Logo_UHH_rgb.tif

Protocol

**Comparison of Biotite and** **Seq\_compat for calculating Sequence-Entropy**

Name: Miriam Busch

Matriculation number: 6926928

Degree course: M.Sc. Molecular Life Sciences

Table of Contents

[1. Introduction 1](#_Toc60224172)

[2. Methods 1](#_Toc60224173)

[3. Results 1](#_Toc60224174)

[4. Discussion 1](#_Toc60224175)

[5. References 1](#_Toc60224176)

# 1. Introduction

The aim of this labrotation was to compare the entropy package of *Seq\_compat* (Torda, 2020), which calculates the per-site entropy in a multiple sequence alignment, to the *Python* package called *Biotite* (Kunzmann & Hamacher, 2018). *Biotite* is a large general package written in Python that can be used to handle a major part of the typical workflow for sequence and biomolecular structure data, whereas the purpose-built entropy package of *Seq\_compat* is very small and written in Go. In order to compare these two programs, the *Biotite* source code had to be adapted to be left with the same functions as the entropy package.   
The calculation of per-site entropy for a multiple sequence alignment allows an assessment of sequence conservation in nucleotide sequences as well as protein sequences (the latter were used for comparison in this labrotation).   
The entropy for each site is calculated as follows:

is the number of different residues. At each site the summation runs over all amino acid types and gives the fraction of each residue that occurs (Schneider, Stormo & Gold, 1986). Hence the entropy range lies between for (or ) and the maximal value that results from all residues being equally likely. This is the case if equals (Sen, Dey & Chowdhury et al., 2019). Low entropy scores imply high sequence conservation.

High sequence conservation in certain protein regions may reflect on important functions and structural roles because mutations in those areas could lead to the protein being non-functional. Identifying conserved sequences can therefore be used for functional annotation and has applications in drug design. Furthermore, sequence conservation is used to generate phylogenetic trees (Torda, 2020).

For graphical representation of sequence composition and conservation sequence logos are frequently used (see figure 1). At each position all symbols that occur are displayed stacked on each other. The relative height of the symbols depicts their relative frequency. The height of each stack indicates its conservation

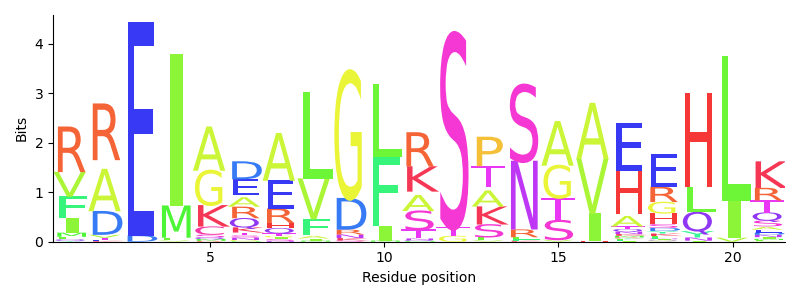


Figure 1

It is the maximum possible Shannon entropy of the alphabet subtracted by the positional entropy.

* Sequence conservation, relationship with sequence logos.  
  The entropy at each site   
  where the summation runs over all amino acid types (Schneider, Stormo, & Gold, 1986).  
  Note that it is traditional to use a small-number correction in sequence logos, but the idea of entropy is similar.  
  The methods differ in that biotite uses to gives bits of information, whereas our code sets to the expected number of amino acid types so as to scale values between 0 and 1.

# 2. Methods

# 3. Results

# 4. Discussion

# 5. References

Kunzmann, P., & Hamacher, K. (2018). Biotite: a unifying open source computational biology framework in Python. BMC Bioinformatics, 19, 346.

Schneider, T. D., Stormo, G. D., & Gold, L. (1986). Information content of binding sites on nucleotide sequences. J. Mol. Biol., 188, 415-431.

Torda, A. (2020). seq\_compat. doi:10.5281/zenodo.433

Sen, S., Dey, A., Chowdhury, S. et al. (2019). Understanding the evolutionary trend of intrinsically structural disorders in cancer relevant proteins as probed by Shannon entropy scoring and structure network analysis. BMC Bioinformatics 19, 549.

Torda, A. (2020). Genome-wide association studies [lecture notes]. University of Hamburg. https://lernen.min.uni-hamburg.de/pluginfile.php/84808/mod\_resource/content/1/4\_gwas\_1-52.pdf