Mol Genet Genomics (2016) 291:473–481
DOI 10.1007/s00438-015-1078-7

METHODOLOGY ARTICLE

repRNA: a web server for generating various feature vectors of RNA sequences

Bin Liu[1,2,3] · Fule Liu[1] · Longyun Fang[1] · Xiaolong Wang[1,2] · Kuo-Chen Chou[3,4]

Received: 24 April 2015 / Accepted: 4 June 2015 / Published online: 18 June 2015

Abstract With the rapid growth of RNA sequences generated in the postgenomic age, it is highly desired to develop a flexible method that can generate various kinds of vectors to represent these sequences by focusing on their different features. This is because nearly all the existing machine-learning methods, such as SVM (support vector machine) and KNN (*k*-nearest neighbor), can only handle vectors but not sequences. To meet the increasing demands and speed up the genome analyses, we have developed a new web server, called "representations of RNA sequences" (repRNA). Compared with the existing methods, repRNA is much more comprehensive, flexible and powerful, as reflected by the following facts: (1) it can generate 11 different modes of feature vectors for users to choose according to their investigation purposes; (2) it allows users to select the features from 22 built-in physicochemical properties and even those defined by users' own; (3) the resultant feature vectors and the secondary structures of the corresponding RNA sequences can be visualized. The repRNA web server is freely accessible to the public at http://bioinformatics.hitsz.edu.cn/repRNA/.

Keywords repRNA · Secondary structure of RNA · PseAAC · PseKNC · repDNA · Physicochemical properties · User-defined properties

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-015-1078-7) contains supplementary material, which is available to authorized users.

✉ Bin Liu
bliu@gordonlifescience.org; bliu@insun.hit.edu.cn

Fule Liu
liufule12@gmail.com

Longyun Fang
dragoncloudest@gmail.com

Xiaolong Wang
wangxl@insun.hit.edu.cn

Kuo-Chen Chou
kcchou@gordonlifescience.org

Introduction

As indicated by a series of recent publications (Chen et al. 2014b; Ding et al. 2014; Guo et al. 2014; Lin et al. 2014; Liu et al. 2014b; Zhong and Zhou 2014; Chou 2015), one of the most challenging problems in computational biology is how to formulate a biological sequence with a discrete model or vector, yet still considerably keep its sequence order information or grasp its core features. This is because almost all the existing algorithms can only handle vectors but not sequence samples (Chou 2015; Liu et al. 2015e).

For protein and peptide sequences, the formulation of pseudo-amino acid composition (Chou 2001, 2005) or Chou's PseAAC (Lin and Lapointe 2013) and its web server generators (Shen and Chou 2008; Du et al. 2012,

[1] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, Guangdong, China

[2] Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, Guangdong, China

[3] Gordon Life Science Institute, Boston, MA 02478, USA

[4] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

2014; Cao et al. 2013) are quite successful in this regard and widely used in nearly all the fields of computational proteomics [see, e.g., (Zhou et al. 2007; Nanni and Lumini 2008; Georgiou et al. 2009; Esmaeili et al. 2010; Sahu and Panda 2010; Mohabatkar et al. 2011, 2013; Mohammad Beigi et al. 2011; Mei 2012; Chen and Li 2013; Georgiou et al. 2013; Liu et al. 2013, 2015a, f; Mondal and Pai 2014; Dehzangi et al. 2015; Khan et al. 2015; Kumar et al. 2015; Mandal et al. 2015; Liu et al. 2015f)] as well as a long list of papers cited in (Chou 2011; Du et al. 2014; Liu et al. 2015h).

For DNA sequences, the kmers (Fletez-Brant et al. 2013) and gapped kmers (Ghandi et al. 2014) were successfully applied to predict regulatory sequences, achieving quite promising outcomes (Lee et al. 2011; Fletez-Brant et al. 2013; Ghandi et al. 2014). Meanwhile, an extension of Chou's PseAAC called PseKNC or "pseudo *K*-tuple nucleotide composition" (Chen et al.) was developed and used to address some important problems in genome analysis [see, e.g., (Chen et al. 2013, 2014a; Guo et al. 2014; Qiu et al. 2014; Liu et al. 2015g)], and the corresponding web server generators have been established as well (Chen et al. 2014b, 2015).

For RNA sequences, however, so far only the standalone tool PseKNC-General (Chen et al. 2015) can be used to generate their feature vectors. But PseKNC-General is limited to certain types of features and a small number of physicochemical properties. Therefore, a comprehensive and flexible web server is urgently needed in this regard.

Here, we are to propose a web server called "Representations of RNA Sequences" (repRNA). To our best knowledge, repRNA is so far the most comprehensive and flexible web server, which can generate various modes of RNA feature vectors by means of the built-in physicochemical properties and even those defined by users themselves. Moreover, repRNA allows users to visualize the resultant feature vectors of the RNA sequences concerned as well as their computed secondary structures, so as to facilitate users to conduct in-depth genome analysis.

Method outline

The repRNA is a web server that can generate 11 different feature vectors for RNA sequences (Table 1), which can be grouped into three categories: oligonucleotide or *K*-tuple nucleotide composition, pseudo-nucleotide composition, and structure composition. The first one is to generate the pseudo-components for the short-range or local sequence order information by counting the occurrence frequencies of the *k* nearest residues along the sequence. The second category is to incorporate the long-range or global sequence order information by counting the correlations of dinucleotides along the sequence as shown in Fig. 1. The third category is to incorporate the local and global sequence-pattern information via the computed secondary structures of RNA sequences, as illustrated in Fig. 2.

K-tuple nucleotide composition

The first category is of oligonucleotide or *K*-tuple nucleotide composition that contains six modes (Zhang et al. 2011): (1) mononucleotide composition, (2) dinucleotide composition, (3) trinucleotide composition, (4) tetranucleotide composition, (5) pentanucleotide composition, and (6) hexanucleotide composition. These features can capture the short-range or local sequence order information (Chen et al. 2014b), and have been extensively used to study the characteristics of RNA sequences (Zhang et al. 2011; Wei et al. 2014).

Suppose an RNA sequence **R** with *L* nucleic acid residues; i.e.,

$$\mathrm{R} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7 \ldots \mathrm{R}_L, \quad (1)$$

Table 1 List of 11 modes of feature vectors for RNA sequences that can be generated by repRNA

Category	Feature vector modes
K-tuple nucleotide composition (Zhang et al. 2011)	Mononucleotide composition
	Dinucleotide composition
	Trinucleotide composition
	Tetranucleotide composition
	Pentanucleotide composition
	Hexanucleotide composition
Pseudo-nucleotide composition (Chen et al. 2015)	Parallel type of pseudo-dinucleotide composition (pPseDNC)
	Series type of pseudo-dinucleotide composition (sPseDNC)
Structure composition (Liu et al. 2015c, d)	Triplet
	Pseudo-structure status composition (PseSSC)
	Pseudo-distance structure status pair composition (PseDPC)

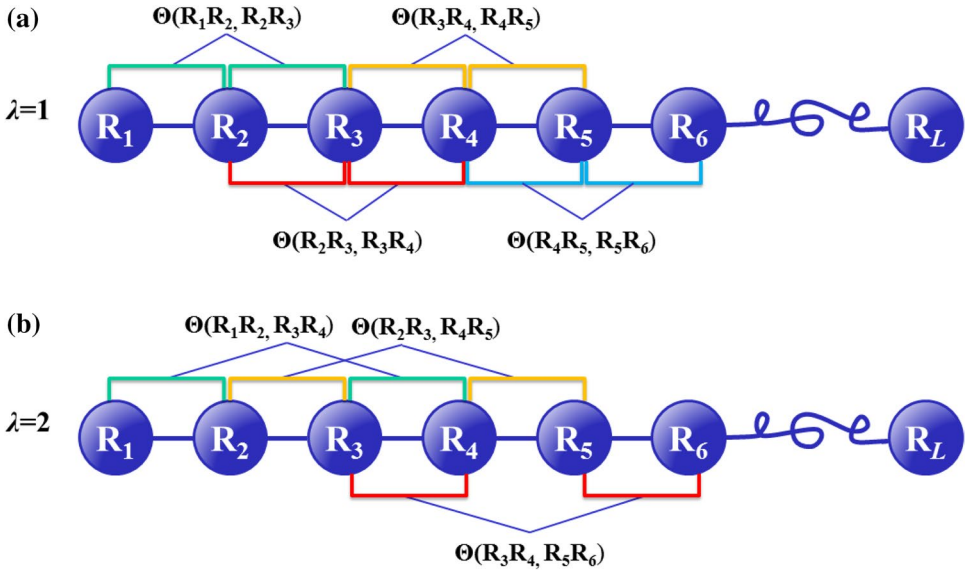

Fig. 1 A schematic illustration to show the correlation of dinucleotides along an RNA sequence. **a** The first-tier correlation reflects the sequence-order mode between all the most contiguous dinucleotide. **b** The second-tier correlation reflects the sequence-order mode between all the second-most contiguous dinucleotide. $\Theta(R_iR_{i+1}, R_jR_{j+1})$ is a coupling factor between the dinucleotide (R_iR_{i+1}) and dinucleotide (R_jR_{j+1}). The parameter λ is an integer, representing the counted rank (or tier) of the correlation along an RNA sequence

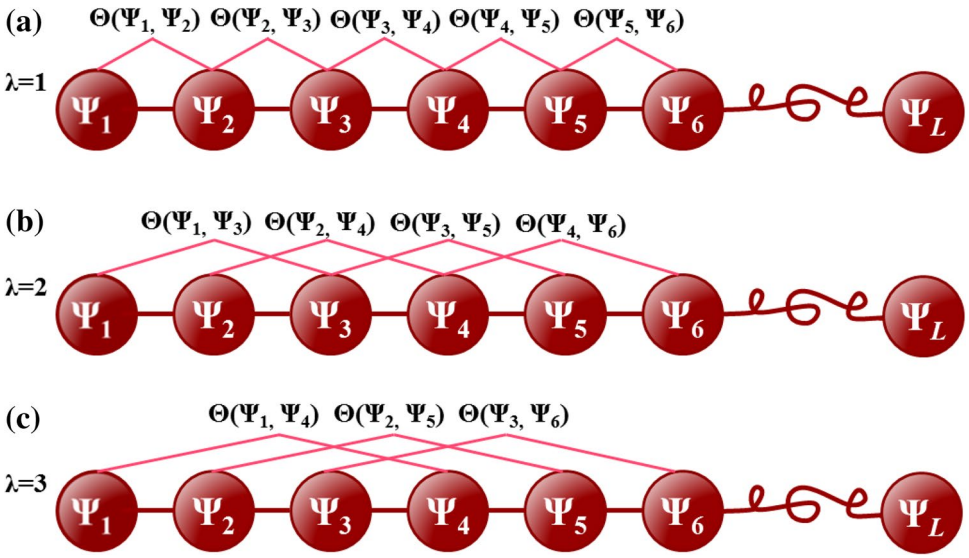

Fig. 2 A schematic drawing to show the correlations of structure statuses along an RNA sequence. **a** The first-tier correlation reflects the structure-order mode between all the most contiguous nucleotides. **b** The second-tier correlation reflects the structure-order mode between all the second-most contiguous nucleotides. **c** The third-tier correlation reflects the structure-order mode between all the third-most contiguous nucleotides. $\Theta(\Psi_i, \Psi_j)$ is a correlation function reflecting the structure-order information between the *i*th structure status and the *j*th structure status. λ is an integer, representing the counted rank (or tier) of the structural correlation along an RNA chain

where

$$R_i \in \{\text{A(adenine), C(cytosine), G(guanine), U(uracil)}\} \quad (2)$$

represents the nucleic acid residue at the sequence position i (= 1, 2, …, L). We can use K-tuple nucleotide composition (substring of RNA sequences containing k neighboring nucleotides) to represent an RNA sequence, then we have 4^k components in the corresponding vector **D** for the RNA sequence, that is,

$$\mathbf{D} = [f_1^{K\text{ - tuple}} \; f_2^{K\text{ - tuple}} \; f_3^{K\text{ - tuple}} \; f_4^{K\text{ - tuple}} \; \ldots \; f_{4^k}^{K\text{ - tuple}}]^{\mathrm{T}}, \quad (3)$$

where the symbol T is the transpose operator, $f^{K\text{ - tuple}}$ represents the frequency of the K-tuple in **R**. For example, by using the dinucleotide composition (2-tuple), the RNA sequence is represented as:

$$\begin{aligned}\mathbf{D} &= [f(\text{AA}) \; f(\text{AC}) \; f(\text{AG}) \; f(\text{AU}) \; \ldots \; f(\text{UU})]^{\mathrm{T}} \\ &= [f_1^{2\text{ - tuple}} \; f_2^{2\text{ - tuple}} \; f_3^{2\text{ - tuple}} \; f_4^{2\text{ - tuple}} \; \ldots \; f_{16}^{2\text{ - tuple}}]^{\mathrm{T}} \quad (4)\end{aligned}$$

where $f_1^{2\text{ - tuple}} = f(AA)$ is the normalized occurrence frequency of AA in the RNA sequence, $f_2^{2\text{ - tuple}} = f(\text{AC})$ is that of AC, $f_3^{2\text{ - tuple}} = f(\text{AG})$ is that of AG, and so forth.

The symbol T is the transpose operator, $f_1^{2\text{ - tuple}} = f(AA)$ is the normalized occurrence frequency of AA in the RNA sequence, $f_2^{2\text{ - tuple}} = f(\mathrm{AC})$ is that of AC, $f_3^{2\text{ - tuple}} = f(\mathrm{AG})$ is that of AG, and so forth. The other five sequence composition modes can be generated by using different K-tuples with $K = 1, 3, 4, 5, 6$ for mononucleotide composition, trinucleotide composition, tetranucleotide composition, pentanucleotide composition, and hexanucleotide composition, respectively.

Pseudo-ribonucleic acid composition

The secondary category is of pseudo-nucleotide composition that contains two modes (Chen et al. 2014b, 2015): parallel type of pseudo-dinucleotide composition (pPseDNC), and series type of pseudo-dinucleotide composition (sPseDNC). See refs. (Chou 2005, 2001) for the definitions of parallel type and series type, respectively, originally used in pseudo-amino acid composition. The feature vectors obtained in this category can contain considerable long-range or global sequence order information via the physicochemical properties of dinucleotides.

Pseudo-dinucleotide composition (pPseNNC)

In parallel type of pseudo-dinucleotide composition (pPseDNC), the users cannot only select the 22 built-in physiochemical indices (Online Supporting Information S1), but also can upload their own indices to generate the pPseDNC feature vector.

Given an RNA sequence **R** (Eq. 1), the pPseDNC feature vector of **R** is defined:

$$\mathbf{R} = \left[\, d_1 \;\; d_2 \;\; \ldots \;\; d_{16} \;\; d_{16+1} \;\; \ldots \;\; d_{16+\lambda} \,\right]^{\mathrm{T}}, \tag{5}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda} \theta_j} & (16+1 \le k \le 16+\lambda) \end{cases}, \tag{6}$$

where f_k ($k = 1, 2, \ldots, 16$) is the normalized occurrence frequency of dinucleotide in the RNA sequence; the parameter λ is an integer, representing the highest counted rank (or tier) of the correlation along an RNA sequence; w is the weight factor ranging from 0 to 1; θ_j ($j = 1, 2, \ldots, \lambda$) is called the j-tier correlation factor reflecting the sequence-order correlation between all the most contiguous dinucleotides along an RNA sequence, which is defined:

$$\begin{cases} \theta_1 = \frac{1}{L-2} \sum_{i=1}^{L-2} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+1}\mathrm{R}_{i+2}) \\ \theta_2 = \frac{1}{L-3} \sum_{i=1}^{L-3} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+2}\mathrm{R}_{i+3}) \\ \theta_3 = \frac{1}{L-4} \sum_{i=1}^{L-4} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+3}\mathrm{R}_{i+4}) \qquad (\lambda < L) \\ \ldots\ldots \\ \theta_\lambda = \frac{1}{L-1-\lambda} \sum_{i=1}^{L-1-\lambda} \Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_{i+\lambda}\mathrm{R}_{i+\lambda+1}) \end{cases} \tag{7}$$

where the correlation function is given by

$$\Theta(\mathrm{R}_i\mathrm{R}_{i+1}, \mathrm{R}_j\mathrm{R}_{j+1}) = \frac{1}{\mu} \sum_{u=1}^{\mu} [P_u(R\mathrm{R}_i\mathrm{R}_{i+1}) - P_u(\mathrm{R}_j\mathrm{R}_{j+1})]^2, \tag{8}$$

where μ is the number of physicochemical indices considered that are listed in the Online Supporting Information S1; $P_u(\mathrm{R}_i\mathrm{R}_{i+1})$ represents the numerical value of the uth ($u = 1, 2, \ldots, \mu$) physicochemical index for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$ at the position i, and so forth.

Series type of pseudo-dinucleotide composition (sPseDNC)

Series type of pseudo-dinucleotide composition is a variant of pPseDNC, which differs in the equations of calculating the correlation factors reflecting the sequence-order correlation between all the most contiguous dinucleotides along an RNA sequence.

Given an RNA sequence R (Eq. 1), the sPseDNC feature vector of **R** is defined:

$$\mathbf{R} = \left[\, d_1 \;\; d_2 \;\; \ldots \;\; d_{16} \;\; d_{16+1} \;\; \ldots \;\; d_{16+\lambda} \;\; d_{16+\lambda+1} \;\; \ldots \;\; d_{16+\lambda\Lambda} \,\right]^{T}, \tag{9}$$

where

$$d_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda\Lambda} \theta_j} & (1 \le k \le 16) \\ \dfrac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\lambda\Lambda} \theta_j} & (16+1 \le k \le 16+\lambda\Lambda) \end{cases}, \tag{10}$$

where f_k ($k = 1, 2, \ldots, 16$) is the normalized occurrence frequency of dinucleotides in the RNA sequence; the parameter λ is an integer, representing the highest counted rank (or tier) of the correlation along an RNA sequence; w is the weight factor ranging from 0 to 1; Λ is the number of physicochemical indices (Online Supporting Information S1); θ_j ($j = 1, 2, \ldots, \lambda$) is called the j-tier correlation factor reflecting the sequence-order correlation between all the most contiguous dinucleotides along an RNA sequence, which is defined:

$$\begin{cases} \theta_1 = \frac{1}{L-3}\sum_{i=1}^{L-3} \mathrm{J}_{i,i+1}^{1} \\ \theta_2 = \frac{1}{L-3}\sum_{i=1}^{L-3} \mathrm{J}_{i,i+1}^{2} \\ \cdots\cdots \\ \theta_\Lambda = \frac{1}{L-3}\sum_{i=1}^{L-3} \mathrm{J}_{i,i+1}^{\Lambda} \qquad \lambda < (L-2) \\ \cdots\cdots \\ \theta_{\lambda\Lambda-1} = \frac{1}{L-\lambda-2}\sum_{i=1}^{L-\lambda-2} \mathrm{J}_{i,i+\lambda}^{\Lambda-1} \\ \theta_{\lambda\Lambda} = \frac{1}{L-\lambda-2}\sum_{i=1}^{L-\lambda-2} \mathrm{J}_{i,i+\lambda}^{\Lambda} \end{cases} \tag{11}$$

The correlation function is given by

$$\begin{cases} \mathrm{J}_{i,i+m}^{u} = P_u(\mathrm{R}_i\mathrm{R}_{i+1}) \cdot P_u(\mathrm{R}_{i+m}\mathrm{R}_{i+m+1}) \\ u = 1,2,\ldots,\Lambda;\ m = 1,2,\ldots,\lambda;\ i = 1,2,\ldots,L-\lambda-2 \end{cases}, \tag{12}$$

where $P_u(\mathrm{R}_i\mathrm{R}_{i+1})$ represents the numerical value of the uth ($u = 1, 2, \ldots, \mu$) physiochemical index for the dinucleotide $\mathrm{R}_i\mathrm{R}_{i+1}$ at position i, and so forth.

Structure composition

The third category is of structure composition that contains three modes (Xue et al. 2005; Liu et al. 2015b, d, e): (1) triplet, (2) pseudo-structure status composition (PseSSC), and (3) pseudo-distance structure status pair composition (PseDPC). The features in this category are derived by using the local and global sequence-pattern information via the computed secondary structures of RNA sequences (Lorenz et al. 2011). The feature vectors thus obtained have been successfully applied to identify microRNA precursors (Liu et al. 2015b, d, e).

Triplet

The Triplet (Xue et al. 2005) is an early approach to use the structure information of RNA sequences, and showed better performance for microRNA identification compared with other sequence-based methods.

Given an RNA sequence **R** (Eq. 1), formulating it according to its secondary structure derived from the Vienna RNA software package (Lorenz et al. 2011) (released 2.1.6), we have

$$\mathbf{R} = \Psi_1\Psi_2\Psi_3\Psi_4\Psi_5\ldots\Psi_L, \tag{13}$$

where Ψ_1 denotes the structure status of R_1, Ψ_2 the structure status of R_2, and so forth.

In the predicted secondary structure, there are only two statuses for each nucleotide, paired or unpaired, indicated by brackets "(“or")" and dots ".", respectively. The left bracket "(" means that the paired nucleotide is located near the 5′-end and can be paired with another nucleotide at the 3′-end, which is indicated by a right bracket ")". We do not distinguish these two situations and use "(" for both situations. For any three adjacent nucleotides, there are 8 (2^3) possible structure compositions: "(((", "((.", "(..", "(.(", ".((", ".(.", "..(" and "…". Considering the middle nucleotide among the 3 adjacent nucleotides, there are 32 (4×8) possible structure–sequence combinations, which we denote as f_A("((("), f_C("((("), etc.

Therefore, Triplet approach formulates a feature vector containing 32 (4×8) components as given by

$$\mathbf{D} = \left[f_{\mathrm{A}}(''((('') \quad f_{\mathrm{A}}(''((.'') \quad \ldots \quad f_{\mathrm{A}}(''\ldots'') \right. \\ \left. f_{\mathrm{C}}(''((('') \quad \ldots \quad f_{\mathrm{U}}(''\ldots'')\right]^{\mathrm{T}}, \tag{14}$$

where f represents the normalized occurrence frequency of the structure–sequence compositions.

Pseudo-structure status composition (PseSSC)

Given an RNA sequence **R** (Eq. 1), we can formulate its secondary structure as Eq. 13. They can be any of the 10 structure statuses; i.e.,

$$\Psi_i \in \{\mathrm{A},\ \mathrm{C},\ \mathrm{G},\ \mathrm{U},\ \mathrm{A}{-}\mathrm{U},\ \mathrm{U}{-}\mathrm{A},\ \mathrm{G}{-}\mathrm{C},\ \mathrm{C}{-}\mathrm{G}, \\ \mathrm{G}{-}\mathrm{U},\ \mathrm{U}{-}\mathrm{G}\} \quad i = 1,2,\ldots,L, \tag{15}$$

where A, C, G, U represent the structure statuses of the four unpaired nucleobases, while A–U, U–A, G–C, C–G, G–U, U–G represent the structure statuses of the six paired bases.

The PseSSC approach formulates a feature vector containing $10^n + \lambda$ components as given by

$$\mathrm{R} = \left[f_1^* \ f_2^* \ f_3^* \ \ldots f_{10^n}^* \ f_{10^n+1}^* \ \ldots f_{10^n+\lambda}^*\right]^T, \tag{16}$$

where

$$f^* = \begin{cases} \frac{f_u}{\sum_{i=1}^{10^n} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le 10^n) \\ \frac{w\theta_{u-10^n}}{\sum_{i=1}^{10^n} f_i + w\sum_{j=1}^{\lambda}\theta_j} & (10^n + 1 \le u \le 10^n + \lambda) \end{cases}, \tag{17}$$

where $f_i (i = 1, 2, \ldots, 10^n)$ represents the normalized occurrence frequency of the structure status combination of n adjacent nucleobases, w is the weight factor used to adjust the effect of the correlation factors, and θ_j the j-tier sequence correlation factor given by

$$\begin{cases} \theta_1 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1}\Theta(\Psi_i,\Psi_{i+1}) \\ \theta_2 = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2}\Theta(\Psi_i,\Psi_{i+2}) \\ \theta_3 = \dfrac{1}{L-3}\displaystyle\sum_{i=1}^{L-3}\Theta(\Psi_i,\Psi_{i+3}) \\ \cdots\cdots \\ \theta_\lambda = \dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda}\Theta(\Psi_i,\Psi_{i+\lambda}) \end{cases} (\lambda < L), \tag{18}$$

where λ is an integer, representing the highest counted rank (or tier) of the structural correlation along an RNA chain; θ_i is the ith-tier correlation factor reflecting the structure-order information between all the ith most contiguous bases along an RNA chain, and the correlation function $\Theta(\Psi_i, \Psi_j)$ is given by

$$\Theta(\Psi_i, \Psi_j) = \left[F(\Psi_i) - F(\Psi_j)\right]^2, \tag{19}$$

where $F(\Psi_i)$ is the free energy of the structure status Ψ_i of the nucleobase at position i, and $F(\Psi_j)$ is the free energy of the structure status Ψ_j of the nucleobase at position j.

Pseudo-distance structure status pair composition (PseDPC)

Given an RNA sequence **R** (Eq. 1), its feature vector (Eq. 13) can also be formulated as follows. In order to capture the structure-order information of the RNA sequence **R**, a concept called the occurrences of "distance structure status pair" or just "distance-pair", $D\left(\Psi_i,\Psi_j|k\right)$ has been proposed, as formulated by

$$\begin{cases} D\left(\Psi_i,\Psi_j|0\right) & \text{if } k = 0 \text{ then } i = j \\ D\left(\Psi_i,\Psi_j|1\right) & \text{if } k = 1 \\ D\left(\Psi_i,\Psi_j|2\right) & \text{if } k = 2 \\ \vdots & \vdots \\ D\left(\Psi_i,\Psi_j|L-1\right) & \text{if } k = L-1 \end{cases} \tag{20}$$

where Ψ_i and Ψ_i can be any of the 10 structure statuses of an RNA chain **R** (cf. Eq. 17), and k ($0 \le k \le L-1$) represents the value counted by the distance between structure statuses the distance between structure and Ψ_j along the RNA chain **R**. Suppose Ψ_i is A–U, Ψ_j is U–G, and $k = 3$, then D(A–U, U–G|3) means the structure status pair (A–U, U–G) with its two counterparts separated by two nucleotides along the RNA chain **R**.

The approach PseDPC formulates a feature vector as below:

$$[d_1\, d_2\, d_3\, \ldots\, d_u\, \ldots\, d_\Omega\, d_{\Omega+1}\, d_{\Omega+1}\, \ldots\, d_{\Omega+\lambda}]^{\mathrm{T}}, \tag{21}$$

where

$$d_u = \begin{cases} \dfrac{f_u}{1+w\sum_{j=1}^{\lambda}\theta_j} & (1 \le u \le \Omega) \\ \dfrac{w\theta_{u-\Omega}}{1+w\sum_{j=1}^{\lambda}\theta_j} & (\Omega+1 \le u \le \Omega+\lambda) \end{cases}, \tag{22}$$

where θ_j is the j-tier sequence correlation factor computed by Eq. 18, w is the weight factor used to adjust the effect of the correlation factors, $\Omega = 10 + 100n$, where n represents the maximum distance between two structure statuses, and f_u is the occurrences of the distance-pairs $D(\Psi_i, \Psi_j|k)$ calculated by

$$f_u = \begin{cases} f\left(D\left(\Psi_i,\Psi_j|0\right)\right) & \text{if } 1 \le u \le 10 \\ f\left(D\left(\Psi_i,\Psi_j|1\right)\right) & \text{if } 11 \le u \le 110 \\ f\left(D\left(\Psi_i,\Psi_j|2\right)\right) & \text{if } 111 \le u \le 210 \\ \vdots & \vdots \\ f\left(D\left(\Psi_i,\Psi_j|n\right)\right) & \text{if } 10+100(n-1) \le u \le 10+100n \end{cases} \tag{23}$$

Description of repRNA server

The repRNA web server has a simple user interface that consists of three input fields: (1) feature-vector mode selection, (2) parameter settings, and (3) sequence to be analyzed. Users should first select any one of the 11 feature vector modes listed in Table 1 according to their need, followed by setting the corresponding parameters. The RNA sequences should be in FASTA format and they can be input by the way of either copying/pasting or uploading a file containing the sequences concerned. To see the input sequence format, click the Example button. The web server accepts as many as 50 input sequences for each submission.

After clicking the submit button, if your input contains any invalid stuff, an error message will be prompted; otherwise, you will see the output on the screen. The output generated by the repRNA server contains a parameter summery and a result section. The former lists all the selected parameters used for deriving the feature vectors; the latter shows the numerical feature vectors thus obtained for the corresponding RNA sequences in the input. The feature vectors can be downloaded into a separate file suitable for downstream computational analyses by various algorithms, such support vector machine, neural network, and covariant discriminant algorithm. For all the 11 modes (Table 1), the resultant feature

vectors can be visualized by an intuitive graphical representation called "heat map" to see the distributions of their feature values. For the three modes in the structure composition category, the secondary structures of the input RNA sequences can also be visualized to see their structure characteristics.

Applications of repRNA

As demonstrated in a series of recent publications (see, e.g., Liu et al. 2014a; Dehzangi et al. 2015; Liu et al. 2015g) in complying with the Chou's 5-step rule (Chou 2011), to establish a really useful statistical predictor for a biological system, one needs to consider the following five guidelines: (1) construct or select a valid benchmark dataset to train and test the predictor; (2) formulate the statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful machine-learning algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the model; (5) establish a user-friendly web server for the predictor that is accessible to the public.

Among the aforementioned five steps, the most difficult and time-consuming job is in the second step; i.e., how to find an effective digit feature vector to represent the RNA sequence concerned. Using the repRNA web server as presented in this article, we can easily generate all these desired feature vectors by just selecting different parameters. For example, in studying the identification of microRNA precursors (Liu et al. 2015c), the authors used the feature vectors of pseudo-structure status composition (PseSSC) to represent the RNA samples after spending a lot of time for mathematical derivations. In contrast, if using the current repRNA web server by selecting 'PseSSC' for the mode and $n = 2$, $\lambda = 13$, $w = 0.5$ for the parameters, we can immediately obtain the exactly same feature vectors as used in Liu et al. (2015c), substantially expediting the process of developing a new method for analyzing RNA sequences. Therefore, the application values of repRNA are self-evident.

Conclusions

To our best knowledge, repRNA is so far the most flexible and comprehensive web server for generating the feature vectors based on the RNA sequence information alone. The performance and efficiency of these feature vectors have been validated by a series of publications (Xue et al. 2005; Zhang et al. 2011; Wei et al. 2014; Liu et al. 2015c, d). Compared with the existing method, repRNA has the following advantages: (1) it contains a total of 14 modes for generating the features; (2) various built-in and user-defined physicochemical properties can be used for the computation; (3) the resultant feature vectors and the secondary structures of the input RNA sequences can be visualized. It is anticipated that repRNA will become a useful high-throughput tool, expediting analysis of uncharacterized RNA sequences.

Acknowledgments The authors wish to thank the three anonymous reviewers for the constructive comments, which were very useful to strengthening the presentation of this paper.

Conflict of interest The authors declare no competing interests.

Funding This work was supported by the National Natural Science Foundation of China (61300112 and 61272383), the Scientific Research Innovation Foundation in Harbin Institute of Technology (Project No. HIT.NSRIF.2013103), the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and Shenzhen Municipal Science and Technology Innovation Council (Grant No. CXZZ20140904154910774), and National High Technology Research and Development Program of China (863 Program) [2015AA015405].

References

Cao DS, Xu QS, Liang YZ (2013) propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29(7):960–962

Chen YK, Li KB (2013) Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. J Theor Biol 318:1–12

Chen W, Feng PM, Lin H (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41(6):e68

Chen W, Feng PM, Lin H (2014a) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Research International (BMRI) 2014:623149

Chen W, Lei TY, Jin DC, Lin H (2014b) PseKNC: a flexible web-server for generating pseudo *K*-tuple nucleotide composition. Anal Biochem 456:53–60

Chen W, Zhang X, Brooker J, Lin H, Zhang L (2015) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 31(1):119–120

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: structure. Function, and Genetics 43:246–255

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21(1):10–19

Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review). J Theor Biol 273(1):236–247

Chou KC (2015) Impacts of bioinformatics to medicinal chemistry. Med Chem 11(3):218–234

Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A (2015) Gram-positive and Gram-negative protein subcellular

localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J Theor Biol 364:284–294
Ding H, Deng EZ, Yuan LF, Liu L, Lin H (2014) iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Research International (BMRI) 2014:286419
Du P, Wang X, Xu C, Gao Y (2012) PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal Biochem 425(2):117–119
Du P, Gu S, Jiao Y (2014) PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. Int J Mol Sci 15(3):3495–3506
Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J Theor Biol 263(2):203–209
Fletez-Brant C, Lee D, McCallion AS, Beer MA (2013) kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic Acids Res 41:W544–W556
Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J Theor Biol 257(1):17–26
Georgiou DN, Karakasidis TE, Megaritis AC (2013) A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory. Open Bioinform J 7:41–48
Ghandi M, Lee D, Mohammad-Noori M, Beer MA (2014) Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput Biol 10(7):e1003711
Guo SH, Deng EZ, Xu LQ, Ding H, Lin H (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo *K*-tuple nucleotide composition. Bioinformatics 30(11):1522–1529
Khan ZU, Hayat M, Khan MA (2015) Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J Theor Biol 365:197–203
Kumar R, Srivastava A, Kumari B, Kumar M (2015) Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. J Theor Biol 365:96–103
Lee D, Karchin R, Beer MA (2011) Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res 21(12):2167–2180
Lin S-X, Lapointe J (2013) Theoretical and experimental biology in one. J Biomed Sci Eng 06(04):435–442
Lin H, Deng EZ, Ding H, Chen W (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo *K*-tuple nucleotide composition. Nucleic Acids Res 42(21):12961–12972
Liu B, Wang X, Zou Q, Dong Q, Chen Q (2013) Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. Mol Inform 32:775–782
Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X (2014a) iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS One 9(9):e106691
Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q (2014b) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 30(4):472–479
Liu B, Chen J, Wang X (2015a) Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis. MGG. doi:10.1007/s00438-00015-01044-00434
Liu B, Fang L, Jie C, Liu F, Wang X (2015b) miRNA-dis: microRNA precursor identification based on distance structure status pairs. Mol BioSyst 11:1194–1204
Liu B, Fang L, Liu F, Wang X, Chen J (2015c) Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS One 10:e0121501
Liu B, Fang L, Liu F, Wang X (2015d) iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J Biomol Struct Dyn. doi:10.1080/07391102.2015.1014422
Liu B, Liu F, Fang L, Wang X (2015e) repDNA: a python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinf 31(8):1307–1309
Liu B, Xu J, Fan S, Xu R, Jiyun Zhou J, Wang X (2015f) PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation. Mol Inform 34:8–17
Liu Z, Xiao X, Qiu WR (2015g) iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal Biochem 474:69–77
Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC (2015h) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. doi:10.1093/nar/gkv458
Lorenz R, Bernhart SH, Siederdissen CHz, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011). ViennaRNA Package 2.0. Algorithms Mol Biol 6(26)
Mandal M, Mukhopadhyay A, Maulik U (2015) Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC. Med Biol Eng Comput 53(4):331–344
Mei S (2012) Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. J Theor Biol 293:121–130
Mohabatkar H, Mohammad Beigi M, Esmaeili A (2011) Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. J Theor Biol 281(1):18–23
Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S (2013) Prediction of allergenic proteins by means of the concept of chou's pseudo amino acid composition and a machine learning approach. Med Chem 9(1):133–137
Mohammad Beigi M, Behjati M, Mohabatkar H (2011) Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. J Struct Funct Genomics 12(4):191–197
Mondal S, Pai PP (2014) Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J Theor Biol 356:30–35
Nanni L, Lumini A (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34:653–660
Qiu WR, Xiao X, Chou KC (2014) iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci 15(2):1746–1766
Sahu SS, Panda G (2010) A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput Biol Chem 34(5–6):320–327
Shen HB, Chou KC (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373(2):386–388
Wei L, Liao M, Gao Y, Ji R, He Z, Zou Q (2014) Improved and promising identification of human microRNAs by incorporating a high-quality negative set. IEEE/ACM Trans Comput Biol Bioinf 11:192–201

Xue C, Li F, He T, Liu GP, Li Y, Zhang X (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics 6:310

Zhang Y, Wang X, Kang L (2011) A k-mer scheme to predict piRNAs and characterize locust piRNAs. Bioinformatics 27(6):771–776

Zhong WZ, Zhou SF (2014) Molecular science for drug development and biomedicine. Int J Mol Sci 15:20072–20078

Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248:546–551