PseKNC-General User Manual

Contents

1. Introduction of PseKNC-General

PseKNC-General (the general form of pseudo k-tuple nucleotide composition) is a cross-platform stand-alone and open-source package that can be used to represent a DNA or RNA sequence with a discrete model or vector yet still keeping considerable sequence-order information, particularly the global or long-range sequence-order information via the physicochemical properties of its constituent oligonucleotides.

2. Installation

The PseKNC-General package can be run on Linux, Mac, and Windows systems. It not only provides a command line environment, but also a user-friendly graphical user interface (GUI).

Download the package from http://lin.uestc.edu.cn/server/pseknc and extract it to a directory, for example, "~/usr".

To execute the PseKNC-General in command line environment, navigate to the "~/usr/pseknc/" directory and you will find two python files, namely "pseknc.py" and "autocorrelations.py". The "pseknc.py" is used for calculating the K-tuple nucleotide compositions and pseudo k-tuple nucleotide compositions. The "autocorrelations.py" is used for calculating Moreau-Broto autocorrelation coefficient, Moran autocorrelation coefficient and Geary autocorrelation coefficient.

To execute the PseKNC-General in GUI, navigate to the "~/usr/gui/" directory, and then double click the "PseKNC.jar" to execute the program.

3. Input/Output formats

3.1. Input format

The input file should be a valid FASTA format that consists of a single initial line beginning with a greater-than symbol (">") in the first column, followed by lines of sequence data. The words right after the ">" symbol in the single initial line are optional and only used for the purpose of identification and description required.

3.2. Output format

The output file formats support three choices that are suitable for downstream computational analyses such as machine learning. The first and the default choice is the csv format. In this format, all data are separated by commas. The second one is the libsvm's sparse data format. For this format, each line contains an instance and is ended by a '\n' character, like <index1>:<value1> <index2>:<value2> ... . The pair <index>:<value> gives a feature (attribute) value: <index> is an integer starting from 1 and <value> is a real number. The third output format is the tab format. This format is similar to the csv format. The only difference is the separation characters between data are tabs.

3.3. Physicochemical Properties Selection

The Physicochemical Properties Selection file is a text file that contains a list of properties used for pseudo k-tuple nucleotide composition and autocorrelation calculations. For example, if you are interested in "Tilt" and "Shift" of DNA dinucleotides, the Physicochemical Properties Selection file should be written as follows,

Tilt Shift

After saving this file as "propNames.txt" and specifying it using the command "-x

propNames.txt", the above two properties will be used in calculations.

A complete list of Physicochemical Properties for DNA and RNA are provided as an **Appendix** of this guide.

4. Commands

4.1 Command line parameters for PseKNC.py

Options	Interpretations
-h ,-?,-help	Display a help screen
-i <input file name> REQUIRED	The input file should be a valid FASTA format that consists of a single initial line beginning with a greater-than symbol (">") in the first column, followed by lines of sequence data. The words right after the ">" symbol in the single initial line are optional and only used for the purpose of identification and description.
-o <output file name> REQUIRED	The results could be found in the output file. The program will overwrite the file if it has already existed.
-x <property file name> REQUIRED	Property file is a text file that contains the names of selected physicochemical properties. Every line of this file contains the name of a physicochemical property.
-t <1 or 2>	Type of PseKNC. 1 - Type 1 PseKNC (default) 2 - Type 2 PseKNC
-k <2 or 3>	Kind of oligonucleotide in PseKNC 2 – Dinucleotide (default) 3 - Trinucleotide
-p	List the text files including the physicochemical properties, for which data is available for use in this program. Nothing needs to be entered after '-p'.
-j <lambda parameter>	The lambda parameter in the PseKNC algorithm can be any integer that smaller than the length of query DNA sequence. (default = 1)
-w <weight factor>	The weight parameter in the PseKNC algorithm can be a value between (0, 1]. (default = 1)
-s	If this argument is entered, it will calculate the k-tuple nucleotide composition of the query sequence. Therefore, the -j, -w and -x arguments are not required any more. But k must be specified. At this case, k can be 1, 2, …, or 6 and its default value is 2.
-f <tab or svm or csv>	The output format can be selected from followings: (1) csv: This format can be loaded into a spreadsheet program. This is designated as the default format. (2) svm: The libSVM training data format. (3) tab: Simple format delimited by TAB. If this option is omitted, the csv format will be default.

4.2 Command line parameters for Autocorrelation.py

Options	Interpretations
-h ,-?,-help	Display a help screen
-i <input file name> REQUIRED	The input file should be a valid FASTA format that consists of a single initial line beginning with a greater-than symbol (“>”) in the first column, followed by lines of sequence data. The words right after the “>” symbol in the single initial line are optional and only used for the purpose of identification and description. required
-o <output file name> REQUIRED	The results could be found in the output file. The program will overwrite the file if it has already existed.
-x <property file name> REQUIRED	Property file is a text file that contains the names of selected physicochemical properties. Every line of this file contains the name of one physicochemical property.
-a <Geary, Moran, and/or Moreau> REQUIRED	Type of autocorrelation, one or more may be entered, separated by commas (ex. -a geary, moran, moreau). Moreau (or moreau) - Normalized Moreau–Broto autocorrelation Moran (or moran) - Moran autocorrelation Geary (or geary) - Geary autocorrelation
-k <2 or 3>	Kind of oligonucleotide 2- Dinucleotide (default) 3- Trinucleotide
-p	List the text files including the physicochemical properties, for which data is available for use in this program. Nothing needs to be entered after '-p'.
-j <lambda parameter>	The lambda parameter in the PseKNC algorithm can be any integer that smaller than the length of query DNA sequence. (default = 1)

4.3 Examples

For user’s convenience, some examples of how to process a query sequence using command line are given below.

Example 1: List the text files including the physicochemical properties.

```
pseknc.py -p
```

Example 2: Calculate the dinucleotide composition of the query sequence and output the result in Libsvm format.

```
pseknc.py -i test.txt -s -k 2 -f svm -o output.txt
```

After running the above command, the following result will be found in “output.txt” file.

```
>Example1
AGTCAGTTATGACATGACACACACAACATAGTCAGATCGACGA
1:0.023 2:0.095 3:0.095 4:0.166 5:0.047 6:0.023 7:0.047 8:0.071 9:0.119 10:0.071 11:0.0
12:0.0 13:0.19 14:0.0 15:0.047 16:0.0
```

Example 3: Calculate the Type 1 pseudo dinucleotide composition of the query sequence and output the result in Libsvm format.

```
pseknc.py -i test.txt -x propNames.txt -k 2 -j 3 -w 0.5 -f svm -o out.txt
```

After running the above command, the following result will be found in "output.txt" file.

```
>Example1
1:0.0 2:0.03 3:0.0 4:0.018 5:0.024 6:0.006 7:0.042 8:0.024 9:0.012 10:0.048 11:0.0 12:0.0
13:0.012 14:0.012 15:0.018 16:0.006 17:0.277 18:0.196 19:0.271
```

Example 4: Calculate the Geary autocorrelation of the query sequence and output the result in Libsvm format.

```
autocorrelations.py -a Geary -i test.txt -o out.txt -x propNames.txt -k 2 -j 3
```

After running the above command, the following result will be found in "output.txt" file.

```
>Example1
AGTCAGTTATGACATGACACACACAACATAGTCAGATCGACGA
---------------------------------
Properties: ['Tilt', 'Shift']
GEARY: [1.268, 1.481]
```

5. GUI usages

For the convenience of users who are not familiar with command line options, the GUI shell is provided. The main interface of the GUI based shell program is shown in **Figure 1**. It includes three modules, namely "PseKNC", "Autocorrelation" and "K-tuple Composition" that can be used to calculate the pseudo k-tuple nucleotide compositions, autocorrelation coefficient and k-tuple nucleotide compositions, respectively.

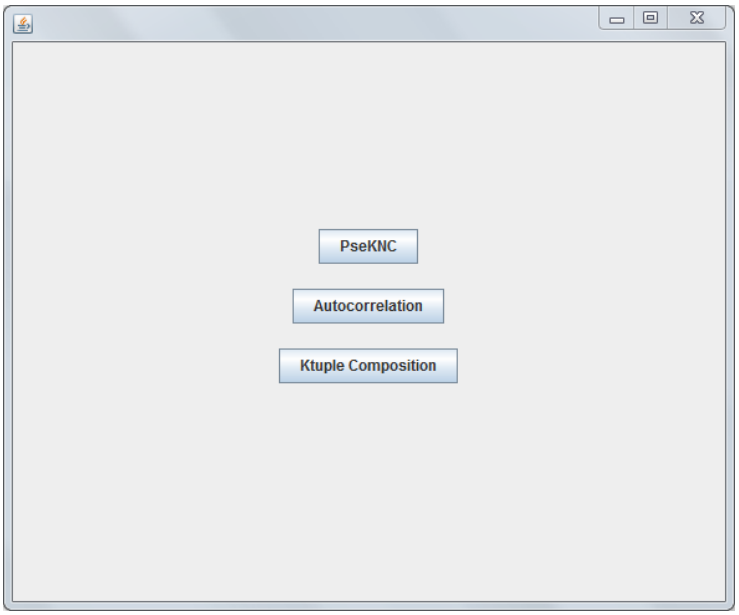

Figure 1. The main interface of GUI shell

Click the "PseKNC" button, the interface for calculating pseudo k-tuple nucleotide compositions will appear as shown in **Figure 2**. Upload the fasta file containing the query sequences (**see section 3.1 for more detail**) and the

physicochemical properties selection file (**see section 3.3 for more detail**). Input the values for the two parameters lambda and weight. Choose the type of pseudo k-tuple nucleotide compositions, Type 1 or Type 2. Click "Get sequence" and "Get Properties", then the names of the query sequences and available properties will be listed on the interface. When all the parameters are set, click the "Calculate" button to generate the desired pseudo k-tuple nucleotide compositions. The results not only can be shown on the screen, but also can be saved into a file with three optional formats, namely "tab", "svm" or "csv" formats.

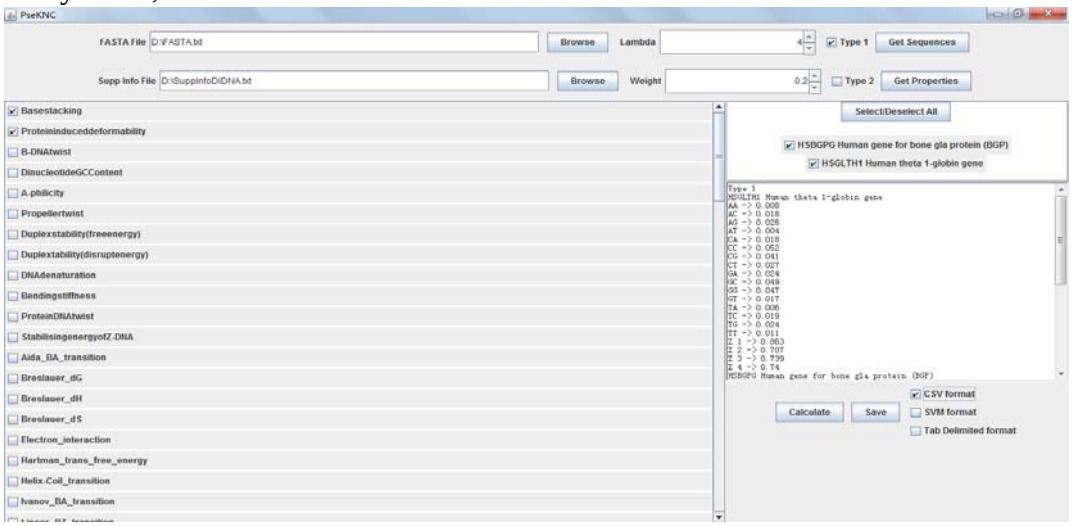

Figure 2. The interface of PseKNC module in the GUI shell

Click the "Autocorrelation" button, the interface for calculating pseudo k-tuple nucleotide compositions will appear as shown in **Figure 3**. Upload the fasta file containing the query sequences (**see section 3.1 for more detail**) and the physicochemical properties selection file (**see section 3.3 for more detail**). Input the value of the interspaces parameter into the box to the right of "Lambda". Click "Get sequence" and "Get Properties", then the names of the query sequences and available properties will be listed on the interface. When all the parameters are set, click the "Moran Autocorrelation", "Geray Autocorrelation", or "Moreau Autocorrelation" button to generate the corresponding types of autocorrelation coefficients. The results not only can be shown on the screen, but also can be saved into a file with three optional formats, namely "tab", "svm" or "csv" formats.

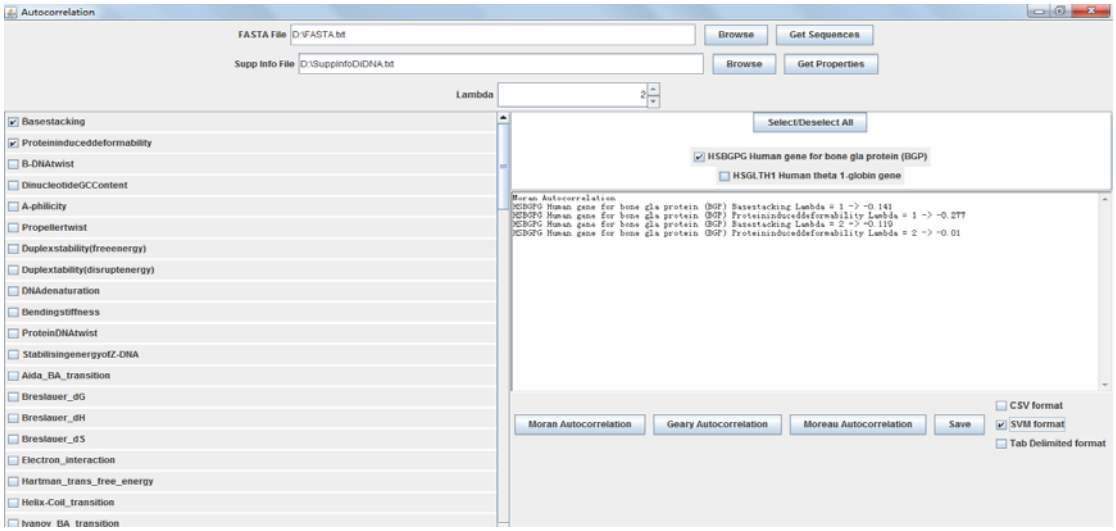

Figure 3. The interface of Autocorrelation module in the GUI shell

Click the "Ktuple Composition" button, the interface for calculating k-tuple nucleotide compositions will appear as shown in **Figure 4**. Upload the fasta file containing the query sequences (**see section 3.1 for more detail**). Input the characters considered, "A", "C", "G" and "T" are accepted for DNA sequences, while "A", "C", "G" and "U" are accepted for RNA sequences. Enter the value of "k" into the box to the right of "size of nucleotide". Click "Get sequence", then the names of the query sequences will be listed on the interface. When all the parameters are well set, click the "Calculate" button to generate the corresponding k-tuple nucleotide compositions. The results not only can be shown on the screen, but also can be saved into a file with three optional formats, namely "tab", "svm" or "csv" formats.

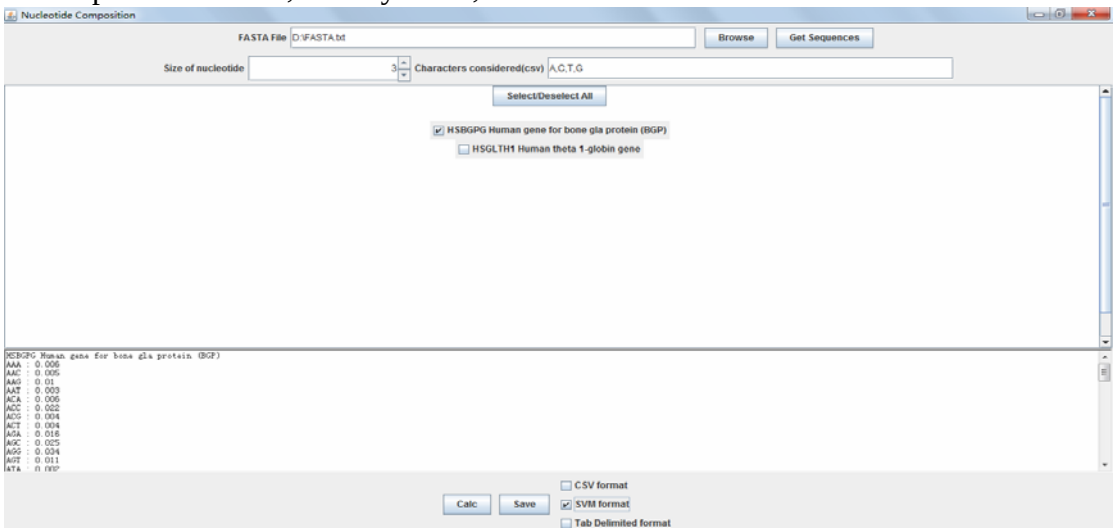

Figure 4. The interface of Ktuple Composition module in the GUI shell

6. Appendix

Table A1. The physicochemical properties for DNA dinucleotides.

Properties	Reference	Properties	Reference	Properties	Reference
Base stacking	[1]	Sugimoto_dS	[18]	Roll_roll	[29]
Protein induced deformability	[2]	Watson-Crick_interaction	[21]	Twist_tilt	[29]
B-DNA twist	[3]	Twist	[22]	Twist_roll	[29]
Dinucleotide GC Content	[4]	Tilt	[22]	Tilt_roll	[29]
A-philicity	[5]	Roll	[22]	Shift_shift	[29]
Propeller twist	[6]	Shift	[22]	Slide_slide	[29]
Duplex stability(free energy)	[7]	Slide	[22]	Rise_rise	[29]
Duplex tability(disrupt energy)	[8]	Rise	[22]	Shift_slide	[29]
DNA denaturation	[9]	Stacking energy	[23]	Shift_rise	[29]
Bending stiffness	[10]	Bend	[24]	Slide_rise	[29]
Protein DNA twist	[2]	Tip	[24]	Twist_shift	[29]
Stabilising energy of Z-DNA	[11]	Inclination	[24]	Twist_slide	[29]
Aida_BA_transition	[12]	Major Groove Width	[24]	Twist_rise	[29]
Breslauer_dG	[8]	Major Groove Depth	[24]	Tilt_shift	[29]
Breslauer_dH	[8]	Major Groove Size	[3]	Tilt_slide	[29]
Breslauer_dS	[8]	Major Groove Distance	[3]	Tilt_rise	[29]
Electron_interaction	[4]	Minor Groove Width	[24]	Roll_shift	[29]
Hartman_trans_free_energy	[13]	Minor Groove Depth	[24]	Roll_slide	[29]
Helix-Coil_transition	[14]	Minor Groove Size	[3]	Roll_rise	[29]
Ivanov_BA_transition	[15]	Minor Groove Distance	[3]	Slide stiffness	[22]
Lisser_BZ_transition	[16]	Persistance Length	[25]	Shift stiffness	[22]
Polar_interaction	[17]	Melting Temperature	[26]	Roll stiffness	[22]
SantaLucia_dG	[18]	Mobility to bend towards major groove	[27]	Rise stiffness	[22]
SantaLucia_dH	[18]	Mobility to bend towards minor groove	[27]	Tilt stiffness	[22]
SantaLucia_dS	[18]	Propeller Twist	[3]	Twist stiffness	[22]
Sarai_flexibility	[19]	Clash Strength	[3]	Wedge	[30]
Stability	[20]	Enthalpy	[7]	Direction	[30]
Stacking_energy	[1]	Free energy	[28]	Flexibility_slide	[31]
Sugimoto_dG	[18]	Twist_twist	[29]	Flexibility_shift	[31]
Sugimoto_dH	[18]	Tilt_tilt	[29]	Entropy	[32]
Sugimoto_dS	[18]	Roll_roll	[29]		

Table A2. The physicochemical properties for RNA dinucleotides.

Properties	Reference	Properties	Reference	Properties	Reference
Shift	[33]	Tilt	[33]	Enthalpy	[22]
Hydrophilicity	[34]	Roll	[33]	Entropy	[35]
Slide	[33]	Twist	[33]	Free energy	[35]
Rise	[33]	Stacking energy	[33]		

Table A3. The physicochemical properties for DNA tinucleotides.

Properties	Reference	Properties	Reference	Properties	Reference
Bendability (DNAse)	[36]	Consensus_roll	[37, 39]	MW-Daltons	[37]
Bendability (consensus)	[36]	Consensus_Rigid	[37, 39]	MW-kg	[37]
Trinucleotide GC Content	[37]	Dnase I	[40]	Nucleosome	[41]
Nucleosome positioning	[38]	Dnase I-Rigid	[40]	Nucleosome-Rigid	[41]

7. Reference

[1] Ornstein RL, Rein R, Breen DL, Macelroy RD: An optimized potential function for the calculation of nucleic acid interaction energies. Biopolymers 1978, 17:2341-2360.

[2] Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB: DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A 1998, 95:11163-11168.

[3] Gorin AA, Zhurkin VB, Olson WK: B-DNA twisting correlates with base-pair morphology. J Mol Biol 1995, 247:34-48.

[4] Vlahovicek K, Kajan L, Pongor S: DNA analysis servers: plot.it, bend.it, model.it and IS. Nucleic Acids Res 2003, 31:3686-3687.

[5] Ivanov VI, Minchenkova LE, Chernov BK, McPhie P, Ryu S, Garges S, Barber AM, Zhurkin VB, Adhya S: CRP-DNA complexes: inducing the A-like form in the binding sites with an extended central spacer. J Mol Biol 1995, 245:228-240.

[6] el Hassan MA, Calladine CR: Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. J Mol Biol 1996, 259:95-103.

[7] Sugimoto N, Nakano S, Yoneyama M, Honda K: Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. Nucleic Acids Res 1996, 24:4501-4505.

[8] Breslauer KJ, Frank R, Blocker H, Marky LA: Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci U S A 1986, 83:3746-3750.

[9] Blake RD: Encyclopedia of Molecular Biology and Molecular Medicine. New York: VCH Publishers; 1996.

[10] Sivolob AV, Khrapunov SN: Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness. J Mol Biol 1995, 247:918-931.

[11] Ho PS, Ellison MJ, Quigley GJ, Rich A: A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. EMBO J 1986, 5:2737-2744.

[12] Aida M: 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 1988, 130:327-335.
[13] Hartmann B, Malfoy B, Lavery R: Theoretical prediction of base sequence effects in DNA. Experimental reactivity of Z-DNA and B-Z transition enthalpies. J Mol Biol 1989, 207:433-444.
[14] Chalikian TV, Volker J, Plum GE, Breslauer KJ: 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 1999, 96:7853-7858.
[15] Ivanov VI, Krilov DY, Shchyolkina AK, Chernov BK, Minchenkov LE: Decimal code controlling the B to A transition of DNA. Journal of Biomolecular Structure and Dynamics 1995, 12:102-108.
[16] Lisser S, Margalit H: Determination of common structural features in Escherichia coli promoters by computer analysis. Eur J Biochem 1994, 223:823-830.
[17] Gromiha MM, Ponnuswamy PK: Hydrophobic distribution and spatial arrangement of amino acid residues in membrane proteins. Int J Pept Protein Res 1996, 48:452-460.
[18] SantaLucia J, Jr., Allawi HT, Seneviratne PA: Improved nearest-neighbor parameters for predicting DNA duplex stability. Biochemistry 1996, 35:3555-3562.
[19] Sarai A, Mazur J, Nussinov R, Jernigan RL: Sequence dependence of DNA conformational flexibility. Biochemistry 1989, 28:7842-7849.
[20] Gotoh O, Tagashira Y: Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles. Biopolymers 1981, 20:1033-1042.
[21] Lewis JP, Sankey OF: Geometry and energetics of DNA basepairs and triplets from first principles quantum molecular relaxations. Biophys J 1995, 69:1068-1076.
[22] Goni JR, Perez A, Torrents D, Orozco M: Determining promoter location based on DNA structure first-principles calculations. Genome Biol 2007, 8:R263.
[23] Sponer J, Gabb HA, Leszczynski J, Hobza P: Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study. Biophys J 1997, 73:76-87.
[24] Karas H, Knuppel R, Schulz W, Sklenar H, Wingender E: Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. Comput Appl Biosci 1996, 12:441-446.
[25] Hogan ME, Austin RH: Importance of DNA stiffness in protein-DNA binding specificity. Nature 1987, 329:263-266.
[26] Gotoh O, Tagashira Y: Stabilities Of Nearest-Neighbor Doublets In Double-Helical DNA Determined by Fitting Calculated Melting Profiles To Observed Profiles. Biopolymers 1981, 20:1033-1042.
[27] Gartenberg MR, Crothers DM: DNA sequence determinants of CAP-induced bending and protein binding affinity. Nature 1988, 333:824-829.
[28] Delcourt SG, Blake RD: Stacking energies in DNA. J Biol Chem 1991, 266:15160-15169.
[29] Lankas F, Sponer J, Langowski J, Cheatham TE, 3rd: DNA basepair step deformability inferred from molecular dynamics simulations. Biophys J 2003, 85:2872-2883.
[30] Shpigelman ES, Trifonov EN, Bolshoy A: CURVATURE: software for the analysis of curved DNA. Comput Appl Biosci 1993, 9:435-440.
[31] Packer MJ, Dauncey MP, Hunter CA: Sequence-dependent DNA structure: dinucleotide conformational maps. J Mol Biol 2000, 295:71-83.

[32] SantaLucia J, Jr., Hicks D: The thermodynamics of DNA structural motifs. Annu Rev Biophys Biomol Struct 2004, 33:415-440.
[33]Perez A, Noy A, Lankas F, Luque FJ, Orozco M: The relative flexibility of B-DNA and A-RNA duplexes: database analysis. Nucleic Acids Res 2004, 32:6144-6151.
[34]Barzilay I, Sussman JL, Lapidot Y: Further studies on the chromatographic behaviour of dinucleoside monophosphates. Journal of Chromatography A 1973, 79:139-146.
[35]Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH: Improved free-energy parameters for predictions of RNA duplex stability. Proc Natl Acad Sci U S A 1986, 83:9373-9377.
[36] Munteanu MG, Vlahovicek K, Parthasarathy S, Simon I, Pongor S: Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena. Trends Biochem Sci 1998, 23:341-347.
[37] Vlahovicek K, Kajan L, Pongor S: DNA analysis servers: plot.it, bend.it, model.it and IS. Nucleic Acids Res 2003, 31:3686-3687.
[38] Goodsell DS, Dickerson RE: Bending and curvature calculations in B-DNA. Nucleic Acids Res 1994, 22:5497-5503.
[39] Gromiha MM, Ponnuswamy PK: Hydrophobic distribution and spatial arrangement of amino acid residues in membrane proteins. Int J Pept Protein Res 1996, 48:452-460.
[40] Brukner I, Sanchez R, Suck D, Pongor S: Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. EMBO J 1995, 14:1812-1818.
[41] Satchwell SC, Drew HR, Travers AA: Sequence periodicities in chicken nucleosome core DNA. J Mol Biol 1986, 191:659-675.