Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 321 (2003) 577–586
PHYSICA A
www.elsevier.com/locate/physa

DNA dimer correlations reflect *in vivo* conditions and discriminate among nearest-neighbor base pair free energy parameter measures

Pedro Miramontes[a,*], Germinal Cocho[b]

[a] *Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de México (UNAM), Cd Universitaria 04510 DF, Mexico*
[b] *Departamento de Sistemas Complejos, Instituto de Física, Universidad Nacional Autónoma de México (UNAM), Cd Universitaria 04510 DF, Mexico*

Received 11 July 2002; received in revised form 29 October 2002

Abstract

The stability of the DNA duplex depends on its sequence rather than on its composition (the relative proportions of the four nucleotides or bases A, C, G and T). This fact was recognized since the eighties by several authors. They calculated the duplex relative stability and temperature-dependent behavior (ΔG, ΔH) for the ten different pairwise interactions. Even accepting that the experimental conditions and theoretical assumptions in these reports differ, one should expect little disagreement if their results were to reflect the same natural phenomenon. Unfortunately, this is not the case: The discrepancies among different teams are beyond acceptable experimental errors. Thermodynamics allows to relate DNA dimer frequencies with *in vivo* DNA interactions. In this paper we follow this approach and settle out the controversy by finding the parameter set consistent with intra-cellular DNA local free energy distribution.

PACS: 87.14.Gg

Keywords: DNA; Stability; Free energy

1. Introduction

DNA is far from being just an information carrier; it is the product of several dynamics involving molecular machines associated to the processes of replication, transcription

* Corresponding author. Fax: +52-55-5622-4859.
E-mail addresses: pmv@fciencias.unam.mx (P. Miramontes), cocho@fisica.unam.mx (G. Cocho).

0378-4371/03/$ - see front matter
doi:10.1016/S0378-4371(02)01808-3

Table 1
Absolute values of the free energy (ΔG_{ij} (kcal/mol)) for all the pairwise DNA interactions. The table was taken from [3]

Interaction	Gotoh	Vologodskii	Breslauer	Delcourt	Sugimoto	Unified
AA/TT	0.43	0.89	1.66	0.67	1.20	1.00
AT/TA	0.27	0.81	1.19	0.62	0.90	0.88
TA/AT	0.22	0.76	0.76	0.70	0.90	0.58
CA/GT	0.97	1.37	1.80	1.19	1.70	1.45
GT/CA	0.98	1.35	1.13	1.28	1.50	1.44
CT/GA	0.83	1.16	1.35	1.17	1.50	1.28
GA/CT	0.93	1.25	1.41	1.12	1.50	1.30
CG/GC	1.70	1.99	3.28	1.87	2.80	2.17
GC/CG	1.64	1.96	2.82	1.85	2.30	2.24
GG/CC	1.22	1.64	2.75	1.55	2.10	1.84

and ribosomal translation into proteins (also nucleosome and chromosomal organization in eukaryotic organisms). The efficiency and accuracy of these molecular machines depend on the DNA sequence and therefore there is a bias towards compatible DNA motifs. On the other hand, the compositional frequency of words (monomers, dimers, trimers, etc.) conflicts with the constraints associated to DNA dynamics. If the strength of those compositional constraints is weak, then averages over long sequences (even complete genomes) are in order to notice their effect. For example, for the shortest words (monomers) in long sequences, the second Chargaff law states that $f_A \approx f_T$ and $f_C \approx f_G$ (where $f_i\,(i \in \{A, C, G, T\})$ are the nucleotide relative proportions over the whole DNA sequence along one DNA strand) and only the percentage of strong bases (C and G) remains free. The frequency f_S of strong bases ($f_S = f_C + f_G$) can be quite different for distinct species and sometimes it is strikingly counterintuitive. Just to mention an example, in hyperthermophile archeabacteria DNA coding for rRNA or tRNA has f_S values ranging from 0.6 to 0.7, which is consistent with the idea of stronger DNA binding in the high temperature environment where these organisms live. However, for protein-coding DNA $f_S \approx 0.3 - 0.4$. In spite of the high temperatures and against the intuition, this DNA is weakly bound. The fact is that there are also nonlocal phenomena (out of the scope of this paper) that are seldom taken into account; for instance, topoisomerases [1] are enzymes that wind or unwind the double helix and their action might induce additional sequence constraints. However, for the vast majority of local DNA process, its stability depends on the way dimers are distributed along the sequence [2].

SantaLucia [3] has compared nearest-neighbor (along one DNA strand) base pair free energy (ΔG_{ij}) parameters from seven laboratories (Table 1) and found that six of them were more or less in agreement among themselves, while the remaining parameter set (Breslauer et al. [4]) differed, concluding that this set was not correct. The laboratories used natural and synthetic DNA polymers and oligomers, they used different salt concentrations and data analysis tools. SantaLucia makes a critical review of the methods employed by each lab comparing the different lab protocols. As it is

not possible to discern from first principles which dataset is the best one, SantaLucia reasoned, quite naturally, that the set of six reports that were in close agreement should be correct.

The reports are in the columns of Table 1. We are including only six out of seven authors because the team led by Benight [5] revised their calculations [6] after the publication of SantaLucia's paper.

2. DNA dimer distribution

To quantify the effect of DNA dimer organization we use the correlation function.

$$C_{ij} = f_{ij} - f_i f_j \ . \tag{1}$$

The sign of C_{ij} indicates whether the dimer ij is under or overrepresented when compared to a random distribution.

Statistical mechanics associate lower frequencies to higher energies (weaker DNA duplex binding). For instance low values of f_{TA} have been reported to be a common trait in practically all organisms.

Specifically, one would expect:

$$C_{ij} \propto \exp\left(-\frac{\Delta G_{ij}}{kT}\right) \ ,$$

where T is the absolute temperature and k is Boltzman's constant.

Given that $\Delta G_{ij}/kT$ is small (DNA is weakly bound), one can approximate C_{ij} by its linear part

$$C_{ij} \propto 1 - \frac{\Delta G_{ij}}{kT}$$

and therefore postulate the linear relationship

$$C_{ij} = a + b\Delta G_{ij} \tag{2}$$

with positive b. However, as it will be shown later, correlation function's sum rules impose constraints that will, in general, conflict with the previous equation, and, at the end, the C_{ij}'s will reflect a compromise between the antagonism of thermodynamics and sum rule constraints.

One would expect that the more physiology-compliant column in Table 1 consistent with *in vivo* conditions would follow, within the sum rule restrictions (see next section), Eq. (2) for low values of the free energy. The precise meaning of "low" will be soon elucidated.

3. Sum rules

The local correlation function (1) has some important properties, among them:

$$\sum_i C_{ij} = \sum_j C_{ij} = 0 \ .$$

Table 2

	$\sum_i \Delta G_{Ti}$	$\sum_i \Delta G_{Ai}$
Gotoh (Ref. [7])	2.55	2.51
Vologodskii (Ref. [8])	4.27	4.21
Breslauer (Ref. [4])	5.63	5.33
Sugimoto (Ref. [9])	5.30	5.10
Unified (Ref. [10])	4.33	4.60
Delcourt (Ref. [11])	3.68	3.74

This is obvious from the definition of the correlation function. In fact

$$\sum_i C_{ij} = \sum f_{ij} - f_j \sum_i f_i = f_j - f_j \, .$$

In particular, for strong (S) and weak (W) bases

$$C_{SS} = C_{WW} = -C_{SW} = -C_{WS} \, .$$

From which, after postulate (2) $C_{ij} = a + b\Delta G_{ij}$, it follows that

$$4a + b \sum_i \Delta G_{ij} = 0$$

and therefore

$$\sum_i \Delta G_{ij} = \text{constant} \, . \tag{3}$$

This analytical constraint is not obeyed by the ΔG_{ij} columns in Table 1, take for example Breslauer's column (the same happens for the rest)

$$\sum_i \Delta G_{Ci} = 9.18$$

while

$$\sum_i \Delta G_{Ti} = 5.63 \, .$$

However, if we exclude from the sums the addends corresponding to strong–strong pairs we get that (3) holds (Table 2) within an acceptable 6% error in the worst case (the unified column in Table 1).

We can thus predict that the linear relationship (2) between C_{ij} and ΔG_{ij} should hold for low values where "low" means all the dimers excepting CC, CG, GC and GG.

4. Results and discussion

We calculated C_{ij} (Table 3) in a sample of 12 organisms representing the domains eukaryota, archaebacteria and eubacteria. The data were taken from the GenBank database as it was in January 15th, 2002. Unless otherwise stated, we worked with complete genomes.

Table 3
Correlation function C_{ij} values ($\times 1000$) for the 12 organisms analyzed

	Hs	Mm	Dm	At	Ec	St	Cj	Mt	Tm	Mj	Ss	Hsp
C_{AA}	10.7	5.7	17.9	13.6	12	12	30.9	1-6	13.5	15.5	3.7	−1.9
C_{AC}	−9.7	−7.6	−9.2	−5.3	−7.3	−9.5	−16	2.4	−7.7	−15	−8.7	14.1
C_{AG}	9.4	14.1	−6.8	1.7	−11	−12	4.6	−11	6.2	5.7	9.8	−11.6
C_{AT}	−10.4	−12.2	−1.8	−9.9	6.2	9.1	−20	7.4	−12	−6.5	−4.8	−0.5
C_{CA}	12.2	14.4	8.9	5.9	7.5	2.2	1.7	6.2	−2.3	1.7	−6.8	−4.2
C_{CC}	9.7	9.2	1.7	1.3	−6.1	−6.5	2.4	−14	−0.6	9.1	7.7	−25.8
C_{CG}	−31.2	−37.3	−3.7	−8.9	10	16	−8.9	20	−4.2	−17	−11	41.5
C_{CT}	9.4	13.7	−7	1.7	−12	−12	4.8	−12	7.1	6	9.6	−11.5
C_{GA}	−0.9	1.3	−6.3	6.6	−4.8	−5.2	−4.7	4.8	25.1	2.9	2.6	18
C_{GC}	0.9	−3	13.7	−2.5	18	21	17.4	7.3	−17	3.2	−1.7	−6.6
C_{GG}	9.7	8.9	1.8	1.3	−6.1	−6.1	2.6	−15	−0.1	9.7	7.9	−25.6
C_{GT}	−9.7	−7.3	−9.2	−5.4	−7.3	−9.7	−15	2.8	−8.3	−16	−8.7	14.1
C_{TA}	−21.9	−21.4	−20.6	−26	−15	−9.3	−28	−13	−36	−20	0.5	−11.9
C_{TC}	−0.9	1.4	−6.2	6.4	−4.9	−5	−4.1	4.4	25	2.4	2.8	18.3
C_{TG}	12.1	14.2	8.7	6	7.1	2	1.7	6.6	−2	1.5	−7.2	−4.3
C_{TT}	10.7	5.8	18.1	13.6	13	12	30.4	1.7	13.3	16.3	3.9	−2.1

Hs stands for *Homo sapiens*, Mm for *Mus musculus* (common mouse), Dm—*Drosophila melanogaster* (fruit fly), At—*Arabidopsis thaliana*, Ec—*Escherichia coli*, St—*Salmonella typhi*, Cj—*Campylobacter jejuni*, Mt—*Mycobacterium tuberculosis*, Tm—*Termotoga maritima*, Mj—*Methanococcus jannaschii*, Ss—*Sulfolobus solfataricus* and Hsp—*Halobacterium sp*. The values were directly calculated using home made software on complete genomes excepting *H. sapiens* and *M. musculus* where only chromosomes 21 and 2 were, respectively, analyzed.

In order to assess our prediction (2) we plotted C_{ij} vs. ΔG_{ij} for the six datasets and the 13 organisms chosen. Due to the overwhelming amount of plots, we show only those corresponding to eukaryotic organisms (*H. sapiens*, *M. musculus*, *D. melanogaster* and *A. thaliana*) merged in a single plot per author (Figs. 1–6). The remaining plots do not show any discordant behavior and, in any case, their informational contents is summarized in Table 4 by its linear regression coefficient r.

The foregoing results show that Breslauer's column in Table 1 agrees with our theoretical considerations. Having no intention of engaging in a discussion about laboratory protocols or techniques, we are just in position to state that Breslauer data is the best (contradicting SantaLucia's [3] conclusions) in the light of genomics and for *in vivo* DNA local interactions.

Our C_{ij} vs. ΔG_{ij} linear relationship is evident in eukaryotic genomes dominated by intergenic sequences, meaning that this relation reflects DNA structural constraints and does not depend on the DNA protein coding requirements. To stress this point, we include the human exonic sequences [12] and plot them (Fig. 7). They look pretty much the same as human intergenic sequences and their linear regression coefficient r are almost identical.

A closer look to Table 4 shows that Breslauer's data are the most consistent with the thermodynamical constraints for all the organisms analyzed with one exception: *Halobacterium sp*. Its r value is remarkably low for Breslauer and Gotoh's data and

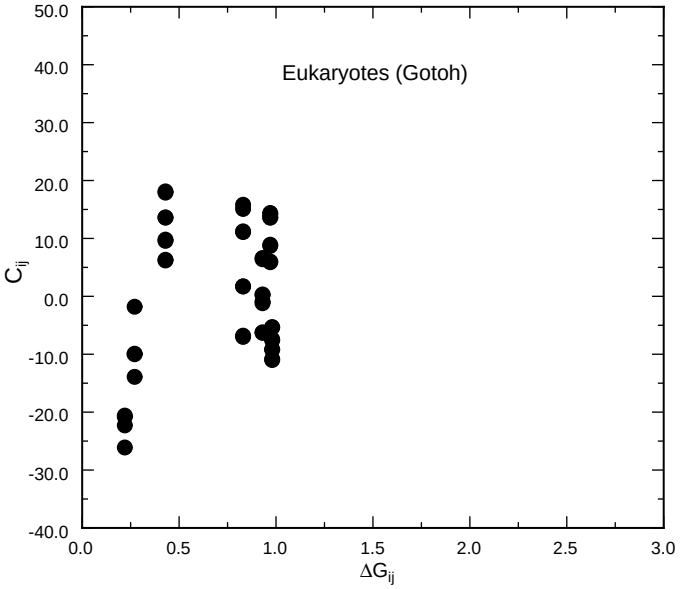

Fig. 1. Free energy (ΔG_{ij}) absolute values vs. local correlation function C_{ij} for every pair ij of overlapping DNA dimers. ΔG_{ij} data from Gotoh [7]. C_{ij} was calculated on *H. sapiens* (chromosome 21), *M. musculus* (chromosome 2), *A. thaliana* (complete genome) and *D. melanogaster* (complete genome).

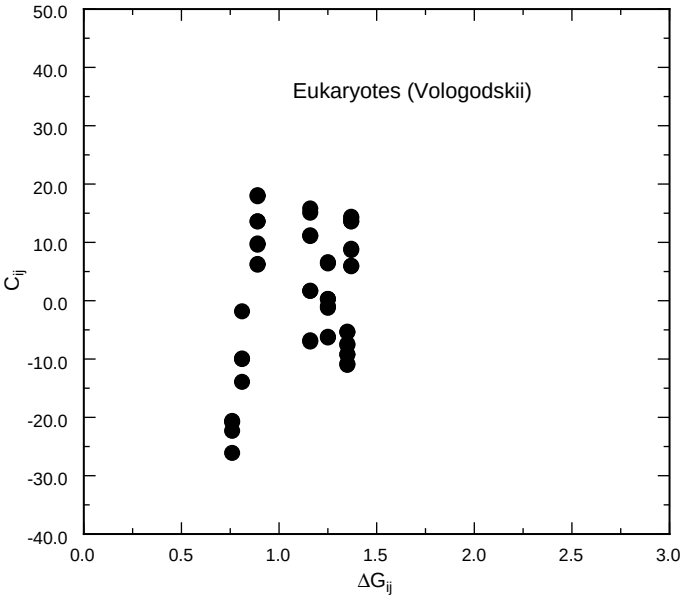

Fig. 2. Free energy (ΔG_{ij}) absolute values vs. local correlation function C_{ij} for every pair ij of overlapping DNA dimers. ΔG_{ij} data from Vologodskii [8]. C_{ij} was calculated on *H. sapiens* (chromosome 21), *M. musculus* (chromosome 2), *A. thaliana* (complete genome) and *D. melanogaster* (complete genome).

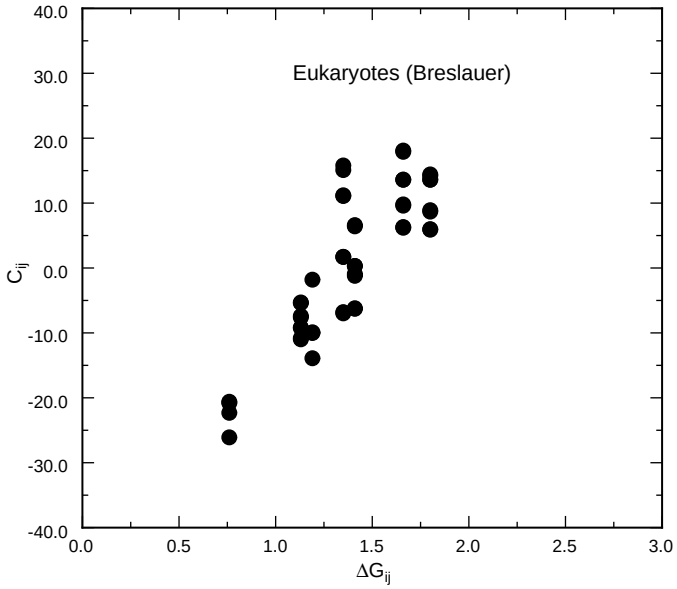

Fig. 3. Free energy (ΔG_{ij}) absolute values vs. local correlation function C_{ij} for every pair ij of overlapping DNA dimers. ΔG_{ij} data from Breslauer [4]. C_{ij} was calculated on *H. sapiens* (chromosome 21), *M. musculus* (chromosome 2), *A. thaliana* (complete genome) and *D. melanogaster* (complete genome).

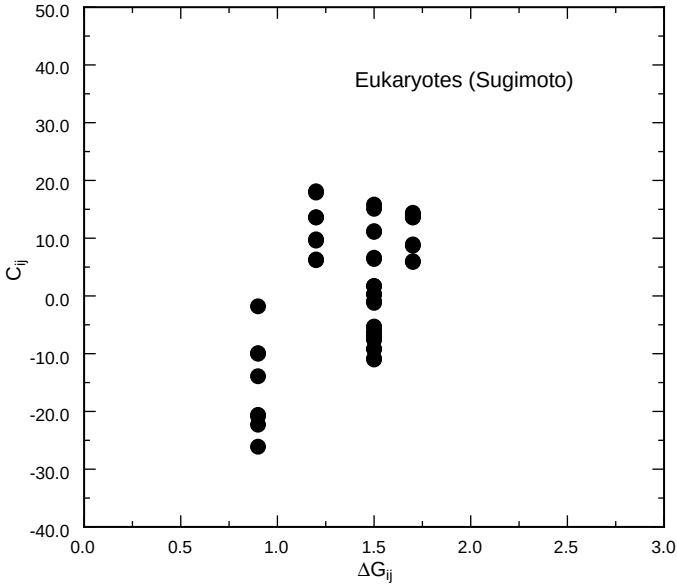

Fig. 4. Free energy (ΔG_{ij}) absolute values vs. local correlation function C_{ij} for every pair ij of overlapping DNA dimers. ΔG_{ij} data from Sugimoto [9]. C_{ij} was calculated on *H. sapiens* (chromosome 21), *M. musculus* (chromosome 2), *A. thaliana* (complete genome) and *D. melanogaster* (complete genome).

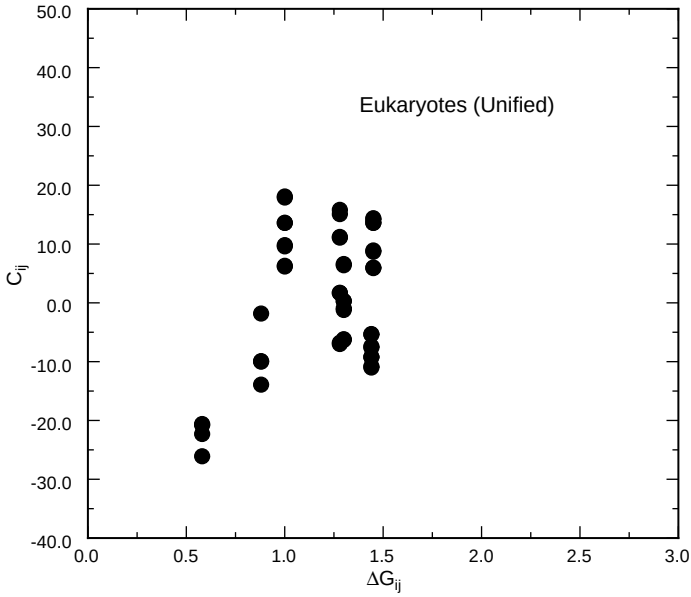

Fig. 5. Free energy (ΔG_{ij}) absolute values vs. the local correlation function C_{ij} for every pair ij of overlapping DNA dimers. ΔG_{ij} data from unified [3]. C_{ij} was calculated on *H. sapiens* (chromosome 21), *M. musculus* (chromosome 2), *A. thaliana* (complete genome) and *D. melanogaster* (complete genome).

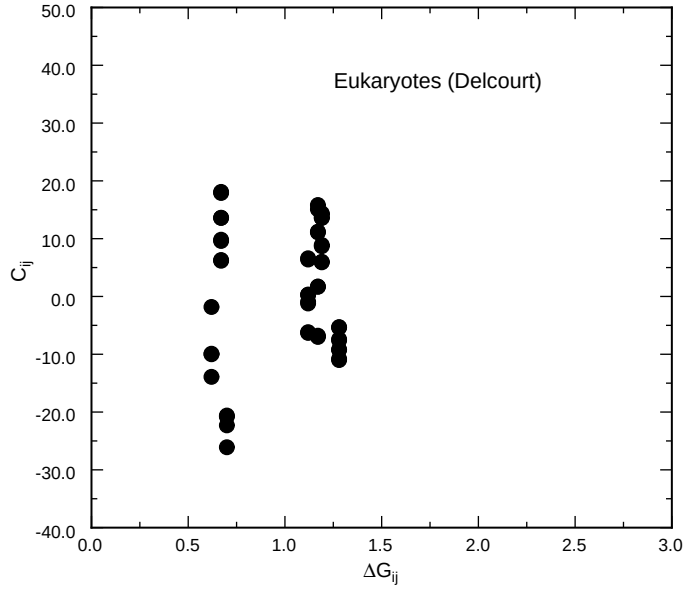

Fig. 6. Free energy (ΔG_{ij}) absolute values vs. local correlation function C_{ij} for every pair ij of overlapping DNA dimers. ΔG_{ij} data from Delcourt [11]. C_{ij} was calculated on *H. sapiens* (chromosome 21), *M. musculus* (chromosome 2), *A. thaliana* (complete genome) and *D. melanogaster* (complete genome).

Table 4
Coefficient r of linear correlation between ΔG_{ij} and C_{ij} for twelve organisms plus human exons. Notice the prevalence of the column corresponding to Breslauer ([4]) data excepting the two last rows

	Gotoh	Vologodskii	Breslauer	Sugimoto	Unified	Delcourt
H. sapiens (Chr 21)	0.0324	0.285	0.896	0.570	0.440	0.235
H. sapiens (exons)	0.409	0.370	0.852	0.655	0.503	0.271
M. musculus (Chr 2)	0.371	0.432	0.830	0.692	0.561	0.382
D. melanogaster	−0.094	−0.070	0.870	0.129	0.104	−0.303
A. thaliana	0.349	0.296	0.875	0.505	0.470	0.120
E. coli	−0.217	−0.084	0.753	−0.314	−0.012	−0.427
S. typhi	−0.490	−0.538	0.593	−0.347	−0.303	−0.618
C. jejuni	−0.040	−0.091	0.763	0.220	−0.389	−0.215
M. tuberculosis	0.237	0.316	0.480	0.217	0.367	0.046
T. maritima	0.392	0.294	0.603	0.439	0.423	0.186
M. jannaschii	−0.105	−0.130	0.762	0.150	0.062	−0.241
S. solfataricus	−0.224	−0.370	0.230	−0.145	−0.286	−0.214
Halobacterium sp.	−0.454	0.477	−0.066	0.254	0.410	0.342

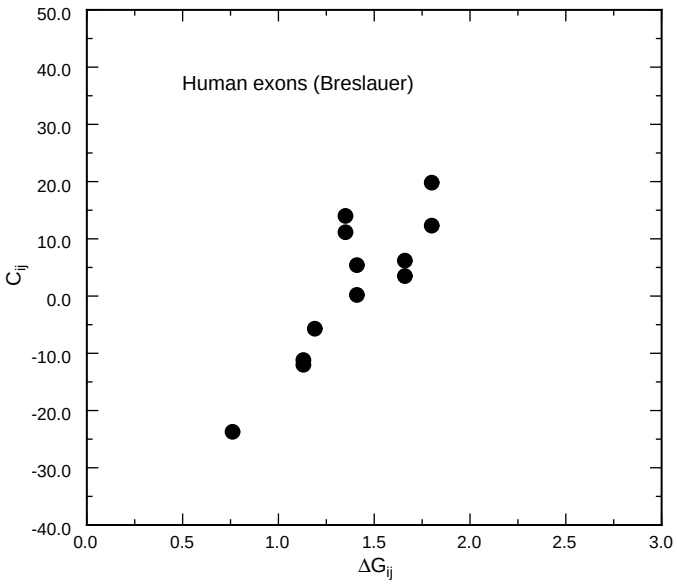

Fig. 7. Free energy (ΔG_{ij}) absolute values vs. local correlation function C_{ij} for every pair ij of overlapping DNA dimers. ΔG_{ij}. C_{ij} was calculated on dimer data taken from Karlin and Mrzek [12].

between 0.254 and 0.477 for the remaining datasets. A plausible explanation for this situation could lie on the fact that this species live in extreme saline environments [13]. The interaction between cations and DNA might imply additional thermodynamic restrictions besides those already discussed.

The free energy differences are related to equilibrium or near-equilibrium thermodynamics where, both, enthalpy and entropy, contribute. The plots C_{ij} vs. ΔH_{ij} (not shown) does not show any linear relationship. This feature suggests that the correlations are not dominated by far-from-equilibrium (molecular machine aspects) or the replication process.

We have shown the convenience of theoretical thinking in biology. The discrepancies among different labs cannot be elucidated within their framework by discussing details about the methods employed. *In vitro* protocols, as perfect as they can be, are just an approximation to *in vivo* reality, to fill the gap conceptual bridges are in demand.

Acknowledgements

We are grateful to the SMM-CONACYT program of Applied Mathematics and Education. PM wishes to thank the IFUNAM hospitality. We thank the useful comments of an anonymous referee.

References

[1] T.A. Brown, Genomes, Wiley, New York, 1999.
[2] P. Miramontes, L. Medrano, C. Cerpa, R. Cedergren, G. Ferbeyre, G. Cocho, J. Mol. Evol. 40 (1995) 698–704.
[3] J. SantaLucia, Proc. Natl. Acad. Sci. USA 95 (1998) 1460–1465.
[4] K.J. Breslauer, F. Frank, H. Blcker, L.A. Marky, Proc. Natl. Acad. Sci. USA 83 (1986) 3746–3750.
[5] M.J. Doktycz, R.F. Goldstein, T.M. Paner, F.J. Gallo, A.S. Benight, Biopolymers 32 (1992) 849–864.
[6] R. Owczarzy, P.M. Vallone, R.F. Goldstein, A.S. Benight, Biopolymers 52 (1999) 29–56.
[7] O. Gotoh, Y. Tagashira, Biopolymers 20 (1981) 1033–1042.
[8] A.V. Vologodskii, B.R. Amirikyan, Y.L. Lyubichenko, M.D. Frank-Kamenetskii, J. Biomol. Struct. Dyn. 2 (1984) 131–148.
[9] N. Sugimoto, S. Nakano, M. Yoneyama, K. Honda, Nucleic Acids Res. 24 (1996) 4501–4505.
[10] H.T. Allawi, J. SantaLucia, Biochemistry 36 (1997) 10 581–10 594.
[11] S.G. Delcourt, R.D. Blake, J. Biol. Chem. 266 (1991) 15 160–15 169.
[12] S. Karlin, J. Mrazek, J. Mol. Biol. 262 (1996) 459–472.
[13] F. Frolow, M. Harel, J.L. Sussman, M. Mevarech, M. Shoham, Nat. Struct. Biol. 3 (1996) 452–458.