Bioinformatics, 31(6), 2015, 871–877
doi: 10.1093/bioinformatics/btu751
Advance Access Publication Date: 12 November 2014
Original Paper

Structural bioinformatics

Optimization method for obtaining nearest-neighbour DNA entropies and enthalpies directly from melting temperatures

Gerald Weber

Departamento de Física, Universidade Federal Minas Gerais, Belo Horizonte-MG, Brazil

Associate Editor: Anna Tramontano

Received on July 28, 2014; revised on October 23, 2014; accepted on November 9, 2014

Abstract

Motivation: Free energy nearest-neighbour (NN) thermodynamics is widely used in DNA biochemistry, ranging from the calculation of melting temperatures to the prediction of secondary structures. Methods to calculate NN parameters require the knowledge of total sequence entropies and enthalpies, which are not always available.

Results: Here, we implement and test a new melting temperature optimization method where we obtain the NN parameters directly from the temperatures. In this way, we bypass the constraints imposed by total sequence entropies and enthalpies. This enabled us to calculate the missing NN entropies and enthalpies for some published datasets, including salt-dependent parameters. Also this allowed us to combine 281 sequences from different types of melting temperature data for which we derived a new set of NN parameters, which have a smaller uncertainty and an improved predictive power.

Availability and implementation: C++ source code and compiled binaries for several Linux distributions are available from https://sites.google.com/site/geraldweberufmg/vargibbs and from OpenSuse build service at https://build.opensuse.org/package/show/home:drgweber/VarGibbs. The software package contains scripts and data files to reproduce all results presented here.

Contact: gweberbh@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The concept of dividing a DNA or RNA sequence into pieces of nearest-neighbour (NN) pairs for predicting melting temperatures (T_m) was pioneered 40 years ago by Borer *et al*. (1974). Breslauer *et al*. (1986) introduced much of the concepts and terminologies that are in use today. A more detailed analysis of the NN representation was given by Gray (1997) and later by Licinio and Guerra (2007). Its simplicity of use and effectiveness has made it into a nearly universal method for applications in oligonucleotide chemistry and molecular biology.

With this technique, predictions of thermodynamic properties for any unknown sequence can be made by summing over NN entropies and enthalpies (ΔP_NN with $P = H, S$). The method which is currently in use for extracting ΔP_NN is essentially based on writing a set of linear equations where each equation represents the total entropy or the total enthalpy variations (ΔP^i_Tot) of a given sequence i. For the remainder of this article, we call this the total entropy and enthalpy variation (TEEV) method. Therefore, it is necessary to first obtain the total ΔP^i_Tot. This can be done either by curve fitting the melting curve or from a linear regression of T_m^{-1} versus $\log C_t$ in which case one needs a number of measurements at different species concentrations C_t. Unfortunately, both curve-fitting or $T_\mathrm{m}^{-1} \times \log C_t$ methods introduce a certain amount of uncertainty into the resulting parameters ΔP_NN. Furthermore, the total ΔP^i_Tot may, for various reasons, not be available at all.

Here, we propose a new melting temperature optimization (MTO) method, which does not require the total parameters ΔP^i_Tot. Instead, we simply fit the ΔP_NN, which best describe the melting

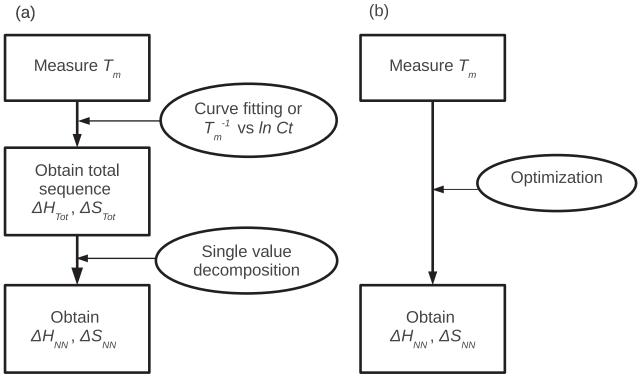
(a)
Measure T_m
Curve fitting or T_m^{-1} vs *ln Ct*
Obtain total sequence ΔH_{Tot}, ΔS_{Tot}
Single value decomposition
Obtain ΔH_{NN}, ΔS_{NN}
(b)
Measure T_m
Optimization
Obtain ΔH_{NN}, ΔS_{NN}

Fig. 1. Schematic workflow for obtaining NN parameters from melting temperatures. (**a**) The usual approach involving total entropies and enthalpies values (TEEV) and (**b**) the method based solely on MTO

temperatures T_m. This direct approach has two advantages over the TEEV method. One is that we fit the much more accurate melting temperatures T_m. The other is that this method can be applied to situations where the total parameters ΔP^i_Tot are unavailable. For example, one can work out the NN parameters from the melting temperatures, which were measured only for a single species concentration for each DNA sequence. Figure 1 summarizes the differences between the methods. Another interesting feature of the model is that the initialization parameters can be obtained in the framework of the MTO.

Numerically, the MTO minimizes the difference between measured and predicted melting temperatures by varying the ΔP_NN parameters freely. Here, we applied the MTO method successfully to the dataset by SantaLucia (1998) and showed that the resulting predictive uncertainty is significantly reduced. We used a similar method for the non-linear Peyrard-Bishop model (Weber *et al*. 2009, 2013a, b). The good results obtained for the non-linear model were one of our main motivations to investigate the applications of the MTO method to the linear NN model.

One key aspect of obtaining the free energy parameters ΔP_NN is the experimental set of melting temperatures. Unsurprisingly, different sets are bound to result in different parameters. Here, we analysed the two sets by SantaLucia *et al*. (1996) and SantaLucia (1998). However, as total sequence entropies and enthalpies are not required, we could also analyse the datasets by Owczarzy *et al*. (2004) (D-OW04). We discuss the significant differences resulting from those datasets, especially regarding initiation parameters. Finally, we combined the datasets D-SL98 and D-OW04 into a new set for which we obtain the best overall melting temperature prediction.

2 Methods

2.1 Data

In this work, we are using five different datasets to test the optimization method. Table 1 presents a summary of these datasets and the notation adopted throughout this work. The D-SL96 includes the non-two-state sequences, otherwise we would not have enough sequences for retrieving all initiation parameters. The D-SL98 dataset was used by SantaLucia (1998) to retrieve NN parameters (P-SL98), which is widely used for temperature prediction. Note that the actual sequences and temperatures were in fact published in the Supplementary Tables of Allawi and SantaLucia (1997), but it has become common practice to refer to them as coming from SantaLucia (1998). D-OW04 are the melting temperatures obtained from ultra-violet absorption measurements at five different salt concentrations by Owczarzy *et al*. (2004).

Table 1. Summary of melting temperature datasets

Dataset code	Reference	[Na$^+$] (mM)	N
D-SL96	SantaLucia *et al*. (1996)	1000	61
D-SL98	SantaLucia (1998); Allawi and SantaLucia (1997)	1000	108
D-OW04	Ultra-violet data from Owczarzy *et al*. (2004)	69–1020	92
D-OW04DSC	DSC data from Owczarzy *et al*. (2004)	69 and 1000	92
D-CMB	Combined data D-SL98, D-OW04 and D-OW04DSC	1000	281

N is the number of sequences provided in each dataset.

The D-OW04DSC dataset comprised the same DNA sequences as in D-OW04 but measured by differential scanning calorimetry (DSC) at low and high salt concentrations. All sequences in D-OW04 and D-OW04DSC are non-self-complementary. D-CMB is the combination of all data from D-SL98, D-OW04 and D-OW04DSC at high salt concentration.

2.2 Temperature calculation

The melting temperature for the *i*th sequence is calculated from

$$T_i' = \frac{\Delta H_i^{\mathrm{tot}}}{\Delta S_i^{\mathrm{tot}} + R\ln\frac{C_t}{fC_0}} \qquad (1)$$

where $f = 1$ for self-complementary sequences and $f = 4$ otherwise. C_t is the species concentration and $C_0 = 1\ \mu\mathrm{M}$ is the reference concentration introduced to insure that the argument to the logarithm is dimensionless. $\Delta H_i^{\mathrm{tot}}$ and $\Delta S_i^{\mathrm{tot}}$ are the total entropy and enthalpy variations for sequence *i*.

2.3 NN models

The various NN models found in the literature differ mainly by the introduction of initiation factors, which try to take into account symmetry and several other specific aspects of nucleotide sequences. The optimization method presented here can be used with any of these methods.

The basic model is simply the summation of all NN parameters *p*

$$\Delta p_i^{\mathrm{tot}} = \sum_\alpha n_{i\alpha}\Delta p_\alpha + \sum_c n_i^c \Delta p^c \qquad (2)$$

where Δp_α is either the enthalpy ($p = H$) or the entropy ($p = S$) variation for $n_{i\alpha}$ occurrences of NN types α in the *i*th sequence. Additional correction factors Δp^c may be added, such as

Table 2. Summary of initiation parameters (iP)

Model	iP-SL96	iP-SL98
ΔS^{symm}	−1.4	−1.4
$\Delta S^{\text{init}}_{\text{CG}}$	−5.9	—
$\Delta S^{\text{init}}_{\text{AT only}}$	−9.0	—
$\Delta H^{\text{term}}_{5'\text{AT}}$	0.4	—
$\Delta H^{\text{term}}_{\text{AT}}$	—	2.3
$\Delta S^{\text{term}}_{\text{AT}}$	—	4.1
$\Delta H^{\text{term}}_{\text{CG}}$	—	0.1
$\Delta S^{\text{term}}_{\text{CG}}$	—	−2.8

iP-SL96 are from SantaLucia *et al.* (1996) and iP-SL98 from SantaLucia (1998).

- Δp^{symm} for a self-complementary sequence
- $\Delta p^{\text{init}}_{\text{CG}}$ if the sequence contains a CG base pair
- $\Delta p^{\text{init}}_{\text{AT only}}$ if the sequence contains solely AT base pairs
- $\Delta p^{\text{term}}_{\text{AT}}$ for each terminal AT base pair
- $\Delta p^{\text{term}}_{\text{CG}}$ for each terminal CG base pair
- $\Delta p^{\text{term}}_{5'\text{AT}}$ for an AT base pair at the 5′-end

Table 2 summarizes the correction factors used in various models.

2.4 Local minimization

The aim of the method is to minimize

$$\chi^2 = \sum_i \left[T_i - T'_i(\{p\}_k)\right]^2 \tag{3}$$

where T_i is the measured melting temperature for the ith sequence. T'_i is the corresponding predicted temperature resulting from the set of tentative parameters

$$\{p\}_k = \{p_{k1}, p_{k2}, p_{k3}, \ldots\}. \tag{4}$$

We start with an initial set of parameters $\{p\}_0$ and let the parameters vary until we minimize as much as possible χ^2. After L steps of minimizations, the resulting set $\{p\}_L$ represents the parameters which best reproduce the melting temperatures for the model.

To minimize simultaneously all parameters, we used the multidimensional Nelder–Mead or downhill simplex method (Press *et al.*, 1988) in a similar way as for the non-linear models (Weber *et al.*, 2009). Each local minimization was carried as follows:

1. Define a set of initial parameters $\{p\}_0$.
2. Carry out the minimization with a characteristic length which should be roughly of the order of magnitude of the parameters (Press *et al.*, 1988). Here, we used a value of 10 for the characteristic length.
3. Stop the minimization either when the variation of χ^2 from one round to the next falls below 0.01 or when a limit of $L = 3000$ steps is reached.
4. Use the newly obtained parameters $\{p\}$ as new initial parameters and repeat the whole procedure five more times.

2.4.1 Local minimization control parameters

In addition to Equation (3), the following parameters are used to assess the quality of the minimization:

Average prediction difference:

$$\langle \Delta T \rangle = \frac{1}{N}\sum_{i=1}^{N} |T_i - T'_i(\{p\}_k)|. \tag{5}$$

This is a simple parameter that allows a more direct comparison of the prediction quality of the results to those by other authors.

Root mean square of prediction differences is closely related to χ^2,

$$\Delta T_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[T_i - T'_i(\{p\}_k)\right]^2} = \sqrt{\frac{\chi^2}{N}}, \tag{6}$$

and is useful for evaluating the relative dispersion of predicted temperatures $T_i{}'$.

2.5 Global minimization

The drawback of the minimization of Equation (3) is that most of the time we will find only local minima. This is especially problematic for minimizations involving experimental data because local minima may be result of noise. One way to overcome this problem is to repeat the local minimization many times with different initial parameters $\{p\}_0$. We will call this procedure global minimization, i.e. the attempt to find the local minima which is closest to the global minimum. In summary, a global minimization is M repetitions of local minimizations by varying certain conditions randomly each time.

2.5.1 Global minimization control parameters

Average parameter

$$\langle p_k \rangle_M = \frac{1}{M}\sum_{i=1}^{M} p_{ki} \tag{7}$$

where p_k is the kth parameter in the set, and M is the total number of complete local minimizations performed.

Standard parameter deviation

$$\sigma(p_k)_M = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(p_{ki} - \langle p_k \rangle)^2} \tag{8}$$

which is used to estimate the uncertainty of each parameter p_k for the melting temperature predictions after M rounds of local minimizations.

Relative average parameter difference

$$\alpha_{k,M} = \left|\frac{\langle p_k \rangle_M - \langle p_k \rangle_{M-1}}{\langle p_k \rangle_{M-1}}\right| \tag{9}$$

which measures how much the average parameter $\langle p_k \rangle_M$ is changing in round M compared with the previous round $M - 1$. One possible stop criterion for the global minimization is to calculate the average of $\alpha_{k,M}$ over all parameters k parameters,

$$\langle \alpha_M \rangle = \sum_k \alpha_{k,M} \tag{10}$$

and set a threshold α_t, such that the minimization stops when

$$\alpha_t < \langle \alpha_M \rangle \tag{11}$$

is satisfied.

Relative standard deviation difference

$$\beta_{k,M} = \left|\frac{\sigma(p_k)_M - \sigma(p_k)_{M-1}}{\sigma(p_k)_{M-1}}\right| \tag{12}$$

Similar to the relative average parameter difference, this can be used as a much stricter criterion for stopping the minimization process, by requiring

$$\beta_t < \langle \beta_M \rangle = \sum_k \beta_{k,M} \tag{13}$$

2.5.2 Global minimization work flow

We carry out the global minimization in two different ways. First we vary randomly the initial parameters $\{p\}_0$ over a large range of values and perform a new local minimization. After a certain number M_1 of local minimizations we keep the set of parameters $\{p\}_k^{\min \chi^2}$ with the smallest χ^2. Here, we repeat the local minimizations 1000 times ($M_1 = 1000$).

For the second block of global minimizations, we take the best parameters from the first block and use them as fixed initial parameters ($\{p\}_0 = \{p\}_k^{\min \chi^2}$), but now we vary randomly the experimental temperatures. We create a new set of modified temperatures $\{T\}_l{}'$, which differ by random amounts from the original set $\{T\}$ in such a way that on average the standard deviation between the two sets is close to the declared experimental uncertainty of 0.3°C. We repeat this procedure using $\beta_t = 0.01$ as stop criterion, see Equation (13), resulting in a variable number M_2 of new local minimizations. From these M_2 results, we calculate the average parameters, which are presented here.

2.6 Parameter correlation

To compare different parameters sets $\{p\}$ and $\{q\}$, we use the Pearson's sample correlation coefficient,

$$r_{p,q} = \frac{\sum_i (p_i - \langle p \rangle) \sum_i (q_i - \langle q \rangle)}{\left[\sum_i (p_i - \langle p \rangle)^2 \sum_i (q_i - \langle q \rangle)^2\right]^{1/2}}. \tag{14}$$

In this work, we will provide the correlation coefficient separately for entropies (r^S) and enthalpies (r^H) and only for NN parameters. Initiation parameters are not included in the parameter correlation. Please note that this coefficient is not part of the MTO method and is used only as a guide to aid the comparison between different parameter sets.

2.7 Summary of notation

DNA sequence and melting temperatures datasets are prefixed by D-, published parameters by P-, initiation parameters by iP- and averaged optimized parameters by AOP-.

3 Results and discussion

The most problematic aspect of the MTO method presented here is the occurrence of local minima which are likely to occur when adjusting experimental data and are mainly caused by fluctuation due to experimental uncertainties. To overcome this problem, our method is applied in two distinct steps. First, we repeat the minimization many times where we vary the initial parameters by a very large amount. From this step, we obtain the parameters with the smallest disagreement from the experimental temperatures. In the second step, we use the previous parameters as input and now vary the experimental temperatures within their declared experimental uncertainty. This last step provides us with an estimate of parameter uncertainty resulting from experimental fluctuations. See section 2.5.2 for details.

We consider here two types of predictions: self-predictions and cross-predictions. Self-predictions are those where the parameters were derived from the same dataset that is being predicted. For instance, predicting the melting temperatures of the sequences in D-SL98 using parameters AOP-SL98. Cross-predictions are predictions where the dataset is different from the one used to obtain the parameters.

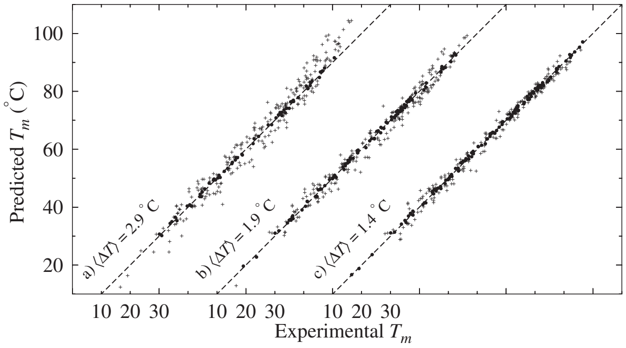

Fig. 2. Scatter plot of predicted melting temperatures compared with the experimental temperatures for dataset D-CMB at 1000 mM [Na^+]. The cross-predictions using the (**a**) original P-SL96, (**b**) the original P-SL98 and (**c**) the self-prediction for optimized AOP-CMB parameters are shown. The closer the points are to the diagonal dashed lines, the better the prediction quality. Predictions are shown as bullets if they are within 1°C of the experimental points, otherwise they are shown as crosses. Data points are horizontally shifted to the right for clarity, and the resulting average temperature deviation $\langle \Delta T \rangle$ is calculated from Equation (5)

Figure 2 shows some examples of melting temperature predictions as scatter plots for the combined dataset D-CMB. The predictions obtained from the older parameter set P-SL96 (SantaLucia *et al.*, 1996), Figure 2a, is one of the poorest. The frequently used set P-SL98 (SantaLucia, 1998) is clearly a significant improvement as shown in Figure 2b but the prediction, especially at higher temperatures, still suffers. Figure 2c shows the self-predictions from the optimized parameter set AOP-CMB with very little dispersion along the diagonal line and a very low average temperature deviation $\langle \Delta T \rangle$. This clearly shows that the MTO method minimizes Equation (3) much more uniformly for all temperatures. Note that it would be very difficult, if not impossible, to treat such a combined set with the usual single value decomposition method. The reason is that for the TEEV method one needs the total entropies and enthalpies of all sequences, which are not available for the D-OW04 and D-OW04DSC datasets.

Evidently, Figure 2 only shows how the MTO successfully minimizes the difference in temperatures between datasets and predictions when compared with other methods. A much more complete overview is presented in Table 3 where we show a summary of all predictions performed in this work. For all optimizations, the average temperature deviation $\langle \Delta T \rangle$ and its ΔT_{RMS} are considerably reduced for self-predictions. Better self-prediction, however, typically results into poorer cross-predictions. For example, the optimized AOP-SL98 results into much better self prediction that the original P-SL98, reducing $\langle \Delta T \rangle$ from 1.9°C to 1.5°C. But for cross-predictions, AOP-SL98 is generally poorer than SL98. The optimized parameters AOP-OW04 and AOP-OW04DSC are exceptionally good at self-predicting melting temperatures but do perform poorly otherwise. This is true even when using AOP-OW04 to predict D-OW04DSC and vice versa, even though both use exactly the same sequences.

With the choice of dataset playing such a crucial choice, we combined the D-SL98, D-OW04 and D-OW04DSC into a single set D-CMB. Once again the self-prediction is very good, at 1.4°C, and all its cross-predictions are second-best if compared with the

Table 3. Summary of temperature predictions

Dataset	P-SL96	AOP-SL96	P-SL98	AOP-SL98	AOP-OW04	AOP-OW04DSC	AOP-CMB
D-SL96	2.7 (4.0)	**2.2** (**2.9**)	*2.4* (3.7)	2.6 (3.7)	4.1 (6.0)	8.6 (16.)	2.4 (*3.5*)
D-SL98	2.5 (3.2)	2.0 (2.6)	1.9 (2.5)	**1.5** (**1.9**)	2.8 (4.0)	6.9 (14.)	*1.7* (2.2)
D-OW04	3.3 (4.2)	3.3 (3.7)	1.6 (2.1)	1.6 (2.0)	**0.57** (**0.73**)	2.5 (2.8)	*1.2* (*1.5*)
D-OW04DSC	2.9 (3.6)	3.5 (3.7)	2.1 (2.5)	4.2 (4.6)	2.8 (2.9)	**0.52** (**0.64**)	*1.3* (*1.5*)
D-CMB	2.9 (3.7)	2.9 (3.3)	*1.9* (*2.4*)	2.3 (3.0)	2.1 (2.9)	3.6 (8.6)	**1.4** (**1.8**)

The resulting $\langle \Delta T \rangle$ and ΔT_{RMS} (in brackets) in °C for datasets at 1000 mM (1020 mM for D-OW04) when predicted with published parameters (P-) or averaged optimized parameters (AOP-) are listed. Best results are indicated in boldface and second best results are in italics.

Table 4. Summary of ranking coefficients for NN parameter sets $\{\Delta H\}$ and $\{\Delta S\}$

	P-SL96	AOP-SL96	P-SL98	AOP-SL98	AOP-OW04	AOP-OW04DSC
AOP-SL96	0.33, 0.26	—				
P-SL98	0.87, 0.80	0.20, −0.06	—			
AOP-SL98	0.66, 0.52	−0.04, −0.13	0.63, 0.47	—		
AOP-OW04	0, −0.13	−0.76, −0.83	0.27, 0.28	0.33, 0.29	—	
AOP-OW04DSC	0.54, 0.46	−0.54, −0.64	0.53, 0.54	0.62, 0.58	0.55, 0.54	—
AOP-CMB	0.71, 0.50	0.40, 0.35	0.84, 0.59	0.70, 0.62	0.21, 0.13	0.22, 0.06

The resulting sample correlation coefficients r^{H} (first number) and r^{S} (second number) are listed, calculated from Equation (14). Initiation factors are not included for calculating sample correlations.

self-predictions of other parameters, Table 3. Compared with the widely used P-SL98 parameters, AOP-CMB reduces $\langle \Delta T \rangle$ by 0.5°C and ΔT_{RMS} by 0.6°C for the D-CMB dataset.

The TEEV method aims at getting close to the total entropy and enthalpy variations, ΔH_{Tot} and ΔS_{Tot}, for all sequences. Nevertheless, the difference between the self-predicted and experimental total parameters for the P-SL98 TEEV parameters is rather large, 7.2% and 8.2% for ΔH_{Tot} and ΔS_{Tot}, respectively. For the MTO parameters AOP-CMB, there is an increase to 12% and 14% when predicting the total parameters ΔH_{Tot} and ΔS_{Tot} of D-SL98.

With the MTO method being clearly able to produce acceptable self-predictions comes the question of how the MTO parameters compare to those obtained from TEEV method. To assess this unambiguously, we calculated the correlation between the parameters using Equation (14), where a good correlation should yield a number close to 1. Table 4 presents correlations comparing all parameters analysed in this work. There is mostly very little correlation between optimized dataset, which explains the generally poor cross-prediction. For the AOP-CMB optimized parameters, there is some correlation for the $\{\Delta H\}$ with the original P-SL98 and a moderate correlation when it comes to $\{\Delta S\}$. Note that initiation parameters were not included in the correlation analysis and that the correlation coefficients may not be a reliable indicator of similarity if non-unique parameters are included.

The complete set of parameters for all optimizations at high salt concentrations is listed in Table 5. In general, the enthalpies have a similar behaviour as for the TEEV parameters (P-SL96 and P-SL98). For entropies, however, important changes are observed with their values covering a much wider range. These wide ranges of entropies are accompanied by large positive variations in initiation factors, in particular for the AOP-OW04 and AOP-OW04DSC with some of them becoming as high as $\langle \Delta S_{\mathrm{AT}}^{\mathrm{term}} \rangle = 35$ cal/mol. One crucial difference from D-OW04(DSC) to all other datasets is the absence of self-complementary sequences. This may explain the important parameter variations, especially for entropies seen for the AOP-OW04(DSC) parameters. The enthalpies of AOP-CMB are clearly closer to the original P-SL98, as already noted from the correlation factors. As the total parameters $\Delta H_i^{\mathrm{tot}}$ and $\Delta S_i^{\mathrm{tot}}$ in P-SL98 are essentially averaging out differences in melting temperatures from various species concentrations C_t, it would seem that a similar averaging effect is occurring for AOP-CMB.

Good self-prediction with poor cross-prediction may sometimes indicate over-fitting, i.e. fitting to noise. Could this be the case for the AOP-OW04(DSC) parameters? The second block of global minimizations, see section 2.5.2, is precisely to assess this type of situation. As we randomly vary the temperatures within the experimental uncertainty, we effectively change the position of the experimental noise. If over-fitting was occurring, we should have seen large parameter uncertainties; however, this not the case as shown in Table 5.

As the datasets D-OW04 and D-OW04DSC are measured at various salt concentrations (Table 1), we can now calculate the parameters as function of [Na$^+$]. Figure 3 shows the parameter dependence with salt concentration for datasets D-OW04 and D-OW04DSC. Note that D-OW04DSC is measured only at the lowest and highest salt concentrations. Only ApT, TpA and ApA (panels c and g) show similar trends for D-OW04 and D-OW04DSC. All remaining parameters display very dissimilar behaviours, and no particular conclusion can be drawn at this stage except that sequence context seems to play an important role for salt concentration corrections. On the other hand, the qualitative trends of enthalpies and entropies are nearly identical, possibly resulting from the form of Equation (1). If the influence of the term ln C_t/fC_0 is small then one would expect ΔH_i^{α} to become largely proportional ΔS_i^{α}. One word of caution is needed here, one should keep in mind that these are self-predictions and contain no self-complementary sequences. Therefore changes to the salt-dependent results should be expected as new datasets become available.

Ideally, large and diverse datasets are best for the MTO method as shown by the results of the AOP-CMB parameters. However, for many practical applications, the available melting temperature data may be limited to only a small collection of sequences, which brings the question of what are the minimal conditions for the MTO method to work properly? Although the non-linearity introduced by Equation (1) makes this difficult to answer precisely, a rule of thumb confirmed by numerical experiments is that there should be more sequences than variables. In general, ΔT_{RMS} should never turn out to be smaller than the experimental uncertainty. Therefore, optimizations that return

Table 5. NN entropies and enthalpies at 1000 mM

Parameter	P-SL96	AOP-SL96	P-SL98	AOP-SL98	AOP-OW04	AOP-OW04DSC	AOP-CMB
$\langle \Delta H_{\mathrm{ApA}} \rangle$	−8.4	−7.2(1)	−7.9	−8.4(1)	−8.6(2)	−8.2(4)	−7.20(4)
$\langle \Delta H_{\mathrm{ApC}} \rangle$	−8.6	−10.4(1)	−8.4	−10.2(1)	−1.8(2)	−8.3(3)	−7.26(4)
$\langle \Delta H_{\mathrm{ApG}} \rangle$	−6.1	−6.8(2)	−7.8	−5.3(1)	−7.0(3)	−6.4(3)	−6.46(5)
$\langle \Delta H_{\mathrm{ApT}} \rangle$	−6.5	−9.5(2)	−7.2	−0.4(2)	−2.2(3)	−3.5(6)	−5.69(7)
$\langle \Delta H_{\mathrm{CpA}} \rangle$	−7.4	−3.2(1)	−8.5	−12.9(2)	−15.0(2)	−8.5(3)	−8.58(6)
$\langle \Delta H_{\mathrm{CpC}} \rangle$	−6.7	−7.49(8)	−8.0	−8.9(2)	−11.0(2)	−7.6(5)	−8.77(6)
$\langle \Delta H_{\mathrm{CpG}} \rangle$	−10	−5.5(1)	−11	−9.3(1)	−14.4(4)	−12.1(6)	−9.70(6)
$\langle \Delta H_{\mathrm{GpA}} \rangle$	−7.7	−2.77(9)	−8.2	−10.6(1)	−9.5(3)	−11.7(3)	−7.10(6)
$\langle \Delta H_{\mathrm{GpC}} \rangle$	−11	−9.1(1)	−9.8	−14.5(2)	−6.8(4)	−12.5(3)	−9.59(8)
$\langle \Delta H_{\mathrm{TpA}} \rangle$	−6.3	1.8(1)	−7.2	−6.1(2)	−13.8(3)	−12.9(5)	−4.08(7)
$\langle \Delta S_{\mathrm{ApA}} \rangle$	−24	−20.3(4)	−22	−23.8(3)	−24.3(6)	−23(1)	−20.2(1)
$\langle \Delta S_{\mathrm{ApC}} \rangle$	−23	−28.6(4)	−22	−28.3(5)	−3.2(7)	−22.2(8)	−19.3(1)
$\langle \Delta S_{\mathrm{ApG}} \rangle$	−16	−17.9(5)	−21	−13.0(4)	−18.5(8)	−16.8(7)	−17.1(2)
$\langle \Delta S_{\mathrm{ApT}} \rangle$	−−19	−26.6(5)	−20	0.9(5)	−5.6(9)	−10(2)	−15.9(2)
$\langle \Delta S_{\mathrm{CpA}} \rangle$	−19	−8.3(4)	−23	−36.6(6)	−42.0(7)	−23.0(8)	−23.2(2)
$\langle \Delta S_{\mathrm{CpC}} \rangle$	−16	−19.7(3)	−20	−22.1(5)	−29.3(7)	−19(1)	−22.7(2)
$\langle \Delta S_{\mathrm{CpG}} \rangle$	−26	−12.9(3)	−27	−22.9(4)	−39(1)	−31(2)	−24.9(2)
$\langle \Delta S_{\mathrm{GpA}} \rangle$	−20	−6.9(3)	−22	−29.4(5)	−26(1)	−32.0(9)	−18.9(2)
$\langle \Delta S_{\mathrm{GpC}} \rangle$	−28	−23.9(3)	−24	−38.9(7)	−16(1)	−32.4(8)	−24.4(2)
$\langle \Delta S_{\mathrm{TpA}} \rangle$	−19	4.6(5)	−21	−17.7(7)	−40.3(8)	−37(2)	−11.3(2)
$\langle \Delta S^{\mathrm{init}}_{\mathrm{AT}} \rangle$	−9.0	−6.0(2)	—	—	—	—	—
$\langle \Delta S^{\mathrm{init}}_{\mathrm{CG}} \rangle$	−5.9	−0.9(1)	—	—	—	—	—
$\langle \Delta H^{\mathrm{term}}_{5'\mathrm{AT}} \rangle$	0.40	−0.40(3)	—	—	—	—	—
$\langle \Delta S^{\mathrm{symm}} \rangle$	−1.4	0.67(6)	−1.4	−0.30(4)	—	—	−0.59(3)
$\langle \Delta H^{\mathrm{term}}_{\mathrm{CG}} \rangle$	—	—	0.10	0.7(1)	4.9(4)	10.6(7)	0.49(4)
$\langle \Delta S^{\mathrm{term}}_{\mathrm{CG}} \rangle$	—	—	−2.8	−1.5(3)	12(1)	29(2)	−0.7(1)
$\langle \Delta H^{\mathrm{term}}_{\mathrm{AT}} \rangle$	—	—	2.3	1.3(2)	6.5(3)	12.6(8)	2.53(7)
$\langle \Delta S^{\mathrm{term}}_{\mathrm{AT}} \rangle$	—	—	4.1	0.2(5)	17(1)	35(3)	5.7(2)

Estimated uncertainties, where available, are displayed in concise notation.

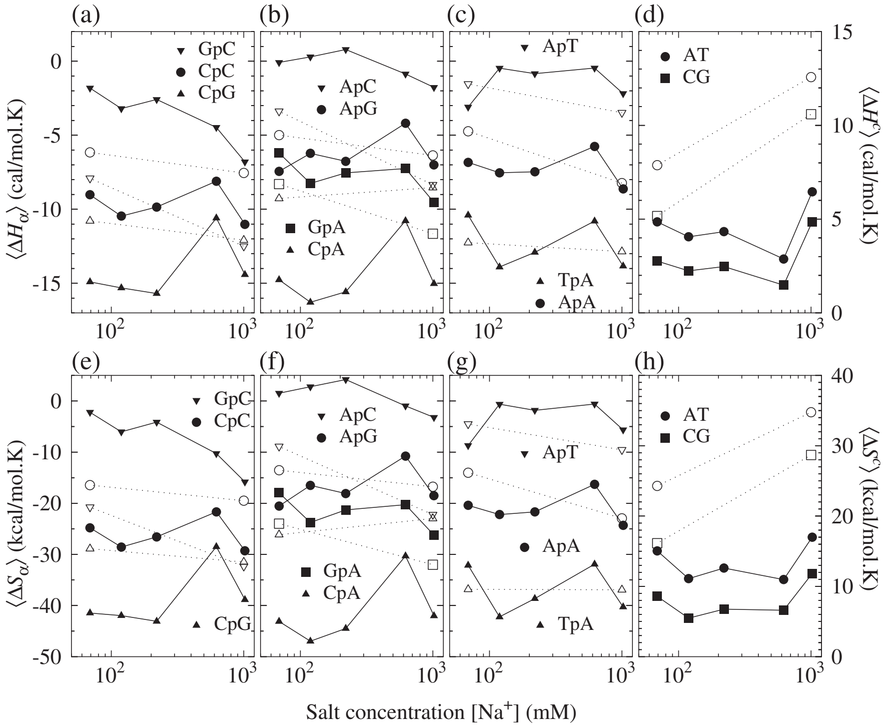

Fig. 3. Calculated NN parameters as function of salt concentration. All parameters were calculated independently for each salt concentration. Filled symbols connected by solids lines are for AOP-OW04 and open symbols connected by dotted lines are for AOP-OW04DSC optimizations. Upper panels (**a–d**) are for enthalpies and lower panels (**e–h**) for entropies. Panels (d) and (h) are for initiation parameters and use the right-hand scale, all other panels refer to the left-hand scale. Parameters are grouped according to the number of hydrogen bonds per NN pair: (a) and (e) with six bonds, (**b**) and (**f**) with five bonds and (**c**) and (**g**) with four bonds. NNs of type YpR are shown as up-pointing triangles, RpY as down-pointing triangles. RpR or YpY are shown either as bullets or boxes, except for panels (d) and (h), which are for base pairs

$\Delta T_{\mathrm{RMS}} \approx 0$ are a clear sign of insufficient number of sequences in the dataset. A practical advice in this case is to reduce the number of parameters to optimize, e.g. by keeping the initiation parameters as constants. When designing new sequences to be used with the MTO method, one should favour longer sequences with a diversity of NN configurations and with a large melting temperature range.

4 Conclusion

We presented a new method for the calculation of NN parameters from experimental melting temperatures. Because this method sidesteps the calculation of total entropies and enthalpies, it enables us to calculate the NN parameters under conditions, which cannot be handled by the traditional methods. It also allows us to combine datasets obtained under very dissimilar experimental conditions. The results of temperature predictions show a marked improvement, especially when larger datasets are used to calculate the parameters.

The software that implements this method is made freely available. A comprehensive user manual is part of the package and included as Supplementary Material. Therefore, the interested reader is provided with ability to calculate the thermodynamic parameters for any desired dataset. Also the availability of the source code allows for modifications for particular variants of the NN model, for instance to include nucleation terms (Oliveira Guerra, 2013).

Funding

This work was supported by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig); and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Conflict of Interest: none declared.

References

Allawi,H.T. and SantaLucia,J., Jr (1997) Thermodynamics and NMR of internal GT mismatches in DNA. *Biochemistry,* **36**, 10581–10594.

Borer,P.N. *et al*. (1974) Stability of ribonucleic acid double-stranded helices. *J. Mol. Biol.,* **86**, 843–853.

Breslauer,K.J. *et al*. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA,* **83**, 3746–3750.

Gray,D.M. (1997) Derivation of nearest-neighbor properties from data on nucleic acid oligomers. i. simple sets of independent sequences and the influence of absent nearest neighbors. *Biopolymers,* **42**, 783–793.

Licinio,P. and Guerra,J.C.O. (2007) Irreducible representation for nucleotide sequence physical properties and self-consistency of nearest-neighbor dimer sets. *Biophys. J.*, **92**, 2000–2006.

Oliveira Guerra,J.C. (2013) Calculation of nucleation free energy for duplex oligomers in the context of nearest neighbor models. *Biopolymers,* **99**, 538–547.

Owczarzy,R. *et al*. (2004) Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. *Biochemistry*, **43**, 3537–3554.

Press,W.H. *et al*. (1988) *Numerical Recipes in C.* Cambridge University Press, Cambridge.

SantaLucia,J., Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA,* **95**, 1460–1465.

SantaLucia,J., Jr *et al*. (1996) Improved nearest-neighbour parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.

Weber,G. (2013a) Mesoscopic model parametrization of hydrogen bonds and stacking interactions of RNA from melting temperatures. *Nucleic Acids Res.*, **41**, e30.

Weber,G. (2013b) TfReg: Calculating DNA and RNA melting temperatures and opening profiles with mesoscopic models. *Bioinformatics,* **29**, 1345–1347.

Weber,G. *et al*. (2009) Probing the microscopic flexibility of DNA from melting temperatures. *Nat. Phys.*, **5**, 769–773.