Análise da via de regulação gênica por ácido retinóico: uma abordagem por bioinformática e biologia estrutural

Tiago José Paschoal Sobreira

Tese apresentada ao Programa de Pós-Graduação Interunidades em Bioinformática, Universidade de São Paulo, para a obtenção de título de Doutor em Bioinformática.

Orientadores:

Prof. Dr. Paulo Sergio Lopes de Oliveira

Prof. Dr. Carlos Alberto de Bragança Pereira

São Paulo

2008

Aos meus pais

A meu pai por ensinar a trabalhar e
sempre fazer o mais bem feito possível.

A minha mãe por ensinar a viver e
sempre viver da melhor maneira possível.

AGRADECIMENTOS

Ao Dr. Paulo Sérgio Lopes de Oliveira pela orientação, formação profissional e amizade.

Ao Prof. Dr. Carlos Alberto de Bragança Pereira pela co-orientação.

Ao Dr. José Xavier-Neto pela constante ajuda e grande contribuição nesse trabalho.

Ao Prof. Dr. José Eduardo Krieger por permitir o desenvolvimento desta tese no Laboratório de Genética e Cardiologia Molecular do Instituto do Coração, Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo.

Ao Prof. Dr. Arthur Gruber pela formação profissional e amizade.

Ao Prof. Dr. Fernando Portella de Luna Marques por disponibilizar o cluster de computadores do Laboratório de Sistemática Molecular do Instituto de Biociências da Universidade de São Paulo (IB-USP).

Ao Programa de Pós-Graduação Interunidades em Bioinformática pela oportunidade de realização do curso de doutorado.

À Sra. Patrícia, secretária do PPG em Bioinformática, pelo excelente apoio.

À Coordenadoria (CAPES) pela concessão da bolsa de doutorado.

As funcionárias do Laboratório de cardiologia molecular Isabel Cristina e Silvana pelas constantes ajudas.

Aos meus irmãos e cunhadas pela eterna amizade.

Aos meus sobrinhos por alegrar o meu caminho.

Ao Fabio Passetti pela grande ajuda no desenvolvimento deste trabalho e pela amizade e companheirismo.

Ao Marcio Almeida pelos excelentes almoços e pelas grandes discussões cientificas e não cientificas.

Ao Tiago Pereira pelos excelentes lanches da tarde com soluções de várias dúvidas cientificas e não cientificas e pelas contribuições nesse trabalho.

A Manuela Gimenes pelas contribuições nesse trabalho e pela grande amizade.

A Daniela Castro pela amizade, carinho, constante companheirismo e por ter aparecido na minha vida!!!

A Elizabeth Gorgone pela amizade, carinho, constante companheirismo.

A Daniel Rabanea pela grande amizade, constante companheirismo.

Aos grandes amigos Daneil Carneiro, Célia Boyadjian, Almir Picanço, Carolina Gois, Cristiane Minussi, Danilo Coelho, Juliana Velasco, Ernesto Nakayasu, Leandro Tambosi, Guilherme Ide, Therry Faria Lima, Mariana Farias, Priscila Rovina e Joana DeCapitani companheiros constantes e queridos da vida.

A Camila Holpert, Denise Gonçalves, Karina Noçais e Renata Estender pela eterna amizade.

RESUMO

Tese (Doutorado) – Programa de Pós-Graduação Interunidades em Bioinformática, Universidade de São Paulo, São Paulo, 2008.

As vias de sinalização celular por meio de moléculas são um dos principais meios de controle funcional de um organismo. O entendimento das funções de moléculas sinalizadoras facilita a compreensão das vias metabólicas de um organismo, assim possibilitando uma melhor compreensão de vários eventos biológicos e também de várias doenças. A sinalização pelo ácido retinóico (AR), e seus derivados, é responsável pelo controle de várias funções, por exemplo: crescimento celular, diferenciação celular, formação da retina, desenvolvimento cardíaco e também relacionado a várias patologias como diabetes, obesidades, cânceres, e doenças cardiovasculares.

A ação do ácido retinóico é controlada em dois níveis: no metabolismo de síntese/degradação e na sua utilização na sinalização para a expressão gênica.

A maquinaria que controla o metabolismo inclui as enzimas de síntese do AR (aldeído desidrogenase ALDH) e as enzimas de degradação do AR (Cyp26), que controlam a distribuição espaço-temporal do AR durante a embriogênese. As ALDHs são enzimas $NAD(P)^+$ dependentes, que oxidam uma ampla gama de aldeídos para os seus correspondentes ácidos carboxílicos, sendo ALDH1A2 a principal enzima na transformação de retinal em ácido retinóico.

A maquinaria da sinalização celular por AR contém os receptores nucleares controlados por AR (RARs) que estão envolvidos com o controle da transcrição gênica. Os mecanismos de controle de expressão mais comuns são os que ocorrem na fase transcricional. Um desses mecanismos envolve proteínas que se ligam às regiões promotoras de transcrição, representadas por trechos de DNA que geralmente estão localizados próximo à região de início da transcrição, mas que também podem estar a centenas ou até milhares de pares de bases desse início. Essas proteínas modulam a maquinaria transcricional, podendo ativá-la ou inibi-la.

A associação de várias técnicas como a biologia molecular, bioinformática, filogenia, análises estruturais de biomoléculas, mecânica molecular e métodos termodinâmicos tem se mostrado uma poderosa abordagem para compreensão de sistemas biológicos simplificando e agilizando o desenvolvimento do conhecimento científico. Nessa direção, esse estudo desenvolveu duas análises: a primeira estudando a evolução das funções das enzimas ALDH, utilizando-se de técnicas de genômica combinatória, filogenia, bioinformática, estrutura de biomoléculas e de biologia do desenvolvimento, tentando compreender o modo como as ALDHs, que apresentam as seqüências de aminoácidos bastante similares, puderam divergir para gerar funções diversas como a destoxificação e a sinalização. Para este estudo foram analisados os genomas de 487 organismos em busca de seqüências de ALDHs e também o genoma do organismo modelo *Branchiostoma floridae*. Foram obtidas 190 seqüências que foram utilizadas em uma análise filogenética para tentar compreender a função primordial e também para definir grupos de aminoácidos candidatos a marcadores das diferentes famílias de ALDHs. Essas 190 seqüências também foram modeladas estruturalmente e analisada a forma e o volume do canal onde se aloja o aldeído a ser oxidado. A partir dessas informações foi possível prever que as ALDHs passaram das funções ancestrais de controle do padrão corporal para algo mais abrangente como funções protetoras.

A segunda análise, utilizando-se das estruturas tridimensionais dos fatores de transcrição ligados ao DNA em diferentes posições e submetendo esses complexos a processos de mecânica molecular, cálculos termodinâmicos e análises das ligações de hidrogênio para tentar prever os mais prováveis sítios de interação entre os receptores e o DNA. O modelo escolhido para essa análise foram os fatores de transcrição regulados por ácido retinóico o RAR e RXR utilizando a região promotora do gene RARE-β2 para avaliar as mais prováveis regiões de ligação desses fatores. Para esse estudo foram construídos 71 complexos proteína-DNA que foram submetidos a processos de mecânica molecular e cálculos termodinâmicos. A partir dessas informações foi possível prever uma região de maior afinidade entre o fator de transcrição e o DNA. As análises de ligações de hidrogênio possibilitaram definir exatamente a região de interação entre os fatores de transcrição e o DNA, e também descrever as interações moleculares responsáveis pela especificidade da interação.

Palavras-chave: Bioinformática. Sinalização celular. Modelagem molecular por homologia. Fatores de transcrição. Mecânica molecular. Ácido retinóico.

ABSTRACT

Thesis (Doctoral) – Programa de Pós-Graduação Interunidades em Bioinformática, Universidade de São Paulo, São Paulo, 2008.

Cellular signaling paths through molecules are one of the main processes of functional control of an organism. The comprehension of signaling molecules functions enables one to understand the metabolic pathways of an organism, along with related biological events and several diseases. The signaling through retinoic acid (RA) and its secondary products is responsible for controlling several functions, such as cellular growth and differentiation, retina's formation and cardio development, and is also related to several pathologies such as diabetes, obesity, cancers and cardiovascular disorders.

There are two levels of control of retinoic acid activity: synthesis/degradation metabolism and its use in gene expression signaling.

The machinery that controls the metabolism includes RA's synthesis (aldehyde dehydrogenase ALDH) and degradation (Cyp26) enzymes, which control the space-temporal distribution of RA during the embryogenesis. The ALDHs are $NAD(P)^+$ dependent enzymes that oxidize many types of aldehydes into the related carboxylic acids, being the ALDH1A2 the main enzyme involved in the process of transformation of retinal into retinoic acid.

The machinery of cellular signaling through RA contains the nuclear receptors controlled by RA (RARs) that are involved in the control of gene transcription. The most common mechanisms of expression control are the ones that occur during the transcriptional phase. One of these mechanisms involves proteins that bind to the transcription promoter regions, represented by DNA sequences that are usually located close to the region where the transcription starts, but can also be hundreds or thousands of base pairs apart from the starting point. These proteins modulate the transcriptional machinery, being responsible for both its activation and inhibition.

The association of several techniques as molecular biology, bioinformatics, phylogeny, structural analysis of biomolecules, molecular mechanics and thermodynamic methods has been shown as a powerful tool for the understanding of biological systems, simplifying and speeding up the production of related scientific knowledge. Facing this direction, the present study developed two analyses. The first one studied the evolution of ALDH enzymes functions, using the techniques of combinatory genomic, phylogeny, bioinformatics, structure of biomolecules and developmental biology, in the attempt of understanding how the ALDHs could diverge and acquire different functions as detoxification and signaling, despite the fact that they have very similar aminoacid sequences. For this study, ALDHs sequences were searched for in the genome of 487 organisms plus the model organism's, *Branchiostoma floridae.* All 190 sequences obtained were used in a phylogenetic analysis, in the attempt of understanding the primordial function of the enzyme and defining possible groups of conserved aminoacids in the different families of ADLHs. These 190 sequences were also structurally modeled and the shape and volume of the channel where the aldehyde is placed to be oxidized were analyzed. Based on this information, it became possible to predict that the ALDHs moved from ancestral functions of corporal pattern control to a wider spectrum of protection functions.

For the second analysis we submitted the complex formed by tridimensional structures of the transcriptional factors bond to DNA in different positions to processes of molecular mechanics, thermodynamic calculi and analysis of the hydrogen bonds, in order to predict the most probable sites of interaction between the receptors and the DNA. The model chosen for this analysis were the transcription factors regulated by retinoic acid, RAR and RXR, using the promoter region of the gene RARE-β2 to assay the most probable binding regions of these factors. For this study, 71 protein-DNA complexes were built and submitted to processes of molecular mechanics and thermodynamic calculi. Based on the resulting data, it became possible to predict a region of greater affinity between the transcription factor and the DNA. The analyses of hydrogen bonds enabled us to define the exact region where the interaction between the transcription factor and the DNA takes place and also enabled us to describe the molecular interactions responsible for the specificity of this interaction.

Key-words: Bioinformatics. Cellular signaling. Molecular homology modeling. Transcription factors. Molecular mechanics. Retinoic acid.

LISTA DE ABREVIATURAS

3D	Tridimensional
Å	Ângstron
ALDH	do inglês *Aldehyde Dehydrogenase*
AR	Ácido retinóico
cDNA	DNA complementar
DCCs	Doenças Cardíacas Congênitas
DNA	Ácido desoxiribonucleico
EST	do Inglês *Expressed Sequence Tag*
mRNA	RNA mensageiro
ns	Nanossegundo
ps	Picossegundo
RAR	*Retinoic acid receptor*
RARE β2	*Retinoic Acid Response Elements* β2
RMN	Ressonância magnética nuclear
RMSD	do inglês *Root Mean Square Deviation*
RNA	Ácido ribonucléico
RXR	*Retinoic X receptor*

LISTA DE TABELAS

Tabela 1: Localização e principal substrato das ALDHs (baseado em Yoshida, 1998)..... 32
Tabela 2: Lista das ALDHs humanas utilizadas na busca de similaridade contra a base de seqüência do organismo de interesse..... 45
Tabela 3: Tabela com exemplos das distâncias entre as seqüências de ALDHs. A seqüência 2259 é uma ALDH obtida pelo método de busca e classificada como ALDH2 (em destaque)..... 46
Tabela 4: Nome do organismo, a classificação da ALDH e a referência da seqüência do organismo por nome do gene e número de acesso no Genbank ou do EMSEMBL. 67
Tabela 5: Comparação entre os comprimentos das seqüências de ALDHs obtidas pelo método desenvolvido com o comprimento das respectivas seqüências de aminoácidos esperado..... 72
Tabela 6: Tabela com quantidade de ALDHs obtidas, separas pela subfamília..... 74
Tabela 7: Relação dos genes e o comprimento da seqüência de aminoácidos traduzida a partir dos transcritos reconstruídos..... 75
Tabela 8: Lista dos fatores de transcrição preditos como possíveis ligantes na região promotora RARE-ß2. 116
Tabela 9: Número de ligações de hidrogênio formadas nas 7100 conformações analisadas. A análise foi feita para cada um dos 71 complexos gerados. Em branco estão as regiões flanqueadoras das hexades de interação. Em cinza claro a hexade de interação do RXR. Em branco entre as hexades do RXR e do RAR a região espaçadora. Em cinza escuro a hexade de interação do RAR. Na parte superior da tabela as ligações de hidrogênio feitas com as bases e com o fosfoaçúcar de fita direta do DNA. Na parte central as ligações de hidrogênio feitas com os átomos das bases e do FA de fita reversa do DNA. Na parte inferior da tabela temos os números totais de ligações de hidrogênio realizadas por cada posição relativa do DNA e também separados pelas ligações realizadas apenas com a base e apenas com o fosfoaçúcar. Foram mostradas apenas as posições que realizavam ligações de hidrogênio. Nas linhas temos as bases do DNA onde ocorre a ligação de hidrogênio. Nas colunas temos as posições relativas do DNA..... 123
Tabela 10: Número de ligações de hidrogênio feitas pelos aminoácidos com o DNA, classificado em interações feitas com o fosfoaçúcar e interações feitas com a base nitrogenada, nas 7100 conformações analisadas dos 71 complexos gerados. Na primeira tabela os aminoácidos do RXR que fazem interação com o DNA. Na segunda tabela os aminoácidos do RAR que fazem interação com o DNA. Na terceira e quarta tabela os resíduos do RXR e RAR respectivamente que fazem interação com as bases discriminando a quantidade de ligações feitas com cada uma das bases. As linhas Número se referem ao número do resíduo e Nome ao nome do resíduo. 129
Tabela 11: Quantidade de bases em cada posição relativa do DNA nas 71 conformações analisadas..... 132
Tabela 12: Quantidade de ligações realizada com as bases, agrupadas pelos aminoácidos. Também são apresentados os resíduos de Pearson mostrando a preferência ou a preterência a uma base. 134
Tabela 13: Quantidade ligações de hidrogênio realizadas nos 71 complexos estudados. Na primeira tabela a quantidade de ligações feitas apenas com os átomos das bases. Na segunda tabela a quantidade de ligações feitas apenas com os átomos do FA e na última

tabela a quantidade total. Nas linhas estão as quantidade de ligações e nas colunas as posições no DNA na qual o complexo esta posicionado. As células em negrito correspondem à melhor posição do complexo nativo e dos complexos com DNA mutados que não alteram a expressão do gene repórter. 138
Tabela 14: Informações referentes ao DNA nativo e as mutações m2 e m5. Na tabela 1 a quantidade de ligações de hidrogênio discriminada por tipo de base nitrogenada. Na tabela 2 a quantidade de ligações de hidrogênio discriminada pelas posições relativas do DNA. E na tabela 3 a quantidade de ligações de hidrogênio discriminado pelo número e nome dos aminoácidos que realizam a ligação. Não foram apresentadas as bases e os aminoácidos que fizeram menos de 3 ligações de hidrogênio..... 139

LISTA DE FIGURAS

Figura 1: Vias da síntese do ácido retinóico partindo do β-caroteno ou do éster retinil. 25
Figura 2: A via de síntese e degradação do ácido retinóico. O retinol é oxidado à retinaldeído pelas álcool desidrogenases (ADHs) em uma etapa reversível (Etapa de formação) e oxidado a ácido retinóico em uma etapa irreversível pelas ALDHs (Etapa posterior), sendo a ALDH1A2 a enzima mais importante neste processo. A etapa de degradação é mediada pelas enzimas da família P450 de citocromo-oxidases, as CYPs (Baseado em Pavam 2007). 27
Figura 3: Árvore filogenética das 17 ALDHs humanas, nome do gene e nome comum da enzima (baseado em Vasiliou, 2004)..... 30
Figura 4: Canais de acesso ao sítio catalítico das ALDHs. Em A) uma visão geral de uma ALDH com uma molécula de Nicotinamida-adenina-dinucleotídeo (NAD) à esquerda e uma molécula de daidzina (ver sessão 2.5.1 Definição dos aminoácidos de interação do sítio de ligação) no canal onde se aloja o aldeído a ser oxidado a direita. Em B), uma vista ampliada do canal onde se aloja o aldeído a ser oxidado com a daidzina e em destaque a cisteína catalítica..... 31
Figura 5: Esquema do processo de seqüenciamento de DNA de um genoma. Em A, temos a etapa de processamento, fragmentação e seqüenciamento do DNA. Em B, temos o processo de sobreposição, posicionamento dos fragmentos e definição da seqüência consenso das seqüências dos fragmentos obtidos na etapa A. 40
Figura 6: Esquema do método desenvolvido para busca de ALDH em organismos não anotados completamente. (A) Ciclo principal de busca. (B) Alinhamento múltiplo das seqüências obtidas para o organismo-alvo. (C) Definição da seqüência de aminoácidos.... 44
Figura 7: Árvore filogenética baseada no alinhamento das ALDHs humanas utilizadas para classificar as seqüências das ALDHs obtidas..... 47
Figura 8: Esquema de entidades de relacionamento criadas para armazenar as informações dos resultados de busca por seqüência de ALDHs. 48
Figura 9: Resíduos da estrutura cristalográfica de PDBid 1OF7, que possuem algum átomo da cadeia lateral a uma distância máxima de 5 Å da molécula de daidzina (verde). 52
Figura 10: Etapas para definição da região de busca pelo canal de acesso ao substrato. Em A) a estrutura cristalográfica de PDBid 1OF7, em azul ALDH2 e em laranja a daidzina. Em B) uma vista aproximada da daidzina na região do canal de acesso ao substrato. Em C) a daidzina e os aminoácidos que estão a uma distância máxima de 5Å. Em D) a caixa de 5Å (verde) definindo a primeira região para a procura pelo canal de acesso ao substrato. Em E) a caixa de 5Å (verde) e a caixa redefina (vermelha) que delimitam a região de busca pelo canal de acesso ao substrato. Em F) apenas a caixa, utilizada neste trabalho, que delimita a região para a busca pelo canal de acesso ao substrato..... 55
Figura 11: Em A) a caixa na região onde se tem interesse em procuras cavidades. Em B) a caixa com o GRID que era utilizado para o posicionamento do átomo de sondagem. 56
Figura 12: Etapas do processo para busca por cavidade. Em A) é mostrado um corte de um dos planos do GRID e em preto a cadeia de carbonos utilizada para esquematizar o processo de busca por cavidades. Em B) um esquema simplificado do desenho em (A), em azul claro a representação dos átomos de carbono e em preto a ligação entre os carbonos, a distância entre as linhas do GRID foi alterada para 2,8Å para simplificar a visualização. Em C) uma representação mostrando quando um ponto é considerado dentro de uma cavidade

(ponto verde) e não é considerado em uma cavidade (ponto vermelho). Em D) todos os pontos aceitos em verde e todos os pontos rejeitados em vermelho e a lista dos pontos aceitos na coordenada relativa do GRID. 58
Figura 13: Representação do processo de agrupamento dos pontos simplificados para um sistema bidimensional. Em A, o início do processo no qual define o primeiro ponto com o valor 0 e procura por pontos adjacentes. Em B, os pontos adjacentes com o mesmo valor do primeiro ponto. Em C, todos os pontos adjacentes ao primeiro ponto definido com o mesmo valor e um segundo ponto definido com o valor 1. Em D, todos os pontos com os seus grupos definidos. Em E, um exemplo do resultado do método de agrupamento de pontos, onde cada cor representa um grupo formado..... 60
Figura 14: As duas possibilidades para calcular o volume. Em A quando é possível formar um cubo. Em B quando não é possível formar um cubo, mas é possível formar tetraedros não sobrepostos. 62
Figura 15: Três vistas do canal de acesso ao substrato. Em A, temos uma vista ampla da proteína e o canal de acesso ao substrato definido no interior em azul escuro. Em B, uma vista ampliada na região do canal de acesso ao substrato e em C, apenas o contorno do canal de acesso ao substrato. 64
Figura 16: Processo de definição dos pontos da superfície da cavidade. Em A, todos os pontos de um corte de uma cavidade. Em B, a etapa de busca por pontos da superfície no eixo X. Em C, a etapa de busca por pontos da superfície no eixo Y. Em D, apenas os pontos considerados da superfície..... 65
Figura 17: Mapeamento dos genes preditos das ALDH1 e ALDH2 sobre os scaffolds do genoma de anfioxo. Foram identificados seis possíveis genes que codificariam a ALDH1, onde cinco desses apresentam-se com duas cópias no genoma. O gene da ALDH2 apresenta apenas uma seqüência com duas cópias no genoma..... 76
Figura 18: Análise filogenética em larga escala das ALDH de eucariotos. Cada subfamília de ALDH é individualmente suportada por uma probabilidade posterior maior que 0,95. . 77
Figura 19: Filogenia diagramática das ALDH1/2s, ALDH1Ls e ALDH8s. As ALDH8s foram utilizadas como grupo externo. ALDH1/2 de metazoários, plantas e fungos formam três clados distintos e bem suportados sugerindo que estes três grupos evoluíram independentemente a partir de uma única ALDH1/2 eucariótica ancestral. As ALDH1/2s metazoários que incluem ALDH1s, ALDH2s e ALDH1B1s, ALDH2 duplicadas específicas de vertebrados. Também está incluída na figura uma representação em três categorias o tamanho do canal. 79
Figura 20: As análises para as ALDH1 duplicadas do Anfioxo sugerem que ao menos alguns genes como ALDH1c e ALDH1d (no scaffold 155), ALDH1a a ALDH1e (no scaffold 560) ou ALDH1a, ALDH1e e ALDH1f (no saffold 31) localizadas no mesmo scaffold e provavelmente ligadas no mesmo cromossomo. As análises filogenética e de sintenia sugerem que as seis ALDH1 de Anfioxo estão relacionadas a duplicação em série do ancestral da ADLH1a. 80
Figura 21: Os canais de acesso ao substrato das estruturas de PDBid 1OF7 (ALDH2) e 1BXS (ALDH1). O canal da ALDH2 apresenta a entrada (próximo ao resíduo 124), o pescoço (próximo ao resíduo 459) e o fundo (próximo ao resíduo 303) mais estreito do que o canal ALDH1..... 81
Figura 22: Diversidade das assinaturas do canal de acesso ao substrato em Anfioxo. Nos vertebrados o canal de acesso ao substrato das ALDH1s e ALDH2s pode ser distinguido pelas posições específicas de 3 aminoácidos: resíduo 124 (boca), 459 (pescoço) e 303

(fundo). As ALDH1s do Anfioxo apresentam um largo espectro de assinaturas variando de aminoácidos pequenos típicos de ALDH1As de vertebrados (Gly124, Val459, Ile303) a aminoácidos grandes típicos de ALDH2 de vertebrados (Met124, Phe459, Cys303). Enquanto a ALDH1a e ALDH1d se parecem mais com as ALDH1As de vertebrados as outras quatro ALDH1As de Anfioxo incorporaram aminoácidos típicos de ALDH2 de vertebrados que também são encontrados nas ALDH2 de Anfioxo..... 84
Figura 23: As ALDH1 duplicadas de tunicados apresentam as mesmas assinaturas das ALDH2. Todas as ALDH1s de tunicados são caracterizadas por um aminoácido grande na posição 459 ("pescoço" do canal) que causam um estreitamento do canal. Entretanto, para as duas assinaturas do canal, existe um contínuo entre ALDH1 de tunicados, que possui os aminoácidos do canal de acesso ao substrato mais relacionados com as ALDH1As de vertebrados (C. savignyi e C. intestinalis ALDH1a), e ALDH1s de tunicados que possui o canal mais relacionado com as ALDH2 (C. savignyi ALDH1b com também C. intestinalis ALDH1b1 e ALDH1b2). As ALDH1a de C. savignyi e C. intestinalis são ambas caracterizadas pela Gly 124, que deixa a boca do canal livre para acomodar a β-ionona do retinaldeído, e pela Thr 303 que ajuda acomodar a porção aldeído do retinaldeído. Em contraste, C. savignyi ALDH1b como a também C. intestinalis ALDH1b1 e ALDH1b2 incorporaram um aminoácido grande, que pode impedir a acomodação da β-ionona. A ALDH1b de C. savignyi e também as ALDH1b1 e ALDH1b2 de C. intestinalis incorporaram a Cys303, que possui o átomo de enxofre que pode causar impedimento com a parte aldeído do retinaldeído. Assim, as ALDH1s duplicadas de tunicados incorporaram aminoácidos típicos de ALDH2 de metazoários sugerindo uma diferenciação das características de ALDH1 ficando mais parecida com as ALDH2. 86
Figura 24: Expressão dos gene de ALDH1/2 de anfioxo. A expressão evidente das ALDH1a é detectável após 12 horas e posteriormente na endoderme após 24 horas. Com 12 horas, as outras cinco ALDH1 são fracamente expressas na região posterior e após 24 horas estão difusamente expressas em todo o embrião. A ALDH2 é expressa posteriormente após 12 horas e após 24 horas é expressa em todo o embrião. Assim, enquanto a ALDH1a provavelmente tem um padrão antero-posterior de posicionamento, as outras ALDH1 possivelmente funcionam como protetora para os diferentes tipos de aldeídos, como a ALDH2. 88
Figura 25: A plasticidade do canal de acesso ao substrato dos deuterostômios. As ALDH2s de vertebrados e equinodermas oferecem mais exemplos da plasticidade evolucionaria do canal das ALHD1/2. Nos vertebrados, as ALDH2s se duplicaram em ALDH2s e ALDH1B1s. Em vertebrados ALDH2s conservaram as assinaturas (MET 124 e LEU 459) mantendo a forma estreita da "boca" e do "pescoço" do canal. A ALDH1B1s incorporou resíduos pequenos (Glu 124 e Val 459) que oferece menos resistência para a acomodação da β-ionona do retinaldeído e também aumentando o canal no "pescoço" elevando o volume do canal (362 ± 30 $Å^3$ contra 451 ± 30 $Å^3$ para ALDH2 e ALDH1B1, respectivamente, $p<0,05$). Uma tendência similar é observada em ouriço do mar (Strongylocentrotus purpuratus) ALDH2s. S. purpuratus ALDH2b incorporou aminoácidos pequenos (Gly 124 e Val 459), que deixa a boca do canal mais aberta possibilitando a acomodação da β-ionona do retinaldeído e aumentando o volume do canal (356 $Å^3$ na ALDH2a para 487 $Å^3$ na ALDH2b). 90
Figura 26: Esquema do processo para o cálculo da energia de ligação entre DNA e proteína. 99

Figura 27: Processo da sobreposição estrutural do RXR sobre o DNA a ser analisado. Em A, na parte superior a estrutura tridimensional do RXR e o DNA da estrutura cristalográfica (PDBid 1DSZ) na parte inferior a estrutura tridimensional do DNA a ser analisado. Em B, a sobreposição dos DNAs. Em C, uma vista detalhada da sobreposição dos DNAs, em destaque os átomos de fósforo. Em D, o RXR sobreposto no DNA a ser analisado. 101
Figura 28: Estrutura do complexo RXR/RAR- RARE-β2 na posição 6 do DNA mostrando que uma posição do DNA representam 17 bases sob influência dos fatores de transcrição. 102
Figura 29: Esquema do clico termodinâmico utilizado para calcular a energia de ligação entre o DNA e os fatores de transcrição. 104
Figura 30: Exemplo mostrando os quatro modos de escolha do valor médio da energia de interação. Em A) as primeiras 40 energias da posição 6 do DNA nativo. Em B) os quatro modos de seleção da melhor janela. Os números acima de cada tabela correspondem a ordem da ordenação dos dados. As janelas ordenadas primeiro por z-escore apresentam energias maiores do que as janelas ordenadas primeiro pela média mas apresentam menor desvio padrão. A comparação das quatro abordagens com os dados experimentais de Vivanco Ruiz *et al.* (1991) irá definir qual é o melhor estimador para busca de sítios..... 108
Figura 31: Gráficos comparando as energias de ligação média entre todos os DNAs analisados. Os gráficos à esquerda são das mutações (m2 e m5) que não alteraram a expressão do gene repórter nos experimentos de Vivanco Ruiz *et al.* Os gráficos centrais são das mutações que diminuiram a expressão do gene repórter (m6 e m9) e os gráficos à direita são das mutações que inibiram a expressão do gene repórter (m3, m4, m7 e m8). Em todos os gráficos é apresentada a energia do complexo proteína-DNA nativo para facilitar a comparação. A primeira trinca de gráficos são referentes aos dados obtidos a partir de janelas temporais de tamanho mínimo de 20 energias, a segunda trinca de gráficos janelas temporais de 30 energias, assim até energias 50. 109
Figura 32: Esquema da nomenclatura relativa às posições das hexades de interação. Em A) vista geral da estrutura e as regiões das hexades e o espaçador. Neste exemplo a hexade do RXR inicia na posição 7 do DNA. Em B) uma visão detalhada das hexades e do espaçador com os nomes das bases na nomenclatura relativa à hexade..... 111
Figura 33: Esquema de um par de nucleotídeos, Adenina e Timina, onde os átomos da base estão em branco e os átomos do fosfoaçúcar estão em preto. 112
Figura 34: Representação esquemática da base de dados relacional com as informações das ligações de hidrogênio. 112
Figura 35: Dados obtidos por Vivanco Ruiz *et al*. mostrando os efeitos de mutações na expressão de um gene repórter ligado a montante da região promotora do gene RARE-ß2 (baseado em Vivanco Ruiz *et al.* 1991)..... 114
Figura 36: Seqüência da região promotora do RARE-ß2 definida no trabalho de Vivanco Ruiz *et al.* (1991). Em destaque as regiões que afetam a transcrição do gene repórter quando são mutadas. 115
Figura 37: Estrutura do complexo RXR/RAR- RARE-ß2 na posição 6 do DNA, após a dinâmica molecular..... 117
Figura 38: Gráficos das energias médias, medianas e mínimas nas 15 posições do DNA nativo. 118
Figura 39: Os quatros modos de análise utilizando o método de janela temporal para o DNA nativo. Na parte superior da figura a legenda para todos os gráficos. 119

Figura 40: Gráficos representantes de cada grupo dos modos de análise utilizando janelas temporais. No canto superior esquerdo da figura está a legenda para todos os gráficos. A primeira coluna de gráficos representa as energias das mutações que não alteram a expressão do gene repórter. A coluna do meio representa as energias das mutações que diminuem a expressão do gene repórter. A terceira coluna representa as energias das mutações que inibem a expressão do gene repórter. 121
Figura 41: Vista estereográfica da região da hexade e da alfa-hélice do RXR que fica inserida na fenda maior do DNA. Em azul escuro a hélice do RXR, em amarelo os resíduos P1, P2, P3 e P4 da fita direta, em laranja os resíduos P3, P4, P4 e P6 da fita complementar e em azul claro os resíduos P5, P6 e P1, P2 da fita direta e complementar, respectivamente. 125
Figura 42: Figura esquemática das ligações de hidrogênio onde o tamanho do fosfoaçúcar e da base representa a quantidade de ligações de hidrogênio feitas. 127
Figura 43: Vista estereográfica da região de contato entre o DNA e o RXR . Em azul escuro a hélice do RXR, em amarelo os resíduos P2 e P3 da fita direta, em laranja os resíduos P2 e P3 da fita complementar e em vermelho umas das possíveis ligações de hidrogênio feitas entre os resíduos e o DNA. 130
Figura 44: Relação das interações específicas que cada aminoácido realiza com o DNA. Cada cor sobre o DNA corresponde às bases que um dado aminoácido interage. 133
Figura 45: Quantidade de ligações de hidrogênio específicas realizada com a base no DNA nativo. 136
Figura 46: Quantidade de ligações de hidrogênio realizada com a base em todos os DNAs analisados. 137

SUMÁRIO

AGRADECIMENTOS **3**
RESUMO **4**
ABSTRACT **6**
LISTA DE ABREVIATURAS **8**
LISTA DE TABELAS **9**
LISTA DE FIGURAS **11**
SUMÁRIO **16**
CAPÍTULO 1: Introdução **18**
2. Organização da tese **20**
3. Objetivos gerais **21**
3.1 Objetivos específicos 22
CAPÍTULO 2: Estudo da via de sínteses do ácido retinóico. **24**
1.1 Ácido retinóico 24
1.2 As famílias da enzima aldeído desidrogenase 29
1.3 Mecanismo catalítico 30
1.4 Família das ALDHs 32
1.5 Ácido retinóico e desenvolvimento cardíaco 35
2. Métodos **38**
2.1 Conceito 38
2.2 Obtenção das seqüências de ALDH 42
2.3 Alinhamento múltiplo das seqüências das ALDHs para análise filogenética 49
2.4 Busca por regiões específica (assinatura) 49
2.5 Obtenção das estruturas cristalográficas das ALDH e modelagem molecular comparativa 50
2.5.1 Definição dos aminoácidos de interação do sítio de ligação 51
2.6 Método para procura por cavidades a partir das estruturas tridimensionais e cálculo do volume 52
2.6.1 Utilização dos programas disponíveis 52
2.6.2 Definição da região de busca por cavidades 53
2.6.3 Busca por cavidades 56
2.6.4 Agrupamentos dos pontos 59
2.6.5 Cálculo do volume 61
2.6.6 Criação da superfície da cavidade 63
2.7 Clonagem dos genes de anfioxo e expressão gênica 65
2.8 Análise estatística 66
3. Resultados **67**
3.1 Busca por seqüências de ALDH em organismos com genomas montados e anotados 67
3.2 Busca por seqüência de ALDH em organismos não montados 70
3.2.1 Análise do genoma do anfioxo 75
3.3 Análises filogenética 76
3.4 Análise da forma e do volume do canal de ligação do substrato 80
3.5 Assinaturas 82
3.5.1 Relação entre as assinaturas e a estrutura 83

3.5.2 Expressão das seqüências putativas de ALDHs de Anfioxo 87
4. Conclusões 91
CAPÍTULO 3: Análise da especificidade na ligação fatores de transcrição, regulados por ácido retinóico, ao DNA. 93
1.1 Análise de sítio de ligação de fatores de transcrição utilizando bases de dados 95
1.2 Proposição de uma abordagem alternativa para análise de sítio de ligação de fatores de transcrição utilizando estruturas tridimensionais de macromoléculas biológicas 96
1.3 Fatores de transcrição regulados por ácido retinóico 96
2. Método 97
2.1 Predição de sítios de ligação baseado na seqüência nucleotídica 97
2.2 Método para busca por sítio de ligação de fatores de transcrição 98
2.3 Construção dos complexos 99
2.3.1 Construção do DNA 99
2.3.2 Posicionamento das proteínas sobre as hexades e preparação para minimização. 100
2.3.3 Minimização de energia e Dinâmica molecular 102
2.3.4 Conceito 103
2.4 Cálculo da energia de ligação 105
2.4.1 Análise das energias obtidas 106
2.5 Análise das ligações de hidrogênio 110
3. Resultados 113
3.1 Predição de sítios de ligação baseado na seqüência nucleotídica 115
3.2 Validação do método de busca por sítios de fatores de transcrição utilizando de estruturas tridimensionais e métodos termodinâmicos 117
3.2.1 Análise das ligações de hidrogênio 122
4. Discussão 140
4.1 Análise da busca por sítios de ligação de fatores de transcrição utilizando matrizes 140
4.2 Análise da busca por sítios de ligação de fatores de transcrição utilizando a estrutura tridimensional e cálculos termodinâmicos 140
4.3 Análises das ligações de hidrogênio. 142
5. Conclusões 148
CONCLUSÕES GERAIS 149
REFERÊNCIAS BIBLIOGRÁFICAS 150
Anexo 1 I
Anexo 2 VII
Anexo 3 X
Anexo 4 XV
Anexo 5 XXIII

CAPÍTULO 1: Introdução

A origem dos cordados e suas inovações no padrão corporal ainda são controversas. Por causa da ação do ácido retinóico (AR) no padrão dos eixos embrionários parecerem restritos aos cordados, foi proposto que a morfogenia regulada pelo ácido retinóico foi uma inovação ligada à origem dos cordados (Canestro, Postlethwait et al. 2006). Nos vertebrados, a ação do ácido retinóico é controlada em dois níveis: no metabolismo do AR e na sinalização pelo AR. A sinalização pelo ácido retinóico, e seus derivados é responsável pelo controle de várias funções biológicas como, por exemplo, crescimento celular, diferenciação celular, formação da retina, desenvolvimento cardíaco e também relacionado a várias patologias como diabetes, obesidade, cânceres, e doenças cardiovasculares (McGrane 2007). A maquinaria que controla o metabolismo inclui as enzimas de síntese do AR (retinaldeído desidrogenase ALDH1A) e as enzimas de degradação do AR (Cyp26), que controlam a distribuição espaço temporal do AR durante a embriogênese (Canestro, Postlethwait et al. 2006). A maquinaria da sinalização celular por AR contém os receptores nucleares controlados por AR (RARs) que estão envolvidos com o controle da transcrição gênica.

A partir das informações de banco de dados de seqüências, de técnicas de filogenia e ferramentas de bioinformática, Cañestro *et. al.* (2006) conseguiram encontrar possíveis seqüências da maquinaria do AR (ALDH1A, CYP26 e RAR) em organismos não cordados mostrando que a utilização do ácido retinóico não é uma inovação que surgiu com os cordados. A associação de várias técnicas como: biologia molecular, bioinformática, filogenia, análises estruturais de biomoléculas, mecânica molecular e métodos

termodinâmicos tem se mostrado uma poderosa abordagem para compreensão de sistemas biológicos, simplificando e agilizando o desenvolvimento do conhecimento científico.

Em humanos, a família de enzimas ALDHs apresentam 17 genes correlacionados. Esses genes são muito conservados tanto em relação à seqüência de aminoácidos quanto a estrutura tridimensional da proteína. As ALDHs são enzimas $NAD(P)^+$ dependentes, que oxidam uma ampla gama de aldeídos para os seus correspondentes ácidos carboxílicos (Yoshida, Rzhetsky et al. 1998; Vasiliou, Pappa et al. 2004). Essas enzimas possuem dois canais, um por onde entra o $NAD(P)^+$ e outro por onde entra o aldeído a ser oxidado, no fundo desse canal encontra-se o sítio catalítico. Cada uma dessas enzimas apresenta especificidade por um grupo de aldeídos. No entanto essa grande semelhança seqüencial e estrutural dificulta o entendimento da especificidade por um determinado aldeído.

Depois de sintetizado, o AR atua ligando-se à receptores nucleares, *retinoid acid receptors* (RARs), que são ativados pelo all-*trans* ácido retinóico e 9-*cis* ácido retinóico (os dois principais isômeros do ácido retinóico), e *retinoid X receptors* (RXRs), que são ativados especificamente pela 9-*cis* ácido retinóico (Simões-Costa 2008). Para desempenhar seu papel como receptores nucleares, os RXRs formam heterodímeros com RARs reconhecendo uma seqüência particular a partir da complementaridade espacial de suas cadeias laterais com a estrutura tridimensional do DNA. Essas regiões geralmente são compostas por 10 a 30 nucleotídeos, conhecidos como *retinoic acid response elements* (RAREs), geralmente localizados próximo à região de início da transcrição do gene. O conhecimento da exata localização dessas regiões promotoras auxilia no entendimento da expressão gênica, que é de extrema importância para a compreensão do funcionamento dos organismos e também das doenças que afetam esses organismos. Existem algumas maneiras de prever a localização das regiões promotoras, a maneira mais eficiente é por

meio de complexos métodos de biologia molecular. Para tentar simplificar essa previsão foram criados métodos computacionais. Os métodos mais comuns utilizam informações pré-existentes em bases de dados de sítios de ligação de fatores de transcrição, buscando sua presença na seqüência nucleotídica, analisada por meio de métodos de comparação de seqüência. Esses métodos apresentam algumas falhas nas predições dos possíveis fatores de transcrição, muitas vezes perdendo sítios verdadeiros ou gerando resultados falso-positivos. Uma outra forma de buscar sítios de ligação de fatores de transcrição é o uso de métodos computacionais que não necessitem, a priori, de informações referentes ao sítio de ligação. A disponibilidade de estruturas tridimensionais de fatores de transcrição, resolvidas por técnicas de cristalografia, torna-se informação importante quando utilizada junto a técnicas de análise estruturais e métodos termodinâmicos. Essa abordagem interdisciplinar auxilia na compreensão de alguns dos aspectos relacionados à interação entre fatores de transcrição e o DNA, auxiliando na compreensão da regulação gênica.

2. Organização da tese

Esta tese foi dividida em duas partes. A primeira parte procurou responder às seguintes perguntas: "*É possível reconstruir as seqüências de genes da ALDH a partir das informações provindas de bancos de seqüências de DNA utilizando genes homólogos como molde?*", "*Existem grupos de aminoácidos que podem ser utilizados para diferenciar a função de uma ALDH*?" e "*É possível diferenciar os grupos de ALDHs existentes a partir do formato e do volume do canal onde se aloja o aldeído a ser oxidado?*". Esse trabalho multidisciplinar foi realizado em colaboração com vários grupos de pesquisa com diferentes especialidades:

- Grupo de Biologia Computacional do Laboratório de Genética e Cardiologia Molecular do Instituto do Coração da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brasil.
- Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, França.
- Grupo de estudo do desenvolvimento cardíaco do Laboratório de Genética e Cardiologia Molecular do Instituto do Coração da Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brasil.
- Departament de Genética, Facultat de Biologia, Universitat de Barcelona, Barcelona, Espanha.
- Division of Biology 139-74, Califórnia Institute of Technology, Pasadena, Califórnia, USA.

A segunda parte deste trabalho procurou responder às seguintes perguntas: "*É possível, por meio do cálculo da energia de interação entre o fator de transição e o DNA, prever qual a melhor região do DNA para a ligação de fatores de transcrição?*" e "*É possível, através das ligações de hidrogênio, tentar descrever os mecanismos moleculares responsáveis pela interação específica entre um fator de transcrição e o DNA ?*".

3. Objetivos gerais

O presente trabalho teve como primeiro objetivo estudar aspectos envolvidos na via de regulação gênica mediada por AR. Isto envolveu a análise da evolução das funções das enzimas ALDH, utilizando técnicas de genômica combinatória, filogenia, bioinformática, estrutura de biomoléculas e de biologia do desenvolvimento, tentando compreender o modo como as ALDHs, que apresentam as seqüências de aminoácidos bastante similares, puderam divergir para gerar funções diversas. O entendimento da interação entre fatores de

transcrição e o DNA, auxilia na compreensão de problemas relacionados à regulação gênica mediada por AR. Foi baseada na análise de estruturas tridimensionais dos fatores de transcrição regulados por ácido retinóico RXR e RAR associada a estudos termodinâmicos.

3.1 Objetivos específicos

Síntese do ácido retinóico:

1. Buscar seqüências de ALDHs em organismos que possuem o genoma seqüenciado e anotado e também em vários organismos basais que ainda não possuam o seu genoma montados e/ou anotados, utilizando as informações provindas dos projetos de seqüenciamento de EST.
2. Realizar a análise filogenética das seqüências das ALDHs, obtidas na primeira etapa, para o estudo evolutivo dos grupos de ALDHs existentes ao longo da evolução dos organismos e análise de suas funções.
3. Buscar assinaturas (região com aminoácidos específicos para cada grupo de ALDH) nas seqüências obtidas na primeira etapa entre os grupos das ALDH2 e ALDH1A para a criação de um possível sistema de classificação de seqüência de aminoácidos quanto ao grupo de ALDH.
4. Construir as estruturas tridimensionais, utilizando a técnica de modelagem molecular por homologia, das seqüências de ALDHs obtidas na primeira etapa e também para o mapeamento das assinaturas definidas na etapa 3 para um possível entendimento da importância dessas regiões.

5. Definir estruturalmente o canal de acesso do substrato e calcular o volume desse canal das enzimas modeladas na etapa 4 para verificar a relação entre a forma e o volume do canal e a função da enzima (sinalização ou destoxificação).
6. Estabelecer uma hipótese sobre evolução das funções das enzimas aldeído desidrogenase.

Sinalização pelo ácido retinóico:

1. A partir de estruturas tridimensionais de fator de transcrição resolvidas por cristalografia de raios-X criar complexos (fatores de transcrição ligado a um DNA) utilizando diferentes seqüências de DNA. Utilizar técnicas de mecânica molecular para explorar as possíveis conformações dos complexos criados.
2. Criação de um método para a identificação de sítios de ligação de fatores de transcrição baseada em algoritmos de análise estrutural de macromoléculas biológicas e métodos termodinâmicos.
3. Análise das ligações de hidrogênio entre os DNAs e os fatores de transcrição regulados por ácido retinóico descrevendo as ligações importantes para a especificidade dessa interação.

CAPÍTULO 2: Estudo da via de sínteses do ácido retinóico.

As vias de sinalização celular por meio de moléculas são um dos principais meios de controle de um organismo. O entendimento das funções de moléculas sinalizadoras facilita a compreensão das vias metabólicas de um organismo, assim possibilitando uma melhor compreensão de vários eventos biológicos e também de várias doenças.

A sinalização pelo ácido retinóico, e seus derivados, é responsável pelo controle de várias funções como, por exemplo, crescimento celular, diferenciação celular, formação da retina, desenvolvimento cardíaco e também relacionado a várias patologias como diabetes, obesidades, cânceres, e doenças cardiovasculares, etc (McGrane 2007; Ziouzenkova and Plutzky 2008). A compreensão dessa via de sinalização pode ajudar a entender vários mecanismos biológicos e também algumas patologias humanas.

1.1 Ácido retinóico

O ácido retinóico (AR) (Figura 1) é um metabólito obtido a partir da vitamina A (retinol). Os mamíferos não conseguem sintetizar a vitamina A sendo então necessário obte-la a partir da dieta na forma de carotenóides provindos das plantas ou na forma de ésteres retinil provindo dos animais (Davidovici, Tuzun et al. 2007; Maden 2007). Estes compostos são armazenados na forma de ésteres retinil (também conhecidos como retinóides) principalmente no fígado, mas também são encontrados nos pulmões, medula e rins (Maden 2007).

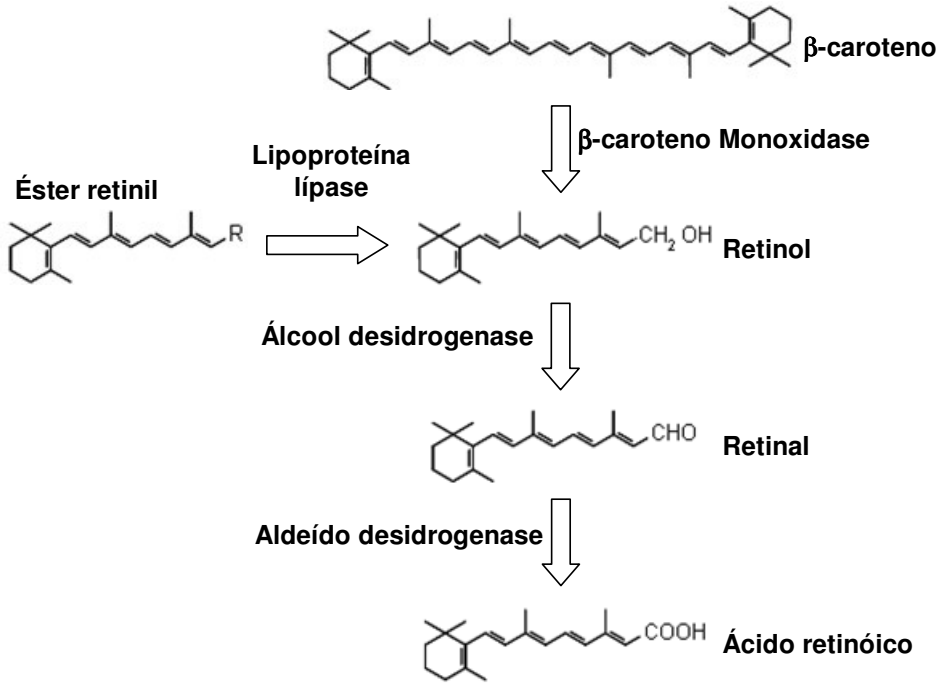


Figura 1: Vias da síntese do ácido retinóico partindo do β-caroteno ou do éster retinil.

A distribuição dos retinóides pelo organismo é feita na forma de retinol associado à Proteína Ligante de Retinol 4 (em inglês *Retinol Binding Protein 4* RBP4). As células que possuem receptores para a RBP4 em sua membrana fazem com que o complexo retinol-RPB4 entre para o citoplasma. No interior da célula o retinol é processado formando o ácido retinóico através de duas etapas oxidativas (Figura *2*) (Vasiliou, Pappa et al. 2004; Maden 2007):

1. Retinol a retinal (Etapa de Formação).
2. Retinal a ácido retinóico (Etapa Posterior).

A Etapa de Formação consiste de uma reação de oxidação reversível e requer a atividade enzimática de diversos membros da família das álcool-desidrogenases (ADH) (Perlmann, Rangarajan et al. 1993; Vasiliou, Pappa et al. 2004). Trabalhos em camundongos mutantes para ADHs indicam que essas enzimas são as principais enzimas na conversão de retinol em retinal (ou retinaldeído) em amniotos (Molotkov, Fan et al. 2002). As enzimas ADHs possuem uma distribuição tecidual bastante ampla. Dessa forma, pode-se considerar que todas as células embrionárias são capazes de oxidar retinol em retinaldeído.

A Etapa Posterior (Figura 2) é irreversível e requer a ação enzimática de membros da família das aldeído desidrogenases (ALDH) (Perlmann, Rangarajan et al. 1993), dentre as quais a ALDH1A2 é a principal enzima sintetizadora de ácido retinóico no embrião em desenvolvimento e durante o qual é expressa na mesoderme embrionária (Zhao, McCaffery et al. 1996; Niederreither, McCaffery et al. 1997; Vasiliou, Pappa et al. 2004). Em contraste com a atividade das ADHs, que são ubíquas devido à expressão onipresente de ADH3 e à ampla distribuição de ADH4 e ADH1, as ALDH1As têm uma distribuição mais restrita (Blentic, Gale et al. 2003) e perfis de expressão que quase completamente coincidem com a ativação da sinalização pelo AR. Dessa forma, pode-se concluir que são as enzimas ALDH1As que delimitam os locais onde o AR é sintetizado (Simoes-Costa, Azambuja et al. 2008).

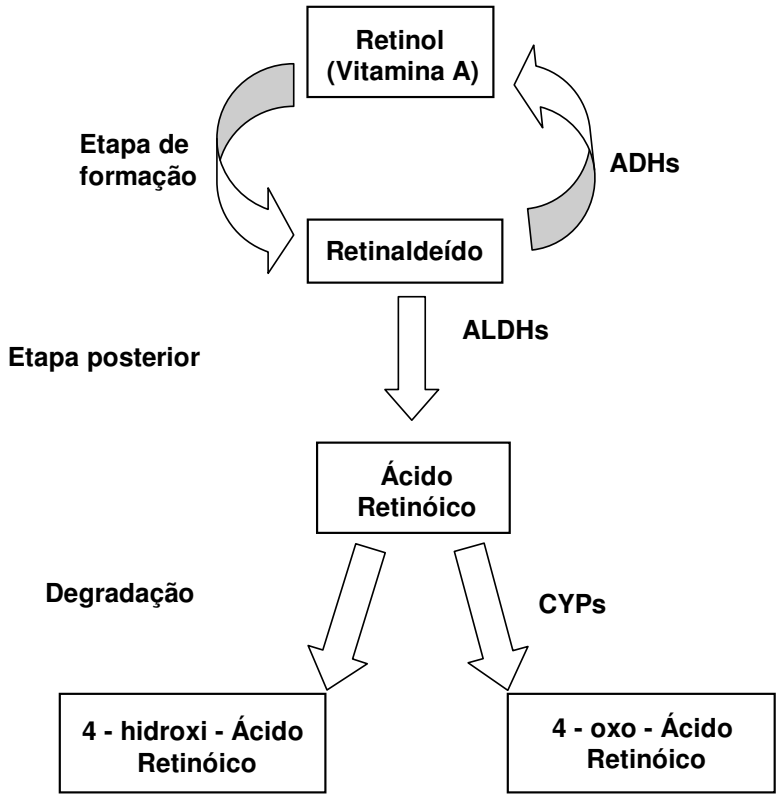


Figura 2: A via de síntese e degradação do ácido retinóico. O retinol é oxidado à retinaldeído pelas álcool desidrogenases (ADHs) em uma etapa reversível (Etapa de formação) e oxidado a ácido retinóico em uma etapa irreversível pelas ALDHs (Etapa posterior), sendo a ALDH1A2 a enzima mais importante neste processo. A etapa de degradação é mediada pelas enzimas da família P450 de citocromo-oxidases, as CYPs (Baseado em Pavam 2007).

O AR é degradado a metabólitos parcialmente ativos ou inativos pelas enzimas P450 tais como CYP26A, B e C (Figura 2). A super-família do citocromo P450 (CYP) catalisa um grande número de importantes reações biológicas. Mais especificamente, a CYP26 é essencial para a produção de 4–hidroxi–ácido retinóico e 4–oxo–ácido retinóico que são produtos de degradação inativos do ácido retinóico (Perlmann, Rangarajan et al. 1993). A função da CYP26, particularmente, parece estar relacionada à proteção dos tecidos contra a alta exposição ao ácido retinóico, pois o seu padrão de expressão é complementar ao da enzima produtora de ácido retinóico (Swindell and Eichele 1999;

Swindell, Thaller et al. 1999). Os animais mutantes homozigotos para o gene CYP26 também apresentam defeitos cardíacos (como tubo cardíaco dilatado, por exemplo) e morrem ainda durante os estágios embrionários ou mesmo logo após o nascimento (Abu-Abed, Dolle et al. 2001; Sakai, Meno et al. 2001). Esta enzima é encontrada em estágios precoces de embriões de camundongo e de galinha e experimentos *in vitro* e *in vivo* mostram, ainda, que sua expressão é regulada pelo ácido retinóico (Ray, Bain et al. 1997; Reijntjes, Gale et al. 2004).

Trabalhos recentes têm revelado novos aspectos da via de síntese do ácido retinóico. A reação que oxida o retinol em retinaldeído é primeiramente catalisada pelos membros da família das álcool-desidrogenases. Entretanto, outra família de enzimas, como a retinol desidrogenase 10 (em inglês *Retinol Dehydrogenase 10*, RDH10), desempenha papel crítico na síntese de ácido retinóico no embrião. Sandell (Sandell, Sanderson et al. 2007) mostra que proteínas mutantes para este gene perdem sua capacidade de oxidar retinol em retinaldeído, acarretando uma sinalização insuficiente pelo ácido retinóico e levando a malformação, inclusive à morte embrionária. Kawaguchi (Kawaguchi, Yu et al. 2007) revela novos aspectos sobre o mecanismo molecular de absorção de vitamina A pelas células, mecanismo pouco conhecido até então. Mesmo com essas novas informações sobre a via de síntese do ácido retinóico, algumas evidências indicam ainda que a etapa posterior da síntese de AR prevalece como sendo a mais importante. Primeiro, o fenótipo do *knockout* de ALDH1A2 é claramente mais severo do que o observado para o mutante RDH10, mostrando que embriões mutantes ALDH1A2 morrem antes que os embriões mutantes RDH10 indicando que os embriões ALDH1A2 sofrem uma maior perda de sinalização pelo AR.

1.2 As famílias da enzima aldeído desidrogenase

O genoma humano contém pelo menos 17 genes que são membros da super-família de enzimas aldeído desidrogenase (número 1.2.1.3 do Enzyme commission (EC), número 53726 do SCOP). Todos esses genes codificam enzimas $NAD(P)^+$ dependentes, que oxidam uma ampla gama de aldeídos para os seus correspondentes ácidos carboxílicos (Yoshida, Rzhetsky et al. 1998; Vasiliou, Pappa et al. 2004). A comparação das seqüências de aminoácidos mostra uma extensa similaridade entre as ALDH de bactéria e as ALDH humanas sugerindo um ancestral comum de aproximadamente 3 bilhões de anos.

As ALDHs são classificadas com base na porcentagem de identidade entre as seqüências de aminoácidos. As proteínas que apresentam menos que 40% de identidade são classificadas em diferentes famílias (números de 1 à 18 até o momento). As proteínas que apresentam mais que 60% de identidade são classificadas na mesma subfamília (letra A ou B) (www.aldh.org). O genoma humano contém 17 ALDHs agrupadas em 10 famílias e 13 subfamílias (Figura 3).

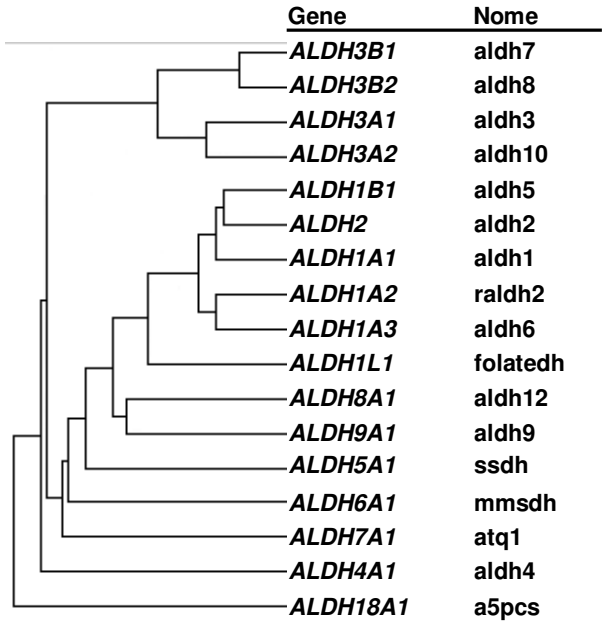


Figura 3: Árvore filogenética das 17 ALDHs humanas, nome do gene e nome comum da enzima (baseado em Vasiliou, 2004).

1.3 Mecanismo catalítico

As enzimas ALDHs possuem um mecanismo geral de catálise baseado na presença de dois canais: um para a ligação e oxidação do aldeído e outro para a ligação de um cofator que captura elétrons liberados no processo (Figura 4 A). No fundo do canal onde se aloja o aldeído a ser oxidado encontra-se a cisteína catalítica responsável pelo ataque nucleofílico ao carbono carbonílico do aldeído a ser oxidado (Ahvazi, Coulombe et al. 2000) (Figura 4 B).

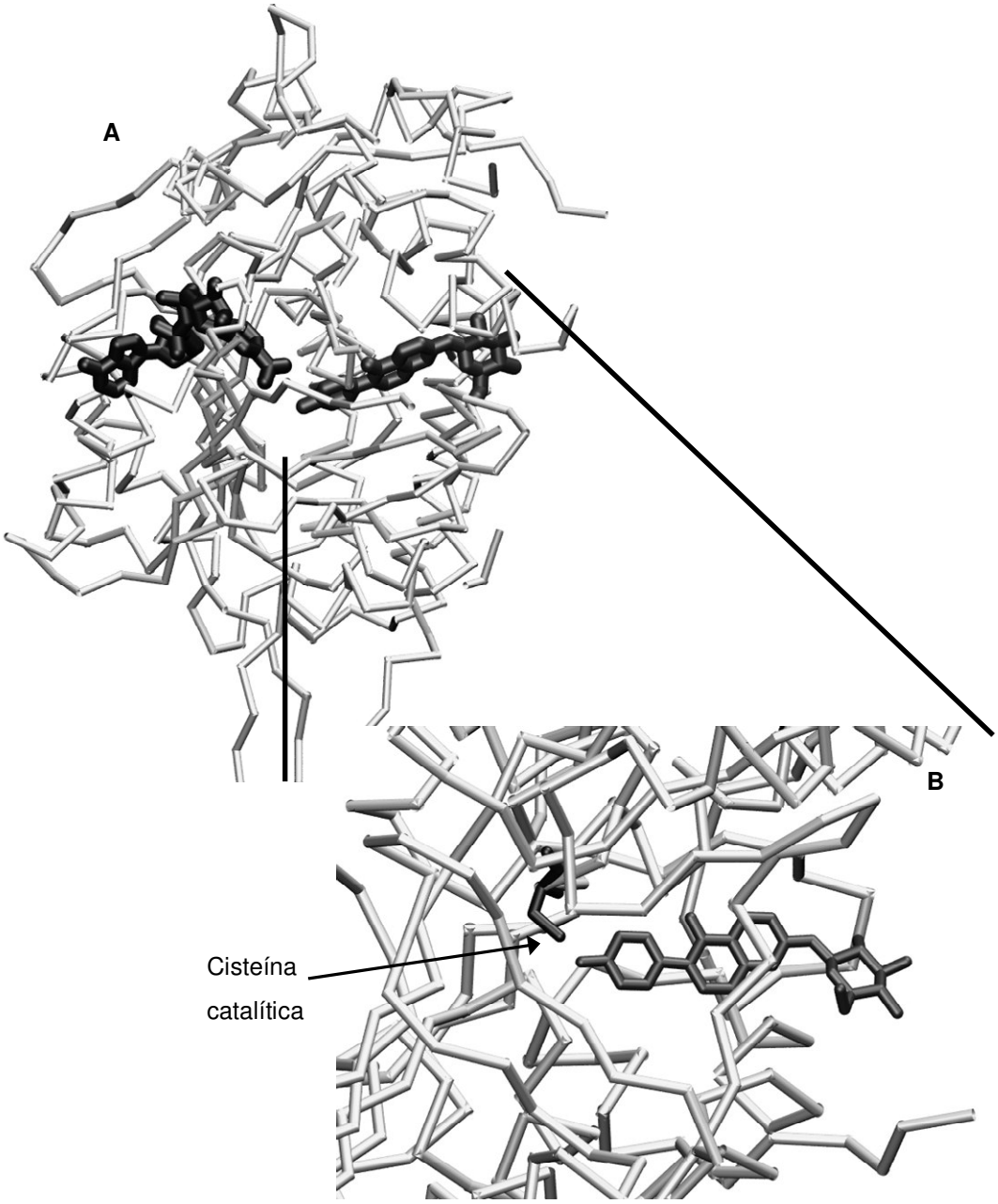


Figura 4: Canais de acesso ao sítio catalítico das ALDHs. Em A) uma visão geral de uma ALDH com uma molécula de Nicotinamida-adenina-dinucleotídeo (NAD) à esquerda e uma molécula de daidzina (ver sessão *2.5.1 Definição dos aminoácidos de interação do sítio de ligação*) no canal onde se aloja o aldeído a ser oxidado a direita. Em B), uma vista ampliada do canal onde se aloja o aldeído a ser oxidado com a daidzina e em destaque a cisteína catalítica.

1.4 Família das ALDHs

As ALDHs apresentam especificidades por diferentes grupos de substratos e também apresentam diferentes localizações celulares dependendo da função. A Tabela 1 tem o resumo dessas informações (Yoshida, Rzhetsky et al. 1998).

Tabela 1: Localização e principal substrato das ALDHs (baseado em Yoshida, 1998).

Proteína	Posição celular	Localização da expressão	Substrato
ALDH1	Citoplasmática	Estômago, fígado, etc.	All-trans e 9-cis retinal
ALDH2	Mitocondrial	Estômago, fígado, etc.	Acetaldeído
ALDH3	Citoplasmática	Estômago, pulmão, etc.	Aldeídos aromáticos de lipossolúveis
ALDH4	Mitocondrial	Fígado e rins	Glutamato
ALDH5	Mitocondrial	Testículos e fígado	Propionaldeido
ALDH6	Citoplasmática	Glândulas salivares, estômago e rins	Aldeídos alifáticos e retinal
ALDH7	Microssomos	Rins e pulmão	Aldeídos aromáticos e alifáticos
ALDH8	Microssomos	Parótida	Não definido
ALDH9	Citoplasmática	Fígado, rins e músculo	Amino aldeído
ALDH10	Microssomos	Fígado, rins e músculo	Aldeídos aromáticos de lipossolúveis

As enzimas ALDH já foram relacionadas à vários eventos biológicos, onde desempenham desde funções essenciais à vida, mas também levam morte quando ocorre uma mutação. Alguns desses eventos serão descritos abaixo para cada família de ALDH.

As ALDH1s são encontradas no citoplasma presentes em vários tecidos. Essa subfamília apresenta três enzimas: A1, A2 e A3. A ALDH1A1 tem alta atividade para

oxidar all-trans retinal e 9-cis retinal em AR, sendo encontrada na parte dorsal da retina. Inicialmente foi relacionada ao seu desenvolvimento, mas posteriormente verificou-se que camundongos ALDH1(-/-) apresentavam desenvolvimento normal sugerindo que as ALDH1A1s não são essenciais na produção de AR para a sinalização celular (Fan, Molotkov et al. 2003). A ALDH1A2 apresenta uma maior atividade de oxidação do all-trans retinal, sendo expressa nos adultos nos órgão reprodutivos e nos tecidos do tronco embrionário de ratos. Experimentos com ratos que não expressam ALDH1A2 durante o desenvolvimento embrionário mostraram que estes animais apresentavam defeitos no desenvolvimento cardíaco acarretando na morte (Vasiliou, Pappa et al. 2004). Outros experimentos com camundongos que mostram um alelo hipomórfico da ALDH1A2 também apresentam a morte prematura com aspectos da síndrome humana de DiGeorge que compreende defeitos no coração e anomalias no arco aórtico da cabeça e artérias do pescoço, defeitos nas cartilagens laríngeo-traqueal e timo/parótida aplasia ou hipoplasia (Vermot, Niederreither et al. 2003). A ALDH1A3 metaboliza apenas all-trans retinal e está relacionada na produção de AR essencial para a formação nasal. Camundongos que não expressam ALDH1A3 morrem logo após o nascimento causado por problemas respiratórios devido a atresia coanal (Dupe, Matt et al. 2003; Vasiliou, Pappa et al. 2004). A ALDH1 apresenta uma alta taxa de destoxificação em usuário de fármacos anti-câncer (oxazaphosphorines) (Vasiliou, Pappa et al. 2004). Está demonstrada que a resistência a esse fármaco está relacionado à expressão da ALDH1. A ALDH1 está presente em alta concentração nas lentes dos olhos e está relacionada com a destoxificação de peróxido de aldeído formado pela absorção de luz ultravioleta (Yoshida, 1998).

As ALDH2 são expressas nas mitocôndrias de vários tecidos, principalmente no fígado (Conklin 2006). A ALDH2 exibe uma alta atividade de oxidação de acetaldeído e é a

maior responsável na destoxificação deste. A sensibilidade clínica ao álcool, descrita pelo aumento da temperatura da pele, aumento da pulsação, enrubescimento facial entre outros sintomas, normalmente em pessoas de origem oriental, é associada a uma deficiência genética da ALDH2 causado por uma mutação pontual (E487K) (Yoshida 1994). Nenhuma deficiência no desenvolvimento ou fisiológica é observada em variantes homozigóticas para ALDH2, que não são funcionais, sugerindo que esta enzima não é essencial para a sobrevivência (Yoshida, Rzhetsky et al. 1998).

A ALDH3 é uma enzima citosólica que oxida aldeídos aromáticos e alifáticos. A ALDH3 é altamente expressa no estômago e pulmão, mas muito pouca expressa no fígado. A associação de resistência a oxazafosforina foi evidenciada por experimentos utilizando o cDNA de ALDH3 de rato ou humano transfectado em células de hamster (Yoshida, Rzhetsky et al. 1998). A ALDH3 é a maior constituinte da córnea humana e de outros animais e como a ALDH1 apresenta uma importante função na destoxificação de peróxido aldeído (Yoshida, 1998; Conklin 2006).

A ALDH4 é uma enzima mitocondrial com uma alta atividade de oxidação de γ-semialdeído como o glutâmico γ-semialdeído. A deficiência genética na ALDH4 interfere na via de degradação de prolina causando a hiperprolemia do tipo II, fazendo com que o portador apresente retardo metal e convulsões (Yoshida, 1998).

A ALDH9 é uma enzima citosólica com alta atividade de oxidação do 4-amino butiraldeído. O gene da ALDH9 é pouco expresso no cérebro adulto, mas altamente expresso nos primeiros estágios de desenvolvimento de cérebro embrionário (menos de 12 semanas de gestação) (Yoshida, Rzhetsky et al. 1998).

A ALDH10 é uma enzima microssomal com alta atividade de oxidação de aldeídos alifáticos. Deficiências genéticas na ALDH10 podem causar a síndrome de Sjögren-Larsson que causa uma alteração no metabolismo de membranas lipídicas, acarretando problemas neurológicos (Yoshida, Rzhetsky et al. 1998).

1.5 Ácido retinóico e desenvolvimento cardíaco

A via de sinalização celular por ácido retinóico está relacionada a vários eventos biológicos. Uma importante função desempenhada pelas ALDH é relacionada ao desenvolvimento cardíaco. Genes pertencentes à via metabólica e de sinalização do ácido retinóico constituem alvos interessantes à investigação como causadores diretos ou modificadores de Doenças Cardíacas Congênitas (DCCs). O AR é importante na padronização dos eixos do embrião e na organogênese em geral. A formação do coração é particularmente afetada pelas variações na sinalização do ácido retinóico onde a falta ou o excesso de vitamina A produz malformações vasculares e cardíacas, sugerindo uma relação direta entre o estado nutricional e a teratogênese. Wilson e Barch (Wilson and Barch 1949) observavam que a privação de vitamina A durante a gestação causava malformações em vários órgãos do embrião. Em 1950, Wilson e Warkany (Wilson and Warkany 1950) demonstraram que a falta de ácido retinóico causava malformação cardíacas congênitas, em animais–modelo, que eram similares àquelas que ocorriam espontaneamente no homem. Sawin e Edmonds (Sawin and Edmonds 1949) demonstraram que fatores genéticos exerciam papel na determinação do arco aórtico em coelhos. Estes fatos ilustravam que o desenvolvimento embrionário poderia ser alterado tanto por fatores ambientais quanto por fatores genéticos. Dersch e Zile (Dersch and Zile 1993) identificaram os períodos no

desenvolvimento embrionário no qual a vitamina A (ácido retinóico) é necessária para o desenvolvimento cardíaco e também que o isômero *all-trans* do ácido retinóico é a forma biologicamente ativa necessária para o desenvolvimento cardíaco embrionário e que, além disso, podia-se recuperar o fenótipo cardíaco normal em embriões deficientes em vitamina A com a administração de ácido retinóico.

Em 1998, Moss (Moss, Xavier-Neto et al. 1998) e colaboradores demonstraram que:

1. Os padrões de expressão do gene ALDH1A2 coincidiam com os locais no coração onde o ácido retinóico era produzido.
2. A progressão da sinalização do AR durante o desenvolvimento cardíaco se dava de maneira dinâmica, de forma que a ação do AR progride caudorostralmente no tubo cardíaco linear do camundongo, estando completamente ausente das porções anteriores do coração.

Estudos com animais transgênicos que possuíam o gene da fosfatase alcalina humana dirigido por um fragmento do promotor do gene *Slow Myosin Heavy-Chain 3* (SMyHC3) de codorna (Xavier-Neto *et al.*, 1999), mostraram que a expressão se sobrepunha aos locais de síntese e resposta ao ácido retinóico. Outros estudos sugerem que a expressão gênica atrial-específica seria controlada pela síntese localizada do ácido retinóico e que a exclusão do AR dos precursores ventriculares é essencial para a correta especificação dos ventrículos (Rosenthal and Xavier-Neto 2000; Xavier-Neto, Rosenthal et al. 2001). Para o entendimento de como e quando o AR afeta os precursores cardíacos, foram realizados experimentos examinando os padrões de expressão do gene ALDH1A2 em galinhas e em camundongoa na mesoderme lateral do embrião em desenvolvimento. A sinalização anteroposterior pelo ácido retinóico, analisada a partir da expressão de ALDH1A2, é caracterizada por um envolvimento progressivo dos precursores cardíacos

por um campo de ALDH1A2 originado na forma de uma onda caudorostral dinâmica e evolucionariamente conservada, na mesoderme lateral (Hochgreb, Linhares et al. 2003; Simoes-Costa, Vasconcelos et al. 2005). Estudos em embriões de camundongos *knockout* para o gene ALDH1A2 mostraram que, além de ser essencial para a sobrevivência do embrião, também é essencial para a morfogênese precoce apresentando malformações cardíacas como a formação de uma única cavidade anterior dilatada, ausência do dobramento cardíaco e diferenciação inapropriada dos cardiomiócitos ventriculares. A administração materna de ácido retinóico resgata a morfogênese cardíaca normal mostrando que esse defeitos resultam do bloqueio da síntese de ácido retinóico (Niederreither, Subbarayan et al. 1999; Niederreither, Vermot et al. 2001). Em experimentos utilizando-se animais que carregavam alelos hipomórficos do gene da ALDH1A2 e que também morriam *in útero*, foi possível demonstrar defeitos cardíacos que exibiam as características da síndrome de Di George humana, como defeitos de septação do trato de saída cardíaco e defeitos aórticos. Este estudo mostrou ainda que a diminuição da disponibilidade de ácido retinóico seja por intervenção genética ou nutricional poderia representar o principal modificador da expressão dos aspectos clínicos da síndrome de Di George (Vermot, Niederreither et al. 2003). Estudos em galinhas mostraram também que não só a privação de ácido retinóico, mas também o excesso de AR pode prejudicar a formação do embrião, levando a uma ampla gama de malformações, desde a completa ausência do coração até desenvolvimento incorreto do eixo anteroposterior (Osmond *et al.*, 1991). Os efeitos teratogênicos do excesso de ácido retinóico também foram observados em ratos, com grandes efeitos na ectoderme e na mesoderme (Morriss 1972).

2. Métodos

Esse trabalho multidisciplinar foi realizado em colaboração com vários grupos de diferentes especialidades. Serão descritos sucintamente os métodos empregados pelos grupos colaboradores e detalhadamente os métodos por nós empregados e desenvolvidos.

2.1 Conceito

A evolução das técnicas de seqüenciamento do DNA e a expansão dos projetos genomas possibilitaram aplicar a busca por seqüência homóloga a uma grande gama de organismos. As técnicas de seqüenciamento do DNA têm utilidade em diferentes campos da ciência, como arqueologia, antropologia, genética, biotecnologia, ciência forense entre outras (Franca, Carrilho et al. 2002). Inicialmente, o seqüenciamento do DNA permitia desvendar apenas as seqüências de bases nucléicas de pequenos fragmentos do DNA. Essas técnicas evoluíram permitindo seqüenciar fragmentos maiores e em maior quantidade. Esses avanços permitiram a realização dos projetos genomas, que têm como objetivo obter todas as bases nucléicas que compõe o genoma de um organismo. No início do ano de 2008, haviam 702 organismos com o genoma totalmente seqüenciado, 639 em fase de finalização e 679 em progresso (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html, Entrez genome project). Outra forma de obter informações contidas no genoma de um organismo são os projetos de seqüenciamento de Etiquetas de Seqüências Expressas (EST, do inglês *Expressed Sequenced Tags*), que consiste em seqüenciar apenas as regiões do genoma que são transcritas. No início do ano de 2008, existiam 843 organismos com 1000 ou mais seqüências disponíveis (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html, dbEST).

Apesar dos métodos de seqüenciamento em larga escala terem evoluído muito nos últimos anos, eles ainda apresentam limitações. Atualmente os métodos mais utilizados (Maxam and Gilbert 1977; Sanger, Nicklen et al. 1977) conseguem seqüenciar trechos de DNA de até 1000 bases. Desta forma, para seqüenciar cadeias mais longas ou um genoma inteiro, o DNA precisa ser fragmentado, manipulado e finalmente cada fragmento ser seqüenciado individualmente (Figura 5 A). Após a obtenção de vários fragmentos da seqüência, são utilizados programas de computador que tentam remontar a seqüência do genoma completo do organismo de interesse (Figura 5 B).

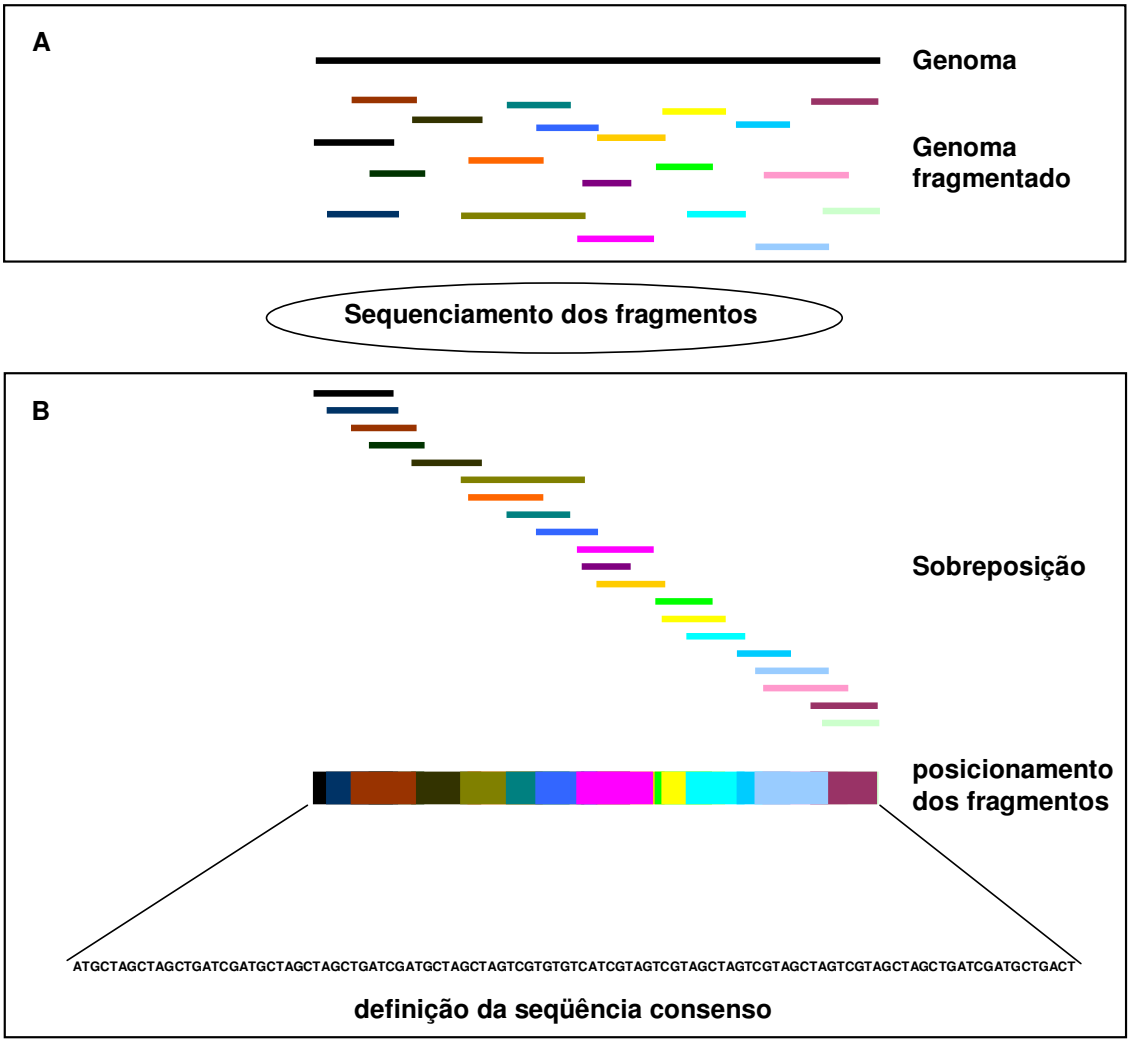


Figura 5: Esquema do processo de seqüenciamento de DNA de um genoma. Em A, temos a etapa de processamento, fragmentação e seqüenciamento do DNA. Em B, temos o processo de sobreposição, posicionamento dos fragmentos e definição da seqüência consenso das seqüências dos fragmentos obtidos na etapa A.

Os programas de montagem de seqüências, mais empregados atualmente, utilizam o método "sobreposição, posicionamento dos fragmentos e definição da seqüência consenso" (em inglês "overlap-layout-consensus") (Figura 5 B). Esse método consiste em alinhar todas as seqüências contra todas as seqüências (sobreposição), determinar o início e o fim das posições de sobreposição (posicionamento dos fragmentos) e definir a seqüência

consenso baseado na sobreposição dos fragmentos. Esse método é muito útil para pequenos laboratórios sem grandes estruturas computacionais, pois, dependendo da quantidade de seqüências a serem montadas, acarreta a uma alta utilização de memória e um alto custo computacional, não sendo viável utilizá-lo na montagem de grande quantidade de seqüências. A montagem de grandes projetos genomas normalmente são realizados por grandes centros de pesquisa, que possuem computadores de alto desempenho e programas específicos como, por exemplo, TIGR assembler (Pertea, Huang et al. 2003).

Com a simplificação e a diminuição de custos do processo de seqüenciamento de DNA, muitas informações biológicas estão sendo geradas, mas essas informações não estão sendo processadas com suficiente rapidez, de forma que vários organismos com o genoma parcialmente ou totalmente seqüenciados possuem poucos genes anotados. Assim, para realizar a busca de algum gene em algum desses organismos é necessário utilizar as informações pré-eliminares disponibilizadas pelos projetos de seqüenciamento. Para os organismos com o genoma já montado, uma busca de similaridade utilizando um gene homologo já possibilita a recuperação da seqüência do gene de interesse. Mas quando apenas as seqüências obtidas estão disponíveis, é necessário montar seqüências para obter o gene de interesse, o que é inviável quando se quer estudar um gene em vários organismos com as seqüências de DNA disponíveis, mas não montadas. Uma alternativa é remontar apenas a região de interesse. Existem dois programas que utilizam dessas informações para a reconstrução de seqüências específicas, *Genotrace* e *Tracembler* (Berezikov, Plasterk et al. 2002; Dong, Wilkerson et al. 2007). Os dois programas partem de uma seqüência semente e, utilizando as informações do *trace archive* (http://www.ncbi.nlm.nih.gov/trace), reconstroem parcialmente ou totalmente a seqüência de interesse sem a necessidade de montar todo o genoma do organismo. Mas esses programas apresentaram vários problemas

na sua utilização, não reconstruindo as seqüências de interesse e também apresentam uma dificuldade quando são utilizadas bases de seqüência diferente do *trace archive*. Assim, a utilização desses programas fica muito limitada impossibilitando a utilização deles nesse trabalho. Sobreira e Gruber (Sobreira and Gruber 2008) desenvolveram um programa que possibilita a montagem de regiões específicas partindo de uma seqüência semente, podendo ser uma seqüência homóloga ao gene de interesse. O método consiste em sucessíveis passos de busca por similaridade contra uma base de seqüência do organismo de interesse, recuperação das seqüências de interesse e montagem das seqüências. Esse método possibilita a reconstrução da região de interesse utilizando as informações provindas de projetos genomas.

2.2 Obtenção das seqüências de ALDH

As seqüências protéicas de ALDH dos organismos já montados e/ou anotados foram obtidas dos repositórios de internet http://www.aldh.org, NCBI Entrez (http://www.ncbi.nlm.nih.gov/sites/gquery), UCSC (http://genome.ucsc.edu). No repositório Trace Archive(http://www.ncbi.nlm.nih.gov/Traces/), foram obtidos os dados genômicos utilizados como base de dados para busca de similaridade utilizando o programa BLAST(Altschul, Madden et al. 1997).

A busca de ALDHs em genomas não montados foi realizada utilizando-se uma versão modificada do programa GenSeed (Sobreira and Gruber 2008). No método desenvolvido (Figura 6), as seqüências de aminoácidos das 19 proteínas da família da ALDH humana (Tabela 2), obtidas no www.aldh.org, foram utilizadas para realizar a busca de similaridade na base de dados de seqüências do organismo de interesse através do

algoritmo tblastn implementado no programa Blast. A partir das informações geradas pelo programa Blast, foram obtidos os nomes das seqüências do organismo de interesse que se alinhavam a algum trecho de alguma ALDH humana. Nesta etapa, foram selecionadas seqüências que apresentavam um *valor de esperança* menor que 10^{-6}, com o tamanho mínimo do bloco de alinhamento superior a 100 pares de bases e a identidade do bloco de alinhamento superior a 60%. As seqüências com as características pré-definidas eram obtidas da base de seqüências do organismo de interesse utilizando o programa fastacmd (do pacote do Blast) e montadas utilizando o programa CAP3 (Huang and Madan 1999). As seqüências utilizadas na montagem dos contig(s) formados pelo programa CAP3 eram armazenadas em uma lista (aqui chamada Lista Final). Após a montagem dos contigs, era realizado um novo ciclo de busca de seqüências homólogas as ALDHs (Figura 6 A), mas ao invés de utilizar as seqüências de ALDHs humana como entrada para o programa Blast eram utilizadas as seqüências dos contigs obtidos. A partir desta etapa, foi utilizado o algoritmo de busca blastn implementado em Blast. Esse ciclo de busca por seqüências era repetido até que nenhuma nova seqüência da base de dados fosse incorporada à Lista Final (Figura 6 A). Na etapa seguinte foi realizada a montagem de todas as seqüências incluídas na Lista Final utilizando o programa CAP3.

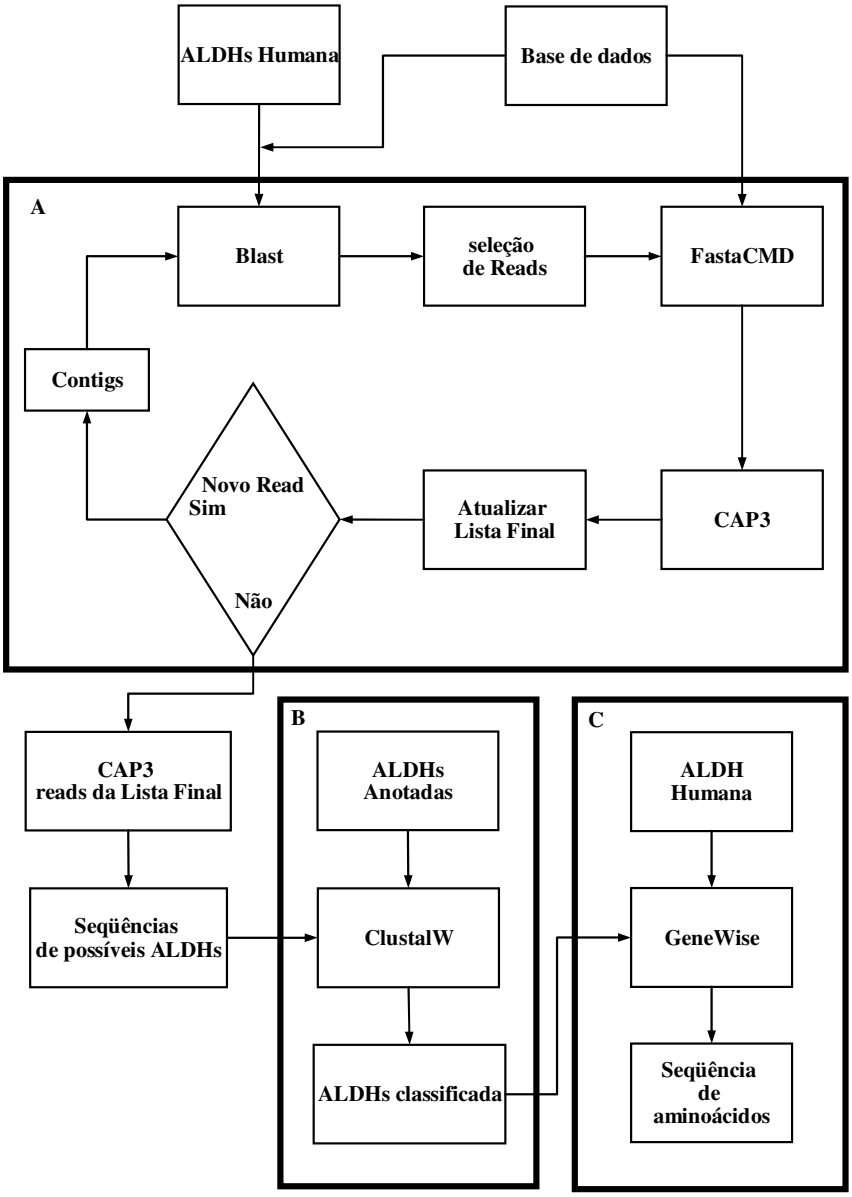


Figura 6: Esquema do método desenvolvido para busca de ALDH em organismos não anotados completamente. (A) Ciclo principal de busca. (B) Alinhamento múltiplo das seqüências obtidas para o organismo-alvo. (C) Definição da seqüência de aminoácidos.

Tabela 2: Lista das ALDHs humanas utilizadas na busca de similaridade contra a base de seqüência do organismo de interesse.

Nome	ref_seq DNA	ref_seq Proteína
ALDH1A1	NM_000689.3	NP_000680.2
ALDH1A2	NM_003888.2	NP_003879.2
ALDH1A3	NM_000693.2	NP_000684,2
ALDH1B1	NM_000692.3	NP_000683.3
ALDH1L1	NM_012190.2	NP_036322.2
ALDH1L2	NM_001034173.2	NP_001029345.1
ALDH2	NM_000690.2	NP_000681.2
ALDH3A1	NM_000691.3	NP_000682.3
ALDH3A2	NM_001031806.1	NP_001026976.1
ALDH3B1	NM_001030010.1	NP_001025181.1
ALDH3B2	NM_001031615.1	NP_001026786.1
ALDH4A1	NM_170726.1	NP_003739.2
ALDH5A1	NM_170740.1	NP_733936.1
ALDH6A1	NM_005589.2	NP_005580.1
ALDH7A1	NM_001182.3	NP_001173.2
ALDH8A1	NM_170771.1	NP_739577.1
ALDH9A1	NM_000696.3	NP_000687.3
ALDH16A1	NM_153329.2	NP_699160.1
ALDH18A1	NM_001017423.1	NP_001017423.1

Os contigs formados com mais de 900 pares de bases foram alinhados individualmente utilizando o programa ClustalW (Thompson, Higgins et al. 1994) (Figura 6 B), ao conjunto de seqüência de nucleotídeos de ALDHs humanas já alinhadas (Figura 7) Após o alinhamento de cada seqüência, era gerada uma distribuição de distância entre as ALDHs. A ALDH obtida pelo método era definida como sendo da classe de ALDH na qual apresentava menor distância. Essa distância é obtida a partir do algoritmo de junção de vizinhos para construção de árvores filogenética. Como exemplo, a Tabela 3 mostra as distâncias entre as seqüências de ALDHs e uma seqüência-problema aqui chamada seqüência 2559. A seqüência 2559 é uma ALDH obtida pelo método que apresenta menor distância com a seqüência de ALDH2, como qual foi classificada.

Tabela 3: Tabela com exemplos das distâncias entre as seqüências de ALDHs. A seqüência 2259 é uma ALDH obtida pelo método de busca e classificada como ALDH2 (em destaque).

	ALDH3A1	ALDH3A2	ALDH3B1	ALDH3B2	ALDH1A1	ALDH1A2	ALDH1B1	ALDH2	ALDH1A3	ALDH1L1	ALDH1L2	ALDH9A1	ALDH8A1	ALDH5A1	ALDH7A1	ALDH6A1	ALDH16A1	ALDH4A1	ALDH18A1	2559
ALDH3A1	0	0,4	0,4	0,5	0,7	0,7	0,6	0,6	0,6	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,6	0,7	0,7	0,7
ALDH3A2	0,4	0	0,6	0,6	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7
ALDH3B1	0,4	0,6	0	0,3	0,7	0,7	0,7	0,6	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7
ALDH3B2	0,5	0,6	0,3	0	0,7	0,7	0,7	0,6	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7
ALDH1A1	0,7	0,7	0,7	0,7	0	0,4	0,5	0,4	0,4	0,7	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,7	0,7
ALDH1A2	0,7	0,7	0,7	0,7	0,4	0	0,5	0,4	0,5	0,7	0,7	0,6	0,6	0,6	0,7	0,6	0,7	0,7	0,7	0,7
ALDH1B1	0,6	0,7	0,7	0,7	0,5	0,5	0	0,4	0,5	0,7	0,7	0,6	0,6	0,7	0,7	0,6	0,6	0,6	0,7	0,7
ALDH2	0,6	0,7	0,6	0,6	0,4	0,4	0,4	0	0,4	0,7	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,6	0,7	0,7
ALDH1A3	0,6	0,7	0,7	0,7	0,4	0,5	0,5	0,4	0	0,7	0,7	0,6	0,6	0,7	0,7	0,6	0,7	0,6	0,7	0,7
ALDH1L1	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0	0,3	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7
ALDH1L2	0,7	0,7	0,7	0,7	0,6	0,7	0,7	0,6	0,7	0,3	0	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7
ALDH9A1	0,7	0,7	0,7	0,7	0,6	0,6	0,6	0,6	0,6	0,7	0,7	0	0,7	0,6	0,7	0,6	0,7	0,7	0,7	0,7
ALDH8A1	0,7	0,7	0,7	0,7	0,6	0,6	0,6	0,6	0,6	0,7	0,7	0,7	0	0,7	0,7	0,7	0,7	0,7	0,7	0,7
ALDH5A1	0,7	0,7	0,7	0,7	0,6	0,6	0,7	0,6	0,7	0,7	0,7	0,6	0,7	0	0,7	0,6	0,7	0,7	0,7	0,7
ALDH7A1	0,7	0,7	0,7	0,7	0,6	0,7	0,7	0,6	0,7	0,7	0,7	0,7	0,7	0,7	0	0,7	0,7	0,7	0,7	0,7
ALDH6A1	0,7	0,7	0,7	0,7	0,6	0,6	0,6	0,6	0,6	0,7	0,7	0,6	0,7	0,6	0,7	0	0,7	0,7	0,7	0,7
ALDH16A1	0,6	0,7	0,7	0,7	0,6	0,7	0,6	0,6	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0	0,7	0,7	0,7
ALDH4A1	0,7	0,7	0,7	0,7	0,6	0,7	0,6	0,6	0,6	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0	0,7	0,7
ALDH18A1	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0,7	0	0,7
2559	0,68	0,7	0,68	0,69	0,68	0,68	0,66	**0,65**	0,68	0,72	0,71	0,67	0,69	0,71	0,7	0,71	0,7	0,74	0,71	0

As seqüências classificadas foram analisadas utilizando o programa Genewise (Birney, Clamp et al. 2004) para obter a seqüência de aminoácidos (Figura 6 C). A seqüência de aminoácidos da ALDH humana a qual foi classificada foi utilizada como proteína de referência para definir a provável seqüência de aminoácidos de cada seqüência-problema.

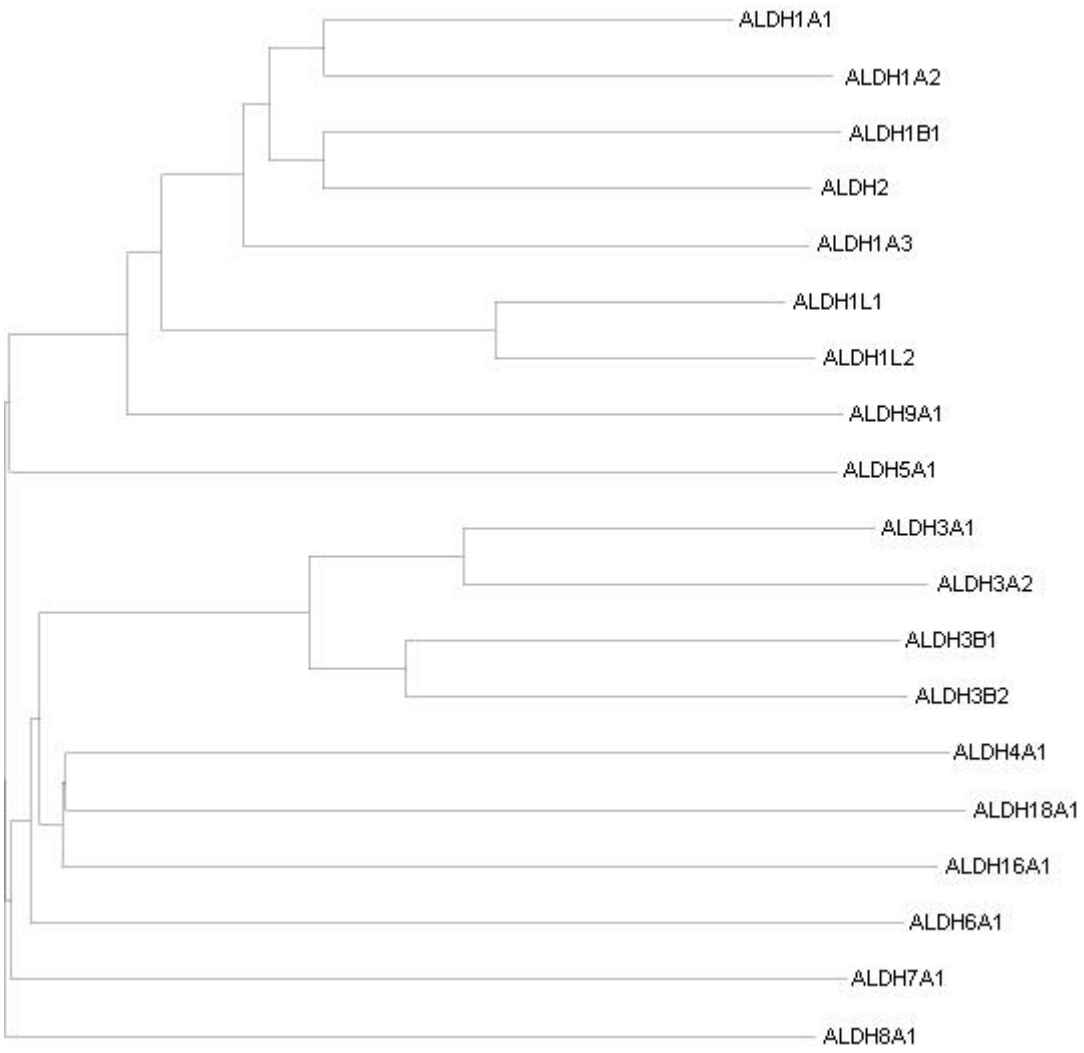


Figura 7: Árvore filogenética baseada no alinhamento das ALDHs humanas utilizadas para classificar as seqüências das ALDHs obtidas.

O método foi aplicado para cada organismo com mais de 1000 seqüências na base de dados de etiquetas de seqüência expressas (EST, do inglês expressed sequence tag) dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) totalizando 498 espécies.

Os resultados foram inseridos em uma base de dados relacional para simplificar a construção de códigos de busca por ALDH específicas (Figura 8). Foram criadas duas

tabelas específicas: uma com as informações referentes às seqüências obtidas e outra com as informações referente à taxonomia dos organismos analisados.

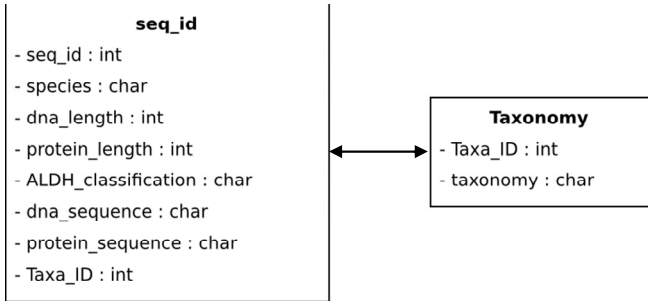


Figura 8: Esquema de entidades de relacionamento criadas para armazenar as informações dos resultados de busca por seqüência de ALDHs.

A relação *Sequence* tem os seguintes atributos: *Seq_id* (Chave primaria com o índex de cada seqüência obtida), *dna_length* (Quantidade de bases nucléica da seqüência obtida pelo método de busca), *protein_length* (Quantidade de aminoácido da seqüência obtida a partir da análise do programa Genewise) *ALDH_classification* (Classificação da seqüência a partir da distância entre ela e a seqüência de ALDH humana), *dna_sequence* (Seqüência de bases nucléicas obtida pelo método de busca), *protein_sequence* (Seqüência de aminoácidos obtida pelo método de busca), *Species* (Nome científico da espécie), *Taxa_ID* (Número de identificação dado a um organismo pelo NCBI).

A tabela *Taxonomy* possui os seguintes campos: *Taxa_ID* (Número de identificação, dado pelo NCBI, utilizado para de identificar uma espécie), *taxonomy* (Todas as classificações do táxon desde reino até o gênero. (Exemplo para o *Homo sapiens* temos: *cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi;*

Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homo/Pan/Gorilla group; Homo).

A tabela Taxonomy foi utilizada para obter os Taxa_ID das espécies de grupo de organismos, como por exemplo cefalocordados.

2.3 Alinhamento múltiplo das seqüências das ALDHs para análise filogenética

As seqüências de ALDH foram alinhadas utilizando o programa MUSCLE (Edgar 2004) seguido de refinamento manual, quando foram retirados os buracos do alinhamento e regiões exclusivas de apenas uma ALDH. As árvores foram calculadas utilizando o programa MrBayes e PhyML (Ronquist and Huelsenbeck 2003) baseado no melhor modelo WAG+I+Γ4 predito por ProtTest (Abascal, Zardoya et al. 2005). Duas rodadas de 5.000.000 de gerações foram realizadas para cada árvore. A convergência era verificada e o período de *burnin* era determinado pelo gráfico do logaritmo da verossimilhança do modelo versus tempo. As árvores consenso e a probabilidade posterior foram calculadas utilizando a regra da maioria de 50%. Essa etapa foi realizada por Frédéric Brunet e Michael Schubert do Institut de Génomique Fonctionnelle de Lyon, Université de Lyon, França.

2.4 Busca por regiões específica (assinatura)

As assinaturas são pequenas seqüências de aminoácidos que podem ser utilizadas para diferenciar uma família de proteínas de outra. As assinaturas de ALDH1A e ALDH2 foram obtidas pelo alinhamento das ALDH1A2 e ALDH2 humanas e depois confirmadas

no alinhamento contendo todas as seqüências de ALDH1A e ALDH2 de vertebrados utilizando o programa Multialin (Corpet 1988). As representações gráficas das freqüências de aminoácidos das assinaturas foram feitas utilizando o programa Weblogo (http://weblogo.berkeley.edu) (Anexo 3). Para simplificar as comparações, os resíduos de aminoácidos foram numerados utilizando como referência a ALDH2 humana onde a cisteína catalítica está localizada na posição 302 (Moore, Baker et al. 1998). Essa etapa foi realizada pelos Drs. José Xavier Neto e Deborah Schechtman.

2.5 Obtenção das estruturas cristalográficas das ALDH e modelagem molecular comparativa

As estruturas tridimensionais das proteínas das famílias ALDH1 e ALDH2 representadas pelas seqüências obtidas nos passos anteriores foram modeladas utilizando como molde as estruturas cristalográficas da ALDH1A1 de ovelha (PDBid 1BXS) ou da ALDH2 humana (PDBid 1OF7). Cada seqüência de ALDH obtida para um organismo de interesse foi alinhada globalmente à seqüência da estrutura molde utilizando o programa ClustalW. Este alinhamento foi utilizado como entrada para o programa de modelagem comparativa Nest do pacote Jackal (Petrey, Xiang et al. 2003). Foram utilizados parâmetros para o refinamento máximo de todas as alças e região de estruturas secundária. Os critérios de escolha de qual estrutura seria utilizada como molde para cada seqüência de ALDH de interesse ser modelada foram:

1) Uso da identidade total entre as seqüência da ALDH de interesse com as seqüências dos moldes e;

2) Em caso de igualdade do primeiro critério, foi usada a identidade dos aminoácidos da região do sítio de ligação (ver seção 2.5.1 Definição dos aminoácidos de

interação do sítio de ligação). A qualidade estereoquímica dos modelos tridimensionais gerados foi avaliada utilizando o programa Procheck (Laskowski, Moss et al. 1993).

2.5.1 Definição dos aminoácidos de interação do sítio de ligação

Na definição dos aminoácidos que interagem na região do sítio de ligação das ALDHs, foram utilizados como referência de posicionamento os átomos da molécula daidzina (Figura 9), que foi cristalizada na região do sítio de ligação da ALDH2 (PDBid 1OF7). Foram selecionados todos os átomos de cadeia lateral da proteína, distribuídos a uma distância máxima de 5 Å de qualquer átomo do composto de referência (Figura 9). Os aminoácidos que apresentaram ao menos um átomo a menos de 5 Å de distância foram considerados como um resíduo da região do sítio de ligação. Para definir os resíduos do sítio de ligação para a estrutura da ALDH1 (PDBid 1BXS) foi realizada a sobreposição estrutural da ALDH1 sobre a ALDH2 e posteriormente removidos os átomos de ALDH2, permitindo o posicionamento de daidzina. Na definição dos aminoácidos que interagem com região do sítio de ligação foi aplicado o mesmo critério utilizado para a enzima ALDH2.

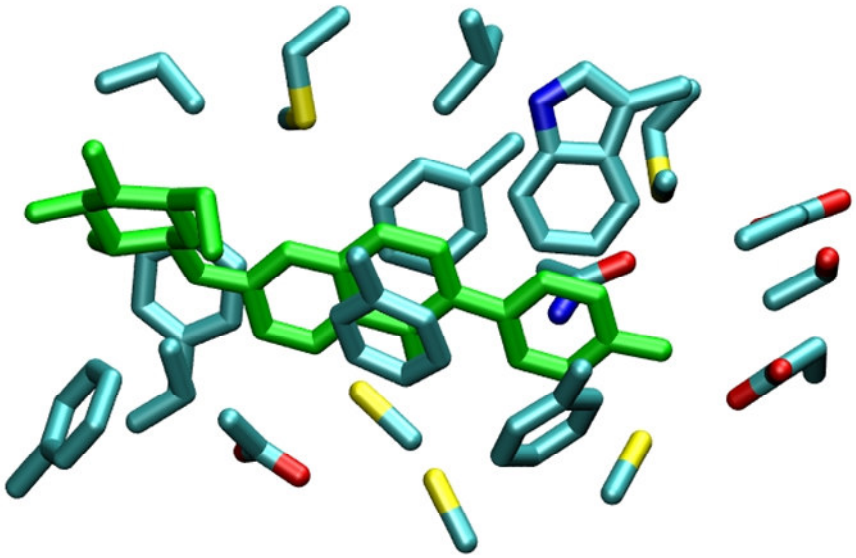


Figura 9: Resíduos da estrutura cristalográfica de PDBid 1OF7, que possuem algum átomo da cadeia lateral a uma distância máxima de 5 Å da molécula de daidzina (verde).

2.6 Método para procura por cavidades a partir das estruturas tridimensionais e cálculo do volume

Nesta seção, descreveremos os métodos aplicados para definir e obter as informações sobre o canal onde o aldeído se insere para ser oxidado.

2.6.1 Utilização dos programas disponíveis

Para definir a região do canal de acesso ao substrato das ALDHs, tentamos utilizar vários programas como: Voidoo (Kleywegt and Jones 1994), Pocket-Finder (Laurie and Jackson 2005), Hole (http://bach.bip.bham.ac.uk/osmart/hole/). Esses programas apresentaram algum dos seguintes problemas:

- O canal de acesso ao substrato em alguma ALDH apresentava-se unido ao canal de acesso do NAD(H), assim impossibilitando o cálculo do volume de apenas do canal de acesso ao substrato.
- O canal de acesso ao substrato era fragmentado em várias cavidades pequenas, assim gerando um erro no cálculo do volume.
- Nenhuma cavidade ou uma cavidade muito pequena era encontrada, não correspondendo a todo o canal de acesso ao substrato.

Como nenhum dos programas disponíveis possibilitou definir corretamente o canal de acesso ao substrato, foi necessário desenvolver um programa com esse objetivo. Foi utilizado o programa VMD (Humphrey, Dalke et al. 1996), que permite a utilização de scripts na linguagem TCL (http://www.tcl.tk/) e programação em PERL (http://www.perl.org/).

2.6.2 Definição da região de busca por cavidades

Para delimitar a região de procura ao canal de acesso ao substrato das ALDHs foram utilizadas as informações provindas da estrutura cristalográfica de PDBid 1OF7. Esta estrutura contém a ALDH2 humana co-cristalizada com a daidzina (Figura 10 A), uma molécula inibidora da atividade da ALDH2 utilizada no tratamento de pessoas dependentes de álcool (Lowe, Gao et al. 2008). A Daidzina interage com a ALDH2 humana ligando-se ao canal de acesso ao substrato (Figura 10 B) e impedindo a sua atividade. Para definir a região onde procurar o canal de acesso ao substrato em outras ALDHs, foi definida uma caixa em torno da daidzina que corresponde às coordenadas máximas e mínimas nos três

eixos (**x** , **y**, **z**), considerando as coordenadas de todos os aminoácidos a uma distância máxima de 5Å da daidzina (Figura 10 C e D). Essa primeira caixa serviu como referência para uma melhor delimitação da região de busca pelo canal. Uma nova caixa foi definida utilizando como referência para o fundo do canal a cisteína catalítica (resíduo 302) e como início do canal a histidina 145 (Figura 10 E e F).

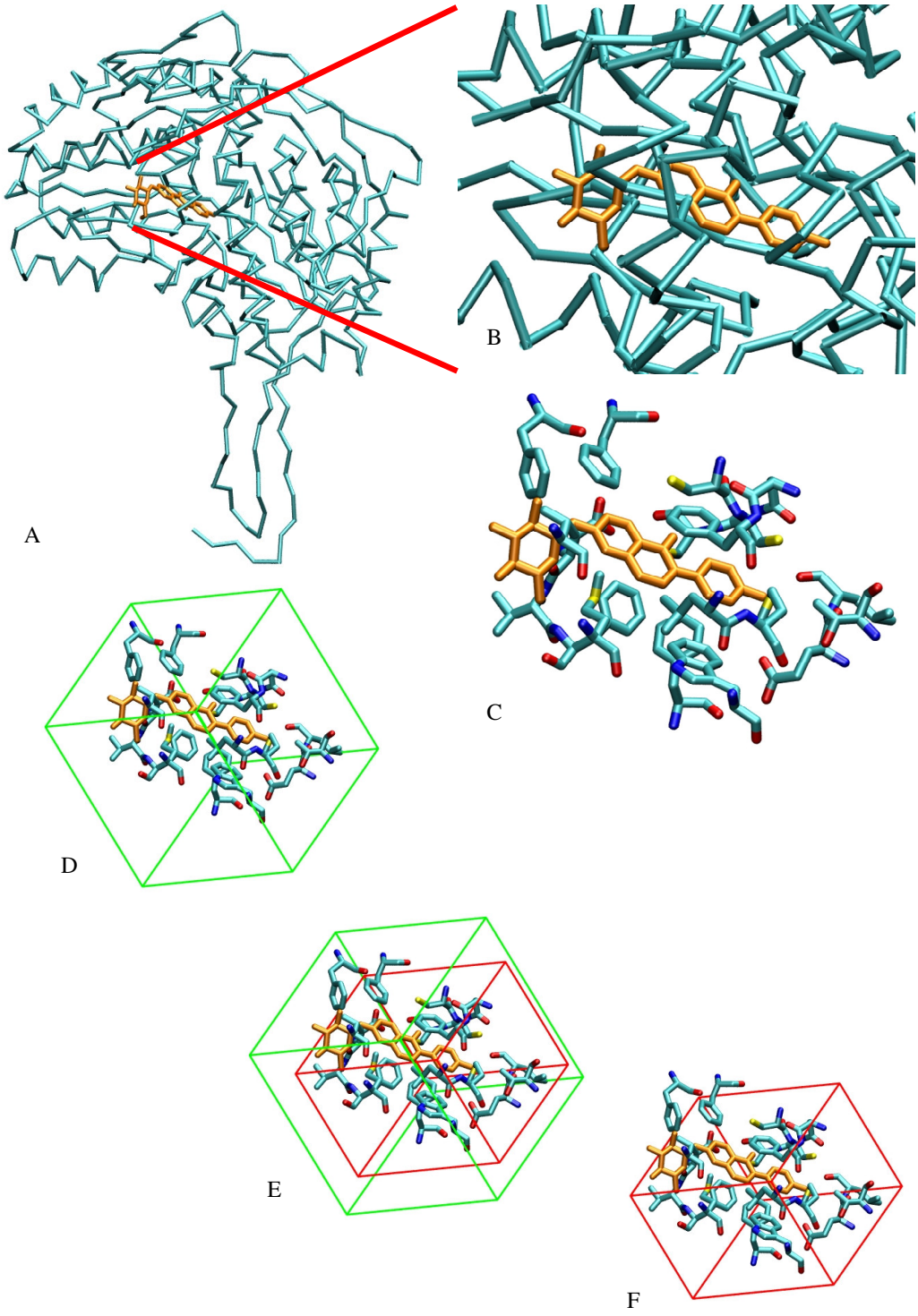


Figura 10: Etapas para definição da região de busca pelo canal de acesso ao substrato. Em A) a estrutura cristalográfica de PDBid 1OF7, em azul ALDH2 e em laranja a daidzina. Em B) uma vista aproximada da daidzina na região do canal de acesso ao substrato. Em C) a daidzina e os aminoácidos que estão a uma distância máxima de 5Å. Em D) a caixa de 5Å (verde) definindo a primeira região para a procura pelo canal de acesso ao substrato. Em E)

a caixa de 5Å (verde) e a caixa redefina (vermelha) que delimitam a região de busca pelo canal de acesso ao substrato. Em F) apenas a caixa, utilizada neste trabalho, que delimita a região para a busca pelo canal de acesso ao substrato.

2.6.3 Busca por cavidades

Após definida a caixa que delimita a região da proteína onde se tem interesse em procurar cavidades (Figura 11 A) era criado um GRID no interior desta caixa, onde a distância entre os pontos do GRID é definida pelo usuário. Nesse trabalho foi utilizado 0,8 Å (Figura 11 B).

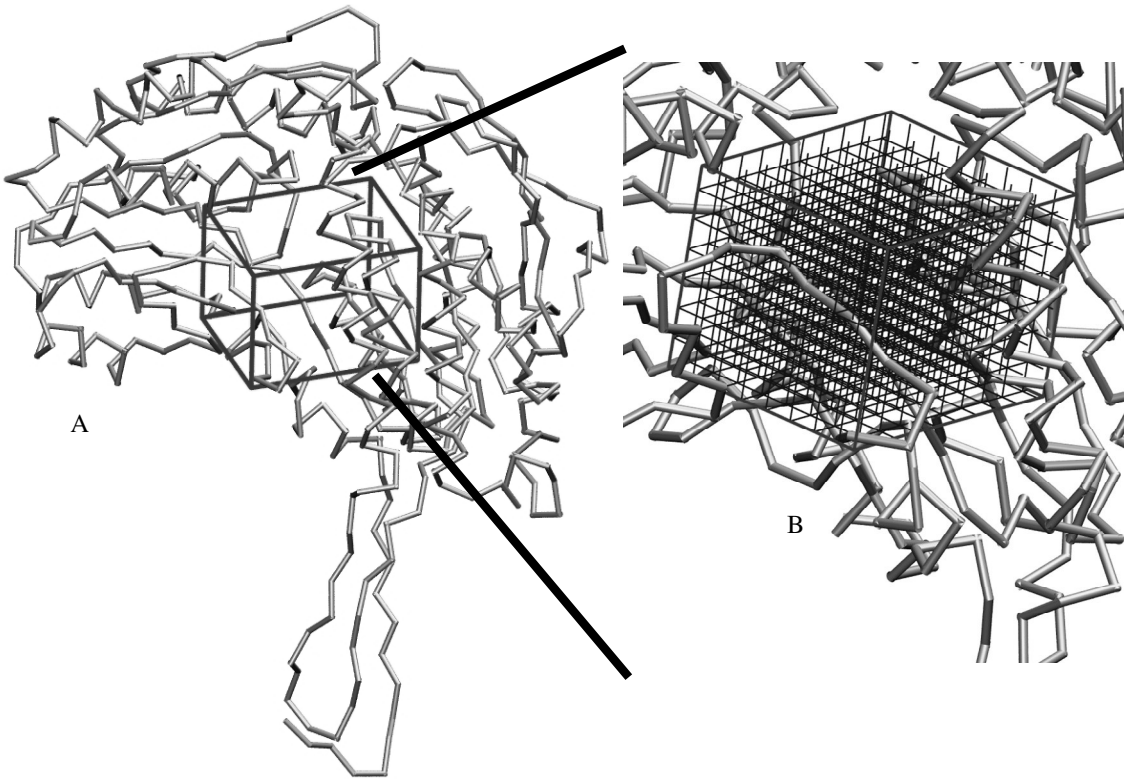


Figura 11: Em A) a caixa na região onde se tem interesse em procuras cavidades. Em B) a caixa com o GRID que era utilizado para o posicionamento do átomo de sondagem.

O processo de busca por cavidades se inicia colocando no primeiro ponto do GRID um átomo de hidrogênio e utilizando uma função implementada no programa VMD que fornece todos os átomos a uma distância mínima. Nesse trabalho foi utilizada a distância mínima de 2,8 Å que corresponde à distância média entre os núcleos dos átomos. Se a função do VMD retornar algum átomo, esse ponto não está dentro de uma cavidade, pois assim existe algum átomo com uma distância menor que 2,8 Å (Figura 12 C). Se a função não retornar nenhum átomo, esse ponto é marcado como pertencente a uma cavidade (Figura 12 C). Esse processo é realizado para todos os pontos do GRID. Após todos os pontos serem processados, obtemos uma lista de todos os pontos pertencentes a alguma cavidade com as suas coordenadas relativas no GRID (Figura 12 D).

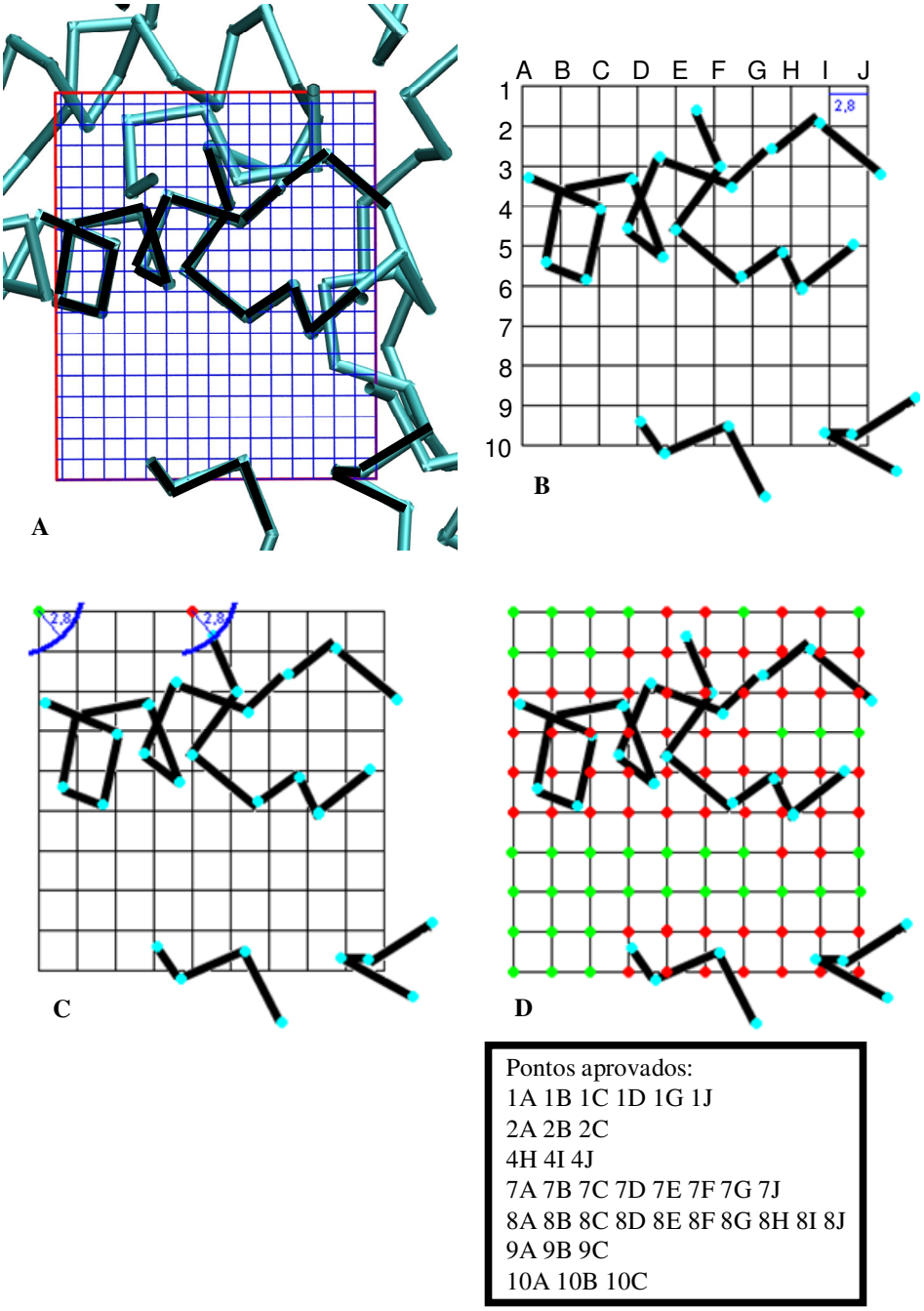


Figura 12: Etapas do processo para busca por cavidade. Em A) é mostrado um corte de um dos planos do GRID e em preto a cadeia de carbonos utilizada para esquematizar o processo de busca por cavidades. Em B) um esquema simplificado do desenho em (A), em azul claro a representação dos átomos de carbono e em preto a ligação entre os carbonos, a distância entre as linhas do GRID foi alterada para 2,8Å para simplificar a visualização. Em C) uma representação mostrando quando um ponto é considerado dentro de uma cavidade (ponto verde) e não é considerado em uma cavidade (ponto vermelho). Em D) todos os

pontos aceitos em verde e todos os pontos rejeitados em vermelho e a lista dos pontos aceitos na coordenada relativa do GRID.

2.6.4 Agrupamentos dos pontos

A partir da lista de pontos gerada na etapa anterior é realizado o agrupamento dos pontos pertencentes à mesma cavidade. Esse procedimento é realizado da seguinte forma, para o primeiro ponto da lista atribui-se a valor 0. Em seguida verifica se existe algum outro ponto adjacente a esse ponto (Figura 13 A). Existindo pontos, eles recebem o mesmo valor do primeiro ponto (Figura 13 B). Para todos os pontos adjacentes, é feita a mesma verificação. Se existirem pontos adjacentes a eles todos esses também recebem o valor do primeiro ponto. Quando não existir nenhum novo ponto adjacente ao grupo do primeiro ponto, passa-se para o segundo ponto da lista. Se esse ponto já tiver sido marcado como pertencente a algum grupo passa-se para o ponto seguinte. Se não tiver sido marcado é iniciado um novo grupo e realizado o mesmo procedimento do primeiro ponto (Figura 13 C). Esse procedimento é realizado para todos os pontos da lista. No final desse processo temos todos os pontos presentes dentro de uma cavidade agrupados pela adjacência (Figura 13 D).

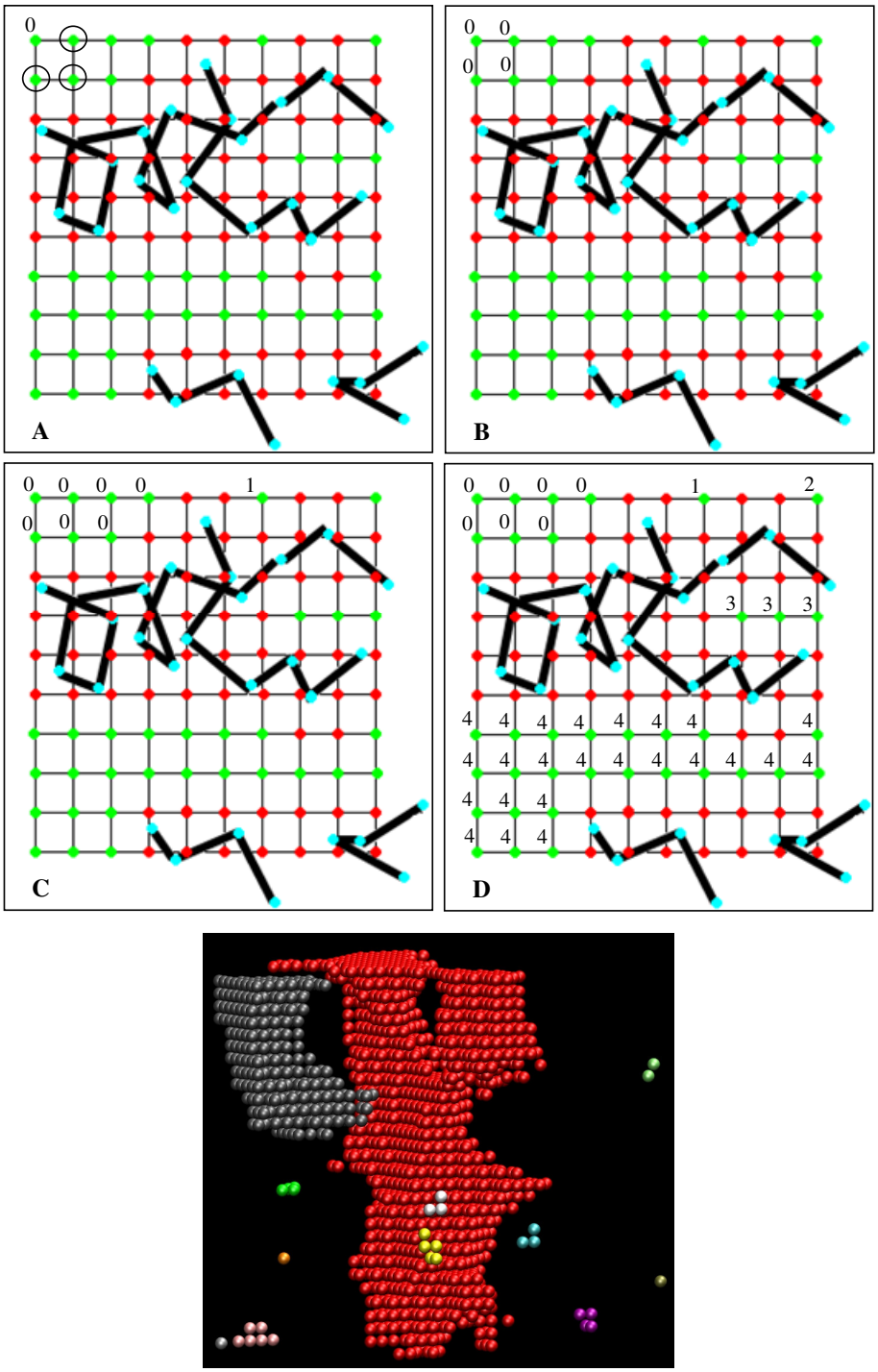


Figura 13: Representação do processo de agrupamento dos pontos simplificados para um sistema bidimensional. Em A, o início do processo no qual define o primeiro ponto com o valor 0 e procura por pontos adjacentes. Em B, os pontos adjacentes com o mesmo valor do primeiro ponto. Em C, todos os pontos adjacentes ao primeiro ponto definido com o mesmo valor e um segundo ponto definido com o valor 1. Em D, todos os pontos com os seus

grupos definidos. Em E, um exemplo do resultado do método de agrupamento de pontos, onde cada cor representa um grupo formado.

2.6.5 Cálculo do volume

Para calcular o volume das cavidades são utilizados todos os pontos do mesmo grupo definidos na etapa anterior. O cálculo do volume é realizado considerando duas situações: uma quando a partir de um ponto é possível formar um cubo (Figura 14 A) ou quando a partir de um ponto não é possível formar um cubo, mas formar tetraedros (Figura 14 B). O segundo caso foi implementado para diminuir o erro do cálculo do volume, pois permite calcular valores parciais da unidade cúbica definida pelo GRID.

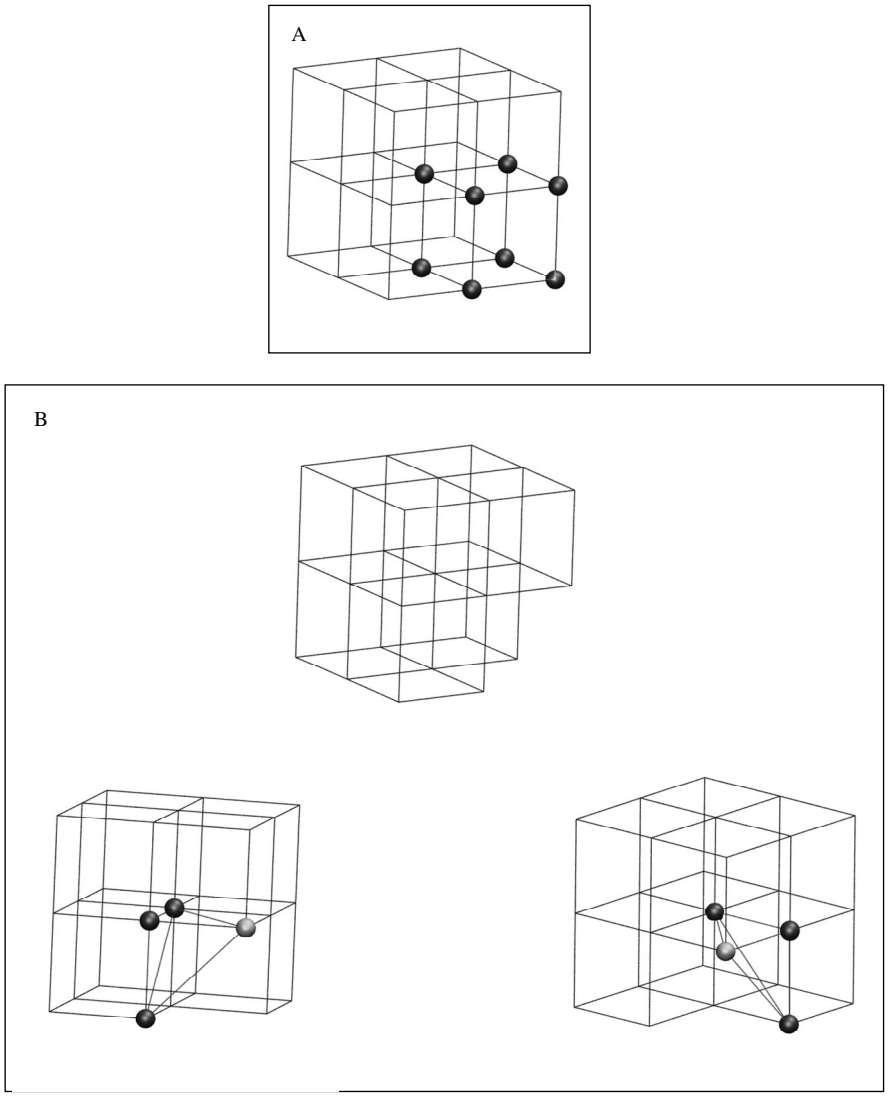


Figura 14: As duas possibilidades para calcular o volume. Em A quando é possível formar um cubo. Em B quando não é possível formar um cubo, mas é possível formar tetraedros não sobrepostos.

No primeiro caso o volume é definido pelo valor da distância entre os pontos do GRID elevado a terceira potência (neste trabalho $0{,}8^3 = 0{,}521$ Å^3). No segundo caso é a soma dos valores dos volumes dos tetraedros formados:

$$\text{Volume}_{\text{tetraedro}} = (\text{Área da base} * \text{Altura})/3$$

Em seguida são utilizados os tetraedros regulares do cubo:

$Volume_{tetraedro} = ([(l * l)/2] * l)/3$ onde

l = lado do cubo = distância entre os pontos do GRID

Neste trabalho o valor do volume do tetraedro foi:

$Volume_{tetraedro} = ([(0,8 * 0,8)/2] * 0,8)/3$

$Volume_{tetraedro} = (0,32 * 0,8)/3$

$Volume_{tetraedro} = 0,0853\ Å^3$

O volume da cavidade é a soma do volume de todos os cubos formados mais o volume de todos os tetraedros formados.

2.6.6 Criação da superfície da cavidade

A criação da superfície da cavidade, para simplificar a visualização (Figura 15), foi gerada utilizando as informações de todos os pontos da cavidade localizada na região de acesso ao substrato.

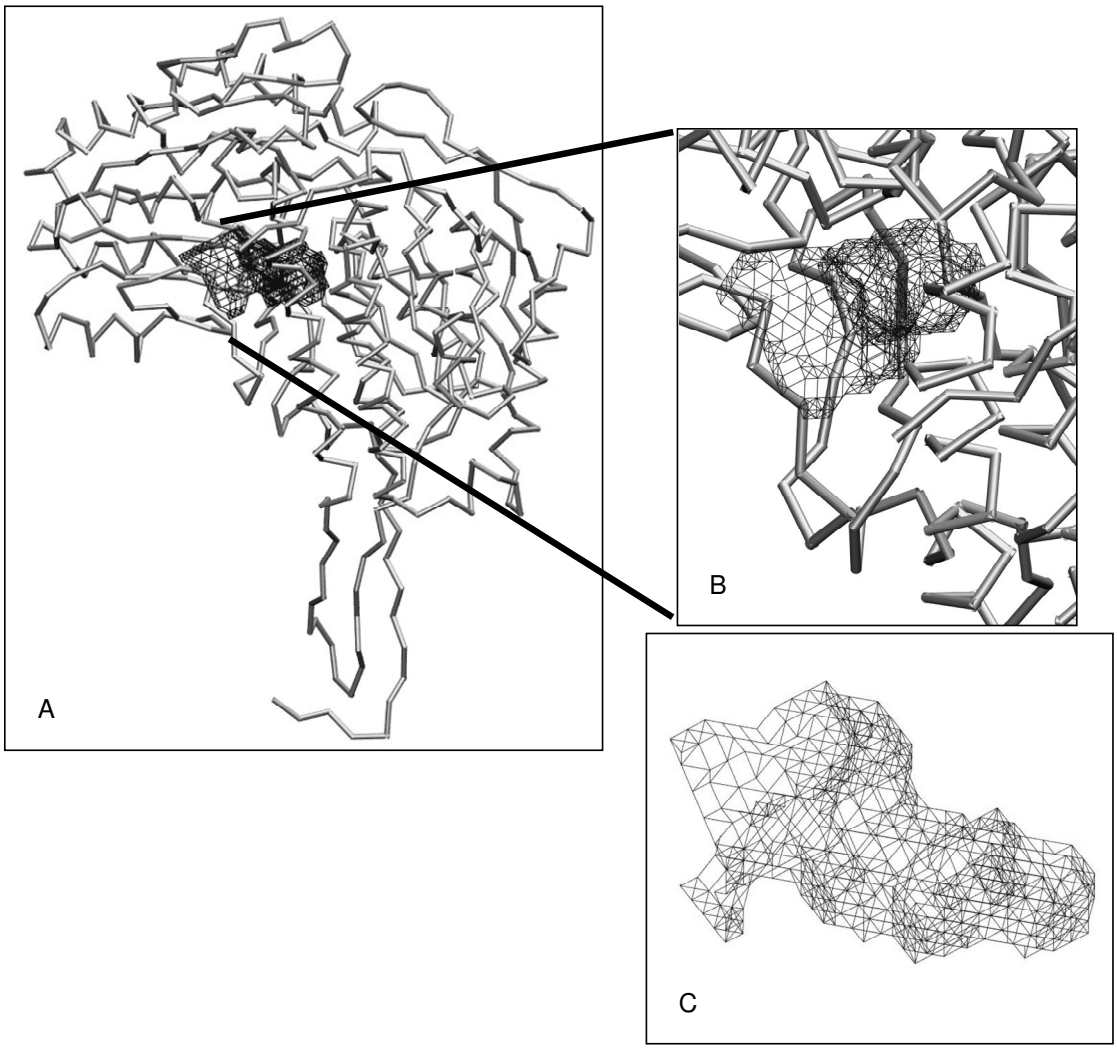


Figura 15: Três vistas do canal de acesso ao substrato. Em A, temos uma vista ampla da proteína e o canal de acesso ao substrato definido no interior em azul escuro. Em B, uma vista ampliada na região do canal de acesso ao substrato e em C, apenas o contorno do canal de acesso ao substrato.

Para definir os pontos da superfície do canal percorre-se uma linha do GRID procurando por um ponto pertencente a uma cavidade, quando encontrado um ponto, é verificado se esse ponto possui apenas pontos à frente ou pontos atrás. Se essa condição é respeitada ele é definido como um ponto da superfície. Esse processo é realizado para todas as linhas nos três eixos (Figura 16).

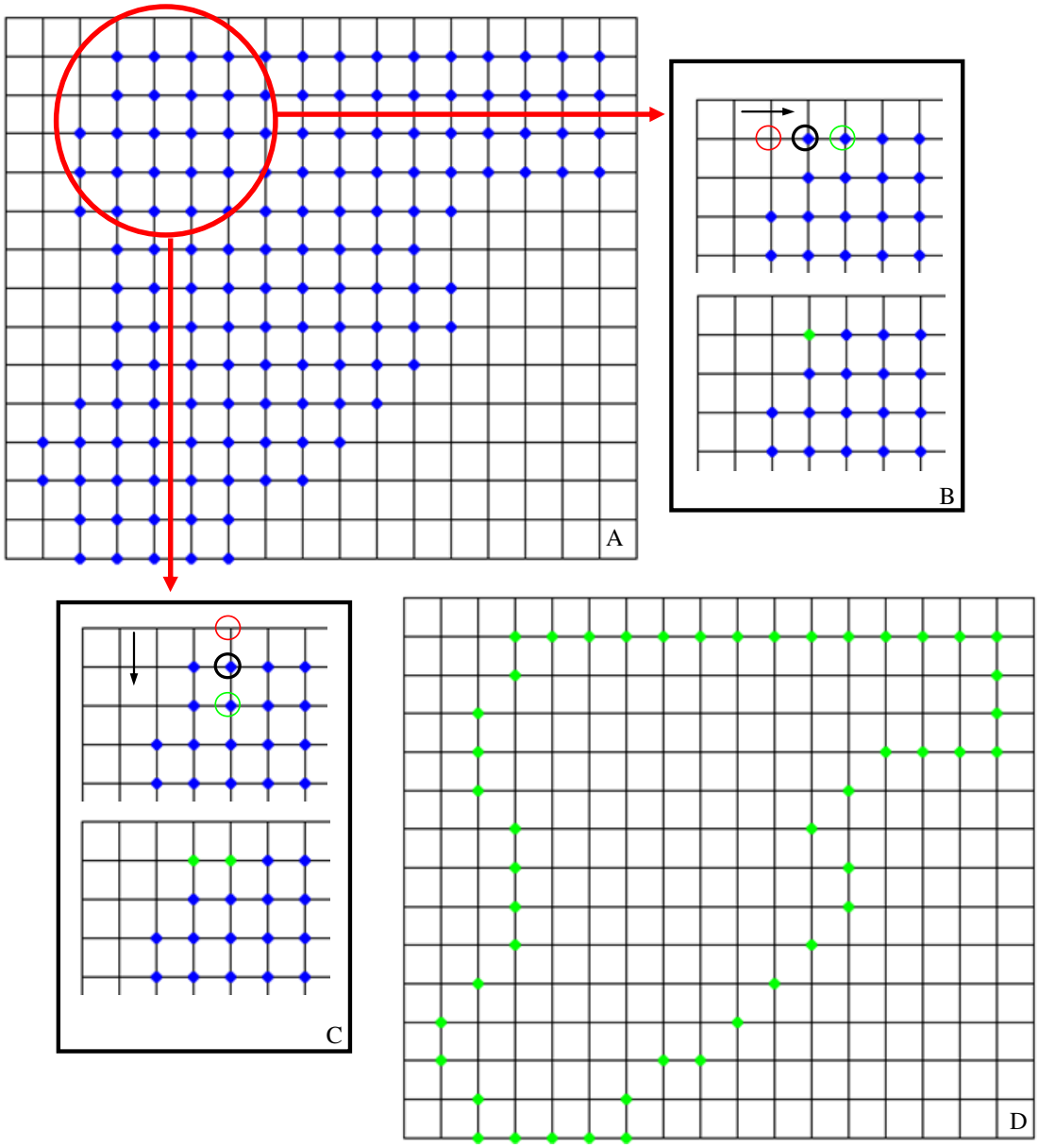


Figura 16: Processo de definição dos pontos da superfície da cavidade. Em A, todos os pontos de um corte de uma cavidade. Em B, a etapa de busca por pontos da superfície no eixo X. Em C, a etapa de busca por pontos da superfície no eixo Y. Em D, apenas os pontos considerados da superfície.

2.7 Clonagem dos genes de anfioxo e expressão gênica

A clonagem dos genes das ALDH1/2 a partir da biblioteca de cDNA e a análise de hibridização *in situ* foi conduzida utilizando os protocolos padrão (Holland 1996).

Para a análise da expressão gênica, foram utilizados anfioxos adultos sexualmente maduros, coletados utilizando pá e peneira em Tampa Bay, Florida. Os gametas foram obtidos por meio de estimulação elétrica dos adultos e, após a fertilização, os embriões e as larvas foram crescidas em laboratório e fixadas em diferentes estágios de desenvolvimento. Cinco bibliotecas de cDNA foram utilizadas para a clonagem dos genes da ALDH1 e ALDH2 de anfioxo. As bibliotecas de cDNA foram feitas a partir de RNA de ovos não fertilizados, gástrulas, nêurulas, larvas e adultos maduros (Yu, Satou et al. 2007). Os experimentos de hibridização *in situ* foram conduzidos utilizando os protocolos padrão (Holland 1996) com temperaturas crescentes (65 °C to 70 °C), e as regiões não conservadas dos diferentes genes como moldes para a síntese das sondas anti-senso para assegurar a especificidade. Experimentos controles foram conduzidos com sondas para verificar a especificidade do modelo de expressão. Após a hibridização *in situ*, os embriões foram montados sobre lâminas de vidro e analisados com microscópio e fotografado inteiro. Estas etapas foram realizadas pelo grupo do Dr. José Xavier Neto.

2.8 Análise estatística

Os volumes dos canais foram analisados em conjuntos não pareados de ALDH1/2 de vertebrados e grupos pareados de ALDH1/2 de vertebrados (ALDH1As, ALDH1B1s e ALDH2) por testes de ANOVA não paramétrica utilizando teste múltiplos de comparação Kruskal-Wallis e Dunn's.

3. Resultados

3.1 Busca por seqüências de ALDH em organismos com genomas montados e anotados

A obtenção de seqüência de ALDHs em organismos com genomas montados e anotados baseada em similaridade foi realizada procurando encontrar as seqüências das enzimas em organismos modelos, para representar a evolução das seqüências de ALDH ao longo do tempo. Foram encontradas 190 seqüências de diferentes famílias da super-família ALDH. Algumas proteínas não apresentam anotação em relação à família de ALDH a qual pertencem e, por isso, não foram classificadas. A maioria das ALDHs não classificadas são de plantas e fungos. Isso decorre pelo fato de que estas enzimas ainda não apresentam uma classificação funcional bem definida. A tabela abaixo (Tabela 4) apresenta os nomes dos organismos e as referências das seqüências de ALDH obtidas. Não foi possível encontrar todas as ALDHs de interesse apenas nos bancos de seqüências genômicas, o que conduziu o trabalho a procurar as seqüências de ALDH nas informações disponíveis dos projetos de seqüenciamento de EST.

Tabela 4: Nome do organismo, a classificação da ALDH e a referência da seqüência do organismo por nome do gene e número de acesso no Genbank ou do EMSEMBL.

	ALDH1Ls		
1	*Caenorhabditis elegans*	ALDH1L1	gi_122127277
2	*Nematostella vectensis*	ALDH1L	17098
3	*Drosophila melanogaster*	ALDH1L1	gi_74869229
4	*Ciona savignyi*	ALDH1L	ALDH1L_Ciosav_07474
5	*Homo sapiens*	ALDH1L1	gi_59802911
6	*Rattus norvegicus*	ALDH1L1	gi_11968144
7	*Mus musculus*	ALDH1L1	gi_27532959
8	*Mus musculus*	ALDH1L2	gi_23956316
9	*Strongylocentrotus purpuratus*	ALDH1L	gi_115913964
10	*Branchiostoma floridae*	ALDH1L	269554
	TUNICADOS ALDH1s		
11	*Ciona savignyi*	ALDH1	3997
12	*Ciona intestinalis*	ALDH1	7939
13	*Ciona savignyi*	ALDH1	1325
14	*Ciona intestinalis*	ALDH1	7835

15	*Ciona intestinalis*	ALDH1	7853
		ECDISOZOA ALDHs	
16	*Caenorhabditis elegans*	ALDH1J2	gi_75024717
17	*Caenorhabditis elegans*	ALDH9B2	gi_74964846
18	*Drosophila melanogaster*	ALDH1M1	gi_74868141
		ALDH1s	
19	*Nematostella vectensis*	ALDH1	26342
20	*Nematostella vectensis*	ALDH1	15570
21	*Xenopus laevis*	ALDH1	gi_50368699
22	*Xenopus tropicalis*	ALDH1A1	gi_73853806
23	*Ornithorhynchus anatinus*	ALDH1A1	ENSOANP00000020
24	*Monodelphis domestica*	ALDH1A1	Mondom_frag_1_peptide_E
25	*Gallus gallus*	ALDH1A1	gi_45383031
26	*Rattus norvegicus*	ALDH1A1	gi_14192935
27	*Mus musculus*	ALDH1A1	gi_14192935
28	*Oryctolagus cuniculus*	ALDH1A1	gi_46240839
29	*Canis familiaris*	ALDH1A1	gi_57035983
30	*Bos taurus*	ALDH1A1	gi_75775505
31	*Ovis aries*	ALDH1A1	gi_57526379
32	*Pongo abelii*	ALDH1A1	gi_55732507
33	*Pan troglodytes*	ALDH1A1	gi_114625012
34	*Macaca mulatta*	ALDH1A1	gi_109111778
35	*Macaca fascicularis*	ALDH1A1	gi_42558919
36	*Homo sapiens*	ALDH1A1	gi_2183299
37	*Oryzias latipes*	ALDH1A2	gi_114325216
38	*Danio rerio*	ALDH1A2	gi_16565463
39	*Danio rerio*	ALDH1A2	gi_18858265
40	*Takifugu rubripes*	ALDH1A2	gi_75832152
41	*Xenopus laevis*	ALDH1A2	gi_118763686
42	*Xenopus tropicalis*	ALDH1A2	gi_113931498
43	*Gallus gallus*	ALDH1A2	gi_92087020
44	*Monodelphis domestica*	ALDH1A2	ENSMODP0000001
45	*Rattus norvegicus*	ALDH1A2	gi_16758767
46	*Mus musculus*	ALDH1A2	NM_009022
47	*Taeniopygia guttata*	ALDH1A2	gi_115529234
48	*Canis familiaris*	ALDH1A2	ENSCAFP0000002431
49	*Bos taurus*	ALDH1A2	gi_119925344
50	*Pan troglodytes*	ALDH1A2	gi_114657247
51	*Macaca mulatta*	ALDH1A2	gi_109081319
52	*Homo sapiens*	ALDH1A2	gi_90109797
53	*Danio rerio*	ALDH1A3	gi_83375892
54	*Xenopus laevis*	ALDH1A3	gi_56684572
55	*Gallus gallus*	ALDH1A3	gi_45382777
56	*Monodelphis domestica*	ALDH1A3	ENSMODP00000015203
57	*Mus musculus*	ALDH1A3	gi_31542123
58	*Rattus norvegicus*	ALDH1A3	gi_52782764
59	*Canis familiaris*	ALDH1A3	ENSCAFP00000015693
60	*Macaca mulatta*	ALDH1A3	gi_109082542
61	*Pan troglodytes*	ALDH1A3	gi_114659192
62	*Homo sapiens*	ALDH1A3	gi_4502041
63	*Mus musculus*	ALDH1A7	gi_127797599
64	*Rattus norvegicus*	ALDH1A4	gi_14485281
65	*Branchiostoma floridae*	ALDH1	261794
66	*Branchiostoma floridae*	ALDH1	118947
67	*Branchiostoma floridae*	ALDH1	86762
68	*Branchiostoma floridae*	ALDH1	124720
69	*Branchiostoma floridae*	ALDH1	113974
70	*Branchiostoma floridae*	ALDH1	206892
71	*Saccoglossus kowalevskii*	ALDH1	ALDH1_Sackow_a gnl|ti|934912767
72	*Saccoglossus kowalevskii*	ALDH1	ALDH1_Sackow_d gnl|ti|934906458 gnl|ti|934885758 gnl|ti|934862769
73	*Saccoglossus kowalevskii*	ALDH1	ALDH1_Sackow_e gnl|ti|934820532 gnl|ti|934856661
74	*Saccoglossus kowalevskii*	ALDH1	ALDH1_Sackow_b gnl|ti|934825314
75	*Saccoglossus kowalevskii*	ALDH1	ALDH1_Sackow_c gnl|ti|934875458
		ALDH2s	
76	*Saccoglossus kowalevskii*		ALDH2_Sackow
77	*Strongylocentrotus purpuratus*	ALDH2	786833

78	*Strongylocentrotus purpuratus*	ALDH2	786787
79	*Branchiostoma floridae*	ALDH2	106815
80	*Drosophila melanogaster*	ALDH1A10	gi_74869734
81	*Ciona savignyi*	ALDH2	1705
82	*Ciona intestinalis*	ALDH2	ALDH2_Cioint_ALDH2
83	*Nematostella vectensis*	ALDH2	15494
84	*Gasterosteus aculeatus*	ALDH2	ALDH2_Gasacu_reconstructed
85	*Danio rerio*	ALDH2	gi_20339358
86	*Tetraodon nigroviridis*	ALDH2	GSTENP00009913001
87	*Xenopus laevis*	ALDH2	gi_50415001
88	*Xenopus tropicalis*	ALDH2	gi_52345722
89	*Ornithorhynchus anatinus*	ALDH2	ENSOANP000000075
90	*Monodelphis domestica*	ALDH2	ENSMODP000000152
91	*Gallus gallus*	ALDH2	gi_118098552
92	*Rattus norvegicus*	ALDH2	gi_14192932
93	*Mus musculus*	ALDH2	gi_142371450
94	*Mesocricetus auratus*	ALDH2	gi_3121992
95	*Canis familiaris*	ALDH2	ALDH2_Canfam_reconstructed
96	*Bos taurus*	ALDH2	gi_2624887
97	*Pongo abelii*	ALDH2	gi_55725752
98	*Pan troglodytes*	ALDH2	gi_114647006
99	*Macaca mulatta*	ALDH2	Macmul_predicted
100	*Homo sapiens*	ALDH2	gi_118504
101	*Apis mellifera*		gi_66530423
102	*Tribolium castaneum*		gi_91077042
103	*Bombyx mori*		gi_114052408
104	*Drosophila pseudoobscura*		gi_125986975
105	*Anopheles gambiae*		gi_118784054
106	*Drosophila melanogaster*	ALDH1A10	gi_74869734
107	*Aedes aegypti*		gi_108874923
108	*Xenopus tropicalis*	ALDH1B1	ENSXETP00000032265
109	*Ornithorhynchus anatinus*	ALDH1B1	ENSOANG2719
110	*Rattus norvegicus*	ALDH1B1	gi_58865517
111	*Mus musculus*	ALDH1B1	gi_21312260
112	*Canis familiaris*	ALDH1B1	ENSCAFP000000034
113	*Bos taurus*	ALDH1B1	gi_122065128
114	*Pongo abelii*	ALDH1B1	gi_75061684
115	*Pan troglodytes*	ALDH1B1	gi_114624620
116	*Macaca mulatta*	ALDH1B1	gi_109110898
117	*Homo sapiens*	ALDH1B1	gi_1263008
	PLANTAS ALDHs GROUPO A		
118	*Oryza sativa*		gi_8163730
119	*Oryza sativa*		gi_75309548
120	*Oryza sativa*		gi_75311498
121	*Zea mays*		gi_19850251
122	*Zea mays*		gi_75282505
123	*Arabidopsis thaliana*		gi_75313899
124	*Arabidopsis thaliana*		gi_118595573
125	*Populus trichocarpa*		estExt_fgenesh4_pm_C_LG_II0915
126	*Populus trichocarpa*		estExt_fgenesh4_pg_C_LG_XII076
127	*Populus trichocarpa*		grail3_0005008802
128	*Sorghum bicolor*		gi_20530131
129	*Sorghum bicolor*		gi_20530129
130	*Nicotiana tabacum*		gi_75279838
131	*Secale cereale*		gi_20530127
	PLANTAS ALDHs GROUP B		
132	*Oryza sativa*		gi_75311485
133	*Zea mays*		gi_19850243
134	*Zea mays*		gi_19850247
135	*Arabidopsis thaliana*		gi_20530143
136	*Populus trichocarpa*		pm_C_LG_XVIII000233_Pt
	FUNGOS ALDHs		
137	*Phycomyces blakesleeanus*		35266
138	*Phycomyces blakesleeanus*		39661
139	*Phycomyces blakesleeanus*		30289
140	*Phanerochaete chrysosporium*		140859

141	*Phanerochaete chrysosporium*		140689
142	*Phanerochaete chrysosporium*		133924
143	*Phanerochaete chrysosporium*		138693
144	*Phanerochaete chrysosporium*		1350
145	*Phanerochaete chrysosporium*		137014
146	*Schizosaccharomyces pombe*	ALDH1R1	gi_74638854
147	*Saccharomyces cerevisiae*	ALDH1G2	gi_1706382
148	*Schizosaccharomyces pombe*	ALDH1S1	gi_30913454
149	*Saccharomyces cerevisiae*	ALDH1F2	gi_20981689
150	*Saccharomyces cerevisiae*	ALDH1D1	gi_1706378
151	*Saccharomyces cerevisiae*	ALDH1D2	gi_2506349
152	*Emericella nidulans*	ALDH57	gi_13309888
153	*Emericella nidulans*		gi_168011
154	*Beauveria bassiana*		gi_110592121
155	*Magnaporthe grisea*		gi_39943778
156	*Chaetomium globosum*		gi_116199967
157	*Phaeosphaeria nodorum*		gi_111064456
		VERTEBRADOS ALDH2s	
158	*Xenopus tropicalis*	ALDH2	gi_52345722
159	*Ornithorhynchus anatinus*	ALDH2	ENSOANP000000075
160	*Monodelphis domestica*	ALDH2	ENSMODP000000152
161	*Rattus norvegicus*	ALDH2	gi_14192932
162	*Mus musculus*	ALDH2	gi_142371450
163	*Canis familiaris*	ALDH2	ALDH2_Canfam_reconstructed
164	*Bos taurus*	ALDH2	gi_2624887
165	*Pongo abelii*	ALDH2	gi_55725752
166	*Pongo abelii*	ALDH2	gi_114647006
167	*Macaca mulatta*	ALDH2	ALDH2_Macmul_predicted
168	*Homo sapiens*	ALDH2	gi_118504
		ALDH1B1s	
169	*Xenopus tropicalis*	ALDH1B1	ENSXETP00000032265
170	*Ornithorhynchus anatinus*	ALDH1B1	ENSOANG2719
171	*Monodelphis domestica*	ALDH1B1	ENSMODP0000003
172	*Rattus norvegicus*	ALDH1B1	gi_58865517
173	*Mus musculus*	ALDH1B1	gi_21312260
174	*Canis familiaris*	ALDH1B1	ENSCAFP000000034
175	*Bos taurus*	ALDH1B1	gi_122065128
176	*Pongo abelii*	ALDH1B1	gi_75061684
177	*Pan troglodytes*	ALDH1B1	gi_114624620
178	*Macaca mulatta*	ALDH1B1	gi_109110898
179	*Homo sapiens*	ALDH1B1	gi_1263008
		ALDH1A1as	
180	*Xenopus tropicalis*	ALDH1A1	gi_73853806
181	*Ornithorhynchus anatinus*	ALDH1A1	ENSOANP00000020
182	*Monodelphis domestica*	ALDH1A1	frag_1_peptide_E
183	*Rattus norvegicus*	ALDH1A1	gi_14192935
184	*Mus musculus*	ALDH1A1	gi_14192935
185	*Canis familiaris*	ALDH1A1	gi_57035983
186	*Bos taurus*	ALDH1A1	gi_75775505
187	*Pongo abelii*	ALDH1A1	gi_55732507
188	*Pan troglodytes*	ALDH1A1	gi_114625012
189	*Macaca mulatta*	ALDH1A1	gi_109111778
190	*Homo sapiens*	ALDH1A1	gi_2183299

3.2 Busca por seqüência de ALDH em organismos não montados

O método para busca de seqüências a partir das informações não montadas dos projetos genoma veio da necessidade de obter seqüências de organismos com o genoma ainda não

montado e/ou anotado. O protocolo utilizado foi desenvolvido a partir de uma adaptação do método desenvolvido por Sobreira & Gruber (2008) foram acrescentadas as etapas de classificação das ALDHs e análise com o programa Genwise para permitir a obtenção de seqüência de aminoácidos das ALDHs de interesse. Para verificar se o método desenvolvido seria capaz de reconstruir as ALDHs dos organismos de interesse com acurácia, foi realizada uma validação utilizando os seguintes organismos:

- *Homo sapiens* – com 8.135.177 seqüências de EST.
- *Mus musculus* – com 4.850.243 seqüências de EST.
- *Ratus novergicus* – com 837.268 seqüências de EST.

Esses organismos estão extensamente anotados, apresentando as enzimas ALDHs (http://alhd.org) mostradas na Tabela 5. A comparação entre o comprimento da seqüência da proteína obtida experimentalmente foi então comparado com o comprimento das ALDHs obtidas pelo método de reconstrução.

Tabela 5: Comparação entre os comprimentos das seqüências de ALDHs obtidas pelo método desenvolvido com o comprimento das respectivas seqüências de aminoácidos esperado.

	Homo sapiens		*Mus musculus*		*Rattus norvegicus*	
	Comprimento obtido	Comprimento esperado	Comprimento obtido	Comprimento esperado	Comprimento Obtido	Comprimento esperado
ALDH1A1	501	501	453	501	501	501
ALDH1A2	485	480	501	499	498	499
ALDH1A3	472	512	512	512	309	512
ALDH1B1	457	517	515	519	519	519
ALDH1L1	902	902	902	902	547	902
ALDH1L2	381	923	802	923	-	923
ALDH2	474	517	519	519	519	519
ALDH3A1	453	453	453	453	453	453
ALDH3A2	485	485	373	484	464	484
ALDH3B1	453	468	468	468	238	468
ALDH3B2	377	385	-	479	-	-
ALDH4A1	473	563	562	562	-	-
ALDH5A1	535	548	523	523	143	523
ALDH6A1	535	535	533	535	533	535
ALDH7A1	511	539	511	539	511	539
ALDH8A1	423	487	487	487	-	-
ALDH9A1	495	518	518	518	509	497
ALDH16A1	802	802	802	802	258	802
ALDH18A1	795	795	793	795	793	793

Pode-se observar que apenas as seqüências da ALDH3B2 de camundongo e ALDH1L2 do rato não foram reconstruídas, todas as outras seqüências foram parcialmente ou completamente reconstruídas. No homem, apenas a ALDH1L2 foi reconstruída parcialmente. No camundongo, a ALDH3B2 não foi reconstruída e as ALDH1L2 e ALDH3A2 foram reconstruídas parcialmente. No rato ALDH1L2 não foi reconstruída e as ALDHs das famílias 1A3, 1L1, 5A1 e 16A1 foram reconstruídas parcialmente. Esses dados confirmam a capacidade de reconstruir as seqüências de ALDH a partir das seqüências disponíveis nos projetos de seqüenciamento de EST. A capacidade de reconstrução está relacionada à quantidade de seqüências disponíveis do organismo de interesse. Pode se observar que quanto maior a quantidade de seqüências disponíveis melhor é a reconstrução dos genes. No homem, com mais de 8 milhões de seqüências, todos os genes foram

reconstruídos. No camundongo e no rato apenas um gene não foi reconstruído. O rato apresentou a maior quantidade de reconstrução parcial, fato que se deve, provavelmente, à menor quantidade de seqüências de EST que cubram as regiões dos genes de interesse, o que pode acarretar a reconstrução de alguma outra ALDH semelhante que apresenta maior quantidade de seqüências desviando a montagem do gene de interesse para outra ALDH. Este problema pode ser corrigido aumentando a estringência dos parâmetros do programa BLAST mas com o custo de perder seqüências importantes em organismos que apresentem seqüências da ALDH muito diferentes da ALDH humana. Como o nosso objetivo era obter seqüências das ALDHs das famílias 1, 2 e 8, optou-se por manter parâmetros menos estringentes o que possibilitou a reconstrução de todos os genes nos três organismos.

A utilização do alinhamento múltiplo para classificar as ALDH mostrou ser muito preciso. As seqüências de ALDHs de humano, camundongo e rato obtidas foram submetidas a uma busca de similaridade de seqüência utilizando o programa BLAST e a base de dados de proteínas do NCBI (http://www.ncbi.nlm.nih.gov/) e todas as seqüências obtidas foram classificadas corretamente de acordo com as anotações disponíveis das seqüências do NP que apresentaram maior identidade.

O Anexo 1 contém uma tabela com todas as ALDHs obtidas e anotadas para cada um dos organismos analisados pelo método. Na Tabela 6 é apresentado o número de ALDHs obtidas, considerando-se os 498 organismos analisados, separadas pela subfamília.

Tabela 6: Tabela com quantidade de ALDHs obtidas, separas pela subfamília.

ALDH1A1	46
ALDH1A2	50
ALDH1A3	35
ALDH1B1	7
ALDH1L1	152
ALDH1L2	49
ALDH2	64
ALDH3A1	122
ALDH3A2	19
ALDH3B1	10
ALDH3B2	14
ALDH4	39
ALDH5	8
ALDH6	68
ALDH7	103
ALDH8	124
ALDH9	5
ALDH16	32
ALDH18	69

Pode-se observar que algumas ALDHs foram reconstruídas em baixo número. Esse fato pode ser relacionado ao diferente nível de expressão de cada um dos genes analisados. A medida que a base de dados de seqüências é formada por ESTs, cuja quantidade de leituras encontradas em uma biblioteca de cDNA não normalizada é proporcional ao nível de expressão de cada gene de interesse, faz com que genes mais expressos apresentem um maior número de leituras da região codificadora do gene do que para genes menos expressos. Isso pode ser notado pelo número de vezes que ALDH1L1 foi encontrado, desde que esse gene representa mais de 1% das proteínas citosólicas solúveis das células do fígado de mamíferos (Krupenko, 2002). ALDH3A1 foi o terceiro gene mais encontrado. De fato, a proteína codificada por esse gene representa 10 a 40% das proteínas hidrossolúveis da córnea de vários organismos (King, 1998).

3.2.1 Análise do genoma do anfioxo

O genoma do anfioxo foi analisado a partir das informações das bases de dados DBest e Trace Archives, para a realização da busca e reconstrução de seqüências de ALDH1 e ALDH2. No início desse trabalho estavam disponíveis apenas as seqüências não havendo nenhuma montagem. Foram reconstruídas então sete seqüências. Comparando essas seqüências com as das ALDHs humanas, foi previsto que seis seqüências eram possíveis ALDH1 e a outra uma possível ALDH2 (Tabela 7).

Tabela 7: Relação dos genes e o comprimento da seqüência de aminoácidos traduzida a partir dos transcritos reconstruídos.

Gene	Comprimento
ALDH1a	422
ALDH1b	369
ALDH1c	497
ALDH1d	497
ALDH1e	495
ALDH1f	497
ALDH2	511

A publicação do genoma do anfioxo (Putnam, Butts et al. 2008) tornou possível mapear as seqüências de ALDH obtidas (Figura 17).

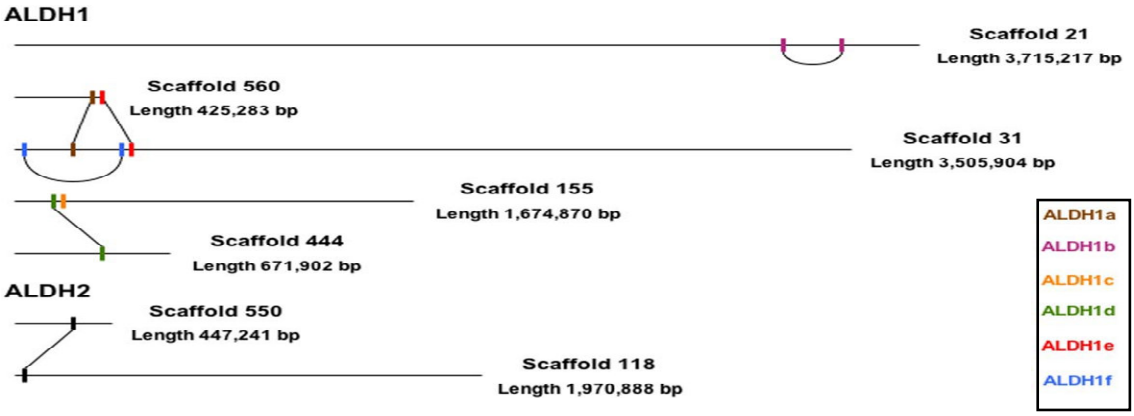


Figura 17: Mapeamento dos genes preditos das ALDH1 e ALDH2 sobre os scaffolds do genoma de anfioxo. Foram identificados seis possíveis genes que codificariam a ALDH1, onde cinco desses apresentam-se com duas cópias no genoma. O gene da ALDH2 apresenta apenas uma seqüência com duas cópias no genoma.

A confirmação das seqüências preditas do Anfioxo ajuda a corroborar a precisão do método de busca por seqüências.

3.3 Análises filogenética

As ALDH1s de vertebrados e ALDH2s compartilham uma alta identidade da seqüência de aminoácidos. Para entender a evolução da afinidade aos substratos tão diferentes entre enzimas tão próximas foi realizada uma análise filogenética em larga escala da super-família das ALDHs (Figura 18).

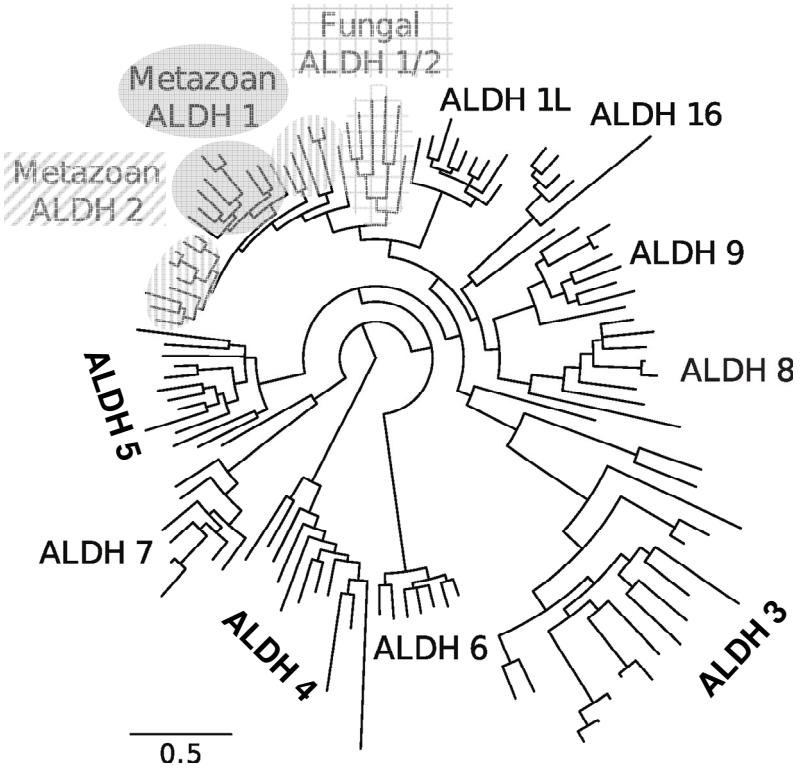


Figura 18: Análise filogenética em larga escala das ALDH de eucariotos. Cada subfamília de ALDH é individualmente suportada por uma probabilidade posterior maior que 0,95.

A habilidade de sintetizar o ácido retinóico a partir do retinaldeído evoluiu independentemente ao menos duas vezes nos metazoários: nas ALDH8 e no ramo que levou às ALDH1A e ALDH2. Em contraste às sugestões anteriores (Vasiliou, Bairoch et al. 1999), as ALDH1s e as ALDH2s são membros de uma única e bem suportada subfamília e estas ALDH2s formam um clado distinto que é agrupado em ALDH1/2 (Figura 18). Assim ALDH1/2s constitui uma exceção a outros grupos fortemente sustentado da subfamília ALDH, onde as enzimas normalmente se agrupam de acordo com a afinidade ao substrato. Interessantemente em contrate com as ALDH2s, as ALDH1s sofrem múltiplas duplicações linhagem específicas (como exemplo, o anfioxo tem seis duplicatas da ALDH1) (Figura 18, Figura 19, Figura 20). A inclusão de seqüências das ALDH1/2 de plantas e fungos na

filogenia baseada na ALDH1/2s, ALDH1Ls (formil tetrahidrofalate desidrogenase) e ALDH8s como grupo externo, mostra que as seqüências das ALDH1/2 de fungo e plantas formam dois clados fortemente sustentados na base das ALDH1/2s dos metazoários. Isso sugere que o último ancestral comum dos eucariotos tem apenas uma única ALDH1/2 e que as de fungos, plantas e metazoários evoluíram independentemente (Figura 19).

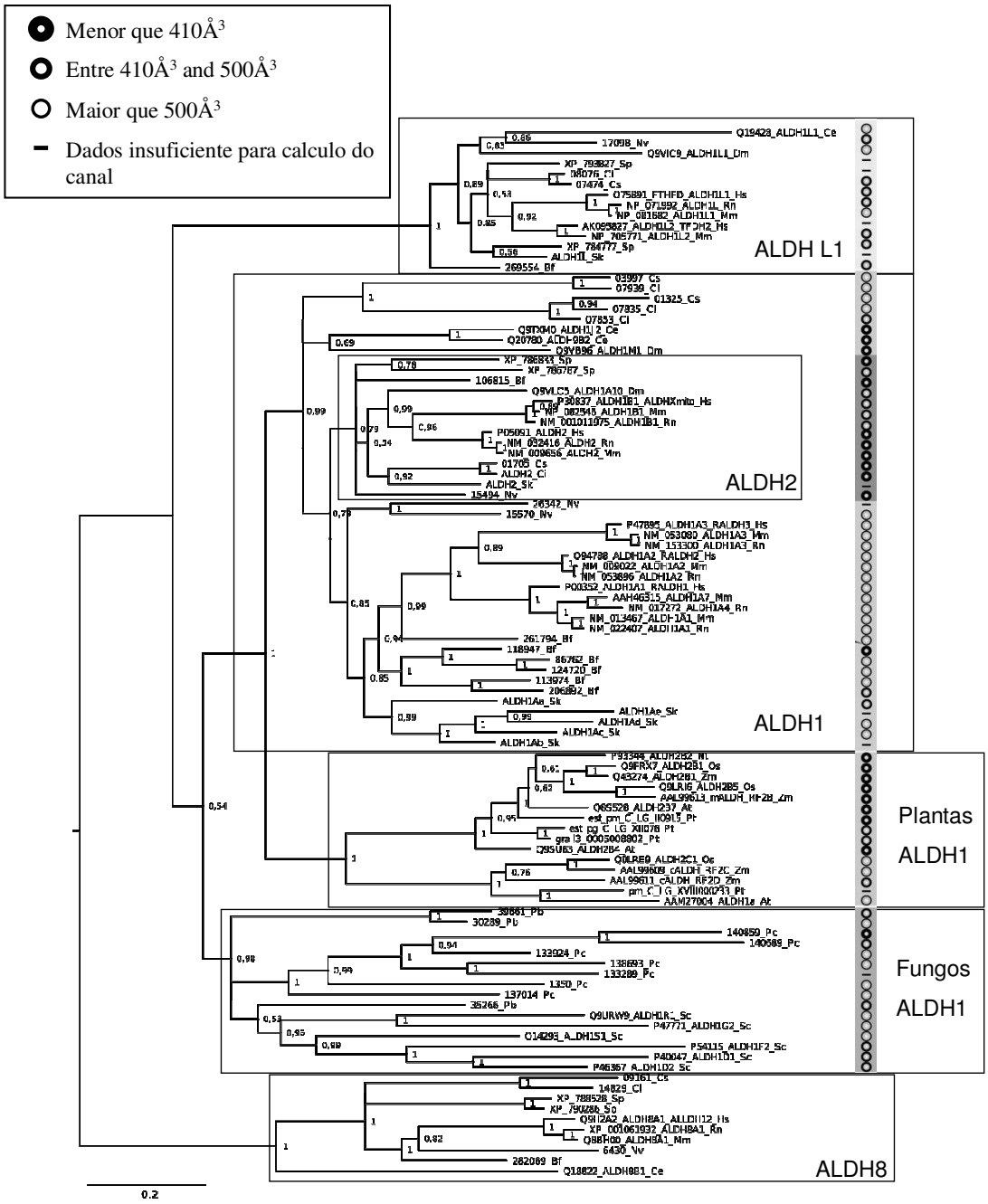


Figura 19: Filogenia diagramática das ALDH1/2s, ALDH1Ls e ALDH8s. As ALDH8s foram utilizadas como grupo externo. ALDH1/2 de metazoários, plantas e fungos formam três clados distintos e bem suportados sugerindo que estes três grupos evoluíram independentemente a partir de uma única ALDH1/2 eucariótica ancestral. As ALDH1/2s metazoários que incluem ALDH1s, ALDH2s e ALDH1B1s, ALDH2 duplicadas específicas de vertebrados. Também está incluída na figura uma representação em três categorias o tamanho do canal.

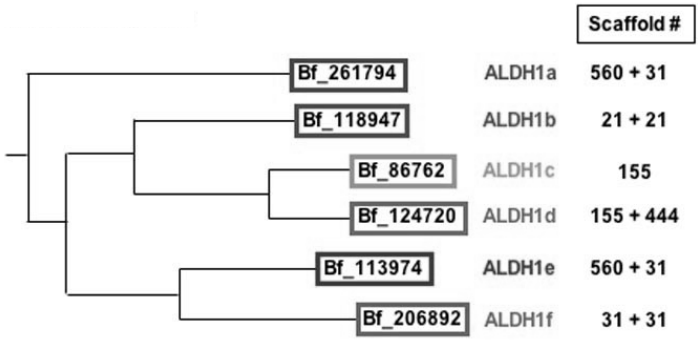


Figura 20: As análises para as ALDH1 duplicadas do Anfioxo sugerem que ao menos alguns genes como ALDH1c e ALDH1d (no scaffold 155), ALDH1a a ALDH1e (no scaffold 560) ou ALDH1a, ALDH1e e ALDH1f (no saffold 31) localizadas no mesmo scaffold e provavelmente ligadas no mesmo cromossomo. As análises filogenética e de sintenia sugerem que as seis ALDH1 de Anfioxo estão relacionadas a duplicação em série do ancestral da ADLH1a.

3.4 Análise da forma e do volume do canal de ligação do substrato

As ALDH1/2s representam um bom modelo para estudar evolução de preferência de substratos e morfogenia-dependente de sinalização, porque sua subfamília é composta de dois grupos relacionados próximos, caracterizados por uma diferença de afinidade aos substratos muito grande. A ALDH2, que catabolíza pequenas moléculas de acetaldeído, apresenta um estreito canal de acesso do substrato com uma abertura pequena enquanto a ALDH1A1 de ovelha, que processa moléculas grandes de retinaldeído, apresentam um largo canal de acesso do substrato com uma grande abertura (Figura 21).

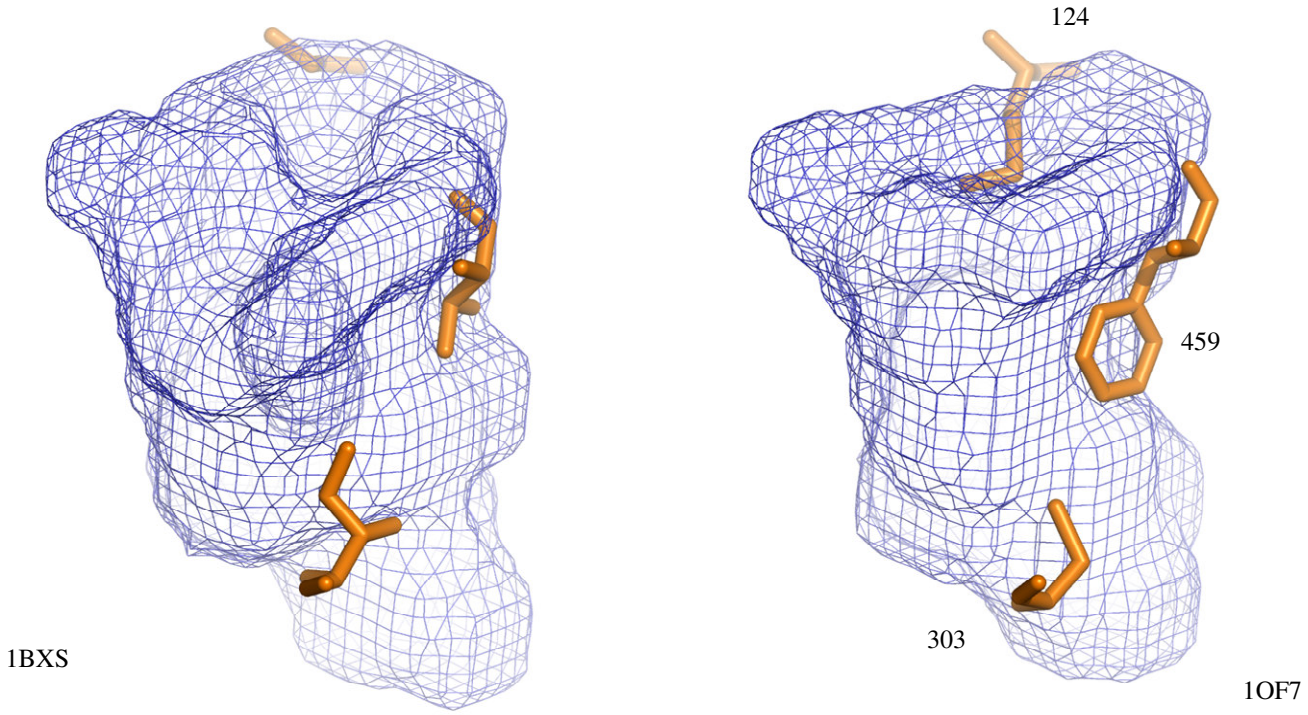


Figura 21: Os canais de acesso ao substrato das estruturas de PDBid 1OF7 (ALDH2) e 1BXS (ALDH1). O canal da ALDH2 apresenta a entrada (próximo ao resíduo 124), o pescoço (próximo ao resíduo 459) e o fundo (próximo ao resíduo 303) mais estreito do que o canal ALDH1.

Estas topologias do canal de acesso influenciam a preferência por substratos das ALDH1/2s. Por exemplo, enquanto retinaldeídos são bons substratos paras as ALDH1s, as ALDH2s de vertebrados apresentam uma atividade muito baixa ou desprezível para grandes aldeídos.

Utilizando as informações estruturais das ALDH1 e ALDH2, foram criados modelos estruturais do canal de acesso ao substrato das ALDH1/2s de eucariotos (Tabela 4) para determinar o volume do canal, implicando na preferência por substratos. No anexo 2 são apresentados os volumes calculados dos canais para cada ALDH modelada e na Figura 19 a filogenia com o tamanho do canal.

As ALDH1s normalmente apresentam volumes maiores (589 ± 59Å^3) que ALDH2 (403 ± 53Å, média ± SD, $p<0,001$). Portanto, o volume do canal representa uma diferença fundamental entre ALDH1s e ALDH2s, refletindo a conservação estrutural requerida associada ao processamento dos diferentes aldeídos.

3.5 Assinaturas

A geometria global do canal de acesso ao substrato determina a preferência enzimática da ALDHs. Analisando o mecanismo mais profundamente conclui-se que a evolução da preferência ao substrato na seqüência é divergente entre ALDH1s e ALDH2s. É possível listar 34 assinaturas de aminoácidos que distinguem e caracteriza as ALDH1As das ALDH2s (anexo 3). Dessas assinaturas, 31 são resíduos do domínio de oligomerização ou exposto na superfície ou dentro da molécula, enquanto 3 delas estão localizadas como um subgrupo dos 27 aminoácidos que formam as cavidades de acesso ao substrato (anexo 3). Surpreendentemente essas 3 assinaturas específicas dos vertebrados não podem ser utilizadas para diferenciar ALDH1s das ALDH2s dos cordados invertebrados sugerindo que tanto as ALDH1s ou ALDH2s dos invertebrados

divergiram das ALDHs homólogas a dos vertebrados. Para determinar se esta divergência é devida à alta freqüência das duplicações linhagem-específicas das ALDH1s, foi utilizado o genoma do anfioxo para analisar em detalhes as assinaturas do canal de seis ALDH1s e uma ALHD2.

3.5.1 Relação entre as assinaturas e a estrutura

A primeira assinatura inclui o aminoácido 124 que está na entrada do canal ("boca") (Figura 21). Na ALDH1A1 de ovelha esse resíduo é uma glicina enquanto que na ALDH2 é uma metionina. Então se a pequena abertura da ALDH2 provavelmente seleciona pequenos substratos, é proposto que resíduo pequeno e resíduo grande na abertura do canal é similar às ALDH1s e ALDH2s respectivamente de invertebrado cordados. Pode-se observar uma leucina na posição 124 da ALDH2 de anfioxo. Já as ALDH1s são heterogêneas, ALDH1a e ALDH1d apresentam glicina na posição 124 que deixam a abertura do canal livre, enquanto que as outras apresentam ácido glutâmico e serina obstruindo parcialmente a entrada do canal (Figura 22). Assim, embora a porção aldeído do retinaldeído possa penetrar no canal das 6 ALDH1s do anfioxo a β-ionona (C13H20O) não é pequena suficiente para entrar nas ALDHs que apresentam ácido glutâmico e serina (Figura 22).

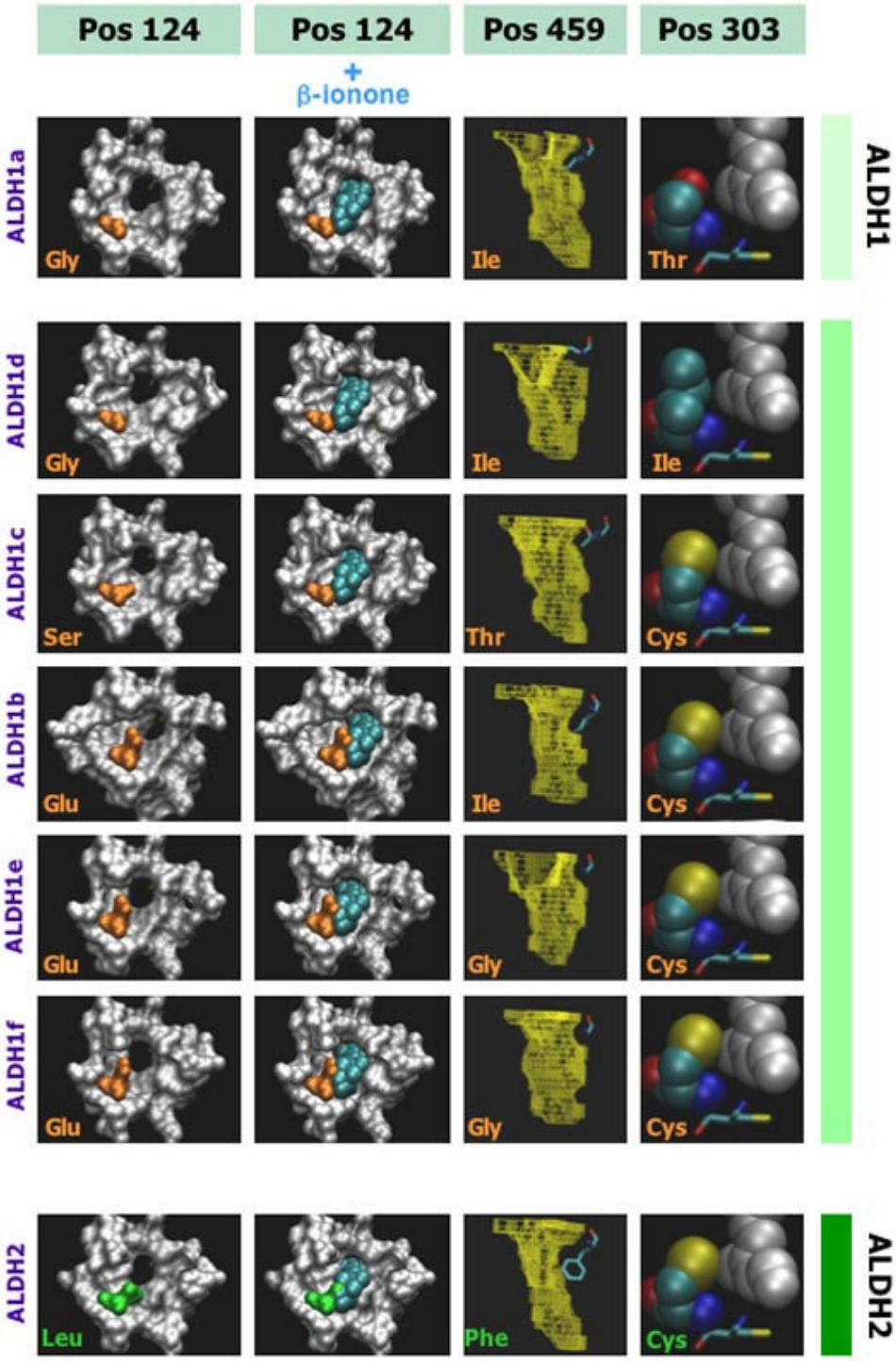


Figura 22: Diversidade das assinaturas do canal de acesso ao substrato em Anfioxo. Nos vertebrados o canal de acesso ao substrato das ALDH1s e ALDH2s pode ser distinguido pelas posições específicas de 3 aminoácidos: resíduo 124 (boca), 459 (pescoço) e 303 (fundo). As ALDH1s do Anfioxo apresentam um largo espectro de assinaturas variando de aminoácidos pequenos típicos de ALDH1As de vertebrados (Gly124, Val459, Ile303) a aminoácidos grandes típicos de ALDH2 de vertebrados (Met124, Phe459, Cys303). Enquanto a ALDH1a e ALDH1d se parecem mais com as ALDH1As de

vertebrados as outras quatro ALDH1As de Anfioxo incorporaram aminoácidos típicos de ALDH2 de vertebrados que também são encontrados nas ALDH2 de Anfioxo.

A segunda assinatura inclui o resíduo 459 que se localiza no primeiro terço do canal (pescoço) (anexo 3). Nas ALDH2s o grande anel aromático da fenilalanina 459 constringe o pescoço do canal. Como nas ALDH1As de vertebrados, em algumas dos anfioxos esse resíduo, é substituído por aminoácidos menores como treonina na ALDH1c ou glicina na ALDH1e e ALDH1f. Mesmo isoleucina não reduz significativamente o diâmetro do canal da ALDH1a e ALDH1b. Mas na ALDH1d o resíduo prolina 458 especifico de Anfioxo (conservados nas ALDH1c, ALDH1d, ALDH1e e ALDH1f) desloca a isoleucina 459, obstruindo parcialmente a entrada do canal e interferindo na acomodação da β-ionona do retinaldeído.

A terceira assinatura inclui o aminoácido 303, próximo da cisteína catalítica 302, no final do canal de acesso do substrato. Nas ALDH2s este aminoácido é uma cisteína. Por meio da modelagem da interação do retinaldeído na ALDH2 de anfioxo foi encontrado choque entre a cisteína 303 e o carbono 20 do retinaldeído. As ALDH1s de anfioxo são heterogêneas, a ALDH1a apresenta uma treonina e a ALDH1d uma isoleucina e nenhum desses apresenta conflito com o retinaldeído, enquanto as outras 4 ALDH1s apresentam o mesmo padrão da ALDH2 (Figura 22).

Enquanto, ao menos a segunda assinatura (aa 459), é conservada entre as ALDHd1/2 de Anfioxo e vertebrados, nenhuma das três assinaturas é conservada entre as ALDH1/2s vertebrados e as ALDH1/2s correspondentes nos invertebrados os tunicados (Figura 23). As assinaturas do canal das ALDH1/2 dos vertebrados diferem das assinaturas dos invertebrados cordados, a maioria das ALDH1s incorporaram aminoácidos e ou motivos estruturais similares ao canal das ALDH2 dos vertebrados.

Em resumo, após a duplicação linhagem específica nas ALDH1 dos invertebrados os cordados apresentaram uma diferença significativa no canal de acesso ao substrato.

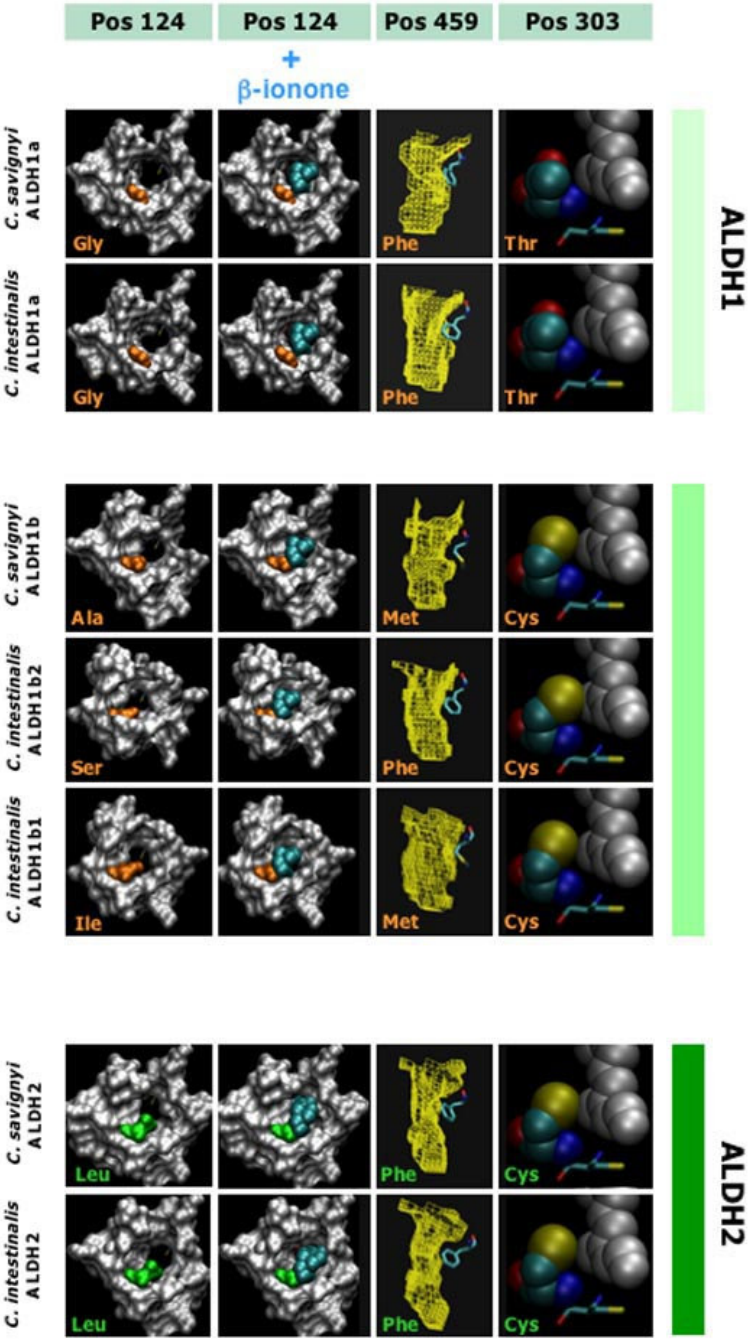


Figura 23: As ALDH1 duplicadas de tunicados apresentam as mesmas assinaturas das ALDH2. Todas as ALDH1s de tunicados são caracterizadas por um aminoácido grande na posição 459 ("pescoço" do canal) que causam um estreitamento do canal. Entretanto, para as duas assinaturas do canal, existe um contínuo entre ALDH1 de tunicados, que

possui os aminoácidos do canal de acesso ao substrato mais relacionados com as ALDH1As de vertebrados (C. savignyi e C. intestinalis ALDH1a), e ALDH1s de tunicados que possui o canal mais relacionado com as ALDH2 (C. savignyi ALDH1b com também C. intestinalis ALDH1b1 e ALDH1b2). As ALDH1a de C. savignyi e C. intestinalis são ambas caracterizadas pela Gly 124, que deixa a boca do canal livre para acomodar a β-ionona do retinaldeído, e pela Thr 303 que ajuda acomodar a porção aldeído do retinaldeído. Em contraste, C. savignyi ALDH1b como a também C. intestinalis ALDH1b1 e ALDH1b2 incorporaram um aminoácido grande, que pode impedir a acomodação da β-ionona. A ALDH1b de C. savignyi e também as ALDH1b1 e ALDH1b2 de C. intestinalis incorporaram a Cys303, que possui o átomo de enxofre que pode causar impedimento com a parte aldeído do retinaldeído. Assim, as ALDH1s duplicadas de tunicados incorporaram aminoácidos típicos de ALDH2 de metazoários sugerindo uma diferenciação das características de ALDH1 ficando mais parecida com as ALDH2.

3.5.2 Expressão das seqüências putativas de ALDHs de Anfioxo

A pergunta seguinte foi se estas mudanças nas assinaturas estão associadas com alterações na função biológica do gene da ALDH1/2. Para esse fim foram clonados os genes da ALDH1 e ALDH2 de anfioxo de bibliotecas de cDNA e foram estudadas as suas expressões durante o desenvolvimento embrionário. Foi observado que a ALDH1a de Anfioxo é muito similar às ALDH1As de vertebrados baseado na expressão gênica, filogenia, assinatura no canal e estrutura (Figura 24).

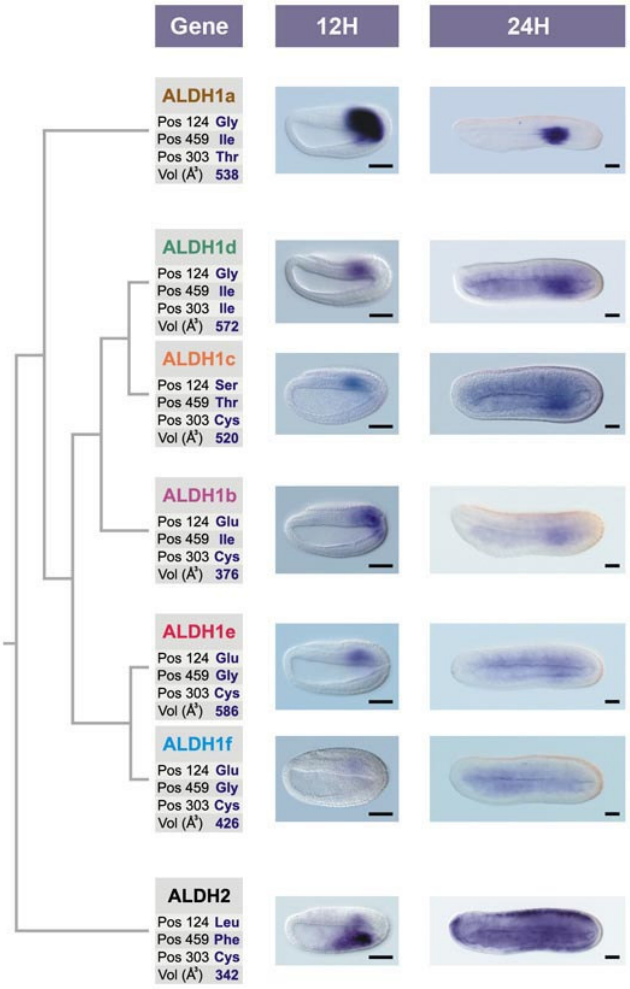


Figura 24: Expressão dos gene de ALDH1/2 de anfioxo. A expressão evidente das ALDH1a é detectável após 12 horas e posteriormente na endoderme após 24 horas. Com 12 horas, as outras cinco ALDH1 são fracamente expressas na região posterior e após 24 horas estão difusamente expressas em todo o embrião. A ALDH2 é expressa posteriormente após 12 horas e após 24 horas é expressa em todo o embrião. Assim, enquanto a ALDH1a provavelmente tem um padrão antero-posterior de posicionamento, as outras ALDH1 possivelmente funcionam como protetora para os diferentes tipos de aldeídos, como a ALDH2.

Na fase de gástrula (12 horas do embrião), a ALDH1a é expressa próxima ao corpo caudal, apresentando-se em uma região limitada. Na fase de nêurula (24 horas), uma forte expressão é detectada em um distinto domínio posterior da endoderme do intestino. Estes padrões são reminiscentes das ALDH1A2s dos vertebrados, que definem a identidade posterior ao longo do eixo antero-posterior de vertebrados. As

outras 5 ALDH1 são fracamente expressas no domínio posterior sobrepondo a ALDH1a (Figura 24). Por 24 horas eles são fracamente expressas e difusas em todo o embrião com uma moderada concentração do sinal das ALDH1d, ALDH1c e ALDH1b na parte posterior da endoderme do intestino. Deste modo, por 24 horas, embora muito fraco, a expressão das cinco ALDH1 é reminiscente a ALDH2 de anfioxo, que é espalhada e conspícua pelo embrião (Figura 24). Também, esse padrão de expressão difuso das quatro ALDH1 (ALDH1c, ALDH1b, ALDH1e e ALDH1f) apresenta a mesma assinatura típica de ALDH2, que resulta em uma constrição do canal, interferindo na acomodação do retinaldeído. Isso sugere que ao menos durante o desenvolvimento esses quatro genes provavelmente influenciam no desenvolvimento embrionário do anfioxo. Após a duplicação linhagem específica, eles devem estar envolvidos em novas funções no processamento de outros aldeídos menores, similar à ALDH2 de Anfioxo. Embora o canal da ALDH1d é superficialmente similar a ALDH1As dos vertebrados, a constrição do canal e também a expressão fraca e difusa no desenvolvimento sugere que esta também tem uma limitada ação no padrão antero-posterior. A partir destes resultados é proposto que ALDH1a é a principal ativadora no padrão antero-posterior do embrião de anfioxo (Figura 24).

Interessantemente, foram achadas evidências de que a troca de aminoácidos na assinatura do canal da ALDH1 de anfioxo reflete-se na tendência geral das ALDH1/2s para alterar a estrutura do acesso ao substrato: algumas das ALDH1/2 acumularam mutações na entrada, pescoço e fundo do canal levando a mudanças da entrada e/ou do volume assim interferindo na especificidade ao substrato (Figura 19). Por exemplo, no ramo das ALDH2s a duplicação de um ancestral vertebrado da ALDH2 com um pequeno canal leva a uma duplicação com a característica de canal pequeno, e outra, a perda de um intron ALDH1B1, com um canal largo (362 ± 30$Å^3$ ALDH2 contra 451 ±

30Å^3 ALDH1B1, p <0,05) pela incorporação de pequenos aminoácidos na boca e no pescoço do canal (Figura 25). Similarmente, o ouriço do mar apresenta duas seqüências filogeneticamente classificadas como ALDH2s (Figura 19): uma com o típico canal estreito da ALDH2 (356 Å^3) e outra com o canal largo (487 Å^3) (Figura 25).

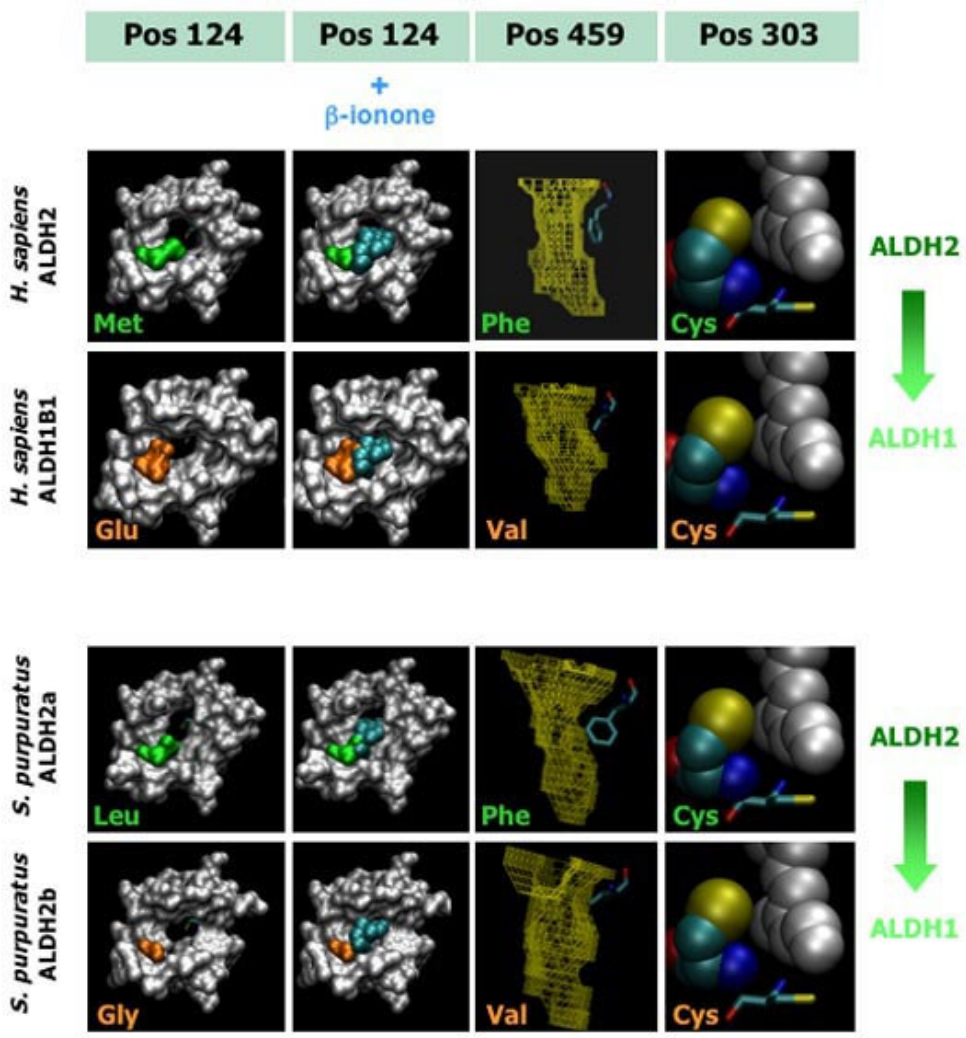


Figura 25: A plasticidade do canal de acesso ao substrato dos deuterostômios. As ALDH2s de vertebrados e equinodermas oferecem mais exemplos da plasticidade evolucionaria do canal das ALHD1/2. Nos vertebrados, as ALDH2s se duplicaram em ALDH2s e ALDH1B1s. Em vertebrados ALDH2s conservaram as assinaturas (MET 124 e LEU 459) mantendo a forma estreita da "boca" e do "pescoço" do canal. A ALDH1B1s incorporou resíduos pequenos (Glu 124 e Val 459) que oferece menos resistência para a acomodação da β-ionona do retinaldeído e também aumentando o canal no "pescoço" elevando o volume do canal (362 ± 30 Å^3 contra 451 ± 30 Å^3 para ALDH2 e ALDH1B1, respectivamente, p<0,05). Uma tendência similar é observada em

ouriço do mar (Strongylocentrotus purpuratus) ALDH2s. S. purpuratus ALDH2b incorporou aminoácidos pequenos (Gly 124 e Val 459), que deixa a boca do canal mais aberta possibilitando a acomodação da β-ionona do retinaldeído e aumentando o volume do canal (356 $Å^3$ na ALDH2a para 487 $Å^3$ na ALDH2b).

4. Conclusões

Essa análise mostra que os membros das ALDH1/2 têm diferenças de tamanho do canal de acesso ao substrato que pode ser categorizado em três classes distintas: pequena, média, grande. No contexto da filogenia da ALDH1/2 (Figura 19) canais pequenos predominam nas ALDH2 de metazoários enquanto canais largos predominam nas ALDH1s de metazoários. Nas plantas, existem dois grandes grupos, um com canal pequeno ou médio e outro com canal grande (345 ± 47$Å^3$ ALDH2 contra 561 ± 78$Å^3$ ALDH1B1, $p < 0,05$) sugerindo que uma simples duplicação separou as plantas em dois clados. Nos fungos, ocorreu uma estratégia diferente, mas com o mesmo resultado: diferentes linhagens fungais parecem ter se duplicado de um único ancestral da ALDH1/2 independentemente. Subseqüente, esta duplicação sofreu alterações no tamanho do canal de acesso ao substrato surgindo ALDH1/2 com tamanhos grandes, médio e pequeno (Figura 19). Interessantemente, membros da subfamília da ALDH1L são caracterizados por canais médio e grande (Figura 19) sugerindo que o último ancestral comum das ALDH1Ls e ALDH1/2s provavelmente tinham canais médios e grandes respectivamente. Portanto, canais pequenos devem estar envolvidos especificamente nas ALDH1/2, o que proveu uma extensa gama de preferência ao substrato permitindo vários tamanhos de aldeídos.

Finalmente, baseado nos resultados do anfioxo, é possível estabelecer uma estrutura do destino dos genes das ALDH1/2 duplicados. Os resultados indicam que os seis genes da ALDH1 originaram-se por duplicação em série do ancestral ALDH1a com funções comparáveis com a ALDH1As de vertebrados: síntese de ácido retinóico para a

sinalização morfogenética (Figura 18 e Figura 19). Também, durante o desenvolvimento, as duplicações da ALDH1 de anfioxo desenvolveram o padrão restritivo ou expansivo reminiscente no padrão de expressão axial versus proteção contra a toxidade dos aldeídos, respectivamente. Esse desenvolvimento é correlato com as estruturas e com o canal de acesso ao substrato das diferentes ALDH1. Portanto, após as duplicações da ALDH1 de anfioxo ocorreram modificações tanto na estrutura da proteína como na regulação gênica, resultando em novas funções, diferentes das ancestrais, envolvendo controle do padrão corporal, para algo mais abrangente, como funções protetoras. Os genes de ALDH1/2 exibem uma funcionalidade não usual após as duplicações e assim são um ótimo modelo de estudo para duplicações gênica e evolução do mecanismo de vias metabólicas e sinalização celular.

CAPÍTULO 3: Análise da especificidade na ligação fatores de transcrição, regulados por ácido retinóico, ao DNA.

A grande quantidade de informações geradas com o seqüenciamento do genoma humano e de outros organismos tornou possíveis estudos mais completos sobre a distribuição e expressão gênica. O grande crescimento no depósito de seqüências, provenientes de projetos de seqüenciamento, nos bancos de dados públicos, tem auxiliado cada vez mais na caracterização dos genes, de suas formas de expressão e de suas funções. Apesar de milhões de pares de bases terem sido seqüenciados, os mecanismos bioquímicos de regulação da transcrição em eucariotos ainda não estão completamente elucidados. Neste sentido, a bioinformática se mostra cada vez mais importante na análise dessa enorme quantidade de dados, uma vez que possibilita análises mais rápidas e de menor custo em estudos de genomas, transcriptomas e proteomas.

Um gene pode ser definido como uma região de um DNA de uma espécie que produz uma molécula funcional em algum estágio do ciclo celular. Em organismos eucariotos, os genes que codificam para proteínas são inicialmente transcritos em moléculas de RNA mensageiros (mRNA), para a sua posterior tradução numa proteína. Em um gene típico, o centro promotor está localizado adjacente e a montante do gene. No centro promotor, ligam-se a RNA-polimerase e os fatores acessórios ("maquinaria transcricional"). Normalmente, só o centro promotor não é o suficiente para iniciar a transcrição. Logo a montante do centro promotor se encontra a região promotora na qual se ligam proteínas chamadas de fatores de transcrição (Carey 2002). Esses fatores também se ligam às regiões chamadas *enhancer*, que se localizam a centenas ou até milhares de pares de base do início do gene, podendo também estar localizadas a jusante

do gene (Carey 2002)). Só depois que todos os fatores de transcrição e a maquinaria transcricional estiverem ligados, a transcrição se inicia.

Na síntese de RNA, em procariotos, a RNA-polimerase se liga precisamente ao promotor. A RNA-polimerase é uma enzima formada por 6 sub-unidades: 2 sub-unidades α, uma β e uma β', uma ω e a sub-unidade σ. Para o reconhecimento do promotor, a RNA polimerase utiliza a sub-unidade σ, que não se liga se fortemente às sub-unidades formadoras do núcleo enzimaticamente ativo $\beta'\beta\alpha 2\omega$. Assim que a RNA polimerase se liga ao promotor, a sub-unidade σ se desliga do núcleo enzimático, que prossegue na síntese do RNA até alcançar um terminador (Lewin 2001). A seqüência do promotor não é única, mas existe um consenso. As seqüências de consenso de promotores da bactéria *Escherichia coli*, chamada caixa de Pribnow, (ou caixa TATA) e caixa -35 são muito conservadas. Pequenas variações nessas seqüências podem reduzir drasticamente a afinidade do fator σ da RNA polimerase pelo promotor (Lewin 2001).

A capacidade de um fator de transcrição se ligar seletivamente a sítios particulares no genoma é o principal fundamento sobre o qual as vias de regulação transcricional são montadas. Nossos conhecimentos de como essas proteínas interagem com as bases do DNA aumentaram muito nos últimos 25 anos (Luscombe, Austin et al. 2000). Os fatores de transcrição funcionam, em sua maioria, como complexos heteroligoméricos, ou seja, para efetuar sua função regulatória necessitam estar acoplados uns aos outros (Carey 2002). Os fatores de transcrição reconhecem uma seqüência particular a partir da complementaridade espacial de suas cadeias laterais com a estrutura tridimensional do DNA. Cada sítio de ligação geralmente é composto por 10 a 30 nucleotídeos e possui uma região de alta conservação, conhecida como centro da seqüência, imprescindível na montagem do complexo DNA-proteína. Todos os sítios já reconhecidos como tendo interação com uma proteína de transcrição em particular

possuem centros de seqüência similares (Werner, Fessele et al. 2003). Baseado nessa alta conservação é possível definir os fatores de transcrição que atuam na regulação de um gene, a partir do estudo da presença desses sítios em uma seqüência de interesse no DNA (Fessele, Maier et al. 2002). Quando considerado um gene em particular, a região de 2kb a montante do primeiro éxon deve ser analisada preferencialmente, pois possui aproximadamente 85% dos sítios dos fatores de transcrição atuantes (Iwama and Gojobori 2004). Os sítios de ligação e os fatores que se ligam a eles constituem, em conjunto com eventos de regulação pós-transcricional, a base da regulação gênica. Em uma seqüência regulatória de um gene, normalmente, são encontrados dois ou mais sítios de ligação, definindo assim o conceito dos módulos regulacionais. Os sítios desses módulos interagem entre si antagonisticamente ou sinergisticamente para a expressão de um gene (Klingenhoff, Frech et al. 2002).

1.1 Análise de sítio de ligação de fatores de transcrição utilizando bases de dados

O aumento do conhecimento da relação entre sítios e fatores de transcrição conduziu em 1995 à publicação da primeira base de dados relacional conhecida como Transfac (Wingender, Dietze et al. 1996). Essa base de dados compila e organiza informações sobre o comportamento dos fatores de transcrição, os modos como estes se ligam aos seus sítios de ligação e em quais genes tais sítios já tiveram sua existência comprovada por ensaios bioquímicos. As matrizes para busca por sítios de transcrição são construídas alinhando todos os sítios de interação com um dado fator de transcrição (Quandt, Frech et al. 1995). A partir desse alinhamento são calculados os valores de conservação de cada uma das posições do sítio de interação e o valor proporcional de conservação de cada nucleotídeo na mesma posição. Esta informação gera uma matriz de valores relativos para cada posição e para cada um dos quatro nucleotídeos, que

representa, em última análise, o padrão de conservação desses sítios. Essa abordagem é extensamente usada por algoritmos de busca de sítios (Cartharius, Frech et al. 2005). As matrizes mostram-se uma boa ferramenta para encontrar as seqüências que possuam sítios de ligação para os fatores de transcrição. Porém, devido a seu caráter pouco restritivo acabam gerando um grande número de sítios falso-positivos.

1.2 Proposição de uma abordagem alternativa para análise de sítio de ligação de fatores de transcrição utilizando estruturas tridimensionais de macromoléculas biológicas

A utilização de métodos computacionais que não necessitam a priori da informação a respeito dos sítios de ligação, bem como as que utilizam parâmetros mais eficientes para a detecção dos resultados, do que escores de pontuação provindos de alinhamentos, podem ser bastante úteis. Nessa direção, esse estudo desenvolve um método no qual são utilizadas as estruturas tridimensionais de fatores de transcrição, completas ou parciais, e do DNA, é realizada uma varredura de todas as possíveis ligações entre o fator de transcrição e o DNA, assim cobrindo todos os possíveis sítios de ligação entre as proteínas e o DNA. Esses complexos são submetidos a processos de mecânica molecular, como minimização de energia e dinâmica molecular. Após essas etapas, é medida a energia de interação entre o DNA e os fatores de transcrição. A partir da energia de interação de todos os possíveis sítios de interação, é possível tentar prever as regiões mais prováveis de interação entre os fatores de transcrição e o DNA.

1.3 Fatores de transcrição regulados por ácido retinóico

O modelo escolhido para análise de regiões promotoras foram os receptores nucleares regulados por ácido retinóico. Estes receptores foram escolhidos por terem as estruturas tridimensionais dos domínios de ligação ao DNA elucidadas

experimentalmente em complexo com o DNA (Rastinejad, Wagner et al. 2000). Outro motivo que levou à escolha destes receptores foi o fato de existirem trabalhos mostrando sua interação com diversas seqüências reguladoras, como a seqüência regulatória RARE-β2, sabida experimentalmente se ligar aos receptores RXR e RAR com cinco bases separando os sítios de ligação (Zelent, Mendelsohn et al. 1991). Utilizamos o modelo experimental descrito no trabalho de Vivanco Ruiz *et al.* (1991), no qual foi realizado um procedimento para verificar qual a mais provável região para ligação dos fatores de transcrição. Vivanco Ruiz utilizou a seqüência nativa da região promotora do gene RARE-β2 ligada a um gene repórter. Esse conjunto foi inserido em células e posteriormente a expressão do gene repórter foi quantificada. Para tentar prever quais regiões eram importantes para a ativação da transcrição Vivanco Ruiz realizou uma varredura com trinca de mutações em RARE-β2 e repetindo o mesmo procedimento da seqüência nativa. Dependendo da mutação a atividade diminuía expressivamente, assim podendo inferir que a região da mutação está relacionada a ligação dos fatores de transcrição. Utilizamos as seqüências descritas no trabalho de Vivanco Ruiz para nossas análises.

2. Método

2.1 Predição de sítios de ligação baseado na seqüência nucleotídica

Foi utilizado o programa Match Search (Kel, Gossling et al. 2003) (www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi) para fazer uma busca de sítios de ligação das seqüências nucleotídicas utilizadas nesse estudo. Foram utilizados como parâmetros de corte a similaridade da seqüência com a matriz e com o core. Os sítios de ligação que apresentaram valores de similaridade com a matriz e núcleo acima

de 0,7 e 0,75, respectivamente, foram aceitos. A seqüência da região promotora do gene RARE-β2 foi utilizada na busca por sítios:

Nativa CGGGTAGGGTTCACCGAAAGTTCACTCGCATATA

2.2 Método para busca por sítio de ligação de fatores de transcrição

A busca por fatores, baseada em estruturas tridimensionais, tem despertado um grande interesse por ser uma possível solução para a busca de sítios (Liu and Bader 2007). Neste trabalho foi desenvolvido um método no qual são construídos os complexos de forma a fazer uma varredura de todas as possíveis ligações entre os receptores nucleares e o DNA. Esses complexos são submetidos a minimizações de energia e dinâmica molecular e após esses passos são medidas as energias de interação entre a proteína e o DNA. Na Figura 26 tem um esquema do método desenvolvido.

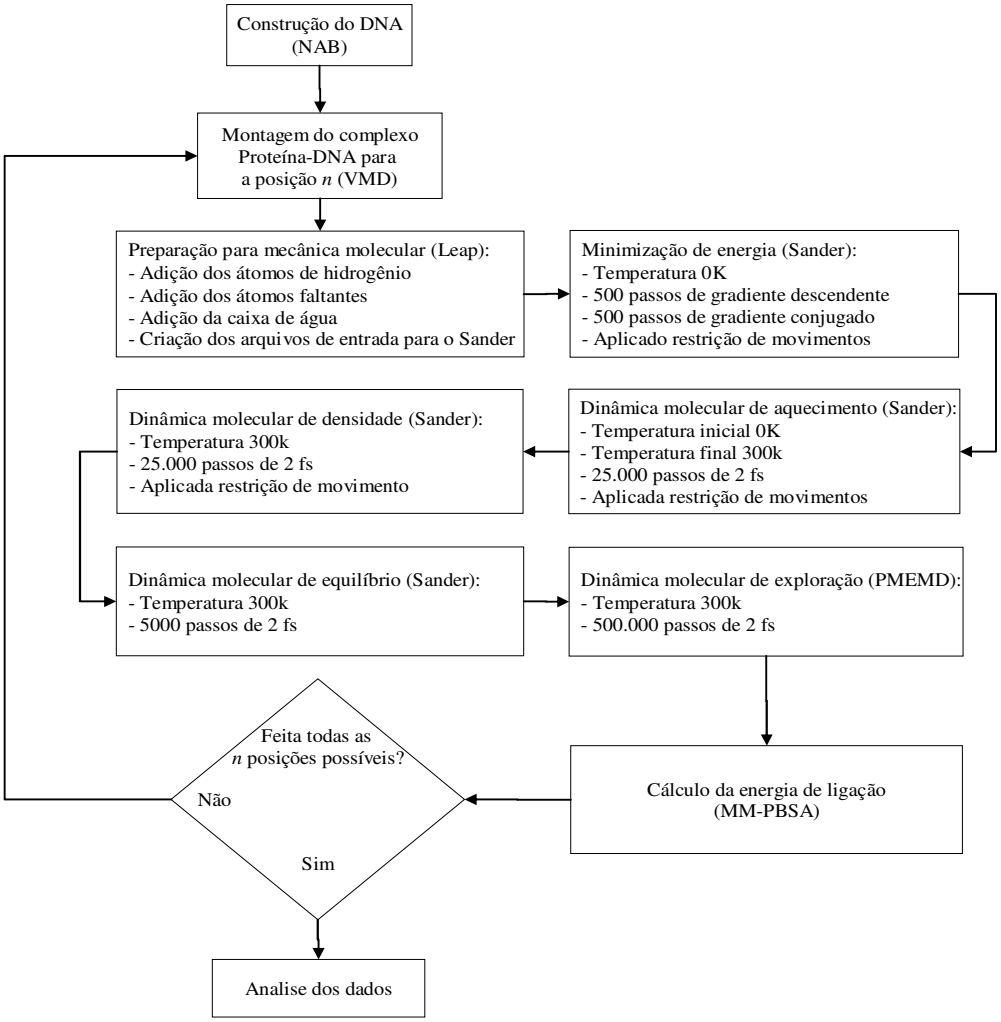


Figura 26: Esquema do processo para o cálculo da energia de ligação entre DNA e proteína.

2.3 Construção dos complexos

2.3.1 Construção do DNA

As estruturas das moléculas de DNA e dos complexos são construídas utilizando uma estrutura predita de um DNA do tipo B, construída utilizando o software NAB (Nucleic Acid Builder, (Case D.A. , Luo et al. 2006), utilizando a seqüência nucleotídica da região a ser analisada. As seqüências do trabalho de Vivanco Ruiz *et al.*

(1991), foram construídas, em negrito destacamos as mutações realizadas no trabalho de Vivanco Ruiz.

```
Nativa CGGGTAGGGTTCACCGAAAGTTCACTCGCATATA

    m2 CGGG**GCT**GGTTCACCGAAAGTTCACTCGCATATA
    m3 CGGGTAG**AAC**TCACCGAAAGTTCACTCGCATATA
    m4 CGGGTAGGGT**CTG**CCGAAAGTTCACTCGCATATA
    m5 CGGGTAGGGTTCA**TAT**AAAGTTCACTCGCATATA
    m6 CGGGTAGGGTTCACCG**CGC**GTTCACTCGCATATA
    m7 CGGGTAGGGTTCACCGAAA**TCG**CACTCGCATATA
    m8 CGGGTAGGGTTCACCGAAAGTT**ACT**TCGCATATA
    m9 CGGGTAGGGTTCACCGAAAGTTCAC**GAT**CATATA
```

2.3.2 Posicionamento das proteínas sobre as hexades e preparação para minimização.

O posicionamento das estruturas dos receptores nucleares sobre o DNA é realizado através de sobreposições das coordenadas dos átomos de fósforo do DNA da estrutura cristalográfica com os respectivos átomos de fósforo das estruturas dos DNAs construídos, utilizando o software LSQMAN (Novotny, Madsen et al. 2004). Assim, por exemplo, para a construção do complexo RXR-DNA, utilizou-se as estruturas 3D do RXR ancorados a um fragmento de DNA (DNA-1DSZ), disponível na estrutura cristalográfica de PDBid 1DSZ (Figura 27 A). Para a ancoragem do RXR, na posição 3` do DNA, os átomos de fósforo dos resíduos 1503 ao 1508 do DNA-1DSZ é sobreposto estruturalmente com os átomos de fósforo dos resíduos 3 a 8 do DNA (Figura 27 B, C e D). Esse mesmo procedimento é repetido para todas as posições de interesse, permitindo a varredura de todas as possíveis ligação.

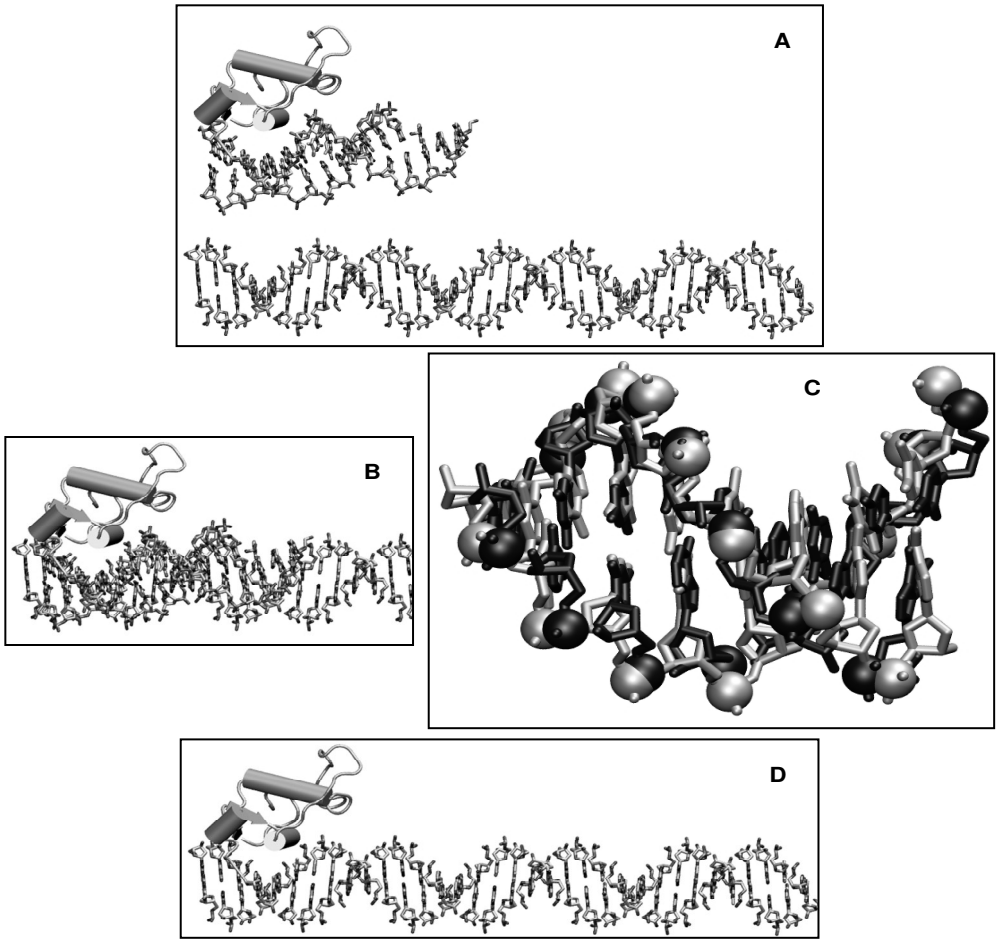


Figura 27: Processo da sobreposição estrutural do RXR sobre o DNA a ser analisado. Em A, na parte superior a estrutura tridimensional do RXR e o DNA da estrutura cristalográfica (PDBid 1DSZ) na parte inferior a estrutura tridimensional do DNA a ser analisado. Em B, a sobreposição dos DNAs. Em C, uma vista detalhada da sobreposição dos DNAs, em destaque os átomos de fósforo. Em D, o RXR sobreposto no DNA a ser analisado.

Neste estudo será utilizado como posição no DNA a posição da primeira base nucleotídica da hexade do RXR. É importante salientar que uma posição no DNA representa pelo menos 17 bases do DNA sob os fatores de transcrição. Na Figura 28 um exemplo do complexo na posição 6 do DNA. A hexade do RXR vai da base 6 a base 11, a região espaçadora vai da base 12 a 16 e a hexade do RAR vai da base 17 a 22. Então, quando falar de uma posição do DNA deve-se imaginar a 17 bases sob influência dos fatores de transcrição.

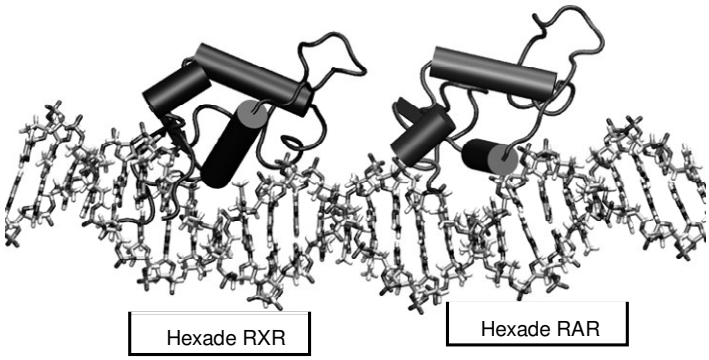


Figura 28: Estrutura do complexo RXR/RAR- RARE-β2 na posição 6 do DNA mostrando que uma posição do DNA representam 17 bases sob influência dos fatores de transcrição.

Todas as estruturas dos complexos são preparadas para minimização utilizando o software Leap do pacote do AMBER 9 (Case D.A. , Luo et al. 2006) onde todos os átomos de hidrogênio e outros átomos faltantes são construídos e também é criada uma caixa de água com 12 Å de distância dos átomos externos. Os parâmetros do campo de força para proteína e para o DNA foram o do parm99 do pacote do AMBER9.

2.3.3 Minimização de energia e Dinâmica molecular

A minimização de energia e as dinâmicas moleculares são conduzidas utilizando o software Sander e PMEMD do pacote AMBER 9. A distância de corte para o cálculo de componentes de energia relacionados com átomos não ligados é de 12 Å.

As estruturas dos complexos são submetidas a uma curta minimização de energia, para eliminar contatos de van der Walls inadequados. Utilizamos 500 passos de Gradiente Descendente e até 500 passos de Gradiente Conjugado. São utilizados passos de 2 fs e o método “Shake” (Ryckaert, Ciccotti et al. 1977; Miyamoto and Kollman 1992) sobre todos os átomos de hidrogênio, este método é utilizado em todas as etapas de mecânica molecular. Foi utilizado também uma restrição de movimento sobre todos

os átomos (proteína e DNA) para evitar movimentos muito amplos que poderiam ser causado por alguma água posicionada muito próxima de algum átomo do complexo peptídeo-proteína.

Uma dinâmica molecular é realizada para elevar a temperatura de 0K para 300K e também eliminar contatos de van der Walls inadequados e distorções na estereoquímica envolvendo comprimento de ligações atômicas e ângulos entre átomos ligados fora dos padrões. São feitos 25000 passos de 2 fs utilizando uma restrição de movimento sobre todos os átomos (proteína e DNA) para evitar movimentos bruscos que poderiam ser causado por alguma água posicionada muito próxima de algum átomo do complexo peptídeo-proteína.

Uma segunda dinâmica molecular, mantendo a restrição de movimento sobre todos os átomos (proteína e DNA), é realizada para atingir o equilíbrio na densidade da caixa de água. São realizados 25.000 passos de 2 fs a 300 K.

Segue-se então uma dinâmica molecular de 5.000 passos de 2 fs a 300 K, para equilibrar o sistema. Finalmente, realiza-se uma dinâmica molecular de 1 ns (500.000 passos de 2 fs) a 300 K, para exploração conformacional. A cada 5000 passos é salva uma conformação para análise posterior. Nessa última etapa foi utilizado o programa PMEMD do pacote do Amber 9.

2.3.4 Conceito

O cálculo da energia de ligação entre os DNA e as proteínas é realizado utilizando a técnica de dinâmica molecular acoplada aos métodos de Poisson-Boltzmann MM-PBSA (Luo, David et al. 2002) e método generalizado de Born MM-GBSA (Case D.A. , Luo et al. 2006), que calculam a diferença de energia livre entre duas condições (ligada e não ligada) de duas moléculas.

Na Figura 29 temos um esquema do ciclo termodinâmico que explica como o cálculo da energia de ligação é realizado.

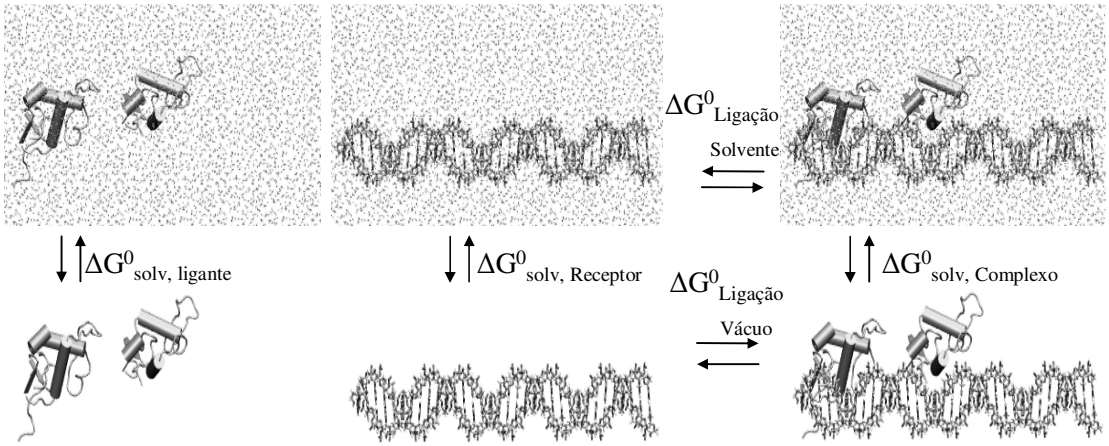


Figura 29: Esquema do clico termodinâmico utilizado para calcular a energia de ligação entre o DNA e os fatores de transcrição.

De acordo com o ciclo termodinâmico acima podemos calcular a energia de ligação da seguinte maneira:

$$\Delta G^0_{\text{Ligação, Solvente}} = \Delta G^0_{\text{Ligação, Vácuo}} + \Delta G^0_{\text{Solv, Complexo}} - (\Delta G^0_{\text{Solv,Ligante}} + \Delta G^0_{\text{Solv, receptor}})$$

No método MM-PBSA as diferentes contribuições para a energia livre de ligação são calculadas como descritas a seguir. As energias livres de solvatação são calculadas pela equação linearizada de Poisson Boltzmann para as três condições (apenas o ligante, apenas o receptor e o complexo receptor/ligante). Este método calcula a contribuição da energia eletrostática para a energia livre de solvatação a adiciona um termo empírico para a contribuição hidrofóbica:

$$\Delta G^0_{\text{Solv,Ligante}} = G^0_{\text{eletrostatica,}\epsilon=80} - G^0_{\text{eletrostatica, }\epsilon=1} + \Delta G^0_{\text{hidrofobico}}$$

$$\Delta G^0_{\text{Solv,Receptor}} = G^0_{\text{eletrostatica,}\epsilon=80} - G^0_{\text{eletrostatica, }\epsilon=1} + \Delta G^0_{\text{hidrofobico}}$$

$$\Delta G^0_{\text{solv,Complexo}} = G^0_{\text{eletrostatica,}\epsilon=80} - G^0_{\text{eletrostatica, }\epsilon=1} + \Delta G^0_{\text{hidrofobico}}$$

O $\Delta G^0_{\text{Ligação, vácuo}}$ é obtido calculando a média da energia de interação entre o receptor e o ligante e utilizando a variação da entropia causada pela ligação se for desejado:

$$\boxed{\Delta G^0_{\text{Ligação, vácuo}} = \Delta E_{\text{mecânica molecular}} - T \cdot \Delta s_{\text{normal análise}}}$$

A contribuição da entropia pode ser calculada realizando uma análise de modo normal nas três estruturas (DNA, proteína e DNA/proteína), mas na prática a contribuição da entropia pode ser negligenciada, pois os DNAs serão comparados com as mesmas proteínas e também essa análise é custosa computacionalmente e possui uma grande margem de erro causando incertezas nos resultados assim optou-se por não utilizar a contribuição entrópica.

A média das energias de interação, entre o receptor e o ligante, é obtida realizando os cálculos de "snapshots" não correlatos coletados de uma dinâmica molecular em equilíbrio. O ideal para realizar o cálculo de energia de ligação seria realizar a dinâmica molecular das três estruturas, mas para otimizar o processo, admitindo que as mudanças na proteína livre e no DNA livre não são muito grandes é realizada apenas a dinâmica molecular do complexo DNA-proteína.

2.4 Cálculo da energia de ligação

O cálculo da energia de interação dos complexos é realizada utilizando-se as trajetórias resultantes dos procedimentos de dinâmicas moleculares. O software MMPBSA do pacote do AMBER, que implementa o método de Poisson-Boltzmann, é utilizado para calcular a energia de interação. A análise da energia de interação entre o DNA e as proteínas dos complexos é realizada utilizando-se as conformações espaciais obtidas ao longo da dinâmicas moleculares. O cálculo da energia de interação ($\Delta\Delta E$) é feito da seguinte forma: mede-se a energia livre do DNA isolado (A), a energia livre dos

receptores nucleares isolados (B) e a energia livre de todo o complexo (AB). A energia de interação é dada por:

$$\Delta\Delta E_{int} = \Delta E_{AB} - (\Delta E_A + \Delta E_B)$$

As medições da energia são realizadas em todas as posições ao longo de toda a dinâmica.

2.4.1 Análise das energias obtidas

Para analisar os dados das energias obtidas foi utilizado o programa R (http://www.R-project.org 2005). A primeira análise foi comparar a média, mediana e menor energia para verificar se havia correspondência com os dados obtidos por Vivanco Ruiz *et al.*.

Neste trabalho também foi utilizada uma segunda abordagem, onde é levada em conta a estabilidade energética do sistema. A utilização de uma varredura utilizando uma janela temporal das energias calculadas a partir das estruturas coletadas durante a dinâmica molecular de exploração conformacional, permite a escolha de um critério termodinâmico onde uma maior estabilidade do sistema por um maior período de tempo recebe uma classificação melhor do que um sistema que apresentou menores energias mas com grande variação. Partindo de um conjunto de energias amostradas em intervalos de tempo consecutivos de 10 ps, onde foram coletadas 100 valores de energias de interação, foram utilizadas janelas variando de 20 a 90 intervalos de tempo, coletando-se parâmetros estatísticos para cada janela, para cada complexo proteína-DNA analisado. Detalhadamente, esse método é realizado da seguinte ordem: inicia-se colocando uma janela sobre os primeiros 20 valores de energia, então é calculada a energia média, o desvio padrão e z-escore (valor absoluto da média/desvio padrão). Desloca-se a janela, passando para o próximo ponto temporal e coletam-se os

parâmetros estatísticos como citados acima, procedendo-se assim até a última posição do vetor de energias apropriada, cujo índice é igual ao último índice do vetor menos o comprimento da janela. Após atingir o último valor de energia apropriado, a janela é aumentada em uma unidade, repetindo-se o procedimento anterior. Esse procedimento é realizado até atingir uma janela de comprimento 90. A partir desses dados foi selecionada a janela que melhor parametriza a interação proteína-DNA. Para tal, a partir dos dados calculados de todas as janelas da varredura, foram utilizadas quatro abordagens diferentes de ordenação das janelas temporais analisadas:

- primeiro pela menor média de energia, segundo pelo maior tamanho da janela temporal e por último o maior z-escore (modo Media Janela Zscore - MJZ).
- primeiro pela menor média de energia, segundo pelo maior z-escore e por último pela maior tamanho da janela temporal (modo Media Zscore Janela - MZJ).
- primeiro pelo maior z-escore, segundo maior tamanho da janela temporal e por último menor média de energia (modo Zscore Janela Media ZJM).
- primeiro pelo maior z-escore, segundo menor média de energia e por último maior tamanho da janela temporal (modo Zscore Media Janela ZMJ).

A Figura 30 apresenta um exemplo dos quatro modos para definir a melhor estimativa da energia de interação utilizando a varredura por janela temporal.

A	6
1	-164,85
2	-187,07
3	-166,76
4	-171,33
5	-172,83
6	-185,26
7	-162,59
8	-175,08
9	-170,01
10	-175,74
11	-167,74
12	-164,69
13	-164,66
14	-153,99
15	-165,07
16	-164,89
17	
18	-163,15
19	-153,47
20	-158
21	-167,04
22	-162,36
23	-166,33
24	-171,2
25	-152,7
26	-153,88
27	-150,41
28	-147,92
29	-152,9
30	-120,78
31	-143,83
32	-121,73
33	-130,14
34	-129,37
35	-130,65
36	-124,62
37	-105,17
38	-136,59
39	-125,42
40	-150,87

B

		3	2		1
inicio	fim	tamanho da janela	media	sd	Zscore
7	26	20	-163	6	23
3	24	22	-166	7	22
3	23	21	-166	7	22
4	24	21	-166	7	22
7	27	21	-163	7	22
3	25	23	-166	7	21
4	25	22	-166	7	21
3	22	20	-166	7	21
4	23	20	-166	7	21
5	24	20	-166	7	21

		2	3		1
inicio	fim	tamanho da janela	media	sd	zscore
7	26	20	-163	6	23
3	24	22	-166	7	22
3	23	21	-166	7	22
4	24	21	-166	7	22
7	27	21	-163	7	22
3	25	23	-166	7	21
4	25	22	-166	7	21
3	22	20	-166	7	21
4	23	20	-166	7	21
5	24	20	-166	7	21

		2	1		3
inicio	fim	tamanho da janela	media	sd	zscore
1	24	24	-167	8	20
2	25	24	-167	8	19
1	23	23	-167	8	20
2	24	23	-167	8	20
1	22	22	-167	8	19
2	23	22	-167	8	19
1	21	21	-167	8	19
2	22	21	-167	8	19
1	20	20	-167	8	18
2	21	20	-167	8	18

		3	1		2
inicio	fim	tamanho da janela	media	sd	zscore
1	24	24	-167	8	20
1	23	23	-167	8	20
2	24	23	-167	8	20
2	25	24	-167	8	19
1	22	22	-167	8	19
2	23	22	-167	8	19
1	21	21	-167	8	19
2	22	21	-167	8	19
1	20	20	-167	8	18
2	21	20	-167	8	18

Figura 30: Exemplo mostrando os quatro modos de escolha do valor médio da energia de interação. Em A) as primeiras 40 energias da posição 6 do DNA nativo. Em B) os quatro modos de seleção da melhor janela. Os números acima de cada tabela correspondem a ordem da ordenação dos dados. As janelas ordenadas primeiro por z-escore apresentam energias maiores do que as janelas ordenadas primeiro pela média mas apresentam menor desvio padrão. A comparação das quatro abordagens com os dados experimentais de Vivanco Ruiz *et al.* (1991) irá definir qual é o melhor estimador para busca de sítios.

Após todos os DNAs serem analisado, foram criados gráficos comparativos entre todos os DNAs, agrupados pelos tamanhos mínimos da janela (20, 30, 40 até 90) (Figura 31). O valor da energia utilizada no gráfico é o valor obtido na janela melhor classificada.

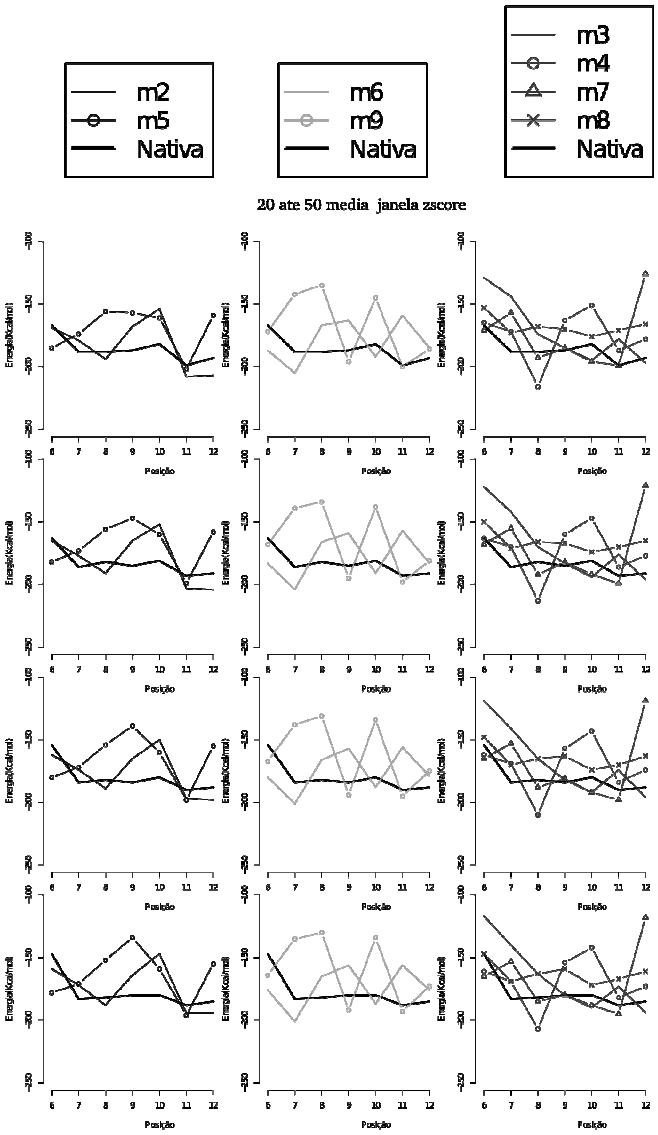


Figura 31: Gráficos comparando as energias de ligação média entre todos os DNAs analisados. Os gráficos à esquerda são das mutações (m2 e m5) que não alteraram a expressão do gene repórter nos experimentos de Vivanco Ruiz *et al.* Os gráficos centrais são das mutações que diminuiram a expressão do gene repórter (m6 e m9) e os gráficos à direita são das mutações que inibiram a expressão do gene repórter (m3, m4, m7 e m8). Em todos os gráficos é apresentada a energia do complexo proteína-DNA nativo para facilitar a comparação. A primeira trinca de gráficos são referentes aos dados obtidos a partir de janelas temporais de tamanho mínimo de 20 energias, a segunda trinca de gráficos janelas temporais de 30 energias, assim até energias 50.

Os gráficos apresentados na Figura 31 foram repetidos para os quatro modos de análise das janelas temporais. A partir destas informações, tentou-se prever qual seria o melhor estimador para definir as regiões de maior chance da ligação de fatores de transcrição ao DNA.

Nas análises das energias de ligação foram retiradas todas as energias que distanciaram uma vez e meia ou mais da mediana de cada conjunto de pontos.

2.5 Análise das ligações de hidrogênio

Para uma melhor compreensão das fontes de variação dos valores de energia medidos no método desenvolvido, foram analisadas as ligações de hidrogênio que ocorrem entre o DNA e as proteínas para verificar se existia uma correlação. Para tal, foi desenvolvido um script em TCL (http://www.tcl.tk/) que possibilitou, juntamente com o programa VMD (Humphrey, Dalke et al. 1996), a obtenção de todas as ligações de hidrogênio, com a distância máxima de 3 Å e ângulo máximo de 30º, formadas entre os átomos do DNA e os átomos das proteínas. Para analisarmos todas as ligações em conjunto, definimos as hexades de DNA (conjuntos de seis nucleotídeos conectados) como referencial. As posições da hexade do RXR foram chamadas de P1D, P2D, P3D, P4D, P5D e P6D para a fita direta e P1R, P2R, P3R, P4R, P5R e P6R para a fita reversa. As posições entre a hexade do RXR e RAR foram chamadas de S1D, S2D, S3D, S4D e S5D para a fita direta e S1R, S2R, S3R, S4R e S5R para fita reversa. As posições da hexade do RAR foram chamadas de P1D', P2D', P3D', P4D', P5D' e P6D' para a fita direta e P1R', P2R', P3R', P4R', P5R' e P6R' para fita reversa (Figura 32).

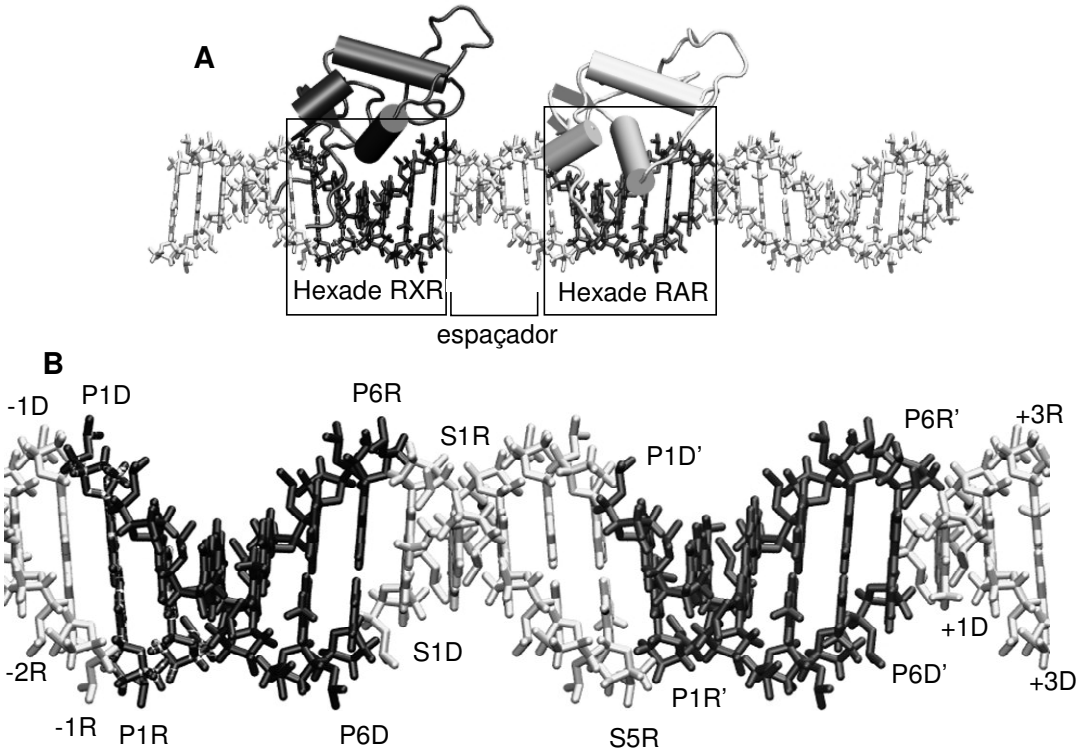


Figura 32: Esquema da nomenclatura relativa às posições das hexades de interação. Em A) vista geral da estrutura e as regiões das hexades e o espaçador. Neste exemplo a hexade do RXR inicia na posição 7 do DNA. Em B) uma visão detalhada das hexades e do espaçador com os nomes das bases na nomenclatura relativa à hexade.

As bases antes da hexade do RXR foram numeradas negativamente (de -1 a -16 D ou R relativo à fita direta ou fita reversa) e as bases após a hexade do RAR foram numeradas positivamente (de 1 a 15 D ou R relativo à fita direta ou fita reversa).

Para verificar se há especificidade nas ligações entre o DNA e a proteína, foi realizada uma análise na qual separamos os átomos do DNA em dois grupos: os fosfatos mais os átomos do açúcar (fosfoaçúcar FA) e os átomos da base nitrogenada. As ligações com o fosfoaçúcar são consideras inespecíficas e as ligações com os átomos da base nitrogenada são consideradas específicas.

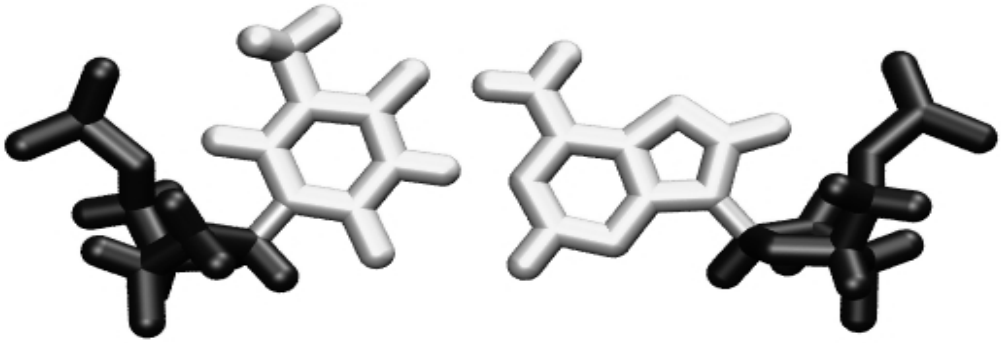


Figura 33: Esquema de um par de nucleotídeos, Adenina e Timina, onde os átomos da base estão em branco e os átomos do fosfoaçúcar estão em preto.

As informações referentes às ligações de hidrogênio foram armazenas em uma base de dados para simplificar a consulta e a quantificação das ligações (Figura 34).

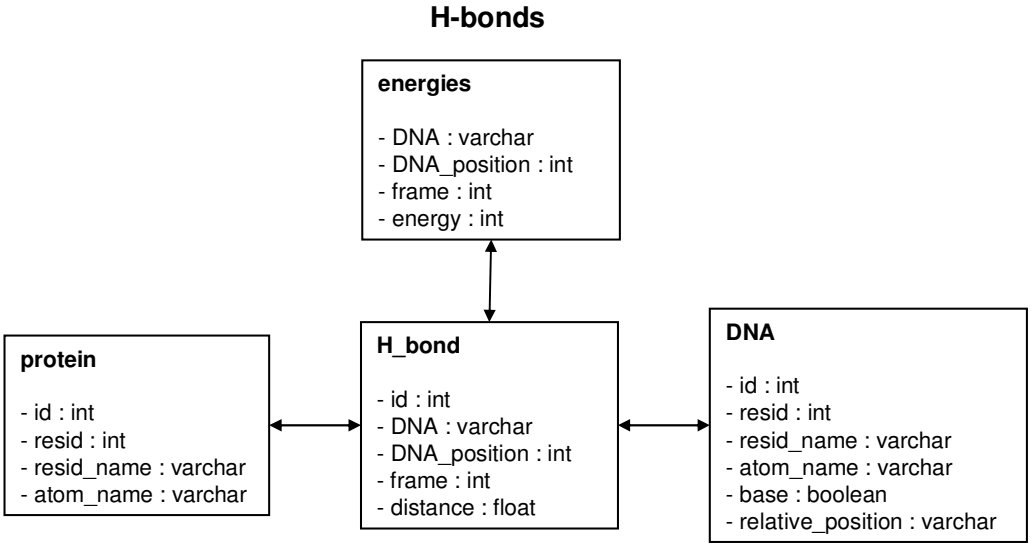


Figura 34: Representação esquemática da base de dados relacional com as informações das ligações de hidrogênio.

A relação H_bonds foi modelada para ter o atributo *id* (identificador interno de H-bonds) e os seguintes atributos: *DNA* (nome do DNA utilizado ex. m2, m3 ou nativo), *DNA_position* (número do resíduo da primeira base da hexade do RXR), *frame* (número da estrutura retirada da dinâmica molecular) e *distance* (distância entre os átomos da ligação de hidrogênio). Outra relação presente nesta base de dados é a *protein*, que

possui os atributos *id* (identificador da ligação), *resid* (número do resíduo no PDB), *resid_name* (nome do resíduo) e *atom_name* (nome do átomo). A relação DNA foi modelada com os seguintes atributos: *id* (identificador da ligação), *resid* (número do resíduo no PDB), *resid_name* (nome do resíduo), *atom_name* (nome do átomo), *base* (0 o átomo faz parte do fosfoaçucar 1 o átomo faz parte da base) e *relative_position* (nome do resíduo na nomenclatura da posição relativa a hexade). A relação *energies* foi modelada com os seguintes atributos: *DNA* (nome do DNA utilizado ex. m2, m3 ou nativo), *DNA_position* (número do resíduo da primeira base da hexade do RXR), *frame* (número da estrutura retirada da dinâmica molecular) e *energy* (energia de interação calculada).

Para verificar se realmente ocorre uma preferência de um aminoácido por uma base foi realizado um teste de qui-quadro de Pearson e calculados os resíduos (tabela 12). Valores de resíduo, em módulo, maior que 2 significam que o aminoácido apresenta uma preferência (valores positivos) ou uma preterência (valores negativos) a uma determinada base.

3. Resultados

Neste trabalho, foram construídos 71 complexos proteína-DNA que foram submetidos a dinâmicas moleculares de 1.000.000 femtosegundos cada, totalizando 71 nanosegundos, ou seja foi avaliado um espaço amostra de 71.000.000 possíveis conformações. De cada dinâmica, foram coletadas 100 estruturas, espaçadas temporalmente, a cada 10.000 femtosegundos, totalizando 7100 estruturas. As análises deste trabalho foram baseadas nessa amostragem de dados.

A validação do método desenvolvido para a predição de sítios de ligação de fatores de transcrição baseado em métodos estruturais e termodinâmicos é baseada no

modelo de ligação dos fatores de transcrição RXR e RAR à região regulatória RARE-β2 , que já foi muito bem estudado experimentalmente e tem muito bem definido os possíveis sítios de ligação (Vivanco Ruiz, Bugge et al. 1991; Zelent, Mendelsohn et al. 1991). Vivanco Ruiz *et al.* (1991) estudaram a região regulatória RARE-β2 demonstrando, por experimentos de *footprinting,* que uma região de aproximadamente 25 pares de base (pb) é responsável pela interação com o heterodimero RXR/RAR. Para verificar a exata localização dos sítios de interação, estes autores geraram nove seqüências dessa região, fazendo trincas de mutações ao longo dela. A seqüência nativa da região e mais as nove seqüências mutadas foram utilizadas em um experimento onde é quantificada a transcrição de um gene repórter, quando cada uma das regiões é inserida individualmente a montante desse gene (Figura 35).

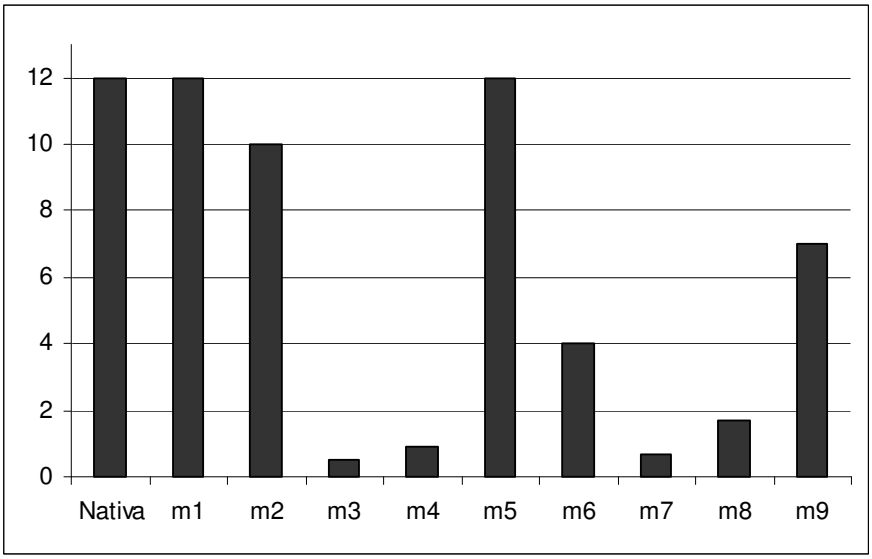


Figura 35: Dados obtidos por Vivanco Ruiz *et al*. mostrando os efeitos de mutações na expressão de um gene repórter ligado a montante da região promotora do gene RARE-ß2 (baseado em Vivanco Ruiz *et al.* 1991).

Os resultados de Vivanco Ruiz *et al.* (1991) mostraram que as mutações m3, m4, m7 e m8 afetam drasticamente a expressão do gene repórter. Essas informações

sugerem que os fatores de transcrição se ligam sobre as posições 8 a 13 e 19 a 24 (Figura 36).

```
1   5    10   15   20   25   30
|   |    |    |    |    |    |
CGGGTAGGGTTCACCGAAAGTTCACTCGCATATA
```

Figura 36: Seqüência da região promotora do RARE-ß2 definida no trabalho de Vivanco Ruiz *et al.* (1991). Em destaque as regiões que afetam a transcrição do gene repórter quando são mutadas.

Trabalhos como os de (Perlmann, Rangarajan et al. 1993; Heery, Pierrat et al. 1994; Zechel, Shen et al. 1994; Rastinejad, Wagner et al. 2000) definiram que a região localizada entre os sítios de ligação dos receptores RXR e do RAR podem ter de 1, 2 e 5 pb. Neste trabalho foram utilizados as seqüências nativa e as mutações m2 a m9 definidas em Vivanco Ruiz *et al.* (1991), sendo os sítios de ligação do RXR e do RAR separadas por 5 pb (DR5) (Figura 28).

3.1 Predição de sítios de ligação baseado na seqüência nucleotídica

Foi utilizado o programa Match Search, que compara as informações do banco de dados de fatores de transcrição TRANSFAC 7.0 (Kel-Margoulis, Kel et al. 2002), para tentar prever os sítios do RXR/RAR na seqüência do receptor RARE-β2. A análise da seqüência nativa identificou 120 regiões (

Tabela **8**) como possíveis sítios de ligação, mas nenhuma dessas regiões está relacionado com o RXR ou RAR.

Tabela 8: Lista dos fatores de transcrição preditos como possíveis ligantes na região promotora RARE-ß2.

Nome	Fita	Posição inicial	Nome	Fita	Posição inicial	Nome	Fita	Posição inicial
P	(-)	1	CdxA	(+)	9	XFD-3	(-)	15
Sp1	(+)	1	AbaA	(-)	9	AP-1	(-)	16
MZF1	(+)	1	C/EBPbeta	(+)	9	Nkx2-5	(+)	16
Pax-4	(-)	1	STATx	(+)	10	AP-1	(-)	16
RORalpha1	(+)	2	TCF11	(+)	10	FOXD3	(+)	16
HOXA3	(+)	3	STATx	(-)	10	c-Myb	(-)	16
STRE	(+)	4	RORalpha1	(-)	10	AbaA	(+)	16
STAT	(+)	4	Elk-1	(+)	10	GAmyb	(-)	16
Kr	(+)	4	c-Myb	(+)	10	C/EBP	(-)	16
XFD-3	(-)	4	USF	(+)	10	v-Myb	(-)	16
GAmyb	(-)	5	STATx	(-)	11	AP-1	(-)	16
c-Myb	(-)	5	MZF1	(-)	11	Pax-4	(+)	17
AP-1	(-)	5	NF-AT	(+)	11	Ik-1	(-)	17
AP-1	(-)	5	Tst-1	(-)	11	SRY	(+)	17
SRY	(-)	5	COUP-TF/HNF-4	(-)	11	En-1	(-)	17
v-Myb	(-)	5	ADR1	(-)	11	C/EBPalpha	(+)	18
AP-1	(-)	5	cap	(+)	11	Zeste	(-)	18
Pax-4	(-)	6	BR-C Z4	(+)	12	GCN4	(+)	19
GR	(-)	6	Barbie Box	(+)	13	HSF2	(-)	19
ER	(-)	6	Dof2	(+)	13	Hairy	(+)	19
Ik-1	(-)	6	PBF	(+)	13	HSF1	(-)	19
Pax-4	(+)	6	RORalpha1	(+)	13	CdxA	(+)	20
AREB6	(+)	6	Dof1	(+)	13	HSF	(-)	20
Ik-3	(-)	6	BR-C Z3	(-)	13	HSF	(-)	20
P	(-)	6	NRF-2	(+)	13	c-Myc/Max	(+)	20
C/EBP	(+)	7	Dof3	(+)	13	MATalpha2	(-)	20
CF1/USP	(+)	7	AML-1a	(-)	13	USF	(+)	21
C/EBPalpha	(+)	7	dl	(+)	13	TCF11	(+)	21
CF1/USP	(+)	7	AREB6	(-)	14	Hairy	(+)	21
C/EBPalpha	(-)	8	CdxA	(-)	14	Oct-1	(+)	21
C/EBP	(-)	8	CdxA	(-)	14	C/EBP	(-)	22
C/EBP	(-)	8	Lmo2 complex	(+)	14	StuAp	(+)	22
HSF	(-)	9	v-ErbA	(+)	14	cap	(+)	22
HSF	(-)	9	GATA-X	(+)	15	GCM	(+)	22
AREB6	(+)	9	HSF2	(-)	15	Nkx2-5	(-)	23
E74A	(+)	9	HSF	(+)	15	StuAp	(-)	24
deltaEF1	(+)	9	HSF1	(+)	15	StuAp	(+)	24
HNF-4	(+)	9	HSF	(+)	15	HSF	(-)	24
C/EBP	(-)	9	HSF1	(-)	15	HOXA3	(-)	26
C/EBPalpha	(-)	9	HSF2	(+)	15	NIT2	(-)	29

Foi possível observar que o programa de predição de sítios de ligação de fatores de transcrição não encontrou os fatores corretos, o que tornaria inviável qualquer estudo dos receptores RXR e RAR relacionado a seqüência promotora do gene RARE-ß2

baseado neste tipo de abordagem. Verificamos posteriormente que os fatores RXR e RAR ainda não foram descritos na base de dados do TRANSFAC 7.0 disponível publicamente.

3.2 Validação do método de busca por sítios de fatores de transcrição utilizando de estruturas tridimensionais e métodos termodinâmicos

O método desenvolvido foi utilizado para tentar prever as regiões de interação mais provável de fatores de transcrição na região promotora do gene RARE-β2. Foram construídos 71 complexos que possuem em média 4728 átomos, sendo em média 2571 átomos do DNA. Essa variação é devida a quantidade de átomos que eram alterados nas substituições dos nucleotídeos de cada mutação, 2157 átomos dos receptores protéicos RXR e RAR mais os átomos da caixa de água que variou de 61000 a 65000. As proteínas foram ancoradas da posição 3 à posição 17 no DNA nativo e da posição 6 a 12 nos DNAs mutados. A Figura 37 mostra uma visão esquemática do complexo na posição 6 após todas as etapas de mecânica molecular.

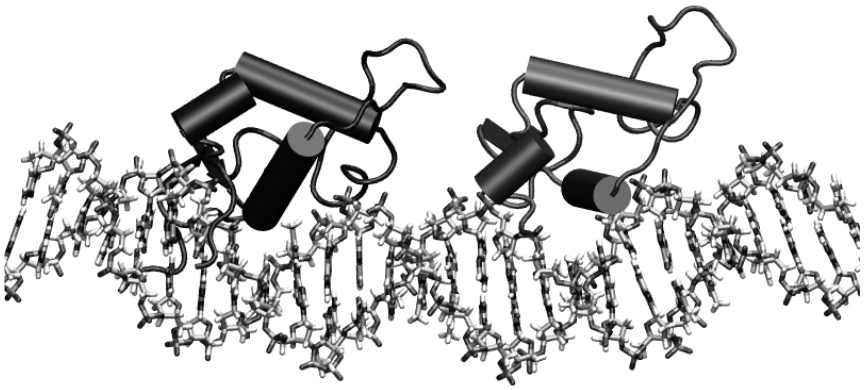


Figura 37: Estrutura do complexo RXR/RAR- RARE-ß2 na posição 6 do DNA, após a dinâmica molecular.

Após os processos de simulação por dinâmica molecular para a exploração do espaço conformacional do DNA nativo, foram calculadas as energias de interação. Para tentar verificar se existe correspondência entre os experimentos de Vivanco Ruiz *et al.* e os nossos dados foram utilizados descritores simples como a média das energias, a mediana e a menor energia (Figura 38).

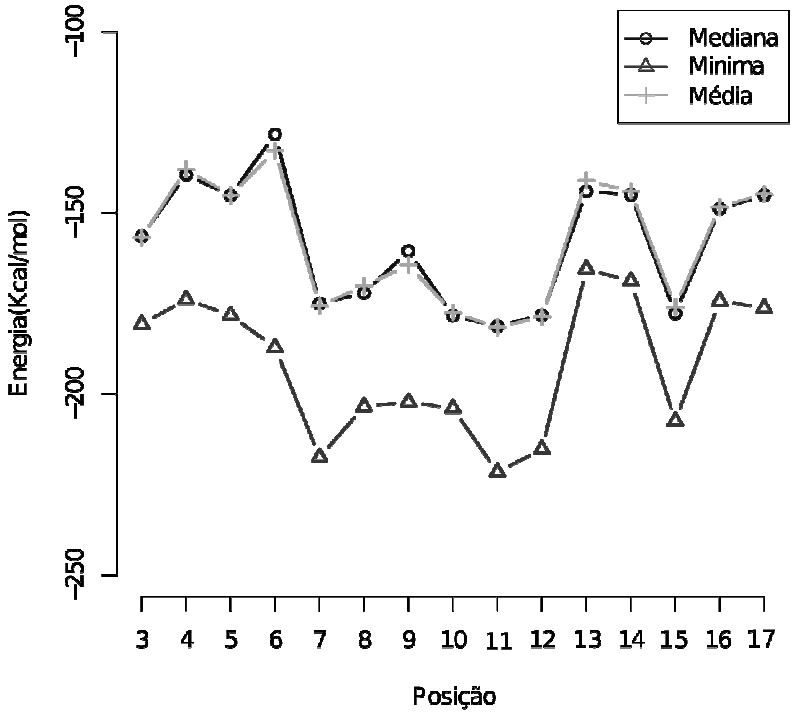


Figura 38: Gráficos das energias médias, medianas e mínimas nas 15 posições do DNA nativo.

Os valores obtidos com as menores energias sugerem que os melhores sítios de ligação do heterodímero RXR-RAR com da região promotora do gene RARE-β2 estão localizado nas posições 11, 12 e 7. Já os dados de média e mediana sugerem 11, 12, 10,15, 7 e 8.

Os resultados utilizando o método de janela temporal para o DNA nativo estão apresentados na Figura 39 e no anexo2, onde foram utilizados tamanhos de janelas mínimas variando de 20 a 90 energias. Os modos MZJ e MJZ apresentaram exatamente os mesmos valores de energia em todas as posições em todos os tamanhos de janela

mínima podendo ser analisados da mesma forma, para simplificar foi chamado de modo M . Os modos ZJM e ZMJ apresentaram diferenças muito pequenas, passando a chamar modo Z para simplificação.

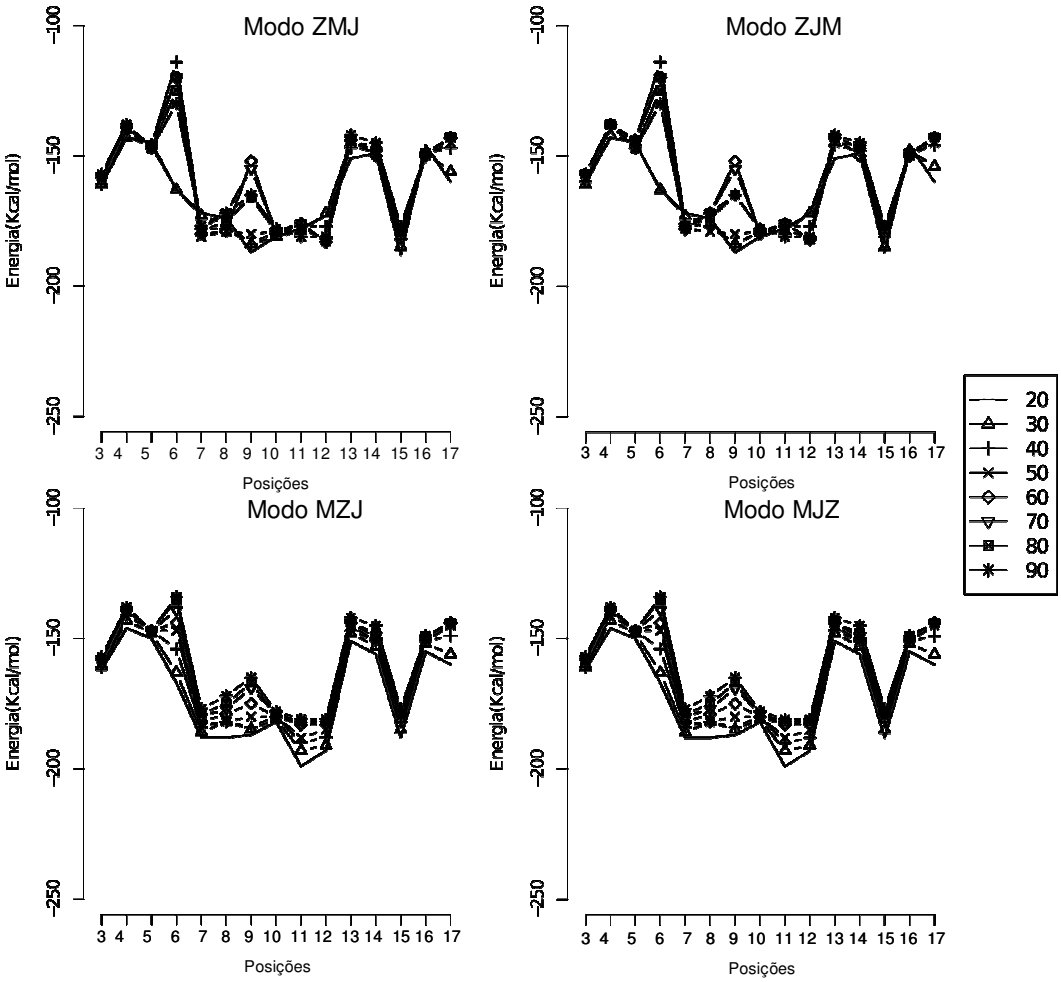


Figura 39: Os quatros modos de análise utilizando o método de janela temporal para o DNA nativo. Na parte superior da figura a legenda para todos os gráficos.

Observou-se que ocorre uma grande variação nos valores de energia nas posições 6 e 9 para diferentes janelas utilizadas no modo Z. Essa variação nos modos M é menor. Percebeu-se também, que as energias do modo Z são maiores que as do modo M. O modo Z sugere que entre a posição 6 à posição 12 e a posição 15 são as regiões

mais prováveis de interação com o fator de transcrição. O modo M também sugere as mesmas regiões como sendo as mais prováveis.

Os dados dos complexos com os DNAs mutados também foram analisados utilizando o método de janelas temporais. Os dados de Vivanco Ruiz *et al.* (1991) sugere que os efeitos das mutações podem ser analisadas utilizando apenas as posições 6 a 12, sendo assim, apenas foram realizados os procedimentos de mecânica molecular nessas posições para os complexos com os DNAs mutados. Os gráficos comparativos dos valores de energias entre o DNA nativo e os DNAs mutados estão no anexo 1.

Foi possível observar que os valores obtidos quando utilizado os modos MZJ e MJZ apresentaram exatamente os mesmos valores. Quando verificado nos dados brutos, as diferenças entre esses modos alterava apenas entre os valores do z-escore e as posições de início e fim da janela, mas nunca o valor da energia média, sendo assim estes modos serão analisados em conjunto sendo chamado de modo M. Os modos ZMJ e ZJM apresentaram diferenças entre as energias apenas com a janela mínima de 20 sendo assim serão também analisados em conjunto sendo chamado de modo Z. Também pode-se observar que, no modo, M quando utilizadas, janelas de 20 a 50 , os perfis dos gráficos permaneceram muito próximos, e janelas de 60 a 90 também apresentam perfis muito similares, sendo possível agrupá-las. Para simplificar a análise foram criados dois grupos do modo M: grupo M1 representado pela janela de 20 (janelas mínimas de 20 a 50) e M2 representada pela janela de 90 (janelas de 60 a 90 energias). Esse mesmo agrupamento também pode ser feito no modo Z, mas neste caso serão criados três grupos: Z1 representando as janelas 20 e 30 , Z2 representado as janelas de 40 e 50 e Z3 representando as janelas de 60 a 90 . Na Figura 40 estão os gráficos selecionados para representar cada grupo.

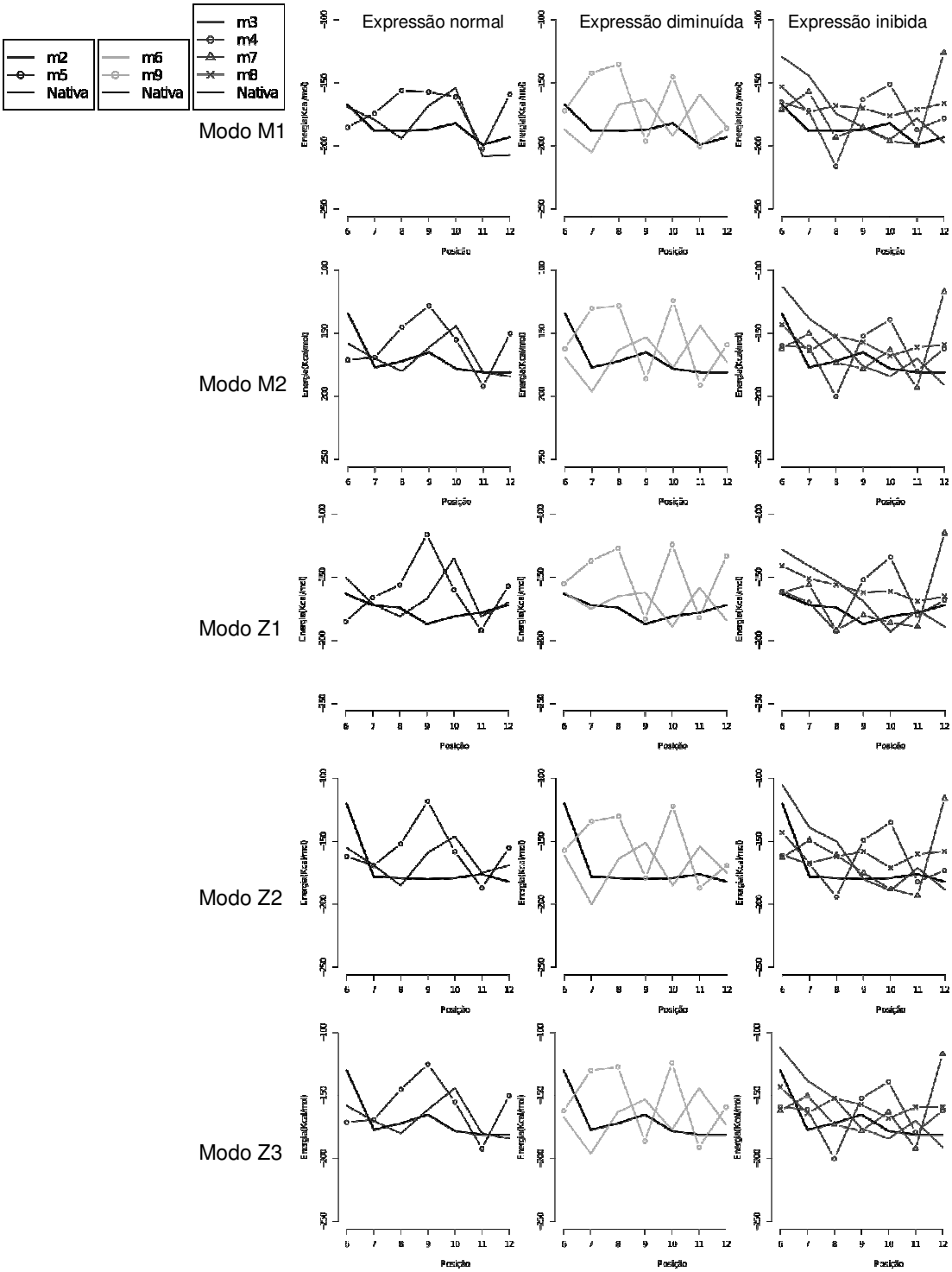


Figura 40: Gráficos representantes de cada grupo dos modos de análise utilizando janelas temporais. No canto superior esquerdo da figura está a legenda para todos os gráficos. A primeira coluna de gráficos representa as energias das mutações que não alteram a expressão do gene repórter. A coluna do meio representa as energias das mutações que diminuem a expressão do gene repórter. A terceira coluna representa as

energias das mutações que inibem a expressão do gene repórter.

É possível observar que no modo de análise M1 em vários pontos os valores de energia dos complexos com DNA apresentando mutações que diminuem ou inibem a expressão do gene repórter são mais baixas que a energia do complexo com DNA nativo, o que também ocorre nos modos M2 e Z3.

No modo Z1 os valores de energia dos complexos apresentando mutações que diminuem a expressão do gene repórter, são maiores que os valores de energia do complexo nativo. Apenas a mutação M6 apresentou uma energia menor na posição 10. Os complexos com mutações que inibem a expressão do gene repórter apresentaram valores de energia pontualmente menores que o complexo nativo, mas, de um modo geral, a energia do complexo nativo apresentou-se mais baixa.

3.2.1 Análise das ligações de hidrogênio

Na tentativa de um melhor entendimento das interações moleculares responsáveis pelas diferenças de energia de interação e, conseqüentemente, pela especificidade, foi examinado o papel de ligações de hidrogênio ao longo da interface proteína-DNA. Analisamos as 7100 conformações retiradas ao longo da dinâmica molecular dos 71 complexos gerados para quantificar (Tabela 9) e classificar as ligações de hidrogênio formadas.

Tabela 9: Número de ligações de hidrogênio formadas nas 7100 conformações analisadas. A análise foi feita para cada um dos 71 complexos gerados. Em branco estão as regiões flanqueadoras das hexades de interação. Em cinza claro a hexade de interação do RXR. Em branco entre as hexades do RXR e do RAR a região espaçadora. Em cinza escuro a hexade de interação do RAR. Na parte superior da tabela as ligações de hidrogênio feitas com as bases e com o fosfoaçúcar de fita direta do DNA. Na parte central as ligações de hidrogênio feitas com os átomos das bases e do FA de fita reversa do DNA. Na parte inferior da tabela temos os números totais de ligações de hidrogênio realizadas por cada posição relativa do DNA e também separados pelas ligações realizadas apenas com a base e apenas com o fosfoaçúcar. Foram mostradas apenas as posições que realizavam ligações de hidrogênio. Nas linhas temos as bases do DNA onde ocorre a ligação de hidrogênio. Nas colunas temos as posições relativas do DNA.

Fita Direta

Base	-3	-2	-1	P1	P2	P3	P4	P5	P6	S1	S2	S3	S4	S5	P1'	P2'	P3'	P4'	P5'	P6'	1	2	3	4
A	0	0	0	188	553	83	0	0	0	0	0	0	0	0	377	434	124	0	0	0	0	0	0	0
T	0	0	0	106	222	73	0	0	0	0	0	0	0	0	149	82	175	36	0	0	0	0	0	0
G	0	1	2	745	1030	30	8	0	0	0	0	0	0	0	433	461	272	110	0	0	0	0	0	0
C	0	0	0	113	67	24	0	0	0	0	0	0	0	2	7	183	51	1	0	0	0	0	0	0
Total	0	1	2	1152	1872	210	8	0	0	0	0	0	0	2	966	1160	622	147	0	0	0	0	0	0

Fosfoaçucar	-3	-2	-1	P1	P2	P3	P4	P5	P6	S1	S2	S3	S4	S5	P1'	P2'	P3'	P4'	P5'	P6'	1	2	3	4
A	35	269	445	1063	2300	2742	27	0	0	0	5	3	47	226	2153	4376	970	0	0	0	0	0	16	49
T	121	298	1340	1193	3605	3107	78	0	0	0	0	91	23	434	1426	2929	131	0	0	0	0	0	55	80
G	181	799	1075	1644	5513	3280	10	0	0	0	1	96	46	216	1158	1918	428	0	0	0	0	0	8	0
C	21	141	184	772	2424	3247	108	0	0	0	5	78	7	150	1022	2599	465	0	0	0	0	1	37	22
Total	358	1507	3044	4672	13842	12376	223	0	0	0	11	268	123	1026	5759	11822	1994	0	0	0	0	1	116	151

Total direta	358	1508	3046	5824	15714	12586	231	0	0	0	11	268	123	1028	6725	12982	2616	147	0	0	0	1	116	151

Fita Reversa

Base	-3	-2	-1	P1	P2	P3	P4	P5	P6	S1	S2	S3	S4	S5	P1'	P2'	P3'	P4'	P5'	P6'	1	2	3	4
A	0	0	1	0	279	509	43	137	0	0	0	0	9	0	0	549	44	281	234	2	0	0	0	0
T	0	0	35	0	163	0	54	0	0	0	0	0	18	0	0	256	8	176	0	0	0	0	0	0
G	0	0	0	2	247	145	863	140	0	0	0	0	0	0	0	310	405	1078	786	0	0	0	0	0
C	0	0	49	2	588	1277	2	0	0	0	0	0	2	0	0	524	268	12	0	0	0	0	0	0
Total base	0	0	85	4	1277	1931	962	277	0	0	0	0	29	0	0	1639	725	1547	1020	2	0	0	0	0

Fosfoaçucar	-3	-2	-1	P1	P2	P3	P4	P5	P6	S1	S2	S3	S4	S5	P1'	P2'	P3'	P4'	P5'	P6'	1	2	3	4
A	109	577	34	0	0	593	1274	3442	1175	19	0	0	29	0	0	0	19	2765	3846	1042	29	0	0	0
T	274	401	91	7	0	60	811	1836	954	120	1	214	35	0	0	0	0	1214	1382	589	1	0	0	0
G	22	179	2	0	0	612	2238	4247	1689	139	0	86	1	0	0	0	12	3499	3491	644	138	0	0	0
C	588	1037	165	1	0	175	444	1941	794	71	2	56	23	0	0	0	0	586	463	754	1	0	0	0
Total	993	2194	292	8	0	1440	4767	11466	4612	349	3	356	88	0	0	0	31	8064	9182	3029	169	0	0	0

Total reversa	993	2194	377	12	1277	3371	5729	11743	4612	349	3	356	117	0	0	1639	756	9611	10202	3031	169	0	0	0
Total fitas	1351	3702	3423	5836	16991	15957	5960	11743	4612	349	14	624	240	1028	6725	14621	3372	9758	10202	3031	169	1	116	151
Total bases	0	1	87	1156	3149	2141	970	277	0	0	0	0	29	2	966	2799	1347	1694	1020	2	0	0	0	0
Total FA	1351	3701	3336	4680	13842	13816	4990	11466	4612	349	14	624	211	1026	5759	11822	2025	8064	9182	3029	169	1	116	151

Observou-se que a região de interação é mais extensa do que apenas as regiões das hexades sugeridas por Zelent *et al.* (1991) e Vivanco Ruiz *et al.* (1991). As interações ocorrem a partir da terceira base antes da primeira hexade até a quarta bases após a segunda hexade. No entanto o número médio de ligações nas regiões das hexades é de 8254 enquanto que nas outras regiões é de apenas 930. Mas, é válido notar que as posições -3, -2, -1, S3 e S5 apresentam um número não negligenciável de ligações de hidrogênio.

As ligações de hidrogênio, nas regiões das hexades do RXR e do RAR, ocorrem preferencialmente nas posições P1, P2, P3 e P4 na fita direta e P3, P4, P5 e P6 na fita reversa. Analisando a estrutura tridimensional foi observado que na região da hexade de interação do DNA tanto RXR como RAR apresentam uma hélice-alfa que fica inserida na fenda maior do DNA, realizando a maior parte dos contatos (Figura 41).

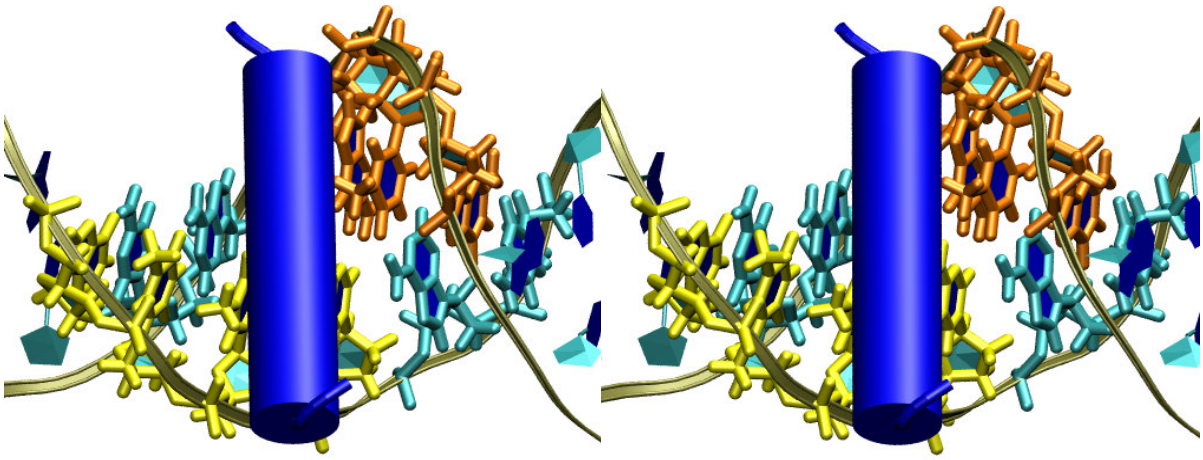


Figura 41: Vista estereográfica da região da hexade e da alfa-hélice do RXR que fica inserida na fenda maior do DNA. Em azul escuro a hélice do RXR, em amarelo os resíduos P1, P2, P3 e P4 da fita direta, em laranja os resíduos P3, P4, P4 e P6 da fita complementar e em azul claro os resíduos P5, P6 e P1, P2 da fita direta e complementar, respectivamente.

Os resíduos carboxiterminais da hélice-alfa ficam mais próxima da fita direta e os resíduos N-terminais ficam mais próximos da fita complementar, explicando por que as

interações ocorrem preferencialmente nas posições de P1 a P4 da fita direta e de P3 a P6 da fita complementar e o fato das posições P1 e P2 da fita direta e P5 e P6 da fita complementar realizarem poucas ou nenhuma ligação com a proteína. Esse mesmo padrão ocorre no modo de ligação do receptor RAR (Figura 42).

Para verificar quais os resíduos são realmente importantes na especificidade da interação entre os fatores de transcrição e o DNA, classificamos as ligações de hidrogênio entre proteínas e o DNA em relação aos átomos do esqueleto de fosfoaçúcar e átomos da base nitrogenada. As duas últimas linhas da Tabela 9 mostram esse dados. Podemos observar que a maioria das ligações são realizadas com os átomos do açúcar ou do fosfato. As posições P1, P2, P3, P5, P1’, P2', P3', P4’ e P5' fazem ligações com as bases nitrogenadas. As posições -1 e S4 que fazem ligações com a base e estão fora das hexades mas em número muito menor que as outras posições (Figura 42).

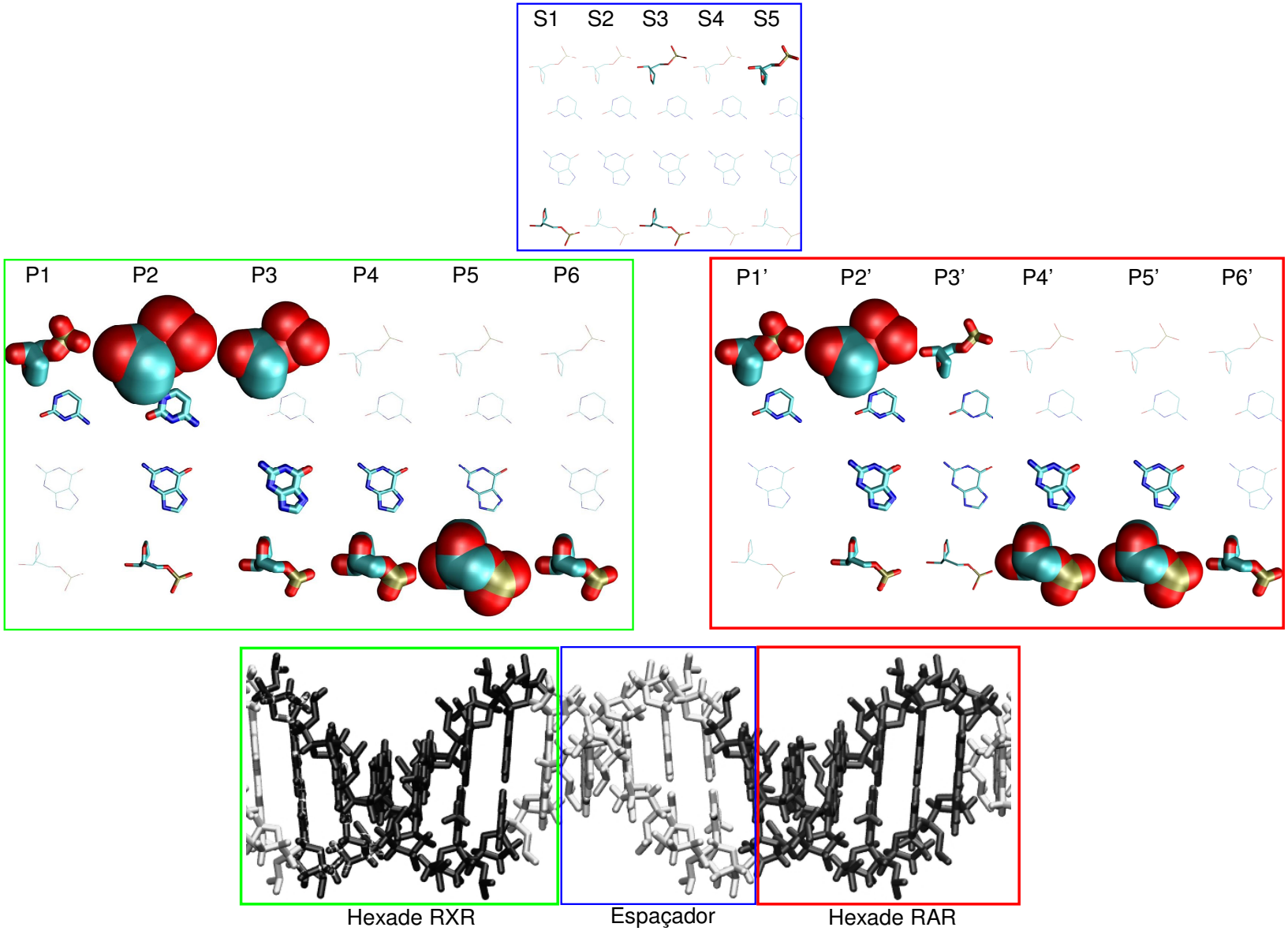


Figura 42: Figura esquemática das ligações de hidrogênio onde o tamanho do fosfoaçúcar e da base representa a quantidade de ligações de hidrogênio feitas.

Analisamos os aminoácidos para tentar estabelecer uma relação entre os resíduos da interação e a especificidade. Classificamos os aminoácidos em dois grupos: os que fazem ligação com os átomos do fosfoaçúcar e os que fazem ligação com os átomos das bases nitrogenadas (Tabela 10).

Tabela 10: Número de ligações de hidrogênio feitas pelos aminoácidos com o DNA, classificado em interações feitas com o fosfoaçúcar e interações feitas com a base nitrogenada, nas 7100 conformações analisadas dos 71 complexos gerados. Na primeira tabela os aminoácidos do RXR que fazem interação com o DNA. Na segunda tabela os aminoácidos do RAR que fazem interação com o DNA. Na terceira e quarta tabela os resíduos do RXR e RAR respectivamente que fazem interação com as bases discriminando a quantidade de ligações feitas com cada uma das bases. As linhas Número se referem ao número do resíduo e Nome ao nome do resíduo.

Número	69	70	73	82	86	87	88	94	97	101	102	105	106	113	123	125	126	127	129	132	147	150	151	152
Nome	GLY	SER	LYS	ARG	LYS	HID	TYR	GLU	LYS	LYS	ARG	ARG	LYS	ARG	ARG	ARG	ASN	ARG	GLN	ARG	GLN	ARG	GLN	ARG
fosfoaçúcar	121	268	1476	1451	3041	1154	9698		205	2792	6814	8120	2206	310	45	3737	1490	254	1977	4970	4270	4736	255	3354
Base						326		3143	2727	197	1246											128		

Número	160	164	166	172	175	179	180	183	184	203	204	205	207	210	223	225	226
Nome	LYS	TYR	TYR	GLU	LYS	ARG	ARG	GLN	LYS	ARG	ASN	ARG	GLN	ARG	SER	ARG	ASN
fosfoaçúcar	45	741	10429			1397	5436	1312	457	3460	1929	266	3659	5446	1585	5294	38
Base				1757	2701	1105	2254										

Número	87	94	97	101	102	150
Nome	HID	GLU	LYS	LYS	ARG	ARG
A	14	1202	257	106	180	34
T	0	0	543	14	61	35
G	274	0	1849	76	1005	6
C	38	1941	78	1	0	53
Total	326	3143	2727	197	1246	128

Número	172	175	179	180
Nome	GLU	LYS	ARG	ARG
A	970	431	138	502
T	0	484	244	151
G	0	1554	700	1601
C	787	232	23	0
Total	1757	2701	1105	2254

É possível observar que apenas os aminoácidos 87, 94, 97, 101, 102 e 150 do RXR e 172, 175, 179 e 180 do RAR fazem ligações com as bases nitrogenadas. Sendo que apenas os aminoácidos 94 do RXR e 172 e 175 do RAR que fazem ligações exclusivamente com a base nitrogenada. Destacamos os aminoácidos 94, 97 e 102 do RXR e os aminoácidos 172, 175, 179 e 180 do RAR como sendo os mais importantes para a especificidade da interação entre os fatores de transcrição e o DNA. Na análise estrutural observamos que os quatro aminoácidos se encontram na região da hélice que fica inserida na fenda maior do DNA (Figura 43).

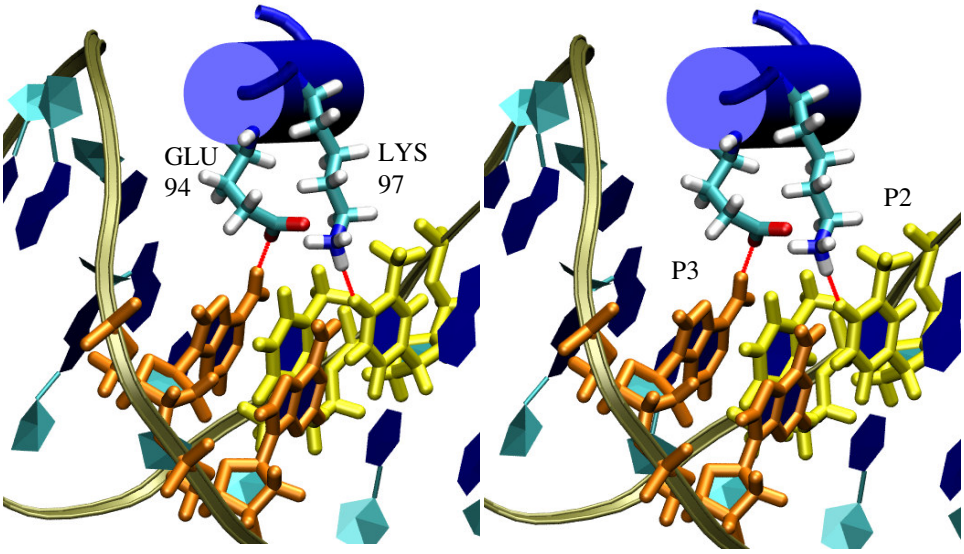


Figura 43: Vista estereográfica da região de contato entre o DNA e o RXR . Em azul escuro a hélice do RXR, em amarelo os resíduos P2 e P3 da fita direta, em laranja os resíduos P2 e P3 da fita complementar e em vermelho umas das possíveis ligações de hidrogênio feitas entre os resíduos e o DNA.

Percebe-se, ainda na Tabela 10, que os resíduos 87, 97, 102, 175, 179 e 180 têm uma preferência por guaninas e os resíduos 94 e 172 apenas se ligam a adeninas e timinas. Para verificar se realmente existe uma preferência dos aminoácidos por um determinado tipo de nucleotídeo, foi analisada a quantidade de bases presentes em cada posição relativa

a hexade (Tabela 11). Pode-se observar que a quantidade de bases por posição não é igualmente distribuídas, sendo assim necessário verificar com quais posições do DNA o aminoácido interage para saber se realmente existe uma preferência.

Para verificar a preferência de cada aminoácido foi analisado com quais bases do DNA um determinado aminoácido interage (Figura 44). Podemos observar que o aminoácido 87 interage apenas com a P1D. O aminoácido 94 interage com as posições P2D, P2R e P3R. O aminoácido 97 interage com as posições P1D, P2D, P2R, P3R. O aminoácido 101 interage com as posições P2D e P3D. O aminoácido 102 interage com as posições P4R e P5R. O aminoácido 150 interage apenas com a posição -1R. O aminoácido 172 do RAR interage com as posições P2D', P2R' e P3R' sendo um aminoácido corresponde ao aminoácido 94 do RXR. O aminoácido 175 seria o correspondente ao 97 do RXR. O aminoácido 179 interage com as posições P3D', P4D' e P4R'. E o aminoácido 180 interage com as posições P4R' e P5R'.

Tabela 11: Quantidade de bases em cada posição relativa do DNA nas 71 conformações analisadas.

Direta

Posição relativa	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	P1	P2	P3	P4	P5	P6
A						1	1	1	1	1	8	8	10	11	11	11	12	13	14	14	12	19
T					1	1	1	1	1	9	10	11	11	11	18	25	19	19	19	18	20	14
G		1	2	3	3	3	12	21	30	31	31	38	45	44	36	29	27	27	19	13	13	12
C	1	1	1	1	1	9	9	9	9	9	10	11	3	4	6	6	13	12	19	26	26	26

Posição relativa	S1	S2	S3	S4	S5	P1'	P2'	P3'	P4'	P5'	P6'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	26	34	27	27	26	26	26	20	13	14	13	13	20	13	21	21	27	27	27	19	19	11	10	2	1	1
T	7	6	6	12	19	18	18	19	27	26	21	14	14	21	19	20	20	19	18	18	10	10	2	1	1	
G	12	11	17	17	18	12	12	11	11	4	11	10	9	9	9	8	8	1	1	1						
C	26	20	21	15	8	15	15	21	20	27	26	34	27	26	19	18	11	10	2	1	1					

Reversa

Posição relativa	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	P1	P2	P3	P4	P5	P6
T						1	1	1	1	1	8	8	10	11	11	11	12	13	14	14	12	19
A					1	1	1	1	1	9	10	11	11	11	18	25	19	19	19	18	20	14
C		1	2	3	3	3	12	21	30	31	31	38	45	44	36	29	27	27	19	13	13	12
G	1	1	1	1	1	9	9	9	9	9	10	11	3	4	6	6	13	12	19	26	26	26

Posição relativa	S1	S2	S3	S4	S5	P1'	P2'	P3'	P4'	P5'	P6'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T	26	34	27	27	26	26	26	20	13	14	13	13	20	13	21	21	27	27	27	19	19	11	10	2	1	1
A	7	6	6	12	19	18	18	19	27	26	21	14	14	21	19	20	20	19	18	18	10	10	2	1	1	
C	12	11	17	17	18	12	12	11	11	4	11	10	9	9	9	8	8	1	1	1						
G	26	20	21	15	8	15	15	21	20	27	26	34	27	26	19	18	11	10	2	1	1					

Residuo	87	Quant.	94	Quant.	97	Quant.	101	Quant.	102	Quant.	150	Quant.
	P1D	326	P2D	432	P1D	755	P2D	67	P4R	956	-1R	85
			P2R	866	P2D	1359	P3D	124	P5R	277		
			P3R	1786	P2R	411						
					P3R	145						

Residuo	172	Quant.	175	Quant.	179	Quant.	180	Quant.
	P2D'	384	P1D'	958	P3D'	587	P4R'	1187
	P2R'	1067	P2D'	750	P4D'	114	P5R'	1020
	P3R'	293	P2R'	572	P4R'	333		
			P3R'	379				

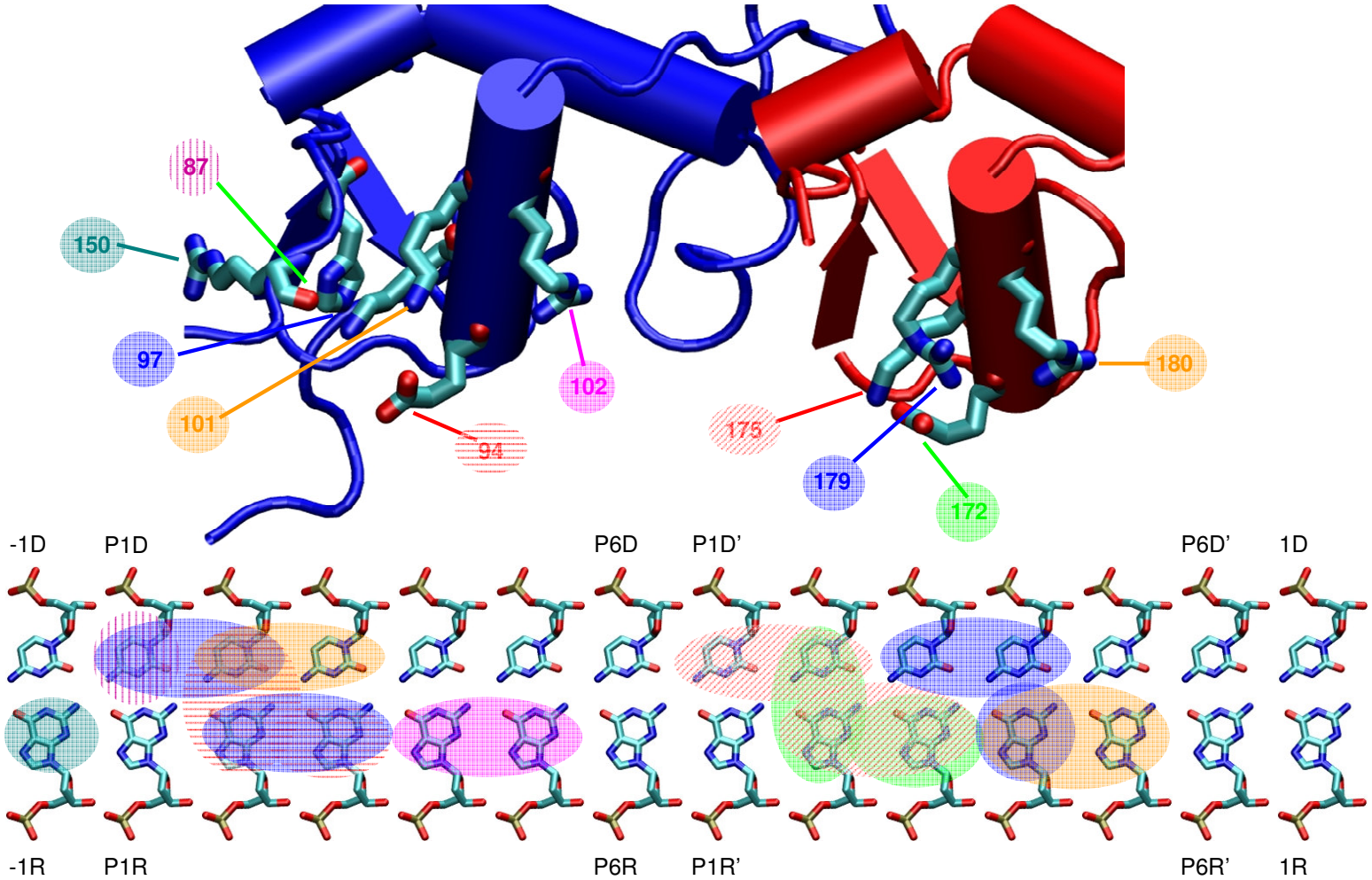


Figura 44: Relação das interações específicas que cada aminoácido realiza com o DNA. Cada cor sobre o DNA corresponde às bases que um dado aminoácido interage.

Com as informações referentes a quantidade de bases em cada posição relativa do DNA e em quais posições relativa um aminoácido pode ligar, foi criada a Tabela 12, onde calculamos os resíduos de um teste qui-quadrado de Pearson para verificar se realmente existe a preferência de uma aminoácido por um resíduo.

Tabela 12: Quantidade de ligações realizada com as bases, agrupadas pelos aminoácidos. Também são apresentados os resíduos de Pearson mostrando a preferência ou a preterência a uma base.

RXR

Resíduo		Base	Pos. relativa	Ligações	resíduo de Pearson
87	HIS	A	P1D	14	-5,537
87	HIS	G	P1D	274	13,474
87	HIS	C	P1D	38	-2,807
94	GLU	A	P2D	414	37,656
94	GLU	C	P2D	18	-6,438
94	GLU	A	P2R	279	3,104
94	GLU	C	P2R	587	3,104
94	GLU	A	P3R	509	1,421
94	GLU	C	P3R	1277	36,55
97	LYS	A	P1D	143	1,363
97	LYS	T	P1D	106	-6,757
97	LYS	G	P1D	467	10,616
97	LYS	C	P1D	39	-8,441
97	LYS	A	P2D	114	-8,547
97	LYS	T	P2D	222	-7,429
97	LYS	G	P2D	985	20,595
97	LYS	C	P2D	38	-12,648
97	LYS	T	P2R	163	10,115
97	LYS	G	P2R	247	21,301
97	LYS	G	P3R	145	17,048
101	LYS	A	P2D	23	3,064
101	LYS	G	P2D	44	3,669
101	LYS	A	P3D	83	11,841
101	LYS	T	P3D	14	-3,33
101	LYS	G	P3D	26	-1,247
101	LYS	C	P3D	1	-5,587
102	ARG	A	P4R	43	-12,806
102	ARG	T	P4R	54	-9,797
102	ARG	G	P4R	859	27,199
102	ARG	C	P4R	0	-13,23
102	ARG	A	P5R	137	6,676
102	ARG	G	P5R	140	3,829
150	ARG	A	-1R	1	-5,288
150	ARG	T	-1R	35	6,016
150	ARG	C	-1R	49	2,424

RAR

Resíduo		Base	Pos. relativa	Ligações	resíduo de Pearson
172	GLU	A	P2D'	384	20,524
172	GLU	A	P2R'	543	16,568
172	GLU	C	P2R'	524	25,591
172	GLU	A	P3R'	42	-4,112
172	GLU	C	P3R'	251	30,516
175	LYS	A	P1D'	376	1,345
175	LYS	T	P1D'	146	-6,216
175	LYS	G	P1D'	433	21,304
175	LYS	C	P1D'	3	-14,016
175	LYS	A	P2D'	36	-14,4
175	LYS	T	P2D'	82	-7,842
175	LYS	G	P2D'	451	28,799
175	LYS	C	P2D'	181	1,791
175	LYS	A	P2R'	6	-11,544
175	LYS	T	P2R'	256	3,215
175	LYS	G	P2R'	310	17,207
175	LYS	A	P3R'	2	-9,872
175	LYS	G	P3R'	360	23,414
179	ARG	A	P3D'	124	-3,216
179	ARG	T	P3D'	171	1,11
179	ARG	G	P3D'	272	18,986
179	ARG	C	P3D'	20	-11,659
179	ARG	T	P4D'	25	-2,787
179	ARG	G	P4D'	88	16,737
179	ARG	T	P4R'	48	-1,661
179	ARG	G	P4R'	285	19,741
180	ARG	A	P4R'	266	-8,726
180	ARG	T	P4R'	128	-6,06
180	ARG	G	P4R'	793	25,082
180	ARG	C	P4R'	0	-13,561
180	ARG	A	P5R'	234	-7,442
180	ARG	G	P5R'	786	19,792

É possível observar que o aminoácido HIS 87 tem preferência por guanina da posição P1D. O aminoácido GLU 94 tem preferência por citosina na posição P3R, por adeninas da posição P2D. Na posição P2R uma leve preferência por adeninas de citosinas. E na posição P3R uma grande preferência pode citosinas, mas ligando a adeninas quando

presente. O aminoácido LYS 97 apresenta uma preferência por guaninas na posição P1D, P2D, P2R e P3R, mas na posição P2R também se liga a timinas. O aminoácido LYS 101 liga a adeninas e guaninas na posição P2D, apresentando uma leve preferência, na posição P3D ele apresenta preferência por adeninas. O aminoácido ARG 102 apresenta uma grande preferência por guaninas na posição P4R e na posição P5R tem preferência por adeninas e guaninas. E o aminoácido ARG 150 interage especificamente com a posição -1R e preferencialmente com timinas.

As análises dos aminoácidos do RAR mostram que o aminoácido GLU 172 tem preferência por adeninas na posição P2D' e na posição P2R' ele se liga tanto com adenina com citosina e na posição P3R' ele prefere citosina. O aminoácido LYS 175 prefere guaninas nas posições P1D', P2D', P2R' e P3R'. O aminoácido ARG 179 tem preferência por guaninas nas posições P3D', P4D' e P4R'. O aminoácido ARG 180 tem preferência por guanina nas posições P4R' e P5R'.

Foram quantificadas as ligações de hidrogênio específicas, ligações realizadas com as bases nas 15 posições analisadas do DNA nativo (Figura 45).

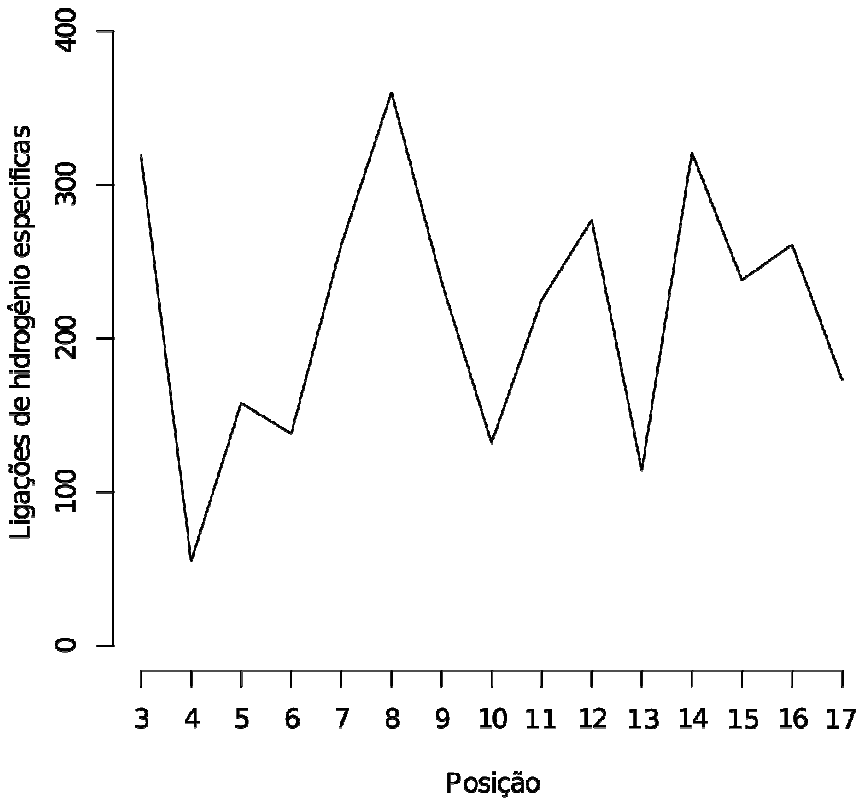


Figura 45: Quantidade de ligações de hidrogênio específicas realizada com a base no DNA nativo.

Foi observado que a posição 8 é a posição que apresenta maior número de ligações específicas. Essa mesma quantificação também foi realizada para os DNAs mutantes (Figura 46).

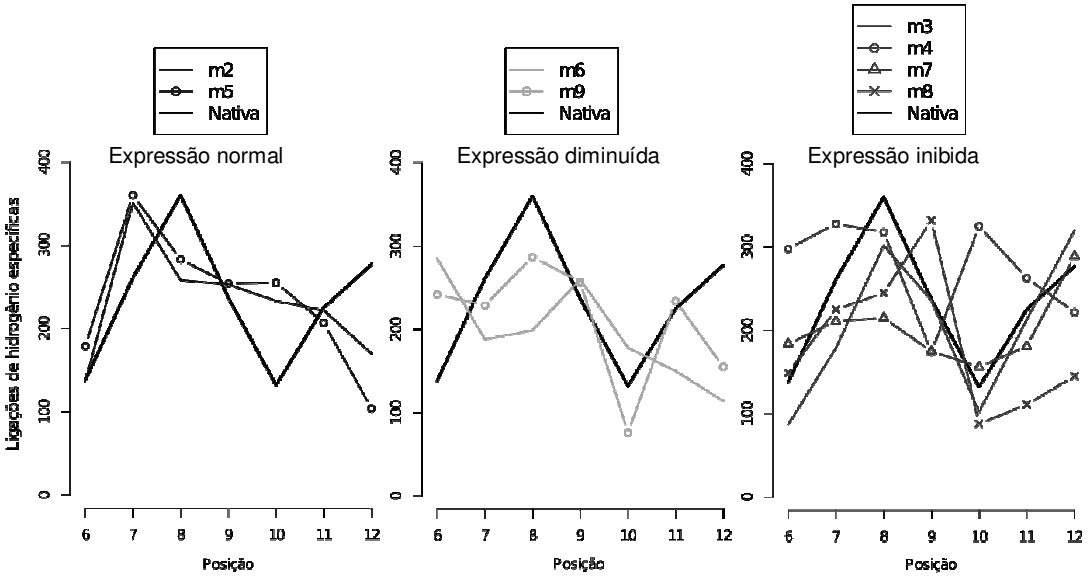


Figura 46: Quantidade de ligações de hidrogênio realizada com a base em todos os DNAs analisados.

Aqui pode ser notado que o complexo envolvendo DNA com as mutações que não alteram a expressão do gene repórter, apresentaram uma quantidade de ligações de hidrogênio que são muito próximas da quantidade de ligações do DNA nativo, mas na posição 6 do DNA. Os complexos com DNA cujas as mutações diminuem ou inibem a expressão do gene repórter apresentaram quantidade menores de ligações de hidrogênio específicas. Na Tabela 13, estão apresentadas as quantidade de ligações feitas apenas com os átomos das bases, quantidade de ligações feitas apenas com os átomos do FA e a quantidade total de ligações de hidrogênio agrupados pelos complexos DNA-proteína.

Tabela 13: Quantidade ligações de hidrogênio realizadas nos 71 complexos estudados. Na primeira tabela a quantidade de ligações feitas apenas com os átomos das bases. Na segunda tabela a quantidade de ligações feitas apenas com os átomos do FA e na última tabela a quantidade total. Nas linhas estão as quantidade de ligações e nas colunas as posições no DNA na qual o complexo esta posicionado. As células em negrito correspondem à melhor posição do complexo nativo e dos complexos com DNA mutados que não alteram a expressão do gene repórter.

Base	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
nat	319	55	158	138	261	**360**	237	132	225	277	114	321	238	261	173
m2				136	**351**	258	253	233	222	170					
m3				87	179	302	236	102	213	320					
m4				298	328	318	174	325	263	222					
m5				179	**360**	283	254	255	207	104					
m6				285	188	199	260	178	150	114					
m7				184	211	215	175	156	181	289					
m8				149	225	245	332	88	111	145					
m9				242	229	287	257	76	234	155					

FA	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
nat	1067	1315	1387	1280	1381	**1122**	1189	1534	1506	1339	1213	1091	1269	1333	1161
m2				1372	**969**	1325	1240	1136	1380	1466					
m3				1104	1012	1106	1304	1547	1169	1207					
m4				1251	1106	1547	1261	776	1272	1245					
m5				1324	**942**	1032	890	1169	1450	1264					
m6				997	1420	1146	992	1364	1218	1314					
m7				1270	1115	1308	1341	1284	1416	873					
m8				1371	1405	1108	1072	1373	1303	1414					
m9				1254	1128	999	1433	1224	1322	1349					

Total	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
nat	1386	1370	1545	1418	1642	1482	1426	1666	1731	1616	1327	1412	1507	1594	1334
m2				1508	1320	1583	1493	1369	1602	1636					
m3				1191	1191	1408	1540	1649	1382	1527					
m4				1549	1434	1865	1435	1101	1535	1467					
m5				1503	1302	1315	1144	1424	1657	1368					
m6				1282	1608	1345	1252	1542	1368	1428					
m7				1454	1326	1523	1516	1440	1597	1162					
m8				1520	1630	1353	1404	1461	1414	1559					
m9				1496	1357	1286	1690	1300	1556	1504					

A quantidade de ligações feitas com o FA pela posição 8 do DNA nativo é a terceira menor sendo apenas maior que a posição 3 e 14. Também, observou-se que complexo envolvendo DNA com as mutações m2 e m5 na posição 7 da tabela 13-FA foram aqueles que apresentaram a menor quantidade de ligação de hidrogênio com FA. Esses dados

sugerem uma possível preferência dos aminoácidos nas posições 7 dos DNAs mutados e posição 8 do complexo nativo de se ligarem aos átomos das bases nitrogenadas em detrimento dos átomos do FA.

Tabela 14: Informações referentes ao DNA nativo e as mutações m2 e m5. Na tabela 1 a quantidade de ligações de hidrogênio discriminada por tipo de base nitrogenada. Na tabela 2 a quantidade de ligações de hidrogênio discriminada pelas posições relativas do DNA. E na tabela 3 a quantidade de ligações de hidrogênio discriminado pelo número e nome dos aminoácidos que realizam a ligação. Não foram apresentadas as bases e os aminoácidos que fizeram menos de 3 ligações de hidrogênio.

1	nat	m2	m5
A	72	77	105
G	163	178	125
C	121	96	130

2	nat	m2	m5
P1D	3	39	31
P2D	56	47	37
P2R	50	13	40
P3R	50	73	79
P1D'	19	18	9
P2D'	44	19	47
P2R'	70		
P3D'	4	86	55
P3R'	3	10	11
P4R'		39	49
P5R'	60		

3		nat	m2	m5
RXR				
87	HID		18	11
94	GLU	100	86	119
97	LYS	59	58	47
101	LYS		16	12
	Soma	159	178	189
RAR				
172	GLU	73	25	50
175	LYS	63	22	12
179	ARG	4	86	60
180	ARG	60	39	49
	Soma	200	172	171

Foi possível observar que não ocorre interação com timinas nos três DNAs analisados e uma ligeira preferência por guaninas e citosinas. Com relação às posições relativas do DNA é possível observar que as posições P1D,P2D, P2Re P3R fazem o maior

número de ligações de hidrogênio específicas na hexade do RXR. Na hexade do RAR as ligações são mais distribuídas, feita por praticamente todas as posições menos a P6'. Em relação aos aminoácidos a maior parte das ligações do RXR foram realizadas pelo ácido glutâmico 94 e pela lisina 97. Já no RAR os aminoácidos: ácido glutâmico 172, lisina 175, arginina 179 e arginina 180 realizaram todas as ligações de hidrogênio específicas.

4. Discussão

4.1 Análise da busca por sítios de ligação de fatores de transcrição utilizando matrizes

As análises utilizando matrizes mostram ser uma ferramenta muito poderosa desde que os fatores de interesse já tenham sido analisados e criada uma matriz de busca pelo fator. O estudo dos fatores RXR e RAR mostrou ser inviável por esse método, pois ainda não estão disponíveis matrizes de busca por esses fatores. Isso demonstra a importância da criação de métodos de busca por sítio de ligação de fatores de transcrição independente de informações a priori.

4.2 Análise da busca por sítios de ligação de fatores de transcrição utilizando a estrutura tridimensional e cálculos termodinâmicos

Os resultados obtidos utilizando as energias médias, medianas e mínimas mostraram uma região, que corresponde da posição 6 a posição 12 do DNA, como uma região de baixa energia sugerindo que essa região é importante para a ligação entre o fator de transcrição e o DNA. Mas não apresentou sensibilidade suficiente para prever exatamente quais bases do

DNA realmente interagem com o fator de transcrição sendo responsável pela especificidade.

O método utilizando as janelas temporais de energias avaliadas por quatro modos de ordenação dos valores obtidos mostrou não haver a necessidade de realizar os modos MJZ e MZJ, já que ambos apresentaram os mesmos valores. Apenas a ordenação pela menor média de energia já é o suficiente para este modo de análise. Os modos ZJM e ZMJ também apresentaram resultados muito semelhantes não interferindo significativamente nas análises dos dados, também podendo utilizar apenas a ordenação pelo z-escore. Cada modo apresentou resultados diferentes em termos valores absolutos, enquanto o modo M apresentou as menores energias, mas não necessariamente as mais estáveis, o modo Z a representou intervalos de tempo de uma maior estabilidade termodinâmica, mas não necessariamente a menor energia do sistema. Estes resultados demonstram que a análise é suscetível ao critério de escolha de um sistema mais estável ou um sistema com menor energia. Neste trabalho, o modo Z apresentou um melhor resultado, onde os valores de energia dos complexos envolvendo DNA com as mutações que diminuem ou inibem a expressão do gene repórter apresentaram energias próximas ou maiores que a energia do complexo nativo, excetuando alguns valores de energia menores, mostrando que esse modo apresenta uma boa sensibilidade, mas ainda com alguns falsos positivos.

Com relação aos tamanhos mínimos das janelas temporais, foi possível agrupá-las em três tamanhos: de 20 a 30, de 40 a 50 e de 60 a 90. Esse agrupamento mostrou que janelas próximas apresentam resultados semelhantes e que, a partir de uma janela temporal de 50, os resultados ficam muito próximos. Neste trabalho as janelas de 40 a 50 mostraram um melhor resultado. Provavelmente, isso pode ter ocorrido porque janelas pequenas sofrem com alterações pontuais ocorridas durante a dinâmica, não mostrando serem

“fotografias” de conformações estáveis do sistema. Assim, para um número menor de valores, variações de amplitude tendem a deslocar a média para valores de energias menores. Em janelas maiores que 60 começam a sofrer com ruídos, pela inclusão de valores de energia muito altos ou muito baixos, introduzindo erros de estimativa. Possivelmente um aumento no tempo de dinâmica poderia melhorar os resultados obtidos, pois alguns dos complexos poderiam ainda não ter chegado à estabilização energética. Os resultados com avaliação energética mostraram ser bastante promissores, mas ainda é necessário aplicar o método com outros fatores de transcrição para validá-lo.

4.3 Análises das ligações de hidrogênio.

O alinhamento das seqüências nucleotídicas de várias regiões que sabidamente ligavam fatores de transcrição definiram uma região de seis pares de bases como sendo o centro de reconhecimento desses fatores, a seqüência AGAACA é preferencialmente reconhecida por fatores de transcrição das famílias dos receptores glucocorticóide, mineralocorticóide, receptores controlados por progesterona e receptores controlados por andrógenos (Glass 1994). A seqüência AG(G/T)TCA é preferencialmente reconhecida por receptores controlados por estrógeno, hormônios da tireóide, vitamina D e ácidos retinóico (Glass 1994). Essa definição de uma seqüência consenso foi baseada apenas nos alinhamentos de seqüências de nucleotídeos, sendo que neste trabalho foi possível verificar que a região de ligação entre os fatores de transcrição e o DNA não se restringe apenas a essas 6 bases, abrangendo da posição -3 a posição S3. Realmente, a região que faz interação específica com o DNA esta inserida no interior da hexade sugerida e, na verdade, pode ser considerada até menor, envolvendo as posições de P1 a P4 no RXR e de P1’ a P5’ no RAR.

As ligações de hidrogênios específicas ocorrem nas duas fitas do DNA, devido ao posicionamento da uma alfa-hélice na fenda maior do DNA. Essas alfa-hélices são formadas pelos resíduos 93 a 107 no RXR e 171 a 185 no RAR, desses aminoácidos apenas os resíduo 94, 97, 101 e 102 do RXR e 171, 175, 179 e 180 realizam ligações específicas com o DNA, e isso se deve apenas porque esses aminoácidos estão voltados para a região das bases nucléicas, enquanto outros aminoácidos estão voltados para o FA ou para o interior da proteína.

O aminoácido GLU 94 interage preferencialmente com a posição P3R, pois essa é a base mais próxima do seu carbono-alfa. Os átomos de oxigênio ligados ao carbono-delta, no carboxilato desse aminoácido, interagem preferêncialmente com os átomos de hidrogênio ligado ao N4 das citosinas, mas também fazem ligações com os átomos de hidrogênio ligado ao N6 das adeninas. O aminoácido GLU 94 também interage com as posições P2D, P2R e P3R sempre com as bases adenina ou citosina. Assim, nucleotídeos com cargas parciais positivas disponíveis (A ou C pelos hidrogênios ligados aos átomos N4 e N6 respectivamente) principalmente na posição P3R, mas podendo ser também na posição P2D e P2R, podem favorecer a especificidade da ligação, sendo doadores em ligações de hidrogênio.

O aminoácido LYS 97 interage preferencialmente com a posição P2D, mas também interage com a P1D, P2R e P3R. Essa interação é feita pelos átomos dos hidrogênios ligados ao NZ da lisina com o átomo O6 de um grupo carbonílico da guanina preferencialmente, mas ocorrendo também interação com o átomo O4 do carbonil de timina e N7 da adenina e guanina. Nucleotídeos com cargas parciais negativas (G e A pelos “lone pairs “ dos átomos O6 e O4 respectivamente) na posição P2D principalmente, mas podendo

ser na posição P1D, P2R e P3R, pode favorecer a especificidade da interação, participando como aceitadores em ligações de hidrogênio.

O aminoácido LYS 101 interage preferencialmente com a posição P3D, mas também interage com a posição P2D. Interagindo preferencialmente com o átomo N7 da adenina, esse aminoácido busca por cargas negativas ou aceitadores de ligação de hidrogênio, interagindo também com o átomo N7 da guanina. Apesar do aminoácido LYS 101 estar localizado próximo a posição P3D e realizar ligações específicas com o DNA, o seu posicionamento em relação ao DNA favorece a interação inespecífica, através da ligação com os átomos dos fosfatos. Sendo assim muito menor é a quantidade de ligações realizada por LYS 101 com os átomos da base. Nucleotídeo com cargas parciais negativas na posição P3D podem favorecer a especificidade, mais aparentemente não sendo essenciais.

Podemos observar que existe uma sobreposição, antagônica, da atuação do aminoácido GLU 94 com os aminoácidos LYS 97 e LYS 101, onde o primeiro tem preferência por cargas parciais positivas e os outros dois por cargas parciais negativas. Considerando que cada um desses aminoácidos esta próximo de uma posição diferente do DNA, a melhor configuração para favorecer a especificidade da ligação seria uma adenina ou guanina na posição P2D, uma guanina na posição P3D e conseqüentemente uma citosina na posição P3R.

O aminoácido ARG 102 interage preferencialmente com a posição P4R, mas também interage com a posição P5R. Interagindo preferencialmente com o átomo O6 da guanina, esse aminoácido busca por aceitadores de ligação de hidrogênio, interagindo também com o átomo N7 de guaninas e adeninas. O carbono-alfa da ARG 102 localiza eqüidistante da posição P4R e P5R, mas ocorre uma preferência pela posição P4R,

provavelmente, devido a melhores ângulos torcionais da cadeia lateral. Assim, a existência de nucleotídeos com carga parcial negativa na posição P4R pode favorecer a especificidade.

Existem também dois aminoácidos que interagem especificamente com o DNA e estão localizados fora da alfa-hélice que se insere na fenda maior do DNA. Esses aminoácidos são HIS 87 e ARG 150.

O aminoácido HIS 87 interage unicamente com a posição P1D preferencialmente com guaninas principalmente com átomo N7, mas também interage com adenina também com o átomo N7. Então a existência de nucleotídeos com carga parcial negativa (G e A) na posição P1D pode favorecer a especificidade.

O aminoácido ARG 150 interage unicamente com a posição -1R com o átomo O2 de grupos carbonil de citosinas e timinas. Sabendo que esse aminoácido fica inserido na fenda menor do DNA, a interação desse aminoácido com o nucleotídeo ocorre no lado oposto ao que fica exposto na fenda maior do DNA. Então, nesse caso, a existência de nucleotídeos aceitadores de ligação de hidrogênio (C e T pelo átomo O2) na posição -1R pode favorecer a especificidade.

No RAR a alfa-hélice, que se insere na fenda maior do DNA, é formada pelos resíduos 171 a 185. Desses aminoácidos apenas os resíduos 172, 175, 179 e 180 interagem especificamente com o DNA. O aminoácido GLU 172 interage preferencialmente com a posição P2R', mas também interage com P2D' e P3R' se ligando preferencialmente com os átomos de hidrogênio ligados ao N4 das citosinas ou com os átomos de hidrogênio ligado ao N6 das adeninas. Nucleotídeos com cargas parciais positivas (A ou C pelos átomos N4 e N6 respectivamente) principalmente na posição P2R', mas podendo ser também na posição

P2D’ e P3R’, podem favorecer a especificidade da ligação, como doadores de ligação de hidrogênio.

O aminoácido LYS 175 interage preferencialmente com a posição P1D’, mas também interagem com a P2D’, P2R’ e P3R’. Essa interação é feita pelos átomos dos hidrogênios ligados ao NZ da lisina com o átomo O6 da guanina preferencialmente, mas ocorrendo também interação com o átomo O4 da timina e N7 da adenina e guanina. Nucleotídeos com cargas parciais negativas (G, A e T pelos átomos O6, N7 e O4 respectivamente) na posição P1D’ principalmente, mas podendo ser na posição P2D’, P2R’ e P3R’ pode favorecer a especificidade da interação.

O aminoácido ARG 179 interage preferencialmente com a posição P3D’, mas também interage com as posições P4D’ e P4R’. Interagindo preferencialmente com o átomo O6 da guanina esse aminoácido busca por cargas negativas, interagindo também com o átomo N7 de guanina e adeninas e O4 das timinas. Então a existência de nucleotídeos com carga parcial negativa (G, A e T pelos átomos O6, N7 e O4 respectivamente) na posição P3D’ pode favorecer a especificidade.

O aminoácido ARG 180 interage preferencialmente com a posição P4R’, mas também interage com a posição P5R’. Interagindo preferencialmente com os átomos N7 e O6 da guanina, esse aminoácido busca por cargas negativas, interagindo também com o átomo N7 de adeninas e O4 das timinas. Assim, a existência de nucleotídeos com carga parcial negativa (G, A e T pelos átomos O6, N7 e O4 respectivamente) na posição P4R’ pode favorecer a especificidade.

A melhor seqüência para uma maior especificidade da ligação do RXR com o DNA, baseado na preferência pela base e pelo por nucleotídeo, a seqüência deve ter as seguintes bases nas seguintes posições:

Direta	-1D	P1D	P2D	P3D	P4D	P5D
		G	G/A			
	T/C		C	C/A	G	G/A
Reversa	-1R	P1R	P2R	P3R	P4R	P5R
Consenso	A/G	G	G	T/G	C	C/T

A posição no complexo com o DNA nativo que apresentou maior quantidade de ligações de hidrogênio específicas foi a posição 8, cuja a seqüência é a seguinte:

```
  -3  P1   P6   S5
   |   |    |    |
CGGGTAGGGTTCACCGAAAGTTCACTCGCATATA
```

A seqüência da melhor posição apresenta cinco das seis bases prevista para ter a melhor especificidade. Apenas ocorre uma divergência na posição P4.

A melhor seqüência para uma maior especificidade da ligação do RAR com o DNA, baseado na preferência pela base e pelo por nucleotídeo, a seqüência deve ter as seguintes bases nas seguintes posições:

Direta	P1D’	P2D’	P3D’	P4D’	P5D’
	G/A	A/G	G		
		A/C	G/C	G	G
Reversa	P1R’	P2R’	P3R’	P4R’	P5R’
Consenso	G/A	A/G/T	C/G	C	C

A posição no complexo com o DNA nativo que apresentou maior quantidade de ligações de hidrogênio específicas foi a posição 8, cuja a seqüência para a região é a seguinte:

```
                 S3 P1’  P6’
                 |  |    |
CGGGTAGGGTTCACCGAAAGTTCACTCGCATATA
```

Podemos observar que a seqüência consenso para a ligação do RXR é mais estringente do que a seqüência para a ligação do RAR, sugerindo que a maior especificidade da ligação é definida pelo RXR.

5. Conclusões

- A falta de informação para alguns fatores de transcrição inviabiliza a utilização do método de busca por sítios de fatores de transcrição utilizando matrizes.

- Foi apresentado um novo método para predição de sítios de ligação de fatores de transcrição, que utiliza das informações estruturais dos fatores de transcrição e do DNA e métodos termodinâmicos. O método conseguiu prever uma região de maior probabilidade em torno da posição onde realmente ligam os fatores de transcrição RXR e RAR na região promotora do gene RARE-ß2. O método não mostrou especificidade suficiente para prever a exata localização do sítio, o que pode ser decorrente de uma limitação metodológica e/ou principalmente, da complexidade dos mecanismos cis-regulatórios que podem apresentar redundância, para garantir a expressão de um dado gene. O aumento do tempo da dinâmica molecular poderia fornecer uma possível resposta para estas questões.

- Foram analisadas as ligações de hidrogênio realizadas entre os fatores de transcrição regulados pelo ácido retinóico e um conjunto de diferentes seqüências de DNA. A partir dessas análises foi possível verificar que apenas cinco bases são responsáveis pela especificidade da interação entre um fator de transcrição e o DNA. Para a região promotora do gene RARE-ß2 foi possível verificar que a especificidade da interação é regida pelo RXR sendo que o RAR possui uma maior tolerância em relação à especificidade da ligação.

CONCLUSÕES GERAIS

Em relação a etapa da síntese do ácido retinóico:

1. O método de buscar por seqüências de ALDHs em organismos que possuem o genoma seqüenciado e anotado e também em vários organismos basais que ainda não possuíam o seus genomas montados e/ou anotados, utilizando as informações provindas dos projetos de seqüenciamento de EST mostrou-se eficaz reconstruindo os genes das ALDHs do anfioxo.
2. A análise filogenética das seqüências das ALDHs para o estudo evolutivo dos grupos de ALDHs mostrou que as ALDH1s e as ALDH2s são membros de uma única e bem suportada subfamília e estas ALDH2s formam um clado distinto que é agrupado em ALDH1/2.
3. Buscar por assinaturas e a modelagem molecular localizou três aminoácidos no interior do canal de acesso ao substrato que auxiliam no entendimento da especificidade das ALDH2 e ALHD1 por diferentes substratos.
4. O cálculo do volume dos canais verificou uma relação entre a forma e o volume do canal e a função da enzima (sinalização ou destoxificação). Canais grandes estão relacionados a sinalização enquanto que canais pequenos estão relacionados a destoxificção.
5. Após as duplicações da ALDH1 de anfioxo ocorreram modificações tanto na estrutura da proteína como na regulação gênica, resultando em novas funções, diferentes das ancestrais, envolvendo controle do padrão corporal, para algo mais abrangente, como funções protetoras.

Em relação a etapa da via de sinalização pelo ácido retinóico:

1. O método para a identificação de sítios de ligação de fatores de transcrição baseada em algoritmos de análise estrutural de macromoléculas biológicas e métodos termodinâmicos foi capaz de prever uma região com maior probabilidade de ligação dos fatores em torno do sítio de ligação mostrando ser uma técnica promissora.
2. Análise das ligações de hidrogênio entre os DNAs e os fatores de transcrição auxiliou no entendimento da especificidade da ligação entre os fatores de transcrição regulados por ácido retinóico e o DNA.

REFERÊNCIAS BIBLIOGRÁFICAS

Abascal, F., R. Zardoya, et al. (2005). "ProtTest: selection of best-fit models of protein evolution." Bioinformatics **21**(9): 2104-5.

Abu-Abed, S., P. Dolle, et al. (2001). "The retinoic acid-metabolizing enzyme, CYP26A1, is essential for normal hindbrain patterning, vertebral identity, and development of posterior structures." Genes Dev **15**(2): 226-40.

Ahvazi, B., R. Coulombe, et al. (2000). "Crystal structure of the NADP+-dependent aldehyde dehydrogenase from Vibrio harveyi: structural implications for cofactor specificity and affinity." Biochem J **349 Pt 3**: 853-61.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Berezikov, E., R. H. Plasterk, et al. (2002). "GENOTRACE: cDNA-based local GENOme assembly from TRACE archives." Bioinformatics **18**(10): 1396-7.

Birney, E., M. Clamp, et al. (2004). "GeneWise and Genomewise." Genome Res **14**(5): 988-95.

Blentic, A., E. Gale, et al. (2003). "Retinoic acid signalling centres in the avian embryo identified by sites of expression of synthesising and catabolising enzymes." Dev Dyn **227**(1): 114-27.

Canestro, C., J. H. Postlethwait, et al. (2006). "Is retinoic acid genetic machinery a chordate innovation?" Evol Dev **8**(5): 394-406.

Carey, M., Smale, S. T. (2002). "Biochemistry and Molecular Biology Education." Biochemistry and Molecular Biology Education , Christopher K. Mathews.

Cartharius, K., K. Frech, et al. (2005). "MatInspector and beyond: promoter analysis based on transcription factor binding sites." Bioinformatics **21**(13): 2933-42.

Case D.A. , T. A. D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R., K. M. M. Luo, D.A. Pearlman, M. Crowley, R.C. Walker, W. Zhang, B. Wang, S., et al. (2006). Amber 9. University of California. U. o. California. California
University of California
0.

Corpet, F. (1988). "Multiple sequence alignment with hierarchical clustering." Nucleic Acids Res **16**(22): 10881-90.

Davidovici, B. B., Y. Tuzun, et al. (2007). "Retinoid receptors." Dermatol Clin **25**(4): 525-30, viii.

Dersch, H. and M. H. Zile (1993). "Induction of normal cardiovascular development in the vitamin A-deprived quail embryo by natural retinoids." Dev Biol **160**(2): 424-33.

Dong, Q., M. D. Wilkerson, et al. (2007). "Tracembler--software for in-silico chromosome walking in unassembled genomes." BMC Bioinformatics **8**: 151.

Dupe, V., N. Matt, et al. (2003). "A newborn lethal defect due to inactivation of retinaldehyde dehydrogenase type 3 is prevented by maternal retinoic acid treatment." Proc Natl Acad Sci U S A **100**(24): 14036-41.

Edgar, R. C. (2004). "MUSCLE: a multiple sequence alignment method with reduced time and space complexity." BMC Bioinformatics **5**: 113.

Fan, X., A. Molotkov, et al. (2003). "Targeted disruption of Aldh1a1 (Raldh1) provides evidence for a complex mechanism of retinoic acid synthesis in the developing retina." Mol Cell Biol **23**(13): 4637-48.

Fessele, S., H. Maier, et al. (2002). "Regulatory context is a crucial part of gene function." Trends Genet **18**(2): 60-3.

Franca, L. T., E. Carrilho, et al. (2002). "A review of DNA sequencing techniques." Q Rev Biophys **35**(2): 169-200.

Glass, C. K. (1994). "Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers." Endocr Rev **15**(3): 391-407.

Heery, D. M., B. Pierrat, et al. (1994). "Homo- and heterodimers of the retinoid X receptor (RXR) activated transcription in yeast." Nucleic Acids Res **22**(5): 726-31.

Hochgreb, T., V. L. Linhares, et al. (2003). "A caudorostral wave of RALDH2 conveys anteroposterior information to the cardiac field." Development **130**(22): 5363-74.

Holland, L. Z. (1996). Molecular Zoology: Advances Strategies, and protocols.

http://www.R-project.org (2005).

http://www.tcl.tk/.

Huang, X. and A. Madan (1999). "CAP3: A DNA sequence assembly program." Genome Res **9**(9): 868-77.

Humphrey, W., A. Dalke, et al. (1996). "VMD: visual molecular dynamics." J Mol Graph **14**(1): 33-8, 27-8.

Iwama, H. and T. Gojobori (2004). "Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network." Proc Natl Acad Sci U S A **101**(49): 17156-61.

Kawaguchi, R., J. Yu, et al. (2007). "A membrane receptor for retinol binding protein mediates cellular uptake of vitamin A." Science **315**(5813): 820-5.

Kel-Margoulis, O. V., A. E. Kel, et al. (2002). "TRANSCompel: a database on composite regulatory elements in eukaryotic genes." Nucleic Acids Res **30**(1): 332-4.

Kel, A. E., E. Gossling, et al. (2003). "MATCH: A tool for searching transcription factor binding sites in DNA sequences." Nucleic Acids Res **31**(13): 3576-9.

Kleywegt, G. J. and T. A. Jones (1994). "Detection, delineation, measurement and display of cavities in macromolecular structures." Acta Crystallogr D Biol Crystallogr **50**(Pt 2): 178-85.

Klingenhoff, A., K. Frech, et al. (2002). "Regulatory modules shared within gene classes as well as across gene classes can be detected by the same in silico approach." In Silico Biol **2**(1): S17-26.

Laskowski, R. A., D. S. Moss, et al. (1993). "Main-chain bond lengths and bond angles in protein structures." J Mol Biol **231**(4): 1049-67.

Laurie, A. T. and R. M. Jackson (2005). "Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites." Bioinformatics **21**(9): 1908-16.

Lewin, B. (2001). Genes VII, Artmed.

Liu, L. A. and J. S. Bader (2007). "Ab initio prediction of transcription factor binding sites." Pac Symp Biocomput: 484-95.

Lowe, E. D., G. Y. Gao, et al. (2008). "Structure of daidzin, a naturally occurring anti-alcohol-addiction agent, in complex with human mitochondrial aldehyde dehydrogenase." J Med Chem **51**(15): 4482-7.

Luo, R., L. David, et al. (2002). "Accelerated Poisson-Boltzmann calculations for static and dynamic systems." J Comput Chem **23**(13): 1244-53.

Luscombe, N. M., S. E. Austin, et al. (2000). "An overview of the structures of protein-DNA complexes." Genome Biol **1**(1): REVIEWS001.

Maden, M. (2007). "Retinoic acid in the development, regeneration and maintenance of the nervous system." Nat Rev Neurosci **8**(10): 755-65.

Maxam, A. M. and W. Gilbert (1977). "A new method for sequencing DNA." Proc Natl Acad Sci U S A **74**(2): 560-4.

McGrane, M. M. (2007). "Vitamin A regulation of gene expression: molecular mechanism of a prototype gene." J Nutr Biochem **18**(8): 497-508.

Miyamoto, S. and P. A. Kollman (1992). "SETTLE: An analytical version of the SHAKE and RATTLE
algorithm for rigid water models." J. Comput. Chem. **13**: 952-962.

Molotkov, A., X. Fan, et al. (2002). "Stimulation of retinoic acid production and growth by ubiquitously expressed alcohol dehydrogenase Adh3." Proc Natl Acad Sci U S A **99**(8): 5337-42.

Moore, S. A., H. M. Baker, et al. (1998). "Sheep liver cytosolic aldehyde dehydrogenase: the structure reveals the basis for the retinal specificity of class 1 aldehyde dehydrogenases." Structure **6**(12): 1541-51.

Morriss, G. M. (1972). "Morphogenesis of the malformations induced in rat embryos by maternal hypervitaminosis A." J Anat **113**(Pt 2): 241-50.

Moss, J. B., J. Xavier-Neto, et al. (1998). "Dynamic patterns of retinoic acid synthesis and response in the developing mammalian heart." Dev Biol **199**(1): 55-71.

Niederreither, K., P. McCaffery, et al. (1997). "Restricted expression and retinoic acid-induced downregulation of the retinaldehyde dehydrogenase type 2 (RALDH-2) gene during mouse development." Mech Dev **62**(1): 67-78.

Niederreither, K., V. Subbarayan, et al. (1999). "Embryonic retinoic acid synthesis is essential for early mouse post-implantation development." Nat Genet **21**(4): 444-8.

Niederreither, K., J. Vermot, et al. (2001). "Embryonic retinoic acid synthesis is essential for heart morphogenesis in the mouse." Development **128**(7): 1019-31.

Novotny, M., D. Madsen, et al. (2004). "Evaluation of protein fold comparison servers." Proteins **54**(2): 260-70.

Perlmann, T., P. N. Rangarajan, et al. (1993). "Determinants for selective RAR and TR recognition of direct repeat HREs." Genes Dev **7**(7B): 1411-22.

Pertea, G., X. Huang, et al. (2003). "TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets." Bioinformatics **19**(5): 651-2.

Petrey, D., Z. Xiang, et al. (2003). "Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling." Proteins **53 Suppl 6**: 430-5.

Putnam, N. H., T. Butts, et al. (2008). "The amphioxus genome and the evolution of the chordate karyotype." Nature **453**(7198): 1064-71.

Quandt, K., K. Frech, et al. (1995). "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data." Nucleic Acids Res **23**(23): 4878-84.

Rastinejad, F., T. Wagner, et al. (2000). "Structure of the RXR-RAR DNA-binding complex on the retinoic acid response element DR1." Embo J **19**(5): 1045-54.

Ray, W. J., G. Bain, et al. (1997). "CYP26, a novel mammalian cytochrome P450, is induced by retinoic acid and defines a new family." J Biol Chem **272**(30): 18702-8.

Reijntjes, S., E. Gale, et al. (2004). "Generating gradients of retinoic acid in the chick embryo: Cyp26C1 expression and a comparative analysis of the Cyp26 enzymes." Dev Dyn **230**(3): 509-17.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-4.

Rosenthal, N. and J. Xavier-Neto (2000). "From the bottom of the heart: anteroposterior decisions in cardiac muscle differentiation." Curr Opin Cell Biol **12**(6): 742-6.

Ryckaert, J. P., G. Ciccotti, et al. (1977). "Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes." Comput. Phys **23**: 327-341.

Sakai, Y., C. Meno, et al. (2001). "The retinoic acid-inactivating enzyme CYP26 is essential for establishing an uneven distribution of retinoic acid along the anterio-posterior axis within the mouse embryo." Genes Dev **15**(2): 213-25.

Sandell, L. L., B. W. Sanderson, et al. (2007). "RDH10 is essential for synthesis of embryonic retinoic acid and is required for limb, craniofacial, and organ development." Genes Dev **21**(9): 1113-24.

Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-7.

Sawin, P. B. and H. W. Edmonds (1949). "Morphogenetic studies of the rabbit; aortic arch variations in relation to regionally specific growth differences." Anat Rec **105**(3): 377-97.

Simoes-Costa, M. S., A. P. Azambuja, et al. (2008). "The search for non-chordate retinoic acid signaling: lessons from chordates." J Exp Zoolog B Mol Dev Evol **310**(1): 54-72.

Simoes-Costa, M. S., M. Vasconcelos, et al. (2005). "The evolutionary origin of cardiac chambers." Dev Biol **277**(1): 1-15.

Sobreira, T. J. and A. Gruber (2008). "Sequence-specific reconstruction from fragmentary databases using seed sequences: implementation and validation on SAGE, proteome and generic sequencing data." Bioinformatics **24**(15): 1676-80.

Swindell, E. C. and G. Eichele (1999). "Retinoid metabolizing enzymes in development." Biofactors **10**(2-3): 85-9.

Swindell, E. C., C. Thaller, et al. (1999). "Complementary domains of retinoic acid production and degradation in the early chick embryo." Dev Biol **216**(1): 282-96.

Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Vasiliou, V., A. Bairoch, et al. (1999). "Eukaryotic aldehyde dehydrogenase (ALDH) genes: human polymorphisms, and recommended nomenclature based on divergent evolution and chromosomal mapping." Pharmacogenetics **9**(4): 421-34.

Vasiliou, V., A. Pappa, et al. (2004). "Role of human aldehyde dehydrogenases in endobiotic and xenobiotic metabolism." Drug Metab Rev **36**(2): 279-99.

Vermot, J., K. Niederreither, et al. (2003). "Decreased embryonic retinoic acid synthesis results in a DiGeorge syndrome phenotype in newborn mice." Proc Natl Acad Sci U S A **100**(4): 1763-8.

Vivanco Ruiz, M. M., T. H. Bugge, et al. (1991). "Functional characterization of a natural retinoic acid responsive element." Embo J **10**(12): 3829-38.

Werner, T., S. Fessele, et al. (2003). "Computer modeling of promoter organization as a tool to study transcriptional coregulation." Faseb J **17**(10): 1228-37.

Wilson, J. G. and S. Barch (1949). "Fetal death and maldevelopment resulting from maternal vitamin A deficiency in the rat." Proc Soc Exp Biol Med **72**(3): 687-93, illust.

Wilson, J. G. and J. Warkany (1950). "Cardiac and aortic arch anomalies in the offspring of vitamin A deficient rats correlated with similar human anomalies." Pediatrics **5**(4): 708-25.

Wingender, E., P. Dietze, et al. (1996). "TRANSFAC: a database on transcription factors and their DNA binding sites." Nucleic Acids Res **24**(1): 238-41.

www.aldh.org.

Xavier-Neto, J., N. Rosenthal, et al. (2001). "Retinoid signaling and cardiac anteroposterior segmentation." Genesis **31**(3): 97-104.

Yoshida, A. (1994). "Genetic polymorphisms of alcohol metabolizing enzymes related to alcohol sensitivity and alcoholic diseases." Alcohol Alcohol **29**(6): 693-6.

Yoshida, A., A. Rzhetsky, et al. (1998). "Human aldehyde dehydrogenase gene family." Eur J Biochem **251**(3): 549-57.

Yu, J. K., Y. Satou, et al. (2007). "Axial patterning in cephalochordates and the evolution of the organizer." Nature **445**(7128): 613-7.

Zechel, C., X. Q. Shen, et al. (1994). "The dimerization interfaces formed between the DNA binding domains of RXR, RAR and TR determine the binding specificity and polarity of the full-length receptors to direct repeats." Embo J **13**(6): 1425-33.

Zelent, A., C. Mendelsohn, et al. (1991). "Differentially expressed isoforms of the mouse retinoic acid receptor beta generated by usage of two promoters and alternative splicing." Embo J **10**(1): 71-81.

Zhao, D., P. McCaffery, et al. (1996). "Molecular identification of a major retinoic-acid-synthesizing enzyme, a retinaldehyde-specific dehydrogenase." Eur J Biochem **240**(1): 15-22.

Ziouzenkova, O. and J. Plutzky (2008). "Retinoid metabolism and nuclear receptor responses: New insights into coordinated regulation of the PPAR-RXR complex." FEBS Lett **582**(1): 32-8.

Anexo 1

Tamanho das seqüências de aminoácidos encontradas nos organismos analisados utilizando o método de busca por seqüências de ALDH.

	ALDH1A1	ALDH1A2	ALDH1A3	ALDH1B1	ALDH1L1	ALDH1L2	ALDH2	ALDH3A1	ALDH3A2	ALDH3B1	ALDH3B2	ALDH4	ALDH5	ALDH6	ALDH7	ALDH8	ALDH9	ALDH16	ALDH18
Acanthoscurria gomesiana					490			342								420			
Acyrthosiphon pisum	457	448			495						519			485	512	499			
Adiantum capillus								392											
Aedes aegypti	488				514			485							511	496			730
Aegilops speltoides								380											
Agrostis stolonifera					348														
Ajellomyces capsulatus					433														
Alligator mississippiensis							379												
Allium cepa		347				358													
Amborella trichopoda					433	473		491						369					
Ambystoma mexicanum					318														
Ambystoma tigrinum									336			343			344				
Ancylostoma caninum		415						462											
Ancylostoma ceylanicum					471											318			
Anopheles gambiae	333			365	486		504	478						338	513	500		531	733
Antirrhinum majus					492														
Antrodia cinnamomea								393											
Apis mellifera					437														
Aplysia californica	390													484					
Aquilegia formosa		456	455		491	470									501				706
Arabidopsis thaliana					492	491		410								463			
Arachis hypogaea		395			354	397		337						463					
Artemia franciscana					470														
Artemisia annua					502			456											346
Asparagus officinalis															391				360
Aspergillus niger					430		483	436								491			
Aureococcus anophagefferens							401	478								497			
Beta vulgaris		461																	
Betula pendula															381				
Blastocladiella emersonii															464				
Bombyx mori	458					494	480	505	471	379	470				370	502		311	375
Bos indicus																317			
Bos taurus	485		468		505	498	404	513		499	422	520	395	502	511	534	800	563	795
Bothrops jararaca							383												
Botryotinia fuckeliana								474						469					
Brachypodium distachyon															501				
Branchiostoma floridae								494	488						509				
Brassica napus		476			497	485	482	496				458		479	502	496			708
Brassica oleracea								484						479	502	496		312	
Brassica rapa		458			480			496	430					479	502	496			693
Bursaphelenchus xylophilus								363											
Caenorhabditis elegans	482	428			465						345				512				708
Caenorhabditis remanei											528				512	507			
Caenorhabditis sp																476			
Callinectes sapidus							439												
Canis lupus	453				506													564	795
Capitella sp					483			486				339			508	499			
Capra hircus					505			513											390
Capsicum annuum					301														
Carcinus maenas					403														
Carthamus tinctorius		455			499									479	501	465			
Catharanthus roseus					320														
Cavia porcellus	455				501														
Centaurea maculosa		473	458		492				473						501				529
Centaurea solstitialis		458			452		367									481			693

	ALDH1A1	ALDH1A2	ALDH1A3	ALDH1B1	ALDH1L1	ALDH1L2	ALDH2	ALDH3A1	ALDH3A2	ALDH3B1	ALDH3B2	ALDH4	ALDH5	ALDH6	ALDH7	ALDH8	ALDH9	ALDH16	ALDH18
Chlamydomonas reinhardtii								472								344			
Cicer arietinum								482											
Cichorium endivia		460			489		377								501	301			367
Cichorium intybus		471			497										501				692
Ciona intestinalis		379			306			598			911	494		465	511			535	727
Ciona savignyi															511				
Citrus clementina					507		399	482							479				706
Citrus sinensis					492									324					
Citrus x														313					570
Clonorchis sinensis								485											
Coccidioides posadasii							497	490								491			
Coffea canephora					486											324			399
Colletotrichum trifolii																		312	
Coprinopsis cinerea								317	329										
Crassostrea gigas					430														
Cryphonectria parasitica		337																	
Cryptococcus neoformans			461	460				475	480						498				
Cryptomeria japonica					495														
Cryptosporidium muris		436					423												
Ctenocephalides felis							430												
Cucumis melo							339								350				
Cucumis sativus			303																
Culex pipiens																308			
Curcuma longa					313		475												
Cycas rumphii					430														
Danaus plexippus					483			495								382			
Danio rerio	482			456	483	518		501	403		904	495	480		511	503		334	743
Daphnia magna			355		417														
Diabrotica virgifera						322													
Dictyostelium discoideum					475										364	507		420	
Drosophila ananassae			364							371									
Drosophila erecta					431											329		378	
Drosophila grimshawi			312					480		319						421		381	
Drosophila melanogaster			483				502	488							510	502		532	730
Drosophila mojavensis	390		429			469		438						449					
Drosophila pseudoobscura							486												
Drosophila simulans								345											
Drosophila virilis								414							421	402			
Drosophila willistoni					315			504							507	502			
Dugesia ryukyuensis					485						442							473	
Echinococcus granulosus					430														
Elaeis guineensis														441					
Epiphyas postvittana					311				319										
Eucalyptus globulus																			351
Euglena gracilis			450				373	500										376	
Euphorbia esula		451				352		373											
Euprymna scolopes					437														
Festuca arundinacea					337		473	483											639
Fragaria vesca		343			508										503				485
Fragaria x																406			
Fundulus heteroclitus	327																		
Gadus morhua			330																
Gallus gallus	483		355		496	482		519	365		453	512	487	496	511	476		510	503
Gasterosteus aculeatus	480				518	501		508		903		497		435	507	441		535	738
Georissus sp																404			
Gibberella moniliformis				448	443	321	481	493				488		506		526		529	
Gibberella zeae							492							477				352	
Gillichthys mirabilis										353						379			
Ginkgo biloba					372			381											

	ALDH1A1	ALDH1A2	ALDH1A3	ALDH1B1	ALDH1L1	ALDH1L2	ALDH2	ALDH3A1	ALDH3A2	ALDH3B1	ALDH3B2	ALDH4	ALDH5	ALDH6	ALDH7	ALDH8	ALDH9	ALDH16	ALDH18
Globodera rostochiensis								302											
Gloeophyllum trabeum																348		331	
Glossina morsitans	444				502			482			582				507	502			730
Glycine max	450				498		491	485						479	464	488			
Glycine soja					482														
Gnetum gnemon						484		476											321
Gossypium arboreum		437				331		490											
Gossypium hirsutum		472	362		504	490								476		354			333
Gossypium raimondii		472			500	487		317							430	502			400
Haematobia irritans		349			406														527
Haemonchus contortus																314			
Halocynthia roretzi	340														514	348			
Hebeloma cylindrosporum																327			
Helianthus annuus					497	490								310	501	350			
Helianthus argophyllus		417			478	329			338					459	351	490			366
Helianthus ciliaris					484									479	400				
Helianthus exilis		383																	418
Helianthus paradoxus					478	363								443	387				
Helianthus petiolaris					510			305						309	501	484			
Helianthus tuberosus			302		492	494								486	467				614
Heliconius erato																		439	
Heliothis virescens							348												
Helobdella robusta	451	447			439		496	480				303				500		546	
Heterodera glycines			377												474	486			415
Heterorhabditis bacteriophora															437	406			
Hippoglossus hippoglossus						323								371					
Hirudo medicinalis		445							464					427					
Homalodisca coagulata					498														
Hordeum vulgare	443		450		473	349	497	502						479	501	518			708
Hydra magnipapillata		392			488							381		484		311			
Hypocrea jecorina							475	518						470					
Ichthyophthirius multifiliis															406	408			
Ictalurus punctatus			409									494			326				
Idiosepius paradoxus					337														
Isochrysis galbana			373													506			
Ixodes scapularis	456						327	506							507	495		343	429
Jakoba bahamiensis								470											
Jakoba libera								482							493				
Juglans hindsii								442											
Karenia brevis	373		429																
Knorringia sibirica							434												
Laccaria bicolor						472								340					
Lactuca perennis			455		479											513			
Lactuca saligna					488														
Lactuca sativa		441			480									453					411
Lactuca serriola		455			455			473							436				408
Lactuca virosa					306											505			512
Laupala kohalensis						327									340				
Leishmania donovani							316												
Lentinula edodes																348			
Leymus cinereus			450		343										501				
Linum usitatissimum															378				
Liriodendron tulipifera					315														313
Locusta migratoria		435			483	455													
Lolium multiflorum															383	463			
Lotus japonicus					354		375								347				
Lutzomyia longipalpis	429							496								500			
Lysiphlebus testaceipes						363													
Macaca fascicularis					433							340							

	ALDH1A1	ALDH1A2	ALDH1A3	ALDH1B1	ALDH1L1	ALDH1L2	ALDH2	ALDH3A1	ALDH3A2	ALDH3B1	ALDH3B2	ALDH4	ALDH5	ALDH6	ALDH7	ALDH8	ALDH9	ALDH16	ALDH18
Macaca mulatta	453				515			376				517							
Macaca nemestrina							411	517											
Macropus eugenii			429																
Macrostomum lignano							313												
Magnaporthe grisea							420												
Malawimonas californiana							482	431											
Malus x	449	449			513	487								478	477	494			708
Manihot esculenta					335											475			
Marchantia polymorpha		453			473			311											
Medicago sativa														430					
Medicago truncatula	338	441			504									485	467	478			692
Meleagris gallopavo															330				
Meloidogyne arenaria							322												
Mesembryanthemum crystallinum																317			
Metarhizium anisopliae							414												
Microbotryum violaceum											439								
Micromonas sp																495			
Misgurnus anguillicaudatus															394				
Molgula tectiformis						500													
Monosiga ovata							370								305				
Mytilus californianus					504							485							
Myzus persicae					487														
Nasonia giraulti	374	456			450			498											
Nasonia vitripennis								489							310				
Nematostella vectensis					487	407		503							507	508			497
Neurospora crassa								310											
Nicotiana langsdorffii					497														
Nicotiana tabacum			465		497	492								489	387	501			
Nuphar advena					446										332	307			
Ocimum basilicum																500			
Oncometopia nigricans															394				
Oncorhynchus mykiss	483		449					518				493			511	432			
Oncorhynchus nerka						520		470				359							
Oncorhynchus tshawytscha																440			
Oryctolagus cuniculus					448														
Oryza sativa	447		468		311	478	501	508						479	501	494			
Oryzias latipes	480		466									493	486	303	511	504		383	
Oscarella carmela					433									362					
Osmerus mordax												493				503			
Ostertagia ostertagi					470														
Ostreococcus lucimarinus							460												
Ostrinia nubilalis					452														
Ovis aries		485			505			513				520		385	364	457	460		539
Panorpa cf								476											
Papio anubis					515												314		469
Paracentrotus lividus	370				644		330	516								505			551
Paracoccidioides brasiliensis							455	494											
Paralabidochromis chilotes					470														
Paramecium tetraurelia		447													401	432			
Pavlova lutheri															310				
Perkinsus marinus					414		453					483				497			
Persea americana					380			315						304		324			
Phaeodactylum tricornutum			439		363									475		485			757
Phaeosphaeria nodorum			455		309											495			
Phakopsora pachyrhizi								470								322		496	
Phalaenopsis equestris		371																	
Phaseolus vulgaris					497				335					317		342			
Phlebotomus papatasi														483	340				310
Phycomyces blakesleeanus					467		477	477						496		499		533	

	ALDH1A1	ALDH1A2	ALDH1A3	ALDH1B1	ALDH1L1	ALDH1L2	ALDH2	ALDH3A1	ALDH3A2	ALDH3B1	ALDH3B2	ALDH4	ALDH5	ALDH6	ALDH7	ALDH8	ALDH9	ALDH16	ALDH18
Physarum polycephalum			436		420		481								345	381		380	
Physcomitrella patens	457				507		317	485						476	500	499			705
Phytophthora infestans							444	326	424			324		490		506			356
Phytophthora parasitica								471						404					
Phytophthora sojae								508				483				505			372
Picea glauca					499			487							501				371
Picea sitchensis					480	333		487											
Pimephales promelas	344						311					461				502	553	551	
Pinus pinaster					479			374											
Pinus taeda	371	471			495	482		315						476	501	489			418
Platichthys flesus												465							
Plodia interpunctella							407												
Poncirus trifoliata					471														
Pongo pygmaeus		360			501					306		412				333			
Populus nigra												436							
Populus tremula					489		478	481	313							316			
Populus trichocarpa					471			487	479						463	391			
Populus x														430					
Porphyra yezoensis								494						319					
Pratylenchus vulnus								416											
Prunus persica	450							488											
Psetta maxima								308											
Pseudotsuga menziesii					448														
Ptyochromis sp												333							
Raphanus raphanistrum		336			497			485		353					501	501			560
Rattus sp																317			
Reclinomonas americana							478									494			
Rhipicephalus appendiculatus					491			492								500			
Rhipicephalus microplus					463			488							511				
Rhizopus oryzae						480									325				
Ribes americanum					349														
Ricinus communis					505			499											
Rosa hybrid					472														
Rutilus rutilus												357							
Saccharum officinarum			452		414	415	483	501	446					479	501	488			708
Saitoella complicata								359										381	
Salmo salar	452					520		518				350		484	511	529		322	
Saruma henryi					479														
Sawyeria marylandensis					349													520	
Schistosoma japonicum												460							
Schistosoma mansoni		453			484														
Schmidtea mediterranea	363				486											507			
Secale cereale																505			
Seculamonas ecuadoriensis								394							308	395			
Selaginella lepidophylla												324							
Solanum habrochaites	457																		
Solanum lycopersicum		457			499			367						510		489			
Solanum tuberosum		465			497	339								510	501	483			
Sorghum bicolor			456		323		497	471						479	501				361
Sorghum propinqum								307											
Spodoptera frugiperda					482		485								511				
Squalus acanthias	447														352				
Stachyamoeba lipophora															457				
Stevia rebaudiana					478														394
Strongylocentrotus purpuratus								500			722								
Syntrichia ruralis							458						472						
Taeniopygia guttata								503											
Takifugu rubripes														428					
Taraxacum kok																532			

	ALDH1A1	ALDH1A2	ALDH1A3	ALDH1B1	ALDH1L1	ALDH1L2	ALDH2	ALDH3A1	ALDH3A2	ALDH3B1	ALDH3B2	ALDH4	ALDH5	ALDH6	ALDH7	ALDH8	ALDH9	ALDH16	ALDH18
Taraxacum officinale		455			434									465					
Tetrahymena thermophila															365	424			
Thellungiella salsuginea					315										357				375
Toxoplasma gondii							472												
Tribolium castaneum		452					496	456						482		498		310	730
Trichophyton rubrum					316			488								355			
Trichosurus vulpecula	377				496	334		325					487		511	532			
Triphysaria pusilla								433								480			
Triphysaria versicolor		330					307									349			
Triticum aestivum	468		447		485	487	484	497				405		479	480	518			709
Triticum monococcum								353								425			
Triticum turgidum															448	314			422
Ustilago maydis																520			
Vitis hybrid					498							452							
Vitis shuttleworthii		471																	
Vitis vinifera	471	468			498	506		456	453			311		486	501	356			
Welwitschia mirabilis						353									414	482			470
Xenopus laevis	483				518	476		498			763	516			511	504			423
Xenopus tropicalis	501			468	518			813		360	902	495	483	484	511	504		541	755
Zea mays	452		449	449	433	483	505	501				482		469	501	479			593
Zingiber officinale	446				475														
uncultured fungus								480											

Anexo 2

Nome do organismo, classificação da seqüência de ALDH, volume do canal em Å3, estrutura utilizada como molde para modelagem molecular.

Organismo	ALDH1Ls	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Caenorhabditis elegans	ALDH1L1	540,1		X	
Nematostella vectensis	ALDH1L	469,0		X	
Drosophila melanogaster	ALDH1L1	505,7		X	
Ciona savignyi	ALDH1L	463,5		X	
Homo sapiens	ALDH1L1	466,2		X	
Rattus norvegicus	ALDH1L1	448,0		X	
Mus musculus	ALDH1L1	521,0		X	
Mus musculus	ALDH1L2	463,1		X	
Strongylocentrotus purpuratus	ALDH1L	466,3	X		
Branchiostoma floridae	ALDH1L	414,1		X	
	Média	475,7			
	SEM	11,6			
	Desvio padrão	36,8			
Organismo	TUNICADOS ALDH1s -a-	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Ciona savignyi	ALDH1	546,0	X		
Ciona intestinalis	ALDH1	640,0	X		
Ciona savignyi	ALDH1	561,0	X		
Ciona intestinalis	ALDH1	514,0	X		
Ciona intestinalis	ALDH1	467,0	X		
	Média	545,6			
	SEM	28,6			
	Desvio padrão	63,9			
Organismo	ECDISOZOA ALDHs -b-	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Caenorhabditis elegans	ALDH1J2	345,3	X		
Caenorhabditis elegans	ALDH9B2	281,0	X		
Drosophila melanogaster	ALDH1M1	399,2	X		
	Média	341,8			
	SEM	34,2			
	Desvio padrão	59,2			
Organismo	ALDH1s -c- -d- -e-	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Nematostella vectensis	ALDH1	-			X
Nematostella vectensis	ALDH1	622,7		X	
Xenopus laevis	ALDH1	664,8		X	
Xenopus tropicalis	ALDH1A1	690,0		X	
Ornithorhynchus anatinus	ALDH1A1	615,3		X	
Monodelphis domestica	ALDH1A1	607,9		X	
Gallus gallus	ALDH1A1	641,8		X	
Rattus norvegicus	ALDH1A1	593,8		X	
Mus musculus	ALDH1A1	551,0		X	
Oryctolagus cuniculus	ALDH1A1	636,8		X	
Canis familiaris	ALDH1A1	561,5		X	
Bos taurus	ALDH1A1	614,7		X	
Ovis aries	ALDH1A1	662,8		X	
Pongo abelii	ALDH1A1	631,3		X	
Pan troglodytes	ALDH1A1	592,6		X	
Macaca mulatta	ALDH1A1	604,0		X	
Macaca fascicularis	ALDH1A1	591,9		X	
Homo sapiens	ALDH1A1	639,5		X	
Oryzias latipes	ALDH1A2	584,5		X	
Danio rerio	ALDH1A2	609,0		X	
Danio rerio	ALDH1A2	527,7		X	
Takifugu rubripes	ALDH1A2	602,9		X	
Xenopus laevis	ALDH1A2	631,0		X	
Xenopus tropicalis	ALDH1A2	577,7		X	
Gallus gallus	ALDH1A2	548,7		X	
Monodelphis domestica	ALDH1A2	611,7		X	
Rattus norvegicus	ALDH1A2	634,0		X	
Mus musculus	ALDH1A2	566,8		X	
Taeniopygia guttata	ALDH1A2	636,0		X	
Canis familiaris	ALDH1A2	609,5		X	
Bos taurus	ALDH1A2	565,8		X	
Pan troglodytes	ALDH1A2	633,9		X	
Macaca mulatta	ALDH1A2	591,5		X	
Homo sapiens	ALDH1A2	590,0		X	
Danio rerio	ALDH1A3	622,2		X	
Xenopus laevis	ALDH1A3	618,9		X	
Gallus gallus	ALDH1A3	575,5		X	
Monodelphis domestica	ALDH1A3	577,4		X	
Mus musculus	ALDH1A3	658,4		X	
Rattus norvegicus	ALDH1A3	610,8		X	
Canis familiaris	ALDH1A3	546,6		X	
Macaca mulatta	ALDH1A3	586,2		X	
Pan troglodytes	ALDH1A3	572,0		X	
Homo sapiens	ALDH1A3	520,1		X	
Mus musculus	ALDH1A7	562,8		X	
Rattus norvegicus	ALDH1A4	600,7		X	
Branchiostoma floridae	ALDH1	538,1	X		
Branchiostoma floridae	ALDH1	375,5	X		
Branchiostoma floridae	ALDH1	520,0	X		
Branchiostoma floridae	ALDH1	571,9		X	
Branchiostoma floridae	ALDH1	585,6	X		
Branchiostoma floridae	ALDH1	425,9		X	
Saccoglossus kowalevskii	ALDH1	425,9		X	
Saccoglossus kowalevskii	ALDH1	567,0		X	
Saccoglossus kowalevskii	ALDH1	673,0	X		
Saccoglossus kowalevskii	ALDH1	-			X

Saccoglossus kowalevskii	ALDH1	-			X
	Média	588,5			
	SEM	7,8			
	Desvio padrão	58,5			
Organismo	ALDH2s	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
		-			X
Strongylocentrotus purpuratus	ALDH2	356,3	X		
Strongylocentrotus purpuratus	ALDH2	487,0	X		
Branchiostoma floridae	ALDH2	341,9	X		
Drosophila melanogaster	ALDH1A10	428,0	X		
Ciona savignyi	ALDH2	321,5	X		
Ciona intestinalis	ALDH2	331,4	X		
Nematostella vectensis	ALDH2	333,8	X		
Gasterosteus aculeatus	ALDH2	397,9	X		
Danio rerio	ALDH2	429,4	X		
Tetraodon nigroviridis	ALDH2	422,3	X		
Xenopus laevis	ALDH2	394,7	X		
Xenopus tropicalis	ALDH2	401,3	X		
Ornithorhynchus anatinus	ALDH2	316,7	X		
Monodelphis domestica	ALDH2	381,4	X		
Gallus gallus	ALDH2	373,0	X		
Rattus norvegicus	ALDH2	404,0	X		
Mus musculus	ALDH2	344,0	X		
Mesocricetus auratus	ALDH2	377,9	X		
Canis familiaris	ALDH2	338,7	X		
Bos taurus	ALDH2	371,0	X		
Pongo abelii	ALDH2	348,9	X		
Pan troglodytes	ALDH2	320,6	X		
Macaca mulatta	ALDH2	380,2	X		
Homo sapiens	ALDH2	377,2	X		
Apis mellifera		445,7	X		
Tribolium castaneum		465,1	X		
Bombyx mori		472,1	X		
Drosophila pseudoobscura		460,7	X		
Anopheles gambiae		395,5	X		
Drosophila melanogaster	ALDH1A10	428,0	X		
Aedes aegypti		323,1	X		
Xenopus tropicalis	ALDH1B1	511,4	X		
Ornithorhynchus anatinus	ALDH1B1	411,0	X		
Rattus norvegicus	ALDH1B1	419,0	X		
Mus musculus	ALDH1B1	460,4	X		
Canis familiaris	ALDH1B1	475,0	X		
Bos taurus	ALDH1B1	423,8	X		
Pongo abelii	ALDH1B1	430,8	X		
Pan troglodytes	ALDH1B1	449,7	X		
Macaca mulatta	ALDH1B1	483,4	X		
Homo sapiens	ALDH1B1	453,1	X		
	Média	403,1			
	SEM	8,2			
	Desvio padrão	53,1			
Organismo	PLANTAS ALDHs GROUPO A -f- -g-	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Oryza sativa		269,4	X		
Oryza sativa		368,5	X		
Oryza sativa		303,7	X		
Zea mays		351,0	X		
Zea mays		309,0	X		
Arabidopsis thaliana		352,5	X		
Arabidopsis thaliana		326,0	X		
Populus trichocarpa		359,8	X		
Populus trichocarpa		447,0	X		
Populus trichocarpa		421,3	X		
Sorghum bicolor		351,0	X		
Sorghum bicolor		341,0	X		
Nicotiana tabacum		320,0	X		
Secale cereale		308,0	X		
	Média	345,0			
	SEM	12,4			
	Desvio padrão	46,6			
Organismo	PLANTAS ALDHs GROUP B	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Oryza sativa		667,0		X	
Zea mays		562,3	X		
Zea mays		484,7	X		
Arabidopsis thaliana		529,4	X		
Populus trichocarpa		-			X
	Média	561,0			
	SEM	38,8			
	Desvio padrão	77,6			
Organismo	FUNGOS ALDHs	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Phycomyces blakesleeanus		431,6	X		
Phycomyces blakesleeanus		439,9	X		
Phycomyces blakesleeanus		601,5		X	
Phanerochaete chrysosporium		378,9	X		
Phanerochaete chrysosporium		469,0	X		
Phanerochaete chrysosporium		531,5	X		
Phanerochaete chrysosporium		524,2	X		
Phanerochaete chrysosporium		503,0		X	
Phanerochaete chrysosporium		521,6		X	
Schizosaccharomyces pombe	ALDH1R1	555,5		X	
Saccharomyces cerevisiae	ALDH1G2	571,1		X	
Schizosaccharomyces pombe	ALDH1S1	547,8		X	
Saccharomyces cerevisiae	ALDH1F2	432,8	X		
Saccharomyces cerevisiae	ALDH1D1	487,3		X	

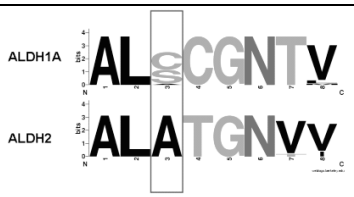
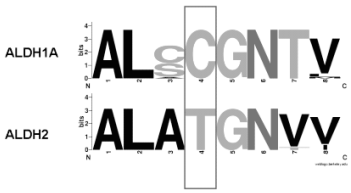
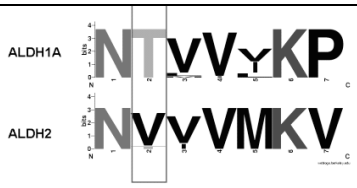
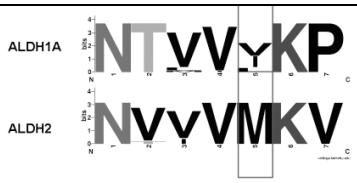
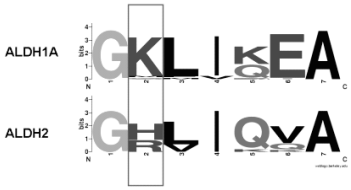
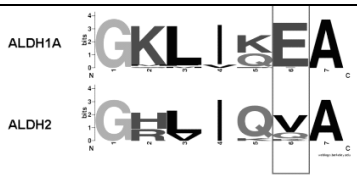
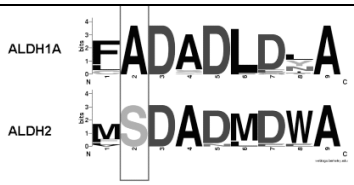
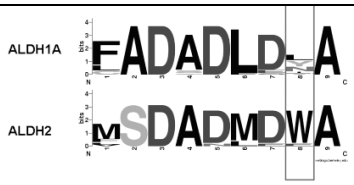
Saccharomyces cerevisiae	ALDH1D2	467,3	X		
Emericella nidulans	ALDH57	417,4	X		
Emericella nidulans		370,2	X		
Beauveria bassiana		437,2	X		
Magnaporthe grisea		391,3	X		
Chaetomium globosum		450,3	X		
Phaeosphaeria nodorum		420,1	X		
	Média	473,8			
	SEM	14,3			
	Desvio padrão	65,7			
Organismo	VERTEBRADOS ALDH2s -h- -i-	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Xenopus tropicalis	ALDH2	401,3	X		
Ornithorhynchus anatinus	ALDH2	316,7	X		
Monodelphis domestica	ALDH2	381,4	X		
Rattus norvegicus	ALDH2	404,0	X		
Mus musculus	ALDH2	344,0	X		
Canis familiaris	ALDH2	338,7	X		
Bos taurus	ALDH2	370,8	X		
Pongo abelii	ALDH2	348,9	X		
Pongo abelii	ALDH2	320,6	X		
Macaca mulatta	ALDH2	380,2	X		
Homo sapiens	ALDH2	377,2	X		
	Média	362,2			
	SEM	9,1			
	Desvio padrão	30,2			
Organismo	ALDH1B1s -j-	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Xenopus tropicalis	ALDH1B1	511,4	X		
Ornithorhynchus anatinus	ALDH1B1	411,0	X		
Monodelphis domestica	ALDH1B1	444,2	X		
Rattus norvegicus	ALDH1B1	419,0	X		
Mus musculus	ALDH1B1	460,4	X		
Canis familiaris	ALDH1B1	475,0	X		
Bos taurus	ALDH1B1	423,8	X		
Pongo abelii	ALDH1B1	430,8	X		
Pan troglodytes	ALDH1B1	449,7	X		
Macaca mulatta	ALDH1B1	483,5	X		
Homo sapiens	ALDH1B1	453,1	X		
	Média	451,1			
	SEM	9,1			
	Desvio padrão	30,3			
Organismo	ALDH1A1as	Volume (Å3)	1OF7	1BXS	Dados insuficiente para modelagem
Xenopus tropicalis	ALDH1A1	690,0		X	
Ornithorhynchus anatinus	ALDH1A1	615,3		X	
Monodelphis domestica	ALDH1A1	607,9		X	
Rattus norvegicus	ALDH1A1	593,8		X	
Mus musculus	ALDH1A1	551,0		X	
Canis familiaris	ALDH1A1	561,5		X	
Bos taurus	ALDH1A1	614,7		X	
Pongo abelii	ALDH1A1	631,3		X	
Pan troglodytes	ALDH1A1	592,6		X	
Macaca mulatta	ALDH1A1	604,0		X	
Homo sapiens	ALDH1A1	639,5		X	
	Média	609,2			
	SEM	11,4			
	Desvio padrão	37,7			

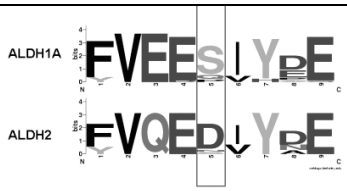
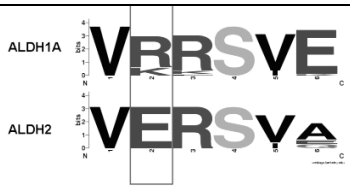
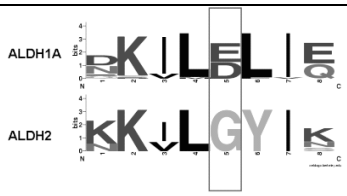
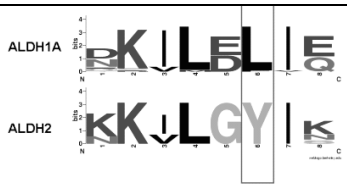
-a- x PLANT ALDHs GROUP A	p<0,01
-b- x ALDH1s	p<0,01
-c- x ALDH2s	p<0,01
-d- x PLANT ALDHs GROUP A	p<0,01
-e- x FUNGAL ALDHs	p<0,01
-f- x PLANT ALDHs GROUP B	p<0,05
-g- x FUNGAL ALDHs	p<0,05
-h- x FUNGAL ALDH1B1s	p<0,05
-i- x ALDH1A1s	p<0,001
-j- x ALDH1A1s	p<0,05

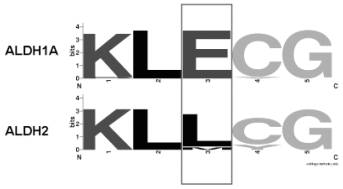
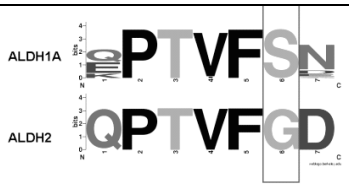
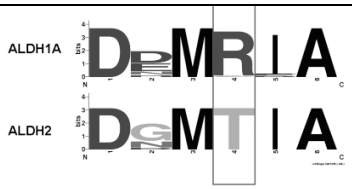
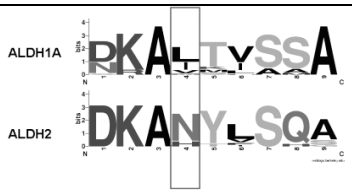
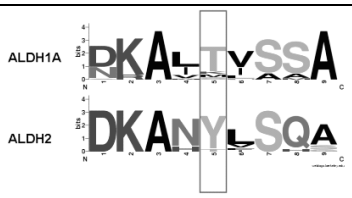
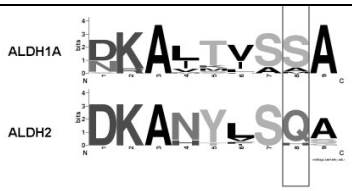
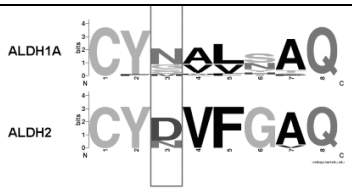
Anexo 3

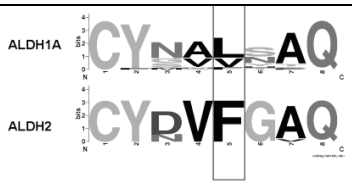
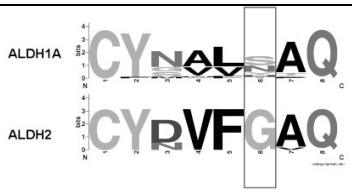
As assinaturas para ALDH1 e ALDH2 de vertebrados. As assinaturas estão mostradas como motivos de quatro a dez aminoácidos. Os motivos são mostrados como pictogramas com a freqüência relativa dos aminoácidos simbolizada pelo tamanho da letra, determinado pelo alinhamento de 24 seqüências de ALDH1 e 12 seqüências de ALDH2. Os aminoácidos não conservados entre as ALDH1 e ALDH2 estão delimitados por um retângulo vermelho.

1	Thr 104-Ala	ALDH1A ALDH2	Polar para alifático, apolar e pequeno	Exposto na superfície (α hélice-B)
2	Gly 110-Asn	ALDH1A ALDH2	Alifático e pequeno para polar	Resíduo enterrado (α hélice-B)
3	Asn 116-Ile	ALDH1A ALDH2	Polar para alifático apolar	Resíduo enterrado (α hélice-C)
4	Gly 123-Asp	ALDH1A ALDH2	Polar para carregado negativamente	Canal (α hélice-C)
5	Gly 124-Met	ALDH1A ALDH2	Alifático, apolar e pequeno para alifático e grande	"boca" do canal (α hélice-C)
6	Thr 128-Cys	ALDH1A ALDH2	Polar para polar thiol-containing	Resíduo enterrado (α hélice-C)
7	Phe 175-Gln	ALDH1A ALDH2	Aromático e grande para polar	Resíduo enterrado (α hélice-D)

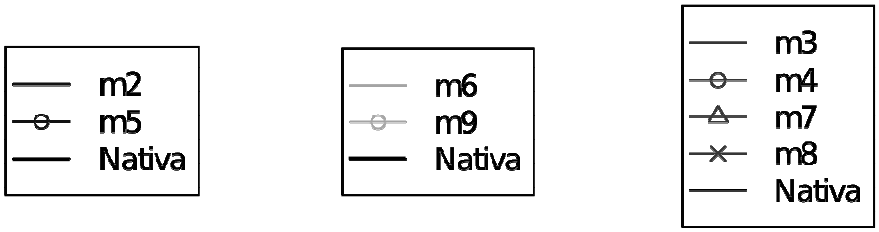
8	Ser 184-Ala		Polar para alifático apolar e pequeno	Resíduo enterrado (α hélice-D)
9	Cys 185-Thr		Thiol-containing para polar	Resíduo enterrado (α hélice-D)
10	Thr 188-Val		Polar para alifático apolar	Resíduo enterrado (folha-β-8)
11	Val 191-Met		Alifático apolar para alifático apolar e grande	Resíduo enterrado (folha-β-8)
12	Pro 193-Val		Apolar cíclico para alifático apolar	Resíduo enterrado (folha-β-8)
13	Lys 251-Arg/His		Carregado positivamente para carregado positivamente imidazolico	Canal (α hélice-G)
14	Glu 255-Val		Negativamente carregado para alifático apolar	Superfície de oligomerização (α hélice-G)
15	Ala 279-Ser		Alifático, apolar e pequeno para polar	Exposto na superfície (Alça de conexão entre a folha-β-12 e α hélice-H)
16	Asn 285-Trp		Polar para aromático grande	Exposto na superfície (α hélice-H)

17	Gln 292-Phe		Polar para aromático	Exposto na superfície (α hélice-H)
18	Gly 293-Ala		Alifático, apolar e pequeno para alifático, apolar e menor	Resíduo enterrado (α hélice-H)
19	Ile 303-Cys		Alifático apolar para thiol-containing	Fundo do Canal (Alça de conexão entre a α hélice-H e folha- β -13)
20	Glu 311-Gln		Negativamente carregado para polar	Exposto na superfície (folha-β-13)
21	Ser 313-Asp		Polar para negativamente carregado	Exposto na superfície (Alça de conexão entre a folha- β -13 e hélice-I)
22	Arg 320-Glu		Positivamente carregado para negativamente carregado	Exposto na superfície (α hélice-I)
23	Asp 355-Gly		Negativamente carregado para alifático, apolar e pequeno	Exposto na superfície (α hélice-J)
24	Leu 356-Tyr		Alifático apolar para aromático grande	Exposto na superfície (α hélice-J)

25	Glu 368-Leu		Negativamente carregado para alifático, apolar e grande	Resíduo enterrado (folha-β-16)
26	Ser 387-Gly		Polar para alifático, apolar e pequeno	Resíduo exposto (folha-β-14)
27	Arg 394-Thr		Positivamente carregado para polar	Resíduo exposto (α hélice-K)
28	Gln 405-Met		Polar para alifático, apolar e grande	Resíduo enterrado (folha-β-16)
29	Ile/Val/Leu 440-Asp		Alifático apolar para negativamente carregado	Superfície de oligomerização (α hélice-M)
30	Thr 441-Tyr		Polar para aromático e grande	Exposto na superfície (α hélice-M)
31	Ser 444-Gln		Polar para polar grande	Exposto na superfície (α hélice-M)
32	Ser 457-Asp		Polar para negativamente carregado	Canal (Alça de conexão entre a folha- β -18 e α hélice-N)

33	Val 459-Phe		Alifático apolar para aromático	“Pescoço” do canal (Alça de conexão entre a folha- β -18 e α hélice-N)
34	Ser 460-Gly		Polar para alifático, apolar e pequeno	Exposto na superfície (Alça de conexão entre a folha- β -18 e α hélice-N)

Anexo 4



Legendas para todas as figuras do anexo 1.

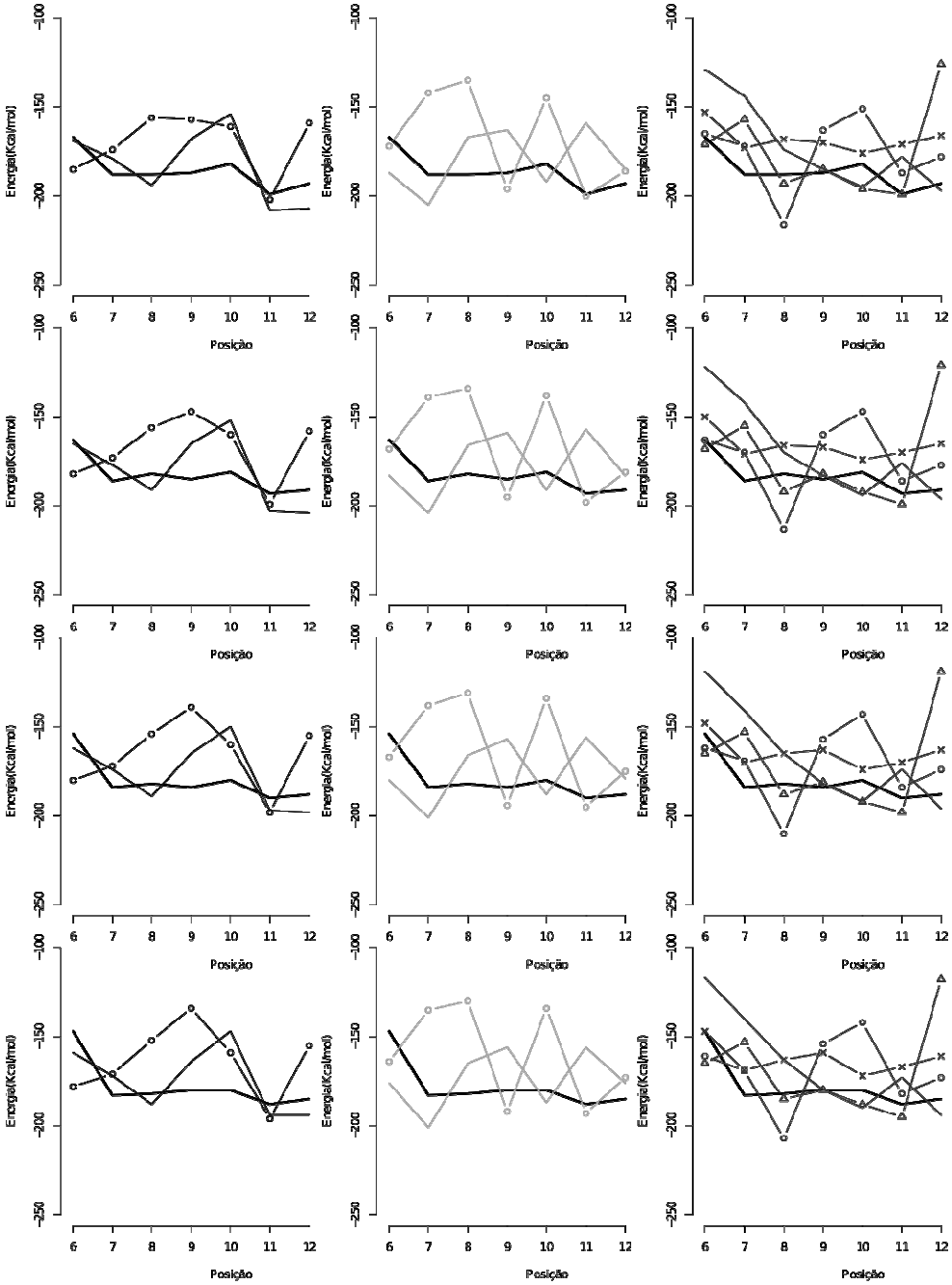


Figura 1: Dados da análise por janela temporal ordenados pela média, tamanho da janela e Zscore. Janelas mínimas de 20 a 50.

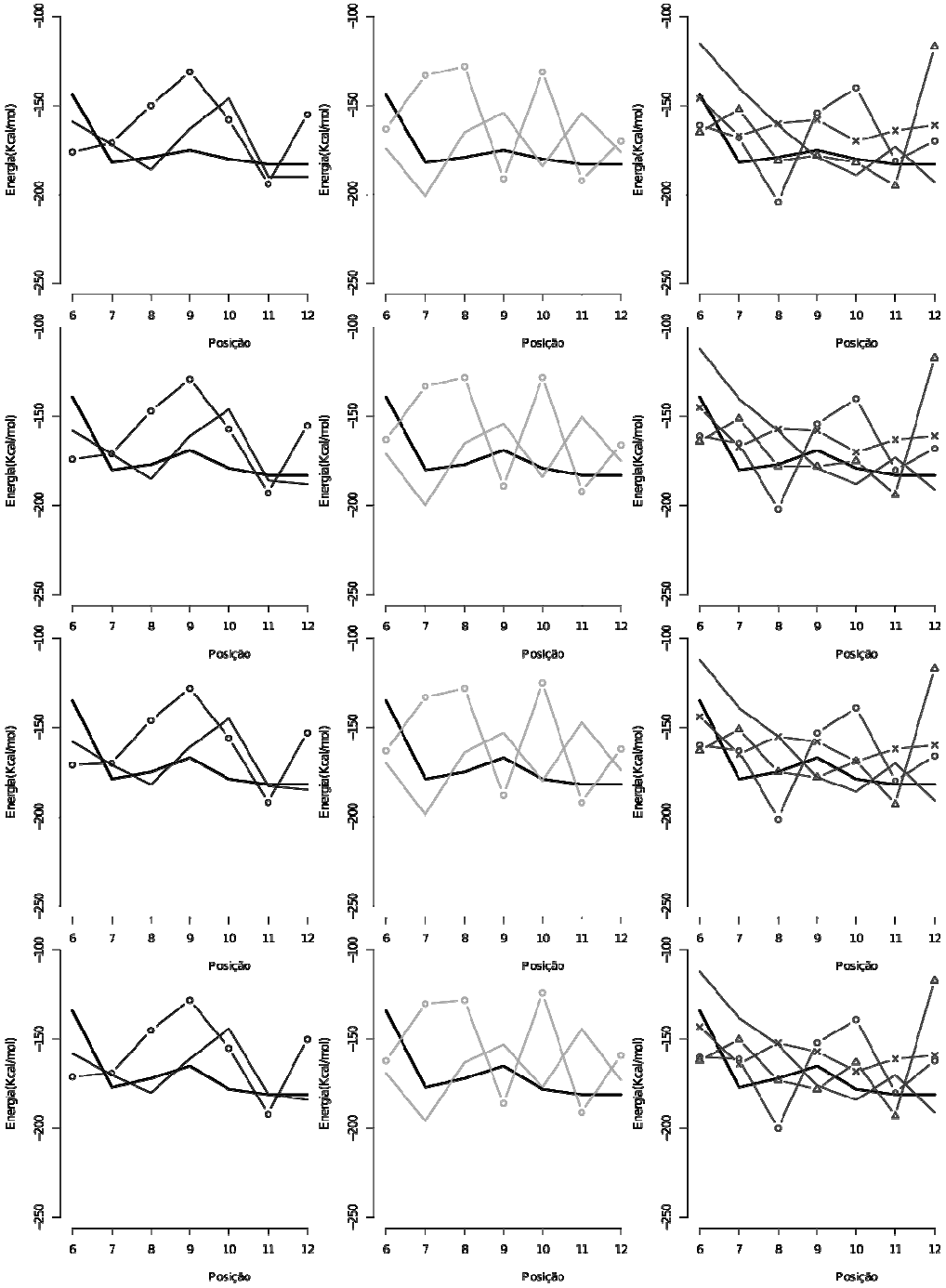


Figura 2: Dados da análise por janela temporal ordenados pela média, tamanho da janela e Zscore. Janelas mínimas de 60 a 90.

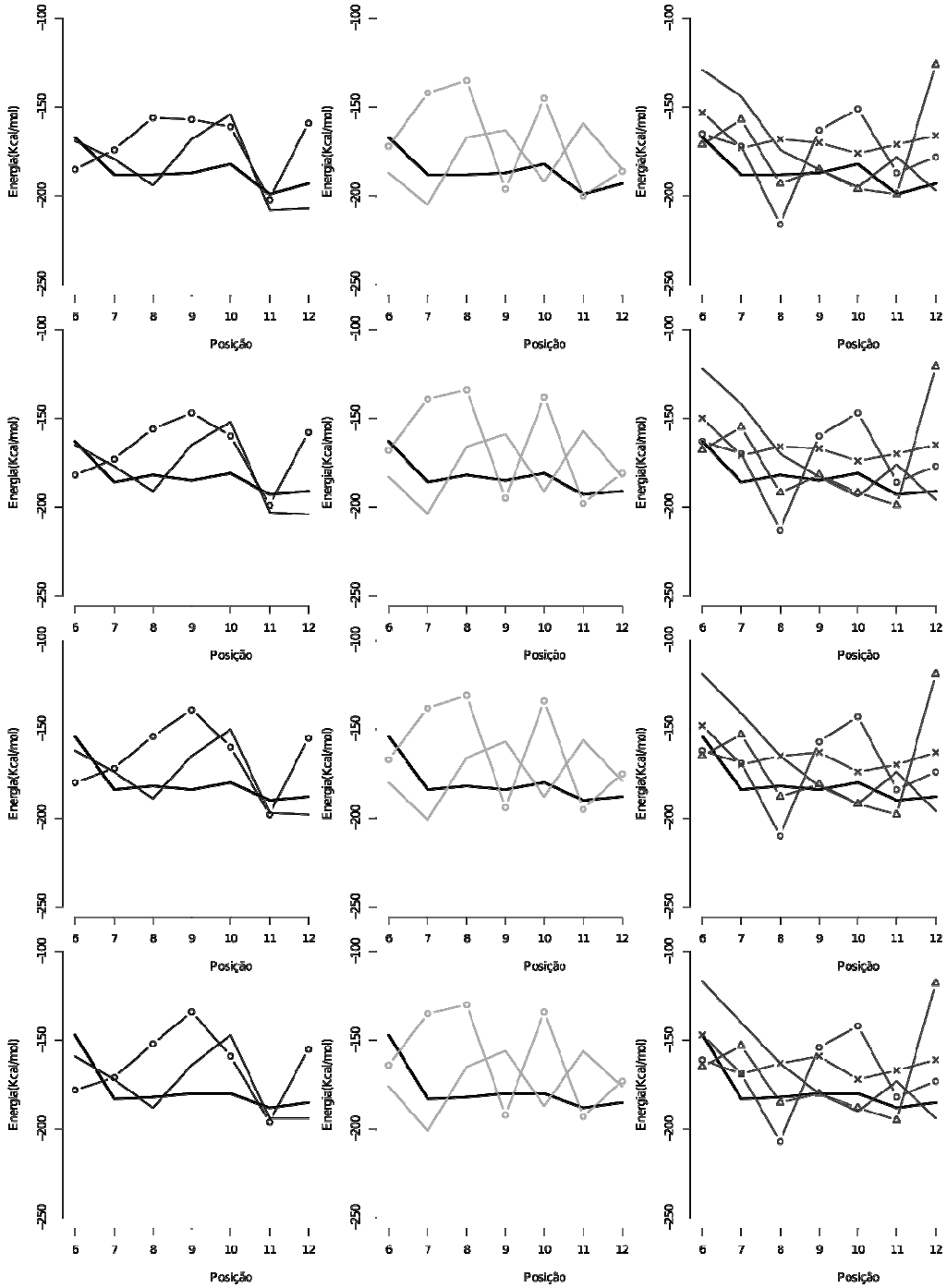


Figura 3: Dados da análise por janela temporal ordenados pela média, Zscore e tamanho da janela. Janelas mínimas de 20 a 50.

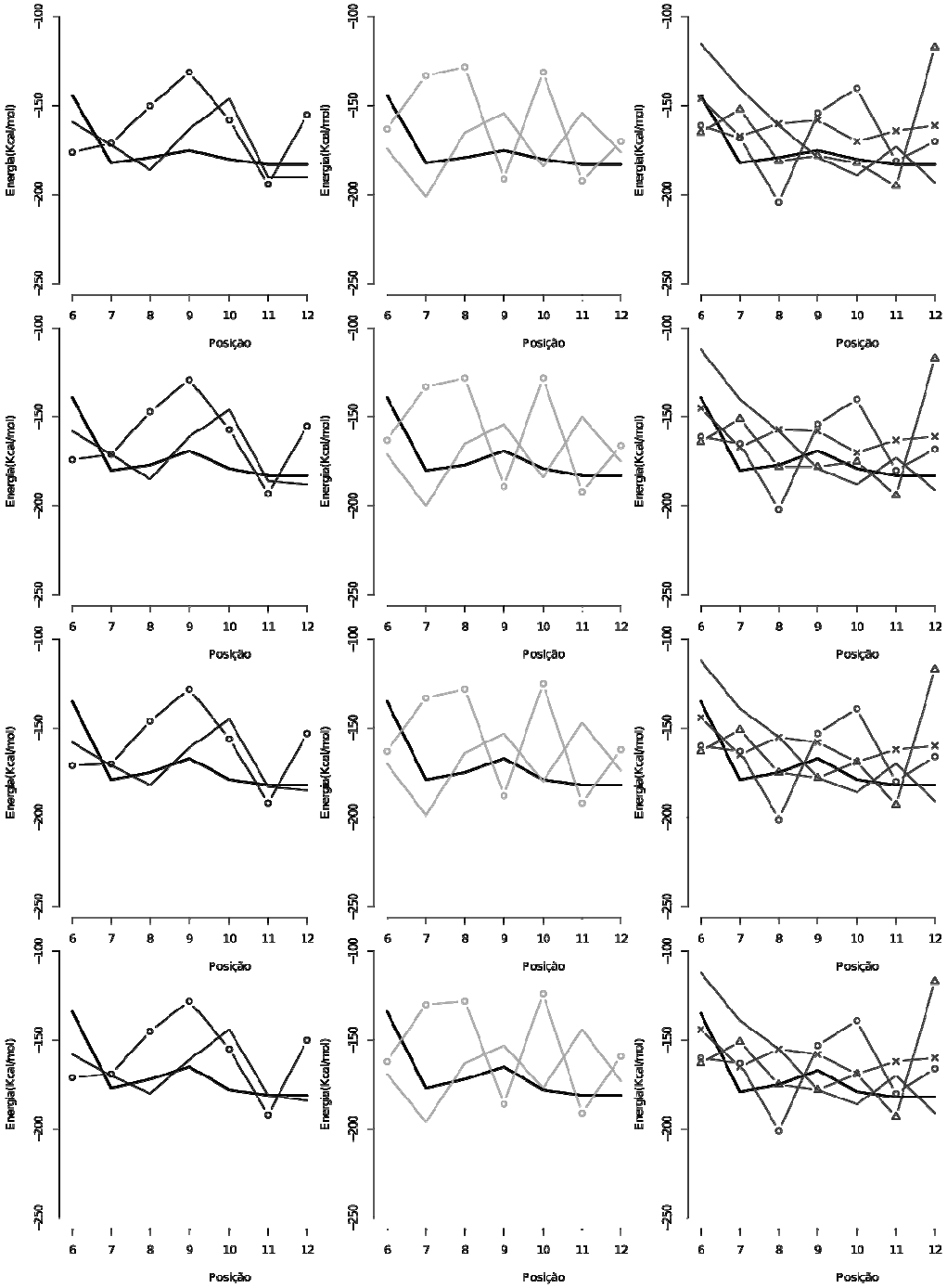


Figura 4: Dados da análise por janela temporal ordenados pela média, Zscore e tamanho da janela. Janelas mínimas de 60 a 90.

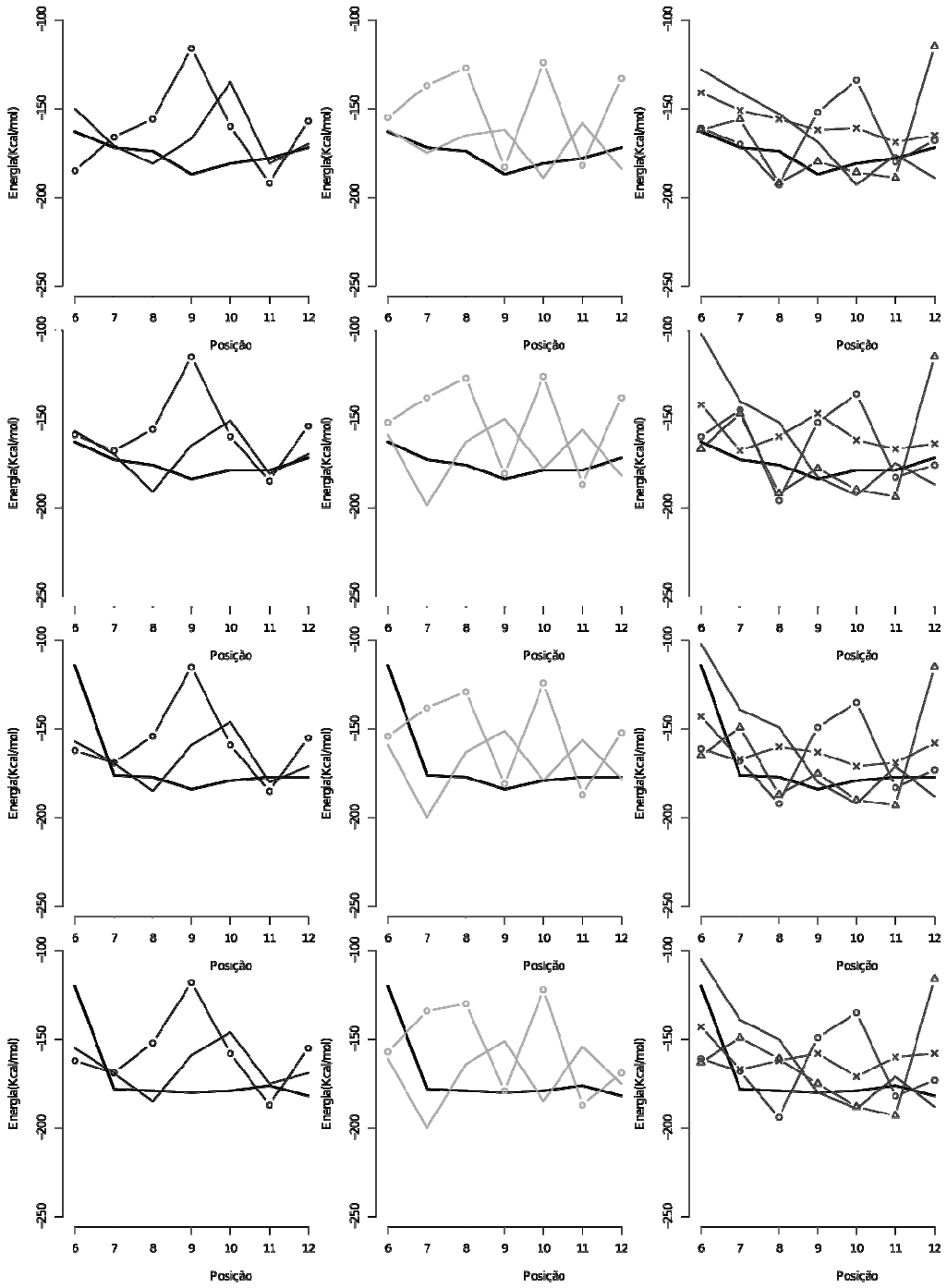


Figura 5: Dados da análise por janela temporal ordenados pela Zscore, tamanho da janela e média. Janelas mínimas de 20 a 50.

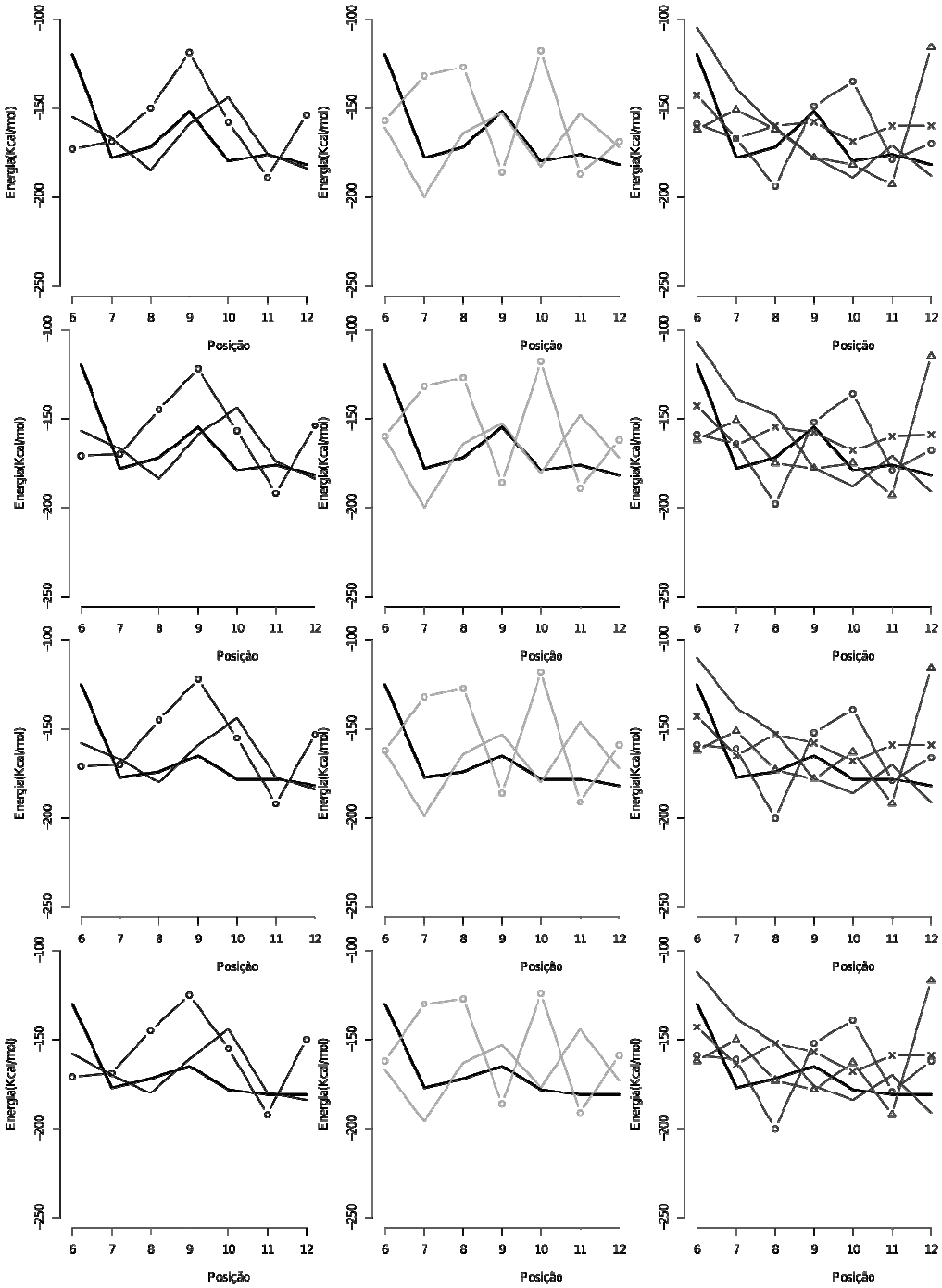


Figura 6: Dados da análise por janela temporal ordenados pela Zscore, tamanho da janela e média. Janelas mínimas de 60 a 90.

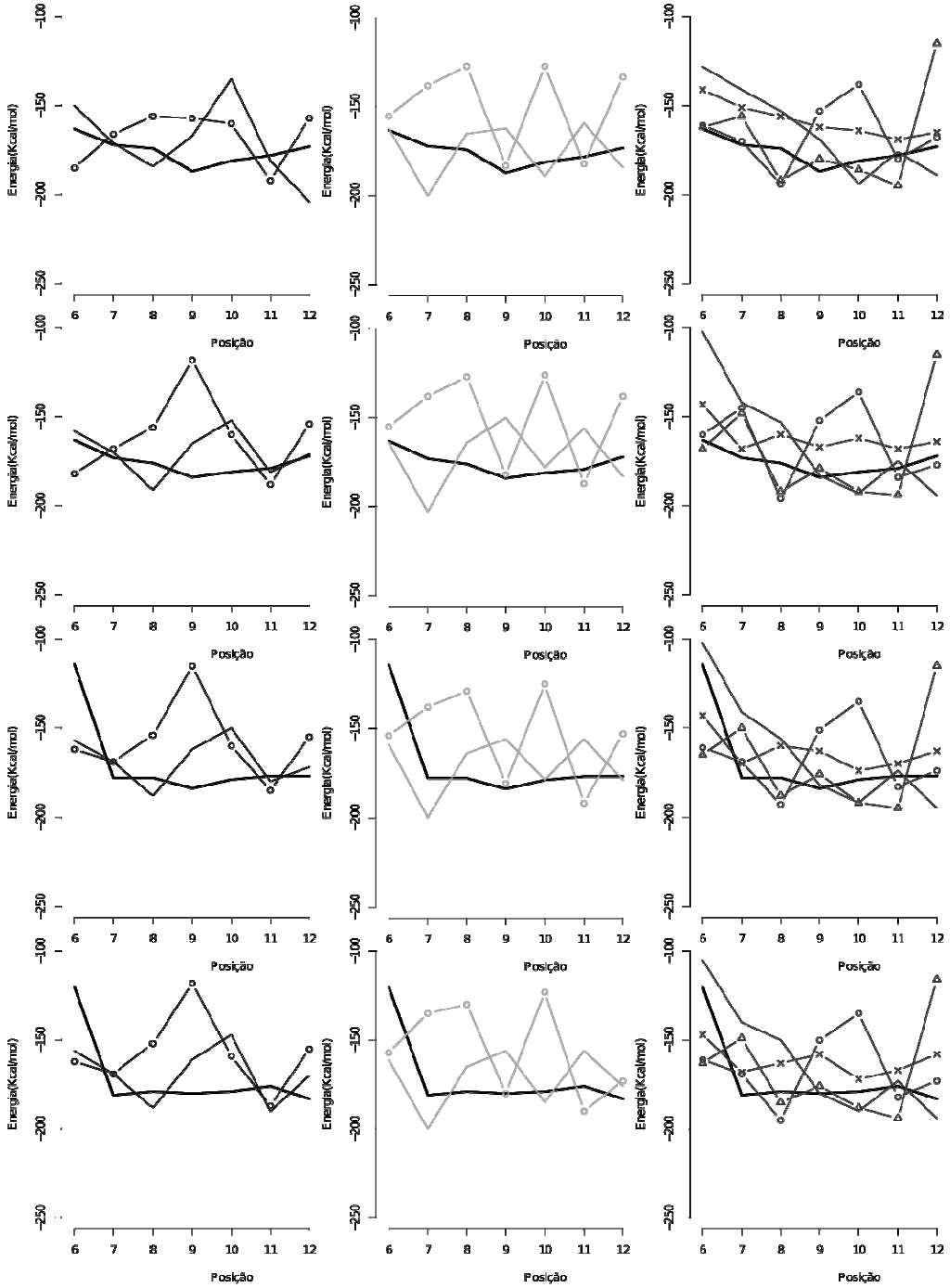


Figura 7: Dados da análise por janela temporal ordenados pela Zscore, média e tamanho da janela. Janelas mínimas de 20 a 50.

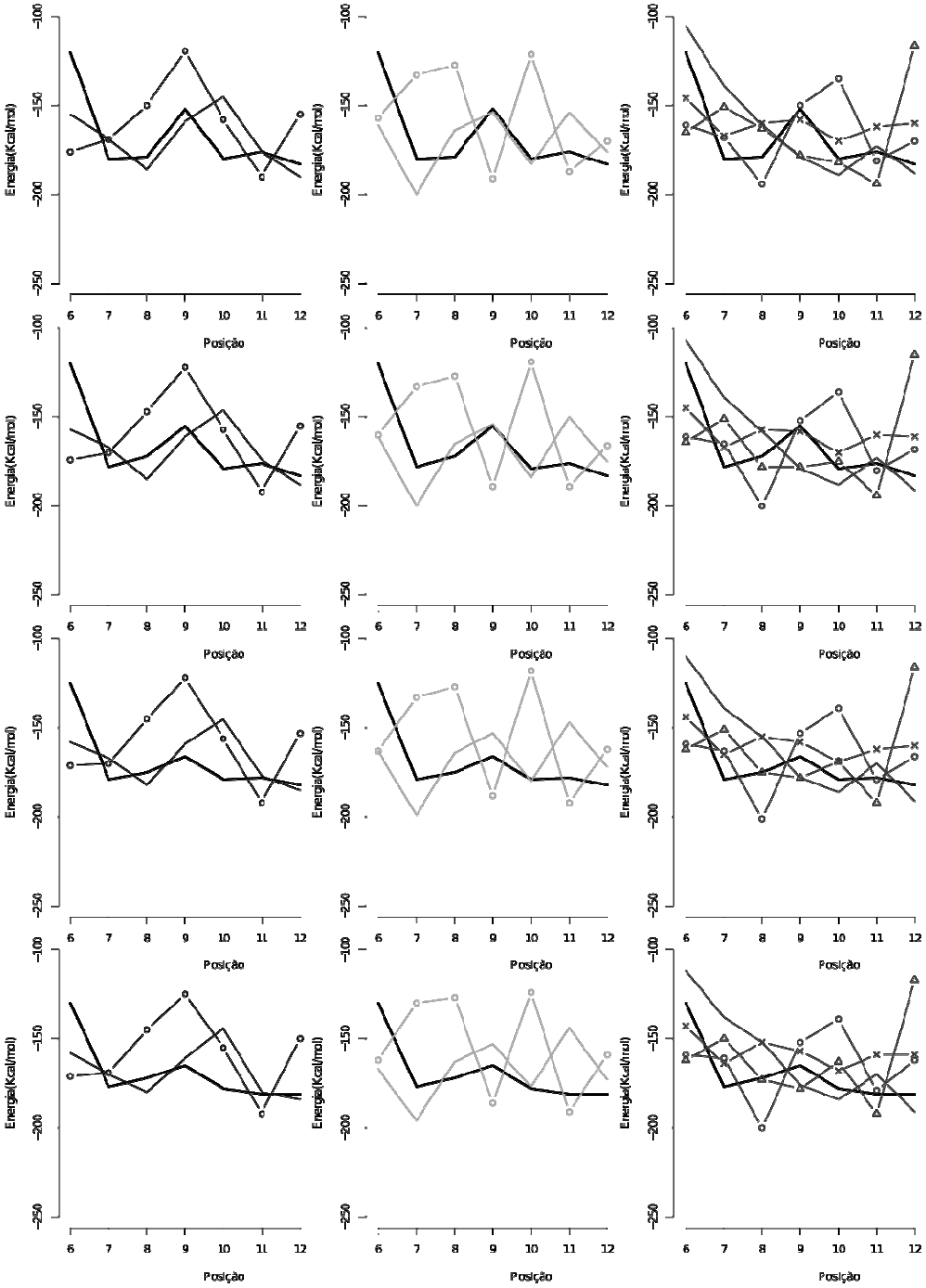


Figura 8: Dados da análise por janela temporal ordenados pela Zscore, média e tamanho da janela. Janelas mínimas de 60 a 90.

Anexo 5

Tabela 1: Energias do DNA nativo coletado nos quatros modos do método de janelas temporais. Nas linhas os diferentes tamanhos mínimos de janela temporal e nas colunas as posições no DNA nativo.

ZMJ	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
20	-160	-143	-145	-163	-172	-174	-187	-181	-178	-173	-151	-149	-188	-146	-160
30	-161	-143	-146	-163	-173	-176	-184	-181	-179	-172	-145	-146	-185	-149	-156
40	-161	-141	-146	-114	-178	-178	-184	-179	-177	-177	-147	-148	-182	-149	-147
50	-159	-139	-146	-120	-181	-179	-180	-179	-176	-183	-145	-146	-180	-150	-145
60	-158	-139	-146	-120	-180	-179	-152	-180	-176	-183	-144	-150	-178	-150	-143
70	-158	-139	-146	-120	-178	-172	-155	-179	-176	-183	-144	-149	-178	-150	-143
80	-158	-139	-147	-125	-179	-175	-166	-179	-178	-182	-144	-148	-179	-149	-143
90	-157	-138	-147	-130	-177	-172	-165	-178	-181	-181	-142	-145	-177	-149	-143

ZJM	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
20	-160	-143	-145	-163	-172	-174	-187	-181	-178	-172	-151	-149	-187	-146	-160
30	-161	-142	-145	-163	-173	-176	-184	-179	-179	-172	-145	-146	-185	-148	-154
40	-160	-139	-144	-114	-176	-177	-184	-179	-177	-177	-146	-148	-181	-149	-146
50	-157	-138	-144	-120	-178	-179	-180	-179	-176	-182	-145	-146	-179	-150	-145
60	-157	-138	-144	-120	-178	-172	-152	-180	-176	-182	-143	-150	-177	-150	-143
70	-157	-138	-144	-120	-178	-172	-155	-179	-176	-182	-143	-147	-177	-149	-143
80	-157	-138	-147	-125	-177	-174	-165	-178	-178	-182	-143	-147	-177	-149	-143
90	-157	-138	-147	-130	-177	-172	-165	-178	-181	-181	-142	-145	-177	-149	-143

MZJ	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
20	-162	-146	-150	-167	-188	-188	-187	-182	-199	-193	-151	-156	-188	-155	-160
30	-161	-143	-148	-163	-186	-182	-185	-181	-193	-191	-148	-153	-185	-152	-156
40	-161	-141	-147	-154	-184	-182	-184	-180	-190	-188	-147	-151	-182	-151	-149
50	-159	-139	-147	-147	-183	-182	-180	-180	-188	-185	-145	-151	-180	-151	-145
60	-158	-139	-147	-144	-182	-179	-175	-180	-183	-183	-144	-150	-179	-150	-144
70	-158	-139	-147	-139	-180	-177	-169	-179	-183	-183	-144	-149	-179	-150	-144
80	-158	-139	-147	-135	-179	-175	-167	-179	-182	-182	-144	-148	-179	-149	-144
90	-157	-138	-147	-134	-177	-172	-165	-178	-181	-181	-142	-145	-177	-149	-144

MJZ	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
20	-162	-146	-150	-167	-188	-188	-187	-182	-199	-193	-151	-156	-188	-155	-160
30	-161	-143	-148	-163	-186	-182	-185	-181	-193	-191	-148	-153	-185	-152	-156
40	-161	-141	-147	-154	-184	-182	-184	-180	-190	-188	-147	-151	-182	-151	-149
50	-159	-139	-147	-147	-183	-182	-180	-180	-188	-185	-145	-151	-180	-151	-145
60	-158	-139	-147	-144	-182	-179	-175	-180	-183	-183	-144	-150	-179	-150	-144
70	-158	-139	-147	-139	-180	-177	-169	-179	-183	-183	-144	-149	-179	-150	-144
80	-158	-139	-147	-135	-179	-175	-167	-179	-182	-182	-144	-148	-179	-149	-144
90	-157	-138	-147	-134	-177	-172	-165	-178	-181	-181	-142	-145	-177	-149	-144