Correction TD3

1. Introduction

a) Character-based tree reconstruction

- Character-based reconstruction algorithms use the (N,M) alignment matrix (n=#species, m=#characters) directly instead of using distance matrix. The goal is to determine what character strings at internal nodes would best explain the character strings for the n observed species. It may be nucleotides, where A, G, C, T are 'states' of this character. Other characters may be the # of eyes or legs or the shape of a beak or a fin.
- Difference with previous algorithms: tree is already given, so are the external leaves. Finding the correct tree is a difficult problem, super-exponential number of trees on n species.

b) Parcimony Principle

- **Principe de parcimonie** : minimum de causes élémentaires pour expliquer un phénomène. Autrement dit, la meilleure explication d'un phénomène est souvent la meilleure. Dans notre cas, on considère que l'étiquettage des noeuds internes donnant le minimum de mutations est le meilleur. Attention ce n'est pas forcément vrai (espèces pour lesquels les ailes sont apparues, puis disparues). D'où un score de parcimonie correspondant au nombre de mutations présentes pour un arbre avec un étiquettage donné, et que l'on cherche à minimiser.
  == Rasoir d'Occam : *Pluralitas non est ponenda sine necessitate*
- **Parcimonie** (large) : Etant donnée n sequences (espèces), trouver l'arbre T (topologie et étiquetage des nœuds internes) avec n feuilles étiquetées par les n séquences (espèces) qui minimise le score de parcimonie S(T)
  Petite parcimonie : Etant donnée la topologie de l'arbre et l'étiquetage interne (mais pas), trouver l'étiquetage interne minimisant S(T).

c) Sankoff-Fitch difference

- **Fitch** : même coût pour toutes les mutations (i.e on considère que les mutations de A à T et de A à G ont le même coût).
- **Sankoff** : chaque mutation à un coût différent, et on peut même avoir un coût différent pour passer de A>T où passer de T>A.

2. Fitch

Algorithme

Soit R(x) l'ensemble des étiquetages possibles pour le nœud interne x. Soient A et B les ensembles des étiquetages possibles pour les deux nœuds fils de x. Alors :

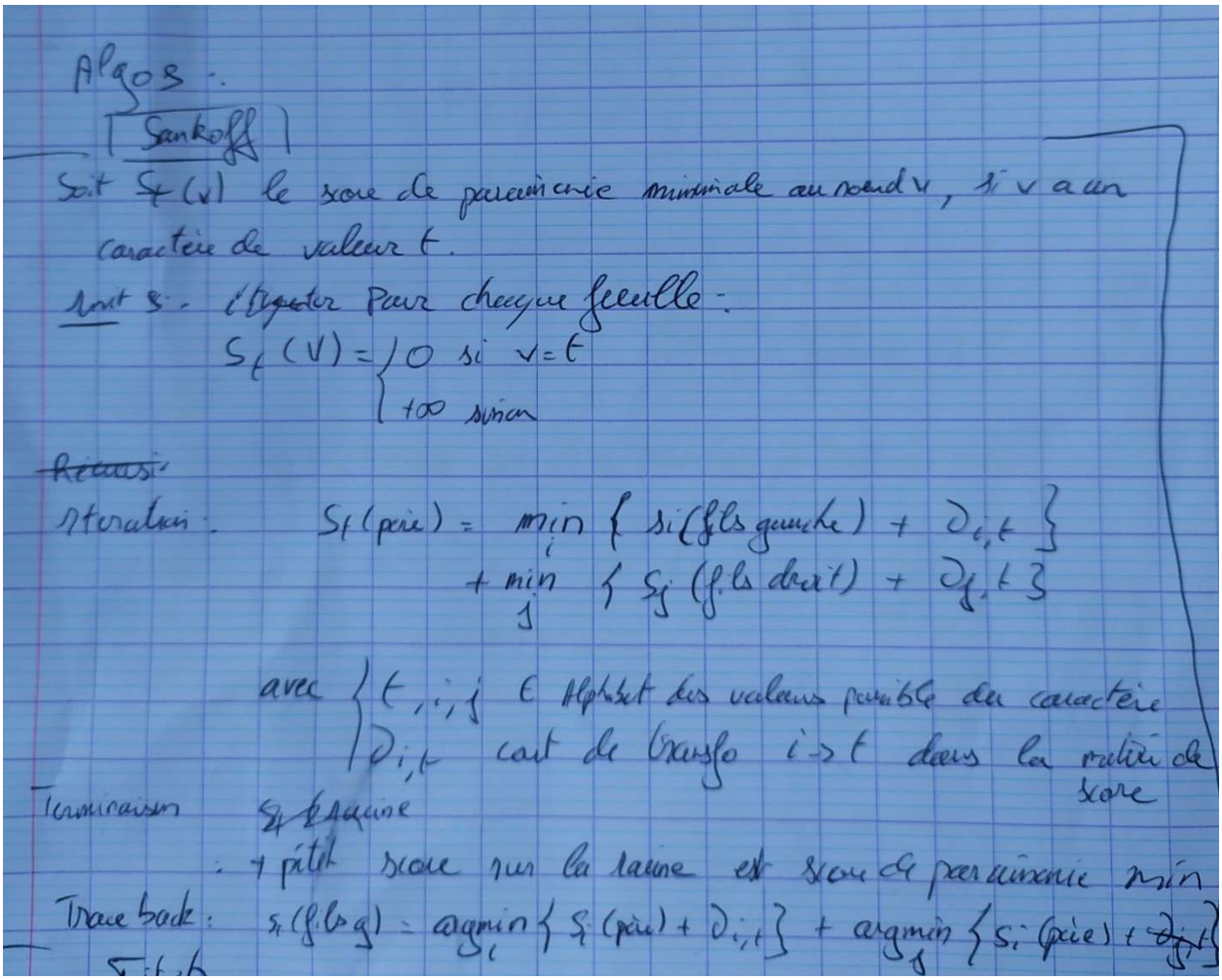
$R(x) = A \cap B$ si $A \cap B$ est non vide.

$R(x) = A \cup B$ sinon.

Score : Nombre d'opérations d'union.

Traceback : A partir de la racine, choix arbitraire parmi les caractères dans l'ensemble. Choix du même caractère chez les fils, arbitraire si pas possible.

3. Sankoff



Correction
(voir CorrigeSankoff.pdf)