Journal of Computers Vol. 37 No. 5, October 2026, pp. 1-12
https://doi.org/10.63367/19911599202610370500 1

Development of Wildlife Identification System Based on Local Alignment Algorithm

Da-Wei Liu[1,3], Peng-Bo Zhao[2], Yong-Wu Zhou[1,3], Sheng-Wen Zhou[2], Jun-Chi Xiao[2], Bo-Yang Fu[1], Yi Liang[2], Shu-Ming Huang[2], and Ming-Yue Qiu[2,3*]

[1] Faculty of Criminal Science & Technology, Nanjing Police University, Nanjing 210023, China
{dwliu, zhouyw}@nfpc.edu.cn, fuby@njpu.edu.cn

[2] Faculty of information Technology, Nanjing Police University, Nanjing 210023, China
{zhaopb, zhousw, Xiaojunc, liangy, huangsm}@njpu.edu.cn, qiumy@nfpc.edu.cn

[3] Key Laboratory of State Forest and Grassland Administration Wildlife Evidence Technology, Nanjing, 210023, China

Received 1 May 2025; Revised 25 July 2025; Accepted 31 July 2025

Abstract. Wildlife crimes seriously threaten biodiversity and public health security. Identifying the species of wild animals and their products involved in such cases is the key to handling them. This study aimed to establish a wildlife DNA identification system based on a local alignment algorithm, thereby providing reliable technical support for public security organizations with high-precision and high-efficiency species identification techniques. The system adopted the Smith–Waterman local alignment algorithm, recognized the local similarity of DNA sequences with dynamic programming, and solved the problem of heavy reliance on complete sequences when using traditional global alignment. The parameters were optimized in combination with the Basic Local Alignment Search Tool, and matrix calculation was accelerated with C-Extensions for Python, thus significantly enhancing the processing efficiency for large-scale data. The database was based on self-dedicated identification data from the College of Criminal Science and Technology, Nanjing Police University. Experts reviewed the data, and the accuracy and safety of data were ensured to avoid the risk of species classification errors in open-source databases. The system innovatively integrated the Streamlit framework to develop the visualization platform and user-friendly interface. The grassroots police officers obtained species-matching results and the related species information by inputting DNA sequences, thereby facilitating case handling. The system provides an efficient technical tool for combating wildlife crimes and maintaining biodiversity, with significant practical value and immense promotion potential.

Keywords: local alignment algorithm, Smith-Waterman algorithm, wildlife

1 Introduction

Wildlife, as an essential natural resource, is a key element for maintaining the biodiversity on Earth and promoting the balance of ecosystems [1]. It can provide direct living materials for humans and indirectly support the realization of various ecological services. Protecting wildlife ensures the health of natural systems essential for human survival, while also preserving the planet's intrinsic biological and cultural value. However, Huge market demand has resulted in rampant illegal wildlife trade, seriously undermining the protection of biodiversity and the construction of Earth's civilization and posing a serious threat to public health and safety through the trade in animals for the purpose of indiscriminate wildlife consumption [2]. In order to effectively curb such crimes and effectively strengthen the protection of precious wildlife resources, it is of key significance to carry out scientific and accurate species identification of the wildlife and their products involved in the case. The identification results can become the core basis for the judicial authorities to handle the relevant cases, providing strong technical support for the fair judgment and legal application of the cases [3]. Traditional identification methods for wild animals mainly rely on observing the morphological and anatomical features and then matching with those of the typical specimens through contrastive analysis [4]. For example, it is possible to identify animal species based on the color of feathers, beak shape and claw shape of birds, or the tooth style and fur color of mammals. However, in the fight against wildlife crimes, staff always seized only animal products, mainly

* Corresponding Author

including colored feathers, tanned leather, carved bones and teeth, and blood. These products show indistinct or even absent morphological features compared with their respective species, making it impossible to identify the species through macroscopic morphological observation. Against this backdrop, DNA barcoding technology, with its outstanding performances and unique advantages, can precisely identify the involved animal samples. This provides strong technical support for public security organizations, enabling them to more effectively combat crimes that endanger animals and protect precious wildlife resources [5]. DNA barcoding technology involves comparing the obtained DNA fragments in the DNA database and drawing the final identification conclusion [6]. At present, the staff in China mainly rely on the GenBank database for comparison. GenBank is an open database where researchers worldwide upload their research results for others to use. Database managers can review the uploaded data; however, they cannot check the species corresponding to that data. Therefore, a certain risk of error exists in the results when using such a database for species identification [7]. The current databases relied upon in wildlife species identification are precisely databases with the risk of error. The present study addressed this issue by establishing its own database, aiming to improve the accuracy of wildlife species identification and provide robust data support for biodiversity conservation and wildlife law enforcement.

The research contents of this manuscript are organized as follows. In Section 2, the application of DNA barcoding in wildlife identification and the related comparison algorithms are introduced in detail, and the relevant research progress worldwide is expounded. The research process of the wildlife identification system based on the local comparison algorithm is described in Section 3. Next, the identification results are demonstrated in Section 4 using 4 frequently involved wildlife species, *Streptopelia orientalis*, *Paguma larvata*, *Nyctereutes procyonoides*, and *Python regius,* as examples. Finally, the findings are summarized to further elaborate on both characteristics and advantages.

2 Related Research

The items involved in the actual process of combating wildlife crimes are always animal remains or processed products. Compared with the original animals, these items have undergone significant modifications or completely lost their morphological features, which cannot be accurately identified by appearance. DNA barcoding is a rapid and accurate biological identification system for species based on standard DNA fragments, with sufficient variation, easy amplification, and relatively small size [8]. Compared with traditional morphological identification, DNA barcoding technology has several advantages, including easy operation, standardization, favorable universality, and high accuracy, without being affected by the macroscopic morphology of species [9]. Moreover, this technology does not require much classical taxonomic knowledge and experience of the appraiser. Currently, DNA barcoding has been extensively applied in the species identification of wild animals. Xie *et al.* provided an effective and accurate method based on DNA mini barcodes of cytochrome C oxidase subunit I (COI) gene fragments for species identification of pangolins and pangolin scale products [10]. Bhattachan *et al*. identified species on three ascidian animals (Ascidiacea) from the northeast of China based on the mitochondrial COI sequences [11]. Zhao *et al.* identified the skeletal sample of a dead fin whale in the sea area of Shanghai with COI, control region, and cytochrome b (Cytb), and found it to be from *Balaenoptera omurai* [12]. The aforementioned species identifications were conducted by comparing the obtained DNA barcode sequences with those in the database. Therefore, the DNA database is the basis for this identification. The database determines the outcome of the appraisal. However, in reality erroneous sequences in the database are often encountered [13]. Hence, the institutions involved in identification can establish an independent and rigorously audited appraisal identification system based on classical taxonomy to provide accurate and rapid identification services for the involved items.

Sequence alignment is a crucial technique in bioinformatics, which involves arranging two or more sequences together and indicating the similarities between them. The use of multi-sequence alignment, such as MAFET, is generally recommended when understanding the similarities among a group of homologous sequences [14]. Double-sequence alignment is an essential method in bioinformatics. Scholars can infer connections between the two sequences in terms of structure, function, and evolution to a certain degree by comparing the similarities of these sequences [15]. Meanwhile, pairwise sequence alignment can also be used in species identification [16]. The alignment methods can be divided into global alignment and local alignment, represented by Needleman–Wunsch (NW) algorithm and Smith–Waterman (SW) algorithm, respectively [17-19]. In actual applications, the obtained sequences may show no remarkable similarity with the whole sequences in the database, despite being quite similar in some local regions. Therefore, the local alignment of pairwise sequences shows more extensive

applications than the global alignment. Local alignment algorithms mainly include accurate dynamic planning algorithms such as SW algorithm and heuristic acceleration algorithms. Heuristic algorithms, such as the Basic Local Alignment Search Tool (BLAST) and fast alignment, can enhance operational speed by simplifying the computational logic. For example, BLAST first adopts the strategy of seed expansion. Then, it performs extension alignment, which is suitable for rapid searches in large-scale databases because of a significant reduction in time complexity [20]. However, heuristic algorithms emphasize speed, which cannot ensure the optimal alignment between queries and basic sequences in most cases [21]. The score matrix is recursively calculated using heuristic algorithms, ensuring that the score drops to zero. Accordingly, the optimal local alignment between sequences can be ensured, thereby reducing the defects of heuristic algorithms [22].

The continuous development in recent years has led to the gradual application of machine learning techniques in sequence alignment. Ren *et al.* developed a no-reference and no-alignment machine learning method DeepVirFinder for identifying viral sequences in metagenomic data using deep learning [23]. The SW algorithm, which was proposed based on the idea of dynamic programming, was featured by the high sensitivity of matching. However, the demands on computational costs for execution time and storage have become intolerable with the large-scale growth of gene sequence databases in recent years [24]. Li *et al.* proposed a DNA double-sequence local alignment algorithm based on variable-length sub-sequence and concluded that the algorithm showed significant improvement in temporal and spatial complexity [25]. In combination with the characteristics of GPU parallel processing, Lin *et al.* improved the parallel steps of SW and proposed a novel GPU-accelerated SW scheme [26]. Oliveira *et al.* implemented the SW algorithm on a field-programmable gate array to accelerate the high-performance execution of the program at a lower utilization rate and a high efficiency [27]. Sajjadinasab *et al*. further optimized and analyzed the SW algorithm with vector extension [28]. Lall *et al*. proposed a pairwise DNA sequence alignment method based on deep reinforcement learning compatible with embedded edge devices. This method could perform pairwise DNA sequence alignment on independent edge devices, reducing the consumption of hard resources [29].

3 Research Contents

3.1 Database

The data in the present database were sourced from the biological specimens examined and identified by the College of Criminal Sciences, Nanjing Police University, China. The College possesses professional identification qualifications and benefits from the guidance of resident experts during the identification process. The data source is, therefore, safe and reliable, with high accuracy and a rapid update rate. Moreover, many new wild species have been constantly identified in the College, which can enrich the database and help in understanding and predicting trends in the smuggling of endangered wild species. This, in turn, provides useful guidance for the development direction of the endangered animal identification system in this study. The data in the present database have been reviewed and edited by relevant Chinese experts to ensure accuracy and avoid the potential unreliability of the data. Meanwhile, this domestic database is managed by Chinese organizations, which helps avoid the data security issues associated with foreign databases.

The data in the present database were obtained from specimens provided by the Department of Criminal Science and Technology at Nanjing Forest Police College after accurate species identification, which ensured both the accuracy and reliability of the source data. The data included the following types: (1) DNA sequence data, including 1904 mitochondrial DNA sequences related to 176 frequently involved animals; (2) textual data, recording in detail the taxonomic status, protection level, morphological characteristics, distribution area, and common trading categories of the species, thus providing intuitive references for police officers to quickly understand the species involved in the case; (3) image data storing typical image paths of various types of species and visually providing appearance characteristics. As an open-source database, the design follows the principle of openness and sharing, supporting external users in submitting new DNA sequences, supplementing text descriptions of species, or uploading high-quality image data via standard interfaces, thereby facilitating convenient data expansion. For example, when grassroots police officers encounter new species during law enforcement or qualified research institutions discover rare species, they can import the data into the database in a standardized format and incorporate them into the existing system. The open-source system not only ensures the dynamic updating of the database in accordance with law enforcement requirements and species research but also aggregates data from multiple resources to continuously expand species coverage.

3.2 Local Alignment Module

Local alignment algorithms aim to find similar fragments between two sequences rather than comparing all sequences. They are particularly suitable for handling damaged or incomplete DNA samples and finding the conserved regions in homologous genes. Two different sequences (p: ACGTC and q: CG) with short lengths were used in this study to facilitate the demonstration of algorithm logic. In order to overcome the memory challenges of large data volumes, the program can consider multiple memory optimizations. One of the more major methods i.e. using DNA character sequence to int array conversion method, i.e. functionsequence_to_int (seqs, dict_DNA= "ATCG") corresponds "ATCG" to "0123", this method can greatly save memory space and memory. In order to be able to fully store as well as manipulate the data, the program uses the@st.cache_datacache decorator to save the data reading time as well as unnecessary repetition, the key code is as follows:

```
Function: load DNA barcode data (file_path)
Input: file path (string type)_
Output: DNA barcode datasheet (DataFrame structure)
Process:
1. Decorator the function with a caching mechanism (@st.cache data
   decorator);
2. Check if the data corresponding to this file is already in the cache:
   ---If the cache hits, directly return the previously loaded data;
   ---If the cache misses, execute the following operations:
   a. Read the DNA barcode data from the specified Excel file;
   b. Store the results in the cache;
   c. Return the loaded datasheet.
```

The scoring rule is of crucial importance in local alignment algorithms. Specifically, *match* (matching score) represents the score when two characters are exactly the same, *mismatch* (mismatching score) represents the penalty score when two characters are different, and *gap* (gap penalty score) represents the penalty score when a character is inserted or deleted in the sequence. The selection of these parameters can directly impact the alignment results. Accordingly, the replacement scoring matrix can be obtained. The results were compared using the BLAST tool for parameter adjustment. When *match* = 15, *mismatch* = –15, and *gap* = –20, the alignment results align well with those obtained using BLAST, as shown in the replacement score matrix in Fig. 1. Some codes can be found in Supplementary Material 1.

<table>
<tr><td></td><td>A</td><td>G</td><td>C</td><td>T</td><td>–</td></tr>
<tr><td>A</td><td>15</td><td>-15</td><td>-15</td><td>-15</td><td rowspan="4">-20</td></tr>
<tr><td>G</td><td>-15</td><td>15</td><td>-15</td><td>-15</td></tr>
<tr><td>C</td><td>-15</td><td>-15</td><td>15</td><td>-15</td></tr>
<tr><td>T</td><td>-15</td><td>-15</td><td>-15</td><td>15</td></tr>
<tr><td>–</td><td colspan="4">-20</td><td></td></tr>
</table>

Fig. 1. Replacement scoring matrix

According to the formula depicted in Fig. 2, the scoring matrix between two sequences was calculated, where $s(i, j)$ denotes the score of the maximum similarity between the prefixes $q[1…j]$ and $p[1…j]$ obtained by replacing the scoring matrix, that is, the corresponding value in the scoring matrix; $w(i, j)$ denotes the score between the characters $q[i]$ and $p[j]$ according to the replacement scoring matrix. The following three conditions were considered during the calculation process.

Diagonal score: It represented the score when $q[i]$ matches or does not match $p[j]$, that is, $w(i, j)$ in the replacement scoring matrix.

Upper score: It represented the score when a space was inserted into the q sequence, that is, $s(i-1, j)$ – gap.

Left score: It represented the score when a space was inserted into the p sequence, that is, $s(i, j-1)$ – gap.

Finally, the score at each position was the maximum under the aforementioned three conditions. An example of a score matrix is illustrated in Fig. 3.

$$
\begin{aligned}
&s(0,0) = 0 \\
&s(0,j) = 0, 1 \le j \le m \\
&s(i,0) = 0, 1 \le i \le n \\
&s(i,j) = \max \begin{cases} 0 \\ s(i-1, j-1) + w(i,j) \\ s(i-1, j) + gap \\ s(i, j-1) + gap \end{cases}
\end{aligned}
$$

Figure 2. Local alignment formula

		0	1	2	3	4	5
			A	C	G	T	C
0		0	0	0	0	0	0
1	C	0	0	15	-5	0	15
2	G	0	0	-5	30	10	10

Fig. 3. Score matrix

The code related to matrix computation is as follows:

```
Input:
    - Sequence 1 — seq_1 (the sequence to be aligned)
    - Sequence 2---- ref_seq (the reference sequence)
    - Scoring parameters: match score match, mismatch score mismatch,
      gap penalty gap
Output:
    - The filled scoring matrix Score
Steps:
1. Iterate over the row index i, from the first row to the last one:
   a.  Start the parallel loop, iterator over the column index j (from
       the first column to the last one):
     i.  Calculate the score in the top-left diagonal direction
         score_diag:
          - If the current bases match, add match;
          - Otherwise, add mismatch;
     ii. Calculat ethe score in the top direction score_up, add gap;
     iii. Select the maximum value from the scores in all directions
          max_score；
   b.  After exiting the parallel loop, acquire GIL (global interpreter
       lock) to write to the score matrix:
       - Store max_score into Score[i, j];
```

Finally, the optimal alignment was obtained by bracketing from the lower right of the score matrix, as shown in Fig. 3. The position with the highest score was first located. Next, the search proceeded toward the upper left along the diagonal until encountering a position where the score no longer increased. During the backtracking

process, the alignment path (e.g., which characters were matched and which positions were inserted by spacing) was recorded. As shown in Fig. 3, the optimal alignment result after backtracking was indicated by a red arrow. The relevant codes are listed below:

```
Steps:
1. Loop over the row index i and the column index j:
   a.  If the value is from the diagonal direction:
     i. Add the dual-sequence base (in cases of identical sequences,
        label as '|');
     ii. Add the identical base indicator (label as | in case of
         identical sequences; otherwise, label as space)
   b.  If the value is from the top direction:
     i. Add the base to the sequence 1 and the space to the sequence 2;
   c.  If the value is from the left part:
     ii. Add the space to the sequence 1 and the base to the sequence 2;
```

Whether the species was the same according to local alignment results was determined by introducing the concepts of similarity and coverage, which were calculated as follows:

Similarity = (length of the same parts in the sequences during comparison/ total length of the reference sequence) × 100%

Coverage = (length of the alignment parts after comparison/ total length of the reference sequence) × 100%

The above formula calculates the code as follows:

```
Reverse the sequence 1 — seq_1 (the sequence to be aligned);
Reverse the sequence 2 ref_seq (the reference sequence);
Reverse the match line ;
Calculate the similarity degree;
Calculate the coverage;
```

The total length of the comparison area should be the value in the local alignment result rather than in the global alignment result. As local alignment focuses on highly similar local regions in the sequences, the local alignment result was used while calculating similarity and coverage. When the similarity after comparison exceeded 98%, it was considered that the sequences originated from the same species. When the coverage exceeded 95%, most regions in the reference sequence or target sequence had corresponding parts in the research sequence fragment. It implied that the detection of the target sequences was comprehensive, with sufficient effective information, thus confirming that these two sequences were from the same species to a large extent. If the coverage was lower than 80%, the matching area between the research sequence fragment and the target sequence was relatively small, with a significant difference. The reasons for this difference should be further analyzed to determine whether the sequences originated from different species, the samples were contaminated, or the sequencing was mistaken.

Comparison sequence algorithm is a very critical function in the species identification system, the calculation speed of the algorithm as well as the correctness of the comparison is directly related to the effect of the comparison and the advantages and disadvantages of the comparison time. Therefore, in the process of designing this system, the selection of the underlying logic into a fast language is an important system implementation method.

When dealing with large amounts of data, especially when executing codes involving a large number of matrix computations, Python can encounter a series of problems, mainly including slow operating speed and high memory consumption. Therefore, choosing C-Extension for Python (Cython) for optimizing and accelerating NumPy matrix calculation has become an effective solution. Cython can significantly enhance computing performance owing to the combination between Python codes and the characteristics of the C language. The use of Cython to accelerate the codes of NumPy has a significant advantage: only a small number of Cython-specific annotations need to be added to the existing Python project, without completely rewriting the codes. Hence, not only can the flexibility and readability of the Python language can be retained, but also the execution efficiency of codes approaches the level of the C language. Cython's function align_sequence (alignsequences) can compare DNA sequences. The function is an improved dynamic programming algorithm, which has advantages in the comparison of large-capacity gene sequence data, and is more suitable for the processing of large amounts of DNA sequence data. By utilizing dynamic programming methods and introducing sequence matching into the

algorithm, which can more comprehensively consider aligning all possible paths and traversing the species in the computational species library, the results of the matching will be more accurate. So the introduction of Cython can remarkably enhance the execution efficiency of big data processing without sacrificing code readability and development speed. The full code to accelerate the calculation using Cython is below:

```
Steps:
Function: align_sequence (Sequence 1 and Sequence 2)
Input:
    Sequence 1: DNA sequence in the format of string such as 'ATCG'
    Sequence 2: DNA sequence in the format of string
Output:
    A tuple containing the alignment results including score or the alignment path
Steps:
1. Encode Sequence 1 into an integer array (call sequence_to_int);
2. Perform local alignment with Smith-Waterman algorithm:
   - Input: encoded sequence 1, encoded sequence 2 and scoring matrix;
   - Call the underlying Cython function for alignment computation;
3. Parse the alignment matrix and output the alignment score or alignment path;
4. Return the final alignment results
Function: sequence_to_int (DNA string)
Input:
    DNA string (for example: ATCG)
Output:
    DNA sequence represented by integer arrays (for exmaple: [0, 1, 2, 3])
Steps:
1. Convert the input DNA string into byte array;
2. Perform character-to-character transformation according to the mapping relation
   betwee character and number (A→0, T→1, C→2, G→3);
3. Create and return the corresponding integer array;
```

To better demonstrate the system framework, the system workflow from the input of sequences to the output of results had been displayed in Fig. 4.

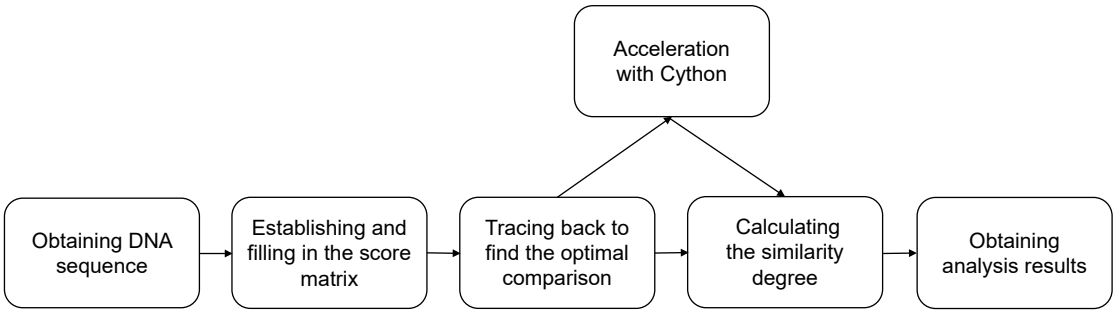


Fig. 4. System's workflow from sequence input to result output

3.3 Visualization Platform Module

This study used a Streamlit framework for visual display to ensure the convenience of a DNA comparison platform for species identification and the smoothness of code operation. Accordingly, the local comparison results and the detailed information of the relevant species were clearly displayed to facilitate the analysis and judgment of intelligence by the public security organizations.

Streamlit is an open-source framework based on Python, especially designed for simplifying the rapid construction of interactive Web applications. Streamlit is particularly applicable to data visualization, display of machine learning models, and development of deployed dashboards owing to the integration of various

functional components including buttons, charts, text input boxes, and file uploaders. Accordingly, users can directly embed these interactive elements in the algorithm codes without deeply modifying Python code files. Further, users can immediately view the results by inputting the sequences to be compared with a simple text box and clicking the “run” operation. This feature can significantly lower the usage threshold of DNA comparison technology so that grassroots policy officers, even without any professional programming background, can easily apply the tool efficiently.

The display interface of the visualization platform module includes the input box, the operation button, and the result display box. When the user clicks on the input box, they can input the local DNA sequence to be detected; when the user clicks on the operation button, it can transmit the local DNA sequence to be detected to the local alignment module and drive local alignment. The result display box is used for displaying the species and detailed information about the endangered animals to be detected.

The present results can be deployed on a public network. Real-time data upload and comparison functions can remarkably enhance the case-handling efficiency of public security organizations. The case-handling personnel can upload the DNA data of specimens to the system through a public network and immediately obtain the comparison results. Public security organizations can use the instant feedback mechanism to rapidly identify the species to which the involved items belong, thus significantly shortening the time for solving cases.

4 Results

For testing, mitochondrial DNA sequences from two sources were selected: one animal included in the existing four databases and one animal not included in the database. This test was designed to display the results of the DNA comparison platform for species identification.

First, the py. file operated on the terminal. It found and ran the code to open the visualization platform module. The local DNA sequence to be detected was input for alignment, and the button for obtaining sequences and comparing (i.e., the operation button) was clicked. The local alignment module was driven to operate. The local alignment results, containing 4 known animals and a negative control, are shown in Fig. 5.

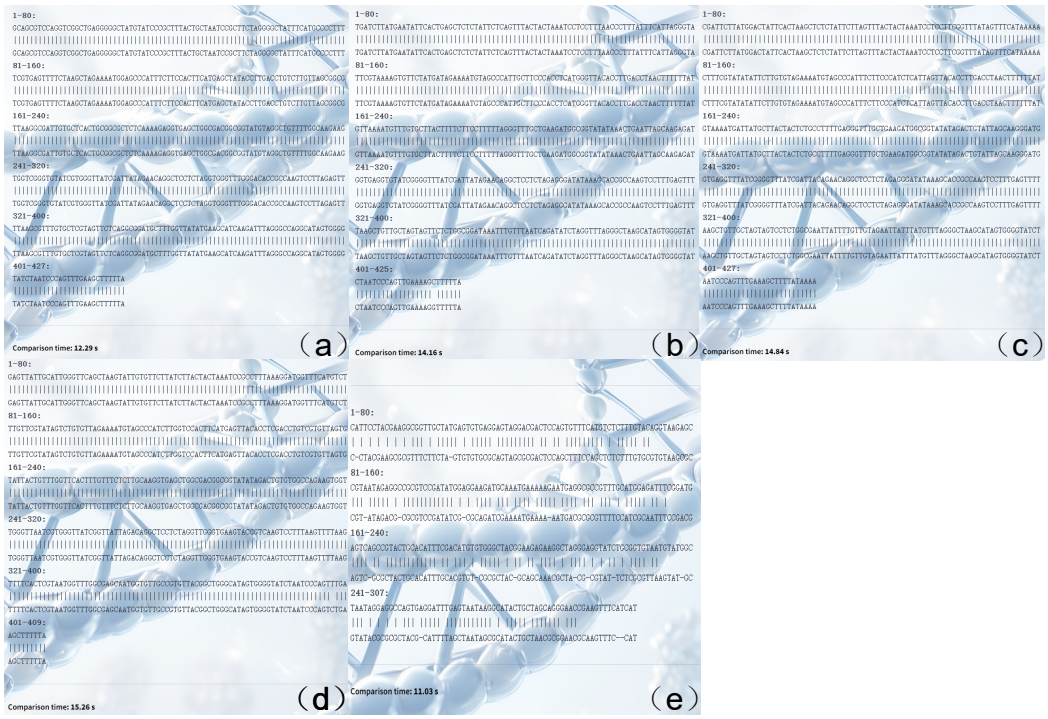


Fig. 5. Local alignment results of five DNA sequences: (a) *S. orientalis*; (b) *P. larvata*; (c) *N. procyonoides*; (d) *P. regius* (e) Negative test

The results of the comparison between the reference sequence and the target sequence are presented in a trilinear format. The trilinear format visualizess the reference sequence, alignment identifier and target sequence with the help of alignment. At the same time, the comparison results take the form of truncation, 80 bases for a truncation result, the truncation result shows the position number of the starting point and the end point of the sequence comparison, and the aligned bases are marked with "|" sign, while the unaligned positions are kept in the blank space, which is convenient for users to view. The three-line format visually presents the three layers of information: reference sequence, alignment identifier and target sequence in the form of alignment, so that users can clearly capture the differences without comparing character by character. When facing long sequence data, the layout logic of 80 character segments avoids overcrowding of information, makes manual verification and screenshot retention more convenient, and take into account both professional analysis and practical operation experience. The program can be used for both professional analysis and practical operation experience.

The visualization module not only displays the local comparison results and the similarity percentage of species in real time but also provides key information about species including the taxonomic status, protection level, morphological characteristics, frequently traded categories, typical forms of illicit exploitation, value, and distribution value (Fig. 6). For example, the *Nyctereutes procyonoides* from wild populations is included in the List of National Wildlife under Key Protection and is classified as a Grade II wildlife under key protection. The value of a single *N. procyonoides* from wild populations is clearly defined as 4,000 RMB. The common types of wildlife traded are dead bodies and skins, and the common forms of utilization are illegal breeding and wild capture. The results provide public security officers a direct understanding of the relevant information about involved wildlife, especially the value and the protection level. This helps officers understand the true character of the case, guides their investigative strategies, and streamlines the investigation process for more effective law enforcement.

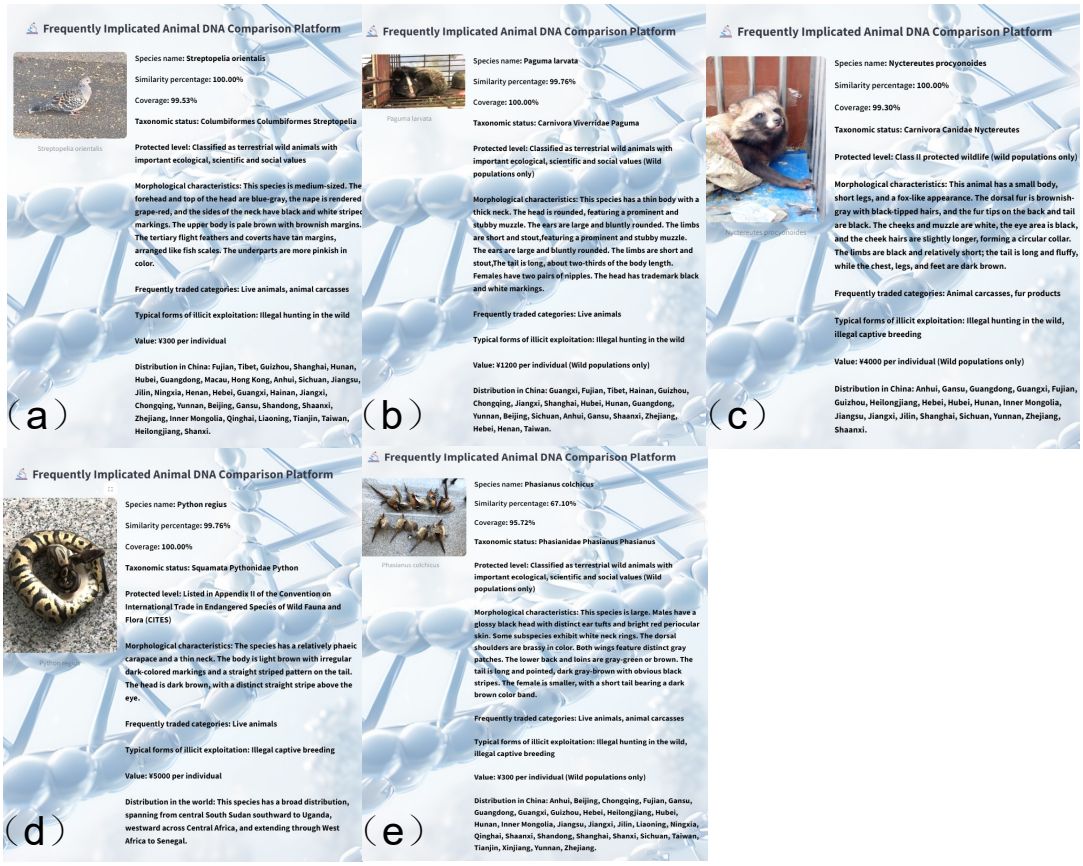


Fig. 6. Corresponding species information according to the alignment results: (a) *S. orientalis*; (b) *P. larvata*; (c) *N. procyonoides* (d) *P. regius*; (e) Negative test

5 Conclusions

Wildlife specimens with missing macroscopic morphology are commonly found in wildlife-related legal cases. Identifying the species to which they belong is an important step in combating such cases. In this study, SW local alignment algorithm was used, and the algorithm parameters were optimized in combination with the BLAST tool. Next, the matrix computation was accelerated with Cython. Finally, the Streamlit framework was integrated to establish a wildlife DNA identification system so as to achieve accurate species identification of various types of specimens. Meanwhile, a user-friendly DNA comparison platform for species identification was constructed to ensure both accuracy and safety of identification.

The specific advantages of this study were as follows.

(1) Establishment of the model from the perspective of public security practices

The present study focused on public security law enforcement practices, emphasizing both the practicality and specificity of the model. It aimed to address DNA identification problems in challenging biological specimens in cases involving endangered species. For this, the study integrated the existing models and algorithms to establish a specialized DNA identification model designed to overcome the challenges in actual public security tasks and ensure its practical application in frontline law enforcement.

(2) Integration of SW algorithm and visualization platform

This study comprehensively used a local alignment algorithm, visualization platform, and database technology to enhance both the accuracy and operational convenience of DNA identification. The model emphasized mining similar fragments between two sequences using the local alignment algorithm. This was especially suitable for dealing with damaged or incomplete DNA specimens, thereby significantly enhancing DNA identifiability in such cases. In addition, the visualization platform module was developed based on the Streamlit framework. Therefore, even non-professional users such as grassroots police officers can easily apply DNA comparison technology, thus significantly lowering the technical threshold. Meanwhile, the findings of the present study contribute to establishing and perfecting the wildlife species protection database and provide strong data support for ecological protection.

(3) Safe and accurate data source

Some actual risks exist in foreign databases such as the inability to guarantee data quality and the lack of supervision in the database background. The databases used in this study were sourced from the biological specimens identified by the College of Criminal Science and Technology, Nanjing Police University, with a safe and reliable source. Further, the database can be integrated with other police colleges and research institutions to better support law enforcement operations.

(4) Enhancement of DNA comparison rate and rapid and accurate species identification

The present study significantly enhanced the operating rate while maintaining extremely high comparison accuracy, thereby displaying remarkable advantages in speed and precision. These advantages not only reflect in daily detection work but also play a crucial role at some key moments such as solving urgent cases.

(5) Visual display of species-related information

The visualization module can clearly display the image of the species with the highest matching similarity along with related information. The comprehensive display function greatly empowered law enforcement officers, enabling them to rapidly and comprehensively grasp the basic information and value evaluation of target species, thereby effectively guiding case investigation direction. The model can be key in combating illegal trade and preserving biodiversity.

This research is oriented to the practical needs of public security, through the innovative integration of SW local matching algorithm and visualization platform, relying on the safe and reliable biological samples identification data source of the School of Criminal Science and Technology of Nanjing Police University, it constructs a DNA matching model and identification platform with high efficiency, accuracy and ease of use. This system can provide solid technical support and data guarantee for the public security department to fight against wildlife crimes and vigorously promote wildlife protection and law enforcement.

Acknowledgement

This study was financially supported by the Special Fund for Basic Research of Central Universities (Grand No.: LGZD202503).

References

[1] Z.G. Jiang, Values and ecological service functions of wildlife, Acta Ecological Sinica 21(11)(2001) 1909-1917. https://doi.org/10.3321/j.issn:1000-0933.2001.11.022

[2] Z.X. Yang, Innovation of wildlife prohibition on eating system—from the analysis of the threefold conception of wild animals, Dongyue Tribune 42(1)(2021) 172-184. https://doi.org/10.15981/j.cnki.dongyueluncong.2021.01.017

[3] D.W. Liu, S.L. Hou, Y.W. Zhou, Y.L. Fei, J. He, DNA barcoding of Artiodactyla in cases involved in wildlife for species identification, Journal of Nanjing Forestry University: Natural Science Edition 44(6)(2020) 27-32. https://doi.org/10.3969/j.issn.1000-2006.202004013

[4] D.W. Liu, G.D. Ying, Y.W. Zhou, Y.L. Fei, W.D Wu, C.P. Xie, S.L. Hou, Identification of a bird remain in a case of destroying wildlife resources based on the COI and 16S rRNA genes, Chinese Journal of Wildlife 43(3)(2022) 715-724. https://doi.org/10.12375/ysdwxb.20220315

[5] M. Staats, A.J. Arulandhu, B. Gravendeel, A. Holst-Jensen, I. Scholtens, T. Peelen, T.W. Prins, E. Kok, Advances in DNA metabarcoding for food and wildlife forensic species identification, Analytical and bioanalytical chemistry 408(17)(2016) 4615-4630. https://doi.org/10.1007/s00216-016-9595-8

[6] B. Dubey, P.R. Meganathan, I. Haque, DNA mini-barcoding: an approach for forensic identification of some endangered Indian snake species, Forensic Science International: Genetics 5(3)(2011) 181-184. https://doi.org/10.1016/j.fsigen.2010.03.001

[7] Y. Shen, X. Chen, R. Murphy, Assessing DNA barcoding as a tool for species identification and data quality control, Plos One 8(2)(2013) e57125. https://doi.org/10.1371/journal.pone.0057125

[8] P.D.N. Hebert, A. Cywinska, S.L. Ball, J.R. deWaard, Biological identifications through DNA barcodes, Proceedings of the Royal Society B 270(1512) (2003) 313-321. https://doi.org/10.1098/rspb.2002.2218

[9] Q.Q. Yang, S. Liu, X. Yu, Research progress on DNA barcoding analysis methods, Chinese Journal of Applied Ecology 29(3)(2018) 1006-1014. https://doi.org/10.13287/j.1001-9332.201803.032

[10] X. Xie, H. Ye, C. Xuan, C. Li, F. Li, E. Tian, Z. Chao, DNA Mini-Barcodes, a Potential Weapon for Conservation and Combating Illegal Trade of Pangolin, Tropical Conservation Science 14(2021) 1-10. https://doi.org/10.1177/19400829211017361

[11] P. Bhattachan, R. Qiao, B. Dong, Identification and population genetic comparison of three ascidian species based on mtDNA sequences, Ecology and evolution 10(8)(2020) 3758-3768. https://doi.org/10.1002/ece3.6171

[12] Z.P. Zhao, J.G. XiangYu, X. He, C.W. Song, K. Shan, W.L. Liu, Identification of Balaenoptera omurai (Balaenopteridae) from Shanghai Seas by using mtDNA Gene Sequences, Chinese Journal of Zoology 54(4)(2019) 493-500. https://doi.org/10.13859/j.cjz.201904006

[13] P. Korning, S. Hebsgaard, P. Rouzé, S. Brunak, Cleaning the GenBank Arabidopsis thaliana data set, Nucleic Acids Research 24(2)(1996) 316-320. https://doi.org/10.1093/nar/24.2.316

[14] K. Katoh, G. Asimenos, H. Toh, Multiple Alignment of DNA Sequences with MAFF, In: D. Posada, (Eds.) Bioinformatics for DNA Sequence Analysis. Methods in Molecular Biology, vol 537, Humana Press, 2009 (pp. 39-64). https://doi.org/10.1007/978-1-59745-251-9_3

[15] E. Kirkness, V. Bafna, A. Halpern, S. Levy, K. Remington, D. Rusch, A. Delcher, M. Pop, W. Wang, C. Fraser, J. Venter, The dog genome: survey sequencing and comparative analysis, Science 301(5641)(2003) 1898-1903. https://doi.org/10.1126/science.1086432

[16] H. Yang, J. Lou, X. Lin, J. Lin, Discovery and description of male individuals of Nilobezzia bamenwana based on DNA barcods (Diptera: Ceratopogonidae), Chinese Journal of Vector Biology and Control 36(1)(2025) 94-98. https://doi.org/10.11853/j.issn.1003.8280.2025.01.017

[17] Y. Liu, B Schmidt, GSWABE: faster GPU-accelerated sequence alignment with optimal alignment retrieval for short dna sequences, Concurrency and Computation: Practice and Experience 27(4) (2015) 958-972. https://doi.org/10.1002/cpe.3371

[18] S. Needleman, C. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology 48(3)(1970) 443-453. https://doi.org/10.1016/0022-2836(70)90057-4

[19] T. Smith, M. Waterman, Identification of common molecular subsequences, Journal of Molecular Biology 147(1)(1981) 195-197. https://doi.org/10.1016/0022-2836(81)90087-5

[20] Y. Jeong, S. Shin, Probabilistic Approach Processing Scheme Based on BLAST for Improving Search Speed of Bioinformatics, Wireless Personal Communications 105(2)(2019) 405-426.

https://doi.org/10.1007/s11277-018-5955-3
[21] D. Eric, T. Nicholas, A. Theophilus, Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA), Journal of Bioinformatics and Sequence Analysis 6(1)(2014) 1-6.
https://doi.org/10.5897/IJBC2013.0086
[22] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics 26(5)(2010) 589-595.
https://doi.org/10.1093/bioinformatics/btp698
[23] J. Ren, K. Song, C. Deng, N.A. Ahlgren, J.A. Fuhrman, Y. Li, X. Xie, R. Poplin, F. Z. Sun, Identifying viruses from metagenomic data using deep learning, Quantitative Biology 8(1)(2020) 64-77.
https://doi.org/10.1007/s40484-019-0187-4
[24] F. Xia, D. Zou, L. Lu, X. Man, C. Zhang, FPGASW: Accelerating Large-Scale Smith–Waterman Sequence Alignment Application with Backtracking on FPGA Linear Systolic Array, Interdisciplinary Sciences: Computational Life Sciences 10(1)(2018) 176-188.
https://doi.org/10.3321/10.1007/s12539-017-0225-8
[25] X. Li, H. Guo, VLSSA: A Pairwise Local Alignment for DNA Sequences based on Variable-Length Segments Splice, Journal of Guangxi Normal University (Natural Science Edition) 26(4)(2008) 53-57.
https://doi.org/10.3321/10.3969/j.issn.1001-6600.2008.04.014
[26] M. Lin, Y. Zhong, J. Lin, Parallel Smith-Waterman algorithm based on GPU, Journal of Fujian Agriculture and Forestry University (Natural Science Edition) 44(4)(2015) 442-448.
https://doi.org/10.3321/10.13323/j.cnki.j.fafu(nat.sci.).2015.04.019
[27] F. Oliveira, L. Dias, M. Fernandes, Proposal of Smith-Waterman algorithm on FPGA to accelerate the forward and backtracking steps, PLoS ONE 17(6)(2022) e0254736.
https://doi.org/10.3321/10.1371/journal.pone.0254736
[28] R. Sajjadinasab, H. Rastaghi, H. Shahzad, S. Arora, U. Drepper, M. Herbordt, Further Optimizations and Analysis of Smith-Waterman with Vector Extensions, in: Proc. 2024 IEEE International Parallel and Distributed Processing Symposium Workshops, 2024.
https://doi.org/10.3321/10.1109/IPDPSW63119.2024.00113
[29] A. Lall, S. Tallur, Deep reinforcement learning-based pairwise DNA sequence alignment method compatible with embedded edge devices, Scientific Reports 13(2023) 2773.
https://doi.org/10.3321/10.1038/s41598-023-29277-6