ELSEVIER

Biophysical Chemistry 110 (2004) 25–38

Biophysical Chemistry

www.elsevier.com/locate/bpc

Coding and non-coding DNA thermal stability differences in eukaryotes studied by melting simulation, base shuffling and DNA nearest neighbor frequency analysis

Dang D. Long[a], Ivo Grosse[b], Kenneth A. Marx[a,*]

[a] *Center for Intelligent Biomaterials, Department of Chemistry, University of Massachusetts Lowell, One University Ave., Lowell, MA 01854, USA*

[b] *Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA*

Received 14 August 2003; received in revised form 8 January 2004; accepted 8 January 2004
Available online 5 May 2004

Abstract

The melting of the coding and non-coding classes of natural DNA sequences was investigated using a program, MELTSIM, which simulates DNA melting based upon an empirically parameterized nearest neighbor thermodynamic model. We calculated T_m results of 8144 natural sequences from 28 eukaryotic organisms of varying F_{GC} (mole fraction of G and C) and of 3775 coding and 3297 non-coding sequences derived from those natural sequences. These data demonstrated that the T_m vs. F_{GC} relationships in coding and non-coding DNAs are both linear but have a statistically significant difference (6.6%) in their slopes. These relationships are significantly different from the T_m vs. F_{GC} relationship embodied in the classical Marmur–Schildkraut–Doty (MSD) equation for the intact long natural sequences. By analyzing the simulation results from various base shufflings of the original DNAs and the average nearest neighbor frequencies of those natural sequences across the F_{GC} range, we showed that these differences in the T_m vs. F_{GC} relationships are largely a direct result of systematic F_{GC}-dependent biases in nearest neighbor frequencies for those two different DNA classes. Those differences in the T_m vs. F_{GC} relationships and biases in nearest neighbor frequencies also appear between the sequences from multicellular and unicellular organisms in the same coding or non-coding classes, albeit of smaller but significant magnitudes.

Keywords: DNA melting simulation; Exon–intron; Sequence shuffling; Nearest neighbor frequencies; Melting temperature

1. Introduction

The process of DNA denaturation, followed by a hybridization event, was one of the earliest and most important technology developments underlying and driving the evolution of biotechnology in the academic and commercial arenas. Historically, the DNA double helix thermal denaturation process has been mostly studied with optical methods [1]. Many early biophysical studies of whole organism genomic DNAs and fractionated purified components such as satellite DNAs, employed experimental DNA

* Corresponding author. Tel.: +1-978-934-3658; fax: +1-978-934-3013
E-mail address: kenneth_marx@uml.edu (K.A. Marx).

0301-4622/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2004.01.001

melting to help understand the structure and organization of these complex DNAs [2–6]. Investigations of the DNA thermodynamic stability have focused on correlations with the DNA base composition [1,7], base substitution, mutation or polymorphism [8–10] and its relationship to the DNA hybridization process [1,11]. It has been long known from various experiments that the melting process of any DNA molecule depends on the solution pH, counterion type and concentration and the presence of solutes in the solution environment, as well as the (G+C) content, the length, and the sequence of the DNA [1,12].

Statistical thermodynamic models of the dsDNA melting process have been proposed [13,14], which have led to algorithms, using empirically derived thermodynamic parameters, for the simulation of the process [15–17]. It is notable that there are differences in the thermodynamic parameters used in these models for simulating long DNA as opposed to oligonucleotide melting [1,11]. This is necessary in order to make the simulations accurate and agree with experimental results [18,19]. Consequently, in simulations it is necessary to treat the DNA melting process separately for DNA and for double helical oligonucleotides. The most often used parameter resulting from the computation of the melting process is the melting temperature (T_m), which is the temperature at which 50% of the base pairs in all sample duplexes have been dissociated.

Classically, the T_m of long DNA sequences has been correlated with the sequence mole fraction of guanine and cytosine bases, F_{GC}, and the ionic strength of the solution environment through a linear equation, the Marmur–Schildkraut–Doty (MSD) equation [20,21]. This equation was refined over three decade time from the empirically measured T_ms of many complex heterogeneous populations of long length DNA sequences from whole organisms. These DNAs covered a wide range of average F_{GC} values with a large, relatively unbiased distribution of nearest neighbor pair frequencies, f_{ij}, where ij denotes the stack of nucleotide pair i on its neighbor j [1]. The MSD equation, while not representing the nearest neighbor pair dependence widely believed to underlie the DNA melting process, shows a linear dependence of T_m on F_{GC} over a wide F_{GC} range ($0.3<F_{GC}<0.7$) [1].

However, the effect of biased nearest neighbor pair frequencies, found in many natural DNA sequences, on the linear relationship of T_m and F_{GC} studied in the MSD equation, is problematic. Especially as studied sequence lengths get smaller and nearest neighbor biases become more pronounced, the MSD equation becomes an oversimplification of more complex dependencies of DNA stability on sequence domains and the physical states of neighboring domains. In the case of extreme nearest neighbor bias in the DNA sequences being melted (for example, d(G·C)·d(G·C), d(A·T)·d(A·T)), quite large deviations have been observed between the experimentally determined and simulated T_ms [22]. In these extreme instances, the disagreement between experiment and simulation is thought to result from the sequences adopting non-B secondary structures with altered base stacking. Given the limitations in the applicability of the MSD equation to sequences with nearest neighbor bias, we decided to investigate the relationship between T_m and F_{GC} in relatively short DNA sequences containing normal bias in neighbor-pair frequencies. For this study, we chose a large number of coding and non-coding sequence elements in eukaryotic genomes, a system that had not been investigated.

Using the MELTSIM program [7,23], we performed a large-scale simulation of the DNA melting of over 8000 single gene-containing DNA sequences obtained from GenBank that were contained within 28 eukaryotic organisms' genome files across the entire %(G+C) range. This type of large-scale survey study of separate gene-containing sequences, containing a wide range of nearest neighbor pair frequencies, has not been described before. MELTSIM is a program based on a statistical mechanics algorithm described by Fixman and Freire [17], in which empirically parameterized nearest neighbor base pair frequencies are used to simulate DNA melting. The program has been shown to simulate accurately the melting processes of a number of specific DNA populations from different organisms that have also been studied experimentally. The T_m values calculated by the MELTSIM program are in agreement with the experimental T_ms of various DNAs with different F_{GC} and lengths [7,22–26]. Here, we compare the T_m values computed from melting simulations of two functionally different kinds of DNA sequences, protein-coding sequences

(exons) and non-coding sequences (introns and gene-flanking DNA) from the single gene-containing sequences. We show that relationships between T_m and F_{GC} for the two functionally different genome regions are both linear like the classic MSD-type relationship, but that they possess a statistically significant 6.6% difference in their slope. Next, via a comparison of T_ms of all sequences to T_ms of their identical F_{GC} shuffled versions, we demonstrate that this exon–intron difference results from a systematic difference in the distribution of nearest-neighbor pairs in coding and non-coding genomic regions. No higher order neighbor pair information is necessary to account for the specific T_m vs. F_{GC} relationships. Finally, we demonstrate significant differences in the T_m and F_{GC} relationship between single-cell and multi-cell organisms, within both their coding and non-coding classes. These differences are due to differential f_{ij} trends found in a subset of nearest neighbor pairs across the F_{GC} range.

2. Materials and methods

2.1. DNA sequence selection

In this study, nucleic DNA sequences coding for single-genes from 28 eukaryotes (Table 1) were included in order to cover the whole range of F_{GC} of typical eukaryotic genomes. These sequences were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi? db=Nucleotide) within the dates April 15–25, 2001. In order to minimize sequence-specific bias, duplicate sequences and alleles

Table 1
List of the 28 organisms studied and composition of their sequences

Organisms	Unicellular (U) or multi-cellular (M)	Average %(G+C)	Number of source sequences and total bps	Number of coding sequences and total bps	Number of non-coding sequences and total bps
Anopheles gambiae	M	48.95	17/60 030	4/7357	10/32 761
Arabidopsis thaliana	M	38.34	885/2 844 052	141/242 310	334/573 281
A. niger	M	52.55	147/400 451	52/84 809	0/0
Caenorhabditis elegans	M	35.92	282/4 757 788	42/81 192	87/190 536
Candida albicans	U	35.25	296/701 662	243/496 201	50/68 702
Danio rerio	M	38.70	177/493 124	29/37 877	75/296 545
D. discoideum	U	26.58	286/824 350	180/430 553	57/82 416
Drosophila melanogaster	M	43.93	599/2 767 024	244/419 398	383/1 644 670
Emericella nidulans	M	50.77	263/1 090 070	149/325 646	110/176 425
Entamoeba histolytica	U	29.76	72/170 130	41/92 986	3/3855
Fugu rubripes	M	45.91	124/1 485 333	33/62 718	271/874 888
Gallus gallus	M	49.96	294/1 103 275	47/80 463	220/467 685
H. sapiens	M	48.37	264/993 148	189/248 222	201/616 788
L. major	U	60.68	103/2 188 745	102/259 197	11/16 461
M. sexta	U	34.18	19/88 840	0/0	19/46 244
Mus musculus	M	47.83	612/2 300 980	142/238 100	317/699 303
Neurospora crassa	M	51.93	250/1 993 520	241/571 197	249/693 151
Oryza sativa	M	44.56	305/1 334 458	64/118 268	197/436 572
P. falciparum	U	24.63	256/790 013	260/702 528	22/48 524
S. cerevisiae	U	38.80	1516/4 015 270	1015/2 100 566	184/265 488
Schistosoma mansoni	M	38.79	55/134 183	10/23 256	14/22 656
Schzosaccharomyces pombe	U	37.71	424/1 177 433	254/524 491	106/167 277
Strongylocentrotus purpuratus	M	41.35	39/95 468	5/7997	17/50 514
T. thermophila	M	27.26	76/168 395	21/26 049	10/15 734
Toxoplasma gondii	U	51.27	69/199 507	26/41 175	21/39 714
Trypanosoma brucei	U	45.49	212/707 014	148/250 084	63/146 887
Xenopus laevis	M	41.51	177/567 675	40/65 735	99/178 261
Zea mays	M	46.67	325/1 181 564	53/105 110	167/345 212
Total			8144/34 633 502	3775/7 643 485	3297/8 200 550

of the same gene were removed using the CLEANUP program [27], which removed from the input sequence collection the shorter of two matching sequences having identity and overlapping percentage above 95%. The non-redundant sequences were then filtered to remove the sequences with over 1% of unidentified nucleotides. Then, from each of the filtered single-gene DNA sequences (source sequence), coding (exon) and non-coding (flanking and intron) regions were separated by the COMPILE program (freely available at http://bioinformatics.org/meltsim). In this process, all the nucleotides that have the same functional structure (coding or non-coding) were placed into one sequence file designated coding or non-coding. Furthermore, in order to minimize variations in our T_m results due to the known end-effect (sequence terminal thermodynamic bias) and the effect of DNA sequence length [1], we selected only the source, coding, and non-coding sequences above 1000-bp length for the melting simulation. The numbers of the sequence files received from each organism and the average %(G+C) in DNA sequences of each organism are summarized in Table 1.

2.2. DNA melting simulation and sequence shuffling

After separating the coding and non-coding nucleotides from each GenBank document into different files, melting was separately simulated for each individual sequence using the MELTSIM program [7,23], which is available for download at http://bioinformatics.org/meltsim. In this paper, we focused on the calculation of melting temperatures (T_m) of the DNA sequences at one condition of the counterion concentration—in this case $[Na^+]$—is equal 0.075 M.

In order to validate the effect of nearest-neighbor pairs on the DNA melting simulation process, we shuffled the sampled sequences separately for each coding or non-coding DNA sequence received after the cleaning and filtering processes. There were three shuffling procedures: (1) randomized shuffling of all nucleotides along the sequence ('random shuffling'); (2) randomized shuffling of nucleotides within the first, the second, and the third positions of a 3-nucleotide frame repeating along the sequence, respectively, ('in-frame shuffling'); and (3) a random shuffling procedure, in which all of the dinucleotide or oligonucleotide frequencies of the sequence are conserved ('conservative shuffling'). In the random shuffling procedure, all the nucleotides' positions in the original sequence were randomly swapped with the help of a random number generator, and a new DNA sequence was created with the same F_{GC} but with a new random distribution of nearest neighbor pair frequencies. In the in-frame shuffling procedure, the nucleotides in each position of a 3-nucleotide repeating frame were randomly swapped within their frame positions. Therefore, the newly created DNA sequence had the same in-frame nucleotide compositions as well as overall F_{GC} as compared to the original sequence. However, this in-frame shuffling procedure still changed randomly the nearest neighbor pair frequencies of the original sequences. The conservative shuffling procedure employed the Euler paths on directed graphs algorithm of Kandell et al. [28], which is implemented by the Shufflet program of Coward [29]. This procedure produced a uniform sample of randomly shuffled sequences, which conserve the exact counts of all words equal to or shorter than a given length k. When we chose k=2, we received randomly shuffled sequences, in which the nearest neighbor pair frequencies (f_{ij}) were all maintained. The melting processes of the shuffled sequences were also simulated by MELTSIM and the T_ms were compared with the results of the original sequences as difference representations. The programs to carry out these procedures are available in the supplementary material. The nearest neighbor pair frequencies, f_{ij}, were calculated by taking the counted number of the nucleotide pair ij in a sequence divided by the total number of all the nearest neighbor pairs. The graphs and statistical procedures were carried out using the SigmaPlot 2002 for Windows v. 8.0 from SPSS (http://www.spssscience.com/) and scripts in R language (www.r-project.org).

3. Results

3.1. The T_m vs. F_{GC} relationship is different in natural coding and non-coding DNA sequences

Natural DNA sequences, which contain single-genes, were retrieved from 28 representative eu-

karyotic organisms (see Table 1). These eukaryotes cover a wide range of GC content in their genomes, from GC-rich, e.g. *Leishmania major*, *Aspergilus niger*, to GC-poor, e.g. *Distyostelium discoideum*, *Plasmodium falciparum*. They also include well-sequenced genomes like *Saccharomyces cerevisiae*, *Homo sapiens*, as well as, sparsely-sequenced genomes like *Tetrahymena thermophila*, *Manduca sexta*. The total number of sequences and base pairs of each class of DNA sequences are shown in Table 1.

There has long been known to be a length-dependence of the T_m of any DNA sequence. Due to thermodynamic end effects—the end melting preferentially in any given sequence—as sequences shorten to hundreds of base pairs length, then T_m begins to decrease significantly. In this investigation, we only considered the natural sequences longer than 1000 bps because it had been shown that when the sequence length (N) is greater than approximately 1000 bps, the $T_{m,N}$-melting temperature of oligomers of length N, approximates the T_m of a 'corresponding' DNA of infinite length with identical nearest-neighbor composition [1]. Nevertheless, for all those selected sequences, we examined the average lengths over the F_{GC} range. These data are shown in Table 2 for five different F_{GC} bins. No consistent average length trend is statistically evident in any of the three kinds of DNA sequences studied in this paper. Therefore, we calculated and compared T_ms of these long sequences without further considering their actual lengths.

Using the MELTSIM program, we calculated the simulated T_ms of the source sequences (original sequences containing both coding and non-coding regions) and plotted them against their F_{GC} in Fig. 1. The data in Fig. 1 were fitted with a linear regression, giving the following equation:

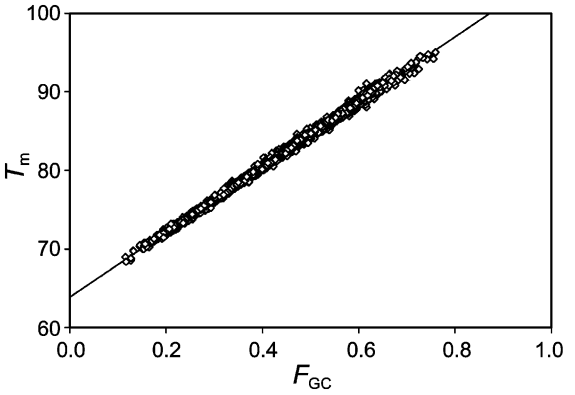

Fig. 1. T_m vs. F_{GC} of the source sequences of the organisms.

$$T_{m\_sr} = 41.44\, F_{GC} + 63.90 \quad (1a)$$

(R^2=0.9969, and the coefficient had P<0.0001). A recent accurate representation of the MSD equation, determined from experimental DNA melting, has the formula:

T_m (°C) = 193.67 − (3.09 − F_{GC})(34.64 − 2.83·ln[Na^+]) [1], and in the specific condition of [Na^+]=0.075 M, we can write:

$$T_m = 41.97 \cdot F_{GC} + 63.98 \quad (1b)$$

Obviously, the relationship in Eq. (1a) agrees quite well with the relationship derived from the MSD equation (Eq. (1b)). However, when we plotted the simulated T_m (°C) vs. F_{GC} of the natural separated coding and non-coding sequences (see Fig. 2), we obtained two different linear relationships that both

Table 2
Length distributions of the DNA sequences studied across their range of F_{GC}

F_{GC}	Whole sequences (bps)		Coding sequences (bps)		Non-coding sequences (bps)	
	Average	(STD)	Average	(STD)	Average	(STD)
0.1–0.2	5026.3	(14 435.5)	2293.3	(981.1)	1610.2	(931.2)
0.2–0.3	2997.7	(3492.3)	2715.1	(1983.6)	1706.0	(1036.4)
0.3–0.4	4193.6	(6794.2)	2288.9	(1366.9)	2528.7	(3456.7)
0.4–0.5	3859.3	(5886.8)	1835.8	(1043.2)	3175.4	(4365.7)
0.5–0.6	4104.5	(8051.9)	1772.8	(1067.5)	2377.7	(3674.4)
0.6–0.7	7280.9	(12819.4)	1365.3	(615.5)	1640.0	(606.8)
0.7–0.8	1551.4	(502.3)	–	–	1727.4	(928.8)

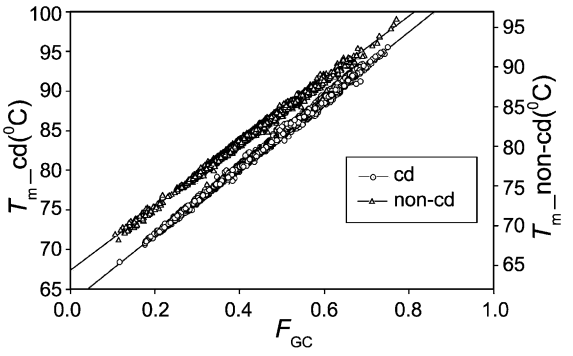

Fig. 2. T_m vs. F_{GC} of the coding and non-coding sequences of all the organisms. Open circles and open triangles represent T_m values of the coding and non-coding sequences, respectively.

differed significantly from the Eq. (1b) MSD-type relationship. The linear regression result for the coding sequences was:

$$T_{m\_cd} = 42.75 \cdot F_{GC} + 63.29 \quad (2a)$$

(R^2=0.9970, P<0.0001); and the result for non-coding sequences was:

$$T_{m\_non-cd} = 40.11 \cdot F_{GC} + 64.39 \quad (2b)$$

(R^2=0.9958, P<0.0001). The slope for coding DNA in Eq. (2a) is 6.6% higher than the slope for non-coding DNA in Eq. (2b). In order to determine whether that difference was significant, we calculated the standard errors of the coefficients in Eqs. (2a) and (2b) by the non-parametric bootstrap method [30]. With 1000 bootstrap cycles, we obtained the standard errors of the slope and the intercept of Eq. (2a) to be 0.04 and 0.0173, respectively. Similarly, standard errors of the slope and the intercept of Eq. (2b) were 0.04 and 0.0172, respectively. Therefore, we could conclude that the difference between these two equations was highly statistically significant as were the differences between these two equations and the Eq. (1a), total source sequence, and Eq. (1b), the MSD relationship.

3.2. The T_m vs. F_{GC} relationship of shuffled DNA sequences demonstrates a nearest neighbor dependence

First, we considered the DNAs obtained from the 'random shuffling' of individual coding and non-coding sequences. This shuffling procedure maintained the overall base composition (constant mononucleotide frequencies, f_A, f_T, f_G, f_C) of the original sequences, but randomly changed every higher order neighbor base frequency in the DNAs (f_{ij}, f_{ijk}, etc.). The T_ms of those random shuffled sequences were computed by the MELTSIM program. In order to ensure that we deal with the properties of a truly randomized sequence, each coding or non-coding sequence was randomly shuffled ten times. Each time a T_m of the resulting sequence was calculated, and then these T_ms were averaged. The average T_ms of the random shuffled sequences were plotted against their F_{GC} in Fig. 3. Here, both of the regression lines are similar. For the random shuffled coding sequences:

$$T_{m\_Sh\_cd} = 41.69 \cdot F_{GC} + 63.83, \quad (3a)$$

for which R^2=0.9993, P<0.0001; and for the random shuffled non-coding sequences:

$$T_{m\_Sh\_non-cd} = 41.69 \cdot F_{GC} + 63.84, \quad (3b)$$

for which R^2=0.9989, P<0.0001. These nearly identical relationships demonstrate that randomized nearest neighbor f_{ij} values have removed all the F_{GC}-dependent melting differences that distinguish coding from non-coding sequences—as we would expect.

In a similar way, each coding and non-coding sequence went through the 'in-frame shuffling' process, which maintained the base composition of

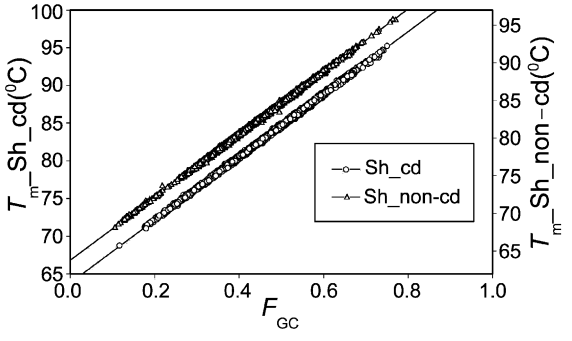

Fig. 3. Average T_m vs. F_{GC} of the randomly shuffled coding and non-coding sequences of all the organisms. Open circles and open triangles represent T_m values of the randomly shuffled coding and non-coding sequences, respectively.

the whole original sequence as well as the base composition in each position of a repeating 3-nucleotide frame along the original sequence. Each sequence also went through this shuffling procedure 10 times, and then an average T_m was calculated. Fig. 4 depicts the relationship between the average T_ms and F_{GC} of those 'in-frame shuffled' sequences. The linear regression results here were very similar with the random shuffled results from Fig. 3. Specifically, for the 'in-frame shuffled' coding DNAs:

$$T_{m\_Sh3\_cd} = 41.59 \cdot F_{GC} + 63.88, \quad (4a)$$

for which R^2=0.9995, P<0.0001; and for the 'in-frame shuffled' non-coding DNAs:

$$T_{m_{Sh3}non-cd} = 41.73 \cdot F_{GC} + 63.82, \quad (4b)$$

for which R^2=0.9994, P<0.0001. These results show that a small residual difference of 0.3% exists between coding and non-coding sequences that is nearly within the standard error limits for these slopes. Therefore, we do not consider these slope differences to be significant.

Finally, the 'conservative shuffling' process, which maintained both the base composition and the nearest neighbor frequencies (f_{ij}) of the original sequences, was applied to the coding and non-coding sequences in the same manner as with the two previous shuffling procedures. Fig. 5 represents the relationship between

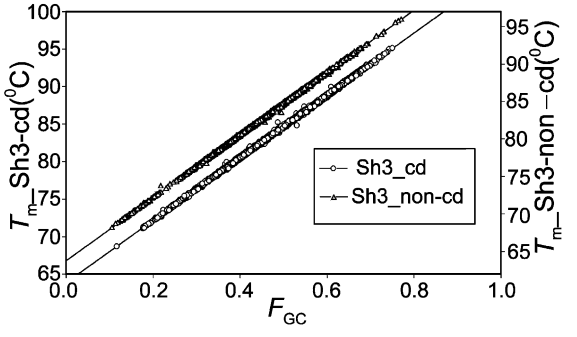

Fig. 4. Average T_m vs. F_{GC} of the in-frame shuffled coding and non-coding sequences of all the organisms. Open circles and open triangles represent T_m values of the in-frame shuffled coding and non-coding sequences, respectively.

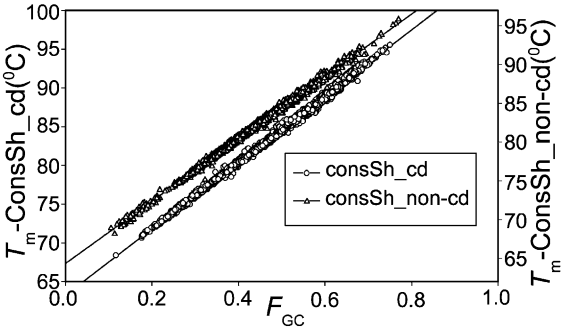

Fig. 5. Average T_m vs. F_{GC} of the dinucleotide frequencies-conservatively shuffled coding and non-coding sequences of all the organisms. Open circles and open triangles represent T_m values of the dinucleotide frequencies-conservatively shuffled coding and non-coding sequences, respectively.

the average T_m and F_{GC} of the 'conservative shuffled' sequences. Those 'conservative shuffled' coding and non-coding sequences both had very good linear relationships but with different coefficients, especially the slopes. For the 'conservative shuffled' coding sequences:

$$T_{m\_ConsSh\_cd} = 42.78 \cdot F_{GC} + 63.27, \quad (5a)$$

for which R^2=0.9970, P<0.0001; and for the 'conservative shuffled' non-coding sequences:

$$T_{m\_ConsSh\_non-cd} = 40.23 \cdot F_{GC} + 64.35, \quad (5b)$$

for which R^2=0.9959, P<0.0001. The slope results in Fig. 5 differ by 6.3% and are, therefore very similar to the results obtained for the original natural coding and non-coding sequences (Eqs. (2a) and (2b)), where a 6.6% difference was observed. In order to confirm that the difference between Eq. (5a) and Eq. (5b) was also statistically significant, we calculated the standard errors of the coefficients in the two equations by the bootstrap method (1000 bootstrap cycles). The standard errors of the slope and the intercept were 0.04 and 0.0172, respectively, for Eq. (5a); and were 0.04 and 0.0173, respectively, for Eq. (5b). Based on these results, we concluded that the 6.3% slope difference between Eq. (5a) and Eq. (5b) was highly statistically significant.

As a way to more easily compare the data in the T_m vs. F_{GC} plots that we presented in Figs. 1–5, the best fit slope values that were determined in Eqs. (1a), (1b),

(2a), (2b), (3a), (3b), (4a), (4b), (5a) and (5b) have been plotted in the Fig. 6 bar chart for each normal and shuffled condition we studied. Here, the statistically significant 6.6% difference in slope between the natural coding (cd) and non-coding (non-cd) sequences that we studied separately is clear. Also clear are the differences between these slopes and the original source sequences (Sr) from which they were obtained, as well as the similarity of the source sequences slope to the MSD equation slope. Then the random shuffled (Sh) and in-frame shuffled (Sh3) sequences results can be seen to produce slopes that are indistinguishable from one another, but are distinctly different than the natural coding and non-coding sequences slopes. However, the conservative shuffled (consSh) sequences reproduce the natural coding and non-coding sequence slopes within the bootstrap determined experimental error.

Another way to compare the T_m vs. F_{GC} relationships in the natural sequences and in the shuffled sequences was to look at the difference in T_m of these sequences across the range of F_{GC}. Fig. 7a,b present the difference in T_m between the natural sequences and the average 'random shuffled' natural sequences of coding and non-coding DNAs, respectively, as a function of their F_{GC}. Although we did not see a good linear regression in either case (in Fig. 7a, R^2=0.1612, and in Fig. 7b, R^2=0.2424), there were clearly two opposite trends for coding (positive slope) and non-coding (negative slope) sequences. As the F_{GC} increased, the difference in T_m of the coding sequences increased significantly, while the difference in T_m of the non-coding sequences decreased by a similar magnitude. However, when we considered the difference

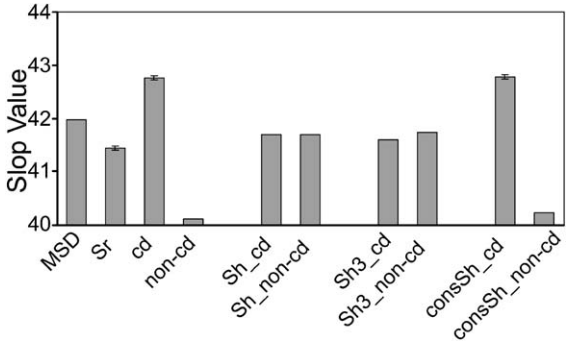

Fig. 6. Bar chart of the best-fit slope values of the linear regression relationships between T_m and F_{GC} of different DNA sequences.

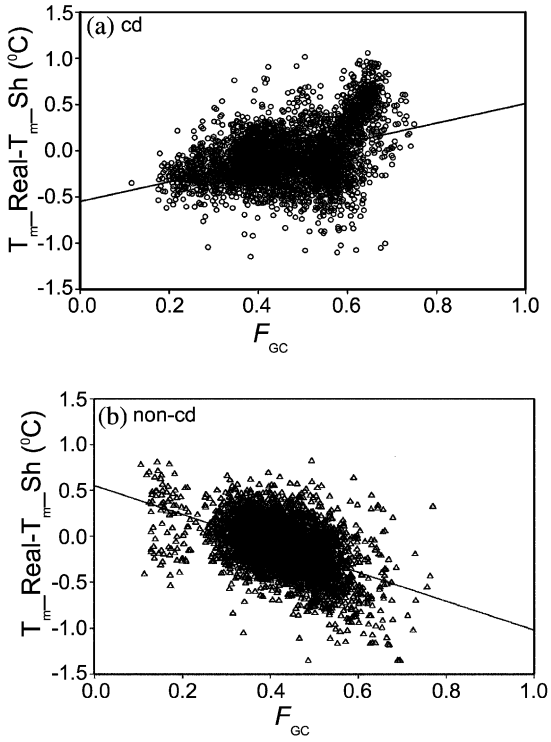

Fig. 7. Plots of the differences between T_ms of the natural sequences and the average T_ms of the randomly shuffled sequences in the cases of (a) coding DNAs and (b) non-coding DNAs against their F_{GC}.

between T_m of the natural sequences and average T_m of the 'conservative shuffled' natural sequences vs. F_{GC} for coding and non-coding DNAs in Fig. 8, we saw two random distributions centered approximately 0 °C over the whole range of F_{GC}. The linear regressions for both coding and non-coding sequences possessed slopes of zero to the third and second decimal places and intercept of approximately 0.01 °C. These data demonstrate the closeness of T_m of the natural sequences to the average T_m of their 'conservative shuffled' sequences for both coding and non-coding DNAs.

3.3. Comparison of the nearest-neighbor frequencies in different classes of DNA

The results presented above demonstrate that the MSD-type relationships between T_m and F_{GC} in coding sequences and in non-coding sequences were substantially different and were only dependent on the

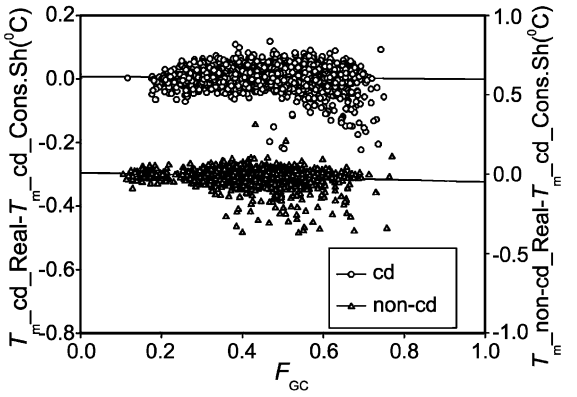

Fig. 8. Plots of the differences between T_ms of the natural coding and non-coding sequences and the average T_ms of the dinucleotide frequencies-conservatively shuffled sequences against their F_{GC}. Open circles and open triangles represent difference T_m values of the coding and non-coding sequences, respectively.

frequencies of the nearest neighbor pairs in DNA. There are 16 possible nearest neighbor combinations (16 dinucleotides), but only 10 nearest neighbors are unique. Therefore, we compared these 10 nearest-neighbor frequencies of the coding and the non-coding sequences in Fig. 9. Here, the nearest-neighbor frequencies of the sequences in five different bins across the range of F_{GC} were averaged with sequence-length weightings, and the weighted averages of the frequencies in the five bins of F_{GC} were compared as the difference between the coding and the non-coding frequencies. On the horizontal axis, the nearest neighbors were arranged in the order of decreasing stacking energy from left to right. We noticed that there were differences between coding and non-coding sequences in every nearest-neighbor frequency across most of the F_{GC} bins. For some nearest neighbor pairs (GG, AG, GA, AT), there was a striking sign reversal in the frequency difference from the lowest four F_{GC} bins going to the fifth and highest 0.6–0.7 F_{GC} bin. A number of nearest neighbors (GC, AC, CG, CA) showed a consistent trend to an increasingly positive frequency difference from the lowest to the highest F_{GC} bins. These cases, and the mostly positive frequency differences observed for GG, AG, GA, and AA would be expected to contribute significantly to the 6.6% higher slope calculated for natural coding compared to natural non-coding sequences from the data presented in Fig. 2, as well as account for the Fig. 7a,b, T_m differences for these DNAs.

3.4. The T_m vs. F_{GC} relationships of multicellular and unicellular organisms are different due to the differences in nearest-neighbor frequencies

As seen in Fig. 7a,b, we can get information about T_m vs. F_{GC} relationships of different classes of DNAs through plots of the differences in T_m between the natural sequences and the average 'random shuffled' natural sequences against their F_{GC}. Considering the varied origins of the natural DNAs, we divided them into two additional classes—the sequences from multicellular organisms and those from unicellular organisms. The differences in T_m between the natural sequences and the average 'random shuffled' natural sequences of coding and non-coding DNAs of these two classes were plotted as a function of their F_{GC} in Fig. 10a–d. As in Fig. 7a,b, the T_m difference between the natural and 'randomly shuffled' sequences vs. F_{GC} relationships have two opposite trends for coding and non-coding DNAs (Fig. 10a,b, [positive slopes] as compared to Fig. 10c,d, [negative slopes]). Furthermore, within each coding or non-coding DNA class, the T_m difference vs. F_{GC} relationship is clearly quantitatively different between the sequences from multicellular organisms and the sequences from unicellular organisms (compare Fig. 10a vs. Fig. 10b, and Fig. 10c vs. Fig. 10d). These results suggest that there is a difference between the T_m vs. F_{GC} MSD relationships of sequences from multicellular and those from

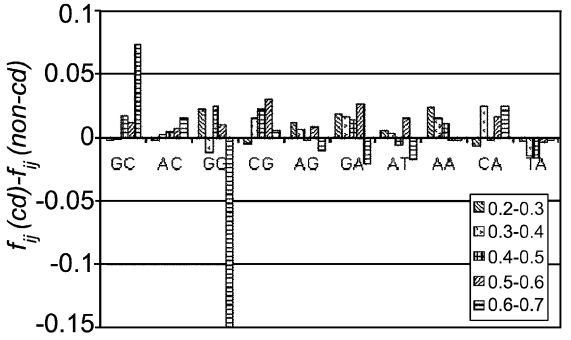

Fig. 9. The distribution of average nearest neighbor frequencies in five different bins of their F_{GC} for all DNA coding and non-coding sequences.

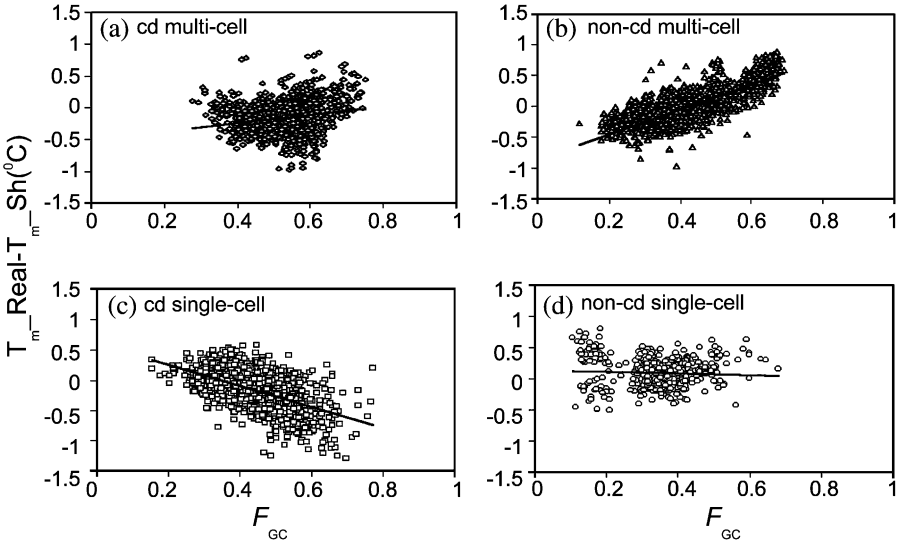

Fig. 10. Plots of the differences between T_ms of the natural sequences and the average T_ms of the randomly shuffled sequences in the cases of (a) coding DNAs from multi-cellular organisms, (b) non-coding DNAs from multi-cellular organisms, (c) coding DNAs from single-cellular organisms, and (d) non-coding DNAs from single-cellular organisms against their F_{GC}.

unicellular organisms in both coding and non-coding DNAs. However, this difference is smaller than the overall difference between coding and non-coding DNA classes. This observation can be explained when we consider the differences in the nearest-neighbor frequencies of the sequences from the two different kinds of organisms.

In a manner similar to Fig. 9, the differences in weighted averages of nearest-neighbor frequencies of the sequences in different bins across the F_{GC} range were plotted for the coding sequences from multicellular organisms (Fig. 11a) and for the non-coding sequences from unicellular organisms (Fig. 11b). The differences in frequencies of several nearest-neighbors (GC, CG, GG, AG, AA, TA) are significant across the range of F_{GC} in both of the figures. However, the most striking features in these two figures are the magnitudes of the frequency differences. Few frequency differences exceed ± 0.01, with the largest few being close to -0.03 (GC, CG). These differences are much smaller than the magnitude of the frequency differences between coding and non-coding sequences in Fig. 9, where many exceed +0.02 and the largest are +0.07 (GC) and -0.15 (GG). That explains why the differences in the T_m vs. F_{GC} relationships between the

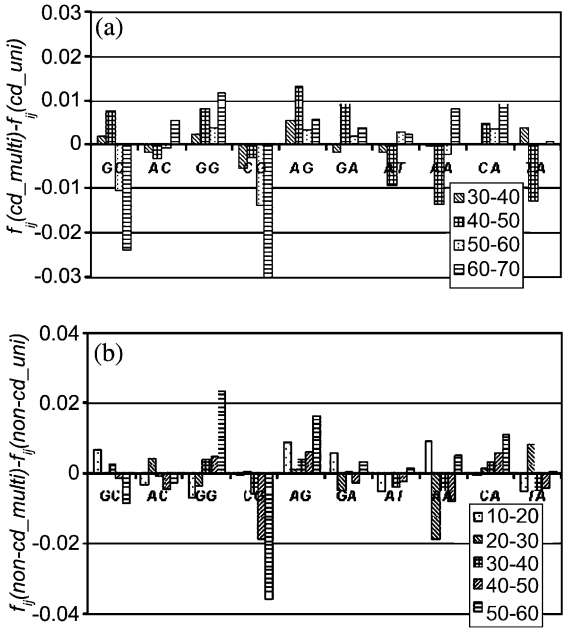

Fig. 11. The differences of average nearest neighbor frequencies between DNAs from (a) coding sequences and (b) non-coding sequences of multi-cellular and single-cellular organisms in different bins of their F_{GC}.

sequences from multicellular organisms and unicellular organisms are much smaller than the differences between coding and non-coding sequences.

4. Discussion

The relationship between the thermodynamic stability of DNA sequences, expressed by the T_m, and the sequence features, most notably the sequences' base content, has been investigated for decades. In various versions of the MSD equation, determined empirically from measured T_ms acquired over three decades, T_m has a linear relationship with the mole fraction of (G+C) in the DNA sequences [1]. The MSD equation has typically been shown to apply to large heterogeneous mixtures of DNA of long sequence lengths. However, for DNA sequences of relatively homogenous sequence composition, large deviations from the MSD equation have been observed and the T_m has been shown to depend more directly on the nearest neighbor frequencies of DNA sequences [22]. In the present investigation, we examined a large collection of natural DNAs that are moderate in length (from 1 to 100 Kbp). Only sequence documents containing single genes were collected from 28 organisms of varying F_{GC}, so that every organism across the range of biological GC content ($0.15<F_{GC}<0.7$) was sampled consistently. Furthermore, using single-gene documents, we obtained a collection of DNAs that were relatively homogeneous in composition and of consistently small size. With this collection of natural DNAs as input, we employed the MELTSIM program, which models the melting process in terms of the nearest neighbor interactions, to simulate the melting and calculate the T_ms.

Comparing the slopes and intercepts of Eqs. (1a) and (1b) and taking into account that the accepted accuracy for the MSD equation is ± 0.15 °C [1], the results of the calculated T_ms of all the natural DNAs gave a T_m vs. F_{GC} relationship, which is similar with that given by the MSD equation. When we separated the natural DNAs into two different functional domains, coding and non-coding, the T_m vs. F_{GC} relationships for these two sequence classes were linear but exhibited significantly different slopes from one another as well as from the original source sequences. Comparing Eq. (2a) for coding sequences with Eq. (1a), we see that the difference between the T_{m_cd} and the T_m of the intact sequences calculated by MELTSIM is higher than the experimental accuracy when F_{GC} is below 0.69. From Eq. (2b) for non-coding sequences and Eq. (1a), the difference between the $T_{m\_non\text{-}cd}$ and the T_m of the intact sequences is also higher than the experimental accuracy when F_{GC} is higher than 0.30. In other words, in most of the natural range of GC content in DNAs, the linear relationships between T_m and F_{GC} in coding DNAs and non-coding DNAs are different. These results could not be accurately described by the MSD equation previously reported. However, the fact that Eqs. (2a) and (2b) both exhibit strong linear relationships, like Eq. (1b), demonstrates that relationships between T_m and F_{GC} in coding DNAs and non-coding DNAs can be described by MSD-like equations for each DNA class, just with different coefficients – in particular a 6.6% difference in slope. Thus, for coding and non-coding sequences of identical or similar F_{GC} values, the very different T_ms that are evident in Fig. 2, are subsumed into a single T_m in Fig. 1 for those coding and non-coding sequence segments that form one single gene sequence in Fig. 1.

The very different coding and non-coding region T_ms that occur within single genes may be manifest as individual cooperative melting domains that retain their functional identity during melting. There is a feature of the MELTSIM program that accurately simulates the thermodynamic domain melting along any DNA sequence [1,7,24–26,31]. Application of this simulation feature would undoubtedly reveal in the original source sequences some of the same behavior exhibited here by the coding and non-coding sequence classes. Such behavior has been previously observed by our laboratory in genes found in *D. discoideum*, comprised of multiple alternating exons and introns [25]. In these cases, MELTSIM demonstrated that the alternating exon and intron sequences melted as distinct thermodynamic cooperative domains, reflecting their nearest neighbor base compositions.

Our experiments with shuffled sequences demonstrate that the biases in nearest neighbor frequencies alone can account for the observed differences in the individual MSD-like relationships between T_m and F_{GC} in coding DNAs and non-coding DNAs—

not their F_{GC} or higher order pair frequencies. Whenever the natural nearest neighbor frequencies in those two classes of DNAs are changed but F_{GC} is maintained by either the 'random' or 'in-frame' shuffling procedures, the relationships between T_m and F_{GC} in the new sequences from the two kinds of DNAs become nearly identical to each other and to the relationship derived for whole gene containing sequences and to the MSD equation reported in the literature [1]. These relationships are clearly seen in Fig. 6. However, when these natural nearest neighbor frequencies are preserved (in the 'conservative' shuffling procedure), the relationships between T_m and F_{GC} in the shuffled sequences from the two kinds of DNAs remain different in slope by 6.3%, which is 95% of the 6.6% slope difference exhibited by the original coding and non-coding sequences (see Eqs. (5a) and (5b), and Fig. 6). Thus, nearest neighbor effects in the large population of relatively homogeneous sequences we studied here can adequately account for the difference observed between the coding and non-coding sequence T_m vs. F_{GC} MSD-like relationships.

However, to understand the underlying physical basis for the 6.6% difference in slopes of these relationships, we investigated whether differences existed in the nearest neighbor frequency distributions between coding and non-coding sequences that are consistent across the entire F_{GC} range. Therefore, we examined the nearest neighbor frequencies of our coding and non-coding sequence populations and presented evidence for the differences in the natural nearest neighbor frequencies of coding and non-coding DNAs across the whole range of their GC content in Fig. 9. These differences are most prominent in the CG, GG, and GC neighbor pairs and less so for AA, GA, and CA neighbor pairs. These differences in nearest neighbor pair frequencies are in agreement both with the results of several published investigations, which surveyed fewer sequences [32,33], as well as the known fact that the triplet code constrains the nearest neighbor pair frequencies of coding DNAs to be different from the neighbor pair frequencies of non-coding DNAs, where no equivalent code exists. Interestingly, the positive differences in CG and GC neighbor pairs reflect the known under-representation of the CG neighbor pair in all eukaryotes and its preferential depletion in non-coding as opposed to coding sequences [34–37]. This restricted occurrence is due to endogenous enzymatic methylation at the 5′ position of cytosine in the CG neighbor pair, which can lead to a deamination reaction generating TG and CA mutations [38].

It can be argued that our simulated T_m results based on the MELTSIM program, which uses only nearest neighbor pairs in the melting simulation process, are obvious based on the known differences in the natural nearest neighbor frequencies of coding and non-coding DNAs. However, arguments have been made for the inclusion of higher order neighbor pair interactions in modeling the melting process [1]. The results here comparing the natural and shuffled sequences support the idea that nearest neighbor frequencies alone adequately determine thermal stability and suggest that for the vast majority of DNA sequences higher order neighbor pair interactions play insignificant roles.

We also demonstrated that there are small differences in the thermal stability of DNAs from multicellular and unicellular eukaryotes that can be accounted for based on small differences in their nearest neighbor frequencies. These differences in nearest neighbor frequencies are consistent with published results, which were derived from far fewer sequences [39]. The large negative differences in CG and GC are also consistent with the overall lower occurrence of methylated cytosine in multicellular organisms as opposed to some unicellular organisms [39].

Acknowledgements

The authors acknowledge Jeff Bizzaro and bioinformatics.org for maintaining and use of the MELTSIM and COMPILE programs used in this study.

References

[1] R.D. Blake, Denaturation of DNA, in: R.A. Meyers (Ed.), Encyclopedia of Molecular Biology and Molecular Medicine 2, VCH Verlagsgesellschaft mbH, Weinheim, 1996, pp. 1–19.

[2] J.E. Hearst, T.C. Cech, K.A. Marx, A. Rosenfeld, J.R. Allen, Characterization of the rapidly renaturing sequences in the main CsCl density bands of Drosophila, mouse and the human DNA, Cold Spring Harb. Symp. Quant. Biol. 38 (1973) 329–339.
[3] K.A. Marx, J.R. Allen, J.E. Hearst, Characterization of the repetitious human DNA families and evidence of their satellite DNA equivalents, Biochim. Biophys. Acta 425 (1976) 139–147.
[4] K.A. Marx, J.R. Allen, J.E. Hearst, Chromosomal localization by in-situ hybridization of the repetitious human DNA families and evidence of their satellite DNA equivalents, Chromosoma 59 (1976) 23–42.
[5] K.A. Marx, I.F. Purdom, K.W. Jones, Primate repetitive DNAs: evidence for new satellite DNAs and similarities in non-satellite repetitive DNA sequence properties, Chromosoma 73 (1979) 153–161.
[6] K.A. Marx, Non-satellite repetitive human DNA families: sequence properties and evidence for occurrence in chimpanzee DNA, Biochim. Biophys. Acta 608 (1980) 232–242.
[7] R.D. Blake, J.W. Bizzaro, J.D. Blake, G.R. Day, S.G. Delcourt, J. Knowles, et al., Statistical mechanical simulation of polymeric DNA melting with MELTSIM, Bioinformatics 15 (1999) 370–375.
[8] D. Riesner, K. Henco, G. Steger, in: A. Chrambach, M.J. Dunn, B.J. Radola (Eds.), Advances in Electrophoresis 4, VCH Verlagsgesellschaft, Weinheim, 1991, pp. 169–250.
[9] L.S. Lerman, S.G. Fischer, I. Hurley, K. Silverstein, N. Lumelsky, Sequence-determined DNA separations, Annu. Rev. Biophys. Bioeng. 13 (1984) 399–423.
[10] G. Steger, Thermal denaturation of double-stranded nucleic acids: prediction of temperatures critical for gradient gel electrophoresis and polymerase chain reaction, Nucleic Acids Res. 22 (1994) 2760–2768.
[11] J. SantaLucia Jr, H.T. Allawi, A. Seneviratne, Improved nearest-neighbor parameters for predicting DNA duplex stability, Biochemistry 35 (1996) 3555–3562.
[12] C.R. Cantor, P.R. Schimmel, Biophysical Chemistry Part III: The Behavior of Biological Macromolecules, W.H. Freeman and Company, San Francisco, 1980, pp. 1109–1181.
[13] D. Poland, H.A. Scheraga, Theory of Helix–Coil Transitions in Biopolymers, Academic Press, New York, 1970.
[14] M.D. Frank-Kamenetskii, Yu.S. Lazurkin, Conformational changes in DNA molecules, Annu. Rev. Biophys. Bioeng. 3 (1974) 127–150.
[15] D. Poland, Recursion relation generation of probability profiles for specific-sequence macromolecules with long-range correlations, Biopolymers 13 (1974) 1859–1871.
[16] Yu.L. Lyubchenko, M.D. Frank-Kamenetskii, A.V. Vologodskii, Yu.S. Lazurkin, G.G. Gause Jr, Fine structure of DNA melting curves, Biopolymers 15 (1976) 1019–1036.
[17] M. Fixman, J.J. Freire, Theory of DNA melting curves, Biopolymers 16 (1977) 2693–2704.
[18] M.J. Doktycz, R.F. Goldstein, T.M. Paner, F.J. Gallo, A.S. Benight, Studies of DNA dumbells. I. Melting curves of 17 DNA dumbbells with different duplex stem sequences linked by T4 endloops: evaluation of the nearest-neighbor stacking interactions in DNA, Biopolymers 32 (1992) 849–864.
[19] J. SantaLucia Jr, A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. USA 95 (1998) 1460–1465.
[20] J. Marmur, P. Doty, Heterogeneity in deoxyribonucleic acids, Nature 183 (1959) 1427–1431.
[21] C.L. Schildkraut, J. Marmur, P. Doty, Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl, J. Mol. Biol. 4 (1962) 430–443.
[22] S.G. Delcourt, R.D. Blake, Stacking energies in DNA, J. Biol. Chem. 266 (1991) 15 160–15 169.
[23] R.D. Blake, S.G. Delcourt, Thermal stability of DNA, Nucleic Acids Res. 26 (1998) 3323–3332.
[24] K.A. Marx, J.W. Bizzaro, I. Assil, R.D. Blake, Proc. Mater. Res. Soc.: Stat. Mech. Phys. Biol. 463 (1997) 147–152.
[25] K.A. Marx, I.Q. Assil, J.W. Bizzaro, R.D. Blake, Comparison of experimental to MELTSIM calculated DNA melting of the (A+T) rich *Dictyostelium discoideum* genome: denaturation maps distinguish exons from introns, J. Biomol. Struct. Dyn. 16 (1998) 329–339.
[26] K.A. Marx, J.W. Bizzaro, R.D. Blake, M. Hsien-Tsai, L. Feng-Tao, Experimental DNA melting behavior of the three major *Schistosoma* species, Mol. Biochem. Parasitol. 15 (2000) 303–307.
[27] G. Grillo, M. Attimonelli, S. Liuni, G. Pesole, CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases, Comput. Appl. Biosci. 12 (1996) 1–8.
[28] D. Kandel, Y. Matias, R. Unger, P. Winkler, Shuffling biological sequences, Discrete Appl. Math. 71 (1996) 171–185.
[29] E. Coward, Shufflet: shuffling sequences while conserving the k-let counts, Bioinformatics 15 (1999) 1058–1059.
[30] B.F.J. Manly, Randomization, Bootstrap and Monte Carlo methods in Biology, 2nd Ed, Chapman and Hall, New York, 1998, p. 34.
[31] J.W. Bizzaro, K.A. Marx, R.D. Blake, Comparison of experimental with theoretical melting of the yeast genome and individual yeast chromosome denaturation mapping using the program MELTSIM, Proc. Mater. Res. Soc. 489 (1998) 73–78.
[32] J.W. Fickett, C.-S. Tung, Assessment of protein coding measures, Nucleic Acids Res. 20 (1992) 6441–6450.
[33] M.Q. Zhang, Statistical features of human exons and their flanking regions, Hum. Mol. Genet. 7 (1998) 919–932.
[34] J. Josse, A.A. Kaiser, A. Kornberg, Enzymatic synthesis of deoxyribonucleic acid VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid, J. Biol. Chem. 236 (1961) 864–875.
[35] M.N. Swartz, T.A. Trautner, A. Kornberg, Enzymatic synthesis of deoxyribonucleic acid XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids, J. Biol. Chem. 237 (1962) 1961–1967.
[36] K.A. Marx, S.T. Hess, R.D. Blake, Characteristics of the large (dA)·(dT) homopolymer tracts in *Distyostelium discoideum* gene flanking and intron sequences, J. Biomol. Struct. Dyn. 11 (1993) 57–66.
[37] H. Musto, H.R. Maseda, F. Alvarez, J. Tort, Possible implica-

tions of CpG avoidance in the flatworm *Schistosoma mansoni*, J. Mol. Evol. 38 (1994) 36–40.

[38] W. Salser, Globin mRNA sequences: analysis of base pairing and evolutionary implications, Cold Spring Harb. Symp. Quant. Biol. 40 (1977) 985–1002.

[39] P. Setlow, Nearest neighbor frequencies in deoxyribonucleic acids, in: G.D. Fasman (Ed.), , 3rd Ed, Handbook of Biochemistry and Molecular Biology Nucleic Acids-II,CRC Press, Cleveland, 1975, pp. 315–318.