Alignements optimaux

Traitement des insertions / délétions

- Optimisation de la comparaison → introduction des insertions/délétions de longueur variable
- Pénalité d'insertion : -1
- Pénalité d'insertion multiple : -1

- Score d'identité : +1
- Pénalité d'insertion : -1
- Pénalité d'insertion multiple : -1

```
A T G T A A T G C A T A

T A T G T G A A T
```

- Alignement optimal par glissement (score = 5)
- Alignement optimal avec une insertion (score = 6)

- Alignement optimal par glissement (score = 5)

```
  A T G T A A T G C A T A
  | | | |   |
T A T G T G A A T
```

- Alignement optimal par insertion

```
  A T G T - A A T G C A T A
  | | | |   | | |
T A T G T G A A T
```

Score = 7 – 1 = 6

La programmation dynamique

- Temps de 2 séquences de longueur N = N^2
- Détermination d'une insertion augmente le temps de 2N
- Programmation dynamique = N^2
- Principe : la plupart des évènements sont rejetés

L'algorithme de Needleman et Wunsch

- Needleman S. B. and Wunch C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443-453.

Comparaison réalisée avec matrice de substitution (ex: PAM250 de Dayhoff)

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-3
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	-5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

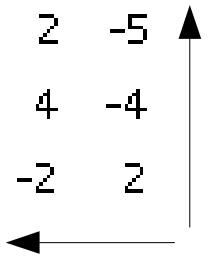
Construction d'une matrice « somme »

$$S(i,j) = se(i,j) + \max(S(x,y))$$

avec $x = i + 1$ et $j < y \leq n$

ou $i < x \leq n$ et $y = j + 1$

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-3
S	-1	1	0	0	0	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2



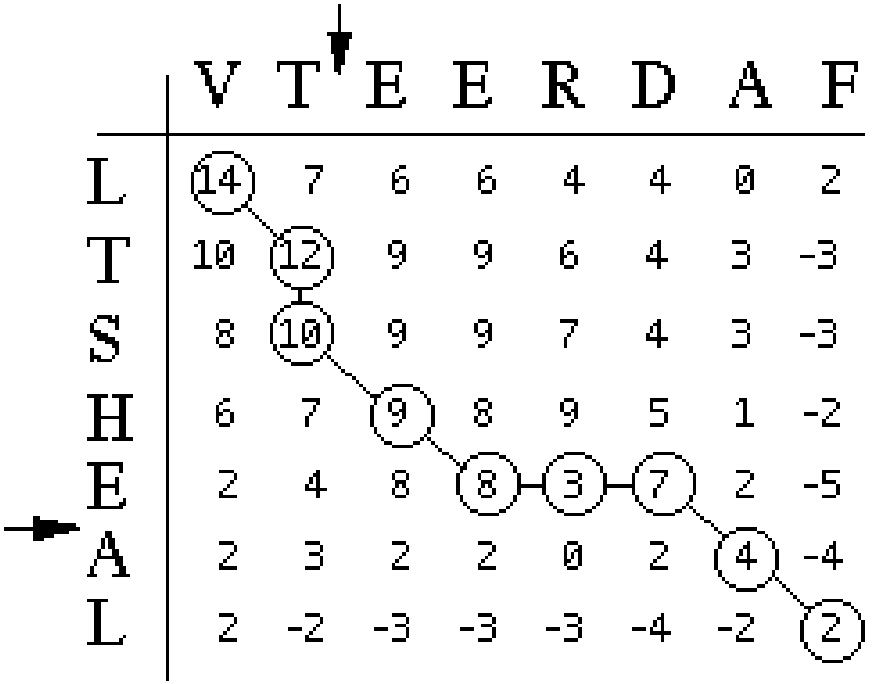
	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-3
S	-1	1	0	0	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

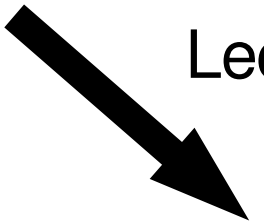
Construction d'une matrice « somme »

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	4	0	2
T	10	12	9	9	6	4	3	-3
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

Trouver le meilleur alignement

Trouver le chemin aux scores les plus élevés






```
VT-EERDAF
LTSHE--AL
```

Résultat de l'alignement

Calcul de la matrice « somme » avec pénalités

$$S(i,j) = se(i,j) + \max \left| \begin{array}{l} S(i+1,j+1) \\ S(x,j+1) - P \\ S(i+1,y) - P \end{array} \right| \quad \begin{array}{ll} \text{avec} & i+2<x\leq m \\ \text{et} & j+2<y\leq n \end{array} \quad (4)$$

P = pénalité due à insertion/délétion

Programmes utilisant cet algorithme

- ALIGN (Dayhoff *et al.*, 1979)
- GAP du logiciel GCG (Devereux *et al.*, 1984)
- EMBOSS : European Molecular Biology Open Software Suite
- BioJava, BioPerl, BioPython

Alignements globaux ou locaux

- alignement global

  ensemble des éléments des 2 séquences est prise en compte

  si longueur différente, incorporation d'insertions pour aligner les extrémités

- alignement local

  recherche de zones les plus similaires entre 2 séquences sans prédétermination de longueurs

  comporte une partie de chacune des séquences et non la totalité

L'algorithme de Smith et Waterman

- Smith T. F. and Waterman M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195-197.

Différences entre Needleman et Wunsch et Smith et Waterman

- N'importe quelle case de la matrice sert de point de départ pour le calcul des scores « somme ».

- Si score « somme » ≤ 0, alors progression stoppée.

Comparaison réalisée avec matrice de substitution
(ex: PAM250 de Dayhoff)

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-3
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	-5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

	V	T	E	E	R	D	A	F
L	2							2
T	0	3	0	0		0	1	
S		1	0	0	0	0	1	
H			1	1	2	1		
E		0	4	4		3	0	
A	0	1	0	0		0	2	
L	2							2

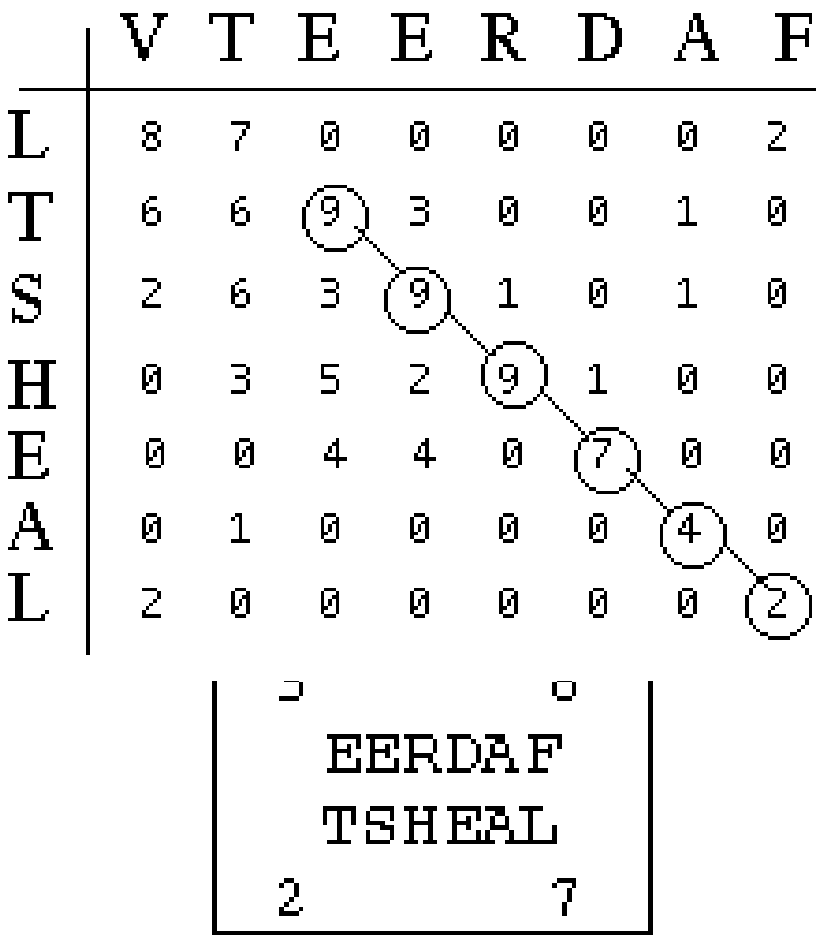
Construction d'une matrice « somme »

$$S(i,j) = \max\begin{cases} se(i,j) + S(i+1, j+1) \\ se(i,j) + S(x,j+1) - P \\ se(i,j) + S(i+1,y) - P \end{cases}$$

avec $i + 2 < x \leq m$

et $j + 2 < y \leq n$

	V	T	E	E	R	D	A	F
L	8	7	0	0	0	0	0	2
T	6	6	9	3	0	0	1	0
S	2	6	3	9	1	0	1	0
H	0	3	5	2	9	1	0	0
E	0	0	4	4	0	7	0	0
A	0	1	0	0	0	0	4	0
L	2	0	0	0	0	0	0	2

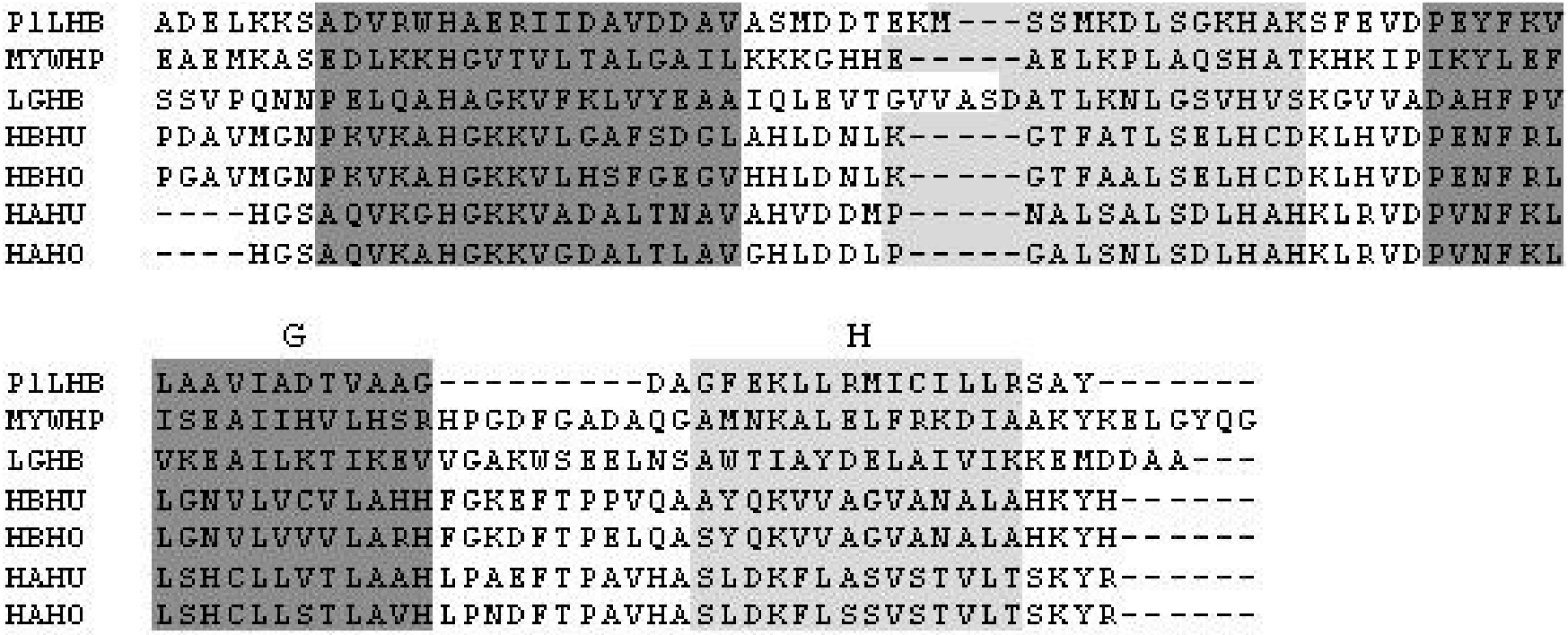


Résultat de l'alignement

Programmes utilisant cet algorithme

- EMBOSS : European Molecular Biology Open Software Suite
- BioJava, BioPerl, BioPython

Alignements de séquences identiques



Zones identiques entre deux séquences

Transcodage numérique des séquences:

- "mot" ou de "motif" pour les segments codés,
- la longueur des mots codés étant référencée comme *uplet* (triplet, quadruplet..) ou "*k-tuple*"

Le code Cx d'un motif Mx débutant en x a pour valeur:

$$Cx = P_1x4^{(n-1)} + P_2x4^{(n-2)} + \ldots + P_ix4^{(n-i)} + \ldots + P_nx4^0 + 1$$

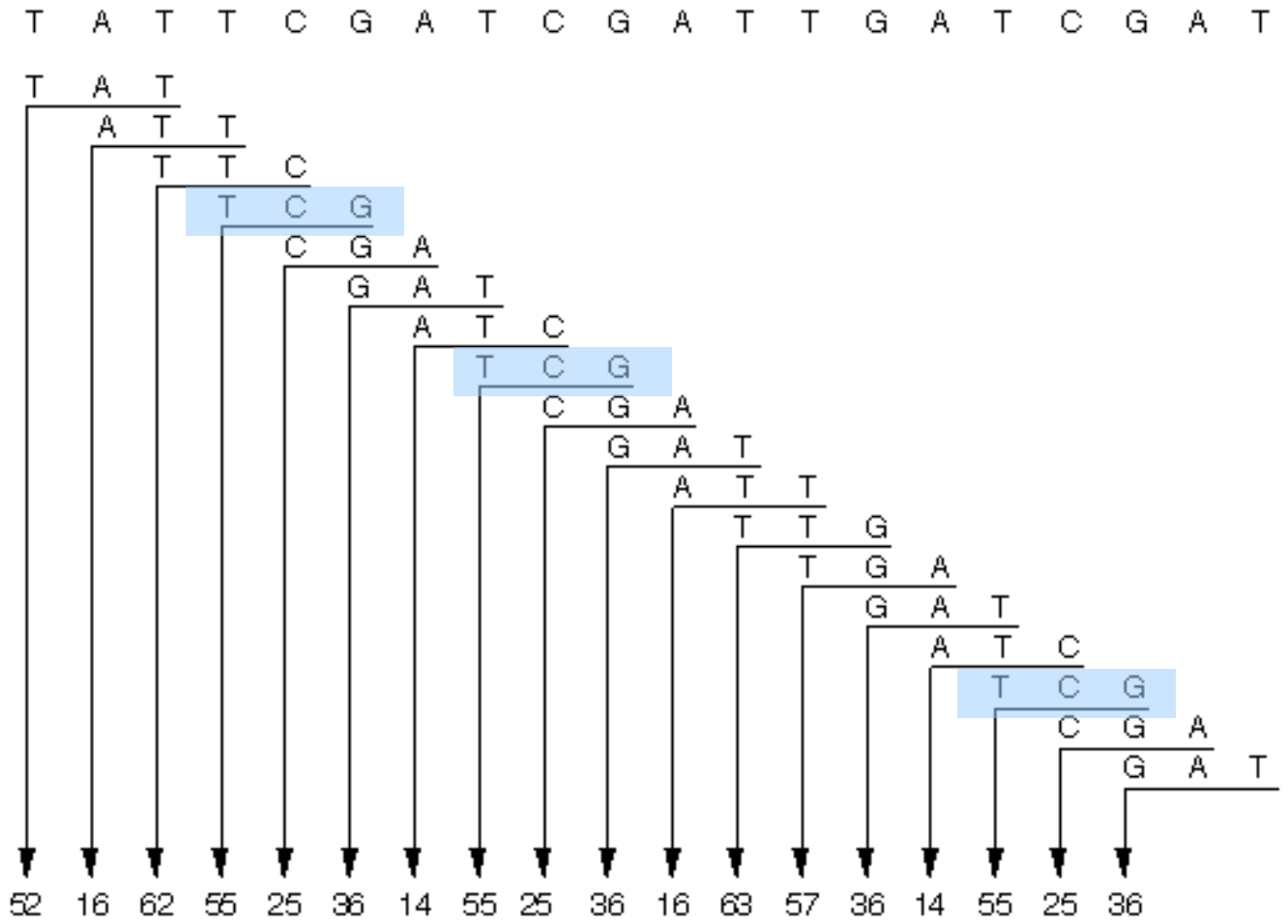
où n = longueur de l'uplet

4 : Nombre de lettres de l'alphabet

P_n: Valeur de la lettre à la position n dans l'uplet.

si n=3, on a $4^3 = 64$ motifs différents

$$Cx = P_1x4^2 + P_2x4^1 + P_3x4^0 + 1$$



Mot4 : TCG

code = $3x4^2 + 1x4^1 + 2x4^0 + 1 = 55$

Avantages

- Très rapide
- Calcul proportionnel à la longueur du *k-tuple.* + long, + résultat grossier.

Inconvénients

- Sensibilité moyenne. Un transcodage de longueur 5 peut ignorer des segments identiques de longueur 4.

Applications

- Recherche de régions répétées
- Localisation de sous-séquences ou motifs
- 1ère étape de programmes de recherche de similitude ou alignements, de comparaison avec banques de données.
- Permet d'éliminer rapidement séquences qui n'ont aucune ressemblance.

Evaluation des résultats

Signification biologique ?

Hasard ?

Méthodes pratiques et empiriques

- Ressemblance forte = relation évidente

- % identité
  - Longueur de la similitude
  - Type de séquence

  protéines: 25% identité pour 100 résidus ou + : ancêtre commun
  ac.nucléiques: 50% identité pour 100 pb : aucune relation biologique

  - Nombre d'insertions

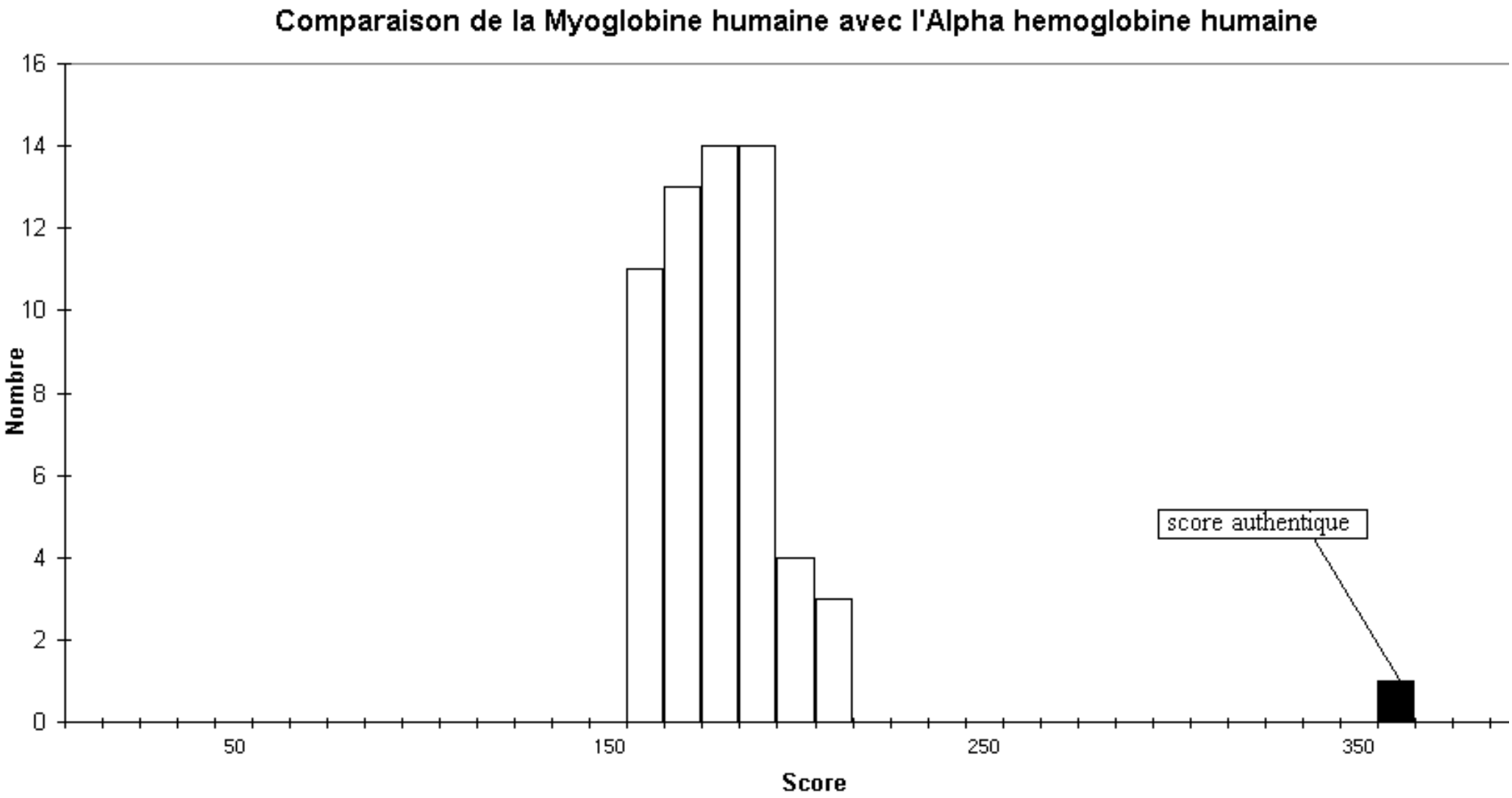
  ?

  + d'une insertion pour 20 résidus
  si changement de pénalités entraîne une modification de l'alignement

Méthode d'analyse de Monte Carlo

- La plus utilisée
- Principe:
  - Construire des séquences aléatoires avec même composition en bases
  - Comparaison avec séquences aléatoires → distribution de scores
  - Score authentique comparé aux autres
  - apprécier l'éloignement de la distribution aléatoire

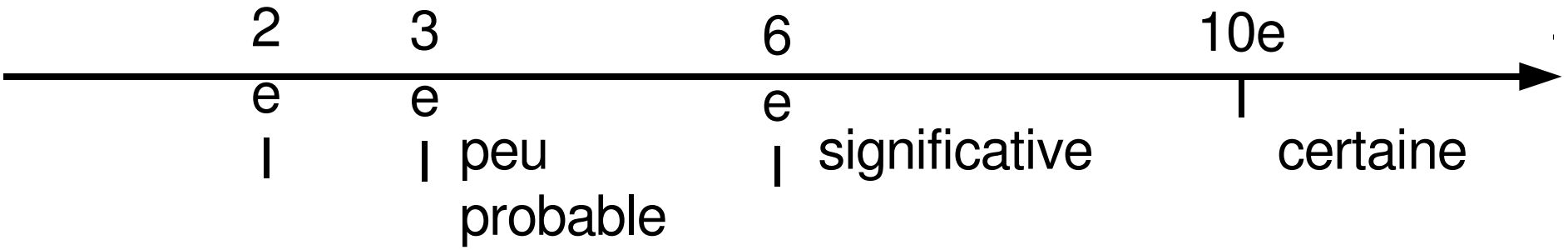
Méthode d'analyse de Monte Carlo



Méthode d'analyse de Monte Carlo

Z = (s -m) / e
avec

s : score authentique
m : moyenne des scores aléatoires
e : écart type des scores aléatoires



Méthode d'analyse de Monte Carlo

Inconvénients

- estimation significativité du score approximative
- programmes simulant séquences pas toujours adaptés
- coûteuse en temps de calcul (au moins 100 tirages par séquence)