Briefings in Bioinformatics, 00(00), 2020, 1–9

doi: 10.1093/bib/bbaa284
Problem Solving Protocol

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa284/5956158 by University of New England user on 06 November 2020

KNIndex: a comprehensive database of physicochemical properties for *k*-tuple nucleotides

Wen-Ya Zhang, Junhai Xu, Jun Wang, Yuan-Ke Zhou, Wei Chen and Pu-Feng Du

Corresponding authors: Junhai Xu, College of Intelligence and Computing, Tianjin University, Weijin Road Campus: No. 92 Weijin Road, Nankai District, Tianjin 300072, China. Tel.: +86-22-27891278; E-mail: jhxu@tju.edu.cn; Wei Chen, School of Life Sciences, North China University of Science and Technology, 21 Bohai Road, Caofeidian Xincheng, Tangshan 063210, China. Tel.: +86-315-8805607; E-mail: greatchen@ncst.edu.cn; Pu-Feng Du, College of Intelligence and Computing, Tianjin University, Weijin Road Campus: No. 92 Weijin Road, Nankai District, Tianjin 300072, China. Tel.: +86-22-23689450; E-mail: pdu@tju.edu.cn

Abstract

With the development of high-throughput sequencing technology, the genomic sequences increased exponentially over the last decade. In order to decode these new genomic data, machine learning methods were introduced for genome annotation and analysis. Due to the requirement of most machines learning methods, the biological sequences must be represented as fixed-length digital vectors. In this representation procedure, the physicochemical properties of *k*-tuple nucleotides are important information. However, the values of the physicochemical properties of *k*-tuple nucleotides are scattered in different resources. To facilitate the studies on genomic sequences, we developed the first comprehensive database, namely KNIndex (https://knindex.pufengdu.org), for depositing and visualizing physicochemical properties of *k*-tuple nucleotides. Currently, the KNIndex database contains 182 properties including one for mononucleotide (DNA), 169 for dinucleotide (147 for DNA and 22 for RNA) and 12 for trinucleotide (DNA). KNIndex database also provides a user-friendly web-based interface for the users to browse, query, visualize and download the physicochemical properties of *k*-tuple nucleotides. With the built-in conversion and visualization functions, users are allowed to display DNA/RNA sequences as curves of multiple physicochemical properties. We wish that the KNIndex will facilitate the related studies in computational biology.

Key words: physicochemical property; *k*-tuple nucleotide; KNIndex; database; web server

Wen-Ya Zhang is a Master Student in the College of Intelligence and Computing, Tianjin University. Her research interests focus on genomic sequence analysis.
Junhai Xu is an Associate Professor in the College of Intelligence and Computing, Tianjin University, His research interests include bioinformatics and biomedical image analysis.
Jun Wang is a Master student in the College of Intelligence and Computing, Tianjin University. His research interests focus on bioinformatics software developing and data visualization.
Yuan-Ke Zhou is a Master Student in the College of Intelligence and Computing, Tianjin University. His research interest focus on functional analysis of genomic sequences.
Wei Chen is a Professor in the School of Life Sciences, North China University of Science and Technology. His research is in the areas of bioinformatics, computational epigenetics and epitranscriptomics.
Pu-Feng Du is an Associate Professor in the College of Intelligence and Computing, Tianjin University. His research interests includes functional analysis of genomic and proteomic sequences.
Submitted: 22 August 2020; **Received (in revised form):** 17 September 2020

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Introduction

With the development and popularization of sequencing technology, the genomic sequences dramatically increased over the past 10 years. In order to decode these new genomic data, a large amount of machine learning based bioinformatics tools has been proposed for genome annotation and analysis. As illustrated in several literatures [1–5], when the machine learning methods are applied in the computational biology, the biological sequence must be firstly represented by fixed-length digital vectors or other forms of discrete representations rather than the sequence itself. Traditional methods to encode genomic sequences were based on k-tuple nucleotide frequency. Although it can include the long-range sequence order information to some extent, the k-tuple nucleotide frequency based method could not reflect the hidden physicochemical patterns buried in the genomic sequences.

In some cases, the physicochemical property is more effective than k-tuple nucleotide for describing the genomic sequences, which can extract the hidden information in the genome. For example, by plotting the six local DNA structural properties (i.e. twist, tilt, roll, shift, slide and rise), Goñi *et al.* [6] obtained the distinctive structural property profiles surrounding the promoters and then proposed a computational method for identifying promoters based on DNA structural properties. The effectiveness of physicochemical properties of k-tuple nucleotide was also demonstrated in analyzing indels [7, 8], replication origins [9], promoters [10, 11], DNA local structures [12–14], and some other functional elements [15–18].

Accordingly, several bioinformatics tools, namely DiProDB [19], PseKNC-General [2], Pse-in-One [3], UltraPse [4] and VisFeature [20], have been proposed to describe and visualize the genomic sequences by using the physicochemical properties of k-tuple nucleotide. Benefitting from these tools, a series of computational methods have been developed for annotating the genomic sequences, such as predicting nucleosome positioning sequences [21], identifying RNA modification sites [22–24], identifying promoters [10, 25], predicting bacterial transcriptional terminators [26], etc. Although significant achievements have been made in decoding genomic sequences by using the physicochemical properties of k-tuple nucleotide, there are still some problems that need to be solved.

At present, to the best of our knowledge, 182 kinds of physicochemical properties for mononucleotide, dinucleotide and trinucleotide have been reported. The only existing database depositing these physicochemical properties is the DiProDB database. However, DiProDB only includes part of these reported physicochemical properties. Most of these physicochemical properties were scattered in different resources. Therefore, it is necessary to build a database for depositing these reported physicochemical properties. The intuitive visualization of genomic sequences by using physicochemical properties will be helpful for researchers to develop a new method for genome annotation. Thus, a human-readable, comprehensive and searchable collection of the physicochemical property of k-tuple nucleotide is demanded.

To this end, we constructed the KNIndex database with web-based user-friendly interfaces, which is available at https://knindex.pufengdu.org. Currently, the KNIndex database deposits the 182 reported physicochemical properties, one of which is for mononucleotide (DNA), 169 for dinucleotide (147 for DNA and 22 for RNA) and 12 for trinucleotide (DNA). The values for each physicochemical property is provided in two different forms, its original values and the standardized values. These physicochemical properties together with their values can be freely downloaded from the well-designed web interface. Additionally, we provide a query interface for users to search the physicochemical properties by their names. In the KNIndex database, it is easy to query, browse, visualize and download the physicochemical properties of k-tuple nucleotides. The KNIndex database also integrates a function that allows the users to convert DNA/RNA sequences into a series of numerical physicochemical property values. We believe that KNIndex will facilitate a wide range of related studies in computational biology.

Materials and methods

Data curation

We collected the physicochemical property of k-tuple nucleotide from two major sources: existing physicochemical property databases and published scientific literatures. Before collecting these data, we first determined the value k of k-tuple nucleotides. Based on the physicochemical properties available so far, we set k values to 1, 2 and 3, which corresponds to mononucleotide, dinucleotide and trinucleotide, respectively. For mononucleotides, according to the data we have collected, there is merely one kind of physicochemical property, called electron-ion interaction pseudopotential (EIIP) [27]. For dinucleotides and trinucleotides, after collecting the values from PseKNC [1], PseKNC-General [2], Pse-in-One [3], UltraPse [4], VisFeature [20] and DiProDB [19], we manually cross-validated the property values from different sources and integrated them together. All physicochemical properties are provided in two different forms, the original and standardized forms. The standardized values are computed from the original values according to the following method.

Considering one kind of physicochemical property of the k-tuple nucleotide, the corresponding physicochemical property set H for the 4^k different k-tuple nucleotides is as following:

$$H = \left\{ h_i | i = 1, 2, \cdots, 4^k \right\}, \tag{1}$$

where h_i is the original value of the physicochemical property for the i-th k-tuple nucleotide.

The standardized physicochemical property value of the i-th type of k-tuple nucleotide is θ_i, which can be calculated as follows:

$$\theta_i = \frac{h_i - m(H)}{\sigma(H)}, \tag{2}$$

where

$$m(H) = \frac{1}{4^k} \sum_{i=1}^{4^k} h_i, \qquad \text{and} \tag{3}$$

$$\sigma(H) = \sqrt{\frac{1}{4^k} \sum_{i=1}^{4^k} [h_i - m(H)]^2}. \tag{4}$$

$m(H)$ is the average value of the physicochemical properties for the 4^k different k-tuple nucleotides and $\sigma(H)$ is the standard deviation.

In addition, the original references for every physicochemical property were also obtained from PubMed or Google Scholar. All values in the KNIndex database are manually rechecked one-by-one against the original literature contents. Finally, 182 physicochemical properties and 47 original references were collected.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa284/5956158 by University of New England user on 06 November 2020

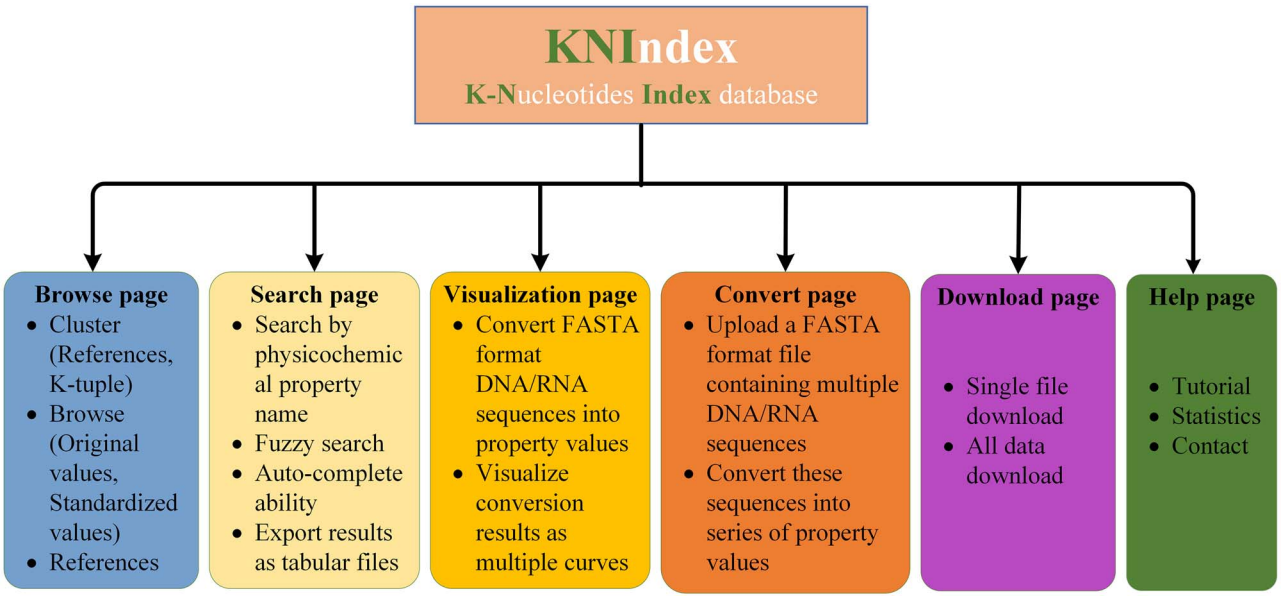

Figure 1. Overview of the architecture of KNIndex.

Database and web interface design

We deposited all physicochemical property values, as well as the corresponding literatures, into a MySQL database. The web interface system of KNIndex is completely implemented by using JavaScript, on both the frontend and backend. We implemented the backend system using JavaScript with node.js script engine and the Express framework. The frontend was implemented using JavaScript with the Vue framework. A user-friendly web-based interface was provided with modern web dynamical effects. We designed six primary functional pages: the 'Browse' page for convenient data browsing, the 'Search' page for quick searching and data retrieving, the 'Visualization' and 'Convert' pages for visualizing and converting DNA/RNA sequences into numerical physicochemical property values, the 'Download' page for downloading all values, and the 'Help' page for guiding users to use the web server. The architecture of KNIndex is illustrated in Figure 1.

Data table design

We designed 10 data tables in the MySQL database to store different types of physiochemical property values, as well as the corresponding literature information. Due to the requirement of efficient database query and data retrieving, we separate physicochemical properties of different k values into different tables. For each physicochemical property record, it contains the following fields: ID, property name, and cross-reference IDs. For different values of k, the number of value fields is different. Table 1 gives a comprehensive list of the table schema for different types of k-tuple nucleotides. Additionally, tables for two different property value forms (e.g. the original values and the standardized values) of the same physicochemical property have the same schema.

Implementation of DNA/RNA sequence conversion

The physicochemical properties of k-tuple nucleotides are primarily used to generate numerical representations of biological sequences. The basic usage of the property values of k-tuple nucleotides is to convert DNA/RNA sequences into a series of numerical physicochemical property values. We implemented this function by integrating an in-house version of the VisFeature program in the backend of our web service. Users should enter DNA/RNA sequences in the textbox or upload a FASTA format file containing multiple DNA/RNA sequences. The backend of our server will divide the DNA/RNA sequences into basic k-mers according to the parameter k. These k-mers will be replaced with the corresponding k-tuple nucleotide physicochemical property values according to the users' choices.

In order to help the users to gain a better understanding of the DNA/RNA sequence conversion module, we take a short DNA sequence 'AATCGAATCGGCTAGTCCAATAGTACGTAG' as an example. The conversion workflow is illustrated in Figure 2. The sequence was first divided into a serial of 2-mers, i.e. AA, AT, TC, . . . , AG. When the six physicochemical properties, including 'Twist,' 'Tilt,' 'Roll,' 'Shift,' 'Slide' and 'Rise' were selected, the conversion process would replace the 2-mers with each of the six corresponding values to generate six series of numbers. The original values of these six properties for 2-tuple nucleotides are given in Table 2. The final conversion results are recorded in Figure 2.

Results

Web server utilities

Browse

This module allows the users to conveniently browse all the manually curated data. We catalogued the physicochemical properties according to different k values ($k = 1, 2, 3$), and different literatures, respectively. The users can browse the database to grasp the basic information on these properties easily. To increase the readability of the 'Browse' page, a sidebar for the navigation purpose was added. It provides the contents hierarchically as in Figure 1. Due to the amount of data transferring from the backend to the frontend, it would takes 3–5 s to be completely loaded.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa284/5956158 by University of New England user on 06 November 2020

Table 1. Basic data schemas of the tables for mononucleotide (DNA), dinucleotide (DNA and RNA) and trinucleotide (DNA)

k-mer types	Common fields	Specific fields
Mononucleotide	ID, PropertyName, ReferID, PubMedID	A, C, T, G,
Dinucleotide(DNA)		AA, AC, AT, AG, CA, CC, CT, CG, TA, TC, TT, TG, GA, GC, GT, GG
Dinucleotide(RNA)		AA, AC, AU, AG, CA, CC, CU, CG, UA, UC, UU, UG, GA, GC, GU, GG
Trinucleotide		AAA, AAC, AAT, AAG, ACA, ACC, ACT, ACG, ATA, ATC, ATT, ATG, AGA, AGC, AGT, AGG, CAA, CAC, CAT, CAG, CCA, CCC, CCT, CCG, CTA, CTC, CTT, CTG, CGA, CGC, CGT, CGG, TAA, TAC, TAT, TAG, TCA, TCC, TCT, TCG, TTA, TTC, TTT, TTG, TGA, TGC, TGT, TGG, GAA, GAC, GAT, GAG, GCA, GCC, GCT, GCG, GTA, GTC, GTT, GTG, GGA, GGC, GGT, GGG

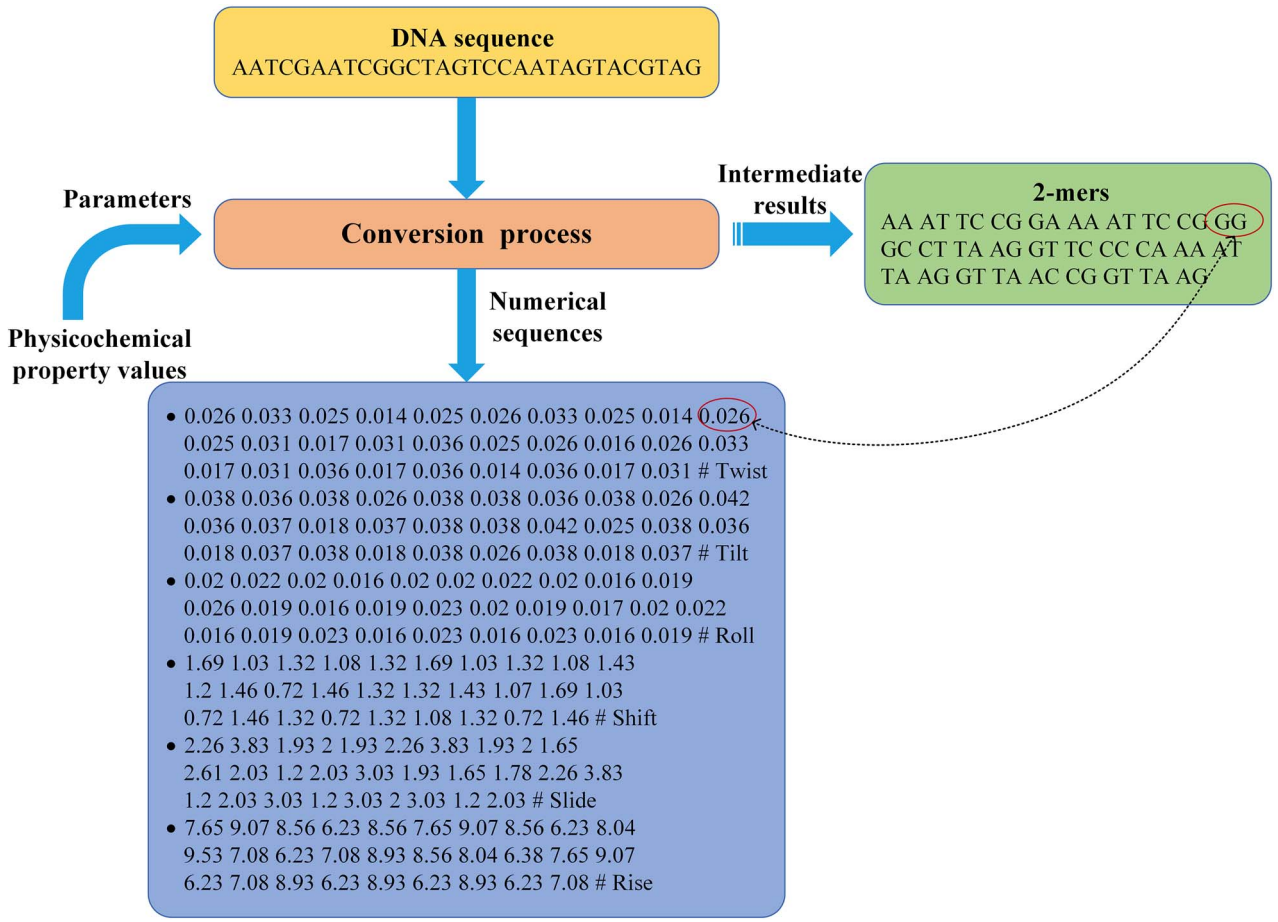

Figure 2. Workflow of the conversion method. The schematic shows a concrete example of the conversion process, taking six physicochemical properties, including 'Twist,' 'Tilt,' 'Roll,' 'Shift,' 'Slide' and 'Rise' into consideration. The DNA sequence was first divided into a serial of 2-mers, like AA, AT, TC, . . . , AG, which are intermediate results of the conversion process. The conversion process would replace the 2-mer with each of the six corresponding property values to generate six numerical sequences when six properties were chosen.

Search

A searchable collection of physicochemical property values is demanded despite that a powerful browsing function has been developed. We developed a smart search interface for conveniently retrieving the physicochemical properties of *k*-tuple nucleotides. On the 'Search' page, the query box supports not only searching by the property name but also wildcards queries using '*' and '%.' Furthermore, the query box has auto-complete abilities. A dropdown menu will be displayed when the content of the box matches the first few letters of a property name (Figure 3A). Consider searching for physicochemical properties that contain 'twist' using 'tw*t' when 'original-di-DNA' is chosen as the type of nucleotides. The search results are presented as tables on the web, with links to the original

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa284/5956158 by University of New England user on 06 November 2020

Table 2. Original values of six physiochemical properties in the example

2-mer	Twist	Tilt	Roll	Shift	Slide	Rise
AA/TT	0.026	0.038	0.020	1.69	2.26	7.65
AC/GT	0.036	0.038	0.023	1.32	3.03	8.93
AG/CT	0.031	0.037	0.019	1.46	2.03	7.08
AT	0.033	0.036	0.022	1.03	3.83	9.07
CA/TG	0.016	0.025	0.017	1.07	1.78	6.38
CC/GG	0.026	0.042	0.019	1.43	1.65	8.04
CG	0.014	0.026	0.016	1.08	2.00	6.23
GA/TC	0.025	0.038	0.020	1.32	1.93	8.56
GC	0.025	0.036	0.026	1.20	2.61	9.53
TA	0.017	0.018	0.016	0.72	1.20	6.23

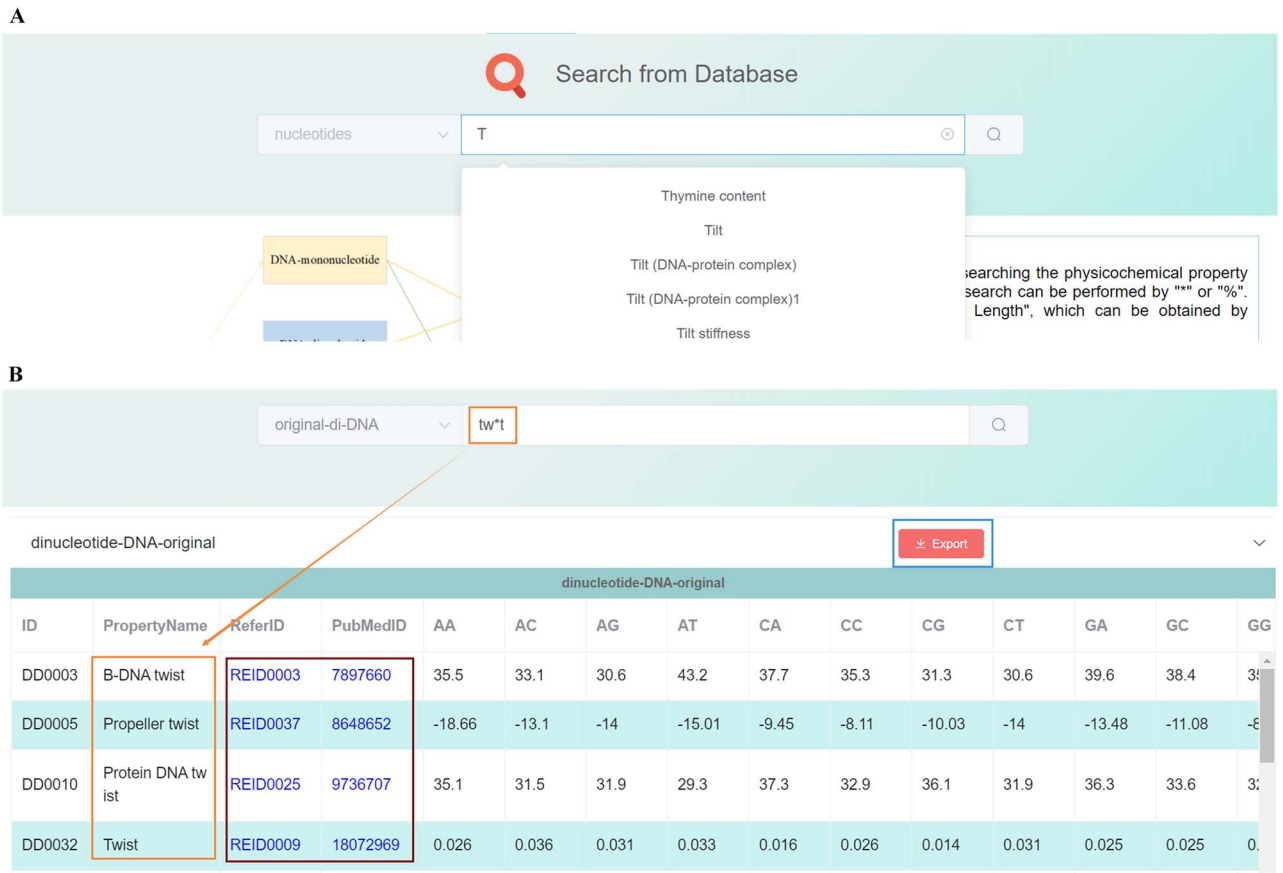

Figure 3. 'Search' page and function illustration in KNIndex. (**A**) The query box has auto-complete abilities. A dropdown menu will be displayed when the content of the box matches the first few letters of a property name. (**B**) The query box supports fuzzy searches using '*' and '%.' Taking 'twist' as an example, use 'tw*t' for fuzzy search when 'original-di-DNA' is selected as the type of nucleotides. The search results are presented as tables, with links to the original references. The results can be exported as tabular files by clicking on the 'Export' button.

references. The results can be exported as tabular files by clicking on the 'Export' button (Figure 3B). In addition, when the nucleotide type is not chosen, and the property name is not entered, all properties will be displayed if users click on the search button.

Visualization and convert

For convenience, the 'Visualization' and 'Convert' pages were implemented in the KNIndex as well. On the 'Visualization' page, the genomic sequences in FASTA format could be converted into a series of property values according to the selected parameters and physicochemical properties. The series of property values can be visualized as multiple curves and saved as images. For a large number of sequences, the users can upload these sequences through a FASTA format file on the 'Convert' page. The conversion process may take some time to process, which depends on the size of the uploaded file. We, therefore, provide the users with a link to download the results after a while. The results are provided as comma-separated vector files. Each row in the file represents the numerical values of a sequence in one type of physicochemical property. The id of the sequence, name

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa284/5956158 by University of New England user on 06 November 2020

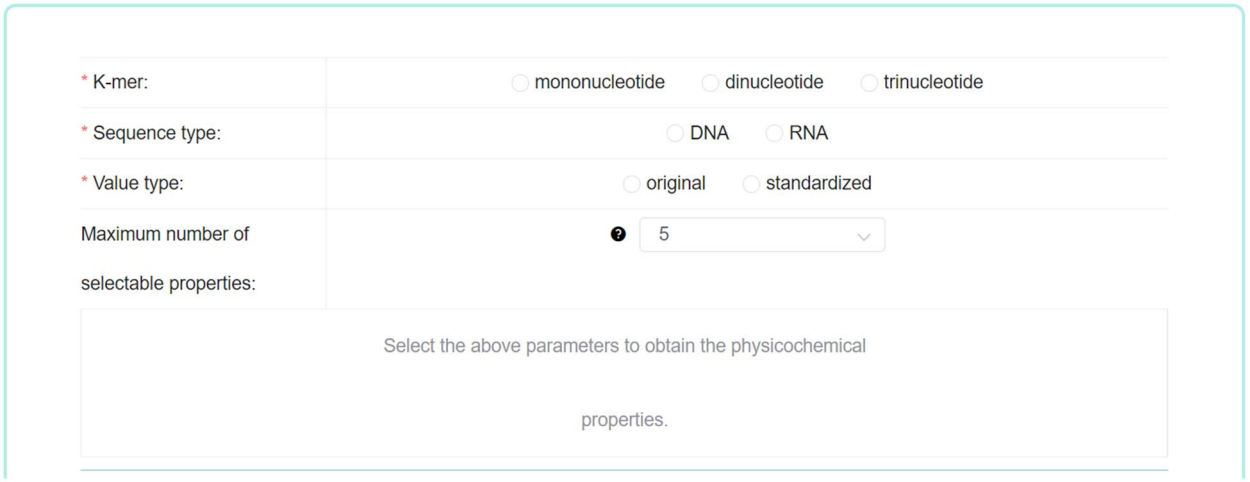

Figure 4. Parameters selection process on the 'Visualization' and 'Convert' pages. After selecting the first three parameters, including 'K-mer,' 'Sequence type' and 'Value type,' corresponding physicochemical properties will be displayed in tables for users to choose. Due to the limited computational resources, users should set the maximum number of selectable physicochemical properties, which is 5 by default.

Figure 5. Case study of visualization. We took the same example in Figure 2 to illustrate a case of visualization. First, the user should select the first three parameters, including dinucleotide, DNA and original. The corresponding physicochemical properties will be displayed for the user to choose. The user can alter the maximum number of selectable physicochemical properties to 10 before selecting the actual properties. After clicking the 'Visualize' button, multiple curves can be displayed as the visualization of results.

and type of the physicochemical properties are marked at the end of the line, with a '#' to separate them from the values. These two utilities have the same parameter selection process (Figure 4).

Download and help

The raw data of the entire KNIndex database can be freely downloaded for academic purposes. On the 'Download' page, users can download the data table in the KNIndex database as tabular files by clicking on the corresponding file name. A file named 'all.zip' can also be downloaded to save all tables in one click.

To make the web server more convenient to use, we designed a help page. The 'Help' page provides three parts. The 'Tutorial' part gives step-by-step instructions on how to use the KNIndex database. The 'Statistics' part lists the basic counts of different types of physicochemical properties. The 'Contact' part provides details on how to contact authors of the database.

Simple visualization case study

We take the same example in Figure 2 to demonstrate the operation on our website in a step-by-step way (Figure 5). After selecting the first three kinds of parameters (dinucleotide, DNA and original), all corresponding properties will be displayed for the

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa284/5956158 by University of New England user on 06 November 2020

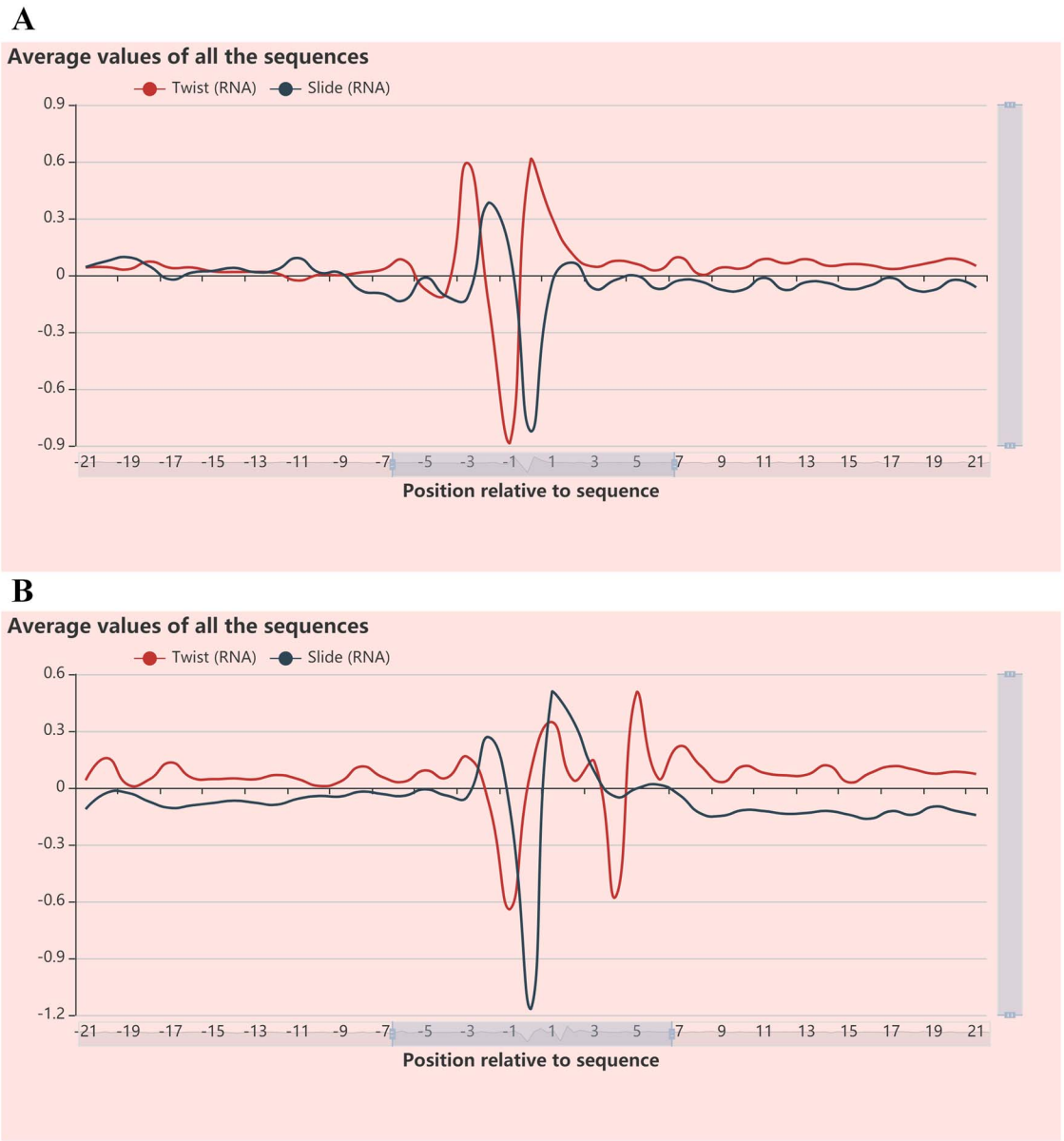

Figure 6. The average property for RNA splicing sites based on 'Twist' and 'Slide.' (**A**) The average property curve of all the splicing acceptor sites. (**B**) The average property of all the splicing donor sites.

user to choose. By altering the maximum number of selectable physicochemical properties to 10, the user can choose up to 10 kinds of properties. By clicking the visualization button, multiple curves will be displayed on a new page. The visualization can appear in either curves or bars. The raw data that generates curves or bars can also be displayed in the data view. The visualization result can be downloaded as an image.

Application case study

To demonstrate the effectiveness of our KNIndex database and its built-in functions, we analyzed RNA splicing sites by using the KNIndex database. The sequences including splicing acceptor and donor sites were obtained from previous work [28]. In this case, we plotted the profiles of 'Twist' and 'Slide' surrounding splicing acceptor and donor sites. Different trends were observed for splicing sites and the surrounding sequences (Figure 6A and B). The curves obtained from KNIndex indicated that the Twist property has two peaks surrounding a deep valley for the acceptor sites. The Slide property pattern of the acceptor sites and the donor sites are just the opposite. The Slide curve shows a first up and then down pattern for the acceptor sites, while the pattern for the donor sites is first down and then up. These hidden patterns might be the signal for splicing factors to accurately recognize splicing sites and will be helpful for computationally identifying RNA splicing sites. These observations consist of the reports in the literature [28], which demonstrate the usefulness and the convenience of the KNIndex database in practical studies.

Self-host availability

For the users who want to host KNIndex themselves for in-house academic purpose, we provide all source codes of KNIndex,

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa284/5956158 by University of New England user on 06 November 2020

which can be obtained from the GitHub repository (https://github.com/wyzhang0401/KNIndex). Users can establish their own mirror sites of KNIndex on their own servers to utilize their own computational powers. A step-by-step instruction has been provided in the GitHub repository.

Conclusion

To the best of our knowledge, KNIndex (https://knindex.pufengdu.org) is the first database focusing on collecting the physicochemical properties of *k*-tuple nucleotides, including mononucleotides, dinucleotides and trinucleotides. We provide the most comprehensive collection of the physicochemical property values of *k*-tuple nucleotides and a user-friendly web-based interface for researchers to obtain all values. With the development of the high-throughput sequencing technology, KNIndex database will greatly promote a wide range of studies that are focusing on the genomic sequences. There will be more physicochemical properties and more built-in functions in the KNIndex database in the future. It is anticipated that our database will be a useful resource and a basic utility in decoding functional information in genomic sequences.

Key Points

- The first comprehensive database, KNIndex, for *k*-tuple nucleotide (mononucleotide, dinucleotide and trinucleotide) physicochemical properties was developed.
- KNIndex curates 182 physicochemical properties including one for mononucleotide (DNA), 169 for dinucleotide (147 for DNA and 22 for RNA) and 12 for trinucleotide (DNA).
- KNIndex provides a function to convert a DNA/RNA sequence into series of property values and visualize the sequence as multiple curves. It also integrates a tool for converting DNA/RNA sequence datasets into series of numerical values of physicochemical properties.
- A user-friendly web-based interface is provided, from which users can browse, search, visualize and download the physicochemical properties of mononucleotides, dinucleotides and trinucleotides.

Funding

This work was supported by National Natural Science Foundation of China [NSFC 61872268 and 31771471]; National Key R&D Program of China [2018YFC0910405]; Natural Science Foundation for Distinguished Young Scholar of Hebei Province [C2017209244]; and Open Project Funding of CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences [CASNDST201705].

References

1. Chen W, Lei T-Y, Jin D-C, *et al*. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal Biochem* 2014;**456**:53–60.
2. Chen W, Zhang X, Brooker J, *et al*. PseKNC-general: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* 2015;**31**:119–20.
3. Liu B, Liu F, Wang X, *et al*. Pse-in-one: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res* 2015;**43**:W65–71.
4. Du P-F, Zhao W, Miao Y-Y, *et al*. UltraPse: a universal and extensible software platform for representing biological sequences. *Int J Mol Sci* 2017;**18**:2400.
5. Chen Z, Zhao P, Li F, *et al*. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**:2499–502.
6. Goñi JR, Pérez A, Torrents D, *et al*. Determining promoter location based on DNA structure first-principles calculations. *Genome Biol* 2007;**8**:R263.
7. Chen W, Zhang L. The pattern of DNA cleavage intensity around indels. *Sci Rep* 2015;**5**:8333.
8. Greenbaum JA, Pang B, Tullius TD. Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res* 2007;**17**:947–53.
9. Li W-C, Deng E-Z, Ding H, *et al*. iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. *Chemom Intel Lab Syst* 2015;**141**:100–6.
10. Lin H, Deng E-Z, Ding H, *et al*. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res* 2014;**42**:12961–72.
11. Abeel T, Saeys Y, Bonnet E, *et al*. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res* 2008;**18**:310–23.
12. Parker SCJ, Hansen L, Abaan HO, *et al*. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 2009;**324**:389–92.
13. Brukner I, Sánchez R, Suck D, *et al*. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J* 1995;**14**:1812–8.
14. Olson WK, Gorin AA, Lu XJ, *et al*. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci USA* 1998;**95**:11163–8.
15. Chen W, Xing P, Zou Q. Detecting N 6-methyladenosine sites from RNA transcriptomes using ensemble support vector machines. *Sci Rep* 2017;**7**:40242.
16. He W, Ju Y, Zeng X, *et al*. Sc-ncDNAPred: a sequence-based predictor for identifying non-coding DNA in Saccharomyces cerevisiae. *Front Microbiol* 2018;**9**:2174.
17. Zeng X, Lin W, Guo M, *et al*. A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput Biol* 2017;**13**:e1005420.
18. Zeng X, Yuan S, Huang X, *et al*. Identification of cytokine via an improved genetic algorithm. *Front Comp Sci* 2015;**9**:643–51.
19. Friedel M, Nikolajewa S, Sühnel J, *et al*. DiProDB: a database for dinucleotide properties. *Nucleic Acids Res* 2009;**37**:D37–40.
20. Wang J, Du P-F, Xue X-Y, *et al*. VisFeature: a stand-alone program for visualizing and analyzing statistical features of biological sequences. *Bioinformatics* 2019;**36**:1277–8.
21. Guo S-H, Deng E-Z, Xu L-Q, *et al*. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* 2014;**30**:1522–9.
22. Chen W, Ding H, Zhou X, *et al*. iRNA (m6A)-PseDNC: identifying N6-methyladenosine sites using pseudo dinucleotide composition. *Anal Biochem* 2018;**561**:59–65.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa284/5956158 by University of New England user on 06 November 2020

23. Akbar S, Hayat M, Iqbal M, *et al*. iRNA-PseTNC: identification of RNA 5-methylcytosine sites using hybrid vector space of pseudo nucleotide composition. *Front Comp Sci* 2020;**14**:451–60.
24. Chen W, Tang H, Ye J, *et al*. iRNA-PseU: identifying RNA pseudouridine sites. *Mol Ther Nucleic Acids* 2016;**5**:e332.
25. Liu B, Yang F, Huang D-S, *et al*. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 2018;**34**:33–40.
26. Feng C-Q, Zhang Z-Y, Zhu X-J, *et al*. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 2019;**35**:1469–77.
27. Nair AS, Sreenadhan SP. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation* 2006;**1**:197–202.
28. Chen W, Feng P-M, Lin H, *et al*. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *Biomed Res Int* 2014;**2014**:623149.

Downloaded from https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa284/5956158 by University of New England user on 06 November 2020