Statistical, Computational and Visualization Methodologies to Unveil Gene Primary Structure Features

M. Pinheiro[1], V. Afreixo[1], G. Moura[2], A. Freitas[3], M. A. S. Santos[2], J. L. Oliveira[1]
[1]IEETA/DET, University of Aveiro, Aveiro, Portugal
[2]Department of Biology, University of Aveiro, Aveiro, Portugal
[3]Department of Mathematics, University of Aveiro, Aveiro, Portugal

Summary

Objectives: Gene sequence features such as codon bias, codon context, and codon expansion (e.g. tri-nucleotide repeats) can be better understood at the genomic scale level by combining statistical methodologies with advanced computer algorithms and data visualization through sophisticated graphical interfaces. This paper presents the ANACONDA system, a bioinformatics application for gene primary structure analysis.

Methods: Codon usage tables using absolute metrics and software for multivariate analysis of codon and amino acid usage are available in public databases. However, they do not provide easy computational and statistical tools to carry out detailed gene primary structure analysis on a genomic scale. We propose the usage of several statistical methods – contingency table analysis, residual analysis, multivariate analysis (cluster analysis) – to analyze the codon bias under various aspects (degree of association, contexts and clustering).

Results: The developed solution is a software application that provides a user-guided analysis of codon sequences considering several contexts and codon usage on a genomic scale. The utilization of this tool in our molecular biology laboratory is focused on particular genomes, especially those from *Saccharomyces cerevisiae, Candida albicans* and *Escherichia coli*. In order to illustrate the applicability and output layouts of the software these species are herein used as examples.

Conclusions: The statistical tools incorporated in the system are allowing to obtain global views of important sequence features. It is expected that the results obtained will permit identification of general rules that govern codon context and codon usage in any genome. Additionally, identification of genes containing expanded codons that arise as a consequence of erroneous DNA replication events will permit uncovering new genes associated with human disease.

Keywords

Bioinformatics software, codon context, codon bias, contingency tables, residual analysis, cluster analysis

Methods Inf Med 2006; 45: 163–8

1. Introduction

Genome sequencing is opening unprecedented ways for understanding how primary gene structure is organized. Two of the most studied open reading frame characteristics are codon usage and codon context. Codons are organized in the open reading frames according to specific rules that determine their usage frequency and context. Since codons are the primary gene structure features that determine the sequence of amino acids in proteins – they interact with the tRNA anticodons during mRNA decoding by the ribosome – understanding the rules that govern codon usage and context is of critical importance in understanding how genes evolve and also how genetic diversity is created. A number of studies have already shown that each genome uses a set of preferred codons and that codon context is not a random event [1-3]. However, the general rules that govern codon usage and context remain largely elusive. So, one may prompt the question: can they be unraveled using genomic scale approaches by combining bioinformatics, statistical and computer visualization tools?

The other important feature of gene primary structure, in particular in eukaryotic genomes, is the existence of tri-nucleotide repeats that are in some cases associated with a number of neurodegenerative diseases, namely Huntington disease [4]. These tri-nucleotide repeats are sometimes organized as tandem repeated codons (ex: CUG) that significantly expand the usage of a single codon and consequently the number of contiguous residues of a single amino acid in a particular protein domain, thus rendering it non-functional [5]. These tri-nucleotide repeats are widespread in the human genome, thus assuming biomedical relevance and making them important targets for bioinformatics and statistical analysis. Such genome wide surveys of tri-nucleotide repeats, identification of genes that contain them and statistical analysis of their distribution can easily be carried out using bioinformatics and biostatistics tools.

Traditional methods used for codon usage and context analysis do not provide user-friendly tools to carry out detailed primary gene structure analysis at a genomic scale. Codon usage tables using absolute metrics are available in public databases for any sequenced gene or genome and freeware software for multivariate analysis (correspondence analysis) of codon and amino acid usage is also readily available, however sophisticated statistical and data visualization tools are clearly lacking.

In this paper we describe a bioinformatics system named ANACONDA that offers a set of statistical tools and visual models for gene sequence analysis.

2. Methods

2.1 Analysis of Contingency Tables

Tests of independence and residual analysis for contingency tables can easily be found in specialized literature surrounding categorical data analysis (e.g. [6-8]). These methodologies have been applied to distinct areas in order to study the association between two variables subdivided in mutually exclusive categories.

For our purpose, the data of gene sequences are processed in contingency

tables. Thus, for example, if we intend to investigate the 3' codon pair context we construct a 64×64 contingency table where the rows correspond to the codons in the P-site and the columns to the codons in the A-site of the ribosome.

Consider an $r \times c$ contingency table with multinomial distribution where N is the total number of observations, n_{ij} the number of classified observations for the cell (i, j), $n_{i\bullet}$ and $n_{\bullet j}$ are the total marginal for the i-th row and for the j-th column, respectively.

The statistical methodology herein proposed involves the following computations: The frequencies expected under independence of the table,

$$e_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{N}$$

The standardized residuals

$$r_{ij} = \frac{\left(n_{ij} - e_{ij}\right)}{\sqrt{e_{ij}}}$$

The adjusted residuals d_{ij},

$$d_{ij} = \frac{r_{ij}}{\sqrt{\left(1 - \frac{n_{i\bullet}}{N}\right)\left(1 - \frac{n_{\bullet j}}{N}\right)}}$$

The hypothesis of independence is tested using the well known Pearson's statistic,

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} r_{ij}^{\,2} .$$

To measure the degree of association on the table we calculate the Cramér coefficient,

$$C = \sqrt{\frac{\chi^2}{N \min(r-1, c-1)}}$$

From [10] it is known that, under independence of the table, the adjusted residuals have a standardized normal probability distribution, and therefore

$$P(-3 \le d_{ij} \le 3) \approx 0.9973, \text{ as } N \to +\infty.$$

This means that, for a 99.73% confidence level, the cell (i, j) is considered responsible for the eventual rejection of independence if $|d_{ij}| > 3$. By this way, we identify the pairs as being highly biased.

A number of different mathematical methodologies have recently been used to study codon context bias (e.g. [2, 11-13]). Most of these are based on z-scores-type tests and give information about preference and rejection. Basically, those methodologies differ with the assumed probabilistic model.

We have performed a comparison between the results obtained by analysis of residues with the results of computations of z-scores based on probabilistic model assumptions for the data considered by [2], [11] and [12]. As a result, we found that all the different models used indicated the same codon pairs as having statistically significant bias. So it is possible to use any of these known methodologies to detect bias.

The advantage of the proposed methodology is that its theory of inference is well known, yielding an analysis that is more sequential, easily interpretable and has more complementary tools of analysis (e.g., measures of association). The adjusted residual gives direct information about preference and rejection in relation to what would be expected on a random basis. Furthermore, its probability distribution, under the hypothesis of independence in the contingency table is determined without simu-

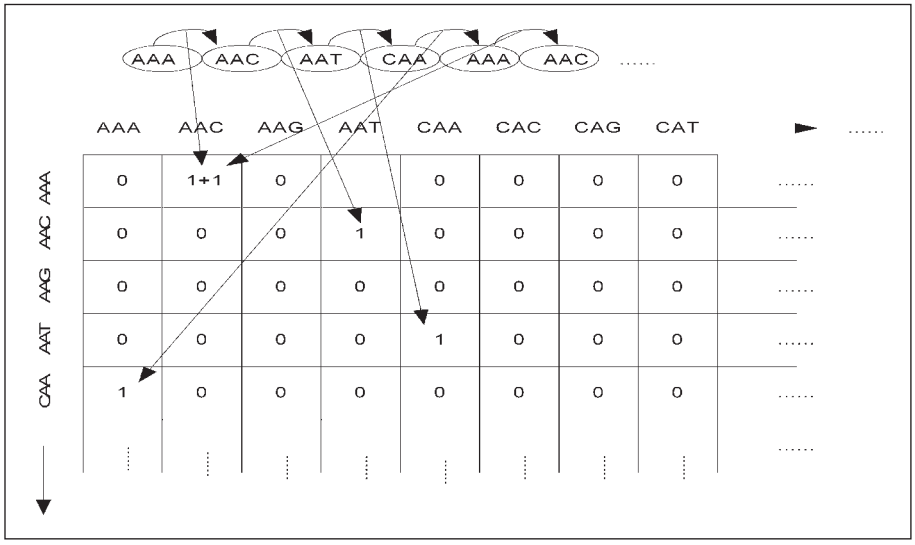

Fig. 1 Schematic representation of the codon context quantification process used in ANACONDA over a hypothetical sequence. The quantification method consists of fixing each codon in the P-site, matching it to the A-site codon and incrementing the cell (P-site, A-site) in the quantification table.

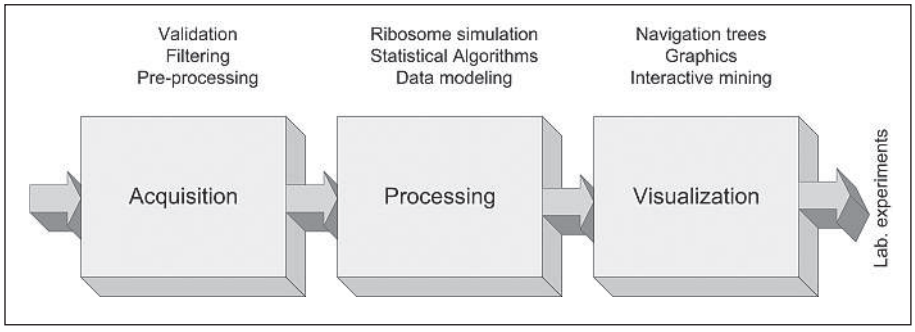

Fig. 2 Architecture of the ANACONDA software. The ANACONDA package contains a data acquisition module that permits downloading raw data from genome databases and filtering it into a local database. This data is then processed using a ribosome simulation algorithm and transferred to a 64×64 table that renders itself to statistical analysis. The processed data is then transferred to the visualization module that has a number of different tools that permit different types of data visualization and analysis.

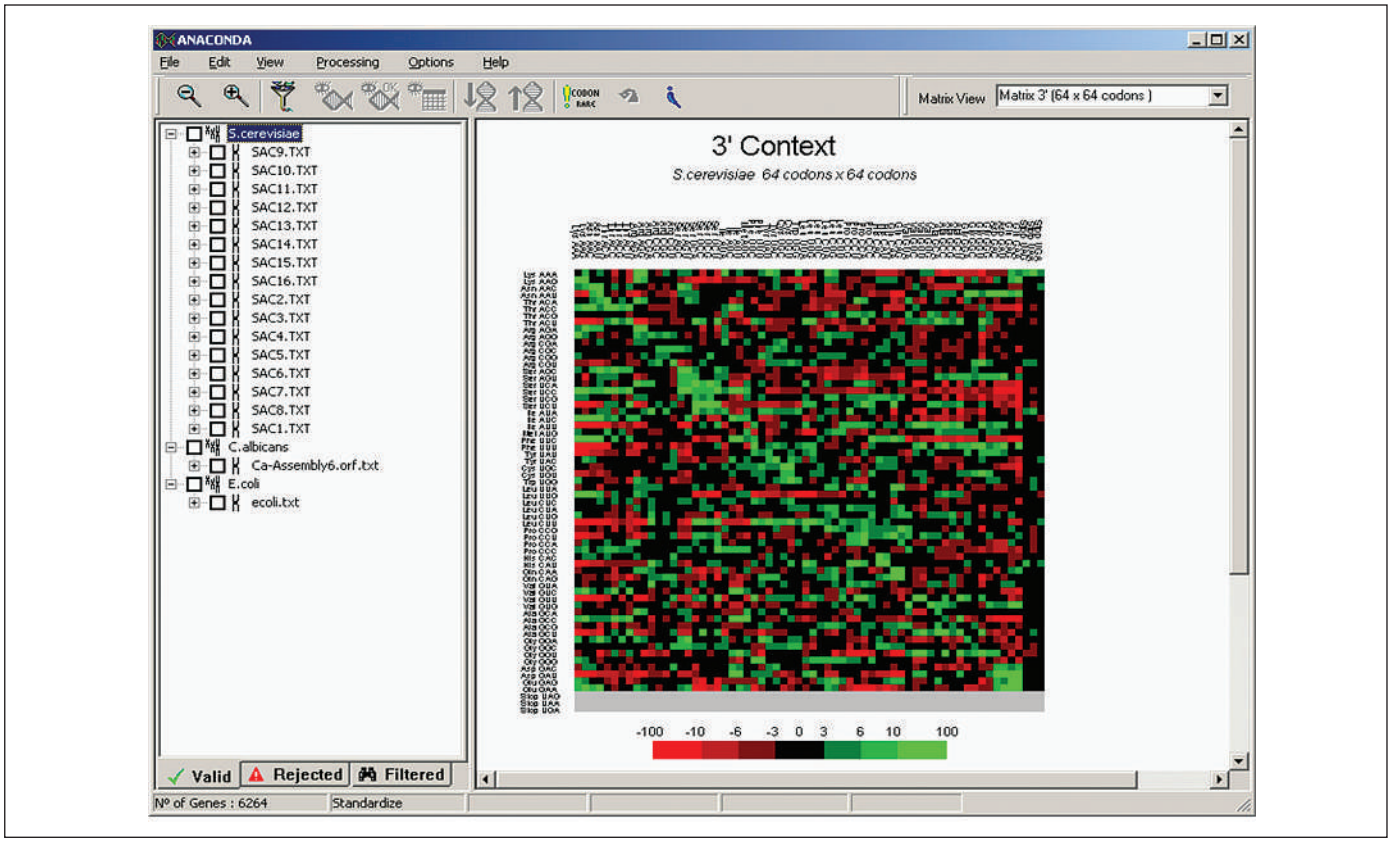

Fig. 3 ANACONDA's main window showing the colored genomic map for codon context analysis of *S. cerevisiae*. The sequences downloaded are controlled on a tree-like structure (left side panel). Each little square of the map corresponds to one individual context composed of one fixed codon (rows) and one 3' codon (columns). The colors visually expose the intensity of the residual values in the contingency table, calculated for each codon (red for negative values and green for positive ones as shown in the scale).

lation techniques ([13] for instance, use a Monte Carlo technique to estimate distribution parameters).

2.2 Multivariate Analysis

In order to find codon context patterns in the contingency tables, lines and columns can be grouped using classifying methodologies such as cluster analysis [14, 15]. These patterns are determined by calculating similarities between two vectors of the contingency table using, for example, Pearson correlation coefficients and applying single linkage clustering.

Having two vectors $X = (X_1, X_2, ..., X_n)$ and $Y = (Y_1, Y_2, ..., Y_n)$ and considering that all the elements of the vectors have the same weight, the Pearson correlation coefficient centered is defined by:

Fig. 4 Graphical display showing two matrices or genomic maps for codon context analysis, one with and the other without clustering. Data clustering permits the identification of patterns of rejected and preferred codon pairs, thus helping the identification of general rules governing evolution.

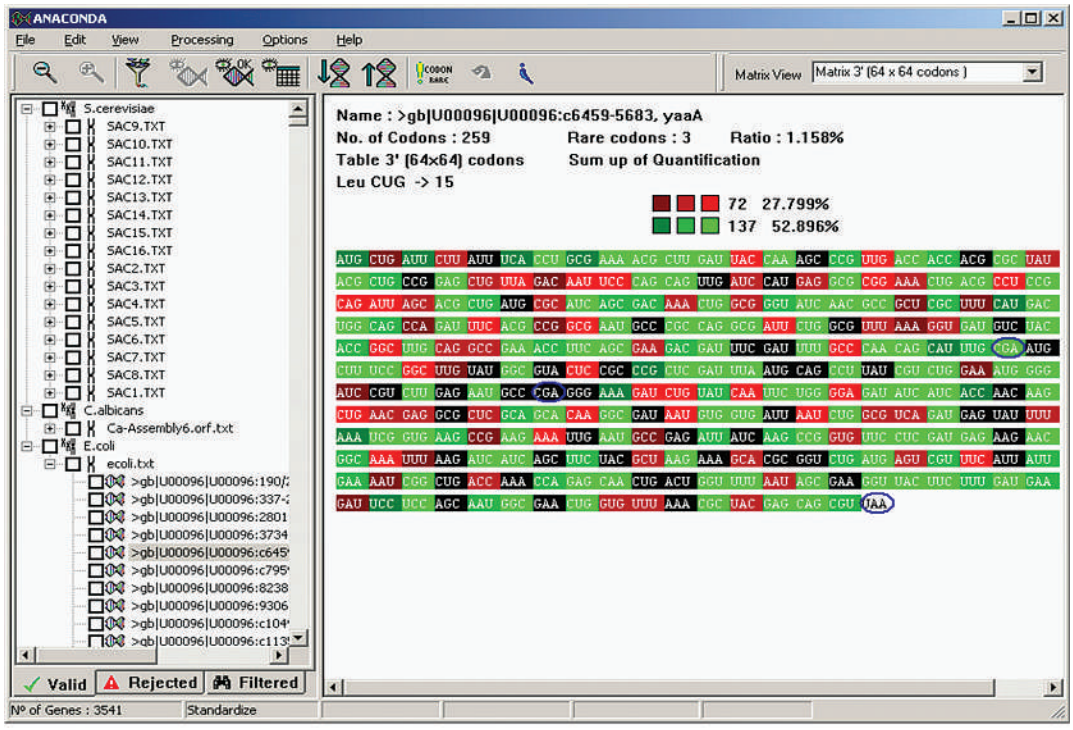

Fig. 5 See legend on opposite page

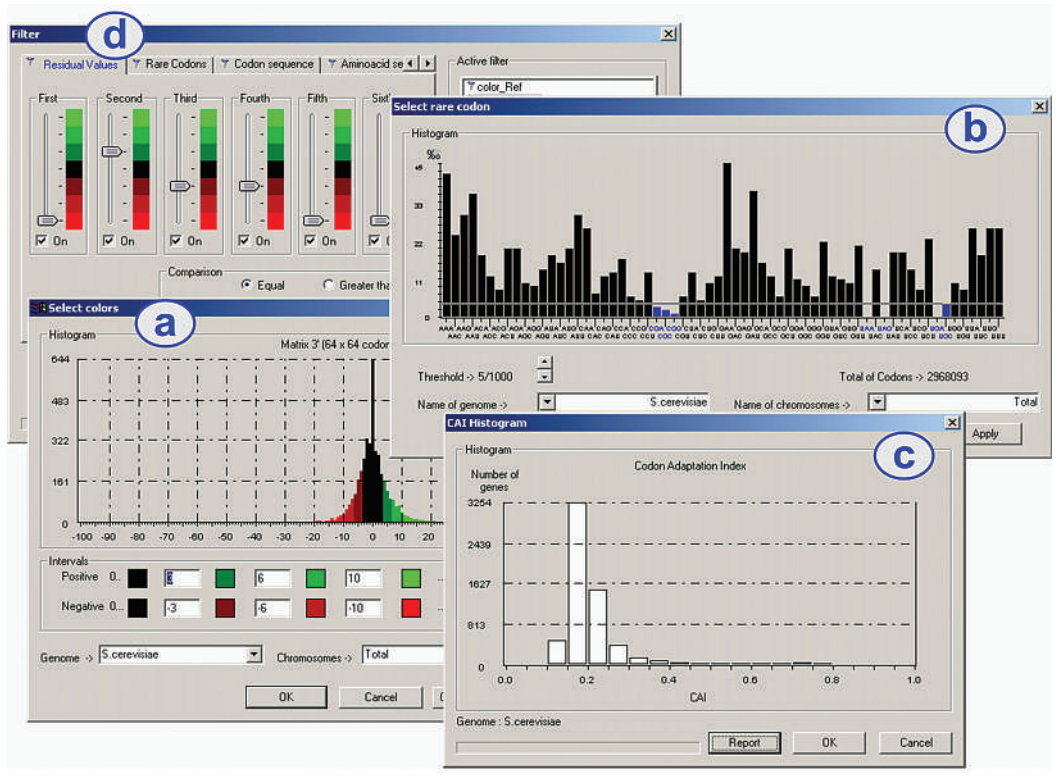

Fig. 6 See legend on opposite page

Fig. 5 ANACONDA's main window showing a particular gene sequence of *E.coli*. The sequence download process is controlled on a tree-like structure (left side panel) where the user can find validated, rejected and specific filtered genes. The colors visually expose the intensity of the residual values in the contingency table, calculated for each codon (red for negative values and green for positive ones).

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \overline{X}}{\sigma_X}\right)\left(\frac{Y_i - \overline{Y}}{\sigma_Y}\right),$$

where $\overline{X}$ and $\overline{Y}$ represent the averages of X and Y components, and σ_x and σ_y represent the standard deviations of X and Y respectively. The Pearson coefficients can assume a value between –1 and 1 ($|r| = 1$ indicates maximum dependency).

Alternatively, the uncentered correlation coefficient is defined by:

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i)^2}}\right)\left(\frac{Y_j}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i)^2}}\right)$$

This correlation is equal to the centered correlation, when the average of the vector components is zero.

The main difference between centered and uncentered Pearson coefficients becomes apparent when we consider two vectors with identical shape but with an offset between them. In this situation we obtain the maximum dependence using the centered coefficient, but not using the uncentered one, which could be an advantage.

This clustering analysis uses a single linkage calculation where the distance between groups is defined as that of the closest pair of individuals, where only pairs consisting of one individual from each group are considered. The single linkage method produces clusters with "chaining effect", i.e. any element of a group is more "similar" to an element of the same group than to any element of another group.

Fig. 6 Multiple overlapped windows that show some of the ANACONDA facilities. a) A histogram graph showing the distribution of negative and positive residues for codon contexts in the 64 x 64 codon matrix. Higher statistical significance of codon context is related to higher distances from the zero value. b) This histogram shows the frequency of each codon in the genome and allows the definition of the maximum value that characterizes a rare codon in the study. c) Codon Adaptation Indexes (CAI) histogram, giving predictions about gene expression levels for the selected genome. d) An editor that allows the definition of specific color patterns to be searched in the genome using algebraic and Boolean combinations.

3. Results

Based on the above mentioned requirements for biological research and aiming to include the proposed statistic methodologies, we have built a software application for codon context analysis (ANACONDA). Its architecture is supported by three main blocks: Acquisition, Processing and Visualization (Fig. 2). Each module was developed in a component-based approach which simplifies replacement, updating or insertion of new modules. The conceptual metaphor of ANACONDA provides seamless navigation, through a tree-like paradigm, over gene sequences, crossing species, chromosomes, genes and codons.

The Acquisition module deals with genome input files, reading and interpreting sequences of complete or partial sets of ORFs from public or private genome databases. FASTA is the main format that can be used, but Genbank or others can easily be integrated due to ANACONDA's plug-in architecture. Several filters can be used in the reading process to ensure that the screened sequences have the best possible quality, and to avoid introducing background noise in the following analysis. In the filters output, genes are separated in classes (valid, rejected) according to user-defined scanning patterns (i.e. check for a nucleotide counting multiple of 3, start with an AUG codon, stop with UAG, UAA or UGA codons, total gene length, etc.).

The Processing module is the core of the application. After the gene sequences acquisition, the generated data is converted into a contingency table that includes the corresponding observed values of Pearson's statistics, the Cramer's coefficient of association, and the matrix of adjusted residues [16]. After this processing, the data becomes available for the Visualization module where the user can investigate the existence of significant bias in the codon context and exploit possible evidence expressed by the matrices of residual values.

The Visualization module is supported by two-sided windows – the navigation tree, on the left side, and the analysis window, on the right side (Fig. 3). The analysis window shows the usual output layout of a residual analysis for codon context, in which each residue of the contingency table is colored according to a given scale (green for preferred contexts and red for rejected ones) to ease interpretation.

A cluster analysis tool also allows calculating similarities between two vectors of the contingency table using, for example, Pearson correlation coefficients and applying single linkage clustering (Fig. 4). This technique is used to group lines and columns (codons) of the correlation matrix, allowing highlight global patterns in the genes. Figure 4 presents and compares two contingency matrixes. In the first, the one behind, both axis indices follow a predefined order, while in the topmost matrix the axis values are defined by the cluster analysis results.

It is also possible to visualize the results of residual analysis at the gene level (Fig. 5), where the individual sequences are presented and colored according to the same scale. This layout is prepared to highlight other important features such as the distribution of rare codons in the ORFs, the ratio of rare codons relative to the total number of codons, the GC% at the 1st, 2nd, and 3rd codon position, the CAI and the effective number of codons [15, 16] of the gene being shown, etc.

This module offers a set of tools that permit carrying out several tasks such as searching pre-defined sequence patterns, visualizing data in histogram format, providing cluster analysis over codon-context data and exporting residual tables or other results for further statistical analysis.

Figure 6 presents a set of facilities that are also available in the software. For instance, a histogram graph (Fig. 6a) shows the distribution of negative and positive residues for codon contexts in the 64 × 64 codon matrix. Dark colors (central zone of the scale) indicate residual values that fall within the –3 to +3 interval, which are statistically non-significant. Another histo-

gram (Fig. 6b) shows the frequency of each codon in the genome. Within this window it is possible to adjust the threshold value that defines a rare codon. Figure 6c presents another tool of the software in which Codon Adaptation Indexes (CAI) are calculated and plotted in a histogram, giving predictions about gene expression levels for the selected genome [14]. There is also a filter tool for searching specific color patterns inside genes (Fig. 6d). By moving the slider bars we can easily define the pattern to be searched and even assume a "don't care" value for some pattern positions. Finally, ANACONDA integrates other special predefined filters, such as the search for rare codon patterns, nucleotide or amino acid motives, genes with a given ratio of negative to positive residues, codon usage indexes other than CAI or rare codon frequencies.

4. Discussion

The ANACONDA software package provides a set of statistical, bioinformatics and data visualization tools for gene primary structure analysis. Full gene sets are automatically downloaded from public databases, namely Genbank. Downloaded sequences can then be analyzed in different ways providing genome scale information about codon context, codon usage, nucleotide repeats within Open Reading Frames and others. More importantly, the data can be processed and visualized in different graphical formats that can reveal new insights from the interpretation of very large data sets.

The statistical tools that are incorporated in the system for data clustering, residual analysis and histogram plotting of calculated indexes allow reaching new conclusions on primary gene structure features at a genomic scale. We expect that the results obtained will permit identifying some general rules that govern codon context and codon usage in any genome. Additionally, the identification of genes containing expanded codons that arise as a consequence of erroneous DNA replication events will permit uncovering new genes associated to human disease.

Ongoing studies are carried out using the yeast genome as a model system to study the effect of translational selection on gene evolution. For this, the complete genomes of *S. cerevisiae* and *C. albicans* are being analyzed with ANACONDA, using the full set of facilities offered in the described package. Our preliminary data confirms previous results obtained in *E. coli* indicating that codon context is highly biased, since each single genome has a characteristic codon context pattern that represents a molecular fingerprint of that species.

The Anaconda software was developed in C++ language, it runs on MS Windows, and it is publicly available for non-commercial usage at http://www.bio.ua.pt/genomica/lab.

Acknowledgments
We are thankful to FCT for financial support (POCTI/BME/39030/2001). Gabriela Moura is supported by the FCT grant SFRH/BPD/7195/2001, and Manuel Santos is supported by an EMBO YIP Award. Adelaide Freitas is partially supported by Unidade de Investigação Matemática e Aplicações of University of Aveiro through POCTI-FCT, co-financed by the European Community fund FEDER.

References

1. Comeron JM, Aguadé M. An evaluation of measures of synonymous codon usage bias. J Mol Evol 1998; 47: 268-74.
2. Boycheva S, Chkodrov G, Ivanov I. Codon pairs in the genome of Escherichia coli. Bioinformatic 2003; 19: 987-98.
3. Berg OG, Silva PJ. Codon bias in Escherichia coli: the influence of codon context on mutation and selection. Nucleic Acids Res 1997; 25; 1397-1404.
4. Margolis RL, Ross CA. Diagnosis of Huntington disease. Clin Chem 2003; 49; 1726-32.
5. Parekh-Olmedo H, Krainc D, Kmiec EB. Targeted gene repair and its application to neurodegenerative disorders. Neoron 2002; 33: 495-8.
6. Avery PJ, Henderson DA. Fitting Markov chain models to discrete state series such as DNA sequences. Applied Statistics 1999; 48: 53-61.
7. Sheskin DJ. Parametric and nonparametric statistical procedures. Chapman & Hall, 2000.
8. Bishop YMM, Fienberg SE, Holland PW. Discrete Multivariate Analysis, Theory and Practice. MIT Press, 1975.
9. Irwin B, Heck JD, Wesley G. Codon Pair Utilization Biases Influence Translational Elongation Step Times. The Journal of Biological Chemistry 1995; 270 (39): 22801-6.
10. Haberman SJ. The analysis of residuals in cross-classified tables. Biometrics 1973; 29 (22): 205-20.
11. Shah AA, Giddings MC, Parvaz JB, Gesteland RF, Atkins JF, Ivanov IP. Computational identification of putative programmed translational frameshift sites. Bioinformatics 2002; 18: 1046-53.
12. Hooper SD, Berg OG. Detection of Genes with Atypical Nucleotide Sequence in Microbial Genomes. J Mol Evol 2002; 54: 365-75.
13. Fedorov A, Saxonov S, Gilbert W. Regularities of context-dependent codon bias in eukaryotic genes. Nucleic Acids Res 2002; 30; 1192-7.
14. Everitt BS. Cluster Analysis. 3rd ed. Edward Arnold, 1998.
15. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. Academic Press, 1994.
16. Everitt BS. The analysis of contingency tables. Chapman and Hall, 1977.
17. Sharp PM, Li WH. The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987; 15 (3): 1281-95.
18. Wright F. The 'effective number of codons' used in a gene. Gene 1990; 87: 23-9.

Correspondence to:
José Luis Oliveira
Universidade de Aveiro, IEETA/DET
3810-193 Aveiro
Portugal
E-mail: jlo@det.ua.pt