MATLAB software for extracting protein name and sequence information from FASTA formatted proteome file

Wenfa Ng

Unaffiliated researcher, Singapore, Email: ngwenfa771@hotmail.com

Abstract

FASTA file format is a common file type for distributing proteome information, especially those obtained from Uniprot. While MATLAB could automatically read fasta files using the built-in function, fastaread, important information such as protein name and organism name remain enmeshed in a character array. Hence, difficulty exists in automatic extraction of protein names from fasta proteome file to help in building a database with fields comprising protein name and its amino acid sequence. The objective of this work was in developing a MATLAB software that could automatically extract protein name and amino acid sequence information from fasta proteome file and assign them to a new database that comprises fields such as protein name, amino acid sequence, number of amino acid residues, molecular weight of protein and nucleotide sequence of protein. Information on number of amino acid residues came from the use of the length built-in function in MATLAB analyzing the length of the amino acid sequence of a protein. The final two fields were provided by MATLAB built-in functions molweight and aa2nt, respectively. Molecular weight of proteins is useful for a variety of applications while nucleotide sequence is essential for gene synthesis applications in molecular cloning. Finally, the MATLAB software is also equipped with an error check function to help detect letters in the amino acid sequence that are not part of the family of 20 natural amino acids. Sequences with such letters would constitute as error inputs to molweight and aa2nt, and would not be processed. Collectively, given that important information such as protein name is enmeshed in a character array in fasta proteome file, this work sets out to develop a MATLAB software that could automatically extract protein name and amino acid sequence information, and assigns them to a new protein database. Using built-in functions, number of amino acid residues, molecular weight and nucleotide sequence of each protein were calculated; thereby, yielding a new protein database with improved functionalities that could support a variety of biology workflows ranging from sequence alignment to molecular cloning.

Keywords: FASTA, proteome, MATLAB, molecular weight, amino acid sequence, nucleotide sequence, character array, parse information, protein database, molecular cloning,

Subject areas: biochemistry, molecular biology, genomics, bioinformatics, computational biology,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27856v1 | CC BY 4.0 Open Access | rec: 13 Jul 2019, publ: 13 Jul 2019

Highlights

1) MATLAB software capable of automatic extraction of protein name and amino acid sequence from fasta proteome file is developed.

2) The software is also able to calculate the number of amino acid residues, molecular weight, and nucleotide sequence of each protein.

3) Error check functionality is incorporated into the software that help checks for letters in amino acid sequence that are not part of the natural set of 20 amino acids.

4) All extracted and calculated information are curated into a proteome database that is output as an Excel file for easy reference by users.

Background

Protein name and amino acid sequence are the two principal types of information in a proteome data file. While amino acid sequence information exists as a defined field that could be easily extracted by bioinformatics software, protein name presents a more difficult challenge as it is enmeshed in a character array that comprise other information such as organism name and protein ID. Hence, dedicated codes must be developed for automatic extraction of protein name after understanding recurring patterns in the character array that help define protein name. Such patterns help delineate the range of characters that should be extracted as a protein name.

One common file format for storing proteome data information is FASTA. In this work, a MATLAB software was developed for automatic extraction of protein name and sequence information from a fasta file and store the extracted information in a new database. As number of residues in protein, molecular weight of protein, and nucleotide sequence of protein are important information, built-in functions in MATLAB would be employed to calculate these parameters whose values would be added to the new proteome database, which would be output as an Excel file for ease of access by the user. In particular, selection of nucleotide sequence of protein for output came about as this information is needed in the synthesis of genes in modern molecular cloning workflow. Given that errors such as wrong letters that are not part of the 20 natural amino acids might be present in the protein sequence extracted from the fasta file, the MATLAB software comes with a built-in error check function that helps filter out protein sequence with letters outside of the family of 20 natural amino acids. This helps prevent errors in executing built-in functions molweight[1] and aa2nt[2] which do not accept letters that are not part of the family of natural amino acids.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27856v1 | CC BY 4.0 Open Access | rec: 13 Jul 2019, publ: 13 Jul 2019

Implementation

In this work, built-in function, fastaread,[3] was first used to extract two principal information from fasta proteome file: Header and Sequence. Header comprises myriad information such as protein name, protein ID and organism name in a meshed character array. On the other hand, sequence information could be more easily handled as it exists as a dedicated field on its own.

Understanding the existence of recurring patterns in the character array was the first step to extracting the protein name. It was observed that the protein name existed after 'ECOLI' and before 'OS' in the character array extracted by fastaread. Using strfind[4] built-in function, the locations in which 'ECOLI' and 'OS' occurred in the character array were identified, which provided placeholders for the automatic extraction of protein name. The same methodology could be applied to extract protein name from fasta proteome file of other bacterial and fungal species.

Next, number of residues in each protein was calculated using built-in function, length,[5] in MATLAB that output the length of each amino acid sequence that corresponds to the number of amino acid residues in each protein. Similarly, built-in function molweight was used to calculate the molecular weight corresponding to each amino acid sequence. On the other hand, aa2nt built-in function helps convert amino acid to nucleotide sequence, which is important for users interested in synthesizing a gene for molecular cloning applications or for sequence similarity check.

Finally, possibility exists of letters not part of the natural set of 20 amino acids being incorporated into the amino acid sequence; thus, an error check function was added to the MATLAB software to help flag incorrect letters that are not part of the 20 amino acid family. Specifically, the error check function scans each letter of the amino acid sequence for letters that are not part of the natural set of amino acids. If a letter that is not part of the family of natural amino acids is detected, the amino acid sequence that contains the letter would not serve as input for molecular weight calculation and amino acid to nucleotide sequence conversion.

Key features of software

Besides automatic extraction of protein name from enmeshed data comprising protein ID, protein name and organism name, other important features of the MATLAB software include: (i) ability to calculate the number of amino acid residue and molecular weight of each protein, as well as (ii) perform amino acid to nucleotide sequence conversion.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27856v1 | CC BY 4.0 Open Access | rec: 13 Jul 2019, publ: 13 Jul 2019

1x4313 struct with 5 fields

Fields	Name	AA_Sequence	No_of_residues	Mol_weight	Nt_Sequence
1	' Lactose op...	'MKPVTLYDVAE...	360	3.8590e+04	'ATGAAACCCG...
2	' D-lactate ...	'MKLAVYSTKQY...	329	3.6534e+04	'ATGAAACTGG...
3	' Probable ...	'MSRSQNLRHNV...	304	3.5397e+04	'ATGTCTCGGAG...
4	' Prolipopro...	'MTSSYLHFPEFD...	291	3.3108e+04	'ATGACATCCTC...
5	' Putative fr...	'MRNLQPNMSR...	28	3.2908e+03	'ATGCGTAATTT...
6	' phe opero...	'MKHIPFFFAFFF...	15	1.9243e+03	'ATGAAGCATAT...
7	' his operon...	'MTRVQFKHHH...	16	2.0813e+03	'ATGACGCGAG...
8	' Putative rh...	'MRSEQISGSSLN...	33	3.7502e+03	'ATGCGTTCCGA...
9	' Lipopolysa...	'MKFKTNKLSLNL...	185	2.0127e+04	'ATGAAATTCAA...
10	' Lipopolysa...	'MATLTAKNLAK...	241	2.6800e+04	'ATGGCAACAC...
11	' Lipopolysa...	'MSKARRWVIIVL...	191	2.1702e+04	'ATGAGCAAGG...
12	' Leucine-re...	'MVDSKKRPGKD...	164	1.8887e+04	'ATGGTAGACTC...
13	' Met repres...	'MAEWSGEYISPY...	105	1.2141e+04	'ATGGCCGAAT...
14	' Methionin...	'MIKLSNITKVFH...	343	3.7788e+04	'ATGATCAAACT...
15	' tRNA-2-m...	'MTKKLHIKTWG...	474	5.3662e+04	'ATGACAAAAA...

Figure 1: Sample output from the MATLAB software that automatically extracts protein name and sequence information from fasta proteome data file, and which calculates number of amino acid residues, molecular weight and nucleotide sequence of each protein.

Figure 1 shows a sample output from the proteome database that is created based on information both extracted and calculated from the raw proteome data. The database serves as input to a subroutine that outputs the same proteome database into an Excel file using the xlswrite[6] function in MATLAB.

Another feature of the MATLAB software is its ability to automatically detect amino acid sequence with letters not belonging to the natural set of 20 amino acids. Such a feature allows the flagging of amino acid sequence with errors, which is important for preventing the use of the problematic amino acid sequence in downstream workflows in molecular biology or proteomics.

Conclusions

Automatic extraction of protein name from a character array read from a fasta proteome file is the main challenge in this work, for which a MATLAB software was developed that could extract protein name from a character array enmeshed with protein ID, protein name and organism name. Besides protein name extraction, other functionalities of the MATLAB software include the calculation of number of amino acid residues, molecular weight and nucleotide sequence of each protein in the proteome. Such information is of relevance to many biology workflows ranging from

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27856v1 | CC BY 4.0 Open Access | rec: 13 Jul 2019, publ: 13 Jul 2019

interpreting results from gel electrophoresis to gene synthesis in molecular cloning. Whether extracted or calculated, all information is stored in a new structured array variable that could be output as an Excel file for easy reference. Overall, the MATLAB software should find ready use in automatic extraction of protein name and amino acid sequence information from fasta proteome file and calculation of pertinent parameters of each protein such as number of amino acid residues, molecular weight and nucleotide sequence.

Source code

```
function Proteome_analysis

    proteome_Ecoli = fastaread('Escherichia coli K-12 proteome.fasta');

    k1 = length(proteome_Ecoli);
    natural_aa = 'GALMFWKQESPVICYHRNDT';

    for i = 1:k1
        header = proteome_Ecoli(i).Header;
        sequence = proteome_Ecoli(i).Sequence;
        sequence_length = length(sequence);
        flag = check_amino_acid_seq(sequence, natural_aa);
        protein_name = extract_protein_name(header);

        if flag == 1

            low_sequence = lower(sequence);
            mol_weight = molweight(low_sequence);
            nt_sequence = aa2nt(low_sequence);
        end

        proteomedb(i).Name = protein_name;
        proteomedb(i).AA_Sequence = sequence;
        proteomedb(i).No_of_residues = sequence_length;

        if flag == 1

            proteomedb(i).Mol_weight = mol_weight;
            proteomedb(i).Nt_Sequence = nt_sequence;
        end
    end

    write_data(proteomedb);

end

function protein_name = extract_protein_name(header)

    index1 = strfind(header, 'ECOLI');
    index2 = strfind(header, 'OS');
    protein_name = header(index1+5: index2-2);
```


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27856v1 | CC BY 4.0 Open Access | rec: 13 Jul 2019, publ: 13 Jul 2019

```
end

function flag = check_amino_acid_seq(sequence,natural_aa)

    k1 = length(sequence);
    flag =1;

    for i=1:k1
        letter = sequence(i);
        if ~contains(natural_aa, letter)
            flag = 0;
            break
        end
    end

end

function write_data(proteomedb)

    k1 = length(proteomedb);

    A{1,1} = 'Name';
    A{1,2} = 'AA_Sequence';
    A{1,3} = 'No_of_residues';
    A{1,4} = 'Mol_weight';
    A{1,5} = 'Nt_Sequence';

    for i = 1:k1
        A{i+1,1} = proteomedb(i).Name;
        A{i+1,2} = proteomedb(i).AA_Sequence;
        A{i+1,3} = proteomedb(i).No_of_residues;
        A{i+1,4} = proteomedb(i).Mol_weight;
        A{i+1,5} = proteomedb(i).Nt_Sequence;
    end

    xlswrite('Proteomedb.xlsx', A)

end
```

Supplementary information

The supplementary information of this manuscript contains a zip file that encapsulates the MATLAB software file and its subroutines.

References

1. Calculate molecular weight of amino acid sequence - MATLAB molweight. Available at:

https://www.mathworks.com/help/bioinfo/ref/molweight.html. (Accessed: 10th July 2019)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27856v1 | CC BY 4.0 Open Access | rec: 13 Jul 2019, publ: 13 Jul 2019

2. Convert amino acid sequence to nucleotide sequence - MATLAB aa2nt. Available at: https://www.mathworks.com/help/bioinfo/ref/aa2nt.html. (Accessed: 10th July 2019)
3. Read data from FASTA file - MATLAB fastaread. Available at: https://www.mathworks.com/help/bioinfo/ref/fastaread.html. (Accessed: 10th July 2019)
4. Find strings within other strings - MATLAB strfind. Available at: https://www.mathworks.com/help/matlab/ref/strfind.html. (Accessed: 10th July 2019)
5. Length of largest array dimension - MATLAB length. Available at: https://www.mathworks.com/help/matlab/ref/length.html. (Accessed: 10th July 2019)
6. (Not recommended) Write Microsoft Excel spreadsheet file - MATLAB xlswrite. Available at: https://www.mathworks.com/help/matlab/ref/xlswrite.html. (Accessed: 10th July 2019)

Conflicts of interest

The author declares no conflicts of interest.

Funding

No funding was used in this work.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27856v1 | CC BY 4.0 Open Access | rec: 13 Jul 2019, publ: 13 Jul 2019