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

José P. Lousado[1], Gabriela R. Moura[2], Manuel A. S. Santos[2], José Luis Oliveira[3]

jlousado@estgl.ipv.pt, gmoura@ua.pt, msantos@bio.ua.pt, jlo@ua.pt

[1] Centro de Estudos em Educação, Tecnologias e Saúde, Instituto Politécnico de Viseu - Campus Politécnico, 3504-510, Viseu, Portugal

[2] CESAM & Departamento. de Biologia, Universidade de Aveiro, 3810-193, Aveiro, Portugal

[3] DETI/IEETA, Universidade de Aveiro, 3810-193, Aveiro, Portugal

Resumo: A descodificação de genomas, em particular do genoma humano, constituiu um marco cientifico 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 muito 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 aminóacidos pelo que existe redundância do código genético, ou seja um mesmo aminóacido 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 sequenciamente 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; *GSWeb* – 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 61x61x61 (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 20x20x20.

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.

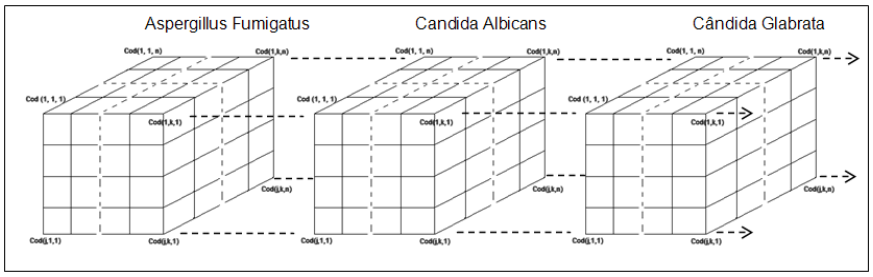

Figura 1 – Representação matricial (4-D Cubo) dos dados resultantes das contagens de tripletos de codões

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).

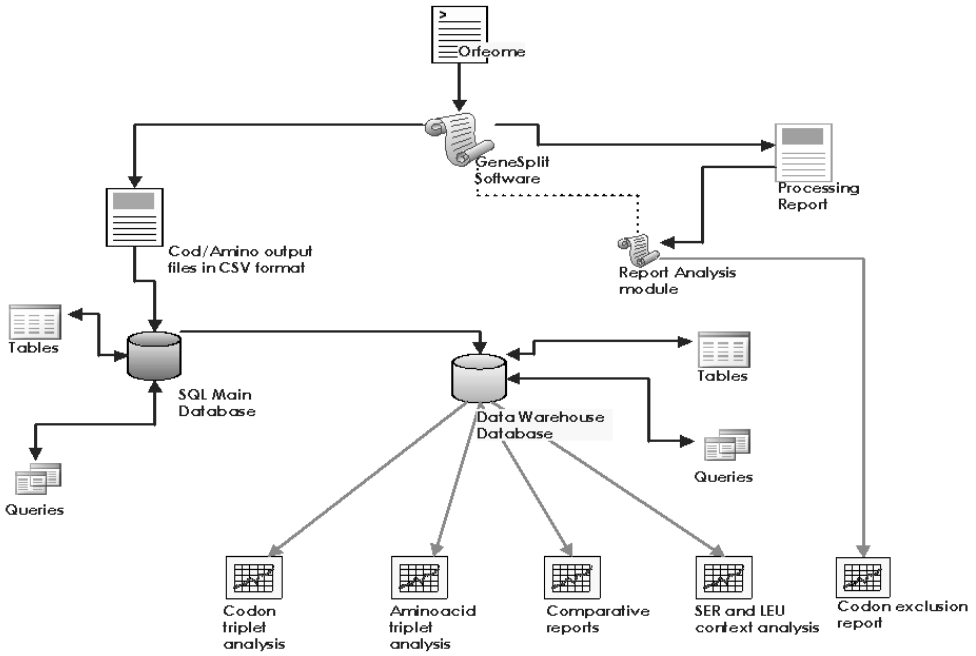

Figura 2 – *Workflow* do projecto, desde o processamento do orfeoma até à disponibilização de resultados.

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 apondador 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 verifiquem. 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 Matrix()
```

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

3.2. Algoritmo para Contagem de Tripletos sem Cadeias de Repetição

O Algoritmo para contagem de tripletos sem cadeias de repetição é em tudo idêntico ao apresentado anteriormente à excepção de que são analisados em cada momento quatro codões e não apenas três como acontece na contagem global. A cada iteração são analisados os codões anteriores à posição n (n-3, n-2 e n-1). Enquanto os quatro codões forem iguais não é alterada a matriz de contagens de codões. No final ficaremos com as contagens onde todas as cadeias de codões iguais, cujo comprimento é superior a 3, contam apenas como uma única ocorrência de um tripleto.

3.3. Algoritmo de Determinação de Cadeias máximas

Quanto ao algoritmo aplicado para efectuar a determinação de cadeias máximas, o processo inicia-se com a utilização do array dos codões do gene. Para cada iteração é analisado o codão anterior, se este for igual ao actual, incrementa-se o contador, até que o codão lido seja diferente do anterior. Nessa iteração lê-se do array de contagens de cadeias máximas, o valor existente para o codão anterior. Se o valor existente for inferior ao contador, actualiza-se o array para este valor.

Para o processamento de cadeias máximas de aminoácidos, são analisados, não o codão, mas sim o aminoácido, mantendo-se o incremento do contador mesmo que o codão seja diferente, desde que este corresponda ao mesmo aminoácido.

4.4. Algoritmo de Determinação de Grupos de repetição

A determinação de grupos de repetição, é efectuada tendo por base o princípio inerente ao algoritmo anterior. No entanto em vez de desprezarmos os valores onde as cadeias encontradas são inferiores, estas são armazenadas numa matriz, permitindo que no final da análise ao orfeoma, tenhamos uma matriz de codões com dimensão *m* x 64, onde m representa o valor máximo de todas as cadeias iguais encontradas, resultando numa matriz contendo o nº de vezes que cada codão aparece em sequências de 1 até *m*.

A descrição do algoritmo para análise de aminoácidos é em tudo análoga à descrição do algoritmo para análise de codões, com a ressalva de que é necessário previamente estabelecer a equivalência entre os codões e os respectivos aminoácidos.

4. GeneSplit

A aplicação GeneSplit foi inicialmente desenhada para efectuar as contagens simples de tripletos de codões (Figura 3). Atendendo aos vários requisitos que resultaram de um processo de desenvolvimento em espiral, várias foram as

funcionalidades acrescentadas ou redefinidas. Como resultado, é possível usar esta ferramenta para obter os seguintes resultados, por orfeoma, gene, cromossoma ou genoma:

- Contagens de tripletos de codões/aminoácidos;
- Determinação de cadeias máximas de codões/aminoácidos;
- Agrupamento de sequências repetidas de codões/aminoácidos;
- Visualização/gravação do relatório de processamento de dados;
- Opção de ignorar cadeias longas (superior a 3) repetidas do mesmo codão;
- Relatórios (informação relativa aos genes que foram desprezados e a causa dessa exclusão; genes que possuem cadeias longas de codões iguais, com a indicação de quantos codões são desprezados);
- Possibilidade de trabalhar com bases códigos genéticos modificados (por exemplo, com ACTG ou com as bases ACUG);
- Processamento em *batch*, permitindo manipular sequencialmente ficheiros de diversas fontes (diferentes pastas);
- Fusão de resultados (Opção "Merge files") para genomas repartidos em vários ficheiros, permitindo que os resultados de análise desses ficheiros indicados num arquivo de processamento *Batch*, possam ser acumulados, minimizando a carga de memória que seria necessária para manipular ficheiros de dados muito grandes.

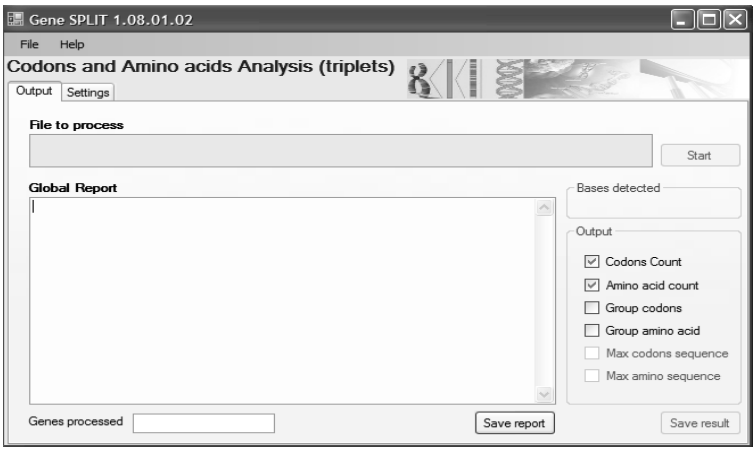

Figura 3 – Interface inicial da aplicação GeneSplit/GSCore

5. Processamento

5.1. Selecção de dados

A selecção de dados das contagens dos tripletos de codões resulta numa tabela com 226.981 linhas para cada genoma. A contagem de tripletos de aminoácidos resulta numa tabela com 8000 linhas, pelo que o resultado da análise de cada genoma exige bastantes recursos e como se compreende, facilmente se atinge uma dimensão de tal forma elevada que impede a utilização de ferramentas de análise tradicionais como as folhas de cálculo.

5.2. Dupla contagem de tripletos de codões

Inicialmente, a contagem de tripletos incidiu sobre a totalidade dos genomas, tendo-se verificado que alguns dos organismos apresentavam no seu genoma elevadas sequências com repetições do mesmo codão. Por esse motivo, foi implementada uma funcionalidade na aplicação informática que permite decidir se a contagem deve reflectir a totalidade do genoma, ou se deve ignorar grandes sequências, entenda-se, as sequências com mais do que três codões iguais. Na contagem de aminoácidos, são considerados iguais, mesmo que o codão de origem seja diferente.

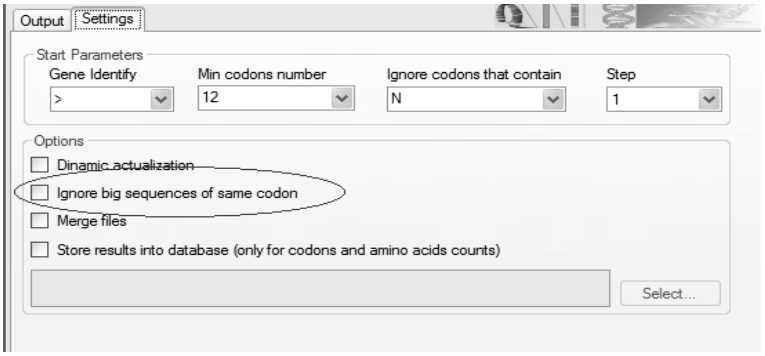

Figura 4 – Parâmetros ajustáveis, com realce para a opção “Ignore big sequences...”

O objectivo dessa diferenciação foi a comparação dos resultados entre as duas contagens de forma a determinar se o comportamento dos organismos, em termos de frequências relativas, foi afectado ou não pela omissão das sequências de repetição.

De acordo com os objectivos globais do estudo, o método foi aplicado da mesma forma à contagem de tripletos de aminoácidos.

5.3. Ontologia para Armazenamento de Dados

Os dados processados são armazenados numa base de dados relacional tendo sido definida uma ontologia de armazenamento de dados através da especificação de cada tabela e os respectivos atributos, conforme apresentado a seguir:

Tabela: tbl_Codon_Amino_Association
Descrição: Associação entre Codões e Aminoácidos
Atributos:

Name	Data type	Length
Cod	Text	3
Amino	Text	3
id	Byte	1

Tabela: tbl_Frequency_Amino
Descrição: Contagens completas das frequências de tripletos de Aminoácidos
Atributos:

Name	Data type	Length
AbrevOrg	Text	10
a1	Text	3
a2	Text	3
a3	Text	3
value	Integer	2
expvalue	Integer	2
freq_rel	Single	4
probb	Single	4
Diff	Single	4
ratio	Single	4

Tabela: tbl_Frequency_Amino_WR
Descrição Contagens das frequências de tripletos de Aminoácidos sem cadeias longas de tripletos iguais
Atributos: (os mesmos que os definidos para a tabela completa)

Tabela: tbl_Frequency_Codon
Descrição: Contagens completas das frequências de tripletos de Codões
Atributos:

Name	Data type	Length
AbrevOrg	Text	10
c1	Text	3
c2	Text	3
c3	Text	3
value	Integer	2
expvalue	Integer	2
freq_rel	Single	4
probb	Single	4
Diff	Single	4
ratio	Single	4

Tabela: tbl_Frequency_Codon_WR
Descrição: Contagens das frequências de tripletos de Codões sem cadeias longas de tripletos iguais
Atributos: (os mesmos que os definidos para a tabela completa)

```
Tabela: tbl_Organisms
Descrição: Organisms dataset
Atributos:
     Name        Data type     Length
     IDOrg        Longint        4
     DescOrg        Text         50
     AbrevOrg       Text         10
     sourceFrom     Memo          -
     dateSource     Text         50
     refID          Byte          1
```

Desta forma qualquer utilizador poderá efectuar a análise de tripletos de genomas e guardar os dados numa base de dados reconhecida pela aplicação, bastando para isso marcar a opção respectiva no separador *Settings* da aplicação GeneSplit.

Os dados das diversas contagens são por defeito guardados em ficheiros no formato CSV, podendo no entanto ser gravadas directamente numa base de dados. A aplicação inclui outros algoritmos que não estão referidos explicitamente, mas que realizam todo o pré-processamento, nomeadamente cálculos estatísticos, tais como médias, frequências relativas, frequência esperada, etc., pelo que o resultado final do processamento da aplicação inclui esses dados pré-processados, no formato e tipo especificado pela ontologia.

6. Portal Web

Paralelamente ao sistema de processamento foi construído um portal Web que possibilita a extracção das contagens efectuadas a partir da base de dados principal (Figura 5). A aplicação pode ser acedida em http://bioinformatics.ua.pt/genesplit, podendo também ser efectuado o *download* da aplicação executável assim como outra documentação, nomeadamente o manual de utilizador.

Actualmente, estão disponíveis para a comunidade 22 orfeomas processados e os respectivos proteomas. Cada orfeoma está disponível em duas versões - com e sem cadeias longas de repetição. A aplicação apresenta uma interface de fácil utilização tendo sido desenhada de acordo com os requisitos da equipa de biólogos envolvidos no projecto. Salienta-se o facto de cada utilizador poder criar um perfil no sistema, de forma a poder gravar as suas sessões no servidor, ficando estas acessíveis para o seu proprietário, a partir de qualquer computador com acesso à internet. Dessa forma o utilizador poderá sempre que quiser, regressar às consultas realizadas para posteriormente continuar o seu trabalho, mesmo que esse decorra distribuído por vários dias. De salientar que não gravados os dados em si mas apenas o conjunto das instruções SQL produzidas em cada query, pelo que o carregamento de cada sessão é extremamente rápido.

As opções de extracção de informação estão separadas em dois grupos: *Standard Queries* e *Advanced Queries*. Na primeira opção o utilizador pode seleccionar um dos vários orfeomas existentes na base de dados, optando de seguida pela selecção

do tipo que mais lhe interessa, se a análise do orfeoma, se a análise do proteoma, com ou sem as repetições referidas anteriormente. De seguida são escolhidos os dados a serem exibidos, em termos de tripletos de codões ou de aminoácidos. Após a submissão do pedido os dados são visualizados no separador *Results*, podendo ser exportados num dos formatos já referidos, ou removido o pedido da lista de sessão.

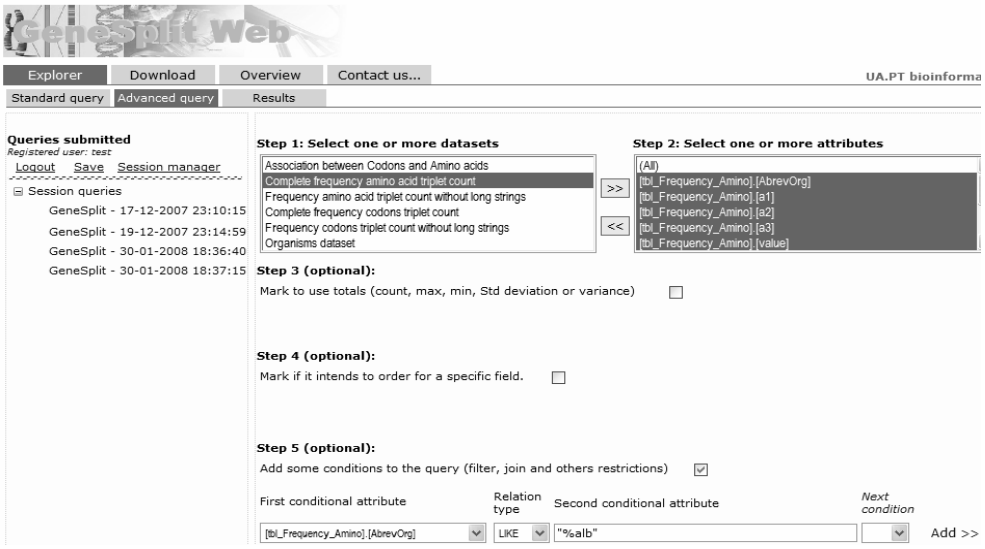

Figura 5 – GeneSplit/GSWeb: interface de consultas avançadas.

A segunda opção (*Advanced*) é especialmente indicada para utilizadores mais experientes que pretendam retirar informação cruzada de mais do que uma tabela, recorrendo para isso a um motor de pesquisa com a inclusão de funções de agregação, ordenação e condições de junção ou filtragem.

7. Conclusão

Os sistemas de informação e de computação têm vindo a assumir um papel cada vez mais importante no desvendar de conhecimento relacionado com as áreas emergentes da biologia molecular, tais como a genómica e proteómica. Um dos problemas que tem vindo a ser estudado envolve os erros de tradução. A implementação de ferramentas informáticas ajudam-nos a perceber, como podemos reduzir esses erros de tradução que conduzem a proteínas aberrantes e como melhorar a eficácia de tradução de proteínas.

Neste artigo apresentámos uma aplicação de software que foi desenvolvida especificamente para análise de associação entre tripletos de codões 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 outros fungos (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