Representação de sistemas biológicos a partir de sistemas dinâmicos: Controle da transcrição a partir do estrógeno

Marcelo Ris

Tese Apresentada ao Programa Interunidades em Bioinformática

da Universidade de São Paulo

para a obtenção do Grau de Doutor em Bioinformática

Área de Concentração: **Bioinformática**
Orientador: **Prof. Dr. Junior Barrera**
Co-orientadora: **Profa. Dra. Helena Brentani**

Durante a elaboração deste trabalho, o autor recebeu apoio financeiro da **CAPES**, da **Texas A&M University** e do **NIH**.

São Paulo
2008

Representação de sistemas biológicos a partir de sistemas dinâmicos: Controle da transcrição a partir do estrógeno

Este exemplar corresponde à redação
da tese, apresentada e defendido por
Marcelo Ris à comissão julgadora.

São Paulo, 14 de abril de 2008.

Banca examinadora:

- Prof. Dr. Junior Barrera (Orientador) - Instituto de Matemática e Estatística, USP, São Paulo(SP), Brasil
- Prof. Dr. Roberto Marcondes César Júnior - Instituto de Matemática e Estatística, USP, São Paulo(SP), Brasil
- Prof. Dr. Alexandre Xavier Falcão - Instituto de Computação, Unicamp, Campinas(SP), Brasil
- Prof. Dr. Roger Chammas - Faculdade de Medicina, Departamento de Radiologia, USP, São Paulo(SP), Brasil
- Profa. Dra. Maria Aparecida Azevedo Koike Folgueira - Faculdade de Medicina, Departamento de Radiologia, USP, São Paulo(SP), Brasil

Orientador:

- Prof. Dr. Junior Barrera (Orientador) - Instituto de Matemática e Estatística, USP, São Paulo, Brasil
- Profa. Dra. Helena Brentani (Co-orientadora) - Hospital do Câncer, São Paulo, Brasil

Agradecimentos

Ao meu pai, Isaac, a quem devo toda a minha educação e tudo o que tive e tenho na minha vida,

Ao Prof. Dr. Junior Barrera, de quem pude contar com a sua valiosa orientação de meu trabalho,

À Profa. Dra. Helena Brentani, de quem pude contar com a sua valiosa orientação na parte biológica do trabalho,

Ao Prof. Dr. Roberto Marcondes Cesar Junior, pela contribuição na parte matemática,

À Banca Examinadora, pela cuidadosa leitura e pelas correções,

Ao David Martins pelos testes do algoritmo aplicado à imagens,

A minha mãe, Léia, e meu segundo pai, Arnaldo, que foram meus grandes conselheiros em todos os momentos da minha vida,

A minha segunda mãe, Helena, que me criou como seu filho e a quem eu tenho o maior orgulho,

Aos meus irmãos, Fernando e Márcia, e meu cunhado, Marcelo, companheiros sempre,

Aos meu sobrinhos, Daniel e Michelle, que amo muito e deixam minha vida mais alegre,

A todos os meus colegas de estudo que me ajudaram muito nestes anos de estudo,

A todos meus amigos que me incentivaram,

A todos os professores dos institutos envolvidos,

E a CAPES, pela bolsa,

o meu *sincero agradecimento.*

Marcelo Ris.

Resumo

Esta pesquisa de doutorado apresenta resultados em três áreas distintas: (i) Ciência da Computação e Estatística – devido ao desenvolvimento de uma nova solução para o problema de seleção de características, um problema conhecido em Reconhecimento de Padrões; (ii) Bioinformática – em razão da construção de um método baseado em um *pipeline* de algoritmos, incluindo o de seleção de características, visando abordar o problema de identificação de arquiteturas de redes de expressão gênica; e (iii) Biologia – ao relacionar o estrógeno com uma nova função biológica, após analisar informações extraídas de séries temporais de *microarrays* pelas novas ferramentas computacionais-estatísticas desenvolvidas.

O estrógeno possui um importante papel nos tecidos reprodutivos. O crescimento das gândulas mamárias e do endométrio durante a gravidez e o ciclo menstrual são estrógeno dependentes. O crescimento das células tumorais nesses órgãos podem ser estimuladas pela simples presença de estrógeno; mais de 300 genes são conhecidos por terem regulação positiva ou negativa devido a sua presença. A motivação inicial desta pesquisa foi a construção de um método que possa servir de ferramenta para a identificação de genes que tenham seu nível de expressão alterado a partir de uma resposta induzida por estrógeno, mais precisamente, um método para modelar os inter-relacionamentos entre os diversos genes dependentes do estrógeno.

Apresentamos um novo *pipeline* de algoritmos que, a partir de dados temporais de *microarray* e um conjunto inicial de genes que compartilham algumas características comuns, denominados de *genes sementes*, devolve como saída a arquitetura de uma rede gênica representada por um grafo dirigido. Para cada nó da rede, uma tabela de predição do gene representado pelo nó em função dos seus genes preditores (genes que apontam para ele) pode ser obtida. O método foi aplicado em estudo de série-temporal de *microarray* para uma cultura de células *T-47D* submetidas a tratamento com estrógeno, e uma possível rede de regulação foi obtida. Encontrar o melhor subconjunto preditor de genes para um dado gene pode ser estudado como um problema de seleção de características, no qual o espaço de busca pode ser representado por um reticulado Booleano e cada um de seus elementos representa um subconjunto candidato. Uma característica importante desse problema é o fato de que para cada elemento existe uma função custo associada, e esta possui forma de curva em U para qualquer cadeia maximal do reticulado. Para esse problema, apresentamos um nova solução, o algoritmo *U-curve*. Esse algoritmo é um método do tipo *branch-and-bound*, o qual utiliza a estrutura do reticulado Booleano e a característica de curva em U da função custo para explorar um subconjunto do espaço de busca equivalente à busca completa. Nosso método obteve excelentes resultados em eficiência e valores quando comparado com as heurísticas mais utilizadas (SFFS e SFS).

A partir de um método baseado no *pipeline* e de um conjunto inicial de genes regulados *diretamente* pelo estrógeno, identificamos uma evidência de envolvimento do estrógeno em um processo biológico ainda não relacionado: a adesão celular. Esse resultado pode direcionar os estudos sobre estrógeno e câncer à investigação de processo metastático, o qual é influenciado por genes relacionados à adesão celular.

Abstract

This Phd. research presents in three distinct areas: (i) Computer Science and Statistics – on the development of a new solution for the feature selection problem which is an important problem in Pattern Recognition; (ii) Bioinformatics – for the construction of a pipeline of algorithms, including the feature selection solution, to address the problem of identification the architecture of a genetic expression network and; (iii) Biology – relating estrogen to a new biological function, from the results obtained by the new computational-statistic tools developed and applied to a time-series microarray data.

Estrogen has an important role in reproductive tissues. The growth mammary glands and endometrial growing during menstrual cycle and pregnancy are estrogen dependent. The growth of tumor cells in those organs can be stimulated by the simple presence of estrogen. Over 300 genes are known by their positive or negative regulation by estrogen. The initial motivation of this research was the construction of a method that can serve as a tool for the identification of genes that have changed their level of expression changed by a response induced by estrogen, more specifically, a method to model the inter-relationships between the several genes dependent on estrogen.

We present a new pipeline of algorithms that from the data of a time-series microarray experiment and from an initial set of genes that share some common characteristics, known as *seed genes*, gives as an output an architecture of the genetic expression network represented by a directed graph. For each node of the network, a prediction table of the gene, represented by the node, in function of its predictors genes (genes that link to it) can be obtained. The method was applied in a study of time-series microarray for a cell line *T-47D* submitted to a estrogen treatment and a possible regulation network was obtained. Finding the best predictor subset of genes for a given gene can be studied as a problem of feature selection where the search space can be represented by a Boolean lattice and each one of its elements represents a possible subset. An important characteristic of this problem is: for each element in the lattice there is a cost function associated to it and this function has a U-shape in any maximal chain of the search space. For this problem we present a new solution, the *U-curve* algorithm. This algorithm is a branch-and-bound solution which uses the structure of the Boolean lattice and U-shaped curves to explore a subset of the search space that is equivalent to the full search. Our method obtained excellent results in performance and values when compared with the most commonly used heuristics (SFFS and SFS).

From a method based on the pipeline of algorithms and from an initial set of genes direct regulated by estrogen, we identified an evidence of involvement of estrogen in a biological process not yet related to estrogen: the cell adhesion. This result can guide studies on estrogen and cancer to research in metastatic process, which is affected by cell adhesion related genes.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Modelar inter-relações de genes em uma *rede genética* específica é um dos estudos de maior desafio na área de sistemas biológicos. Existem inúmeros estudos que tentam modelar essas redes. Um resumo destes pode ser encontrado no trabalho de Dejong [21]. As redes podem ter fundamento na análise de regiões promotoras ou na análise de expressões gênicas (*microarray* e *SAGE*), e a partir desses dados a rede é construída. Uma rede genética (ou gênica) é caracterizada quando se conhece como a expressão de um conjunto de genes em um certo instante, afeta a expressão de outro em um instante subseqüente. Modelos de redes genéticas são representados por *sistemas dinâmicos*. O grafo orientado definido pelas componentes da função de transição que caracteriza o sistema dinâmico é chamado *arquitetura da rede genética*.

A motivação inicial deste trabalho foi a identificação de genes que tenham seu nível de expressão alterado a partir de uma resposta induzida por estrógeno. Além disso, desenvolver um método para modelar os inter-relacionamentos entre os diversos genes dependentes do estrógeno e, assim, obter uma rede genética. Modelando as redes genéticas por uma classe de sistemas dinâmicos estocásticos, denominado *Redes Genéticas Probabilísticas* (*PGN*s) [3], desenvolvemos, aqui, um novo método, descrito por um *pipeline* de algoritmos, para identificar a arquitetura de redes genéticas.

Estimar inter-relacionamentos entre genes nos remete a um problema de otimização combinatória, no qual encontrar um "bom" subconjunto de genes preditores de um outro, significa minimizar uma função custo no espaço completo de todos os subconjuntos dos genes que formam a rede. Esta função custo é uma medida estatística sobre distribuições conjuntas de probabilidade que permite caracterizar dependência (por exemplo: entropia). Esse é um problema exponencial no qual as heurísticas conhecidas atingem um resultado sub-ótimo. Com o objetivo de desenvolver uma nova solução para esse problema, apresentamos um novo algoritmo *branch-and-bound*, denominado *algoritmo U-curve*, que executa uma busca completa sem percorrer o espaço total de possibilidades.

Dividiremos nosso trabalho nas seguintes partes:

- No Capítulo 2 apresentamos os conceitos empregados em nosso trabalho e como eles se encontram atualmente com suas bibliografias relevantes. Ele se compõe basicamente de: (i) uma introdução às redes genéticas; (ii) discussão dos experimentos de *microarrays* e

séries-temporais; (iii) uma intodução aos sistemas dinâmicos, discutindo sobre o conceito de Redes Genéticas Probabilísticas e identificação de uma rede; e (iv) um resumo sobre o estrógeno.

- No Capítulo 3, nossa solução para o problema de otimização combinatória é apresentado em detalhes. Resultados comparativos entre o nosso método e os mais utilizados atualmente são abordados e discutidos.

- No Capítulo 4, apresentamos nosso método com base no *pipeline* de algoritmos desenvolvido. Uma aplicação do método em dados de série-temporal de *microarray* resulta em uma rede genética representada por um grafo dirigido.

- No Capítulo 5, ao utilizar parte do *pipeline*, apresentado anteriormente, e partindo-se de genes regulados diretamente por estrógeno, pudemos evidenciar um novo marcador gênico para o estrógeno: a *adesão celular*. O processo utilizado, assim como a relevância deste resultado são, também, discutido ao final deste Capítulo.

- No Capítulo de conclusão, apontamos: (i) a contribuição deste trabalho em: Ciência da Computação e Estatística, Bioinformática e Biologia; (ii) um resumo dos resultados obtidos; e (iii) perspectivas de novas frentes de estudo que este trabalho oferece.

- No Apêndice A, descrevemos o programa que implementa o algoritmo *U-curve* com seus parâmetros e apresentamos uma parte do código implementado.

- No Apêndice B, apresentamos o banco de dados implementado e uma lista completa de genes-sementes para as iterações do *pipeline*.

- No Apêndice C, apresentamos a lista completa dos genes preditos pelo conjunto de genes regulados diretamente pelo estrógeno.

- No Apêndice D, listamos o conteúdo do DVD anexo.

- No Apêndice E, descrevemos os trabalhos apresentados e artigos submetidos e publicados.

1.1 Objetivos

Os principais objetivos deste trabalho podem ser enumerados a seguir:

- Definição do sub-sistema de genes afetados pelo estrógeno a partir de dados de expressão gênica;

- Obtenção da rede gênica para os genes do sub-sistema;

- Para cada gene do sub-sistema identificar o possível mecanismo transcricional que ele pertence;

- Encontrar novos possíveis genes afetados pelo estrógeno ainda não identificados.

Modelar inter-relacionamentos entre genes nos remete a um problema de otimização combinatória, onde encontrar um "bom" subconjunto de genes preditores de um outro, significa minimizar uma função custo no espaço completo de todos os subconjuntos possíveis. Este é um problema exponencial no qual as heurísticas conhecidas atingem um resultado sub-ótimo. Um objetivo secundário deste trabalho é desenvolver uma nova solução para este problema.

Capítulo 2

Estado da arte

2.1 Sistemas dinâmicos

Um *sistema dinâmico* é representado por um vetor de funções de transição que descreve a evolução temporal de um vetor denominado *estado*. As funções de transição podem, também, receber um sinal independente chamado de *entrada*. Em geral, não são os estados que são observados, e sim a transformação deles, a qual se dá por meio de um vetor de funções de transformação que resulta como resultado um *vetor de saída*. Para modelar uma *rede de expressão gênica* vamos utilizar sitemas dinâmicos finitos, discretos no tempo e finitos no número de estados. Um estado $x[t]$ em um instante t é um vetor de dimensão n, onde cada elemento $x_i[t], i = 1, .., n$ representa a expressão do gene i no instante t. A expressão pode ser discretizada por um conjunto R, por exemplo, $R = \{-1, 0, 1\}$, onde -1 indica sub-expressão, 0 expressão padrão, 1 super-expressão do gene, sendo assim, a função de transição ϕ é uma função de R^n em R^n, e mapeia um estado $x[t]$ no próximo estado $x[t + 1]$. Um sistema dinâmico finito (no número de estados), em cada instante t, é representado por:

$$x[t + 1] = \phi(x[t]),$$

onde $x[t] \in R^n$, para todo $t \geq 0$.

Se a função de transição ϕ é a mesma para todos os instantes t, denominamos o sistema de *translação-invariante no tempo*.

2.2 Redes genéticas

Genes codificam proteínas, sendo que algumas delas, por sua vez, regulam outros genes. Esse processo nos remete às denominadas *redes genéticas*. Redes genéticas combinam regulações gênicas com interações protéicas e podem ser extremamente complexas, isto é, um enorme número de variáveis (genes e proteínas) e um pequeno número de amostras tornam o estudo de todo o sistema inviável e impraticável. Existem um série de estudos [21, 10] que tentam modelar as redes genéticas. Esses estudos são baseados na análise da região promotora do gene por projetos de seqüenciamento ou no perfil de expressão do gene pela análise de dados de *microarray*

ou de *SAGE*. O objetivo desses estudos é obter uma rede a partir dos dados de entrada indicando a maneira como a expressão de cada gene afeta a expressão de outros. Algumas redes podem indicar se um gene reprime ou ativa outros, enquanto outras indicam que a expressão de um gene em um certo período pode predizer a expressão de outro no período subseqüente. A Figura 2.1 mostra um exemplo de arquitetura de uma rede genética: os nós indicam os genes, as arestas dirigidas indicam que um gene ativa o outro, por exemplo, conectando o gene A ao gene B, indica que A ativa B, e as arestas bloqueadas indicam que um gene reprime o outro, por exemplo, conectando o gene C ao gene B, indica que C reprime B. A Figura 2.2 mostra um exemplo de uma rede genética, representando os preditores de um gene ao longo do tempo, isto é, a expressão de B em um instante $(t+1)$ de tempo pode ser obtida pela expressão dos genes A e C no instante de tempo anterior (t). Uma *tabela de predição* também pode ser estimada para as redes dependentes de tempo. As tabelas de predição normalmente não indicam se um gene reprime ou ativa outro. A razão disso se dá devido ao fato de que a expressão de um gene depende, freqüentemente, não apenas de um único gene mas da expressão conjunta de um subconjunto destes.

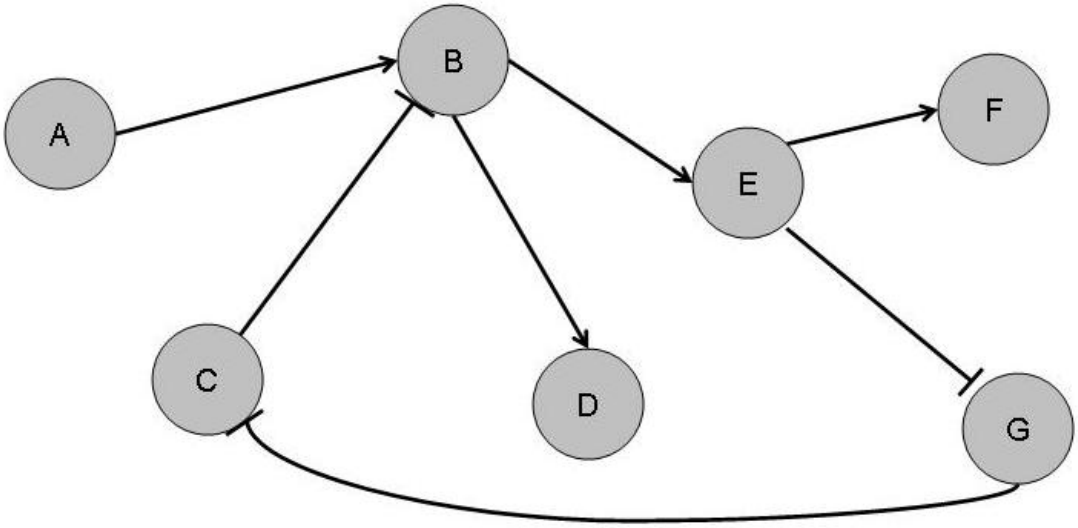


Figura 2.1: Um exemplo de arquitetura de uma rede genética composta por 7 genes

2.3 Rede genética probabilística

Podemos dizer que um sistema dinâmico é um *processo estocástico* quando a função de transição ϕ é uma *função estocástica*, isto é, o próximo estado $\phi(x[t])$ é a realização de um vetor aleatório X_t. Considere a seqüência de vetores aleatórios $X_0, X_1, X_2, ...$ assumindo valores $x[0], x[1], x[2]$ em R^n.

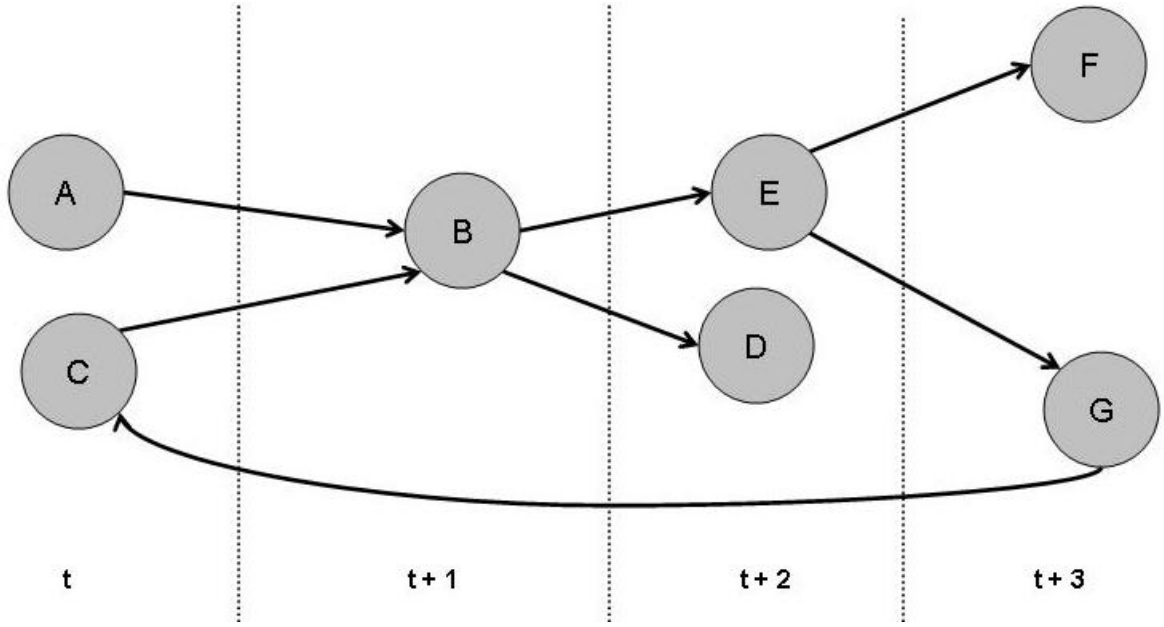


Figura 2.2: Um exemplo de uma rede genética de 7 genes para 4 períodos de tempo

Uma *cadeia de Markov*, de primeira ordem, é definida por uma seqüência de estados aleatórios $(X_t)_{t=0}^{\infty}$, se para cada instante $t \geq 1$,

$$P(X_t = x[t]|X_0 = x[0], ..., X_{t-1} = x[t-1]) = P(X_t = x[t]|X_{t-1} = x[t-1]),$$

ou seja, a probabilidade de um evento futuro depende apenas do último evento, e não de todos os eventos passados. Cadeias de Markov podem ser caracterizadas por uma *matriz de transição* $\pi_{Y|X}$ de probabilidades condicionais entre estados, em que cada elemento representa a probabilidade condicional $p_{y|x}$, onde x e y são dois estados da cadeia, e uma condição inicial de estados π_0. Com isso, a função de transição estocástica ϕ no instante $t \geq 1$ é dada por:

$\phi(x[t]) = y$,

onde y é a realização de um vetor aleatório com distribuição de probabilidades $p_{\bullet|x[t]}$.

Rede Booleana Probabilística (*PBN*) [36, 37] é uma extensão de *Redes Booleanas* [23], nas quais a função Booleana para determinar o próximo estado na rede é *não-determinística* e pode variar de acordo com uma função probabilística, isto é, uma função Booleana é sorteada a partir de uma distribuição de probabilidades e usada para determinar o próximo estado da rede. Em certos casos é desejável trabalhar não somente com valores Booleanos mas com valores discretizados em múltiplos níveis, por exemplo, $x_i[t] \in \{-1, 0, 1\}$ para expressões gênicas com: -1, indicando sub-expressão; 0, expressão normal (igual a um valor de referência); ou 1, super-expressão, e, nesses casos, a teoria deve ser extendida [44].

Rede Genética Probabilística (*PGN*) [3] é um modelo usado para representar redes genéticas. Ela é derivadas da *Rede Booleana Probabilística* (*PBN*), nas quais a função de transição não é determinística e os estados são compostos por valores discretos. Em outras palavras, *PGN* é uma

cadeia de Markov $(\pi_{Y|X}, \pi_0)$, na qual cada estado é um vetor de expressões gênicas discretizadas por alguns axiomas:

- a função de transição é *independente do tempo* ($\pi_{Y|X}$ é *homogêneo*): a probabilidade de um estado dado um estado anterior não varia no tempo ($p_{y|x}$ é independente de t);
- todas as probabilidades de transição possuem um valor positivo maior que 0 ($p_{y|x} > 0, \forall x, y \in R^n$): todos os estados, dado um estado anterior, tem probabilidade de ocorrer;
- a função de transição ($\pi_{Y|X}$) é *condicionalmente independente*: a probabilidade de um estado é dada pelo produto das probabilidades de suas componentes dado um estado anterior ($\forall x, y \in R^n, p_{y|x} = \prod_{i=1}^{n} p(y_i|x)$);
- a função de transição ($\pi_{Y|X}$) é *quase-determinística*: existe um estado com probabilidade muito maior de ocorrer do que os outros ($\forall x \in R^n, \exists i \in \{1, ..., n\}$, tal que, $|p_{y_i|x} \approx 1$);
- existe um subconjunto "pequeno" das componentes do estado que, quando a função de transição é restrita a ele, possui resultados iguais à função de transição aplicada ao estado completo. Em outras palavras, o próximo estado pode ser determinado por um "pequeno" subconjunto dos componentes do estado anterior, isto é: $\forall x \in R^n, \forall i \in \{1, ..., n\}$, existe um sub-espaço de dimensão $j, j << n$, tal que: $p_{y_i|x'} \approx p_{y_i|x}$, onde x' é a projeção de x neste sub-espaço.

Esses axiomas são motivados pela comum falta de dados e enorme número de variáveis em experimentos biológicos, isto é, existem poucas amostras de *microarray* para milhares de genes em um experimento série-temporal *microarray*, e implicam que cada gene pode ser caracterizado por um vetor de coeficientes a e um vetor de funções estocáticas g_j que leva um número inteiro em $\mathbb{Z}$ a um elemento de R. O vetor a^j indica os genes que afetam o gene j, isto é, se $a_i^j > 0$, o gene i *excita* o gene j (regulação positiva), se $a_i^j < 0$, o gene i *inibe* o gene j (regulação negativa), se $a_i^j = 0$, o gene i *não afeta* o gene j. Logo, o gene j é previsto apenas pelos genes i que possuem $a_i^j \neq 0$. Podemos assim decompor a função de transição estocástica ϕ em componentes ϕ_j que são obtidos pela composição da função estocástica g_j aplicada a combinação linear de a^j e o estado anterior $x[t]$, isto é, para todo instante $t \geq 1$:

$$\phi_j(x[t]) = g_j(\sum_{i=1}^{n} a_i^j x_i[t]),$$

onde $g_j(\sum_{i=1}^{n} a_i^j x_i[t])$ é a realização de uma variável aleatória em R com distribuição $p_{\bullet|\sum_{i=1}^{n} a_i^j x_i[t]}$.

2.4 Identificação do sistema

Nesta seção estudaremos um método para a identificação da PGN (Rede Genética Probabilística) e, para isso, iremos introduzir alguns conceitos estatísticos necessários. Os conceitos definidos a seguir são aplicados a variáveis aleatórias discretas, as quais podem assumir valores discretos no conjunto $R = \{1, 2, ..., n\}$:

A *Entropia* $H(X)$ de uma variável aleatória X é uma medida aplicada à distribuição p_X, dada por:

$$H(X) = \sum_{i=1}^{n} p_i \log p_i,$$

onde $p_i = P(X = i), i \in R$. A entropia é uma função aplicada a distribuição de probabilidades que possui algumas importantes características:

1. qualquer distribuição $p_{X'}$ formada por uma permutação da distribuição p_X possui mesmo valor de entropia;

2. quanto maior a concentração de probabilidade em algum valor, isso implica em um menor valor de entropia. Podemos dizer, com isso, que a entropia mede o grau de aleatoriedade da variável X. Pode-se provar que a máxima entropia é atingida na distribuição uniforme, e a mínima é atingida quando a probabilidade está toda concentrada em apenas um valor[1].

A *Informação Mútua* $I(X, Y)$ entre duas variáveis aleatórias X e Y, é obtida por:

$$I(X, Y) = H(Y) - H(Y|X),$$

onde $I(X, Y) \geq 0$. Essa medida aufere o grau de independência entre X e Y valendo 0 se e somente se X e Y são independentes. A entropia condicional $H(Y|X)$ mede a concentração de massa de p_Y em $p_{Y|X}$, ou seja, quanto maior a concentração de massa, menor seu valor, atingindo seu pico em $H(Y)$; já a informação mútua descreve caminho inverso, quanto maior a concentração de massa, maior seu valor.

Como vimos na seção anterior, a hipótese quase-determinística assumida para a PGN e a hipótese de que um subconjunto pequeno de genes é necessário para a previsão de um gene-alvo nos remete a um valor alto para a esperança da informação mútua deste em relação ao subconjunto de genes. Com isso, o problema de identificação da PGN (encontrar tais subconjuntos de genes preditores para cada gene-alvo) pode ser interpretado como um problema de maximização da esperança da informação mútua $E[I(X, Y)]$, dada por:

$$E[I(X, Y)] = H(Y) - E[H(Y|X)],$$

onde a variável aleatória Y representa a expressão $y_j[t+1]$ do gene j no instante $t+1$, e a variável X representa o vetor de variáveis aleatórias X_i, que é a ponderação da expressão do gene i no instante t pela componente a_i do vetor de coeficiente inteiro a, visto na seção anterior. No nosso caso, assumiremos que $a_i \in \{0, 1\}$ valendo 1 se o gene i afeta (positivamente ou negativamente) o gene j, e 0 caso contrário. O problema então se transforma em um outo: encontrar qual vetor a (ou vetores a's) maximiza (maximizam) a informação mútua, ou de modo análogo minimiza (minimizam) a *entropia condicional média*. Aplicando o método para cada gene j podemos construir a rede de conexão entre os n genes estudados, onde cada gene j pode ser predito pelos genes que se ligam a ele, segundo a estimação da distribuição de probabilidade $p_{y_j|a}$, obtida

sempre dos dados amostrais. Esse problema pode ser interpretado na área de *Reconhecimento de Padrões* [9] como um problema de *seleção de características*, onde estamos interessados em selecionar os genes (características) que melhor predizem um outro gene (classe), segundo uma *função critério* que para nosso caso será a minimização da entropia condicional média estimada. Para a estimação da entropia condicional média, podem ser usados vários métodos, sendo o empregado em nossos experimentos a estimação utilizada por Barrera et al. [3].

2.5 O problema de seleção de características

Um algoritmo de otimização combinatória seleciona, a partir de uma coleção finita de objetos (chamada de *espaço de busca*), aquele objeto que possui custo mínimo de acordo com uma dada função custo. A arquitetura mais simples para esse algoritmo, chamada de *busca completa*, percorre todos os objetos do espaço de busca, porém, é impraticável para espaços de busca de tamanho grande. Para esses casos, é possível percorrer alguns objetos e escolher o de mínimo custo, baseando-se nas medidas observadas. Algoritmos baseados em *heurísticas* e *branch-and-bound* são dois tipos de soluções para este problema. Os algoritmos baseados em heurísticas não possuem garantia formal em encontrar o objeto de custo mínimo, enquanto os *branch-and-bound* possuem propriedades matemáticas que garantem encontrá-lo.

O problema aqui estudado é um problema de otimização combinatória onde o espaço de busca é composto por 2^n objetos, organizados como um reticulado Booleano, e a função custo possui um formato de curva em U para qualquer cadeia maximal do espaço de busca.

Esse tipo de estrutura é encontrado em alguns problemas conhecidos, tais como: seleção de características em Reconhecimento de Padrões [9, 19] e projeto de *W-operadores* em morfologia matemática [22]. Nesses problemas, um subconjunto mínimo de características, que é suficiente para representar um objeto, deve ser obtido de um conjunto de n características. No projeto de W-operadores, as características são pontos de um retângulo em Z^2 chamado de janela. As funções com formato de curva em U são formadas pelo *erro de estimação* dos classificadores ou W-operadores projetados. Esse é um fenômeno muito conhecido na área de Reconhecimento de Padrões: para um conjunto de treinamento fixado, o aumento no número de características consideradas para a construção do classificador implica na redução do erro de classificador, pelo aumento na separação entre as classes. Isso ocorre até que os dados disponíveis se tornam insuficientes para cobrir o domínio do classificador, e o aumento do erro de estimação implica em aumento do erro do classificador. Os métodos conhecidos para esse problema são na sua maioria heurísticas, sendo o SFS e o SFFS [32] dois casos de relativos sucessos.

Existe uma gama de algoritmos *branch-and-bound* na literatura e a maioria deles são baseados na premissa de monotonicidade da função custo [12, 27, 40, 43]. No trabalho em [39] temos um resumo da literatura existente de algoritmos *branch-and-bound*. Se a real distribuição das probabilidades conjuntas entre os padrões e classes são conhecidas, dimensionalidades maiores implicam em menores erros de classificação. Na prática, porém, essas distribuições não são conhecidas e devem ser estimadas. O problema na premissa de monotonicidade da função custo se dá, devido ao fato de que ela não considera o erro de estimação implícito para espaços de dimensões grandes ("*curse of dimensionality*", também conhecido como "*problema U-curve*" ou "*fenômeno peaking*" [19]).

2.6 O estrógeno

O *estrógeno* não é apenas um *hormônio*, mas, sim, um grupo de hormônios que pode ser encontrado principlamente em três principais formas no corpo humano: *estrona* (*E1*), *estradiol* (*E2*) e *estriol*(*E3*), sendo a forma $17-\beta$ estradiol a mais abundante no corpo humano. O estrógeno possui um papel muito importante nos tecidos reprodutivos [45]: o crescimento das *glândulas mamárias* [14] e do *endométrio* durante a gravidez e o ciclo menstrual são *estrógeno-dependentes*. Os efeitos biológicos dos estrógenos são mediados por meio de seu receptor (ER) α e β. Esses receptores são membros de uma super-família de receptores nucleares e agem como um *fator de transcrição* quando ativados pela molécula ligante. Os receptores de estrógeno podem influenciar a expressão gênica por meio de quatro vias clássicas conhecidas [28, 4]:

1. via clássica (nuclear): *dímero E2-ER* se liga ao *elemento de resposta* na região promotora dos genes;
2. via independente de elemento de resposta: o *dímero E2-ER* se liga por uma conexão proteína-proteína a um fator de transcrição que, por sua vez, se liga na região promotora do gene.
3. via independente de estrógeno: *fatores de crescimento* (*GF*) ativam via de proteína-kinase, a qual fosforila e ativa os receptores de estrógeno quando ligados aos seus elementos de resposta.
4. via não-genômica: o *dímero E2-ER* ativa via de proteína-kinase, a qual pode alterar funções de proteínas no citoplasma ou fosforilar e ativar fatores de transcrição.

A Figura 2.3 mostra uma representação esquemática das quatro vias descritas.

É importante notar que os receptores de estrógeno contêm duas *funções de ativação*(FA) transcricional independentes: uma autônoma, presente na porção *N-terminal*, e outra dependente de ligantes. Assim, a atividade dos ERs pode ser controlada por várias outras proteínas. Essa ativação de ERs dependente de ligantes necessita de *co-ativadores* como TIF2 e SRC-1. A ativação transcricional dos ERs também é regulada por *fosforilação*, seja pela *MAPK* na presença de fatores de crescimento ou pela *Cdk7* na forma dependente de ligantes.

Existem dois tipos de receptores de estrógeno ERα e ERβ concentrados e localizados em diferentes partes do organismo, e ambos podem se ligar ao estrógeno ou a outras moléculas como vimos, tornando o complexo ativo ou inativo. A Figura 2.4 mostra a estrutura tridimensional dos receptores de estrógeno, quando ativados (ligados ao estrógeno) ou desativados (ligados a uma molécula antagonista).

O estrógeno exibe um papel biológico importante nos tecidos dos *órgãos reprodutores*. O crescimento da *glândula mamária* como também do *endométrio* durante a gravidez e durante o ciclo menstrual são dependentes de estrógeno. Em adição aos *efeitos proliferativos* normais do estrógeno, temos os efeitos estimulantes para iniciação e promoção de tumores nestes mesmos órgãos. O crescimento estimulado pelo estrógeno nas células tumorais, assim como, nas normais requer a presença de receptores de estrógeno. A expressão dos receptores de estrógeno em uma célula tumoral pode classificar o tumor em dois sub-grupos: os tumores *ER+* (ou *ER positivos*)

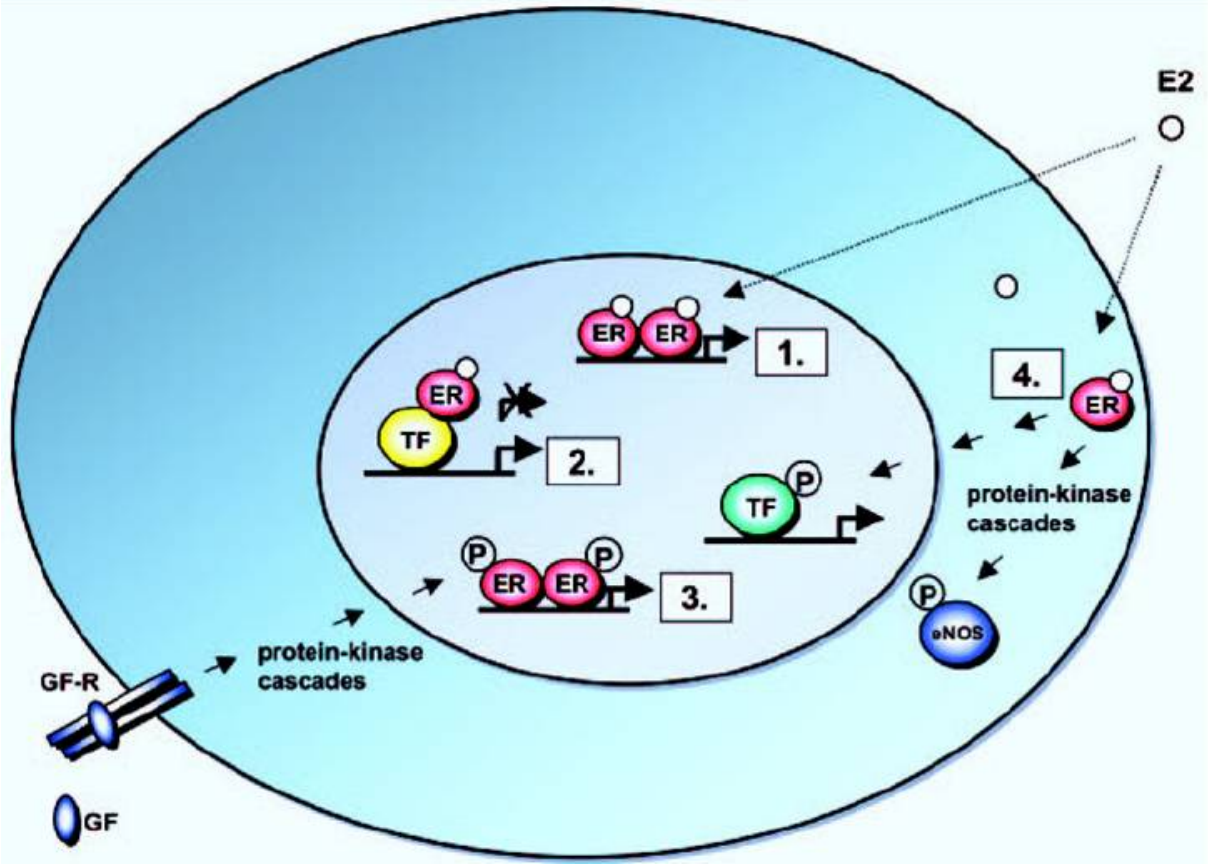


Figura 2.3: Representação esquemática das 4 vias clássicas de ação dos receptores de estrógeno (Extraído do trabalho em [4])

e os tumores *ER-* (ou *ER negativos*). Sabe-se que os tumores *ER-* são mais agressivos e possuem pouco sucesso nos tratamentos conhecidos até o momento. Nas clínicas médicas são usadas terapias tanto com *drogas anti-estrogênicas* como com *drogas inibidoras de aromatases*. Espera-se que o *Tamoxifen*, uma droga anti-estrogênica, se ligue ao ER tornando-o não funcional, enquanto que os inibidores de aromatase reduzem os níveis de estrógeno. A maioria dos tumores *ER+* responde ao tratamento com Tamoxifen, no entanto uma grande parcela destes pacientes com o tempo adquire resistência ao tratamento além de não suportarem o tratamento pelos efeitos colaterais das drogas. Os pacientes ER negativos, por não apresentarem o receptor de estrógeno, usualmente não respondem a terapia anti-estrogênica.

Existem mais de 300 genes conhecidos regulados positivamente ou negativamente pelo estrógeno, segundo esta via direta ou usando outras indiretas. Entre os genes regulados positivamente, podemos citar: *IGFBP4 (Insulin-like growth factor binding protein 4)*, *GREB1*, *PGR (Progesterone receptor)*; e os negativamente: *NMA (Putative transmembrane protein)*, *BMP7 (Bone morphogenetic protein 7)*. Os estudos, atualmente, estão direcionados de modo a caracterizar o melhor possível a via metabólica do estrógeno, ou seja:

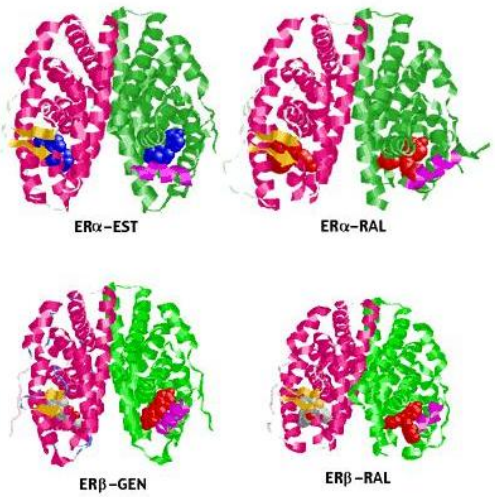


Figura 2.4: Representação tridimensional dos receptores de estrógeno (ERα e ERβ). Ligados ao estrógeno: ERα-EST (Estradiol) ou ERβ-GEN (Genisteína); e ao Raloxifene (antagonista de estrógeno): ERα-RAL ou ERβ-RAL

- identificação dos genes regulados positivamente ou negativamente pelo estrógeno;
- identificação de vias alternativas ao estrógeno, bem como novos fatores de transcrição alternativos aos receptores de estrógeno, e, assim, explicar a regulação nos tumores ER-;
- identificação da *rede de expressão gênica*, que caracteriza a inter-regulação entre os genes encontrados.

Existem vários estudos feitos na identificação dos genes regulados pelo estrógeno, entre eles, podemos citar:

- *V. X. Jin et al.* [20] utilizam uma nova técnica denominada *ChIP-on-chip* para a identificação de seqüência promotora relativa ao receptor ERα; 70 genes candidatos foram obtidos;
- *B. S. Katzenellenbogen et al.* [13] obtêm 438 genes regulados, dos quais 70 % são reprimidos pelo estrógeno, a partir de dados de *microarray* em células *MCF-7* (células de tumor de mama *ER+*);
- *E. T. Liu et al.* [26] obtêm uma lista de 386 genes responsivos ao estrógeno, dos quais 137 são regulados pelo estrógeno e 89 diretamente regulados (59 positivamente e 30 negativamente), a partir de experimentos de série-temporal de *microarray* em céluas *T47-D* (células de tumor de mama *ER+* – expressam ER em maior quantidade que as células MCF-7) submetidas ao estrógeno e componentes anti-estrógeno e inibidor de síntese protéica;

- *A. Weisz et al.* [41] identificam um conjunto de 61 genes que caracterizam células ERα positivas, a partir de dados de *microarray* em culturas celulares *MCF-7* e *ZR-75.1* (células de tumor de mama ER-positivas) induzidas por estrógeno;
- *K. R. Coser et al.* [7] estudam a sensitividade dos genes em resposta ao estrógeno, por meio de análise de *microarray* em cultura celular *MCF-7* induzidas por diferentes níveis de concentração de estrógeno, e o resultado obtido sugere diferenças em sensitividade entre os genes responsivos ao estrógeno;
- *A. S. Levenson* [25] procuram identificar genes regulados por estrógeno por meio de análise de *microarray* em cultura celular *MCF-7* induzida por diferentes complexos entre ERα e seus diversos ligantes, tais como: estrógeno, raloxifene e tamoxifene.

Os genes responsivos ao estrógeno podem ser agrupados em razão de suas funções biológicas. Podemos citar dois métodos de padronização das funções biológicas do genes: (i) o *Gene Ontology Consortium* [2] que relaciona o gene com processos biológicos de qualquer organismo; (ii) o banco de dados *KEGG* [29], que associa um gene a uma *via de regulação*. Genes que apresentam funções biológicas associadas à proliferação celular podem ser relacionados ao câncer [42] e o estrógeno, por sua vez, regula positivamente ou negativamente esses genes [13]. Entre as funções biológicas relacionadas à proliferação celular dos genes regulados pelo estrógeno temos: *ciclo celular*, *apoptose*, *fatores de crescimento*, *citoquinas*, *hormônios*, *receptores*, *transdução de sinais*, *fatores de transcrição*, *fatores de crescimento*.

2.7 *Microarrays*

A tecnologia de *microarray* [34] permite estudos baseados em larga escala de dados, tais como *genômica funcional* e *sistemas biológicos*, pela medição de níveis de expressão de milhares de genes simultaneamente. Os dados resultantes são a base para uma gama de estudos de perfis moleculares e para o entendimento dos mecanismos biológicos e de regulação para um organismo específico ou uma cultura celular [35]. Esses estudos podem ser executados a partir de um único experimento, resultando no agrupamento (*clustering*) de genes que possuem sinais de expressões semelhantes. Quando os estudos necessitam dos perfis de expressão em um período de tempo, em vez de um único intervalo, o experimento de *série-temporal de microarray* (*time-course microarray*) é utilizado. A análise dos dados de expressão dessa técnica pode agrupar os genes com perfis de expressão temporal semelhantes [31], mas, também, podem ser a base para a construção de uma rede genética [17]. A Figura 2.5 mostra uma representação esquemática de um experimento de série-temporal de *microarray*: n genes em m períodos de tempo e cada vetor $x[i]$ representa o experimento de *microarray* no instante i.

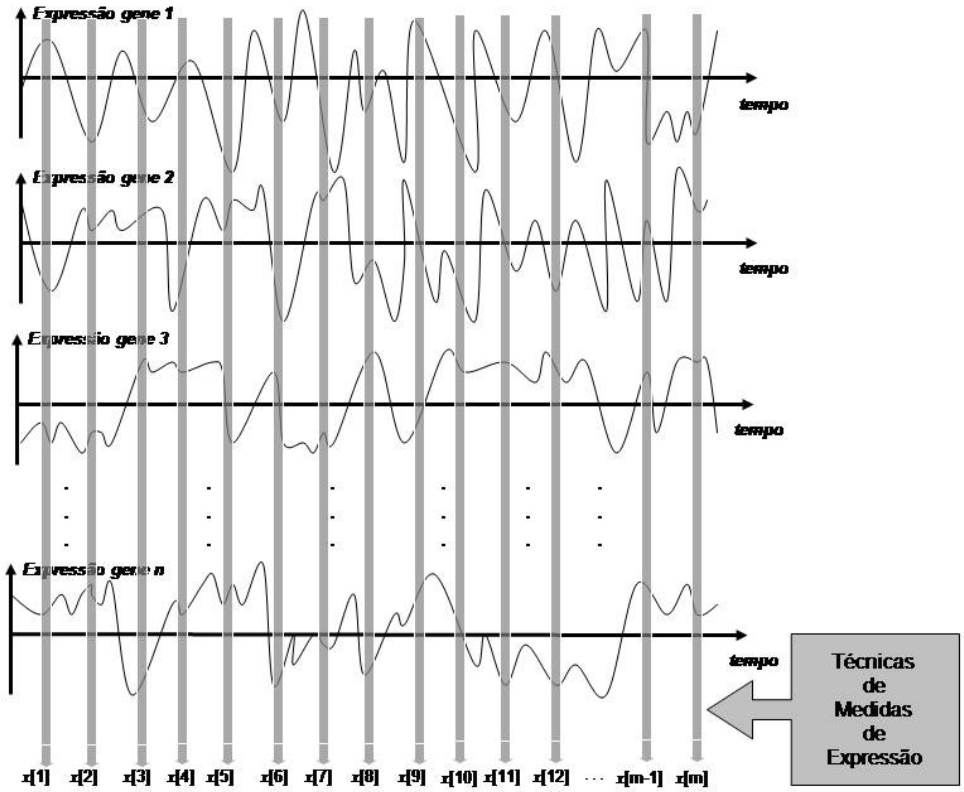


Figura 2.5: Representação esquemática de um experimento de série-temporal de *microarray*

Capítulo 3

O algoritmo *U-curve*

Neste Capítulo apresentaremos um probema de otimização combinatória com as seguintes características:

i. o espaço de busca é composto por 2^n objetos organizados como um *reticulado Booleano*;

ii. a função custo descreve uma *curva em U* quando aplicada a todos os elementos de qualquer cadeia maximal do reticulado.

Essa formulação pode ser caracterizada por um problema de *seleção de características* no contexto de *Reconhecimento de Padrões*. Os estudos conhecidos para esse problema recaem em heurísticas que exploram o espaço de maneira parcial, não sendo equivalente a busca completa ou outros algoritmos *branch-and-bound* que dependem da monotonicidade da função custo para atingirem a solução ótima com um tempo computacional bem menor que o obtido pela busca exaustiva.

Apresentamos uma solução *branch-and-bound*, a qual usa a estrutura do reticulado Booleano e as curvas em U da função custo para explorar um subconjunto do espaço de busca equivalente à busca completa. Algumas aplicações no projeto de *W-operadores* e na identificação da arquitetura de redes gnéticas ilustam os resultados aqui apresentados. Novas propriedades sobre reticulados Booleanos foram descobertas e aplicadas no desenvolvimento de uma estrutura de dados adequada, para representar e atualizar a parte não explorada do espaço de busca.

Seguindo esta Introdução, a Seção 3.1 apresenta a formalização do problema estudado. A Seção 3.2 descreve o algoritmo *branch-and-bound* de maneira estrutural. A Seção 3.3 apresenta as propriedades matemáticas que dão suporte aos passos do algoritmo. A seção 3.4 apresenta algumas aplicações do algoritmo para o projeto de W-operadores e para a identificação da arquitetura de redes genéticas. Finalmente, a Seção 3.5 discute as contribuições desta solução e propões algumas próximas etapas possíveis desta pesquisa.

3.1 Otimização de curvas em U

O espaço de busca é composto por 2^n objetos, organizados em um reticulado Booleano. Seja W um subcojunto finito, $\mathscr{P}(W)$ a coleção de todos os subconjuntos de W, $\subseteq$ a relação usual de inclusão sobre conjuntos e $|W|$ denotando a cardinalidade de W.

O *conjunto parcialmente ordenado* (*poset*) $(\mathscr{P}(W), \subseteq)$ é um *reticulado Booleano completo* $\mathcal{L}$ de grau $|W|$ onde: o menor e o maior elemento são, respectivamente, $\emptyset$ e W; a soma e o produto são, respectivamente, a união e intersecção usuais sobre conjuntos e o *complemento de um conjunto* X em $\mathscr{P}(W)$ é o seu complemento em relação a W, denotado por X^c.

Subconjuntos de W serão representados por cadeias de 0's e 1's, com 0 indicando que o ponto não pertence ao subconjunto e 1 indicando que ele pertence. Por exemplo, se $W = \{(-1,0),(0,0),(+1,0)\}$, o subconjunto $\{(-1,0),(0,0)\}$ será representado por 110. Como abuso de linguagem, ocorrências de $X = 110$, por exemplo, indica que X é o conjunto representado por 110.

Uma *cadeia* em $\mathcal{X} \subseteq \mathscr{P}(W)$ é uma coleção $\mathcal{A} = \{A_1, A_2, \ldots, A_k\} \subseteq \mathcal{X}$, tal que, $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_k$. Uma cadeia $\mathcal{M} \subseteq \mathcal{X}$ é *maximal* em $\mathcal{X}$ se não existe outra cadeia $\mathcal{C} \subseteq \mathcal{X}$ tal que $\mathcal{C}$ contêm propriamente $\mathcal{M}$.

Seja c uma *função custo* definida de $\mathscr{P}(W)$ em $\mathbb{R}$, dizemos que c é *decomponível em curvas em U* se, para qualquer cadeia maximal $\mathcal{M} \subseteq \mathscr{P}(W)$, a restrição de c em $\mathcal{M}$ descreve uma curva em U, isto é, para qualquer $A, X, B \in \mathcal{M}$, $A \subseteq X \subseteq B \Rightarrow \max(c(A), c(B)) \geq c(X)$.

A Figura 3.1 aprsenta um reticulado Booleano completo $\mathcal{L}$ de grau 4 e uma fução de custo c decomponível em curvas em U. Nessa figura, uma cadeia maximal com sua função custo em $\mathcal{L}$ é enfatizada. A Figura 3.2 apresenta as curvas da mesma função custo restrita a algumas cadeias maximais em $\mathcal{L}$ e em $\mathcal{X} \subseteq \mathcal{L}$. Note a forma em U das curvas da Figura 3.2.

Nosso problema consiste em encontrar o elemento (ou elementos) de *custo mínimo* no reticulado Booleano de ordem $|W|$. A busca completa nesse espaço é um problema exponencial, pois o espaço é composto por $2^{|W|}$ elementos. Sendo assim, para valores moderadamente grandes de $|W|$, a busca completa se torna inviável.

3.2 Descrição do método *U-curve*

Buscar subconjutos de custo mínimo pelo espaço de busca completo é um problema combinatório muito custoso. O formato de curva em U da restrição da função custo para qualquer cadeia maximal é a premissa para a elaboração de um algoritmo *branch-and-bound*, denominado, aqui, de *algoritmo U-curve*, para lidar com esse tipo de problema. O algoritmo *U-curve* realiza a busca total sobre o reticulado Boolenao completo $\mathcal{L}$ com uma considerável redução no processamento computacional em relação à busca exaustiva.

Sejam A e B elementos do reticulado Booleano $\mathcal{L}$. Um *intervalo* $[A, B]$ de $\mathcal{L}$ é um subconjunto de $\mathcal{L}$ dado por $[A, B] = \{X \in \mathcal{L} : A \subseteq X \subseteq B\}$. Seja R um elemento de $\mathcal{L}$. Neste trabalho, intervalos do tipo $[\emptyset, R]$ e $[R, W]$ são chamados, respectivamente, de *constrição* inferior e superior. A extremidade esquerda de uma restrição inferior e a extremidade direita de uma restrição

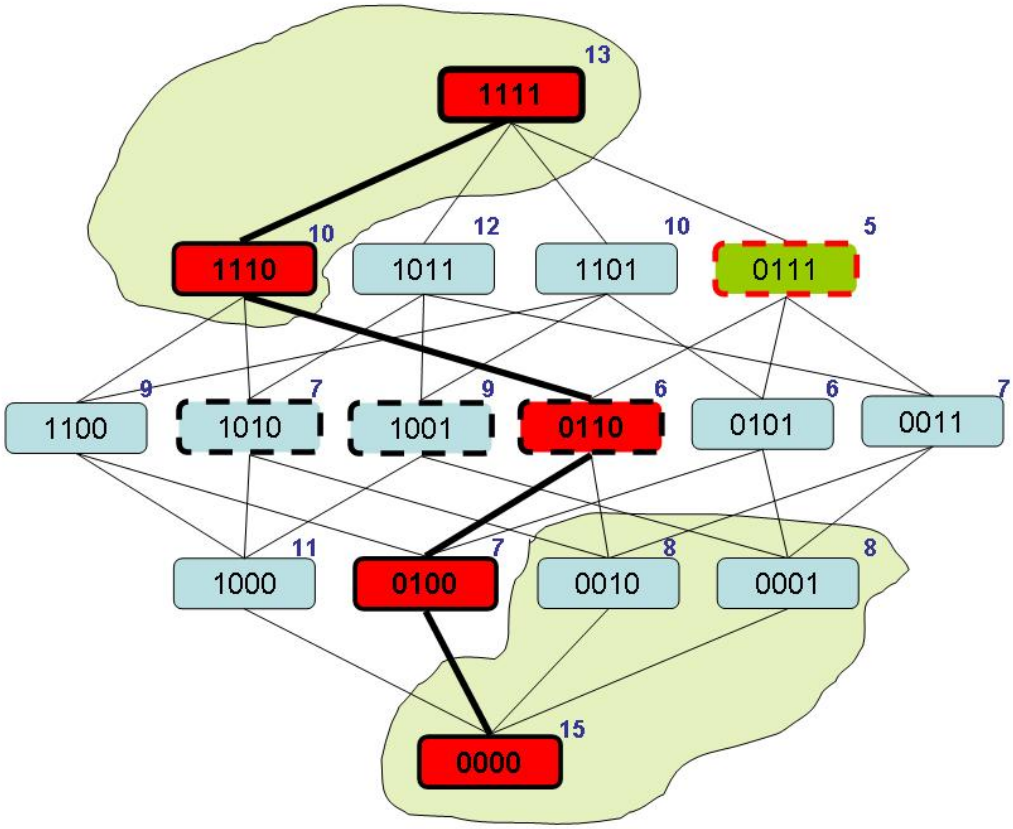


Figura 3.1: O espaço de busca em um reticulado Booleano de ordem 4. $\mathcal{X}$ é um *poset* obtido de $\mathcal{L}$, onde $\mathcal{X} = \mathcal{L} - \{0000, 0010, 0001, 1110, 1111\}$

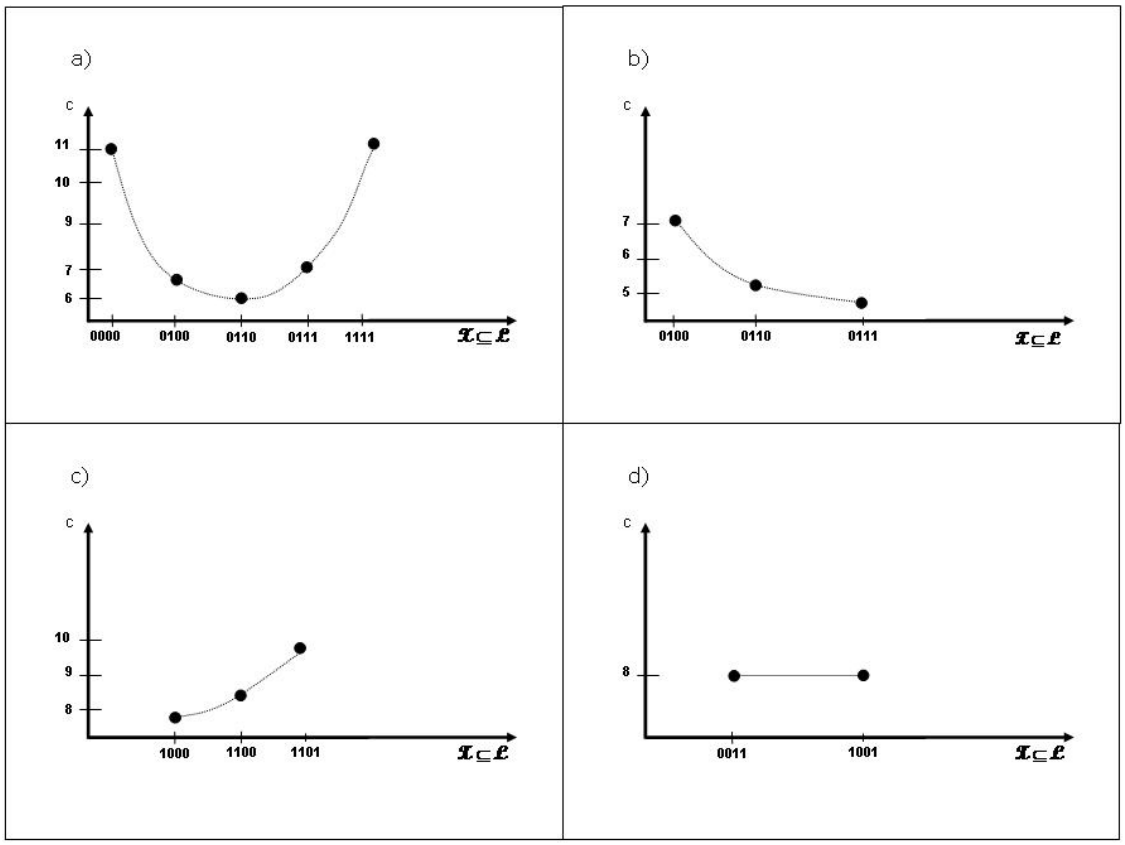


Figura 3.2: As quatro possíveis representações da função custo c restritas a algumas cadeias maximais em $\mathcal{L}$ e em $\mathcal{X} \subseteq \mathcal{L}$ da Figura 3.1

superior são chamadas, respectivamente, de *restrição* inferior e superior. Sejam $\mathcal{R}_L$ e $\mathcal{R}_U$ as representações, respectivamente, das coleções das restrições inferiores e superiores, o espaço de busca será o *poset* $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ obtido eliminando-se a coleção de restrições inferiores e superiores do reticulado $\mathcal{L}$, isto é, $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U) = \mathcal{L} - \bigcup\{[\emptyset, R] : R \in \mathcal{R}_L\} - \bigcup\{[R, W] : R \in \mathcal{R}_U\}$. Em casos onde apenas as restrições inferiores ou superiores são eliminadas, o espaço resultante é denotado, respectivamente, por $\mathcal{X}(\mathcal{R}_L)$ e $\mathcal{X}(\mathcal{R}_U)$, e dado, respectivamente, por $\mathcal{X}(\mathcal{R}_L) = \mathcal{L} - \bigcup\{[\emptyset, R] : R \in \mathcal{R}_L\}$ e $\mathcal{X}(\mathcal{R}_U) = \mathcal{L} - \bigcup\{[R, W] : R \in \mathcal{R}_U\}$.

O espaço de busca é explorado por um algoritmo iterativo que, a cada iteração, explora um pequeno subconjunto de $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$, calcula um *mínimo local*, atualiza a lista de elementos mínimos encontrados e extende ambos conjuntos de restrições, eliminando com isso a região recém explorada. O algoritmo é iniciado com uma lista vazia de elementos mínimos, assim como os conjunto de restrições inferiores e superiores. A execução do algoritmo é feita até que o espaço de busca seja totalmente explorado, isto é, até que $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ fique vazio. O subconjunto de $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ explorado a cada iteração é uma cadeia que pode ser contruída em sentido inferior-superior ou superior-inferior do reticulado. O algoritmo 1 descreve esse processo. O processo de *seleção de direção* (linha 5) é um procedimento que pode usar um método aleatório ou adaptativo. Para método randômico, uma probabilidade estática é definida para seleção de uma das direções possíveis. Já o método adaptativo calcula uma nova probabilidade a cada uma das direções, atribuindo maior valor para a direção inferior-superior ou superior-inferior se a maioria dos mínimos locais, encontrados até o momento, estão, respectivamente, mais perto da base ou do topo do reticulado.

Algoritmo 1 Algoritmo-U-curve()

```
1:  M ⇐ ∅
2:  R_L ⇐ ∅
3:  R_U ⇐ ∅
4:  while X(R_L, R_U) ≠ ∅ do
5:     direcao ⇐ Selecao-Direcao()
6:     if direcao is UP then
7:        Inferior-Superior(R_L, R_U)
8:     else
9:        Superior-Inferiror(R_L, R_U)
10:    end if
11: end while
```

Um elemento C do *poset* $\mathcal{X} \subseteq \mathcal{L}$ é chamado de *elemento minimal* de $\mathcal{X}$ se não existe outro elemento C' de $\mathcal{X}$ com $C' \subset C$. Um elemento D do *poset* $\mathcal{X} \subseteq \mathcal{L}$ é chamado de *elemento maximal* de $\mathcal{X}$ se não existe outro elemento D' de $\mathcal{X}$ com $D \subset D'$. Na Figura 3.1, os elementos minimais de $\mathcal{X}(\mathcal{R}_L)$ são: 1000, 0100 e 0011, e os maximais são: 1011, 1101 e 0111. Se a direção inferior-superior é selecionada o processo Inferior-Superior é executado (Algoritmo 2):

- O procedimento Elemento-Minimal calcula um *elemento minimal* B do *poset* $\mathcal{X}(\mathcal{R}_L)$. Somente o conjunto de restrições inferiores é utilizado para o cálculo desse elemento minimal. Um elemento B é *coberto* pelo conjunto de restrições inferiores $\mathcal{R}_L$ se $\exists R \in \mathcal{R}_L : B \subseteq R$, e B é *coberto* pelo conjunto de restrições superiores $\mathcal{R}_U$ se $\exists R \in \mathcal{R}_L : R \subseteq B$. Quando o elemento B calculado é coberto por uma restrição superior este é descartado, isto é,

o conjunto de restrições inferiores é atualizado com B e uma nova iteração do algoritmo pode ser iniciada (linhas 1-5).

- O processo de construção da cadeia na direção inferior-superior começa com um elemento minimal B e adiciona novos elementos ao selecionar randomicamente um elemento na lista de *elementos adjacentes* superiores ao último adicionado e que pertencem ao *poset* atual $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$. Esse processo continua até que a *condição U-curve* seja encontrada, isto é, até que o último elemento a ser adicionado à cadeia (B) tenha custo maior que o anterior (M) (linhas 7-11).

- Nesse ponto, o elemento M é o elemento mínimo da cadeia construída e A e B são, respectivamente, os elementos adjacentes inferior e superior ao elemento M, isto é, $A \subset M \subset B$ e, por construção, $c(A) \leq c(M) \leq C(B)$. Note que A é equivalente ao conjunto vazio ou B ao conjunto W, se M não têm, respectivamente, elemento adjacente inferior ou superior em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$. Podemos provar que qualquer elemento C de $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$, com $C \subset A$, possui custo maior que A, e qualquer elemento D de $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$, com $B \subset D$, possui custo maior que B. Usando essa propriedade, os conjuntos de restrições inferiores e superiores podem ser atualizados, respectivamente, por A e B (linhas 12-17). A Figura 3.3 apresenta uma representação esquemática de uma primeira iteração do algoritmo e os elementos contidos nas constrições $[\emptyset, A = 1 \ldots 1010 \ldots 0]$ e $[B = 1 \ldots 11110 \ldots 0, W]$.

- A lista de resultados pode ser atualizada com M (linha 18) , isto é, M será incluído na lista de resultados se ele tiver custo menor (ou igual) aos elementos já armazenados nela. A lista de resultados pode ainda guardar não apenas os elementos de custo mínimo mas, também, uma lista com um número parametrizado de elementos com os custos mais baixos.

- Visando prevenir o reprocessamento do elemento M, um procedimento recursivo denominado de *esgotamento do mínimo* é executado (linha 19).

Se a direção superior-inferior for a selecionada, o processo Superior-Inferior é executado (Algoritmo 3). Note a dualidade desse processo com o anterior. Ele inicia com um elemento *maximal* B e constrói uma cadeia em sentido superior-inferior a partir desse elemento.

Um elemento é denominado de *mínimo esgotado* em $\mathcal{L}$ se todos seus elementos adjacentes (superiores e inferiores) possuem custo maior que ele. Essa definição pode ser extendida ao *poset* $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$, isto é, todos os elementos adjacentes (superiores e inferiores) em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ possuem custo maior que ele. Na Figura 3.1 podemos ver que os elementos 1010, 1001 e 0111 são mínimos esgotados em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$, mas 1001 não é um mínimo esgotado em $\mathcal{L}$. Neste trabalho, o termo mínimo esgotado será aplicado sempre se referindo ao *poset* atual $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$.

O procedimento *esgotamento do mínimo* (Algoritmo 4) é um processo recursivo que visita todos os elementos vizinhos de um dado elemento M e os transforma em mínimos esgotados no *poset* atual $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$. Para a recursão, o processo usa uma pilha $\mathcal{S}$. A pilha $\mathcal{S}$ é iniciada empilhando-se M à ela e o processo é executado enquanto $\mathcal{S}$ não for vazia (linhas 2-22). Em cada iteração, o procedimento processa o elemento T no topo de $\mathcal{S}$: todos os elementos adjacentes (superiores e inferiores) a T em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ que ainda não estão empilhados em $\mathcal{S}$ são checados. Se o custo de um elemento A adjacente a T é menor (ou igual) ao custo de T, então esse elemento é empilhado em $\mathcal{S}$. Se o custo de A for maior que o custo de T, então um dos conjuntos de

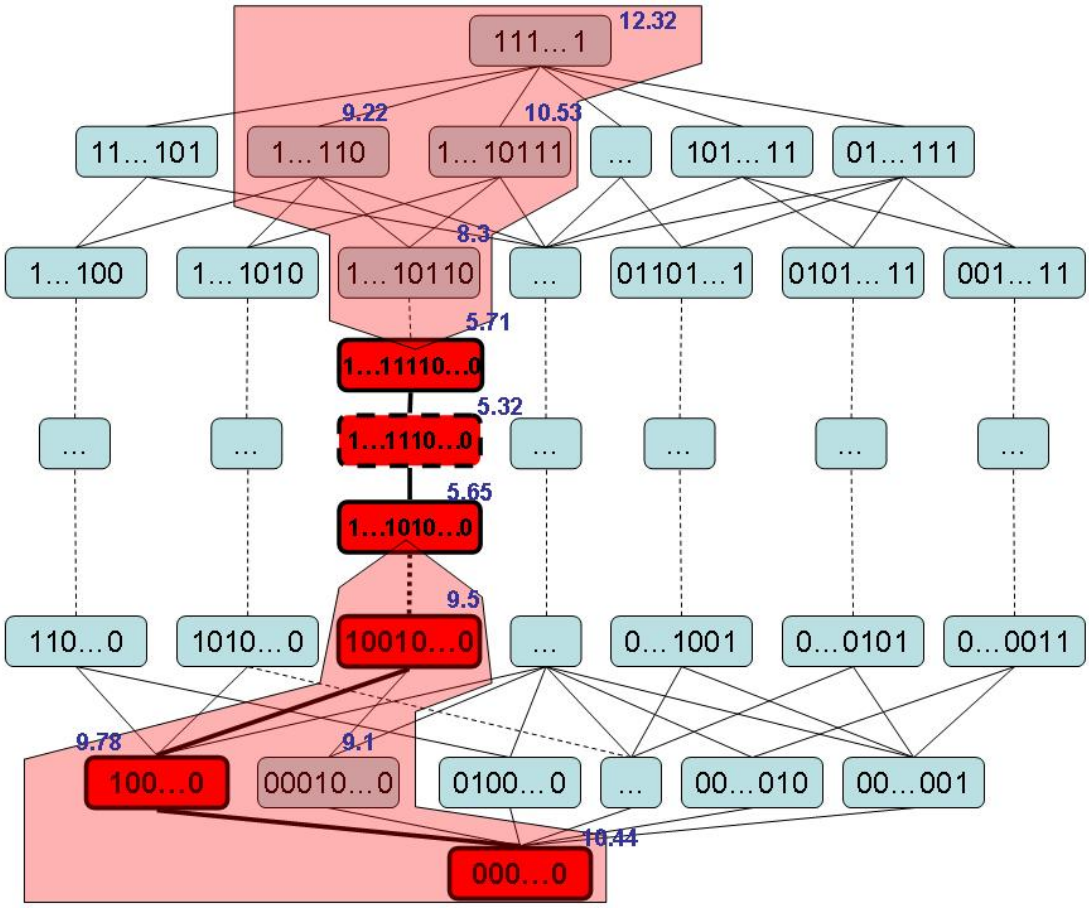


Figura 3.3: Representação esquemática de um passo do algoritmo *U-curve*. As áreas em destaque representam os elementos contidos nas restrições inferiores e superiores

Algoritmo 2 Inferior-Superior(ElementSet $\mathcal{R}_L$, ElementSet $\mathcal{R}_U$)

1: $B \Leftarrow$ Elemento-Minimal($\mathcal{R}_L$)
2: **if** B é coberto por $\mathcal{R}_U$ **then**
3: Atualiza-Restricao-Inferior(B, $\mathcal{R}_L$)
4: **return**
5: **end if**
6: $M \Leftarrow$ **null**
7: **repeat**
8: $A \Leftarrow M$
9: $M \Leftarrow B$
10: $B \Leftarrow$ Seleciona-Adjacente-Superior(M, $\mathcal{R}_L$, $\mathcal{R}_U$)
11: **until** $c(B) > c(M)$ or $B =$ **null**
12: **if** $A \neq$ **null then**
13: Atualiza-Restricao-Inferior(A, $\mathcal{R}_L$)
14: **end if**
15: **if** $B \neq$ **null then**
16: Atualiza-Restricao-Superior(B, $\mathcal{R}_U$)
17: **end if**
18: Atualiza-Lista-Resultados(M)
19: Esgotamento-Minimo(M, $\mathcal{R}_L$, $\mathcal{R}_U$)

Algoritmo 3 Superior-Inferior(ElementSet $\mathcal{R}_L$, ElementSet $\mathcal{R}_U$)

1: $B \Leftarrow$ Elemento-Maximal($\mathcal{R}_U$)
2: **if** B é coberto por $\mathcal{R}_L$ **then**
3: Atualiza-Restricao-Superior(B, $\mathcal{R}_L$)
4: **return**
5: **end if**
6: $M \Leftarrow$ **null**
7: **repeat**
8: $A \Leftarrow M$
9: $M \Leftarrow B$
10: $B \Leftarrow$ Seleciona-Adjacente-Inferior(M, $\mathcal{R}_L$, $\mathcal{R}_U$)
11: **until** $c(B) > c(M)$ or $B =$ **null**
12: **if** $A \neq$ **null then**
13: Atualiza-Restricao-Superior(A, $\mathcal{R}_U$)
14: **end if**
15: **if** $B \neq$ **null then**
16: Atualiza-Restricao-Inferior(B, $\mathcal{R}_L$)
17: **end if**
18: Atualiza-Lista-Resultados(M)
19: Esgotamento-Minimo(M, $\mathcal{R}_L$, $\mathcal{R}_U$)

Algoritmo 4 Esgotamento-Minimo(Element M, ElementSet $\mathcal{R}_L$, ElementSet $\mathcal{R}_U$)

```
1: Empilha M em S
2: while S não é vazio do
3:    T ⇐ Topo(S)
4:    MinimoEsgotado ⇐ verdadeiro
5:    for all A adjacente de T em X(R_L, R_U) e A ∉ S do
6:       if c(A) ≤ c(T) then
7:          Empilha A em S
8:          MinimoEsgotado ⇐ falso
9:       else
10:         if A é adjacente superior a T then
11:            Atualiza-Restricao-Superior(A, R_U)
12:         else
13:            Atualiza-Restricao-Inferior(A, R_L)
14:         end if
15:      end if
16:   end for
17:   if MinimoEsgotado then
18:      Desempilha T de S
19:      Atualiza-Lista-Resultados(T)
20:      Atualiza-Restricao-Inferior(T, R_L)
21:      Atualiza-Restricao-Superior(T, R_U)
22:   end if
23: end while
24: return
```

restrições pode ser atualizada com A: o conjunto de restrições inferiores, se A é adjacente inferior a T, e o conjunto de restrições superiores, se A é adjacente superior a T (linhas 5-16). Se T é mínimo esgotado em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$, isto é, não existe elemento adjacente A em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ com custo menor do que T, então T é desmpilhado de $\mathcal{S}$, e ambos, os conjuntos de restrições e a lista de resultados, são atualizados com T (linhas 19-21). Ao final desse procedimento todos os elementos processados se transformam em mínimos esgotados no *poset* atualizado $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$.

A Figura 3.4 mostra uma representação gráfica do processo de esgotamento do mínimo. A Figura 3.4-A mostra o processo de construção de uma cadeia na direção inferior-superior. Essa cadeia possui suas arestas enfatizadas na figura. O elemento $M = 010101$ (em laranja) possui o custo mínimo sobre a cadeia. Os elementos em preto são os elementos eliminados do espaço de busca pelas restrições obtidas pelos elementos adjacentes (inferiores e superiores) ao mínimo local M. A pilha é iniciada com o elemento M. A Figura 3.4-B mostra a primeira iteração do processo de esgotamento do mínimo. As arestas em vermelho e os elementos em vermelho indicam os elementos adjacentes a M (topo da pilha) que possuem custo menor (ou igual) a ele. Esses elementos (010001 e 010111), por sua vez, são empilhados. Os elementos adjacentes a M, com custo maior do que ele, atualizam os conjuntos de restrições, isto é, o elemento adjacente inferior 000101 atualiza o conjunto de restrições inferiores, e o elemento adjacente superior 000101 atualiza o conjunto de restrições superiores. A Figura 3.4-C mostra a segunda iteração: os elementos adjacentes 010011 e 000111 com custos menores (ou iguais) ao novo topo 010111 são empilhados, e os elementos adjacentes 010110 e 011111 com custo maior que 010111 atualizam, respectivamente, os conjuntos de restrições inferiores e superiores. Na Figura 3.4-D o elemento 000111 é um mínimo esgotado (em cinza) em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ e ele é desempilhado. Na Figura 3.4-E os elementos eliminados pela novas constrições $[\emptyset, 000111]$ e $[000111, W]$ são destacados na cor preta. A partir desse ponto, 010011 é um mínimo esgotado (em cinza) em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ e ele é desempilhado. Da Figura 3.4-F à Figura 3.4-H, todos os elementos da pilha são desempilhados e os elementos removidos do espaço de busca pelas novas restrições são destacados em preto. A Figura 3.4-H mostra todos os elementos removidos (em preto) do espaço de busca pela execução do processo de esgotamento do mínimo uma única vez.

Os procedimentos para o cálculo dos elementos minimais e maximais e para a atualização dos conjuntos de restrições inferiores e superiores serão discutidos na próxima Seção.

3.3 Fundamentos Matemáticos

Esta Seção introduz os fundamentos matemáticos de alguns módulos do algoritmo *U-curve*.

3.3.1 Procedimento de obtenção dos elementos minimais e maximais

Um grande número de elementos minimais e maximais dos *posets* $\mathcal{X}(\mathcal{R}_L)$ e $\mathcal{X}(\mathcal{R}_U)$, respectivamente, podem ser obtidos em um passo do algoritmo. Aqui apresentaremos uma solução simples para obtê-los.

Seja $\mathcal{L}$ o reticulado Booleano completo de ordem n e A um elemento de $\mathcal{L}$. O elemento A^c representa o *conjunto complementar* ao conjunto representado pelo elemento A em $\mathcal{L}$, isto é, $A^c = W \setminus \{A\}$. Por exemplo, se $A = 00101$ então $A^c = 11010$.

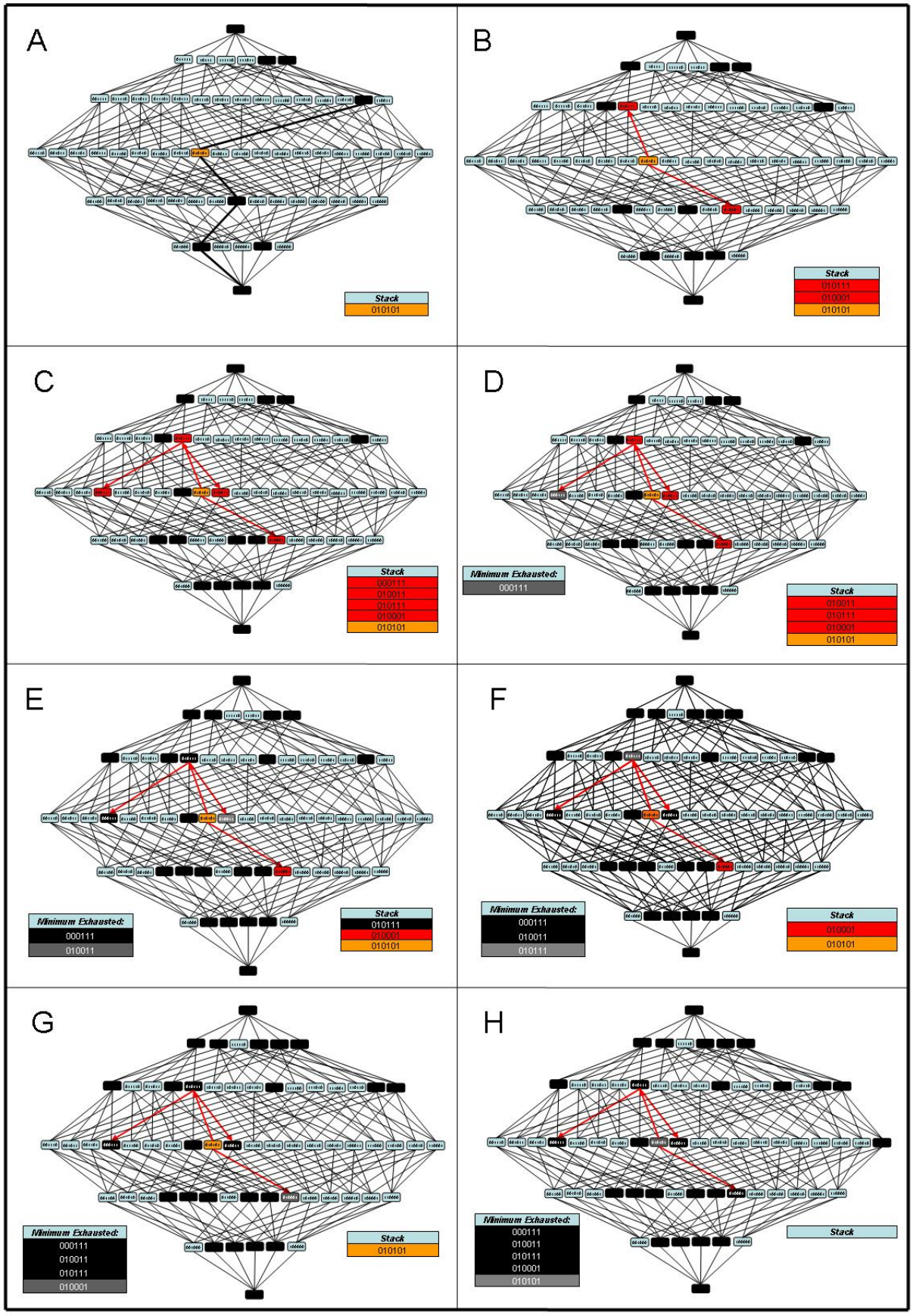


Figura 3.4: Representação gráfica do processo de esgotamento do mínimo

Teorema 1. *Para todo* $A \in \mathcal{X}(\mathcal{R}_L)$,

$$A \in \mathcal{X}(\mathcal{R}_L) \Leftrightarrow A \cap R^c \neq \emptyset, \forall R \in \mathcal{R}_L.$$

Demonstração.

$$\begin{aligned} A \in \mathcal{X}(\mathcal{R}_L) &\Leftrightarrow A \in \mathcal{L} - \bigcup\{[\emptyset, R] : R \in \mathcal{R}_L\} \\ &\Leftrightarrow A \notin \bigcup\{[\emptyset, R] : R \in \mathcal{R}_L\} \\ &\Leftrightarrow A \notin [\emptyset, R], \forall R \in \mathcal{R}_L \\ &\Leftrightarrow A \not\subseteq R, \forall R \in \mathcal{R}_L \\ &\Leftrightarrow A \cap R^c \neq \emptyset, \forall R \in \mathcal{R}_L \end{aligned}$$

□

O algoritmo 5 implementa o procedimento de *construção do minimal*. Ele constrói um elemento minimal C do *poset* $\mathcal{X}(\mathcal{R}_L)$. O processo começa com $C = (\underbrace{1 \ldots 1}_{n})$ e $S = (\underbrace{1 \ldots 1}_{n})$, e executa n iterações (linhas 3-16) tentando remover componentes de C. A cada iteração, uma componente $k, k \in \{1, \ldots, n\}$ é selecionada de S. S previne multi-seleção de uma mesma componente. Se um elemento C' resultante de C pela remoção da componente k está contido em $\mathcal{X}(\mathcal{R}_L)$, então C é atualizado por C' (linhas 7-15).

Algoritmo 5 Elemento-Minimal(ElementSet $\mathcal{R}_L$)

```
1:  C ⇐ 1...1 (n)
2:  S ⇐ 1...1 (n)
3:  while S ≠ 0...0 (n) do
4:     k ⇐ índice aleatório em {1,...,n} onde S[k] = 1
5:     S[k] ⇐ 0
6:     C' ⇐ C \ {k}
7:     RemoveComponente ⇐ verdadeiro
8:     for all R in R_L do
9:        if R^c ∩ C' = ∅ then
10:          RemoveComponente ⇐ falso
11:       end if
12:    end for
13:    if RemoveComponente then
14:       C ⇐ C'
15:    end if
16: end while
17: return C
```

O elemento minimal calculado é igual a $\underbrace{1\ldots1}_{n}$ quando $\mathcal{R}_L = \{\underbrace{1\ldots1}_{n}\}$. Neste ponto, o *poset* $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ é vazio e o algoritmo pára na próxima iteração.

O próximo teorema demonstra a validade do Algoritmo 5.

Teorema 2. *O elemento C de $\mathcal{X}(\mathcal{R}_L)$ obtido pelo processo de construção do minimal (Algoritmo 5) é um elemento minimal em $\mathcal{X}(\mathcal{R}_L)$.*

Demonstração. Analisando os passos do processo de construção do minimal temos:

- As linhas 7-15 garantem que em qualquer iteração do processo o elemento resultante C está contido em $\mathcal{X}(\mathcal{R}_L)$, isto é, ele só é alterado quando o elemento C' satisfaz a condição do Teorema 1.
- Sejam $C_1, \ldots, C_n$ a seqüência dos elementos resultantes a cada iteração i ($i = 1, \ldots, n$) do processo e $C_0 = \underbrace{1\ldots1}_{n}$ o elemento inicial. Como um índice k é escolhido para ser removido de C_{i-1}(linhas 4-6) a cada iteração i, isso implica que $C_n \subseteq C_{n-1} \subseteq \ldots \subseteq C_0$.
- Provar que o elemento resultante C_n é minimal em $\mathcal{X}(\mathcal{R}_L)$ é equivalente a provar que $\forall l \in C_n, C_n \setminus \{l\} \notin \mathcal{X}(\mathcal{R}_L)$.
- Seja $k = l, l \in C_n$ e i a iteração do processo quando o índice l é escolhido para ser removido de C_{i-1}. $C_n \subseteq C_i$ e $l \in C_n$, implica que $l \in C_i$, isto é, l não pode ser removido de C_{i-1} ao final da iteração i. Isso é evitado pelo algoritmo (linhas 8-12) quando existe um elemento $R \in \mathcal{R}_L$ com $R^c \cap (C_{i-1} \setminus \{l\}) = \emptyset$. Como $C_n \setminus \{l\} \subseteq C_{i-1} \setminus \{l\}$, então $R^c \cap (C_n \setminus \{l\}) = \emptyset$ e, pelo Teorema 1, $C_n \setminus \{l\} \notin \mathcal{X}(\mathcal{R}_L)$. Isso implica que C_n é um elemento minimal em $\mathcal{X}(\mathcal{R}_L)$.

□

O processo para a obtenção dos elementos maximais é *dual* ao minimal.

Teorema 3. *Para todo $A \in \mathcal{X}(\mathcal{R}_U)$,*

$$A \in \mathcal{X}(\mathcal{R}_U) \Leftrightarrow A^c \cap R \neq \emptyset, \forall R \in \mathcal{R}_U.$$

Demonstração.

$$\begin{aligned}
A \in \mathcal{X}(\mathcal{R}_U) &\Leftrightarrow A \in \mathcal{L} - \bigcup\{[R, W] : R \in \mathcal{R}_U\} \\
&\Leftrightarrow A \notin \bigcup\{[R, W] : R \in \mathcal{R}_U\} \\
&\Leftrightarrow A \notin [R, W], \forall R \in \mathcal{R}_U \\
&\Leftrightarrow R \not\subseteq A, \forall R \in \mathcal{R}_U \\
&\Leftrightarrow R \cap A^c \neq \emptyset, \forall R \in \mathcal{R}_U
\end{aligned}$$

□

O algoritmo 6 implementa o procedimento de *construção do maximal*. Podemos notar a sua dualidade com o da construção do minimal. O processo começa com $C = (\underbrace{0 \ldots 0}_{n})$ e $S = (\underbrace{1 \ldots 1}_{n})$, e executa n iterações (linhas 3-16) tentando adicionar componentes ao elemento C. A cada iteração, uma componente $k, k \in \{1, \ldots, n\}$ é selecionada de S. Se um elemento C' resultante de C pela remoção da componente k está contido em $\mathcal{X}(\mathcal{R}_U)$, então C é atualizado por C' (linhas 7-15).

Algoritmo 6 Elemento-Maximal(ElementSet $\mathcal{R}_U$)

1: $C \Leftarrow \underbrace{0 \ldots 0}_{n}$
2: $S \Leftarrow \underbrace{1 \ldots 1}_{n}$
3: **while** $S \neq \underbrace{0 \ldots 0}_{n}$ **do**
4: $k \Leftarrow$ índice aleatório em $\{1, \ldots, n\}$ onde $S[k] = 1$
5: $S[k] \Leftarrow 0$
6: $C' \Leftarrow C \cup \{k\}$
7: *AdicionaComponente* $\Leftarrow$ verdadeiro
8: **for all** R in $\mathcal{R}_U$ **do**
9: **if** $R \cap C'^c = \emptyset$ **then**
10: *AdicionaComponente* $\Leftarrow$ falso
11: **end if**
12: **end for**
13: **if** *AdicionaComponente* **then**
14: $C \Leftarrow C'$
15: **end if**
16: **end while**
17: **return** C

Também de forma *dual* o Teorema 4 demonstra a validade do Algoritmo 6.

Teorema 4. *O elemento C de $\mathcal{X}(\mathcal{R}_U)$ obtido pelo processo de construção do maximal (Algoritmo 6) é um elemento maximal em $\mathcal{X}(\mathcal{R}_U)$.*

Demonstração. Analisando os passos do processo de construção do maximal temos:

- As linhas 7-15 garantem que em qualquer iteração do processo o elemento resultante C está contido em $\mathcal{X}(\mathcal{R}_U)$, isto é, ele só é alterado quando o elemento C' satisfaz a condição do Teorema 3.
- Sejam $C_1, \ldots, C_n$ a seqüência dos elementos resultantes a cada iteração i $(i = 1, \ldots, n)$ do processo e $C_0 = \underbrace{0 \ldots 0}_{n}$ o elemento inicial. Como um índice k é escolhido para ser adicionado a C_{i-1}(linhas 4-6) a cada iteração i, isso implica que $C_0 \subseteq C_1 \subseteq \ldots \subseteq C_n$.
- Provar que o elemento resultante C_n é maximal em $\mathcal{X}(\mathcal{R}_L)$ é equivalente a provar que $\forall l \notin C_n, C_n \cup \{l\} \notin \mathcal{X}(\mathcal{R}_U)$.

- Seja $k = l, l \notin C_n$ e i a iteração do processo quando o índice l é escolhido para ser adicionado a C_{i-1}. $C_i \subseteq C_n$ e $l \notin C_n$ implica que $l \notin C_i$, isto é, l não pode ser adicionado a C_{i-1} ao final da iteração i. Isso é evitado pelo algoritmo (linhas 8-12) quando existe um elemento $R \in \mathcal{R}_L$ com $R \cap (C_{i-1} \cup \{l\})^c = \emptyset$. Como $C_{i-1} \cup \{l\} \subseteq C_n \cup \{l\}$ e $(C_{i-1} \cup \{l\})^c \supseteq (C_n \cup \{l\})^c$, então $R \cap (C_n \cup \{l\})^c = \emptyset$ e, pelo Teorema 3, $C_n \cup \{l\} \notin \mathcal{X}(\mathcal{R}_U)$. Isso implica que C_n é um elemento maximal em $\mathcal{X}(\mathcal{R}_U)$.

□

3.3.2 Atualização dos conjuntos de restrições

Os conjuntos de restrições $\mathcal{R}_L$ e $\mathcal{R}_U$ representam o espaço de busca. Sendo assim, eles são atualizados após cada nova busca pela seguite regra: um elemento A é adicionado ao conjunto de restrições inferiores (ou superiores) se todos os elementos contidos em $[\emptyset, A]$ (ou $[A, W]$) possuem custo maior ou igual a A.

Os conjuntos de restrições são os responsáveis pela redução significativa do espaço de busca. O conjunto de restrições inferiores $\mathcal{R}_L = \{R_1, \ldots, R_m\}$ transforma o espaço de busca $\mathcal{L}$ no espaço restringido $\mathcal{X}(\mathcal{R}_L) = \mathcal{L} - \bigcup\{[\emptyset, R] : R \in \mathcal{R}_L\}$.

O próximo teorema estabelece a condição *U-curve* , a qual permite que o processo de construção da cadeia pare e os conjuntos de restrições sejam atualizados.

Teorema 5. *Sejam $C_0, ..., C_{k-1}, C_k$ a cadeia construída pelo Algoritmo 2 (ou sua versão dual). Seja c a função custo do reticulado $\mathcal{L}$ em $\mathbb{R}$ decomponível em curvas em U e $c(C_k) > c(C_{k-1})$. É verdade que:*

$$\forall A \in \mathcal{L}, C_k \subseteq A \Rightarrow c(A) \geq c(C_k).$$

Demonstração. Suponha que $\exists B \in \mathcal{L}, C_k \subseteq B$ e $c(B) < c(C_k)$. Por hipótese, c é uma função decomponível em curvas em U e como $C_{k-1} \subseteq C_k \subseteq B$, então $\max(c(C_{k-1}), c(B)) > c(C_k)$. Por outro lado, $\max(c(C_{k-1}), c(B))$ é ou $c(C_{k-1})$ ($c(C_{k-1}) < c(C_k)$, por hipótese) ou $c(B)$ ($c(B) < c(C_k)$, por suposição), logo $\max(c(C_{k-1}), c(B)) < c(C_k)$, contradizendo a hipótese. □

Com uma demonstração semelhante a apresentada para o Teorema 5, pode-se provar, também, que todos os elementos de $\mathcal{L}$ contidos em C_{k-2} possuem custo maior ou igual a ele. A Figura 3.3 mostra a cadeia obtida pelo processo de contrução da cadeia e o *poset* resultante. Os elementos em destaque possuem custos maiores que os elementos $C_k = (1 \ldots 11110 \ldots 0)$ ou $C_{k-2} = (1 \ldots 1010 \ldots 0)$.

O Algoritmo 7 descreve o processo de atualização do conjunto de restrições inferiores a partir de um elemento A. Se A já é coberto por $\mathcal{R}_L$, isto é, existe um elemento de $\mathcal{R}_L$ que contém A, então o processo se encerra (linhas 1-3). Caso contrário, todos elementos em $\mathcal{R}_L$ contidos em A são removidos de $\mathcal{R}_L$ e A é adicionado a $\mathcal{R}_L$ (linhas 4-9). Esse procedimento possui o efeito de compactar o conjunto de restrições sem alterar o *poset* resultante $\mathcal{X}(\mathcal{R}_L)$, já que as restrições removidas estão contidas em A.

Algoritmo 7 Atualiza-Restricao-Inferior(Element A, ElementSet $\mathcal{R}_L$)

1: **if** existe R de $\mathcal{R}_L$ onde $A \subseteq R$ **then**
2: **return**
3: **end if**
4: **for all** R in $\mathcal{R}_L$ **do**
5: **if** $R \subseteq A$ **then**
6: $\mathcal{R}_L = \mathcal{R}_L \setminus \{R\}$
7: **end if**
8: **end for**
9: $\mathcal{R}_L = \mathcal{R}_L \cup \{A\}$
10: **return**

A atualização do conjunto de restrições superiores é *dual* ao de inferiores e, nesse caso, procuramos elementos que estejam contidos em A ao invés de elementos que contêm A para indicar se A deve ser inserido em $\mathcal{R}_L$. O Algoritmo 8 descreve o processo de atualização do conjunto de restrições superiores a partir de um elemento A.

Algoritmo 8 Atualiza-Restricao-Superior(Element A, ElementSet $\mathcal{R}_U$)

1: **if** existe R de $\mathcal{R}_U$ onde $A \supseteq R$ **then**
2: **return**
3: **end if**
4: **for all** R in $\mathcal{R}_U$ **do**
5: **if** $R \supseteq A$ **then**
6: $\mathcal{R}_U = \mathcal{R}_U \setminus \{R\}$
7: **end if**
8: **end for**
9: $\mathcal{R}_U = \mathcal{R}_U \cup \{A\}$
10: **return**

3.3.3 Esgotamento do mínimo

O cálculo computacional da função custo é em geral muito custoso. Sendo assim, é desejável que cada elemento seja visitado e seu custo computado uma única vez. Uma maneira de evitar esse reprocessamento é aplicar o procedimento de esgotamento do mínimo. Esse procedimento é uma função recursiva (Algoritmo 4). Ele usa uma pilha $\mathcal{S}$ para processar recursivamene todos os elementos próximos de um dado elemento M contidos no *poset* atual $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$. Em cada recursão, os elementos adjacentes (superiors e inferiores) ao elemento do topo de $\mathcal{S}$, em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$, e não em $\mathcal{S}$, são visitados. Os elementos adjacentes com custo maior que T são elementos que satisfazem a condição *U-curve* e podem atualizar os respectivos conjuntos de restrições e, conseqüentemente, serem removidos do espaço de busca. Os elementos adjacentes com custo menor (ou igual) a T são empilhados em $\mathcal{S}$ para serem processados depois. Note que esses elementos não são reprocessados ao longo do método: eles ou estão em um conjunto de restriçõe, ou não se repetem em $\mathcal{S}$. Se T é um elemento mínimo esgotado em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$, então T é removido de $\mathcal{S}$. Depois que o processo é finalizado, todos os elementos são removidos do

poset resultante $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$, o que previne seus reprocessamentos. O fato de um elemento não poder ser reprocessado implica no fato de que o procedimento termina após alguns passos (sem entrar em ciclo infinito), isto é, o número de elemenos em $\mathcal{X}(\mathcal{R}_L, \mathcal{R}_U)$ é finito (limitado pelo número de elementos do espaço total), e ele é um limitante superior para o número de passos do procedimento. Em alguns casos, esse procedimento pode ter que processar um número muito grande de elementos, e algumas heurísticas devem ser implementadas para lidar com eles. Por exemplo: parar a recursão após alguns passos sem a alteração do valor encontrado para o custo mínimo.

O procedimento de esgotamento do mínimo adiciona uma outra propriedade interessante ao algoritmo *U-curve*. Se a função custo em uma cadeia maximal descreve uma curva em U com *oscilações*, como as ilustradas na Figura 3.5-A, o algoritmo *U-curve* pode perder algum mínimo local (no caso, o elemento mínimo local após as oscilações tem custo menor que o anterior). Apesar disso, esse mínimo não é perdido se existir outra cadeia, com uma função custo descrevendo realmente uma curva em forma de U, contendo os dois mínimos locais. A Figura 3.5-B mostra uma cadeia alternativa (em vermelho) para alcançar o elemento mínimo (em preto). Note que o mínimo local (em amarelo) está contido em ambas as cadeias. Essas *cadeias alternativas* para atingir os mínimos locais são possíveis pelo procedimento de esgotamento exatamente quando aplicada ao primeiro mínimo. Sendo assim, o procedimento de esgotamento do mínimo permite relaxar a classe de problemas abordados pelo algoritmo *U-curve*.

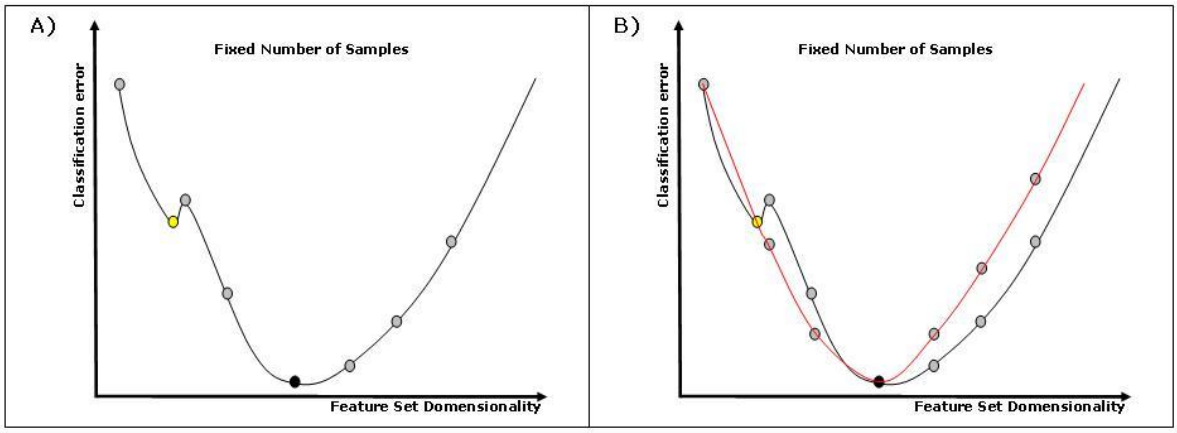


Figura 3.5: Exemplos de curvas de erro com oscilações e caminhos alternativos

3.4 Resultados experimentais

Nesta Seção, apresentaremos alguns resultados de aplicações do algoritmo proposto para a seleção de características. Seu desempenho foi comparado com o desempenho do SFFS ("*Sequential Floating Forward Selection*") [32]. Duas aplicações foram estudadas: desenho de *W-operadores* [22] e *identificação de preditores* para redes genéticas. Em ambos os casos, foi atribuído valor 2 para o parâmetro delta do SFFS.

A função custo é a entropia condicional média com o mesmo método de estimativa usado em [3] e dado por:

$$\hat{E}[H(Y|\mathbf{X}_{\mathcal{Z}})] = \frac{N}{t} + \sum_{\mathbf{x}_{\mathcal{Z}} \in \mathbf{X}_{\mathcal{Z}}: P(\mathbf{x}_{\mathcal{Z}}) > \frac{1}{t}} P(\mathbf{x}_{\mathcal{Z}}) H(Y|\mathbf{x}_{\mathcal{Z}}),$$

onde $H(X) = -\sum_{x \in X} P(x) log_c P(x)$ (entropia de Shannon com base logarítmica igual ao número de possíveis classes c), t é o número total de amostras, $\mathcal{Z}$ é o subconjunto dos índices de características e N é o número total de instâncias com $P(\mathbf{x}_{\mathcal{Z}}) = \frac{1}{t}$. Nessa equação, as instâncias raramente observadas são aquelas com apenas uma observação no conjunto de treinamento, o que significa que a probabilidade de todas essas instâncias é $\frac{N}{t}$. Sem essa penalização, tais instâncias teriam entropia condicional igual a zero (mínima possível), pois a massa de probabilidades estaria concentrada em apenas uma classe. Essa penalização consiste em atribuir, em conjunto com as classes, uma distribuição uniforme, já que a entropia da distribuição uniforme é máxima (utilizando logaritmo na base c, a máxima entropia é 1). Essa função custo descreve uma curva em forma de U pelo fato de que: para uma dimensão suficientemente grande, o número de instâncias com apenas uma observação começa a crescer, crescendo a penalização e, conseqüentemente, crescendo o valor da função custo (novas componentes não adicionam informação suficiente para compensar o erro de estimação).

O problema de seleção de características pode ter funções custo com cadeias que apresentam oscilações e não existe nenhuma garantia teórica da existência de caminhos alternativos para alcançar os mínimos locais além das oscilações. Todavia, esses casos foram testados experimentalmente, e para todos os casos observados, o procedimento de esgotamento do mínimo pode atingir os elementos mínimos locais usando caminhos alternativos. Testamos 100.000 curvas aleatórias para ambos os problemas. Para o problemas desenho de W-operadores, uma média de 24.000 curvas (24%) contém partes com oscilações, e para o problema de classificação biológica uma média de 15.000 curvas (15%) contêm partes com oscilações. Para todas essas curvas com oscilações o procedimento de esgotamento do mínimo encontra o mínimo local por cadeias alternativas .

Os resultados do algoritmo *U-curve* foram divididos em dois conjuntos: UC – até atingir (ou melhorar pela primeira vez) o resultado do SFFS e UCC – até o espaço total de busca ser completamente processado. O algoritmo *U-curve* é *estocástico* e, com isso, ele pode atingir o melhor resultado em diferentes tempos de processamento. Para cada teste, processamos o *U-curve* cinco vezes e os resultados da coluna UC indicam uma média dos cinco processamentos. A máquina usada para os testes foi um AMD Turion 64^{TM} com 2Gb de memória RAM.

O problema de desenho de W-operadores consiste em procurar subconjuntos de uma janela de tamanho n que possuem o menor erro de estimação (custo), isto é, a imagem obtida pelo processamento do operador calculado possui o melhor resultado desejável dentre todos os operadores possíveis. Para nossos testes, foram usadas imagens binárias como amostras com janela de tamanho 16 (4×4) e duas classes (0 e 1). A Tabela 3.1 mostra o resultado do algoritmo *U-curve* em comparação ao SFFS para 20 testes. A segunda coluna mostra qual processo (*U-curve* ou SFFS) atinge o melhor resultado (elemento de menor custo). O algoritmo *U-curve*

melhora o resultado do SFFS em 8 dos 20 testes e atinge resultados iguais nos outros 12 testes (para esses casos, o SFFS atinge o elemento mínimo global). Em todos os testes, o algoritmo *U-curve* atinge o resultado do SFFS (UC) processando um menor número de nós em um tempo menor. A busca completa (UCC) freqüentemente requer mais nós processados (17/20) e tempo de execução (19/20) para processar o espaço inteiro.

Teste	Vencedor	Nós			Tempo(seg.)		
		SFFS	UC	UCC	SFFS	UC	UCC
1	IGUAL	358	73	373	8	2	393
2	IGUAL	333	31	154	7	1	392
3	IGUAL	417	17	137	10	1	393
4	UC	435	58	5,965	9	1	541
5	UC	357	101	223	7	3	385
6	UC	384	66	345	9	2	399
7	IGUAL	302	111	266	6	4	392
8	UC	1,217	158	13,963	21	2	591
9	UC	330	31	274	8	1	385
10	IGUAL	406	113	825	10	4	408
11	IGUAL	329	70	544	7	2	387
12	IGUAL	336	17	17	8	0.5	0.5
13	IGUAL	310	26	26	8	1	384
14	UC	328	67	67	8	4	421
15	IGUAL	425	66	671	8	1	391
16	UC	333	31	151	8	1	377
17	IGUAL	1,257	659	11,253	31	16	717
18	UC	336	39	218	7	1	385
19	IGUAL	296	32	137	6	2	379
20	IGUAL	323	31	151	8	2	376

Tabela 3.1: Comparação entre o resultado do SFFS e o algoritmo *U-curve* em nós calculados e tempo de execução para o desenho de W-operadores

Um problema de classificação biológica pode ser visto como o problema de se encontrar um bom subconjunto de genes preditores de um específico gene-alvo em um experimento temporal de *microarrays*. Os dados utilizados para esses testes foram obtidos do artigo em [26]. Eles foram normalizados e discretizados em 3 níveis, utilizando o mesmo método de [3], O subconjunto de preditores foi obtido de um conjunto de 27 genes e os resultados foram comparados ao SFFS. A Tabela 3.2 mostra os resultados de 10 testes utilizando diferentes genes-alvos e suas comparações com os resultados em SFFS. As colunas apresentadas são as mesmas da Tabela 3.1. A busca completa para esses testes é relativamente grande (2^{27} nós). A heurística SFFS alcançou o melhor elemento, processando um número menor de nós que o algoritmo *U-curve* em apenas 3/10 vezes, e o algoritmo *U-curve* processou o espaço completo encontrando o melhor elemento (7/10) processando mais nós, porém com um tempo de processamento aceitável.

O Apêndice A contém uma parte da fonte do algoritmo desenvolvido em C++ e a descrição dos parâmetros que podem ser utilizados para a sua execução. No DVD anexado ao traba-

Teste	Vencedor	Nós			Tempo(seg.)		
		SFFS	UC	UCC	SFFS	UC	UCC
1	IGUAL	135	777	9.964	0,5	0,6	3,1
2	UC	135	9252	30.724	0,5	2,1	11,2
3	UC	135	1037	9.410	0,5	0,6	3,1
4	UC	164	786	9.276	0,5	0,6	3,1
5	UC	281	247	6.126	0,5	0,6	1,5
6	IGUAL	135	2675	11.031	0,5	0,7	7,3
7	IGUAL	135	998	10.836	0,5	0,6	6,9
8	UC	135	463	5.381	0,5	0,5	1,5
9	UC	135	246	4.226	0,5	0,5	1,5
10	UC	191	474	8.930	0,5	0,5	2,9

Tabela 3.2: Comparação entre o resultado do SFFS e o algoritmo *U-curve* em nós calculados e tempo de execução para o problema de classificação biológica

lho, encontramos: (i) as fontes em C++; (ii) o programa executável para rodar em ambiente Microsoft Windows^{TM}; (iii) os arquivos de dados utilizados nos testes apresentados e (iv) os resultados obtidos.

3.5 Discussão

Desevolvemos uma nova técnica para abordar os problemas de seleção de características: o algoritmo *U-curve*, que retorna elementos de custo mínimo para funções custo decomponíveis em curvas em U (com oscilações de um certo modo ou não) para cadeias maximais. Essa técnica difere das outras, como SFFS, pois ela possui um modelo formal para a estruturação do problema de seleção de características, incluindo um reticulado Booleano para o domínio, o formato de curva em U da função custo para cadeias maximais e eventuais oscilações dessas funções.

O problema de otimização é resolvido por um algoritmo estocástico do tipo *branch-and-bound*, que explora o domínio e a função custo de estruturas particulares. A natureza Booleana do domínio permite representar o espaço de busca por uma coleção de restrições inferiores e superiores. Em cada iteração, um nó para início de cadeia é computado a partir das restrições do espaço de busca. A cadeia explorada a cada iteração é construída a partir desse nó pela escolha de nós adjacentes (superiores ou inferiores). A escolha do início da cadeia geralmente possui vários candidatos, e um deles é selecionado randomicamente. A função custo e a estrutura do domínio permitem cortes no espaço de busca, quando um mínimo local é encontrado. Após a localização de um mínimo local, todos os mínimos locais atingíveis por ele são computados pelo procedimento de esgotamento do mínimo, e os correspondentes cortes são executados. As relações de adjacência e conectividade adotadas são as do espaço de busca do *diagrama de Hesse*: um grafo no qual a conectividade é induzida pela relação de ordem parcial. O procedimento de esgotamento do mínimo evita que um nó seja visitado mais de uma vez e generaliza o algoritmo para funções custo decomponíveis em alguma classe de funções decomponíveis em formas de U com alguma oscilação. Os procedimentos empregados dentro do algoritmo *U-curve* são sustentados por

resultados formais.

O algoritmo do *U-curve* foi aplicado para problemas práticos e comparados com o SFFS. Os experimentos foram concentrados em dois problemas: desenho de W-operadores e redes genéticas. Para o desenho de W-operadores, os mínimos estão situados em níveis intermediários ou superiores do reticulado que representa o espaço de busca, enquanto para redes genéticas os mínimos se encontram em níveis inferiores. Em ambos problemas, os resultados médios do algoritmo *U-curve* foram melhores do que os obtidos pelo SFFS em precisão e, na maior parte das vezes, em desempenho. Os resultados do algoritmo *U-curve* considerados para comparação são compostos por uma média de várias execuções para um mesmo dado de entrada, pois, sendo ele um algoritmo estocástico, ele pode ter diferentes desempenhos para cada execução.

A eficiência do algoritmo *U-curve* depende da posição relativa dos mínimos locais no espaço de busca. O algoritmo é mais eficiente quando os mínimos locais estão próximos às extremidades do espaço de busca. A maior ineficiência é verificada quando os mínimos locais estão próximos ao meio do reticulado.

Os resultados obtidos até o momento são encorajadores, mas a presente versão do algoritmo *U-curve* não é uma solução rápida para problemas de dimensões altas com mínimos locais presentes no meio do reticulado que representa o espaço de busca. Para uma abordagem eficiente desses problemas, a solução da otimização *U-curve* proporciona alguns novos tópicos a serem estudados em pesquisas futuras: desenvolvimento de novos cortes à formulação *branch-and-bound*; desenho e estimação das distribuições dos parâmetros randômicos utilizados na seleção do início de cadeias ou elementos adjacentes para sua construção, com o objetivo de atingir os mínimos com menos processamento; construir versões paralelizáveis do algoritmo; analizar possíveis erros cometidos pela técnica *U-curve* nos problemas de seleção de características, devido à parcial aderência às hipóteses do *U-curve*; e outros.

Com tudo isso, a técnica de otimização do *U-curve* constitui um novo *framework* para o estudo de problemas de seleção de características.

Capítulo 4

O *pipeline* de algoritmos

Utilizando o conceito de PGNs para modelar redes genéticas, desenvolvemos, aqui, um novo *pipeline* de algoritmos para construir redes genéticas. Os dados de entrada são baseados em experimentos de *série-temporal de microarray*. O método é iniciado com *genes-sementes*, os quais compartilham algumas características comuns, e procura sobre o banco de dados completo os genes melhores preditos por um subconjunto dos genes-sementes para o próximo período de tempo. O processo pode ser executado um certo número de iterações, obtendo, desse modo, um grafo representando a rede. Uma tabela de predição pode ser obtida para cada nó da rede, indicando a relação existente entre os genes.

Nosso método é uma solução robusta na modelagem de redes genéticas, utilizando dados temporais de expressão gênica. O grafo obtido é uma excelente ferramenta para o pesquisador investigar a rede em questão. Com o desenvolvimento de novas técnicas para a obtenção de dados temporais, as redes construídas podem ser mais representativas e úteis para estudos futuros. Nosso método possui, também, um forte suporte matemático em comparação com os métodos existentes, ou seja, a partir de um simples experimento de série-temporal de *microarray*, conseguimos reproduzir a "melhor" rede possível, se considerarmos que apenas os sinais temporais são conhecidos. O suporte matemático vêm de conceitos já conhecidos na área de Reconhecimento de Padrões, tais como entropia e seleção de características. Sendo assim, a rede obtida pelo nosso *pipeline* é uma representação bem característica do experimento de entrada e serve de ferramenta para possíveis estudos futuros.

Na próxima Seção serão apresentados os detalhes do processo completo e algumas alternativas para ele. Um exemplo da aplicação e seus resultados também serão apresentados. Para esse exemplo, os dados de entrada são: um experimento série-temporal de *microarray* de resposta ao *estrógeno* em cultura de células *T-47D* [26] tratada com estrógeno (E2) durante 24 horas. Um subgrupo de genes regulados pelo estrógeno é selecionada para iniciar a rede. A saída é composta por um grafo representando a interação entre os genes e seus preditores, assim como suas tabelas de predição.

4.1 Descrição do método

A motivação deste trabalho é desenvolver uma ferramenta útil que pode modelar uma rede genética a partir de um experimento específico de série-temporal de *microarrays*. A Figura 4.1 mostra o esquema da representação do *pipeline*:

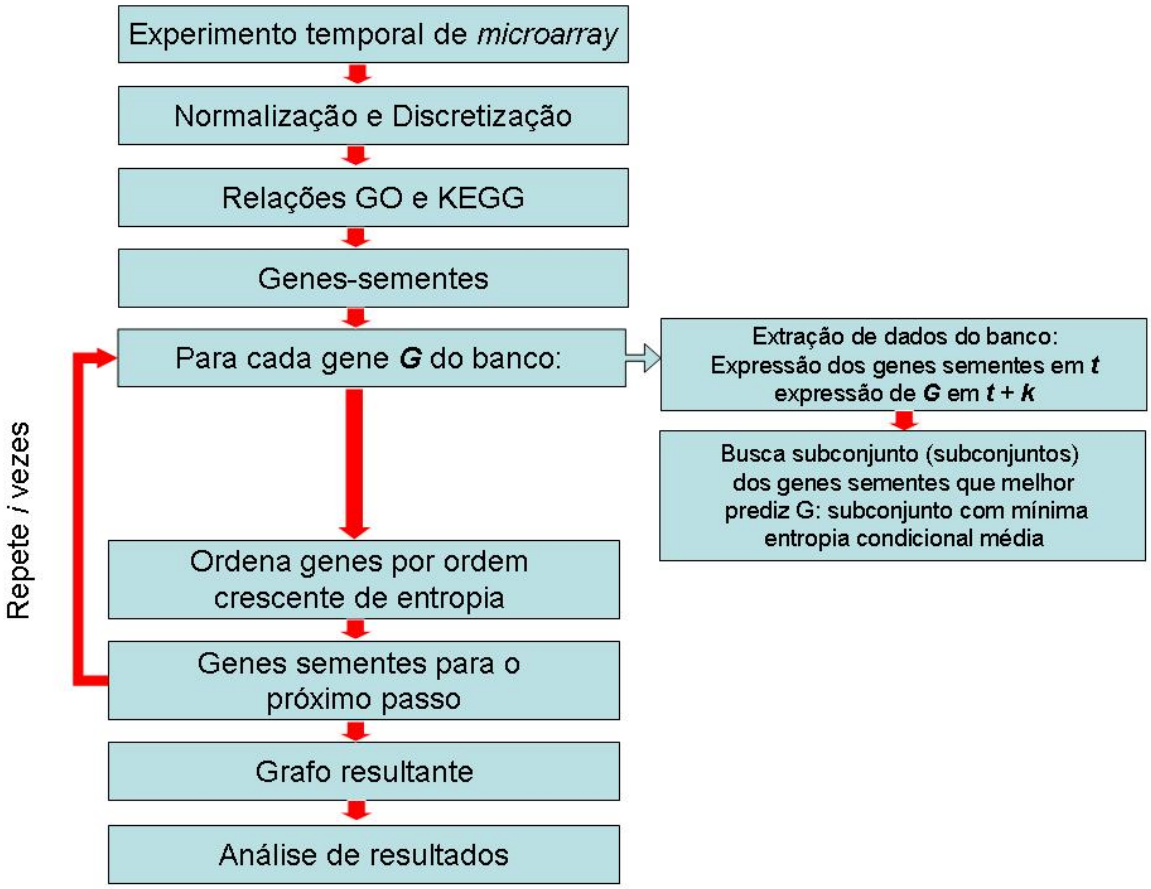


Figura 4.1: Representação esquemática do *pipeline* de algoritmos

1. A entrada é composta por uma série-temporal de *microarray*: ela pode ser vista como uma matriz $n \times m$, medindo-se o sinal de expressão de uma coleção fixa de n genes (geralmente, $n > 10.000$) durante m instantes (geralmente, a cada hora, durante 24 horas) para uma cultura celular ou tecido de algum organismo, submetido a uma condição específica. Cada experimento da série é comparado com um sinal de controle (logarítmo da razão). O sinal de controle é obtido, geralmente, a partir do mesmo experimento (mesma cultura celular ou tecido) submetido a condições normais [24, 15]. O número resultante no experimento é um número real: positivo se o gene na condição específica é mais expressso do que nas normais, negativo se menos expresso e próximo de 0 se igual às normais. Algumas vezes, uma expressão não pode ser medida por algum motivo (por exemplo: um sinal com ruídos), e para essas entradas o valor numérico é indefinido.

2. O *pipeline* requer uma entrada discretizada e, para isso, a matriz $n \times m$ de números reais deve ser normalizada e discretizada. A *normalização* é excutada para transformar o sinal temporal (m instantes) em um sinal com distribuição normal, com média (esperança) igual a 0 e desvio padrão igual a 1. A *discretização* transforma o sinal de números reais em valores discretos (inteiros), por exemplo, $\{-1, 0, +1\}$, com -1, indicando sub-expressão; 0, expressão normal, e $+1$, super-expressão.

3. O experimento de *microarray* contém um grande número de genes, sendo que alguns deles possuem funções biológicas indefinidas (não conhecidas). Identificar funções biológicas dos genes contidos no experimento é desejável para o processo como um todo. Muitos estudos têm produzido procedimentos de classificação para padronizar as funções biológicas dos genes, tais como: o *Gene Ontology Consortium* [2] e o banco de dados *KEGG* [29]. Enquanto a ontologia gênica relaciona um gene com os processos biológicos aos quais ele está envolvido, o banco KEGG relaciona um gene com uma *via de regulação*. Um processo para relacionar cada um dos n genes com suas ontologias e vias de regulação KEGG é processado procurando-se cada um dos genes (ou seus *aliases*) em cada um dos bancos (Gene Ontology e KEGG).

4. O *pipeline* é iniciado com um subconjunto dos n genes, denominado de *genes-sementes*, que compartilham alguma característica comum. O tipo de experimento é a base para a seleção desse subconjunto. Em experimentos, por exemplo, onde o efeito de algum hormônio sobre a célula é testado, um bom candidato para genes-sementes são os genes regulados diretamente por esse hormônio, isto é, genes cuja expressão é afetada nos primeiros momentos quando da presença do hormônio. A ontologia ou, também, as vias KEGG podem ser usadas para restringir a seleção dos genes-sementes. Esses genes serão o ponto de início da rede resultante ao final do *pipeline*.

5. Para cada gene G dos n genes da matriz M:

   - Uma sub-matriz $(m-k) \times (s+1)$ contendo a expressão dos correntes s genes-sementes nos instantes t $(t = 1, \ldots, m-k)$ e a expressão do gene G nos instantes $t+k$ $(t = 1, \ldots, m-k)$ é extraído de M. Note que k é qualquer número positivo menor que m, e ele pode ser ajustado de maneira a refletir o processo biológico, isto é, o tempo que uma mudança na expressão de um gene requer para afetar a expressão de outro gene.
   - O subconjunto (ou subconjuntos) dos genes-sementes atuais que melhor predizem o gene G é procurado entre todo o espaço de subconjuntos possíveis. Esse processo pode ser visto como um problema de *seleção de características* na área de *Reconhecimento de Padrões*, e é equivalente a buscar o subconjunto (ou subconjuntos) de menor *entropia condicional média* [22]. O subconjunto (ou subconjuntos) encontrado(s), seus valores de entropia associados e suas tabelas de predições são armazenadas no banco de dados para cada gene G. Os valores de entropia que relacionam o gene G ao subconjunto preditor serão denominados de *custo* do gene G.

6. Os n genes são, então, ordenados por seu custo de modo crescente (*ranking*). Os primeiros genes na ordenação são os *melhores preditos* pelo conjunto de genes-sementes atual, isto é, o subconjuto dos genes-sementes atuais tem um valor baixo para entropia condicional média.

7. Um novo subconjunto de genes-sementes é escolhido dentre os genes no topo da ordenação (genes de menor custo). Para essa seleção pode ser aplicada as restrições semelhantes das usadas para selecionar o conjunto inicial.

8. O processo pode ser iterado i vezes.

9. Para cada iteração, um conjunto de genes (genes-sementes) atua como entrada, e a saída é composta por um conjuto de genes preditos com seus subconjuntos de preditores. A partir desses dados, um grafo, como o da Figura 2.2, pode ser contruído.

10. Todos os dados resultantes podem ser analizados: simulação, validação, comprovação ds relações em bancada, entre outros.

4.2 Normalização e discretização

O *pipeline* requer que as expresssões sejam valores discretos em vez de números reais contidos na matriz que representa o experimento de série-temporal de *microarray*. Existem vários métodos para a tranformação de números reais em valores discretos. Seja M a matriz resultante $n \times m$ do experimento de série-temporal de *microarray*. O método baseado em [3] foi implementado. Ele se baseia em duas etapas:

- *Normalização* da matriz M para a matriz M_N. Esse processo consiste em normalizar o sinal de cada gene em um sinal com distribuição normal de média (esperança) igual a 0 e desvio padrão igual a 1. Com isso, todos os genes terão a mesma distribuição e suas expressões poderão ser comparadas. A normalização consiste em: (i) calcular o valor esperado E_i e o desvio padrão σ_i para o sinal de cada gene i dados, respectivamente, por $E_i = \frac{\sum_{k=1}^{m} M[i,k]}{m}$ e $\sigma_i = \sqrt{\frac{\sum_{k=1}^{m}(M[i,k]-\mu_i)^2}{m-1}}$; (ii) os elementos resultantes da matriz normalizada M_N são dados por $M_N[i,k] = \frac{M[i,k]-E_i}{\sigma_i}, i = 1, \ldots, n$ e $k = 1, \ldots, m$.

- *Discretização* da matriz M_N para a matriz M_Q. Esse processo equivale a mapear os sinais normalizados de M_N para um conjunto de níveis qualitativos de expressão. Aqui usaremos três níveis qualitativos de expressão $\{-1, 0, +1\}$, indicando, respectivamente: sub-expresso, nulo, e sobre-expresso em relação ao valor de referência. O *mapeamento* para a quantização foi feito usando-se dois *limitantes* como em [3]. Para cada gene i, um limitante inferior l_i e outro superior u_i foram calculados, respectivamente, por:

$$l_i = \frac{\sum_{M_N[i,k]<0} M_N[i,k]}{|\{M_N[i,k] : M_N[i,k] < 0\}|} \tag{4.1}$$

$$u_i = \frac{\sum_{M_N[i,k]>0} M_N[i,k]}{|\{M_N[i,k] : M_N[i,k] > 0\}|} \tag{4.2}$$

  Em outras palavras, l_i e u_i são, respectivamente, o valor esperado dos sinais negativos e dos positivos. Os elementos da matriz quantizada M_Q são dados por:

$$M_Q[i,k] = \begin{cases} -1 & \text{se } M_N[i,k] < l_i \\ 0 & \text{se } M_N[i,k] >= l_i \text{ and } M_N[i,k] >= l_i \\ +l & \text{se } M_N[i,k] >= u_i \end{cases}$$

para $i = 1, \ldots, n$ e $k = 1, \ldots, m$.

O número de limitantes deve ser incrementado quando a discretização for composta por mais de três níveis. Por exemplo, se os níveis de discretização são $\{-2, -1, 0, +1, +2\}$, os valores -2 e $+2$ podem, respectivamente, ser associados aos limitantes $2 \times l_i$ e $2 \times u_i$. O número de níveis de discretização pode ser determinado pelo número de amostras (instantes) que o experimento de série-temporal possui. Para valores pequenos de amostras (m), um número superior a três níveis de discretização, geralmente, não produz bons resultados, isto é, poucas amostras para um número grande de possíveis estados.

Algumas entradas (genes) do experimento de série-temporal devem ser filtradas. A razão para isso pode estar em um dos dois casos: (i) o sinal da expressão não pode ser determinado, ou (ii) a expressão do gene é constante ao longo de todo experimeento, isto é, seu valor se mantém igual a -1, 0 ou $+1$ durante todos os instantes (1 a m). Para o primeiro caso, simplesmente atribuímos o valor nulo para a entrada ($M_Q[i,k] =$ nulo). Para o segundo caso, os genes com expressões constantes podem ser preditos por quaisquer combinações de outros genes, ou seja, a entropia condicional média sempre vale 0, dado qualquer subconjunto de preditores. Sendo assim, esses genes não agregam nenhuma nova informação à rede resultante e, por essa razão, eles devem ser removidos do banco de dados inicial.

4.2.1 Análise do conjunto de genes-sementes

A cada passo o *pipeline* requer um conjunto de genes, denominado de *genes-sementes*. Esse conjunto é formado pelos possíveis preditores a serem encontrados para cada um dos genes contidos no banco de dados completo do experimento. Existem diversas possibilidades na definição do método de seleção dos genes-sementes. A rede resultante do *pipeline* é uma rede dependente do tempo (Figura 2.2) e, para a seleção dos genes que iniciarão a rede, seria conveniente que estes compartilhassem alguma característica comum. Experimentos nos quais, a regulação de algum fator biológico é testada, um bom candidato para o conjunto de genes-sementes iniciais pode ser formado pelo genes regulados diretamente pelo fator biológico. O termo "*regulado diretamente*" por um fator biológico significa que o fator biológico participa diretamente na regulação da expressão do gene, em outras palavras, o fator biológico é um *fator de transcrição* do gene. Por exemplo, para células tratadas com estrógeno para testar sua rede de regulação, um bom candidato para o conjunto de genes-sementes pode ser os genes regulados diretamente pelo estrógeno, isto é, genes onde o estrógeno atua como fator de transcrição. Esses genes podem ser obtidos em estudos prévios ou a partir de métodos de agrupamento (*clustering*). Os genes-sementes para os passos subseqüentes serão compostos pelos genes melhores preditos pelo conjunto de genes-sementes do passo anterior. Um subconjunto $\mathcal{A}$ dos genes-sementes prediz a expressão do gene A no instante seguinte melhor do que um subconjunto $\mathcal{B}$ prediz o gene B, se a tabela de predição que associa $\mathcal{A}$ para A é *mais determinística* que a que associa $\mathcal{B}$ para B. Em outras palavras, o vetor de expressões $\mathbf{A}$ dos genes em $\mathcal{A}$ determina a expressão do gene A no próximo instante com menos erro que os genes em $\mathcal{B}$ determinam a expressão do gene B.

O tamanho do conjunto de genes-sementes ideal para o problema, também, pode ser função do tamanho m do experimento. Quanto maior o número m, maior será o subconjunto de predição obtido, pois os erros de estimação associados aos subconjuntos serão relevantes apenas para subconjuntos com tamanhos superiores. Em nossos testes, utilizamos conjuntos de genes-semente da ordem de 30 elementos para um experimento da ordem de $m = 15$.

4.2.2 Função custo

Para cada gene, denominado de *gene-alvo*, do banco de dados e um dado conjunto de genes-sementes, tentamos encontrar, entre todos os subconjuntos possíveis, aquele que melhor prediz a expressão do gene-alvo em questão. *Erro de classificação* é um conceito matemático que pode ser aplicado para medir o quanto uma variável aleatória A é determinada dado um vetor aleatório $\mathbf{A}$. Se o erro de classificação de A dado $\mathbf{A}$ é pequeno, dizemos que A é *bem determinado* por $\mathbf{A}$; e se o erro de classificação é grande, dizemos que A é *mal determinado* por $\mathbf{A}$. Para nosso problema, a variável aleatória A representa a expressão do gene-alvo no instante t, e o vetor aleatório $\mathbf{A}$ representa a expressão de um conjunto de genes em $t - k$. Existem inúmeras maneiras de estimar o erro de classificação: *MAE* (erro absoluto médio), *entropia condicional média* [22], *CoD* (coeficiente de determinação) [8, 16] e *Bolstered Error* [5, 38] são alguns exemplos deles. Enquanto as medidas de erro MAE, CoD e *Bolstered* indicam a probabilidade média da classificação incorreta, a entropia condicional média é uma medida inversamente proporcional à quantidade de informação obtida para a classificação dado o subconjunto de entrada. Todas essas medidas possuem valores pequenos se o subconjunto de entrada conduz a bons resultados de classificação; no nosso caso, se ele é um bom preditor do gene-alvo.

Em nossos testes, usamos a *entropia condicional média* como função custo. A *entropia* mede a dispersão de uma função de distribuição de probabilidade, isto é, a entropia possui valores pequenos para distribuições com massa concentrada em um dos valores possíveis, e possui o maior valor para uma distribuição uniforme. A Figura 4.2 mostra três exemplos de funções de distribuição de probabilidades para uma variável aleatória discreta com três níveis de discretização: -1, 0 e 1. A entropia máxima ocorre para a varáivel aleatória Y (distribuição uniforme), e as entropias das variáveis Y' e Y'' são equivalentes e com valores pequenos, já que suas massas estão concentradas em um dos valores: 1 para Y' e -1 para Y''.

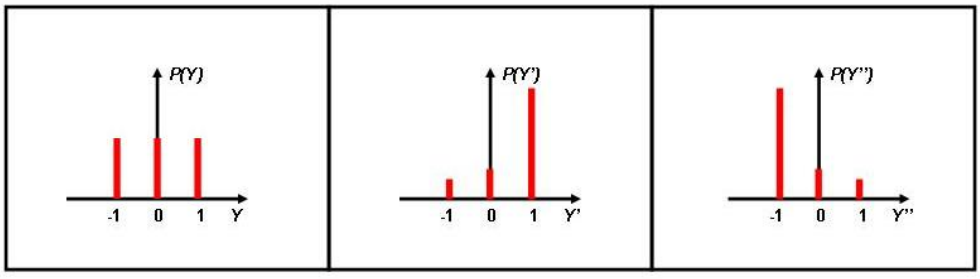


Figura 4.2: Três exemplos de funções distribuição de probabilidades para uma variável aleatória discreta com três níveis de discretização

Em nosso estudo estamos interessados na probabilidade de uma variável aleatória discreta G, que representa a expressão do gene-alvo, dado um vetor aleatório discreto $\mathbf{A}$ representando a expressão de um subconjunto dos genes-sementes. Essa probabilidade é denotada por $P_{G|\mathbf{A}}$. Seja $\mathcal{Q}$ o conjunto de valores discretos usados para diferenciar as expressões dos genes, por exemplo, $\mathcal{Q} = \{-1, 0, +1\}$. A *entropia condicional média* $E[H(G|\mathcal{A})]$ é dada por:

$$E[H(G|\mathbf{A})] = \sum_{\mathbf{a}\in\mathcal{Q}^{|\mathbf{A}|}} P_{\mathbf{A}}(\mathbf{A}=\mathbf{a}) \left(-\sum_{g\in\mathcal{Q}} P_{G|\mathbf{A}}(G=g|\mathbf{A}=\mathbf{a}) \log_{|\mathcal{Q}|} P_{G|\mathbf{A}}(G=g|\mathbf{A}=\mathbf{a}) \right) \quad (4.3)$$

Baseando-se apenas nos dados de entrada, podemos simplesmente calcular uma estimativa da função custo, e não o seu valor real. A *entropia condicional média estimada* $\hat{E}[H(G|\mathcal{A})]$ é dada por:

$$\hat{E}[H(G|\mathbf{A})] = \sum_{\mathbf{a}\in\mathcal{Q}^{|\mathbf{A}|}} \hat{P}_{\mathbf{A}}(\mathbf{A}=\mathbf{a}) \left(-\sum_{g\in\mathcal{Q}} \hat{P}_{G|\mathbf{A}}(G=g|\mathbf{A}=\mathbf{a}) \log_{|\mathcal{Q}|} \hat{P}_{G|\mathcal{A}}(G=g|\mathbf{A}=\mathbf{a}) \right) \quad (4.4)$$

A equação (4.4) requer a estimativa de $P_{\mathbf{A}}$ e $P_{G|\mathbf{A}}$. Existem vários estimadores possíveis para essas medidas. Para o nosso trabalho, adotaremos os mesmos estimadores em [3]. Seja f um número inteiro, denominado de *freqüência limitante*, usado para separar instâncias $\mathbf{a}$ do vetor aleatório $\mathbf{A}$ que ocorrem com pouca freqüência. Seja N^+ a soma das freqüências das instâncias $\mathbf{a}$ de $\mathbf{A}$ que ocorrem um número de vezes maior ou igual a f nas amostras, e N^- a soma das freqüências das instâncias $\mathbf{a}$ de $\mathbf{A}$ que ocorrem menos do que f vezes nas amostras. Isto equivale a:

$$N^+ = \sum_{\#(\mathbf{A}=\mathbf{a})\geq f, \forall \mathbf{a}\in\mathcal{Q}^{|\mathbf{A}|}} \#(\mathbf{A}=\mathbf{a}) \quad (4.5)$$

$$(4.6)$$

$$N^- = \sum_{\#(\mathbf{A}=\mathbf{a})< f, \forall \mathbf{a}\in\mathcal{Q}^{|\mathbf{A}|}} \#(\mathbf{A}=\mathbf{a}) \quad (4.7)$$

Seja $\mathbf{a} \in \mathcal{Q}^{|\mathbf{A}|}$ uma instância de $\mathbf{A}$, e $g \in \mathcal{Q}$ uma instância de G, o estimador $\hat{P}_{G|\mathbf{A}}$ de $P_{G|\mathbf{A}}$ é dado por:

$$\hat{P}_{G|\mathbf{A}}(G=g|\mathbf{A}=\mathbf{a}) = \begin{cases} \frac{\#(G=g\wedge\mathbf{A}=\mathbf{a})}{\#(\mathbf{A}=\mathbf{a})}, & \text{se } \#(\mathbf{A}=\mathbf{a}) \geq f, \\ \\ \frac{\#(G=g)}{N^+ + N^-} & \text{se } \#(\mathbf{A}=\mathbf{a}), < f \end{cases}$$

e o estimador $\hat{P}_{\mathbf{A}}$ de $P_{\mathbf{A}}$ dado por:

$$\hat{P}_{\mathbf{A}}(\mathbf{A}=\mathbf{a}) = \begin{cases} \frac{N^+}{N^+ + N^-} \times \frac{\#(\mathbf{A}=\mathbf{a})}{N^+}, & \text{se } \#(\mathbf{A}=\mathbf{a}) \geq f \\ \\ \frac{N^-}{N^+ + N^-} \times \frac{1}{2^{|\mathbf{A}|} - |\{\mathbf{a}'\in\mathcal{Q}^{|\mathbf{A}|}:\#(\mathbf{A}=\mathbf{a}'\geq f)\}|}, & \text{se } \#(\mathbf{A}=\mathbf{a}) \geq f \end{cases}$$

O estimador $\hat{P}_{\mathbf{A}}$ distribui uniformemente a freqüência N^- entre todas as instâncias que não ocorrem ou ocorrem menos que f vezes. A Figura 4.3 mostra graficamente a representação do estimador $\hat{P}_{\mathbf{A}}$.

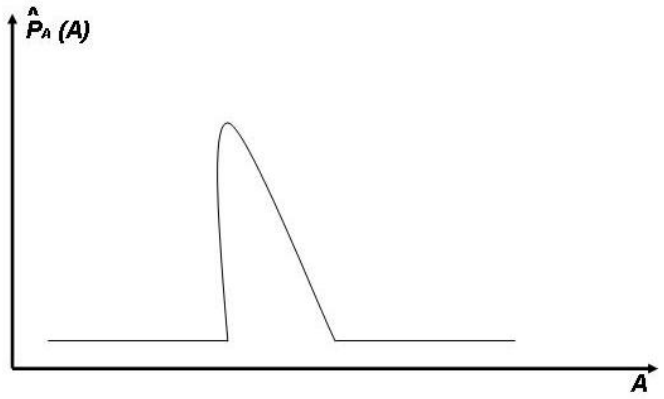


Figura 4.3: Representação gráfica do estimador $\hat{P}_{\mathbf{A}}$ de uma função distribuição de probabilidade para um vetor de expressões $\mathbf{A}$

4.2.3 Melhor subconjunto preditor

O problema de encontrar o melhor subconjunto de predição (subconjunto de menor custo) entre a família de subconjuntos de um dado conjunto é um problema combinatório de grande complexidade (*exponencial*). Seja $\mathcal{S}$ o conjunto de genes-sementes usados para predizer a expressão de um gene G. O número de possíveis subconjuntos de $\mathcal{S}$ é exponencial ($2^{|\mathcal{S}|}$), e para grandes valores de $|\mathcal{S}|$, a busca completa se torna impraticável. Esse problema pode ser estudado como um problema de *seleção de características* no contexto de *Reconhecimentos de Padrão* [9, 19]. Algumas *heurísticas* foram desenvolvidas para lidar com esses tipos de problema. Elas exploram o espaço de possibilidades sem ser equivalente a busca completa, isto é, elas não garantem encontrar o melhor subconjunto (menor custo). Podemos exemplificar com algumas heurísticas de relativo sucesso empregadas: *SFS* (*Sequential Forward Selection*), *SBS* (*Sequential Backward Selection*), *SFFS* (*Sequential Floating Backward Selection*) e *SFFS* (*Sequential Floating Forward Selection*) [32].

O espaço de busca pode ser organizado por um *reticulado Booleano completo* de ordem $|\mathcal{S}|$, no qual cada elemento representa um subconjunto de $\mathcal{S}$. A Figura 4.4 mostra um reticulado Booleano completo de ordem 4. Seja $\mathcal{S} = \{G_1, G_2, G_3, G_4\}$, o elemento 1010 do reticulado Booleano representa o subconjunto $\{G_1, G_3\}$ de $\mathcal{S}$, isto é, 1 ou 0 na posição k indica, respectivamente, que o gene G_k está ou não está contido no subconjunto representado pelo elemento. Além disso, cada elemento do reticulado Booleano possui um valor para a função custo associado ao subconjunto representado por ele. O elemento 0111 (verde) representa o subconjunto de mínimo custo (5). Uma característica importante desse problema é que a função custo restrita a qualquer cadeia do reticulado Booleano tem o formato de *curva em U*. Uma cadeia (0000, 0100, 0110, 1110, 1111) é enfatizada (vermelho) na Figura 4.4, e podemos ver a curva em U formada pelo custo restrito à ela (9, 7, 6, 10, 11). Essa característica pode ser usada para desenvolver soluções do tipo *branch-and-bound*. Um exemplo é o *algoritmo U-curve* [33], o qual se equivale à busca completa sem visitar o espaço total. Isso ocorre pela remoção de subespaços (*posets*) do espaço

de busca, a cada momento em que um elemento de custo mínimo é encontrado na cadeia. Na Seção 3, o algoritmo *U-curve* é apresentado em detalhes.

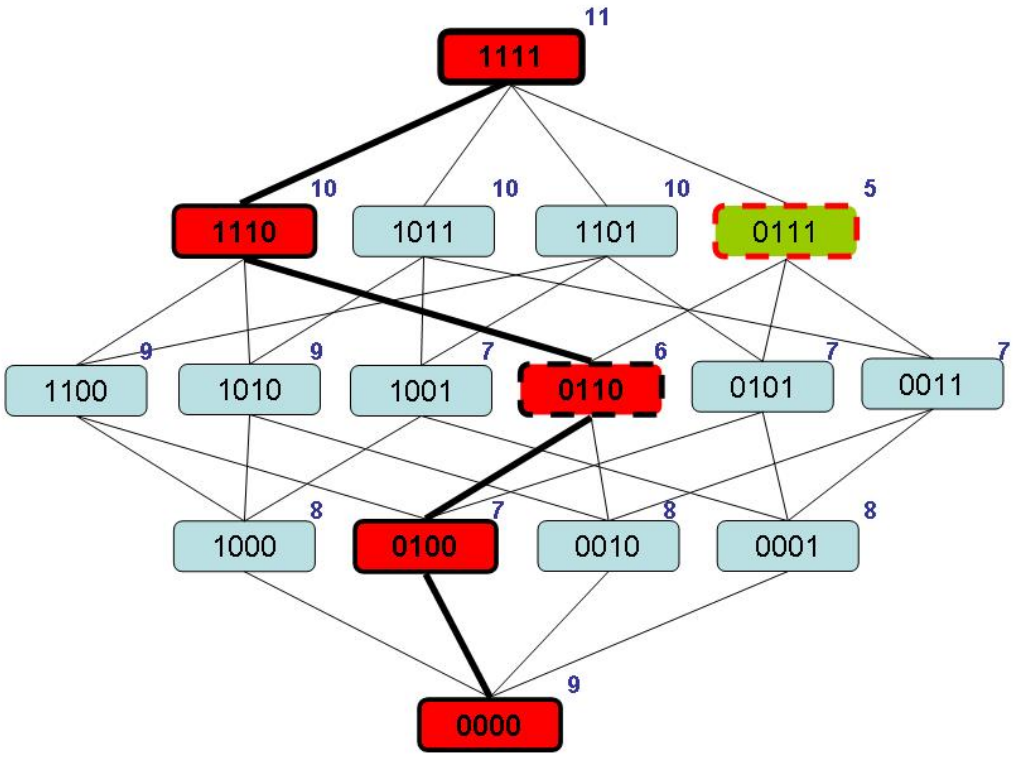


Figura 4.4: Representação de um reticulado Booleano de ordem 4 e a função custo associada a cada um de seus elementos

A cardinalidade do subconjuto de custo mínimo está relacionada à cardinalidade $|\mathcal{S}|$ do conjunto $\mathcal{S}$ e ao tamanho m do experimento (amostras) usado para estimar a função custo. O custo de um subconjunto é formado em parte pelo erro de classificação real e parte pelo erro de estimação. Quando m é pequeno e $|\mathcal{S}|$ é grande, os subconjuntos de $\mathcal{S}$ com mínimo custo tendem a ter poucos componentes em comparação ao conjunto inteiro $\mathcal{S}$. Observando o reticulado Booleano (Figura 4.4), esses subconjuntos estarão mais perto da base (elemento 0000) do que do topo (elemento 1111). Para esses casos, uma busca exaustiva limitada a um tamanho de subconjunto definido pode ser executada. Por exemplo, limitar a busca a subconjuntos de tamanho 3 reduz o espaço de busca para um tamanho de $\binom{|\mathcal{S}|}{0} + \binom{|\mathcal{S}|}{1} + \binom{|\mathcal{S}|}{2} + \binom{|\mathcal{S}|}{3}$ elementos.

O subconjuto de menor custo nem sempre representa o melhor subconjunto a ser considerado. Isso pode acontecer, pois experimentos com poucas amostras tornam o erro de estimação como a principal componente da função custo e, sendo assim, subconjuntos com custos próximos podem ser bons candidatos ao melhor subconjunto de preditores. O método usado para encontrar o

melhor subconjunto preditor pode retornar não somente o melhor, mas, sim, uma lista dos melhores subconjutos encontrados, ordenados pelos seus respectivos custos. Essa lista pode fornecer ao pesquisador uma variedade maior de subconjuntos a serem analisados.

As Tabelas 4.1 e 4.2 mostram dois exemplos de tabelas de predições e suas funções funções custos: entropia condicional média com freqüência limitante $f = 1$. A primeira coluna das tabelas mostra a freqüência da ocorrência de cada uma das configurações dos genes do subconjunto (TBX21,FRAT2) em um experimento de série-temporal de *microarray* com 15 instantes. Pode-se notar que na Tabela 4.1 a expressão do gene-alvo TGFBP3 é quase sempre determinada (5 de 6 vezes) pela expressão dos genes TBX21 e FRAT2; já na Tabela 4.2 a expressão do gene-alvo TNFSF8 é determinada um número menor de vezes (4 de 6 vezes) pelos genes POU1F1 e EPAS1. Com isso, o custo associado à Tabela 4.1 $(0,066667)$ é menor em comparação ao associado à Tabela 4.2 $(0,217674)$.

Subconjunto dos genes-sementes			**TGFBP3**		
Freqüência	**TBX21**	**FRAT2**	**-1**	**0**	**+1**
2	-1	-1	0,00 %	**100,00 %**	0,00 %
3	-1	0	**100.00 %**	0,00 %	0,00 %
2	0	-1	0,00 %	**100,00 %**	0,00 %
5	0	0	0,00 %	**100,00 %**	0,00 %
2	0	1	0,00 %	0,00 %	**100,00 %**
1	1	1	33,33 %	33,33 %	33,33 %

Tabela 4.1: Exemplo de uma tabela de predição para o subconjunto de genes TBX21 e FRAT2 ao gene TGFBP3.

Subconjunto dos genes-sementes			**TNFSF8**		
Freqüência	**POU1F1**	**EPAS1**	**-1**	**0**	**+1**
2	-1	0	0,00 %	**100,00 %**	0,00 %
4	0	-1	0,00%	75,00 %	25,00 %
3	0	0	0,00 %	**100,00 %**	0,00 %
2	0	1	0,00 %	**100,00 %**	0,00 %
1	1	0	33,33 %	33,33 %	33,33 %
3	1	1	**100,00 %**	0,00 %	0,00 %

Tabela 4.2: Exemplo de uma tabela de predição para o subconjunto de genes POU1F1 e EPAS1 ao gene TNFSF8.

4.2.4 Ordenação dos resultados ("*ranking*")

O melhor subconjuto preditor dos genes-sementes e o custo associados a cada um dos genes do banco de dados são armazenados ao longo do processo. Dizemos que um gene G é *melhor predito* que um gene G' pelo conjunto de genes-sementes se o custo do melhor subconjunto de preditores

dos genes-sementes para G é menor que o custo do melhor subconjunto para G'. Dispor os genes do banco de dados em ordem crescente do custo associado aos seus melhores subconjuntos de predição produz uma lista em que os elementos iniciais são os genes melhores preditos pelos genes-sementes no próximo instante k. Esse é o procedimento principal na obtenção de uma nova lista de genes-sementes para o próximo passo do *pipeline*. Muitos métodos podem ser empregados para extrair o próximo conjunto de genes-sementes. Apresentamos, aqui, alguns deles:

- Definir de um valor s limitante para a função custo, isto é, extrair os genes com melhor subconjunto de preditores de custo menor que s. Esse valor limitante pode ser alterado a cada iteração, dependendo do número de genes extraídos por esse valor. À medida que s cresce, o melhor subconjunto preditor para alguns genes se torna menos deteminístico, sendo assim, um gene com custo alto não será bem determinado pelo conjunto de genes-sementes, não tornando-se um bom candidato a ser incluído no próximo conjunto de genes-sementes.
- Extrair um número fixo de genes do topo da lista de ordenação que possuem alguma função biológica em comum (GO e/ou KEGG). Por exemplo, extrair os primeiros 30 genes que possuem uma das funções biológicas: divisão celular, proliferação celular e/ou ciclo celular.
- A partir de uma dada lista inicial de genes, extrair um número fixo de genes do topo da ordenação que pertence a lista.

Esses métodos podem ser agrupados para a obtenção do próximo conjunto de genes-sementes. Por exemplo, escolher um valor limitante s e uma série de funções biológicas para extrair um conjunto de genes-sementes que possuem melhor subconjunto preditor com custo menor que s e compartilham essas mesmas funções biológicas.

4.2.5 Resultados experimentais

A entrada para os resultados experimentais foi obtida a partir de um experimento de série-temporal de *microarray* (*Compugen 19K human oligonucleotide array*) de células *T-47D*, tratadas com *estrógeno* (*E2*) [26], com 16 experimentos, durante 24 horas. Alguns dos genes foram removidos do processo completo: (i) genes não encontrados no banco de dados "*Stanford Micro-Array*", ou (ii) genes com sinais de expressão constantes ou com a maioria (mais do que metade) dos sinais não definidos. Para os sinais não definidos, as entradas foram marcadas com valores nulos e ignoradas no processo. As vias *KEGG* e as funções biológicas (*GO*) foram pesquisadas e armazenadas no banco de dados para cada gene.

Todos os dados foram normalizados e discretizdos usando-se o mesmo método em [3] com três níveis de discretização $\{-1, 0, +1\}$ e utilizando os limitantes l_i e u_i, descritos nas Equações 4.1 e 4.2, para cada gene i.

O conjunto inicial dos genes-sementes foi obtido pela pesquisa no banco de dados dos 386 genes que possuem resposta ao estrógeno (E2) (trabalho em [26]) e que contêm uma das seguintes funções biológicas em suas entradas no *GO*: *cell cycle* (ciclo celular), *cell proliferation* (proliferação celular), *DNA* ou *cell diferentiation* (diferenciação celular). Essas funções biológicas,

usadas como filtro para os genes-sementes iniciais, foram obtidas do trabalho em [13], como as categorias funcionais de genes estimulados ou reprimidos pelo estrógeno. Essa condição foi usada como filtro para extração dos genes-sementes para todas as iterações. O conjunto resultante inicial foi composto de 33 genes. O melhor subconjunto preditor dos genes-sementes para cada gene foi obtido usando, como função custo a entropia condicional média estimada descrita anteriormente, e a busca exaustiva limitada em um subconjunto de tamanho 3. A submatriz 15×34 das expressões dos genes-sementes no instante t ($t = 1, \ldots, 15$) e de cada um dos genes-alvos no instante $t + 1$ ($k = 1$) foi a entrada para a busca exaustiva do melhor subconjunto preditor do gene-alvo. O conjunto dos genes-sementes, extraídos na segunda iteração, foi obtido dos genes filtrados (funções biológicas utilizados no conjunto inicial) do topo da ordenação (*ranking*) entre os melhores preditos pelos genes-sementes iniciais (custo < 0.13), resultando em um conjunto de 38 novos genes-sementes. O mesmo processo para encontrar o melhor subconjunto preditor para esse novo conjunto de genes-sementes foi executado. O conjunto de genes-sementes para a terceira iteração foi extraído entre os genes filtrados no topo da ordenação (custo < 0.09), resultando em um novo conjunto de 37 genes-sementes. Os melhores genes preditos após a terceira iteração foram, também, extraídos entre os genes filtrados no topo da ordenação (custo < 0.09), resultando em um conjunto de 38 genes-sementes.

Os resultados do processo completo foram organizados e processados para gerar um grafo representando a rede. A Figura 4.5 mostra o grafo que representa a rede obtida. As arestas direcionadas verdes, vermelhas e azuis representam, respectivamente, as primeiras, segundas e terceiras iterações.

O grafo resultante é apresentado como uma página *html* e, para cada nó, representando um gene, existem redirecionamentos (*links*) para: (i) o banco de dados "*Stanford MicroArray*" (Figura 4.6), relacionando o gene com seus possíveis *aliases*, e informações conhecidas a respeito dele; (ii) suas tabelas de predições, relacionando-o a seus melhores subconjuntos de predições (Figure 4.7); e (iii) um gráfico do sinal de expressões do gene-alvo e de seus preditores ao longo do experimento (Figura 4.8).

4.3 Discussão

O *pipeline* de algoritmos desenvolvido possui vários parâmetros a serem explorados. A simples mudança em seus valores ou nos métodos empregados para a obtenção deles podem produzir resultados diferentes. Implementamos, neste trabalho, alguns dos métodos discutidos e utilizamos alguns valores para os parâmetros, produzindo resultados que podem ser analisados pelo pesquisador em trabalhos futuros. A qualidade dos resultados e os parâmetros utilizados para sua obtenção estão extremamente relacionados com o tipo do experimento inicial de entrada. Em um experimento com poucas amostras, isto é, um experimento de série-temporal de *microarray* com um número pequeno de medidas, a qualidade do resultado pode ser ruim, mas, de qualquer modo, o resultado produzido é um dos melhores que podem ser obtidos a partir do experimento de entrada. Um parâmetro importante a destacar é a diferença de tempo k entre o instante da expressão medida para os genes preditores e o gene-alvo predito, isto é, o período de tempo que uma mudança na expressão de um gene afeta a expressão de outro. Para um experimento de série-temporal no qual as medidas são feitas a cada hora, o valor mínimo para k é 1, mas não existe garantia formal de que esta é a melhor escolha. Outros parâmetros que podem ser

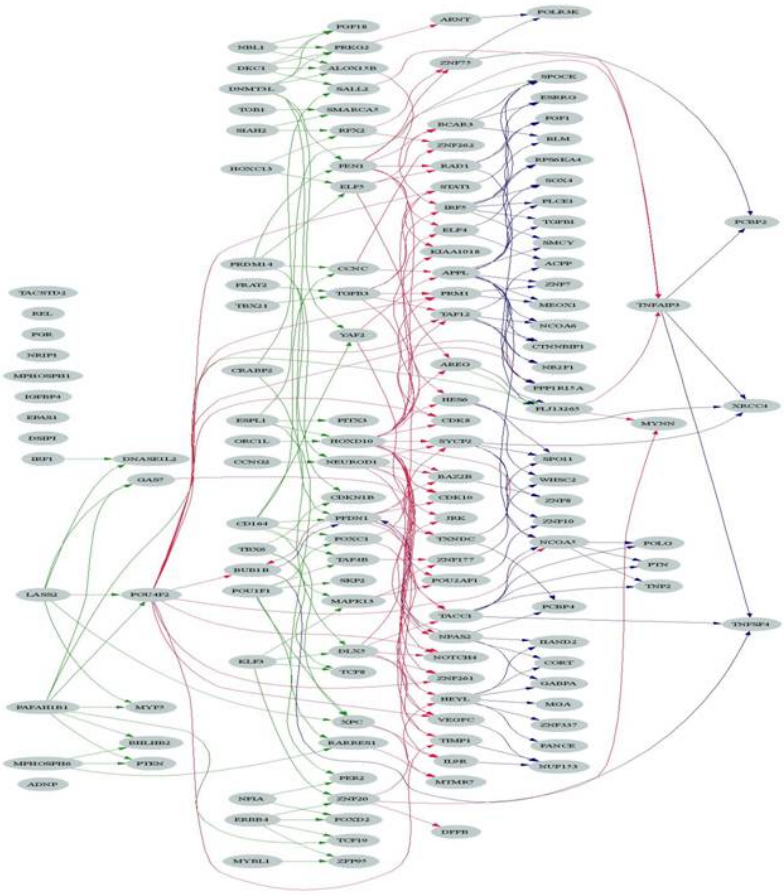


Figura 4.5: Grafo resultante obtido pela aplicação do *pipeline* ao experimento de série-temporal de *microarray* em [26]

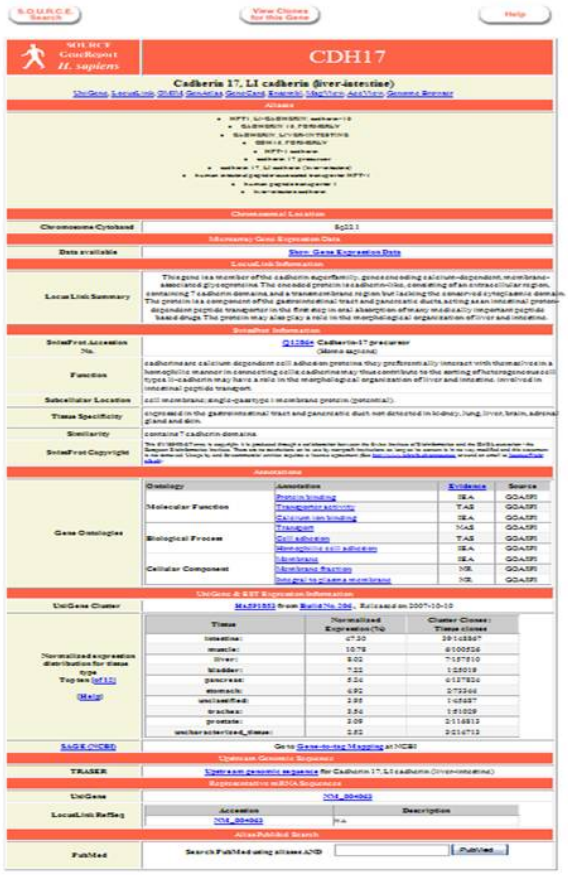


Figura 4.6: Página do banco de dados de "*Stanford MicroArray*" para o gene CDH17

Subset Entropy: **0.126185**
Target Entropy: 0.826074
Expression Plot: PLOT - TESTE1_STEP2_GO

ORC1L[CC]	ESPL1[CC]	HOXD10		
-1	-1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
-1	0	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	0 (50.00 %) [1]	1 (50.00 %) [1]

Subset Entropy: **0.083333**
Target Entropy: 0.783581
Expression Plot: PLOT - TESTE1_STEP3_GO

HOXD10	NEUROD1	PFDN1	SYCP2		
-1	0	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	-1	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
0	0	-1	-1 (33.00 %) [0]	0 (33.00 %) [1]	1 (33.00 %) [0]
0	0	0	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	0	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	0	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]

Subset Entropy: **0.083333**
Target Entropy: 0.783581
Expression Plot: PLOT - TESTE1_STEP3_GO

TGFB3[MAPK][CC][CYT][TGF]	POU4F2	HOXD10	CDK8		
-1	0	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	0	-1 (33.00 %) [0]	0 (33.00 %) [1]	1 (33.00 %) [0]
0	0	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	1	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
1	0	0	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]

Subset Entropy: **0.083333**
Target Entropy: 0.783581
Expression Plot: PLOT - TESTE1_STEP3_GO

HOXD10	NEUROD1	PFDN1	BAZ2B		
-1	0	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
0	-1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	-1	-1 (33.00 %) [0]	0 (33.00 %) [1]	1 (33.00 %) [0]
0	0	0	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	0	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	0	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]

Figura 4.7: Quatro exemplos de tabelas de predições como uma página *html*, para uma parte da rede: gene HOXD10 com seus preditores e preditos

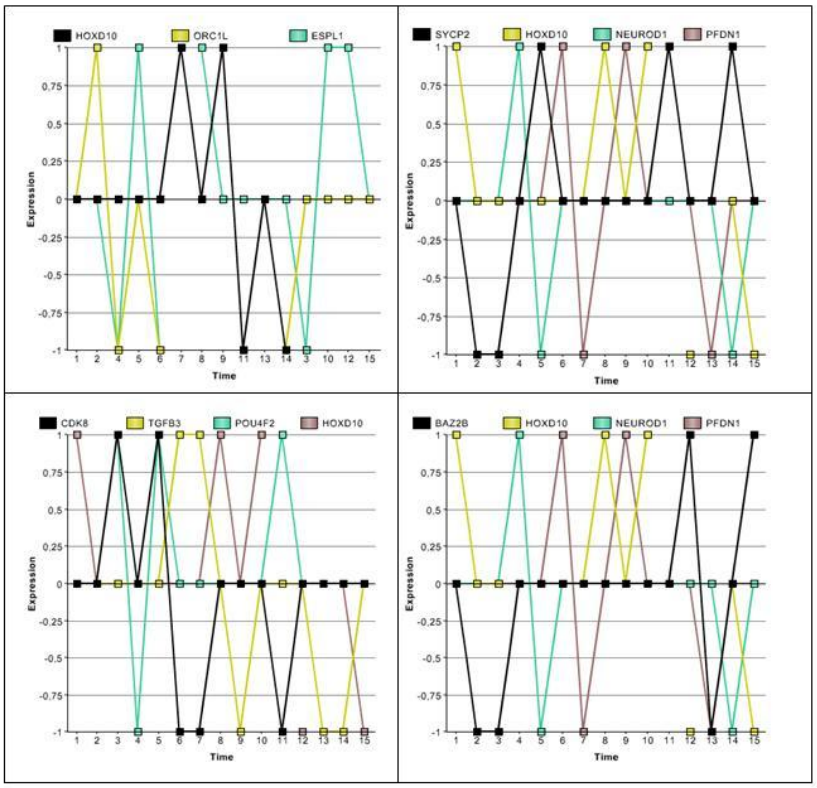


Figura 4.8: Gráficos das expressões dos genes relacionados a cada uma das quatro tabelas de predições da Figura 4.7

analisados em trabalhos futuros são:

- métodos alternativos para filtrar os genes do banco de dados;
- normalização e discretização dos sinais de expressões dos genes;
- estimação para a função custo;
- métodos para a extração dos genes-sementes;
- algoritmos para seleção de características aplicada a seleção de melhor conjunto preditor.

A rede produzida pelo processo completo pode também ser simulada e testada em comparação com dados reais. Isso pode ser feito a partir da entrada de uma configuração dada para os genes-sementes iniciais e simulada pelas tabelas de predições de cada um dos genes preditos. Esse processo não é apresentado neste trabalho, mas pode ser implementado e testado, em trabalhos futuros, para uma validação da rede obtida.

4.4 Materiais e métodos

Nesta seção, apresentaremos os materiais e métodos empregados pelo processo completo na rede gênica obtida nos resultados experimentais.

4.4.1 Série-temporal de *microarray*

Os dados de entrada de nosso trabalho são baseados nos resultados do experimento em Liu et. al (2004) [26]. O experimento é composto por uma série-temporal de *microarray* de células *T47-D* tratadas com estrógeno (17β-estradiol (E2)) durante 24 horas. O sinal de controle é dado pelo mesmo experimento em células *T-47D* não tratadas com estrógeno durante as mesmas 24 horas. O número total de experimentos é de 16 pares (tratados com estrógeno e seu sinal de controle): os primeiros 8 foram medidos a cada hora e os restantes 8, a cada duas horas. Esse trabalho obteve uma lista de 386 genes responsivos ao estrógeno, e uma lista de 89 genes diretamente regulados por estrógeno. Os dados foram inteiramente armazenados em um banco de dados MySql$^{\text{TM}}$ (<`http://www.mysql.com`>). O banco de dados completo pode ser obtido em <`http://www.vision.ime.usp.br/~mris/pipeline/mysql_data.zip`>. O Apêndice B do presente trabalho, apresenta uma descrição completa das tabelas do banco de dados.

Os dados foram normalizados e discretizados pelo mesmo método utilizado em [3], com três níveis de discretização $\{-1, 0, +1\}$ e com limitantes l_i e u_i (Equações 4.1 e 4.2) para cada gene i do banco de dados.

Cada gene do banco de dados foi pré-processado:

- Localização do gene no banco de dados "*Stanford MicroArray*" e identificação de seus *aliases*. Quando um gene não foi encontrado, ele foi marcado, para ser removido do processo completo.

- Procura de todos os processos biológicos dos quais o gene está envolvido a partir de "*Gene Ontology Database*" (<`http://www.geneontology.org`>) e armazenamento destes em nosso banco de dados MySql^{TM}.
- Procura de todas as vias às quais o gene está envolvido a partir de "KEGG: Kyoto Encyclopedia of Genes and Genomes" (<`http://www.genome.jp/kegg`>) e armazenamento destas mesmas em nosso banco de dados MySql^{TM}.
- Se o sinal do gene é constante ou indefinido em mais da metade dos experimentos (mais de 8 vezes), o gene foi marcado para ser removido do processo completo.

4.4.2 Algoritmos

O *pipeline* é composto pelos seguintes algoritmos:

- Cada iteração foi processada por um algoritmo desenvolvido em Microsoft Visual Stusio 7.0^{TM} que recebeu os genes-sementes como entrada e o banco de dados MySql^{TM}. Para cada gene G do banco de dados, o algoritmo executa os seguintes processos:
  - extração da submatriz contendo a expressão dos genes-sementes no instante t e o gene-alvo G no instante $t + k$;
  - execução do algoritmo para encontrar o melhor conjunto preditor;
  - armazenamento dos resultados no banco de dados MySql^{TM}.
- O algoritmo para encontrar o melhor conjunto preditor foi desenvolvido em *C++*. Ele pode executar um dos seguintes métodos: algoritmo *U-curve* [33] ou busca exaustiva limitada a um determinado tamanho de subconjunto. A saída é composta por uma tabela de predição que contém os melhores subconjuntos de predição utilizando como função custo a entropia condicional média estimada descrita anteriormente.
- Os resultados foram organizados e processados por um algoritmo desenvolvido em Microsoft Visual Studio 7.0^{TM} que produzem a entrada para a contrução do grafo que representa a rede.
- O pacote *graphviz* (<`http://www.graphviz.org`>) foi utilizado na contrução da imagem do grafo que representa a rede.
- Os redirecionamentos para as tabelas de predições, assim como os gráficos dos sinais de expressão, foram gerados por um algoritmo desenvolvido em Adobe ColdFusion^{TM} (<`http://www.adobe.com/products/coldfusion`>).

Os resultados experimentais foram processados em uma máquina AMD Turion 64^{TM} com 2Gb de memória RAM. O grafo resultante e os genes-sementes para cada iteração podem ser obtidos em <`http://www.vision.ime.usp.br/~mris/pipeline/graph.zip`>.

Capítulo 5

O Estrógeno e a adesão celular

Com base em estudos de série-temporal de *microarray* na presença de estrógeno e iniciando-se com uma lista de genes diretamente regulados por estrógeno "*genes-sementes*", usamos um *pipeline* de algoritmos para a busca de novos genes preditos pelos "genes-sementes". O resultado de nossa análise nos direcionou a um grupo de genes com um novo processo biológico: *adesão celular*. Uma lista completa dos genes e suas tabelas de predição relacionando esses genes aos seus preditores foi obtida.

A partir deste estudo, pudemos obter uma evidência do estrógeno regulando genes relacionados à adesão celular. Esse processo biológico ainda não possuia nenhum estudo relacionando esses genes ao estrógeno, e nossos resultados possuem um valor de significância alto para essa relação. Esse resultado nos direciona a uma nova vertente de estudos sobre o estrógeno em tratamentos de câncer, podendo relacionar ele, também, ao estudo de metástase [30].

Em nosso trabalho, procuramos encontrar genes responsivos ao estrógeno de maneira não-direta, isto é, genes cuja expressão é predita por genes regulados diretamente pelo estrógeno. Os genes regulados diretamente por estrógeno são aqueles que participam de algumas das vias clássicas descritas anteriormente [4].

5.1 O processo

A partir de um experimento de série-temporal de *microarray* obtido do trabalho em [26], no qual foram caracterizados 89 genes regulados diretamente pelo estrógeno e utilizando o modelo para redes genéticas, denominado de *redes genéticas probabilísticas* (*PGNs*) [3], obtivemos um conjunto de genes melhores preditos pelo conjunto dos genes diretamente regulados pelo estrógeno. Para isso, um processo baseado no *pipeline* de algoritmos, visto na Seção 4, e fundamentado no modelo de *PGNs* foi desenvolvido. Esse processo pode ser resumido pelas seguintes etapas:

- A entrada é a série-temporal de *microarray* em [26]: um experimento que mede a expressão dos genes ($> 19,000$ genes) de uma cultura celular *T-47D* submetidas ao estrógeno durante 16 instantes em 24 horas. Essas expressões foram comparadas a expressão de uma cultura não submetida ao estrógeno (controle) durante as mesmas 24 horas. Uma lista com 89 genes foi obtida no experimento indicando regulação direta do estrógeno sobre eles.

- Localização do gene no banco de dados "*Stanford MicroArray*" e identificação de seus *aliases*. Quando um gene não foi encontrado, ele foi marcado para ser removido do processo completo.

- Identificação das funções biológicas de todos os genes do experimento. Esse processo foi feito pela submissão do identificador dos genes ao banco de dados de ontologia genética (*Gene Ontology Database*).

- Normalização e discretização dos dados do experimento. Os dados do experimento de entrada podem ser vistos como uma matriz onde cada linha representa o sinal de expressão de um gene do experimento. Esse sinal é um número real: positivo se o gene é super-expresso, próximo a 0 se o gene possui expressão normal, e negativo se o gene é sub-expresso. Esses sinais foram normalizados e discretizados em três níveis, $\{-1, 0, +1\}$, pelo mesmo método empregado em [3].

- Filtragem dos genes. Os genes com expressão constante, ou os que não foram localizados no banco de dados de *microarray* ("*Stanford MicroArray*"), foram descartados do processo. Após essa filtragem, o conjunto de 89 genes, regulados diretamente pelo estrógeno em [26], foi reduzido a um conjunto de 53 genes. Esse subconjunto é denominado de "*genes-sementes*".

- Para cada um dos genes (já filtrados) procura-se, entre todos os subconjuntos possíveis (2^{53} possibilidades) dos genes-sementes, o "melhor" subconjunto preditor. Esse subconjunto tem a seguinte característica: a distribuição conjunta estimada das expressões dos genes que compõe o subconjunto e o gene analisado possui menor *entropia condicional média* [3]. Encontrar o melhor conjunto preditor pode ser visto como um problema de seleção de características em Reconhecimentos de Padrão. Várias heurísticas, como: *SFFS* e *SFFS* [32], e algoritmos, como: *o algoritmo U-curve* [33], existem para tratar esse problema. Neste trabalho, usamos a busca exaustiva limitada a um subconjunto de tamanho 3.

- Ordenação dos genes em ordem crescente de custo (*ranking*). Os genes no topo da lista são os genes melhores preditos pelo conjunto de genes-sementes.

- Seleção dos genes melhores preditos. Para essa seleção são escolhidos os genes onde o melhor subconjunto preditor associado a ele possue menor custo (entropia condicional média – valor $< 0,08$).

- Agrupamento dos genes por suas funções biológicas.

A Figura 5.1 é uma representação esquemática do processso.

5.1.1 Genes regulados diretamente por estrógeno

O experimento em [26] obtém 386 genes regulados pelo estrógeno. Para a obtenção dos genes regulados diretamente pelo estrógeno mediu-se os sinais de expressão para duas novas séries-temporais: uma tratada com estrógeno (E2) mais um componente anti-estrógeno (*ICI*), e outra tratada com etrógeno mais um componente inibidor de síntese de proteína (*CHX-cycloheximide*). O primeiro experimento filtra os genes que não se expressaram mais na presença do componente

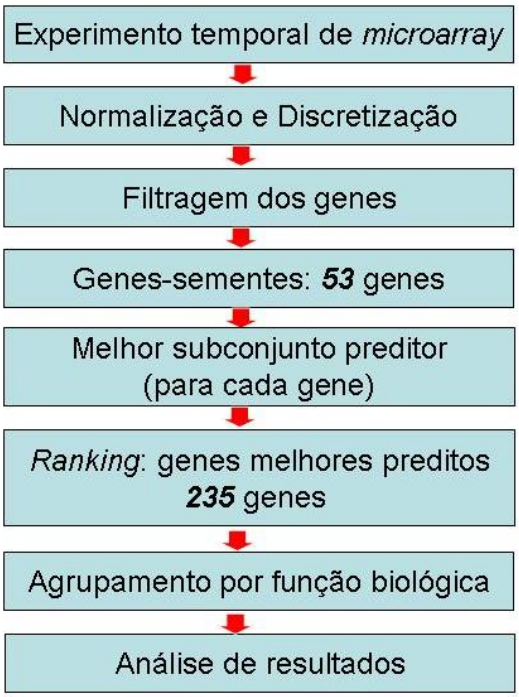


Figura 5.1: Representação esquemática do processo utilizado no experimento

anti-estrógeno (*ICI*), ou seja, insensíveis ao componente, resultando em 137 genes. O segundo experimento filtra os genes que não se expressaram mais na presença do componente inibidor de proteína (*CHX*), ou seja, ficando somente os genes expressos sem a presença de nenhum outro fator de transcrição subseqüente, resultando em 89 genes. Desses 89 genes filtramos, para o nosso experimento, os genes que possuem expressão constante (não são informativos para nosso processo) e os genes não encontrados no banco de dados de *microarray* (*Stanford MicroArray*), resultando em um conjunto de 53 genes.

A Tabela 5.1 apresenta os 53 genes com suas respectivas entradas *KEGG* e funções biológicas (ontologias). A Tabela 5.2 e a Figura 5.2 agrupam os genes-sementes entre as funções biológicas mais freqüentes (> 2 ocorrências). Podemos verificar que os genes com funções biológicas de transdução de sinal (*signal transduction*) e regulação da trasncrição (*regulation of transcription, DNA-dependent*) estão entre as funções mais freqüentes. As funções de via de sinalização intracelular (*intracellular signaling cascade*) e coagulação sangüínea (*blood coagulation*) merecem destaque nesse conjunto, já que são bem representativas e possuem alto valor de qui-quadrado ($\mathcal{X}^2$), respectivamente, $28,320$ e $64,993$.

5.1.2 Genes resultantes

Os genes melhores preditos pelo conjunto de genes-sementes foi obtido pela ordenação (*ranking*) em ordem crescente de custo (entropia condicional média) relativo ao melhor subconjunto preditor do gene. Foram filtrados os genes cujos custos eram menores que $0,08$, obtendo-se uma lista

Gene	KEGG	Funções Biológicas (Ontologia)
ABCA3		ATP binding ; ATPase activity ; ATPase activity, coupled to transmembrane movement of substances ; integral to membrane ; membrane fraction ; nucleotide binding ; response to drug ; transport
ADCY9		adenylate cyclase activity ; cAMP biosynthesis ; integral to plasma membrane ; intracellular signaling cascade ; isomerase activity ; magnesium ion binding ; peptidyl-prolyl cis-trans isomerase activity ; protein folding
AFG3L2		ATP binding ; integral to membrane ; membrane ; metalloendopeptidase activity ; mitochondrion ; nucleoside-triphosphatase activity ; nucleotide binding ; protein catabolism ; proteolysis and peptidolysis ; unfolded protein binding ; zinc ion binding
ALOX12B		arachidonate 12-lipoxygenase activity ; electron transport ; electron transporter activity ; epidermis development ; iron ion binding ; leukotriene biosynthesis ; lipid metabolism ; lipoxygenase activity ; oxidoreductase activity
AMD1		adenosylmethionine decarboxylase activity ; cellular component unknown ; lyase activity ; spermidine biosynthesis ; spermine biosynthesis
ARPP-21		biological process unknown ; cellular component unknown ; nucleic acid binding
BMP7	[CYT][TGF]	cell differentiation ; cytokine activity ; growth ; growth factor activity ; skeletal development
BRI3BP		
CAP350		cytoskeleton
CCNG2		cell cycle checkpoint ; cytokinesis ; mitosis ; regulation of cell cycle
CD7		T cell activation ; calcium ion transport ; cellular defense response ; integral to membrane ; membrane fraction ; plasma membrane ; receptor activity ; transmembrane receptor protein tyrosine kinase signaling pathway
CENTG1		GTP binding ; GTPase activator activity ; GTPase activity ; kinase activity ; nucleus ; regulation of GTPase activity ; signal transduction ; small GTPase mediated signal transduction
CISH	[JAK]	cellular component unknown ; intracellular signaling cascade ; molecular function unknown ; regulation of cell growth
CRABP2		epidermis development ; lipid binding ; regulation of transcription, DNA-dependent ; retinoid binding ; signal transduction ; transport ; transporter activity
CTBS		carbohydrate metabolism ; hydrolase activity, acting on glycosyl bonds ; lysosome ; metabolism
CTSD		cathepsin D activity ; extracellular region ; lysosome ; pepsin A activity ; peptidase activity ; proteolysis and peptidolysis
DGKZ		ATP binding ; diacylglycerol binding ; diacylglycerol kinase activity ; diacylglycerol kinase activity ; intracellular signaling cascade ; kinase activity ; nucleus ; protein kinase C activation ; transferase activity
ELOVL2		endoplasmic reticulum ; fatty acid biosynthesis ; integral to membrane
EPHA4		ATP binding ; ephrin receptor activity ; integral to plasma membrane ; membrane ; protein amino acid phosphorylation ; receptor activity ; signal transduction ; transferase activity ; transmembrane receptor protein tyrosine kinase signaling pathway
F10		blood coagulation ; calcium ion binding ; chymotrypsin activity ; coagulation factor Xa activity ; extracellular region ; peptidase activity ; proteolysis and peptidolysis ; trypsin activity
FLJ13710		extracellular matrix ; metalloendopeptidase activity ; peptidase activity
FLJ20986		
FLJ22269		integral to membrane
GALNT4		Golgi apparatus ; carbohydrate metabolism ; integral to membrane ; manganese ion binding ; polypeptide N-acetylgalactosaminyltransferase activity ; sugar binding ; transferase activity, transferring glycosyl groups
GREB1		biological process unknown ; cellular component unknown ; molecular function unknown
HIG2		integral to membrane ; molecular function unknown ; response to stress
HSPC111		nucleus
IGFBP4		DNA metabolism ; cell proliferation ; extracellular region ; insulin-like growth factor binding ; regulation of cell growth ; signal transduction ; skeletal development
IGSF4		
IL6ST	[CYT][JAK]	
JAK1	[JAK]	ATP binding ; Janus kinase activity ; cytoskeleton ; intracellular signaling cascade ; protein amino acid phosphorylation ; protein-tyrosine kinase activity ; transferase activity
KCNG1		cation transport ; membrane ; membrane fraction ; potassium ion transport ; protein binding ; voltage-gated potassium channel activity ; voltage-gated potassium channel complex
LMCD1		biological process unknown ; cellular component unknown ; zinc ion binding
LOR		insoluble fraction ; structural constituent of cytoskeleton

Gene	KEGG	Funções Biológicas (Ontologia)
MPP3		guanylate kinase activity ; integral to plasma membrane ; protein binding ; signal transduction
NFIA		DNA replication ; electron transport ; heme binding ; nucleus ; nucleus ; regulation of transcription, DNA-dependent ; regulation of transcription, DNA-dependent ; transcription ; transcription factor activity ; transcription factor activity ; viral genome replication
NIF3L1		
NOL5A		RNA binding ; nucleolus ; rRNA processing
NRIP1		nucleus ; regulation of transcription, DNA-dependent ; transcription ; transcription coactivator activity
OLFM1		development ; endoplasmic reticulum ; latrotoxin receptor activity ; membrane ; neurogenesis
PAFAH1B1		astral microtubule ; cell cortex ; cell cycle ; cell differentiation ; cell motility ; cytokinesis ; cytoskeleton ; dynein binding ; establishment of mitotic spindle orientation ; kinetochore ; lipid metabolism ; microtubule associated complex ; microtubule-based process ; mitosis ; neurogenesis ; nuclear membrane ; signal transduction
PGR		cell-cell signaling ; nucleus ; regulation of transcription, DNA-dependent ; signal transduction ; steroid binding ; steroid hormone receptor activity ; steroid hormone receptor activity ; transcription ; transcription factor activity ; transcription from RNA polymerase II promoter
PTGES		antimicrobial humoral response (sensu Vertebrata) ; integral to membrane ; isomerase activity ; membrane fraction ; prostaglandin metabolism ; prostaglandin-E synthase activity ; signal transduction
RBBP8		
RFPL2		protein binding ; protein ubiquitination ; ubiquitin ligase complex ; ubiquitin-protein ligase activity ; zinc ion binding
SCN1B		integral to membrane ; ion channel activity ; ion transport ; membrane fraction ; sodium ion transport ; synaptic transmission ; voltage-gated sodium channel activity
SERPINE1		blood coagulation ; extracellular region ; plasminogen activator activity ; serine-type endopeptidase inhibitor activity
SIAH2		apoptosis ; cell cycle ; cytoplasm ; development ; ligase activity ; nucleus ; protein ubiquitination ; small GTPase mediated signal transduction ; transcription corepressor activity ; ubiquitin ligase complex ; ubiquitin-dependent protein catabolism ; ubiquitin-protein ligase activity ; zinc ion binding
SLC38A1		amino acid transport ; amino acid-polyamine transporter activity ; integral to membrane ; integral to membrane ; membrane ; neutral amino acid transport ; neutral amino acid transporter activity ; sodium:amino acid symporter activity ; transport
STC2		cell surface receptor linked signal transduction ; cell-cell signaling ; extracellular region ; hormone activity ; response to nutrients
THBS1	[TGF][COMM]	blood coagulation ; calcium ion binding ; cell adhesion ; cell motility ; development ; endopeptidase inhibitor activity ; extracellular region ; heparin binding ; neurogenesis ; protein binding ; signal transducer activity ; structural molecule activity
TPD52L1		biological process unknown ; cellular component unknown ; molecular function unknown
UGCGL1		UDP-glucose:glycoprotein glucosyltransferase activity ; endoplasmic reticulum ; posttranslational protein folding ; protein amino acid glycosylation ; protein binding ; transferase activity

Tabela 5.1: Lista dos 53 genes-sementes

Grupo	Freqüência (C)	Total	Valor Esperado (E)	$\mathcal{X}^2 : \frac{(\|C-E\|-0.5)^2}{E}$
signal transduction	8 [20.51 %]	862 [4.74 %]	1.850	20.443
regulation of transcription, DNA-dependent	5 [12.82 %]	800 [4.40 %]	1.717	6.277
intracellular signaling cascade	4 [10.26 %]	208 [1.14 %]	0.446	28.287
biological process unknown	4 [10.26 %]	162 [0.89 %]	0.348	38.364
neurogenesis	3 [7.69 %]	185 [1.02 %]	0.397	17.063
blood coagulation	3 [7.69 %]	59 [0.32 %]	0.127	64.199
transcription	3 [7.69 %]	601 [3.31 %]	1.290	2.267
transport	3 [7.69 %]	335 [1.84 %]	0.719	7.236
proteolysis and peptidolysis	3 [7.69 %]	274 [1.51 %]	0.588	9.892
development	3 [7.69 %]	297 [1.63 %]	0.637	8.756

Tabela 5.2: Distribuição dos genes-sementes nos vários grupos de funções biológicas

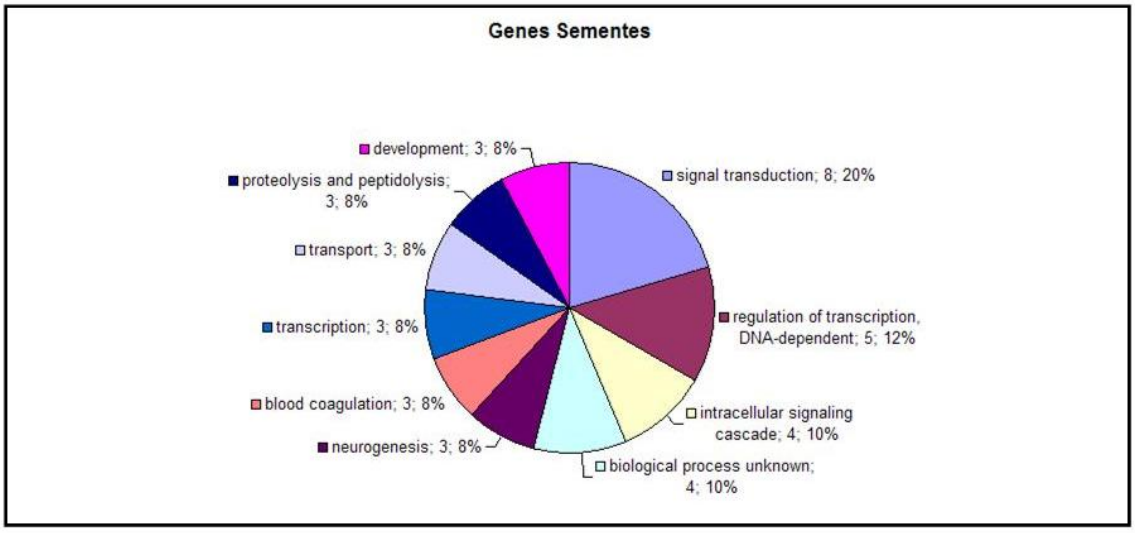


Figura 5.2: Gráfico da distribuição dos 53 genes-sementes por suas funções biológicas

de 235 genes. A Tabela 5.3 apresenta as funções biológicas mais freqüentes (> 2 ocorrências) na lista dos 235 genes filtrados, e a Figura 5.3 dá uma visão gráfica da distribuição.

As funções biológicas relacionadas ao estrógeno, já conhecidas, aparecem em destaque: transdução de sinal (*signal transduction*), regulação da transcrição (*regulation of transcription, DNA-dependent*)e fosforilação (*phosphorylation*) aparecem com freqüência entre os genes preditos. Podemos ressaltar a função de *adesão celular* (*cell adhesion*), a qual aparece com destaque no conjunto dos genes preditos: 19 genes (8 %) com grande significância ($\mathcal{X}^2 = 28.906$). Esse resultado nos leva a concluir que o estrógeno pode estar relacionado também com genes de adesão celular. A Tabela 5.4 mostra os 19 genes de adesão celular obtidos. A coluna de preditores indica os melhores cinco conjuntos preditores para cada gene.

Grupo	Freqüência (C)	Total	Valor Esperado (E)	$\mathcal{X}^2 : \frac{(\lvert C-E\rvert-0.5)^2}{E}$
signal transduction	29 [7.51 %]	862 [4.74 %]	18.311	5.670
regulation of transcription, DNA-dependent	25 [6.48 %]	800 [4.40 %]	16.994	3.315
transcription	20 [5.18 %]	601 [3.31 %]	12.767	3.551
cell adhesion	19 [3.63 %]	276 [1.52 %]	5.863	28.906
protein amino acid phosphorylation	19 [4.92 %]	352 [1.94 %]	7.477	16.251
transport	14 [3.63 %]	335 [1.84 %]	7.116	5.727
G-protein coupled receptor protein signaling pathway	11 [2.85 %]	256 [1.41 %]	5.438	4.712
cell differentiation	10 [2.59 %]	154 [0.85 %]	3.271	11.862
proteolysis and peptidolysis	10 [2.59 %]	274 [1.51 %]	5.820	2.327
development	10 [2.59 %]	297 [1.63 %]	6.309	1.614
electron transport	10 [2.59 %]	173 [0.95 %]	3.675	9.233
immune response	10 [2.59 %]	184 [1.01 %]	3.909	7.997
cell proliferation	10 [2.59 %]	199 [1.10 %]	4.227	6.578
cell cycle	9 [2.33 %]	160 [0.88 %]	3.399	7.655
neurogenesis	9 [2.33 %]	185 [1.02 %]	3.930	5.314
intracellular signaling cascade	8 [2.07 %]	208 [1.14 %]	4.418	2.150
cell motility	8 [2.07 %]	89 [0.49 %]	1.891	16.637
carbohydrate metabolism	7 [1.81 %]	148 [0.81 %]	3.144	3.582
cell-cell signaling	7 [1.81 %]	208 [1.14 %]	4.418	0.981
lipid metabolism	7 [1.81 %]	137 [0.75 %]	2.910	4.429
ion transport	7 [1.81 %]	218 [1.20 %]	4.631	0.754
sodium ion transport	6 [1.55 %]	72 [0.40 %]	1.529	10.313
protein biosynthesis	6 [1.55 %]	180 [0.99 %]	3.824	0.735
metabolism	5 [1.30 %]	219 [1.21 %]	4.652	0.005
inflammatory response	5 [1.30 %]	116 [0.64 %]	2.464	1.682
protein amino acid glycosylation	5 [1.30 %]	40 [0.22 %]	0.850	15.674
potassium ion transport	5 [1.30 %]	120 [0.66 %]	2.549	1.493
transcription from RNA polymerase II promoter	5 [1.30 %]	139 [0.76 %]	2.953	0.810
spermatogenesis	5 [1.30 %]	73 [0.40 %]	1.551	5.607
apoptosis	5 [1.30 %]	175 [0.96 %]	3.717	0.165

Grupo	Freqüência (C)	Total	Valor Esperado (E)	$\mathcal{X}^2 : \frac{(\lvert C-E\rvert -0.5)^2}{E}$
sensory perception	4 [1.04 %]	116 [0.64 %]	2.464	0.436
phosphate metabolism	4 [1.04 %]	18 [0.10 %]	0.382	25.450
biological process unknown	4 [1.04 %]	162 [0.89 %]	3.441	0.001
morphogenesis	4 [1.04 %]	84 [0.46 %]	1.784	1.651
protein folding	4 [1.04 %]	138 [0.76 %]	2.931	0.110
phosphate transport	4 [1.04 %]	45 [0.25 %]	0.956	6.770
regulation of cyclin dependent protein kinase activity	4 [1.04 %]	28 [0.15 %]	0.595	14.183
chemotaxis	3 [0.78 %]	48 [0.26 %]	1.020	2.147
regulation of transcription from RNA polymerase II promoter	3 [0.78 %]	132 [0.73 %]	2.804	0.033
phospholipase C activation	3 [0.78 %]	13 [0.07 %]	0.276	17.921
intracellular protein transport	3 [0.78 %]	120 [0.66 %]	2.549	0.001
epidermal growth factor receptor signaling pathway	3 [0.78 %]	15 [0.08 %]	0.319	14.911
regulation of cell growth	3 [0.78 %]	46 [0.25 %]	0.977	2.374
DNA repair	3 [0.78 %]	92 [0.51 %]	1.954	0.153
cell surface receptor linked signal transduction	3 [0.78 %]	109 [0.60 %]	2.315	0.015
protein amino acid dephosphorylation	3 [0.78 %]	109 [0.60 %]	2.315	0.015
negative regulation of cell cycle	3 [0.78 %]	60 [0.33 %]	1.275	1.177
DNA replication	3 [0.78 %]	77 [0.42 %]	1.636	0.456
glycogen biosynthesis	3 [0.78 %]	9 [0.05 %]	0.191	27.914
regulation of cell cycle	3 [0.78 %]	162 [0.89 %]	3.441	0.001
positive regulation of cytosolic calcium ion concentration	3 [0.78 %]	34 [0.19 %]	0.722	4.379
visual perception	3 [0.78 %]	121 [0.67 %]	2.570	0.002
ubiquitin cycle	3 [0.78 %]	104 [0.57 %]	2.209	0.038
digestion	3 [0.78 %]	34 [0.19 %]	0.722	4.379

Tabela 5.3: Distribuição dos 235 genes preditos por suas funções biológicas

Gene	Preditores	Entopia
CLDN18	(0.00000000):RFPL2;IL6ST[CYT][JAK] —— (0.09090900):TPD52L1;RFPL2;IL6ST[CYT][JAK] —— (0.09090900):RFPL2;PAFAH1B1;IL6ST[CYT][JAK] —— (0.23076900):JAK1[JAK];SIAH2;PAFAH1B1 —— (0.23076900):ABCA3;FLJ20986;	0.000
L1CAM	(0.00000000):ALOX12B —— (0.07692300):NRIP1;ALOX12B —— (0.07692300):F10;ALOX12B —— (0.07692300):AMD1;ALOX12B —— (0.07692300):PTGES;ALOX12B	0.000
CDH17	(0.00000000):CAP350;BRI3BP —— (0.00000000):BRI3BP —— (0.07692300):F10;BRI3BP —— (0.07692300):PTGES;BRI3BP —— (0.07692300):AFG3L2;PAFAH1B1;BRI3BP	0.000
CDH16	(0.00000000):ALOX12B —— (0.07692300):NRIP1;ALOX12B —— (0.07692300):F10;ALOX12B —— (0.07692300):AMD1;ALOX12B —— (0.07692300):PTGES;ALOX12B	0.000
LSAMP	(0.06666600):BMP7[CYT][TGF];HSPC111 —— (0.06666600):BMP7[CYT][TGF];GALNT4 —— (0.11587600):HSPC111 —— (0.11587600):GALNT4;KCNG1 —— (0.13333300):HSPC111;JAK1[JAK]	0.067
ERBB2IP	(0.06666600):THBS1[TGF][COMM];CRABP2 —— (0.14285700):TPD52L1;THBS1[TGF][COMM];FLJ22269 —— (0.15079000):AMD1;CRABP2 —— (0.20316200):CTSD;THBS1[TGF][COMM] —— (0.21745700):F10;CRABP2	0.067
SPOCK	(0.06666600):CTSD;DGKZ —— (0.08412300):CTBS;NOL5A —— (0.11587600):DGKZ —— (0.13333300):CTBS;NOL5A;DGKZ —— (0.13333300):CTBS;DGKZ	0.067
COL4A6	(0.06666600):SIAH2;STC2 —— (0.08333300):TPD52L1;RFPL2 —— (0.08412300):NRIP1;PTGES —— (0.08412300):PTGES;STC2 —— (0.09013200):CAP350;BRI3BP	0.067
COL17A1	(0.06666600):SCN1B;CISH[JAK] —— (0.15079000):NRIP1;PTGES;STC2 —— (0.18254200):UGCGL1;PTGES —— (0.20000000):NFIA;SCN1B;CISH[JAK] —— (0.20000000):NRIP1;PTGES	0.067
PKP4	(0.06666600):CAP350;CD7 —— (0.15079000):SERPINE1;HIG2 —— (0.18254200):PTGES;CD7 —— (0.20000000):BMP7[CYT][TGF];PTGES;CD7 —— (0.20000000):RBBP8;CTSD;CAP350	0.067
TINAG	(0.07142800):EPHA4;IL6ST[CYT][JAK] —— (0.12415200):NFIA;SCN1B —— (0.14285700):NFIA;EPHA4;IL6ST[CYT][JAK] —— (0.14285700):NIF3L1;EPHA4 —— (0.14285700):NIF3L1;OLFM1	0.071
SLIT1	(0.07142800):NOL5A;TPD52L1 —— (0.16156100):PTGES;TPD52L1 —— (0.16156100):TPD52L1;FLJ13710 —— (0.16267300):TPD52L1 —— (0.18848800):TPD52L1;RFPL2	0.071
PTPRS	(0.07142800):AFG3L2;PAFAH1B1;BRI3BP —— (0.13333300):AFG3L2;PAFAH1B1 —— (0.20000000):AFG3L2;LMCD1;PAFAH1B1 —— (0.20000000):AFG3L2;CCNG2;PAFAH1B1 —— (0.21428500):PAFAH1B1;BRI3BP;KCNG1	0.071
F8	(0.07142800):NFIA;SCN1B —— (0.12415200):NFIA —— (0.12415200):NFIA;FLJ20986 —— (0.12415200):FLJ20986 —— (0.12415200):FLJ20986;EPHA4	0.071
CNTNAP2	(0.07692300):FLJ13710;ALOX12B —— (0.15079000):CRABP2;FLJ13710 —— (0.15079000):FLJ13710;HIG2 —— (0.17398900):SERPINE1;ALOX12B —— (0.18254200):GREB1;CD7	0.077
NEO1	(0.07692300):HSPC111;NRIP1 —— (0.08333300):THBS1[TGF][COMM];BRI3BP —— (0.09706600):NRIP1;PTGES —— (0.13370300):HSPC111 —— (0.13370300):HSPC111;IL6ST[CYT][JAK]	0.077
COL8A1	(0.07692300):SERPINE1;ALOX12B —— (0.21062600):FLJ13710;ALOX12B —— (0.21428500):ABCA3;FLJ13710 —— (0.21745700):STC2;FLJ13710 —— (0.23076900):SERPINE1;FLJ13710;ALOX12B	0.077
NRP2	(0.07692300):AMD1;ALOX12B —— (0.14285700):AMD1;ABCA3 —— (0.15079000):THBS1[TGF][COMM];KCNG1 —— (0.18254200):JAK1[JAK];AMD1 —— (0.18254200):CTSD;AMD1	0.077
KITLG	(0.07692300):NRIP1;ALOX12B —— (0.15384600):NRIP1;JAK1[JAK];ALOX12B —— (0.23076900):NRIP1;LMCD1;ALOX12B —— (0.23076900):NRIP1;ELOVL2;ALOX12B —— (0.23076900):NRIP1;CCNG2;ALOX12B	0.077

Tabela 5.4: Lista dos 19 genes de adesão celular obtidos

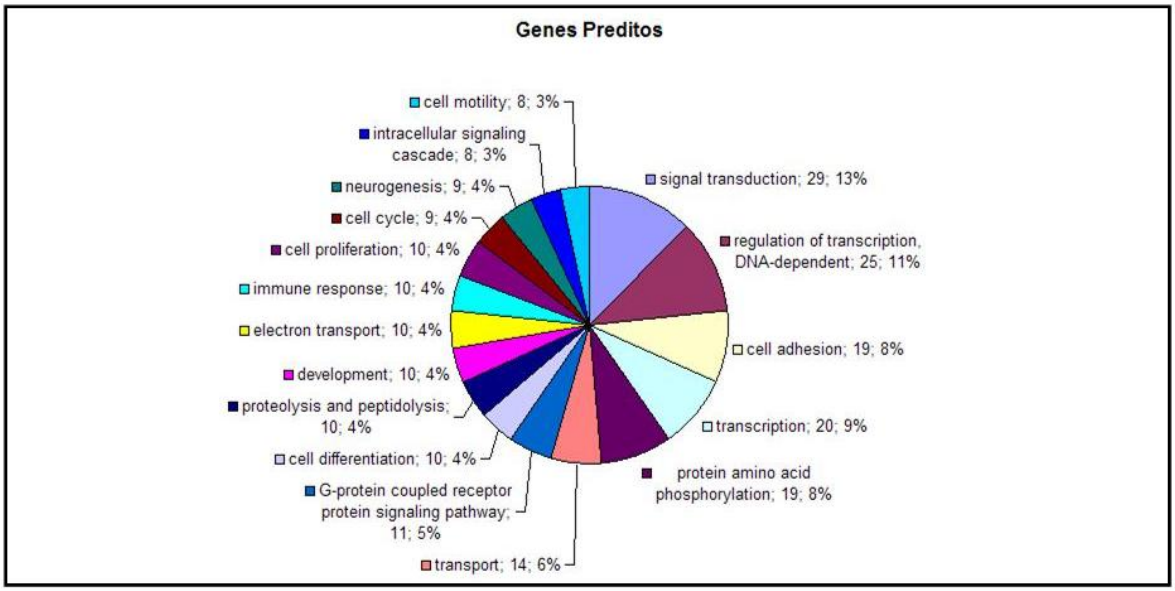


Figura 5.3: Gráfico da distribuição dos 235 genes melhores preditos por suas funções biológicas

No Apêndice C temos: (i) a tabela completa dos 235 genes preditos nessa etapa; (ii) as tabelas de predições para cada um dos 19 genes de adesão celular; e (iii) os gráficos do sinal de expressão destes genes com seus preditores.

5.2 Discussão

O processo de metástase consiste em uma complexa seqüência de etapas que envolvem as células tumorais e propriedades do organismo hospedeiro [11]. O *descolamento* das células tumorais do tumor primário é considerada como a primeira e mais importante etapa no processo metastático. As células tumorais podem mais facilmente ser descoladas de um tecido tumoral compacto do que as células normais nas proximidades de um tecido normal [6]. Esse descolamento das células tumorais é regulada pela propriedade de *adesão celular* do tumor. A função biológica de *adesão celular* é empregada aos genes relacionados a moléculas de adesão, as quais atuam como moduladores positivos ou negativos no processo de *metástase* [18, 30]. Apesar do rápido avanço no entendimento da biologia da adesão celular, os dados disponíveis na literatura tornam difícil a proposição de um modelo simples, no qual pode-se relacionar moléculas de adesão ao crescimento dos tumores e a metástase. Isso pode estar relacionado a um número considerável de fatores. Alguns dos resultados experimentais, aparentemente conflitantes, que demonstram tanto aumento como diminuição da adesão das células tumorais durante a progressão do tumor podem ser atribuídos aos sistemas usados. Aqueles estudos que injetam células tumorais de maneira intravenosa têm, em geral, mostrado que um aumento na função de adesão dessas células possui correlação positiva com a habilidade mestastática. Esses estudos possuem um viés em determinar que a alta função de adesão dessas células faz com que elas tenham facilidade

em se juntar com células da circulação e serem depositadas em regiões diversas do organismo. Entretanto, estudos que implantam tecidos tumorais em organismos e permitem que cresçam e sofram metástase espontânea têm, geralmente, demonstrado uma relação inversa da função de adesão celular e a habilidade metastática.

A relação do estrógeno como regulador de genes relacionados a adesão celular não é muito destacada na literatura. Neste trabalho, conseguimos obter, a partir de um experimeto de série-temporal de *microarray* com células submetidas a estrógeno, uma evidência forte de regulação do estrógeno sobre genes relacionados a essa função. A partir de uma lista inicial de 53 genes regulados diretamente pelo estrógeno [26], uma lista com 20 genes relacionados a adesão celular foi obtida. Para cada um dos genes, uma tabela de predição, relacionando o gene aos genes iniciais diretamente regulados pelo estrógeno, é determinada. Esses resultados, indicam uma forte relação entre o estrógeno e genes de adesão celular e, por conseqüente, a metástase. Os genes obtidos neste trabalho nos dão uma forte evidência da relação entre o estrógeno e a função de adesão celular. Esses dados, porém, precisam ser validados em laboratório. A lista completa dos genes candidatos com suas tabelas de predições são fontes valiosas ao pesquisador para a validação dos resultados. O método empregado em nossos resultados possui forte suporte matemático, ou seja, dado o conjunto de entrada: o experimento e os genes-sementes iniciais – o melhor subconjunto preditor é sempre obtido pela minimização da entropia condicional média (custo) associada à distribuição conjunta do gene e o subconjunto preditor. Sendo assim, à medida que o conjunto de entrada cresce (mais instantes amostrais), o resultado se torna mais confiável. Apesar de, no nosso caso, o experimento de entrada ser pequeno (16 instantes), acreditamos que a evidência do estrógeno influenciando genes de adesão celular é fortemente comprovada pelos resultados.

5.3 Materiais e métodos

5.3.1 Série-temporal de *microarray*

Os dados de entrada de nosso trabalho são baseados nos resultados do experimento em Liu et. Al (2004) [26]. O experimento é composto por uma série-temporal de *microarray* de células *T47-D* tratadas, durante 24 horas, com: (i) estrógeno (17β-estradiol (E2)); (ii) estrógeno (E2) + ICI (componente anti-estrógeno); iii. estrógeno (E2) + *CHX-Cycloheximide* (componente inibidor de síntese de proteína). O sinal de controle é dado pelo mesmo experimento em células *T-47D* não tratadas com estrógeno durante as mesmas 24 horas. O número total de experimentos é de 16 pares (tratados com estrógeno e seu sinal de controle): os primeiros oito foram medidos a cada hora e, os oito restantes a cada duas horas. O experimento obteve: 385 genes, responsivos ao estrógeno; 139 genes, responsivos ao estrógeno e sensíveis a ICI; 89 genes, responsivos ao estrógeno, sensíveis a ICI e insensíveis a CHX. Esses genes foram denominados como *diretamente regulados* pelo estrógeno. Os dados foram inteiramente armazenados em um banco de dados MySql^{TM} (disponível em <`http://www.mysql.com`>). O banco de dados completo pode ser obtido em: <`http://www.vision.ime.usp.br/~mris/pipeline/mysql_data.zip`> e no *DVD* em anexo. Uma descrição completa das tabelas do banco é obtida no Apêndice B.

Os dados foram normalizados e discretizados pelo mesmo método utilizado em [3] com três níveis de discretização $\{-1, 0, +1\}$ com limitantes l_i e u_i (descritos anteriormente) para cada

gene i do banco de dados.

Cada gene do banco de dados foi pré-processado:

- Localização do gene no banco de dados "*Stanford MicroArray*" e identificação de seus *aliases*. Quando um gene não foi encontrado, ele foi marcado para ser removido do processo completo.
- Procura de todos os processos biológicos dos quais o gene está envolvido a partir de "*Gene Ontology Database*" (<`http://www.geneontology.org`>) e armazenamento dos mesmos em nosso banco de dados MySql^{TM}.
- Procura de todas as vias as quais o gene está envolvido a partir de "KEGG: Kyoto Encyclopedia of Genes and Genomes" (<`http://www.genome.jp/kegg`>) e armazenamento das mesmas em nosso banco de dados MySql^{TM}.
- Se o sinal do gene é constante ou indefinido em mais da metade dos experimentos (mais de oito vezes), o gene foi marcado para ser removido do processo completo.

Os processos foram desenvolvidos a partir de:

- Adobe ColdFusion^{TM} (<`http://www.adobe.com/products/coldfusion`>): agrupamento em funções biológicas, tabelas de predições, gráficos de sinais e resultados.
- Microsoft Visual Studio 7.0^{TM}: normalização, discretização, processamento e preparação de cada gene para obtenção do melhor subconjunto preditor.
- *C++*: algoritmo para encontrar o melhor conjunto preditor

Os resultados experimentais foram processados em uma máquina AMD Turion 64^{TM} com 2Gb de memória *RAM*.

Capítulo 6

Conclusão

Foi apresentado, neste trabalho, resultados em três áreas distintas: Ciência da Computação e Estatística, Bioinformática e Biologia.

Na área da Ciência da Computação e Estatística, desenvolvemos uma solução para o problema de seleção de características em Reconhecimentos de Padrões. Esse problema se caracteriza por buscar, no espaço de subconjuntos das carcaterísticas, o subconjunto com menor função custo associada. No contexto de identificação de redes de expressão gênica, o problema se caracteriza por, dado um conjunto de genes inicial, encontrar o melhor subconjunto preditor para um gene-alvo. Nossa solução, denominada algoritmo *U-curve*, executa uma busca completa no espaço total de possibilidades sem a necessidade de percorrer o espaço total. Obtivemos resultados consideráveis quando comparamos nosso método com as heurísticas mais usadas para esse problema. Apesar dos resultados obtidos, a presente versão do algoritmo *U-curve* não é uma solução rápida para problemas de dimensões altas, proporcionando algumas novas frentes de pesquisa, tais como: (i) construção de soluções paralelizáveis do algoritmo; (ii) desenvolvimento de novos cortes à formulação *branch-and-bound*; (iii) desenho e estimação das distribuições das características. Com tudo isso, nossa técnica abre um novo *framework* de pesquisa para problemas de seleção de características.

Na área da Bioinformática, descrevemos um método abrangente para modelar redes genéticas utilizando como entrada os dados de um experimento de série-temporal de *microarray*. Esse método é baseado em um *pipeline* de algoritmos, no qual cada um deles foi descrito em detalhes e uma solução foi implementada. Obtivemos, como resultado experimental, uma rede de predição representada por um grafo dirigido em que cada nó representa um gene da rede. Esse grafo possui uma apresentação em formato *html* onde, para cada gene, existem redirecionamentos para: (i) suas tabelas de predições; (ii) gráfico das expressões do gene-alvo e seus preditores; (iii) informações do gene e seus *aliases* no banco em "*Stanford MicroArray*" (<`http://genome-www5.stanford.edu`>); (iv) ontologia; e (v) via *KEGG* relacionadas. O *pipeline*, quando empregado a um experimento biológico, é uma ferramenta muito útil ao pequisador, podendo, a partir da rede obtida, desenvolver novas frentes de estudos. Como pesquisas futuras, podemos citar: (i) simulação e validação das redes obtidas; e (ii) construção de um utilitário como ferramenta para a análise de experimentos biológicos.

Para a área da Biologia, evidenciamos que os marcadores dependentes da proliferação ce-

lular do estrógeno estão basicamente relacionados às seguintes funções biológicas: regulação de transcrição (*regulation of transcription – DNA-dependent*), ligação com DNA (*DNA binding*), transdução de sinal (*signal transduction*) e regulação do ciclo celular (*regulation of cell cycle*). Isso foi obtido agrupando-se os genes regulados diretamente pelo estrógeno dos resultados em [26] em suas diversas funções biológicas. A partir desses mesmos genes e utilizando o nosso *pipeline* para uma iteração, obtivemos um conjunto de 235 genes preditos. Este conjunto, quando agrupado segundo suas funções biológicas, apresentou: a *adesão celular* (*cell adhesion*) – 19 genes com uma alta significância ($\mathcal{X}^2 = 28.906$) – como uma função de relevância entre aquelas obtidas pelos genes preditos. A lista desses genes, com suas tabelas de predições associadas, é uma fonte de dados promissora para estudos que relacionam estrógeno como fator de proliferação celular ao processo de adesão celular e, conseqüentemente, à metástase.

Apêndice A

Algoritmo U-curve

Neste Apêndice, apresentaremos o programa desenvolvido em C++, que implementa o algoritmo *U-curve*. O programa é executado pela linha de comando (por exemplo: *uc -iamostras.dat -lamostras.log*) e pode receber como entrada os seguintes parâmetros:

- *-i<nome>*: nome do arquivo texto de entrada – uma matriz $n \times m$ separada por espaços e fim de linha onde cada linha representa uma amostra com $n-1$ colunas representando os valores para cada uma das componentes do conjunto de preditores e a última o valor para a componente predita;
- *-o<nome>*: nome do arquivo que contém os resultados (padrão = *amostras.dat*);
- *-l<nome>*: nome do arquivo de *log*. Este arquivo contém dados sobre todo o processamento, servindo como análise do processo (padrão = *uc_res.txt*);
- *-L< n >*: nível de informações (n) sobre o processamento – 0 (sem informação) a 3 (informação completa) (padrão = 1);
- *-t< t >*: tipo de processamento (t) – 0 (algumas heurísticas); 1 (U-curve sem esgotamento do mínimo); 2 (U-curve com esgotamento do mínimo); 3 (busca exaustiva); 4 (SFFS) (padrão = 1);
- *-c< c ><t>*: forma de cálculo da função custo (c) – 0 (custo sem penalização pela freqüência limitante), 1 (custo de teste com mínimos no meio do reticulado), 2 (custo de teste com dois mínimos no meio do reticulado), 3 (custo de teste com mínimos em lugares aleatórios), 4 (custo com penalização pela freqüência limitante) (padrão = 4); tipo da função custo (t) – 0 (entropia condicional média), 1 (Coeficiente de determinação)¿ (padrão = 0);
- *-r< p >*: probabilidade (p) de se começar a procura em sentido de Inferior-Superior (0) ou Superior-Inferior (1) (padrão = $0,50$);
- *-d< f >*: relachamento da curva em U, para continuar a construção da cadeia para diferenças entre o mínimo encontrado e o último menores que *fração* (padrão = 0);
- *-e< i >*: valor que indica se o procedimento de esgotamento do mínimo deve ser suspenso se não encontrar um elemento menor que o último encontrado nas próximas i recursões (padrão = 0, não pára);
- *-s< s >*: valor que indica se o algoritmo deve ser suspenso se não encontrar um elemento menor que o último encontrado nas próximas s iterações (padrão = 0, não pára);
- *-S< S >*: numéro S de iterações até o algoritmo parar (padrão = 0, não pára);
- *-#< m >*: valor mínimo para o custo m, ao atingir este valor o algoritmo pára (padrão = 0, vai até o final);
- *-n< f >*: freqüência limitante f (Seção 4.2.2) para a estimação do custo (padrão = 2);
- *-k< s >*: cadeia de 0s e 1s indicando para retornar apenas o custo associado ao elemento que s representa;

- -*f*< n >: f contém um dos seguintes valores: 0 (não roda o SFFS antes), 1 (roda o SFFS antes do U-curve), 2 (roda o SFFS e pára) (padrão = 0);
- -*F*< d >: parâmetro delta para o SFFS (padrão = 3);
- -*E*< n >: executa busca exautiva limitando o nível em n, valores negativos começam no sentido Superior-Inferior;
- -*M*< m >: número máximo m de elementos no resultado final (padrão = 5);
- -*w*: processa uma heurística de pesos para cada componente a ser inserida, ou seja, componentes que melhoram o resultado possuem mais chance de serem selecionadas;
- -*u*< n >: testa oscilações para n curvas.

O código a seguir apresenta o código fonte para a construção de uma cadeia (`FCL_ProcessChain`) e para o esgotamento do mínimo (`FCL_ExhaustProcess`).

```
/* ************************************************************************   */
/* PROCEDIMENTO de construção da cadeia                                       */
/*            PLP_IniNode:  início da cadeia (elemento minimal ou maximal)    */
/*            PLC_Type: 'U' (sentido Inferior-Superior)                       */
/*                      'D' (sentido Superior-Inferior)                       */
/* ************************************************************************   */
int ProcessClassExhaustive::FCL_ProcessChain(ReticNode *PLP_IniNode, char PLC_Type)
{
   ReticNode *VLP_Aux, *VLP_Min;
   ReticNode VLS_Node1;
   long int VLN_Attrib;
   long int VLN_StepsForMin;

   if(VGN_LogLevel >= 2)
   {
      fprintf(VGF_Log, "\n\n---------------------");
      fprintf(VGF_Log, "\nChain Process BEGINS\n");
      printf("\n\n-------------------------------");
      printf("\nChain Process BEGINS\n");
   }

   this->VCO_Path.FCL_FreeList();

   VLP_Aux = PLP_IniNode;
   if(VGN_LogLevel >= 2)
   {
      if(PLC_Type == 'U')
      {
```

```
      fprintf(VGF_Log, "\n\nGoing UP: Processing Exhaustive\n");
      printf("\n\nGoing UP: Processing Exhaustive\n");
      fprintf(VGF_Log, "\n\nInitial Element from %ld nodes:\n",
            this->VCO_Lower.VCN_Nodes);
      printf("\n\nInitial Element from %ld nodes:\n",
            this->VCO_Lower.VCN_Nodes);
      this->VCO_Upper.FCL_ShowElement(VLP_Aux);
   }
   else
   {
      fprintf(VGF_Log, "\n\nGoing DOWN: Processing Exhaustive\n");
      printf("\n\nGoing DOWN: Processing Exhaustive\n");
      fprintf(VGF_Log, "\n\nInitial Element from %ld nodes:\n",
            this->VCO_Upper.VCN_Nodes);
      printf("\n\nInitial Element from %ld nodes:\n",
            this->VCO_Upper.VCN_Nodes);
      this->VCO_Upper.FCL_ShowElement(VLP_Aux);
   }
}


/* begins the path */
VLP_Aux->FCN_CalcCost();
this->VCO_Path.FCL_AddNode(VLP_Aux->attrib, VLP_Aux->cost);
if(VGN_LogLevel >= 3)
{
   fprintf(VGF_Log, "\nElements In the Path:\n");
   printf("\nElements In the Path:\n");
}

VLN_StepsForMin = this->VCN_StepsForMin;
VLP_Aux = this->VCO_Path.VCP_End;
VLP_Min = VLP_Aux;
while(this->FCL_LookForNode(this->VCO_Path.VCP_End, &VLS_Node1,
            &VLN_Attrib, PLC_Type))
{
   VLP_Aux = this->VCO_Path.VCP_End;
   this->VCO_Path.FCL_AddNode(VLS_Node1.attrib, VLS_Node1.cost);
   this->FCL_AdjustWeights(this->VCO_Path.VCP_End->level,
            this->VCO_Path.VCP_End->attrib, PLC_Type,
            VLP_Aux->cost - this->VCO_Path.VCP_End->cost);
   if(VGN_LogLevel >= 3)
   {
      this->VCO_Path.FCL_ShowElement(this->VCO_Path.VCP_End);
   }
   if(VLP_Aux->cost >= this->VCO_Path.VCP_End->cost)
```

```
   {
      VLP_Min = this->VCO_Path.VCP_End;
      VLN_StepsForMin = this->VCN_StepsForMin;
   }
   else
   {
      VLN_StepsForMin--;
      if(VLN_StepsForMin <= 0) break;
   }
}

if(VGN_LogLevel >= 3)
{
   fprintf(VGF_Log, "\nMinimum Found In the Path at:\n");
   printf("\nMinimum Found In the Path  at:\n");
   this->VCO_Path.FCL_ShowElement(VLP_Min);
}

if((this->VCN_MinCost == EMPTYCOST) ||
            (VLP_Min->cost < this->VCN_MinCost))
{
   this->VCN_StepEqual = this->VCN_StepsEqual;
   this->VCN_MinCost = VLP_Min->cost;
}
if(this->VCO_Result.FCL_AddNodeinOrder(VLP_Min->attrib,
        VLP_Min->cost, this->VCN_MaxResults))
          this->VCN_ResultChange = 1;
this->FCL_BeatSFFS();

if(PLC_Type == 'U')
{
   if(VLP_Min->next)
      this->VCO_Upper.FCL_AddNode(VLP_Min->next->attrib,
                        VLP_Min->next->cost);
   else
      this->VCO_Upper.FCL_AddNode(VLP_Min->attrib,
                        VLP_Min->cost);

   if(VLP_Min->prev)
      this->VCO_Lower.FCL_AddNode(VLP_Min->prev->attrib,
                              VLP_Min->prev->cost);
   else
      this->VCO_Lower.FCL_AddNode(VLP_Min->attrib,
                              VLP_Min->cost);
}
else
```

```
   {
      if(VLP_Min->next)
           this->VCO_Lower.FCL_AddNode(VLP_Min->next->attrib,
                               VLP_Min->next->cost);
      else
           this->VCO_Lower.FCL_AddNode(VLP_Min->attrib,
                               VLP_Min->cost);

      if(VLP_Min->prev)
            this->VCO_Upper.FCL_AddNode(VLP_Min->prev->attrib,
                               VLP_Min->prev->cost);
      else
            this->VCO_Upper.FCL_AddNode(VLP_Min->attrib,
                               VLP_Min->cost);
   }

   if(this->VCN_ExecExhaust)
   {
      return this->FCL_ExhaustProcess(VLP_Min, PLC_Type);
   }

   if(VGN_LogLevel >= 2)
   {
      fprintf(VGF_Log, "\nChain Process ENDS\n");
      fprintf(VGF_Log, "----------------------------\n\n");
      printf("\nChain Process ENDS\n");
      printf("-------------------------------------\n\n");
   }
   return 1;
}



/* ************************************************************************   */
/* PROCEDIMENTO de esgotamento do mínimo                                      */
/*            PLP_IniNode: elemento inicial                                   */
/*            PLC_Type: 'U' (sentido Inferior-Superior)                       */
/*                      'D' (sentido Superior-Inferior)                       */
/* ************************************************************************   */
int ProcessClassExhaustive::FCL_ExhaustProcess(ReticNode *PLP_IniNode, char PLC_Type)
{
   ReticNode *VLP_Aux, *VLP_Min, *VLP_AuxBefore;
   ReticNode VLS_Node1, VLS_Node2;
   ReticNodeList *VLP_List;
   long int VLN_Attrib;
   int VLL_HasLower;
```

```
int VLL_Stop;
char *VLC_Message;
long int VLN_StepsForMin;
time_t start, finish;

if(VGN_LogLevel >= 2)
{
   fprintf(VGF_Log, "\n\n-----------------------------------");
   fprintf(VGF_Log, "\nMinimum Exhausting Process BEGINS\n");
   printf("\n\n-----------------------------------------------");
   printf("\nMinimum Exhausting Process BEGINS\n");
}
time(&start);
this->VCO_Stack.FCL_FreeList();
this->VCO_Stack.FCL_AddNode(PLP_IniNode->attrib, PLP_IniNode->cost);

this->VCO_Path.FCL_FreeList();

VLP_AuxBefore = RETICNODE_NIL;
VLL_Stop = 0;
if(VGN_LogLevel >= 3)
{
   fprintf(VGF_Log, "\n\nUpper List: %lu elements \n", this->VCO_Upper.VCN_Nodes);
   printf("\n\nUpper List: %lu elements \n", this->VCO_Upper.VCN_Nodes);
   this->VCO_Upper.FCL_ShowList();
   fprintf(VGF_Log, "\n\nLower List: %lu elements \n", this->VCO_Lower.VCN_Nodes);
   printf("\n\nLower List: %lu elements \n", this->VCO_Lower.VCN_Nodes);
   this->VCO_Lower.FCL_ShowList();
}
while(this->VCO_Stack.VCN_Nodes)
{
   if(VGN_LogLevel >= 3)
   {
      fprintf(VGF_Log, "\n\nStack List: %lu elements \n",
                    this->VCO_Stack.VCN_Nodes);
      printf("\n\nStack List: %lu elements \n", this->VCO_Stack.VCN_Nodes);
      this->VCO_Stack.FCL_ShowList();
   }

   VLP_Aux = this->VCO_Stack.VCP_End;
   if(VGN_LogLevel >= 3)
   {
      fprintf(VGF_Log, "\nCentral Element:\n");
      printf("\nCentral Element:\n");
      this->VCO_Stack.FCL_ShowElement(VLP_Aux);
   }
```

```
if(true) /*!this->FCL_InRestrictions(VLP_Aux->attrib))*/
{
   VLL_HasLower = 0;
   if(!VLL_Stop)
   {
      if((!VLP_AuxBefore) || (VLP_AuxBefore->level > VLP_Aux->level))
      {
         VLL_HasLower = this->FCL_LookLowerUp(VLP_Aux, VLP_AuxBefore,
                                        &VLS_Node1);
         if(VLL_HasLower)   this->VCO_Stack.FCL_AddNode(VLS_Node1.attrib,
                                                  VLS_Node1.cost);
         VLP_AuxBefore = RETICNODE_NIL;
      }
      if(!VLL_HasLower)
      {
         if((!VLP_AuxBefore) || (VLP_AuxBefore->level < VLP_Aux->level))
         {
            VLL_HasLower = this->FCL_LookLowerDown(VLP_Aux, VLP_AuxBefore,
                                          &VLS_Node1);
            if(VLL_HasLower)
                 this->VCO_Stack.FCL_AddNode(VLS_Node1.attrib,
                                       VLS_Node1.cost);
            VLP_AuxBefore = RETICNODE_NIL;
         }
      }
   }

   if(!VLL_HasLower)
   {
      if(VGN_LogLevel >= 3)
      {
         fprintf(VGF_Log, "\nMinimum Completed: %s (%d), Cost: %lf\n",
                VLP_Aux->attrib, VLP_Aux->level, VLP_Aux->cost);
         printf("\nMinimum Completed: %s (%d), Cost: %lf\n",
                VLP_Aux->attrib, VLP_Aux->level, VLP_Aux->cost);
      }
      if(VLP_Aux->level < this->VCN_MinLevel)
                     this->VCN_MinLevel = VLP_Aux->level;
      if(VLP_Aux->level > this->VCN_MaxLevel)
                     this->VCN_MaxLevel = VLP_Aux->level;

      /* if cost of the element found is lower than Minumum,
                                inserts it in the result list */
      if(this->VCN_MinCost == EMPTYCOST)
                  this->VCN_MinCost = VLP_Aux->cost + 1;
```

```
if(!VLL_Stop)
{
   if(VLP_Aux->cost >= this->VCN_MinCost)
   {
      if(this->VCN_StepsEqual > 0)
      {
         this->VCN_StepEqual--;
         if(this->VCN_StepEqual <= 0)
         {
            if(VGN_LogLevel >= 3)
            {
               fprintf(VGF_Log, "\n\nProcess Stopped - staying %lu
                  steps with the same result\n", this->VCN_StepsEqual);
               printf("\n\nProcess Stopped - staying %lu steps with
                       the same result\n", this->VCN_StepsEqual);
            }
            VLL_Stop = 1;
         }
         else
         {
            if(VGN_LogLevel >= 3)
            {
               fprintf(VGF_Log, "\n\nGreater than minimum - steps to
                                 stop: %lu\n", this->VCN_StepEqual);
               printf("\n\nGreater than minimum - steps to stop:
                             %lu\n", this->VCN_StepEqual);
            }
         }
      }
   }
}
if(VLP_Aux->cost <= this->VCN_MinCost)
{
   if(VLP_Aux->cost < this->VCN_MinCost)
   {
      this->VCN_StepEqual = this->VCN_StepsEqual;
      this->VCN_MinCost = VLP_Aux->cost;
   }
   if(this->VCO_Result.FCL_AddNodeinOrder(VLP_Aux->attrib,
          VLP_Aux->cost,this->VCN_MaxResults))
                  this->VCN_ResultChange = 1;
   this->FCL_BeatSFFS();

   if(VGN_LogLevel >= 3)
   {
      this->VCO_Result.FCL_ShowElement(VLP_Aux);
```

```
            }
         }

         this->VCO_Upper.FCL_AddNode(VLP_Aux->attrib, VLP_Aux->cost);
         this->VCO_Lower.FCL_AddNode(VLP_Aux->attrib, VLP_Aux->cost);
         memcpy(&VLS_Node2, VLP_Aux, sizeof(ReticNode));
         VLP_AuxBefore = &VLS_Node2;
         this->VCO_Stack.FCL_RemoveNode(VLP_Aux);
      }
   }
   else
   {
      this->VCO_Upper.FCL_AddNode(VLP_Aux->attrib, VLP_Aux->cost);
      this->VCO_Lower.FCL_AddNode(VLP_Aux->attrib, VLP_Aux->cost);
      memcpy(&VLS_Node2, VLP_Aux, sizeof(ReticNode));
      VLP_AuxBefore = &VLS_Node2;
      this->VCO_Stack.FCL_RemoveNode(VLP_Aux);
   }

   if(VGN_LogLevel >= 3)
   {
      fprintf(VGF_Log, "\n\nUpper List: %lu elements \n",
                              this->VCO_Upper.VCN_Nodes);
      printf("\n\nUpper List: %lu elements \n",
                              this->VCO_Upper.VCN_Nodes);
      this->VCO_Upper.FCL_ShowList();
      fprintf(VGF_Log, "\n\nLower List: %lu elements \n",
                              this->VCO_Lower.VCN_Nodes);
      printf("\n\nLower List: %lu elements \n",
                              this->VCO_Lower.VCN_Nodes);
      this->VCO_Lower.FCL_ShowList();
   }
}
time(&finish);
if(VGN_LogLevel >= 2)
{
   fprintf(VGF_Log, "\n\nNodes Allocated: %lu;
            Nodes calculated: %lu,
            Time Processing: %.0f seconds\n",
            this->FCN_NodesAllocated(), VGO_Cost->VCN_Calcs,
            difftime( finish, start ));
   printf("\n\nNodes Allocated: %lu; Nodes calculated: %lu,
            Time Processing: %.0f seconds\n",
            this->FCN_NodesAllocated(), VGO_Cost->VCN_Calcs,
            difftime( finish, start ));
   fprintf(VGF_Log, "Minimum Exhausting Process ENDS\n");
```

```
      fprintf(VGF_Log, "-----------------------------------\n\n");
      printf("Minimum Exhausting Process ENDS\n");
      printf("-------------------------------------------\n\n");
   }
   return 1;
}
```

Apêndice B

Pipeline de Algoritmos

Veremos agora as tabelas componentes do banco de dados MySql^{TM} usado no *pipeline*:

- **edliugenes** – tabela principal do banco de dados, contêm os seguintes campos:

  1. *genesymbol* (varchar(30)) – identificador do gene, nome mais comum pelo qual ele é encontrado na literatura (por exemplo: *ADNP* para o gene *Activity-dependent neuroprotector homeobox*);
  2. *kegg_symb* (varchar(50)) – agrupamento dos identificadores das vias do KEGG (por exemplo: [CC][AP]), onde cada identificador significa:

  ACTIN - regulação de actina (*regulation of actin cytoskeleton*);
  AP - apopotose (*apoptosis*);
  CC - ciclo celular (*cell cycle*);
  COMM - comunicação celular (*cell communication*);
  CYT - interação do receptor de citoquina-citoquina (*cytokine-cytokine receptor interaction*);
  JAK - via de sinalização *Jak-STAT* (*Jak-STAT signaling pathway*);
  KILL - mediador natutal de citotoxidade de morte celular (*natural killer cell mediated cytotoxicity*);
  MAPK - via de sinalização de MAPK (*MAPK signaling pathway*);
  TCELL - via de sinalização de receptor de célula T (*T cell receptor signaling pathway*);
  TGF - via de sinalização de TGF-beta (*TGF-beta signaling pathway*);
  TOLL - via de sinalização de receptor tipo *TOLL* (*Toll-like receptor signaling pathway*);
  WNT - via de sinalização de *Wnt* (*Wnt signaling pathway*);

  3. *kegg_descr* (varchar(200)) – agrupamento das descrições dos identificadores das vias do KEGG (por exemplo: *Cell Communication,TGF-beta signaling pathway*);
  4. *marker* (varchar(50)) – marcações para identificar os em que grupo os genes estão (por exemplo: [386][139]):

  386 - pertence a lista de 386 genes responsivos ao estrógeno obtidos em [26];
  139 - pertence a lista de 139 genes regulados pelo estrógeno obtidos em [26];
  89 - pertence a lista de 89 genes diretamente regulados por estrógeno obtidos em [26];

- **edliuexpr** – tabela que armazena os sinais de expressão dos genes, contêm os seguintes campos:

  1. *genesymbol* (varchar(30)) – identificador do gene;
  2. *expression* (text) – contém a o sinal de expressão do gene obtido diretamente do experimento (por exemplo: *-0.19084621,0.18350670,0.08434014*);
  3. *expression_n* (text) – contém a o sinal de expressão normalizado do gene (por exemplo: *-0.80893650,1.13035355,0.61663314*);
  4. *expression_nq* (text) – contém a o sinal de expressão normalizado e discretizado do gene (por exemplo: *0,1,0*);

- **edliugengo** – tabela que relaciona um gene com suas funções biológicas (ontologia), contêm os seguintes campos:

  1. *genesymbol* (varchar(30)) – identificador do gene;
  2. *genegroup* (varchar(100)) – função biológica do gene (por exemplo: *cell cycle*);
  3. *genetype* (varchar(200)) – tipo de função biológica (*biological process*, *molecular function*, *cellular component*);
  4. *gonumber* (varchar(10)) – número identificador da ontologia em "*Gene Ontology Database*" (<`http://www.geneontology.org`>);
  5. *evidence* (varchar(20)) – código de evidência da ontologia (por exemplo: *IC, ND, NAS*) em "*Gene Ontology Database*" (<`http://www.geneontology.org`>);

- **edliutest** – armazena os testes (iterações) do *pipeline*, contêm os seguintes campos::

  1. *testid* (varchar(30)) – identificador do teste (por exemplo: *TESTE1_STEP1*);
  2. *testdescr* (varchar(200)) – descrição do teste (por exempo: *Test 1 (386+kegg+go), Step 1, Skip 1*);
  3. *predictors* (text) – texto contendo o símbolos do conjunto de genes-sementes (por exemplo: *ADNP;AREG;CCNG2;CD164;CRABP2;DKC1*);
  4. *sementeselect* (text) – texto como comaando *SQL* para selecionar o conjunto de genes-sementes;

- **edliutest_res** – armazena os resultados dos testes para cada gene, contêm os seguintes campos:

  1. *genesymbol* (varchar(30)) – identificador do gene;
  2. *testid* (varchar(30)) – identificador do teste;
  3. *entropy* (double) – valor para o custo (entropia condicional média) relativo ao melhor subconjunto preditor;
  4. *predictors* (varchar(200)) – armazena os subconjuntos preditores com suas entropias relativas (por exemplo: *(0.00000000):SCN1B;GREB1, (0.06666600):NIF3L1;GREB1*, indica o subconjunto preditor (SCN1B,GREB1), com entropia 0; e o subconjunto (NIF3L1,GREB1), com entropia 0,06666600);
  5. *reshtml* (text) – código *html* para a página com as taabelas de predição.

A Figura B.1 é uma representação esquemática do banco de dados com suas tabelas e relacionamentos.

A Tabela B.1 contêm os 33 genes-sementes do resultado do *pipeline* (Seção 4.2.5). A Tabela B.2 contêm os 38 genes melhores preditos com seus conjuntos de preditores no primeiro passo, a Tabela B.3 contêm os 37 do segundo passo, e a Tabela B.4 contêm os 38 do terceiro passo.

ADNP		regulation of transcription, DNA-dependent
CCNG2		cell cycle checkpoint
CD164		negative regulation of cell proliferation
CRABP2		regulation of transcription, DNA-dependent
DKC1		cell proliferation
DNMT3L		DNA methylation
DSIPI		regulation of transcription, DNA-dependent
EPAS1		regulation of transcription, DNA-dependent
ERBB4		cell proliferation
ESPL1	[CC]	regulation of cell cycle
FRAT2	[WNT]	cell proliferation
HOXC13		regulation of transcription, DNA-dependent
IGFBP4		DNA metabolism
IRF1		cell cycle
KLF3		regulation of transcription, DNA-dependent
LASS2		regulation of transcription, DNA-dependent
MPHOSPH1		cell cycle arrest
MPHOSPH6		M phase of mitotic cell cycle
MYBL1		regulation of transcription, DNA-dependent
NBL1		cell cycle
NFIA		DNA replication
NRIP1		regulation of transcription, DNA-dependent
ORC1L	[CC]	DNA replication
PAFAH1B1		cell cycle
PGR		regulation of transcription, DNA-dependent
POU1F1		negative regulation of cell proliferation
PRDM14		regulation of transcription, DNA-dependent
REL		regulation of transcription, DNA-dependent
SIAH2		cell cycle
TACSTD2		cell proliferation
TBX21		regulation of transcription, DNA-dependent
TBX6		regulation of transcription, DNA-dependent
TOB1		negative regulation of cell proliferation

Tabela B.1: Tabela com os 33 genes-sementes iniciais do resultado do *pipeline*

Gene	Preditores	Entropia	KEGG	Função Biológica
FOXC1	(0.00000000):POU1F1;CD164 , (0.07142800):REL;POU1F1;CD164 , (0.07142800):NRIP1;CD164 , (0.07142800):POU1F1;LASS2;CD164 , (0.07142800):LASS2;CD164	0.000		regulation of transcription, DNA-dependent
BHLHB2	(0.06666600):PAFAH1B1;MPHOSPH6 , (0.13333300):DKC1;ADNP , (0.13333300):DKC1;MPHOSPH6 , (0.20000000):PRDM14;ADNP , (0.20000000):MYBL1;MPHOSPH6	0.067		regulation of transcription, DNA-dependent
CCNC	(0.06666600):PRDM14;CD164 , (0.26666600):PRDM14;TOB1;CD164 , (0.28412300):ERBB4;PGR;CD164 , (0.31483800):TOB1;AREG , (0.33333300):NFIA;TOB1;AREG	0.067		regulation of cell cycle
CDKN1B	(0.06666600):PRDM14;TBX6 , (0.06666600):REL;TBX6 , (0.13333300):TBX6;TBX21;FRAT2[WNT] , (0.13333300):TBX6;FRAT2[WNT] , (0.13333300):TBX6;IGFBP4	0.067	[CC]	cell cycle arrest
ELF5	(0.06666600):PRDM14;HOXC13 , (0.15079000):PRDM14;CD164 , (0.21745700):TACSTD2;DKC1 , (0.26666600):DNMT3L;PRDM14 , (0.26666600):DNMT3L;PRDM14;NBL1	0.067		DNA binding
F2	(0.06666600):SIAH2;DKC1 , (0.06666600):SIAH2;AREG , (0.13333300):SIAH2;NBL1 , (0.13333300):SIAH2;NBL1;DKC1 , (0.15079000):DNMT3L;REL;NBL1	0.067	[ACTIN]	regulation of cell cycle
FEN1	(0.06666600):DNMT3L;PRDM14 , (0.06666600):DNMT3L;PRDM14;NBL1 , (0.20000000):DNMT3L;PRDM14;DKC1 , (0.23175200):DNMT3L;NBL1;DKC1 , (0.23175200):DNMT3L;DKC1	0.067		DNA replication
FLJ13265	(0.06666600):PAFAH1B1;AREG , (0.13333300):DSIPI;PAFAH1B1;AREG , (0.13333300):PAFAH1B1;CRABP2 , (0.18254200):DSIPI;PAFAH1B1 , (0.20000000):NFIA;PAFAH1B1;AREG	0.067		regulation of cell cycle
MAPK13	(0.06666600):TBX6;KLF3 , (0.13333300):DNMT3L;TBX6 , (0.16824700):TBX6 , (0.18254200):TBX6;AREG , (0.19558100):ESPL1[CC];AREG	0.067	[MAPK][TOLL]	cell cycle
NEUROD1	(0.06666600):PRDM14;CCNG2 , (0.11587600):PRDM14 , (0.11587600):PRDM14;AREG , (0.13333300):PRDM14;TACSTD2 , (0.13333300):PRDM14;DSIPI	0.067		DNA binding
PER2	(0.06666600):NFIA;POU1F1 , (0.06666600):NFIA;TOB1 , (0.06666600):NFIA;EPAS1 , (0.11587600):NFIA , (0.11587600):NFIA;AREG	0.067		regulation of transcription, DNA-dependent
PFDN1	(0.06666600):DNMT3L;CD164 , (0.16824700):KLF3;CRABP2 , (0.20000000):DNMT3L;REL;CRABP2 , (0.20000000):DNMT3L;REL;CD164 , (0.20000000):DNMT3L;KLF3;CRABP2	0.067		cell cycle

Gene	Preditores	Entropia	KEGG	Função Biológica
PTEN	(0.06666600):PAFAH1B1;MPHOSPH6 , (0.21745700):DSIPI;CD164 , (0.21745700):PAFAH1B1;PGR , (0.26666600):ERBB4;PAFAH1B1;PGR , (0.29841800):ERBB4;PGR	0.067		cell cycle
RFX2	(0.06666600):SIAH2;HOXC13 , (0.18254200):REL;HOXC13 , (0.20000000):REL;SIAH2;HOXC13 , (0.20000000):SIAH2;PAFAH1B1;HOXC13 , (0.23491400):PRDM14;CD164	0.067		DNA binding
SALL2	(0.06666600):DKC1;CRABP2 , (0.13333300):DKC1;CCNG2 , (0.13333300):DKC1;ADNP , (0.13333300):DKC1;MPHOSPH6 , (0.13333300):CCNG2;AREG	0.067		regulation of transcription, DNA-dependent
SMARCA5	(0.06666600):SIAH2;TOB1 , (0.13333300):SIAH2;TOB1;IGFBP4 , (0.20000000):REL;SIAH2;TOB1 , (0.21745700):SIAH2;ADNP , (0.23298900):SIAH2;ESPL1[CC]	0.067		DNA binding
TGFB3	(0.06666600):TBX21;FRAT2[WNT] , (0.13333300):TBX6;TBX21;FRAT2[WNT] , (0.13333300):PAFAH1B1;PGR , (0.20000000):TBX21;FRAT2[WNT];CCNG2 , (0.21745700):PAFAH1B1;CRABP2	0.067	[MAPK][CC][CYT][TGF]	cell proliferation
YAF2	(0.06666600):DNMT3L;CD164 , (0.20000000):DNMT3L;REL;CRABP2 , (0.20000000):DNMT3L;REL;CD164 , (0.23491400):DNMT3L;REL , (0.23491400):DNMT3L;REL;NBL1	0.067		regulation of transcription, DNA-dependent
ZFP95	(0.06666600):MYBL1;ERBB4 , (0.13333300):TACSTD2;DSIPI , (0.18254200):TBX21;DSIPI , (0.20000000):PRDM14;TACSTD2;AREG , (0.20000000):TBX21;TACSTD2;DSIPI	0.067		regulation of transcription, DNA-dependent
RARRES1	(0.07142800):ESPL1[CC];MPHOSPH6 , (0.20316200):NFIA;TOB1 , (0.21428500):NFIA;ESPL1[CC];MPHOSPH6 , (0.21428500):MPHOSPH1;ESPL1[CC];MPHOSPH6 , (0.21745700):NFIA;TOB1;AREG	0.071		negative regulation of cell proliferation
TCF19	(0.07142800):ERBB4;PAFAH1B1 , (0.12415200):ERBB4 , (0.14285700):MYBL1;ERBB4 , (0.16156100):KLF3;ERBB4 , (0.16156100):ERBB4;CD164	0.071		cell proliferation
TCF8	(0.07142800):TBX6;KLF3 , (0.14285700):DNMT3L;TBX6 , (0.14285700):TBX6;DSIPI , (0.15384600):ESPL1[CC];DSIPI , (0.18026500):TBX6	0.071		cell proliferation
XPC	(0.07142800):POU1F1;LASS2;CD164 , (0.07142800):LASS2;CD164 , (0.21428500):IRF1;LASS2;CD164 , (0.21428500):TBX21;LASS2;CD164 , (0.21428500):TACSTD2;CD164	0.071		damaged DNA binding
ZNF20	(0.08333300):KLF3;ERBB4 , (0.14484500):ERBB4 , (0.16666600):NFIA;KLF3 , (0.16666600):NFIA;KLF3;ERBB4 , (0.16666600):NFIA;ERBB4	0.083		DNA binding
FOXD2	(0.09013200):NFIA;ERBB4 , (0.09013200):KLF3;ERBB4 , (0.09013200):ERBB4 , (0.14285700):NFIA;IRF1;ERBB4 , (0.14285700):NFIA;ERBB4;TOB1	0.090		regulation of transcription, DNA-dependent

Gene	Preditores	Entropia	KEGG	Função Biológica
PRKG2	(0.09090900):DNMT3L;NBL1;DKC1 , (0.09090900):DNMT3L;DKC1 , (0.09090900):DNMT3L;DKC1;IGFBP4 , (0.20562300):KLF3;LASS2 , (0.20562300):KLF3;LASS2;ERBB4	0.091		regulation of cell cycle
ALOX15B	(0.11587600):DNMT3L;NBL1;DKC1 , (0.11587600):DNMT3L;DKC1 , (0.11587600):NBL1;DKC1 , (0.20000000):NBL1;DKC1;PAFAH1B1 , (0.21428500):ORC1L[CC];NBL1;DKC1	0.116		negative regulation of cell cycle
DLX5	(0.11587600):KLF3;CRABP2 , (0.13333300):CRABP2;CD164 , (0.15079000):DNMT3L;CD164 , (0.15079000):SIAH2;CD164 , (0.20000000):DNMT3L;REL;CD164	0.116		regulation of transcription, DNA-dependent
DNASE1L2	(0.11587600):IRF1;LASS2 , (0.18254200):IRF1;TBX21 , (0.18254200):IRF1;TBX21;LASS2 , (0.20000000):IRF1;POU1F1;LASS2 , (0.22061900):KLF3;CRABP2	0.116		DNA binding
FGF18	(0.11587600):DNMT3L;NBL1;DKC1 , (0.11587600):DNMT3L;DKC1 , (0.20000000):DNMT3L;PRDM14;DKC1 , (0.20420000):NBL1;DKC1 , (0.21745700):FRAT2[WNT];PAFAH1B1	0.116	[MAPK][ACTIN]	positive regulation of cell proliferation
GAS7	(0.11587600):LASS2;PAFAH1B1 , (0.20000000):TBX21;PAFAH1B1;AREG , (0.20316200):KLF3;LASS2 , (0.22061900):KLF3;CRABP2 , (0.23491400):SIAH2;DKC1	0.116		cell cycle arrest
MYF5	(0.11587600):LASS2;PAFAH1B1 , (0.13333300):FRAT2[WNT];PAFAH1B1 , (0.13333300):DKC1;PAFAH1B1 , (0.15079000):DSIPI;PAFAH1B1 , (0.18254200):PAFAH1B1;MPHOSPH6	0.116		DNA binding
POU4F2	(0.11587600):LASS2;PAFAH1B1 , (0.18254200):POU1F1;LASS2;HOXC13 , (0.18254200):KLF3;LASS2 , (0.18254200):LASS2;HOXC13 , (0.24920900):IRF1;PAFAH1B1	0.116		regulation of transcription, DNA-dependent
SKP2	(0.11587600):POU1F1 , (0.11587600):POU1F1;LASS2 , (0.11587600):POU1F1;CD164 , (0.13333300):POU1F1;CCNG2 , (0.15079000):NFIA;POU1F1	0.116	[CC]	G1/S transition of mitotic cell cycle
PITX3	(0.12415200):ESPL1[CC] , (0.13333300):IRF1;SIAH2 , (0.14285700):DNMT3L;ESPL1[CC] , (0.15079000):SIAH2;DKC1 , (0.16156100):ORC1L[CC];ESPL1[CC]	0.124		regulation of transcription, DNA-dependent
SPOCK	(0.12415200):ESPL1[CC] , (0.13333300):DNMT3L;TBX6 , (0.14285700):DNMT3L;ESPL1[CC] , (0.15079000):TBX6;KLF3 , (0.15079000):DSIPI;PAFAH1B1	0.124		cell proliferation
TAF4B	(0.12415200):POU1F1;CD164 , (0.16156100):REL;POU1F1;CD164 , (0.25169400):POU1F1;EPAS1 , (0.26701000):IRF1;ADNP , (0.26701000):POU1F1;LASS2;CD164	0.124		regulation of transcription, DNA-dependent
HOXD10	(0.12618500):ORC1L[CC];ESPL1[CC] , (0.18181800):FRAT2[WNT];IGFBP4 , (0.20000000):ORC1L[CC];POU1F1;ESPL1[CC] , (0.22942900):TBX21;FRAT2[WNT] , (0.22942900):FRAT2[WNT]	0.126		regulation of transcription, DNA-dependent

Tabela B.2: Tabela com os 38 genes melhores preditos no primeiro passo

Gene	Preditores	Entropia	KEGG	Função Biológica
ARNT	(0.00000000):PRKG2 , (0.09090900):FEN1;PRKG2 , (0.09090900):ALOX15B;PRKG2 , (0.09090900):SKP2[CC];PRKG2 , (0.09090900):SPOCK;PRKG2	0.000		regulation of transcription, DNA-dependent
DFFB	(0.00000000):ZNF20 , (0.00000000):ZNF20;GAS7 , (0.08333300):TGFB3[MAPK][CC][CYT][TGF];ZNF20 , (0.08333300):TAF4B;ZNF20 , (0.08333300):ZNF20;FLJ13265	0.000	[AP]	DNA fragmentation during apoptosis
HEYL	(0.00000000):FOXC1;ZNF20;HOXD10 , (0.12500000):DLX5;TCF8;PRKG2 , (0.12500000):POU4F2;TCF8;PRKG2 , (0.12500000):TCF8;SKP2[CC];PRKG2 , (0.12500000):TCF8;SPOCK;PRKG2	0.000		regulation of transcription, DNA-dependent
IL9R	(0.00000000):XPC;PFDN1 , (0.09090900):PRKG2;FOXD2 , (0.14285700):XPC;NEUROD1;PFDN1 , (0.15801200):PRKG2 , (0.18026500):XPC	0.000	[CYT][JAK]	cell proliferation
NCOA5	(0.00000000):POU4F2;HOXD10 , (0.10000000):DLX5;POU4F2;HOXD10 , (0.10000000):DLX5;HOXD10 , (0.10000000):POU4F2;HOXD10;SKP2[CC] , (0.20000000):TGFB3[MAPK][CC][CYT][TGF];DLX5;HOXD10	0.000		regulation of transcription, DNA-dependent
PPP1R15A	(0.00000000):POU4F2;HOXD10 , (0.08333300):TGFB3[MAPK][CC][CYT][TGF];POU4F2;HOXD10 , (0.08333300):DLX5;POU4F2;HOXD10 , (0.08333300):DLX5;HOXD10 , (0.08333300):DLX5;HOXD10;PFDN1	0.000		cell cycle arrest
BUB1B	(0.06666600):POU4F2;PFDN1 , (0.08333300):DLX5;HOXD10;PFDN1 , (0.08333300):POU4F2;HOXD10;PFDN1 , (0.13333300):DLX5;POU4F2;PFDN1 , (0.13333300):POU4F2;YAF2;PFDN1	0.067	[CC]	cell cycle
IRF5	(0.06666600):TGFB3[MAPK][CC][CYT][TGF];ALOX15B , (0.06666600):TGFB3[MAPK][CC][CYT][TGF];ALOX15B;SALL2 , (0.06666600):ALOX15B;SALL2 , (0.06666600):ALOX15B;SALL2;MYF5 , (0.06666600):ALOX15B;MYF5	0.067		regulation of transcription, DNA-dependent
TNFAIP3	(0.06666600):FEN1;FLJ13265;SALL2 , (0.14484500):TGFB3[MAPK][CC][CYT][TGF];HOXD10 , (0.16666600):TGFB3[MAPK][CC][CYT][TGF];HOXD10;CDKN1B[CC] , (0.16666600):BHLHB2;POU4F2;HOXD10 , (0.16666600):BHLHB	0.067		DNA binding
ZNF262	(0.06666600):RFX2;CCNC , (0.20000000):RFX2;SKP2[CC];CCNC , (0.26666600):RFX2;ELF5;CCNC , (0.26666600):RFX2;GAS7;CCNC , (0.26666600):RFX2;YAF2;CCNC	0.067		DNA binding
ZNF75	(0.06666600):FEN1;RFX2 , (0.12500000):ZNF20;HOXD10;PRKG2 , (0.13333300):FEN1;RFX2;FLJ13265 , (0.20562300):PRKG2;PER2 , (0.21745700):FEN1;PER2	0.067		DNA binding
APPL	(0.07142800):YAF2;CCNC , (0.14285700):SKP2[CC];YAF2;CCNC , (0.14285700):RARRES1;YAF2 , (0.18026500):YAF2 , (0.19558100):RFX2;YAF2	0.071		cell cycle

Gene	Preditores	Entropia	KEGG	Função Biológica
ZNF177	(0.07142800):TAF4B;MAPK13[MAPK][TOLL] , (0.07692300):TAF4B;TCF8 , (0.07692300):TAF4B;TCF8;MAPK13[MAPK][TOLL] , (0.09090900):PRKG2;PER2 , (0.16666600):FEN1;HOXD10;TCF8	0.071		DNA binding
AREG	(0.07692300):FOXC1;ELF5 , (0.18181800):FEN1;PRKG2;NEUROD1 , (0.18181800):PRKG2;NEUROD1;PER2 , (0.21745700):ELF5;SKP2[CC];PFDN1 , (0.22618500):FOXC1;HOXD10;NEUROD1	0.077		cell proliferation
BAZ2B	(0.08333300):HOXD10;NEUROD1;PFDN1 , (0.14484500):TGFB3[MAPK][CC][CYT][TGF];HOXD10 , (0.16666600):TGFB3[MAPK][CC][CYT][TGF];HOXD10;CDKN1B[CC] , (0.18848800):TGFB3[MAPK][CC][CYT][TGF];POU4F2;HOXD10 —	0.083		DNA binding
CDK10	(0.08333300):HOXD10;NEUROD1;PFDN1 , (0.16666600):ZFP95;HOXD10;NEUROD1 , (0.18181800):FEN1;PRKG2;NEUROD1 , (0.18181800):HOXD10;TCF19;NEUROD1 , (0.18978500):HOXD10;NEUROD1	0.083		negative regulation of cell proliferation
CDK8	(0.08333300):TGFB3[MAPK][CC][CYT][TGF];POU4F2;HOXD10 , (0.14484500):TGFB3[MAPK][CC][CYT][TGF];HOXD10 , (0.16666600):BHLHB2;POU4F2;HOXD10 , (0.16666600):POU4F2;HOXD10;SALL2 , (0.18254200):TGFB3[MAP	0.083		regulation of cell cycle
ELF4	(0.08333300):FEN1;HOXD10 , (0.16666600):FEN1;HOXD10;TCF8 , (0.16666600):FEN1;HOXD10;MAPK13[MAPK][TOLL] , (0.17398900):TAF4B;TCF8 , (0.17398900):TAF4B;TCF8;MAPK13[MAPK][TOLL]	0.083		regulation of transcription, DNA-dependent
HES6	(0.08333300):FEN1;HOXD10 , (0.10000000):FOXC1;HOXD10 , (0.10000000):FOXC1;HOXD10;NEUROD1 , (0.10515400):ZNF20 , (0.10515400):ZNF20;GAS7	0.083		regulation of transcription, DNA-dependent
KIAA1018	(0.08333300):FEN1;HOXD10 , (0.16666600):FEN1;ZNF20 , (0.16666600):FEN1;HOXD10;TCF8 , (0.16666600):FEN1;HOXD10;MAPK13[MAPK][TOLL] , (0.21745700):FEN1;RFX2;FLJ13265	0.083		DNA binding
MTMR7	(0.08333300):XPC;RARRES1 , (0.08333300):XPC;RARRES1;FOXD2 , (0.16666600):XPC;RARRES1;NEUROD1 , (0.20000000):XPC;FOXC1;RARRES1 , (0.20000000):DNASE1L2;FOXC1;RARRES1	0.083		cell cycle
MYNN	(0.08333300):ZNF20;FLJ13265 , (0.08333300):ZNF20;FLJ13265;GAS7 , (0.08333300):ZNF20;FLJ13265;CCNC , (0.08333300):ZNF20;CCNC , (0.15079000):FEN1;ELF5;FGF18[MAPK][ACTIN]	0.083		regulation of transcription, DNA-dependent
NOTCH4	(0.08333300):DLX5;POU4F2;HOXD10 , (0.08333300):DLX5;HOXD10 , (0.08333300):DLX5;HOXD10;PFDN1 , (0.08333300):POU4F2;HOXD10;PFDN1 , (0.16666600):DLX5;ZFP95;HOXD10	0.083		positive regulation of transcription, DNA-dependent
NPAS2	(0.08333300):HOXD10;NEUROD1;PFDN1 , (0.18848800):TGFB3[MAPK][CC][CYT][TGF];ZNF20 , (0.20562300):TGFB3[MAPK][CC][CYT][TGF];ZNF20;TCF19 , (0.25000000):TGFB3[MAPK][CC][CYT][TGF];ZNF20;ZFP95 , (0.2500	0.083		regulation of transcription, DNA-dependent

Gene	Preditores	Entropia	KEGG	Função Biológica
PRM1	(0.08333300):TGFB3[MAPK][CC][CYT][TGF];POU4F2;HOXD10 , (0.14484500):TGFB3[MAPK][CC][CYT][TGF];HOXD10 , (0.18848800):DLX5;HOXD10;PFDN1 , (0.18848800):POU4F2;HOXD10;PFDN1 , (0.18978500):POU4F2;HOXD1	0.083		DNA binding
STAT1	(0.08333300):TGFB3[MAPK][CC][CYT][TGF];POU4F2;HOXD10 , (0.14484500):TGFB3[MAPK][CC][CYT][TGF];HOXD10 , (0.16666600):BHLHB2;POU4F2;HOXD10 , (0.16666600):POU4F2;HOXD10;SALL2 , (0.18848800):DLX5;HOXD	0.083	[JAK][TOLL]	regulation of cell cycle
SYCP2	(0.08333300):HOXD10;NEUROD1;PFDN1 , (0.24920900):TGFB3[MAPK][CC][CYT][TGF];FLJ13265;MYF5 , (0.25000000):TGFB3[MAPK][CC][CYT][TGF];HOXD10 , (0.25000000):TGFB3[MAPK][CC][CYT][TGF];HOXD10;NEUROD1 , (	0.083		DNA binding
TACC1	(0.08333300):DLX5;HOXD10;PFDN1 , (0.08333300):POU4F2;HOXD10;PFDN1 , (0.16666600):FEN1;HOXD10;TCF8 , (0.16666600):FEN1;HOXD10;MAPK13[MAPK][TOLL] , (0.21745700):ELF5;ALOX15B	0.083		cell cycle
TAF12	(0.08333300):TGFB3[MAPK][CC][CYT][TGF];POU4F2;HOXD10 , (0.14484500):TGFB3[MAPK][CC][CYT][TGF];HOXD10 , (0.18181800):ALOX15B;SKP2[CC];PRKG2 , (0.18848800):DLX5;HOXD10;PFDN1 , (0.18848800):POU4F2;HO	0.083		DNA binding
TIMP1	(0.08333300):DLX5;POU4F2;HOXD10 , (0.08333300):DLX5;HOXD10 , (0.08333300):DLX5;HOXD10;PFDN1 , (0.08333300):POU4F2;HOXD10;PFDN1 , (0.08333300):HOXD10;TCF8	0.083		positive regulation of cell proliferation
TXNDC	(0.08333300):HOXD10;NEUROD1;PFDN1 , (0.25000000):HOXD10;SKP2[CC];NEUROD1 , (0.25525000):HOXD10;NEUROD1 , (0.26589700):TCF8;GAS7 , (0.26589700):TCF8;GAS7;MAPK13[MAPK][TOLL]	0.083		DNA replication
VEGFC	(0.08333300):DLX5;POU4F2;HOXD10 , (0.08333300):DLX5;HOXD10 , (0.08333300):DLX5;HOXD10;PFDN1 , (0.08333300):POU4F2;HOXD10;PFDN1 , (0.12500000):FOXC1;ZNF20;HOXD10	0.083	[CYT]	cell proliferation
ZNF261	(0.08333300):DLX5;POU4F2;HOXD10 , (0.08333300):DLX5;HOXD10 , (0.08333300):DLX5;HOXD10;PFDN1 , (0.08333300):POU4F2;HOXD10;PFDN1 , (0.16156100):XPC;RARRES1	0.083		DNA binding
BCAR3	(0.08412300):FEN1;ELF5 , (0.08412300):DLX5;ELF5 , (0.15079000):FEN1;BHLHB2 , (0.15079000):FEN1;ELF5;FGF18[MAPK][ACTIN] , (0.15079000):DLX5;ELF5;SKP2[CC]	0.084		regulation of cell cycle
JRK	(0.08412300):DLX5;YAF2 , (0.15079000):DLX5;YAF2;PFDN1 , (0.20000000):DLX5;FLJ13265;YAF2 , (0.21428500):DLX5;YAF2;FOXD2 , (0.21745700):DLX5;POU4F2;YAF2	0.084		DNA binding
POU2AF1	(0.08412300):GAS7;MAPK13[MAPK][TOLL] , (0.09013200):TCF8;GAS7 , (0.09013200):TCF8;GAS7;MAPK13[MAPK][TOLL] , (0.13333300):GAS7;SPOCK;MAPK13[MAPK][TOLL] , (0.14285700):TCF8;GAS7;SPOCK	0.084		regulation of transcription, DNA-dependent
RAD1	(0.08412300):FEN1;ELF5 , (0.15079000):FEN1;BHLHB2 , (0.15079000):FEN1;ELF5;FGF18[MAPK][ACTIN] , (0.18254200):FEN1;MYF5 , (0.18848800):FEN1;HOXD10	0.084		DNA repair

Tabela B.3: Tabela com os 37 genes melhores preditos no segundo passo

Gene	Preditores	Entropia	KEGG	Função Biológica
CORT	(0.00000000):HEYL;NPAS2 , (0.09090900):RAD1;HEYL;NPAS2 , (0.09090900):BCAR3;HEYL;NPAS2 , (0.18181800):PPP1R15A;HEYL;NPAS2 , (0.18181800):CDK8;HEYL;NPAS2	0.000		DNA binding
FANCE	(0.00000000):HEYL , (0.00000000):HEYL;NPAS2 , (0.09090900):RAD1;BCAR3;HEYL , (0.09090900):RAD1;HEYL , (0.09090900):RAD1;HEYL;NPAS2	0.000		DNA repair
GABPA	(0.00000000):HEYL;NPAS2 , (0.09090900):RAD1;HEYL;NPAS2 , (0.09090900):BCAR3;HEYL;NPAS2 , (0.15079000):POU2AF1;ZNF261;NPAS2 , (0.15079000):BUB1B[CC];NPAS2	0.000		regulation of transcription, DNA-dependent
HAND2	(0.00000000):HEYL;NPAS2 , (0.09090900):RAD1;HEYL;NPAS2 , (0.09090900):BCAR3;HEYL;NPAS2 , (0.14285700):PPP1R15A;APPL;NPAS2 , (0.14285700):APPL;ZNF261;NPAS2	0.000		DNA binding
MGA	(0.00000000):HEYL , (0.00000000):HEYL;NPAS2 , (0.09090900):RAD1;BCAR3;HEYL , (0.09090900):RAD1;HEYL , (0.09090900):RAD1;HEYL;NPAS2	0.000		regulation of transcription, DNA-dependent
PCBP2	(0.00000000):ZNF75;TNFAIP3 , (0.11111100):TAF12;MTMR7;TNFAIP3 , (0.11111100):ZNF75;MTMR7;TNFAIP3 , (0.11111100):MTMR7;PPP1R15A;TNFAIP3 , (0.11111100):MTMR7;CDK8;TNFAIP3	0.000		DNA binding
ZNF337	(0.00000000):HEYL , (0.00000000):HEYL;NPAS2 , (0.09090900):RAD1;BCAR3;HEYL , (0.09090900):RAD1;HEYL , (0.09090900):RAD1;HEYL;NPAS2	0.000		DNA binding
ACPP	(0.06666600):TAF12;ELF4 , (0.13333300):TAF12;ELF4;TNFAIP3 , (0.13333300):TAF12;ELF4;JRK , (0.18254200):ELF4;KIAA1018 , (0.20000000):TAF12;MYNN;STAT1[JAK][TOLL]	0.067		regulation of cell cycle
BLM	(0.06666600):ELF4;BCAR3 , (0.11587600):ELF4;ZNF262 , (0.13333300):ELF4;IL9R[CYT][JAK];AREG , (0.13333300):ELF4;BCAR3;AREG , (0.13649500):ELF4;AREG	0.067		ATP-dependent DNA helicase activity
FLJ13265	(0.06666600):RAD1;AREG , (0.11587600):ARNT;RAD1 , (0.13333300):MYNN;RAD1;AREG , (0.13333300):TACC1;RAD1;AREG , (0.13333300):RAD1;BCAR3;AREG	0.067		regulation of cell cycle
NUP153	(0.06666600):VEGFC[CYT];IL9R[CYT][JAK];TIMP1 , (0.27086700):TACC1;VEGFC[CYT];IL9R[CYT][JAK] , (0.27086700):TACC1;IL9R[CYT][JAK];TIMP1 , (0.28412300):VEGFC[CYT];ZNF262;TIMP1 , (0.29841800):PPP1R15A	0.067		DNA binding
PFDN1	(0.06666600):BUB1B[CC];NPAS2 , (0.11471400):HEYL , (0.11471400):HEYL;NPAS2 , (0.11587600):TXNDC;NPAS2 , (0.15079000):TXNDC;ZNF261	0.067		cell cycle

Gene	Preditores	Entropia	KEGG	Função Biológica
POLR3K	(0.06666600):ZNF75;ARNT , (0.07692300):ARNT;VEGFC[CYT];NCOA5 , (0.07692300):ARNT;NCOA5 , (0.13333300):SYCP2;ARNT;HES6 , (0.13333300):ARNT;CDK10;HES6	0.067		DNA-directed RNA polymerase III complex
SPOCK	(0.06666600):IRF5;RAD1;BCAR3 , (0.06666600):IRF5;BCAR3 , (0.07142800):IRF5;APPL , (0.07142800):IRF5;APPL;BCAR3 , (0.11471400):HEYL;NPAS2	0.067		cell proliferation
WHSC2	(0.06666600):ZNF177;HES6 , (0.11471400):HEYL , (0.11471400):HEYL;NPAS2 , (0.18181800):PPP1R15A;CDK8;HEYL , (0.18181800):PPP1R15A;ZNF261;HEYL	0.067		regulation of transcription, DNA-dependent
ZNF10	(0.06666600):POU2AF1;SYCP2 , (0.08412300):POU2AF1;NPAS2 , (0.11471400):HEYL , (0.11471400):HEYL;NPAS2 , (0.13333300):POU2AF1;STAT1[JAK][TOLL]	0.067		DNA binding
ZNF8	(0.06666600):BAZ2B;SYCP2 , (0.06666600):BAZ2B;SYCP2;NPAS2 , (0.06666600):SYCP2;TXNDC;NPAS2 , (0.06666600):SYCP2;NPAS2 , (0.11587600):SYCP2;TNFAIP3	0.067		DNA binding
CTNNBIP1	(0.07142800):TAF12;APPL , (0.14285700):PPP1R15A;APPL , (0.14285700):PPP1R15A;APPL;ZNF261 , (0.14285700):PPP1R15A;APPL;NPAS2 , (0.14285700):PPP1R15A;APPL;PRM1	0.071	[WNT]	cell proliferation
ESRRG	(0.07142800):BCAR3;STAT1[JAK][TOLL] , (0.14285700):TNFAIP3;BCAR3;STAT1[JAK][TOLL] , (0.14285700):BCAR3;STAT1[JAK][TOLL];AREG , (0.18026500):SYCP2;TXNDC , (0.19558100):STAT1[JAK][TOLL];AREG	0.071		positive regulation of transcription, DNA-dependent
MEOX1	(0.07142800):APPL;PRM1 , (0.13333300):POU2AF1;PRM1 , (0.14285700):POU2AF1;APPL;PRM1 , (0.14285700):PPP1R15A;APPL;PRM1 , (0.14285700):APPL;ZNF261;PRM1	0.071		regulation of transcription, DNA-dependent
NCOA6	(0.07142800):TAF12;APPL , (0.11471400):HEYL;NPAS2 , (0.18181800):HEYL;NOTCH4;NPAS2 , (0.20000000):TAF12;TNFAIP3 , (0.20000000):SYCP2;MTMR7;APPL	0.071		DNA recombination
NR2F1	(0.07142800):TAF12;APPL , (0.11587600):TAF12 , (0.11587600):TAF12;CDK8 , (0.11587600):TAF12;TNFAIP3 , (0.15079000):TAF12;ARNT	0.071		regulation of transcription, DNA-dependent
PLCE1	(0.07142800):IRF5;APPL , (0.07142800):IRF5;APPL;BCAR3 , (0.11471400):HEYL;NPAS2 , (0.14285700):IRF5;APPL;HES6 , (0.18181800):HEYL;NOTCH4;NPAS2	0.071		cell proliferation
PPP1R15A	(0.07142800):TAF12;APPL , (0.11471400):HEYL;NPAS2 , (0.11587600):TAF12 , (0.11587600):TAF12;CDK8 , (0.11587600):TAF12;TNFAIP3	0.071		cell cycle arrest
RPS6KA4	(0.07142800):IRF5;APPL , (0.07142800):IRF5;APPL;BCAR3 , (0.11471400):HEYL;NPAS2 , (0.14285700):IRF5;APPL;HES6 , (0.18181800):HEYL;NOTCH4;NPAS2	0.071	[MAPK]	regulation of transcription, DNA-dependent

Gene	Preditores	Entropia	KEGG	Função Biológica
SMCY	(0.07142800):IRF5;APPL , (0.07142800):IRF5;APPL;BCAR3 , (0.11471400):HEYL;NPAS2 , (0.14285700):IRF5;APPL;HES6 , (0.18181800):HEYL;NOTCH4;NPAS2	0.071		DNA binding
SOX4	(0.07142800):IRF5;APPL , (0.07142800):IRF5;APPL;BCAR3 , (0.14285700):IRF5;APPL;HES6 , (0.14285700):PPP1R15A;APPL;NPAS2 , (0.14285700):APPL;ZNF261;NPAS2	0.071		regulation of transcription, DNA-dependent
TGFBI	(0.07142800):IRF5;APPL , (0.07142800):IRF5;APPL;BCAR3 , (0.11471400):HEYL;NPAS2 , (0.14285700):IRF5;APPL;HES6 , (0.18181800):HEYL;NOTCH4;NPAS2	0.071		cell proliferation
XRCC4	(0.07142800):SYCP2;TNFAIP3;HES6 , (0.12415200):SYCP2;TNFAIP3 , (0.22618500):TAF12;MTMR7;CDK8 , (0.22618500):MTMR7;PPP1R15A;CDK8 , (0.22618500):MTMR7;CDK8	0.071		DNA recombination
POLG	(0.07692300):TACC1;NCOA5 , (0.15079000):SYCP2;TNFAIP3;HES6 , (0.15079000):CDK8;TNFAIP3;HES6 , (0.15079000):CDK8;HES6 , (0.15079000):TNFAIP3;HES6	0.077		DNA binding
PTN	(0.07692300):TACC1;NCOA5 , (0.13333300):MYNN;PPP1R15A;TACC1 , (0.13333300):MYNN;TACC1;IL9R[CYT][JAK] , (0.13333300):ZNF75;HES6;KIAA1018 , (0.13333300):ZNF75;KIAA1018	0.077		cell proliferation
SPO11	(0.07692300):SYCP2;TXNDC;HES6 , (0.07692300):SYCP2;CDK10;HES6 , (0.07692300):SYCP2;TNFAIP3;HES6 , (0.07692300):SYCP2;HES6 , (0.07692300):ZNF177;HES6	0.077		DNA binding
TNFSF4	(0.07692300):TACC1;BUB1B[CC];TNFAIP3 , (0.15384600):TAF12;TNFAIP3;STAT1[JAK][TOLL] , (0.15384600):VEGFC[CYT];CDK8;TNFAIP3 , (0.15384600):CDK8;TNFAIP3;STAT1[JAK][TOLL] , (0.21062600):PPP1R15A;VEGFC	0.077	[CYT]	positive regulation of cell proliferation
TNP2	(0.07692300):TACC1;NCOA5 , (0.15384600):MYNN;TACC1;NCOA5 , (0.15384600):TACC1;VEGFC[CYT];NCOA5 , (0.15384600):TACC1;IL9R[CYT][JAK];NCOA5 , (0.16666600):ARNT;APPL;NCOA5	0.077		DNA binding
ZNF7	(0.07692300):APPL;PRM1 , (0.14285700):POU2AF1;PRM1 , (0.15384600):POU2AF1;APPL;PRM1 , (0.18181800):PPP1R15A;HEYL;PRM1 , (0.18181800):CDK8;HEYL;PRM1	0.077		DNA binding
NCOA5	(0.08333300):TACC1;APPL , (0.09706600):TAF12;TXNDC , (0.15384600):TACC1;BUB1B[CC] , (0.16666600):IRF5;TACC1;APPL , (0.16666600):TACC1;APPL;BCAR3	0.083		regulation of transcription, DNA-dependent
FGF1	(0.08412300):IRF5;RAD1 , (0.13333300):IRF5;TACC1;RAD1 , (0.15079000):IRF5;RAD1;BCAR3 , (0.15079000):ZNF75;TACC1 , (0.15079000):ZNF75;TACC1;TIMP1	0.084	[MAPK][ACTIN]	cell proliferation
PCBP4	(0.08412300):TXNDC;NPAS2 , (0.15079000):SYCP2;TXNDC;CDK10 , (0.15079000):SYCP2;TXNDC;NPAS2 , (0.15079000):TXNDC;CDK10 , (0.15079000):TXNDC;CDK10;HES6	0.084		DNA binding

Tabela B.4: Tabela com os 38 genes melhores preditos no terceiro passo

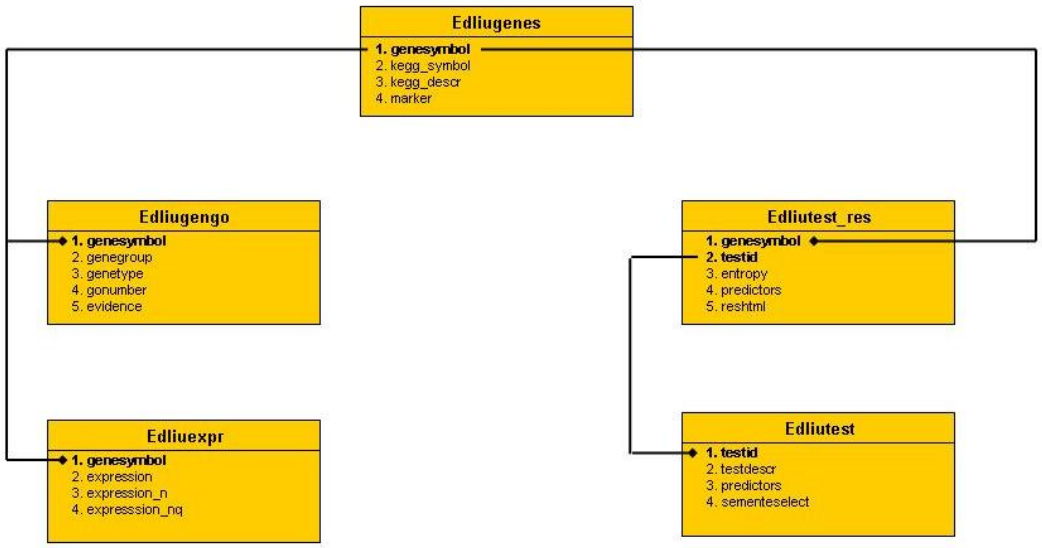


Figura B.1: Representação esquemática do banco de dados

Apêndice C

Adesão celular

A Tabela C.1 contém os 235 genes melhores preditos (custo < 0.08) pelo conjunto de 53 genes-sementes contidos na Tabela 5.1. A coluna preditores indica os melhores 5 conjuntos preditores para cada gene.

Gene	Preditores	Entropia	KEGG	Função Biológica
B4GALT2	(0.00000000):SCN1B;GREB1, (0.06666600):NIF3L1;GREB1, (0.06666600):GREB1;CD7, (0.13333300):NOL5A;GREB1, (0.15079000):NOL5A;CTSD	0.000		carbohydrate metabolism
C8G	(0.00000000):NRIP1;PTGES, (0.00000000):PTGES;STC2, (0.06666600):NRIP1;PTGES;IGFBP4, (0.06666600):NRIP1;PTGES;STC2, (0.06666600):NRIP1;IGFBP4	0.000		complement activation, alternative pathway
CDH16	(0.00000000):ALOX12B, (0.07692300):NRIP1;ALOX12B, (0.07692300):F10;ALOX12B, (0.07692300):AMD1;ALOX12B, (0.07692300):PTGES;ALOX12B	0.000		cell adhesion
ERG	(0.00000000):ADCY9;MPP3, (0.18254200):AMD1;PAFAH1B1, (0.21745700):BMP7[CYT][TGF];UGCGL1, (0.23491400):UGCGL1;PTGES, (0.24920900):AMD1;ADCY9	0.000		cell proliferation
HGNT-IV-H	(0.00000000):CISH[JAK];FLJ13710, (0.09090900):CISH[JAK];FLJ13710;HIG2, (0.09090900):STC2;FLJ13710;HIG2, (0.09090900):FLJ13710;HIG2, (0.15801200):AMD1;CISH[JAK]	0.000		carbohydrate metabolism
SFRS6	(0.00000000):MPP3;KCNG1, (0.13333300):HSPC111;MPP3, (0.13333300):CTSD;MPP3, (0.13333300):F10;OLFM1, (0.13333300):F10;PAFAH1B1	0.000		mRNA splice site selection
CLDN18	(0.00000000):RFPL2;IL6ST[CYT][JAK], (0.09090900):TPD52L1;RFPL2;IL6ST[CYT][JAK], (0.09090900):RFPL2;PAFAH1B1;IL6ST[CYT][JAK], (0.23076900):JAK1[JAK];SIAH2;PAFAH1B1, (0.23076900):ABCA3;FLJ20986;	0.000		calcium-independent cell-cell adhesion
L1CAM	(0.00000000):ALOX12B, (0.07692300):NRIP1;ALOX12B, (0.07692300):F10;ALOX12B, (0.07692300):AMD1;ALOX12B, (0.07692300):PTGES;ALOX12B	0.000		cell adhesion
PES1	(0.00000000):SCN1B;GREB1, (0.06666600):GREB1;CD7, (0.13333300):SIAH2;SCN1B, (0.13333300):FLJ20986;SCN1B, (0.13333300):SCN1B;CD7	0.000		cell proliferation
MCM6	(0.00000000):NRIP1;PTGES, (0.06666600):NRIP1;PTGES;IGFBP4, (0.06666600):NRIP1;PTGES;STC2, (0.15079000):UGCGL1;PTGES, (0.17398900):NRIP1;ALOX12B	0.000	[CC]	DNA replication
ABCB8	(0.00000000):LOR;BRI3BP, (0.09090900):AFG3L2;LOR;BRI3BP, (0.15801200):AFG3L2;BRI3BP, (0.16666600):NOL5A;LOR, (0.18181800):NOL5A;LOR;BRI3BP	0.000		transport
CDH17	(0.00000000):CAP350;BRI3BP, (0.00000000):BRI3BP, (0.07692300):F10;BRI3BP, (0.07692300):PTGES;BRI3BP, (0.07692300):AFG3L2;PAFAH1B1;BRI3BP	0.000		cell adhesion

Gene	Preditores	Entropia	KEGG	Função Biológica
FAP	(0.00000000):CENTG1;SERPINE1, (0.21428500):CENTG1;SERPINE1;EPHA4, (0.21428500):CENTG1;SERPINE1;PGR, (0.24830500):CENTG1, (0.25169400):CENTG1;CRABP2	0.000		cell-cell signaling
MAP3K1	(0.00000000):SCN1B;GREB1, (0.06666600):NIF3L1;GREB1, (0.06666600):GREB1;CD7, (0.13333300):NOL5A;GREB1, (0.15079000):NOL5A;CTSD	0.000	[MAPK] [AP] [TCELL]	protein amino acid phosphorylation
KIAA1463	(0.06666600):CTBS;ELOVL2;CCNG2, (0.11587600):CTBS;ELOVL2, (0.13333300):CTBS;LMCD1;ELOVL2, (0.13649500):CTBS;CCNG2, (0.15079000):CTBS;CCNG2;CRABP2	0.067		metabolism
JRKL	(0.06666600):PTGES;ADCY9, (0.20000000):F10;ADCY9, (0.20000000):PTGES;LMCD1;ADCY9, (0.21745700):RBBP8;ADCY9, (0.21745700):ADCY9;IGFBP4	0.067		central nervous system development
AKR1C4	(0.06666600):CCNG2;LOR;CRABP2, (0.06666600):LOR;CRABP2, (0.11587600):ELOVL2;LOR, (0.11587600):CCNG2;LOR, (0.13333300):ELOVL2;LOR;CRABP2	0.067		androgen metabolism
GUCA1B	(0.06666600):NRIP1;PTGES;STC2, (0.11587600):NRIP1;PTGES, (0.11587600):PTGES;STC2, (0.13333300):PTGES;IGFBP4;STC2, (0.13333300):IGFBP4;STC2	0.067		cell-cell signaling
DELGEF	(0.06666600):GREB1;CD7, (0.11587600):SCN1B;GREB1, (0.20000000):RBBP8;SCN1B, (0.21062600):FLJ13710;ALOX12B, (0.21745700):SIAH2;SCN1B	0.067		signal transduction
ADRA2B	(0.06666600):THBS1[TGF] [COMM];CRABP2, (0.15079000):NOL5A;CTSD, (0.15079000):CTSD;AMD1, (0.15079000):CTSD;DGKZ, (0.15079000):AMD1;CRABP2	0.067		G-protein coupled receptor protein signaling pathway
MX2	(0.06666600):LOR;THBS1[TGF] [COMM], (0.08412300):RBBP8;NOL5A, (0.13333300):RBBP8;NOL5A;DGKZ, (0.13333300):RBBP8;LMCD1;EPHA4, (0.13333300):LMCD1;EPHA4	0.067		immune response
OSBPL2	(0.06666600):CTSD;AMD1, (0.26666600):CENTG1;SIAH2, (0.26666600):JAK1[JAK];CTSD;AMD1, (0.26666600):AMD1;PGR;KCNG1, (0.26666600):AMD1;CRABP2	0.067		lipid transport
HLA-DPA1	(0.06666600):HSPC111;NRIP1, (0.13333300):HSPC111;UGCGL1, (0.15079000):CENTG1;SIAH2, (0.15079000):JAK1[JAK];GREB1, (0.16156100):IL6ST[CYT] [JAK];BRI3BP	0.067		antigen presentation, exogenous antigen
FUT6	(0.06666600):SERPINE1;HIG2, (0.12415200):ABCA3;SERPINE1, (0.15079000):CAP350;CD7, (0.15079000):FLJ13710;HIG2, (0.17398900):SERPINE1;ALOX12B	0.067		L-fucose catabolism
BAT8	(0.06666600):JAK1[JAK];PAFAH1B1, (0.06666600):PAFAH1B1;GALNT4, (0.13333300):JAK1[JAK];GREB1;PAFAH1B1, (0.14285700):HSPC111;TPD52L1;IL6ST[CYT] [JAK], (0.18254200):HSPC111;IL6ST[CYT] [JAK]	0.067		chromatin modification
MASA	(0.06666600):F10;FLJ20986, (0.13333300):F10;FLJ20986;IGFBP4, (0.20000000):F10;FLJ20986;SLC38A1, (0.23491400):PTGES;FLJ20986, (0.25169400):F10;TPD52L1	0.067		metabolism

Gene	Preditores	Entropia	KEGG	Função Biológica
CENTB2	(0.06666600):HSPC111;IL6ST[CYT] [JAK], (0.06666600):HSPC111;IGSF4, (0.11587600):HSPC111, (0.13333300):HSPC111;RBBP8, (0.13333300):HSPC111;RBBP8;DGKZ	0.067		regulation of GTPase activity
GYS1	(0.06666600):NIF3L1;CTSD, (0.15079000):NIF3L1;GREB1, (0.20316200):NRIP1;LOR, (0.23076900):NRIP1;RFPL2, (0.23076900):NRIP1;RFPL2;KCNG1	0.067		glycogen biosynthesis
NIT2	(0.06666600):JAK1[JAK];AMD1, (0.20000000):JAK1[JAK];AMD1;ELOVL2, (0.21062600):AMD1;ALOX12B, (0.21745700):GREB1;IL6ST[CYT] [JAK];FLJ22269, (0.23491400):F10;FLJ20986	0.067		nitrogen compound metabolism
GUCY2D	(0.06666600):SIAH2;STC2, (0.15079000):ADCY9;GALNT4, (0.15384600):SIAH2;RFPL2;STC2, (0.20000000):BMP7[CYT] [TGF];EPHA4, (0.21062600):RFPL2;STC2	0.067		cGMP biosynthesis
CCNT2	(0.06666600):NRIP1;PTGES;STC2, (0.06666600):NRIP1;STC2, (0.15079000):RBBP8;CAP350, (0.16824700):NRIP1;PTGES, (0.16824700):PTGES;STC2	0.067		cytokinesis
FLJ20485	(0.06666600):AMD1;CRABP2, (0.14285700):AMD1;ABCA3, (0.14285700):ABCA3;OLFM1, (0.20316200):CTSD;THBS1[TGF] [COMM], (0.21428500):CTSD;ABCA3;THBS1[TGF] [COMM]	0.067		tRNA processing
HGF	(0.06666600):NFIA;PAFAH1B1, (0.06666600):FLJ20986;SERPINE1, (0.06666600):FLJ20986;SERPINE1;EPHA4, (0.06666600):FLJ20986;SLC38A1, (0.06666600):GREB1;SLC38A1	0.067	[CYT]	mitosis
ESRRB	(0.06666600):SCN1B;GALNT4, (0.16824700):CTBS;NOL5A, (0.20000000):SCN1B;GALNT4;KCNG1, (0.21745700):CTBS;LMCD1;ELOVL2, (0.21745700):SCN1B;DGKZ	0.067		regulation of transcription, DNA-dependent
BNIP3L	(0.06666600):FLJ20986;SLC38A1, (0.15079000):NIF3L1;GREB1, (0.15079000):F10;FLJ20986, (0.18254200):GREB1;PAFAH1B1, (0.20000000):F10;FLJ20986;SLC38A1	0.067		apoptosis
ZFX	(0.06666600):CENTG1;SIAH2, (0.21062600):AMD1;ALOX12B, (0.21428500):CENTG1;TPD52L1, (0.23491400):SIAH2;CAP350;IGFBP4, (0.24920900):NIF3L1;DGKZ	0.067		regulation of transcription
CIDEB	(0.06666600):JAK1[JAK];AMD1, (0.20000000):JAK1[JAK];AMD1;ELOVL2, (0.21745700):BMP7[CYT] [TGF];NRIP1, (0.26666600):JAK1[JAK];CTSD;AMD1, (0.26666600):JAK1[JAK];AMD1;CCNG2	0.067		DNA damage response, signal transduction resulting in induction of apoptosis
NFATC4	(0.06666600):FLJ20986;SLC38A1, (0.06666600):GREB1;SLC38A1, (0.13333300):HSPC111;OLFM1, (0.13333300):HSPC111;MPP3, (0.13333300):THBS1[TGF] [COMM];PAFAH1B1	0.067	[MAPK] [WNT] [KILL] [TCELL]	inflammatory response
SPOCK	(0.06666600):CTSD;DGKZ, (0.08412300):CTBS;NOL5A, (0.11587600):DGKZ, (0.13333300):CTBS;NOL5A;DGKZ, (0.13333300):CTBS;DGKZ	0.067		cell adhesion
MRPL4	(0.06666600):NRIP1;PTGES;IGFBP4, (0.11587600):NRIP1;PTGES, (0.13333300):RBBP8;F10;PTGES, (0.13333300):F10;PTGES, (0.13333300):F10;PTGES;IGFBP4	0.067		protein biosynthesis

Gene	Preditores	Entropia	KEGG	Função Biológica
CSPG2	(0.06666600):NIF3L1;GREB1, (0.18254200):NIF3L1;ADCY9, (0.20000000):NIF3L1;FLJ13710, (0.20000000):NIF3L1;KCNG1, (0.20000000):LMCD1;PAFAH1B1;STC2	0.067		cell recognition
KIFAP3	(0.06666600):CD7;ADCY9, (0.21745700):CD7;STC2, (0.23595200):RBBP8;NOL5A, (0.24920900):RBBP8;PTGES;DGKZ, (0.24920900):PTGES;DGKZ	0.067		microtubule-based process
RIPK3	(0.06666600):NIF3L1;GREB1, (0.21745700):F10;PTGES;STC2, (0.21745700):PTGES;PAFAH1B1, (0.21745700):CAP350;PAFAH1B1, (0.22061900):PTGES;STC2	0.067		protein amino acid phosphorylation
TNK1	(0.06666600):NIF3L1;GREB1, (0.15079000):NIF3L1;CTSD, (0.25091200):RFPL2;IL6ST[CYT] [JAK], (0.26666600):CTBS;SIAH2, (0.26666600):NIF3L1;CTSD;KCNG1	0.067		autophosphorylation
IL9R	(0.06666600):HSPC111;NRIP1, (0.07692300):F10;ALOX12B, (0.07692300):PTGES;ALOX12B, (0.08412300):NRIP1;PTGES, (0.08412300):PTGES;STC2	0.067	[CYT] [JAK]	cell proliferation
KEL	(0.06666600):GREB1;CD7, (0.11587600):SCN1B;GREB1, (0.15079000):RBBP8;PTGES, (0.15079000):RBBP8;PTGES;IGFBP4, (0.15079000):NIF3L1;GREB1	0.067		N-linked glycosylation
STK4	(0.06666600):CCNG2;LOR;CRABP2, (0.06666600):LOR;CRABP2, (0.06666600):CRABP2;FLJ22269, (0.11587600):ELOVL2;LOR, (0.11587600):CCNG2;LOR	0.067	[MAPK]	apoptosis
SSR2	(0.06666600):IL6ST[CYT] [JAK];GALNT4, (0.16156100):CAP350;BRI3BP;SLC38A1, (0.20000000):IL6ST[CYT] [JAK];IGSF4;GALNT4, (0.20000000):IL6ST[CYT] [JAK];GALNT4;KCNG1, (0.20316200):BMP7[CYT] [TGF];GALNT	0.067		cotranslational protein-membrane targeting
FIBP	(0.06666600):FLJ20986;SLC38A1, (0.06666600):GREB1;SLC38A1, (0.06666600):PAFAH1B1;SLC38A1, (0.13333300):SIAH2;SLC38A1, (0.20000000):HSPC111;SLC38A1	0.067		fibroblast growth factor receptor signaling pathway
HMGCS1	(0.06666600):JAK1[JAK];AMD1, (0.20000000):JAK1[JAK];AMD1;ELOVL2, (0.21062600):AMD1;ALOX12B, (0.21745700):GREB1;IL6ST[CYT] [JAK];FLJ22269, (0.23491400):F10;FLJ20986	0.067		acetyl-CoA metabolism
GPR64	(0.06666600):CENTG1;SIAH2, (0.06666600):CENTG1;CRABP2, (0.13333300):CENTG1;RBBP8, (0.16824700):CENTG1;SERPINE1, (0.19558100):ABCA3;LOR	0.067		neuropeptide signaling pathway
PCCA	(0.06666600):JAK1[JAK];AMD1, (0.20000000):JAK1[JAK];AMD1;ELOVL2, (0.21062600):AMD1;ALOX12B, (0.21745700):HSPC111;OLFM1, (0.21745700):HSPC111;MPP3	0.067		fatty acid metabolism
DOC2B	(0.06666600):NIF3L1;UGCGL1, (0.13333300):NIF3L1;SLC38A1, (0.16824700):NIF3L1, (0.16824700):NIF3L1;KCNG1, (0.18254200):NIF3L1;ADCY9	0.067		transport
NEU3	(0.06666600):HSPC111;F10, (0.06666600):HSPC111;F10;IGFBP4, (0.06666600):HSPC111;IGFBP4, (0.13333300):HSPC111;PTGES, (0.13333300):HSPC111;PTGES;IGFBP4	0.067		carbohydrate metabolism

Gene	Preditores	Entropia	KEGG	Função Biológica
WNT3	(0.06666600):AFG3L2;FLJ13710, (0.15079000):IL6ST[CYT] [JAK];GALNT4, (0.20000000):NFIA;AFG3L2;FLJ13710, (0.20000000):AFG3L2;IGFBP4;FLJ13710, (0.20316200):ADCY9;GALNT4	0.067	[WNT]	cell-cell signaling
ALAD	(0.06666600):NOL5A;AMD1, (0.13333300):NOL5A;AMD1;FLJ22269, (0.15079000):NIF3L1;ADCY9, (0.15079000):ADCY9;GALNT4, (0.17398900):RFPL2;STC2	0.067		heme biosynthesis
JRK	(0.06666600):GREB1;CD7, (0.11587600):SCN1B;GREB1, (0.13333300):NFIA;FLJ20986;PAFAH1B1, (0.15079000):NFIA;SCN1B, (0.15079000):SCN1B;OLFM1	0.067		biological process unknown
SLC34A1	(0.06666600):PAFAH1B1;IL6ST[CYT] [JAK], (0.15079000):AMD1;PAFAH1B1, (0.18848800):TPD52L1;RFPL2, (0.20316200):NFIA;PAFAH1B1, (0.21745700):THBS1[TGF] [COMM];PAFAH1B1	0.067		fluid secretion
SPRR1B	(0.06666600):PTGES;CD7, (0.15079000):NIF3L1;GREB1, (0.18254200):CAP350;CD7, (0.18254200):THBS1[TGF] [COMM];CRABP2, (0.20000000):BMP7[CYT] [TGF];PTGES;CD7	0.067		epidermis development
ARF5	(0.06666600):NRIP1;PTGES;STC2, (0.13333300):PTGES;PAFAH1B1, (0.16824700):PTGES;STC2, (0.21745700):F10;PAFAH1B1, (0.23071400):SCN1B;KCNG1	0.067		intracellular protein transport
GJA7	(0.06666600):RBBP8;THBS1[TGF] [COMM], (0.06666600):LOR;THBS1[TGF] [COMM], (0.08412300):FLJ20986;EPHA4, (0.11587600):CCNG2;LOR, (0.11587600):THBS1[TGF] [COMM]	0.067	[COMM]	cell communication
ABCE1	(0.06666600):BMP7[CYT] [TGF];KCNG1, (0.20000000):BMP7[CYT] [TGF];IGFBP4;KCNG1, (0.20000000):BMP7[CYT] [TGF];IGSF4;KCNG1, (0.20000000):BMP7[CYT] [TGF];GALNT4;KCNG1, (0.20000000):UGCGL1;IL6ST[CYT] [J	0.067		electron transport
PEX11B	(0.06666600):GREB1;SLC38A1, (0.20000000):JAK1[JAK];GREB1;SLC38A1, (0.20000000):GREB1;PAFAH1B1;SLC38A1, (0.21745700):NRIP1;GREB1, (0.23491400):FLJ20986;SLC38A1	0.067		peroxisome division
PROCR	(0.06666600):JAK1[JAK];GREB1, (0.06666600):JAK1[JAK];CISH[JAK], (0.11587600):JAK1[JAK], (0.13333300):CENTG1;JAK1[JAK], (0.13333300):NFIA;JAK1[JAK]	0.067		blood coagulation
BRMS1	(0.06666600):BMP7[CYT] [TGF];HSPC111, (0.06666600):BMP7[CYT] [TGF];GALNT4, (0.15079000):BMP7[CYT] [TGF];F10, (0.15079000):BMP7[CYT] [TGF];F10;IGFBP4, (0.15079000):BMP7[CYT] [TGF];PTGES	0.067		cell cycle
MFN1	(0.06666600):ADCY9;GALNT4, (0.11587600):GALNT4;KCNG1, (0.13333300):SERPINE1;GALNT4;KCNG1, (0.13333300):PGR;GALNT4, (0.18254200):IL6ST[CYT] [JAK];GALNT4	0.067		mitochondrial fusion
AKR1B1	(0.06666600):IL6ST[CYT] [JAK];FLJ13710, (0.13333300):AMD1;IL6ST[CYT] [JAK];FLJ13710, (0.14285700):TPD52L1;IL6ST[CYT] [JAK];FLJ13710, (0.18254200):IL6ST[CYT] [JAK];GALNT4, (0.20000000):UGCGL1;IL6ST	0.067		carbohydrate metabolism
DOC1	(0.06666600):NIF3L1;CTSD, (0.13333300):CTSD;SCN1B, (0.15079000):CTSD;DGKZ, (0.20000000):BMP7[CYT] [TGF];HSPC111;F10, (0.20000000):CTSD;PTGES;SCN1B	0.067		biological process unknown

Gene	Preditores	Entropia	KEGG	Função Biológica
G6PC	(0.06666600):ADCY9;GALNT4, (0.13333300):BMP7[CYT] [TGF];NOL5A, (0.15079000):BMP7[CYT] [TGF];LMCD1, (0.15079000):BMP7[CYT] [TGF];LMCD1;CCNG2, (0.15079000):BMP7[CYT] [TGF];GALNT4	0.067		glucose metabolism
NBL1	(0.06666600):CTSD;DGKZ, (0.20000000):RBBP8;CTSD;DGKZ, (0.23298900):TPD52L1;THBS1[TGF] [COMM];FLJ22269, (0.23491400):NIF3L1;GREB1, (0.23491400):SCN1B;GALNT4	0.067		cell cycle
ETFB	(0.06666600):F10;FLJ20986, (0.13333300):F10;FLJ20986;IGFBP4, (0.19558100):CAP350;BRI3BP;SLC38A1, (0.20000000):F10;FLJ20986;SLC38A1, (0.21745700):F10;PAFAH1B1	0.067		electron transport
CDKN1B	(0.06666600):RBBP8;LMCD1, (0.06666600):RBBP8;LMCD1;IGFBP4, (0.06666600):RBBP8;THBS1[TGF] [COMM], (0.06666600):RBBP8;EPHA4, (0.06666600):LMCD1;IGFBP4	0.067	[CC]	cell cycle arrest
RPA2	(0.06666600):NOL5A;CTSD, (0.15384600):CTSD;RFPL2, (0.16156100):NOL5A;TPD52L1, (0.19558100):ABCA3;LOR, (0.20000000):NOL5A;CTSD;LMCD1	0.067		DNA-dependent DNA replication
MAD2L2	(0.06666600):UGCGL1;PTGES, (0.15079000):NRIP1;PTGES;STC2, (0.15079000):SCN1B;GALNT4, (0.16824700):ADCY9;MPP3, (0.17398900):NRIP1;ALOX12B	0.067	[CC]	cell cycle
BLM	(0.06666600):UGCGL1;PTGES, (0.06666600):SCN1B;FLJ22269, (0.08412300):GALNT4;KCNG1, (0.11587600):UGCGL1, (0.11587600):UGCGL1;IL6ST[CYT] [JAK]	0.067		DNA recombination
PALM	(0.06666600):ADCY9;GALNT4, (0.13333300):GREB1;GALNT4, (0.15079000):RBBP8;IGFBP4;DGKZ, (0.15079000):RBBP8;DGKZ;IGSF4, (0.15079000):FLJ20986;SLC38A1	0.067		cell motility
TBX4	(0.06666600):ADCY9;FLJ22269, (0.13333300):HSPC111;AFG3L2;ADCY9, (0.13333300):HSPC111;ADCY9, (0.13333300):RBBP8;AMD1, (0.13333300):RBBP8;ADCY9	0.067		development
PDCD2	(0.06666600):CD7;ADCY9, (0.15079000):F10;EPHA4, (0.21745700):F10;EPHA4;IGFBP4, (0.26666600):GREB1;CD7, (0.28571400):TPD52L1;IL6ST[CYT] [JAK];FLJ22269	0.067		apoptosis
IRX4	(0.06666600):CCNG2;LOR;CRABP2, (0.11587600):CCNG2;LOR, (0.13333300):PAFAH1B1;PGR, (0.13333300):PAFAH1B1;CRABP2, (0.15079000):LOR;THBS1[TGF] [COMM]	0.067		heart development
CHST7	(0.06666600):CRABP2;FLJ22269, (0.11587600):ELOVL2;LOR, (0.15079000):ELOVL2;CCNG2;LOR, (0.17398900):AMD1;ALOX12B, (0.18254200):AFG3L2;ELOVL2;LOR	0.067		N-acetylglucosamine metabolism
PPIG	(0.06666600):CRABP2;FLJ13710, (0.06666600):FLJ13710;HIG2, (0.07692300):SERPINE1;ALOX12B, (0.07692300):FLJ13710;ALOX12B, (0.11587600):CISH[JAK];FLJ13710	0.067		RNA splicing

Gene	Preditores	Entropia	KEGG	Função Biológica
PKP4	(0.06666600):CAP350;CD7, (0.15079000):SERPINE1;HIG2, (0.18254200):PTGES;CD7, (0.20000000):BMP7[CYT] [TGF];PTGES;CD7, (0.20000000):RBBP8;CTSD;CAP350	0.067		cell adhesion
HSPE1	(0.06666600):NIF3L1;GREB1, (0.21745700):F10;PTGES;STC2, (0.21745700):F10;PAFAH1B1, (0.22061900):PTGES;STC2, (0.23491400):RBBP8;F10;IGFBP4	0.067		protein folding
CORO2B	(0.06666600):NFIA;PAFAH1B1, (0.06666600):FLJ20986;SERPINE1, (0.06666600):FLJ20986;SERPINE1;EPHA4, (0.06666600):FLJ20986;SLC38A1, (0.06666600):GREB1;SLC38A1	0.067		actin cytoskeleton organization and biogenesis
STK19	(0.06666600):JAK1[JAK];AMD1, (0.20000000):JAK1[JAK];AMD1;ELOVL2, (0.21062600):AMD1;ALOX12B, (0.21745700):GREB1;IL6ST[CYT] [JAK];FLJ22269, (0.23491400):F10;FLJ20986	0.067		protein amino acid phosphorylation
CORO1C	(0.06666600):NIF3L1;CTSD, (0.18254200):CTSD;OLFM1, (0.18254200):PTGES;OLFM1, (0.20000000):F10;PTGES;OLFM1, (0.20316200):CTSD;F10	0.067		phagocytosis
HPN	(0.06666600):ADCY9;GALNT4, (0.21745700):PGR;GALNT4, (0.26666600):NOL5A;AFG3L2, (0.26666600):UGCGL1;ADCY9;GALNT4, (0.26666600):EPHA4;IL6ST[CYT] [JAK]	0.067		proteolysis and peptidolysis
SFRP5	(0.06666600):NOL5A;CTSD, (0.16156100):NOL5A;TPD52L1, (0.20000000):NOL5A;CTSD;LMCD1, (0.20000000):SCN1B;GREB1, (0.20000000):SCN1B;GREB1;STC2	0.067	[WNT]	Wnt receptor signaling pathway
MAT2A	(0.06666600):LOR;THBS1[TGF] [COMM], (0.13333300):F10;SIAH2, (0.13333300):F10;SIAH2;IGFBP4, (0.20000000):HSPC111;NRIP1;IGSF4, (0.20000000):HSPC111;F10;SIAH2	0.067		one-carbon compound metabolism
LTB	(0.06666600):NIF3L1;UGCGL1, (0.11587600):UGCGL1;IL6ST[CYT] [JAK], (0.21745700):UGCGL1;SLC38A1, (0.21745700):ADCY9;SLC38A1, (0.23076900):UGCGL1;ALOX12B	0.067	[CYT]	cell-cell signaling
GGPS1	(0.06666600):PAFAH1B1;GALNT4, (0.15079000):FLJ20986;SLC38A1, (0.15079000):GREB1;SLC38A1, (0.20000000):JAK1[JAK];PAFAH1B1;GALNT4, (0.20000000):FLJ20986;PAFAH1B1;SLC38A1	0.067		isoprenoid biosynthesis
PPYR1	(0.06666600):JAK1[JAK];AMD1, (0.06666600):CTSD;AMD1, (0.06666600):AMD1;CRABP2, (0.07692300):AMD1;ALOX12B, (0.11587600):AMD1	0.067		G-protein coupled receptor protein signaling pathway
DCI	(0.06666600):BMP7[CYT] [TGF];LMCD1, (0.06666600):BMP7[CYT] [TGF];LMCD1;CCNG2, (0.11587600):BMP7[CYT] [TGF];CCNG2, (0.13333300):BMP7[CYT] [TGF];NOL5A, (0.15079000):BMP7[CYT] [TGF];ELOVL2	0.067		fatty acid metabolism
EDNRA	(0.06666600):CTSD;AMD1, (0.26666600):CENTG1;SIAH2, (0.26666600):JAK1[JAK];CTSD;AMD1, (0.26666600):AMD1;PGR;KCNG1, (0.26666600):AMD1;CRABP2	0.067		G-protein coupled receptor protein signaling pathway
SSR3	(0.06666600):SCN1B;PAFAH1B1, (0.20000000):NIF3L1;SCN1B;KCNG1, (0.20000000):SCN1B;PAFAH1B1;STC2, (0.24920900):SIAH2;SCN1B, (0.24920900):SCN1B;LOR	0.067		cotranslational protein-membrane targeting

Gene	Preditores	Entropia	KEGG	Função Biológica
ADAM29	(0.06666600):SCN1B;KCNG1, (0.07692300):NRIP1;ALOX12B, (0.08412300):NRIP1;PTGES, (0.13333300):CTSD;SCN1B, (0.13333300):PTGES;SCN1B	0.067		proteolysis and peptidolysis
CYP11B1	(0.06666600):NRIP1;LOR, (0.19558100):LOR;BRI3BP, (0.20000000):LOR;MPP3, (0.23298900):LOR;PAFAH1B1;BRI3BP, (0.23491400):F10;FLJ20986	0.067		C21-steroid hormone biosynthesis
CRMP1	(0.06666600):JAK1[JAK];FLJ22269, (0.06666600):SCN1B;FLJ22269, (0.06666600):CRABP2;FLJ22269, (0.08412300):ELOVL2;LOR, (0.08412300):CCNG2;LOR	0.067		neurogenesis
SLC17A3	(0.06666600):BMP7[CYT] [TGF];AMD1, (0.17398900):AMD1;ALOX12B, (0.20000000):BMP7[CYT] [TGF];NRIP1;AMD1, (0.20000000):NRIP1;AMD1, (0.23076900):FLJ20986;ALOX12B	0.067		ion transport
PTCH	(0.06666600):GREB1;FLJ22269, (0.13333300):GREB1;IL6ST[CYT] [JAK];FLJ22269, (0.15079000):RBBP8;CAP350, (0.15079000):NOL5A;FLJ22269, (0.15079000):NRIP1;PTGES;STC2	0.067		cell cycle
SLC12A4	(0.06666600):SCN1B;OLFM1, (0.08412300):SCN1B;GREB1, (0.13333300):NIF3L1;SCN1B, (0.13333300):SCN1B;CD7, (0.13333300):SCN1B;IL6ST[CYT] [JAK]	0.067		amino acid transport
PER2	(0.06666600):NFIA;RBBP8, (0.11587600):NFIA, (0.11587600):NFIA;FLJ20986, (0.13333300):NFIA;AMD1, (0.13333300):NFIA;AFG3L2	0.067		circadian rhythm
FLJ14249	(0.06666600):NIF3L1;GREB1, (0.26666600):NIF3L1;GREB1;LMCD1, (0.26666600):NIF3L1;GREB1;IL6ST[CYT] [JAK], (0.28412300):BMP7[CYT] [TGF];PTGES;CD7, (0.29841800):NRIP1;HIG2	0.067		intracellular signaling cascade
KCNK7	(0.06666600):UGCGL1;PTGES, (0.06666600):CTSD;AMD1, (0.06666600):AMD1;CRABP2, (0.15079000):PTGES;ADCY9, (0.15384600):AFG3L2;RFPL2;ADCY9	0.067		ion transport
PLA2R1	(0.06666600):RBBP8;CAP350, (0.11587600):SIAH2;CAP350, (0.13333300):RBBP8;CAP350;IGFBP4, (0.13333300):RBBP8;CAP350;DGKZ, (0.13333300):RBBP8;CAP350;IGSF4	0.067		endocytosis
TNP1	(0.06666600):SCN1B;CISH[JAK], (0.18254200):NFIA;SCN1B, (0.20000000):NFIA;SCN1B;CISH[JAK], (0.22061900):NRIP1;PTGES, (0.23076900):PTGES;CCNG2;ALOX12B	0.067		cell differentiation
SIAT8E	(0.06666600):ELOVL2;CRABP2, (0.11587600):ELOVL2;LOR, (0.13333300):CTBS;ELOVL2;CRABP2, (0.13333300):ELOVL2;CCNG2;CRABP2, (0.13333300):ELOVL2;LOR;CRABP2	0.067		carbohydrate metabolism
C17orf31	(0.06666600):NIF3L1;GREB1, (0.21745700):F10;PTGES;STC2, (0.21745700):PTGES;PAFAH1B1, (0.21745700):CAP350;PAFAH1B1, (0.22061900):PTGES;STC2	0.067		telomerase-dependent telomere maintenance
CHDH	(0.06666600):NIF3L1;GREB1, (0.14285700):AFG3L2;LOR;BRI3BP, (0.15079000):GREB1;IL6ST[CYT] [JAK], (0.15079000):IL6ST[CYT] [JAK];GALNT4, (0.18254200):NIF3L1;CTSD	0.067		electron transport

Gene	Preditores	Entropia	KEGG	Função Biológica
RPS16	(0.06666600):CAP350;CD7, (0.20000000):CTSD;CAP350;CD7, (0.20000000):F10;SCN1B, (0.20000000):FLJ20986;CD7, (0.20000000):SCN1B;CAP350	0.067		protein biosynthesis
GYG2	(0.06666600):PAFAH1B1;KCNG1, (0.13333300):HSPC111;KCNG1, (0.13333300):NIF3L1;PAFAH1B1, (0.13333300):NIF3L1;PAFAH1B1;KCNG1, (0.13333300):CTSD;KCNG1	0.067		carbohydrate biosynthesis
CHRAC1	(0.06666600):PTGES;CD7, (0.18254200):CAP350;CD7, (0.20000000):BMP7[CYT] [TGF];PTGES;CD7, (0.20000000):PTGES;CD7;STC2, (0.20000000):PTGES;STC2	0.067		chromatin remodeling
NFIA	(0.06666600):CTBS;ELOVL2;CCNG2, (0.11587600):CTBS;ELOVL2, (0.13333300):CTBS;LMCD1;ELOVL2, (0.13649500):CTBS;CCNG2, (0.15079000):CTBS;CCNG2;CRABP2	0.067		DNA replication
PKMYT1	(0.06666600):GREB1;FLJ22269, (0.13333300):GREB1;IL6ST[CYT] [JAK];FLJ22269, (0.15079000):NOL5A;AMD1, (0.15079000):NOL5A;FLJ22269, (0.15079000):ADCY9;FLJ22269	0.067	[CC]	cell cycle
TNFAIP1	(0.06666600):NOL5A;NRIP1, (0.14285700):NOL5A;BRI3BP, (0.20000000):NIF3L1;NOL5A, (0.20000000):NIF3L1;NOL5A;ADCY9, (0.20316200):NRIP1;LOR	0.067		potassium ion transport
FLJ22222	(0.06666600):CTSD;DGKZ, (0.08412300):RBBP8, (0.08412300):RBBP8;NOL5A, (0.08412300):RBBP8;IGFBP4, (0.08412300):RBBP8;IGSF4	0.067		protein metabolism
MYCN	(0.06666600):GREB1;SLC38A1, (0.20000000):JAK1[JAK];GREB1;SLC38A1, (0.20000000):GREB1;PAFAH1B1;SLC38A1, (0.21745700):NRIP1;GREB1, (0.23491400):FLJ20986;SLC38A1	0.067		regulation of transcription from RNA polymerase II promoter
AATK	(0.06666600):HSPC111;NRIP1, (0.20000000):HSPC111;NRIP1;UGCGL1, (0.20000000):HSPC111;NRIP1;IGFBP4, (0.20000000):HSPC111;NRIP1;IGSF4, (0.24920900):HSPC111;UGCGL1	0.067		protein amino acid phosphorylation
TAF6	(0.06666600):JAK1[JAK];CTSD;SERPINE1, (0.13333300):JAK1[JAK];CTSD;CAP350, (0.13333300):CTSD;CAP350;SERPINE1, (0.19558100):F10;BRI3BP, (0.21428500):THBS1[TGF] [COMM];BRI3BP	0.067		regulation of transcription factor activity
COL17A1	(0.06666600):SCN1B;CISH[JAK], (0.15079000):NRIP1;PTGES;STC2, (0.18254200):UGCGL1;PTGES, (0.20000000):NFIA;SCN1B;CISH[JAK], (0.20000000):NRIP1;PTGES	0.067	[COMM]	cell-matrix adhesion
SNRPF	(0.06666600):UGCGL1;PTGES, (0.22061900):NRIP1;PTGES, (0.23076900):CTBS;RFPL2;ELOVL2, (0.23491400):NRIP1;PTGES;STC2, (0.26666600):NRIP1;UGCGL1;PTGES	0.067		nuclear mRNA splicing, via spliceosome
LSAMP	(0.06666600):BMP7[CYT] [TGF];HSPC111, (0.06666600):BMP7[CYT] [TGF];GALNT4, (0.11587600):HSPC111, (0.11587600):GALNT4;KCNG1, (0.13333300):HSPC111;JAK1[JAK]	0.067		cell adhesion
EEF1A2	(0.06666600):NIF3L1;GREB1, (0.11587600):SCN1B;GREB1, (0.15079000):GREB1;CD7, (0.18254200):NIF3L1;ADCY9, (0.21062600):FLJ13710;ALOX12B	0.067		protein biosynthesis

Gene	Preditores	Entropia	KEGG	Função Biológica
COL4A6	(0.06666600):SIAH2;STC2, (0.08333300):TPD52L1;RFPL2, (0.08412300):NRIP1;PTGES, (0.08412300):PTGES;STC2, (0.09013200):CAP350;BRI3BP	0.067	[COMM]	cell adhesion
JAK2	(0.06666600):FLJ20986;SERPINE1, (0.06666600):FLJ20986;SERPINE1;EPHA4, (0.06666600):SERPINE1;EPHA4, (0.15079000):ADCY9;GALNT4, (0.20000000):FLJ20986;SERPINE1;PAFAH1B1	0.067	[JAK]	JAK-STAT cascade
ADCY8	(0.06666600):BMP7[CYT] [TGF];GALNT4, (0.11587600):GALNT4;KCNG1, (0.13333300):PTGES;GALNT4, (0.15079000):HSPC111;IGSF4, (0.15079000):SCN1B;GALNT4	0.067		cAMP biosynthesis
MMP26	(0.06666600):SCN1B;GALNT4, (0.18254200):SCN1B;FLJ22269, (0.20000000):BMP7[CYT] [TGF];HSPC111;GALNT4, (0.20000000):SCN1B;GALNT4;KCNG1, (0.21745700):UGCGL1;SCN1B	0.067		collagen catabolism
RPS6KA5	(0.06666600):PAFAH1B1;IL6ST[CYT] [JAK], (0.18848800):TPD52L1;RFPL2, (0.19558100):AFG3L2;PAFAH1B1;BRI3BP, (0.20316200):PAFAH1B1;SLC38A1, (0.21745700):F10;PAFAH1B1	0.067	[MAPK]	DNA damage induced protein phosphorylation
UCHL1	(0.06666600):GREB1;CD7, (0.11587600):SCN1B;GREB1, (0.15079000):RBBP8;PTGES, (0.15079000):RBBP8;PTGES;IGFBP4, (0.15079000):NIF3L1;GREB1	0.067		protein deubiquitination
ERBB2IP	(0.06666600):THBS1[TGF] [COMM];CRABP2, (0.14285700):TPD52L1;THBS1[TGF] [COMM];FLJ22269, (0.15079000):AMD1;CRABP2, (0.20316200):CTSD;THBS1[TGF] [COMM], (0.21745700):F10;CRABP2	0.067		basal protein localization
RARRES2	(0.06666600):NOL5A;CTSD, (0.11587600):SCN1B;GREB1, (0.15079000):GREB1;CD7, (0.16156100):NOL5A;TPD52L1, (0.20000000):NOL5A;CTSD;LMCD1	0.067		retinoid metabolism
PTGFR	(0.06666600):SCN1B;GALNT4, (0.11587600):GALNT4;KCNG1, (0.13333300):HSPC111;OLFM1, (0.13333300):HSPC111;MPP3, (0.15079000):BMP7[CYT] [TGF];GALNT4	0.067		G-protein coupled receptor protein signaling pathway
HAO1	(0.06666600):IL6ST[CYT] [JAK];GALNT4, (0.20000000):JAK1[JAK];PAFAH1B1;GALNT4, (0.20000000):IL6ST[CYT] [JAK];IGSF4;GALNT4, (0.20000000):IL6ST[CYT] [JAK];GALNT4;KCNG1, (0.20316200):BMP7[CYT] [TGF];G	0.067		electron transport
RGS3	(0.06666600):SCN1B;KCNG1, (0.20000000):NIF3L1;SCN1B;KCNG1, (0.20000000):SCN1B;GREB1, (0.20000000):SCN1B;GALNT4;KCNG1, (0.23491400):SCN1B;GALNT4	0.067		inactivation of MAPK
PLA2G6	(0.06666600):CRABP2;FLJ13710, (0.06666600):FLJ13710;HIG2, (0.07692300):SERPINE1;ALOX12B, (0.07692300):FLJ13710;ALOX12B, (0.11587600):CISH[JAK];FLJ13710	0.067	[MAPK]	lipid catabolism
MYST1	(0.06666600):NRIP1;PTGES;IGFBP4, (0.06666600):NRIP1;IGFBP4, (0.11587600):NRIP1;PTGES, (0.13333300):NRIP1;F10, (0.13333300):NRIP1;F10;IGFBP4	0.067		chromatin assembly or disassembly
BRCA2	(0.06666600):JAK1[JAK];GREB1, (0.13333300):JAK1[JAK];GREB1;PAFAH1B1, (0.15079000):JAK1[JAK];AMD1, (0.18254200):JAK1[JAK];PAFAH1B1, (0.20000000):NFIA;JAK1[JAK];PAFAH1B1	0.067		DNA repair

Gene	Preditores	Entropia	KEGG	Função Biológica
NEUROD1	(0.06666600):BMP7[CYT] [TGF];LMCD1, (0.06666600):BMP7[CYT] [TGF];LMCD1;CCNG2, (0.06666600):CTBS;LMCD1, (0.06666600):CTBS;LMCD1;CCNG2, (0.06666600):AFG3L2;LMCD1;CCNG2	0.067		cell differentiation
AK5	(0.07142800):AFG3L2;PAFAH1B1;BRI3BP, (0.07142800):PAFAH1B1;BRI3BP, (0.14285700):CAP350;PAFAH1B1;BRI3BP, (0.14285700):LOR;PAFAH1B1;BRI3BP, (0.18026500):CAP350;BRI3BP	0.071		ADP biosynthesis
GLE1L	(0.07142800):CAP350;BRI3BP;SLC38A1, (0.13333300):CAP350;SLC38A1, (0.18254200):CAP350;CD7, (0.23491400):CENTG1;CRABP2, (0.26666600):CENTG1;SIAH2	0.071		mRNA-nucleus export
BMX	(0.07142800):FLJ20986;SERPINE1, (0.07142800):FLJ20986;SERPINE1;EPHA4, (0.12415200):FLJ20986;EPHA4, (0.14285700):EPHA4;PAFAH1B1, (0.16156100):RBBP8;EPHA4	0.071		intracellular signaling cascade
PTPN12	(0.07142800):SIAH2;TPD52L1, (0.13333300):FLJ20986;LMCD1, (0.14285700):RBBP8;SIAH2;TPD52L1, (0.14285700):RBBP8;TPD52L1;DGKZ, (0.14285700):SIAH2;TPD52L1;IL6ST[CYT] [JAK]	0.071		protein amino acid dephosphorylation
STAR	(0.07142800):FLJ20986;SERPINE1, (0.09090900):TPD52L1;RFPL2, (0.14285700):SIAH2;STC2, (0.14285700):FLJ20986;SERPINE1;EPHA4, (0.16156100):RBBP8;NOL5A	0.071		C21-steroid hormone biosynthesis
SLC22A7	(0.07142800):CAP350;BRI3BP;SLC38A1, (0.13333300):CAP350;SLC38A1, (0.18254200):CAP350;CD7, (0.23491400):CENTG1;CRABP2, (0.26666600):CENTG1;SIAH2	0.071		ion transport
RALBP1	(0.07142800):CRABP2;FLJ13710, (0.07142800):FLJ13710;GALNT4, (0.07142800):FLJ13710;HIG2, (0.08333300):FLJ13710;ALOX12B, (0.12415200):CISH[JAK];FLJ13710	0.071		chemotaxis
AVPR1B	(0.07142800):CAP350;BRI3BP;SLC38A1, (0.07142800):BRI3BP;SLC38A1, (0.12415200):CAP350;BRI3BP, (0.12415200):BRI3BP, (0.14285700):GALNT4;BRI3BP	0.071		G-protein coupled receptor protein signaling pathway
LRP1	(0.07142800):TPD52L1;THBS1[TGF] [COMM], (0.14285700):NFIA;TPD52L1;THBS1[TGF] [COMM], (0.14285700):TPD52L1;THBS1[TGF] [COMM];FLJ22269, (0.20000000):CTBS;GREB1, (0.20000000):HSPC111;FLJ20986	0.071		cell proliferation
GNRHR	(0.07142800):CAP350;BRI3BP;SLC38A1, (0.13333300):CAP350;SLC38A1, (0.15079000):PTGES;CD7, (0.17398900):NRIP1;ALOX12B, (0.18254200):CAP350;CD7	0.071		G-protein coupled receptor protein signaling pathway
SEMG2	(0.07142800):PAFAH1B1;KCNG1, (0.14285700):NIF3L1;PAFAH1B1, (0.14285700):NIF3L1;PAFAH1B1;KCNG1, (0.17398900):AFG3L2;PAFAH1B1;BRI3BP, (0.23076900):PTGES;CD7;TPD52L1	0.071		sexual reproduction
KERA	(0.07142800):JAK1[JAK];FLJ22269, (0.07142800):SCN1B;FLJ22269, (0.07692300):ABCA3;FLJ22269, (0.09013200):ELOVL2;LOR, (0.12415200):FLJ22269	0.071		eye morphogenesis (sensu Mammalia)
GPR57	(0.07142800):ABCA3;LOR, (0.07142800):LOR;BRI3BP, (0.08412300):ELOVL2;LOR, (0.08412300):CCNG2;LOR, (0.13333300):CENTG1;LOR	0.071		G-protein coupled receptor protein signaling pathway

Gene	Preditores	Entropia	KEGG	Função Biológica
FOXD3	(0.07142800):CAP350;BRI3BP;SLC38A1, (0.13333300):CAP350;SLC38A1, (0.18026500):CAP350;BRI3BP, (0.20000000):SIAH2;AFG3L2;ADCY9, (0.20000000):SIAH2;ADCY9	0.071		development
PFKFB4	(0.07142800):CAP350;CD7, (0.07692300):CAP350;BRI3BP;SLC38A1, (0.09013200):BMP7[CYT] [TGF];PTGES, (0.09013200):BMP7[CYT] [TGF];PTGES;IGFBP4, (0.14285700):BMP7[CYT] [TGF];PTGES;CD7	0.071		fructose 2,6-bisphosphate metabolism
CRSP6	(0.07142800):CTBS;NOL5A, (0.16156100):PTGES;SERPINE1, (0.21428500):CTBS;RBBP8;NOL5A, (0.21428500):CTBS;NOL5A;LMCD1, (0.21428500):CTBS;NOL5A;CCNG2	0.071		androgen receptor signaling pathway
VAMP4	(0.07142800):CAP350;BRI3BP;SLC38A1, (0.07142800):BRI3BP;SLC38A1, (0.08412300):SLC38A1, (0.13333300):NIF3L1;SLC38A1, (0.13333300):F10;SLC38A1	0.071		protein complex assembly
DYRK3	(0.07142800):CAP350;BRI3BP;SLC38A1, (0.13333300):CAP350;SLC38A1, (0.18026500):CAP350;BRI3BP, (0.21745700):NFIA;FLJ13710, (0.23076900):RFPL2;PAFAH1B1;IL6ST[CYT] [JAK]	0.071		protein amino acid phosphorylation
PAFAH2	(0.07142800):BMP7[CYT] [TGF];ABCA3, (0.12415200):CAP350;BRI3BP, (0.19558100):CAP350;BRI3BP;SLC38A1, (0.20000000):BMP7[CYT] [TGF];HSPC111;JAK1[JAK], (0.20316200):BMP7[CYT] [TGF];JAK1[JAK]	0.071		lipid catabolism
KCNC4	(0.07142800):IL6ST[CYT] [JAK];BRI3BP, (0.08412300):ADCY9;MPP3, (0.13333300):FLJ22269;SLC38A1, (0.15079000):BMP7[CYT] [TGF];HSPC111, (0.15079000):BMP7[CYT] [TGF];AMD1	0.071		cation transport
ARPC1B	(0.07142800):ABCA3;LOR, (0.14285700):NIF3L1;ABCA3, (0.14285700):SIAH2;TPD52L1;FLJ22269, (0.15079000):CENTG1;SIAH2, (0.15079000):CENTG1;CRABP2	0.071	[ACTIN]	cell motility
SEMA7A	(0.07142800):CAP350;BRI3BP;SLC38A1, (0.13333300):CAP350;SLC38A1, (0.18254200):CAP350;CD7, (0.23491400):CENTG1;CRABP2, (0.26666600):CENTG1;SIAH2	0.071		cell differentiation
TM4SF3	(0.07142800):SERPINE1;EPHA4, (0.12415200):EPHA4, (0.12415200):EPHA4;IL6ST[CYT] [JAK], (0.14285700):UGCGL1;EPHA4, (0.14285700):UGCGL1;EPHA4;SLC38A1	0.071		protein amino acid glycosylation
PTPRS	(0.07142800):AFG3L2;PAFAH1B1;BRI3BP, (0.13333300):AFG3L2;PAFAH1B1, (0.20000000):AFG3L2;LMCD1;PAFAH1B1, (0.20000000):AFG3L2;CCNG2;PAFAH1B1, (0.21428500):PAFAH1B1;BRI3BP;KCNG1	0.071		cell adhesion
KLHL1	(0.07142800):RBBP8;IGFBP4, (0.12415200):F10;IGFBP4, (0.12415200):PTGES;IGFBP4, (0.12415200):IGFBP4, (0.14285700):CTBS;PTGES;IGFBP4	0.071		actin cytoskeleton organization and biogenesis
TINAG	(0.07142800):EPHA4;IL6ST[CYT] [JAK], (0.12415200):NFIA;SCN1B, (0.14285700):NFIA;EPHA4;IL6ST[CYT] [JAK], (0.14285700):NIF3L1;EPHA4, (0.14285700):NIF3L1;OLFM1	0.071		Malpighian tubule morphogenesis
GPR58	(0.07142800):F10;TPD52L1, (0.07142800):TPD52L1;THBS1[TGF] [COMM], (0.14285700):NFIA;TPD52L1;THBS1[TGF] [COMM], (0.14285700):GREB1;TPD52L1, (0.14285700):TPD52L1;THBS1[TGF] [COMM];FLJ22269	0.071		G-protein coupled receptor protein signaling pathway

Gene	Preditores	Entropia	KEGG	Função Biológica
VPREB1	(0.07142800):TPD52L1;THBS1[TGF] [COMM], (0.13333300):CAP350;THBS1[TGF] [COMM], (0.14285700):NFIA;TPD52L1;THBS1[TGF] [COMM], (0.14285700):TPD52L1;THBS1[TGF] [COMM];FLJ22269, (0.19558100):TPD52L1;CI	0.071		immune response
PAPOLA	(0.07142800):IL6ST[CYT] [JAK];FLJ22269, (0.14285700):SCN1B;IL6ST[CYT] [JAK], (0.19558100):GREB1;IL6ST[CYT] [JAK], (0.21062600):NOL5A;TPD52L1, (0.21428500):GREB1;IL6ST[CYT] [JAK];FLJ22269	0.071		mRNA polyadenylylation
EGR1	(0.07142800):AFG3L2;PAFAH1B1;BRI3BP, (0.13333300):AFG3L2;PAFAH1B1, (0.15079000):FLJ20986;SERPINE1;EPHA4, (0.15384600):RFPL2;PAFAH1B1, (0.15384600):RFPL2;PAFAH1B1;STC2	0.071		regulation of transcription, DNA-dependent
RPS3	(0.07142800):LOR;THBS1[TGF] [COMM], (0.19558100):CCNG2;LOR;CRABP2, (0.21428500):HSPC111;F10;PTGES, (0.21428500):LMCD1;LOR;THBS1[TGF] [COMM], (0.21878600):CCNG2;LOR	0.071		protein biosynthesis
AKT2	(0.07142800):BMP7[CYT] [TGF];BRI3BP, (0.15079000):BMP7[CYT] [TGF];GALNT4, (0.18254200):JAK1[JAK];CTSD;SERPINE1, (0.18254200):IL6ST[CYT] [JAK];GALNT4, (0.23410200):IL6ST[CYT] [JAK];BRI3BP	0.071	[MAPK] [AP] [JAK] [TCELL] [TOLL]	protein amino acid phosphorylation
C2	(0.07142800):SIAH2;TPD52L1, (0.14285700):RBBP8;SIAH2;TPD52L1, (0.14285700):RBBP8;ABCA3, (0.14285700):RBBP8;TPD52L1;DGKZ, (0.14285700):SIAH2;TPD52L1;IL6ST[CYT] [JAK]	0.071		complement activation, classical pathway
SLC5A4	(0.07142800):AFG3L2;ADCY9, (0.15384600):AFG3L2;RFPL2;ADCY9, (0.16156100):NIF3L1;ADCY9, (0.19558100):SCN1B;GREB1, (0.21428500):HSPC111;AFG3L2;ADCY9	0.071		ion transport
SLC2A1	(0.07142800):JAK1[JAK];AMD1, (0.07142800):CTSD;AMD1, (0.07142800):AMD1;PGR, (0.07142800):AMD1;CRABP2, (0.12415200):AMD1	0.071		carbohydrate transport
IL16	(0.07142800):BMP7[CYT] [TGF];BRI3BP, (0.15079000):IGFBP4;CISH[JAK], (0.18254200):JAK1[JAK];FLJ22269, (0.21745700):RBBP8;IGFBP4;CISH[JAK], (0.21745700):F10;IGFBP4;CISH[JAK]	0.071		chemotaxis
VILL	(0.07142800):SIAH2;TPD52L1, (0.07142800):TPD52L1;THBS1[TGF] [COMM], (0.14285700):NFIA;TPD52L1, (0.14285700):NFIA;TPD52L1;THBS1[TGF] [COMM], (0.14285700):HSPC111;TPD52L1	0.071		cytoskeleton organization and biogenesis
IL1R1	(0.07142800):SCN1B;DGKZ, (0.12415200):DGKZ, (0.14285700):CTBS;GREB1, (0.14285700):JAK1[JAK];DGKZ, (0.14285700):CTSD;DGKZ	0.071	[MAPK] [AP] [CYT]	cell surface receptor linked signal transduction
RAMP1	(0.07142800):PTGES;ADCY9, (0.14285700):PTGES;LMCD1;ADCY9, (0.19558100):UGCGL1;PTGES, (0.21428500):F10;EPHA4;IGFBP4, (0.21428500):F10;ADCY9	0.071		intracellular protein transport
SLIT1	(0.07142800):NOL5A;TPD52L1, (0.16156100):PTGES;TPD52L1, (0.16156100):TPD52L1;FLJ13710, (0.16267300):TPD52L1, (0.18848800):TPD52L1;RFPL2	0.071		axon guidance
F8	(0.07142800):NFIA;SCN1B, (0.12415200):NFIA, (0.12415200):NFIA;FLJ20986, (0.12415200):FLJ20986, (0.12415200):FLJ20986;EPHA4	0.071		acute-phase response

Gene	Preditores	Entropia	KEGG	Função Biológica
MRPS15	(0.07142800):PTGES;ADCY9, (0.16156100):BMP7[CYT] [TGF];CTSD;F10, (0.16156100):BMP7[CYT] [TGF];F10, (0.16156100):ADCY9;IGFBP4, (0.21428500):F10;ADCY9	0.071		protein biosynthesis
DAP13	(0.07142800):IL6ST[CYT] [JAK];BRI3BP, (0.16824700):ADCY9;MPP3, (0.24920900):AFG3L2;IL6ST[CYT] [JAK], (0.25091200):RFPL2;IL6ST[CYT] [JAK], (0.26666600):BMP7[CYT] [TGF];AMD1	0.071		electron transport
MAP3K4	(0.07692300):SERPINE1;ALOX12B, (0.21062600):PTGES;ALOX12B, (0.22061900):NRIP1;PTGES, (0.23076900):SERPINE1;FLJ13710;ALOX12B, (0.23076900):SERPINE1;ALOX12B;KCNG1	0.077	[MAPK]	JNK cascade
B3GALT2	(0.07692300):FLJ13710;ALOX12B, (0.15079000):CRABP2;FLJ13710, (0.15079000):FLJ13710;HIG2, (0.17398900):SERPINE1;ALOX12B, (0.18254200):GREB1;CD7	0.077		protein amino acid glycosylation
PSMC3	(0.07692300):SERPINE1;ALOX12B, (0.20000000):MPP3;IL6ST[CYT] [JAK], (0.20000000):MPP3;FLJ13710, (0.21745700):MPP3;IGFBP4, (0.21745700):MPP3;IGFBP4;KCNG1	0.077		protein catabolism
PDE3A	(0.07692300):FLJ13710;ALOX12B, (0.15079000):CRABP2;FLJ13710, (0.15079000):FLJ13710;HIG2, (0.17398900):SERPINE1;ALOX12B, (0.18254200):GREB1;CD7	0.077		lipid metabolism
SLC13A3	(0.07692300):SERPINE1;ALOX12B, (0.07692300):FLJ13710;ALOX12B, (0.15079000):AMD1;PAFAH1B1, (0.15079000):ADCY9;FLJ22269, (0.15079000):CRABP2;FLJ13710	0.077		ion transport
C5R1	(0.07692300):AMD1;ALOX12B, (0.15079000):THBS1[TGF] [COMM];KCNG1, (0.15079000):OLFM1;KCNG1, (0.20000000):NFIA;CTSD;FLJ20986, (0.21745700):AMD1;PGR	0.077		G-protein coupled receptor protein signaling pathway
TXN2	(0.07692300):F10;ALOX12B, (0.15079000):NRIP1;PTGES;STC2, (0.20000000):NRIP1;PTGES, (0.23076900):F10;ELOVL2;ALOX12B, (0.23076900):F10;CCNG2;ALOX12B	0.077		electron transport
RAG2	(0.07692300):NRIP1;ALOX12B, (0.15079000):HSPC111;NRIP1, (0.15384600):NRIP1;JAK1[JAK];ALOX12B, (0.15384600):JAK1[JAK];ALOX12B, (0.16824700):NRIP1;PTGES	0.077		DNA recombination
ASB2	(0.07692300):SCN1B;GALNT4, (0.13370300):GALNT4;KCNG1, (0.15384600):PAFAH1B1;DGKZ;KCNG1, (0.15384600):DGKZ;KCNG1, (0.17398900):HSPC111;KCNG1	0.077		intracellular signaling cascade
CNTNAP2	(0.07692300):FLJ13710;ALOX12B, (0.15079000):CRABP2;FLJ13710, (0.15079000):FLJ13710;HIG2, (0.17398900):SERPINE1;ALOX12B, (0.18254200):GREB1;CD7	0.077		cell adhesion
IL18R1	(0.07692300):NRIP1;ALOX12B, (0.08412300):NRIP1;PTGES, (0.15079000):NRIP1;PTGES;IGFBP4, (0.15079000):NRIP1;PTGES;STC2, (0.15079000):UGCGL1;PTGES	0.077	[CYT]	immune response
BLNK	(0.07692300):SCN1B;DGKZ, (0.15384600):SCN1B;PAFAH1B1;DGKZ, (0.15384600):ADCY9;DGKZ, (0.15384600):DGKZ;STC2, (0.15749500):DGKZ	0.077		B cell differentiation

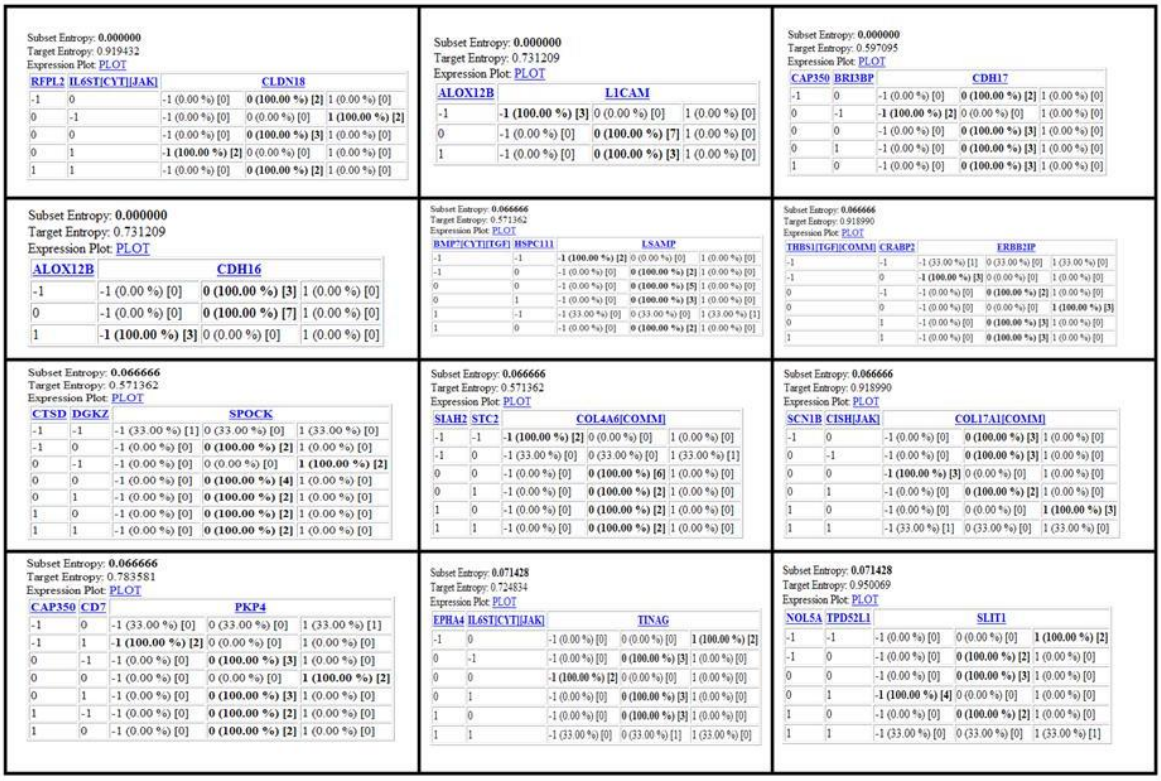
Gene	Preditores	Entropia	KEGG	Função Biológica
KITLG	(0.07692300):NRIP1;ALOX12B, (0.15384600):NRIP1;JAK1[JAK];ALOX12B, (0.23076900):NRIP1;LMCD1;ALOX12B, (0.23076900):NRIP1;ELOVL2;ALOX12B, (0.23076900):NRIP1;CCNG2;ALOX12B	0.077	[CYT]	cell adhesion
WRN	(0.07692300):CAP350;TPD52L1, (0.13370300):TPD52L1;THBS1[TGF] [COMM], (0.14285700):F10;EPHA4, (0.14285700):THBS1[TGF] [COMM];EPHA4, (0.14285700):THBS1[TGF] [COMM];EPHA4;FLJ22269	0.077		DNA metabolism
NEO1	(0.07692300):HSPC111;NRIP1, (0.08333300):THBS1[TGF] [COMM];BRI3BP, (0.09706600):NRIP1;PTGES, (0.13370300):HSPC111, (0.13370300):HSPC111;IL6ST[CYT] [JAK]	0.077		cell adhesion
KIAA1404	(0.07692300):SERPINE1;ALOX12B, (0.16824700):CTBS;NOL5A, (0.19558100):PTGES;BRI3BP, (0.20000000):CENTG1;SERPINE1, (0.20000000):FLJ20986;EPHA4	0.077		regulation of transcription, DNA-dependent
ZNF16	(0.07692300):BMP7[CYT] [TGF];ABCA3, (0.07692300):CTSD;ABCA3, (0.07692300):ABCA3;PAFAH1B1, (0.13370300):ABCA3, (0.13370300):ABCA3;SERPINE1	0.077		regulation of transcription, DNA-dependent
ELF5	(0.07692300):LMCD1;ALOX12B, (0.15384600):LMCD1;MPP3;ALOX12B, (0.15384600):MPP3;ALOX12B, (0.18254200):NIF3L1;CTSD, (0.21745700):CENTG1;MPP3	0.077		cell proliferation
UBD	(0.07692300):RFPL2;ELOVL2, (0.13370300):RFPL2, (0.15384600):AFG3L2;RFPL2;ELOVL2, (0.15384600):RFPL2;CCNG2, (0.18848800):TPD52L1;RFPL2	0.077		antimicrobial humoral response (sensu Vertebrata)
ZNF277	(0.07692300):RFPL2;STC2, (0.08333300):TPD52L1;RFPL2, (0.13370300):RFPL2, (0.15079000):PAFAH1B1;KCNG1, (0.15384600):RBBP8;RFPL2	0.077		regulation of transcription, DNA-dependent
NPEPL1	(0.07692300):NFIA;EPHA4;IL6ST[CYT] [JAK], (0.13370300):EPHA4;IL6ST[CYT] [JAK], (0.17398900):PAFAH1B1;GALNT4, (0.21062600):SIAH2;IL6ST[CYT] [JAK], (0.23076900):SIAH2;IL6ST[CYT] [JAK];FLJ13710	0.077		protein metabolism
PITPNB	(0.07692300):AMD1;ALOX12B, (0.23076900):AMD1;ELOVL2;ALOX12B, (0.28754900):CTBS;ELOVL2;ALOX12B, (0.28754900):CTBS;ALOX12B, (0.30261900):BMP7[CYT] [TGF];AMD1	0.077		lipid metabolism
DEFA4	(0.07692300):RFPL2;STC2, (0.08333300):TPD52L1;RFPL2, (0.08412300):RBBP8;NOL5A, (0.13333300):AMD1;STC2, (0.13370300):RFPL2	0.077		defense response to bacteria
GOSR1	(0.07692300):F10;ALOX12B, (0.07692300):PTGES;ALOX12B, (0.15384600):HSPC111;ALOX12B, (0.15384600):CD7;ALOX12B, (0.21062600):NRIP1;ALOX12B	0.077		ER to Golgi transport
IFNG	(0.07692300):LMCD1;ALOX12B, (0.15079000):CTSD;LMCD1, (0.15079000):LMCD1;PAFAH1B1, (0.15384600):RFPL2;LMCD1, (0.15384600):LMCD1;MPP3;ALOX12B	0.077	[CYT] [JAK] [KILL] [TGF] [TCELL]	cell motility
NOV	(0.07692300):NRIP1;ALOX12B, (0.13333300):CENTG1;JAK1[JAK], (0.13333300):JAK1[JAK];SLC38A1, (0.15384600):NRIP1;JAK1[JAK];ALOX12B, (0.15384600):JAK1[JAK];ALOX12B	0.077		regulation of cell growth

Gene	Preditores	Entropia	KEGG	Função Biológica
ITPKA	(0.07692300):SERPINE1;ALOX12B, (0.07692300):FLJ13710;ALOX12B, (0.15079000):AMD1;PAFAH1B1, (0.15079000):ADCY9;FLJ22269, (0.15079000):CRABP2;FLJ13710	0.077		signal transduction
WBP2	(0.07692300):AFG3L2;RFPL2, (0.07692300):AFG3L2;RFPL2;BRI3BP, (0.07692300):RFPL2;BRI3BP, (0.15384600):AFG3L2;RFPL2;ELOVL2, (0.15384600):AFG3L2;RFPL2;ADCY9	0.077		biological process unknown
CARD15	(0.07692300):F10;ALOX12B, (0.15079000):CENTG1;CRABP2, (0.15384600):CD7;ALOX12B, (0.20000000):UGCGL1;IL6ST[CYT] [JAK], (0.23076900):F10;ELOVL2;ALOX12B	0.077		regulation of apoptosis
NPY	(0.07692300):FLJ13710;ALOX12B, (0.15079000):CRABP2;FLJ13710, (0.15079000):FLJ13710;HIG2, (0.17398900):SERPINE1;ALOX12B, (0.18254200):GREB1;CD7	0.077		G-protein signaling, coupled to cyclic nucleotide second messenger
ANGPTL2	(0.07692300):AMD1;ALOX12B, (0.13333300):AMD1;HIG2, (0.15079000):JAK1[JAK];AMD1, (0.15079000):CTSD;AMD1, (0.15079000):AMD1;CRABP2	0.077		development
CHIC2	(0.07692300):NRIP1;ALOX12B, (0.13333300):JAK1[JAK];PTGES, (0.13333300):JAK1[JAK];KCNG1, (0.15384600):NRIP1;JAK1[JAK];ALOX12B, (0.15384600):JAK1[JAK];ALOX12B	0.077		biological process unknown
HOXD3	(0.07692300):NRIP1;ALOX12B, (0.15384600):NRIP1;JAK1[JAK];ALOX12B, (0.22061900):NRIP1;PTGES, (0.23076900):NRIP1;LMCD1;ALOX12B, (0.23076900):NRIP1;ELOVL2;ALOX12B	0.077		morphogenesis
FUT4	(0.07692300):SERPINE1;ALOX12B, (0.12415200):ABCA3;SERPINE1, (0.13333300):CTBS;ELOVL2;CRABP2, (0.13333300):SERPINE1;STC2, (0.15079000):CTBS;CCNG2;CRABP2	0.077		L-fucose catabolism
SAFB	(0.07692300):NRIP1;ALOX12B, (0.15384600):NRIP1;JAK1[JAK];ALOX12B, (0.20000000):NRIP1;PTGES, (0.21062600):PTGES;ALOX12B, (0.23076900):NRIP1;LMCD1;ALOX12B	0.077		establishment and/or maintenance of chromatin architecture
SMARCA2	(0.07692300):AMD1;ALOX12B, (0.21745700):NRIP1;FLJ22269, (0.23076900):AMD1;ELOVL2;ALOX12B, (0.23298900):CAP350;PAFAH1B1;BRI3BP, (0.23491400):BMP7[CYT] [TGF];KCNG1	0.077		cell cycle
DMAP1	(0.07692300):NRIP1;ALOX12B, (0.13333300):CENTG1;JAK1[JAK], (0.13333300):JAK1[JAK];SLC38A1, (0.15384600):NRIP1;JAK1[JAK];ALOX12B, (0.15384600):JAK1[JAK];ALOX12B	0.077		DNA methylation
PASK	(0.07692300):TPD52L1;CISH[JAK];FLJ13710, (0.13370300):TPD52L1;CISH[JAK], (0.16156100):RBBP8;LMCD1;IGFBP4, (0.16156100):AMD1;HIG2, (0.19413200):TPD52L1;FLJ22269	0.077		protein amino acid phosphorylation
GCNT3	(0.07692300):FLJ13710;ALOX12B, (0.15079000):CRABP2;FLJ13710, (0.15079000):FLJ13710;HIG2, (0.17398900):SERPINE1;ALOX12B, (0.18254200):GREB1;CD7	0.077		O-linked glycosylation
UBE2D2	(0.07692300):RBBP8;IL6ST[CYT] [JAK], (0.07692300):RBBP8;IL6ST[CYT] [JAK];DGKZ, (0.07692300):IL6ST[CYT] [JAK];DGKZ, (0.15384600):RBBP8;FLJ20986;IL6ST[CYT] [JAK], (0.15384600):RBBP8;IL6ST[CYT] [JAK];	0.077		ubiquitin cycle

Gene	Preditores	Entropia	KEGG	Função Biológica
SULT2A1	(0.07692300):FLJ13710;ALOX12B, (0.11587600):SCN1B;GREB1, (0.15079000):PTGES;CD7, (0.15079000):CAP350;CD7, (0.15079000):CRABP2;FLJ13710	0.077		bile acid catabolism
HOOK1	(0.07692300):SERPINE1;ALOX12B, (0.12415200):ABCA3;SERPINE1, (0.15079000):SERPINE1;KCNG1, (0.20000000):CENTG1;SERPINE1;GALNT4, (0.21428500):ABCA3;SERPINE1;KCNG1	0.077		cell differentiation
DNAJB5	(0.07692300):AMD1;ALOX12B, (0.14285700):AMD1;ABCA3, (0.15079000):THBS1[TGF] [COMM];KCNG1, (0.18254200):JAK1[JAK];AMD1, (0.18254200):CTSD;AMD1	0.077		protein folding
FBXO21	(0.07692300):LMCD1;ALOX12B, (0.13370300):ALOX12B, (0.15079000):CTSD;LMCD1, (0.15079000):LMCD1;PAFAH1B1, (0.15384600):RBBP8;DGKZ;ALOX12B	0.077		ubiquitin cycle
COL8A1	(0.07692300):SERPINE1;ALOX12B, (0.21062600):FLJ13710;ALOX12B, (0.21428500):ABCA3;FLJ13710, (0.21745700):STC2;FLJ13710, (0.23076900):SERPINE1;FLJ13710;ALOX12B	0.077		cell adhesion
IMPA2	(0.07692300):F10;ALOX12B, (0.15384600):CD7;ALOX12B, (0.23076900):F10;ELOVL2;ALOX12B, (0.23076900):F10;CCNG2;ALOX12B, (0.23076900):CD7;ELOVL2;ALOX12B	0.077		phosphate metabolism
MST1R	(0.07692300):PTGES;ALOX12B, (0.16824700):NRIP1;PTGES, (0.21062600):SERPINE1;ALOX12B, (0.22061900):PTGES;STC2, (0.23076900):HSPC111;PTGES;ALOX12B	0.077		cell motility
ADCYAP1R1	(0.07692300):FLJ13710;ALOX12B, (0.15079000):NOL5A;AMD1, (0.15079000):NOL5A;FLJ22269, (0.15079000):CRABP2;FLJ13710, (0.15079000):FLJ13710;HIG2	0.077		G-protein coupled receptor protein signaling pathway
NMI	(0.07692300):SERPINE1;ALOX12B, (0.23076900):SERPINE1;FLJ13710;ALOX12B, (0.23076900):SERPINE1;ALOX12B;KCNG1, (0.23076900):ELOVL2;FLJ13710;ALOX12B, (0.23076900):CCNG2;FLJ13710;ALOX12B	0.077		JAK-STAT cascade
NRP2	(0.07692300):AMD1;ALOX12B, (0.14285700):AMD1;ABCA3, (0.15079000):THBS1[TGF] [COMM];KCNG1, (0.18254200):JAK1[JAK];AMD1, (0.18254200):CTSD;AMD1	0.077		angiogenesis
DLL3	(0.07692300):AMD1;ALOX12B, (0.21428500):CTSD;ABCA3;THBS1[TGF] [COMM], (0.21428500):ABCA3;THBS1[TGF] [COMM], (0.23076900):AMD1;ELOVL2;ALOX12B, (0.24920900):AMD1;HIG2	0.077		Notch signaling pathway
GTF2H1	(0.07692300):AFG3L2;RFPL2, (0.07692300):AFG3L2;RFPL2;BRI3BP, (0.07692300):RFPL2;BRI3BP, (0.13370300):RFPL2, (0.14285700):TPD52L1;THBS1[TGF] [COMM];FLJ22269	0.077		DNA repair
ESRRG	(0.07692300):BMP7[CYT] [TGF];BRI3BP, (0.09013200):BMP7[CYT] [TGF];CCNG2, (0.09013200):BMP7[CYT] [TGF];KCNG1, (0.14285700):BMP7[CYT] [TGF];UGCGL1, (0.14624500):BMP7[CYT] [TGF]	0.077		development
DPH2L1	(0.07692300):F10;ALOX12B, (0.16156100):F10;TPD52L1, (0.20316200):HSPC111;F10, (0.20316200):HSPC111;F10;IGFBP4, (0.21428500):F10;ABCA3	0.077		protein biosynthesis

Tabela C.1: Lista dos 235 genes melhores preditos pelos genes-sementes da Tabela 5.1

A Figura C.1 mostra as tabelas de predições para os genes relacionados à função de adesão celular com os genes de seu melhor subconjunto de predição.



Subset Entropy: **0.000000**
Target Entropy: 0.919432
Expression Plot: PLOT

RFPL2	IL6ST[CYT][JAK]	CLDN18		
-1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	-1	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	1	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
1	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]

Subset Entropy: **0.000000**
Target Entropy: 0.731209
Expression Plot: PLOT

ALOX12B	L1CAM		
-1	**-1 (100.00 %) [3]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	-1 (0.00 %) [0]	**0 (100.00 %) [7]**	1 (0.00 %) [0]
1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]

Subset Entropy: **0.000000**
Target Entropy: 0.597095
Expression Plot: PLOT

CAP350	BRI3BP	CDH17		
-1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	-1	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]

Subset Entropy: **0.000000**
Target Entropy: 0.731209
Expression Plot: PLOT

ALOX12B	CDH16		
-1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	-1 (0.00 %) [0]	**0 (100.00 %) [7]**	1 (0.00 %) [0]
1	**-1 (100.00 %) [3]**	0 (0.00 %) [0]	1 (0.00 %) [0]

Subset Entropy: **0.066666**
Target Entropy: 0.571362
Expression Plot: PLOT

BMP7[CYT][TGF]	HSPC111	LSAMP		
-1	-1	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
-1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [5]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
1	-1	-1 (33.00 %) [0]	0 (33.00 %) [0]	1 (33.00 %) [1]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]

Subset Entropy: **0.066666**
Target Entropy: 0.918990
Expression Plot: PLOT

THBS1[TGF][COMM]	CRABP2	ERBB2IP		
-1	-1	-1 (33.00 %) [1]	0 (33.00 %) [0]	1 (33.00 %) [0]
-1	0	**-1 (100.00 %) [3]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [3]**
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
1	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]

Subset Entropy: **0.066666**
Target Entropy: 0.571362
Expression Plot: PLOT

CTSD	DGKZ	SPOCK		
-1	-1	-1 (33.00 %) [1]	0 (33.00 %) [0]	1 (33.00 %) [0]
-1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	-1	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [4]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]

Subset Entropy: **0.066666**
Target Entropy: 0.571362
Expression Plot: PLOT

SIAH2	STC2	COL4A6[COMM]		
-1	-1	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
-1	0	-1 (33.00 %) [0]	0 (33.00 %) [0]	1 (33.00 %) [1]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [6]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]

Subset Entropy: **0.066666**
Target Entropy: 0.918990
Expression Plot: PLOT

SCN1B	CISH[JAK]	COL17A1[COMM]		
-1	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	0	**-1 (100.00 %) [3]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [3]**
1	1	-1 (33.00 %) [1]	0 (33.00 %) [0]	1 (33.00 %) [0]

Subset Entropy: **0.066666**
Target Entropy: 0.783581
Expression Plot: PLOT

CAP350	CD7	PKP4		
-1	0	-1 (33.00 %) [0]	0 (33.00 %) [0]	1 (33.00 %) [1]
-1	1	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
1	-1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]

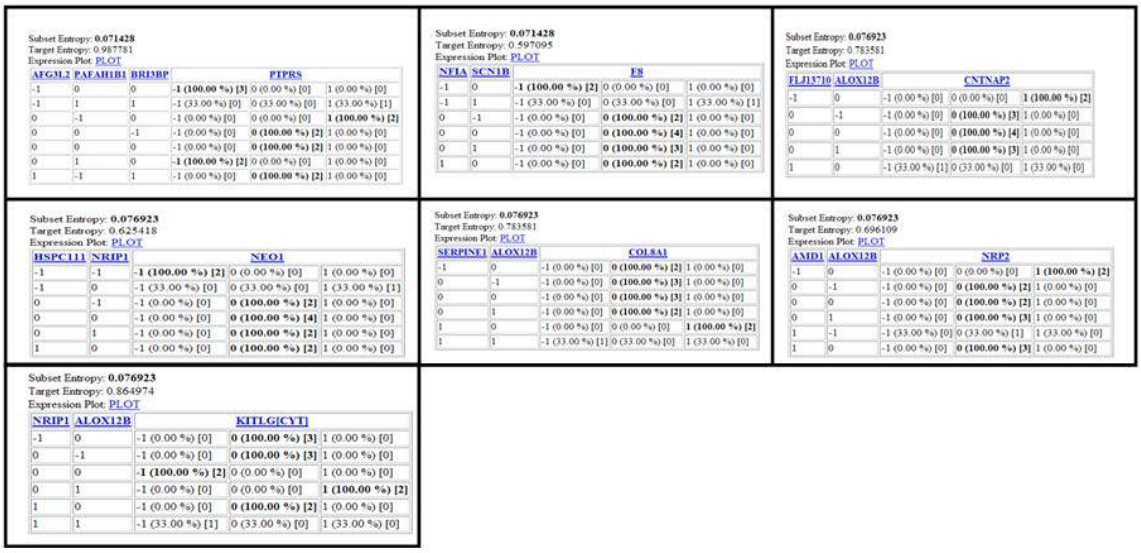
Subset Entropy: **0.071428**
Target Entropy: 0.724834
Expression Plot: PLOT

EPHA4	IL6ST[CYT][JAK]	TINAG		
-1	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	0	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
1	1	-1 (33.00 %) [0]	0 (33.00 %) [1]	1 (33.00 %) [0]

Subset Entropy: **0.071428**
Target Entropy: 0.950069
Expression Plot: PLOT

NOL5A	TPD52L1	SLIT1		
-1	-1	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
-1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	1	**-1 (100.00 %) [4]**	0 (0.00 %) [0]	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	1	-1 (33.00 %) [0]	0 (33.00 %) [0]	1 (33.00 %) [1]

A Figura C.2 mostra os gráficos dos sinais dos genes relacionados à função de adesão celular e os genes do seu melhor subconjuntos de predição.



Subset Entropy: **0.071428**
Target Entropy: 0.987781
Expression Plot: PLOT

AFG3L2	PAFAH1B1	BRI3BP	PTPRS		
-1	0	0	**-1 (100.00 %) [3]**	0 (0.00 %) [0]	1 (0.00 %) [0]
-1	1	1	-1 (33.00 %) [0]	0 (33.00 %) [0]	1 (33.00 %) [1]
0	-1	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
0	0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	1	0	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
1	-1	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]

Subset Entropy: **0.071428**
Target Entropy: 0.597095
Expression Plot: PLOT

NFIA	SCN1B	E8		
-1	0	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
-1	1	-1 (33.00 %) [0]	0 (33.00 %) [0]	1 (33.00 %) [1]
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [4]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]

Subset Entropy: **0.076923**
Target Entropy: 0.783581
Expression Plot: PLOT

FLJ13710	ALOX12B	CNTNAP2		
-1	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [4]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
1	0	-1 (33.00 %) [1]	0 (33.00 %) [0]	1 (33.00 %) [0]

Subset Entropy: **0.076923**
Target Entropy: 0.625418
Expression Plot: PLOT

HSPC111	NRIP1	NEO1		
-1	-1	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
-1	0	-1 (33.00 %) [0]	0 (33.00 %) [0]	1 (33.00 %) [1]
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [4]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]

Subset Entropy: **0.076923**
Target Entropy: 0.783581
Expression Plot: PLOT

SERPINE1	ALOX12B	COL8A1		
-1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
1	1	-1 (33.00 %) [1]	0 (33.00 %) [0]	1 (33.00 %) [0]

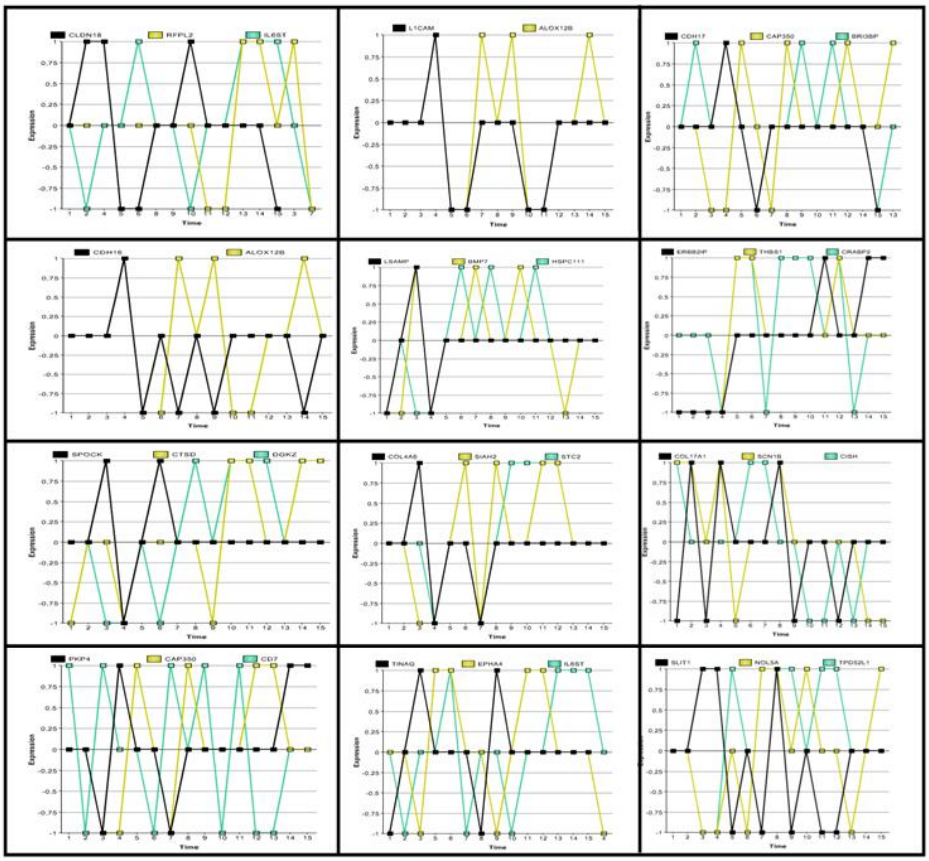
Subset Entropy: **0.076923**
Target Entropy: 0.696109
Expression Plot: PLOT

AMD1	ALOX12B	NRP2		
-1	0	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
1	-1	-1 (33.00 %) [0]	0 (33.00 %) [1]	1 (33.00 %) [0]
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]

Subset Entropy: **0.076923**
Target Entropy: 0.864974
Expression Plot: PLOT

NRIP1	ALOX12B	KITLG[CYT]		
-1	0	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	-1	-1 (0.00 %) [0]	**0 (100.00 %) [3]**	1 (0.00 %) [0]
0	0	**-1 (100.00 %) [2]**	0 (0.00 %) [0]	1 (0.00 %) [0]
0	1	-1 (0.00 %) [0]	0 (0.00 %) [0]	**1 (100.00 %) [2]**
1	0	-1 (0.00 %) [0]	**0 (100.00 %) [2]**	1 (0.00 %) [0]
1	1	-1 (33.00 %) [1]	0 (33.00 %) [0]	1 (33.00 %) [0]

Figura C.1: Tabelas de predições dos genes relacionados à adesão celular



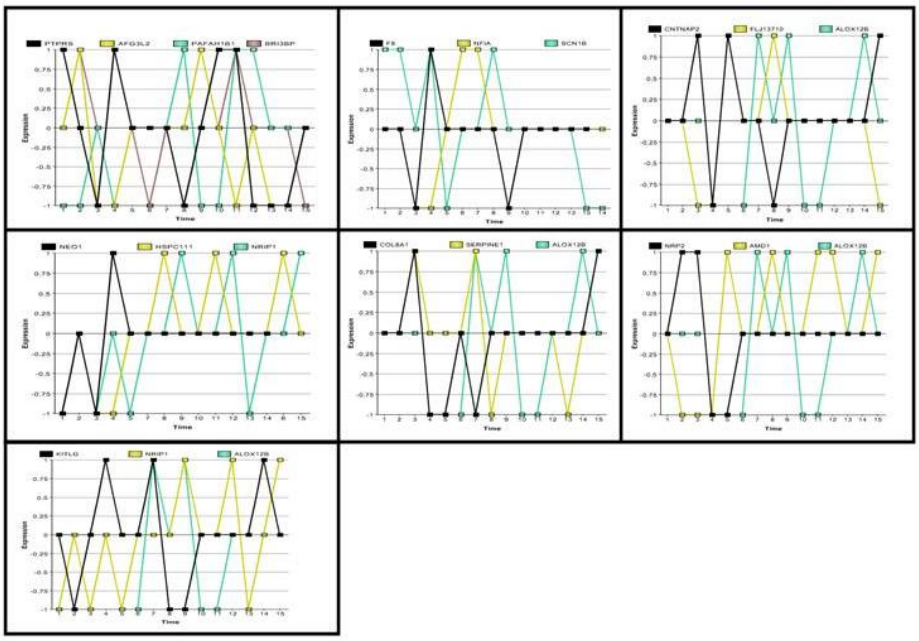


Figura C.2: Gráfico dos sinais dos genes relacionados à adesão celular

Apêndice D

Conteúdo do DVD

O DVD em anexo é composto pelos seguintes pastas:

- **bibliografia** – as referências bibliográficas deste trabalho;
- **trabalhos e apresentacoes** – todos os trabalhos, apresentações orais, pôsteres em congressos e conferências, e artigos científicos submetidos. Estes trabalhos estão dispostos em sub-pastas com o nome contendo o ano de execução como prefixo do nome (por exemplo: 2007_ISMM, significa conferência ISMM em 2007);
- **ucurve** – programa executável *ucurve.exe* para ser rodado em ambiente Microsoft Windows^{TM} com as seguintes sub-pastas:
  - **source** – todos os programas fontes com seus códigos em Microsoft C++;
  - **testes** – testes feitos na Seção 3.4, separados em: edliu (testes biológicos), e W-operador (testes de imagens). Os arquivos com extensão ".dat" contêm os dados de entrada, ".log" os dados de saída e os arquivos ".bat" são arquivos de execução em lote;
- **pipeline** – programas relativos ao pipeline de algoritmos com as seguintes sub-pastas:
  - **database** – script para geração do banco de dados MySql^{TM};
  - **genes sementes** – arquivos contendo o comando sql para obtenção dos genes-sementes e a lista dos genes-sementes para três passos do aalgoritmo;
  - **resultados** – o grafo resultante (**teste1_go.html**) obtido do teste executado na Seção 4.2.5, com redirecionamentos para a tablas de predição e gráficos dos sinais de expressão;
  - **source** – programas fontes do pipeline desenvolvidos em Microsoft Visual Studio 7.0^{TM}:
    - **csv2dbf.prg** - converte um arquivo separado por ";" para um arquivo DBF (usado para converter o arquivo edliu.csv em edliu.dbf);
    - **dbf_norm_new.prg** - programa de normalização e discretização dos dados (armazenando os resultados em um banco DBF);
    - **go2dbf.prg** – converte o arquivo de retorno **genes_go.txt** da pesquisa em "*Gene Ontology Database*" (<`http://www.geneontology.org`>) para um banco de dados DBF (**edliugo.dbf**);
    - **uploadgo.prg** – move os dados do banco de dados DBF (**edliugo.dbf**) para o banco de dados MySql^{TM};
    - **edliugo2kegg.prg** – processa o arquivo de retorno (**edliugo_kegg.txt**) do "KEGG: Kyoto Encyclopedia of Genes and Genomes" (<`http://www.genome.jp/kegg`>) inserindo marcadores KEGG no banco de dados;
    - **test_insert.prg** – inserção de um teste novo ou uma nova iteração para o *pipeline*, passando dados do conjunto de genes-sementes;
    - **edliunrm_process.prg** – processamento do teste para todos os genes do banco de dados;

    - **test2dot.prg** – a partir de um arquivo contendo todas as iterações dos testes (por exemplo: **teste1_go.txt**), gera um arquivo ".dot" que serve de entrada para o pacote *graphviz* (<`http://www.graphviz.org`>);
    - **map2html.prg** – a partir das saídas ".imap" (por exemplo: **teste1_go.imap**) e ".gif" (por exemplo: **teste1_go.gif**) do pacote *graphviz*, gera o arquivo *html* (por exemplo: **teste1_go.html**) contendo o grafo resultante;
  - **source_web** – programas fontes de análise desnvolvidos para a *WEB* em em Macromedia ColdFusion™:
    - **edliu.cfm** – para analisar o banco de dados resultante;
    - **edliu_res.cfm** – gera a apresentação das tabelas de predições a partir do banco de dados;
    - **edliugo_plot.cfm** – gera a apresentação dos gráficos dos sinais de expressão;

- **adhesion** – arquivos *html* com genes-sementes (**genes_sementes.html**) e seus agrupamentos (**genes_sementes_agrupamento.html**); genes preditos (**genes_preditos.html**) e seus agrupamentos (**genes_peditos_agrupamento.html**); e genes de adesão celular obtidos (**genes_adesao.html**);

- **tese** – arquivos eletrônicos do texto deste trabalho, fontes *Latex* e imagens.

Apêndice E

Publicações e apresentações em eventos científicos

E.1 Eventos científicos

Os seguintes trabalhos foram publicados ou apresentados em eventos científicos durante o curso de doutorado:

"Biological System Representation By Dynamic Systems: the retinoic acid transcription control" Marcelo Ris, Junior Barrera, Helena Brentani. Pôster apresentado em: *ICoBiCoBi: 1st. International Conference on Bioinformatics and Computational Biology*, Maio de 2003, Riberão Preto, São Paulo, Brasil.

"Identification of Estrogen Transcription Control Network Using U-curve Algorithm" Marcelo Ris, Junior Barrera, Helena Brentani. Pôster apresentado em: *ICoBiCoBi II: 2nd. International Conference on Bioinformatics and Computational Biology*, Outubro de 2004, Angra dos Reis, Rio de Janeiro, Brasil.

"U-curve Search for Biological States Characterization and Genetic Network Design" Marcelo Ris, Junior Barrera, Helena Brentani. Publicação de resumo e apresentação oral em: *GENSIPS 2005: IEEE International Workshop on Genomic Signal Processing and Statistics 2005*, Maio de 2005, New Port, Rhode Island, USA.

"Representação de Sistemas Biológicos a Partir de Sistemas Dinâmicos: Controle da Transcrição a Partir do Estrógeno" Marcelo Ris, Junior Barrera, Helena Brentani. Publicação de resumo e apresentação oral em: *I Simpósio de Iniciação Científica e Pós-Graduação do IME-USP*, Outubro de 2005, São Paulo, São Paulo, Brasil.

"Dynamical System Modeling of the Estrogen Transcription Control Network" Marcelo Ris, Junior Barrera, Helena Brentani. Pôster apresentado em: *X-Meeting - 1st. International Conference of the AB3C* (Associação Brasileira de Bioinformática e Biologia Computacional) Outubro de 2005, Caxambu, Minas Gerais, Brasil.

"Dynamical System Modeling of the Estrogen Transcription Control Network" Marcelo Ris, Junior Barrera, Helena Brentani. Pôster apresentado em: *ISMB 2006 - 14th. Annual International Conference On Intelligent Systems For Molecular Biology* Agosto de 2006, Fortaleza, Ceará, Brasil.

"A branch-and-bound optimization algorithm for U-shaped cost functions on Boolean lattices" Marcelo Ris, Junior Barrera. Publicação de resumo e pôster apresentado em: *SSP 2007 - IEEE Statistical Signal Processing Workshop 2007* Agosto 2007, Madison, WI, USA.

"Dynamical System Modeling of the Estrogen Transcription Control Network" Marcelo Ris, Junior Barrera, Helena Brentani. Pôster apresentado em: *10th International Meeting of MGED* Setembro de 2007, Brisbane, Australia.

"A branch-and-bound optimization algorithm for U-shaped cost functions on Boolean lattices" Marcelo Ris, Junior Barrera. Publicação de resumo e pôster apresentado em: *ISMM 2007 - 8th. International Symposium on Mathematical Morphology* Outubro de 2007, Rio de Janeiro, Rio de Janeiro, Brasil.

"A New Algorithm Pipeline for Transcription Control Network Modeling" Marcelo Ris, Junior Barrera. Pôster apresentado em: *X-Meeting - 3rd. International Conference of the AB3C* (Associação Brasileira de Bioinformática e Biologia Computacional) Novembro de 2007, São Paulo, São Paulo, Brasil.

E.2 Arigos em revistas

"A branch-and-bound optimization algorithm for U-shaped cost functions on Boolean lattices" Marcelo Ris, Junior Barrera, David C. Martins Jr. Artigo submetido para: *IEEE Transactions Pattern Analysis and Machine Intelligence (PAMI)*

"A New Algorithm Pipeline to Construct Gene Networks" Marcelo Ris, Junior Barrera, Helena Brentani. Artigo submetido para: *BMC-Bioinformatics*

"Cellular Adhesion Evidence of Estrogen Regulation Network" Marcelo Ris, Junior Barrera, Helena Brentani. Artigo submetido para: *Genome Biology*

Índice Remissivo

Referências Bibliográficas

[1] N. Abramson. *Information Theory and Coding*. McGraw-Hill Book Co., Inc., New York, second edition, 1963.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[3] J. Barrera, R. M. Cesar-Jr, D. C. Martins-Jr, R. Z. N. Vencio, E. F. Merino, M. M. Yamamoto, F. G. Leonardi, C. A. B. Pereira, and H. A. del Portillo. *Constructing probabilistic genetic networks of Plasmodium falciparum from dynamical expression signals of the intraerythrocytic development cycle*, chapter 2, pages 11–26. Springer, 2006.

[4] Linda Björnström and Maria Sjöberg. Mechanisms of estrogen receptor signaling: Convergence of genomic and nongenomic actions on target genes. *Molecular Endocrinology*, 19(4):833–842, 2005.

[5] U. Braga-neto and E. R. Dougherty. Bolstered error estimation. *Pattern Recognition*, 37:1267–1281, 2004.

[6] D. R. Coman. Adhesiveness and stickiness: Two independent properties of the cell surface. *Cancer Research*, 1:1436–1438, 1961.

[7] K. R. Coser, J. Chesnes, J. Hur, S. Ray, K. J. Isselbacher, and T. Shioda. Global analysis of ligand sensitivity of estrogen inducible and suppressible genes in mcf7/bus breast cancer cells by dna microarray. *PNAS*, 100(24):13994–9, 2003.

[8] E. R. Dougherty, S. Kim, and Y. Chen. Coefficient of determination in nonlinear signal processing. *Signal Processing*, 80:2219–2235, 2000.

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*, volume 1, pages 1–19. Wiley-Interscience, 2nd edition, 2000.

[10] D. Endy and R. Brent. Modelling cellular behaviour. *Nature*, 409:391–395, 2001.

[11] J. I. Fidler. Origin and biology of cancer metastasis. *Cytometry*, 10:673–680, 1989.

[12] A. Frank, D. Geiger, and Z. Yakhini. A distance-based branch and bound feature selection algorithm. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 241–248, San Francisco, CA, 2003. Morgan Kaufmann Publishers.

[13] J. Frasor, J. M. Danes, B. Komm, K. C. N. Chang, C. R. Lyttle, and B. S. Katzenellenbogen. Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: Insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype. *Endocrinology*, 144(10):4562–4574, 2003.

[14] C. Förster, S. Mäkela, A. Wärri, Silke Kietz, David Becker, Kjell Hultenby, Margaret Warner, and J. Gustafsson. Involvement of estrogen receptor β in terminal differentiation of mammary gland epithelium. *PNAS*, 99(24):15578–15583, 2002.

[15] G. Hardiman. Microarray platforms - comparison and contrasts. *Pharmacogenics*, 5:487–502, 2004.

[16] R. F. Hashimoto, E. R. Dougherty, M. Brun, Z. Z. Zhou, M. L. Bittner, and J. M. Trent. Efficient selection of feature sets possessing high coefficients of determination based on incremental determinations. *Signal Processing Special issue: Genomic signal processing*, 83(4):695–712, 2003.

[17] R. F. Hashimoto1, S. Kim, I. Shmulevich, W. Zhang, M. L. Bittner, and E. R. Dougherty. Growing genetic regulatory networks from seed genes. *Bioinformatics*, 20:1241–1247, 2004.

[18] R. O. Hynes. Integrins: Versatility, modulation, and signaling in cell adhesion. *Cell*, 69:11–25, 1992.

[19] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[20] V. X. Jin, Y. W. Leu, S. Liyanarachchi, H. Sun, M. Fan, K. P. Nephew, T. H. Huang, and R. V. Davuluri. Identifying estrogen receptor α target genes using integrated computational genomics and chromatin immunoprecipitation microarray. *Nucleic Acids Research*, 32(22):6627–35, 2004.

[21] H. De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9:67–103, 2002.

[22] D. C. Martins Jr, R. M. Cesar Jr, and J. Barrera. W-operator window design by minimization of mean conditional entropy. *Pattern Analysis & Applications*, 9:139–153, 2006.

[23] S. A. Kauffman. *The Origins of Order, Self-Organization and Selection in Evolution*, pages 441–520. Oxford University Press, New York, 1st edition, 1993.

[24] W. P. Kuo, T. K. Jenssen, A. J. Butte, L. O. Machado, and I. S. Kohane. Analysis of matched mrna measurements from two different microarray technologies. *Bioinformatics*, 18:405–412, 2002.

[25] A. S. Levenson, I. L. Kliakhandler, K. M. Svoboda, K. M. Pease, S. A. Kaiser, J. E. Ward III, and V. C. Jordan. Molecular classification of selective oestrogen receptor modulators on the basis of gene expression profiles of breast cancer cells expressing oestrogen receptor α. *British Journal of Cancer*, 87:449–456, 2002.

[26] C. Lin, A. Ström, V. B. Vega, S. L. Kong, A. L. Yeo, J. S. Thomsen, W. C. Chan, B. Doray, D. K. Bangarusamy, A. Ramasamy, L. A. Vergara, S. Tang, A. Chong, V. B. Bajic, L. D. Miller, J. Gustafsson, and E. T. Liu. Discovery of estrogen receptor α target genes and response elements in breast tumor cells. *Genome Biology*, 5(9):1–18, 2004.

[27] S. Nakariyakul and D. P. Casasent. Adaptive branch & bound algorithm for selecting optimal features. *Pattern Recognition Letters*, 28:1415–1427, 2007.

[28] S. Nilsson, S. Mäkelä, E. Treuter, M. Tujague, J. Thomsen, G. Andersson, E. Enmark, K. Pettersson, M. Warner, and J. Gustafsson. Mechanisms of estrogen action. *Physiological Reviews*, 81(4):1535–1565, 2001.

[29] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27:29–34, 1999.

[30] H. Oka, H. Shiozaki, K. Kobayashi, M. Inoue, H. Tahara, T. Kobayashi, Y. Takatsuka, N. Matsuyoshi, S. Mirano, M. Takeichi, and T. Mori. Expression of e-cadherin cell adhesion molecules in human breast cancer tissues and its relationship to metastasis. *Cancer Research*, 53:1696–1701, 1993.

[31] S. Peddada, E. Lobenhofer, L. Li, C. Afshari, C. Weinberg, and D. M. Umbach. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Science*, 19:834–841, 2003.

[32] P. Pudil, J. Novovicová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.

[33] M. Ris and J. Barrera. A branch-and-bound optimization algorithm for u-shaped cost functions on boolean lattices. In *Proceedings of the 8th International Symposium on Mathematical Morphology*, volume 2, pages 55–56, Rio de Janeiro, RJ, 2007.

[34] M. Schena. *DNA microarrays: A Practical Approach*. Oxford University Press, 1st edition, 1999.

[35] M. Schena, D. Shalon, R. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.

[36] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks: a rule-based uncertainty. *Bioinformatics*, 18(2):261–274, 2002.

[37] I. Shmulevich, E. R. Dougherty, and W. Zhang. Gene perturbation and intervention in probabilistic boolean networks. *Bioinformatics*, 18(2):1319–1331, 2002.

[38] C. Sima, U. Braga-Neto, and E. R. Dougherty. Superior feature-set ranking for small samples using bolstered error estimation. *Bioinformatics*, 21(7):1046–1054, 2005.

[39] P. Somol and P. Pudil. Fast branch & bound algorithms for optimal feature selection. *PAMI*, 26(7):900–912, July 2004.

[40] Z. Wang, J. Yang, and G. Li. An improved branch & bound algorithm in feature selection. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing: 9th International Conference*, Lecture Notes in Computer Science, pages 549–556, Chongqing, China, May 2003. Springer Berlin / Heidelberg.

[41] A. Weisz, W. Basile, C. Scafoglio, L. Altucci, F. Bresciani, A. Facchiano, P. Sismondi, L. Cicatiello, and M. Bortoli. Molecular identification of erα-positive breast cancer cells by the expression profile of an intrinsic set of estrogen regulated genes. *Journal Cellular Physiology*, 200(3):440–450, 2004.

[42] M. L. Whitfield, L. K. George, G. D. Grant, and C. M. Perou. Common markers of proliferation. *Nature Reviews Genetics*, 6:99–106, 2006.

[43] S. Yang and P. Shi. Bidirectional automated branch and bound algorithm for feature selection. *Journal of Shanghai University*, 9(3):244–248, 2005.

[44] X. Zhou, X. Wang, and E. R. Dougherty. Construction of genomic networks using mutual-information clustering and reversible-jump markov-chain-monte-carlo predictor design. *Signal Processing*, 83:745–761, 2003.

[45] B. T. Zhu and A. H. Conney. Functional role of estrogen metabolism in target cells: review and perspectives. *Carcinogenesis*, 19(1):1–27, 1998.