# Sequence Processing Software User Manual

## Table of Contents

# 1. Introduction

Welcome to the **Sequence Processing Software**, an integrated GUI tool for conserved region analysis, sequence export, difference and identity calculations, length statistics, NCBI data retrieval, sequence optimization, primer design, and 1D/2D code (barcode/QR code) generation. This manual provides step- by- step instructions for installation, interface navigation, module operation, and troubleshooting to help you get up and running quickly and efficiently.

# 2. System Requirements and Installation

1. **Operating Systems**: Windows 10/11, macOS 10.15+ , Linux (with PyQt5 support)
2. **Python Version**: Python 3.8 or higher
3. **Required Libraries**:
    - PyQt5
    - Biopython
    - numpy
    - pandas
    - scikit- learn
    - seaborn
    - matplotlib
    - python- docx
    - qrcode
    - pillow
4. **Installation**:

    ```
    pip install PyQt5 biopython numpy pandas scikit-learn seaborn matplotlib python-docx qrcode pillow
    ```

5. **Program Files**:
    - Place `module9.py` and `软件图标.png` in the same folder. Ensure the icon file is accessible by the application.

# 3. Launching the Application

1. Open a terminal or command prompt.
2. Change directory:

    ```
    cd /path/to/software/folder
    ```

3. Run the program:

    ```
    python module9.py
    ```

4. The main window will appear, ready for input.

# 4. Interface Overview

- **Top Bar**: Language selector (English/中文) on the left and Module selector (Module 1–9) in the center.
- **Main Panel**: Displays input controls for the selected module using a stacked widget.
- **Run Button**: Each module panel has its own “Run” button at the bottom.
- **Styling**: Gradient background, Arial font, and a layered color scheme for controls.

# 5. Detailed Module Guide

Each module’s section below describes its purpose, required inputs, usage steps, and outputs.

## 5.1 Module 1: Conserved Region Analysis

**Purpose**: Identify highly conserved segments across multiple FASTA sequences.

**Inputs**:

- **Input Folder**: Directory containing FASTA files (`.fasta`/`.fas`).
- **Output Folder**: Directory for saving results.
- **Minimum Length (bp)**: Shortest conserved segment to report.
- **Maximum Length (bp)**: Longest conserved segment to report.
- **Gap Threshold (0.0–1.0)**: Maximum allowed gap proportion per column (default 0.1).
- **Region Threshold (0.0–1.0)**: Conservation frequency cutoff per position (default 0.8).

**Steps**:

1. Click **Browse…** next to “Select folder containing FASTA files” and choose your input folder.
2. Click **Browse…** next to “Select folder to save Conserved Region results” and choose an output folder.
3. Enter numeric values for minimum length, maximum length, gap threshold, and region threshold.
4. Click **Run Module 1**.

**Outputs**:

- For each FASTA file, a `conserved_sequences_<filename>.txt` file listing each conserved fragment, its start position, validation flag, and overall SVM cross‑ validation accuracy.

## 5.2 Module 2: Sequence Export

**Purpose**: Split a combined FASTA text file into test and conserved segments for downstream analysis.

**Inputs**:

- **Input FASTA File**: A .txt file formatted in FASTA style.
- **Output Folder**: Directory for saving split segments.
- **Split Index**: Integer index (e.g. 1) indicating how many sequences are "test" sequences; remaining are conserved.

**Steps**:

1. Browse and select the input FASTA .txt file.
2. Choose an output folder.
3. Enter the split index.
4. Click **Run Module 2**.

**Outputs**:

- A set of _segment_<n>.txt files, where each file contains all sequences for that segment.

## 5.3 Module 3: Difference Calculation

**Purpose**: Compute average nucleotide- level difference between test sequences and species segments, output distribution tables and heatmaps.

**Inputs**:

- **Test Sequence File**: The .txt file with test sequences.
- **Species Folder**: Folder containing _segment_<n>.txt files.
- **Output Excel File**: Path for saving results (.xlsx).

**Steps**:

1. Select the test sequence .txt file.
2. Select the species segment folder.
3. Specify an output .xlsx file path.
4. Click **Run Module 3**.

**Outputs**:

1. **Excel Workbook** with three sheets:
    - Avg Nuc Diff (%): Average difference per segment.
    - Species Diff (%) (Prop): Proportion distribution for thresholds 1–99%.
    - Species Diff (%) (Num): Count distribution for bases thresholds.

2. **Heatmaps**: Saved as <prefix>_species_diff_prop_heatmap.svg/png and <prefix>_species_diff_num_heatmap.svg/png.

## 5.4 Module 4: Identity Calculation

**Purpose**: Compute average nucleotide-level identity against species consensus, with distribution details and heatmaps.

**Inputs**: Same as Module 3.

**Steps**:

- Provide test sequence and species folder paths.
- Enter output Excel path.
- Click **Run Module 4**.

**Outputs**:

1. **Excel Workbook** with:
   - Avg Nuc Identity (%)
   - Species Identity (%) (Prop)
   - Species Identity (%) (Num)
2. **Heatmaps**: <prefix>_species_identity_prop_heatmap.svg/png, <prefix>_species_identity_num_heatmap.svg/png.

## 5.5 Module 5: Sequence Length Calculation

**Purpose**: Calculate the length of each sequence in a FASTA file.

**Inputs**:

- **Input FASTA File**: .fasta or .fas.
- **Output Excel File**: Path to .xlsx.

**Steps**:

1. Browse for the FASTA file.
2. Specify output Excel path.
3. Click **Run Module 5**.

**Outputs**:

1. An Excel sheet with columns: Title and Length.

## 5.6 Module 6-1: Fetch Genome from NCBI

**Purpose**: Batch-download complete genome FASTA for given accession numbers via Entrez.

**Inputs**:

- **Accession List File**: .txt file listing one accession number per line.
- **Output Folder**: Destination for .fasta files.

**Steps**:

1. Select the accession list file.
2. Choose an output folder.
3. Click **Run Module 6-1**.

**Outputs**:

- Individual <accession>.fasta files for each number in the list.

## 5.7 Module 6-2: Fetch CDS from NCBI

**Purpose**: Batch- download GenBank records, extract CDS features, and save them as text.

**Inputs**: Same as Module 6-1.

**Steps**:

1. Provide the accession list file.
2. Select an output directory.
3. Click **Run Module 6-2**.

**Outputs**:

- <accession>_cds.txt files containing CDS feature entries.

## 5.8 Module 7: Sequence Optimization (SeqRefine)

**Purpose**: Remove degenerate bases and optionally rename sequence titles in FASTA files.

**Inputs**:

1. **Input Folder**: Contains .fasta or .fas files.
2. **Modify Titles**: Check to use a custom title.
3. **New Title**: Text (without >) if modifying titles.
4. **Rename to Filename**: Check to rename each sequence header to its file basename.

**Steps**:

- Choose the FASTA folder.
- (Optional) Tick **Modify Titles** and enter a new title.

- (Optional) Tick **Rename to Filename**.
- Click **Run Module 7**.

**Outputs**:

- FASTA files overwritten in place: degenerate bases removed and headers updated accordingly.

## 5.9 Module 8: Primer Design

**Purpose**: Automatically design forward and reverse primers, calculate GC content and melting temperature.

**Inputs**:

- **Input FASTA File**: .fasta or .fas.
- **Output Excel File**: Path for .xlsx.
- **Primer Length**: Integer (default 18 bp).

**Steps**:

- Browse to select the input FASTA.
- Specify output Excel path.
- Enter desired primer length.
- Click **Run Module 8**.

**Outputs**:

1. Excel table listing region names, forward/reverse primer sequences, GC%, and Tm.

## 5.10 Module 9: 1D/2D Code Generation

**Purpose**: Convert text or DOCX content to stylized barcodes/QR codes and generate a Word document with base- replacements.

**Inputs**:

- **Input File**: .txt or .docx.
- **Output File**: .docx for final document.

**Steps**:

1. Select the input text or Word file.
2. Set the output .docx path.
3. Click **Run Module 9**.

**Outputs**:

- Stylized PNG barcode/QR images saved to the output folder.
- A Word document containing replaced base symbols and embedded codes.

# 6. Common Operations

## 6.1 Language Switching

Use the top-left combo box to switch between **English** and 中文—all UI text updates instantly.

## 6.2 Module Switching

Select any module (1–9) from the top-center combo box; the main panel will display that module's inputs.

## 6.3 File and Folder Browsing

Click each panel's **Browse…** button to open file/folder dialogs. Selected paths populate the adjacent text fields.

# 7. Output Description

- All generated files (Excel, TXT, PNG, SVG, DOCX) appear in user-specified directories.
- Filenames include module-specific prefixes/suffixes.
- Heatmaps: files ending with _prop_heatmap (percentage) or _num_heatmap (count).

# 8. Troubleshooting

1. **Missing Dependencies**: Ensure all Python packages are installed.
2. **Invalid Paths**: Verify file/folder existence and read/write permissions.
3. **Excel Export Errors**: Close open Excel instances or change the output filename.
4. **NCBI Fetch Failures**: Check Internet connectivity and set Entrez.email in the code.
5. **GUI Unresponsive**: Restart the application; confirm Python and PyQt5 compatibility.

# 9. Technical Support and Feedback

For questions or feature requests, contact our support team:

- **Email**: gs2022@hnu.edu.cn

Thank you for using the Sequence Processing Software. We wish you success in your research!