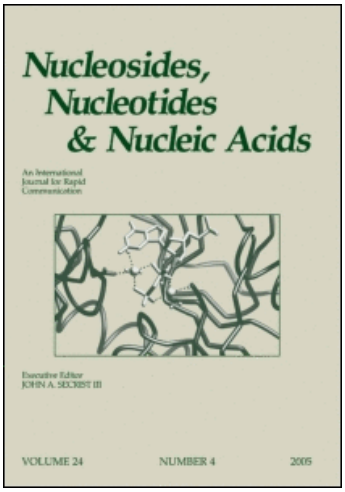

Nucleosides, Nucleotides and Nucleic Acids
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597286

A NEURAL NETWORK TO PREDICT MELTING TEMPERATURE (T_m) OF RNA DUPLEX

Changmei Cheng[a]; Lixin[b]; Hiromi Miyajima[b]; Piet Herdewijn[a]
[a] Laboratory of Medicinal Chemistry, Rega Institute, Catholic University of Leuven, Leuven, Belgium [b] Department of Information Science, Faculty of Engineering, Kagoshima University, Kagoshima City, Japan

Online publication date: 28 February 2001

To cite this Article Cheng, Changmei , Lixin, Miyajima, Hiromi and Herdewijn, Piet(2001) 'A NEURAL NETWORK TO PREDICT MELTING TEMPERATURE (T_m) OF RNA DUPLEX', Nucleosides, Nucleotides and Nucleic Acids, 20: 3, 261 — 269
To link to this Article: DOI: 10.1081/NCN-100002086
URL: http://dx.doi.org/10.1081/NCN-100002086

NUCLEOSIDES, NUCLEOTIDES & NUCLEIC ACIDS, 20(3), 261–269 (2001)

A NEURAL NETWORK TO PREDICT MELTING TEMPERATURE (T_m) OF RNA DUPLEX

Changmei Cheng,[1] Lixin Ma,[2] Hiromi Miyajima,[2] and Piet Herdewijn[1,*]

[1]Laboratory of Medicinal Chemistry, Rega Institute, Catholic University of Leuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium
[2]Department of Information Science, Faculty of Engineering, Kagoshima University, Korimoto 1-21-40, 890 Kagoshima City, Japan

ABSTRACT

A straightforward method to predict RNA duplex stability by neural network is described. The best network consists of three layers in which the input layer units are 12 (frequencies of 10 nearest-neighbors and 2 terminals), the hidden layer units are 3 and the output layer unit is 1 (measured T_m). This method to predict T_m has the advantage that the determinations of thermodynamic parameters is not needed.

Melting temperature (T_m) determination is the easiest and fastest method to evaluate the stability of a RNA duplex. It is the first experiment that is carried out when a oligonucleotide is evaluated for its potential use as antisense construct or in other molecular biological studies. The predictability of the stability of an RNA duplex may be of utmost importance in designing sequences for microarrays. A procedure to predict T_m based on thermodynamic parameters (TP) of nearest-neighbor model has been established and improved several times (1). This method has been used succesfully for predicting the stability of oligonucleotide duplexes (2). The prediction of the stability of modified oligonucleotides such as phosphorothioate oligomers or peptide-nucleic acid (PNA) has not been carried out using this methodology because the thermodynamic data of this kind of helixes are not

*Corresponding author. Fax: +32-16-33 73 87; E-mail: Piet.Herdewijn@rega.kuleuven.ac.be

available yet. Prediction of T_ms (and even ΔG) using neural network might become an interesting alternative in the future. Neural network (NN) has been proven to be a pattern recognition methodology with excellent capabilities to predict functional properties based on some well selected parameters (3). The method consists in collecting large numbers of known data, followed by training and testing of a model, and finally, use it to predict the results of similar experiments. This method has been applied in chemistry, for example, in structure-activity relationship studies. Therefore we investigated the possibility to use NN for the predictions of T_ms of RNA duplexes directly from the sequences. The predictions, of course, are limited to well defined secondary studies in well defined circumstances. This communication might be considered as a proof of concept of this new methodology for T_m prediction.

EXPERIMENTAL

The input data (frequencies of nearest-neighbors and terminals) of the network are given in Table 1–3 and were obtained from the sequences by simple calculations following the method described in literature (1). All the supervisor data (measured T_ms) were obtained from literature (4). The computations were performed in Silicon Graphics O2 R5000 work station using the artificial neural network program developed by the group of Hiromi Miyajima (5).

RESULTS AND DISCUSSIONS

In order to be able to compare the NN method and TP method, we selected the same 109 sequences used in TP method (4) and divided them into 90 training sequences and 19 testing sequences. In our neural network, back-propagation (BP) (5,6) algorithm is used to minimize the mean squared error between the desired and actual outputs of the network. The activation function to be used in those networks is a sigmoid function. The 10 nearest-neighbors and 2 terminals are defined as in INN-HB model (1,4) and the frequencies of them for a given sequence are used as input data of the network. The symmetry effect and terminal phosphate groups present in some oligomers are ignored.

We constructed 4 three-layered traditional neural networks in which the input units are 12, the output unit is 1 (T_m) and the hidden layer units are 5, 4, 3, and 2 respectively (called 5H, 4H, 3H and 2H network) (Fig. 1). The training and testing processes are performed as followed. a) Measured data are divided into two groups of training data and testing data. Among the 109 oligomers, 99 are normal two-state-transition sequences and 10 are non-two-state-transition sequences. We used 90 two-state-transition sequences for training and the others for testing. b) In the process of training those networks, the initial weight values are set as random real numbers and back propagation (BP) method is used to optimize the final weights

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table 1. Frequencies of Nearest Neighbor and Terminals of RNA

	(Two-State Sequences for Training)										Terminals	
Sequences	AA UU	AU UA	CG GC	CU GA	GA CU	GC CG	GG CC	GU CA	UA AU	UG AC	A U	C G
CCGG	0	0	1	0	0	0	2	0	0	0	0	2
CGCG	0	0	2	0	0	1	0	0	0	0	0	2
GCGC	0	0	1	0	0	2	0	0	0	0	0	2
GGCC	0	0	0	0	0	1	2	0	0	0	0	2
ACGCA/	0	0	1	0	0	1	0	1	0	1	2	0
AGCGA/	0	0	1	1	1	1	0	0	0	0	2	0
CACAG/	0	0	0	1	0	0	0	1	0	2	0	2
GCACG/	0	0	1	0	0	1	0	1	0	1	0	2
GCUCG/	0	0	1	1	1	1	0	0	0	0	0	2
ACCGGUp	0	0	1	0	0	0	2	2	0	0	2	0
AGCGCU	0	0	1	2	0	2	0	0	0	0	2	0
AGGCCU	0	0	0	2	0	1	2	0	0	0	2	0
CACGUG	0	0	1	0	0	0	0	2	0	2	0	2
CAGCUGp	0	0	0	2	0	1	0	0	0	2	0	2
CCAUGG	0	1	0	0	0	0	2	0	0	2	0	2
CCGCGG	0	0	2	0	0	1	2	0	0	0	0	2
CCUAGG	0	0	0	2	0	0	2	0	1	0	0	2
CGCGCGp	0	0	3	0	0	2	0	0	0	0	0	2
CGGCCGp	0	0	2	0	0	1	2	0	0	0	0	2
CUGCAGp	0	0	0	2	0	1	0	0	0	2	0	2
GACGUC	0	0	1	0	2	0	0	2	0	0	0	2
GAGAGA/	0	0	0	2	3	0	0	0	0	0	1	1
GAGCUC	0	0	0	2	2	1	0	0	0	0	0	2
GAGGAG/	0	0	0	2	2	0	1	0	0	0	0	2
GCAACG/	1	0	1	0	0	1	0	1	0	1	0	2
GCAUCG/	0	1	1	0	1	1	0	0	0	1	0	2
GCAUGC/	0	1	0	0	0	2	0	0	0	2	0	2
GCCGCG/	0	0	2	0	0	2	1	0	0	0	0	2
GCCGGCp	0	0	1	0	0	2	2	0	0	0	0	2
GCGCCG/	0	0	2	0	0	2	1	0	0	0	0	2
GCGCGCp	0	0	2	0	0	3	0	0	0	0	0	2
GCGCGG/	0	0	2	0	0	2	1	0	0	0	0	2
GCGGCG/	0	0	2	0	0	2	1	0	0	0	0	2
GCGUCG/	0	0	2	0	1	1	0	1	0	0	0	2
GCUACG/	0	0	1	1	0	1	0	1	1	0	0	2
GCUAGC	0	0	0	2	0	2	0	0	1	0	0	2
GGAUCC	0	1	0	0	2	0	2	0	0	0	0	2
GGCGCCp	0	0	1	0	0	2	2	0	0	0	0	2
GGCGCG/	0	0	2	0	0	2	1	0	0	0	0	2
GGUACC	0	0	0	0	0	0	2	2	1	0	0	2
GUCGAC	0	0	1	0	2	0	0	2	0	0	0	2
GUGCAC	0	0	0	0	0	1	0	2	0	2	0	2
GUGGUG/	0	0	0	0	0	0	1	2	0	2	0	2
GUGUCG/	0	0	1	0	1	0	0	2	0	1	0	2
UCAUGA	0	1	0	0	2	0	0	0	0	2	2	0

(*continued*)

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table 1. Continued

	(Two-State Sequences for Training)										Terminals	
Sequences	AA UU	AU UA	CG GC	CU GA	GA CU	GC CG	GG CC	GU CA	UA AU	UG AC	A U	C G
UCCGGAp	0	0	1	0	2	0	2	0	0	0	2	0
UCGCGA	0	0	2	0	2	1	0	0	0	0	2	0
UCUAGA	0	0	0	2	2	0	0	0	1	0	2	0
UGAUCA	0	1	0	0	2	0	0	0	0	2	2	0
UGCGCA	0	0	1	0	0	2	0	0	0	2	2	0
AAGGAGG/	1	0	0	2	1	0	2	0	0	0	1	1
ACUGUCA/	0	0	0	1	1	0	0	2	0	2	2	0
AGUCUGA/	0	0	0	2	2	0	0	1	0	1	2	0
GACUCAG/	0	0	0	2	2	0	0	1	0	1	0	2
GAGUGAG/	0	0	0	2	2	0	0	1	0	1	0	2
GUCACUG/	0	0	0	1	1	0	0	2	0	2	0	2
AACUAGUU	2	0	0	2	0	0	0	2	1	0	2	0
AAUGCAUUp	2	2	0	0	0	1	0	0	0	2	2	0
ACCUUUGC/	2	0	0	1	0	1	1	1	0	1	1	1
ACUAUAGU	0	1	0	2	0	0	0	2	2	0	2	0
ACUUAAGU	2	0	0	2	0	1	0	1	1	0	2	0
AGAGAGAG/	0	0	0	4	3	0	0	0	0	0	1	1
AGAUAUCU	0	2	0	2	2	0	0	0	1	0	2	0
AGUUAACU	2	0	0	2	0	0	0	2	1	0	2	0
AUACGUAU	0	2	1	0	0	0	0	2	2	0	2	0
AUCUAGAU	0	2	0	2	2	0	0	0	1	0	2	0
AUGCGCAUp	0	2	1	0	0	2	0	0	0	2	2	0
AUGUACAUp	0	2	0	0	0	0	0	2	1	2	2	0
CAAAAAAG/	5	0	0	1	0	0	0	0	0	1	0	2
CAUGCAUGp	0	2	0	0	0	1	0	0	0	4	0	2
CGACGCA/	0	0	2	1	1	1	0	1	0	1	0	2
CUCGCACA/	0	0	1	1	1	1	0	1	0	2	1	1
GAACGUUC	2	0	1	0	2	0	0	2	0	0	0	2
GAUAUAUC	0	3	0	0	2	0	0	0	2	0	0	2
GAUGCAUCp	0	2	0	0	2	1	0	0	0	2	0	2
GGCUUCAA/	2	0	0	1	1	1	1	0	0	1	1	1
GUAUAUAC	0	2	0	0	0	0	0	2	3	0	0	2
GUCUAGAC	0	0	0	2	2	0	0	2	1	0	0	2
GUUCGAAC	2	0	1	0	2	0	0	2	0	0	0	2
UAGAUCUA	0	1	0	2	2	0	0	0	2	0	2	0
UAUGCAUAp	0	2	0	0	0	1	0	0	2	2	2	0
UCCUUGCA/	1	0	0	1	1	1	1	0	0	2	2	0
UCUAUAGA	0	1	0	2	2	0	0	0	2	0	2	0
UGACCUCA/	0	0	0	1	2	0	1	1	0	2	2	0
UUCCGGAA	2	0	1	0	2	0	2	0	0	0	2	0
UUGCGCAA	2	0	1	0	0	2	0	0	0	2	2	0
UUGGCCAA	2	0	0	0	0	1	2	0	0	2	2	0
UUGUACAA	2	0	0	0	0	0	0	2	1	2	2	0
CAAAAAAA	6	0	0	1	0	0	0	0	0	1	0	2
AAGGUUGG	3	0	0	1	1	0	2	1	0	1	2	0

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table 2. Frequencies of Nearest Neighbors and Terminals of RNA

	(Two-State Sequences for Testing)										Terminals	
Sequences	AA	AU	CG	CU	GA	GC	GG	GU	UA	UG	A	C
CAUGCG/	0	1	1	0	0	1	0	0	0	2	0	2
GAGCUG/	0	0	0	2	1	1	0	0	0	1	0	2
GCUGAG/	0	0	0	2	1	1	0	0	0	1	0	2
GUGCAG/	0	0	0	1	0	1	0	1	0	2	0	2
UAAGGUA/	1	0	0	1	0	0	1	1	2	0	2	0
GAGAUCUC	0	1	0	2	4	0	0	0	0	0	0	2
GCCAUGGC	0	1	0	0	0	2	2	0	0	2	0	2
GCUGCGAC/	0	0	1	1	1	2	0	1	0	1	0	2
UCCGCGCA/	0	0	2	0	1	2	1	0	0	1	2	0

Table 3. Frequencies of Nearest Neighbors and Terminals of RNA

	(Non-Two-State Sequences for Testing)										Terminals	
Sequences	AA	AU	CG	CU	GA	GC	GG	GU	UA	UG	A	C
CACUG/	0	0	0	1	0	0	0	1	0	2	0	2
AGAGAG/	0	0	0	3	2	0	0	0	0	0	1	1
AUGCAUp	0	2	0	0	0	1	0	0	0	2	2	0
CGUACG	0	0	2	0	0	0	0	2	1	0	0	2
UGGCCAp	0	0	0	0	0	1	2	1	0	1	2	0
GCAACGA/	1	0	1	0	1	1	0	1	0	1	1	1
AGUAUACU	0	1	0	2	0	0	0	2	2	0	2	0
GAGAGAGA/	0	0	0	3	4	0	0	0	0	0	1	1
GUGAUCAC	0	1	0	0	2	0	0	2	0	2	0	2
A_7U_7	12	1	0	0	0	0	0	0	0	0	2	0

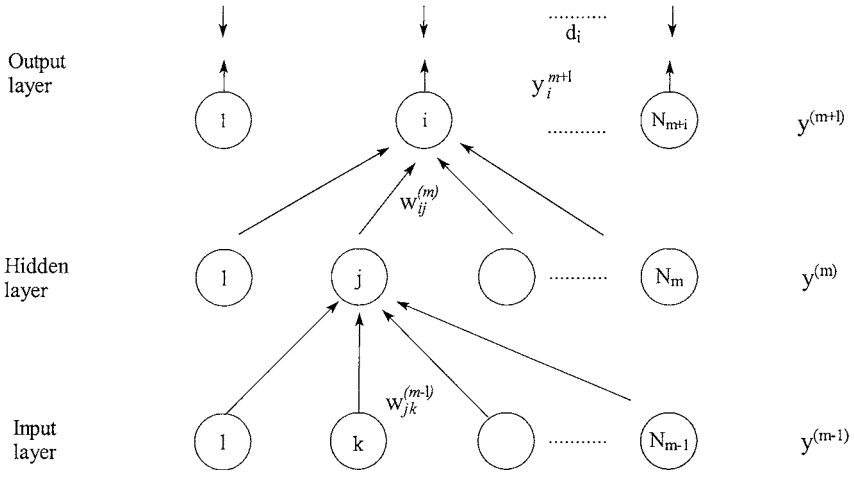

Figure 1. Three-layer networks.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table 4. Predictions of T_m by NN and TP Method

Sequence	Measured T_m (K)	Predicted T_m by NN (K)	Predicted T_m by TP (K)
Two-State-Transition Sequences for Training			
CCGG	300.35	302.12 (−1.77)	298.45 (1.9)
CGCG	292.45	296.16 (−3.71)	291.95 (0.5)
GCGC	299.75	300.62 (−0.87)	302.25 (−2.5)
GGCC	307.45	308.70 (−1.25)	308.05 (−0.6)
ACGCA/	302.55	302.70 (−0.15)	302.15 (0.4)
AGCGA/	303.35	303.27 (0.08)	303.05 (0.3)
CACAG/	297.65	295.61 (2.04)	297.25 (0.4)
GCACG/	310.65	308.88 (1.77)	309.85 (0.8)
GCUCG/	310.35	309.96 (0.39)	310.55 (−0.2)
ACCGGUp	327.05	324.16 (2.89)	324.95 (2.1)
AGCGCU	325.15	324.78 (0.37)	324.75 (0.4)
AGGCCU	328.45	328.00 (0.45)	329.05 (−0.6)
CACGUG	315.95	315.10 (0.85)	315.55 (0.4)
CAGCUGp	316.25	317.06 (−0.81)	319.75 (−3.5)
CCAUGG	319.55	320.57 (−1.02)	319.75 (−0.2)
CCGCGG	332.95	334.28 (−1.33)	335.35 (−2.4)
CCUAGG	323.15	322.11 (1.04)	321.25 (1.9)
CGCGCGp	330.95	332.50 (−1.55)	331.65 (−0.7)
CGGCCGp	336.35	334.28 (2.07)	335.35 (1.0)
CUGCAGp	318.45	317.06 (1.39)	319.75 (−1.3)
GACGUC	319.35	317.54 (1.81)	318.25 (1.1)
GAGAGA/	313.75	311.72 (2.03)	313.75 (0.0)
GAGCUC	321.85	321.71 (0.14)	322.15 (−0.3)
GAGGAG/	324.05	322.45 (1.60)	321.35 (2.7)
GCAACG/	315.75	313.87 (1.88)	315.75 (0.0)
GCAUCG/	317.25	317.48 (−0.23)	317.05 (0.2)
GCAUGC	318.85	318.89 (−0.04)	321.45 (−2.6)
GCCGCG/	337.05	335.19 (1.86)	335.85 (1.2)
GCCGGCp	340.35	337.33 (3.02)	339.75 (0.6)
GCGCCG/	338.35	335.19 (3.16)	335.85 (2.5)
GCGCGCp	335.25	334.39 (0.86)	336.25 (−1.0)
GCGCGG/	335.05	335.19 (−0.14)	335.85 (−0.8)
GCGGCG/	334.95	335.19 (−0.24)	335.85 (−0.9)
GCGUCG/	326.85	326.67 (0.18)	325.05 (1.8)
GCUACG/	318.15	319.24 (−1.09)	318.05 (0.1)
GCUAGC	322.45	321.72 (0.73)	322.95 (−0.5)
GGAUCC	320.75	321.38 (−0.63)	322.05 (−1.3)
GGCGCCp	338.35	337.33 (1.02)	339.75 (−1.4)
GGCGCG/	334.75	335.19 (−0.44)	335.85 (−1.1)
GGUACC	319.75	320.77 (−1.02)	323.05 (−3.3)
GUCGAC	318.45	317.54 (0.91)	318.25 (0.2)
GUGCAC	320.85	320.07 (0.78)	321.55 (−0.7)
GUGGUG/	320.55	319.87 (0.68)	320.75 (−0.2)
GUGUCG/	316.85	316.41 (0.44)	316.95 (−0.1)
UCAUGA	300.35	304.35 (−4.00)	302.65 (−2.3)

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table 4. Continued

UCCGGAp	323.25	326.39 (−3.14)	325.85 (−2.6)
UCGCGA	317.75	319.10 (−1.35)	321.45 (−3.7)
UCUAGA	304.15	302.56 (1.59)	303.65 (0.5)
UGAUCA	305.75	304.35 (1.40)	302.65 (3.1)
UGCGCA	326.25	323.71 (2.54)	325.15 (1.1)
AAGGAGG/	329.35	328.03 (1.32)	328.25 (1.1)
ACUGUCA/	321.35	320.58 (0.77)	321.85 (−0.5)
AGUCUGA/	318.85	321.03 (−2.18)	322.15 (−3.3)
GACUCAG/	325.15	327.12 (−1.97)	325.55 (−0.4)
GAGUGAG/	326.85	327.12 (−0.27)	325.55 (1.3)
GUCACUG/	324.25	326.16 (−1.91)	325.35 (−1.1)
AACUAGUU	318.85	316.61 (2.24)	314.35 (4.5)
AAUGCAUUp	318.15	315.09 (3.06)	313.25 (4.9)
ACCUUUGC/	329.45	332.12 (−2.67)	331.55 (−2.1)
ACUAUAGU	317.15	316.71 (0.44)	317.35 (−0.2)
ACUUAAGU	313.45	320.06 (−6.61)	314.35 (−0.9)
AGAGAGAG/	332.75	334.48 (−1.73)	332.05 (0.7)
AGAUAUCU	314.55	316.33 (−1.78)	316.85 (−2.3)
AGUUAACU	314.25	313.37 (0.88)	314.35 (−0.1)
AUACGUAU	315.15	315.28 (−0.13)	313.55 (1.6)
AUCUAGAU	318.25	316.33 (1.92)	316.85 (1.4)
AUGCGCAUp	333.45	333.43 (0.02)	331.85 (1.6)
AUGUACAUp	314.85	315.88 (−1.03)	316.15 (−1.3)
CAAAAAAG/	301.75	300.94 (0.81)	302.05 (−0.3)
CAUGCAUGp	328.05	329.78 (−1.73)	327.65 (0.4)
CGACGCAG/	340.25	339.83 (0.42)	339.85 (0.4)
CUCGCACA/	338.05	337.32 (0.73)	337.95 (0.1)
GAACGUUC	325.45	324.33 (1.12)	325.65 (−0.2)
GAUAUAUC	312.25	310.97 (1.28)	312.55 (−0.3)
GAUGCAUCp	330.35	332.18 (−1.83)	329.05 (1.3)
GGCUUCAA/	332.25	333.12 (−0.87)	331.75 (0.5)
GUAUAUAC	311.45	311.57 (−0.12)	312.75 (−1.3)
GUUCGAAC	323.55	324.34 (−0.79)	325.65 (−2.1)
UAGAUCUA	318.45	316.93 (1.52)	318.25 (0.2)
UCCUUGCA/	333.85	336.82 (−2.97)	335.55 (−1.7)
UCUAUAGA	316.65	316.93 (−0.28)	318.25 (−1.6)
UGACCUCA/	337.75	335.09 (2.66)	335.95 (1.8)
UUCCGGAA	335.55	334.60 (0.95)	332.25 (3.3)
UUGCGCAA	334.35	333.78 (0.57)	331.75 (2.6)
UUGGCCAA	338.45	336.53 (1.92)	335.25 (3.2)
UUGUACAA	316.75	317.52 (−0.77)	314.65 (2.1)
CAAAAAAAG/	306.95	307.37 (−0.42)	307.85 (−0.9)
AAGGUUGGAA/	339.65	340.08 (−0.43)	339.35 (0.3)
Two-State-Transition Sequences for Testing			
CAUGCG/	316.05	315.37 (0.68)	315.65 (0.4)
GAGCUG/	318.65	319.45 (−0.8)	321.05 (−2.4)
GCUGAG/	319.35	319.45 (−0.10)	321.05 (−1.7)
GUGCAG/	319.15	318.74 (0.41)	320.65 (−1.5)

(*continued*)

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Table 4. Continued

Sequence	Measured T_m (K)	Predicted T_m by NN (K)	Predicted T_m by TP (K)
Two-State-Transition Sequences for Training			
UAAGGUA/	315.35	314.42 (0.93)	310.55 (4.8)
GAGAUCUC	329.65	333.15 (−3.50)	329.55 (0.1)
GCCAUGGC	344.55	342.03 (2.52)	345.35 (−0.8)
GCUGCGAC/	341.05	340.96 (0.09)	343.15 (−2.1)
UCCGCGCA/	346.35	341.02 (5.33)	347.45 (−1.1)
Non-Two-State-Transition Sequences for Testing			
CACUG/	289.55	295.61 (−6.06)	297.55 (−8.0)
AGAGAG/	313.85	311.97 (1.88)	312.05 (1.8)
AUGCAUp	303.25	301.12 (2.13)	301.25 (2.0)
CGUACG	307.75	312.48 (−4.73)	312.55 (−4.8)
UGGCCAp	328.45	328.18 (0.27)	329.45 (−1.0)
GCAACGA/	323.35	324.29 (−0.94)	325.45 (−2.1)
AGUAUACU	317.25	316.71 (0.54)	317.35 (−0.1)
GAGAGAGA/	330.75	334.09 (−3.34)	332.75 (−2.0)
GUGAUCAC	327.55	330.89 (−3.34)	329.35 (−1.8)
A_7U_7	314.15	311.54 (2.61)	314.55 (−0.4)

while the supervisor is the measured T_m data and the mean square error (MSE) is set as 0.000025. It was found that 3H network (46 weights) gives the best result whereas 5H, 4H and 2H networks take more time to converge and the predicting average deviations are larger than for 3H (data not shown). c) By using the frozen weights obtained from the best network, the ability to predict T_m data of dsRNA sequences was evaluated. Again, the 3H network shows the best accuracy. In order to test the repeatability of 3H network, we undertook the training and testing processes twice using random real numbers as initial weights and obtained very similar results. The predicted data are listed in Table 4. The data for prediction of T_m using TP method are taken from literature (4). The data in parentheses are the deviations between measured temperatures and predicted temperatures (measured T_m-predicted T_m).

In the above training and testing procedure, the melting temperatures are expressed in degree Kelvin (K). When we performed the same process but expressing the T_m data in degree Centigrade (°C) in 3H network, the testing accuracy becomes very bad although the training accuracy is good. It demonstrates that the quantitative relationship between oligomer sequence and T_m only exists when the thermal stability of the duplexes is described in degree Kelvin. This conclusion is in agreement with the results from the TP method. The 3H neural network analysis seems to give reliable results in very good agreement with experimental data.

As shown in Table 4, all T_m's of the training set, predicted by the NN method, agree very well with the experimental data [average deviation (AD) = 1.34 K]. The accuracy of the data for the testing sequences are of similar quality (AD = 1.59 K for two-state-transition sequences and AD = 2.58 K for non-two-state-transition

sequences respectively). In the TP method, the above ADs are 1.28 K, 1.67 K and 2.40 K respectively. The results obtained for both methods are very similar.

CONCLUSIONS

We provide here a simple method to predict the melting temperatures directly from the sequences of dsRNA. The results, of course, can only be used for the given strand concentration and salt concentration. It is, likewise, not a method to predict RNA secondary structures. Indeed, it is well documented that the neural network method can be used to obtain a good prediction for a known trained sequence, but that it is not usefull for the prediction of data generated in other circumstances. Therefore, the method has rigorous limitations. This method may, however, be applied for the prediction of T_ms of longer dsRNA sequences and for T_m prediction of dsDNA and DNA/RNA duplexes. In addition, this methodology might be especially useful for predicting T_m of duplexes where one or both strands consists of modified oligonucleotides such as phosphorothioates and PNA because the thermodynamic parameters for these oligomers are not available yet.

REFERENCES

1. Freier, S.; Kierzek, M.R.; Jaeger, J.A.; Sugimoto, N.; Caruthers, M.H.; Neilson, T.; Turner, D.H. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 9373.
2. Mathews, D.J.; Sabina, J.; Zuker, M.; Turner, D.J. *J. Mol. Biol.* **1999**, *288*, 911–940.
3. Zupan, J.; Gasteiger, J. *Neural Networks for Chemistry* **1993**, VCH, Cambridge.
4. Xia, T.; SantaLucia, J. Jr.; Burkard, M.E.; Kierzek, M.R.; Schroeder, S.J.; Jiao, X.; Cox, C.; Turner, D.H. *Biochemistry* **1998**, *37*, 14719.
5. Ma, L.; Kawabata, S.; Miyajima, H. *NOLTA'98*, **1998** *Switzerland*, Sept. 507.
6. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. *Nature* **1986**, *323*, 533.

Received October 2, 2000
Accepted December 5, 2000

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Request Permission or Order Reprints Instantly!