Recherche de motifs dans les arbres phylogénétiques

Pattern matching in phylogenetic trees

Jean-François Dufayard [†] Laurent Duret [‡] Guy Perrière [‡] François Rechenmann [†]

[†] INRIA Rhône-Alpes, 655 avenue de l'Europe, F-38334 Montbonnot, Saint Ismier Cedex
[‡] Laboratoire BBE - UMR CNRS 5558, 16 rue Raphaël Dubois, F-69622 Villeurbanne Cedex

Courriel : {Jean-Francois.Dufayard,Francois.Rechenmann}@inrialpes.fr, {duret,perriere}@biomserv.univ-lyon1.fr

Résumé

Les bases de données de familles de gènes homologues, telles qu'HOBACGEN ou HOVERGEN, contiennent des dizaines de milliers d'arbres phylogénétiques. Jusqu'à présent, la seule possibilité pour retrouver des familles en fonction de critères phylogénétiques (c'est-à-dire en fonction de la topologie des arbres) était de parcourir manuellement ces bases. Du fait de leur taille, il s'agit d'un travail long et sensible aux erreurs de saisie. Afin de répondre de façon automatique à ce type de requêtes, nous avons développé un algorithme de recherche de motifs non-ordonnés dans les arbres. La méthode prend en compte de nombreux paramètres, tels que les nœuds de duplication, les incertitudes de topologie et les taxons de haut niveau. L'algorithme est intégré à l'interface FAMFETCH, qui permet d'accéder aux différentes bases de familles de gènes développées au sein du Pôle Bioinformatique Lyonnais (HOBACGEN, HOVERGEN, NureBase, RTKdb). Un éditeur graphique permet de dessiner un motif (en spécifiant des contraintes sur les branches), dont toutes les occurrences en tant que sous-arbre seront recherchées dans la base d'arbres phylogénétiques choisie. Il est ainsi possible, en décrivant le motif approprié, de rechercher toutes les familles de gènes orthologues dont les membres proviennent de taxons donnés. Il est aussi possible de rechercher des familles de gènes paralogues, et de contraindre la date relative de la duplication supposée.

Mots-clés : *phylogénie, arbres, motifs d'arbres.*

Abstract

Phylogenetic tree databases, such as HOBACGEN or HOVERGEN, are often explored manually in order to retrieve genes using phylogenetic criteria. This type of database may contain several thousands of trees, so this search process is time consuming and error prone. An algorithm for unordered tree pattern matching has been developed, in order to automatically solve this type of request. This method takes into account numerous parameters such as duplication nodes, unresolved topologies and high-level taxa. The algorithm is integrated into the interface FAMFETCH, which permits to access the gene family databases developed at the Pôle Bioinformatique Lyonnais *(HOBACGEN, HOVERGEN, NureBase, RTKdb). A graphic editor allows to draw a pattern (by specifying constraints on branches), and each of its occurrences as sub-tree will be retrieved in the chosen phylogenetic tree database. So, it is possible to search each orthologous gene families defined by given taxa, simply by editing the appropriate pattern. It is also possible search paralogous gene families, and to force the relative date of the supposed duplication.*

Keywords: *phylogeny, trees, tree patterns.*

1 Introduction

Les bases d'arbres phylogénétiques, telles que HOVERGEN [3] ou HOBACGEN [10], contiennent des grandes quantités de données. Le nombre important d'arbres et la taille de ceux-ci rendent l'exploration manuelle de ces bases extrêmement difficile. Il n'existe aucun outil adapté à la recherche d'information dans ce type de structure et extraire les gènes potentiellement intéressants pour la biologie, selon une requête précise, est un travail long et très sensible à l'erreur humaine.

Le type de question qu'est susceptible de se poser un utilisateur biologiste peut se formuler comme suit : « Je désire retrouver trois gènes homologues : un chez l'homme, un chez le rat et un chez la souris. Je souhaite que le gène du rat et le gène de la souris soient orthologues, et que, outre leur ancêtre commun, aucun nœud qui les relie ne soit une duplication. Je désire que le gène de l'homme soit paralogue aux deux premiers, et que la duplication à l'origine de la paralogie se soit passée après l'avènement des mammifères. »

Une telle requête peut être traduite sous forme d'un arbre (fig. 1), assimilable à un motif. Il possède les mêmes caractéristiques de base qu'un arbre phylogénétique, mais il apparaît évident que ces dernières ne suffisent pas à

formuler précisément les paramètres de la requête. Le motif recherché est plus raffiné et ses nœuds et arcs sont décorés de diverses contraintes.

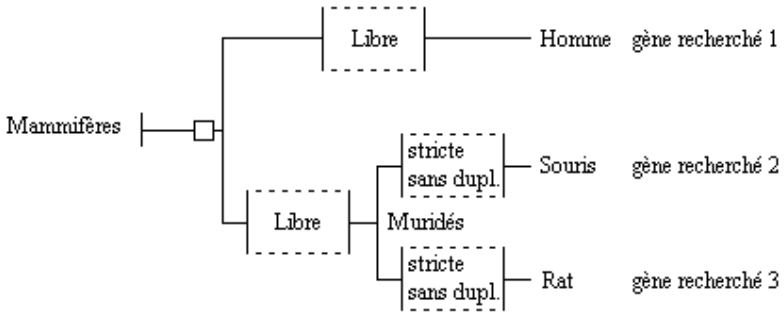

FIG. 1 – Un exemple de motif susceptible d'être recherché dans une base de données. Il s'agit de trouver un ou plusieurs arbres possédant un gène du rat, un gène de la souris et un gène de l'homme. Le gène du rat et celui de la souris doivent être strictement orthologues entre eux, c'est-à-dire sans aucun nœud de duplication entre leur ancêtre commun et eux. L'ancêtre commun aux trois gènes, placé avant le nœud d'origine des mammifères, doit être une duplication (carré blanc sur le motif).

2 La recherche de motifs non-ordonnés dans les arbres

Le problème de la recherche de motifs est un problème de comparaison d'arbres. Il s'agit, intuitivement, de déterminer si un arbre appelé motif est inclus dans un second arbre appelé arbre cible. La définition de l'inclusion varie en fonction du sous problème considéré.
Rechercher un motif revient à déterminer s'il existe des nœuds et feuilles étiquetés de la même façon et avec les mêmes liens de parenté dans un arbre cible donné. On constate que sa présence peut être détectée sans pour autant que les liens de parenté soient directs : dans l'exemple de la figure 2, les arcs du motif, symbolisés en les surlignant sur l'arbre cible, passent par des nœuds qui ne sont pas inclus dans le motif, tels que B ou C ; ils sont appelés nœuds intermédiaires.

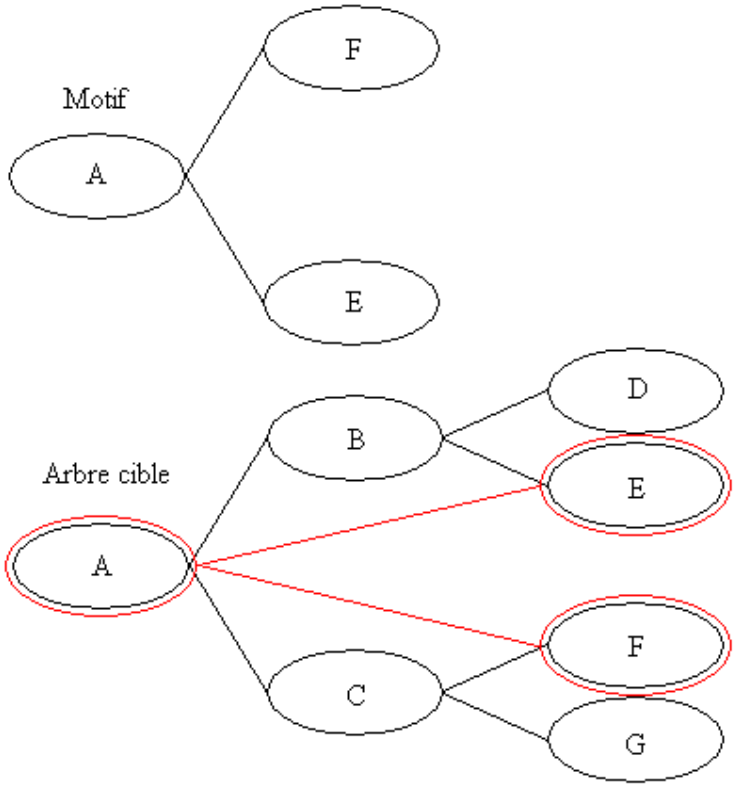

FIG. 2 – Le motif est présent dans l'arbre cible selon les contraintes du problème de la recherche de motifs non-ordonnés, ou *unordered tree pattern matching*. En effet, ni les nœuds intermédiaires (B, C), ni l'ordre des fils du motif (E en bas et F en haut) ne sont pris en compte.

Il existe deux sous problèmes au problème de la recherche de motifs dans les arbres.
La recherche de motifs ordonnés (ou *ordered tree pattern matching*), en plus de prendre en compte les liens de parentés, respecte aussi l'ordre des fils du motif. Dans l'exemple ci-dessus, le motif n'est pas présent dans l'arbre cible selon les contraintes de ce sous-problème, car, selon l'ordre vertical des nœuds sur la figure, E est au-dessus de F sur l'arbre cible, alors qu'il est en dessous sur le motif. Par contre, la recherche de motifs non-ordonnés, dans l'exemple de la figure 2, ne tient pas compte de l'ordre des fils du motif. C'est un problème sensiblement plus difficile : sa complexité n'est pas polynomiale.
La recherche de motifs dans les arbres phylogénétiques relève, sans conteste possible, de la recherche de motifs non-ordonnés. En effet, l'ordre des fils d'un arbre phylogénétique n'est pas informatif : il est totalement dépendant de la méthode de reconstruction de l'arbre à partir de sa matrice de distances, et indépendant des séquences elles-mêmes.
Il est possible de formaliser le problème de la recherche de motifs non-ordonnés dans les arbres comme suit [6,7]:
Soit les arborescences A et M :
A = (V, E, racine(A)) où V est l'ensemble des nœuds de A et E l'ensemble de ses arcs.
M = (W, F, racine(M)) où W est l'ensemble des nœuds de M et F l'ensemble de ses arcs
La question qui est posée est de savoir si M est un sous-arbre de A, M étant assimilé au motif et A à l'arbre cible.
M est un motif de A si et seulement s'il existe une fonction injective f telle que :

i) $f : w \in W \rightarrow v \in V$, c'est-à-dire que pour tout nœud de M est associé un nœud de A.
ii) f(u) = f(v) si et seulement si u = v
iii) étiquette(u) = étiquette(f(u)), c'est-à-dire que u et f(u) sont équivalents.
iv) u est un ancêtre de v si et seulement si f(u) est un ancêtre de f(v).

La recherche de motifs non-ordonnés est dans la classe de complexité des problèmes NP-complets (voir [6,7]). Néanmoins, en tenant compte des caractéristiques particulières des arbres phylogénétiques, il est possible d'effectuer ce type de recherche dans des temps tout à fait raisonnables.

3 Application aux arbres phylogénétiques

La recherche de motifs dans les arbres a été souvent appliquée à des données différentes des arbres phylogénétique. Mais seuls des arbres ordonnés sont explorés dans ces problématiques. Les méthodes utilisées ne peuvent donc pas être appliquées aux arbres phylogénétiques [1,4,7].
De plus, le problème générique de la recherche de motifs non-ordonnés ne permet pas de répondre à des requêtes intéressantes pour l'utilisateur biologiste. Aussi, afin d'appliquer ces méthodes aux arbres phylogénétiques, de nombreuses possibilités ont été ajoutées pour enrichir le concept de motif voir [2].

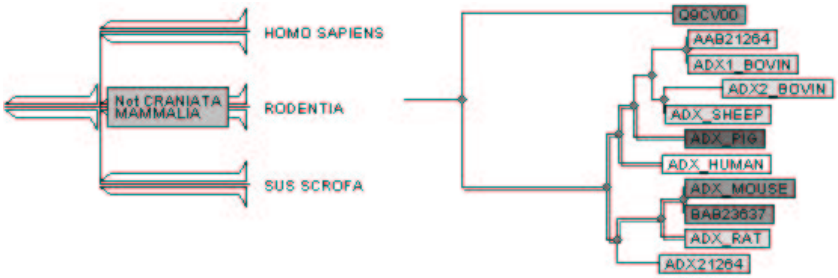

FIG. 3 – Le motif de gauche est destiné à rechercher l'ensemble des familles de gènes contenant un membre chez l'homme, un membre chez un rongeur quelconque, et un membre chez le porc (*Sus scrofa*). Contrairement aux arbres phylogénétiques, ce motif n'est pas binaire. Cela signifie que n'importe quelle topologie binaire sera tolérée dans l'arbre cible, figuré à droite. Ici, se sont les gènes issus de rongeurs qui sont à l'extérieur du motif (surligné).

La recherche de motifs non-ordonnés est bien adaptée aux topologies mal résolues sur les arbres phylogénétiques (fig. 3). Lorsque l'on recherche une famille de gènes dont on ne souhaite pas contraindre toute ou partie de la topologie, il est possible d'y placer un nœud non-binaire, ou râteau. Par exemple, en recherchant le motif ci-dessus dont la racine possède trois fils, on tolère n'importe quelle topologie binaire sur les arbres cible. Ici, ce sont les gènes issus de rongeurs qui divergent en premier dans l'arbre de droite.
Il est possible de placer des contraintes sur les taxons tolérés à divers points d'un motif. Dans l'exemple de la figure 3, ces contraintes servent à limiter les espèces représentées dans la partie de l'arbre cible contenant le motif. La partie de l'arbre sous la racine du motif est contrainte à ne contenir aucun gène de crâné (les crânés sont la racine de l'arbre des espèces contenues dans HOVERGEN), si ce n'est des gènes de mammifères.

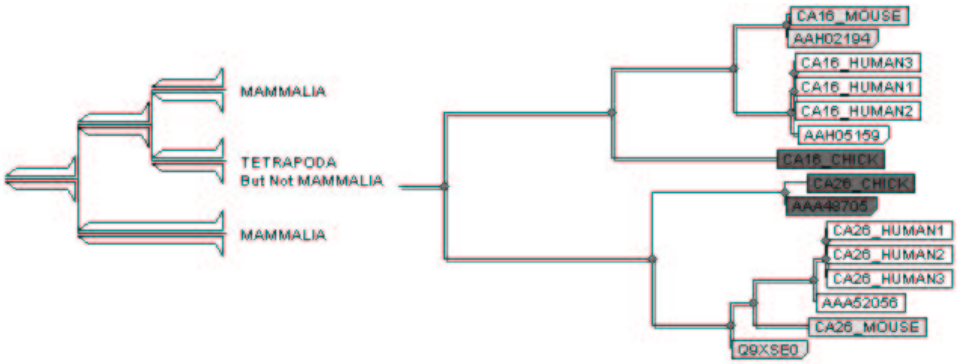

FIG. 4 – Ce motif est destiné à rechercher des familles de gènes paralogues chez les mammifères, issus d'une duplication antérieure à la divergence entre les mammifères et les autres tétrapodes. Dans ce cas, le nœud de l'arbre cible correspondant à la racine du motif pourra être interprété comme un nœud de duplication.

Il est possible de rechercher des gènes de façon plus fine que par désignation d'un simple taxon. Dans l'exemple de la figure 4, le motif recherché contient au moins deux gènes de mammifères et au moins un gène d'un tétrapode non mammalien, tels que le nœud reliant ce tétrapode au mammifère le plus proche soit postérieur au nœud reliant les deux mammifères. Ainsi, ce motif permet de retrouver des familles de gènes paralogues chez les mammifères, issus d'une duplication antérieure à la divergence entre les mammifères et les autres tétrapodes.
Notons que la recherche de motifs ne peut être effectuée pertinemment que sur des arbres enracinés. Les bases développées au Pôle Bioinformatique Lyonnais contiennent des arbres enracinés selon la méthode du point central. Cela permet d'obtenir des résultats plus fiables lorsque les motifs se limitent aux nœuds proches des feuilles. En effet, les nœuds plus profonds peuvent être victimes d'erreurs d'enracinement.
Contenant des gènes issus de vertébrés, l'arbre des espèces représentées dans HOVERGEN est relativement bien connu. Grâce à cet arbre, il est possible de proposer une prédiction de l'emplacement des nœuds de duplication sur les arbres phylogénétiques d'HOVERGEN. La méthode utilisée pour placer les duplications est une méthode automatique dérivée de la réconciliation d'arbres phylogénétiques (voir [5,8,9]). Chaque arbre est comparé à l'arbre des espèces correspondantes, en prenant en compte les redondances, les incertitudes de topologies et les longueurs de branches. Notons que la réconciliation d'arbre tend à enraciner les arbres phylogénétiques plus précisément que par la méthode du point central. Ainsi, dans le cadre de la recherche de motifs, les arbres d'HOVERGEN sont enracinés en minimisant le nombre de duplications engendrées sur l'ensemble des racines possibles. Il est donc possible d'enrichir les motifs recherchés dans HOVERGEN, en spécifiant si les nœuds du motifs correspondent à des spéciations ou à des duplications (voir fig. 5). Il est aussi envisageable de contraindre les nœuds intermédiaires à n'être que des spéciations, dans le but de détecter, par exemple, des orthologies strictes.

4 Intégration à FAMFETCH

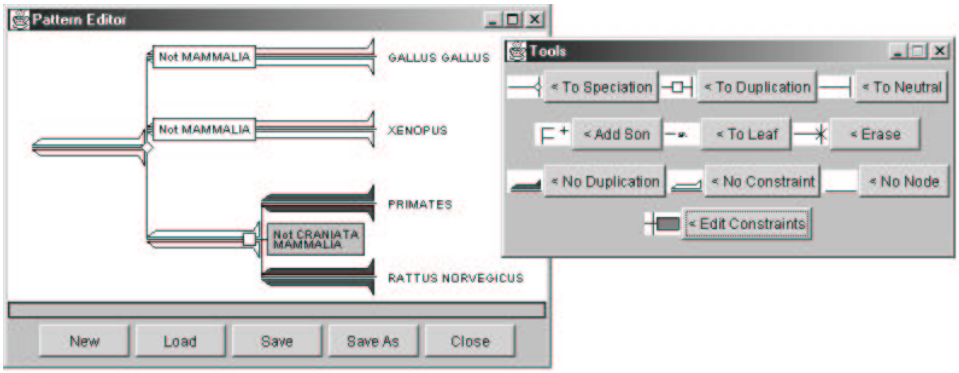

FIG. 5 – L'éditeur de motifs offre tous les outils nécessaires à la formulation de requêtes. Les motifs peuvent être chargés et sauvegardés. Il est alors possible, depuis cette fenêtre, de lancer la recherche de motifs dont les résultats seront consultables et enregistrables depuis la fenêtre principale. Dans cet exemple, les nœuds reliant Gallus, Xenopus et les mammifères (rattus, primates) sont contraints à être des spéciations (losange blanc). Par contre, le nœud reliant les primates et le rat (Rattus norvegicus) est contraint à être une duplication (carré blanc). Les branches entre ce nœud et les feuilles (primates, rattus) sont contraintes à ne contenir aucun nœud de duplication (branches sombres).

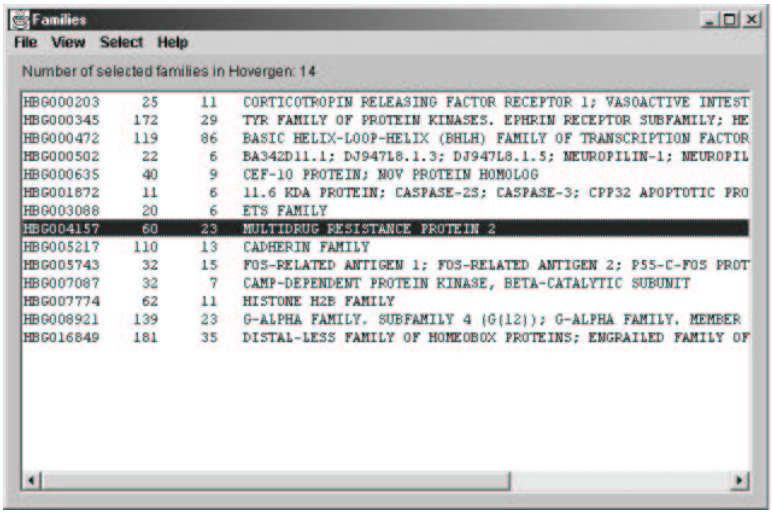

FIG. 6 – La fenêtre principale présente toutes les familles sélectionnées dans la base courante. Ici, 14 familles d'HOVERGEN ont été sélectionnées suite à la recherche de motifs décrite plus haut. Les résultats de la recherche sont enregistrables dans un fichier texte. L'arbre de chaque famille peut être visualisé en double-cliquant sur la ligne correspondante.

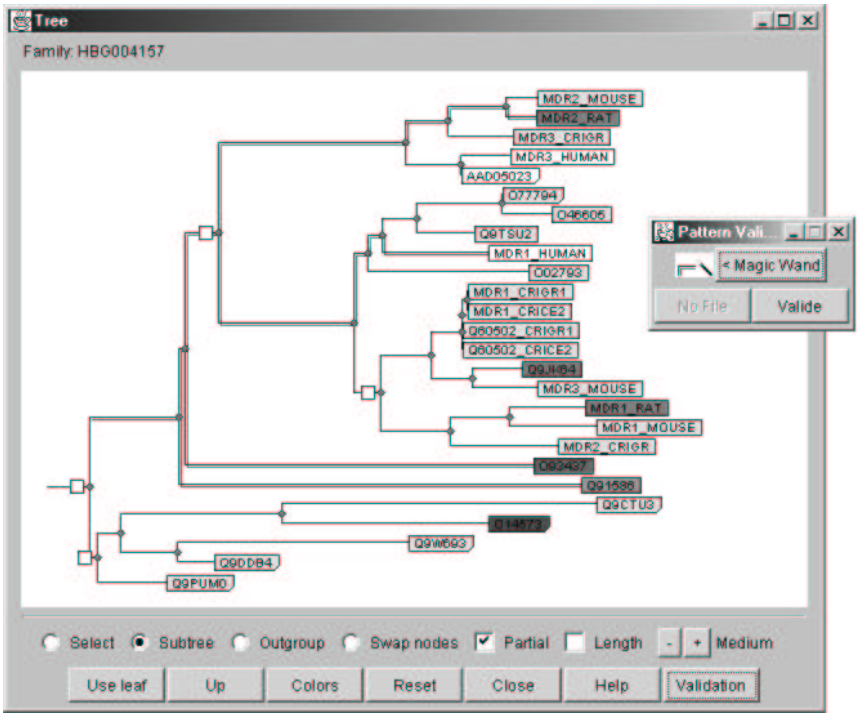

FIG. 7 – En double-cliquant sur une famille dans la fenêtre principale, la fenêtre d'arbre est ouverte et affiche l'arbre phylogénétique correspondant. Si la famille a été sélectionnée par recherche de motifs d'arbres, le ou les motifs retrouvés sont surlignés directement sur la figure. Une boîte à outils spécialisée permet de modifier les motifs et de corriger en conséquence le fichier contenant les résultats.

FAMFETCH est l'interface qui permet l'accès aux bases de familles de gènes homologues développées au Pôle Bioinformatique Lyonnais [11].

L'intégration des méthodes de recherche de motifs a été faîte selon un modèle client-serveur : l'édition du motif et la gestion des résultats sont assurées par le client, alors que les calculs de recherches de motifs, plus lourds, sont délégués à un serveur adapté à ce type de traitements.

La fenêtre principale de FAMFETCH (voir fig. 6) présente l'ensemble des familles de gènes homologues de la base choisie (HOBACGEN, HOVERGEN…). Plusieurs outils permettent déjà de sélectionner des familles sur divers critères : les espèces représentées, le nom des gènes recherchés, ou encore les tailles des familles. Un nouveau moyen de sélection a donc été ajouté : la recherche de motifs dans les arbres.

En sélectionnant l'outil « recherche de motifs », l'éditeur s'ouvre (voir fig. 5). Il est alors possible de formuler une nouvelle requête, sous la forme d'un motif d'arbre destiné à être recherché dans la base courante. Les motifs peuvent aussi être chargés et sauvegardés. Une fois le motif choisi, la recherche peut être effectuée sur la base courante, les résultats étant consultables et enregistrables depuis la fenêtre principale (voir fig. 6).
La fenêtre principale présente toutes les familles sélectionnées dans la base courante. Ici, 14 familles d'HOVERGEN ont été sélectionnées suite à la recherche de motifs décrite plus haut. Les résultats de la recherche sont enregistrables dans des fichiers textes. Ces derniers présentent la liste de tous les gènes sélectionnés, joignant le numéro de famille d'origine et le numéro du motif dans sa famille d'origine (si cette dernière en contient plusieurs). L'arbre de chaque famille peut être visualisé en double-cliquant sur la ligne correspondante.
La fenêtre affichant l'arbre phylogénétique sert à modifier les résultats de la recherche de motifs (voir fig. 7). En effet, si la famille a été sélectionnée par recherche de motifs d'arbres, le ou les motifs retrouvés sont surlignés directement sur la figure. Une boîte à outils spécialisée permet de modifier les motifs et de corriger en conséquence le fichier contenant les résultats. Il est possible, après expertise, d'effacer ou de rajouter certains motifs, mais aussi de corriger des motifs déjà identifiés. Le fichier résultat obtenu depuis la fenêtre principale est simplement modifié par rapport à sa version initiale.

5 Résultats

Le recherche de motifs, intégrée à FAMFETCH, est un outil très polyvalent. Il trouve des applications intéressantes pour toutes les bases développées au Pôle Bioinformatique Lyonnais.
Par rapport à la recherche manuelle, la recherche automatique de motifs présente des avantages et des inconvénients. L'expertise manuelle permet de détecter les anomalies éventuelles présentes sur les arbres phylogénétiques et apporte donc une plus grande souplesse quant aux données retrouvées. En effet, la formulation offerte, aussi riche soit-elle, ne pourra jamais satisfaire parfaitement l'objectif initial de l'utilisateur biologiste. L'algorithme est aussi parfois dépendant des artefacts de reconstruction, en particulier les erreurs d'enracinement sur les nœuds profonds.
En contrepartie, la recherche de motifs s'avère une opération très rapide. Le parcours d'une base d'arbres entière est parfaitement compatible avec une application interactive. Aussi, la correspondance avec des résultats obtenus manuellement, pour une recherche ayant le même objectif, est très bonne. La méthode fait autant d'erreurs qu'elle en répare en retrouvant les motifs oubliés par l'expert.
Enfin, la possibilité de corriger les résultats de l'algorithme est un bon moyen de tirer avantage de la méthode. Rechercher automatiquement des motifs est un excellent pré-traitement pour parcourir de façon précise et systématique une base d'arbres phylogénétiques.

Références

[1] A. Aho, S. Tjiang and M. Ganapathi, *Code generation using tree matching and dynamic programming*, AT&T Bell Laboratories, Stanford university, ACM Trans. Program. Lang. Syst. 11, pp. 491-516, 1989.

[2] JF. Dufayard, sous la direction de Laurent Duret et François Rechenmann, *Recherche de motifs dans les arbres*, UFR-IMA, Rapport de DEA « informatique : systèmes et communication », 2001.

[3] L. Duret, D. Mouchiroud and M. Gouy, *HOVERGEN, a database of homologous vertebrate genes*, Nucleic Acids Res. 22, 2360-2365, 1994.

[4] J. A. Eisen, *Phylogenomics : improving functional predictions for uncharacterized genes by evolutionnary analysis*, Genome Research 8, pp. 163-167, 1998.

[5] O. Eulenstein, B. Mirkin et M. Vingron, *Comparison of annotating duplication, tree mapping, and copying as methods to compare gene trees with species trees*, Mathematical hierarchies and biology, DIMACS series in discrete mathematics and theoretical computer science, Vol. 37, pp. 71-93, 1997.

[6] P. Kilpeläinen, *Tree matching problems with application to structured text databases*, University of Helsinki, departement of computer science, report A-1992-6, 1992.

[7] P. Kilpeläinen and H. Mannila, *Retrieval from hierarchical texts by partial patterns*, Proceedings of the sixteenth annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 214 – 222, 1993.

[8] R. D. M. Page et M. A. Charleston, *From gene to organismal phylogeny : reconciled trees and the tree/species tree problem*, Molecular phylogenetics and evolution, Vol. 7, No2, april, pp.231-240, 1997.

[9] R. D. M. Page et E. C. Holmes, *Molecular evolution, a phylogenetic approach*, Blackwell Science, 1998.

[10] G. Perrière, L. Duret and M. Gouy, *HOBACGEN: database system for comparative genomics in bacteria*, Genome Res. 10, 379-385, 2000.

[11] Le logiciel FAMFETCH est diponible sur le site du Pôle Bioinformatique Lyonnais à l'adresse suivante : http://pbil.univ-lyon1.fr/hobacgen/client.html