The Journal of Open Source Software

Python Implementation of Codon Adaptation Index

Benjamin D. Lee[1]

1 School of Engineering and Applied Sciences, Harvard University

DOI: 10.21105/joss.00905

Software

- Review
- Repository
- Archive

Submitted: 06 June 2018
Published: 03 October 2018

Summary

Amino acids, the building blocks of proteins, are encoded in DNA by triplets of nucleotides called codons. Notably, the synonymous codons for an amino acid are not used in equal proportions in coding DNA sequences. Rather, they exhibit a bias which varies from organism to organism. The codon adaptation index (CAI) is a measurement of this bias with respect to a set of reference genes (Sharp & Li, 1987). It has been used in the context of heterologous protein expression (Grote et al., 2005), virus attenuation (Eschke, Trimpert, Osterrieder, & Kunec, 2018), and cotranslational protein folding prediction (Rodriguez, Wright, Emrich, & Clark, 2017).

`CAI` is a Python package for the efficient calculation of this metric, along with the associated relative synonymous codon usage (RSCU) and relative adaptiveness metrics. In addition, `CAI` includes a command line interface for the calculation of CAI from FASTA files containing DNA sequences. For example, to find the CAI of the sequence within `sequence.fasta` with respect to the sequences within `reference.fasta`, one need only run:

```
$ CAI -s sequence.fasta -r reference.fasta
0.24948128951724224
```

Similarly, using the Python API:

```
>>> from CAI import CAI
>>> from Bio import SeqIO # to parse FASTA files
>>> reference = [seq.seq for seq in SeqIO.parse("reference.fasta", "fasta")]
>>> sequence = SeqIO.read("sequence.fasta", "fasta")
>>> CAI(sequence, reference=reference)
0.24948128951724224
```

In comparison to other Python implementations of the CAI metric (Cock et al., 2009), `CAI` features a CLI, supports multiple genetic codes, can yield the RSCU of reference sets, and correctly handles the case of missing codons in the reference set. Moreover, on a benchmark to determine the CAI of 100 genes consisting of 3,000 random base pairs each with respect to highly expressed genes in *Escherichia coli*, `CAI` performed 39.6% faster than Biopython's implementation.

In conclusion, `CAI` is a significantly faster and more versatile method to determine the CAI, RSCU, and relative adaptiveness of DNA sequences.

Acknowledgements

The author would like to thank Paul Gamble, Karl Ni, and Todd Stavish for their feedback on the manuscript and documentation of this project.

References

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., et al. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. doi:10.1093/bioinformatics/btp163

Eschke, K., Trimpert, J., Osterrieder, N., & Kunec, D. (2018). Attenuation of a very virulent mareks disease herpesvirus (MDV) by codon pair bias deoptimization. *PLOS Pathogens*, *14*(1), e1006857. doi:10.1371/journal.ppat.1006857

Grote, A., Hiller, K., Scheer, M., Munch, R., Nortemann, B., Hempel, D. C., & Jahn, D. (2005). JCat: A novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Research*, *33*(Web Server), W526–W531. doi:10.1093/nar/gki376

Rodriguez, A., Wright, G., Emrich, S., & Clark, P. L. (2017). %MinMax: A versatile tool for calculating and comparing synonymous codon usage and its impact on protein folding. *Protein Science*, *27*(1), 356–362. doi:10.1002/pro.3336

Sharp, P. M., & Li, W.-H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295. doi:10.1093/nar/15.3.1281