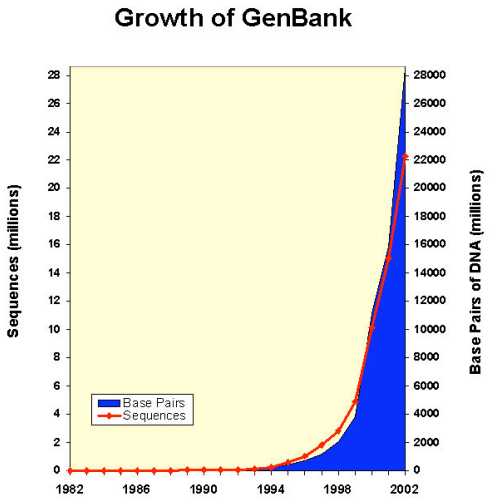
Recherche dans une banque de séquences

Bases de données

Bases de données de séquences : beaucoup d'information.

Exemple : GenBank

- 28 milliards de nucléotides ; 22 millions de séquences
- croissance exponentielle (taille doublée tous les 14 mois)



Recherche dans une BD

On a une séquence : trouver son occurrence dans la BD.

Deux problèmes :
- occurrences exactes
- occurences similaires

Problème algorithmique :
On a une séquence (courte) P et une séquence (longue) T :
trouver les occurrences [exactes ou similaires] de P dans T.

On a vu un algorithme [Alignements, page 17] qui prend temps $O(nm)$ pour $|T| = m$, $|P| = n$.

Est-ce qu'il y a des meilleures solutions ? — Oui !

Banques de données

NCBI : «National Center for Biotechnology Information» — États-Unis

Interface [Entrez] à plusieurs bases de données :

– séquences d'acides nucléiques
– séquences protéiques
– PubMed : publications
– structures
– taxonomie
– …

GenBank

Séquences d'ADN : GenBank (É-U), DDBJ (Japon), EMBL (Europe)

GenBank «flatfile» : exemple (HUMXIHB :

```
LOCUS       HUMXIHB                  458 bp    mRNA    linear   PRI 14-JAN-1995
DEFINITION  Human zeta hemoglobin mRNA, complete cds.
ACCESSION   M24173
VERSION     M24173.1  GI:340391
KEYWORDS    zeta-globulin.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 458)
  AUTHORS   Cohen-Solal,M.M., Authier,B., deRiel,J.K., Murnane,M.J. and
            Forget,B.G.
  TITLE     Cloning and nucleotide sequence analysis of human embryonic
            zeta-globin cDNA
  JOURNAL   DNA 1 (4), 355-363 (1982)
  MEDLINE   83182021
   PUBMED   6963223
COMMENT     Original source text: Human erythroleukemia cell line K562, cDNA to
            mRNA, clones 1 (1g7-8), 2 (4p7-7), and 3 (5a3-3).
```

GenBank - champs 1

`LOCUS`

– 1–10 caractères alphanumériques ; jadis l'identificateur de la séquence (p.e. l'abbreviation du gène), préservée pour compatibilité seulement.
– longueur et type de la séquence (DNA, mRNA, tRNA, rRNA)
– code de la division (p.e. PRI) et date de dernière modification.

`DEFINITION` «sommaire» de la séquence : espèce et le nom de la séq

`ACCESSION` nombre d'accession : clé dans la base de donnée. Identificateur unique parmi les BDs. Forme AA999999. L'accno est généré automatiquement lors de la soumission d'une séquence à la BD.

GenBank - champs 2

KEYWORDS et SOURCE : moins d'importance

VERSION donne ⟨accno⟩.⟨version⟩ et gi : identificateur de GenInfo. Ce sont des identificateurs des *séquences* (qui peut changer pour le même accno).

REFERENCE

GenBank — exemple cont.

```
FEATURES             Location/Qualifiers
     source          1..458
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /map="16p13.3"
     gene            1..458
                     /gene="HBZ"
     CDS             30..458
                     /gene="HBZ"
                     /note="zeta hemoglobin"
                     /codon_start=1
                     /protein_id="AAA61306.1"
                     /db_xref="GI:340392"
                     /db_xref="GDB:G00-119-302"
                     /translation="MSLTKTERTIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYF
                     PHFDLHPGSAQLRAHGSKVVAAVGDAVKSIDDIGGALSKLSELHAYILRVDPVNFKLL
                     SHCLLVTLAARFPADFTAEAHAAWDKFLSVVSSVLTEKYR"
BASE COUNT       80 a    173 c    127 g     78 t
ORIGIN      79 bp upstream of BglII site.
        1 actccagtgc agctgcccac cctgccgcca tgtctctgac caagactgag aggaccatca
          .......
      421 cggtcgtatc ctctgtcctg accgagaagt accgctga
//
```

GenBank - champs 3

FEATURES annotation de la séquence : un «feature» comprend un mot-clé, sa position, et des qualifieurs

position : sous-mot [p.e., `2..280`], entre deux bases [p.e., `91^92`], `...`,
et opérations : `complement(·)`, `join(·,...,·)`

mots-clé :

- `source` information taxonomique
- `CDS` partie traduite en une séquence protéique
- `exon`, `intron`, `gene`
- `repeat_region`
- `...`

Exemple : U96726

GenBank - entrées virtuelles

Exemple : U00089

```
LOCUS       U00089                816394 bp    DNA     circular CON 06-DEC-2002
DEFINITION  Mycoplasma pneumoniae M129, complete genome.
...
CONTIG      join(AE000016.2:1..19313,AE000015.2:59..17535,AE000014.2:22..12521,
            AE000013.2:53..10328,AE000012.2:59..10228,AE000011.2:59..15387,
            ... [plusieurs lignes]
            AE000019.2:59..10270,AE000018.2:59..11147,AE000017.2:62..15963)
//
```

NCBI

GenBank «flatfile» généré automatiquement à partir des bases de données.

Entrez : interface integré : recherche par identificateurs, mots clés, auteurs, etc.

BLAST : famille d'outils pour trouver des occurrences inexactes d'une séquence P dans le «texte» T

choix de T : `nr`, `est`, `month`, etc.

BLAST

BLAST : recherche par hachage + théorie de probabilités pour alignements locaux

Hachage — idée principale :
pour trouver l'occurrence inexacte de P en T

1. fixer $w > 0$
2. comparer chaque sous-mot de longueur w de P avec ceux de T
3. extension des matches pour obtenir un alignement local entre P et T

$\Rightarrow$ on trouvera rapidement les alignements qui contiennent w matches consécutifs

Hachage

Pour un sous-mot S (séquence ADN) de longueur $|S| = w$, on calcule

$$\mathcal{F}(S) = \sum_{i=1}^{w} e(S[i])4^{w-i},$$

avec e : $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} \mapsto \{0, 1, 2, 3\}$.

On a donc $0 \leq \mathcal{F}(S) < 4^w$.

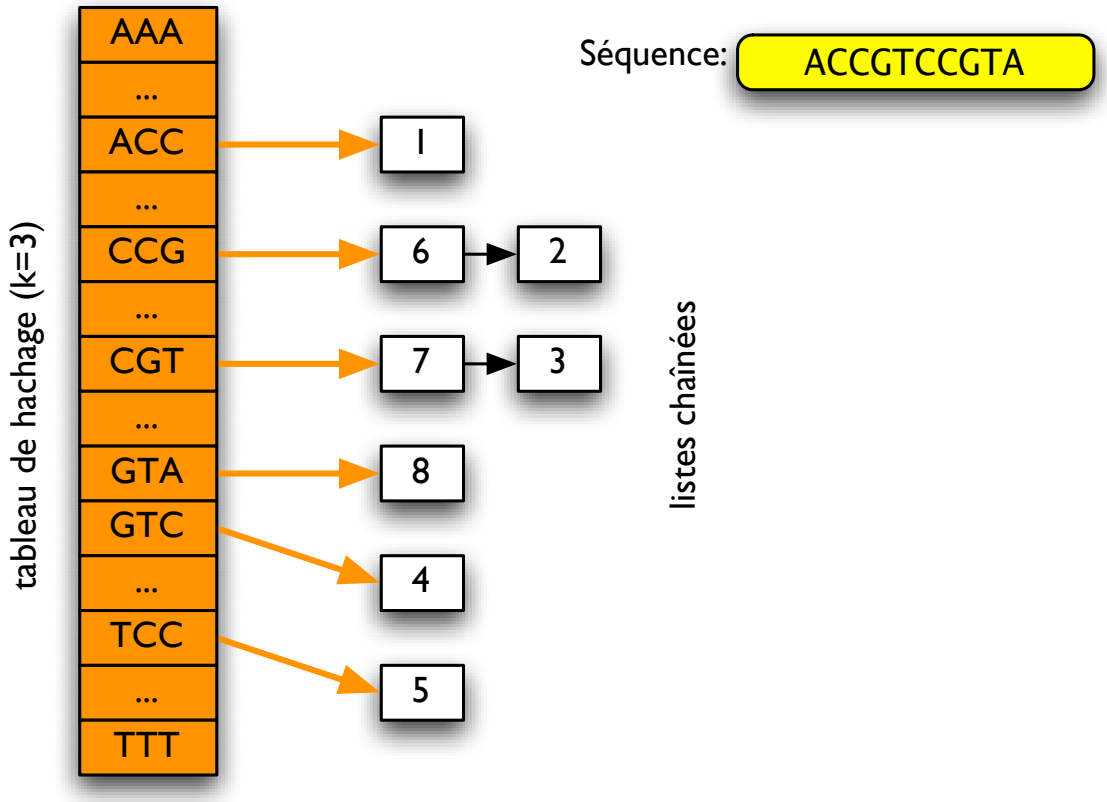
On remplit un tableau en calculant $\mathcal{F}(T[k..k + w - 1])$ pour chaque position k. Dans rangée t, se trouve la liste L_t des positions $k_1, k_2, \ldots$ pour lesquelles $\mathcal{F}(T[k_i..k_i + w - 1]) = t$.

Hachage 2

Structure de données pour le tableau (idée de base) :
arbre de recherche pour les rangées+liste chaînée pour chaque rangée

Implantation facile en Java : on peut utiliser les sous-mots comme clés directement, `Hashtable` calcule une sorte de $\mathcal{F}(\cdot)$ pour nous automatiquement

Tableau de k-mers



Hachage 3

- tableau pour T est calculé en avance.

- si $|T| \ll 4^w$, la plupart des rangées sont vides

- pour un P donné, on calcule $p_i = \mathcal{F}(P[i..i+w-1])$ pour $i = 1, \ldots, |P| - w + 1$

- *hits* : $\bigcup_i \{p_i\} \times L_{p_i}$: paires de positions où on a un match

- extension d'un *hit* dans les deux extrémités jusqu'à ce que la valeur de l'alignement tombe trop bas (ou simplement < 0)

Nombre de hits

Temps de calcul déterminé par nombre de hits : amélioration de l'ordre $|\Sigma|^w$ sur PD naïve montrée par thm suivant.

Thm. Le nombre moyen de hits entre deux séquences aléatoires avec $w \ll |P|, |T|$ est $\frac{|P| \cdot |T|}{|\Sigma|^w}$.

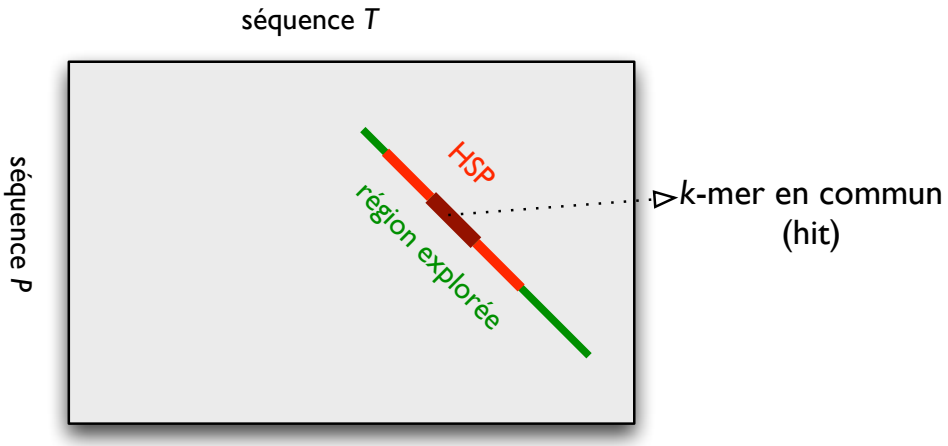
Preuve. Soit $m = |\Sigma|$ ($m = 4$ pour ADN), et soit $t \in \Sigma^w$ un w-mer. Probabilité qu'on le voit en une position fixe : $1/m^w$. Nombre moyenne de positions dans T où on le voit : $|T|/m^w$. (En fait, c'est $\frac{|T|-w+1}{m^w}$ mais $w \ll |P|, |T|$). De façon similaire, le nombre d'occurrences de t en P est $|P|/m^w$. Donc, le mot t génère $\frac{|P| \cdot |T|}{m^w}$ hits. Sommation sur $t \in \Sigma^w$ donne le résultat du théorème. □

Extension d’un hit

Techniques :

– extension rapide sur une diagonale
– X-drop
– programmation dynamique dans une bande

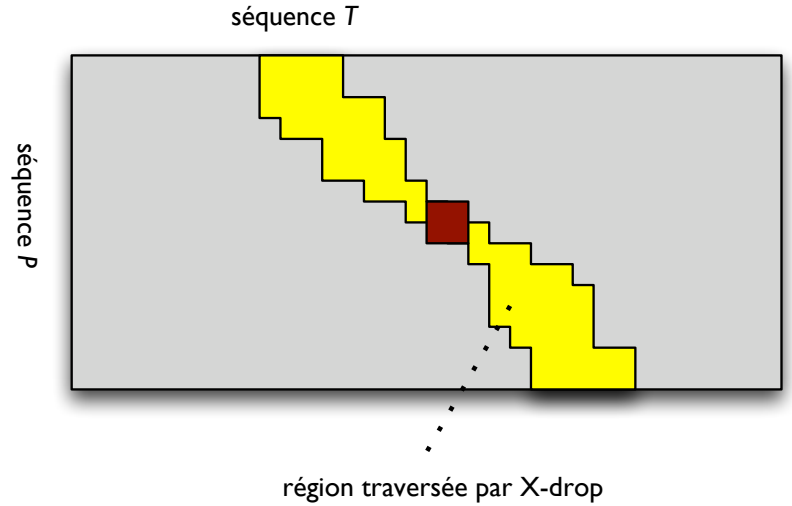
Extension rapide



Rester sur la même diagonale ; explorer jusqu'à ce que le score devienne 0, prendre le meilleur segment (*high-scoring segment pair*, HSP)

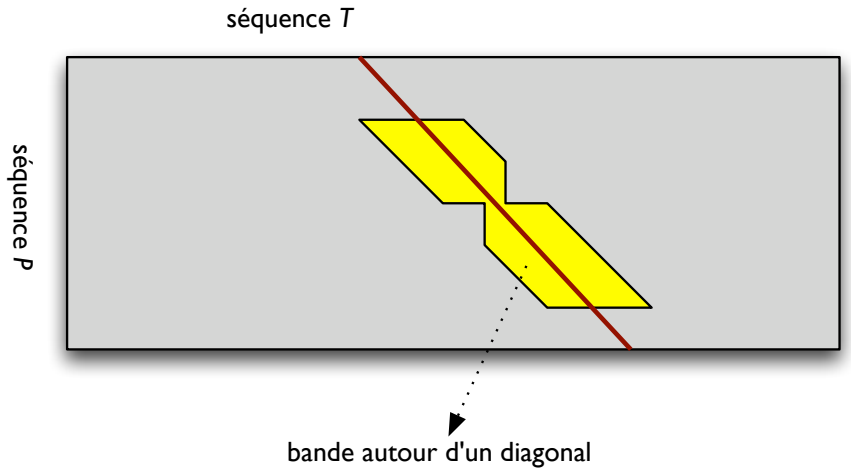
Altschul et al, *Nucleic Acids Res.* 25 : 3389. .

X-drop



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

PD dans une bande



Bande de diagonales $D \pm k$: $\{[i,j] : |(i-j) - D| \leq k\}$

temps de calcul : $O(k \cdot \min\{|P|, |T|\})$

Détails : extension rapide

[vers sud-est]

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 **tant que** $i \leq |S|, j \leq |T|$, score ≥ 0
ER5 score $\leftarrow$ score $+ \mathbf{C}[P[i], T[j]]$
ER6 **si** score $\geq$ meilleur **alors** meilleur $\leftarrow$ score, extension $\leftarrow j - j_0$
ER7 $i \leftarrow i + 1, j \leftarrow j + 1$
ER8 **retourner** meilleur, extension

Détails : X-drop

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

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

[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 $A^* \leftarrow s_0$, $\text{col}_\text{g} \leftarrow j_0, \text{col}_\text{d} \leftarrow |T|$
XD3 $i \leftarrow i_0$
XD4 **tant que** $i \leq |S|, \text{col}_\text{g} \leq \text{col}_\text{d}$
XD5 $j \leftarrow \text{col}_\text{g}$
XD6 **tant que** $j \leq \min\{\text{col}_\text{d} + 1, |T|\}$
XD7 calculer $A(i, j)$
XD8 **si** $A(i, j) > A^*$ **alors** $A^* \leftarrow A(i, j)$
XD9 **si** $A(i, j) < A^* - X$ **alors** $A(i, j) \leftarrow -\infty$
XD10 $j \leftarrow j + 1$
XD11 **tant que** $\text{col}_\text{g} \leq \text{col}_\text{d}$ et $A(i, \text{col}_\text{g}) = -\infty$, $\text{col}_\text{g} \leftarrow \text{col}_\text{g} + 1$
XD12 $\text{col}_\text{d} \leftarrow \text{col}_\text{d} + 1$; **tant que** $\text{col}_\text{d} \geq \text{col}_\text{g}$ et $A(i, \text{col}_\text{d}) = -\infty$, $\text{col}_\text{d} \leftarrow \text{col}_\text{d} - 1$
XD13 $i \leftarrow i + 1$
XD14 retourner A^*

Détails : X-drop 3

Calcul en ligne XD7 : pondération par $\mathbf{C}$

- si $i = i_0$ et $j > j_0$, alors $A(i,j) \leftarrow A(i,j-1) + \mathbf{C}[-, T[j]]$
- si $i > i_0$ et $j = \text{col}_\text{g}$ alors $A(i,j) \leftarrow A(i-1,j) + \mathbf{C}[P[i], -]$
- si $i > i_0$ et $j = \text{col}_\text{d} + 1$ alors $A(i,j) \leftarrow \max\Big\{A(i-1,j-1) + \mathbf{C}[P[i], T[j]], A(i,j-1) + \mathbf{C}[-, T[j]]\Big\}$
- sinon $A(i,j) \leftarrow \max\Big\{A(i-1,j) + \mathbf{C}[P[i], -], A(i-1,j-1) + \mathbf{C}[P[i], T[j]], A(i,j-1) + \mathbf{C}[-, T[j]]\Big\}$

Détails : bande

B1 **Entrée** i_0, j_0 départ de l'extension, s_0 score initial, k épaisseur
B2 $A^* \leftarrow s_0, i \leftarrow i_0, D \leftarrow i_0 - j_0$
B3 **tant que** $i \leq |S|$
B4 $j \leftarrow \max\{j_0, i - D - k\}$
B5 **tant que** $j \leq \min\{|T|, i - D + k\}$
B6 calculer $A(i,j)$; **si** $A^* < A(i,j)$ **alors** $A^* \leftarrow A(i,j)$
B7 $j \leftarrow j + 1$
B8 **si** $\forall j : A(i,j) \leq 0$ **alors** sauter à Ligne B10
B9 $i \leftarrow i + 1$
B10 reporter A^*

Détails : bande 2

Calcul en ligne B6 : pondération par $\mathbf{C}$

- si $i = i, j = j_0$, alors $A(i, j) = s_0$
- si $i = i_0$, et $j > j_0$, alors $A(i, j) = A(i, j-1) + \mathbf{C}[-, T[j]]$
- si $i > i_0$ et $j = \max\{j_0, i - D - k\}$, alors $A(i, j) = A(i-1, j) + \mathbf{C}[P[i], -]$
- si $i > i_0$ et $j < \min\{|T|, i - D + k\}$, alors $A(i, j) = \max\Big\{A(i-1, j) + \mathbf{C}[P[i], -], A(i-1, j-1) + \mathbf{C}[P[i], T[j]], A(i, j-1) + \mathbf{C}[-, T[j]]\Big\}$
- si $i > i_0$ et $j = \min\{|T|, i - D + k\}$, alors $A(i, j) = \max\Big\{A(i-1, j-1) + \mathbf{C}[P[i], T[j]], A(i, j-1) + \mathbf{C}[-, T[j]]\Big\}$

BLAST - 2

quelques mots de longueur w qui sont trop fréquents sont exclus de la recherche

extension de hits :
- aucun trou : match/mismatch seulement
- extension jusqu'à un seuil de v sur la valeur de l'alignement : l'extension à valeur maximale est choisie

résultat de la recherche : liste de paires de segments («high-scoring segment pairs» ou HSPs)

probabilités : signification — P-valeur

Signification

Supposons qu'on a trouvé une occurrence de P en T. Est-ce que c'est par chance ou non ?

P-valeur : test de l'hypothèse nulle
H_0 : «P apparaît dans T par chance»
H_1 : «l'occurrence de P dans T 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.

Signification - 2

Exemple : «pas tous les mots sont créés égaux» (P-valeur pour occurrence exacte)

Exemple : T est une séquence de longueur n «au hasard»
au hasard : chaque caractère de T est 0 ou 1 avec probabilités $\frac{1}{2}$-$\frac{1}{2}$.

Quelle est la probabilité de trouver $P = 00$ ou $P = 01$?

BLAST - 3

P-valeur pour un HSP avec valeur v

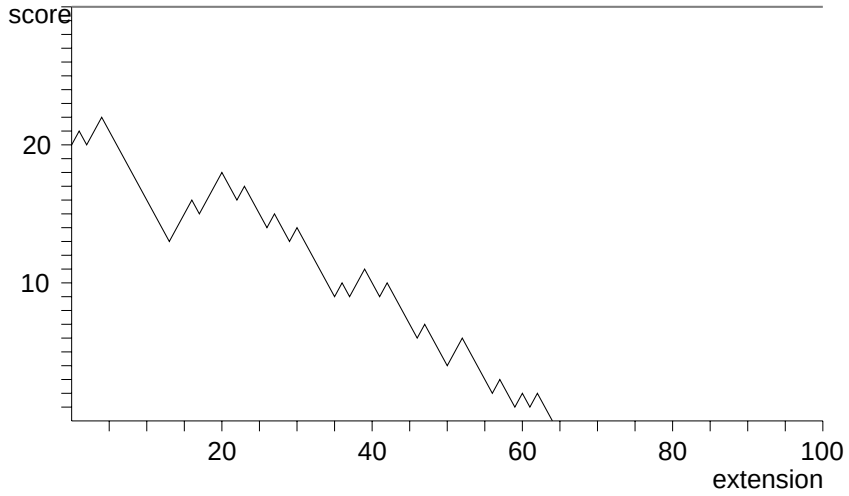
H_0 : HSP entre P et T donne un score aussi grand que v

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 $p = \sum_{\sigma \in \Sigma} \pi_\sigma^2$ 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 $H_i(x) = \min\{j \geq i \colon X_j = x\}$. On veut savoir la probabilité $P(b) = \mathbb{P}\Big\{ H_1(m) < H_1(f) \;\Big|\; X_1 = b \Big\}$.

Notez 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. □

Rélévance pour HSPs

Pour une pondération de substitutions par **C** et des séquences aléatoires, on a

$$G(\lambda) = \sum_{\sigma,\sigma' \in \Sigma} \pi_\sigma \pi_{\sigma'} \exp\Big(\lambda \mathbf{C}[\sigma, \sigma']\Big)$$

BLAST 4

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 - 5

De E à P-valeur :
nombre η de HSPs est une variable aléatoire : distribution de Poisson avec espérance E :

$$\mathbb{P}\{\eta \geq 1\} = 1 - \mathbb{P}\{\eta = 0\} = 1 - e^{-E} \approx E,$$

si $E \ll 1$.

(calcul un peu plus compliqué quand le même segment est aligné avec plusieurs autres dans l'autre séquence)

BLAST - 6

Correction pour une BD de longueur L : la BD contient plusieurs séquences, notre P aligne avec une de ses séquences est $(1 - e^{-E})$.

Espérance de séquences en BD avec lesquelles P aligne :

$$\mathsf{Expect} = \frac{\left(1 - e^{-E}\right) L}{\ell},$$

où ℓ est la longueur de la séquence qui contient le match. (Idée : comme si L/ℓ était le nombre total de séquences dans la BD)

P-valeur : $\approx 1 - \exp(-\mathsf{Expect})$.

BLAST - 7

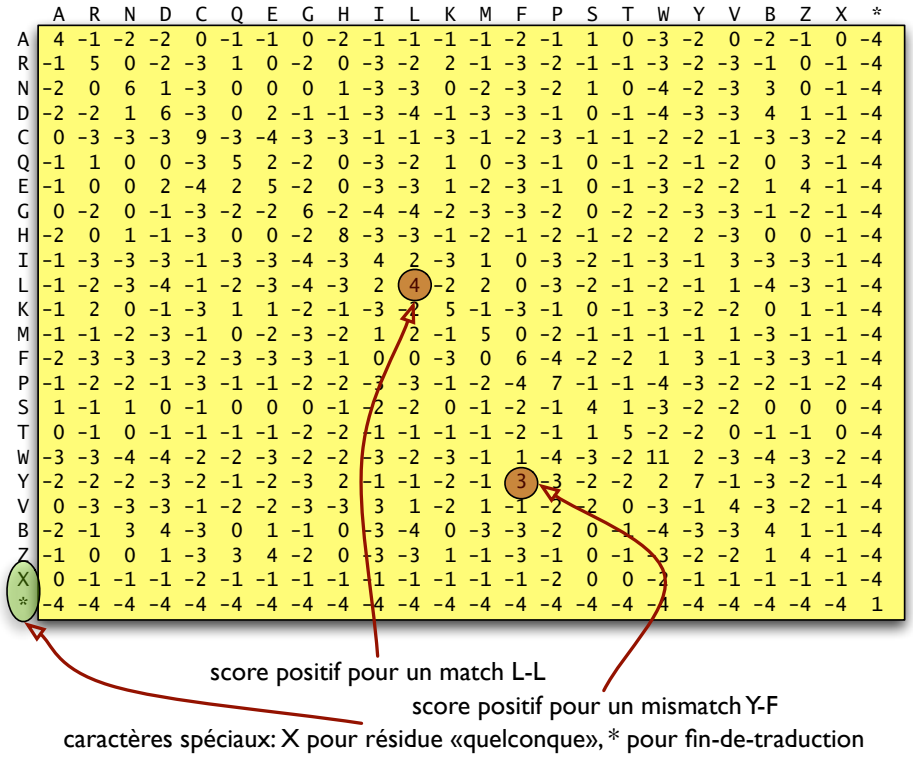
```
Sequences producing significant alignments:                          (bits) Value
gi|14336674|gb|AE006462.1|  Homo sapiens 16p13.3 sequence se...   119   4e-25
gi|14523048|ref|NG_000006.1|  Homo sapiens genomic alpha glo...   119   4e-25
gi|20379745|gb|BC027892.1|  Homo sapiens, hemoglobin, zeta, ...   119   4e-25
...
>gi|14336674|gb|AE006462.1|   Homo sapiens 16p13.3 sequence section 1 of 8
          Length = 258002
 Score =  119 bits (60), Expect = 4e-25
 Identities = 60/60 (100%)
 Strand = Plus / Plus
Query: 1      actccagtgcagctgcccaccctgccgccatgtctctgaccaagactgagaggaccatca 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 142880 actccagtgcagctgcccaccctgccgccatgtctctgaccaagactgagaggaccatca 142939
...
Lambda     K      H
    1.37    0.711     1.31

Matrix: blastn matrix:1 -3
```

Types de recherche

	nucléique	protéique	nucléique traduit
nucléique	`blastn`	`blastp`	
protéique		`blastp`	`tblastn`
nucléique traduit		`blastx`	`tblastx`

Alignement de protéines

Matrice BLOSUM62



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Alignement de protéines 2

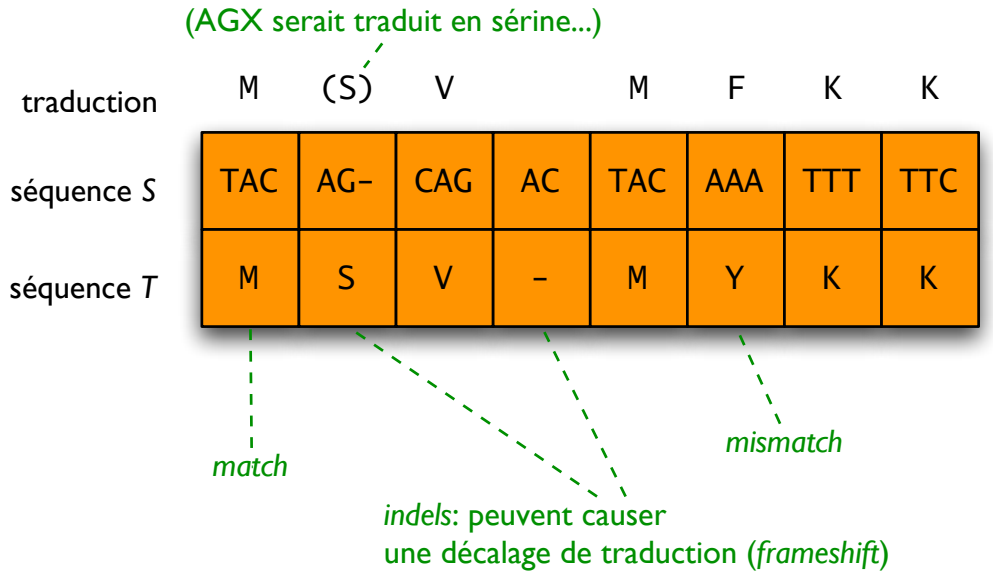
Identité typique de séquences cca. 10%, scores positifs pour mismatch $\Rightarrow$ «hit» défini par le score (au lieu d'identité)

On construit le tableau de hachage sur T ($k = 3, 4, 5$) ssont typiques pour longueur de sous-mots)

Pour trouver les matchs entre $w = P[j..j + k - 1]$ et le tableau, on considère tous les mots w' dans le tableau avec $\mathsf{score}(w, w') \geq H$

(p.e. $H = 20$).

Alignement d'une séquence nucléique et une séquence protéique



Alignement ADN-protéine

1. match/mismatch : score entre codon traduit et acide aminé donné par BLOSUM62, PAMx, . . .
2. trous : peuvent changer la «phase» de la traduction (*frameshift*)
3. sens de la traduction inconnue

2.+3. : tableau de hachage pour six séquences traduites (deux sens fois trois décalages)