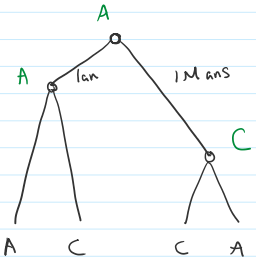
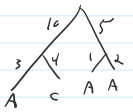
Inférence statistique

18 novembre 2022 10:24

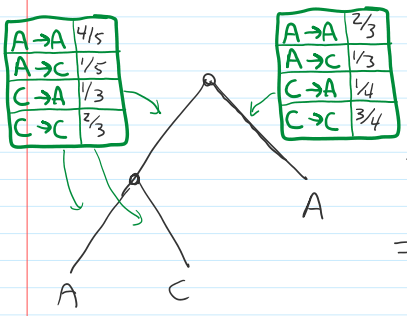


En principe, chaque branche doit avoir ses propres coûts, qui dépendent de leur longueur.
probabilité

On suppose qu'on nous donne
- un arbre T, feuille / caractère
- longueurs sur branches
- chaque branche a une matrice de proba. de mutation qui lui est propre



$\forall uv \in E(T)$, $Pr_{uv}[a,b]$ = prob. que a mute en b sur uv



Supposons $\Sigma = \{A, C\}$

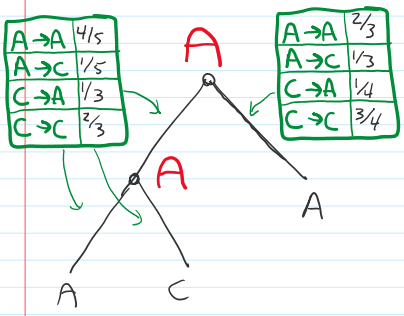
vraisemblance $(T) \approx Pr[\text{séquences} \mid T]$

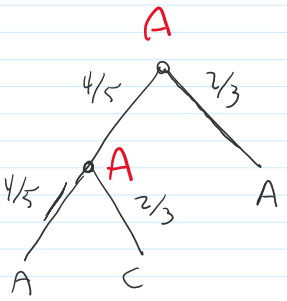
$= \sum_{\text{étiquetages possibles de T}} Pr[\text{l'étiquetage a eu lieu}]$

$= Pr\left[\begin{smallmatrix} & A & \\ A & & C \end{smallmatrix}\right] + Pr\left[\begin{smallmatrix} & A & \\ C & & A \end{smallmatrix}\right] +$

$Pr\left[\begin{smallmatrix} & C & \\ A & & C \end{smallmatrix}\right] + Pr\left[\begin{smallmatrix} & C & \\ C & & A \end{smallmatrix}\right]$

ex:





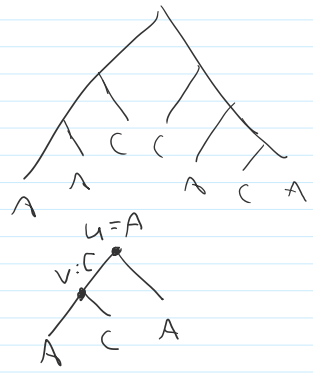
Prob. que tous les changements prédits sur les branches surviennent = $\frac{2}{3} \cdot \frac{4}{5} \cdot \frac{4}{5} \cdot \frac{1}{5} \stackrel{?}{=} \frac{32}{375}$

$Pr\left[\begin{smallmatrix} & A & \\ A & & A \end{smallmatrix}\right] = \frac{4}{5} \cdot \frac{2}{3} \cdot \frac{4}{5} \cdot \frac{2}{3} \cdot \frac{1}{2}$

prob. d'avoir A à la racine

Il faut aussi considérer la prob que A apparaisse à la racine
Souvent, on attribue de façon uniforme, ie. $Pr[A \text{ racine}] = \frac{1}{2} = \frac{1}{|\Sigma|}$

- En général, T a n feuilles (1 caractère) accès à $Pr_{uv}[a,b]$
- Soit $\mathcal{E}$ l'ensemble des étiquetages possibles des noeuds internes de T
  ↳ $s \in \mathcal{E} \to \forall u \in V(T)$, $s(v)$ = caractère de v
- $\forall a \in \Sigma$, $\pi(a)$ = Prob. d'observer a à la racine



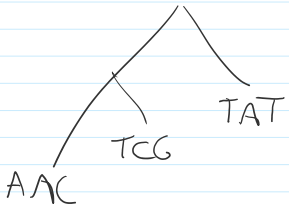
$s(u)=A$ $s(v)=C$

- On cherche

$$\sum_{s \in \mathcal{E}} \left(\prod_{uv \in E(T)} Pr_{uv}[s(u), s(v)] \right) \cdot \pi(s(\text{racine}))$$

produit itéré

Ceci est pour un seul caractère : hypothèse d'indépendance entre les positions
→ la vraisemblance de T = produit des vraisemblances par caractère



Calculer $\sum_{e \in \mathcal{E}} (\dots)$ peut prendre un temps $O(|\Sigma|^I)$ où I = #noeuds internes

naïvement

Il existe un algo en temps polynomial pour calculer $\sum_{e \in \mathcal{E}} (\dots)$ par prog. dynamique (algo de Felsenstein)