Wilberd van der Kallen Celebration 14-3-2008

Jan Draisma

I "Ideal approach to phylogenetics" (SIAM News)

ultimate goal: given genetic data of n species, find most likely evolutionary tree.

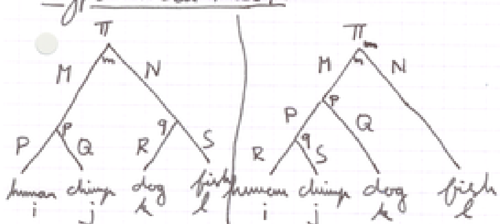
intermediate goal: given genetic data of n species and a hypothetical tree, decide whether tree fits data.

data: n strings over $\{A, C, G, T\}$, aligned:

human	A A A C G T G A C	gives empirical distribution on $\{A, C, G, T\}^n$.
chimp	A A A G G A G A C	
dog	A A C C G A A T C	
fish	C C G A G G C A C	

eg. $Pr(A,A,A,C) = \frac{2}{9}$, $Pr(T,A,A,G) = \frac{1}{9}$.

hypothetical tree T:



gives statistical model = family of prob. distr on $\{A, C, G, T\}^n$ parameterised by

① initial distr. π

② transition mats M, N, P, Q, R, S.

$$\boxed{Pr(i,j,k,l) = \sum_{m,p,q} \pi_m M_{pm} N_{lm} P_{qp} Q_{kp} R_{iq} S_{jq}}$$

$$Pr(i,j,k,l) = \sum_{m,p,q} \pi_m M_{pm} P_{ip} Q_{jp} N_{qm} R_{kq} S_{lq}.$$

(2)

<u>problem</u>: $\exists ?\ \pi, M, N, P, Q, R, S$ that give rise (almost) to the empirical distribution?

<u>more precisely</u>: we have a polynomial map

$$\Phi_T : \Delta^3 \times \{4\times 4 \text{ stoch. mats}\}^{\text{edges of } T} \longrightarrow \Delta^{4^n - 1}$$

and an emp. distr. $\in \Delta^{4^n-1}$, and want to test if emp. distr. $\in \operatorname{im} \Phi$.

<u>idea</u>: find the ideal of $\operatorname{im} \Phi$, and test equations on the emp. distr.

<u>1st simplification</u>: forget about non-neg. conds 1, and allow complex numbers:

$$\tilde{\Phi}_T : \{\pi \in \mathbb{C}^4 \mid \Sigma \pi_i = 1\} \times \{4\times 4 \text{ mats} / \mathbb{C} \text{ w col sums } 1\}^{\text{edges}(T)}$$

$$\longrightarrow \{\sigma \in \mathbb{C}^{(4^n)} \mid \Sigma \sigma_{\bar{j}\bar{k}\bar{l}} = 1\}. \qquad I(\operatorname{im}\Phi) = I(\operatorname{im}\tilde{\Phi})$$

<u>2nd simplification</u>: forget about π and col sums:

$$\Psi_T : \{4\times 4\text{-mats} / \mathbb{C}\}^{\text{edges}(T)} \longrightarrow \mathbb{C}^{(4^n)}.$$

<u>ex</u>: $\sum_{m,p,q} M_{pm} P_{ip} Q_{jp} N_{qm} R_{kq} S_{lq}$

<u>lm</u> the cone over $\operatorname{im} \tilde{\Phi}$ lies dense in $\operatorname{im} \Psi$.

<u>goal</u>: find the ideal of $\operatorname{im} \Psi$, given the tree T.

<u>thm</u> (Allman - Rhodes) From set-theoretic equations for  one can construct set-theoretic equations for all trees in which all valencies ≤ 3.

<u>thm</u> (Draisma - Kuttler) Same is true w set-theoretic replaced by scheme-theoretic.

(2½) G finite group (for ex. $G = \{1\}$)

def a G-spaced tree consists of :

- a finite tree T
- a finite G-set B_p for every vertex p of T.

set $V_p := \mathbb{C} B_p$, w bilin. form $\langle b, b' \rangle = \delta_{b,b'}$.

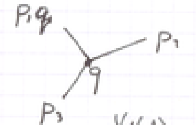
def a representation of a G-tree is collection
$A = (A_{p,q})_{p \sim q}$
$(V_p \otimes V_q)^G$.

def $A \in rep^G(T)$ then $\Psi(A) \in \bigotimes_{p \in leaf(T)} V_p$

defined as follows:

· start with $\bigotimes_{p \sim q} A_{p,q} \in \bigotimes_{p \in vertex(T)} V_p^{\otimes degree(p)}$

· contract with $\bigotimes_{p \in internal(T)} \left(\sum_{b \in B_p} b^{\otimes degree(p)} \right)$.

ex: 

$\sum_{b \in B_q} (A_{qp_1}, b) \otimes (A_{qp_2}, b) \otimes (A_{qp_3}, b)$.

· leaves elt $\Psi(A)$ of $\bigotimes_{p \in leaf(T)} V_p$ (actually, G-invariant)

def $EM(T) = \{ \Psi(A) \mid A \in rep^G(T) \}$
(Equivariant Model)

thm (D-K) $I(EM(T))$ can be constructed from $I(EM(T'))$ where T' runs over the induced subtrees of T of diam ≤ 2.

thm $I(EM(T)) = \sum_{q \in vertex(T)} I(EM(\sigma_q T))$.

③

<u>major tool:</u>

<u>def</u> if $V \subseteq M_{k,l}(\mathbb{C})$, $W \subseteq M_{l,m}(\mathbb{C})$ varieties

then $V \cdot W = \overline{\{MN \mid M \in V, N \in W\}} \subseteq M_{k,m}$.

<u>prop</u> if $V \subseteq M_{k,l}(\mathbb{C})$ $W \subseteq M_{l,m}(\mathbb{C})$ ~~then~~ and $V = V \cdot M_{l,l}$, $W = M_{l,l} \cdot W$, then

$$I(V \cdot W) = I(V \cdot M_{l,m}) + I(M_{k,l} \cdot W)$$

<u>proof</u> uses Weyl's FFT for Gl_l!

<u>cor</u> $I(V \cdot M_{l,m}) = I(V') + I(R_l)$

$$V' = \{M \in M_{l,m} \mid M \cdot M_{m,l} \subseteq V\}$$
$$R_l = \{M \in M_{k,m} \mid \operatorname{rk} M \leq l\}.$$

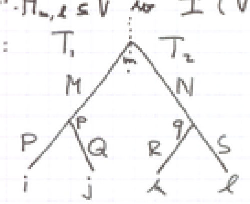
<u>pf of cor</u> set $W' := M_{m,l} \cdot M_{l,m}$

apply ~~for~~ prop. to (k, m, m, V', W'):

<u>cor</u> $I(V \cdot W)$ can be expressed in $I(V)$, $I(W)$.

$V' \cdot M_{m,l} \subseteq V$ so $I(V \cdot M_{l,m}) \subseteq I(V' \cdot M_{m,l} \cdot M_{l,m}) = I(V' \cdot W') = I(V') + I(\underbrace{M_{k,m} \cdot W'}_{R_l})$

<u>use</u>:



$$EM(T) = EM(T_1) \cdot EM(T_2):$$

$$\sum_{m,p,q} M_{pm} P_{ip} Q_{jp} N_{qm} R_{kq} S_{lq}$$

$$= \sum_m \underbrace{\Big(\sum_p M_{pm} P_{ip} Q_{jp}\Big)}_{\substack{(i,j,m)\text{-entry} \\ \text{von } \Psi_{T_1}(M,P,Q)}} \underbrace{\Big(\sum_q N_{qm} R_{kq} S_{lq}\Big)}_{\substack{(k,l,m)\text{-entry} \\ \text{von } \Psi_{T_2}(N,R,S)}}.$$

④

<u>BAD NEWS</u> : we don't know set- scheme-theoretic eqs for EM () with $G = \{1\}$,

alle B_p of size 4. Well, we know <u>some</u> but not if they suffice. (eg. a 1728-dim space of quintics)

<u>Luckily</u>, we do know such eqs ~~..~~ for models w more symmetry (larger G) imposed (Sturmfels - Sullivant - Casanellas - ....).
Our result explains ~~..~~ in a unified way how to construct all eqs for larger trees.

<u>Proof of prop</u> :

$$f \in I(V \cdot W) \rightsquigarrow \tilde{f} : M_{k,l} \times M_{l,m} \to \mathbb{C}, \quad (M, N) \mapsto f(MN)$$

$\tilde{f} \in I(V \times W)$ so

$$\tilde{f} = \tilde{f}_1 \otimes h_1 + \tilde{h}_2 \otimes f_2$$

where $f_1 \in I(V)$, $f_2 \in I(W)$, $h_1 \in \mathbb{C}[M_{l,m}]$, $h_2 \in \mathbb{C}[M_{k,l}]$

$\tilde{f}$ is GL_l-inv. thus

$$\tilde{f} = \rho(\tilde{f}) = \underbrace{\rho(f_1 \otimes h_1)}_{\substack{\| \\ \tilde{g}_1}} + \underbrace{\rho(h_2 \otimes f_2)}_{\substack{\| \\ \tilde{g}_2}}$$

$\tilde{g}_1 \in I(V) \otimes \mathbb{C}[M_{l,m}]$ (as V is GL_l-inv)
$\tilde{g}_2 \in \mathbb{C}[M_{k,l}] \otimes I(W)$ (as W ________)

FFT : $\exists g_1, g_2 \in \mathbb{C}[M_{k,m}] : \tilde{g}_i(MN) = g_i(MN)$, $i = 1, 2$.
$f = g_1 + g_2 \in I(V \cdot M_{l,m}) + I(M_{k,l} \cdot W)$. □