

Un modèle probabiliste d'évolution de protéines

Fabien Campillo, Lorie Dudoignon

▶ **To cite this version:**

Fabien Campillo, Lorie Dudoignon. Un modèle probabiliste d'évolution de protéines. RR-4332, INRIA. 2001. inria-00072255

HAL Id: inria-00072255
https://inria.hal.science/inria-00072255

Submitted on 23 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Un modèle probabiliste d'évolution de protéines

Fabien Campillo — Lorie Dudoignon

N° 4332

Décembre 2001

THÈME 4



ISSN 0249-6399 ISRN INRIA/RR--4332--FR

INRIA
SOPHIA ANTIPOLIS

Un modèle probabiliste d'évolution de protéines

Fabien Campillo* , Lorie Dudoignon†

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet SYSDYS

Rapport de recherche n° 4332 — Décembre 2001 — 34 pages

Résumé : Nous proposons un modèle probabiliste pour décrire l'évolution de protéines. Les paramètres de ce modèle sont identifiés par maximum de vraisemblance. Pour un exemple numérique nous étudions l'hétérogénéité du taux d'évolution le long de la séquence protéique et son rapport avec la structure secondaire.

Mots-clés : protéine, évolution, phylogénie, modèle de Markov, maximum de vraisemblance

* Projet SYSDYS (INRIA/LATP), IMT, 38 rue F. Joliot-Curie, 13451 Marseille Cedex 20 (France), Fabien.Campillo@sophia.inria.fr

† Projet SYSDYS (INRIA/LATP), IMT, 38 rue F. Joliot-Curie, 13451 Marseille Cedex 20 (France), Lorie.Dudoignon@sophia.inria.fr

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, B.P. 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : 04 92 38 77 77 - International : +33 4 92 38 77 77 — Fax : 04 92 38 77 65 - International : +33 4 92 38 77 65

A probabilistic model for protein evolution

Abstract: We propose a probabilistic model to describe the evolution of proteins. The parameters of this model are identified by the maximum likelihood criteria. For a numerical example, we study the heterogeneity of the rate of evolution along the proteinic sequence and its relationship with the secondary structure.

Key-words: protein, evolution, phylogeny, Markov model, maximum likelihood

Table des matières

1 Introduction **5**

2 Modélisation **7**
2.1 Arbre phylogénétique 7
2.2 Un modèle d'évolution de séquence de protéine 14
2.3 Loi d'un alignement 16
2.4 Calcul de la vraisemblance 17
2.5 Hypothèse de réversibilité 18
2.6 Vraisemblance pour un arbre en étoile 19
2.7 Mise en œuvre 20

3 Exemple **21**
3.1 Les données utilisées 21
3.2 Recherche du maximum de vraisemblance 22
3.3 Résultats 23
3.4 Comparaison des différentes topologies 25

4 Discussion **31**

Bibliographie **33**

1 Introduction

La phylogénie consiste à décrire l'évolution des espèces sous forme d'arbres dits phylogénétiques. Les récents progrès de la génétique donnent accès à un nouveau type de données (séquences de gènes, de protéines) qui apportent de nouvelles perspectives de recherche dans ce domaine.

L'utilisation de ces données a conduit au développement de différentes méthodes de construction d'arbres phylogénétiques, dites moléculaires (par opposition aux méthodes dites classiques utilisées précédemment). Pour avoir un aperçu de l'ensemble de ces méthodes, on peut consulter Durbin [4], Swofford et al [13], Waterman [15], Lange [10].

Dans ce rapport, nous nous intéressons aux méthodes fondées sur la *vraisemblance*, pour des données *protéiques*. L'utilisation de la vraisemblance a été introduite en phylogénie par Felsenstein [6]. Pour comprendre les protéines et leurs mécanismes d'évolution, on peut se reporter à Branden–Tooze [2] et Risler [12]. Dans [1] (paragraphe 1.3), J.L. Thorne et N. Goldman proposent un tour d'horizon récent des modèles probabilistes pour l'étude de l'évolution des protéines.

Au § 2, nous proposons une modélisation du processus d'évolution des protéines à partir d'articles écrits, le plus souvent, d'un point de vue biologique. Aussi, nous est–il apparu nécessaire d'en donner une représentation probabiliste plus formalisée. Un exemple particulier est proposé au § 3 afin d'étudier l'hétérogénéité du processus d'évolution le long de la séquence protéique.

2 Modélisation

On dispose d'un alignement :

$$\mathsf{y} = \mathsf{y}_{1:K,1:N}$$

([1]) de K séquences de protéines (homologues) de longueur N ; $\mathsf{y}_{k,n}$ désigne la valeur prise au site n de la séquence k dans l'alphabet des 20 acides aminés $\mathcal{A}$. Ainsi $\mathsf{y} \in \mathcal{A}^{K \times N}$.

Dans un premier temps nous allons proposer un modèle probabiliste pour ce genre de données. Ces dernières sont une réalisation d'une variable aléatoire Y à valeurs dans $\mathcal{A}^{K \times N}$.

L'hypothèse « d'homologie » (i.e. les séquences ont une « origine » commune) conduit à l'utilisation d'arbres phylogénétiques (voir § 2.1). Le modèle d'évolution de séquences de protéine proposé au § 2.2 décrira le phénomène de mutation des acides aminés le long des séquences (les autres phénomènes, comme l'insertion, la délétion ne seront pas pris en compte ici). On pourra alors, dans ce cadre, déterminer la loi des observations (§ 2.3) et ensuite leur vraisemblance (§ 2.4).

2.1 Arbre phylogénétique

Nous introduisons d'abord la notion d'arbre « enraciné. » Par la suite nous verrons que nous aurons besoin de la notion d'arbre « déraciné, » nous l'introduisons donc également ici. Nous définissons également la notion d'arbre « en étoile » qui sera utilisé sur le plan numérique.

1. Notation : $\mathsf{y}_{1:K,1:N}$ désigne la famille suivante :

$$\mathsf{y}_{1:K,1:N} = \left(\mathsf{y}_{k,n} \, ; \, 1 \leq k \leq K \, , \; 1 \leq n \leq N\right) , .$$

Si K désigne un ensemble d'indices on notera de la même façon $\mathsf{y}_{\mathsf{K},1:N} = (\mathsf{y}_{k,n} \, ; \, k \in \mathsf{K} \, , \; 1 \leq n \leq N)$.

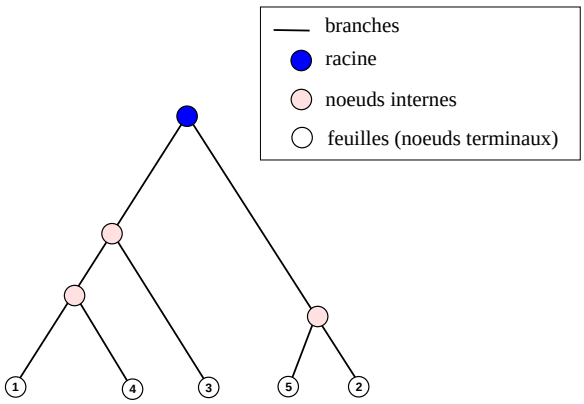


FIG. 1: Arbre enraciné T : comprend trois types de nœuds, la *racine*, les *nœuds internes* et les *feuilles*. Ces nœuds sont reliés par des branches. La racine est notée $\mathsf{r}(T)$, l'ensemble des nœuds $\mathsf{N}(T)$ (racine et feuilles comprises), l'ensemble des feuilles $\mathsf{F}(T)$, l'ensemble des temps de divergence $\mathsf{t}(T)$.

Arbres enracinés

Définition

Nous allons définir l'ensemble $\mathcal{T}_K$ des arbres (enracinés) phylogénétiques possédant K feuilles :

> *$\mathcal{T}_K$ est l'ensemble des arbres binaires à K feuilles étiquetées et auxquels on attribue des longueurs de branches.*

La Figure 1 donne un exemple d'un tel arbre correspondant à un alignement de 5 séquences (5 feuilles) : tous les nœuds correspondent à des séquences, i.e. des éléments de $\mathcal{A}^N$.

Les étiquettes associées aux feuilles correspondent aux numéros des *séquences observées*. Les longueurs de branches correspondent à des *temps de divergence*. Les nœuds internes ne sont pas étiquetés car ils correspondent à des séquences non observées (contrairement aux feuilles).

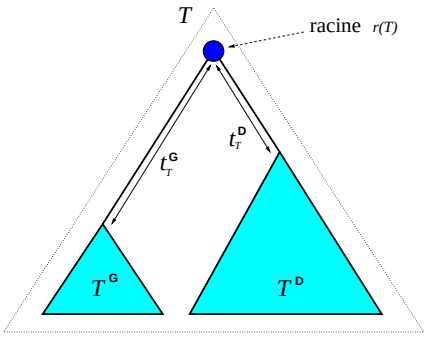


FIG. 2: Opérateurs sur $\mathcal{T}_K$, qui à un élément $T \in \mathcal{T}_K$ associe T^{G} l'arbre de descendance gauche, T^{D} l'arbre de descendance droit, $\mathsf{r}(T)$ la racine, t_T^{G} le temps de divergence de la racine à l'arbre de descendance gauche, t_T^{D} le temps de divergence de la racine à l'arbre de descendance droit.

Sur $\mathcal{T}_K$ nous définissons les opérateurs suivants (cf. Figure 2) :

(*i*) $T \mapsto T^{\mathsf{G}}$ (resp. $T \mapsto T^{\mathsf{D}}$) qui associe à $T \in \mathcal{T}_K$ son *arbre de descendance gauche* (resp. *droit*). Dans ce cas, il existe K' et K'' tels que $T^{\mathsf{G}} \in \mathcal{T}_{K'}$, $T^{\mathsf{D}} \in \mathcal{T}_{K''}$ avec $K' + K'' = K$.

(*ii*) $T \mapsto t_T^{\mathsf{G}} > 0$ (resp. $T \mapsto t_T^{\mathsf{D}} > 0$) qui associe à $T \in \mathcal{T}_K$ le *temps de divergence de la racine à l'arbre de descendance gauche* (resp. *droit*).

(*iii*) $T \mapsto \mathsf{r}(T)$ qui associe à $T \in \mathcal{T}_K$ sa *racine*.

(*iv*) $T \mapsto \mathsf{t}(T)$ qui associe à $T \in \mathcal{T}_K$ *l'ensemble des temps de divergence*.

Deux arbres T_1 et $T_2 \in \mathcal{T}_K$ seront dits égaux si :

$$T_1^{\mathsf{D}} = T_2^{\mathsf{D}}\ ,\ T_1^{\mathsf{G}} = T_2^{\mathsf{G}}\ ,\ t_1^{\mathsf{D}} = t_2^{\mathsf{D}}\ ,\ t_1^{\mathsf{G}} = t_2^{\mathsf{G}}\ .$$

ou

$$T_1^{\mathsf{D}} = T_2^{\mathsf{G}}\ ,\ T_1^{\mathsf{G}} = T_2^{\mathsf{D}}\ ,\ t_1^{\mathsf{D}} = t_2^{\mathsf{G}}\ ,\ t_1^{\mathsf{G}} = t_2^{\mathsf{D}}\ .$$

Topologie

Nous allons définir la topologie d'un arbre comme une classe d'équivalence sur $\mathcal{T}_K$. Le fait que deux arbres enracinés T_1 et T_2 possèdent la même topologie sera noté $T_1 \sim T_2$. Il s'agit d'une relation d'équivalence que nous définissons par récurrence :

$$\left.\begin{array}{l} \text{Soit } T_1 \text{ et } T_2 \in \mathcal{T}_K \ (K > 1),\ T_1 \sim T_2 \text{ si} \\ \qquad (T_1^{\mathsf{D}} \sim T_2^{\mathsf{D}} \text{ et } T_1^{\mathsf{G}} \sim T_2^{\mathsf{G}}) \quad \text{ou} \quad (T_1^{\mathsf{D}} \sim T_2^{\mathsf{G}} \text{ et } T_1^{\mathsf{G}} \sim T_2^{\mathsf{D}})\,, \\ \text{si } T_1 \text{ et } T_2 \in \mathcal{T}_1,\ T_1 \sim T_2 \text{ si} \\ \qquad T_1 = T_2\,. \end{array}\right\} \qquad (1)$$

Les classes d'équivalence seront appelée *topologies*. On note $\tilde{\mathcal{T}}_K = \mathcal{T}_K / \sim$ l'ensemble des topologies et, pour un arbre $T \in \mathcal{T}_K$, sa topologie sera notée $\tilde{T}$.

La topologie $\tilde{T}$ correspond à la « forme » de l'arbre T, i.e. l'arbre sans tenir compte des longueurs des branches. La topologie représente les liens de parenté entre les séquences observées (les feuilles), abstraction faite des distances.

Propriétés

Proposition 2.1 *Soit $T \in \mathcal{T}_K$ un arbre enraciné à K feuilles. Il existe :*

(*i*) $K-1$ *nœuds internes (dont la racine) et donc* $2K-1$ *nœuds,*

(*ii*) $2K-2$ *branches, ainsi* $|\mathsf{t}(T)| = 2K-2$,

(*iii*) $(2K-3)!!$ *topologies* ([2]) *si* $K > 1$ *et 1 si* $K = 1$.

Preuve Nous allons calculer le cardinal de $\tilde{\mathcal{T}}_K$. Cette formule a fait l'objet de la publication Felsenstein [5] de 1978 (corrigée en 1981). Nous pouvons en donner ici une version plus simple, toutefois, notre démonstration consiste à démontrer la récurrence (3) ce qui est plus aisé après avoir lu le travail de Joe Felsenstein...

2. Avec $(2K-3)!! = (2K-3)\,(2K-5)\cdots 1$

D'après la définition de la topologie d'un arbre enraciné, il est simple de vérifier que :

$$|\tilde{\mathcal{T}}_K| = \frac{1}{2}\sum_{k=1}^{K-1} C_K^k \, |\tilde{\mathcal{T}}_k| \, |\tilde{\mathcal{T}}_{K-k}| \text{ avec } |\tilde{\mathcal{T}}_1| = 1. \tag{2}$$

En effet, l'étiquetage des feuilles introduit le terme C_K^k, le $1/2$ est dû à la « symétrie » des descendances gauches et droites (cf. Définition 1).

Nous allons en déduire par récurrence la relation suivante :

$$|\tilde{\mathcal{T}}_{K+1}| = (2K-1)\,|\tilde{\mathcal{T}}_K|\,. \tag{3}$$

Supposons que $|\tilde{\mathcal{T}}_{k+1}| = (2k-1)\,|\tilde{\mathcal{T}}_k|$ pour $k = 1, \ldots, K-1$, démontrons (3) :

$$\begin{aligned}
(2K-1)\,|\tilde{\mathcal{T}}_K| &= \frac{1}{2}\sum_{k=1}^{K-1} C_K^k \, |\tilde{\mathcal{T}}_k| \, (2K-1)\,|\tilde{\mathcal{T}}_{K-k}| \\
&= \frac{1}{2}\Big(\sum_{k=1}^{K-1} C_K^k \, |\tilde{\mathcal{T}}_k| \, (2K-2k-1)\,|\tilde{\mathcal{T}}_{K-k}| \\
&\qquad + \sum_{k=1}^{K-1} C_K^k \, (2k-1)\,|\tilde{\mathcal{T}}_k| \, |\tilde{\mathcal{T}}_{K-k}| + \sum_{k=1}^{K-1} C_K^k \, |\tilde{\mathcal{T}}_k| \, |\tilde{\mathcal{T}}_{K-k}| \Big) \\
&= \frac{1}{2}\Big(\sum_{k=1}^{K-1} C_K^k \, |\tilde{\mathcal{T}}_k| \, |\tilde{\mathcal{T}}_{K+1-k}| + \sum_{k=1}^{K-1} C_K^k \, |\tilde{\mathcal{T}}_{k+1}| \, |\tilde{\mathcal{T}}_{K-k}| + |\tilde{\mathcal{T}}_K| \Big) \\
&= \frac{1}{2}\Big(\sum_{k=1}^{K-1} C_K^k \, |\tilde{\mathcal{T}}_k| \, |\tilde{\mathcal{T}}_{K+1-k}| + \sum_{k=2}^{K} C_K^{k-1} \, |\tilde{\mathcal{T}}_k| \, |\tilde{\mathcal{T}}_{K+1-k}| + |\tilde{\mathcal{T}}_K| \Big) \\
&= \frac{1}{2}\Big(\sum_{k=2}^{K-1} (C_K^k + C_K^{k-1}) \, |\tilde{\mathcal{T}}_k| \, |\tilde{\mathcal{T}}_{K+1-k}| + (C_K^1 + C_K^{K-1} + 1) \, |\tilde{\mathcal{T}}_1| \, |\tilde{\mathcal{T}}_K| \Big) \\
&= \frac{1}{2}\sum_{k=1}^{K} C_{K+1}^k \, |\tilde{\mathcal{T}}_k| \, |\tilde{\mathcal{T}}_{K+1-k}| \\
&= |\tilde{\mathcal{T}}_{K+1}|\,.
\end{aligned}$$

□

On introduit l'ensemble $\mathsf{N}(T)$ des nœuds de l'arbre T, on a la relation de récurrence suivante :

$$\mathsf{N}(T) \leftarrow \{\mathsf{r}(T)\} \cup \mathsf{N}(T^{\mathsf{G}}) \cup \mathsf{N}(T^{\mathsf{D}}) \,.$$

Enfin on pose $\mathsf{F}(T)$ l'ensemble des feuilles de T ($\mathsf{F}(T) \subset \mathsf{N}(T)$).

Arbres déracinés

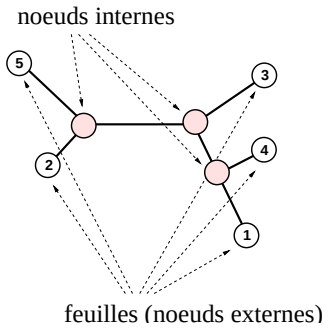


FIG. 3: Exemple d'arbre déraciné pour 5 séquences.

Comme nous le verrons au § 2.5, sous l'hypothèse de *réversibilité*, il n'est plus possible de positionner la racine de l'arbre de phylogénie. Nous serons donc amenés à faire appel à des arbres « déracinés » (cf. Figure 3). Nous proposons ici une construction de tels arbres à partir de la notion d'arbre enraciné.

À tout arbre enraciné T on associe un arbre déraciné, noté $T^{\text{dér}}$: pour un alignement de 5 séquences, à l'arbre enraciné de la Figure 1, on associe l'arbre déraciné de la Figure 3. Il est clair que des arbres enracinés différents peuvent donner le même arbre déraciné.

On définit $\mathcal{T}_K^{\text{dér}}$ l'ensemble des arbres déracinés correspondant à un alignement de K séquences :

$$\mathcal{T}_K^{\text{dér}} = \left\{ T^{\text{dér}} \,;\, T \in \mathcal{T}_K \right\} .$$

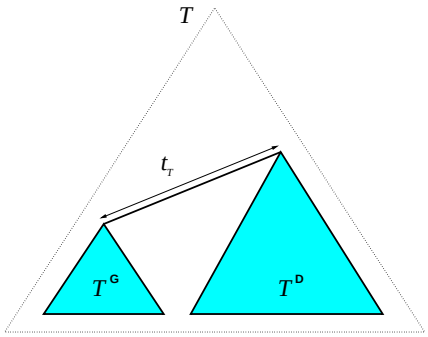


FIG. 4: Construction d'une arbre déraciné à partir d'un arbre enraciné (cf. Figure 2). Opérateurs sur $\mathcal{T}_K^{\text{dér}}$, qui à un élément $T \in \mathcal{T}_K^{\text{dér}}$ associe $T^{\mathsf{G}} \in \mathcal{T}_{K'}$ l'arbre de descendance gauche, $T^{\mathsf{D}} \in \mathcal{T}_{K''}$ l'arbre de descendance droit (avec $K'+K''=K$), t_T le temps de divergence entre les descendances gauche et droite. Attention, T' et T'' sont des arbres enracinés.

Étant donné $R \in \mathcal{T}_K^{\text{dér}}$ il existe un arbre enraciné $T \in \mathcal{T}_K$ tel que $T^{\text{dér}} = R$ (l'égalité d'arbres déracinés se construit de la même façon que celle des arbres enracinés). On peut associer à R (cf. Figure 4) :

- un arbre de descendance gauche (resp. droit) $R^{\mathsf{G}} = T^{\mathsf{G}}$, (resp. $R^{\mathsf{D}} = T^{\mathsf{D}}$), ce sont des arbres enracinés.
- le temps de divergence $t_R = t_T^{\mathsf{G}} + t_T^{\mathsf{D}}$ entre les arbres droit et gauche.

Ces objets dépendent du choix de l'arbre enraciné T. On peut également définir $\mathsf{t}(R)$, l'ensemble des temps de divergence.

La notion de topologie d'arbre déraciné se définit de façon équivalente à celle des arbres enracinés.

Proposition 2.2 *Soit $R \in \mathcal{T}_K^{\text{dér}}$ un arbre déraciné à K feuilles. Il existe :*

(*i*) *$K-2$ nœuds internes et donc $2K-2$ nœuds,*

(*ii*) *$2K-3$ branches, ainsi $|\mathsf{t}(R)| = 2K-3$,*

(*iii*) *$(2K-5)!!$ topologies si $K \geq 3$ et 1 si $K = 1$ ou 2.*

Arbres en étoile

En pratique nous serons amené à utiliser des arbres dits en étoile : ils ne sont plus binaires et possèdent une racine à laquelle toutes les feuilles sont directement rattachées (cf. 5). Ce type d'arbre ne permet pas de description phylogénétique. Toutefois, il permet de simplifier considérablement le calcul des fonctions de vraisemblance et, comme nous le verrons en pratique, de rendre compte du phénomène d'hétérogénéité le long de la séquence.

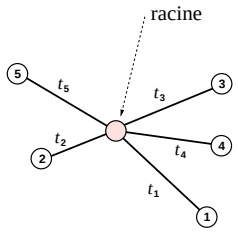


FIG. 5: Arbre en étoile pour un alignement de 5 séquences.

Pour un alignement de K séquences, il existe une seule topologie d'arbre en étoile ! La longueur de la branche liant la racine à la feuille k est notée t_k.

2.2 Un modèle d'évolution de séquence de protéine

Nous allons travailler sous l'hypothèse suivante :

Hypothèse 2.3 (« Indépendance des sites ») *L'évolution est mutuellement indépendante de site à site.*

Il suffit donc de décrire le mode d'évolution sur un site n fixé. Nous verrons que malgré cette hypothèse plutôt restrictive, nous obtenons déjà une classe suffisamment riche de modèles conduisant à des résultats convainquants.

On se fixe un arbre enraciné $T \in \mathcal{T}_K$: il décrit la descendance des K séquences actuelles (feuilles) depuis une séquence ancêtre (racine) via des séquences intermédiaires (nœuds internes), voir Figure 6.

On se donne une loi *a priori* μ_n sur la racine $\mathsf{r}(T)$ de T, la propagation de cette loi le long de l'arbre se fait à l'aide d'un processus de Markov sur $\mathcal{A}^N$

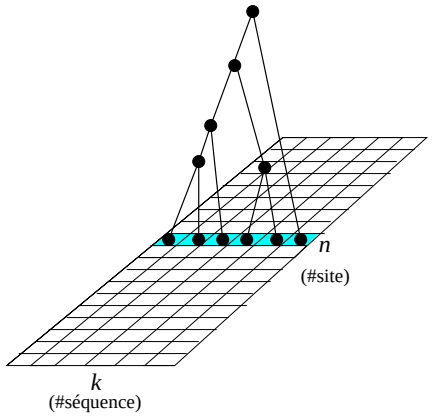


FIG. 6: Le même arbre $T \in \mathcal{T}_K$ représentera l'homologie des K séquences, i.e. T ne dépend pas de l'indice de site $n = 1, \ldots, N$.

le long des branches. Ce processus est indépendant de branche en branche. Soit $T \in \mathcal{T}_K$, on notera Q_n le générateur infinitésimal associé à la descendance gauche et à la descendance droite (le générateur est le même sur toutes les branches).

Indépendamment, on associe alors une loi μ_n^{G} (resp. μ_n^{D}) à la racine de T^{G} (resp. T^{D}). La propagation de la loi initiale μ_n se fait indépendamment le long des branches.

Hypothèse 2.4 (« Indépendance des branches ») *Le processus d'évolution est indépendant de branche en branche. La loi d'évolution est identique de branche en branche et ne dépend que de l'indice du site.*

Hypothèse 2.5 (Loi d'évolution le long des branches) *Le processus de mutation le long de l'arbre est markovien, homogène, de générateur infinitésimal Q_n, i.e. en un site fixé, c'est le même générateur qui décrit le processus de mutation le long des branches. On suppose que tous les états communiquent.*

D'après l'Hypothèse 2.5, le processus d'évolution associé à Q_n admet une probabilité invariante et une seule $\bar{\mu}_n$:

$$(\bar{\mu}_n)^* \, Q_n = 0 \, . \tag{4}$$

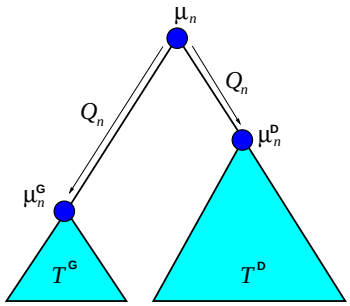


FIG. 7: Modèle d'évolution sur un arbre enraciné T : une loi μ est associée à la racine $\mathsf{r}(T)$, un générateur infinitésimal Q est associé à la descendance gauche ainsi qu'à la droite (le même générateur pour les deux). Indépendamment, on associe alors une loi μ^{G} (resp. μ^{D}) à la racine T^{G} (resp. T^{D}). La propagation de la loi initiale μ se fait indépendamment le long des branches de T.

Nous supposerons que la loi de la séquence à la racine $\mathsf{r}(T)$ est la loi produit $\bar{\mu} = \bar{\mu}_1 \otimes \cdots \otimes \bar{\mu}_N$. Le processus d'évolution est donc stationnaire.

Remarque 2.6 *Le processus de Markov étant pris homogène, Q_n ne dépend pas du temps, i.e, le schéma de remplacement des acides aminés ne dépend pas de l'endroit où on se trouve dans l'arbre.*

L'hypothèse de stationnarité signifie que le processus est à l'équilibre, ce qui n'est pas une hypothèse restrictive au niveau biologique.

L'Hypothèse 2.4 est sous-jacente à la représentation de l'évolution sous forme d'arbre qu'utilisent les biologistes. En effet, pour eux, la séparation en deux branches différentes au niveau d'un nœud, correspond à une divergence entre deux espèces distinctes, et, donc qui ne peuvent plus interagir l'une avec l'autre.

2.3 Loi d'un alignement

Soit $T \in \mathcal{T}_k$ un arbre enraciné fixé.

Nous notons $\mathsf{X}_{m,n}$ la valeur (variable aléatoire à valeur dans $\mathcal{A}$) au nœud $m \in \mathsf{N}(T)$ et au site n $(1 \leq n \leq N)$. Ainsi l'alignement Y apparaît comme la

marginale de $\mathsf{X} = \mathsf{X}_{\mathsf{N}(T),1:N}$ sur l'ensemble des feuilles $\mathsf{F}(T)$:

$$\mathsf{Y}_{\mathsf{F}(T),1:N} = \mathsf{X}_{\mathsf{F}(T),1:N} .$$

On se donne une loi initiale $\bar{\mu}$ sur la racine $\mathsf{r}(T)$: une loi produit, i.e. indépendante de site à site, sur $\mathcal{A}^N$. Décrire la loi de X, et donc de Y, revient à décrire comment la loi initiale $\bar{\mu}$ se propage le long de l'arbre via le processus de mutation décrit au § 2.2.

On a la relation de récurrence :

$$\mathrm{loi}(\mathsf{X}_{\mathsf{N}(T),n}) \leftarrow \mathrm{loi}(\mathsf{X}_{\mathsf{r}(T),n}) \otimes \mathrm{loi}(\mathsf{X}_{\mathsf{N}(T^{\mathsf{G}}),n}) \otimes \mathrm{loi}(\mathsf{X}_{\mathsf{N}(T^{\mathsf{D}}),n}) \tag{5}$$

qui signifie que, par hypothèse d'indépendance des branches, la loi de X sur un arbre est le produit de la loi de X sur la racine de cet arbre, de la loi sur l'arbre de descendance gauche et de la loi sur l'arbre de descendance droit. Cette relation va permettre le calcul effectif de la vraisemblance.

2.4 Calcul de la vraisemblance

On se donne un alignement de K séquences de longueur N.

L'hypothèse « d'indépendance de sites » permet d'écrire la vraisemblance d'un arbre $T \in \mathcal{T}_K$ et de la famille $Q = Q_{1:N} \in \mathcal{Q}^N$ ($\mathcal{Q}$ est l'espace des générateurs de taille N admettant une mesure invariante) comme :

$$\mathcal{L}(T, Q) = \prod_{n=1}^{N} \mathcal{L}_n(T, Q_n)$$

où, $\mathcal{L}_n(T, Q_n)$, la vraisemblance d'un arbre $T \in \mathcal{T}_K$ et du générateur Q_n au site $n \in \{1, \ldots, N\}$ est donnée par :

$$\mathcal{L}_n(T, Q_n) = \sum_{a \in \mathcal{A}} \bar{\mu}_n(a)\, \mathcal{L}_n(T, Q_n | \mathsf{X}_{r(T),n} = a)$$

et la vraisemblance de T au site n, sachant que sa racine est a, est donnée par (d'après l'hypothèse « d'indépendance des branches ») :

$$\mathcal{L}_n(T, Q_n | \mathsf{X}_{r(T),n} = a) = \begin{cases} \sum_{b \in \mathcal{A}} (e^{t_T^{\mathsf{D}} Q_n})_{a,b} \, \mathcal{L}_n(T^{\mathsf{D}}, Q_n | \mathsf{X}_{r(T^{\mathsf{D}}),n} = b) \times \\ \qquad \times \sum_{c \in \mathcal{A}} (e^{t_T^{\mathsf{G}} Q_n})_{a,c} \, \mathcal{L}_n(T^{\mathsf{G}}, Q_n | \mathsf{X}_{r(T^{\mathsf{G}}),n} = c) \\ \qquad\qquad \text{lorsque } T \text{ n'est pas une feuille,} \\ \mathbf{1}_{\mathsf{X}_{r(T),n} = a} = \mathbf{1}_{\mathsf{Y}_{r(T),n} = a} \\ \qquad\qquad \text{lorsque } T \text{ est une feuille.} \end{cases} \tag{6}$$

Il s'agit de l'algorithme « d'élagage » (*pruning*) de Felsenstein [6].

2.5 Hypothèse de réversibilité

Hypothèse 2.7 (Réversibilité) *On suppose que le processus d'évolution défini au § 2.2, est* réversible, *c'est–à–dire :*

$$\bar{\mu}_n(a) \, [Q_n]_{a,b} = \bar{\mu}_n(b) \, [Q_n]_{b,a} \,, \ \forall a, b \in \mathcal{A} \,.$$

Cette hypothèse signifie, qu'en un site n donné, la probabilité qu'il y ait une mutation de l'acide aminé a vers l'acide aminé b en un temps t est égale à la probabilité qu'il y ait une mutation de l'acide aminé b vers l'acide aminé a dans le même laps de temps.

On suppose à présent que Q_n appartient à l'espace $\mathcal{Q}^{\text{rév}}$ des générateurs de dimension N réversibles.

D'après la section précédente :

$$\begin{aligned}
\mathcal{L}_n(T, Q_n) &= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{A}}\sum_{c\in\mathcal{A}} \bar{\mu}_n(a)\,(e^{t_T^{\mathsf{D}}\,Q_n})_{a,b}\,(e^{t_T^{\mathsf{G}}\,Q_n})_{a,c} \\
&\qquad\qquad \mathcal{L}_n(T^{\mathsf{D}}, Q_n|\mathsf{X}_{r(T^{\mathsf{D}}),n} = b)\,\mathcal{L}_n(T^{\mathsf{G}}, Q_n|\mathsf{X}_{r(T^{\mathsf{G}}),n} = c) \\
&= \sum_{b\in\mathcal{A}}\sum_{c\in\mathcal{A}}\Big(\sum_{a\in\mathcal{A}} \bar{\mu}_n(b)\,(e^{t_T^{\mathsf{D}}\,Q_n})_{b,a}\,(e^{t_T^{\mathsf{G}}\,Q_n})_{a,c}\Big) \\
&\qquad\qquad \mathcal{L}_n(T^{\mathsf{D}}, Q_n|\mathsf{X}_{r(T^{\mathsf{D}}),n} = b)\,\mathcal{L}_n(T^{\mathsf{G}}, Q_n|\mathsf{X}_{r(T^{\mathsf{G}}),n} = c) \\
&= \sum_{b\in\mathcal{A}}\sum_{c\in\mathcal{A}} \bar{\mu}_n(b)\,(e^{(t_T^{\mathsf{D}}+t_T^{\mathsf{G}})\,Q_n})_{b,c} \\
&\qquad\qquad \mathcal{L}_n(T^{\mathsf{D}}, Q_n|\mathsf{X}_{r(T^{\mathsf{D}}),n} = b)\,\mathcal{L}_n(T^{\mathsf{G}}, Q_n|\mathsf{X}_{r(T^{\mathsf{G}}),n} = c)
\end{aligned}$$

Ainsi, il n'est plus possible d'estimer t_T^{D} et t_T^{G} mais seulement leur somme. En d'autres termes, cela signifie que l'arbre d'évolution sous–jacent est *déraciné*. En effet, si on prend $R = T^{\text{dér}}$, la fonction de vraisemblance est :

$$\begin{aligned}
\mathcal{L}_n(T, Q_n) = \mathcal{L}_n(R, Q_n) = \sum_{b\in\mathcal{A}}\sum_{c\in\mathcal{A}} \bar{\mu}_n(b)\,(e^{t_R\,Q_n})_{b,c}\,\mathcal{L}_n(R^{\mathsf{D}}, Q_n|\mathsf{X}_{r(R^{\mathsf{D}}),n} = b) \\
\mathcal{L}_n(R^{\mathsf{G}}, Q_n|\mathsf{X}_{r(R^{\mathsf{G}}),n} = c)
\end{aligned}$$

En fait, rien ne nous permet de positionner la racine dans l'arbre.

2.6 Vraisemblance pour un arbre en étoile

D'après les hypothèses précédentes, la vraisemblance d'un arbre en étoile $T^{\text{ét}}$ est :

$$\mathcal{L}(T^{\text{ét}}, Q) = \prod_{n=1}^{N} \mathcal{L}_n(T^{\text{ét}}, Q_n)$$

avec

$$\mathcal{L}_n(T^{\text{ét}}, Q_n) = \sum_{a\in\mathcal{A}}\prod_{k=1}^{K}(e^{t_k\,Q_n})_{a,\mathsf{Y}_{n,k}}\,.$$

Estimer l'arbre revient ici à estimer le vecteur des distances $\mathsf{t}(T^{\text{ét}}) = t_{1:K}$.

2.7 Mise en œuvre

On s'intéresse à présent à la mise en œuvre pratique de l'estimation des paramètres (Q, T) par maximum de vraisemblance, pour un alignement de K séquences de longueur N.

On pose :

$$\mathcal{L}(T) = \mathcal{L}(T, \hat{Q}(T)) = \max_{Q \in \mathcal{Q}^N} \mathcal{L}(T, Q)\,,$$

et on cherche à déterminer l'arbre qui maximise $\mathcal{L}(T)$.

Il n'est pas simple de chercher l'arbre qui maximise cette fonction. En effet, il nous faut parcourir l'ensemble des topologies $\tilde{\mathcal{T}}_K$ et pour chaque topologie $\tilde{T} \in \tilde{\mathcal{T}}_K$ optimiser les longueurs des branches, i.e. chercher $\hat{T} \in \tilde{T}$ tel que :

$$\mathcal{L}(\hat{T}) = \max_{T \in \tilde{T}} \mathcal{L}(T)\,.$$

Ensuite, on choisit la topologie optimale.

En résumé, on cherche l'arbre qui réalise le maximum suivant :

$$\max_{\tilde{T} \in \tilde{\mathcal{T}}_K} \max_{T \in \tilde{T}} \mathcal{L}_n(T)\,.$$

Il faut noter que le parcours de $\tilde{\mathcal{T}}_K$ pose un problème de combinatoire (cf. $|\tilde{\mathcal{T}}_K|$ calculé à la Proposition 2.1). Pour l'application numérique, nous n'abordons pas ce problème : on calcule la vraisemblance de toutes les topologies (en se plaçant dans un cas où K est relativement faible).

Remarque 2.8 *D'après la forme de la fonction de vraisemblance, on ne pourra estimer Q qu'à un facteur près. En effet, on aura la même valeur de $\mathcal{L}(T, Q)$ pour les estimateurs $(\hat{T}, \hat{Q})$ et $(\tilde{T}, \tilde{Q})$ si il existe $\lambda \in \mathbb{R}_+$ tel que :*

$$\mathsf{t}(\tilde{T}) = \lambda\, \mathsf{t}(\hat{T}) \text{ et } \tilde{Q}_n = \frac{1}{\lambda}\, \hat{Q}_n\,,\ 1 \leq n \leq N\,.$$

3 Exemple

Dans ce modèle simple, on associe à chaque site un taux de substitution. Le but modélisation est de « quantifier » la variation des taux d'évolution le long de l'alignement afin de mettre en relation cette variation avec la structure secondaire de la protéine étudiée. Nous voulons mettre en évidence qu'aux sites où la protéine est plus structurée (conformation en hélices α, en feuillets β ...), la séquence est plus conservée qu'aux autres sites.

On prend pour matrice de substitution au site n une matrice de la forme :

$$Q_n = \tau_n Q \,,$$

i.e, seul le taux de substitution τ_n varie de sites à sites. On a donc $\bar{\mu}_n$ qui ne dépend pas de n. On ajoute une contrainte de normalisation sur $\tau = \tau_{1:N}$ pour résoudre le problème exposé dans la Remarque 2.8 : $\sum_{1 \leq n \leq N} \tau_n = 1$.

La matrice Q est donnée, on prend une matrice du type Dayhoff [3] qui est une matrice de substitution empirique calculée sur un grand nombre de séquences de protéines. On choisit ici celle de Jones–Taylor–Thornton [9], parmi les plus utilisées. Il est à noter que toutes ces matrices conduisent, par construction, à des processus d'évolution réversibles et donc à l'utilisation d'arbres déracinés.

Pour ce modèle, il nous faudra donc estimer l'arbre de maximum de vraisemblance ainsi que le vecteur τ.

3.1 Les données utilisées

Pour faciliter notre recherche nous utilisons le serveur SRS de *l'European Bioinformatics Institute* de Cambridge[3] qui permet de faire une recherche simultanée dans plusieurs banques, sélectionnées au préalable. Sur ce serveur, on trouve notamment une description détaillée de chaque banque de données ainsi que de leur format.

Dans notre cas, nous sélectionnons la banque de données *Prosite*, dans la section « InterPro&Related » qui contient des protéines homologues. Nous utilisons 5 séquences de la famille des `HSP70` (*Heat Shock Protein*) : `HS70_BRUMA`

3. `http://srs.ebi.ac.uk`

(nématode : *Brugia malayi*), HS71_DROME (drosophile), HS71_HUMAN (homme), HS7C_BOVIN (bovin) et HS7C_MOUSE (souris). Pour chacune de ces séquences, nous récupérons les fichiers PDB (Protein Data Bank) qui en donnent une description détaillée.

Ces séquences sont alignées à l'aide de Clustal W ([4]). On obtient ainsi un alignement de longueur 384. La structure utilisée pour les comparaisons est celle du bovin.

Remarque 3.1 *Les séquences utilisées ne sont, la plupart du temps, pas de même longueur. Car, en plus des phénomènes de mutation, il existe des phénomènes d'insertion et de délétion. Aussi, quelque soit la méthode utilisée pour construire l'alignement, nous sommes amenés à introduire des gaps dans certaines séquences (noté −). Or, comme nous l'avons vu précédemment le modèle proposé ne tient pas compte de ces gaps. En pratique, deux solutions simples sont envisageables. La première consiste à supprimer tous les sites où il y auraient des gaps de l'alignement. La deuxième permet de conserver toute l'information en traitant les gaps comme des valeurs manquantes. C'est à dire qu'on suppose qu'à la place, il y a en réalité un acide aminé mais qu'on ne sait pas lequel (on fait donc la somme sur toutes les possibilités dans l'Équation (6)). Dans les résultats numériques qui suivent c'est cette deuxième solution que nous avons choisie.*

3.2 Recherche du maximum de vraisemblance

On cherche à maximiser la log–vraisemblance :

$$\ell(T,\tau) = \log \mathcal{L}(T,\tau) = \log \mathcal{L}(T,\mathsf{Q})\,.$$

On rappelle que dans le modèle utilisé ici on a $Q_n = \tau_n Q$ pour $1 \leq n \leq N$ avec Q donnée.

Pour cela, on parcourt l'ensemble des topologies $\tilde{\mathcal{T}}_5$ de manière exhaustive (ce qui est encore possible puisque $|\tilde{\mathcal{T}}_5| = 15$). Pour chaque topologie $\tilde{T} \in \tilde{\mathcal{T}}_5$ on cherche l'arbre T et le vecteur τ optimaux. C'est à dire que l'on cherche le

4. http://www.sgi.com/solutions/sciences/chembio/resources/clustalw/parallel_clustalw.html

maximum suivant :

$$\max_{T\in\tilde{T}} \max_{\tau\in\mathbb{R}^N} \ell(T,\tau),$$

sous la contrainte $\sum_{1\leq n\leq N} \tau_n = 1$.

Ce qui revient en fait à optimiser selon les deux vecteurs $\mathsf{t}(T)$ et τ ($\mathsf{t}(T)$ désigne les longueurs de branche de T).

En pratique, nous utilisons une méthode d'optimisation alternée[5] qui présente l'avantage de ne pas nécessiter le calcul du gradient (qui est ici très coûteux).

3.3 Résultats

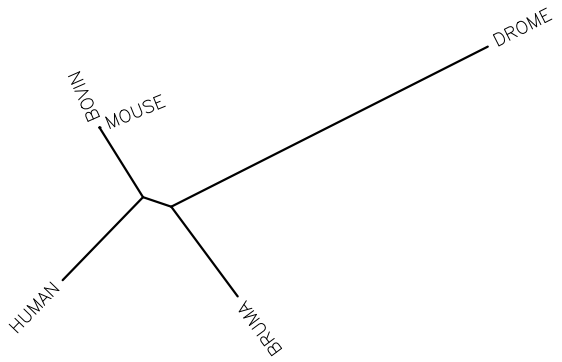


FIG. 8: Arbre optimisé.

L'arbre de maximum de vraisemblance est représenté sur la Figure 8.

5. Méthode de Powell, pp. 394–455 de *Numerical Recipes in C, the Art of Scientific Computing, Second Edition, Cambridge University Press 1992.*

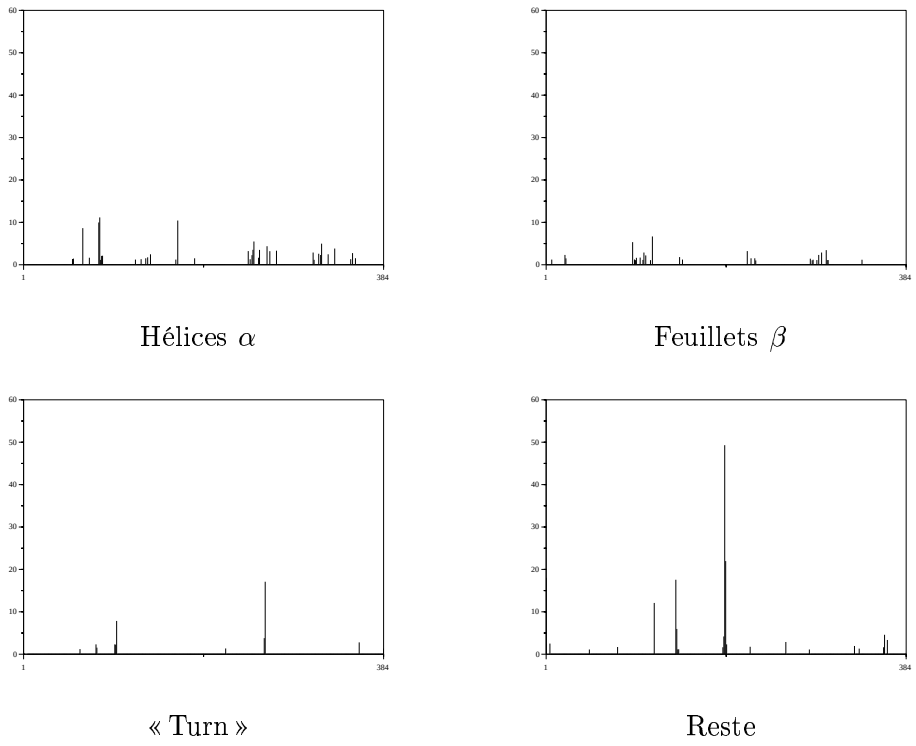


FIG. 9: Taux d'évolution relatifs le long de la séquence en distinguant 4 catégories de structure secondaire : les hélices α, les feuillets β, les « turn » et le « reste ». Pour mieux analyser ses résultats il faut se reporter aux Tableaux 1 et 2.

La structure utilisée pour les comparaisons est celle de la séquence du bovin. Sur la Figure 9 nous avons représenté les taux d'évolution relatifs le long de la séquence en distinguant 4 catégories de structure secondaire : les hélices α, les feuillets β, les « turn » et le « reste ». En nous limitant aux « trois types classiques » de structure secondaire (hélices α, feuillets β et « reste »), nous ne pouvons pas mettre en évidence de relation significative entre la structure et l'hétérogénéité des taux.

Le Tableau 1 donne le nombre de sites concernés ainsi que la moyenne des taux d'évolution relatifs pour chacune des catégories. Dans le Tableau 2,

catégorie	nombre de sites	taux relatif moyen
α hélices	151	0,79
feuillets β	110	0,53
« turn »	51	0,87
reste	*72*	*2,25*
total	384	1

TAB. 1: Nombre de sites par catégorie et taux d'évolution relatif associé.

catégorie	nombre de sites	pourcentage	taux relatif moyen
α hélices	37	24.50%	3,21
feuillets β	29	26.36%	1,97
« turn »	11	21.57%	3,98
reste	*24*	*33.33%*	*6,75*
total	101	26.30%	3,78

TAB. 2: Par catégorie, nombre de sites qui ont changé, pourcentage de sites qui ont changé par rapport au nombre de sites de la catégorie et taux d'évolution relatif des sites qui ont changés.

nous ne considérons que les sites pour lesquels nous observons une différence entre les cinq séquences étudiées. Pour chacune des catégories, nous donnons le nombre de sites concernés, leurs pourcentages par rapport au nombre total de sites dans la catégorie et la moyenne des taux d'évolution relatifs.

Nous constatons sur cet exemple que les zones très structurées (hélices α, feuillets β et « turn » ont un taux d'évolution moyen plus faible que le reste de la séquence et que leurs pourcentages de sites où l'on observe des différences est aussi plus faible.

3.4 Comparaison des différentes topologies

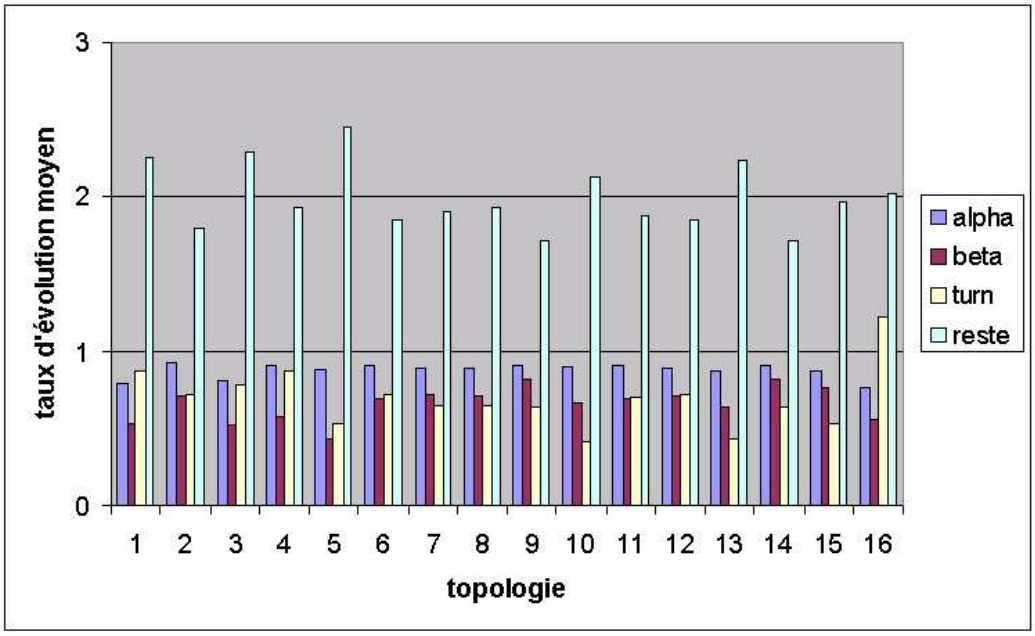
Toujours sur le même jeu de données, il est possible de considérer les 15 topologies possibles, T1 (celle du maximum de vraisemblance), T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12, T13, T14, T15, ainsi que la topologie en étoile

Et (également noté T16). Cette dernière forme d'arbre, bien que non réaliste, simplifie grandement le calcul des expressions de vraisemblance et peut donc avoir un intérêt si les résultats varient peu.

La comparaison de toutes ces topologies permet de vérifier la robustesse des résultats selon la topologie. Ce qui nous intéresse est de voir si le fait de travailler sur une « mauvaise » topologie modifie considérablement les estimations du vecteur r par rapport à la « bonne » topologie. Si tel n'est pas le cas, nous pourrons alors travailler sur des alignements plus importants sans nous poser le problème du choix de la topologie, qui devient, comme nous l'avons signalé auparavant, très compliqué pour un grand nombre de séquences.

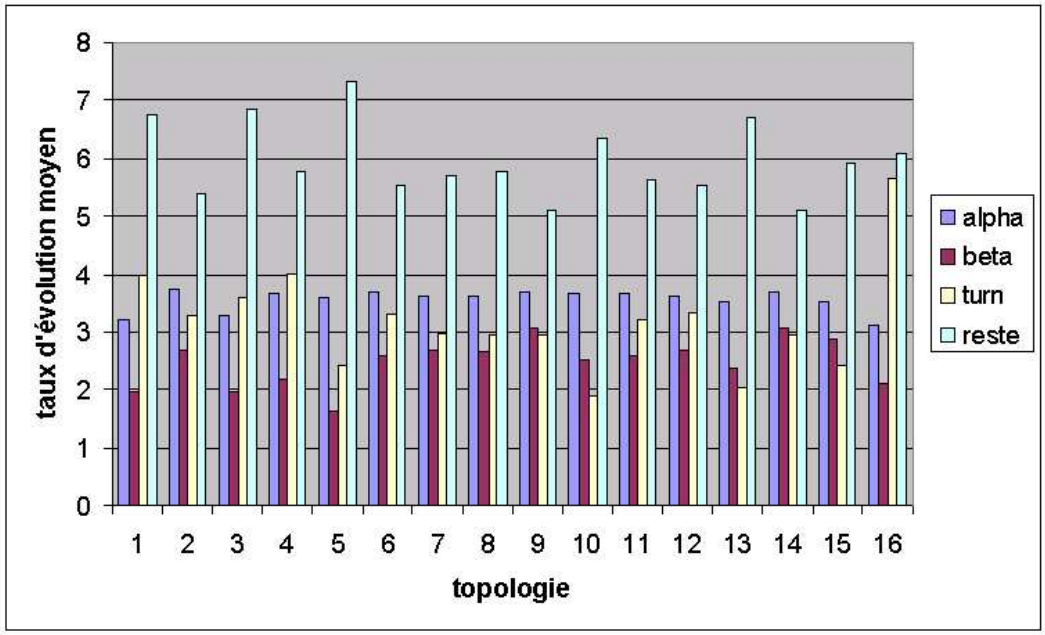
Aux vues des Tableaux 3 et 4, on voit que les principales tendances sont conservées : quelque soit la topologie considérée, les parties structurées de la protéine évoluent beaucoup moins que le reste de la protéine. De plus, d'après le Tableau 5, le choix de la topologie n'affecte pas considérablement l'estimation du taux d'évolution.

Arbre	*α Hélices*	*Feuillets β*	*« Turn »*	*Reste*	*Séquence entière*
T1	0,79	0,53	0.87	2,25	1
T2	0,92	0,71	0,72	1,80	1
T3	0,81	0,52	0,78	2,29	1
T4	0,91	0,57	0,87	1,93	1
T5	0,88	0,43	0,53	2,45	1
T6	0,91	0,69	0,72	1,85	1
T7	0,89	0,71	0,64	1,91	1
T8	0,89	0,71	0,64	1,93	1
T9	0,91	0,82	0,64	1,71	1
T10	0,90	0,67	0,41	2,13	1
T11	0,91	0,69	0,70	1,88	1
T12	0,89	0.71	0,72	1,86	1
T13	0,87	0,63	0,44	2,24	1
T14	0,91	0,82	0,64	1,71	1
T15	0,87	0,76	0,53	1,98	1
Et	0,76	0,56	1,22	2.02	1



TAB. 3: Taux d'évolution moyen par catégorie de structure secondaire et selon la topologie.

Arbre	*α Hélices*	*Feuillets β*	*« Turn »*	*Reste*	*Séquence entière*
T1	3,21	1,97	3,98	6,75	3,78
T2	3,71	2,68	3,28	5,40	3,77
T3	3,27	1,96	3,60	6,84	3,78
T4	3,68	2,17	4,01	5,77	3,78
T5	3,58	1,66	2,42	7,32	3,79
T6	3,69	2,59	3,30	5,53	3,77
T7	3,62	2,70	2,97	5,72	3,78
T8	3,61	2,67	2,96	5,76	3,78
T9	3,71	3,09	2,94	5,12	3,79
T10	3,68	2,52	2,90	6,37	3,79
T11	3,67	2,60	3,23	5,61	3,78
T12	3,62	2,68	3,33	5,56	3,78
T13	3,53	2,38	2,02	6,70	3,79
T14	3,71	3,09	2,94	5,13	3,79
T15	3,52	2,88	2,43	5,92	3,79
Et	3,10	2,11	5,64	6,07	3,80



TAB. 4: Taux d'évolution moyen par catégorie de structure secondaire et selon la topologie pour les sites où l'on observe un changement.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	Et
T1	1	0,78	0,99	0,85	0,84	0.81	0,78	0,79	0,56	0.71	0,80	0,79	0,74	0.56	0,69	0,64
T2	.	1	0,77	0,84	0.75	0,997	0,99	0,99	0.88	0,87	0,997	0,996	0.88	0,88	0,85	0,77
T3	.	.	1	0,86	0.83	0,81	0,79	0,80	0.57	0,71	0,80	0,80	0.74	0,57	0,69	0,62
T4	.	.	.	1	0.67	0,86	0,79	0,79	0.73	0,61	0,84	0,82	0.65	0,73	0,66	0,57
T5	.	.	.	.	1	0,78	0,80	0,80	0.66	0,86	0,78	0,77	0.85	0,66	0,88	0,70
T6	.	.	.	.	.	1	0,99	0,99	0.88	0,88	0,999	0,996	0.89	0,88	0,87	0,78
T7	.	.	.	.	.	.	1	0,999	0.88	0,93	0,99	0,995	0.94	0,88	0,90	0,80
T8	.	.	.	.	.	.	.	1	0,88	0.93	0,99	0,994	0,94	0.88	0,90	0,80
T9	.	.	.	.	.	.	.	.	1	0,80	0,87	0,88	0.77	1	0,91	0,70
T10	.	.	.	.	.	.	.	.	.	1	0,90	0,90	0.99	0,80	0,92	0,83
T11	.	.	.	.	.	.	.	.	.	.	1	0,997	0.90	0,87	0,86	0,78
T12	.	.	.	.	.	.	.	.	.	.	.	1	0.91	0,88	0,87	0,80
T13	.	.	.	.	.	.	.	.	.	.	.	.	1	0,77	0,89	0,83
T14	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0,91	0,69
T15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	0,77
TEt	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1

TAB. 5: Corrélations des taux d'évolution le long de la séquence entre topologies.

4 Discussion

L'approche que nous proposons ici (cf. Section 3) afin d'étudier l'hétérogénéité des taux d'évolution le long de l'alignement, présente l'avantage d'être simple. Les premiers résultats sont encourageants. Mais il faut tout de même signaler que ce modèle est limité par son trop grand nombres de paramètres à optimiser. Le temps de calcul s'en fait grandement ressentir ainsi que la qualité des estimations obtenues.

En ce qui concerne l'étude de l'hétérogénéité des taux d'évolution le long de l'alignement, on peut citer les travaux de Yang [16, 18] qui opte pour une approche bayésienne dont l'avantage réside dans le nombre peu élevé de paramètres introduits pour tenir compte de l'hétérogénéité, ce qui lui permet ensuite, à peu de frais de s'affranchir de l'hypothèse, très contraignante du point de vue biologique, d'indépendance des sites (cf. Yang [17]). Les modèles proposés par Z. Yang sont essentiellement consacrés aux séquences d'ADN, mais s'adaptent facilement aux séquences protéiques.

On peut aussi citer les travaux de Goldman et al. [14, 8, 11] qui proposent des modèles pour des séquences de protéines, dans lesquels l'hétérogénéité des taux est supposée liée à la structure secondaire. Ils utilisent pour cela des modèles de Markov cachés (cf. Felsenstein [7]).

Références

[1] D.J. Balding, M. Bishop, and C. Cannings, editors. *Handbook of Statistical Genetics*. John Wiley, 2001.

[2] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing Inc., second edition, 1999.

[3] M. O. Dayhoff, R. V. Eck, and C. M. Park. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, volume 5, pages 89–99, Washington DC, 1972. National Biomedical Fundation.

[4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

[5] J. Felsenstein. The number of evolutionary trees. *Systematic Zoology*, 27:27–33, 1978. (Correction, vol. 30, p. 122, 1981).

[6] J. Felsenstein. Evolutionnary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[7] J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate evolution. *Molecular Biology and Evolution*, 13:93–104, 1996.

[8] N. Goldman, J. L. Thorne, and D. T. Jones. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149:445–458, 1998.

[9] D. T. Jones, W. R. Taylor, and J.M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Cabios*, 8:275–282, 1992.

[10] K. Lange. *Mathematical and Statistical Methods for Genetic Analysis*. Springer Verlag, 1997.

[11] P. Liò, N. Goldman, J. L. Thorne, and D. T. Jones. PASSML: Combining evolutionary inference and protein secondary structure prediction. *Bioinformatics*, 14:726–733, 1998.

[12] J. L. Risler. Evolution des protéines. *Le dictionnaire du darwinisme*, 1996.

[13] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. *Molecular Systematics (D.M. Hillis, C. Moritz, B.K. Mable eds.)*, chapter 11 : Phylogenetic inference, pages 407–514. Sinauer, second edition, 1996.

[14] J. L. Thorne, N. Goldman, and D. T. Jones. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13(5):666–673, 1996.

[15] M. S. Waterman. *Introduction to Computational Biology.* Chapman and Hall, 1995.

[16] Z. Yang. Maximum–likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10:1396–1401, 1993.

[17] Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139:993–1005, 1995.

[18] Z. Yang. Among–site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, 11(9):367–372, 1996.



Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 Villers lès Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot St Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, B.P. 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399