5004–5006 *Nucleic Acids Research, 1998, Vol. 26, No. 21*

A formula for thermal stability (T_m) prediction of PNA/DNA duplexes

Ursula Giesen, Wilhelm Kleider, Christoph Berding, Albert Geiger, Henrik Ørum[1] and Peter E. Nielsen[2,*]

Boehringer Mannheim GmbH, Roche Diagnostics, Business Unit Laboratory Systems, Bahnhofstr. 9-15, 82327 Tutzing, Germany, [1]PNA Diagnostics, Rønnegade 2, 2100 Copenhagen Ø, Denmark and [2]Center of Biomolecular Recognition, The Panum Institute, IMBG, Department of Biochemistry B, Blegdamsvej 3C, 2200 Copenhagen N, Denmark

Received September 11, 1998; Revised and Accepted September 18, 1998

ABSTRACT

An empirical formula for thermal stability (T_m) prediction of PNA/DNA duplexes has been derived. The model is based on the T_m as calculated for the corresponding DNA/DNA duplex employing a nearest neighbour approach, by including terms for the pyrimidine content and length of the PNA to take into account the increased thermostability of PNA/DNA hybrids and the asymmetry of the PNA–DNA heteroduplex. The predictive power of the T_m prediction formula was challenged with an independent data set not used for model building. The T_m of >90% of the sequences was predicted within 5 K; 98% of the predicted T_ms differ by not more than 10 K from the experimentally determined T_m.

Peptide nucleic acid (PNA) is a DNA mimic with potential use for both diagnostic and therapeutic purposes (1–4). In PNA, the negatively-charged sugar phosphate backbone of DNA is replaced with an uncharged pseudo-peptide backbone. The two strands of a PNA/DNA hybrid therefore lack the electrostatic repulsion as observed for DNA/DNA duplexes, giving rise to virtually ionic strength independent thermal stability (T_m) (5,6). Furthermore, PNA/DNA duplexes generally have higher T_m than the corresponding DNA/DNA duplexes.

It is therefore of paramount importance for most applications of PNA to establish a correlation between T_m and PNA/DNA sequence and length. For DNA duplexes, two formulae for T_m prediction are employed. For oligonucleotides (maximum length ~20–30 bp), the nearest neighbour model has been successfully applied in the past (7–10). For longer DNAs with high T_m, the Marmur formula and refined versions of it are applied (11) that determine T_m as a linear function of GC-content and salt concentration.

The nearest neighbour model describes the T_m on the basis of experimentally accessible thermodynamic terms: the standard enthalpy ΔH^0 and the entropy ΔS^0 of next neighbour bases in a given sequence (7–10). In DNA, there are 10 such values each. The T_m is then calculated as:

$$T_m\ (\text{calc}) = \Sigma\Delta H^0/(\mathbf{R}\ln(\mathbf{C_t}/\mathbf{n}) + \Sigma\Delta S^0)$$

in which **R** is the gas constant, $\mathbf{C_t}$ is the total strand concentration and **n** reflects the symmetry factor, which is 1 in the case of self-complementary strands and 4 in the case of non-self-complementary strands. Additionally, there are penalties for initiation and terminal fray (7–10).

PNA/DNA duplexes are known to be helices comparable to a B-helix (12–13). With analogous underlying structure, one does expect similar stabilising base interactions to take place. We therefore decided to use the T_m calculated for the corresponding DNA/DNA duplex using a nearest neighbour model as a starting point and first dependent variable for a phenomenological model.

For model building, 316 T_m observations were used. These comprised PNA/DNA antiparallel, fully matched duplexes of lengths ranging from 6 to 23 bp and a pyrimidine content in the PNA strand from 0 to 90% (triplex-forming homopyrimidine PNAs were excluded). Approximately 9% of the PNAs were chemically modified (such as His, biotin) at their N- and C-terminals (control experiments using set of PNAs ± modification showed that the effect of these modifications was <2°C). The N-terminal base was A in 23.2%, C in 25.9%, G in 27.1% or T in 23.8% of the PNA sequences. The C-terminal base was A in 18.3%, C in 29.0%, G in 24.7% and T in 28.0% of the PNA sequences. The total number of A-bases in the PNA sequences was 990, of C, 1107, of G, 963, and of T, 1144. The amount of (16) nearest neighbour base pairs present in the data set used for model building was AT: 254, TA: 222, AA: 194, TT: 249, CA: 285, TG: 266, GT: 239, AC: 252, CT: 324, AG: 230, GA: 213, TC: 315, GG: 206, CC: 231, CG: 172, GC: 224.

Generalised linear models were constructed and analysed using Statistical Analysis Software. These models relate the measured T_m to selected independent variables, e.g. length of sequence **L**, number of GC-pairs $\mathbf{N_{GC}}$, number of purine bases $\mathbf{N_p}$ etc:

$$T_m = \mathbf{f}\ (\mathbf{L}, \mathbf{N_{GC}}, \mathbf{N_P}, \ldots\ldots)$$

The coefficients of these models are estimated by means of standard regression techniques so as to minimise the differences between measured and predicted T_m. Good predictive power of such a model requires that all relevant independent factors have been taken into account, representative data have been chosen for coefficient estimation, the relationship between dependent and independent variables is captured by a model up to second order, and that noise is sufficiently small. The quality of each model was initially judged by the regression coefficient, and the significance of contributing factors was analysed by F-tests.

*To whom correspondence should be addressed. Tel: +45 3532 7762; Fax: +45 3539 6042; Email: pen@imbg.ku.dk

If a nearest neighbour approach is to be successful, it must be demonstrated that the measured T_m of PNA/DNA hybrids correlates with the T_m as calculated for the corresponding DNA/DNA duplex using a nearest neighbour model. This was found to be the case (not shown). However, as expected, the T_m prediction formula yielded a T_m that is significantly lower than the experimentally determined T_m of the PNA/DNA duplex. Therefore, suitable variables have to be identified that will correct this.

Another variable must account for the asymptotic behaviour of the PNA/DNA T_ms in terms of oligomer length. The nearest neighbour model for DNA/DNA duplexes (7–10) is employed for DNAs having T_ms lower than ~333 K where there is still a linear contribution of an additional nearest neighbour pair to the T_m. For longer DNA/DNA duplexes, simple empirical formulae like the Marmur formula are employed (11).

A third correction function must reflect the fact the PNA/DNA duplex is an asymmetric molecule, whereas a DNA/DNA duplex is symmetric. Due to the symmetry of the DNA/DNA helix, only 10 of the $2^4 = 16$ possible nearest neighbour interactions are considered for DNA duplexes. An appropriate nearest neighbour model for a PNA/DNA duplex must describe all 16 possible nearest neighbours. The 10 DNA nearest neighbour values are already incorrect for a PNA/DNA duplex (otherwise their T_ms would be identical to those of the corresponding DNA/DNA duplex). Additionally, we are lacking six values completely. Hence, one will have to describe a variable that—in a phenomenological fashion—corrects for these unknown changes in ΔH^0 and ΔS^0 and further thermodynamic contributions due to asymmetry. In addition to the nearest neighbour derived T_m, fractional pyrimidine (or purine) content (cf. 14) and length are required to predict the T_m of PNA/DNA hybrids correctly.

The linear model for the melting temperature prediction of PNA/DNA duplexes thus reads:

$$T_{m_{pred}} = c_0 + c_1 * T_{m_{nnDNA}} + c_2 * f_{pyr} + c_3 * \text{length}$$

in which $T_{m_{nnDNA}}$ is the melting temperature as calculated using a nearest neighbour model for the corresponding DNA/DNA duplex applying ΔH^0 and ΔS^0 values as described by SantaLucia *et al.* (9). f_{pyr} denotes the fractional pyrimidine content, and length is the PNA sequence length in bases.

The constants were determined to be $c_0 = 20.79$, $c_1 = 0.83$, $c_2 = -26.13$ and $c_3 = 0.44$.

Within the data set used to construct the model its quality described by statistical figures is: $R^2 = 0.87$

Parameter	Coefficient estimate	Pr>|T|	Std error of estimate
Intercept	20.79	0.0001	1.43
Length	0.44	0.0030	0.15
f_{pyr}	–26.13	0.0001	1.66
$T_{m_{nnDNA}}$	0.83	0.0001	0.03

An $R^2 = 0.87$ is very good for such a model, and the F-test shows high significance for the three parameters. If the probability that a certain parameter contributes to the observed T_m is >95% then the figure would be 0.04. A result of Pr>|T| = 0.0001 and 0.0030 shows very high significance of the contributing factors.

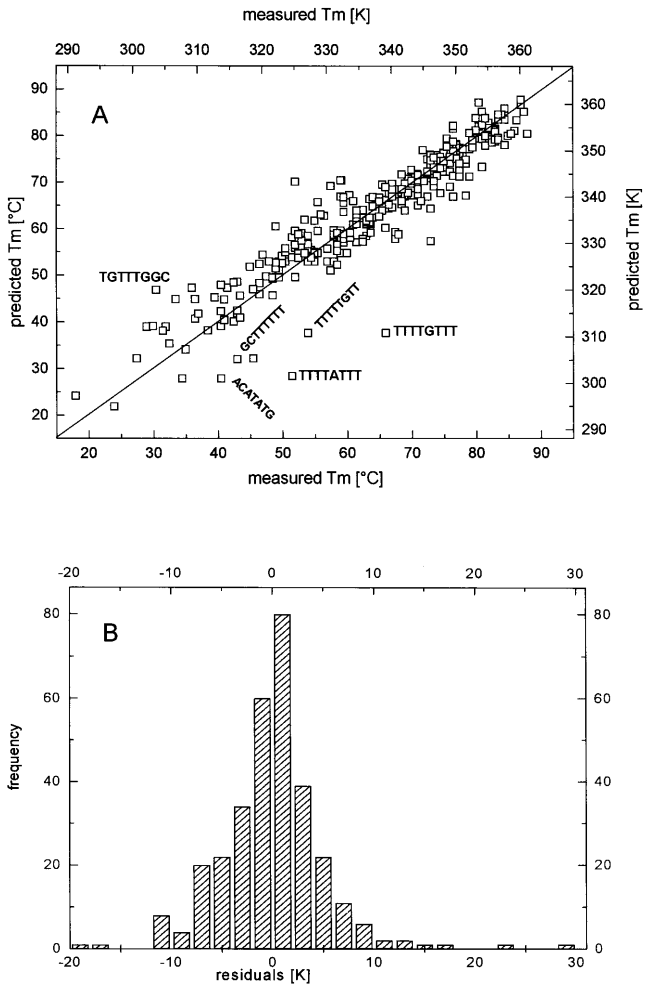

Figure 1. (**A**) Comparison between T_ms predicted by empirical linear model and measured T_ms. Most of the measured T_ms are predicted ±10 K. (**B**) Distribution of residuals of model data set. Eighty-one percent of the data yield residuals ≤5 K. T_m determination: absorbance versus temperature profiles were obtained at 260 nm with a Gilford Response spectrophotometer. Quartz cells had path lengths of 10 mm. The sample was heated at 0.5°C per step (~0.7°C/min) in 10 mM phosphate buffer pH 7, 100 mM NaCl, 0.1 mM EDTA. Total single strand concentration was 4 µM so that absorbance would be ~0.8–1. T_m was determined as the maximum of the first derivative of the absorbance versus temperature curve.

The correlation between the measured and the calculated T_ms for the data set used in the model building is presented in Figure 1. These results show that 81% of the data yield residuals <5 K. It is noted that outliers become more prominent below 320 K, and that the majority of these are very pyrimidine rich (75–96%) in the PNA strand, which gives high propensity for forming thermostable triplexes.

The predictive power of the above derived formula was verified using an independent data set (not employed for model building) comprising the T_ms of 44 different PNA/DNA duplexes. Figure 2 shows the result: >90% of the sequences are predicted within 5 K, and 98% of the predicted T_ms differ by not more than 10 K from the experimentally determined T_m.

The present results indicate that there is a physico-chemical explanation behind the two modifying functions (which also

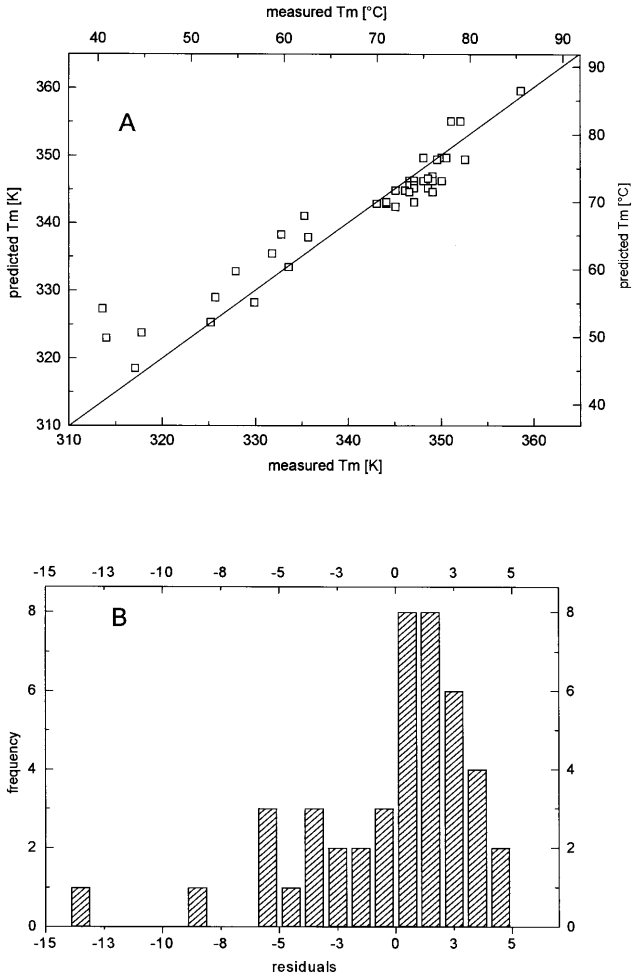

Figure 2. Validation of T_m formula. (**A**) Comparison of measured versus predicted T_ms of the independent challenge data set. (**B**) Distribution of residuals.

implies stability of the formula with respect to its predictive power). Length probably models the asymptotic behaviour of the melting temperature above 363 K. The 'length' parameter functions as an analogue to a more elegant solution which would weight the DNA nearest neighbour contributions entering the $T_{m_{nnDNA}}$ according to the number of nearest neighbour bases: the more interactions there are, the smaller the incremental contribution of the nearest neighbour base.

The fractional pyrimidine content almost certainly reflects wrong assignment of ΔH^0 and ΔS^0 values to the PNA/DNA nearest neighbours and further contributions due to the asymmetry of the molecule. Thus it should be possible by doing the reverse exercise to determine the 'missing' ΔH^0 and ΔS^0 values of the PNA/DNA nearest neighbours, and work along this line is in progress.

Finally, the small salt dependence of PNA should ensure that the T_m formula is valid over a broad salt concentration range, and that incorporation of an ionic correction factor should be rather trivial.

During the preparation of this manuscript a paper describing an approach to predicting PNA–DNA duplex stabilities based on DNA–DNA nearest neighbour ΔG^0 values was published (15). However, this model only relies on 11 measured T_ms of which 10 are 9mers and it is tested on an additional two complexes. Furthermore, this model does not take the very important aspect of PNA–DNA duplex asymmetry into account.

ACKNOWLEDGEMENTS

Ane Lester and Annette Jørgensen are thanked for technical assistance; Dr Frank Bergmann for providing part of the data set, and Dr Christoph Kessler for helpful discussions.

REFERENCES

1 Nielsen,P.E., Egholm,M., Berg,R.H. and Buchardt,O. (1991) *Science*, **254**, 1497–1500.
2 Nielsen,P.E. and Ørum,H. (1995) *Molecular Biology: Current Innovations and Future Trends*, Part 2, pp. 73–89.
3 Dueholm,K. and Nielsen,P.E. (1996) *New J. Chem.*, **21**, 19–31.
4 Good,L. and Nielsen,P.E. (1997) *Antisense Nucleic Acid Drug Dev.*, **7**, 431–437.
5 Egholm,M., Buchardt,O., Christensen,L., Behrens,C., Freier,S.M., Driver,D.A., Berg,R.H., Kim,S.K., Nordén,B. and Nielsen,P.E. (1993) *Nature*, **365**, 556–568.
6 Tomac,S., Sarkar,M., Ratilainen,T., Wittung,P., Nielsen,P.E., Nordén,B. and Gräslund,A. *J. Amer. Chem. Soc.*, **118**, 5544–52.
7 Marky,L.A. and Breslauer,K.J. (1987) *Biopolymers*, **26**, 1601–1620; .
8 Breslauer,K.J., Frank,R., Blocker,H. and Marky,L.A. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.
9 SantaLucia,J., Allawi,H.T. and Seneviratne,P.A. (1995) *Biochemistry*, **35**, 3555–3562.
10 Sugimoto,N., Nakano,S., Yoneyama,M. and Honda,K. (1996) *Nucleic Acids Res.*, **24**, 4501–4505.
11 Marmur,J. and Doty,P. (1962) *J. Mol. Biol.*, **5**, 109.
12 Eriksson,M. and Nielsen,P.E. (1996) *Nat. Struct. Biol.*, **3**, 410–413.
13 Eriksson,M. and Nielsen,P.E. (1996) *Quart. Rev. Biophysics*, **29**,369–394.
14 Nielsen,P.E. and Christensen,L. (1996) *J. Amer. Chem. Soc.*, **118**, 2287–88.
15 Griffin,T.J. and Smith,L.M. (1998) *Anal. Biochem.*, **260**, 56–63.