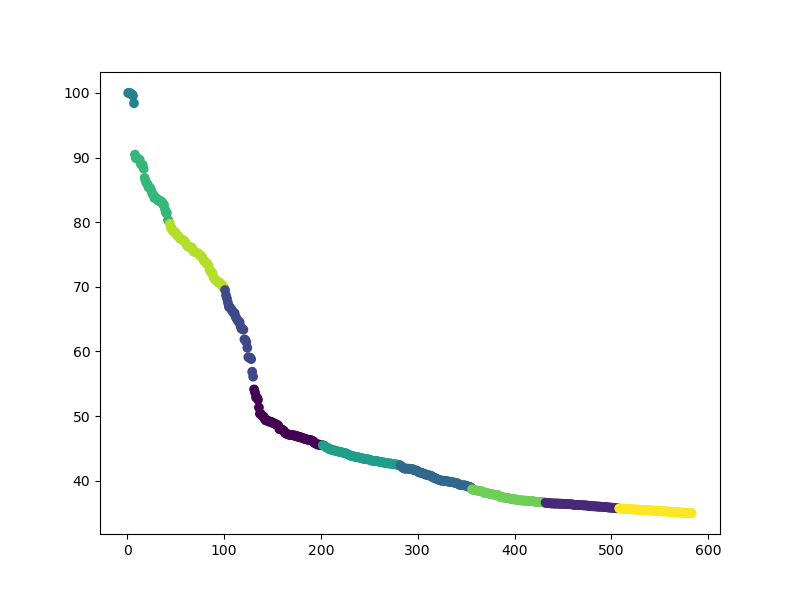
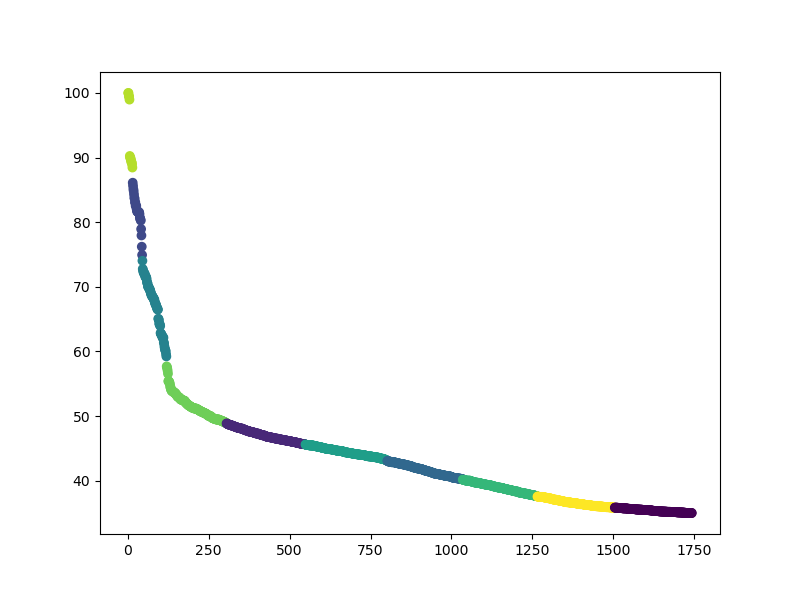
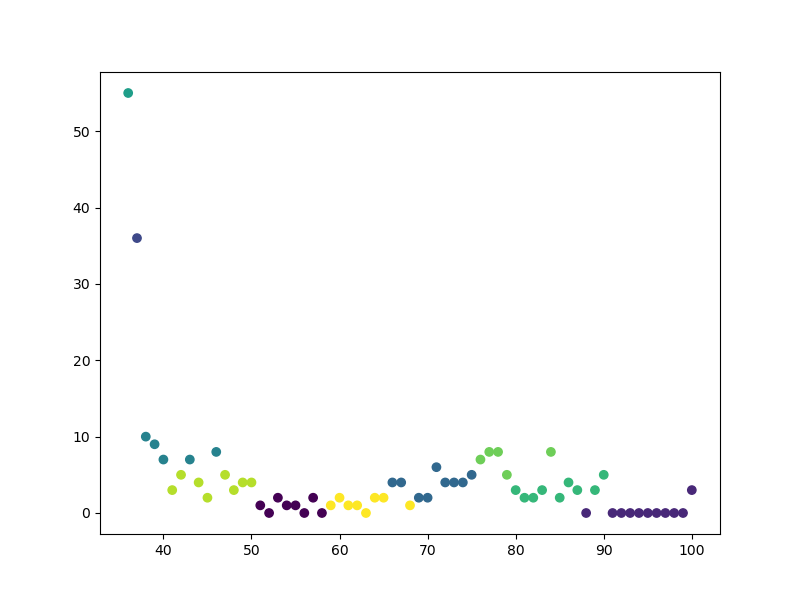
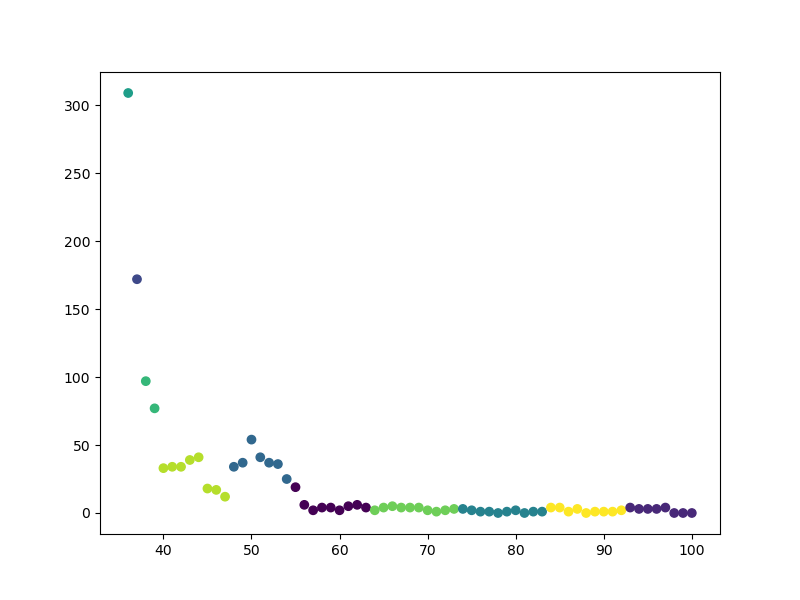
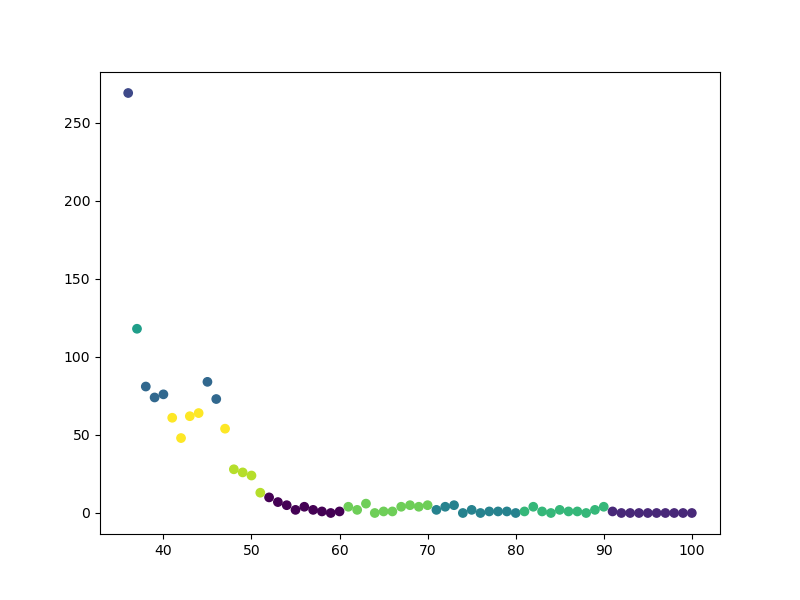
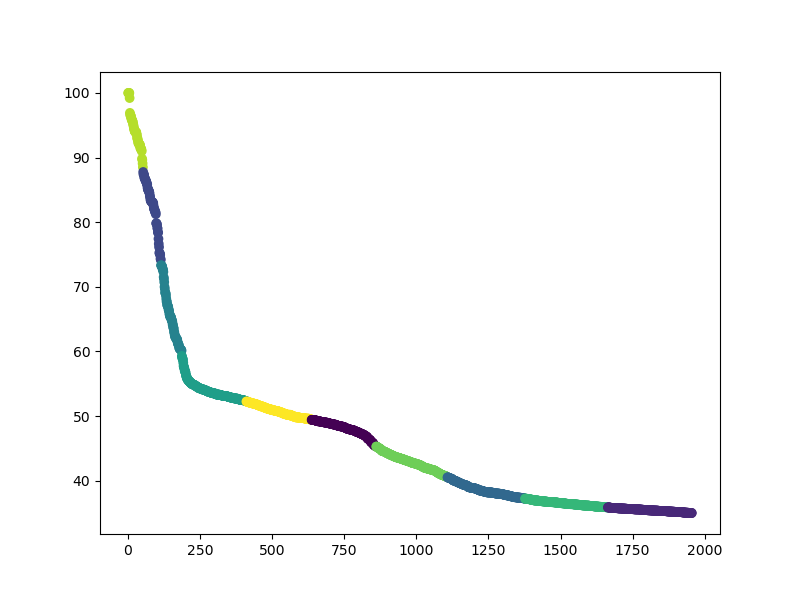
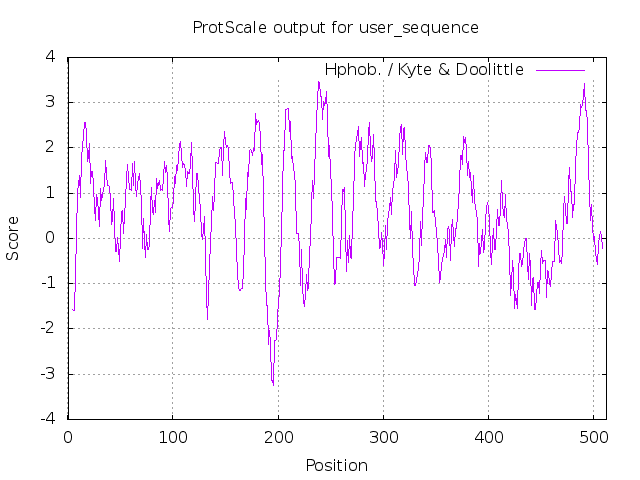
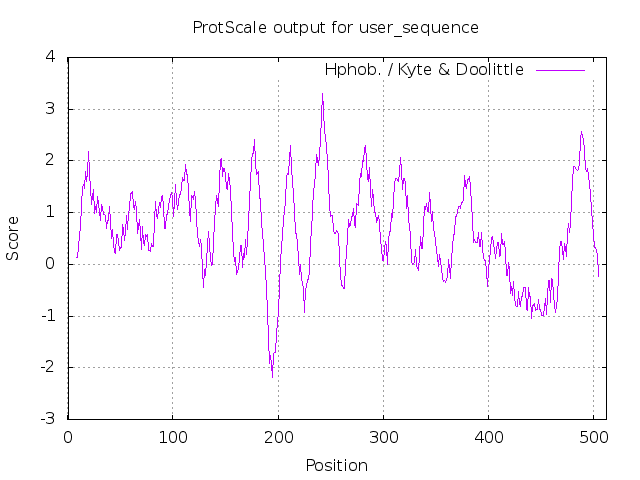
**Supplementary Data**

**K-means clustering**



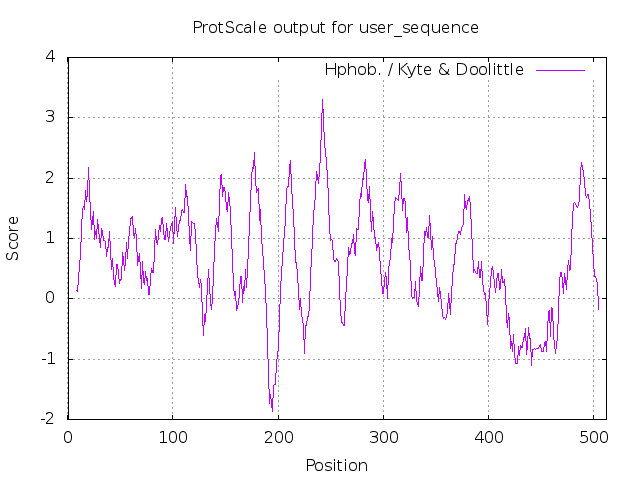
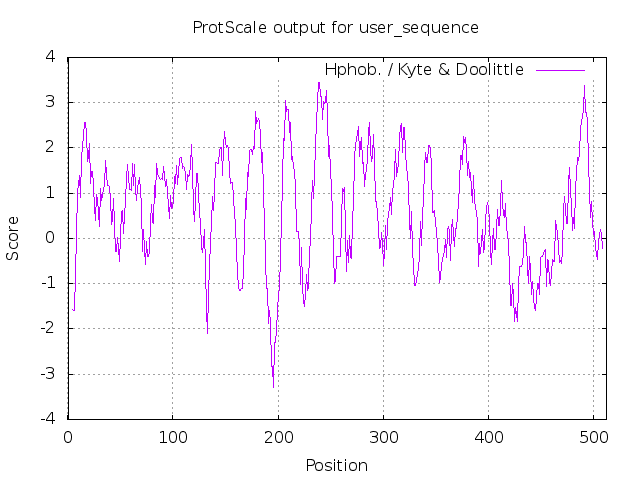
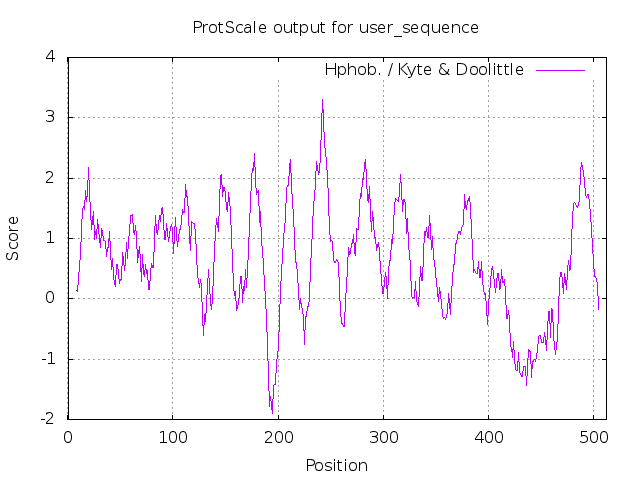
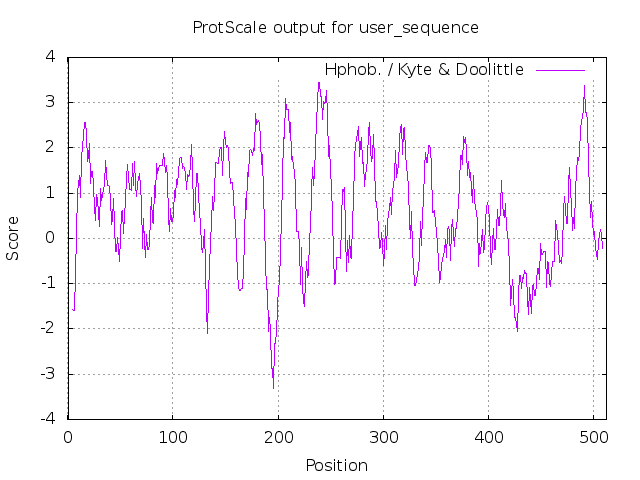
*Figure 1: K-means clustering analysis of frequency of hits in each cluster of identity and overall count of hits for bacterial species with homologs of EmrA (top left for frequency and top right for count), EmrB (middle left for frequency and middle right for count), and TolC (bottom left for frequency and bottom right for count). The Y-axis of clustering by frequency represents the frequency of hits per cluster, with the X-axis representing the identity values (%). They Y-axis of clustering by count represents the identity (%), with X-axis representing count.*

**Hydropathy Plots (ExPaSy’s ProtScale)**

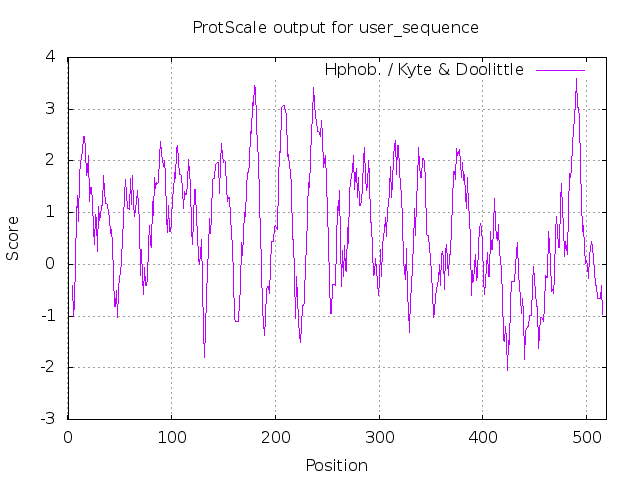
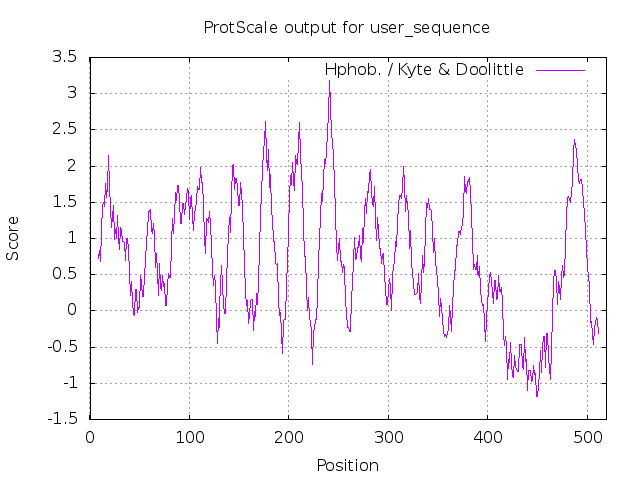


*Figure 2: Kyte-Doolittle hydropathy plots of K. pneumoniae EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*

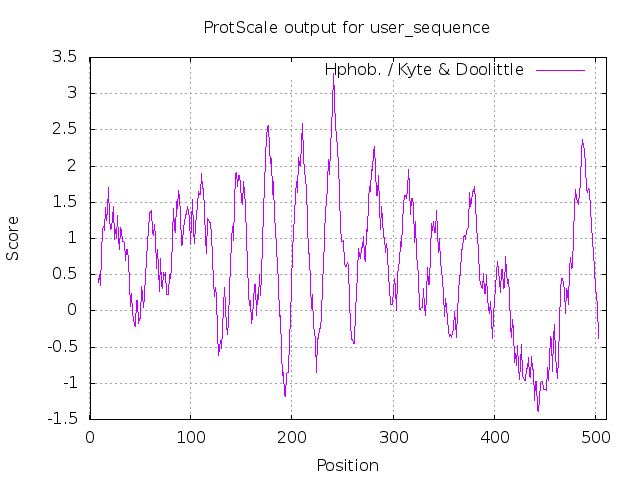
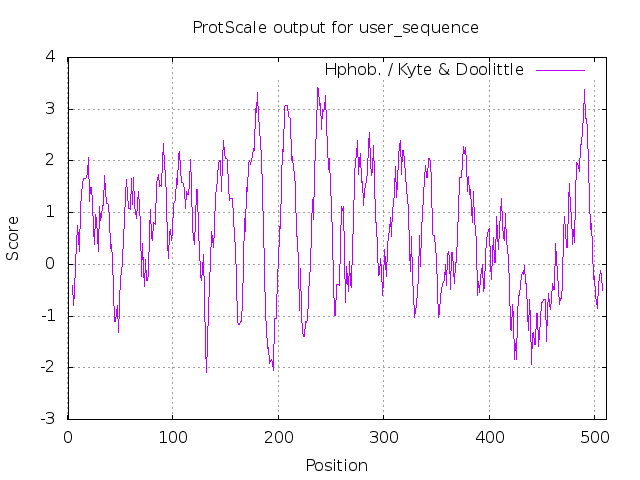
*Figure 3: Kyte-Doolittle hydropathy plots of C. koseri EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



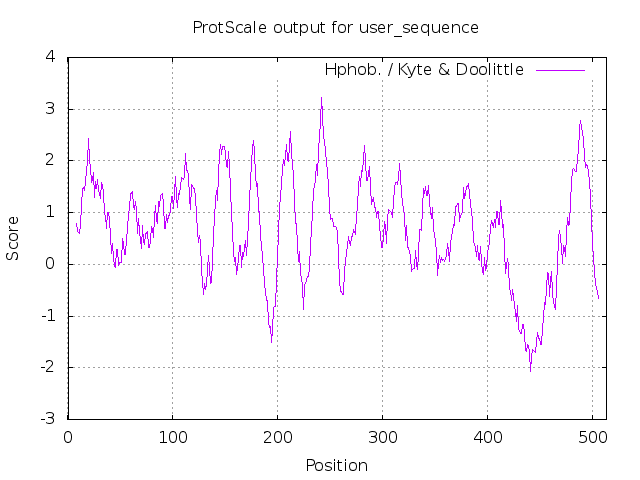
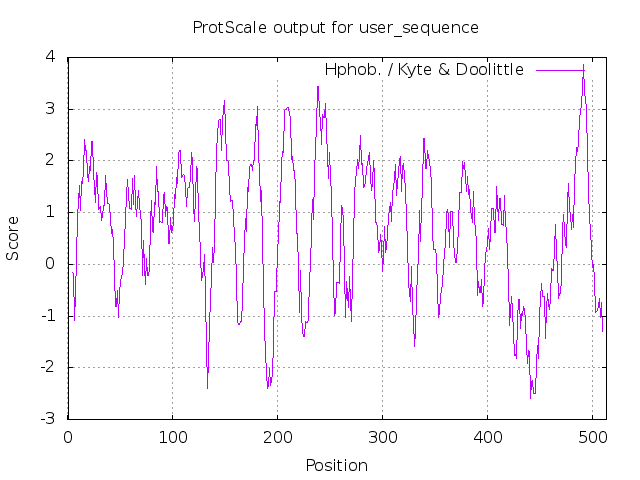
*Figure 4: Kyte-Doolittle hydropathy plots of S. typimurium EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



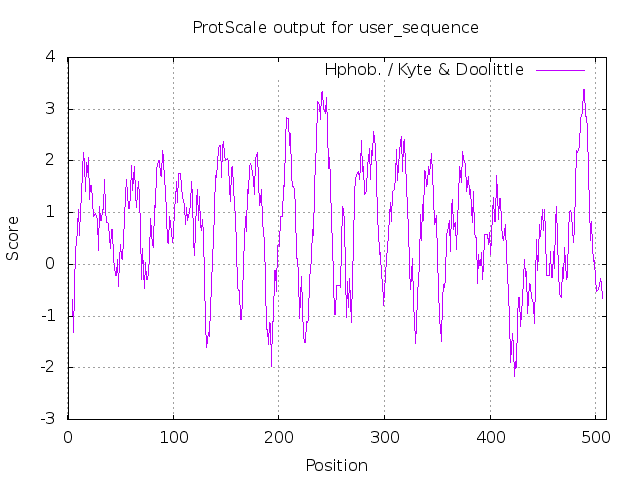
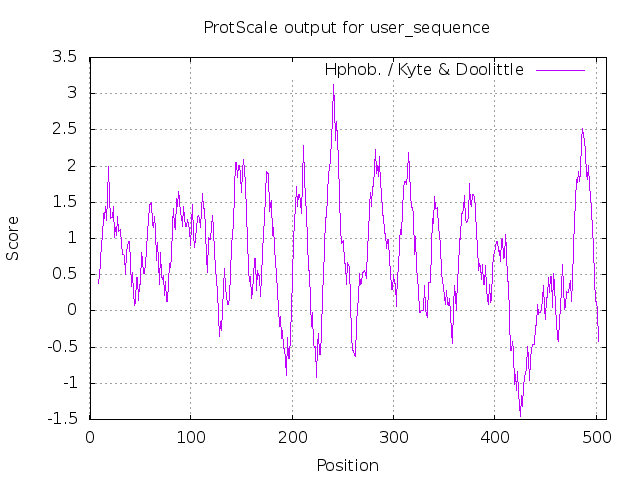
*Figure 5: Kyte-Doolittle hydropathy plots of Erwinia sp. Leaf53 EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



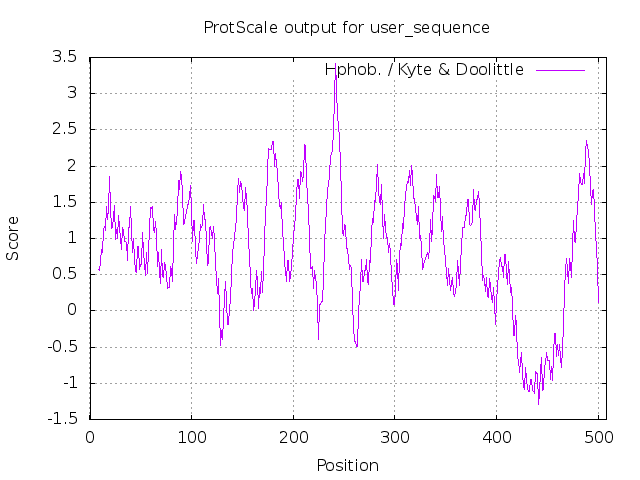
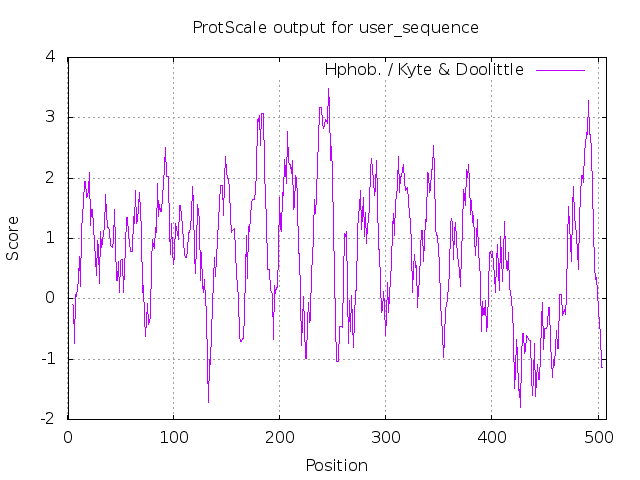
*Figure 6: Kyte-Doolittle hydropathy plots of Y. pestis EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



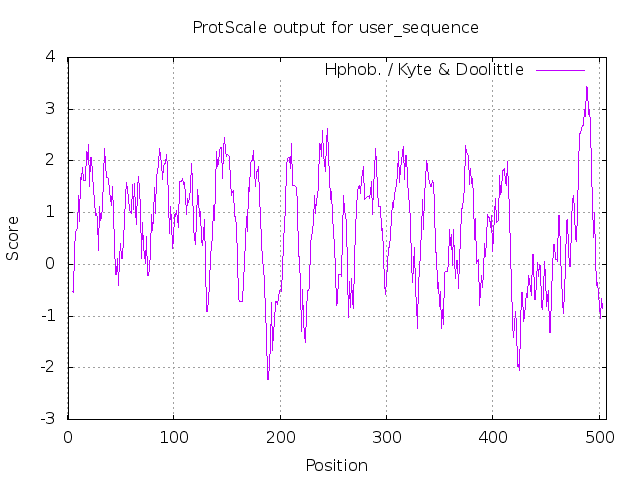
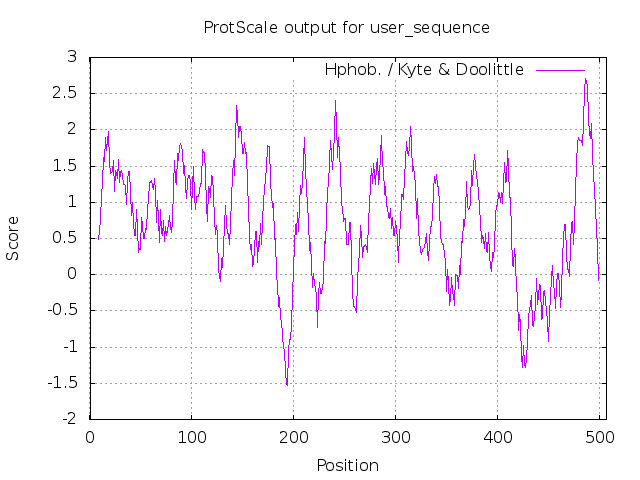
*Figure 7: Kyte-Doolittle hydropathy plots of X. bovienii EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



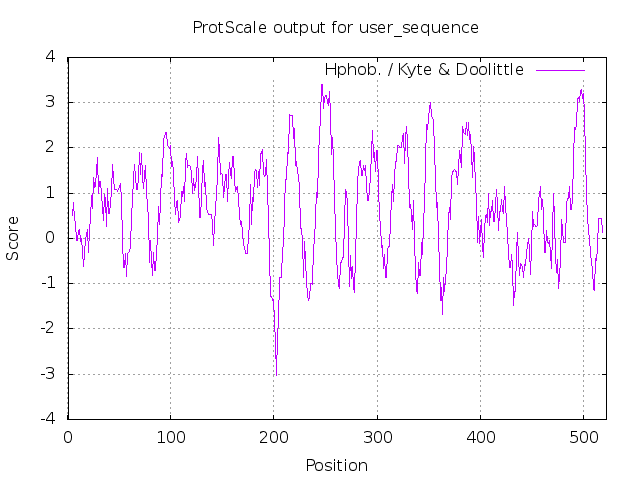
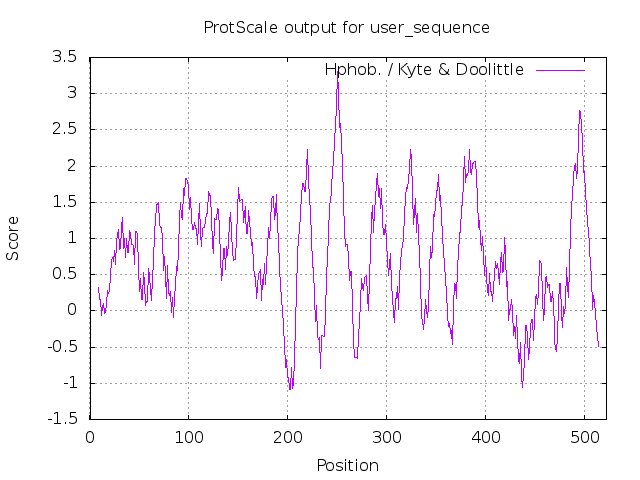
*Figure 8: Kyte-Doolittle hydropathy plots of Chromobacterium LK11 EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



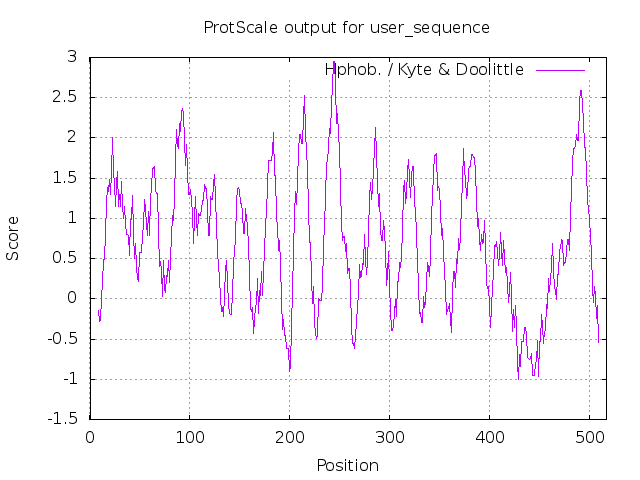
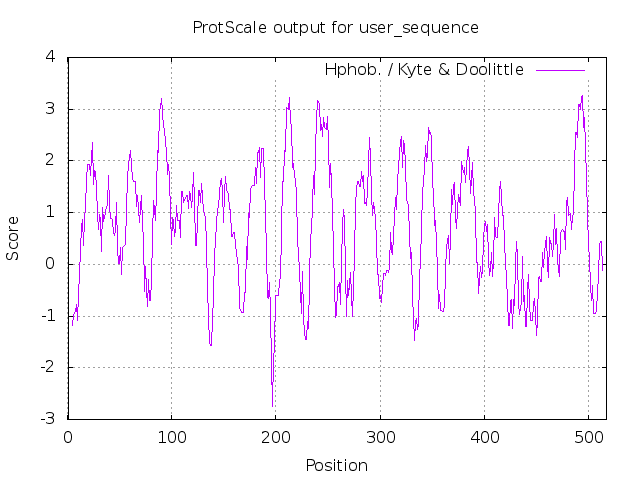
*Figure 9: Kyte-Doolittle hydropathy plots of P. heimbachae EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



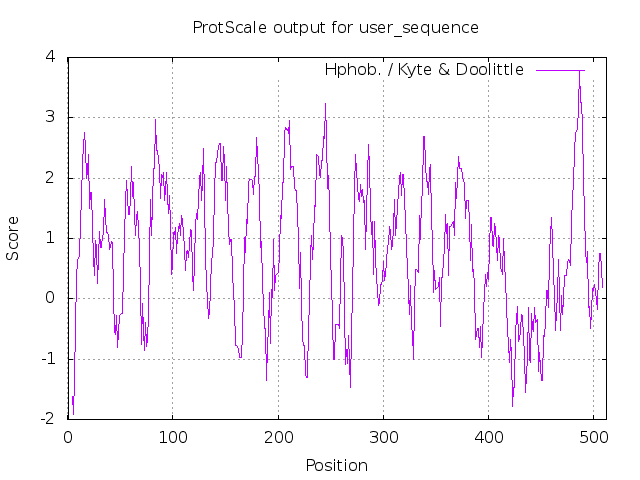
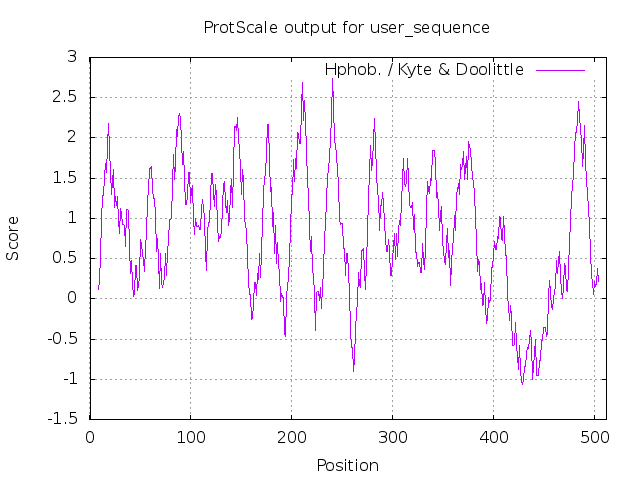
*Figure 10: Kyte-Doolittle hydropathy plots of N. gonorrhoea EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



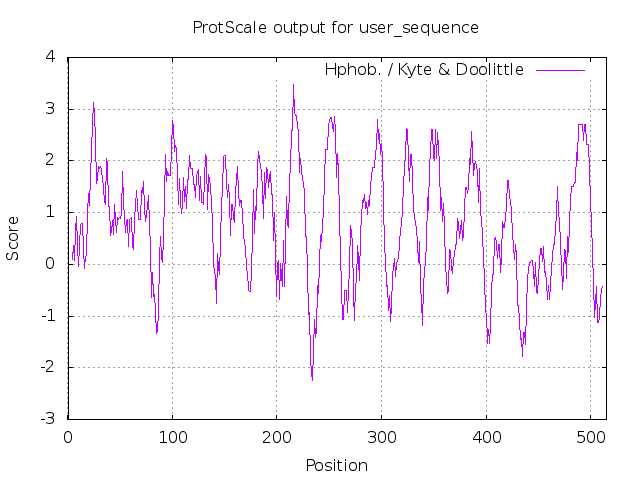
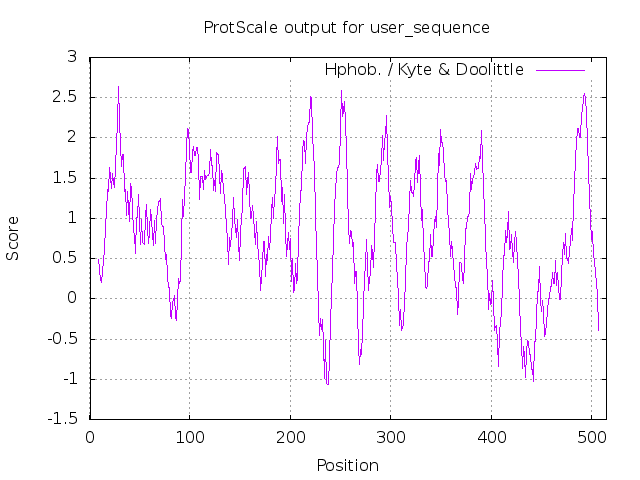
*Figure 11: Kyte-Doolittle hydropathy plots of Polaromonas sp. CF318 EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



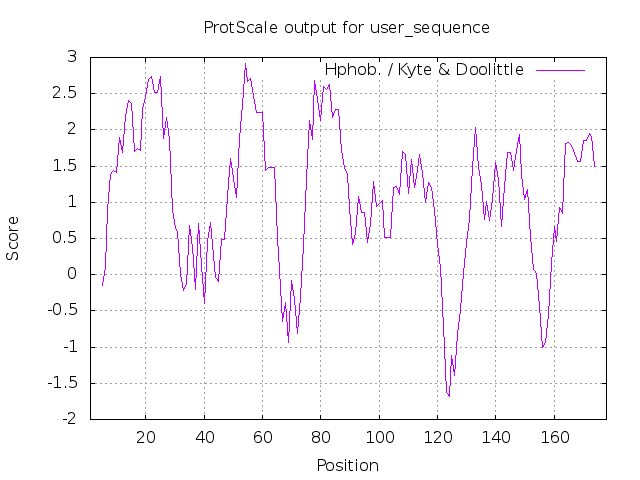
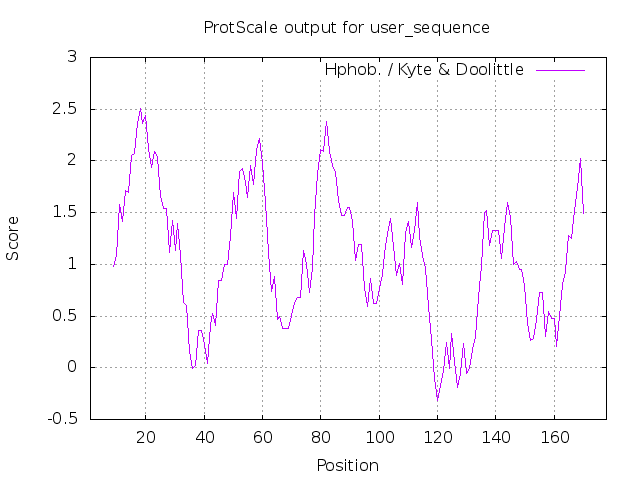
*Figure 12: Kyte-Doolittle hydropathy plots of Burkholderia sp. lig30 EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



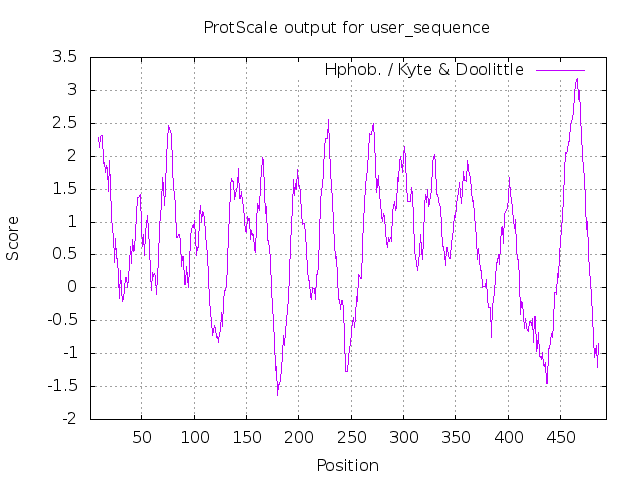
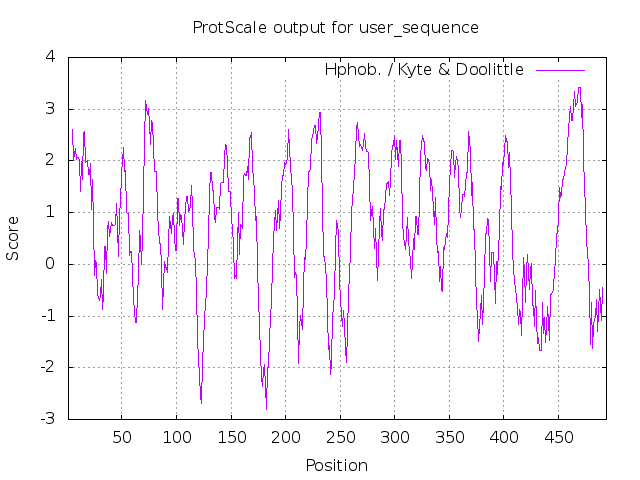
*Figure 13: Kyte-Doolittle hydropathy plots of Vibrio rumoiensis EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



*Figure 14: Kyte-Doolittle hydropathy plots of Sphingobium japonicum EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*

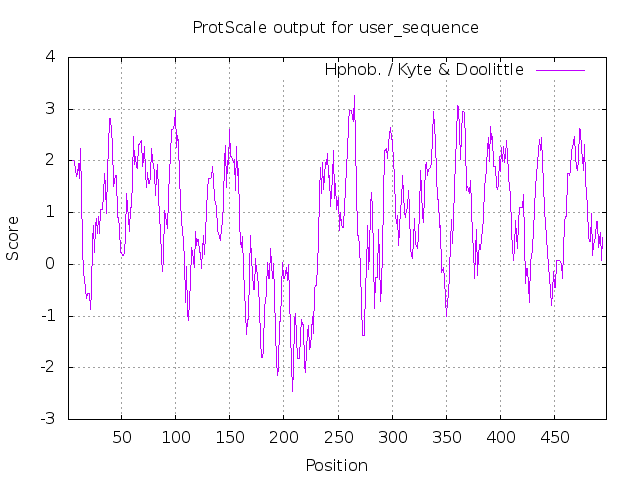
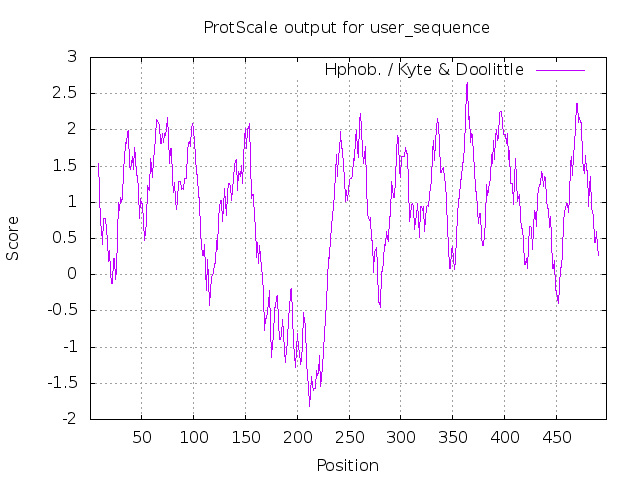


*Figure 15: Kyte-Doolittle hydropathy plots of Streptomyces spongiae EmrB fragment (matched by BLAST) with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



*Figure 16: Kyte-Doolittle hydropathy plots of Streptomyces spongiae complete EmrB (obtained from UniProt) with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*

*Figure 17: Kyte-Doolittle hydropathy plots of Actinoplanes brasiliensis EmrB with window size 17 (left) and 9 (right). With window size 9, all 12-14 transmembrane regions (peak score >1.6) can be seen.*



**Frequency Plots**

Figure 18: Frequency of hits in each cluster of identity for bacterial species with homologues of EmrA (orange), EmrB (grey), and TolC (blue). As the stringency is reduced, the more hits can be seen. EmrA and EmrB have increased number of hits on the order of hundreds below a 50% identity threshold, after which the number of aligned homologues soars. TolC, however, only showed significant increase in hits below 40% identity threshold. It is likely that many bacterial species have TolC homologues that serve, as TolC of E. coli does, multiple efflux pumps, being coded for separately. Expectedly, there are few hits between 90-100% identity for all but EmrB, suggesting unique mutations and changes to EmrA and TolC acquired by other bacterial species, including close relatives of E. coli, in accord with their environmental demands. The highest overall hits were acquired for EmrB, which could be attributed to its central role in the transport of substrates across both gram-positive and gram-negative orthologs and paralogs. The second-most abundant protein proved to be EmrA, which also forms an important part of a given efflux pump, bridging the gap between the transporter and exit channel in gram-negative species, which make up the vast majority of hits in this alignment. EmrA and EmrB are often coded for in one gene cluster and work in tandem to serve a bacterium’s physiological needs, which is likely why their numbers are similar. Meanwhile, TolC is often found in a distant genomic location to any efflux pump, especially in E. coli, and can be incorporated in several efflux pumps as a ubiquitous exit channel, possibly explaining the relatively low overall number of hits for this protein.

The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model ([Jones et al., 1992](#Jones)). The tree with the highest log likelihood (-10069.48) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 16 amino acid sequences. There were a total of 471 positions in the final dataset. Evolutionary analyses were conducted in MEGA-X ([Kumar et al., 2018](#Kumar3)).

The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model ([Jones et al., 1992](#Jones)). The tree with the highest log likelihood (-10435.22) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 16 amino acid sequences. There were a total of 583 positions in the final dataset. Evolutionary analyses were conducted in MEGA X ([Kumar et al., 2018](#Kumar3)).