SPDE: A Multi-functional Software for Sequence Processing and Data Extraction

Dong Xu^{abc}, Zhuchou Lu^{ac}, Kangming Jin^{ac}, Wenmin Qiu^{ac}, Guirong Qiao^{ac}, Xiaojiao Han^{ac*},
Renyong Zhuo^{ac*}

a, State Key Laboratory of Tree Genetics and Breeding, Chinese Academy of Forestry, Beijing,
100091, P. R. China

b, Nanjing forestry university, Nanjing, 210037, Jiangsu, P. R. China

c, Key Laboratory of Tree Breeding of Zhejiang Province, The Research Institute of Subtropical of
Forestry, Chinese Academy of Forestry, Hangzhou, Zhejiang, 311400, P. R. China

Abstract:

Efficiently extracting information from biological big data can be a huge challenge for people (especially those who lack programming skills). We developed Sequence Processing and Data Extraction (SPDE) as an integrated tool for sequence processing and data extraction for gene family and omics analyses. Currently, SPDE has seven modules comprising 100 basic functions that range from single gene processing (e.g., translation, reverse complement, and primer design) to genome information extraction. All SPDE functions can be used without the need for programming or command lines. The SPDE interface has enough prompt information to help users run SPDE without barriers. In addition to its own functions, SPDE also incorporates the publicly available analyses tools (such as, NCBI-blast, HMMER, Primer3 and SAMtools), thereby making SPDE a comprehensive bioinformatics platform for big biological data analysis.

Availability: SPDE was built using Python and can be run on 32-bit, 64-bit Windows and macOS systems. It is an open-source software that can be downloaded from <https://github.com/simon19891216/SPDEv1.2.git>.

Contact: xudongzhuanyong@163.com

Introduction

Extracting and organizing valid information from biological big data are the two foundations of bioinformatics analyses. In recent years, the rapid accumulation of genome data^{1,2} indicates that more efficient and reliable tools are needed to extract and organize the resultant biological big data. Although a lot of bioinformatic software^{3,4} have been developed, many are distributed in Linux systems or require users to have a knowledge of programming language. This poses huge challenges, especially for Windows users, and limits the ability of researchers to extract useful information from big data. Therefore, data extraction software is urgently needed to make up for this deficiency.

Gene family and omics analyses are two import aspects in bioinformatics analysis^{5,6}. Extracting sequences (e.g., full-length gene sequences, promoter sequences and untranslated regions sequences (UTR)) from big data is the basis of gene family analysis⁷. The coding sequences (CDS) and protein sequences can be obtained easily (as the partial results of the genome sequencing), for example from Phytozome, the Plant Genomics Resource (<https://phytozome.jgi.doe.gov/pz/portal.html>). However, obtaining promoter sequences, full-length gene sequences and UTR can be difficult and time-consuming, especially when a gene family has more than 100 members. The extraction of effective information is the premise of omics analyses. For example, in transcriptome analysis, the main focus is often on genes that are differentially expressed under different experimental treatments⁸. It is also important to generate data in a file format that can be input to downstream software. For example, Circos⁹, which is a visualization tool for genomic information, has strict requirements for the format of input files (http://circos.ca/documentation/tutorials/configuration/data_files/). Although many Circos tutorials are publicly available, very few software can be used directly to generate files in the correct format.

The analysis software, NCBI blast+ (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>), ClustalW¹⁰, Muscle¹¹, MAFFT¹², HMMER (<http://hmmer.org/>), MCSanX³, and Primer3¹³, are important for sequence alignment, detecting homologous genes, collinearity analysis, and primer design; however, they can be used only through programming or command lines. Moreover, the input file of Primer3 (not the online version) needs to be in a specific format. Therefore, to utilize them skillfully and effectively can be a challenge for researchers, especially for wet-lab biologists.

To address the practical problems in sequence processing and data extraction, we developed

Sequence Processing and Data Extraction (SPDE) as an integrated tool for sequence processing and data extraction. The main goal was to minimize the amount of time required to deal with big data files. Thus, except for the first module, the other six modules in SPDE use batch processing. For example, except for the required files, users need to provide only the gene identifiers (ids) to extract the full-length gene sequences and their promoter sequences from genomic data. In addition, we simulated the awk, sed, and other Linux commands to develop a new extraction engine. To extract information from a big data file, users need to provide only rows/columns ids or other filter criteria, then the module will write required the rows or columns into a new file. The interfaces of NCBI blast+, ClustalW, Muscle, MAFFT, HMMER, MCScanX, and Primer3 also have been integrated in SPDE. The aim was to provide a tool that allows researchers to focus on the meaning of the bioinformatics data rather than on how to obtain relevant information.

Methods and Features

Overview of SPDE functions

The modules and some functions of SPDE are shown in Fig. 1. Currently, SPDE contains nine modules comprising 100 functions that range from processing a single gene sequence to extracting genomic information.

The functions of ‘Gene’ module can help biologists mine promoter sequences, full-length sequences and UTR sequences from genomic files, based on gene ID. Then, ‘design primers’ can be used for generating primers. Meanwhile, sequence processing (such as, reverse complement and translation) was also set in this module (Supplementary video 1). In the ‘Genes’ module, SPDE performs batch processing of multi-gene files, such as batch primer design and translation (Supplementary video 2). In the ‘Alignment/HMMER’ module, we developed a user-friendly interface for Blast, HMMERsearch, ClustalW, Muscle, MAFFT (for Windows users) and MCScanX. The 13 functions in this module can be used to complete alignment, collinearity analysis, and homologous genes identification (Supplementary videos 3.1 and 3.2). The ‘Genome’ module can be used to compile statistics of genome information, such as counting the numbers and positions of genes on each chromosome, and to extract promoter and full-length gene sequences in batches (Supplementary video 4). The ‘For circos’ module automatically generates files that are formatted for use in Circos software from input files in a different format (Supplementary videos 5.1 and 5.2). The ‘File operation’ module allow users to replace and delete the content of input files, and to extract,

rearrange, and cut file contents, according to their requirements. All of these functions can be achieved by setting simple parameters (Supplementary videos 6.1 and 6.2). Although strict file formats have been specified for bioinformatics applications (such as fasta and General Feature Format (gff)), we found subtle differences in the format of files, even among files of the same format type. Unfortunately, these subtle differences seriously affected the processing efficiency and accuracy of SPDE. Therefore, we developed the ‘Formatting’ module so that users can format their files according to the prompts of the SPDE interface, before using SPDE for their analyses. The last module is ‘Format_conversion and others’. This module supports conversion between files in different formats (Supplementary videos 7).

Module name	Some functions	Module name	Some functions
Gene	extract promoter/full length sequences from genome files base on gene ID	File operation	extract or rearrange contents of your files based on simple parameters
Genes	extract genes based on IDs, design primers in batch	Format	formatting files to ensure the accuracy of the results and save time
Alignment/hmmer	blast/hmmersearch/MCScanX/mafft/Clustalw/muscle	Format conversion	conversion between GBK, GFF, GTF, SAM, BAM, VCF, PHYLIP
Genome	extract promoter/full-length sequences in batch based on gene IDs	Help	users can obtain the information about how to use SPDE
For circos	preparation documents for Circos software		

Fig. 1 Overview of SPDE functions

Functions of nine modules in SPDE are ranging from sequence mining to organizing genome information, which mainly play roles in gene family analysis and genome analysis.

Information extraction

Three types of information can be extracted using SPDE: sequence information, genomic information, and rows/columns information. For sequence extraction, except for the required files, users need to provide only gene ids to obtain the CDS, UTR, promoter, and full-length gene sequences from CDS files (Supplementary video, Genes), genome-sequence files and gff/gff3 files (Supplementary videos, 1 and 4), respectively. Extracted genomic information includes chromosome/scaffold ids, length, GC content, starting and end positions of genes, and numbers of

introns, exons, and genes (Supplementary video4). To focus on partial chromosomes/scaffolds instead of the whole genome, we provided related functions that allow users to extract information from a gff/gff3 file by inputting the chromosome/scaffold id (Fig. 2A). The output file can be used as the input file to extract the information of a specific chromosome/scaffold (Fig. 2B).

The rows/columns extraction module allows the same document to be interpreted from different perspectives because files will often contain a lot of information that is not used in subsequent analysis. For example, the focus may be on differentially expressed genes before and after a particular experimental treatment. Furthermore, opening big data files in the existing Office software of Windows can be difficult, let alone dealing with them, and big data files are becoming more and more common in omics. Using the rows/columns extraction module, according to their own needs, users can easily design parameters to extract information from their files instead of opening them (Supplementary videos 6.1 and 6.2). Moreover, we set the function of big file preview to help users determine the required ids (Fig. S1). There are nine functions in this module. For example, keywords can be set up (Fig. 2C) to extract the required content (Fig. 2D). In the input file, the only requirement is that the columns are separated by a Tab. There are eight ways to complete the extraction process. To extract continuous columns/rows, the parameters should be separated by a space (Fig. 2E), whereas, to extract discontinuous columns/rows, the parameters should be separated by a comma (Fig. 2F). To extract rows by setting threshold values, for example, to obtain differentially expressed genes with expression levels above or below a particular threshold cutoff, threshold values can be set as shown in Fig. 2G. To sort the content of a new file after extracting the required rows, the parameters can be set as shown in Fig. 2H.

A extract information of a chr/scaffold/contig, based on gff/gff3
the chr/scaffold ID
save location and name your file
extract

B input GFF/GFF3 file for the following three functions
chr(scaffold or contig)ID/gene ID/start/end
save location and name your file
extract the start and end positions of genes

C input the target file
input the file containing keywords
save location and name your file
extract information based on keywords in other files(1,2,3)

D

Chr18	EVM	gene	1867718	1869325	.	+	.	ID=evm.TU.000008F.44
Chr18	.	gene	2872358	2879410	.	-	.	ID=evm.TU.000008F.28
Chr18	.	gene	3248146	3250643	.	+	.	ID=evm.TU.000008F.19
Chr23	transdecoder	mRNA	4305406	4308713	.	-	.	ID=Novel
Chr23	transdecoder	mRNA	4149488	4151327	.	-	.	ID=Novel
Chr23	transdecoder	mRNA	3978539	3980890	.	-	.	ID=Novel

E extract and sort values based on the threshold value
input the target file
extract rows/columns threshold value sort, based on column id
row 2,3
save location and name your file
extract contents from the second row to the fifth row

F extract and sort values based on the threshold value
input the target file
extract rows/columns threshold value sort, based on column id
row,2,5,7
save location and name your file
extract the second, fifth, seventh rows

G extract the rows with their third column values that more than 1e-20
extract rows/columns threshold value sort, based on column id
3 >= 1e-20
save location and name your file
extract

H sort rows by ascending order, according to the values of the third column
extract rows/columns threshold value sort, based on column id
3 positive
save location and name your file
extract

Fig. 2 Information mining

(A) SPDE can extract partial chromosomes/scaffolds information from gff/gff3 files by inputting the chromosome/scaffold identifier (id). (B) The output file from (A) can be used as the input file to extract information about the genes that are located in the extracted chromosome/scaffold. (C, D) A file containing keywords can be set up to extract information from user files. (E, F) Setting parameters separated by a space or a comma will extract continuous or discontinuous rows/columns, respectively. (G) Setting a threshold value will extract eligible rows. (H) Setting 'positive/negative' will sort the content of the file obtained in (G) in ascending/descending order.

Creating input files suitable for downstream software

Our files sometimes may not be suitable for downstream software analysis because of (1) subtle differences in same-type files (Fig. S2A); (2) differences in format between the original files and input file of downstream software (Fig. S2B); and (3) extra information in original the files that is not required for the follow-up analysis (Fig. S2C). In SPDE, four functions were set up to solve such problems (Fig. 3). To delete specific strings, the function can be completed as shown in Fig. 3A (Supplementary video 6.1). To remove extra information from the original file, the functions can be set up as showed in Fig. 3B and C. To reorganize the file content to meet the format requirement of downstream software, the related parameters can be set up as shown in Fig. 3D (Supplementary videos 6.2).

In particular, considering the importance of the Circos software in displaying genomic

information, we created a ‘For circos’ module in SPDE to generate the output files which meet the format required for Circos input files (Supplementary videos 5.1 and 5.2). As shown in Fig. 3E, this module can produce six types of files (karyotype, links, gene density, heatmap, text, SNP) that can be used directly as input files for Circos.

Fig. 3 Creating input files for downstream software

(A) Function to remove specific strings from original files. (B, C) Function to remove the extra information from original files. (D) Function to reorganize user files to meet the format requirement of downstream software. (E) Functions for generating input files for the Circos software.

Data format conversion

Specific types and formats of the input data should meet the requires of different data extraction and machine learning algorithms¹⁴. Thus, format conversion is an operation that we often need to perform in the process of bioinformatics analysis. In SPDE, nine functions were set up for helping users convert their files to another format (Fig. 4). For example, the function of ‘convert fastq to

fasta' (Fig. 4A) can change the input file to a fasta file (Supplementary video 2). Two kinds of files (GenBank files and gff/gff3 files) can also be converted to each other (Fig. 4B, Supplementary video 6.1). With the help of SAMtools¹⁵, Sequence Alignment/Map format files (sam files) will be changed to Binarysequence Alignment/Map format files (bam files) or Variant Call Format files (VCF files, Fig. 4C). Furthermore, the VCF files can also be transformed to another format files¹⁶ (in Windows versions, Fig. 4D, Supplementary video 7).

A Transform

input fastq file (it should be a gzip file) ...

save location and name your file ...

convert fastq to fasta

B GBK and GFF

input gff file ...

input genome file ...

save to and save as ...

transform gff to GBK (gff and genome files required)

input NCBI GBK file ...

save to and save as ...

transform GBK to gff extract protein seqs from GBK

C

convert *.tree to *.nwk

input the Phylogenetic tree file ...

save location and name your file ...

convert *.tree to *.nwk extract gene ids from the input file

convert gtf/gff to gtf/gtf

input gtf/gff file ...

save location and name your file ...

convert gtf to gff convert gff to gtf

convert *.sam, *.bam, and *.vcf

input the sam file (1)

input the bam file (2)

reference file (sequences of the target genome/chr/scaffold) (3)

save location and name your file (4)

sam to bam (1,4)

bam to vcf (2,3,4)

D

convert *.vcf to *.phylip/NEXUS/NEXUS-binary/FASTA_alignments

input vcf file ...

The format of the output file

☐ phylip ☐ phylip, fasta, nexus and binary_nexus

☐ Phylip and resolved by IUPAC ☐ phylip and fasta ☐ Only nexus

transform vcf (the results showed in the folder containing the input file)

Fig. 4 File format conversion

A, fastq to fasta; B, conversion between (GenBank files and gff/gff3 files); C, phylogenetic tree file to NEWICK file, conversion among sam files, bam files and VCF files; D, VCF files to another format file.

Using the user-friendly SPDE interface

The user-friendly SPDE interface used point-and-click operations. Although the software NCBI blast+, HMMERsearch, ClustalW, Muscle, MCScanX, and Primer3 (Fig. 5) are excellent tools for sequence alignment, homologous genes identification, collinearity analysis, and primer design, the required command-line operations can make them difficult for users to implement. For this reason, we developed interfaces for these programs in SPDE whereby users can obtain the results by inputting their files according to the prompts of the interface (Supplementary videos 3.1 and 3.2).

Considering that the design of primers for multiply genes or sequences is a common task, we created the ‘Genes’ module to finish it (Supplementary video 2). The settings for the functions in this module are shown in Fig. 5. The ‘create control file (for primers,4,5,6)’ and ‘design primers (4,5,6,7)’ buttons are used to generate primers. A total of five pairs of possible primers is recommend. Then, the ‘extract the best primers’ button can be used to extract the best primers for a particular purpose (Fig. S3).

Fig. 5 SPDE incorporates NCBI blast+, HMMERsearch, MCScanX, ClustalW, Muscle and Primer3

HMMERsearch, MCScanX, ClustalW, Muscle, NCBI blast+ and Primer3 can be used without the need for programming or command lines. All users need to do is entering the corresponding content as prompted and clicking the relative buttons.

Usability of SPDE

There is a ‘Help’ button or ‘About SPDE’ button at the SPDE interface that opens a list of module names that can be clicked to obtain the help document for the selected module, and article attachments that contain videos showing the use of SPDE also can be downloaded (Supplementary video 8). Although, the interface of different versions (32-bit, 64-bit Windows and macOS version) may have slight differences, there is enough prompt information on the software interface to help users apply SPDE. For example, the prompt ‘input formatted genome sequences’ tells users where the formatted genome sequences should be input (Fig. 6, Supplementary video 4). Users can use functions in the ‘Formatting’ module to format genome sequences that have not been formatted.

Although formatting files may take some time because the operation objects are big data files, we found that ‘formatting files’ was an effective way to improve the efficiency of SPDE processing.

Statistical analysis

N50/N90/Chrid/GC_ratio/N_ratio/Length/the number of N/A/T/C/G

input formatted genome sequences ...

save location and name your file ...

extract basic information of genome extract chrid/length for karyotype

input GFF/GFF3 file for the following three functions ...

chr(scaffold)ID/gene ID/start/end (generate gff file for MCScanX)

save location and name your file ...

extract the start and end positions of genes

extract information of a chr/scaffold/contig, based on gff/gff3

the chr/scaffold ID

save location and name your file ...

extract

numbers of genes/introns/CDSs in a chromosome/genome

save location and name your file ...

extract information based on gff/gff3 file

Fig. 6 Enough prompt information on the software interface

To remove barriers in the use of SPDE, three kind of prompt information were left in the interface. The red character is the title or the warning message; the blue character is the function introduction or the contents of results; the grey character is the content that should be input in the box.

Conclusions

SPDE has an important role in facilitating sequence processing and data extraction. As we gradually improve its functions, SPDE will become a powerful tool for biologists in processing biological big data.

Acknowledgments

This work was supported by a grant from National Transformation Science and Technology Program (2018ZX08020002-005-003), the Zhejiang Science and Technology Major Program on

Agricultural New Variety Breeding (No. 2016C02056-1), the National Natural Science Foundation of China (No.31870647 and NO.31872168).

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Reference

- 1 Wei, Q. *et al.* A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Horticulture Research* **7**, 153, doi:10.1038/s41438-020-00391-0 (2020).
- 2 Chen, S.-P. *et al.* The Phoebe genome sheds light on the evolution of magnoliids. *Horticulture Research* **7**, 146, doi:10.1038/s41438-020-00368-z (2020).
- 3 Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49, doi:10.1093/nar/gkr1293 (2012).
- 4 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652, doi:10.1038/nbt.1883 (2011).
- 5 Schilling, S., Kennedy, A., Pan, S., Jermin, L. S. & Melzer, R. Genome-wide analysis of MIKC-type MADS-box genes in wheat: pervasive duplications, functional conservation and putative neofunctionalization. *New Phytologist* **225**, 511-529, doi:10.1111/nph.16122 (2020).
- 6 Li, S. *et al.* The curvature of cucumber fruits is associated with spatial variation in auxin accumulation and expression of a YUCCA biosynthesis gene. *Horticulture Research* **7**, 135, doi:10.1038/s41438-020-00354-5 (2020).
- 7 Liu, M. *et al.* Genome-wide identification, phylogeny, evolutionary expansion and expression analyses of bZIP transcription factor family in tartary buckwheat. *BMC Genomics* **20**, 483, doi:10.1186/s12864-019-5882-z (2019).
- 8 Oh, D.-H. *et al.* Cell type-specific responses to salinity – the epidermal bladder cell transcriptome of *Mesembryanthemum crystallinum*. *New Phytologist* **207**, 627-644, doi:10.1111/nph.13414 (2015).
- 9 Krzywinski, M. I. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Research*, doi:10.1101/gr.092759.109 (2009).
- 10 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).
- 11 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797, doi.org/10.1093/nar/gkh340 (2004).
- 12 Katoh, K., Misawa, K., Kuma, K. i. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066, doi:10.1093/nar/gkf436 (2002).
- 13 Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods of Molecular Biology* **132**, 365-386, doi:10.1385/1-59259-192-2:365 (2000).
- 14 Ahmed, S. *et al.* Modern Data Formats for Big Bioinformatics Data Analytics. *International Journal of Advanced Computer Science and Applications* **8**, 366-377, doi:10.14569/IJACSA.2017.080450 (2017).
- 15 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

374 16 Ortiz, E.M. 2019. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for
375 phylogenetic analysis. doi:10.5281/zenodo.2540861(2019).