Exploiting codon-triplets association for genome primary structure analysis

José P. Lousado, Gabriela R. Moura, Manuel A. S. Santos and José Luis Oliveira

Abstract— **The way evolution shapes the arrangement of synonymous codons within open reading frames (ORF) for fine tuning mRNA decoding efficiency is not yet understood. Since the ribosome has 3 tRNA binding sites, the context of triplets should be relevant to decoding fidelity and efficiency. We have developed a software application for large-scale analysis of codon-triplet associations to shed new light into this problem. The developed algorithms identify codon-triplets context biases, allowing for large scale comparative codon-triplet analysis and for identification of alterations to the standard genetic code.**

***Index Terms*—Bioinformatics, comparative genomics, data mining**

I. Introduction

The degeneracy of the genetic code permits synthesis of identical proteins from mRNAs with very different primary structures. The latter are shaped by tRNA abundance, genome G+C pressure, codon-pair context effects, codon-anticodon interactions, or other DNA replication-, transcription- and translation-driven biases [2]. Also, the arrangement of two consecutive codons in Open Reading Frames (ORFs) has a direct effect on decoding efficiency. For example, altering the 3´context from G to U in the insertion sequence IS911 (A-AAA-AAG) of *E. coli* increases frameshifting from 10% to 60% [1], while, on the other hand, the under-represented codon-pair ACC-CUG is translated faster than its over-represented synonymous pair ACG-CUG [3]. These codon context effects suggest, therefore, that codon-pair biases are strong modulators of mRNA translation speed and accuracy. However, the ribosome has 3 rather than 2 decoding sites, i.e. A-, P- and E-sites, and codon-pairs alone cannot provide a full picture of the selective pressure imposed by the translational machinery on mRNA sequences.

The role of A- and P-sites is well established in aminoacyl-tRNA (aa-tRNA) selection and translocation during ribosomal translation, while, at least from a structural perspective, the way the E-site interferes on mRNA translation is not so clear. However, the presence of a tRNA at the E-site changes allosterically the affinity of the A-site to select the incoming aa-tRNAs [6]. The evidence about this allosteric interaction between the E- and A-sites, together with ribosome crystallography and cryo-EM studies [4], suggest that the 3 tRNAs accommodated in the ribosome at each translocation site are critical for decoding efficiency. In other words, codon-triplets determine the type of tRNAs that are present in the A-, P- and E-sites and it is likely that this will have an important impact on decoding efficiency and accuracy, as is the case of codon-pairs.

GeneSplit is a software framework that highlights codon-triplets patterns in genomes and in complete sets of ORFs (ORFeomes). It helps finding specific patterns or bias that may help elucidate whether codon-triplets associations influences mRNA decoding error.

II. METHODS

In order to perform statistical analyses of codon-triplet biases at the ORFeome level, we have developed GeneSplit to simulate the ribosome during translation. This is achieved by moving the reading window 3 x 3 nucleotides at a time, while reading ORFs from the ATG initiation codon until a stop codon is encountered. The algorithm memorizes all codon-triplets, which represent A-, P- and E-sites during mRNA decoding. Since the analysis of codon triplets generates 3-dimensional 613 data sets for each ORFeome, and the processing time associated with large sequences is typically high, all pre-processed data is stored in a relational database. These large data sets can then be analyzed using data mining tools or direct database queries.

A similar methodology is used to count amino acid triplets. For this, codons are translated into the respective amino acids using standard genetic code rules or using non-standard decoding when necessary.

Repetitions of more than 3 consecutive codons can be excluded from the analysis since they can introduce noise in the codon-triplet analysis. For this, the algorithm ignores the presence of identical codons following three consecutive identical triplets until a different codon appears in the ORF sequence. This methodology can also be used for amino acid triplet counting.

GeneSplit was built in a three-tiered approach: a) Data Repository, i.e. data warehouse where the processed ORFeomes are stored; b) GeneSplit core (GScore), that processes the ORFeomes and uploads the repository; and c) GeneSplit web (GSweb), that provides a web interface to the framework.

978-0-7695-3191-5/08 $25.00 © 2008 IEEE
DOI 10.1109/BIOTECHNO.2008.21

A. *GScore*

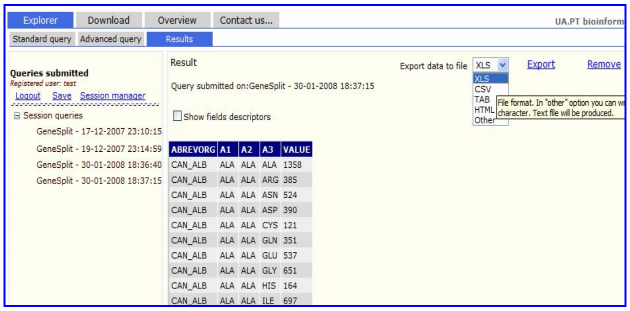
GeneSplit core is the base component responsible for all the data warehouse construction and maintenance. First, it imports genome files (Fasta or Genbank format) and produces, according to the user parameterization, aggregated or per-genome statistics. The results are then stored in CVS format and uploaded into the Data Repository where a post-processing step is performed in order to enhance the warehouse performance and reduce the response time to user queries. This strategy was necessary due to the high volume of data and the time consuming algorithms that are necessary.

Also, this method follows the Data Marts architecture, where several predefined subsets of the overall data warehouse are built to cope with complexity and processing requirements. As a result, and according to the user questions, an adequate selection of the data mart is performed, in order to answer those questions.

This component works primary as a back office application to support the web portal, but it can also be used in standalone mode if one wants to conduct their own personal research in a monolithic way.

B. *GSweb*

The web component handles the user interaction with the data warehouse. It allows both basic and advanced queries (Figure 1). Basic queries make use of a set of predefined questions that can be selected, while the advanced queries allow for building any kind of database queries through the help of wizards that hide the technical complexity of the SQL language and the database schema. An important feature of this component is the Project concept. All the questions that are submitted to the system, by each user, are stored and presented permanently during the working session (Figure 2). With this, the user keeps track of his/her interactions with the repository, both the queries and the results, simplifying the research management.



Queries submitted
Registered user: test
Logout Save Session manager
⊟ Session queries
GeneSplit - 17-12-2007 23:10:15
GeneSplit - 19-12-2007 23:14:59
GeneSplit - 30-01-2008 18:36:40
GeneSplit - 30-01-2008 18:37:15

Result

Query submitted on:GeneSplit - 30-01-2008 18:37:15

Export data to file XLS Export Remove
XLS
CSV
TAB
HTML
Other

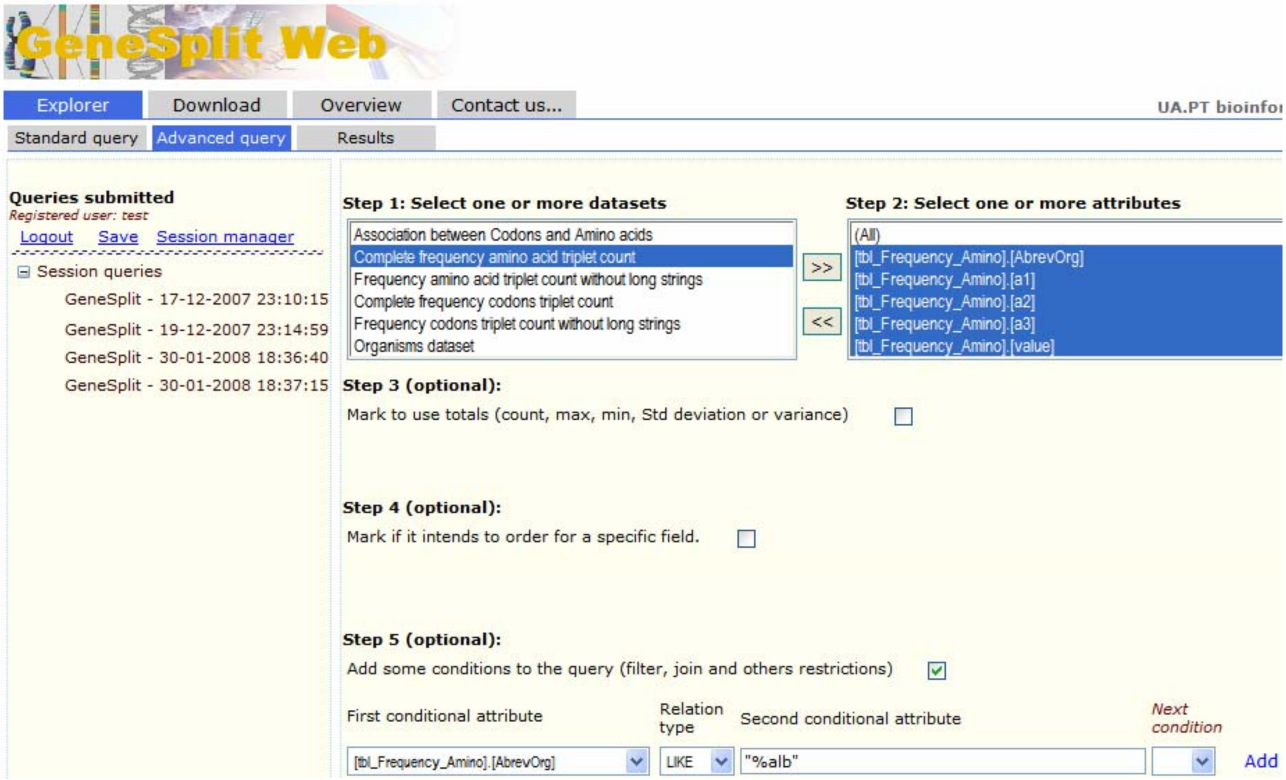
File format. In "other" option you can w
character. Text file will be produced.

□ Show fields descriptors

ABREVORG	A1	A2	A3	VALUE
CAN_ALB	ALA	ALA	ALA	1358
CAN_ALB	ALA	ALA	ARG	385
CAN_ALB	ALA	ALA	ASN	524
CAN_ALB	ALA	ALA	ASP	390
CAN_ALB	ALA	ALA	CYS	121
CAN_ALB	ALA	ALA	GLN	351
CAN_ALB	ALA	ALA	GLU	537
CAN_ALB	ALA	ALA	GLY	651
CAN_ALB	ALA	ALA	HIS	164
CAN_ALB	ALA	ALA	ILE	697

Figure 2 – Session Queries.

The session management is provided for the registered user which only requires introduction of username, password and e-mail. After logging into the system the list of sessions already created and recorded by this user are displayed (**Figure 3**) and each session history can be opened or deleted..



Queries submitted
Registered user: test
Logout Save Session manager
⊟ Session queries
GeneSplit - 17-12-2007 23:10:15
GeneSplit - 19-12-2007 23:14:59
GeneSplit - 30-01-2008 18:36:40
GeneSplit - 30-01-2008 18:37:15

Step 1: Select one or more datasets

Association between Codons and Amino acids
Complete frequency amino acid triplet count
Frequency amino acid triplet count without long strings
Complete frequency codons triplet count
Frequency codons triplet count without long strings
Organisms dataset

Step 2: Select one or more attributes

(All)
[tbl_Frequency_Amino].[AbrevOrg]
[tbl_Frequency_Amino].[a1]
[tbl_Frequency_Amino].[a2]
[tbl_Frequency_Amino].[a3]
[tbl_Frequency_Amino].[value]

Step 3 (optional):

Mark to use totals (count, max, min, Std deviation or variance) □

Step 4 (optional):

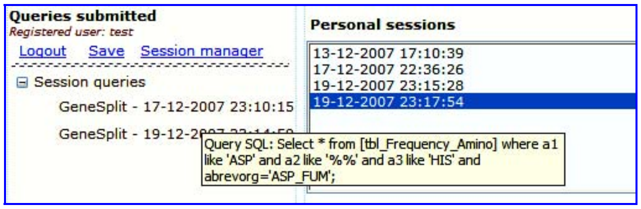
Mark if it intends to order for a specific field. □

Step 5 (optional):

Add some conditions to the query (filter, join and others restrictions) ☑

First conditional attribute	Relation type	Second conditional attribute	*Next condition*
[tbl_Frequency_Amino].[AbrevOrg]	LIKE	"%alb"	

Figure 1 – Queries Wizards.



Queries submitted
Registered user: test
Logout Save Session manager

⊟ Session queries

GeneSplit - 17-12-2007 23:10:15

GeneSplit - 19-12-20

Query SQL: Select * from [tbl_Frequency_Amino] where a1 like 'ASP' and a2 like '%%' and a3 like 'HIS' and abrevorg='ASP_FUM';

Personal sessions

13-12-2007 17:10:39
17-12-2007 22:36:26
19-12-2007 23:15:28
19-12-2007 23:17:54

Figure 3 – Sessions Management.

The portal already has 22 processed orfeomes. Each orfeome is available in two versions – with and without long repetition chains.

The application was designed with the goal to provide a user-friendly interface. Each user can create its one profile and store projects, sessions, search queries, revisiting later the previous studies, repeating or modifying them, in an ubiquitous utilization allowed by the Web interface. Since the recorded information are composed by SQL statements produced only, the internal management is very efficient. Since pre-processed data is handled, most of the queries response are provided in an expedite time.

The retrieval of information from the database can be realized in two different ways: Standard and Advanced Queries (Figure 4).

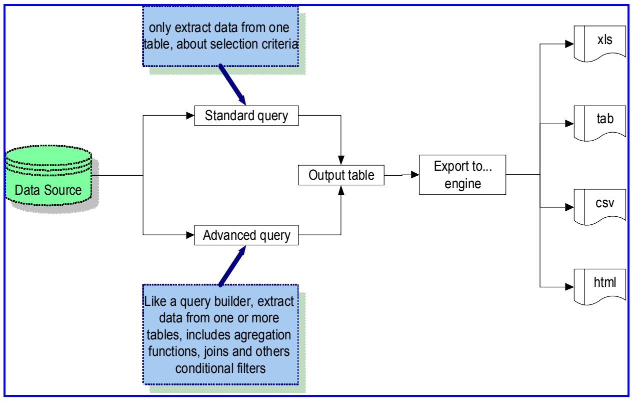
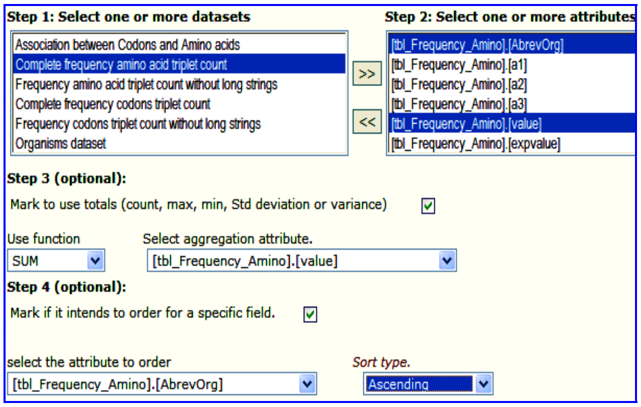

only extract data from one table, about selection criteria
Standard query
Data Source
Output table
Export to... engine
Advanced query
Like a query builder, extract data from one or more tables, includes agregation functions, joins and others conditional filters
xls
tab
csv
html

Figure 4 – GeneSplit workflow.

In the first option, the user may select one of several Orfeomes stored in the database, followed by the selection of characteristics of interests, whether the analysis is over the Orfeome or over the proteome, with or without repetition chains. The database has a particular ontology that allows the user to have access both to the database attributed and to a more descriptive and friendly metadata information. Using this feature one can use, in the list boxes, not the field names but its more meaningful descriptor.

After the submission of the query, the resulting data is displayed on the *Results* tab, to be analyzed or exported in several formats, and stored or removed from the query list. GeneSplit generates data in XLS, CSV, TAB and HTML formats.

The second option is particularly suited for more skilled users that are familiar with SQL and wish to get information in a more complex and versatile way – for instance, joining several tables together and using additional aggregation, sorting and conditions features. For example, to query the total triplets, grouped and sorted by organism (**Figure 5**), it is only necessary to select the fields and the function one wants to use.



Step 1: Select one or more datasets

Association between Codons and Amino acids
Complete frequency amino acid triplet count
Frequency amino acid triplet count without long strings
Complete frequency codons triplet count
Frequency codons triplet count without long strings
Organisms dataset

Step 2: Select one or more attributes

[tbl_Frequency_Amino].[AbrevOrg]
[tbl_Frequency_Amino].[a1]
[tbl_Frequency_Amino].[a2]
[tbl_Frequency_Amino].[a3]
[tbl_Frequency_Amino].[value]
[tbl_Frequency_Amino].[expvalue]

Step 3 (optional):

Mark to use totals (count, max, min, Std deviation or variance) ☑

Use function: SUM

Select aggregation attribute.: [tbl_Frequency_Amino].[value]

Step 4 (optional):

Mark if it intends to order for a specific field. ☑

select the attribute to order: [tbl_Frequency_Amino].[AbrevOrg]

Sort type.: Ascending

Figure 5 – Sample SUM Query.

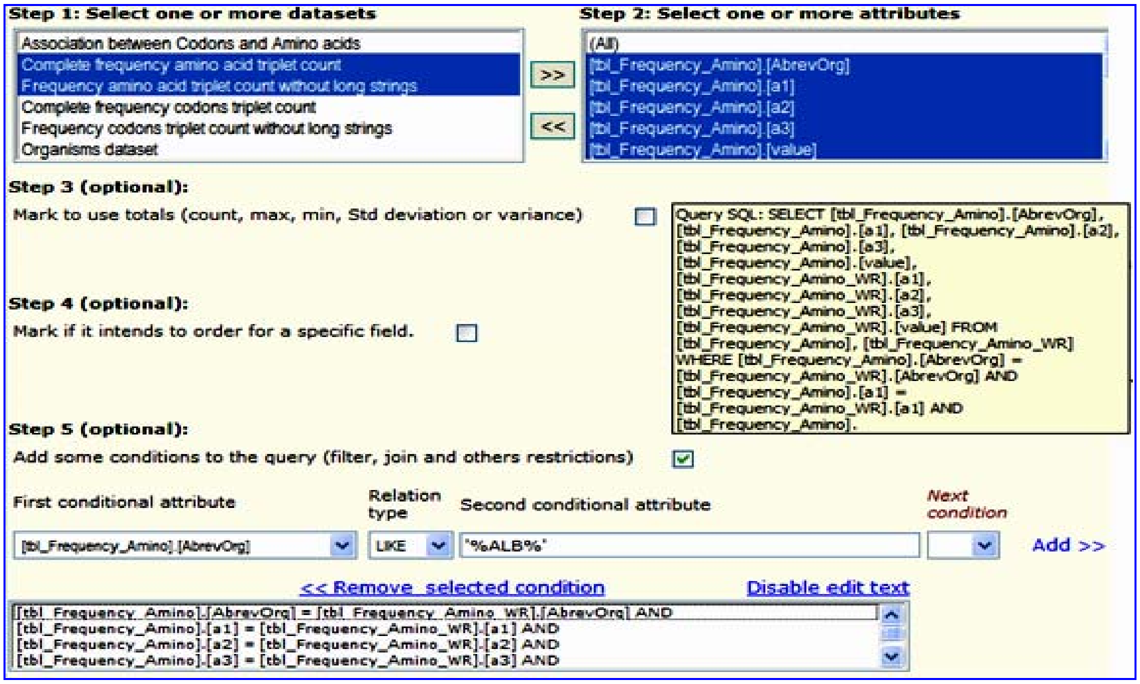
As an example, if one wants to compare two counts, with and without repetition chains, for a given species (*Candida albicans*) it is only necessary to select the two datasets - 'Complete frequency amino acid triplet count' and 'Frequency amino acid triplet count without long strings', the fields of interest and adding the extract conditions (Figure 6).

This specification will produce a standard SQL query, and can thus be easily understood and edited by those who have basic knowledge in this language. This is not a necessary condition to operate with the software. The interaction can be achieved through the wizard.

III. Results

GeneSplit provides a set of functionalities for comparative genomics, such as: codon-triplets and tandem codon repetition analyses; filtering and specific counting (gene, chromosome); sessions management; pre-processing of large datasets for faster results; user friendly interface for database information extraction (predefined queries and advanced queries).

The system has already been used in a case study that highlighted major differences at the level of codon-triplets and tandem codon repetitions between twelve fungal species [5].



Step 1: Select one or more datasets

- Association between Codons and Amino acids
- Complete frequency amino acid triplet count
- Frequency amino acid triplet count without long strings
- Complete frequency codons triplet count
- Frequency codons triplet count without long strings
- Organisms dataset

Step 2: Select one or more attributes

- (All)
- [tbl_Frequency_Amino].[AbrevOrg]
- [tbl_Frequency_Amino].[a1]
- [tbl_Frequency_Amino].[a2]
- [tbl_Frequency_Amino].[a3]
- [tbl_Frequency_Amino].[value]

Step 3 (optional):

Mark to use totals (count, max, min, Std deviation or variance) □

Step 4 (optional):

Mark if it intends to order for a specific field. □

Step 5 (optional):

Add some conditions to the query (filter, join and others restrictions) ☑

First conditional attribute: [tbl_Frequency_Amino].[AbrevOrg]
Relation type: LIKE
Second conditional attribute: '%ALB%'
Next condition

```
Query SQL: SELECT [tbl_Frequency_Amino].[AbrevOrg],
[tbl_Frequency_Amino].[a1], [tbl_Frequency_Amino].[a2],
[tbl_Frequency_Amino].[a3],
[tbl_Frequency_Amino].[value],
[tbl_Frequency_Amino_WR].[a1],
[tbl_Frequency_Amino_WR].[a2],
[tbl_Frequency_Amino_WR].[a3],
[tbl_Frequency_Amino_WR].[value] FROM
[tbl_Frequency_Amino], [tbl_Frequency_Amino_WR]
WHERE [tbl_Frequency_Amino].[AbrevOrg] =
[tbl_Frequency_Amino_WR].[AbrevOrg] AND
[tbl_Frequency_Amino].[a1] =
[tbl_Frequency_Amino_WR].[a1] AND
[tbl_Frequency_Amino].
```

```
[tbl_Frequency_Amino].[AbrevOrg] = [tbl_Frequency_Amino_WR].[AbrevOrg] AND
[tbl_Frequency_Amino].[a1] = [tbl_Frequency_Amino_WR].[a1] AND
[tbl_Frequency_Amino].[a2] = [tbl_Frequency_Amino_WR].[a2] AND
[tbl_Frequency_Amino].[a3] = [tbl_Frequency_Amino_WR].[a3] AND
```

Figure 6 – Sample JOIN Query.

IV. CONCLUSIONS

The set of tools that were presented in this paper enable researchers to study the characteristics of codon and amino acids triplets in any genome, and permit revealing hidden patterns in the genome primary structure. A special care was put in the usability and on the presentation of a simple web user interface to query the database. Each user session can be stored for later utilization, for reviewing queries and results, for increasing new requests and studies.

The GeneSplit software package is publicly available at http://bioinformatics.ua.pt/genesplit.

V. ACKNOWLEDGEMENTS

GM was supported by an FCT grant SFRH/BPD/7195/2001 and MASS was supported by an EMBO YIP Award.

VI. REFERENCES

[1] Bertrand, C. et al., “Influence of the stacking potential of the base 3' of tandem shift codons on -1 ribosomal frameshifting used for gene expression”, Rna, 8, 2002, 16-28.

[2] Dong, H., et al., “Co-variation of tRNA abundance and codon usage in Es-cherichia coli at different growth rates”, J Mol Biol, 260, 1996, 649-663.

[3] Irwin, B., et al., “Codon pair utilization biases influence translational elongation step times”, J Biol Chem, 270, 1995, 22801-22806.

[4] Korostelev, A., et al., “Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements”, Cell, 126, 2006, 1065-1077.

[5] Moura, G., et al., “Codon-triplet context unveils unique features of the Candida albicans protein coding genome”, BMC Genomics, 4, 2007.

[6] Nierhaus, K.H. “Decoding errors and the involvement of the E-site”, Biochimie, 88, 2006, 1013-1019.

[7] Yarus, M. et. al., “Origins of the Genetic Code: The Escaped Triplet Theory”, Annual Review of Biochemistry, 74, 2005, 179-198