

# GeneSplit – Uma Aplicação para o Estudo de Associações de Codões e de Aminoácidos em ORFeomas

José P. Lousado<sup>1</sup>, Gabriela R. Moura<sup>2</sup>, Manuel A. S. Santos<sup>2</sup>, José Luis Oliveira<sup>3</sup>

[jlousado@estgl.ipv.pt](mailto:jlousado@estgl.ipv.pt), [gmoura@ua.pt](mailto:gmoura@ua.pt), [msantos@bio.ua.pt](mailto:msantos@bio.ua.pt), [jlo@ua.pt](mailto:jlo@ua.pt)

<sup>1</sup> Centro de Estudos em Educação, Tecnologias e Saúde, Instituto Politécnico de Viseu - Campus Politécnico, 3504-510, Viseu, Portugal

<sup>2</sup> CESAM & Departamento. de Biologia, Universidade de Aveiro, 3810-193, Aveiro, Portugal

<sup>3</sup> DETI/IEETA, Universidade de Aveiro, 3810-193, Aveiro, Portugal

**Resumo:** A descodificação de genomas, em particular do genoma humano, constituiu um marco científico extremamente importante nas últimas décadas e veio abrir caminho a novas áreas de investigação como a genómica e a proteómica. Espera-se que os avanços de conhecimento introduzidos por estas áreas tragam novas perspectivas sobre a forma como são diagnosticadas e tratadas muitas das doenças actuais, nomeadamente as que têm uma associação clara com disfunções ao nível do genotipo.

Neste artigo, apresentamos uma aplicação computacional que permite estudar associações anormais de codões em orfeomas, i.e. em sequências responsáveis pela construção de proteínas. Os resultados biológicos já obtidos mostram claramente a utilidade prática do software desenvolvido, que é disponibilizado de uma forma pública para a comunidade científica.

**Palavras-chave:** Bioinformática; Data Warehouse; Genómica.

## 1. Introdução

A sequenciação e anotação de genomas tem sido das áreas de investigação na biologia molecular em que mais se tem investido nos últimos anos. As bases de dados de genomas crescem diariamente, dando origem, como em muitas outras áreas, ao mesmo problema: perante tantos dados, como extrair informação relevante? Para responder a essa questão são necessárias ferramentas de bioinformática e de bioestatística cada vez mais eficazes e também cada vez mais específicas. Implícita está também a utilização de técnicas de descoberta de conhecimento em base de dados, nomeadamente com recurso a mineração de

dados. Assim, cabe às aplicações informáticas resolverem parte do problema. Por fazer fica ainda muita trabalho de análise e interpretação dos resultados.

Os genomas são originalmente representados em ficheiros de texto, no formato FASTA, onde cada oligonucleotido (base do DNA) é representado por uma letra (A – Adenina, C – Citosina, T – Timina e G – Guanina). O genoma é por sua vez subdividido em genes que representam regiões que podem expressar proteínas. Uma *Open Reading Frame* é uma sub-região do gene que é sujeita ao processo de tradução. Existem regras biologicamente estabelecidas que nos indicam se um gene, constituído por centenas ou milhares de bases é ou não válido, ajudando a confirmar a validade dos resultados experimentais de sequenciação.

A associação de cada três bases constitui um codão sendo cada codão do orfeoma traduzido para um aminoácido que é o elemento base da proteína. Existem 64 combinações diferentes de codões ( $4^3$ ) existindo somente 20 aminoácidos pelo que existe redundância do código genético, ou seja um mesmo aminoácido pode ser traduzido por diferentes codões. Isto levanta a questão de saber se existem codões preferenciais no processo de tradução.

A tradução dos genes em proteínas é realizada através de um mecanismo biológico designado por ribossoma. As consequências de eventuais erros de tradução são inúmeras, algumas com pouco impacto ou mesmo sem efeito no organismo, outras mais graves tais como o envelhecimento precoce, diversos tipos de cancro ou doenças raras.

No últimos anos temos vindo a estudar com sucesso as associações estatísticas entre pares de codões (Moura G., et. al. 2005). Na sequência destes trabalhos, e tendo em conta que o ribossoma se liga sequencialmente a 3 codões, a análise de associações entre tripletos surgiu da necessidade de responder a novas questões, nomeadamente associadas à evolução das espécies.

Neste contexto, e perante a existência de um grande número de genomas já descodificados, entre os quais o humano, desenvolvemos uma ferramenta de software que permite o estudo das relações estatísticas entre codões consecutivos, nomeadamente tripletos. Esta aplicação, denominada por GeneSplit, é composta por dois módulos independentes, mas que se complementam: *GScore* – a parte de *backoffice*, que permite o processamento de genomas isoladamente ou em larga escala; *GWeb* – a componente de disponibilização on-line das ferramentas para extrair a informação nas bases de dados produzidas pelo *GScore*, com a possibilidade de importação total ou parcial dos dados, em diversos formatos.

## **2. Metodologia**

A aplicação utiliza os orfeomas (parte codificante dos genomas) que são disponibilizados em bases de dados públicas. Estas sequências são pré-analisadas,

sendo ignorados os genes que não verificam as condições de validade. Para esse efeito, basta que ocorra no gene uma das seguintes condições de rejeição:

- Não iniciar por ATG;
- Comprimento não múltiplo de 3;
- Não terminar com TAA ou TAG ou TGA;
- Conter TAA ou TAG ou TGA sem ser no fim do gene;
- Conter nucleótidos desconhecidos, indicados pela letra N.

O algoritmo de contagens de tripletos de codões e de respectivos aminoácidos para cada organismo, efectua essa filtragem, sendo exibido no final o número de genes contados, o número de genes considerados e o nº de genes que foram ignorados, mostrando nesse caso, quais as condições pelas quais foram excluídos. É contemplada ainda a degeneração do código genético para alguns organismos. Por exemplo em *Candida Albicans* e *Debaryomyces Hansenii* o codão CTG, que normalmente codifica o aminoácido Leucina, nestes organismos codifica o aminoácido Serina (Santos M.A., et. al., 1997). Este tipo de variação está previsto no processamento de acordo com as evidências científicas correntes. A representação dos tripletos de codões é obtida pelas três posições em que cada codão aparece, pelo organismo de origem e pela contagem de ocorrências do tripleto. Os codões do início foram excluídos da contagem, iniciando-se esta no segundo codão e terminando no penúltimo codão, sendo portanto ignoradas as contagens em que estão incluídos os codões de finalização.

Durante o processo são criadas várias matrizes tri-dimensionais, cuja dimensão individual é dada por  $61 \times 61 \times 61$  (são ignorados os codões de terminação da região codificante), sendo usadas para armazenar os dados resultantes nas contagens  $cod(i,j,k)$ , onde  $i, j, k$  representam os codões que se encontram na 1ª, 2ª, e 3ª posição respectivamente, sendo o valor armazenado o nº de vezes que um determinado tripleto aparece no orfeoma (Figura 1). Analogamente, para armazenar as contagens de aminoácidos, é criada uma matriz tridimensional cuja dimensão é dada por  $20 \times 20 \times 20$ .

De forma a facilitar o processo de contagens, recorre-se à programação dinâmica, sendo criados dois arrays contendo um todos os codões (64), e outro todos os aminoácidos (20), assim como um terceiro array contendo os aminoácidos, nas posições dos respectivos codões.

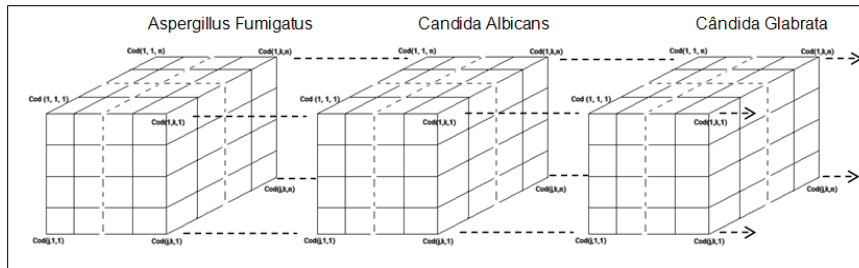


Figura 1 – Representação matricial (4-D Cubo) dos dados resultantes das contagens de tripletos de códons

Uma vez armazenados os resultados das contagens, o sistema incorpora várias consultas de pós-processamento dos dados, para que posteriormente possam ser aplicados em software de análise de dados (Figura 2).

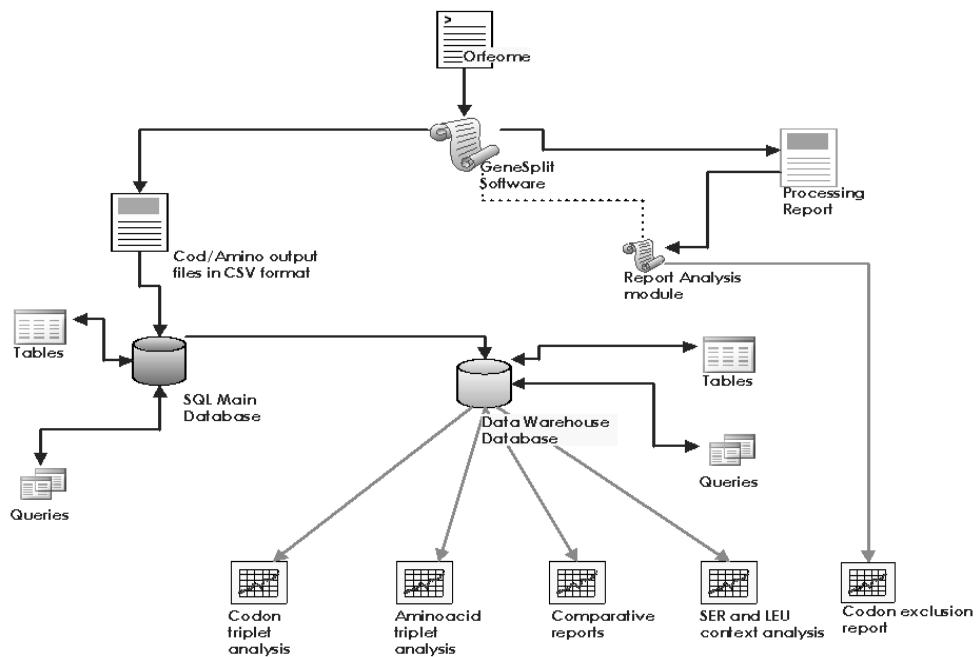


Figura 2 – Workflow do projecto, desde o processamento do orfeoma até à disponibilização de resultados.

### 3. Algoritmos

#### 3.1. Algoritmo Contagem Global de Tripletos

O algoritmo para contagem de tripletos de codões, inicia-se com a leitura gene a gene, sendo aplicados os critérios de filtragem referidos anteriormente. Se o gene lido for considerado válido, é separado em tripletos, sendo criado um array contendo todos os codões do gene. A contagem inicia-se no 2º codão, sendo assim ignorados o 1º e o último codões.

Simulando o processo de tradução do ribossoma, o apontador do codão inicial posiciona-se na 4ª posição do array de codões e identifica os codões anteriores adjacentes à posição n, ou seja posições n-2 e n-1, sendo armazenado o respectivo tripleto no cubo respectivo de valor acumulados. Em seguida o apontador desloca-se um codão e o processo repete-se.

O processo de contagem de aminoácidos é análogo, utilizando para o efeito a matriz de contagem dos tripletos de aminoácidos.

As contagens são efectuadas em duas passagens e com dois resultados distintos. Na segunda contagem são ignoradas as cadeias contendo codões iguais em número superior ou igual a 4, sendo apenas contada uma ocorrência sempre que estas repetições de cadeias longas se verificarem. O objectivo é permitir também o estudo sem o enviesamento causado por estas sequências.

O algoritmo aplicado para efectuar as contagens é, de uma forma simplificada, o seguinte:

```
OPEN sourcefile
WHILE NOT Sourcefile.EOF
  gene=Sourcefile.Readgene()
  IF ClearGene(gene) THEN
    WITH gene DO
      Codgene=SPLIT(gene,3)
      FOR i=4 to UBOUND(Codgene)-1
        xcod=poscodon(Codgene(i-3))
        ycod=poscodon(Codgene(i-2))
        zcod=poscodon(Codgene(i-1))
        matrix(xcod, ycod, zcod) = matrix(xcod, ycod, zcod) + 1
      END FOR
    END WITH
  ELSE
    PRINT TO FILE "Excluded..."
  END IF
END WHILE
PRINT TO FILE DataMatrix()
```

A partir daqui, desencadeia-se todo o processo de análise estatística e probabilística.

Neste artigo apresentamos uma aplicação de software que foi desenvolvida especificamente para análise de associação entre tripletos de códons consecutivos. Este trabalho permitiu já a obtenção de alguns resultados científicos nomeadamente através da análise comparativa de 11 genomas, foi possível detectar um padrão particular da espécie “*Candida Albicans*” relativamente a outras bactérias (Moura, G., et al., 2007).

## **8. Referências**

- Bertrand, C. et al. (2002) Influence of the stacking potential of the base 3' of tandem shift codons on -1 ribosomal frameshifting used for gene expression, *Rna*, 8, 16-28.
- Dong, H., et al. (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates, *J Mol Biol*, 260, 649-663.
- Irwin, B., et al. (1995) Codon pair utilization biases influence translational elongation step times, *J Biol Chem*, 270, 22801-22806.
- Korostelev, A., et al. (2006) Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements, *Cell*, 126, 1065-1077.
- Moura, G., et al. (2007) Codon-triplet context unveils unique features of the *Candida albicans* protein coding genome, *BMC Genomics*, 8:444.
- Moura, G., et al. (2005) Comparative context analysis of codon pairs on an ORFeome scale, *Genome Biology*, 6:R28.
- Nierhaus, K.H. (2006) Decoding errors and the involvement of the E-site, *Biochimie*, 88, 1013-1019.
- Santos, M.A., et al. (1997) The non-standard genetic code of *Candida* spp.: an evolving genetic code or a novel mechanism for adaptation?, *Molecular Microbiology* 26 (03) , 423–431
- Yarus, M. et al., (2005) Origins of the Genetic Code: The Escaped Triplet Theory, *Annual Review of Biochemistry*, 74, 179-198