# USING A JAVA-BASED PROGRAM TO PRODUCE REDUCED AMINO ACID ALPHABET VERSIONS OF QUERY SEQUENCES AND DATABASES

Breeanna Burkinshaw, Alex Hamilton, Sharmin M. Sikich

Department of Chemistry, Doane University, Crete, Nebraska 68333

***ABSTRACT***

*Reduced amino acid alphabets are used to show similarity in sequence between two proteins that may not otherwise be detected. This can occur if slight changes or mutations have occurred over time through evolution and divergence of species. In this case the structure would remain similar, but the sequence would appear to be much different. It is important to explore different methods in which to determine the reduced amino acid alphabet that produces the most accurate results. Here we describe a program we created to convert individual protein sequences and entire databases to a user-defined reduced amino acid alphabet. An optimal reduced alphabet would need to conserve the fold of the original protein while also showing similarities in alignments that were not previously shown. Through the use of this Java-based program developed and BLAST we have begun to explore and make easier the use of reduced amino acid alphabets.*

***KEYWORDS***

*Reduced amino acid alphabet,fold conservation, processing, sequence alignment*

## 1. INTRODUCTION

### 1.1 Computational Biology Tools Reduced AminoAcid Alphabets

There are 20 different naturally occurring amino acids that all have differing physicochemical properties (physical and chemical properties). This is important because each amino acid has a different shape and makes a different contribution to folding. Reduced amino acid alphabets(reduced AAA) have paired similar amino acids together with the best results being produced using 10-12 letter reduced alphabets [1]. The best results show no significant decrease in fold conservation [1]. Reduced amino acid alphabets are used to determine homology, structural similarities and fold recognition between distantly related proteins. This is important becauseprotein fold is influenced by how the side chains of the amino acids interact with each other. Non-covalent interactions between amino acids help to form secondary structure of proteins. Hydrophobic clustering causes hydrophobic amino acids to go toward the inside of a structure and hydrophilic amino acids with go on the outside. In an aqueous solution protein folding prediction has been studied using reduced alphabets and shows no signs of significant changes in the fold when the alphabet drops from 20 to 10 amino acids [1]. If the fold is not

conserved when the reduced AAA is used, then it would no longer have the same fold and function as the original protein.

Figure 1 shows an example of how amino acids could be grouped together for a reduced AAA. These reduced AAA will vary based on what groupings the user wants to explore, and will produce sequences with different outcomes.

| Representative Letter | L | F | S | P | H | Y | N | K | Q | C | G | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Groupings | L<br>V<br>I<br>M | F | S<br>T | P<br>A | H | Y | N<br>D | K | Q<br>E | C<br>R | G | W |

Figure 1: An example of how amino acids could be grouped in a reduced AAA and what the representative letter would be for the new groupings.

In recent years several tools have been developed for computational biology[2-4]. Here we describe a very simple program that could introduce students to topics of coding, amino acid properties, and evolution of amino acids. Processing, a java-based programing tool, was used to create a program that will transform a protein or protein sequence database into a reduced amino acid alphabet form that has been selected by the user.The program currently stores four different alphabet options that can be manipulated. The code can be easily modified in order to create a new alphabet. Having the option of four alphabets will allow the user to have multiple alphabets accessible at the click of a button instead of manipulating the program every time a second alphabet is necessary. Each of these four options will take in a protein sequence containing the original 20 letter amino acid sequence and convert it into a reduced AAA so only the representative letters are shown. It does this by pairing together amino acids in groups and choosing one of the letters to now represent the whole group. The program code can be written to allow for variation in alphabet length. For example, two of the alphabets could be 10 letters, one could be 12 letters, and the other could be 5 letters. This program can be used to reduced a single query sequence, or an entire database, allowing the user to create a custom database. The ability to choose a specific database will allow the study to be directed toward a particular species.

## 2. MATERIALS AND METHODS

### 2.1 Program Creation

The program was created using Processing. The program was saved into a folder with the same name as the program. Inside this new folder a folder named “Data” was created (see Figure 2). Inside of the data folder there were files with code forLucidaSans-24.vlw, and SansSerif.plain-14.vlw, the two fonts that are used in the program, as well as the Doane University logo (which can be found in the supplemental documents). It is important that they are placed into the correct folder for the program to work. Any other folders or files in the Data folder are up to the

user'sdiscretion. To recreate the program, the code can be found in the supplementary data and would need to be copied into an empty java file. The user could then begin manipulating the code for their own reduced AAA and add their own logo for the user interface. Figure 2 displays the user interface with the Doane University logo.

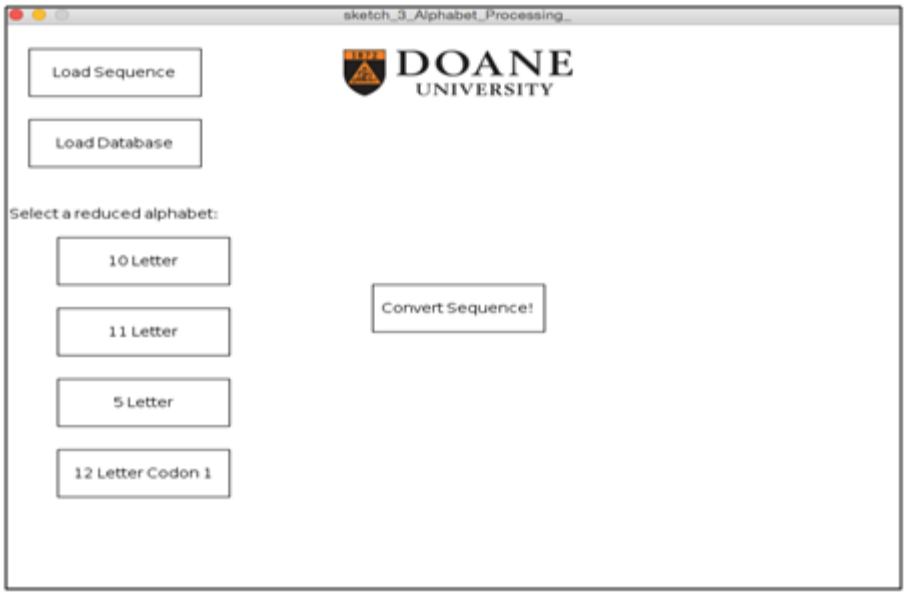


Figure 2. The user interface pop-up window that appears when the program is initiated.

## 2.2 Input and Output

An example of putting methionine and cytosine together in a new group now represented by M, is accomplished by the following in the code within the program:

```
        // Converts M and C to M
        if (toBeConverted.substring(i, i+1).equals("M") ||
        toBeConverted.substring(i,i+1).equals("C")) {
outputString += "M";
```

This allows the sequence to be converted to the reduced amino acid alphabet form as this is done for each of the groupings. An example of a query sequence that was copied and pasted into the program and output as a reduced amino acid alphabet is seen below (Figure 3).The Sequences shown below illustrate the conversion of an original sequence to the new user defined reduced alphabet.

P C R I I L T R L K A G E V D M L E E Q L G H L T S L
P C C L L L S C L K P G Q L N L L Q Q Q L G H L S S L

Figure 3. Conversion of a query sequence. The original amino acid sequence is shown on top and the sequence after it has been reduced is shown on the bottom.Changes in the sequence are highlighted. The reduced alphabet used is pictured in Figure 1. For example, Isoleucine, valine, methinione, and leucine are now represented by "L".

## 2.3 Running the Program

The Java-based program was downloaded and run on a MacBook Air 13-inch with 1.6 GHz intel core i5, 4 GB, and 1600MHz DDR3 on Macintosh HD. The program itself takes up 14KB of space with another 105 KB taken in the data folder with necessary files for the program to run. Once the program has converted the sequences, the output filesare saved in the outermost folder as FASTA files, see the example in Figure 4 "1fwp 5 letter reduced.fasta". For converting single polypeptide sequences, the button titled "Load Sequence" is clicked. A pop-up window allowing the user to copy and paste in the sequence appears (Figure 5, left).

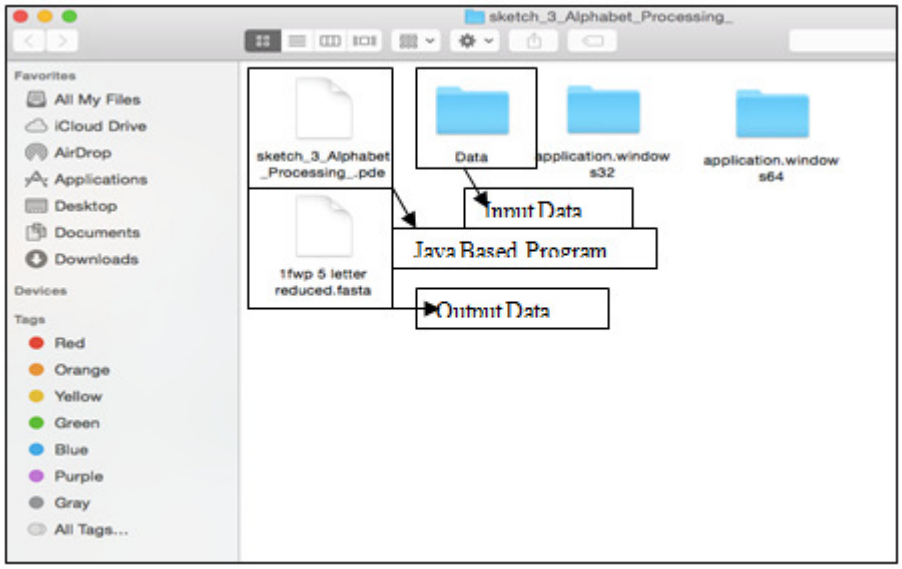


Figure 4.The folder organization so that the program will run correctly as well as a representative output file after conversion by the program.

Before a database can be converted by the program, it must first be saved into the "Data" folder. Databases can be obtained from multiple sources; however, all databases used thus far were saved from UniProt (http://www.uniprot.org/) in FASTA form into the "Data" folder [2]. When finding a database from UniProt the user can search by taxonomy and slowly narrow down to the specific database desired, or simply use the broad options supplied by the website. These options include the reviewed and unreviewed databases.

When "Load Database" is selected, then a different pop-up window appears and asks for the FASTA filename of the database file the user would like to use(Figure 5, right). The entire file name of the database must be entered including the ".fasta" end of the filename. When reducing the databases, it is important to note the file size before and after the sequence is converted by the program to ensure that the entire database was reduced and the program was allowed to run to completion. Conversion may take several seconds to several minutes depending on the size of the database and the computer specifications on which the program runs. The following video shows how to get a database from UniProt, reducing the database using the program, and making sure that the entire database has been convertedhttps://www.youtube.com/watch?v=uZUJjdEqVl4.

## 2.4 Using Stand Alone Blast

Once all sequences and databases were reduced, stand-alone BLAST was downloaded and installed. BLAST will then search the converted database for proteins with a similar sequence as the converted query sequence of interest[3]. Using stand-alone BLAST allows for the use of custom databases. The instructions for how to use stand alone BLAST and commands to use once it is installed are on the following link: https://www.ncbi.nlm.nih.gov/books/NBK52637/[4]. To use stand-alone BLAST, all files including the database and query sequences will need to be put into the correct bin, which can be seen in the following video: https://www.youtube.com/watch?v=FYlIcYQEbIY&t=4s Next, the database was formatted. The following video outlines all of the necessary steps to format databases for use in the stand-alone BLAST: https://www.youtube.com/watch?v=yM5wRdUeSxg&t=11s. Running BLAST is seen in the following video: https://www.youtube.com/watch?v=FWDwZ7KLfwo. BLAST results from both the reduced and non-reduced query sequences and databases were compared to see if the same results were obtained after using the reduced AAA.

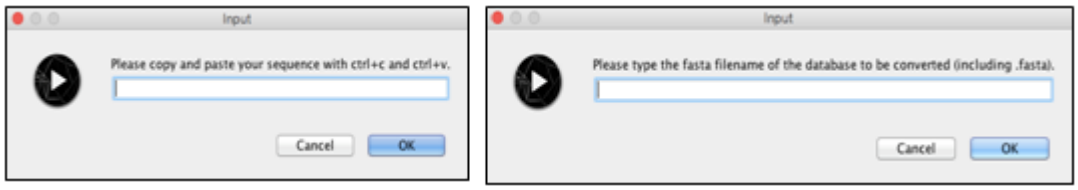


Figure 5. Left: The pop-up window that appears when the "Load Sequence" button is selected. Right: The window that appears when the user has selected to load a database.

## 2.5 Future Direction

Users can reduce sequences using custom alphabets to determine which reduced AAA will produce sequences that fold the same as the original sequence. Reduced AAA sequences can be entered into structure prediction programs to determine structure and compare the prediction of the reduced alphabet structures to the original structures [5]. This is important to be able to determine if the structure has remained the same after the sequence was reduced. Once a reduced AAA has been found to produce consistent results then it could be used to help predict unknown structures and can be used to search for distantly related proteins using custom reduced databases with BLAST. This program could also be used for bioinformatics-based projects in undergraduate research or undergraduate biochemistry courses to introduce coding and concepts of protein sequence and structure similarity, conservation, and amino acid properties.

## REFERENCES

[1] Murphy, L., Wallqvist, A. and Levy, R. (2017). Simplified amino acid alphabets for protein fold recognition and implications for folding. Oxford Academic.

[2] M. Imani, and U.M. Braga-Neto, "Optimal finite-horizon sensor selection for Boolean Kalman filter." 2017 51th Asilomar Conference on Signals, Systems and Computers. IEEE. 2017

[3] S.F.Ghoreishi, D. Allaire, "Adaptive Uncertainty Propagation for Coupled Multidisciplinary Systems, AIAA Journal (2017). (URL: https://arc.aiaa.org/doi/abs/10.2514/1.J055893)

[4] M. Imani, and U.M. Braga-Neto, " Control of Gene Regulatory Networks with Noisy Measurements and Uncertain Inputs," IEEE Transactions on Control of Network Systems (TCNS), 2018.

[5] UniProt: the universal protein knowledgebase. (2016). Nucleic Acids Research, 45(D1), pp.D158-D169.

[6] Altschul, S. (1990). Basic Local Alignment Search Tool. Journal of Molecular Biology, 215(3), pp.403-410.

[7] BLAST Manual. (2017). BLAST® Command Line Applications User Manual. NCBI.

[8] Roy, A., Kucukural, A. and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols, 5(4), pp.725-738.