scientific reports

OPEN

A novel method for identifying key genes in macroevolution based on deep learning with attention mechanism

Jiawei Mao[1,3], Yong Cao[1,3], Yan Zhang[2], Biaosheng Huang[1] & Youjie Zhao[1]✉

Macroevolution can be regarded as the result of evolutionary changes of synergistically acting genes. Unfortunately, the importance of these genes in macroevolution is difficult to assess and hence the identification of macroevolutionary key genes is a major challenge in evolutionary biology. In this study, we designed various word embedding libraries of natural language processing (NLP) considering the multiple mechanisms of evolutionary genomics. A novel method (IKGM) based on three types of attention mechanisms (domain attention, kmer attention and fused attention) were proposed to calculate the weights of different genes in macroevolution. Taking 34 species of diurnal butterflies and nocturnal moths in Lepidoptera as an example, we identified a few of key genes with high weights, which annotated to the functions of circadian rhythms, sensory organs, as well as behavioral habits etc. This study not only provides a novel method to identify the key genes of macroevolution at the genomic level, but also helps us to understand the microevolution mechanisms of diurnal butterflies and nocturnal moths in Lepidoptera.

Traces of macroevolution are widespread in nature, such as the aquatic to terrestrial evolution of vertebrates[1–3], the warm-bloodedness of mammals[4], the origin of bird wings[5], the origin of animal taste organs[6], etc. During the evolutionary process that occurs over long time scales, macroevolution was generally the result of synergistic action of complex molecular mechanisms at the genome level in addition to environmental factors[7,8]. However, among these possible molecular mechanisms, it is difficult to identify the key genes that drive the macroevolution of taxa. Therefore, it has become an important but unsolved problem in evolutionary biology to quantify the weight of key functional genes that cause macroevolution of ancestral species.

Previous studies[9–11] have shown that polyploidization (or whole-genome duplication, WGD) is the major driver for species formation and macroevolution in plants. However, WGD events are relatively rare and have only been found in a few animal taxa, such as euryhaline fishes[12] and Arachnida[13,14]. The molecular mechanisms involved in animal macroevolution are mainly as follows: (1) Contraction and expansion of gene families; for example, the emergence of epithelial tubular organs in vertebrates was associated with contraction and expansion of the Claudins gene family[15]; the evolution of functional plough nose organs in mammals was associated with contraction and expansion of the OR gene family[16]. (2) Selective evolution of genes in response to environmental stress; for example, the evolutionary rate of genes involved in energy metabolism, low-oxygen adaptation and skeletal development was significantly faster in ground tits that occur at high-altitudes compared to the closely related species[17]; studies of visual proteins in Lepidopteran insects showed that visual genes associated with brighter environments evolved faster and were under positive selection in insects from diurnal taxa[18]. (3) Structure variation in genomes, such as the evolution of butterfly wing mimicry duo to the inversions[19].

The above-mentioned studies on the molecular mechanisms of macroevolution are mainly based on traditional bioinformatics and statistical methods[20–22]. Considering the complex mechanisms of multiple key genes in the process of macroevolution, it is a challenge to systematically identify the key genes of macroevolution at the genome level. In order to obtain new knowledge from huge genomic data, machine learning (ML) has become a widely used and successful approach[23–25]. The correct performance of traditional ML algorithms relies heavily on data representations called features, and different features often need to be constructed for different task objectives. Moreover, deep learning (DL), a subfield of ML, can automatically learn features and patterns from data

[1]College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming 650224, China. [2]College of Mathematics and Physics, Southwest Forestry University, Kunming 650224, China. [3]These authors contributed equally: Jiawei Mao and Yong Cao. ✉email: bioala@swfu.edu.cn

without the need for manual feature engineering. DL has been applied to various aspects of biological research and has shown powerful capabilities[26–30], such as the analysis of gene expression data, and DNA and protein sequence data using natural language processing (NLP) related techniques with recurrent neural networks as the cornerstone. Despite the excellent results achieved by DL in several areas of bioinformatics, the inference process of DL is agnostic. In some bioinformatics scenarios, interpretable inference processes are often as important or even more important than excellent results. The attention mechanism (AM)[31] can compute different weights for different parts of the training sample during the training process. During the inference step without additional computation, these weights are generated and considered as the importance of that part to the model. The part with high weight was always focused on in the training process of model, and this can explain the inference process of DL. Previous studies showed that AM has been applied in the more and more fields of bioinformatics, such as the prediction of enhancers[32], and the prediction of protein interactions[33], etc. However, it is still a challenge to use AM to identify the different weights of key genes in the macroevolution of taxa.

In this paper, we develop IKGM, a method based on deep learning with attention mechanisms for identifying key genes in the macroevolution of biological taxa, which allows attaching different weights to genes to characterize the importance of these genes in macroevolutionary processes. Using 34 species of diurnal butterflies and nocturnal moths as an example, we used IKGM to mine the key genes with high weights and performed KEGG enrichment analysis based on these genes. These results should help us to understand the mechanisms of macroevolution in Lepidoptera.

Materials and methods

Data source

All the protein sequences of 34 Lepidoptera species were downloaded from *InsectBase*[34] and protein-coding genes were used as original samples. These species were labeled into two groups (nocturnal or diurnal) according to the diel behavior information in previous studies[18,35] (Supplementary-file2: Table S1). The proteins of these species were annotated to obtain the domain information by Pfam database (http://pfam.xfam.org/).

Pipeline of IKGM

In this study, the diel behavior information of 34 Lepidoptera species is used as the classification labels of experimental samples. The macroevolution phenomenon of nocturnal moths and diurnal butterflies is modeled as a classification problem of protein sequences. NLP was used to construct the word embedding libraries based on these sequences, and then AM is added to the classification network to compute the weight of different genes in the classification process. The pipeline of this paper mainly consists of four important parts (Fig. 1): Data pre-processing, Classification Model, Weights calculation, and Evaluation. In this study, three types of attention mechanisms (domain attention, kmer attention and fused attention) were developed to calculate the weights (weight 1, weight 2 and weight 3 in Fig. 1) of different genes. The details of each part are described in the following subsections. In addition, some important symbols in the method are shown in Table 1.

Data pre-processing

Data augmentation

Considering the prevalence of single nucleotide polymorphisms (SNP), insertion and deletion (InDel) and structural variants (SV) in different populations for each species, the original samples of 34 Lepidoptera species were expanded by simulating sequence variants to meet the sample requirements of the deep learning algorithm. In addition, we refer to the Neutral Theory of Molecular Evolution (Neutral Theory) proposed by Motoo Kimura, and attempt to make random small-scale mutations without selection bias while not changing the overall phenology of the species. We performed the simulation in several ways: (1) gene rearrangement, we divide

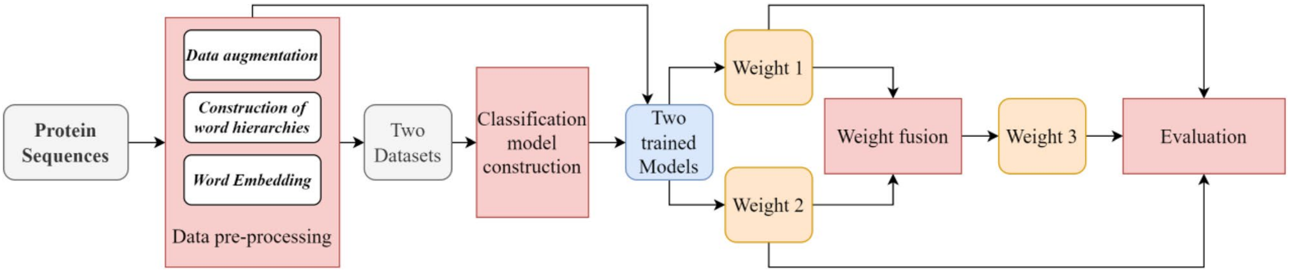

Figure 1. Pipeline of IKGM in this study. The red rectangles (Data pre-processing, Classification model construction, Weights fusion, and Evaluation) indicate the four key components of IKGM. Data pre-processing part consists of three main sub-modules, which use NLP technology to complete the pre-processing of raw protein sequences and feature construction. Classification model construction is the modeling of macroscopic evolutionary processes using deep learning and hierarchical attention mechanisms. Weight fusion is the process of fusing two different sets of gene weights together using a certain computational strategy (AM), where the gene weights represent the importance of the gene in the classification process, also known as the contribution to macroevolution. Evaluation is the process of annotation and KEGG enrichment for these genes with high weights.

Symbols	Description
HAN	Hierarchical Attention Network for genome classification
$\alpha_{i,t}$	Attention score of the t th word in the i th sentence in the hierarchical attention network
α_i	Attention score of the i th sentence in the hierarchical attention network
v	Hierarchy of words
$PAS_{v,s}$	Protein attention scores for species s when using word hierarchy v
$GAS_{v,T}$	Gene attention scores for taxon T when using word hierarchy v
DAS_v	Attention scores of genes causing differences between the two extant taxa when using word hierarchy v

Table 1. Description of the meaning of some important symbols.

the entire genome into multiple parts and perform random interchanges between the multiple parts without changing the gene order within each part; and (2) sequence mutations, i.e., random mutations of amino acids in a portion of the protein sequence, with an overall frequency of mutations of less than 1% of the genome. In addition, considering that the number of samples from the two taxa is uneven, the data are augmented separately for the two categories of genomes. In order to ensure that the mutated genomes are as diverse as possible, the number of amplifications is kept consistent for each original genome, the number of genomes before and after augmentation of each taxon and the number of each individual genome augmented by mutation (as shown in Supplementary-file2: Table S2).

Construction of word hierarchies

In natural language, multiple words are arranged in a certain word order to form a sentence with semantic meaning, while multiple sentences arranged in a certain order can form a text with rich semantic meaning. By analogy with natural language, the protein sequences of a single species can be considered as a text, while a single protein sequence can be considered as a sentence. However, word hierarchies are not clearly represented in proteins, and using an amino acid character as a word not only does not reflect the molecular mechanisms that may lead to macroevolution, but also leads to excessively long sentences. To address this problem and construct biologically meaningful word hierarchies to characterize the evolutionary mechanisms at different scales, we propose two methods for constructing word hierarchies:

1) Word hierarchy construction based on domain name ($v = 1$);
   Given that contractions and expansions of gene family often lead to quantitative differences in functional domains, and Pfam annotation of the original samples was performed. The annotated functional domain names are then used as word hierarchies to express contractions and expansions of gene families that may occur, as shown in Fig. 2a.
2) Word hierarchy construction based on variant kmer ($v = 2$);
   The selective evolution of genes is often reflected in the sequence differences of amino acids. So, a sliding window is performed on all sequences according to the fixed length k and frequency statistics are performed on the obtained *kmer.* Then the high-frequency (greater than the quartile) *kmer* is selected as the segment marker, segment operation is performed on all the protein sequences, and all unequal short sequences

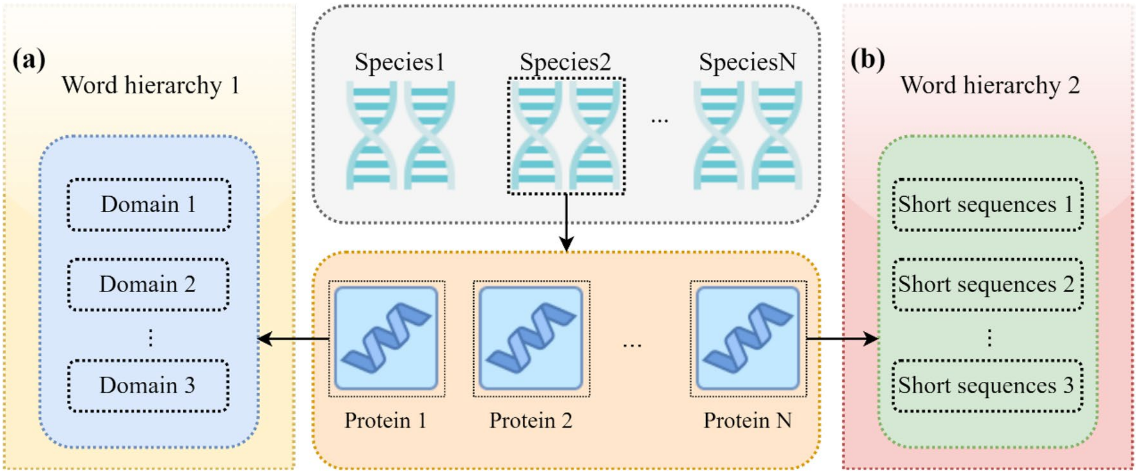

Figure 2. Word hierarchy construction. (**a**) Use the name of the Pfam functional domain contained in a protein as a word hierarchy to characterize the protein. (**b**) Short sequences obtained using the variant-based *kmer* method were used as word hierarchies to characterize the proteins.

obtained were used as word hierarchies to characterize the possible selective evolution of genes (as shown in Supplementary -file 1: Fig. S1). These short sequences can be viewed as unequal short peptides, as shown in Fig. 2b.

Word embedding

As mentioned in Sect. "Introduction", in traditional machine learning methods, it is often necessary to require feature engineering, which is usually based on a statistical approach to the original sequence. Their traditional features require manual construction, such as counting the spectrum of *kmer* in the sequence as a feature of the input sequence[36]. However, this approach does not fully represent all the information contained in the original sequence in particular, nor does it reflect the key contextual relationships. With the rise of deep learning, feature methods now focus more on the original sequence itself by directly encoding the original sequence to vectorize input features, such as one-hot encoding. However, the one-hot encoded vector is too sparse and does not express the correlation between the meanings of words in the original sequence. Unlike one-hot coding, a technique called word embedding captures the semantic association between words and can help obtain a better and more specific representation of sequence features. In this paper, for the data represented by the two types of word hierarchies mentioned above, the word embedding pre-training is performed using the *Skip-gram* algorithm[37] to obtain a vector representation of sequences with embedding dimension 200. The internal words of each protein sequence are replaced with the word vector representation obtained from the pre-trained model, which is converted into a feature matrix by concatenating all word embedding vectors in that protein sequence. Similarly, the feature matrix corresponding to each protein sequence is concatenated to obtain a complete vector representation of all protein sequences in a species.

Classification model construction and weight calculation

Considering the different protein information using different word hierarchy representations, as shown in Fig. 3a, and the two classification networks were trained. After that, the two representations of the original protein sequences of each species were input to the two trained models to obtain the domain attention scores ($PAS_{1,s}$) and kmer attention scores ($PAS_{2,s}$) for each species. Considering the hierarchical structure of the samples and the need for attention scores, this paper uses a hierarchical attention classification network to capture the weights at different hierarchies. The architecture of the hierarchical attention classification network is shown in Fig. 3b. The two sets of $PAS_{v,s}$($PAS_{1,s}$ and $PAS_{2,s}$) were fused to obtain the fused attention scores ($PAS_{3,s}$) using the self-attention mechanism to reflect the variation mechanisms captured by different word hierarchies simultaneously, as shown in Fig. 3c. The details of the hierarchical attention classification network and the self-attention fusion module will be described in the next subsections.

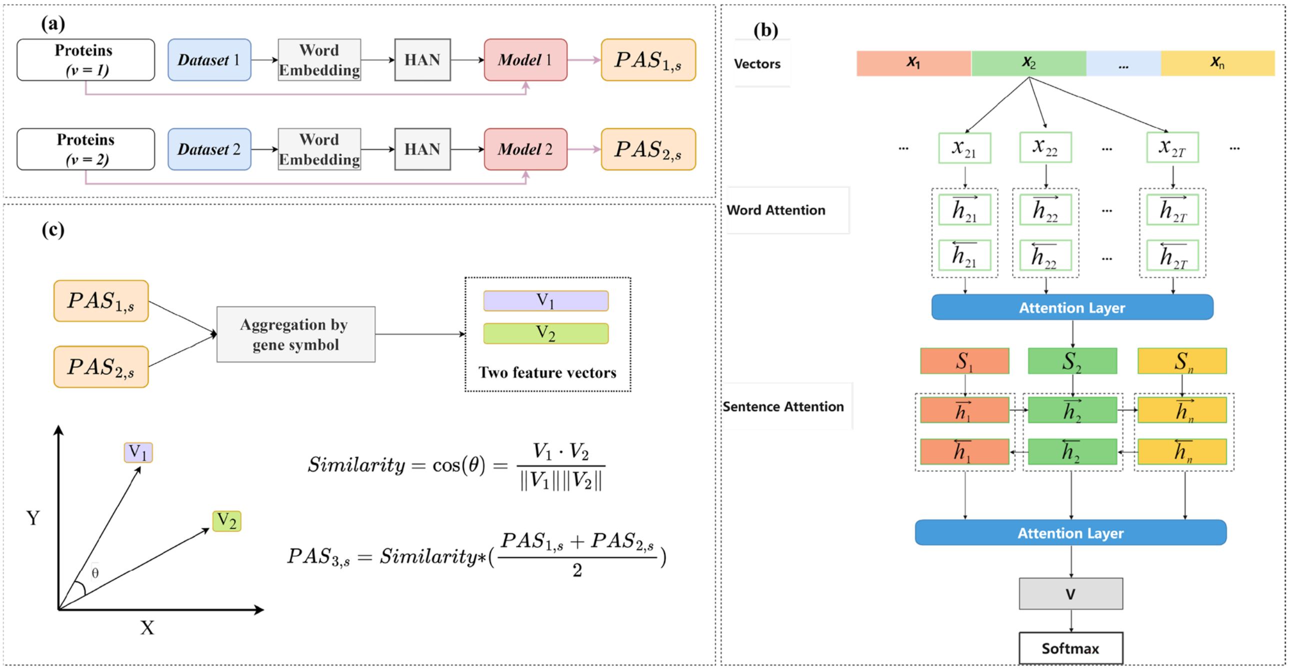

Figure 3. Schematic diagram of classification model and weight calculation. (**a**) The main process of classification, after data pre-processing using two type word hierarchical representations of two data sets were pre-trained with word embedding through a hierarchical attention classification network to obtain two classification models, and the original protein sequences of each specie were input into the trained classification model to obtain the domain attention scores ($PAS_{1,s}$) and kmer attention scores ($PAS_{2,s}$). (**b**) The network architecture of the hierarchical attention classification network (HAN) with two attention layers added to the basic classification network. (**c**) Fused attention scores ($PAS_{3,s}$) based on two sets of weights ($PAS_{1,s}$ and $PAS_{2,s}$).

Hierarchical attention classification network

In the classification process, the abstract hidden layer representation of each data segment is obtained by passing the data through *BiLSTM*. The $x_{i,t}$ is the *i*th word vector in the sequence at time t , and the bidirectional $\overrightarrow{h_{i,t}}$ and $\overleftarrow{h_{i,t}}$ represent the forward and backward hidden layer states of the $\overrightarrow{h_{i,t-1}}$ and $\overleftarrow{h_{i,t+1}}$ in the *BiLSTM* (Eqs. (1), (2)). The bidirectional information is integrated to obtain the bidirectional hidden layer state $h_{i,t}$(Eq. (3)).

$$\overrightarrow{h_{i,t}} = \overrightarrow{LSTM}\left(x_{i,t}, \overrightarrow{h_{i,t-1}}\right) \tag{1}$$

$$\overleftarrow{h_{i,t}} = \overleftarrow{LSTM}\left(x_{i,t}, \overleftarrow{h_{i,t+1}}\right) \tag{2}$$

$$h_{i,t} = \left[\overrightarrow{h_{i,t}}, \overleftarrow{h_{i,t}}\right] \tag{3}$$

The bidirectional information of the word vector $x_{i,t}$ is combined into $h_{i,t}$ and it is inputted into an MLP for activation to obtain the nonlinear hidden layer representation $u_{i,t}$ (Eq. (4), $h_{i,t}^{\intercal}$represents the transpose of $h_{i,t}$) of the word vector $x_{i,t}$. In general, different words have different emotional color biases for the sentence, that is, they have different importance for a sentence, as shown in Supplementary-file 1: Fig. S2a, in order to make the BiLSTM focus on important words, we design the attention mechanism based on word vectors, as formulated in Eqs. (5) and (6).

$$u_{i,t} = tanh\left(W_w h_{i,t}^{\intercal} + b_w\right) \tag{4}$$

$$\alpha_{i,t} = \frac{exp\left(w_w u_{i,t}\right)}{\sum_t exp\left(w_w u_{i,t}\right)} \tag{5}$$

$$s_i = \sum\nolimits_t \alpha_{i,t} h_{i,t} \tag{6}$$

where W_w and b_w are the weight matrix and the bias term in the tanh function (Eqs. (4)), respectively. To better represent the word importance represented by the attention weights, we normalized the attention weights (Eqs. (5)) where w_w represents the vector of the context of $u_{i,t}$, and $\alpha_{i,t}$ denotes the attention weight of the word vector $x_{i,t}$. The sentence vector s_i represents the weight sum of the product of the word weights and the hidden layer state information $h_{i,t}$. Similar to word-level attention, different sentences have different importance for a text, as shown in Supplementary-file 1: Fig. S2b, sentence-level attention is designed so that the classification network focuses on the important sentences based on the attention weights.

$$\overrightarrow{h_i} = \overrightarrow{LSTM}\left(s_i, \overrightarrow{h_{i-1}}\right) \tag{7}$$

$$\overleftarrow{h_i} = \overleftarrow{LSTM}\left(s_i, \overleftarrow{h_{i+1}}\right) \tag{8}$$

$$h_i = \left[\overrightarrow{h_i}, \overleftarrow{h_i}\right] \tag{9}$$

where $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ are the forward and backward implicit variables of $\overrightarrow{h_{i-1}}$and$\overleftarrow{h_{i+1}}$ for the *i*th sentence in the text, respectively. h_i is the implicit variable for the *i* th sentence in both directions. Similar to the attention mechanism at the word level, a sentence attention mechanism is designed for the sentence level, and it is calculated as follows:

$$u_i = tanh\left(W_s h_i^{\intercal} + b_s\right) \tag{10}$$

$$\alpha_i = \frac{exp(w_s u_i)}{\sum_i exp(w_s u_i)} \tag{11}$$

$$P = \sum\nolimits_t \alpha_i h_i \tag{12}$$

where W_s and b_s are the weight matrix and bias vector of the tanh function at the sentence level, respectively. w_s represents the vector of the context of u_i, α_i is the normalized attention weight of the *i* th sentence. And P is the weight sum of all sentences in the text(genome) and the attention of the sentence is its weight. Finally, P is fed to a fully connected layer to calculate the output classification probability $\widehat{y}$.

$$\widehat{y} = softmax(W_c P + b_c) \tag{13}$$

Weight acquisition and fusion

The two sets of $PAS_{v,s}$ ($PAS_{1,s}$ and $PAS_{2,s}$) from different v were aggregated according to gene symbols and the aggregation results were considered as two sets of feature vectors as shown in Fig. 3c. The similarity between these two sets of feature vectors is calculated using cosine similarity (Eq. (14)) and use the multiplication of this similarity and the mean of the first two sets of $PAS_{v,s}$ as $\mathrm{PAS}_{3,s}$ (Eq. (15)).

$$similarity = cos(\theta) = \frac{V_1 \cdot V_2}{\|V_1\|\|V_2\|} \tag{14}$$

$$PAS_{3,s} = similarity * \left(\frac{PAS_{1,s}+PAS_{2,s}}{2}\right) \tag{15}$$

Evaluation and KEGG enrichment

The normalized $PAS_{v,s}$ corresponding to protein sequences with the same annotation name are summed as the two total contributions of the protein sequences to the taxon $GAS_{v,T}$. The normalized $GAS_{v,T}$ of the two taxa are summed as the total contribution to the classification process(Eq. (16)).

$$DAS_v = \left[normalize\left(GAS_{(v,diurnal)}\right) + normalize\left(GAS_{(v,nocturnal)}\right)\right] \tag{16}$$

In this paper, we ranked the genes according to DAS_v (weight of genes) from high to low, and the top 1% of genes were used as the key genes for the macroevolution of Lepidoptera. In order to explore the difference of high-weight genes between butterflies and moths, we analyzed the evolution of ninaB, GNB1l and eys genes obtained by different types of attention mechanisms (domain attention, kmer attention and fused attention). The phylogeny tree of 18 butterflies and 13 moths was obtained from Timetree (http://www.timetree.org/) (three butterfly species are missing in Timetree). NinaB genes were identified based on the annotation of InsectBase (http://v2.insect-genome.com/). Sequence of GNB1l genes was aligned by MegaX[38]. Domains of eys genes were annotated by Pfam (http://pfam.xfam.org/). To verify the accuracy of the key genes leading to macroevolution in Lepidoptera identified in this paper, the enrichment analysis of the KEGG metabolic pathway and the search of the corresponding gene functions were performed. If the enriched metabolic pathways are significantly associated with differences in Diel behavior or if the functions of certain genes are associated with certain macroscopic phenotypes of the two major taxa, then the approach of this paper is proven to be effective.

Result

Data pre-processing results

After Pfam annotation of all protein sequences of the original 34 Lepidopteran insect species, only proteins with functional domains were retained as shown in Fig. 4.

The results of word hierarchy construction are as follows ($v = 1$): After counting the results of Pfam annotation, there were 6,448 functional domains in all Lepidoptera proteins, and all the functional domain names were recorded as a word list. The differences in the number of kmer at different k values can lead to differences in the number of split tokens, and thus leads to differences in segmentation results. In order to reduce the occurrence of very few words to preserve the integrity of the corpus as much as possible, this paper investigates the retention of the corpus under various k-values. as shown in Supplementary-file2: Table S3. In the case of $k = 3$, the corpus is retained to the highest degree, which is about 95% of the original corpus, so the new corpus obtained in the case of $k = 3$ is chosen for word embedding in this paper. Then, Skip-gram algorithm was used to pretrain the word embedding of the augmented corpus with an embedding dimension of 200. Then, K-means clustering was adopted to the embedded word vector ($K = 3$, representing tripartite clusters, namely two specific taxa words, and their intersection words), and used the PCA algorithm to downscale the word vector to two dimensions as shown in Fig. 5. It is obvious from Fig. 5 that the word vectors have a clear separation trend after clustering at $v = 1$, while they do not show a similar separation trend at $v = 2$. This is caused by the larger scale and more obvious features of the functional domain.

Acquisition and analysis of weights (*$PAS_{v,s}$, $GAS_{v,T}$* and *DAS_v*)

The two sets of $PAS_{v,s}$ ($PAS_{1,s}$ and $PAS_{2,s}$)were obtained by inputting all protein sequences of each real species based on two word hierarchical representations corresponding to two classification models(the training process of the two models is shown in Supplementary-file1: Fig. S3) respectively. Then $PAS_{3,s}$ was obtained by fusing $PAS_{1,s}$ and $PAS_{2,s}$ through the self-attention mechanism. The $GAS_{v,T}$ obtained after clustering based on protein information annotation are shown in Supplementary-file1: Fig. S4. The distribution of top 1% DAS_v is shown in Fig. 6 and detail genetic information and weights at different v are available in Supplementary file 2: Tables S4–S6. Several genes with the highest DAS_1 values were FPS, GGPS1 and ninaB, corresponding to DAS_1 of 0.58, 0.48, 0.21, respectively. Several genes with the highest DAS_2 values were nfil3 and GNBIL, corresponding to DAS_2 of 2.05 and 0.16 respectively. Several genes with the highest DAS_3 values were nfil3 and Fbxo42, corresponding to DAS_3 of 1.03 and 0.38 respectively. In particular, nine genes (Plc21C[39,40] EP300[41], Timeless[42], foxo[43], norpA[44], nfil3[45], to[46], dyw[47] and Nup153[48]) were found to relate to circadian rhythms in previous studies (Fig. 6a–c).

Furthermore, we compared the gene number of ninaB with the high DAS_1 values ($v = 1$) between butterflies and moths. There are more than two ninaB genes in most species of butterflies, while only one was found in each moth species (Fig. 7a). The results suggest that the method of word hierarchy ($v = 1$) could reflect the quantity or position changes of domains, including the contraction and expansion of gene families. Meanwhile, we compared the sequence variation of GNB1l gene with the high DAS_2 values ($v = 2$) between butterflies and moths (Supplementary-file3). A few of specific variable sites causing the kmer weight changes were found in GNB1l gene between butterflies and moths, which contains several non-synonymous mutations (Fig. 7b). The variant kmer method of word hierarchy ($v = 2$) could reflect the sequence variation, or small InDel in the macroevolution process of Lepidoptera. For the eys gene with the high DAS_3 values ($v = 3$), hEGF domain showed different insertion or deletion between butterflies and moths (Fig. 7c). These diverse domains not only affect the weight of

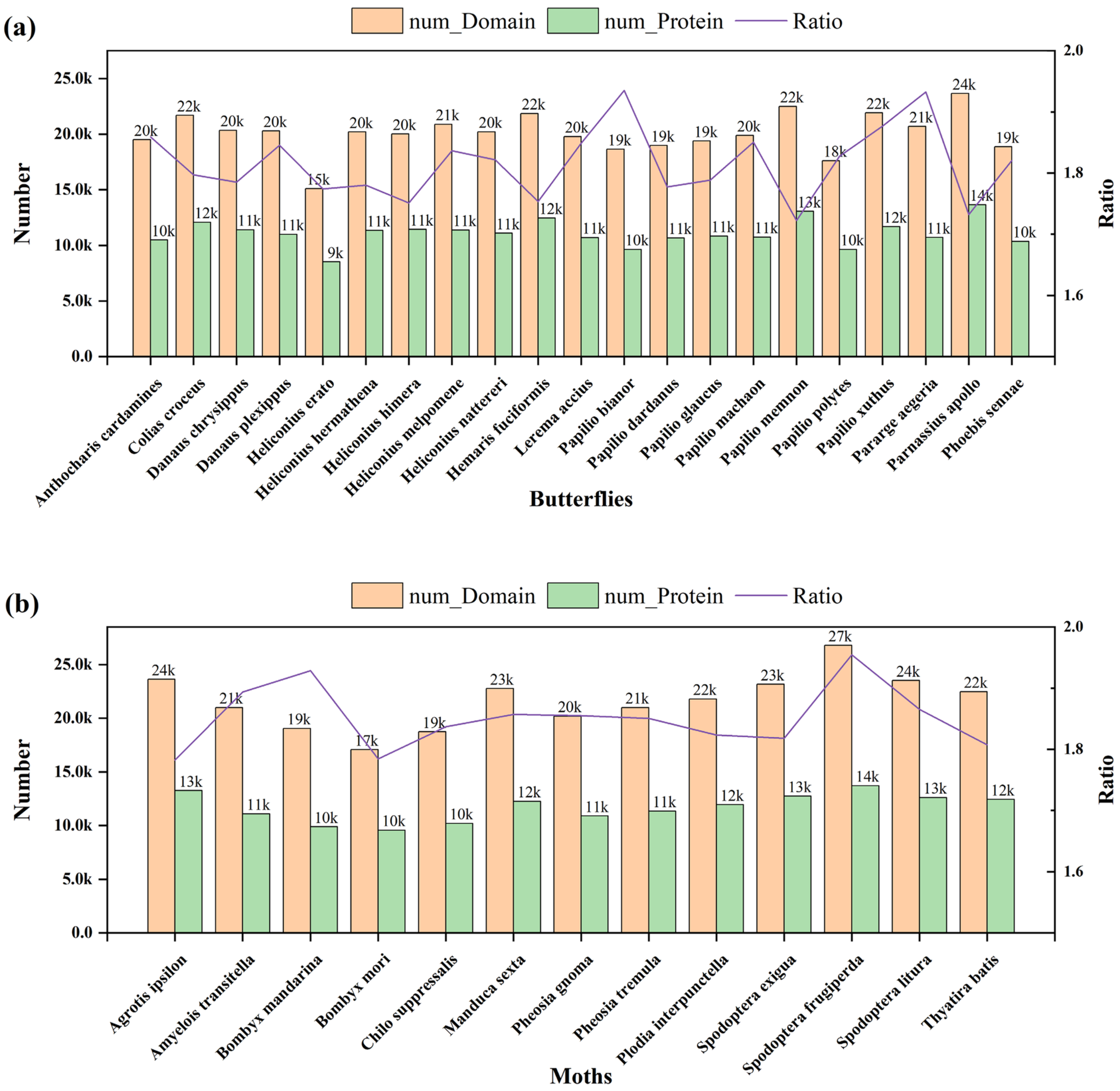

Figure 4. Pfam annotation results for protein sequences in 34 Lepidoptera species. (**a**) represent the Pfam annotation results for diurnal butterflies. (**b**) represent the Pfam annotation results for nocturnal moths. *num_Protein* and *num_Domain* are the number of coding genes and the number of functional domains retained in the Lepidopteran genome after Pfam annotation, respectively. *Ratio = num_Protein/num_Domain*, which is the (average) number of functional domains contained in a single coding gene per species.

the domain (*v* = *1*) but also the weight of the kmer (*v* = *2*), so this fusion method (*v* = *3*) may reflect the synergistic effects of the above two (v = 1 and *v* = *2*).

KEGG enrichment of key genes with high weights

The above genes (Top 1%) obtained by three types of attention mechanisms (domain attention, kmer attention and fused attention) were regarded as the key genes in macroevolution of Lepidoptera. These genes were taken for the KEGG enrichment (as shown in Fig. 8). It can be seen that the three groups of genes have some commonality and are all enriched in some specific pathways, such as *Phototransduction—fly* (where Negative logarithmic P-value is 5.91, 2.37, 0.52, respectively, for a total of 8.80), *Phosphatidylinositol signaling system* (where Negative logarithmic P-value is 6.96, 3.86, 1.30, respectively, for a total of 12.12), *Drug metabolism—other enzymes* (where Negative logarithmic P-value is 10.695, 4.588, 0.21, respectively, for a total of 5.29), *Fanconi anemia pathway* (where Negative logarithmic P-value is 7.3, 2.15, 1.12, respectively, for a total of 10.57), *Terpenoid backbone biosynthesis* (where Negative logarithmic P-value is 6.48, 0.14, 0.62, respectively, for a total of 7.24), etc.

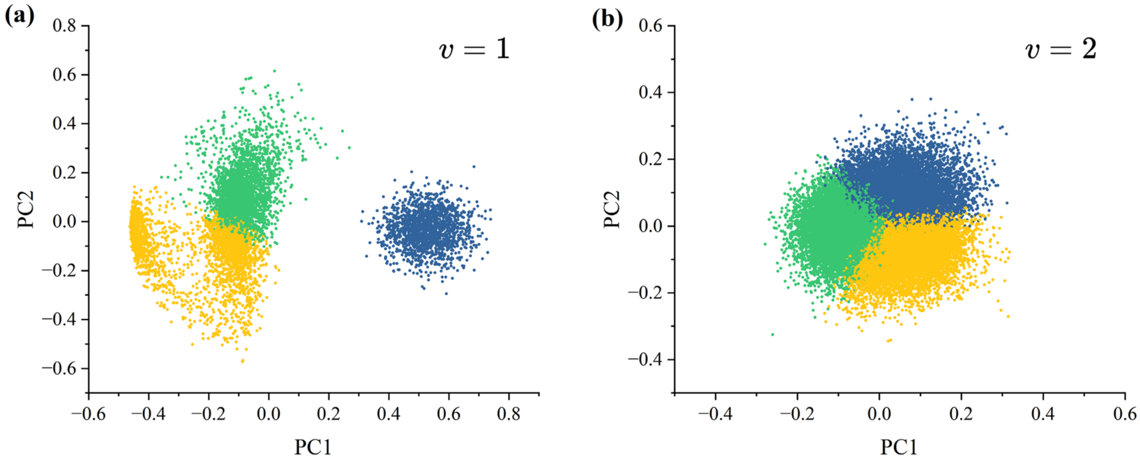

Figure 5. Word embedding results graph. (**a**) represents the result of word embedding when using v = 1, i.e., using the functional domain name as the word hierarchy. (**b**) represent the word embedding results of short sequences obtained as word hierarchies (v = 2) using k(k = 3) as the sliding window length.

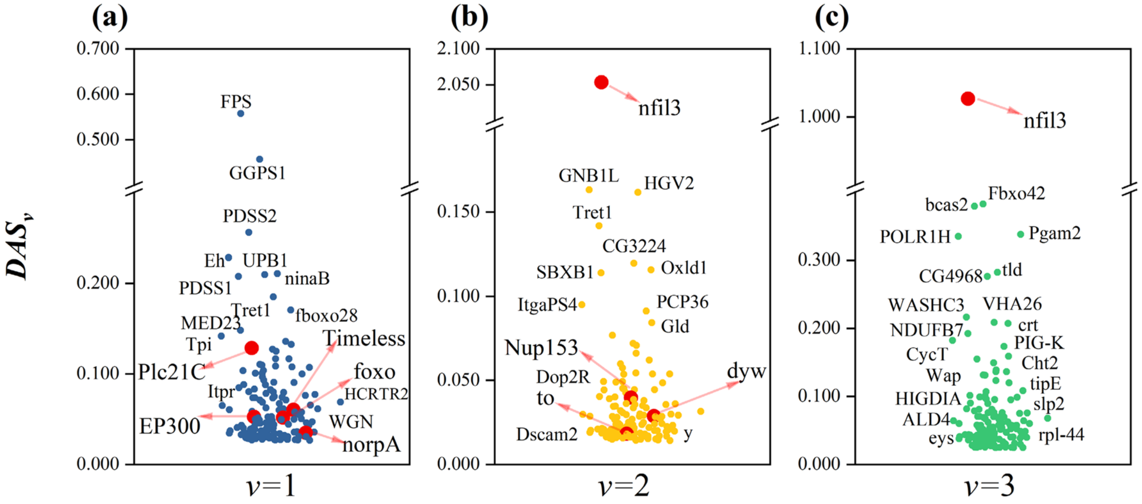

Figure 6. The distribution of Top 1% DAS_v values based on different word hierarchies (v). (**a**) represent the distributions of Top 1% DAS_1 by domain attention (v = 1); (**b**) represent the distributions of Top 1% DAS_2 by kmer attention (v = 2); (**c**) represent the distributions of Top 1% DAS_3 by fused attention (v = 3). The genes with red color in this figure have been reported to be associated with circadian rhythms in previous studies.

Discussion

Phenotypic differences of macroevolution usually represent the synergistic action of multiple key genes in evolutionary biology[49–51]. However, there are still some challenges to establish a universal method or model for exploring these key genes of macroevolution[49,52]. The first challenge comes from the diversity of biological sequences (DNA and proteins). A central issue for machine learning methods is how to design a good representation for the biological sequences[53,54]. The word embeddings can capture the semantic correlation between words and reflect the contextual relationship of the original sequence. We used two characteristic word-level construction methods including functional domain ($v = 1$) and variable length kmer ($v = 2$). The functional domain embedding can well reflect the domain variations at a large scale, such as gene duplication[15,16], structure variation[19], etc. The kmer embedding can well reflect the sequence variation, such as selective evolution[17,18], small InDel, etc. Instead of the common kmer method[55–58], the variable length kmer word embedding comprehensively consider the kmer with different lengths based on probability. The embedding and partitioning method can better reserve the diversity of gene sequences. Therefore, these two methods of word embedding take into account various scales of gene variation as well as various relationships with biological sequence location and context. In addition to the word hierarchies at both scales ($v = 1$ and $v = 2$), a fusion method ($v = 3$) is proposed to capture the combined molecular mechanisms influenced by domain and variant kmer. The second challenge is to clarify the inference process of DL algorithms[59]. This study focuses on the molecular mechanisms behind macroevolution and models it into a computer classification problem using the genomes of taxa that have undergone macroevolution. The aim is not just to create a classification model with high accuracy, but to understand the inference process itself, specifically which genes are important for classification. To achieve this, the study proposes the inclusion of

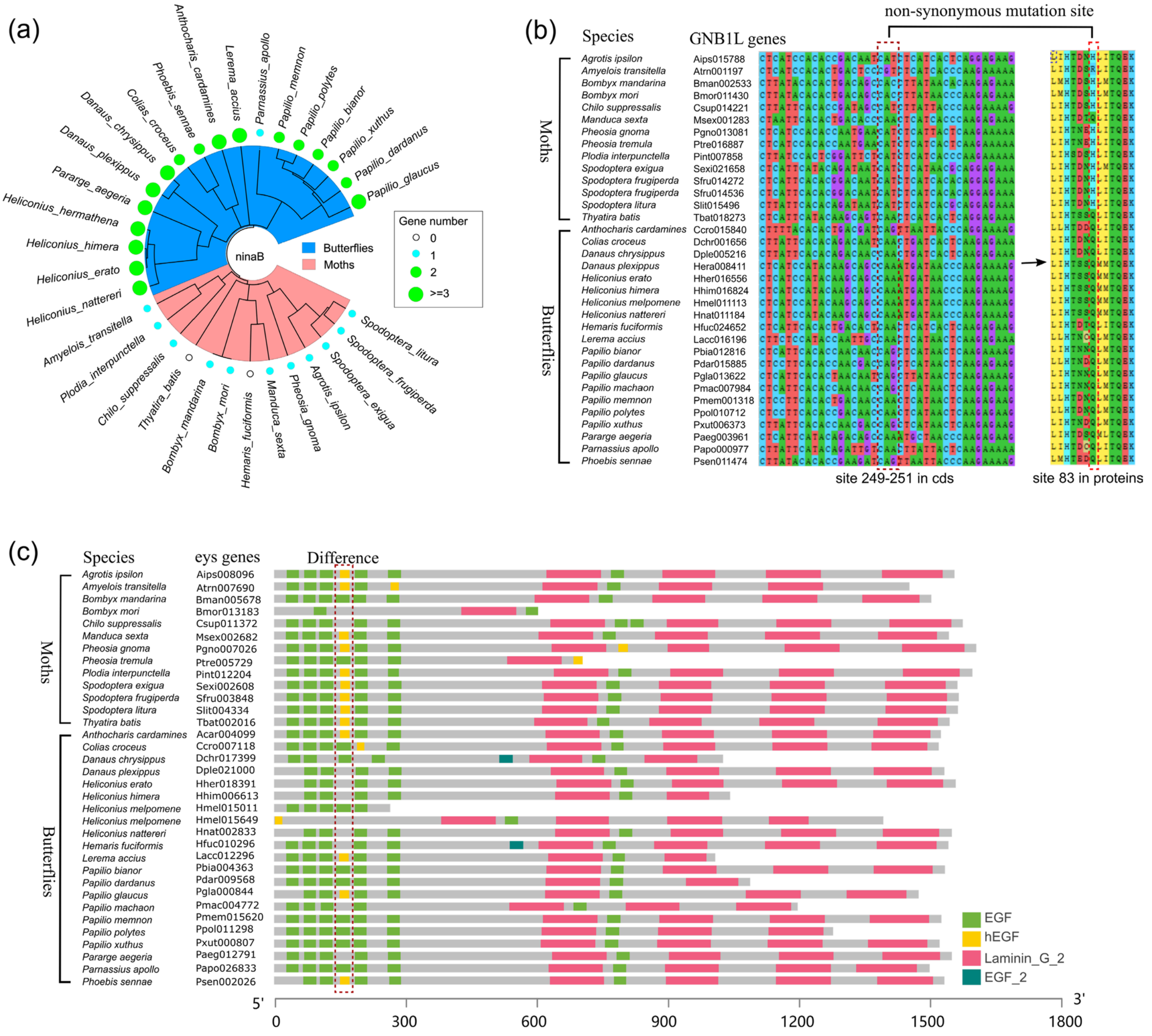

Figure 7. Examples of high-weight genes identified by three types of attention mechanisms. (**a**) Gene number of ninaB in butterflies and moths (identified by domain attention). (**b**) Non-synonymous mutation site of GNB1L genes in butterflies and moths (identified by kmer attention). (**c**) Domain difference of eys genes in butterflies and moths (identified by fused attention).

AM as a feasible strategy[60]. Combining with the hierarchical structural properties of biological sequences, this study incorporates a hierarchical AM into the deep classification model, so that the model can focus not only on important "words" (Domain/short sequences) but also on important "sentences" (proteins). Interestingly, three types of attention mechanisms (domain attention, kmer attention and fused attention) maybe stand for different molecular mechanisms of macroevolution in evolutionary biology (Fig. 7).

Previous studies indicated that the two major taxa of butterflies and moths showed significant differences in circadian rhythm[61]. Our results identified a number of genes with high weights, which were mainly enriched in Phototransduction—fly, Phosphatidylinositol signaling system, Inositol phosphate metabolism, Wnt signaling pathway, MAPK signaling pathway—fly, Notch signaling pathway as well as FoxO signaling pathway etc. Most of these genes have been reported to be associated with the control of circadian rhythms in insects, such as dyw (daywake)[47], to (takeout)[46], EP300 (E1A binding protein p300)[41], Plc21C (Phospholipase C at 21C)[39,40], norpA (no receptor potential A)[44], Nup153 (Nucleoporin 153kD)[48], Nfil3 (nuclear factor, interleukin 3, regulated)[45], foxo (forkhead box, sub-group O)[43] and Timeless (timeless circadian regulator)[42] (Figs. 6, 7). These reported circadian rhythm-related genes with high weights proved the validity of our method. Moreover, it is suggested that some of the other high-weight genes identified in this paper may also play important roles in the macroevolution of Lepidopterans. We found some high weights genes were reported to be associated with senses in Lepidopteran insects[18,62], such as ninaB (neither inactivation nor afterpotential B)[63], eys(eyes shut)[64], and Dscam2 (Down syndrome cell adhesion molecule 2)[65] related to axonal tiling of the insect visual system, the aop (anterior open) gene related to the photoreceptor rhabdomere[66], Itpr (Inositol 1,4,5,-trisphosphate receptor) related to visual and

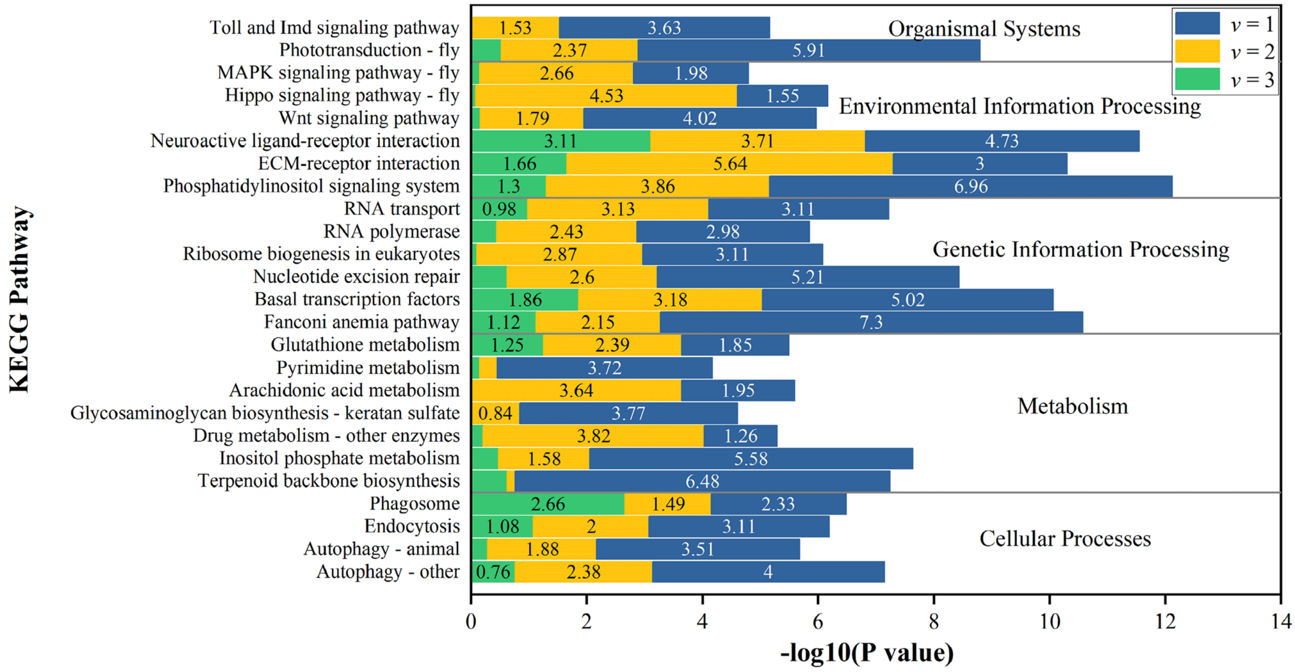

Figure 8. KEGG enrichment analysis results of high-weight genes (top 1%). The reference species is silkworm. Blue color stands for the domain attention; Yellow color stands for the kmer attention; Green color stands for the fused attention.

olfactory transduction[67], as well as WGN (Wengen) related to photoreceptor cell axon guidance[68]. Additionally, we also identified many genes that may be involved in these behavioral differences in butterflies and moths, for example, y (yellow) and Dop2R (Dopamine 2-like receptor) are involved in male courtship behavior of insects[69,70], HCRTR2(hypocretin receptor 2) may be involved in regulating feeding and sleep behavior[71,72]. The above genes related to circadian rhythms, sensory organs, and behavioral habits should help us to explain the macroscopic differences of diurnal butterflies and nocturnal moths in Lepidoptera.

Conclusion

This paper proposes a new method for identifying the important genes of macroevolution using deep learning and attention mechanism. Based on this new method, we mined a few of key genes related to the phenotypic differences (circadian rhythms, sensory organs, as well as behavioral habits etc) of diurnal butterflies and nocturnal moths in Lepidoptera. It not only provides a novel method to identified the key genes of macroevolution at the genomic level, but also helps us to understand the microevolution mechanisms of diurnal butterflies and nocturnal moths in Lepidoptera.

Data availability

The source data and experiment code for our implementation are available for public access and can be found in GitHub (https://github.com/JiaweiMao12135/IKGM). The code is written in Python and serves as a reference for the experiments conducted in this paper. We encourage collaboration and feedback from the community to improve the code and foster future advancements. Feel free to report any issues or suggest improvements by creating an issue in the GitHub repository's issue tracker. Please note that while we have taken measures to thoroughly test the code, unforeseen issues or limitations may still exist. We appreciate your understanding and assistance in refining the codebase. By sharing our code, we aim to contribute to the open research community and promote reproducibility, allowing others to validate our results and build upon our work.

Received: 27 June 2023; Accepted: 9 November 2023
Published online: 13 November 2023

References

1. Fish, F. E. Transitions from drag-based to lift-based propulsion in mammalian swimming. *Am. Zool.* **36**, 628–641 (1996).
2. Ashley-Ross, M. A., Hsieh, S. T., Gibb, A. C. & Blob, R. W. Vertebrate land invasions-past, present, and future: An introduction to the symposium. *Integr. Comp. Biol.* **53**, 192–196 (2013).
3. Zimmer, C. *At the Water's Edge: Fish with Fingers, Whales with Legs, and How Life Came Ashore but Then Went Back to Sea* (Simon and Schuster, 2014).
4. Ruiz-Herrera, A. & Robinson, T. J. Chromosomal instability in Afrotheria: Fragile sites, evolutionary breakpoints and phylogenetic inference from genome sequence assemblies. *BMC Evol. Biol.* **7**, 199 (2007).
5. Dececchi, T. A. & Larsson, H. C. E. Body and limb size dissociation at the origin of birds: Uncoupling allometric constraints across a macroevolutionary transition. *Evolution* **67**, 2741–2752 (2013).

6. Behrens, M., Di Pizio, A., Redel, U., Meyerhof, W. & Korsching, S. I. At the Root of T2R Gene Evolution: Recognition Profiles of Coelacanth and Zebrafish Bitter Receptors. Genome Biol Evol 13, evaa264 (2021).
7. Hannisdal, B. & Peters, S. E. Phanerozoic Earth system evolution and marine biodiversity. *Science* **334**, 1121–1124 (2011).
8. Mayhew, P. J., Bell, M. A., Benton, T. G. & McGowan, A. J. Biodiversity tracks temperature over time. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 15141–15145 (2012).
9. Doyle, J. J. & Egan, A. N. Dating the origins of polyploidy events. *New Phytol.* **186**, 73–85 (2010).
10. Clark, J. W. & Donoghue, P. C. J. Whole-genome duplication and plant macroevolution. *Trends Plant Sci.* **23**, 933–945 (2018).
11. Clark, J. W., Puttick, M. N. & Donoghue, P. C. J. Origin of horsetails and the role of whole-genome duplication in plant macroevolution. *Proc. Biol. Sci.* **286**, 20191662 (2019).
12. Guo, B., Wagner, A. & He, S. Duplicated gene evolution following wholegenome duplication in teleost Fish. *Gene Duplic.* **27**, 36 (2011).
13. Schwager, E. E. *et al.* The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biol.* **15**, 62 (2017).
14. Fan, Z. *et al.* A chromosome-level genome of the spider Trichonephila antipodiana reveals the genetic basis of its polyphagy and evidence of an ancient whole-genome duplication event. *Gigascience* **10**, giab016. https://doi.org/10.1093/gigascience/giab016 (2021).
15. Baumholtz, A. I., Gupta, I. R. & Ryan, A. K. Claudins in morphogenesis: Forming an epithelial tube. *Tissue Barriers* **5**, e1361899 (2017).
16. Hughes, G. M. *et al.* The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Mol. Biol. Evol.* **35**, 1390–1406 (2018).
17. Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau | Nature Communications. https://www.nature.com/articles/ncomms3071.
18. Sondhi, Y., Ellis, E. A., Bybee, S. M., Theobald, J. C. & Kawahara, A. Y. Light environment drives evolution of color vision genes in butterflies and moths. *Commun. Biol.* **4**, 177 (2021).
19. Timmermans, M. J. T. N., Srivathsan, A., Collins, S., Meier, R. & Vogler, A. P. Mimicry diversification in Papilio dardanus via a genomic inversion in the regulatory region of engrailed-invected. *Proc. Biol. Sci.* **287**, 20200443 (2020).
20. Hayward, A., Cornwallis, C. K. & Jern, P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 464–469 (2015).
21. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).
22. Li, H. *et al.* Panoramic insights into microevolution and macroevolution of a prevotella copri-containing lineage in primate guts. *Genom. Proteom. Bioinform.* **20**, 334–349 (2022).
23. Larrañaga, P. *et al.* Machine learning in bioinformatics. *Brief. Bioinform.* **7**, 86–112 (2006).
24. Leung et al. Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. (2015).
25. Hroza & Jiří. Protein secondary structure prediction by machine learning methods. Bioinformatics 14, 892–893 (2005).
26. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Brief. Bioinform.* **18**, 851–869 (2017).
27. Li, Y. *et al.* Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods* **166**, 4–21 (2019).
28. Wang, W. & Gao, X. Deep learning in bioinformatics. *Methods* **166**, 1–3 (2019).
29. Li, H. *et al.* Modern deep learning in bioinformatics. *J. Mol. Cell Biol.* **12**, 823–827 (2020).
30. Berrar, D. & Dubitzky, W. Deep learning in bioinformatics and biomedicine. *Brief. Bioinform.* **22**, 1513–1514 (2021).
31. Attention is all you need Proceedings of the 31st International Conference on Neural Information Processing Systems. https://doi.org/10.5555/3295222.3295349.
32. Hong, J., Gao, R. & Yang, Y. CrepHAN: Cross-species prediction of enhancers by using hierarchical attention networks. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab349 (2021).
33. Fergadis, A., Baziotis, C., Pappas, D., Papageorgiou, H. & Potamianos, A. Hierarchical bi-directional attention-based RNNs for supporting document classification on protein-protein interactions affected by genetic mutations. *Database (Oxford)* https://doi.org/10.1093/database/bay076 (2018).
34. Mei, Y. *et al.* InsectBase 2.0: A comprehensive gene resource for insects. *Nucleic Acids Res.* **50**, D1040–D1045 (2022).
35. Barber, J. Diel behavior in moths and butterflies: A synthesis of data illuminates the evolution of temporal activity. *Organ. Divers. Evol.* https://doi.org/10.1007/s13127-017-0350-6 (2018).
36. Chen, L., Fish, A. E. & Capra, J. A. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS Comput. Biol.* **14**, e1006484 (2018).
37. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. Preprint at http://arxiv.org/abs/1310.4546 (2013).
38. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
39. Ogueta, M., Hardie, R. C. & Stanewsky, R. Non-canonical phototransduction mediates synchronization of the drosophila melanogaster circadian clock and retinal light responses. *Curr. Biol.* **28**, 1725-1735.e3 (2018).
40. Ogueta, M., Hardie, R. C. & Stanewsky, R. Light sampling via throttled visual phototransduction robustly synchronizes the drosophila circadian clock. *Curr. Biol.* **30**, 2551-2563.e3 (2020).
41. Curtis, A. M. *et al.* Histone acetyltransferase-dependent chromatin remodeling and the vascular clock. *J. Biol. Chem.* **279**, 7091–7097 (2004).
42. Cai, Y. D. & Chiu, J. C. Timeless in animal circadian clocks and beyond. *FEBS J.* **289**, 6559–6575 (2022).
43. Zheng, X., Yang, Z., Yue, Z., Alvarez, J. D. & Sehgal, A. FOXO and insulin signaling regulate sensitivity of the circadian clock to oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 15899–15904 (2007).
44. Saint-Charles, A. *et al.* Four of the six Drosophila rhodopsin-expressing photoreceptors can mediate circadian entrainment in low light. *J. Comp. Neurol.* **524**, 2828–2844 (2016).
45. Liu, W. *et al.* Dibutyl phthalate disrupts conserved circadian rhythm in Drosophila and human cells. *Sci. Total Environ.* **783**, 147038 (2021).
46. So, W. V. *et al.* takeout, a novel Drosophila gene under circadian clock transcriptional regulation. *Mol. Cell. Biol.* **20**, 6935–6944 (2000).
47. Yang, Y. & Edery, I. Daywake, an anti-siesta gene linked to a splicing-based thermostat from an adjoining clock gene. *Curr. Biol.* **29**, 1728-1734.e4 (2019).
48. Jang, A. R., Moravcevic, K., Saez, L., Young, M. W. & Sehgal, A. Drosophila TIM binds importin α1, and acts as an adapter to transport PER to the nucleus. *PLoS Genet.* **11**, e1004974 (2015).
49. Pagel, M., O'Donovan, C. & Meade, A. General statistical model shows that macroevolutionary patterns and processes are consistent with Darwinian gradualism. *Nat. Commun.* **13**, 1113 (2022).
50. Molecular phylogeny and macroevolution of Chaitophorinae aphids (Insecta: Hemiptera: Aphididae). Systematic Entomology (2021) doi:https://doi.org/10.1111/syen.12531.
51. Bagchi, B. *et al.* Sexual conflict drives micro- and macroevolution of sexual dimorphism in immunity. *BMC Biol.* **19**, 114 (2021).

52. Alencar, L. R. V. & Quental, T. B. Exploring the drivers of population structure across desert snakes can help to link micro and macroevolution. *Mol. Ecol.* **28**, 4529–4532 (2019).
53. Zou, Q., Xing, P., Wei, L. & Liu, B. Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* **25**, 205–218 (2019).
54. Hoinka, J. & Przytycka, T. M. Embedding gene sets in low-dimensional space. *Nat. Mach. Intell.* **2**, 367–368 (2020).
55. Wen, J., Chan, R. H. F., Yau, S.-C., He, R. L. & Yau, S. S. T. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* **546**, 25–34 (2014).
56. Fletez-Brant, C., Lee, D., McCallion, A. S. & Beer, M. A. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res.* **41**, W544-556 (2013).
57. Zhu, Z. *et al.* Predicting the receptor-binding domain usage of the coronavirus based on kmer frequency on spike protein. *Infect. Genet. Evol.* **61**, 183–184 (2018).
58. Villacrés-Vallejo, J. *et al.* Using full chloroplast genomes of 'red' and 'yellow' Bixa orellana (achiote) for kmer based identification and phylogenetic inference. *BMC Genom.* **21**, 544 (2020).
59. Sheehan, S. & Song, Y. S. Deep learning for population genetic inference. *PLoS Comput. Biol.* **12**, e1004845 (2016).
60. Xuan, P., Cao, Y., Zhang, T., Kong, R. & Zhang, Z. Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes. *Front. Genet.* https://doi.org/10.3389/fgene.2019.00416 (2019).
61. Brady, D., Saviane, A., Cappellozza, S. & Sandrelli, F. The circadian clock in lepidoptera. *Front. Physiol.* **12**, 776826 (2021).
62. Vogt, R. G., Große-Wilde, E. & Zhou, J.-J. The Lepidoptera Odorant Binding Protein gene family: Gene gain and loss within the GOBP/PBP complex of moths and butterflies. *Insect Biochem. Mol. Biol.* **62**, 142–153 (2015).
63. Voolstra, O. *et al.* NinaB is essential for Drosophila vision but induces retinal degeneration in opsin-deficient photoreceptors. *J. Biol. Chem.* **285**, 2130–2139 (2010).
64. Husain, N. *et al.* The agrin/perlecan-related protein eyes shut is essential for epithelial lumen formation in the Drosophila retina. *Dev. Cell* **11**, 483–493 (2006).
65. Millard, S. S., Flanagan, J. J., Pappu, K. S., Wu, W. & Zipursky, S. L. Dscam2 mediates axonal tiling in the Drosophila visual system. *Nature* **447**, 720–724 (2007).
66. Nam, S.-C. & Choi, K.-W. Interaction of Par-6 and Crumbs complexes is essential for photoreceptor morphogenesis in Drosophila. *Development* **130**, 4363–4372 (2003).
67. Yoshikawa, S. *et al.* Molecular cloning and characterization of the inositol 1,4,5-trisphosphate receptor in Drosophila melanogaster. *J. Biol. Chem.* **267**, 16613–16619 (1992).
68. Ruan, W., Unsain, N., Desbarats, J., Fon, E. A. & Barker, P. A. Wengen, the sole tumour necrosis factor receptor in Drosophila, collaborates with moesin to control photoreceptor axon targeting during development. *PLoS One* **8**, e60091 (2013).
69. Massey, J. H., Chung, D., Siwanowicz, I., Stern, D. L. & Wittkopp, P. J. The yellow gene influences Drosophila male mating success through sex comb melanization. *Elife* **8**, e49388 (2019).
70. Love, C. R., Gautam, S., Lama, C., Le, N. H. & Dauwalder, B. The Drosophila dopamine 2-like receptor D2R (Dop2R) is required in the blood brain barrier for male courtship. *Genes Brain Behav.* **22**, e12836 (2023).
71. Sakurai, T. *et al.* Orexins and orexin receptors: A family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92**, 573–585 (1998).
72. Yin, J. *et al.* Structure and ligand-binding mechanism of the human OX1 and OX2 orexin receptors. *Nat. Struct. Mol. Biol.* **23**, 293–299 (2016).

Author contributions

J.W.M. and Y.C. designed this study and wrote the main manuscript text. Y.J.Z. offered the related devices. and B.S.H. and Y.Z. collected the data, analyzed the data and prepared figures. Y.J.Z. and Y.C. helped for interpreting the results. Y.J.Z. and Y.Z. helped for editing the language. All of the authors contributed to the interpretation of the results and the writing of the manuscript.

Funding

This work was funded by National Natural Science Foundation (31960142, 61962055 and 32360388).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-47113-9.

Correspondence and requests for materials should be addressed to Y.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.