Thermal equivalence of DNA duplexes for probe design

2009 J. Phys.: Condens. Matter 21 034106

(http://iopscience.iop.org/0953-8984/21/3/034106)

IOP Publishing Journal of Physics: Condensed Matter

J. Phys.: Condens. Matter **21** (2009) 034106 (11pp) doi:10.1088/0953-8984/21/3/034106

Thermal equivalence of DNA duplexes for probe design

Gerald Weber[1], **Niall Haslam**[2], **Jonathan W Essex**[3] **and Cameron Neylon**[3,4]

[1] Department of Physics, Federal University of Ouro Preto, Ouro Preto-MG, Brazil
[2] The European Molecular Biology Laboratory, Heidelberg, Germany
[3] School of Chemistry, University of Southampton, Southampton, UK
[4] STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, UK

Received 31 May 2008, in final form 27 August 2008
Published 17 December 2008
Online at stacks.iop.org/JPhysCM/21/034106

Abstract

We present the theory of thermal equivalence in the framework of the Peyrard–Bishop model and some of its anharmonic variants. The thermal equivalence gives rise to a melting index τ which maps closely the experimental DNA melting temperatures for short DNA sequences. We show that the efficient calculation of the melting index can be used to analyse the parameters of the Peyrard–Bishop model and propose an improved set of Morse potential parameters. With this new set we are able to calculate some of the experimental melting temperatures to $\pm 1.2\,^{\circ}\mathrm{C}$. We review some of the concepts of sequencing probe design and show how to use the melting index to explore the possibilities of gene coverage by tuning the model parameters.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

The cost of identifying and determining genetic sequences is rapidly coming down due to a wide range of intense technological efforts [1, 2]. Many of the most widely used methods for determining a DNA sequence use an approach based on sequencing by hybridization (SBH) [3, 4]. This family of techniques exploit the complementary nature of DNA to indirectly read a sequence. The best known implementation of this technology is in DNA microarrays where a library of selected short oligonucleotide probes of specifically designed sequence composition is immobilized on a substrate [5]. In some other technologies, such as the bead arrays from Illumina [6], the position is not known but can be determined. The probes are then brought into contact with a set of target unknown sequences which has usually been fragmented and labelled with a fluorescent dye. Successful hybridization of the probes is then detected via fluorescence or some other signal. As the identity and sequence of the probe in each position is known, successful hybridization at a specific position indicates the presence of the DNA sequence complementary to the probe in the sample.

A key piece of information for the design of a library of probes is the stability of the double helix formed when each probe hybridizes to its complementary sequence [7]. The stability of this interaction will determine both whether a signal is observed and its intensity. The most common measure of DNA duplex stability is its melting temperature, the temperature at which half of a sample of duplex DNA will have de-hybridized. The melting temperature is strongly dependent on the DNA sequence as well as being affected by common solution conditions such as salt and DNA concentration. Identifying probes that will melt within a narrow range of temperatures is often essential to avoid both false positives and false negatives in the SBH experiment. A false negative, i.e., the apparent non-occurrence of a target which is in fact present, may occur if the experimental conditions are such that the temperature is higher than the melting temperature of a correct probe–target duplex. On the other hand, false positives may occur if the experimental conditions lead to a non-complementary target hybridizing to a probe, e.g. if the experiment is carried out below the melting temperature for the formation of the correct duplex [8].

Currently, where probe design is influenced by melting temperature analysis, researchers rely on empirical regression models based on Gibbs free energies [9–12] for the prediction of melting temperatures of unknown DNA sequences. Alternative melting temperature analysis are available through the use of the Poland–Scheraga method [13]. These models

adjust a large number of parameters from experimental datasets and are accurate for most applications. However, for these empirical models the minimized parameters are not linked to any specific aspect of the DNA structure or energetics. This makes it difficult to adapt the parameters to experimental conditions for which an empirical dataset has not yet been fully determined. For this to be possible, one needs models where the parameters representing the molecular interactions can be changed at an intuitive level. In addition, to be useful from a bioinformatics point of view, these models need to allow the fast computation of melting temperatures for thousands or even millions of probes.

One statistical physics model that in principle allows such intuitive modelling is the Peyrard–Bishop model [14]. This model describes, in a simplified way, the hydrogen bond and the stacking interaction in double stranded DNA and was used successfully in numerous applications such as transcription bubble analysis [15, 16], energy localization [17] or to calculate solitonic speed [18]. Recently, we showed that this model can also be used for fast calculation of melting temperatures via the concept of thermal equivalence [19]. The thermal equivalence may even be used in place of the melting temperature. As a result, by changing the conditions of the molecular interaction, say by reducing the strength of the hydrogen bonds between paired bases, it is possible to gain an insight into the behaviour of melting temperatures under untested experimental conditions.

In this paper, we will show how to use the thermal equivalence index for probe design under hybridization conditions for which there are no available experimental data. For this purpose we use as hypothetical experimental environments changed salt concentrations and solvent conditions [20] and show how this affects the probe coverage for a given gene. This paper is organized as follows: in section 2 we will briefly review the 1D Peyrard–Bishop model and some of its variants, where we describe some of our recent contributions on solvent interactions [20]. In section 3 we describe the partition function expansion for non-homogeneous DNA proposed by Zhang *et al* [21] which is essential to the understanding of the thermal equivalence outlined in section 4. In section 5 we discuss the comparison of the thermal equivalence with experimental melting temperatures and how to optimize the parameters of the Peyrard–Bishop model. Also, in this section we present a set of improved Morse potentials. The methods for probe design are discussed in section 6 and we draw our conclusions in section 7.

2. The Peyrard–Bishop model

For Hamiltonians of the type proposed by Peyrard and Bishop [14] to model the denaturation of a homogeneous DNA double helix, the configurational part is written generically as,

$$U = w(y_i, y_{i-1}) + V(y_i), \tag{1}$$

where y_i is the displacement of the ith base pair from equilibrium, w is the stacking interaction of the nearest neighbours i and $i+1$, and V the interaction of the ith base pair [14]. Therefore, this model mixes nearest-neighbour interactions with base-pair interactions, in contrast for instance to Gibbs free energy models which only consider nearest-neighbour free energies [9, 10].

The interaction of the ith base pair is written as a Morse potential of the form

$$V_{\text{Morse}}(y_i) = D(\mathrm{e}^{-ay_i} - 1)^2, \tag{2}$$

and the nearest-neighbour stacking interaction as a harmonic oscillator

$$w_{\text{harm}}(y_i, y_{i-1}) = \frac{k}{2}(y_i - y_{i-1})^2. \tag{3}$$

The advantage of this formalism, which we call the harmonic-Morse (HM) model, is that it describes both the base-pair and stacking interaction with just one variable y, and as such can be calculated easily within the formalism of the transfer integral method (for a detailed discussion see also [21] and [22]).

2.1. Sharp transitions

Unfortunately, in the form of equation (3) this model does not give rise to a sharp first-order-like denaturation which are observed experimentally. This issue was addressed later by Dauxois, Peyrard and Bishop [23, 24] by the addition of an anharmonic term to the stacking interaction

$$w_{\text{an.}}(y_i, y_{i-1}) = [1 + \rho \mathrm{e}^{-\alpha(y_i + y_{i-1})}] w_{\text{harm}}(y_i, y_{i-1}), \tag{4}$$

in this paper we call this important model the anharmonic-Morse model (AM).

Similar anharmonic formulations have been put forth by Joyeux and Buyukdagli [25] with the introduction of a finite stacking potential

$$w_{\text{fin.}}(y_i, y_{i-1}) = \frac{\Delta H}{2}[1 - \mathrm{e}^{-b(y_i - y_{i-1})^2}] + \frac{K_{\text{b}}}{2}(y_i - y_{i-1})^2, \tag{5}$$

where ΔH is a finite stacking energy and the harmonic potential is used with a much smaller elastic constant K_{b} (about three orders of magnitude smaller than k for the HM model). Also, Saccomandi and Sgura [26] proposed the addition of a nonlinear term to the stacking potential of polynomial form

$$w_{\text{pol.}}(y_i, y_{i-1}) = w_{\text{harm}}(y_i, y_{i-1}) + \frac{k_{\text{nl}}}{4}(y_i - y_{i-1})^4, \tag{6}$$

to obtain the similar effect of sharpened transitions, where they introduced the additional parameter k_{nl}.

The addition of anharmonic terms to the nearest-neighbour interaction potential w is not the only form to obtain a sharp DNA denaturation. We showed that the addition of a solvent potential to the base-pair, interaction

$$V(y_i) = V_{\text{Morse}}(y_i) - f_{\text{s}} D[\tanh(y_i/\lambda_{\text{s}}) + 1], \tag{7}$$

also causes such a sharp denaturation [20]. The second term is a solvent interaction potential, adapted from Drukker *et al* [27], which simulates the formation of hydrogen bonds with the solvent once the base-pair hydrogen bonds are displaced by more than λ_{s} from their equilibrium values. For $y_n > \lambda_{\text{s}}$ the base pairs are pulled away from each other until the bond with the solvent is established. Once the bases are bonded to the freely moving solvent molecule they are no longer pushed to any particular direction, a situation which is represented by the potential plateau for $y_n > \lambda_{\text{s}}$ (see also figure 1 in [20]).

2.2. Divergence of the partition function

A numerical divergence problem with the partition function was analysed by Zhang *et al* [21]. Usually, this problem is simply avoided by introducing a finite integration interval for the partition function integral [21], which in practice is equivalent to placing a potential barrier for sufficiently large y. Another way to avoid this divergence was proposed by Theodorakopoulos *et al* [28] who added a small stress term to the base-pair interaction potential. The finite stacking potential of equation (5) proposed by Joyeux and Buyukdagli [25] has a similar effect but was added to the nearest-neighbour potential. Similarly, we proposed an alternative form for the harmonic potential [19, 20], rewriting equation (3)

$$w_{\mathrm{harm}}(y_i, y_{i-1}) = \frac{k}{2}(y_i^2 - 2y_i y_{i-1}\cos\theta + y_{i-1}^2), \qquad (8)$$

where θ is the twist angle between neighbouring base pairs. This is motivated by 3D helicoidal models such as proposed by Barbi *et al* [29], as well as torsional potentials used in molecular dynamics [27]. This formulation is particularly convenient as it can be readily introduced in existing models, such as the anharmonic-Morse model [23, 24], without significantly modifying existing analytical calculations. For an angle of $\theta = 0$ the usual harmonic stacking interaction term [14] is obtained and would represent the situation of perfectly parallel neighbouring bonds. Evidently, the base pairs can only denaturate when the double helix is largely unwound and therefore we use a small, but non-zero, fixed angle of $\theta = 0.01$ rad for the calculations presented in this work.

3. Partition function for inhomogeneous DNA

To apply the Peyrard–Bishop model to realistic DNA sequences, i.e. DNA containing actual genomic information, Zhang *et al* [21] proposed an expansion into orthonormal functions for the integration of the partition function Z. In this section we will describe parts of the method which are relevant for the concept of thermal equivalence. First however, we will introduce a few key aspects of the partition function calculation for homogeneous sequences which will aid the understanding of the calculation for non-homogeneous DNA.

3.1. Homogeneous DNA sequences

The partition function for homogeneous sequences can be written as [14, 21]

$$Z_y = \int \mathrm{d}y_1 \int \mathrm{d}y_2 \cdots \int \mathrm{d}y_N K(y_1, y_2)K(y_2, y_3) \times \cdots K(y_{N-1}, y_N)K(y_N, y_1), \qquad (9)$$

with the kernel function defined as

$$K(y_i, y_{i+1}) = \exp\left(-\frac{1}{kT}\{w(y_i, y_{i+1}) + \tfrac{1}{2}[V(y_i) + V(y_{i+1})]\}\right), \qquad (10)$$

where k is the Boltzmann constant and T the thermodynamic temperature. The partition function can be solved by introducing the integral equation

$$\int K(x, y)\phi(y)\,\mathrm{d}y = \lambda\phi(x), \qquad (11)$$

in which case it reduces to

$$Z = \sum_{n=1}^{\infty} \lambda_n^N. \qquad (12)$$

For very long sequences the partition function can be further simplified, leading to the remarkable result

$$Z \approx \lambda_1^N, \qquad (13)$$

since all eigenvalues are less than unity and λ_1 is the largest eigenvalue.

To calculate the eigenvalues λ and eigenfunctions ϕ we discretize the integration in equation (11) for M points over an interval $[y_i, y_f]$,

$$\int_{y_i}^{y_f} K(x, y)\phi(y)\,\mathrm{d}y = \sum_{k=1}^{M} r_k K(x, y_k)\phi(y_k) = \lambda\phi(x), \qquad (14)$$

where r_k are integration weights given by the specific quadrature chosen for this integration. We write equation (14) for each point x_l, over the same interval and number of points as for y_k, which results in a matrix equation,

$$\mathsf{KR}\Phi = \Lambda\Phi, \qquad (15)$$

with the $M \times M$ matrix

$$\mathsf{K} = \begin{pmatrix} K(x_1, y_1) & \dots & K(x_1, y_M) \\ \vdots & \ddots & \vdots \\ K(x_M, y_1) & \dots & K(x_M, y_M) \end{pmatrix}, \qquad (16)$$

and

$$\Phi = \begin{pmatrix} \phi(x_1) \\ \vdots \\ \phi(x_M) \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_M \end{pmatrix}. \qquad (17)$$

The integration quadrature is represented by the diagonal matrix

$$\mathsf{R} = \begin{pmatrix} r_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & r_M \end{pmatrix}. \qquad (18)$$

The matrix product KR is not symmetric but can be symmetrized [30],

$$(\mathsf{R}^{1/2}\mathsf{K}\mathsf{R}^{-1/2})\mathsf{D} = \Lambda\mathsf{D}, \qquad (19)$$

with $\mathsf{D} = \mathsf{R}^{1/2}\Phi$. The new matrix product $\mathsf{R}^{1/2}\mathsf{K}\mathsf{R}^{-1/2}$ is symmetric and results in real eigenvalues and eigenfunctions, which can be evaluated by standard numerical eigenvalue techniques [30]. In this work we use a Gauss–Legendre quadrature with 400 points over an interval $[-0.1, 20.0]$ nm.

3.2. Inhomogeneous DNA

For non-homogeneous sequences the model parameter may change from site to site, therefore an additional index representing the nearest neighbours needs to be added to the kernel function K in the partition function of equation (9) [21]

$$Z_y = \int \mathrm{d}y_1 \int \mathrm{d}y_2 \cdots \int \mathrm{d}y_N K^{(1,2)}(y_1, y_2) K^{(2,3)}(y_2, y_3) \times \cdots K^{(N-1),N}(y_{N-1}, y_N) K^{(N,1)}(y_N, y_1), \quad (20)$$

the kernel function K from equation (10) now takes the form

$$K^{(i,i+1)}(y_i, y_{i+1}) = \exp\left(-\frac{1}{kT}\{w^{(i,i+1)}(y_i, y_{i+1}) + \tfrac{1}{2}[V^{(i)}(y_i) + V^{(i+1)}(y_{i+1})]\}\right), \quad (21)$$

where w and V may have different model parameters from site to site.

Zhang *et al* [21] proposed the use of a site-independent set of orthonormal functions ϕ to expand the function $K^{(i,i+1)}(y_i, y_{i+1})$,

$$K^{(i,i+1)}(x, y) = \sum_{n,m=1}^{P} C_{nm}^{(i,i+1)} \phi_n(x) \phi_m(y), \quad (22)$$

which is truncated to P terms and the coefficients $C_{nm}^{(i,i+1)}$ are calculated from

$$C_{nm}^{(i,i+1)} = \int \int K^{(i,i+1)}(x, y) \phi_n(x) \phi_m(y)\, \mathrm{d}x\, \mathrm{d}y. \quad (23)$$

Note that for homogeneous DNA sequences the matrix C_{nm} reduces to a diagonal matrix with eigenvalues $C_{nn} = \lambda_n$ as in equation (11). The partition function equation (20) is written as successive multiplications of $P \times P$ square matrices,

$$Z = \mathrm{Tr}\,(\mathsf{C}^{(1,2)}\mathsf{C}^{(2,3)} \cdots \mathsf{C}^{(N,1)}), \quad (24)$$

where each of the matrices $\mathsf{C}^{(i,i+1)} = [C_{nm}^{(i,i+1)}]$ represents the interaction between neighbouring base pairs i and $i+1$, and Tr is the matrix trace. The last matrix, $\mathsf{C}^{(N,1)}$, represents the boundary condition which links the first and the last base pair. The boundary conditions can either be periodic, where the DNA sequence is considered as a ring, or open ended where the stacking interaction is neglected, i.e., $w^{(N,1)} = 0$. From a numerical point of view, this partition function is advantageous. After selecting the parameters it is sufficient to calculate the matrices $\mathsf{C}^{(a,b)}$ for each type of nearest neighbours (a, b) just once, then for any arbitrary DNA sequence the calculation of equation (24) is a simple matter of rearranging the matrix multiplication. As a consequence, for periodic boundary conditions, a circular permutation of the base pairs leaves the partition function unchanged. The numerical efficiency of the matrix multiplication is defined by the dimension of the matrices C. In this work, matrices of size $P = 30$ are sufficiently accurate for our purposes. For longer sequences the matrix multiplication may become numerically intensive. However, in this case one may still achieve sufficient numerical accuracy with smaller matrices since the first few matrix elements may dominate the final value of the partition function as shown for homogeneous sequences (see section 3.1).

The central problem of this method is to find a suitable set of orthonormal functions ϕ. Zhang *et al* [21] tested several sets and concluded that the one that works best is a set of functions obtained for a homogeneous DNA sequence. For instance, we may choose a DNA duplex formed only of CG base pairs such that the matrix $\mathsf{C}^{(\mathrm{CG,CG})}$ is simply

$$\mathsf{C}^{(\mathrm{CG,CG})} = \Lambda^{(\mathrm{CG,CG})} \quad (25)$$

where $\Lambda^{(\mathrm{CG,CG})}$ is calculated from equation (19). Using this set of functions, the non-homogeneous matrices $\mathsf{C}^{(\mathrm{AT,AT})}$ and $\mathsf{C}^{(\mathrm{CG,AT})}$ are then readily calculated from equation (23).

4. Thermal equivalence

The Hamiltonian in equation (1) contains all terms of the molecular interaction for a particular model and the partition function from equation (24) contains all information regarding the composition of the DNA sequence through the ordering of the matrix multiplication. Therefore, the Hamiltonian and the order of the matrix multiplication completely define the outcome of any subsequent calculation such as the average base-pair displacement $\langle y \rangle$. The main idea of the thermal equivalence concept is to calculate an intermediate physical quantity which would allow mapping it to measured melting temperatures. This physical quantity could then act as an melting index which to be effective should not be dependent, or at least not strongly dependent, on the temperature for which the calculation is carried out.

4.1. Rewriting the partition function

For the remaining part of this paper we assume that the expansion of the partition function uses as a basis set a homogeneous CG-sequence, equation (25). Therefore, the matrices in the partition function of equation (24) are diagonal if they represent CG–CG nearest neighbours, and non-diagonal otherwise. We can now define a non-diagonal matrix $\Delta^{(a,b)}$ such that

$$\mathsf{C}^{(a,b)} = \Lambda + \Delta^{(a,b)}, \qquad \Lambda = \mathsf{C}^{(\mathrm{CG,CG})} \quad (26)$$

where $\Delta^{(a,b)}$ represents the difference of the interaction between neighbours of type (a, b) and neighbours of type (CG, CG). The partition function is then rewritten as,

$$Z = \mathrm{Tr}\,[(\Lambda + \Delta^{(1,2)})(\Lambda + \Delta^{(2,3)}) \ldots (\Lambda + \Delta^{(N,1)})]. \quad (27)$$

Carrying out the matrix multiplication and using common properties of the trace we obtain

$$Z = \sum_{\omega=0}^{N} Z_\omega(\Lambda) = \sum_{\omega=0}^{N} \mathrm{Tr}[M(\Lambda^\omega)], \quad (28)$$

where $M(\Lambda^\omega)$ are all terms containing ω multiplications of the matrix Λ. Unfortunately, factorizing each term in Λ analytically is not possible due the non-commutativity of the matrix multiplication. Lower orders of Λ are obtained

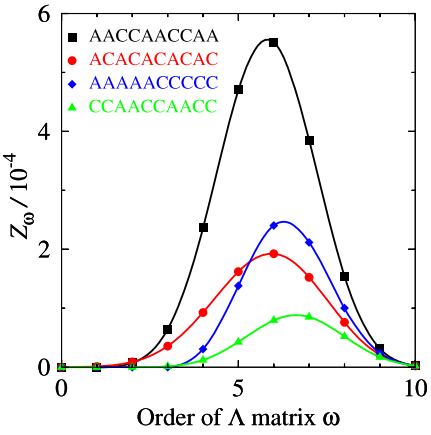

Figure 1. Partition function Z_ω as a function of the order ω of the diagonal matrix Λ for several short sequences with similar CG-content. The continuous line is the Gaussian interpolation $\mathcal{F}(\omega)$.

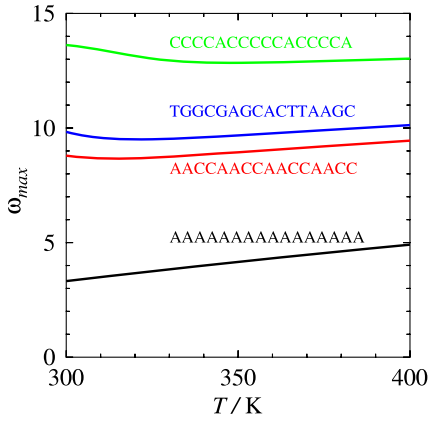

Figure 2. Maximal ordering parameter $\omega_{\max}$ versus temperature. Several example sequences of length 16 bp were considered.

depending on the amount of non-null Δ matrices. Therefore the weight of each order of Λ^ω represents a measure of sequence composition as well as the relative importance of the base-pair and the stacking interaction.

To evaluate all terms as an order of Λ^ω, we track numerically the number of matrices Λ for each term in the matrix multiplication of equation (27). In figure 1 we show the contribution of each order Λ^ω to the partition function. We notice a distinctive Gaussian shape for $Z_\omega(\Lambda)$ and the curves are Gaussian interpolation $\mathcal{F}(\omega)$ through the calculated values $Z_\omega(\Lambda)$. The Gaussian or normal distribution is indeed characteristic of the algebraic binomial expansion, and in the limit of long sequences the Gaussian regression $\mathcal{F}(\omega)$ is $Z_\omega(\Lambda)$ itself, i.e. $\lim_{N\to\infty} Z_\omega(\Lambda) = \mathcal{F}(\omega)$. Evidently, the partition function has a strong temperature dependence, and using it to compare different sequences would not be practical. However, unlike the partition function, the maxima $\omega_{\max}$ of the interpolated Gaussian function $\mathcal{F}(\omega)$ has no such strong temperature dependence, and its change with sequence composition is well behaved as shown in figure 2.

One should note that for very long sequences calculating the maximal order of Λ by numerically tracking the matrix multiplication becomes impractical. However, it is possible to speed up the calculation by reducing the dimension of the matrices for equation (28) while giving up some numerical accuracy. For the extreme case of keeping only the first element of the matrices it is even possible to obtain an analytical result for the partition function as we will show in section 4.2.

4.2. Approximate thermal equivalence

If we use only the first elements of the matrices in equation (27), the partition function in equation (28) simplifies to

$$Z = \prod_{a,b=\mathrm{CG,AT}} (\lambda_1 + \delta_{a,b})^{N_{a,b}}, \tag{29}$$

where $\delta_{a,b} = [\Delta^{(a,b)}]_{1,1}$, and $N_{a,b}$ is the number of nearest neighbours of type (a, b). For homogeneous stacking parameters we can use $\delta_{\mathrm{CG,AT}} = \delta_{\mathrm{AT,CG}}$ and obtain

$$Z = \lambda_1^{N_{\mathrm{CG,CG}}} (\lambda_1 + \delta_{\mathrm{AT,AT}})^{N_{\mathrm{AT,AT}}} \times (\lambda_1 + \delta_{\mathrm{CG,AT}})^{N_{\mathrm{AT,CG}}+N_{\mathrm{AT,CG}}}. \tag{30}$$

Using the binomial expansion of $(p + q)^N$ and the property that the maximum of the binomial distribution

$$\binom{N}{n} p^n q^{N-n}, \tag{31}$$

is given by

$$n_{\max} \approx \frac{N}{1 + q/p}, \tag{32}$$

we can write an approximate thermal equivalence

$$\omega_{\max} \approx N_{\mathrm{CG,CG}} + \frac{N_{\mathrm{AT,AT}}}{1 + \delta_{\mathrm{AT,AT}}/\lambda_1} + \frac{N_{\mathrm{CG,AT}} + N_{\mathrm{AT,CG}}}{1 + \delta_{\mathrm{CG,AT}}/\lambda_1}. \tag{33}$$

If the parameters of the stacking interactions (equations (4), (5) and (8)) for different nearest neighbours of type a, b are known, i.e., for non-homogeneous stacking interactions, the approximate thermal equivalence can be generalized to

$$\omega_{\max} \approx \sum_{a,b} \frac{N_{a,b}}{1 + \delta_{a,b}/\lambda_1} = \sum_{a,b} \frac{\lambda_1 N_{a,b}}{C_{1,1}^{a,b}}. \tag{34}$$

We tested this approximation numerically varying the sequence lengths and verified that the difference between the approximate and exact $\omega_{\max}$ (calculated from the single matrix element partition function of equation (29)) is less than 1% for sequences longer than 9 bp. This difference drops to less than 0.2% for sequences longer than 60 bp for any amount of CG-content. Therefore, for the sequence lengths considered in this work the approximate $\omega_{\max}$ from equation (33) yields the same results as calculating the partition function with matrices of size $P = 1$ in equations (28) or (29). However, they do not become more accurate for longer sequences when compared to calculations with larger matrix dimension ($P \geqslant 2$) since important off-diagonal elements are absent.

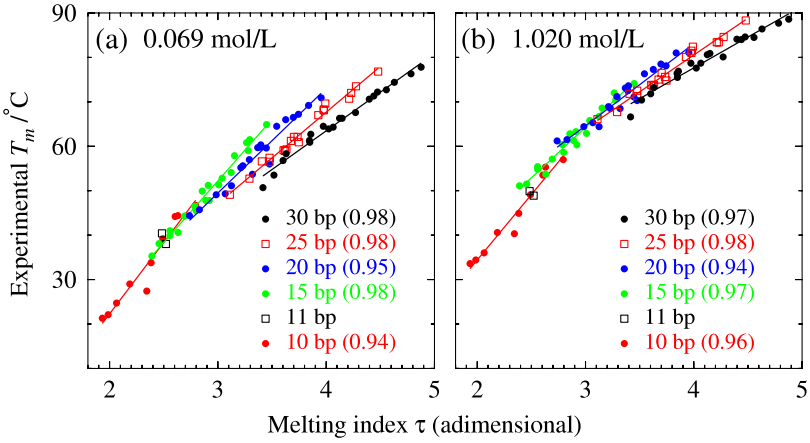

Figure 3. Experimental melting temperatures T_m as a function of the melting index τ. Experimental data are from [12]. Salt concentrations are [Na^+] are (a) 0.069 mol l^{-1} and (b) 1.020 mol l^{-1} [12]. Straight lines are linear regressions through the data points for each group of same size, except for the two data points of size 11 bp, and with average temperature deviation $\langle \Delta T \rangle$ of (a) 2.2, 1.2, 1.7, 0.9, 1.0 °C and (b) 1.7, 1.1, 1.4, 0.9, 1.1 °C, for increasing sequence length. Values in parenthesis are the linear correlation coefficients R^2. Melting indexes were calculated with the AM4 model (table 1).

Table 1. Summary of the model parameters studied in this paper for the harmonic-Morse model (HM), the anharmonic-Morse model (AM), the harmonic-Morse-solvent model (HMS) and the Finite Stacking model (FS). D_{AT} and D_{CG} are given in meV. Except for the FS model we used $\theta = 0.01$ rad throughout. χ^2 and $\langle \Delta T \rangle$ are the merit function and average temperature deviation defined in section 5.3. The values in parenthesis are χ^2 and $\langle \Delta T \rangle$ calculated for the approximate thermal equivalence of equation (33).

Model (Eqs)	D_{AT}	D_{CG}	D_{CG}/D_{AT}	Remaining parameters	χ^2 (K^2)	$\langle \Delta T \rangle$ (K)
Harmonic models with optimized Morse potentials						
HM ((2) + (8))	39	80	2.05	$\lambda_{AT} = 0.0333$ nm $\lambda_{CG} = 0.0125$ nm $k = 2.5$ eV nm^{-2}	1062 (1429)	1.18 (1.38)
HMS1 ((7) + (8))	45	98	2.18	As in HM and $\lambda_s = 0.1$ nm $f_s = 0.1$	1057 (1374)	1.17 (1.36)
HMS2 ((7) + (8))	39	80	2.05	As in HM and $\lambda_s = 1$ nm $f_s = 0.1$	1062 (1482)	1.18 (1.41)
HMS3 ((7) + (8))	27.7	74.5	2.69	As in HMS1 $\lambda_{AT} = 0.0666$ nm $\lambda_{CG} = 0.0250$ nm	—	—
Anharmonic models						
AM1 ((2) + (4))	38	42	1.105	As in [21] figure 10	50 236	9.0
AM2 ((2) + (4))	50	75	1.5	As in [31]	1938 (1117)	1.6 (1.21)
AM3 ((2) + (4))	50	80	1.6	As in [19]	1445	1.4
Anharmonic models with optimized Morse potentials						
AM4 ((2) + (4))	39	77	1.97	As in [19]	1066 (1097)	1.18 (1.20)
FS ((2) + (5))	9.7	48	4.94	As in [25]	1058 (2111)	1.18 (1.74)

5. Comparison with experimental melting temperatures

In this section we describe the comparison with experimental melting temperatures. A summary of the various models and parameters which will be discussed in the following sections are shown in table 1.

5.1. Mapping the equivalence index to T_m

Owczarzy *et al* [12] reported on a carefully measured set of melting temperatures for sequence lengths between 10 and 30 bp and various salt concentrations. Figure 3 shows these experimental melting temperatures for two salt concentrations [12] ordered as a function of the calculated $\omega_{max}^{1/2} = \tau$ for each DNA sequence. It is immediately clear from figure 3 that the square root of ω_{max} represents a convenient quantity for comparing DNA sequences, i.e., sequences with close $\omega_{max}^{1/2}$ should have similar melting temperatures. Based on this observation we call ω_{max} the thermal equivalence and introduce the melting index $\tau = \omega_{max}^{1/2}$. The data points show a clear linear dependence with τ, and linear regressions through each group of same sequence size provide a linear fit with standard deviation between 0.9 and 2.2 °C. The linear correlation coefficient varies from $R^2 = 0.94$ to 0.98 representing an excellent fit to a straight line. The

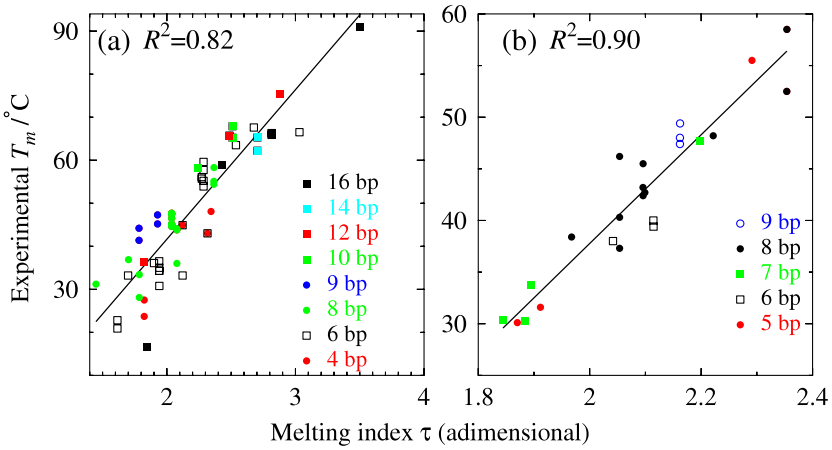

Figure 4. Experimental melting temperatures T_m as a function of the melting index τ for very short DNA sequences. Experimental data are from (a) SantaLucia [32] and (b) Wu *et al* [33]. Straight lines are linear regressions through the complete set of data points and linear correlation coefficients R^2 are shown on the graph. Melting indexes were calculated with the AM4 model (table 1).

results presented in figure 3 are for the two extremes of salt concentration reported in [12]. Similar results are obtained for the remaining salt concentrations reported by Owczarzy *et al* [12] (data not shown).

For short sequences of 10 bp the fit is less optimal with the poorest linear correlation coefficient and largest average temperature difference of up to 2.2 °C, as shown in figure 3 (see also figure caption). This is also observed for even shorter sequences from other sources [32, 33]. In figure 4 we show the melting index τ for very short sequences mapped against the experimental melting temperatures from SantaLucia [32] and Wu *et al* [33]. Unlike the data for longer sequences, we have not separated the melting temperatures into groups of same length. Instead, we performed the linear regression for the whole set of melting temperatures for which we obtained linear regression coefficients between $R^2 = 0.82$ and 0.90. One likely source for the poorer correlation coefficient for short sequences may be the end effects such as fraying at the open ends of the DNA duplex which are not yet considered by Peyrard–Bishop models.

5.2. Calculating the melting temperature from the melting index

In the previous section we showed that for the experimental melting temperatures measured by Owczarzy *et al* [12] each group of sequences of the same length N could be fitted to a straight line as a function of the melting index τ (figure 3). Once we know the linear coefficients we may predict the melting temperature T_p for an unknown sequence with a calculated melting index τ,

$$T_p = a_0(N, [Na^+]) + a_1(N, [Na^+])\tau, \qquad (35)$$

where $a_{0,1}$ are the linear regression coefficients calculated as function of the sequence length N and salt concentration $[Na^+]$. Fortunately, for the melting temperatures reported by Owczarzy *et al* [12] we found that the linear coefficients $a_{0,1}$ are also linear functions of $N^{1/2}$

$$a_k(N, [Na^+]) = b_{0,k}([Na^+]) + b_{1,k}([Na^+])N^{1/2}, \qquad (36)$$

which allows us to calculate the melting temperature T_p for sequences with lengths different from those of the experimental dataset. To calculate the melting temperatures for unknown salt concentrations one may also calculate the coefficients $b_{j,k}$ as linear regressions as a function of the logarithm of the salt concentration

$$b_{j,k}([Na^+]) = c_{0,j,k} + c_{1,j,k} \log[Na^+]. \qquad (37)$$

One of the key aspects of using the melting index τ as a predictor for melting temperatures is that the DNA sequence melts over a very narrow range of temperatures, i.e. that its melting can be described as a two-state helix-coil denaturation. This is generally true for the short sequences used in this work [12, 32, 33] but not for long sequences. Also, for long sequences calculating the melting index τ becomes numerically intensive as discussed in section 4.2. Therefore, the method for calculating melting temperatures described in this work is generally restricted to short sequences.

5.3. Improving the model parameters

Despite the popularity of the Peyrard–Bishop model, only few attempts were made to improve on the model parameters. Campa and Giansanti [31] carried out complete partition function calculation to investigate the parameters for the AM model by comparing them with experimental melting profiles. To date, these parameters by Campa and Giansanti [31] are the most widely used, including for applications where the precise knowledge of these parameters is of crucial importance [34–38]. Here we describe the optimization of these parameter by comparing the predicted melting temperatures them to large sets of measured DNA melting data [12].

The use of the melting index τ allows temperatures to be calculated several orders of magnitude faster than a complete temperature calculation with Peyrard–Bishop-type models [21]. Therefore, we are now in position to address the question of optimizing the model parameters to provide a

better fit with large sets of experimental data. First, we define the merit functions

$$\chi^2 = \sum (T_\mathrm{p} - T_\mathrm{m})^2, \qquad \langle \Delta T \rangle = N^{-1} \sum |T_\mathrm{p} - T_\mathrm{m}| \quad (38)$$

where T_m are the experimental melting temperatures and T_p the predicted temperatures from equations (35)–(37). We then minimize χ^2 as a function of the two Morse potentials D_CG and D_AT. In table 1 we summarize the optimized Morse potentials for the models discussed in this paper. Typically we are able to reduce χ^2 from 1445 K^2 to less than 1070 K^2 and $\langle \Delta T \rangle$ from 1.4 to 1.18 °C. However, for the various models considered there is yet no clear indication which one would be the best to use. This may come from the fact that all models which try to improve over the simpler HM model aim at producing sharp melting transitions. These sharp first-order-like transitions occur mainly for longer sequences, while the experimental data [12, 32, 33] are generally for very short sequences of 30 bp or less. In figure 5 we show a map of the merit function χ^2 as a function of the two Morse potential D_CG and D_AT for the AM model. Within a $\chi^2 \leqslant 1080$ K^2 there is a region corresponding to a ratio $D_\mathrm{CG}/D_\mathrm{AT} \approx 2$. Along this region any combination of Morse potentials with this ratio yield an acceptably low χ^2. This finding may be of importance to reconcile the order of magnitude of the Morse potentials to those used molecular dynamics [27].

At present, optimizing only the Morse potentials D_CG and D_AT, which alone takes about 5 h of computation on a standard 2 GHz processor, does not allow us to favour one specific Peyrard–Bishop variant over the other. This would at least require the optimization of the complete set of parameters for any given model. However this is no simple undertaking, even for the simplest HM model we would need to optimize 14 parameters requiring considerable computational effort even within the simplified scheme of the thermal equivalence. We are currently working on this important problem and hope to report on its progress in the near future.

Table 1 also shows the values for the merit function χ^2 and $\langle \Delta T \rangle$ calculated for the approximate equivalence index of equation (33). For the AM4 parameters the average temperature deviation calculated with the approximate thermal equivalence is practically as good as for the complete calculation. The accuracy of the approximate melting index is mainly determined by how dominant the first eigenvalue of the matrix equation (23) is. Generally, we found that the approximate melting index is closest to the exact melting index for the anharmonic models (AM).

6. Probe design

Probe library design is the selection of probes to be used in sequencing by hybridization (SBH) techniques [39]. The design depends on numerous factors such as the amount of probes that can be placed onto a microarray as well as on its intended application. Most of these applications depend on previously acquired knowledge of the target sequence [40], especially if it is to be applied to single nucleotide polymorphism (SNAP) genotyping [41–43] and resequencing or tracking the evolution of a target sequence [44]. Therefore,

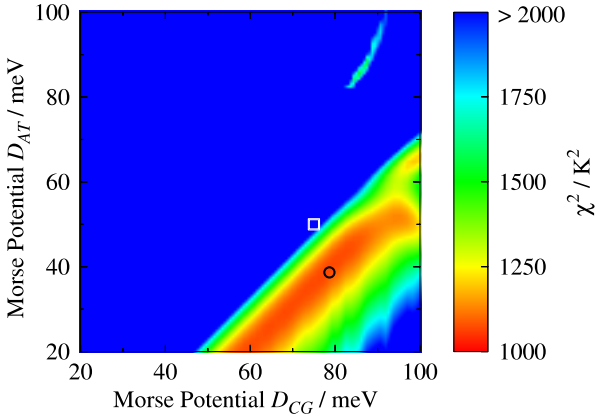

Figure 5. Map of the merit function χ^2 as a function of the Morse potentials D_CG and D_AT. The black circle shows the best value for χ^2, corresponding to AM4, while the white box shows the position of the Morse parameters AM2 which are currently used in the literature. The dark blue region shows merit function values higher than 2000 K^2 which, at some points, reach values up to 10^{11} K^2.

the pre-existing target sequences need to be analysed, e.g. for their genetic variability. A set of probes of fixed length, simply called 'the library,' is generated from this combinatorial analysis which then needs to be further refined especially for melting temperatures. Ideally, all the probes in a library should hybridize within a narrow range of temperatures [7]. If some of the probes can only hybridize with the target at temperatures above the library average, presumably the temperature for which the experiment is carried out, these probes may never detect the target resulting in false negatives. Conversely, if the probe hybridizes at lower temperatures, there is an increased probability of cross-hybridization, i.e., the hybridization to targets that are not entirely complementary to the probe. In this case the probe 'detects' a target whose sequence composition does not really correspond to that of the probe and this results in false positives [8].

Restrictions of the microarray fabrication process usually require all probes to be of the same length. The length of the probe set can be varied for the library as a whole but not generally from probe to probe. Therefore, one aspect that probe design needs to take into account is the optimal probe length for a sequencing library. Recently, we analysed libraries for highly varying genes [45] and found no general rule for the optimal probe length of a library. For each specific target the library needs to be entirely redesigned and new optimal probe lengths need to be established. One main reason for this is precisely the probe melting temperatures: probes of different lengths result in completely different distributions of hybridization temperatures which in turn affect how much of a gene can be covered by the library.

It is beyond the scope of this paper to describe the intricate details of the combinatorial analysis for probe design, these can be found in [45] and references therein. Instead, we base our discussion on an example set of 2434 probes of length 20 nt generated to sequence the Influenza *np*5 gene of size 1588 that was reported in [45]. Multiple probes are used to cover highly variable regions of the gene and this is one of the reasons why

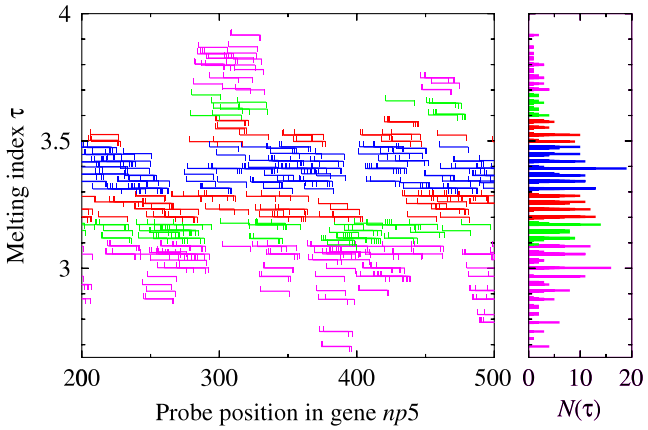

Figure 6. Distribution of probes along a section of a specific variant of the *np5* gene and their respective melting index τ which was calculated using the HMS1 model parameters (see table 1). To the right the frequency distribution of the probes $N(\tau)$ is shown. Short up and down dashes represent the beginning and the end of each probe. (Online only) Probes are colour coded according to their distance to the most frequent melting index $\Delta = |\tau - \tau_f|$: blue for $\Delta < 0.1$; red for $\Delta < 0.2$; green for $\Delta < 0.3$ and magenta otherwise.

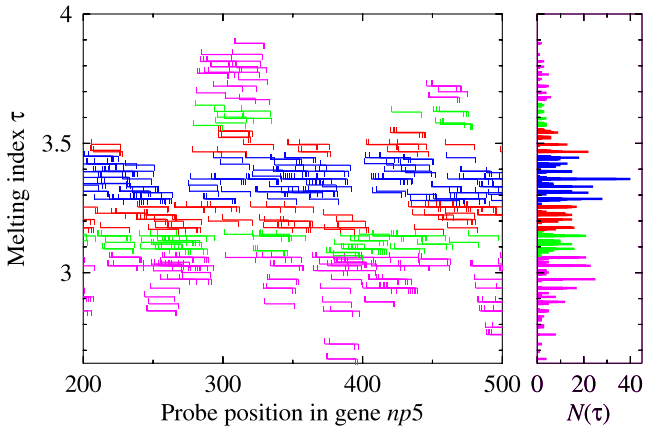

Figure 7. Distribution of probes along a section of a specific variant of the *np5* gene and their respective melting index τ which was calculated using the HMS1 model parameters with modified elastic constant $k = 3$ eV nm^{-2} (see table 1). To the right the frequency distribution of the probes $N(\tau)$ is shown. (Online only) Colour coding of the probes follows the same rule as in figure 6.

there are more probes than sequence positions. In figure 6 we show how these probes would cover a section of the *np5* variant (AX350192.1) as a function of the melting index τ. An isothermic library is generated by removing all probes within a narrow interval of the melting index τ. From the analysis of figure 6 it is clear that, regardless of the choice of the melting index range, most of the probes will be removed from the library. Also, for those highly variable sites for which there are large number of probes covering the same region, their number will be severely reduced therefore leaving important parts of the gene uncovered. This leads to the following question: can we change the number of probes that are likely to be removed due to isothermal requirements by changing the hybridization conditions?

Salt concentration changes the melting temperature of DNA strands and therefore has significant effects over the hybridization condition [12]. One key effect of the salt concentration is to change the elastic properties of DNA [46, 47] as well as other structural properties such as the torsional writhe of circular DNA [48]. Generally, DNA stiffens with increasing salt concentration [49] which means that its effect could be simulated by increasing the elastic constant k of the Peyrard–Bishop Hamiltonian [50]. In figure 7 we show the results of increasing the elastic constant k of the HMS1 data set by 20% which simulates an increasing stiffness due to higher salt concentrations. This indicates that a possible outcome of increasing the salt concentration for this particular set of probes would be to concentrate the largest number of probes at lower melting indexes when compared to figure 6.

Another way to change the hybridization conditions would be to change the solvent. Generally, nonaqueous solvents lower the temperature of the DNA hybridization [51, 52]. For instance for glycerol the relative optical absorbance at 260 nm shows a sharper melting transition and at lower temperatures

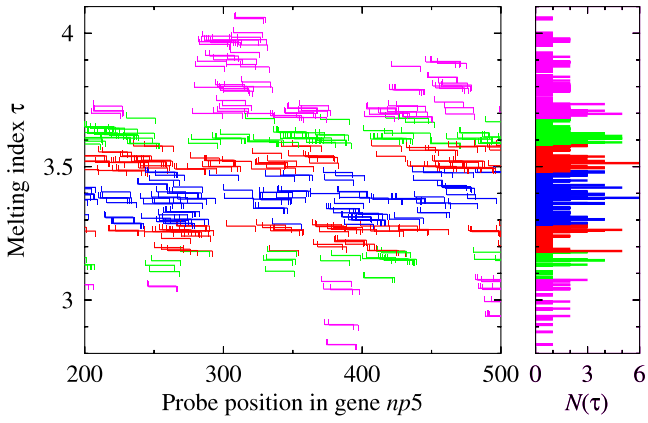

Figure 8. Distribution of probes along a section of the *np5* gene and their respective melting index τ which was calculated using the HMS3 model with $D_{AT} = 27.7$ meV and $D_{CG} = 74.5$ meV, corresponding to a reduction of 20% and 10% respectively to the values for HMS1 (table 1). Values of $\lambda_{AT} = 0.0666$ nm and $\lambda_{CG} = 0.0250$ nm are twice the value for HMS1. Short up and down dashes represent the beginning and end of each probe. (Online only) Colour coding of the probes follows the same rule as in figure 6.

than in aqueous solutions [51]. Similarly, aqueous solutions of diethylsulfoxide also lower the melting temperature with the additional property of an increased DNA denaturation rate [53]. Experimental melting temperatures for these solvents are scarce and do not allow us at present to make an analysis similar to that presented in section 5. We may consider that under certain solvent conditions the hydrogen bonding is decreased. An increase of the interstrand distance has also been observed for mixed solvents by Hammouda and Worcester [52]. To test how different solvent conditions may affect the thermal distribution of the probes we modify the conditions of the hydrogen bond Morse potential. In figure 8 we show the dependence with melting index for the same probe library as in figure 6 by changing some of the model parameters: we reduced the Morse potentials to $D_{AT} = 27.7$ meV and $D_{CG} = 74.5$ meV, a reduction of 20% and

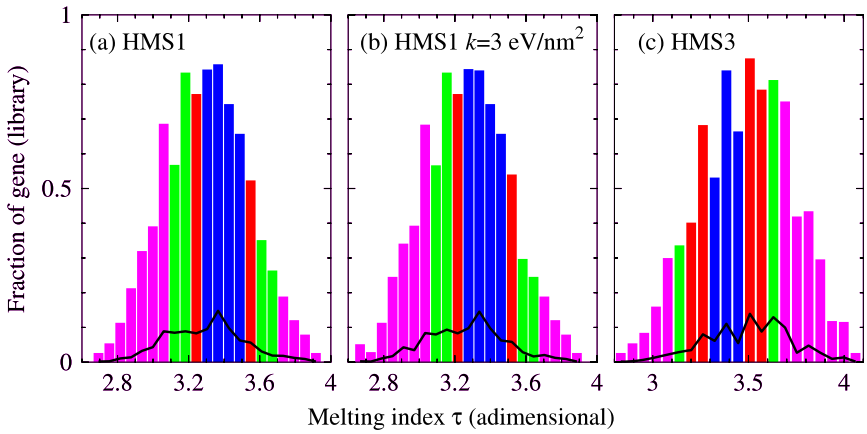

Figure 9. Histogram showing the fraction of the *np*5 gene which is covered by isothermal probes for (a) the HMS1, (b) HMS1 with modified elastic constant $k = 3$ eV nm^{-2} and (c) HMS3 data sets. The curve shows the fraction of the original probe library that remains after thermal filtering. The histogram interval is $\Delta\tau = 0.06$ and for part (a) this corresponds to temperature intervals between 1.35 °C for 0.069 mol l^{-1} and 1.1 °C for 1.020 mol l^{-1} [12]. (Online only) Colour coding corresponds to figure 6 for part (a) and figure 7 for part (b) and figure 8 for part (c).

10% respectively, and doubled the characteristic length $\lambda_{AT} = 0.0666$ nm and $\lambda_{CG} = 0.0250$ nm (HMS3 in table 1). These are relatively minor changes as can be seen by comparing them to other parameter sets in table 1 and corresponds to reducing the melting temperature, since the hydrogen bond are now considerably weaker. Also, we allow for a larger interstrand distance by increasing the Morse potential width λ. As a result, the probes now present a quite different melting index distribution. The difference between the HMS1 and HMS3 parameter set can be better appreciated in figure 9 where we show the fraction of the gene that is covered by the isothermal library. For instance, for the HMS3 parameters there is a much better coverage for the probes with higher melting index. This would provide some cover for the highly variable gene position at 300 of the *np*5 gene (see figure 6).

7. Conclusions

We presented the theory of the thermal equivalence in the framework of the Peyrard–Bishop model. The thermal equivalence gives rise to a melting index τ which maps closely the experimental DNA melting temperatures for short DNA sequences. We used the efficient calculation of the melting index to analyse the Hamiltonian parameters of the Peyrard–Bishop model and propose an improved set of Morse parameters which predicts the experimental melting temperatures within ± 1.2 °C. We showed how the melting index τ can be used as an prospective tool to explore alternative hybridization conditions. By changing these conditions alternative isothermal probe libraries can be designed which open the possibility of new technological pathways for methods based on the sequencing-by-hybridization concept.

Acknowledgments

This work was supported by Research Councils UK through the Basic Technology Programme and Fapemig/Brazil.

References

[1] Service R F 2006 Gene sequencing: the race for the $1000 genome *Science* **311** 1544–6
[2] Shendure J, Mitra R D and Church G M 2004 Advanced sequencing technologies: methods and goals *Nat. Rev. Gen.* **5** 335–44
[3] Chan E Y 2005 Advances in sequencing technology *Mutat. Res.* **573** 13–40
[4] Hacia J G 1999 Resequencing and mutational analysis using oligonucleotide microarrays *Nat. Genet.* **21** 42–7
[5] Berger M F, Philippakis A A, Qureshi A M, He F S, Estep P W III and Bulyk M L 2006 Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities *Nat. Biotechnol.* **24** 1429–35
[6] Fan J B, Gunderson K L, Bibikova M, Yeakley J M, Chen J, WickhamGarcia E, Lebruska L L, Laurent M, Shen R and Barker D 2006 Illumina universal bead arrays *Methods Enzymol.* **410** 57–73
[7] Charbonnier Y, Gettler B, Francois P, Bento M, Renzoni A, Vaudaux P, Schlegel W and Schrenzel J 2005 A generic approach for the design of whole-genome oligoarrays, validated for genomotyping, deletion mapping and gene expression analysis on *Staphylococcus aureus* *BMC Genomics* **6** 95
[8] Okoniewski M and Miller C 2006 Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations *BMC Bioinformatics* **7** 276
[9] Breslauer K J, Frank R, Blocker H and Marky L A 1986 Predicting DNA duplex stability from the base sequence *Proc. Natl Acad. Sci. USA* **83** 3746–50
[10] SantaLucia J Jr, Allawi HT and Seneviratne PA 1996 Improved nearest-neighbour parameters for predicting DNA duplex stability *Biochemistry* **35** 3555–62
[11] Owczarzy R, Vallone P M, Gallo F J, Paner T M, Lane M J and Benight A S 1998 Predicting sequence dependent melting stability of short duplex DNA oligomers *Biopolymers* **44** 217–39
[12] Owczarzy R, You Y, Moreira B G, Manthey J A, Huang L, Behlke M A and Walder J A 2004 Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures *Biochemistry* **43** 3537–54
[13] Dimitrov R A and Zuker M 2004 Prediction of hybridization and melting for double-stranded nucleic acids *Biophys. J.* **87** 215–26

[14] Peyrard M and Bishop A R 1989 Statistical mechanics of a nonlinear model for DNA denaturation *Phys. Rev. Lett.* **62** 2755–7
[15] Alexandrov B S, Wille L T, Rasmussen K O, Bishop A R and Blagoev K B 2006 Bubble statistics and dynamics in double-stranded DNA *Phys. Rev.* E **74** 050901
[16] Theodorakopoulos N 2008 DNA denaturation bubbles at criticality *Phys. Rev.* E **77** 031919
[17] DeLuca J, Drigo Filho E, Ponno A and Ruggiero J R 2004 Energy localization in the Peyrard–Bishop DNA model *Phys. Rev.* E **70** 026213
[18] Zdravkovic S and Sataric M V 2008 Solitonic speed in DNA *Phys. Rev.* E **77** 031906
[19] Weber G, Haslam N, Whiteford N, Prügel-Bennett A, Essex J W and Neylon C 2006 Thermal equivalence of DNA duplexes without melting temperature calculation *Nat. Phys.* **2** 55–9
[20] Weber G 2006 Sharp DNA denaturation due to solvent interaction *Europhys. Lett.* **73** 806–11
[21] Zhang Y-L, Zheng W-M, Liu J-X and Chen Y Z 1997 Theory of DNA melting based on the Peyrard–Bishop model *Phys. Rev.* E **56** 7100–15
[22] Cuesta J A and Sánches A 2004 General non-existence theorem for phase transitions in one-dimensional systems with short-range interactions, and physical examples of such transitions *J. Stat. Phys.* **115** 869–93
[23] Dauxois T, Peyrard M and Bishop A R 1993 Entropy-driven DNA denaturation *Phys. Rev.* E **47** R44–7
[24] Dauxois T and Peyrard M 1995 Entropy-driven transition in a one-dimensional system *Phys. Rev.* E **51** 4027–40
[25] Joyeux M and Buyukdagli S 2005 Dynamical model based on finite stacking enthalpies for homogeneous and inhomogeneous DNA thermal denaturation *Phys. Rev.* E **72** 051902
[26] Saccomandi G and Sgura I 2006 The relevance of nonlinear stacking interactions in simple models of double-stranded DNA *J. R. Soc. Interface* **3** 655–67
[27] Drukker K, Wu G and Schatz G C 2001 Model simulations of DNA denaturation dynamics *J. Chem. Phys.* **114** 579–90
[28] Theodorakopoulos N, Dauxois T and Peyrard M 2000 Order of the phase transition in models of DNA thermal denaturation *Phys. Rev. Lett.* **85** 6–9
[29] Barbi M, Lepri S, Peyrard M and Theodorakopoulos N 2003 Thermal denaturation of a helicoidal DNA model *Phys. Rev.* E **68** 061909
[30] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1988 *Numerical Recipes in C* (Cambridge: Cambridge University Press)
[31] Campa A and Giansanti A 1998 Experimental tests of the Peyard–Bishop model applied to the melting of very short DNA chains *Phys. Rev.* E **58** 3585
[32] SantaLucia J Jr 1998 A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics *Proc. Natl Acad. Sci. USA* **95** 1460–5
[33] Wu P, Nakano S and Sugimoto N 2002 Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation *Eur. J. Biochem.* **269** 2821–30
[34] Choi C H, Kalosakas G, Rasmussen K O, Hiromura M, Bishop A R and Usheva A 2004 DNA dynamically directs its own transcription initiation *Nucleic Acids Res.* **32** 1584–90
[35] Kalosakas G, Rasmussen K O, Bishop A R, Choi C H and Usheva A 2004 Sequence-specific thermal fluctuations identify start sites for DNA transcription *Europhys. Lett.* **68** 127133
[36] van Erp T S, Cuesta-Lopez S, Hagmann J-G and Peyrard M 2005 Can one predict DNA transcription start sites by studying bubbles? *Phys. Rev. Lett.* **95** 218104
[37] Choi C H, Usheva A, Kalosakas G, Rasmussen K Ø and Bishop A R 2006 Comment on Can one predict DNA transcription start sites by studying bubbles? *Phys. Rev. Lett.* **96** 239801
[38] van Erp T S *et al* 2006 *Phys. Rev. Lett.* **96** 239802 (reply)
[39] Błazewicz J and Kasprzak M 2003 Complexity of DNA sequencing by hybridization *Theor. Comput. Sci.* **290** 1459–73
[40] Mockler T C and Ecker J R 2005 Applications of DNA tiling arrays for whole-genome analysis *Genomics* **85** 1–15
[41] Kozal M J, Shah N, Shen N, Yang R, Fucini R, Merigan T C, Richman D D, Morris D, Hubbell E, Chee M and Gingeras T R 1996 Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays *Nat. Med.* **2** 753–9
[42] Favis R, Day J P, Gerry N P, Phelan C, Narod S and Barany F 2000 Universal DNA array detection of small insertions and deletions in BRCA1 and BRCA2 *Nat. Biotechnol.* **18** 561–4
[43] Zhang M, Gong Y, Osiowy C and Minuk G Y 2002 Rapid detection of hepatitis B virus mutations using real-time PCR and melting curve analysis *Hepatology* **36** 723–8
[44] Wong C W, Albert T J, Vega V B, Norton J E, Cutler D J, Richmond T A, Stanton L W, Liu E T and Miller L D 2004 Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays *Genome Res.* **14** 398–405
[45] Haslam N, Whiteford N, Weber G, Prügel-Bennett A, Essex J W and Neylon C 2008 Optimal probe length varies for targets with high sequence variation: implications for probe library design for resequencing highly variable genes *PLoS ONE* **3** e2500
[46] Baumann C G, Smith S B, Bloomfield V A and Bustamante C 1997 Ionic effects on the elasticity of single DNA molecules *Proc. Natl Acad. Sci. USA* **94** 6185–90
[47] Wenner J R, Williams M C, Rouzina I and Bloomfield V A 2002 Salt dependence of the elasticity and overstretching transition of single DNA molecules *Biophys. J.* **82** 3160–9
[48] Harris S A, Laughton C A and Liverpool T B 2008 Mapping the phase diagram of the writhe of DNA nanocircles using atomistic molecular dynamics simulations *Nucleic Acids Res.* **36** 2129
[49] Rief M, Clausen-Schaumann H and Gaub H E 1999 Sequence-dependent mechanics of single DNA molecules *Nat. Struct. Biol.* **6** 346–9
[50] Dong R, Yan X and Liu S 2004 The salt dependence of the stretching transition of double-stranded DNA molecules *J. Phys.: Condens. Matter* **37** 4977–84
[51] Bonner G and Klibanov A M 2000 Structural stability of DNA in nonaqueous solvents *Biotechnol. Bioeng.* **68** 339–44
[52] Hammouda B and Worcester D 2006 The denaturation transition of DNA in mixed solvents *Biophys. J.* **91** 2237–42
[53] Markarian S A, Asatryan A M, Grigoryan K R and Sargsyan H R 2006 Effect of diethylsulfoxide on the thermal denaturation of DNA *Biopolymers* **82** 1–5