Supporting online material for:

ptRNApred: Computational identification and classification of post-transcriptional RNA

Yask Gupta, Mareike Witte, Steffen Möller, Ralf J. Ludwig, Tobias Restle, Detlef Zillikens and Saleh M. Ibrahim

Table of Contents for Supporting Online Material

Material

Table S1: Table of selected dinucleotide properties.

Property Name	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
Shift[1]	-0.08	0.23	-0.04	-0.06	0.11	-0.01	0.3	-0.04	0.07	0.07	-0.01	0.23	-0.02	0.07	0.11	-0.08
Hydrophilicity[2]	0.023	0.083	0.035	0.09	0.118	0.349	0.193	0.378	0.048	0.146	0.065	0.16	0.112	0.359	0.224	0.389
GC_content[3]	0	1	1	0	1	2	2	1	1	2	2	1	0	1	1	0
Keto_content[4]	0	0	0	1	0	0	1	1	1	1	2	2	1	1	1	2
Adenine_content[5]	2	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0
Guanine_content[6]	0	0	1	0	0	0	1	0	1	1	2	1	0	0	1	0
Cytosine_content[7]	0	1	0	0	1	2	1	1	0	1	0	0	0	1	0	0
Slide[1]	-1.27	-1.43	-1.5	-1.36	-1.46	-1.78	-1.89	-1.5	-1.7	-1.39	-1.78	-1.43	-1.45	-1.7	-1.46	-1.27
Rise[1]	3.18	3.24	3.3	3.24	3.09	3.32	3.3	3.3	3.38	3.22	3.32	3.24	3.26	3.38	3.09	3.18
Tilt[1]	-0.8	0.8	0.5	1.1	1	0.3	-0.1	0.5	1.3	0	0.3	0.8	-0.2	1.3	1	-0.8
Roll[1]	7	4.8	8.5	7.1	9.9	8.7	12.1	8.5	9.4	6.1	12.1	4.8	10.7	9.4	9.9	7
Twist[1]	31	32	30	33	31	32	27	30	32	35	32	32	32	32	31	31
Stacking_energy[8]	-13.7	-13.8	-14	-15.4	-14.4	-11.1	-15.6	-14	-14.2	-16.9	-11.1	-13.8	-16	-14.2	-14.4	-13.7
Entropy_1[9]	-18.4	-26.2	-19.2	-15.5	-27.8	-29.7	-19.4	-19.2	-35.5	-34.9	-29.7	-26.2	-22.6	-26.2	-19.2	-18.4
Entropy_2[10]	-19	-29.5	-27.1	-26.7	-26.9	-32.7	-26.7	-27.1	-32.5	-36.9	-32.7	-29.5	-20.5	-32.5	-26.9	-19

1. Perez, A. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. Nucleic Acids Res. 32, 6144–6151 (2004).

2. Weber, A. L. & Lacey, J. C., Jr. Genetic code correlations: amino acids and their anticodon nucleotides. J. Mol. Evol. 11, 199–210 (1978).

3. Friedel, M. Each C or G counts +1.

4. Friedel, M. G and T (U) counts +1.

5. Friedel, M. Each A counts +1.

6. Friedel, M. Each G counts +1.

7. Friedel, M. Each C counts +1.

8. Encyclopedia of Life Sciences. (John Wiley & Sons, Ltd, 2001).

9. Freier, S. M. et al. Improved free-energy parameters for predictions of RNA duplex stability. Proc. Natl. Acad. Sci. U. S. A. 83, 9373–9377 (1986).

10. Xia, T. et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry (Mosc.) 37, 14719–14735 (1998).

Data: The table shows the dinucleotide properties selected as vectors for the SVM. 15 distinct properties (left column), ranging from the shift score to the entropy, are assigned to the16 possible dinucleotides: AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CC, CG, GA, GT, GC and GG. These properties have been described in previous experimental or computational work. All information was derived from DiProGB (1). Further information is provided in Section S1.

Table S2: Performances of ptRNApred on each of the different ptRNA subclasses using a random forest classification according to Breiman.

The results are presented in a confusion matrix. As a result, implementation of Random Forest yields an overall accuracy of 82%. In comparison, our multi-class classifier developed using LibSVM yields an accuracy of 91% (data not shown).

		Predicted class						
		RNase MRP	RNAse P	snoRNA	snRNA	telomerase RNA	YRNA	Accuracy
Actual class	RNase MRP	3	2	1	0	0	0	50
	RNAse P	0	100	6	0	3	0	91,7
	snoRNA	0	11	1123	75	1	0	92,8
	snRNA	0	4	194	713	2	0	78,1
	telomerase RNA	0	4	4	2	8	0	44,4
	YRNA	0	0	10	1	0	0	0
								82

Table S3: Table of properties for discrimination of ptRNA.

Table S3 summarizes all 91 features integrated into the algorithm of ptRNApred. The properties are ordered the way they appear in the Perl-script for training and testing. Furthermore, the table depicts the F-score corresponding to every feature, as well as the Gini-Index derived from random forest calculation according to Breiman. A detailed description of the feature selection is provided in Section S1.

Property number	Property description	F-Score	Gini-Index
1	Adenine_content[1]	0,064268	2,41E-44
2	Cytosine_content[1]	0,083006	6,03E-15
3	Entropy_1[1]	0,088785	4,34E-19
4	Entropy_2[1]	0,09256	5,93E-21
5	GC_content[1]	0,106083	5,43E-11
6	Guanine_content[1]	0,12987	2,75E-18
7	Hydrophilicity[1]	0,098519	4,37E-23
8	Keto_content[1]	0,131688	4,12E-16
9	Rise[1]	0,091877	8,37E-20
10	Roll[1]	0,098188	6,50E-26
11	Shift[1]	0,057281	5,62E-49
12	Slide[1]	0,09489	4,93E-19
13	Stacking_energy[1]	0,088663	1,80E-30
14	Tilt[1]	0,060113	6,03E-20
15	Twist[1]	0,090472	3,98E-27
16	value_in_3rd_rnafold[2]	0,039964	2,48E-51
17	count_star_bracket_{[2]	0,032699	0,002720791
18	count_comma_in_rnafold[2]	0,032914	0,000482739
19	value_line_no_3_RNAfold[2]	0,037566	2,54E-21
20	value_line_no_3_RNAfold(second value)[2]	0,034666	7,25E-18
21	value_line_number_4[2]	0,037892	1,68E-18
22	value_line_number_4(second value) [2]	0,034268	6,11E-32
23	frequency_of_mfe_structure_in_ensemble[2]	0,035204	1,12E-16
24	ensemble diversity[2]	0,039227	1,21E-11
25	value_MFE_RNAfold[2]	0,038989	1,18E-34
26	Number_of_loops[2]	0,03857	1,12E-12
27	(((2	0,032912	3,18E-29

28	((.^{2}	0,037534	7,90E-11
29	(..^{2}	0,039972	3,42E-07
30	...^{2}	0,029854	7,49E-10
31	.((^{2}	0,038714	4,07E-11
32	..(^{2}	0,040519	0,000102735
33	.(.^{2}	0,023965	0,996225164
34	(.(^{2}	0,031401	0,003872273
35	A((((^{2}	0,003259	5,34E-15
36	A((.^{2}	0,003767	1,49E-06
37	A(.(^{2}	0,00501	0,037528552
38	A(..^{2}	0,014225	2,74E-05
39	A.((^{2}	0,001732	4,20E-06
40	A.(.^{2}	0,001482	0,119541752
41	A..(^{2}	0,015074	2,38E-12
42	A...^{2}	0,029134	5,04E-12
43	U((((^{2}	0,003228	1,41E-09
44	U((.^{2}	0,00721	2,26E-05
45	U(.(^{2}	0,00737	0,000196773
46	U(..^{2}	0,010258	4,59E-09
47	U.((^{2}	0,002741	0,00024243
48	U.(.^{2}	0,000326	0,996205638
49	U..(^{2}	0,017021	1,90E-21
50	U...^{2}	0,093364	3,83E-43
51	G((((^{2}	0,021702	3,72E-27
52	G((.^{2}	0,024863	0,000494941
53	G(.(^{2}	0,005776	3,58E-08
54	G(..^{2}	0,010109	6,84E-14
55	G.((^{2}	0,003169	0,000144641
56	G.(.^{2}	0,001119	0,403681085
57	G..(^{2}	0,010458	2,26E-06
58	G...^{2}	0,006942	1,12E-30
59	C((((^{2}	0,044132	4,38E-07
60	C((.^{2}	0,009389	2,98E-09
61	C(.(^{2}	0,001738	0,98836726
62	C(..^{2}	0,002454	1,76E-08
63	C.((^{2}	0,016575	0,017213739
64	C.(.^{2}	0,000654	0,000405832
65	C..(^{2}	0,006828	0,003771578
66	C...^{2}	0,019666	2,19E-12
67	number_of_AU^2	0,018431	1,55E-20
68	number_of_CG^2	0,008113	1,63E-10
69	number_of_GU^2	0,005051	4,01E-09
70	number_of_mistatches_in_sec_struc^2	0,036243	1,48E-12
71	number_of_buldges_in_sec_struc^2	0,06064	4,57E-09
72	A_in_buldges^2	0,059369	2,94E-13
73	G_in_buldges^2	0,064056	0,168681863
74	C_in_buldges^2	0,037917	2,02E-05

75	U_in_buldges[2]	0,035724	1,48E-09
76	length_of_hairpin[2]	0,029	4,29E-12
77	number_of_sub_sec_structure[2]	0,025913	1,59E-11
78	number_of_A_hairpin[2]	0,026994	8,85E-09
79	number_of_G_hairpin[2]	0,010075	0,021895501
80	number_of_C_hairpin[2]	0,027103	0,006452735
81	number_of_U_hairpin[2]	0,066878	1,63E-12
82	number_of_A_purine[2]	0,021688	6,41E-07
83	number_of_A_pyramidine[2]	0,055579	1,70E-14
84	number_of_A_in_first_complementary_strand[2]	0,002319	1,32E-16
85	number_of_G_in_first_complementary_strand[2]	0,005858	4,21E-14
86	number_of_C_in_first_complementary_strand[2]	0,001151	1,07E-07
87	number_of_U_in_first_complementary_strand[2]	0,004656	1,13E-34
88	number_of_A_in_second_complementary_strand[2]	0,003052	2,39E-14
89	number_of_G_in_second_complementary_strand[2]	0,003099	0,00106611
90	number_of_C_in_second_complementary_strand[2]	0,012551	2,20E-08
91	number_of_U_in_second_complementary_strand[2]	0,035877	9,40E-16

[1] dinucleotide properties as shown in Table S1
[2] properties derived from the secondary structure provided by RNAfold

Table S4: Table of properties ranked by their importance for discrimination among ptRNA according to F-score and Gini-Index.

The table depicts the 25 most discriminative properties according to their F-score and Gini-Index. Interestingly, dinucleotide properties achieve high ranks: 9 of the 10 most discriminative features according to the F-score are composed of dinucleotide properties. Furthermore, all of the 15 dinucleotide properties can be found among the 25 most discriminative properties. According to the Gini-Index, 12 properties can be found among the 25 most discriminative properties, whereas only 3 of them can be found among the top 10.

Rank of importance for discrimination	Property ranked by F-score	Property ranked by Gini-Index
1	Keto_content	value_in_3rd_rnafold
2	Guanine_content	Shift
3	GC_content	Adenine_content
4	Hydrophilicity	U...
5	Roll	number_of_U_in_first_complementary_strand
6	Slide	value_MFE_RNAfold
7	U...	value_line_number_4(second value)
8	Entropy_2	G...
9	Rise	Stacking_energy
10	Twist	(((
11	Entropy_1	G(((
12	Stacking_energy	Twist
13	Cytosine_content	Roll

14	number_of_U_hairpin	Hydrophilicity
15	Adenine_content	U..(
16	G_in_buldges	value_line_no_3_RNAfold
17	number_of_buldges_in_sec_struc	Entropy_2
18	Tilt	number_of_AU
19	A_in_buldges	Tilt
20	Shift	Rise
21	number_of_A_pyramidine	Entropy_1
22	C(((	Slide
23	..(	value_line_number_4
24	(..	Guanine_content
25	value_in_3rd_rnafold	value_line_no_3_RNAfold(second value)

Section S1: Features for classification.

Feature selection and SVM training was performed using two sets of input parameters: The first set is based on the primary sequence and the second set considers the secondary structure which is predicted with RNAfold. All training steps were automated by a Perl script.

Set 1: Dinucleotide properties

Each sequence was divided into its dinucleotides, using the sliding window approach (window size: 2 nucleotides). In total, 16 different dinucleotides are possible: AA, AT, AC, AG, TA, TT, TC, TG, CA, CT, CC, CG, GA, GT, GC and GG. Each of the 16 dinucleotides can be assigned distinct properties, ranging from thermodynamic (e.g. stacking energy, free energy), structural (e.g. twist, roll) to other properties (e.g. sequence based). These properties have been described in previous experimental or computational work. The dinucleotide property database (DiProDB) (1) contains information on dinucleotides and a collection of more than 100 published dinucleotide property sets. In order to determine whether different ptRNA-subclasses can be distinguished via specific dinucleotide properties, 125 dinucleotide properties were abstracted from DiProDB and individually correlated with every ptRNA-subclass. Properties were clustered and a representative property was selected from each of the 16 resulting clusters (Table S1).

Set 2: Secondary structure properties

Secondary structures of every sequence were calculated via RNAfold (2), accessing the Vienna RNA Package (3).

RNAfold provides structures according to different parameters. Various properties were derived from the RNAfold output:

A) The Minimum free energy (MFE) structure

The MFE structure of an RNA sequence is the secondary structure that contributes a minimum of free energy. For MFE structure prediction, RNAfold uses a loop-based energy model and the dynamic

programming algorithm introduced by Zuker et al. (4). As an RNA secondary structure can be uniquely decomposed into loops and external bases the loop-based energy model treats the free energy of an RNA secondary structure as the sum of the contributing free energies of the loops contained in the secondary structure. According to the chosen energy parameter set and a given temperature (defaults to 37 °C) the secondary structure that minimizes the free energy of the secondary structure is computed. The minimum free energy was selected as property in our SVM. Additional features were deducted from the MFE structure, which is denoted by brackets '(or')' and dots '.' Brackets indicate paired nucleotides, whereas dots represent unpaired nucleotides. The left bracket '(means the paired nucleotide is located near the 5′-end and can be paired with another nucleotide at the 3′-end, which is indicated as a right bracket ')'. In our script, we do not distinguish these two situations and use '(' for both situations. Brackets and dots were counted within each sequence, yielding two additional properties.

Different features were selected combining secondary structure and primary sequence. In this context, the number of bulges and hairpins were counted, as well as the four nucleotides A, G, C and U in every bulge and every hairpin, yielding ten additional properties. Furthermore, purine and pyrimidine contents were examined and the number of mismatches was determined. Moreover, paired bases were considered alongside, counting AU, CG and GU pairs. All in all, this section of sequence examination yields 18 properties.

Further information was gained of every three adjacent nucleotides, which we call triplet elements for the convenience of discussion. Eight additional properties were given by the counting of the eight possible triplet element compositions '(((', '((.', '(..', '…', '.((', '..(', '.(.' and '(.(' within every sequence. The nucleotide composition of the triplet elements was not regarded and the compositions were counted using the sliding window approach.

32 further triplet element properties were derived from miPred, a triplet SVM for the classification of miRNA (41). MiPred considers the middle nucleotide among the triplet elements, resulting in 32 (4 × 8) possible combinations, which are denoted as 'U(((', 'A((.', etc. The number of appearance of each triplet element is counted for each hairpin to produce the 32-dimensional feature vector and used as input features for SVM.

B) The ensemble free energy

RNAfold provides an ensemble structure, considering probabilities of the presence of certain base pairs. Bases with a strong preference (more than 2/3) to pair upstream (with a partner further 3'), pair downstream or not pair and represented by the usual symbols '(', ')' or '.'. Additional symbols '{', '}' or ',' reflect bases with a weaker preference and thus are a weaker version of the above and '|' represents a base that is mostly paired but has pairing partners both upstream and downstream. In this case open and closed brackets need not match up. This pseudo bracket notation is followed by the ensemble free energy. The numbers of '{', '}' and ',' as well as the ensemble free energy were taken as features for our SVM.

C) The centroid structure

RNAfold further provides a centroid structure that is given by is the secondary structure with minimal base pair distance to all other secondary structures in the Boltzmann ensemble (5). The values of the centroid structure's free energy as well as its distance to the ensemble were taken as features for our SVM.

D) The maximum expected accuracy (MEA) structure

RNAfold further outputs a MEA structure, in which each base pair (i,j) gets a score 2*gamma*p_ij and the score of an unpaired base is given by the probability of not forming a pair. Subsequently, the expected accuracy is computed from the pair probabilities. The MEA as well as the MEA structure's free energy serve as additional features for our SVM.

E) The frequency of the MFE representative in the complete ensemble of secondary structures and the ensemble diversity

Two additional features are given by the frequency of the MFE representative in the complete ensemble of secondary structures and the ensemble diversity.

Altogether, ptRNApred uses 91 features for classification. A complete table of features is shown in Table S3.

Figure S1: Input of ptRNApred.

The user can either paste the sequence into the dedicated area or upload a FASTA file.

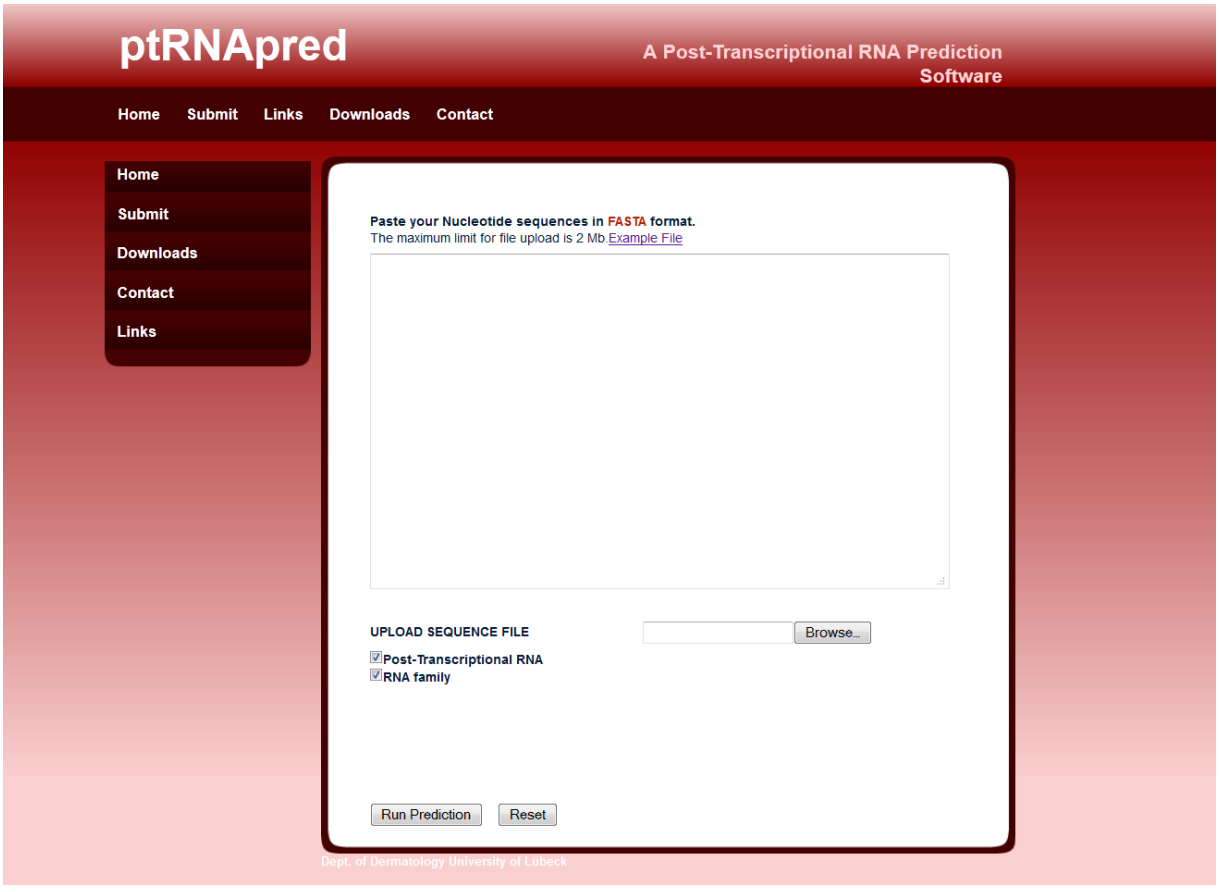


Figure S2. Output of ptRNApred.

The output of ptRNApred includes two predictions: 1. a prediction, whether or not the input sequence will be ptRNA and 2. the prediction of the ptRNA-subclass. Additionally, it displays the minimum free energy as well as the secondary structure, using VARNA. The output can directly be downloaded at the bottom of the page.

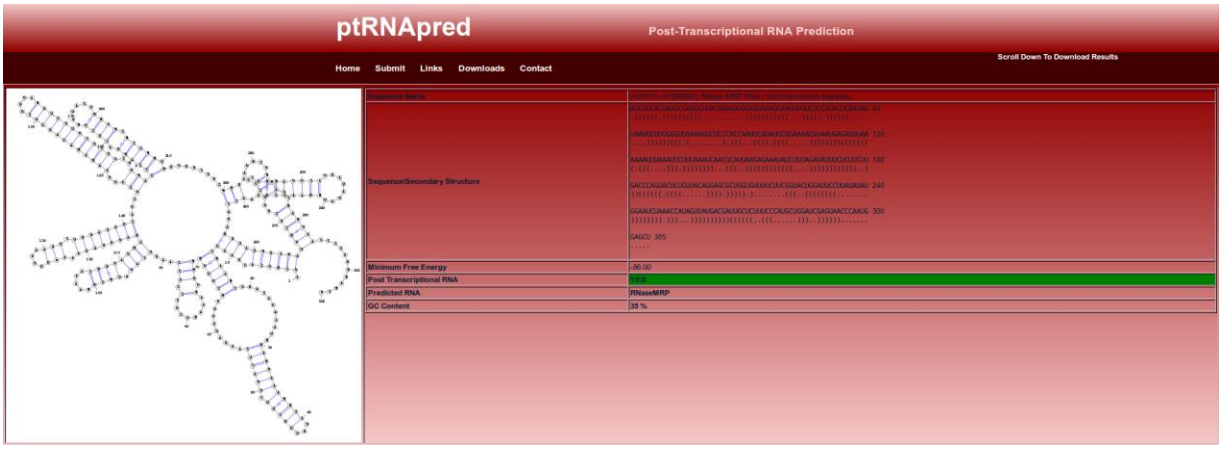
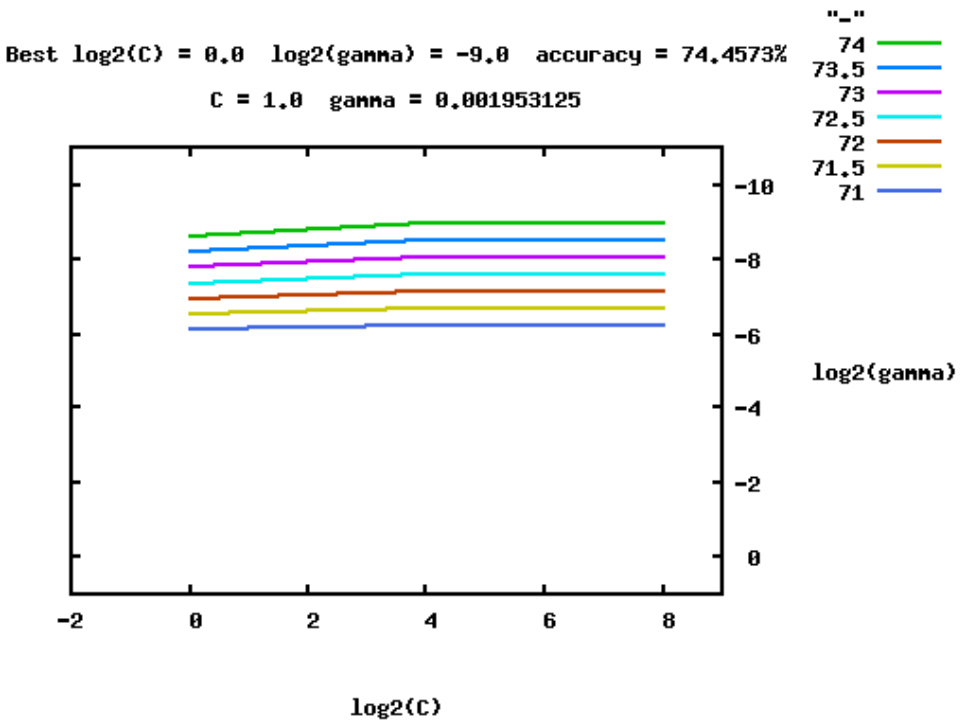


Figure S3. C and γ determination and 5 fold cross validation when using 78 instead of 91 features.

The green graph represents the optimal values for C and gamma. In this case, the highest 5 fold cross validation accuracy (74.46%) is achieved when C=1 and γ=0.002.



References for Supporting Online Material

1. Friedel, M., Nikolajewa, S., Sühnel, J. and Wilhelm, T. (2009) DiProGB: the dinucleotide properties genome browser. *Bioinforma. Oxf. Engl.*, **25**, 2603–2604.

2. Ding, Y. (2006) Statistical and Bayesian approaches to RNA secondary structure prediction. *RNA N. Y. N*, **12**, 323–331.

3. Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

4. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.

5. Chan, C.Y. and Ding, Y. (2008) Boltzmann ensemble features of RNA secondary structures: a comparative analysis of biological RNA sequences and random shuffles. *J. Math. Biol.*, **56**, 93–105.

6. Moran, N.E., Rogers, R.B., Lu, C.-H., Conlon, L.E., Lila, M.A., Clinton, S.K. and Erdman, J.W., Jr (2013) Biosynthesis of highly enriched 13C-lycopene for human metabolic studies using repeated batch tomato cell culturing with 13C-glucose. *Food Chem.*, **139**, 631–639.