UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE de ODONTOLOGIA de PIRACICABA

ROBERTO BENTO WOLF JÚNIOR

AVALIAÇÃO DO VALOR DE PREDIÇÃO DE CLIVAGEM QUÍMICA DE RADICAIS HIDROXILA EM REGIÕES PROMOTORAS DE GENES LIGADOS AO DESENVOLVIMENTO CRANIOFACIAL.

DISSERTAÇÃO de MESTRADO APRESENTADA a FACULDADE de ODONTOLOGIA de PIRACICABA da UNICAMP para obtenção do título de MESTRE em BIOLOGIA BUCO-DENTAL na área de HISTOLOGIA e EMBRIOLOGIA

ORIENTADOR-Prof. Dr. Sergio Roberto Peres Line.
CO-ORIENTADOR-Prof. Dr. Marcelo Rocha Marques.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE/DISSERTAÇÃO DEFENDIDA PELO ALUNO, E ORIENTADA PELO PROF.DR. SERGIO ROBERTO PERES LINE.

_____
Assinatura do Orientador

PIRACICABA, 2012

FICHA CATALOGRÁFICA ELABORADA POR
MARILENE GIRELLO – CRB8/6159 - BIBLIOTECA DA
FACULDADE DE ODONTOLOGIA DE PIRACICABA DA UNICAMP

Wolf Junior, Roberto Bento, 1962-
W831a Avaliação do valor de predição de clivagem química de radicais hidroxila em regiões promotoras de genes ligados ao desenvolvimento craniofacial / Roberto Bento Wolf Junior. -- Piracicaba, SP : [s.n.], 2012.

Orientador: Sergio Roberto Peres Line.
Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Odontologia de Piracicaba.

1. DNA. I. Line, Sergio Roberto Peres, 1963- II. Universidade Estadual de Campinas. Faculdade de Odontologia de Piracicaba. III. Título.

Informações para a Biblioteca Digital

Título em Inglês: Evaluation of predictive value for chemical cleavage of hydroxyl radicals in the promoter regions of genes related to craniofacial development

Palavras-chave em Inglês:

DNA

Área de concentração: Histologia e Embriologia

Titulação: Mestre em Biologia Buco-Dental
Banca examinadora:

Sergio Roberto Peres Line [Orientador]

Ana Paula de Souza Pardo
Maria Cristina Leme Godoy dos Santos

Data da defesa: 28-02-2012
Programa de Pós-Graduação: Biologia Buco-Dental

UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Odontologia de Piracicaba

A Comissão Julgadora dos trabalhos de Defesa de Dissertação de Mestrado, em sessão pública realizada em 28 de Fevereiro de 2012, considerou o candidato ROBERTO BENTO WOLF JUNIOR aprovado.

Prof. Dr. SERGIO ROBERTO PERES LINE

Profa. Dra. MARIA CRISTINA LEME GODOY DOS SANTOS

Profa. Dra. ANA PAULA DE SOUZA PARDO

Dedicatória:

A *DEUS*, o Criador da vida; a ***ELE*** toda glória, toda honra, todo louvor; obrigado por tudo.

Aos meus pais,
Roberto (in memoriam) e ***Eunice,*** pelo amor, cuidado, sabedoria, paciência, preocupação, provisão, incentivo e pela torcida.

Aos meus irmãos,
Rosely e ***Rinaldo***, pela amizade, convivência, apoio, alegria e companheirismo.

A esposa,
Susane, pelo amor, cumplicidade, paciência, amizade, disposição, dedicação, incentivo e pelos filhos abençoados.

Aos filhos,

Willy, Franz e ***Stefany,*** pelo amor, alegria, compreensão, esforço, respeito, energia, sinceridade e por dar continuidade a nossa história.

Aos meus avôs, (in memoriam)

Oscar e ***Elizabete***, ***Aristides*** e ***Geni,*** pelo ensinamento e alegrias na infância.

Agradecimento especial

Ao meu orientador,

Prof. Dr. ***Sergio Roberto Peres Line***, pela amizade desde a graduação, oportunidade, dedicação, confiança, respeito, apoio, aprendizado, sabedoria, experiência, paciência e persistência**.**

Ao meu coorientador,

Prof. Dr. ***Marcelo Rocha Marques***, pela amizade, paciência, dicas, conversas, incentivo.

Agradecimentos

Ao Reitor da UNICAMP, Prof.Dr. Fernando Ferreira Costa.

A Faculdade de Odontologia de Piracicaba FOP-UNICAMP, na pessoa de seu diretor Prof.Dr. Jacks Jorge Junior, lugar onde tive o privilégio de cursar a graduação e especialização, e por proporcionar a oportunidade desta pesquisa.

Ao coordenador de Pós-graduação Prof.Dr. Renata Cunha Matheus Rodrigues Garcia

Ao coordenador do programa de Pós-graduação em Biologia Buco-dental Prof.Dr. Ana Paula de Souza Pardo.

Aos professores do departamento de morfologia, Área de Histologia e Embriologia da Faculdade de Odontologia de Piracicaba FOP-UNICAMP, Prof.Dr. Pedro Duarte Novaes, Profa. Dra. Darcy de Oliveira Tosello.

Aos colegas de laboratório,
Gustavo Narvaes Guimarães, Luciana Souto Mofatto, Aline Cristiane Planello, Mariana Martins Ribeiro, Simone Caixeta de Andrade, Glaucia de Camargo Pereira, Denise Carleto Andia.

Aos técnicos de laboratório,
Eliene Aparecida Orsini Narvaes e Maria Aparecida Varella

A todos que de forma direta ou indireta contribuíram para a realização deste trabalho.

Criou *DEUS*, pois, o homem à sua imagem,
À imagem de *DEUS* o criou;
homem e mulher os criou"

Gênesis 1: 27

Resumo

O inicio da transcrição gênica é um fenômeno complexo, provavelmente a principal etapa onde ocorre o controle da expressão gênica. A transcrição ocorre pela ligação de proteínas denominadas de fatores de transcrição com sequências específicas do DNA, chamadas de regiões ou seqüências "cis". A interação DNA - proteína pode depender da estrutura do DNA nestas seqüências. A interação dos fatores de transcrição com os sítios no DNA não só depende das bases onde ocorre o contato, mas pode depender também das bases vizinhas e da conformação do DNA. O presente trabalho teve por objetivo investigar a importância da conformação estrutural do DNA dupla fita de seqüências "cis" evolutivamente conservadas, localizadas em regiões promotoras (entre os nucleotídeos -1 e -500) em genes relacionados ao desenvolvimento craniofacial de espécies de mamíferos. Foi analisado se a variação na estrutura do DNA causada por variações genéticas em regiões filogeneticamente conservadas difere da variação na estrutura em regiões menos conservadas. As seqüências de DNA com 500 pares de bases das regiões 5' de 22 genes foram obtidas no site Ensembl (http://www.ensembl.org/index.html. Foram utilizadas seqüências de cinco espécies de mamíferos. A estrutura do DNA (secundária) foi estimada pela predição do padrão de clivagem química dos radicais hidroxila do DNA dupla fita, utilizando-se o algoritmo Orchid (http://dna.bu.edu/orchid/). As seqüências das cinco espécies selecionadas foram alinhadas no programa Clustalw (http://www.ebi.ac.uk/Tools/clustalw2/index.html) e as regiões conservadas com entropia 0,1(conservação de 80%) e 0,2 (70%) com tamanho mínimo de 15 pares de bases e sem lacunas (*gaps*) foram obtidas pelo programa Bioedit (http://www.mbio.ncsu.edu/bioedit/bioedit.html). Para avaliação dos valores de predição de clivagem química foi desenvolvido um algoritmo em linguagem Ruby versão 1.91 (anexo 1), onde os valores obtidos correspondiam aos valores absolutos da subtração dos valores de predição de clivagem química das bases que diferiam nas seqüências conservadas e não conservadas entre duas espécies. Comparadas regiões conservadas e não conservadas, não apresentaram diferenças estatisticamente significantes no padrão de clivagem química quando substituímos nucleotídeos na região promotora.

Palavras chave: DNA, algoritmo Ruby, sequências "cis", transcrição, fenantrolina.

Abstract

The initiation of the gene transcription is a very complex phenomenon, probably the principal stage where the control of the gene expression takes place. In general terms the transcription takes place with the connection of proteins called "transcription factors" with specific sequences of the DNA, called of regions or sequences "cis ". The interaction DNA - protein can depend on the configuration of the DNA in these sequences. The interaction of the transcription factors with the site in the DNA depends not only on the bases where the contact takes place, but they can depend also on the nearby bases and configuration of DNA. The aim of this work was to investigate the importance of the structural double band configuration of the DNA of conserved sequences in promoter region of genes (between the nucleotídeos-1 and-500), in genes involved in mammalian craniofacial development .We analyzed if the variation in the structure of the DNA caused by genetic variations in regions phylogenetically preserved differs from the variation in the structure in less preserved regions. The 500 bp sequences of DNA of the regions 5' of the 22 genes, were obtained of the site Ensembl (http: // www.ensembl.org/index.html). We used sequences of the five species of mammals. The structure (secondary) of DNA was estimated by the prediction hydroxyl radical cleavage of the DNA double strand, by the algorithm Orchid (http: // dna.bu.edu/orchid/). The sequences of five selected species were aligned in the program Clustalw (http: // www.ebi.ac.uk/Tools/clustalw2/index.html)and the regions when 0, 1 (conservation of 80 %) and 0, 2 (70 %) with at least 15 bases and without (*gaps*) were obtained using the program Bioedit (http: // www.mbio.ncsu.edu/bioedit/bioedit.html). To evaluate the predictive value of chemical cleavage, was developed an algorithm in language Ruby version 1.91 (annex), where they obtained values were corresponding to the absolute values of the subtraction of the values of prediction of fracture chemistry of the bases that were differing in the sequences preserved and not preserved between two species. Comparing conserved and not conserved regions, no statistically significant differences in standard chemical cleavage when substituted nucleotides in promoter region.

Keywords: DNA, cleavage chemistry, Ruby algorithm.

Sumário:

1-Introdução

Talvez 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 esta 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. Conforme o International Human Genome Sequencing Consortium (2004) aproximadamente 2% do genoma humano possui seqüências codificadoras (i.e. codificam proteínas), (www.genome.gov). Apesar das regiões não codificadoras terem sido consideradas de menor importância, trabalhos recentes mostram que seqüências contidas nestas regiões podem ter um papel importante na regulação da expressão gênica. A regulação da transcrição gênica pode ocorrer em várias etapas, sendo as principais a regulação da tradução e transcrição. São etapas interessantes e ainda pouco compreendidas, a regulação e 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, a ativação ocorre pela ligação de proteínas denominadas de fatores de transcrição com sequências específicas do DNA, chamadas de regiões ou seqüências "cis" (Vellozo, NC. 2010). São seqüências específicas da região promotora reconhecidas por fatores de transcrição; sítios de ligação do complexo basal de início da transcrição. Estes fatores de transcrição podem se ligar ao DNA e à 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 cofatores 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 e análise filogenética comparativa, haja visto que devido a sua importância funcional estas regiões são conservadas entre espécies. Além disso, tais análises podem ser feitas também por meio de métodos laboratoriais feitos "*in vitro*" utilizando-se fragmentos de DNA contendo a região desejada em ensaios que medem a expressão de genes marcadores ou 'reporter genes' como a luciferase do vagalume (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 à 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 doenças. 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 (Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J.).

As seqüências "cis" têm tamanho pequeno, variando entre 5 e 20 pares de bases (Nikolay *et al* 2007) e podem se localizar próximas 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 "enhancers", que aumentam a taxa de transcrição, podem ser

encontrados acima ou abaixo do gene no meio dos introns ou na região intragênica; ou "silencers" que diminuem ou inibem a transcrição por três processos: 1- competição com o ativador no sítio de transcrição; 2-interação com o domínio de ativação; 3-interação com fatores de transcrição geral(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). A clivagem ou corte de DNA foi inicialmente observada em células bacterianas relacionadas ao mecanismo de proteção contra DNA exógeno; parasitose viral; inativando o organismo invasor fragmentando o material genético do mesmo. A clivagem pode gerar dois tipos de extremidades, abruptas ou cegas, quando a clivagem ocorre no centro de simetria, ou coesivas, quando a clivagem ocorre fora do centro de simetria, podendo gerar uma extremidade saliente.

O presente trabalho teve por objetivo investigar a importância da conformação estrutural do DNA dupla fita de seqüências "cis" evolutivamente conservadas, localizadas em regiões promotoras (entre os nucleotídeos -1 e -500) em genes relacionados ao desenvolvimento craniofacial de espécies de mamíferos. Desta maneira, foi analisado se a variação na estrutura do DNA causada por variações genéticas em regiões filogeneticamente conservadas difere da variação na estrutura em regiões menos conservadas.

2-Revisão da literatura

Conformação do DNA e regulação de transcrição

A hélice dupla fita de DNA é resultado de interações entre pares de bases adjacentes. Essa interação consiste de atrações de Van der Waals e faz com que os pares de base se unam a fim de que eles mantenham contato próximo. O resultado é uma pilha de bases, uma em cima da outra, sem praticamente nenhum espaço entre elas.

O interior da dupla fita de DNA é muito hidrofóbico, essencial, pois estabiliza e protege as ligações de hidrogênio entre as bases. Estas ligações de hidrogênio não se formariam se eles fossem cercados por moléculas de água, pois cada um deles poderia ser facilmente substituído por ligações de hidrogênio com a água. A distância entre um par de bases e os próximos é de 0,33 nm, em média. Alguns dos mais fortes reflexos no padrão de difração de raios-X de DNA são devido a essa repetição de 0,33 nm.

O comprimento das moléculas dupla fita de DNA é freqüentemente expressa em termos de pares de base (Pb). As estruturas são medidas em milhares de pares bases, Kilobase, abreviada Kb. Genomas bacterianos mais consistem de uma única molécula de DNA com milhares de Kb. As moléculas de DNA nos cromossomas de mamíferos têm centenas de milhares de Kb de comprimento. O genoma humano contém 3 200 000 Kb (3 x 10^{9} pares de bases) de DNA.

A dupla hélice possui dois sulcos de largura desigual, esses sulcos são chamados sulco maior e sulco menor. Dentro de cada sulco, grupos funcionais nas bordas dos pares de bases estão expostos a água. Cada par de bases tem um padrão distinto de grupos químicos nos sulcos. As moléculas que interagem com pares de bases podem identificá-los, sem interromper a hélice. Isto é particularmente importante para as proteínas que devem se ligar a dupla fita de DNA e "ler" uma seqüência específica. Horton *et al* (2006)

Regulação da transcrição pode ser dividida em três grandes eixos de influência; genética (interação direta de um fator de controle com o gene), modulação (interação de um fator de controle com os mecanismos de) e alterações epigenéticas (sequência não-estrutura do DNA que influência da transcrição).

A interação direta com o DNA é o método mais simples e direto de uma proteína poder alterar os níveis de transcrição, genes têm frequentemente sitios para proteínas de ligação ao redor da região de codificação com a função específica de regulação da transcrição. Existem muitas classes de sítios de DNA regulatórios conhecidos como potenciadores, isoladores, repressores e silenciadores. Os mecanismos de regulação da transcrição são muito variadas, desde bloquear sítios de ligação no DNA para RNA polimerase, para atuar como um ativador e promover a transcrição, auxiliando RNA polimerase.

A atividade de fatores de transcrição é também modulada por sinais intracelulares causando modificações pós-tradução de proteínas. Essas alterações influenciam a capacidade de um fator de transcrição de ligar, direta ou indiretamente. A membrana nuclear em eucariotos permite uma regulação maior dos fatores de transcrição Ela é regulada por alterações reversíveis em sua estrutura e pela ligação de outras proteínas. Os estímulos ambientais ou sinais endócrinos podem causar modificações das proteínas reguladoras provocando uma cascata de sinais intracelulares, que resultam na regulação da expressão gênica.

Metilação do DNA é um mecanismo que influência epigenéticamente a expressão do gene, é visto em bactérias e eucariotos e tem papéis no silenciamento da transcrição hereditárias e regulação da transcrição. Em eucariotos a estrutura da cromatina, controlada pelo código de histonas, regula o acesso ao DNA com impactos significativos sobre a expressão de genes em áreas de eucromatina e heterocromatina.

Algumas proteínas especiais são necessárias para a transcrição em eucariotos, sendo os principais elementos associados ao processo de regulação da transcrição. Estas proteínas são os fatores transcricionais. Os fatores transcricionais são capazes de modular o funcionamento genético por apresentarem uma das seguintes capacidades: presença de domínios de ligação ao DNA; presença de domínios de interação a proteínas que se ligam ao DNA; presença de domínios funcionais associados à condensação do DNA.

A interação DNA - proteína pode depender da conformação do DNA, que pode ser influenciada pelas bases vizinhas. É sabido que a dupla fita de DNA forma uma estrutura tridimensional que varia conforme a composição de bases (Rohs *et al* 2009). 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). Os valores de predição de clivagem química de radicais hidroxila foram baseados em experimentos de clivagem de uma biblioteca de fragmentos de DNA dupla fita pela fenantrolina. Apesar de não mostrar a estrutura espacial tridimensional do DNA o padrão de clivagem química dos radicais hidroxilas 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 ás seqüências de nucleotídeos (Greenbaum *et al* 2007). Análises utilizando este algoritmo mostraram que seqüências diferentes no DNA dupla fita podem ter estruturas semelhantes e que mutações que alteram mais severamente a estrutura do DNA em regiões regulatórias alteram mais severamente a ligação de fatores de transcrição nestas regiões e a atividade transcricional (Parker *et al* 2009).

Bioinformática

Com a realização de estudos envolvendo químicos e geneticistas foi concluído que a molécula de DNA é que armazena toda a informação genética.

Watson e Crick foram os primeiros a apresentar a estrutura da molécula de DNA (1953), depois foram desenvolvidos novos métodos de seqüenciamento dos polímeros de DNA, permitindo o estudo das formas mais simples que o compõe.

Seqüenciadores automáticos surgiram por volta dos anos 90, aumentando muito a quantidade de informações, sendo necessários novos métodos, programas e bancos de dados para armazenamento.

A bioinformática envolve diversas áreas de conhecimento como: química, matemática, física, biologia molecular, ciência da computação. Novos estudos e técnicas na

área de genética e desenvolvimento de novos softwares ajudaram que novas frentes de pesquisa fossem iniciadas.

Começaram a serem simulados ambientes virtuais de organismos semelhantes aos reais, analisando seu comportamento e obtendo dados que antes só eram obtidos em organismos reais.

A inteligência artificial tem evoluído proporcionando análises cada vez mais completas dos dados biológicos e através de algoritmos genéticos busca a melhor solução para um problema específico. O projeto genoma alavancou a bioinformática, devido à necessidade de análise e armazenamento de grande quantidade de informações geradas por esse projeto.

No Brasil, iniciou-se em 1992 com a BBNET (BRAZILIANBIONET), na Embrapa recursos genéticos e biotecnologia.

Em 2002, na UNICAMP a construção do NBI núcleo de bioinformática, sofisticado laboratório, que executa simulações com uso de software nacional chamado de Sting, permitindo o estudo do genoma estrutural e genoma funcional.

Utilizamos nesse trabalho um algoritmo de linguagem Ruby versão 1.91.

Um algoritmo é uma descrição que mostra passo a passo os procedimentos necessários para a resolução de um problema ou tarefa; é uma seqüência lógica, finita e definida de instruções que devem ser seguidas para resolver um problema ou executar uma tarefa.

Genes Envolvidos no desenvolvimento Craniofacial

Mais de 200 genes foram identificados no desenvolvimento do dente em mamíferos (Jernvall, Thesleff, 2000, Parr, Macmahon, 1994), alguns desses genes pertencem à família HOX que contem uma região denominada homeobox, uma seqüência de DNA conservada durante a evolução. Na região homeobox encontra-se a região de homeodomínio um ligante do DNA onde se associam os fatores de transcrição (Davidson, 1995, Manzanares; Krunlauf, 2001)

A expressão combinada de membros de grupos da família de genes homeóticos, genes Hox, reguladores, desempenham importante papel na especificação posicional de estruturas no sistema esquelético. Análises de expressão em embriões mostraram que genes Hox têm uma combinação específica de alguns de seus genes, como um código Hox, responsável pela embriogênese de determinadas regiões do embrião (Hunt *et al* 1991). O código Hox responsável pela modelagem da região craniofacial de vertebrados inclui membros dos grupos de genes Muscle segment (Msx), Distal-less (Dlx), Goosecoid (Gsc) e Paired (Pax). Esses grupos de genes atuam nas regiões onde se formarão os incisivos, molares e caninos, na mandíbula em desenvolvimento (Sharpe *et al*, 1995). No desenvolvimento do dente alem dos fatores de transcrição, existem outras moléculas que tem papel importante no processo como; fatores de crescimento e moléculas da matriz extracelular (ECM); (Vaahtokari, Vainio, Thesleff, 1991). Fatores de transcrição são moléculas que interagem com o DNA modulando a expressão de um gene. A inativação no gene que codifica um fator de transcrição pode modificar o fenótipo, especialmente se sua expressão ocorrer nos estágios iniciais da embriogênese (Satokata, Maas, 1994).

Abaixo, descrição dos genes envolvidos no desenvolvimento craniofacial:

DLX- Distal-less homeobox é composto por seis membros no genoma de mamíferos que são arranjados em três pares mais estreitamente ligados: Dlx1, Dlx2-Dlx7, Dlx3-Dlx6, Dlx5(Weiss, Stock, Zhao, 1998). Os genes Dlx1 e Dlx2 são expressos nos processos mandibulares e maxilares. O domínio de expressão do Dlx1 nos arcos mandibulares é mais distalmente restringida do que o Dlx2, Dlx5 e Dlx6. A expressão dos genes Dlx3 e Dlx7 no arco mandibular são coincidentes com a área formadora de elementos dentais (Zhao, 2000). O papel do Dlx3 na amelogênese foi comprovado por uma mutação em humanos que foi associada com hipoplasia do esmalte (Price *et al*, 1998). A inativação do Dlx5 afeta a maturação do esmalte dental (Depew *et al*, 1999).

Lymphoid enhancer-binding factor 1; Lef1 é um fator de transcrição expresso em linfócitos de ratos adultos , em células da crista neural, germes dentais, folículos pilosos, e outros tecidos durante a embriogênese (Travis *et al,* 1991; Waterman *et al*, 1991; Oster Wegel, *et al,* 1993; Vangederen *et al*, 1994; Zhou et al, 1995). Esse gene foi mapeado no cromossomo humano 4.

Muscle segment Box, pertence a um grupo de genes altamente conservados dentro da escala evolutiva, família chamada de genes homeóticos. Em humanos existem os genes Msx1 e Msx2, que não se localizam no mesmo cromossoma (Davidson, 1995) e estão estreitamente envolvidos na odontogênese (Maas, Bei, 1997). O gene Msx1 foi localizado no cromossomo 4p16. 3-p16. 1, enquanto o Msx2 esta no cromossomo 5q34-q35. Nos humanos o papel dos genes Msx no desenvolvimento crânio facial tem sido esclarecido em estudos que identificaram mutações nesses genes associadas a alterações da normalidade. A transversão na região homeobox do gene Msx1 resulta na substituição de uma arginina por uma prolina, em um domínio de conservação da proteína, causando uma forma de agenesia dental, no entanto, mutações nos Msx não explicam todas as formas de agenesia dental (Scarel, *et al*, 2000).

Paired Box, é formado de 9 membros em mamíferos, que apresentam uma região conservada ligante do DNA , nomeada domínio Paired (Underhill, 2000), sendo que os membros mais ligados a odontogênese são Pax1 e Pax9. Mutações nos genes Pax causam profundos defeitos no desenvolvimento em organismos diversos como moscas, ratos e humanos (Epsten, 2002), ligados a oligodontia em humanos, afetando principalmente os posteriores da dentição permanente (Stockton *et al*, 2000) . O gene Pax9 em humanos é encontrado no cromossomo 14q12-q13. Mutação nesse gene esta ligada a agenesia dental de uma família com oligodontia autossômica dominante, onde foram afetados pré-molares e molares (Stockton *et al*,2000)

Fatores de crescimento são importantes moléculas sinalizadoras. Um fator de crescimento produzido e liberado por uma célula pode afetar o desenvolvimento de outra célula na vizinhança (Scarel *et al*, 2003). Os efeitos dos fatores de crescimento são sempre diretamente mediados pela ligação de receptores de superfície em células específicas (Thesleff, 1995). Os sinais mais estudados são os da família Fgf, Egf, Tgf, cada família consiste de vários sinais codificados por diferentes genes (Thesleff, 2000).

Fator de crescimento de fibroblasto é uma grande família de proteínas que tem efeitos morfogenéticos potentes em vários órgãos e são também potentes estimuladores da proliferação celular. Eles induzem a divisão celular no mesênquima e no epitélio em vários

estágios da morfogênese do dente (Jernvall, 1994, Kettunen & Thesleff, 1998). Elas também previnem apoptose no mesênquima dental (Vaahtokari, Aberg; Thesleff, 1996).

Fator de crescimento epidermal, são mitógenos para a ectoderme, endoderme e mesoderme, estimulando a proliferação das células do embrião durante a morfogênese (Carpenter; Cohen, 1979).

Fator de crescimento transformante, regulam a proliferação celular, diferenciação e apoptose, controlando o desenvolvimento e manutenção de vários tecidos (HSU *et al*, 2002; Peres et al 2004) , atua durante os estágios iniciais do desenvolvimento do dente(Shimo *et al*, 2002; Vaahtokari, Vainio, Thesleff, 1991).

Proteínas morfogenética do osso, regulam o desenvolvimento dos ossos e cartilagens (Deconto *et al*, 2004), família de Bmp em mamíferos formada por 8 membros.Bmp2 e Bmp4 têm 95% de similaridade entre si, e podem ser um fator chave para iniciação e morfogênese do dente (Jernvall, 1994, Kettunen & Thesleff 1998). Moléculas da matriz extracelular estão envolvidas na interação epitélio-mesênquima durante a morfogênese e diferenciação do dente. A odontogênese pode ser alterada por mutações de genes do colágeno e proteoglicanas (Maas, Bei, 1997). A integridade da membrana basal é pré-requisito para morfogênese epitelial do dente (Sahlberg, Akhil, Thesleff, 2001). Na formação do elemento dental, a membrana basal do epitélio contem colágeno do tipo 1, 3 e 4, laminina, fibronectina e proteoglicanas. (Thesleff *et al*, 1981). Essas moléculas são expressas ao mesmo tempo em que a interação mediada pela membrana basal regula a diferenciação de células ectomesenquimais em odontoblastos (Lesot *et al*, 1990).A sinalização molecular no desenvolvimento do dente é expressa nos diferentes estágios da odontogênese.A expressão do Msx1 no mesênquima dental é inicialmente induzida por derivados epiteliais de Bmps e Fgfs. No estágio de botão de desenvolvimento, a regulação do Bmp4 no mesênquima de molar, causa a regulação de Lef1 e Dlx2 no botão epitelial.

A seguir os genes pesquisados com suas principais funções e localização:

ACAN- codifica proteína que integra a matriz extracelular em cartilagem, dando resistência à compressão, adesão; quando ocorre mutação nesse gene pode ocorrer displasia óssea e degeneração espinhal. Localiza-se no cromossoma 15.

AMBN-(ameloblastin)- codifica proteína da matriz extracelular; envolvido na calcificação do esmalte, quando ocorre mutação pode causar dentinogênese imperfeita, amelogênese imperfeita, ameloblastoma ou tumor odontogênico. Localiza-se no cromossoma 4.

AMELX- codifica a proteína amelogenina; envolvido na formação do esmalte, essencial ao desenvolvimento dentário; quando ocorre mutação pode causar amelogênese imperfeita, ligado ao cromossoma x afetando com mais severidade os homens.

AMTN-(amelotin)- codifica proteína de adesão celular envolvida na maturação do esmalte, ligado ao cromossoma 4.

COL1A1- codifica parte da grande molécula chamada colágeno tipo 1, proteína que dá resistência e suporte para a maioria dos tecidos. Mutação nesse gene pode causar alteração em cartilagem, osso, tendões, pele, osteogênese imperfeita. Localiza-se no cromossoma 17.

COL1A2-codifica parte da grande molécula de colágeno tipo 1,mutação nesse gene causa osteogênese imperfeita. Localiza-se no cromossoma 7.

COL3A1-codifica parte da molécula de colágeno tipo 3,envolvido na formação da pele, pulmão, paredes do intestino, parede dos vasos sanguíneos. Localiza-se no cromossoma 2.

FN1- codifica a fibronectina, uma glicoproteína presente no plasma, superfície da célula, matriz extracelular, envolvida também na adesão celular e processos de migração incluindo embriogênese, coagulação sanguínea, defesa e metástase. Localiza-se no cromossoma 2.

LAMA1-codifica laminina alpha 1, responsável pelo encaixe, migração e organização das células no tecido. Localiza-se no cromossoma 18.

LAMB2-codifica laminina beta 2, glicoproteina da matriz extracelular, é o maior não colágeno constituinte da membrana basal,responsável pela adesão celular, diferenciação, migração.Localiza-se no cromossoma 3.

MMP3- codifica enzima metolopeptidase 3- remodela a matriz extracelular em processos de normalidade como o desenvolvimento embriológico, reprodução; pode degradar: fibronectina, laminina, colágeno, proteoglicanas. Localiza-se no cromossoma 11.

MMP7- codifica enzima metolopeptidase 7- remodela a matriz extra celular, pode degradar: caseína, fibronectina. Localiza-se no cromossoma 11.

MMP9- codifica enzima metolopeptidase 9- remodela a matriz extra celular , ligada a migração de leucócitos, reabsorção óssea; mutação pode causar artrite, ulcera, encefalomielite e câncer.Localiza-se no cromossoma 20.

MMP10- codifica enzima metolopeptidase 10, pode degradar: fibronectina, colágenos. Localiza-se no cromossoma 11.

MMP12- codifica enzima metolopeptidase 12, remodela a matriz em tecido traumatizado. Localiza-se no cromossoma 11.

MMP13-codifica metolopeptidase 13- degrada colágeno tipo 1; tem ação sobre caseína.Localiza-se no cromossoma 11.

MMP20- codifica metolopeptidase 20, degrada amelogenina, a maior componente da matriz do esmalte. Localiza-se no cromossoma 11.

MMP27-codifica enzima metolopeptidase 27, degrada proteínas componentes da matriz extracelular como fibronectina, laminina, colágeno. Localiza-se no cromossoma 11.

OSTCAL- (osteocalcin), a mais abundante proteína não colágena do tecido ósseo encontrada em osteoblasto, condroblasto e odontoblastos; presente na dentina, regulação da formação óssea, envolvida na mineralização e homeostasis de íon cálcio. Quando mutado pode causar dentinogênese imperfeita.

TIMP2- codifica enzima metolopeptidase inibidora 2, inibidora de metaloproteinases da matriz extracelular, suprime a proliferação de células endoteliais. Localiza-se no cromossoma 11. Mutação pode causar hipodontia.

TNC- codifica proteína tenascin C, proteína da matriz extracelular envolvida no direcionamento da migração de neurônios durante a fase de desenvolvimento, plasticidade sináptica e regeneração neuronal. Localiza-se no cromossoma 4.

TUFT1- codifica proteína tuftelin envolvida na mineralização do esmalte. Formada durante a amelogênese. Localiza-se no cromossoma 1.Mutação causa amelogênese imperfeita .

3- Proposição

O presente trabalho teve por objetivo investigar a importância da conformação estrutural do DNA dupla fita de seqüências "cis" evolutivamente conservadas, localizadas em regiões promotoras (entre os nucleotídeos -1 e -500) em genes de espécies de mamíferos. Desta maneira, foi analisado se a variação na estrutura do DNA causada por variações genéticas em regiões filogeneticamente conservadas difere da variação na estrutura em regiões menos conservadas.

4- Material e métodos

As seqüências de DNA 500 pares de bases das regiões 5' dos genes: *ACAN,AMBN,AMELX,AMTN,COL1a1,COL1a2,COL3a1,FN1,LAMA1,LAMB2,MMP3,MMP7,MMP9,MMP10,MMP12,MMP13,MMP20,MMP27,OSTCAL,TIMP2,TNC,TUFT1*, foram obtidas do site Ensembl (http://www.ensembl.org/index.html). Este site contém seqüências de DNA curadas o que confere um alto grau de confiabilidade às mesmas (Figura 1). Foram coletadas somente seqüências de DNA de animais mamíferos. Após a obtenção das seqüências de DNA foram selecionadas apenas aquelas onde o seqüenciamento era completo, sem lacunas ou falhas. Foram utilizadas as seqüências das seguintes espécies: *Homo sapiens* (homem), *Canis familiaris* (cão) ou na ausência desta *Bos taurus* (boi), *Macaca mulata* (macaco rhesus) ou na ausência desta *Pongo pygmaeus* (orangotango), *Callithrix jacchus* (sagüi), *Mus musculus* (camundongo) ou na ausência desta espécie *Rattus norvegicus* (rato).

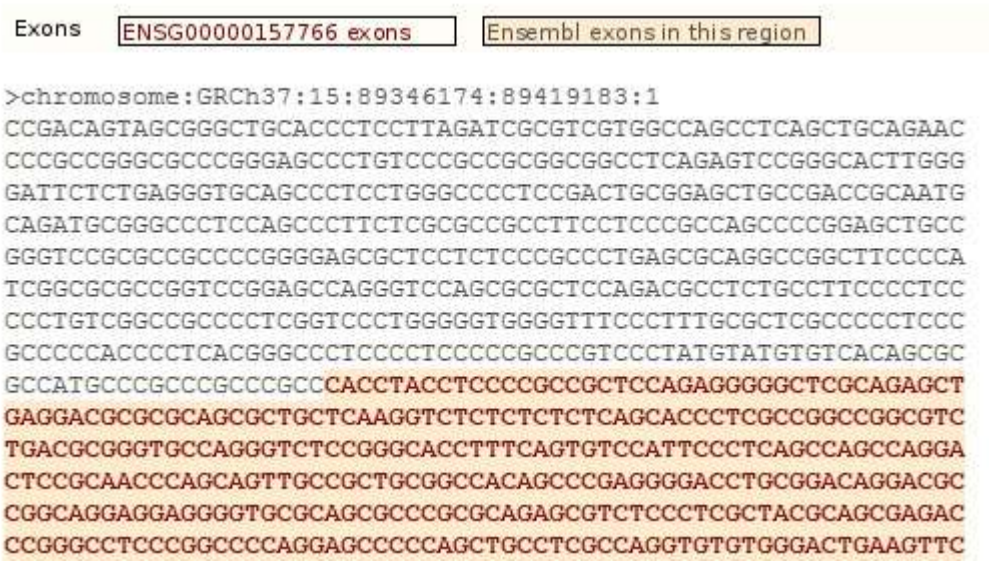


Figura 1. Figura obtida do site Ensembl mostrando seqüência de 500 pares de bases da região promotora do gene ACAN seguida do primeiro exon do gene (em destaque).

A predição do padrão de clivagem química dos radicais hidroxila do DNA dupla fita das espécies listadas acima foi feita utilizando-se o algoritmo Orchid (http://dna.bu.edu/orchid/) (Price e Tullius 1992, Greenbaum *et al* 2007).

As seqüências das 5 espécies selecionadas foram alinhadas no programa Clustalw (http://www.ebi.ac.uk/Tools/clustalw2/index.html) e as regiões conservadas com entropia 0,1(conservação de 80%) e 0,2 (70%) com tamanho mínimo de 15 pares de bases e sem lacunas (*gaps*) foram obtidas pelo programa Bioedit (http://www.mbio.ncsu.edu/bioedit/bioedit.html, Figura 2.).

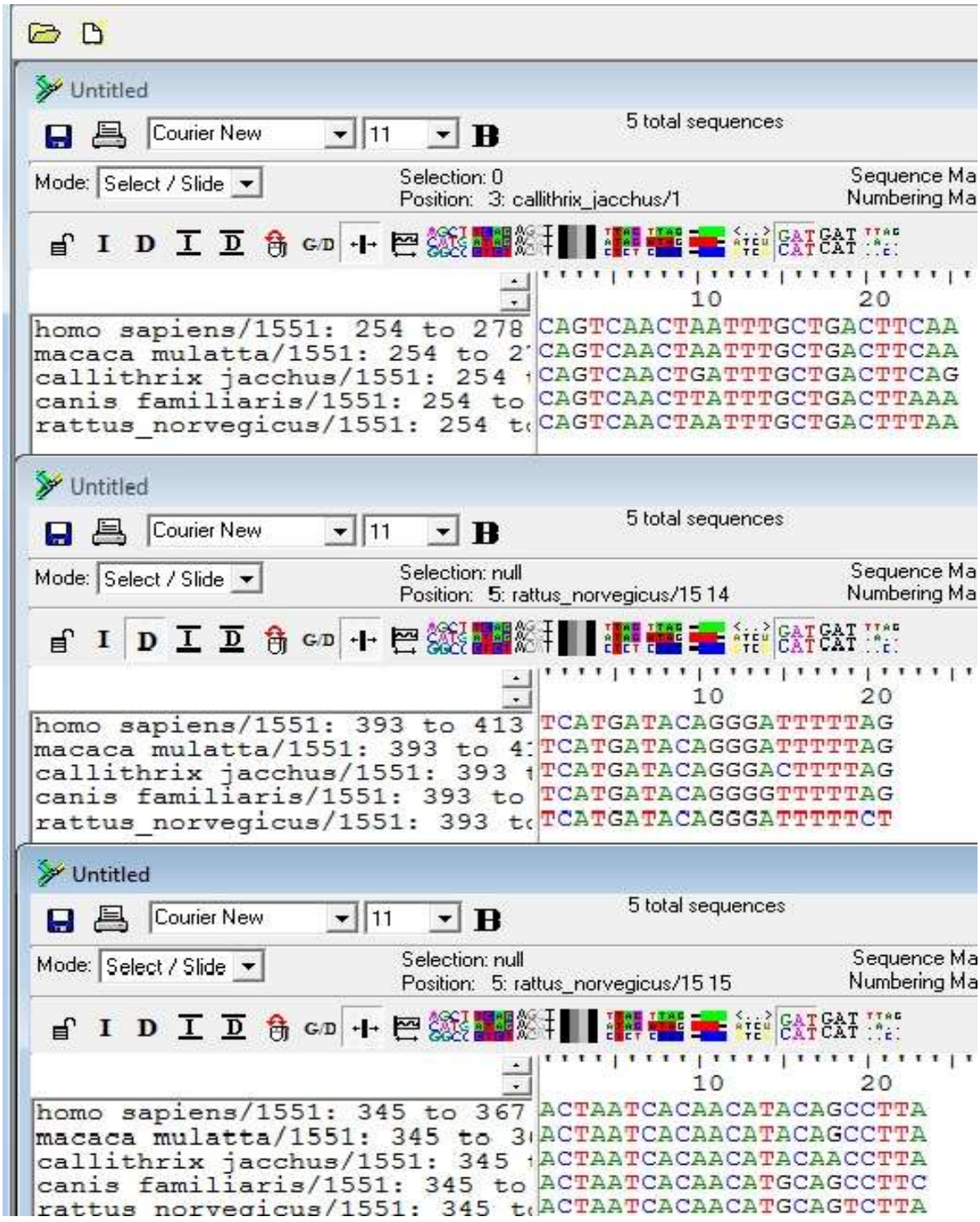
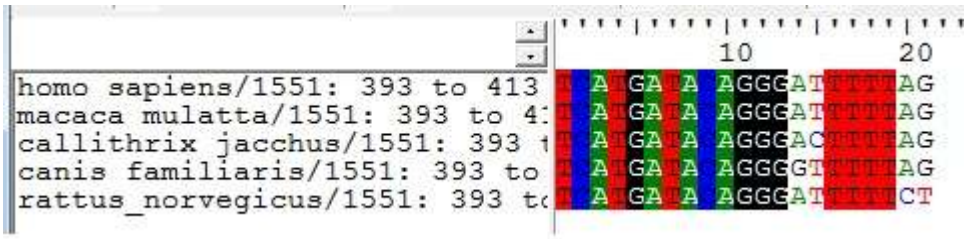


Figura 2.. Obtenção de seqüências conservadas no programa Bioedit. Observar três seqüências conservadas (entropia mínima de 0,1) entre as 5 espécies listadas os valores 254 to 278, 393 to 413 e 345 to 367, listados nas 3 seqüências correspondem a localização destas seqüências dentro das seqüências alinhadas.

A obtenção dos valores de predição de clivagem química dos radicais hidroxila nas regiões conservadas e não conservadas foi desenvolvido um algoritmo em linguagem Ruby versão 1.91 (anexo 1), onde os valores obtidos correspondiam aos valores absolutos da subtração dos valores de predição de clivagem química das bases que diferiam nas seqüências conservadas e não conservadas entre duas espécies. A figura 3 abaixo permitirá um melhor entendimento desta metodologia:

A



B-exemplo:

Homo sapiens	T	C	A	T	G	A	T	A	C	A	G	G	G	A	T	T	T	T	T	A	G
Rattus norvegicus	T	C	A	T	G	A	T	A	C	A	G	G	G	A	T	T	T	T	T	C	T
Homo sapiens	1	2	1	3	2	3	4	1	1	2	2	2	3	2	1	3	2	2	4	2	1
Rattus norvegicus	2	2	1	3	2	3	4	1	2	1	1	1	2	3	2	2	3	1	3	1	2
RES	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1

Figura 3. Na parte superior da figura (A) figuras acima aparecem seqüências de DNA conservadas de 21 pares de bases de cinco espécies no programa Bioedit. (B) exemplo- Seqüências *Homo sapiens e Rattus norvegicus* da figura 3A e os valores de predição de clivagem química de cada base das respectivas espécies. A linha RES mostra os valores finais que correspondem aos valores absolutos da subtração dos valores de predição de clivagem química dos radicais hidroxila somente nos locais onde as seqüências de DNA diferem (i.e. neste caso nas últimas 2 bases não conservadas entre *Homo sapiens e Rattus norvegicus*). Número de bases diferentes é igual a 2, **n** é igual a 2.

As significâncias estatísticas entre os valores entre as seqüências conservadas e não conservadas foi determinada pelo teste Mann-Whitney.

5- Resultados

As figuras 4 e 5 mostram a distribuição das diferenças dos valores de predição de clivagem química das bases não conservadas nas regiões conservadas (entropia 0,1 ou 0,2) entre *Homo sapiens* e *Callithrix jacchus* e também *Homo sapiens* e *Mus musculus* (ou *Rattus norvegicus*, na ausência das seqüências de *Mus musculus* para um determinado gene). As figuras mostram que a maioria dos valores obtidos está abaixo de 0,60 e que os valores médios das espécies *murinas* estão abaixo dos valores obtidos para *Callithrix.*

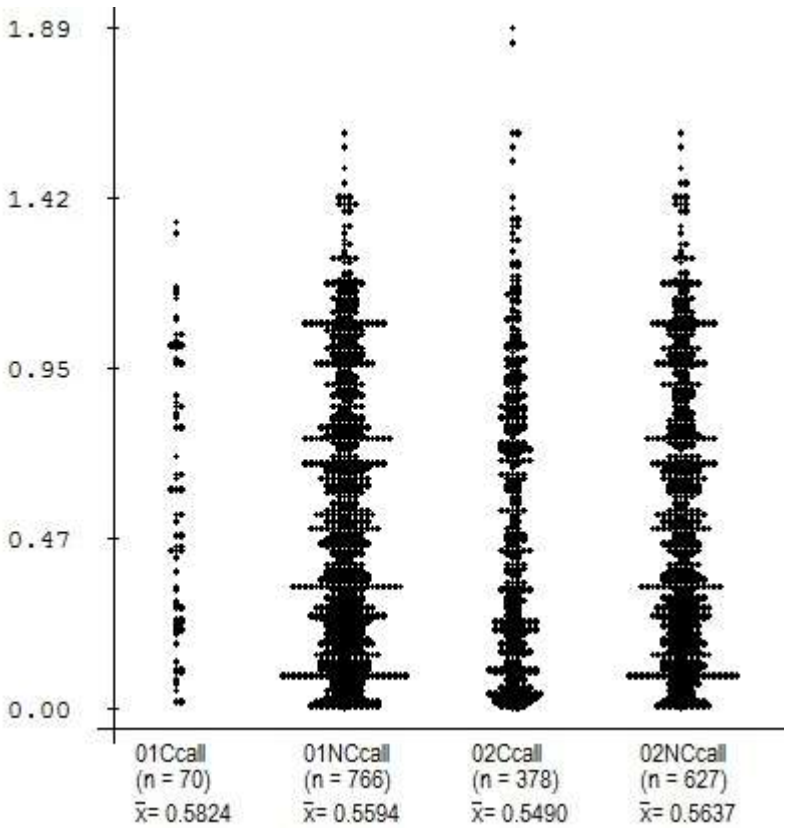


Figura 4. Gráfico da distribuição das variações dos valores absolutos de predição de clivagem química entre *Homo sapiens e Callithrix jacchus* (eixo Y). 01Ccall = valores obtidos das bases variantes dentro de regiões conservadas com entropia menor ou igual a 0,1. 01NCcall = valores obtidos das bases variantes das seqüências conservadas com entropia maior do que 0,1. 02Ccall = valores obtidos das bases variantes dentro de regiões conservadas com entropia menor ou igual a 0,2. 02NCcall = valores obtidos das bases variantes das seqüências conservadas com entropia maior do que 0,2. **n=** número amostral (número de bases diferentes) e ẍ= valores médios.

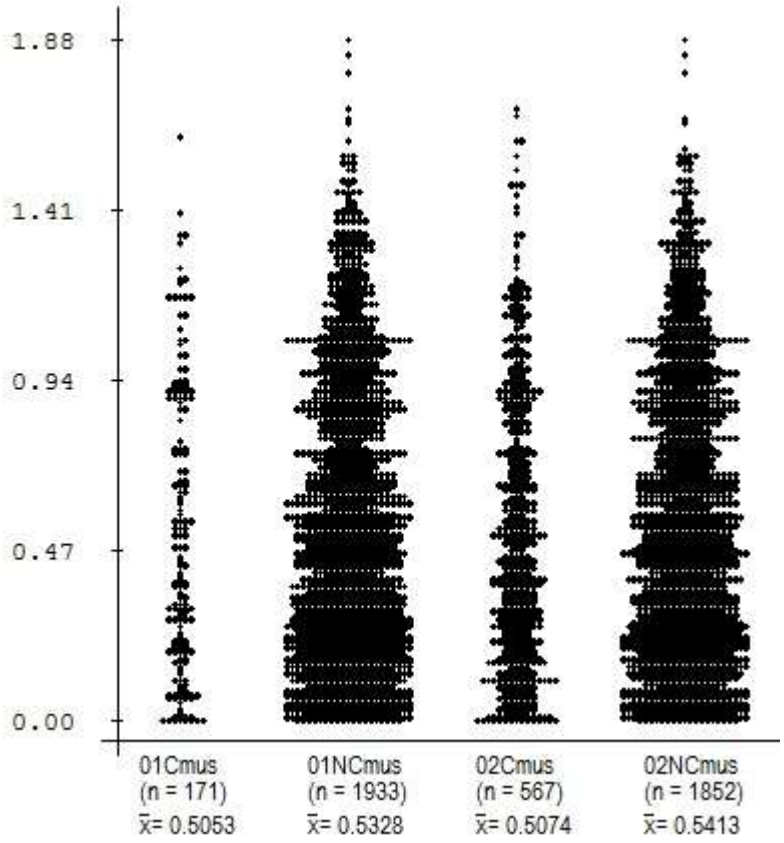


Figura 5. Gráfico da distribuição das variações dos valores absolutos de predição de clivagem química entre *Homo sapiens e Rattus norvegicus* ou *Mus musculus* (eixo Y). 01Cmus = valores obtidos das bases variantes dentro de regiões conservadas com entropia menor ou igual a 0,1. 01NCmus = valores obtidos das bases variantes das seqüências conservadas com entropia maior do que 0,1. 02Cmus = valores obtidos das bases variantes dentro de regiões conservadas com entropia menor ou igual a 0,2. 02NCmus = valores obtidos das bases variantes das seqüências conservadas com entropia maior do que 0,2. n= número amostral (número de bases diferentes) e ẍ= valores médios

Antes de se fazer as comparações dos valores entre as regiões conservadas foram feitos testes estatísticos para determinação de normalidade (i.e. se os valores obtidos em cada análise apresentavam distribuição normal), pois valores com distribuição normal podem ser analisados por testes paramétricos. O teste utilizado foi o de D'Agostino-Pearson, onde a hipótese de nulidade prediz que os valores estão normalmente distribuídos. Os resultados, apresentados na figura 3 mostram que para 7 das 8 análises o valor de ***p*** foi menor do que 0,05, rejeitando-se a hipótese de nulidade. Desta maneira empregamos teste não para métrico de Mann-Whitney para a comparação entre os grupos.

Análise de D'Agostino-Person dos dados obtidos:

	01Ccall	01NCcall	02Ccall	02NCcall	01Cmus	01NCmus	02Cmus	
Resultados	- 1 -	- 2 -	- 3 -	- 4 -	- 5 -	- 6 -	- 7 -	- 8 -
Tamanho da amostra =	70	766	171	1933	378	627	567	1852
G1 =	0.2484	0.3715	0.5562	0.6004	0.5385	0.3832	0.6753	0.5679
G2 =	-1.1620	-0.8889	-0.7312	-0.5119	-0.3826	-0.8409	-0.3099	-0.5354
Zg1 =	0.8930	1.5228	2.0979	2.3779	2.1099	1.5623	2.5957	2.2642
Zg2 =	1.7497	3.7053	1.7305	3.7704	1.4432	3.2533	1.4370	3.8281
K2 =	3.8587	16.0482	7.3956	19.8708	6.5344	13.0245	8.8027	19.7809
Graus de liberdade =	2	2	2	2	2	2	2	2
p =	0.1452	< 0.0001	0.0248	< 0.0001	0.0381	0.0015	0.0123	< 0.0001

Figura 6. Teste de D'Agostino-Pearson para os valores dos 8 grupos apresentados nas figuras 4 e 5. Notar que apenas o primeiro valor, que corresponde ao grupo 01Ccall não apresentou distribuição normal. Os quatro primeiros números correspondem aos valores da figura 4, e os quatro seguintes correspondem aos valores da figura 5.

As comparações entre as bases variantes entre *Homo sapiens* e *Callithrix jacchus* nas regiões conservadas (C) e não conservadas (NC) nas entropias 0,1 e 0,2 não mostraram diferenças significativas. Quando os valores das regiões com entropia menor ou igual a 0,1 (conservadas) foram comparadas com os valores das regiões com entropia maior a 0,1(não conservadas), obteve-se um valor de $\boldsymbol{p}$ = 0,6. Para a entropia menor ou igual a 0,2(conservadas) quando comparadas com regiões com entropia maior que 0,2 (não conservadas) o valor de $\boldsymbol{p}$ foi 0,31. Figura 7.

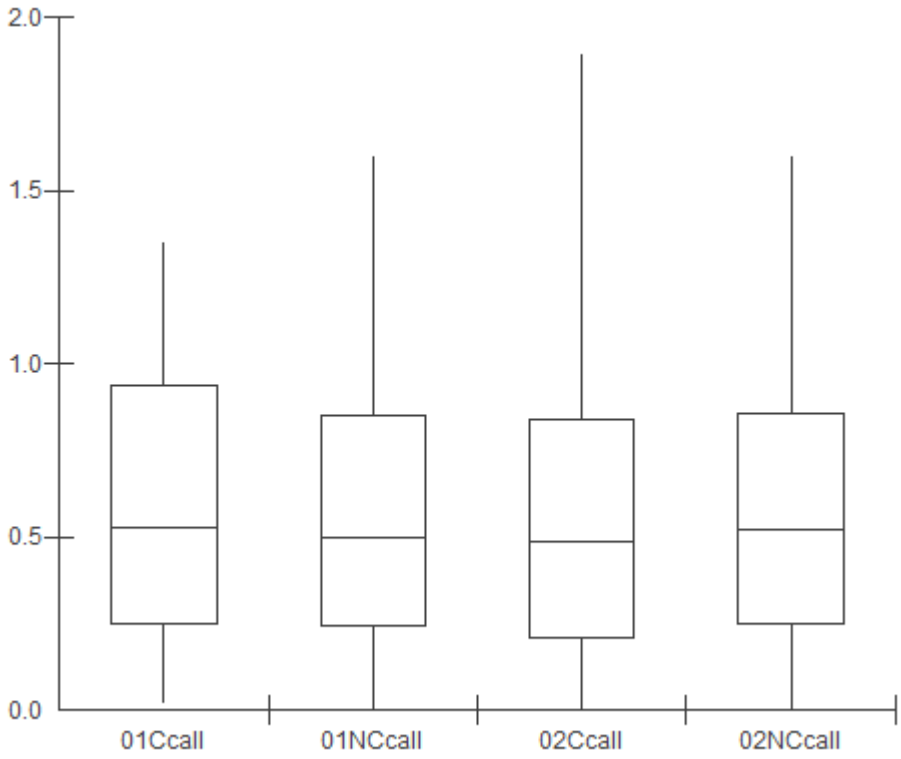


Figura 7. Gráficos mostrando medianas e intervalos interquartílicos dos valores absolutos de predição de clivagem química entre *Homo sapiens e Callithrix* jacchus (eixo Y). 01Ccall= valores obtidos das bases variantes dentro de regiões conservadas com entropia menor ou igual a 0,1. 01NCcall = valores obtidos das bases variantes das seqüências conservadas com entropia maior do que 0,1. 02Ccall = valores obtidos das bases variantes dentro de regiões conservadas com entropia menor ou igual a 0,2. 02NCcall = valores obtidos das bases variantes das seqüências conservadas com entropia maior do que 0,2.

As comparações entre as bases variantes entre *Homo sapiens* e *Rattus norvegicus ou Mus musculus* nas regiões conservadas (C) e não conservadas (NC) nas entropias 0,1 e 0,2 não mostraram diferenças significativas. Quando os valores das regiões com entropia menor ou igual a 0,1 (conservadas) foram comparadas com os valores das regiões com entropia maior a 0,1(não conservadas) obteve-se um valor de $\boldsymbol{p}$ = 0,32. Para a entropia menor ou igual a 0,2 (conservada) comparada com os valores das regiões maiores que 0,2(não conservada) o valor de $\boldsymbol{p}$ foi 0,67. Figura 8,

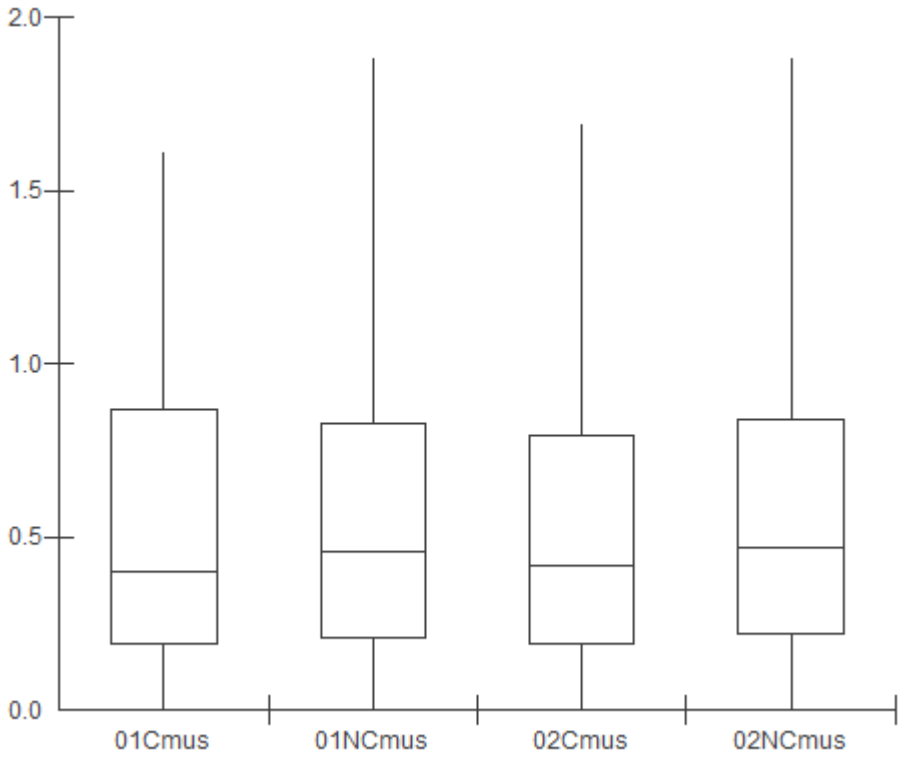


Figura 8. Gráficos mostrando medianas e intervalos interquartílicos dos valores absolutos de predição de clivagem química entre *Homo sapiens e Rattus norvegicus* ou *Mus musculus* (eixo Y). 01Cmus = valores obtidos das bases variantes dentro de regiões conservadas com entropia menor ou igual a 0,1. 01NCmus = valores obtidos das bases variantes das seqüências conservadas com entropia maior do que 0,1. 02Cmus = valores obtidos das bases variantes dentro de regiões conservadas com entropia menor ou igual a 0,2. 02NCmus = valores obtidos das bases variantes das seqüências conservadas com entropia maior do que 0,2.

6- Discussão

A bioinformática é considerada uma ciência interdisciplinar que envolve biologia molecular, química molecular, física, matemática, ciência da computação, entre outros (Shen & Tuszynski, 2008). O desenvolvimento de ferramentas de bioinformática tem permitido uma análise precisa e rápida de seqüências de proteínas, DNA e RNA.

Um algoritmo pode ser definido como um conjunto previamente especificado de passos procedimentais referentes a um problema computacional delimitado incluído em um software específico (Pressman, 1995). Várias linguagens de programação (como C++, Java, Python, Perl e Ruby) têm sido vastamente utilizadas para a produção de algoritmos de bioinformática. Os resultados apresentados no presente trabalho foram obtidos com ferramentas de bioinformática obtidos de sites na internet e também em programa desenvolvido em nosso laboratório. O programa foi desenvolvido em linguagem Ruby, que possui as vantagens de ser orientado a objetos, multi-plataforma, dinâmica e interpretado (Black, 2006); livre, rápida, *open-source* (código aberto), fácil de realizar download e de instalar, fácil de aprender e ler (Berman, 2008). Esta linguagem foi criada por Yukihiro Matsumoto em 1995, tornando-se mundialmente conhecida a partir de 2000 (Black, 2006). A linguagem Ruby é freqüentemente utilizada para processar arquivos de texto e tarefas de manutenção do sistema, que são ferramentas necessárias em bioinformática (Goto *et al*, 2003). Ruby é uma linguagem estritamente Orientada a Objetos. Isto significa que tudo em Ruby é um objeto com atributos definidos que podem ser manipulados utilizando-se métodos associados. Objetos são instâncias de classes, (Berman, 2008). Já foram publicadas várias aplicações em bioinformática usando Ruby descritas pela literatura. Prince & Marcotte (2008) utilizaram esta linguagem para realizar uma análise proteômica de espectrometria de massa. Goto e colaboradores (2003) criaram uma biblioteca de classes adequada para aplicações em bioinformática, no qual descreveram e manipularam complicadas estruturas de dados biológicos. Essa biblioteca foi denominada BIORUBY e possui suas classes e métodos capazes de: manipular seqüências biológicas, acessar bancos de dados biológicos, analisar entradas em banco de dados, executar aplicações de análise

biológica e analisar os seus resultados (Nakao *et al*, 2004). O programa desenvolvido para o presente trabalho permite a entrada de dados a partir de três arquivos em formato texto (. txt, anexo 1). Sendo que, um dos arquivos, contem as seqüências alinhadas em formato fasta (inicia com >), outro arquivo contém as seqüências conservadas em formato fasta com a entropia colocada na primeira linha e na última seqüência contém os valores obtidos no site Orchid (anexos 2, 3, 4). Os resultados são escritos em arquivo em formato texto de onde são copiados e inseridos no programa de análise estatística.

Começamos o trabalho com várias espécies, mas não tínhamos uma seqüência completa de todas, fomos então reduzindo o número. No caso de *Mus musculus/Rattus norvegicus* em alguns genes, não obtivemos a seqüência completa então usamos do outro.

De maneira interessante as médias e medianas das regiões conservadas (entropias 0,1 e 0,2) obtidas entre *Homo sapiens* e *Mus musculus/Rattus norvegicus* foram menores do que as regiões não conservadas (entropias 0,1 e 0,2), enquanto que as medidas obtidas entre *Homo sapiens* e *Callithrix jacchus* foram muito semelhantes. Apesar de não serem estatisticamente significantes ($\boldsymbol{p} > 0,05$), se mantida esta tendência a significância estatística poderia ser obtida se aumentássemos o número amostral analisando um número maior de genes. Desta maneira, poderíamos especular que as variações nas regiões conservadas entre as espécies *murinas* e humanos resultaram em maior variação na estrutura de regiões conservadas. Este dado é compatível com a maior distância filogenética entre estas espécies e possivelmente também reflete maiores diferenças funcionais entre estas espécies do que entre humanos e calitriquídeos. Vale lembrar que humanos e calitriquídeos divergiram entre 23,5 e 34 milhões de anos enquanto que humanos e murinos divergiram provavelmente entre 61,7 e 100,5 milhões de anos (http://www.fossilrecord.net/dateaclade/, Donoghue, 2007)

Sabe-se que a estrutura tridimensional da molécula do DNA genômico, mais especificamente o formato da porção glicosídica e sulcos podem ser afetadas por variações nos nucleotídeos (Hoede *et al* 2006). Isto por sua vez pode interferir na ligação de proteínas ao DNA e afetar a expressão protéica e conseqüentemente o fenótipo celular (Rohs *et al* 2009). Baseado em predição de clivagem química por radicais hidroxila, Parker *et al*

(2009) desenvolveram um algoritmo que compara similaridade da estrutura do DNA entre espécies. Estes autores mostraram que 12% do genoma humano é estruturalmente conservado, dobrando o número de bases conservadas quando se usa apenas padrão de conservação de bases. Além disto, as regiões com topografia conservada correlacionaram-se com elementos não codificantes funcionais (fatores de transcrição), importantes para a regulação da expressão gênica, de maneira mais freqüente do que regiões identificadas somente por similaridade de bases. No entanto, nossos resultados mostraram que as substituições de nucleotídeos na região promotora dos genes estudados não causaram diferenças estatisticamente significantes no padrão de clivagem química por radicais de hidroxila quando as regiões conservadas foram comparadas com regiões não conservadas entre *Homo sapiens* e *Mus musculus* ou *Rattus norvegicus e entre Homo sapiens e Callithrix jacchus.* Diferente do estudo de Parker *et al* (2006) as seqüências analisadas no presente estudo foram previamente separadas por similaridade de bases. É sabido que seqüências de nucleotídeos diferentes podem ter estruturas semelhantes (Greenbaum *et al* 2007, Greenbaum *et al* 2007a). Neste sentido é possível que regiões com bases não conservadas, mas com topologia semelhante, tenham influenciado nos resultados.

7- Conclusão

Quando comparadas regiões conservadas e não conservadas entre *Homo sapiens e Mus musculus ou Rattus norvegicus e entre Homo sapiens e Callithrix jacchus*, não apresentaram diferenças estatisticamente significantes no padrão de clivagem química quando substituímos nucleotídeos na região promotora.

8- Referências Bibliográficas

Adamson E D. Growth Factors and Development. In: Developmental of epidermal growth factor receptor. New York: John Wiley; 1990: cap. 1, p. 1-30.

Bei M, Maas R. FGFs and BMP4 induce both Msx1- independent and Msx1-dependent signaling pathways in early tooth development. Development, 1998 Dec; v. 125 p. 4325-33.

Bergqvist, L P. The role of teeth in mammal history. Brazilian Journal of Oral Sciences, Piracicaba, v. 2, n.6, p. 249-257, July/ sep. 2003.

Berman JJ. Ruby Programming for Medicine and Biology. 1st edition. Jones and Bartllet Publishers. 2008.

Black DA. Ruby for Rails: Ruby techniques for Rails developers. 1st edition. Manning Publications. 2006.

Carpenter G, Cohen S. Epidermal growth factor. Ann Rev Biochem, Nashville, Tennessee, v. 48, p.193-216, aug. 1979.

Das P, Stockton D W, Bauer C, Shaffer L G, D´Souza R N, Writh J T, et al.

Elisangela R S, Peres R C R, Scarel R M, De Conto F, Line S R P. Absence of. Mutations in the promoter region of the LEF1 gene in Patients with hypodontia. Brazilian Journal of oral Sciences, Piracicaba, v. 2, n. 4, p. 144-146, jan/mar. 2003.

Haploinsufficiency of PAX9 is associated with autosomal dominant hypodontia. Hum. Genet Houston, v. 110, p. 371-6, sep. 2002.

Goto N, Nakao MC, Kawashima S, Katayama T, Kanehisa M. BioRuby: Open-Source Bioinformatics Library. Genome Informatics. 2003; 14: 629-630.

Greenbaum JA, Pang B, Tullius TD. Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res. 2007 Jun; 17(6):947-53.

Greenbaum JA, Parker SC, Tullius TD. Detection of DNA structural motifs in functional genomics elements. Genome Res. 2007(a) Jun; 17(6): 940-6.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, *et al.* Transcriptional regulatory code of a eukaryotic genome. Nature. 2004 Sep 2; 431(7004):99-104.

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

International Human Genome Sequencing Consortium. Nature. 2004; 431:931

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

Jernvall J, Kettunen P, Karavanova I, Martin L B, Thesleff I. Evidence for the role of the enamel knot as a control center in mammalian tooth cusp formation: Non-dividing cells express growth stimulating FGF-4 gene. Int J Dev Biol., Helsinki, Finland, v. 38, p. 463-9, Nov. 1994.

Jernvall J, Aber G T, Kettunen P, Keranen S, Thesleff I, Jernvall J, *et al*. The life History of an embryonic signaling center: BMP-4 induces p21 and is associated with apoptosis in the mouse tooth enamel knot. Development. Helsinki, Finland, v. 125, p. 161-9, may 1998.

Kellis M, Patterson N, Endrizzi M, Birren B & Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423, 241−254 (2003).

Kettunen P, Thesleff I. Expression and function of FGFs 4, 8 and 9 suggest functional redundancy and repetitive use as epithelial signals during tooth development. Dev. Dyn., Helsinki, Finland, v. 211, p. 256-68, sep. 1998.

Lodish H**,** Berk A**,** Zipursky SL**,** Matsudaira P**,** Baltimore D**,** Darnell J**.** (2000) Molecular Cell Biology. W.H.Freeman and Company, N.Y. (4ª Edição),

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

Maas R, Bei M. The genetic control of early tooth development. Crit. Rev. Oral Biol. Boston,
Massachusetts, v. 8, p. 4-39, oct. 1997.

Nakao M, Goto N, and Katayama, T. Bioinformatics recipes with BioRuby. GIW 2004 Poster Abstracts, S011 (2004). http://bioruby.org/

Nikolay A Kolchanov, Tatyana I Merkulova, Elena V Ignatieva, Elena A Ananko, Dmitry Yu Oshchepkov, *et al.* Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes. Briefings in Bioinformatics Advance Access published July 12, 2007 (on line).

Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. Local DNA topography correlates with functional no coding regions of the human genome. Science. 2009 Apr 17; 324(5925): 389-92.

Pressman RS. Engenharia de Software. 6ª edição. Editora McGraw-Hill. 2006.

Price MA and Tullius. 1992. Using hydroxyl radical to probe DNA structure. Methods Enzymol. 212: 194–219.

Prince JT, Marcotte EM. Mspire: mass spectrometry proteomics in Ruby. Bioinformatics. 2008 Dec 1; 24(23):2796-7. Epub 2008 Oct 16.

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

Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. (2009) the role of DNA shape in protein-DNA recognition. Nature. 461(7268):1248-53.

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

Shen S, Tuszynski JA. Theory and Mathematical Methods for Bioformatics. Springer. 2008

Thesleff I. Homeobox genes and growth factors in regulation of craniofacial and tooth morphogenesis. Acta Odontol Scand., Helsinki, Finland, v. 53, p. 129-34, Jan. 1995.

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.

Vellozo N C. 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
Piracicaba, SP: [s.n.], 2010. 47f. : il.

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

Wright BE, Reimers JM, Schmidt KH, and Reschke DK. (2002) hyper mutable bases in the p53 cancer gene are at vulnerable positions in DNA secondary structures. 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.

Lista de figuras

Anexo 1. Software para obtenção dos valores de diferença de predição de clivagem química.

```
a = 'timp2' #********COLOCAR O NOME DO GENE ENTRE "
    Conservação = '0,2' #*****MUDAR A CONSERVACAO QUE DESEJA 0,1 0U 0,2
******
    orchid2 = File.readlines ("#{a}-orchid.txt") # valores orchid do excel
    sequence = File.readlines("#{a}align.txt") #  seqüências alinhadas
    linhas = File.readlines("#{a}-Clustalw-es.txt") # arquivo com as seqüências
conservadas 0,1 e 0,2

    homo_s = []; pan_s = []; pongo_s = []; bos_s = []; macaca_s = []; canis_s = [];
gorilla_s = []; ancestral_s = []; callitrix_s = []; equus_s = []; mus_s = []; rattus_s = []

    homo_o = []; pan_o = []; pongo_o = []; bos_o = []; macaca_o = []; canis_o = [];
gorilla_o = []; ancestral_o = []; callitrix_o = []; equus_o = []; mus_o = []; rattus_o = []
#valores 0rchid

    homo_r = []; pan_r = []; pongo_r = []; bos_r = []; macaca_r = []; canis_r = [];
gorilla_r = []; ancestral_r = []; callitrix_r = []; equus_r = []; mus_r = []; rattus_r = [] #array
contendo valores orchid alinhados

    homo_con = []; pan_con = []; pongo_con = []; bos_con = []; macaca_con = [];
canis_con = []; gorilla_con = []; ancestral_con = []; callitrix_con = []; equus_con = [];
mus_con = []; rattus_con = [] # contem valores orchid das seq conservadas

    homo_seqcon = []; pan_seqcon = []; pongo_seqcon = []; bos_seqcon = [];
macaca_seqcon = []; canis_seqcon = []; gorilla_seqcon = []; ancestral_seqcon = [];
callitrix_seqcon = []; equus_seqcon = []; mus_seqcon = []; rattus_seqcon = [] # contem
valores orchid das seq conservadas
```

```
a0=[]; a1=[]; a2=[]; a3=[]; a4=[]
a = [a0, a1, a2, a3, a4]
sequencia = []
Orchid = []
orchid1 = []
teste ="
teste1=[]
#orchid2 = File.readlines ('acan-o.txt')
#p orchid2

orchid2.each{|i|
orchid1<< i.gsub("\n",",").gsub("\t",",")}
#orchid1.each{|i|
#orchid<< i.gsub("\t",",")}
orchid = orchid1.join.split(",")
#p orchid.size#ver como fica o array orchid

count = 0

while count < orchid.size
a0<<orchid[count]; count += 1
a1<<orchid[count]; count += 1
a2<<orchid[count]; count += 1
a3<<orchid[count]; count += 1
a4<<orchid[count]; count += 1



end
#p 'a'
```

```
#p a0
#p a1
#p a2
#p a3
#p a4
contador = 0
contador1 = 0

while contador < a.size #sao os arrays com os valores orchid
if a[contador][0][0,3] == 'hom'
                homo_o = a[contador].drop(1)
                contador += 1
                elsif a[contador][0][0,3] == 'pan'
                pan_o = a[contador].drop(1)
                contador += 1
                elsif a[contador][0][0,3] == 'pon'
                pongo_o = a[contador].drop(1)
                contador += 1
                elsif a[contador][0][0,3] == 'bos'
                bos_o = a[contador].drop(1)
                contador += 1
                elsif a[contador][0][0,3] == 'mac'
                macaca_o = a[contador].drop(1)
                contador += 1
                elsif a[contador][0][0,3] == 'can'
                canis_o = a[contador].drop(1)
                contador += 1
                elsif a[contador][0][0,3] == 'gor'
                gorilla_o = a[contador].drop(1)
                contador += 1
```

```
            elsif a[contador][0][0,3] == 'cal'
            callitrix_o = a[contador].drop(1)
            contador += 1
            elsif a[contador][0][0,3] == 'equ'
            equus_o = a[contador].drop(1)
            contador += 1
            elsif a[contador][0][0,3] == 'mus'
            mus_o = a[contador].drop(1)
            contador += 1
            elsif a[contador][0][0,3] == 'rat'
            rattus_o = a[contador].drop(1)
            contador += 1
end
end

#p homo_o
#p callitrix_o
#p mus_o
#p macaca_o
#p canis_o
#p rattus_o

#sequence = File.readlines('acan-s.txt')
sequence.each{|i|
sequencia << i.gsub("\n","")}
#puts sequencia

marcador = []
sequencia.each{|i|
```

```
if i[0,4] == '>hom'
                marcador = i[0,4]
end
if i[0,4] == '>pan'
                marcador = i[0,4]
end
if i[0,4] == '>pon'
                marcador = i[0,4]
end
if i[0,4] == '>bos'
                marcador = i[0,4]
end
if i[0,4] == '>mac'
                marcador = i[0,4]
end
if i[0,4] == '>can'
                marcador = i[0,4]
end
if i[0,4] == '>gor'
                marcador = i[0,4]
end
if i[0,4] == '>anc'
                marcador = i[0,4]
end
if i[0,4] == '>cal'
                marcador = i[0,4]
end
if i[0,4] == '>equ'
                marcador = i[0,4]
end
```

```
if i[0,4] == '>mus'
                marcador = i[0,4]
end
if i[0,4] == '>rat'
                marcador = i[0,4]
end
if marcador == '>hom'
                homo_s << i
end
if marcador == '>pan'
                pan_s << i
end
if marcador == '>pon'
                pongo_s << i
end
if marcador == '>bos'
                bos_s << i
end
if marcador == '>mac'
                macaca_s << i
end
if marcador == '>can'
                canis_s << i
end
if marcador == '>gor'
                gorilla_s << i
end
if marcador == '>cal'
                callitrix_s << i
end
```

```
if marcador == '>equ'
                equus_s << i
end
if marcador == '>mus'
                mus_s << i
end
if marcador == '>rat'
                rattus_s << i
end
}

rattus_s = rattus_s.drop(1).join #strings contendo sequencias alinhadas das especies
homo_s = homo_s.drop(1).join
pan_s = pan_s.drop(1).join
pongo_s = pongo_s.drop(1).join
bos_s = bos_s.drop(1).join
macaca_s = macaca_s.drop(1).join
canis_s = canis_s.drop(1).join
gorilla_s = gorilla_s.drop(1).join
callitrix_s = callitrix_s.drop(1).join
equus_s = equus_s.drop(1).join
mus_s = mus_s.drop(1).join
#p 'homos'
#p homo_s
#p pongo_s
#p equus_s

a = [homo_s, pan_s ,pongo_s ,bos_s, macaca_s, canis_s, gorilla_s, callitrix_s,
equus_s, mus_s, rattus_s]
```

```
b = [homo_o, pan_o ,pongo_o ,bos_o, macaca_o, canis_o, gorilla_o, callitrix_o,
equus_o, mus_o, rattus_o]
c = [homo_r, pan_r ,pongo_r ,bos_r, macaca_r, canis_r, gorilla_r, callitrix_r, equus_r,
mus_r, rattus_r]

count = 0 #contador especie
contador = 0 #contador sequencia dna com -
contador_o = 0 #contador do orchid

while count < a.size
        if
                a[count] != [] or b[count] != []
                contador  = 0
                        while contador <
        a[count].size #contador da string com '

        if a[count][contador] [/[A-Za-z]/]
                        c[count] << b[count][contador_o]
                        contador += 1
                        contador_o += 1
                                elsif
a[count][contador] == '-'
                        c[count] << '-'
                        contador += 1
                                else
                                c[count] << "N"
                        contador += 1
                                end
                        end
                        count += 1
```

```
                                contador  = 0
        else
                                                count += 1
                                end
                                contador_o = 0
end
#x = 2 #colocar posição da sequencia
#y = 19

c.each{|i|
if i[0] == nil
        i = nil
end}

#p "homo_r = #{homo_r.join(' ,')}"

linhas1 =[]
naocon= [0, homo_s.size - 1] #array contendo as posicoes nao conservadas
conserv = []
linhas.each{|i|
linhas1 << i.gsub("\n","").gsub("\t","")}
marcador = ''
linhas1.each{|i|
if i.split[-1] == '0,1'
        marcador  = '0,1'
end
if i.split[-1] == '0,2'
        marcador  = '0,2'
end
if i[0..3] == '>hom' and conservacao == marcador
```

```
                naocon << i.split[1].to_i
                naocon << i.split[3].to_i
 end
  if conservacao == marcador
                conserv << i
 end
}
naocon.sort!
#p 'conserv'
#puts conserv
#p 'naocon'
#puts naocon

count = 0 #contador da linha
contador = 0 #contador da posicao da base na linha

marcador = []
conserv.each{|i|
if i[0,4] == '>hom'
                marcador = i[0,4]
                homo_con << homo_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>pan'
                marcador = i[0,4]
                pan_con << pan_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>pon'
                marcador = i[0,4]
                pongo_con << pongo_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
```

```
if i[0,4] == '>bos'
                marcador = i[0,4]
                bos_con << bos_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>mac'
                marcador = i[0,4]
                macaca_con << macaca_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>can'
                marcador = i[0,4]
                canis_con << canis_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>gor'
                marcador = i[0,4]
                gorilla_con << gorilla_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>anc'
                marcador = i[0,4]
                ancestral_con << ancestral_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>cal'
                marcador = i[0,4]
                callitrix_con << callitrix_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>equ'
                marcador = i[0,4]
                equus_con << equus_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>mus'
                marcador = i[0,4]
```

```
            mus_con << mus_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end
if i[0,4] == '>rat'
            marcador = i[0,4]
            rattus_con << rattus_r[i.split[1].to_i - 1..i.split[3].to_i - 1]
end



if marcador == '>hom' and i[0,4] != '>hom'
            homo_seqcon << i.strip# array c as seqs conservadas
end
if marcador == '>pan' and i[0,4] != '>pan'
            pan_seqcon << i.strip
end
if marcador == '>pon' and i[0,4] != '>pon'
            pongo_seqcon << i.strip
end
if marcador == '>bos' and i[0,4] != '>bos'
            bos_seqcon << i.strip
end
if marcador == '>mac' and i[0,4] != '>mac'
            macaca_seqcon << i.strip
end
if marcador == '>can' and i[0,4] != '>can'
            canis_seqcon << i.strip
end
if marcador == '>gor' and i[0,4] != '>gor'
            gorilla_seqcon << i.strip
end
if marcador == '>cal' and i[0,4] != '>cal'
```

```
                callitrix_seqcon << i.strip
end
if marcador == '>equ' and i[0,4] != '>equ'
                equus_seqcon << i.strip
end
if marcador == '>mus' and i[0,4] != '>mus'
                mus_seqcon << i.strip
end
if marcador == '>rat' and i[0,4] != '>rat'
                rattus_seqcon << i.strip
end
if marcador == 'anc' and i[0,4] != 'anc'
                ancestral_seqcon << i.strip
                end
}
p 'SEQUENCIAS CONSERVADAS homo, rat, mus, call, mac, pongo'
#p homo_con#.size
p homo_seqcon
#p rattus_con
p rattus_seqcon
#p mus_con
p mus_seqcon
#p callitrix_con
p callitrix_seqcon
p macaca_seqcon
p pongo_seqcon


array_con_mus = []
array_con_rattus = []
```

```
array_con_callitrix = []
array_con_pongo = []
array_con_macaca = []

array_mus = []
array_rattus = []
array_callitrix = []
array_pongo = []
array_macaca = []


contador = 0

while contador <  homo_seqcon.size #while 1
count = 0
while count < homo_seqcon[contador].size and mus_o != [] #while 2
        if homo_seqcon[contador][count] != mus_seqcon[contador][count]
                array_con_mus<< ((mus_con[contador][count].to_f) -
(homo_con[contador][count].to_f)).abs
                count += 1
        else count += 1
        end
end #end 2
count = 0
while count < homo_seqcon[contador].size and rattus_o != [] #while 2
        if homo_seqcon[contador][count] != rattus_seqcon[contador][count]
                array_con_rattus<< ((rattus_con[contador][count].to_f)
- (homo_con[contador][count].to_f)).abs
                count += 1
        else count += 1
```

```
                end
        end
        count = 0
        while count < homo_seqcon[contador].size and callitrix_o != [] #while 2
                if homo_seqcon[contador][count] != callitrix_seqcon[contador][count]
                        array_con_callitrix<<
((callitrix_con[contador][count].to_f) - (homo_con[contador][count].to_f)).abs
                        count += 1
                else count += 1
                end
        end
        count = 0
        while count < homo_seqcon[contador].size and pongo_o != [] #while 2
                if homo_seqcon[contador][count] != pongo_seqcon[contador][count]
                        array_con_pongo<<
((pongo_con[contador][count].to_f) - (homo_con[contador][count].to_f)).abs
                        count += 1
                else count += 1
                end
        end
        count = 0
        while count < homo_seqcon[contador].size and macaca_o != [] #while 2
                if homo_seqcon[contador][count] != macaca_seqcon[contador][count]
                        array_con_macaca<<
((macaca_con[contador][count].to_f) - (homo_con[contador][count].to_f)).abs
                        count += 1
                else count += 1
                end
        end
        contador += 1
```

```
end #end 1


File.open('resultado_con_mus.txt', 'w')            do |f|
                if array_con_mus.empty? == false
                 f.puts array_con_mus
                end
end
File.open('resultado_con_rattus.txt', 'w')         do |f|
                if array_con_rattus.empty? == false
                 f.puts array_con_rattus
                end
end
File.open('resultado_con_callitrx.txt', 'w')       do |f|
                if array_con_callitrix.empty? == false
                 f.puts array_con_callitrix
                end
end
File.open('resultado_con_pongo.txt', 'w')          do |f|
                if array_con_pongo.empty? == false
                 f.puts array_con_pongo
                end
end
File.open('resultado_con_macaca.txt', 'w')         do |f|
                if array_con_macaca.empty? == false
                 f.puts array_con_macaca
                end
end
puts "resultado conservado mus = #{array_con_mus}"
puts "resultado conservado rat = #{array_con_rattus}"
puts "resultado conservado calitrix = #{array_con_callitrix}"
```

```
puts "resultado conservado maca = #{array_con_macaca}"
puts "resultado conservado pon = #{array_con_pongo}"

homo_si = []; pan_si = []; pongo_si = []; bos_si = []; macaca_si = []; canis_si = []; gorilla_si = []; callitrix_si = []; equus_si = []; mus_si = []; rattus_si = []
#p homo_s.join

homo_ri = []; pan_ri = []; pongo_ri = []; bos_ri = []; macaca_ri = []; canis_ri = []; gorilla_ri = []; callitrix_ri = []; equus_ri = []; mus_ri = []; rattus_ri = []

count = 0
#p 'homo-r'
#p homo_r
while count < naocon.size - 1

homo_si += homo_s.split('').values_at(naocon[count]..naocon[count + 1])
pan_si += pan_s.split('').values_at(naocon[count]..naocon[count + 1])
pongo_si += pongo_s.split('').values_at(naocon[count]..naocon[count + 1])
bos_si += bos_s.split('').values_at(naocon[count]..naocon[count + 1])
macaca_si += macaca_s.split('').values_at(naocon[count]..naocon[count + 1])
canis_si += canis_s.split('').values_at(naocon[count]..naocon[count + 1])
gorilla_si += gorilla_s.split('').values_at(naocon[count]..naocon[count + 1])
callitrix_si += callitrix_s.split('').values_at(naocon[count]..naocon[count + 1])
equus_si += equus_s.split('').values_at(naocon[count]..naocon[count + 1])
mus_si += mus_s.split('').values_at(naocon[count]..naocon[count + 1])
rattus_si += rattus_s.split('').values_at(naocon[count]..naocon[count + 1])

homo_ri += homo_r.values_at(naocon[count]..naocon[count + 1])
pan_ri += pan_r.values_at(naocon[count]..naocon[count + 1])
pongo_ri += pongo_r.values_at(naocon[count]..naocon[count + 1])
```

```
	bos_ri += bos_r.values_at(naocon[count]..naocon[count + 1])
	macaca_ri += macaca_r.values_at(naocon[count]..naocon[count + 1])
	canis_ri += canis_r.values_at(naocon[count]..naocon[count + 1])
	gorilla_ri += gorilla_r.values_at(naocon[count]..naocon[count + 1])
	callitrix_ri += callitrix_r.values_at(naocon[count]..naocon[count + 1])
	equus_ri += equus_r.values_at(naocon[count]..naocon[count + 1])
	mus_ri += mus_r.values_at(naocon[count]..naocon[count + 1])
	rattus_ri += rattus_r.values_at(naocon[count]..naocon[count + 1])
	count += 2
	end
	#p homo_si
	#p mus_si
	#p 'homo_ri'
	#p homo_ri
	#p mus_ri
	#p mus_r.size
	count = 0
	while count < homo_si.size and mus_o.empty? == false
		if mus_si[count] != homo_si[count] and homo_si[count] != '-' and
mus_si[count] != '-'
			array_mus << ((mus_ri[count].to_f) -
(homo_ri[count].to_f)).abs
			count += 1
		else count += 1
		end
	end

	count = 0
	while count < homo_si.size and rattus_o.empty? == false
```

```
                if rattus_si[count] != homo_si[count] and homo_si[count] != '-' and
rattus_si[count] != '-'
                        array_rattus << ((rattus_ri[count].to_f) -
(homo_ri[count].to_f)).abs
                        count += 1
                else count += 1
                end
        end
        count = 0
        while count < homo_si.size and callitrix_o.empty? == false
                if callitrix_si[count] != homo_si[count] and homo_si[count] != '-' and
callitrix_si[count] != '-'
                        array_callitrix << ((callitrix_ri[count].to_f) -
(homo_ri[count].to_f)).abs
                        count += 1
                else count += 1
                end
        end
        count = 0
        while count < homo_si.size and macaca_o.empty? == false
                if macaca_si[count] != homo_si[count] and homo_si[count] != '-' and
macaca_si[count] != '-'
                        array_macaca << ((macaca_ri[count].to_f) -
(homo_ri[count].to_f)).abs
                        count += 1
                else count += 1
                end
        end

        count = 0
```

```
	while count < homo_si.size and pongo_o.empty? == false
			if pongo_si[count] != homo_si[count] and homo_si[count] != '-' and
pongo_si[count] != '-'
				array_pongo << ((pongo_ri[count].to_f) -
(homo_ri[count].to_f)).abs
				count += 1
			else count += 1
			end
	end

	File.open('result_nao_cons_mus.txt', 'w')	do |f|
			if array_mus.empty? == false
				f.puts array_mus

			end
	end
	File.open('result_nao_cons_rat.txt', 'w')	do |f|
			if array_rattus.empty? == false
				f.puts array_rattus

			end
	end
	File.open('result_nao_cons_call.txt', 'w')	do |f|
			if array_callitrix.empty? == false
				f.puts array_callitrix

			end
	end
	File.open('result_nao_cons_maca.txt', 'w')	do |f|
			if array_macaca.empty? == false
```

```
                                f.puts array_macaca

                end
        end

        File.open('result_nao_cons_pon.txt', 'w')          do |f|
                if array_pongo.empty? == false
                                f.puts array_pongo

                end
        end

        puts "res nao cons mus = #{array_mus}"
        puts "res nao cons rat = #{array_rattus}"
        puts "res nao cons call = #{array_callitrix}"
        puts "res nao cons maca = #{array_macaca}"
        puts "res nao cons pongo = #{array_pongo}"
```

Anexo 2. Exemplo de arquivo contendo as seqüências alinhadas em formato fasta (inicia com >).

>*Homo sapiens*

-CCGACAGTAGCGGGCTGCACCCTCCTTAGATCGCGTCGTGGCCAGCCTC

AGCTGCAGAACCCC-GCCGGGCGCCCGGGAG---CCCTG--TCCC--GCC

GCGGC---GGCC-----T----------CAGAGTCCGG-------GC

ACTTGGGGA------TTCTCTGAGGGTGCAGCCCTCCTGGGCCCCTCCGA

CTGCGGAGCTGCCGACCGCAATGCAGATGCGGGCCCTCCAGCCCTTCTC

G

CGCCGCCTTCCTCCCGCCAGCCCCGGAGCTGCCGGGTCCGCGCCGCCCC
G

GGGAGCGCTCCTCTCCCGCCC-TGAGCGCAGGCCGGCTTCCCCATCGGCG

CGCCGGTCCGGAGCCAGGGTCCAGCGCGCTCCAGACGCCTCTGCCTTCC
C

CTCCCC---CTGTCGGCCGCCCCTCGG--TCCCTGGGGGTGGGGTTTCCC

TTTGCGCTCGCCCCCTCCCGCCCCCA-CCCCTCACGGGCCCTCCCCTCCC

CCGCCCGTCCCTATGTATGTGTCACAGCGCGCCATGCCCGCCCGCCCGCC

C----

>*Pongo pygmaeus*

-CCGACAGTAGGGGGCTGCACCCTCCTTAGATCGCGTCGTGCCCAGCTTC

AGCTGCAGACCCCC-GCCGGGCGCCCGGGAG---CCCTG--TCAG--GCC

GCGGC---GGCC-----T----------CAGAGTCCGG-------GC

ACTCGGGGA------TACTCTGAGGGTGCAGCCCTCCCGGGCCCCTCCGA

CTGCGGAGCTGCCGACCGCAATGCAGACGCGGGCCCTCCAGCCCTTCTC
G

CGCCGCTTTCCTCCCGCCGGCCTCGGAGCTGCCGGGTCCGCGCCGCCGC
G

GGGAGCGCTCCTCTCGCGCCC-TGAGCGCAGGGCGGCTTCCCCATCGGCG

CGCCGGTCCGGAGCCAGGGTCCAGCGCGCTCCAGACGCCTCTGCCTTCC
C

CTCCCC---CTGTCGGCCGCCCCTCGG--TCCCCGGGGGTGGGGTTTCCC

TTTGCGCTCGCCCCCTCCCGCCCCCAACCCCTCGCGGGCCCTCCCCTCCC

CCGCCCGTCCCTATGTATGTGTCACAGCGCGCCATGCCCGCCCGCCCGCC

C----

>*Callithrix jacchus*

-CCGAGAGTAGGGGGCCGCACCCTCCCTTGATCGCGTCGTGCCCAGCCTC

AGCTGCAGACCCCA-GCCCAGCGCCCGGGAG---CCCCG--TCAG--GCC

GCGGC---GGCC-----C----------CAGAGCCCGG-------GC

GCTCAGGAT------TGCTCTGAAGGTGCAGTCCTCC-GGGCCCCTCCGA

CTGCGGAGCTGCCGACCGCAATGCAGACGCCGGCCCTCCAGTCCTCCTC

G

CGCCACCTTCCTCCCGCCGGCCCCGGAGC-GCCGGGTCCGCGCCGCCGCG

GGGAGCGTTCCTCTCACGCCC-TGAGCGCAGG-CGGCTTCCCCTTCCGCG

CGCCTATCCGGAGCCAGGGTTCAGCGCGCTCCAGACGCCTCTGCCTCCC

C

CTCCCC---CTGCCGGCCGCCCCTCGG--TCCCCGGGGGTGGGGTTTCCC

TTCGCGCTCGCCCCTCCTGCCCCCCCACCCCTCACGG--CCCCCCCTCCC

CTGCCCGTCCCTATGTATGTGTCACCGCGCGCCATGCCCGCCCGCCCGCC

C----

>*Mus musculus*

-CCCAGAGTAGGGGGTCGTCCTGTGC---GATGGCATCTTTCTGGGTATC

AGCTGCAAATCGCA-TCCCGATACCAGGGAAACACCCCGACCCCA--ACC

CCAGCACCGGAC-----TGCTACCCTGACAGCAGGAGCCGG-------GT

GCCTTCGG-------CACTCTGCGGGCGCAGCCCTCCGGGGCCGCCTGGA

CTGCGGAGCTGCCGACCGCAATGCAGACGCGGGCCCTCCAGTGCTGCCC

G

CACAGCTTTCCTCC-GCGGCCCCAGGAGCTGCGGGGTCCGCGC-------

--------TCCTCTCGTGCCC-TGCGCGCCCGGAGCCTTCCCCAGCTGAG

CGC-GGTCCCCAGCCCGGGTCCTGCGCGCTCCGGACGTTTCTGCCTTCCC

CTCCCC---CCGCAGA-------TTGG---CCCCGGGGGTGGGGTTTCCC

TGTGCGCTCGCCCCCACC---------CCTCGTGTGTGCCCTCCCCTCCC

CCGCCCGCCCCTATGTATGTGTCACCGCGCACCATTCCCGCCC-------

>*Canis familiaris*

CCCGGGAGCAGCGGG-CGCGCCCCGCC-GGGTGGCTTC-CGCGGGGCGTC

CGCTGCACTCCCCGCGCCGGGCGCCCGGGAG---CCGCGG-CCGGTCACC

GCGGGCGCGGGTGCGGGTGCGGGTGCGGGCGCGGGCGCCGGTGCCGGT

GC

GGGTGCGGGCCCCCGCACCCGGGCGACGCAGCCCTCCAGGGCCCCTGGG

A

CTGCGGAGCCGCCGACCGCATTGCAG----------CCCTTCCCG

CGCCGC-TGCCTCCCGCGGGCCCGGG----------CCGCGT-------

------GCTCCTCTCGCGCCCCTGCGC-CCCGGCGGCTTCCCC-GCGGCG

CGCCGGTCCGGAGCCGGGGTCCAGCGCTCTCCAGACGCCTCCGCCCTCC

C

CGCCCCTCCCCCCCGGCCGCCCCTCGGGTCTCCCGGGGGTGGGGTTTCCC

TCGGCGCCGGCC----------TCCCCTCCC

CCGCCCGCCCCTATGTATGTGTCACCGCGCGCCATGCCCGCCCGCCCGCC

CGCCC

Anexo 3. Exemplo de arquivo contendo as seqüências conservadas em formato fasta com a entropia colocada na primeira linha.

>*Homo sapiens*: 493 to 543 0,1

CCCCTCCCCCGCCCGTCCCTATGTATGTGTCACAGCGCGCCATGCCCGCC
C

>*Pongo pygmaeus*: 493 to 543

CCCCTCCCCCGCCCGTCCCTATGTATGTGTCACAGCGCGCCATGCCCGCC
C

>*Callithrix_jacchus*: 493 to 543

CCCCTCCCCTGCCCGTCCCTATGTATGTGTCACCGCGCGCCATGCCCGCC
C

>*Mus musculus*: 493 to 543

CCCCTCCCCCGCCCGCCCCTATGTATGTGTCACCGCGCACCATTCCCGCC
C

>*Canis familiaris*: 493 to 543

CCCCTCCCCCGCCCGCCCCTATGTATGTGTCACCGCGCGCCATGCCCGCC
C

>*Homo sapiens*: 435 to 452

GGGGGTGGGGTTTCCCTT

>*Pongo pygmaeus*: 435 to 452

GGGGGTGGGGTTTCCCTT

>*Callithrix jacchus*: 435 to 452

GGGGGTGGGGTTTCCCTT

>*Mus musculus*: 435 to 452

GGGGGTGGGGTTTCCCTG

>*Canis familiaris*: 435 to 452

GGGGGTGGGGTTTCCCTC

>*Homo sapiens*: 199 to 226

GACTGCGGAGCTGCCGACCGCAATGCAG

>*Pongo pygmaeus*: 199 to 226

GACTGCGGAGCTGCCGACCGCAATGCAG

>*Callithrix jacchus:* 199 to 226

GACTGCGGAGCTGCCGACCGCAATGCAG

>*Mus musculus*: 199 to 226

GACTGCGGAGCTGCCGACCGCAATGCAG

>*Canis familiaris*: 199 to 226

GACTGCGGAGCCGCCGACCGCATTGCAG

>*Homo sapiens*: 493 to 543 0, 2

CCCCTCCCCCGCCCGTCCCTATGTATGTGTCACAGCGCGCCATGCCCGCC
C

>*Pongo pygmaeus*: 493 to 543

CCCCTCCCCCGCCCGTCCCTATGTATGTGTCACAGCGCGCCATGCCCGCC
C

>*Callithrix jacchus*: 493 to 543

CCCCTCCCCTGCCCGTCCCTATGTATGTGTCACCGCGCGCCATGCCCGCC
C

>*Mus musculus*: 493 to 543

CCCCTCCCCCGCCCGCCCCTATGTATGTGTCACCGCGCACCATTCCCGCC
C

>*Canis familiaris*: 493 to 543

CCCCTCCCCCGCCCGCCCCTATGTATGTGTCACCGCGCGCCATGCCCGCC
C

>*Homo sapiens*: 432 to 462

CCTGGGGGTGGGGTTTCCCTTTGCGCTCGCC

>*Pongo pygmaeus*: 432 to 462

CCCGGGGGTGGGGTTTCCCTTTGCGCTCGCC

>*Callithrix jacchus*: 432 to 462

CCCGGGGGTGGGGTTTCCCTTCGCGCTCGCC

>*Mus musculus*: 432 to 462

CCCGGGGGTGGGGTTTCCCTGTGCGCTCGCC

>*Canis familiaris*: 432 to 462

CCCGGGGGTGGGGTTTCCCTCGGCGCCGGCC

>*Homo sapiens*: 362 to 406

AGCCAGGGTCCAGCGCGCTCCAGACGCCTCTGCCTTCCCCTCCCC

>*Pongo pygmaeus*: 362 to 406

AGCCAGGGTCCAGCGCGCTCCAGACGCCTCTGCCTTCCCCTCCCC

>*Callithrix jacchus*: 362 to 406

AGCCAGGGTTCAGCGCGCTCCAGACGCCTCTGCCTCCCCCTCCCC

>*Mus musculus*: 362 to 406

AGCCCGGGTCCTGCGCGCTCCGGACGTTTCTGCCTTCCCCTCCCC

>*Canis familiaris*: 362 to 406

AGCCGGGGTCCAGCGCTCTCCAGACGCCTCCGCCCTCCCCGCCCC

>Homo sapiens: 199 to 226

GACTGCGGAGCTGCCGACCGCAATGCAG

>*Pongo pygmaeus*: 199 to 226

GACTGCGGAGCTGCCGACCGCAATGCAG

>*Callithrix jacchus*: 199 to 226

GACTGCGGAGCTGCCGACCGCAATGCAG

>*Mus musculus:* 199 to 226

GACTGCGGAGCTGCCGACCGCAATGCAG

>*Canis familiaris:* 199 to 226

GACTGCGGAGCCGCCGACCGCATTGCAG

Anexo 4. Exemplo de arquivo contendo os valores obtidos no site Orchid para cada espécie analisada. Devido à grande extensão do arquivo apenas alguns valores foram colocados. Os valores para cada espécie são colocados em colunas abaixo do respectivo nome.

Homo Pongo Callithrix Musmus Canis

0.63	0.63	0.63	-0.24	0.43
0.30	0.30	0.35	-0.03	0.08
0.60	0.60	0.78	0.45	0.01
0.29	0.29	0.26	-0.26	0.55
0.89	0.89	0.97	1.23	0.93

-0.40	-0.40	0.04	0.10	0.90
1.28	1.28	0.94	0.94	-0.08
0.76	0.77	0.63	0.63	0.94
-0.36	-0.46	-0.46	-0.46	0.81
0.92	0.60	0.60	0.60	-0.39
0.61	0.60	0.60	0.60	1.20
0.19	0.58	0.58	0.53	0.64