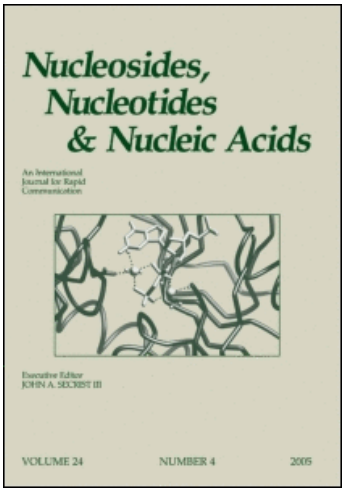

Nucleosides, Nucleotides and Nucleic Acids
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597286

A Neural Network for Predicting the Stability of DNA/DNA Duplexes

Xiaohong Liu[a]; Lixin[b]; Changmei Cheng[a]; Yong Wang[b]; Hiromi Miyajima[a]; Yufen Zhao[a]
[a] Key Laboratory of Bioorganic Phosphorus Chemistry, Ministry of Education, Department of Chemistry, School of Life Sciences and Engineering, Tsinghua University, Beijing, P. R. China [b] Artificial Intelligence Institute, Shanghai University of Electric Power, Shanghai, P. R. China

To cite this Article Liu, Xiaohong , Lixin, Cheng, Changmei , Wang, Yong , Miyajima, Hiromi and Zhao, Yufen(2005) 'A Neural Network for Predicting the Stability of DNA/DNA Duplexes', Nucleosides, Nucleotides and Nucleic Acids, 24: 3, 199 — 209
To link to this Article: DOI: 10.1081/NCN-200055718
URL: http://dx.doi.org/10.1081/NCN-200055718

Nucleosides, Nucleotides, and Nucleic Acids, 24 (3):199–209, (2005)

ISSN: 1525-7770 print/ 1532-2335 online
DOI: 10.1081/NCN-200055718

Taylor & Francis
Taylor & Francis Group

A NEURAL NETWORK FOR PREDICTING THE STABILITY OF DNA/DNA DUPLEXES

Xiaohong Liu ▫ *Key Laboratory of Bioorganic Phosphorus Chemistry, Ministry of Education, Department of Chemistry, School of Life Sciences and Engineering, Tsinghua University, Beijing, P. R. China*

Lixin Ma ▫ *Artificial Intelligence Institute, Shanghai University of Electric Power, Shanghai, P. R. China*

Changmei Cheng ▫ *Key Laboratory of Bioorganic Phosphorus Chemistry, Ministry of Education, Department of Chemistry, School of Life Sciences and Engineering, Tsinghua University, Beijing, P. R. China*

Yong Wang ▫ *Artificial Intelligence Institute, Shanghai University of Electric Power*

Hiromi Miyajima ▫ *Faculty of Engineering, Kagoshima University Korimoto, Kagoshima, Japan*

Yufen Zhao ▫ *Key Laboratory of Bioorganic Phosphorus Chemistry, Ministry of Education, Department of Chemistry, School of Life Sciences and Engineering, Tsinghua University, Beijing, P. R. China*

▫ *A back-propagation neural network method has been developed to predict the stability of DNA/DNA duplexes. Calculated T_m with the present parameters fits the experimental values within reasonable errors (AD = 1.86 K, SEP = 1.99151, R2 = 0.9894 for NN1; AD = 1.59667 K, SEP = 2.03824, R2 = 0.99371 for NN2), and it has the advantage that the determinations of thermodynamic parameters are not needed.*

Keywords Melting Temperature, DNA Duplex, Neural Network

Received 1 September 2004, accepted 31 January 2005.

This work was supported by the ShangHai Natural Science Foundation (No.02ZF14049) and Chinese National Natural Science Foundation (No. 20132020 and No. 20175026).

Address correspondence to Changmei Cheng, Key Laboratory of Bioorganic Phosphorus Chemistry, Ministry of Education, Department of Chemistry, School of Life Sciences and Engineering, Tsinghua University, Beijing 100084, P. R. China; Fax: ++ 86-10-62781659; E-mail: chengcm@mail.tsinghua.edu.cn

INTRODUCTION

Accurate prediction of DNA thermal denaturation is important for several molecular biological techniques including PCR,[1] sequencing by hybridization,[2] antigene targeting,[3] and Southern blotting.[4] In these techniques, choice of a nonoptimal sequence or temperature can lead to amplification or detection of wrong sequences.[5] Furthermore, knowledge of the sequence dependence of DNA stability is important for understanding the details of DNA replication, mutation, repair, and transcription.[6,7] Melting temperature (T_m) determination is the easiest and fastest method to evaluate the stability of a duplex, and a procedure to predict T_m of RNA/DNA,[8] RNA/RNA,[9] and DNA/DNA[10,11] based on thermodynamic parameters (TP) of a nearest-neighbor model has been established. Although this method can predict the melting temperature in good accuracy, it is obvious that the determination of enthalpy and entropy changes is very troublesome and time consuming. This drawback also limits its application to the prediction of modified nucleotides such as phosphorothioate oligomer (S-oligomer) and peptide nucleic acid (PNA), because the thermodynamic data for these helixes are not available yet. On the other hand, the neural network (NN) has been developed as a successful pattern recognition methodology for predicting functional properties based on some well-selected parameters. This method has been applied to chemistry,[12] e.g., structure and activity relationship studies, which prompts us to apply the NN method to the prediction of T_ms of DNA duplexes directly from the sequences. These networks consist of a series of connecting layers made up of nodes which relay information to different layers. In our previous work, a prediction method of T_m value of RNA/RNA and DNA/RNA duplexes using neural networks (NN) has been developed.[13–15] In this article, we will report a study of prediction of the T_m of DNA duplexes using neural networks.

EXPERIMENTAL

Predicting Melting Temperature (T_m) of Oligo-2-deoxyribonucleotide Duplex by NN

The nearest neighbors are defined as in Xia et al.[16] For example, the nearest neighbors of the following sequence are indicated by the arrows:

```
  ↓   ↓   ↓   ↓
A G A T A T C T
T C T A T A G A
    ↑   ↑   ↑
```

Mathematically, there are 16 nearest neighbors when the four bases A, G, C, and T are used to make two-element arrays. However, in the case of DNA sequences, six pairs of these are identical (AA/TT = TT/AA, AG/TC = CT/GA, AC/TG = GT/CA, GA/CT = TC/AG, GG/CC = CC/GG, CA/GT = TG/AC), hence the number

TABLE 1 Frequencies of Nearest Neighbor and Terminals of DNA Duplex

		Sequences for training										Terminals	
		AA	AT	CG	CT	GA	GC	GG	GT	TA	TG	A	C
DNA sequences		TT	TA	GC	GA	CT	CG	CC	CA	AT	AC	T	G
1	AGCCG	0	0	1	1	0	1	1	0	0	0	1	1
2	ACCGCA	0	0	1	0	0	1	1	1	0	1	2	0
3	ATGCGC	0	1	1	0	0	2	0	0	0	1	1	1
4	CGGTGC	0	0	1	0	0	1	1	1	0	1	0	2
5	CGTGCC	0	0	1	0	0	1	1	1	0	1	0	2
6	TGCGCA	0	0	1	0	0	2	0	0	0	2	2	0
7	AATACCG	1	1	1	0	0	0	1	1	1	0	1	1
8	AGCCGTG	0	0	1	1	0	1	1	1	0	1	1	1
9	AGCTTCA	1	0	0	2	1	1	0	0	0	1	2	0
10	GGACTTA	1	0	0	1	1	0	1	1	1	0	1	1
11	CGATATCG	0	2	2	0	2	0	0	0	1	0	0	2
12	GAAGCTTC	2	0	0	2	2	1	0	0	0	0	0	2
13	GGAATTCC	2	1	0	0	2	0	2	0	0	0	0	2
14	CAAACAAAG	4	0	0	1	0	0	0	1	0	2	0	2
15	AAAAAAAA	7	0	0	0	0	0	0	0	0	0	2	0
16	AAGCGTAG	1	0	1	2	0	1	0	1	1	0	1	1
17	AATCCAGT	1	1	0	1	1	0	1	1	0	1	2	0
18	ACATATGT	0	2	0	0	0	0	0	2	1	2	2	0
19	ACCTAGTC	0	0	0	2	1	0	1	2	1	0	1	1
20	ACGACCTC	0	0	1	1	2	0	1	2	0	0	1	1
21	AGAGAGAG	0	0	0	4	3	0	0	0	0	0	1	1
22	AGCGTAAG	1	0	1	2	0	1	0	1	1	0	1	1
23	AGTCCTGA	0	0	0	2	2	0	1	1	0	1	2	0
24	ATGCGCAT	0	2	1	0	0	2	0	0	0	2	2	0
25	CACGGCTC	0	0	1	1	1	1	1	1	0	1	0	2
26	CCATATGG	0	2	0	0	0	0	2	0	1	2	0	2
27	CGCGTATA	0	1	2	0	0	1	0	1	2	0	1	1
28	CGCTGTAA	1	0	1	1	0	1	0	1	1	1	1	1
29	CTAGTGGA	0	0	0	2	1	0	1	1	1	1	1	1
30	CTCACGGC	0	0	1	1	1	1	1	1	0	1	0	2
31	CTGAGTCC	0	0	0	2	2	0	1	1	0	1	0	2
32	GAATATTC	2	2	0	0	2	0	0	0	1	0	0	2
33	GACTAGTC	0	0	0	2	2	0	0	2	1	0	0	2
34	GAGTACTC	0	0	0	2	2	0	0	2	1	0	0	2
35	GATTAATC	2	2	0	0	2	0	0	0	1	0	0	2
36	GCATATGC	0	2	0	0	0	2	0	0	1	2	0	2
37	GCCAGTTA	1	0	0	1	0	1	1	1	1	1	1	1
38	GGTGCCAA	1	0	0	0	0	1	2	1	0	2	1	1
39	GTCGAACA	1	0	1	0	2	0	0	2	0	1	1	1
40	GTCTAGAC	0	0	0	2	2	0	0	2	1	0	0	2
41	TAGGCCTA	0	0	0	2	0	1	2	0	2	0	2	0
42	TATGCATA	0	2	0	0	0	1	0	0	2	2	2	0
43	AAAAAAAAA	8	0	0	0	0	0	0	0	0	0	2	0
44	ATAACTGGC	1	1	0	1	0	1	1	1	1	1	1	1
45	ATCTATCCG	0	2	1	1	2	0	1	0	1	0	1	1
46	CGCTGTTAC	1	0	1	1	0	1	0	2	1	1	0	2
47	GCCAGTTAA	2	0	0	1	0	1	1	1	1	1	1	1

(*continued*)

TABLE 1 Frequencies of Nearest Neighbor and Terminals of DNA Duplex. Continued

		Sequences for training										Terminals	
		AA	AT	CG	CT	GA	GC	GG	GT	TA	TG	A	C
DNA sequences		TT	TA	GC	GA	CT	CG	CC	CA	AT	AC	T	G
48	AAAAAAAAAA	9	0	0	0	0	0	0	0	0	0	2	0
49	CGGCAAGCGC	1	0	2	1	0	3	1	0	0	1	0	2
50	TAGGTTATAA	2	1	0	1	0	0	1	1	3	0	2	0
51	ACGTATTATGC	1	2	1	0	0	1	0	2	2	1	1	1
52	ATTGGATACAAA	3	2	0	0	1	0	1	1	1	2	2	0
53	ACATTATTATTACA	3	3	0	0	0	0	0	2	3	2	2	0
54	GGATATCC	0	2	0	0	2	0	2	0	1	0	0	2
55	CACAG	0	0	0	1	0	0	0	1	0	2	0	2
56	CAACCAACCAAC	3	0	0	0	0	0	2	3	0	3	0	2
57	CTTCCTTCCTTC	3	0	0	3	3	0	2	0	0	0	0	2
58	CCGG	0	0	1	0	0	0	2	0	0	0	0	2
59	CGCG	0	0	2	0	0	1	0	0	0	0	0	2
60	GCGC	0	0	1	0	0	2	0	0	0	0	0	2
61	CCGCGG	0	0	2	0	0	1	2	0	0	0	0	2
62	CGATCG	0	1	2	0	2	0	0	0	0	0	0	2
63	CGCGCG	0	0	3	0	0	2	0	0	0	0	0	2
64	CGGCCG	0	0	2	0	0	1	2	0	0	0	0	2
65	CGTACG	0	0	2	0	0	0	0	2	1	0	0	2
66	GACGTC	0	0	1	0	2	0	0	2	0	0	0	2
67	GCATGC	0	1	0	0	0	2	0	0	0	2	0	2
68	GCCGGC	0	0	1	0	0	2	2	0	0	0	0	2
69	GCGAGC	0	0	1	1	1	2	0	0	0	0	0	2
70	GCTAGC	0	0	0	2	0	2	0	0	1	0	0	2
71	GGATCC	0	1	0	0	2	0	2	0	0	0	0	2
72	GGCGCC	0	0	1	0	0	2	2	0	0	0	0	2
73	GGGACC	0	0	0	0	1	0	3	1	0	0	0	2
74	GTGAAC	1	0	0	0	1	0	0	2	0	1	0	2
75	CAAAAAAG	5	0	0	1	0	0	0	0	0	1	0	2
76	CAAGCTTG	2	0	0	2	0	1	0	0	0	2	0	2
77	CATCGATG	0	2	1	0	2	0	0	0	0	2	0	2
78	GCGCGC	0	0	2	0	0	3	0	0	0	0	0	2
79	CGTCGACG	0	0	3	0	2	0	0	2	0	0	0	2
80	GATCGATC	0	2	1	0	4	0	0	0	0	0	0	2
81	GATGCATC	0	2	0	0	2	1	0	0	0	2	0	2
82	GGACGTCC	0	0	1	0	2	0	2	2	0	0	0	2
83	GGAGCTCC	0	0	0	2	2	1	2	0	0	0	0	2
84	GGTATACC	0	1	0	0	0	0	2	2	2	0	0	2
85	GTACGTAC	0	0	1	0	0	0	0	4	2	0	0	2
86	GTAGCTAC	0	0	0	2	0	1	0	2	2	0	0	2
87	GTTGCAAC	2	0	0	0	0	1	0	2	0	2	0	2
88	GCGAATTCGC	2	1	2	0	2	2	0	0	0	0	0	2
89	CGCGTACGCGTACGCG	0	0	6	0	0	3	0	4	2	0	0	2
90	CCATCGCTACC	0	1	1	1	1	1	2	1	1	1	0	2
91	GCGAAAAGCG	3	0	2	1	1	2	0	0	0	0	0	2
92	CCATTGCTACC	1	1	0	1	0	1	2	1	1	2	0	2
93	CTGACAAGTGTC	1	0	0	2	2	0	0	3	0	3	0	2
94	CATATGGCCATATG	0	4	0	0	0	1	2	0	2	4	0	2

(*continued*)

TABLE 1 Continued

		Sequences for training										Terminals	
		AA	AT	CG	CT	GA	GC	GG	GT	TA	TG	A	C
DNA sequences		TT	TA	GC	GA	CT	CG	CC	CA	AT	AC	T	G
95	TCATGA	0	1	0	0	2	0	0	0	0	2	2	0
96	TGATCA	0	1	0	0	2	0	0	0	0	2	2	0
97	AAAAAAAA	7	0	0	0	0	0	0	0	0	0	2	0
98	TAGATCTA	0	1	0	2	2	0	0	0	2	0	2	0
99	TCTATAGA	0	1	0	2	2	0	0	0	2	0	2	0
100	ATGAGCTCAT	0	2	0	2	2	1	0	0	0	2	2	0
101	TTTTATAATAAA	6	2	0	0	0	0	0	0	3	0	2	0
102	CAACTTGATATTATTA	4	3	0	1	1	0	0	1	3	2	1	1

of nearest neighbors is 10. For the above sequence, the frequencies of each nearest neighbor are as follows:

AA	AT	CG	CT	GA	GC	GG	GT	TA	TG
TT	TA	GC	GA	CT	CG	CC	CA	AT	AC
0	2	0	2	2	0	0	0	1	0

In this calculation, each base pair in the middle of the sequence is used twice (it belongs to both left nearest neighbor and right nearest neighbor), but each base pair at the terminals is used only once. Therefore, we proposed two terminal parameters A/T (or T/A) and C/G (or G/C) as a correction. In the above sequence the frequencies of the terminal parameters are A/T = 2 and G/C = 0. In our method, the symmetry effect, the initiation parameter, and the 3′-terminal phosphate present

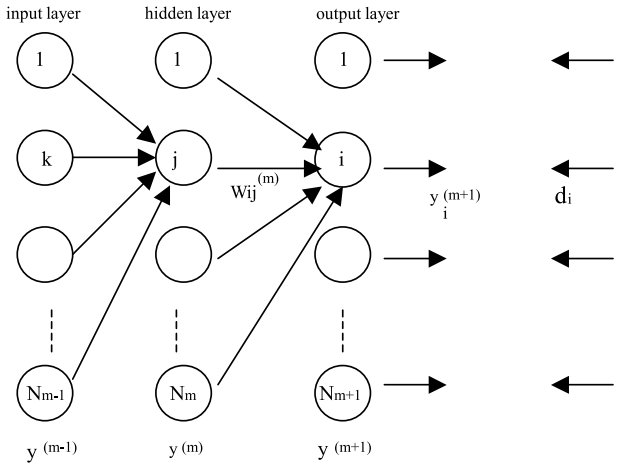

FIGURE 1 Three-layer networks.

TABLE 2 Predictions of T_m (K) by NN and TP Method for Training Set

	Sequence	Measured T_m (K)	Predicted T_m by NN1 (K)	Predicted T_m by NN 2 (K)	Predicted T_m by TP (K)
1	AGCCG	302.45	302.14 (− 0.31)	301.89 (0.56)	300.95 (91.5)
2	ACCGCA	310.95	309.00 (− 1.95)	308.84 (2.11)	313.05 (− 2.1)
4	CGGTGC	313.95	311.53 (− 2.42)	312.77 (1.18)	313.05 (0.9)
5	CGTGCC	312.55	311.55 (− 1.00)	312.77 (− 0.22)	313.05 (− 0.5)
6	TGCGCA	319.35	320.15 (0.80)	319.21 (0.14)	318.65 (0.7)
8	AGCCGTG	321.85	320.22 (− 1.63)	320.55 (1.3)	321.35 (0.5)
10	GGACTTA	304.55	303.16 (− 1.39)	303.78 (0.77)	302.75 (1.8)
11	CGATATCG	318.15	317.22 (− 0.93)	317.20 (0.95)	319.95 (− 1.8)
12	GAAGCTTC	316.05	314.60 (− 1.45)	315.49 (0.56)	317.25 (− 1.2)
13	GGAATTCC	319.75	319.49 (− 0.26)	318.97 (0.78)	315.45 (4.3)
15	AAAAAAAA	300.15	302.09 (1.94)	301.71 (− 1.56)	302.15 (− 2)
16	AAGCGTAG	318.55	318.93 (0.38)	318.74 (− 0.19)	320.05 (− 1.5)
17	AATCCAGT	313.45	313.27 (− 0.18)	313.81 (− 0.36)	312.75 (0.7)
18	ACATATGT	309.65	308.33 (− 1.32)	308.58 (1.07)	307.35 (2.3)
19	ACCTAGTC	310.35	314.37 (4.02)	314.28 (− 3.93)	313.35 (− 3)
20	ACGACCTC	322.15	323.08 (0.93)	322.14 (0.01)	322.55 (− 0.4)
21	AGAGAGAG	315.95	314.61 (− 1.34)	315.22 (0.73)	314.15 (1.8)
22	AGCGTAAG	316.35	318.91 (2.56)	318.74 (− 2.39)	320.05 (− 3.7)
23	AGTCCTGA	315.65	317.03 (1.38)	318.01 (− 2.36)	317.75 (− 2.1)
25	CACGGCTC	327.65	325.06 (− 2.59)	325.51 (2.14)	327.55 (0.1)
26	CCATATGG	315.15	315.46 (0.31)	315.32 (− 0.17)	314.25 (0.9)
28	CGCTGTAA	317.35	316.23 (− 1.12)	316.28 (1.07)	320.65 (− 3.3)
29	CTAGTGGA	314.85	317.39 (2.54)	316.84 (− 1.99)	314.45 (0.4)
31	CTGAGTCC	317.65	320.48 (2.83)	320.30 (− 2.65)	317.75 (− 0.1)
32	GAATATTC	303.05	303.39 (0.34)	304.82 (− 1.77)	301.85 (1.2)
33	GACTAGTC	315.75	313.06 (− 2.69)	313.15 (2.6)	312.55 (3.2)
34	GAGTACTC	310.45	313.09 (2.64)	313.15 (− 2.7)	312.55 (− 2.1)
36	GCATATGC	322.65	323.46 (0.81)	323.67 (− 1.02)	316.85 (5.8)
37	GCCAGTTA	318.65	319.07 (0.42)	319.29 (− 0.64)	316.95 (1.7)
38	GGTGCCAA	322.15	323.17 (1.02)	324.44 (− 2.29)	323.25 (− 1.1)
39	GTCGAACA	319.55	318.96 (− 0.59)	318.42 (1.13)	319.05 (0.5)
40	GTCTAGAC	314.55	313.05 (− 1.51)	313.15 (1.4)	312.55 (2)
41	TAGGCCTA	324.85	320.98 (− 3.87)	320.46 (4.39)	320.15 (4.7)
42	TATGCATA	311.75	312.98 (1.23)	312.54 (− 0.79)	308.55 (3.2)
44	ATAACTGGC	323.05	322.72 (− 0.33)	323.50 (− 0.45)	320.75 (2.3)
45	ATCTATCCG	320.05	318.75 (− 1.30)	317.97 (2.08)	321.05 (− 1)
46	CGCTGTTAC	324.35	326.71 (2.36)	326.73 (− 2.38)	326.35 (− 2)
47	GCCAGTTAA	321.25	323.81 (2.56)	324.16 (− 2.91)	321.95 (− 0.7)
48	AAAAAAAAAA	313.05	313.69 (0.64)	313.49 (− 0.44)	314.35 (− 1.3)
50	TAGGTTATAA	312.25	312.93 (0.68)	312.70 (− 0.45)	315.55 (− 3.3)
51	ACGTATTATGC	326.55	327.97 (1.42)	328.12 (− 1.57)	328.95 (− 2.4)
52	ATTGGATACAAA	326.35	328.66 (2.31)	328.80 (− 2.45)	328.45 (− 2.1)
53	ACATTATTATTACA	326.75	325.42 (− 1.33)	324.22 (2.53)	328.75 (− 2)
54	GGATATCC	309.35	310.76 (1.41)	310.68 (− 1.33)	311.95 (− 2.6)
55	CACAG	285.25	286.48 (1.23)	286.14 (− 0.89)	279.05 (6.2)
57	CTTCCTTCCTTC	338.15	335.32 (− 2.83)	334.11 (4.04)	335.75 (2.4)
59	CGCG	296.85	297.35 (0.50)	298.20 (− 1.35)	298.15 (− 1.3)
60	GCGC	300.65	300.32 (− 0.33)	299.43 (1.22)	299.35 (1.3)
61	CCGCGG	328.35	330.58 (2.23)	329.92 (− 1.57)	328.15 (0.2)
62	CGATCG	307.45	308.79 (1.34)	309.04 (− 1.59)	309.55 (− 2.1)

(*continued*)

TABLE 2 Continued

	Sequence	Measured T_m (K)	Predicted T_m by NN1 (K)	Predicted T_m by NN 2 (K)	Predicted T_m by TP (K)
63	CGCGCG	328.85	327.91 (− 0.94)	328.69 (0.16)	330.45 (− 1.6)
64	CGGCCG	332.75	330.56 (− 2.19)	329.92 (2.83)	328.15 (4.6)
65	CGTACG	308.15	308.29 (0.14)	310.12 (− 1.97)	309.75 (− 1.6)
66	GACGTC	309.25	310.80 (1.55)	310.35 (− 1.1)	310.05 (− 0.8)
67	GCATGC	309.65	309.20 (− 0.45)	309.45 (0.2)	311.15 (− 1.5)
69	GCGAGC	306.35	310.73 (4.38)	311.00 (− 4.65)	305.15 (1.2)
70	GCTAGC	307.45	306.64 (− 0.81)	305.95 (1.5)	305.75 (1.7)
72	GGCGCC	327.05	329.12 (2.07)	329.22 (− 2.17)	328.55 (− 1.5)
73	GGGACC	318.05	317.26 (− 0.79)	317.02 (1.03)	318.45 (− 0.4)
74	GTGAAC	306.35	304.17 (− 2.18)	302.97 (3.38)	305.15 (1.2)
75	CAAAAAAG	310.05	309.40 (− 0.65)	309.51 (0.54)	312.55 (− 2.5)
76	CAAGCTTG	317.75	317.80 (0.05)	317.32 (0.43)	319.15 (− 1.4)
77	CATCGATG	320.55	318.08 (− 2.47)	317.94 (2.61)	317.45 (3.1)
78	GCGCGC	329.35	328.38 (− 0.97)	328.51 (0.84)	328.55 (0.8)
79	CGTCGACG	331.45	331.03 (− 0.42)	330.71 (0.74)	333.15 (− 1.7)
80	GATCGATC	320.75	320.09 (− 0.66)	319.73 (1.02)	317.25 (3.5)
81	GATGCATC	319.65	320.01 (0.36)	320.08 (− 0.43)	317.95 (1.7)
82	GGACGTCC	328.25	329.96 (1.71)	330.08 (− 1.83)	332.15 (− 3.9)
83	GGAGCTCC	327.55	327.28 (− 0.27)	326.52 (1.03)	329.75 (− 2.2)
84	GGTATACC	309.15	308.15 (− 1.00)	308.84 (0.31)	313.35 (− 4.2)
85	GTACGTAC	318.25	318.08 (− 0.17)	316.71 (1.54)	317.05 (1.2)
86	GTAGCTAC	318.05	315.50 (− 2.55)	315.87 (2.18)	313.85 (4.2)
87	GTTGCAAC	320.85	322.68 (1.83)	322.24 (− 1.39)	322.35 (− 1.5)
89	CGCGTACGCGTACGCG	364.15	363.40 (− 0.75)	364.06 (0.09)	362.45 (1.7)
91	GCGAAAAGCG	338.35	339.06 (0.71)	339.38 (− 1.03)	340.35 (− 2)
93	CTGACAAGTGTC	338.85	336.15 (− 2.70)	337.03 (1.82)	339.35 (− 0.5)
95	TCATGA	295.95	294.68 (− 1.27)	294.63 (1.32)	290.45 (5.5)
96	TGATCA	294.05	294.71 (0.66)	294.63 (− 0.58)	290.45 (3.6)
97	AAAAAAAA	304.35	301.88 (− 2.47)	301.71 (2.64)	308.35 (− 4)
98	TAGATCTA	306.55	304.29 (− 2.26)	304.50 (2.05)	297.65 (8.9)
99	TCTATAGA	301.25	304.34 (3.09)	304.50 (− 3.25)	297.65 (3.6)
101	TTTTATAATAAA	309.45	309.54 (0.09)	310.16 (− 0.71)	314.75 (− 5.3)
102	CAACTTGATATTATTA	331.95	332.36 (0.41)	332.63 (− 0.68)	331.25 (0.7)
9	AGCTTCA	306.95		307.26 (− 0.31)	309.05 (− 2.1)
24	ATGCGCAT	328.35		327.15 (1.2)	326.95 (1.4)
56	CAACCAACCAAC	334.35		337.43 (− 3.08)	335.55 (− 1.2)
71	GGATCC	303.95		305.38 (− 1.43)	303.75 (0.2)
90	CCATCGCTACC	340.75		338.06 (2.69)	342.65 (− 1.9)
92	CCATTGCTACC	336.65		335.73 (0.92)	338.15 (− 1.5)
100	ATGAGCTCAT	331.25		332.18 (− 0.93)	327.55 (3.7)
3	ATGCGC	310.65	310.76 (0.11)		310.65 (0)
7	AATACCG	305.65	305.78 (0.13)		306.95 (− 1.3)
27	CGCGTATA	320.25	319.34 (− 0.91)		321.45 (− 1.2)
30	CTCACGGC	325.45	325.07 (− 0.38)		327.55 (− 2.1)
35	GATTAATC	302.85	303.39 (0.54)		301.85 (1)
43	AAAAAAAAA	308.25	307.91 (− 0.34)		308.85 (− 0.6)
49	CGGCAAGCGC	335.85	337.47 (1.62)		345.9 (− 10.1)
58	CCGG	289.75	290.79 (1.04)		275.55 (14.2)
68	GCCGGC	330.85	329.13 (− 1.72)		328.55 (2.3)
88	GCGAATTCGC	340.95	340.89 (− 0.06)		338.85 (2.1)

(*continued*)

TABLE 2 Predictions of T_m (K) by NN and TP Method for Training Set. Continued

	Sequence	Measured T_m (K)	Predicted T_m by NN1 (K)	Predicted T_m by NN 2 (K)	Predicted T_m by TP (K)
94	CATATGGCCATATG	338.45	338.64 (0.19)		339.55 (− 1.1)
SEP			1.73872	1.65278	2.18411
AD			1.45	1.34033	1.75667
R^2			0.99075	0.9884	0.9751

AD, average deviation; R^2, relation coefficient; SEP, standard error of prediction (SEP = $[1/N\Sigma\ (T_i\text{-}O_i)^2]^{1/2}$).

on some oligomers are ignored. All frequencies of 12 parameters for 103 sequences are listed in Table 1.

Back-Propagation (BP) Method

In our neural network, a back-propagation (BP) 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 sigmoidal function. The neural network contained three layers. The 10 nearest neighbors and 2 terminals as is shown in Table 1 are defined as in the INN-HB model[10,16] and the frequencies of them for a given sequence are used as input data of the network. The middle layer, or hidden layer, interconnects with the output and input layers as shown in Figure 1. The initial weight values are randomly generated in the first step. This value is summed for all the connections made from the lower layer. Each node in the hidden layer has a specific threshold. The node is turned on once the threshold is breached. The hidden layer nodes are then multiplied by the weights that connect them to the output node. All the supervisor data (measured T_ms) were obtained from literature, and all the computations were performed on a Silicon Graphics O2

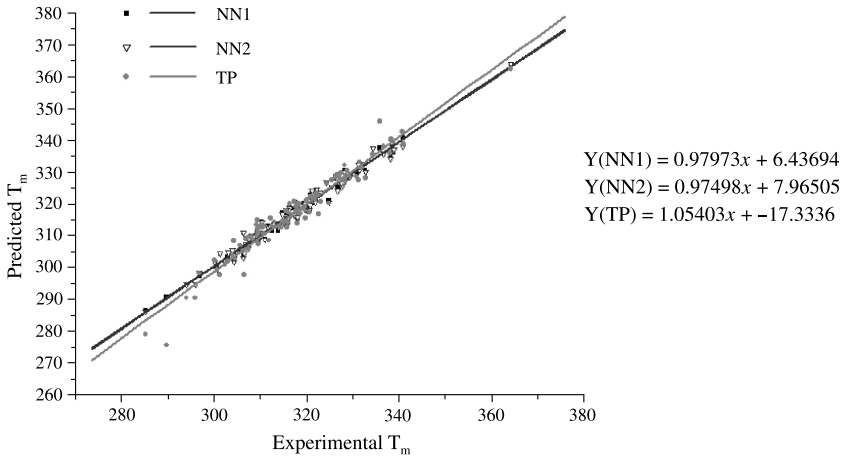

FIGURE 2 Comparison of NN1, NN2 and TP methods for training set.

R5000 workstation using the artificial neural network program developed by the Artificial Intelligence Institute, Shanghai University of Electric Power.

RESULTS

In order to compare the NN method and TP method, we adopted the same 102 sequences which have been used in the TP method. We divided them into 94 sequences as the learning set and 8 sequences as the test set in NN1, and 90 sequences as learning set and 12 sequences as test set in NN2. All the melting temperatures are expressed in degrees Kelvin (K).

As shown in Table 2 and Figure 2, all the predicted T_ms of the learning set agree with the experimental data very well for the NN method (AD = 1.45 K, SEP = 1.73872, R^2 = 0.99075 for NN1; AD = 1.34033 K, SEP = 1.652785, R^2 = 0.9884 for NN2). Table 3 and Figure 3 show that the prediction accuracy for the test set is good (AD = 1.86 K, SEP = 1.99151, R^2 = 0.9894 for NN1; AD = 1.59667 K, SEP = 2.03824, R^2 = 0.99371 for NN2). For comparison, the prediction accuracy of the TP method is AD = 1.75667 K, SEP = 2.18411, R^2 = 0.9751 for the learning set and AD = 2.54474 K, SEP = 4.29205, R^2 = 0.97553 for the test set. From the above results we can see that the network for the test set NN1 and NN2 demonstrated good prediction accuracy, with some indices being better than for TP methods. Thus, we can conclude that the neural

TABLE 3 Predictions of T_m (K) by NN and TP Method for Test Set

	Sequence	Measured Tm (K)	Predicted Tm by NN 1 (K)	Predicted Tm by NN 2 (K)	Predicted Tm by TP (K)
9	AGCTTCA	306.95	305.79 (1.16)		309.05 (− 2.1)
24	ATGCGCAT	328.35	326.23 (2.12)		326.95 (1.4)
56	CAACCAACCAAC	334.35	337.22 (− 2.86)		335.55 (− 1.2)
71	GGATCC	303.95	305.07 (− 1.12)		303.75 (0.2)
90	CCATCGCTACC	340.75	337.69 (3.05)		342.65 (− 1.9)
92	CCATTGCTACC	336.65	334.92 (1.72)		338.15 (− 1.5)
100	ATGAGCTCAT	331.25	330.02 (1.23)		327.55 (3.7)
14	CAAACAAAG	316.95	318.54 (− 1.58)	318.99 (− 2)	316.55 (0.4)
3	ATGCGC	310.65		311.69 (− 1)	310.65 (0)
7	AATACCG	305.65		306.11 (− 0.5)	306.95 (− 1.3)
27	CGCGTATA	320.25		319.14 (1.1)	321.45 (− 1.2)
30	CTCACGGC	325.45		325.51 (− 0.1)	327.55 (− 2.1)
35	GATTAATC	302.85		304.82 (− 1.97)	301.85 (1)
43	AAAAAAAAA	308.25		307.27 (0.98)	308.85 (− 0.6)
49	CGGCAAGCGC	335.85		337.56 (− 1.7)	345.9 (− 10.1)
58	CCGG	289.75		295.07 (− 5.3)	275.55 (14.2)
68	GCCGGC	330.85		329.22 (1.6)	328.55 (2.3)
88	GCGAATTCGC	340.95		339.98 (0.96)	338.85 (2.1)
94	CATATGGCCATATG	338.45		340.32 (− 1.86)	339.55 (− 1.1)
SEP			1.99151	2.03824	4.29205
AD			1.86	1.59667	2.54474
R^2			0.9894	0.9937	0.97553

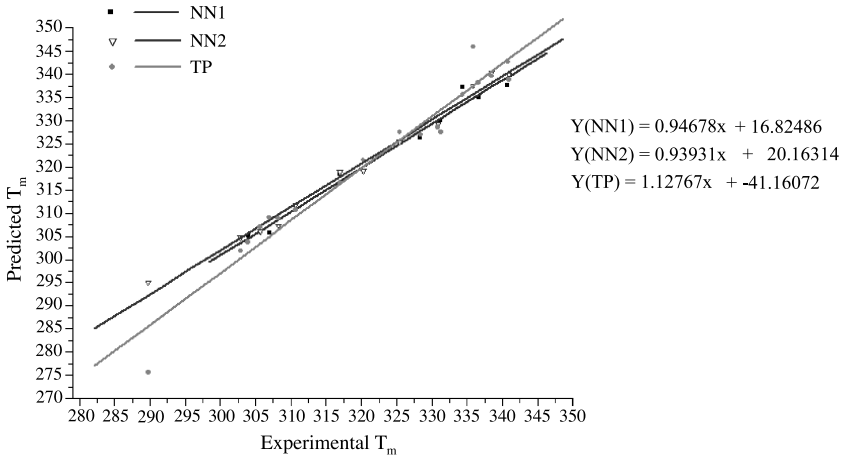

FIGURE 3 Comparison of NN and TP methods for test set.

network method is a reliable way of predicting melting temperatures of DNA duplexes. The greatest advantage of this method is that measurements of thermodynamic parameters are not required. Therefore, this method can be smoothly applied to the prediction of melting temperatures of DNA/DNA duplexes.

CONCLUSION

We provide here a simple and reliable method to predict the melting temperature directly from the sequences of DNA/DNA. The main merit of this method is that the measurements of thermodynamic parameters are omitted.

REFERENCES

1. Petruska, J.; Goodman, M.F.; Boosalis, M.S.; Sowers, L.S.; Cheong, C.; Tinoco, I. On comparison between DNA melting thermodynamics and DNA polymerase fidelity. Proc. Natl. Acad. Sci. U. S. A. **1988**, 85, 6252–6256.
2. Mendelman, L.V.; Boosalis, M.S.; Petruska, J.; Goodman, M.F. On nearest neighbor influences on DNA polymerase insertion fidelity. J. Biol. Chem. **1989**, 264, 14415–14423.
3. Steger, G. On thermal denaturation of double-stranded nucleic acids: prediction of temperatures critical for gradient gel electrophoresis and polymerase chain reaction. Nucleic Acids Res. **1994**, 22, 2760–2768.
4. Southern, E.M. On detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **1975**, 98, 503–517.
5. Freier, S.M. Antisense Research and Applications. Crooke, S.T., Lebleu, B., Eds.; CRC Press: Boca Raton, FL, 1993; 67–82.
6. Fodor, S.P.A.; Rava, R.P.; Huang, X.C.; Pease, A.C.; Holmes, C.P.; Adams, C.L. On multiplexed biochemical assays with biological chips. Nature **1993**, 364, 555–556.
7. Saiki, R.K.; Gelfand, D.H.; Stoffel, S.; Scharf, S.; Higuchi, R.H.; Horn, G.T.; Mullis, K.B.; Erlich, H.A. On primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **1988**, 239, 487–494.
8. Sugimoto, N.; Nakano, S.; Katoh, M.; Matsumura, A.; Nakamuta, H.; Ohmichi, T.; Yoneyama, M.; Sasaki, M. On thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry **1995**, 34, 11211–11216.

9. Breslauer, K.J.; Frank, R.; Blocker, H.; Marky, L.A. On predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. U. S. A. **1986**, 83, 3746–3750.
10. Freier, S.M.; Kierzek, R.; Jaeger, J.A.; Sugimoto, N.; Caruthers, M.H.; Neilson, T.; Turner, D.H. On improved free-energy parameters for predictions of RNA duplex stability. Proc. Natl. Acad. Sci. U. S. A. **1986**, 83, 9373–9377.
11. John, S.L.J.; Hatim, T.A.; Seneviratne, P.A. On improved nearest-neighbor parameters for predicting DNA duplex stability. Biochemistry **1996**, 35, 3555–3562.
12. Zupan, J.; Gasteiger, J. Neural Networks for Chemistry. VCH: Cambridge, 1993.
13. Cheng, C.M.; Ma, L.X.; Miyajima, H.; Herdewijn, P. A neural network to predict melting temperature (T_m) of RNA duplex. Nucleosides Nucleotides Nucleic Acids **2001**, 20(3), 261–269.
14. Ma, L.X.; Cheng, C.M.; Miyajima, H.; Zhao, Y.F.; Herdewijn, P. On predicting melting temperature (T_m) of oligoribonucleotide duplex by neural network. J. Chemom. **2002**, 16, 75–80.
15. Ma, L.X.; Cheng, C.M.; Liu, X.H.; Zhao, Y.F.; Wang, A.L.; Herdewijn, P. On a neural network for predicting the stability of RNA/DNA hybrid duplexes. Chemometr. Intell. Lab. Syst. **2004**, 70, 123–128.
16. Xia, T.; SantaLucia, J., Jr.; Burkard, M.E.; Kierzek, M.R.; Schroeder, S.J.; Jiao, X.; Cox, C.; Turner, D.H. On thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry **1998**, 37, 14719–14735.