Aging in Sequence Alignment

Enzo Marinari

(Roma *La Sapienza*, Italy)

1. An introduction to sequence alignment (SA).
2. Gapless SA.
3. Gaped SA.
4. Aging.
5. Dynamics.
6. (DNA).

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.

Parma, May 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).

Global alignment.

a global alignment of two sequences $\{P_i\}$ and $\{Q_j\}$ is defined as:

an ordered set of pairings (P_i, Q_j) and of unpaired elements $(P_i, -)$ and $(-, Q_j)$ called gaps.

Each letter $\{P_i\}$, $\{Q_j\}$ belongs exactly to one pairing or to one gap.

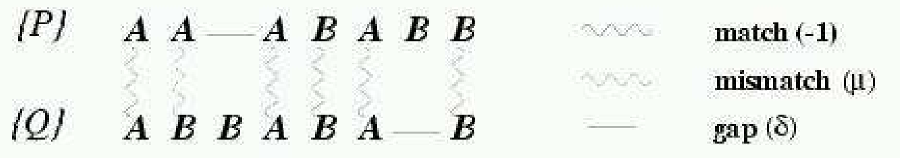


FIG. 1. One possible alignment of two binary sequences, $\{P_i\} = AAABABB$ and $\{Q_j\} = ABBABAB$, with six pairings (five matches, one mismatch) and two gaps.

(figure from T. Hwa and M. Lässig, Phys. Rev. Lett. **76** (1996) 2591, cond-mat/9511072)

Alphabet length: DNA, $\Lambda = 4$; proteins, $\Lambda = 20$.

The optimal alignment of the two sequences is determined by minimization of an "energy" function E.

E favors matches (M) $(P_i = Q_j)$ over mismatches MM $(P_i \neq Q_j)$ and gaps G $(P_i, -; -, Q_j)$

Simple common energy function:

$$\sum_{M,MM,G} \begin{cases} -1 & \text{for matches} \\ \mu > 0 & \text{for mismatches} \\ \delta > 0 & \text{for gaps} \end{cases}$$

Choice of energy (cost) function: crucial issue in biology. We will not discuss that much here. Proteins: BLOSUM, PAM.

Relation among

$$\begin{array}{c} \text{Transfer} \\ \text{matrix} \\ \text{algorithm} \\ \text{for} \\ \text{SA} \end{array} \rightarrow \left\{ \begin{array}{c} \text{Partition function of} \\ \text{a directed polymer} \\ \text{in a random medium} \\ \\ \text{KPZ surface growth} \end{array} \right.$$

$\longrightarrow$ Scaling Laws:

Scaling laws can help in tuning parameters to optimal values

$$\left. \begin{array}{c} \text{Universal} \\ \text{and} \\ \text{Non-Universal} \end{array} \right\} \text{features}$$

Assessment of alignment significance.

Obviously an optimal alignment does not reflect necessarily a sequence similarity.

I align two random sequences.

I (obviously) get a best score.
This does not mean that these sequences have something in common.

When does a best alignment really reflect a meaningful similarity?

Start from the random case:

probability of getting a given by chance.

Distribution of alignment score for random sequences:

$\longrightarrow$ Gumbel

$\longrightarrow$ universal

$\longrightarrow$ 2 parameters

(role of extremal statistics).

Simplest problem: (local) gapless alignment.

(BLAST has a very effective code for that)

We consider an alphabet of size Λ, and 2 sequences

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

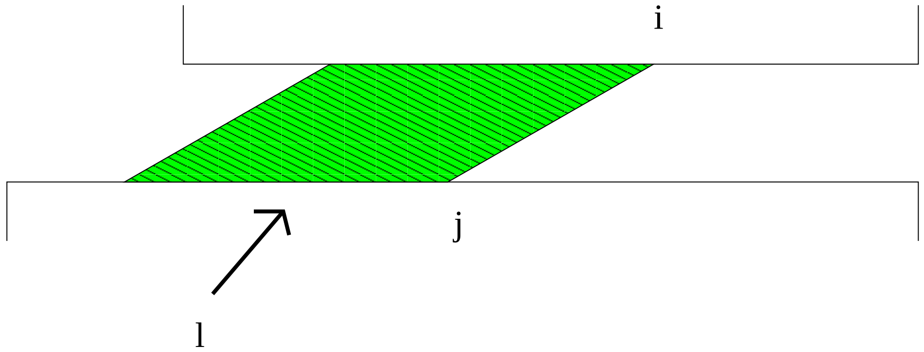
respectively of lenght M and N.

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

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

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).

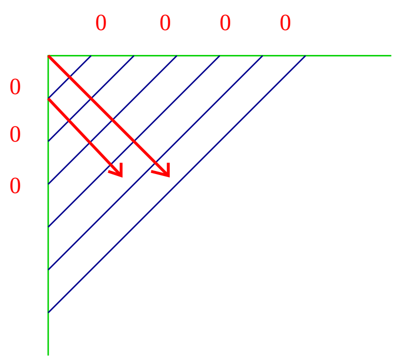
Our goal is: for a given scoring matrix we want to find the highest total score

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

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

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

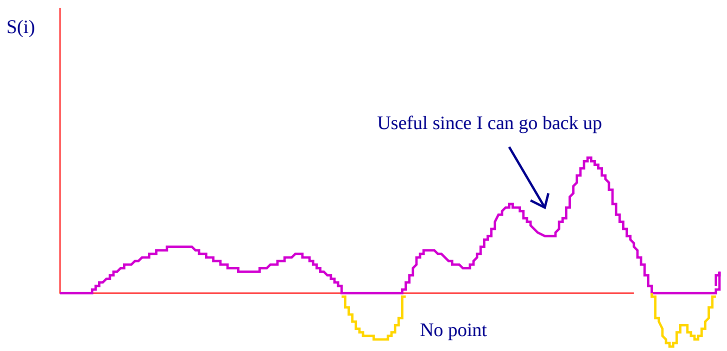
with "initial conditions" $\sigma_{0,k} = \sigma_{k,0} = 0$.



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:

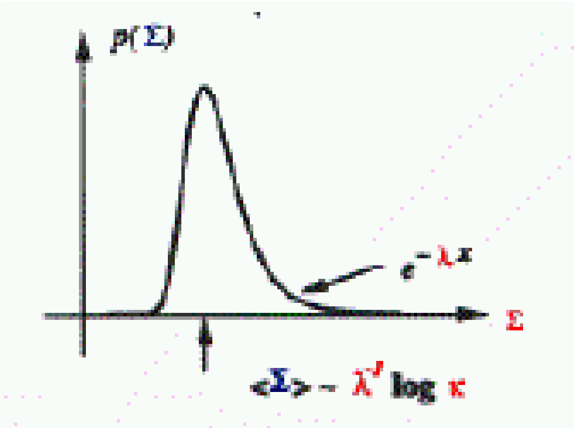
$$\boxed{\Sigma = \max_{i,j} \sigma_{i,j}}$$

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: $\langle \Sigma \rangle = \frac{1}{\lambda} \log K$.

We will go back to that.

Gapless alignment: one can compute λ and K.

λ: unique solution of

$$\langle e^{\lambda s} \rangle = 1 \ ,$$

i.e.

$$\sum_{a,b} p_a p_b \ e^{\lambda s_{a,b}} = 1 \ ,$$

and $K = \tilde{K}(s_{a,b}, p_a) \cdot M \cdot N$.

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 $-\mu)$.

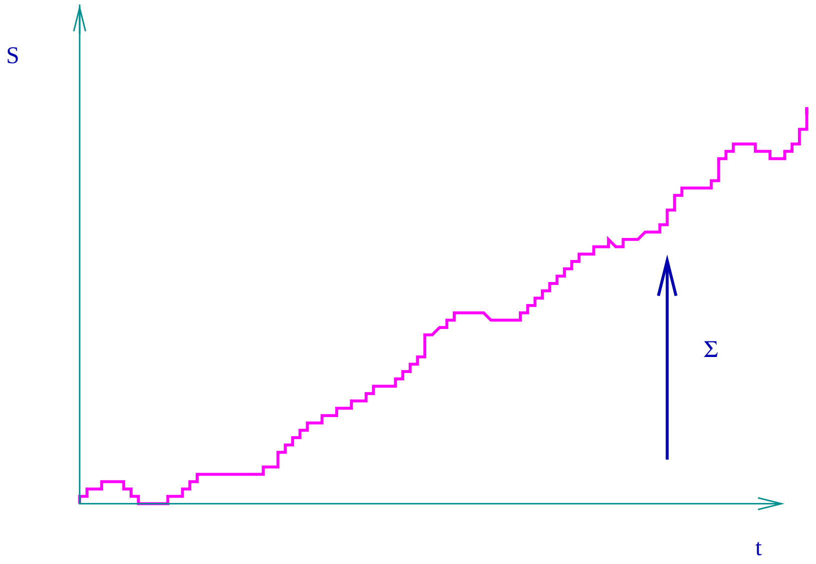
$$\begin{aligned} \sigma(t) &= \max\{S(t) + s(t), 0\} \\ \Sigma &= \max_t \sigma(t) \end{aligned}$$

This is a random walk with lower boundary.

There are two phases. The quantity that selects them is the local similarity score

$$\sum_{a,b} p_a p_b \; s_{a,b} \; .$$

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



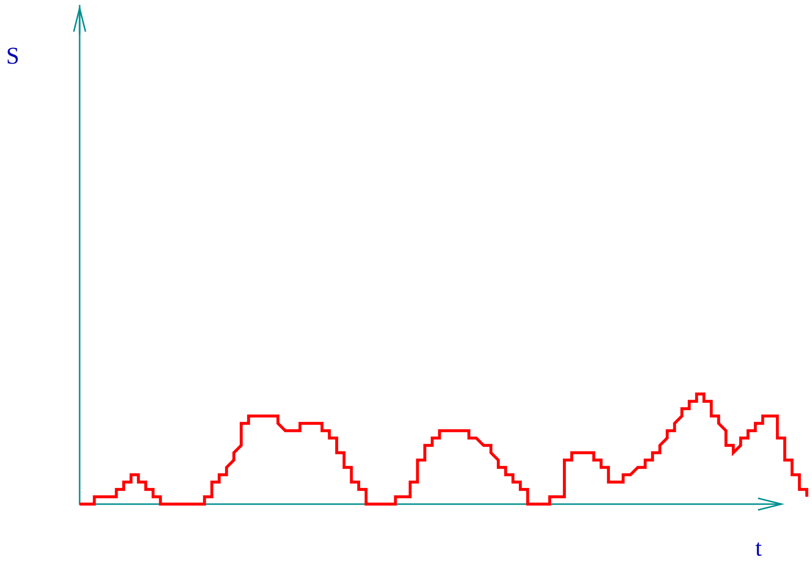
$$\langle \Sigma \rangle \simeq N \langle s \rangle$$

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:

$$P(\sigma_k > \sigma) \simeq A e^{-\lambda \sigma}$$

λ: typical scale of the maximal island score.

How do we show that it is Poisson?

$$P(\sigma) = \langle \delta \left(\sigma - \sum_{t=1}^{L} s(t) \right) \rangle$$

$$= \langle \frac{1}{2\pi} \int dk \ e^{-ik\sigma + ik \sum_{t=1}^{L} s(t)} \rangle$$

$$= \frac{1}{2\pi} \int dk \ e^{-ik\sigma} \langle e^{iks} \rangle^{L}$$

we assume $\sigma = \alpha L$, i.e. that the optimal score of an island is proportional to its length.

$$P(\sigma) = \frac{1}{2\pi} \int dk \ e^{-ik\alpha L} \langle e^{iks} \rangle^{L}$$

$$= \frac{1}{2\pi} \int dk \ e^{-L\left(ik\alpha - \log \langle e^{iks} \rangle\right)}$$

Saddle point for $k = k^*$

$$1 = \frac{i\langle s e^{ik^* s}\rangle}{\langle e^{ik^* s}\rangle \alpha}$$

On the saddle point

$$P(\sigma) \simeq e^{-\lambda \sigma} \;,$$

with

$$\lambda = ik^* - \frac{1}{\alpha}\log\left[\langle e^{ik^* s}\rangle\right]$$

Minimize over α: $\langle e^{ik^* s}\rangle = 1 \Longrightarrow \lambda = ik^*$.

All together I get

$$\boxed{\langle e^{\lambda s}\rangle = 1}$$

$$\boxed{\langle s e^{\lambda s}\rangle = \alpha}$$

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).

$$P(\Sigma < S) = P(\max\{\sigma_1, \cdots, \sigma_k\} < S)$$

$$= P(\sigma < S)^k = \left(1 - Ae^{-\lambda S}\right)^k$$

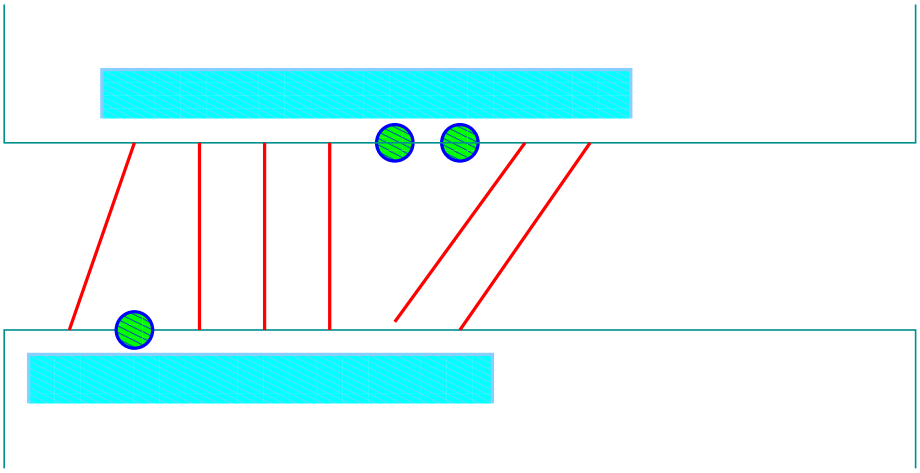
$$\simeq \left(e^{-Ae^{-\lambda S}}\right)^k \simeq e^{-\mathcal{K}e^{-\lambda S}}$$

with $\mathcal{K} \equiv Ak$.

Gumbel distribution: theory of extremal statistics.

Bouchaud and Mézard work about connection of RSB in Derrida REM model and Gumbel

Alignment with gaps



Biological (and non) sequences can have repetitions, duplications, deletions.

Allow gaps. Pay a penalty for that.

TAGGC, TAGC

$$\longrightarrow \begin{array}{l} \text{TAGGC} \\ \text{TAG}-\text{C} \end{array}$$

Now the two sequences to be aligned can have different lenghts

$$S = \sum_{a,b} s_{a,b} - N_{\text{gap}}\ \delta$$

δ: gap penalty.

One can use more complicated gap costs. For example the cost of deleting three G can be smaller than three times the cost of deleting a single G.

Now we look for a best alignment allowing gaps. Again a very effective transfer matrix method exists.

Alignment path representation.

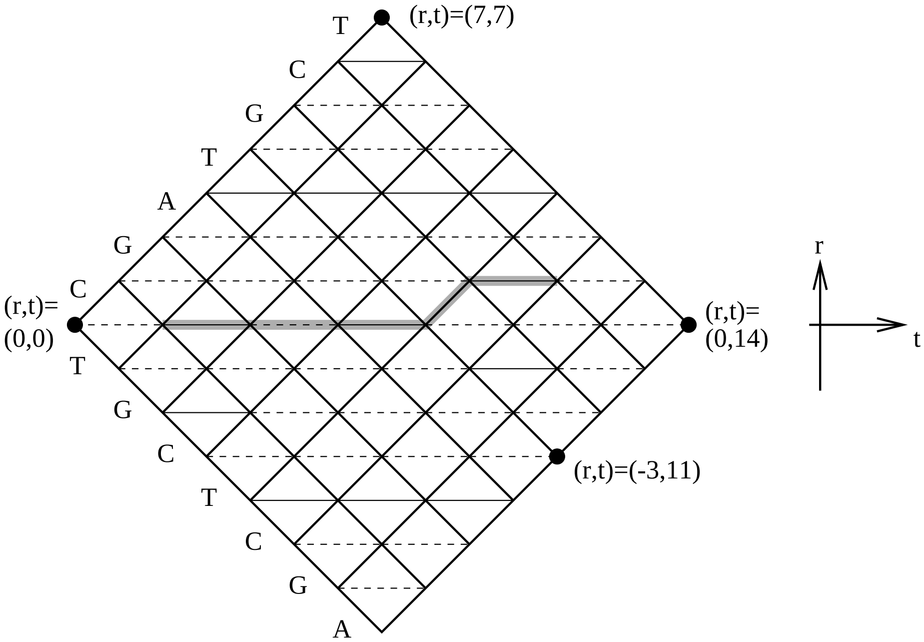


figure from Bundschuh, cond-mat/9911386

Horizontal bonds: gain $s_{a,b}$.

Diagonal bonds: pay $-\delta$.

Each directed path on the lattice stands for a possible alignment.

$$(r,t) = (i-j, i+j-1)$$

Best score for global alignment is for a path (the best) going from $(0,0)$ to $(0,2N)$.

Call $h(r,t)$ the best score of the path $(0,0) \to (r,t)$.

Needleman-Wunsch transfer matrix algorithm:

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

Analogies:

Directed polymer in random potential $\{s_{a,b}\}$ at $T=0$. h: polymer energy.

h: spatial height profile of a growing surface (KPZ).

Now local alignment. Use the same trick than in the gapless case: we cutoff unfavorable scores. Find out islands by

$$S(r,t+1) = \max \begin{cases} S(r,t-1) & + & s(r,t) \\ S(r+1,t) & - & \delta \\ S(r-1,t) & - & \delta \\ \quad 0 & & \end{cases}$$

and now select the island with maximal score:

$$\Sigma = \max_{r,t} S(r,t) \ .$$

Also with gaps there is a phase transition.

- linear phase, $\langle\Sigma\rangle \simeq N$
- logarithmic phase, $\langle\Sigma\rangle \simeq \log(N)$

Again: in the linear phase because of the score growth the zero option does not play any role.

Here even with $\langle s\rangle < 0$ you can be in the linear phase (gaps can be used to connect regions with positive score).

Now need $u(\{s\},\delta) + \langle s\rangle < 0$ (that defines a critical line).

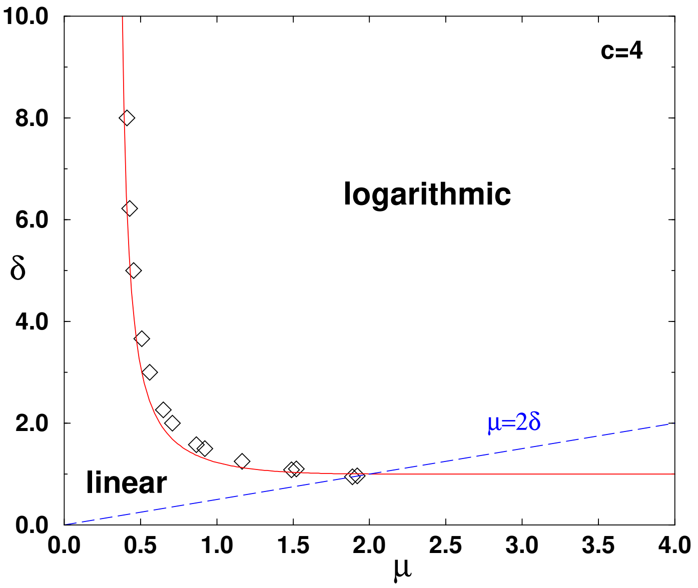


figure from Bundschuh, `cond-mat/9911386`

Empirical finding: distribution is 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.

The $T \to 0$ limit gives back the original result:

$$h = \lim_{\beta \to \infty} \frac{1}{\beta} \log W$$

There is a straightforward generalization also for the gaped case.

Finite T is interesting. There is maybe more to learn, in DNA sequencing and elsewhere, than from ground state.

Maybe even that the ground state is not very relevant in some cases.

Multiple alignment: complex case, important motivation for $T > 0$.

We will introduce a local dynamics, compute numerically correlation functions, note relevant analogies, ...

But, before that, ..., a few words about aging.

Complex, glassy systems.

Low T: very long relaxation times (even in experiments that last three years you can check that you do not reach equilibrium).

Does it have any meaning to describe equilibrium?

17

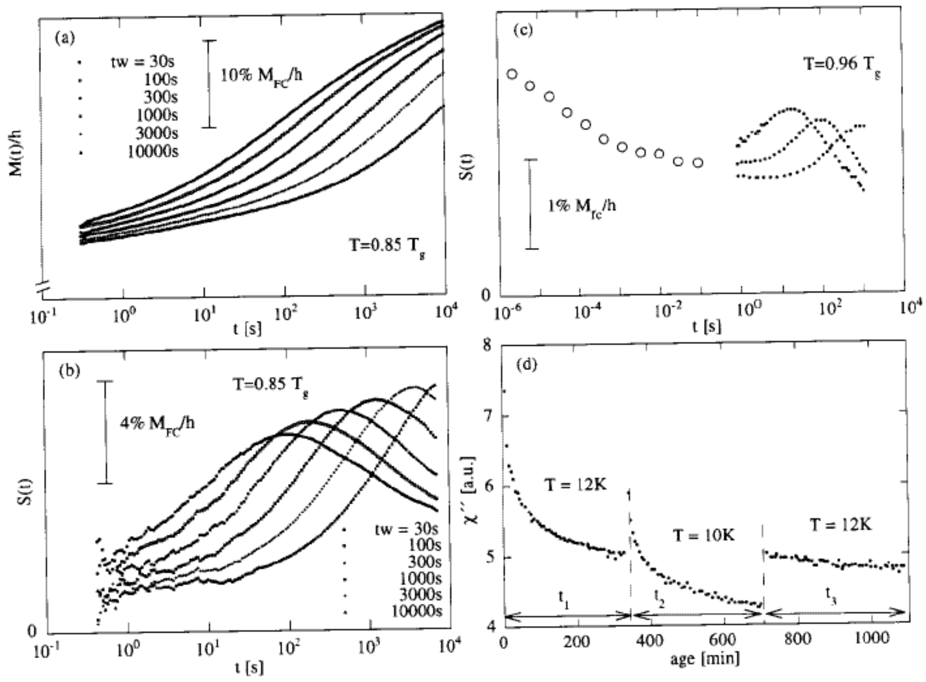


Figure 9: The influence of ageing on the susceptibility of spin glasses. a) $M_{ZFC}(t_w, t)/h$ vs. $logt$ for a Cu(Mn) spin glass at $T/T_g = 0.85$, b) the relaxation rate of the curves in a), c) relaxation rate in a wider time window for an amorphous metallic spin glass, $T/T_g = 0.96$ and d) relaxation of $\chi''(\omega, t_a)$ of an insulating spin glass, $CdCr_{1.7}In_{0.3}S_4$, at different temperature. The sample is first cooled to 12 K, after t_1 the temperature is decreased to 10 K and the relaxation is measured during t_2 and finally the sample is heated back to 12 K and χ'' is recorded a time t_3. Frequency $\omega/2\pi = 0.01$ Hz. The figure is reproduced using data from Ref. [37].

figure from Nordblad-Svendlish, in P. Young book

So: compute $C(t, t_w)$.

Decay rate depends on t_w. One finds a dynamical universality: different scaling time regions.

Equilibrium: $C(t, t_w) = C(t - t_w)$

Aging: $C(t, t_w)$ true function of two parameters.

Typical spin glass experiment:

- $t = 0$: fast quench in $h = 0$, from $T \gg T_g$ to $T_1 < T_g$
- wait t_w, then measure χ, ac susceptibility, with small oscillating field. One finds aging:

$$\chi = \chi(\omega, t_w)$$

i.e. the response to a perturbation depends on thermal history.

Phenomenological:

$$\chi(\omega, t_w) = \chi_{ST}(\omega) + \frac{A(T)}{(\omega t_w)^b}$$

Also: TRM (thermo-remanent magnetization) experiments

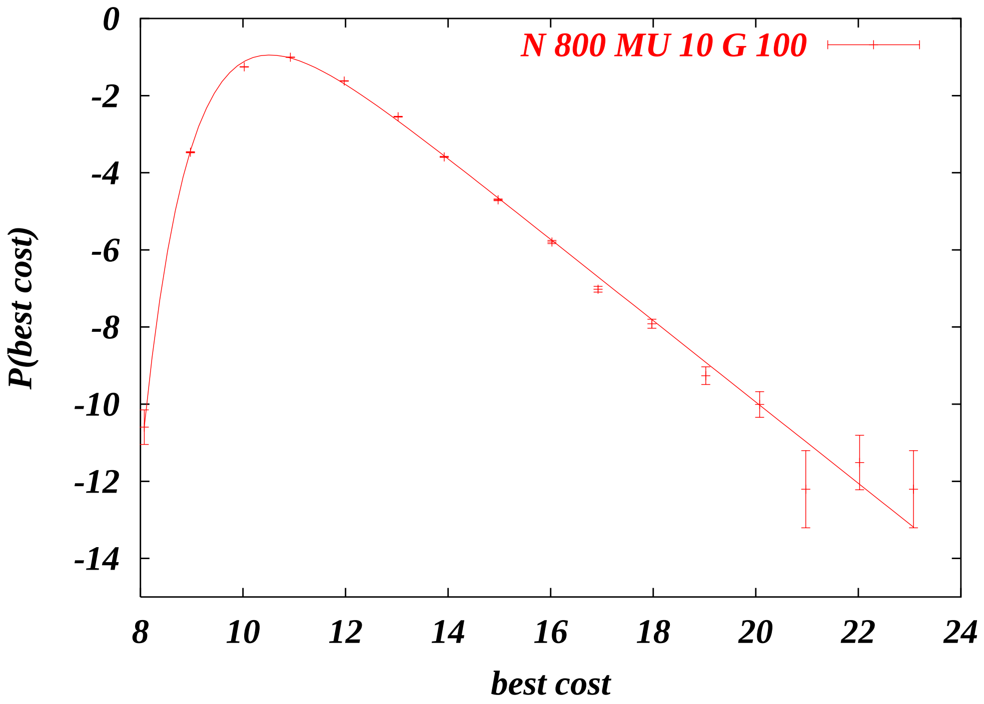
$$M(t_w + \tau) = M_{ST}(\tau) + M_{AG}(t_w + \tau)$$

where here t_w is the time where I switch off a small field. $M_{AG}(t_w + \tau) \simeq f(\frac{\tau}{t_w})$, and t_w is connected with the level of equilibration reached by the system.

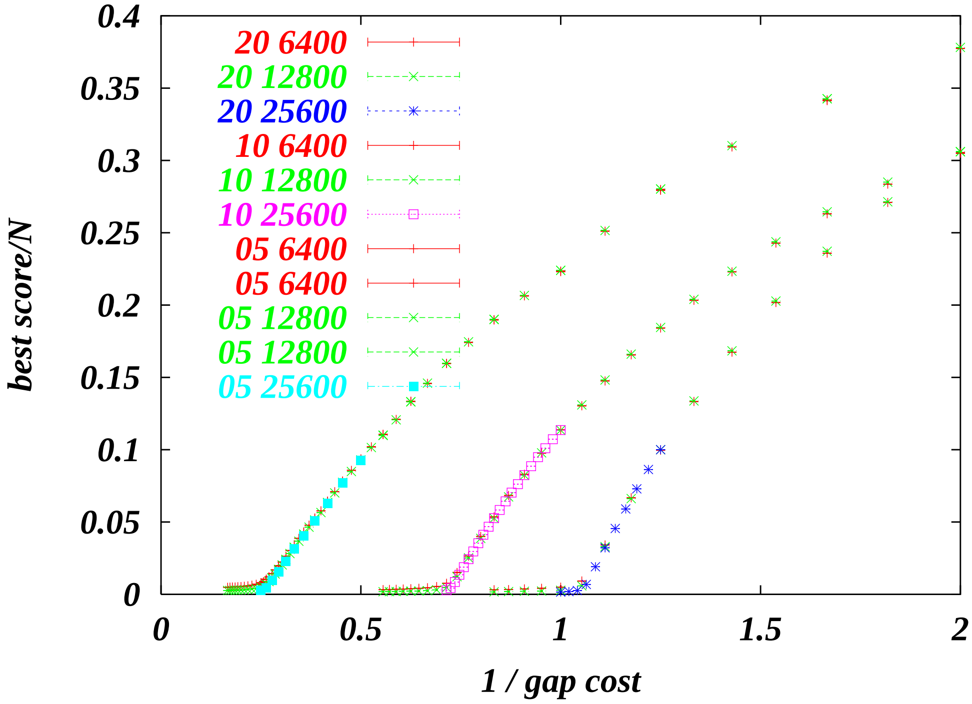
Again, detecting aging gives hints about some intrinsic glassiness of the system.

Implement transfer matrix algorithm. Solve “many”, “large” samples.

Data really give a Gumbel distribution.

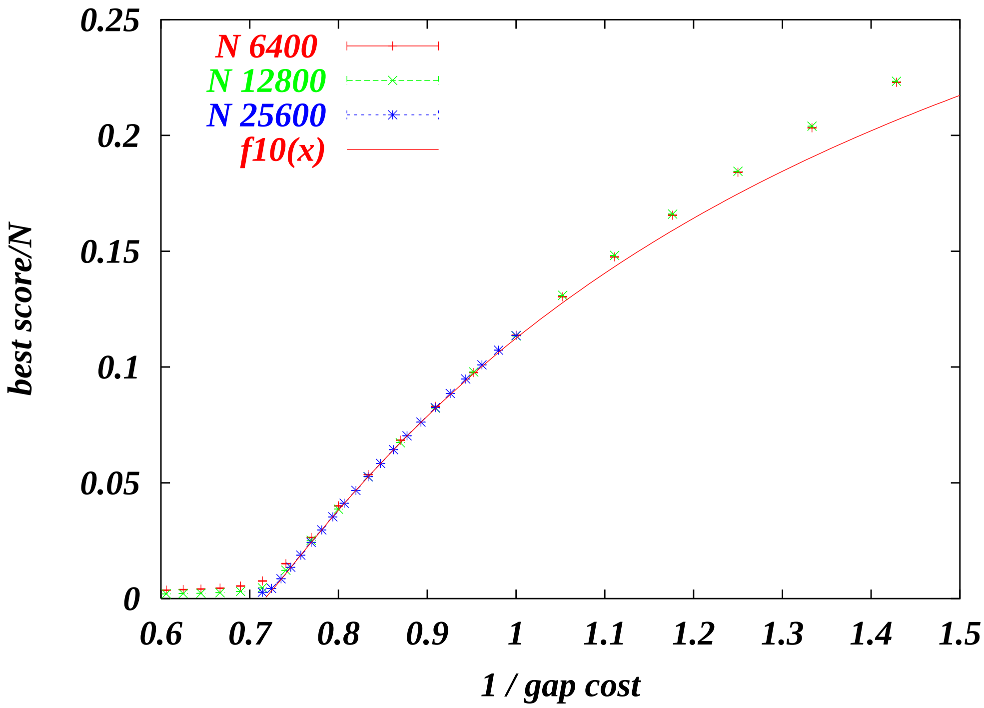


There is a phase transition in the gap cost from the linear phase to the log phase (see Hwa, Bundschuh, Lässig, Muñoz).



Matrix 1, μ. Different μ values and sequence size.

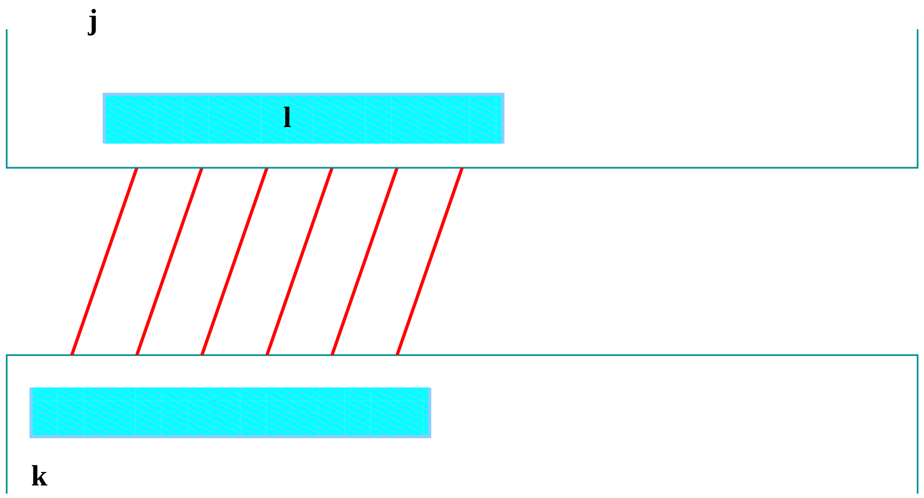
Best fit to best score as a function of the gap cost.



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.

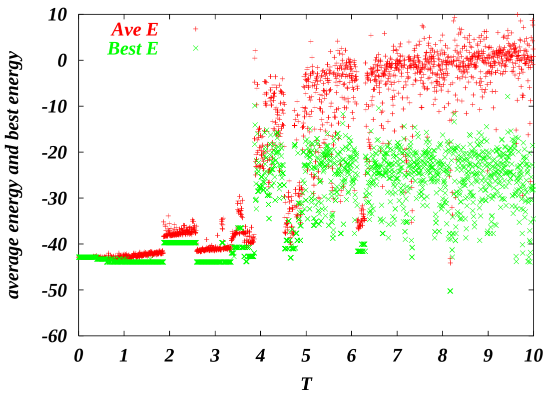
Hints for slow dynamics (see later).

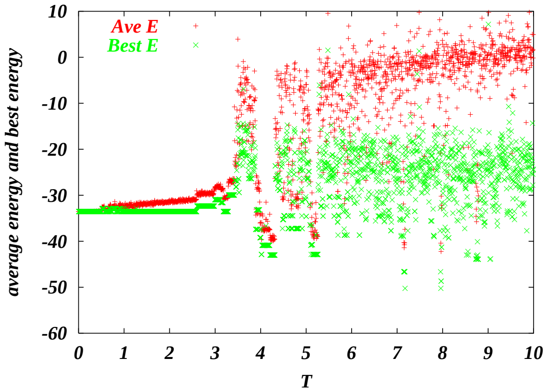
For the gaped case introduce "gap" variables, $\Gamma_i = 0$ if site i is gaped, 0 if it is connected.
Same kind of results.

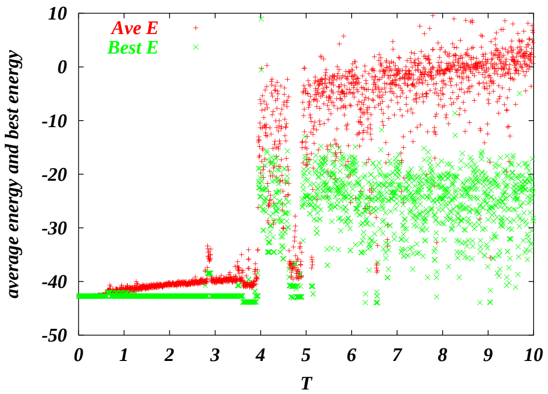
Gapless local alignment.

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

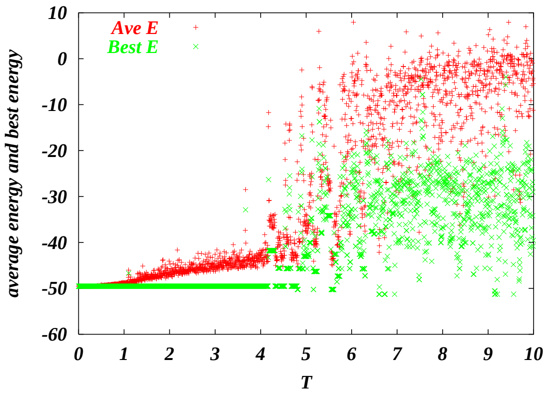
Note traps. In the last run the GS is not found.

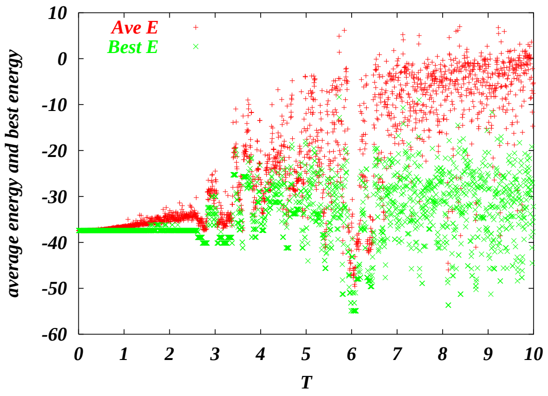


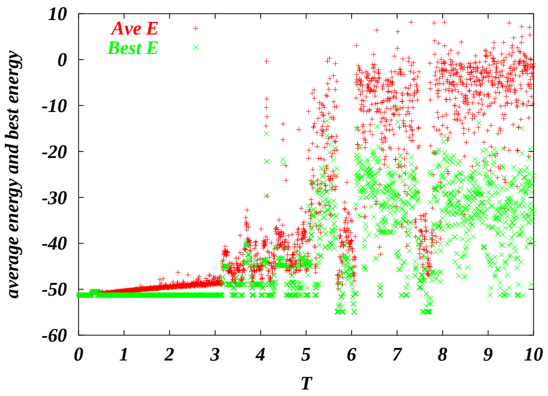


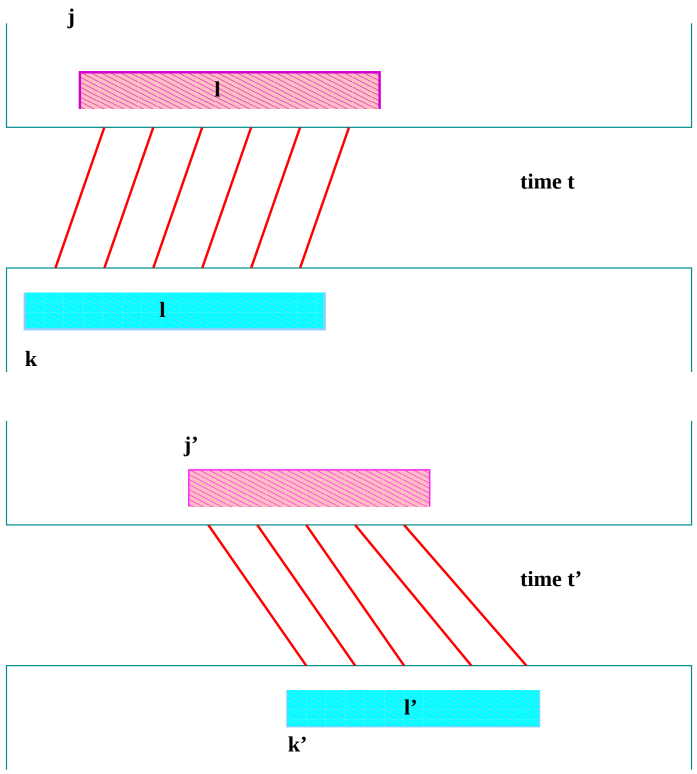


Here three further runs for a different score matrix.









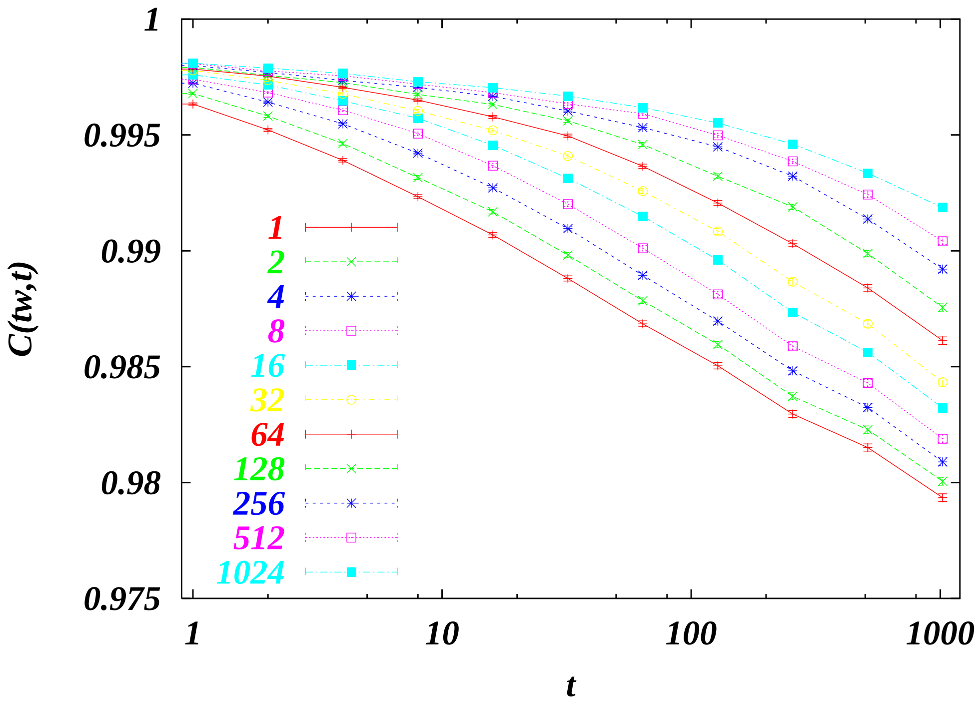
Compare the matched part of the sequence $a(t)$ at time t with the matched part of $a(t')$ at time t' and $b(t)$ at time t with the matched part of $b(t')$ at time t' (two separate correlation functions).

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

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

$$\begin{aligned} &\sum_i \sigma_i(t)\sigma_i(t') \\ &= \sum_i \left(1 - 2\ \eta_i(t) - 2\ \eta_i(t') + 4\ \eta_i(t)\eta_i(t')\right) \\ &= N - 2\ l(t) - 2\ l(t') + 4\ \sum_i \eta_i(t)\eta_i(t') \end{aligned}$$

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.

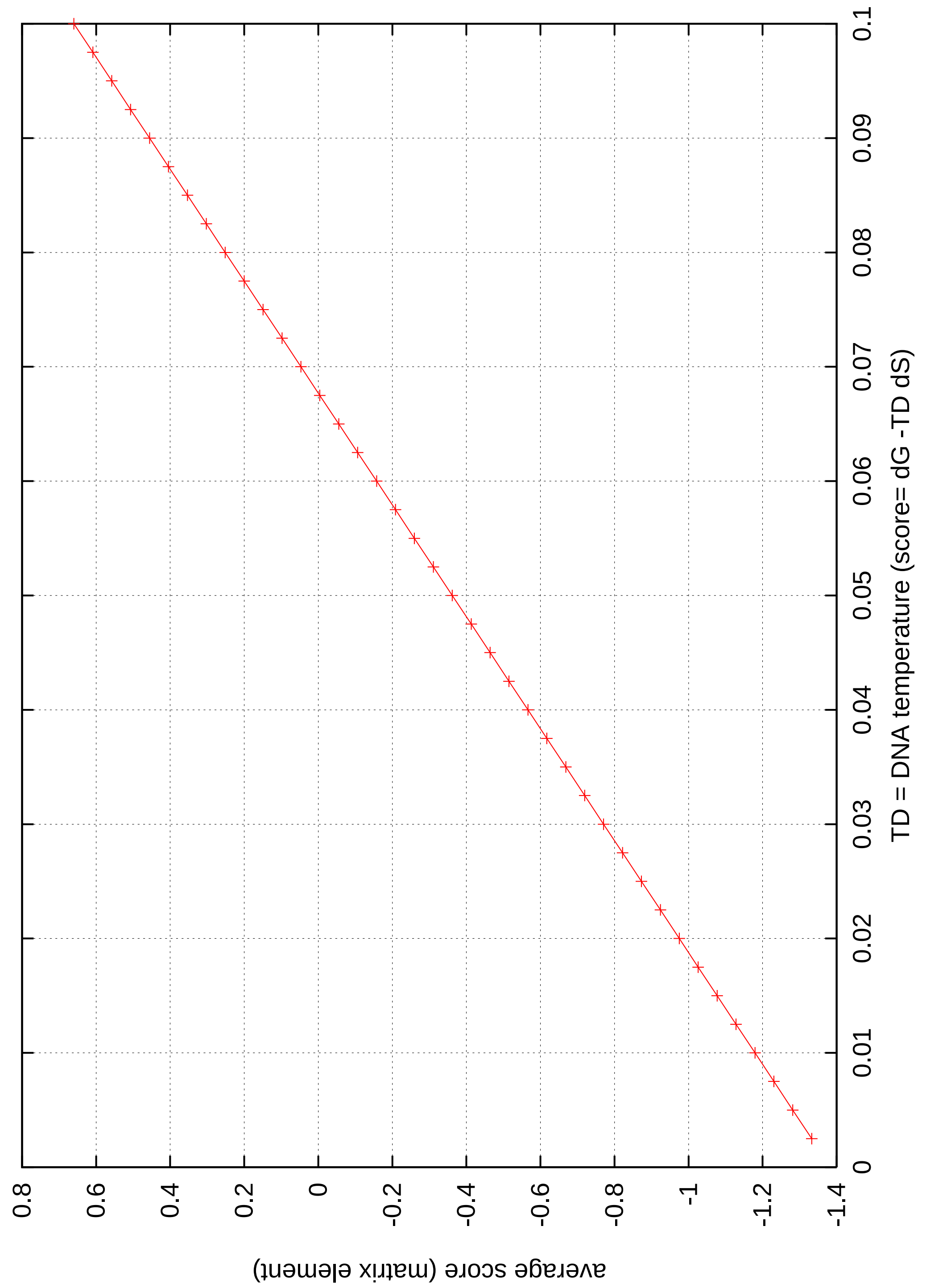
Scores from experimental values.

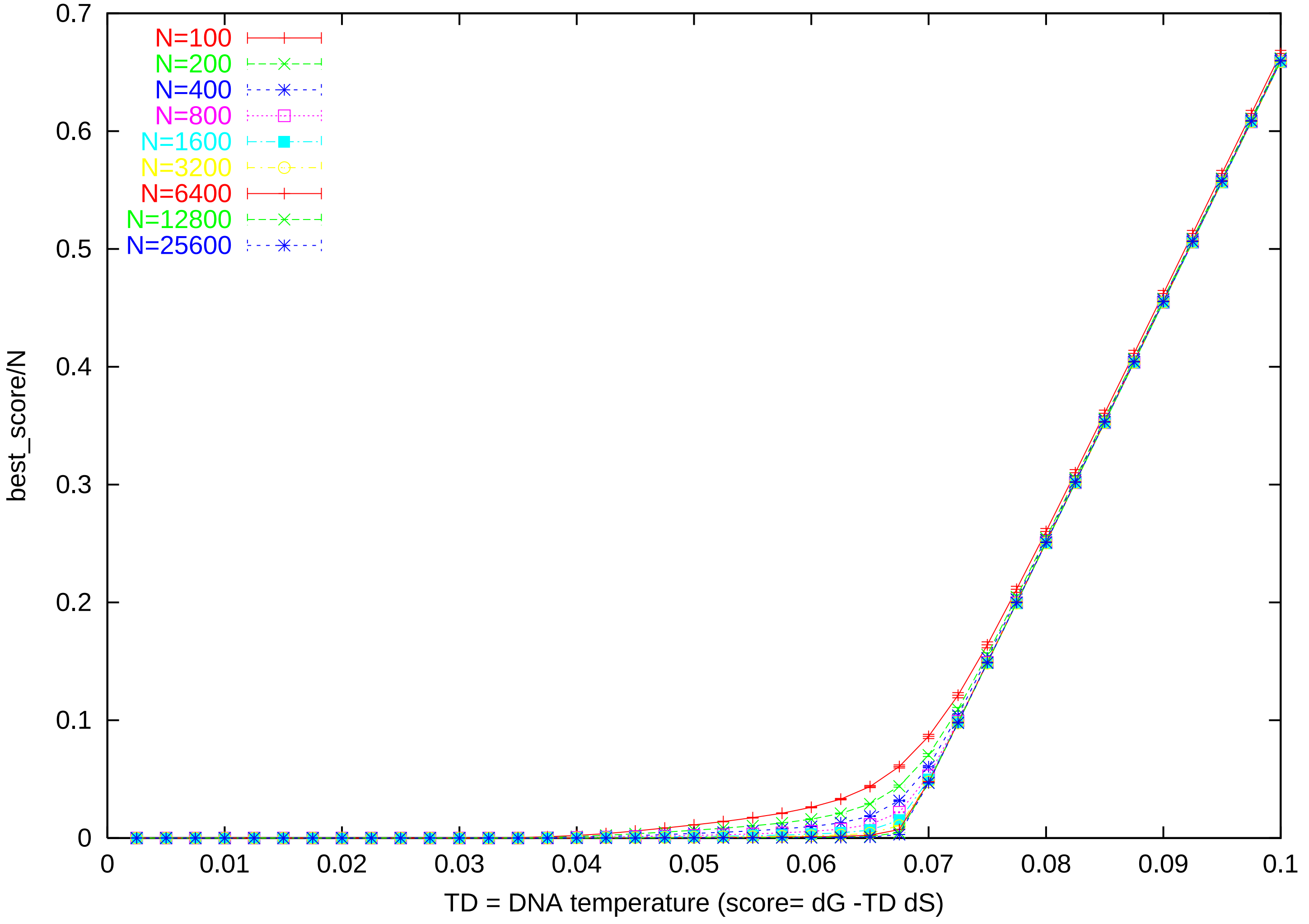
Unzipping experiments: increase T and get opening bubbles.

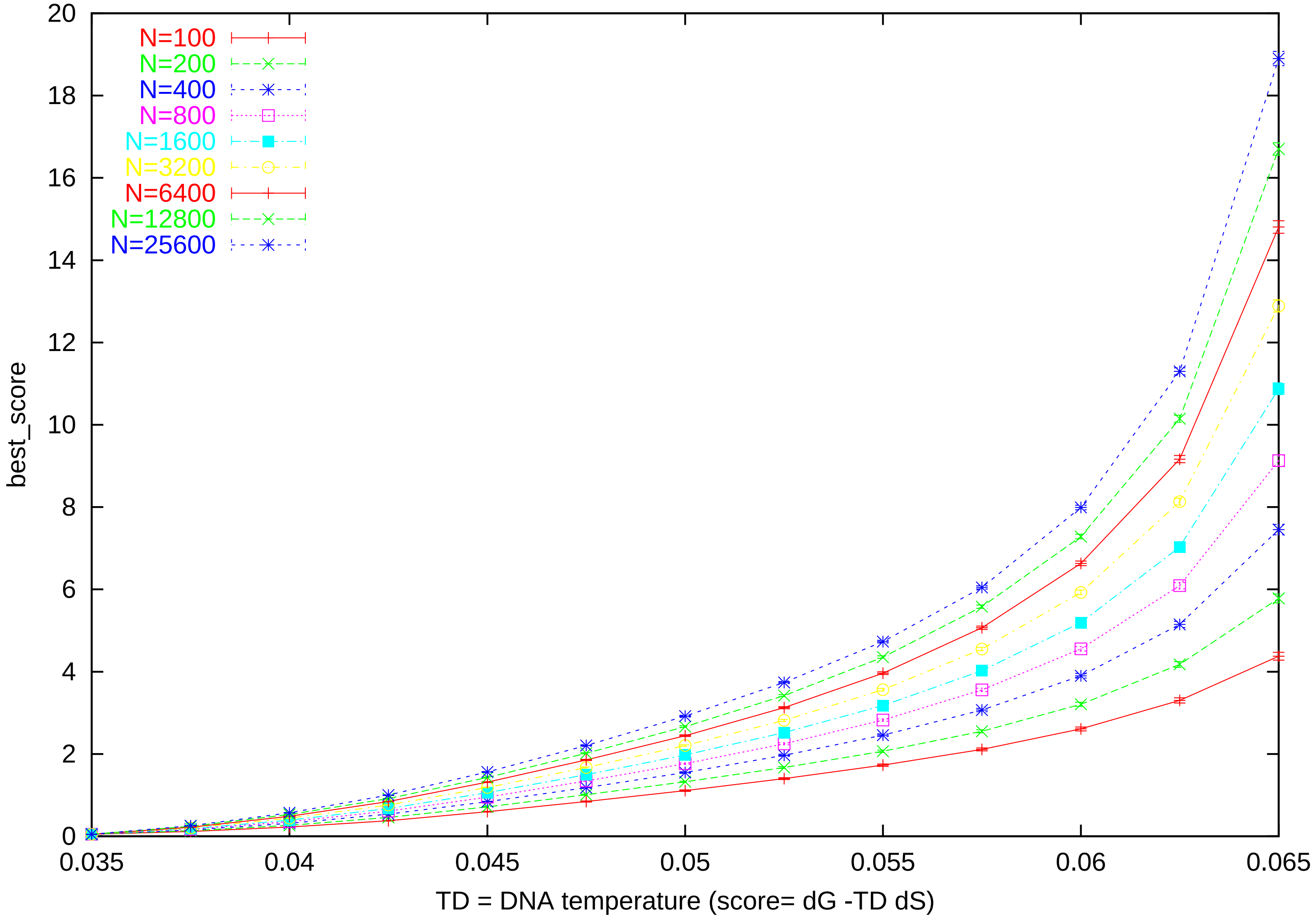
You expect aging!

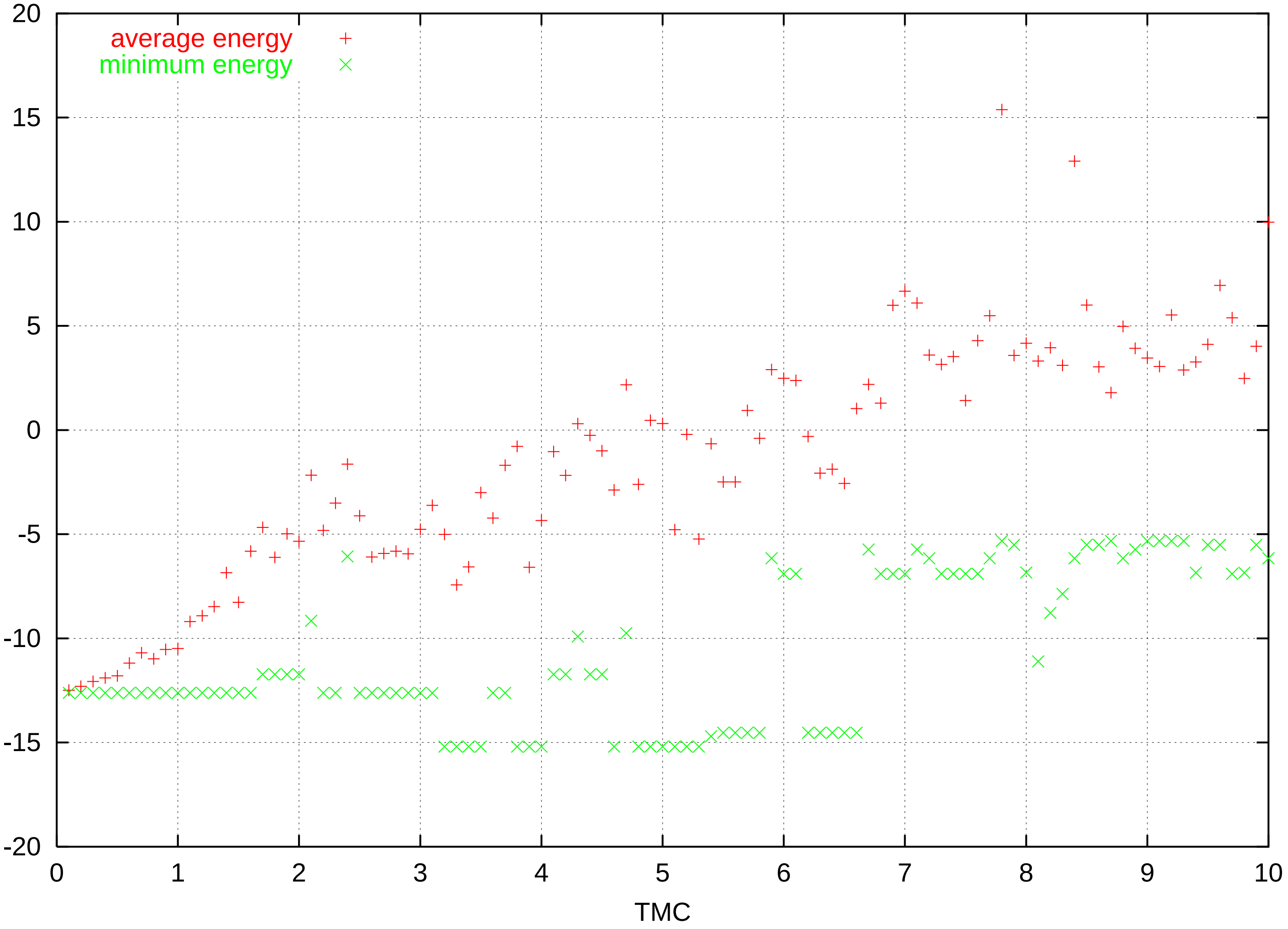
Here I just give a few details.

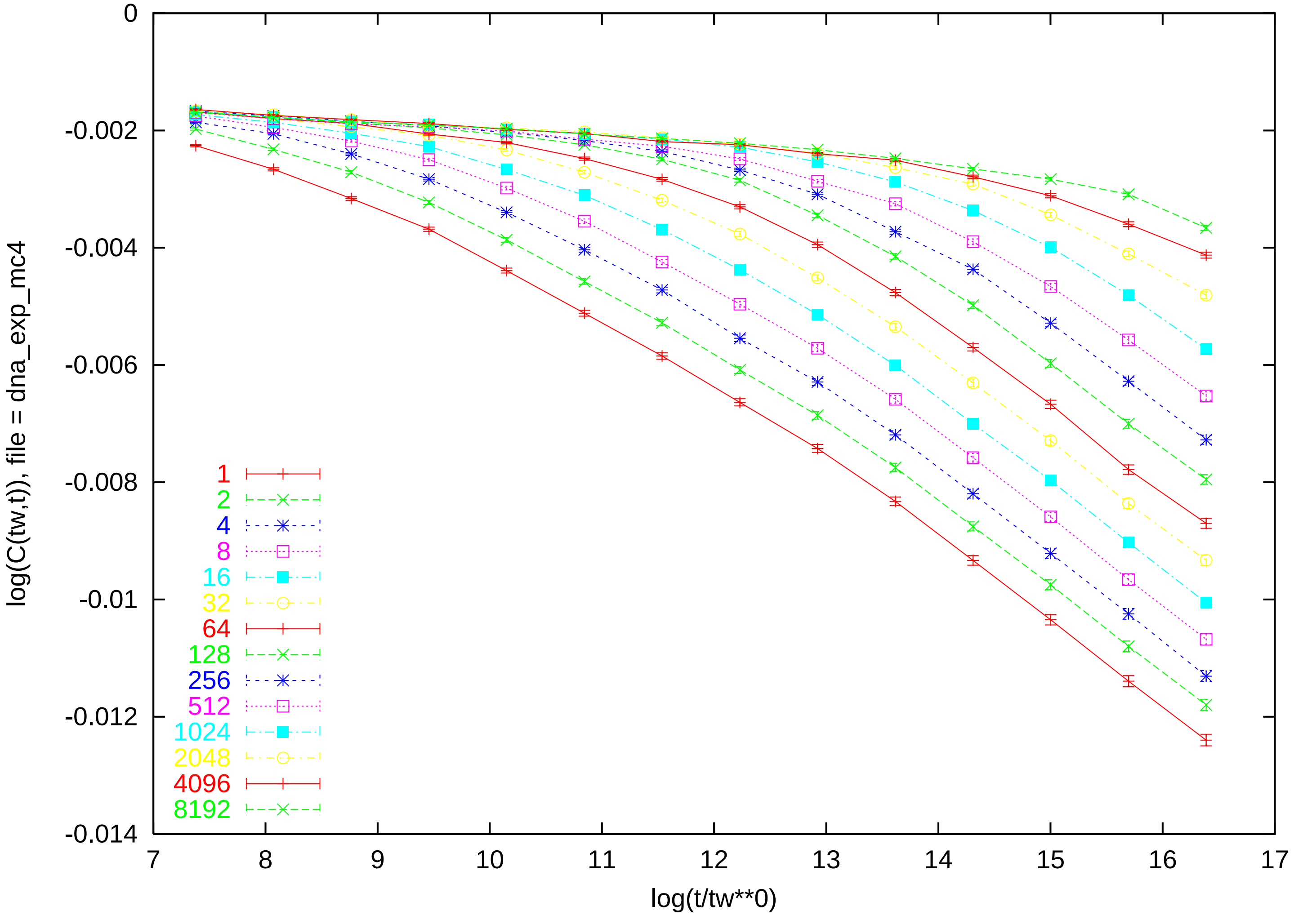
DNA


average score (matrix element)
TD = DNA temperature (score= dG -TD dS)
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1

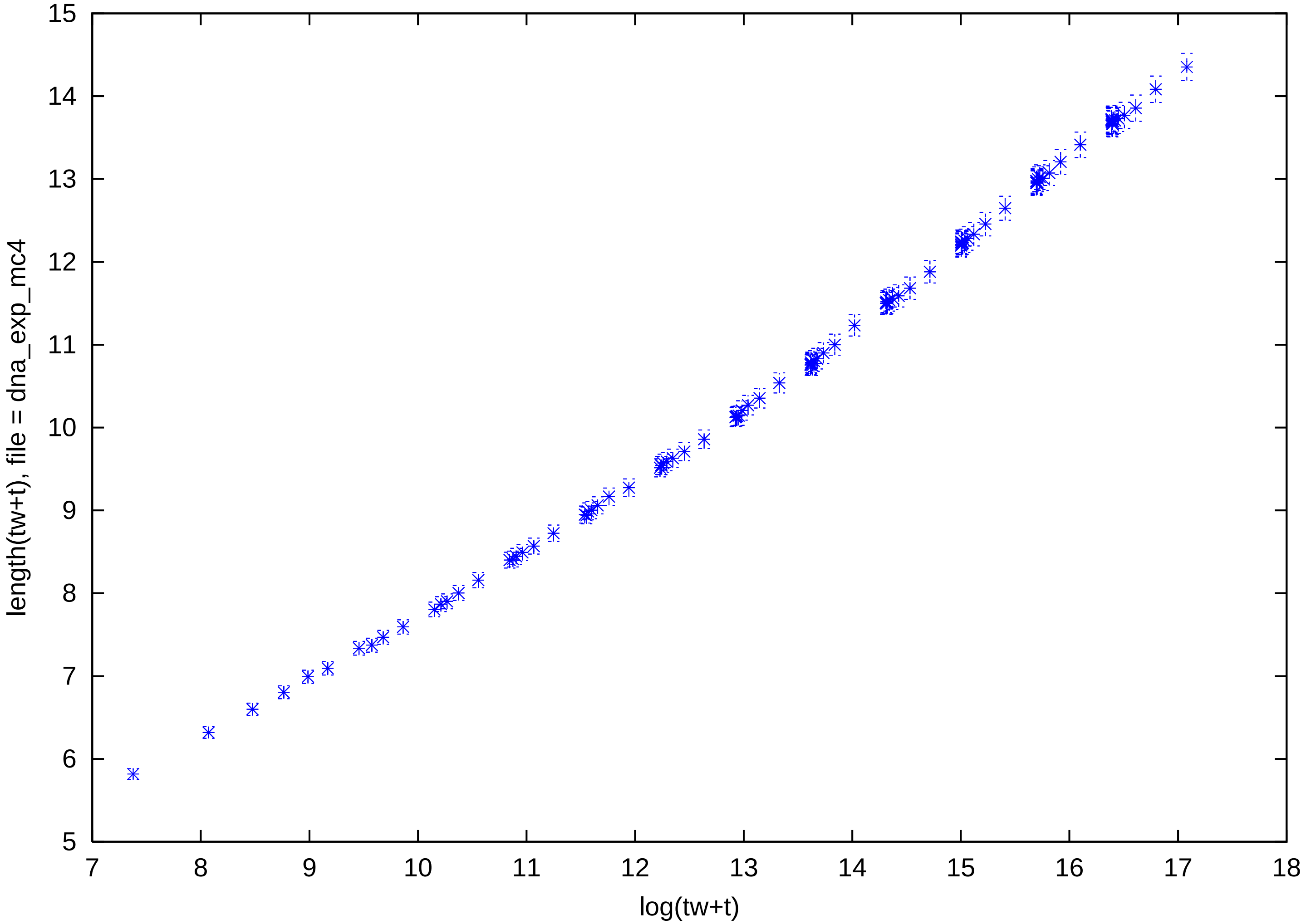

N=100
N=200
N=400
N=800
N=1600
N=3200
N=6400
N=12800
N=25600
best_score/N
TD = DNA temperature (score= dG -TD dS)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09


N=100
N=200
N=400
N=800
N=1600
N=3200
N=6400
N=12800
N=25600
best_score
20
18
16
14
12
10
8
6
4
2
0
0.035
0.04
0.045
0.05
0.055
0.06
0.065
TD = DNA temperature (score= dG -TD dS)


average energy
minimum energy
20
15
10
5
0
-5
-10
-15
-20
0
1
2
3
4
5
6
7
8
9
10
TMC



log(C(tw,t)), file = dna_exp_mc4
log(t/tw**0)
0
-0.002
-0.004
-0.006
-0.008
-0.01
-0.012
-0.014
7
8
9
10
11
12
13
14
15
16
17
1
2
4
8
16
32
64
128
256
512
1024
2048
4096
8192


length(tw+t), file = dna_exp_mc4
log(tw+t)
15
14
13
12
11
10
9
8
7
6
5
7
8
9
10
11
12
13
14
15
16
17
18