Genome wide application of DNA melting analysis

2009 J. Phys.: Condens. Matter 21 034108

(http://iopscience.iop.org/0953-8984/21/3/034108)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **21** (2009) 034108 (14pp) doi:10.1088/0953-8984/21/3/034108

Genome wide application of DNA melting analysis

Daniel Jost and Ralf Everaers

Laboratoire de Physique de l'École Normale Supérieure de Lyon, Université de Lyon, CNRS UMR 5672, 46 Allée d'Italie 69364 Lyon Cedex 07, France

E-mail: daniel.jost@ens-lyon.fr

Received 10 June 2008, in final form 24 September 2008
Published 17 December 2008
Online at stacks.iop.org/JPhysCM/21/034108

Abstract

Correspondences between functional and thermodynamic melting properties in a genome are being increasingly employed for *ab initio* gene finding and for the interpretation of the evolution of genomes. Here we present the first systematic genome wide comparison between biologically coding domains and thermodynamically stable regions. In particular, we develop statistical methods to estimate the reliability of the resulting predictions. Not surprisingly, we find that the success of the approach depends on the difference in GC content between the coding and the non-coding parts of the genome and on the percentage of coding base-pairs in the sequence. These prerequisites vary strongly between species, where we observe no systematic differences between eukaryotes and prokaryotes. We find a number of organisms in which the strong correlation of coding domains and thermodynamically stable regions allows us to identify putative exons or genes to complement existing approaches.

In contrast to previous investigations along these lines we have not employed the Poland–Scheraga (PS) model of DNA melting but use the earlier Zimm–Bragg (ZB) model. The Ising-like form of the ZB model can be viewed as an approximation to the PS model, with averaged loop entropies included into the cooperative factor $\sigma_{ZB} = \sigma_{PS}\bar{N}^{-c}$. This results in a speed-up by a factor of 20–100 compared to the Fixman–Freire algorithm for the solution of the PS model. We show that for genomic sequences the resulting systematic errors are negligible compared to the parameterization uncertainty of the models. We argue that for limited computing resources, available CPU power is better invested in broadening the statistical base for genomic investigations than in marginal improvements of the description of the physical melting behavior.

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

1. Introduction

Since the discovery of the principles underlying the storage and replication of the genetic information [1] and of the DNA nucleotides triplet code [2], extreme efforts have been made to determine [3, 4] and analyze [5] the genomes of organisms. Several techniques exist to identify the location of coding sequences or genes: extrinsic approaches [6, 7] using proteins with known amino-acid sequences, comparative methods [8] between genomes, or *ab initio* approaches [9–14] which try to predict biologically functional domains along the genome on the basis of sequence effects. Both genetic information [15, 16] and the physical properties [17–19] of DNA are correlated with the GC content. This leads to the question of whether there is a causal relation or if, for practical purposes, the former can be identified using the latter.

Since processes like replication and transcription require a partial opening of the double-helix, it is natural to consider the thermal denaturation of DNA in this context [13, 14, 20–33]. To exploit this idea, one has to choose (1) a source of information on the physical melting properties and (2) a method for inferring biologically relevant information from a physical melting profile. For example, in his investigations Yeramian [23, 24] calculated physical melting profiles from the Poland–Scheraga (PS) model of DNA melting [39] and identified coding sequences with regions of the genome which melt above a suitably chosen temperature.

In this paper we focus on improving the computational efficiency of the physical annotation of genomes and, in particular, on devising methods to quantify the reliability of the resulting predictions. Clearly, it is not enough to investigate specific examples chosen on the basis of previous knowledge. Rather a comparison needs to be done statistically, requiring the investigation of isochores [15], chromosomes, entire genomes and, possibly, inter-species comparisons. This need to broaden the statistical base explains our interest in judging the merits of different physical models of DNA melting in relation to their computational cost (section 2). In particular, we find that one of the oldest models of DNA melting, the Zimm–Bragg (ZB) model [34, 35], performs extremely well when compared to the Poland–Scheraga (PS) model [39–42] used by most previous investigations along this line. In section 3, we show that the systematic error due to ZB approximation of the PS model is *smaller* than the error due to the parameterization uncertainty of the PS model [44]. In particular, section 4 demonstrates that using the ZB model we fully reproduce the results of Yeramian and Jones [14] and Carlon and Blossey [25] on the physical annotation of genomic DNA. In section 5, we apply the ZB model to several complete genomes and study the correlation between melting and coding properties at the base-pair and the domain level. As a final step, we estimate the confidence level of the resulting identification of putative exons or genes (section 6). In section 7 we give a brief conclusion.

2. The Zimm–Bragg model

Since it is difficult to experimentally resolve the separation of individual base-pairs, the physical melting analysis is carried out using models of DNA denaturation [13, 14, 20–33]. Several models exist: the ZB model [34, 35], the nearest-neighbor (NN) model [36–38], the PS model [39–42, 44], a corresponding lattice model [45] and the independent Peyrard–Bishop–Dauxois (PBD) model [46, 47]. Mostly, these models view the denaturation process as the successive opening of domains in the sequence. As a consequence of the highly cooperative nature of DNA melting, they define precisely delimited regions which can be compared to biologically functional domains along the genome. All models approximate the local specificity of DNA by nearest-neighbor (pairing and stacking) interactions. Non-local (in terms of chemical not spatial distance) interactions are included to different degrees, using results or models describing generic polymer properties of DNA. The lattice model [45] takes into account both *intra*- and *inter*-loop excluded volume interactions. The PS model neglects [39] or approximates [56] the long-range inter-bubble excluded volume interactions. The ZB and PBD models also neglect the intra-loop excluded volume. Treating the PBD model as a case apart, the relation between parameters in different models is well understood [45, 44]. The values of the parameters were determined in extensive comparisons to experiment [42, 38] and we have recently shown how the remaining uncertainties in the parameterization of the PS model propagate to the level of the predicted melting profiles [44]. In the following, we concentrate on two models:

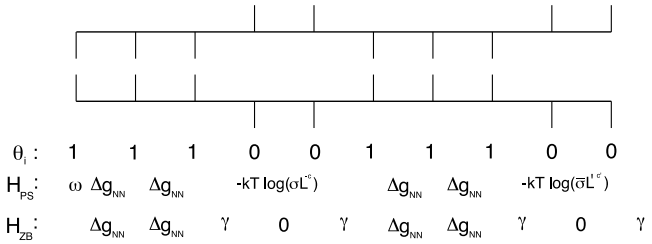

Figure 1. Example of 1D-Ising-like configuration and description of the free energy contributions in the Hamiltonian of the PS model (with end effects) and of the ZB model (with periodic boundary conditions).

the PS model used in most previous investigations along the present lines [13, 14, 23–25] and the computationally cheaper ZB model.

The PS model [39, 40] considers the denaturation of the double-helical complex as a sequence of cooperative and successive openings of internal domains. Double-stranded DNA is viewed at the secondary structure level and interactions are modeled as the free energy of a base-pair step formation (10 possible steps) $\Delta g_{\mathrm{NN}}(T) = \Delta h_{\mathrm{NN}} - T\Delta s_{\mathrm{NN}}$, a capping free energy $\omega(T)$, and an entropic loop factor σ (cooperativity) and entropic free-end factor $\bar{\sigma}$ [43, 44] (see figure 1). The contribution of an internal loop (size L) to the total partition function is σL^{-c}, the contribution of a free end (size L') is $\sigma' L'^{c'}$, while the contribution of a helical stem is $\exp(-\beta \sum \Delta g_{\mathrm{NN}})$ and the contribution of a closed end is $\exp(-\beta\omega)$. L^{-c} and $L'^{c'}$ account for polymeric effects of loops and free ends [43, 52]. Within the parameterization uncertainties, a recent version of the PS model [44] predicts the melting behavior of DNA strands with arbitrary strand length, strand concentration and ionic strength of the buffer solution. The typical uncertainty of the predicted local melting temperature is of the order of 2 K. Predictions for the domain structure are extremely robust (see figures 2(A) and (B)).

The PS model is solved using recursion relations for conditional and unconditional probabilities of base-pair openings [41, 48] or partition functions [49, 50, 43, 44]. The computation of the PS model could be accelerated using the Fixman–Freire algorithm [53].

The earliest model of DNA denaturation, the ZB model, has the same mathematical structure as the heterogeneous 1D Ising model [35, 21, 54, 55] and can be viewed as an approximation to the PS model. The loop-length-dependent cooperative entropic factor σL^{-c} is replaced by a constant contribution $\sigma \bar{L}^{-c} \equiv \exp(-2\gamma/(k_{\mathrm{B}}T))$ where $\bar{L}$ is a typical size and γ is a boundary (forking) free energy. At first glance, this approximation is very crude as it neglects generic consequences of the polymeric nature of DNA. In particular, for homogeneous DNA, the ZB model fails to predict the existence of a first-order melting transition at a finite temperature [56]. Sequence heterogeneity masks these effects to some extent [57]. Indeed, for natural sequences, the exponent c (from polymer theory) does not seem to influence the melting properties if the value of σ is modified appropriately [52].

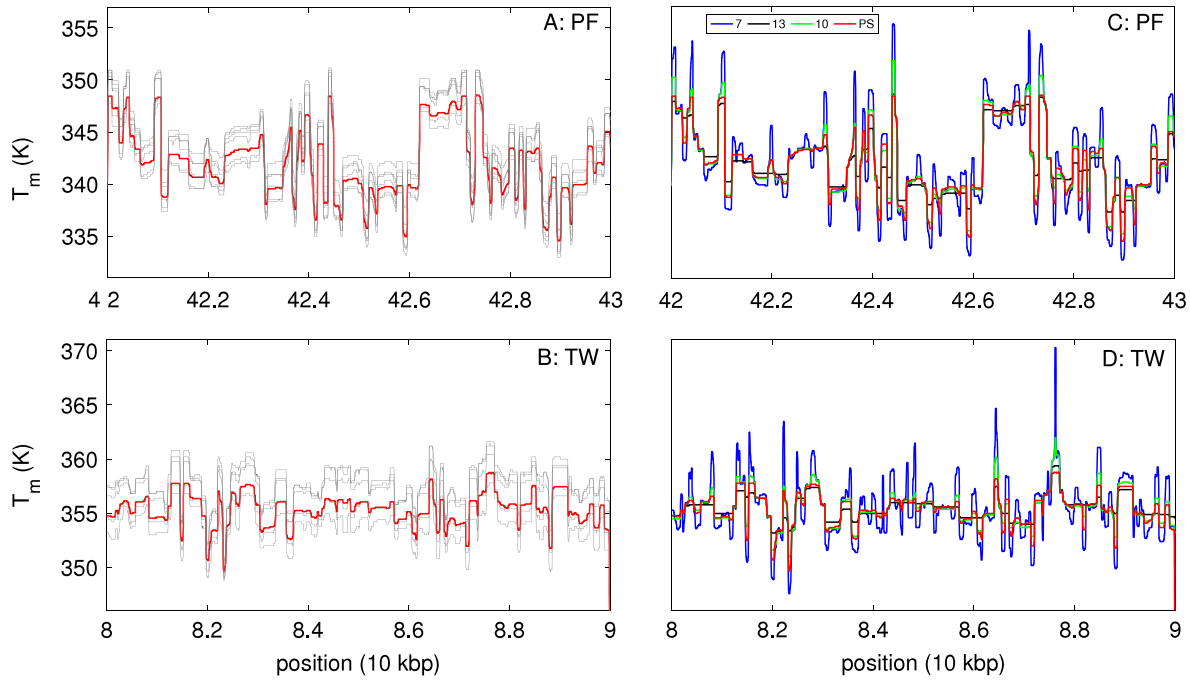

Figure 2. Evolution of $T_m(i)$ for a piece of the genome of *Plasmodium falciparum* (A, C) and *Tropheryma whipplei* (B, D) ($[Na^+] = 0.1$ M) calculated with the PS model using standard parameters (red) or different random sets of parameters in the confidence limit (gray) (A, B) and with the ZB model with various values of γ (C, D).

If this approximation is made, the Hamiltonian of the system can be written as

$$\begin{aligned}\mathcal{H} &= \sum_i \{\Delta g_{NN}(i, i+1)\,\theta_i\,\theta_{i+1} \\ &\quad + \gamma[\theta_i(1-\theta_{i+1}) + \theta_{i+1}(1-\theta_i)]\} \\ &= \sum_i \{2\gamma\theta_i + (\Delta g_{NN}(i, i+1) - 2\gamma)\,\theta_i\,\theta_{i+1}\} \end{aligned} \quad (1)$$

where $\theta_i = 0(1)$ if base-pair i is open (closed). In equation (1), the first term is the sum over the free energies of base-pair step formation, the second term represents the forking contribution at a stem–loop boundary. The partition function is defined as

$$\mathcal{Z} = \sum_{\theta_i} \exp\left(-\beta\mathcal{H}(\{\theta_j\})\right). \quad (2)$$

The different free energy contributions are illustrated in figure 1.

We are interested in melting. The basic properties of interest are the temperature-dependent probabilities that base-pair i is closed, $\Theta(i, T) \equiv \langle\theta_i\rangle(T)$, and the local melting temperature $T_m(i)$, defined as $\Theta(i, T_m(i)) = 1/2$. Due to the formulation as a simple 1D Ising-like model, resolution of the model and calculation of observables can be easily performed using a simple $\mathcal{O}(N)$ transfer-matrix method algorithm [34, 58]:

$$\mathcal{Z} = \text{Trace}\left(\prod_{i=1}^{N} T_i\right) \quad (3)$$

$$\Theta(i, T) = \frac{1}{\mathcal{Z}}\text{Trace}\left[\left(\prod_{j=1}^{i-1} T_j\right) T_i' \left(\prod_{j=i+1}^{N} T_j\right)\right] \quad (4)$$

Table 1. Typical values for γ and $\sigma\bar{L}^{-c}$ for various values of $\bar{L}$.

$\bar{L}$	$\gamma/(k_B T)$	$\sigma\bar{L}^{-c}$
10	7.0	8.9×10^{-7}
170	10.0	2.0×10^{-9}
500	11.2	2.0×10^{-10}
1000	11.9	4.5×10^{-11}
5000	13.6	1.4×10^{-12}
10000	14.4	3.2×10^{-13}

with

$$T_i = \begin{pmatrix} e^{-\beta\Delta g_{NN}(i,i+1)} & e^{-2\beta\gamma} \\ 1 & 1 \end{pmatrix}$$

$$T_i' = \begin{pmatrix} e^{-\beta\Delta g_{NN}(i,i+1)} & e^{-2\beta\gamma} \\ 0 & 0 \end{pmatrix}.$$

Capping terms are neglected and periodic boundary conditions are used. The parameter values are the same as those used for the PS model in [44]. The value for γ depends on the choice of $\bar{L}$ (see table 1).

For a sequence of N base-pairs (bp), the ZB model can be solved by a $\mathcal{O}(N)$ algorithm (transfer-matrix method). While the exact solution of the PS model requires $\mathcal{O}(N^2)$ operations, the Fixman–Freire algorithm [53], by approximating the non-local (and limiting factor) correction to the loop entropy L^{-c} by a multiexponential expansion $L^{-c} \approx \sum_{i=1}^{I} a_i \exp(-b_i L)$, reduces the PS computation to a $\mathcal{O}(N \times I)$ algorithm, where $I \propto \log(L_{max})$, the maximal loop length approximated by the FF approximation. Typically, for loop lengths up to 5 kbp, $I \approx 14$, for loop lengths up to 10^8 bp, $I \approx 21$.

For cache and numerical reasons, the PS model computation can be further accelerated by slicing the sequence

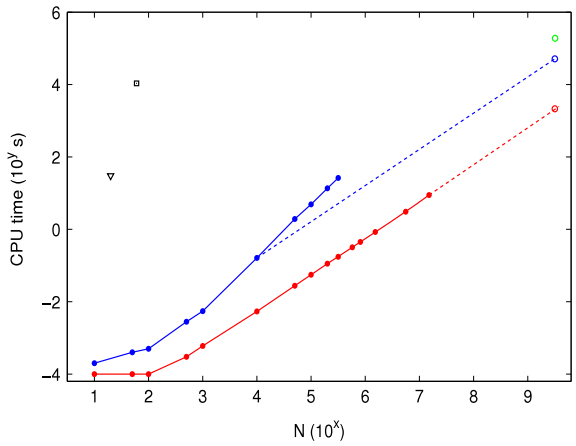

Figure 3. CPU time (in seconds) used to compute melting properties for one temperature as a function of the sequence length in a log–log plot on a 2.4 GHz Intel Core 2 Duo processor using the PS model (solid blue line) with $N_s = L_{max} = N$ and the ZB model (solid red line). The circles represent the total CPU times needed to study the entire human genome using the PS model (green: $N_s \sim 10^8$ bp, $L_{max} = 10^8$ bp, by Liu *et al* [33]; blue: $N_s = 10$ kbp, $L_{max} = 10$ kbp, extrapolated time) or the ZB model (red, extrapolated time). The black square corresponds to typical resolution time of the PBD model [28] and the black triangle to the lattice model.

into blocks of a given size N_s. For a genomic application, the size of the relevant domains is in the kbp range or below (see figure 11). Therefore, $N_s >$ kbp and $L_{max} \sim$ kbp. For example, Yeramian *et al* in [14] take $N_s = 10$ kbp and $L_{max} = 5$ kbp; in [33], Liu *et al* take $N_s \sim 10^8$ bp and $L_{max} = 10^8$ bp. For numerical convenience, we solve the ZB model for sequence slices of 100 kbp with overlapping windows to prevent boundary effects.

Figure 3 compares the CPU time needed by the PS-FF, the PBD and the ZB models. The ZB model allows us to speed up the melting computations by about a factor of 30 compared to the PS model with $N_s = 10$ kbp and $L_{max} = 10$ kbp and about a factor of 90 with $N_s = 100$ kbp and $L_{max} = 10$ kbp. For example, the calculation of the human melting map by Liu *et al* [33] ($N_s \sim 10^8$ bp, $L_{max} = 10^8$ bp) took 22 CPU days on a HP SuperDome (64× Itanium 2 processors, 1.5 GHz, 6 MB cache) whereas it should take about 6 days for the same machine to solve the PS model ($N_s = 10$ kbp, $L_{max} = 10$ kbp) and about only 5 h for the ZB model.

On the same level, resolutions of the PBD model using a direct numerical integration method [28, 51] or of the lattice model take hours to calculate the melting properties of a 60 bp sequence for one temperature. In the same computing time, sequences of length up to 5×10^8 bp could be investigated with the PS-FF model and up to 10^{10} bp with the ZB model.

3. Validation I: DNA melting

In figures 2(C) and (D), we have plotted melting profiles $T_m(i)$ obtained with the PS model and with the ZB model for three different values of γ (corresponding to $\bar{L} \sim 10, 150, 2500$) for two genomic DNA sequences: one extracted from the chromosome 11 of *Plasmodium falciparum* (PF, accession number NC_004315 [59]) and one extracted from the genome of *Tropheryma whipplei* TW08/27 (TW, accession number BX251411). The behavior of $T_m(i)$ over the sequence is a good guideline for observing melting domains along the chain. Results illustrate the cooperativity of the ZB model: larger interfacial energies γ increase the size of the melting domains. The ZB model reproduces the PS melting profiles reasonably well. For low $\bar{L}$, it predicts small bubbles which are not present in the PS computation, and for high $\bar{L}$, it groups together some PS domains. The curve for $\gamma = 10 k_B T$ is in excellent agreement with the prediction of the PS model. To be more precise about the systematic error introduced by the ZB approximation, we calculate

$$\langle |\Delta T| \rangle = \frac{1}{N} \sum_i |T_m(i, \mathrm{ZB}) - T_m(i, \mathrm{PS})| \qquad (5)$$

as a function of γ. In figure 4(A), we show that there exists, for each organism, a minimum which corresponds to the typical bubble size in the sequence. The optimal $\bar{L}$ is approximately 150–200 bp and the minimized error is very small ($\langle |\Delta T| \rangle_{min} \approx 0.3\ K$). By comparison, errors due to parameterization are about 1–2 K (figures 2(A) and (B)). For the sequences from figure 2, the correlation between T_m(ZB) and T_m(PS) approaches 1, as shown in figure 4(B).

4. Validation II: genomic applications

4.1. Gene finding in genomic DNA

The sharpness of the helix-loop transition predicted by melting models allows us to identify domains along the sequence. Matching these domains with the genome properties of DNA is therefore an interesting way to compare genomic and physical properties in natural sequences. As G · C base-pairs are more stable than A · T base-pairs, the melting temperatures depend on the local GC content f_{GC} [18]. Generally the coding regions of genomes are GC-rich compared to the average GC content [60, 61] (table 2). For this reason, it was proposed to compare melting domains and coding parts of the genome [13, 14, 20–24]. For several temperatures, Yeramian *et al* [14], plot $\Theta(i, T)$ as a function of the base-pair position. For each temperature, the evolution of $\Theta(i, T)$ along the sequence highlights successive regions which are compared with coding sequences (CDS). We perform this comparison for the entire genome of PF and TW. As a confirmation, figure 5 underlines the agreement between results with PS and ZB. Moreover, as in [14], a significant mapping between coding and closed domains is found for PF at 68 or 70 °C but no clear evidence of correspondence is observed for TW.

4.2. Exon boundaries in complementary DNA

Carlon *et al* [25] studied the melting properties of complementary DNA (cDNA). cDNA is the reverse transcription from the single-stranded mRNA, it contains no introns. Figure 6 shows the differential melting curve $-\mathrm{d}\Theta_t/\mathrm{d}T$ (where $\Theta_t(T) = (1/N) \sum_i \Theta(i, T)$) for three human cDNAs. Again,

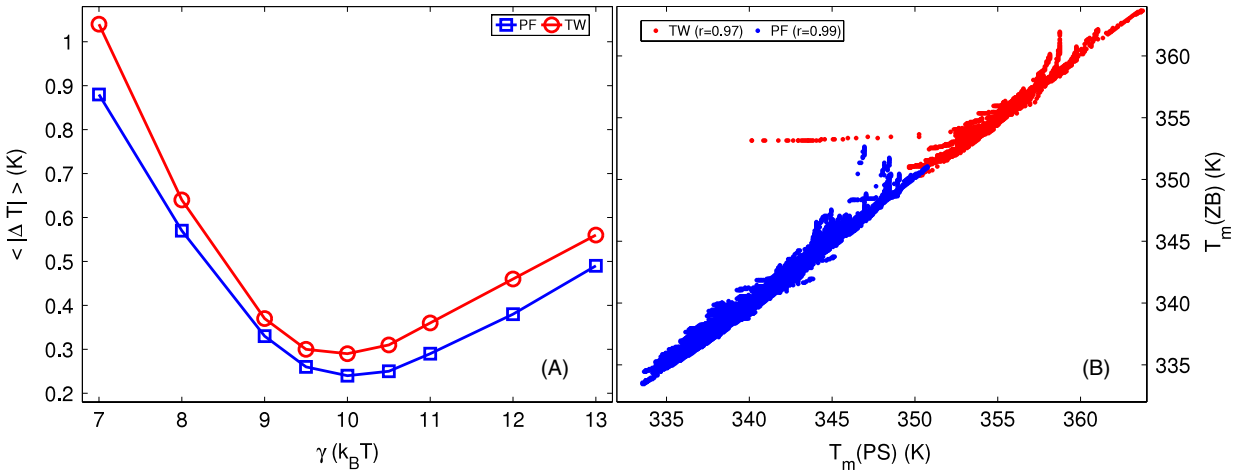

Figure 4. (A) Mean error on temperatures $\langle|\Delta T|\rangle$ as a function of γ for several sequences. (B) Scatter plot of $T_{\rm m}$(ZB) calculated with the ZB model versus $T_{\rm m}$(PS) calculated with the PS model. Correlation factor r for PF and TW is given in the legend box ($\gamma = 10k_{\rm B}T$).

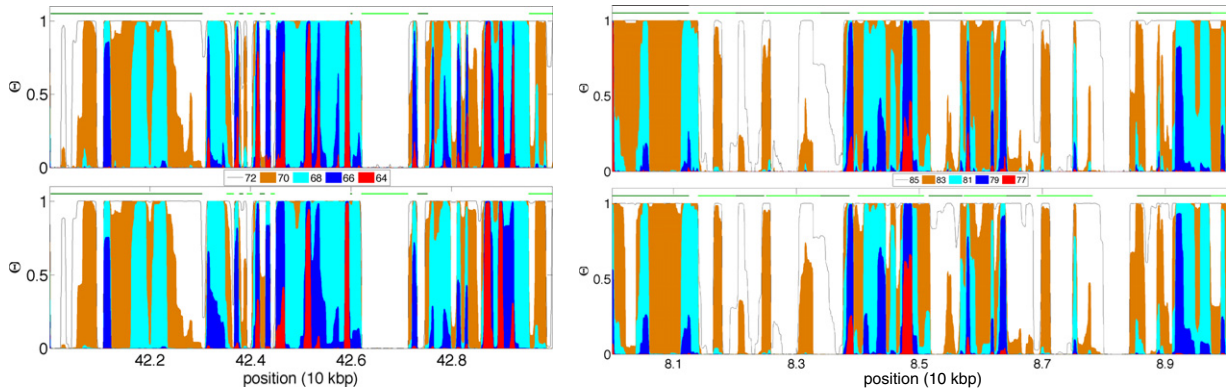

Figure 5. Probabilities of helix opening along the sequence for PF (left) and TW (right) at different temperatures (see the color legend) calculated with the PS (top) or ZB (bottom) model ($[{\rm Na}^+] = 0.1M$, $\gamma = 10k_{\rm B}T$). Coding domains are represented in green at the top of the curves.

Table 2. Structural properties for several prokaryote (†) and eukaryote (∗) genomes.

Species	N (Mbp)	%CDS	$f_{\rm GC}^{\rm tot}$	$f_{\rm GC}^{\rm cds}$	$f_{\rm GC}^{\rm rest}$	$\Delta\tau_{\rm max}$
P. falciparum∗ (PF)	22.86	52.5	0.190	0.237	0.146	0.33
S. pombe∗ (SP)	12.57	56.4	0.36	0.397	0.314	0.30
T. whipplei† (TW)	0.93	84.4	0.463	0.464	0.458	0.03
E. coli† (EC)	4.64	88.0	0.508	0.519	0.43	0.10
S. cerevisiae∗ (SC)	12.07	72.4	0.383	0.396	0.348	0.14
D. melanogaster∗ (DM)	120.38	18.6	0.424	0.534	0.403	0.17
C. elegans∗ (CE)	100.26	25.2	0.354	0.426	0.330	0.16

the results corroborate the validity of the ZB approximations and show for some exon boundaries a good correspondence with melting properties. For complex melting curves, it can be convenient to compute the temperature range over which a boundary between melting domains exists at particular positions in the cDNA in comparison to exon–exon boundaries. Figure 7 confirms the possible coincidence between melting domains and exons boundaries in human cDNA.

5. Genome wide melting analysis

In section 4 we demonstrated the utility of the ZB model in terms of numerical accuracy and computational performance. We now extend the analysis proposed in [13, 14, 23, 24] to the entire genomes of seven species (*Plasmodium falciparum, Schizosaccharomyces pombe, Tropheryma whipplei, Escherichia coli, Saccharomyces cerevisiae, Drosophila melanogaster, Caenorhabditis elegans*, see table 2 for details and abbreviations) partially analyzed before (PF, TW, SC, SP, DM) or known as model organisms (SC, EC, CE). These examples include both prokaryotic (TW, EC) and eukaryotic (PF, SP, SC, DM, CE) organisms with CDS densities ranging from 18% to 88% and with GC content varying between 19% and 51% (see table 2).

5.1. Correlations on the base-pair level

We start with a more quantitative comparison of coding and melting properties. For each studied temperature T, we can attribute to each base-pair i a predicted state: if $T_{\rm m}(i) > T$ (i.e. the base-pair is closed at T), the predicted state of i is

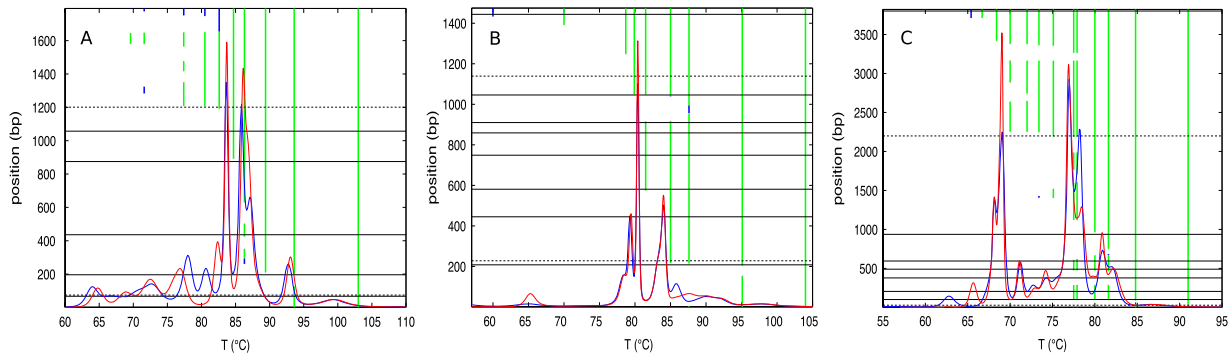

Figure 6. Differential melting curve $-\mathrm{d}\Theta_t/\mathrm{d}T$ and melting domains for: (A) human β-actin cDNA (accession number NM_001101), (B) human CDK4 cDNA (NM_000075) and (C) the human gene EHHADH (NM_001966) in a 0.05 M [Na^+]-buffer ($\gamma = 10k_BT$) obtained with the PS model (blue) and the ZB model (red). Vertical bars indicate the base-pairs for which $T_m(i) > T$ for the PS model strictly (blue), for the ZB model strictly (red) or for both models (green). Horizontal bars are the exon–exon boundaries (solid lines) and the boundaries between coding sequences (CDS) and untranslated regions (UTRs) (dashed lines).

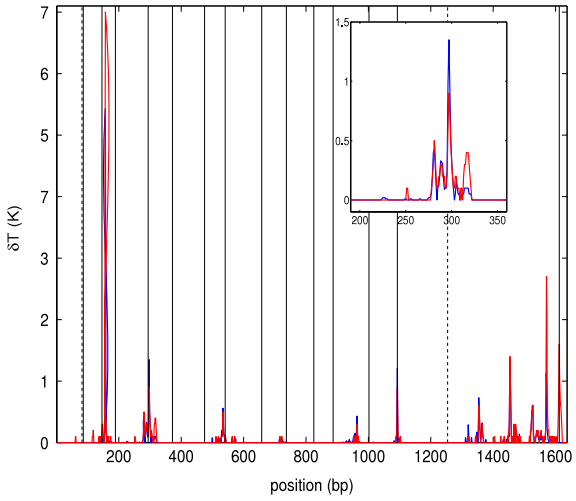

Figure 7. The temperature range δT over which a particular base-pair is a boundary between melting domains for the human interleukin enhancer binding factor 2 (ILF2) cDNA (accession number NM_004515) in a 0.05 M [Na^+]-buffer ($\gamma = 10k_BT$) obtained with the PS model (blue) and the ZB model (red). Vertical lines represent the exon–exon boundaries (solid lines) or the CDS–UTR boundaries (dashed lines).

coding otherwise the base is open and its predicted state is *non-coding*. For the studied genomes, we obtain as a function of temperature the number of true positive predictions N_{TP} (i is correctly predicted as a *coding* base), of true negative predictions N_{TN}, of false positive predictions N_{FP} and of false negative predictions N_{FN}. From these numbers, we deduce the value of statistical indicators: the sensitivity $\beta \equiv N_{TP}/(N_{TP} + N_{FN})$ measuring how well the coding base-pairs of the genome are identified as closed, the specificity $\alpha \equiv N_{TN}/(N_{TN} + N_{FP})$ evaluating how well the non-coding base-pairs are identified as open, and the correlation rate $\tau \equiv (N_{TP} + N_{TN})/(N_{TP} + N_{FN} + N_{TN} + N_{FP})$ measuring the fraction of well predicted base-pairs. The mapping between coding and closed domains is more and more accurate as the three indicators approach *together* 1. Figures 8(A) and (B) represent α, β and τ for PF and TW as a function of the temperature and compare results with α^r, β^r and τ^r calculated from a random distribution of coding base-pairs along the sequence. At low temperatures, all the base-pairs are closed and seen as coding. Therefore the sensitivity tends to 1 (all the coding bases are recognized), the specificity tends to 0 (no non-coding bases are recognized) and τ tends to the percentage of CDS in the genome. In the high temperature limit, the effects are reversed, $\beta \to 0$, $\alpha \to 1$ and τ tends to the percentage of non-coding base-pairs in the sequence. In figure 8(C), we observe the temperature dependence of the correlation increase relative to the random case $\Delta\tau \equiv \tau - \tau^r$. As τ^r corresponds to the average random case, fluctuations σ_{τ^r} around the mean value are evaluated by generating several different random coding annotations for the genome in question and by comparing them with the melting properties (see legend of figure 8). While for TW, no clear difference appears with the random distribution, for PF a large range of temperatures exists where the difference is significant (compared to σ_{τ^r}) and the correlation between coding and thermodynamic domains is high.

Concerning the sensitivity and the specificity, it is possible to show (see appendix) that

$$\Delta\beta = \frac{1}{2 \times \%\mathrm{CDS}}\Delta\tau \tag{6}$$

$$\Delta\alpha = \frac{1}{2 \times (1 - \%\mathrm{CDS})}\Delta\tau \tag{7}$$

where %CDS is the percentage of coding sequences in the genome (see table 2). Equations (6) and (7) imply directly that, for a given species, $\Delta\tau$, $\Delta\beta$ and $\Delta\alpha$ are maximal at the same temperature (defined as the optimal temperature T_{opt}) and are proportional. Figure 9(A) indicates that the maximal matching occurs around $T \approx T_{mid}$ (where $T_{mid} = (\langle T_m^{cds}\rangle + \langle T_m^{rest}\rangle)/2$), i.e. between the average melting temperature of coding regions $\langle T_m^{cds}\rangle$ and the average melting temperature of the rest of the genome $\langle T_m^{rest}\rangle$.

The reason for the different rates of success of the melting analysis becomes apparent in figure 9(B). It shows the normalized melting temperature distributions for the coding

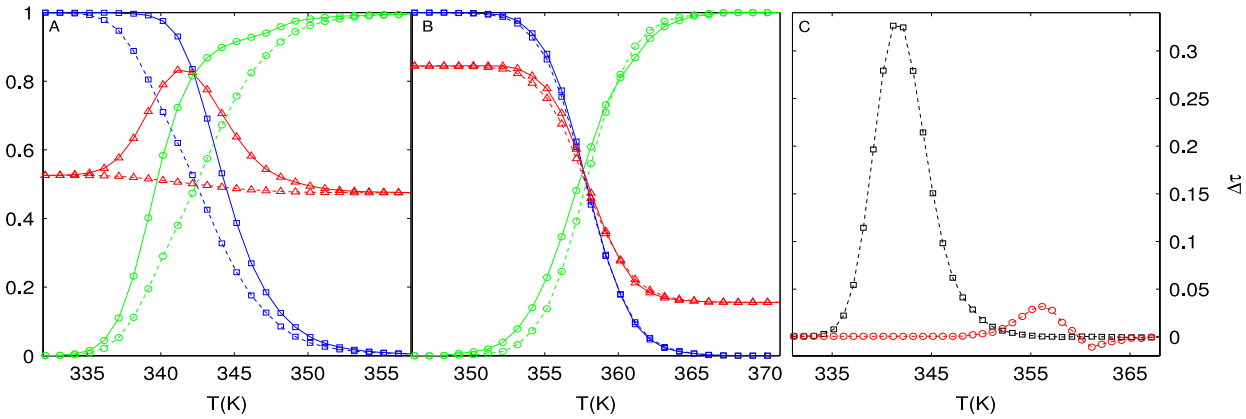

Figure 8. (A, B) Statistical indicators α (green circles), β (blue squares) and τ (red triangles) (see text) for PF (A) and TW (B) with the real coding annotations (full lines) or for a random distribution of the coding base-pairs (dashed lines). (C) $\Delta\tau \equiv \tau - \tau^{\mathrm{r}}$ for PF (black squares) and TW (red circles). Fluctuations of τ^{r} are smaller than the symbol sizes: $\sigma_{\tau^{\mathrm{r}}} = 1 \times 10^{-3}$ (PF) and 5×10^{-3} (TW).

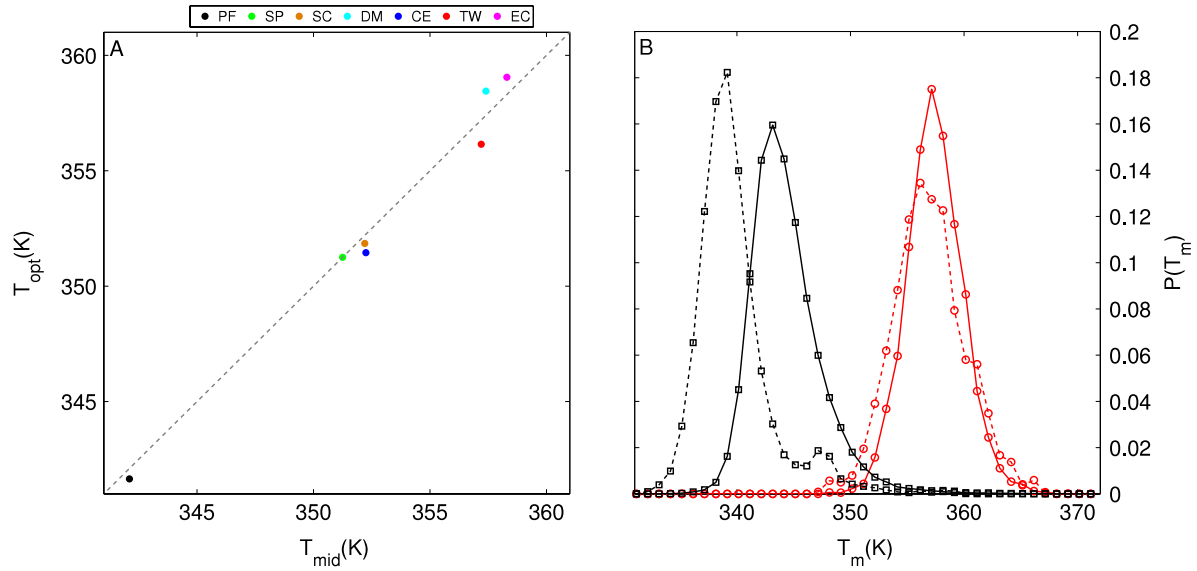

Figure 9. (A) Temperature T_{opt} where $\Delta\tau$ is maximal as a function of the middle temperature T_{mid} between the average melting temperature of coding and non-coding regions for several species (see legend box and table 2). (B) Normalized distribution $P(T_{\mathrm{m}})$ of the melting temperatures for the coding base-pairs (solid lines) and the non-coding base-pairs (dashed lines) for PF (black squares) and TW (red circles).

and non-coding base-pairs for PF and TW. The overlap of the two distributions appears to be organism-dependent and it can be evaluated via the parameter $O_{T_{\mathrm{m}}} \equiv \Delta T/\sqrt{w_{\mathrm{cds}}^2 + w_{\mathrm{rest}}^2}$, where $\Delta T \equiv \langle T_{\mathrm{m}}^{\mathrm{cds}}\rangle - \langle T_{\mathrm{m}}^{\mathrm{rest}}\rangle$ and w_{cds} and w_{rest} are, respectively, the standard deviations of the coding and the non-coding distributions of melting temperatures. In the poorly matching example TW, the two distributions are quasi-identical: $\Delta T \equiv \langle T_{\mathrm{m}}^{\mathrm{cds}}\rangle - \langle T_{\mathrm{m}}^{\mathrm{rest}}\rangle$ is very small ($\approx$0.5 K) and $O_{T_{\mathrm{m}}} = 0.124$. In contrast, for PF, the two distributions are largely separated ($\Delta T \approx 4.5$ K) and overlap much less ($O_{T_{\mathrm{m}}} = 0.965$).

In figure 10, we show how the normalized overlap correlation is connected with the success of the melting analysis though the maximal values of $\Delta\tau$, $\Delta\beta$ and $\Delta\alpha$. In figure 10(A), we plot $\Delta\tau_{\max} \equiv \Delta\tau(T_{\mathrm{opt}})$ as a function of the overlap parameter $O_{T_{\mathrm{m}}}$. The larger the overlap (i.e. the weaker $O_{T_{\mathrm{m}}}$ is), the smaller the predictive power of the melting analysis becomes. If we compare those results with the structural data given in table 2 showing the different GC content of the coding and of the non-coding parts of the genomes, we observe that the mapping between biological and thermodynamic properties works better the more strongly the GC content differs between the coding and the non-coding sequences (see inset in figure 10(A)), underlying the clear relation between melting behavior and the GC composition. Figures 10(B) and (C) show the behavior of $\Delta\beta_{\max}$ and $\Delta\alpha_{\max}$ as a function of $O_{T_{\mathrm{m}}}/(\%\mathrm{CDS})$ and $O_{T_{\mathrm{m}}}/(1 - \%\mathrm{CDS})$ (parameters inspired by equations (6) and (7)). Generally, we remark on an important difference from the random case in the sensitivity and a weaker difference for the specificity for gene-poor genomes (DM, CE) and the reverse observations can be made for gene-rich genomes (SC, EC).

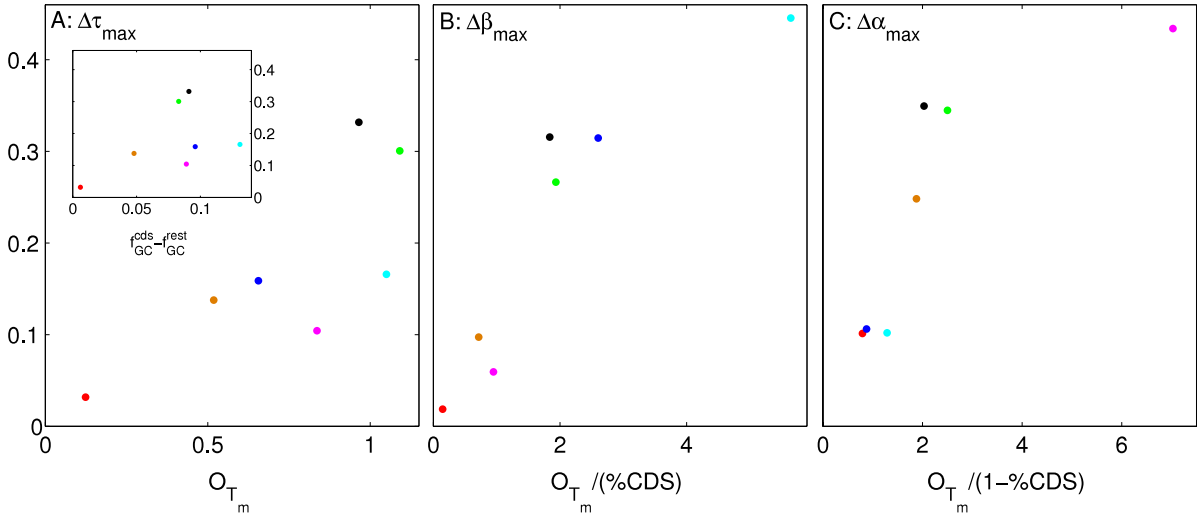

Figure 10. Differences $\Delta\tau_{max}$ (A), $\Delta\beta_{max}$ (B) and $\Delta\alpha_{max}$ (C) for several species (same color legend as in figure 9(A)) as a function of the overlap of the distributions of the melting temperatures for the coding and non-coding parts of genomes (see figure 9(B)). The inset in (A) represents $\Delta\tau_{max}$ as a function of the difference between the GC content of the coding f^{cds}_{GC} and the non-coding f^{rest}_{GC} parts of the genomes.

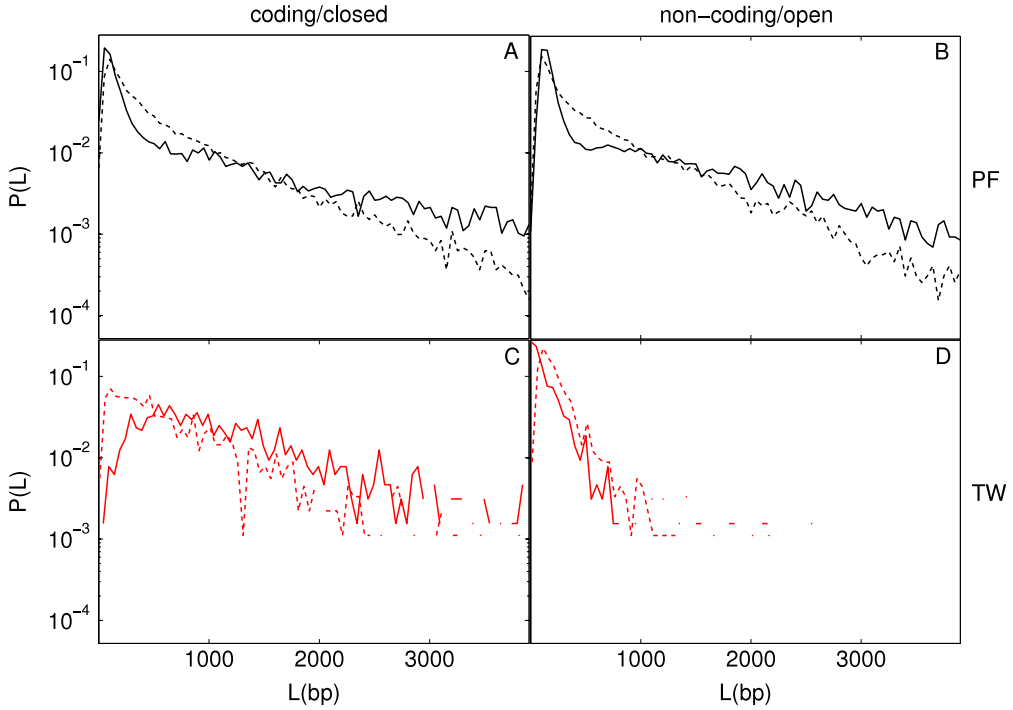

Figure 11. Normalized distributions of the coding domain sizes (solid lines) and of the closed domain sizes (dashed lines) for PF (A) and TW (C), and the normalized distribution of the non-coding domain sizes (solid lines) and of the open domain sizes (dashed lines) for PF (B) and TW (D).

5.2. Correlations on the domain level

In the preceding section we studied the relation between coding and melting properties at the base-pair level. In the following we investigate if the melting analysis recognizes biologically functional *domains*. To remove the temperature dependence of the results, we study each species at the optimal temperature T_{opt} identified above.

Figure 11 shows the normalized distributions of the sizes of the different domain types (coding/non-coding, closed/open) for PF and TW. For PF, the closed and coding distributions, as well as the open and non-coding distributions, are identically peaked around the same domain size, but the distributions of the melting domains are more concentrated around the small region lengths. For TW, distributions extend over a similar size range, but show larger deviations in the limit of small domain sizes. To be more precise about domain identification, for each region we define the local sensitivity β_{loc} (for coding domains) or the local specificity α_{loc} (for non-

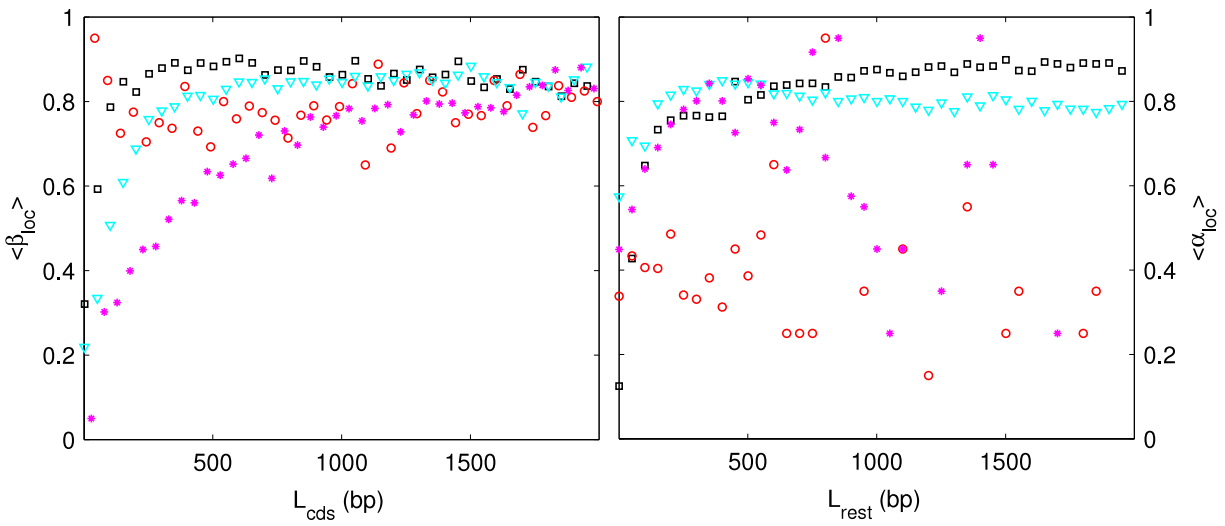

Figure 12. (A) Mean value $\langle\beta_{\rm loc}\rangle$ of the local sensitivity for coding domains as a function of their lengths $L_{\rm cds}$ for PF (black squares), TW (red circles), DM (cyan triangles) and EC (purple stars). (B) Mean value $\langle\alpha_{\rm loc}\rangle$ of the local specificity for non-coding domains as a function of their lengths $L_{\rm rest}$.

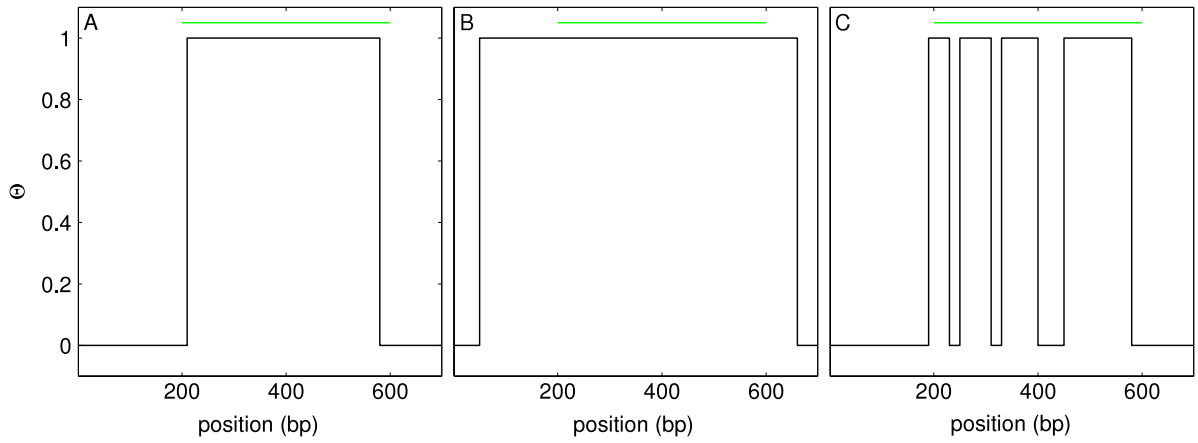

Figure 13. Possible repartitions of closed ($\Theta = 1$) and open ($\Theta = 0$) domains (black lines) around a 400-bp coding regions (green lines). $\beta_{\rm loc} = 0.93$ (A), 1 (B) or 0.73 (C).

coding domains) as the rate of well predicted base-pairs in a single region. Figure 12 shows the mean value of $\beta_{\rm loc}$ (and $\alpha_{\rm loc}$) as a function of the size of the coding (respectively non-coding) regions $L_{\rm cds}$ (respectively $L_{\rm rest}$). We remark that, at the optimal temperature, the short domains are generally poorly identified. For larger domains, the values of $\langle\beta_{\rm loc}\rangle$ and $\langle\alpha_{\rm loc}\rangle$ are approximately constant around the average sequence values $\beta(T_{\rm opt})$ and $\alpha(T_{\rm opt})$. The short size effect means that cooperativity dominates for domain sizes below the cooperative length $\bar{L}$. The constant values of $\langle\beta_{\rm loc}\rangle$ and $\langle\alpha_{\rm loc}\rangle$ for domains with sizes between 500 and 2000 bp indicates that there is no preferential length for the success of the melting analysis in this length range. This is in itself an interesting and non-trivial result. Moreover, we can remark that locally, even for TW (whose $\Delta\beta_{\rm max}$, $\Delta\alpha_{\rm max}$ and $\Delta\tau_{\rm max}$ are very low), $\langle\beta_{\rm loc}\rangle$ and $\langle\alpha_{\rm loc}\rangle$ are not negligible. Nevertheless, a good value for $\langle\beta_{\rm loc}\rangle$ or $\langle\alpha_{\rm loc}\rangle$ is not inevitably a signature of good domain recognition. Indeed, in figure 13, we represent three possible repartitions of closed and open domains around a coding region. In all cases, $\beta_{\rm loc}$ is high, but only in the first one is the coding domain well identified.

Superficial inspection of most indicators presented so far does not reveal striking differences in predictive power for PF and TW, even if figure 8(C) shows the random behavior of the melting analysis for TW. To go further in the domain analysis, we plot in figure 14 the joined probability distributions for a base-pair to be simultaneously in a coding/non-coding domain of size $L_{\rm cds}/L_{\rm rest}$ and in a closed/open region of size $L_{\rm closed}/L_{\rm open}$. In the case of TW (figure 14(B)), no clear correlation between closed (open) and coding (non-coding) domains appears. The distribution of points is diffuse and highlights the poorness and randomness of the domain recognition for TW. In the case of PF (figure 14(A)), the good correlation between $L_{\rm closed}$ and $L_{\rm cds}$ and between $L_{\rm open}$ and $L_{\rm rest}$ proves the successful identification of the coding and non-coding regions. For longer coding (or non-coding) domains, we observe the loss of the correlations with the closed (or open) regions, and the distribution of points is nearly homogeneous

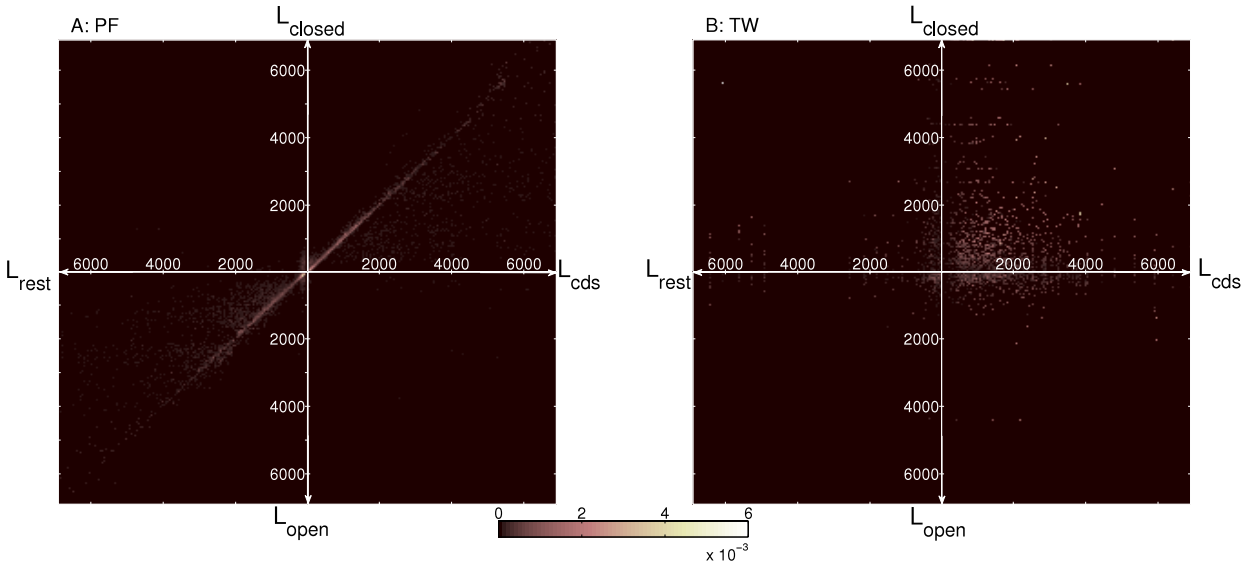

Figure 14. Joined probability distributions for a base-pair to be simultaneously in a coding/non-coding domain of size $L_{\rm cds}/L_{\rm rest}$ and in a closed/open region of size $L_{\rm closed}/L_{\rm open}$, for PF (A) and TW (B).

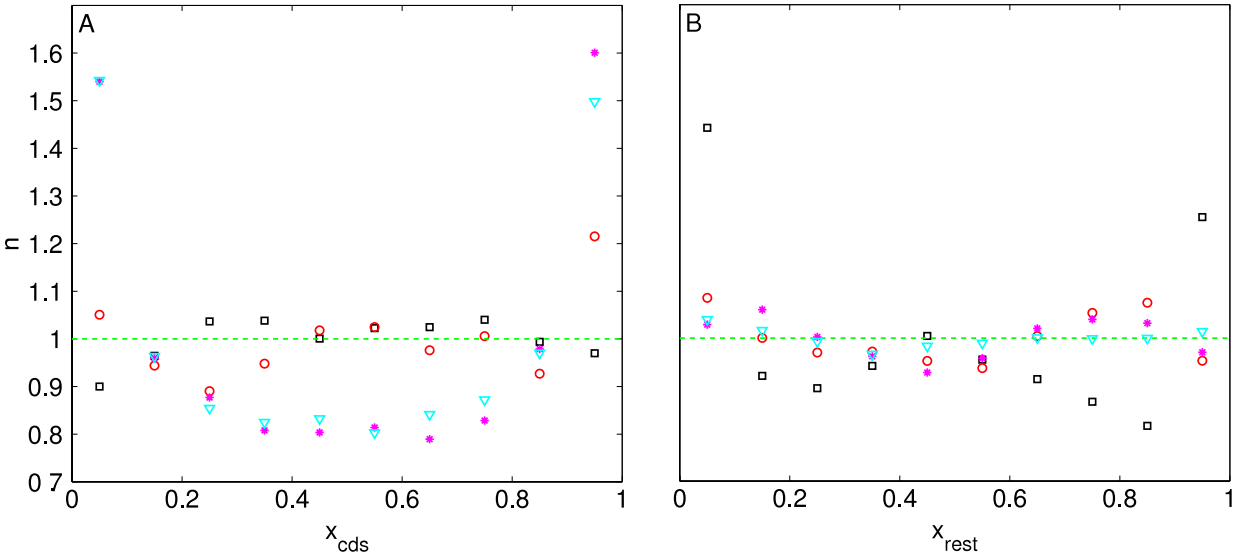

Figure 15. Normalized histograms $n_{\rm cds}(x)$ and $n_{\rm rest}(x)$ of the relative distances x between an error situated in a coding (A) or non-coding (B) domain and the left boundary of the domain (see text), for several species (same legend as in figure 12). The green dashed lines represent the case when the errors are randomly distributed in the domains.

for $L_{\rm closed} < L_{\rm cds}$ (or $L_{\rm open} < L_{\rm rest}$) and quasi-null for $L_{\rm closed} > L_{\rm cds}$ (or $L_{\rm open} > L_{\rm rest}$). This means that, for PF, long biological domains are divided into smaller melting regions (as in the third example of figure 13) revealing some discrepancies in the relation with thermodynamic properties or some putative multi-exon genes [24] which have not been identified by standard gene finding methods [6–12] (see section 6). For other organisms, the same representation reveals that good correlations between closed and coding domains are observed for species with a high $\Delta\beta_{\rm max}$ value and good correlations between open and non-coding domains are observed for species with a high $\Delta\alpha_{\rm max}$ value (data not shown).

As a final step, we investigate the location of incorrectly identified base-pairs in coding and non-coding domains. In each coding or non-coding domain between base-pairs i_1 and i_2, for each error i (i.e. for each open or closed base-pair), we calculate $x = (i - i_1)/(i_2 - i_1)$. $x \sim 0$ or $x \sim 1$ imply that the error i is situated near a boundary. Following [25], in figure 15, we compute the normalized histograms $n_{\rm cds}(x)$ and $n_{\rm rest}(x)$ of the relative distances x observed in coding and non-coding domains. No clear rules or dependences appear. We could just remark than, in general, while $n_{\rm cds}(x)$ has a boundary close distribution, $n_{\rm rest}(x)$ is relatively flat and random, and vice versa. If the errors are more or less situated around the boundaries of a domain, we can consider that the identification of domains is good.

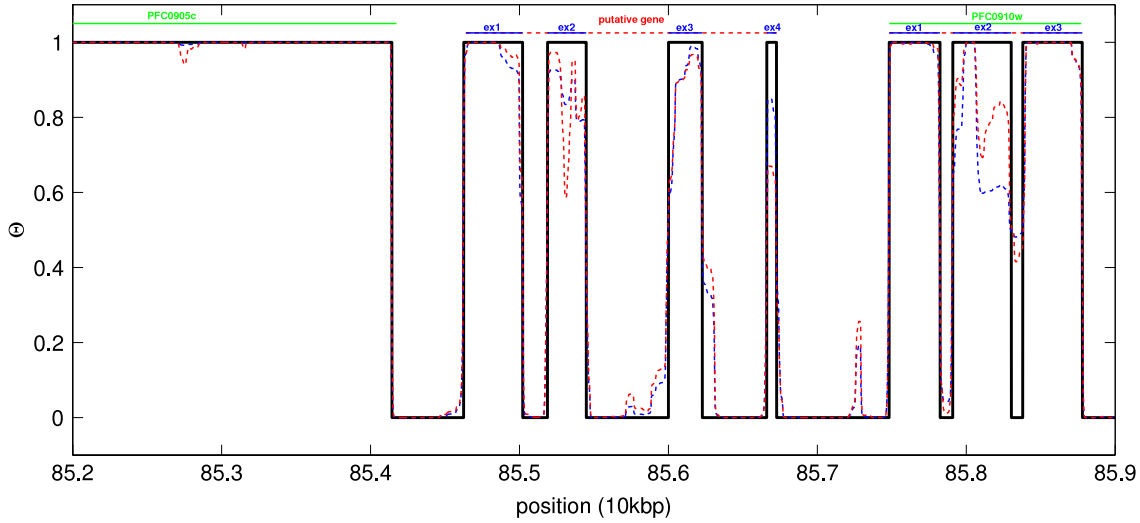

Figure 16. Coding (green) and melting (black) domains at $T_{\rm opt}$ for part of chromosome 3 of PF. Identification of a putative gene composed of four exons and of a possible division of PFC0910w into three exons. The probability Θ computed at $T_{\rm opt}$ with the ZB (blue dashed line) and the PS (red dashed line) models shows a very weak model-dependent behavior.

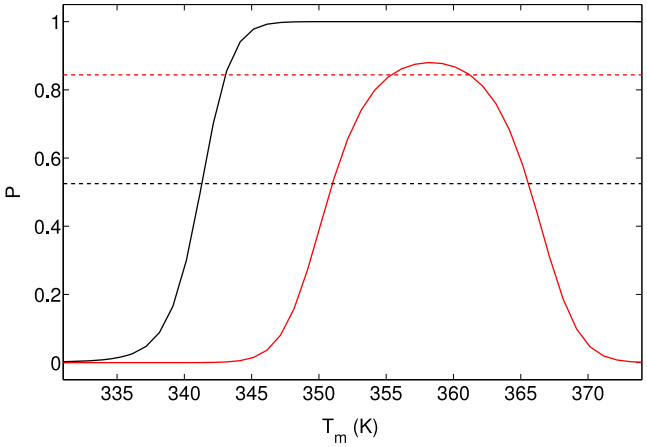

Figure 17. Probability for a base-pair to be part of a coding sequence as a function of the local melting temperature for PF (solid black line) and TW (solid red line). The fractions of coding base-pairs %CDS are shown with dashed lines.

6. Identification of putative exons and genes

As a final step, we estimate the reliability of predictions based on the melting analysis by a statistical comparison to independent annotations present in the databases. The latter are for the most part produced by combining several *ab initio* bioinformatic gene finding codes. Programs like GlimmerM [11] or phat [62] are based on motifs research and Markov models [63]. Their results do not constitute absolute truth and, instead of being errors, the discrepancies we pointed out in the preceding sections could be the signature of putative exons or genes unidentified and unspecified in the databases. Clearly, there is little reason to trust the melting annotation of the TW genome, while in the case of PF some, but probably not all, of the predictions might indeed turn out to be relevant. As a example, in figure 16, we represent part of chromosome 3 of PF where the melting analysis identifies a putative gene composed of four exons between genes PFC0905c and PFC0910w. Similarly PFC0910w, described in the database as a single-exon gene, is possibly composed of three exons. The inspection of the same chromosome segment using the full PS model shows that the proposed domains are not artifacts of the ZB model (see figure 16). If, during the ZB fast analysis of the genome, doubts appear concerning part of the proposed segmentation, one could easily go back to the PS model, for specific sections, to verify or adjust the partition of the domains.

In the following, we determine a position-dependent confidence level for the results of the melting analysis. We base our estimate on the distribution of local melting temperatures, $P(T_{\rm m}|\text{coding})$ and $P(T_{\rm m}|\text{non-coding})$, in parts of the genome identified as coding and non-coding by other annotation methods (figure 9(B)). From these, we can determine the probability that a base-pair with a melting temperature $T_{\rm m}$ belongs to a coding or a non-coding sequence:

$$P(\text{coding}|T_{\rm m}) = \%\text{CDS} \times P(T_{\rm m}|\text{coding})/\{\%\text{CDS} \times P(T_{\rm m}|\text{coding}) + (1 - \%\text{CDS}) \times P(T_{\rm m}|\text{non-coding})\} \quad (8)$$

$$P(\text{non-coding}|T_{\rm m}) = 1 - P(\text{coding}|T_{\rm m}). \quad (9)$$

These quantities provide a convenient local measure of confidence in the predictions of the melting analysis. For simplicity, and in order to obtain robust estimates, we determine them from Gaussian approximations of $P(T_{\rm m}|\text{coding})$ and $P(T_{\rm m}|\text{non-coding})$.

We first consider the case of PF, where we expect the melting analysis to work well. $P(\text{coding}|T_{\rm m})$ exhibits a sigmoidal form in line with Yeramian's original argument (see figure 17). Base-pairs with a high (low) melting temperature have a high probability of being coding (non-coding). Figure 18 represents the same part of chromosome

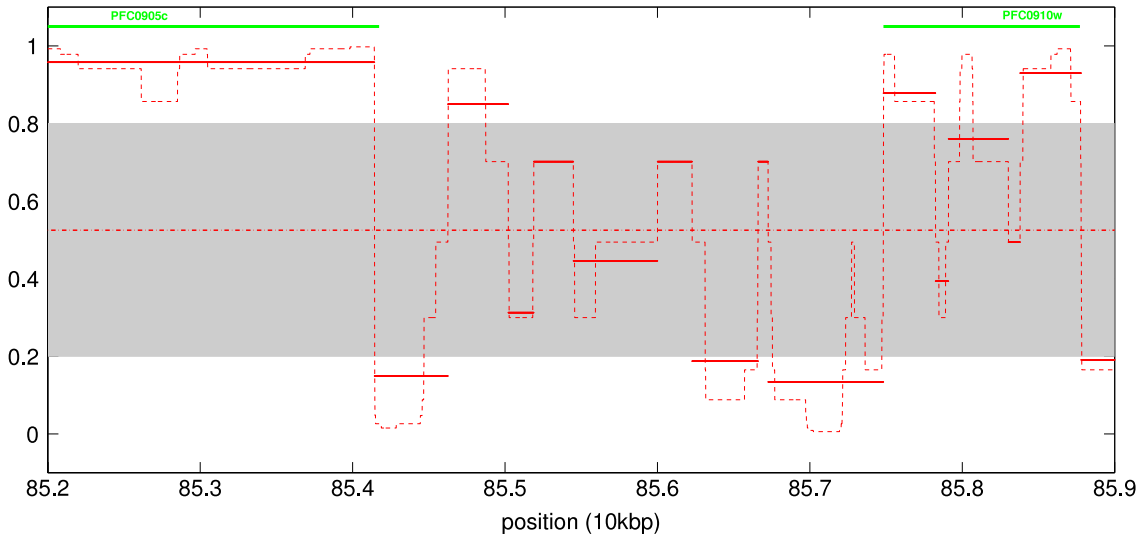

Figure 18. Individual base-pair (red dashed lines) or average melting domain (red solid lines) value of $P(\text{coding}|T_m)$ for a pat of chromosome 3 of PF. The red dotted-dashed line is the random probability %CDS of predicting a coding base-pair. The gray zone ($0.2 < P(\text{coding}|T_m) < 0.8$) is representative of the non-confident predictions. The green lines corresponds to coding exons present in the database.

3 of PF as figure 16 for which we plot $P(\text{coding}|T_m)$ (base-pair and domain mean values). We remark that for the putative gene introduced before, only the first exon is predicted reliably ($P(\text{coding}|T_m) > 0.8$). For the putative division of PFC0910w into three exons, the confidence level of the predicted introns is also relatively low.

In figures 19(A), (B), we compare the standard and melting annotations for the complete PF genome as a function of the confidence we have in our results. As expected from the results presented in section 5, the results largely coincide. The majority of putative coding sequences are detected by both methods with high confidence levels for the melting analysis. Deviations can be clearly grouped into two classes: (1) *failures*, where the melting analysis proposes a deviating annotation at low confidence levels and (2) *predictions* made at high levels of confidence. The latter correspond to the small peak in $P(T_m|\text{non-coding})$ around the average melting temperature of coding sequences in the PF genome, which is visible in figure 9(B). The putative corresponding sequences detected by the melting analysis are listed in the supplementary materials.

The corresponding results for TW are strikingly different. Panels (C) and (D) in figure 19 show that the high confidence levels for predicted coding sequences simply reflect the high proportion of coding sequences in the genome. In fact, a value of $P(\text{coding}|T_m) = \%\text{CDS}$ corresponds to the confidence level of a random annotation reproducing the *average* density of coding sequences. This confidence level can be quite high (%CDS(TW) = 84%), but, of course, one does not learn anything from the exercise. Figure 19(D) shows that the melting analysis does not fare any better in the case of TW. In particular, the analysis cannot predict any *non-coding* sequences with reasonable confidence. We argued in section 5 that the problem is the high overlap of the distributions of melting temperatures for coding and non-coding sequences (figure 9(B)). Closer inspection of $P(\text{coding}|T_m)$ in figure 19 shows that the situation is even worse. The TW genome presents a counter example to the working hypothesis underlying the melting analysis: due to the larger spread of melting temperatures in non-coding sequences, base-pairs with large melting temperatures are more likely to be non-coding than to be coding! Note, however, that this is not a major problem. As long as coding and non-coding sequences differ in their characteristics, it is possible to exploit this difference to construct an annotation scheme along the lines explored in this section.

7. Conclusion

In this paper we have addressed two independent questions concerning the identification of coding sequences in genomes on the basis of thermodynamic melting behavior: (1) which model should be used for the generation of melting profiles and (2) how to quantify the reliability of the predictions for the biological information content.

In a first part we showed that one of the earliest models of DNA thermal denaturation, the ZB model, makes surprisingly reliable predictions for position-dependent melting temperatures even though loop entropies are treated incorrectly compared to the PS model. This underlines the importance of sequence heterogeneity for the physical properties of genomic DNA. The low computational costs of the ZB model make it possible to investigate melting profiles of entire genomes ($\sim 10^8$ bp per hour) on a personal computer.

In the main part of the paper we investigated correlations between the coding and physical melting properties of DNA on a genome wide scale. In particular, we developed a method to estimate the confidence level of the physical annotation based on a statistical comparison to independent results. For some species, the correlations are strong enough to allow us to identify new putative genes and introns with a high level of

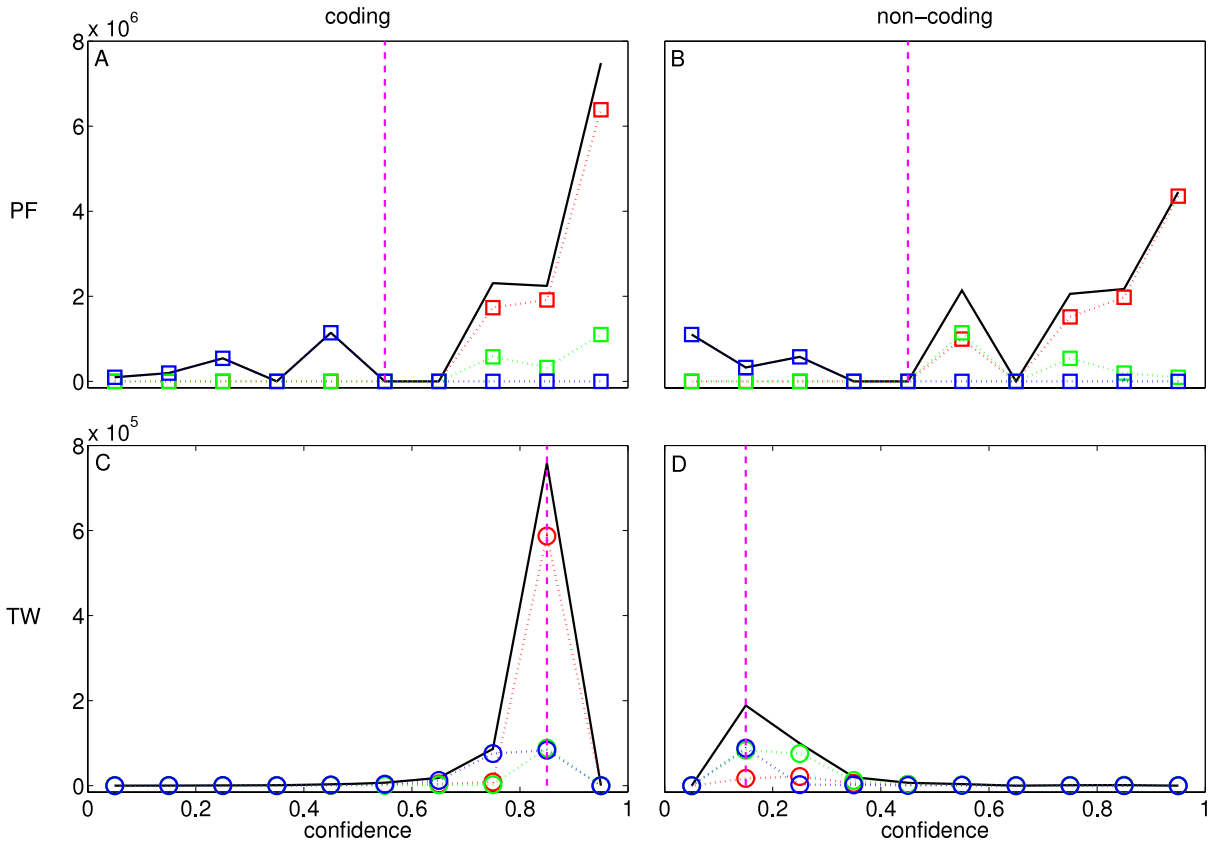

Figure 19. Number of base-pairs predicted as coding (A, C) (or non-coding (B, D)) by the melting analysis or by other methods (solid black lines), by the melting analysis and by other methods (dotted red lines), by the melting analysis only (dotted green lines) and by other methods only (dotted blue lines), as a function of the confidence (P(coding$|T_m$) for the coding predictions and P(non-coding$|T_m$) for the non-coding predictions), for PF (A, B) and TW (C, D). The purple dashed lines represent the confidence in the random case.

confidence (listed in the supplementary materials), suggesting that the coupling of a physically based approach with standard gene finding methods could improve the genomic annotation process. In other cases, the melting analysis yielded no reliable predictions. The key determinant is the difference in the GC content between the coding and the non-coding parts of the genome.

Qualitatively, this mixed success had already been noted in earlier studies [23] and was interpreted as a signature of the strong influence of the physical (melting) properties on the organization of archaic genomes, which was partially lost during later stages of evolution. It could be interesting to use our methods to systematically investigate this hypothesis. Our current results show no significant trend between the studied genomes of eukaryotes and prokaryotes or within the same phylogenetic class (SP, SC).

It is only through an extensive comparison with the results of independent annotation schemes that one can judge the pertinence of the physical approach and the reliability of the deviating predictions. Unfortunately, this validation cannot be done once and for all, but needs to be repeated for each (part of a) genome. Our use of the ZB model is not essential to the results we have obtained in the second part of the paper. Obviously, melting profiles generated on the basis of the PS model could be analyzed in exactly the same way given the necessary computing power. We have seen no evidence that this would lead to significantly altered results. Therefore, for given, limited computing resources, we tend to advocate a broadening of the statistical base of the analysis, at the expense of a possible, but (for the present purposes) marginal, improvement of the description of the physical melting behavior.

Appendix. Random distributions

For a random distribution of the coding base-pairs in the sequence, at a given temperature T, the number of closed base-pairs which actually correspond to a coding one N^r_{TP} is

$$N^r_{TP} = N_{closed} \times \frac{N_{coding}}{N} \tag{10}$$

where N_{closed} is the number of closed base-pairs and $N_{coding} = \%CDS \times N = N^r_{TP} + N^r_{FN}$. Then, $\beta^r = N_{closed}/N$. Similarly, we can show that $\alpha^r = N_{open}/N$ and $\tau^r = (N_{coding}N_{closed} + N_{noncoding}N_{open})/N^2$ where $N_{open} = N - N_{closed}$ and $N_{noncoding} = N - N_{coding}$.

Now, we can derive the equations (6) and (7). Using definitions of $N_{noncoding}$, N_{open}, N, α, β and τ in terms of N_{TP}, N_{TN}, N_{FP} and N_{FN}, we derive

$$\alpha = \frac{1}{2N_{noncoding}}\left(N\tau + N_{noncoding} + N_{open} - N\right) \tag{11}$$

$$\beta = \frac{1}{2N_{\text{coding}}}\left(N\tau - N_{\text{noncoding}} - N_{\text{open}} + N\right). \qquad (12)$$

By replacing τ by $\Delta\tau + \tau^{\text{r}}$, and by subtracting α^{r} from equation (11) and β^{r} from equation (12), we finally obtain

$$\Delta\alpha = \frac{N}{2N_{\text{noncoding}}}\Delta\tau \qquad (13)$$

$$\Delta\beta = \frac{N}{2N_{\text{coding}}}\Delta\tau. \qquad (14)$$

References

[1] Watson J D and Crick F H C 1953 *Nature* **171** 737–8
[2] Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottman F and O'Neal C 1965 *Proc. Natl Acad. Sci. USA* **53** 1161–8
[3] Venter J C *et al* 2001 *Science* **16** 1304–51
[4] The International Human Genome Mapping Consortium 2001 *Nature* **409** 860–921
[5] The International Human Genome Mapping Consortium 2001 *Nature* **409** 934–41
[6] Gelfand M S, Mironov A A and Pevzner P A 1996 *Proc. Natl Acad. Sci. USA* **93** 9061–6
[7] Zhang M Q 1997 *Proc. Natl Acad. Sci. USA* **94** 565–8
[8] Brent M R and Guigo R 2004 *Curr. Opin. Struct. Biol.* **14** 264–72
[9] Burge C and Karlin S 1998 *Curr. Opin. Struct. Biol.* **8** 346–54
[10] Besemer J and Borodovsky M 1999 *Nucleic Acids Res.* **27** 3911–20
[11] Salzberg S L, Pertea M, Delcher A L, Gardner M J and Tettelin H 1999 *Genomics* **59** 24–31
[12] Hiller M, Pudimat R, Bush A and Backhofen R 2006 *Nucleic Acids Res.* **34** e117
[13] Yeramian E, Bonnefoy S and Langsley G 2002 *Bioinformatics* **18** 1–4
[14] Yeramian E and Jones L 2003 *Nucleic Acids Res.* **31** 3843–9
[15] Bernardi G 2000 *Gene* **241** 3–17
[16] Oliver J L, Bernaola-Galvan P, Carpena P and Roman-Roldan R 2001 *Gene* **276** 47–56
[17] Olson W K, Gorin A A, Lu X J, Hock L M and Zhurkin V B 1998 *Proc. Natl Acad. Sci. USA* **95** 11163–8
[18] Frank-Kamenetskii M 1971 *Biopolymers* **10** 2623–4
[19] Audit B, Vaillant C, Arneodo A, d'Aubenton-Carafa Y and Thermes C 2004 *J. Biol. Phys.* **30** 33–81
[20] Tong B Y and Battersby S J 1979 *Nucleic Acids Res.* **6** 1073–9
[21] Gotoh O 1983 *Adv. Biophys.* **16** 1–52
[22] Suyama A and Wada A 1983 *J. Theor. Biol.* **105** 133–45
[23] Yeramian E 2000 *Gene* **255** 139–50
[24] Yeramian E 2000 *Gene* **255** 151–68
[25] Carlon E, Malki M L and Blossey R 2005 *Phys. Rev. Lett.* **94** 178101
[26] Choi C H, Kalosakas G, Rasmussen K O, Hiromura M, Bishop A R and Usheva A 2004 *Nucleic Acids Res.* **32** 1584–90
[27] Kalosakas G, Rasmussen K O, Bishop A R, Choi C H and Usheva A 2004 *Europhys. Lett.* **68** 127–33
[28] van Erp T S, Cuesta-Lopez S, Hagmann J G and Peyrard M 2005 *Phys. Rev. Lett.* **95** 218104
[29] Benham C J 1993 *Proc. Natl Acad. Sci. USA* **90** 2999–3003
[30] Benham C J 1996 *J. Mol. Biol* **255** 425–34
[31] Benham C J and Bi C 2004 *J. Comput. Biol.* **11** 519–43
[32] Ak P and Benham C J 2005 *PLoS Comput. Biol.* **1** e7
[33] Liu F *et al* 2007 *PLos Comput. Biol.* **3** e93
[34] Zimm B H and Bragg J K 1959 *J. Chem. Phys.* **31** 526–35
[35] Wartell R M and Montroll E W 1972 *Adv. Chem. Phys.* **22** 129–203
[36] Crothers D M and Zimm B H 1964 *J. Mol. Biol.* **9** 1–9
[37] DeVoe H and Tinoco I Jr 1962 *J. Mol. Biol.* **4** 500–17
[38] SantaLucia J Jr 1998 *Proc. Natl Acad. Sci. USA* **95** 1460–5
[39] Poland D and Scheraga H A 1966 *J. Chem. Phys.* **45** 1456–64
[40] Poland D and Scheraga H A 1970 *Theory of Helix-coil Transition in Biopolymers* (New York: Academic)
[41] Poland D 1974 *Biopolymers* **13** 1859–71
[42] Wartell R M and Benight A S 1985 *Phys. Rep.* **126** 67–107
[43] Garel T and Orland H 2004 *Biopolymers* **75** 453–67
[44] Jost D and Everaers R 2008 *Biophys. J.* at press (doi:10.1529/biophysj.108.134031)
[45] Everaers R, Kumar S and Simm C 2007 *Phys. Rev.* E **75** 041918
[46] Peyrard M and Bishop A R 1989 *Phys. Rev. Lett.* **62** 2755
[47] Dauxois T, Peyrard M and Bishop A R 1993 *Phys. Rev.* E **47** 684
[48] Lacombe R H and Simha R 1973 *J. Chem. Phys.* **58** 1043–53
[49] Yeramian E, Schaeffer F, Caudron B, Clavierie P and Buc H 1990 *Biopolymers* **30** 481–97
[50] Tostesen E, Liu F, Jenssen T-K and Hovig E 2003 *Biopolymers* **70** 364–76
[51] van Erp T S, Cuesta-Lopez S and Peyrard M 2006 *Eur. Phys. J.* E **20** 421–34
[52] Blossey R and Carlon E 2003 *Phys. Rev.* E **68** 061911
[53] Fixman M and Freire J J 1977 *Biopolymers* **16** 2693–704
[54] Palmeri J, Manghi M and Destainville N 2007 *Phys. Rev. Lett.* **99** 088103
[55] Palmeri J, Manghi M and Destainville N 2008 *Phys. Rev.* E **77** 011913
[56] Kafri Y, Mukamel D and Peliti L 2000 *Phys. Rev. Lett.* **85** 4988
[57] Garel T and Monthus C 2005 *J. Stat. Mech.: Theor. Exp.* P06004
[58] Montroll E W 1941 *J. Chem. Phys.* **9** 706–21
[59] National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov
[60] Sumner A T, de la Torre J and Stuppia L 1993 *J. Mol. Evol.* **37** 117–22
[61] Oliver J L and Marin A 1996 *J. Mol. Evol.* **43** 216–23
[62] Cawley S E, Wirth A I and Speed T P 2001 *Mol. Biochem. Parasitol.* **118** 167–74
[63] Salzberg S L and Delcher A L 2004 *Microbial Genome* (Totowa, NJ: Humana Press)