UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Pedro Rafael Costa

Determinação de genes mórbidos e drogáveis a partir da construção e análise da rede integrada de interações moleculares entre genes humanos

Botucatu – SP

2010

Pedro Rafael Costa

Determinação de genes mórbidos e drogáveis a partir da construção e análise da rede integrada de interações moleculares entre genes humanos

Monografia apresentada ao Instituto de Biociências da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu, para a obtenção do título de Bacharel em Física Médica.

Orientador:
Prof. Dr. Ney Lemke

Bacharelado em Física Médica
Departamento de Física e Biofísica
Instituto de Biociências
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Campus de Botucatu

Botucatu – SP

2010

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: *ROSEMEIRE APARECIDA VICENTE*

Costa, Pedro Rafael.
Determinação de genes mórbidos e drogáveis a partir da construção e análise da rede integrada de interações moleculares entre genes humanos / Pedro Rafael Costa. – Botucatu, 2010.

Trabalho de conclusão de curso (bacharelado – Física Médica) – Universidade Estadual Paulista, Instituto de Biociências de Botucatu.
Orientador: Ney Lemke
Assunto CAPES: 20000006

1. Física médica. 2. Genes.

Palavras-chave: Aprendizagem de máquina; Genes drogáveis; Genes mórbidos; Redes biológicas.

À meus pais, Cláudio e Roselene Costa,
exemplos de honestidade e esforço,
que tornaram possível esta conquista.

Agradecimentos

Meus mais sinceros agradecimentos à todos que me ajudaram na elaboração desse trabalho:

– Ao Professor Doutor Ney Lemke, pela orientação e incentivo;

– À equipe do Laboratório de Bioinformática e Biofísica Computacional do Departamento de Física e Biofísica do IBB-Unesp e agregados de Laboratórios vizinhos, em especial ao colega doutorando e "co-orientador" Marcio Luis Acencio, pela ajuda e incentivo para sempre fazer o melhor possível e ao colega Carlos Alexandre Henrique Fernandes, vulgo "Pituta", pela disposição em tirar dúvidas e curiosidades;

– À meus pais e à minha namorada Elaine Cristina Galhardo, pelo carinho, paciência e disposição para a revisão desta monografia;

– À todos os colegas da IV Turma de Física Médica da Unesp de Botucatu, com os quais dividi momentos de alegria e de desespero durante os 4 anos de graduação, e que foram capazes de me suportar durante todo esse tempo.

– À Fundação de Amparo à Pesquisa do Estado de São Paulo, pelo apoio recebido durante os dois anos de trabalho;

"'Would you tell me, please, which way I ought to go from here?'
'That depends a good deal on where you want to get to,' said the Cat."
Lewis Carroll, Alice's Adventures in Wonderland

Resumo

O processo de descoberta de novos genes mórbidos e de novas proteínas alvo para drogas é atualmente muito custoso e laborioso. Visando diminuir os custos e acelerar esse processo, propomos, neste trabalho, um método *in silico* para determinação do *grau de drogabilidade* e do *grau de morbidade* de um gene, medidas da probabilidade da proteína codificada pelo dado gene possuir características que a tornariam alvo para novas drogas e da probabilidade de em caso de ocorrência de alteração do dado gene, o fenótipo observado caracterizar uma doença com base genética. Para determinar essas características, construímos, analisamos e determinamos os dados da topologia da rede integrada de interações moleculares entre genes humanos, contendo interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional, e incluímos outros dados, como nível de transcrição gênica e localização celular da proteína codificada pelo gene. Os índices de acerto obtidos para os conjuntos de teste foram iguais ou superiores aos obtidos por métodos semelhantes encontrados na literatura. Finalmente, buscamos na literatura biomédica evidências de que os genes classificados com os 10 maiores graus de drogabilidade e morbidade, excluídos os já conhecidamente drogáveis/mórbidos, possuíam potencial para tal característica, encontrando-as em 73% e 90% dos casos, respectivamente.

Palavras-chave: Redes Biológicas, Aprendizagem de Máquina, Genes Drogáveis, Genes Mórbidos.

Abstract

The discovering process of new morbid genes and new target proteins for drugs have been shown to be very costly and laborious. Having in view cutting costs and speeding up this process, we propose, in this work, a new method to determine the gene *druggability score* and *morbidity score*, the probabilities of the protein encoded by the gene have the characteristics that make it a new target for drugs and in case of an alteration in that gene, we observed a phenotype that characterizes a genetic based illness. To determine these characteristics, we built, analyzed and determined the characteristics of the topology of the integrated molecular interactions network among human genes containing physical interactions between proteins, metabolic interactions and interactions of transcriptional regulation, and included other data such as level of gene transcription and cellular localization of the protein encoded by the gene. We tested our model in training sets and achieved results equal or better than the ones achieved by similar methods in the literature. Finally, with the purpose of investigating whether the assigned scores resembles the potential druggabilities and morbities of the previously unclassified genes, we looked for evidences in biomedical literature supporting the potential druggability and morbidity status of genes with the 10 highest scores. We found clear evidences for 73% and 90% of potential druggable and morbid genes respectively.

Key Words: Biological Networks, Machine Learning, Druggable Genes, Morbity Genes.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Biologia de Sistemas

O método reducionista de dissecação dos sistemas biológicos em suas partes constituintes tem ajudado no esclarecimento do funcionamento de muitos processos biológicos. Porém, tais processos são extremamente complexos e possuem propriedades emergentes que não podem ser explicadas ou mesmo previstas através do estudo de suas partes individuais (REGENMORTEL, 2004). Tal limitação imposta pelo método reducionista é um dos fatores preponderantes na falta de uma melhor compreensão e desenvolvimento de terapias eficazes para doenças complexas (AHN et al., 2006).

Para suplantar esses limites do reducionismo, pesquisadores têm usado um conjunto de métodos que tratam os processos biológicos de forma integrada. Esta nova área da biologia é conhecida como biologia de sistemas. A biologia de sistemas objetiva a compreensão das interações não-lineares entre os múltiplos componentes dos processos biológicos (AHN et al., 2006). Tais interações são geralmente representadas por um objeto matemático chamado *grafo* ou *rede* (BOLLOBÁS, 1979), dado por um conjunto de nodos G (componentes) e por um conjunto de arestas A (interações) que conectam cada dois componentes no conjunto G. As arestas de um grafo podem ser direcionadas, indicando uma fonte (ponto de partida) e um alvo (ponto terminal), ou não direcionadas.

O trabalho de Jeong et. al. (JEONG et al., 2000) foi o pioneiro em utilizar grafos para representação de sistemas biológicos (no caso, vias metabólicas de diversos organismos), analisar suas propriedades e interpretá-las biologicamente. Esse método vem sendo utilizado com maior freqüência para elucidação de propriedades emergentes de processos biológicos complexos (PERROUD et al., 2006; RHODES et al., 2005), devido a capacidade computacional e grande quantidade de dados disponíveis.

1.1.1 Características Topológicas dos Grafos

Existem várias características topológicas que podem ser obtidas através da análise da distribuição das arestas de determinado nodo ou nodos. Todas possuem um caráter abstrato, e seu significado varia de acordo com o que o grafo representa. As principais estão citadas abaixo:

- **Grau ou Conectividade, $k(g)$**: Número de arestas que determinado nodo g possui. Caso o grafo seja direcionado, teremos um *grau de entradas,* $k_{in}(g)$, para o número de arestas direcionadas para g, e um *grau de saídas,* $k_{out}(g)$, para o número de arestas que partem do nodo g.

- **Coeficiente de agregação, $c(g)$**: sua definição está relacionada com os ciclos de comprimento três (triângulos de arestas). É dado pela fórmula

$$c(g) = \frac{2n(g)}{k(g)[k(g)-1]} \tag{1.1}$$

  $n(g)$ é o número total de arestas que os vizinhos de g possuem, e $k(g)$ seu grau (WATTS; STROGATZ, 1998).

- **Número de caminhos mais curtos, $\sigma_{g_i g_j}$**: Número de conjuntos de nodos capazes de interligar dois nodos distintos g_i e g_j utilizando o menor número de nodos possível.

- **Grau de intermediação, $inbet(g)$**: relação entre o número de conjuntos de caminhos mais curtos que a rede possui e o número desses conjuntos que determinado nodo g pertence, ou seja:

$$inbet(g) = \sum_{g_i \neq g \neq g_j} \frac{\sigma_{g_i g_j}(g)}{\sigma_{g_i g_j}} \tag{1.2}$$

  sendo que $\sigma_{g_i g_j}$ é o número de caminhos mais curtos entre os nodos g_i e g_j e $\sigma_{g_i g_j}(g)$ é o número de caminhos mais curtos entre g_i e g_j que passam por g (ANTHONISSE, 1971; FREEMAN, 1977).

- **Grau de proximidade, $cent(g)$**: Dada pela fórmula

$$cent(g) = \frac{n}{\sum_{g_j} d(g, g_j)} \tag{1.3}$$

  onde $d(g, g_j)$ é a menor distância, em número de arestas, entre os genes g e g_j e n é o número de nodos presentes na rede (SABIDUSSI, 1966).

- **Número de Sósias, *ident*(g)**: Número de nodos que possuem exatamente as mesmas características topológicas de interesse do dado nodo g.

1.2 Genes Humanos

Inicialmente tido como uma entidade abstrata por Gregor Mendel em meados do século XIX, o *gene* passou a ser definido como sendo a unidade funcional do DNA, ou seja, uma sequência específica de ácidos nucleicos, que constituem uma região que pode ser transcrita em RNA e uma outra que determina sua regulação (GRIFFITHS et al., 2001). Em seres humanos, encontramos 46 cromossomos, sendo 22 pares de autossomos e 2 cromossomos sexuais, que comportam quase 3 bilhões de pares de bases de DNA, contendo em torno de 25.000 genes codificadores de proteínas (STUMPF et al., 2008).

Cada gene possui uma série de características individuais que são responsáveis por boa parte do fenótipo observado nos mais diferentes organismos. O estudo dessas características acarretam em grande desenvolvimento tanto na área da saúde, com a descoberta das raízes de várias doenças e com o desenvolvimento de drogas para combatê-las, como na área de melhoramento genético, onde, por exemplo, controla-se a regulação de determinado gene para obter um novo fenótipo. Algumas dessas características são detalhadas nos tópicos a seguir.

1.2.1 Interações Moleculares entre Genes

As proteínas codificadas pelos genes podem interagir com outras proteínas. Essas interações merecem destaque pois a alteração de um determinado gene pode acarretar numa cascata de acontecimentos que podem gerar um novo fenótipo, desejável ou indesejável. Exitem vários tipos de interações moleculares entre genes, entre elas:

- **Interações proteína-proteína**: interações físicas entre duas proteínas codificadas pelos respectivos genes;
- **Interações metabólicas**: ocorrem quando temos uma reação metabólica catalisada por uma enzima codificada por um determinado gene. O reagente pode ser produto de um outro gene, ou o produto dessa reação pode ser utilizado por outra proteína;
- **Interações regulatórias entre fatores de transcrição**: O gene pode ser regulado por um fator transcricional, produto de um outro gene, ou seu produto pode ser um fator transcricional para outro gene.

Organizando os dados a respeito dessas interações em uma única *rede*, obtemos uma *rede integrada* de interações entre os genes de determinado organismo, e podemos analisar suas propriedades afim obter informações para cada um dos genes constituintes.

1.2.2 Nível de Expressão Gênica em Diferentes Tecidos e Localização Celular das Proteínas Codificadas pelos Genes

Organismos pluricelulares possuem basicamente o mesmo material genético em suas células constituintes. Os diferentes níveis de regulação gênica são os maiores responsáveis pela diferenciação dos tecidos, pois acarretam em diferentes quantidades de proteínas naquela célula. Com exceção dos chamados genes *housekeeping*, expressos em todas as células no mínimo em níveis basais, os demais possuem um nível específico de expressão, utilizando uma grande gama de ferramentas para tal. Esse nível de expressão está diretamente ligado à função da célula no momento, e pode variar dependendo dos estímulos externos.

As proteínas traduzidas nos ribossomos presentes no citoplasma celular ou aderidos no retículo endoplasmático podem ser recrutadas para cumprirem seus papéis em qualquer região intracelular, ou mesmo extracelular. Podemos separá-las utilizando esse conceito em proteínas *citoplasmáticas*, *nucleares*, *membranares*, *organelares* e *extracelulares*.

1.2.3 Genes Mórbidos e Genes Drogáveis

Quando uma alteração em determinado gene é necessária para a manifestação de uma doença, dizemos que esse gene é *mórbido*. Mais de 4500 doenças catalogadas possuem base genética e pelo menos um gene responsável pelo fenótipo observado (AMBERGER et al., 2009). A descoberta de novos genes mórbidos requer um laborioso estudo do padrão de hereditariedade de determinada doença em diversas famílias e a análise do genótipo desses indivíduos para identificar o gene candidato envolvido (PRASAD et al., 2009).

Genes *drogáveis* são aqueles que codificam proteínas que quando reguladas por compostos químicos incitam determinado fenótipo desejável. De modo similar à descoberta de genes mórbidos, descobrir novos alvos para drogas requer um grande esforço envolvendo genômica, proteômica, associações genéticas e técnicas de genética direta e inversa (LINDSAY, 2003), e assim produzir drogas capazes de modular o desenvolvimento de determinada doença.

1.3 Mineração de Dados

Trata-se do processo de exploração de grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados (WITTEN; FRANK, 2000). Esses padrões são obtidos geralmente a partir de amostras dos dados, portanto a verificação e validação dos padrões obtidos em outras amostras de dados é extremamente importante.

Para se obter as regras de associação para um grande quantidades de dados, geralmente são utilizados *algoritmos de aprendizagem de máquina*, programas que melhoram seu desempenho por meio de experiência. São capazes de gerar hipóteses a partir dos dados, identificando padrões complexos que maximizam o índice de acerto da mineração.

Artigos que utilizam mineração de dados para elucidação de um problema são encontrados em grande quantidade e variedade na literatura. Com escopo biológico, essa ferramenta já foi utilizada com sucesso para determinação de essencialidade de um gene em *Escherichia coli* (SILVA et al., 2008) e em *Saccharomyces cerevisiae* (ACENCIO; LEMKE, 2009).

1.3.1 Definições em Mineração de Dados

Alguns termos são utilizados com frequência durante a mineração de dados, e faz-se necessário definí-los corretamente:

- **Instância**: Objeto a ser classificado, independente do conceito a ser aprendido;
- **Atributos**: Características que descrevem determinado conjunto de instâncias. Quando várias instâncias apresentam determinado atributo com mesmo valor, dizemos que tais instâncias pertencem à mesma *Classe* para aquele atributo;
- **Dado**: Sequência de símbolos quantificados ou quantificáveis, para determinado atributo. Determina a *classificação* da respectiva instância;
- **Treino**: Etapa na qual o algoritmo busca as regras de associação entre os dados disponibilizados;
- **Modelo**: Conjunto de regras que buscam determinar corretamente a classificação de determinada instância;

- **Verdadeiros Positivos (*Vp*)**: Instâncias corretamente classificadas como pertencentes a determinada classe;
- **Verdadeiros Negativos (*Vn*)**:Instâncias corretamente classificadas como não pertencentes a determinada classe;
- **Falsos Positivos (*Fp*)**: Instâncias erroneamente classificadas como pertencentes a determinada classe;
- **Falsos Negativos (*Fn*)**: Instâncias erroneamente classificadas como não pertencentes a determinada classe.

1.3.2 Estatísticas Utilizadas na Mineração de Dados

1.3.2.1 Valores de Desempenho do Classificador

Os algoritmos de mineração de dados retornam valores que representam o desempenho da classificação. Para um resultado robusto, devem visar o equilíbrio entre essas medidas. As de maior representatividade são:

- **Precisão**: dada pela soma dos verdadeiros positivos obtidos para todas as classes dividida pela soma de todos os verdadeiros positivos e falsos positivos;
- **Recall**: razão entre o número de verdadeiros positivos de determinada classe e número total de exemplos daquela classe;
- **ASC – Área sob a curva ROC (*"Receiver operating characteristic"*)**: A curva ROC plota a fração de verdadeiros positivos pela fração de falsos positivos, sendo que área abaixo dessa curva é numericamente igual a probabilidade de uma determinada instância ser corretamente classificada.

1.3.2.2 Processo de Validação Cruzada

Como costuma-se trabalhar com amostras, é interessante dispormos de ferramentas que possam estatisticamente validar os valores obtidos e os modelos gerados. O método de validação cruzada por k vezes consiste no particionamento aleatório da amostra em k subamostras (geralmente, $k = 10$). Uma única subamostra é separada para o teste do modelo, enquanto as restantes são utilizadas para o treino. O processo é repetido até que as k subamostras tenham sido utilizadas para teste. Os valores de precisão, *recall* e ASC que o algoritmo retorna é a média obtida para os k testes (PICARD; COOK, 1984).

1.3.2.3 Índice Kappa – κ

Para determinar o quanto o modelo obtido se diferencia de um modelo com regras aleatórias, pode-se calcular o respectivo índice estatístico kappa de Cohen's (κ) (COHEN, 1960), que varia de -1 a 1. Para $\kappa = 1$, temos que o modelo classificou perfeitamente todas as instâncias; valores entre 0 e 1 indicam que o algoritmo encontrou relações com desempenho superior ao modelo aleatório;se $\kappa = 0$, obteve-se o mesmo resultado do modelo aleatório; e se $\kappa < 0$, o desempenho obtido é pior que o obtido aleatoriamente.
O valor de κ é obtido pela seguinte fórmula:

$$\kappa = \frac{\sum_{i=1}^{n} Vp_i - \sum_{i=1}^{n} (Fn_i + Vp_i)(Fp_i + Vp_i)}{T - \sum_{i=1}^{n} (Fn_i + Vp_i)(Fp_i + Vp_i)} \tag{1.4}$$

onde n é o número de classificações possíveis e T o número total de instâncias classificadas. Vp_i, Fn_i e Fp_i, representam respectivamente os verdadeiros positivos, os falsos negativos e os falsos positivos obtidos para cada classificação i.

1.3.2.4 Teste de Wilcoxon – W

O *teste de postos com sinais de Wilcoxon* (WILCOXON, 1947) é utilizado para comparação entre resultados de dois modelos e possui prestígio dentro da comunidade de aprendizagem de máquina. Esse teste não-paramétrico é recomendado para qualquer tipo de distribuição de dados. A hipótese nula diz que os resultados obtidos pelos dois classificadores são iguais.

Pareados os resultados obtidos para cada conjunto de dados i, calcula-se d_i como sendo a diferença entre esses desempenhos. Os valores d_i são ordenados de forma crescente de acordo com seu módulo, e recebem um valor $r(d_i)$ igual a sua colocação na lista ordenada. Caso existam dois ou mais valores iguais, o valor considerado para $r(d_i)$ desses termos passa a ser a média entre as colocações que os termos ocupam. Se existir um número ímpar de $d_i = 0$, ignora-se um dos respectivos valores de $r(d_i)$. Calcula-se então, R^+ e R^-, dados pelas seguintes fórmulas:

$$R^+ = \sum_{d_i>0} r(d_i) + \frac{1}{2}\sum_{d_i=0} r(d_i) \qquad R^- = \sum_{d_i<0} r(d_i) + \frac{1}{2}\sum_{d_i=0} r(d_i) \tag{1.5}$$

Determinamos, então, o valor de W, dado por $W = \min(R^+, R^-)$. Se possuirmos mais de 15 diferenças, excluindo um termo caso o número de $d_i = 0$ seja ímpar, segue-se o teste

de hipótese nula calculando o valore de z, dado por:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)N(2N+1)}} \tag{1.6}$$

onde N é o número final de diferenças utilizadas. Se considerarmos $\alpha = 0.05$ como probabilidade da hipótese nula ser verdadeira, podemos descartá-la $z < -1,96$ (DEMSAR, 2006).

Caso $N < 25$, costuma-se comparar o valor de W com os valores críticos para determinado α, W_c, propostos por Wilcoxon em seu artigo (WILCOXON, 1947). Se $W \leq W_c$, a hipótese nula pode ser rejeitada. Alguns desses valores estão apresentados na Tabela 1.

Tabela 1: Valores de W_c dependendo do número de diferenças utilizadas N, com com probabilidade $\alpha < 0,05$ e $\alpha < 0,01$ para hipótese nula.

$\boldsymbol{N}$	$\boldsymbol{W_c}$ $\boldsymbol{\alpha < 0,05}$	$\boldsymbol{W_c}$ $\boldsymbol{\alpha < 0,01}$
8	3	0
9	5	1
10	8	3
11	10	5
12	13	7

2 Objetivos

2.1 Organização do Banco de Dados de Genes Humanos

1. Construção da rede integrada de interações moleculares entre genes humanos, contendo interações físicas entre as proteínas codificadas por esses genes, interações metabólicas entre enzimas codificadas por esses genes e interações regulatórias entre fatores de transcrição e seus respectivos genes alvos;

2. Determinação dos valores das propriedades topológicas para cada gene na rede obtida, e organização dos resultados em forma de banco de dados;

3. Inclusão de dados sobre a expressão nos diferentes tecidos, localização celular das proteínas codificadas, morbidade e drogabilidade para cada um dos genes.

2.2 Mineração dos Dados

1. Utilização de ferramentas para mineração de dados, visando encontrar relações capazes de identificar tanto os genes mórbidos, como os genes drogáveis presentes na rede, determinando seu desempenho, validando-o e comparando-o com resultados obtidos através de métodos semelhantes na literatura;

2. Idem, mas para identificação dos genes drogáveis presentes na rede.

2.3 Aplicação dos Modelos

1. Aplicação dos modelos gerados, tanto para morbidade, quanto para drogabilidade, no banco de dados sem a respectiva classificação, obtendo, dessa forma, uma *probabilidade* do gene possuir a determinada característica, que denominaremos de *Grau de Morbidade* e de *Grau de Drogabilidade* de um gene.

2. Busca na literatura por evidências experimentais que confirmem que os genes que apresentam maiores *graus de morbidade* e de *drogabilidade* possam realmente possuir a respectiva característica.

3 Métodos

3.1 Organização do Banco de Dados de Genes Humanos

3.1.1 Construção da rede integrada de interações moleculares de genes humanos

A rede integrada de interações moleculares foi definida por um grafo, onde os nodos representam os genes e suas arestas podem representar três tipos interações entre eles:

- Interações físicas entre proteínas codificadas pelos genes correspondentes;
- Interações metabólicas entre as enzimas codificadas pelos genes correspondentes;
- Interações regulatórias entre fatores de transcrição e seus genes alvos.

Todas as arestas do grafo são direcionadas. As interações físicas entre proteínas foram representadas utilizando sempre duas arestas entre os nodos, uma para cada sentido. Para as interações regulatórias, o sentido das arestas indicará o controle de um fator de transcrição sobre o gene indicado. Para as interações metabólicas, o sentido das arestas indicará o fluxo de metabólitos entre as enzimas.

As interações moleculares da rede integrada foram obtidas a partir dos seguintes bancos de dados:

- Interações físicas entre as proteínas: BIOGRID (BREITKREUTZ et al., 2008), DIP (XENARIOS et al., 2002), HPRD (PRASAD et al., 2009), INTACT (HERMJAKOB et al., 2004), MINT (CHATR-ARYAMONTRI et al., 2007) e MIPS (PAGEL et al., 2005);
- Interações metabólicas entre enzimas: BiGG (DUARTE et al., 2007). Para humanos, a rede metabólica reconstruída provém do trabalho de (DUARTE et al., 2007);

- Interações regulatórias entre fatores de transcrição e seus genes alvos: TRED (JIANG et al., 2007);

As interações presentes em cada um dos bancos de dados acima foram baixadas e gravadas localmente. Todos os nomes dos genes foram convertidos para seu número identificador único, GeneID, fornecido pelo banco de dados Entrez Gene (WHEELER et al., 2008). Todos os arquivos foram modificados para um formato de texto tabulado com duas colunas, onde cada linha representa uma interação e cada coluna contém a lista de interagentes, isto é, os genes. Na construção da rede de interações metabólicas, as interações foram determinadas com base no fluxo de metabólitos entre duas enzimas. Todos os metabólitos e as interações geradas a partir de metabólitos muito comuns, como H_2O, H^+, ATP, ADP e NAD, foram retirados da rede. As três redes individuais de interações foram integradas pelos GeneIDs dos genes humanos. Essa integração foi realizada utilizando-se o pacote Combinatorica (SRIRAM; STEVE, 2003) para o programa Mathematica (Wolfram Research. Inc.). Nossa rede final dispunha de mais de 10 mil genes e mais de 70 mil interações.

3.1.2 Determinação das propriedades topológicas da rede integrada

As propriedades topológicas abaixo apresentadas foram calculadas utilizando-se o pacote Combinatorica (SRIRAM; STEVE, 2003) para o programa Mathematica 7.0 (Wolfram Research. Inc.) e o pacote NetworkX 0.99 para a linguagem Python.

- **Grau relacionado com interações físicas entre proteínas, $ppi(g)$**: Número de ligações não direcionadas presentes gene g, representando o número de interações físicas entre proteínas.

- **Grau de entradas relacionadas com interações metabólicas, $met_{in}(g)$**: Número de ligações direcionadas para o gene g, representando número de reagentes que participam numa reação metabólica catalisada por uma enzima codificada por g.

- **Grau de saídas relacionadas com interações metabólicas, $met_{out}(g)$**: Número de ligações direcionadas que partem do gene g, representando número de produtos gerados por uma reação metabólica catalisada por uma enzima codificada por g.

- **Grau de entradas relacionadas com interações transcricionais regulatórias, $reg_{in}(g)$**: Número de ligações direcionadas para o gene g, representando o número de fatores transcricionais que regulam g.

- **Grau de saídas relacionadas com interações transcricionais regulatórias, $reg_{out}(g)$**: Número de ligações direcionadas que partem do gene g, representando o número de genes regulados pelo fator transcricional codificado por g.

- **Grau total, $K(g)$**: Dado pela soma dos graus $ppi(g)$, $met_{in}(g)$, $met_{out}(g)$, $reg_{in}(g)$, $reg_{out}(g)$.

- **Grau total de entradas, $K_{in}(g)$**: Dado pela soma dos graus $ppi(g)$, $met_{in}(g)$, $reg_{in}(g)$.

- **Grau total de saídas, $K_{out}(g)$**: Dado pela soma dos graus $ppi(g)$, $met_{out}(g)$, $reg_{out}(g)$.

- **Coeficiente de Agregação, $c(g)$**: Conforme explicado na seção 1.1.1, considerando os três tipos de interação para os cálculos.

- **Grau de intermediação**: Conforme explicado na seção 1.1.1, foram calculados 4 tipos de graus de intermediação:

  - $inbet(g)$: Considerando todos os tipos de interações.
  - $inbet_{ppi}(g)$: Considerando somente as interações físicas entre proteínas.
  - $inbet_{met}(g)$: Considerando somente as interações metabólicas.
  - $inbet_{reg}(g)$: Considerando somente as interações regulatórias transcricionais.

- **Grau de proximidade, $cent(g)$**: Conforme explicado na seção 1.1.1.

- **Número de Sósias, $ident(g)$**: Número de genes que possuem mesmo valor de ppi, $metin$, $metout$, $regin$, $regout$ que o gene g.

Para observar o comportamento da rede e classificá-la, calculamos a *distribuição do número de genes que possuem k ligações*, $P(k)$, equivalente a probabilidade de encontrarmos um nodo com k arestas em nosso grafo, considerando todos os todos os tipos de interação. Matematicamente, temos

$$P(k) = \frac{1}{N} \sum_{g=1}^{N} \delta_{k,K(g)} \tag{3.1}$$

onde N é o número total de nodos e $\delta_{k,K(g)}$ é uma *delta de Kronecker*, ou seja, $\delta_{k,K(g)} = 1$ se $k = K(g)$, e zero caso contrário. Calculamos $P_{in}(k)$, substituindo $K(g)$ por $K_{in}(g)$ e $P_{out}(k)$, utilizando $K_{out}(g)$. Calculamos também a *distribuição dos valores médios dos coeficientes de agregação de acordo com k, $C(k)$,* matematicamente dada por:

$$C(k) = \frac{1}{N_k} \sum_{g=1}^{N} \delta_{k,K(g)} \cdot c(g) \tag{3.2}$$

onde N_k é o número total de nodos que possuem k ligações.

3.1.3 Construção de um banco de dados com informações a respeito dos genes humanos

Utilizando o pacote DatabaseLink do software Mathematica, criamos um banco de dados para genes humanos contendo o GeneID, e as características calculadas na seção 3.1.2. Adicionamos ao banco os dados a respeito da morbidade, retirados do *Morbid Map* presente no OMIM (AMBERGER et al., 2009), e da drogabilidade dos genes (YILDIRIM et al., 2007), sendo que para morbidade calculamos a distância entre o gene e o gene mórbido mais próximo (zero caso o gene seja mórbido). Além disso, foram incluídos dados sobre o número de tecidos onde o gene possui um nível de expressão maior que 5 transcrições por milhão em média (tpm, 32 tecidos estudados), nível de expressão médio nesses tecidos, em tpm (REVERTER; INGHAM; DALRYMPLE, 2008) e dados sobre a localização celular da proteína codificada pelo gene, utilizando a ferramenta *QuickGO*, presente na página do *Gene Ontology, GO,* associado com o banco de dados integrado para famílias de proteínas, *InterProt*, do Instituto Europeu de Bioinformática (BINNS et al., 2009). Para isso selecionamos um conjunto generalizando os termos utilizados pelo GO, classificando a localização das proteínas de acordo com as seguintes opções: *Citoplasma*, *Retículo endoplasmático*, *Mitocôndria*, *Núcleo*, *Complexo de Golgi*, *Membrana plasmática* e *Espaço extracelular*. As proteínas ainda foram classificadas como *Outros locais*, quando localizada em locais diferentes dos citados, *Componente celular*, quando não determinada sua localização exata na célula, ou *Desconhecida*, caso o GO não possuísse os dados a respeito do respectivo gene.

3.2 Mineração dos dados

3.2.1 Criação dos conjuntos de treino

Como não temos um conjunto negativo, pois genes não presentes em (YILDIRIM et al., 2007) e no OMIM não podem ser classificados como não-drogáveis ou não-mórbidos, foram organizados 10 conjuntos para cada um dos objetivos. Cada conjunto é formado por 80% dos genes conhecidamente drogáveis/mórbidos aleatórios, e pelo mesmo número de genes com a respectiva característica desconhecida, escolhidos aleatoriamente, com todos os valores calculados anteriormente. Para o conjunto de drogáveis, retiramos os dados sobre morbidade, e vice-versa.

Para constatar a diferença entre nossa classificação e uma classificação totalmente aleatória, permutamos aleatoriamente entre os genes seus dados sobre drogabilidade e morbidade, e criamos mais 10 conjuntos para drogáveis/mórbidos, ou seja, 10 conjuntos permutados para cada objetivo.

3.2.2 Análise dos dados e validação dos resultados obtidos

De posse dos conjuntos, utilizamos o software *Weka* (*Waikato Environment for Knowledge Analysis*), um conjunto de algoritmos de aprendizado de máquina desenvolvido pela Universidade de Waikato, na Nova Zelândia, buscando obter informações para mineração de dados para determinação do *grau de drogabilidade* e de *morbidade* de um gene. Utilizamos o *método de validação cruzada por 10 vezes*, e uma combinação dos algoritmos de aprendizagem *REPtree*, *naive bayes tree*, *random tree, random forest*, *J48 (C4.5)*, *best-first decision tree*, *logistic model tree* e *alternating decision tree* através do meta-classificador *Vote* para geração dos modelos (WITTEN; FRANK, 2000). Além disso, aplicamos a técnica de bootstrap aggregating (bagging) para cada classificador. Para gerar a árvore de decisão, um método de visualização de como foi realizada a classificação das instâncias, foi utilizado o algoritmo J48.

A saída do programa retornou os valores de *Recall*, de *Área sob a curva ROC*, de *Precisão*, e o valor da estatística *kappa* obtidos para cada uma das listas teste. Calculamos então a média e o desvio padrão desses para os 10 conjuntos de drogáveis, para os 10 conjuntos de drogáveis permutados, para os 10 conjuntos de mórbidos e para os 10 conjuntos de mórbidos permutados.

Calculamos a média e o desvio padrão para cada um dos quatro tipos de modelos

gerados, e selecionamos os genes conhecidamente drogáveis/mórbidos para observarmos o comportamento da distribuição de probabilidades.

3.3 Aplicação dos modelos gerados

Aplicamos os 10 modelos de cada caso na lista com todos os genes da rede, mas sem a informação de interesse (substituída por uma interrogação). O WEKA, então, retorna para cada um dos genes a probabilidade do mesmo ser drogável ou mórbido a partir dos respectivos modelos, que denominamos aqui de **Grau de drogabilidade** e **Grau de morbidade** de um gene.

Para os resultados com os dados não permutados, selecionamos, após retirados das listas os genes conhecidamente drogáveis/mórbidos, os 10 genes que possuíam maior probabilidade de possuírem a respectiva característica, dada pela mediana dos dados (recomendado pois a distribuição desses dados pode não ser normal), e buscamos na literatura indícios de drogabilidade, considerando os genes claramente denominados como potenciais alvos para drogas na literatura biomédica, e de morbidade, considerando os genes que estão claramente associados a alguma doença no banco de dados HuGENnet (LIN et al., 2006). Nessa tabela, incluímos o valor de N e de W do teste de Wilcoxon, comparando se a probabilidade do resultado obtido ser aleatório é menor que 5%.

4 Resultados e Discussão

4.1 Rede Integrada de Interações Moleculares em Humanos

O grafo obtido para a rede integrada de humanos está representado na Figura 1, onde os nodos foram coloridos e dimensionados de acordo com o grau de intermediação: quanto maior seu valor, mais vermelho o nodo (vindo do amarelo) e maior. As interações regulatórias estão representadas em verde, as metabólicas em azul e as físicas em preto. Podemos observar a complexidade inerente à rede, e sem o uso de ferramentas computacionais, pouco podemos inferir sobre ela.

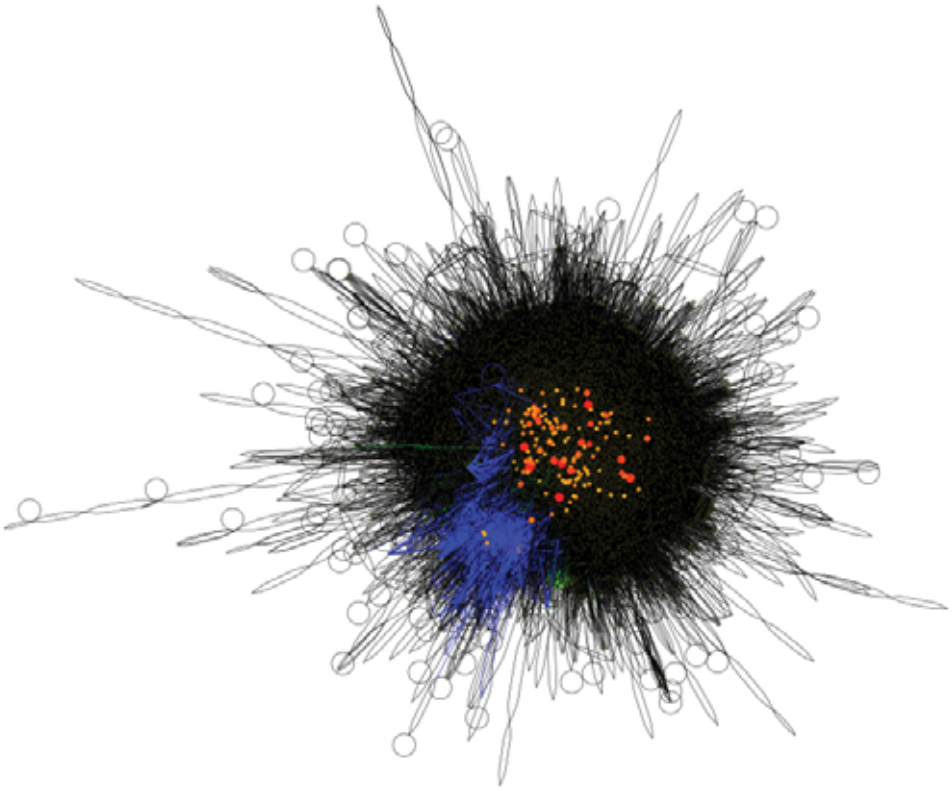

Figura 1: Grafo da rede integrada das interações moleculares em humanos

As distribuições $P(k)$ e $C(k)$ da rede integrada estão representadas nas Figuras 2 e 3. As equações das curvas que aproximam os gráficos de $P(k)$ foram:

$$P_{in}(k) = 1,3k^{-1,3} \text{ (em azul)} \qquad \text{e} \qquad P_{out}(k) = 1,3k^{-1.7} \text{ (em roxo).}$$

O comportamento das distribuições classifica a rede como sendo *livre de escala*, pois muitos genes participam de poucas interações e poucos genes participam de muitas interações. Biologicamente, isso torna o organismo mais resiliente a perturbações, pois a perda aleatória de um gene (mutação) acarreta em baixa probabilidade de inviabilidade do indivíduo (ALBERT; JEONG; BARABASI, 2000).

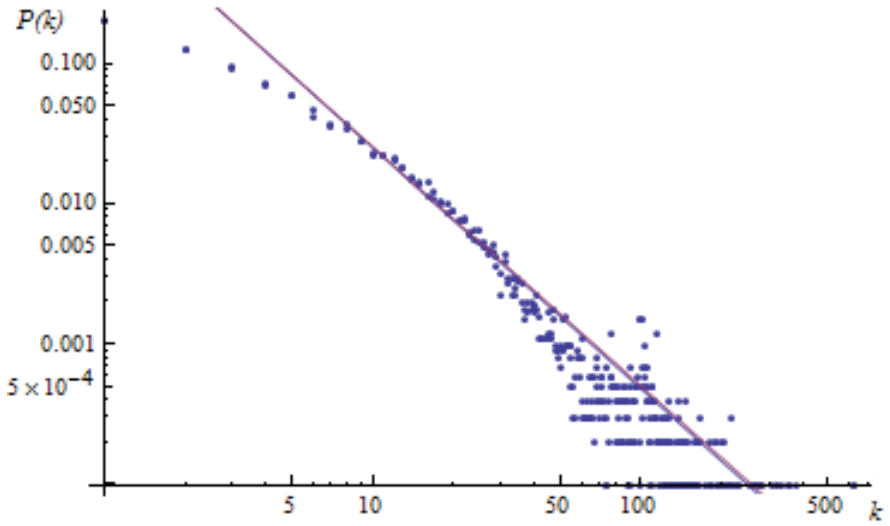

Figura 2: Gráfico da distribuição $P(k)$

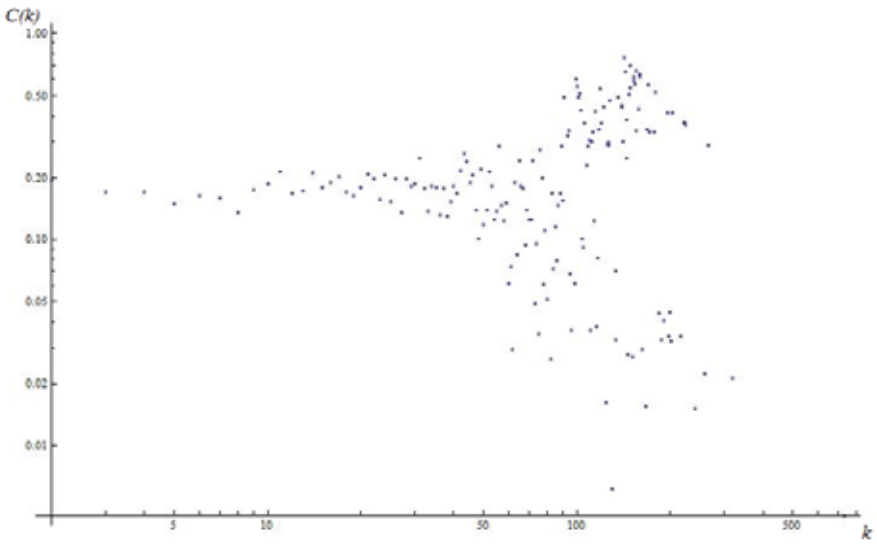

Figura 3: Gráfico da distribuição $C(k)$

4.2 Treinamento dos modelos

O algoritmo *j48* retorna *árvores de decisão* para mostrar quais as regras utilizadas para sua classificação. Para cada conjunto teste, a árvore de decisão possuía detalhamento diferente, mas as Figuras 4 e 5 mostram as árvores mais significativas para cada caso. *Verd.* significa que o algoritmo identificou os genes como portadores da característica (para Figura 4, drogável, e para a Figura 5, mórbido), e *Falso* que indica que o algoritmo identificou os genes como não portadores da característica. Entre parênteses temos o número de acertos seguido do número de erros que o algoritmo cometeu seguindo o critério anterior no fluxograma. Para a Figura 4, temos que o algoritmo considerou necessários os dados de apenas 3 características para fazer a previsão a respeito da drogabilidade: presença da proteína codificada pelo gene na membrana plasmática (*PlasmaMembrane*), o grau de intermediação relacionado a interações regulatórias transcricionais (*InBetReg*) e o número de metabólitos que são utilizados como reagentes em uma reação metabólica catalisada por uma enzima codificada por esse gene (entradas metabólicas, *metin*). Para a Figura 5, foram necessários dados de 5 características: número de fatores de transcrição que controlam o gene (*regin*), grau de intermediação relacionado a interações metabólicas (*InBetMet*), se as proteínas codificadas pelo gene estão presentes no meio extracelular (*Extracellular*) e na membrana plasmática (*PlasmaMembrane*) e seu valor de coeficiente de agregação.

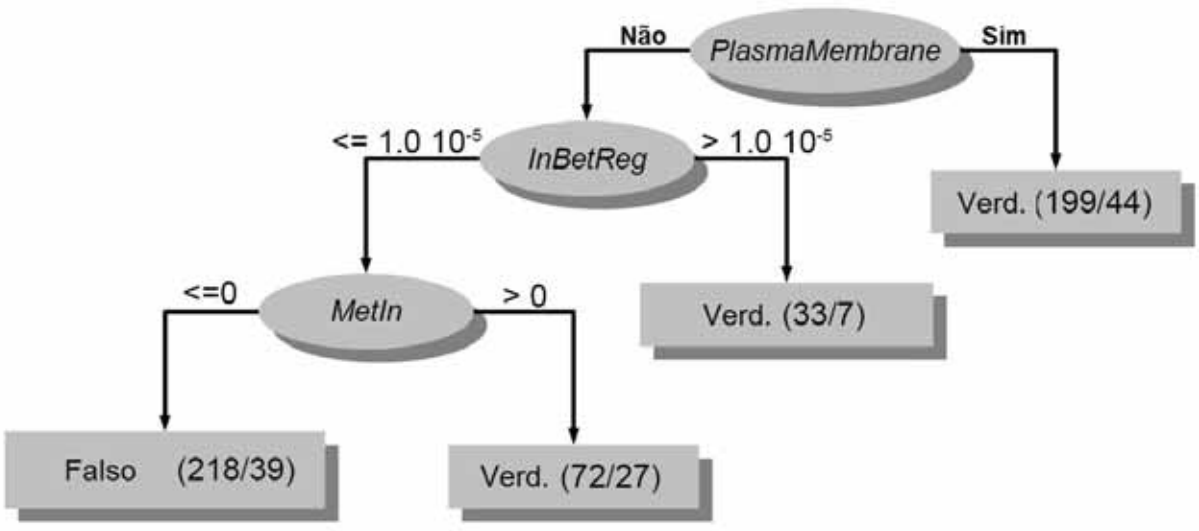

Figura 4: Árvore de decisão para drogabilidade

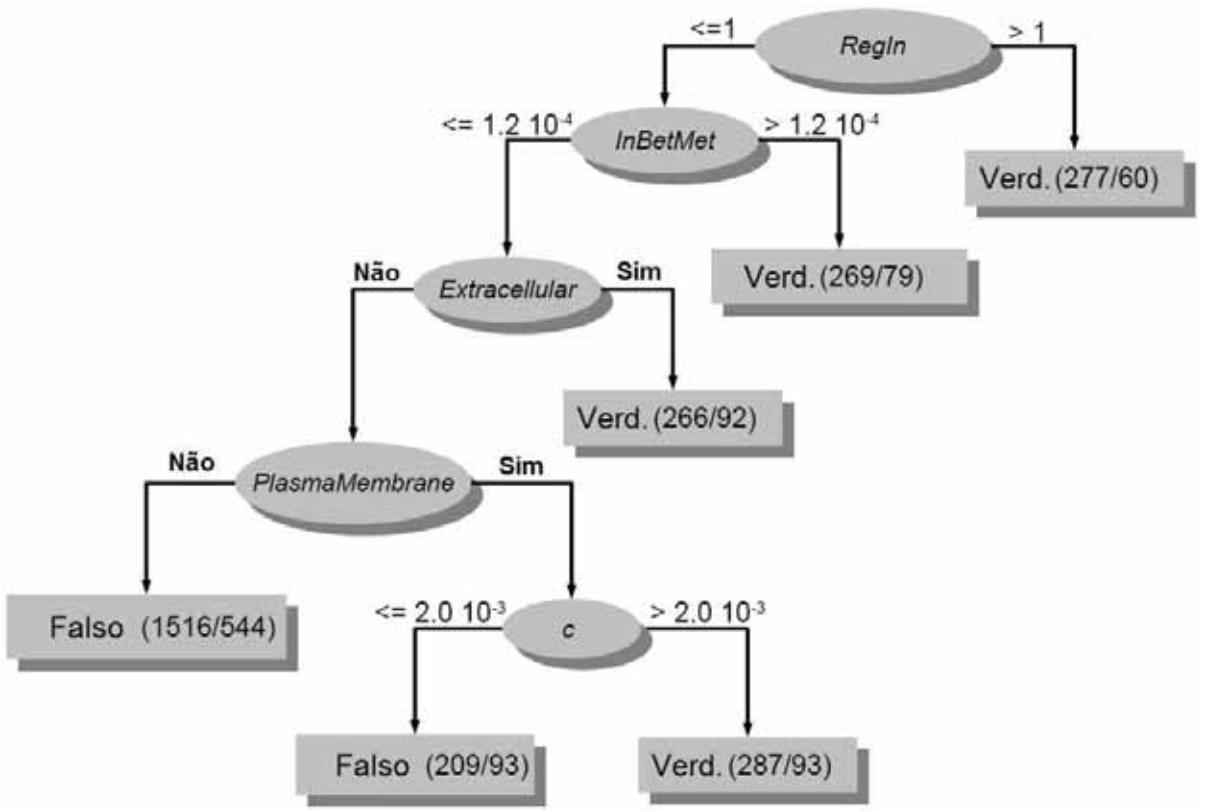

Figura 5: Árvore de decisão para morbidade

Na Tabela 2, temos os valores encontrados para *recall*, precisão e área sob a curva ROC para cada um dos 4 modelos (10 listas para cada um), e seu respectivo valor estatístico de Kappa. Os valores apresentados representam a média e o desvio padrão obtido. Podemos observar que as listas permutadas obtiveram resultados aleatórios, conforme o esperado.

Tabela 2: Tabela com os resultados obtidos durante o treinamento do modelo.

Dados	Área Sob Curva ROC (ASC) (%)	Recall (%)	Precisão (%)	Kappa
Drogabilidade	82 ± 1	78 ± 2	75 ± 1	0.51 ± 0.02
Drogabilidade Permutada	50 ± 3	50 ± 4	50 ± 3	0.00 ± 0.06
Morbidade	72 ± 1	65 ± 1	66 ± 1	0.32 ± 0.02
Morbidade Permutada	50 ± 2	50 ± 2	50 ± 2	-0.01 ± 0.03

Os resultados obtidos para *recall* e precisão médios dos modelos gerados para determinação da *morbidade* (65% e 66%, respectivamente) indicam grande quantidade de ruído nos conjuntos de teste, devido a possíveis características compartilhadas entre os genes classificados como "mórbidos" e "não mórbidos", que induziram o classificador ao erro. Isso se deve ao fato de não existir um conjunto negativo para essa classificação, ou seja, não existem evidências de que os genes apresentados como "não mórbidos" possam

ser classificados como tal. Logo, possíveis genes mórbidos participaram do conjunto de treinamento.

Além disso, a rede criada, apesar de integrar os dados experimentais de interações disponíveis na literatura, ainda está incompleta. Por exemplo, (STUMPF et al., 2008) estimaram que encontraremos em humanos cerca de 650 mil interações físicas entre proteínas, mas nossa rede dispunha de em torno de 43 mil. Os resultados obtidos para os valores topológicos poderão ser diferentes se novas interações forem incluídas, e as semelhanças encontradas pelo algoritmo para genes "mórbidos" e "não mórbidos" poderiam desaparecer. A existência de características comuns também afetou os resultados encontrados para *drogabilidade*, mas não com mesmo impacto.

Nas Figuras 6 e 7 temos os gráficos das distribuições de probabilidades de drogabilidade e morbidade respectivamente para os genes conhecidamente drogáveis/mórbidos, onde podemos observar o comportamento normal das listas permutadas em comparação com os modelos gerados, que possuem uma concentração da distribuição das probabilidades para valores acima de 0,5. O pequeno desvio do centro visto no gráfico para drogabilidade permutada pode ser explicado pelo fato de poucos genes participarem dessa lista. A tendência dessa curva é se aproximar cada vez mais do centro com o aumento do número de genes, como ocorre na curva para morbidade permutada, que possui maior número de componentes, o que classificaria a distribuição como gaussiana.

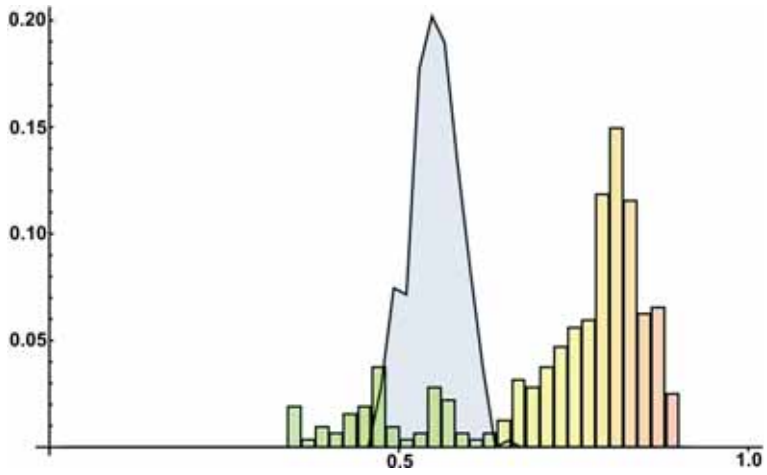

Figura 6: Barras: distribuição das probabilidades de drogabilidade (eixo das abscissas, intervalos de 0,2) de genes conhecidamente drogáveis para nosso modelo. Curva: distribuição das probabilidades de drogabilidade para genes conhecidamente drogáveis para a lista permutada.

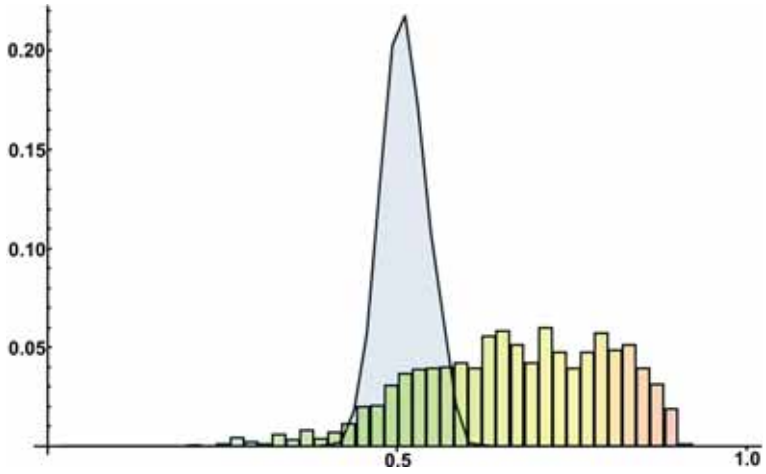

Figura 7: Barras: distribuição das probabilidades de morbidade (eixo das abscissas, intervalos de 0,2) de genes conhecidamente mórbidos para nosso modelo. Curva: distribuição das probabilidades de morbidade para genes conhecidamente mórbidos para a lista permutada.

4.3 Comparação com métodos semelhantes

Existem uma série de métodos computacionais para predição de genes mórbidos, mas a maioria deles se concentra em pequenos conjuntos de genes candidatos a doenças específicas, como o ENDEAVOUR (AERTS et al., 2006) e o ToppGene (CHEN et al., 2009). Nosso método foi construído levando todo o genoma, portanto podemos compará-lo com os desenvolvidos por (ADIE et al., 2005), PROSPECTR, e por (X JIANG R, 2008), CIPHER. Nosso método superou o CIPHER, que obteve uma precisão de 10%, sem declarar valores de *recall* ou ASC, e é comparável com o PROSPECTR, que obteve 62% de precisão, 70% de *recall* e 70% de ASC. Apesar desse último ter obtido um valor maior de *recall*, consideramos os resultados comparáveis já que obtivemos um valor maior para a precisão em para a ASC. Além disso, nossos resultados foram obtidos a partir de 10 listas que passaram por validação cruzada por 10 vezes, enquanto o resultado apresentado pelo PROSPECTR foi obtido a partir de uma única lista que passou por validação cruzada por 10 vezes.

Para a predição de genes drogáveis, assim como para a predição de genes mórbidos, nosso método só pode ser comparado com aqueles que consideraram todo os genoma para tal. Cumprindo esse requisito, encontramos o método desenvolvido por (SUGAYA; IKEDA, 2009), que utilizando *máquinas de suporte vetorial* (BURGES, 1998) treinadas em conjuntos com 69 atributos diversos, obteve uma precisão média de 70%, *recall* médio de 75%, e ASC média de 72%, valores comparáveis com nosso método utilizando meta-classificadores.

4.4 Aplicação dos modelos gerados

Apresentamos nas Tabelas 3 e 4:

- A lista dos 10 genes com maior probabilidade de serem drogáveis/mórbidos segundo nosso modelo;
- O valor mediano dos 10 resultados obtidos e valor mínimo e máximo obtido para cada gene, excluídos os conhecidamente drogáveis/mórbidos, para nosso modelo e para o modelo permutado;
- N e W para o teste de Wilcoxon;
- O resultado do teste está representado na coluna "$p < 0,05$", que indica se a probabilidade do resultado obtido pelo nosso modelo ser igual ao modelo permutado é menor que 5%, de acordo com a Tabela 1;
- Indícios encontrados na literatura confirmando o potencial do gene ser drogável/mórbido, representado pelo PubMed ID do respectivo artigo. Os indícios foram encontradas em 73% dos casos para drogabilidade 90% dos casos para morbidade. Diante dos resultados, podemos inferir que os genes restantes, ou seja, sem indícios, possuem grande probabilidade de atender aos pré-requisitos necessários para, a partir da análise convencional, que suas respectivas proteínas sejam alvos de drogas, ou no caso da morbidade, alterações nesses genes possam causar doenças hereditárias.

Tabela 3: Lista dos genes que possuem os 10 maiores *graus de drogabilidade* e indícios encontrados na literatura que confirmem essa condiçã.o

Gene	Grau de drogabilidade (Mediana [min,máx])		N	W	$p < 0,05$?	Indício
	Normal	Permutado				
HLA-F	0.887[0.803,0.915]	0.530[0.427,0.584]	10	0	Sim	–
PLAU	0.886[0.808,0.907]	0.561[0.387,0.675]	10	0	Sim	19301652
CD8A	0.885[0.871,0.902]	0.56[0.37,0.664]	10	0	Sim	–
CD19	0.880[0.751,0.907]	0.562[0.38,0.628]	10	0	Sim	19509168
ITGAM	0.878[0.614,0.887]	0.534[0.36,0.656]	10	1	Sim	11931348
THBS1	0.875[0.53,0.9]	0.532[0.293,0.592]	10	0	Sim	17878288
ITGAX	0.873[0.784,0.897]	0.539[0.422,0.691]	10	0	Sim	–
CXCR5	0.871[0.755,0.895]	0.537[0.49,0.59]	10	0	Sim	17652619
EBI3	0.871[0.801,0.888]	0.529[0.391,0.626]	10	0	Sim	19556516
IL6	0.87[0.766,0.893]	0.591[0.361,0.643]	10	0	Sim	17465721
TIMP2	0.869[0.645,0.916]	0.584[0.34,0.701]	10	0	Sim	10985804

Tabela 4: Lista dos genes que possuem os 10 maiores *graus de morbidade* e indícios encontrados na literatura que confirmem essa condição.

Gene	Grau de morbidade (Mediana [min,máx])		N	W	$p < 0,05$?	Indício
	Normal	Permutado				
TFRC	0.880 [0.576,0.939]	0.568 [0.447,0.678]	10	1	Sim	5941956
ITGA5	0.875 [0.635,0.916]	0.491 [0.377,0.631]	10	0	Sim	–
LTF	0.868 [0.803,0.913]	0.509 [0.356,0.642]	10	0	Sim	19258923
SFTPD	0.866 [0.618,0.923]	0.565 [0.458,0.682]	10	2	Sim	19590686
THBS1	0.865 [0.831,0.918]	0.511 [0.354,0.566]	10	0	Sim	18178577
TIMP2	0.860 [0.603,0.92]	0.574 [0.388,0.609]	10	0	Sim	19933216
TGFB2	0.857 [0.565,0.918]	0.526 [0.407,0.707]	10	3	Sim	19258923
CGA	0.856 [0.62,0.916]	0.535 [0.283,0.656]	10	0	Sim	19730683
SPP1	0.856 [0.577,0.887]	0.564 [0.34,0.696]	10	0	Sim	15868370
FLT1	0.854 [0.61,0.931]	0.527 [0.424,0.715]	10	3	Sim	19741061
NOL3	0.850 [0.647,0.875]	0.576 [0.31,0.651]	10	1	Sim	19773279

5 Conclusão

A identificação de novos genes mórbidos e de genes drogáveis é feita experimentalmente, e necessitando de muito tempo e gasto com materiais. Visando acelerar esse processo e diminuir seus custos, propomos uma método computacional baseado em aprendizagem de máquina, utilizando características topológicas da rede integrada de genes humanos, dados sobre expressão gênica e localização celular da respectiva proteína, para predição de genes mórbidos/drogáveis.

Nosso método possui três limitações:

– Dependência dos bancos de dados disponíveis, que tendem a priorizar estudos à respeito de genes de interesse para a área da saúde, ou seja, os genes mórbidos e os genes drogáveis;

– A construção da rede integrada depende de uma grande quantidade de dados experimentais, sendo que dispúnhamos de uma quantidade limitada de dados sobre os genes humanos – nossa rede, por exemplo, dispunha de em torno de 25% dos genes humanos conhecidos.

– Não existência de conjunto de dados negativos para o treinamento dos modelos, uma vez que não podemos confirmar a existência de genes "não drogáveis" ou "não mórbidos".

A despeito dessas limitações, demonstramos que nosso método possui desempenho moderado à alto nas listas de treinamento (Tabela 2), estatisticamente diferentes do resultado aleatório, e comparáveis ou superiores à outros métodos disponíveis na literatura. Além disso, genes drogáveis/mórbidos tendem a possuir *graus* superiores à $0,5$, conforme mostrado nas Figuras 6 e 7. Também foi possível, a partir das árvores de decisão geradas (Figuras 4 e 5), definir as principais características que são levadas em conta para essa predição.

Finalmente, a aplicação dos modelos gerados na lista contendo todos os genes e suas respectivas características, mostrou a validade de nosso método, uma vez que foram encontrados indícios que confirmam os resultados para a grande maioria dos genes que possuíam os 10 maiores *graus de drogabilidade* e *graus de morbidade* (Tabelas 3 e 4).

Referências

ACENCIO, M. L.; LEMKE, N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. **BMC Bioinformatics**, v. 10, 2009.

ADIE, E. et al. Speeding disease gene discovery by sequence based candidate prioritization. **BMC Bioinformatics**, v. 6, p. 55, 2005.

AERTS, S. et al. Gene prioritization through genomic data fusion. **Nature Biotechnology**, v. 24, n. 5, p. 537–544, 2006.

AHN, A. C. et al. The limits of reductionism in medicine: Could systems biology offer an alternative? **Plos Medicine**, v. 3, n. 6, p. 709–713, 2006.

ALBERT, R.; JEONG, H.; BARABASI, A. Error and attack tolerance of complex networks. **Nature**, v. 406, n. 6794, p. 378–382, 2000.

AMBERGER, J. et al. McKusick's Online Mendelian Inheritance in Man (OMIM (R)). **Nucleic Acids Research**, v. 37, p. D793–D796, 2009.

ANTHONISSE, J. **The rush in a directed graph**. Amsterdam: Stichting Mathematisch Centrum, 1971. (Technical Report BN, 9/71).

BINNS, D. et al. QuickGO: a web-based tool for Gene Ontology searching. **Bioinformatics**, v. 25, n. 22, p. 3045–3046, 2009.

BOLLOBÁS, B. **Graph theory: an introductory course**. New York: Springer, 1979.

BREITKREUTZ, B. J. et al. The BioGRID interaction database: 2008 update. **Nucleic Acids Research**, v. 36, p. D637–D640, 2008.

BURGES, C. A Tutorial on Support Vector Machines for Pattern Recognition. Boston: Kluwer Academic Publishers, 1998.

CHATR-ARYAMONTRI, A. et al. MINT: the molecular INTeraction database. **Nucleic Acids Research**, v. 35, p. D572–D574, 2007.

CHAURASIA, G. et al. UniHI: an entry gate to the human protein interactome. **Nucleic Acids Research**, v. 35, p. D590–D594, 2007.

CHEN, J. et al. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. **Nucleic Acids Research**, v. 37, p. W305–311, 2009.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, p. 37–46, 1960.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v. 7, p. 1–30, 2006.

DUARTE, N. C. et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. **Proceedings of the National Academy of Sciences of the United States of America**, v. 104, n. 6, p. 1777–1782, 2007.

ERDOS, P.; RENYI, A. On the evolution of random graphs. **Bulletin of the International Statistical Institute**, v. 38, n. 4, p. 343–347, 1960.

FREEMAN, L. A set of measures of centrality based on betweenness. **Sociometry**, v. 40, n. 1, p. 35–41, 1977.

GRIFFITHS, G. et al. **Genética Moderna**. Rio de Janeiro: Guanabara, 2001.

HERMJAKOB, H. et al. IntAct: an open source molecular interaction database. **Nucleic Acids Research**, v. 32, p. D452–D455, 2004.

JEONG, H. et al. The large-scale organization of metabolic networks. **Nature**, v. 407, n. 6804, p. 651–654, 2000.

JIANG, C. et al. TRED: a transcriptional regulatory element database, new entries and other development. **Nucleic Acids Research**, v. 35, p. D137–D140, 2007.

LIN, B. et al. Tracking the epidemiology of human genes in the literature: The HuGE Published Literature database. **American Journal of Epidemiology**, v. 164, n. 1, p. 1–4, 2006.

LINDSAY, M. Innovation: Target discovery. **Nature Reviews Drug Discovery**, v. 2, n. 10, p. 831–838, 2003.

MA, H. et al. The Edinburgh human metabolic network reconstruction and its functional analysis. **Molecular Systems Biology**, v. 3, 2007.

MATYS, V. et al. TRANSFAC (R) and its module TRANSCompel (R): transcriptional gene regulation in eukaryotes. **Nucleic Acids Research**, v. 34, p. D108–D110, 2006.

PAGEL, P. et al. The MIPS mammalian protein-protein interaction database. **Bioinformatics**, v. 21, n. 6, p. 832–834, 2005.

PERROUD, B. et al. Pathway analysis of kidney cancer using proteomics and metabolic profiling. **Molecular Cancer**, v. 5, 2006.

PICARD, R.; COOK, R. Cross-validation of regression-models. **Journal of the American Statistical Association**, v. 79, n. 387, p. 575–583, 1984.

PRASAD, T. S. K. et al. Human Protein Reference Database-2009 update. **Nucleic Acids Research**, v. 37, p. D767–D772, 2009.

REGENMORTEL, M. V. Reductionism and complexity in molecular biology. **Embo Reports**, v. 5, n. 11, p. 1016–1020, 2004.

REVERTER, A.; INGHAM, A.; DALRYMPLE, B. Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes. **BioData Mining**, v. 1, n. 1, p. 8, 2008.

RHODES, D. et al. Probabilistic model of the human protein-protein interaction network. **Nature Biotechnology**, v. 23, n. 8, p. 951–959, 2005.

RUAL, J. et al. Towards a proteome-scale map of the human protein-protein interaction network. **Nature**, v. 437, n. 7062, p. 1173–1178, 2005.

SABIDUSSI, G. The centrality index of a graph. **Psychometrika**, v. 31, n. 4, p. 581–603, 1966.

SHANNON, P. et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. **Genome Research**, v. 13, n. 11, p. 2498–2504, 2003.

SILVA, J. P. Muller da et al. In silico network topology-based prediction of gene essentiality. **Physica A-Statistical Mechanics and Its Applications**, v. 387, n. 4, p. 1049–1055, 2008.

SRIRAM, P.; STEVE, S. **Computational Discrete Mathematics:** Combinatorics and Graph Theory with Mathematica. Cambridge: Cambridge University Press, 2003.

STUMPF, M. P. H. et al. Estimating the size of the human interactome. **Proceedings of the National Academy of Sciences**, v. 105, n. 19, p. 6959–6964, 2008.

SUGAYA, N.; IKEDA, K. Assessing the druggability of protein-protein interactions by a supervised machine-learning method. **BMC Bioinformatics**, v. 10, p. 263, 2009.

WATTS, D.; STROGATZ, S. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440–442, 1998.

WHEELER, D. L. et al. Database resources of the national center for biotechnology information. **Nucleic Acids Research**, v. 36, p. D13–D21, 2008.

WILCOXON, F. Probability tables for individual comparisons by ranking methods. **Biometrics**, v. 3, n. 3, p. 119–122, 1947.

WITTEN, I.; FRANK, E. **Data Mining:** Practical Machine Learning Tools and Techniques with Java Implementations. 2ª. ed. São Francisco: Morgan Kaufmann Publishers, 2000.

X JIANG R, Z. M. L. S. W. Network-based global inference of human disease genes. **Molecular Systems Biology**, v. 4, p. 189, 2008.

XENARIOS, I. et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. **Nucleic Acids Research**, v. 30, n. 1, p. 303–305, 2002.

YILDIRIM, M. A. et al. Drug-target network. **Nature Biotechnology**, v. 25, n. 10, p. 1119–1126, 2007.

Anexos

Trabalhos baseados nos resultados obtidos

1.Os resultados obtidos por esse projeto foram submetidos em formato de artigo para a revista BMC Genomics;

2.Primeiro autor do trabalho apresentado em seção oral: COSTA, P. R.; ACENCIO, M. L.; LEMKE, N.. Network topology-based prediction of morbid and druggable genes. **International Workshop on Network Science 2009** (NetSci 2009), Veneza, 2009.

3.Apresentação de pôster: COSTA, P. R.; ACENCIO, M. L.; LEMKE, N.. Discovering gene druggability by Topological features in Human integrated network of gene interactions. **XII Encontro Nacional de Física da Matéria Condensada** (ENFMC XII), Águas de Lindóia, 2009;

4.Apresentação de pôster: COSTA, P. R.; ACENCIO, M. L.; LEMKE, N.. Topological Features Predict Druggability in Human Integrated Network of Gene Interactions. **5th International Conference of the Brazilian Association for Bioinformatics and Computional Biology** (V X-Meeting), Angra dos Reis, 2009;

5. Co-autor do trabalho: ACENCIO, M. L.; COSTA, P. R.; NOLLI, D. A.; LEMKE, N.. Network topology information-based prediction of human disease genes. **International Workshop and Conference on Network Science 2008** (NetSci 2008), Norwich, 2008

6. Apresentação de pôster: COSTA, P. R.; ACENCIO, M. L.; LEMKE, N.. Discovering gene druggability by Topological features in Human integrated network of gene interactions. **8th International Symposium on Mathematical and Computional Biology** (8th BioMat), Campos do Jordão, 2008

7. Apresentação oral: COSTA, P. R.; ACENCIO, M. L.; LEMKE, N.. Predição de genes alvo para drogas a partir da topologia da rede integrada do H. sapiens. **IV Congresso de Física Aplicada à Medicina** (IV CONFIAM), Botucatu, 2008.