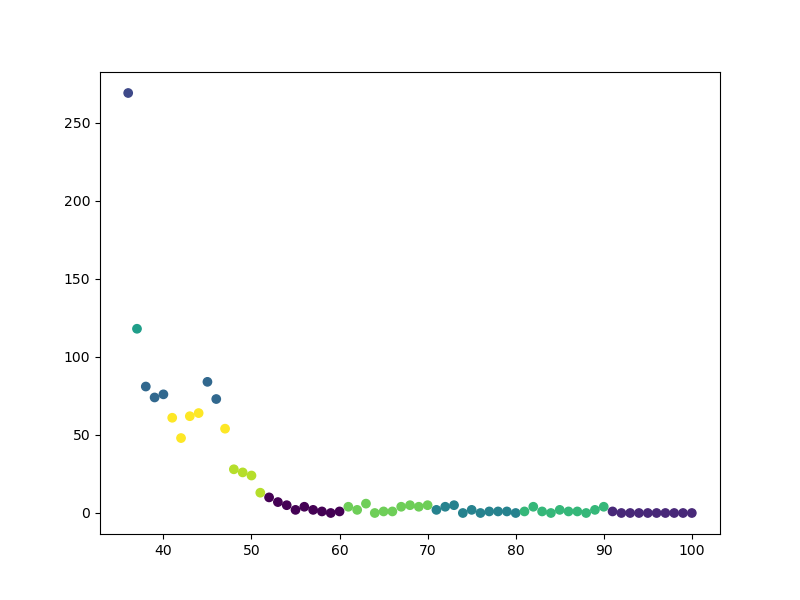
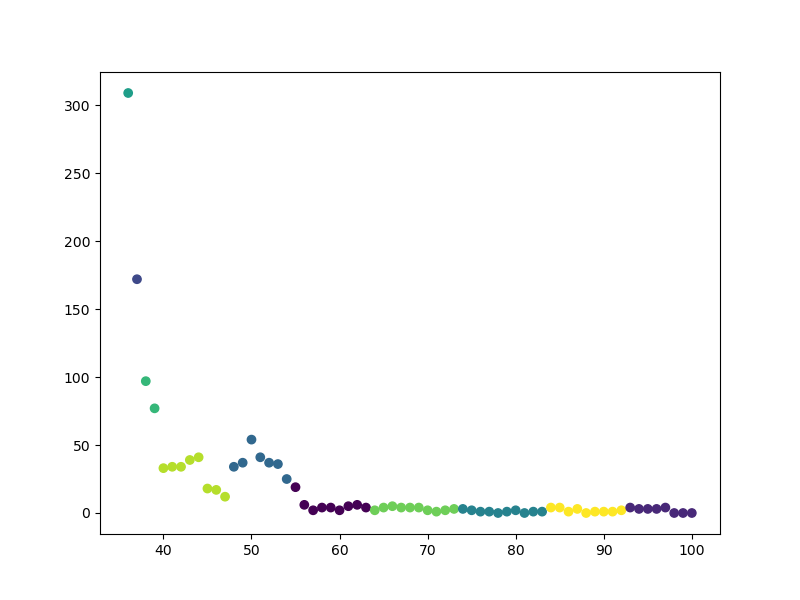
**Supplementary Data**

**Page 1: K-means clustering plots made with python.  
Pages 2-9: Hydropathy analysis made with ExPaSy’s ProtScale tool.  
Pages 10-11: Frequency plots made with Microsoft Excel.**

**K-means clustering**

EmrB

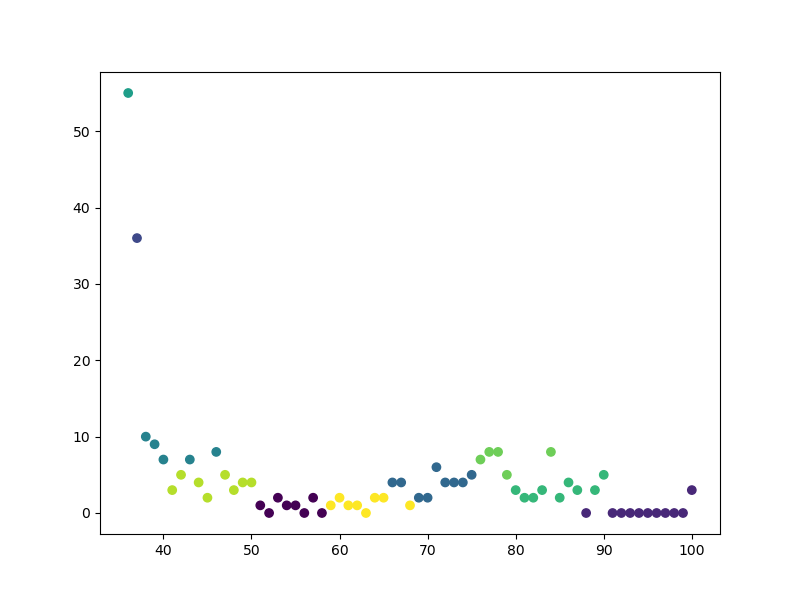
EmrA

Identity (%)

Identity (%)

Frequency of hits

Frequency of hits

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

Identity (%)

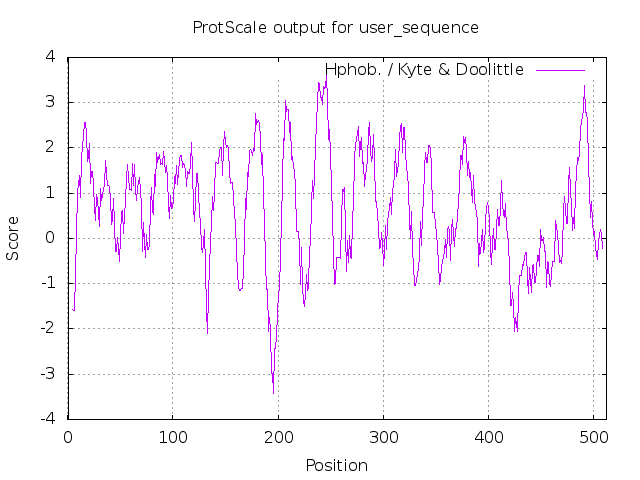
TolC

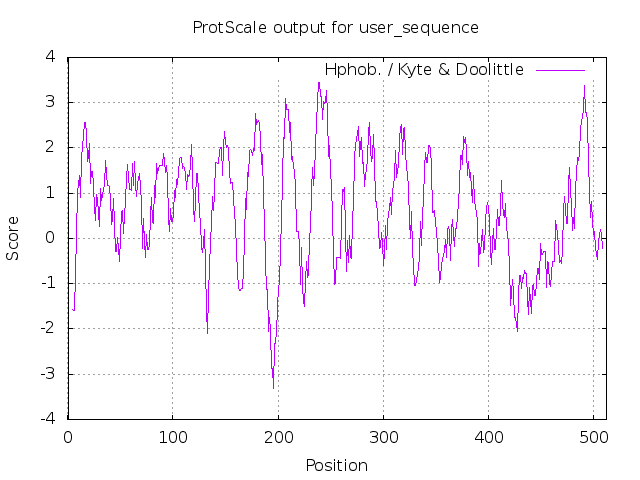
Frequency of hits

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

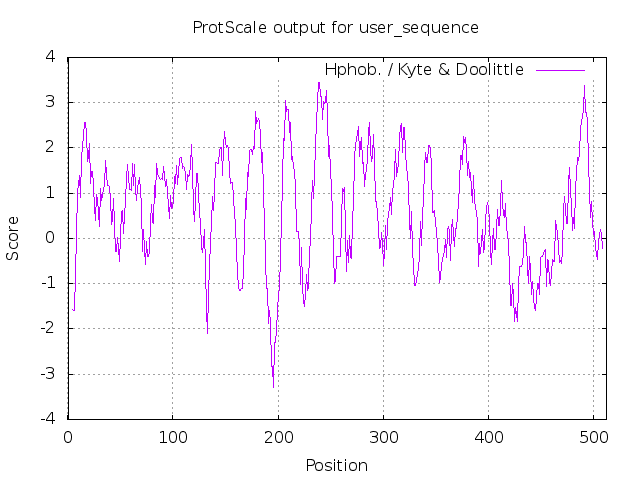
*Table 1: Summary of all hydropathy analyses made for EmrB homologues of representative species and E. coli EmrB, showing the number of transmembrane regions found for each.*

|  |  |
| --- | --- |
| **Bacterial species** | **Transmembrane domains (helices)** |
| *Escherichia coli (strain K12)* | 14 |
| *Citrobacter koseri (strain ATCC BAA-895 CDC 4225-83 SGSC4696)* | 13 |
| *Salmonella typhimurium (strain LT2 SGSC1412 ATCC 700720)* | 13 |
| *Klebsiella pneumoniae subsp. pneumoniae (strain ATCC 700721 MGH 78578)* | 13 |
| *Erwinia sp. Leaf53* | 13 |
| *Yersinia pestis* | 13 |
| *Xenorhabdus bovienii str. kraussei Quebec* | 13-14 |
| *Chromobacterium sp. LK11* | 13 |
| *Providencia heimbachae ATCC 35613* | 12-13 |
| *Neisseria gonorrhoeae* | 13-14 |
| *Polaromonas sp. CF318* | 13-14 |
| *Burkholderia sp. lig30* | 12 |
| *Vibrio rumoiensis 1S-45* | 13 |
| *Sphingobium japonicum BiD32* | 12-13 |
| *Streptomyces spongiae* | 13 (full sequence) |
| *Actinoplanes brasiliensis* | 12 |

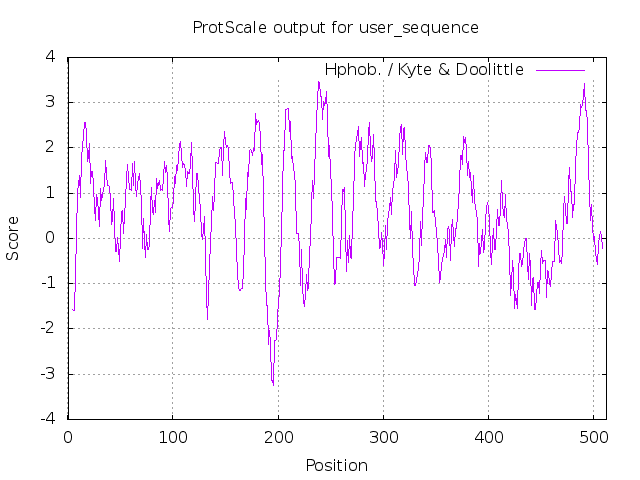
*Figure 2: Kyte-Doolittle hydropathy plot of E. coli EmrB with window size 9. All 14 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*



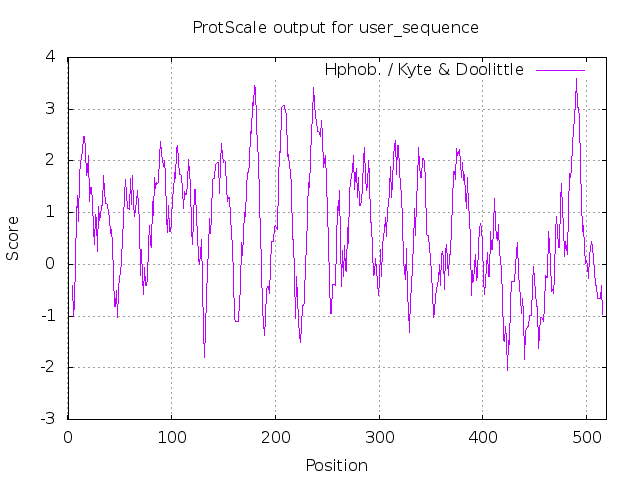
*Figure 3: Kyte-Doolittle hydropathy plot of* *C. koseri EmrB with window size 9. All 13 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*



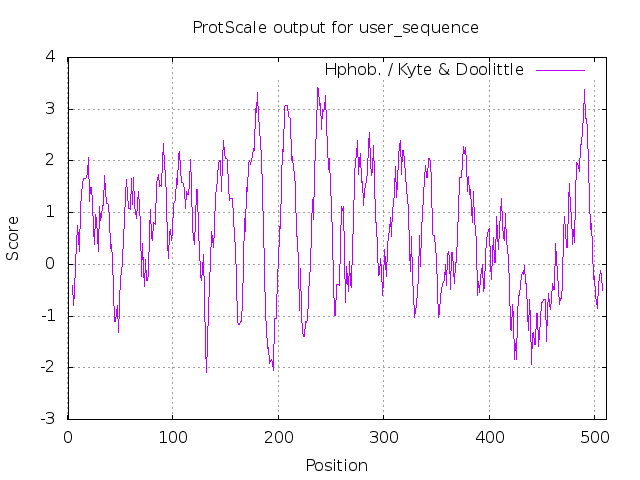
*Figure 4: Kyte-Doolittle hydropathy plot of* *S. typimurium EmrB with window size 9. All 13 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*

****

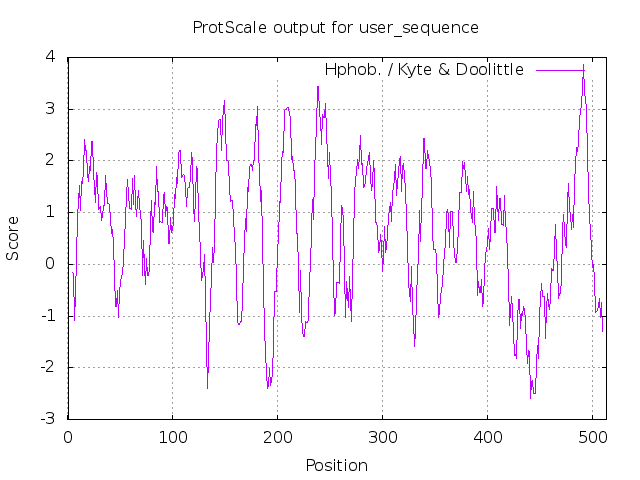
*Figure 5: Kyte-Doolittle hydropathy plot of* *K. pneumoniae EmrB with window size 9. All 13 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*

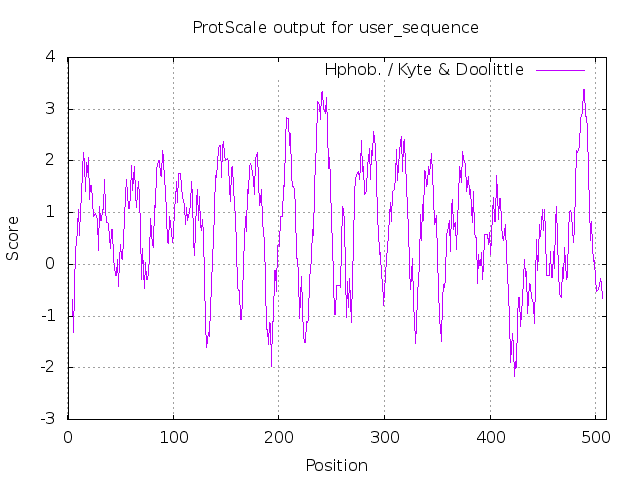


*Figure 6: Kyte-Doolittle hydropathy plot of* *Erwinia sp. Leaf53 EmrB with window size 9. All 13 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*

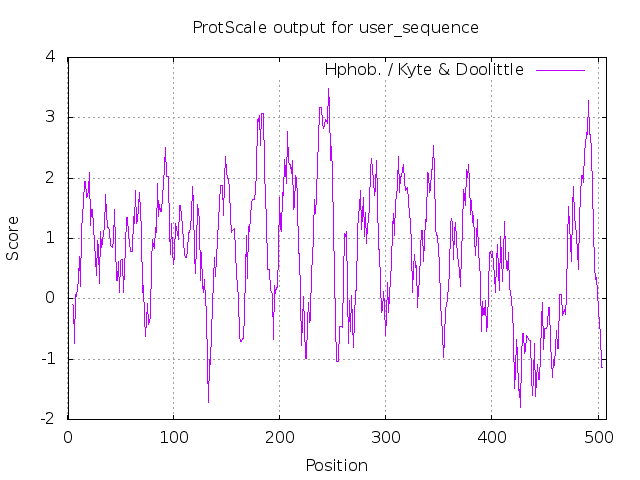


*Figure 7: Kyte-Doolittle hydropathy plot* *of Y. pestis EmrB with window size 9. All 13 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*

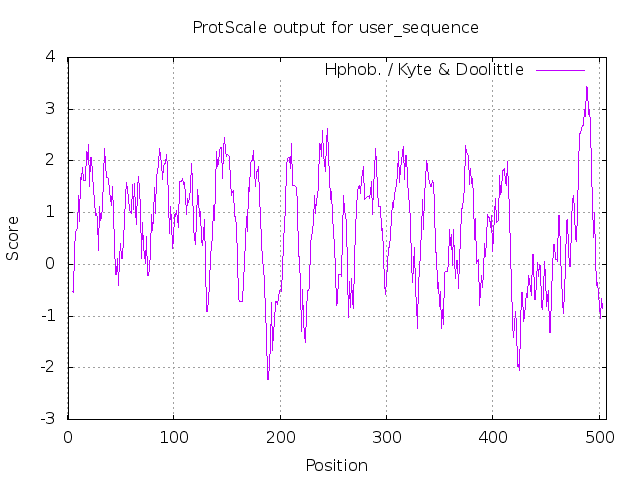
 *Figure 8: Kyte-Doolittle hydropathy plot of* *X. bovienii EmrB with window size 9. All 13-14 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*



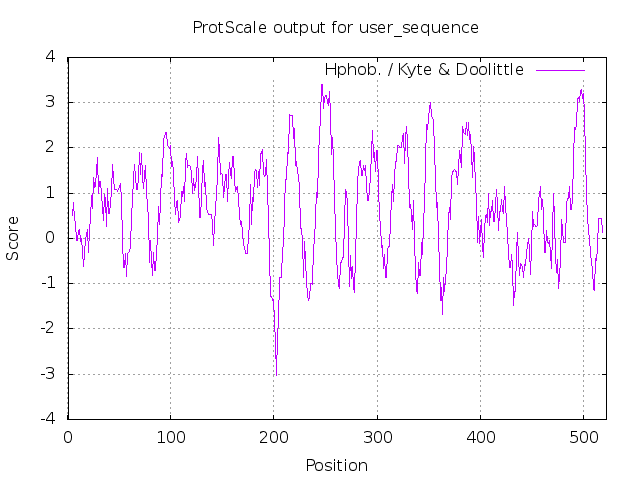
*Figure 9: Kyte-Doolittle hydropathy plot of* *Chromobacterium LK11 EmrB with window size 9. All 12-13 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*



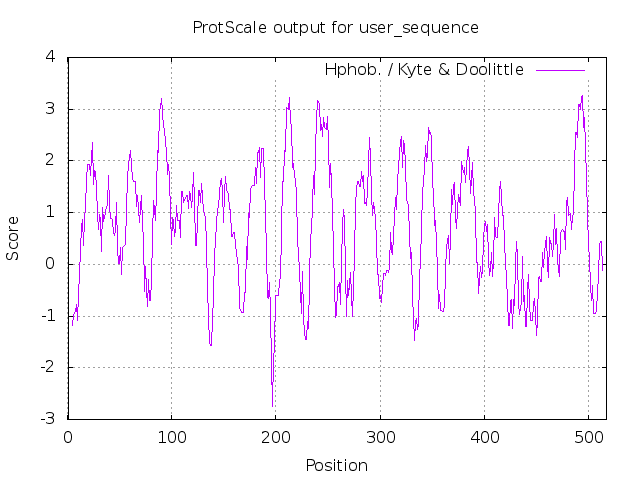
*Figure 10: Kyte-Doolittle hydropathy plot of* *P. heimbachae EmrB with window size 9. All 13 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*



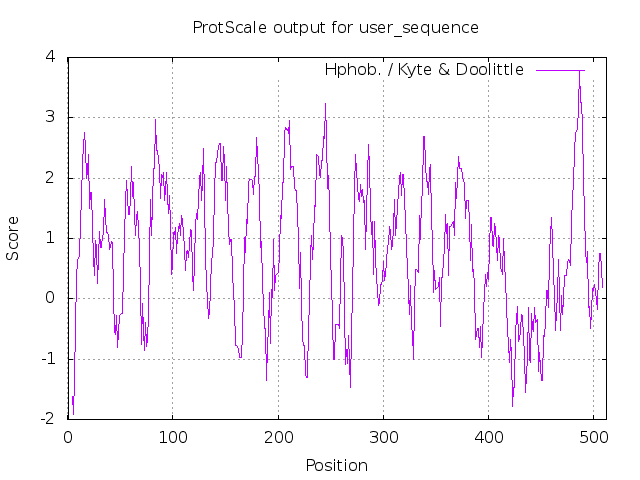
*Figure 11: Kyte-Doolittle hydropathy plot of* *N. gonorrhoea EmrB with window size 9. All 13-14 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*



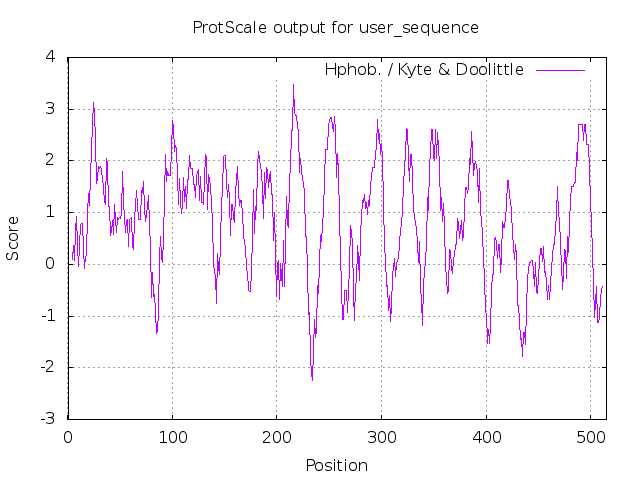
*Figure 12: Kyte-Doolittle hydropathy plot of* *Polaromonas sp. CF318 EmrB with window size 9. All 13-14 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*



*Figure 13: Kyte-Doolittle hydropathy plot of* *Burkholderia sp. lig30 EmrB with window size 9. All 12 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*

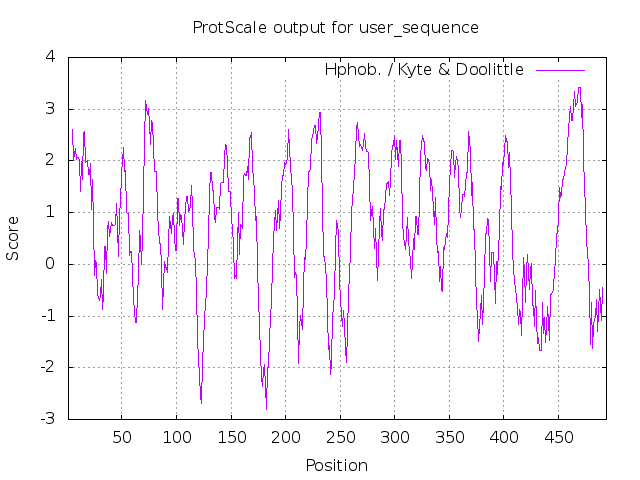
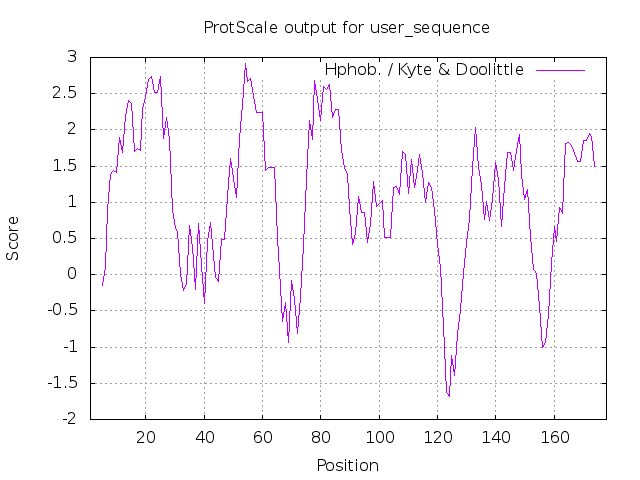


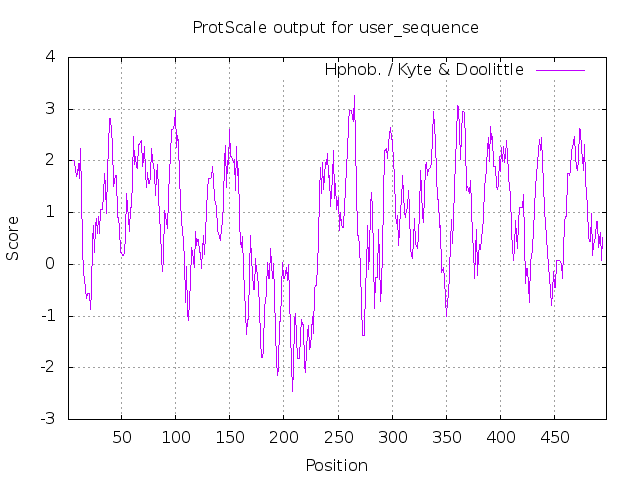
*Figure 14: Kyte-Doolittle hydropathy plot of* *Vibrio rumoiensis EmrB with window size 9. All 13 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*



*Figure 15: Kyte-Doolittle hydropathy plot of* *Sphingobium japonicum EmrB with window size 9. All 12-13 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*

*Figure 16: Kyte-Doolittle hydropathy plots of* *Streptomyces spongiae EmrB fragment (left, matched by BLAST) and full S. spongiae EmrB homologue (right, obtained from UniProt) with window size 9. 4-6 transmembrane regions (peak score >1.6) can be seen for the fragment and all 13 transmembrane regions can be seen for the plot of the full sequence, as compared to sequence analysis at UniProt.*





*Figure 17: Kyte-Doolittle hydropathy plots of* *Actinoplanes brasiliensis EmrB with window size 9. All 12 transmembrane regions (peak score >1.6) can be seen, as compared to sequence analysis at UniProt.*

**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.