



UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ODONTOLOGIA DE PIRACICABA



Comparação das variações no valor da predição do padrão de clivagem química de radicais hidroxila de DNA dupla fita em sítios mutantes funcionais de regiões codificantes e não codificantes.

Neliane Cora Vellozo

Piracicaba

2010

Faculdade de Odontologia de Piracicaba - UNICAMP

Neliane Cora Vellozo

Comparação das variações no valor da predição do padrão de clivagem química de radicais hidroxila de DNA dupla fita em sítios mutantes funcionais de regiões codificantes e não codificantes.

Monografia apresentada ao curso de Odontologia da Faculdade de Odontologia de Piracicaba – UNICAMP, para a obtenção do diploma de Cirurgião-Dentista

Orientador: Professor Sergio Roberto Peres Line

Piracicaba
2010

**FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA FACULDADE DE ODONTOLOGIA DE PIRACICABA**
Bibliotecária: Elis Regina Alves dos Santos – CRB-8^a. / 8099

V546c

Vellozo, Neliane Cora.

Comparação das variações no valor da predição do padrão de clivagem química de radicais hidroxila de DNA dupla fita em sítios mutantes funcionais de regiões codificantes e não-codificantes / Neliane Cora Vellozo. -- Piracicaba, SP: [s.n.], 2010.

47f. : il.

Orientador: Sérgio Roberto Peres Line.

Monografia (Graduação) – Universidade Estadual de Campinas, Faculdade de Odontologia de Piracicaba.

1. Genética. 2. Mutação. I. Line, Sérgio Roberto Peres. II. Universidade Estadual de Campinas. Faculdade de Odontologia de Piracicaba. III. Título.

(eras/fop)

Sumário

	Página
Resumo.....	5
Abstract.....	7
Introdução.....	9
Revisão da Literatura.....	11
Proposição.....	12
Material e Métodos.....	13
Resultados e Discussão.....	23
Conclusão.....	43
Referências.....	46

Resumo

A ativação da transcrição gênica é um fenômeno bastante complexo, sendo provavelmente a principal etapa onde ocorre o controle da expressão gênica. Em termos gerais a transcrição ocorre pela ligação de proteínas denominadas de fatores de transcrição com seqüências específicas do DNA, chamadas de regiões ou seqüências “cis”. A interação DNA-proteína pode depender da conformação do DNA nestas seqüências. O presente trabalho tem por objetivo buscar evidências da importância da conformação estrutural do DNA dupla fita de seqüências “cis” localizadas em regiões não codificantes na expressão gênica através da análise da conservação filogenética dos padrões de clivagem. Além de estudar o efeito de mutações genéticas na estrutura do DNA dupla fita e investigar possível associação entre a estrutura do DNA dupla fita e a propensão para o aparecimento de mutações genéticas.

Foram utilizadas mutações do tipo substituição em regiões não codificadoras (até - 2000 pares de bases do início da tradução) de genes que causam alterações fenotípicas evidentes resultando em estados patológicos em genes humanos e as seqüências flanqueadoras destes sítios mutantes. Foram também obtidas seqüências do tipo substituição de mutações nas regiões codificadoras destes genes. Utilizaram-se as bases de dados (Pubmed e HGMD <http://www.hgmd.cf.ac.uk/ac/index.php>). A importância da estrutura do DNA dupla fita foi avaliada por três análises distintas: 1- Análise da conservação filogenética dos padrões de clivagem. 2- Influência do tipo de mutação na alteração da estrutura do DNA. 3- Influência da estrutura do DNA dupla fita na propensão para mutações.

Os resultados mostraram que as mutações que causam alterações fenotípicas em humanos não alteram a estrutura do DNA de forma significativamente, diferente das variações que ocorrem entre humanos e outros animais. Também não foram encontradas diferenças significativas quando se comparou a estrutura do DNA normal e mutado nas regiões codificadoras e promotoras. O DNA se torna mais propenso a mutação quando ocorre a formação de fita simples durante a transcrição ou replicação. Além disso, a propensão para mutagenicidade também pode ser dependente da estrutura do DNA dupla fita. E as diferenças na proporção das bases no genoma (Chargaff et al 1949) são principalmente atribuídas a maior mutagenicidade da base C. Nossos resultados sugerem que a menor probabilidade de sofrer mutações nas bases A e T em relação às bases C e

G, (principalmente a base A), pode também ter sido um fator que contribuiu para que a proporção destas bases seja maior do que a proporção das bases C e G.

A análise estatística mostrou que a variação no valor de predição de clivagem química do DNA diferem do valor esperado (zero) ($p < 0,0001$), sendo o valor amostral médio de -0.2 e erro padrão de 0,05. Isto indica que o padrão das mutações não ocorre ao acaso, havendo uma diminuição média de 0,2 na probabilidade de clivagem química das bases mutantes em relação às bases normais.

Palavras-chave: DNA, mutações, padrão de clivagem.

Abstract

The activation of gene transcription is a very complex phenomenon, being probably the main stage where there is the control of gene expression. Overall transcription occurs through binding proteins called transcription factors with specific sequences of DNA, called regions or sequences "cis". The DNA-protein interaction may depend on the conformation of DNA sequences. This study aims to find evidence of the importance of structural conformation of double-stranded DNA sequences "cis" located in non-coding regions on gene expression through analysis of phylogenetic conservation of cleavage patterns. Besides studying the effect of genetic mutations in the structure of double-stranded DNA and investigate possible association between double-stranded DNA structure and the propensity for the emergence of genetic mutations.

Substitution mutations in non-coding regions (up to -2000 bp of the start of translation) were used of genes that cause obvious phenotypic changes resulting in pathological states in human genes and flanking sequences of these sites mutated. We also obtained sequences of the type of replacement mutations in coding regions of genes. We used the databases (Pubmed and <http://www.hgmd.cf.ac.uk/ac/index.php> HGMD). The importance of the structure of double-stranded DNA was assessed by three different tests: 1 - Analysis of phylogenetic conservation of cleavage patterns. 2 - Influence of mutation in altering the structure of DNA. 3 - Influence of the structure of double-stranded DNA in the propensity for mutations.

The results showed that mutations that cause phenotypic changes in humans alter the structure of DNA in a manner significantly different from the variations that occur between humans and other animals. We also found no significant differences when comparing the structure of normal and mutated DNA in coding regions and promoters. The DNA becomes more prone to mutation occurs when the formation of single stranded during transcription or replication. Moreover, the propensity for mutagenicity may also be dependent on the structure of DNA double strand. And the differences in the proportion of bases in the genome (Chargaff et al 1949) are mainly attributed to higher mutagenicity of base C. Our results suggest that less likely to mutate A and T bases in relation to the bases C and G, (especially the base), may also have been a contributing factor to the proportion of these bases is greater than the proportion bases C and G.

Statistical analysis showed that the variation in the predictive value of chemical cleavage of DNA differ from the expected value (zero) ($p < 0.0001$) with the average

sample value of -0.2 and standard error of 0.05. This indicates that the pattern of mutations occur randomly, with an average decrease of 0.2 in the probability of chemical cleavage of mutant bases in relation to the normal bases.

Keywords: DNA, mutations, cleavage pattern.

1. Introdução

O maior avanço na área científica nos últimos anos foi o seqüenciamento do genoma humano e de várias outras espécies. Boa parte do conhecimento gerado está sendo catalogado e depositado em bases de dados com livre acesso à comunidade científica. Este fato abriu o leque de possibilidades de estudos na área de genética médica e molecular. Uma das áreas mais interessante e ainda pouco compreendida é a regulação da iniciação da transcrição gênica. A ativação da transcrição gênica é um fenômeno bastante complexo, sendo provavelmente a principal etapa onde ocorre o controle da expressão gênica. Em termos gerais, ocorre pela ligação de proteínas denominadas de fatores de transcrição com seqüências específicas do DNA, chamadas de regiões ou seqüências “cis”. Estes fatores de transcrição podem se ligar ao DNA e a DNA polimerase ou a outros fatores. Estas interações vão, em última análise, influenciar a função da enzima DNA polimerase determinando não só a iniciação da transcrição, mas também a taxa de transcrição (i.e. quanto RNA vai ser transcrito num espaço de tempo). A ligação dos fatores de transcrição ao DNA e as outras proteínas do complexo transcricional pode ser influenciada por fatores externos. Desta maneira, alguns fatores de transcrição precisam se ligar a co-fatores como a vitamina A, vitamina D ou hormônios corticóides, provenientes do ambiente externo ou de outras células, para se ligarem ao DNA. Outros fatores precisam sofrer alterações estruturais que ocorrem pela fosforilação de aminoácidos específicos (Raven et al 2004).

Em sua grande maioria a determinação das seqüências “cis” é feita por métodos computacionais funcionais, análise filogenética comparativa, ou por meio de análises laboratoriais feitas “in vitro” utilizando-se fragmentos de DNA contendo a região desejada (Kelis et al 2003, Harbison et al 2004, Nikolay et al 2007). No entanto, a maneira mais direta e confiável de se determinar a importância de uma seqüência na regulação da transcrição é simplesmente observar se mutações genéticas nestas seqüências produzem alterações fenotípicas. Desta maneira, se uma mutação em uma região reguladora da transcrição (não codificadora) está associada a alterações fenotípicas, pode-se concluir que aquela região é importante para a regulação da transcrição gênica. Mutações em seqüências regulatórias que causam alterações fenotípicas evidentes são bem menos comuns do que mutações em regiões codificadoras. Dados de estudos relatando associação entre polimorfismos genéticos em regiões não codificadoras e tendência para desenvolvimento de doenças devem ser interpretados com cautela, no que diz respeito a

influência da região polimórfica na regulação da transcrição. Na grande maioria das vezes estes estudos mostram associações entre polimorfismos genéticos e tendência para o desenvolvimento de doença. Nestes casos o efeito da região polimórfica não é evidente em todos os indivíduos que carregam o alelo que confere susceptibilidade, sendo dependente do background genético. Desta maneira, espera-se que polimorfismos genéticos em regiões não codificadoras tenham um efeito na atividade transcricional menos acentuado do que mutações que são transmitidas como caráter mendeliano e que causam alterações fenotípicas evidentes em todos os casos onde estão presentes.

As seqüências “cis” têm tamanho pequeno, variando entre 5 e 20 pares de bases (Nikolay et al 2007) e podem se localizar próximos ao sítio de iniciação de transcrição em uma região chamada de promotora ou a milhares de pares de base do início da transcrição em regiões denominadas de “enhancer”, que aumentam a taxa de transcrição ou “silencer” que diminui ou inibem a transcrição (Raven et al 2004). A interação dos fatores de transcrição com os sítios no DNA depende não só das bases onde ocorre o contato, mas pode depender também das bases vizinhas (Sarai e Kono 2005, Faiger et al 2007). Isto pode ser explicado pelo fato de que a interação DNA-proteína pode depender da conformação do DNA, que pode ser influenciada pelas bases vizinhas. O crescente interesse pela conformação estrutural do DNA tem levado ao desenvolvimento de ferramentas computacionais capazes de transformar padrões estruturais do DNA dupla fita em valores numéricos. Uma destas ferramentas denominada de ORChID (<http://dna.bu.edu/orchid/>) faz a predição do padrão de clivagem químico de radicais hidroxila de DNA dupla fita (Price e Tullius 1992, Greenbaum et al 2007). Apesar de não mostrar a estrutura espacial tridimensional do DNA o padrão de clivagem química dos radicais hidroxila reflete a área superficial acessível ao solvente durante a reação química, que por sua vez é um importante parâmetro estrutural (Balasubramanian et al. 1998). Desta maneira o padrão de clivagem relaciona-se com a estrutura tridimensional da molécula de DNA e como esta varia em relação às seqüências de nucleotídeos (Greenbaum et al 2007).

O presente trabalho teve como objetivo investigar a importância da conformação estrutural do DNA dupla fita de seqüências “cis” localizadas em regiões não codificadoras na expressão gênica e na propensão para desenvolvimento de mutações. Foi estudado ainda se mutações em diferentes bases afetariam de forma diferente a estrutura do DNA dupla fita. Também foram analisados os valores de predição de clivagem química das seqüências que não estavam pareadas.

2. Revisão da Literatura

Não foi encontrado na literatura nenhum trabalho correlacionando a estrutura do DNA dupla fita com a propensão para mutação. No entanto, existe farta literatura mostrando que a estrutura do DNA pode influenciar na propensão para mutações durante a transcrição gênica. A mutabilidade do DNA é maior quando o mesmo encontra-se na forma de fita simples durante a transcrição (Wright et al. 2002, Wright et al. 2003, Todd and Glickman 1982, Kruisselbrink et al. 2008). Desta maneira, a estrutura secundária do DNA simples fita influencia na mutabilidade principalmente em regiões codificadoras. Em regiões onde ocorre o pareamento com formação de fita dupla intracadeia o DNA é 'protegido' possuindo menor tendência para mutagenicidade. Além disto, durante a transcrição, a enzima RNA polimerase desestabiliza o DNA dupla fita introduzindo um aumento na torção da dupla fita do DNA (supercoiling) (Wright 2000). O 'supercoiling' pode se estender por até 800 pares de base acima e abaixo da RNA polimerase (Krasilnikov et al. 1999), aumentando a probabilidade de mutações em regiões não traduzidas próximas ao sítio de início da transcrição. Experimentos feitos com DNA plasmideano mostram que existe uma tendência para o aparecimento de estruturas conhecidas como 'stem loop' em DNA dupla fita com estrutura 'supercoiled'. As estruturas tipo 'stem loop' possuem DNA simples fita em sua extremidade, que por sua vez tem maior propensão para mutações. Usando como modelo a bactéria *E. coli*, (Hoede et al. 2006) mostraram que existe uma pressão seletiva para a diminuição do número de regiões com tendência a sofrer mutação durante a transcrição. Isto é, existe uma frequência maior do que o esperado de regiões capazes de formar pareamento intracadeia, sendo esta tendência maior nos genes ativamente transcritos.

Obviamente as mutações não ocorrem somente durante a transcrição, se assim fosse não existiriam mutações em regiões não transcritas. É plausível imaginar que a estrutura do DNA dupla fita pode influenciar na tendência para o desenvolvimento de mutações. Desta maneira, as alterações na estrutura do DNA dupla fita causadas por mutações (excluindo-se aquelas que possuem tendência para estar na forma de fita simples durante a transcrição), tenderiam a seguir uma mesma tendência no padrão na variação na estrutura do DNA (i.e. mutações em DNA dupla fita tendem a aumentar ou tendem a diminuir a probabilidade de clivagem química dos radicais hidroxila). Esta hipótese pôde ser testada analisando-se se existe um padrão na variação dos valores obtidos no programa OrChid das bases normais e as respectivas bases mutadas.

3. Proposição

O presente trabalho teve como objetivo investigar a importância da conformação estrutural do DNA dupla fita de seqüências “cis” localizadas em regiões não codificadoras na expressão gênica e na propensão para desenvolvimento de mutações. Foi estudado ainda se mutações em diferentes bases afetariam de forma diferente a estrutura do DNA dupla fita. Também foram analisados os valores de predição de clivagem química das seqüências que não estavam pareadas.

4. Material e Métodos

Foram utilizadas mutações do tipo substituição em regiões não codificadoras (até -2000 pares de bases do início da tradução) de genes que causam alterações fenotípicas evidentes resultando em estados patológicos em genes humanos e as seqüências flanqueadoras destes sítios mutantes. Foram também obtidas seqüências do tipo substituição de mutações nas regiões codificadoras destes genes. As seqüências foram obtidas usando-se bases de dados (Pubmed e HGMD <http://www.hgmd.cf.ac.uk/ac/index.php>). Foram utilizadas as seqüências dos seguintes genes: F9 Haemophilia XIST, TH, Porphyria, MSH2, RET, ABCC8, PIGM, APOC2, RB1, ALAS2, Acetylcolina, Gamaglobina A, Gamaglobina G, Epidermolysis bolhosa, AT7B, FGA, Myocilin, Proteína C, APOA1, CETP, CRTF, ELN, IRF6, PRKAG2, PRKAR1A, KCNJ11 e CDKN2A, ADSL, EDN3, ELA2, F7, FMR1, LDLR, CYP17A1, SLC26A4, GJB1, HNF1, SPINK1, ATM provenientes do site HGMD (<http://www.hgmd.cf.ac.uk/ac/all.php>). Foram obtidas as seqüências codificadoras e promotoras. Nas regiões promotoras as seqüências tanto normais quanto mutantes já são oferecidas prontas pelo site, já nas codificadoras estas vêm com o local da mutação especificado de forma numérica assim como o códon que traz a base normal e seu respectivo códon com a base mutada, foi preciso então montar as seqüências normal e mutante com 61 pares de bases feita pela seqüência de cDNA também obtida neste site.

Exemplo: Gene ELASTASE 2 – ELA2

CR041558	TCCCAGCACAGGGCTATAAGAGGAGCCGGG(C- A)GGGCACGGAGGGGCAGAGACCCCGGAGCCC -41 relative to initiation codon	Neutropaenia, congenital	Bellanne- Chantelot (2004) Blood 103, 4119
----------	---	-----------------------------	--

Região Promotora

CM041328	CCC-CTC	Pro-Leu	13	Neutropaenia, congenital	Bellanne-Chantelot (2004) Blood 103, 4119
----------	---------	---------	----	--------------------------	---

Região Codificadora

```

-29  ATG ACC CTC GGC CGC CGA CTC GCG TGT CTT TTC CTC GCC TGT GTC -15
-14  CTG CCG GCC TTG CTG CTG GGG GGC ACC GCG CTG GCC TCG GAG ATT 1
2    GTG GGG GGC CGG CGA GCG CGG CCC CAC GCG TGG CCC TTC ATG GTG 16
17   TCC CTG CAG CTG CGC GGA GGC CAC TTC TGC GGC GCC ACC CTG ATT 31
32   GCG CCC AAC TTC GTC ATG TCG GCC GCG CAC TGC GTG GCG AAT GTA 46
47   AAC GTC CGC GCG GTG CGG GTG GTC CTG GGA GCC CAT AAC CTC TCG 61
62   CGG CGG GAG CCC ACC CGG CAG GTG TTC GCC GTG CAG CGC ATC TTC 76
77   GAA AAC GGC TAC GAC CCC GTA AAC TTG CTC AAC GAC ATC GTG ATT 91
92   CTC CAG CTC AAC GGG TCG GCC ACC ATC AAC GCC AAC GTG CAG GTG 106
107  GCC CAG CTG CCG GCT CAG GGA CGC CGC CTG GGC AAC GGG GTG CAG 121
122  TGC CTG GCC ATG GGC TGG GGC CTT CTG GGC AGG AAC CGT GGG ATC 136
137  GCC AGC GTC CTG CAG GAG CTC AAC GTG ACG GTG GTG ACG TCC CTC 151
152  TGC CGT CGC AGC AAC GTC TGC ACT CTC GTG AGG GGC CGG CAG GCC 166
167  GGC GTC TGT TTC GGG GAC TCC GGC AGC CCC TTG GTC TGC AAC GGG 181
182  CTA ATC CAC GGA ATT GCC TCC TTC GTC CGG GGA GGC TGC GCC TCA 196
197  GGG CTC TAC CCC GAT GCC TTT GCC CCG GTG GCA CAG TTT GTA AAC 211
212  TGG ATC GAC TCT ATC ATC CAA CGC TCC GAG GAC AAC CCC TGT CCC 226
227  CAC CCC CGG GAC CCG GAC CCG GCC AGC AGG ACC CAC TGA

```

Seqüência Gênica Completa

A importância da estrutura do DNA dupla fita foi avaliada por três análises distintas:

1. Análise da conservação filogenética dos padrões de clivagem.

Nesta análise foi realizada a comparação de seqüências em regiões não codificadoras (5' UTR e promotora) e regiões codificadoras. A comparação da conservação da estrutura foi feita utilizando-se o seguinte protocolo:

a) A seqüência humana contendo a mutação e 60 bases flanqueadoras (30 de cada lado) foi comparada com seqüências de diversas espécies mamíferos utilizando-se o UCSC genome browser na opção Genome Blat/Browser (<http://genome.ucsc.edu/>). Esta ferramenta nos permite obter seqüências de várias espécies animais.

b) Foram selecionadas a partir das seqüências obtidas para a análise no item 2 (pelo menos 500 seqüências codificadoras e aproximadamente 200 seqüências não codificadoras) 30 seqüências em regiões mutadas não codificadoras e 30 em regiões codificadoras. Nos mamíferos placentários não humanos só foram selecionadas as seqüências onde ocorreu substituição de um único nucleotídeo (como ocorre na mutação humana) na região de 6 bases que flanqueiam a região correspondente a mutação

humana (3 pares de bases em cada lado, totalizando 7 bases). Quando mais de uma espécie apresentou mutação conforme os critérios estabelecidos acima as mutações foram comparadas separadamente com a seqüência humana. A região de 7 bases foi determinada devido ao fato de ser um tamanho razoável para a ligação de um fator de transcrição ao DNA determinado por ensaio de “footprinting” (Maston et al. 2006). Os valores de predição de clivagem química da região que engloba 3 pares de bases que flanqueiam o sítio mutado humano juntamente com a região que engloba 3 pares de bases do sítio mutado nas outras espécies de mamíferos foram obtidos (ver Figura 1). Foram excluídas as seqüências das espécies que tiverem mais de uma substituição em relação à espécie humana nesta seqüência. Os valores de predição de clivagem química do alelo humano normal foram comparados com o humano mutado utilizando-se o programa ORCHID ([HTTP://dna.bu.edu/orchid](http://dna.bu.edu/orchid)) (Tabela 1), e posteriormente com a seqüência da espécie não humana pelo método estatístico de Kolmogorov-Smirnov (Figura 2). Este teste estatístico é usado para determinar se duas distribuições (no caso, valores de predição de clivagem de radicais hidroxila de duas sequencias de DNA) diferem uma da outra ou pertencem a uma mesma população. Para cada análise foi obtido o valor de desvio máximo bilateral (quanto maior a diferença dos valores de predição de clivagem das duas distribuições maior o desvio) e o valor de p unilateral (quanto maior a diferença dos valores de predição de clivagem das duas distribuições menor o valor de p). Os valores de cada par de comparação (para desvio máximo e p) foram comparados pelo teste dos sinais (teste não paramétrico onde se define em cada par a diferença dos escores baseando-se somente no sentido das diferenças, registrados sob a forma de sinais de + ou -). Desta maneira testamos se as mutações que causam alterações fenotípicas graves alteram a estrutura do DNA de maneira diferente dos polimorfismos genéticos que não causam alterações fenotípicas graves.

Humano normal ATGCGTGCAGTGACGAT...

Humano mutado ...ATGCGT**C**CAGTGACGAT...

Seqüência rato ATGCGTGCA**A**TGACGAT...

Figura 1. Os valores de predição de clivagem química das regiões sublinhadas serão obtidos e comparados pelo método estatístico de Kolmogorov-Smirnov. Notar que as seqüências de 9 pares de bases englobam 3 nucleotídeos que flanqueiam a mutação humana (negrito e sublinhado) e a mutação na espécie não humana (negrito e sublinhado).

Tabela 1. A linha marcada em vermelho corresponde a mutação e em azul as 3 bases vizinhas. Gene ELASTASE 2 – ELA2 (CR041558).

Sequência Normal		Sequência Mutante	
Nucleotide	Prediction	Nucleotide	Prediction
T	-0.07	T	-0.08
C	-1.07	C	-1.10
C	-1.13	C	-1.15
C	0.22	C	0.22
A	-0.69	A	-0.70
G	1.62	G	1.64
C	-0.22	C	-0.23
A	-1.91	A	-1.95
C	0.66	C	0.66
A	-2.03	A	-2.07
G	0.88	G	0.89
G	0.55	G	0.55
G	0.41	G	0.41
C	1.50	C	1.52
T	-0.12	T	-0.13
A	-0.20	A	-0.21
T	2.30	T	2.33
A	-1.02	A	-1.04
A	-0.77	A	-0.79
G	0.48	G	0.48
A	-1.09	A	-1.11
G	0.49	G	0.50
G	0.55	G	0.55
A	-1.04	A	-1.07
G	1.45	G	1.47
C	0.11	C	0.11
C	-1.42	C	-1.45
G	-0.38	G	-0.18
G	0.71	G	0.86
G	0.65	G	0.79
C	0.26	A	-0.60
G	-1.03	G	0.57
G	1.22	G	0.93

G	1.27	G	0.72
C	0.63	C	0.63
A	-2.21	A	-2.26
C	1.00	C	1.01
G	-0.41	G	-0.43
G	0.98	G	0.99
A	-0.24	A	-0.25
G	0.24	G	0.24
G	0.12	G	0.12
G	0.23	G	0.23
G	0.84	G	0.85
C	0.87	C	0.88
A	-1.68	A	-1.71
G	1.34	G	1.35
A	-0.77	A	-0.79
G	1.75	G	1.78
A	-0.20	A	-0.21
C	0.64	C	0.64
C	-0.92	C	-0.94
C	-0.87	C	-0.89
C	-0.65	C	-0.67
G	-0.21	G	-0.22
G	0.72	G	0.72
A	-0.73	A	-0.75
G	0.68	G	0.69
C	-0.44	C	-0.45
C	-0.76	C	-0.78
C	-1.05	C	-1.07

Humano normal tcccagcacagggtataagaggagccggg-c-gggcacggaggggcagagaccccgagccc

Humano mutado tcccagcacagggtataagaggagccggg-a-gggcacggaggggcagagaccccgagccc

Seqüência rhesus tcttgccacagggtataagagaagccggg-t-gggcacggaggggcagagaccccgagacc

Figura 2. Os valores de predição de clivagem química das regiões sublinhadas foram obtidos e comparados pelo método estatístico de Kolmogorov-Smirnov.

Da mesma maneira como foi exposta acima foi comparado o efeito das mutações nas regiões não codificadoras e codificadoras. Isto se deve ao fato de que nas regiões reguladoras da transcrição o DNA dupla fita possui estruturas conservadas. A conservação estrutural possivelmente reflete parâmetros funcionais (Greambaum et al 2007). A menor frequência de motivos estruturais em regiões codificadoras pode ser explicada pelo fato de que nestas regiões a seqüência de nucleotídeos (formando códons) é mais relevante do que a estrutura. Desta maneira, a hipótese testada é de que as mutações nas regiões não codificadoras e nas regiões codificadoras alteram a estrutura do DNA de maneira diferenciada.

2. Influência do tipo de mutação na alteração da estrutura do DNA.

Juntamente com as análises filogenéticas descritas acima observamos a influência do tipo de mutação na alteração da estrutura do DNA.

As análises foram feitas separadamente em dois grupos sendo:

1. mutação em região não codificadora (inclui região 5' UTR e promotora)

2. mutação em região codificadora.

As análises mostraram que as mutações de G-A (de G para A) ou A-G produzem a maior alteração na estrutura do DNA (i.e. predição de clivagem química de radicais hidroxila) e a mutação C-T/T-C o menor valor. No caso da A-G/G-A o valor de p foi menor do que 0,05 quando no teste estatístico Kruskal Wallis/ método de Dunn. No caso da alteração C-T/T-C a maioria dos valores de p foram menores que 0,05, a única exceção foi a comparação com a mutação C-C/C-G, porém neste caso o z calculado foi próximo do z crítico sugerindo que com uma amostragem maior o valor de p passe a ser menor que 0,05. Na Tabela 2 abaixo são mostrados os resultados para a região promotora (n=116).

Mutação	Posto médio
1 (C-T/T-C)	134.2320
5 (A-G/G-A)	350.4970
9 (C-G/G-C)	189.9583
13 (A-C/C-A)	271.0513
17 (A-T/T-A)	258.7093
21 (G-T/T-G)	209.8091

Tabela 2. A maior parte das comparações os grupos 1 (C-T/T-C) e 5 (G-A/A-G) diferiram dos outros grupos ($p < 0.05$).

Comparações (método Dunn)	Diferentes Postos	z calculado	z crítico	p
Postos médios 1 e 5	216.2650	12.9305	2.935	< 0.05
Postos médios 1 e 9	55.7263	2.5059	2.935	ns
Postos médios 1 e 13	136.8193	5.2682	2.935	< 0.05
Postos médios 1 e 17	124.4773	4.9725	2.935	< 0.05
Postos médios 1 e 21	75.5771	3.2987	2.935	< 0.05
Postos médios 5 e 9	160.5387	7.5387	2.935	< 0.05
Postos médios 5 e 13	79.4457	3.1566	2.935	< 0.05
Postos médios 5 e 17	91.7877	3.7930	2.935	< 0.05
Postos médios 5 e 21	140.6879	6.3958	2.935	< 0.05
Postos médios 9 e 13	81.0929	2.7844	2.935	ns
Postos médios 9 e 17	68.7510	2.4301	2.935	ns
Postos médios 9 e 21	19.8508	0.7510	2.935	ns
Postos médios 13 e 17	12.3420	0.3942	2.935	ns
Postos médios 13 e 21	61.2422	2.0661	2.935	ns
Postos médios 17 e 21	48.9002	1.6965	2.935	ns

Foi analisado também o efeito da mutação na região de 7 pares de bases, que inclui o sítio mutado e bases flanqueadoras. Serão excluídas mutações em regiões 5' UTR que criarem códon de início (ATG). A análise da variação na região de 7 bases foi feita pela somatória dos módulos dos valores de predição de clivagem cada base normal

subtraída do valor da respectiva base mutada, ou seja $\sum |\text{OrChid normal} + \text{OrChid mutado}|$. A análise estatística foi feita pelo teste de Kruskal-Wallis como mostrado acima.

3. Influência da estrutura do DNA dupla fita na propensão para mutações.

As mutações não ocorrem somente durante a transcrição, se assim fosse não existiriam mutações em regiões não transcritas. É plausível imaginar que a estrutura do DNA dupla fita pode influenciar na tendência para o desenvolvimento de mutações. Desta maneira, as alterações na estrutura do DNA dupla fita causadas por mutações (excluindo-se aquelas que possuem tendência para estar na forma de fita simples durante a transcrição), tenderiam a seguir uma mesma tendência no padrão na variação na estrutura do DNA (i.e. mutações em DNA dupla fita tendem a aumentar ou tendem a diminuir a probabilidade de clivagem química dos radicais hidroxila). Esta hipótese foi testada analisando-se se existe um padrão na variação dos valores obtidos no programa OrChid das bases normais e as respectivas bases mutadas. A análise da influência da estrutura do DNA na predisposição para mutações foi feita seguindo o seguinte protocolo:

a) As seqüências normais das regiões não codificadoras obtidas na análise do item 2 (61 bases) foram submetidas à análise de conformação de DNA simples fita no DNA mfold Server (<http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi>) (Zuker, 2003). Só foram incluídas nas análises posteriores seqüências onde a mutação estivesse pareada na estrutura de menor ΔG (maior valor de ΔG negativo). Isto foi feito para excluir as bases que possuíssem maior probabilidade de estarem como fita simples induzido pelo 'supercoiling' durante a transcrição (Wright et al 2003). As regiões 5' UTR e codificadoras foram excluídas da análise por estarem na forma de fita simples durante a transcrição. Foram obtidos os valores de predição de clivagem química no OrChid das seqüências normais e mutantes.

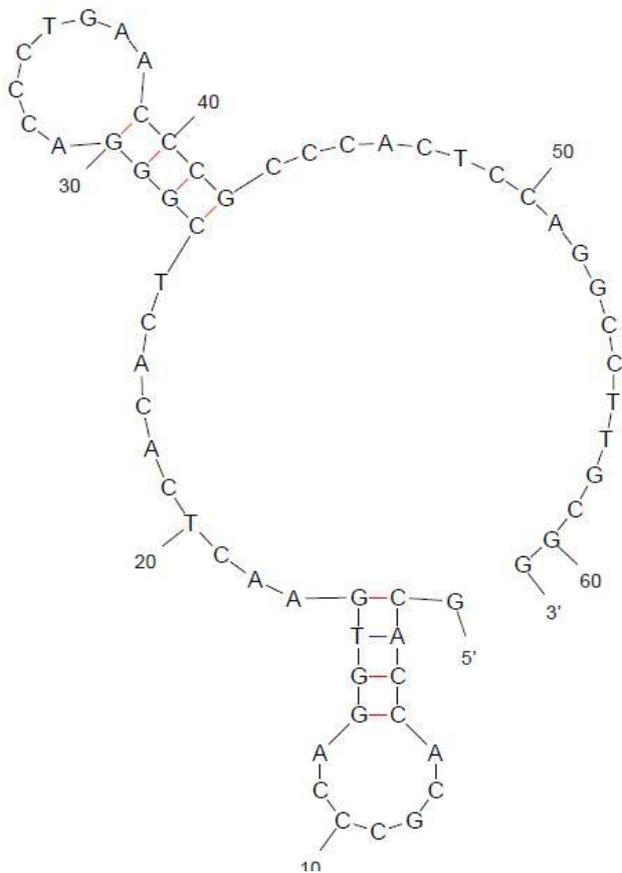


Figura 3. Exemplo da análise de conformação de DNA no DNA mfold Server: Polimorfismo - Gene ABCC8 - Sequência RS 36234280 – Notar que a base na posição 31, que corresponde a base do polimorfismo, aparece na forma não pareada.

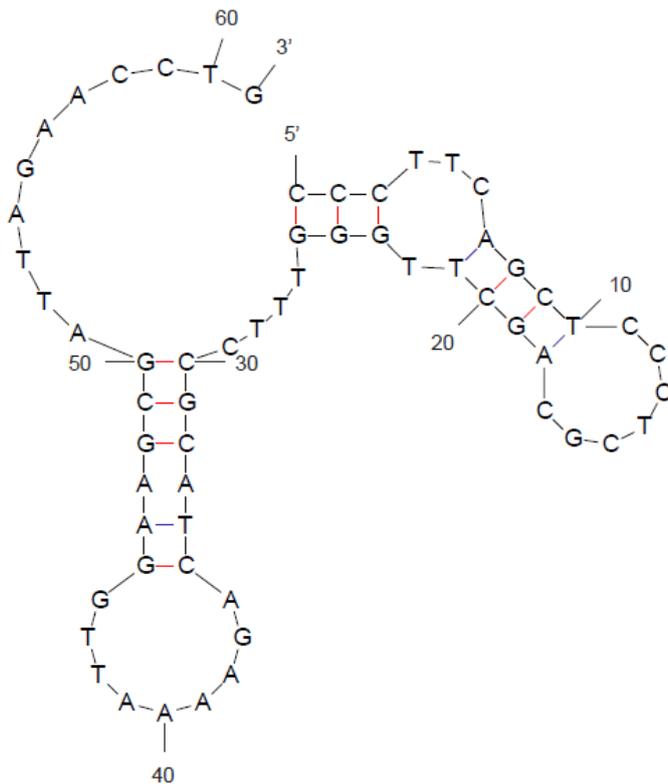


Figura 4. Exemplo da análise de conformação de DNA no DNA mfold Server: Polimorfismo - Gene ABCC8 - Sequência RS 12270649 – Notar que a base na posição 31, que corresponde a base do polimorfismo, aparece na forma pareada.

b) O aumento ou diminuição nos valores de predição de clivagem dos radicais hidroxila foram obtidos pela somatória dos valores de $(-N+M)$ (onde N corresponde ao valor OrChid da base normal e M da base mutante), para cada par de alelo normal e respectivo mutante. Por exemplo, se o valor OrChid da base normal for 3 e da mutada -1, teremos uma variação de -4, ou seja diminuição de 4 unidades quando ocorre a mutação. A análise estatística foi feita pelo teste t de Student para uma amostra admitindo-se que o valor médio é zero, ou seja, na população estudada não deve existir aumento ou diminuição dos valores na somatória de $(-N+M)$. Além desta análise os valores de N e M de cada par foram tabulados e os dados para todas as seqüências analisadas foram comparados pelo teste dos sinais (teste não paramétrico onde se define em cada par a diferença dos escores baseando-se somente no sentido das diferenças, registrados sob a forma de sinais de + ou -) e teste de t de Student para duas amostras relacionadas (ou Wicoxon, caso os dados sejam não paramétricos).

5. Resultados e Discussão

1. Análise da conservação filogenética dos padrões de clivagem.

Quando os valores de desvio máximo das regiões promotoras das mutações em humanos que causam alterações fenotípicas (obtidos no teste de Kolmogorov-Smirnov) foram comparados com os valores obtidos de variações entre as seqüências humanas e de outros mamíferos placentários (Teste t amostras relacionadas) não foram observadas diferenças significativas ($p= 0,09$). Também não foram encontradas diferenças estatisticamente significantes quando foram comparados os valores de p ($p= 0,42$, teste t para amostras relacionadas).

Quando os valores de desvio máximo das regiões codificadoras das mutações em humanos que causam alterações fenotípicas (obtidos no teste de Kolmogorov-Smirnov) foram comparados com os valores obtidos de variações entre as seqüências humanas e de outros mamíferos placentários (teste t para amostras relacionadas) não foram observadas diferenças significativas ($p= 0,075$). Também não foram encontradas diferenças estatisticamente significantes quando foram comparados os valores de p ($p=0,05$, teste t para amostras relacionadas) (Figura 5).

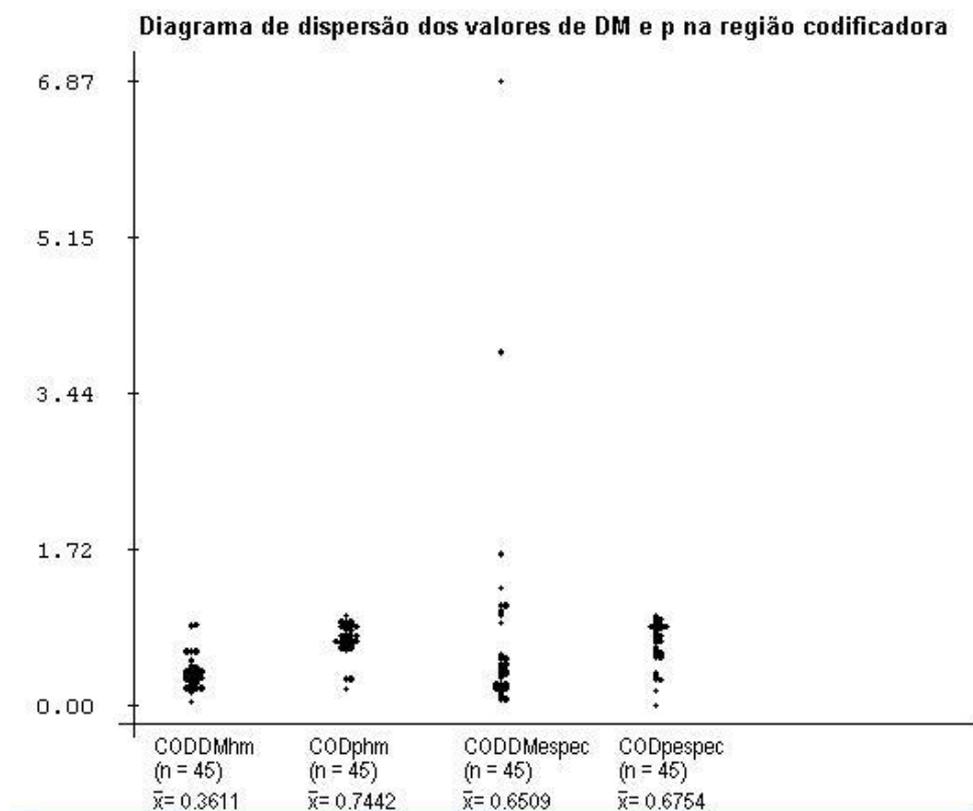


Figura 5. Diagrama de dispersão com os valores de desvio máximo DM e p nas regiões codificadoras. CODDMhm= desvio máximo da mutação humana.

CODDMespec= desvio máximo da comparação entre humano e animais. CODp_{hm}= valores de p da mutação humana. COD_{pespec}= valores de p da comparação entre humano e animais.

Quando os valores de desvio máximo das regiões promotoras e codificadoras das mutações em humanos que causam alterações fenotípicas (obtidos no teste de Kolmogorov-Smirnov) foram comparados (Teste de Mann-Whitney) não foram observadas diferenças significativas ($p= 0,61$). Também não foram encontradas diferenças estatisticamente significantes quando foram comparados os valores de p ($p=0,62$, Mann-Whitney) (Figura 6).

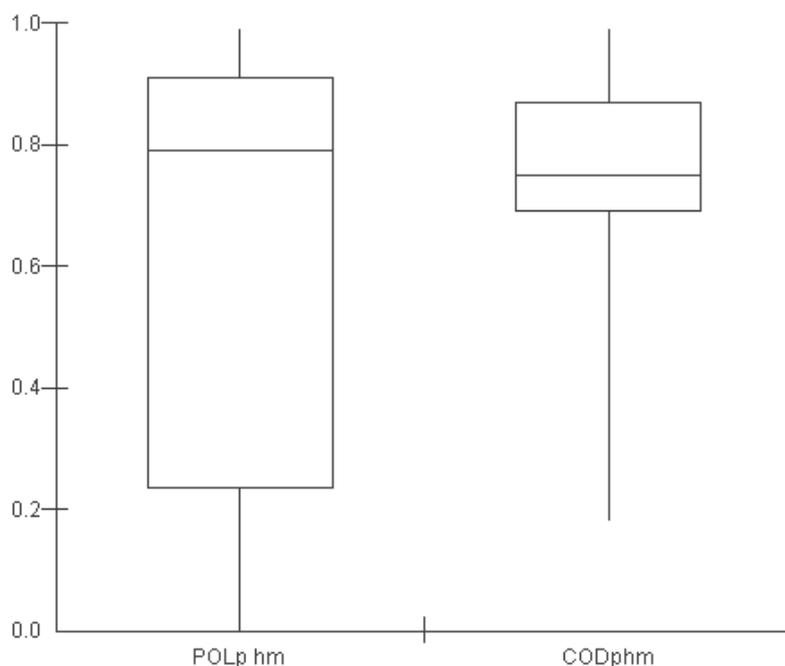


Figura 6. Valores de mediana a quartis dos valores das regiões promotoras humanas (POL_phm) e codificadoras (COD_phm). Observar que apesar de não existir diferença estatisticamente significativa ($p > 0,05$) a variação dos valores na região promotora foi maior do que na região codificadora.

Estes resultados mostraram que as mutações que causam alterações fenotípicas em humanos não alteram a estrutura do DNA de forma significativamente diferente das variações que ocorrem entre humanos e outros animais. Também não foram encontradas diferenças significativas quando se comparou a estrutura do DNA normal e mutado nas regiões codificadoras e promotoras.

2. Influência do tipo de mutação na alteração da estrutura do DNA.

Foram obtidos valores de padrão de clivagem de 868 seqüências. Quando se comparou a variação no padrão de clivagem na base mutada (normal-mutante) entre os tipos de mutação obteve-se uma diferença significativa entre a mutação AG/GA com as demais seqüências. Isto mostra que as mutações GA/AG produzem as maiores variações na estrutura do DNA nas regiões codificadoras.

Resultados

H = 64.0552

Graus de liberdade = 5

(p) Kruskal-Wallis = < 0.0001

AC/CA (posto médio 1) = 474.2652

AG/GA (posto médio 2) = 516.6890

CG/GC (posto médio 3) = 354.1550

CT/TC (posto médio 4) = 368.5368

GT/TG (posto médio 5) = 386.1435

AT/TA (posto médio 6) = 461.6250

Comparações (método de Dunn)	Dif. Postos	z calculado	z crítico	p
Postos médios 1 e 2	42.4239	1.2411	2.935	ns
Postos médios 1 e 3	120.1102	3.0208	2.935	< 0.05
Postos médios 1 e 4	105.7284	3.0214	2.935	< 0.05
Postos médios 1 e 5	88.1216	2.2496	2.935	ns
Postos médios 1 e 6	12.6402	0.2959	2.935	ns
Postos médios 2 e 3	162.5340	5.5927	2.935	< 0.05
Postos médios 2 e 4	148.1522	6.7057	2.935	< 0.05
Postos médios 2 e 5	130.5455	4.6212	2.935	< 0.05
Postos médios 2 e 6	55.0640	1.6686	2.935	ns
Postos médios 3 e 4	14.3818	0.4792	2.935	ns
Postos médios 3 e 5	31.9885	0.9194	2.935	ns
Postos médios 3 e 6	107.4700	2.7734	2.935	ns
Postos médios 4 e 5	17.6067	0.6024	2.935	ns
Postos médios 4 e 6	93.0882	2.7508	2.935	ns
Postos médios 5 e 6	75.4815	1.9788	2.935	ns

Quando se comparou a variação no padrão de clivagem na região de 7 bases que inclui base mutada e 3 bases flanqueadoras de cada lado as diferenças entre os postos médios foram menores.

Resultados

H = 13.2581

Graus de liberdade = 5

(p) Kruskal-Wallis = 0.0211

AC/CA (posto médio 1) = 450.5530

AG/GA (posto médio 2) = 472.7010

CG/GC (posto médio 3) = 400.9050

CT/TC (posto médio 4) = 400.0087

GT/TG (posto médio 5) = 429.5324

AT/TA (posto médio 6) = 430.1597

Comparações (método de Dunn)	Dif. Postos	z calculado	z crítico	p
Postos médios 1 e 2	22.1480	0.6479	2.935	ns
Postos médios 1 e 3	49.6480	1.2487	2.935	ns
Postos médios 1 e 4	50.5444	1.4444	2.935	ns
Postos médios 1 e 5	21.0206	0.5366	2.935	ns
Postos médios 1 e 6	20.3933	0.4773	2.935	ns
Postos médios 2 e 3	71.7960	2.4705	2.935	ns
Postos médios 2 e 4	72.6924	3.2902	2.935	< 0.05
Postos médios 2 e 5	43.1686	1.5281	2.935	ns
Postos médios 2 e 6	42.5413	1.2891	2.935	ns
Postos médios 3 e 4	0.8963	0.0299	2.935	ns
Postos médios 3 e 5	28.6274	0.8228	2.935	ns
Postos médios 3 e 6	29.2547	0.7550	2.935	ns
Postos médios 4 e 5	29.5237	1.0102	2.935	ns
Postos médios 4 e 6	30.1511	0.8910	2.935	ns
Postos médios 5 e 6	0.6273	0.0164	2.935	ns

Estes resultados mostram que a alteração na estrutura do DNA ocorre principalmente na base mutada não se expandindo significativamente para as bases vizinhas.

Além do experimento acima, foi feita também a análise de mutações separadas, ou seja, verificamos se mutação AG produz alteração na estrutura do DNA diferente da

mutação GA e com os outros tipos de mutação possíveis, isto foi feito somente com o valor no local das bases normais e mutadas. Nesta análise os valores das regiões codificadoras e não codificadoras foram agrupados, haja vista que os valores destas não diferiram significativamente ($p > 0,05$) quando comparados pelo teste de Mann-Whitney. Desta maneira os valores de predição de clivagem das mutações AG nas regiões codificadoras não diferiram significativamente dos valores nas regiões não codificadoras, o mesmo também ocorreu para as outras seqüências.

H = 672.1832

Graus de liberdade =	11
(p) Kruskal-Wallis =	< 0.0001
AC (posto médio 1) =	193.3167
AG (posto médio 2) =	89.9560
AT (posto médio 3) =	166.8200
CA (posto médio 4) =	728.4245
CG (posto médio 5) =	271.2500
CT (posto médio 6) =	434.4448
GA (posto médio 7) =	795.2277
GC (posto médio 8) =	537.5446
GT (posto médio 9) =	636.7264
TA (posto médio 10) =	688.7255
TC (posto médio 11) =	414.3333
TG (posto médio 12) =	272.8939

Comparações (método de Dunn)	Dif. Postos	z calculado	z crítico	p
Postos médios 7 e 8	103.3606	1.7275	3.126	ns
Postos médios 7 e 9	26.4967	0.3443	3.126	ns
Postos médios 7 e 10	535.1079	8.2409	3.126	< 0.05
Postos médios 7 e 11	77.9333	1.2393	3.126	ns
Postos médios 7 e 12	241.1281	4.2514	3.126	< 0.05
Postos médios 7 e 13	601.9110	10.8937	3.126	< 0.05
Postos médios 7 e 14	344.2280	5.3534	3.126	< 0.05
Postos médios 7 e 15	443.4097	6.8287	3.126	< 0.05
Postos médios 7 e 16	495.4088	7.5760	3.126	< 0.05
Postos médios 7 e 17	221.0167	3.8001	3.126	< 0.05
Postos médios 7 e 18	79.5773	1.2716	3.126	ns

Postos médios 8 e 9	76.8640	1.1977	3.126	ns
Postos médios 8 e 10	638.4685	13.0015	3.126	< 0.05
Postos médios 8 e 11	181.2940	3.9102	3.126	< 0.05
Postos médios 8 e 12	344.4888	9.1674	3.126	< 0.05
Postos médios 8 e 13	705.2716	19.9628	3.126	< 0.05
Postos médios 8 e 14	447.5886	9.2728	3.126	< 0.05
Postos médios 8 e 15	546.7704	11.1342	3.126	< 0.05
Postos médios 8 e 16	598.7694	12.0447	3.126	< 0.05
Postos médios 8 e 17	324.3773	8.1660	3.126	< 0.05
Postos médios 8 e 18	182.9379	3.9813	3.126	< 0.05
Postos médios 9 e 10	561.6045	8.1445	3.126	< 0.05
Postos médios 9 e 11	104.4300	1.5580	3.126	ns
Postos médios 9 e 12	267.6248	4.3672	3.126	< 0.05
Postos médios 9 e 13	628.4077	10.4860	3.126	< 0.05
Postos médios 9 e 14	370.7246	5.4231	3.126	< 0.05
Postos médios 9 e 15	469.9064	6.8147	3.126	< 0.05
Postos médios 9 e 16	521.9055	7.5217	3.126	< 0.05
Postos médios 9 e 17	247.5133	3.9527	3.126	< 0.05
Postos médios 9 e 18	106.0739	1.5893	3.126	ns
Postos médios 10 e 11	457.1745	8.6615	3.126	< 0.05
Postos médios 10 e 12	293.9797	6.4954	3.126	< 0.05
Postos médios 10 e 13	66.8032	1.5388	3.126	ns
Postos médios 10 e 14	190.8799	3.5047	3.126	< 0.05
Postos médios 10 e 15	91.6981	1.6610	3.126	ns
Postos médios 10 e 16	39.6990	0.7121	3.126	ns
Postos médios 10 e 17	314.0912	6.6748	3.126	< 0.05
Postos médios 10 e 18	455.5306	8.6902	3.126	< 0.05
Postos médios 11 e 12	163.1948	3.8610	3.126	< 0.05
Postos médios 11 e 13	523.9777	13.0078	3.126	< 0.05
Postos médios 11 e 14	266.2946	5.1207	3.126	< 0.05
Postos médios 11 e 15	365.4764	6.9242	3.126	< 0.05
Postos médios 11 e 16	417.4755	7.8259	3.126	< 0.05
Postos médios 11 e 17	143.0833	3.2382	3.126	< 0.05
Postos médios 11 e 18	1.6439	0.0330	3.126	ns
Postos médios 12 e 13	360.7829	12.1272	3.126	< 0.05
Postos médios 12 e 14	103.0998	2.3248	3.126	ns
Postos médios 12 e 15	202.2816	4.4694	3.126	< 0.05
Postos médios 12 e 16	254.2807	5.5380	3.126	< 0.05

Postos médios 12 e 17	20.1115	0.5770	3.126	ns
Postos médios 12 e 18	161.5509	3.8637	3.126	< 0.05
Postos médios 13 e 14	257.6830	6.0688	3.126	< 0.05
Postos médios 13 e 15	158.5013	3.6512	3.126	< 0.05
Postos médios 13 e 16	106.5022	2.4153	3.126	ns
Postos médios 13 e 17	380.8943	11.7495	3.126	< 0.05
Postos médios 13 e 18	522.3337	13.1226	3.126	< 0.05
Postos médios 14 e 15	99.1818	1.8211	3.126	ns
Postos médios 14 e 16	151.1808	2.7483	3.126	ns
Postos médios 14 e 17	123.2113	2.6680	3.126	ns
Postos médios 14 e 18	264.6507	5.1255	3.126	< 0.05
Postos médios 15 e 16	51.9991	0.9328	3.126	ns
Postos médios 15 e 17	222.3931	4.7261	3.126	< 0.05
Postos médios 15 e 18	363.8325	6.9409	3.126	< 0.05
Postos médios 16 e 17	274.3922	5.754	0.126	< 0.05
Postos médios 16 e 18	415.8316	7.8480	3.126	< 0.05
Postos médios 17 e 18	141.4394	3.2328	3.126	< 0.05

Os resultados mostraram claramente que existe diferença na variação do padrão de clivagem química nos diversos tipos de mutação. Sendo que as maiores variações foram encontradas nas mutações GA e CA e os menores AG e AC. É possível que as diferenças nas variações de predição de clivagem podem ter implicações na expressão genética.

Estes resultados nos motivaram a procurar uma possível relação entre a base normal e a variação no valor de clivagem química, assim como relação entre a base mutada e a variação no valor de clivagem química. Desta maneira foram comparados os valores de predição de clivagem das bases normais, mostrados na tabela abaixo.

Resultados:

H = 207.1345

Graus de liberdade = 3

(p) Kruskal-Wallis = < 0.0001

A (posto médio 1) = 213.0070

C (posto médio 2) = 488.3729

G (posto médio 3) = 585.9217

T (posto médio 4) = 404.8422

Comparações (método de Dunn)	Dif. Postos	z calculado	z crítico	p
Postos médios 19 e 20	275.3658	9.7999	2.635	< 0.05
Postos médios 19 e 21	372.9147	13.9309	2.635	< 0.05
Postos médios 19 e 22	191.8352	6.7653	2.635	< 0.05
Postos médios 20 e 21	97.5488	4.2768	2.635	< 0.05
Postos médios 20 e 22	83.5307	3.3884	2.635	< 0.05
Postos médios 21 e 22	181.0795	7.8307	2.635	< 0.05

O padrão de clivagem foi também analisado agrupando-se as bases mutadas segundo o tipo de mutação. Os resultados são mostrados na tabela abaixo.

Resultados

H = 265.1929

Graus de liberdade = 3

(p) Kruskal-Wallis = < 0.0001

A (posto médio 1) = 313.6705

C (posto médio 2) = 510.9606

G (posto médio 3) = 702.6523

T (posto médio 4) = 571.0107

Comparações (método de Dunn)	Dif. Postos	z calculado	z crítico	p
Postos médios 27 e 28	197.2901	7.8689	2.635	< 0.05
Postos médios 27 e 29	388.9818	15.3667	2.635	< 0.05
Postos médios 27 e 30	257.3403	10.7106	2.635	< 0.05
Postos médios 28 e 29	191.6917	6.7373	2.635	< 0.05
Postos médios 28 e 30	60.0501	2.1985	2.635	ns
Postos médios 29 e 30	131.6416	4.7808	2.635	< 0.05

Os resultados mostrados nas duas tabelas acima evidenciam principalmente que as bases normais A tem um menor valor de predição de clivagem, portanto, uma menor probabilidade de sofrer mutação. É interessante notar que esta tendência ocorreu nas

análises com as seqüências normais e com as seqüências mutadas. No entanto, pode-se notar que houve um aumento nos valores de predição de clivagem química quando os valores foram agrupados pelas bases mutantes. A comparação com cada tipo de base (i.e. base normal x base mutante) pelo método de Kruskal-Wallis, Teste de Dumm mostrou diferença significativa para as bases C ($p= 0,01$) e T (0,04). Pode-se inferir das análises feitas que além do tipo de mutação a variação na estrutura do DNA predita pelos valores de predição de clivagem química também são influenciadas pelas bases vizinhas, confirmando resultados obtidos por outros autores (Hirons ET al 2006).

Foi realizado também o teste dos sinais para verificar se existia uma tendência para aumento ou diminuição dos valores de predição de clivagem para cada tipo de mutação. O teste de sinais é um teste que compara grupos de amostras relacionadas fazendo a subtração dos valores de predição de clivagem da base normal (selvagem) do valor da base mutada. Desta maneira se a resultante for positiva significa que o valor normal é maior que o mutado e vice versa. A tabela 4 abaixo mostra os resultados agrupados de duas maneiras. No grupo contendo as 12 linhas superiores os valores foram agrupados pela base normal e no grupo contendo as 12 linhas inferiores os mesmos dados foram agrupados segundo as bases mutantes.

Tabela 4. Nas 12 linhas superiores os valores foram agrupados pela base normal e no grupo contendo as 12 linhas inferiores os mesmos dados foram agrupados segundo as bases mutantes.

mutação	positivo	negativo	neutro	p
AC	1	27	2	<0.0001
AG	0	91	0	<0.0001
AT	0	25	0	<0.0001
CA	51	2	0	<0.0001
CG	14	49	1	<0.0001
CT	80	73	2	<0.0001
GA	221	2	1	<0.0001
GC	37	19	0	0.023
GT	45	8	0	<0.0001
TA	47	4	0	<0.0001
TC	54	60	3	0.68

TG	13	52	1	<0.0001
CA	51	2	0	<0.0001
GA	221	2	1	<0.0001
TA	47	4	0	<0.0001
AC	1	27	2	<0.0001
GC	37	19	0	0.023
TC	54	60	3	0.68
AG	0	91	0	<0.0001
CG	14	49	1	<0.0001
TG	13	52	1	<0.0001
AT	0	25	0	<0.0001
CT	80	73	2	<0.0001
GT	45	8	0	<0.0001

A tabela mostra que existe uma forte tendência de quando a base A sofre mutação ocorrer uma diminuição na probabilidade de clivagem química do DNA, independente da base mutante. O inverso também é verdadeiro, ou seja, quando a base mutante é A existe uma forte tendência para um aumento da probabilidade de clivagem química. Os diagramas pontuais com os valores médios são mostrados abaixo (Figuras 7 e 8).

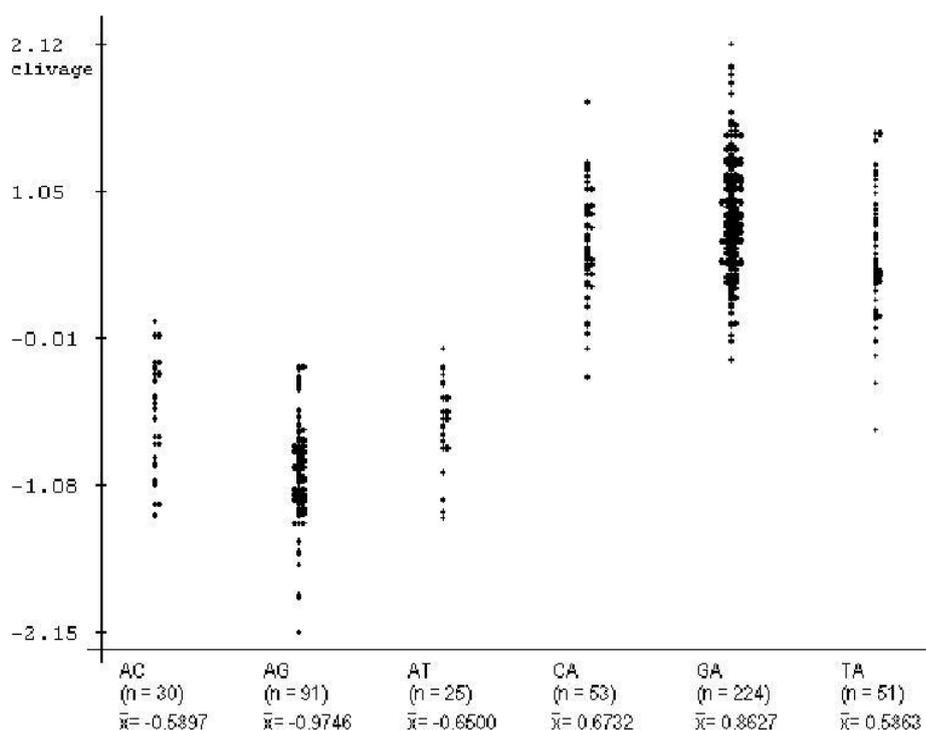


Figura 7. Distribuição e valores médios das mutações cuja base normal é A (3 primeiras) e cuja base mutada é A (3 últimas) no eixo X e os valores de predição de clivagem obtidos no programa orchid (eixo Y). Observar que os valores de predição de clivagem das mutações cuja base normal é A é negativo.

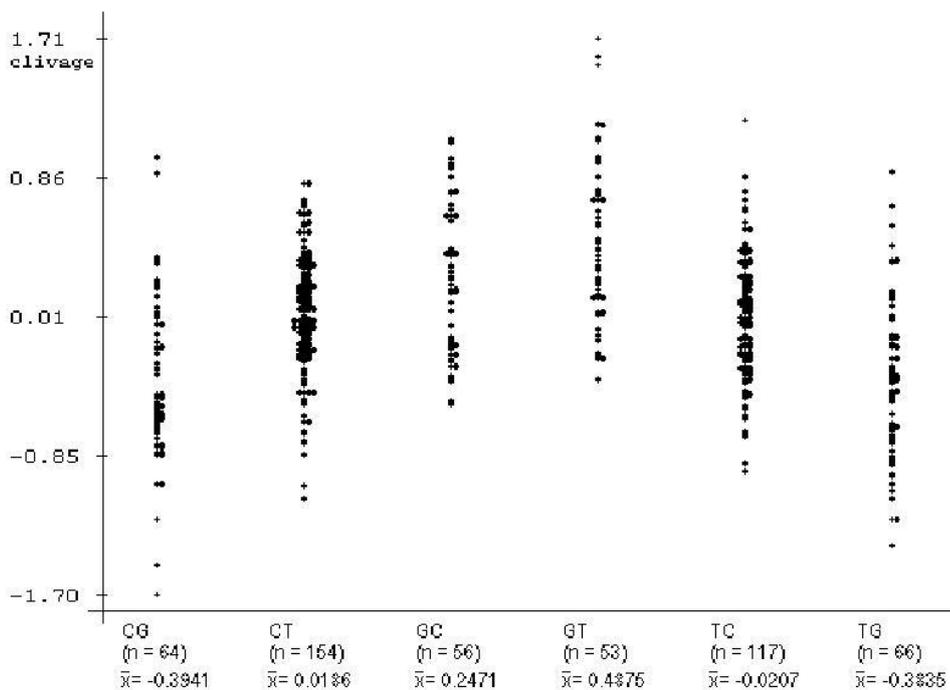


Figura 8. Distribuição e valores médios das mutações CG, CT, GC, GT, TC, TG (eixo X) e os valores de predição de clivagem obtidos no programa orchid (eixo Y).

Com a finalidade de se correlacionar os valores de predição de clivagem química com a freqüência das mutações foram catalogados os tipos de mutações das regiões codificadoras e promotoras, cujos resultados são mostrados na figura 9 abaixo:

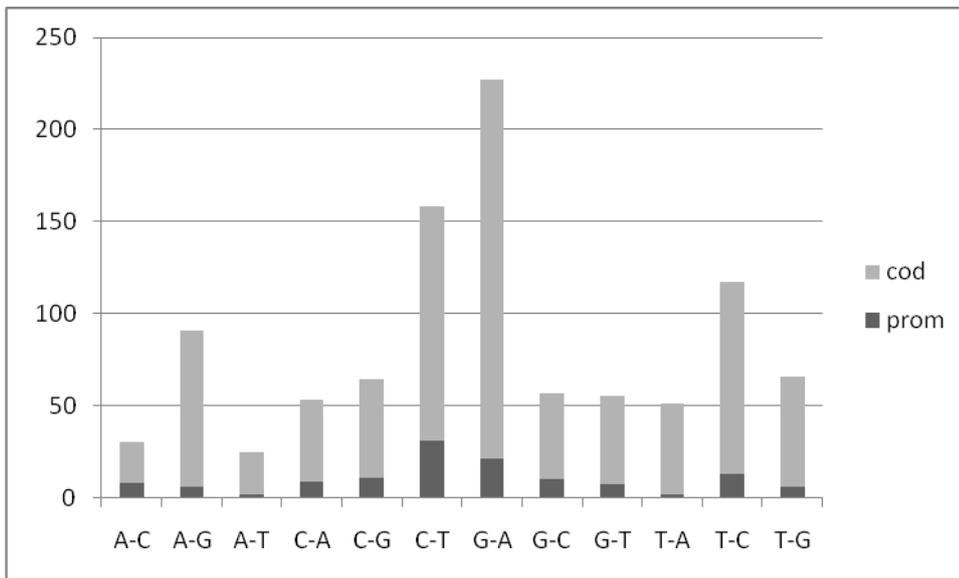


Figura 9. Mutações nas regiões promotoras (barra escura) e codificadoras (barra cinza). O tipo de mutação é mostrado no eixo X. O número de mutações é mostrado no eixo Y.

Nota-se pela figura acima que a maior parte das mutações é do tipo transição, causadas por pequenas mudanças químicas nas bases nitrogenadas. A transição mais comum encontrada nos genomas em geral é a mutação CT, que é causada pela desaminação da metilcitosina (Gardiner-Garden and Frommer 1987, Leader et al 1995).

Foi feita também a comparação da frequência das mutações entre as regiões promotora e codificadora pelo teste do qui-quadrado, sendo os resultados obtidos;

Mutação AC foi mais freqüente na região promotora ($p=0,02$)

Mutação AG não diferiu entre os grupos $p=0,08$

Mutação AT não diferiu entre os grupos $p=0,47$

Mutação CA não diferiu entre os grupos $p=0,33$

Mutação CG não diferiu entre os grupos $p=0,26$

Mutação CT foi mais freqüente na região promotora ($p=0,027$)

Mutação GA não diferiu entre os grupos $p=0,075$

Mutação GC não diferiu entre os grupos $p=0,25$

Mutação GT não diferiu entre os grupos $p=0,99$

Mutação TA não diferiu entre os grupos $p=0,053$

Mutação TC não diferiu entre os grupos $p=0,58$

Mutação TG não diferiu entre os grupos $p=0,36$

É interessante comentar que a mutação mais freqüente nas regiões codificadoras foi AG enquanto que nas regiões não codificadoras foi CT. Este fato provavelmente se deve ao fato de que a estrutura do DNA pode influenciar na propensão para mutações durante a transcrição gênica. A mutabilidade do DNA é maior quando o mesmo encontra-se na forma de fita simples durante a transcrição (Wright et al. 2002, Wright et al. 2003, Todd and Glickman 1982, Kruisselbrink et al. 2008). Desta maneira, a estrutura secundária do DNA simples fita influencia na mutabilidade principalmente em regiões codificadoras. Em regiões onde ocorre o pareamento com formação de fita dupla intracadeia o DNA é 'protegido' possuindo menor tendência para mutagenicidade. Além disto, durante a transcrição, a enzima RNA polimerase desestabiliza o DNA dupla fita introduzindo um aumento na torção da dupla fita do DNA (supercoiling) (Wright 2000). O 'supercoiling' pode se estender por até 800 pares de base acima e abaixo da RNA polimerase (Krasilnikov et al. 1999), aumentando a probabilidade de mutações em regiões não traduzidas próximas ao sítio de início da transcrição. A torção do DNA ocorre de maneira oposta nas regiões upstream e downstream ocasionado torções horário e anti-horário respectivamente. Isto explica porque na fita sense, que foi a fita analisada em nossos experimentos a mutação CT é mais nas regiões não codificadoras enquanto que esta mutação é mais comum na fita anti-sense nas regiões codificadoras (onde a mutação AG é mais comum na fita sense). Experimentos feitos com DNA plasmideano mostram que existe uma tendência para o aparecimento de estruturas conhecidas como 'stem loops' em DNA dupla fita com estrutura 'supercoiled'. As estruturas tipo 'stem loop' possuem DNA simples fita em sua extremidade, que por sua vez tem maior propensão para mutações.

Finalmente, verificamos a existência ou não de uma correlação entre os valores de predição de clivagem das bases A, C, G e T e a freqüência das mutações nestas bases. As mutações CT e GA foram as mais freqüentes na nossa análise, isto já era esperado devido ao fato que a mutação CT é influenciada pela metilação da citosina, e a mutação GA também é freqüente devido a metilação do C na fita complementar. Assim sendo, estas mutações foram removidas da análise. Como estas mutações são do tipo transição purina-purina (G-A) ou pirimidina-pirimidina (C-T) decidimos fazer as análises apenas com mutações tipo purina-pirimidina e pirimidina-purina. A análise de correlação linear de Pearson mostrou uma correlação de 0,89 e um valor de $p=0,11$. Como pode ser visto na tabela abaixo a base A, que apresentou os menores valores de predição de clivagem química também foi a mutação menos freqüente, no entanto, a base G apresentou uma

freqüência menor que esperada. Isto pode ser em parte devido ao fato de que as bases C e G são menos freqüentes no genoma, devido à alta freqüência de mutações CT. Desta maneira os valores de freqüência foram corrigidos dividindo-se pela freqüência das bases obtidas de Chargaff et al (1949), sendo a freqüência das bases A= 3.09 e T = 2.94 e das bases C = 1.98 e G = 1,99. Com esta correção houve uma melhora na correlação e significância dos resultados ($r= 0.91$, $p= 0.0876$), o que é bastante significativo para uma análise com apenas quatro pontos. Além disso, a análise de comparação de das distribuições feita pelo teste de qui-quadrado comparando-se a freqüência das mutações com os valores esperados de freqüência das bases A, C, G e T (Chargaff ET al 1949) mostrou que a freqüência da base A esta bem abaixo do esperado ($P<0,0001$), enquanto que a outras freqüências encontram-se dentro do esperado ($p> 0,05$), o que também é compatível com os valores de predição de clivagem química obtidos no programa OrChid. Estes resultados sugerem que é possível prever a freqüência ou probabilidade de uma mutação em um determinado local baseando-se na probabilidade de clivagem química dada pelo programa Orchid. No nosso conhecimento este é o primeiro experimento que valida o programa OrChid usando dados de mutações reais (Greenbaum ET al 2007).

Tabela 6. Posto médio da predição de clivagem, (PMOrchid), freqüência das mutações e correção das freqüências baseadas na regra de Chargaff

	PM Orchid	Frequencia mutação	Correção Chargaff
A	88.3	55	17.8
C	210.7	117	59.1
G	257.2	109	54.8
T	200.1	117	39.8

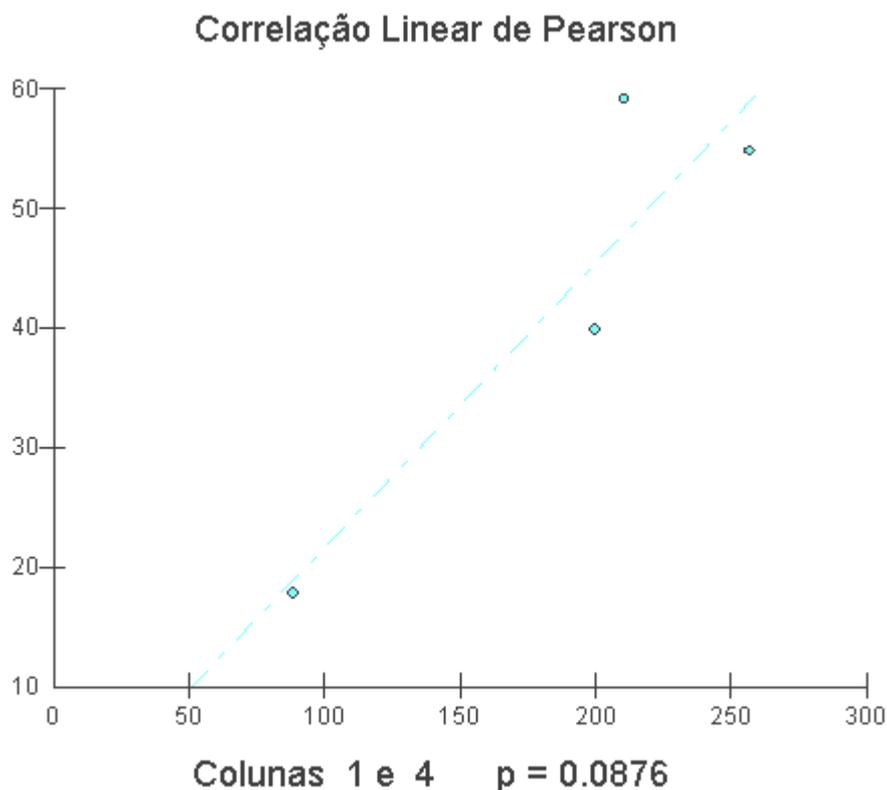


Figura 10. Correlação linear entre valores de posto médio de predição de clivagem do programa *Orchid* (eixo X) e frequência de mutações tipo transversão.

Os resultados acima são interessantes sob dois aspectos. O primeiro é que se atribui a mutagenicidade do DNA a formação de fita simples durante a transcrição ou replicação. Nossos resultados mostram que a propensão para mutagenicidade também pode ser dependente da estrutura do DNA dupla fita. O segundo aspecto refere-se à evolução do genoma. Como ressaltado anteriormente as diferenças na proporção das bases no genoma (Chargaff et al 1949) são principalmente atribuídas a maior mutagenicidade da base C. Usando como modelo a bactéria *E. coli*, Hoede et al. (2006) mostraram que existe uma pressão seletiva para a diminuição do número de regiões com tendência a sofrer mutação durante a transcrição. Isto é, existe uma frequência maior do que o esperado de regiões capazes de formar pareamento intracadeia, sendo esta tendência maior nos genes ativamente transcritos. Da mesma maneira, nossos resultados sugerem que a menor probabilidade de sofrer mutações nas bases A e T em relação às bases C e G, (principalmente a base A), pode também ter sido um fator que contribuiu para que a proporção destas bases seja maior do que a proporção das bases C e G. Ou seja, uma vez que uma mutação transformou bases C ou G em base A, esta por

ter uma menor probabilidade de sofrer mutação tende a permanecer como base A e não sofrer mutações posteriores. Isto progressivamente aumentaria a proporção desta base e da sua base complementar (T) no genoma.

3. Influência da estrutura do DNA dupla fita na propensão para mutações.

As seqüências normais das regiões não codificadoras obtidas na análise do item 2 (61 bases) foram submetidas à análise de conformação de DNA simples fita no DNA mfold Server (<http://mfold.bioinfo.rpi.edu/cgi-bin/dna-form1.cgi>) (Zuker, 2003). Só foram incluídas nas análises posteriores seqüências onde a mutação estava pareada na estrutura de menor ΔG (maior valor de ΔG negativo). Isto foi feito para excluir as bases que possuem maior probabilidade de estarem como fita simples induzido pelo 'supercoiling' durante a transcrição (Wright et al 2003). As regiões 5' UTR e codificadoras foram excluídas da análise por estarem na forma de fita simples durante a transcrição. Foram obtidos os valores de predição de clivagem química no OrChid das seqüências normais e mutantes. Como somente 38 seqüências mutantes foram obtidas procuramos aumentar a amostragem obtendo também variações genéticas de polimorfismos encontrados em humanos, onde as formas ancestrais (presentes no chimpanzé e macaco resus) foram agrupadas juntamente com as seqüências normais e a forma nova (não presentes nos dois primatas) foi agrupada com as seqüências mutantes. Desta maneira a amostragem aumentou para 140 seqüências. É interessante notar que inicialmente foram analisadas 367 seqüências sendo que destas 227 (61,85%) foram descartadas por estarem não pareadas quando na forma de simples fita, ou seja, durante a duplicação e transcrição do DNA. Estes dados estão de acordo com outros trabalhos da literatura que mostram que seqüências que estão na forma não pareada tendem a mutar mais freqüentemente do que as formas pareadas (Wright et al. 2002, Wright et al. 2003, Todd and Glickman 1982, Kruisselbrink et al. 2008).

O aumento ou diminuição nos valores de predição de clivagem dos radicais hidroxila foram obtidos pela somatória dos valores de $(-N+M)$ (onde N corresponde ao valor OrChid da base normal e polimorfismo ancestral e M da base mutante e polimorfismo novo), para cada par de alelo normal e respectivo mutante. Por exemplo, quando o valor OrChid da base normal foi 3 e da mutada -1, teremos uma variação de -4, ou seja diminuição de 4 unidades quando ocorre a mutação. A análise estatística foi feita pelo teste t de Student para uma amostra admitindo-se que o valor médio é zero, ou seja, na população estudada não deve existir aumento ou diminuição dos valores na somatória

de $(-N+M)$. A análise estatística mostrou que a variação no valor de predição de clivagem química do DNA diferem do valor esperado (zero) ($p < 0,0001$), sendo o valor amostral médio de $-0,2$ e erro padrão de $0,05$ (Figura 11). Isto indica que o padrão das mutações não ocorre ao acaso, havendo uma diminuição média de $0,2$ na probabilidade de clivagem química das bases mutantes em relação às bases normais.

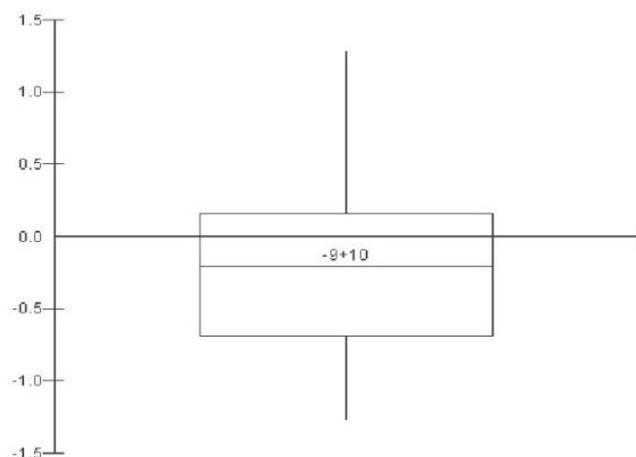


Figura 11. Gráfico de barras mostrando valores médios, desvio padrão e intervalos interquartílicos dos valores obtidos da fórmula $-N+M$, onde N representa os valores de predição de clivagem química das bases normais e M das bases mutantes.

Além desta análise os valores de N e M de cada par foram tabulados e os dados para todas as seqüências analisadas foram comparados pelo teste dos sinais (teste não paramétrico onde se define em cada par a diferença dos escores baseando-se somente no sentido das diferenças, registrados sob a forma de sinais de $+$ ou $-$) e teste de t de Student para duas amostras relacionadas (ou Wilcoxon, caso os dados sejam não paramétricos). A análise do teste dos sinais mostrou que houve uma diferença significativa ($p=0,0066$), sendo a maioria das diferenças com sinal positivo (Figura 12).

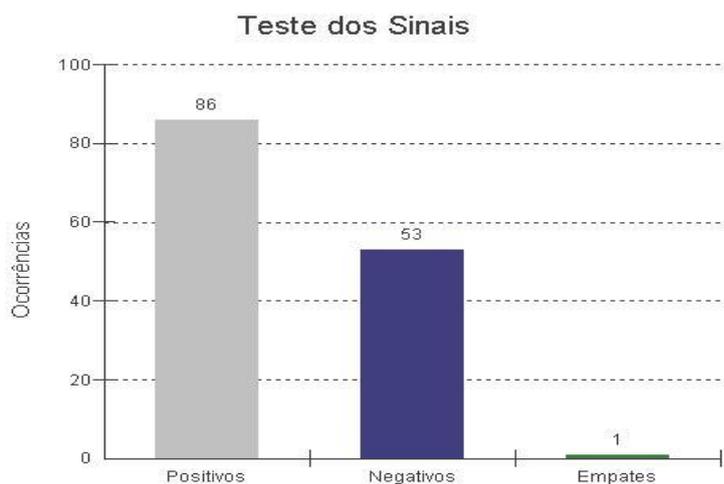


Figura 12. Gráfico de barras mostrando resultados do teste dos sinais para os valores de predição de clivagem química subtraindo-se os valores obtidos nas bases normais das mutantes.

A análise feita pelo teste de Wilcoxon (teste não paramétrico) para amostras pareadas também mostrou diferença significativa quando as sequencias normais e mutantes foram comparadas, sendo as médias das sequências normais 0,38 e das mutantes 0,18 ($p < 0,0001$, Figura 13).

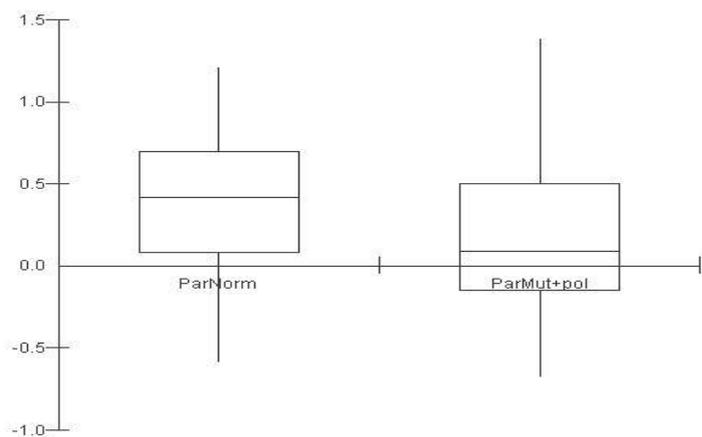


Figura 13. Mediana dos valores de predição de clivagem química obtidos no programa OrChid das sequencia normais (ParNorm) e mutantes (ParMut+Pol).

Estes resultados são realmente interessantes, pois mostram que quando ocorre uma mutação no DNA dupla fita existe uma tendência para que forma mutada seja mais estável, ou seja, tenha uma menor probabilidade para sofrer mutação do que a forma ancestral. Trabalho recentemente publicado por (Hoede et al 2006) mostrou que existe uma pressão seletiva para a diminuição do número de regiões com tendência a sofrer mutação durante a transcrição. Foi demonstrado que existe uma frequência maior do que o esperado de regiões capazes de formar pareamento intracadeia durante a transcrição, sendo esta tendência maior nos genes ativamente transcritos. A diferença do nosso trabalho em relação ao trabalho de Hoede et al (2006) é que as regiões analisadas não ficam totalmente na forma simples fita durante a transcrição gênica, porém exibem uma tendência devido a forças torcionais do DNA que está sendo transcrito

Foram também analisados os valores de predição de clivagem química das sequências que não estavam pareadas. Diferente das sequências pareadas, onde foi encontrada diferença significativa entre os grupos normal e mutante, não foram encontradas diferenças entre os grupos não pareados (Figura 14, teste de Wilcoxon). Isto mostra que as variações na estrutura do DNA que ocorrem devido a mutações e polimorfismos genéticos podem ser influenciadas pela tendência do DNA permanecer como dupla ou simples fita durante a duplicação ou transcrição. Desta maneira mutações em sequências que tendem a permanecer na forma de dupla fita causam maior variação na estrutura do DNA. Este fator pode ter implicações funcionais, pois é sabido que a estrutura secundária do DNA pode influenciar na ligação de proteínas que vão modular a expressão gênica (Rohs et al 2009). Assim existe uma maior probabilidade de alteração no padrão de ligação de proteínas ao DNA quando uma mutação ocorre em uma região que permanece na forma de dupla fita durante a transcrição gênica. Esta diferença entre as regiões pareadas e não pareadas provavelmente se deve ao fato de que o programa Orchid analisa sequências de DNA dupla fita. Desta maneira a tendência para diminuição da probabilidade de clivagem química ocorre somente nas sequências de DNA que tendem a permanecer na forma de dupla fita, por pareamento intracadeia, durante a replicação do DNA ou durante a tendência de abertura das fitas provocadas pelas forças torcionais que ocorrem nas regiões 5' durante a transcrição. As sequências que tendem a ficar na forma simples fita seguem tendência distinta mostrada por Hoede et al (2006).

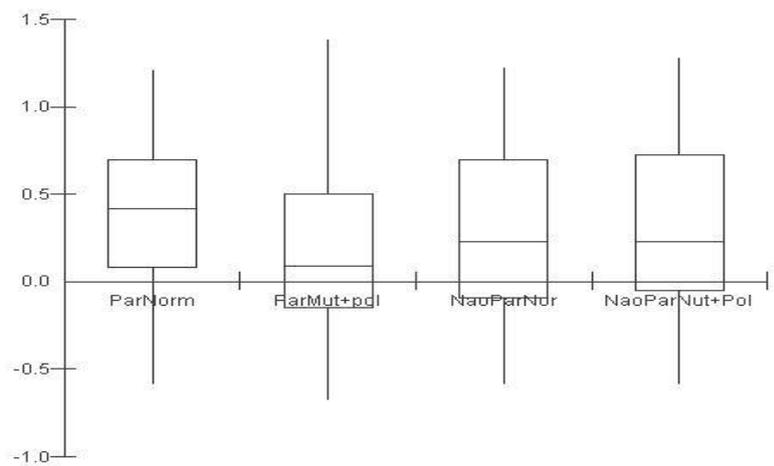


Figura 14. Mediana dos valores de predição de clivagem química obtidos no programa OrChid das sequencia normais (ParNorm) e mutantes (ParMut+Pol) pareadas e das sequencia normais (NaoParNorm) e mutantes (NaoParMut+Pol) não pareadas.

6. Conclusão

Os resultados mostraram que as mutações que causam alterações fenotípicas em humanos não alteram a estrutura do DNA de forma significativamente diferente das variações que ocorrem entre humanos e outros animais. Também não foram encontradas diferenças significativas quando se comparou a estrutura do DNA normal e mutado nas regiões codificadoras e promotoras. Existe diferença na variação do padrão de clivagem química nos diversos tipos de mutação. Sendo que as maiores variações foram encontradas nas mutações GA e CA e os menores AG e AC. É possível que as diferenças nas variações de predição de clivagem podem ter implicações na expressão genética.

As bases normais A tem um menor valor de predição de clivagem, portanto, uma menor probabilidade de sofrer mutação. É interessante notar que esta tendência ocorreu nas análises com as seqüências normais e com as seqüências mutadas. No entanto, pode-se notar que houve um aumento nos valores de predição de clivagem química quando os valores foram agrupados pelas bases mutantes.

Existe uma forte tendência de quando a base A sofre mutação ocorrer uma diminuição na probabilidade de clivagem química do DNA, independente da base mutante. O inverso também é verdadeiro, ou seja, quando a base mutante é A existe uma forte tendência para um aumento da probabilidade de clivagem química.

A mutação mais freqüente nas regiões codificadoras foi AG enquanto que nas regiões não codificadoras foi CT. Este fato provavelmente se deve ao fato de que a estrutura do DNA pode influenciar na propensão para mutações durante a transcrição gênica. A mutabilidade do DNA é maior quando o mesmo encontra-se na forma de fita simples durante a transcrição. Desta maneira, a estrutura secundária do DNA simples fita influencia na mutabilidade principalmente em regiões codificadoras. Em regiões onde ocorre o pareamento com formação de fita dupla intracadeia o DNA é 'protegido' possuindo menor tendência para mutagenicidade. Além disto, durante a transcrição, a enzima RNA polimerase desestabiliza o DNA dupla fita introduzindo um aumento na torção da dupla fita do DNA (supercoiling). O 'supercoiling' pode se estender por até 800 pares de base acima e abaixo da RNA polimerase, aumentando a probabilidade de mutações em regiões não traduzidas próximas ao sítio de início da transcrição. A torção do DNA ocorre de maneira oposta nas regiões upstream e downstream ocasionado torções horário e anti-horário respectivamente. Isto explica porque na fita sense, que foi a fita analisada em nossos experimentos a mutação CT ocorre mais nas regiões não

codificadoras enquanto que esta mutação é mais comum na fita anti-sense nas regiões codificadoras (onde a mutação AG é mais comum na fita sense).

Estes resultados sugerem que é possível prever a frequência ou probabilidade de uma mutação em um determinado local baseando-se na probabilidade de clivagem química dada pelo programa Orchid. No nosso conhecimento este é o primeiro experimento que valida o programa OrChid usando dados de mutações reais.

Esses resultados são interessantes sob dois aspectos. O primeiro é que se atribui a mutagenicidade do DNA a formação de fita simples durante a transcrição ou replicação. Nossos resultados mostram que a propensão para mutagenicidade também pode ser dependente da estrutura do DNA dupla fita. O segundo aspecto refere-se à evolução do genoma. Como ressaltado anteriormente as diferenças na proporção das bases no genoma são principalmente atribuídas a maior mutagenicidade da base C. Nossos resultados sugerem que a menor probabilidade de sofrer mutações nas bases A e T em relação às bases C e G, (principalmente a base A), pode também ter sido um fator que contribuiu para que a proporção destas bases seja maior do que a proporção das bases C e G. Ou seja, uma vez que uma mutação transformou bases C ou G em base A, esta por ter uma menor probabilidade de sofrer mutação tende a permanecer como base A e não sofrer mutações posteriores. Isto progressivamente aumentaria a proporção desta base e da sua base complementar (T) no genoma.

Quando ocorre uma mutação no DNA dupla fita existe uma tendência para que forma mutada seja mais estável, ou seja, tenha uma menor probabilidade para sofrer mutação do que a forma ancestral.

As variações na estrutura do DNA que ocorrem devido a mutações e polimorfismos genéticos podem ser influenciadas pela tendência do DNA permanecer como dupla ou simples fita durante a duplicação ou transcrição. Desta maneira mutações em sequências que tendem a permanecer na forma de dupla fita causam maior variação na estrutura do DNA. Este fator pode ter implicações funcionais, pois é sabido que a estrutura secundária do DNA pode influenciar na ligação de proteínas que vão modular a expressão gênica. Assim existe uma maior probabilidade de alteração no padrão de ligação de proteínas ao DNA quando uma mutação ocorre em uma região que permanece na forma de dupla fita durante a transcrição gênica. Esta diferença entre as regiões pareadas e não pareadas provavelmente se deve ao fato de que o programa Orchid analisa sequências de DNA dupla fita. Desta maneira a tendência para diminuição da probabilidade de clivagem química ocorre somente nas sequências de DNA que tendem a permanecer na forma de dupla fita, por pareamento intracadeia, durante a replicação do DNA ou durante a

tendência de abertura das fitas provocadas pelas forças torcionais que ocorrem nas regiões 5' durante a transcrição.

Referências

Chargaff E, Visher E, Doniger R, Green C, Misani F: The composition of the desoxyribonucleic acids of thymus and spleen. *J. Biol. Chem*; 1949. 177: 405-416.

Gardiner-Garden M, Frommer M: CpG islands in vertebrate genomes. *J Mol Biol*; 1987. 196: 261-282.

Hana Faiger, Marina Ivanchenko, Tali E. Haran: Nearest-neighbor non-additivity versus long-range non-additivity in TATA-box structure and its implications for TBP-binding mechanism. *Nucleic Acids Research*; 2007, Vol. 35, No. 13 4409–4419.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA.: Transcriptional regulatory code of a eukaryotic genome. *Nature*; 2004 Sep 2;431(7004):99-104.

Hirons L, Gardiner EJ, Hunter CA, Willett P: Structural DNA profiles: single sequence queries. *J Chem Inf Model*; 2006, 46(2):743-52.

Hoede C, Denamur E, Tenailon O: Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. *PLoS Genet*; 2006 Nov 3;2(11):e176.

Jason A. Greenbaum, Bo Pang and Thomas D. Tullius: Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res*; 2007 17: 947-953.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S: Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*; 2003. 423, 241–254.

Krasilnikov, A. S, Podtelezhnikov A, Vologodskii A, Mirkin S: Large-scale effects of transcriptional DNA supercoiling *in vivo*. *J. Mol. Biol*; 1999. 292:1149-1160.

Kruisselbrink E, Guryev V, Brouwer K, Pontier DB, Cuppen E, Tijsterman M: Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCD1-defective *C. elegans*. *Curr Biol*. 2008 Jun 24;18(12):900-5. Epub 2008 Jun 5.

Leader DP, Peter B, Ehmer B: Analysis of CpG dinucleotide frequency in relationship to translational reading frame suggests a class of genes in which mutation of this dinucleotide is asymmetric with respect to DNA strand. *FEBS Lett*. 1995 376(3):125-9.

Maston GA, Evans SK, and Green MR: Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet*; 2006. 7: 29–59.

Nikolay A. Kolchanov, Tatyana I. Merkulova, Elena V. Ignatieva, Elena A. Ananko, Dmitry Yu. Oshchepkov, Viktor G. Levitsky, Gennady V. Vasiliev, Nataly V. Klimova, Vasily M. Merkulov and T. Charles Hodgman: Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes. *Briefings in Bioinformatics Advance Access published 2007, July 12 (on line)*.

Price MA, Tullius TD: Using hydroxyl radical to probe DNA structure. *Methods Enzymol*; 1992 212: 194–219.

Raven, Peter H, George B Johnson, Susan R Singer, Jonathan Losos: *Biology*, 7th edition, New York: McGraw-Hill; 2004, January 8. 1250 pages. ISBN 978-0072921649.

Rohs R., West M. S., Sosinsky A., Liu P., Richard S. Mann & Barry Honig: The role of DNA shape in protein–DNA recognition. *Nature*; 2009. 461:1248-1254.

Sarai, A. and Kono, H: Protein-DNA recognition patterns and predictions. *Ann. Rev. Biophys. Biomol. Struct.*; 2005 34, 379–398.

Todd PA, Glickman BW: Mutational specificity of UV light in *Escherichia coli*: indications for a role of DNA secondary structure. *Proc Natl Acad Sci U S A*. 1982 Jul;79(13):4123-7.

Wright BE: A Biochemical Mechanism for Nonrandom Mutations and Evolution. *Journal of Bacteriology*; 2000, p. 2993-3001, Vol. 182, No. 11.

Wright BE, Reimers JM, Schmidt KH, Reschke DK: Hypermutable bases in the p53 cancer gene are at vulnerable positions in DNA secondary structures; 2002 *Cancer Res* 62: 5641–5644.

Wright BE, Reschke DK, Schmidt KH, Reimers JM, Knight W: Predicting mutation frequencies in stem-loop structures of derepressed genes: implications for evolution. *Mol Microbiol.* 2003 Apr;48(2):429-41.

Zuker M: Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*; 2003. 31 (13), 3406-15.