Michael Hart

Bioinformatics

Final Project Report

**Investigating a *C. auris* protein that could contribute to surface adhesion in clinical settings**

*C. auris* has recently emerged as a lethal threat in clinical settings due to its multidrug resistance. This danger is compounded by strong adhesive properties that make it difficult to remove from medical equipment (Kean et al. 2018). The proteins that give this adhesiveness have not yet been fully investigated, limiting the range of possible countermeasures (Muñoz et al. 2018; Kean et al. 2018).

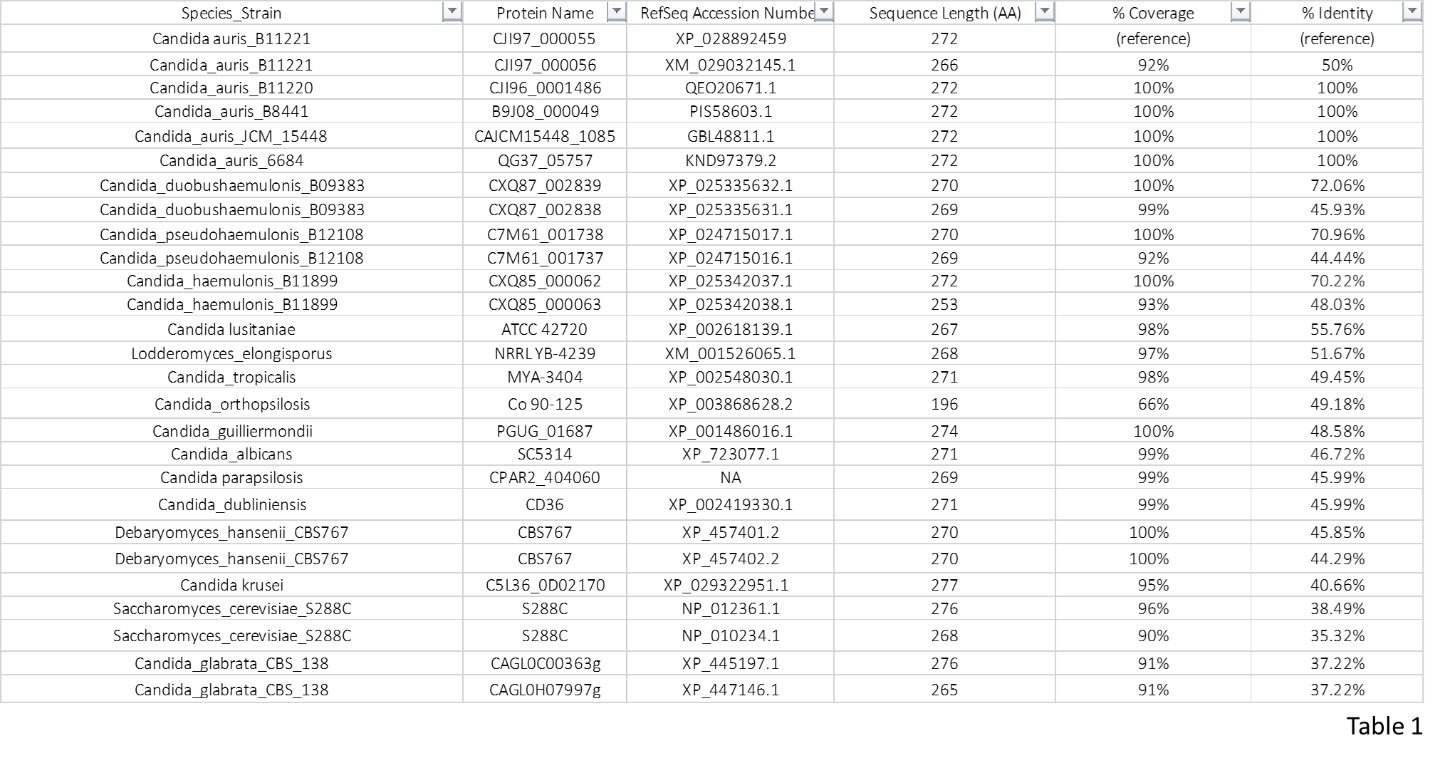
My research aim was to investigate the uncharacterized *C. auris* protein CJI97\_000055 to determine if it is an adhesin that contributes to *C. auris*’s strong adhesive properties. An InterProScan search revealed that CJI97\_000055 contains a large segment belonging to the glycophosphatidylinositol (GPI)-anchored superfamily, which has been implicated as a unique contributor to C. auris’s tenacious stickiness to medical devices (Kean et al. 2018; Muñoz et al. 2018). The GPI anchor is a modification added to the C-terminus of proteins after translation that allows the modified protein to be anchored in the cell membrane’s outer leaflet, in some cases contributing to cell-cell contact and adhesion (Paulick and Bertozzi, 2008). Several of the analyses I undertook will be detailed in this paper, including a multiple sequence alignment of homologs, phylogenetic trees, 3D protein structure prediction, and motif discovery.

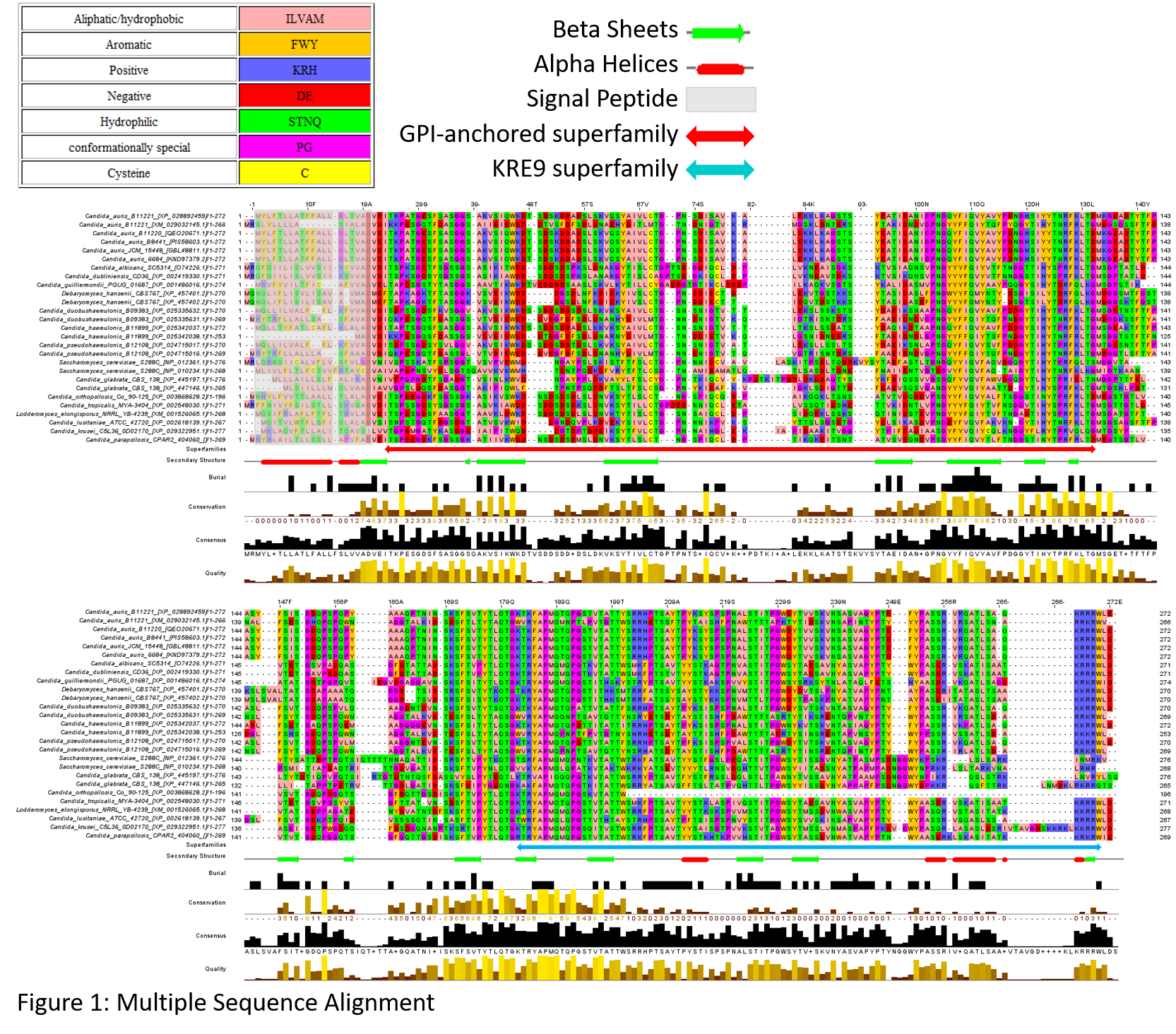
I found a good amount of evidence supporting the hypothesis that the protein is a contributor to *C. auris* adhesion. According to an InterPro search of CJI97\_000055, the protein contains one domain belonging to the GPI-anchored superfamily (implicated in surface adhesion) and another domain belonging to the KRE9 superfamily (involved in cell wall synthesis). My multiple sequence alignment shows relatively high conservation in both of these domains. Phylogenetic trees built from the multiple sequence alignment suggest that five different strains of *C. auris* carry identical homologs of CJI97\_000055, suggesting the importance of this protein for *C. auris*. Several other analyses add more weight to the theory that CJI97\_000055 is an important adhesin, but definite conclusions are hampered by a lack of structural characterization of the protein and limited experimental evidence from gene knockout experiments.

**Gene Structure and Function:**

**Multiple Sequence Alignment**

Running BLAST searches of CJI97\_000055’s protein sequence revealed the homologs listed in Table 1, which I put into a multiple sequence alignment using MUSCLE in the program MEGA-X. Figure 1 is the multiple sequence alignment built in Jalview with the Zappo color configuration (which colors by amino acid properties) and secondary structures predicted by JPred. Signal peptide and superfamily annotation was added based on the InterProScan results for CJI97\_000055.



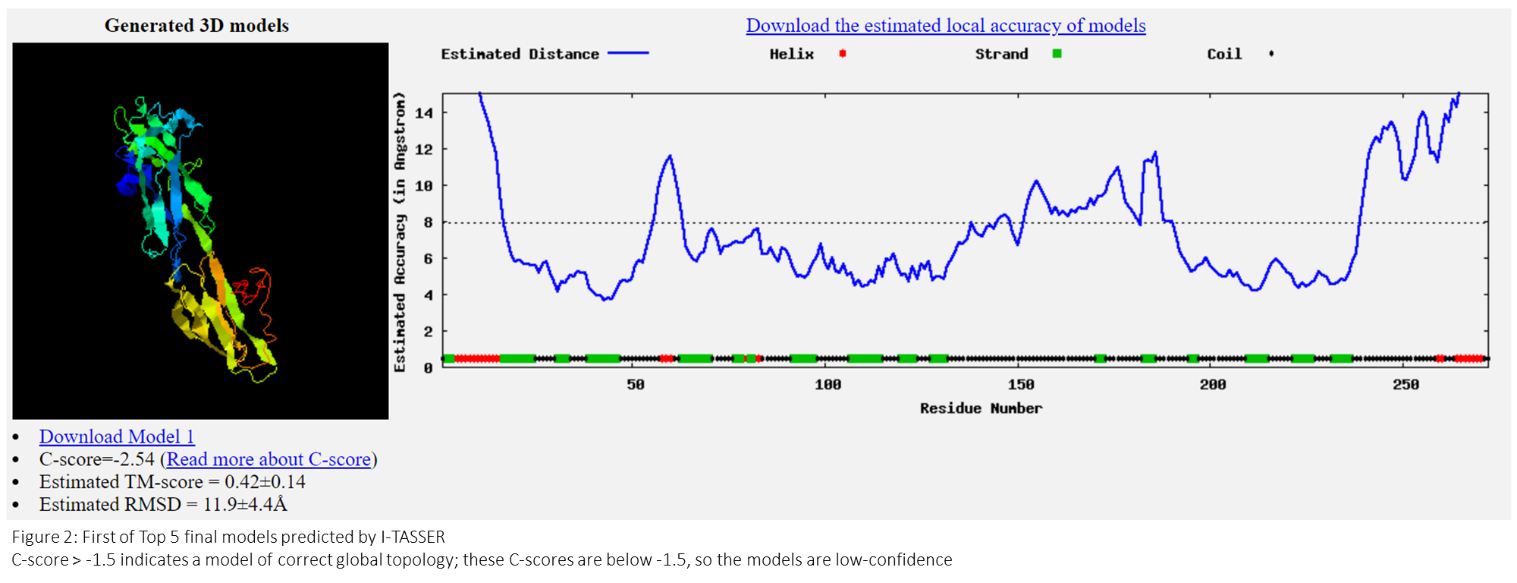


The multiple sequence alignment shows good conservation for most of the non-gap regions of the GPI-anchored domain, with some amino acids retaining 100% conservation for many residues (ex. the proline at 26 bp, the tryptophan at 45 bp, and the leucine at 73 bp). For the KRE9 domain the conservation quickly drops halfway through. The signal peptide region at the beginning is mostly hydrophobic and the KRE9 domain is largely hydrophilic (which suggests that this region is not entirely buried in the cell wall), but the GPI-anchored domain contains a broad mix of amino acid properties. There are about 12 predicted beta sheets and about four predicted alpha helices. The GPI-anchored domain is mostly beta sheets, and the KRE9 domain is a mix of beta sheets and alpha helices.

**Gene Structure and Function:**

**Protein Structure Prediction**

CJI97\_000055 is an uncharacterized protein and did not return any good template matches from BLAST Protein Database or Swiss Model searches, so I used the protein threading method of the program I-TASSER to generate predicted models of the protein’s structure.



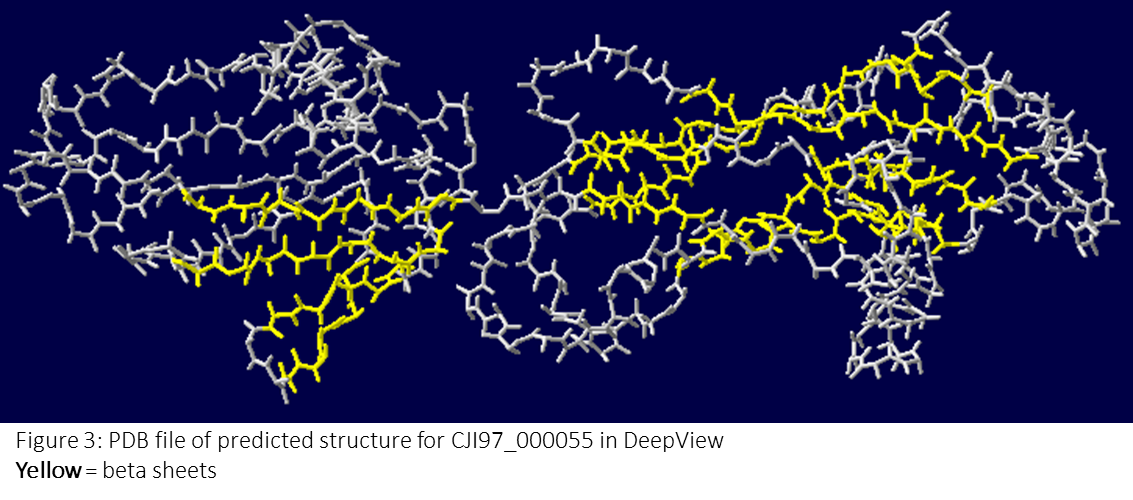


Figure 2 is the first of the “Top 5” final models predicted by I-TASSER. The chart to the right of the protein shows the estimated accuracy (in angstroms) of each region of the secondary structure, revealing higher predicted precision for most of the beta sheets but lower predicted precision for the alpha helices. Figure 3 is the image produced by SwissProt DeepView when given the PDB file of the first of the “Top 5” final models predicted by I-TASSER. I colored the secondary structures, revealing eight beta sheets (yellow) clustered in fours on each side of the protein. I was initially surprised that the predicted alpha helices at the beginning and end of the protein were not depicted in the PDB DeepView file; this may be because those regions have lower conservation and confidence in their structure prediction.

I-TASSER also provides a report of gene ontology terms predicted to be associated with CJI97\_000055. The gene ontology terms predicted from the top three templates were “cell wall”, “binding”, “pathogenesis”, and “cholesterol binding”. The gene ontology terms associated with the consensus prediction of all of the templates included “sterol binding”, “multi-organism process”, and ”cell wall”. Some of these gene ontology terms (“binding”, “cell wall”) make sense for CJI97\_000055 because it is known to contain a GPI-anchored domain involved in adhesion and the KRE9 domain involved in cell wall synthesis, but it is difficult to determine how the other terms may or may not relate to surface adhesion in *C. auris*.

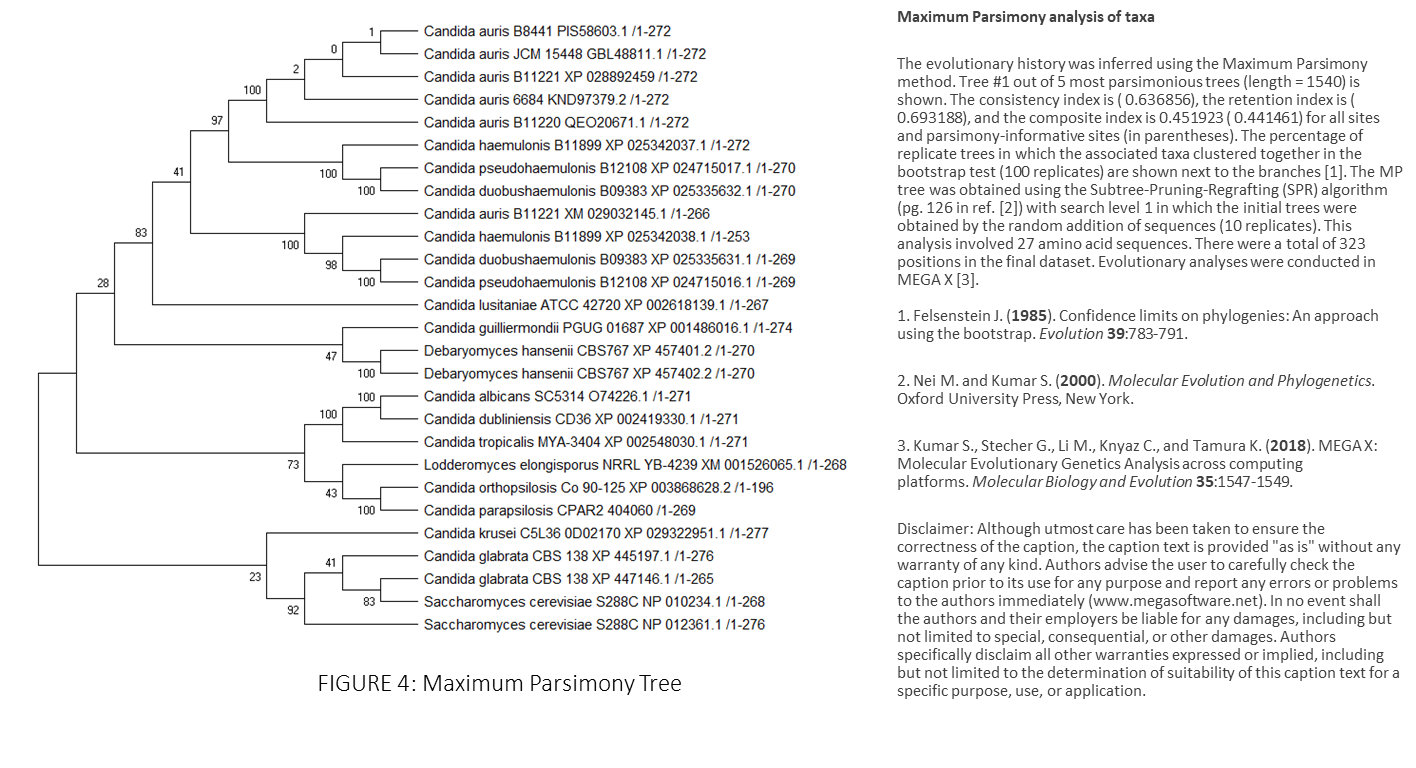
Regarding all of the results from I-TASSER, the confidence-related scores (Z-score, C-score, etc.) were all relatively poor, and either barely met a confidence threshold or fell below it. For example, for Figure 2 the C-Score is measured between 0-1 and my C-scores are all 0.1 or lower, so the results are extremely low-confidence. This indicates that the protein CJI97\_000055 is a hard target to model and that the I-TASSER results should be cited with caution.

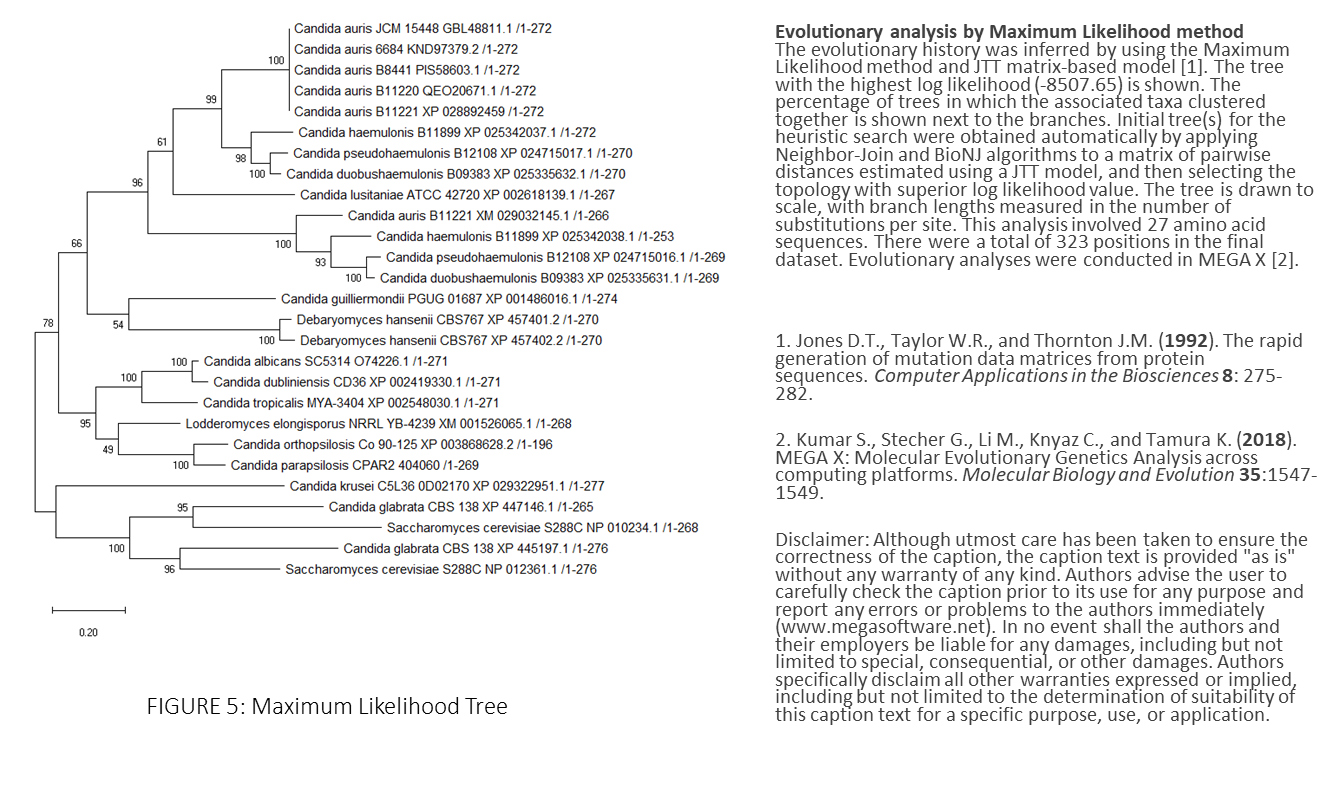
**Gene & Genome Evolution:**

**Phylogenetic Analysis**

I undertook a phylogenetic analysis to gather evidence regarding when and in what species CJI97\_000055 expanded compared to closely related Candida species. Figure 4 and Figure 5 are maximum parsimony and maximum likelihood trees, respectively. They were generated from a multiple sequence alignment of CJI97\_000055 homologs in the MEGA-X program. Both my maximum parsimony tree and maximum likelihood trees match fairly well with Fig. 3 of Muñoz et al. 2018 (which is a maximum likelihood tree) when rooted on the branch containing *C. krusei*, *C. glabrata*, and *S. cerevisiae*, which is assumed to be the outgroup. One striking feature of both trees is that protein sequences from five strains of *C. auris* (B11221, B11220, B8441, JCM\_15448, 6884) are grouped in one clade with a bootstrap value of 100; referencing the multiple sequence alignment in Figure 1 also reveals that there are zero amino acid substitutions in these five sequences. This suggests that the protein is critical for functionality in the *C. auris* life cycle and cannot tolerate amino acid substitutions, thus preventing any fixation of new mutations.

My phylogenetic trees both suggest that the CJI97\_000055 protein underwent duplication events followed by speciation events in strains of the species *C. auris*, *C. duobushaemulonis*, *C. haemulonis*, *S. cerevisiae*, and *C. glabrata*. The most noticeable difference between the trees is that the maximum likelihood tree has S. cerevisiae and C. glabrata display a duplication followed by two speciation events, whereas the maximum parsimony tree shows a speciation followed by one duplication. The maximum likelihood tree has higher bootstrap values for these branchings and requires fewer evolutionary events, so according to the parsimony principle the maximum likelihood tree is in fact more likely to be correct.

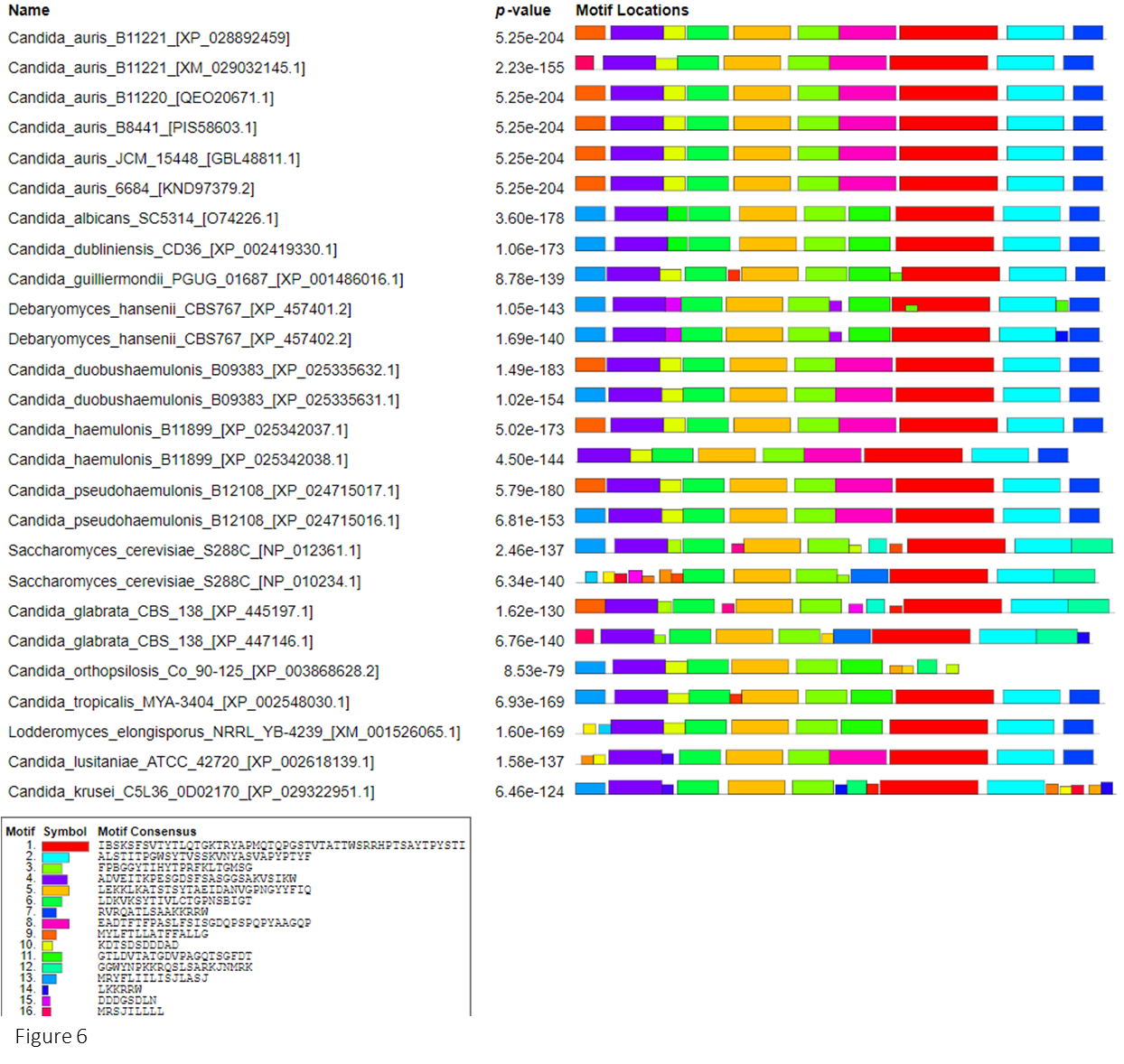




**Global Analyses & Models:**

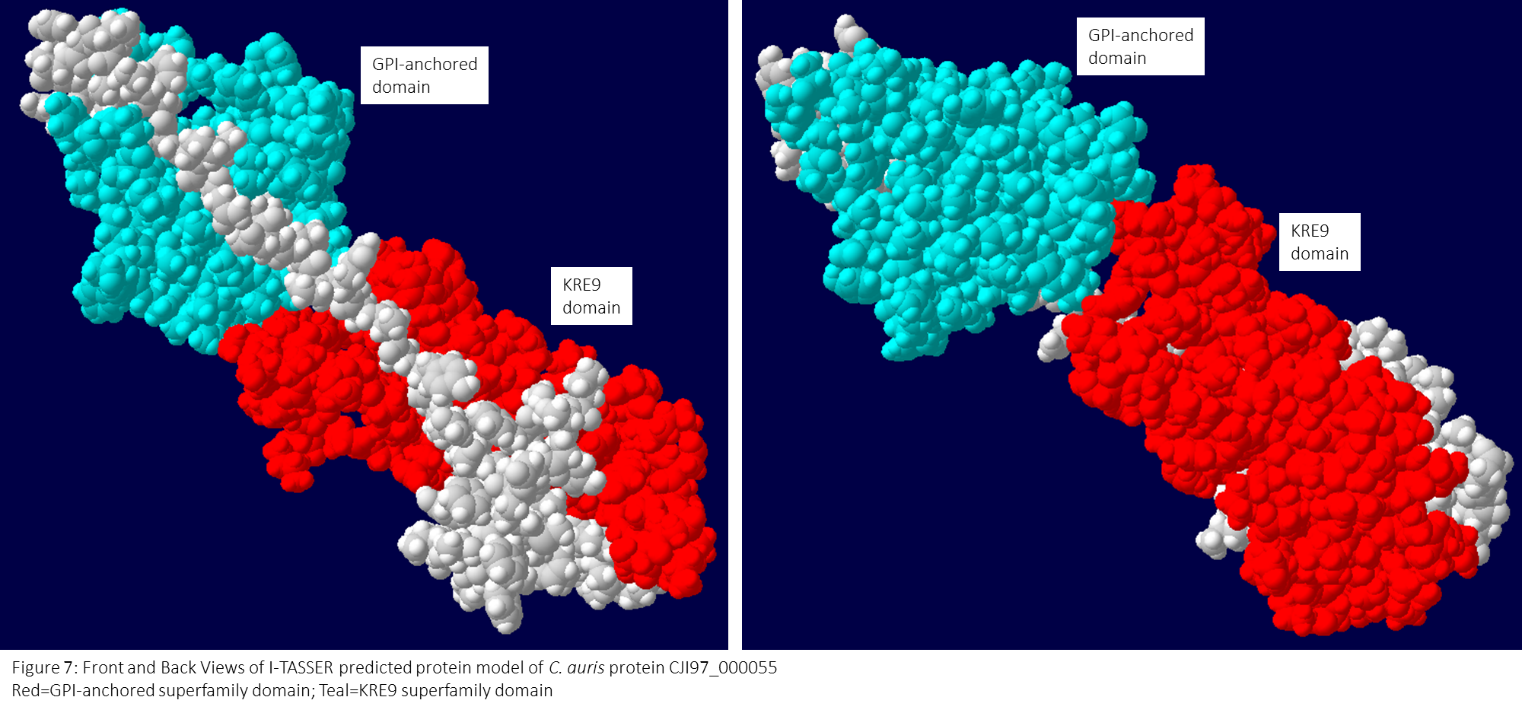
**Motif Discovery**

Figure 6 shows the motif discovery output from the MEME program. When searching for a total of 100 motifs between 6 and 50 amino acids long, MEME returned many motifs that are conserved in every sequence from the multiple sequence alignment. The most highly conserved motifs appear in the ~20 to ~130 range and ~180 to ~280 range, which overlaps with the predicted location of the GPI-anchored superfamily domain and KRE9 superfamily domain. Serine and threonin (hydrophobic) appear the most in the MSA, with frequences of 0.119 and 0.125, respectively. The largest motif (1) appears in the KRE9 domain.

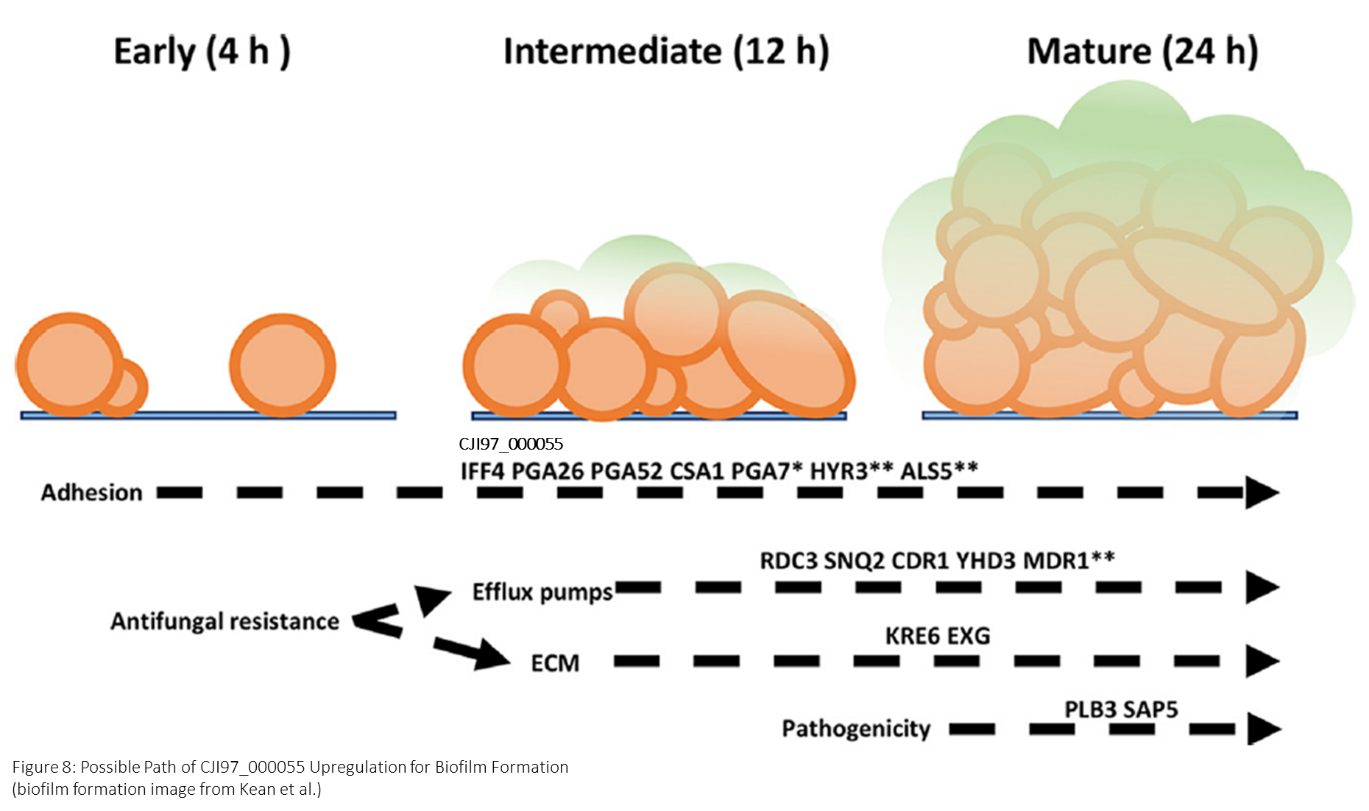


**Global Analyses & Models:**

**Cartoon Summary Model**



Experiments with *C. auris* have shown that adhesin-related GPI-anchored cell wall genes experience upregulation throughout the formation of biofilm, suggesting that GPI-anchored domains play an important role within cellular adhesion (Kean et al. 2018). Figure 7 is an I-TASSER-predicted PDB file in DeepView with the GPI-anchored domain colored teal and the KRE9 domain colored red, shown from two different angles. Although my protein is not confirmed to be an adhesin, the cartoon summary model in Figure 8 (adapted from an original figure presented in Kean et al.) indicates that CJI97\_000055 upregulation should be seen at the intermediate (12 hour) timepoint during *C. auris* adhesion.



Discussion

CJI97\_000055 probably contributes to *C. auris*’s strong adhesive properties. The strongest evidence to support this is the fact that multiple databases and algorithms recognized CJI97\_000055 as containing a GPI-anchored domain, and proteins that have this domain have been shown to be upregulated during *C. auris* biofilm formation (Kean et al. 2018). Additionally, FungalRV’s “Adhesin Predictor” program scored the CJI97\_000055 amino acid sequence as 1.1209, which is above the threshold for being classified as an adhesin. The I-TASSER results are less convincing because although the CJI97\_000055 protein model and its predicted templates sequences were associated with gene ontology terms such as “cell wall” and “binding” that could be related to biofilm adhesion, the relatively low confidence-related scores associated with the I-TASSER models indicate that these results should be cited with caution and highlight the need for actual structural characterization of the CJI97\_000055 protein to confirm the results.

There are several wet lab experiments that could be performed to determine if CJI97\_000055 is actually a *C. auris* adhesin. It has been observed that in *C. auris* some GPI-linked cell wall proteins are upregulated in early biofilm formation which highlights their importance in the first stage of adherence (Kean et al. 2018). If CJI97\_000055 is indeed an adhesin, it should also be upregulated during biofilm formation as demonstrated in figure 8. This could be tested by a replication of the Kean et al. experimental procedure that also measures levels CJI97\_000055 translation. Additionally, simple gene knockout experiments could be performed on CJI97\_000055 to see if the absence of the protein leads to less effective biofilm formation and surface adhesion. Since there are many proteins involved in GPI-anchored adhesion, knocking out different amounts of the proteins simultaneously could help determine which proteins play the most critical role in surface adhesion.

References

Kean R., Delaney C., Sherry L., Borman A., Johnson EM., Richardson MD., Rautemaa-Richardson R., Williams C., Ramage G. Transcriptome assembly and profiling of Candida auris reveals novel insights into biofilm-mediated resistance. mSphere 3:e00334-18. (2018).

Muñoz J.F., Gade L., Chow N.A. et al. Genomic insights into multidrug-resistance, mating and virulence in Candida auris and related emerging species. Nat Commun 9, 5346 (2018) doi:10.1038/s41467-018-07779-6.

Paulick M. G., Bertozzi C. R. The glycosylphosphatidylinositol anchor: a complex membrane-anchoring structure for proteins. Biochemistry. 2008;47(27):6991–7000. doi: 10.1021/bi8006324.

Bailey T. and Elkan C. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers". Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California. (1994).