Sistema bioinformático para análise do mecanismo molecular de descodificação da informação genética

Miguel Pinheiro, Gaspar Dias, Vera Afreixo[1], Adelaide Valente[1],
Gabriela Moura[2], Manuel A. S. Santos[2], José Luís Oliveira

[1]Departamento de Matemática, [2]Departamento de Biologia
Universidade de Aveiro, 3810-193 Aveiro.

***Resumo* – A descodificação do genoma humano e de outros organismos veio oferecer uma oportunidade única para elucidar o funcionamento e evolução dos seres vivos. Contudo, a enorme quantidade de informação contida nos genomas criou novos desafios e questões cuja resolução requer o desenvolvimento de metodologias matemáticas e ferramentas bioinformáticas capazes de lidar com grandes volumes de informação.**
Este artigo descreve uma aplicação de software baseada em modelos matemáticos, estatísticos e visuais para processamento e análise da informação contida nos genomas.

***Abstract* – The decoding of the genome of humans and of other organisms opened a unique opportunity for understanding genome function, structure and evolution. However, the enormous volume of information contained in the genome created new challenges and questions whose resolution requires new mathematical models and bioinformatics tools able to deal with large volumes of information.**
In here, we describe a software application based on mathematical, statistical and visual models, which offers a set of tools for genetic information processing and analysis.

I. INTRODUÇÃO

Toda a informação necessária à perpetuação, evolução e funcionamento da vida está contida no genoma [1]. Este é constituído por longas cadeias de 4 bases químicas, designadas, por questões de simplificação, pelas letras A, C, T e G. O século XX, trouxe-nos a descoberta da estrutura tridimensional do DNA (uma dupla hélice), a sequenciação dos genomas, a elucidação da sua organização em regiões codificantes (genes) e não codificantes e, ainda, a identificação dos mecanismos que asseguram o fluxo de informação genética do gene à proteína (Figura 1).
O fluxo de informação genética é assegurado por um conjunto de motores moleculares que dão "vida" ao genoma. Isto é, o genoma é o repositório da informação que comanda todas as funções vitais, contudo esta só é útil se for lida (descodificada) e usada para sintetizar as proteínas e outras biomoléculas que asseguram o funcionamento celular. A descodificação do genoma assume, assim, um papel central no funcionamento da vida, sendo, por esta razão, objecto de intenso estudo.

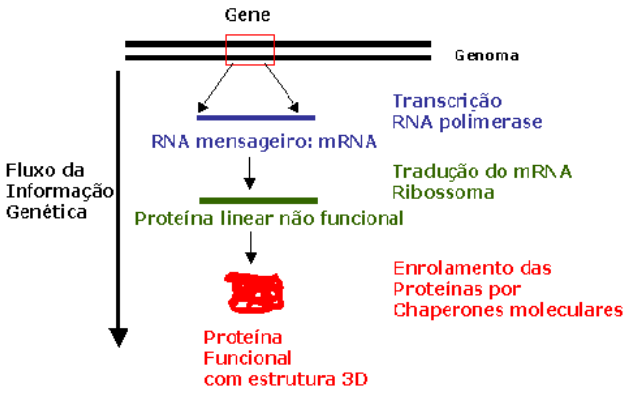

Figura 1 - Os genomas contêm toda a informação genética necessária ao funcionamento dos seres vivos. Contudo, a informação neles contida só é útil se for descodificada. Para tal, existem descodificadores moleculares que convertem a informação contida nos genes em proteínas. Estas são as moléculas que fazem todo o trabalho necessário à manutenção da vida. O diagrama ilustra de um modo simples o fluxo de informação do gene à proteína.

Um dos aspectos mais fascinantes da descodificação do genoma é a enorme velocidade e fidelidade (10^{-5} a 10^{-6} erros por letra descodificada) da maquinaria molecular (RNA polimerases) de leitura dos genes e, também, da síntese de proteínas (ribossomas) que o fazem a uma velocidade de 8 amino ácidos por segundo com um erro de 10^{-4} a 10^{-5}. O mesmo acontece com os motores moleculares (DNA polimerases) de replicação (duplicação) dos genomas cujo erro é de 10^{-6} a 10^{-7} à velocidade de leitura de 1000 letras s^{-1}.
Curiosamente, este baixo erro de replicação e descodificação dos genomas tem duas faces aparentemente contraditórias. Por um lado, é fundamental para a evolução de novas espécies, criando diversidade genética. Por outro, é a causa do envelhecimento dos seres vivos e de enumeras doenças. Um ser vivo cujo erro de replicação e descodificação do seu genoma fosse zero, seria incapaz de evoluir, criando um colapso evolutivo que impediria a evolução de novas espécies. Note-se que a perda de controlo da fidelidade de leitura dos genomas é catastrófico originando, nos seres humanos, cancros e doenças graves.
A importância do erro associado à descodificação da informação genética levou-nos a desenvolver um sistema bioinformático para identificar as leis gerais que governam a fidelidade de descodificação dos genomas ao nível da

síntese das proteínas que ocorre no ribossoma, isto é, durante a descodificação do mRNA (Figura 1). Foram desenvolvidas metodologias matemáticas e algoritmos para análise de grandes volumes de informação genética e interfaces para visualização da informação que são descritas neste artigo.

II. O CÓDIGO GENÉTICO

Tal como qualquer código matemático, informático ou de linguagem, o código genético é constituído por um conjunto de regras que asseguram a conversão de um tipo de informação noutro tipo de informação. Assim, o código genético define as regras de conversão da informação contida nos genes, que é escrita com base num alfabeto de quatro letras (A, C, T, G), na informação contida nas proteínas, que é baseada num alfabeto de 22 letras (amino ácidos). Ou seja, o código genético estabelece a relação entre as bases do DNA e os amino ácidos das proteínas utilizando as regras definidas na (Figura 2).

UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Figura 2 - O código genético define as regras de atribuição dos 61 tripletos das bases A, U, G, e C aos 22 amino ácidos que constituem as proteínas. A tradução destes tripletos contidos na sequência linear do mRNA, que resulta da cópia de uma das cadeias dos genes, é feita por uma máquina molecular (descodificador) chamada ribossoma.

A existência deste código levanta duas questões fundamentais: Como é que é feita a conversão da informação de um alfabeto de 4 letras num alfabeto de 22 letras?, e, qual o erro associado à transferência de informação entre os dois alfabetos?

III. O MECANISMO DE DESCODIFICAÇÃO DO CÓDIGO GENÉTICO.

Os genomas contêm uma parte codificante e outra não codificante (cuja utilidade se desconhece), contudo são ambas escritas com base no mesmo alfabeto de 4 letras. Assim, a organização das letras na parte codificante (genes) deverá ser diferente da parte não codificante. Por exemplo, sequências específicas de letras definem o inicio e o fim dos genes, a sua estrutura primária e os sinais que controlam o seu funcionamento.

Quando um gene é activado, as RNA polimerases identificam o gene, lêem a sua sequência de letras e utilizando uma das cadeias da dupla hélice do DNA como molde sintetizam uma molécula de RNA chamada RNA mensageiro (mRNA) - O RNA é uma molécula quimicamente semelhante ao DNA sendo a única diferença a troca da letra T (timidina) por U (uracilo) no RNA – O mRNA tem como função transportar a informação contida no gene para ribossoma (Figura 1). Este passo intermédio do fluxo de informação genética é necessário porque o ribossoma é incapaz de ler os genes directamente no genoma.

A organização diferencial das 4 letras do alfabeto genético torna-se evidente na fase-2 do fluxo de informação genética, isto é, durante a leitura do mRNA pelo ribossoma que é o centro de descodificação que transforma o alfabeto de 4 letras (A, U, C e G) do mRNA no alfabeto de 22 amino ácidos. O ribossoma lê conjuntos de 3 letras e não uma letra de cada vez. Ou seja, o código genético é organizado em arranjos das 4 letras 3 a 3. Por exemplo, a sequência do mRNA AACGGCCCACUG é lida como **AAC-GGC-CCA-CUG,** sendo que o tripleto AAC codifica o amino ácido asparagina, GGC glicina, CCA prolina e CUG leucina (Figura 2). Assim, existem 64 arranjos diferentes das 4 letras, sendo 3 dos tripletos utilizados como sinais de terminação e 1 como sinal de iniciação da síntese proteica. Esta organização do código genético implica que ele é redundante, ou seja há vários tripletos para cada amino ácido porque há 64 tripletos diferentes e apenas 22 amino ácidos. A razão desta discrepância não é conhecida e constitui um dos grandes mistérios da biologia e da origem da vida.

A leitura dos tripletos de bases (codões) é feita no ribossoma por um adaptador molecular chamado RNA de transferência (tRNA). Este é um pequeno RNA que tem duas propriedades fundamentais: i) numa das extremidades tem uma sequência de letras complementar da do codão do mRNA chamada anticodão – a complementaridade das letras é estabelecida pelas propriedades químicas das bases que permite o emparelhamento específico entre as letras A e T(U) e G e C – e ii) na outra extremidade é carregado com amino ácidos por um grupo de enzimas chamadas aminoacil-tRNA sintetases (Figura 3).

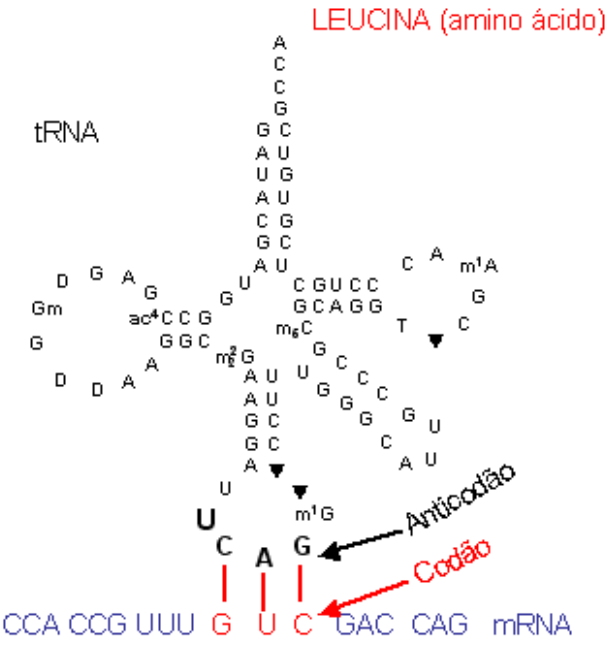

Figura 3 - O código genético é estabelecido através de um adaptador molecular chamado RNA de transferência (tRNA). Este não é mais do que uma cabeça de leitura dos tripleto de bases existentes no mRNA (codões). A sua capacidade de estabelecer o código genético deve-se ao facto de numa das suas extremidades ser carregado com amino ácidos e na outra ter um tripleto de bases chamado anticodão que é complementar ao tripleto de bases do mRNA (codão). Nos seres vivos existe pelo menos um tRNA para cada amino ácido, ou seja, 22 tRNAs diferentes. A interacção codão anticodão ilustrada na figura ocorre no ribossoma (ver texto).

O tRNA é de facto uma cabeça de leitura da informação contida no mRNA que também transporta os amino ácidos para o ribossoma, tendo este a função de os ligar formando as proteínas que são cadeias longas de amino ácidos (Figura 4).

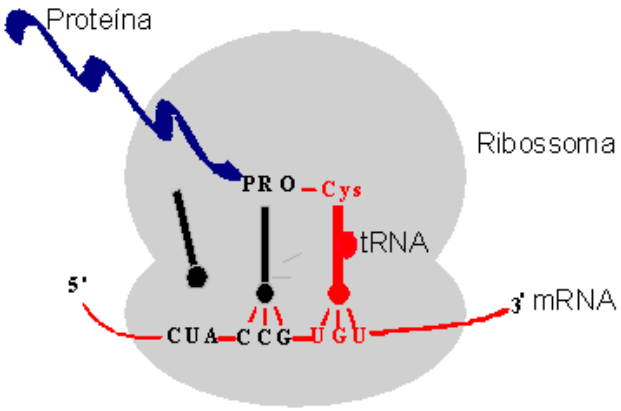

Figura 4 - Esquema geral da síntese de proteínas no ribossoma evidenciando a descodificação dos codões do mRNA pelos anticodões do tRNA. O ribossoma é o centro de descodificação e o centro de síntese das proteínas mas é o tRNA que estabelece a relação entre o código de 4 letras do mRNA e o código de 22 amino ácidos das proteínas.

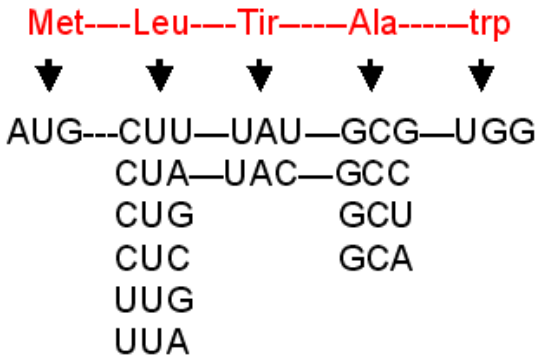

Figura 5 - O código genético é degenerado e redundante. Isto é, alguns amino ácidos são codificados por 6 tripletos de bases, outros apenas por 4, 2 ou apenas 1 tripleto de bases. Teoricamente o código genético poderá ser expandido de 22 para 63 amino ácidos, contudo issso não acontece nos seres vivos e não se sabe qual a razão biológica para a existência de apenas 22 amino ácidos.

Um tRNA específico para o amino ácido alanina deverá ser capaz de ler os 4 codões que codificam alanina (GCG, GCC, GCU e GCA), ou alternativamente deverá haver mais do que 1 tRNA para o amino ácido alanina (Figura 5). As duas situações existem na natureza, ou seja, há organismos que têm 4 tRNAs para alanina (1 anticodão para cada um dos 4 codões) e outros que utilizam 1, 2 ou 3 tRNAs. A redundância do código genético e a existência de um número variável de tRNAs em diferentes organismos levanta duas questões pertinentes, a saber: será que os diferentes codões de um determinado amino ácido são utilizados indiscriminadamente nos genes? Ou há restrições à sua utilização?.

A análise da frequência dos codões nos genes mostra que cada organismo (genoma) favorece a utilização de determinados codões e reprime a utilização de outros. Por outro lado, codões utilizados frequentemente são descodificados por tRNAs cuja concentração celular é elevada, sugerindo a existência de uma relação entre a frequência de utilização dos codões nos genes e a abundância dos tRNAs que os descodificam. Para além desta restrição, a análise dos codões vizinhos de um determinado codão (contexto do codão) sugere que os codões influenciam os seus vizinhos a montante e a jusante. Estas observações levantam a hipótese de que o contexto dos codões influencia a velocidade e fidelidade de descodificação do mRNA pelos anticodões do tRNA no ribossoma e sugere que a análise do contexto dos codões à escala genómica poderá evidenciar leis gerais que governam a fidelidade de descodificação do código genético.

IV. Objectivos

Este projecto teve como objectivo principal desenvolver metodologias e ferramentas informáticas para o estudo do contexto dos codões à escala genómica. O sistema bioinformático desenvolvido lê todos os genes de um genoma (independentemente do seu número total), e simula a leitura dos codões pelos tRNAs no ribossoma. Ao

faze-lo memoriza os codões vizinhos e constrói uma tabela de frequências de contexto que pode ser tratada estatisticamente. O sistema constrói automaticamente várias tabelas de contingência, calculando à posteriori os valores residuais das mesmas. Estes são visíveis através de vários níveis de cores, atribuindo uma cor a um intervalo do valor residual.

A informação (sequências de letras) de cada genoma é organizada de acordo com a estrutura dos próprios genomas (Figura 6). Ou seja, do genoma ao gene, sendo o genoma constituído por cromossomas e estes por genes (as sequencias não codificantes não são analisadas).

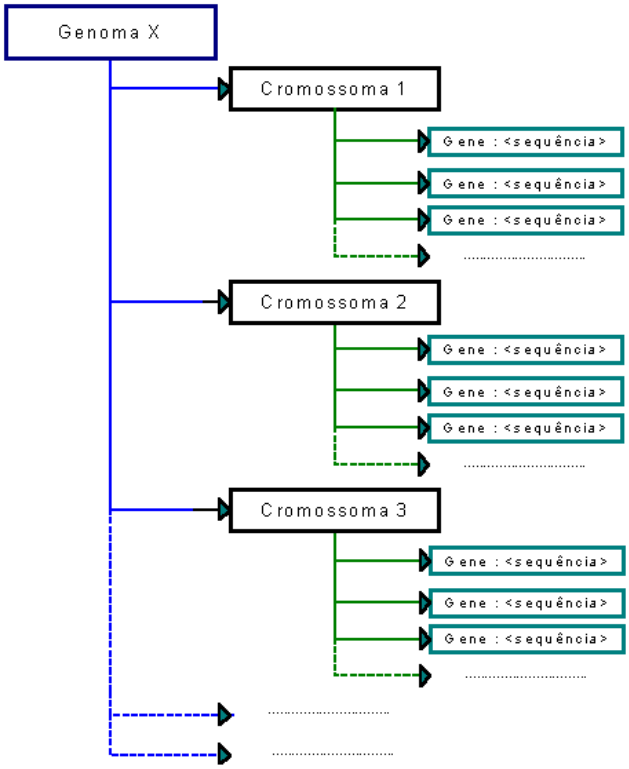

Figura 6 - Organização dos genes e cromossomas.

As sequências dos genomas já sequenciados encontram-se depositadas em bases de dados internacionais públicas em diversos formatos. O formato FASTA, um dos mais populares, é o único formato aceite pelo sistema que desenvolvemos. É um simples ficheiro texto contendo o nomes dos genes e a sua sequência de letras.

O sistema permite a leitura de vários genomas em simultâneo e a navegação pelos mesmos, assente no paradigma do explorador do Windows. Permite também a comparação dos genomas através das tabelas residuais.

Na visualização dos genes têm-se a possibilidade de fazer corresponder as cores da tabela dos valores residuais ao fundo dos codões correspondentes. Existe também a possibilidade de visualizar os codões raros (um codão diz-se raro quando a sua existência num genoma é inferior a cinco por mil), entre outros dados considerados importantes para análise biológica. É possível, ainda, procurar padrões específicos de organização das letras definidos pelo utilizador.

V. METODOLOGIAS MATEMÁTICAS

A análise do contexto dos codões faz-se através de tabelas de contingência [2] com dimensões de 64x64, dado que existem 64 codões diferentes. A quantificação da frequência de utilização dos codões no genoma faz-se preenchendo a tabela de contingência de acordo com o esquema apresentado na Figura 7.

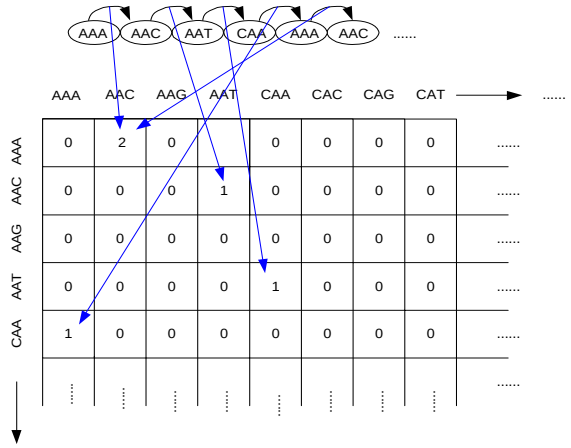

Figura 7 - Quantificação da frequência de utilização dos codões. Na parte superior indica-se uma sequência de um gene hipotético. Fixando um codão faz-se corresponder o mesmo a uma linha da tabela e o seguinte a uma coluna da tabela.

Por convenção, fixando um codão qualquer no mRNA o codão a jusante diz-se que está a 5' e o codão a montante a 3´ daquele. Assim, podemos ter tabelas de contexto a 5´ e a sua transposta a 3´.

Face à rejeição da independência através da aplicação do teste do qui-quadrado para tabelas de contingência e com o propósito de identificar responsáveis pelo grande valor das estatísticas de teste, efectuamos uma análise de resíduos, utilizando os valores residuais sugeridos por Haberman (1973) [3].

O método envolve cálculo dos resíduos normalizados, e_{ij}, dado por:

$$e_{ij} = \frac{(n_{ij} - E_{ij})}{\sqrt{E_{ij}}}$$

onde E_{ij}

$$E_{ij} = \frac{n_{i.} n_{.j}}{N}$$

O $n_{i.}$, $n_{.j}$, e N corresponde o total da linha i, total da coluna j, e total de todas as linhas correspondentemente.

A variância de e_{ij} é dada por:

$$v_{ij} = \left(1 - \frac{n_{i.}}{N}\right)\left(1 - \frac{n_{.j}}{N}\right)$$

Cada célula da tabela de contingência irá conter o valor residual ajustado, d_{ij}, dado por:

$$d_{ij} = \frac{e_{ij}}{\sqrt{v_{ij}}}$$

Quando as células da tabela de contingência são independentes, obedecem aproximadamente a uma distribuição normal com média zero e desvio padrão 1 [2].

Tendo encontrado os valores residuais tivemos necessidade de agrupar as linhas e colunas recorrendo a metodologias da análise classificatória, com intuito de encontrar padrões nas tabelas de contingência.

Estes padrões são detectados calculando as semelhanças, a partir dos coeficiente de Pearson, entre cada dois vectores, linha com linha ou coluna com coluna, da tabela de contingência.

Dados dois conjuntos X = {x1, x2,..,xn} e Y = {y1, y2, ...yn} e considerando que todos os elementos dos vectores têm o mesmo peso, o coeficiente de correlação de Pearson centrado r é definido por:

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

onde $\overline{X}$ e $\overline{Y}$ representam as médias de X e Y, e σ_x e σ_y representam os desvios padrão de X e Y respectivamente.

Os coeficientes de Pearson estão compreendidos no intervalo de 1 a –1, para r igual a 0 os vectores não estão associados e para |r| igual a 1 estão perfeitamente associados em que o sinal indica o tipo de direcção de associação. A título de exemplo dados dois vectores um múltiplo do outro o coeficiente de correlação é 1.

Também é usada como medida de semelhança o coeficiente de correlação não centrado, dado por:

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

Esta correlação é semelhante à centrada, assumindo que a média do vector é 0 mesmo quando não o é. Uma das diferenças surge quando temos dois vectores com a mesma forma mas tendo "offset", a correlação de *Pearson* (centrada) que seria de um enquanto a correlação não centrada será inferior a um.

Calculada a medida de semelhança entre indivíduos, o coeficiente de correlação, escolhe-se a ligação simples como critério de agregação entre grupos [4].

VI. Descrição da Aplicação

A aplicação tinha como principal requisito ser "*user friendly*". Para tal apoiámo-nos no paradigma do Explorador do Windows, apresentando uma árvore de navegação no lado esquerdo, pelos vários genomas admitidos, tendo acesso aos seus cromossomas. Existe no entanto algum abuso nesta comparação, fazendo crer que cada ficheiro contendo um certo número de genes corresponde a um cromossoma, e um certo número de ficheiros lidos simultaneamente corresponde a um genoma. Isto depende do conteúdo dos ficheiros analisados que nem sempre respeitam esta estrutura. No entanto, esta restrição limita apenas a abrangência do estudo. Os resultados das análises realizadas, são apresentados no lado direito.

Como se pode ver na (Figura 8) encontra-se a janela principal do software Anaconda (Análise do Contexto de Codões), contendo vários cromossomas (ficheiros) admitidos e com a visualização de uma matriz 64x64 codões. Cada cor da matriz representa um determinado intervalo nos valores residuais. Quanto mais negativo são os valores residuais mais vermelho (■) é a sua representação, ocorrendo o mesmo no verde (■) mas para valores residuais positivos.

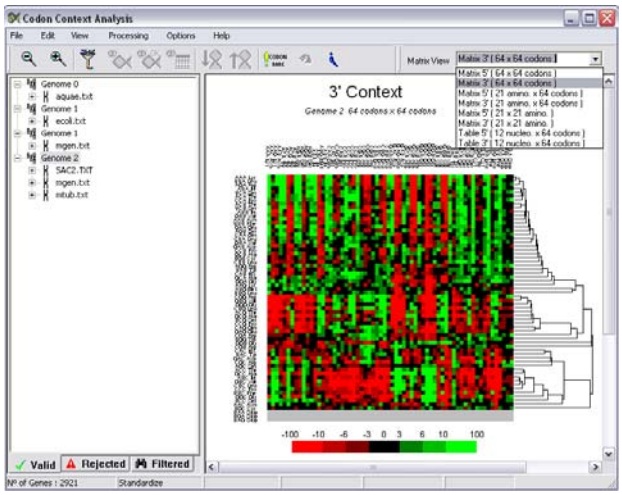

Figura 8 – Janela principal do software.

O software disponibiliza uma ferramenta que possibilita a visualização de uma histograma que corresponde à ocorrência dos valores residuais de uma matriz 64x64 (Figura 9). Assim o utilizador poderá definir os intervalos para cada cor, ou mesmo alterar a cor.

Recorrendo a uma *Combo Box* situada na barra de ferramentas poderemos escolher qual a matriz, com os valores residuais, que se pretende visualizar, entre as especificadas pelos requisitos.

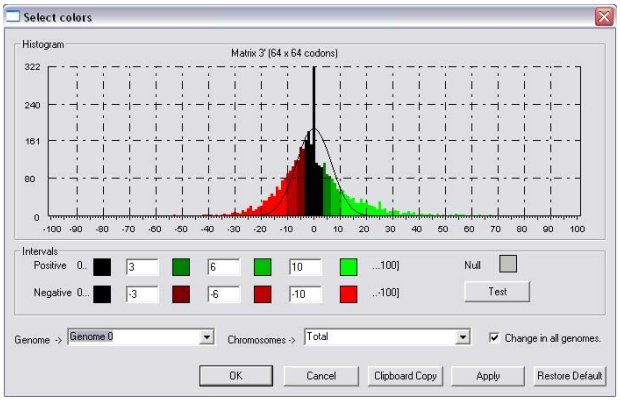

Figura 9 – Histograma da matriz (64x64 codões).

Os ficheiros que correspondem aos cromossomas têm que obedecer ao formato FASTA, estando disponíveis nas bases de dados internacionais, normalmente com a extensão ".ffn". Para agrupá-los num só genoma, os ficheiros terão que ser abertos em simultâneo. Aquando da abertura dos respectivos ficheiros tem-se a possibilidade de impor certos requisitos aos genes, como se pode ver pela (Figura 10). Permite-nos também a quantificação ou não das sequências admitidas. Se não se quantificarem as sequências, poderemos à posteriori redireccionar as tabelas de valores residuais pertencentes a outros genomas, para analisar como se comportam determinadas sequências genómicas quando inseridas noutros genomas.

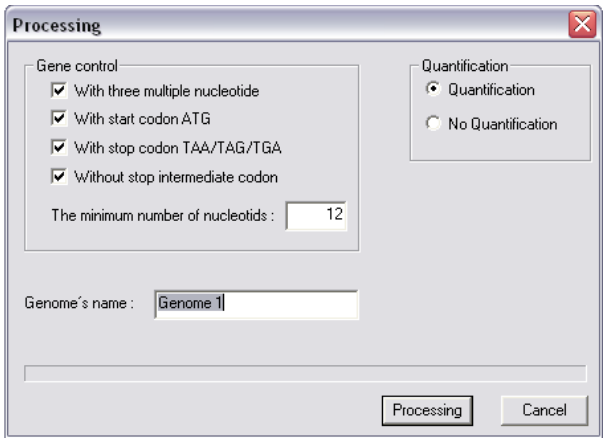

Figura 10 – Janela de controlo na admissão dos genes.

Na parte esquerda da área central da janela principal temos um "*Tab Control*", no qual estão incluídos três *Tab´s*, os quais contém:

- *Valid* : árvore onde ficam representados os genes, que preenchem os requisitos de entrada, ficando agrupados aos cromossomas a que pertencem.
- *Rejected* : genes que não preenchem os requisitos de entrada impostos pelo utilizador.
- *Filtered* : genes que foram filtrados por uma ferramenta, que impõem certos requisitos aos genes. Esta ferramenta pode ser parametrizada pelo utilizador, baseada nas tabelas residuais.

Na (Figura 11) pode-se ver a sequência genómica de um gene pertencente a um determinado cromossoma. Na sua visualização fazem-se corresponder os valores residuais da matriz, nos codões correspondentes. Por exemplo, se o codão AAA com a codão CAT na matriz dos valores residuais, corresponder o valor -30.05, significa que o codão AAA aparece com o fundo vermelho, aquando da sua visualização no gene. Esta situação indica que o codão AAA tem pouca afinidade pelo codão CAT. Se o codão CAT com o codão CCA corresponder o valor 10.0, o codão CAT aparece com o fundo em verde. Esta situação indica que esta sequência é preferida em relação às combinações com valores residuais negativos. A combinação é tanto mais aceite quanto maior for o seu valor residual.

Os codões raros são apresentados com uma elipse azul à sua volta.

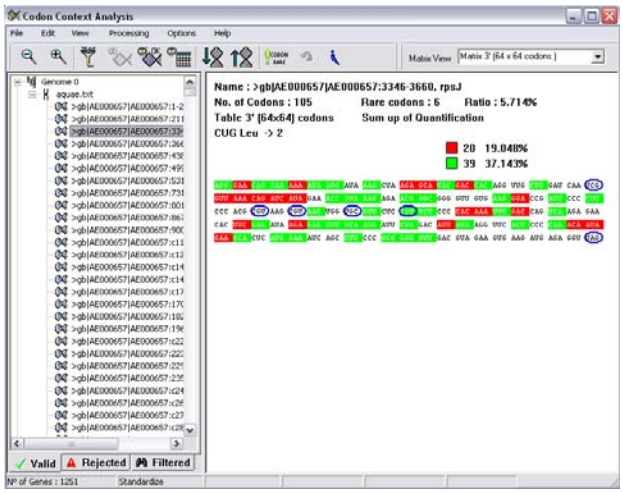

Figura 11 – Gene seleccionado.

Temos também a possibilidade de ver informação variada do gene em questão, como o número de codões raros, o seu rácio relativo ao número total de codões, a percentagem de codões "vermelhos", a percentagem de codões "verdes", e o número de ocorrências de um determinado codão no gene corrente, definindo as possibilidades de visualização através da caixa de dialogo.

Numa outra caixa de diálogo, quando o gene se encontra seleccionado, pode-se obter várias informações do mesmo, tais como, percentagem de codões com G e C na terceira posição, índice *Codon Adaptation Index* (CAI) [5], número efectivo de codões, entre outras informações.

O sistema permite o acesso a uma ferramenta de procura de padrões de cores nos genes de um determinado genoma seleccionado pelo utilizador (Figura 12). Movendo os *slider bars* podemos alterar o padrão a procurar, podendo mesmo obter um "*don´t care*" se desactivarmos esse codão. A mesma ferramenta permite também o acesso a outros filtros, tais como, procura padrões de codões raros,

sequências genónimas, percentagens de codões raros nos genes e percentagem de certos valores residuais.

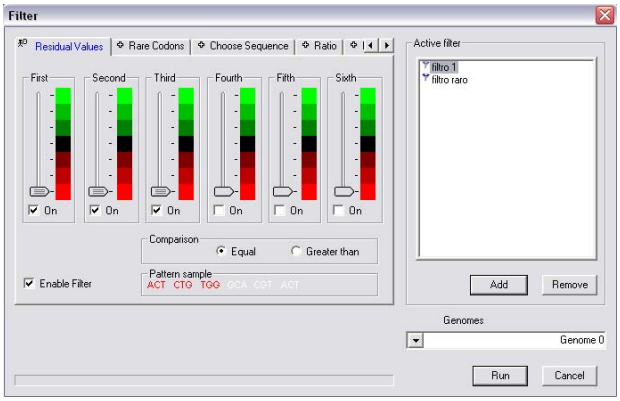

Figura 12 – Ferramenta para filtrar os genes.

Existe também a possibilidade de gravar os filtros para sua posterior reutilização.

Os genes que correspondem ao filtro, serão carregados no *Tab Filtered* para posterior visualização.

Uma outra ferramenta permite o acesso a um histograma, contendo a ocorrência de cada codão num determinado genoma ou cromossoma (Figura 13). Permite também redefinir o nível que indica se um determinado codão é raro ou não.

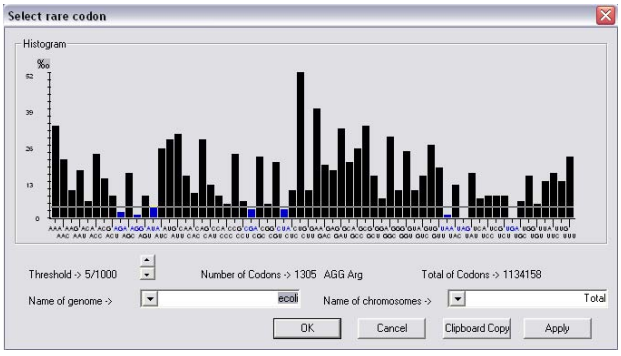

Figura 13 – Histograma de ocorrência de codões.

O software disponibiliza a análise *cluster* como ferramenta, podendo ser usada para agrupar linhas e colunas de uma determinada matriz, permitindo assim a formação de padrões para posterior análise.

Como podemos ver na (Figura 14) e na (Figura 15), temos duas imagens com e sem análise *cluster*.

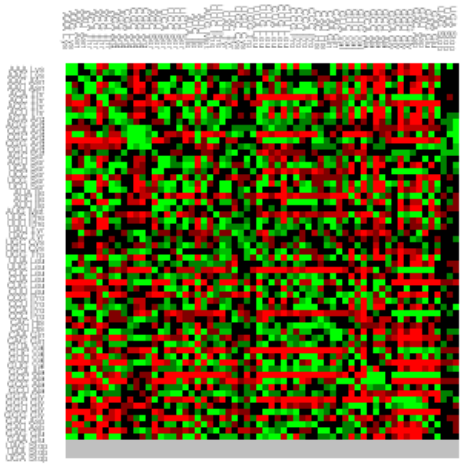

Figura 14 – Matriz 64x64 codões sem análise cluster.

Como se pode ver existem formações de padrões, zonas onde uma determinada cor é predominante (Figura 15). Isto revela que o índice linha tem ou não afinidade pelo índice coluna, consoante a cor seja verde ou vermelha respectivamente.

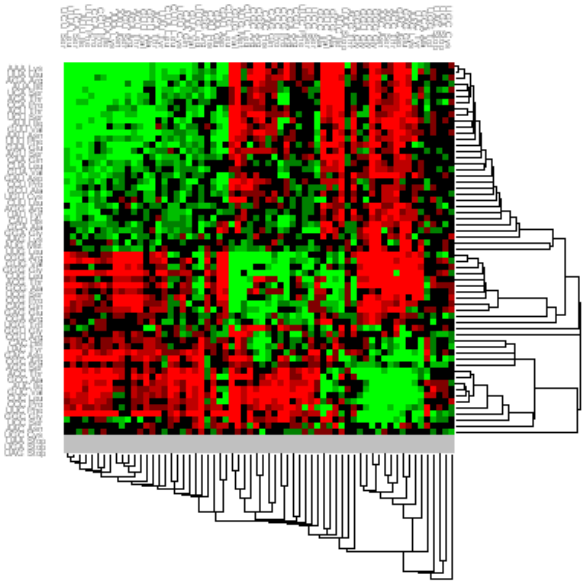

Figura 15 – Matriz 64x64 codões com análise cluster.

Através da ferramenta de análise *cluster* pode-se construir ainda árvores filogenéticas de todos os genomas carregados no software. A (Figura 16) mostra a árvore filogenética que corresponde aos genomas lidos pelo sistema . Os braços da árvore são proporcionais à distância que separa um genoma de outro.

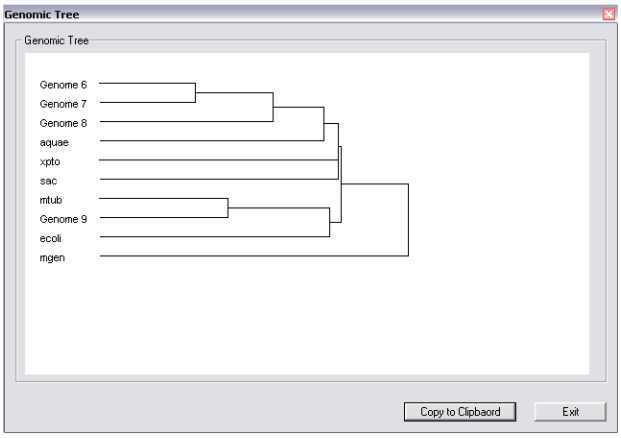

Figura 16 – Árvore filogenética.

O sistema permite também, através de uma caixa de diálogo, a possibilidade de o utilizador procurar certas sequências nos genes, tais como:

- Codões sequencialmente repetidos;
- Aminoácidos sequencialmente repetidos;
- Valores residuais sequencialmente repetidos;

Posteriormente a esta análise é possível a formação de gráficos, para codões ou aminoácidos sequencialmente repetidos.

O utilizador tem também a informação da quantidade de sequências que foram encontrados, e a sua posterior visualização, entre outras informações.

Uma das grandes potencialidades do sistema é a possibilidade de gerar histogramas com os índices CAI de um determinado genoma ou cromossoma (Figura 17).

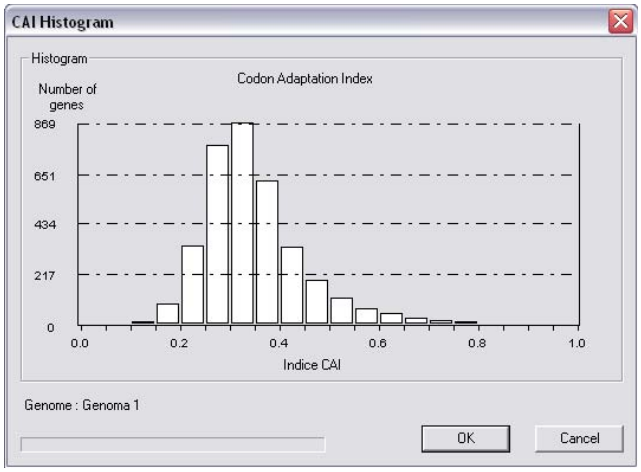

Figura 17 – Histograma CAI.

O investigador tem também à sua disposição uma ferramenta, com a qual poderá obter informação variada (Figura 18). Número de codões, número de aminoácidos, valores *Relative Synonymous Codon Usage* (RSCU)[1][5], entre outras informações, poderão ser obtidas e copiadas através do *clipboard*. Esta informação está disponível para genes, cromossomas e genomas.

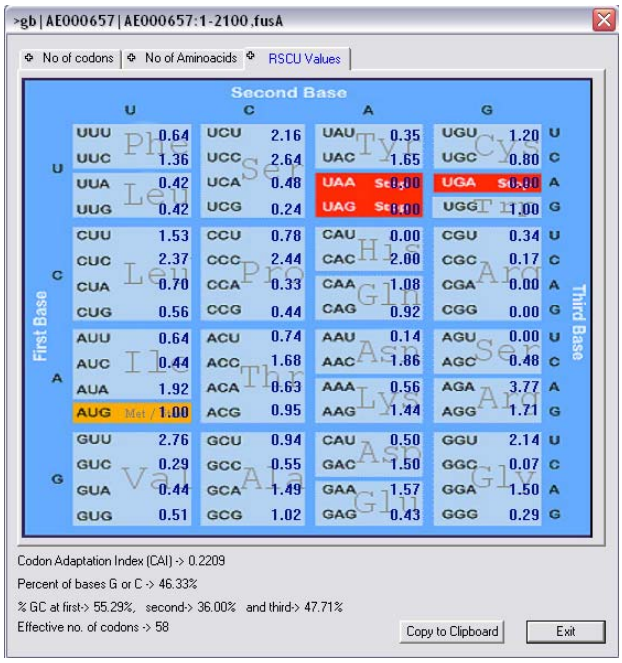

Figura 18 – Informação sobre um gene.

VII. CONCLUSÕES

O sistema desenvolvido apresenta as seguintes funcionalidades:

- Análise do contexto de codões.
- Capacidade de ler vários genomas em simultâneo.
- Capacidade de simular a leitura feita pelo ribossoma.
- Possibilidade de rejeitar genes que não correspondam às especificações desejadas.
- Cada cromossoma contém um grupo de oito tabelas residuais, sendo elas:
    1. matriz 3’ de 64 por 64 codões;
    2. matriz 5’ de 64 por 64 codões;
    3. matriz 3’ de 64 codões por 12 nucleótidos;
    4. matriz 5’ de 64 codões por 12 nucleótidos;
    5. mariz 3’ de 21 por 21 amino-ácidos;
    6. matriz 5’ de 21 por 21 amino-ácidos;
    7. matriz 3’ de 64 codões por 21 amino-ácidos;
    8. matriz 5’ de 64 codões por 21 amino-ácidos;

[1] RSCU é o número de vezes que determinado codão é observado relativamente ao numero de vezes que ele deveria ser observado.

- Atribuição de cores aos diferentes níveis da análise residual, e sua posterior visualização.
- Possível reagrupamento das linhas e colunas, para maior realce da dependência entre codões, através de análise classificatória.
- Possível visualização dos codões raros aquando da visualização dos genes.
- Possibilidade de copiar as matrizes com os valores residuais de um genoma ou cromossoma para um outro genoma "redireccionamento".
- Visualizar os genes, onde a incidência de erro de leitura é maior, através de filtros que o próprio utilizador pode parametrizar.
- Possibilidade de sobrepor as matrizes de dois genomas, de modo a realçar as suas maiores diferenças.
- Possibilidade de aperfeiçoar os genes, substituindo codões pelos seu correspondentes, a nível de aminoácidos, onde a incidência de erro de leitura nas sequências do mRNA é superior à média, diminuindo também os codões raros no gene.

O software desenvolvido disponibiliza um conjunto de ferramentas para o estudo do contexto dos codões à escala genómica.
Os dados obtidos validam a hipótese de trabalham identificando pontos específicos do mRNA, cujo erro de leitura é superior à média de 10^{-4} a 10^{-5} erros por codão, e, também, potenciais pontos de regulação da expressão genética.

Para além disto, o software disponibiliza várias informações sobre a estrutura primária dos genes, nomeadamente índices CAI, codões raros e codões ou aminoácidos sequencialmente repetidos. Os resultados são visualizadas através de um paradigma bastante familiar, facilitando a interacção com o utilizador.
O programa está em exploração há cerca de seis meses tendo sido já obtidos alguns resultados promissores sobre as leis gerais que governam a fidelidade de descodificação do código genético. A informação obtida está a ser validada *in vivo* no laboratório de genómica funcional.

VIII. Referências

[1] Robert F. Weaver, *Molecular Biology*, New York McGraw-Hill, 2001.

[2] Brian S. Everitt, *The Analysis of Contingency Tables*, CRC Press, 2nd edition, 1992.

[3] M. Beaver, *Introduction to Probability and Statistics*, 9th ed. Duxbury.

[4] Brian S. Everitt, Sabine Landau, Morven Leese, Cluster Analysis, Edward Arnold; 4th edition , 2001.

[5] P.M. Sharp and W.-H. Li "The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications" Nucleic Acids Res., 15, 1281-1295 (1987)