ScaffoldSeq: Software for characterization of directed evolution populations

Daniel R. Woldring, Patrick V. Holec, and Benjamin J. Hackel

University of Minnesota – Twin Cities

Software Walkthrough

Overview

ScaffoldSeq is software designed for the numerous applications – including directed evolution analysis – in which a user generates a population of DNA sequences encoding for partially diverse proteins with related functions and would like to characterize the single site and pairwise amino acid frequencies across the population. Importantly, the software provides tools to cluster similar protein families, dampen the impact of dominant clones, remove background, and evaluate diversity.

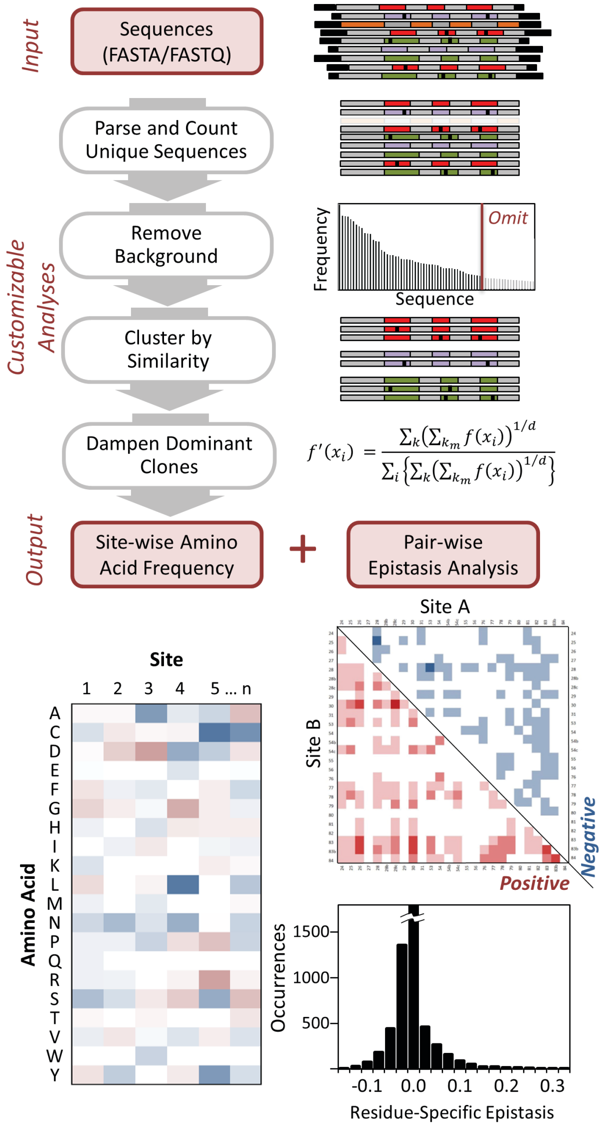

Input
Sequences (FASTA/FASTQ)
Parse and Count Unique Sequences
Customizable Analyses
Remove Background
Frequency
Omit
Sequence
Cluster by Similarity
Dampen Dominant Clones
$f'(x_i) = \frac{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}}{\sum_i \left\{\sum_k \left(\sum_{k_m} f(x_i)\right)^{1/d}\right\}}$
Output
Site-wise Amino Acid Frequency
+
Pair-wise Epistasis Analysis
Site
1 2 3 4 5 … n
Amino Acid
A C D E F G H I K L M N P Q R S T V W Y
Site A
Site B
Positive
Negative
Occurrences
1500
1000
500
-0.1 0.0 0.1 0.2 0.3
Residue-Specific Epistasis

Figure 1: ScaffoldSeq workflow.

Workflow

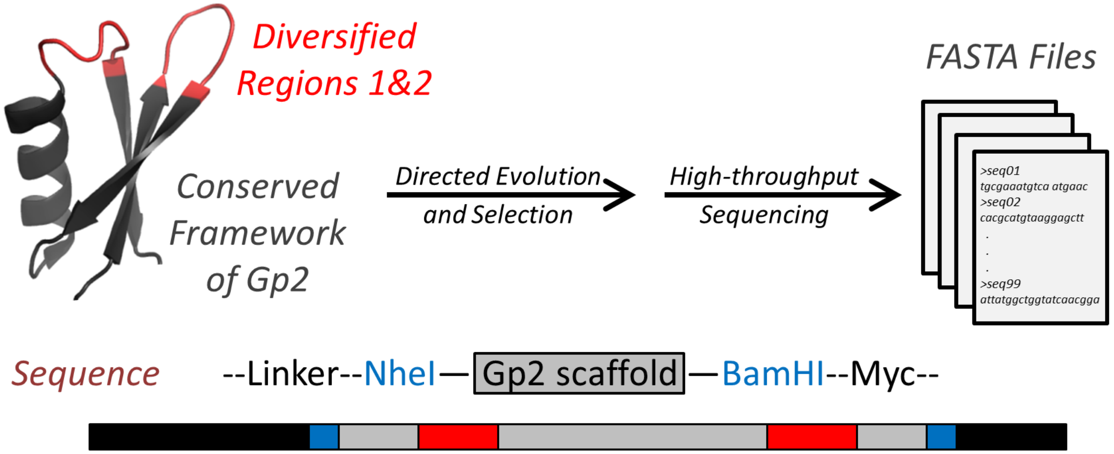
1. ScaffoldSeq reads high-throughput DNA sequences from FASTA/FASTQ files.
2. Regions of interest are parsed; unique sequences are enumerated.
3. Background sequences (i.e. the rarest clones) are quantified and omitted from analysis, if desired.
4. Highly similar clones are clustered.
5. Dampen dominant clones.
6. Output graphical and tabular results for (a) site-wise amino acid frequency and (b) pair-wise epistasis analysis.

Downloads *(Two Options)*

http://research.cems.umn.edu/hackel

https://github.com/HackelLab-UMN

Example Aim: Evaluate sitewise and pairwise amino acid frequencies within the evolved regions in a population of synthetic ligands



Amino Acid

GGGGSGGGGSGGGGSASKFWATVESSEHSFEVPVYAETLDEALELAEWQYVPAGFEVTRVRPGSEQKLISEEDL

Nucleotide

...GGTTCTGCTAGCAAATTTTGGGCGACTGTAGAATCCTCTGAACACAGCTTCGAGGTTCCGGTTTATGCTGAAACCCTGGACGAAGCACTGG
AACTGGCCGAATGGCAGTACGTTCCGGCTGGTTTCGAAGTGACCCGCGTGCGTCCGGGATCCGAACAA...

ScaffoldSeq Input

5' Anchor: GCTAGC
3' Anchor: GGGATCC
Gene Start: AAATTTTGGGCGACTGTA
DNA After Region 1: TTCGAGGTTCCGGTTTATGCTGAAACCCTGGACGAAGCACTGGAACTGGCCGAATGGCAGTAC
DNA After Region 2: GTGACCCGCGTGCGTCC

Figure 2: Representative sequence analysis scenario. The Gp2 scaffold[1] was analyzed using an in-development version of ScaffoldSeq, similar to a previous study[2]. From the 45-amino acid parental domain (PDB: 2WNM), a combinatorial library was employed whereby the two solvent exposed loops (red) were diversified in genetic sequence as well as length, with the inclusion of 6, 7, or 8 residues within each of the two regions. Populations of high-affinity binding clones evolved from this library were sequenced across the entire indicated gene (Illumina MiSeq, paired-end). Raw sequences were groomed using PANDAseq[3], producing FASTA files of full-length reads (see Paired-end Assembly section). Using the FASTA files, ScaffoldSeq evaluated the sitewise and pairwise diversity throughout the two regions of interest (red). To be included in the analysis, an entry within the FASTA file must contain matching segments for both the 5' / 3' anchors (blue) as well as the framework regions (gray) adjacent to the diversified regions (red). Default anchor and framework matching thresholds are 100% and 80%, respectively. This identifies the appropriate genes and localizes the analysis to the intended regions even within a diverse population. Note that the conserved framework positions (gray) are excluded from all future analysis. To analyze the full gene sequences, simply specify the anchor and framework sequences to be directly adjacent to, but not overlapping with the gene region. In the following walkthrough, analysis parameters were set at 0.25 for dampening (1/*d*) with a similarity clustering threshold of 0.8.

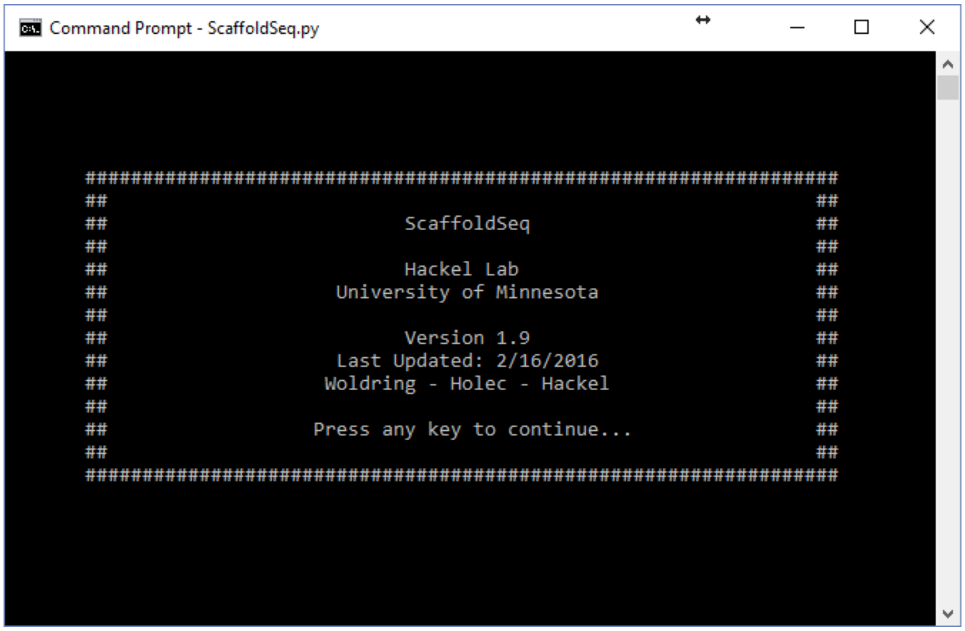
ScaffoldSeq.py is compatible with any common operating system (Windows 7/8/10, Mac OS X or Linux OS) that has Python 2.7 installed.

The software package can be downloaded from either the Hackel Lab research page (http://research.cems.umn.edu/hackel/Hackel/Publications.html) or the GitHub repository (https://github.com/HackelLab-UMN).

The script is intended to run via the operating system command line or Python terminal, rather than IDLE. Start by ensuring that both the sequence data file and ScaffoldSeq.py are located in the same directory. Then navigate to this directory using the command prompt and load the program, for example:

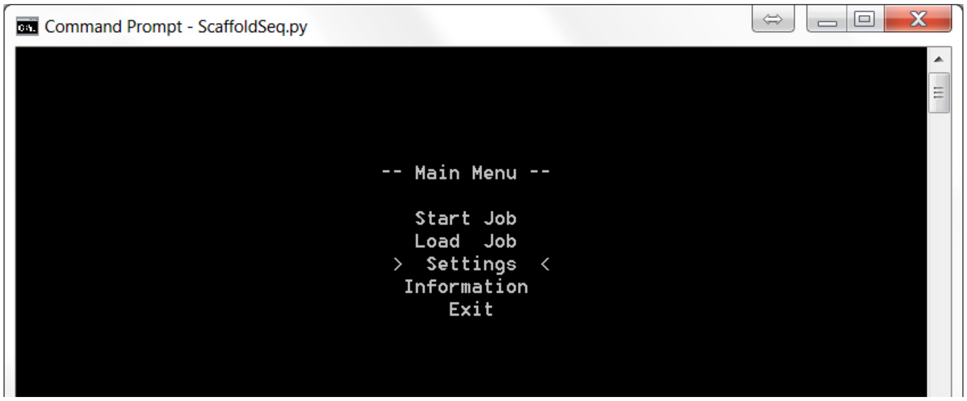
```
C:\User\Profile\GitHub\ScaffoldSeq>ScaffoldSeq.py
```

An introduction screen will then be shown.



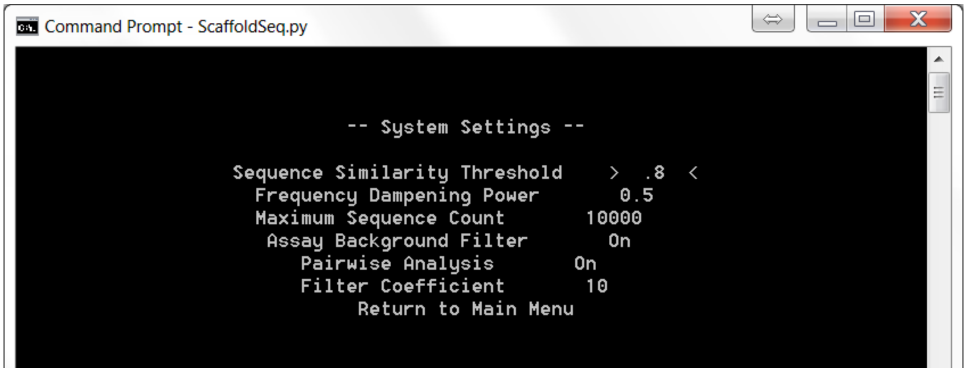
```
##################################################################
##                                                              ##
##                         ScaffoldSeq                          ##
##                                                              ##
##                          Hackel Lab                          ##
##                   University of Minnesota                    ##
##                                                              ##
##                         Version 1.9                          ##
##                   Last Updated: 2/16/2016                    ##
##                  Woldring - Holec - Hackel                   ##
##                                                              ##
##                   Press any key to continue...               ##
##                                                              ##
##################################################################
```

Press any key to continue. The Main Menu is navigated using keyboard arrows (←→↑↓), then pressing Enter. You can exit at any time by pressing Esc.



```
              -- Main Menu --

                 Start  Job
                 Load   Job
               >  Settings  <
                Information
                    Exit
```

Sequence analysis parameters can be specified within *Settings*.



```
                  -- System Settings --

        Sequence Similarity Threshold    >  .8  <
          Frequency Dampening Power        0.5
          Maximum Sequence Count          10000
           Assay Background Filter          On
              Pairwise Analysis         On
              Filter Coefficient          10
                   Return to Main Menu
```

Sequence Similarity Threshold specifies the minimum fraction of site-wise amino acid matches required to place two sequences of the same region into a common cluster. *Frequency Dampening Power* (1/*d*) operates on the individual family clusters by applying a weight to the total count of each residue-position pair, as shown in Equation 1:

$$f'(x_i) = \frac{\Sigma_k\left(\Sigma_{k_m} f(x_i)\right)^{1/d}}{\Sigma_i\left\{\Sigma_k\left(\Sigma_{k_m} f(x_i)\right)^{1/d}\right\}} \tag{1}$$

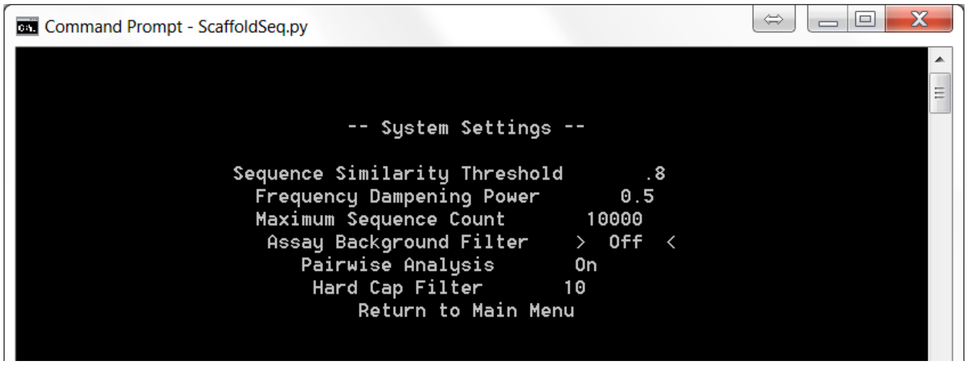
where $f_{i,j}$ is the observed occurrence of amino acid *i* at site *j within* the m^{th} sequence of family *k*; and *f′* is the dampened frequency with d^{th} root dampening. Traditional sequence analysis often treats each sequence as a distinct solution to a problem. However, within a population, two non-identical, but highly similar sequences may share a common structural or functional motif, akin to providing comparable solutions to the same problem. By lowering the *Sequence Similarity Threshold*, the ScaffoldSeq algorithm defines a broader range of related sequences to be a common solution. The contribution of each common solution (i.e. dominant clones and their common-motif variants) can be tuned to suit the needs of the analysis by using family clustering in combination with dampening.

The *Frequency Dampening Power* (1/*d*) will typically be within the range of 0.25 – 1. As this value approaches zero, the data set will be treated as though all duplicate sequences were removed. A value of 1 has the effect of weighting all sequences equally and, consequently, negates all impact of clustering, irrespective of the *Sequence Similarity Threshold*. *Frequency Dampening Power* of 0.5 is suggested for sequence data sets that contain a relatively high number of occurrences for a few dominant clones.

Maximum Sequence Count sets an upper limit for the number of sequences included in the analysis. This can be set to limiting values to speed analysis time for preliminary explorations.

Background sequences or noise should be accounted for based on the specific experiments that yielded the sequence set. *Assay Background Filter* refers to a quantifiable, assay-specific level of false positives or background. The *Filter Coefficient* is the ratio of total events to false positives. If this ratio is unknown, the *Assay Background Filter* can be turned off using the left and right arrow keys. When turned off (see below), the minimum number of occurrences (*Hard Cap Filter*) can be specified. All unique sequences that are observed less often than this value will be neglected in the analysis. Toggling *Pairwise Analysis* On/Off gives the option of performing this computationally expensive feature.

Selecting *Return to Main Menu* will save these settings.



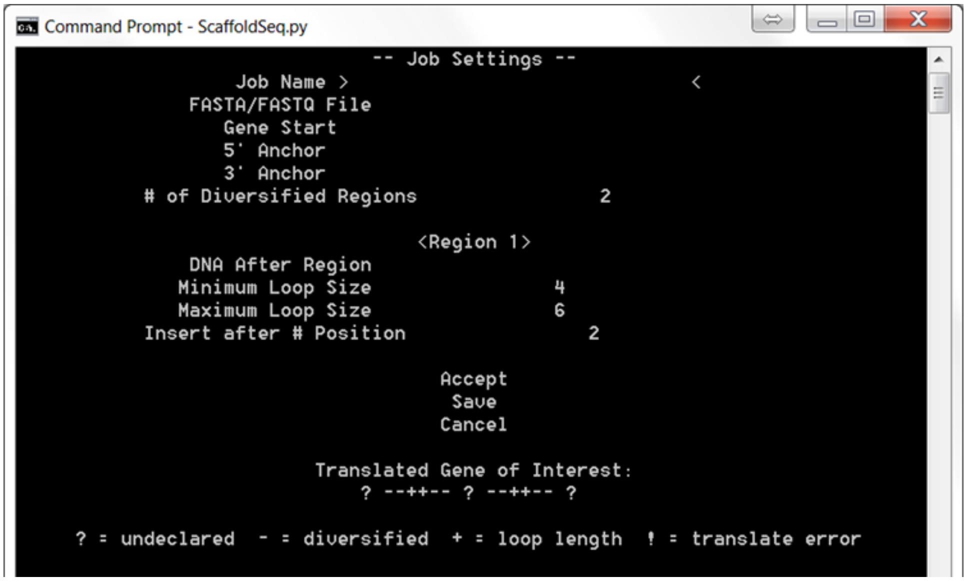
```
                    -- System Settings --

        Sequence Similarity Threshold        .8
          Frequency Dampening Power        0.5
          Maximum Sequence Count        10000
           Assay Background Filter    >  Off  <
              Pairwise Analysis       On
               Hard Cap Filter       10
                   Return to Main Menu
```

From the Main Menu, the *Start Job* option leads to a screen where the FASTA/FASTQ input file and scaffold specific DNA information are entered. *Job Name* will become the leading name of the output files. The input *FASTA/FASTQ File* is specified on the second row.

DNA sequences should be entered using all caps. Sequences can be typed in manually or pasted into the appropriate field by first selecting the field using the arrow keys, then right clicking and selecting *paste.* The Windows keyboard shortcut *ctrl-c* will likely not be accessible within the Python terminal.

Gene Start refers to all nucleotides within the gene of interest that lead up to the first diversified region. The *5' Anchor* and *3' Anchor* fields should be conserved nucleotide sequences that directly precede and follow the gene of interest, respectively. These are commonly in the form of restriction enzyme cut sites or adapter sequences that were part of the amplicon sample preparation prior to high-throughput sequencing run. The *Gene Start* nucleotides should be a conserved region that directly follows the *5' Anchor* and directly precedes the first diversified region. If multiple Diversified Regions are being investigated, the *DNA After Region* field should include all nucleotides directly following the selected Region and lead up to the next Diversified Region. The nucleotides that follow the final Diversified Region must begin directly following the diversified region and end at the nucleotide preceding the *3' Anchor* sequence. Note that both the *Anchor* matching threshold and the framework threshold are adjustable global variables within the script (ScaffoldSeq.py): *adaptor_tolerance* (default: 100%) and *framework_match_threshold* (default: 80%), respectively.

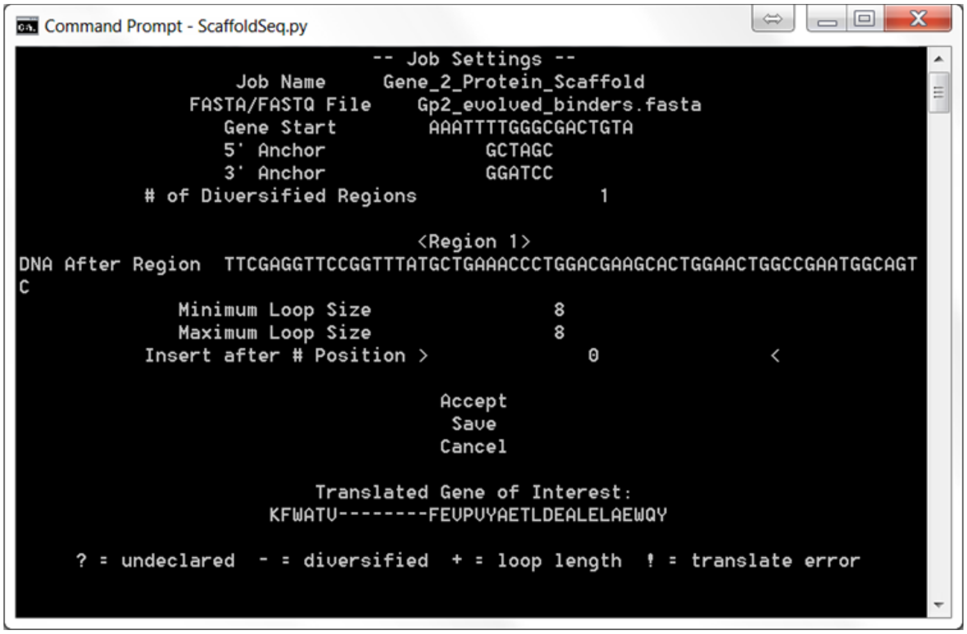
The following example coincides with the scenario shown in *Figure 2*.



```
                               -- Job Settings --
                   Job Name >                              <
               FASTA/FASTQ File
                  Gene Start
                  5' Anchor
                  3' Anchor
           # of Diversified Regions                2

                                   <Region 1>
               DNA After Region
              Minimum Loop Size                4
              Maximum Loop Size                6
           Insert after # Position                2

                                     Accept
                                      Save
                                     Cancel

                          Translated Gene of Interest:
                              ? --++-- ? --++-- ?

     ? = undeclared  - = diversified  + = loop length  ! = translate error
```

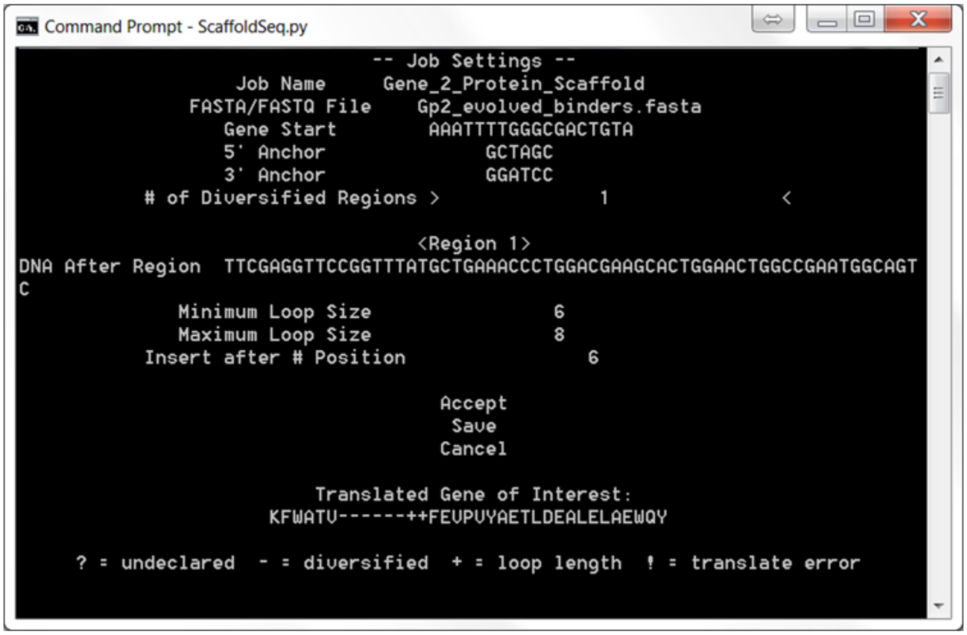
As nucleotides are entered into the *Gene Start* and *DNA After Region* fields, the *Translated Gene of Interest* section will populate.

When analyzing Diversified Regions, the nucleotides within that region should not be keyed in; however, the total number of amino acids within the diversified region must be specified with the *Minimum* and *Maximum Loop Size* fields. The Diversified Regions will be displayed as dashes, “-“, at the bottom of the screen. The amino acids that are displayed as letters at the bottom will not be included in the site-wise or pair-wise analysis. They are displayed to assist the user in ensuring the diversified regions are accurately positioned for the analysis.



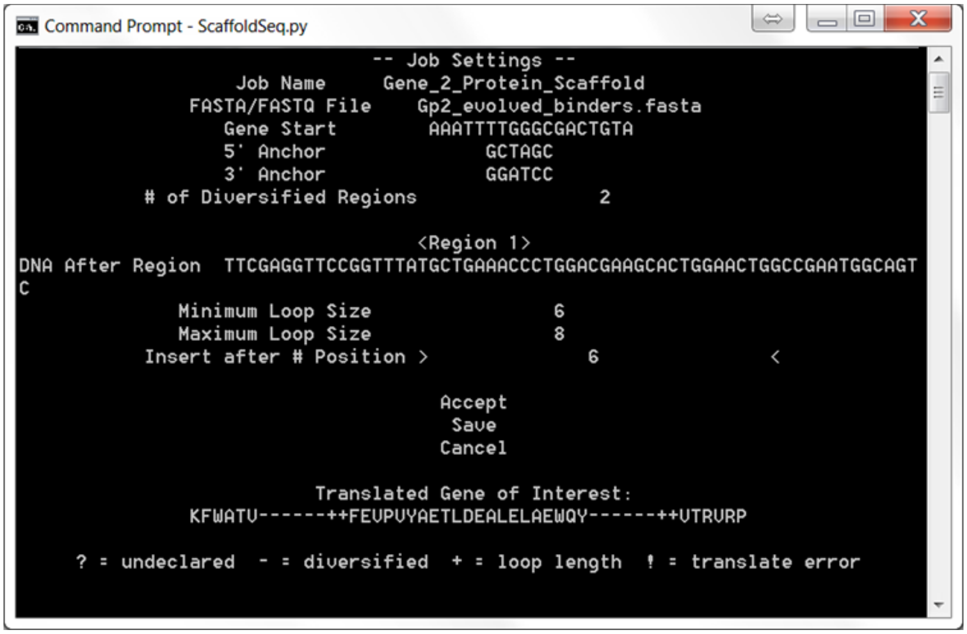
```
                         -- Job Settings --
             Job Name     Gene_2_Protein_Scaffold
         FASTA/FASTQ File    Gp2_evolved_binders.fasta
            Gene Start        AAATTTTGGGCGACTGTA
            5' Anchor              GCTAGC
            3' Anchor              GGATCC
      # of Diversified Regions               1

                             <Region 1>
DNA After Region  TTCGAGGTTCCGGTTTATGCTGAAACCCTGGACGAAGCACTGGAACTGGCCGAATGGCAGT
C
          Minimum Loop Size               8
          Maximum Loop Size               8
      Insert after # Position >              0               <

                              Accept
                               Save
                              Cancel

                    Translated Gene of Interest:
                KFWATV--------FEVPVYAETLDEALELAEWQY

     ? = undeclared  - = diversified  + = loop length  ! = translate error
```

For analyses that harbor loop length diversity within the diversified region, the position of insertion can be selected following the *Minimum* and *Maximum Loop Size* fields. The loop length diversity sites are indicated as '+' within the translated sequence.



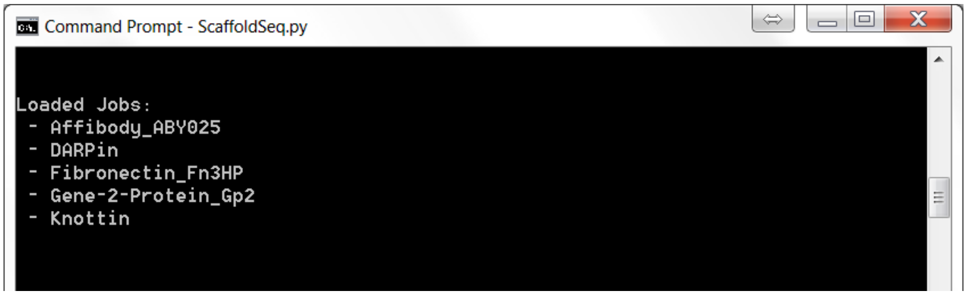
```
                         -- Job Settings --
             Job Name     Gene_2_Protein_Scaffold
         FASTA/FASTQ File    Gp2_evolved_binders.fasta
            Gene Start        AAATTTTGGGCGACTGTA
            5' Anchor              GCTAGC
            3' Anchor              GGATCC
      # of Diversified Regions >             1               <

                             <Region 1>
DNA After Region  TTCGAGGTTCCGGTTTATGCTGAAACCCTGGACGAAGCACTGGAACTGGCCGAATGGCAGT
C
          Minimum Loop Size               6
          Maximum Loop Size               8
      Insert after # Position                6

                              Accept
                               Save
                              Cancel

                    Translated Gene of Interest:
                KFWATV------++FEVPVYAETLDEALELAEWQY

     ? = undeclared  - = diversified  + = loop length  ! = translate error
```

Additional sequence regions can be included in the analysis by using the keyboard arrows to specify the # of diversified regions. Selecting *Save* will store the settings to a file in the working directory.



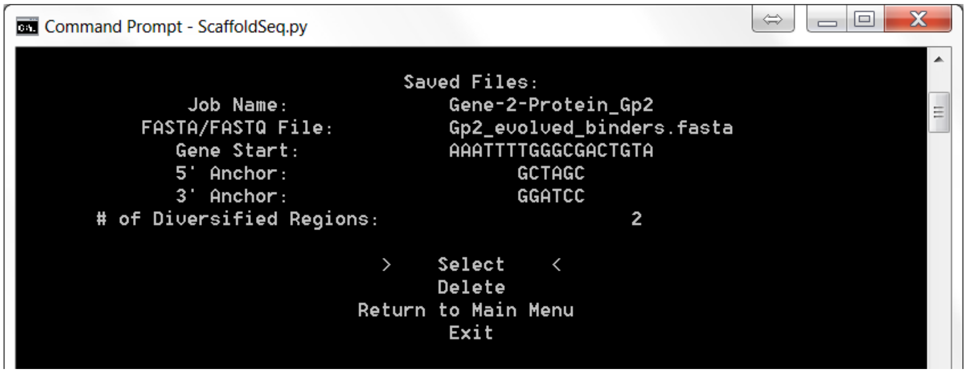
```
                        -- Job Settings --
             Job Name     Gene_2_Protein_Scaffold
         FASTA/FASTQ File    Gp2_evolved_binders.fasta
            Gene Start        AAATTTTGGGCGACTGTA
            5' Anchor              GCTAGC
            3' Anchor              GGATCC
     # of Diversified Regions                 2

                            <Region 1>
DNA After Region  TTCGAGGTTCCGGTTTATGCTGAAACCCTGGACGAAGCACTGGAACTGGCCGAATGGCAGT
C
           Minimum Loop Size                 6
           Maximum Loop Size                 8
     Insert after # Position >                  6               <

                              Accept
                               Save
                              Cancel

                       Translated Gene of Interest:
              KFWATU------++FEVPVYAETLDEALELAEWQY------++VTRVRP

     ? = undeclared  - = diversified  + = loop length  ! = translate error
```

All custom settings saved by the user will be located within the *Load Job* menu for future use.



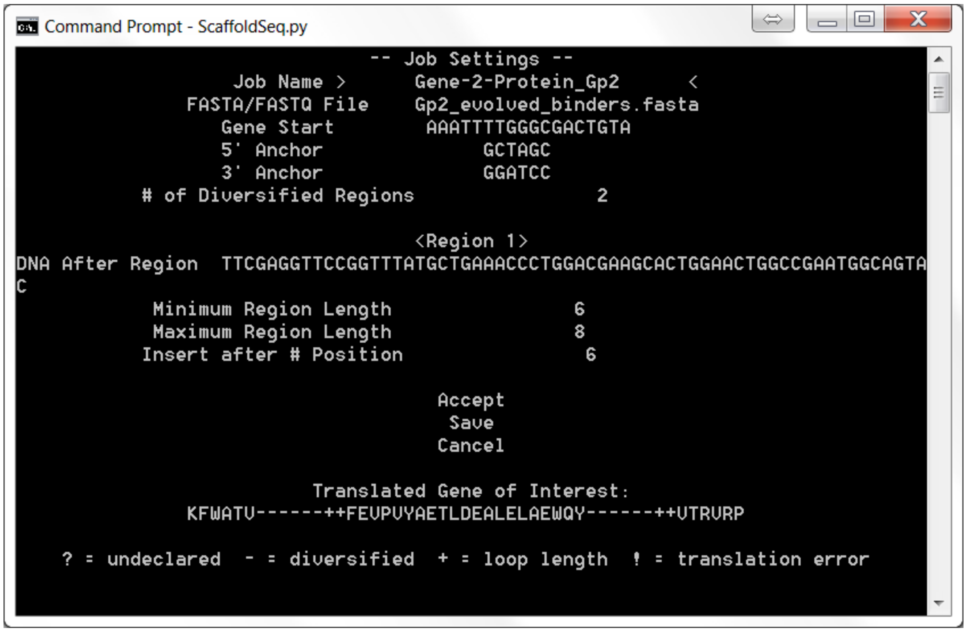
```
Loaded Jobs:
 - Affibody_ABY025
 - DARPin
 - Fibronectin_Fn3HP
 - Gene-2-Protein_Gp2
 - Knottin
```

Use the arrow keys to browse summaries for each of the saved jobs.



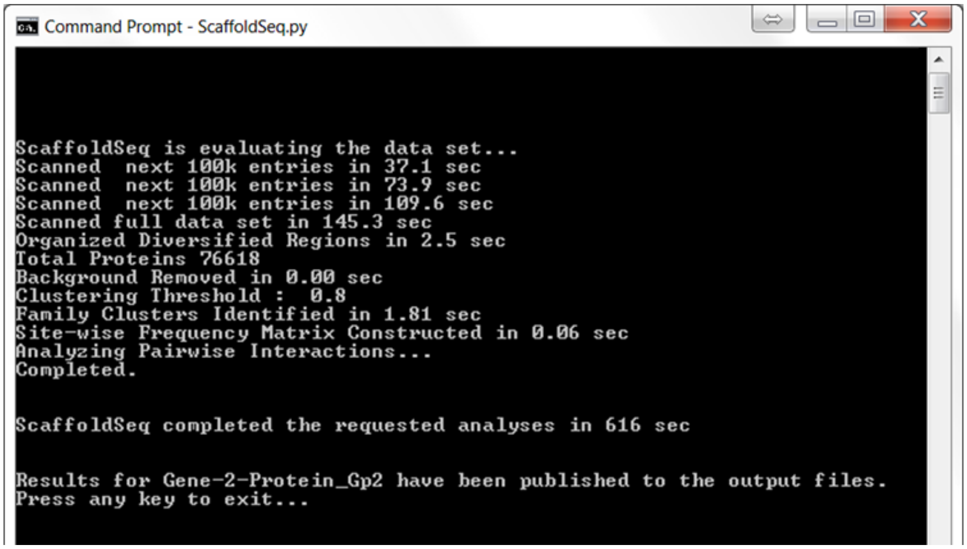
```
                                 Saved Files:
             Job Name:               Gene-2-Protein_Gp2
         FASTA/FASTQ File:           Gp2_evolved_binders.fasta
            Gene Start:              AAATTTTGGGCGACTGTA
            5' Anchor:                     GCTAGC
            3' Anchor:                     GGATCC
     # of Diversified Regions:                      2

                              >    Select    <
                                   Delete
                            Return to Main Menu
                                    Exit
```

Upon selecting a job from with the *Saved Files*, you enter the *Job Settings* environment. At this point, press *Accept* to start the job. Confirm by pressing any key or Esc to abort.



```
                               -- Job Settings --
                   Job Name >      Gene-2-Protein_Gp2      <
               FASTA/FASTQ File    Gp2_evolved_binders.fasta
                  Gene Start        AAATTTTGGGCGACTGTA
                  5' Anchor              GCTAGC
                  3' Anchor              GGATCC
           # of Diversified Regions                2

                                   <Region 1>
DNA After Region  TTCGAGGTTCCGGTTTATGCTGAAACCCTGGACGAAGCACTGGAACTGGCCGAATGGCAGTA
C
            Minimum Region Length               6
            Maximum Region Length               8
           Insert after # Position               6

                                     Accept
                                      Save
                                     Cancel

                          Translated Gene of Interest:
               KFWATV------++FEVPVYAETLDEALELAEWQY------++VTRVRP

    ? = undeclared  - = diversified  + = loop length  ! = translation error
```

As the script is performing each task, a brief status is delivered to the user as shown in the next image. Note that both the sequence file and ScaffoldSeq.py should be contained within the current working directory. Failure to do so will result in an error at which point the user must verify that (a) the sequence data file is located within the current directory and (b) the file name was entered properly. Proper implementation will produce incremental status updates resembling the following:



```
ScaffoldSeq is evaluating the data set...
Scanned  next 100k entries in 37.1 sec
Scanned  next 100k entries in 73.9 sec
Scanned  next 100k entries in 109.6 sec
Scanned full data set in 145.3 sec
Organized Diversified Regions in 2.5 sec
Total Proteins 76618
Background Removed in 0.00 sec
Clustering Threshold :  0.8
Family Clusters Identified in 1.81 sec
Site-wise Frequency Matrix Constructed in 0.06 sec
Analyzing Pairwise Interactions...
Completed.


ScaffoldSeq completed the requested analyses in 616 sec


Results for Gene-2-Protein_Gp2 have been published to the output files.
Press any key to exit...
```

Output files consists of sitewise amino acid frequency heatmaps (shown below) and tabular summaries (*.csv) of family clusters for each region of interest. If *Pairwise Analysis* was selected, an additional tabular summary (*.csv) is output, which includes mutual information (Equation 2) – with and without average product correction[4] (Equation 3) – and epistasis (amino acid-specific components of mutual information). Supporting equations for epistasis are shown in equations 5-7.

$$MI(x,y) = \sum_i \sum_j f(x_i, y_j)\ \log_2 \frac{f(x_i,y_j)}{f(x_i)f(y_j)} \tag{2}$$

$$MI_p(x,y) = MI(x,y) - \frac{MI(x,*)*MI(*,y)}{MI(*,*)} \tag{3}$$

where MI(x,*) and MI(*,y) are the mean mutual information values of site-pairs involving site x and y, respectively. MI(*,*) is the mean mutual information values across all site-pairs.

$$\textit{Residue–Specific Epistasis: } RSE(x_i y_j) = f(x_i y_j) log \left\{\frac{f(x_i y_j)}{f(x_i)f(y_j)}\right\} \tag{4}$$

$$\text{RSE, corrected: } RSE_C(x_i y_j) = RSE(x_i y_j) - \frac{RSE(x_i,*)*RSE(*,y_j)}{RSE(*,*)} \tag{5}$$

$$RSE(x_i,*) = \frac{1}{\rho(\sigma-1)} \sum_{y,y\neq x} \sum_j f(x_i y_j) log \left\{\frac{f(x_i y_j)}{f(x_i)f(y_j)}\right\} \tag{6}$$

$$RSE(*,*) = \frac{2}{\rho^2(\sigma^2-\sigma)} \sum_{x,y,y\neq x} \sum_{i,j} f(x_i y_j) log \left\{\frac{f(x_i y_j)}{f(x_i)f(y_j)}\right\} \tag{7}$$

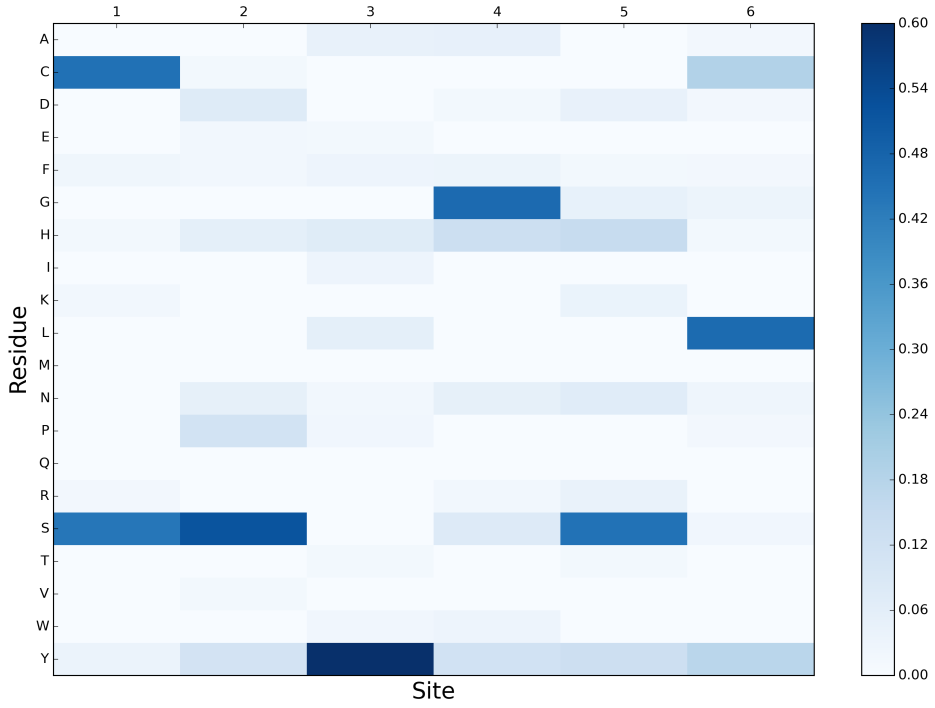
where x and y are individual positions, i and j are amino acids, ρ is the number of residue options at each position and σ is the total number of sites. Additional algorithms that improve upon mutual information have been described by others[5–9].

Representative Output Figures

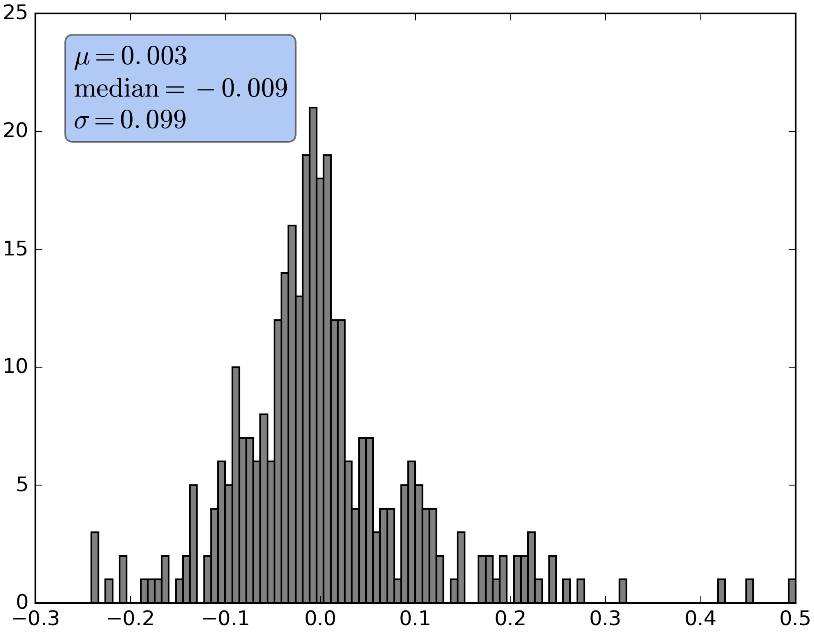
The figures below demonstrate the graphical output provided by ScaffoldSeq. The visualization modules (matplotlib, pandas, numpy) required for figure output are not included within default Python install. These are easily installed using pip via command line (pip.pypa.io/en/stable/reference/pip_install/).

Sitewise amino acid frequency heatmaps are shown for each region of interest with a color scale bar having default range 0 – 60%. Pairwise analysis is summarized by a histogram which includes mutual information using the average product correction method.

Sitewise Frequency Analysis - Region 1



Mutual Information (MI_P) 325 Site-pairs



Silent Mode

To run the software in the absence of the dynamic interface, ScaffoldSeq can be executed from the command prompt in silent mode. This is done by including a text file (e.g. jobname.txt) as the single argument for ScaffoldSeq:

```
C:\User\Profile\GitHub\ScaffoldSeq>ScaffoldSeq.py jobname.txt
```

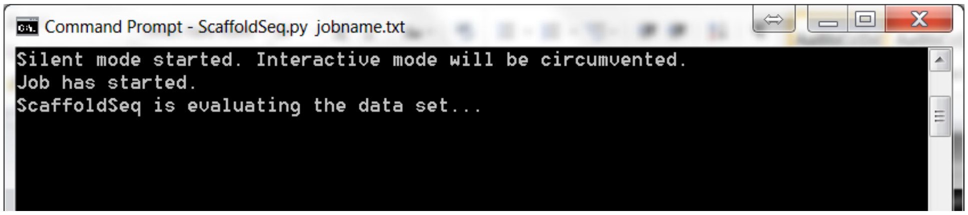
The job text file must include a complete list of predetermined settings and parameters, shown below:

```
Job Name:
FASTA/FASTQ File:
Gene Start:
5' Anchor:
3' Anchor:
# of Diversified Regions:
DNA After Region:
Minimum Region Length:
Maximum Region Length:
Insert after # Position:
Sequence Similarity Threshold:
Frequency Dampening Power:
Maximum Sequence Count:
Assay Background Filter:
Pairwise Analysis:
Filter Coefficient:
```

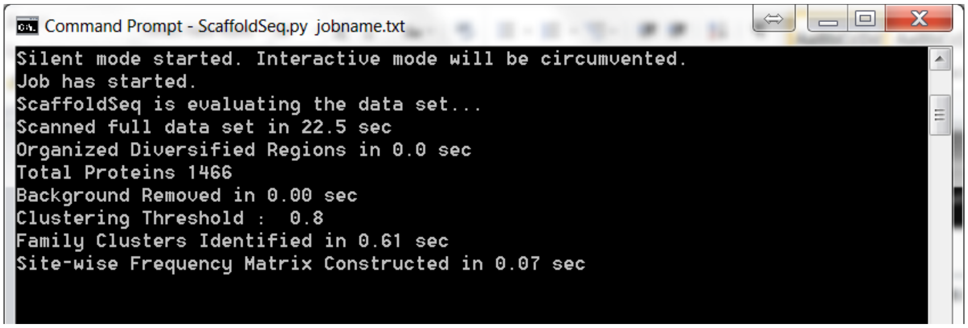
A sample job file is included in the software package. The file contents are shown below:

```
Job Name: Fibronectin_Fn3HP
FASTA/FASTQ File: High_affinity.fasta
Gene Start:
TCCTCCGACTCTCCGCGTAACCTGGAGGTTACCAACGCAACTCCGAACTCTCTGACTATTTCTTGG
5' Anchor: GCTAGC
3' Anchor: GGATCC
# of Diversified Regions: 3
DNA After Region:
TACCGTATCACCTACGGCGAAACTGGTGGTAACTCCCCGAGCCAGGAATTCACTGTTCCG,GCGACCATC
AGCGGTCTGAAACCGGGCCAGGATTATACCATTACCGTGTACGCTGTA,CCAATCAGCATCAATTATCGC
ACCGAAATCGACAAACCGTCTCAG
Minimum Region Length: 6,3,6
Maximum Region Length: 11,7,12
Insert after # Position: 3,1,3
Sequence Similarity Threshold: 0.8
Frequency Dampening Power: 1
Maximum Sequence Count: 10000
Assay Background Filter: On
Pairwise Analysis: Off
Filter Coefficient: 10
```

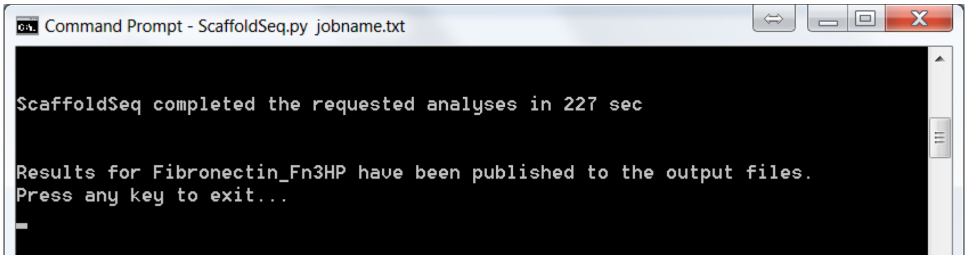
Upon running a job using silent mode, the window should display brief descriptions of progress:



```
Silent mode started. Interactive mode will be circumvented.
Job has started.
ScaffoldSeq is evaluating the data set...
```



```
Silent mode started. Interactive mode will be circumvented.
Job has started.
ScaffoldSeq is evaluating the data set...
Scanned full data set in 22.5 sec
Organized Diversified Regions in 0.0 sec
Total Proteins 1466
Background Removed in 0.00 sec
Clustering Threshold :  0.8
Family Clusters Identified in 0.61 sec
Site-wise Frequency Matrix Constructed in 0.07 sec
```



```


ScaffoldSeq completed the requested analyses in 227 sec


Results for Fibronectin_Fn3HP have been published to the output files.
Press any key to exit...
_
```

Output files will be exported to the working directory.

Runtime and Memory Requirements

The analyses discussed in this walkthrough were conducted on a standard desktop PC (Windows 10, Intel i5 4590 @3.3GHz, 16GB RAM). The RAM requirements are dictated by the total number of unique sequences being processed, while the overall runtime governed by the total number of clusters. As a representative high-RAM test case mimicking the analysis of a naïve or unselected population not requiring clustering, 1x10^7 unique clones were pseudo-randomly generated in the framework of the Gp2 scaffold with NNK codons at each diversified position. This analysis required 2 hours to run with a peak RAM usage of 4.2GB. With 10-fold fewer sequences, this process takes only 12 minutes (400MB). As a representative long-runtime test case mimicking the analysis of broadly diverse population of matured clones, 1x10^7 sequences (10% unique) were generated in the framework of the Gp2 such that 1x10^5 family clusters were organized. The more computationally demanding tasks of clustering added 4 hours to the total runtime. When allowing for 1x10^4 clusters, this results in an 18-fold reduction in time required for clustering.

Paired-end Assembly

Multiple algorithms[10–13] exist for processing paired-end reads. The above examples used PANDAseq[3] for assembling quality sequences into FASTA files. Below, is a basic template for using this method. The forward [-f] and reverse [-r] reads are input as separate FASTQ files. Multi-threading [-T] can be enabled for CPU-bound situations. A sequence quality threshold [-t] can be adjusted to reduce the presence of low-quality reads. This process generates an output file [-w].

```
module load pandaseq
FOR = HACKEL_S1_L001_R1_001.fastq
REV = HACKEL_S1_L001_R2_001.fastq
OUT = panda_assembled.fasta
pandaseq -f $FOR -r $REV -T 4 -t 0.99 –w >$OUT
```

References

1. Kruziki MA, Bhatnagar S, Woldring DR, Duong VT, Hackel BJ. A 45-Amino-Acid Scaffold Mined from the PDB for High-Affinity Ligand Engineering. *Chem Biol*. 2015;22(7):946-956.
2. Woldring DR, Holec P V, Zhou H, Hackel BJ. High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. *PLoS One*. 2015;10(9):e0138956.
3. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*. 2012;13(1):31.
4. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*. 2008;24(3):333-340.
5. Brown CA, Brown KS. Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! *PLoS One*. 2010;5(6).
6. Little DY, Chen L. Identification of Coevolving Residues and Coevolution Potentials Emphasizing Structure, Bond Formation and Catalytic Coordination in Protein Evolution. Shiu S-H, ed. *PLoS One*. 2009;4(3):e4762.
7. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins Struct Funct Genet*. 2004;56(2):211-221.
8. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins Struct Funct Genet*. 2002;48(4):611-617.
9. Durani V, Magliery TJ. *Protein Engineering and Stabilization from Sequence Statistics: Variation and Covariation Analysis.* Vol 523. 1st ed. Elsevier Inc.; 2013.
10. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*. 2014;30(5):614-620.
11. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011;27(21):2957-2963.
12. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120.
13. Cole JR, Wang Q, Fish JA, et al. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res*. 2014;42(D1):1-10.