BMC Genomics

RESEARCH Open Access

Distribution rules of 8-mer spectra and characterization of evolution state in animal genome sequences

Xiaolong Li[1], Hong Li[1*], Zhenhua Yang[2] and Lu Wang[1]

Abstract

Background Studying the composition rules and evolution mechanisms of genome sequences are core issues in the post-genomic era, and k-mer spectrum analysis of genome sequences is an effective means to solve this problem.

Result We divided total 8-mers of genome sequences into 16 kinds of XY-type due to XY dinucleotides number in 8-mers. Previous works explored that the independent unimodal distributions observed only in three CG-type 8-mer spectra, while non-CG type 8-mer spectra have not the universal phenomenon from prokaryotes to eukaryotes. On this basis, we analyzed the distribution variation of non-CG type 8-mer spectra across 889 animal genome sequences. Following the evolutionary order of animals from primitive to more complex, we found that the spectrum distributions gradually transition from unimodal to tri-modal. The relative distance from the average frequency of each non-CG type 8-mers to the center frequency is different within a species and among different species. For the 8-mers contain CG dinucleotides, we further divided these into 16 subsets, where each 8-mer contains both CG and XY dinucleotides, called XY1_CG1 subsets. We found that the separability values of XY1_CG1 spectra are closely related to the evolution and specificity of animals. Considering the constraint of Chargaff's second parity rule, we finally obtained 10 separability values as the feature set to characterize the evolution state of genome sequences. In order to verify the rationality of the feature set, we used 14 common classification algorithms to perform binary classification tests. The results showed that the accuracy (*Acc*) ranged between 98.70% and 83.88% among birds, other vertebrates and mammals.

Conclusion We proposed a credible feature set to characterizes the evolution state of genomes and obtained satisfied results by the feature set on large scale classification of animals.

Keywords Animal genomes, 8-mer spectra, Feature set, Evolution state of genome, Species classification

*Correspondence:
Hong Li
ndlihong@imu.edu.cn
[1]Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China
[2]School of Economics and Management, Inner Mongolia University of Science and Technology, Baotou 014010, China

Introduction

Mining the biological information contained in nucleotide sequences is an eternal theme of molecular biology. The composition of nucleotide sequences is non-random, and this non-randomness can be expressed by the non-random distribution of the occurrence frequency of k-mers. K-mer is a nucleotide fragment with length *k* in a nucleotide sequence, and the preference of certain k-mers is an important means to explore the composition structure and evolution information of nucleotide sequences.

Since the 1980s, k-mer feature analysis has gradually become an effective means for exploring functional fragments and evolution relationships of nucleotide sequences [1]. In the field of nucleotide sequence analysis, the frequency distribution of k-mers not only reveals the basic composition of sequences but also deeply reflects the uniqueness of their structure and function. Researchers constructed Markov transition matrices [2] and artificial neural networks (ANN) [3] using k-mer frequencies (k=2, 3, 4) to predict and identify gene promoter regions. The support vector machine (SVM) algorithm was used to predict enhancer regions of gene sequences based on k-mer features (k=6, 7) [4]. Through clustering analysis of rare k-mers, researchers predicted CpG island sequences and promoter regions [5]. Based on k-mer frequency, the incremental diversity and quadratic discriminant analysis (IDQD) was employed to predict the potential formation and nucleosome positioning [6]. Combining the abundance and location information of k-mers, various machine learning models were used to predict the interactions between RNA and proteins, as well as the subcellular localization of lncRNA and MicroRNA [7–13]. By analyzing the k-mer features (k=2, 3, 4) in helix and loop regions, simulated selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) to predict RNA secondary structure [14].

From sequence alignment to genome assembly, k-mers play an important role. The sequence comparison algorithms created by k-mer frequency information can quickly identify sequence contamination and illegal sequences [15]. Additionally, combining k-mer frequency information with various correction techniques can effectively improve the correction quality of sequencing data and genome assembly indicators [16]. Direct analysis of the abundance of different k-mers in sequences could help identify the repetitive structure of genome sequences and estimate the genome size [17]. In RNA-seq data analysis, k-mers with varying abundances could help detect the variation of data sets [18].

In the fields of biodiversity research and microbiology, k-mer analysis is increasingly important for non-sequence alignment methods. It has been found that different microbial genomes can be distinguished by using the relative abundance of 2-mers [19]. The genome barcode constructed from k-mer frequency distribution not only helps solve the problem of metagenome binning but also plays an important role in species identification and horizontal transfer gene identification [20–22]. Additionally, the multifunctional database KGCAK based on k-mers as a kind of genome element provided a new non-sequence alignment method for constructing phylogenetic relationships [23]. The concept of k-mer distance between the same chromosome of two species opened up a new way to analyze the evolution relationships of prokaryotes and eukaryotes [24].

In the interdisciplinary field of medicine and bioinformatics, the application of k-mers has been extended to the diagnosis, treatment of diseases and oncology research. Using k-mers in mixed logistic regression model, researchers could simulate the DNA methylation sensitivity of different cell types [25], which is essential for understanding the epigenetic state of cells. The k-mer counting method has been used to infer the evolution and lineage of tumor cells in single-cell sequencing data [26]. The non-invasive prenatal testing (NIPT) method based on k-mers opened up a new way for early diagnosis and treatment of genetic diseases [27]. K-mer decomposition reading technology could detect single-base mutations, insertion, deletion mutations and fusion and other types of genetic variations [28, 29], which is significant for precision medicine and early disease diagnosis. Additionally, the supervised machine learning models based on k-mers have been applied to predict the antibiotic resistance of bacteria [30], which is strategically important for combating resistant bacteria. In virology research, the comparative analysis of k-mer frequency distribution has been used to deeply understand the pathogenicity, origin, transmission and evolution of coronavirus, which provides a scientific basis for the global prevention and control of coronavirus epidemic [31–33].

Since the implementation of the Human Genome Project, an increasing number of whole genome sequences have been obtained. Exploring species evolution and diversity at the genome level has become an essential issue. It is known that the occurrence frequency of k-mers in genome sequence effectively characterizes the composition and evolution information of genome sequence. Therefore, k-mer spectrum analysis has become an important tool in many biological fields, such as genome information interpretation, species classification and identification, and evolutionary history research. Annotating k-mers in a single species to find statistically significant ones [34]. This provides valuable insights into gene regulation and expression. Comparing genomes through whole genome k-mer signature analysis could reveal significant differences in genomic signatures and find monophyletic group for specific species [35]. These findings are

crucial for understanding the evolutionary history and classification of species at the whole genome data level. Using the concept of k-mer spectra and its related mathematical properties [36], unknown species can be identified through metagenomic fragments reconstruction, and the differences between metagenomic samples can be calculated through k-mer spectra, thereby the composition differences and evolution information of the intestinal microbiomes can be analyzed easily in different populations [37]. Additionally, researchers found that the k-mer spectra of genome sequences are unimodal and non-normal distribution in prokaryotes, and their distribution regularity is closely related to the G+C content of genomes [38]. In yeast and zebrafish, the k-mer spectra of genome sequences are unimodal, while the k-mer spectra of the human genome sequence is multimodal [39]. Furthermore, researchers found that the k-mer spectra of tetrapod animals are all multimodal, and the causes of non-normal distribution were studied [40]. These results shown that the k-mer spectra of genome sequences are closely related to the evolution levels of species.

Our research group conducted further studies on the k-mer spectra of genome sequences. To reveal the unimodal and multimodal phenomenon of 8-mer spectra, we divided total 8-mers into three subsets of XY2, XY1 and XY0 due to XY dinucleotide number in 8-mers, and discussed the spectrum distribution features of these 8-mer subsets. Based on nearly a thousand species genomes from prokaryotes to eukaryotes, we found that only the 8-mer spectra of CG2, CG1 and CG0 subsets form independent unimodal distributions in genome sequences, this phenomenon has species universality. We called the phenomenon as CG independent selection rule. We found that the CG independent selection intensity is closely related to the species evolution. Additionally, we found that TA independent selection intensity is also related to the species evolution, and that CG and TA independent selection intensities exist the mutual inhibition relation. Thus, we proposed an evolution mechanism of genome sequences as following: CG and TA independent selection intensities as well as their mutual inhibition relation characterize the evolution state of genome sequences [41].

The 8-mers containing CG dinucleotides are not only related to species evolution but also are functional motifs in nucleosome and CpG island sequences [42, 43]. Based on the evolution mechanism of genome sequences, we constructed an objective feature set based on the 8-mer relative frequency, and for the first time constructed the evolution relationships of animals at the genomic level, and achieved satisfactory results [44]. In previous research [41], we obtained two features to characterize the evolution state of genome sequences, but the two features only show the basic properties about genome evolution. In this study, we will analyze the spectrum distribution patterns of all XY-type 8-mers in detail within animal genome sequences, and explore the relationship between spectrum distributions of XY-type 8-mers and the evolution as well as the composition differences of species genomes. We will try to give a feature set based on the 8-mer spectra to characterize the evolution state of genome sequences, provide theoretical supports for constructing the evolution relationship at large-scale and cross-species level, and provide new perspectives and clues for understanding the evolution mechanism of genome sequences.

Results

Distribution features of 8-mer spectra in genome sequences

The k-mer frequency set is a window for external display of genome sequence information, and the k-mer spectra show the composition and evolution features of genome sequences. In our study, we chose k=8 for the following reasons. First, there is an empirical formula in statistics for k-mer selection. The chosen k value must ensure that the rare k-mer frequency must be guaranteed to meet the statistical significance in a given DNA sequence. Chor proposed a formula $k=0.7log_4L$ to estimate the minimum k value, L is the length of the given DNA sequence [40]. In eukaryote genomes, the yeast genome is short and the calculated k value is 8.9. Without loss of generality, 8-mer was selected in eukaryotic genomes. Second, the base composition of the genome sequence is long range correlated. If the k value is too small, a lot of information will be lost. An appropriate k value makes k-mers contain more information of the DNA sequences. Third, in our previous research, we tried k values from 3 to 11 and found that the spectrum distributions of k-mers tends to be stable when k>6. In addition, researchers have conducted research based on 8-mers and obtained significant research results [34, 35]. Taking the above reasons into consideration, we finally chose $k=8$.

Previous work mainly explored the spectrum distribution features of the CG-type 8-mers and their relationship to species evolution. Here, we further analyzed the spectrum distribution features of all 16 XY-type 8-mers in animal genome sequences. Only the spectrum distributions of all 16 XY-type 8-mers in human genome sequence were shown in Fig. 1.

It can be seen that the total 8-mer spectrum of human genome sequence shows a tri-modal distribution. The three peaks of the 8-mer spectrum are referred to as peak 2, peak 1 and peak 0 from low to high frequency. We found that only the 8-mer spectra of the three CG-type subsets CG2, CG1 and CG0 each form an independent unimodal distribution, and these three independent unimodal distributions coincided respectively with peak 2,

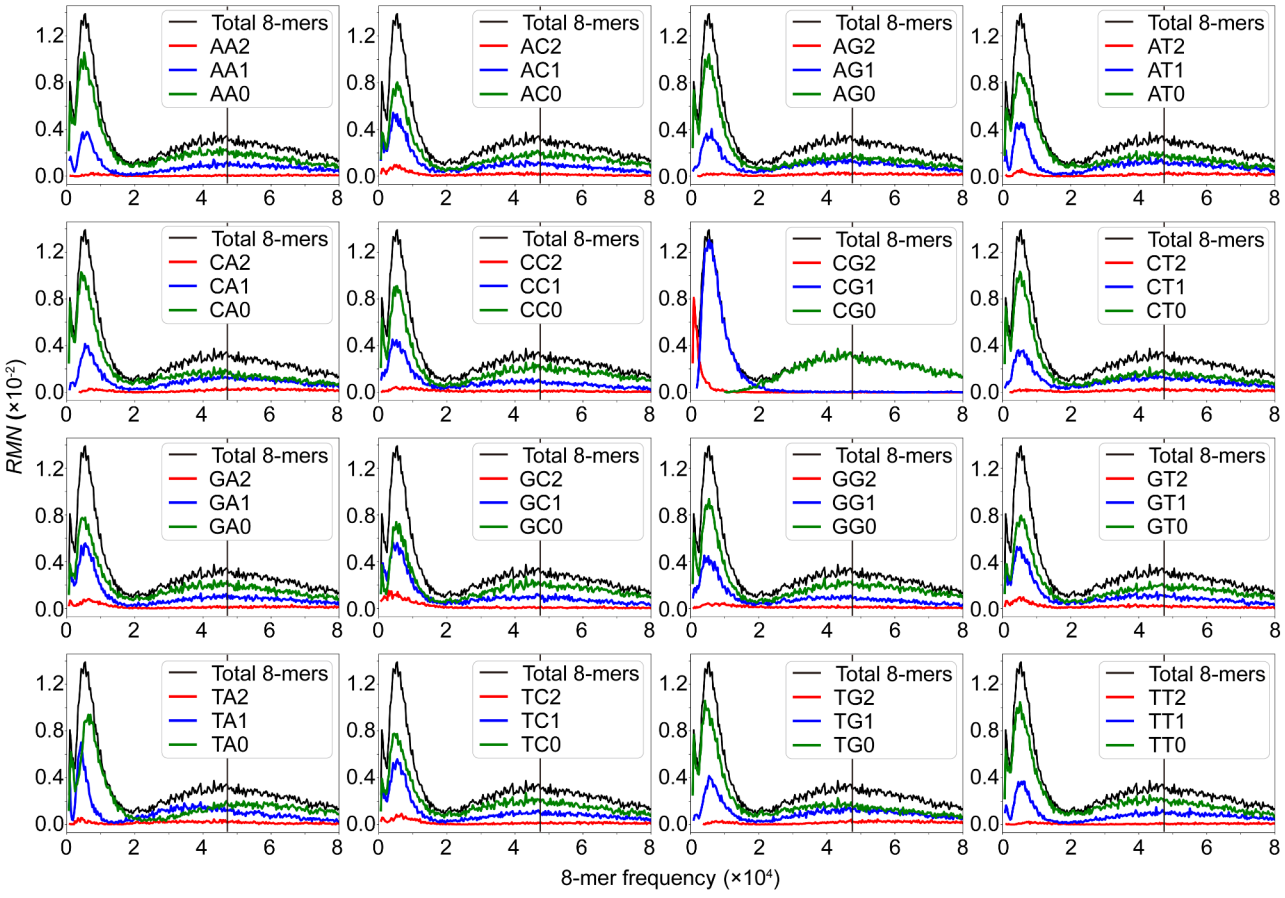

Fig. 1 Spectrum distributions of XY-type 8-mers in human genome sequence. The black curves represent the spectrum of total 8-mers, the red curves represent the spectrum of XY2 8-mers, the blue curves represent the spectrum of XY1 8-mers, and the green curves represent the spectrum of XY0 8-mers. The vertical line is the average frequency of total 8-mers, called center frequency

peak 1 and peak 0 of the total 8-mer spectrum. The 8-mer spectra of the other XY2, XY1 and XY0 subsets are still tri-modal distributions. This is one of the most important composition features of genome sequences. Previous work only studied the spectrum distribution features of the 3 CG-type 8-mers and the relationships between the separability values of CG-type 8-mer spectra and the evolution of species genomes [41, 44]. The spectrum distribution features of the other 15 XY-type 8-mers were not discussed. In this paper, we discussed the spectrum distribution features of all 16 XY-type 8-mers. Based on 889 animal genome sequences, the animals were divided into invertebrates, fishes, amphibians, reptiles, birds, other mammals, rodents and primates. We found that the spectrum distributions of each XY-type 8-mers of genome sequences gradually evolved from unimodal to tri-modal distribution according to the species evolution levels from primitive to more complex. We inferred that the spectrum evolving of the other 15 XY-type 8-mers is caused by the separation of the 3 CG-type 8-mers and the spectrum features of these subsets reflect the information about the composition and evolution of genome sequences in more details.

Position preference for spectrum distributions of XY-type 8-mers

By observing the positions of spectrum distributions of XY-type 8-mers in all analyzed animal genomes, we found that the spectrum positions for each XY-type 8-mers are different. To verify these differences, we calculated the average frequency of each XY-type 8-mers of 889 animal and 59 green algae genome sequences. We added green algae for the following reasons. We want to more clearly show the changing process of CG and TA independent selection modes that relate to the evolutionary process of organisms from simple eukaryotes to complex eukaryotes. Because green algae have higher TA independent selection intensity and lower CG independent selection intensity, on the contrary, mammals have higher CG independent selection intensity and lower TA independent selection intensity. Considering that the number of XY2 8-mers is much smaller than that of XY1 8-mers (**Method**), we only discussed the average

frequency of XY1 and XY0 8-mers. The average frequency of XY1 and XY0 8-mers of four representative species *Raphidocelis subcapitata* (Green Algae), *Aethina tumida* (Invertebrate), *Mauremys reevesii* (Reptile) and *Homo sapiens* (Primate) were shown in Fig. 2.

In the analyzed animals and green algae, we found that the average frequency of 16 XY-type 8-mers varied within and among different genome sequences. The average frequency of XY0 8-mers is always opposite to that of XY1 8-mers, due to non-random sampling. In green algae and invertebrates, the 16 average values of XY1 8-mers are more dispersed relative to the central frequency. According to the evolutionary order of green algae, invertebrates, other vertebrates and mammals, except for CG1 subset, the average values of the other XY1 8-mers gradually approach the central frequency. The average values of CG1 8-mers showed obvious regularity in different species. According to the evolutionary order, the average values of CG1 8-mers change from the high frequency end and far from the central frequency (green algae) to the low frequency end and far from the central frequency (mammals). This phenomenon revealed that CG1 8-mers are sensitive to species evolution. Conversely, the average values of TA1 8-mers change from the low frequency end and far from the central frequency to the high frequency end and far from the central frequency, and from invertebrates onwards, the average values again gradually approach the central frequency. The results showed that during the evolution process of eukaryotes from unicellular (green algae) to multicellular (invertebrate), the usage frequency of CG1 and TA1 8-mers maintains an opposite evolutionary trend. This relation reflects that there is obvious mutual inhibition relationship in green algae and invertebrates. From other vertebrates to mammals, the average frequency of TA1 8-mers is close to the central frequency, but the average frequency of CG1 8-mers keep increasing. It indicated that the mutual inhibition relationship gradually disappears. We guess that the evolution mechanism of higher animals has improved. In addition to CG1 and TA1 8-mers, the other XY1 8-mers also showed a certain degree of co-evolution regularity in genome sequences. In green algae and invertebrates, CC1, GG1 and GC1 8-mers have similar change trends to CG1 8-mers, but this trend is weakened in vertebrates and mammals. AA1, TT1 and AT1 8-mers have similar change trends to TA1 8-mers. The average frequencies of the other XY1 8-mers have no obvious relationship with the evolution of species. In four pairs of reverse complementary XY1 8-mers (AC1/GT1, AG1/CT1, CA1/TG1, GA1/TC1) and two pairs of forward and reverse complementary XY1 8-mers (AA1/TT1, CC1/GG1), we found that the average frequencies of each pair of XY1 8-mers are the same. For example, the average frequency of AC1 8-mers is the same as that of GT1 8-mers. This

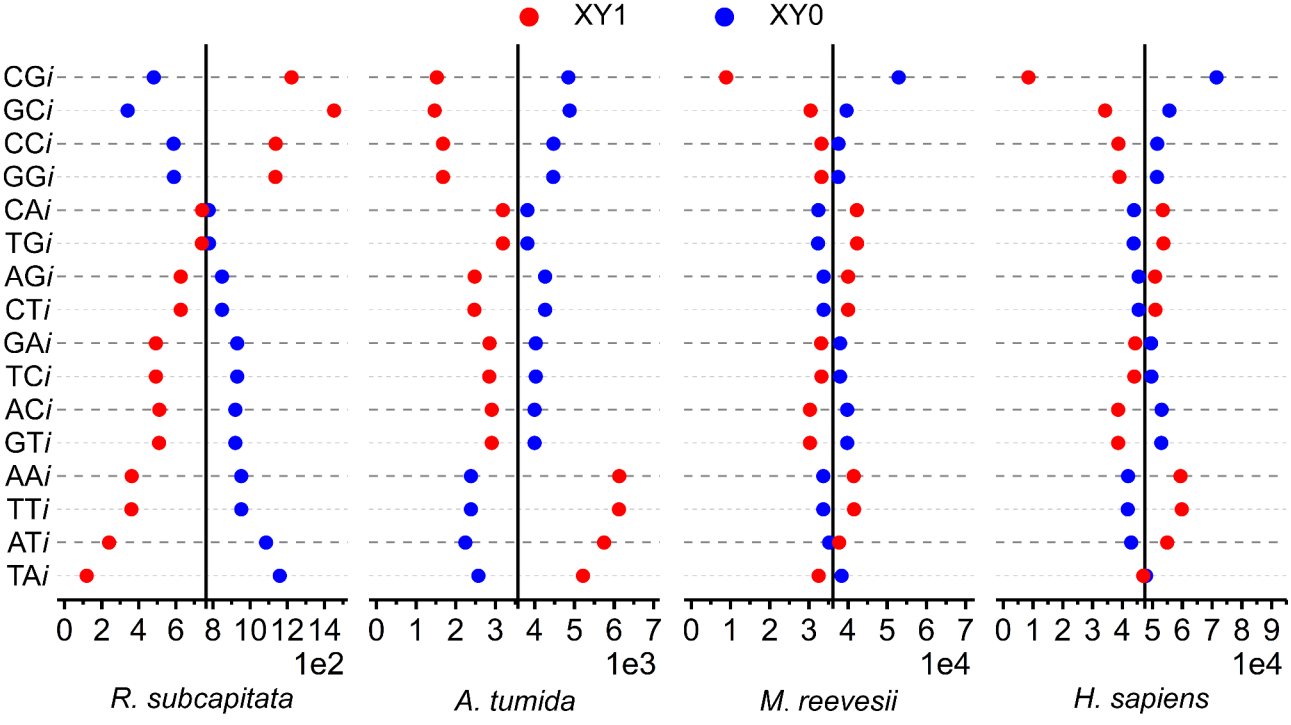

Fig. 2 The position relationship between the average frequency of 16 XY-type 8-mers and the center frequency. X-axis represents the frequency values. Red represents the average frequency of the XY1 8-mers and blue represents the average frequency of the XY0 8-mers. The black vertical line represents the central frequency of total 8-mers

conclusion verified the correctness of Chargaff's second parity rule at the length of 8 nucleotides. Chargaff's second parity rule [45–47] states that the frequency of nucleotides and oligonucleotides with reverse complementary pairing structure is statistically the same in a sufficiently long single-stranded nucleotide sequence, while the frequency of nucleotides and oligonucleotides with forward complementary pairing structure is different, which is called the phenomenon of strand symmetry.

In summary, during the evolutionary process of animals from primitive to more complex, different XY*i* 8-mers have varying effects on the evolution of genome sequences. TA1 8-mers are constrained by CG1 8-mers, and a mutual inhibition relationship is formed between them. The other XY1 8-mers showed varying degrees of co-evolution relationship under the constraint of CG1 8-mers.

Spectrum distribution features of XY *i* 8-mer subsets

By analyzing the spectrum distributions of XY*i* 8-mers of 889 animal genome sequences, we found that the average frequency of each XY*i* 8-mers varies within a species and among different species. This indicated that there are differences for each XY*i* 8-mers in response to the evolution and composition of genome sequences, which reflects both the commonality and the specificity of species evolution. In addition to the CG*i* 8-mers, we found that the spectrum distributions of the other XY*i* 8-mers are similar to that of the total 8-mers that gradually evolve from unimodal to tri-modal distribution according to the evolutionary order of animals (see Fig. 3 first row). We thought that the multimodal distributions of the other XY*i* 8-mers must include more information about composition and evolution of genome sequences. To deeply analyze the spectrum composition features of these XY*i* 8-mers, we further classified them. According to whether each XY*i* subset contains MN (M, N = A, C, G, T) dinucleotides, they were divided into XY*i*_MN*j* subsets (*i*, *j* = 0, 1), and the spectrum distribution of each 8-mer subset was given. Here, XY1 and XY2 were merged into one subset, called XY1 subset. We selected four species as representatives from animals with different evolution levels, they are *Aphidius gifuensis* (Invertebrate), *Cyprinus carpio* (Fish), *Naja naja* (Reptile) and *Homo sapiens* (Primate). The spectrum distributions of total 8-mers and some XY1_CG*j* 8-mers of genome sequences in these four species are shown in Fig. 3. Their spectra represented the distribution patterns from unimodal to tri-modal. The spectrum distributions of all XY1_CG*j* 8-mer subsets are shown in Supplementary Figures S1, S2, S3 and S4.

We analyzed the spectrum distributions of XY*i*_MN*j* 8-mers in all animal genome sequences. When the spectrum distributions of total 8-mers of genome sequences transition from unimodal to multimodal, we found that, in addition to the spectrum distributions of XY*i*_CG*j* 8-mers, the spectrum distributions of the other XY*i*_MN*j* 8-mers also transition from unimodal to multimodal. This indicated that this classification method could not divide the spectrum distributions of XY*i* 8-mers into two independent distributions. Only the spectrum distributions of XY*i*_CG*j* 8-mers could divided the spectrum distributions of XY*i* 8-mers into two independent distributions, and these two distributions showed obvious separation phenomenon (Fig. 3). It indicated that 8-mers containing CG dinucleotides occupy a core position in composition and evolution of genome sequences. We thought that XY*i*_CG1 and XY*i* _CG0 8-mer subsets belong to two composition units with different biological features, which can reflect the deeper composition and evolution features of genome sequences.

Spectrum separability of XY1_CG1 8-mer subsets

Due to the constraint of non-random sampling, there is a correlation between the spectrum features of XY1 and XY0 8-mers. Here, we only discussed the classification and related properties of XY1 8-mer subsets. The spectrum separability value δ_{XY1_CG1} of each XY1_CG1 8-mer subset in different species were calculated according to Eq. (2), which reflects the relative deviation degree of the average frequency of XY1_CG1 8-mer subsets from the central frequency of total 8-mers. The spectrum separability distributions of 16 kinds of XY1_CG1 8-mers in 8 groups of animal genome sequences were shown in Fig. 4.

According to the overall spectrum distribution trend of XY1_CG1 8-mers, as the evolution levels of species increased, the average values of the spectrum separability of each XY1_CG1 8-mers gradually increased (Fig. 4A). This indicated that the spectrum separability of XY1_CG1 8-mers correlates positively with the evolution of genome sequences, which characterizes the commonality of species evolution. Analyzing the spectrum separability distributions of 16 kinds of XY1_CG1 8-mers, we found that the distribution features of XY1_CG1 8-mer subsets are almost the same as those of X'Y'1_CG1 8-mer subsets, where X'Y' and XY dinucleotides are reverse complementary or forward and reverse complementary in 8 groups of animals, such as the average value and variance of the separability distributions (Fig. 4B). And the distribution features of X'Y'1_CG1 8-mer subsets with forward complementary to XY1_CG1 8-mer subsets are obviously different. This indicates that the spectrum separability features of the 8-mer subsets also follow the theory of Chargaff's second parity rule. Specifically, the separability values of four pairs of 8-mer subsets with reverse complementary δ_{AC1_CG1} and δ_{GT1_CG1}, δ_{AG1_CG1} and δ_{CT1_CG1}, δ_{CA1_CG1} and δ_{TG1_CG1}, δ_{GA1_CG1} and δ_{TC1_CG1}, and the separability values of two pairs of

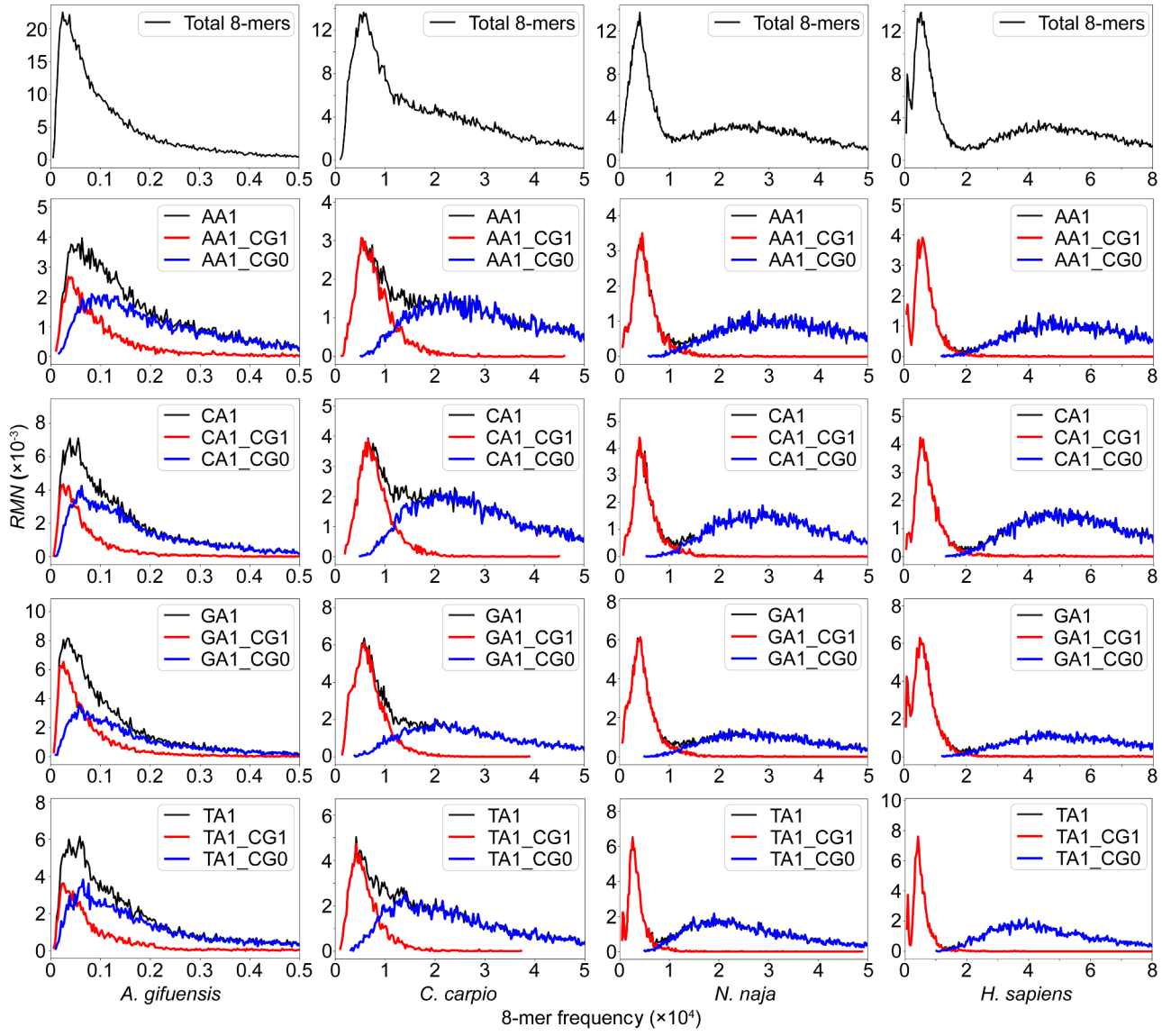

Fig. 3 Spectrum distributions of total 8-mers and part of XY1_CG*j* 8-mers for the four representative species genome sequences. The first row represents the spectrum distributions of total 8-mers. In the next four rows, the black curve represents the spectrum distributions of XY1 8-mers, the red curve represents the spectrum distributions of XY1_GC1 8-mers, and the blue curve represents the spectrum distributions of XY1_GC0 8-mers

8-mer subsets with forward and reverse complementary δ_{AA1_CG1} and δ_{TT1_CG1}, δ_{CC1_CG1} and δ_{GG1_CG1} are basically the same. And the separability values of two pairs of 8-mer subsets with forward complementary δ_{AT1_CG1} and δ_{TA1_CG1}, δ_{CG1_CG1} and δ_{GC1_CG1} are different.

We found that the spectrum separability of CG1_CG1 8-mers in vertebrates is significantly higher than that of the other XY1_CG1 8-mers, followed by TA1_CG1 8-mers. This indicates that these two 8-mer subsets are more sensitive motif sets for characterizing the evolution of vertebrate genomes. In invertebrates, the separability values of these two 8-mer subsets fall within the separability range of the other XY1_CG1 8-mer subsets. Compared with vertebrates, this distribution feature highlights the fundamental difference between vertebrates and invertebrates. For the other 14 kinds of XY1_CG1 8-mer subsets, the separability values vary within each species group and among different species groups. For example, the separability value of GC1_CG1 8-mer subset is relatively high in invertebrates, while it is relatively low in birds and other mammals. In the XY1 8-mer subset containing CG dinucleotides, the above conclusions indicated that CG1 and TA1 8-mer subsets mainly reflect the commonality of the evolution of genome sequences, while the other 14 kinds of XY1 8-mer subsets mainly reflect the specificity of the composition of genome sequences.

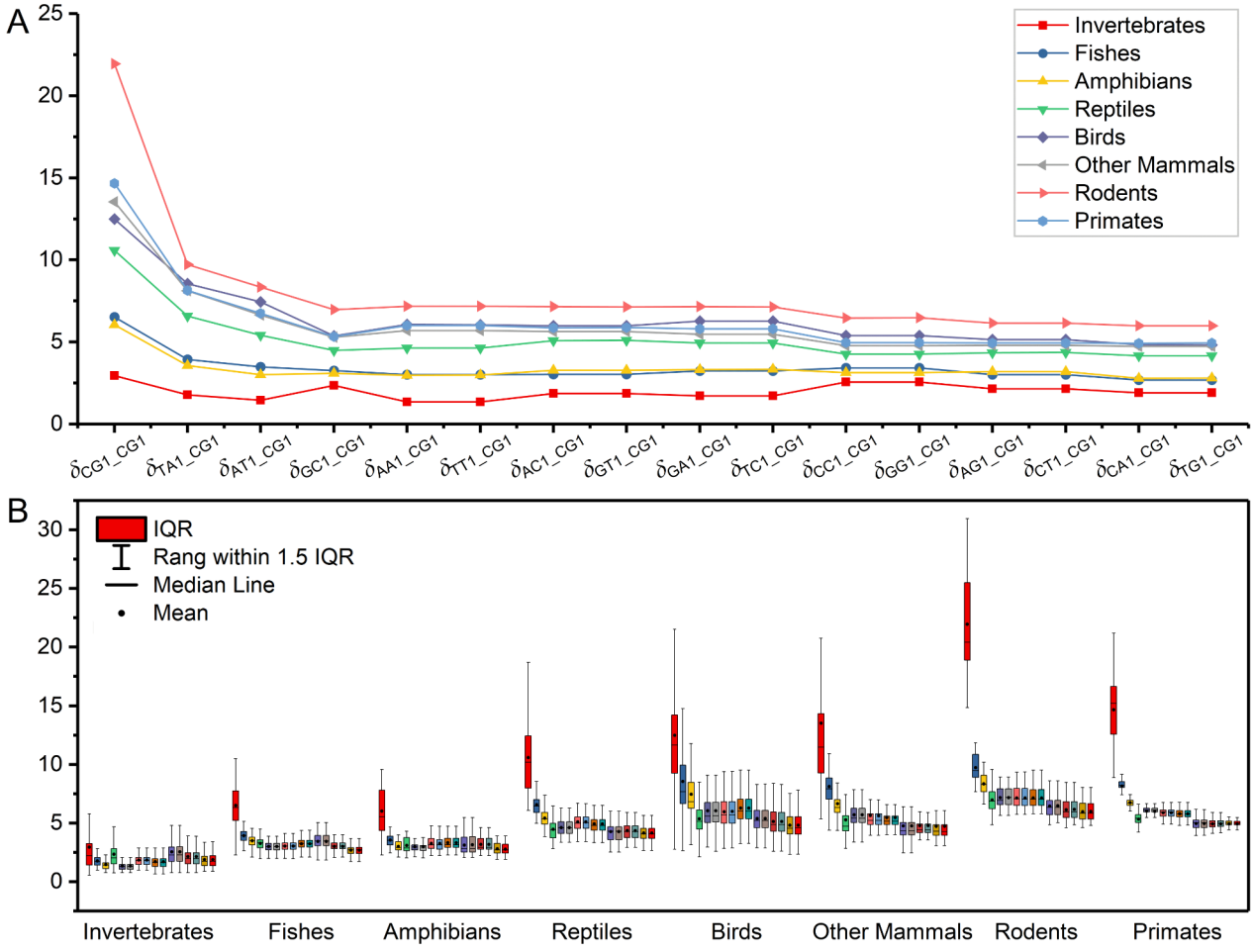

Fig. 4 Spectrum separability distributions of XY1_CG1 8-mer subsets of genome sequences in 8 animal groups. **A**. Average values of the spectrum separability of 16 kinds of XY1_CG1 8-mers. **B**. Spectrum separability range of 16 kinds of XY1_CG1 8-mer subsets in each animal group. The individual colored boxes within a group corresponds to the x-axis parameters of Fig. 4A from left to right. The y-axis represents the spectrum separability values

We thought that the separability values of the 16 kinds of XY1 subsets containing CG dinucleotides can be used as a feature set to characterize the evolution state of genome sequences and the feature set can be used to explore the evolution relationships of species. Considering the constraint of Chargaff's second parity rule, we selected four 8-mer subsets from four reverse complementary subset pairs, two 8-mer subsets from two forward and reverse complementary subset pairs and four 8-mer subsets from four forward complementary subsets. We used the separability of these 10 subsets as the feature value. Finally, we obtained 10 separability values as the feature set to characterize the evolution state of a genome sequence. This feature set includes δ_{AT1_CG1}, δ_{AA1_CG1}, δ_{GA1_CG1}, δ_{TA1_CG1}, δ_{AC1_CG1}, δ_{CA1_CG1}, δ_{AG1_CG1}, δ_{CC1_CG1}, δ_{GC1_CG1} and δ_{CG1_CG1}. Based on whole genome sequences, constructing the evolution relationships in animals, especially in higher animals, has not yet achieved satisfactory results [44]. Additionally, constructing the evolution relationship across large-scale and cross-species is still a current challenge. Based on the k-mer spectra distribution rule of genome sequences, we obtained a feature set to characterize the evolution state of genome sequences. We believed that this feature set can efficiently distinguish the species differences in large-scale and cross-species.

Difference analysis of species genome sequences

The main difficulty in constructing evolution relationships at large-scale and cross-species is the feature set selection with high quality. To test the rationality and quality of the feature set we gave, we analyzed the differences between various animal genome sequences. By examining the spectrum separability distributions of XY1_CG1 8-mer subsets in 8 groups of animal genome sequences, we found that the separability distribution range in birds is broader and covers the distribution ranges of fishes, amphibians and reptiles, and nearly covers mammals (Fig. 4B). This indicated that the composition and evolution of bird genomes are more diverse. If

we can effectively distinguish birds from other vertebrate and mammal genomes, it indicates that the quality of our feature set is excellent. For this purpose, we divided 8 groups of animals into four large-scale categories, namely invertebrates, other vertebrates (fish, amphibians and reptiles), birds and mammals (other mammals, rodents and primates). We used machine learning algorithms to test the classification quality of birds versus other vertebrates and birds versus mammals. To avoid the randomness of the test results, we used 14 machine learning algorithms for binary classification analysis in our classification test. The results are shown in Fig. 5, with the 14 machine learning algorithms detailed in the Method, and the results of each algorithm presented in the Supplementary Table S2.

Overall, the classification results are excellent. Between birds and other vertebrates, the highest *Acc* value is 98.70% in SVM algorithm and the lowest *Acc* value is 87.06% in SGD algorithm. Between birds and mammals, the highest *Acc* value is 96.50% in QDA algorithm and the lowest *Acc* value is 83.88% in DT algorithm. These results indicated that, although we only provided 10 feature values to characterize the evolution state of genome sequences, the feature set could effectively distinguish differences among different species groups in large-scale animal classification, and had a high classification ability. According to the composition and evolution rules of genome sequences, we obtained the feature set that can characterize the evolution state of genome sequences. We thought that our feature set is objective, and the spectrum separability values of XY1 8-mers containing CG dinucleotides reflect the core information of whole genome sequences.

In order to evaluate the reasonability of the 10 features in 14 machine learning algorithms and validate the importance for each feature of the feature set, we conducted a sensitivity analysis. To ensure the independence and importance of the selected features, we utilized the F-score method to reorder their features and utilized the IFS strategy for redundant feature detection, the results are shown in Fig. 6A and B.

We order the features according to its importance in classification algorithms, if the feature order is 1-2-3-4-5-6-7-8-9-10 between birds and other vertebrates, but the feature order is 10-3-1-8-6-9-7-4-5-2 between birds and mammals. Results show that when the number of selected features is increased to 5, the classification accuracy (*Acc*) tends to be stable for the two classifications. That means the former 5 features are important. But we found that the 5 important features are different. For example, the most important feature is feature 1 and the most unimportant feature is feature 10 between birds and other vertebrates, but the most important feature is feature 10 and the most unimportant feature is feature 2 between birds and mammals. That means all of 10 features, taken as a whole, are indispensable. It indicates that these features play different roles in distinguishing different species. Therefore, we believe that the 10 features currently selected are all important features and there is no significant redundancy.

In the experiment of large-scale species classification among mammals, other vertebrates and birds, we

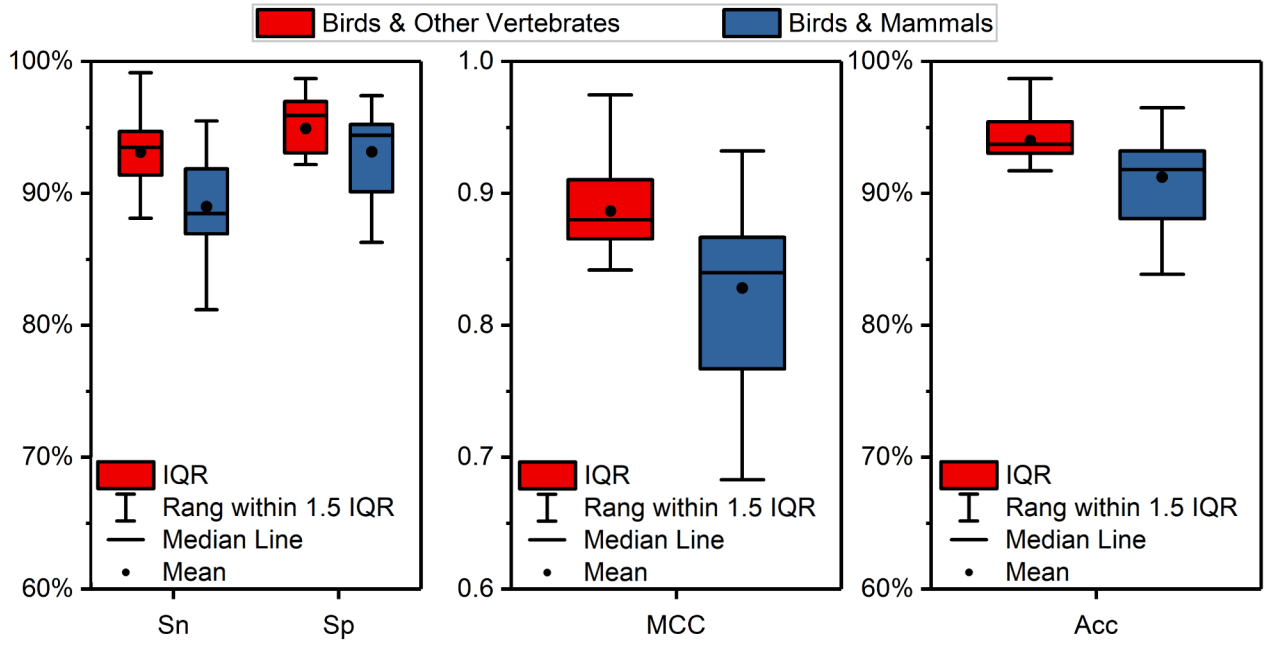

Fig. 5 Binary classification test results of 14 machine learning algorithms

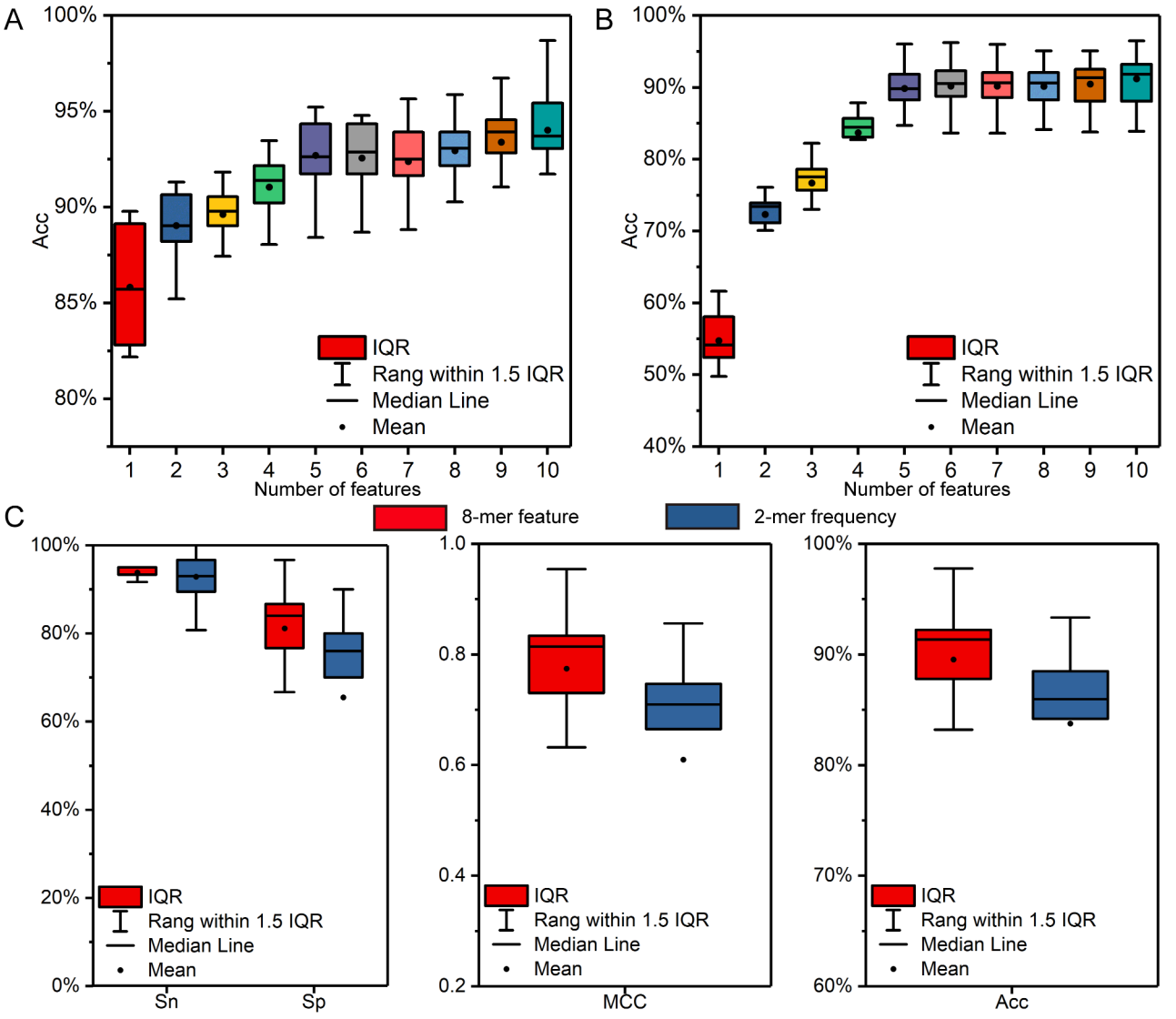

Fig. 6 Binary classification test results of 14 machine learning algorithms based on different feature numbers and different features. **A**. Binary classification test results between birds and other vertebrates based on different feature numbers. **B**. Binary classification test results between birds and mammals based on different feature numbers. **C**. Binary classification test results between primates and rodents based on different features

achieved satisfying classification results using the 10 features. We considered that the 10 features also apply to order-level species classification. Here we did the classification on order-level between primates and rodents, the results are shown in Fig. 6C. Results showed that the Interquartile Range (IQR) for *Acc* value is between 87.78%~92.22% and the IQR for *MCC* value is between 0.73 ~ 0.83 (Fig. 6C, red boxes) in 14 machine learning algorithms, the highest *Acc* value is 97.78% in QDA algorithm. That means the 10 features obtained by the spectra of 8-mers containing CG dinucleotide can apply to order-level species classification.

Some researchers thought that 2-mers were suitable features. Here we did a classification comparison between primates and rodents by 2-mer and 8-mer information. Results showed that all of the four scores *Sn*, *Sp*, *Acc* and *MCC* are poor by 2-mer frequencies (Fig. 6C, blue boxes). The IQR for *Acc* value is between 84.17%~88.47% and the IQR for *MCC* value is between 0.67 ~ 0.75 in 14 machine learning algorithms. It showed that the accuracy and the robustness of classification results obtained by 8-mer information are better than that obtained by 2-mer information. It indicated that the base composition of genome sequences is long range correlation. Long DNA fragments can reflect more accurate information of the composition and evolution of genome sequences.

Conclusion and discussion

Based on 889 animal genomes, we analyzed the spectrum distribution rules of various 8-mer subsets of genome sequences, explored the relationship between the spectrum distribution features of 8-mer subsets and the composition and evolution of species genome sequences, and provided a feature set that characterizes the evolution state of genome sequences. Using the bootstrap sampling method, total 8-mers were divided into 16 kinds of XY2, XY1 and XY0 8-mer subsets, and the spectrum distributions of XY-type 8-mers were given. Previous studies pointed out that only the spectra of CG2, CG1 and CG0 8-mer subsets form independent unimodal distributions respectively, while the spectrum distributions of the other 15 XY-type 8-mers did not have this feature. According to the evolutionary order of animals from primitive to more complex, we found that the spectrum distributions of the other 15 XY-type 8-mers gradually transition from unimodal to tri-modal. The differences between the average frequency of each XY-type 8-mers and the central frequency vary within a species and among different species. The most obvious difference appears in CG1 and TA1 8-mer subsets, and this difference is closely related to the evolution levels of animals. In the other XY-type 8-mer subsets, the differences still have a certain relationship with species evolution. The results indicated that CG1 and TA1 8-mer subsets are sensitive motif sets in the evolution of genome sequences, and the other XY-type 8-mer subsets show the co-evolution relationship under the constraint of CG1 and TA1 8-mer subsets. To explore the contribution of all 16 XY-type 8-mer subsets to the composition and evolution of genome sequences, each XY-type 8-mer subsets were divided into two 8-mer subsets according to whether they contained CG dinucleotides. We found that these two 8-mer subsets form two independent distributions, and the spectrum separability of 16 XY-type 8-mer subsets containing CG dinucleotides are different within and across species. This indicated that the multimodal phenomenon of spectrum distributions of XY-type 8-mer subsets is caused by the separation of CG-type 8-mer subsets. Analyzing the spectrum separability of the 16 kinds of 8-mer subsets in 8 groups of animals, we found that the spectrum separability of CG1 8-mers containing CG dinucleotides correlates obviously and positively with the evolution of species genomes, followed by that of TA1 8-mers containing CG dinucleotides, and the spectrum separability of the other 14 kinds of XY1 8-mers containing CG dinucleotides also have a certain positive correlation with the evolution of species genomes. Further analysis showed that the spectrum separability of the other 14 kinds of XY1 8-mers containing CG dinucleotides mainly reflected the specificity of the composition of species genome sequences. The above results indicated that the spectrum separability of 16 kinds of XY1 8-mers containing CG dinucleotides not only characterize the evolution information of genomes but also show the composition information of genome sequences. We thought that these 16 separability values can be used as a feature set to characterize the evolution state of genome sequences. Considering the constraint of Chargaff's second parity rule, 6 separability values with similarity were filtered out, and finally, 10 separability values were obtained as the feature set to characterize the evolution state of genome sequences. To verify the accuracy of the feature set, vertebrates were divided into birds, other vertebrates and mammals, and 14 common classification algorithms were used for binary classification analysis. The results showed that, between birds and other vertebrates, the highest classification accuracy *Acc* value is 98.70% (SVM), and the lowest value is 87.06% (SGD). Between birds and mammals, the highest classification accuracy *Acc* value is 96.50% (QDA), and the lowest value is 83.88% (DT). The results showed that a high-quality classification effect was obtained with only 10 feature values, which indicated that the feature set we provided can objectively characterize the evolution state of genome sequences.

Compared with other vertebrates, the classification accuracy between birds and mammals is relatively lower, which is inconsistent with our known knowledge. We thought that the lower classification accuracy is caused by the change of evolution mechanism of rodents and primates in mammals. Previous studies [41, 44] have shown that there are two evolution modes in species genome sequences, CG and TA, and there is a mutual inhibition relationship between the two evolution modes. However, the TA independent selection mode gradually disappeared in rodent and primate genome sequences. Under the original evolution mechanism, these two groups of animals adopted a more advanced evolution mode. Therefore, using the separability values to construct the feature set is insufficient to characterize the evolution state of rodent and primate genome sequences. Explore the advanced evolution mechanism of mammal genomes is an interesting topic for future research.

In subsequent research, we will further consider the following issues: (1) In this study, we only considered the effect of the spectrum separability of XY1 8-mers containing CG dinucleotides. The effects of the spectrum separability of XY1 8-mers not containing CG dinucleotides need to be considered further. (2) Considering only the separability feature of XY1 8-mer subsets is insufficient. Since using the average value of a spectrum alone does not fully characterize the spectrum distribution feature, the variance and non-normality features of a spectrum distribution should also be considered. In subsequent work, we will explore the relationship between the variance and non-normality of the

spectrum distribution and the evolution state of genome sequences. By combining the average value feature of the spectrum distribution, we will construct a more comprehensive feature set to characterize the evolution state of genome sequences. (3) This study only analyzed the genome sequences of a part species within a larger taxonomic category, and did not cover all species. In the next step, we will analyze the genome sequences of plants, fungi and prokaryotes to explore whether the feature set we provided has species universality. Our ultimate goal is to reveal the composition and evolution rules of genome sequences, and provide novel ideas and methods for constructing evolution relationships at large-scale and cross-species level.

Materials and methods

Dataset

The whole genome sequences and annotated information of all species involved in this study were obtained from NCBI (https://www.ncbi.nlm.nih.gov/). The selected species number is 948. This includes 889 animals and 59 green algae. The species genome taxonomy was shown in Table 1, see Supplementary Table S1 for detailed information.

K-mer spectrum of the genome sequence

The genome sequence can be viewed as a linear string of *N* bases in length consisting of A, C, G and T. K-mer refers to a substring consisting of *k* consecutive nucleotides in the sequence. For a given genome sequence, *k* bp is used as the sliding window and 1 bp is used as the step to calculate the occurrence frequency of all k-mers. If the number of k-mers that occurs *i* times is N_i, the relative motif number (*RMN*) is defined as:

$$RMN = \frac{N_i}{4^k} \tag{1}$$

The distribution of *RMN* value with k-mers frequency is called the k-mer spectrum of the genome sequence.

XY dinucleotide classification method

For $k=8$, we classified the 8-mer set into different subsets according to the compositional features of 8-mers. The 8-mers containing zero XY (X, Y = A, T, C, G) dinucleotide were called the XY0 subset, those containing one XY dinucleotide were called the XY1 subset, and those containing two or more XY dinucleotides were called the XY2 subset. For example, for CG-type 8-mers, the 8-mer CGTACGAT have 2 CG dinucleotides, it belongs to CG2 8-mer subset. The 8-mer CCTACGAT have 1 CG dinucleotides, it belongs to CG1 8-mer subset. The total number of 8-mers is 4^8=65,536. Theoretically, when X ≠ Y, the numbers of XY0 8-mer subsets is 40,545, of XY1 8-mer subsets is 21,468, and of XY2 8-mer subsets is 3523. When X = Y, the numbers of XY0 8-mer subsets is 44,631, of XY1 8-mer subsets is 14,931, and of XY2 8-mer subsets is 5974. This is called the XY dinucleotide classification method. Total 8-mers are divided into three subsets, XY2, XY1 and XY0, which can be classified in 16 ways.

Table 1 Number of species genome sequences

Species	Number	Species	Number
Invertebrates	235	Birds	232
Fishes	125	Other mammals	106
Amphibians	29	Rodents	29
Reptiles	74	Primates	59
Green algae	59		

Spectrum separability of 8-mer subset

For a given 8-mer spectrum, the average value of the spectrum distribution is used to characterize its distribution characteristic. In order to eliminate the influence of different genome sizes and show the relative position difference of 8-mer spectra in different 8-mer subsets, the separability value (δ_i) was defined:

$$\delta_i = \frac{\bar{x}}{\bar{x}_i} \tag{2}$$

Where $\bar{x}$ is the average frequency of total 8-mers, called the center frequency. $\bar{x}_i$ is the average frequency of 8-mers of subset *i*. δ_i represents the separability for the distribution position of 8-mer spectrum of subset *i* relative to the center frequency. If $\delta_i > 1$, it indicates that 8-mer spectrum of subset *i* is located at the low frequency end and is away from the center frequency. If $\delta_i = 1$, it indicates that the location of 8-mer spectrum of subset *i* is the same as that of the center frequency.

In this definition, the separability value is independent of genome size and the absolute position of the subset spectrum. Additionally, this parameter can compare not only the distribution difference of different 8-mer subsets within a genome sequence but also the distribution difference of 8-mer subsets among genome sequences.

Machine learning algorithms and classification performance evaluation

This study used 14 machine learning algorithms provided by a comprehensive and automated machine learning platform *iLearnPlus* [48], including multilayer perceptron (MLP), *K*-nearest neighbors (KNN), Adaptive boosting (AdaBoost), Gradient boosting decision tree (GBDT), Light gradient boosting machine (LightBGM), Random forest (RF), Extreme gradient boosting (XGBoost), Logistic regression (LR), Decision tree (DT), Stochastic gradient descent (SGD), Quadratic discriminant analysis

(QDA), Bagging (Bagging), Linear discriminant analysis (LDA) and Support vector machine (SVM). Use the 10-fold cross-validation method to evaluate the performance of the machine learning model. And we use 8 commonly indicators to evaluate and compare the classification performance of models, including sensitivity (*Sn*), specificity (*Sp*), precision (*Pre*), accuracy (*Acc*), Matthews correlation coefficient (*MCC*), *F*1 score (*F*1), the area under ROC curve (AUROC) and the area under the PRC curve (AUPRC) [49–52], which are defined as:

$$Sn = \frac{TP}{TP + FN} \tag{3}$$

$$Sp = \frac{TN}{TN + FP} \tag{4}$$

$$Pre = \frac{TP}{TP + FP} \tag{5}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$MCC = \frac{TP \times TN + FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{7}$$

$$F1 = 2 \times \frac{Pre \times Sn}{Pre + Sn} \tag{8}$$

Where *TP*, *FP*, *TN* and *FN* represent the numbers of true positives, false positives, true negatives and false negatives, respectively. The AUROC and AUCPRC values, which range between 0 and 1. The closer these values are to 1, the better the classification performance of the model.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10786-1.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Acknowledgements
We are very grateful for the computational resources provided by the Theoretical Biophysics Laboratory of Inner Mongolia University.

Author contributions
L.H. and L.X.L. conceived and designed the study and critically revised the manuscript. L.X.L. analyzed the data and drafted the manuscript. Y.Z.H. helped in the study design. W.L. helped with data collation and prepared figures S1–S4. All authors contributed to the article and approved the submitted version.

Funding
This work was supported by the National Natural Science Foundation of China (31860304). The funding agency played no role in research design, data collection, analysis and interpretation, and manuscript writing.

Data availability
All genome sequences and the corresponding annotation information were obtained from NCBI (https://www.ncbi.nlm.nih.gov/). See Supplementary Table S1 for detailed information.

Declarations

Ethics approval and consent to participate
Not applicable to this study.

Consent for publication
Not applicable to this study.

Competing interests
The authors declare no competing interests.

Conflict of interest
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 January 2024 / Accepted: 9 September 2024
Published online: 12 September 2024

References

1. Brendel V, Beckmann JS, Trifonov EN. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. J Biomol Struct Dyn. 1986;4:11–21.
2. Audic S, Claverie JM. Detection of eukaryotic promoters using Markov transition matrices. Comp Chem. 1997;21:223–7.
3. Bhukya R, Kumari A, Amilpur S, Dasari CM. PPred-PCKSM: a multi-layer predictor for identifying promoter and its variants using position based features. Comput Biol Chem. 2022;97:107623.
4. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 2011;21:2167–80.
5. Mohamed Hashim EK, Abdullah R. Rare k-mer DNA: identification of sequence motifs and prediction of CpG island and promoter. J Theor Biol. 2015;387:88–100.
6. Zhao X, Pei Z, Liu J, Qin S, Cai L. Prediction of nucleosome DNA formation potential and nucleosome positioning using increment of diversity combined with quadratic discriminant analysis. Chromosome Res. 2010;18:777–85.
7. Kirk JM, Kim SO, Inoue K, Smola MJ, Lee DM, Schertzer MD, Wooten JS, Baker AR, Sprague D, Collins DW, et al. Functional classification of long non-coding RNAs by k-mer content. Nat Genet. 2018;50:1474–82.
8. Gudenas BL, Wang L. Prediction of LncRNA subcellular localization with deep learning from sequence features. Sci Rep. 2018;8:16385.
9. Ahmad A, Lin H, Shatabda S. Locate-R: subcellular localization of long non-coding RNAs using nucleotide compositions. Genomics. 2020;112:2583–9.
10. Su Z-D, Huang Y, Zhang Z-Y, Zhao Y-W, Wang D, Chen W, Chou K-C, Lin H. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. 2018;34:4196–204.
11. Cheng S, Zhang L, Tan J, Gong W, Li C, Zhang X. DM-RPIs: Predicting ncRNA-protein interactions using stacked ensembling strategy. Comput Biol Chem. 2019;83:107088.
12. Asim MN, Malik MI, Zehe C, Trygg J, Dengel A, Ahmed S. MirLocPredictor: a ConvNet-Based Multi-label MicroRNA subcellular localization predictor by incorporating k-Mer positional information. Genes. 2020;11:1475.
13. Kirk JM, Sprague D, Calabrese JM. Classification of long noncoding RNAs by k-mer content. Methods Mol Biol. 2021;2254:41–60.

14. Montaseri S, Zare-Mirakabad F, Ganjtabesh M. Evaluating the quality of SHAPE data simulated by k-mers for RNA structure prediction. J Bioinform Comput Biol. 2017;15:1750023.
15. Miller C, Gurd J, Brass A. A RAPID algorithm for sequence database comparisons: application to the identification of vector contamination in the EMBL databases. Bioinformatics. 1999;15:111–21.
16. Liu Y, Schröder J, Schmidt B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinformatics. 2013;29:308–15.
17. Williams D, Trimble WL, Shilts M, Meyer F, Ochman H. Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. BMC Genom. 2013;14:537.
18. Audoux J, Philippe N, Chikhi R, Salson M, Gallopin M, Gabriel M, Le Coz J, Drouineau E, Commes T, Gautheret D. DE-kupl: exhaustive capture of biological variation in RNA-seq data through k-mer decomposition. Genome Biol. 2017;18:243.
19. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 1995;11:283–90.
20. Zhou F, Olman V, Xu Y. Barcodes for genomes and applications. BMC Bioinform. 2008;9:546.
21. Wei C, Wang G, Chen X, Huang H, Liu B, Xu Y, Li F. Identification and typing of human enterovirus: a genomic barcode approach. PLoS ONE. 2011;6:e26296.
22. Meher PK, Sahu TK, Rao AR. Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier. Gene. 2016;592:316–24.
23. Wang D, Xu J, Yu J. KGCAK: a K-mer based database for genome-wide phylogeny and complexity evaluation. Biol Direct. 2015;10:53.
24. Kafri A, Chor B, Horn D. Inter-chromosomal k-mer distances. BMC Genom. 2021;22:644.
25. Yang Y, Nephew K, Kim S. A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters. BMC Bioinform. 2012;13:S15.
26. Subramanian A, Schwartz R. Reference-free inference of tumor phylogenies from single-cell sequencing data. BMC Genom. 2015;16:S7.
27. Sauk M, Žilina O, Kurg A, Ustav EL, Peters M, Paluoja P, Roost AM, Teder H, Palta P, Brison N, et al. NIPTmer: rapid k-mer-based software package for detection of fetal aneuploidies. Sci Rep. 2018;8:5616.
28. Audemard EO, Gendron P, Feghaly A, Lavallée VP, Hébert J, Sauvageau G, Lemieux S. Targeted variant detection using unaligned RNA-Seq reads. Life Sci Alliance. 2019;2:e201900336.
29. Lee H, Shuaibi A, Bell JM, Pavlichin DS, Ji HP. Unique k-mer sequences for validating cancer-related substitution, insertion and deletion mutations. NAR Cancer. 2020;2:zcaa034.
30. Jaillard M, Palmieri M, van Belkum A, Mahé P. Interpreting k-mer-based signatures for antibiotic resistance prediction. Gigascience. 2020;9:giaa110.
31. Naghibzadeh M, Savari H, Savadi A, Saadati N, Mehrazin E. Developing an ultra-efficient microsatellite discoverer to find structural differences between SARS-CoV-1 and Covid-19. Inf Med Unlocked. 2020;19:100356.
32. Zhang Y, Wen J, Li X, Li G. Exploration of hosts and transmission traits for SARS-CoV-2 based on the k-mer natural vector. Infect Genet Evol. 2021;93:104933.
33. Sung I, Lee S, Pak M, Shin Y, Kim S. AutoCoV: tracking the early spread of COVID-19 in terms of the spatial and temporal patterns from embedding space by K-mer based deep learning. BMC Bioinform. 2022;23:149.
34. Cserháti M, Turóczy Z, Dudits D, Györgyey J. The rice word landscape—a detailed catalog of the rice motif content in the noncoding regions. OMICS. 2011;15:819–28.
35. Cserhati M. A tail of two pandas— whole genome k-mer signature analysis of the red panda (Ailurus fulgens) and the Giant panda (Ailuropoda melanoleuca). BMC Genomics. 2021;22:228.
36. Bonnici V, Franco G, Manca V. Spectral concepts in genome informational analysis. Theor Comput Sci. 2021;894:23–30.
37. Dubinkina VB, Ischenko DS, Ulyantsev VI, Tyakht AV, Alexeev DG. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. BMC Bioinform. 2016;17:38.
38. Huimin X, Bailin H. Aug. Visualization of K-tuple distribution in procaryote complete genomes and their randomized counterparts. In: Proceedings IEEE Computer Society Bioinformatics Conference: 16–16 2002. 2002;2002:31–42.
39. Chen YH, Nyeo SL, Yeh CY. Model for the distributions of k-mers in DNA sequences. Physi Rev E. 2005;72:011908.
40. Chor B, Horn D, Goldman N, Levy Y, Massingham T. Genomic DNA k-mer spectra: models and modalities. Genome Biol. 2009;10:R108.
41. Yang ZH, Li H, Jia Y, Zheng Y, Meng H, Bao T, Li XL, Luo LF. Intrinsic laws of k-mer spectra of genome sequences and evolution mechanism of genomes. BMC Evol Biol. 2020;20:157.
42. Jia Y, Li H, Wang J, Meng H, Yang Z. Spectrum structures and biological functions of 8-mers in the human genome. Genomics. 2019;111:483–91.
43. Zheng Y, Li H, Wang Y, Meng H, Zhang Q, Zhao X. Evolutionary mechanism and biological functions of 8-mers containing CG dinucleotide in yeast. Chromosome Res. 2017;25:173–89.
44. Li XL, Li H, Yang ZH, Wu Y, Zhang MC. Exploring objective feature sets in constructing the evolution relationship of animal genome sequences. BMC Genom. 2023;24:634.
45. Rudner R, Karkas JD, Chargaff E. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. Proc. Natl. Acad. Sci. U.S.A. 1968;60:921–922.
46. Prabhu VV. Symmetry observations in long nucleotide sequences. Nucleic Acids Res. 1993;21:2797–800.
47. Yamagishi MEB. Mathematical Grammar of Biology. Springer Cham; 2017.
48. Chen Z, Zhao P, Li C, Li F, Xiang D, Chen YZ, Akutsu T, Daly Roger J, Webb Geoffrey I, Zhao Q, et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. Nucleic Acids Res. 2021;49:e60.
49. Liu B, Fang L, Long R, Lan X, Chou KC. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics. 2016;32:362–9.
50. Liu B, Yang F, Huang DS, Chou KC. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics. 2018;34:33–40.
51. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. Nucleic Acids Res. 2019;47:e127.
52. Chen Z, Zhao P, Li F, Marquez-Lago TT, Leier A, Revote J, Zhu Y, Powell DR, Akutsu T, Webb GI, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Brief Bioinf. 2020;21:1047–57.

Publisher's note