

BioInformatique : Évolution et Phylogénèse

Plan du cours — J.-M. Steyaert

1. Historique : évolution des idées et concepts
2. Arbre phylogénique : pourquoi ? comment ?
3. Critères de comparaison
4. Construction des arbres : les difficultés
5. Le point de vue de l'algorithmicien

Historique

Voir la page Web de Jean-Baptiste Ferdy

http://162.38.181.25/~jbf/svc/svc_cm7.html

Une histoire des idées sur la classification du vivant

Critères de comparaison

Comparaison de l'hémoglobine (beta chain) de plusieurs espèces : la séquence est de 150 acides aminés environ. On a tabulé le nombre de différences. On aurait aussi pu (dû ?) considérer la matrice des différences.

Human	0	Gray kangaroo	38
Gorilla	1	Chicken	45
Gibbon	2	Frog	67
Rhesus monkey	8	Lamprey	125
Dog	15	Sea slug	127
Horse, Cow	25	Soybean	124
Mouse	27		

Comparaison des Cytochromes C

		1					6				10	
Human		Gly	Asp	Val	Glu	Lys	Gly	Lys	Lys	Ile	Phe	Ile
Pig		-	-	-	-	-	-	-	-	-	-	Val
Chicken		-	-	Ile	-	-	-	-	-	-	-	Val
Dogfish		-	-	-	-	-	-	-	-	Val	-	Val
Drosophila	<<<<	-	-	-	-	-	-	-	-	Leu	-	Val
Wheat	<<<<	-	Asn	Pro	Asp	Ala	-	Ala	-	-	-	Lys
Yeast	<<<<	-	Ser	Ala	Lys	-	-	Ala	Thr	Leu	-	Lys
		12		14			17	18		20		
Human		Met	Lys	Cys	Ser	Gln	Cys	His	Thr	Val	Glu	Lys
Pig		Gln	-	-	Ala	-	-	-	-	-	-	-
Chicken		Gln	-	-	-	-	-	-	-	-	-	-
Dogfish		Gln	-	-	Ala	-	-	-	-	-	-	Asn
Drosophila		Gln	Arg	-	Ala	-	-	-	-	-	-	Ala
Wheat		Thr	-	-	Ala	-	-	-	-	-	Asp	Ala
Yeast		Thr	Arg	-	Glu	Leu	-	-	-	-	-	-

Un arbre phylogénétique ainsi obtenu (W. M. Fitch et E. Margoliash)

charger CytoCphylo.ps

Remarquer la position bizarre de la branche Kangourou par rapport à celle des Primates

Phylogénie : Définitions

Arbre phyllogénétique :

- feuilles pour les espèces
- nœuds internes pour les ancêtres virtuels
- arêtes pour les lignes de filiation (valuées ou non)

Arbres enracinés ou non, planaires ou non.

En général arbres binaires si enracinés, ternaires si à plat.

Autres caractéristiques :

- garder la proximité pour des groupes voisins de leurs traits partagés
- privilégier les groupes monophylétiques
- éviter les groupes para et poly phylétiques
- (cladiste) état plésiomorphe (ancêtre) et apomorphe (dérivé)
- homologie et homoplasie dans l'évolution (conjointe ou indépendante)

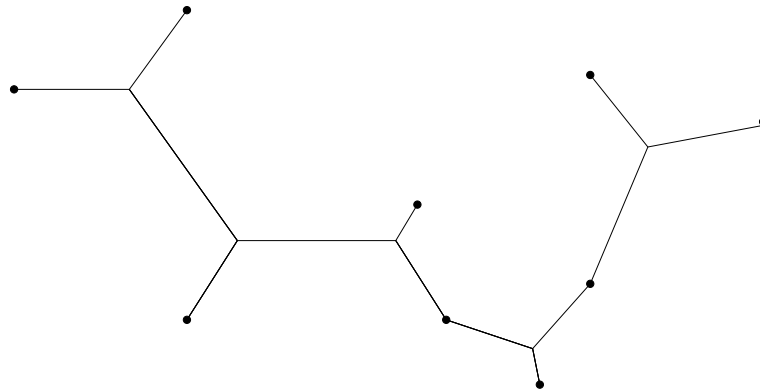
Principe de parcimonie : la Nature fait les choses à moindre coût.

Arbres de Steiner : modèle géométrique planaire.

Déf : (Arbre de Steiner d'un ensemble fini de points)

1. (Norme euclidienne) Soit X un ensemble de n points placés dans le plan. Trouver un arbre $T = (Y, A)$, dont les sommets couvrent les points de $X \subset Y$ et dont la somme des longueurs des arêtes soit minimale.
2. (Distance de Hamming) Soient A un ensemble fini et X un ensemble de n points placés dans A^n : la distance de deux points est le nombre de leurs coordonnées qui diffèrent. Trouver un arbre $T = (Y, A)$, dont les sommets couvrent les points de $X \subset Y$ et dont la somme des longueurs des arêtes soit minimale.

Exemple géométrique

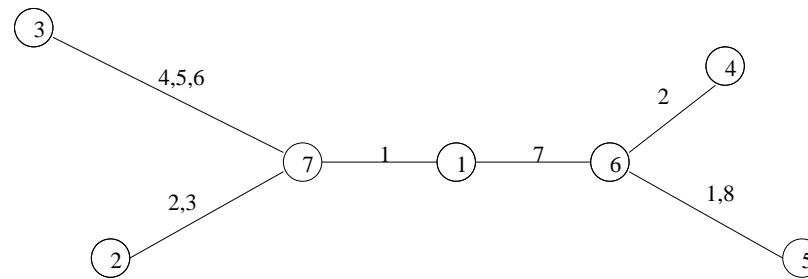


Un arbre « presque optimal » de Steiner dans le cas euclidien ; les sommets de l'arbre appartenant à X sont figurés en gras, alors que les autres restent des points.

Génétique : 5 espèces représentées par des « codes » à 8 bases (!)

	1	2	3	4	5	6	7	8
E_1	A	G	T	G	T	T	A	A
E_2	C	A	A	G	T	T	A	A
E_3	C	G	T	C	C	G	A	A
E_4	A	A	T	G	T	T	C	A
E_5	C	G	T	G	T	T	C	T

E_1 et E_4 sont à distance 2, puisqu'il faut effectuer deux substitutions (en positions 2 et 7) pour les identifier ; E_2 et E_5 sont à distance 4.



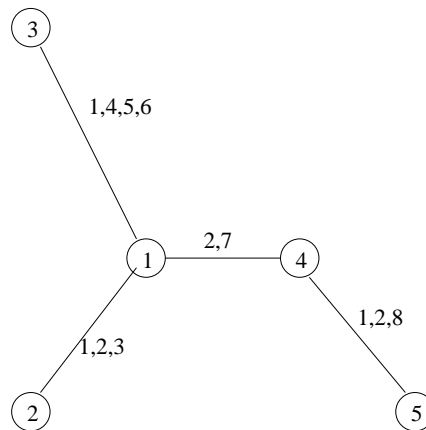
Un arbre de Steiner phylogénétique, optimal pour la distance de Hamming ; les arêtes portent les positions sur lesquelles on effectue des substitutions et les deux espèces fictives ou non attestées 6 et 7 sont introduites dans l'arbre

Le problème est NP-complet. (non résoluble en temps polynomial)

Méthode heuristique pour arbre de Steiner et distance de Hamming

Idée : se ramener à la recherche d'un arbre couvrant de poids minimal dont les sommets sont **tous** les génomes possibles (4^n) !

Point de départ :



Un arbre de Steiner philogénétique, en fait arbre de Trémaux du graphe initial : les substitutions sont indiquées par le rang des bases modifiées

La matrice des distances entre les génomes n'est pas obtenue par sommation des longueurs des arêtes, car des substitutions peuvent s'annuler le long d'un chemin.

	E_2	E_3	E_4	E_5
E_1	3	4	2	3
E_2		5	3	4
E_3			6	5
E_4				3

On procède en faisant grossir le graphe des sommets virtuels lentement (il en faut le nombre de génomes moins 1), en essayant de rapprocher le poids du graphe de la matrice des distances entre génomes.

	E_2	E_3	E_4	E_5
E_1	3	4	2	3
E_2		5	3	4
E_3			6	5
E_4				3

Réduction du poids :

- factoriser une modification sur deux arêtes issues d'un même sommet ($E3, E1, E2 \rightarrow E7$)
- ne pas faire une substitution et son inverse en créant un nouveau sommet ($E1, E4, E5 \rightarrow E6$).

Dans le cas étudié l'arbre obtenu est optimal, car il faut effectuer au moins 10 substitutions.

Les méthodes « probabilistes » (voir le cours sur les alignements)

On reprend les mêmes idées avec des distances qui sont des logarithmes de probabilités et qui donnent des arêtes additives... Comme dans le logiciel Clustal.

On recherche des propriétés particulières sur la longueur des arêtes (notion de distance) :

- $d(i, i) = 0, d(i, j) > 0, d(i, j) = d(j, i)$
- distance : $d(i, j) = 0 \Rightarrow i = j$
- métrique : $d(i, j) \leq d(i, k) + d(k, j)$
- ultramétrique : $d(i, j) \leq \max(d(i, k), d(k, j))$

Construction hiérarchique montante : regrouper les sous-arbres deux à deux

UPGMA : la méthode du lien moyen qui choisit les deux sommets les plus proches (k, l) et leur en associe un nouveau t : $d(i, t) = (n_k d(i, k) + n_l d(i, l)) / (n_k + n_l)$.

Adaptée à une ultramétrie : horloge moléculaire

NJ : pour minimiser la longueur totale des branches (crée un arbre aussi équilibré que possible). (On part d'un arbre à $n + 1$ sommets qu'un transforme en arbre binaire selon le principe de parcimonie ci-dessus). On répète l'opération sur le gros sommet restant, etc.

- Les résultats sont fluctuants pour plusieurs raisons :
- les critères retenus ne couvrent pas tous les paramètres
 - le temps est très mal représenté
 - les mesures de distance sont incertaines

Conclusions et perspectives

Autres pistes :

- modifier les notions de distance sur de nouveaux critères : minisatellites, mutations avec recopie, renversement des gènes, duplication
- combiner les critères ; l'univers des gènes est-il plat ?

Tant va l'autruche à l'eau qu'à la fin elle se palme
Raymond Queneau