

Ocorrências de Ilhas CpG em Sequências de DNA

Cleonis V. Figueira^{1†}, Sílvia Regina C. Lopes²

¹ Departamento de Matemática, Universidade Tecnológica Federal do Paraná- PR (UTFPR-PR).

² Departamento de Matemática e Estatística, Universidade Federal do Rio Grande do Sul - RS (UFRGS - RS).

Resumo: O presente trabalho apresenta uma comparação entre três cromossomos de diferentes organismos: cromossomo 12 do *Saccharomyces cerevisiae*, o cromossomo 22 do *Homo sapiens* e o cromossomo 3L da *Drosophila melanogaster*. O objetivo é analisar a incidência de dinucleotídeos CpG em intervalos da sequência de DNA de cada cromossomo, procurando a existência ou não de ilhas CpG (ICpG). Segundo a literatura, a *D. melanogaster* e o *S. cerevisiae* não apresentam metilação ou trazem forte supressão de metilação, uma alteração da citosina do dinucleotídeo CpG que ocorre de forma significativa nas ICpG. A existência de metilação nestas ilhas é questão importante para a pesquisa no âmbito das mutações e patologias relacionadas a diferentes tipos de doenças. Algumas simulações de Monte Carlo da ocorrência de metilação em ICpG são também apresentadas.

Palavras-chave: Sequências de DNA, Ilhas CpG, Metilação, Índice de Incidência de CpG.

Abstract: This paper presents a comparison among three chromosomes from different organisms: *Saccharomyces cerevisiae* chromosome 12, *Homo sapiens* chromosome 22 and *Drosophila melanogaster* chromosome 3L. The aim is to analyze the incidence of dinucleotides CpG at intervals of DNA sequence of each chromosome, seeking for the existence of CpG islands (ICpG). According to the literature, *D. melanogaster* and *S. cerevisiae* show no methylation or bring strong suppression of methylation, a modification of the cytosine dinucleotide ICpG occurring significantly in the ICpG. The existence of these islands methylation are important for research in the context of mutations and pathologies related to different types of diseases. Some Monte Carlo simulations for the occurrence of methylation in ICpG are also presented.

Keywords: DNA Sequences, CpG Islands, Methylation, CpG Incidence Rate.

Introdução

Após sua descoberta, em 1869, pelo bioquímico alemão Johann Friedrich Miescher, o DNA tem sido alvo de estudos em diferentes abordagens e níveis de complexidade. Neste trabalho comparam-se sequências de cromossomos de três diferentes organismos com o objetivo em identificar a ocorrência ou não de ilhas CpG: cromossomo 12 do *Saccharomyces cerevisiae*, o cromossomo 22 do *Homo sapiens* e o cromossomo 3L da *Drosophila melanogaster*. Informações quantitativas e qualitativas sobre estas ilhas são de primordial importância para o estudo de diferentes mutações e tipos de doenças.

Material e métodos

Para as análises consideramos as sequências de DNA em nucleotídeos (pirimidinas são compostas por citosina (C) e timina (T), enquanto que purinas, por adenina (A) e guanina (G)), em formato FASTA, de três organismos do banco de dados do Projeto Ensembl obtidas do site <ftp://ftp.ensembl.org>. As análises foram realizadas através do uso do programa estatístico R (R CORE TEAM, 2013) e seus pacotes.

[†] autor correspondente: cleonis@utfpr.edu.br.

Foram escolhidos o cromossomo 12 do *S. cerevisiae* (fermento de pão), o cromossomo 22 do *H. sapiens* e o cromossomo 3L da *D. melanogaster* (mosca da fruta). As escolhas da mosca da fruta e do fermento de pão foram determinadas pela propriedade de que tais genomas não apresentam metilação ou apresentam forte supressão de metilação em dinucleotídeos CpG (na literatura, o ordenamento do nucleotídeo CG é comumente representado por CpG) e que esta característica é observada em grande parte dos organismos tanto procariontes como eucariontes (ver Bird et al., 1995).

A metilação é uma alteração química/enzimática que afeta, em procariontes, apenas a citosina e é específica para uma sequência CpG, nesta ordem. Ou seja, a citosina encontra-se na posição 5 e a guanina na posição 3 na sequência de nucleotídeos. Segundo Model et al. (2009), as regiões de metilação do DNA podem ser um vasto leque de possibilidades utilizadas para diagnóstico de doenças.

A região de incidência de CpGs é chamada de ilha CpG (denotada aqui por ICpG). A definição mais difundida para uma ICpG é aquela apresentada por Gardiner-Garden e Frommer (1987). A existência de uma ICpG ocorre em uma região com pelo menos 200 pares de bases (pb), com proporção de C+G maior do que 50% e razão de CpG observado e CpG esperado acima de 0,6. A taxa ou índice de ocorrência de ICpG será denotada por λ . Outras variações para a definição de ICpG e discussões sobre estas variações podem ser encontradas em Wu et al. (2010).

A razão λ é obtida dividindo a proporção de dinucleotídeos CpG na região pela proporção esperada caso os nucleotídeos fossem assumidos como resultados independentes de uma distribuição multinomial (ver Wu et al., 2010). Matematicamente, o índice de ocorrência de ICpG pode ser dado por

$$\lambda = \frac{f_D/N}{f_C/N \times f_G/N}, \quad (1)$$

onde N é o número de pares de bases (pb), f_i é a frequência da base $i \in \{C, G\}$ e f_D é a frequência do dinucleotídeo CpG no segmento de DNA considerado.

São apresentadas algumas simulações de Monte Carlo para a análise da ocorrência de metilação em ICpG. Para as simulações das sequências de DNA utilizamos o programa estatístico R e um modelo baseado na distribuição multinomial, onde cada subsequência é um experimento de $N = 200$ ensaios.

Cada ensaio possui cinco possíveis resultados, sendo eles os quatro nucleotídeos C, A, T e G e um quinto elemento, $C_m pG$, representando a ocorrência de citosina metilada em um dinucleotídeo CpG.

As probabilidades de ocorrência de cada nucleotídeo são dadas, respectivamente, por $p_C = r(1 - m)/2$, $p_A = (1 - r)/2$, $p_T = (1 - r)/2$, $p_G = (1 - m)r/2$ e $p_{C_m pG} = rm/2$, onde m representa o índice de metilação, r indica a taxa de C+G e ambos os valores são dados em porcentagens.

O índice m de metilação em mamíferos está entre 70% e 80% das citosinas em ICpGs (ver Antequera e Bird, 1993 e Model et al., 2009). Consideramos o cenário onde $C+G \geq 0,5$ e intervalo de subsequências de $N = 200$ pb, para satisfazer as exigências da definição de uma ICpG.

Resultados e discussões

A Figura 1 apresenta a composição em nucleotídeos e dinucleotídeos de cada cromossomo em estudo. Em todos os casos pode-se perceber que há menor incidência de nucleotídeos C e G, sendo que, no caso da *D. melanogaster* (mosca da fruta), esta característica não é tão acentuada. Entretanto, para o caso em que se observam os dinucleotídeos, há evidência de menor incidência de CpG nos cromossomos do *S. cerevisiae* (fermento de pão) e do *H. sapiens*.

A Figura 2 apresenta a incidência de CpGs, (linhas verticais) em um espaço de 10000pb de nucleotídeos. Tem-se uma relativa homogeneidade das barras verticais nos cromossomos de *S. cerevisiae* e *D. melanogaster*, o que não ocorre no intervalo de 10000pb do cromossomo de *H. sapiens*.

Assim, já do indicativo da baixa concentração relativa de dinucleotídeos CpG nos cromossomos (ver Figura 1) procura-se verificar como estes dinucleotídeos estão dispersos nos cromossomos em análise e se chegam a se organizar de forma a caracterizar uma região de ICpG.

Dessa forma, desejamos identificar os sítios de ocorrência das CpGs. Para tal, precisa-se, inicialmente que haja no intervalo de 200pb a taxa de pelo menos 50% de nucleotídeos C ou G, representada por C+G, que está apresentada na Figura 3.

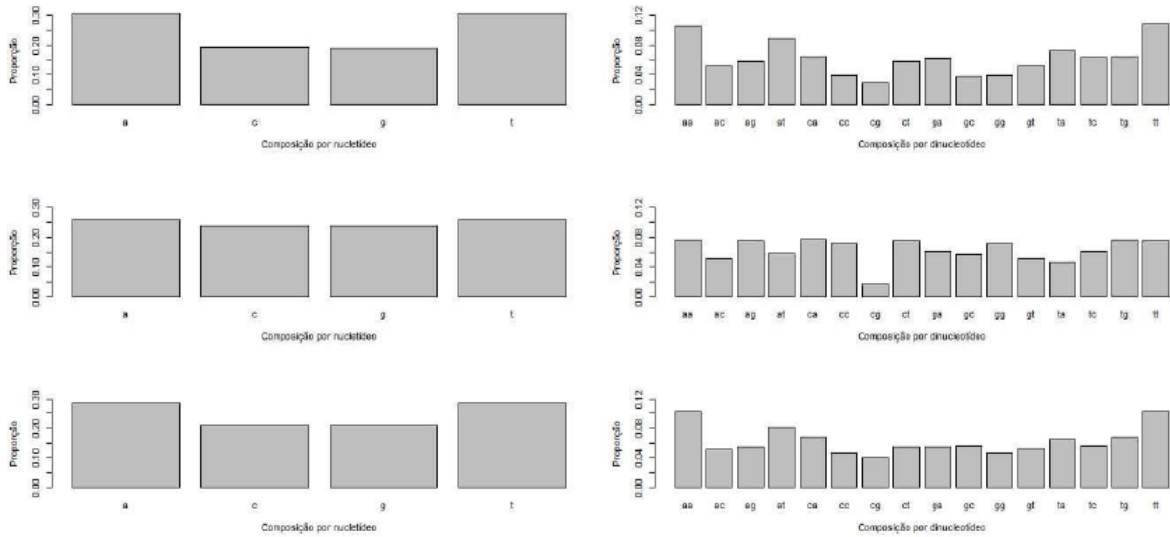


Figura 1: Composição dos Cromossomos: cromossomo 12 do *S. cerevisiae* (fermento de pão, no topo); cromossomo 22 do *H. sapiens* (no centro); cromossomo 3L da *D. melanogaster* (mosca da fruta).

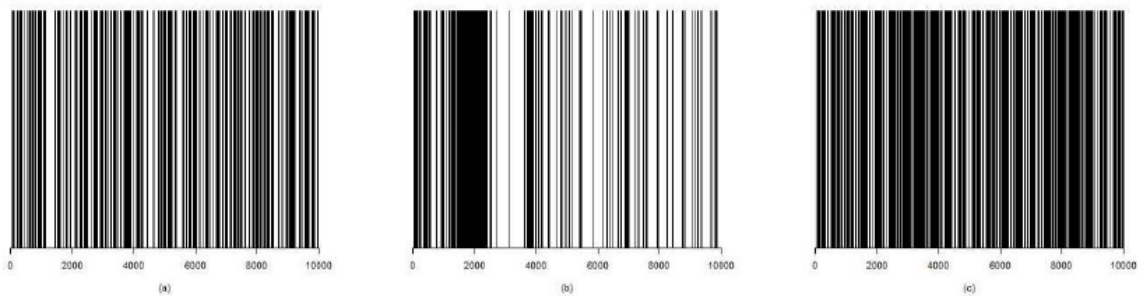


Figura 2: Sítios de CpG em intervalo de 10000pb: (a) cromossomo 12 do *S. cerevisiae* (fermento do pão); (b) cromossomo 22 do *H. sapiens*; (c) cromossomo 3L da *D. melanogaster* (mosca da fruta).

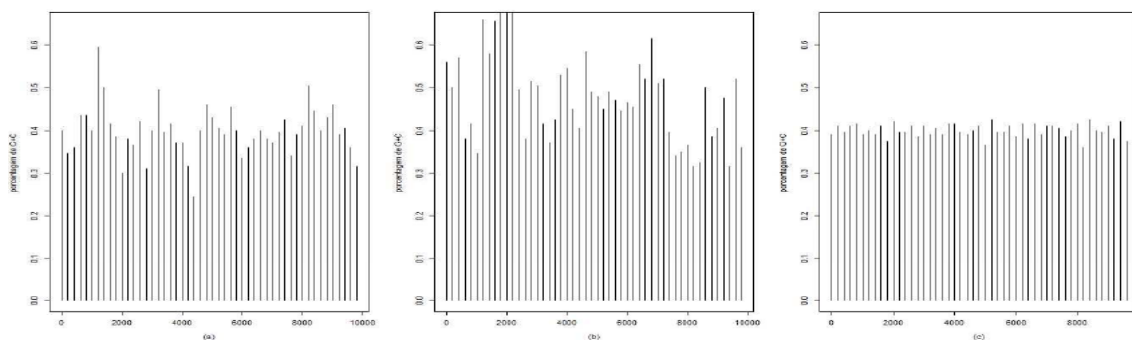


Figura 3: Proporção de C+G: (a) cromossomo 12 do *S. cerevisiae* (fermento do pão); (b) cromossomo 22 do *H. sapiens*; (c) cromossomo 3L da *D. melanogaster* (mosca da fruta).

Com base nas mesmas sequências de 10000pb para cada cromossomo em estudo, investigamos a incidência de ICpG. Utilizamos a fórmula dada pela expressão (1), com $N = 200pb$. A cada 200pb, obteve-se a frequência de C, G e CpG e, desta forma, foi calculado o índice λ de ocorrência de ICpG. Estas informações estão listadas nas Tabelas 2 e 3 (cada linha, $n \in \{1, \dots, 25\}$ para a Tabela 2 e $n \in \{26, \dots, 50\}$ para a Tabela 3, representa uma subsequência 200pb, $f_i(j)$ é a frequência do nucleotídeo/dinucleotídeo $i \in \{C, G, CpG\}$) para o cromossomo $j \in \{(a) S. cerevisiae, (b) H. sapiens, (c) D. melanogaster\}$.

As Tabelas 2 e 3 apresentam a análise da existência ou não de ICpG para um intervalo total de 10000pb divididos em subintervalos de 200pb dos três cromossomos em estudo. Sendo que os primeiros 5000pb estão apresentados na Tabela 2 e os 5000pb subsequentes estão na Tabela 3.

As informações destacadas em negrito, nas Tabelas 2 e 3, trazem locais onde ocorreram, de forma simultânea, a proporção de C + G acima de 50% e o índice λ de ocorrência de ICpG acima de 0,6.

Foram simuladas sequências com $N = 200pb$, com índice m de metilação, para $m \in \{0,5; 0,7; 0,8\}$, e com índice r de C+G, para $r \in \{0,5; 0,6; 0,7\}$. Tais simulações, e respectivas sequências, são representadas na Figura 4.

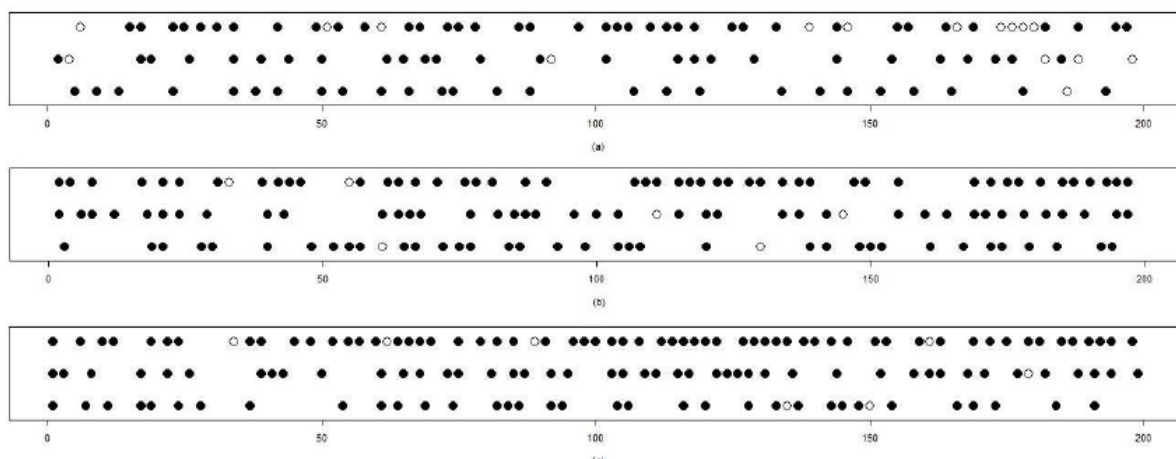


Figura 4: Ocorrências de CpG em intervalo de 200pb: (a) $m = 0,5$; (b) $m = 0,7$; (c) $m = 0,8$. Cada gráfico apresenta $r = 0,7$ (topo), $r = 0,6$ (centro) e $r = 0,5$ (embaixo). CpG é representada por \circ , $C_m pG$ é representada por \bullet .

A Figura 4 representa a simulação de sequências de nucleotídeos com $m = 0,5$ para o gráfico (a), $m = 0,7$ para o gráfico (b) e $m = 0,8$ para o gráfico (c). Em todos os gráficos são apresentados três valores para a taxa de C + G: $r = 0,5$ (embaixo), $r = 0,6$ (centro) e $r = 0,7$ (topo). As Figuras 4(a) e 4(c) mostram que o aumento de r de 0,5 para 0,7 resulta em aumento ou estabilidade da ocorrência de CpG não metilada. Para a Figura 4(c) houve a diminuição da ocorrência de CpG quando comparada com as sequências do centro ($r = 0,6$) e topo ($r = 0,7$).

A Tabela 1 apresenta o índice λ de ocorrência de ICpG, dado pela expressão (1), para cada caso apresentado na Figura 4. Em todos os casos, o valor obtido ficou acima de 0,6 evidenciando que as sequências simuladas correspondem a regiões de ICpG.

Tabela 1: Resultados do índice λ de ocorrência de ICpG, para cada um dos casos simulados.

$m \setminus r$	$r = 0,5$	$r = 0,6$	$r = 0,7$
$m = 0,5$	6,109	3,726	3,654
$m = 0,7$	7,650	11,728	7,650
$m = 0,8$	15,652	19,287	11,532

Tabela 2: Resultado do índice de ocorrência de ICpG (λ): (a) cromossomo 12 do *S.serevisiae* (fermento do pão); (b) cromossomo 22 do *H. sapiens*; (c) cromossomo 3L da *D. melanogaster* (mosca da fruta) em um intervalo total de 5000pb e subintervalos de tamanho $N = 200pb$.

n	$f_{C^{(a)}}$	$f_{G^{(a)}}$	$f_{D^{(a)}}$	$\lambda_{(a)}$	$C + G^{(a)}$	$f_{C^{(b)}}$	$f_{G^{(b)}}$	$f_{D^{(b)}}$	$\lambda_{(b)}$	$C + G^{(b)}$	$f_{C^{(c)}}$	$f_{G^{(c)}}$	$f_{D^{(c)}}$	$\lambda_{(c)}$	$C + G^{(c)}$
1	81	13	3	0,570	0,400	52	59	8	0,522	0,555	32	46	10	1,359	0,390
2	57	21	5	0,835	0,350	49	52	6	0,471	0,505	46	36	9	1,087	0,410
3	55	17	5	1,070	0,355	50	63	9	0,571	0,565	36	43	8	1,034	0,395
4	54	40	9	0,833	0,440	43	33	2	0,282	0,380	40	42	10	1,190	0,410
5	66	36	13	1,094	0,435	55	29	6	0,752	0,420	44	39	7	0,816	0,415
6	52	32	4	0,481	0,400	46	23	2	0,378	0,345	38	40	10	1,316	0,390
7	62	29	7	0,779	0,595	92	39	10	0,557	0,655	44	36	9	1,136	0,400
8	44	36	7	0,884	0,500	60	57	12	0,702	0,585	36	42	8	1,058	0,390
9	34	58	3	0,304	0,410	70	61	10	0,468	0,655	44	38	9	1,077	0,410
10	29	67	6	0,618	0,385	76	68	18	0,697	0,720	36	39	8	1,140	0,375
11	35	70	6	0,490	0,300	82	69	19	0,672	0,755	42	42	8	0,907	0,420
12	56	42	6	0,510	0,380	79	86	20	0,589	0,825	38	41	8	1,027	0,395
13	59	34	7	0,698	0,365	37	62	4	0,349	0,495	40	39	11	1,410	0,395
14	57	28	6	0,752	0,425	37	39	1	0,139	0,380	45	37	7	0,841	0,410
15	46	40	5	0,543	0,305	60	43	0	0,000	0,515	40	37	9	1,216	0,385
16	61	43	12	0,915	0,400	55	46	1	0,079	0,505	44	38	7	0,837	0,410
17	50	34	6	0,706	0,500	48	35	0	0,000	0,415	32	46	10	1,359	0,390
18	54	46	10	0,805	0,390	30	45	0	0,000	0,375	46	35	9	1,118	0,405
19	51	46	12	1,023	0,415	42	43	7	0,775	0,425	35	43	7	0,930	0,390
20	47	48	18	1,596	0,370	51	54	6	0,436	0,525	40	43	11	1,279	0,415
21	46	38	9	1,030	0,370	63	46	2	0,138	0,545	45	38	7	0,819	0,415
22	46	45	9	0,870	0,315	49	42	3	0,292	0,455	38	41	10	1,284	0,395
23	37	40	8	1,081	0,245	41	39	1	0,125	0,400	43	35	8	1,063	0,390
24	45	36	10	1,235	0,405	69	48	2	0,121	0,585	37	43	9	1,131	0,400
25	33	31	3	0,587	0,455	58	40	3	0,259	0,490	44	38	9	1,077	0,410

Nota: cada linha $n \in \{1, \dots, 25\}$ representa uma subsequência 200pb, $f_i(j)$ é a frequência do nucleotídeo/dinucleotídeo $i \in \{C, G, CpG\}$ para o cromossomo $j \in \{(a)S. cerevisiae, (b)H. sapiens, (c)D. melanogaster\}$.

Tabela 3: Resultado do índice de ocorrência de ICpG (λ): (a) cromossomo 12 do *S. cerevisiae* (fermento do pão); (b) cromossomo 22 do *H. sapiens*; (c) cromossomo 3L da *D. melanogaster* (mosca da fruta) em um intervalo total de 5000pb e subintervalos de tamanho $N = 200pb$.

n	$f_{C^{(a)}}$	$f_{G^{(a)}}$	$f_{D^{(a)}}$	$\lambda_{(a)}$	$C+G_{(a)}$	$f_{C^{(b)}}$	$f_{G^{(b)}}$	$f_{D^{(b)}}$	$\lambda_{(b)}$	$C+G_{(b)}$	$f_{C^{(c)}}$	$f_{G^{(c)}}$	$f_{D^{(c)}}$	$\lambda_{(c)}$	$C+G_{(c)}$
26	41	39	8	1,001	0,435	51	45	4	0,349	0,480	35	38	7	1,053	0,365
27	49	38	13	1,396	0,400	48	43	1	0,097	0,455	43	42	9	0,997	0,425
28	52	30	8	1,026	0,395	65	32	2	0,192	0,485	38	41	8	1,027	0,395
29	77	16	4	0,649	0,450	48	47	0	0,000	0,475	40	39	11	1,410	0,395
30	62	21	4	0,614	0,400	36	53	1	0,105	0,445	45	37	7	0,841	0,410
31	50	14	5	1,429	0,340	45	47	1	0,095	0,460	39	38	9	1,215	0,385
32	57	37	10	0,948	0,360	53	38	2	0,199	0,455	45	38	8	0,936	0,415
33	60	42	12	0,952	0,375	43	68	1	0,068	0,555	31	45	10	1,434	0,380
34	62	27	7	0,836	0,405	55	50	3	0,218	0,525	47	36	9	1,064	0,415
35	56	31	6	0,691	0,380	70	52	10	0,549	0,610	35	43	7	0,930	0,390
36	44	40	5	0,568	0,365	49	53	1	0,077	0,510	39	43	11	1,312	0,410
37	42	50	5	0,476	0,400	56	48	3	0,223	0,520	45	37	6	0,721	0,410
38	31	72	7	0,627	0,425	41	38	3	0,385	0,395	39	42	11	1,343	0,405
39	40	59	5	0,424	0,335	31	37	0	0,000	0,340	43	34	8	1,094	0,385
40	59	39	5	0,435	0,395	41	29	1	0,168	0,350	36	44	9	1,136	0,400
41	53	29	6	0,781	0,410	33	40	1	0,152	0,365	45	38	9	1,053	0,415
42	61	34	8	0,771	0,505	25	39	1	0,205	0,320	34	38	7	1,084	0,360
43	49	44	8	0,742	0,440	32	32	2	0,391	0,320	43	42	10	1,107	0,425
44	52	41	8	0,750	0,400	49	51	3	0,240	0,500	39	41	8	1,001	0,400
45	57	33	8	0,851	0,435	32	45	1	0,139	0,385	40	39	11	1,410	0,395
46	50	51	8	0,627	0,460	40	41	2	0,244	0,405	45	37	7	0,841	0,410
47	53	46	14	1,148	0,390	38	58	5	0,454	0,480	38	38	9	1,247	0,380
48	48	40	15	1,563	0,405	27	36	0	0,000	0,315	46	38	8	0,915	0,420
49	43	41	7	0,794	0,355	47	57	3	0,224	0,520	31	44	10	1,466	0,375
50	42	45	11	1,164	0,320	32	40	2	0,313	0,360	46	37	9	1,058	0,415

Nota: cada linha $n \in \{26, \dots, 50\}$ representa uma subsequência 200pb, $f_i(j)$ é a frequência do nucleotídeo/dinucleotídeo $i \in \{C, G, CpG\}$ para o cromossomo $j \in \{(a)S. cerevisiae, (b)H. sapiens, (c)D. melanogaster\}$.

Conclusões

Com base no exposto, e observando as Figuras 2 e 3 e as Tabelas 2 e 3, para o cromossomo 12 da *S.serevisiae* os valores obtidos para λ nos intervalos de pares de bases $n = 7$ e $n = 42$ ocorrem de forma esparsa não caracterizando uma região de ICpG. Para o caso do cromossomo humano considerado, existe evidência da ocorrência de ICpGs no intervalo entre $n = 8$ e $n = 11$ pois há porcentagem de C+G acima de 50%, a razão $\lambda_{(b)}$ ocorre acima de 0,6 e os intervalos são próximos ou adjacentes. Há também indicativo da supressão desta característica para o cromossomo 3L da *D. melanogaster*, no intervalo de 10000pb em análise, pois não há nenhum intervalo n que satisfaça as exigências para a existência de uma ICpG. Pelos resultados apresentados na Tabela 1, todas as simulações realizadas, baseadas no modelo multinomial, correspondem à definição de ICpG.

Agradecimentos

Cleonis V. Figueira agradece ao PPG-MAT/UFRGS pelo apoio para a participação da 58ª Reunião Anual da RBras e 15º SEAGRO. Sílvia R.C. Lopes agradece ao CNPq-Brasil e ao INCT em Matemática.

Referências

ANTEQUERA, F.; BIRD, A. Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, v. 90, p. 11995-11999, 1993.

BIRD, A.; TATE, P.; NAN, X.; CAMPOY, J.; MEEHAN, R.; CROSS, S.; TWEEDIE, S.; CHARLTON, J.; MACLEOD, D. *Studies of DNA methylation in animals. Journal of Cell Science*, v. 19, p. 37-39, 1995.

GARDINER-GARDEN, M.; FROMMER, M. CpG islands in vertebrate genomes. *Journal of Molecular Biology*, v. 196, p. 261-282, 1987.

MODEL, F.; LEWIN, J.; LOFTON-DAY, C.; WEISS, G. *Analysis of DNA methylations in cancer. in: Wiuf, C. and Andersen, C.L. (orgs.), Statistics and Informatics in Molecular Cancer Research. Oxford: Oxford University Press. 2009. 217p.*

R CORE TEAM. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.*

WU, H.; CAFFO, B.; JAFFEE, H.A.; IRIZARRY, R.A.; FIENBERG, A.P. *Redefining CpG islands using hidden Markov models. Biostatistics, Oxford Journals*, v. 11, n. 3, p. 499-514, 2010.