Programa interunidades de Bioinformática
Universidade de São Paulo

Dissertação de doutorado

**“Aplicação de métodos estatísticos e computacionais para o
estudo da cis-regulação da expressão gênica”**

Aluno: Márcio Augusto Afonso de Almeida
Orientador: Paulo Sérgio Lopes de Oliveira

Laboratório de Genética e Cardiologia Molecular – InCor- FMUSP

Dedicatória

Esse trabalho é dedicado a toda a minha família, em especial aos meus pais e irmãs pelo apoio incondicional durante todos esses anos de estudo. A minha esposa Andréia por ter sido uma guerreira nas horas difíceis e grande companheira que tenho para toda a vida. E a Deus por tudo que sou e por ter colocado grandes amigos, colaboradores e familiares que tanto me auxiliaram.

Índice

Resumo.....	9
Summary.....	11
1. Introdução.....	13
2, Estado de arte.....	14
2.1. Regulação pré-transcricional.....	19
2.2. Corregulação e análise de expressão.....	24
2.3. Epigenética	25
2.4. Bases de dados dedicadas ao estudo de fatores de transcrição.....	27
2.5. Técnicas de predição computacional	29
2.6. Limitações dos algoritmos de predição de TFBS	31
2.6.1. Limitações biológicas	31
2.6.2. A limitação combinatorial da busca computacional de TFBS.	32
2.7. Seleção de seqüências candidatas	34
2.8. Algoritmos mistos de predição de sítios de ligação	39
2.9. Referências dessa seção	42
3. Objetivos.....	45
3.1. Objetivos gerais	46
3.2. Objetivos específicos	47
4. Justificativa	49
5. Métodos	51
5.1. Seleção de sítios de alto valor biológico	51
5.1.1. Classificação dos TFBS em famílias	51
5.1.2. Criação de uma ferramenta Perl modular para a busca de TFBS. ..	53
5.1.3. Montagem de uma base local de sintênia	56
5.1.4. Identificação de seqüências conservadas interespecies	57
5.1.5. Identificação de famílias de fatores de transcrição conservadas evolutivamente	60
5.2. Análise de enriquecimento	62
5.2.1. Criação de distribuições nulas para testes de enriquecimento de sítios em uma seqüência de interesse.....	62
5.2.2. Definição do teste de significância estatística aplicado para a identificação do enriquecimento de sítios	63

5.3. Análise de agrupamento de genes a partir das informações preditas em suas regiões regulatórias	65
5.3.1. Identificação de seqüências conservadas flanqueadoras a genes de interesse e predição computacional de sítios de interação nas mesmas. 65	
5.3.2. Normalização das contagens de TFBS em seqüências genômicas de interesse.....	67
5.3.3. Agrupamento de genes a partir das contagens de famílias de TFBS preditas em seus blocos sintênicos	67
5.4. Referências dessa seção	68
6. Resultados.....	70
6.1. Seção 1 - Aplicação dos métodos computacionais desenvolvidos para o estudo de uma possível SIT (Seqüência Indutora de Transcrição) próxima ao gene Aldh1a2 de camundongo	70
6.1.1. Motivação biológica	70
6.1.2. Objetivos dessa seção	73
6.1.3. Justificativa dessa seção	73
6.1.4. Resultados dessa seção	74
6.1.4.1 Identificação de uma possível SIT (Seqüência Indutora de Transcrição) associada à expressão do gene aldh1a2 de camundongo.....	74
6.1.4.2. Predição computacional de sítios de interação de fatores de transcrição na seqüência SIT 1G	76
6.1.4.3. O modelo funcional da SIT 1G	85
6.1.4.4. Testes de enriquecimento	86
6.1.4.5. Resultados - Distribuição de seqüências aleatórias	87
6.1.4.6. Resultados - Distribuição biológica	90
6.1.4.7. - Resultados e Discussão dessa seção	93
6.1.4.8. - Referências utilizadas nessa seção	98
6.2. Agrupamento de genes baseando-se unicamente em sinas transcricionais preditos em suas regiões regulatórias	99
6.2.1. Algoritmos de análise filogenética	99
6.2.2. Motivação biológica - Desenvolvimento neurológico de murinos	102
6.2.3. Objetivo principal dessa seção	105
6.2.3.1. Objetivos específicos dessa seção	105

6.2.4. Justificativa dessa seção	106
6.2.5. Resultados	106
6.2.5.1. Determinação dos genes cujos produtos já foram associados com processos de especialização de neurônios	108
6.2.5.2. Determinação do padrão de expressão desses genes no desenvolvimento neurológico embrionário	111
6.2.5.2.1. Estabelecendo relações entre os genes selecionados com base em seu padrão de expressão tecidual	114
6.5.2.3. Agrupamento de genes baseando-se unicamente em sítios de ligação conservados em suas regiões regulatórias	122
6.5.2.3.1 Determinação de blocos conservados flanqueando esses genes	123
6.5.2.4. Predição computacional de sítios para fatores de transcrição nos blocos conservados flanqueando genes de interesse	128
6.5.2.4.1. Predição de sítios usando a ferramenta MATCH®	128
6.5.2.4.2. Contagem absoluta de sítios preditos nos blocos conservados	129
6.5.2.4.3. Agrupamento de genes com base na matriz de características	130
6.5.2.4.4. Sítios preditos agrupados em famílias e agrupamento usando a matriz de características resultante	132
6.5.2.4.5. Predição de TFBS nas regiões regulatórias desses genes usando o algoritmo busca de TFBS desenvolvido localmente	134
6.5.2.4.6. Determinação de sítios de alto valor biológico e construção de uma matriz de características com as famílias de fatores de transcrição.	135
6.5.2.4.7. Agrupamento da matriz de características construída com a utilização de nosso algoritmo e normalização com base nos tamanho dos blocos sintênicos identificados	135
6.5.2.4.8. Determinação da efetividade da normalização z-score para o posterior agrupamento de genes com base nos sinais transcrpcionais	138
6.5.2.5. Avaliação funcional da árvore solução obtida do agrupamento de genes envolvidos no desenvolvimento neurológico de vertebrados	141
6.5.2.5.1. Análise comparativa dos sítios preditos nos	

elementos organizados na sub-árvore 1	145
6.5.2.5.2. Análise das famílias discriminantes e	
determinantes para os grupamentos das árvores 3 e 4	149
6.5.2.6. Referências dessa seção	152
Discussão	157
Conclusão	162
Anexos	163

Resumo

Resumo

Ferramentas de bioinformática têm se tornado a escolha para auxiliar pesquisadores tanto para a anotação de novos genes, como para estudar genes em condições fisiológicas de interesse. Dentre essas ferramentas destacam-se os algoritmos de agrupamento filogenético e os algoritmos de predição de padrões curtos de DNA, como, por exemplo, predições de sítios para ligação de fatores de transcrição. Desenvolver uma abordagem mista com o objetivo de agrupar genes baseando-se unicamente nos sinais transcrpcionais preditos em suas seqüências é um desafio de difícil transposição. No presente trabalho, apresentamos nossos resultados para tentar superar tal limitação que podem ser subdivididos em duas seções: a primeira aonde desenvolvemos uma abordagem para a melhoria das predições computacionais de sítios de ligação e a segunda, onde passamos a agrupar genes com base nos seus sinais transcrpcionais preditos em seqüências conservadas flangeadoras. A primeira seção de nosso trabalho foi focada no estudo de uma seqüência de indução de transcrição próxima ao gene *Aldh1a2* de camundongo aonde foram preditos sítios para fatores de transcrição que foram posteriormente testados biologicamente e se mostraram associados ao controle da expressão desse gene. A partir de uma profunda pesquisa bibliográfica, nós determinamos um grupo de 57 fatores de transcrição já associados com a especialização de subpopulações de neurônios durante o desenvolvimento neuro-embrionário de vertebrados. Nossa abordagem de seleção de sítios de alto valor biológico foi agora testada em seqüências conservadas próximas a cada um desses genes que codificam esses fatores de transcrição associados e os sítios de ligação para fatores de transcrição foram preditos. Tais sítios foram contabilizados e utilizados com entrada para nossa abordagem de agrupamento. A análise dos resultados do agrupamento determinou que, nossa abordagem se mostrou suficientemente sensível para construir uma árvore solução com boas relações com os padrões, já conhecidos, de expressão para esses genes agrupados. Essa abordagem poderá ser utilizada tanto para anotar funcionalmente genes de interesse quanto para prospectar informações dentro de um grupo de genes previamente selecionado.

Summary

Summary

Bioinformatics tools are becoming the choice to gene annotation and for studying genes in physiological conditions of interest. Among those representative efforts, phylogenetics clustering algorithms and tools for predicting short DNA patterns, such as binding sites for transcription factor, are highlighted. To develop a hybrid procedure that merges these two distant bioinformatics fields of research is challenging. In this study, we present a computational solution to surpass this barrier and, it could be easily subdivided in two distinct sections: Initially, we develop a procedure to improve the computational prediction of binding site for transcription factors and the second one, where genes of interest were grouped based only in their predicted transcriptional patterns in their conserved flanking sequences. The first section of the present study was developed during the study of an enhancer near *Aldh1a2* gene in mouse. The protein coded by this gene is fundamental to the metabolism of retinoic acid, an important mediator to neurogenesis in vertebrate embryos. A set of biologically meaningful binding sites were computationally predicted and latter tested and a group of sites showed strong influence in controlling the expression of target gene. By a comprehensive bibliographic research, we determined a group of 57 transcription factors which were already associated with neuron subpopulations specialization during the development in vertebrates. Our algorithm for selection of high biological value binding sites was applied in conserved flanking sequence in each of these genes and a group of binding sites was determined. These predicted sites were counted and their sum, were used as an input for our clustering routine. The results showed that our clustering procedure were sufficiently sensible to construct a solution tree related to the accumulated knowledge of the expression patterns of these genes. This proposed procedure could be used for functionally annotation of genes of interest and for data mining in the information provided by a group of previously determined genes of interest.

Introdução

1. Introdução

A regulação pré-transcricional da expressão de um gene é, em última análise, intimamente relacionada com a interação entre fatores de transcrição e o complexo de mediação de transcrição. A presença de tais fatores e a sua perfeita localização são dependentes da ligação de seus domínios de ligação a sítios de ligação nas regiões regulatórias dos genes. Tais sítios podem ser interpretados como curtas seqüências de DNA, em geral menores do que 10 pares de base, cuja estrutura tridimensional de seus nucleotídeos é complementar aos domínios de ligação de fatores de transcrição. Diversas abordagens computacionais foram desenvolvidas para a predição de tais sítios de ligação e em conjunto com algoritmos de agrupamento filogenético são largamente utilizadas tanto para a caracterização funcional de seqüências genômicas de interesse como para a mineração de dados de interesse biológico. Unir tais informações tão distantes é um desafio de difícil transposição. Tendo como objetivo tentar transpor tal limitação, o presente projeto me foi proposto pelo inicialmente colaborador e posteriormente orientador Dr. Paulo Sergio Lopes de Oliveira. O projeto nos apresentou uma grande dificuldade em seu início pela ausência de um algoritmo computacional que permitisse a análise serial de um grande número de seqüências genômicas, o que nos levou a desenvolver localmente um algoritmo para tal fim.

Com esse algoritmo em mãos nos deparamos com provavelmente a maior dificuldade em se tratando de predições computacionais de padrões curtos de DNA que se trata do impressionante número de predições falsas positivas com pouca ou nenhuma relevância biológica. Foram testadas diferentes abordagens para a seleção de seqüências candidatas, com máxima probabilidade de apresentarem tais sítios, porém nós optamos por desenvolver e testar uma nova abordagem para a melhoria das predições computacionais. Essa abordagem faz uso da combinação de alinhamentos múltiplos de seqüências conservadas ortólogas com a redução do ruído nas predições computacionais proporcionada pelo agrupamento dos fatores de transcrição em famílias. Esse procedimento foi testado em uma seqüência de camundongo altamente conservada próxima ao gene *Aldh1a2* e a análise dos resultados destacou a validade de nossa proposta. Tais sítios considerados de alto valor biológico predito computacionalmente foram testados com o auxílio de mutações sítio dirigidas e a sua relevância biológica foi comprovada no controle da expressão desse gene. Esses resultados promissores nos permitiram gerar a primeira publicação desse projeto.

Tendo determinado uma boa abordagem para a melhoria das predições computacionais de sítios de ligação de fatores de transcrição, nós passamos a desenvolver uma abordagem que respondesse à uma questão biológica relevante:

- “Existe relação entre as estruturas cis-regulatórias de genes participantes de eventos fisiológicos comuns?”

- “Genes com expressão similar poderiam ser agrupados hierarquicamente segundo suas estruturas cis-regulatórias similares, permitindo estimar quais seriam os módulos regulacionais que controlam um grupo particular de genes em uma condição específica?”

Dessa forma, nossos esforços foram focados em desenvolver uma abordagem, estritamente computacional, para agrupar genes com base nos sítios de ligação relevantes preditos em suas seqüências regulatórias putativas. Nós selecionamos um grupo de fatores de transcrição envolvidos diretamente no desenvolvimento de subpopulações de neurônios na medula espinhal de embriões de vertebrados. Tal modelo foi escolhido por dois motivos: o primeiro motivo é a forte pressão seletiva para a manutenção das funções desses genes entre cordados, e o segundo é o fato de que esse modelo já é estudado funcionalmente pelo grupo de pesquisa liderado pelo Dr. José Xavier, grande colaborador no desenvolvimento desse projeto. Inicialmente determinamos uma árvore solução a partir de informações sobre os padrões de expressão de cada um dos genes que codificam tais fatores de transcrição. Tais informações foram coletadas a partir de uma extensão revisão bibliográfica dentro da literatura científica específica. A nossa abordagem de agrupamento foi testada e avaliada com base nessa árvore de semelhanças, o que permitiu o seu desenvolvimento e aperfeiçoamento. A análise da árvore solução determinada destacou a eficiência de nossa abordagem de seleção em separar genes baseando-se unicamente nos sinais transcrpcionais preditos em suas regiões regulatórias. Genes com padrões de expressão similar foram agrupados de forma sustentada em ramos da árvore solução, separando-os de genes com padrões díspares de expressão localizada.

A organização desta tese

Para desenvolver os métodos propostos nós dividimos esta tese em cinco partes principais:

- Parte 1: Trata do estado de arte dos métodos computacionais utilizados até o momento para o estudo de estruturas cis-regulatórias.
- Parte 2: Os principais objetivos e justificativas do presente projeto
- Parte 3: Apresenta os métodos utilizados para a elaboração do modelo computacional e estatístico utilizado para a predição de sítios de ligação e as estratégias utilizadas para o agrupamento de genes por similaridade de suas estruturas cis-regulacionais.
- Parte 4: Resultados e discussão do modelo de predição de sítios regulacionais para uma sequência indutora de transcrição em medula do gene *aldh1a2* de camundongo, como prova de conceito para o poder do método computacional.
- Parte 5: Resultados e discussão para o modelo de agrupamento de genes participantes do processo de diferenciação de neurônios na medula espinhal embrionária, baseado em suas estruturas cis-regulacionais.

Estado de arte

2.1. Regulação pré-transcricional

Proteínas são constituídas de longas cadeias de aminoácidos que se acomodam e assumem conformações estruturais tridimensionais secundárias, terciárias e quaternárias. Essas macromoléculas se destacam pela sua grande versatilidade em um sistema biológico atuando praticamente em todas as vias de sinalização intra e extracelular, resposta imunológica e regulação gênica. Proteínas envolvidas no tênue controle da expressão dos genes são conhecidas como fatores de transcrição e formam uma das classes de proteínas mais conservadas dentro de um genoma moderno [1]. Genomas de eucariotos superiores, como mamíferos ou aves, compartilham significativa proporção de proteínas de regulação como bactérias ou fungos. Essa impressionante conservação se deve a coevolução das proteínas de regulação e a seqüências de nucleotídeos de um genoma. Essa forte pressão seletiva positiva permite a manutenção do tênue controle da expressão de um gene. Diversas patologias são ocasionadas pela perda desse controle [2]. Doenças de caráter complexo, como câncer, são caracterizadas por grandes mudanças no transcriptoma da célula acometida, provavelmente refletindo mudanças nos perfis de expressão de fatores de transcrição basais que podem alterar significativamente a expressão de um grande número de genes alvos [3].

As cadeias laterais de um fator de transcrição se acomodam e se complementam tridimensionalmente com curtas e específicas seqüências de DNA conhecidas como sítios de ligação ou TFBS (Transcription Factor Binding Site). A interação com seu respectivo sítio de ligação permite ao fator de transcrição se localizar a uma distância ótima que, por sua vez, permite a interação do fator para a regulação do gene em questão [4]. Muito se conhece sobre as principais estruturas protéicas responsáveis pela ligação ao DNA, domínios de ligação clássicos como dedo de zinco (Zinc Finger), Homeodomínio (Homeodomain) e hélice-volta-hélice básico (basic helix-loop-helix) são compartilhados por grande parte dos fatores conhecidos (Figura 1). Cada um desses domínios é caracterizado por domínios tradicionais específicos que se liga a seqüências de DNA similares entre si. Um destaque imediato nas ilustrações da figura 1 é a impressionante complementaridade entre fatores de transcrição e seus respectivos sítios. O domínio Zinc finger se caracteriza pela estabilização de um íon zinco por resíduos de aminoácidos como histidina e cisteína, o que forma um “gancho” na seqüência do

peptídeo e permite sua ligação a seqüências específicas de DNA [5]. O homeodomínio é uma estrutura de ligação ao DNA composta por um domínio hélice-volta-hélice específico de 60 aminoácidos, na qual três alfas-hélice são conectadas por curtas regiões loop. As hélices N-terminais são antiparalelas e a mais longa hélice C-terminal é aproximadamente perpendicular ao eixo formado pelas outras duas hélices [6]. O motivo hélice-volta-hélice básico é composto por duas alfas hélices conectados por uma pequena seqüência de aminoácidos, fatores de transcrição que incluem esse domínio em geral são diméricos, cada unidade possui uma hélice contendo os resíduos de aminoácidos responsáveis por permitir a sua ligação a sítios de DNA. Em geral essas proteínas têm forte afinidade por seqüências conhecidas como E-Box que apresentam uma seqüência consenso (CANNTG)[1] [7].

Figura 1 - Representação "Cartoon" da interação entre fatores de transcrição e seus respectivos sítios. Em azul a sequência de DNA do sítios de ligação em vermelho as proteínas de transcrição. A-) Fator de transcrição com domínio de ligação do tipo "Zinc Finger". B-) Fator de transcrição com domínio de ligação do tipo homeodomínio. C-) Fator de transcrição que apresenta o domínio de ligação do tipo hélice-volta-hélice tradicional.

O domínio de interação de um fator de transcrição possui uma estrutura tridimensional específica e é responsável pela interação com outros fatores de transcrição, com o complexo de moderador de transcrição e diretamente com a própria RNA-Polimerase [8]. A interação combinada dos fatores com o complexo de mediação da transcrição altera significativamente a ligação da RNA-Polimerase com o promotor de iniciação, essa alteração na afinidade da ligação permite o fino

controle do gene controlado por esse complexo de transcrição. Dada essa natureza de ligação, os fatores de transcrição e seus respectivos sítios de ligação são organizados, de modo não aleatório, em unidades de regulação conhecidas como módulos de regulação (Figura 2). Essas estruturas regulacionais estão organizadas de modo a permitir que conjuntos de fatores de transcrição se relacionem entre si sinergisticamente ou antagonisticamente [9]. Análises estatísticas recentes determinaram que aproximadamente 85% dos sítios de interação se localizam em uma janela de no máximo 2kb (2000 pb) a jusante do TSS ou a montante do códon de terminação, incluindo exons e introns [10]. Fatores ligados a sítios consideravelmente distantes podem interagir ativamente com a expressão de um gene principalmente devido ao grande tamanho dessas proteínas e a estrutura parcialmente ou totalmente enovelada da cromatina durante diferentes períodos do ciclo celular. Embora essa grande janela de pares de base seja considerada a região preferencial para a busca de um TFBS, já foram identificados sítios funcionais de ligação a distâncias superiores a um milhão de pares de base em relação ao gene alvo em espécies murinas. No organismo modelo mosca-da-fruta (*Drosophila melanogaster*) já foi identificado um sítio de ligação funcional em um cromossomo diferente ao gene alvo [11]. Essa impressionante plasticidade na regulação da expressão de genes tem sido alvo para diversos estudos recentes na área. A ausência de um fator de transcrição em uma condição fisiológica específica ou uma simples mutação pontual nas bases de seus sítios de posicionamento pode alterar significativamente a expressão de um gene.

Figura 2 - Ligação conjunta de diversos fatores de transcrição e seu papel efetor na regulação de expressão de um gene alvo. Os fatores são representados por figuras geométricas e esses são perfeitamente localizados por seus respectivos sítios de ligação totalmente complementares ao domínio de ligação dessas proteínas. Figura obtida no [sítio](http://www.genwaybio.com/images/gw_static/gw_products/gw_products_transcription_factor_2081_3.jpg). (http://www.genwaybio.com/images/gw_static/gw_products/gw_products_transcription_factor_2081_3.jpg)

Ensaio bioquímico específico, como *DNA footprint* e Chip-chip, são amplamente utilizados para a identificação de sítios de ligação para fatores de transcrição[12]. A técnica *DNA footprint* consiste em efetuar clivagem enzimática, com o auxílio de proteínas de restrição em uma sequência conhecida na presença e ausência do fator de transcrição de interesse. Com o uso de géis de eletroforese, o padrão de clivagem é comparado e trechos de DNA que deixaram de ser clivados na presença do fator de transcrição podem ser considerados sítios putativos para o fator. O fator de transcrição quando devidamente acoplado em sua sequência alvo impede a ação enzimática da proteína de restrição alterando os padrões de bandas observados (Figura 3). Esses experimentos têm sido largamente utilizados tanto para a identificação *de novo* de sítios como a validação dos mesmos preditos computacionalmente ou por experimentos de varredura genômica como Chip-Chip.

Figura 3 - Experimento DNA footprint – A figura 3A traz um fluxograma com os passos utilizados para a construção de um experimento. A figura 3B traz uma representação esquemática dos resultados gerados, por um experimento desse tipo. A coluna Ctrl e c/fator representam respectivamente o controle (sem presença do fator de transcrição) e a coluna teste (com a presença do fator de transcrição). Os fragmentos resultantes de uma reação de restrição são “separados” com o auxílio de um gel de eletroforese. Os padrões de bandas de restrição são comparados e região onde são observadas mudanças nesses padrões (elipses vermelhas) representam possíveis sítios de docking do fator de transcrição.

A técnica Chip-chip ou chip on chip combina dois passos totalmente distintos para a identificação de sítios de ligação para fatores de transcrição a partir de uma varredura genômica. Inicialmente uma célula de interesse é lisada e o DNA extraído é complexado com um conjunto de fatores de transcrição sob investigação. Anti-corpos especializados para cada um dos fatores de transcrição são adicionados ao complexo DNA-Fatores e através de microarranjos genômicos e técnicas de imunoprecipitação são determinados e localizados genomicamente sítios de ligação para fatores de transcrição [13]. Esse é uma técnica que, a exemplo dos microarranjos de análise de expressão, privilegiam a sensibilidade

em detrimento à especificidade. Em geral esses resultados são confirmados através de experimentos DNA-footprint. Essas técnicas de imunoprecipitação aceleram grandemente a velocidade de identificação do conjunto de sítios de ligação para fatores de transcrição e já foram utilizadas com sucesso para o entendimento dos eventos de regulação que genes estão expostos em uma condição fisiológica de interesse [2].

2.2. Corregulação e análise de expressão global

Genes de uma mesma via de sinalização ou participantes de uma cascata gênica são em geral coordenados por grupo de fatores de transcrição intimamente relacionados [14]. Esse tipo de controle coordenado permite que proteínas sejam transcritas de modo organizado para efetuar suas funções biológicas. Essa interessante característica já foi identificada em vias de sinalização como o desenvolvimento cardiológico embrionário, aonde os fatores de transcrição da família NKX e GATA são respectivamente responsáveis diretos por controlar a expressão de diversos genes das cascatas de especialização[15]. Esses fatores de transcrição após terem sido transcritos a partir de seus genes, se ligam a sítios próximos a outros genes de uma cascata e alteram positivamente ou negativamente a expressão dos mesmos. Essa impressionante coordenação permite a melhoria do controle da expressão dos genes, pois permite a utilização de uma quantidade menor de proteínas para a regulação coordenada de um grande número de genes. Em última análise, estudar genes de interesse corregulados pode permitir a identificação desses fatores de transcrição responsáveis pelo evento de corregulação, seja por análise computacional ou experimentação biológica.

Plataformas de análise de expressão global permitem o monitoramento simultâneo de milhares de transcritos nucleares e a identificação de novos genes expressos em uma célula em uma condição fisiológica específica. Essas técnicas fazem uso da quantificação das concentrações de moléculas de RNAm codificados por genes expressos em uma condição sob estudo. Os microarrays (Microarrajós) são constituídos por milhares de sondas de DNA fixadas em uma lâmina de vidro e permite a mensuração simultânea da expressão de milhares de transcritos nucleares a partir da hibridização direta de cDNA marcados. Cada uma das seqüências fixadas identifica univocamente um dos genes de um genoma. Essa plataforma é largamente

utilizada nos diferentes ramos da investigação científica, com destaque a suas aplicações médicas [16]. Doenças de caráter complexo, como câncer, ocasionam grandes alterações no transcriptoma da célula acometida pela neoplasia e o uso de microarrays permitiu identificar alguns dos eventos moleculares responsáveis pela malignização desse tecido [17]. Outras abordagens como SAGE (Serial Analysis of Gene Expression) e ESTs (Expressed Sequence Tags) fazem uso de sequenciamento direto de bibliotecas de transcritos.

As ESTs fazem uso da construção de uma biblioteca de cDNA, cujos transcritos individuais são seqüenciados diretamente. Esse sequenciamento, a exemplo do sequenciamento genômico, torna-se pouco específico em seqüências longas (>200pb) e no caso das ESTs o uso de amplificações com primers randômicos permitem a melhoria na sensibilidade da técnica [18]. Experimentos SAGE guardam alguma semelhança com os experimentos de ESTs, porém fazem uso de um passo intermediário entre a retrotranscrição e o sequenciamento. Nesse passo intermediário, as seqüências de DNA são digeridas por enzimas de restrição específicas que geram pequenas seqüências de 8 a 10 pares de base, chamadas de etiquetas (do inglês tags). Essas pequenas etiquetas são ligadas entre si de forma aleatória formando longas seqüências de mais de mil pares de bases conhecidas como long tags. Essas longas seqüências são seqüenciadas e cada etiqueta de uma long tag é identificada computacionalmente e, é determinada a abundância relativa de cada etiqueta. A partir da abundância relativa de uma etiqueta em relação ao conjunto completo de etiquetas de um tecido, é possível determinar se houve aumento ou diminuição da expressão do gene identificado de modo unívoco por essas etiquetas individuais [19].

A incrível quantidade de dados gerados por esses experimentos de análise de expressão global exigiu o desenvolvimento de bases de dados dedicadas totalmente a organizar e facilitar a consulta nesses dados, dentre essas se destacam as bases NCBI GEO (Gene Expression Omnibus) e a SMD (Stanford Microarray Database). A base NCBI GEO de domínio público oferecida pelo NCBI cataloga informações de aproximadamente 10000 experimentos de expressão global realizados com aproximadamente 300000 amostras em plataformas de análise [20]. O banco SMD oferecido publicamente pela Stanford University organiza somente dados gerados em plataformas microarray [21]. Todas as informações organizadas e, ambas as bases estão organizadas segundo os padrões MIAME (Minimal Information About

Microarray Experiment) de exportação de dados de microarray. A iniciativa MIAME definiu um consenso sobre a forma de exportação e anotação dos dados gerados experimentalmente [22]. Analisar de modo compreensivo essa imensa quantidade de dados gerada por tais experimentos exigiu o desenvolvimento de novas abordagens computacionais e modelagens estatísticas especializadas.

O banco de dados público CoexpDB (Coexpression Data Base) tem como principal objetivo oferecer os resultados de uma ampla análise estatística aplicada aos dados organizados na base NCBI GEO [23]. Essa análise estatística fez uso de uma nova métrica estatística desenvolvida pelos autores para determinar grupos de genes expostos a eventos de correção ou cuja expressão é correlacionada. A métrica proposta pode ser interpretada como uma adaptação da correlação de Pearson e permite determinar o grau de correção entre dois genes quaisquer de um genoma. Cada gene é comparado contra o conjunto complementar de genes e os resultados da análise par a par são ordenados de modo a prover uma lista ordenada de genes mais corrigidos aos menos em relação ao gene base. A análise foi aplicada no conjunto completo de dados de expressão gerados em espécies como homens, ratos e camundongos. Em uma última atualização o banco NCBI GEO passou a oferecer nativamente uma análise de correção entre dois genes baseando-se em uma análise de ranking simplista quando comparada a oferecida pela base CoexpDB. Nela genes são ordenados segundo as suas expressões absolutas e rankeados em cada um dos experimentos disponíveis na base. Genes que apresentam forte semelhança entre seus rankings são considerados bons alvo para eventos de correção [20]. Em ambas as abordagens oferecidas genes corrigidos são pares de genes cujas expressões estão intimamente relacionadas e conseqüentemente a variação de expressão em um deles infere fortemente a variação no outro par.

2.3. Epigenética

As longas seqüências de um cromossomo apresentam diferentes níveis de enovelamento durante o ciclo de vida celular, esse controle é realizado basicamente por pequenas estruturas conhecidas como nucleossomos. Nucleossomos são compostos por dois quartetos idênticos de histonas (H2A, H2B, H3 e H4) estruturadas em um octâmero compacto. Através de estáveis ligações covalentes

,147 pares de base DNA se enrolam em aproximadamente 1,67 voltas em torno desse complexo protéico. Os nucleossomos são as estruturas mínimas utilizadas por genomas de eucariotes para enovelar seus DNA. Os nucleossomos se organizam entre si em grandes complexos que apresentam grande nível de compactação do DNA enrolado em cada uma das unidades. Esse alto grau de compactação permite que e aproximadamente dois metros de DNA se organizem em um núcleo celular de apenas 10 nm de diâmetro (Figura 4). Genomas de espécies modernas apresentam longos trechos de seqüência altamente enovelados que impede totalmente a entrada de fatores de transcrição e a conseqüente expressão de genes localizados nesses trechos. Proteínas conhecidas como Acetilases e Deacetilases alteram completamente a organização desses nucleossomos através da adição e/ou remoção de radicais acetil ao carboxi-terminal da histona H4, alterando a organização tridimensional das histonas e conseqüentemente a estrutura dos nucleossomos [24]. Essa mudança estrutural é tão significativa que permite que fatores de transcrição disponíveis no soluto nuclear localizem e interajam com seus respectivos sítios de ligação permitindo a expressão de um gene próximo.

Esse tipo de regulação é extremamente importante para a correta regulação da expressão temporal e posicional dos genes de um genoma, porém identificar de modo eficiente esses trechos susceptíveis a esse tipo de regulação é um desafio [24]. O DNA genômico, depois de extraído, sofre um processo de desnaturação seguido por uma reação de fragmentação, seja química ou mecânica, o que retira totalmente as informações sobre a localização e enovelamento de nucleossomos que se organizavam anteriormente nessas seqüências. A falta dessas informações dificulta muito a detecção *in vitro* de sítios de ligação para fatores de transcrição, já que seqüências de DNA que apresentaram forte interação para um fator de transcrição específico podem simplesmente estar total ou parcialmente ocluídas por nucleossomos em uma situação *in vivo*, tornando esse um resultado falso positivo. Algoritmos computacionais recentemente desenvolvidos permitem predizer sítios de ligação das histonas de um nucleossomo em seqüências de DNA fazendo uso de modelos estatísticos. Em geral são utilizados modelos HMM (Hidden Markov Model) treinados em um conjunto de seqüências que apresentaram interação com histonas dos nucleossomos e esse padrão é aplicado em novas seqüências de interesse. Tais algoritmos, embora promissores, se mostram expostos a predições falsas positivas, em outras palavras, sítios de histona são preditos computacionalmente em

seqüências que não os apresentam in vivo [25]. A principal razão dessa dificuldade é o paradoxo combinatorial, o qual algoritmos de predição de padrões curtos de DNA estão expostos quando tais padrões são preditos em seqüências de DNA muito longas. Essa limitação será mais bem discutida no decorrer desse texto.

Figura 4 - A -) Representação esquemática dos octâmeros de histona e o enovelamento do DNA em um nucleossomo. B-) Estrutura de empilhamento de nucleossomos dando destaque ao impressionante poder de compactação dessas estruturas

2.4. Base de dados dedicadas ao estudo de fatores de transcrição

O acúmulo exponencial de conhecimento levou em 1995 a publicação da primeira base de dados totalmente dedicada ao estudo de fatores de transcrição, conhecida como Transfac® (<http://www.biobase-international.com/pages/index.php?id=transfac>) [26]. Essa base de dados cataloga informações sobre fatores de transcrição e seus respectivos sítios já identificados e em sua versão 6.0 traz descrições completas de 3660 fatores de transcrição de todos os principais organismos modelos de fungos a humanos. Essa base se destaca pelo estridente critério de qualidade exigido para adição de novas

informações na base, todos os sítios e fatores são curados manualmente e validações biológicas são exigidas. O constante estudo dessas proteínas determinou que fatores de transcrição sempre se ligam a um conjunto de TFBS com alta semelhança, porém não idênticos apresentando alguma variação entre suas bases. Essa é uma variação esperada de qualquer fenômeno biológico conhecido. Essa interessante característica dos conjuntos de sítios para um fator de transcrição levou a criação das matrizes monoposicionais que tentam a partir de uma modelagem estatística determinar quais são as preferências e restrições de um fator de transcrição ao efetuar a sua ligação (As matrizes serão melhores explicadas no próximo item- “Técnicas de predição computacional”).

A base de dados EPD (Eukaryotic Promoter Database), de domínio publico, organiza seqüências promotoras identificadas em homens, camundongos, ratos e outros animais (<http://www.epd.isb-sib.ch>). A base EPD se destaca por utilizar critérios restritivos para a adição de entradas em sua base, o que permitiu esse banco apresentar uma grande coleção de promotores não redundantes de alto valor biológico [27]. Inicialmente são determinados os sítios de iniciação de transcrição (TSS) dos genes de um genoma, a partir do mapeamento de transcritos identificados em experimentos SAGE ou EST, por exemplo[28], Esse banco é intimamente relacionado com outros bancos de dados públicos focados em organizar dados de expressão e outros serviços bioinformática dedicados à oferecer soluções para busca de TFBS em seqüências de interesse.

A base JASPAR, também de domínio publico, organiza somente as matrizes monoposicionais de detecção para grupos de fatores de transcrição [29]. Os autores dessa base se utilizaram de critérios biológicos para classificar fatores de transcrição disponíveis na base de dados Transfac® em famílias de fatores intimamente relacionados. Após o seu agrupamento, o conjunto completo de sítios para uma família de fatores é alinhado e uma matriz monoposicional é construída de forma muito similar a implementada nas matrizes disponíveis no banco Transfac®. Essas matrizes familiares podem ser interpretadas como o resumo das preferências de ligação não para um fator isolado e sim para um conjunto de fatores relacionados [29]. A principal vantagem do procedimento é diminuir consideravelmente o número de predições falsas positivas simplesmente diminuindo o número de matrizes de busca embora tal procedimento pouco ou nada altere a razão sinal/ruído/ das predições computacionais.

2.5. Técnicas de predição computacional

Abordagens estatísticas foram desenvolvidas para modelar a variabilidade observada no conjunto de sítios com interação com um fator de transcrição específico. Modelos estatísticos, determinados por essas abordagens, são utilizados para a predição computacional de TFBS (Transcription Factor Binding Sites) em seqüências de interesse biológico. Durante o desenvolvimento dessa tese, O conceito de predição computacional se refere exclusivamente a determinação de seqüências de DNA que têm a máxima probabilidade de apresentarem um TFBS válido. As matrizes monoposicionais fazem uso das informações disponibilizadas nas bases de dados Transfac e Jaspar e constroem um modelo HMM (Hidden Markov Model) adaptado [30]. Inicialmente o conjunto de sítios para um fator de transcrição é alinhado com auxílio de uma ferramenta dedicada de alinhamento múltiplo. O alinhamento dos sítios é usado como entrada para uma rotina que determina as probabilidades de emissão de cada base possível já identificada em uma dada posição do alinhamento (As principais características dessa modelo de treinamento são discutidas detalhadamente no item métodos do presente texto) [30]. Em cada um das posições, calcula-se um score de “conservação” que define o quanto essa posição é restritiva a trocas de nucleotídeos. À medida que esses estudos foram sendo desenvolvidos, ficou claro que em geral, sítios alinhados para um mesmo fator apresentavam um quarteto de nucleotídeos consecutivos altamente restritivos. Essa pequena seqüência ficou conhecida como centro ou “core” da matriz e é imprescindível para a montagem do complexo DNA - proteína, trocas pontuais nessas posições inviabilizam totalmente a ligação do fator de transcrição aos sítios alvo. De modo complementar, posições específicas aonde foram identificadas proporções quase iguais de uma base, acrescentam muito pouco a especificidade modelo. Substituições ou trocas de nucleotídeos nessas posições têm pouco efeito na ligação do fator de transcrição a seu sitio específico.

Essa interessante característica inerente a todos os fatores de transcrição é utilizada para a predição computacional de TFBS. Uma seqüência ao ser interrogada pelo modelo apresenta um score que é somatório das probabilidades de emissão de seus nucleotídeos em cada uma das posições do modelo multiplicada pela conservação dessa posição. Após a construção do modelo, é determinada a

seqüência de nucleotídeos que assumiria o máximo valor possível no modelo, essa seqüência é composta pelos nucleotídeos mais comuns em cada uma das posições de uma matriz monoposicional. Essa seqüência é denominada consenso e o seu valor é considerado um “score” base de 1 e todos os resultados dos somatórios das seqüências sob investigação são comparados com esse valor máximo. O mesmo procedimento é realizado com o centro das seqüências e a partir da definição *a priori* de um limite mínimo de corte, sítios são identificados ou não nessas seqüências. O somatório de probabilidades de uma seqüência testada é comparado com o valor máximo que o modelo pode assumir. Em geral utilizam-se valores arbitrários dentro do intervalo de [0,8-0,9], corte maiores levam a uma grande proporção de sítios falsos negativos e abaixo, desse valor mínimo, muitos falsos positivos. A escolha arbitrária de um ponto de corte talvez seja o ponto mais crucial no uso desses modelos e em geral utilizam-se outras abordagens para melhorar o modelo de detecção.

As matrizes monoposicionais são consideradas um modelo HMM adaptado por não modelar as probabilidades de transição entre bases consecutivas, assumindo o valor máximo (1) para todas as transições. Essa simplificação torna o modelo mais sensível perdendo consideravelmente sua especificidade, essa troca em muitos casos é vantajosa embora determine um modelo com grandes chances de predição de padrões falsos positivos. Essa limitação se torna ainda mais latente quando um grande número de seqüências ou seqüências muito longas são usadas como entrada para o algoritmo. Com objetivo de melhorar essa indesejável característica estatística as matrizes biposicionais e triposicionais foram desenvolvidas[31]. Esses modelos fazem uso do mesmo modelo de treinamento, mas inferem tanto os valores de conservação quando as probabilidades de emissão a partir de pares ou trios de nucleotídeos. Essa simples mudança permite indiretamente modelar as probabilidades de transição através da observação da probabilidade conjunta de um par ou um trio de bases consecutivas. Outros modelos fazem uso de uma abordagem markoviana pura treinada nos sítios disponíveis na base Transfac [32]. O uso desses modelos mais complexos altera sensivelmente a proporção sinal/ruído das predições computacionais de TFBS, modelos mais restritivos, como matrizes triposicionais, exigem uso cuidadoso, pois muitos sítios de bom valor biológico são rejeitados[32].

2.6. Limitações dos algoritmos de predição de TFBS

2.6.1. Limitações biológicas

Modelos computacionais preditivos quando aplicados a dados biológicos sofrem com nosso limitado conhecimento biológico atual, pouco ainda se conhece sobre os eventos moleculares responsáveis pelo controle de expressão de genes. O desenvolvimento de novas tecnologias de identificação biológica de sítios de ligação e o seu uso extensivo por parte de grupos de pesquisa tem levado ao contínuo aumento do conhecimento sobre essas proteínas [26]. As bases de dados mais importantes têm sido atualizadas constantemente com as adições de novos fatores de transcrição e novos sítios para fatores já conhecidos. A adição de novos sítios para um fator específico leva a reconstrução da matriz monoposicional e consequentemente pode alterar significativamente a sensibilidade e especificidade do modelo que será aplicado em seqüências de interesse. Uma outra grande limitação dos algoritmos de detecção de TFBS (Transcription Factor Binding Site) é a grande dificuldade de se determinar quais são os fatores de transcrição cujos sítios preditos podem explicar o fenômeno observado[33]. Isso se deve basicamente ao pouco conhecimento que temos dos eventos moleculares mesmo em vias fisiológicas muito conhecidas. Tal dificuldade levou a publicação de algoritmos específicos que busca a partir do conjunto de sítios preditos determinar quais desses seriam ligados por fatores direto ou indiretamente ligados com fatores de transcrição em contexto com o fenótipo fisiológico sob estudo[34]. Faz pouco sentido se ater a fatores de transcrição cuja expressão ainda não foi identificada na célula sob estudo e esses podem ser, senão desprezados, aos menos, observados com menos atenção.

As longas seqüências dos cromossomos ao serem seqüenciadas perdem totalmente as informações sobre o seu status de enovelamento. As informações seqüenciadas as quais estamos acostumados a trabalhar podem ser interpretadas como uma abstração muito simplista da complexa organização tridimensional dessas seqüências em uma condição *in vivo*. A figura 4 traz uma representação de umas dessas superestruturas se organizam e é fácil perceber que grandes trechos de DNA estariam localizados em regiões de difícil acesso para fatores de transcrição. Os

nucleossomos podem e são re-allocados para outras regiões de um genoma para permitir a expressão de um gene, mas pouco pode ser inferido baseando-se apenas em material seqüenciado. Algoritmos de predição de sítios de histonas guardam a promessa de permitir definir a partir de uma seqüência DNA quais são as regiões mais expostas à ligação de nucleossomos. Esses programas guardam fortes semelhanças com as matrizes monoposicionais embora façam em geral uso de modelos HMM completos. Inicialmente o conjunto total de sítios para histonas nucleossômicas é alinhado e um padrão estatístico é determinado na tentativa de modelar as preferências de ligação a partir de probabilidades de emissão e transição markovianas [25]. Esse padrão, novamente a exemplo das matrizes monoposicionais, é aplicado em seqüências de interesse. Atualmente já existem tentativas em tentar unificar as matrizes monoposicionais com as predições de sítios de histona, de modo a dar um peso que favoreça binding sites em regiões não ocupadas ou com pequena possibilidade de ocupação [24].

2.6.2. A limitação combinatorial na busca computacional de TFBS

O problema combinatorial advém das limitações da busca de padrões curtos de DNA (normalmente menores que 15 pb) em longas seqüências genômicas seqüenciadas. O limitado alfabeto do DNA composto de apenas quatro bases nitrogenadas (Adenina, Guanina, Citosina e Timina) implica que exista uma chance não desprezível que longas seqüências de DNA apresentem esses curtos padrões de modo não específico. Existe uma probabilidade combinatorial mínima descrita como uma em um universo amostral do tamanho de quatro (04) elevado ao tamanho do padrão de busca. Supondo um sítio de ligação extremamente restritivo de apenas seis bases (pequeno para padrões biológicos), aonde somente um nucleotídeo é aceito pelo fator em cada um das posições, denotado pela letra grega Ω . Devido ao limitado alfabeto do DNA existe a chance combinatorial de encontrarmos uma seqüência perfeitamente idêntica a Ω a cada 4^6 pares de bases de DNA. Portanto, supondo uma seqüência perfeitamente aleatória, a cada 24576 pares de bases únicas existirá um sexteto de bases de DNA, idêntica a Ω , provavelmente com valor biológico discutível. Em geral regiões genômicas candidatas de mais de 100000

pares de base são usadas como entrada para algoritmos de predição, o que impõe que pelo menos quatro sítios sejam identificados tenham eles valor biológico ou não. É importante salientar que, a suposição de um sítio perfeitamente restritivo é totalmente ilustrativa, em geral um fator de transcrição é específico para no máximo 4 nucleotídeos, o que torna o problema combinatorial ainda mais desafiador.

O problema combinatorial se amplifica geometricamente quando pouco se sabe sobre quais fatores de transcrição estão associados ao fenótipo de interesse. Em geral não se sabe exatamente qual fator interage com o gene em questão, o que obriga a predição de sítios não para apenas um fator específico, mas para grupos de 500 ou mais fatores o que leva a predição de um enorme número de sítios em uma pequena sequência de interesse. Se tivermos evidências *a priori* de quais fatores são potenciais candidatos a participarem na regulação do gene alvo, a predição pode ficar restrita apenas a esse limitado grupo [33]. Esse filtro biológico, embora ideal, muitas vezes não pode ser aplicado, já que muitas vezes pouco se sabe sobre o fenótipo de interesse. O problema combinatorial é especialmente desafiador na predição de sítios de ligação e de splicing, em ambos os casos os sítios são em geral pequenos (<14 pb) e as regiões de busca são grandes. Esse é um problema de difícil solução e tentativas de alterar os limites de detecção dos modelos os expõem a muitos eventos falsos negativos ou falsos positivos.

As matrizes monoposicionais organizam um modelo estatístico que tenta representar as preferências de um fator de transcrição com relação aos sítios que se ligam. Porém, esse modelo torna-se falho, por não permitir que seja testada sua especificidade em um conjunto de sítios, que sabidamente não apresentem interação com o fator sob análise. Em geral os sítios que são identificados por um fator representam uma ínfima porção dos sítios possíveis de se organizarem combinatorialmente e testar sua especificidade no conjunto complementar de sequências traria poucos resultados práticos. Essas limitações impõem a necessidade de desenvolver estratégias *a priori* (seleção de sequências candidatas) ou *a posteriori* (data mining e testes de enriquecimento) para a melhoria na predição de TFBS em sequências de interesse.

2.7. Seleção de seqüências candidatas

Devido às limitações dos algoritmos de predição computacional de TFBS, algumas abordagens foram desenvolvidas para tentar identificar quais seriam as seqüências que, *a priori*, têm maiores probabilidades de apresentarem um elemento funcional [35]. A busca de seqüências conservadas se baseia na premissa que seqüências funcionalmente relevantes são mais conservadas do que seqüências de pouca importância. A teoria neutral da evolução de nucleotídeos determina que a chance de ocorrência de mutações é a mesma em qualquer uma das posições de um genoma e a sua fixação ou eliminação depende exclusivamente do seu efeito posterior [10]. Seqüências fortemente conservadas devem apresentar elementos funcionais justificando, portanto, a sua seleção positiva durante a evolução. Mutações ocorridas nesses trechos conservados podem acarretar forte pressão seletiva negativa ao indivíduo carreador dessa variante. Essa teoria é evidenciada pela grande conservação entre exons de genes homólogos em genomas próximos e principalmente pela ainda mais significativa conservação das proteínas codificadas pelos mesmos. Genes homólogos de espécies distintas muitas vezes guardam enorme semelhança entre si, quando alinhados demonstram grandes trechos com perfeita conservação (Figura 5). Esse conceito, já comprovado, foi estendido de modo a determinar que seqüências conservadas próximas a genes de interesse têm uma maior probabilidade de apresentarem elementos funcionais já que mutações ocorridas nesses trechos foram selecionadas negativamente durante milhões de anos de evolução [36]. Mutações simples e, a primeira vista, inofensivas podem inviabilizar a ligação de um fator de transcrição imprescindível para expressão de um gene com serias alterações fisiológicas. Estudos recentes determinaram, a partir de alinhamentos globais, que uma grande proporção, aproximadamente 5%, das bases do genoma humano é constituída por seqüências ultra conservadas [37]. Quando esse nível de conservação é comparado com a expectativa de que apenas 1,5% das bases sejam exons codificantes, leva à conclusão de que ainda existem 3,5% de bases altamente conservadas ainda sem uma satisfatória anotação funcional e que provavelmente carregam informações importantes. Já foram identificados longos trechos intrônicos, aparentemente sem função, que são compartilhados por praticamente todas as espécies que tiveram seus genomas seqüenciados [36].

Algoritmos de alinhamento dedicados foram desenvolvidos com o objetivo de alinhar genomas inteiros e permitir a identificação de trechos altamente conservados compartilhados por essas espécies [35]. Genomas que apresentavam bons níveis de cobertura de sequenciamento como humano, chimpanzé, rato, camundongo e galinha foram alinhados entre si com o auxílio do algoritmo BlastZ [38]. Esse algoritmo é uma variante do algoritmo de alinhamento Smith-Waterman [39] clássico e faz uso de heurísticas de alinhamento, um tratamento eficiente da memória virtual utilizada e parâmetros de alinhamento adaptados para buscar de modo eficiente tais trechos. Os resultados gerados por esses alinhamentos ao serem analisados compreensivamente denotaram não apenas a impressionante conservação de sequência entre genomas de espécies próximas como homens e chimpanzés, mas também a manutenção da organização de genes e exons em longos trechos genômicos entre espécies de maior distância evolutiva [38]. Esses blocos altamente conservados foram chamados de blocos sintênicos e sua existência é considerada um bom indício que essas sequências dentro desses blocos apresentem elementos funcionais. Homens e camundongos compartilham aproximadamente 80% de seus genomas com grande conservação de sequência e organização de genes [38].

Figura 5 - Alinhamento múltiplo entre seqüências de aminoácidos do gene Gata1 em diferentes espécies. Os picos cinzas são uma representação do nível de conservação de cada uma das posições individuais.

A Universidade da Califórnia – Santa Cruz (<http://genome.ucsc.edu/>) oferece publicamente bases de dados completas com alinhamentos de genomas pareados entre diversas espécies com diferentes níveis de parentesco evolutivo. Genomas de espécies modelos, como o humano, foram alinhados contra dezenas de espécies diferentes desde chipanzés a fungos. Esses resultados têm sido utilizados de modo isolado ou combinados entre si para a construção de filtros filogenéticos para a identificação de seqüências conservadas compartilhadas entre genomas. O uso de tais filtros de seleção quando associados à algoritmos de busca de sítios de interação permitiu a efetiva identificação de sítios de transcrição de alto valor biológico [40]. Em 2007, Davies et al investigando a expressão de genes em tecidos

cartilaginosos utilizou abordagens de busca de seqüências conservadas para a melhoria da predição de TFBS. Sítios para fatores da família SOX, com fortes implicações funcionais com o fenótipo sob investigação, foram identificados e posteriormente validados usando estratégias de mutação sítio dirigida [41].

Paralelamente a busca de seqüências conservadas em bases de sintênia, outras abordagens puramente estatísticas foram desenvolvidas para o mesmo fim. O RegPot (Regulatory Potential) faz uso de dois modelos HMM treinados respectivamente em seqüências sabidamente funcionais e em seqüências repetitivas de baixo valor biológico [42]. Esses modelos foram testados em grupos de seqüências funcionais e em seqüência repetitivas, não utilizadas para o treinamento inicial do modelo, para a determinação da ordem markoviana mais eficiente para a classificação das seqüências teste [42]. Após rigorosos testes foi determinado que modelos Markovianos de ordem seis apresentam as melhores relações entre falsos positivos e verdadeiros positivos. Esse modelo de classificação composto foi aplicado em janelas deslizantes de cem pares de base obtidas diretamente de cada um dos cromossomos do genoma do rato, camundongo e homem. Cada janela recebe um valor referente a razão entre a probabilidade *a posteriori* dessas seqüência apresentarem um elemento funcional (HMM treinado em seqüências funcionais) e a probabilidade obtida com o uso do modelo HMM de seqüências aleatórias não funcionais (Figura 6). Seqüências que apresentam valores altos são consideradas de bom valor e possíveis candidatas a apresentarem sítios. Essa informação foi utilizada de modo combinado a algoritmos de detecção de sítios de ligação melhorando consideravelmente a qualidade das predições dos mesmos. Toda essa grande quantidade de informações é de domínio público e está acessível no seguinte link (Potencial regulacional do genoma humano em sua montagem HG18- <ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/regPotential7X/>).

Figura 6 – Potencial regulacional - Fluxograma que determina os passos utilizados para a determinação do potencial regulacional de um dada janela de 100 pares de bases de um cromossomo.

O banco PhastCons também oferecido publicamente pela UCSC faz uso de algoritmos de PhyloHMM (Phylogenetic Hidden Markov Model) que associado a critérios matemáticos rigorosos permite a determinação de seqüências altamente conservadas[43]. A grande utilidade de tal modelo é determinar todos os parâmetros de busca a partir de técnicas de máxima verossimilhança em cima das próprias seqüências candidatas. Os resultados das análises conduzidas com o auxílio do algoritmo PhastCons destacam a existência de diversas regiões ultra conservadas dentro dos genomas humano, rato e camundongo. Essas regiões ultra conservadas ou HCE (Highly Conserved Elements) muitas vezes são mais conservados do que as seqüências de exons de genes próximos e provavelmente devem organizar imprescindíveis informações sobre sítos de Splicing, sítios de docking para micro-RNA ou o para o correto posicionamento de fatores de transcrição com auxílio de sítios de ligação. Embora efetivo, o algoritmo utilizado tem como principal defeito aplicar um modelo constante de substituição de nucleotídeos para todo um genoma. Isso se distância fortemente da realidade biológica aonde trechos do genoma evoluem usando taxas de substituição de nucleotídeos quase locais. (<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phastCons17way/>).

2.8. Algoritmos mistos para a predição de sítios de ligação

A expressão de um gene é em última análise relacionado à ligação seletiva de fatores de transcrição a seus relativos sítios. Tendo isso em mente, autores focaram seus esforços científicos em tentar identificar fatores de transcrição que possam explicar, ao menos parcialmente, eventos de coregulação observados em experimentos de análise de expressão global. Veerla & Hoglund em 2006, partiram de um grupo de genes fortemente coregulado no desenvolvimento da leucemia mieloide aguda. (AML) e estudaram detalhadamente seqüências próximas [-600, +100] ao TSS desses genes[44]. A predição de sítios foi realizada com o auxílio do programa Match [45] e sítios considerados significantes foram identificados segundo a proposição de Aerts et al [46] e foram comparados usando distância de Jaccard. O uso dessa distância permite analisar a composição de sítios que constituem tais promotores, genes demonstrando uma maior semelhança de suas regiões regulatórias são agrupados em ramos próximos de uma árvore filogenética. A análise permitiu a identificação da presença de sítios para o fator MYC/MAX em

todos os genes agrupados em um dos ramos da árvore solução. Dentre esses genes se destacam *MCM5* e *CDC2* cujo papel em promover a expressão de fatores de transcrição E2F cuja expressão é intimamente relacionada à expressão do fator MYC/MAX. [47].

Outros programas como CONFAC fazem uso das informações geradas em um experimento de microarray e as combinam com informações sobre sítios de ligação preditos para a identificação de sítos cujas proporções anormais podem denotar uma possível atuação no fenótipo sob investigação [48]. A proposta do CONFAC se baseia na comparação entre o número de total de sítios para um fator de transcrição em um grupo de genes de interesse contra o número obtido em um grupo de genes de mesmo tamanho selecionado aleatoriamente de um genoma. Os valores são comparados diretamente com o uso de teste de Mann-Whitney e fatores cujo número de sítios seja significativamente diferente ($P < 0,05$) entre os grupos são identificados. O CONFAC pode ser interpretado como uma tentativa de mineração de dados para a identificação de relações funcionais de valor biológico.

Eventos de coregulação na expressão de um grupo de genes são intimamente relacionados com a ligação de fatores de transcrição á seus respectivos sítios Porém pouco ainda se conhece sobre a evolução e organização desses elementos cis-regulatórios (sítios de ligação) nas regiões regulatórias que coordenam tal coexpressão. Um estudo recente testou a ação de 77 combinações de sítios de ligação pré-estabelecidas na coexpressão de um conjunto de 19 genes da espécie *Ciona intestinalis*[49]. O estudo foi realizado com mutações sítio-dirigidas em cada um dos motivos regulatórios e a expressão de cada um dos genes foi mensurada com a utilização de embriões em desenvolvimento. Diferentes modelos de interação entre os motivos foram testados como o modelo aditivo (cada sítios tem uma ação aditiva aos outros) e multiplicativo (cada sítios presente multiplica a expressão do gene). Os resultados sugerem que exista uma grande interdependência entre sítios para fatores de transcrição e em alguns casos, sítios podem substituir outros para a mesma função. Essa impressionante plasticidade levou ao desenvolvimento de algoritmos de predição computacional de sítios de ligação que buscam, a partir de um grupo de genes coregulados, identificar sítios compartilhados ou com grande super representação. A busca por sítios super representados em regiões regulatórias se baseia na premissa que sítios para fatores fundamentais para a expressão de um gene em questão estão agrupados em pequenos trechos de seqüências

O algoritmo DIRE (Distant Regulatory Elements of coregulated genes) faz uso de dados de expressão, genômica comparativa e conjuntos de TFBS preditos em regiões flangeadoras de genes coregulados de interesse [14]. Outras tentativas tentam estudar eventos de coregulação com base na organização dos sítios preditos em suas regiões regulatórias com o objetivo de definir motivos compartilhados que podem ao menos parcialmente definir [50]. O algoritmo Cluster-Buster tenta identificar super-representações de motivos regulacionais em seqüências regulatórias de interesse usando testes de razão de verossimilhança. Esse teste compara o número de motivos esperados em uma distribuição nula de seqüências com o número de motivos preditos na seqüência sobre investigação apresentando um valor que representa o quanto a seqüência de interesse se distancia da distribuição sem apresentar significância. A proposta de identificar grandes proporções de sítios para fatores de interesse permitiu recentemente identificar uma grande proporção de sítios E-Box próximos ao gene *Math1*. Tais sítios se mostraram imprescindíveis para a expressão desse gene no modelo testado[51].

2.9. Referências dessa seção

1. Osborne, T.F. and P.J. Espenshade, *Evolutionary conservation and adaptation in the mechanism that regulates SREBP action: what a long, strange tRIP it's been*. Genes Dev, 2009. 23(22): p. 2578-91.
2. de Vooght, K.M. and W.W. van Solinge, *Gene promoter analysis in molecular diagnostics: do or don't?* Expert Rev Mol Diagn, 2009. 9(5): p. 403-5.
3. Suzuki, H.I., et al., *Modulation of microRNA processing by p53*. Nature, 2009. 460(7254): p. 529-33.
4. Sandelin, A. and W.W. Wasserman, *Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics*. J Mol Biol, 2004. 338(2): p. 207-15.
5. Sera, T., *Zinc-finger-based artificial transcription factors and their applications*. Adv Drug Deliv Rev, 2009. 61(7-8): p. 513-26.
6. Mann, R.S., K.M. Lelli, and R. Joshi, *Hox specificity unique roles for cofactors and collaborators*. Curr Top Dev Biol, 2009. 88: p. 63-101.
7. Nakazaki, H., et al., *Key basic helix-loop-helix transcription factor genes Hes1 and Ngn2 are regulated by Pax3 during mouse embryonic development*. Dev Biol, 2008. 316(2): p. 510-23.
8. Guo, J., Z.C. Li, and Y.H. Feng, *Expression and activation of the reprogramming transcription factors*. Biochem Biophys Res Commun, 2009. 390(4): p. 1081-6.
9. Link, K.A., F.S. Chou, and J.C. Mulloy, *Core binding factor at the crossroads: determining the fate of the HSC*. J Cell Physiol. 222(1): p. 50-6.
10. Iwama, H. and T. Gojobori, *Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network*. Proc Natl Acad Sci U S A, 2004. 101(49): p. 17156-61.
11. Li, X.Y., et al., *Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm*. PLoS Biol, 2008. 6(2): p. e27.
12. de Vooght, K.M., R. van Wijk, and W.W. van Solinge, *Management of gene promoter mutations in molecular diagnostics*. Clin Chem, 2009. 55(4): p. 698-708.
13. Collas, P., *The state-of-the-art of chromatin immunoprecipitation*. Methods Mol Biol, 2009. 567: p. 1-25.
14. Gotea, V. and I. Ovcharenko, *DiRE: identifying distant regulatory elements of co-expressed genes*. Nucleic Acids Res, 2008. 36(Web Server issue): p. W133-9.
15. Putt, M.E., et al., *Evidence for coregulation of myocardial gene expression by MEF2 and NFAT in human heart failure*. Circ Cardiovasc Genet, 2009. 2(3): p. 212-9.
16. Allison, D.B., et al., *Microarray data analysis: from disarray to consolidation and consensus*. Nat Rev Genet, 2006. 7(1): p. 55-65.
17. Yong, A.S. and J.V. Melo, *The impact of gene profiling in chronic myeloid leukaemia*. Best Pract Res Clin Haematol, 2009. 22(2): p. 181-90.
18. Sakabe, N.J., et al., *ORESTES are enriched in rare exon usage variants affecting the encoded proteins*. C R Biol, 2003. 326(10-11): p. 979-85.
19. Takamura, T., et al., *SAGE application in the study of diabetes*. Curr Pharm Biotechnol, 2008. 9(5): p. 392-9.
20. Barrett, T., et al., *NCBI GEO: archive for high-throughput functional*

- genomic data. *Nucleic Acids Res*, 2009. 37(Database issue): p. D885-90.
21. Demeter, J., et al., *The Stanford Microarray Database: implementation of new analysis tools and open source release of software*. *Nucleic Acids Res*, 2007. 35(Database issue): p. D766-70.
22. Brazma, A., *Minimum Information About a Microarray Experiment (MIAME)--successes, failures, challenges*. *ScientificWorldJournal*, 2009. 9: p. 420-3.
23. Obayashi, T., et al., *COXPRESdb: a database of coexpressed gene networks in mammals*. *Nucleic Acids Res*, 2008. 36(Database issue): p. D77-82.
24. Teif, V.B. and K. Rippe, *Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities*. *Nucleic Acids Res*, 2009. 37(17): p. 5641-55.
25. Segal, E., et al., *A genomic code for nucleosome positioning*. *Nature*, 2006. 442(7104): p. 772-8.
26. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles*. *Nucleic Acids Res*, 2003. 31(1): p. 374-8.
27. Schmid, C.D., et al., *EPD in its twentieth year: towards complete promoter coverage of selected model organisms*. *Nucleic Acids Res*, 2006. 34(Database issue): p. D82-5.
28. Carninci, P., et al., *The transcriptional landscape of the mammalian genome*. *Science*, 2005. 309(5740): p. 1559-63.
29. Bryne, J.C., et al., *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update*. *Nucleic Acids Res*, 2008. 36(Database issue): p. D102-6.
30. Quandt, K., et al., *MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data*. *Nucleic Acids Res*, 1995. 23(23): p. 4878-84.
31. Keich, U. and P.A. Pevzner, *Finding motifs in the twilight zone*. *Bioinformatics*, 2002. 18(10): p. 1374-81.
32. da Fonseca, P.G., K.S. Guimaraes, and M.F. Sagot, *Efficient representation and P-value computation for high-order Markov motifs*. *Bioinformatics*, 2008. 24(16): p. i160-6.
33. Yang, E., et al., *Context specific transcription factor prediction*. *Ann Biomed Eng*, 2007. 35(6): p. 1053-67.
34. Kel, A., et al., *Beyond microarrays: Finding key transcription factors controlling signal transduction pathways*. *BMC Bioinformatics*, 2006. 7 Suppl 2: p. S13.
35. Elnitski, L., et al., *Distinguishing regulatory DNA from neutral sites*. *Genome Res*, 2003. 13(1): p. 64-72.
36. Taylor, J., et al., *ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements*. *Genome Res*, 2006. 16(12): p. 1596-604.
37. Katzman, S., et al., *Human genome ultraconserved elements are ultraselected*. *Science*, 2007. 317(5840): p. 915.
38. Schwartz, S., et al., *Human-mouse alignments with BLASTZ*. *Genome Res*, 2003. 13(1): p. 103-7.
39. Waterman, M.S., *Sequence alignments in the neighborhood of the optimum with general application to dynamic programming*. *Proc Natl Acad Sci U S A*, 1983. 80(10): p. 3123-3124.
40. Castillo, H.A., et al., *Insights into the organization of dorsal spinal cord*

- pathways from an evolutionarily conserved raldh2 intronic enhancer. Development. 137(3): p. 507-518.*
41. Davies, S.R., et al., *Computational identification and functional validation of regulatory motifs in cartilage-expressed genes. Genome Res, 2007. 17(10): p. 1438-47.*
 42. King, D.C., et al., *Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res, 2005. 15(8): p. 1051-60.*
 43. Fang, F. and M. Blanchette, *FootPrinter3: phylogenetic footprinting in partially alignable sequences. Nucleic Acids Res, 2006. 34(Web Server issue): p. W617-20.*
 44. Veerla, S. and M. Hoglund, *Analysis of promoter regions of co-expressed genes identified by microarray analysis. BMC Bioinformatics, 2006. 7: p. 384.*
 45. Kel, A.E., et al., *MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res, 2003. 31(13): p. 3576-9.*
 46. Aerts, S., et al., *Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res, 2003. 31(6): p. 1753-64.*
 47. Ohtani, K., J. DeGregori, and J.R. Nevins, *Regulation of the cyclin E gene by transcription factor E2F1. Proc Natl Acad Sci U S A, 1995. 92(26): p. 12146-50.*
 48. Karanam, S. and C.S. Moreno, *CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets. Nucleic Acids Res, 2004. 32(Web Server issue): p. W475-84.*
 49. Brown, C.D., D.S. Johnson, and A. Sidow, *Functional architecture and evolution of transcriptional elements that drive gene coexpression. Science, 2007. 317(5844): p. 1557-60.*
 50. Werner, T., et al., *Computer modeling of promoter organization as a tool to study transcriptional coregulation. Faseb J, 2003. 17(10): p. 1228-37.*
 51. Krizhanovsky, V., et al., *Math1 target genes are enriched with evolutionarily conserved clustered E-box binding sites. J Mol Neurosci, 2006. 28(2): p. 211-29.*

Objetivos

3.1. Objetivos Gerais

➤ Desenvolver e uma abordagem para a melhoria da predição computacional de sítios de ligação de fatores de transcrição a partir do uso de matrizes monoposicionais presentes na base de dados Transfac® 6.0. Essa abordagem se baseia no alinhamento de sinais preditos para famílias de fatores de transcrição em sequência homologas de diferentes espécies. Essa técnica foi testada em uma sequência indutora de transcrição que se mostrou associada à expressão do gene. Sítios preditos nessa sequência foram contados e testados para a possibilidade de estarem enriquecidos nessa sequência, estabelecendo uma associação entre a abundância de sítios para um dado fator de transcrição e a regulação da expressão do gene Aldh1a2 na placa tectal da medula espinhal embrionárias.

➤ Desenvolver uma abordagem de agrupamento de genes de interesse baseada unicamente em seus sinais transcrpcionais preditos em seqüências conservadas flanqueadoras . As seqüências conservadas foram determinadas com base na combinação de alinhamentos pareados entre espécies com diferentes elos evolutivos. Essa abordagem foi testada no conjunto completo de genes que transcrevem proteínas da família fosfolipase e em um conjunto de genes que transcrevem fatores de transcrição e proteínas de sinalização envolvidas diretamente nos processos de especialização de neurônios em cordados.

3.2 Objetivos específicos

Para cumprir os objetos gerais propostos foram propostos alguns objetivos específicos:

- Desenvolvimento de uma ferramenta Perl modular para a predição computacional de TFBS que permite a fácil alteração de padrões de entrada e saída. Essa ferramenta precisa ser facilmente adicionada em análises computacionais seriais.
- Construção de uma base de dados local que organize os alinhamentos pareados entre genomas oferecidos publicamente pela UCSC (University Califórnia Santa Cruz). Essas informações serão utilizadas para a determinação de eficientes filtros filogenéticos para a seleção de seqüências conservadas.
- Classificar o conjunto de fatores da base de dados Transfac 6.0 em famílias de fatores intimamente relacionados.
- Determinação de famílias de TFBS conservados em uma seqüência regulatória putativa a partir de um alinhamento múltiplo de seqüências homólogas.
- Construção de duas distribuições nulas de seqüência para a aplicação de testes de enriquecimento: distribuição aleatória e distribuição biológica.
- Desenvolvimento e utilização de algoritmos de agrupamento para a organização de genes em árvores de similaridade baseando-se unicamente na predição de TFBS em suas regiões regulatórias conservadas. Essa abordagem permitirá a identificação de sítios e conseqüentemente fatores responsáveis pelos agrupamentos.

Justificativa

Justificativa

Esse trabalho nasceu da oportunidade de criar abordagens computacionais que melhorem a razão sinal/ruído das predições computacionais de TFBS. Essas predições são grandemente prejudicadas pelo nosso limitado conhecimento biológico e pela limitação combinatorial imposta por predições de padrões curtos de DNA em longas seqüências genômicas. Predições de alto valor biológico quando associadas à validações funcionais específicas permitem o entendimento dos eventos transcricionais, aos quais, genes estão submetidos. Estudos funcionais em promotores putativos são prejudicados pela pouca informação disponível *a priori* e pela dificuldade em determinar quais trechos da seqüência são de interesse funcional. Tal dificuldade pode ser superada ou minimizada pela utilização de ferramentas de bioinformática que quando associadas às técnicas de “phylogenetic footprint” e mineração de dados permitem uma sensível melhora na predição. Algoritmos de agrupamento permitem que elementos sejam classificados com base em uma característica definida pelo usuário e nesse projeto, genes foram classificados com base na composição dos elementos transcricionais preditos em suas seqüências regulatórias. As abordagens desenvolvidas foram testadas em modelos biológicos para a avaliação de sua eficiência, mostrando-se úteis na formulação de modelos de funcionamento de seqüências estimuladoras/repressoras de transcrição em sistemas biológicos específicos. Assim, nós esperamos estender esse modelo para outros genes e, posteriormente, disponibilizar os algoritmos desenvolvidos para uso da comunidade científica.

Métodos

5 - Métodos

5.1. – Seleção de sítios de alto valor biológico

5.1.1 - Classificação dos TFBS em famílias

Devido a grande similaridade entre suas matrizes de detecção, sítios de fatores de transcrição funcionalmente semelhantes são preditos computacionalmente de forma redundante em uma mesma sequência de interesse. Essa grande redundância se deve a tênue relação evolutiva entre fatores de transcrição paralogos, os quais são originados a partir de antigas duplicações de fatores de transcrição ancestrais. Essa destacada conservação quando adicionada a já conhecida forte pressão seletiva positiva na região da interação proteína-DNA, leva essas proteínas a apresentarem matrizes de detecção muito semelhantes [1]. Para diminuir a redundância de detecção e permitir a identificação de informações biologicamente relevantes, cada um dos 442 fatores de transcrição de vertebrados da base de dados pública TRANSFAC® 6.0 foi classificado em 51 famílias pré-definidas de fatores de transcrição. A definição dessas famílias e a classificação dos fatores foram realizadas após exaustiva revisão bibliográfica na literatura especializada e consulta nas informações disponibilizadas pela própria base de dados.

A tabela 1 ilustra os resultados da classificação realizada, na mesma existe a descrição de três famílias de fatores de transcrição e seus respectivos elementos constituintes. A família GATA é constituída por um grupo de sete fatores de transcrição paralogos com padrões de ligação de DNA muito similares, todos esses fatores têm como principal característica a ligação com grande afinidade a quartetos de nucleotídeos de sequência “gata”. A família LEUCINE_ZIPER é funcionalmente mais heterogênea do que a família GATA, mas todos os seus elementos possuem domínios de ligação ao DNA muito similares entre si e consequentemente se a ligam a sítios de DNA muito parecidos. Essa grande semelhança, por sua vez, infere que a predição de sítios para um dos fatores levará a sobreposição com sítios preditos para outros fatores também classificados a essa família. A família HEAT_SCHOCK é composta por apenas três fatores que são envolvidos diretamente nas adaptações celulares de eucariotos e procariotos a

mudanças extremas de temperatura. Os três fatores são proteínas paralogas de divergência recente e possuem domínios de ligação ao DNA muito parecidos.

GATA	LEUCINE_ZIPPER	HEAT_SHOCK
GATA-1	E4BP4	HSF
GATA-2	XBP-1	HSF1
GATA-3	HTF	HSF2
GATA-4	Bach2	
GATA-5	Bach1	
GATA-6	NRF2	
GATA-X	VMAF	
	HLF	
	NFE2	

Tabela 1 - Tabela que ilustra três famílias de fatores de transcrição e seus respectivos elementos

A figura 1 destaca a grande semelhança nos padrões de ligação dos fatores de transcrição organizados na família GATA. Essa ilustração traz uma representação gráfica do tipo Weblogo que a partir de um alinhamento múltiplo destaca o grau de conservação de cada uma das posições alinhadas [3]. Todos os sítios já identificados para cada um dos elementos da família GATA foram identificados e alinhados com o auxílio da ferramenta dedicada de alinhamento múltiplo ClustalW [4, 5]. A análise dos padrões de cada um dos elementos destaca a grande semelhança entre a preferência de ligação de seus sítios e em geral sítios a predição computacional para sítios para esses fatores são altamente redundantes. Um destaque imediato dessa ilustração, são as posição de 6 a 9 do alinhamento aonde os sítios para fatores GATA apresentam predileção a ligação a sequência de nucleotídeos com seqüências “gata”.

Figura 1 - Cada uma das colunas representa uma posição consecutiva do conjunto de TFBS de uma proteína de transcrição e o tamanho relativo dos nucleotídeos da uma exata representação de quais são os mais comuns e mais raros em cada uma as posições de um TFBS.

5.1.2 - Criação de uma ferramenta Perl modular para a busca de TFBS

Atualmente existem diversos algoritmos computacionais disponíveis para a predição de TFBS que implementam diferentes estratégias computacionais estatísticas para a detecção de padrões curtos de DNA [5, 6]. Esses algoritmos, embora meritórios, são destinados para análises de seqüências curtas de DNA e não permitem sua fácil integração a protocolos seriais de análise de uma grande quantidade de dados genômicos (pipeline de análise). Para suprir essa carência, nós desenvolvemos uma ferramenta perl modular que implementa o algoritmo MatInspector [7] e permite sua fácil adição em protocolos de análises genômicas (Pipelines). O algoritmo foi desenvolvido de modo a permitir a fácil alteração de parâmetros de busca e de entrada e saída de dados. Nosso algoritmo pode ser subdividido em dois módulos totalmente independentes: aprendizado e de busca. A rotina de aprendizado utiliza-se das matrizes monoposicionais de cada um dos 442 fatores de transcrição de espécies vertebradas disponíveis na base de dados TRANSFAC® 6.0 para a determinação de um modelo estatístico que represente os padrões de seqüências de um conjunto de sítios já identificados para uma dada proteína de transcrição [8]. A rotina de busca aplica os modelos construídos pelo módulo de aprendizado em seqüências de interesse e determina a probabilidade

de uma sequência de DNA apresentar um sítio para o fator de transcrição modelado pela sua matriz de alinhamento de sítios. O código completo e arquivos com instruções de uso podem ser obtidos http://www.ime.usp.br/~maugusto/TESE_DOUTORADO/.

A figura 2 traz um fluxograma que ilustra os passos necessários para a construção do modelo probabilístico para um fator de transcrição que posteriormente será utilizado para a busca computacional de TFBS para o fator em questão. O modelo probabilístico se baseia na determinação da máxima probabilidade de uma sequência de DNA apresentar um sítio para o fator de transcrição de interesse. Para a determinação desse valor máximo, são calculados os escores de conservação em cada uma das posições do alinhamento múltiplo de sítios, esse valor descreve o quanto essa posição é restritiva ou permissiva a troca entre bases nitrogenadas. A sequência de DNA com máxima probabilidade é composta pelas bases mais comuns em cada uma das posições do alinhamento e esse é chamada de matriz consenso. Cada sequência de interesse a ser testada tem o seu somatório de probabilidades determinado e comparado ao valor máximo da matriz consenso. Em geral, se aceita seqüências que apresentem no mínimo 85 ou 90% do escore máximo que uma seqüência pode assumir. O algoritmo implementado na ferramenta Match ® faz uso de um escore adicional determinado pelo quarteto de bases mais conservado dentro do alinhamento múltiplo de sítios que é denominado “core” ou centro da seqüência. Essa abordagem permite identificar um grupo de seqüência com altíssima probabilidade de apresentarem sítios de interação DNA-Proteína funcional. A utilização das matrizes monoposicionais nas busca de TFBS têm se tornado cada vez mais difundida embora tal modelo quando aplicado em longas seqüências de DNA, com pouca informação a priori, leva a identificação de uma grande proporção de TFBS falsos positivos. Uma possível solução seria identificar seqüências que apresentam uma maior chance de apresentarem tais elementos e após essa pré-seleção aplicar ferramentas de busca com as matrizes monoposicionais.

Figura 2 - Fluxograma destacando os passos necessários para a construção do modelo probabilístico e a sua posterior aplicação para a predição computacional de TFBS.

5.1.3 - Montagem de uma base de dados local de sintênia.

Algoritmos de detecção de padrões curtos de DNA em longas seqüências de um genoma são expostos a apresentarem baixa especificidade e uma conseqüente predição de grande número de TFBS com pouco valor biológico. Uma forma muito eficaz de identificar seqüências com grande chance de apresentarem sítios funcionais e determinar trechos de DNA conservados durante a evolução de duas espécies distintas. Algoritmos de alinhamentos dedicados ao alinhamento de genomas completos permitiram a publicação de bases de sintênia que organizam e permitem a identificação de seqüências genômicas conservadas durante a evolução de duas espécies. A UCSC (Universidade da Califórnia Santa Cruz) disponibiliza publicamente resultados do alinhamento par a par de diversos genomas modernos (<http://genome.ucsc.edu/>) gerados com o algoritmo BlastZ [9]. As seqüências foram selecionadas por apresentarem bom nível de similaridade e por um serem específicas ($E.value < 10^{-8}$). Através da utilização de protocolos transferência de arquivos, nós obtivemos as bases de dados com as informações obtidas a partir do alinhamento do genoma do camundongo contra os genomas humano, rato e de galinha. As informações obtidas foram organizadas e, devido à enorme quantidade de dados de cada uma delas, foram inseridas em um banco de dados PostgreSQL (<http://www.postgresql.org/>) de alta performance permitindo um acesso eficiente as informações organizadas nesses bancos.

A tabela 2 traz as medidas estatísticas de resumo dos tamanhos dos trechos conservados organizados nas bases de dados utilizadas e sua análise reflete fielmente, como esperado, a filogenia aceita entre as espécies envolvidas. Quando o genoma do camundongo foi alinhado contra o genoma do rato, diversos longos trechos foram detectados contrastando contra o número significativamente menor de trechos alinhados contra o genoma humano ($p.value < 10^{-16}$). Vale a pena salientar que devido a publicação incompleta(Draft) do genoma da galinha esses resultados podem alterar substancialmente com a publicação da versão completamente montada do genoma da ave.

Alinhamentos contra o genoma do camundongo	Rato	Homem	Galinha
Número de alinhamentos	1474875	1783266	333869
Média do tamanho dos alinhamentos	1321,71	669,78	314,47
Desvio padrão do tamanho dos alinhamentos	1734,73	787,12	318,36

Tabela 2 - Medidas resumo do tamanho dos alinhamentos interespecies tendo com base o genoma do camundongo.

5.1.4 - Identificação de seqüências conservadas interespecies

Foi desenvolvido um script Perl que a partir de do módulo DBI (Data Base Interface) faz seleções dentro da base local de sintênia permitindo a identificação de trechos de seqüência conservados a uma distância pré-determinada a genes de interesse. Tais trechos conservados são, em última análise, seqüências compartilhadas por um par de genomas analisados evidenciando forte seleção positiva para a manutenção dessas seqüências. Em um primeiro momento, os genes de interesse são mapeados na montagem mm9 do genoma do camundongo com o auxílio das informações disponibilizadas tanto pela USCS e NCBI (National Center for Biotechnology information). Uma vez mapeados esses genes são determinados, a partir de seleções da base de sintênia, trechos compartilhadas por um par ou um grupo de genomas de interesse. Em geral utiliza-se uma combinação de três ou mais espécies para a construção de um filtro filogenético eficiente. A figura 3 traz um fluxograma com os passos utilizados para a identificação de tais seqüências conservadas.

Tais seqüências identificadas foram devidamente anotadas, exportadas e organizadas localmente. A ferramenta de busca foi desenvolvida de modo a ser facilmente inserida em análises seriais e permitir a fácil alteração de parâmetros de busca. Esse restritivo critério permitiu a identificação de trechos de DNA altamente conservados durante a evolução dessas espécies. As seqüências de DNA mapeadas em exons foram excluídas, essa exclusão se baseia no fato que devido a sua já reconhecida conservação funcional, se essas seqüências funcionais fossem aceitas integralmente enormes trechos seriam aceitos como conservados para a posterior busca computacional de TFBS. Essa dificuldade é evidenciada principalmente quanto um grande número de genes com distintos níveis de conservação exônica são analisados serialmente.

Figura 3 - Fluxograma ilustrando os passos para a determinação de seqüências conservadas interespecies.

A figura 4 traz uma representação gráfica dos escores de conservação em regiões flangeadoras ao gene *Aldh1a2* do camundongo. Esses escores são calculados a partir de janelas deslizantes de 100 pb, trechos altamente similares têm escores mais altos do que outros. A análise da figura destaca a existência de trechos com conservação significativa contra outros três genomas alinhados (Homem, rato e galinha). A seta vermelha destaca uma pequena região de 888 pares de base altamente conservados que apresentou fortes evidências de se tratar de uma SIT (Sequência Indutora de Transcrição) funcionalmente relacionada com a expressão do gene em questão. Altos níveis de conservação podem ser interpretados como possíveis indícios de conservação funcional já que a ocorrência de mutações pode alterar substancialmente a ligação de um fator de transcrição imprescindível para o correto controle da expressão do gene em questão. Esse procedimento é conhecido como “footprint” filogenético e tem sido utilizado extensivamente na busca de elementos funcionais, principalmente curtos, em longas seqüências genômicas sem informação a priori de regiões propensas a apresentarem tais elementos.

Figura 4 - Representação gráfica dos trechos conservados interespécie próximos ao gene *aldh1a2* do Camundongo (seta vermelha). No eixo Y estão os escores de conservação da região em janela de 100 pares de base, um trecho considerado conservado deve apresentar picos de conservação contra todos os genomas analisados.

5.1.5 - Identificação de famílias de fatores de transcrição conservadas evolutivamente

Algoritmos de predição de TFBS, devido a sua já discutida baixa especificidade, são extremamente suscetíveis a eventos falsos positivos. Essa grande proporção de sítios indesejáveis torna a identificação de sítios relevantes biologicamente um desafio. Para identificar tais sítios, nós desenvolvemos um script Perl que identifica famílias de fatores de transcrição que se mantiveram presentes e com forte conservação posicional durante a evolução de genes ortólogos em espécies distintas. Para cumprir tal objetivo, as seqüências sintênicas conservadas identificadas nas bases de dados de sintênia foram alinhadas com o auxílio da ferramenta de alinhamento múltiplo Clustal [4]. Em cada uma das seqüências alinhadas foram preditos TFBS utilizando um corte mínimo de 90% de estringência em relação ao centro da matriz (quarteto de nucleotídeos mais conservados) e a matriz consenso completa, cada sítio predito foi classificado em sua respectiva família, a partir do critério descrito no item “Definição de classificação dos TFBS em famílias” descrito anteriormente. Só foram consideradas conservadas, famílias presentes na seqüência do camundongo e presente em todas as outras seqüências em uma mesma posição relativa ao alinhamento múltiplo das mesmas.

Esse procedimento restritivo foi extremamente eficiente para identificar famílias de fatores de transcrição relacionados cujos sítios de ligação se mantiveram presentes durante a evolução paralela de diferentes genomas modernos. A conservação posicional desses sítios é fundamental para que o fator de transcrição ligado possa interagir ativamente com a maquinaria de transcrição diretamente ou através do complexo de moderação [10]. Ao respeitar as limitações tridimensionais existentes na interação DNA-Proteína, o procedimento desenvolvido permitiu a criação de um restritivo filtro filogenético. Esse procedimento quando associado ao uso de seqüências conservadas interespecies permite a identificação de famílias altamente conservadas e funcionalmente relevantes. Na figura 5 traz um fluxograma que destaca passo a passo esse procedimento e a sua implementação computacional

Figura 5 - Fluxograma destacando o procedimento para a identificação de famílias de fatores conservadas evolutivamente

5.2- Análise de enriquecimento de sítios

5.2.1. - Criação de distribuições nulas para testes de enriquecimento de sítios em uma sequência de interesse.

O principal objetivo de um teste de enriquecimento é definir se existe uma proporção incomum de sítios preditos em uma sequência de interesse. Para efetuarmos esse teste construímos duas distribuições nulas: distribuição randômica e distribuição biológica. A distribuição randômica foi construída a partir de um milhão de seqüências de 1000 pares de base geradas computacionalmente e em cada uma delas foram preditos TFBS usando 90% de estringência utilizando o algoritmo de busca Perl desenvolvido anteriormente. Todos os sítios identificados (aproximadamente 26 GB de informação) foram organizados em uma base de dados relacional PostgreSQL, essa implementação foi fundamental para permitir múltiplas consultas simultâneas assim como testes de enriquecimento combinados para mais de um sítio. Essa distribuição visa modelar a chance de múltiplas ocorrências de um mesmo fator em uma condição estritamente aleatória. Embora tenha sua validade, um modelo puramente aleatório não descreve satisfatoriamente os eventos e pressões evolutivas de um genoma e por isso desenvolvemos paralelamente uma distribuição biológica.

A distribuição biológica foi construída a partir de um milhão de seqüência de 1000 pb, obtidos diretamente das regiões flaqueadoras do conjunto completo de genes conhecidos no genoma do camundongo (montagem mm9). Essa distribuição visa aproximar o modelo nulo aos padrões de seqüência selecionados durante a evolução do genoma murino evitando assim um possível enviesamento do modelo completamente aleatório. Em cada uma das seqüências selecionadas foram preditos TFBS usando a mesma estringência utilizada no modelo aleatório. De modo similar à distribuição aleatória, o grande conjunto de sítios preditos foi organizado em uma base de dados relacional dedicada. A implementação de ambas as distribuições foi feita de tal forma a permitir a busca de tanto sítios individuais como combinação de elementos.

5.2.2. - Definição do teste de significância estatística aplicado para a identificação do enriquecimento de sítios.

Nós desenvolvemos uma rotina de análise para a construção de uma distribuição nula robusta para cada sítio individual ou combinação dos mesmos a partir das distribuições nula e biológica construídas anteriormente. Essa distribuição específica é construída a partir da seleção de sítios preditos em 100000 trechos de seqüência com o mesmo tamanho da seqüência de interesse a ser analisada. O número total de sítios individuais ou para grupos de sítios identificados em cada trecho é determinado e o conjunto total do número de sítios preditos é utilizado para a construção da distribuição nula para esse sítio. Essa distribuição construída visa representar o número de sítios que se esperaria prever em uma seqüência puramente aleatória ou em seqüência biológicas extraídas diretamente de um genoma moderno. O mesmo procedimento pode ser estendido para a combinação de sítios ou de famílias dos mesmos permitindo alterações nos padrões de busca e mineração de dados.

Uma vez tendo sido construída uma distribuição nula Gaussiana padrão a significância do desvio entre o número de sítios observados em sua seqüência de interesse e o número esperado dentro da distribuição é facilmente obtida. Tanto a base de dados de seqüências como os algoritmos de seleção na base e determinação de significância podem ser obtidos http://www.ime.usp.br/~maugusto/TESE_DOUTORADO/. A figura 6 destaca os passos necessários para a construção de uma distribuição nula de sítios e de que forma é determinada a significância do observado. O mesmo procedimento foi aplicado para ambas as distribuições de seqüências: biológica e aleatória.

> Random Seq 1
TACTAGAAGTTTTGCGACGTTCTAAGAGTTGGACGAAATGTTTC
GCGACCTAGGATGAGGTCGCCCTAGAAAATAGATTTCTGCTACT
CTCCTCATAAGCAGTCCGGTGTATCGAAAGTACAAGACTCCACG
GGCGCTCGGCGTCTTGGCTAATCCCCGTACATGTTGTTATAAAT
AATCAGTAGAACTCTGTGTTAGAGGGTGGAGTGACCTTAAATC

Uma sequência aleatória de 1000
pares de base é construída

ACTCTTTTAGACCAAATAAGGGCAAATTA

5' 3'

Predição computacional de sítios
de transcrição(TFBS)

Família	CONTAGEM
GATA	3
PAX	1
STAT	2

Contagem do número de
sítios para cada família de TFBS

Aplicação do teste de enriquecimento comparando o
número de sítios preditos na sequência de interesse
com o esperado da distribuição nula. A área
vermelha sob a curva permite a determinação exata
do p.valor da observação

Construção de uma distribuição nula do número
de sítios preditos em 100000 trechos de DNA
com o tamanho a sequência de interesse.

Figura 6 - Fluxograma destacando os passos utilizados para a determinação da
distribuição nula sítios específica e de que forma é obtida a significância do observado.

5.3. - Agrupamento de genes a partir das informações preditas em suas regiões regulatórias

5.3.1. - Identificação de seqüências conservadas flanqueadoras a genes de interesse e predição computacional de sítios de interação nas mesmas.

Com o auxílio das ferramentas Perl desenvolvidas anteriormente, blocos sintênicos altamente conservados foram identificados a uma distância máxima de 30kb a montante ao TSS (Transcription Start Site) e a jusante ao códon de terminação de genes de interesse definidos anteriormente. Devido a existência de múltiplas variantes de splicing para cada um dos genes, nós utilizamos a montagem referência mais longa, também conhecida como RefSeq de cada um. Sem dúvida outras estruturas regulacionais poderiam ter sido identificadas a partir do uso de outras montagens de splicing gênicas.

Como critério de seleção, as seqüências foram consideradas conservadas se e somente se apresentassem alinhamentos significantes contras os genomas humano, da galinha e rato em suas montagens mais recentes. Blocos sintênicos conservados identificados em cada um dos genes foram organizados localmente e cada um deles foi utilizado com entrada para o algoritmo de detecção de TFBS, foi utilizada a estringência mínima de detecção de 90% tanto para a matriz completa e core. Os sítios identificados no conjunto total de blocos sintênicos conservados de um gene foram classificados em famílias, segundo o critério familiar descrito anteriormente. As famílias preditas foram organizadas em strings respeitando sua ordem de predição (sempre no sentido 5' -> 3' do DNA genômico). A partir desse momento o trecho de DNA correspondente a um bloco sintênico foi abstraído para uma seqüência de famílias de fatores de transcrição cujos sítios foram ordenadamente preditos no trecho genômico. A figura 7 traz um fluxograma que destaca e ilustra os passos utilizados para a construção das strings representando cada um dos blocos sintênicos conservados identificados.

Figura 7 - Fluxograma descrevendo os passos utilizados para a construção das strings de famílias de fatores preditas em um bloco sintênico conservado próximo a um gene de interesse

5.3.2. - Normalização das contagens de TFBS em seqüências genômicas de interesse.

Após determinação das famílias de fatores (Item 5.9), ficou claro que algumas das famílias determinadas são constituídas por elementos cujas matrizes de detecção são consideravelmente mais degeneradas do que de outras famílias. Para testar se essa grande heterogeneidade no número de TFBS preditos para essas famílias levaria algumas famílias a se tornarem artificialmente mais representativas do que outras, nós corrigimos as contagens de uma determinada família a partir de uma normalização z-score. Para essa correção foram determinados a média e o desvio padrão do conjunto de contagens para uma mesma família no grupo de genes de interesse. Com essa simples correção as contagens de uma determinada família foram representadas por seu z-score dentro de uma distribuição Gaussiana padrão.

5.3.3. - Agrupamento de genes a partir das contagens de famílias de TFBS preditas em seus blocos sintênicos.

O conjunto de dados foi organizado em uma matriz, na qual em cada uma de suas linhas estão as contagens corrigidas de cada uma das 51 famílias para cada um dos genes definidos anteriormente. Os genes foram agrupados em uma árvore de semelhança através de um algoritmo de máxima verossimilhança que faz uso da correlação de Pearson como métrica de similaridade, entre os elementos. A sustentabilidade de cada ramo da árvore solução foi determinada com o auxílio do pacote R dedicado PvcLust e para isso foram utilizados um milhão de reamostragens bootstrap parciais [11]. A pacote PvcLust se utiliza de métodos de verossimilhança para a construção de uma distribuição de possíveis árvores sem a necessidade de uma árvore solução inicial para ser reamostrada [12]. Esse procedimento permite que as reamostragens bootstrap parciais construam uma distribuição não enviesada para a definição da sustentabilidade dos ramos. A árvore determinada como solução é aquela que demonstrou os maiores valores de sustentabilidade em seus ramos. Uma das grandes vantagens desse pacote é permitir a comparação entre os valores de sustentabilidade obtida pelo método não enviesado e a obtida pela abordagem bootstrap clássica.

5.4. Referências dessa seção

1. Portales-Casamar, E., et al., JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**(Database issue): p. D105-10.
2. Sepulveda, J.L., et al., GATA-4 and Nkx-2.5 coactivate Nkx-2 DNA binding targets: role for regulating early cardiac gene expression. *Mol Cell Biol*, 1998. **18**(6): p. 3405-15.
3. Crooks, G.E., et al., WebLogo: a sequence logo generator. *Genome Res*, 2004. **14**(6): p. 1188-90.
4. Larkin, M.A., et al., Clustal W and Clustal X version 2.0. *Bioinformatics*, 2007. **23**(21): p. 2947-8.
5. Chen, Q.K., G.Z. Hertz, and G.D. Stormo, MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci*, 1995. **11**(5): p. 563-6.
6. Kel, A.E., et al., MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*, 2003. **31**(13): p. 3576-9.
7. Cartharius, K., et al., MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 2005. **21**(13): p. 2933-42.
8. Matys, V., et al., TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 2003. **31**(1): p. 374-8.
9. Kolbe, D., et al., Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res*, 2004. **14**(4): p. 700-7.
10. Brown, C.D., D.S. Johnson, and A. Sidow, Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science*, 2007. **317**(5844): p. 1557-60.
11. Suzuki, R. and H. Shimodaira, Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 2006. **22**(12): p. 1540-2.
12. Shimodaira, H., An approximately unbiased test of phylogenetic tree selection. *Syst Biol*, 2002. **51**(3): p. 492-508.

Resultados - Seção 1

6. Resultados

6.1. Secção 1 - Aplicação dos métodos computacionais desenvolvidos para o estudo de uma possível SIT (Seqüência Indutora de Transcrição) próxima ao gene Aldh1a2 de camundongo.

6.1.1 Motivação biológica:

O contínuo uso de modelos embriológicos permitiu a identificação de uma miríade de fatores de transcrição envolvidos com esses processos de maturação e comprometimento de células neurais. Tipicamente esses fatores se organizam de forma hierárquica aonde fatores expressos anteriormente são responsáveis diretos pela expressão de outros fatores localizados “abaixo” na casca gênica [1]. A correta identificação desses fatores e o entendimento da fisiologia da interação entre os mesmos são fundamentais para entender os elementos responsáveis pela determinação e especialização das diferentes subpopulações do tubo neural [2]. Predições computacionais de TFBS (Transcription Factor Binding Site) têm sido utilizadas com sucesso para a pré-identificação de possíveis fatores de transcrição envolvidos com o desenvolvimento neural, o uso do termo pré-identificação advém do fato de que todos esses sites precisam ser posteriormente confirmados em ensaios biológicos [3]. Os sítios de ligação para fatores de transcrição, em geral, se organizam em pequenas regiões genômicas conhecidos como SIT (Seqüência Indutora de Transcrição) ou, em inglês, enhancers e a sua identificação e estudo permitem um melhor entendimento das vias fisiológicas os quais os genes estão inseridos.

A sinalização realizada pela vitamina A, derivada diretamente do ácido retinóico, é parte integral dos mecanismos que modelam o desenvolvimento embriológico de estruturas neurais do tubo neural embriológico [4]. Em vertebrados, a sinalização de ácido retinóico (AR) é realizada basicamente por duas famílias de genes que codificam genes para retináldeido desidrogenases. Os genes raldh1, raldh2 e raldh3 são parálogos da família ALDH1 e o gene raldh4 pertencente à família ALDH8. Dentre esses genes, destaca-se a proteína RALDH2 que demonstra a maior afinidade ao substrato retináldeido, precursor do ácido retinóico, sendo a primeira a ser devidamente identificada em embriões de galinha e camundongo [5]. Dentre outras funções, o AR é fundamental para a correta formação do tubo

neural, participando de diferentes programas de desenvolvimento do sistema nervoso central [6]. Entender os eventos de regulação responsáveis pelo controle de expressão temporal e de localização permite entender o papel desse importante mediador no desenvolvimento e especialização das estruturas neurológicas do tubo neural embrionário. Um dos principais mecanismos utilizados é uma gradação na concentração do fator de transcrição SHH (Sonic Hedgehog Homologue) e das BMP (Bone Morphogenetic Proteins). O efeito competitivo dessas duas classes de proteína é responsável pela ativação de outros fatores de transcrição que disparam programas de especialização irreversíveis. Atualmente já se identificaram 66 fatores de transcrição relacionados a esse processo, compostos por 42 fatores homeodomínio, 11 fatores com o motivo hélice-*loop*-hélice básico, 8 com domínios de ligação *Zinc finger* e o outros cinco subdivididos em classes menos representativas[7]. O tubo neural embrionário pode ser subdividido em secções radiais aonde diferentes fatores de transcrição, auxiliados por proteínas de sinalização, dão origem às células progenitoras tanto de neurônios sensoriais como moto-neurônios (Figura 1) [8]. A figura 1A destaca as diferentes concentrações das proteínas BMP e SSH no eixo dorso-ventral do embrião, as concentrações combinadas dessas duas classes gera “micro” ambientes propícios que produzem diferentes células progenitoras. A figura 1B destaca, a partir da ação combinada de fatores de transcrição, a formação de domínio progenitores (retângulos pontilhados) que são responsáveis diretos por especializar diferentes populações de neurônios (formas elípticas). As populações comprometidas de neurônios podem se subdividir em dois grandes grupos: neurônios dorsais e ventrais e essas respectivamente dão origem a neurônios sensoriais e a moto-neurônios [9].

Figura 1 - Desenvolvimento de diferentes populações de neurônios a partir da gradação das concentrações dos fatores SSH e BMP. A -) Gráfico que ilustra as diferentes concentrações de SSH e BMP. B -) Esquema que demonstra a partir de um corte sagital de uma medula espinhal embrionária como os domínio progenitores (retângulos) dão origem a diferentes populações de neurônios (elipses). Cada uma das elipses da origem a uma subpopulação de neurônios específica. Essa ilustração foi adaptado de Lewis et al, 2006. [8]

6.1.2. Objetivos dessa seção

- Identificar inicialmente possíveis SITs atuantes na expressão do gene ALDH1a2 através de prospecção de seqüências conservadas.
- Predizer computacionalmente sítios de interação DNA - Proteína usando uma restritiva abordagem de *footprint* filogenético.
- Analisar o papel regulacional de fatores de transcrição candidatos na expressão temporal e espacial do gene através de modelos de expressão.
- Testar a ocorrência de trechos de DNA que apresentam uma proporção incomum de fatores de transcrição individuais ou família de fatores.

6.1.3. Justificativa dessa seção

O entendimento de como as estruturas neurológicas são desenvolvidas nos permite entender melhor o papel de diferentes fatores de transcrição responsável pela especialização tecidual de um organismo pluricelular. O uso criterioso de ferramentas de bioinformática para a detecção de sítios de ligação abre novas possibilidades para a análise a um custo experimental muito menor e quando associados à testes funcionais permite a criação e confirmação de modelos de controle a expressão do gene alvo. Os testes de enriquecimento permitem a identificação de sítios de ligação de fatores de transcrição que estão anormalmente concentrados em curtas seqüências de interesse. Nossa abordagem desenvolvida nesse trabalho pode ser estendida para outros fenótipos de interesse.

6.1.4. - Resultados dessa seção

6.1.4.1- Identificação de uma possível SIT (Seqüência Indutora de Transcrição) associada à expressão do gene *aldh1a2* de camundongo.

Com o uso de ferramenta de prospecção Genoma browser oferecido publicamente pela Universidade Califórnia Santa Cruz (<http://genome.ucsc.edu/>), foram identificados 72 trechos de DNA conservados nas regiões flangeadoras e intrônicas do gene *Aldh1a2* (Figure 2). Tais trechos são denominados conservados devido a grande semelhança de suas bases com seqüências ortólogas de espécies distantes evolutivamente. Dentre esses blocos se destacam três com destacada conservação (Figura 2 - Triângulos vermelhos), o que denota uma forte pressão seletiva positiva para a manutenção dessas seqüências e consequentemente um possível papel funcional dessas bases de DNA.

Experimentos com o uso de embriões transientes permitem com auxílio de genes repórteres, estudarem a expressão de genes de interesse em um modelo *in vivo* [10]). O uso dessa abordagem permite que experimentos de análise de expressão se aproximem da realidade biológica a qual tais seqüências e fatores de transcrição estão submetidos durante o desenvolvimento embrionário. Através do uso de embriões de camundongos transientes foram determinados os papéis de cada um dos três blocos de seqüência conservados em promover e controlar a expressão do gene *Aldh1a2* (Figura 3). O produto gênico marcado pode ser identificado em todos os embriões desenvolvidos e a sua localização dorso-caudal está em acordo com o conhecimento prévio sobre a expressão do gene *Aldh1a2* na condição sob estudo (Figura 3 A-E). A expressão do gene é limitada a placa tectal da medula espinhal embrionária e a análise das Figuras 3 F e G corroboram o papel do bloco de seqüência conservado ser realmente um SIT, atuando no controle da expressão quantitativa e espacial do gene sob estudo. Essa seqüência de DNA foi nomeada SIT 1G e foi objeto de estudo usando ferramentas computacionais para a predição de possíveis sítios de ligação em suas bases de DNA e técnicas de eletroporação para, através de ensaios biológicos, determinarem um modelo de ação para fatores de transcrição agindo nessa possível SIT.

Figura 2- Trechos conservados multi-espécie- As barras magenta representam seqüências com significativo alinhamento de sequências de diferentes espécies (citar aqui as espécies) com o genoma do camundongo. Dos blocos conservados se destacam três que se mantiveram conservados até o genoma de peixes e esses foram testados posteriormente em embriões de camundongo transientes (indicados pelas setas no trecho superior).

Figura 3 - Embriões transientes- Fig A-D Quatro embriões em posição lateral aonde pode se observar a expressão do gene *Aldh1a2* guiado exclusivamente pela presença da SIT 1G. E- F e G corte longitudinal do embrião aonde destaca a expressão do gene na placa tectal corroborando com a esperada localização da expressão do gene sob estudo. H e I

6.1.4.2.- Predição computacional de sítos de interação de fatores de transcrição na sequência SIT 1G.

Dentre esses blocos, a SIT 1G composta de 888 pares de base foi o único trecho que apresentou atividade de modulação na expressão do gene *Aldh1a2* e foi posteriormente analisada computacionalmente. A análise evolutiva filogenética determinou que a sequência dessa SIT putativa está presente nos genomas de animais marsupiais e placentários, mas foi perdida nos monotremados e em peixes cartilaginosos. A sequência conservada do camundongo foi alinhada contra sequências de 10 organismos que apresentaram desatacada conservação com a sequência candidata a SIT. A escolha desses organismos se baseou em um critério puramente evolutivo aonde foram selecionados dois representantes de cada um dos grandes clados evolutivos (Figura 4). Foram selecionados sete mamíferos, duas aves e 2 anfíbios cujos genomas estavam com bom nível de cobertura. Os representantes mamíferos podem ser novamente subdivididos em eutérios e marsupiais: os eutérios têm gestação completamente interna enquanto marsupiais fazem uso de uma bolsa materna para completar o desenvolvimento do feto. Foram selecionados dois representantes marsupiais: canguru e gambá e cinco eutérios que se subdividem novamente em dois representantes primatas (homem e chimpanzé), em dois representantes da ordem Rodentia (rato e camundongo) e em um representante dos cingulados (tatu-galinha). A figura 4 traz uma tabela com o nome popular, nome científico e a classificação de cada um dos animais alinhados. Para melhor representar a escolha desses organismos nós nos utilizamos de um cladograma evolutivo que ilustra as relações evolutivas entre os representantes. A classificação se baseou em publicações de autores recentes que, com o uso de análises genômicas, determinaram uma árvore de espécies mamíferas consistente usando abordagens de relógio molecular [11].

As sequências conservadas dessas 11 espécies foram alinhadas com o auxílio da ferramenta de alinhamento múltiplo dedicada ClustalW [12]. Foram usados parâmetros adaptados para o alinhamento desse tipo de sequência aonde foi usada uma penalidade alta para a abertura de um gap e uma penalização baixa para a extensão do gap. Esses parâmetros combinados têm sido amplamente utilizados em alinhamentos genômicos ou de sequências curtas com grande distância evolutiva. A figura 5 traz o trecho inicial do alinhamento múltiplo das sequências e destaca a forte conservação inter-espécie dessa sequência. Um grande trecho de aproximadamente 600 pares de bases (entre posições 51 e 680 na sequência do camundongo) apresentou destacada conservação, com grandes blocos que apresentaram alinhamento perfeito (Figura 5).

Figura 4 - Animais os quais tiveram trechos de seus genomas alinhados com a SIT 1G de camundongo. O cladograma traz uma representação dos clados envolvidos e seus respectivos representantes

Figura 5 - Trecho do alinhamento múltiplo de seqüências conservadas de 10 espécies de interesse com a seqüência da SIT 1G sob estudo

Usando o alinhamento múltiplo como base, cada uma das seqüências alinhadas foi usada como entrada para o algoritmo de detecção de TFBS usando um corte de estrinência 90%. Esse algoritmo implementa o algoritmo MATCH@[12] de detecção de sítios de ligação e foi programado de forma modular para permitir o seu uso irrestrito (Mais detalhes no item 5.1.2., página 53 do presente texto). Sítios preditos na seqüência do camundongo foram classificados em famílias, seguindo o critério descrito anteriormente, e só foram consideradas famílias de valor biológico aquelas que estavam presentes em uma mesma posição relativa ao alinhamento múltiplo em ao menos um representante dos clados descritos na figura 4. As famílias conservadas identificadas foram organizadas levando-se em conta a posição de detecção e se a predição foi feita na fita molde ou complementar. O uso do critério de filogenético de conservação familiar associado à retirada de sítios total ou parcialmente sobrepostos permitiu identificar 131 sítios de altíssimo valor biológico. A tabela 1 destaca o número de sítios preditos na seqüência da SIT 1G. A figura 6 destaca um trecho da SIT e as famílias de fatores de transcrição preditas. Um detalhe importante é que embora esse trecho seja proporcionalmente curto, este apresenta uma grande quantidade de predições, fruto provável da combinação única de seus nucleotídeos, e isto foi mais bem explorado posteriormente nos testes de enriquecimento.

FAMILIAS	CONTAGEM	FAMILIAS	CONTAGEM
AP1	2	MSX	1
Cdx	12	MYB	10
E-BOX	2	NF-AT	1
ETS	7	NKX	3
FORKHEAD	12	PDF	1
GATA	15	PRX	2
GEN_INI	9	Pax	15
HOMEODOMAIN	7	STAT	15
LIM_HOMEODOMAIN	1	Sox	10
LMO	1	TCF	1
MADS	1	TEF	1
MEIS	1	ZINC	1
		TOTAL	131

Tabela 1 - Famílias de fatores de transcrição e as contagens de sítios conservados preditos na SIT 1G para cada uma das classes.

Figura 6 - Trecho conservado da SIT e as famílias de fatores de transcrição preditas na mesma. Cada traço colorido acompanhado pelo nome da família dá a exata localização da predição

Esse interessante caráter evolutivo da SIT e a sua associação com o controle da expressão do gene *Alh1a2*, a tornaram objeto de estudos funcionais e exploratórios em nosso grupo de pesquisa (Laboratório de Genética e Cardiologia Molecular). Experimentos de eletroporação permitem a análise detalhada tanto da localização quanto à quantificação da expressão de um gene repórter. Nesses experimentos, um vetor de expressão é construído, com o trecho de interesse, e posteriormente analisado através da expressão de um gene repórter marcado com o fluoróforo eGFP. Esse tipo de experimento tem sido combinado com mutações sítio-dirigidas para a determinação dos efeitos da presença e/ou ausência de sítios de ligação para fatores de transcrição. Essa abordagem foi utilizada para a elucidação da regulação de genes envolvidos em condições fisiológicas de interesse.

Inicialmente foi construído um vetor com todos os 888 pares de base constituintes da SIT sob investigação e o resultado obtido foi utilizado como um controle (Figura 7A). A sequência da SIT foi subtraída através de deleções aninhadas de seus trechos e foram construídos vetores de expressão com essas seqüências. Cada um dos vetores teve determinada a sua ação na expressão de um gene repórter através de experimentos de eletroporação. Essas construções e seus principais resultados serão melhor explorados no decorrer desse item da tese apresentada. Inicialmente um trecho de 279 pares de base foi subtraído da extremidade 5' da seqüência, o que resultou na expressão generalizada do gene repórter e inferiu a possível presença de um fator de transcrição com forte caráter repressivo para a expressão do gene repórter nas porções mais ventrais da notocorda do embrião (Figura 7B). Dentre os sítios preditos computacionalmente nesse trecho destaca-se um sítio de ligação para o fator de transcrição LIM (Lhx3) com já conhecido papel repressivo. Esse sítio de ligação foi alvo de um experimento de mutação sítio-dirigida para a determinação de seu papel no controle da expressão do gene repórter. A análise do resultado que essa mutação recapitulou o resultado observado no figura 7B (Figura 7C). O trecho restante composto por 609 pares de base foi subdividido em quatro trechos [A, B, C, D] que individualmente foram testados com auxílio de vetores de expressão (Figura 7 D-H). Os resultados determinaram que o trecho B e D foram responsáveis pela expressão localizada na placa tectal da notocorda do embrião, mas não na porção dorsal, o que evidencia o papel de um outro possível repressor além do fator LIM. A predição computacional de sítios nesses trechos identificou a presença de três sítios combinados: aonde 2 sítios próximos para fatores das famílias TCF/CDX (um no trecho B e um na porção 5' do trecho D) e dois sítios próximos para fatores das famílias TCF/TGIF (porção 3' do trecho D). Devido a presença desses sítios alvos potenciais, e com o objetivo de analisar o seu papel individualizado, o

trecho D foi subdividido em D1 e D2.

Figura 7 – Eletroporações- ilustrações à esquerda, coradas em verde, eletroporações das construções repórteres e a direita, coradas em vermelho, eletroporações controle para determinar a efetividade do processo de eletroporação. A -) Expressão tectal do gene repórter guiado pela sequência da SIT 1G completa B-) Expressão ubíqua do gene repórter guiado pelo SIT incompleta C-) Mutação sítios dirigida ao sítios LIM e a replicação da expressão ubíqua D-) Sequência é subdividida em quatro trechos [A-D] que serão testados individualmente E a H -) Análise do papel de cada um dos subtrechos no controle da expressão de gene repórter.

Uma vez determinados os papéis dos trechos das sequências, sítios de interesse preditos computacionalmente foram testados através de mutações sítio dirigidas, cada um dos sítios foi mutado individualmente e de forma combinada em duplas e trios (Figura 8). Os resultados sugerem que a expressão do gene repórter é guiada pela presença dos dois sítios TCF/CDX (D1) e reprimida pelos sítios TGIF (D2). Interessantemente, na presença do repressor, um sítio sozinho não é capaz de guiar satisfatoriamente a expressão generalizada do gene repórter limitando-a apenas a região da placa tectal da medula espinhal (Figuras 8 A-C). Quando ambos os sítios promotores são mutados, somente o sítio TCF do sítio combinado TCF/TGIF guia somente a expressão tectal do gene repórter (Figura 8D). A mutação tripla de todos os sítios eliminou completamente a expressão do gene repórter

(Figura 8G). Todos os fatores candidatos foram alvos para uma profunda busca na literatura científica especializada e tanto a localização da expressão do gene repórter quando sua expressão quantitativa confirma o prévio conhecimento sobre a localização e ação desses fatores de transcrição. Esses resultados combinados nos permitiram criar um modelo de ação entre fatores de transcrição e o seu papel na modulação da expressão do gene *aldh1a2* representado pela construção repórter.

Figura 8 – Mutações sítios dirigidas - ilustrações à esquerda, coradas em verde, eletroporações das construções repórteres e a direita, coradas em vermelho, eletroporações controle para determinar a efetividade do processo de eletroporação. A-C -) Mutações sítios dirigidas individuais e o seu resultado na expressão do gene repórter D-E -) Mutações sítios dirigidas de duplas de sítios e o seu resultado na expressão do gene repórter G-) Mutações sítio dirigida tripla e o seu resultado na expressão do gene repórter

6.1.4.3. O modelo funcional para a sequência SIT 1G

A partir dos resultados dos experimentos de eletroporação e dos modelos computacionais de predição foi possível construir um modelo para a ação de fatores de transcrição em sítios de ligação presentes na sequência da SIT 1G e o seu papel no controle da expressão do gene alvo *Aldh1a2*. Foram identificados sítios para quatro fatores de transcrição que podem ser subdivididos em dois grupos: Fatores com papel indutor composto por fatores CDX e TCF e fatores com papel de repressão à expressão composto por fatores TGIF e LIM (Figura 9). Os repressores são responsáveis diretos por limitar a expressão do gene repórter apenas a placa tectal da medula espinhal do embrião à exemplo da já conhecida expressão do gene estudado (Figura 7A). O modelo proposto explicou de forma satisfatória as observações experimentais e as já conhecidas funções dos fatores, partícipes do modelo, em outros modelos biológicos relevantes [8]. A expressão do fator de transcrição TCF é fortemente influenciada pela presença do fator de transcrição Wnt cuja expressão é localizada ubiquamente e quando o repressor forte LIM se mostra ausente estimula a expressão do gene alvo nas regiões centro-ventrais e dorsais da medula do embrião. De modo complementar nós estudamos padrões estatísticos dentro das predições de TFBS para identificar outras especializações dessa interessante sequência funcional.

Figura 9 - Modelo proposto para a ação de fatores de transcrição atuando na sequência da SIT 1G e a sua consequência para expressão do gene *Aldh1a2*. Nesse gene atuam dois ativadores Cdx e Tcf esse último intermediando a ação do fator de transcrição Wnt. Os dois repressores TGIF e LIM são responsáveis tanto pela localização quando o fino controle da expressão do gene.

6.1.4.4. Testes de enriquecimento

A SIT 1G de apenas 888 pares de base organiza sítios para fatores de transcrição com destacado papel no controle da expressão do gene *Aldh1a2* no desenvolvimento embrionário de camundongos. Adicionalmente as abordagens descritas nos itens anteriores, nós desenvolvemos e efetuamos testes de enriquecimento para a identificação de TFBS cuja a frequência de ocorrência na SIT 1G fosse maior que um valor esperado. Essa interessante característica biológica, já explorada anteriormente por outros autores [13], é baseada na premissa que sítios para fatores de transcrição imprescindíveis na expressão de um gene específico devem estar concentrados nas regiões regulatórias próximas ao gene em que atuam. Essa super-representação facilitaria muito a localização e interação do fator de transcrição com seu gene alvo. Devido a seu importantíssimo papel no controle da expressão do gene *Aldh1a2*, foi testada a hipótese de que SIT 1G tivesse se especializado para a ligação de fatores de transcrição específicos ou para uma família de fatores de transcrição.

Em um primeiro momento, o número de sítios preditos computacionalmente para cada fator individual foi determinado. Não foi utilizado o critério de seleção filogenética de famílias conservadas, pois o mesmo não pode ser aplicado às seqüências da distribuição. Esses resultados destacam as 702 predições de sítios realizados nas 888 pares de base da SIT 1G aonde foi predito 1 sítio a cada 1,26 pares de base. As contagens individuais e familiares foram comparadas com os valores esperados determinados pelas distribuições nulas de sítios randômica e biológica. Tais distribuições são constituídas de 1 milhão de seqüências de 1000 pares de bases gerados randomicamente ou obtidos diretamente de trechos conservados próximos de genes mapeados no genoma do camundongo (A construção dessas distribuições foi mais bem discutida no item 5.2.1, página 62). Foram determinados dois diferentes cenários para a aplicação dos testes de enriquecimento: teste para cada fator de transcrição isoladamente e teste para o número de contagens de sítios para uma determinada família.

Para efetuarmos os testes de enriquecimento foram selecionados nas distribuições nulas organizadas em bases de dados relacionais de alta eficiência 100000 seqüências de mesmo tamanho a SIT (888 pares de base). Essas seqüências foram geradas computacionalmente (Distribuição aleatória considerando igual frequência para as quatro bases de DNA) ou retiradas aleatoriamente do genoma do camundongo (Distribuição Biológica). A partir dessa amostra foram obtidos o valor médio e o desvio padrão esperados da distribuição nula para o número de sítios para um fator individual ou agrupado em

famílias. Para o teste de enriquecimento foi usado um teste Gaussiano aonde são comparados os valores observados e esperados para a determinação da significância do desvio entre as medidas (O teste de enriquecimento aplicado é explicado no item “Teste de enriquecimento”, item 5.2.2., página 63).

6.1.4.5. Resultados - Distribuição de seqüências aleatórias

Inicialmente o teste foi aplicado utilizando-se as seqüências organizadas na distribuição aleatória nos dois cenários descritos anteriormente e as variáveis comparadas cujos p.valores observados eram inferiores a 0.05 foram selecionadas e estão organizadas respectivamente nas tabelas 3,4 e 5.

Família	# Observados	μ (Distribuição)	σ (Distribuição)	P.value
AP1	7	1,999	1,893	0,00411
CDX	28	4,271	2,533	0,00000
E-BOX	8	0,833	1,051	0,00000
ETS	25	2,959	3,958	0,00000
FORKHEAD	22	0,821	1,412	0,00000
GATA	66	8,456	8,996	0,00000
HOMEODOMAIN	21	1,713	2,341	0,00000
KRUPEL_LIKE	1	0,113	0,347	0,00527
LIM_HOMEODOMAIN	1	0,082	0,293	0,00085
MADS	4	1,087	1,068	0,00319
MSX	3	1,026	1,028	0,02739
MYB	20	5,960	5,614	0,00619
NKX	11	0,980	1,160	0,00000
PDF	2	0,346	0,593	0,00263
POU	2	0,308	0,584	0,00189
Pax	42	6,730	8,930	0,00004
STAT	138	12,499	7,155	0,00000
Sox	20	1,520	1,785	0,00000
TEF	8	3,227	1,910	0,00622

Tabela 3 – Testes de enriquecimento para família de fatores - Resultados significantes ($P \leq 0,05$) dos testes de enriquecimento para famílias de fatores de transcrição usando distribuição nula randômica e um teste normal para a definição da significância do observado.

A tabela 3 destaca os resultados da comparação do número de sítios combinados de uma dada família, os resultados destacam três famílias KRUPEL_LIKE, LIM_HOMEODOMAIN e STAT cujos números de sítios observados na SIT são significativamente superiores aos esperados na distribuição nula. Foi identificado apenas um

sítio para cada uma das famílias KRUPEL_LIKE e LIM_HOMEODOMAIN na SIT 1G sob estudo computacional, mas devido ao impressionante caráter restritivo das matrizes monoposicionais de detecção isso é aproximadamente 10 vezes mais do que se esperaria na distribuição nula. Em outras palavras, se espera a ocorrência de um sítio da família LIM em apenas 8% das seqüências geradas computacionalmente e essa média e o desvio padrão observado determinam o destacado enriquecimento desses sítios. O fator Lhx3, constituinte único da família LIM_HOMEODOMAIN na base de dados TRANSFAC usada neste trabalho, se mostrou fundamental para a expressão localizada do gene Aldh1a2 na placa tectal do embrião de camundongo e o seu enriquecimento destaca ainda mais o papel exercido pelo fator no desenvolvimento neural de embriões[14]. A família KRUPEL_LIKE, composta apenas pelo fator de transcrição Gklf, teve apenas um sítio identificado na SIT 1G. Esse fator, por sua vez, o único representante dessa família dentro da base Transfac ® 6.0 utilizada nas análises. Essa família de fatores de transcrição está fortemente associada a processos de regeneração de axônios em neurônios adultos [15]. Os fatores da família STAT (Signal Transducers and Activator of Transcription) participam ativamente da transdução de sinal citoplasmático e já foram associados a muitos aspectos de processos como crescimento, sobrevivência e diferenciação celular em neurônios [16]. Foram identificados nos 888 pares de base da SIT 1G , 131 sítios para fatores da família STAT, muito acima do que se esperava na distribuição nula (apenas 12).

[M1] Comentário: HERE!!!!!!

SÍTIO	FAMÍLIA	# Observados	μ (Distribuição)	σ (Distribuição)	P.value
V\$CDP_02	HOMEODOMAIN	1	0,123	0,351	0,00627
V\$CDXA_01	CDX	17	5,389	2,554	0,00000
V\$CDXA_02	CDX	11	3,166	1,956	0,00003
V\$CHX10_01	HOMEODOMAIN	1	0,129	0,363	0,00821
V\$E47_01	E-BOX	1	0,173	0,415	0,02324
V\$ETS1_B	ETS	1	0,200	0,448	0,03705
V\$FOXD3_01	FORKHEAD	1	0,089	0,300	0,00119
V\$FOXJ2_01	FORKHEAD	1	0,132	0,364	0,00857
V\$FOXO4_01	FORKHEAD	4	1,624	1,296	0,03335
V\$FREAC7_01	FORKHEAD	2	0,225	0,474	0,00009
V\$GKLF_01	KRUPEL LIKE	1	0,117	0,353	0,00619
V\$HFH8_01	FORKHEAD	1	0,131	0,361	0,00813
V\$HMGY_Q6	Sox	13	3,546	2,002	0,00000
V\$HNF3B_01	FORKHEAD	2	0,213	0,462	0,00006
V\$IPF1_Q4	PDF	2	0,349	0,594	0,00271
V\$LHX3_01	LIM_HOMEODOMAIN	1	0,081	0,292	0,00082
V\$MEF2_Q6	MADS	4	1,085	1,062	0,00302
V\$MSX1_01	MSX	3	1,028	1,027	0,02748
V\$NKX62_Q2	NKX	8	1,818	1,411	0,00001
V\$PAX2_02	Pax	40	18,308	5,870	0,00011
V\$STAT1_03	STAT	23	12,740	4,348	0,00914
V\$STAT3_02	STAT	11	6,337	2,788	0,04719
V\$TCF1P_Q6	TCF	2	0,558	0,752	0,02761
V\$TEF1_Q6	TEF	8	3,247	1,914	0,00652
V\$TST1_01	POU	1	0,127	0,354	0,00678
V\$XFD2_01	FORKHEAD	1	0,099	0,315	0,00209

Tabela 4 – Testes de enriquecimentos individuais - Resultados significantes ($P \leq 0,05$) dos testes de enriquecimento para fatores de transcrição individuais usando distribuição nula randômica e um teste normal para a definição da significância do observado.

De modo complementar aos testes de enriquecimento de famílias de fatores, foram efetuados testes de enriquecimento para cada um dos fatores de transcrição preditos na SIT 1G. A soma total de sítios foi comparada com os valores obtidos em uma amostra de 100 mil seqüências de 888 pares de base e a partir de um teste Gaussiano foi determinada à significância do desvio entre sítios observados e esperados (Tabela 4). A análise de enriquecimento nesse cenário permitiu identificar um total de 26 fatores de transcrição cujos números de sítios observados na SIT 1G estavam significativamente distantes dos valores esperados da distribuição nula. A análise da tabela 4 destaca além dos sítios individuais Lhx3 e Gklf, responsáveis enriquecimento das famílias aos quais estão inseridos, a família FORKHEAD. Essa família de fatores já foi associada a processos de desenvolvimento embrionário e seus elementos apresentam um domínio de ligação específico e em geral precisam de outros fatores da mesma família ou não para efetuarem ativamente o seu papel na regulação de um gene [17, 18]. O fator de transcrição CHX10 é um fator *homeobox* crítico para eventos de proliferação celular e cuja expressão já foi associada diretamente com a de fatores de transcrição da família POU, os quais por sua vez são fundamentais para a

especialização de subpopulações específicas de neurônios [19]. A SIT 1G apresenta sítios de bom valor biológico para ambos os fatores e esses resultados em conjunto com os resultados de experimentação biológica destacam a combinação de nucleotídeos altamente direcionada presente na SIT 1G sob estudo.

6.1.4.6. Resultados - Distribuição biológica

Embora promissores, os resultados dos testes de enriquecimento baseados em uma distribuição de seqüências aleatórias geradas computacionalmente, podem estar seriamente enviesados pela utilização de seqüências que não refletem exatamente os milhões de eventos evolutivos que um genoma moderno já sofreu. Com o objetivo de testar essa hipótese de enviesamento, foi criada uma distribuição nula biológica, de mesmo tamanho, composta por trechos de 1000 pares de base retirados diretamente do genoma do camundongo e os testes Gaussianos de enriquecimento foram repetidos. Foram testados os mesmos dois cenários: teste de família de fatores agrupados e teste individual para cada um dos fatores de transcrição com sítios preditos no SIT sob análise. Os resultados significantes ($P \leq 0,05$) estão apresentados nas tabelas 5 e 6.

Família	# Observados	μ (Distribuição)	σ (Distribuição)	P.value
AP1	7	2,305	2,055	0,01118
AP2	10	3,456	3,608	0,03487
CDX	28	7,647	5,629	0,00015
E-BOX	8	1,084	1,315	0,00000
ETS	25	3,841	5,109	0,00002
FORKHEAD	22	1,551	2,483	0,00000
GATA	66	7,203	7,896	0,00000
HOMEODOMAIN	21	1,572	2,300	0,00000
MSX	3	0,874	1,006	0,01732
MYB	20	4,982	5,128	0,00170
NKX	11	1,607	2,438	0,00006
PDF	2	0,359	0,628	0,00449
Pax	42	11,686	17,173	0,03876
STAT	138	17,547	10,235	0,00000
Sox	20	2,618	3,097	0,00000
TCF	3	0,998	1,194	0,04679
TEF	8	3,273	2,043	0,01034

Tabela 5 - Teste de enriquecimento para família de fatores - Resultados significantes ($P \leq 0,05$) dos testes de enriquecimento para famílias de fatores de transcrição usando distribuição nula biológica e um teste normal para a definição da significância do observado.

Os resultados da tabela 5 destacam um total de 16 famílias de fatores consideradas enriquecidas no SIT 1G. Em uma comparação com os resultados na tabela 3 houveram algumas mudanças bem significativas, como a não inclusão de família como LIM_HOMEODOMAIN ($\mu = 0,2161$ e $\sigma = 0.675$) e KRUEPEL_LIKE ($\mu = 0,871$ e $\sigma = 2,517$) e a inclusão das famílias AP2 e TCF entre os resultados significantes. A família TCF em conjunto com fatores da família CDX tiveram preditos dois sítios de ligação combinados preditos em duas regiões da SIT 1G. Os experimentos de eletroporação determinaram que ambos os sítios têm central papel na estimulação da expressão do gene *Aldh1a2*. Experimento de mutação dirigida em ambos aboliu completamente a expressão gene repórter. Um detalhe interessante é que embora não exista uma grande diferença entre as famílias significantes em ambas as distribuições, o número médio de sítios preditos (μ) na distribuição biológica é constantemente maior do que o mesmo predito na distribuição aleatória. Essa diferença se deve principalmente a seleção positiva de motivos de DNA cuja proporção de nucleotídeos individuais é muito diferente de 25%.

SÍTIO	FAMÍLIA	# Observados	μ (Distribuição)	σ (Distribuição)	P.value
V\$CDP_02	HOMEODOMAIN	1	0,087	0,301	0,00121
V\$CHX10_01	HOMEODOMAIN	1	0,138	0,389	0,01327
V\$HMGY_Q6	Sox	13	5,247	3,042	0,00541
V\$IPF1_Q4	PDF	2	0,356	0,628	0,00445
V\$MSX1_01	MSX	3	0,869	0,991	0,01579
V\$TCF1P_Q6	TCF	2	0,430	0,668	0,00938
V\$TEF1_Q6	TEF	8	3,254	2,003	0,00891
V\$TST1_01	POU	1	0,179	0,435	0,02971

Tabela 6 - Testes de enriquecimentos individuais - Resultados significantes ($P < 0,05$) dos testes de enriquecimento para fatores de transcrição individuais usando distribuição nula biológica e um teste normal para a definição da significância do observado.

Os testes de enriquecimento individual, organizados na tabela 6, destacam um total de 9 fatores de transcrição cujo número de sítios observados na SIT 1G se mostrou significativamente distante ($P < 0,05$) dos valores esperados dentro da distribuição biológica de seqüências. Esse número é muito menor do que os 26 fatores de transcrição que seriam considerados enriquecidos baseando-se apenas nos valores obtidos na distribuição aleatória de seqüências. A exemplo do observado no teste aplicado em famílias combinadas, os valores médios esperados na distribuição biológica são em geral superiores aos observados na distribuição aleatória. Essa distribuição pode ser interpretada como mais restritiva do que a aleatória exigindo maiores números de sítios observados para a identificação do

enriquecimento. O sítio TST1 permite a ligação de um fator da família POU que já se mostrou imprescindível para a migração cortical de neurônios no desenvolvimento do córtex cerebral de mamíferos [17]. Os fatores família TCF se ligam fortemente aos 8 sítios TCF1P presentes na SIT 1G e participam de vias de sinalização para a especialização neuronal em geral mediados pela expressão do fator Wnt [18]. É interessante notar que esses interessantes sítios se mostraram enriquecidos mesmo usando uma distribuição mais restritiva.

6.1.4.7. - Resultados e Discussão dessa seção

O uso de ferramentas computacionais para a predição computacional de TFBS permitiu a identificação de famílias de fatores de transcrição cujos sítios preditos puderam auxiliar o entendimento da sequência de indução de transcrição 1G associada ao gene *Aldh1a2*. A predição foi auxiliada por critérios restritivos de seleção de seqüências conservadas e a posterior determinação de famílias conservadas a partir do alinhamento múltiplo de seqüências sintênicas. A SIT 1G se mostrou rica em potenciais sítios preditos, apresentando total de 702 sítios de interação que puderam ser resumidos para 131 sítios para famílias de fatores de transcrição de alto valor biológico. Essa diminuição se deve ao uso de critérios filogenéticos de seleção e a diminuição da redundância da predição, aonde sítios sobrepostos para o mesmo fator são resumidos a apenas um. Essa grande “poder” regulacional está intimamente relacionado à fortíssima conservação de sua seqüência de nucleotídeos aonde mutações sofridas por essa seqüência sofreram pressão seletiva negativa. Essa pressão é importante para a manutenção da combinação de suas bases de DNA, o que por si só evidencia uma possível função regulacional importante. O papel central da enzima conversora de ácido retinóico no desenvolvimento neurológico de embriões já foi destacado por outros autores [19]. Devido a sua importância funcional, a expressão temporal e espacial desse gene é efetuada por um tênue, mas imprescindível controle realizado por fatores de transcrição. Análises genômicas de prospecção identificaram pequenos trechos conservados próximos ao gene *Aldh1a2* e a relação entre esses trechos e o controle da expressão do gene foi realizada com o auxílio do uso de embriões transientes. Os resultados combinados determinaram que apenas um dos três trechos testados se mostrou importante para a expressão de um gene repórter em uma condição *in vivo* específica. Esse pequeno trecho de apenas 888 pares de base foi nomeado SIT (Seqüência Indutora de Transcrição) 1G e passou a ser alvo de estudos computacionais e ensaios biológicos para a determinação exata de seu papel na expressão do gene sob estudo.

Dentre os sítios preditos se destacam quatro sítios anteriormente correlacionados ao desenvolvimento do sistema nervoso central embrionário: dois sítios combinados para fatores Cdx /TCF, um sítio para o fator Lhx3 e um sítio para o TGIF. Para analisar o papel de cada um desses sítios dentro da regulação do gene, a seqüência da SIT 1G foi dividida em trechos e o papel isolado de cada trecho e de cada um dos sítios candidatos foi determinado através do uso de construções repórteres e análises de eletroporações. O uso de mutações sítio

dirigidas permitiu a elucidação de diversos eventos de regulação e quando associada às predições computacionais de sítios de ligação permitem a determinação dos eventos de regulação de genes de interesse. A análise dos resultados e a busca na literatura especializada nos permitiram a construção de um modelo consistente da atuação dos quatro fatores de transcrição e seus respectivos sítios na regulação da expressão quantitativa e espacial do gene *Aldh1a2* [14].

A construção das distribuições nulas e a sua utilização para testes de enriquecimento permitiu identificar sítios para fatores individuais ou para famílias de fatores que estavam anormalmente concentrados em região regulatórias de genes de interesse. A ferramenta computacional CONFAC faz uso da premissa de enriquecimento, mas os números de sítios não são comparados contra distribuições de seqüências e sim com o auxílio de uma amostra de genes controle escolhidos aleatoriamente[13] . Embora interessante tal abordagem pode ser comprometida pela escolha da amostra controle, o que exige procedimentos de reamostragens parciais ou totais para a determinação da sustentabilidade do resultado. Em um trabalho recente Krizhanovsky, V., et al construiu uma distribuição empírica composta por 1000 seqüências geradas computacionalmente e usando testes de hipótese efetuou testes de enriquecimento em regiões de interesse próximas ao gene *Math1*[3]. Essa abordagem é prejudicada pelo limitado número de seqüências utilizadas para a construção da distribuição nula e pelo uso de seqüências aleatórias computacionais que pouco refletem as informações organizadas nas bases de um genoma. Usando uma abordagem similar, nós construímos duas grandes distribuições com aproximadamente 1 milhão de seqüências de 1000 pb cada uma. A primeira é composta por seqüências geradas computacionalmente e a segunda de trechos conservados retirados diretamente do genoma do camundongo, os parâmetros principais das distribuições são utilizados para a aplicação de testes de enriquecimento em variáveis de interesse.

Todos os sítios individuais foram testados em ambas a distribuições individualmente e agrupados em famílias de fatores e os resultados demonstram claramente que a SIT 1G tem uma composição de bases única e especializada. Grandes quantidades absolutas de sítios foram preditas para fatores de transcrição associados direta ou indiretamente com o papel desse pequeno trecho de DNA em controlar a expressão do gene *Aldh1a2*. A análise de fatores agrupados em famílias destacou famílias como CDX, TCF, LIM_HOMEODOMAIN cujos elementos foram testados, em experimentos de eletroporação, e apresentaram importante papel na regulação temporal e espacial da expressão do gene. Testes realizados nas contagens de famílias de fatores permitiram identificar importantes relações que não

seriam detectadas nos testes de fatores individuais.

Os testes individuais permitiram identificar sítios em contexto biológico com o fenótipo estudado que não foram considerados candidatos interessantes no começo da análise computacional. A combinação dos resultados do testes de enriquecimento usando ambas as distribuições se mostrou uma boa ferramenta para a prospecção de proporções anormais de sítios e o seu cruzamento com a literatura especializada destacou interessantes relações funcionais existentes entre famílias de fatores ou fatores individuais. A comparação dos resultados baseados nas duas distribuições de seqüências validou o valor do uso de um modelo mais próximo à realidade biológica, o que permitiu continuar identificando os sítios mais interessantes, mas eliminou possíveis relações de pouco valor. As abordagens computacionais de predição de TFBS e identificação de sítios enriquecidos podem ser aplicadas em outros fenótipos de interesse. A determinação prévia de sítios de ligação interessantes poupa tempo e recurso por permitir uma análise específica com pouca perda de sensibilidade.

6.1.4.8. - Referências utilizadas nessa seção

1. Bertrand, N., F. Medevielle, and F. Pituello, *FGF signalling controls the timing of Pax6 activation in the neural tube*. Development, 2000. **127**(22): p. 4837-43.
2. Muhr, J., et al., *Convergent inductive signals specify midbrain, hindbrain, and spinal cord identity in gastrula stage chick embryos*. Neuron, 1999. **23**(4): p. 689-702.
3. Krizhanovsky, V., et al., *Math1 target genes are enriched with evolutionarily conserved clustered E-box binding sites*. J Mol Neurosci, 2006. **28**(2): p. 211-29.
4. Altmann, C.R. and A.H. Brivanlou, *Neural patterning in the vertebrate embryo*. Int Rev Cytol, 2001. **203**: p. 447-82.
5. Moss, J.B., et al., *Dynamic patterns of retinoic acid synthesis and response in the developing mammalian heart*. Dev Biol, 1998. **199**(1): p. 55-71.
6. Wilson, L., et al., *Retinoic acid and the control of dorsoventral patterning in the avian spinal cord*. Dev Biol, 2004. **269**(2): p. 433-46.
7. Kioussi, C. and M.K. Gross, *How to build transcriptional network models of mammalian pattern formation*. PLoS One, 2008. **3**(5): p. e2179.
8. Lewis, K.E., *How do genes regulate simple behaviours? Understanding how different neurons in the vertebrate spinal cord are genetically specified*. Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1465): p. 45-66.
9. Lewis, K.E. and J.S. Eisen, *Hedgehog signaling is required for primary motoneuron induction in zebrafish*. Development, 2001. **128**(18): p. 3485-95.
10. Joshi, R.P. and K.H. Schoenbach, *Electroporation dynamics in biological cells subjected to ultrafast electrical pulses: a numerical simulation study*. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics, 2000. **62**(1 Pt B): p. 1025-33.
11. Springer, M.S. and W.J. Murphy, *Mammalian evolution and biomedicine: new views from phylogeny*. Biol Rev Camb Philos Soc, 2007. **82**(3): p. 375-92.
12. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
13. Karanam, S. and C.S. Moreno, *CONFAC: automated application of comparative genomic promoter analysis to DNA microarray datasets*. Nucleic Acids Res, 2004. **32**(Web Server issue): p. W475-84.
14. Castillo, H.A., et al., *Insights into the organization of dorsal spinal cord pathways from an evolutionarily conserved raldh2 intronic enhancer*. Development. **137**(3): p. 507-18.
15. Moore, D.L., et al., *KLF family members regulate intrinsic axon regeneration ability*. Science, 2009. **326**(5950): p. 298-301.
16. Foshay, K.M. and G.I. Gallicano, *Regulation of Sox2 by STAT3 initiates commitment to the neural precursor cell fate*. Stem Cells Dev, 2008. **17**(2): p. 269-78.
17. Taylor, J.J., T. Wang, and K.L. Kroll, *Tcf- and Vent-binding sites regulate neural-specific geminin expression in the gastrula embryo*. Dev Biol, 2006. **289**(2): p. 494-506.
18. Morikawa, Y., et al., *Detailed Expression Pattern of Foxp1 and Its Possible Roles in Neurons of the Spinal Cord during Embryogenesis*. Dev Neurosci, 2009. **31**(6): p. 511-522.
19. McEvelly, R.J., et al., *Transcriptional regulation of cortical neuron migration by*

POU domain factors. Science, 2002. **295**(5559): p. 1528-32.

Resultados - Seção 2

6.2.- Agrupamento de genes baseando-se unicamente em sinas transcrpcionais preditos em suas regiões regulatórias

6.2.1- Algoritmos de análise filogenética

Algoritmos de agrupamento têm sido utilizados com sucesso em áreas díspares como engenharia, biologia e matemática. O principal objetivo desses algoritmos é particionar uma amostra em grupo de elementos similares entre si e significativamente diferentes de outros grupos. Esse particionamento é realizado através de uma função critério determinada *a priori* e está usada para classificar os elementos do conjunto amostral (Ex: Classificar um tecido em normal ou canceroso). No caso específico da biologia molecular tais algoritmos são largamente utilizados para a análise de dados gerados a partir de experimentos de mensuração de expressão global como Microarrays e SAGE [1].

Um tipo especial de algoritmo de agrupamento são os algoritmos filogenéticos amplamente utilizados no estudo de genes e proteínas. Algoritmos filogenéticos fazem uso de um alinhamento múltiplo de seqüências (aminoácidos ou bases de DNA) e a partir de uma matriz de substituições determinam uma árvore solução que acomode as evidências observadas nas seqüências individuais [2] (Figura 1). A figura 1A traz um alinhamento múltiplo entre os aminoácidos da proteína codificada pelos genes parálogos *Gata1*, esse alinhamento foi construído com o auxílio do algoritmo de alinhamento múltiplo dedicado ClustalW [3]. A análise desse alinhamento destaca a grande conservação existente entre essas seqüências, principalmente no trecho final das mesmas. A figura 1B traz um agrupamento filogenético construído a partir das evidências observadas no alinhamento múltiplo. Esse resultado foi obtido através do algoritmo de agrupamento do tipo “Neighbor Joining” disponível na ferramenta ClustalW, tal agrupamento tenta construir um dendograma que aproxima seqüência semelhantes e as separa de seqüências divergentes de aminoácidos. Como esperado a árvore solução respeita o parentesco evolutivo entre as espécies representadas pelas seqüências de aminoácidos dos genes ortólogos *Gata1*.

Esse tipo de abordagem tem sido utilizado para a classificação de novos genes recentemente identificados, em estabelecer parentescos evolutivos entre espécies e para estabelecer parentescos evolutivos entre genes de uma mesma espécie. Os genes, a exemplo das espécies, são frutos de um longo processo evolutivo e consequentemente o

estudo dos padrões de suas seqüências permite estabelecer uma relação de parentesco entre genes de interesse. Genes paralogos são frutos de uma duplicação de um gene ancestral, aonde diferentes pressões evolutivas podem estar sendo exercidas em cada um dos genes. Recentemente os fatores de transcrição *Dbx1* e *Dbx2*, genes paralogos, foram alvo de um estudo funcional que determinou que embora esses genes fossem evolutivamente relacionados, apresentaram padrões de expressão diferentes dentro do processo de desenvolvimento neurológico de peixe-zebra [4].

Algoritmos filogenéticos fazem uso de abordagens estatísticas e computacionais para encontrar a melhor árvore solução em tempo eficiente. A árvore-solução resultante pode ser interpretada como a que apresentou a melhor subdivisão de elementos levando-se em conta as informações individuais [2]. Essa busca exige que um grande número de árvores seja inspecionado e a cada novo elemento adicionado, a complexidade do algoritmo aumenta exponencialmente, assim, o número de árvores a serem inspecionadas em um agrupamento de 50 elementos é incrivelmente grande. Esse grande universo de árvores levou ao desenvolvimento de heurísticas computacionais eficientes para diminuir o custo computacional da análise, entre elas destacam-se os processos de reamostragens parciais e totais [5].

Em geral, ao invés de se analisar todo o universo de árvores possíveis, se determina algumas árvores solução *a priori* que, por sua vez, têm a sustentabilidade de seus ramos determinada através de reamostragens. O processo de reamostragem conhecido como bootstrap retira um elemento da análise e repete o agrupamento com os genes restantes e ramos presentes na maioria das reamostragens são considerados consistentes e de bom valor [5]. Essa solução é largamente utilizada dentro da literatura científica, mas sofre de uma limitação de difícil solução. Essa limitação advém do fato que a primeira árvore solução, que será posteriormente reamostrada, é uma boa solução, mas não a melhor para o conjunto de dados sob análise. Com o objetivo de apresentar uma solução a essa limitação, Shimodaira et al, propôs uma abordagem que visa diminuir esse enviesamento através de uma busca heurística de árvores solução em um espaço amostral de árvores[6]. Somente árvores consideradas de altíssimo valor serão reamostradas, diminuindo ao mesmo tempo o custo computacional e o enviesamento da escolha inicial da árvore solução[7].

Figura 2 - Agrupamento filogenético- A-) Alinhamento múltiplo entre seqüências homologas do gene Gata1 obtidas em diferentes espécies B-) Resultado do agrupamento filogenético dessas seqüências usando o algoritmo Neighbor Joining de reconstrução filogenética.

Algoritmos de filogenia têm sido aplicados em diversas áreas dentro da biologia molecular e medicina e a sua associação a outras fontes de informação permitiu a elucidação de doenças de interesse. Essas abordagens mistas muitas vezes fazem uso de informações como: dados de expressão global ou informações fenéticas para a construção de árvores solução que representem fielmente o fenótipo sob investigação. Um dos objetivos dessa seção é desenvolver uma abordagem de agrupamento filogenética baseada unicamente nos sinais transcrpcionais preditos nas regiões regulatórias de genes. Em um primeiro momento, se faz necessário determinar um bom modelo biológico que será posteriormente explorado por essa abordagem computacional.

6.2.2.- Motivação biológica - Desenvolvimento neurológico de murinos

Entender e compreender o funcionamento de diferentes estruturas neurológicas em vertebrados nos permite satisfazer tanto nossa curiosidade intelectual quanto ajudar a elucidar doenças neurológicas. Muito têm se entendido sobre o funcionamento de neurônios cerebrais e como estímulos primitivos como dor, tato, calor e frio são conduzidos para a medula espinhal e de que forma tais estímulos são traduzidos em respostas motoras controladas [8]. Esse grande desenvolvimento advém do fato que esses estímulos primitivos podem ser estudados, com muito sucesso, em modelos animais simples como peixe-zebra, batráquios e lampreias. A medula espinhal é um órgão neurológico primário associado fortemente a respostas instintivas de um ser vivo, como, por exemplo, afastar imediatamente um membro em contato com calor excessivo. O entendimento de que modo os neurônios medulares trabalham recebendo estímulos e gerando respostas aos mesmos passa pela análise da estrutura histológica desse tecido. Para entendermos as diferentes subpopulações celulares organizadas na medula espinhal de um vertebrado é imprescindível estudarmos as populações de células progenitoras embrionárias e de que forma essas se especializam. Processos de desenvolvimento e especialização requerem significativas mudanças no perfil transcricional das células envolvidas, sendo essas mudanças guiadas principalmente por proteínas sinalizadoras e fatores de transcrição [9]. Fatores de transcrição quando localizados em seus respectivos sítios interagem ativamente com a maquinaria de transcrição, alterando a expressão do gene em questão.

Neurônios formam uma população celular extremamente especializada e são responsáveis diretos por coletar e transmitir as informações dentro de um sistema nervoso. Diversas subpopulações especializadas de neurônios podem ser encontradas na medula espinhal de um vertebrado adulto [10]. Nos últimos 10 anos, diversos experimentos foram conduzidos e identificaram ao menos 14 subpopulações de neurônios pós mitóticos em embriões de galinha e camundongo. Cada uma delas se desenvolve em uma região específica da medula espinhal embrionária (Figura 2A, página 72). Um dos principais mecanismos utilizados para essa impressionante especialização celular são as gradações nas concentrações tanto do fator de transcrição SHH (Sonic Hedgehog Homologue) quanto das proteínas BMP (Bone Morphogenetic Proteins) (Figura 2B, página 72). O efeito competitivo dessas duas classes de proteína é responsável pela ativação em cascata de outros fatores de transcrição que, por sua vez, disparam programas de especialização irreversíveis. Cada região, bem delimitada, da medula embrionária apresenta uma combinação única de fatores de transcrição e proteínas sinalizadoras, que são responsáveis diretos por desencadear complexas mudanças no transcriptoma dessas células e consequentemente especializando-as.

A medula embrionária pode ser subdivida em duas grandes regiões: os domínios progenitores aonde as células neurológicas iniciam o seu processo de especialização e as células pós mitóticas, células pré-especializadas, que continuam a se especializar sob a ação de outros fatores de transcrição. A medula pode ser novamente subdivida em domínios dorsais e ventrais aonde conjuntos totalmente diferentes de fatores de transcrição agem. A região ventral da medula embrionária pode ser subdivida em cinco pequenos domínios progenitores: p0, p1, p2, pMn e p3 que, por suas vez, dão origem a cinco populações de células pós mitóticas respectivamente V0,V1,V2, VMn e V3 aonde se destacam as ações dos fatores de transcrição Gata-2, Gata-3, Chx10 e Lhx3 [11]. As populações de neurônios ventrais resultantes desse processo de especialização podem ser subdivididas claramente através de marcadores moleculares como a expressão dos fatores de transcrição Evx1 e Pax2 [12]. A porção dorsal da medula pode ser subdividida em seis domínios progenitores: dp1, dp2, dp3, dp4, dp5, dp6 aonde a ação dos fatores de transcrição como Lhx1, Lhx5, Lhx2 e fatores da família Pax leva ao desenvolvido de seis subpopulação de neurônios pós mitóticos: dl1, dl2, dl3, dl4, dl5 e dl6. Essas subpopulações de neurônios quando totalmente maduras são especialmente envolvidas em processar e entregar estímulos nervosos a estruturas neurológicas periféricas. O estudo funcional das subpopulações dorsais de neurônios é prejudicado pela migração neuronal

dessas células para as estruturas neurológicas periféricas como músculos ou a epiderme [13].

Os processos de desenvolvimento e maturação de populações de neurônios são, a exemplo de outros eventos de desenvolvimento, totalmente associados à ação coordenada de fatores de transcrição. A expressão desses fatores de transcrição é intimamente relacionada com os padrões transcricionais (sítios de ligação para esses fatores) presentes em sua região regulatória. Identificar computacionalmente tais sítios de ligação e estudar tais predições permitirá entender melhor o papel dos fatores de transcrição da expressão tecido específica desses genes.

6.2.3. Objetivo principal dessa secção

Desenvolver uma abordagem de agrupamento de genes baseando-se unicamente em sítios de ligação para fatores de transcrição preditos em seqüências conservadas flanqueando genes de interesse. Os agrupamentos foram realizados usando como métricas a distância Euclidiana e a correlação de Pearson aplicadas diretamente às contagens absolutas de sítios para famílias de fatores de transcrição preditos nas regiões regulatórias desses genes.

6.2.3.1. Objetivos específicos dessa secção

- A partir de uma profunda pesquisa na literatura científica da área determinar um conjunto de fatores e proteínas sinalizadoras associadas a processos de especialização das diferentes subpopulações de neurônios na medula espinhal embrionária.
- Mapear no genoma do camundongo (montagem Mm9) os genes que codificam tais fatores de transcrição.
- Determinar através de filtros filogenéticos trechos conservados próximos aos genes selecionados no item anterior. Tais trechos devem compartilhar suas seqüências com homens, ratos e galinhas.
- Predizer computacionalmente sítios de ligação para fatores de transcrição disponíveis na base de dados pública Transfac 6.0® nos trechos considerados conservados dentro de uma janela com tamanho pré-estipulado.
- Contabilizar o número de fatores preditos para cada um das famílias de fatores de transcrição em cada um dos genes.
- Normalizar o conjunto de contagens das famílias de famílias através de uma correção z-score.
- Agrupar os genes com base em seus sinais transcripcionais preditos com o auxílio de heurística de reamostragens não enviesada
- Determinar o efeito das predições realizadas nas fitas complementar e molde individualmente em relação à montagem referência desses genes no agrupamento desses elementos.

- Analisar a árvore solução e determinar famílias de fatores de transcrição que sejam responsáveis total ou parcialmente pelos agrupamentos observados.
- Comparar os ramos da árvore solução com dados funcionais e de coregulação para a validação ou exclusão de ramos da árvore.

6.2.4. Justificativa dessa seção

O correto desenvolvimento das estruturas neurológicas é fundamental para a manutenção da vida de um ser vivo. Tais estruturas ocupam papéis centrais na coleta de informações, processamento e coordenação motora de um cordado. Tais funções vitais impõem uma grande pressão seletiva positiva para manutenção da regulação de genes envolvidos nesses processos, mutações nesses genes podem acarretar a inviabilidade fetal. Tentar entender quais são os fatores envolvidos direta ou indiretamente na especialização de diferentes subpopulações de neurônios trará relevantes informações acerca desses processos. O uso de algoritmo de agrupamento de sinais transcricionais preditos traz novas possibilidades para a análise de grupos de genes envolvidos em patologias de interesse. Genes com grande grau de coregulação ou expressão tecido específica são considerados bons alvos para a abordagem aqui proposta.

6.2.5. - Resultados

Construção de uma abordagem de agrupamentos baseada em sinais transcrpcionais preditos.

Com o objetivo de construirmos e testarmos um método de agrupamento nós definimos uma serie de quatro passos ou objetivos que serão seguidos para a construção e posterior teste dessa abordagem. A figura 2 traz um fluxograma aonde estão graficamente representados a ordem desses objetivos. Cada um dos passos destacados será mais bem explicado tanto em seus objetivos como seus resultados durante esse capítulo. Esse fluxograma será utilizado dentro desse capítulo para uma melhor compreensão e acompanhamento do leitor.

Figura 2- Fluxograma guia – Ilustração com os passos que serão utilizados para a construção e teste da abordagem de agrupamento proposta

6.2.5.1 – Determinação dos genes cujos produtos já foram associados com processos de especialização de neurônios.

Através de uma extensa busca na literatura especializada foram identificados 57 fatores de transcrição, os quais já foram determinados experimentalmente como apresentando papéis centrais no processo de especialização das subpopulações de neurônios. Todos esses genes foram mapeados no genoma do camundongo usando como base a sua montagem referência (Montagem Mm9 disponibilizada pela UCSC – <http://genome.ucsc.edu>). A tabela 1 organiza os resultados desse cruzamento e as referências científicas utilizadas para a determinação desse grupo de genes.

Figura 3 - Fluxograma guia- Determinação de um grupo de fatores já envolvidos com os processos de desenvolvimento nerurológico

Símbolo	Gene ID	RefSeq	Cromossomo	Início	Fim	Fita	Referência
<i>Gli1</i>	14632	NM_010296.2	chr10	126766937	126778635	-	[14]
<i>Gli2</i>	14633	NM_001081125.1	chr1	120730637	120950196	-	[15]
<i>Gli3</i>	14634	NM_008130.2	chr13	15555067	15806947	+	[16]
<i>Gata3</i>	14462	NM_008091.3	chr2	9778704	9799611	-	[17]
<i>Evx1</i>	14028	NM_007966.4	chr6	52263491	52268372	+	[18]
<i>Evx2</i>	14029	NM_007967.2	chr2	74493672	74497476	-	[18]
<i>Barhl1</i>	54422	NM_019446.3	chr2	28763203	28771942	-	[19]
<i>Gbx1</i>	231044	NM_015739.2	chr5	24010243	24032666	-	[20]
<i>Tlx1</i>	21908	NM_021901.2	chr19	45225204	45231433	+	[21]
<i>Sim1</i>	20464	NM_011376.3	chr10	50615456	50708958	+	[22]
<i>cux</i>	13047	NM_198602.2	chr5	136724004	137041698	-	[23]
<i>GATA2</i>	14461	NM_008090.4	chr6	88143884	88157023	+	[17]
<i>Prrxl1(drg11)</i>	107751	NM_001001796.4	chr14	33413112	33462432	+	[24]
<i>NKX2-2</i>	18088	NM_010919.2	chr2	147008887	147012138	-	[25]
<i>NKX6-2</i>	14912	NM_183248.2	chr7	146765197	146768692	-	[26]
<i>GSH1</i>	14842	NM_008178.2	chr5	148000271	148002522	+	[27]
<i>GSH2</i>	14843	NM_133256.2	chr5	75471625	75473918	+	[28]
<i>Ascl1(MASH1)</i>	17172	NM_008553.4	chr10	86953785	86956405	-	[21, 27]
<i>Neurog1(NGN1)</i>	18014	NM_010896.2	chr13	56351859	56353524	-	[23]
<i>Neurog2(NGN2)</i>	11924	NM_009718.2	chr3	127336062	127338549	+	[29]
<i>Pax2</i>	18504	NM_011037.3	chr19	44831883	44910517	+	[30]
<i>Lhx1</i>	16869	NM_008498.2	chr11	84332880	84339036	-	[31]
<i>Lhx5</i>	16873	NM_008499.5	chr5	120881909	120891466	+	[31]
<i>Hlxb9</i>	15285	NM_019944.1	chr5	29799801	29805010	-	[32]
<i>Dbx1</i>	13172	NM_001005232.1	chr7	56886868	56892205	-	[33]
<i>Dbx2</i>	223843	NM_207533.2	chr15	95453993	95485202	-	[4]
<i>isl2</i>	104360	NM_027397.3	chr9	55388955	55393985	+	[34]
<i>irx3</i>	16373	NM_008393.2	chr8	94322423	94325273	-	[35]
<i>Barhl2</i>	104382	NM_001005477.1	chr5	106881541	106887185	-	[36]
<i>scl</i>	21349	NM_011527.2	chr4	114730415	114744360	+	[37]
<i>Math-1</i>	11921	NM_007500.4	chr6	64679150	64681229	+	[29]
<i>Pax3</i>	18505	NM_008781.3	chr1	78099266	78193711	-	[38]
<i>Pax7</i>	18509	NM_011039.2	chr4	139293995	139388883	-	[38]
<i>BarH1</i>	54422	NM_019446.3	chr2	28763203	28771942	-	[39]
<i>Brn3a</i>	18996	NM_011143.4	chr14	104861540	104867216	-	[40]
<i>Lhx2</i>	16870	NM_010710.3	chr2	38194800	38225248	+	[41]
<i>Lhx9</i>	16876	NM_010714.3	chr1	140721762	140744156	-	[41]
<i>CDX(caudal)</i>	12590	NM_009880.2	chr18	61178518	61195853	-	[42]
<i>Engrailed</i>	13798	NM_010133.2	chr1	122499063	122504568	+	[30]
<i>Ets</i>	23871	NM_011808.2	chr9	32503626	32565405	+	[43]
<i>Isl1</i>	16392	NM_021459.4	chr13	117088488	117099893	-	[44]
<i>Lim3</i>	16871	NM_001039653.1	chr2	26055731	26058386	-	[45]
<i>Mnx1</i>	15285	NM_019944.1	chr5	29799801	29805010	-	[32]
<i>Nkx6.1</i>	18096	NM_144955.2	chr5	102088215	102093730	-	[26]
<i>Pax6</i>	18508	NM_013627.4	chr2	105509052	105537226	+	[46]
<i>Msx</i>	17703	NM_010836.2	chr7	147232155	147234985	-	[47]
<i>Wnt</i>	22420	NM_009526.3	chr1	74818465	74831893	+	[48]
<i>Zic</i>	22771	NM_009573.3	chr9	91255182	91260612	-	[49]
<i>Hoxc9</i>	15427	NM_008272.3	chr15	102807462	102814875	+	[50]
<i>Lmx1a</i>	110648	NM_033652.5	chr1	169619688	169778864	+	[34]
<i>Mafb</i>	16658	NM_010658.2	chr2	160189496	160192738	-	[51]
<i>Msx1</i>	17701	NM_010835.2	chr5	38211729	38215824	-	[47]

<i>Msx2</i>	17702	NM_013601.2	chr13	53562249	53568149	-	[52]
<i>Pax3</i>	18505	NM_008781.3	chr1	78099266	78193711	-	[53]
<i>Pax7</i>	18509	NM_011039.2	chr4	139293995	139388883	-	[54]
<i>Zic2</i>	22772	NM_009574.3	chr14	122874605	122879550	+	[49]
<i>Zic4</i>	22774	NM_009576.2	chr9	91263809	91275824	+	[49]
<i>Cux2</i>	13048	NM_007804.2	chr5	122310224	122316074	-	[23]
<i>Nrp-2</i>	18187	NM_010939.2	chr1	62749891	62865266	+	[55]

Tabela 1 - Conjunto de fatores de transcrição ou proteínas de sinalização associados aos processos de desenvolvimento da espinha dorsal de cordados. A última coluna coleta as referências utilizadas para a determinação desse grupo de genes.

Esse grupo de genes compreende basicamente os principais marcadores para a descrição de células neurais progenitoras ventrais e dorsais. Os fatores da família LIM são fortes repressores para a expressão de genes ventrais e por isso suas expressões estão localizadas nos progenitores dorsais, principalmente subpopulações de neurônios dl1 e dl2. As expressões dos genes *Evx1*, *Pax2* e *Pax7* são consideradas bons indicadores para a identificação de uma célula progenitora ventral. Os outros genes organizados na tabela 1, embora não atuando diretamente como marcadores, já foram associados direta ou indiretamente com a especialização dessas estruturas neurais. Genes *Mnx1* (Moto neuron homeobox1) e seu paralog *Mnx2* são expressos exclusivamente na região progenitora de moto-neurônios e as suas expressões em neurônios maduros são consideradas ótimos marcadores para a identificação de células originárias nessa região da medula embrionária.

6.2.5.2 - Determinação do padrão de expressão desses genes no desenvolvimento neurológico embrionário.

A partir de um extensivo cruzamento com a literatura científica e com informações disponibilizadas diretamente pela base Transfac® foram determinados os padrões de expressão espacial dos genes selecionados dentro do desenvolvimento da medula embrionária. As tabelas 2 e 3 organizam as informações obtidas e nessas tabelas, as regiões específicas (colunas) aonde a expressão de um dado gene já foi observada foi preenchida com o algarismo 1 e região aonde a expressão não foi detectada ou não existem informações disponíveis foi preenchida com o algarismo 0. A tabela 2 organiza as informações em regiões progenitoras dorsais e a tabela 3 nas regiões ventrais da mesma e as referências utilizadas para construção de ambas as tabelas.

Figura 4 - Fluxograma guia – Determinação dos padrões de expressão desses genes e construção de uma árvore solução que organize tais informações

	RP	dP1	dP2	dP3	dP4	dP5	dP6	dl1	dl1B	dl2	dl3	dl4	dl5	dl6
<i>ascl1</i>	0	0	0	1	1	1	0	0	0	0	0	0	0	0
<i>barhl1</i>		0	0	0	0	0	0	1	0	0	0	0	0	0
<i>barhl2</i>	0	0	0	0	0	0	0	1	0	0	0	0	0	0
<i>cdx1</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1
<i>cux1</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1
<i>cux2</i>	1	0	0	0	0	0	0	1	0	1	1	0	0	0
<i>dbx-1</i>	0	0	0	0	0	0	1	0	0	0	0	0	0	0
<i>dbx-2</i>	0	0	0	0	0	0	1	0	0	0	0	0	0	0
<i>en1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>evx1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>evx2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>gata2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>gata3</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>gbx1</i>	0	0	0	1	1	1	1	0	0	0	0	0	0	0
<i>gli1</i>	0	0	0	1	1	1	1	0	0	0	1	1	1	1
<i>gli2</i>	0	0	1	1	1	1	1	0	0	0	0	1	1	1
<i>gli3</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1
<i>gsx1</i>	0	0	0	0	0	0	0	0	0	0	0	1	1	0
<i>gsx2</i>	0	0	0	0	0	0	0	0	0	0	1	1	1	0
<i>hoxc9</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1
<i>irx3</i>	0	1	1	1	1	1	1	0	0	0	0	0	0	0
<i>isl-1</i>	0	0	0	0	0	0	0	0	0	1	1	1	1	1
<i>isl2</i>	0	0	0	0	0	0	0	0	0	0	1	0	0	0
<i>lbx1</i>	0	0	0	0	0	0	0	0	0	0	0	1	1	1
<i>lhx1</i>	0	0	0	0	0	0	0	1	0	1	1	1	0	1
<i>lhx2</i>	0	0	0	0	0	0	0	1	1	0	0	0	0	0
<i>lhx3</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>lhx5</i>	0	0	0	0	0	0	0	1	0	1	1	1	0	1
<i>lhx9</i>	0	0	0	0	0	0	0	1	1	1	1	1	1	1
<i>lmx1a</i>	1	0	0	0	0	0	0	1	0	0	0	0	0	0
<i>mafb</i>	1	1	0	0	0	0	0	1	0	0	0	0	0	0
<i>mnx1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>msx1</i>	1	1	1	1	1	1	0	0	0	0	0	0	0	0
<i>msx2</i>	1	1	1	1	1	1	0	1	0	1	1	0	0	0
<i>neurog1</i>	0	0	1	1	1	1	1	0	0	0	0	0	0	0
<i>neurog2</i>	0	0	1	1	1	1	0	0	0	0	0	0	0	0
<i>nkx6-1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>nkx6-2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>olig3</i>	0	1	1	1	0	0	0	0	0	0	0	0	0	0
<i>pax2</i>	0	0	0	0	0	0	0	0	0	0	0	1	0	1
<i>pax3</i>	1	1	1	1	1	1	1	1	0	1	1	1	1	1
<i>pax6</i>	0	0	0	0	1	1	1	0	0	0	0	0	0	0
<i>pax7</i>	0	1	1	1	1	1	1	0	0	0	0	0	0	0
<i>pou4f1</i>	0	0	0	0	0	0	0	1	0	1	1	0	1	0
<i>prrx1</i>	0	0	0	0	0	0	0	0	0	0	1	0	1	0
<i>sim1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>tal1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>tlx1</i>	0	0	0	0	0	0	0	0	0	0	0	0	1	0
<i>tlx3</i>	0	0	0	0	0	0	0	0	0	1	1	1	1	0
<i>vsx2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<i>zic1</i>	1	1	1	1	1	1	0	1	0	1	1	1	0	0
<i>zic2</i>	1	1	1	1	0	0	0	1	0	1	1	0	0	0
<i>zic4</i>	1	1	1	1	1	1	1	1	0	1	1	1	1	1

	p1	p2	pMN	p3	V0	V0v	V0d	V1	V2	V2a	V2b	MN	V3	FP	DRG	dILA	dILB	Ref.
<i>asc1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[56]
<i>barhl1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[57]
<i>barhl2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[56]
<i>cdx1</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	[42]
<i>cux1</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	[8]
<i>cux2</i>	0	0	0	0	1	0	0	1	1	0	0	1	1	0	1	0	0	[58]
<i>dbx-1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[8]
<i>dbx-2</i>	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[8]
<i>en1</i>	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	[59]
<i>evx1</i>	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	[8]
<i>evx2</i>	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	[8]
<i>gata2</i>	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	[60]
<i>gata3</i>	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	[60]
<i>gbx1</i>	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	[20]
<i>gli1</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	[61]
<i>gli2</i>	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	[61]
<i>gli3</i>	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	1	1	[61]
<i>gsx1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[56]
<i>gsx2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[56]
<i>hoxc9</i>	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	[50]
<i>irx3</i>	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[10]
<i>isl-1</i>	0	0	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0	[62]
<i>isl2</i>	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	[62]
<i>lhx1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	[13]
<i>lhx2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[63]
<i>lhx3</i>	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	[63]
<i>lhx5</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	[63]
<i>lhx9</i>	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	[63]
<i>lhx1a</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[13]
<i>mafb</i>	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	[61]
<i>mnx1</i>	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	[61]
<i>msx1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[47]
<i>msx2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[47]
<i>neurog1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[61]
<i>neurog2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[61]
<i>nkx6-1</i>	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	[8]
<i>nkx6-2</i>	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	[8]
<i>olig3</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[61]
<i>pax2</i>	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	[61]
<i>pax3</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[61]
<i>pax6</i>	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[61]
<i>pax7</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	[54]
<i>pou4f1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	[64]
<i>prrx1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	[24]
<i>sim1</i>	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	[61]
<i>tal1</i>	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	[60]
<i>tlx1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[8]
<i>tlx3</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	[8]
<i>vsx2</i>	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	[61]
<i>zic1</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[65]
<i>zic2</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[61]
<i>zic4</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	[66]

Tabela 2 – Padrões de expressão dos genes selecionados dentro dos domínios da medula.

6.5.2.2.1. – Estabelecendo relações entre os genes selecionados com base em seu padrão de expressão tecidual.

As informações organizadas nas tabelas 2 e 3 foram utilizadas como entrada para um algoritmo de agrupamento dedicado para a determinação de genes com padrões de expressão similares agrupando-os em ramos proximais de uma árvore solução. Tais algoritmos fazem uso de abordagens estatísticas e computacionais para a determinação de elementos similares. O pacote R *pvclust* foi utilizado para a construção da árvore o agrupamento de genes utilizando a distância euclidiana como métrica de dissimilaridade e um total de dez mil reamostragens bootstrap para a determinação da sustentabilidade dos ramos[7]. O resultado dessa comparação está graficamente representado na figura 1, aonde podem ser observados grupos de genes: porém não foi possível observar agrupamentos claros entre genes com expressão tipicamente dorsal, genes com expressão tipicamente ventral e genes com expressão ubíqua (tanto dorsal como ventral). É interessante notar que genes paralogos, como os genes *Lhx3*, *Lhx5*, *Lhx1* e *Lhx2* têm padrões de expressão espacial diferentes entre si. Embora esses genes sejam resultados de uma duplicação de um gene ancestral, esses podem ser subdivididos em genes dorsais (*Lhx1* e *Lhx5*) e ventrais (*Lhx3* e *Lhx2*). Essa árvore solução pode ser interpretada como uma representação do atual conhecimento que temos para expressão desses genes e, por isso, será utilizada para a avaliação de nossos agrupamentos baseados apenas nos sinais transcricionais preditos em seqüências regulatórias conservadas.

Figura 2 - Árvore solução obtida a partir do agrupamento de genes com base em seu padrão de expressão.

Os valores em vermelho se referem aos valores de bootstrap não enviesado.

De modo complementar construímos uma representação heatmap (Mapa de calor) representado as informações da tabela 2. Esse tipo de gráfico é largamente utilizado em análises globais de expressão, como Microarrays. Tal representação faz uso de um algoritmo de agrupamento bidimensional aonde são ao mesmo tempo agrupados: genes e os tecidos da medula embrionária. A sua análise permitiu definir um grupo de cinco genes (Msx1, Msx2, Zic1, Zic4 e Pax3) que apresentaram praticamente os mesmos padrões de expressão espacial. A forte relação entre esses genes pode denotar sinais transcricionais compartilhados por esses elementos que permitiram a correção entre os mesmo. A análise dos resultados também determinou os mesmos agrupamentos observados na figura 1 (agrupamento vertical da figura). Tais agrupamentos não apresentam uma clara separação entre os genes analisados.

Figura 3 – Mapa de calor (heatmap) com a representação das informações organizadas na tabela 2.

Retângulos laranja - escuros representam tecidos (colunas) e genes (linhas) já tiveram sua expressão confirmada.

Retângulos laranja - claros determinam a ausência de evidências da expressão do gene no tecido de interesse.

Retângulos brancos representam genes com expressão ampla e portanto presentes em todos os tecidos analisados..

Com o objetivo de analisar os dados e minerar por possíveis relações entre os genes foi construída uma nova tabela de características aonde os resultados de expressão dos genes nas células neurais pós-mitóticas foram agrupados diretamente com os resultados dos respectivos domínios progenitores. Portanto a evidência de expressão de um gene pode ser observada em qualquer um dos tecidos intimamente relacionados. Esse procedimento visa permitir uma aprofundada mineração de dados e uma melhor separação entre genes de padrão primariamente dorsal ou ventral de expressão dentro da medula embrionária. A matriz resultante foi utilizada com entrada para o algoritmo de agrupamento Pvcust e os resultados podem ser observados na figura 4.

Figura 4 – Agrupamento de genes baseados em seus padrões de expressão com base em seu padrão dorso-ventral de expressão. Nesse agrupamento neurônios pós mitóticos foram agrupados à seus domínios progenitores. Os valores em vermelho se referem aos valores de bootstrap não enviesado.

O agrupamento dos dados de expressão espacial desses genes permitiu a construção de uma árvore solução composta por três grandes ramos. O ramo da direita organiza 13 genes (do gene *Cux2* a *Cux1*), todos com expressão ampla (dorsal ou ventral) ou exclusivamente dorsal. O ramo da esquerda é subdividido em outros dois ramos: o ramo esquerdo composto por 15 genes (do gene *Pax2* a *Gata3*) e o ramo direito composto por 18 genes (do gene *Lbx1* a *Gsx2*). Esses dois sub-ramos são compostos em sua maioria por genes de expressão ventral e dorsal respectivamente. Esse fato será mais bem ilustrado com uma representação heatmap com os mesmos padrões de expressão agrupados na figura 4 (figura 5). Como esperado essa representação ilustra muito bem esses padrões e destaca genes que apresentam padrões de expressão muito parecidos entre si. Essa grande semelhança entre esses padrões pode denotar eventos de coregulação por outros fatores de transcrição agindo para a expressão conjunta desses genes que por sua vez transcrevem para outros fatores de transcrição.

Figura 5 – Mapa de calor (heatmap) com a representação das informações organizadas na tabela 2 agrupadas a partir da expressão espacial desses genes no eixo dorso-ventral da medula embrionária. Retângulos laranja - escuros representam tecidos (colunas) e genes (linhas) já tiveram sua expressão confirmada. Retângulos laranja - claros determinam a ausência de evidências da expressão do gene no tecido de interesse. Retângulos brancos representam genes com expressão ampla e portanto presentes em todos os tecidos analisados.

De modo complementar as informações organizadas na tabela 2 foram novamente agrupadas em apenas cinco classes de expressão: expressão em neurônios pós-mitóticos (Neuron), expressão em domínios progenitores (PROG), na placa tectal (RP), na placa basal (FP) e região progenitora de células gliais (DRG). Esse agrupamento tenta minerar novas informações a partir dos dados de expressão e uma nova tabela com as evidências de expressão para esse conjunto de genes foi construída. Essa tabela foi utilizada com entrada para o algoritmo de agrupamento PvClust e os resultados podem ser consultados na figura 6. Essa significativa diminuição nas informações de expressão espacial levou a

construção de uma árvore solução aonde podem ser observados dois grandes grupos de genes com expressão tipicamente em neurônios ou em células progenitoras. Essa tabela com as informações agrupadas foi utilizada para a construção de um mapa de calor (figura 7) cujos resultados demonstram a separação entre os genes.

Figura 6 – Agrupamento de genes baseados em seus padrões de expressão com base em seu padrão expressão com base na expressão desses genes nas cinco classes de tecidos da medula embrionária.. Nesse agrupamento neurônios pós mitóticos foram agrupados à seus domínios progenitores. Os valores em vermelho se referem aos valores de bootstrap não enviesado.

Figura 7 – Mapa de calor (heatmap) com a representação das informações organizadas na tabela 2 agrupadas a partir da expressão espacial desses genes com base na expressão nas cinco classes de tecidos da medula embrionária. Retângulos laranja - escuros representam tecidos (colunas) e genes (linhas) já tiveram sua expressão confirmada. Retângulos laranja - claros determinam a ausência de evidências da expressão do gene no tecido de interesse. Retângulos brancos representam genes com expressão ampla e portanto presentes em todos os tecidos analisados.

Como pode ser observado na Tabela 2 e nos gráficos construídos, existem genes que são expressos em domínios específicos da medula embrionária, tais com o mnx1, em moto-neurônios e lmx1a, presente apenas na placa tectal. Existem genes expressos apenas no território dorsal ou ventral da medula, bem como genes cuja expressão já foi observada em ambas as regiões. Foi possível observar que existem genes que são expressos apenas nas

linhagens precursoras e aqueles expressos somente em sub-populações pós-mitóticas de neurônios. Esse evento de regulação da expressão espacial dos genes é fortemente associado aos eventos regulatórios estruturais relacionados, como a ligação de fatores de transcrição a seus respectivos sítios. Tendo esses padrões de expressão, várias análises foram conduzidas, através das estruturas regulatórias preditas computacionalmente, as quais serão posteriormente agrupadas por algoritmos por nós desenvolvidos.

6.5.2.3 - Agrupamento de genes baseando-se unicamente em sítios de ligação conservados em suas regiões regulatórias.

Uma vez tendo determinado uma árvore solução que acomode os padrões de expressão individuais de cada um dos genes, nós começamos a construção de uma abordagem de agrupamento que: permita a detecção de sítios de alto valor biológico e o seu posterior agrupamento computacional. Os agrupamentos resultantes serão comparados à luz das figuras (3 a 7) que representam nosso atual conhecimento sobre o comportamento de expressão desses genes no fenótipo sob investigação.

Figura 8- Fluxograma Guia – Comparação entre a ferramenta Match e nosso algoritmo em prever TFBS de alto valor biológico.

6.5.2.3.1 - Determinação de blocos conservados flanqueando esses genes.

A predição computacional de TFBS é prejudicada pela baixa sensibilidade dos atuais algoritmos e uma solução é determinar *a priori* seqüências conservadas que apresentam uma maior possibilidade de apresentarem tais sítios. Foram determinados blocos conservados de seqüências localizados dentro de uma janela de 30 Kb a montante e a jusante da montagem referência (RefSeq) desses genes na atual montagem do genoma do camundongo. A determinação do uso de blocos presentes nessas grandes regiões deve ao fato que nessa região se concentram a grande maioria de sítios de ligação validados funcionalmente. Tais seqüências são ditas conservadas por apresentarem alinhamentos significantes com trechos presentes nas montagens atuais dos genomas: humano, rato e galinha. A escolha da combinação de alinhamentos é um processo arbitrário e nós nos utilizamos do seguinte critério: a presença de um genoma muito próximo (Rato) que validaria predições no genoma do camundongo e um genoma mais distante (Galinha) cujos alinhamentos seriam filtros filogenéticos eficientes. O uso do genoma humano se deve a permitir aproximar os resultados preditos no genoma do murino aos padrões regulatórios encontrados em nosso genoma. Devido a já conhecida alta conservação entre exons de genes ortólogos, seqüências conservadas que sobrepunham exons foram excluídas. O mapeamento desses exons foi feito com base no mapa de exons da montagem referência desses genes. Somente foram selecionados blocos de seqüências conservadas com tamanho igual ou maior de cinco pares de base, esse critério de seleção visa eliminar seqüências conservadas muito fragmentadas e curtas com pouco valor biológico. A figura 9 traz um bloco determinado conservado identificado em uma seqüência à jusante na montagem referência do gene *neurog2*. Cada um dos blocos identificados no genoma do camundongo, a exemplo do bloco apresentado na figura 3, apresentam um significativo alinhamento contra os genomas das três espécies. Esse procedimento permite a construção de eficientes filtros filogenéticos para a identificação de seqüências com máxima probabilidade de apresentarem um elemento funcional.

Figura 9 – Bloco conservado próximo ao gene *Neurog2*. Sequências determinadas conservadas são organizadas no formato FASTA padrão.

Com o objetivo de entender as principais características desses blocos conservados, foram realizadas algumas análises estatísticas descritivas. A Tabela 3 organiza as informações obtidas a partir da identificação desses blocos, onde estão organizados o número de blocos e o tamanho total desses blocos em cada um dos genes analisados. A análise das informações dessa tabela destaca a grande variabilidade existente entre o tamanho dos blocos identificados em cada um dos genes. Foram identificados genes que apresentavam desde pequenos trechos conservados de menos de 500 pares de base a genes com mais de 300 mil pares de base compartilhados com outras espécies. Essa grande variabilidade podendo levar a grandes distorções no número de sítios possivelmente preditos em tais sequências.

Simbolo	Num	Tam	Simbolo	Num	Tam
<i>Tlx1</i>	3	495	<i>Zic2</i>	19	3367
<i>Lhx9</i>	4	532	<i>Cyp26a1</i>	25	3402
<i>Isl2</i>	9	898	<i>Neurog2</i>	15	3620
<i>Gsh1</i>	6	908	<i>Barhl2</i>	12	3880
<i>Gdf7</i>	6	952	<i>Dlx1</i>	15	4156
<i>Sox9</i>	7	1002	<i>Hoxc9</i>	20	4169
<i>Gata2</i>	9	1042	<i>Barhl1</i>	25	4248
<i>Evx1</i>	4	1045	<i>Pou4f1</i>	19	4502
<i>Lhx2</i>	4	1071	<i>Zic1</i>	15	4596
<i>Bmp4</i>	4	1162	<i>Raldh2</i>	21	4750
<i>Ascl1</i>	5	1174	<i>Pax6</i>	14	4799
<i>Sox8</i>	10	1223	<i>Zic4</i>	18	5355
<i>Neurog1</i>	7	1266	<i>Fgf8</i>	23	5520
<i>Isl1</i>	9	1306	<i>Nkx6-1</i>	25	5658
<i>Msx3</i>	8	1416	<i>Msx2</i>	18	6019
<i>Lbx1</i>	10	1462	<i>Evx2</i>	22	6405
<i>En1</i>	6	1463	<i>Gbx1</i>	37	6412
<i>Matf</i>	7	1501	<i>Pax2</i>	34	6655
<i>Msx1</i>	7	1504	<i>Gli3</i>	29	7778
<i>Lmx1a</i>	8	1506	<i>Cutl2</i>	50	10214
<i>Bmp2</i>	4	1531	<i>Sim1</i>	28	10448
<i>Foxd3</i>	8	1604	<i>Prrxl1</i>	31	11942
<i>Olig3</i>	7	1638	<i>Bmp7</i>	73	13380
<i>Bmp6</i>	14	1729	<i>Atp8b2</i>	95	13571
<i>Hlxb9</i>	8	1730	<i>Gata3</i>	70	14017
<i>Tal1</i>	8	1735	<i>Cdx1</i>	54	14431
<i>Dbx2</i>	8	1797	<i>Wnt3a</i>	54	16572
<i>Nkx6-2</i>	14	1950	<i>Irx3</i>	78	23497
<i>Wnt1</i>	12	2033	<i>Sema3c</i>	119	33165
<i>Dbx1</i>	9	2112	<i>Cntrn2</i>	136	36417
<i>Lhx5</i>	13	2422	<i>Gli1</i>	133	38999
<i>Tcfap2a</i>	11	2454	<i>Pax7</i>	240	63077
<i>Wnt6</i>	11	2609	<i>Gli2</i>	243	68720
<i>Gsh2</i>	14	2683	<i>Pax3</i>	210	75810
<i>Lhx1</i>	14	3027	<i>Notch1</i>	256	94171
<i>Chx10</i>	20	3133	<i>Scube2</i>	396	100873
<i>Tlx3</i>	14	3149	<i>Dcc</i>	689	226767
<i>Ets1</i>	20	3263	<i>Cutl1</i>	1360	323638
<i>Lhx3</i>	25	3359			

Tabela 3 – Conjunto de blocos conservados identificados em cada um dos genes analisados. A coluna “símbolo” traz o símbolo oficial de cada um dos genes, a coluna “Num” traz o número de blocos identificados e a coluna “Tam” o tamanho total dos blocos identificados.

A tabela 4 traz estatísticas descritivas do conjunto completo de blocos conservados identificados próximos a esses genes. A análise inicial de tais estatísticas determinam que em geral os blocos conservados têm em média 265 pares de base de tamanho e proporcionalmente pouco dispersos em relação a sua média. Essa pequena dispersão é evidenciada pelos valores de primeiro quartil, terceiro quartil e desvio padrão que são muito próximos ao valor da média.

Estatística Resumo	Valor
Mínimo	5
1st Quartil	104
Mediana	191
Media	265
3rd Quartil	334
Máximo	2501
Desvio Padrão	258

Tabela 4 - Medidas resumo do conjunto completo de blocos sintênicos identificados próximos aos genes envolvidos com desenvolvimento neurológico

Com o objetivo de melhor caracterizarmos a distribuição dessa variável nós nos utilizamos de representações Boxplot (Gráficos de Tuckey) e histogramas (Figura 10). O gráfico 10A traz um conjunto de boxplots representando a distribuição dos tamanhos de blocos identificados nos genes analisados. Cada um dos boxplots da figura determina a dispersão do tamanho dos blocos de um gene em relação a seus quartis, quando menores os boxplots menos dispersos são os tamanhos dos blocos e, portanto os blocos sintênicos conservados identificados em um gene têm aproximadamente o mesmo tamanho. Essa análise, embora preliminar, destacou que esse grupo de genes, proporcionalmente grande, apresentou blocos conservados aproximadamente do mesmo tamanho. O gráfico 10B traz um histograma que representa a distribuição de tamanhos do conjunto completo de blocos identificados nos 57 genes sob análise e destaca que a grande maioria dos blocos identificados nesses genes tem entre 50 e 500 pares de base. Pouquíssimos blocos tinham tamanho superior a mil pares de base. Os genes, seus respectivos blocos conservados identificados e as seqüências ortólogas conservadas de homens, ratos e galinhas foram organizados localmente. Essas seqüências serão utilizadas para a posterior identificação computacional de sítios de ligação de fatores de transcrição de alto valor biológico.

Figura 10 – Tamanho dos blocos sintêmicos conservados - A-) O gráfico A traz Boxplots (Gráficos de Tuckey) que representam o conjunto dos blocos identificados em cada um dos genes de interesse. B-) O gráfico B traz um histograma do conjunto completo de blocos identificados

6.5.2.4. – Predição computacional de sítios para fatores de transcrição nos blocos conservados flanqueando genes de interesse.

6.5.4.2.1 – Predição de sítios usando a ferramenta pública Match®

A ferramenta Match é disponibilizada publicamente no domínio (http://www.gene-regulation.com/cgi-bin/pub/programs/match_bin/match.cgi) e faz uso, em sua versão pública, das matrizes monoposicionais presentes na base de dados Transfac 7.0. Com o objetivo de compararmos a eficiência de nossa abordagem proposta, foram preditos computacionalmente sítios de ligação para fatores de transcrição. Tais sítios foram preditos utilizando como entrada os blocos conservados no genoma do camundongo, determinados no item anterior e foi utilizado como parâmetro de sensibilidade o corte de detecção 0,9 em relação à matriz consenso. O conjunto completo de sítios preditos nos blocos conservados de cada um dos genes foi organizado localmente. A tabela 5 destaca o conjunto completo de sítios para fatores de transcrição preditos no bloco conservado próximo ao gene *neurog2* (destacado na figura 2). Foram preditos 22 sítios nos quase 300 pares de base desse bloco próximo ao gene *neurog2*, aproximadamente 1 sítios a cada 13 pares de base. Esse proporcionalmente pequeno número de sítios se deve, entre outras razões, ao algoritmo de limpeza interna do algoritmo MATCH que elimina sítios sobrepostos menos representativos. Dentre os sítios preditos destacam-se os sítios para o fator Barbie box (2 sítios) e os oito sítios para fatores da família GATA (GATA-1, GATA-2, GATA-3 e GATA-X) que são intimamente relacionados com o desenvolvimento neurológico.

Matrix ID	Pos	Strand	Core	Matrix	Sequence	Factor
V\$COUP_01	31	(-)	0,988	0,901	catcacaaaGTTCA	COUP-TF
V\$AHRARNT_01	122	(-)	1	0,923	ctaCACGCactaggcc	AhR/Arnt
V\$GFI1_01	151	(-)	1	0,956	aggcttggcgaTGATTttaaataca	Gfi-1
V\$BRN2_01	160	(-)	1	0,919	gatgATTTTaaatcac	Brn-2
V\$BRN2_01	161	(+)	0,993	0,905	atgattTAAATcact	Brn-2
V\$GFI1_01	162	(+)	1	0,97	tgattttaAATCActgggctccca	Gfi-1
V\$GATA3_03	167	(-)	0,968	0,916	ttAAATCact	GATA-3
V\$USF_Q6	171	(+)	0,945	0,913	aTCACTgggc	USF
V\$GATA1_02	187	(+)	0,993	0,947	ttctTGATAacagc	GATA-1
V\$GATA1_03	187	(+)	1	0,961	ttctTGATAacagc	GATA-1
V\$GATA1_05	189	(+)	1	0,978	cttGATAAca	GATA-1
V\$GATA1_06	189	(+)	1	0,944	cttGATAAca	GATA-1
V\$GATA2_02	189	(+)	1	0,966	cttGATAAca	GATA-2
V\$GATA2_03	189	(+)	1	0,928	cttGATAAca	GATA-2
V\$LMO2COM_02	190	(+)	0,985	0,94	tTGATAAca	Lmo2
V\$GATA_C	191	(+)	1	0,943	tGATAAcagcc	GATA-X
V\$BARBIE_01	197	(-)	1	0,903	cagcccGCTTTgagc	Barbie Box
V\$BARBIE_01	204	(-)	1	0,907	cttgaGCTTTgtt	Barbie Box
V\$CHOP_01	228	(-)	1	0,915	taggaTTGCAcac	CHOP-C/EBPalpha
V\$ER_Q6	257	(+)	1	0,927	gatgaacagttTGACCctt	ER
V\$AHRARNT_01	282	(-)	1	0,943	gttCACGCtatggaca	AhR/Arnt

Tabela 5 - Predições de TFBS usando o algoritmo MATCH no bloco conservado destacado na figura 9

6.5.2.4.2 – Contagem absoluta de sítios preditos nos blocos conservados

O conjunto completo de sítios preditos nos blocos conservados de um gene foi utilizado como entrada para um simples algoritmo Perl de contagem que determinou o número de sítios preditos para cada um dos fatores de transcrição preditos nessas seqüências sintênicas. Genes que não apresentaram blocos conservados ou que apresentavam blocos aonde não foram preditos sítios de ligação foram eliminados da análise posterior. A partir dessas contagens foi construída uma matriz de características (Gene X Sítios preditos em suas regiões regulatórias) que descreve computacionalmente os principais eventos transcricionais que agem nesse grupo de genes. A tabela 6 traz um pequeno trecho da matriz construída, nela podemos observar os genes (nas linhas) e o número de sítios preditos nas regiões regulatórias desses genes para cada um dos fatores (nas colunas).

GENE FATOR	V\$AHRARNT_01	V\$AHRARNT_02	V\$AHR_01	V\$AP1_C	V\$AP1_Q2	V\$AP1_Q4	V\$AP4_01	V\$AREB6_01
Ascl1	1	0	0	3	9	6	0	1
Atoh1	2	1	0	0	3	2	0	0
Barhl1	3	0	0	0	18	4	0	3
Barhl2	13	0	0	2	38	18	0	2
Cdx1	1	0	0	4	25	16	1	12
Cux1	7	0	0	22	116	68	3	41
Cux2	9	1	1	2	17	7	1	20
Dbx1	4	0	0	5	20	11	0	9
Dbx2	1	0	0	3	8	5	0	8
En1	2	0	0	0	7	4	0	4
Ets1	1	0	0	4	13	9	1	4
Evx1	1	0	0	3	7	3	0	8
Evx2	9	0	0	4	37	18	0	18
Gata2	2	0	0	0	12	6	1	1
Gata3	2	0	0	5	21	12	0	7

Tabela 6 - Matriz de características - As contagens de sítios preditos para cada um dos fatores presentes nos trechos conservados são determinadas a partir dos dados dos sítios preditos com o auxílio da ferramenta MATCH®.

6.5.2.4.2. - Agrupamento de genes com base na matriz de características

A matriz de características construída no item anterior organiza as informações preditas computacionalmente em trechos fortemente conservados de genes de interesse. Essa valiosa fonte de informação é usada como entrada para algoritmos de agrupamento que visam, em última análise, determinar grupos de genes com semelhança entre suas contagens de sítios preditos e separa-los de genes muito dispares com relação a essa fonte de informação. As principais limitações impostas a esses algoritmos são informações insuficientes e a complexidade computacional para a determinação de uma árvore solução que acomode as características observadas nas matrizes de características. Esses problemas são satisfatoriamente superados ou minimizados pelo uso de reamostragens aplicadas aos dados que permitem, em última análise, determinar a confiabilidade dos ramos de um dendrograma. Com esse objetivo nós utilizamos o pacote R dedicado a análises de agrupamento PvcLust [7]. Esse pacote faz uso de duas diferentes abordagens para a determinação da sustentabilidade dos ramos de uma árvore solução: a primeira é a abordagem *bootstrap* padrão proposta por Felsenstein [67] e a segunda, a abordagem *bootstrap* não enviesada proposta por Shimodaira et al[6].

A figura 12 traz a árvore resultante obtida através de 10 mil reamostragens nos dados organizados na matriz de características determinada a partir das predições computacionais de TFBS geradas com auxílio da ferramenta MATCH. Quando a árvore foi confrontada com resultados funcionais pouca correlação foi encontrada, genes com expressão tipicamente ventral, como *Evx1* e *Lhx3* foram separados em dois ramos

6.5.2.4.4. – Sítios preditos agrupados em famílias e agrupamento usando a matriz de características resultante

Fatores de transcrição, como todos os genes conhecidos, são originados a partir de duplicações e especializações de versões ancestrais dos mesmos. Devido a esse critério evolutivo presente, fatores de transcrição presentes na base pública Transfac 7.0 foram classificados em famílias de fatores intimamente relacionados (Essa classificação é mais bem descrita no item 5.1.1.dos métodos do presente texto, página 51). Usando essa classificação os fatores individuais presentes na matriz de característica determinada no item 6.5.2.4.1. foram classificados e as contagens absolutas para cada uma das famílias foram obtidas. A tabela 7 traz uma representação parcial dessa nova matriz de características. Um detalhe que pode ser destacado dessa tabela é o grande número de predições observadas nas regiões regulatórias do gene Cux1 o que evidencia a existência de grandes blocos conservados próximos a esse gene.

GENE FAMILY	AP1	AP2	AREB	ARID	BARBIE	CDX	CP2	CREB/ATF
Ascl1	18	0	5	0	0	0	0	0
Atoh1	5	0	0	0	1	0	0	0
Barhl1	22	0	10	0	1	0	0	10
Barhl2	58	0	9	0	9	0	2	7
Cdx1	45	0	37	0	8	0	1	8
Cux1	206	0	108	0	15	0	8	30
Cux2	26	0	60	0	6	0	4	6
Dbx1	36	0	26	0	7	0	1	3
Dbx2	16	0	17	0	1	0	0	2
En1	11	0	5	0	1	0	0	1
Ets1	26	0	8	0	1	0	2	0
Evx1	13	0	19	0	5	0	0	0
Evx2	59	0	35	0	9	0	4	25
Gata2	18	0	11	0	2	0	1	0
Gata3	38	0	12	0	7	0	0	6

Tabela 7 - Matriz de características - Contagem absoluta de sítios preditos para diferentes famílias de fatores de transcrição em cada um dos genes de interesse

A matriz completa de características composta por sítios de fatores classificados em famílias de sítios, exemplificada na tabela 6, foi utilizada como entrada para o algoritmo dedicado de agrupamento PvClust. O agrupamento desses dados foi realizado utilizando como métrica de similaridade a correlação de Pearson entre as contagens e a sustentabilidade dos ramos foi determinada a partir de 10 mil reamostragens. A árvore solução está apresentada na figura 12 e destaca uma melhor sustentabilidade entre os

ramos da árvore, o que pode indicar uma melhor “resolução” dos padrões transcricionais desses genes. Porém, uma outra indicação de que nem todos os sítios de valor biológico foram preditos é o não agrupamento de genes paralogos em ramos finais da árvore. Esses genes são duplicações de genes ancestrais e espera-se que ainda compartilhem alguns sinais transcricionais e esses provavelmente não foram preditos com o uso da ferramenta MATCH. Genes paralogos como os fatores da família LIM (*Lhx1*, *Lhx2*, *Lhx5* e *Lhx9*) ou genes *Dbx* (*Dbx1* e *Dbx2*) foram separados nas dicotomias finais da árvore solução analisada. Diferentemente da árvore de expressão (figura 1), o gene *Lhx2* foi agrupado com o gene *Lhx5*, sendo que ambos têm padrão díspar de expressão localizada na medula embrionária. Novamente, a exemplo da figura 11, os genes paralogos *neurog1* e *neurog2*, genes com expressão espacial idêntica, foram separados em ramos diferentes da árvore solução. Esses resultados podem indicar que a ferramenta MATCH® apresentou uma relação sinal/ruído desfavorável, o que levou a predição de sítios de baixo valor biológico ou a ausência de sítios que possam explicar melhor os padrões de expressão espacial desses genes. Esses resultados serão confrontados com os obtidos a partir do uso da nossa abordagem para a seleção de sítios de alto valor biológico.

Figura 13 - Árvore solução resultante do agrupamento das informações contidas na matriz de características de famílias de fatores de transcrição. Sítios preditos com o auxílio da ferramenta MATCH. Os valores em vermelho se referem aos valores de bootstrap não enviesado.

6.5.2.4.5. - Predição de TFBS nas regiões regulatórias desses genes usando o algoritmo busca de TFBS desenvolvido localmente

Blocos conservados flanqueando genes de interesse identificados no item 2 foram utilizados como entrada para o nosso algoritmo de detecção de sítios de ligação de fatores de transcrição. Esse algoritmo implementa livremente o algoritmo de detecção da ferramenta MATCH, mas não faz uso de algoritmos de limpeza de sítios parcial ou totalmente sobrepostos. Tais blocos conservados delimitam seqüências compartilhadas pelos seguintes genomas: humano, rato, camundongo e galinha. As seqüências ortólogas desses quatro organismos identificadas em nossos bancos de dados de sintênia são utilizadas como entrada para o algoritmo de detecção de sítios de ligação para fatores de transcrição usando o mesmo corte de detecção (0,9) da ferramenta MATCH (Esse

procedimento é descrito com mais detalhes na seção “Métodos”, item 5.1.5., página 60). O conjunto completo de predições foi organizado localmente aonde todos os sítios possíveis de serem preditos, sobrepostos ou não, foram identificados e organizados localmente. Com o objetivo de permitir um detalhado dos padrões transcricionais encriptados nos blocos conservados desses genes foram preditos sítios tanto na fita molde como complementar.

6.5.2.4.6. – Determinação de sítios de alto valor biológico e construção de uma matriz de características com as famílias de fatores de transcrição

Cada um dos blocos conservados identificados no genoma do camundongo e as seqüências ortólogas das outras espécies foram alinhadas como o auxílio do algoritmo dedicado ClustalX [3].. Um algoritmo Perl especializado em mineração de dados foi desenvolvido para a determinação de sítios para famílias de fatores de transcrição presentes em uma mesma posição relativa, como base no alinhamento múltiplo das seqüências ortólogas. Sítios conservados foram identificados e organizados localmente (esse processo de seleção de sítios é mais bem explicado no item 5.1.5. do presente texto, página 60). Sítios de alto valor biológico para as famílias de fatores de transcrição foram contabilizados, de forma similar as predições MATCH, e uma matriz de características construída, organizando o número de predições para cada uma das famílias no grupo de genes de interesse. A análise inicial da matriz resultante destacou o grande aumento de predições computacionais de nossa abordagem quando comparada às predições da ferramenta MATCH. Isso se deve a ausência do algoritmo de filtragem implementado em MATCH e a aplicação de nossa abordagem para a determinação de sítios de alto valor biológico.

6.5.2.4.7. – Agrupamento da matriz de características construída com a utilização de nosso algoritmo e normalização com base nos tamanho dos blocos sintênicos identificados.

A matriz de características determinada no item 3.2.2. foi normalizada levando-se em conta o tamanho dos blocos conservados identificados nas regiões regulatórias desses genes. A normalização por tamanho dos blocos conservados corrigiu as contagens de sítios para uma família para um fragmento de 1000 pares de base. Para isso foi determinado o tamanho total de todos os blocos sintênicos conservados identificados e as contagens para cada uma das famílias foram normalizadas como se os blocos tivessem

um tamanho total de 1000pb. Como já observado na figura 4A os genes sob análise tem em média tamanhos similares de regiões sintênicas, mas pelos menos 25 % de blocos presentes nesses genes são de tamanho muito grande, variando entre 334 a mais de 2500 pares de base (ver boxplots nas figuras 4A). Essas distorções podem prejudicar consideravelmente as análises de agrupamento subseqüentes, o que dificultaria a mineração de dados a partir de uma árvore solução.

O uso dessa simples correção por tamanho de regiões conservadas permitiu corrigir grandes distorções observadas nas predições brutas realizadas com a ferramenta MATCH entre os genes sob análise. Alguns genes do conjunto apresentavam blocos sintênicos cuja somada ultrapassava os 1000 pb e esses tiveram sua representatividade diminuída e enquanto genes com blocos menores que 1000, tiveram sua representatividade aumentada. Essa normalização pode ser interpretada como uma tentativa de permitir o estudo da proporção de sítios para famílias de fatores de transcrição dentro dos blocos conservados de um gene. A tabela 8 organiza um pequeno trecho da matriz de características normalizada resultante.

GENE FAMILY	AP1	AP2	AREB	ARID	BARBIE	CDX	CP2	CREB/ATF
Ascl1	4,258943782	1,703577513	5,110732538	0	0	3,407155026	0	0
Barhl1	1,177024482	3,29566855	3,531073446	0	0	6,120527307	0	0
Barhl2	2,577319588	2,835051546	3,608247423	0	0,257731959	9,278350515	0	0
Cdx1	1,247314808	2,771810685	2,078858014	0	0	0,623657404	0,069295	0,069295267
Cux1	1,08763495	3,649138853	1,881732059	0	0,018539232	2,576953263	0,00309	0,154493601
Cux2	0,783238692	3,035049931	1,076953201	0	0	2,056001566	0	0,293714509
Dbx1	1,893939394	1,420454545	3,787878788	0	0	4,734848485	0	0
Dbx2	5,008347245	1,112966055	3,338898164	0	0	12,79910963	0	0
En1	0,683526999	3,417634997	1,367053999	0	0,683526999	13,67053999	0	0
Ets1	1,225865768	2,758197977	1,838798652	0	0	2,145265093	0	0
Evx1	0,956937799	0	1,913875598	0	0	2,870813397	0	2,870813397
Evx2	1,8735363	1,717408275	3,278688525	0	0	9,679937549	0	0,156128025
Gata2	0,959692898	1,919385797	1,919385797	0	0	3,838771593	0	0
Gata3	2,140258258	2,568309909	2,782335735	0	0	3,70978098	0	0,214025826

Tabela 8 - Trecho da matriz de características construída a partir de sítios de alto valor biológico definidos com o uso da abordagem de conservação filogenética de famílias de fatores de transcrição

A matriz de características normalizada apenas pelos tamanhos dos blocos foi utilizada como entrada para o algoritmo de agrupamento PvClust. Foram utilizados os mesmos parâmetros de reamostragens e métrica de dissimilaridade e os resultados estão apresentados na figura 13. A análise inicial dessa figura e a sua comparação com a figura12 (construída a partir das predições da ferramenta MATCH®) determinaram que, essas árvores apresentam topologias bem diferentes. A árvore da figura 6 apresentou uma grande quantidade de genes paralogos agrupados em ramos dicotômicos finais,

como, por exemplo, *Zic1* e *Zic4*. Esses dois genes apresentam padrões de expressão idênticos e esse fator ao ser associado à história evolutiva desses genes é uma forte indicação que esses genes compartilham regiões regulatórias similares. Os genes *Barhl1* e *Barhl2*, agrupados anteriormente com o uso da ferramenta MATCH®, foram novamente agrupados, porém em um ramo superior. Um outro destaque dessa árvore é o agrupamento composto pelos genes *Lbx1*, *En1*, *Lhx3* e *Mnx1*, genes com expressão tipicamente ventral. Genes com expressão ampla (ventral e dorsal) na medula embrionária como: *Cdx1*, *Cux1*, *Cux2*, *Gli1* e *Gli2*, foram agrupados de forma sustentada em um ramo dicotômico final. Os genes paralogos *neurog1* e *neurog2*, por sua vez, não foram agrupados em um ramo dicotômico final, mas sim agrupados em um ramo dicotômico mais profundo. Aparentemente a árvore solução gerada a partir das predições com nossa ferramenta têm uma correlação funcional superior a gerada anteriormente. No próximo passo, nós testaremos se uma normalização adicional traria benefícios ao agrupamento gerado ou o tornaria menos específico.

Figura 13 - Árvore solução obtida do agrupamento da matriz de características normalizada com base apenas nos tamanhos dos genes.

6.5.2.4.8. – Determinação da efetividade da normalização z-score para o posterior agrupamento de genes com base nos sinais transcricionais.

A normalização de Z-escore tenta representar as contagens de uma dada família em uma distribuição gaussiana padrão com os valores de média e desvio padrão obtidos diretamente do conjunto de contagens para uma dada família de fatores de transcrição. A normalização z-escore, aplicada diretamente as contagens absolutas das famílias, tenta diminuir as distorções ocasionadas por famílias de fatores cujos elementos são mais comuns do que elementos de outras famílias. Esse simples procedimento coloca, em teoria, as características da matriz (famílias de fatores de transcrição) em igualdade de condições, diminuindo as distorções causadas por famílias constituídas por fatores mais facilmente preditos do que outros (A normalização por Z-escore é mais bem explicada nos métodos do presente texto, item 5.3.2., página 67). Com o objetivo de testar tal procedimento às contagens dos sítios as famílias de fatores organizadas parcialmente na tabela 8 foram novamente normalizadas a partir de uma normalização z-escore. Um pequeno trecho da matriz de característica de ambas as fitas conjuntas com as contagens normalizadas está representado na tabela 9. Após a aplicação dessa normalização as contagens de uma família são representadas pelo seu respectivo z-score dentro da distribuição normal determinada. Esse processo, embora aproxime as contagens das diferentes famílias, pode levar a perda de informações relevantes. Essa perda levaria a construção de árvores solução pouco sustentadas ou distantes da realidade biológica.

GENE FAMILY	AP1	AP2	AREB	ARID	BARBIE	CDX	CP2	CREB/ATF
Ascl1	2,499888	1,374261	0,211766	-0,41412	-0,36766	0,524399	-0,36219	0,380774043
Barhl1	-0,32764	3,41588	-0,25833	NaN	-1,01494	-0,79484	-1,11813	NaN
Barhl2	1,47884	6,102857	0,124568	1,50132	-1,63901	-0,35776	NaN	-0,532221341
Cdx1	-0,19566	-0,38886	NaN	0,171893	-0,56181	-0,753	-0,71712	-0,481000229
Cux1	-0,3193	-0,3903	0,197311	0,754594	0,06752	-0,28815	0,586057	-0,362188081
Cux2	-0,80357	0,84485	-0,27688	-0,25833	NaN	-0,79368	-0,59654	0,16644882
Dbx1	-0,2978	0,458142	-0,36894	1,538602	0,070498	-0,97737	-0,47178	NaN
Dbx2	-0,35273	-0,19566	0,45012	NaN	-0,66865	-0,55841	-0,68265	0,18740258
En1	-0,35325	-0,3193	0,922879	1,131466	-1,42219	0,208693	-0,60313	-0,728903299
Ets1	-0,45388	-0,16864	0,505629	0,970136	-0,25833	NaN	-0,67084	-0,900283788
Evx1	-0,54133	-0,2978	0,508008	-0,36894	-1,36495	0,271646	-1,45474	-0,471783473
Evx2	NaN	-0,35273	-0,19566	-0,50647	NaN	0,13176	-0,69484	0,200250045
Gata2	-0,67135	-0,38406	1,597359	-0,32438	-0,75765	-0,34407	0,266373	-0,603132778
Gata3	0,380023	2,559144	0,570288	0,459056	-0,70797	-0,25833	NaN	-0,363727218

Tabela 9 - Matriz de características com valores normalizados. Os sítios de ligação foram preditos com o auxílio de nossa ferramenta de detecção e filtrados com o uso de nossa abordagem de conservação filogenética de famílias de fatores de transcrição.

A tabela 9 traz um pequeno trecho da matriz de características normalizada por z-score, essa matriz completa foi utilizada como entrada para o algoritmo de alinhamento PvClust. A árvore solução foi determinada a partir do uso dos mesmos parâmetros de reamostragens e similaridade entre elementos e está apresentada na figura 7. A análise dessa figura e a sua comparação com a árvore gerada a partir das contagens não normalizadas destacou uma grande diferença entre as topologias de ambas as árvores. A árvore da figura 15, com normalização z-score, apresentou aumento da distância e diminuição da sustentabilidade entre vários ramos dicotômicos, tanto superiores como inferiores. Essa característica da árvore solução evidencia a perda de informação na matriz de características normalizada e isso levou a gerar incertezas na construção da árvore solução.

Tanto a árvore solução não-normalizada (figura 13) e a árvore normalizada (figura 14) foram comparadas à luz das informações conhecidas sobre a localização da expressão desses genes no desenvolvimento neurológica da medula espinhal de murinos (figuras 2 a 6). Os ramos que organizam genes com expressão ampla (ventral e dorsal) e expressão exclusivamente ventral foram desfeitos e poucos pares de parálogos foram observados nas dicotomias finais da árvore normalizada, exceção feita aos genes *barhl1* e *barhl2* (figura 14). Outros parálogos de expressão similar como *cux1* e *cux2* foram separados de forma sustentada em ramos dicotômicos diferentes. Essa comparação determinou que a árvore solução não normalizada se mostrou mais correlacionada com os padrões de expressão desses genes. Esse resultado sugere que a utilização do normalizador z-score, embora teoricamente bem vinda, levou a perda de importantes informações transcricionais, o que ocasionou a construção de uma árvore com pouca relação com o conhecimento biológico atual.

Figura 14 – Agrupamento de genes com base na matriz de característica normalizada construída a partir de sítios preditos em ambas as fitas.

6.5.2.5. – Avaliação funcional da árvore solução obtida do agrupamento de genes envolvidos no desenvolvimento neurológico de vertebrados.

Tendo definido que a melhor relação sinal/ruído, nós avaliamos funcionalmente os resultados organizados na árvore solução da figura 13. Cada um dos genes agrupados em seus ramos teve determinado o seu padrão de expressão com base nas informações disponibilizadas pela tabela 2 e pelas ilustrações 2 a 7. Os genes foram classificados em três classes: expressão majoritariamente ventral, expressão majoritariamente dorsal e expressão ampla. Os resultados dessa classificação estão apresentados na figura 16.

Figura 15 - Fluxograma guia – Cruzamento funcional da árvore solução e identificação de famílias de fatores determinantes para o agrupamento de dois ramos.

Um interessante detalhe desse gráfico é que o agrupamento de boa parte dos genes com expressão ampla (cor amarela) nos ramos à esquerda na árvore solução. Dentre esses genes destacam-se os genes *cdx1*, *cux1* e *cux2*, todos os genes presentes em todos os domínios progenitores ou neurônios pós-mitóticos no eixo dorso-ventral da medula embrionária. Genes com expressão típica ventral (cor azul) foram agrupados em sua maioria em dois grupos bem estabelecidos: no ramo a esquerda em conjunto com genes de expressão ampla caso dos genes *Vsx2*, *Tal1*, *Evx2* e *Gata2* e no ramo na extrema direita, caso dos genes *Lbx1*, *En1*, *Lhx3* e *Mnx1*. Genes com expressão dorsal formam a maioria dos genes analisados e foram preferencialmente organizados nos ramos centrais da árvore solução. Nossa abordagem de detecção de sítios aparentemente identificou um conjunto de sítios de bom valor biológico cujo agrupamento permitiu, ao menos parcialmente, construir uma árvore solução bem sustentada e que permitiu a separação de genes com padrões de expressão distintos.

Figura 16 – Padrão de expressão dos genes agrupados em cada um dos ramos da árvore solução

Essa separação dos genes guarda muita semelhança com a organização dicotômica proposta pelo agrupamento dos sinais transcricionais preditos nas regiões regulatórias desses genes. Tendo isso em mente, a árvore solução foi subdividida em quatro pequenas sub-árvores respeitando a separação proposta pela árvore solução (figura 17). Cada uma dessas sub-árvores terá os seus elementos inspecionados para o entendimento dos padrões transcricionais responsáveis pelos agrupamentos e por suas separações em relação a outros ramos. Inicialmente foi construída uma tabela de contingência para a determinação se essa separação observada é aleatória ou guarda dependência entre as variáveis (Tabela 10). Inicialmente construímos uma tabela de contingência com as contagens de genes das diferentes classes propostas em cada uma das sub-árvores determinadas. A análise inicial dessa tabela destaca a ausência de genes com padrão dorsal de expressão nas sub-árvores dois e quatro e pequena proporção de genes com expressão ventral na árvore 3.

Figura 17 - Subdivisão proposta a árvore solução- A árvore solução com as informações da expressão espacial desses genes foi subdividida em quatro sub-árvores respeitando a divisão dicotômica proposta pela árvore solução inicial.

As informações organizadas na tabela 10 foram analisadas com um teste qui-quadrado de independência e os resultados sugerem a forte dependência entre as variáveis ($\chi^2 \sim 25,7974$, $Gl=6$, $p.value \leq 0,0002428$). Essa dependência infere que a separação entre os genes das três diferentes classes nas quatro sub-árvores não é aleatória. Isso sugere que tanto nossa abordagem de seleção quanto o agrupamento de genes foram sensíveis o suficiente para a identificação de relações entre genes de interesse no fenótipo sob estudo. Em um próximo passo, nós começamos a minerar os dados gerados para um melhor entendimento dos eventos transcricionais aos quais esses genes estão expostos.

	Dorsal	Ventral	Ampla
Arvore1	1	0	2
Arvore2	0	4	7
Arvore3	23	5	8
Arvore4	0	4	1

Tabela 10 - Tabela de contingência com as contagens de genes para cada uma das três classes propostas na

6.5.2.5.1. – Análise comparativa dos sítios preditos nos elementos organizados na sub-árvore 1.

A partir da matriz de características inicial não normalizada foram criadas duas pequenas matrizes compostas apenas por genes localizados em sub-árvores sob investigação. Essas matrizes organizam somente as contagens de famílias para os elementos de interesse desses sub-agrupamentos. No agrupamento apresentado na figura 18 foi utilizada como métrica de similaridade a correlação de Pearson entre as contagens de famílias dos genes de interesse. A correlação de Pearson é, em última análise, uma medida que determina, no nosso caso, a relação entre as contagens das famílias preditas em ambos os genes com base na média do conjunto completo de contagens dos genes ou grupos analisados. A figura 18 traz a formulação matemática da correlação de Person e fica claro que sendo a correlação um somatório dos desvios entre contagens das famílias em relação a média de contagens, é possível determinar a contribuição de cada uma das famílias de fatores de transcrição para a correlação final entre dois genes ou grupos de genes.

$$\sum \frac{(X_i - \mu_X)(Y_i - \mu_Y)}{\sigma_X \sigma_Y}$$

Figura 18 – Correlação linear de Pearson. Em nosso agrupamento os valor X_i e Y_i representam as contagens da família i no gene X e Y. Os valores μ_X e σ_X referenciam respectivamente a média e o desvio do conjunto completo de contagens do gene X. Os valores μ_Y e σ_Y referenciam respectivamente a média e o desvio do conjunto completo de contagens do gene Y.

Inicialmente foram comparadas as matrizes de características dos elementos organizados na sub-árvore 1 compostas por três ramos independentes com apenas um elemento, respectivamente os genes *Tlx1*, *Lhx9* e *Isl2*. A tabela 11 traz as contagens de sítios preditos para famílias de fatores nesses genes e evidencia a grande semelhança

entre as contagens. A análise da tabela destaca que não foram preditos, em nenhum gene, sítios para 34 das 58 famílias de fatores de transcrição (58%), o que os distanciou desses três elementos do conjunto de genes. Esse exíguo número de predições é totalmente relacionado ao limitado tamanho das regiões conservadas próximas a esses três genes principalmente quando esses genes são confrontados com genes com regiões conservadas muito maiores (Tabela 3). Dentre as ausências de predições destaca-se a ausência de predições de sítios para fatores da família GATA no gene *Lhx9* e de sítios para fatores AP1 nos genes *Tlx1* e *Lhx9*. Ambas as famílias são compostas por fatores cuja predição de sítios é extremamente comum em praticamente todas as seqüências regulatórias testadas dentro do genoma do camundongo.

Os genes *Lhx9* e *Isl2* têm padrões amplos de expressão e a predição de sítios para fatores de transcrição da família CDX, dentre os fatores participantes dessa família estão os fatores de transcrição parálogos Cdx1 e Cdx2 com expressão ampla. O gene *Tlx1* é expresso de maneira localizada somente na região d5 e já foi associado a processos de especialização e manutenção de neurônios sensoriais, responsáveis pela modulação da dor em vertebrados. O gene *Isl2*, por sua vez, apresentou um padrão de expressão localizado em regiões da medula como: as células progenitoras tanto dorsais como ventrais respectivamente d3 e MN, e na região pericussora de células gliais DRGA. Em nossas análises, esse padrão difuso de expressão desses dois genes provavelmente está relacionado aos padrões preditos nas regiões regulatórias de ambos. Esses genes apresentaram padrões tão diferentes que foram separados em ramos exclusivos bem separados do resto do conjunto de genes analisados. O gene *Lhx9* apresentou um padrão de expressão ubíquo, sem um padrão dorsal ou ventral determinante. E a exemplo dos genes *Tlx1* e *Isl2* foi agrupado em um ramo exclusivo distante do resto dos genes.

Família GENE	<i>Tlx1</i>	<i>Lhx9</i>	<i>Isl2</i>
AP1	0	0	1,113585746
AP2	0	3,759398	0
AREB	2,02020202	1,879699	0
ARID	0	0	0
BARBIE	0	0	0
CDX	4,04040404	1,879699	1,113585746
CP2	0	0	0
CREB,ATF	0	0	0
E,BOX	0	5,639098	0
E2F	0	0	0
ELK	0	0	0
ENGRAILED	0	0	0
ETS	0	0	4,454342984
FORKHEAD	2,02020202	0	1,113585746
GATA	2,02020202	0	3,340757238
GCM	0	0	0
HEAT_SHOCK	0	0	0
HETEROCHROMATIN	0	0	0
HIF	0	0	0
HOMEODOMAIN	6,060606061	1,879699	2,227171492
IKAROS	0	0	0
IPT_COE	0	0	0
IRF	0	0	3,340757238
KRUPEL_LIKE	0	0	0
LEUCINE_ZIPPER	0	0	0
LIM_HOMEODOMAIN	0	0	1,113585746
LMO	0	1,879699	0
MADS	0	0	1,113585746
MEIS	0	0	1,113585746
MSX	2,02020202	0	0
MYB	0	0	2,227171492
Major_t_Antigen	0	0	0
Myc,Max	0	0	0
NF,AT	0	0	0
NF_KAPPA	0	0	0
NKX	2,02020202	3,759398	5,567928731
NRSF	0	0	0
NUCLEAR_RECEPTOR	0	0	0
P53	0	0	1,113585746
PDF	0	0	0
PITX	0	0	0
POU	0	0	0
PRIMASE	0	0	0
PRX	0	0	5,567928731
Pax	2,02020202	3,759398	3,340757238
Pbx	0	0	0
RFX	0	0	0
Ring_Finger	0	0	0
SMAD	0	0	0
SREBP	2,02020202	0	0
STAT	2,02020202	3,759398	5,567928731
Skeletal_muscle	0	0	0
Sox	2,02020202	1,879699	2,227171492
TBOX	0	0	0
TCF	2,02020202	0	2,227171492
TEF	0	0	0
ZIC	0	0	0
ZINC	0	0	0

Tabela 11 - Contagens normalizadas de sítios de ligação para fatores de transcrição preditos nas regiões regulatórias dos genes *Tlx1*, *Lhx9* e *Isl2*. Esses genes estão localizados respectivamente nas sub-árvores 1,2 e 3.

Com o objetivo de determinar as contribuições de cada uma das famílias para a separação entre dois elementos individuais ou para grupos de elementos, foi desenvolvido um algoritmo dentro do ambiente estatístico R chamado de “Ranking_family”. Esse algoritmo basicamente organiza em um vetor as parcelas de cada uma das famílias e posteriormente ordena os valores absolutos dessas contribuições parciais e os apresenta em uma lista ordenada. Os valores organizados na tabela 11 foram utilizados como entrada para esse algoritmo comparando os genes *Lhx9* e *Isl2* em comparação ao gene *Tlx1*. As dez famílias de fatores de transcrição mais representativas nessas comparações estão respectivamente organizadas nas figuras 19A (*tlx1* X *Lhx9*) e 19B (*tlx1* X *isl2*). Vale a pena salientar que a ausência de famílias com contribuições negativas para a correlação se deve ao uso de valores absolutos durante o processo de ordenação. Os resultados organizados na figura 12 destacam o interessante fato de que as mesmas quatro famílias (Homeodomain, STAT, NKX e Pax) foram as que mais contribuíram para a separação entre esses ramos. Quando esses resultados foram cruzados com as quantificações organizadas na tabela 11, ficou claro que, os três genes (*Tlx1*, *Lhx9* e *Isl2*) apresentaram predições para as quatro famílias, mas com contagens muito distantes entre si. Quando os genes *tlx1* e *Lhx9* foram comparados ganhou destaque à contribuição da família CDX para a separação desses dois genes cujos elementos constituintes apresentam padrões de expressão similares aos observados para o gene *lhx9*. A presença de sítios para fatores da família TCF está totalmente em contexto com a expressão tipicamente dorsal do gene *Tlx1* e expressão ampla do gene *Isl2*.

Esse tipo de análise estatística deve ser feito com algum cuidado, pois o desvio de contagens entre famílias precisa ser sempre levado em conta durante a análise de dados de agrupamento. Famílias cujos elementos possuem matrizes monoposicionais pouco específicas tendem a se destacar nesse tipo de análise, exemplos das famílias STAT e GATA. A mesma abordagem de comparação foi testada em duas sub-árvores compostas por grupos maiores de genes e que foram agrupadas bem distantes uma da outra. Essa grande distância entre esses agrupamentos pode evidenciar diferentes coleções de sítios preditos em suas regiões regulatórias.

Figura 19 - As 10 famílias que mais contribuíram para a separação entre os ramos da sub-árvore 1. A-) Resultados da comparação entre os genes *tlx1* e *lhx9*. B-) Resultados da comparação entre os genes *tlx1* e *isl2*.

6.2.5.2.2 – Análise das famílias discriminantes e determinantes para os agrupamentos das árvores 3 e 4

As sub-árvores 3 e 4 organizam em seus ramos elementos, como já discutidos anteriormente, que apresentam padrões de expressão majoritariamente dorsal e ventral respectivamente. A sub-árvore 3 é composta por 36 genes, onde se destacam os 23 genes com expressão dorsal. A sub-árvore 4 é composta por 5 genes dos quais quatro apresentam expressão exclusivamente ventral no eixo dorso-ventral da medula embrionária. Essa significativa separação entre os elementos já foi comprovada estatisticamente, mas pouco se sabe sobre quais foram os sinais transcricionais preditos computacionalmente responsáveis por essa separação. Para essa análise, inicialmente, foram separados os dois conjuntos de predições para os genes constituintes de ambas as árvores analisadas. O conjunto de contagens para uma determinada família foi organizado em um vetor e será representado a partir de sua mediana. O uso da mediana em detrimento a média se deve ao seu menor enviesamento para a representação de uma amostra de pequeno tamanho (poucos genes). A tabela 12 traz os valores das medianas para ambas as sub-árvores e a razão entre elas. Os valores Inf referem-se a valores que tendem a um valor infinito devido a divisão por zero e, por isso, não serão analisados. Famílias cujos sítios não foram preditos em nenhum dos elementos de ambas as árvores

também foram excluídas por não adicionarem nenhum tipo de informação a análise proposta.

Família	Grupo3	Grupo4	Razão	Família	Grupo3	Grupo4	Razão
AP1	1,664224845	0,959462701	1,734538	Myc,Max	0,031707278	0	Inf
AP2	1,995349714	0,683994528	2,917201	NF,AT	1,058222283	1,190830604	0,888642
AREB	2,893797469	3,358119453	0,861732	NKX	3,869218166	6,936416185	0,557812
CDX	7,662860241	13,67053999	0,560538	NUCLEAR_RECEPTOR	0,436149326	0,47973135	0,909153
E.BOX	1,929784479	2,050580998	0,941092	P53	0,278830715	0	Inf
ELK	0,05152461	0	Inf	PDF	0,296902112	0	Inf
ETS	5,726005355	3,572491813	1,602804	PITX	0,019786308	0	Inf
FORKHEAD	5,433127429	3,417634997	1,589733	POU	0,522843306	1,367053999	0,38246
GATA	17,70016067	18,46785226	0,958431	PRIMASE	0,207712987	0	Inf
HEAT_SHOCK	0,271244021	0	Inf	PRX	1,992859266	4,787961696	0,416223
HOMEODOMAIN	5,866639888	8,670520231	0,676619	Pax	15,75313368	22,57181943	0,697912
IRF	1,331600687	2,735978112	0,4867	Pbx	0,294249572	0,683526999	0,430487
LEUCINE_ZIPPER	0,062052715	0	Inf	SREBP	0,332603767	0,47973135	0,693313
LIM_HOMEODOMAIN	0,161278072	0	Inf	STAT	15,0813953	15,04787962	1,002227
LMO	1,069871623	0,297707651	3,593699	Sox	5,289199466	6,155950752	0,859201
MADS	0,70876866	2,051983584	0,345407	TCF	1,247714461	1,439194051	0,866954
MEIS	0,441710301	0,578034682	0,764159	TEF	0,935237483	1,367989056	0,683659
MSX	0,188165904	0,239865675	0,784464	ZINC	0,989162209	1,199328376	0,824763
MYB	6,074924409	4,784688995	1,269659				

Tabela 12 - Tabela de contagens - Os valores observados se referem a mediana entre as contagens de sítios para uma mesma família de fatores no genes agrupados em cada uma das árvores

As famílias de fatores organizadas na tabela 12 foram classificadas em família determinantes e discriminantes. As famílias determinantes foram responsáveis diretas pelo agrupamento dos genes constituintes de uma das árvores analisada e as famílias discriminantes foram responsáveis pela separação entre os grupos. Essa classificação foi feita da seguinte forma: famílias discriminantes apresentam razão entre as medidas, para ambos os grupos, superior a 1,5 ou inferior a 0,5 e famílias determinantes apresentam razão entre 0,5 e 1,5 entre as medidas resumo das contagens. A partir dessa classificação foram determinadas discriminantes as seguintes famílias de fatores: AP1, AP2, ETS, FORKHEAD, IRF, LMO, MADS, POU e PRX. É interessantes notar que o baixo número de sítios preditos para 9 famílias de fatores de transcrição nos genes agrupados na árvore 4 levou a mediana das contagens a apresentar um valor nulo, dentre essas famílias destaca-se a família LIM_HOMEODOMAIN cujos fatores se mostram associados à expressão ventral do gene *aldh1a2* [68]. A família LEUCINE_ZIPER, composta por um grande número de fatores de transcrição, teve sítios preditos apenas no único gene que apresentou expressão ampla (*hoxc9*) dentre os genes da sub-árvore 4. As famílias determinantes podem ser interpretadas como pouco informativas para a separação de duas sub-árvores e, portanto, importantes para os agrupamentos internos de cada uma das árvores. Foram descritas como determinantes um conjunto de 16 famílias: AREB, CDX, E.BOX, GATA,

HOMEODOMAIN, MEIS, MYB, NF-AT, NUCLEAR_RECEPTOR, Pax, SREBP, STAT, Sox, TCF, TEF, ZINC. Dentre essas famílias destacam-se as famílias GATA e ZINC cujos elementos são facilmente encontrados em praticamente qualquer região regulatória.

Referências utilizadas nessa secção

1. Do, J.H. and D.K. Choi, *Clustering approaches to identifying gene expression patterns from DNA microarray data*. Mol Cells, 2008. **25**(2): p. 279-88.
2. Wilgenbusch, J.C. and D. Swofford, *Inferring evolutionary trees with PAUP**. Curr Protoc Bioinformatics, 2003. **Chapter 6**: p. Unit 6 4.
3. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX*. Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2 3.
4. Gribble, S.L., O.B. Nikolaus, and R.I. Dorsky, *Regulation and function of Dbx genes in the zebrafish spinal cord*. Dev Dyn, 2007. **236**(12): p. 3472-83.
5. Baser, O., W.H. Crown, and C. Pollicino, *Guidelines for selecting among different types of bootstraps*. Curr Med Res Opin, 2006. **22**(4): p. 799-808.
6. Shimodaira, H., *An approximately unbiased test of phylogenetic tree selection*. Syst Biol, 2002. **51**(3): p. 492-508.
7. Suzuki, R. and H. Shimodaira, *Pvclust: an R package for assessing the uncertainty in hierarchical clustering*. Bioinformatics, 2006. **22**(12): p. 1540-2.
8. Lewis, K.E., *How do genes regulate simple behaviours? Understanding how different neurons in the vertebrate spinal cord are genetically specified*. Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1465): p. 45-66.
9. Briscoe, J. and J. Ericson, *Specification of neuronal fates in the ventral neural tube*. Curr Opin Neurobiol, 2001. **11**(1): p. 43-9.
10. Lewis, K.E. and J.S. Eisen, *From cells to circuits: development of the zebrafish spinal cord*. Prog Neurobiol, 2003. **69**(6): p. 419-49.
11. Karunaratne, A., et al., *GATA proteins identify a novel ventral interneuron subclass in the developing chick spinal cord*. Dev Biol, 2002. **249**(1): p. 30-43.
12. Burrill, J.D., et al., *PAX2 is expressed in multiple spinal cord interneurons, including a population of EN1+ interneurons that require PAX6 for their development*. Development, 1997. **124**(22): p. 4493-503.
13. Gross, M.K., M. Dottori, and M. Goulding, *Lbx1 specifies somatosensory association interneurons in the dorsal spinal cord*. Neuron, 2002. **34**(4): p. 535-49.
14. Dakubo, G.D., et al., *Control of glial precursor cell development in the mouse optic nerve by sonic hedgehog from retinal ganglion cells*. Brain Res, 2008. **1228**: p. 27-42.
15. Aronica, E., et al., *Expression of Id proteins increases in astrocytes in the hippocampus of epileptic rats*. Neuroreport, 2001. **12**(11): p. 2461-5.
16. Blaess, S., D. Stephen, and A.L. Joyner, *Gli3 coordinates three-dimensional patterning and growth of the tectum and cerebellum by integrating Shh and Fgf8 signaling*. Development, 2008. **135**(12): p. 2093-103.
17. Ferreira, R., et al., *Dynamic regulation of Gata factor levels is more important than their identity*. Blood, 2007. **109**(12): p. 5481-90.
18. Minguillon, C. and J. Garcia-Fernandez, *Genesis and evolution of the Evx and Mox genes and the extended Hox and ParaHox gene clusters*. Genome Biol, 2003. **4**(2): p. R12.
19. Li, S., et al., *Barhl1 regulates migration and survival of cerebellar granule cells by controlling expression of the neurotrophin-3 gene*. J Neurosci, 2004. **24**(12): p. 3104-14.
20. Rhinn, M., et al., *Isolation and expression of the homeobox gene Gbx1 during mouse development*. Dev Dyn, 2004. **229**(2): p. 334-9.
21. Lonyai, A., et al., *The promise of Hox11+ stem cells of the spleen for treating autoimmune diseases*. Horm Metab Res, 2008. **40**(2): p. 137-46.
22. Kublaoui, B.M., et al., *Oxytocin deficiency mediates hyperphagic obesity of Sim1*

- haploinsufficient mice*. Mol Endocrinol, 2008. **22**(7): p. 1723-34.
23. Cubelos, B., et al., *Cux-1 and Cux-2 control the development of Reelin expressing cortical interneurons*. Dev Neurobiol, 2008. **68**(7): p. 917-25.
 24. Rebelo, S., et al., *DRG11 immunohistochemical expression during embryonic development in the mouse*. Dev Dyn, 2007. **236**(9): p. 2653-60.
 25. Raum, J.C., et al., *FoxA2, Nkx2.2, and PDX-1 regulate islet beta-cell-specific mafA expression through conserved sequences located between base pairs -8118 and -7750 upstream from the transcription start site*. Mol Cell Biol, 2006. **26**(15): p. 5735-43.
 26. Nelson, S.B., A.E. Schaffer, and M. Sander, *The transcription factors Nkx6.1 and Nkx6.2 possess equivalent activities in promoting beta-cell fate specification in Pdx1+ pancreatic progenitor cells*. Development, 2007. **134**(13): p. 2491-500.
 27. Mizuguchi, R., et al., *Ascl1 and Gsh1/2 control inhibitory and excitatory cell fate in spinal sensory interneurons*. Nat Neurosci, 2006. **9**(6): p. 770-8.
 28. Kriks, S., et al., *Gsh2 is required for the repression of Ngn1 and specification of dorsal interneuron fate in the spinal cord*. Development, 2005. **132**(13): p. 2991-3002.
 29. Zordan, P., et al., *Comparative analysis of proneural gene expression in the embryonic cerebellum*. Dev Dyn, 2008. **237**(6): p. 1726-35.
 30. Michibata, H., et al., *Inhibition of mouse GPM6A expression leads to decreased differentiation of neurons derived from mouse embryonic stem cells*. Stem Cells Dev, 2008. **17**(4): p. 641-51.
 31. Pillai, A., et al., *Lhx1 and Lhx5 maintain the inhibitory-neurotransmitter status of interneurons in the dorsal spinal cord*. Development, 2007. **134**(2): p. 357-66.
 32. Wilson, J.M., A.I. Cowan, and R.M. Brownstone, *Heterogeneous electrotonic coupling and synchronization of rhythmic bursting activity in mouse Hb9 interneurons*. J Neurophysiol, 2007. **98**(4): p. 2370-81.
 33. Hirata, T., et al., *Identification of distinct telencephalic progenitor pools for neuronal diversity in the amygdala*. Nat Neurosci, 2009. **12**(2): p. 141-9.
 34. Huang, M., et al., *Diverse expression patterns of LIM-homeodomain transcription factors (LIM-HDs) in mammalian inner ear development*. Dev Dyn, 2008. **237**(11): p. 3305-12.
 35. Jorgensen, J.S. and L. Gao, *Irx3 is differentially up-regulated in female gonads during sex determination*. Gene Expr Patterns, 2005. **5**(6): p. 756-62.
 36. Mo, Z., et al., *Role of the Barhl2 homeobox gene in the specification of glycinergic amacrine cells*. Development, 2004. **131**(7): p. 1607-18.
 37. Gothert, J.R., et al., *NOTCH1 pathway activation is an early hallmark of SCL T leukemogenesis*. Blood, 2007. **110**(10): p. 3753-62.
 38. Thompson, J.A., et al., *Pax7 is requisite for maintenance of a subpopulation of superior collicular neurons and shows a diverging expression pattern to Pax3 during superior collicular development*. BMC Dev Biol, 2008. **8**: p. 62.
 39. Li, S. and M. Xiang, *Barhl1 is required for maintenance of a large population of neurons in the zonal layer of the superior colliculus*. Dev Dyn, 2006. **235**(8): p. 2260-5.
 40. D'Souza, C.A., et al., *Identification of a set of genes showing regionally enriched expression in the mouse brain*. BMC Neurosci, 2008. **9**: p. 66.
 41. Wilson, S.I., et al., *A molecular program for contralateral trajectory: Rig-1 control by LIM homeodomain transcription factors*. Neuron, 2008. **59**(3): p. 413-24.
 42. Pilon, N., et al., *Wnt signaling is a key mediator of Cdx1 expression in vivo*. Development, 2007. **134**(12): p. 2315-23.
 43. Gerhauser, I., S. Alldinger, and W. Baumgartner, *Ets-1 represents a pivotal transcription factor for viral clearance, inflammation, and demyelination in a mouse model of multiple sclerosis*. J Neuroimmunol, 2007. **188**(1-2): p. 86-94.
 44. Sun, Y., et al., *A central role for Islet1 in sensory neuron development linking sensory and spinal gene regulatory programs*. Nat Neurosci, 2008. **11**(11): p. 1283-93.

45. Hume, C.R., D.L. Bratt, and E.C. Oesterle, *Expression of LHX3 and SOX2 during mouse inner ear development*. Gene Expr Patterns, 2007. **7**(7): p. 798-807.
46. Osumi, N., et al., *Concise review: Pax6 transcription factor contributes to both embryonic and adult neurogenesis as a multifunctional regulator*. Stem Cells, 2008. **26**(7): p. 1663-72.
47. Liu, Y., A.W. Helms, and J.E. Johnson, *Distinct activities of Msx1 and Msx3 in dorsal neural tube development*. Development, 2004. **131**(5): p. 1017-28.
48. Joksimovic, M., et al., *Wnt antagonism of Shh facilitates midbrain floor plate neurogenesis*. Nat Neurosci, 2009. **12**(2): p. 125-31.
49. Inoue, T., et al., *Zic1 and Zic3 regulate medial forebrain development through expansion of neuronal progenitors*. J Neurosci, 2007. **27**(20): p. 5461-73.
50. Dasen, J.S., J.P. Liu, and T.M. Jessell, *Motor neuron columnar fate imposed by sequential phases of Hox-c activity*. Nature, 2003. **425**(6961): p. 926-33.
51. Nishimura, W., et al., *Preferential reduction of beta cells derived from Pax6-MafB pathway in MafB deficient mice*. Dev Biol, 2008. **314**(2): p. 443-56.
52. Goupille, O., et al., *Msx1 and Msx2 are expressed in sub-populations of vascular smooth muscle cells*. Dev Dyn, 2008. **237**(8): p. 2187-94.
53. Burren, K.A., et al., *Gene-environment interactions in the causation of neural tube defects: folate deficiency increases susceptibility conferred by loss of Pax3 function*. Hum Mol Genet, 2008. **17**(23): p. 3675-85.
54. White, R.B. and M.R. Ziman, *Genome-wide discovery of Pax7 target genes during development*. Physiol Genomics, 2008. **33**(1): p. 41-9.
55. Bannerman, P., et al., *Peripheral nerve regeneration is delayed in neuropilin 2-deficient mice*. J Neurosci Res, 2008. **86**(14): p. 3163-9.
56. Kuhn, R.M., et al., *The UCSC Genome Browser Database: update 2009*. Nucleic Acids Res, 2009. **37**(Database issue): p. D755-61.
57. Bulfone, A., et al., *Barhl1, a gene belonging to a new subfamily of mammalian homeobox genes, is expressed in migrating neurons of the CNS*. Hum Mol Genet, 2000. **9**(9): p. 1443-52.
58. Iulianella, A., G. Vanden Heuvel, and P. Trainor, *Dynamic expression of murine Cux2 in craniofacial, limb, urogenital and neuronal primordia*. Gene Expr Patterns, 2003. **3**(5): p. 571-7.
59. Perrier, A.L., et al., *Derivation of midbrain dopamine neurons from human embryonic stem cells*. Proc Natl Acad Sci U S A, 2004. **101**(34): p. 12543-8.
60. Smith, E., et al., *Coexpression of SCL and GATA3 in the V2 interneurons of the developing mouse spinal cord*. Dev Dyn, 2002. **224**(2): p. 231-7.
61. Wilson, L. and M. Maden, *The mechanisms of dorsoventral patterning in the vertebrate neural tube*. Dev Biol, 2005. **282**(1): p. 1-13.
62. Li, Y., et al., *Cloning and expression of a novel human gene, Isl-2, encoded a LIM-homeodomain protein*. Mol Biol Rep, 2007. **34**(1): p. 19-26.
63. Zhadanov, A.B., et al., *Expression pattern of the murine LIM class homeobox gene Lhx3 in subsets of neural and neuroendocrine tissues*. Dev Dyn, 1995. **202**(4): p. 354-64.
64. Mu, X. and W.H. Klein, *A gene regulatory hierarchy for retinal ganglion cell specification and differentiation*. Semin Cell Dev Biol, 2004. **15**(1): p. 115-23.
65. Sun Rhodes, L.S. and C.S. Merzdorf, *The zic1 gene is expressed in chick somites but not in migratory neural crest*. Gene Expr Patterns, 2006. **6**(5): p. 539-45.
66. Gaston-Massuet, C., et al., *Zic4, a zinc-finger transcription factor, is expressed in the developing mouse nervous system*. Dev Dyn, 2005. **233**(3): p. 1110-5.
67. Felsenstein, J., *Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method*. Genet Res, 1992. **60**(3): p. 209-20.
68. Castillo, H.A., et al., *Insights into the organization of dorsal spinal cord pathways from*

- an evolutionarily conserved raldh2 intronic enhancer*. Development. **137**(3): p. 507-18.
69. Krizhanovsky, V., et al., *Math1 target genes are enriched with evolutionarily conserved clustered E-box binding sites*. J Mol Neurosci, 2006. **28**(2): p. 211-29.

Discussão

Discussão

Algoritmos de agrupamento são largamente utilizados em diversas áreas do conhecimento humano e tem se tornado uma efetiva escolha para a mineração de dados. Algoritmos de filogenia formam uma classe especial de algoritmos de agrupamento, aonde seqüências de aminoácidos ou bases de DNA são utilizadas como entrada para a determinação de seus possíveis parentescos evolutivos ou para a anotação de uma nova seqüência de interesse. Tais algoritmos, em geral, usam como entrada um alinhamento múltiplo entre seqüências de interesse e a partir de algoritmos de reconstrução determinam uma árvore solução que melhor acomode as evidências individuais de cada seqüência. No presente trabalho, desenvolvemos uma abordagem de agrupamentos baseando-se unicamente nos sinais computacionalmente preditos em seqüências conservadas próximas a genes de interesse. Tais seqüências conservadas são compartilhadas por espécies evolutivamente distantes como: camundongo, rato, galinha e homem. Esse alto grau de conservação dessas seqüências, a exemplo de exons e introns, pode indicar um evento de seleção positiva para a manutenção das informações presentes nesses blocos sintênicos.

Mesmo com o uso de seqüências conservadas inter-espécie uma grande proporção de sítios de ligação para fatores de transcrição de pouco valor biológico é predita nesses blocos conservados. Isso se deve basicamente à limitação combinatorial imposta pela desfavorável relação entre o tamanho dos padrões preditos (sítios de ligação) e o tamanho dos blocos conservados identificados. Com o objetivo de diminuir a redundância das predições computacionais os sítios de ligação para fatores de transcrição que apresentavam matrizes monoposicionais de detecção na base pública Transfac® 6.2 foram classificados em famílias de fatores intimamente relacionados. Foi desenvolvida e testada uma nova abordagem para a determinação de sítios de altíssimo valor biológico. Foram selecionados unicamente os sítios para fatores de transcrição cuja família de fatores esteja presente em uma mesma posição relativa em todas as seqüências conservadas identificadas anteriormente compartilhadas por espécies de interesse. Com o objetivo de melhorar a relação sinal/ruído das predições computacionais de TFBS foram desenvolvidas diversas outras abordagens que fazem uso de estratégias estatísticas (Potencial Regulacional e PhastCons) e estratégias puramente filogenéticas que fazem uso da identificação de seqüências conservadas inter-espécies. No presente trabalho, nós desenvolvemos uma abordagem mista que faz uso de critérios filogenéticos de seleção de seqüências associada à seleção de sítios conservados para família de fatores de

transcrição. A seleção de sítios conservados, por sua vez, pode ser interpretada como um filtro estatístico, uma vez que a predição computacional de um sítio de ligação pode ser interpretada como um evento probabilístico markoviano.

Em um primeiro momento, através de uma profunda busca na literatura, foram determinados genes cujos produtos já foram direta ou indiretamente associados ao desenvolvimento das subpopulações de neurônios oriundos da medula embrionária. A escolha dessa modelo se deve tanto a forte seleção positiva aplicada para a manutenção como o fato de que esse modelo já é estudado funcionalmente pelo grupo liderado pelo professor Dr. Xavier Neto. Cada um desses genes teve o seu padrão de expressão determinado e essas informações foram utilizadas para a construção de árvores solução que acomodem o conhecimento atual em relação a expressão desses genes. As informações obtidas a partir dessa árvore solução em conjuntos com os mapas de calor foram utilizadas para a validação funcional das árvores determinadas por nossa abordagem de agrupamento. Inicialmente, foi comparada a eficiência de nossas abordagens seleção de sítios a partir da comparação das árvores solução geradas a partir dos sítios preditos por nosso algoritmo e o algoritmo MATCH®, largamente utilizado na literatura científica. Sítios selecionados como sendo de bom valor biológico pela nossa abordagem ou preditos com o auxílio da ferramenta MATCH® foram organizados em matrizes de característica. Tais matrizes tabulam o número de sítios para uma dada família de fatores de transcrição em cada um dos genes sob análise. Ambas as matrizes foram utilizadas como entrada para o algoritmo de agrupamento PvClust utilizando como métrica de similaridade a correlação de Pearson entre os elementos da matriz. As árvores solução dos dois agrupamentos foram comparadas e determinaram que nossa abordagem de filtro filogenético selecionou sítios de alto valor biológico que permitiram a construção de uma árvore solução mais próxima a realidade biológica. Foram agrupados de forma sustentada genes paralogos de divergência recente e genes com padrões de expressão espacial similar com base nas árvores de expressão construídas inicialmente.

Com o objetivo de refinar e melhorar nossa abordagem de agrupamento foram analisados os efeitos de uma normalização aplicada diretamente nas contagens das famílias de sítios para fatores de transcrição (Z-escore). Essa normalização se baseia na premissa de que é possível diminuir as distorções entre as contagens das famílias de fatores de transcrição. Essas distorções advêm do fato que algumas famílias são constituídas por fatores cujos sítios são mais facilmente preditos, essa facilidade está diretamente relacionada com a restritividade da matriz de busca para sítios desses fatores.

As contagens normalizadas foram agrupadas em uma matriz de características que foi posteriormente agrupada pelo algoritmo pvclust usando os mesmos parâmetros da matriz não normalizada. Os resultados do agrupamento das contagens normalizadas não foram animadores levando a construção de uma árvore solução com pouca correlação biológica e valores baixos de sustentabilidade. Uma provável limitação do uso dessa normalização é que embora ela teoricamente permita uma melhor comparação entre as contagens, ela leva a uma considerável perda de informações.

Uma vez determinada a melhor árvore solução, essa foi estudada detalhadamente através de cruzamentos com o conhecimento sobre o padrão de expressão desses genes. A análise dos resultados determinou o sucesso de nossa abordagem em identificar sítios de alto valor biológico cujo agrupamento permitiu, ao menos parcialmente, mimetizar a expressão espacial desses genes de interesse. A árvore foi subdividida em quatro sub-árvores respeitando totalmente a subdivisão cladística proposta pelo agrupamento. Com o objetivo de testar a aleatoriedade dos agrupados propostos foi construída uma tabela de contingência e nela foi aplicado um teste qui-quadrado de independência. O resultado de tal teste determinou que existe uma forte dependência entre os elementos agrupados em cada uma das árvores ($p\text{-value} \leq 10^{-4}$) e, portanto, nossa abordagem promoveu um agrupamento não aleatório dos genes. Cada um das quatro sub-árvores determinadas apresentou um agrupamento majoritário de genes de uma mesma classe de expressão (expressão dorsal, expressão ventral e expressão ampla). Dentre essas se destacam as sub-árvores 2, 3 e 4, onde respectivamente organizam-se majoritariamente genes com expressão ampla, dorsal e ventral.

Tendo construído uma boa árvore solução com boas relações funcionais entre seus elementos, nossos esforços foram voltados para entender quais foram os padrões preditos, representados por suas contagens, responsáveis tanto pelos agrupamentos como separações entre as quatro sub-árvores observadas. Para efetuar essas comparações duas métricas foram utilizadas: a primeira se baseia em ordenar as famílias com base em suas contribuições para a correlação de Pearson entre dois elementos e a segunda aonde são definidas famílias determinantes e discriminantes com base na razão entre as contagens das famílias. Ambas as contagens se mostraram efetivas para a seleção de famílias cujos fatores de transcrição, se mostram funcionalmente associados com os padrões de expressão espacial desses genes. Outras métricas podem e devem ser avaliadas para essa comparação entre as contagens das famílias dentre essas: a proporção de sítios para uma determinada família e a média de contagens corrigida por sua

dispersão.

Nossa proposta pode ser comparada com outras desenvolvidas anteriormente por outros grupos para a mineração de dados e extração de conhecimento a partir de padrões transcricionais preditos. Em geral tais abordagens são focadas em determinar sítios individuais ou conjuntos de sítios com proporção anormal dentro de um grupo de genes de interesse [69]. Um ponto que merece ser mais avaliado futuramente é o uso de outros filtros filogenéticos construídos a partir da combinação de alinhamentos contra outras espécies. Tais filtros podem permitir a seleção de conjuntos diferentes de sítios que podem alterar sensivelmente tanto a seleção de sítios de bom valor biológico como o agrupamento posterior de tais sinais. Porém, uma grande dificuldade imposta para a utilização de espécies mais distantes para a construção de filtros filogenéticos mais restritivos e o nível de coberturas desses projetos genoma. Os projetos genomas para pequenos mamíferos (não rodentios) e répteis estão ainda incompletos e o uso dessas informações pode levar a exclusão de seqüências funcionais, apenas pelo simples fato de que suas seqüências homólogas sequer foram depositadas ainda.

Nossa abordagem permitiu agrupar computacionalmente genes que apresentam padrões similares de sítios de transcrição seja esse evento resultante da ancestralidade comum entre os genes (parálogos) ou genes com pouca relação de ancestralidade, mas com expressão espacial similar no desenvolvimento neurológico de murinos. Tais resultados, embora iniciais, são promissores e podem ser estendidos para outros grupos de genes e condições fisiológicas de interesse. Como um próximo passo, uma interface gráfica intuitiva está sendo desenvolvida para permitir a simples utilização das abordagens propostas no presente texto. Essa interface permitirá uma melhor visualização da organização dos padrões transcricionais preditos nos blocos conservados em cada um dos genes e quais são os módulos regulacionais compartilhados por genes com expressão similar.

Conclusão

Conclusões

A regulação da expressão de um gene é intrinsecamente relacionada com a ação combinada de fatores de transcrição. Tais proteínas reconhecem padrões na estrutura do DNA, que constituem-se em sítios de ligação ou TFBS (Transcription Factor Binding Sites), que permitem que a interação entre os fatores de transcrição e o complexo de transcrição. Esse trabalho foi fundamentado em três grandes objetivos: Desenvolver uma abordagem para a melhoria da predição computacional de sítios de ligação, construir distribuições nulas para a aplicação de testes de enriquecimentos de sítios e construir uma abordagem que permite agrupar genes com base em seus padrões transcrpcionais preditos em suas regiões regulatórias.

O primeiro objetivo foi cumprido a partir do estudo funcional de uma sequência indutora de transcrição (SIT) próxima ao gene *aldh1a2*. A partir da utilização de nosso algoritmo de predição de TFBSs foi possível predizer sítios de bom valor biológico cuja função foi comprovada através de ensaios bioquímicos. Esses sítios foram utilizados para a construção de um modelo de interação para os fatores de transcrição que controlam a expressão do gene estudado. Essa pequena sequência de nucleotídeos foi também testada para o enriquecimento de sítios em suas bases de DNA. Os resultados dessa análise corroboraram grandemente com o modelo de expressão proposto para esse gene. A abordagem de seleção de sítios de bom valor biológico foi novamente testada agora para a predição computacional de sinais transcrpcionais nas regiões regulatórias conservadas próximas a genes de interesse. Tais sinais foram contabilizados e posteriormente agrupados e a árvore solução resultante recapitulou bem os padrões de expressão desses genes no processo de desenvolvimento neurológico de vertebrados.

Esses resultados analisados em conjunto permitiram concluir que as mesmas forças de seleção que agem na seleção positiva de exons, também agem ativamente na purificação dos padrões transcrpcionais das regiões regulatórias dos genes. Esse estrito controle permite que genes sejam estudados a partir desses sinais de regulação transcrricional. Essas abordagens de seleção de sítios serão estendidas e, permitirão a predição *in silico* de sequências funcionais em regiões genômicas de genes de interesse.

Anexos