OPEN ACCESS Freely available online

PLoS one

A Phenomenological Model for Predicting Melting Temperatures of DNA Sequences

Garima Khandelwal[1,2], Jayaram Bhyravabhotla[1,2,3]*

1 Department of Chemistry, Indian Institute of Technology Delhi, New Delhi, India, **2** Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology Delhi, New Delhi, India, **3** School of Biological Sciences, Indian Institute of Technology Delhi, New Delhi, India

Abstract

We report here a novel method for predicting melting temperatures of DNA sequences based on a molecular-level hypothesis on the phenomena underlying the thermal denaturation of DNA. The model presented here attempts to quantify the energetic components stabilizing the structure of DNA such as base pairing, stacking, and ionic environment which are partially disrupted during the process of thermal denaturation. The model gives a Pearson product-moment correlation coefficient (r) of ~0.98 between experimental and predicted melting temperatures for over 300 sequences of varying lengths ranging from 15-mers to genomic level and at different salt concentrations. The approach is implemented as a web tool (www.scfbio-iitd.res.in/chemgenome/Tm_predictor.jsp) for the prediction of melting temperatures of DNA sequences.

Citation: Khandelwal G, Bhyravabhotla J (2010) A Phenomenological Model for Predicting Melting Temperatures of DNA Sequences. PLoS ONE 5(8): e12433. doi:10.1371/journal.pone.0012433

Editor: Sudhindra Gadagkar, Midwestern University, United States of America

Received December 9, 2009; **Accepted** August 2, 2010; **Published** August 26, 2010

Funding: The funding for the work has been provided by the Department of Biotechnology, India. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bjayaram@chemistry.iitd.ac.in

Introduction

Several physico-chemical factors such as base stacking, hydrogen bonding, hydrophobic, electrostatic and van der Waals interactions etc. stabilize the DNA molecule [1]. Base stacking and hydrogen bonding are considered to be the dominant of all these forces [2–4]. These diverse forces stabilizing DNA act in concert to protect the genetic code against external perturbations. But if these forces render the DNA to be static, the coding bases will not be directly accessible to the expression of genetic code. DNA, however, is a dynamic entity and the forces do get disrupted and the coding bases exposed to enzymes [5] as in replication of DNA, transcription into m-RNA etc.. How DNA opens up in response to intrinsic sequence effects and extrinsic local environment is thus a matter of considerable interest in deciphering molecular details of gene expression in particular and genome organization in general. We have been interested in understanding the sequence effects on the structure and energetics of DNA [6–8]. Here we focus on the stability of DNA of varying lengths and base composition and constitution from a melting perspective.

DNA denaturation (melting) is the process of separation of ds-DNA into two single strands. This cooperative unwinding is also known as helix-coil or melting transition [9]. DNA melting occurs over a small range of temperature and results in changes in its physical properties [10]. It has been known since the 1950s, that heating a DNA solution above room temperature results in the separation of strands. The temperature at which half of the DNA molecule is denatured, i.e. one half is in double helical form and the other half in a random coil state, is termed as the melting temperature of the DNA, Tm [9]. The melting temperature depends on a variety of factors, such as the length of DNA [11,12] (shorter pieces tend to melt more easily, [13]), the nucleotide sequence composition [14–16], salt concentration (ionic strength of the added salt) [14–15,17] and generally lies between 50°C and 100°C. DNA can be denatured not only by heating, but by other methods as well, eg. use of organic solvents such as formamide [18] and dimethyl sulfoxide, ligands [19], increasing the pH of the solution, lowering the salt concentration [20] etc.

DNA 'breathes' even at normal cell temperatures [21,22] and local regions of a few tens of base pairs become temporarily unwound and form a bubble, in which stacking and hydrogen bonding are partially disrupted [23–25]. It is easier for the proteins (RNA polymerase, and origin binding proteins) to create locally unwound regions on DNA in A/T rich regions, which could be one of the reasons for DNA replication origins and transcription initiation bubbles to have such regions [26]. In G/C rich regions, the strands do not unwind until higher temperatures are reached. When all of the base interactions are broken, the two strands separate. This is called denaturation. Local unwinding however, is not denaturation but an essential prerequisite.

DNA melting is measured by the absorbance of UV light (260 nm) by the DNA solution, where the amount of UV light absorbed is proportional to the fraction of non-bonded base pairs. This UV absorbance is due to the π-π^* electronic transition in both purine and pyrimidine bases, which reflects a change in the electronic configuration of the bases due to the decrease in double helical stacking and base paring upon melting. As the temperature increases, melting of the double-stranded DNA is initiated and the absorbance of UV-light increases through a series of sharp jumps. The absorbance increases by 30–40% depending on the DNA sample. [9]. The middle-point of the temperature range over which the strands of DNA separate gives the melting temperature [10].

Earlier theories on DNA melting have incorporated stacking and hydrogen bonding within the framework of models for transitions in polypeptides: (i) Zimm-Bragg theory; where stacking is modeled as a nearest-neighbor interaction; [27] (ii) Lifson-Roig theory; where conformational restriction due to hydrogen bonding is taken into account [28]. The role of stacking against the background of hydrogen bonding has been investigated within the context of Generalized Model of Polypeptide Chain (GMPC) [29]. Other descriptions of melting have also been advanced [30–32]. Theories addressing the helix-coil transitions are not widely used for the prediction of melting temperatures [32]. One of the reasons for this could be the difficulty in calculations, which are computation-intensive and require adjustment of many parameters [33].

Many attempts have been made to predict the melting temperatures of short nucleotide sequences, which is of particular interest in primer design. The earliest of these methods used a simple formula to calculate Tm based on the GC content of the sequence [17]. Subsequently, this formula was modified to include the effect of salt concentration of the solution [20]. The next set of methods utilized the nearest neighbor (NN) model to calculate Tm, which requires a set of thermodynamic parameters. Many groups have provided these parameters [14,34,35] and, it was noted that there was a consensus among these methods [35]. While the ranges of energy determined in different studies are similar, the values for individual NN pairs show discrepancies [36]. Also, the coefficients obtained by these methods from fitting the data are non-unique and defy simple interpretation [4]. Taking the research efforts a step further towards a reliable predictive model, we report in this work, a phenomenological model to predict the melting temperature of DNA, accounting for the physico-chemical events taking place in the melting process. In particular, the model introduced here accounts quantitatively and explicitly for disruption in stacking interactions, breakage of hydrogen bonding, salt effects and the nucleotide strand concentration in the melting of DNA.

Materials and Methods

Dataset

The accuracy benchmark dataset compiled by Panjkovich & Melo [37] is adopted here for the study. The dataset is made up of 348 data points comprising 108 unique oligonucleotide sequences at various salt concentrations. This dataset is divided into two parts: (i) A training set consisting of 123 oligomers for obtaining the best fit equation giving the minimum possible error and (ii) a test dataset consisting of 225 oligomers, to assess the quality of prediction on independent data. Both the datasets represents the complete data space (Figures S1, S2 and S3). We have also examined the performance of the method on an additional dataset of 100 short nucleotide sequences (15mers) [38]. Subsequently, we investigated the validity of the model on 20 genomic sequences.

Methodology

Melting of DNA necessitates the disruption of stacking interactions between the two base pairs within each dinucleotide step. During the process, cross strand stacking interactions are completely lost while intra-strand stacking interactions are disrupted partially. The dinucleotide steps are assembled into four groups on the basis of their possible interactions as RR, RY, YR and YY, where R and Y denote a purine and a pyrimidine respectively. RY has the highest stacking as known from experiments [39] and simulations [40]. Various combinations of values were tried out to give the least possible error for the training dataset. Finally, the four dinucleotide groups (RY, RR, YY, YR) were assigned values as 5, 3, 3, 2, keeping in mind that the values should be relative to the values for H-bonding as well as to each other.

The melting of DNA also requires the breakage of Watson-Crick hydrogen bonds (H-bonds) and it is well known that GC pairs (3 H-bonds) are stronger than AT pairs (2 H-bonds). Based on this, and the knowledge of interaction energies of H-bonded pairs [7,41], values of 4 and 1 are assigned to GC and AT base pairs respectively. On the basis of hydrogen bonding between the bases, the double helical dinucleotide steps can be divided into three groups: (a) Group with 6 H-bonds, (b) Group with 5 H-bonds and (c) Group with 4 H-bonds; the corresponding H-bond energy values being 8, 5 and 2 respectively.

The contribution of H-bond energy and stacking energy is almost equivalent in the stabilization of duplex DNA, as discerned from various studies on modified bases [42], and dangling bases [39] and is of the order of 1–2 kcal. Also, it has been observed that the rise in melting temperature due to the addition of a single H-bond is about 2–6°C [43], while it is approximately 2°C due to increase in stacking energy per added base pair [44]. The H-bonding and stacking energy values are assigned considering all these observations. The DNA strength parameter for each double helical dinucleotide step can be then developed as a sum of stacking and hydrogen bonding values proposed above. For example, in case of GC, which belongs to RY group, the value of stacking is 5 while two triple H-bonds add up to a value of 8. So, the DNA strength parameter for a GC step is given as: 5+8 = 13.

A total of 16 dinucleotide combinations are possible of which only 10 are unique when read in the $5' \rightarrow 3'$ direction. These are arranged here in the decreasing order of DNA strength parameter value (Table1): (i) GC, (ii) CC = GG, (iii) CG, (iv) AC = GT, (v) TC = GA, (vi) CT = AG, (vii) TG = CA, (viii) AT, (ix) TT = AA, (x) TA. The above assignment of DNA strength parameter values is also found to be consistent with the observations on relative stabilities of dinucleotides [25], the molecular interpretation of the conjugate rule [45] and some recent molecular dynamics simulations [40]. These values are found to be in overall agreement with the calculated free energies [14,15,34,46] and melting free energy parameters [36] with a few exceptions.

The value of DNA strength parameter for the whole sequence is accumulated by adding the values (Table 1) for each dinucleotide step which is referred to here as the cumulative DNA strength parameter. This would go on increasing with the length, so to delineate the effect of length, the DNA strength parameter (E) is derived on a per unit (base pair) basis as given below:

$$\textbf{DNA strength pararmeter per base } (\mathbf{E}) = \frac{\text{Cumulative DNA strength parameter}}{\text{Length of the DNA sequence}}$$

Table 1. Values of DNA strength parameter for each dinucleotide step.

Stack	5	3	3	2
H-bond	RY	YY	RR	YR
4+4	GC = 13	CC = 11	GG = 11	CG = 10
1+4	AC = 10	TC = 8	AG = 8	TG = 7
4+1	GT = 10	CT = 8	GA = 8	CA = 7
1+1	AT = 7	TT = 5	AA = 5	TA = 4

doi:10.1371/journal.pone.0012433.t001

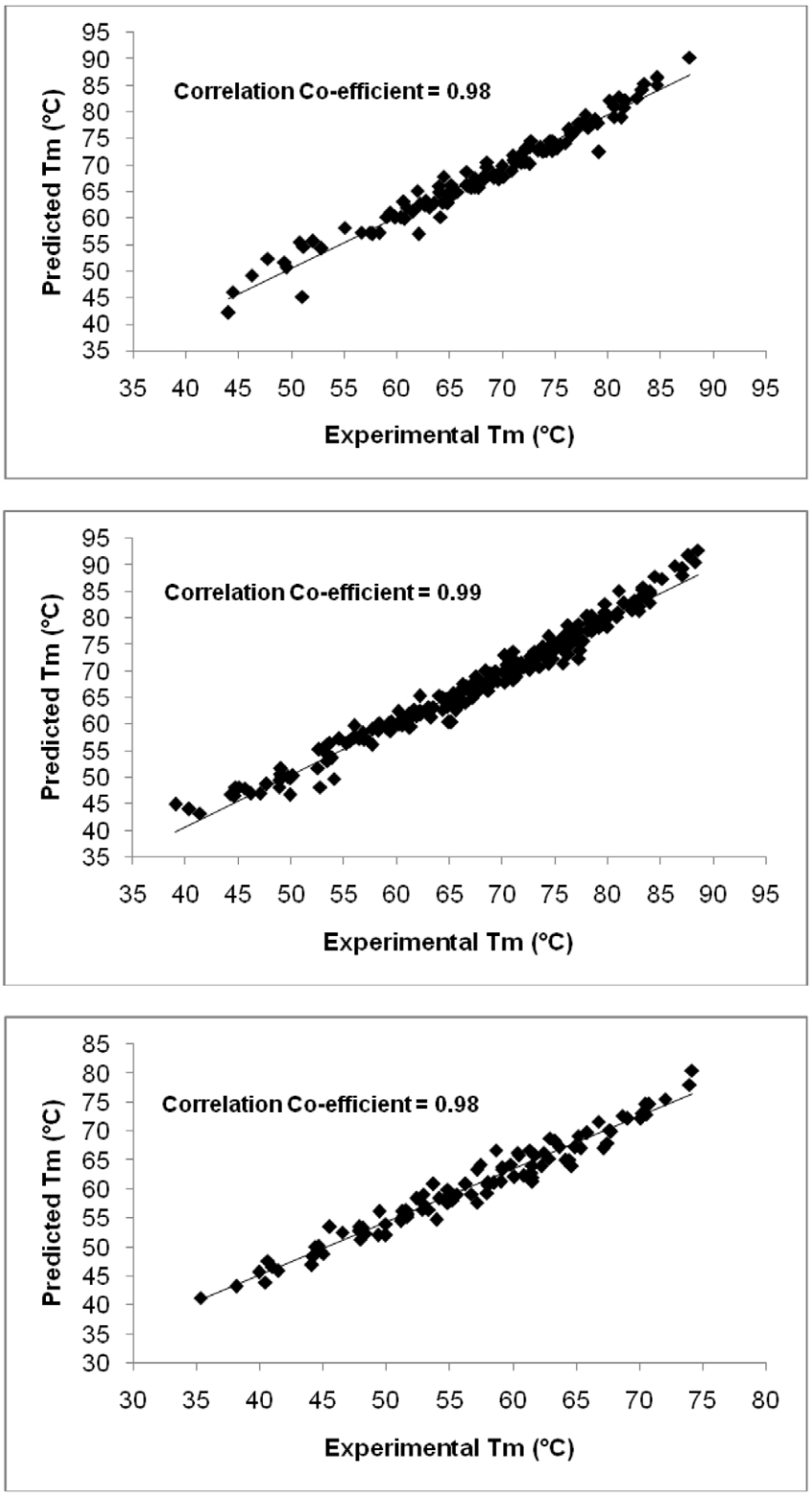

Figure 1. Correlation plots between the experimental and predicted melting temperatures. Figure 1(a). Correlation between predicted and experimental melting temperatures for the training dataset of 123 oligomers **Figure 1(b).** Correlation between predicted and experimental melting temperatures for the test dataset of 225 oligomers **Figure 1(c).** Correlation between predicted and experimental melting temperatures for an additional dataset of 100 oligomers (15-mers) adapted from Ref. 38.
doi:10.1371/journal.pone.0012433.g001

The salt effects are taken into account on the basis of $[Na^+]$ concentration in the solution, implemented as a natural logarithmic variable, which is in accordance with previous work [38,47]. Similarly borrowing from the electrostatic behavior of DNA from the literature [6], the length of the sequence is also accounted for via a natural logarithmic function. Length considerations via a variable such as $(n-1)/n$, (n = length of oligonucleotide sequence) were reported earlier to account for the decrease in Tm with decreasing length of the oligomer [47]. The concentration units for oligonucleotides and genomic sequences are typically reported as molar and μg/ml respectively in experimental studies. The nucleotide strand concentration parameter is implemented using a natural logarithmic function.

All the above contributors are pooled into a simple equation and processed through the multiple regression analysis method of Analyse-It software package [48], to derive the best fitting equation predicting the Tm values for the training dataset. Residual values and the standard error of estimate were also calculated. The good-ness of fit is critically evaluated by various statistical techniques such as the normal probability plots of residual, residual distribution plots (Figures S4 and S5 respectively). The final equation derived after the multiple regression is:

$$\mathrm{Tm}(^{\circ}\mathrm{C}) = (7.35 \times \mathrm{E}) + [17.34 \times \ln(\mathrm{Len})] + [4.96 \times \ln(\mathrm{Conc})] + [0.89 \times \ln(\mathrm{DNA})] - 25.42 \quad (1)$$

Tm = Predicted melting temperature
E = DNA strength parameter per base
Len = Length of nucleotide sequence (number of base pairs)
Conc = $[Na^+]$ concentration of the solution (Molar)
DNA = Total nucleotide strand concentration.

The r^2 obtained from this equation on the training dataset is 0.96. The equation to predict the melting temperature, without the use of nucleotide strand concentration (DNA) as one of the parameters is provided in the supporting information (Supporting Text S1).

The use of eq. (1) is illustrated below. Consider for example a 15 bp long sequence GACGACAAGACCGCG, taken at 0.22 M salt concentration and 0.000002 nucleotide strand [38]. The melting temperature for this sequence is calculated as follows.

Step 1: Read the sequence from 5′ end to 3′ end and add up the DNA strength parameter given in Table 1 for each dinucleotide step, moving one base at a time as: GA = 8, AC = 10, CG = 10, GA = 8, AC = 10, CA = 7 and so on. (For the given sequence of 15 base pairs, 14 dinucleotide steps are obtained). So, The DNA strength parameter for the given sequence is: 8+10+10+8+10+7+5+8+8+10+11+10+13+10 = 128. The DNA strength parameter per base (E) is then calculated as: 128/15 = 8.53

Step 2: Substituting all the values in eq. (1),

$$\mathrm{Tm}(^{\circ}\mathrm{C}) = (7.35 \times 8.53) + [17.34 \times \ln(15)] + [4.96 \times \ln(0.22)] + [0.89 \times \ln(0.000002)] - 25.42$$

Predicted Tm = 65.04°C
Reported Experimental Tm = 64.4°C [38]

For genomic sequences, the Tm is first calculated by computing the cumulative strength parameter of a melting unit of 70 bp from the start which is then derived per base and employed in eq. (1). This window is translated by one base pair and a new Tm is calculated and the procedure is repeated till the end of the sequence. The Tm for the whole genomic sequence is then developed as the average of overlapping melting units of length 70 bp, a number arrived at empirically which appears to have biological significance as discussed below.

Results and Discussion

In this study, a phenomenological model is developed on the basis of a theoretical appraisal of the events occurring during the process of DNA thermal denaturation. The model was trained on a dataset of 123 oligomers to achieve a best fit equation (1); (Figure 1), which gave a correlation coefficient (r) of 0.98 and an average error of 1.36°C (data provided in Table S1). This equation (1) was used to predict the melting temperatures for a test dataset of 225 oligonucleotide sequences whose experimental melting temperatures were known; (Figure 1), where a correlation coefficient (r) of 0.99 and an average error of 1.31°C was obtained (data provided in Table S2). Subsequently the model was validated on 100 15-mers compiled by Owczarzy [38]. The results are depicted in Fig. 1(c), which indicate that even for shorter sequences not occurring in the training set, the correlation between the predicted and the experimental Tm on a large dataset of 100 sequences is quite high (correlation coefficient, r = 0.98, data provided in Table S3). A further verification of the viability of the current method was undertaken by considering three oligonucleotide sequences of 40 base pair length, taken at two different salt concentrations [38]. The average error of prediction for these sequences is 1.48°C (data provided in Table S4). The significance of the model was checked by means of Anova (Table S5).

The correlation coefficients with experimental melting temperatures for the four parameters used in the model, as a single entity and in all possible combinations are shown in Table 2. As clear from Table 2, the strength parameter appears to be the main

Table 2. Correlation coefficients for all possible combinations of the four parameters used in eq. (1).

Parameter	Correlation Coefficient (r)
E	0.77
Len	0.49
Conc	0.44
DNA	−0.21
E + Len	0.83
E + Conc	0.93
E + DNA	0.71
Len + Conc	0.65
Len + DNA	0.49
Conc + DNA	0.50
E + Len + Conc	0.98
E + Len + DNA	0.84
E + Conc + DNA	0.93
Len + Conc + DNA	0.66
E + Len + Conc + DNA	0.98

E = DNA strength parameter per base; Len = Length of nucleotide sequence (number of base pairs); Conc = [Na+] concentration of the solution (Molar); DNA = Total nucleotide strand concentration (Molar).
doi:10.1371/journal.pone.0012433.t002

Table 3. Experimental and predicted melting temperatures of a few genomic DNA sequences.

S. No.	Genome	NCBI ID	Length (bp)	Na^+ Conc. (M)	DNA Conc. (g/ml)	Exp. Tm (°C)	Pred. Tm (°C)	Exp. – Pred. Tm(°C)
1.	*Cytophaga hutchinsonii*	NC_008255	4433218	0.016	0.00002	70.2[49]	73	−2.8
2.	*Lactobacillus acidophilus*	NC_006814	1993560	0.016	0.00002	67.9[49]	71.1	−3.2
3.	*Lactobacillus bulgaricus*	NC_008054	1864998	0.016	0.00002	74.9[49]	77.6	−2.7
4.	*Lactobacillus fermenti*	NC_010610	2098685	0.016	0.00002	75.6[49]	78.4	−2.8
5.	*Leptospira interrogans*	NC_004343	358943	0.016	0.00002	68.4[49]	71.1	−2.7
6.	*Leptospira borgpetersenii*	NC_008508	3614446	0.016	0.00002	72.4[49]	73.3	−0.9
7.	*Mycoplasma arthritidis*	NC_011025	820453	0.016	0.00002	65.9[49]	69.3	−3.4
8.	*Micrococcus luteus*	NC_012803	2501097	0.016	0.00002	84.9[49]	87.9	−3
9.	*Nitrobacter winogradskyi*	NC_007406	3402093	0.016	0.00002	81.0[49]	83.2	−2.2
10.	*Pseudoalteromonas atlantica*	NC_008228	5187005	0.016	0.00002	71.2[49]	75.6	−4.4
11.	*Pseudomonas pseudomallei*	NC_006350	4074542	0.016	0.00002	84.3[49]	85.8	−1.5
12.	*Stenotrophomonas maltophilia*	NC_010943	4851126	0.016	0.00002	83.1[49]	85.2	−2.1
13.	*Pseudomonas fluorescens*	NC_004129	7074893	0.016	0.00002	80.1[49]	83.7	−3.6
14.	*Shewanella putrefaciens*	NC_009438	4659220	0.016	0.00002	73.2[50]	75.5	−2.3
15.	*Bacillus subtilis*	NC_000964	4214630	0.0732	0.00005	82.1 [12]	83.3	−1.2
16.	*Clostridium perfringens*	NC_003366	3031430	0.0732	0.00005	75.1[12]	76.7	−1.6
17.	*Micrococcus luteus*	NC_012803	2501097	0.0732	0.00005	94.5[12]	96.3	−1.8
18.	*Pseudomonas fluorescens*	NC_004129	7074893	0.0732	0.00005	89.8[12]	92.1	−2.3
19.	*Bacillus subtilis*	NC_000964	4214630	0.15	0.00002	87[13]	86	1
20.	*Deinococcus radiodurans*	NC_001263	2648638	0.15	0.00002	97[13]	96.4	0.6
21.	*Mycobacterium leprae*	NC_002677	3268203	0.15	0.00002	93[13]	92.5	0.5
22.	*Saccharomyces cerevisiae*	NC_001133 to NC_001148	12057500	0.15	0.00002	82.5[13]	83.8 Ω	−1.3
23.	*Ureaplasma urealyticum*	NC_011374	874478	0.15	0.00002	78[13]	78.4	−0.4

Ω Average melting temperature for the 16 chromosomes.
Exp. Tm = Experimental melting temperature.
Pred. Tm = Predicted melting temperature.
doi:10.1371/journal.pone.0012433.t003

Table 4. Experimental and predicted melting temperatures of *Escherichia coli* DNA at various salt concentrations.

S. No.	Genome	Na^+ Conc. (M)	DNA Conc. (g/ml)	Experimental Tm (°C)	Predicted Tm (°C) #	Experimental – Predicted Tm (°C)
1.	*Escherichia coli*	0.015	0.000018	70.7[20]	77.9	−7.2
2.	*Escherichia coli*	0.016	0.00002	75.7[49]	78.3	−2.6
3.	*Escherichia coli*	0.0732	0.00005	85.7[12]	86.6	−0.9
4.	*Escherichia coli*	0.075	0.000018	83.3[20]	85.9	−2.6
5.	*Escherichia coli*	0.01	0.000018	68.7[20]	75.8	−7.1
6.	*Escherichia coli*	0.02	0.000018	73.4[20]	79.3	−5.9
7.	*Escherichia coli*	0.035	0.000018	77.1[20]	82.1	−5
8.	*Escherichia coli*	0.05	0.000018	80.0[20]	83.8	−3.8
9.	*Escherichia coli*	0.1	0.000018	86.5[20]	87.3	−0.8
10.	*Escherichia coli*	0.12	0.000018	86.0[20]	88.2	−2.2
11.	*Escherichia coli*	0.195	0.000018	88.7[20]	90.6	−1.9
12.	*Escherichia coli*	0.6	0.000018	93.9[20]	96.2	−2.3

#*Escherichia coli* K-12 genome sequence (4639675 base pairs) obtained from NCBI (NC_000913) is used for these calculations.
doi:10.1371/journal.pone.0012433.t004

driving force in the melting of DNA. The length of the nucleotide sequence as well as the concentration of the solution also play a substantial role in the melting of DNA, where the effect of concentration is more pronounced than that of length when combined with the strength parameter, but even both of them together do not reach up to the mark of strength parameter taken alone. Although the correlation achieved after adding the strand concentration (DNA) does not improve much, the average error between the experimental and predicted Tm comes down marginally; hence it is retained in the model.

The following methods were reported earlier in the literature for melting temperature predictions: (i) Basic method [17]; (ii) Salt corrected method [20]; (iii) NN method using Breslauer's parameters [14]; (iv) NN method using Santa Lucia's parameters [35]; (v) NN method using Sugimoto's parameters [34] and (vi) Consensus method [37]. On the basis of a previous comparison of various Tm prediction methods, it was observed that the best methods were the Nearest Neighbor methods based on thermodynamic properties, but the major drawback with these methods was that they applied well primarily to oligomers ranging from 4 to 20 bp [37]. Panjkovich and Melo [37] after an extensive study, observed that under certain experimental conditions of salt and oligonucleotide concentration, even a very simple method that did not take into account these parameters could give results similar to the more complex methods, but under variable salt and oligonucleotide concentrations, the thermodynamic methods outperformed the simpler ones. We infer from the results presented here that a simple model [eq. (1)] developed on the basis of a quantification of forces destabilized during melting shows satisfactory performance for any length of the oligonucleotide sequence, salt concentration and base composition.

Extension of the methodology to genomes

The melting temperatures of 20 genomes were also calculated using eq. (1) as described in the methods section. The results are compared with the experimental data [12,13,49,50] and presented in Table 3.

The melting of large and genomic level sequences can be modeled as a cooperative phenomenon, occurring simultaneously at various places along the DNA sequence, where each melting region can be described as a "melting unit" [51]. The size of the melting unit has been a centre of attention for many years. Many estimates have been provided in the literature on the size of the unit specific to a given sequence [52–53], but there has been no molecular level explanation towards the number of base pairs present in a melting unit. Moreover, the size of the melting unit estimated is highly variable. We have investigated the melting temperature for large DNA sequences in terms of melting units of various sizes ranging from 40 bp all the way upto 100 bp and found the predictions to converge well for units of size 60–70 base pairs. Thus a choice of 70 base pairs as a melting unit is made in this study. This is also found to be in accord with the literature regarding packaging of DNA in a compact form with the help of bacterial HU proteins (58 bp [54]), archaeal histones (60 bp [55]; 80 bp [56]) and eukaryal histones (70 bp [54]; 70 bp [57]). These proteins adapt themselves to open the double stranded DNA into single stranded DNA, forming a bubble of approximately the same length as the melting unit, to perform the necessary molecular

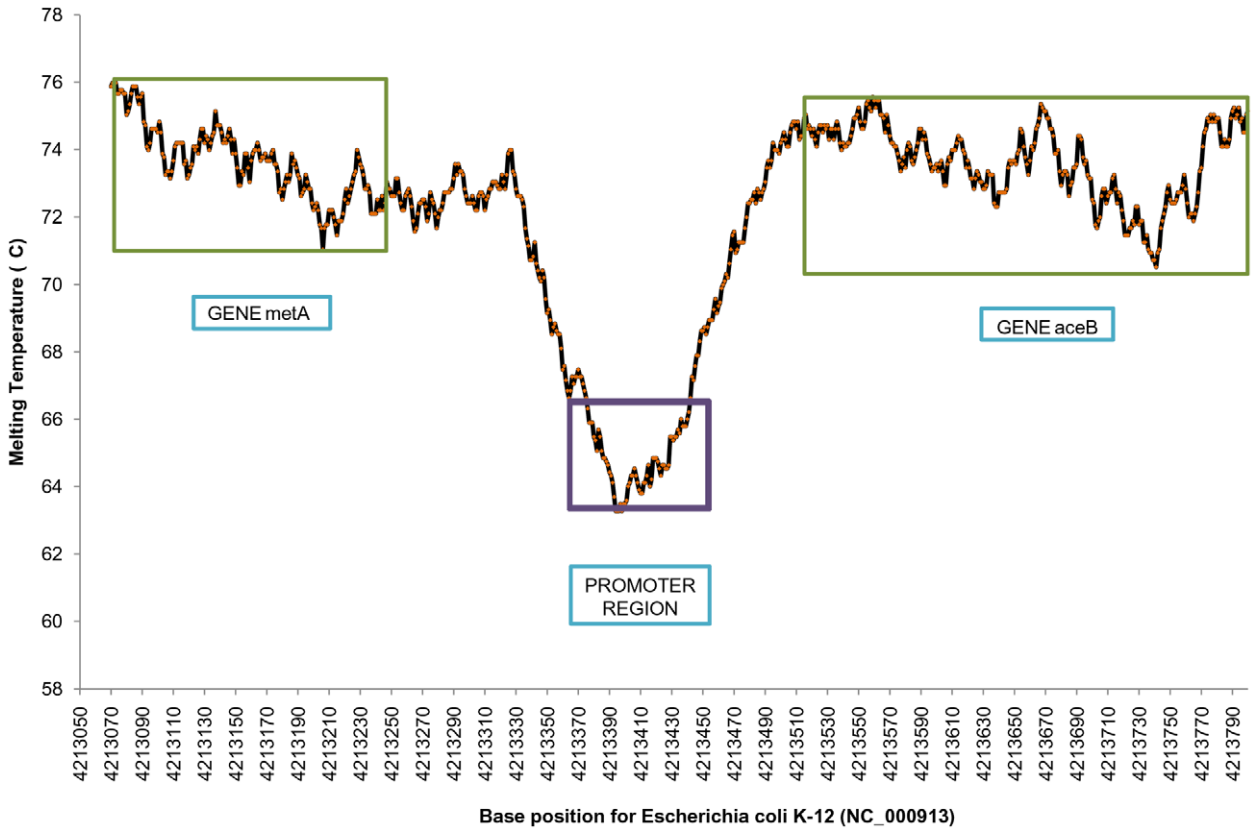

Figure 2. Melting profile of a promoter and its flanking genes. Melting profile for a stretch of 731 base pairs containing a promoter sequence from Ref. 59 and its corresponding experimentally verified gene sequence for *Escherichia coli* K-12 genome (NC_000913).
doi:10.1371/journal.pone.0012433.g002

tasks such as transcription [54–56] and replication of DNA. Our choice (hypothesis) of 70 base pairs seems to be validated by the results presented in Table 3 where the correlation between experimental and predicted values is excellent (correlation coefficient, r = 0.98; average error of prediction = 2.0°C). The last column of Table 3 depicting the difference between experimental and predicted melting temperatures does not show any obvious pattern.

The melting temperatures of *Escherichia coli* at various salt concentrations are calculated and reported in Table 4. It may be seen from the 1st entry (Experimental Tm = 70.7°C) and the 2nd entry (Experimental Tm = 75.7°C) of the table that there are discrepancies in the experimental melting temperature values derived by various methods at nearly the same salt and nucleotide concentrations. Allowing for this difference, it may be noted that the calculations are in general accord with experiment.

In a nutshell, the phenomenological model presented here for melting temperature prediction covers a large range of salt concentration, GC content and length of DNA sequence and could pave the way for a deeper molecular-level understanding of DNA melting.

Potential application of the methodology to genome annotation

Previous work has shown that there appears to be an underlying energy basis for the discrimination of genic and non-genic regions in prokaryotic genomes [57,58]. As the proposed model of Tm prediction is based on the energetics of DNA, it is tempting to examine the melting temperature variations (Tm profiles) along genomic sequences. An illustrative genome profile of a part (4213070–4213801 bp) of *Escherichia coli* genome (NC_000913) is depicted in Figure 2, where a promoter region [59] is clearly differentiated from the gene region. The Tm profile of a gene (GBSS1, Gene Id: FJ235783.1) of *Oryza sativa* is shown in Figure 3, which shows discrimination of the exonic and intronic regions. Thus the methodology shows the ability to discriminate various functional units present on a genome sequence. The lower melting temperature of promoter regions could be due to the requirement of structural adaptation by DNA to facilitate specific binding of regulatory proteins, while the lower melting temperatures of introns relative to corresponding exons might be due to their low thermodynamic stability, as also observed independently by Wada and Suyama two and half decades ago [60]. Clearly, further investigations are required to utilize the strength of the methodology for genome annotation.

Description of the web utility

The melting temperature prediction method presented here is also presented by means of a web utility: www.scfbio-iitd.res.in/chemgenome/Tm_predictor.jsp. The utility has an input box wherein the user can paste the sequence. Alternatively, the user can input the sequence with the help of buttons provided in the utility. In case of large DNA sequences, the user can also upload the sequence file through the browse option provided. The

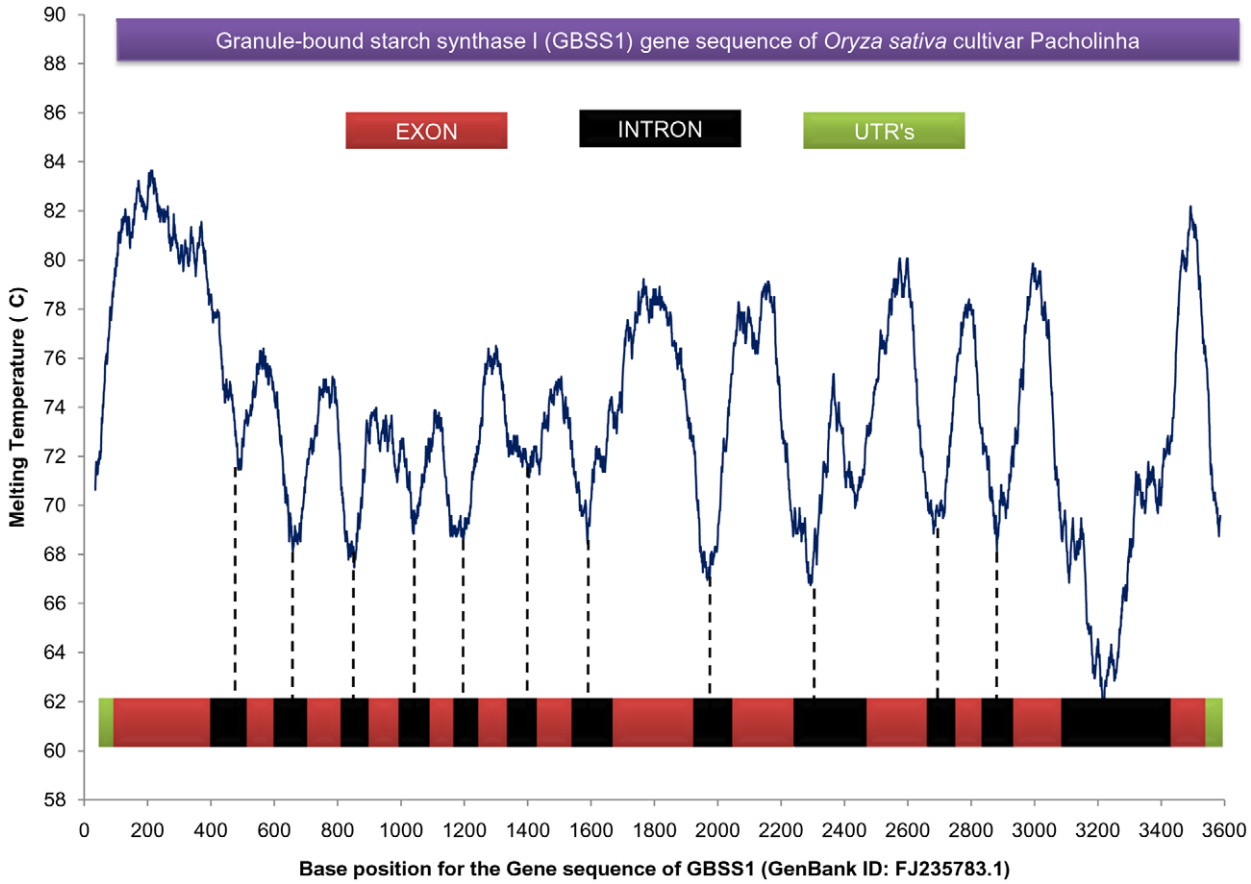

Figure 3. Melting profile of an *Oryza sativa* gene. Melting profile of Granule bound starch synthase I (GBSS1) gene (Length = 3621 base pairs) *of Oryza sativa* cultivar Pacholinha (GenBank ID: FJ235783.1), showing a clear discrimination of exons from introns and Un-translated regions (UTR's).
doi:10.1371/journal.pone.0012433.g003

calculated Tm is reported either on the web page (for smaller sequences) or on the email-id provided by the user (for large sequences). The utility also provides the option of calculating melting temperatures at various salt and DNA concentrations. The training and test datasets and a tutorial to calculate Tm for a small sequence manually are also provided.

Conclusion

A simple phenomenological model is developed for predicting the melting temperatures of DNA sequences based on stacking and hydrogen bonding interactions, length of the sequence, salt and nucleotide strand concentration. The model is applicable to a wide range of sequence lengths including genomic sequences, base composition and salt concentrations. This method thus overcomes the limitations noted earlier of predictive models giving good results in a limited sequence and length data space and smaller range of salt concentration. Work is in progress to develop melting profiles of complete genomes in pursuit of genome annotation to eventually facilitate a molecular level understanding of genome organization.

Supporting Information

Figure S1 Data space representation for the length parameter.
Found at: doi:10.1371/journal.pone.0012433.s001 (0.25 MB TIF)

Figure S2 Data space representation for the salt concentration parameter.
Found at: doi:10.1371/journal.pone.0012433.s002 (0.23 MB TIF)

Figure S3 Data space representation for the %GC content of the sequence.
Found at: doi:10.1371/journal.pone.0012433.s003 (0.48 MB TIF)

Figure S4 Normal probability plot of residuals for the training dataset.
Found at: doi:10.1371/journal.pone.0012433.s004 (0.22 MB TIF)

Figure S5 Distribution of residuals with the predicted melting temperatures.
Found at: doi:10.1371/journal.pone.0012433.s005 (0.37 MB TIF)

Text S1 The equation to predict the melting temperature of DNA without the use of the nucleotide strand concentration.
Found at: doi:10.1371/journal.pone.0012433.s006 (0.02 MB DOC)

Table S1 Experimental and predicted melting temperatures for the training dataset of 123 oligomers.
Found at: doi:10.1371/journal.pone.0012433.s007 (0.22 MB DOC)

Table S2 Experimental and predicted melting temperatures for the test dataset of 225 oligomers.
Found at: doi:10.1371/journal.pone.0012433.s008 (0.38 MB DOC)

Table S3 Experimental and predicted melting temperatures for a dataset of 15-mers.
Found at: doi:10.1371/journal.pone.0012433.s009 (0.16 MB DOC)

Table S4 Experimental and predicted melting temperatures for 40 base pair long oligonucleotide sequences.
Found at: doi:10.1371/journal.pone.0012433.s010 (0.03 MB DOC)

Table S5 Analysis of variance for the regression equation (1) derived from the training dataset.
Found at: doi:10.1371/journal.pone.0012433.s011 (0.03 MB DOC)

Acknowledgments

The web-enabling of the melting temperature prediction utility by Ms. Vandana Shekhar and Mr. Bharat Lakhani and useful suggestions from Prof. D. L. Beveridge, Dr. S. K. Khare and the editor are gratefully acknowledged.

Author Contributions

Conceived and designed the experiments: BJ. Performed the experiments: GK. Analyzed the data: GK. Wrote the paper: GK BJ.

References

1. Cantor CR, Schimmel PR (1980) Biophysical Chemistry Part III: The Behavior of Biological Macromolecules. W H Freeman, San Francisco.
2. Doktycz MJ, Morris MD, Dormady SJ, Beattie KL, Jacobson KB (1995) Optical melting of 128 octamer DNA duplexes. J Biol Chem 270: 8439–8445.
3. Sundaralingam M, Ponnuswamy PK (2004) Stability of DNA duplexes with Watson-Crick base pairs: A predicted model. Biochemistry 43: 16467–16476.
4. Rezac J, Hobza P (2007) On the nature of DNA-duplex stability. Chemistry A Eur J 13: 2983–2989.
5. Peyrard M, Dauxois T, Hoyet H, Willis CR (1993) Biomolecular dynamics of DNA: statistical mechanics and dynamical models. Physica D 68: 104–115.
6. Jayaram B, Beveridge DL (1990) Free energy of an arbitrary charge distribution imbedded in coaxial cylindrical dielectric continua: Application to conformational preferences of DNA in aqueous solutions. J Phys Chem 94: 4666–4671.
7. Arora N, Jayaram B (1998) Energetics of base pairs in B-DNA in solution: An appraisal of potential functions and dielectric treatments. J Phys Chem B 102: 6139–6144.
8. Jayaram B, Jain T (2004) The role of water in protein-DNA recognition. Annu Rev Biophys Biomol Struct 33: 343–61.
9. Wartell RM, Benight AS (1985) Thermal denaturation of DNA molecules: A comparison of theory with experiment. Physics Reports 126: 67–107.
10. Lewin B (2004) Gene VIII. Pearson Prentice Hall, NJ.
11. Porschke, D (1971) Cooperative non-enzymic base recognition II. Thermodynamics of the helix-coil transition of oligoadenylic + oligouridylic acids. Biopolymers 10: 1989–2013.
12. Blake RD (1987) Cooperative lengths of DNA during melting. Biopolymers 26: 1063–1074.
13. Ussery DW (2001) DNA denaturation. Academic Press.
14. Breslauer KJ, Frank R, Blocker H, Marky LA (1986) Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci USA 83: 3746–3750.
15. Delcourt SG, Blake RD (1991) Stacking energies in DNA. J Biol Chem 266: 15160–15169.
16. Lafontaine I, Lavery R (2000) Optimization of nucleic acid sequences. Biophys J 79: 680–685.
17. Marmur J, Doty P (1962) Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. J Mol Biol 5: 109–118.
18. Blake RD, Delcourt SD (1996) Thermodynamics effect of formamide on DNA stability. Nucleic Acid Res 24: 2095–2103.
19. Shaikh SA, Jayaram B (2007) A swift all-atom energy based computational protocol to predict DNA ligand binding affinity and ΔTm. J Med Chem 50: 2240–2244.
20. Schildkraut C, Lifson S (1965) Dependence of the melting temperature of DNA on salt concentration. Biopolymers 3: 195–208.
21. Frank-Kamenetskii MD (1987) How the double helix breathes. Nature 328: 17–18.
22. Barbi M, Cocco S, Peyrard M (1999) Helicoidal model for DNA opening. Phys Lett A 253: 358–369.
23. Kim J-Y, Jeon J-H, Sung W (2008) A breathing wormlike chain model on DNA denaturation and bubble: Effects of stacking interactions. J Chem Phys 128: 055101–055101-6.
24. Kohandel M, Ha B-Y (2006) Thermal denaturation of double-stranded DNA: Effect of base stacking. Physical Review E 73: 011905.
25. Yakovchuk P, Protozanova E, Frank-Kamenetskii MD (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res 34: 564–574.
26. Kanhere A, Bansal M (2005) A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinformatics 6: 1.
27. Zimm BH (1960) Theory of "Melting" of the helical form in double chains of the DNA type. J Chem Phys 33: 1349–1356.

28. Lifson S, Roig A (1961) On the theory of helix-coil transition in polypeptides. J Chem Phys 34: 1963–1974.
29. Grigoryan AV, Mamasakhlisov ESh, Buryakina TYu, Tsarukyan AV, Benight AS, et al. (2007) Stacking heterogeneity: A model for the sequence dependent melting cooperativity of duplex DNA. J Chem Phys 126: 165101–165101-9.
30. Dauxois T, Peyrard M, Bishop AR (1993) Entropy-driven DNA denaturation. Physical Review E 47: R44–R47.
31. Zhang Y-L, Zheng W-M, Liu J-X, Chen YZ (1997) Theory of DNA melting based on the Peyrard-Bishop model. Phys Rev E 56: 7100–7115.
32. Weber G, Haslam N, Whiteford N, Prugel-Bennett A, Essex JW, et al. (2006) Thermal equivalence of DNA duplexes without calculation of melting temperature. Nature Physics 2: 55–59.
33. Campa A, Giansanti A (1998) Experimental tests of the Peyrard-Bishop model applied to the melting of very short DNA chains. Physics Review E 58: 3585–3588.
34. Sugimoto N, Nakano S, Yoneyama M, Honda K (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. Nucleic Acids Res 24: 4501–4505.
35. SantaLucia J, Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci 95: 1460–1465.
36. Protozanova E, Yakovchuk P, Frank-Kamenetskii MD (2004) Stacked-unstacked equilibrium at the nick site of DNA. J Mol Biol 342: 775–785.
37. Panjkovich A, Melo F (2005) Comparison of different melting temperature calculation methods for short DNA sequences. Bioinformatics 21: 711–722.
38. Owczarzy R, You Y, Moreira BG, Manthey JA, Huang L, et al. (2004) Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. Biochemistry 43: 3537–3554.
39. Guckian KM, Schweitzer BA, Ren RX-F, Sheils CJ, Paris PL, et al. (1996) Experimental measurement of aromatic stacking in the context of duplex DNA. J Am Chem Soc 118: 8182–8183.
40. Dixit SB, Beveridge DL, Case DA, Cheatham TE, 3rd, Giudice E, et al. (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. Biophys J 89: 3721–3740.
41. Pullman B, Claverie P, Caillet J (1966) On the exclusivity of hydrogen-bonded pairing between the Watson-Crick complementary bases. J Mol Biol 22: 373–375.
42. Jiranusornkul S, Laughton CA (2008) Destabilization of DNA duplexes by oxidative damage at guanine: implications for lesion recognition and repair. J R Soc 5: 191–198.
43. Haaima A, Hansen HF, Christensen L, Dahl O, Nielsen PE (1997) Increased DNA binding and sequence discrimination of PNA oligomers containing 2,6-diaminopurine. Nucleic Acids Res 25: 4639–4643.
44. Eldrup AB, Christensen C, Haaima G, Nielsen PE (2002) Substituted 1,8-naphthyridin-2 (1H)-ones are superior to thymine in the recognition of adenine in duplex as well as triplex structures. J Am Chem Soc 124: 3254–3262.
45. Jayaram B (1997) Beyond the wobble: The rule of conjugates. J Mol Evol 45: 704–705.
46. Licino P, Guerra JCO (2007) Irreducible representation for nucleotide sequence physical properties and self consistency of nearest-neighbor dimer sets. Biophys J 92: 2000–2006.
47. Petruska J, Goodman MF (1995) Enthalpy-entropy compensation in DNA melting thermodynamics. J Biol Chem 270: 746–750.
48. Analyse-it for Microsoft Excel (version 220) Analyse-it Software, Ltd http://wwwanalyse-itcom/.
49. Mandel M, Igambi I, Bergendahl J, Dodson ML, Jr., Scheltgen E (1970) Correlation of melting temperature and cesium chloride buoyant density of bacterial deoxyribonucleic acid. J Bacteriology 101: 333–338.
50. Owen RJ, Jackman PJH (1982) The similarities between Pseudomonas paucimobilis and allied bacteria derived from analysis of deoxyribonucleic acids and electrophoretic protein patterns. J Gen Microbiol 128: 2945–2954.
51. Wada A, Yabuki S, Husimi Y (1980) Fine structure in the thermal denaturation of DNA: high temperature-resolution spectrophotometric studies. CRC Crit Rev Biochem 9: 87–144.
52. Movileanu L, Benevides JM, Thomas GJ, Jr. (2002) Determination of base and backbone contributions to the thermodynamics of premelting and melting transitions in B-DNA. Nucleic Acids Res 30: 3767–3777.
53. Vamosi G, Clegg RM (2008) Helix-coil transition of a four-way DNA junction observed by multiple fluorescence parameters. J Phys Chem B 112: 13136–13148.
54. Sinden RR (1994) DNA structure and function. Academic Press Inc, California.
55. Reeve JN, Sandman K, Daniels CJ (1997) Archaeal histones, nucleosomes and transcription initiation. Cell 89: 999–1002.
56. Sandman K, Reeve JN (2000) Structure and functional relationships of archaeal and eukaryal histones and nucleosomes. Arch Microbiol 173: 165–169. First published on January 20, 2000, 101007/s002039900122.
57. Dutta S, Singhal P, Agrawal P, Tomer R, Kritee, et al. (2006) A physicochemical model for analyzing DNA sequences. J Chem Inf Model 46: 78–85.
58. Singhal P, Jayaram B, Dixit SB, Beveridge DL (2008) Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. Biophys J 94: 4173–4183.
59. Lisser S, Margalit H (1993) Compilation of E coli mRNA promoter sequences. Nucleic Acids Res 21: 1507–1516.
60. Wada A, Suyama A (1986) Local stability of DNA and RNA secondary structure and its relation to biological functions. Prog Biophys Mol Biol 47: 113–157.