LAB #3 – MODELOS PROBABILÍSTICOS

Grupo I

a) Identificação de *CpG islands* e influência da variação de parâmetros

Com vista a caracterizar a sequência genómica em estudo, procedeu-se à identificação de potenciais *CpG Islands*, regiões genómicas onde se encontram com elevada frequência *CpG sites* – dinucleotido CG, em que "p" denota a ligação fosfodiéster entre os nucleótidos. Tipicamente localizadas nos promotores ou regiões de início de transcrição de genes expressos em organismos vertebrados, a identificação de *CpG Islands* permite prever a localização de potenciais genes com transcrição ativa nas sequências genómicas, assim como caracterizar a transcrição de genes sequenciados, especulando-se que a metilação dos *CpG sites* (processo através do qual um grupo metilo (R-CH_3) é adicionado a uma molécula de DNA por enzimas denominadas metiltransferases de DNA, existindo uma alta probabilidade de ocorrer uma mutação do nucleótido C em T [1]) nos promotores dos genes esteja associada ao silenciamento dos mesmos.

Deste modo, para detetar regiões ricas no padrão CpG, recorreu-se a dois softwares distintos*: CpG Islands*, disponível na coleção de programas *Sequence Manipulation Suite* [2]; *CpG plot*, disponível no site do EMBL-EBI [3].

- ***CpG Islands – Sequence Manipulation Suite***

A ferramenta *CpG Islands* permite identificar potenciais regiões de *CpG Islands*, de acordo com o método descrito por *Gardiner-Garden and Frommer* (1987), que define uma *CpG Island* como sendo uma sequência de nucleótidos com pelo menos 200 pares de base (pb), com conteúdo C+G de pelo menos 50% e um rácio de dímeros CpG, observados por esperados (Obs/Exp), superior a 0.6 [4].

Tendo em conta os critérios estabelecidos, o software computa os resultados para subsequências delimitadas por uma janela de 200 bp que desliza sobre a sequência X em intervalos de 1 pb (*step size*), analisando-se $X^k = (x_{k+1}, \ldots, x_{k+l})$ subsequências, com $0 \leq k \leq L - l$, L o comprimento da sequência genómica X e l o comprimento da janela (200 bp) [1]. Para cada subsequência considerada, o número esperado de dímeros CpG (equação (1)) é calculado como o número de 'C's multiplicado pelo número de 'G's, dividido pelo comprimento da janela (N), i.e, número total de nucleótidos na subsequência (neste caso, N = 200 pb). Como tal, o rácio de CpG Obs/Exp é dado pela equação (2).

$$\#\ CpG \text{ esperados} = \frac{\#\ \mathrm{C} \times \#\ \mathrm{G}}{\mathrm{N}} \qquad (1)$$

$$CpG\ \mathrm{Obs/Exp} = \frac{\#\ CpG\ observados}{\#\ CpG\ esperados} = \frac{\#\ CpG}{\#\ \mathrm{C} \times \#\ \mathrm{G}} \times \mathrm{N} \qquad (2)$$

Após aplicação da ferramenta *CpG islands* à sequência FASTA em análise, 440 potenciais *CpG Islands* foram identificadas, considerando os parâmetros fixos envolvidos na sua classificação, isto é, comprimento da janela (200 bp), o rácio Obs/Exp (> 0,6) e o conteúdo %CG (> 50%). Parte dos resultados obtidos encontram-se discriminados na Figura 1.

CpG Islands results

Results for 10009 residue sequence "Untitled" starting "GCTATCGTAG"

```
CpG island detected in region 160 to 359 (Obs/Exp = 0.63 and %GC = 50.50)
CpG island detected in region 161 to 360 (Obs/Exp = 0.63 and %GC = 50.50)
CpG island detected in region 162 to 361 (Obs/Exp = 0.62 and %GC = 51.00)
CpG island detected in region 163 to 362 (Obs/Exp = 0.68 and %GC = 51.50)
CpG island detected in region 164 to 363 (Obs/Exp = 0.69 and %GC = 51.00)
CpG island detected in region 165 to 364 (Obs/Exp = 0.71 and %GC = 50.50)
CpG island detected in region 168 to 367 (Obs/Exp = 0.71 and %GC = 50.50)
CpG island detected in region 170 to 369 (Obs/Exp = 0.71 and %GC = 50.50)
CpG island detected in region 171 to 370 (Obs/Exp = 0.71 and %GC = 50.50)
CpG island detected in region 172 to 371 (Obs/Exp = 0.71 and %GC = 50.50)
CpG island detected in region 173 to 372 (Obs/Exp = 0.71 and %GC = 50.50)
CpG island detected in region 174 to 373 (Obs/Exp = 0.71 and %GC = 50.50)
CpG island detected in region 469 to 668 (Obs/Exp = 0.66 and %GC = 50.50)
CpG island detected in region 470 to 669 (Obs/Exp = 0.66 and %GC = 50.50)
CpG island detected in region 471 to 670 (Obs/Exp = 0.66 and %GC = 50.50)
CpG island detected in region 473 to 672 (Obs/Exp = 0.66 and %GC = 50.50)
CpG island detected in region 474 to 673 (Obs/Exp = 0.66 and %GC = 50.50)
```

Figura 1 - Parte dos resultados obtidos pela aplicação da ferramenta CpG Islands, disponível na coleção de programas Sequence Manipulation Suite. Para cada CpG Island detetada, discrimina-se a pb de início e fim, o respetivo rácio CpG Obs/Exp e o conteúdo %CG.

Como o comprimento mínimo das *CpG Islands* não é definido, o algoritmo assume que as *CpG islands* têm pelo menos o comprimento da janela (200 bp), sendo que os resultados obtidos correspondem às janelas em que os requisitos para a definição duma *CpG Island* são verificados. Devido ao *step size* aplicado, as *CpG Islands* obtidas apresentam elevada sobreposição. Deste modo, para melhor delimitar as regiões que contêm potenciais *CpG Islands*, procedeu-se à concatenação das regiões que apresentavam sobreposição, tendo-se atribuído o rácio Obs/Exp e conteúdo %CG mínimos entre as *CpG Islands* concatenadas. Os resultados obtidos após concatenação são os seguintes:

- **Island 1** - nucleótidos 160-373 (Min Obs/Exp =0.62 and min %GC= 50.50)
- **Island 2** - nucleótidos 469-730 (Min Obs/Exp =0.65 and min %GC= 50.50)
- **Island 3** - nucleótidos 2315-2569 (Min Obs/Exp =0.80 and min %GC= 50.50)
- **Island 4** - nucleótidos 3651-3904 (Min Obs/Exp =0.61 and min %GC= 50.50)

- **Island 5** - nucleótidos 4745-5137 (Min Obs/Exp =0.63 and min %GC= 50.50)
- **Island 6** - nucleótidos 7461-7837 (Min Obs/Exp =0.71 and min %GC= 50.50)
- **Island 7** - nucleótidos 8403-8872 (Min Obs/Exp =0.61 and min %GC= 50.50)

Dadas as potenciais *CpG Islands*, quanto maior o rácio Obs/Exp, menor a probabilidade de a região ser um falso positivo, isto é, maior a probabilidade de corresponder de facto a uma *CpG Island*.

É de salientar que certas *CpG Island* detetadas têm um valor de rácio superior a 1. Tal pode sugerir que os valores obtidos por esta ferramenta estejam sobrevalorizados, identificando como potenciais *CpG Islands* regiões da sequência que na verdade não cumprem com os requisitos mínimos definidos acima

- ***CpG Plot***

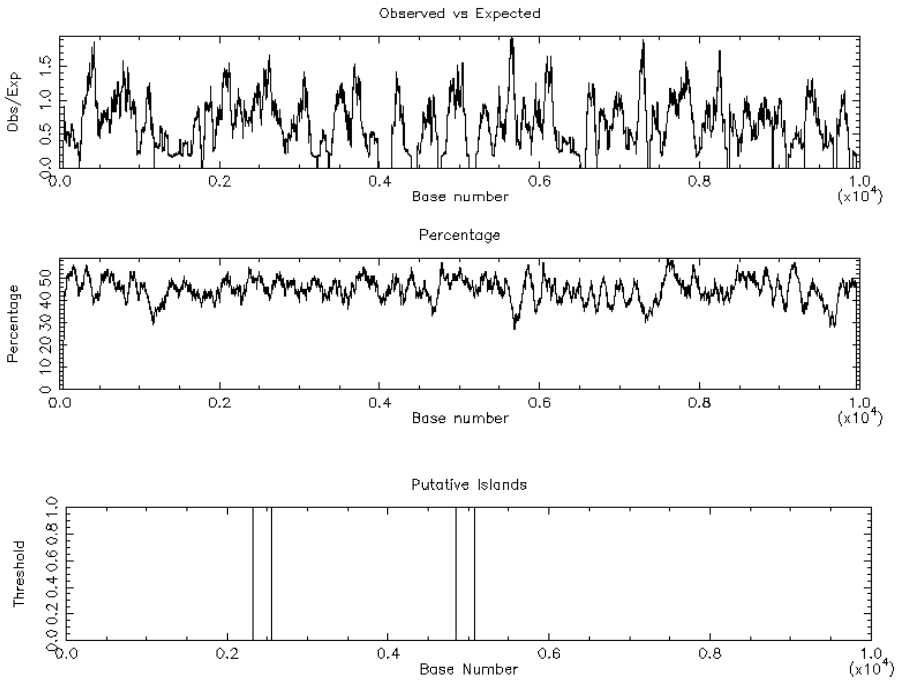
Ao contrário da ferramenta anterior, o software *CpG plot* do EMBL-EBI permite ao utilizador manipular os diversos parâmetros envolvidos na classificação de potenciais CpG Islands:

- Tamanho da janela que delimita as subsequências a analisar;
- Dimensão mínima de uma potencial *CpG Island*;
- Percentagem mínima do conteúdo %CG;
- Rácio mínimo CpG's Obs/Exp.

Novamente, o software computa os resultados para subsequências delimitadas pela janela de tamanho especificado pelo utilizador, a qual desliza sobre a sequência X a intervalos de 1 pb (*step size*). Contudo, apenas reporta uma potencial *CpG Island* se a média dos resultados obtidos em 10 janelas consecutivas cumprir os critérios impostos pelo utilizador, verificando-se esta como sendo a principal diferença relativamente ao software anterior.

Pretendendo-se simular os resultados anteriormente obtidos com a ferramenta *CpG Islands*, definiram-se os parâmetros do *CpG plot* idênticos aos do *software CpG Islands*, obtendo-se os resultados representados na Figura 2.

Realizando uma análise cruzada dos resultados obtidos pelos dois softwares, conclui-se que as anteriormente identificadas 2ª, 3ª, 5ª, 6ª e 7ª potenciais *CpG Islands* através da ferramente *CpG Islands* têm correspondência com as obtidas através do *CpG Plot*. Visto a deteção das *CpG islands* para este último software envolver a computação de uma média móvel de 10 subsequências, é mais rígido e exigente na seleção das regiões correspondentes a *CpG Islands (*exige que os critérios relativos ao conteúdo %CG e rácio Obs/Exp mínimos sejam verificados não só na janela devolvida como ainda no conjunto das 9 janelas a jusante da primeira). Ora, as 1ª e 4ª *CpG Islands* obtidas através do software *CpG Islands* são das que apresentam um rácio Obs/Exp menor, ou seja, apresentam maior probabilidade de corresponderem a falsos positivos, sendo esta hipótese apoiada pela sua não deteção com o CpG plot.




```
CPGPLOT islands of unusual CG composition
EMBOSS_001 from 1 to 10009

     Observed/Expected ratio > 0.60
     Percent C + Percent G > 50.00
     Length > 200

 Length 221 (2326..2546)

 Length 232 (4840..5071)
```

Figura 2 - Resultados obtidos pela aplicação da ferramenta CpG plot, disponível no site do EMBL-EBI, para os seguintes parâmetros: dimensão da janela de 200 pb, comprimento mínimo da sequência de 200 pb, rácio CpG observados/esperados superior a 0.6 e conteúdo %CG superior a 50%.

Pretendendo-se avaliar o impacto dos parâmetros nos resultados obtidos, isoladamente, procedeu-se à variação de cada um dos parâmetros. Enquanto parâmetros de referência, estabelecerem-se os definidos anteriormente (dimensão da janela: 200 bp; comprimento mínimo das CpG Islands: 200 bp; conteúdo %CG: 50%; rácio Obs/Exp: 0.6). Os resultados obtidos encontram-se no Anexo A.

Após a análise dos diferentes resultados obtidos para diferentes parâmetros do software, foi possível retirar algumas conclusões:

- A diminuição do comprimento mínimo da sequência das *CpG Islands* conduz ao aumento sucessivo do número de ilhas reconhecidas pelo programa, uma vez que os critérios para considerar uma *CpG Island* tornam-se menos rígidos, considerando todos os outros parâmetros coincidentes com os de referência. A relação inversa foi também verificada, pela mesma razão mencionada.

- A variação do rácio Obs/Exp e conteúdo %CG mínimos seguem a mesma evolução, sendo que da sua diminuição resulta o aumento do número de ilhas encontradas. Visto o critério para existir uma *CpG Island* se tornar mais abrangente (menos rígido), aumenta a probabilidade de uma região ser assinalada como uma potencial *CpG Island*. A relação inversa foi igualmente verificada para ambos os parâmetros.
- A variação da dimensão das janelas conduz à alteração dos perfis do rácio Obs/Exp e conteúdo %CG estimados ao longo da sequência, sendo que o impacto que tem nos resultados obtidos apresenta uma maior dependência com as características da sequência em causa. Para os resultados obtidos, a redução da dimensão das janelas resultou numa diminuição do número de *CpG Islands* identificadas. Tal poderá dever-se ao facto de se restringir a dimensão da janela, mantendo os restantes parâmetros fixos, levado a que sejam excluídas as *CpG Islands* com conteúdo %CG menos uniforme ao longo de toda a sua extensão, isto é, que contêm regiões mais ricas em CG e outras menos ricas.

Por fim, remetendo para a comparação feita entre os resultados obtidos através dos dois softwares realizada para os parâmetros de referência conclui-se que a escolha dos recursos algorítmicos é um aspeto crítico a considerar, ao produzirem-se inconsistências nos resultados obtidos para os mesmos parâmetros ajustáveis pelo utilizador.

- ***DNA Stats***

À sequência nucleotídica em análise também se aplicou a ferramenta *DNA Stats*, disponível na coleção de programas *Sequence Manipulation Suite*. Esta ferramenta fornece o número de ocorrências de cada nucleótido, bem como de diferentes dímeros na sequência em análise, quantificando a sua percentagem relativa [2]. Na Tabela 1 encontram-se os resultados obtidos.

Tabela 1 – Resultados obtidos pela aplicação da ferramenta DNA stats à sequência em análise (apenas se apresentam os resultados relevantes).

Padrão	Nº de vezes encontradas	Percentagem
G	1682	16.80
A	2964	29.61
T	2550	25.48
C	2813	28.10
CG	302	3.02
G, C	4495	44.91

Examinando a Tabela 1, constata-se que a probabilidade de se encontrar um dímero CG é de 3,02% (valor observado), sendo inferior à estimada a partir das probabilidades independentes dos nucleótidos C (28,10%) e G (16,80%), que atribuiria 4,72% (valor esperado) de probabilidade à ocorrência de *CpG sites*. Tal deve-se à elevada taxa de mutação de *CpG sites* metilados em organismos vertebrados, resultando na sub-representação dos dinucleótidos CpG no genoma face ao estimado através das probabilidades independentes dos nucleótidos C e G (consequência da conversão, por períodos evolutivamente representativos, dos dinucleótidos CG em TG). Ainda assim, o valor esperado e observado das probabilidades não dista significativamente, sugerindo que a sequência em causa não foi extensivamente metilada, pelo

que *CpG islands* são suscetíveis de ser encontradas, apesar do conteúdo em guaninas e citosinas não ser elevado (44,91%, esperando-se valores superiores a 60% numa *CpG Island*).

- ***ORF Finder***

Por fim, recorreu-se ao software *ORF Finder*, disponível na coleção de programas *Sequence Manipulation Suite*, o qual percorre a sequência nucleotídica fornecida de modo a identificar potenciais *Open Reading Frames* (ORFs), isto é, segmentos da sequência que potencialmente codificam proteínas, por se encontrarem entre um codão de iniciação (geralmente AUG) e um codão de terminação (geralmente UAA, UAG ou UGA) [2] [5]. Cada ORF identificada é caracterizada quanto ao seu pb inicial e final, assim como a correspondente sequência de aminoácidos mais provável.

1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

Figura 3 - Sequência de DNA com três possíveis Reading Frames. Os codões de iniciação estão destacados a roxo, e os codões de terminação estão destacados a vermelho. [5]

O programa permite que o operador possa estabelecer alguns parâmetros que vão determinar como é realizada a procura de ORFs:

- Tipo de codão em que a ORF se inicia (qualquer um, ATG, ou ATG, GTG, CTG e TTG);
- *Reading Frame* (1,2,3 ou todas);
- Sequência a analisar (direta ou a inversa);
- Tamanho mínimo de uma ORF;
- Tipo de código genético utilizado na pesquisa.

Considerando-se pertinente analisar para todos os *Reading Frames*, modificaram-se os parâmetros *default*. Nesse sentido, a pesquisa foi feita com os seguintes parâmetros: inicialização da ORF em qualquer codão, com o *Reading Frame* 1, 2 e 3 para sequência direta, um tamanho mínimo da ORF de 30 codões, considerando-se o código genético standard.

Tal como já foi mencionado, as *CpG Islands* situam-se tipicamente em promotores ou locais de iniciação de transcrição dos genes de organismos vertebrados. Tal implica que as ORFs geralmente precedam as respetivas *CpG Islands*, pois o codão de iniciação (início de uma ORF) encontra-se após uma região promotora. Considerando-se a típica correlação espacial entre ORFs e *CpG Islands*, na Figura 4 discriminam-se parte dos resultados obtidos, nomeadamente os que dizem respeito a ORFs consistentes com as *CpG islands* identificadas pela aplicação da ferramenta CpG Plot com os parâmetros default ([2326…2546] e [4840...5071] – Figura 11 do Anexo A.I).

```
>ORF number 10 in reading frame 1 on the direct strand extends from base 2632 to base 2730.
TCCAGGTCGGTTTCTATCTATGACGCAATCCTTTTTCAGTACGAAAGGACCAAAAAGAAG
AGGCCCCTGTTACAAACACGCCTCACCCCAACTCGCTGA

>Translation of ORF number 10 in reading frame 1 on the direct strand.
SRSVSIYDAILFQYERTKKKRPLLQTRLTPTR*

>ORF number 24 in reading frame 1 on the direct strand extends from base 5365 to base 5544.
TCTGCTTCTTCAGATTTGCAATCTGAAATGTTAAATACACCTCAGGGCTGGCAAGAAGAG
GACTTGAACCTCTGTACATGGGGCTACAATCCACCGCTTAACCCTCAGCCATCTTACCCG
TGGCAATTACACGTTGATTCTTCTCAACTAACCATAAAGACATCGGCACCCTTTACCTAA

>Translation of ORF number 24 in reading frame 1 on the direct strand.
SASSDLQSEMLNTPQGWQEEDLNLCTWGYNPPLNPQPSYPWQLHVDSSQLTIKTSAPFT*

>ORF number 7 in reading frame 2 on the direct strand extends from base 2576 to base 2749.
GGGTTCGTTTGTTCAACGATTAAAATCCTACGTGATCTGAGTTCAGACCGGAGTAATCCA
GGTCGGTTTCTATCTATGACGCAATCCTTTTTCAGTACGAAAGGACCAAAAAGAAGAGGC
CCCTGTTACAAACACGCCTCACCCCAACTCGCTGAACCCAACTCAAGCGAATAA

>ORF number 12 in reading frame 3 on the direct strand extends from base 2655 to base 2945.
CGCAATCCTTTTTCAGTACGAAAGGACCAAAAAGAAGAGGCCCCTGTTACAAACACGCCT
CACCCCAACTCGCTGAACCCAACTCAAGCGAATAAAGAGGTGCCTCATCCCGTCAAAGAA
CATGACATATTAAGGTGGCAGAGCCCGGACATTGCAAAAGACCTAAGCCCTTTCTACAGA
GGTTCAAGTCCTCTCCTTAATTATGATCTCAACCCTTATTACCCACGTGATCTCCCCCCT
GGCTTTCATCGTCCCTATTCTTCTGGCAGTAGCCTTCCTCACCTTAGTTGA

>Translation of ORF number 12 in reading frame 3 on the direct strand.
RNPFSVRKDQKEEAPVTNTPHPNSLNPTQANKEVPHPVKEHDILRWQSPDIAKDLSPFYR
GSSPLLNYDLNPYYPRDLPPGFHRPYSSGSSLPHLS*

>ORF number 21 in reading frame 3 on the direct strand extends from base 5271 to base 5390.
TTAACAGCTAAGCGCTCAAACCAGCGAGCATCCATCTACTTTCCCCCGCCTAGCAAAAAC
AAAATAGGCGGGGGAAAGCCCCGGCAGACGTTAATCTGCTTCTTCAGATTTGCAATCTGA

>Translation of ORF number 21 in reading frame 3 on the direct strand.
LTAKRSNQRASIYFPPPSKNKIGGGKPRQTLICFFRFAI*
```

Figura 4 - Parte dos resultados (revelantes) devolvidos pela ferramenta ORF finder, disponível na coleção de programas Sequence Manipulation Suite, com os parâmetros impostos: inicialização da ORF em qualquer codão; reading frame 1, 2 e 3 para sequência direta; tamanho mínimo da ORF de 30 codões; código genético standard. Para cada ORF identificada, o programa retorna ao pb inicial e final de cada ORF e a correspondente sequência de aminoácidos mais provável.

Segundo a metodologia aplicada, estando a identificação das ORFs associada a *CpG Islands* previstas pelo software *CpG plot*, as mesmas constituem unicamente evidência acerca da localização de genes com transcrição ativa, não permitindo aferir acerca da localização de genes silenciados, por não estarem associados a *CpG Islands*.

Posto isto, se pretendido validar os resultados obtidos, seria necessário confirmar experimentalmente a tradução das sequências identificadas em proteínas ou em RNA não-codificante (ncRNA), ou alternativamente proceder a uma base de dados como a *protein BLAST*, com o objetivo de se encontrar proteínas com sequências primárias semelhantes. Sem proceder a nenhum dos procedimentos referidos, as ORF identificadas não constituem prova suficiente e conclusiva da presença de um gene ou de ncRNA, sendo uma das maiores limitações da ferramenta ORF Finder a não distinção entre os codões ATG que constituem codões de iniciação e aqueles que não constituem, por não serem precedidos de um promotor.

b) Plataforma *GenBank Nucleotide* e identificação do genoma

Pretendendo-se analisar os resultados obtidos quanto à sua validade, procedeu-se à comparação da sequência nucleótida fornecida com as disponíveis na *GenBank Nucleotide database*, de forma a avaliar a correlação das *CpG Islands* previamente identificadas com a distribuição dos genes devidamente caracterizada na sequência de maior homologia.

Para esse fim, recorreu-se ao programa *BLAST* (*Basic Local Alignment Search Tool*), nomeadamente à ferramenta *nucleotide BLAST*, efetuando o alinhamento local da sequência fornecida com as contidas na *GenBank Nucleotide database*, avaliando o grau de homologia entre as mesmas.

Considerando os parâmetros *default*, obteve-se um total de 100 sequências que produzem alinhamentos locais significativos, como denotado pela elevada *percent identity* (80%-100%) e o *E value* (0%), apresentando-se na figura 5 os primeiros resultados obtidos, por ordem decrescente da *percent identity*.

select all *100 sequences selected* GenBank Graphics Distance tree of results

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Abalistes stellaris mitochondrial DNA, complete genome	18484	18484	100%	0.0	100.00%	AP009202.1
Abalistes stellaris mitochondrion, complete genome	12497	17328	100%	0.0	96.82%	MG515171.1
Balistapus undulatus mitochondrial DNA, complete genome	12338	12338	99%	0.0	88.96%	AP009203.1
Rhinecanthus aculeatus mitochondrial DNA, complete genome	12059	12059	100%	0.0	88.44%	AP009210.1
Pseudobalistes flavimarginatus mitochondrial DNA, complete genome	11897	11897	100%	0.0	88.17%	AP009209.1
Odonus niger mitochondrial DNA, complete genome	11865	11865	99%	0.0	88.11%	AP009208.1
Balistes vetula mitochondrial DNA, complete genome	11814	11814	100%	0.0	88.00%	AP009204.1
Pseudobalistes fuscus mitochondrion, complete genome	11786	11786	100%	0.0	87.95%	KU985150.1
Xanthichthys auromarginatus mitochondrial DNA, complete genome	11707	11707	100%	0.0	87.82%	AP009211.1
Xenobalistes tumidipectoris mitochondrial DNA, complete genome	11673	11673	100%	0.0	87.76%	AP009182.1
Balistoides conspicillum mitochondrial DNA, complete genome	11642	11642	99%	0.0	87.73%	AP009205.1
Sufflamen fraenatum mitochondrial DNA, complete genome	11638	11638	100%	0.0	87.69%	AP004456.1
Melichthys vidua mitochondrial DNA, complete genome	11585	11585	99%	0.0	87.61%	AP009207.1
Thamnaconus modestus mitochondrial DNA, complete genome	8835	8835	99%	0.0	82.72%	AP009185.1

Figura 5 - Parte das sequências obtidas que produzem alinhamentos significativos com a sequência nucleótida em estudo, devolvidas pela ferramenta nucleotide BLAST.

Para os alinhamentos realizados, a sequência correspondente ao genoma mitocondrial da espécie *Abalistes stellaris* diz respeito à que apresenta maior homologia com a sequência em estudo, com o respetivo alinhamento a ser o único com *percent identity* de 100%, isto é, as sequências alinhadas apresentam nucleótidos idênticos na mesma posição em toda a extensão do alinhamento. O alinhamento obtido apresenta um grau de significância máximo, isto é, *E value* de 0.0. Adicionalmente, o alinhamento em questão apresenta *Query Cover* de 100%, indicando que a sequência da base de dados engloba a totalidade da sequência submetida, tendo a primeira 16502 pb e a segunda 10009 pb. O referido sugere que a sequência em estudo

corresponde aos primeiros 10009 pb do genoma mitocondrial da espécie *Abalistes stellaris*, apresentando-se na figura 6 uma descrição gráfica da sequência.

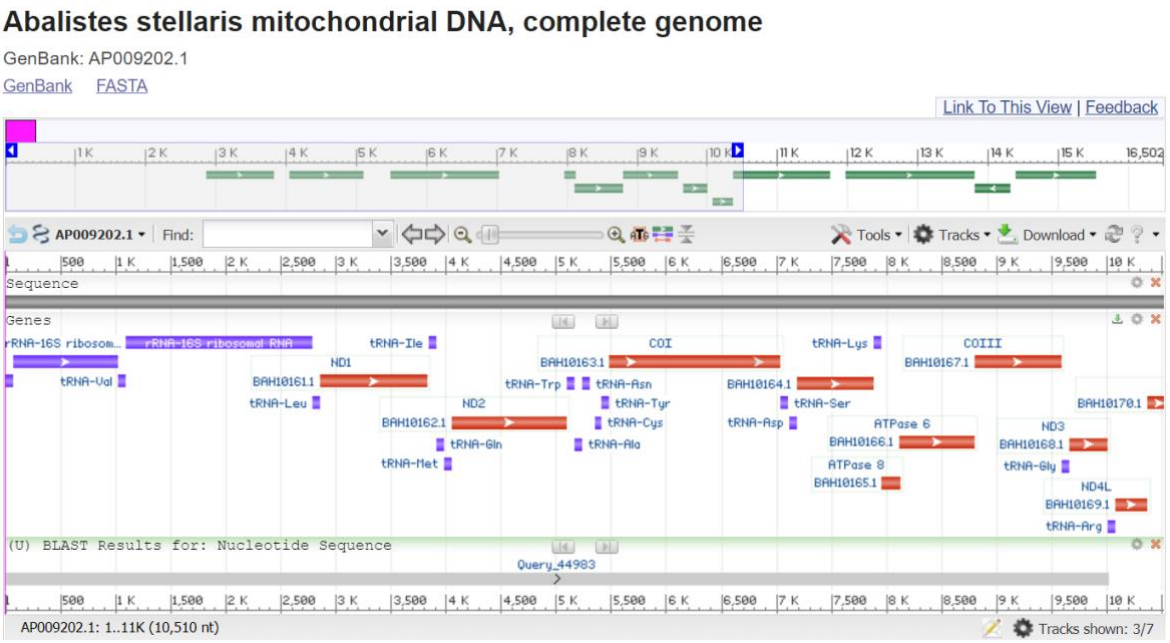


Figura 6 - Descrição gráfica do genoma mitocondrial da espécie Abalistes stellaris (comprimento de 16503 pb), discriminando-se espacialmente os genes contidos no fragmento 1-10009 bp.

Acedendo à anotação[1] disponível para o genoma mitocondrial da espécie *Abalistes stellaris* na *Genbank Nucleotide database,* no fragmento 1-10009 bp identificam-se 7 genes, bem como 15 regiões codificadoras de t-RNA e 2 de rRNA no início do genoma. Na tabela 2 descrevem-se os 7 genes referidos quanto à sua localização e comprimento.

Tabela 2 - Genes contidos no fragmento 1-10009 bp do genoma mitocondrial da espécie Abalistes stellaris, anotados na Genbank Nucleotide database.

Genes	bp Inicial	bp Final	Comprimento (bp)
1º Gene	2857	3831	975
2º Gene	4048	5092	1045
3º Gene	5494	7034	1551
4º Gene	7956	8123	168
5º Gene	8114	8796	683
6º Gene	8797	9581	785
7º Gene	9654	10002	349

1 https://www.ncbi.nlm.nih.gov/nucleotide/AP009202.1?report=genbank&log$=nucltop&blast_rank=1&RID=TGEF9FCR016

De forma a correlacionar os genes identificados no genoma e as regiões assinaladas como potenciais *CpG Islands* pelas ferramentas previamente aplicadas, torna-se importante considerar o já mencionado conhecimento *a priori* acerca da distribuição preferencial das *CpG Islands* nas regiões promotoras ou de início de transcrição dos genes de organismos vertebrados.

Para fins de visualização da análise a realizar, na figura 7 esquematiza-se a correlação espacial entre os genes anotados para o fragmento 1-10009 bp e as potenciais *CpG Islands* identificadas pelos *softwares CpG plot* e *CpG Islands* do SMS, tendo em conta os parâmetros nomeados como referência (dimensão da janela: 200 bp; comprimento mínimo das CpG Islands: 200 bp; conteúdo %CG: 50%; rácio Obs/Exp: 0.6).

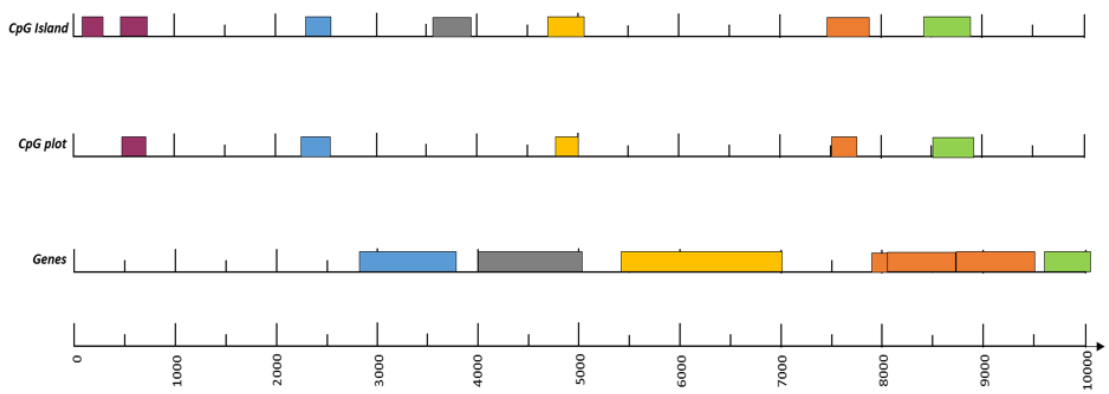


Figura 7 - Esquematização da correlação espacial entre os genes anotados na Genbank Nucleotide database para o fragmento 1-10009 bp e as potenciais CpG Islands identificadas pelos softwares CpG plot e CpG Islands do SMS, tendo em conta os parâmetros nomeados como referência (dimensão da janela: 200 bp; comprimento mínimo das CpG Islands: 200 bp; conteúdo %CG: 50%; rácio Obs/Exp: 0.6). Atribui-se aos genes a mesma cor que as potenciais CpG Islands com papel regulador na sua transcrição, por se localizarem na proximidade da sua região regulatória.

Remetendo para a Figura 7, as *CpG Islands* assinaladas a roxo não se encontram na proximidade da região regulatória de nenhum gene anotado, encontrando-se sobrepostas a uma região de RNA funcional não-codificante (69 – 1022 bp) – gene 12s rRNA –, que codifica a subunidade 12S do ribossoma mitocondrial[2]. Com base nisto, conjetura-se a hipótese de que as *CpG Islands* em causa participam na regulação da transcrição desta molécula de rRNA.

Relativamente ao 1º gene anotado, assinalado a azul, o mesmo encontra-se a 300 pb após uma *CpG Island* estimada por ambos os softwares, sendo consistente com a localização de uma região promotora desse gene. Quanto aos 2º, 3º e 7º genes anotados, respetivamente assinalados a cinzento, amarelo e verde, as potenciais *CpG Islands* identificadas encontram-se sobrepostas com outros genes, sugerindo que estas correspondam a falsos positivos, apesar de apresentarem elevado conteúdo %CG. Tal como discutido na alínea a), modificando os parâmetros do cgplot, poder-se-ia obter uma melhor modulação da real distribuição das *CpG Islands.*

[2] https://en.wikipedia.org/wiki/MT-RNR1

Por fim, dada a proximidade dos 4º, 5º e 6º genes, assinalados a laranja, considera-se a hipótese de existir uma região regulatória comum, onde estaria localizada a *CpG Island* assinalada a laranja e estimada por ambos os softwares.

Tendo em conta as correlações estabelecidas, os 1º, 4º, 5º e 6º genes anotados reúnem evidências de serem genes com transcrição ativa, visto terem-se identificado potenciais *CpG Islands* consistentes com a localização das suas regiões regulatórias. Quanto aos restantes 3 genes, o facto de não se terem reunido as condições para considerar *CpG Islands* nas respetivas regiões regulatórias poderia sugerir que os *CpG sites* teriam sido metilados, com a consequente conversã dos dinucleótidos CG em TG, documentando-se evidência de que sob estas condições a expressão dos genes esteja silenciada. Contudo, ao contrário dos genes nucleares, os quais frequentemente têm associados múltiplos promotores, a transcrição dos genes mitocondriais encontra-se essencialmente associada a promotores contidos na região reguladora D-loop (15677 – 16502 bp), o principal segmento não-codificante do genoma mitocondrial. Como tal, para os genes não associados a *CpG Islands*, estas podem estar contidas no D-loop, não tendo sido identificadas visto a análise restringir-se ao fragmento 1-10009 bp da sequência em estudo [6] [7].

Grupo II

a) Markov Chain Model de 1ªordem

Considerando um modelo de primeira ordem da cadeia de Markov, ou seja, a distribuição da probabilidade do símbolo X_i depende apenas do símbolo X_{i-1} (propriedade denominada *memorylessness*), com probabilidades de transição, a_{st}, dadas por:

$$(3) \quad a_{st} = P(X_i = t | X_{i-1} = s)$$

, com s e t pertencentes um determinado alfabeto A, a soma das probabilidades de todas as sequências possíveis de estados de comprimento L pode ser escrita da seguinte forma:

$$\sum_x P(X) = \sum_{xL} \dots \sum_{x_2} \sum_{x_1} P(x_1) \prod_{i=2}^{L} a_{x_{i-1}x_i}$$

Rearranjando a ordem, obtém-se o seguinte:

$$(4) \quad \sum_x P(X) = \sum_{x_L} \dots \sum_{x_2} \sum_{x_1} P(x_1) \prod_{i=2}^{L} a_{x_{i-1}x_i} = \sum_{x_L} \dots \sum_{x_2} \prod_{i=3}^{L} a_{x_{i-1}x_i} \sum_{x_1} P(x_1) a_{x_1 x_2}$$

Aplicando a lei da probabilidade total à soma interna em x_1, bem como considerando a definição de a_{st} dada pela equação (3), deriva-se:

$$(5) \quad P(x_2) = \sum_{x_1} P(x_1, x_2) = \sum_{x_1} P(x_1)P(x_2|x_1) = \sum_{x_1} P(x_1)a_{x_1x_2}$$

, tendo também em consideração a definição de probabilidade condicional:

$$P(x_2|x_1) = P(x_1, x_2)/P(x_1) \Leftrightarrow P(x_1, x_2) = P(x_2|x_1)\, P(x_1)$$

Aplicando a equação (5) na equação (4), esta última pode ser simplificada:

$$\sum_x P(X) = \sum_{x_L} \dots \sum_{x_2} \prod_{i=3}^{L} a_{x_{i-1}x_i} P(x_2) = \sum_{x_L} \dots \sum_{x_3} \prod_{i=4}^{L} a_{x_{i-1}x_i} \sum_{x_2} P(x_2)P(x_3|x_2) =$$
$$= \sum_{x_L} \dots \sum_{x_3} \prod_{i=4}^{L} a_{x_{i-1}x_i} P(x_3)$$

Aplicando o mesmo procedimento a todos os símbolos, ou seja, após L - 1 passos, obtém-se:

$$\sum_x P(X) = \sum_{x_L} P(x_L)$$

Uma vez que a soma considera as probabilidades de todos os estados possíveis X_L assumidos na posição L da Cadeia de Markov, conclui-se que

$$\sum_x P(X) = 1 \qquad c.q.d$$

, provando-se que o modelo de Markov Chain tem uma distribuição de probabilidade adequada no espaço das sequências com comprimento L.

b) Hidden Markev Model (HMM)

Num HMM, $\mathrm{P}(x, \pi)$ representa a probabilidade conjunta de se observar uma sequência x e uma sequência de estados π. Define-se o caminho mais provável π^*como:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathrm{P}(x, \pi)$$

Aplicando a definição de probabilidade condicionada (definida em (6)), tem-se que:

$$\mathrm{P}(x, \pi) = \mathrm{P}(\pi|x)\mathrm{P}(x)$$

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathrm{P}(\pi|x)\mathrm{P}(x)$$

Dado que a probabilidade da sequência x não depende do caminho oculto π, demonstra-se que a definição inicial é equivalente a:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathrm{P}(\pi|x) \qquad c.q.d$$

c) Markov Chain Model de 1ªordem – transição *two-step*

Considerando uma cadeia de Markov de 1ª ordem com probabilidades de transição $a_{st} = P(X_{i+1} = t|X_i = s)$ em que s e t pertencem a um determinado alfabeto A, a probabilidade de uma transição *two-step* $a_{su} = P(X_{i+2} = u|X_i = s)$ pode ser derivada recorrendo-se à definição de probabilidade condicionada e ao Teorema de Bayes:

$$P(X_{i+2} = u|X_i = s) = \sum_{x_{i+1}} P(x_{i+2}, x_{i+1}|x_i)$$
$$= \sum_{x_{i+1}} P(x_{i+2}|x_{i+1}, x_i).\, P(x_{i+1}|x_i)$$

Aplicando a propriedade de *memorylessness,* tem-se que:

$$P(X_{i+2} = u|X_i = s) = \sum_{x_{i+1}} P(x_{i+2}|x_{i+1}).\, P(x_{i+1}|x_i) = \sum_{x_{i+1}} a_{x_{i+1},x_{i+2}}.\, a_{x_i,x_{i+1}}$$
$$= \sum_{x_{i+1}} a_{x_{i+1},u}.\, a_{s,x_{i+1}}$$

Visto $a_{st} = P(X_{i+1} = t|X_i = s)$, demonstra-se que:

$$a_{s,u} = P(X_{i+2} = u|X_i = s) = \sum_{t \in A} a_{st} a_{tu} \quad c.q.d$$

Grupo III

a) Representação Gráfica do Modelo HMM

O modelo estudado, composto por três *hidden states* e pelo alfabeto Σ = {A,T,C,G}, foi representado graficamente tendo em conta as probabilidades de transição e de emissão referidas no enunciado:

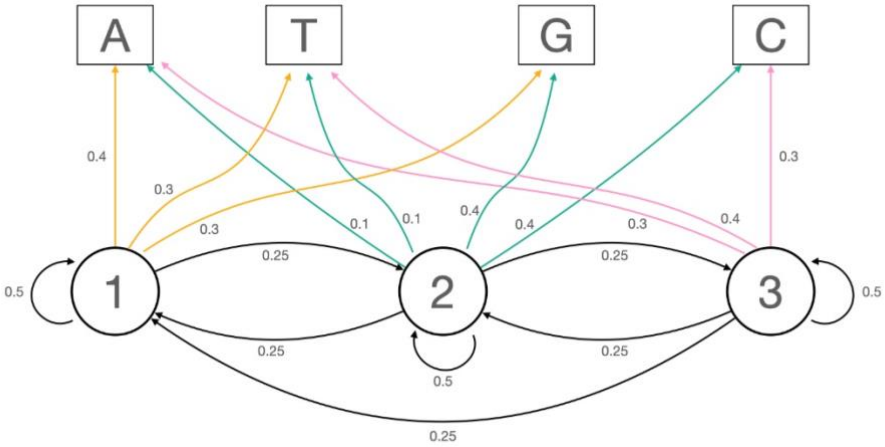


Figura 8 - Representação Gráfica do Hidden Markov Model com estados Q = {1,2,3}, alfabeto Σ = {A,T,C,G} e probabilidades de emissão e transmissão explicitadas no enunciado. As probabilidades de emissão nula não são representadas, assim como não se inclui o hidden state inicial, cuja probabilidade de transição para os estados Q é igualitária (1/3).

b) Processo de Decodificação – Algoritmo de Viterbi

Tendo em conta o modelo HMM previamente apresentado e a sequência de DNA S = CATGCGGGTTATAAC, pretende-se decodificar a sequência de estados π* que maximiza a probabilidade de S ser gerada, isto é, encontrar a probabilidade P(x,π) máxima para todos os caminhos π possíveis.

Para tal, recorreu-se ao algoritmo de Viterbi, que engloba a construção de um *Manhattan* composto por Q linhas e N colunas, sendo Q o número de estados e N o tamanho da sequência dada como input. Este corresponde a um grafo acíclico onde a aresta que tem origem na entrada (k,i) e dirigida para (l, i+1) tem um peso dado por $e_l(x_{i+1}).a_{kl}$ e em que a única direção possível é da esquerda para a direita. Assim, o problema da decodificação fica reduzido a encontrar o maior caminho no grafo que maximiza a probabilidade pretendida.

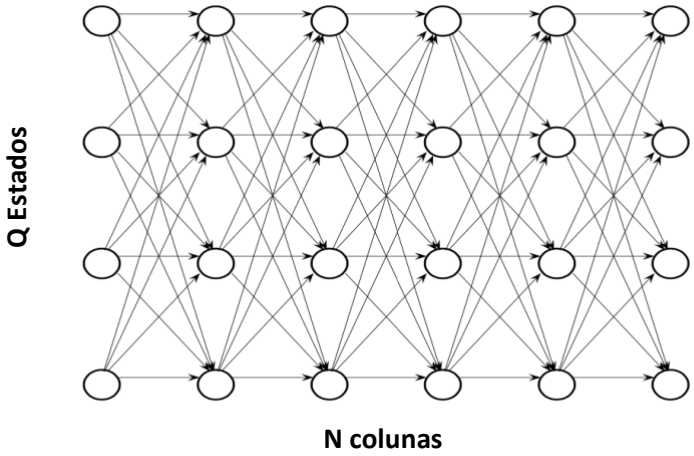


Figura 9 – Manhattan *com N colunas e Q estados. Adaptado de* ***Invalid source specified.***

Considerando que a probabilidade $v_k(i)$ do caminho mais provável acabar no estado k com observação i é conhecida para todos os k, as mesmas probabilidades para a observação x_{i+1} podem ser computadas recursivamente:

$$(7) \qquad v_l(i+1) = e_l(x_{i+1}).max_{k\epsilon Q}\{v_k(i).a_{kl}\}$$

De forma a modular o início da sequência, considera-se um estado inicial fictício (estado 0), impondo-se que todas as possíveis sequências comecem no estado 0, pelo que as condições iniciais são dadas por $v_0(0)$= 1 e $v_k(0)$ = 0. Deste modo a probabilidade para o caminho ótimo π* é dado por:

$$\mathrm{P}\,(\mathrm{x}, \pi^*) = max_{k\epsilon Q}\{v_{k,N}.a_{k,fim}\}$$

No caso analisado tem-se $a_{k,fim} = 1$ para todo k ϵ Q, ou seja, a sequência pode acabar em qualquer um dos três estados.

A abordagem computacional foi desenvolvida no ambiente de programação MATLAB, tendo consistido na construção da função *viterbi*. Esta recebe uma sequência de nucleótidos *x* e devolve o caminho *pi_star*, que tendo em conta o HMM previamente apresentado, maximiza a probabilidade P(x,π) para todos os caminhos π possíveis. O algoritmo baseia-se em duas matrizes. A matriz V contém os valores v_k(i), computados a partir da equação recursiva (7). A segunda, *m_arrows*, tem a mesma dimensão e é utilizada para guardar a origem da aresta que aponta para a entrada de índice respetivo da matriz *V*, ou seja, o estado k-1 que antecede o estado k no caminho gerado. De forma a evitar erros de *underflow* foram guardados os logaritmos das probabilidades de transmissão e de emissão nas respetivas matrizes *a* e *e*. A probabilidade P(x,π) máxima corresponde ao valor máximo da última coluna da matriz V, a partir do qual se faz o *traceback* na matriz *m_arrows,* de forma a obter o caminho final *pi_star*, sendo P(x, π^*) e pi_star os outputs do algoritmo.

Para a sequência de DNA dada S= **CATGCGGGTTATAAC**, o caminho de *hidden states* π* obtido foi:

$$\pi^* = 211222221111112$$

Analisando a sequência π^*, constata-se que o estado 3 não está incluído no processo de gerar a sequência S pelo caminho mais provável. Tendo em conta as probabilidades de transição do modelo, isto seria expectável. Considerando o significado dos estados Q, a sequência de DNA em análise tem maior probabilidade de ser composta exclusivamente por regiões de exões e *start site signals*, não contendo intrões:

- *Exon region* composta pelo nucleótido C;

- Um *Start Site signal* composto pelo dinucleótido AT;

- *Exon region* constituída pelos nucleótidos GCGGG;

- *Start Site signal* com a sequência TTATAAA;

- *Exon region* composta pelo nucleótido C.

c) Processo de avaliação– Algoritmo forward

Dado o mesmo HMM e a mesma sequência de DNA, pretendeu-se obter a probabilidade do modelo gerar S. Este trata-se do problema de avaliação, que quantifica quão bem-adaptado o modelo em causa está à realidade, ou seja, qual a sua capacidade de, com as probabilidades de emissão e transição dadas, produzir uma sequência igual a uma observação real.

De forma a resolver corretamente o problema, consideram-se todos os diferentes caminhos que geram S, pelo que a probabilidade P(S) consiste na soma de todas as probabilidades de todos os caminhos possíveis:

$$P(S) = \sum_{\pi} P(x|\pi)$$

A implementação algorítmica da solução assemelha-se ao algoritmo de Viterbi, substituindo-se em cada iteração a maximização pela soma de todos os estados possíveis, de forma a obter a probabilidade final. O descrito é implementado por um algoritmo de programação dinâmica designado algoritmo Forward, cuja fórmula de recursão é dada por:

$$(8) \qquad f_k(i) = e_k(x_i).\sum_{l \in Q} f_l(i-1)\,.\,a_{lk},$$

sendo $f_{k,i}$ a probabilidade do modelo HMM emitir o prefixo $x_1 ... x_i$ e alcançar o estado π_i=k. Novamente, tem-se em conta as condições inicias $f_0(0)$= 1 e $f_l(0)$ = 0.

Foi implementado o programa *forward*, uma versão alterada da função *viterbi*, onde não são guardadas as arestas dos caminhos e onde a matriz V passa denominar-se F com entradas calculadas pela equação anterior. Deste modo, o resultado pretendido, e que é devolvido, resume-se à soma de todas as entradas da última coluna da matriz F. Para o caso apresentado, com S= **CATGCGGGTTATAAC**, obteve-se:

$$\boldsymbol{P}(\boldsymbol{S}) = 9.386459970616895e - 10$$

Optou-se pelo algoritmo *forward* em detrimento do de Viterbi pois este último apenas considera os caminhos mais prováveis, ocultando todos os restantes com menores

probabilidades, pelo que subestimar-se-ia a probabilidade de se observar a sequência S. Uma outra solução passaria por usar o algoritmo *backward*.

d) Probabilidades posteriores – Decodificação posterior

Pode ser posta outra questão: dada uma sequência de DNA *x* e um modelo HMM, qual a probabilidade $P(\pi_i=k|x)$ do HMM se encontrar no estado k no instante i, designada por probabilidade posterior.

Considerando a definição de probabilidade condicionada e as propriedades do HMM, tem-se:

$$P(\pi_i = k|S) = \frac{P(s_1 \dots s_i, \pi_i = k)P(s_{i+1} \dots s_N|\pi_i = k)}{P(S)}$$

O primeiro termo do numerador corresponde a $f_k(i)$, sendo computado pelo algoritmo *forward* já implementado. O segundo termo é designado por $b_k(i)$ e consiste na probabilidade de o modelo emitir o sufixo $x_{i+1}...x_N$ estando no estado $\pi_i=k$. Esta é obtida por outro algoritmo de programação dinâmica, o algoritmo *backward*. Este é análogo ao *forward*, aplicando uma fórmula recursiva na direção inversa, isto é, começando pelo final da sequência:

$$b_{k,i} = \sum_{l \in Q} e_l(x_{i+1}).\, b_{l,i+1}\,.\, a_{kl}$$

Assim a probabilidade $P(\pi_i=k|x)$ resume-se a:

$$\mathrm{P}(\pi_i = k|x) = \frac{f_k(i).\, b_k(i)}{P(x)}$$

com o estado mais provável a gerar s_i dado por:

$$\hat{\pi}_i = \underset{k}{\operatorname{argmax}}\, P(\pi_i = k|S)$$

Foi então implementado o programa *backward*, uma versão modificada da função *forward* que utiliza a equação previamente apresentada para obter $b_{k,i}$ e construir a matriz de propgramção dinâmica *b*, a qual é devolvida como um dos outputs da função. Deste modo, recorrendo aos programas *forward* e *backward* implementados e acedendo às entradas de interesse das matrizes f e b, é possível calcular as probabilidades pretendidas.

Na figura seguinte ilustram-se os comandos necessários para obter $\mathrm{P}(\pi_2 = 1|S)$:

```
>> [p1,f]=forward('CATGCGGGTTATAAC');
>> [p2,b]=backward('CATGCGGGTTATAAC');
>> (f(1,2)*b(1,2))/p1

ans =

   0.404872262414548

>>
```

Figura 10 – Janela de comandos para cálculo de $P(\pi_2 = 1|S)$

Procedendo da mesma forma para as restantes probabilidades, obteve-se:

$$\mathrm{P}(\pi_2 = 1|S) = 0.404872262414548$$

$$P(\pi_2 = 2|S) = 0.118078799996503$$

$$P(\pi_2 = 3|S) = 0.477048937588949$$

$$P(\pi_9 = 1|S) = 0.493933599819848$$

$$P(\pi_9 = 2|S) = 0.213854021722129$$

$$P(\pi_9 = 3|S) = 0.292212378458023$$

Analisando os resultados obtidos, conclui-se que o modelo HMM estima que s_2 é gerado com maior probabilidade pelo estado 3, seguindo-se o estado 1 e, por último, o 2. Comparando com o previsto na 2ª posição do caminho mais provável π^*, tem-se que π_2^* corresponde ao estado 1. Tendo em que conta que a sequência de estados $\hat{\pi}$ não resulta da maximização da probabilidade conjunta $\mathrm{P}(\mathrm{S}, \pi)$, a mesma pode não ser particularmente provável como um caminho ao longo do modelo todo, podendo até corresponder a um caminho impossível devido às transições do modelo HMM considerado.

No que diz respeito à 9ª posição, o estado 1 apresenta a probabilidade *posteriori* mais elevada, em comparação com os estados 2 e 3. Neste caso, $\hat{\pi}_9$ coincide com π_9^*, reforçando que, dada toda a sequência de DNA, o nucleótido T em análise encontra-se associado a uma região *Start Site signal*.

Bibliografia

[1] "Probabilistic Models" - slides das aulas teóricas disponíveis no fénix 1º semestre 2020/2021.

[2] http://www.bioinformatics.org/sms2/cpg_islands.html - Consultado a 25/10/2020.

[3] https://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/ - Consultado a 25/10/2020.

[4] Takai D., Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A.. Março de 2002..

[5] https://en.wikipedia.org/wiki/Open_reading_frame#cite_note-1 - consultado a 28/10/2020.

[6] https://physoc.onlinelibrary.wiley.com/doi/pdf/10.1113/eph8802514 - Consultado a 30/10/2020.

[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2926779/ - Consultado a 30/10/2020.

[8] N. C. Jones e P. A. Pevzner, “An Introduction to Bionformatics Algorithms,” MIT Press, 2004.

Anexo A

A.I Variação do comprimento da janela

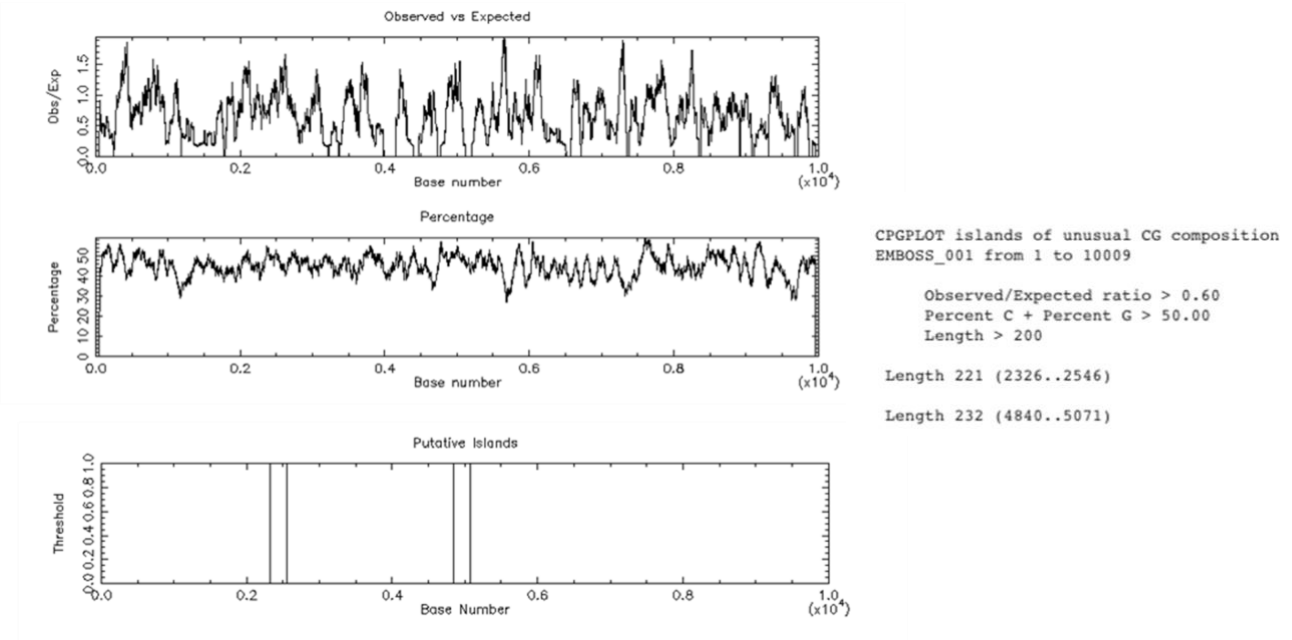


Figura 11 - Resultados obtidos pela aplicação da ferramenta CpG plot, disponível no site do EMBL-EBI, para os seguintes parâmetros: dimensão da janela de 100 bp, comprimento mínimo da sequência de 200 bp, rácio CpG Obs/Exp superior a 0.6 e conteúdo %CG superior a 50%.

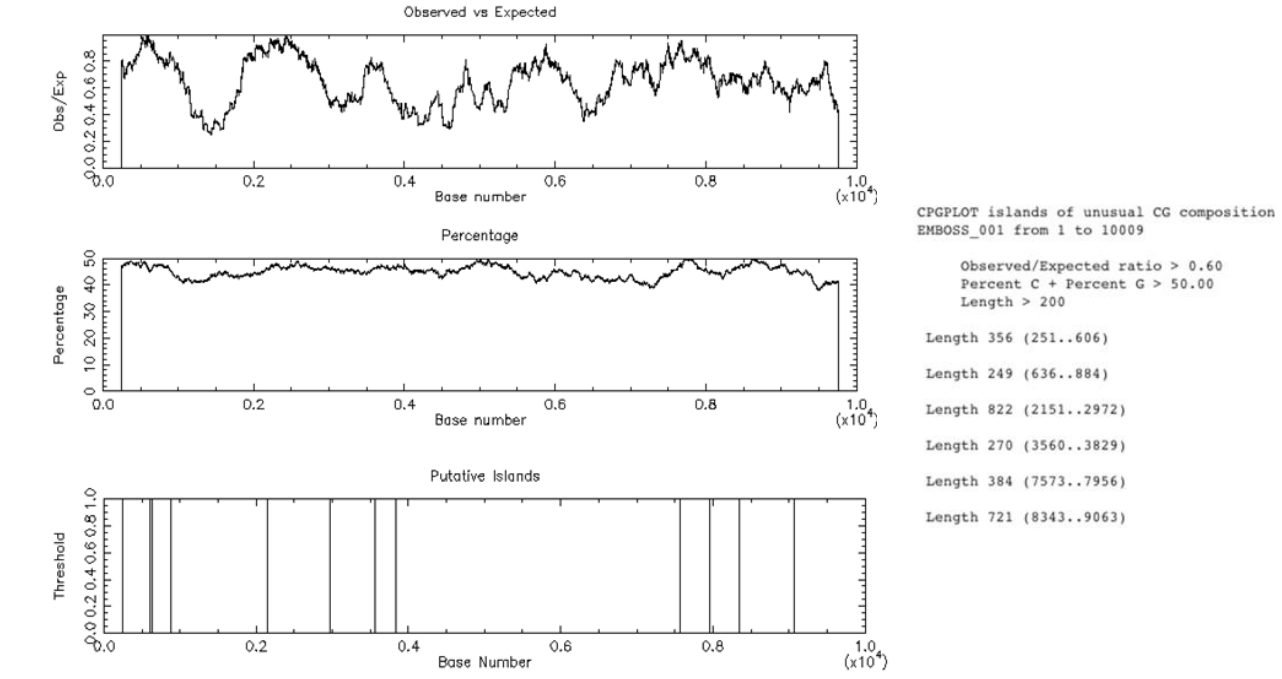


Figura 12 - Resultados obtidos pela aplicação da ferramenta CpG plot, disponível no site do EMBL-EBI, com parâmetros: dimensão da janela de 500 bp, comprimento mínimo da sequência de 200 bp, rácio CpG observados/esperados superior a 0.6 e conteúdo %CG superior a 50%.

A.II Variação do Comprimento Mínimo da CpG Island

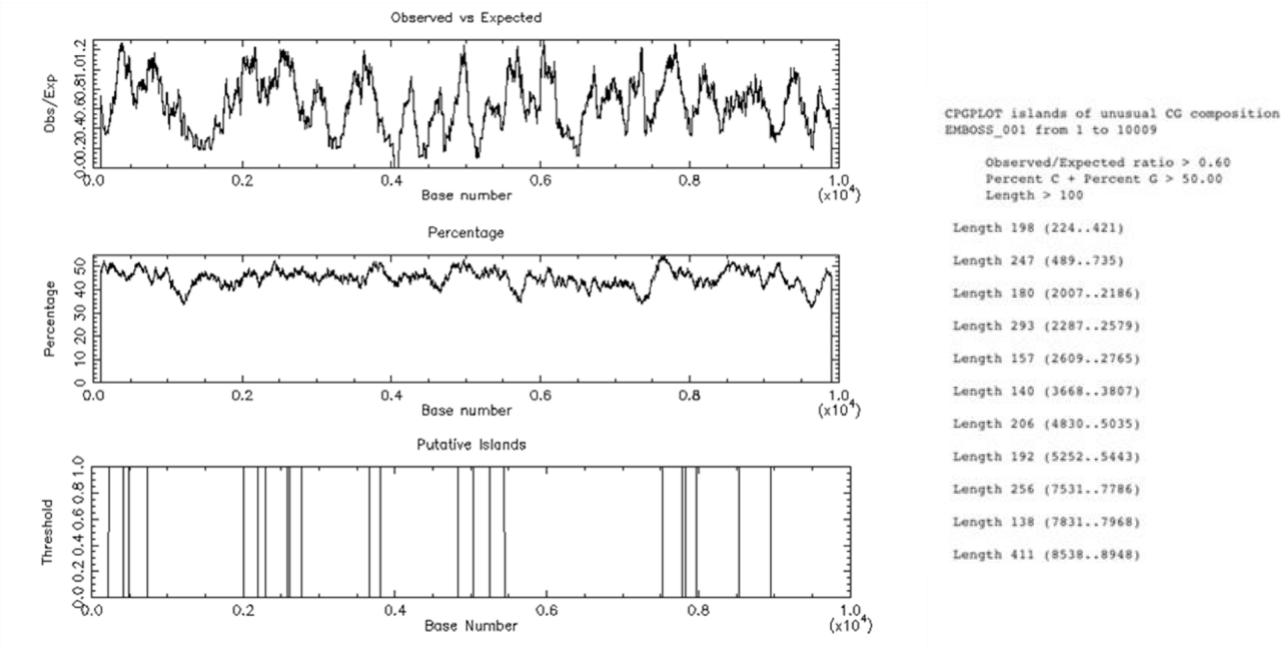
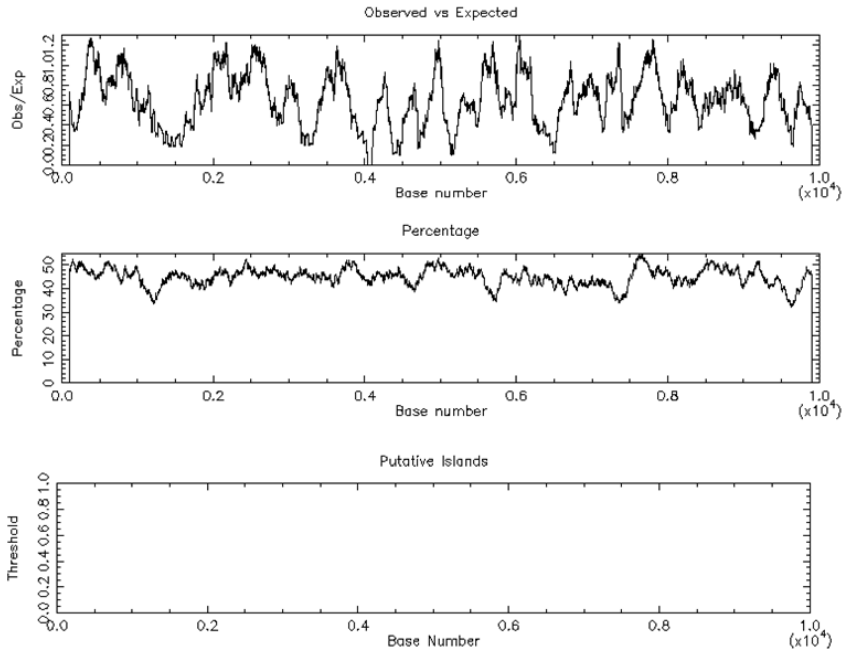


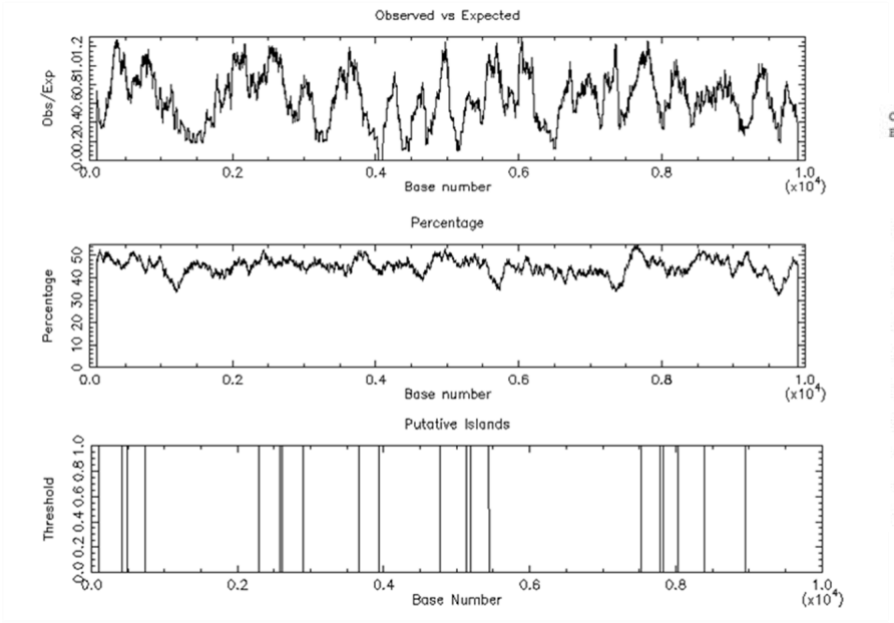
Figura 13 - Resultados obtidos pela aplicação da ferramenta CpG plot, disponível no site do EMBL-EBI, com parâmetros: dimensão da janela de 200 bp, comprimento mínimo da sequência de 100 bp, rácio CpG observados/esperados superior a 0.6 e conteúdo %CG superior a 50%.




```
CPGPLOT islands of unusual CG composition
EMBOSS_001 from 1 to 10009

     Observed/Expected ratio > 0.60
     Percent C + Percent G > 50.00
     Length > 500
```

Figura 14 - Resultados obtidos pela aplicação dal ferramenta CpG plot, disponível no site do EMBL-EBI, com parâmetros: dimensão da janela de 200 pb, comprimento mínimo da sequência de 500 pb, rácio CpG observados/esperados superior a 0.6 e conteúdo %CG superior a 50%.

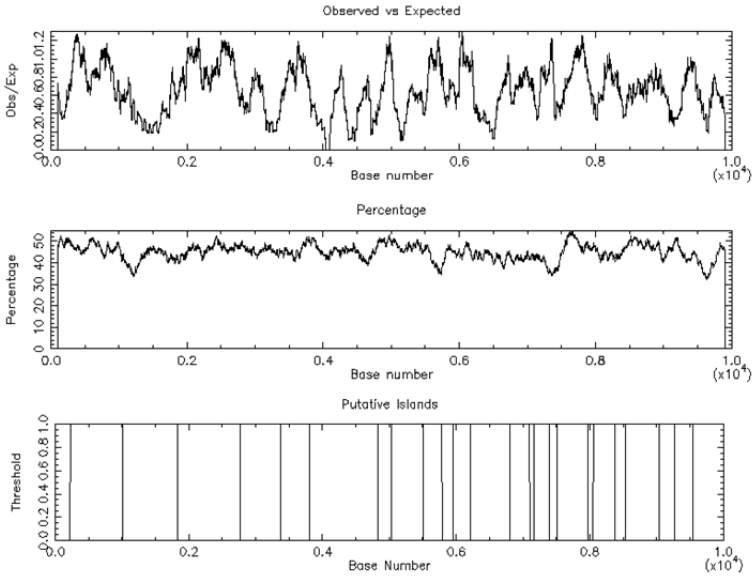
A.III Variação do rácio mínimo observado/esperado




```
CPGPLOT islands of unusual CG composition
EMBOSS_001 from 1 to 10009

     Observed/Expected ratio > 0.20
     Percent C + Percent G > 50.00
     Length > 200

Length 321 (101..421)

Length 247 (489..735)

Length 293 (2287..2579)

Length 287 (2609..2895)

Length 273 (3668..3940)

Length 362 (4769..5130)

Length 251 (5193..5443)

Length 256 (7531..7786)

Length 201 (7831..8031)

Length 556 (8393..8948)
```

Figura 15 - Resultados obtidos pela aplicação da ferramenta CpG plot, disponível no site do EMBL-EBI, com parâmetros: dimensão da janela de 200 pb, comprimento mínimo da sequência de 200 pb, rácio CpG observados/esperados superior a 0.2 e conteúdo %CG superior a 50%.

A.IV Variação da % de conteúdo (C+G)




```
CPGPLOT islands of unusual CG composition
EMBOSS_001 from 1 to 10009

     Observed/Expected ratio > 0.60
     Percent C + Percent G > 20.00
     Length > 200

Length 781 (224..1004)

Length 937 (1829..2765)

Length 427 (3381..3807)

Length 206 (4830..5035)

Length 278 (5512..5789)

Length 259 (5956..6214)

Length 298 (6804..7101)

Length 220 (7173..7392)

Length 452 (7517..7968)

Length 326 (8054..8379)

Length 510 (8538..9047)

Length 278 (9266..9543)
```

Figura 16 - Resultados obtidos pela aplicação da ferramenta CpG plot, disponível no site do EMBL-EBI, com parâmetros: dimensão da janela de 200 bp, comprimento mínimo da sequência de 200 bp, rácio CpG observados/esperados superior a 0.6 e conteúdo %CG superior a 20%.