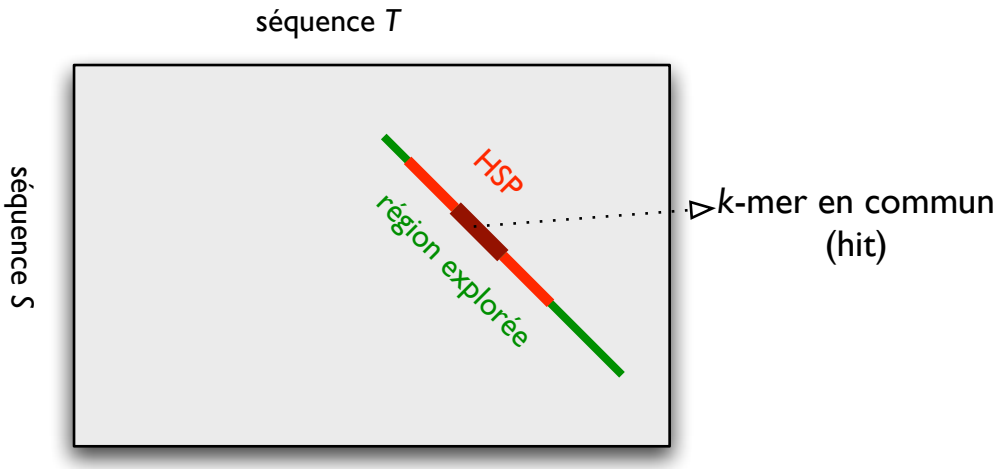
Alignement rapide : extension

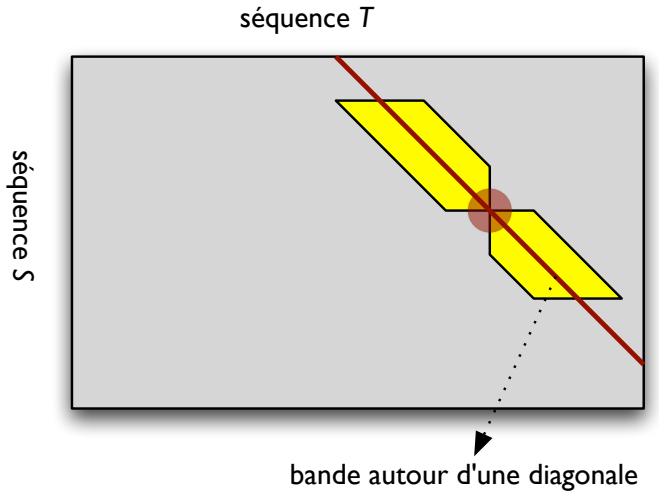
Extension rapide



Rester sur la même diagonale ; explorer jusqu'à ce que le score tombe sous un seuil t, prendre le meilleur segment (*high-scoring segment pair*, HSP)

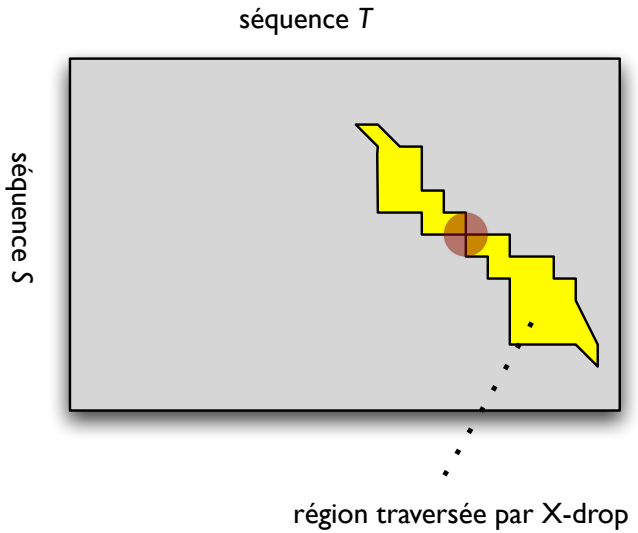
X-drop : explorer jusqu'à ce que le score tombe par un seuil X

PD dans une bande



Bande de $\pm d$ sommets proches de la diagonale D : $\{v_{i,j} : |(i-j) - D| \leq d\}$

X-drop



à partir d'une case initiale, explorer vers $v_{0,0}$ et $v_{|S|,|T|}$; arrêter si le score tombe par X

(exploration de toute une région ou quelques [même 1] diagonales)

Altschul et al, *Nucleic Acids Res.* 25 : 3389.

Détails : extension rapide [vers sud-est]

extension sur une diagonale :

ER1 **Entrée** i_0, j_0 départ de l'extension, s_0 score initial
ER2 meilleur $\leftarrow s_0$; extension $\leftarrow 0$
ER3 $i \leftarrow i_0 + 1$; $j \leftarrow j_0 + 1$; score $\leftarrow s_0$
ER4 **while** $i \leq |S|, j \leq |T|$, score ≥ 0
ER5 score $\leftarrow$ score $+ \mathbf{C}\begin{bmatrix} S[i] \\ T[j] \end{bmatrix}$
ER6 **if** score $\geq$ meilleur **then** meilleur $\leftarrow$ score, extension $\leftarrow j - j_0$
ER7 $i \leftarrow i + 1, j \leftarrow j + 1$
ER8 **return** meilleur, extension

$\mathbf{C}\begin{bmatrix} S[i] \\ T[j] \end{bmatrix}$: pénalisation d'un match/mismatch entre $S[i]$ et $T[j]$

Détails : bande [vers sud-est]

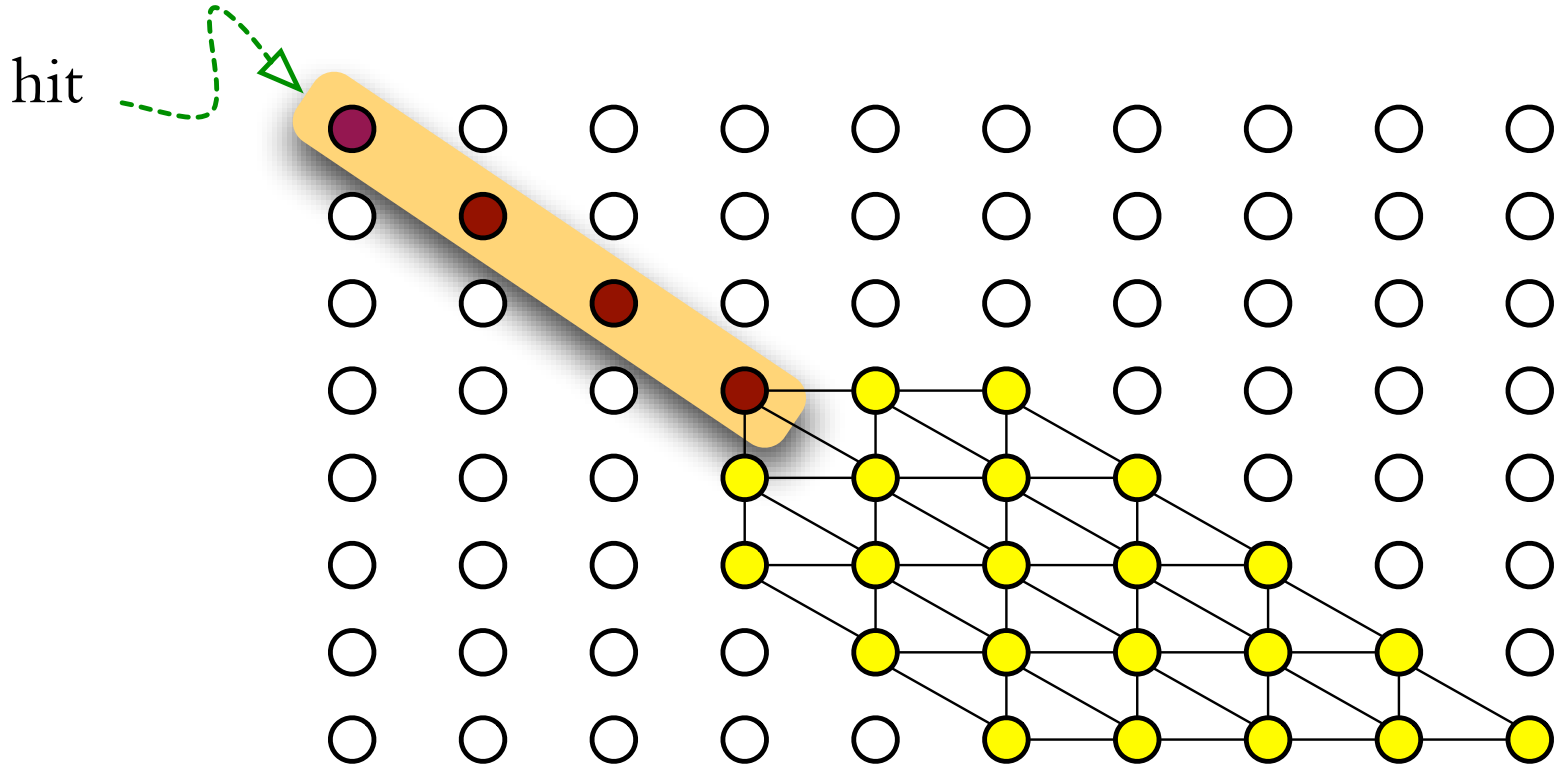
$A(i, j)$: score du meilleur alignement entre $S[i_0..i]$ et $T[j_0..j]$

A^* : meilleur score

B1 **Entrée** i_0, j_0 départ de l'extension, s_0 score initial, d épaisseur
B2 $A^* \leftarrow s_0, D \leftarrow i_0 - j_0$
B3 **for** $i \leftarrow i_0..|S|$
B4 **for** $j \leftarrow \max\{j_0, i - D - d\}..\min\{|T|, i - D + d\}$
B5 calculer $A(i, j)$
// récurrence usuelle avec $A(i-1, j-1)$, $A(i-1, j)$ et $A(i, j-1)$
B6 **if** $A^* > A(i, j)$ **then** $A^* \leftarrow A(i, j)$
B7 **if** $\forall j : A(i, j) \leq 0$ **then** sortir de la boucle i
B8 reporter A^*

Détails : bande 2 — graphe

Calcul en ligne B5 : programmation dynamique pour alignements restreints à la bande (ici $d = 2$)



Détails : X-drop

Idée : maintenir A^* score du meilleur alignement et ne pas continuer l'extension sur la bande de (i, j) si $A(i, j) < A^* - X$

Stocker col_g, col_d : colonnes de la dernière rangée que l'on a explorée.

(Ou code plus simple si l'exploration est dans une bande seulement : on n'a pas besoin de col_g, col_d)

[Code pour extensions vers sud-est seulement]

Détails : X-drop 2

XD1 **Entrée** i_0, j_0 départ de l'extension, s_0 score initial, X
XD2 $A^* \leftarrow s_0$, $\mathsf{col_g} \leftarrow j_0$, $\mathsf{col_d} \leftarrow |T|$
XD3 $i \leftarrow i_0$
XD4 **while** $i \leq |S|, \mathsf{col_g} \leq \mathsf{col_d}$
XD5 $\quad j \leftarrow \mathsf{col_g}$
XD6 $\quad$ **while** $j \leq \min\{\mathsf{col_d} + 1, |T|\}$
XD7 $\quad\quad$ calculer $A(i,j)$ // *récurrences usuelles*
XD8 $\quad\quad$ **if** $A(i,j) > A^*$ **then** $A^* \leftarrow A(i,j)$
XD9 $\quad\quad$ **if** $A(i,j) < A^* - X$ **then** $A(i,j) \leftarrow -\infty$
XD10 $\quad\quad$ $j \leftarrow j + 1$
XD11 $\quad$ **while** $\mathsf{col_g} \leq \mathsf{col_d}$ et $A(i, \mathsf{col_g}) = -\infty$, $\mathsf{col_g} \leftarrow \mathsf{col_g} + 1$
XD12 $\quad$ $\mathsf{col_d} \leftarrow \mathsf{col_d} + 1$; **while** $\mathsf{col_d} \geq \mathsf{col_g}$ et $A(i, \mathsf{col_d}) = -\infty$, $\mathsf{col_d} \leftarrow \mathsf{col_d} - 1$
XD13 $\quad i \leftarrow i + 1$
XD14 reporter A^*

Extension vorace

minimiser la distance d'édition :
pénalisation de mismatch = pénalisation d'indel = $+1$; match = 0

$$D(i,j) = \min\Big\{ D(i-1,j)+1, \\ D(i,j-1)+1, \\ D(i-1,j-1)+\{S[i] \neq T[j]\} \Big\}.$$

meilleure extension sur diagonale k avec δ erreurs :

$$R(\delta,k) = \max\Big\{ j : D(k+j,j) = \delta \Big\}$$

Extension vorace 2

boucle rapide pour trouver $R(\delta, k)$ à partir d'une cellule avec $D(k+j, j) = \delta$:

Algorithme EXTENSION(i, j)

E1 **while** $i < |S|$ et $j < |T|$ et $S[i+1] = T[j+1]$ **do**

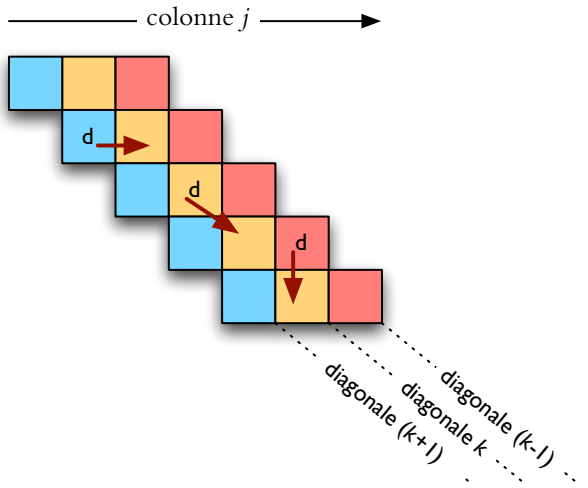
E2 $\quad j \leftarrow j+1; i \leftarrow i+1$

E3 $R \leftarrow j$ // *extension maximale sur la diagonale à partir de* (i, j)

Thm. Soit

$$j^* = \max\{\, R(\delta, k-1),\; R(\delta, k)+1,\; R(\delta, k+1)+1 \,\}$$

Alors $D(j^*+k, j^*) = \delta + 1$.



Extension vorace 3

(alignment entre $S[i..]$ et $T[1..m]$, avec $\leq \Delta$ erreurs)

Algorithme Glouton

G1 $k \leftarrow i - 1$

G2 initialiser $R(0, k) \leftarrow \mathsf{Extension}(k, 0)$; assumer $R(d, t) = -\infty$ pour tout autre d, t

G3 **if** $R(0, k) = |T|$, **then return** 0

G4 **for** $\delta \leftarrow 1, 2, \ldots, \Delta$ **do** // *avec $\delta \leq \Delta$ erreurs*

G5 **for** $t \leftarrow k - d..k + d$ **do** // *bande $\pm d$*

G6 $j \leftarrow \max\{R(\delta - 1, t - 1), R(\delta - 1, t) + 1, R(\delta - 1, t + 1) + 1\}$

G7 $R(\delta, t) \leftarrow \mathsf{Extension}(j + t, j)$

G8 **if** $R(\delta, t) = |T|$ **then return** δ

G9 **return** $\Delta + 1$ // trop de différences

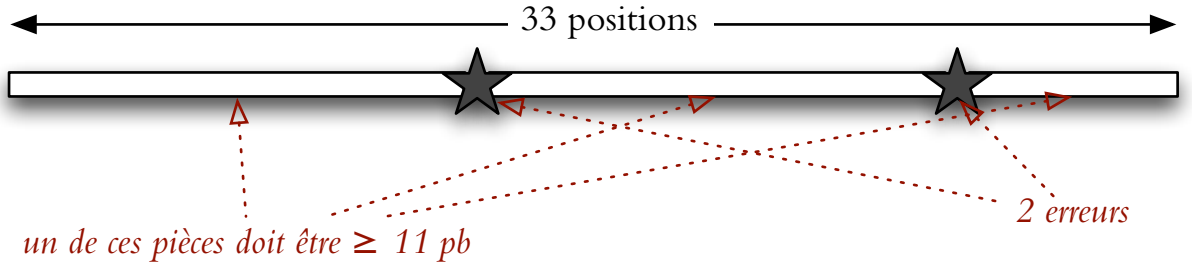
Temps de calcul : $O(m\Delta)$ au pire, $O(m + \Delta d)$ en moyenne

(Remarque : initialisation de $R(\delta, t)$ pour $t \neq k$ omise en G5/G6)

Peu d'erreurs : sensitivité de 100%

si on a $\leq \Delta$ erreurs dans une séquence de longueur ℓ et $k \leq \frac{\ell}{\Delta+1}$:
on a 100% de sensitivité dans seed-and-extend

(*lossless filtration*)



Mais k est trop petit : $\ell = 33, \Delta = 2 \Rightarrow k \leq 11$
$\rightarrow$ il vaut mieux utiliser plusieures graines espacées

ZOOM — mapping avec graines espacées

ZOOM [Lin et al 2008]

méthode : utiliser multiples graines, tableaux de hachage sur les séquences (*reads*), fenêtre glissant sur la référence

```
111111111111100000000000000000000
000000011111111111110000000000000
000000000000000000001111111111111
111111000000111111000000000000000
```

≤ 2 erreurs dans 33 bp permises avec des graines de poids 13 !

Sensitivité de 100% avec graines

Nombre minimal de graines pour 100% sensitivité ?

Table 1. The exact number of spaced seeds required and sufficient to detect up to two mismatches for each read length, at full sensitivity

Weight	Read Length											
	25	26	27	28	29	30	31	32	33	34	35	36
9	4	4	3	3	3	3	3	3	3	3	3	3
10	4	4	4	4	4	3	3	3	3	3	3	3
11	5	5	5	4	4	4	4	4	3	3	3	3
12	6	6	5	5	5	4	4	4	4	4	4	3
13	7	6	6	6	6	5	5	5	4	4	4	4
14			7	6	6	6	6	5	5	5	4	4
15					7	6	6	6	6	5	5	5
16							7	6	6	6	6	5

. . . nécessite bcp de graines pour séquences longues et ++ erreurs

(p.e., 9 graines de poids 14 pour 50-pb séquences et 4 erreurs)

Lin & al *Bioinformatics* 24 :2431 (2008)

Signification

Supposons qu'on a trouvé un alignement local. Est-ce que c'est par chance ou non ?

P-valeur : test de l'hypothèse nulle
H_0 : «alignement par chance»
H_1 : «alignement correspond à qqch importante»

Probabilité de H_0 donne la P-valeur : si elle est petite, on ne rejette pas l'hypothèse H_1.

$\Rightarrow$ on a besoin d'un modèle probabiliste pour calculer la probabilité de H_0.

Score de l'alignement

P-valeur pour un HSP avec valeur v

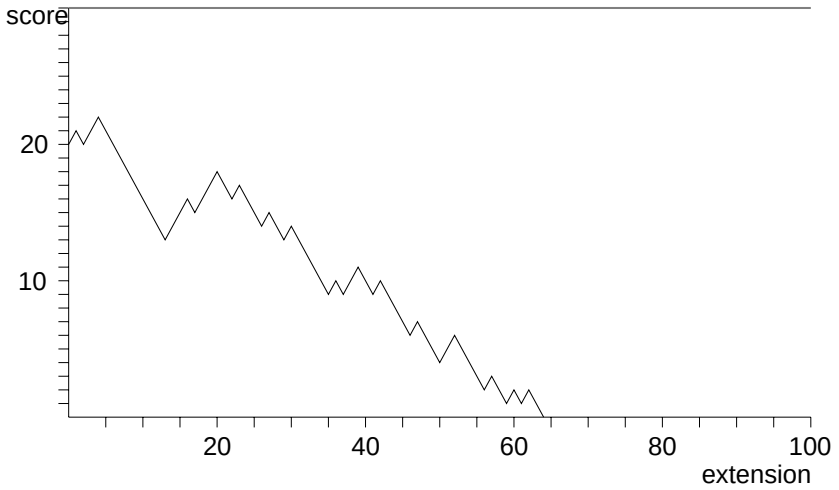
H_0 : HSP entre S et T donne un score aussi grand que v

modèle probabiliste : chaque caractère est choisi au hasard avec probabilités $p_{\mathtt{A}}, p_{\mathtt{C}}, p_{\mathtt{G}}, p_{\mathtt{T}}$

pondération de match/mismatch par une matrice $\mathbf{C}$

score d'une extension : marche aléatoire

Marche aléatoire



Pondération simple : $+1$ pour match, -1 pour subsitution

Score entre deux séquences aléatoires : $+1$ avec probabilité $p = \sum_{\sigma \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_\sigma^2$ et -1 avec probabilité $q = 1 - p$.

Évaluation d'un score v : quelle est la probabilité que l'extension atteint v avant 0 pour des séquences aléatoires ?

Problème de l'ivrogne

b bar, m maison, f fossé ($f < b < m$) ; X_t position en temps t

$$X_1 = b \qquad X_{t+1} = \begin{cases} X_t + 1 & \text{avec probabilité } p \\ X_t - 1 & \text{avec probabilité } q = 1 - p. \end{cases}$$

Définir $H_i(x) = \min\{j \geq i : X_j = x\}$ (*hitting time*)

On veut savoir la probabilité $P(b) = \mathbb{P}\Big\{H_1(m) < H_1(f) \;\Big|\; X_1 = b\Big\}$.

Noter que l'on peut décaler l'échelle du temps :

$P(b) = \mathbb{P}\Big\{H_i(m) < H_i(f) \;\Big|\; X_i = b\Big\}$

Ivrogne 2

On a $P(m) = 1$, $P(f) = 0$, et
$P(x) = pP(x+1) + qP(x-1)$ pour $f < x < m$
$\Rightarrow$ équation de récurrence linéaire homogène d'ordre 2 avec conditions initiales...

Deviner la solution : $P(x) = \alpha^x$

Trouver α [en utilisant que $q = 1 - p$]

$$\begin{aligned} \alpha^b &= p\alpha^{b+1} + q\alpha^{b-1} \\ 0 &= p\alpha^2 - \alpha + q \\ 0 &= (p\alpha - q)(\alpha - 1). \end{aligned}$$

Donc on a deux solutions $\alpha_1 = 1$ et $\alpha_0 = \frac{q}{p}$

Ivrogne 3

Solution intermédiaire : $P(x) = A\alpha_0^x + B\alpha_1^x$; trouver A et B en imposant $P(m) = 1, P(f) = 0$.

Solution finale avec $\alpha = \frac{q}{p}$:

$$P(x) = \frac{\alpha^x - \alpha^f}{\alpha^m - \alpha^f}.$$

Notez que pour $x = 0$, $f = -1$, $m \gg 0$, on a $P(x) \approx (1 - \alpha^{-1})\alpha^{-m}$ $\rightarrow$ signification de $(1 - r)r^v$ pour atteindre le score v avec $r = \alpha^{-1}$ — c'est la distribution géometrique : comportement typique des P-valeurs pour alignements.

Ivrogne tout confus

Taille de pas en temps t : variable aléatoire Z_t
on a $X_t = b + \sum_{i=1}^{t-1} Z_t$. (On utilisera $b = 0$ d'ici.)

Supposons que Z_t sont iid avec $\mathbb{P}\{Z = k\} = p_k$
où $k = -c, -c+1, \ldots, 0, 1, \ldots, d-1, d$; $c, d > 0$.

Récurrence : $P(x) = \sum_{k=-c}^{d} p_k P(x+k)$. Supposons que $P(x) = \alpha^x$.

Il faut trouver la solution α pour

$$1 = \sum_{k=-c}^{d} p_k \alpha^k$$

Ivrogne 4

On écrit $\alpha = e^\lambda$: donc on veut résoudre $G(\lambda) = 1$ où

$$G(\lambda) = \sum_{k=-c}^{d} p_k e^{k\lambda}.$$

Thm. Si $\mathbb{E}Z \neq 0$, il existe exactement une solution réelle $G(\lambda) = 1$ avec $\lambda \neq 0$. Si $\mathbb{E}Z < 0$, alors $\lambda > 0$.

Preuve. On a $G(0) = 1$, $G'(0) = \mathbb{E}Z$, et $G''(\lambda) > 0$ pour tout λ. □

Pertinence pour alignements

Pour une pondération de substitutions par $\mathbf{C}$ et des séquences aléatoires, on a

$$G(\lambda) = \sum_{\sigma,\sigma' \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_\sigma \pi_{\sigma'} \exp\left(\lambda \mathbf{C}\begin{bmatrix}\sigma \\ \sigma'\end{bmatrix}\right)$$

Alignement local entre S et T :

Thm. L'espérance du nombre de HSPs qui satisfont H_0 est

$$E = K|S||T|e^{-\lambda v},$$

où K est une constante et λ est la solution de $G(\lambda) = 1$.

BLAST

```
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value
gi|6633805|ref|NM_005332.2|  Homo sapiens hemoglobin, zeta (...    80   5e-13
gi|14523048|ref|NG_000006.1|  Homo sapiens alpha globin regi...    80   5e-13
gi|183790|gb|J00182.1|HUMHBA1  Homo sapiens hemoglobin zeta ...    80   5e-13
...
>gi|6633805|ref|NM_005332.2|   Homo sapiens hemoglobin, zeta (HBZ), mRNA
          Length = 589
 Score = 79.8 bits (40), Expect = 5e-13
 Identities = 56/60 (93%), Gaps = 1/60 (1%)
 Strand = Plus / Plus
Query: 1  actccagtgcag-tgcctaccctgcggccatttctctgaccaagactgagaggaccatca 59
          |||||||||||| |||| ||||||| ||||| ||||||||||||||||||||||||||
Sbjct: 27 actccagtgcagctgcccaccctgccgccatgtctctgaccaagactgagaggaccatca 86
...
Lambda     K      H
    1.37    0.711     1.31
Matrix: blastn matrix:1 -3
```