dna_melting: a code to compute DNA melting temperature extimations

Lara Querciagrossa

Contents

1 Introduction

Various ways for computing melting temperature for a DNA sequence have been published. This code will use 7 different methods and compute melting curves with three of them.

2 Input

The input file is composed by:

- the sequence itself;
- salt concentration (M);
- total nucleotide strand concentration (M).

The ideal $[Na^+]$ concentration is 0.05 M, while the ideal nucleotide strand concentration is $5 \cdot 10^{-8}$ M (even if such value usually ranges between $1 \cdot 10^{-9}$ and $1 \cdot 7^{-6}$).

3 Melting temperature

3.1 Wallace Rule

The first method used takes into account two different equations, depending on sequence length [1]. For sequences shorter than 15 base pairs ($N_{bp} \leq 15$), the correct formula is:

$$T_m[^\circ C] = 2(n_A + n_T) + 4(n_C + n_G) \tag{1}$$

while for longer sequences, another equation is used:

$$T_m[^\circ C] = 69.3 + \left[\frac{41.0(n_C + n_G)}{N_{bp}} - \frac{350}{N_{bp}}\right] \tag{2}$$

These equations are valid for a salt concentration $[Na^+]$ =50 mM, for a nucleotide concentration of 50 nM and at pH=7.0.

3.2 Salt–adjusted method

This method takes into account a term that adjust for the GC content ($\frac{820.0}{n_A+n_T+n_C+n_G}$) and another one for the sequence length ($16.6log([Na^+])$). The following equation [2] is valid for a nucleotide concentration of 50 nM at pH=7.0.

$$T_m[^\circ C] = 100.5 + 41.0\left(\frac{n_C + n_G - 16.4}{n_A + n_T + n_C + n_G}\right) - \left(\frac{820.0}{n_A + n_T + n_C + n_G}\right) + 16.6log([Na^+]) \tag{3}$$

where $[Na^+]$ is in molar units [M].

3.3 Khandelwal method

The Khandelwal method [3] attempts to quantify the energetic components stabilizing the structure of DNA such base pairing, stacking, and ionic environment. The equation used for computing melting temperature for a DNA sequence is:

$$T_m[^\circ C] = 7.35 \cdot E + 17.34\ln(N_{bp}) + 4.96\ln([Na^+]) + 0.89\ln([DNA]) - 25.42 \tag{4}$$

where $[DNA]$ is the total nucleotide strand concentration [M] and E is the strength parameter per base ($strength/N_{bp}$). The *strength* parameter is the sum of the strength for dinucleotide step, which is typical for each pair of subsequent bases, as shown inTable 1.

For example, the sequence GACGACAAGACCGCG ($[Na^+] = 0.22M, [DNA] = 0.000002g/mL$) shows an E value of $E = 128/15 = 8.53$ and a computed melting temperature $T_{m,calc} = 65.04^\circ C$and an experimental one of $T_{m,exp} = 64.4^\circ C$.

3.4 Nearest neighbor (NN) model

This model uses the following equation[4]:

$$T_m[K] = \frac{\sum(\Delta H_d) + \Delta H_i}{\sum(\Delta S_d) + \Delta S_i + \Delta S_{self} + R \cdot \ln\left(\frac{[DNA]}{b}\right)} + 16.6\log([Na^+]) \tag{5}$$

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

Table 1: Strength for dinucleotide step. Note that R stands for purine and Y for pyrimidine, while the value typical of a CG base pair is set to 4 and that of an AT base pair 1.

which is valid in particular for a total strand concentration ($[DNA]$) of 50 nM, salt concentration $[Na^+]$ of 50 mM and at pH=7.0. b is a values which is 4 for non–self–complementary sequences and 1 for self–complementary strands or for duplexes when one strand is in significant excess. ΔS_{self} is an entropic penalty for self–complementary sequences and ΔH_i and ΔS_i are sums of initiation enthalpy and entropy. ΔH_d and ΔS_d are calculated over all internal nearest–neighbor doublets. The Gibbs energy contribution to ΔS_{self} for duplex formed from a self–complementary sequence is 0.4 kcal and 0 kcal for a simple complementary sequence. ΔH_i is equal to zero for a sequence length comprised between 1 and 19 base pairs. $DeltaS_i$ is characterized by a free energy term of 5 kcal for duplex containing GC base pair and of 6 kcal for duplex composed exclusively of AT base pairs. Different parameters have been proposed (Table 2).

Propagation sequence	Breslauer 86 [4]			SantaLucia 96 [5]			Sugimoto 96 [6]		
	ΔH	ΔS	ΔG	ΔH	ΔS	ΔG	ΔH	ΔS	ΔG
AA/TT	-9.1	-24.0	-1.9	-8.4	-23.6	-1.02	-8.0	-21.9	-1.2
AT/TA	-8.6	-23.9	-1.5	-6.5	-18.8	-0.73	-5.6	-15.2	-0.9
TA/AT	-6.0	-16.9	-0.9	-6.3	-18.5	-0.60	-6.6	-18.4	-0.9
CA/GT	-5.8	-12.9	-1.9	-7.4	-19.3	-1.38	-8.2	-21.0	-1.7
GT/CA	-6.5	-17.3	-1.3	-8.6	-23.0	-1.43	-9.4	-25.5	-1.5
CT/GA	-7.8	-20.8	-1.6	-6.1	-16.1	-1.16	-6.6	-16.4	-1.5
GA/CT	-5.6	-13.5	-1.6	-7.7	-20.3	-1.46	-8.8	-23.5	-1.5
CG/GC	-11.9	-27.8	-3.6	-10.1	-25.5	-2.09	-11.8	-29.0	-2.8
GC/CG	-11.1	-26.7	-3.1	-11.1	-28.4	-2.28	-10.5	-26.4	-2.3
GG/GG	-11.0	-26.6	-3.1	-6.7	-15.6	-1.77	-10.9	-28.4	-2.1
Any GC pair?	0.0	-16.77	5.0	0.0	-5.9	1.82	0.6	-9.0	3.4
Only AT pairs?	0.0	-20.13	6.0	0.0	-9.0	2.8	0.6	-9.0	3.4
Symmetry correction	0.0	-1.34	0.4	0.0	-1.4	0.4	0.0	-1.4	0.4
4'–terminal– –TA–3bp	0.0	0.0	0.0	0.4	0.0	0.4	0.0	0.0	0.0

Table 2: Thermodynamics parameters for DNA helix initiation and propagation in 1M NaCl.

3.5 Consensus method

The consensus method [7] gives a quantitative comparison of the similarities and differences among the three previously explained nearest–neighbor models. Such comparison was carried out for a large set of short oligonucleotide sequences ranging from 16 to 30 nucleotide base pairs long, which

span the whole range of GC–content. Results show that significant differences were observed in all methods, which in some cases depend on the oligonucleotide length and the GC–content in a non trivial manner. This method reports regions of consensus and disagreement of various models and the equation for average T_m computation (Figure 1).

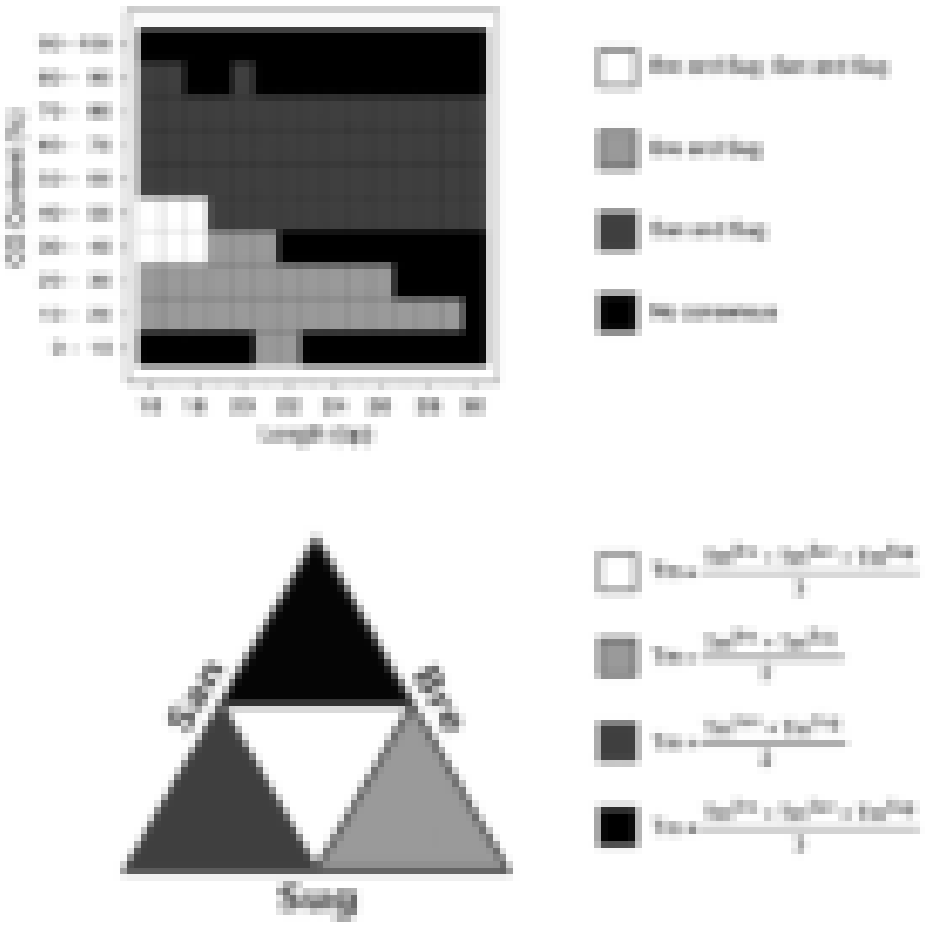

Figure 1: Consensus T_m estimation method.

3.6 Plotting melting curves

Finally, this code gives three output text files containing information for plotting melting curves. The equation used is the following:

$$f = \frac{1 + [DNA] \cdot K_{eq} - \sqrt{1 + 2 \cdot [DNA] \cdot K_{eq}}}{[DNA] \cdot K_{eq}} \tag{6}$$

where

$$K_{eq} = e^{[\frac{\Delta S}{R} - \frac{\Delta H}{RT}]} \tag{7}$$

Curves are obtained using the three NN method (Breslauer, SantaLucia and Sugimoto) for computing ΔS and ΔH. Figure 2 shows a plot of such three output melting curves.

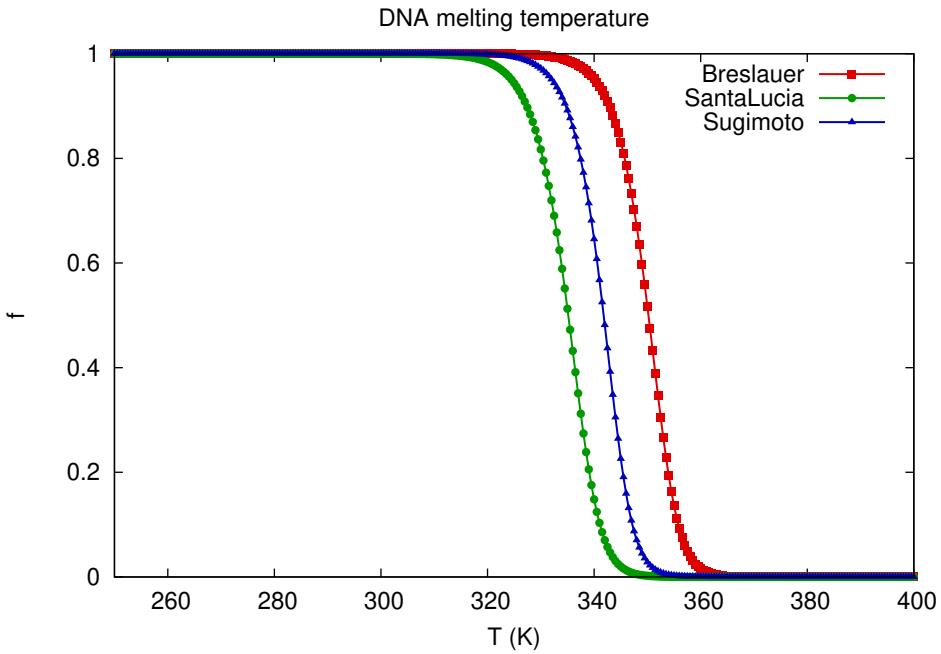

Figure 2: Melting curves.

4 Output

Besides the three melting curves, "dna_melting" prints a series of information:

- sequence itself,
- sequence length,
- number of adenine, cytosine, guanine and thymine,
- GC–content,
- total molecular weight.

The molecular total weight is computed summing those of bases: guanine weights 329.2 Da, adenine 313.2 Da, thymine 304.2 Da and cytosine 298.2 Da.
Then it follows a list of melting temperature computed with the seven models previously described:Wallace rule, salt adjusted method, Khandelwal method, Breslauer method, SantaLucia method, Sugimoto method and consensus method. It follows an example of a typical output:

```
AT.inp
INFO
Sequence.............. TTGTAGTCAT
Length................ 10
Number of Adenine.... 2
Number of Cytosine... 1
Number of Guanine.... 2
Number of Thymine.... 5
GC content........... 30%
Molecular weigth..... 3230 Da

EXTIMATED MELTING TEMPERATURES
1. Wallace rule
Tm: 26C  =  299.15 K

2. Salt adjusted method
Tm: 15.3842C  =  288.534 K

3. Khandelwal method
Tm: 40.5447C  =  313.695 K

4. Breslauer method
Tm: 6.38488C  =  279.535 K

5. SantaLucia method
Tm: 15.0968C  =  288.247 K

6. Sugimoto method
Tm: 19.0703C  =  292.22 K

7. Consensus method
Non-consensus sequence
Tm: 13.5173C  =  286.667 K

Melting curve files written: bre_melting_curve.out, san_melting_curve.out and sug_melting_curve.out
```

References

[1] J. Marmur and P. Doty, J. Mol. Bio. **5**, 109 (1962).

[2] P. Howley, M. Israel, M. Law and M. Martin, J. Biol. Chem. **254**, 4876 (1979).

[3] G. Khandelwal and J. Bhyravabhotla, PlosOne **5, issue 8**, 1 (2010).

[4] K. Breslauer, R. Frank, H. Bl"ocker and L. Marky, Proc. Natl. Acad. Sci. USA **83**, 3746 (1986).

[5] J. J. SantaLucia, H. Allawi and P. Seneviratne, Biochemistry **35**, 3555 (1996).

[6] N. Sugimoto, S. Nakano, M. Yoneyama and K. Honda, Nucl. Acids Res. **24**, 4501 (1996).

[7] A. Panjkovich and F. Melo, Bioinformatics **21, n 6**, 711 (2005).