Large Scale Comparative Genomics of Codon Context

José Paulo Ferreira Lousado[1], Gabriela Moura[2], Miguel Pinheiro[3], Manuel A. S. Santos[2] and José Luís Oliveira[3]
[1]*Instituto Politécnico de Viseu and Escola Superior de Tecnologia e Gestão de Lamego*
[2]*Department of Biology and CESAM,* [3]*Institute of Electronics and Telematics Engineering of Aveiro (IEETA) Email: jlo@ieeta.ua.pt*

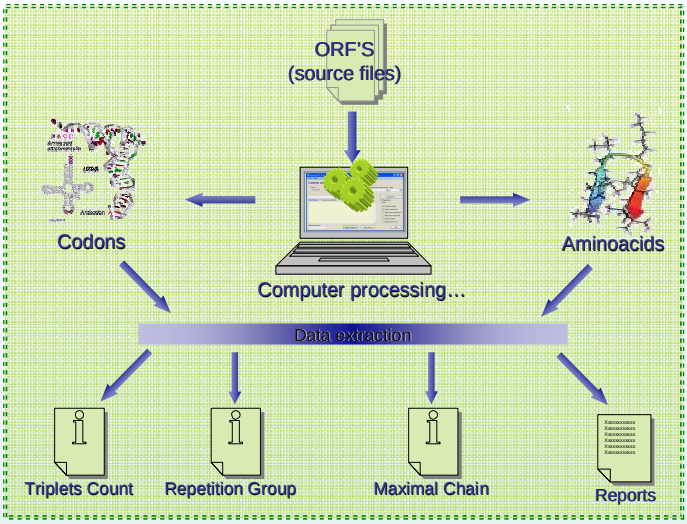
departamento de biologia
universidade de aveiro

Introduction

The efficiency of protein synthesis is highly dependent on codon usage and codon context. Indeed, the choice of particular synonymous codons is constrained by neighbour codons (codon context) to optimize mRNA decoding speed and accuracy. This is related to spatial (steric) effects created by the need to accommodate 3 tRNAs in the ribosome A-, P- and E-sites. Since these tRNAs interact with each other, with their cognate codons and with various structural domains of rRNAs, the structure of the 6 nucleotide RNA helix formed by the anticodon-codon interactions is strongly influenced by the type of codons and tRNAs present in the ribosome decoding centre. To ensure proper tRNA selection and correct codon decoding the rRNA monitors the structure of the codon-anticodon RNA helix.

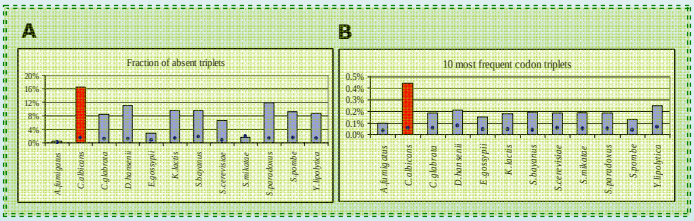
We hypothesized that large scale comparative analysis of 3 consecutive codons, corresponding to the ribosome A-, P- and E-sites codons, would unveil novel codon biases and "bad" codon combinations that may increase decoding error. For this, we have built a software package that counts codon triplets in complete assemblies of open reading frames (ORFeomes) and used the ORFeome sequences of 12 fungal species, including *Aspergillus fumigatus*, *Saccharomyces cerevisiae*, and *Candida albicans* to validate our working hypothesis. We have used data mining methodologies to explore this large dataset of 220,000 combinations of 3 consecutive codons, and extracted the most biased contexts.

Data Flow



Data Analysis

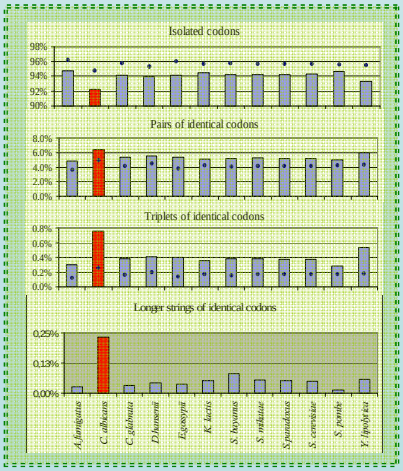
The human pathogen *Candida albicans* has a highly biased codon triplet usage.



In order to better characterize codon triplet distributions in the 12 fungal species we have calculated the percentage of codon triplets that never appear in each ORFeome (**A**), and determined the 10 most frequent codon triplets (**B**). In both cases, *C. albicans* (red bars) has the strongest codon triplet bias. Bars represent the observed percentages and blue dots show expected values. Expected values were obtained by calculating the frequency of each codon in the whole ORFeome.

***C. albicans*' codons appear less times isolated and more times as repeated triplets or longer strings.**

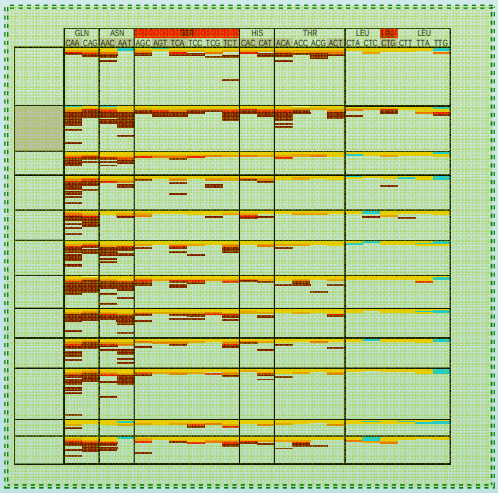
Since codon repeats strongly influence on codon triplets, we have quantified separately the frequency of isolated codons, codons repeated in pairs, in triplets or in strings longer than 3 codons. In *C. albicans*, the frequency of isolated codons is lower than in the other fungal species, however codon repetition biases increases dramatically from pairs to strings of identical codons in this species. Interestingly, isolated codons appear less frequently than expected, as shown by the relative position of observed (bars) and expected values (blue circles).



Codon repeation biases are species and amino acid dependent.

In order to determine whether codon repeats were composed of a special set of codons and/or amino acids, we have quantified all repeats present of each ORFeome for each individual codon. Differences are shown in the diagram below. The existence of bias is highlighted by a color scale in which light blue corresponds to repressed repeats and preferred ones are colored from yellow (low bias) to brown (high bias). For each species, the first line from the top corresponds to all cases in which the codon appears isolated from other identical codons. The second line shows the results of isolated pairs of identical codons and so on, so that the lower the line the longer will the string detected be.

Interestingly, the novel serine decoding CTG codon in *C. albicans* is also preferred in long strings, as is the case for the other standard serine codons. Overall, in *C. albicans* codon repeats is mainly related with a defined set of codons, namely: CAA-Gln; AAC-Asn; AAT-Asn; AGT-Ser; TCA-Ser; TCT-Ser; CAC-His; CAT-His; ACA-Thr; ACT-Thr; and CTG-Ser/Leu.



Conclusions

Our large scale codon triplet analysis show that three-codon contexts are species-specific, although major context rules could also be found. Biases arising from DNA replication and transcription, namely trinucleotide repeats, play an important role in the evolution of ORFeomes. *Candida albicans* revealed unique features and very strong triplet context biases. For example, it has a very high number of consecutive triplet codon repeats, which comprise up to 0,23% of the total ORFeome.

GM is supported by an FCT grant (SFRH/BPD/7195/2001) and MASS is an EMBO Young Investigator. This work is supported by the FCT project POCTI/BME/39030/2001, IEETA and the II-UA (CTS-12).