Variation du taux d'évolution le long de séquences de protéines

Lorie Dudoignon[1] Élisabeth Remy[2] Jean–Loup Risler[2] Fabien Campillo[1]

[1]INRIA/LATP
IMT
38, rue F. Joliot–Curie
13451 Marseille cedex 20
{ldudoignon|campillo}@sophia.inria.fr

[2]Laboratoire Génome et Informatique
Batiment Buffon
45, avenue des Etats-Unis
78035 Versailles
{remy|risler}@genetique.uvsq.fr

Résumé

Il s'agit d'analyser les variations du taux d'évolution le long de séquences de protéines et de déterminer s'il existe un lien avec la structure secondaire de ces protéines. Pour décrire l'évolution nous utilisons des modèles de Markov. Nous mettons en évidence qu'aux sites où la protéine est plus structurée (conformation α–hélices, brins–β, les *turn*), la séquence est plus conservée qu'aux autres sites.

1 Présentation du problème

Nous cherchons à déterminer s'il existe une relation entre la structure d'une protéine et la variation des taux d'évolution le long de la séquence. Il est communément admis que la structure est plus conservée que la séquence, on peut donc s'attendre à une hétérogénéité du taux d'évolution du fait des contraintes physico–chimiques liées à la structure.

Nous proposons une modélisation probabiliste simple des séquences protéiques (sites indépendants). Pour le modèle d'évolution, on utilise un modèle de Markov en temps continu (homogène, stationnaire, réversible) à valeurs dans l'alphabet des vingt acides aminés (on suppose que les séquences ne présentent pas de brèches, du moins on n'en tient pas compte).

On considère des séquences alignées, pour lesquelles la structure est connue (au moins la structure de l'une d'entre elles). Les séquences sont alignées à l'aide de CLUSTAL qui fournit un arbre de distances. Nous nous baserons sur la topologie de cet arbre (ainsi nous n'abordons pas le délicat problème de la détermination de la topologie de l'arbre par maximum de vraisemblance). Nous supposons évidemment que l'arbre phylogénétique (sa topologie mais aussi la longueur de ses branches) sera commun à tous les sites. À chaque site nous associons un taux d'évolution différent. Nous estimons donc les longueurs des branches ainsi que ces taux par maximum de vraisemblance (cf. Felsenstein [1]). Enfin, les taux estimés sont comparés à la structure connue.

Les premières méthodes de maximum de vraisemblance se contentaient de décrire l'évolution au niveau d'un site, puis, de supposer que tous les sites d'une même séquence suivent indépendamment le même processus d'évolution. Cette dernière hypothèse semble toutefois être irréaliste. Aussi de nombreux modèles ont été développés ces dernières années dans le but de mieux décrire l'évolution sur l'ensemble de la séquence et de "mesurer" l'hétérogénéité des taux d'évolution le long de celle–ci.

Certains modèles conservent l'hypothèse d'indépendance (cf. [5], [3]). D'autres préfèrent supprimer cette hypothèse d'indépendance en faisant appel à des modèles de Markov cachés (cf. [6], [2]).

Nous pouvons aussi consulter les travaux de *Thorne et al* [4] qui proposent des modèles pour des séquences protéiques permettant de tenir compte de la structure des protéines.

Dans la Section 2 nous détaillons le modèle et l'algorithme utilisés. Dans la Section 3 nous présentons un exemple numérique.

2 Modélisation et algorithme

Nous considérons un alignement de K séquences de longueur N : $\{\mathsf{y}_{k,n}\,;\, 1 \leq k \leq K,\, 1 \leq n \leq N\}$, $\mathsf{y}_{k,n}$ désigne la valeur prise au site n de la séquence k dans l'alphabet des 20 acides aminés, noté $\mathcal{A}$.

On dispose d'un arbre phylogénétique (déraciné) de $2K-3$ branches de longueurs $t_1, t_2, \ldots, t_{2K-3}$ (cf. Figure 1 pour un exemple de 4 séquences).

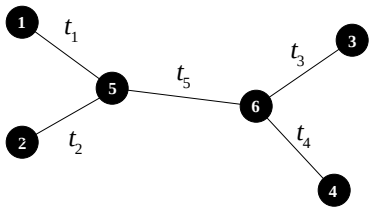

FIG. 1: Exemple d'arbre déraciné pour 4 séquences. Ici nous considérons que la topologie est donnée et que les paramètres $\mathbf{t} = (t_i)$ sont inconnus.

Sur un site n fixé, notons τ_n le taux d'évolution. Nous décrivons l'évolution de ce site par un processus de Markov $\{X_t\}_{t\geq 0}$ en temps continu, homogène, réversible, stationnaire. Notons $P^{\tau_n}_{a,b}(t)$ la probabilité de transition de a à b en un temps t :

$$P^{\tau_n}_{a,b}(t) = \mathbb{P}(X_t = b | X_0 = a)\,,\ a, b \in \mathcal{A}\,,\ t \geq 0\,.$$

Nous appelons Q_n le générateur infinitésimal de ce processus et μ_n sa mesure invariante. Nous supposons que ce générateur est de la forme :

$$Q_n = \tau_n \times Q$$

où Q est donné (matrice de Dayhoff). Sous cette hypothèse, nous avons $P^{\tau_n}_{a,b}(t) = P^1_{a,b}(t\,\tau_n)$ et $\mu_n = \mu$ (i.e. ne dépend pas de n). Désignons P^1 par P. Pour un arbre, nous supposons que l'évolution est décrite par ce même processus le long de chaque branche et ce indépendemment des branches.

Avec ces notations, la fonction de vraisemblance pour les paramètres $\mathbf{t} = (t_i)_{1\leq i\leq 2K-3}$ et $\boldsymbol{\tau} = (\tau_n)_{1\leq n\leq N}$ s'écrit :

$$L(\mathbf{t}, \boldsymbol{\tau}) = \prod_{n=1}^{N} \mathbb{P}_{\tau_n}(\mathsf{y}_{1,n}, \mathsf{y}_{2,n}, \ldots, \mathsf{y}_{K,n})$$

où $\mathbb{P}_{\tau_n}(\mathtt{y}_{1,n}, \mathtt{y}_{2,n}, \ldots, \mathtt{y}_{K,n})$ est la loi des observations au site n des K séquences pour l'arbre considéré, pour l'exemple de la Figure 1 nous avons :

$$\begin{aligned}\mathbb{P}_{\tau_n}(\mathtt{y}_{1,n}, \ldots, \mathtt{y}_{4,n}) &= \sum_{a_6 \in \mathcal{A}} \mu_{a_6} P_{a_6,y_{3,n}}(t_3\,\tau_n)\, P_{a_6,y_{4,n}}(t_4\,\tau_n) \\ &\quad \times \Big(\sum_{a_5 \in \mathcal{A}} P_{a_6,a_5}(t_5\,\tau_n)\, P_{a_5,y_{1,n}}(t_1\,\tau_n)\, P_{a_5,y_{2,n}}(t_2\,\tau_n) \Big)\end{aligned}$$

Pour une formulation de cette dernière expression dans le cas général voir [1].

Nous nous intéressons qu'aux taux d'évolution relatifs, i.e. nous rajoutons une contrainte : $\frac{1}{N}\sum_{n=1}^{N} \tau_n = 1$. Nous avons donc en tout $(2K-3)+(N-1)$ paramètres libres ($\boldsymbol{\tau} = \{\tau_n\}_{1 \le n \le N-1}$).

Pour maximiser la log–vraisemblance $\ell(\mathtt{t}, \boldsymbol{\tau}) = \log L(\mathtt{t}, \boldsymbol{\tau})$ on utilise un algorithme d'optimisation alterné en $\mathtt{t}$ et $\boldsymbol{\tau}$. L'optimisation selon chacun de ces paramètres vectoriels se fait également par méthode alternée[1]. L'intérêt de cette méthode est de ne pas avoir à calculer la valeur du gradient, le calcul de la valeur de la (log–)vraisemblance est déjà très coûteux en temps !

3 Quelques résultats numériques

Les données utilisées dans cet exemple simple sont un alignement de 4 séquences de longueur 383 (séquences partielles) de la famille des `HSP70` (*Heat Shock Protein*). Les 4 séquences sont celles du bovin, de la souris, de l'homme et la drosophile.

L'arbre de maximum de vraisemblance pour le modèle simple de Dayhoff (sites i.i.d.) est représenté sur la Figure 2.

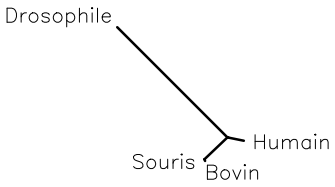

FIG. 2: Arbre optimisé.

La structure utilisée pour les comparaisons est celle de la séquence du bovin. Sur la Figure 3 nous avons représenté les taux d'évolution relatifs le long de la séquence en distinguant 4 catégories de structure secondaire : les hélices α, les brins β, les *turn* et le "reste". En nous limitant aux "trois types classiques" de structure secondaire (hélices α, brins 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, nous ne considérons que les sites pour lesquels nous observons une différence entre les quatre 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 clairement sur cet exemple que les zones très structurées (hélices α, brins β et *turn*) ont un taux d'évolution moyen nettement plus faible que le reste de la séquence et que leurs pourcentages de sites où l'on observe des différences est aussi plus faible.

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

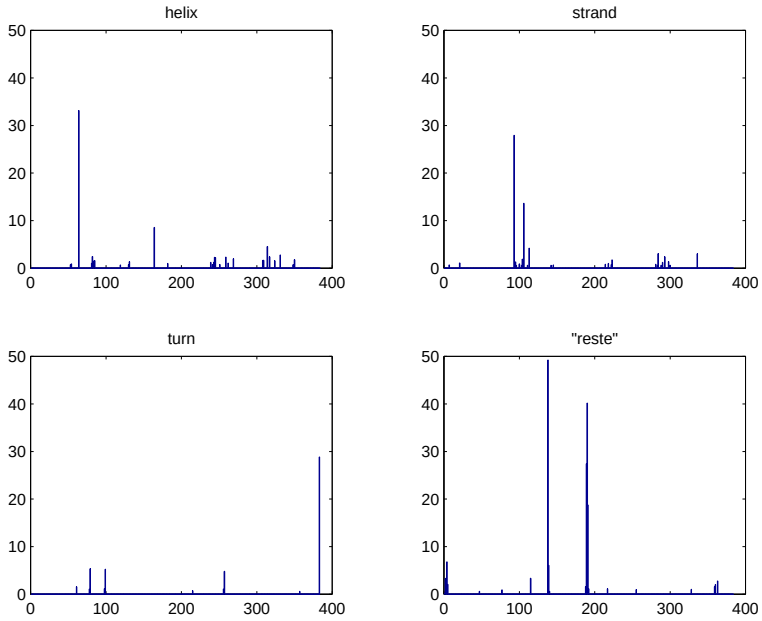

FIG. 3: Taux d'évolution relatifs le long de la séquence en distinguant 4 catégories de structure secondaire : les hélices α, les brins β, les *turn* et le "reste". Pour mieux analyser ses résultats il faut se reporter aux Tableaux 1 et 2.

catégorie	*nombre de sites*	*taux relatif moyen*
hélices	151	0.57
brins	110	0.67
turn	50	1.02
reste	*72*	*2.39*
total	383	1

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

4 Conclusions

Cette première étape semble plutôt encourageante mais il serait naturellement nécessaire de faire une étude plus systématique sur différents exemples.

Il est difficile de trouver des jeux de données qui conviennent à ce type d'étude car il nous faut connaître la structure décrite de manière suffisamment précise. Nous pourrions également nous intéresser aux relations possibles entre la variation du taux d'évolution et l'hydrophilie.

Par la suite, nous souhaiterions nous pencher sur la validité de l'hypothèse d'indépendance des sites pour ce problème et la pertinence de modèles plus sophistiqués le long de la séquence.

Références

[1] J. FELSENSTEIN, (1981), *Evolutionnary trees from DNA sequences : a maximum likelihood approach*, Journal of Molecular Evolution, 17, pp. 368–376.

[2] J. FELSENSTEIN AND G. A. CHURCHILL, (1996), *A hidden markov model approach to variation among sites in rate evolution*, Molecular Biology and Evolution, 13, pp. 93–104.

catégorie	*nombre de sites*	*pourcentage*	*taux relatif moyen*	*taux relatif médian*
hélices	30	19.87%	2.86	1.56
brins	27	24.55%	2.73	0.92
turn	10	20.00%	5.07	1.38
reste	*21*	*29.17%*	*8.19*	*2.01*
total	88	22.98%	4.34	1.35

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

[3] C. KELLY AND J. RICE, (1995), *Modeling nucleotide evolution : a heterogeneous rate analysis*, Mathematical Biosciences, 133, pp. 85–109.

[4] J. L. THORNE, N. GOLDMAN, AND D. T. JONES, (1996), *Combining protein evolution and secondary structure*, Molecular Biology and Evolution, 13, pp. 666–673.

[5] Z. YANG, (1993), *Maximum–likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites*, Molecular Biology and Evolution, 10, pp. 1396–1401.

[6] ——, (1995), *A space-time process model for the evolution of DNA sequences*, Genetics, 139, pp. 993–1005.