Analytical Biochemistry 456 (2014) 53–60

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition

Wei Chen [a,b,*], Tian-Yu Lei [a], Dian-Chuan Jin [a], Hao Lin [b,c,*], Kuo-Chen Chou [a,b,d,*]

[a] School of Sciences, and Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China
[b] Gordon Life Science Institute, Belmont, MA 02478, USA
[c] Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[d] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 29 December 2013
Received in revised form 20 March 2014
Accepted 1 April 2014
Available online 13 April 2014

Keywords:
Pseudo oligonucleotide composition
DNA sequence representation
PseAAC
Global sequence order information
Physicochemical properties
Web server

ABSTRACT

The pseudo oligonucleotide composition, or pseudo K-tuple nucleotide composition (PseKNC), can be used to represent a DNA or RNA sequence with a discrete model or vector yet still keep considerable sequence order information, particularly the global or long-range sequence order information, via the physicochemical properties of its constituent oligonucleotides. Therefore, the PseKNC approach may hold very high potential for enhancing the power in dealing with many problems in computational genomics and genome sequence analysis. However, dealing with different DNA or RNA problems may need different kinds of PseKNC. Here, we present a flexible and user-friendly web server for PseKNC (at http://lin.uestc.edu.cn/pseknc/default.aspx) by which users can easily generate many different modes of PseKNC according to their need by selecting various parameters and physicochemical properties. Furthermore, for the convenience of the vast majority of experimental scientists, a step-by-step guide is provided on how to use the current web server to generate their desired PseKNC without the need to follow the complicated mathematical equations, which are presented in this article just for the integrity of PseKNC formulation and its development. It is anticipated that the PseKNC web server will become a very useful tool in computational genomics and genome sequence analysis.

With the avalanche of DNA sequences generated in the post-genomic age, it is highly desirable to develop computational methods for rapidly and accurately identifying their biological features or attributes based on the sequence information alone. So far, many methods have been proposed to decode the complicated genomes or DNAs therein (see, e.g., Refs. [1,2]). Although considerable progress has been made in this regard, all these methods were merely based on the nucleic acid composition alone without taking into account the sequence order effect. Obviously, this is a compromise with the difficulty in dealing with the huge number of different sequence orders that a DNA sequence may possibly have, as elaborated below.

DNA sequences consist of four nucleotides (A, C, G, and T). Thus, for a DNA sequence of only 100 nucleotides, the number of different sequence order combinations would be $4^{100} = 10^{100\log 4} > 1.6065 \times 10^{60}$. Actually, the length of DNA sequences is much longer than 100, and hence the number of different combinations will be $\gg 1.6065 \times 10^{60}$. For such an astronomical number, it is impracticable to construct a reasonable training dataset to statistically cover all the possible different sequence order patterns. Besides, DNA sequences vary widely in length, which poses an additional difficulty for incorporating the sequence order information in both the benchmark dataset construction and the algorithm formulation. Faced with such a dilemma, can we find an approach to partially incorporate the sequence order effects?

Actually, a similar problem existed in computational proteomics as well. To address it, the pseudo amino acid composition [3], or Chou's PseAAC[1] [4], was proposed. Ever since the concept of PseAAC was introduced in 2001, it has penetrated into nearly all the fields of computational proteomics such as classifying enzyme

* Corresponding authors at: Gordon Life Science Institute, Belmont, MA 02478, USA.
E-mail addresses: wchen@gordonlifescience.org, chenweiimu@gmail.com (W. Chen), hlin@uestc.edu.cn (H. Lin), kcchou@gordonlifescience.org (K.-C. Chou).

[1] *Abbreviations used:* PseAAC, pseudo amino acid composition; GPCR, G-protein-coupled receptor; PseKNC, pseudo K-tuple nucleotide composition; PseDNC, pseudo dinucleotide composition; PseTNC, pseudo trinucleotide composition.

http://dx.doi.org/10.1016/j.ab.2014.04.001
0003-2697/

families [5], identifying submitochondrial localization of proteins [6], detecting remote homologous proteins [7], predicting G-protein-coupled receptor (GPCR) classes [8], predicting protein structural classes [9], identifying risk type of human papillomaviruses [10], predicting $GABA_A$ receptor proteins [11], predicting protein supersecondary structure [12], identifying bacterial virulent proteins [13], predicting anticancer peptides [14], identifying human GPCR N-linked glycosylation sites [15], identifying PTM (posttranslational modification) sites in proteins [16,17], discriminating outer membrane proteins [18], and predicting allergenic proteins [19] (see a long list of the references cited in Chou's review article [20]). Owing to its wide and increasing applications, recently three powerful software programs—PseAAC–Builder [21], propy [22], and PseAAC–General [23]—were established for generating various different modes of PseAAC in addition to the web server PseAAC [24] built in 2008.

Encouraged by the success of using the pseudo amino acid composition idea to deal with protein sequences, the corresponding approaches were proposed recently to deal with DNA sequences such as using the pseudo dinucleotide composition [25] and pseudo trinucleotide composition [26] to identify recombination spots, using pseudo trinucleotide composition [27] to identify promoters, and using pseudo K-tuple nucleotide composition to identify nucleosomes [28]. The pseudo oligonucleotide composition can be generally expressed as

$$\text{PseKNC} = \begin{cases} \text{Pseudo dinucleotide composition (PseDNC)}, & \text{when K} = 2 \\ \text{Pseudo trinucleotide composition (PseTNC)}, & \text{when K} = 3 \\ \vdots & \vdots \end{cases} \tag{1}$$

As we can see from above, there are many different ways to formulate PseKNC or pseudo K-tuple nucleotide composition. To deal with different problems in DNA, different modes of PseKNC may be needed to optimize the outcomes. In view of this, it would be very useful to provide a flexible web server by which users can generate various modes of pseudo K-tuple nucleotide composition as desired. The current study was initiated in an attempt to realize this.

Pseudo K-tuple nucleotide composition

Suppose a DNA sequence **D** with L nucleotides, that is,

$$\mathbf{D} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_\mathrm{L}, \tag{2}$$

where

$$\mathrm{R}_i \in \{\mathrm{A}\ (\text{adenine}), \mathrm{C}\ (\text{cytosine}), \mathrm{G}\ (\text{guanine}), \mathrm{T}\ (\text{thymine})\} \tag{3}$$

denotes the nucleic acid residue at the sequence position i (=1,2, $\cdots$,L). When the DNA sequence is represented by the dinucleotide composition, we have

$$\begin{aligned} \mathbf{D} &= [f(\mathrm{AA}) \quad f(\mathrm{AC}) \quad f(\mathrm{AG}) \quad f(\mathrm{AT}) \quad \cdots \quad f(\mathrm{TT})]^{\mathbf{T}} \\ &= \left[f_1^{\mathrm{di}} \quad f_2^{\mathrm{di}} \quad f_3^{\mathrm{di}} \quad f_4^{\mathrm{di}} \quad \cdots \quad f_{16}^{\mathrm{di}}\right]^{\mathbf{T}}, \end{aligned} \tag{4}$$

where the symbol T is the transpose operator, $f_1^{\mathrm{di}} = f(\mathrm{AA})$ is the normalized occurrence frequency of AA in the DNA sequence, $f_2^{\mathrm{di}} = f(\mathrm{AC})$ is that of AC, $f_3^{\mathrm{di}} = f(\mathrm{AG})$ is that of AG, and so forth. When the DNA sequence is represented by the trinucleotide composition, we have

$$\begin{aligned} \mathbf{D} &= [f(\mathrm{AAA}) \quad f(\mathrm{AAC}) \quad f(\mathrm{AAG}) \quad f(\mathrm{AAT}) \quad \cdots \quad f(\mathrm{TTT})]^{\mathbf{T}} \\ &= \left[f_1^{\mathrm{tri}} \quad f_2^{\mathrm{tri}} \quad f_3^{\mathrm{tri}} \quad f_4^{\mathrm{tri}} \quad \cdots \quad f_{64}^{\mathrm{tri}}\right]^{\mathbf{T}}, \end{aligned} \tag{5}$$

where $f_1^{\mathrm{tri}} = f(\mathrm{AAA})$ is the normalized occurrence frequency of AAA in the DNA sequence, $f_2^{\mathrm{tri}} = f(\mathrm{AAC})$ is that of AAC, and so forth. Generally speaking, if a DNA sequence is represented by the K-tuple nucleotide composition [29], the corresponding vector **D** for the DNA sequence will contain 4^{K} components, that is,

$$\mathbf{D} = \left[f_1^{\text{K-tuple}} \quad f_2^{\text{K-tuple}} \quad f_3^{\text{K-tuple}} \quad f_4^{\text{K-tuple}} \quad \cdots \quad f_{4^{\mathrm{K}}}^{\text{K-tuple}}\right]^{\mathbf{T}}. \tag{6}$$

As we can see from Eqs. (4)–(6), by increasing the value of K, although the base sequence order information within a local or short range could be gradually included, none of the global or long-range sequence order information would be reflected by this kind of oligonucleotide composition. To incorporate the global or long-range sequence order information for a DNA sequence, we need to use the pseudo oligonucleotide composition or PseKNC (see Eq. (1)) to represent DNA sequences, as in the case of using the pseudo amino acid composition or PseAAC [3,30] to represent protein sequences.

It has been proved that DNA physicochemical properties play important roles in gene expression regulation [31–33]. For example, DNA physicochemical property is evolutionarily more constrained than the underlying actual sequence, and the topography-informed constrained regions usually correlate with functional noncoding elements such as enhancers [34]. Accordingly, it is reasonable to use the physicochemical properties of nucleotides to formulate PseKNC for DNA sequences, just like using the physicochemical properties of amino acids to formulate PseAAC [3,30] for protein sequence.

Listed in Tables 1 and 2 are 38 dinucleotide physicochemical properties and 12 trinucleotide physicochemical properties that can be used to generate various different modes of pseudo dinucleotide composition (PseDNC) and pseudo trinucleotide composition

Table 1
List of 38 physicochemical properties of dinucleotides in DNA.

Number	Description	Reference
1	Base stacking	[37]
2	Protein-induced deformability	[38]
3	B-DNA twist	[39]
4	Dinucleotide GC content	[40]
5	A-philicity	[41]
6	Propeller twist	[42]
7	Duplex stability (free energy)	[43]
8	Duplex stability (disrupt energy)	[44]
9	DNA denaturation	[45]
10	Bending stiffness	[46]
11	Protein–DNA twist	[38]
12	Stabilizing energy of Z-DNA	[47]
13	Aida_BA_transition	[48]
14	Breslauer_dG	[44]
15	Breslauer_dH	[44]
16	Breslauer_dS	[44]
17	Electron interaction	[40]
18	Hartman_trans_free_energy	[49]
19	Helix–coil_transition	[50]
20	Ivanov_BA_transition	[41]
21	Lisser_BZ_transition	[51]
22	Polar_interaction	[52]
23	SantaLucia_dG	[53]
24	SantaLucia_dH	[53]
25	SantaLucia_dS	[53]
26	Sarai_flexibility	[54]
27	Stability	[55]
28	Stacking_energy	[37]
29	Sugimoto_dG	[53]
30	Sugimoto_dH	[53]
31	Sugimoto_dS	[53]
32	Watson–Crick_interaction	[56]
33	Twist	[57]
34	Tilt	[57]
35	Roll	[57]
36	Shift	[57]
37	Slide	[57]
38	Rise	[57]

Table 2
List of 12 physicochemical properties of trinucleotides in DNA.

Number	Description	Reference(s)
1	Bendability (DNase)	[58]
2	Bendability (consensus)	[58]
3	Trinucleotide GC content	[40]
4	Nucleosome positioning	[59]
5	Consensus–roll	[40,52]
6	Consensus–rigid	[40,52]
7	DNase I	[31]
8	DNase I–rigid	[31]
9	MW–daltons	[40]
10	MW–kg	[40]
11	Nucleosome	[60]
12	Nucleosome–rigid	[60]

(PseTNC), respectively. Because so far not much physicochemical property data are available for the K-tuple nucleotides when K = 4 (tetranucleotides) and above, the current study was limited to PseDNC and PseTNC only. Nevertheless, the formulations presented here can be easily extended to cover the case of PseKNC with $K \geqslant 4$ when the corresponding physicochemical property data are available.

In addition, like the PseAAC web server [24] that could generate two different types of pseudo amino acid composition for protein sequences—(i) the parallel correlation type [3] or type 1 PseAAC and (ii) the series correlation type [30] or type 2 PseAAC—here we make the PseKNC web server able to generate these two types of pseudo K-tuple nucleotide composition as well.

Type 1 PseKNC

Type 1 PseKNC, also called parallel correlation PseKNC, is used to represent a DNA sequence with a vector containing $(4^K + \lambda)$ components, as formulated below.

The sequence order effect for a DNA sequence (Eq. (1)) can be approximately reflected with a set of sequence order correlated factors defined by

$$\begin{cases} \theta_1 = \frac{1}{L-K}\sum_{i=1}^{L-K}\Theta_{i,i+1} \\ \theta_2 = \frac{1}{L-K-1}\sum_{i=1}^{L-K-1}\Theta_{i,i+2} \\ \theta_3 = \frac{1}{L-K-2}\sum_{i=1}^{L-K-2}\Theta_{i,i+3} \\ \cdots\cdots\cdots \\ \theta_\lambda = \frac{1}{L-K-\lambda+1}\sum_{i=1}^{L-K-\lambda+1}\Theta_{i,i+\lambda} \end{cases} \quad (\lambda < L-K), \tag{7}$$

where θ_1 is the first-tier correlation factor that reflects the sequence order correlation between all the most contiguous K-tuple nucleotides along the DNA sequence (Fig. 1A), θ_2 is the second-tier correlation factor that reflects the sequence order correlation between all the second most contiguous K-tuple nucleotides (Fig. 1B), θ_3 is the third-tier correlation factor that reflects the sequence order correlation between all the third most contiguous K-tuple nucleotides (Fig. 1C), and so forth. The number λ is an integer used to reflect the correlation rank (or tier) and, hence, must be smaller than L – K.

In Eq. (7), the correlation function is given by

$$\begin{cases} \Theta_{i,i+j} = \frac{1}{\Lambda}\sum_{\xi=1}^{\Lambda}\left[H_\xi(R_iR_{i+1}\cdots R_{i+K-1}) - H_\xi\left(R_{i+j}R_{i+j+1}\cdots R_{i+j+K-1}\right)\right]^2, \\ i = 1,2,\cdots,\ L-K+1; j = 1,\ 2,\ ,\cdots,\lambda; \lambda < L-K \end{cases} \tag{8}$$

where R_i and all the other symbols of its kind can be any valid nucleic acid A, C, G, or T (cf. Eq. (3)), $H_\xi(R_iR_{i+1}\cdots R_{i+K-1})$ is the numerical value of the ξ-th physicochemical property for the K-tuple nucleotide $R_iR_{i+1}\cdots R_{i+K-1}$ in a DNA sequence, $H_\xi(R_{i+j}R_{i+j+1}\cdots R_{i+j+K-1})$ is the corresponding value for the K-tuple nucleotide $R_{i+j}R_{i+j+1}\cdots R_{i+j+K-1}$, and Λ is the total number of correlation functions counted. Note that before substituting the values of physicochemical property into Eq. (8), they were all subjected to a standard conversion as described by Eq.(9) or (10) below.

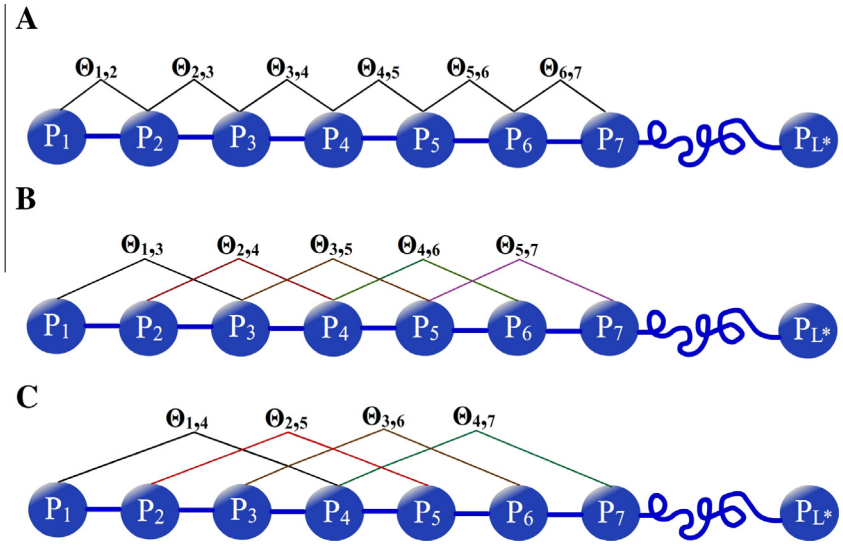

Fig.1. Schematic drawing to show the first tier (A), second tier (B), and third tier (C) sequence order correlation mode along a DNA sequence in the type 1 PseKNC. Panel A reflects the correlation mode between all the most contiguous K-tuple nucleotides, panel B reflects that between all the second most contiguous K-tuple nucleotides, and panel C reflects that between all the third most contiguous K-tuple nucleotides. P_1 represents the first K-tuple nucleotide (i.e., $R_1R_2\cdots R_K$) along the DNA sequence, P_2 represents the second K-tuple nucleotide ($R_2R_3\cdots R_{K+1}$), P_3 represents the third K-tuple nucleotide ($R_3R_4\cdots R_{K+2}$), and so forth. $L^* = L - K$ is the maximum number allowed for the K-tuple nucleotides in the L-bp-long DNA sequence.

$$H_\xi(R_1R_2\cdots R_K)=\frac{H^0_\xi(R_1R_2\cdots R_K)-\sum_\Pi H^0_\xi(R_1R_2\cdots R_K)/4^K}{\sqrt{\left[\sum_\Pi\left[H^0_\xi(R_1R_2\cdots R_K)-\sum_\Pi H^0_\xi(R_1R_2\cdots R_K)/4^K\right]^2/4^K\right]}},\tag{9}$$

where $R_1R_2\cdots R_K$ can be either $R_iR_{i+1}\cdots R_{i+K-1}$ or $R_{i+j}R_{i+j+1}\cdots R_{i+j+K-1}$ and the operator Π means counting all the different combinations of A, C, G, and T (cf. Eq. (3)) for $R_1R_2\cdots R_K$; when $K=2$ we have $4^2=16$ different combinations (cf. Eq. (4)), when $K=3$ we have $4^3=64$ different combinations (cf. Eq. (5)), and so forth. Actually, Eq. (9) can be written as [35]

$$H_\xi(R_1R_2\cdots R_K)=\frac{H^0_\xi(R_1R_2\cdots R_K)-\left\langle H^0_\xi(R_1R_2\cdots R_K)\right\rangle}{SD\left\langle H^0_\xi(R_1R_2\cdots R_K)\right\rangle},\tag{10}$$

where the symbol $\langle\cdot\rangle$ means taking the average of the quantity therein over the 4^K different combinations of A, C, G, and T (cf. Eq. (3)) for $R_1R_2\cdots R_K$ and SD is the corresponding standard deviation.

In the above equations, $H^0_\xi(R_1R_2\cdots R_K)$ $(\xi=1,2,\cdots,\Lambda)$ are the original physicochemical property values for the oligonucleotides, which can be obtained from Ref. [31] as well as the references given in Tables 1 and 2 for dinucleotides and trinucleotides, respectively. The advantage of using the converted physicochemical property values obtained via Eq. (9) or (10) is that they will have a zero mean value over the 4^K different K-tuple nucleotides and will remain unchanged if they go through the same conversion procedure again. For readers' convenience, the original values for both the 38 physicochemical properties of dinucleotides and the 12 physicochemical properties of trinucleotides are given in Online Supporting Information S1 and S2, respectively (see online Supplementary material), and their corresponding standard-converted values are given in Online Supporting Information S3 and S4, respectively.

As we can see from Fig. 1, the sequence order effect of a DNA can be, to some extent, reflected through a set of sequence correlation factors $\theta_1,\theta_2,\theta_3,\cdots,\theta_\lambda$ as defined by Eq. (7). To incorporate such information, instead of Eq. (6) we use a $(4^K+\lambda)$-D vector to represent a DNA sequence as given by

$$\mathbf{D}^{\mathrm{I}}_{\mathrm{PseKNC}}=\left[d_1\ \cdots\ d_{4^K}\ d_{4^K+1}\ \cdots\ d_{4^K+\lambda}\right]^{\mathbf{T}},\quad(\lambda<L-K),\tag{11}$$

where

$$d_u=\begin{cases}\dfrac{f_u^{K\text{-tuple}}}{\sum_{i=1}^{4^K}f_i^{K\text{-tuple}}+w\sum_{j=1}^{\lambda}\theta_j} & (1\le u\le 4^K)\\ \dfrac{w\theta_{u-4^K}}{\sum_{i=1}^{4^K}f_i^{K\text{-tuple}}+w\sum_{j=1}^{\lambda}\theta_j} & (4^K+1\le u\le 4^K+\lambda)\end{cases},\tag{12}$$

where $f_i^{K\text{-tuple}}$ is the normalized occurrence frequency of the i-th K-tuple nucleotide in the DNA sequence, θ_j is the j-tier sequence correlation factor computed according to Eqs. (7)–(10) for the DNA sequence, and w is the weight factor used to adjust the effect of the pseudo nucleotide component.

As shown in Eqs. (11) and (12), the first 4^K components reflect the effect of the K-tuple nucleotide composition, whereas the components from 4^K+1 to $4^K+\lambda$ reflect the effect of long-range sequence order. A set of such $4^K+\lambda$ components as formulated in Eqs. (11) and (12) is called the type 1 PseKNC for a DNA sequence with L nucleotides.

Type 2 PseKNC

Type 2 PseKNC, also called series correlation PseKNC, can be used to express a DNA sequence with a vector containing $(4^K+\lambda\cdot\Lambda)$ components, where Λ is the number of physicochemical properties counted (cf. Eq. (8)). In a way parallel to the approach as elaborated in Refs. [30,36] for protein sequences, the sequence order effect of a DNA with L nucleotides can be approximately reflected with a set of sequence order correlated factors as defined below:

$$\begin{cases}\tau_1=\frac{1}{L-K-1}\sum_{i=1}^{L-K-1}J^1_{i,i+1}\\ \tau_2=\frac{1}{L-K-1}\sum_{i=1}^{L-K-1}J^2_{i,i+1}\\ \cdots\cdots\cdots\cdots\\ \tau_\Lambda=\frac{1}{L-K-1}\sum_{i=1}^{L-K-1}J^\Lambda_{i,i+1}\quad\lambda<(L-K),\\ \cdots\cdots\cdots\cdots\\ \tau_{\lambda\Lambda-1}=\frac{1}{L-K-\lambda}\sum_{i=1}^{L-K-\lambda}J^{\Lambda-1}_{i,i+\lambda}\\ \tau_{\lambda\Lambda}=\frac{1}{L-K-\lambda}\sum_{i=1}^{L-K-\lambda}J^{\Lambda}_{i,i+\lambda}\end{cases}\tag{13}$$

where

$$\begin{cases}J^\xi_{i,i+m}=H_\xi(R_iR_{i+1}\cdots R_{i+K-1})\cdot H_\xi(R_{i+m}R_{i+m+1}\cdots R_{i+m+K-1})\\ \xi=1,2,\cdots,\Lambda;\quad m=1,2,\cdots,\lambda;\quad i=1,2,\cdots,L-K-\lambda\end{cases}.\tag{14}$$

In the above equation, the function H_ξ has exactly the same meaning as defined in Eqs. (8)–(10). Thus, similar to the case in dealing with proteins [30], after incorporating the $\lambda\Lambda$ sequence order correlation factors from Eq. (13) into the conventional K-tuple nucleotide composition of Eq. (6), we obtain the type 2 K-tuple pseudo nucleotide composition or type 2 PseKNC (Fig. 2), which is actually a vector with $(4^K+\lambda\Lambda)$ components as given below:

$$\mathbf{D}^{\mathrm{II}}_{\mathrm{PseKNC}}=\left[d_1\ \cdots\ d_{4^K}\ d_{4^K+1}\ \cdots\ d_{4^K+\lambda}\ d_{4^K+\lambda+1}\ \cdots\ d_{4^K+\lambda\Lambda}\right]^{\mathbf{T}},\tag{15}$$

where

$$d_u=\begin{cases}\dfrac{f_u^{K\text{-tuple}}}{\sum_{i=1}^{4^K}f_i^{K\text{-tuple}}+w\sum_{j=1}^{\lambda\Lambda}\tau_j} & (1\leqslant u\leqslant 4^K)\\ \dfrac{w\tau_{u-4^K}}{\sum_{i=1}^{4^K}f_i^{K\text{-tuple}}+w\sum_{j=1}^{\lambda\Lambda}\tau_j} & (4^K+1\leqslant u\leqslant 4^K+\lambda\Lambda)\end{cases},\tag{16}$$

where all the terms have exactly the same meanings as those in Eq. (12) except for τ_j and Λ: the former is given by Eq. (13), and the latter is the total number of the correlation functions counted (cf. Eq. (8)). As we can see from Eq. (15), in comparison with type 1 PseKNC (Eq. (11)) containing $(4^K+\lambda)$ components, type 2 PseKNC is a vector with $(4^K+\lambda\Lambda)$ components.

Web server guide

For the convenience of the vast majority of experimental scientists, below we give a step-by-step guide on how to generate their desired pseudo K-tuple nucleotide composition without the need to follow the complicated mathematic equations in the previous section ("Pseudo K-tuple nucleotide composition") that were presented just for the integrity in developing the current flexible web server.

Step 1

Open the web server at http://lin.uestc.edu.cn/pseknc/default.aspx, and you will see the top page of PseKNC on your computer screen, as shown in Fig. 3.

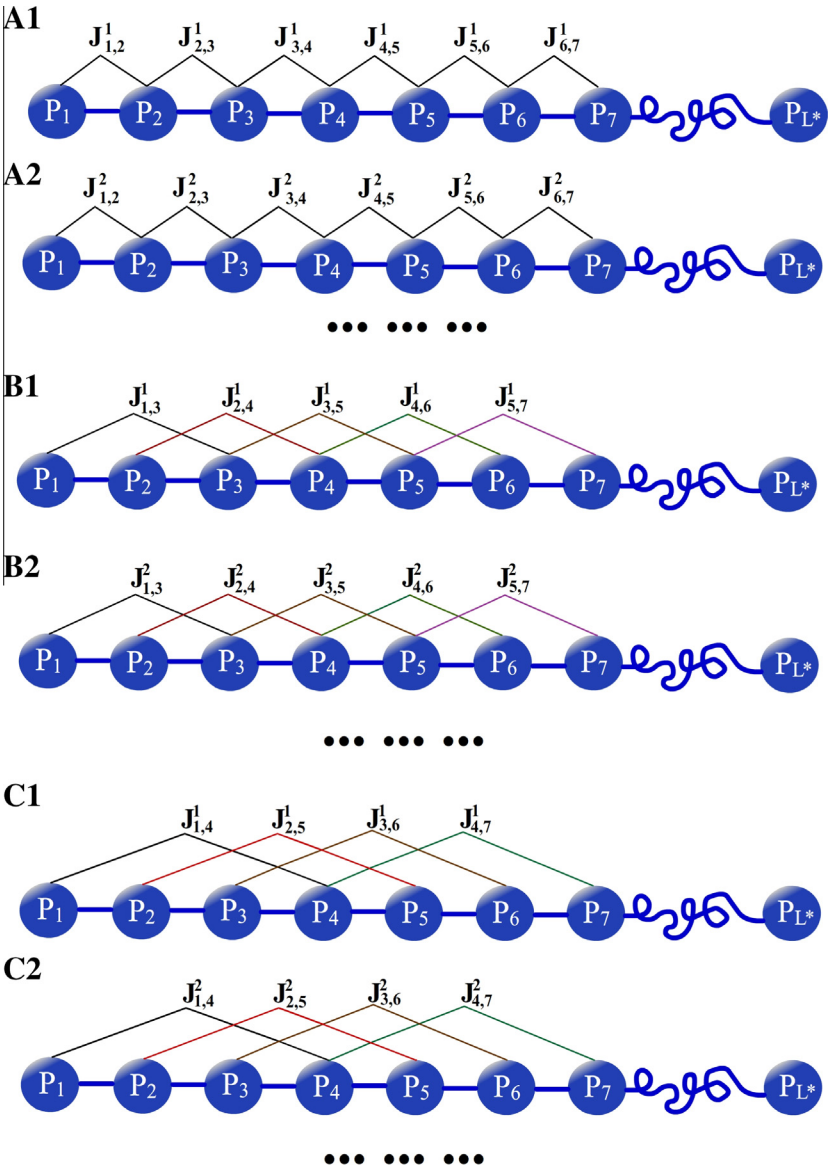

Fig.2. Schematic drawing to show the first tier (A1/A2), second tier (B1/B2), and third tier (C1/C2) sequence order correlation mode along a DNA sequence for the type 2 PseKNC based on the first/second physicochemical properties, respectively (cf. Eqs. (13) and (14)). Panel A1/A2 reflects the correlation mode between all the most contiguous K-tuple nucleotides, panel B1/B2 reflects that between all the second most contiguous K-tuple nucleotides, panel C1/C2 reflects that between all the third most contiguous K-tuple nucleotides, and so forth. The symbols P_1, P_2, P_3, and so forth have exactly the same meaning as in Fig. 1. See text for further explanation.

Step 2

Select which type of PseKNC by using the drop-down menu: type 1 or type 2.

Step 3

Select which kind of oligonucleotide: Dinucleotide **or** Trinucleotide. If selecting the former, you will see 38 physicochemical properties shown on the screen (Fig. 3); if selecting the latter, you will see 12 physicochemical properties shown on the screen (Fig. 4).

Step 4

Select the physicochemical property/properties considered by clicking its/their left button/buttons; suppose that the number of selected properties is $\Lambda = 3$.

Step 5

Select the value for the weight factor w from 0.1 to 1.0 by using the drop-down menu. Its optimal value is determined via an optimization procedure as described in our previous studies [25,26,28]. For a detailed and rigorous description of how to determine the optimal value of the weight factor w with other parameters together, see Ref. [35].

Step 6

Enter the desired parameter for λ into the box to the right of "λ parameter." The parameter reflects the sequence correlation ranks counted. Its value must be a non-negative integer and smaller than $(L-2)$ or $(L-3)$ for the case of dinucleotide or trinucleotide, respectively, where L is the length of the DNA sequence concerned. When $\lambda = 0$, the outcome will be reduced to the conventional dinucleotide or trinucleotide composition, respectively.

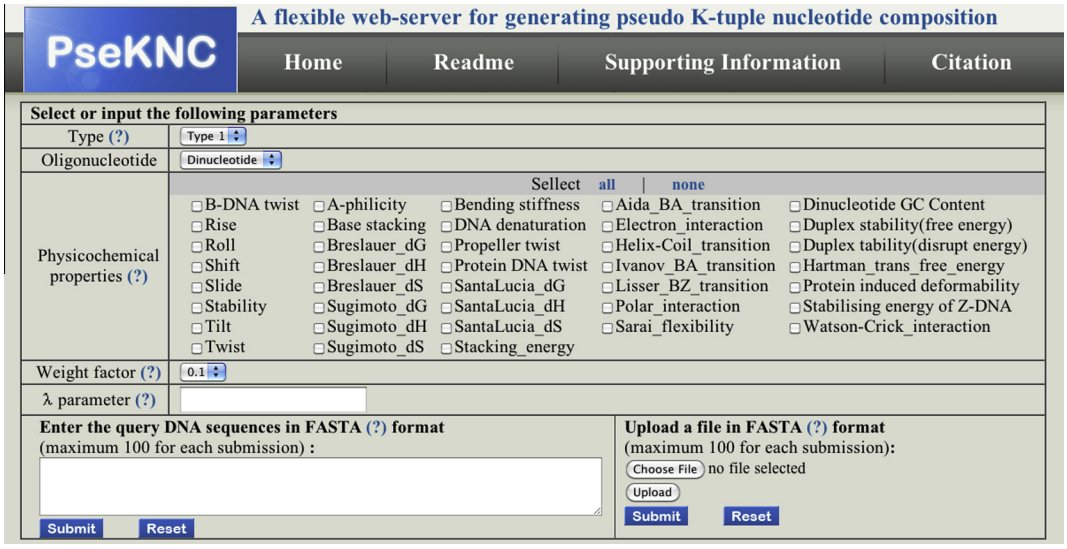

Fig.3. Partial screenshot showing the top page of the PseKNC. Its website address is http://lin.uestc.edu.cn/pseknc/default.aspx. In the top page, users can select type 1 or type 2 PseKNC by using the dropdown menu in the Type module. Once the kind of oligonucleotide is selected by using the drop-down menu in the Oligonucleotide module, the related physicochemical properties will be shown on the right side of the Physicochemical Properties module and users can select the physicochemical properties considered by clicking their left buttons. The two parameters weight factor w and λ can be assigned in the Weight Factor and λ Parameter modules, respectively. The query DNA sequences can be directly entered into the input box or uploaded via the Browse button with a FASTA format file. Users can also get help by clicking on the **?** button in each of the aforementioned modules.

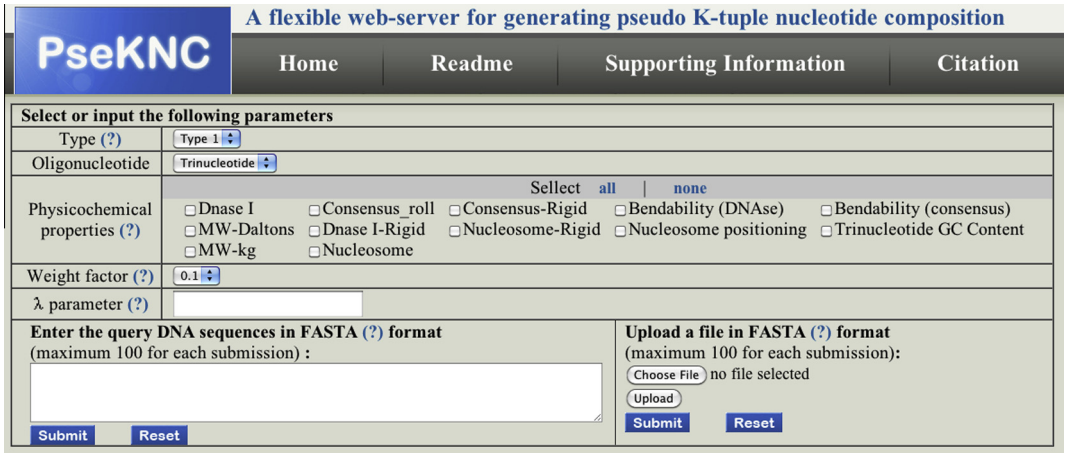

Fig.4. Partial screenshot showing the web page of PseKNC after selecting Trinucleotide via the drop-down menu. See the legend of Fig. 3 for further explanation.

Step 7

Either directly enter your query DNA sequences into the input box or upload them via the "Browse" button with a data file. The input sequences should be in the FASTA format. The maximum number of DNA sequences allowed for each submission is 100.

Step 8

Click on the Submit button to see the results. (i) If selecting type 1 and Dinucleotide and entering 10 for λ in step 6, you will get $(16 + 10) = 26$ discrete numbers for each of the DNA sequences inputted, where the 10 pseudo dinucleotide components are colored in red. (ii) If replacing type 1 with type 2 and keeping all the others the same, you will get $(16 + 30) = 46$ numbers, of which $\lambda\Lambda = 10 \times 3 = 30$ are the pseudo dinucleotide components. (iii) If replacing Dinucleotide with Trinucleotide in the above two cases, you will get $(64 + 10) = 74$ and $(64 + 30) = 94$ for the type 1 and type 2 cases, respectively. (iv) Generally speaking, you will get $(4^K + \lambda)$ or $(4^K + \lambda\Lambda)$ numbers for type 1 or type 2 PseKNC, respectively, where K = 2 or K = 3 for the case of dinucleotide or trinucleotide, respectively, whereas Λ and λ are the number of physicochemical properties and that of sequence correlated ranks selected by you. It takes a few seconds before the result appears on your computer screen.

Step 9

Click on Supporting Information to download the 38 and 12 physicochemical properties for dinucleotides and trinucleotides, respectively, covered by the current PseKNC generator.

Step 10

Click on Citation to see the papers that document the PseKNC generator or closely associate with its development.

Caveat

Only the valid characters (A, C, G, and T) for DNA sequences are allowed for your input; otherwise, a warning message will be prompted on your screen.

Discussion

Genome is a very complicated system, and hence it will need many different PseKNC modes to deal with various different problems in genome analysis. In this regard, the current web server provides a very flexible tool with extremely high capacity.

This can be seen from how many different modes of pseudo K-tuple nucleotide compositions the current web server can generate. As shown in Tables 1 and 2, there are 38 physicochemical properties for dinucleotides and 12 physicochemical properties for trinucleotides that users can select to generate different modes of PseKNC. Accordingly, the total possible different modes for PseDNC would be

$$\begin{aligned} &C(38,1)+C(38,2)+C(38,3)+\cdots+C(38,38)\\ &=\frac{38!}{(38-1)!1!}+\frac{38!}{(38-2)!2!}+\frac{38!}{(38-3)!3!}+\cdots+\frac{38!}{(38-38)!38!}\\ &=274,877,906,943>2.74\times10^{11} \end{aligned} \tag{17}$$

and that for PseTNC would be

$$\begin{aligned} &C(12,1)+C(12,2)+C(12,3)+\cdots+C(12,12)\\ &=\frac{12!}{(12-1)!1!}+\frac{12!}{(12-2)!2!}+\frac{12!}{(12-3)!3!}+\cdots+\frac{12!}{(12-12)!12!}\\ &=4,095 \end{aligned} \tag{18}$$

The above figures, plus various selections for the weight factor w, the parameter λ, and the types of PseKNC, fully indicate that the current PseKNC web server is very flexible and allows users to have many choices to generate their desired pseudo oligonucleotide compositions in dealing with various different DNA or RNA systems.

Although the current PseKNC generator can generate only pseudo dinucleotide composition (K = 2) and pseudo trinucleotide composition (K = 3), with more experimental physicochemical property data available for K-tuple nucleotides (K = 4, 5, or higher) in the future, the current formulation can be easily extended to generate pseudo tetranucleotide composition, pseudo pentanucleotide composition, or other pseudo oligonucleotide compositions with higher numbers of K as well.

Acknowledgments

This work was supported by the National Nature Scientific Foundation of China (61202256 and 61100092) and the Nature Scientific Foundation of Hebei Province (C2013209105).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ab.2014.04.001.

References

[1] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J.M. Claverie, O. Gascuel, Phylogeny.fr: robust phylogenetic analysis for the non-specialist, Nucleic Acids Res. 36 (2008) W465–W469.
[2] T.L. Bailey, M. Boden, F.A. Buske, M. Frith, C.E. Grant, L. Clementi, J. Ren, W.W. Li, W.S. Noble, MEME SUITE: tools for motif discovery and searching, Nucleic Acids Res. 37 (2009) W202–W208.
[3] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255 (erratum: 44 (2001) 60).
[4] S.X. Lin, J. Lapointe, Theoretical and experimental biology in one, J. Biomed. Sci. Eng. 6 (2013) 435–442.
[5] X.B. Zhou, C. Chen, Z.C. Li, X.Y. Zou, Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, J. Theor. Biol. 248 (2007) 546–551.
[6] L. Nanni, A. Lumini, Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization, Amino Acids 34 (2008) 653–660.
[7] B. Liu, X. Wang, Q. Zou, Q. Dong, Q. Chen, Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation, Mol. Inf. 32 (2013) 775–782.
[8] J.D. Qiu, J.H. Huang, R.P. Liang, X.Q. Lu, Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform, Anal. Biochem. 390 (2009) 68–73.
[9] S.S. Sahu, G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction, Comput. Biol. Chem. 34 (2010) 320–327.
[10] M. Esmaeili, H. Mohabatkar, S. Mohsenzadeh, Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses, J. Theor. Biol. 263 (2010) 203–209.
[11] H. Mohabatkar, M. Mohammad Beigi, A. Esmaeili, Prediction of $GABA_A$ receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine, J. Theor. Biol. 281 (2011) 18–23.
[12] D. Zou, Z. He, J. He, Y. Xia, Supersecondary structure prediction using Chou's pseudo amino acid composition, J. Comput. Chem. 32 (2011) 271–278.
[13] L. Nanni, A. Lumini, D. Gupta, A. Garg, Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information, IEEE/ACM Trans. Comput. Biol. Bioinf. 9 (2012) 467–475.
[14] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, H. Mohabatkar, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, J. Theor. Biol. C 341 (2013) 34–40.
[15] H.L. Xie, L. Fu, X.D. Nie, Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC, Protein Eng. Des. Sel. 26 (2013) 735–742.
[16] Y. Xu, J. Ding, L.Y. Wu, K.C. Chou, ISNO–PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, PLoS ONE 8 (2013) e55844.
[17] Y. Xu, X.J. Shao, L.Y. Wu, N.Y. Deng, K.C. Chou, ISNO–AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, Peer J. 1 (2013) e171. https://peerj.com/articles/171.pdf.
[18] M. Hayat, A. Khan, Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC, Protein Pept. Lett. 19 (2012) 411–421.
[19] H. Mohabatkar, M.M. Beigi, K. Abdolahi, S. Mohsenzadeh, Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach, Med. Chem. 9 (2013) 133–137.
[20] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review), J. Theor. Biol. 273 (2011) 236–247.
[21] P. Du, X. Wang, C. Xu, Y. Gao, PseAAC–Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions, Anal. Biochem. 425 (2012) 117–119.
[22] D.S. Cao, Q.S. Xu, Y.Z. Liang, Propy: a tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (2013) 960–962.
[23] P. Du, S. Gu, Y. Jiao, 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 (2014) 3495–3506.
[24] H.B. Shen, K.C. Chou, PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition, Anal. Biochem. 373 (2008) 386–388.
[25] W. Chen, P.M. Feng, H. Lin, K.C. Chou, IRSpot–PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (2013) e69.
[26] W.R. Qiu, X. Xiao, K.C. Chou, IRSpot–TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. 15 (2014) 1746–1766.
[27] X. Zhou, Z. Li, Z. Dai, X. Zou, Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform, J. Theor. Biol. 319 (2013) 1–7.
[28] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, W. Chen, K.C. Chou, iNuc–PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo K-tuple nucleotide composition, Bioinformatics, in press. http://dx.doi.org/10.1093/bioinformatics/btu083.
[29] I. Ioshikhes, A. Bolshoy, K. Derenshteyn, M. Borodovsky, E.N. Trifonov, Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences, J. Mol. Biol. 262 (1996) 129–139.
[30] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[31] I. Brukner, R. Sanchez, D. Suck, S. Pongor, Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides, EMBO J. 14 (1995) 1812–1818.

[32] Y. Fukue, N. Sumida, J. Tanase, T. Ohyama, A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance, Nucleic Acids Res. 33 (2005) 3821–3827.
[33] D.M. Gowers, S.E. Halford, Protein motion from non-specific to specific DNA by three-dimensional routes aided by supercoiling, EMBO J. 22 (2003) 1410–1418.
[34] S.C. Parker, L. Hansen, H.O. Abaan, T.D. Tullius, E.H. Margulies, Local DNA topography correlates with functional noncoding regions of the human genome, Science 324 (2009) 389–392.
[35] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.
[36] K.C. Chou, Y.D. Cai, Prediction of membrane protein types by incorporating amphipathic effects, J. Chem. Inf. Model. 45 (2005) 407–413.
[37] R.L. Ornstein, R. Rein, D.L. Breen, R.D. Macelroy, An optimized potential function for the calculation of nucleic acid interaction energies, Biopolymers 17 (1978) 2341–2360.
[38] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, V.B. Zhurkin, DNA sequence-dependent deformability deduced from protein–DNA crystal complexes, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 11163–11168.
[39] A.A. Gorin, V.B. Zhurkin, W.K. Olson, B-DNA twisting correlates with base-pair morphology, J. Mol. Biol. 247 (1995) 34–48.
[40] K. Vlahovicek, L. Kajan, S. Pongor, DNA analysis servers: plot.it, bend.it, model.it, and IS, Nucleic Acids Res. 31 (2003) 3686–3687.
[41] V.I. Ivanov, L.E. Minchenkova, B.K. Chernov, P. McPhie, S. Ryu, S. Garges, A.M. Barber, V.B. Zhurkin, S. Adhya, CRP–DNA complexes: inducing the A-like form in the binding sites with an extended central spacer, J. Mol. Biol. 245 (1995) 228–240.
[42] M.A. el Hassan, C.R. Calladine, Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA, J. Mol. Biol. 259 (1996) 95–103.
[43] N. Sugimoto, S. Nakano, M. Yoneyama, K. Honda, Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes, Nucleic Acids Res. 24 (1996) 4501–4505.
[44] K.J. Breslauer, R. Frank, H. Blocker, L.A. Marky, Predicting DNA duplex stability from the base sequence, Proc. Natl. Acad. Sci. U.S.A. 83 (1986) 3746–3750.
[45] R.D. Blake, Encyclopedia of Molecular Biology and Molecular Medicine, VCH, New York, 1996.
[46] A.V. Sivolob, S.N. Khrapunov, Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness, J. Mol. Biol. 247 (1995) 918–931.
[47] P.S. Ho, M.J. Ellison, G.J. Quigley, A. Rich, A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences, EMBO J. 5 (1986) 2737–2744.
[48] M. Aida, An ab initio molecular orbital study on the sequence-dependency of DNA conformation: an evaluation of intra- and inter-strand stacking interaction energy, J. Theor. Biol. 130 (1988) 327–335.
[49] B. Hartmann, B. Malfoy, R. Lavery, Theoretical prediction of base sequence effects in DNA: experimental reactivity of Z-DNA and B–Z transition enthalpies, J. Mol. Biol. 207 (1989) 433–444.
[50] T.V. Chalikian, J. Volker, G.E. Plum, K.J. Breslauer, A more unified picture for the thermodynamics of nucleic acid duplex melting: a characterization by calorimetric and volumetric techniques, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 7853–7858.
[51] S. Lisser, H. Margalit, Determination of common structural features in *Escherichia coli* promoters by computer analysis, Eur. J. Biochem. 223 (1994) 823–830.
[52] M.M. Gromiha, P.K. Ponnuswamy, Hydrophobic distribution and spatial arrangement of amino acid residues in membrane proteins, Int. J. Pept. Protein Res. 48 (1996) 452–460.
[53] J. SantaLucia Jr., H.T. Allawi, P.A. Seneviratne, Improved nearest-neighbor parameters for predicting DNA duplex stability, Biochemistry 35 (1996) 3555–3562.
[54] A. Sarai, J. Mazur, R. Nussinov, R.L. Jernigan, Sequence dependence of DNA conformational flexibility, Biochemistry 28 (1989) 7842–7849.
[55] O. Gotoh, Y. Tagashira, Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles, Biopolymers 20 (1981) 1033–1042.
[56] J.P. Lewis, O.F. Sankey, Geometry and energetics of DNA basepairs and triplets from first principles quantum molecular relaxations, Biophys. J. 69 (1995) 1068–1076.
[57] J.R. Goni, A. Perez, D. Torrents, M. Orozco, Determining promoter location based on DNA structure first-principles calculations, Genome Biol. 8 (2007) R263.
[58] M.G. Munteanu, K. Vlahovicek, S. Parthasarathy, I. Simon, S. Pongor, Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena, Trends Biochem. Sci. 23 (1998) 341–347.
[59] D.S. Goodsell, R.E. Dickerson, Bending and curvature calculations in B-DNA, Nucleic Acids Res. 22 (1994) 5497–5503.
[60] S.C. Satchwell, H.R. Drew, A.A. Travers, Sequence periodicities in chicken nucleosome core DNA, J. Mol. Biol. 191 (1986) 659–675.