Annotation of the *Kytococcus sedentarius* Genome from DNA Coordinates 05500 to 05520

Victoria Fabrizi*, Kailey Ferger*, Larissa Gaul* and Samantha Evans

Iroquois High School and The Western New York Genetics in Research Partnership *Indicates equal contribution

Abstract

A group of 3 consecutive genes from the microorganism *Kytococcus sedentarius* (Ksed_05500 – Ksed_05520) were annotated using the collaborative genome annotation website GENI-ACT. The Genbank proposed gene product name for each gene was assessed in terms of the general genomic information, amino acid sequence-based similarity data, structure-based evidence from the amino acid sequence, cellular localization data, potential alternative open reading frames, and the possibility of horizontal gene transfer. In general, it was found that the data obtained manually matched the computer's data, such as corresponding gene product names. Thus, the Genbank proposed gene product name did not differ significantly from the proposed gene annotation for each of the genes in the group and as such, the genes appear to be correctly annotated by in the database.

Introduction

This research project has several major objectives. Primarily, it aims to educate high school students about careers associated with bioinformatics. Furthermore, it hopes to promote the implementation of this subject material in the traditional science curriculum. The project also aims to instruct students on the usage of various gene annotation tools, including but not limited to, NCBI, BLAST, and T-COFFEE.
By exposing students to these scientific tools, it provides insight into future career paths, especially those in the STEM fields, and may provide valuable experience for future educational endeavors. Finally, to ensure increased accuracy, the *Kytococcus sedentarius* genome must be annotated manually and compared with the computer 's data (Western New York Genetics in Research Partnership).

Previously, the computer found the locus of each gene of *Kytococcus sedentarius* as well as the coordinates, the length, the gene products, the protein family, and the nucleotide and amino acid sequences of the genes. Before this research was carried out, bacterial organisms that were closely related to *Kytococcus sedentarius* had not had their genomes sequenced. Additionally, the conservation of particular regions of the amino acid sequence were unknown (GENI-ACT).

Currently, the study of *Kytococcus sedentarius* is being conducted partly by students through the Western New York Genetics in Research Partnership. This bacteria is being studied because it directly affects humans through pitted keratolysis, causing skin afflictions on the soles of the feet. The understanding of this bacteria's genome and characteristics will help provide information to combat this problem and similar bacterial problems. Furthermore, the study of this particular bacteria may lead to increased understanding of similar bacteria's genomes and characteristics.

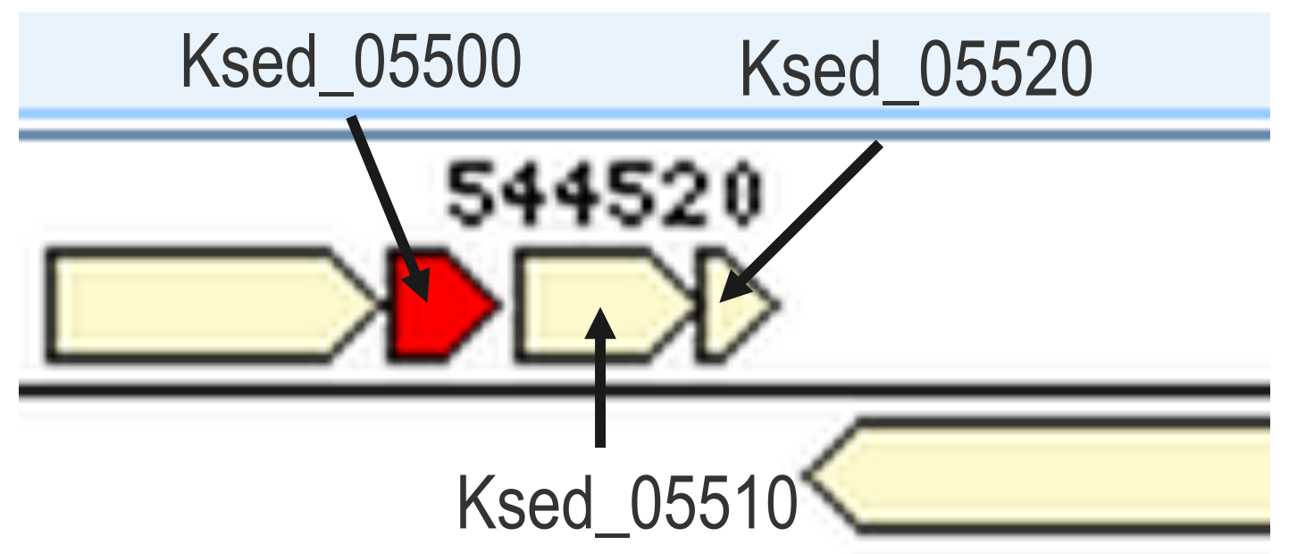

Figure I – Gene Neighborhood for Annotated Genes: Ksed_05500, Ksed_05510 and Ksed_05520

Methods and Materials

Modules of the GENI-ACT (http://www.geni-act.org/) were used to complete *Kytococcus sedentarius* genome annotation . The modules are described below:

Modules	Activities	Questions Investigated
Module 1- Basic Information Module	DNA Coordinates and Sequence, Protein Sequence	What is the sequence of my gene and protein? Where is it located in the genome?
Module 2- Sequence-Based Similarity Data	Blast, CDD, T-Coffee, WebLogo	Is my sequence similar to other sequences in Genbank?
Module 3- Cellular Localization Data	Gram Stain, TMHMM, SignalP, PSORT, Phobius	Is my protein in the cytoplasm, secreted or embedded in the membrane?
Module 4- Alternative Open Reading Frame	IMG Sequence Viewer For Alternate ORF Search	Has the amino acid sequence of my protein been called correctly by the computer?
Module 5- Structure-Based Evidence	TIGRfam, Pfam, PDB	Are there functional domains in my protein?
Module 8- Evidence for Horizontal Gene Transfer	Phylogenetic Tree	Has my gene co-evolved with other genes in the genome?

Results

Kytococcus sedentarius 05500: The computer lists this gene's protein family as GrpE. The Pfam database also determined this protein to be GrpE. Therefore, the computer's annotation appears to be correct. Furthermore, in reference to Figure II a and b, the HMM logo, there is a high amount of conservation at the end of the amino sequence. There is also a moderate amount of conservation toward the middle of the sequence. Due to the presence of a Shine-Dalgarno sequence 5-15 spaces before the start codon in the DNA coordinates and the fact that there are no other potential start codon locations, the computer is correct.

Kytococcus sedentarius 05510: The computer lists this gene's protein family as DnaJ domain with DnaJ C terminal region. The Pfam database also determined this protein to be Chaperone DnaJ. Therefore, the computer's annotation appears to be correct. Furthermore, in reference to Figure III a and b, the HMM logo, there is a high amount of conservation at the beginning and middle of the amino sequence. There is also a moderate amount of conservation toward the end of the sequence. Due to the presence of a Shine-Dalgarno sequence 5-15 spaces before the start codon in the DNA coordinates and the fact that there are no other potential start codon locations, the computer is correct.

Kytococcus sedentarius 05520: The computer lists this gene's protein family as MerR family regulatory protein. The Pfam database also determined this protein to be MerR HTH familiy regulatory protein. Therefore, the computer's annotation is correct. Furthermore, in reference to Figure IV, the HMM logo, there is a high amount of conservation toward the first third of the sequence as well as in the middle of the amino sequence. Due to the presence of a Shine-Dalgarno sequence 5-15 spaces before the start codon in the DNA coordinates and the fact that there are no other potential start codon locations, the computer is correct.

In these figures, the large letters represent high conservation for a particular section of the amino acid sequence. If there are several areas where large letters are present, it represents higher conservation and thus greater alignment with the computer's results. The variance in color represents different biochemical properties of amino acids present in the sequence.

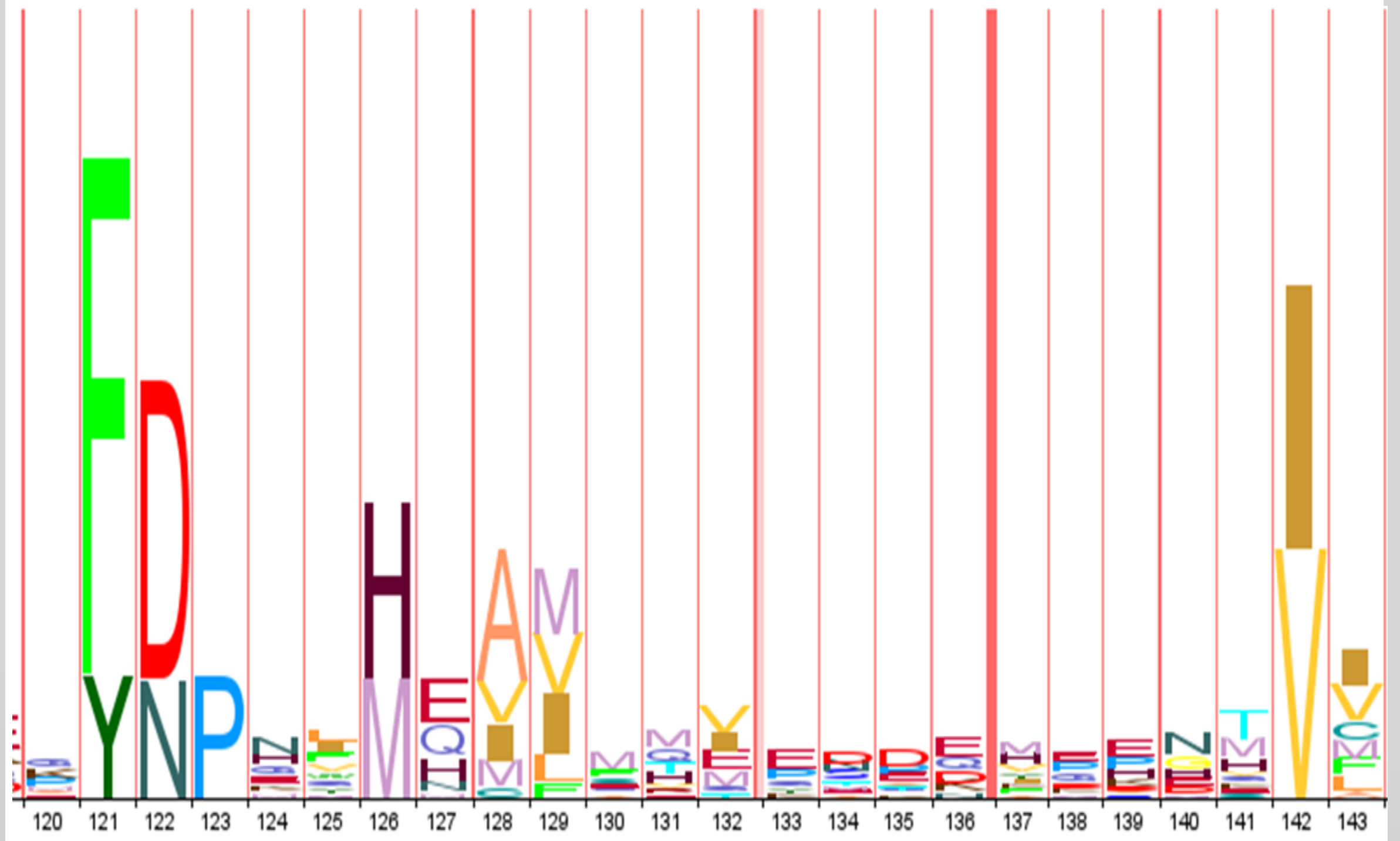

Figure II-a HMM Logo from Ksed_05500

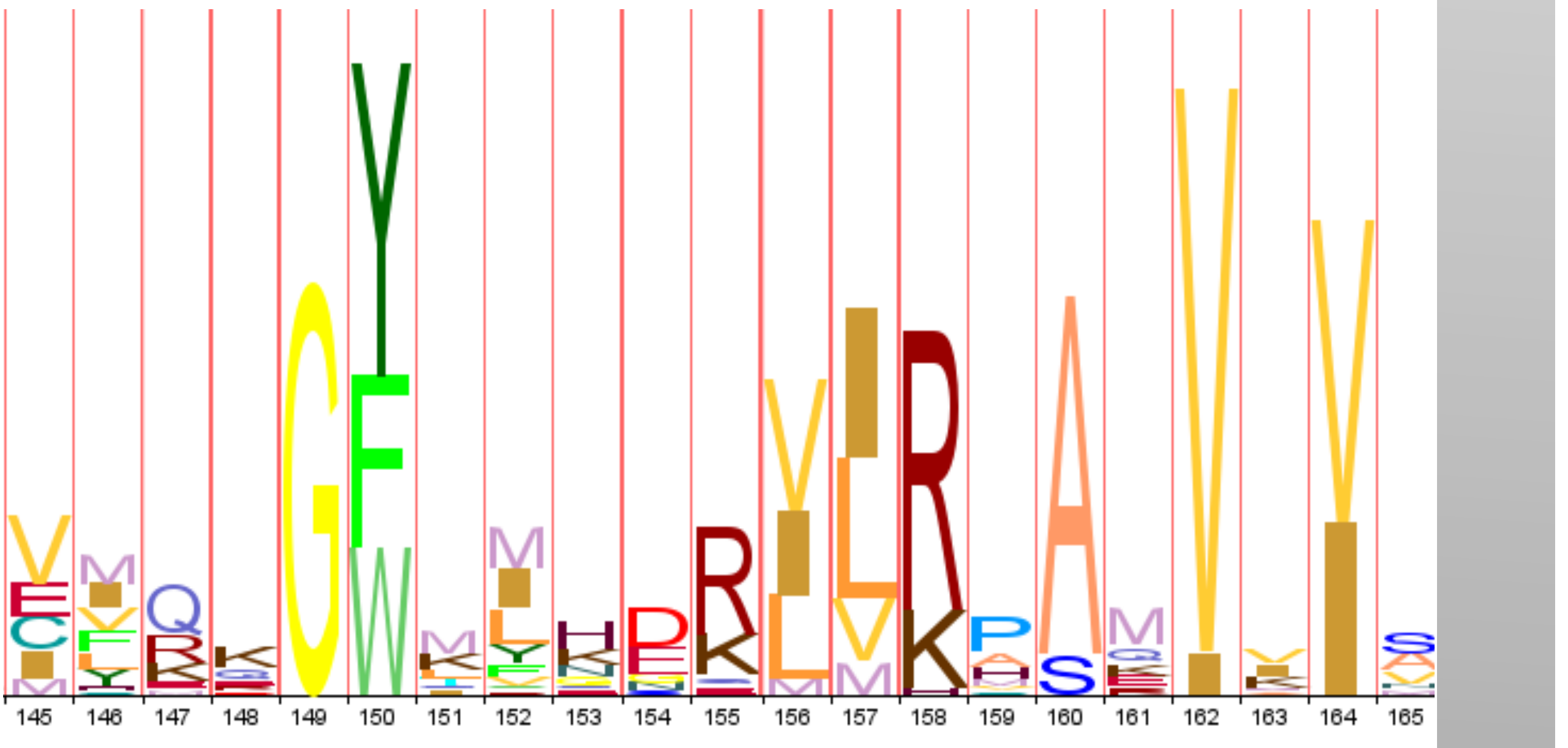

Figure II-b HMM Logo from Ksed_05500

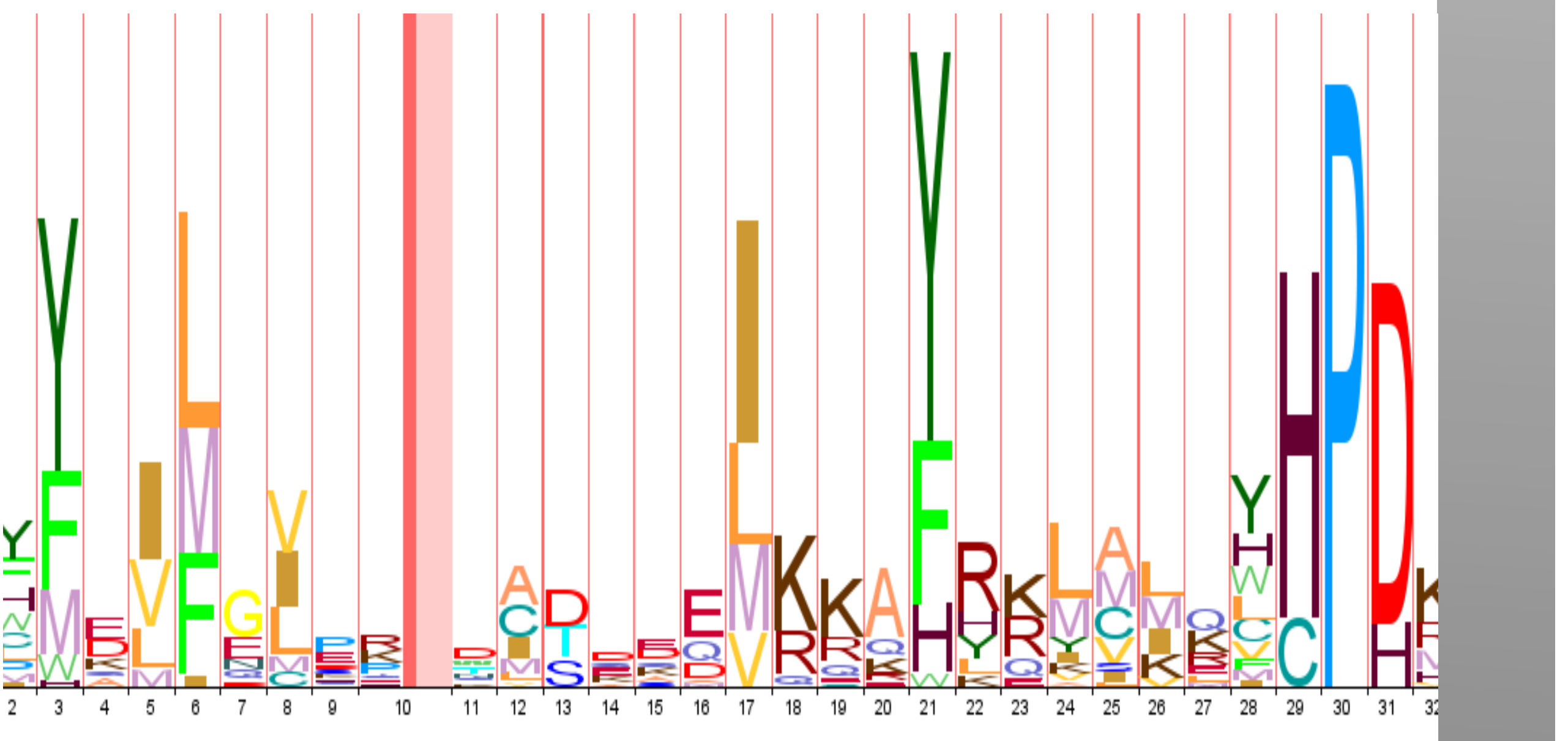

Figure III-a HMM Logo from Ksed_05510

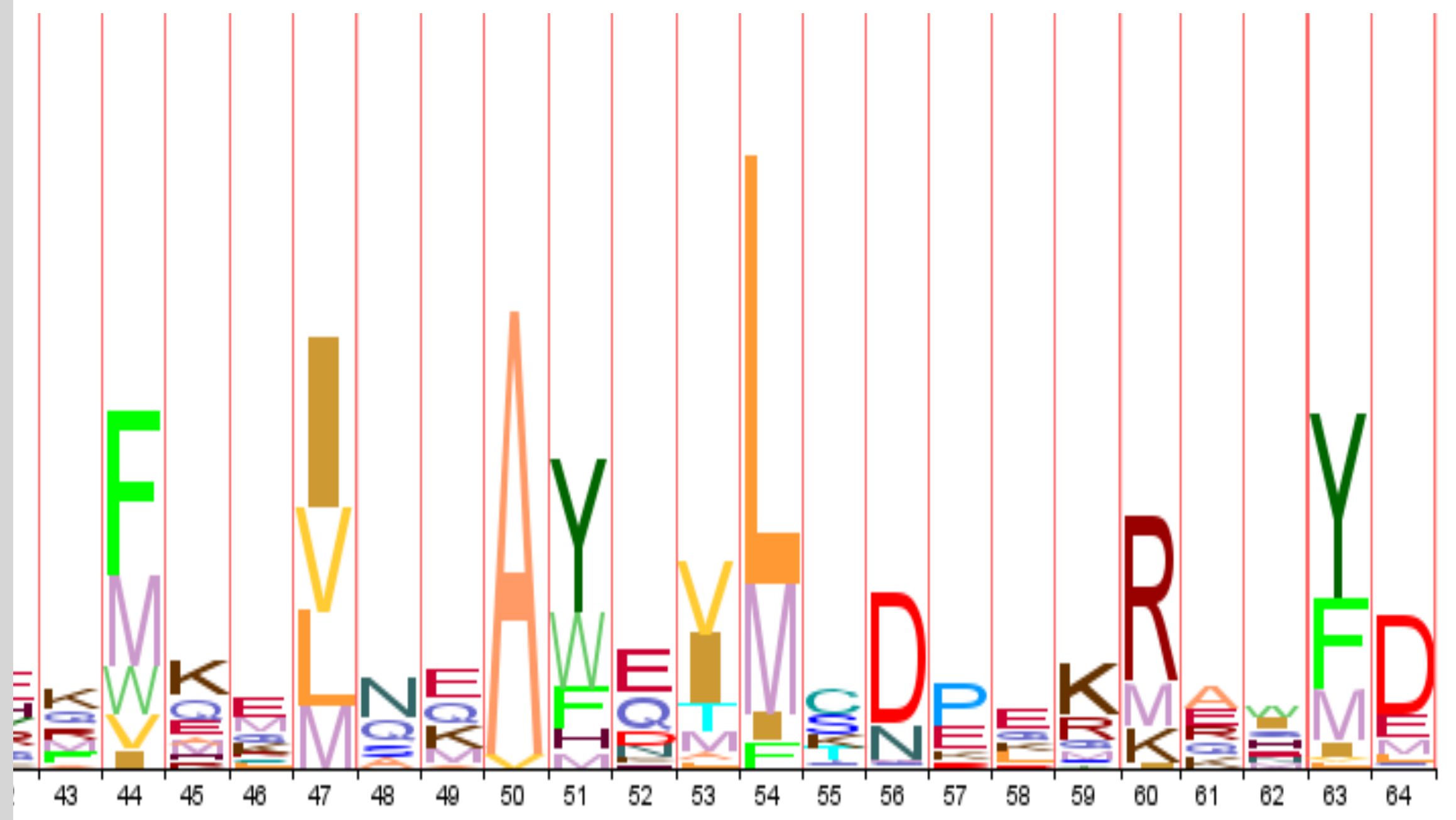

Figure III-b HMM Logo from Ksed_05510

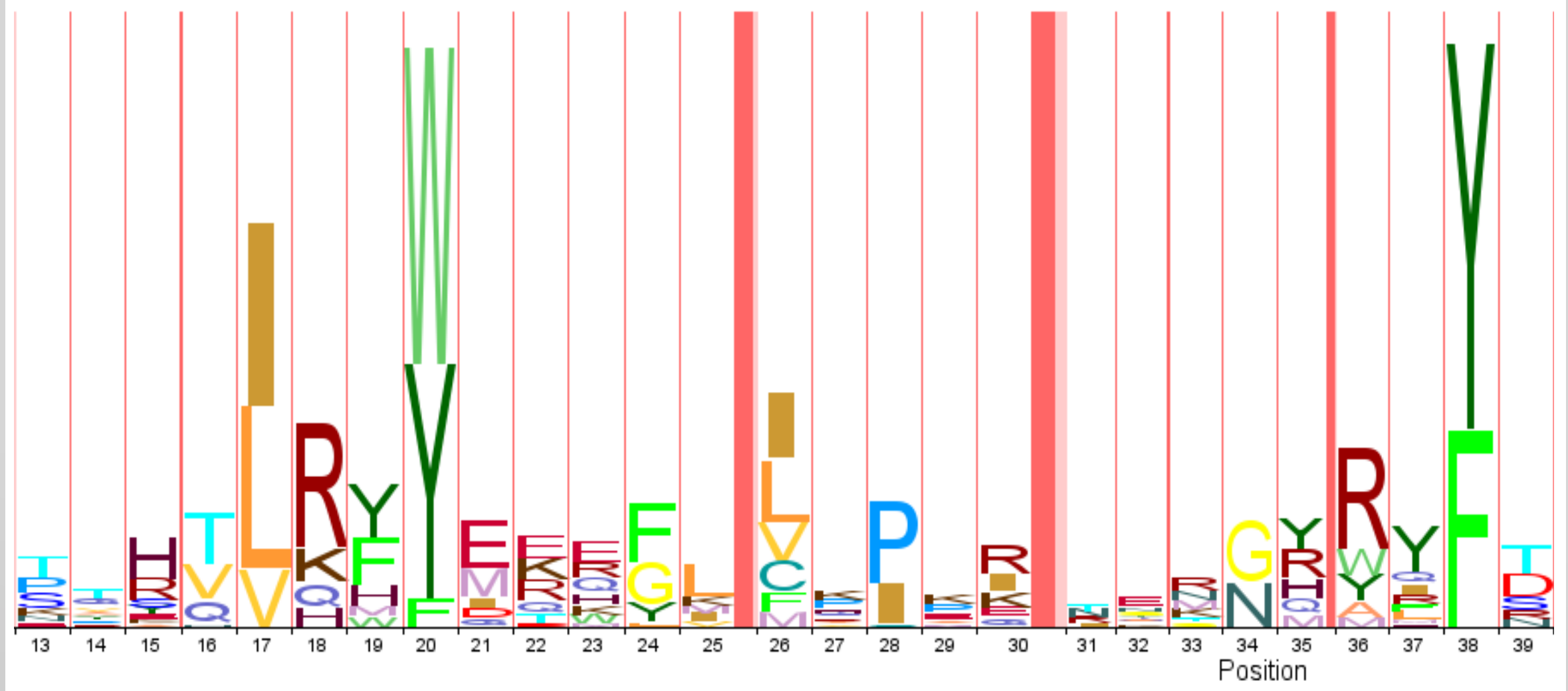

Figure IV HMM Logo from Ksed_05520

Conclusion

The GENI-ACT proposed gene product did not differ significantly from the proposed gene annotation for each of the genes in the group and as such, the genes appear to be correctly annotated by the computer database.

Gene Locus	Geni-Act Gene Products	Proposed Annotation
05500	Molecular chaperone GrpE (heat shock protein)	Molecular chaperone GrpE (heat shock protein)
05510	DnaJ-class molecular chaperone with C-terminalZn finger domain	Chaperone protein DnaJ 2
05520	Predicted transcriptional regulator	HTH-type transcriptional regulator GlnR

References

Western New York Genetics in Research Partnership. Home. Web. 2 June 2015.

Acknowledgements

Special thanks to Dr. Stephen Koury and Dr .Rama Dey-Rao for their assistance with this project. This work was supported by the National Science Foundation ITEST Strategies Award Number 1311902.