References

Bae, T., & Schneewind, O. (2003). The YSIRK-G/S motif of staphylococcal protein A and its role in efficiency of signal peptide processing. Journal of bacteriology, 185(9), 2910–2919. https://doi.org/10.1128/JB.185.9.2910-2919.2003

Lusetti, S. L., & Cox, M. M. (2002). The bacterial RecA protein and the recombinational DNA repair of stalled replication forks. Annual review of biochemistry, 71(1), 71-100.

Appendix

Appendix 1: Python Script

Useful for finding average and longest read in a FASTA file.

```python
# usage
# cat predict/protein.fa.orf | python av_length.py
import sys

def read_stdin():
    lines = []
    for line in sys.stdin:
        line = line.strip()

        if line.startswith(";"):
            continue

        if line.startswith(">") or len(line) > 0:
            lines.append(line)

    return lines

def parse_genes(lines):
    gene = ""
    genes = []
    for line in lines:
        if line.startswith(">"):
            genes.append(gene)
            gene = ""
        else:
            gene += line
    return genes

def av_length(genes):
    total = sum(map(lambda x: len(x), genes))
```

```
    return round(total / len(genes))

def find_longest(genes):
    best = genes[0]
    for gene in genes:
        if len(gene) >= len(best):
            best = gene
    return best

lines = read_stdin()
genes = parse_genes(lines)
mean = av_length(genes)
print(f"Mean length of genes: {mean}")

longest = find_longest(genes)
print(f"Longest gene: {len(longest)}. {longest}")
```