Université des sciences de Montpellier 2

Rapport de projet

Master 2 STIC & SANTE, parcours Bioinformatique, Connaissances, Données

Bibliographie sur la reconstruction de relations ancestrales

Encadrants : Séverine Bérard, Vincent Berry et Eric Tannier

JEAN Pierre-Antoine

2012-2013

Sommaire

Table des figures :

I. Introduction :

Jusqu'au XVIII^{ème} siècle, toute manifestation vivante était interprétée comme le résultat de l'œuvre du créateur. C'est à la fin du XIX^{ème} siècle que certains scientifiques ont proposé une autre vision du monde et ont mis fin à la théorie du fixisme. Jean-Baptiste Lamarck (1744 – 1829) a ouvert la voie avec sa théorie du transformisme, démontrant que la diversité des êtres vivants est le fruit d'une longue transformation et que les caractères acquis sont ensuite transmis à la descendance. Puis, à partir de 1859 avec la publication « L'origine des espèces » par Charles Darwin (1809 – 1882), la théorie de l'évolution s'est imposée dans la communauté scientifique. C'est le docteur Ernst Haeckel (1834 – 1919) le premier à employer le mot « phylogénie » en 1866. La phylogénie est la science qui établit des relations de parenté entre les organismes aussi appelés taxons. Elle attise la communauté scientifique, des multitudes de recherches basées sur l'évolution ont été menées pour tenter d'expliquer l'histoire des espèces. Dans ses débuts, la phylogénie consistait à prendre en note les caractères morphologiques des espèces. Mais depuis les années 80 et grâce à l'amélioration des techniques de séquençage, la phylogénie moléculaire a fait son apparition, et la communauté scientifique a commencé à chercher des liens de parenté entre les espèces grâce à l'étude de leurs séquences biologiques, nucléotidiques (gènes) et protéiques (protéines). L'intérêt de la phylogénie est de comprendre l'origine de la vie, d'étudier la biodiversité, de déterminer l'origine géographique des espèces et aussi de comprendre les mécanismes d'évolution des organismes, par exemple, le phénomène de spéciation a permis l'évolution des espèces ainsi que la construction d'une filiation sous forme d'arbre. De plus des recherches sont actuellement faites sur la phylogénie du virus du VIH, pour comprendre comment ce virus évolue afin de mieux s'en protéger. Une étude classique d'inférence phylogénétique considère l'évolution des gènes indépendamment les uns des autres. Donc les méthodes classiques d'inférence phylogénétique sont clairement insuffisantes pour décrire des événements qui affectent plusieurs gènes à la fois, parce qu'elles font cette hypothèse d'indépendance. De plus, les méthodes classiques ne reconstruisent parfois pas le bon arbre parce que la séquence nucléotidique ou protéique manque de signal, il faut donc des informations supplémentaires. Parmi les nombreuses méthodes de reconstruction de l'histoire évolutive, certaines d'entre elles, incorporent ces informations supplémentaires à la phylogénie. Nous nous intéresserons plus particulièrement à deux types d'informations : l'adjacence des gènes sur le chromosome et l'interaction du produit des gènes, les protéines, dans l'organisme. L'intérêt est de reconstruire les relations (adjacences, interactions, régulations,...) ancestrales et retracer leur évolution pour comprendre le processus historique qui est à l'origine de la structure et du fonctionnement des génomes actuels. De nos jours, nous avons à notre disposition un nombre important de séquences de génomes, rendant possible la reconstruction de génomes anciens, qui ont subit des réarrangements génomiques à grande échelle. Le but de cette recherche bibliographique est de recenser les outils et algorithmes capables de retracer l'évolution de relations entre gènes.

II. Descriptif du déroulement du stage :

Le temps imparti pour le projet a été divisé en 2. La première moitié a été dédiée à une vaste recherche d'article dans le but de repérer dans la bibliographie tous les articles traitant de l'évolution en tenant compte des informations données par la phylogénie classique, les arbres de gènes, mais aussi d'autres types de relations impliquant plusieurs gènes à la fois, les synténies, les interactions protéines ou autres. L'idée était de rechercher des méthodes se rapprochant de l'article scientifique de [S. Bérard, 2012]. Nous avons utilisé les moteurs de recherches spécialisés (Pubmed) ou non (Google) pour trouver d'autres articles faisant référence aux mots clés, adjacence, phylogénie, réseaux d'interaction protéine-protéine, synténie. Cette première partie,

nous a permis de lire rapidement une vingtaine d'articles scientifiques en référence au domaine d'étude. À partir de ces articles, nous avons réalisé un tri afin d'en récupérer les meilleurs. Pour trier les articles nous avons utilisé une grille de lecture (disponible en annexe), cela nous a permis de comparer les articles de manière efficace. Suite à cette première phase on a choisit 2 types de relations: les synténies et les réseaux d'interactions protéiques, qui sont les deux types d'interactions les plus présentes dans la bibliographie. Puis on a revu parmi les articles, lesquels étaient les plus intéressants, à cette étape il y en avait encore une dizaine, les références sont données en annexe. Parmi ces articles nous en avons sélectionnés trois qui nous paraissait les plus pertinents, deux traitent des synténies et un des réseaux d'interaction protéine-protéine. La seconde moitié du temps a été consacrée à l'étude plus détaillée des articles sélectionnés. Nous avons décrit les méthodes et les algorithmes mis en place, afin de réaliser une comparaison entre les articles.

Ainsi ce rapport s'articule en 3 parties principales. La première permet de décrire et définir le cadre bioinformatique et phylogénétique, partie III page 5. La deuxième correspond à la description de 3 algorithmes extraits d'articles différents, partie IV page 9. Finalement la troisième partie fera office de conclusion où une comparaison entre les algorithmes sera établie, partie V page 20.

III. Définitions fondamentales :

Une étude classique d'inférence phylogénétique permet d'obtenir, à partir d'alignement de séquences d'ADN ou d'acides aminés, des arbres de gènes représentant les relations ancestrales à l'intérieur des familles multigéniques, ces familles représentent un ensemble de gènes ayant des homologies de séquences au sein du même génome. Il existe de nombreuses méthodes, que nous détaillerons, permettant d'inférer un arbre de gènes. Dans un premier temps nous allons définir les événements de transformation qui affectent les nucléotides et les portions d'ADN. Puis dans un second temps nous détaillerons les étapes d'une inférence phylogénétique.

Définition 1 : Événements de transformation des nucléotides :

La différence entre deux séquences est décomposée en différences élémentaires. Ces différences constituent des opérations de transformation élémentaires pour passer d'une séquence à une autre. Classiquement on considère :

– les substitutions, correspondant aux remplacements d'un caractère par un autre.
– et les délétions et insertions, correspondant au retrait ou à l'ajout d'un caractère.

Définition 2 : Événements de transformation des portions d'ADN (gènes) :

À l'échelle des gènes, il existe également des événements de transformation, ces événements peuvent impliquer des adjacences entre deux gènes (*cf* définition 4, adjacences page 8) :

– la duplication ;

Processus par lequel un gène est copié et transposé à un autre endroit dans le génome, possiblement à côté.

– le transfert ;

Introduction dans le génome d'un gène provenant d'un autre organisme, ou du même organisme, par exemple en plusieurs exemplaires pour renforcer son expression.

– la perte ;

Cet événement correspond à la disparition d'un gène qui peut être due à la transformation d'un

gène en un pseudo-gène ou à la suppression du gène par réarrangement génomique.

– l'inversion ;

L'inversion des gènes *a* b_1 b_2 b_3 *c* donne *a* b_3 b_2 b_1 *c*.

III.1. L'alignement :

La réalisation d'une phylogénie nécessite une bonne connaissance des séquences que l'on analyse en s'assurant de leur qualité et de leur cohérence, la qualité de l'alignement multiple des séquences en dépend. L'alignement est une étape cruciale qui permet de choisir les sites (colonne de l'alignement) qui seront utilisés dans les analyses phylogénétiques, cf Figure 1. Il prend en paramètre des séquences et un schéma de score puis réalise une mise en correspondance des séquences.

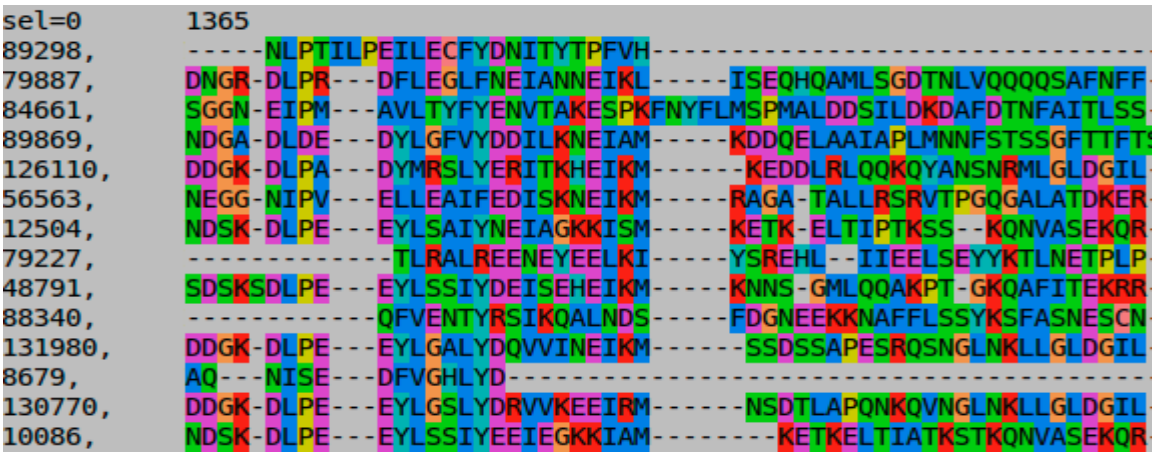

Figure 1 : Présentation d'un alignement multiple [1]

Le but de l'alignement en phylogénie est de déterminer des gènes homologues, deux gènes sont homologues lorsqu'ils descendent d'un même ancêtre commun. L'homologie se distingue en deux groupes : l'orthologie et la paralogie. Deux séquences homologues de deux espèces différentes sont orthologues si elles descendent d'une séquence unique présente dans le dernier ancêtre commun aux deux espèces et deux séquences homologues au sein d'une espèce (ou d'espèces différentes) sont paralogues si elles résultent d'une duplication génique.

III.2. Les méthodes d'inférence phylogénétique :

Il existe plusieurs méthodes permettant d'inférer un arbre de gènes à partir d'un alignement de séquences nucléiques ou protéiques.

III.2.a. Le maximum de parcimonie :

Le maximum de parcimonie est une approche qui minimise le nombre ou le coût de différents événements. En d'autres termes, étant donnés une matrice S de m caractères et un coût pour chaque transformation de l'état d'un caractère C(x→y) , chercher la (ou les) phylogénie(s) la (ou les) moins coûteuse(s). Par exemple : parcimonie de Fitch (C(x→y) = 1) [Fitch W, 1971], de Sankoff [Sankoff D, 1975] (C(x→y) = C(y →x))...

III.2.b. Les méthodes de distance :

Les méthodes de distance permettent de trouver l'arbre dont la distance entre chaque paire de feuilles correspond aux meilleures distances observées entre les gènes. La distance peut être définie comme le nombre de substitution par site. Il existe plusieurs algorithmes qui permettent de reconstruire un arbre phylogénétique qui correspond aux distances estimées, par exemple l'algorithme d'agglomération, Neighbor Joining [Saitou N. 1987].

III.2.c. Les approches probabilistes :

Il existe deux approches probabilistes, le maximum de vraisemblance qui maximise la probabilité d'observer les données selon un modèle et un arbre donné et l'inférence bayésienne qui est la probabilité postérieure maximale qui permet de déduire l'incertitude correspondant à l'arbre et à ses paramètres. Ainsi cette deuxième méthode retourne l'ensemble des arbres les plus probables, avec une estimation de leur probabilité postérieure respective.

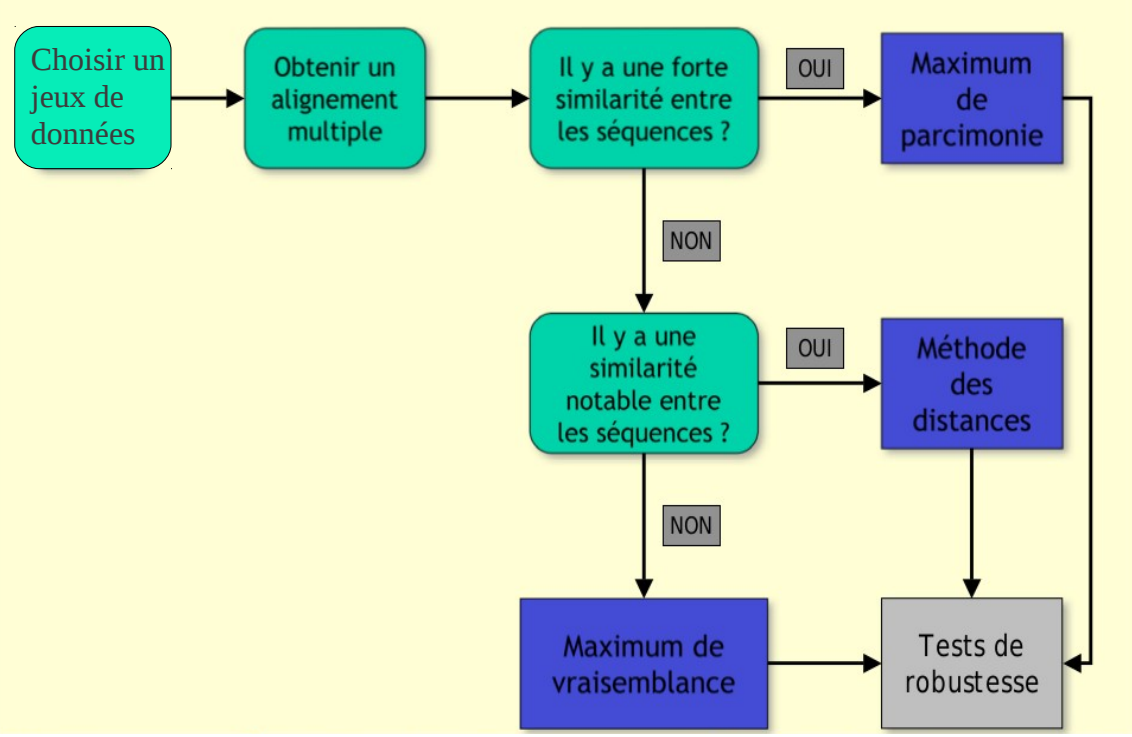

Figure 2 : Choisir la méthode d'analyse phylogénétique la mieux adaptée

La figure 2, soumet les méthodes d'inférences phylogénétiques à des tests de robustesses. On peut citer en exemple la méthode du bootstrap. Elle recrée un grand nombre de jeux de données par ré-échantillonnage numérique à partir du jeu de données réel. Puis elle construit un arbre phylogénétique pour chaque jeu de données simulé en utilisant l'une des méthodes de reconstruction phylogénétique et compte la proportion de ces arbres obtenus par bootstrap qui montrent un même regroupement phylogénétique (un même clade dans l'arbre). Cela permet d'avoir une notion de confiance sur une arête de l'arbre.

III.3. L'arbre phylogénétique :

Une fois que l'on a choisit la méthode adéquate à nos données, nous pouvons calculer ou estimer l'arbre phylogénétique. La définition d'un arbre phylogénétique est similaire à celle d'un arbre de la théorie des graphes. C'est un graphe connexe non cyclique, il est orienté car toutes les arêtes partent d'un sommet « ancien » vers un sommet plus « récent ». Le plus ancien de tous les sommets est la racine. Les nœuds internes correspondent ainsi à des gènes ou des espèces ancestraux et le nœud racine, à l'ancêtre commun de tous les gènes ou espèces. Les arêtes sont orientées du passé vers le présent et peuvent avoir une valeur en fonction du temps ou du nombre de mutations séparant les différents nœuds. Lorsqu'un arbre de gènes, *cf* Figure 3, est inféré à partir d'un ensemble de gènes appartenant à plusieurs espèces, ses nœuds internes correspondent à des événements de duplication ou de spéciation, éventuellement de transfert. Tandis que pour un arbre des espèces tous les nœuds correspondent à des événements de spéciations.

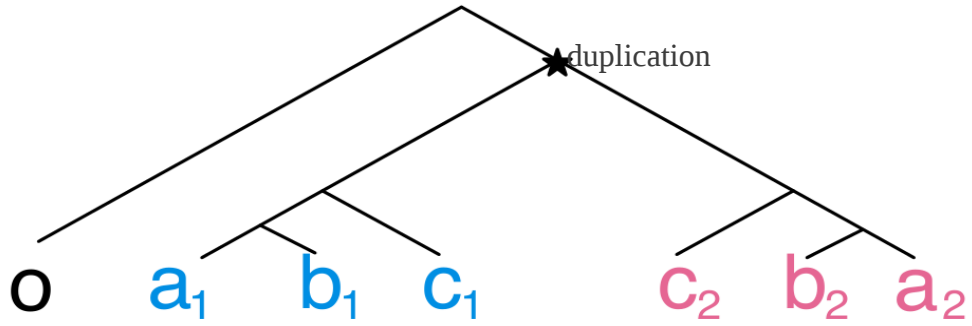

Figure 3 : Représentation d' un arbre phylogénétique, o représente le gène de l'espèce outgroup et a_1 correspond a l'espèce « a » et au gène 1 [2].

De plus, lorsque la phylogénie des espèces considérées est connue, il est possible d'établir une correspondance explicite entre les nœuds de l'arbre de gènes et les événements évolutifs, en "réconciliant" l'arbre de gènes avec l'arbre des espèces à l'aide d'une méthode appropriée.

Définition 3 : La Réconciliation :

Cela consiste à ajuster l'arbre de gènes dans l'arbre des espèces, et à en déduire une histoire d'événements évolutifs spécifiques aux gènes, la plus parcimonieuse possible, permettant d'expliquer la non-congruence (dissymétrie) éventuelle entre les deux arbres. Un arbre des espèces a pour feuille l'ensemble des espèces que l'on étudie.

III.4. Les relations entre gènes :

Il existe plusieurs types de relations ente les gènes. Pour ce projet nous avons décidé de nous concentrer sur deux types en particuliers, l'adjacence entre les gènes et les réseaux d'interactions protéine-protéine.

Définition 4 : L'adjacence des gènes :

L'adjacence signifie que les gènes ne sont séparés par aucun autre gène sur un même chromosome. Certains articles insèrent également la notion de distance entre les gènes dans la définition de l'adjacence.

Définition 5 : Les réseaux d'interaction protéine-protéine :

Une interaction protéine–protéine apparaît lorsque deux ou plusieurs protéines se lient entre elles, le plus souvent pour mener à bien leur fonction biologique. Les interactions protéine-protéine sont un aspect essentiel des processus biologiques. Elles sont fortement impliquées dans la formation de structures macromoléculaires, dans la signalisation, dans la régulation et dans les différentes voies métaboliques. Leur étude est donc cruciale pour la compréhension des réseaux d'interaction protéiques, but majeur dans l'étude des systèmes biologiques. Les interactions protéine-protéine ont un rôle conséquent dans l'induction de beaucoup d'états pathologiques et dans les processus importants pour la pathogenèse des infections bactériennes et virales [ML. Gervais, 2010].

IV. Présentation des articles :

Notre recherche bibliographique, nous a permis de sélectionner 3 articles intéressants, dont deux proposent un algorithme basé sur les synténies de gène, partie IV.1 page 9 et l'autre sur les réseaux d'interactions protéine-protéine, partie IV.2 page 17.
Le rôle principal des gènes est d'induire la production de protéines. Les gènes ne sont pas des entités indépendantes les unes des autres. Ils interagissent entre eux par l'intermédiaire des protéines qui vont activer ou inhiber leur activité. C'est pour cette raison qu'il est important de prendre en compte les notions de synténie et de réseau protéique.

IV.1. La synténie :

Le concept de synténie se définit en terme d'adjacences, d'α-adjacences ou d'intervalle. Soit un segment nucléotidique g={a,b,c,d} constitué de 4 gènes, alors l'adjacence gauche de b est a et l'adjacence droite de b est c, de la même manière {b} est une 1-adjacence droite, {b, c} sont des 2-adjacences droites et {b, c, d} sont des 3-adjacences droites de a.

IV.1.1. Inferring the evolutionary history of gene clusters from phylogenetic and gene order data – Mathieu. Lajoie, Denis Bertrand et Nadia El-Mabrouk (2010) :

IV.1.1.a : Introduction au modèle de duplication en tandem classique :

Les méthodes classiques d'inférence phylogénétique ne se préoccupent pas de l'ordre des gènes sur les chromosomes, pourtant cet ordre constitue une information précieuse qui permet d'améliorer l'inférence de l'histoire. Fitch en 1977 [Fitch W., 1977], propose un modèle de duplication en tandem pour inférer l'histoire évolutive des groupes de gènes répétés en tandem (GRT). Une duplication en tandem a pour effet de placer le segment d'ADN répliqué adjacent au segment d'origine et dans la même orientation que celui-ci. Ce modèle permet d'étudier l'évolution d'un cluster de GRT débutant avec un unique gène ancestral et se poursuivant par une séquence de duplications en tandem appelée histoire de duplication. Il permet également de prendre en compte l'ordre des gènes sur le chromosome. Avec l'ordre des gènes, l'arbre résultant est appelé

arbre de duplication.

IV.1.1.b : Présentation de DILTAG :

M. Lajoie présente dans sa thèse 3 publications, chacune d'entre elles présente une extension du modèle de Fitch. Dans [M. Lajoie, 2007] il intègre les inversions au modèle de duplications en tandem simple et propose un algorithme exact pouvant être appliqué à des familles multigéniques contenant des gènes dans les deux orientations transcriptionnelles. Dans [D. Bertrand, 2007] il généralise ce modèle pour permettre l'étude d'un ensemble de clusters orthologues dans plusieurs espèces. Ces deux extensions se limitent aux duplications simples (duplication d'un seul gène). Finalement, il propose l'algorithme DILTAG qui est une heuristique permettant d'inférer l'histoire évolutive d'un cluster de GRT en tenant compte d'un large spectre d'événements évolutifs pouvant impliquer plusieurs gènes à la fois : les duplications en tandem, les duplications inversées, les inversions et les délétions menant à des pertes de gènes. Voici un exemple de différents événements évolutifs, figure 4.

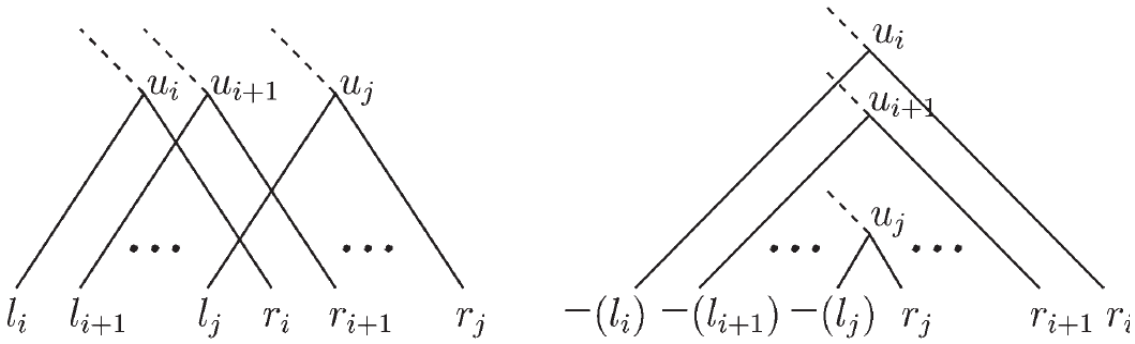

Figure 4: Deux types de duplications agissant sur une sous-séquence (u_i, u_{i+1},...,u_j) d'un arbre de gène ordonné. (À gauche) Une duplication en tandem (À droite) une duplication inversée [M. Lajoie et al., 2010].

IV.1.1.c : Définitions et méthodes employées dans DILTAG :

Un cluster de GRT est représenté par un arbre de gènes ordonné, dénoté (T, O), où T est un arbre de gènes (binaire et enraciné) représentant les relations ancestrales entre les gènes, et O=(v_1, . . . , v_n) est une permutation des feuilles de T correspondant à l'ordre des gènes sur le chromosome. Une permutation de *n* objets distincts rangés dans un certain ordre, correspond à un changement de l'ordre de succession de ces *n* objets. Une cerise de T est une paire de feuilles (g, d) séparée par un unique nœud appelé sa racine.

Définition 6 :

Soit (T, O) un arbre de gènes ordonné. Une duplication en tandem a pour effet de remplacer une sous-séquence (v_i , v_{i+1} , . . ., v_j) de O par une séquence de nouveaux éléments (g_i , g_{i+1} , . . . , g_j , d_i , d_{i+1} , . . ., d_j). De plus, chaque feuille de T étiquetée par v_x , pour i ≤ x ≤ j, est substituée par la « cerise » (g_x , d_x).

Définition 7 :

Une histoire évolutive est une suite d'arbres de gènes ordonnés H = ((T1 , O1), (T2, 02), . . . , (Tn, On)) telle que :

1. T1 est égal à un arbre v, constitué d'une unique feuille et O1 = (v).
2. Pour 1 ≤ k < n, (Tk+1 , Ok+1) peut être obtenu en effectuant soit une duplication en tandem ou inversée soit un événement d'inversion ou de délétion, sur (Tk , Ok).

Définition 8 :

Un arbre de gènes ordonné (T, O) est un arbre de duplication si et seulement si il existe une histoire de duplication H = ((T1, O1), (T2 , 02), . . ., (Tn , On)) telle que (Tn , On) = (T, O).

Les définitions 6, 7 et 8 sont reprises de [M. Lajoie, 2010].

Voici la représentation schématique de la construction d'un scénario évolutif, *cf* Figure 5.

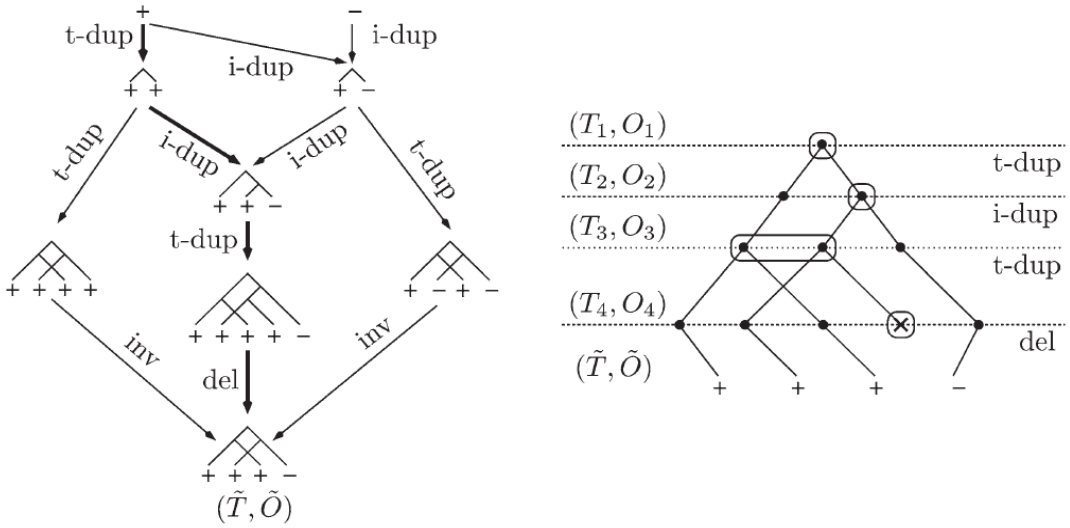

Figure 5 :(À gauche) Le graphe des histoires, les segments correspondent aux événements d'évolution connectant différents arbres de gènes ordonnés intermédiaires. Le label « inv » est pour inversion, « t-dup » pour duplication en tandem, « i-dup » pour duplication inversée et « del » pour délétion. (À droite) L'histoire évolutive correspondant à la voie définie par les segments en gras dans le graphe des histoires [M. Lajoie et al., 2010].

Pour chaque événement d'un type donné t agissant sur une sous-séquence (v_i, v_{i+1},...,v_j) de O, il associe un coût $C_t(n)=\alpha_t+(n \times \beta_t)$, où n=j-i+1 est la taille de l'événement et $\alpha_t>0$ et $\beta_t>0$ sont les constantes choisies pour refléter les probabilités de chaque type d'événement. Le coût de l'histoire évolutive est la somme des coûts associés avec ces événements.
Ainsi pour retrouver l'histoire évolutive optimale, on doit à travers le graphe des histoires, retrouver un chemin minimisant les coûts des événements évolutifs pour permettre d'obtenir à partir de (T1,O1), (T~,O~). Leur procédé se rapproche de la programmation dynamique, on résout des sous problèmes, en l'occurrence on construit un sous-arbre et on lui ajoute un événement spécifique à chaque étape jusqu'à retrouver l'histoire évolutive optimale.

IV.1.1.e : En résumé :

Les paramètres de leur algorithme sont, l'arbre de gène qui contient tous les gènes actuels comme feuille pour une seule espèce et l'ordre des gènes sur le chromosome. Il tente de

reconstruire l'histoire évolutive en partant d'un unique gène ancestral. Une histoire évolutive correspond à un ensemble d'arbre de gènes ordonnés ou (T1,O1) est un arbre avec une seule feuille et progressivement on lui ajoute différents événements pour le faire évoluer et le faire correspondre à l'arbre entré en paramètre. Ensuite ils choisissent le chemin le plus parcimonieux dans l'arbre pour connaître l'histoire optimale en partant de (T~, O~). Ils prennent donc en compte les relations d'adjacences des gènes et les types d'événements traités sont la duplication en tandem et inversée ainsi que les inversions et les duplications. Leur algorithme à une complexité en $O(n^3)$, il n'est pas en $O(n^2)$ qui est classiquement la complexité d'un algorithme de programmation dynamique car il initialise la procédure d'alignement dans les deux directions pour chaque cerise d'un groupe.

IV.1.2. DUPCAR : Reconstructing Contiguous Ancestral Regions with Duplications – Jian Ma et al. (2008) :

DUPCAR est une approche heuristique basée sur la méthode InferCar [J. Ma, 2006] qui incorpore les événements de duplications dans les prédictions de l'ordre ancestral des gènes. La méthode requiert pour tout gène « a » un arbre de gènes T_a représentant l'évolution de la famille formée de toutes les occurrences du gène « a » dans tout génome G, appartenant à l'ensemble des génomes modernes.

Définition 9 : CAR :

Chaîner de manière optimale des synténies ancestrales permet d'obtenir des régions ancestrales contiguës (CAR).

IV.1.2.a. Méthodologie de InferCar :

InferCar est une méthode basée sur les synténies. Elle considère les adjacences immédiates de gènes et prend en paramètre les informations d'adjacences entre les segments conservés des espèces modernes afin de déduire l'ordre des segments dans le génome ancestral. Ils encodent chaque gène avec un nombre, ainsi la représentation du chromosome est vue comme une permutation signée ou le signe correspond à l'orientation (brin) des gènes encodés. Par exemple : C = 1 -4 -3 5 2 représente un chromosome de 5 éléments génomiques dans lesquels les gènes 1, 5 et 2 sont sur le brin positif et les gènes 3 et 4 sont sur le brin négatif. Par contre ils ne prennent pas en compte les duplications c'est à dire, deux élémentaires avec la même valeur absolue. Chaque élément g_i dans un chromosome $C = g_1\ g_2\ \dots\ g_{i-1},\ g_i,\ g_{i+1}\dots\ g_n$ a un unique prédécesseur g_{i-1} et successeur g_{i+1} et g_1 a 0 prédécesseur et g_n a 0 successeur. Il suppose que la phylogénie est connue et que $\Sigma_A = \Sigma$, comme le contenu en gène est égal dans tous les génomes de Γ.

Abréviations utilisées :
G : Génome
Σ : Alphabet représentant les familles de gènes
Σ_G : Familles de gènes présentent dans le génome G, $\Sigma_G \subseteq \Sigma$
Γ : Ensemble de génomes modernes
F_a : Famille formée de toutes les occurrences du gène a dans tout $G \in \Gamma$
S : Phylogénie pour les espèces associées aux génomes de Γ

InferCar se réalise en 3 grandes étapes, on peut les résumer ainsi :

Étape 1 :

Pour chaque gène a ∈ Σ les adjacences ancestrales potentielles aux nœuds internes de S sont inférées par une méthode semblable à la parcimonie de Fitch. Soit AD(a,u) l'ensemble des adjacences droites potentielles de « a » à un nœud interne u, et soit v, w les fils gauche et droit respectivement de u. En faisant un parcours postfixe de S, cf Figure 6.

$$AD(a,u) = \begin{cases} AD(a,u) \text{ si u est une feuille,} \\ AD(a,v) \cup AD(a,w) \text{ si } AD(a,v) \text{ et } AD(a,w) \text{ sont disjoints,} \\ AD(a,v) \cap AD(a,w) \text{ autrement.} \end{cases}$$

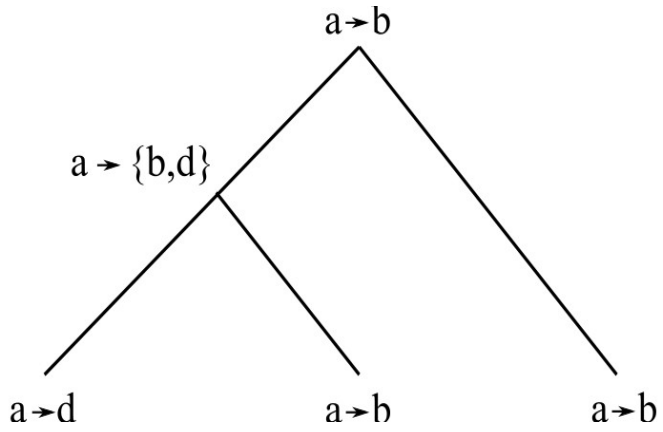

Figure 6 :Représentation de l'étape 2 de la méthode de Ma et al. pour déterminer les adjacences droites potentielles du gène « a » aux nœuds internes de S. a → b indique que b est l'adjacence droite de « a » [Y. Gagnon, 2012].

De façon similaire, l'ensemble des adjacences gauches est calculé. Ainsi nous obtenons deux graphes, l'un pour les prédécesseurs et l'autre représentant les successeurs. Le graphes des prédécesseurs et des successeurs ont les mêmes feuilles mais les nœuds internes sont généralement différents, bien qu'ils ont typiquement une partie commune. Ils réalisent par la suite l'intersection entre les deux graphes, ou ils prennent toutes les arêtes en commun entre les deux graphes. Dans le graphe obtenu, un élément peut être impliqué dans 3 sortes d'ambiguïtés, *cf* Figure 7.

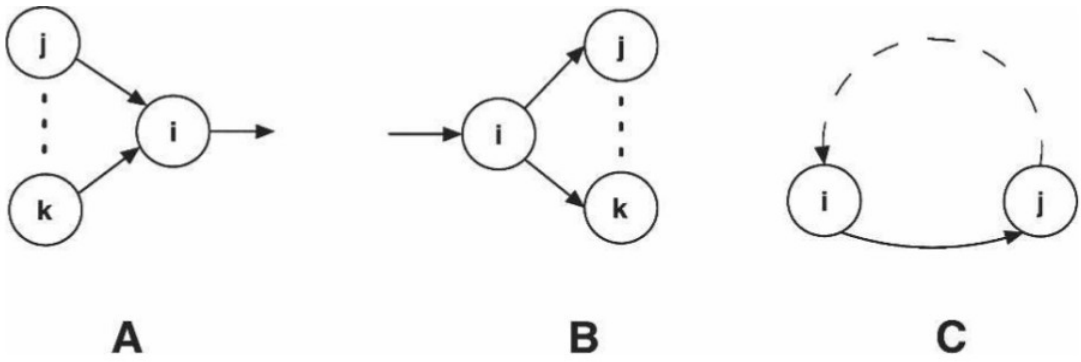

Figure 7 : 3 cas d'ambiguïtés potentielles. (A) i a plusieurs prédécesseurs possibles ; (B) i a plusieurs successeurs possibles ; (C) i forme un cycle avec j (circularisation) [J. Ma, 2006].

Si aucune de ces ambiguïtés est présente alors le graphe constitue l'ensemble des voix que couvre tous les éléments et fournit une structure de génome ancestral reconstruit.

Étape 2 :

Dans le cas où il y a ambiguïté sur les adjacences de a, un poids est attribué aux adjacences de « a » afin de les départager. Le poids p_u (a, b) d'une adjacence entre deux gènes a et b à un nœud u est également calculé par un parcours postfixe de S. Plus précisément,

$$p_u(a,b) = \begin{cases} 1 & \text{si } u \text{ est une feuille et l'adjacence (a,b) existe à } u \\ 0 & \text{si } u \text{ est une feuille et l'adjacence (a,b) n'existe pas à } u \\ \frac{L_v p_v(a,b) + L_w p_w(a,b)}{L_v + L_w} & \text{autrement} \end{cases}$$

où L_v , L_w sont les longueurs de branches entre le nœud u et ses deux fils (la longueur de branche représente la distance évolutive ; InferCar ne la calcule pas et doit être fournie avec l'arbre d'espèce en entrée). Une schématisation et un exemple de la procédure sont présentées sur la figure 8.

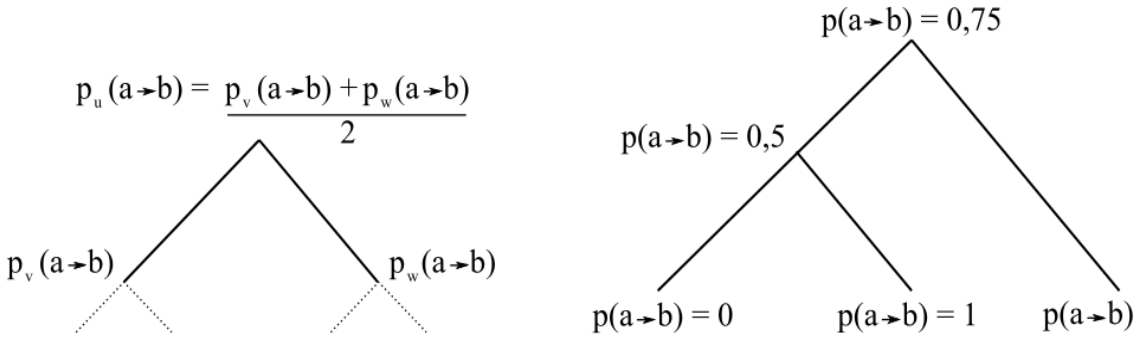

*Figure 8 : (*Gauche) : Schématisation de la récurrence permettant de calculer par programmation dynamique le score d'une adjacence selon la méthode InferCar, étant donné un sous-arbre de S enraciné au nœud u. (Droite) : Valeurs du poids de l'adjacence (a, b), étant donné l'ensemble des adjacences droites potentielles de « a » aux nœuds internes à la figure 6 et en suivant la procédure décrite dans le texte. Les longueurs de branches sont fixées à une valeur de 1 [Y. Gagnon, 2012].

Les étapes 1 et 2 sont les étapes reprises de l'algorithme de Fitch avec la phase ascendante dans un premier temps et la phase descendante dans un deuxième temps.

Étape 3 :

Enfin, une heuristique gloutonne permet de construire les régions ancestrales contiguës (morceau de chromosome ; CAR), en ajoutant dans l'ordre décroissant, les adjacences de poids élevés et rejetant celles de poids plus faible causant des ambiguïtés. Plus spécifiquement, la solution ne doit contenir aucun des trois cas suivant : (A) un gène possède plus d'une adjacence gauche, (B) un gène possède plus d'une adjacence droite, (C) le CAR est circularisé par l'adjacence ajoutée (*cf* Figure 7). Si l'adjacence ajoutée crée l'un de ces trois cas, elle est ignorée et la prochaine adjacence dans la liste des adjacences potentielle est considérée.

IV.1.2.b. Méthodologie de DUPCAR :

DUPCAR est donc l'extension de la méthode ci-dessus. Le principal inconvénient d'InferCar est qu'elle ne traite pas les duplications. Pourtant les duplications (segmentales et en tandem) ont un impact important sur l'évolution du génome. La méthode prend en paramètres, un ensemble de génomes modernes G, un arbre des espèces T qui décrit la phylogénie de génomes modernes et un ensemble d'arbre de gènes Ta_i pour chaque famille de gène a_i qui définit les relations entre tous les gènes dans une famille.

La première étape consiste en une réconciliation des arbres de gènes avec l'arbre des espèces. Une réconciliation entre un arbre de gène T_a et un arbre d'espèce S permet d'expliquer les divergences entre les topologies de Ta et de S, *cf* Figure 9. Les feuilles de T sont associées aux feuilles de S correspondant à leur espèce de provenance. Les nœuds internes de T_a représentants des spéciations sont associés aux nœuds internes correspondants à la même spéciation dans S. Les nœuds internes de T_a représentant des duplications sont associés à la branche de S où a eu lieu la duplication. Le processus de réconciliation permet de connaître la multiplicité des gènes aux nœuds internes de l'arbre des espèces, puisque l'on sait par la réconciliation où les événements de duplications ont eu lieu dans S.

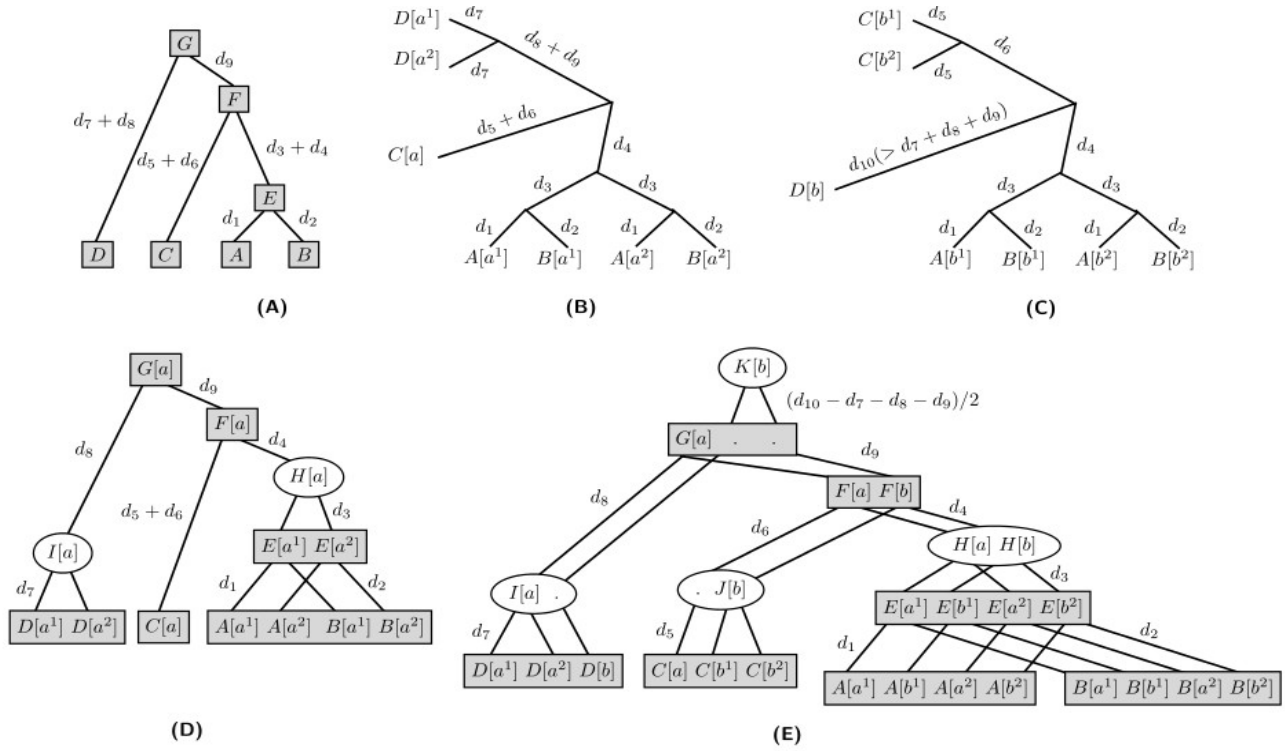

Figure 9 : (A) Un arbre des espèces T. (B) Arbre de gène T_a. (C) Arbre de gène T_b. (D) Arbre réconcilié R après que T_a soit fusionné. (E) Arbre réconcilié après que T_a et T_b soit fusionnés.

Une fois la réconciliation terminée, ils applique leur fonction AugmentGeneTree(R, T_a) avec R l'arbre réconcilié, cette fonction va additionner des nœuds le long des branches de chaque arbre de gènes, représentant les formes intermédiaires qui sont présumées avoir existé mais qui ne sont pas apparues dans l'arbre des gènes original, *cf* Figure 10.

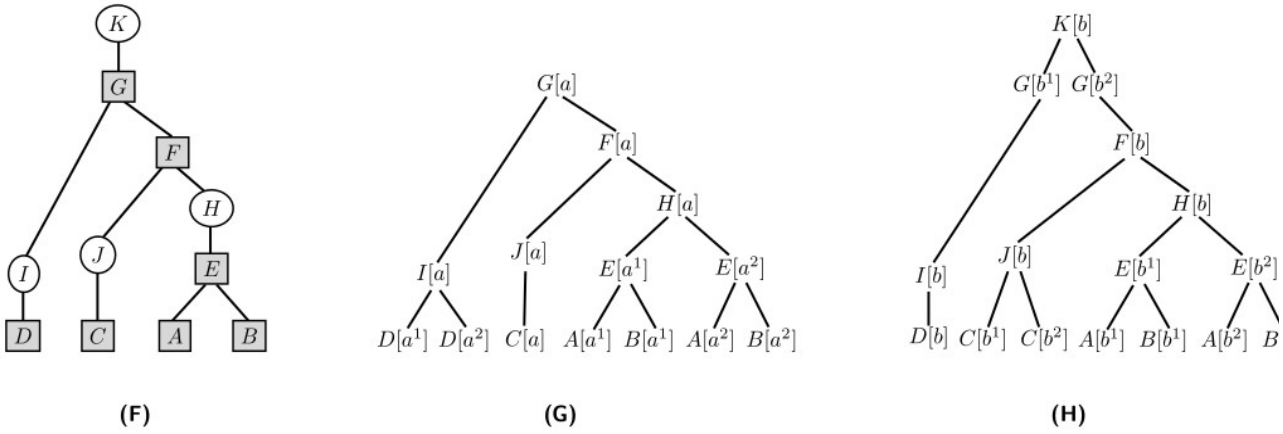

Figure 10 : (F) Version simplifiée de R, ou I, J, K et H montre quatre événements de duplication, K est une ancienne duplication arrivée avant G. (G) Arbre de gènes augmenté T_a*. (H) Arbre de gènes augmenté T_b*.

Après avoir obtenu l'arbre des gènes réconcilié R et les arbres de gènes augmentés pour toutes les familles de gènes, le but est de déterminer une liste contenant l'ordre des gènes qui se

rapprochent de la structure du génome ancestral α. La reconstruction des adjacences ancestrales est réalisée en deux phases :

– La première étape est une procédure *« bottom-up fashion »* des feuilles vers la racine. Voici l'idée générale. Pour chaque nœud x dans l'arbre des gènes augmenté, soit g le génome contenant x et soit u et v ses deux enfants dans l'arbre. L'algorithme calcule l'ensemble des prédécesseurs selon les règles suivantes : Si x est une feuille alors P(x) (ensemble des prédécesseurs de x) se compose d'un unique prédécesseur. Autrement, P(x) est égal à l'intersection ou à l'union de $A_g(P(u))$ et $A_g(P(v))$ (« A » correspond à l'ancêtre direct de P(u) ou P(v)), selon que $A_g(P(u))$ et $A_g(P(v))$ sont disjoints ou pas. De la même manière l'algorithme déduit l'ensemble des successeurs.

– Dans la deuxième phase, il propage l'information vers le bas de l'arbre, *« Top-down fashion »*. Pour chaque nœud x dans l'arbre des gènes augmenté, soit w le parent de x dans l'arbre. Il propage P(w) vers le bas de l'arbre pour ajuster P(x). S(x) (ensemble des successeurs de x) est ajusté de manière similaire.

Ensuite ils vont construire l'arbre des adjacences ancestrales, sur la base des prédécesseurs et des successeurs calculés. Cet arbre indique des relations de prédécesseurs et de successeurs cohérents. Cependant il peut tout de même présenter des cas ambiguës, cf Figure 7. À ce moment il utilise l'algorithme glouton *« FindCAR »* permettant de construire les régions ancestrales contiguës, en ajoutant dans l'ordre décroissant, les adjacences de poids élevés et rejetant celles de poids plus faible causant des ambiguïtés. Cette étape est similaire à InferCar.

Leur algorithme s'exécute en O(mn) ou m est le nombre de nœud dans l'arbre de gènes augmenté et n est le nombre total de feuille dans l'arbre de gènes augmenté.

IV.2. Les réseaux d'interaction protéiques :

Les réseaux d'interactions proteine-proteine d'organismes divergents peuvent être comparés, afin d'expliquer comment les différents événements, comme la duplication de gène ou bien la délétion, ont façonnés l'évolution des structures contemporaines. Actuellement, les techniques n'ont pas de contexte phylogénétique, nécessaire pour la reconstruction d'histoire évolutive de réseaux [JW. Pinney, 2007].

IV.2.1. Reconstruction of ancestral protein interaction networks for the bZIP transcription factors – JW. Pinney, GD. Amoutzias, M. Rattray, DL. Robertson(2007) :

Dans cet article, les auteurs utilisent un modèle d'évolution de réseau pour reconstruire le réseau d'interaction entre les facteurs de transcription bZIP (basic zipper) chez les ancêtres de cordées. Leur méthode suppose qu'il est possible à la fois de reconstruire une phylogénie fiable pour une famille de gène d'intérêt et de réconcilier cette phylogénie avec l'arbre des espèces qui est connu. Ils débutent par la construction de l'arbre des gènes, réalisé grâce à l'alignement de séquences protéiques bZip de quatre cordés, puis ils réconcilient cet arbre, avec l'arbre des espèces. Donc chaque nœud est identifié comme un événement de duplication ou de spéciation. Ensuite, ils considèrent l'ensemble possible des homo- et hétéro-dimères entre toutes les paires de protéines bZip et la façon dont elles sont liées par la duplication des gènes, pour transformer l'arbre des gènes en une représentation arborescente des interactions. Chaque nœud dans ce nouvel arbre représente une interaction potentielle entre une paire de protéine.

Définition 10 : Arbre des interactions :

Un arbre des interactions est raciné, orienté et ses nœuds sont de degré 3 au maximum. Il décrit l'évolution des interactions de protéine. Les nœuds représentent les interactions protéiques possibles et les branches représentent les effets de la duplication, spéciation et de la perte des protéines sur l'évolution de l'interaction protéine-protéine [A. Rajaraman , 2009].

Un modèle graphique probabiliste pour l'évolution du réseau d'interaction protéine-protéine est ensuite construit, il est basé directement sur l'arbre d'interaction, *cf* Figure 11. Dans le modèle graphique probabiliste un ensemble de feuilles supplémentaires est utilisé, indiquant la présence ou l'absence des interactions dans l'espèce existante, cette information est généralement calculé en utilisant les techniques d'alignement de séquences, dans le papier ils ont utilisé le logiciel de Fong et Singh [JH. Fong , 2004]. Chaque nœud interne est également binaire, pour représenter l'absence ou la présence d'une interaction potentielle. Les duplications multiples de gènes survenant entre les espèces sont supposées avoir lieu dans l'ordre indiqué en considérant la longueur relative des branches.

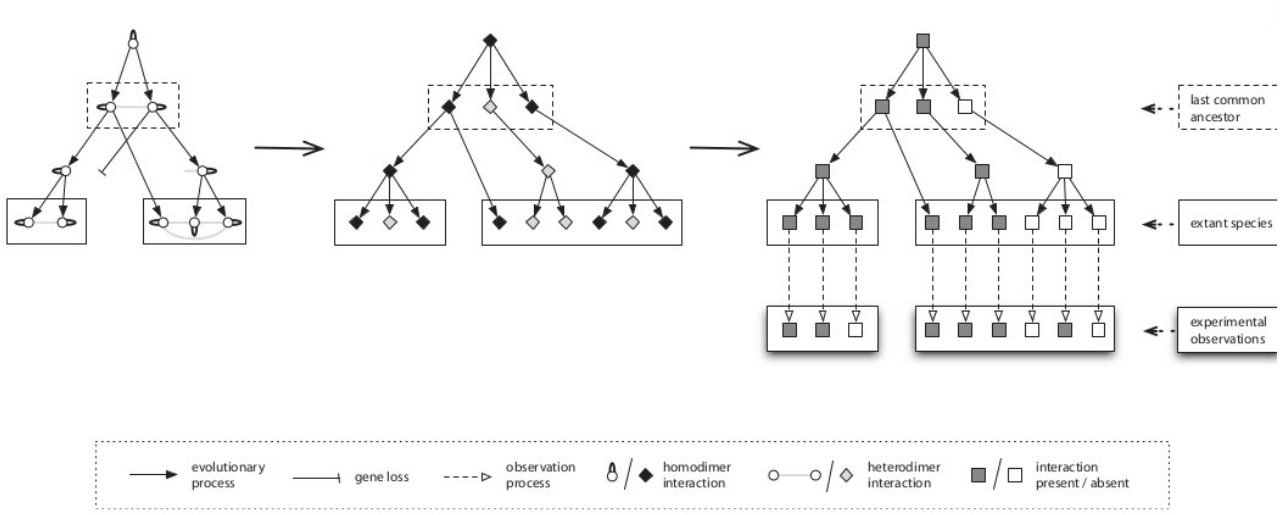

Figure 11 : (À gauche) L'arbre des gènes réconcilié. (Au milieu) L'arbre des interactions. (À droite) Le modèle graphique probabiliste [JW Pinney, 2007].

La longueur de branche est utilisés comme un paramètre pour calculer la probabilité de gain et de perte de chaque segment. Ce modèle peut être ensuite utilisé pour déduire la probabilité de la force d'interaction aux nœuds internes de l'arbre des interactions donc, pour reconstruire un réseau d'interaction de protéine pour chaque espèce ancestrale, *cf* Figure 12. L'histoire évolutive du réseau d'interaction de bZip chez les cordés est déduit grâce à leur modèle probabiliste.

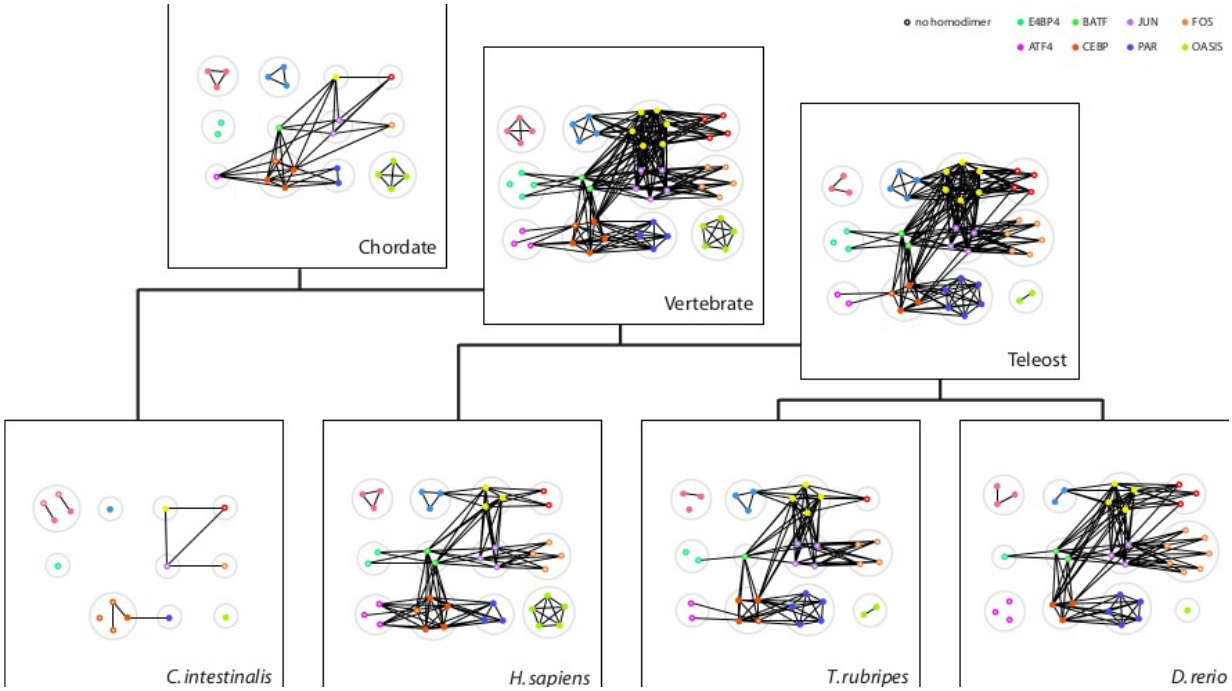

Figure 12 : Représentation de l'histoire évolutive des interactions. Chaque protéine est représentée comme un nœud coloré et un segment est dessiné entre deux protéines si elles ont >50 % de chance de partager une interaction forte [JW Pinney, 2007].

Donc cet article tente de reconstruire les interactions des protéiques ancestrales de la famille bZIP en utilisant un réseau bayésien modélisé par l'arbre d'interaction.

V. Conclusion :

Un génome n'est plus seulement un ensemble de gène. C'est aussi un ensemble structuré en chromosomes et organisé en réseaux d'interactions. Les études prenant en compte des informations phylogénétiques et des informations de structures ne sont pas très nombreuses [E. Tannier, 2011]. Depuis le programme défini il y a plus de dix ans par Sankoff & El-Mabrouk [D. Sankoff, 2000], les succès sont toujours rares. Lajoie et al. [M. Lajoie, 2007], Bertrand et al. [D. Bertrand, 2008] et Lajoie et al. [M. Lajoie, 2010] ont construit des modèles d'évolution pour des groupes de gènes se dupliquant en tandem. Ma et al. [J. Ma, 2008] utilisent une méthode de parcimonie avec des phylogénies de gènes très contraintes, interdisant les pertes par exemple. On peut également citer Bertrand et al. (2010) ou la thèse de Muffato et al. [M. Muffato, 2010] qui utilisent des phylogénie dans un processus d'inférence d'adjacences ancestrales, sans modéliser la complexité des histoires de gènes. Dans ce projet nous avons traité deux types d'informations supplémentaires, les synténies et les réseaux protéiques. Mais d'autres informations peuvent également être couplées aux données phylogénétiques, on peut par exemple citer les informations sur les réseaux métaboliques. Le tableau 1, ci-dessous est une comparaison entre les algorithmes vus le long de ce rapport.

	DILTAG	DUPCAR	Réseaux d'interaction protéine-protéine
Algorithme exact ou heuristique	- Heuristique	- Heuristique	- Modèle graphique probabiliste construit avec Bayes Net Toolbox → méthode heuristique de recherche de structures dans l'espace des solutions
Paramètres	- Arbres de gènes - Ordre des gènes sur le chromosome	- Ensemble de génomes modernes - Arbres de gènes pour chaque famille de gènes - Arbre des espèces	- Séquences protéiques de la région « Leucine *Zipper* » de toutes les espèces - Arbre des espèces
Méthodes employées	- Parcimonie de Fitch	- Parcimonie de Fitch	- Construction d'un réseau bayésien modélisé par l'arbre d'interaction
Sortie de l'algorithme	- Histoire évolutive optimale menant a nos adjacences actuelles	- Assemblage d'une région ancestrale contiguë	- Histoires évolutives du réseau d'interactions de bZip
Complexité	$O(n^3)$	$O(n^2)$	Non mentionné Ce n'est pas vraiment un algorithme qui est décrit mais une méthodologie utilisant plusieurs outils
Événements évolutifs prient en compte	- Duplications en tandem - Duplications inversées - Inversions - Pertes de gènes	- Duplications segmentales - Duplications en tandem	- Gain des interactions - Perte des interactions - Multiplication des interactions après la duplication de gène - Perte de gènes
Utilisation de la réconciliation	Non	Oui	Oui
Disponibilité du logiciel	Diponible on line à l'adresse http://www-lbit.iro.umontreal.ca/DILTAG/	Téléchargement de InferCar à l'adresse http://compbio.bioen.illinois.edu/software.html	/

*Tableau 1 : Synthèse de la c*omparaison des algorithmes

Chacune des méthodes sont, ou utilisent une heuristique. DUPCAR et DILTAG sont deux approches basées sur la programmation dynamique. Par contre DILTAG prend en compte des événements évolutifs supplémentaires par rapport à DUPCAR et l'étape de réconciliation de DUPCAR augmente le nombre de données en paramètre par rapport a DILTAG. On peut également remarquer que les transferts ne sont pas prient en compte dans aucun des trois algorithmes. Un travail intéressant serait d'implémenter une méthode pour prendre en compte cet événement dans nos phylogénies.

Références :

S. Bérard , C. Gallien , B. Boussau , GJ. Szöllosi , V. Daubin et E. Tannier : Evolution of gene neighborhoods within reconciled phylogenies , Vol. 28 ECCB 2012, pages i382–i388 doi:10.1093/bioinformatics/bts374 , 2012.

D. Bertrand, M. Lajoie et N. El-Mabrouk : Inferring ancestral gene orders for a family of tandemly arrayed genes. Journal of Computational Biology, 15 (8):1063-1077, 2008.

O. Elemento, O. Gascuel et M.-P. Lefranc : Reconstructing the duplication history of tandemly repeated genes. Molecular Biology and Evolution, 19:278- 288, 2002.

W. Fitch : Phylogenies constrained by cross-over process as illustrated by human hemoglobins and a thirteen-cycle, eleven amino-acid repeat in human apolipoprotein A-I. Genetics, 86:623-644, 1977.

W. Fitch : Toward defining the course of evolution ; minimum change for a specified tree topology. Sys. Zool., 20, 406-416, 1971.

JH. Fong, AE. Keating, M. Singh : Predicting specificity in bZIP coiled-coil protein interactions Genome Biol 5:R11, 2004.

Y. Gagnon. Algorithme pour la reconstruction de génomes ancestraux, 2012.

ML. Gervais : Étude des interactions proteine-proteine par double hybride bactérien : applications en agro-alimentaire et en santé, 2010.

M. Lajoie : Approches algorithmiques pour l'inférence d'histoires de duplication en tandem avec inversions et délétions pour des familles multigéniques , 2010.

M. Lajoie, D. Bertrand et N. El-Mabrouk : Inferring the evolutionary history of gene clusters from phylogenetic and gene order data. Molecular Biology and Evolution, p. 761-772, 2010.

M. Lajoie, D. Bertrand et N. El-Mabrouk et O. Gascuel : Duplication and inversion history of a tandemly repeated gene family. Journal of Computational Biology, p. 462-478, 2007.

J. Ma, L. Zhang, B.B. Suh, B.J. Raney, R.C. Burhans, W.J. Kent, M. Blanchette, D. Haussler et W. Miller. Reconstructing contiguous regions of an ancestral genome. Genome Research, 16:1557-1565, 2006.

M. Muffato : Reconstruction de génomes ancestraux chez les vertébrés , 2010.

JW. Pinney, GD. Amoutzias, M. Rattray, DL. Robertson : Reconstruiction of ancestral protein interaction networks for the bZIP transcription factors, pnas.0706339104, 2007.

A. Rajaraman : Inference of ancestral protein-protein interactions using methods from algebraic statistics, 2009.

N. Saitou, M. Nei : The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, volume 4, issue 4, pp. 406-425, July 1987.

D. Sankoff and N. El-Mabrouk : Minimal mutation trees of sequences . SIAM J. Appl. Math, 28, 35, 1975.

D. Sankoff & N. El-Mabrouk : Duplication, rearrangement and reconciliation. In Comparative Genomics : Empirical and Analytical Approaches to Gene Order Dynamics, Map alignment and the Evolution of Gene Families (D. Sankoff & J. H. Nadeau, eds.), vol. 1 of Computational Biology. Kluwer Academic Press, 2000.

E. Tannier : Évolution combinatoire, Algorithme des chromosomes, 2011.

[1] http://modernmodelorganism.wordpress.com/2011/02/15/creating-phylogenetic-trees/

[2] Cours d'introduction à la phylogénie par Céline Scornavacca

Annexe :

- Grille de lecture utilisée pour trier les articles susceptibles de nous intéresser :

Titre
Auteurs et date de parution
Critères que l'algorithme présenté cherche à optimiser
Description brève de la méthode employée dans l'article
Paramètres de la méthode
Types d'interactions prient en compte (pertes de gènes, duplications...)
Jeu de données sur lequel l'algorithme a été testé
Implémentation disponible (y/n)

- Références de quelques articles passés en revus pendant la phase de trie :

J. Dutkowski et al. : Phylogeny-guided interaction mapping in seven eukaryotes, 2009.

L. Goodstadt et al. : Phylogenetic reconstruction of orthology, paralogy and conserved synteny for dog and Human, 2006.

R. Patro et al. : Parsimonious reconstruction of network evolution, 2011.

F. Pazos et al. : Similarity of phylogenetic trees as indicator of protein-protein interaction, 2001.

O. Tremblay Savard et al. : Evolution of orthologous tandemly arrayed gene clusters, 2011.

I. Wapinski et al. : Automatic genome-wide reconstruction of phylogenectic gene trees, 2007.

X. Zang et al. : B.M.E : Boosting the performance of inference algorithms for transcriptional regulatory networks using a phylogenetic approach, 2008.

X. Zang et al. : Rafining transcriptional regulatory network evolutionary models and gene histories, 2010.