



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”

Ivan Pagotto

**Determinação de genes potencialmente
responsivos à radiação ionizante através de
aprendizado de máquina**

Botucatu – SP
2013

Ivan Pagotto

**Determinação de genes potencialmente responsivos à
radiação ionizante através de aprendizado de máquina**

Departamento de Física e Biofísica
Instituto de Biociências
Universidade Estadual "Júlio de Mesquita Filho"
Campus de Botucatu

Orientador: Dr. Marcio Luis Acencio

Botucatu – SP
2013

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE - CRB 8/5651

Pagotto, Ivan.

Determinação de genes potencialmente responsivos à radiação ionizante através de aprendizado de máquina / Ivan Pagotto. - Botucatu, 2013

Trabalho de conclusão de curso (bacharelado - Física Médica) - Universidade Estadual Paulista, Instituto de Biociências de Botucatu

Orientador: Marcio Luis Acencio

Coorientador: Ney Lemke

Capes: 10501002

1. Radioterapia. 2. Cancer - Tratamento.. 3. Radiação ionizante. 4. Marcadores biológicos de tumor.

Palavras-chave: Aprendizado de máquina; Genes; Radiação ionizante.

À meus pais, Wilson e Sueli Pagotto, que tornaram possível esta conquista. À minha namorada Marcela por me apoiar nos momentos difíceis e a todos os meus amigos, que, unespianos ou não, me ajudaram a superar os obstáculos.

Agradecimentos

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

- Ao Doutor Marcio Luis Acencio e ao Professor Doutor Ney Lemke, pela orientação e incentivo;

- À meus pais e à minha namorada Marcela de Camargo Trentini, pelo carinho e paciência;

- À equipe do Laboratório de Bioinformática e Biofísica Computacional do Departamento de Física e Biofísica do IBB-Unesp e agregados de Laboratórios vizinhos.

“You don’t need a reason to help people.”

Zidane Tribal

“How do you prove that you exist ...? Maybe we don’t exist...”

Vivi Ornitier

Resumo

A predição da resposta do tumor à radioterapia é a questão mais importante durante o tratamento de pacientes com câncer. Como consequência, a predição de genes que sejam responsivos à radiação ionizante é uma possibilidade para a melhoria dos resultados clínicos e a otimização das doses as quais os pacientes são submetidos ao longo do tratamento. Juntamente com esses dados, é possível obter respostas sobre os mecanismos de resistência à radiação dos tumores e até mesmo a identificação de biomarcadores responsáveis pela resistência à radiação ionizante que podem ser potenciais para o desenvolvimento de novas drogas visando a proteção de tecidos saudáveis. A determinação experimental dos genes que sejam responsivos à radiação ionizante é algo caro e que demanda muito tempo e trabalho; porém, se utilizarmos uma forma computacional de direcionar os estudos experimentais diretamente aos genes que têm mais potencial para serem responsivos à radiação ionizante, as pesquisas podem ser mais direcionadas e específicas. Para determinar essa característica, 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. Os dados topológicos foram utilizados como atributos de treinamento para o aprendizado de máquina, no qual os genes conhecidamente responsivos à radiação ionizante foram apresentados a um algoritmo de árvore de decisão que gerou modelos de predição com índices de sensibilidade e precisão de 85% e 72%, respectivamente. Os índices de acerto obtidos para os conjuntos de teste foram satisfatórios, retornando 91% dos genes conhecidos como responsivos à radiação ionizante utilizados para o treinamento da árvore de decisão. Nós aplicamos o modelo de predição na rede integrada e atribuímos probabilidades de resposta à radiação ionizante a todos os genes da rede. Nós analisamos os genes potencialmente responsivos à radiação ionizante que apresentaram os maiores valores e pudemos constatar que esses genes participam de processos biológicos ligados direta ou indiretamente à resposta à radiação ionizante. Esses resultados indicam que o modelo de predição criado pode ser considerado uma ferramenta útil na seleção de genes que participam da resposta da célula à radiação ionizante.

Sumário

Sumário	9
1 Introdução	10
2 Objetivos	12
3 Metodologia	13
3.1 Seleção dos atributos de treinamento	13
3.2 Construção da <i>RIGH</i>	13
3.3 Cálculo das medidas de centralidade	14
3.4 Seleção dos algoritmos de aprendizado para predição	15
3.5 Construção dos grupos de treinamento	15
3.6 Avaliação dos preditores	16
3.7 Aplicação dos preditores para previsão	17
4 Resultados e Discussão	18
4.1 Avaliação e análise dos preditores	18
4.2 Classificação dos genes pelo modelo	19
5 Conclusão	22
6 Referências	23

1 Introdução

Muitos comportamentos manifestados pelos processos biológicos e pelos seus componentes são propriedades emergentes, isto é, propriedades que surgem a partir das interações entre os componentes dos processos. Por conta dessa natureza, as propriedades emergentes não podem ser explicadas ou mesmo previstas através do estudo de cada componente individualmente (REGENMORTEL, 2004). Embora a dissecação dos sistemas biológicos em suas partes constituintes pelo método reducionista tem ajudado o esclarecimento do funcionamento de alguns aspectos relacionados aos processos biológicos, somente uma abordagem holística é capaz de revelar como as interações entre os componentes de um sistema organizam-se para o surgimento das propriedades emergentes (AHN et al., 2006).

Essa abordagem holística aplicada aos sistemas biológicos faz parte de um campo relativamente novo na biologia conhecido como biologia sistêmica. A biologia sistêmica tem como objetivo determinar como as propriedades emergentes manifestadas pelos componentes de um sistema biológico surgem a partir das interações não-lineares entre esses componentes. De forma geral, a etapa inicial em um estudo baseado em biologia sistêmica é a organização das interações relacionadas ao sistema biológico de interesse em forma de grafo ou rede, um objeto matemático formado pelo conjunto de nodos (componentes) e um conjunto de arestas que conectam cada dois nodos (BARABASI; OLTVAI, 2004).

Estudos realizados pelo nosso grupo demonstram que, através da utilização de uma abordagem fundamentada nos princípios da biologia sistêmica, é possível gerar hipóteses biologicamente plausíveis a respeito de certos processos biológicos de interesse (DA SILVA et al., 2008; ACENCIO; LEMKE, 2009; COSTA; ACENCIO; LEMKE, 2010). Para essa demonstração, foram utilizadas as etapas de montagem da rede até a previsão e descrição de novos comportamentos (geração de hipóteses), sendo que as principais técnicas computacionais utilizadas nessa demonstração foram (i) a construção de uma rede integrada, isto é, uma rede contendo não somente dados de interatoma de proteínas, mas uma rede contendo também interações de regulação transcricional e interações metabólicas, (ii) cálculo de medidas de centralidade de rede – medidas que medem, como o próprio nome diz, a centralidade ou a importância de um nodo na rede – e (iii) o aprendizado de máquina.

O aprendizado de máquina é uma subárea da inteligência artificial dedicado ao desenvolvimento de algoritmos que permitam ao computador aprender e extrair padrões relevantes a um certo problema (WITTEN; FRANK, 2000). De forma geral, a utilização do aprendizado de máquina envolve a criação de um preditor através da seleção de atributos de treinamento, isto é, características associadas às instâncias (no nosso caso, os genes) analisadas pelos algoritmos de aprendizado (AA) para extração de padrões, da seleção de um AA ou de uma combinação de AAs, da construção de um grupo de treinamento, isto é, grupo de instâncias cuja classificação nas classes de interesse se conhece e, finalmente,

do treinamento do AA ou combinação de AAs com os grupos de treinamento e geração do preditor.

No trabalho (COSTA; ACENCIO; LEMKE, 2010), foi demonstrado que, através da utilização de aprendizado de máquina em uma rede integrada de genes humanos, medidas de centralidade da rede associadas a outros dados sistêmicos são atributos de treinamento satisfatoriamente capazes de prever genes mórbidos e drogáveis e de descrever regras celulares importantes para a morbidade e a drogabilidade de um gene (COSTA; ACENCIO; LEMKE, 2010). Com base no sucesso obtido na previsão e descrição de genes essenciais em *E. coli* e *S. cerevisiae* (DA SILVA et al., 2008; ACENCIO; LEMKE, 2009) e genes mórbidos e drogáveis em humanos (COSTA; ACENCIO; LEMKE, 2010), uma nova questão sobre a utilização do aprendizado de máquina em redes integradas surgiu: será que é possível prever genes responsivos à radiação ionizante em humanos utilizando essa abordagem?

A resposta a essa questão é pertinente por que a identificação experimental de novos genes que sejam responsivos à radiação ionizante pode tornar o conhecimento sobre os mecanismos de resposta dos tecidos aos danos causados pela radiação ionizante e, com isso, propiciar o desenvolvimento de possíveis fármacos para a otimização dos tratamentos com radioterapia com o intuito de aumentar a resistência à radiação ionizante dos tecidos saudáveis e diminuir a resistência à radiação ionizante dos tecidos tumorais (WEST; BARNETT, 2011). Ainda, como estudos experimentais com o propósito de determinar os genes que reagem à radiação ionizante são trabalhosos e requerem muito tempo de execução, o desenvolvimento de uma técnica computacional capaz de prever genes responsivos à radiação ionizante é altamente recomendável para tal fim.

Com o intuito de desenvolver uma técnica computacional capaz de prever genes responsivos à radiação ionizante e simultaneamente investigar se aprendizado de máquina e dados de redes biológicas seriam capazes de tal predição, nós desenvolvemos neste projeto um método computacional baseado nesses princípios para a determinação do potencial de resposta à radiação ionizante dos genes humanos.

2 Objetivos

Para o desenvolvimento do método computacional baseado em aprendizado de máquina e propriedades das redes biológicas para a determinação de genes potencialmente responsivos à radiação ionizante é necessário os seguintes passos:

1. Montagem da rede integrada de interações entre genes humanos (RIGH);
2. Avaliar o desempenho dos modelos de predição;
3. Desenvolver um modelo final de classificação para genes responsivos à radiação ionizante;
4. Determinar as condições para que um gene seja responsivo à radiação ionizante;
5. Atribuir um valor de probabilidade de responsividade à radiação ionizante (p_{rad}) a todos os genes da RIGH;
6. Selecionar os genes com maiores valores de p_{rad} e interpretar sua função biológicas.

3 Metodologia

De forma geral, todos os objetivos foram alcançados mediante criação de classificadores ou preditores. A criação de um preditor envolve *(i)* a seleção de atributos de treinamento (nesse caso, várias medidas de centralidade da rede [ver Tabela 1]), isto é, características associadas às instâncias (nesse caso, os genes) analisadas pelos algoritmos de aprendizado (AA) para extração de padrões; *(ii)* seleção de um AA ou de uma combinação de AAs, *(iii)* a construção de um grupo de treinamento, isto é, grupo de instâncias com classificação conhecida e seus atributos de aprendizagem e, finalmente, *(iv)* o treinamento do AA ou combinação de AAs com os grupos de treinamento e geração do preditor. Assim, com o preditor gerado, avaliamos seu desempenho e finalmente o utilizamos para o objetivo a que se destina.

3.1 Seleção dos atributos de treinamento

Utilizamos como atributos de treinamento as medidas de centralidade da rede. Para a obtenção dos dados referentes às medidas de centralidade, as etapas da seleção dos atributos de treinamento foram *(i)* a construção da rede integrada de interações gênicas de Humanos (*RIGH*) e *(ii)* o cálculo de medidas de centralidade a partir da *RIGH*.

3.2 Construção da *RIGH*

A *RIGH* foi construída como descrito em (COSTA; ACENCIO; LEMKE, 2010). Para a construção da *RIGH*, dois genes, g_1 e g_2 , que codificam, respectivamente, as proteínas p_1 e p_2 , são considerados interagentes se *(i)* p_1 e p_2 interagem fisicamente (interação física entre proteínas, *ppi*), *(ii)* o fator de transcrição p_1 regula diretamente a transcrição de g_2 , isto é, p_1 se liga à região promotora de g_2 (interação de regulação transcricional, *reg*) ou *(iii)* as enzimas p_1 e p_2 compartilham metabólitos, isto é, o produto gerado por uma reação catalisada pela enzima p_1 é usado como reagente em uma reação catalisada pela enzima p_2 (interação metabólica, *met*). Esses três tipos de interações foram coletados a partir de diferentes bancos de dados, como descrito adiante, sendo consideradas para a construção da *RIGH* somente as interações verificadas experimentalmente. As *ppis* foram obtidas a partir do *Human Integrated Protein-Protein Interaction reference (HIPPIE)* (SCHAEFER et al., 2012), um repositório de *ppis* e associações entre interações gênicas experimentalmente verificadas de humanos. As *regs* foram obtidas a partir do banco de dados *The Human Transcriptional Regulation Interactions database (HTRIdb)* (BOVOLenta; ACENCIO; LEMKE, 2012), um repositório de *regs* verificadas

experimentalmente em Humanos e para obter as *mets* experimentalmente verificadas, utilizamos o banco de dados *Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions (BiGG)* (SCHELLENBERGER et al., 2010).

3.3 Cálculo das medidas de centralidade

Com a *RIGH* construída, calculamos várias medidas de centralidade (ver a Tabela 1 para uma breve descrição de cada medida) utilizadas como os atributos de treinamento. Para isso utilizamos o pacote *Networkx* (HAGBERG; SCHULT; SWART, 2008) para Python. O *Networkx* é um pacote que contém centenas de algoritmos para criação, manipulação e análise da estrutura, dinâmica e funções de redes complexas.

Tabela 1 – Funções e descrições das 17 medidas de centralidade utilizadas como atributos de treinamento neste projeto.

Centralidade	Função	Descrição
Grau de conectividade geral e para as interações físicas entre proteínas, regulatórias e metabólicas.	deg	Número de interações do gene g .
Grau de conectividade de entrada relacionado à interação metabólica	met_in	Número de interações de entrada de g que representam os reagentes que participam de uma reação metabólica catalisada pela enzima codificada por g
Grau de conectividade de saída relacionado à interação metabólica	met_out	Número de interações de saída de g que representam os produtos gerados pela reação metabólica catalisada pela enzima codificada por g
Grau de conectividade de entrada relacionado à interação de regulação transcricional	reg_in	Número de interações de entrada de g que representam o número de fatores de transcrição que regulam g
Grau de conectividade de saída relacionado à interação de regulação transcricional	reg_out	Número de interações de saída de g que representam o número de genes regulados pelo fator de transcrição codificado por g
Coefficiente de agrupamento	$clusterG = \frac{2n_g}{k_g(k_g - 1)}$	n_g é o número de interações que conectam os vizinhos de g e k_g é o número de interações que conectam g aos seus vizinhos.
Grau de intermediação geral e para as interações físicas entre proteínas, regulatórias e metabólicas.	$bet = \sum_{g_i \neq g \neq g_j} \frac{\sigma_{g_i g_j}(g)}{\sigma_{g_i g_j}}$	$\sigma_{g_i g_j}$ é o número de caminhos mais curtos entre 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 .
Grau de proximidade geral e para as interações físicas entre proteínas, regulatórias e metabólicas.	$clo = \frac{n}{\sum_{g_j} d(g, g_j)}$	$d(g, g_j)$ é a distância mais curta via todos os tipos de interações entre g and g_j ; n é o número de genes na rede.

3.4 Seleção dos algoritmos de aprendizado para predição

Selecionamos o AA usando o *WEKA* (*Waikato Environment for Knowledge Analysis*), programa escrito em JAVA desenvolvido na Universidade de Waikato, Nova Zelândia, que agrega ferramentas de visualização e algoritmos de análise de dados, incluindo centenas de algoritmos de aprendizagem de máquina (WITTEN; FRANK, 2000).

O AA selecionado para prever os genes responsivos à radiação ionizante, isto é, atribuir aos genes a probabilidade estimada desses genes pertencerem à classe dos responsivos à radiação ionizante ($D(i)$), foi o J48, AA que gera uma árvore de decisão, isto é, um grafo de condições e suas possíveis consequências. A estrutura de uma árvore de decisão consiste em um nodo-raíz que representa a condição mais importante para distinguir as diferentes classes sob estudo, nodos internos que representam condições adicionais para a distinção das classes e nodos-folhas que representam a classificação final. Portanto, além de atribuir valores de $D(i)$ para as instâncias, é possível também determinar as condições para classificar as instâncias nas diferentes classes sob estudo. Para isso, deve-se seguir o caminho do nodo-raiz até os nodos-folhas (KINGSFORD; SALZBERG, 2008).

3.5 Construção dos grupos de treinamento

Para a predição de genes que são responsivos a radiação ionizante, construímos 1000 grupos "balanceados" de treinamento contendo 97 genes conhecidamente responsivos à radiação ionizante de acordo com o banco de dados Gene Ontology (CONSORTIUM, 2012) com 97 genes selecionados aleatoriamente da *RIGH* cuja resposta à radiação ionizante ainda é desconhecido. Como atualmente não é possível afirmar inequivocamente que um gene não seja responsivo à radiação ionizante, tivemos que considerar como genes não responsivos à radiação ionizante todos os genes da *RIGH* exceto os conhecidamente responsivos selecionados no Gene Ontology. Esses grupos foram denominados "treinamento normal". A partir desses 1000 grupos redistribuímos aleatoriamente a classe dos genes em cada grupo 10 vezes e, dessa forma, obtivemos 10000 grupos com a classe dos genes distribuída aleatoriamente entre os integrantes de cada grupo. Esses grupos foram denominados de "treinamento permutado". A construção de um grupo de treinamento no qual as classes são aleatoriamente atribuídas às instâncias serve para verificar se o AA treinado com o grupo de treinamento normal aprendem características realmente associadas às classes de interesse em vez de características associadas a quaisquer subgrupos aleatórios de gene ou interações. Ainda, esses grupos de treinamento são ditos "balanceados" por que eles contêm o mesmo número de instâncias. O treinamento de um algoritmo de aprendizagem com a mesma quantidade de instâncias positivas e negativas é essencial já que, quando essa diferença é muito grande, o AA pode encontrar dificuldades em gerar regras para a classe minoritária (VISA; RALESCU, 2005) e isso degrada o desempenho

desses algoritmos. Como há muito mais genes cuja resposta à radiação ionizante é desconhecido do que genes conhecidamente responsivos à radiação ionizante, foi necessário balancear os grupos para evitar o problema supracitado

3.6 Avaliação dos preditores

O desempenho dos preditores foi avaliado através da determinação da precisão, da sensibilidade e da área sob a curva ROC (*Receiver operating characteristic*) desses preditores.

Dada duas classes, i e j , precisão é a proporção entre instâncias realmente pertencentes à classe i que são corretamente classificadas como i (VP) e todas as instâncias classificadas como i ($VP + FP$):

$$Precisão = \frac{VP}{VP + FP} \quad (3.1)$$

VP significa “verdadeiros positivos” e representa a quantidade de instâncias realmente pertencentes à classe i que são corretamente classificadas como i . FP significa “falsos positivos” e representa a quantidade de instancias realmente pertencentes à classe j que são incorretamente classificadas como i .

Sensibilidade é a proporção entre instâncias realmente pertencentes à classe i que são corretamente classificadas como i (VP) e todas as instâncias realmente pertencentes à classe i ($VP + FN$):

$$Sensibilidade = \frac{VP}{VP + FN} \quad (3.2)$$

FN significa “falsos negativos” e representa a quantidade de instâncias realmente pertencentes à classe i que são incorretamente classificadas como j .

A área sob a curva ROC (ASC) plota a fração de verdadeiros positivos pela fração de falsos positivos, sendo que a área abaixo dessa curva é numericamente igual a probabilidade de uma determinada instância ser corretamente classificada.

As precisões, sensibilidades e ASCs dos preditores foram estimadas através da técnica de validação cruzada com a utilização do WEKA. Nessa técnica, o grupo de treinamento original é aleatoriamente dividido em v subgrupos. Desses v subgrupos, um é mantido como o subgrupo de validação para testar o modelo de predição e os $v - 1$ remanescentes são usados como o subgrupo de treinamento. O processo de validação cruzada é repetido v vezes, sendo que cada um dos v subgrupos são utilizados exatamente e somente uma vez como subgrupo de validação. As medidas de desempenho finais são as médias das medidas de cada etapa da validação cruzada. Estimamos as precisões e as sensibilidades de nossos preditores, utilizando $v = 10$. A escolha desse valor deve-se à observação de que a variação das medidas de desempenho estimadas entre os v subgrupos é menor para $v = 10$ do que para outros valores de v quando se utiliza indutores de árvores de decisão (KOHAVI, 1995).

As precisões, sensibilidades e ASCs estimadas pela validação cruzada foram expressas como medianas das 1000 medidas de desempenho dos 1000 preditores gerados pelo treinamento dos 1000 grupos de treinamento do conjunto normal e das 10000 medidas de desempenho dos 10000 preditores gerados pelo treinamento os 10000 grupos de treinamento do conjunto permutado.

3.7 Aplicação dos preditores para previsão

Os 1000 preditores construídos a partir dos grupos de treinamento normal foram concatenados em um único preditor utilizando "Vote", um meta-algoritmo implementado no WEKA que combina as D(i)s de cada preditor através de diferentes regras (KITTLER et al., 1998). Nós aplicamos esse preditor combinado para atribuir a todos os genes da RIGH uma probabilidade de resposta à radiação ionizante (p_{rad}).

4 Resultados e Discussão

4.1 Avaliação e análise dos preditores

Na Tabela 2, temos os valores encontrados para sensibilidade, precisão e área sob a curva ROC para os treinamentos normal e permutado. Os valores apresentados representam a média e o desvio padrão obtidos. Podemos observar que o treinamento permutado obteve resultados aleatórios, conforme o esperado.

Tabela 2 – Performance dos modelos de predição

Modelo	Sensibilidade (%)	Precisão (%)	ASC (%)
Normal	85 ± 6	72 ± 3	75 ± 3
Permutado	42 ± 27	49 ± 12	48 ± 3

Como podemos observar na Figura 1 as distribuições dos valores de ASC do treinamento permutado e do treinamento normal são diferentes e essa observação indica que o algoritmo encontra padrões que correspondem às características intrínsecas aos genes responsivos à radiação ionizante em vez de subgrupos aleatórios de genes..

Os resultados obtidos do treinamento normal demonstram que as propriedades de redes na forma de medidas de centralidade aparentemente são capazes de recuperar 85% dos genes responsivos à radiação ionizante com uma precisão de 72%. Isso demonstra que existe uma ligação direta entre as propriedades de rede relacionadas aos genes responsivos a radiação ionizante.

Além da predição em si, o algoritmo J48 retorna árvores de decisão para mostrar quais as regras utilizadas para a classificação das instâncias de interesse (no nosso caso, a classificação dos genes em responsivos ou não-responsivos à radiação ionizante). Foram geradas 1000 árvores de decisão e a Figura 2 mostra uma árvore representativa que possui os atributos que aparecem com mais frequência nessas árvores.

Como pode ser observado na Figura 2, temos que o algoritmo considerou necessários os dados de apenas 3 características para fazer a previsão: **deg_ppi**, que representa a quantidade de interações físicas entre proteínas que os genes codificam; o **reg_in**, que quantifica as conexões de entrada relacionados com a regulação transcricional; e o **clo_ppi**, que representa a proximidade entre os caminhos da rede para ocorrer interações físicas com a proteína. Essa árvore indica que os genes potencialmente responsivos à radiação ionizante codificam proteínas com alta conectividade, têm sua transcrição altamente regulada (regulação por mais de três fatores de transcrição) e as funções de suas proteínas têm forte influência sobre todas as outras proteínas da rede (grau de proximidade maior que 0,3).

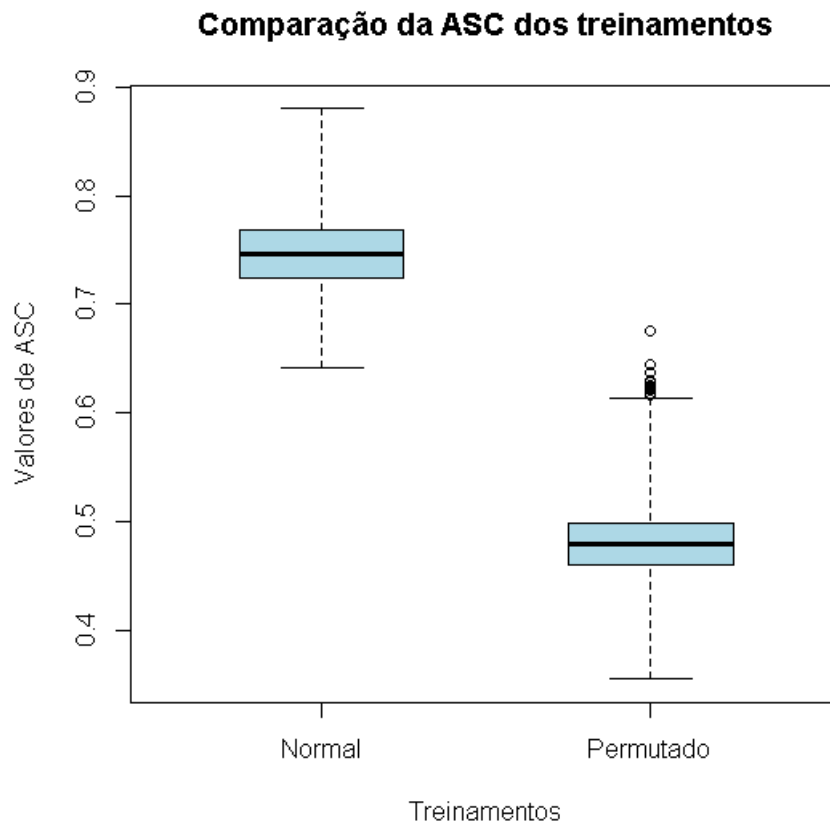


Figura 1 – *Boxplot* da distribuição dos valores da ASC dos 1000 valores do modelo normal comparada com os 10000 valores do modelo permutado.

4.2 Classificação dos genes pelo modelo

Para classificar os genes a partir do treinamento normal, criamos um modelo genérico com base na média entre todos os valores obtidos para cada gene de cada um dos 1000 modelos utilizando o meta-algoritmo "Vote" implementado no WEKA de forma que a classificação é feita em todas as árvores e a probabilidade de resposta à radiação ionizante (p_{rad}) é a média aritmética dos valores obtidos. Para a avaliação dos genes classificados como responsivos à radiação ionizante tomamos como valor de corte 0,75, que é a mediana dos valores classificados para os 97 genes tomados como verdadeiramente responsivos à radiação ionizante. Dos 19589 genes na rede, 1250 receberam valores de p_{rad} maior que 0,75.

A distribuição da Figura 3 é resultado da classificação do modelo normal para todos os genes presentes na rede construída. Podemos observar que nosso modelo separou toda a rede em dois grupos bem distintos: um com baixa probabilidade de resposta à radiação ionizante e o de alta probabilidade de resposta à radiação ionizante. O fato de nenhum gene ter recebido valores de p_{rad} intermediários indica que há um grupo bem definido de 1250 genes na rede que controla a resposta das células à radiação ionizante.

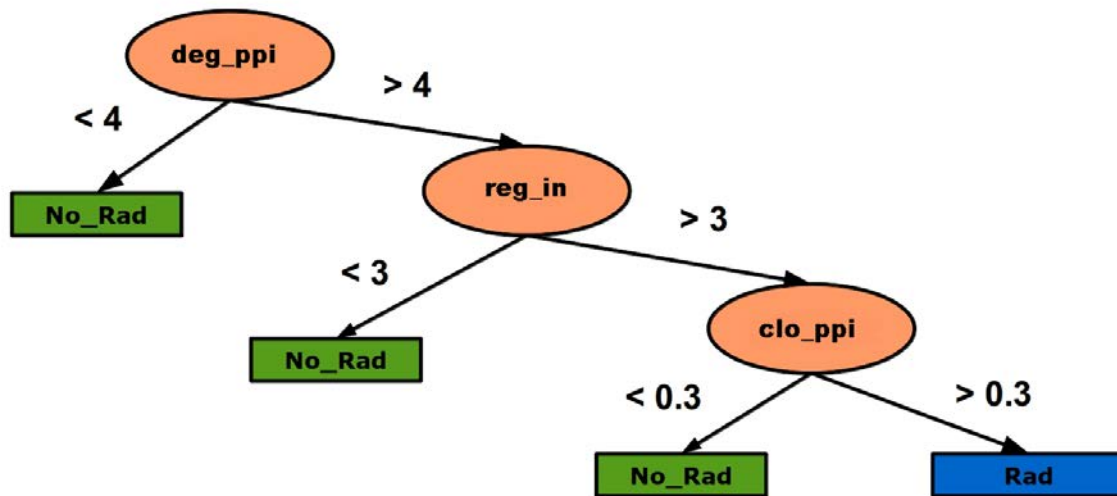


Figura 2 – Árvore de decisão genérica do modelo normal. A classificação Rad significa que o algoritmo identificou os genes como potencialmente responsivos à radiação ionizante, e No_Rad que o algoritmo identificou o gene como não responsivo à radiação ionizante. **deg_ppi**: quantidade de interações físicas entre proteínas que os genes codificam; **reg_in**: quantifica as conexões de entrada relacionadas com a regulação transcricional; e **clo_ppi**: representa a proximidade entre os caminhos da rede através de interações físicas entre proteínas.

Tabela 3 – Lista de genes com os dois maiores valores de probabilidade de resposta à radiação ionizante

Gene	GeneId	p_rad (mediana)	
		Normal	Permutado
ESR1	2099	0.792	0.499
SP1	6667	0.792	0.499
AR	367	0.792	0.499
STAT1	6772	0.791	0.5
CREB1	1385	0.791	0.501
RARA	5914	0.791	0.501
HIF1A	3091	0.791	0.499
NFKB1	4790	0.791	0.5
E2F1	1869	0.791	0.499

Desses 1250 genes considerados responsivos à radiação ionizante, nós analisamos aqueles que receberam os dois maiores valores de p_rad (Tabela 3). Curiosamente, todos os nove genes considerados por nosso modelo como os mais responsivos à radiação ionizante codificam fatores de transcrição, isto é, proteínas que se ligam às regiões promotoras dos genes e que através dessa ligação regulam a expressão desses genes. Segundo o banco de dados Gene Ontology (CONSORTIUM, 2012), esses fatores de transcrição estão envolvi-

dos com proliferação celular e apoptose, processos biológicos intimamente relacionados à resposta à radiação ionizante (HARRINGTON; JANKOWSKA; HINGARONI, 2007).

Observando os dados de alguns fatores de transcrição em particular, é possível notar que nosso preditor classificou como responsivos à radiação ionizante genes relevantes. O fator H1F1A, por exemplo, é um fator de transcrição induzido por hipóxia – baixa concentração de oxigênio – e sabe-se que hipóxia tem relação direta com a resistência à radiação ionizante (WACHSBERGER; BURD; DICKER, 2003). O fator NFkB1, por sua vez, é induzido, segundo dados do Gene Ontology, por radiação ultravioleta e está envolvido com apoptose. E, finalmente, o fator E2F1, de acordo também com o Gene Ontology, está envolvido com a apoptose em resposta ao dano no DNA, efeito principal da radiação ionizante nas células.

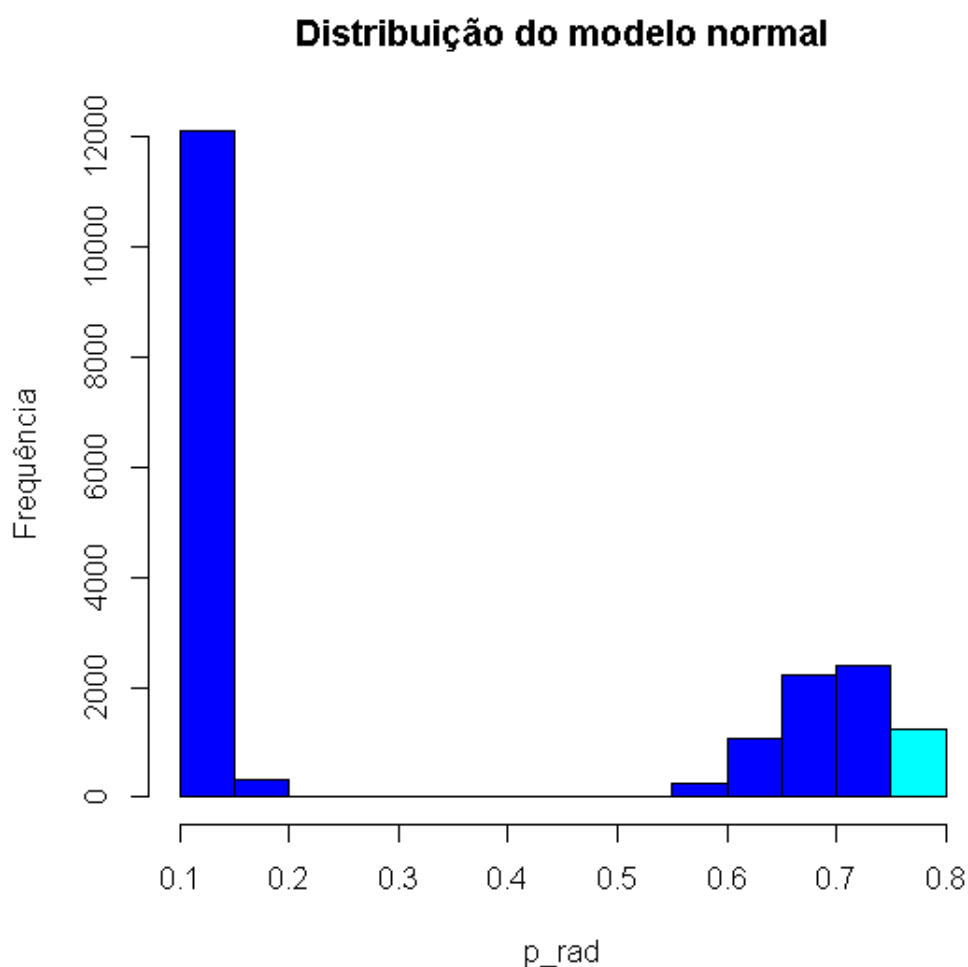


Figura 3 – Distribuição dos genes classificados pelo modelo normal, onde a coloração mais clara é a porção dos genes que foi classificada como responsivos à radiação ionizante, contendo 1250 genes. Os 97 genes conhecidamente responsivos à radiação ionizante usados para o treinamento também estão nessa região mais clara.

5 Conclusão

A identificação experimental de novos genes responsivos à radiação é um método caro e que demanda tempo. Para a redução de custo e tempo podemos utilizar o método computacional para filtrar as possibilidades e otimizar o tempo de obtenção de resultados. Neste trabalho, nós justamente desenvolvemos um método computacional baseado em aprendizado de máquina e propriedades topológicas da rede para prever genes responsivos à radiação ionizante

Como todo método computacional o sistema de predição não é absoluto. A falta de bancos de dados disponíveis com informações sobre a resposta gênica à radiação ionizante dificulta a predição, pois o grupo de treinamento com validação na literatura é pequeno e não podemos confirmar a não resposta à radiação ionizante, pois não temos nenhum grupo que possa ser usado como treinamento.

Mesmo com essas dificuldades o modelo preditor se mostrou eficiente para recuperar os genes conhecidamente responsivos à radiação ionizante. Em relação aos classificados como possíveis responsivos à radiação ionizante, a análise funcional dos genes que receberam os maiores valores de p_{rad} reforça a eficiência de nosso preditor já que todos esses genes têm algum envolvimento com resposta à radiação ionizante.

A aplicação do nosso modelo na rede revelou 1250 genes altamente responsivos à radiação ionizante. Esses genes parecem formar um grupo bem distinto de genes que, de acordo com as condições indicadas pela árvore de decisão da Figura 2, parecem ter um papel importante na célula: além de codificarem proteínas com alta conectividade, esses genes parecem ser altamente regulados e possuir funções centrais nas células.

6 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, p. 290, 2009.
- AHN, A. C. et al. The limits of reductionism in medicine: could systems biology offer an alternative? *PLoS Med*, v. 3, n. 6, p. e208, 2006.
- BARABASI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, v. 5, n. 2, p. 101–113, 2004.
- BOVOLENTA, L. A.; ACENCIO, M. L.; LEMKE, N. Htridb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, 2012.
- CONSORTIUM, G. O. The gene ontology: enhancements for 2011. *Nucleic Acids Res.*, 2012.
- COSTA, P. R.; ACENCIO, M. L.; LEMKE, N. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC Genomics*, v. 11, p. S9, 2010.
- DA SILVA, J. P. M. et al. In silico network topology-based prediction of gene essentiality. *Physica A*, v. 387, p. 1049–1055, 2008.
- HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. Exploring network structure, dynamics, and function using NetworkX. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA USA: [s.n.], 2008. p. 11–15.
- HARRINGTON, K.; JANKOWSKA, P.; HINGARONI, M. Molecular biology for the radiation oncologist: the 5Rs of radiobiology meet the Hallmarks of cancer. *Clinical Oncology*, 2007. p. 561–571
- KINGSFORD, C.; SALZBERG, S. L. What are decision trees? *Nat Biotechnol*, v. 26, n. 9, p. 1011–1013, 2008.
- KITTLER, J. et al. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 20, n. 3, p. 226–239, 1998.
- KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*. [S.l.]: Morgan Kaufmann, 1995. p. 1137–1143.
- REGENMORTEL, M. H. V. V. Reductionism and complexity in molecular biology. scientists now have the tools to unravel biological and overcome the limitations of reductionism. *EMBO Rep*, v. 5, n. 11, p. 1016–1020, 2004.
- SCHAEFER, M. et al. Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS One*, 2012.

SCHELLENBERGER, J. et al. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 2010.

VISA, S.; RALESCU, A. Issues in mining imbalanced data sets - a review paper. In: *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference*. [S.l.: s.n.], 2005. p. 67–73.

WACHSBERGER, P.; BURD, R.; DICKER, A.P. Tumor response to ionizing radiation combined with antiangiogenesis or vascular targeting agents: Exploring mechanisms of interaction. 2003 *Clinical Cancer Research*, 2003.

WEST, C. M.; BARNETT, G. C. Genetics and genomics of radiotherapy toxicity: towards prediction. *Genome Medicine*, 2011.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann, 2000.