Aging in Sequence Alignment

Enzo Marinari

(Roma *La Sapienza*, Italy)

1. Sequence Alignment (SA).
2. $T = 0$ and transfer matrix algorithms.
3. Finite T.
4. Dynamics and Aging.

This work: E.M.; T. Hwa and E.M..

SA: T. Smith, M. Waterman, S. Karlin, S. Altschul;

SA and St. Mech.: T. Hwa, R. Bundschuh, M. Lässig, M. Muñoz.

Creta, June 2001

Sequence Alignment

Simple model system for pattern matching $\longrightarrow$ one of the most commonly used computational tools in molecular biology.

- Identification of the function of newly sequenced genes;
- Construction of phylogenic trees.

Computational biology:

compare sequences via a transfer matrix algorithm to find an optimal alignment.

"Evaluate similarity between long strings of the alphabet"

(see also: compare copies of a message sequence ruined by imperfect transmission).

Simplest problem: (local) gapless alignment.

(BLAST has a very effective code for that)

We consider an alphabet of size Λ, and 2 sequences

$$\begin{aligned} \vec{a} &= \{a_1, a_2, \cdots, a_M\} \\ \vec{b} &= \{b_1, b_2, \cdots, b_N\} \ , \end{aligned}$$

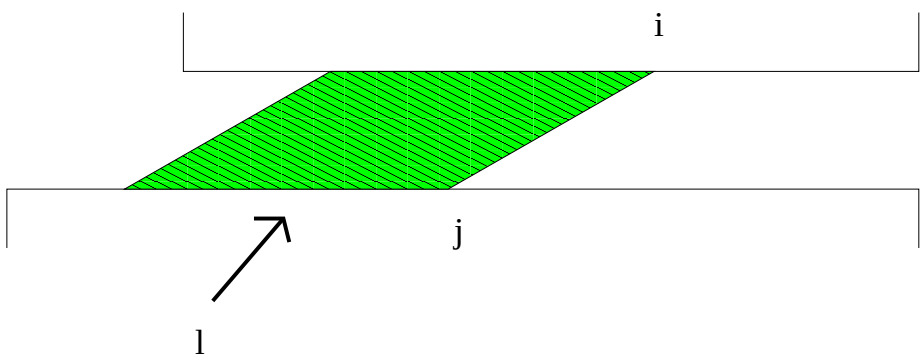
respectively of lenght M and N.

For example for DNA $\Lambda = 4$, alphabet $= \{A, C, G, T\}$. For proteins: twenty letters. Frequency of the letters: natural frequency of aminoacids.

A local gapless alignment of two sequences is done of two substrings of length l

$$\begin{array}{llll} a_{i-l+1}, & \cdots, & a_{i-1}, & a_i \\ b_{j-l+i}, & \cdots, & b_{j-1}, & b_j \end{array}$$

The (gapless) alignment can be characterized by the three variables i, j and l.



In this way each alignment gets a score

$$S(i,j,l) \equiv \sum_{k=0}^{l-1} s_{a_{i-k},b_{j-k}}$$

$s_{a_{i-k},b_{j-k}}$: scoring matrix.

The typical example is the match-mismatch matrix that we have already described, with $s_{a,b}$ equal to 1 for $a = b$ and to $-\mu$ for $a \neq b$ (here the gapless case, no δ).

$$\begin{pmatrix} 1 & -\mu & -\mu & \cdots \\ -\mu & 1 & -\mu & \cdots \\ -\mu & -\mu & 1 & \cdots \\ \cdots & & & \end{pmatrix}$$

This scheme is used for DNA. Most complex schemes (Pam 20 x 20 or BLOSUM are used for proteins, accounting for many issues like for example hydrophobicity).

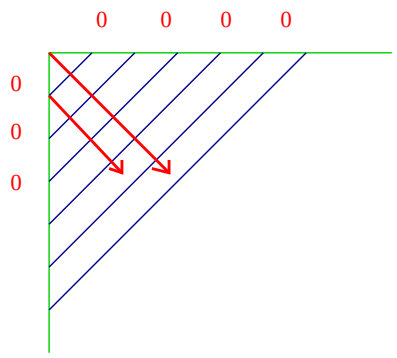
Our goal is: for a given scoring matrix we want to find the highest total score

$$\Sigma \equiv \max_{i,j,l} S(i,j,l) \ .$$

Transfer matrix algorithm: allows to compute Σ in $O(N^2)$ instead than in $O(N^3)$ steps.

$$\sigma_{i,j} = \max\left\{\sigma_{i-1,j-1} + s_{a_i,b_j} \ , \ 0\right\} \ ,$$

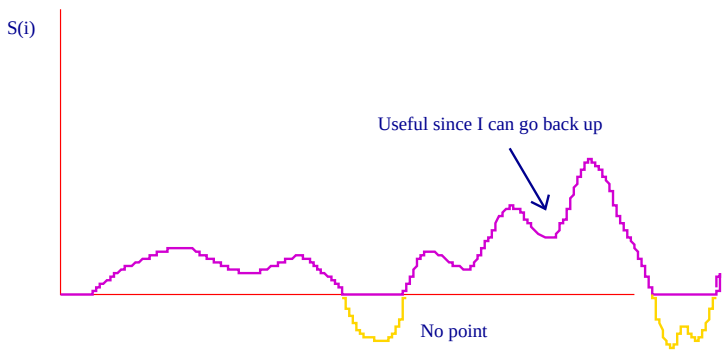
with "initial conditions" $\sigma_{0,k} = \sigma_{k,0} = 0$.



If in a given matrix site I reach a score ≤ 0 I can get a better score starting the matching from this point (i.e. matching a shorter string).

In a given site:

- optimal score zero $\Longrightarrow$ optimal l equal to zero;
- optimal score larger than zero $\Longrightarrow$ optimal l larger than zero.



Traveling on diagonal islands.

Basically: random walk with increments $s_{a,b}$, with cutoff in zero.

Optimal score:

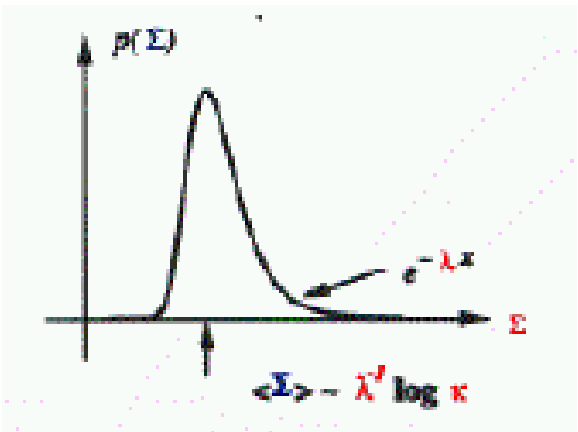
$$\boxed{\Sigma = \max_{i,j} \sigma_{i,j}}$$

To judge about the significance of a match we need to know Σ for two random sequences: we do that with same scores $s_{a,b}$ and using the observed frequencies p_a.

It has been derived rigorously (Karlin-Dembo, Karlin-Altschul) that for suitable scoring parameters

$$\boxed{P\{\Sigma < S\} = e^{-K e^{-\lambda S}}}$$

Gumbel extreme value distribution.



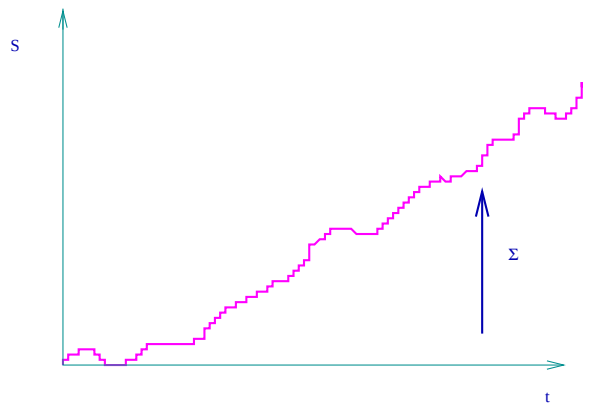
Parameters λ and K. λ: tail. K: $\langle\Sigma\rangle = \frac{1}{\lambda}\log K$.

A simple starting point and approximation: random sequences. Take $i = j$ without loss of generality (all diagonals are born equal...):
$S_{i,j} \to S_{i,i} \to S(t)$; $s(a,b) \to s(t)$; ($s(t) = 1$ with probability p and $-\mu$).

$$\begin{aligned} \sigma(t) &= \max\{S(t) + s(t), 0\} \\ \Sigma &= \max_t \sigma(t) \end{aligned}$$

This is a random walk with lower boundary. There are two phases.

$\langle s \rangle > 0 \Longrightarrow S(t)$ will increase in average (after a while the zero option becomes immaterial).

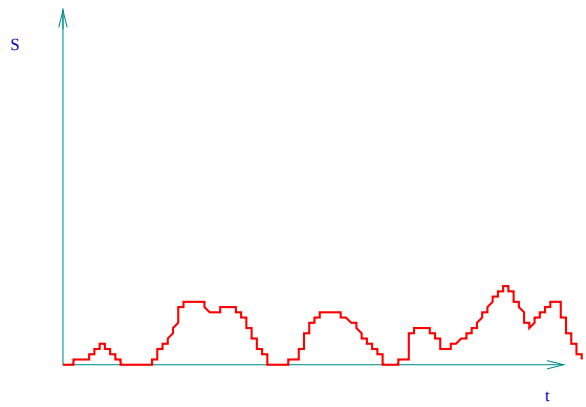


$$\langle \Sigma \rangle \simeq N \langle s \rangle$$

linear phase of local alignment.

- No match of subsequences (always match it all).
- Σ is distributed as a Gaussian random variable (central limit): not extreme valued distribution.

If $\langle s \rangle < 0$ it is all different. Now the cutoff at zero is crucial: it always comes back to play a role, even for very large sequences.



When $S(t) > 0$: random walk with independent increments. Typically it comes back to zero, since $\langle s \rangle < 0$ means a negative drift.

Large number of islands, statistically independent. Sea: part with $S(t) = 0$.

Distribution of island peak scores σ_k for continuous time and Gaussian $s(t)$ is asymptotically Poisson:
$P(\sigma_k > \sigma) \simeq Ae^{-\lambda\sigma}$

λ: typical scale of the maximal island score.

The global optimal score Σ. Take $K = \frac{N}{\langle l \rangle}$ islands.
$\Sigma = \max_k \{\sigma_k\}$ (that will turn out to be extreme valued).

$P(\Sigma < S) =\simeq e^{-\mathcal{K}e^{-\lambda S}}$

Gumbel distribution: theory of extremal statistics.
Bouchaud and Mézard work about connection of RSB in Derrida REM model and Gumbel.

Now we know a lot about the best alignment.

But what about good alignments?

Excited states $\longrightarrow$ finite T problem.

Basically: count score of all islands, and weight

$$\sum_k e^{-\beta E_k}$$

For example (Y-K Yu) T=0 Needleman-Wunsch transfer matrix algorithm:

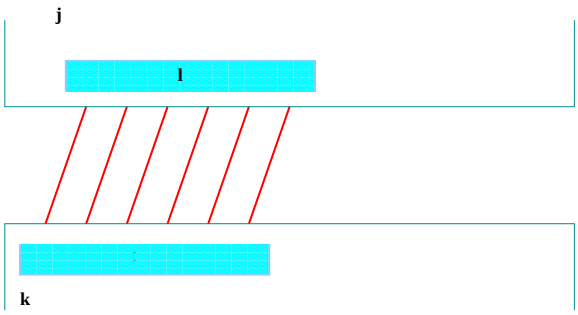
$$h(r,t+1) = \max \begin{cases} h(r,t-1) & + & s(r,t) \\ h(r+1,t) & - & \delta \\ h(r-1,t) & - & \delta \end{cases}$$

becomes at $T \neq 0$

$$W(r,t+1) = e^{-\beta\delta}\left(W(r+1,t) + W(r-1,t)\right) + e^{-\beta s(r,t)} W(r,t)$$

finite T generalization of NW-TM.

We introduce a local dynamics. Situation is very simple for the local gapless case. We can describe the configuration with three variables: j, k, l.



Now we propose the basic moves:

$$j \to \begin{cases} j+1 \\ j-1 \end{cases} ; \; k \to \begin{cases} k+1 \\ k-1 \end{cases} ; \; l \to \begin{cases} l+1 \\ l-1 \end{cases}$$

it the matching does not pass the boundary and if the length does not become smaller than zero. Energy is defined as $E = -\sum_{a=j,j+l} \sum_{b=k,k+l} s_{a,b}$

Boltzmann: $P(C) \simeq e^{-\beta E(C)}$, $\beta = \frac{1}{T}$. Use simple Metropolis algorithm. Thermal histories and annealing.

Annealing: start from high T; reduce T; compute observables for different T values: for example average score and best score found (typical of annealing optimization).

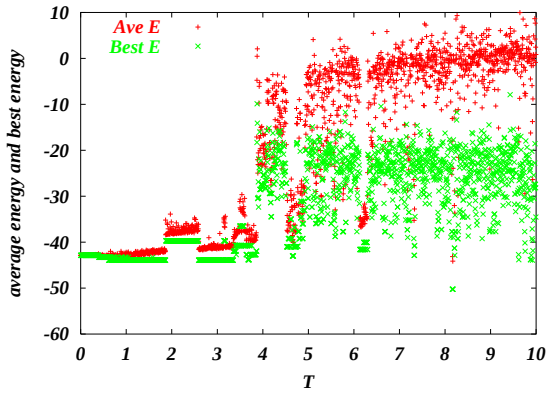
High complexity. Traps. Hints for slow dynamics.

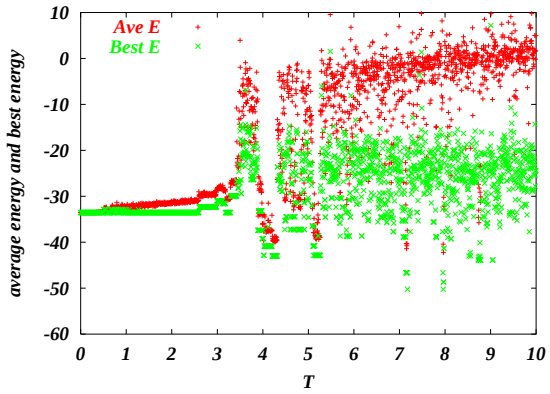
For the gaped case introduce "gap" variables, $\Gamma_i = 0$ if site i is gaped, 0 if it is connected. Same kind of results.

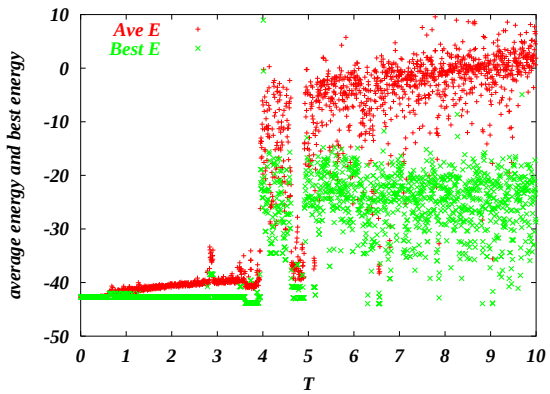
Gapless local alignment.

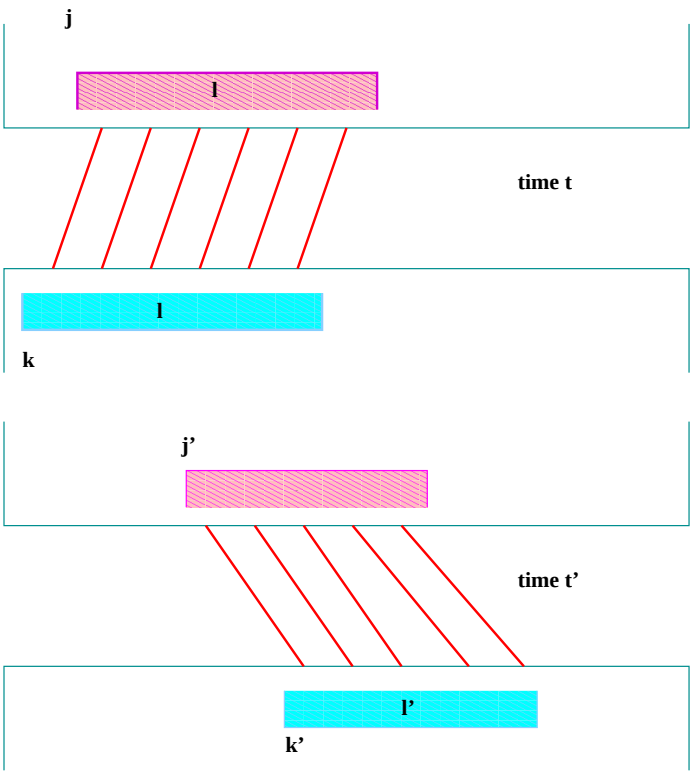
Here random quenched score matrix. 4 letter alphabet. -51 is the true ground state energy (computed via the transfer matrix method).

Note traps. In the last run the GS is not found.









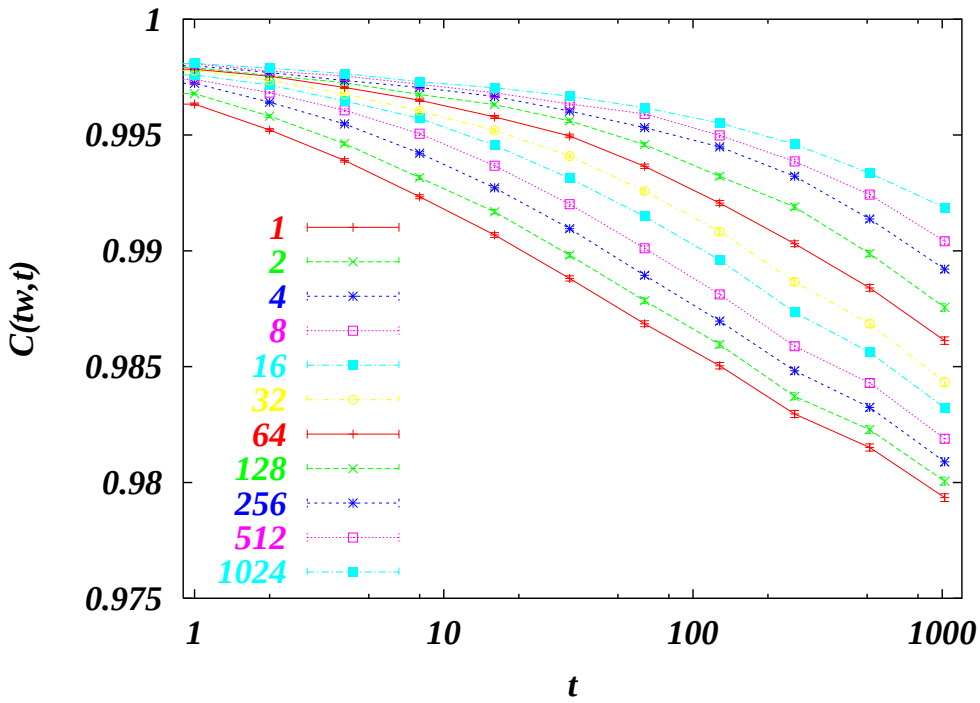
Compare the matched part of the sequence $a(t)$ at time t with the matched part of $a(t')$ at time t' and $b(t)$ at time t with the matched part of $b(t')$ at time t' (two separate correlation functions).

$\eta_i^{(a)}(t) = 0, 1$, $i = 1, \cdots N$, 0 if not matched, 1 if matched.

$\sigma_i \equiv 1 - 2\ \eta_i = \pm 1$, and

$$\sum_i \sigma_i(t)\sigma_i(t')$$
$$= \sum_i \Big(1 - 2\ \eta_i(t) - 2\ \eta_i(t') + 4\ \eta_i(t)\eta_i(t')\Big)$$
$$= N - 2\ l(t) - 2\ l(t') + 4\ \sum_i \eta_i(t)\eta_i(t')$$

Very clear aging. No time translation invariance.



Two regimes. First decay for local wandering (stay inside a valley). Second decay region determined by length change.

Application to DNA is very relevant.

Complementary sequences.

Scores from experimental values.

Unzipping experiments: increase T and get opening bubbles.

You expect aging in experimental conditions!