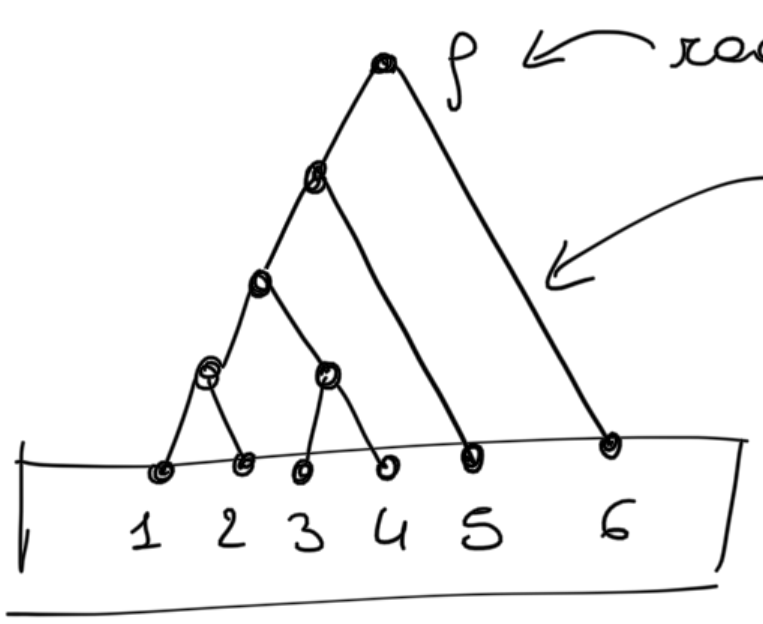
GROUP-BASED MODELS

07-05-21
Francesca

We buied on the shoulders of phylogenetic models



every edge $e \in E(T)$ has a transition matrix associated $\exp(Qt) = M^e$

← labeled leaves corresponding to extant species

ex 1: CFN model

$$Q^{CFN} = \begin{pmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{pmatrix} \longrightarrow \exp(Q^{CFN} t) = \begin{bmatrix} \frac{1}{2}(1+e^{-2\alpha t}) & \frac{1}{2}(1-e^{-2\alpha t}) \\ \frac{1}{2}(1-e^{-2\alpha t}) & \frac{1}{2}(1+e^{-2\alpha t}) \end{bmatrix}$$

↑ rate matrix

↑ transition matrix

ex 2: K3P model

$$Q^{K3P} = \begin{bmatrix} -\alpha-\beta-\gamma & \beta & \alpha & \gamma \\ \beta & -\alpha-\beta-\gamma & \gamma & \alpha \\ \alpha & \gamma & -\alpha-\beta-\gamma & \beta \\ \gamma & \alpha & \beta & -\alpha-\beta-\gamma \end{bmatrix}$$

$$\exp(Q^{K3P} t) = \begin{bmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{bmatrix}$$

$$a = \frac{1}{4}\left(1 + e^{-2(\alpha+\beta)t} + e^{-2(\alpha+\gamma)t} + e^{-2(\beta+\gamma)t}\right)$$

$$b = \frac{1}{4}\left(1 - e^{-2(\alpha+\beta)t} - e^{-2(\alpha+\gamma)t} + e^{-2(\beta+\gamma)t}\right)$$

$$c = \frac{1}{4}\left(1 - e^{-2(\alpha+\beta)t} + e^{-2(\alpha+\gamma)t} - e^{-2(\beta+\gamma)t}\right)$$

$$d = \frac{1}{4}\left(1 + e^{-2(\alpha+\beta)t} - e^{-2(\alpha+\gamma)t} - e^{-2(\beta+\gamma)t}\right)$$

i We end up with a model that is not so easy to describe!

<u>What</u>? The story of a simplification

<u>How</u>? Put some additional structure on the model in order to have a "nicer" description.

<u>Def</u>: A phylogenetic model is called a GROUP-BASED model if for every transition matrix M^e associated to $e \in E(T)$ there is a function $f^e: G \longrightarrow \mathbb{R}$ from a finite abelian group $(G, +)$ s.t

$$M^e_{gh} = M^e(g,h) = f^e(g-h)$$

<u>Note</u>: In most of the cases $G = \mathbb{Z}_2$ or $G = \mathbb{Z}_2 \times \mathbb{Z}_2$

<u>ex 1</u>: CFN is a group based model

$$\exp(Q^{CFN}t) = \begin{bmatrix} \frac{1}{2}(1+e^{-2\alpha t}) & \frac{1}{2}(1-e^{-2\alpha t}) \\ \frac{1}{2}(1-e^{-2\alpha t}) & \frac{1}{2}(1+e^{-2\alpha t}) \end{bmatrix}$$

$\{A, C\} \longrightarrow 0$

$\{G, T\} \longrightarrow 1$

Define $f: \mathbb{Z}_2 \longrightarrow \mathbb{R}$ as

$$f(0) = \frac{1}{2}(1+e^{-2\alpha t}) \qquad f(1) = \frac{1}{2}(1-e^{-2\alpha t})$$

$$\Rightarrow M_{00} = f(0-0) = f(0) \qquad M_{01} = f(0-1) = f(1) = M_{10}$$

$$M_{11} = f(1-1) = f(0)$$

<u>ex 2</u>: K3P is a group-based model

$$\exp(Q^{K3P}t) = \begin{bmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{bmatrix}$$

$A \longrightarrow (0,0)$
$C \longrightarrow (0,1)$
$G \longrightarrow (1,0)$
$T \longrightarrow (1,1)$

$$a = \frac{1}{4}\left(1 + e^{-2(\alpha+\beta)t} + e^{-2(\alpha+\gamma)t} + e^{-2(\beta+\gamma)t}\right)$$

$$b = \frac{1}{4}\left(1 - e^{-2(\alpha+\beta)t} - e^{-2(\alpha+\gamma)t} + e^{-2(\beta+\gamma)t}\right)$$

$$c = \frac{1}{4}\left(1 - e^{-2(\alpha+\beta)t} + e^{-2(\alpha+\gamma)t} - e^{-2(\beta+\gamma)t}\right)$$

$$d = \frac{1}{4}\left(1 + e^{-2(\alpha+\beta)t} - e^{-2(\alpha+\gamma)t} - e^{-2(\beta+\gamma)t}\right)$$

Define $f: \mathbb{Z}_2 \times \mathbb{Z}_2 \longrightarrow \mathbb{R}$ as

$$f(0,0) = a, \quad f(0,1) = b, \quad f(1,0) = c, \quad f(1,1) = d$$

$$H_{(0,1),(1,1)} = f\big((0,1) - (1,1)\big) = f(1,0) = c$$

Why not $G = \mathbb{Z}_4$? For example, $\bar{f}: \mathbb{Z}_4 \longrightarrow \mathbb{R}$ s.t

$$\bar{f}(0) = a, \quad \bar{f}(1) = b, \quad \bar{f}(2) = c, \quad \bar{f}(3) = d$$

$$H_{01} \neq \bar{f}(0-1) = \bar{f}(-1) = \bar{f}(3) = d$$

↑ the issue is that elements in $\mathbb{Z}_4$ are not inverses of themselves, so the symmetry of the matrix is not guaranteed.

Def: A 1-dimensional character of a group G is group homomorphism $\chi: G \longrightarrow \mathbb{C}^*$
(↑ additive, ↑ multiplicative)

Prop: $\hat{G} = \{\chi : G \longrightarrow \mathbb{C}^* \mid \chi \text{ homo}\}$ is a group with the operation $(\chi_1 \cdot \chi_2)(g) = \chi_1(g)\chi_2(g)$.
Moreover, $G \cong \hat{G}$ if G is a finite abelian group.

Pf
— It is not difficult to check that the operation is well defined and that $\hat{G}$ is a group.
We now show that $G \cong \hat{G}$. Given that G is a finite abelian group, $G \cong \mathbb{Z}_{p_1}^{\alpha_1} \times \cdots \times \mathbb{Z}_{p_r}^{\alpha_r}$, for $p_1, \ldots p_r$ primes and $\alpha_1, \ldots \alpha_r \in \mathbb{N}$.
It is also easy to see that $\widehat{G \times H} = \hat{G} \times \hat{H}$, hence it

suffices to show the isomorphism for $G \cong \mathbb{Z}_n$.
In this case, G is cyclic and $\chi : G \longrightarrow \mathbb{C}^*$ is characterized by where it sends a generator, which will go to a n-th root of unity (since $\chi(g)^n = \chi(ng) = \chi(0) = 1 \Rightarrow \chi(g)$ n-th root of 1). This correspondence describe the isomorphism desired.

Def: Discrete Fourier transform of $f : G \longrightarrow \mathbb{C}$ is $\hat{f} : \hat{G} \longrightarrow \mathbb{C}$ defined by

$$\hat{f}(\chi) = \sum_{g \in G} f(g)\chi(g)$$

Intuition: $f \longleftrightarrow A = (a_g)_{g \in G} \in \mathbb{C}^G \Rightarrow a_g \in \mathbb{C}$

$\hat{f} \longleftrightarrow B = (b_\chi)_{\chi \in \hat{G}} \in \mathbb{C}^G \Rightarrow b_\chi \in \mathbb{C}$

$$B = \begin{bmatrix} \chi_1 \\ \vdots \\ \chi_{|G|} \end{bmatrix} A$$

$\nwarrow$ $|G| \times |G|$ matrix

Fourier coordinates
$\updownarrow$
phase space

Th: $$f(g) = \frac{1}{|G|} \sum_{\chi \in \hat{G}} \hat{f}(\chi)\overline{\chi(g)}$$

Def: $f_1, f_2 : G \longrightarrow \mathbb{C}$ functions

$$(f_1 * f_2)(g) = \sum_{h \in G} f_1(h) f_2(g-h)$$

Prop: $\widehat{f_1 * f_2}(\chi) = \hat{f}_1(\chi)\hat{f}_2(\chi)$

pf

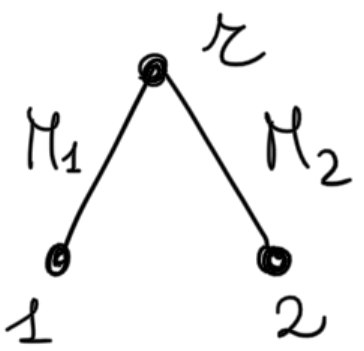
$$\widehat{f_1 * f_2}(\chi) = \sum_{g \in G} \sum_{h \in G} f_1(h) f_2(g-h)\chi(g) = \quad (g - h = g')$$

$$= \sum_{g' \in G} \sum_{h \in G} f_1(h) f_2(g')\chi(g'+h) =$$

$$= \left(\sum_{h \in G} f_1(h)\chi(h) \right) \left(\sum_{g' \in G} f_2(g')\chi(g') \right) =$$

$$= f_1(\chi) f_2(\chi)$$

ex T_2 :



Let π be the root distribution
$\pi : G \longrightarrow \mathbb{R}$

The probability of observing (g_1, g_2) on the leaves :

$$p(g_1, g_2) = \sum_{h \in G} \pi(h) M_1(h, g_1) M_2(h, g_2) =$$

$$= \sum_{h \in G} \pi(h) f_1(g_1 - h) f_2(g_2 - h)$$

From the phase space :

$$\begin{cases} g_1 - h = g_1' \\ g_2 - h = g_2' \end{cases}$$

$$\hat{p}(\chi_1, \chi_2) = \sum_{(g_1, g_2) \in G^2} \sum_{h \in G} \pi(h) f_1(g_1 - h) f_2(g_2 - h) \chi_1(g_1) \chi_2(g_2) =$$

$$= \sum_{g_1' \in G} \sum_{g_2' \in G} \sum_{h \in G} \pi(h) f_1(g_1') f_2(g_2') \chi_1(g_1' + h) \chi_2(g_2' + h) =$$

$$= \left(\sum_{h \in G} \pi(h) \chi_1(h) \chi_2(h) \right) \left(\sum_{g_1' \in G} f_1(g_1') \chi_1(g_1) \right) \left(\sum_{g_2' \in G} f_2(g_2') \chi_2(g_2) \right) =$$

$\uparrow$ $\chi_i(g_i' + h) = \chi_i(g_i') \chi_i(h), \quad \forall i = 1, 2$

$$= \hat{\pi}(\chi_1 \chi_2) \hat{f}_1(\chi_1) \hat{f}_2(\chi_2)$$

$\uparrow$ In the phase space the probability $p(g_1, g_2)$ has this "monomial" form, which leads to the following result.

Prop : Consider the map

$$\varphi_G : \mathbb{C}^G \times \mathbb{C}^G \times \mathbb{C}^G \longrightarrow \mathbb{C}^{G \times G}$$

$$\left((a_g)_{g \in G}, (b_g)_{g \in G}, (c_g)_{g \in G} \right) \longmapsto \left(a_{g_1} b_{g_2} c_{g_1 + g_2} \right)_{(g_1, g_2) \in G^2}$$

The Zariski closure of the image of φ_G is equal to the Zariski closure of the cone over the image of the group-based model of ex T_2.

Rem: φ_G is monomial map, so the Zariski closure of its image is a toric variety.

pf (idea)

Observe that in the phase space

$$\hat{p}(\chi_1, \chi_2) = \hat{f}_1(\chi_1)\hat{f}_2(\chi_2)\hat{\pi}(\chi_1\chi_2)$$

which is exactly of the same form of the image of φ_G. However, f_1, f_2, π are subjects to the following constraints:

$$\sum_{g\in G} f_1(g) = \sum_{g\in G} f_2(g) = \sum_{g\in G} \pi(g)$$

$$\Rightarrow \hat{f}_1(id) = \hat{f}_2(id) = \hat{\pi}(id) = 1, \text{ since}$$

$id(g) = 1 \quad \forall g \in G$ and thus

$$\hat{f}_i(id) = \sum_{g\in G} f_i(g)\, id(g) = 1 \quad \text{(same for } \pi)$$

$$\hat{f}_1 = \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}, \quad \hat{f}_2 = \begin{bmatrix} 1 \\ b_1 \\ \vdots \\ b_m \end{bmatrix}, \quad \hat{\pi} = \begin{bmatrix} 1 \\ c_1 \\ \vdots \\ c_m \end{bmatrix} \in U_0 \cong \mathbb{R}^m$$

$$\| \quad \{(x_0 : \_ : x_m) \mid x_0 \neq 0\}$$

So we actually have a map of the form of φ_G restricted to a dense subset $U_0 \times U_0 \times U_0$.

In order for it to be well-defined, we need to take the cone of the image (or in other terms to "projectivize"). Then we can conclude by taking the Zariski closures of both and conclude that they must be equal.

Up to now we have considered only group-based model with no additional assumptions on the root distribution or on the transition matrices. They can be collected under the name of GENERIC group-based models.

ex: Consider the generic group-based model on $\mathbb{Z}_2 \times \mathbb{Z}_2$

with identification:

$A = (0,0)$, $C = (0,1)$, $G = (1,0)$, $T = (1,1)$

$$\varphi_G(a_A, a_C, a_G, a_T, b_A, b_C, b_G, b_T, c_A, c_C, c_G, c_T) =$$

$$= \begin{bmatrix} a_A b_A c_A & a_A b_C c_C & a_A b_G c_G & a_A b_T c_T \\ a_C b_A c_C & a_C b_C c_A & a_C b_G c_T & a_C b_T c_G \\ a_G b_A c_G & a_G b_C c_T & a_G b_G c_A & a_G b_T c_C \\ a_T b_A c_T & a_T b_C c_G & a_T b_G c_C & a_T b_T c_A \end{bmatrix}$$

Macaulay 2 → vanishing ideal, to check it is toric

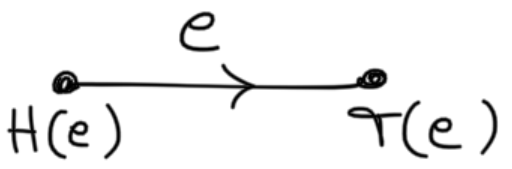
Example of generator: $q_{AA} q_{CG} q_{TC} - q_{TA} q_{CC} q_{AG}$ in Fourier coordinates.

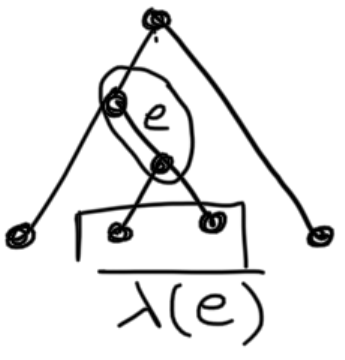
It is possible to go back to standard coordinate, for example:

$$q_{CG} = \sum_{(i,j),(k,\ell) \in \mathbb{Z}_2 \times \mathbb{Z}_2} \overbrace{(-1)^{j+k}}^{\chi_C(i,j)\chi_G(k,\ell)} p_{(i,j),(k,\ell)}$$

with $\chi_A(i,j) = 1$, $\chi_C(i,j) = (-1)^j$, $\chi_G(i,j) = (-1)^i$, $\chi_T(i,j) = (-1)^{i+j}$

We generalise the case  to the model represented by a rooted tree with m-leaves T.

- $\pi : G \longrightarrow \mathbb{R}$ root distribution
- $M^e(g,h)$ transition matrix
- $\forall e \in E(T)$ 

- $\lambda(e)$ = set of leaves lying below e 

- $V_{int} = V(T) \setminus \{\text{leaf}\}$

The probability of seeing the tuple $(g_1, \dots g_m)$ on the leaves is:

$$p(g_1, \dots g_m) = \sum_{g \in G^{V_{int}}} \pi(g^r) \overbrace{\prod_{e \in E(T)}} M^e(g_{T(e)}, g_{H(e)}) =$$

$$= \sum_{g \in G^{\text{int}}} \pi(g_r) \prod_{e \in E(T)} f^e(g_{H(e)} - g_{T(e)})$$

Prop: Let $p : G^m \longrightarrow \mathbb{R}$ denote the probability distribution of a group-based model on a m-leaf tree. Then the Fourier transform of p is

$$\hat{p}(\chi_1, \_ \chi_m) = \hat{\pi}\left(\prod_{i=1}^{m} \chi_i\right) \widehat{\prod_{e \in E(T)}} \hat{f}^e\left(\widehat{\prod_{i \in \lambda(e)}} \chi_i\right)$$

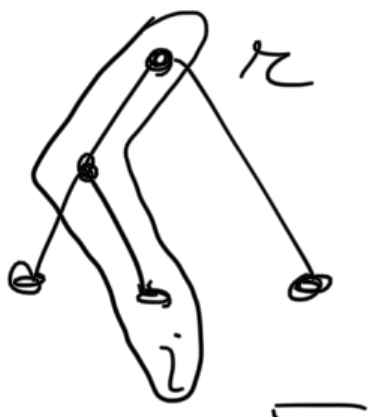
pf

$$\hat{p}(\chi_1, \_ \chi_m) = \sum_{(g_1, \_ g_m) \in G^m} \sum_{g \in G^{\text{int}}} \pi(g_r) \widehat{\prod_{e \in E(T)}} f^e\left(g_{H(e)} - g_{T(e)}\right) \widehat{\prod_{i=1}^{m}} \chi_i(g) =$$

change of variable: $g_e = g_{H(e)} - g_{T(e)} \quad \forall e \in E(T)$

$$\Rightarrow g_i = g_r + \sum_{\varepsilon \in P(i)} g_\varepsilon$$

where $P(i) = \{$edges connecting the root the the i-th leaf$\}$

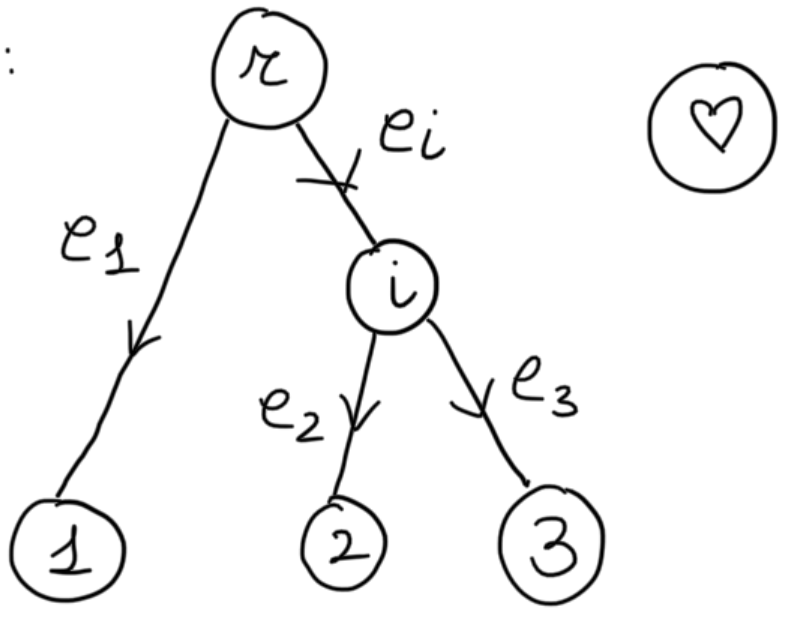


$$= \sum_{g_r \in G} \sum_{g \in G^{E(T)}} \pi(g_r) \widehat{\prod_{e \in E(T)}} f^e(g_e) \widehat{\prod_{i=1}^{m}} \chi_i\left(g_r + \sum_{\varepsilon \in P(i)} g_\varepsilon\right) =$$

$$\overset{*}{=} \left(\sum_{g_r \in G} \pi(g_r) \widehat{\prod_{i=1}^{m}} \chi_i(g_r)\right)\left(\prod_{e \in E(T)} \sum_{g_e \in G} f^e(g_e) \widehat{\prod_{i \in \lambda(e)}} \chi_i(g_e)\right) =$$

$$= \hat{\pi}\left(\prod_{i=1}^{m} \chi_i\right) \widehat{\prod_{e \in E(T)}} \hat{f}^e\left(\prod_{i \in \lambda(e)} \chi_i\right)$$

ex:



♡

$$\hat{p}(\chi_1,\chi_2,\chi_3)=\sum_{g_r\in G}\sum_{g_{e_i},g_{e_1},g_{e_2},g_{e_3}}\pi(g_r)f^{e_i}(g_{e_i})f(g_{e_1})\cdot$$

$$\cdot f^{e_2}(g_{e_2})f^{e_3}(g_{e_3})\chi_1(g_r+g_{e_1})\chi_2(g_r+g_{e_i}+g_{e_2})\chi_3(g_r+g_{e_i}+g_{e_3}$$

$$\overset{*}{=}\left(\sum_{g_r\in G}\pi(g_r)\chi_1(g_r)\chi_2(g_r)\chi_3(g_r)\right)\cdot$$

$$\cdot\left(\sum_{g_{e_i}}f^{e_i}(g_{e_i})\chi_2(g_{e_i})\chi_3(g_{e_i})\right)\cdot$$

$$\cdot\left(\sum_{g_{e_1}}f^{e_1}(g_{e_1})\chi_1(g_{e_1})\right)\left(\sum_{g_{e_2}}f^{e_2}(g_{e_2})\chi_2(g_{e_2})\right)\left(\sum_{g_{e_3}}f^{e_3}(g_{e_3})\chi_3\right)$$

$$=\hat{\pi}(\chi_1\chi_2\chi_3)\hat{f}^{e_i}(\chi_2\chi_3)\hat{f}^{e_1}(\chi_1)\hat{f}^{e_2}(\chi_2)\hat{f}^{e_3}(\chi$$

Cor: $\forall M$ the Fourier coordinates, the Zariski closure of the cone over the generic group-based model on n-leaves is equal to the Zariski closure of the image of

$$\varphi^{G,T}:\mathbb{C}^{|G|\times(|E(T)|+1)}\longrightarrow\mathbb{C}^{|G|^n}$$

$$\left((a^e_g)_{e\in E(T),g\in G},(a^r_g)_{g\in G}\right)\longmapsto a^r_{\sum_{i\in[n]}g_i}\overline{\prod}_{e\in E(T)}a^e_{\sum_{i\in\chi(e)}g_i}$$

ex: $T=$ 

$$\varphi^{G,T}(a)=a^r_{g_1+g_2+g_3}a^{e_i}_{g_2+g_2}a^{e_1}_{g_1}a^{e_2}_{g_2}a^{e_3}_{g_3}.$$

Up to now we have studied the generic group-based model. Now we see what happens when we add an hp on the root distribution. $\forall$ particular, it is often the case that the root distribution is assumed to be the stationary distribution.
It can be seen that in all the cases that we have encountered it coincides with the uniform distribution

Def: A stationary distribution π is s.t $\pi\exp(Qt)=\pi$

$\forall \alpha$.

ex: Consider the JC^{69} model with the uniform distribution

$$\pi(A) = \pi(C) = \pi(G) = \pi(T) = \frac{1}{4}$$

$$\left[\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\right] \frac{1}{4} \begin{pmatrix} 1+3e^{-4t\alpha} & 1-e^{-4t\alpha} & 1-e^{-4t\alpha} & 1-e^{-4t\alpha} \\ 1-e^{-4t\alpha} & 1+3e^{-4t\alpha} & 1-e^{-4t\alpha} & 1-e^{-4t\alpha} \\ 1-e^{-4t\alpha} & 1-e^{-4t\alpha} & 1+3e^{-4t\alpha} & 1-e^{-4t\alpha} \\ 1-e^{-4t\alpha} & 1-e^{-4t\alpha} & 1-e^{-4t\alpha} & 1+3e^{-4t\alpha} \end{pmatrix} =$$

$$= \left[\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\right]$$

" $\exp(Q^{JC}t)$

Prop: $f : G \longrightarrow \mathbb{C}$ constant defined by $f(g) = \frac{1}{|G|}$

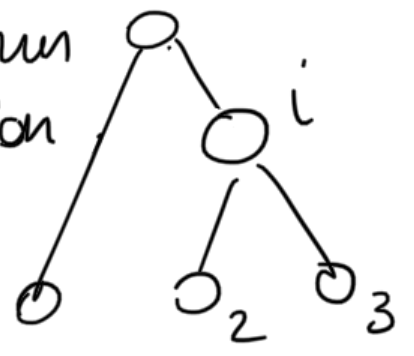
Then $\hat{f}(\chi) = \begin{cases} 1 & \text{if } \chi = id \\ 0 & \text{otherwise} \end{cases}$

Cor: In the Fourier coordinates, the Zariski closure of the cone over the generic group-based model with uniform root distribution is equal to the Zariski closure of the image of

$$\varphi^{G,T} : \mathbb{C}^{|G| \times E(T)} \longrightarrow \mathbb{C}^{|G|^m}$$

$$(a_g^e)_{\substack{e \in E(T) \\ g \in G}} \longrightarrow \begin{cases} \prod_{e \in E(T)} a^e_{\sum_{i \in \lambda(e)} g_i} & \text{if } \sum_{i \in [m]} g_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

ex: $T = \heartsuit$ with JC with uniform root distribution



$f^e(A) = a$

$f^e(C) = f^e(G) = f^e(T) = b$

a and b are functions of α and α

$\Rightarrow \hat{f}^e(A) = a + 3b \;,\quad \hat{f}^e(C) = \hat{f}^e(G) = \hat{f}^e(T) = a - b$

$\Rightarrow a_C^e = a_G^e = a_T^e \qquad \forall e \in E(T)$

Using these constraints we get linear invariants.

$$q_{CGT} = a_C^{e_1} a_G^{e_2} a_T^{e_3} a_{\underbrace{G+T}_{C}}^{e_i} =$$

$$= a_C^{e_1} a_C^{e_2} a_C^{e_3} a_C^{e_i}$$

$C = (0,1)$
$G = (1,0)$
$T = (1,1)$

$$q_{GTC} = a_G^{e_1} a_T^{e_2} a_C^{e_3} a_{\underbrace{T+C}_{G}}^{e_i} = a_C^{e_1} a_C^{e_2} a_C^{e_3} a_C^{e_i}$$

$\Rightarrow$ $q_{CGT} - q_{GTC}$ is a linear invariant for the model