Bioinformatique M1: Lecture 4

P. Derreumaux

ALIGNEMENT DE DEUX SEQUENCES:

1. Algorithmes
2. Evaluation des résultats

1.1. Définitions

Alignement : représentation

• Opérations élémentaires d'édition : opérations permettant de « passer » d'une séquence à une autre ;

• insertions (i) :

```
A A - B C A A
* *   * * * *
A A C B C A A
```

• délétions (d) :

```
A A B C A A
* *   * * *
A A - C A A
```

INsertion / DELétion
INDEL

• substitutions (s) :

```
A A B C A A
* *   * * *
A A C C A A
```

. Etant donné une matrice de scores avec un score de gap, le problème est de trouver l'alignement optimal qui maximise le score total.

The sequence alignment problem

```
1.  THESESENTENSESALIGN--NICELY      2
    |||||  ||  |    |||||  ||||||   19
2.  THESEQENCE----ALIGNEDNICELY      4

1.  THESESENTENSESALIGN--NICELY      2
    |||||    ||  | |||||  ||||||   19
2.  THESE-Q--ENCE-ALIGNEDNICLEY      4

1.  THESESENTENSESALIGN--NICELY      2
    |||  ||  ||  | |||||  ||||||   19
2.  THE--SEQ-ENCE-ALIGNEDNICLEY      4
```

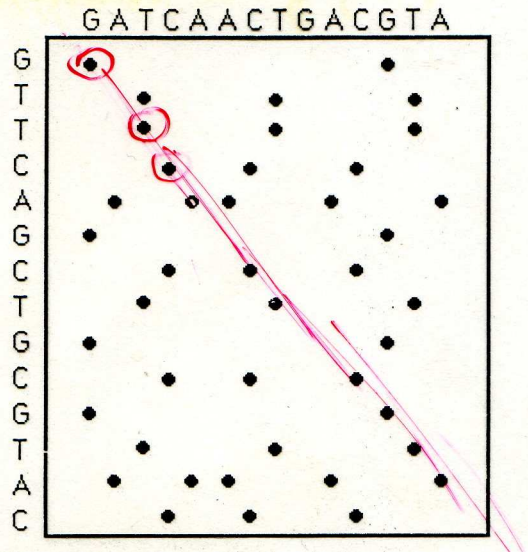
The gap parameters are optimized along with the substitution matrices (# PAM, BLOSUM; BLOSUM30#BLOSUM60)

1.2 Recherche de segments similaires : DOT PLOT

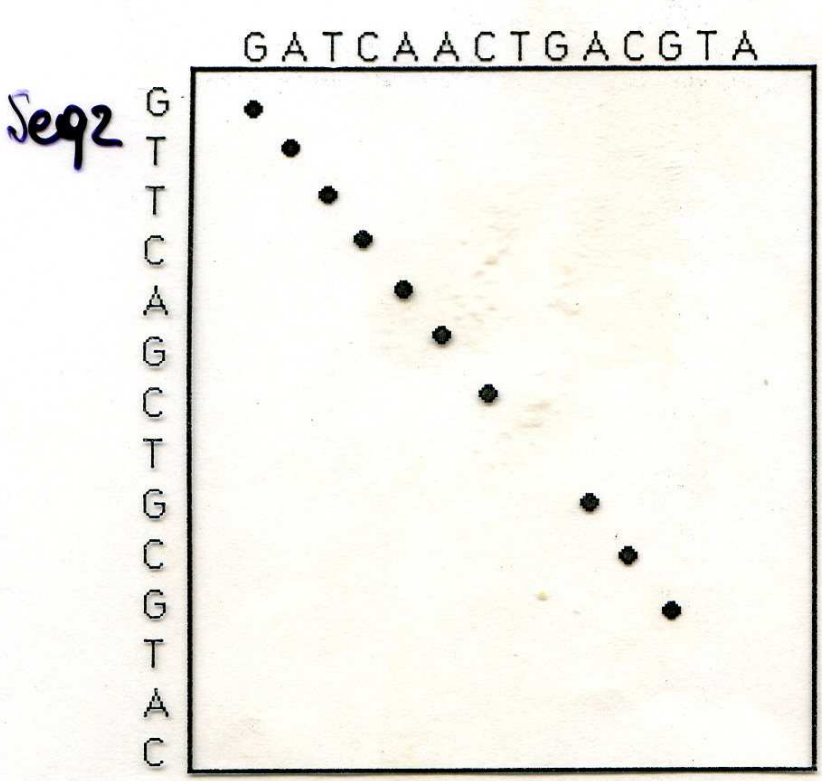
(without insertions and deletions)

```
GTCCTG GGCCACCTT VT
   GTG GGCCATCTT AL
```

Simple data



using 1 base and 100% identity score



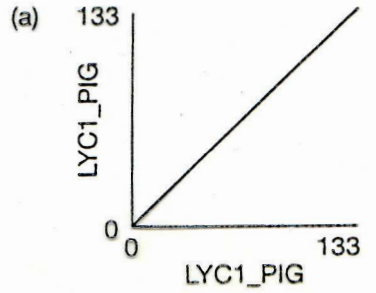
Using 4 bases and

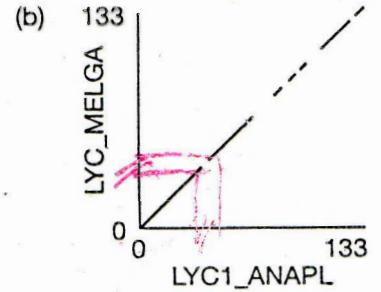
75% identity score

- Calculate a score within a window
- move the window

Seq1 GATC A:A

Seq2 GTTC A:G





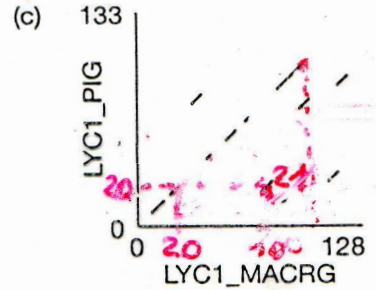


Figure 6.6 Graphical representation of dotplots, showing comparisons of (a) two identical sequences; (b) two highly similar sequences; and (c) two different, but related sequences.

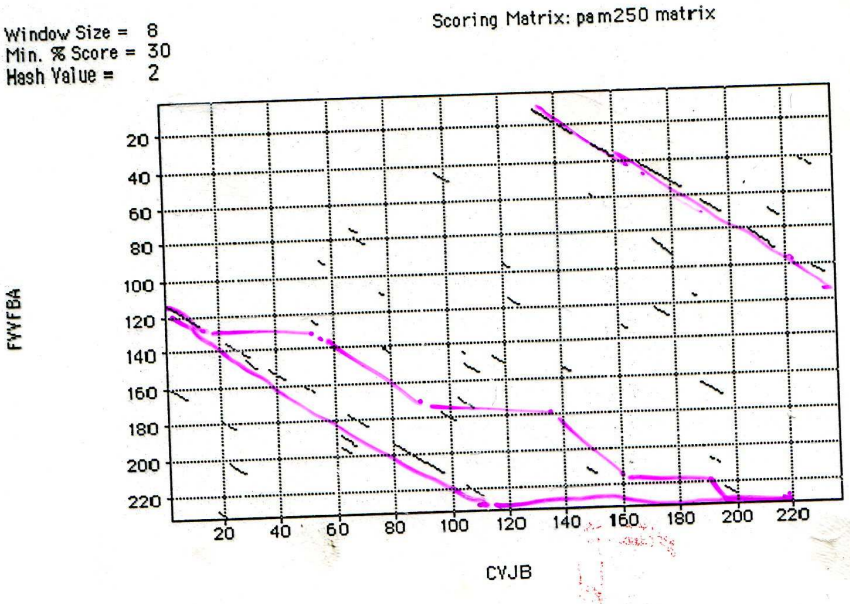


Fig 4. A sample dot plot of two 230 AA sequences

problems

- window length
- treshold score

better than alignments

- rearrangements
- repeated sequences

1.3 Recherche d'alignements optimaux.

• Insertions et deletions

$$P = O + n E$$
$$(a + n b)$$

O, E ?

alignement = f(P, scoring matrix)

Optimum gap and length extension penalties and alignment algorithm found using the training set

Matrix	Opn.	Ext.	Alignment algorithm
H3P2	3.0	0.3	GLOLOC
GONNET	3.0	0.08	GLOGLO
PAM250	2.0	0.1	GLOGLO
BLOSUM62	8.0	0.8	GLOGLO

Using the training set, we evaluated a range of parameters for each matrix. The parameters given in the Table maximize the area under the SENS-SPEC curve. The second column (Opn.) gives the gap opening penalty and the third column (Ext.) gives the gap length extension penalty. The range evaluated, GLOLOC, and GLOGLO are explained in the text.

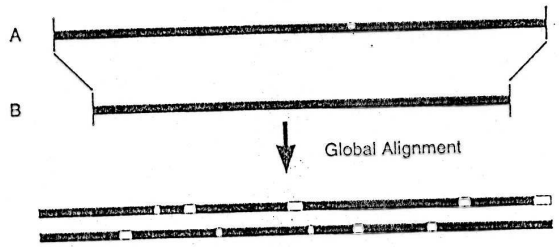
. (Programmation dynamique) dynamic programming algorithms provide a rigorous mathematical approach towards sequence alignment. Best (optimal) alignment of two sequences with n and m amino acids is found on the order of nm steps.

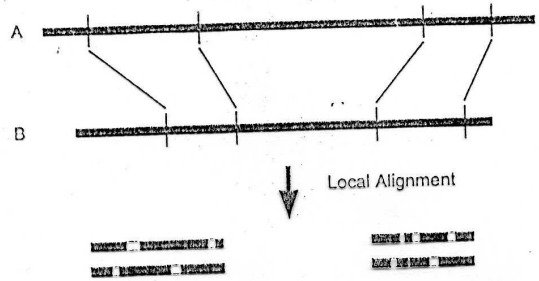
A rigorous method:
- no approximation
- all alignments are considered with all possible gaps at any positions

Sequence Matching

Dynamic Programming Alignment

- **Global vs local methods**
  - Global - uses all positions of both sequences
  - Local - best region of alignment
- **Use global if** → NW (Needleman, Wunsch)
  - You expect and want the entire sequence to match
- **Use local** → SW (Smith, Waterman)
  - You expect only part of the sequences to match
  - Sequences are very different in length — best when scanning database.





Global Alignment: NW Algorithm

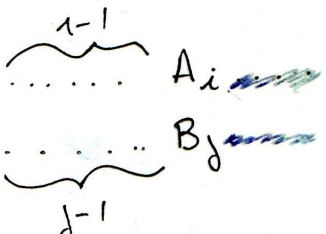
Two Sequences $A = (A_1, A_2, \dots A_m)$, $B = (B_1, B_2, \dots B_m)$

At each aligned position: 3 events
- $\sigma(A_i, B_j)$ substitution
- $\sigma(A_i, -)$ gap / insertion in B
- $\sigma(-, B_j)$

The optimal alignment is found by the recurrence relation for $1 \le i \le n$ and $1 \le j \le m$

$$V_{i,j} = \max \begin{cases} V_{i-1,j-1} + \sigma(A_i, B_j) & (1) \\ V_{i-1,j} + \sigma(A_i, -) & (2) \\ V_{i,j-1} + \sigma(-, B_j) & (3) \end{cases}$$

(1) Aligning A_i with B_j



Score = score of aligning $i-1$ elements of A with $j-1$ elements of B + $\sigma(A_i, B_j)$

(2) $V_{i-1,j} + \sigma(A_i, -)$

Aligning A_i with a space character in B

$\overbrace{\ldots}^{i-1} A_i$

$\underbrace{\ldots B_j}_{j} -$

Score = Score aligning $i-1$ of A with j elements of B
$+ \sigma(A_i, -)$

(3) $V_{i,j-1} + \sigma(-, B_j)$

Aligning B_j with a space character in A.

$\overbrace{\ldots A_i}^{i} -$

$\underbrace{\ldots\ldots B_j}_{j-1}$

Rmq 3 cells $(V_{i-1,j-1} ; V_{i-1,j} ; V_{i,j-1})$
are required for computing V_{ij}

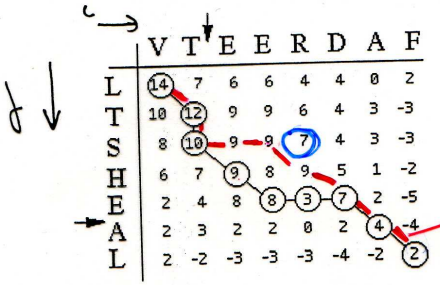
complexity : $O(nm)$

All possible solutions are represented by a 2D-matrix

Alignement optimal

- Détermination d'un chemin qui correspond au passage des scores les plus élevées en autorisant 3 types de mouvements à partir du score maximum :

  - mouvement diagonal (substitution, privilégié)
  - mouvement vertical (insertion gap sur seq en i)
  - mouvement horizontal (insertion —— en j)



Alignement Optimal

Seq1 VT-E
Seq2 LTSH
mais autres solutions

VT-EERDAF
LTSHE--AL

Résultat de l'alignement

Local Alignment: SW algorithm.

- Locates the best alignment between subregions of A and B. There may be a large number of distinct local alignments.

$$V_{i,j} = \max \begin{cases} (1) \\ (2) \\ (3) \end{cases} \text{NW}. \\ 0$$

1.4. Recherche de segments identiques

(used by FASTA)

- basée sur la codification numérique des séquences
  (chaîne de caractères → entiers)
- longueur des mots (uplet ex quadruplet)
  'k-tuplé'
- vitesse exécution = $f(\ell)$

  sensibilité = $\frac{1}{f(\ell)}$
- souvent utilisée pour similarité avec bases de données (élimination rapide)

2. Evaluation of results
(Applicable when scanning database)

Question: Biological significance
vs.
Random.

1st case: strong similarity between your seq and a seq with known structure.

Expected Topology is conserved, but surface loops (and thus alignment) may vary

Alignment is not accurate if INDELS within the core β-strands and α-helices.

2nd case: no similarity

↓

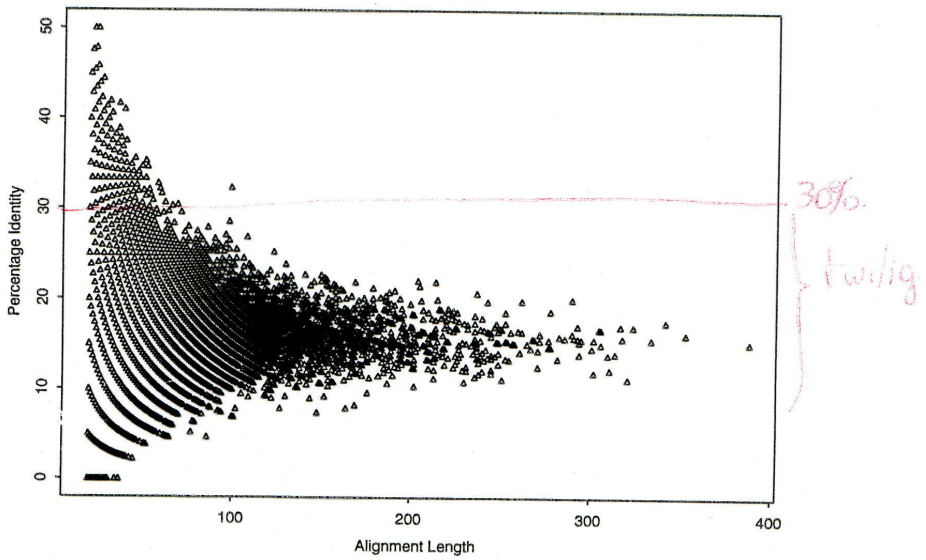
Emprical methods.

2.1 Empirical Methods

eg. Alignment flabitelity
- number of insertions (> 10%)
- alignment varies with small variation of P (± 10%)

Percentage Identity

— %id = f(length of alignment)



% id found for a large number of locally optimal alignments of differing length between proteins known to be of unrelated 3D structure

— %id = f(proteins, nucleic acids)

— RMSD = f(%id)

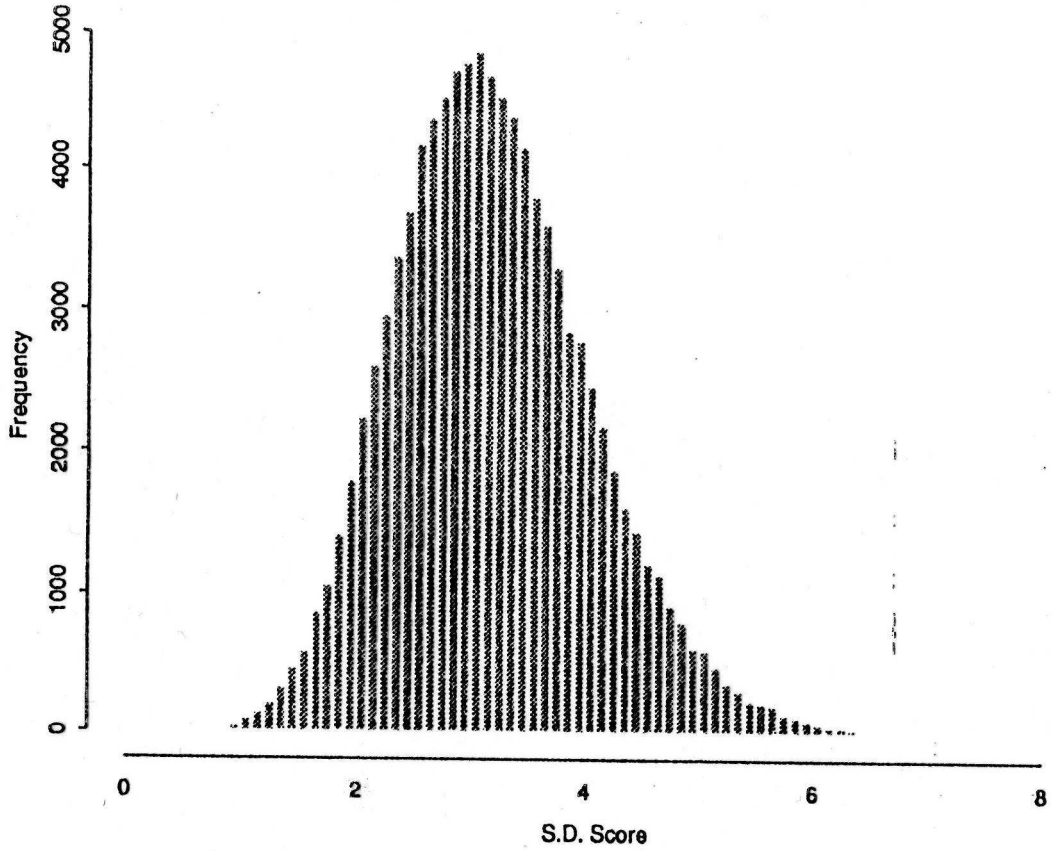


Figure 2. Distribution of SD scores obtained for 100 000 alignments of length > 20 between unrelated proteins. The SD scores were calculated from 100 randomizations using a global alignment method (13), PAM250 matrix with eight added to each element, and length-independent gap penalty of eight.

Figure 3. A local alignment found between citrate synthase (Brookhaven code: 2cts) and transthyritin (2paba). The SD score for this alignment is 7.55, its length is 54 residues, and the identity is 25.9%. Despite this apparently high similarity, the sequences are of completely different secondary structure.

2.2 Simulations pour alignements globaux et locaux

Seq Native → Seq Aléatoires avec la même composition AA.

S ⟷ comparaison. $S_{A_1}, S_{A_2}, S_{A_3}$

$$\text{Score } Z = \frac{S - \mu}{\sigma} \qquad \mu = \sum_{\text{Seq Aléa}}^{N} \frac{S_i}{N}$$

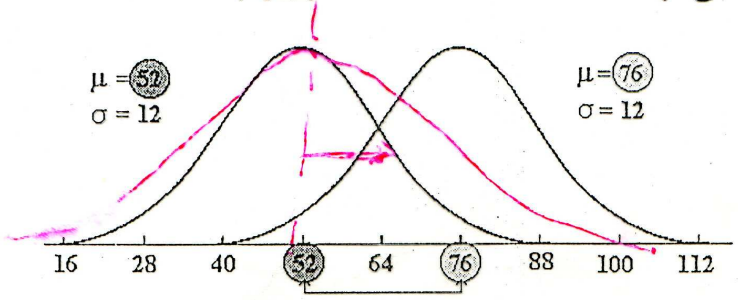
$$\sigma = \sum_{\text{Seq Aléa}}^{N} \frac{(S_i - \mu)^2}{N-1}.$$

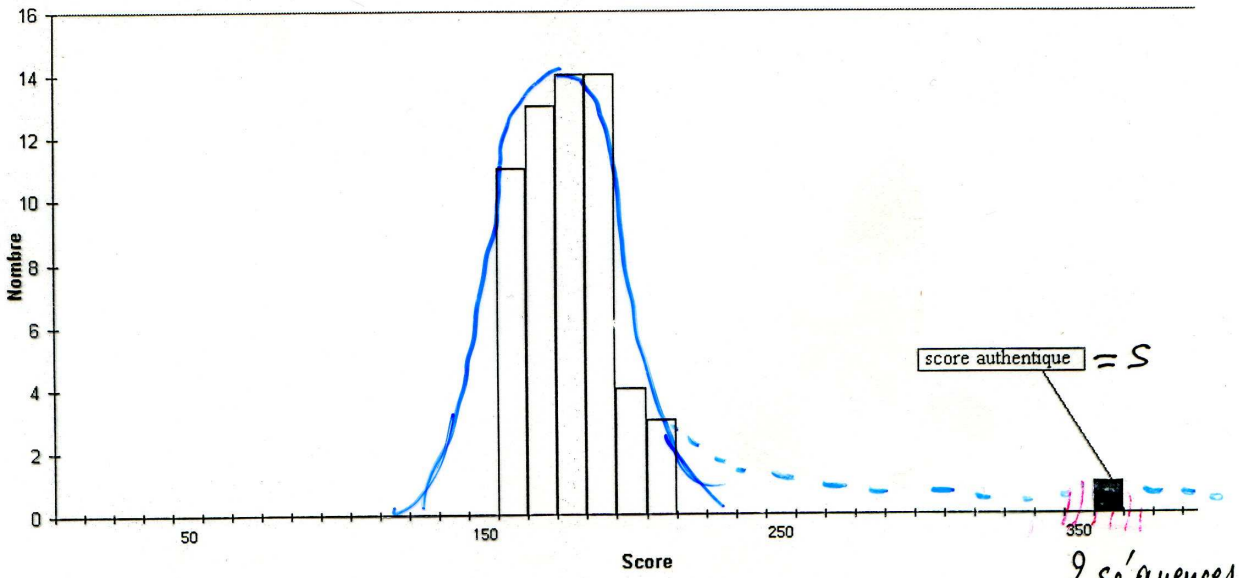
Score Z ≡ nombre de déviations standard au dessus de la valeur moyenne.

Rmq Score Z suppose que la distribution des scores aléatoires suit une loi Normale centrée Réduite

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \mu = \sum_{i=1}^{n} x_i \, p(x_i)$$

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2 p(x_i)}{n-1}$$





n séquences aléatoires

2 séquences native.

$$\text{Score } Z = \frac{S - \mu}{\sigma}$$

$$\mu = \sum_{\text{seq aléa}}^{n} \frac{x_i}{n} \quad , \quad \sigma = \sum_{i}^{n} \frac{(x_i - \mu)^2}{n-1.}$$

exemple précédent $\mu = 188$, $\sigma = 12.8$

$$Z = \frac{367 - 188}{12.8} = 14 \text{ SD}$$

Mais les distributions suivent une loi de distribution de valeurs extrêmes

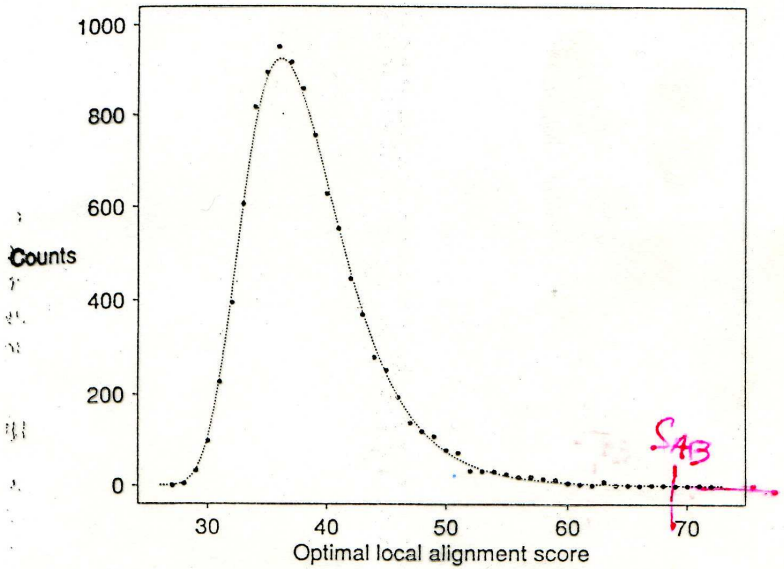


Figure 6. The distribution of optimal local alignment scores from the comparison of a position-specific score matrix with 10 000 random protein sequences. The score matrix was constructed by PSI-BLAST from the 128 local alignments with *E*-value ≤0.01 found in a search of SWISS-PROT using as query the length-567 influenza A virus hemagglutinin precursor (27) (SWISS-PROT accession no. P03435). The random sequences, each of length 567, were generated using the amino acid frequencies of Robinson and Robinson (20). Optimal local alignment scores were calculated using the position-specific matrix in conjunction with $10 + k$ gap costs. The extreme value distribution that best fits the data (3,15) is plotted. A χ^2 goodness-of-fit test with 34 degrees of freedom has value 41.8, corresponding to a *P*-value of 0.20.

→ Z score not correct but used.
(Z score ≥ 12 SD)

2.3 Random Models. Analytical results

- statistics of global seq comparison not well understood
- statistics of local seq comparison well understood.

2.3.1. Local alignments without gaps

→ between 2 seq of length n and m

Two important concepts : E-value and P-value

- **E-value for the score S :**

number of HSP (High-scoring segment pair) with score at least S expected to occur by chance is

$$E\text{-value}(S) = K\, m\, n\, e^{-\lambda S}$$

K, λ = f(scoring matrix, gaps) are determined by some formula.

NB 'E-value' pour le score S est le nombre de scores égal ou supérieur à S par chance.

- **P-value for the Score S = $1 - e^{-(E\text{-value})}$**

is the probability of achieving an equal or higher score S at random

$$E \to 0 \Rightarrow P \to 0$$

2.4. Conclusions

∃ batterie techniques → simples (% ident
→ complexes (HC, Modèles aléatoi

De manière générale, il faut utiliser ces évaluations statistiques avec prudence

→ Score non optimal (incertitude matrice de subst, gap costs

→ pb mathématiques pour les statistiques alig global

→ structure 3D de 2 protéines peut être conservée alors que les séquences 1D n'ont pas de ressemblance significative.