pORFect: ***a not so perfect predictor, contrary to what name suggests***

Hesselman, Maria Carmen; Ropat, Maryia

Introduction

The analysis of sequenced genomes is imperative for translating nucleotide sequences into knowledge that can be used in biological research. With the increase in the number of available genome sequences, it is necessary to develop tools that can make sense of genomic sequences and analyse different characteristics. One of the most important analyses that must be done on sequenced genomes is the prediction of genes. Finding genes in a genome provides knowledge that is useful for understanding more about the organism and its evolution as genes can be annotated with functions and compared to homologous genes to find evolutionary relationships. Apart from finding genes in a genome, there are other analyses the can improve our understanding of an organism's genome.

DNA is made up of four bases: adenine, thymine, cytosine and guanine. However, the ratio between AT and CG content is not always 1:1. In fact, this ratio varies from species to species (1). Therefore, calculating the GC content of a genome is useful for comparing genomes and it can even be used as a measure of evolutionary distance between genomes (1).

For this project, we received five genomes to analyse with different strategies. We calculated the GC content, along with the dinucleotide content. The distinction between GC content and dinucleotide content is that GC content is the total number of C's and G's in the genome (regardless of their position), divided by the total number of nucleotides in the genome. On the other hand, the dinucleotide frequencies indicate how many times two given nucleotides appear next to each other in the genome. Additionally, we created a gene predictor to find genes in the five genomes. The results of our predictions were compared to the reference proteomes obtained from UniProt. Finally, we present a phylogenetic tree based on the distances between our genomes in terms of GC-content.

Materials and Methods

We received four prokaryotic genomes and one eukaryotic genome to work with for this project. The genomes are summarized in table 1.

Genome	**Organism**	**Size of genome**	**Domain**
04.fa.txt	B. japonicum,	9,224,208	Prokaryote
05.fa.txt	Ch. trachomatis	1,042,588	Prokaryote
08.fa.txt	D. turgidum	1,855,560	Prokaryote
16.fa.txt	Rh. baltica	7,145,576	Prokaryote
34.fa.txt	S.cerevisiae XIV	784,333	Eukaryote

Table 1. Information about the used genomes.

The first step in the genome analysis was to calculate the nucleotide composition, GC-content (eq. [1]) and dinucleotide composition (eq. [2]) of the whole genomes.

$$GC\ content\ = ((G + C) \div (A + T + C + G)) \times 100 \qquad [1]$$

$$Di - nucleotide\, frequency\ =\ XX \div (total\, X - 1) \qquad [2]$$

The GC contents of the five genomes were used as distance measures between the genomes (eq.[3]). The distance was calculated and used to build a phylogenetic tree using Belvu with neighbor-joining because it does not assume constant evolutionary rates (5).

$$D\ =\ \sqrt{(GC\ genome\ 1\ -\ GC\ genome\ 2)^{2}} \qquad [3]$$

Additionally, we wrote an open reading frame (ORF) prediction tool in Python. Our ORF predictor reads in a genome file and predicts ORFs on the basis of simple assumptions about gene structure. It searches for stretches of DNA starting with a start codon (ATG) and ending with a stop codon (TAG, TAA or TGA) in all six reading frames (3). We further filtered the ORFs based on Shannon entropy (eq.[4]) of the prediction (4). We calculated the Shannon entropy of our predicted ORFs and compared them to the entropy of the entire genome.

$$H\ \ =\ \sum P(Dinucleotide\ in\ sequence) \times\ log\ _{2}P(Dinucleotide\ in\ genome)\ [4]$$

Where P is the probability of finding a given dinucleotide.

We used an entropy cutoff to filter out sequences. The entropy is higher for sequences in which unexpected combinations of dinucleotides that are in the genome are present. When departing from assumption that coding sequences are not completely random, we wanted to see if information entropy could help predict and refine our results.

We used the entropy metric to select the best potential transcript out of overlapping ones (within same reading frame) of different lengths. If the entropy would be higher in the shorter overlapping transcript, we would elect to truncate it to that size. This approach also warrants that we analyze nucleotide sequences of at least >225 nt as it yielded better results, both due to minimum protein length and in order for the large enough sample size to properly represent the entropy.

After predicting ORFs, we translated them to protein sequence and calculated the amino acid frequencies for all five predicted proteomes (eq. [5]). We also calculated the diamino acid frequencies in all of our predicted proteomes (eq. [6]).

$$aa\, frequency\ =\ (amino\ acid/(total\ number\ of\ amino\ acids)) \times 100 \qquad [5]$$

$$di - aa\, frequency\ =\ (di\ amino\ acid/\ (length\ of\ sequence\ - 1) \qquad [6]$$

Results

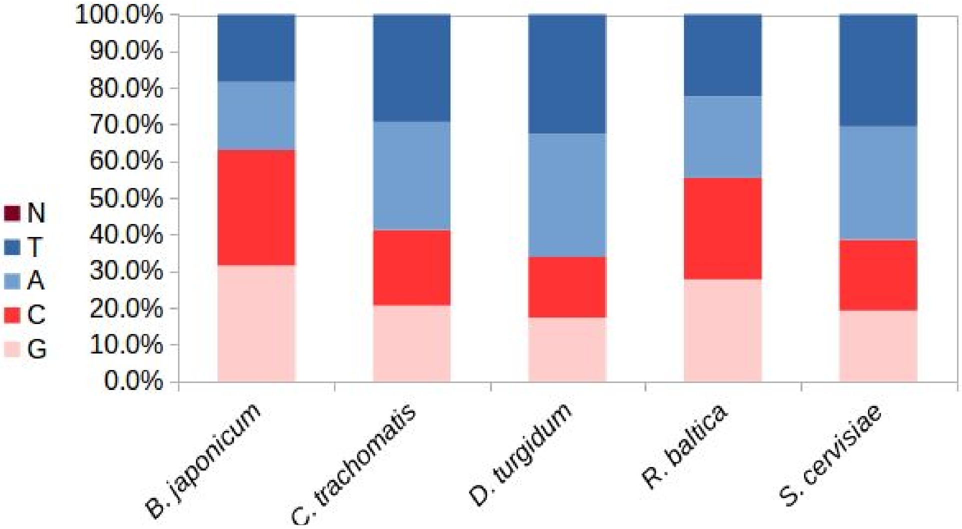


Figure 2. Nucleotide composition of the studied genomes. The GC content is shown in red and pink.

The analysis of the GC content of the five genomes showed that *B. japonicum* has the highest percentage of GC's out of our genomes (Figure 1). *B. japonicum* and *R. baltica* have a higher percentage of C than the rest of the genomes and *C. trachomatis* and *D. turgidum* have a higher percentage of C (Figure 1). The frequencies of dinucleotides for all genomes are shown in figure 2.

A and CG are overrepresented in *B. japonicum.* The genomes of C. trachomatis and S. cerevisiae show a more even distribution of dinucleotide frequencies (figure 2).

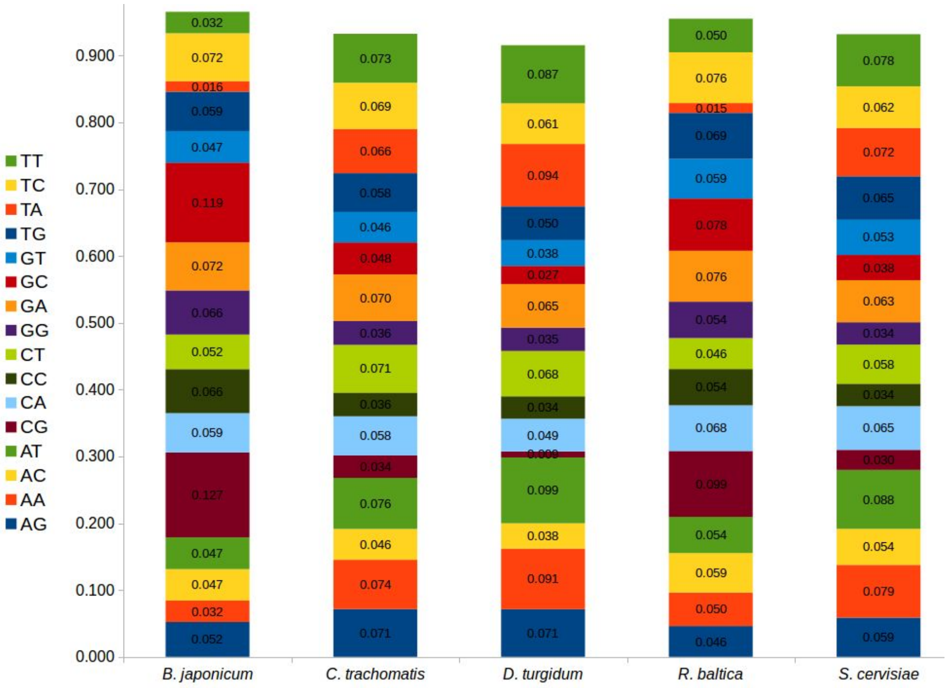


Figure 3. Dinucleotide frequency of the studied genomes.

ORF predictor

The performance of our ORF predictor is summarized in table 2. While the true positive rate is quite low in all genomes, the true negative rates are higher, meaning that the filter removes false ORFs efficiently. Our predictor generally performs better on genomes with lower GC contents.

The results of our predictor are shown in figure 4 as a histogram of protein lengths found by our predictor. The distributions of protein length follow the distributions in a previous study (6).

Species	TP	FP	TN	FN	Ref. orfs	Predicted
B. japonicum	35%	18%	82%	64%	8621	3070
C. trachomatis	63%	14%	84%	37%	935	588
D. turgidum	49%	14%	87%	51%	1865	906
R. baltica	39%	16%	82%	61%	7405	2898
S.cerevisiae	72%	15%	85%	28%	418	302

Table 2. Confusion matrix of ORF predictor.

The performance of the predictor was estimated using all possible fragments in all reading frames which terminate with a stop codon as initial dataset. BLAST database was created from reference proteomes of each organism, and predicted ORFs would be queried against the database. The scores indicate how well our refinement strategy isolates the ORFs which are also included in the reference proteomes.

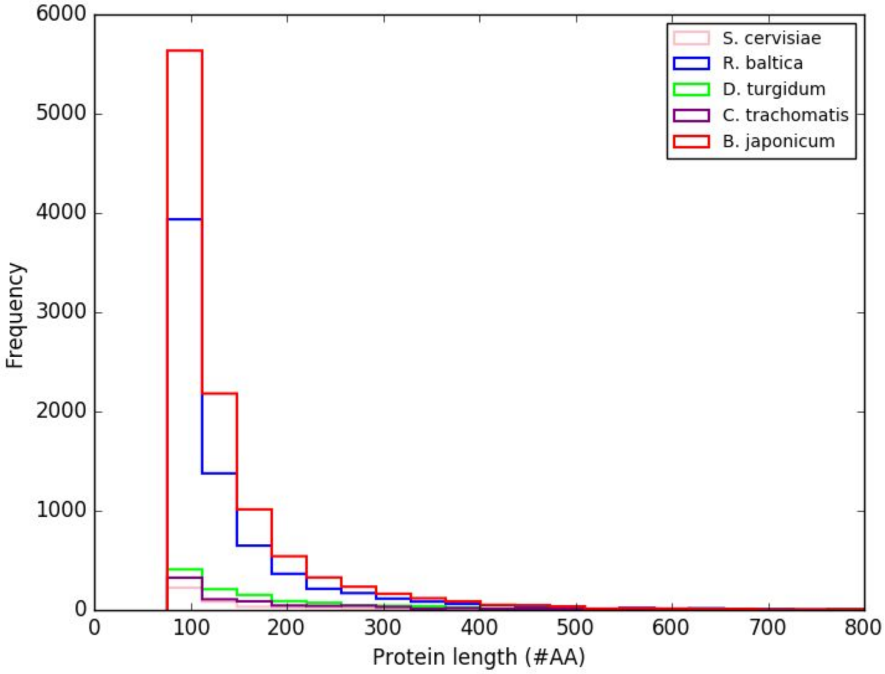


Figure 4. Histogram of protein length distribution in our predicted ORFs.

After using our predictor to find genes in the genomes, we calculated the amino acid frequencies in the translated ORFs (figure 5). Leucine (L) and alanine (A) are overrepresented in all proteomes. *C. trachomatis* and *S. cerevisia*e show similar amino acid frequencies.

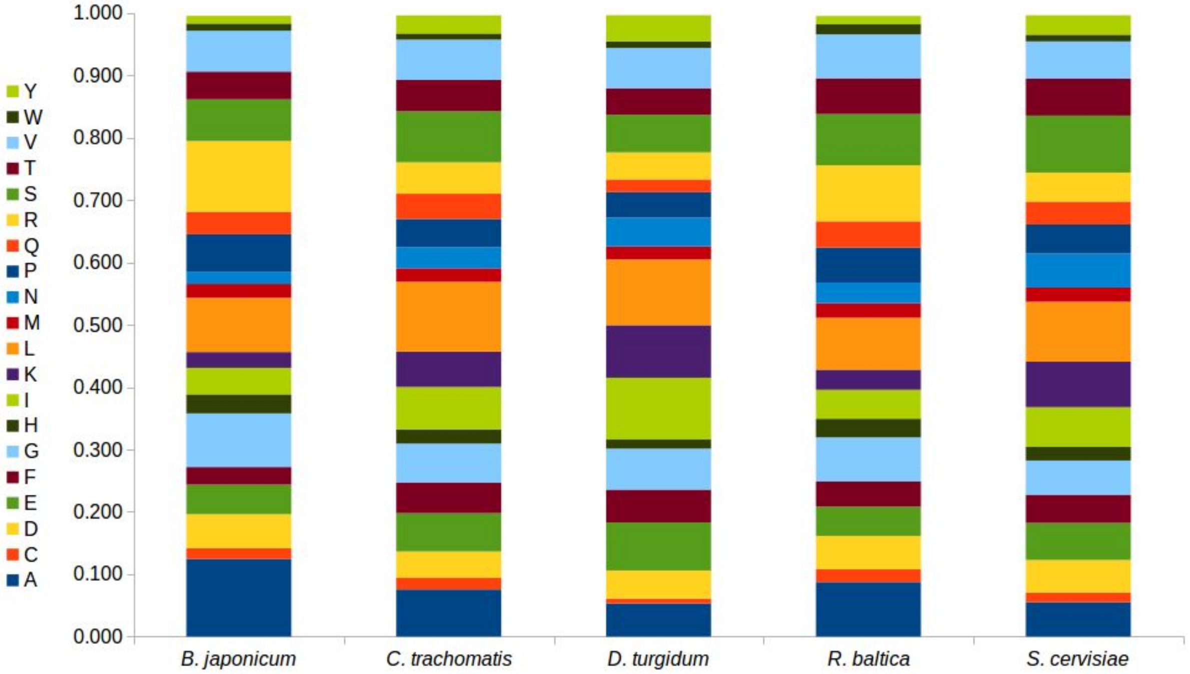


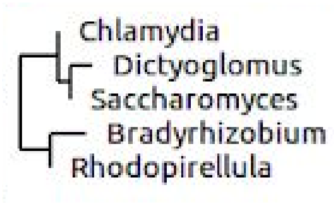
Figure 5. Amino acid frequencies in translated predicted ORFs.

Using our predicted proteomes, we also calculated the di-amino acid frequencies (results not shown).

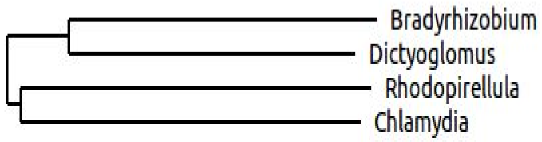
Distance tree comparison

Below, we present phylogenetic trees based on different evidence, which we constructed in the span of the course:

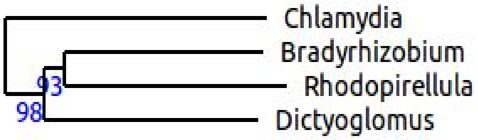
GC-content



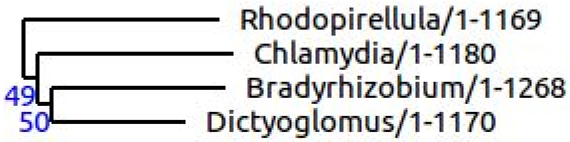
Synteny



10 orthologs alignment



Ribosome alignment



Surprisingly, trees constructed using these different methods, often contradict each other.
We can expect that evidence from GC-content, ortholog clusters and ribosome alignment will be somewhat related and not completely independent of each other, yet the tree architecture still varies. Synteny-based tree was based on gene order conservation, and we found a group of 32 consecutive orthologs where gene order was preserved in all species. This group had gene order preserved between B. japonica and D. turgidum, and the inverted order between the remaining organisms. Such artifact is very unlikely to appear by chance, but suggests an entirely different evolutionary relationship compared to other methods.

Discussion

Out of the five genomes that we analyzed, the genomes of C. trachomatis and S. cerevisiae have many similarities. First, their GC-content is very similar, as well as the distribution of nucleotide compositions, where they both have a relatively unbiased nucleotide composition. These factors lead to a better performance when predicting the ORFs of these two organisms because it is more likely to find all of the dinucleotide combinations in a given gene when the distributions of these are even. This, in turn, leads to a higher entropy and more correctly predicted genes. We also observed that the translated ORFs for these two organisms had a more uniform distribution of amino acids than the other organisms’.
The performance of our ORF predictor varies for different organisms, which could depend on many factors.
For Prokaryotes, ORFs could be verified by looking upstream from the translation initiation codon (ATG in 85% cases, but prokaryotes can have alternative translation initiation codons), in order to find a potential promoter binding motif or TATA box. The TATA motif would be present before a fraction of sequences, but is not a good general rule to base all refinement strategies on. This can not easily by applied for Eukaryotes, as such sites could be anywhere between 100b or 2kb away from the coding region, making this not at all a reliable prediction strategy. Furthermore, “ATG” is a good marker of potential transcription start sites in Eukaryotes, but but Prokaryotes, alternative sites could be used, and our predictor currently is not equipped to evaluate those. Additionally, the gene regulation strategies have diverged in many ways between kingdoms, and different assumptions should be made about what a coding sequence should look like, and how it may be regulated.
These enormous differences mean that prediction software is usually designed for a specific fraction of organisms, and there is no general solution that works for every organism equally well.
Other popular methods, namely GLIMMER (7) and GENSCAN (8) would use a combination of machine learning and information entropy to determine boundaries of coding ORFs. These

predictors are trained on proposed ORFs in sequenced genomes thats were curated, experimentally supported, and showed evidence of conservation in other species, and very specific assumptions about genes were made based on experimental evidence, and perform the best on organisms most similar to training data. Generally, they predict few false positives, but will not find all the annotated orfs.

Our predictor performs better on low-GC genomes, and the best on Yeast chromosome. This could have something to do with the extent of how organisms utilize the reading frames of the same genomic region to encode different proteins, as the high entropy overlapping transcripts would all have a higher chance to be selected by our algorithm.

In the future, we would continue to utilize the information entropy differences in the genome, but in a more refined way, by calculating the entropy of specific region in a sliding window, and assessing local increases as a suggestion for possible transcription start sites(9).

Given more time, our predictor would be supplemented to be customized according to statistics extracted from genome. Utilizing upstream TATA-box, promoter binding sequences and alternative start codons could provide more information and make ORF predictions better in Prokaryotes.

If machine learning is to be utilized, supervised methods could help our predictions to be as similar to annotation - but there might be limited use for that, as with current popularity of sequencing, experimentally verified data is easier to obtain and is more reliable. Instead, we would like to explore unsupervised approaches to more objectively assess information content of given sequences, to find functional, non-coding regions of potential scientific interest.

-

References

(1) Romiguier, J., & Roux, C. (2017). Analytical Biases Associated with GC-Content in Molecular Evolution. *Frontiers in Genetics*, *8*, 16. http://doi.org/10.3389/fgene.2017.00016

(2) Madigan,MT. and Martinko JM. (2003). *Brock biology of microorganisms* (10th ed.). Pearson-Prentice Hall.

(3) F.H.C. Crick, The origin of the genetic code, Journal of Molecular Biology, Volume 38, Issue 3, 1968, Pages 367-379, ISSN 0022-2836, https://doi.org/10.1016/0022-2836(68)90392-6.

(4) Shannon, Claude E. (July–October 1948). "A Mathematical Theory of Communication". Bell System Technical Journal. 27 (3): 379–423.

(5) "Scoredist: A simple and robust protein sequence distance estimator" Erik LL Sonnhammer and Volker Hollich BMC Bioinformatics 6:108 (2005)

(6) Tiessen, A., Pérez-Rodríguez, P., & Delaye-Arredondo, L. J. (2012). Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. *BMC Research Notes*, *5*, 85. http://doi.org/10.1186/1756-0500-5-85

(7) Salzberg, S. L.; Delcher, A. L.; Kasif, S.; White, O. (1998). "Microbial gene identification using interpolated Markov models". Nucleic Acids Research. 26 (2): 544–548. doi:10.1093/nar/26.2.544. PMC 147303 Freely accessible. PMID 9421513.

(8) Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78-94.

(9) David Koslicki; Topological entropy of DNA sequences, Bioinformatics, Volume 27, Issue 8, 15 April 2011, Pages 1061–1067, https://doi.org/10.1093/bioinformatics/btr077