Exploring Homology Using the Concept of Three-State Entropy Vector

Armando J Pinho[1], Sara P Garcia[1], Paulo JSG Ferreira[1], Vera Afreixo[2], Carlos AC Bastos[1], António JR Neves[1] and João MOS Rodrigues[1]

[1]Signal Processing Lab / IEETA
Dep. of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal

[2]Signal Processing Lab / IEETA
[2]Dep. of Mathematics, University of Aveiro, Portugal

PRIB 2010

Outline

Introduction

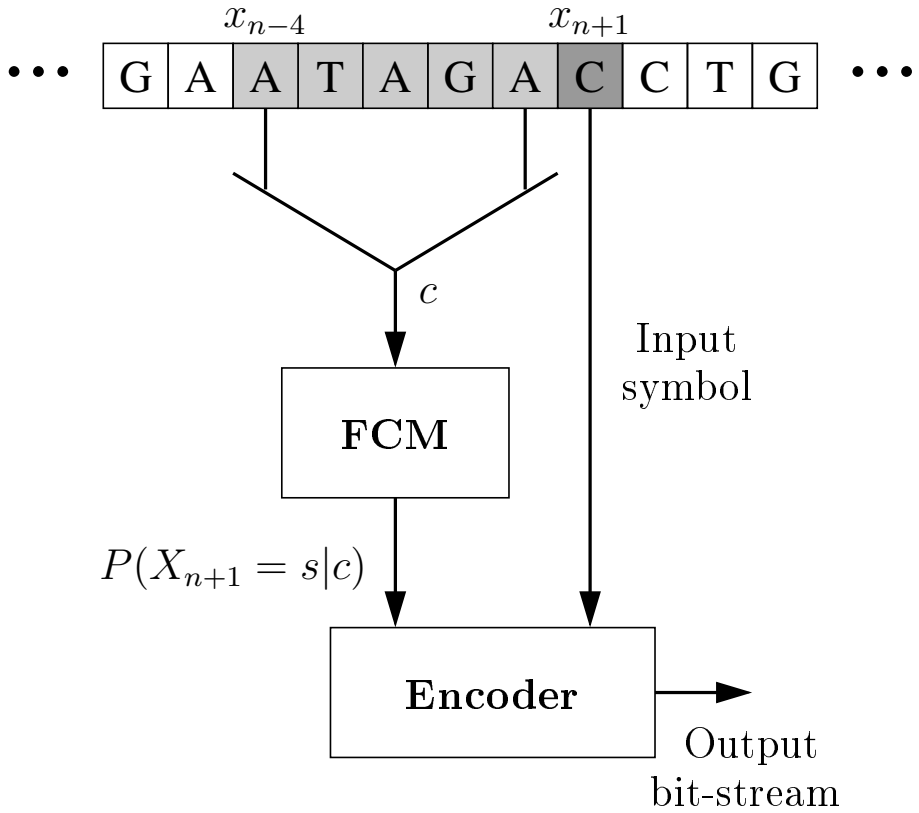
- It is well-known that there are **periodicities** in the DNA sequences.
- The strongest is generally associated with the **period three** that can be found in the exons of prokaryotes and eukaryotes.
- In fact, this period three periodicity has been used, for example, for predicting potential protein-coding regions.
- One of the consequences of this periodicity is that the **entropy** of each of the three bases of the codon varies.

Introduction

- Even more interesting, this entropy pattern seems to vary "continuously" **from species to species**.
- In this work, we further investigated this variation in the entropy of each of the three base positions of the codon.
- For that, we used three-state finite-context (Markov) models of the protein-coding regions.
- Although still preliminary, the results obtained suggest that we are able to construct low-dimensional **entropy vectors** that are useful for **clustering species**.

Finite-context models

A finite-context model is a computational model that assigns probability estimates to symbols s of an alphabet $\mathcal{A}$, according to a **conditioning context** computed over the past k outcomes (order-k finite-context model).



Finite-context models

The bitrate average of the finite-context model after encoding n symbols is given by

$$H_n = -\frac{1}{n}\sum_{i=1}^{n} \log_2 P(X_i = x_i|c) \quad \text{bpb},$$

where the probabilities are obtained using the estimator

$$P(X_{n+1} = s|c) = \frac{n_s^c + \alpha}{\sum_{a\in\mathcal{A}} n_a^c + \alpha|\mathcal{A}|},$$

with n_s^c representing the number of times that, in the past, the source generated symbol s having c as the conditioning context.

Finite-context models

How they are implemented: an example

Each row of the table represents a probability model at a given instant. This example illustrates an order-5 context.

Context, c	n_{A}^c	n_{C}^c	n_{G}^c	n_{T}^c	$\sum_{a \in \mathcal{A}} n_a^c$
AAAAA	23	41	3	12	79
⋮	⋮	⋮	⋮	⋮	⋮
ATAGA	16	6	21	15	58
⋮	⋮	⋮	⋮	⋮	⋮
GTCTA	19	30	10	4	63
⋮	⋮	⋮	⋮	⋮	⋮
TTTTT	8	2	18	11	39

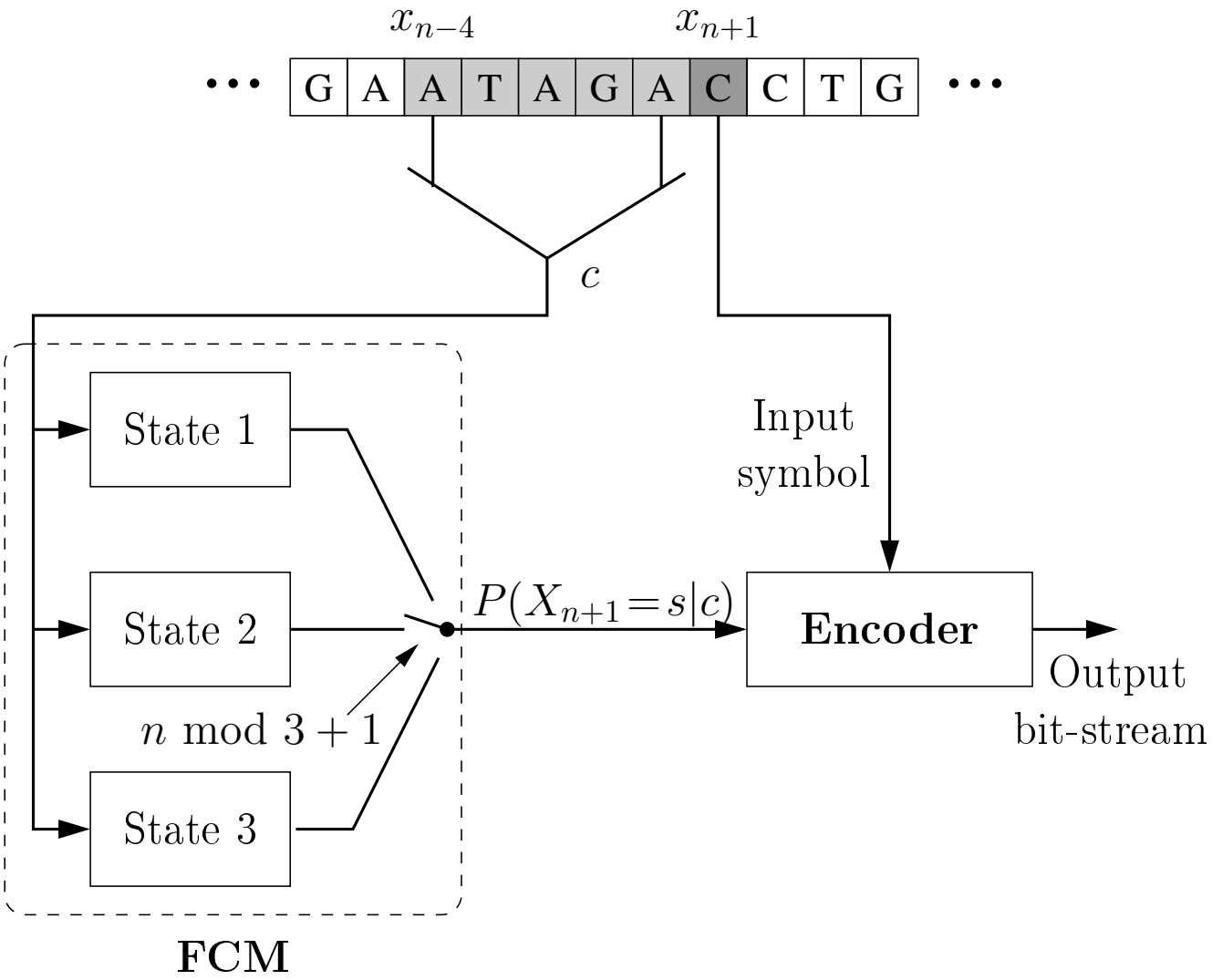
Finite-context models

How they are implemented: an example

Therefore, if the last symbols (context symbols) were "ATAGA", i.e., c =ATAGA, and if the next symbol to encode were, for example, a "C", then the counters would become

Context, c	n_{A}^c	n_{C}^c	n_{G}^c	n_{T}^c	$\sum_{a\in\mathcal{A}} n_a^c$
AAAAA	23	41	3	12	79
⋮	⋮	⋮	⋮	⋮	⋮
ATAGA	16	**7**	21	15	**59**
⋮	⋮	⋮	⋮	⋮	⋮
GTCTA	19	30	10	4	63
⋮	⋮	⋮	⋮	⋮	⋮
TTTTT	8	2	18	11	39

The three-state finite-context model

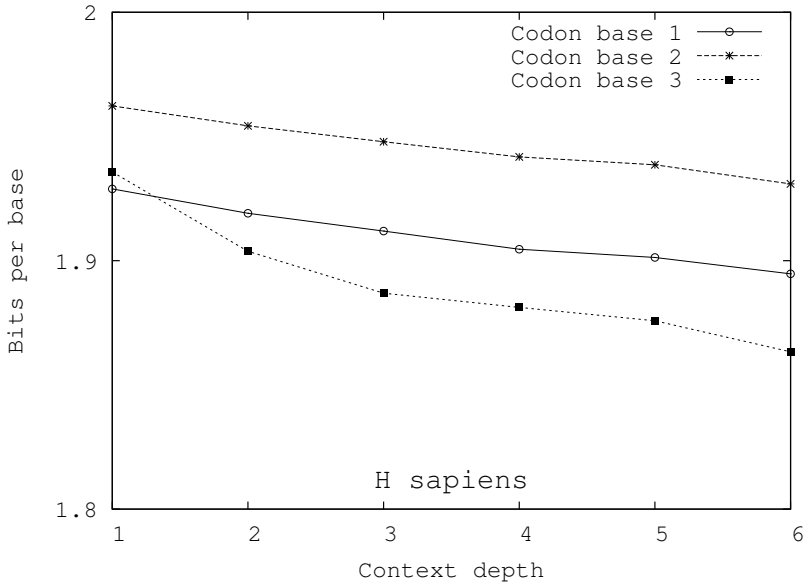


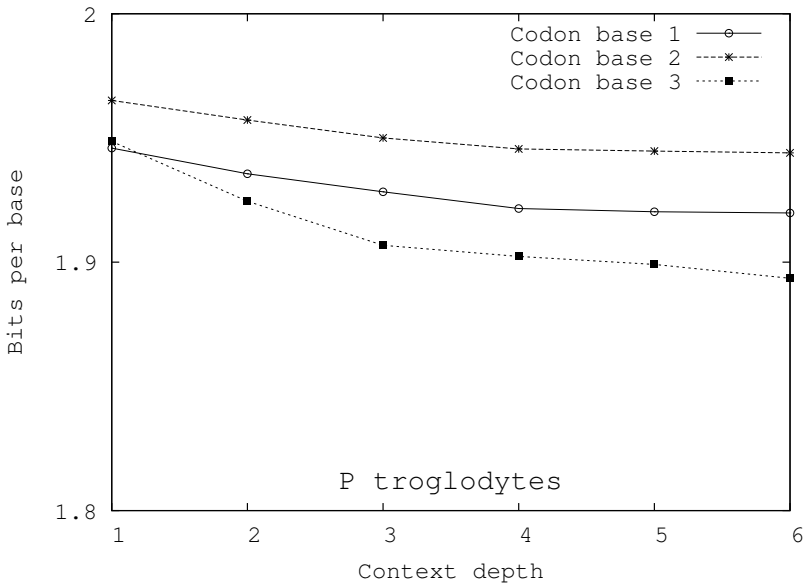
Experimental results

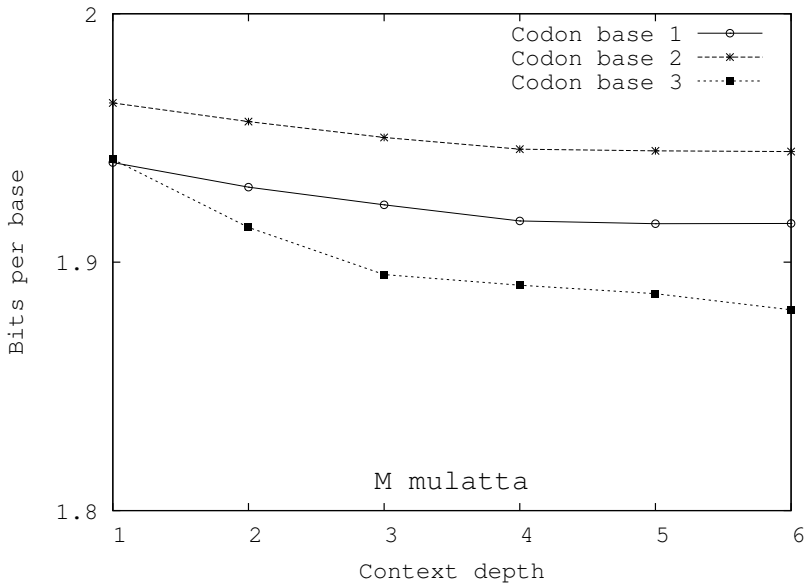
Organisms

Organism	Reference
Homo sapiens (human)	Build 37.1
Pan troglodytes (chimpanzee)	Build 2.1
Macaca mulatta (rhesus macaque)	Build 1.1
Mus musculus (mouse)	Build 37.1
Rattus norvegicus (brown rat)	Build 4.1
Arabidopsis thaliana (thale cress)	NC003070/1/4/5/6
Populus trichocarpa (black cottonwood)	Version 2.0
Vitis vinifera (grape vine)	Build 1.1
Ricinus communis (castor oil plant)	Release 0.1
Streptococcus pneumoniae strain ATCC 700669	NC011900
Chlamydia trachomatis strain D/UW-3/CX	NC000117
Mycoplasma genitalium strain G-37	NC000908
Streptococcus mutans strain UA159	NC004350

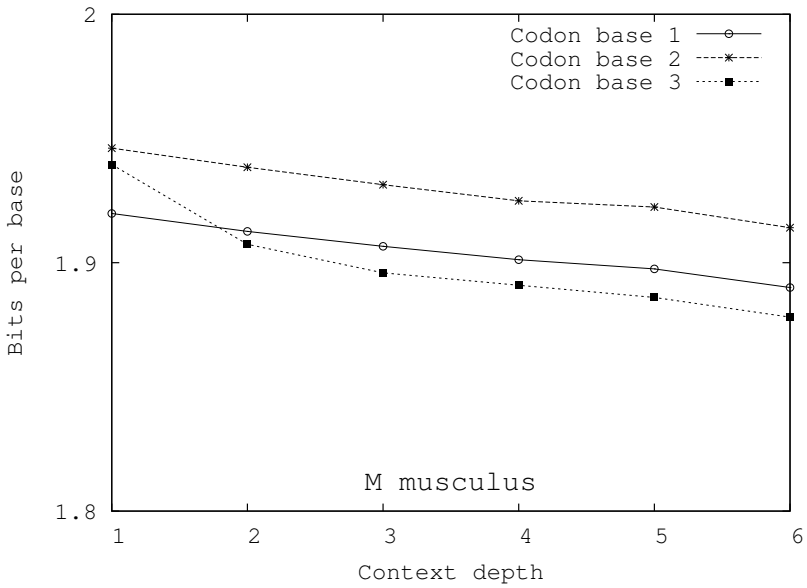
Experimental results

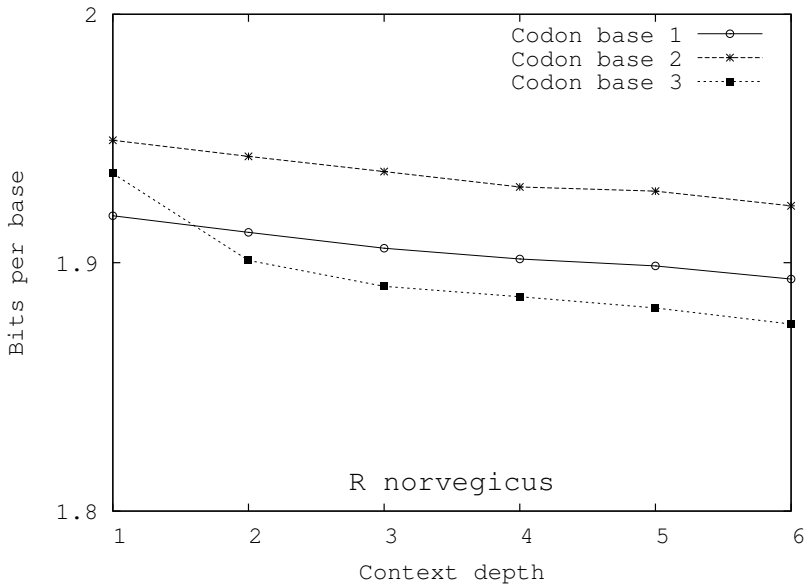




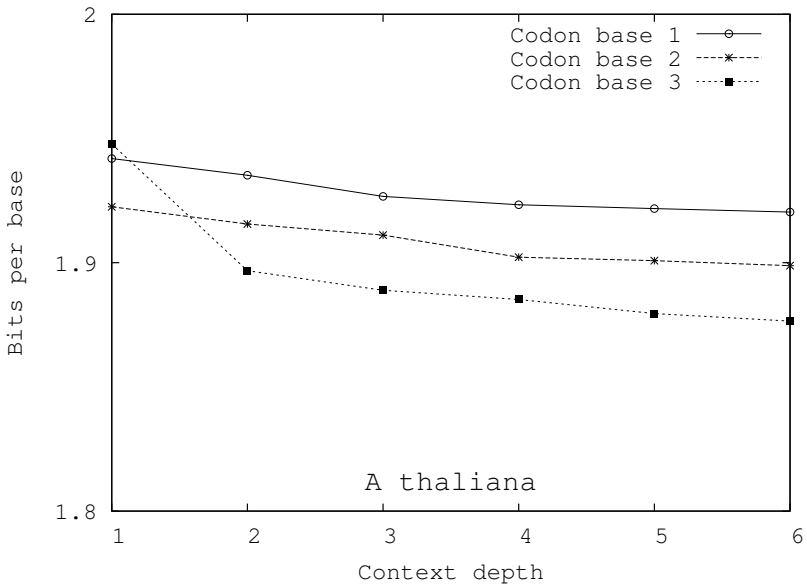


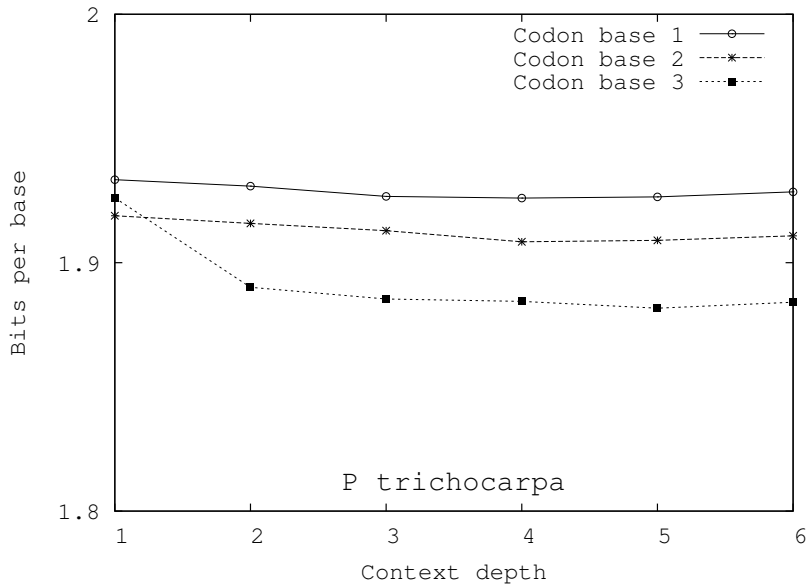
Experimental results

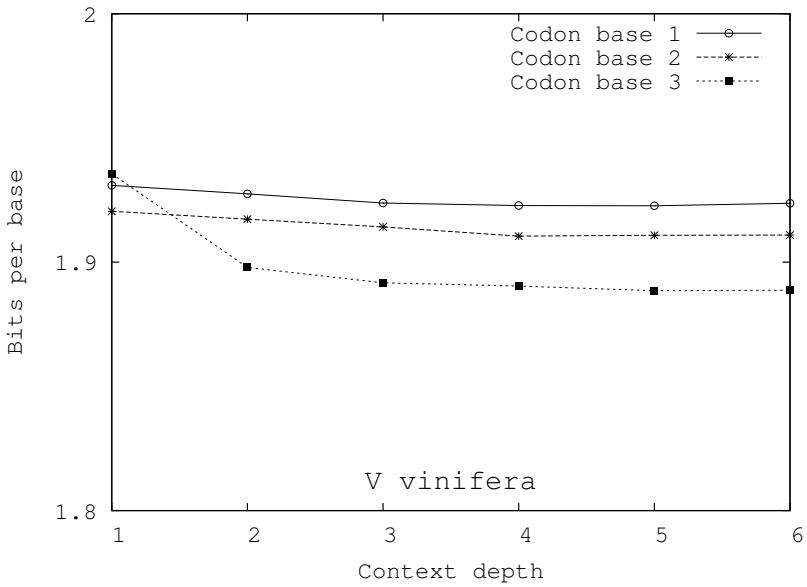


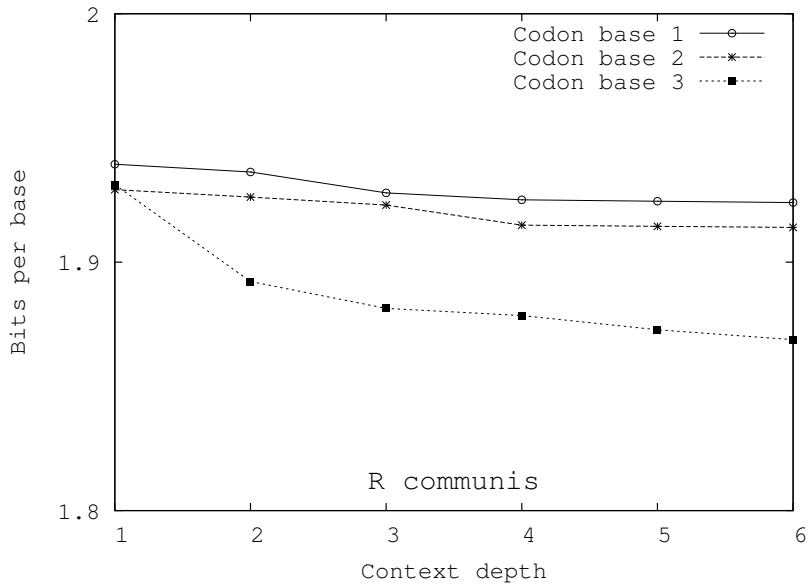


Experimental results

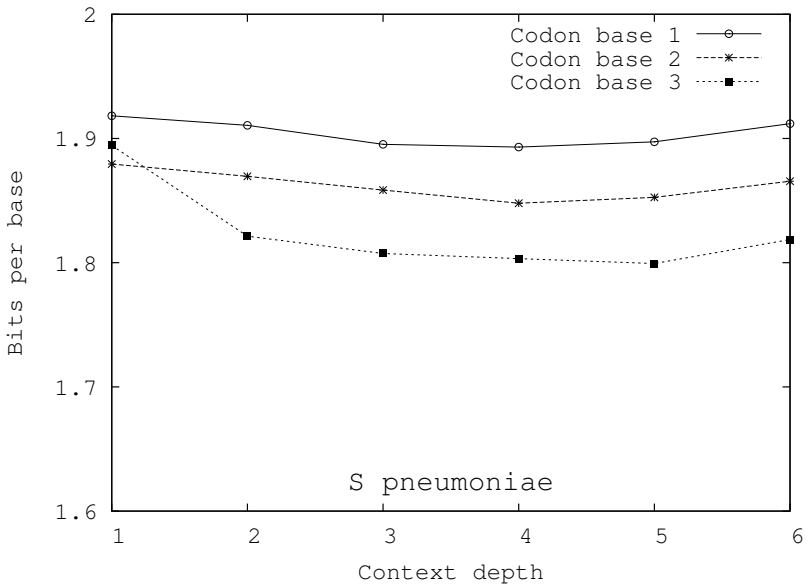


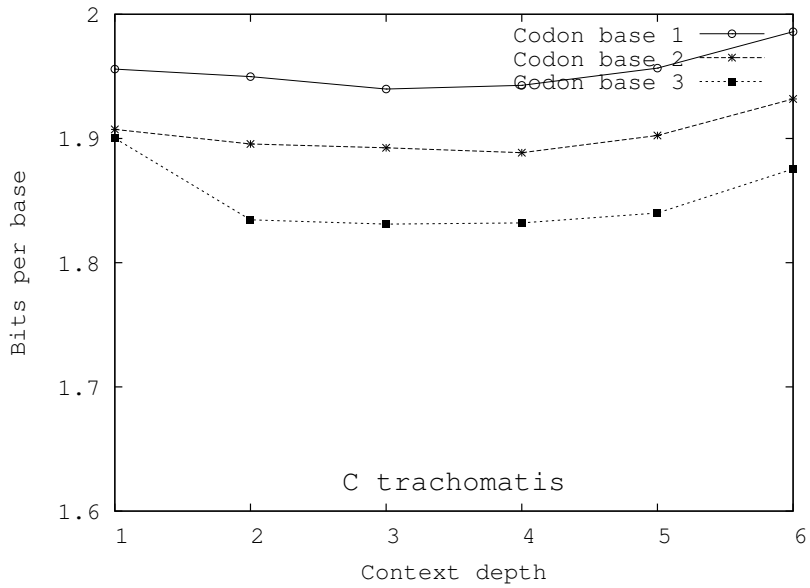


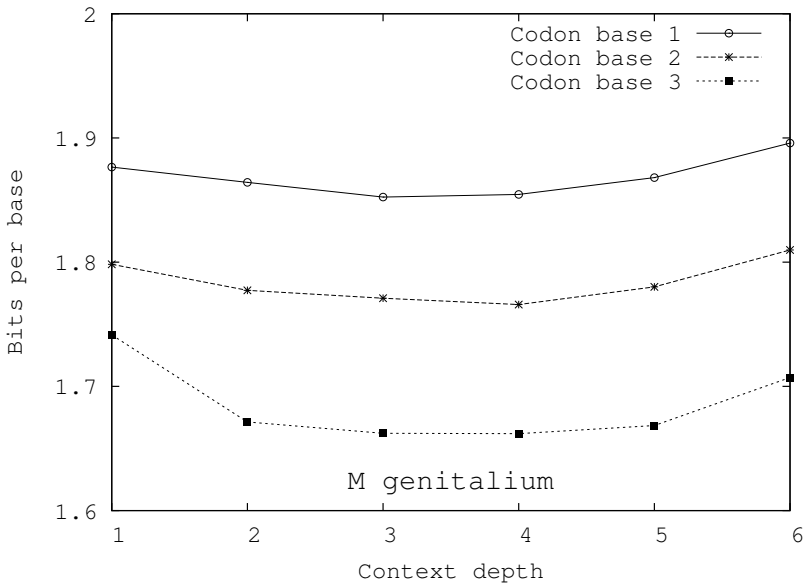


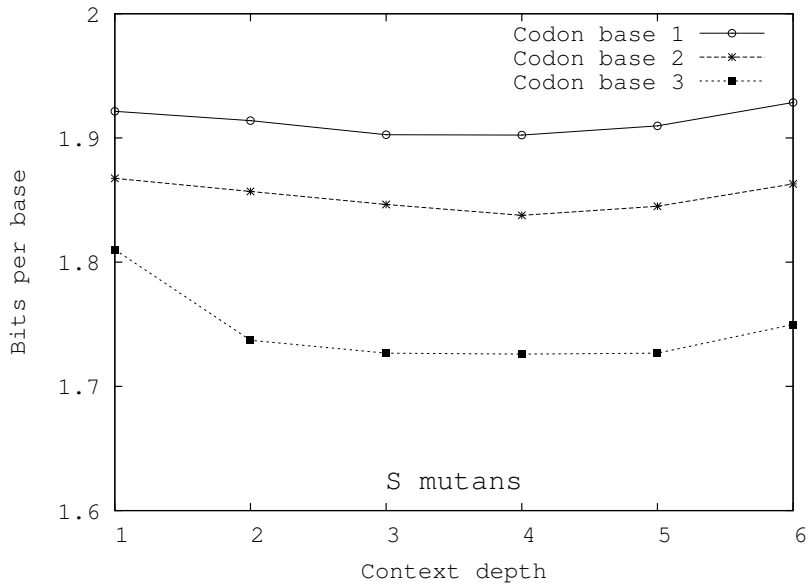


Experimental results



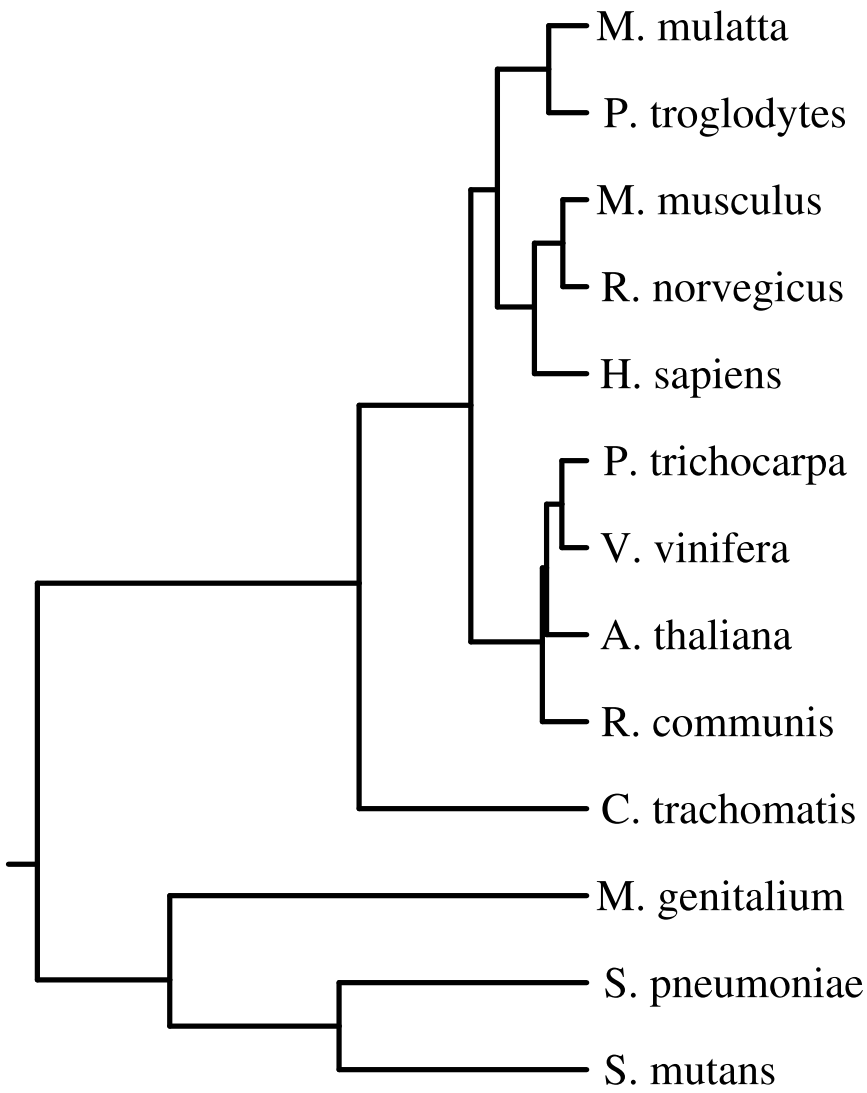






Experimental results

Dendogram



Conclusion

- We built **three-state entropy vectors** for several organisms.
- Based on these vectors, we tried to cluster species.
- The preliminary results suggest that the information gathered from these three-state entropy vectors seems to be useful for building meaningful dendograms.
- Further study is needed, including adding more species and having a better control over the quality of the data.
- Large differences in data size is a problem that needs to be solved.