2. Profils phylogénétiques

Profil phylétique : vecteur $x[1..n]$ qui caractérise la distribution de gène x dans génomes $1, 2, \ldots, n$

- ⋆ présence (1) — absence (0)
- ⋆ nombre de homologues (0,1,2,...)
- ⋆ *best hit* : chercher homologue avec BLASTP et utiliser *E-value*.[1] Souvent[2], on transforme une valeur ϵ à

$$x[i] = \begin{cases} \frac{1}{-\lg \epsilon} & \{\epsilon < 1/2\} \\ 1 & \{\epsilon \geq 1/2\} \end{cases}$$

[1] E-value : nombre de séquences similaires en espérance ; petite valeure indique une similarité spécifique (on choisit typiquement $10^{-6}, 10^{-9}, 10^{-12}$ pour chercher de vrais homologues).

[2] Date, S. V. and Marcotte, E. M. (2003). Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology*, **21**, 1055–1062

Corrélations entre profils

Définition 2.1. ***Information mutuelle*** *entre deux variables aléatoires X, Y discrètes :*

$$\begin{aligned} I(X;Y) &= \underbrace{H(X)}_{\text{entropie de } X} - \underbrace{H(X|Y)}_{\text{entropie conditionnelle}} \\ &= H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \\ &= \sum_{\alpha,\beta} \mathbb{P}\{X=\alpha, Y=\beta\} \lg \frac{\mathbb{P}\{X=\alpha, Y=\beta\}}{\mathbb{P}\{X=\alpha\} \cdot \mathbb{P}\{Y=\beta\}} \end{aligned}$$

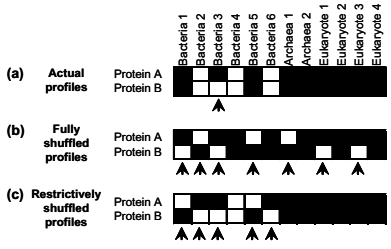
(Si X et Y sont indépendantes, alors $I(X;Y) = 0$.) On peut mesurer la corrélation entre deux profils $x[1..n]$ et $y[1..n]$ en considérant $(x[1], y[1]), (x[2], y[2]), \ldots, (x[n], y[n])$ comme des observations (indépendantes) du paire (X, Y) :

1. pour absence-présence, il suffit de compter les paires $0-0$, $0-1$, $1-0$, $1-1$.
2. pour valeurs réelles (comme $-1/\log E$), la solution usuelle est de discrétiser ...

Comparer avec corrélation entre profils randomisés

[3] Jothi, R., Przytycka, T. M., and Aravind, L. (2007). Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, **8**, 173

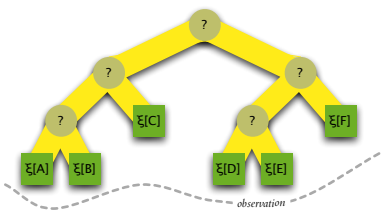
Best practices. [évaluation de prédiction de chemins dans KEGG[3]] (1) Faire attention à la spécificité des profils lors de la randomisation (2) Marche mieux avec actéries qu'avec eucaryotes. (3) Choisir des génomes diverses mais pas trop proches l'un à l'autre. (4) Bonne méthode mais non pas superbe (spécificité — sensitivité).



Profils et phylogénies

Si on a une phylogénie T, on peut penser à réconstruire l'histoire d'un profil Φ. Méthode de parcimonie pour étiquetage $\xi[u] \in \mathcal{X}$ de tout nœud u :

- ⋆ étiquettes $\xi[u] = \Phi[u]$ connues aux feuilles $u \in \mathcal{F}$
- ⋆ pénalisation fixe $\Delta(x \to y)$ pour changement d'étiquette entre parent et enfant



⋆ déterminer les étiquettes aux nœuds ancestraux qui minimisent

$$\underbrace{\sum_{uv\in T} \Delta\big(\xi[u] \to \xi[v]\big)}_{\text{somme de pénalités sur les arêtes}}$$

Variations : Dollo ($0 \to 1$ juste 1 fois), Fitch ($\Delta(a \to b) = 1$ pour tout $a \neq b$), Wagner ($\Delta(a \to b) = |a - b|$), carré ($\Delta(a \to b) = (a-b)^2$)

Programmation dynamique[4] : définir la fonction du *coût de sous-arbre* $f_u \colon \mathcal{X} \mapsto [0, \infty)$ comme le minimum de pénalités dans le sous-arbre de u implié par l'étiquette de u :

$$f_u(x) = \min \sum_{vw\in T_u; \xi[u]=x} \Delta(\xi[v] \to \xi[w]).$$

(On prend le minimum parmi tous les étiquetages du sous-arbre T_u, mais on garde l'étiquette de u fixe à x.) On a les récurrences :

$$f_u(x) = \begin{cases} 0 & \text{si } x = \Phi[u]\,; \\ \infty & \text{if } x \neq \Phi[u]\,; \end{cases} \qquad \{u \text{ terminal}\} \qquad (2.1a)$$

$$f_u(x) = \sum_{v\in \mathsf{enfants}(u)} \min_{y\in\mathcal{X}} \big(\Delta(x \to y) + f_v(y)\big) \qquad \{u \text{ ancestral}\} \qquad (2.1b)$$

```
    ANCESTRAL(u)                                   // (calculer f_u pour nœud u)
A1  si u est terminal alors f_u(x) ← {x = Φ[u]}?0 : ∞      // (terminal)
A2  sinon
A3      pour v ∈ enfants(u) faire
A4          f_v ← ANCESTRAL(v)
A5          calculer h_uv(x) ← min_{y∈X} (Δ(x → y) + f_v(y))
A6      mettre f_u(x) ← Σ_{v∈enfants(u)} h_uv(x)
A7  retourner f_u
```

```
    LABELING(v)                                    // (calculer un étiquetage optimal )
L1  si v est la racine alors ξ[v] = arg min_x f_v(x)
L2  sinon
L3      u ← parent de v ; x ← ξ[u]
L4      ξ[v] ← arg min_{y∈X} (Δ(x → y) + f_v(y))
L5  pour w ∈ enfants(v) faire LABELING(w)
```

Si on a un nombre fini d'étiquettes (comme 0 et 1), on trouve le minimum en ligne A5 en examinant toutes les valeurs possibles. En même temps, on peut stocker les minimums pour backtracking dans lignes L1 et L4. Dans d'autres cas (parcimonie de Wagner ou carrée), on peut implanter les algorithmes par la manipulation symbolique des fonctions f, h.

[4] Sankoff, D. and Rousseau, P. (1975). Locating the vertices of a Steiner tree in arbitrary metric space. *Mathematical Programming*, **9**, 240–246