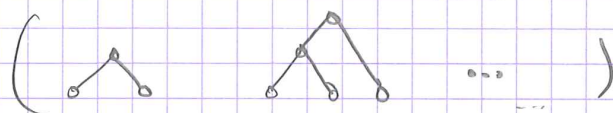
$$\begin{aligned} \text{INT} &= \{i \in T \mid i \text{ not leaf}\} \\ &= \{i \in T \mid \exists j, i = a(j)\} \\ &= \text{internal nodes} \end{aligned}$$

$$\Rightarrow P(D \mid \omega, \mu, T) = \sum_{\substack{\{S_i = A, C, G, T\} \\ i \in \text{INT}}} P(\{S_i\}_{i \in T} \mid \omega, \mu, T)$$

(remember: $D = \{s_i \mid i \notin INT\}$)
 = nt on leaves of T

(2.5)

problem binary tree with M leaves
 has $M-1$ internal nodes

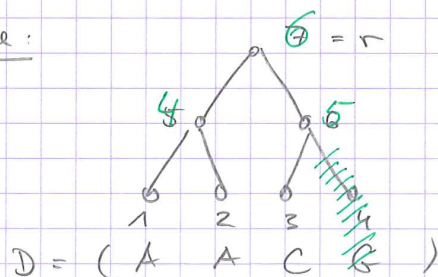


- Sum over 4^{M-1} nt configs
- infeasible for large M

Solution: efficient implementation by
 dynamic programming
 [Felsenstein '81]

→ recursive procedure over
 subtrees

example:



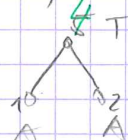
→ introduce subtree T_i of node i
 (e.g. $T_5 =$)

and probs of leaf config in T_i given
 sth nt a in root node i

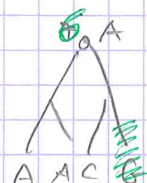
→ $g_i(a)$

e.g.

$g_4(T) =$



$g_7(A) =$



(2.6)

In particular:

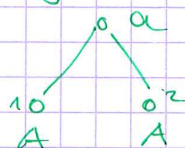
$$g_1(a) = \delta_{a,A} = \begin{cases} 1 & \text{if } a=A \\ 0 & \text{else} \end{cases}$$

$$g_2(a) = \delta_{a,A}$$

$$g_3(a) = \delta_{a,C}$$

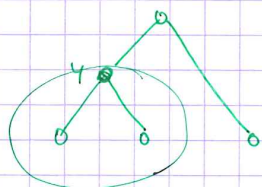
~~$$g_4(a) = \delta_{a,A}$$~~

→ fixed by leaves

Now: $g_4(a)$ 

$$= P(A|a, t_1) \cdot P(A|a, t_2)$$

$$= \left(\sum_b P(b|a, t_1) \cdot g_1(b) \right) \times \left(\sum_c P(c|a, t_2) \cdot g_2(c) \right)$$



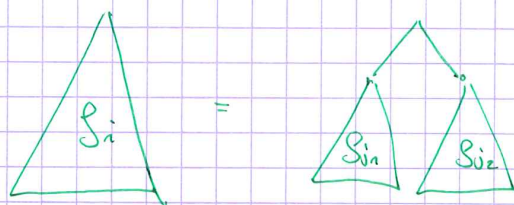
$$\Rightarrow g_6(a) = \left(\sum_b P(b|a, t_4) g_4(b) \right) \times \left(\sum_c P(c|a, t_3) g_3(c) \right)$$

$$P(AAC | w, \mu, T) = \sum_a w(a) \cdot g_6(a)$$

root node = r

leaf nodes $1 \dots M$ with $a_1^1, \dots, a_1^M$ GENERAL TREEInitialize: i leaf: $g_i(a) = \delta_{a, a^i}$ Iteration: $i \in \text{INT}$:

$$g_i(a) = \prod_{\substack{j \in C(i) \\ \text{children of } i}} \left(\sum_{b \in \{A, C, G, T\}} P(b|a, t_j) g_j(b) \right)$$



→ Iterate from leaves to root

$$P(a^1, \dots, a^m \mid w, \mu, T) = \sum_b w(b) g_r(b)$$

↑
root

Back to full data set $\Rightarrow$ L columns

D_i = column i of D

$$= \begin{pmatrix} a_i^1 \\ \vdots \\ a_i^m \end{pmatrix}$$

$$w_i = (w_i(A), w_i(C), w_i(G), w_i(T))$$

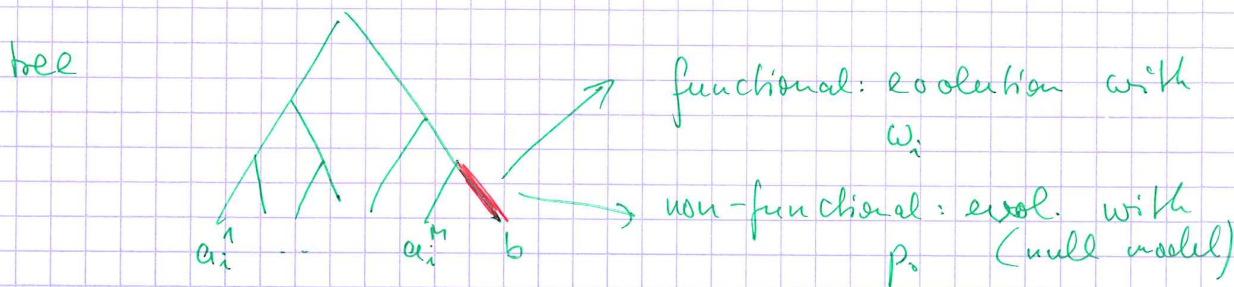
weights for pos. i

$$\Rightarrow \boxed{P(D \mid w, \mu, T) = \prod_i P(D_i \mid w_i, \mu_i, T)}$$

• Scoring a sequence

$b = (b_1, \dots, b_e)$ from orthologous regulatory region

functional or not



$$P_{\text{functional}}(\underline{b} \mid D, w, \mu, T) = \frac{P(b, D \mid w, \mu, T)}{P(D \mid w, \mu, T)} \quad (2.8)$$

$$- P(b_i | a_i, t) = e^{-\mu t} \delta_{ab_i} + (1 - e^{-\mu t}) w_i(b_i)$$

→ functional site evolution

$$P_0(\underline{b} \mid D, w, \mu, T) = \frac{P(b, D \mid w, \mu, T, p_0)}{P(D \mid w, \mu, T)}$$

$$- P(b_i | a_i, t) = e^{-\mu t} \delta_{ab_i} + (1 - e^{-\mu t}) p_0(b_i)$$

⇒ likelihood score

$$S(\underline{b}) = \sum_{i=1}^n \log \frac{P_{\text{funct.}}(b_i, D_i \mid w_i, \mu_i, T)}{P_0(b_i, D_i \mid w_i, \mu_i, p_0, T)}$$

Finding a motif → estimation of w

⇒ max likelihood

$$\hat{w}^{\text{ML}} = \underset{\text{PWM } w}{\operatorname{argmax}} P(D \mid w, \mu, T)$$

→ needs numerical solution

Reconstruction of ancestral sequences

→ Do not sum over one internal ~~site~~ ^{node} $\frac{1}{2}$
to determine its probability for each
type of nt.