Statistique pour la bio-informatique

Séance 5 - 5 Novembre 2003

BLAST

1 Présentation de BLAST

Les programmes de comparaison de séquences ont pour but de repérer les endroits où se trouvent des régions identiques ou très proches entre deux séquences. Elles permettent de déduire les régions significatives correspondant à un sens biologique de celles qui sont observées par hasard.

La conception de l'algorithme BLAST (Basic Local Alignment Search Tool) s'appuie sur un modèle statistique dû à Karlin et Altschul (PNAS 1990, 1993). L'unité fondamentale de BLAST est le HSP (High-scoring Segment Pair). Il s'agit d'un couple de fragments identifiés de longueurs égales mais non prédéfinies sur chacune des séquences comparées et possédant un score significatif. Les régions similaires vont être repérées par une longue excursion de la marche aléatoire associée aux scores, et de fait un score cumulé important. Les méthodes statistiques de Karlin et Altschul sont appliquées pour déterminer la significativité de ces scores. Nous allons en décrire le détail dans des cas simples. La généralisation dépasse les objectifs de ce cours.

2 Loi des valeurs extrêmes

Nous commençons par un prérequis général sur les lois des valeurs extrêmes. Soit X_1, X_2, …, X_n, n variables aléatoires de fonction de répartition F. Nous nous intéressons à la loi de

$$X_{\max} = \max\{X_i\,,\, i = 1, \ldots, n\}.$$

Par ailleurs, nous définissons la loi de Gumbel de densité

$$\forall x \in \mathbb{R}, \quad f_W(x) = e^{-x-e^{-x}}.$$

Sa fonction de répartition est

$$\forall t \in \mathbb{R}, \quad F_W(t) = e^{-e^{-t}}.$$

Cette loi est souvent appelée *loi des valeurs extrêmes*. Un exercice classique (à faire une fois) permet de calculer l'espérance et la variance de la loi de Gumbel

$$E[W] = \gamma \qquad Var[W] = \pi^2/6$$

où γ est la constante d'Euler. Pour cela, on peut calculer la fonction génératrice de la loi

$$\varphi(t) = E[e^{tW}]$$

et se rappeler de quelques résultats classiques d'intégration

$$\Gamma'(1) = \int_0^\infty e^{-x} \ln x dx = -\gamma \qquad \Gamma''(1) = \frac{\pi^2}{6}.$$

Le résultat important concernant les valeurs extrêmes peut s'énoncer de la manière suivante.

Théoreme 2.1 *Soient $X_1, X_2, \ldots, X_n$, n variables aléatoires réelles continues indépendantes et de même densité ayant pour support $(A, +\infty)$ pour un $A \in \mathbb{R}$. Supposons de plus que les moments de tous ordres soient finis et notons $E[X_{\max}] = \mu_{\max}$ et $Var[X_{\max}] = \sigma_{\mathrm{m}}^2$. Alors, la suite de variables aléatoires*

$$W_n = \frac{\pi}{\sigma_{\mathrm{m}}\sqrt{6}}(X_{\max} - \mu_{\max}) + \gamma$$

converge en loi vers la loi de Gumbel.

La démonstration de ce résultat peut être trouvée dans de nombreux ouvrages et en particulier dans celui de Karlin et Taylor (1981). Pour les applications nous concernant, une situation particulièrement intéressante est le cas où

$$\mu_{\max} \approx c_1 \log n,$$

et

$$\sigma_{\mathrm{m}}^2 \approx c_2.$$

Dans cette situation, pour t raisonnablement proche de $\mu_{\max}$, nous pouvons utiliser l'approximation suivante

$$\mathrm{P}(X_{\max} \geq t) \approx 1 - e^{-e^{-\pi(t-\mu_{\max})/(\sigma_{\mathrm{m}}\sqrt{6})-\gamma}}.$$

À titre d'exercice et afin de mieux comprendre le résultat, examinons l'exemple de la loi du maximum de n variables aléatoires *iid* de loi exponentielle de paramètre λ. La fonction de répartition de cette loi est

$$\forall t \geq 0, \quad F(t) = (1 - e^{-\lambda t})^n.$$

Afin de calculer l'espérance et la variance de $X_{\max}$, notons qu'il est plus facile d'utiliser la loi du minimum et la propriété d'absence de mémoire de la loi exponentielle que de tenter de faire le calcul direct. Le résultat est (exercice)

$$E[X_{\max}] = \mu_{\max} = \sum_{k=1}^{n} \frac{1}{\lambda k} \approx (\gamma + \log n)/\lambda$$

et

$$Var[X_{\max}] = \sigma^2_{\mathrm{m}} = \sum_{k=1}^{n} \frac{1}{(\lambda k)^2} \approx \frac{\pi^2}{6\lambda^2}.$$

Lorsque l'on applique le résultat du théorème, on obtient

$$\mathrm{P}(X_{\max} \geq t) \approx 1 - e^{-ne^{-\lambda t}}.$$

Lorsque $t \approx (\gamma + \log n)/\lambda$, on retrouve bien la loi de Gumbel

$$\mathrm{P}(X_{\max} \geq t) \approx 1 - e^{-e^{-t}}.$$

D'autre part, le calcul direct montre que (exercice)

$$\mathrm{P}(X_{\max} \leq (t + \log n)/\lambda) = \left(1 - \frac{e^{-t}}{n}\right)^n \approx e^{-e^{-t}}.$$

Cela justifie que la bonne normalisation à considérer est

$$W = \lambda X_{\max} - \log n.$$

Cette normalisation apparaitra de manière naturelle dans le logiciel BLAST.

Dans la théorie de BLAST, les variables aléatoires jouant un rôle central sont discrètes et ont une queue de distribution “géométrique”

$$P(Y \geq t) \sim Ce^{-\lambda t}, \quad \text{pour} \ \ t \to \infty,$$

où C et λ sont des constantes positives. Dans ce cas, nous considérons, pour t grand, que

$$\mathrm{P}(Y_{\max} \geq t) \approx 1 - e^{-nCe^{-\lambda t}}.$$

Pour la loi géométrique de paramètre $p = e^{-\lambda}$, la valeur précédente (appelée p-value) est égale à

$$p\text{-value} \approx 1 - e^{-ne^{-\lambda t}} = 1 - e^{-e^{-\lambda s}}$$

où l'on renormalise de sorte que $s = t - \log n/\lambda$.

3 Marches aléatoires

Dans cette section, nous donnons le prérequis nécessaire concernant les excursions des marches aléatoires. *Nous discutons de la marche aléatoire simple* en détails. Les marches aléatoires intervenant en bio-informatique sont plus complexes, et liées à des matrices de scores biologiquement pertinentes dont l'élaboration mérite une discussion séparée. Toutefois, les outils nécessaires pour approfondir les résultats restent globalement analogue à ceux présentés dans le cas simple. Karlin et Dembo (1992) présentent la théorie asymptotique des marches aléatoires dans le cas général.

La marche aléatoire simple est une chaîne de Markov (S_ℓ) à valeurs dans un intervalle $[a, b]$ telle que

$$S_0 = h, \quad h \in [a, b],$$

et

$$\forall \ell \geq 1, \quad S_\ell = S_{\ell-1} + X_\ell$$

où (X_ℓ) est une suite de variables aléatoires iid vérifiant

$$\mathrm{P}(X_\ell = +1) = p = 1 - \mathrm{P}(X_\ell = -1).$$

Les questions que l'on se pose sont les suivantes.

1. Quelle est la probabilité pour que la marche aléatoire atteigne b avant a ?
2. Quelle est le temps moyen nécessaire pour atteindre a où b ?

Il est possible de répondre sans trop de difficultés à ces deux questions en utilisant la théorie des chaînes de Markov. Dans cette théorie, la question des temps d'atteinte est centrale et bien étudiée (voir par exemple Kemeny et Snell). L'approche adoptée ici est celle de la *fonction génératrice des moments* (transformée de Laplace), aussi utilisée dans le cas des marches générales.

3.1 Probabilités d'atteinte

Commençons par un résultat qui se démontre simplement.

Théoreme 3.1 *Soit Y une variable aléatoire discrète à valeurs dans $\mathbb{Z}$ et de transformée de Laplace*

$$m_Y(\theta) = E[e^{\theta Y}] = \sum_{k \in \mathbb{Z}} p_k e^{\theta k}.$$

On supppose que

1. *Y prend au moins une valeur négative $-a < 0$ (avec une probabilité $p_{-a} > 0$) ;*
2. *Y prend au moins une valeur positive $b > 0$ (avec une probabilité $p_b > 0$) ;*
3. *$E[Y] \neq 0$.*

Alors, il existe une unique solution $\theta^* \neq 0$ à l'équation

$$m(\theta) = 1.$$

Démonstration. Exercice. Utiliser la convexité de $m(\theta)$, le comportement à l'infini et la pente en 0 pour caractériser l'aspect du graphe de cette fonction. $\square$

Nous considérons désormais le déplacement D de la marche aléatoire à l'instant d'atteinte de a ou b. Cet instant est noté N et nous avons

$$D = S_N - h.$$

Notons τ_a et τ_b les temps d'atteinte de a et b respectivement. Le déplacement $D = D_N$ prend la valeur $a - h$ avec la probabilité

$$v_h = \mathrm{P}(\tau_a < \tau_b)$$

et la valeur $(b - h)$ avec la probabilité

$$w_h = \mathrm{P}(\tau_a > \tau_b).$$

Ainsi, nous avons

$$m_D(\theta) = v_h e^{(a-h)\theta} + w_h e^{(b-h)\theta}.$$

D'aprés le théorème précédent, il existe un unique θ^* tel que

$$m_D(\theta^*) = 1. \tag{1}$$

Ainsi, on obtient, pour cette valeur θ^*

$$\mathrm{P}(\tau_b < \tau_a) = \frac{e^{h\theta^*} - e^{a\theta^*}}{e^{b\theta^*} - e^{a\theta^*}}.$$

Pour finir, remarquons que le déplacement au pas ℓ

$$D_\ell = S_\ell - h = \sum_{k=1}^{\ell} X_k$$

admet pour fonction génératrice

$$m(\theta) = (pe^{\theta} + qe^{-\theta})^{\ell}.$$

Si θ^* est solution de

$$pe^{\theta^*} + qe^{-\theta^*} = 1$$

alors, il est encore solution de l'équation (1). Un petit peu d'arithmétique permet de calculer, pour $p \neq 1/2$,

$$e^{\theta^*} = \frac{1-p}{p}.$$

3.2 Temps moyen d'atteinte

Considérons à nouveau la marche aléatoire simple démarrant au point h, ($a < h < b$). Notons encore N le nombre de pas nécessaires pour atteindre soit a, soit b. Nous avons

$$D = D_N = \sum_{k=1}^{N} X_k$$

Avec ces notations, nous admettons le résultat suivant.

Théoreme 3.2 *(Identité de Wald) Pour tout θ en lequel $m_X(\theta)$ est défini, nous avons*

$$E[m_X(\theta)^{-N} e^{\theta D_N}] = 1.$$

La démonstration de ce résultat très puissant peut être trouvé dans l'ouvrage classique de Karlin et Taylor (1975, p.264). Voyons maintenant comment il peut être utile. La dérivation des deux membres de l'identité

$$E[\frac{d}{d\theta} m_X(\theta)^{-N} e^{\theta D_N}] = 0,$$

conduit, en prenant $\theta = 0$, à la nouvelle identité

$$E[D_N] = E[\sum_{k=1}^{N} X_k] = E[N]E[X]$$

qui est tout à fait classique et intuitive. Par ailleurs, dans la section précédente, nous avons observé que

$$E[D_N] = v_h(a-h) + w_h(b-h).$$

Puisque $E[X] = p - q = 2p - 1$, nous obtenons pour $p \neq 1/2$,

$$E[N] = \frac{v_h(a-h) + w_h(b-h)}{2p-1}.$$

Exercice. Reprendre les calculs précédents pour la marche aléatoire dont le pas X_ℓ vérifie

$$\mathrm{P}(X_\ell = +1) = p = 1 - \mathrm{P}(X_\ell = -2).$$

4 Excursions

Dans la marche aléatoire, on dira qu'un point est un *point-échelle* s'il est le plus bas jamais atteint. Dans la figure 1, les points-échelle correspondent aux pas $\ell = 5, 8, 19$, etc. Une excursion de la marche aléatoire est une trajectoire qui relie deux points-échelle.

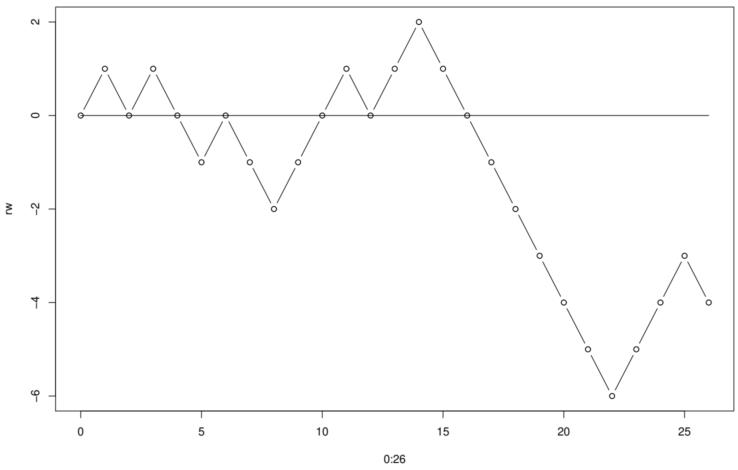

FIG. 1 – *Points échelles : $\ell = 5, 8, 19$, etc. La plus longue excursion est entre les points 8 et 19.*

Dans BLAST, il s'agira d'étudier les marches aléatoires partant de $h = 0$ et se terminant en $a = -1$. La barrière supérieure est $b = \infty$. Il est nécessaire de calculer deux grandeurs fondamentales

1. la loi de probabilité de la valeur cumulée des scores la plus élevée atteinte lors d'une excursion ;

2. la longueur moyenne de l'excursion, notée A dans les documentations du logiciel.

Afin de répondre à la première question, notons Y la valeur cumulée des scores la plus élevée atteinte lors d'une excursion. Nous supposons que le système de score est simple, c'est à dire que les scores sont $+1$ ou -1. Imaginons une barrière artificielle en $y \geq 1$. Si $h = 0$ et $a = -1$, les équations des sections précédentes conduisent à

$$\mathrm{P}(Y \geq y) = \frac{1 - e^{-\theta^*}}{e^{\theta^* y} - e^{-\theta^*}}.$$

Pour de grandes valeurs de Y, nous obtenons un comportement asymptotique de type géométrique

$$\mathrm{P}(Y \geq y) \sim Ce^{-\theta^* y}, \quad y \to \infty$$

où $C = 1 - e^{-\theta^*}$. Concernant la constante A, un raisonnement similaire montre que

$$A = \frac{1}{1 - 2p}.$$

5 Principe de BLAST

Nous nous contenterons de décrire une théorie sommaire de BLAST, nous limitant à la comparaison de deux séquences alignées sans indel (ou gap). Notre objectif est de décrire les principes de fonctionnement de la méthode (et donc ses limites) à partir de ce cas bien compris. Les généralisations concernent la recherche des meilleurs alignements locaux, ou le criblage d'une banque de données.

L'hypothèse nulle H_0 à tester est que *dans chaque paire d'acides aminés (ou de nucléotides) présents lors d'un appariement de deux séquences, les symboles ont été générés indépendamment*, c'est à dire que si l'on trouve l'acide A avec la probabilité p_A dans la première séquence, et l'acide B avec la probabilité q_B dans la seconde, l'appariement des deux symboles présents dans l'alignement survient avec la probabilité $p_A q_B$

$$H_0 : \quad \forall (i,j), \quad p(i,j) = p_i q_j$$

Lorsque l'hypothèse alternative est spécifiée, nous avons de plus

$$H_1 : \quad \forall (i,j), \quad p(i,j) = q(i,j)$$

Numérotons les positions des sites de l'alignement de 1 à m. À chaque position, un score est associé. Ce score prend la valeur $S(i,j)$ lorsque l'appariement (i,j) est observé. Cette valeur est traditionnellement issue d'une matrice de substitution. Nous supposerons, afin d'appliquer la théorie, qu'un élément $S(i,j)$ au moins de la matrice de substitution est strictement positif, et que le score moyen

$$E[S] = \sum_{i,j} S(i,j) p_i q_j$$

est négatif sous l'hypothèse nulle. Nous supposons de plus que les scores sont, dans leur ensemble, premiers entre eux.

Comme dans les sections précédentes, le score cumulé au site ℓ constitue une marche aléatoire, dont la dérive (sous H_0) est négative et qui va donc générer une suite d'excursions. Soit $Y_1, Y_2, \ldots$ les valeurs respectives des maxima atteints lors des excursions et posons

$$Y_{\max} = \max_{i=1,\ldots,n} Y_i, \quad n : \text{nombre d'excursions.}$$

Cette statistique est celle qui est effectivement utilisée dans l'algorithme BLAST pour tester l'hypothèse H_0. La théorie de Karlin et Dembo montre que les (Y_i) sont indépendantes et de distribution asymptotiquement géométrique

$$\mathrm{P}(Y \geq y) \sim C e^{-\theta^* y}.$$

La loi de $Y_{\max}$ dépend donc de constantes C, θ^* et de n le nombre d'excursions de la marche associée. Tous ces paramètres sont eux-même fonction de la matrice de substitution utilisée et des lois empiriques p et q. Voici de quelle manière sont calculés ces paramètres dans BLAST.

Calcul de θ^* Le paramètre θ^* se calcule numériquement comme la solution (unique) de l'équation

$$m_S(\theta^*) = \sum_{i,j} p_i q_j e^{\theta^* S(i,j)} = 1.$$

Calcul de A La longueur moyenne d'une excursion A est égale à

$$A = -\frac{\sum_{j \geq 1} j R_{-j}}{E[S]}.$$

où R_{-j} est la probabilité pour que la marche finisse en $-j$.

Calcul de n Puisque les séquences comparées sont de longueur m et que la longueur moyenne entre deux points-échelle est égale à A, une approximation raisonnable du nombre d'excursions est

$$n = \frac{m}{A}.$$

Calcul de C Nous ne rentrons pas dans les détails du calcul de la constante C. Elle découle des résultats sur les marches aléatoires générales de Karlin et Dembo

$$C = \frac{(1 - \sum_j R_{-j} e^{-j\theta^*})^2}{(1 - e^{-\theta^*}) A E[S e^{\theta^* S}]}$$

Il existe des algorithmes de calcul de ces probabilités. Dans ces cas suffisamment simples, il est parfois possible de les calculer *à la main*.

Calcul des p-values En utilisant les valeurs précédentes et la loi des extrêmes, il devient possible de donner des bornes asymptotiques concernant la distribution de $Y_{\max}$ sous l'hypothèse nulle. Notons que dans les sorties du logiciel BLAST, on utilise plutôt les notations

$$\lambda = \theta^*$$

et

$$K = \frac{C}{A} e^{-\lambda}.$$

Nous obtenons, pour t suffisament grand,

$$1 - e^{-Ke^{-\lambda t}} \leq \mathrm{P}(Y_{\max} > t + \log m/\lambda) \leq 1 - e^{-Ke^{-\lambda(t-1)}}$$

En utilisant le score normalisé

$$S' = \lambda Y_{\max} - \log(mK),$$

nous obtenons la p-value

$$\mathrm{P}(S' \geq s) \approx 1 - e^{-e^{-s}} \approx e^{-s}, \quad \text{pour} \quad s \to \infty.$$

Expect values Une grandeur cruciale à interpréter lors d'une sortie de BLAST est la valeur E appelée *Expect*. Cette valeur correspond au nombre moyen de segments de score maximal (HSPs). Puisque la probabilité pour que le score associé à une excursion dépasse la valeur $y_{\max}$ est

$$C e^{-\lambda y_{\max}}$$

et que le nombre d'excursions est $n = m/A$, nous avons

$$E \approx mK e^{-\lambda y_{\max}}$$

et donc la relation

$$S' = -\log E.$$

Cela permet d'exprimer la p-value sous la forme utilisée dans BLAST

$$p\text{-value} = 1 - e^{-E}.$$

Exercice. Calculer les grandeurs précédentes pour la marche aléatoire associée à

$$\mathrm{P}(X_\ell = +1) = p = 1 - \mathrm{P}(X_\ell = -2).$$

La théorie complète du calcul des p-values dont nous n'avons donné qu'une vue partielle se fonde sur l'approximation poissonienne ("clumping heuristic"). On peut trouver une description des démonstrations dans l'ouvrage de Waterman (1995) et une justification plus intuitive dans les cours de T. Speed.