UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Marcio Luis Acencio

Construção e análise da rede integrada de interações entre genes humanos envolvida com a regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular

Botucatu – SP

2011

Marcio Luis Acencio

Construção e análise da rede integrada de interações entre genes humanos envolvida com a regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular

Tese apresentada ao Instituto de Biociências de Botucatu da Universidade Estadual Paulista "Júlio de Mesquita Filho" para obtenção de título de Doutor em Ciências Biológicas (Genética).

Orientador:
Prof. Dr. Ney Lemke

Botucatu – SP

2011

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação
Divisão Técnica de Biblioteca e Documentação - Campus De Botucatu - UNESP
Bibliotecária responsável: *Sulamita Selma Clemente Colnago – CRB 8/4716*

Acencio, Marcio Luis.
Construção e análise da rede integrada de interações entre genes humanos envolvida com a regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular / Marcio Luis Acencio. - Botucatu, 2011

Tese (doutorado) - Instituto de Biociências de Botucatu, Universidade Estadual Paulista, 2011
Orientador: Ney Lemke
Capes: 31301037

1. Câncer - Aspectos genéticos. 2. Genética - Processamento de dados. 3. Bioinformática

Palavras-chave: Aprendizado de máquina; Bioinformática; Biologia sistêmica; Câncer; Ciclo celular; Metástase, Redes complexas

Àqueles cujo amor incondicional foi fundamental para que eu chegasse até aqui: minha esposa Denise, meus pais, Edison e Maisa e meu irmão Marcel

Agradecimentos

Meus mais sinceros agradecimentos a todos que me ajudaram na elaboração desta tese:

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

À minha esposa, Denise, pelo companheirismo, incentivo, compreensão e encorajamento, tudo isso com muito amor!

Aos meus pais, Edison e Maisa, e ao meu irmão, Marcel, por tudo que me ensinaram nessas três décadas de convivência regada de muito amor!

A todos do Laboratório de Bioinformática e Biofísica Computacional do Departamento de Física e Biofísica, em especial ao Luiz Augusto Bovolenta e ao Pedro Rafael Costa, pelos momentos acalorados de discussões científicas e também pelos momentos divertidos de descontração;

A todos do Laboratório de Biologia Molecular Estrutural, em especial à Agnes, Carlos e Juliana, pela amizade e pelos ensinamentos que me auxiliaram muito na consolidação da minha visão sobre ciência e sobre o meu trabalho em si;

Ao Pituta, também conhecido por Carlos Alexandre Henrique Fernandes, pela amizade e pela parceria, regada de muitas discussões, em prol daquilo que mais gostamos de fazer: ciência;

Aos Professores Doutores Joel Mesa Hormaza e Roberto Morato Fernandez, pela amizade e pelo apoio constante;

Aos Professores Doutores Marcos Antonio de Rezende e Marcos Roberto Mattos Fontes, chefes do Departamento de Física e Biofísica durante o período de desenvolvimento desta tese, pelo apoio técnico e burocrático;

A todos do Programa de Pós-Graduação em Ciências Biológicas em Genética, das secretárias aos membros do Conselho de Pós-Graduação do Programa, pelo apoio técnico e burocrático;

A todos da Diretoria do Instituto de Biociências de Botucatu, das secretárias aos diretores, pelo apoio técnico, financeiro e burocrático;

Aos membros do Núcleo de Computação Científica da UNESP (NCC/GridUNESP) pela disponibilização dos seus recursos computacionais;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro recebido durante os primeiros seis meses de trabalho;

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro recebido durante os últimos três anos de trabalho;

A todos os meus familiares e amigos, pela torcida e carinho;

Por fim, à cidade dos bons ares e boas escolas, Botucatu, que me recebeu de braços abertos e me adotou como um de seus filhos.

"Tudo é loucura ou sonho no começo. Nada do que o homem fez no mundo teve início de outra maneira – mas já tantos sonhos se realizaram que não temos o direito de duvidar de nenhum"

Monteiro Lobato (Mundo da Lua, 1923)

Resumo

Virtualmente, todas as células normais, com exceção das células hematopoieticas, precisam estar aderidas à matriz extracelular para que elas possam se proliferar. Na ausência de adesão, essas células não se proliferam mais e acabam sofrendo apoptose. Porém, após transformação oncogênica, as células adquirem a capacidade de proliferação na ausência de adesão à matriz extracelular. Essa capacidade, cuja base molecular está na regulação anormal da transição G1/S do ciclo celular pela adesão, é uma das propriedades fundamentais das células cancerosas e também requisito para que essas células adquiriam sua capacidade metastática. Como as metástases correspondem a aproximadamente 90% das mortes por câncer, a elucidação dos mecanismos moleculares subjacentes à regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular é, portanto, essencial para o desenvolvimento de drogas que possam inibir a formação das metástases. Com o intuito de elucidar esses mecanismos, nós adotamos neste trabalho uma abordagem estritamente computacional baseada em teoria das redes e aprendizado de máquina através do desenvolvimento de novos métodos de *(i)* construção de redes que representam a provável regulação entre dois diferentes processos (nesse caso, regulação da transição G1/S pela adesão à matriz extracelular), *(ii)* predição de interações oncogênicas, *(iii)* determinação de sub-redes de vias de sinalização oncogênica entre dois genes de interesse em uma rede (batizado de *graph2sig*) e *(iv)* predição de potenciais alvos de drogas. A rede potencialmente envolvida na regulação da transição G1/S do ciclo celular pela matriz extracelular construída (G_{ccam}) possui $\approx$ 2000 genes e $\approx$ 20.000 interações e representa $\approx$ 78% dos processos biológicos conhecidamente envolvidos nessa regulação. A aplicação do *graph2sig* na G_{ccam}, sendo os genes de interesse o *EGFR* e o *CDC6*, genes cujas proteínas codificadas parecem exercer um importante papel na proliferação independente de adesão à matriz extracelular, nos permitiu levantar a seguinte hipótese sobre a proliferação das células cancerosas sem adesão à matriz extracelular: parte dessa capacidade, pelo menos para células cancerosas que carregam a proteína EGFR continuamente ativada, deve-se à estabilização da proteína CDC6 pela CDKN1A. A aplicação do método de predição de alvos de droga na G_{ccam}, mais especificamente sobre a sub-rede de vias de sinalização oncogênica gerada entre os genes *EGFR* e *CDC6* pelo *graph2sig*, nos permitiu, por sua vez, indicar os genes *CDKN1A*, *JUN*, *SMAD3*, *SMAD4*, *CAV1*, *CCND1* e *CTNNB1* como potenciais alvos terapêuticos no tratamento contra o câncer.

Abstract

Virtually all normal cells, excluding the hematopoietic cells, require anchorage to the extracellular matrix for their proliferation and survival. When such cells are deprived of anchorage, they arrest in the G1 phase of the cell cycle and eventually undergo apoptosis. Cancer cells, on the other hand, acquire the ability to perform anchorage-independent proliferation as a result of the disruption of the regulation of the G1/S cell cycle transition by adhesion to extracellular matrix. Anchorage-independent proliferation is the foundation for tumorigenicity and metastatic capability of cancer cells. As metastases are the cause of 90% of human cancer deaths, it is crucial to decipher the molecular mechanisms underlying the regulation of the G1/S cell cycle transition by the adhesion to extracellular cell matrix. In order to decipher such mechanisms, we developed in this present work machine learning and graph theory-based computational methods for the *(i)* construction of networks representing the regulatory relationships between two biological processes of interest, *(ii)* prediction of oncogenic interactions, *(iii)* extraction of oncogenic signaling subnetworks between two genes and *(iv)* prediction of druggable genes. The network representing the regulatory relationships between G1/S cell cycle transition and adhesion to extracellular matrix, G_{ccam}, is comprised by $\approx$ 2,000 genes and $\approx$ 20,000 interactions. Moreover, $\approx$ 78% of known biological process involved in the regulation of G1/S cell cycle transition by adhesion to extracellular matrix are embedded in G_{ccam}. Through the prediction of oncogenic interactions and the extraction of oncogenic signaling subnetworks between *EGFR* and *CDC6*, genes that encode proteins likely to be relevant to anchorage-independent proliferation, we postulate the following hypotheses for the molecular mechanisms underlying the anchorage-independent proliferation: cancer cells bearing constitutively-activated EGFR are able to proliferate without adhesion to extracellular matrix partly due to CDC6 stabilization by CDKN1A. The prediction of potential druggable genes in the G_{ccam} and in the oncogenic signaling subnetworks between *EGFR* and *CDC6* revealed that the full or partial suppression of the anchorage-independent proliferation is likely to be achieved by drugs targeting proteins encoded by genes *CDKN1A*, *JUN*, *SMAD3*, *SMAD4*, *CAV1*, *CCND1* and *CTNNB1*.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução p. 15

1.1 Considerações iniciais e objetivos da tese . . . . . p. 15

1.2 Estrutura da tese . . . . . p. 20

2 Regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular, redes biológicas e aprendizado de máquina: conceitos p. 21

2.1 Regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular p. 21

2.2 Redes biológicas . . . . . p. 25

2.3 Aprendizado de máquina . . . . . p. 28

3 Análise da estrutura da rede integrada de interações gênicas envolvidas com o controle da transição G1/S pela adesão à matriz extracelular (G_{ccam}) p. 32

3.1 Introdução . . . . . p. 32

3.2 Métodos . . . . . p. 33

3.2.1 Construção da rede integrada de interações entre genes humanos (*RIGH*) p. 33

3.2.2 Construção da G_{ccam} . . . . . p. 35

3.2.3 Cálculo das medidas de centralidade . . . . . p. 39

3.3 Resultados e discussão . . . . . p. 39

3.3.1 Características gerais da *RIGH* . . . . . p. 39

3.3.2 Características gerais da G_{ccam} . . . . . p. 40

3.3.3 Análise das estruturas globais da *RIGH* e da G_{ccam} . . . . . p. 42

3.3.4 Conclusões . . . . . p. 44

4 Predição de potenciais vias de sinalização na G_{ccam} usando o *graph2sig* p. 46

4.1 Introdução . . . . . p. 46

4.2 Métodos . . . . . p. 47

4.2.1 Construção da rede biológica de interesse e cálculo das medidas de centralidade . . . . . p. 47

4.2.2 Geração dos pesos das interações . . . . . p. 48

4.2.3 Busca de caminhos potencialmente oncogênicos . . . . . p. 53

4.2.4 Construção das sub-redes de vias de sinalização . . . . . p. 54

4.3 Resultados e discussão . . . . . p. 55

4.3.1 Análise do desempenho dos preditores utilizados no *graph2sig* . . . . p. 55

4.3.2 Validação do *graph2sig*: extração da sub-rede global de vias de sinalização oncogênica . . . . . p. 57

4.3.3 Aplicação do *graph2sig* na G_{ccam} . . . . . p. 59

4.4 Conclusões . . . . . p. 63

5 Predição de alvos para drogas na G_{ccam} p. 65

5.1 Introdução . . . . . p. 65

5.2 Métodos . . . . . p. 66

5.2.1 Geração dos atributos de treinamento . . . . . p. 66

5.2.2 Construção e avaliação dos preditores . . . . . p. 67

5.2.3 Predição de novos alvos de drogas . . . . . p. 69

5.2.4 Comparações estatísticas . . . . . p. 69

5.3 Resultados e discussão . . . . . p. 70

5.3.1 Avaliação do desempenho dos preditores . . . . . p. 70

5.3.2 Predição de potenciais alvos de drogas na *RIGH* . . . . . p. 72

5.3.3 Predição de potenciais alvos de drogas na G_{ccam} . . . . . p. 74

5.4 Conclusões . . . . . p. 77

6 Considerações finais p. 79

Referências Bibliográficas p. 81

Apêndices p. 87

Apêndice A - Teste estatístico de Wilcoxon . . . . . p. 87

Apêndice B - *GeneTrail* . . . . . p. 88

Apêndice C - Análise de enriquecimento de vias do *KEGG PATHWAY* no *pred_ONCO* pelo *GeneTrail* . . . . . p. 89

Apêndice D - Trabalho publicado no periódico *Physica A* . . . . . p. 92

Apêndice E - Trabalho publicado no periódico *BMC Bioinformatics* . . . . . p. 100

Apêndice F - Trabalho publicado no periódico *BMC Genomics* . . . . . p. 119

Lista de Figuras

1.1 Esquema cíclico e iterativo de funcionamento da biologia sistêmica . . . . . p. 16

2.1 Ciclo celular em eucariotos . . . . . p. 22

2.2 Eventos moleculares durante a transição G1/S do ciclo celular . . . . . p. 23

2.3 Modelo alternativo de regulação da transição G1/S pela adesão à matriz extracelular . . . . . p. 24

2.4 Exemplos de redes . . . . . p. 25

2.5 Tipos de redes e suas distribuições de graus de conectividade e de coeficientes de agrupamento médios . . . . . p. 27

2.6 Etapas da construção de preditores . . . . . p. 29

2.7 Exemplo de árvore de decisão . . . . . p. 30

3.1 Parte da rede das relações hierárquicas entre termos da categoria *biological process* (processo biológico) do *Gene Ontology* . . . . . p. 36

3.2 Esquema de funcionamento do algoritmo *busca_cg* . . . . . p. 38

3.3 Distribuições dos graus de conectividade da *RIGH* e da G_{ccam} . . . . . p. 43

3.4 Distribuições dos coeficientes de agrupamento médios, $C(k)$, em relação à conectividade k . . . . . p. 44

4.1 Esquema representativo da construção dos grupos de treinamento no *graph2sig* p. 50

4.2 Rede hipotética para ilustrar o *REA* . . . . . p. 54

4.3 Esquema da determinação do valor de corte v no *graph2sig* . . . . . p. 55

4.4 Distribuição de frequências de interações conhecidamente oncogênicas em intervalos de valores de p_{carc} . . . . . p. 58

4.5 Coeficientes de agrupamento médios das sub-redes construídas a partir de m caminhos entre *EGFR* e *CDC6* com $nF(c)$ acima de diferentes valores . . . . p. 59

4.6 Sub-rede de vias de sinalização oncogênica entre *EGFR* e *CDC6* extraída da G_{ccam} pelo *graph2sig* . . . . . p. 60

5.1 Distribuição de frequências de genes conhecidamente drogáveis em intervalos de valores de grau de drogabilidade . . . . . p. 72

5.2 Genes conhecidamente e potencialmente drogáveis na sub-rede $EGFR-CDC6$ p. 77

Lista de Tabelas

3.1 Termos da categoria *biological process* do *GO* relacionados com a transição da fase G1 para a fase S do ciclo celular e adesão à matriz extracelular utilizados para selecionar os g_{cc} e g_{am} . . . . . p. 37

3.2 Termos GO relacionados a processos biológicos conhecidamente envolvidos com a regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular . . . . . p. 41

4.1 Vias de sinalização envolvidas em câncer no *Netpath* e no *KEGG PATHWAY* . p. 51

4.2 Medidas de desempenho dos preditores de interações oncogênicas . . . . . p. 56

5.1 Medidas de desempenho dos preditores de genes drogáveis . . . . . p. 70

5.2 Genes da *RIGH* com os 10 maiores graus de drogabilidade . . . . . p. 73

5.3 Descrição e função dos genes da *RIGH* com os 10 maiores graus de drogabilidade . . . . . p. 74

5.4 Genes da G_{ccam} com os 10 maiores graus de drogabilidade . . . . . p. 75

6.1 Valores críticos (W_c) para o teste estatístico de Wilcoxon . . . . . p. 88

1 Introdução

1.1 Considerações iniciais e objetivos da tese

Muitos comportamentos manifestados pelos sistemas ou processos biológicos e seus componentes são propriedades sistêmicas ou emergentes, isto é, propriedades que surgem a partir das interações entre os componentes. Devido a essa natureza, as propriedades emergentes não podem ser explicadas ou mesmo previstas através do estudo de cada componente individualmente (REGENMORTEL, 2004), como preconiza o reducionismo. Embora a dissecação dos sistemas biológicos em suas partes constituintes pelos métodos reducionistas tradicionais vem sendo inegavelmente eficaz e útil para 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 por um sistema biológico e seus componentes 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) (ver o Capítulo 2 para mais detalhes sobre redes). A etapa subsequente consiste na geração de um modelo a partir da análise estrutural ou dinâmica da rede construída. Dado uma certa propriedade emergente e um sistema biológico de interesse, a biologia sistêmica funciona de forma cíclica e iterativa como mostrado na Figura 1.1.

O pioneiro em afirmar que os processos biológicos eram sistemas, isto é, conjuntos de elementos interconectados, foi Ludwig von Bertalanffy. Além de determinar que os processos biológicos poderiam ser representados por redes, von Bertalanffy mostrou que as proprieda-

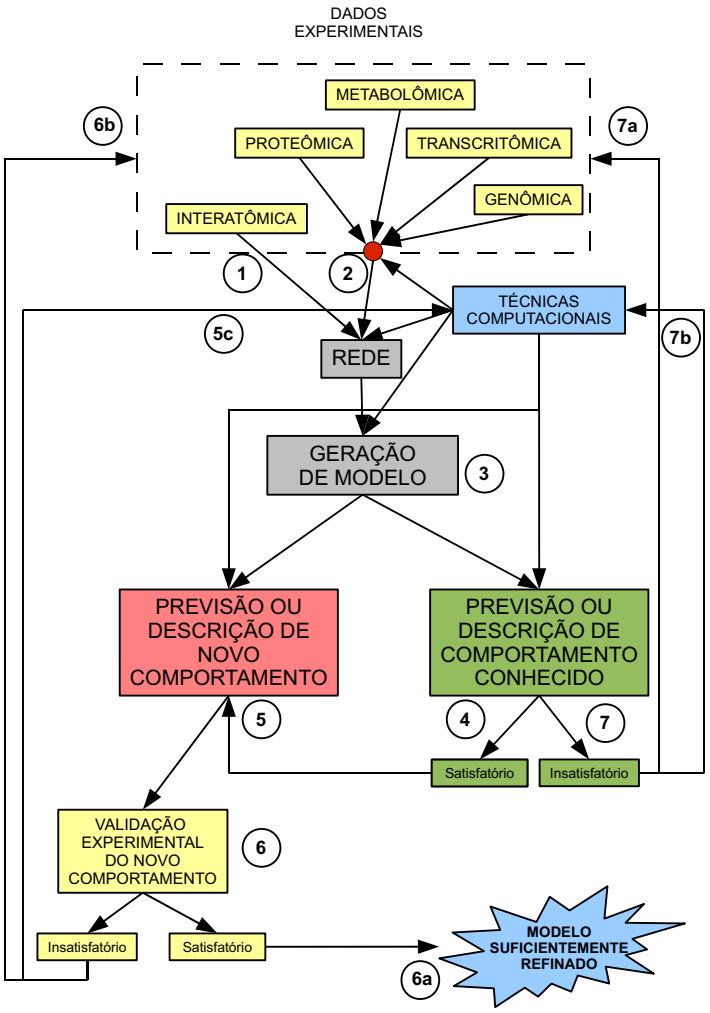

Figura 1.1: Esquema cíclico e iterativo de funcionamento da biologia sistêmica. (1) Construção da rede a partir de dados de interatoma; (2) Integração de dados gerados em estudos de genoma, transcritoma, proteoma e metaboloma; (3) Determinação da estrutura ou dinâmica da rede construída (geração de um modelo); (4) Verificação se o modelo gerado descreve ou prevê satisfatoriamente comportamentos conhecidos, isto é, surgimento experimentalmente verificado de propriedades emergentes a partir da interação de certos componentes do sistema; (5) Utilização do modelo gerado para descrição ou previsão de novos comportamentos – surgimento de propriedades emergentes a partir da interação de certos componentes do sistema ainda sem comprovação experimental – se o resultado da etapa anterior for satisfatório; (6) Verificação experimental da etapa anterior; (6a) Consolidação do modelo se a confirmação experimental do item 5 for satisfatória; (6b) Busca ou geração de mais dados experimentais para o refinamento do modelo se a confirmação experimental do item 5 for insatisfatória; (6c) Desenvolvimento de novas técnicas computacionais para o refinamento do modelo se a confirmação experimental do item 4 for insatisfatória; (7) Descrição ou previsão insatisfatória da etapa 3; (7a) Busca ou geração de novos dados experimentais; (7b) Desenvolvimento de novas técnicas computacionais.

des emergentes desses processos eram resultantes das interações entre seus componentes e que essas propriedades poderiam ser explicadas através da modelagem matemática dos processos biológicos (BERTALANFFY, 1968). Seguindo a mesma linha de von Bertalanffy, Stuart Kauffman publicou, em 1969, um trabalho onde ele propôs que as principais características dos seres

vivos, tais como o tempo de replicação de uma célula e a diferenciação celular, poderiam ser previstas ou descritas a partir da análise da dinâmica das redes de interações regulatórias entre genes conectados aleatoriamente entre si (KAUFFMAN, 1969).

Após os trabalhos pioneiros de von Bertalanffy e de Kauffman, a modelagem de interações entre componentes biológicos em forma de rede para previsão ou descrição de certas propriedades emergentes só foi retomada de forma mais vigorosa no final da década de 90. Em 1999, Bhalla e Iyengar (BHALLA; IYENGAR, 1999) modelaram em forma de rede as reações bioquímicas entre proteínas de vias de sinalização conhecidamente envolvidas, até aquele momento, no fenômeno de potenciação de longa duração. Através da análise quantitativa da rede, Bhalla e Iyengar sugeriram que a integração de sinais em diferentes escalas de tempo, a geração de respostas distintas de acordo com a intensidade e a duração dos estímulos e a formação de ciclos de retroalimentação auto-sustentáveis são propriedades emergentes das redes de vias de sinalização bioquímica (BHALLA; IYENGAR, 1999).

Outro importante trabalho que marcou a retomada da modelagem de interações entre componentes biológicos em forma de rede para prever alguma propriedade emergente foi publicado em 2001 por Jeong e colaboradores (JEONG et al., 2001). Nesse trabalho, os investigadores modelaram em forma de rede as interações físicas entre proteínas da levedura *Saccharomyces cerevisiae* e, através da análise das características estruturais dessa rede, demonstraram que a consequência fenotípica da eliminação da proteína no organismo depende de sua posição na rede. Esse trabalho foi especialmente importante por que mostrou explicitamente que as medidas de centralidade, medidas que representam numericamente a posição de um nodo na rede (mais detalhes no Capítulo 2), podem indicar a importância de um dado componente biológico em um determinado contexto. Nesse caso, uma das medidas de centralidade, o grau de conectividade, conseguiu indicar a importância de uma proteína para a sobrevivência da levedura.

Desde os trabalhos publicados por Bhalla e Iyengar (BHALLA; IYENGAR, 1999) e por Jeong e colaboradores (JEONG et al., 2001), outros milhares de trabalhos que utilizaram a estratégia da análise da estrutura ou da dinâmica da rede para a descrição ou previsão da manifestação de uma propriedade emergente por um sistema ou processo biológico ou pelos seus componentes já foram publicados. Por exemplo, com o objetivo de verificar quais proteínas são mais influentes para a deflagração da asma – nesse caso, a propriedade emergente em estudo –, Hwang e colaboradores (HWANG et al., 2008) construíram uma rede de interações contendo proteínas com reconhecida influência sobre a asma e outras proteínas cuja influência sobre a asma ainda não tinha sido determinada. A partir da análise das características estruturais da rede, Hwang e colaboradores (HWANG et al., 2008) conseguiram confirmar a influência da

maioria das proteínas conhecidamente envolvidas com asma e sugerir a existência de outras proteínas potencialmente e biologicamente relevantes para a gênese dessa doença.

Outro exemplo também recente de modelagem de interações em forma de rede para a previsão ou descrição do surgimento de propriedades emergentes é o trabalho de Barberis e colaboradores (BARBERIS et al., 2007). Nesse trabalho, foi construída uma rede de reações bioquímicas entre proteínas envolvidas na transição G1/S do ciclo celular da levedura *S. cerevisiae* e foram integradas às interações equações diferencias ordinárias descrevendo as dinâmicas das reações entre as proteínas. O modelo dinâmico gerado – um sistema de equações que descreve toda a dinâmica da rede – previu satisfatoriamente os valores do tamanho crítico da célula para início da fase S – a propriedade emergente em estudo – em diferentes condições de crescimento e confirmou que esse tamanho crítico é realmente uma propriedade emergente das interações entre os componentes da rede construída (BARBERIS et al., 2007).

Nos trabalhos citados acima, foram mostradas basicamente duas técnicas utilizadas para a previsão de propriedades emergentes a partir das interações entre os componentes do sistema: a análise das características estruturais da rede (JEONG et al., 2001; HWANG et al., 2008) e a resolução de sistema de equações diferenciais ordinárias (BHALLA; IYENGAR, 1999; BARBERIS et al., 2007). Outra técnica que tem sido utilizada para essa finalidade é 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) (Ver Capítulo 2 para uma descrição mais detalhada sobre aprendizado de máquina). Em dois trabalhos publicados recentemente pelo nosso grupo (ver Apêndices D e E) (DA SILVA et al., 2008; ACENCIO; LEMKE, 2009), foram utilizados algoritmos de aprendizado de máquina para verificar se a essencialidade de um gene é uma propriedade emergente das interações entre todos os genes dos organismos estudados. Tanto na bactéria *Escherichia coli* (DA SILVA et al., 2008) quanto na levedura *Saccharomyces cerevisiae* (ACENCIO; LEMKE, 2009), observou-se que esses algoritmos conseguem extrair padrões a partir das características estruturais da rede e prever satisfatoriamente a essencialidade de um gene a partir desses padrões. Portanto, os algoritmos de aprendizado de máquina conseguiram demonstrar que a essencialidade parece ser uma propriedade emergente das interações entre os genes dos organismos estudados.

O sucesso obtido na previsão e descrição de genes essenciais em *E. coli* e *S. cerevisiae* (DA SILVA et al., 2008; ACENCIO; LEMKE, 2009) instigou a seguinte questão: será que a utilização de aprendizado de máquina para extrair padrões a partir das características estruturais da rede poderia ser útil para a geração de hipóteses sobre o funcionamento de algum processo

biológico de interesse em humanos?

O objetivo principal do trabalho desenvolvido nesta tese foi justamente responder essa questão. Para isso, foi selecionado como processo biológico de interesse a regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular (ver detalhes no Capítulo 2). Esse processo foi selecionado por causa de sua importância estratégica no tratamento do câncer: enquanto a maioria das células normais dos organismos só se proliferam quando aderidas à matriz extracelular, as células cancerosas, de forma geral, conseguem se proliferar sem adesão. Como esse fenótipo, que pode ser considerado como resultado da regulação anormal da transição G1/S do ciclo celular pela adesão à matriz extracelular, parece ser pré-requisito para a aquisição da capacidade metastática por parte das células cancerosas (ver Capítulo 2 para mais detalhes) (CIFONE, 1982; FREEDMAN; SHIN, 1974; STEIN, 1979; MORI et al., 2009), elucidar seus mecanismos moleculares subjacentes pode ser relevante para o desenvolvimento de tratamentos mais eficazes.

Portanto, nesta tese, foram utilizadas técnicas estritamente computacionais fundamentadas em biologia sistêmica e aprendizado de máquina para gerar hipóteses sobre o funcionamento da regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular em células cancerosas. Dentro desse escopo, os objetivos específicos foram:

1. Modelagem das interações entre genes humanos na forma de uma rede integrada contendo interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional (*RIGH*);

2. Construção da subrede envolvida com o controle da transição G1/S do ciclo celular pela adesão à matriz extracelular (G_{ccam}) a partir da *RIGH*;

3. Análise das características estruturais da G_{ccam};

4. Desenvolvimento de um método para construção de sub-redes de vias de sinalização envolvidas em processos biológicos de interesse a partir de dois genes de interesse;

5. Aplicação do método desenvolvido no objetivo anterior para a identificação da sub-rede de vias de sinalização potencialmente envolvidas em câncer entre as proteína CDC6 e EGFR na G_{ccam};

6. Identificação de potenciais alvos para drogas presentes na G_{ccam}.

1.2 Estrutura da tese

Esta tese está estruturada em seis capítulos e um conjunto de Apêndices. Além do presente capítulo, que introduz conceitos básicos sobre a biologia sistêmica e os objetivos do trabalho desenvolvido, esta tese contém:

Capítulo 2: contém conceitos e fundamentos sobre o ciclo celular, redes biológicas e aprendizado de máquina; importante para o entendimento dos capítulos seguintes;

Capítulo 3: contém as descrições dos métodos e resultados referentes à modelagem das interações entre genes humanos na forma de uma rede integrada (*RIGH*), à construção da subrede envolvida com o controle da transição G1/S do ciclo celular pela adesão à matriz extracelular (G_{ccam}) e à análise da estrutura da G_{ccam};

Capítulo 4: contém introdução sobre a identificação de vias de sinalização em redes e as descrições do desenvolvimento de um novo método (*graph2sig*) para construção de sub-redes de vias de sinalização envolvidas em algum processo de interesse a partir de dois genes de interesse e da aplicação desse novo método para a extração da sub-rede de vias de sinalização potencialmente envolvidas em câncer entre as proteínas EGFR e CDC6 na G_{ccam};

Capítulo 5: contém introdução sobre as vantagens de se desenvolver um método computacional para a identificação de alvos para drogas e a descrição do método desenvolvido e sua aplicação na identificação de potenciais alvos para drogas presentes na G_{ccam};

Capítulo 6: contém conclusões e considerações finais sobre os resultados mostrados na tese e perspectivas futuras;

Apêndices: contém os detalhes sobre o teste estatístico de Wilcoxon e o *GeneTrail*, o resultado da análise de enriquecimento de vias do *KEGG PATHWAY* na sub-rede global de vias de sinalização oncogênica extraída pelo *graph2sig* em relação à *RIGH* e os trabalhos submetidos e publicados em periódicos internacionais durante o período de doutoramento.

2 *Regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular, redes biológicas e aprendizado de máquina: conceitos*

Neste Capítulo estão apresentados *(i)* os eventos moleculares que ocorrem na transição G1/S do ciclo celular conhecidamente envolvidos na aquisição da capacidade de proliferação sem adesão à matriz extracelular pelas células cancerosas, *(ii)* as evidências da existência de um sistema de regulação da transição G1/S pela adesão distinto ao já conhecido que também tornam as células cancerosas capazes de se proliferarem independentemente da adesão, sistema esse que foi o alvo de investigação neste trabalho, e *(iii)* as ferramentas utilizadas para a investigação desse sistema (redes biológicas e aprendizado de máquina).

2.1 Regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular

O ciclo celular é uma série de eventos celulares sequenciais que preparam a célula para a divisão celular. Em eucariotos, o ciclo celular é tradicionalmente dividido em quatro fases sequenciais: as fases G1, S, G2 e M. A fase G1 é quando a célula sintetiza proteínas reguladoras necessárias para a replicação do ácido desoxirribonucleico (DNA), a fase S é quando ocorre a duplicação do DNA, a fase G2 é quando a célula sintetiza proteínas e outras moléculas necessárias para a divisão celular e a fase M é quando ocorre a divisão da célula em duas células-filhas (Figura 2.1) (ALBERTS et al., 2002).

A progressão do ciclo celular ao longo das fases G1, S, G2 e M é controlada principalmente por interações entre duas classes de proteínas: as ciclinas e as quinases dependentes de ciclinas (CDKs). As ciclinas e as CDKs formam heterodímeros nos quais as ciclinas são as subunidades reguladoras e as CDKs são as subunidades catalíticas. Quando ativadas pelas ciclinas, as CDKs

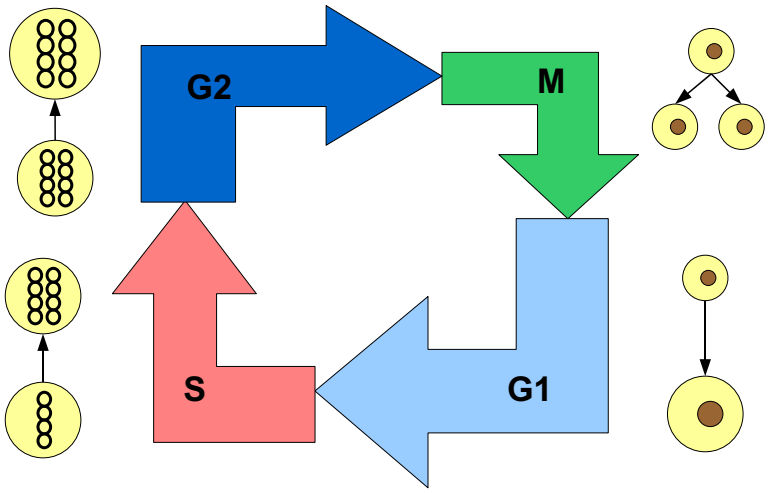

Figura 2.1: Ciclo celular em eucariotos.

regulam a transição coordenada entre uma fase e outra do ciclo celular através da ativação ou inativação de certas proteínas alvos. Combinações de diferentes tipos de ciclinas e CDKs determinam as proteínas a serem ativadas ou inativadas (ALBERTS et al., 2002).

A transição da fase G1 para a fase S ocorre, respectivamente, mediante a ativação das quinases dependentes de ciclinas CDK4 e CDK6 por ciclinas do tipo D (CCND1, CCND2 e CCND3) e pela ativação da CDK2 pelas ciclinas do tipo E (CCNE1 e CCNE2) (BALDIN et al., 1993; HENGSTSCHLAGER et al., 1999) (Figura 2.2). A CDK4 e a CDK6, quando ativadas, fosforilam proteínas da família do retinoblastoma (RB1, RBL1 e RBL2) e essa fosforilação causa a dissociação entre as proteínas da família do retinoblastoma e os complexos formados pelos fatores de transcrição da família E2F (E2F1 a E2F8) e da família DP (TFDP1 e TFDP2) (Figura 2.2). Os complexos E2F-DP, quando dissociados das proteínas do retinoblastoma, induzem a expressão de genes que codificam as proteínas CCNE1 e CCNE2 e genes essenciais para o início da fase S do ciclo celular, como os genes *CDC6*, *CCNA2*, *PCNA*, *MCM1*, *MCM2*, *MCM3*, *MCM5* e *MCM7* e *POLA1* (OBAYA; SEDIVY, 2002; HARBOUR; DEAN, 2000) (Figura 2.2). As proteínas CCNE1 e CCNE2 formam complexos com a CDK2 que mantêm as proteínas da família do retinoblastoma hiperfosforiladas para garantir a transcrição de todos os genes necessários para a entrada na fase S (Figura 2.2).

Enquanto a expressão das CDKs é constitutiva, a expressão das ciclinas, por sua vez, ocorre mediante sinais mitogênicos, isto é, sinais geralmente deflagrados pela ativação de receptores de fatores de crescimento, e sinais de adesão da célula à matriz extracelular, isto é, sinais deflagrados pelo ativação de integrinas por componentes da matriz extracelular (ASSOIAN, 1997). A ausência de adesão à matriz extracelular reprime a expressão das ciclinas do tipo D e induz

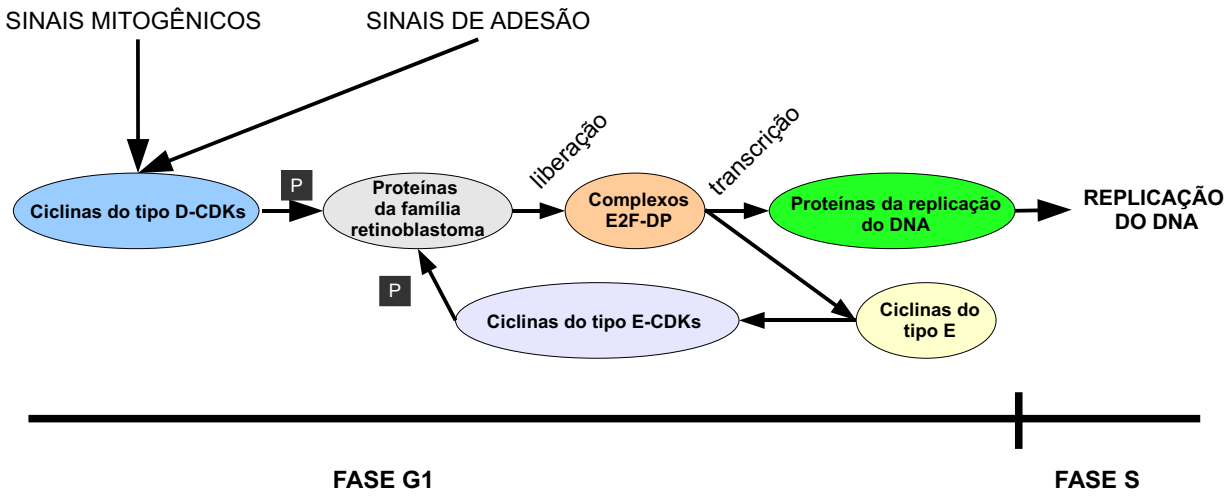

Figura 2.2: Eventos moleculares durante a transição G1/S do ciclo celular. As ciclinas do tipo D, cuja expressão é induzida conjuntamente por sinais mitogênicos e de adesão, ativam quinases dependentes de ciclinas (CDKs) que, por sua vez, fosforilam proteínas da família do retinoblastoma (RB1, RBL1 e RBL2). Fosforiladas, essas proteínas se dissociam de complexos transcricionais (complexos de proteínas da família E2F e da família DP) que, livres, ativam a transcrição dos genes que codificam ciclinas do tipo E e dos genes que codificam proteínas importantes para o início da replicação.

a expressão de um dos inibidores de CDKs, a CDKN1B (também conhecida como p27) (ZHU et al., 1996; KAWADA et al., 1997). Consequentemente, ocorre a inativação das CDKs e as proteínas da família do retinoblastoma não se dissociam dos complexos E2F-DP, fazendo com que os genes essenciais para o início da fase S do ciclo celular deixem de ser expressos. Com isso, a célula fica estacionada na fase G1 do ciclo celular (ZHU et al., 1996; KAWADA et al., 1997).

A inibição da expressão das ciclinas do tipo D, como mostrado acima, é a forma bem mais estudada de regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular. Porém, estudos têm demonstrado que há outras vias de sinalização independentes da expressão das ciclinas do tipo D e da fosforilação das proteínas do retinoblastoma envolvidas nessa regulação. Jinno e colaboradores (JINNO et al., 1999) mostraram que células normais sem adesão à matriz, mas com CDKs constitutivamente ativadas, são incapazes de transitarem para a fase S, mesmo com a dissociação dos complexos E2F-DP das proteínas da família do retinoblastoma. Quando essas células sofrem transformação oncogênica através da expressão constitutiva da proteína EGFR, porém, elas conseguem progredir para a fase S na ausência de matriz (JINNO et al., 1999).Tais resultados sugerem, portanto, que a regulação da transição G1/S pela adesão à matriz sofre influência de um sistema que é independente do eixo das ciclinas e proteínas do retinoblastoma e que é ativado por sinais oncogênicos. A existência desse sistema é reforçada pela demonstração de que a dependência da célula em fatores de crescimento para progredir para a fase S está restrita ao início da fase G1, enquanto que a dependência da

célula na adesão à matriz extracelular é necessária em toda a fase G1 (GAD et al., 2004).

O alvo principal desse sistema parece ser a proteína CDC6 (JINNO et al., 2002), uma proteína que exerce um papel crítico no início da fase S do ciclo celular ao recrutar o complexo de proteínas MCM para as origens de replicação ligadas ao complexo de proteínas ORC promovendo, portanto, o início da replicação do DNA (DEPAMPHILIS et al., 2006). Dentre os fatores essenciais para o início da fase S, foi demonstrado que somente a CDC6 teve sua expressão reprimida, tanto transcricionalmente quanto pós-traducionalmente, na ausência de adesão à matriz extracelular com expressão constitutiva concomitante das ciclinas do tipo D e, consequentemente, ativação também constitutiva das CDKs (JINNO et al., 2002). Porém, quando ocorre transformação oncogênica da célula sem adesão por ativação contínua da EGFR, esse silenciamento da expressão de CDC6 não ocorre e a célula consegue avançar para a fase S mesmo sem adesão à matriz extracelular (JINNO et al., 2002).

Esse sistema independente da expressão das ciclinas do tipo D e das proteínas da família do retinoblastoma que participa da regulação da transição G1/S pela adesão à matriz extracelular e é ativado por sinais oncogênicos deflagrados pela ativação constitutiva da EGFR (Figura 2.3) foi o alvo de investigação neste estudo dentro do escopo de proposta da geração de hipóteses sobre o funcionamento da regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular em células cancerosas.

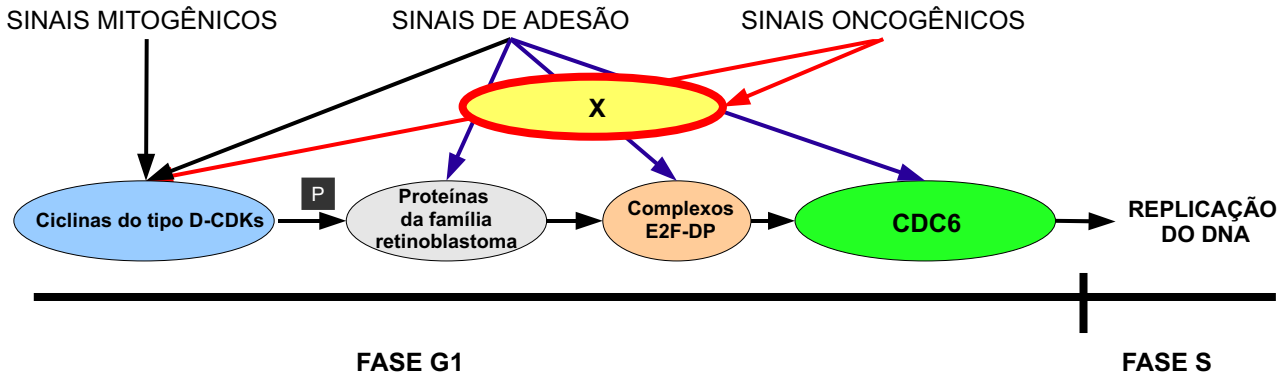

Figura 2.3: Modelo alternativo de regulação da transição G1/S pela adesão à matriz extracelular. Evidências sugerem a existência de um sistema distinto ao do eixo das ciclinas-CDKs e proteínas do retinoblastoma – sistema "X" – que é ativado por sinais oncogênicos e influencia a regulação da transição G1/S pela adesão à matriz extracelular

2.2 Redes biológicas

Os grafos ou redes são objetos matemáticos formados pelo conjunto de vértices ou nodos (componentes) e um conjunto de arestas ou interações que conectam cada dois nodos (BARABASI; OLTVAI, 2004) (Figura 2.4). Uma rede biológica é, portanto, uma representação abstrata de um sistema biológico em forma de rede onde o conjunto de nodos é o conjunto de componentes biológicos (proteínas, genes, metabólitos etc) e o conjunto de arestas é conjunto de interações de natureza biológica (interação física entre proteínas, interações metabólicas, interações de regulação transcricional etc) que conectam cada dois componentes biológicos (BARABASI; OLTVAI, 2004).

As redes podem ser não-direcionadas ou direcionadas. Redes não-direcionadas são aquelas nas quais as interações não têm uma direção definida e as direcionadas são aquelas nas quais as interações têm uma orientação bem definida (Figura 2.4). Como exemplo biológico de rede não-direcionada podemos citar as redes de interação física entre proteínas. Nessas redes, as interações entre as proteínas não têm uma orientação definida: se a proteína A se liga à proteína B, então a proteína B também se liga à proteína A. As redes metabólicas e de regulação transcricional, por outro lado, são exemplos de redes direcionadas: nas redes metabólicas, cada interação representa a direção do fluxo material de um substrato para um produto em uma reação metabólica enquanto que nas redes de regulação transcricional, cada interação representa a direção do fluxo de informação entre o fator de transcrição e seu gene alvo.

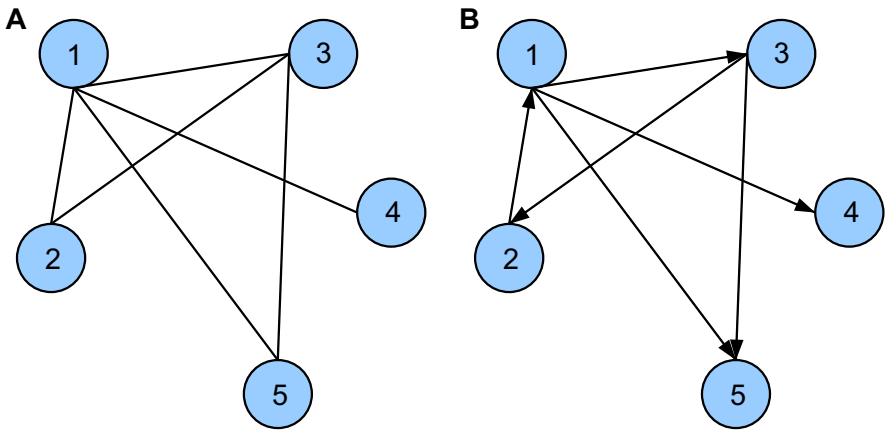

Figura 2.4: Exemplos de redes: a rede A é uma rede não-direcionada formado por um conjunto de cinco vértices ou nodos e um conjunto de seis arestas ou interações; a rede B é igual à rede A, mas suas interações são direcionadas.

A posição de um determinado nodo na rede em relação aos outros nodos pode ser determinada por funções chamadas de medidas de centralidade. As medidas de centralidade medem, como o próprio nome diz, a centralidade ou a importância de um nodo na rede. Em redes biológicas, as medidas de centralidade poderiam indicar, por exemplo, quais enzimas são as

mais importantes em uma determinada rede metabólica ou quais fatores de transcrição são os reguladores globais em uma determinada rede de regulação transcricional. As medidas de centralidade mais utilizadas na análise de redes biológicas são:

- *Grau de conectividade,* $k(g)$: número de interações que determinado nodo g possui. Caso a rede seja direcionado, há um *grau de entrada,* $k_{in}(g)$, para o número de interações direcionadas para g, e um *grau de saída,* $k_{out}(g)$, para o número de interações que partem do nodo g.

- *Coeficiente de agrupamento,* $c(g)$: proporção do número de interações entre os nodos vizinhos do nodo g e o número de todas as possíveis interações entre esses nodos:

$$c(g) = \frac{2n(g)}{k(g)[k(g)-1]} \tag{2.1}$$

  $n(g)$ é o número total de interações que os vizinhos de g possuem e $k(g)$ é seu grau de conectividade. O coeficiente de agrupamento é uma medida da coesividade local da rede (WATTS; STROGATZ, 1998).

- *Grau de intermediação,* $inbet(g)$: relação entre o número total de caminhos geodésicos – sequências de nodos adjacentes entre os nodos g_i e g_j contendo o menor número de nodos entre todas as sequências existentes entre g_i e g_j – da rede e o número de caminhos geodésicos que passam por g, ou seja:

$$inbet(g) = \sum_{g_i \neq g \neq g_j} \frac{\sigma_{g_i g_j}(g)}{\sigma_{g_i g_j}} \tag{2.2}$$

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

- *Grau de proximidade,* $cent(g)$: mede o quanto um nodo está próximo ou pode ser alcançado pelos demais nodos na rede:

$$cent(g) = \frac{n}{\sum_{g_j} d(g, g_j)} \tag{2.3}$$

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

Além de indicar a importância de um nodo, as medidas de centralidade podem ser utilizadas para revelar a estrutura global de uma rede. A análise da distribuição das frações de nodos na

rede com k interações, $P(k)$, indica, por exemplo, se a rede é aleatória (a maioria dos nodos possui o mesmo grau de conectividade, seguindo a distribuição de Poisson) ou livre de escala (enquanto a maioria dos nodos tem baixo grau de conectividade, alguns poucos nodos possuem alto grau de conectividade) seguindo uma lei de potência $P(k) = Ak^{-\gamma}$ (BARABASI; OLTVAI, 2004) (Figura 2.5).

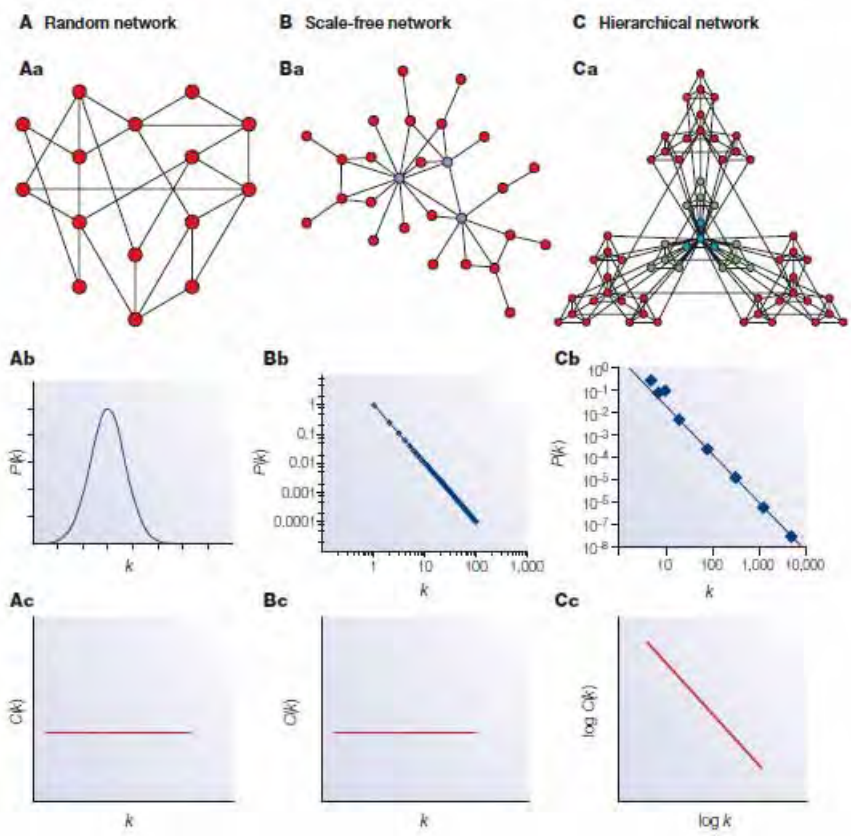

Figura 2.5: Tipos de redes e suas distribuições de graus de conectividade e de coeficientes de agrupamento médios. A: rede aleatória; B: rede livre de escala não hierárquica; C: rede livre de escala hierárquica. Figura retirada de (BARABASI; OLTVAI, 2004).

A análise da dependência dos coeficientes de agrupamento médios em relação aos graus de conectividade, $C(k)$, por sua vez, indica se o agrupamento dos nodos de uma determinada rede em módulos – sub-redes onde há mais conexões entre seus nodos do que com nodos fora dessas sub-redes – depende do grau de conectividade desses nodos, ou seja, nodos com baixo grau de conectividade tendem a formar módulos mais coesos do que nodos com alto grau de conectividade e vice-versa. Se a rede for livre de escala, o $C(k)$ indica se essa rede é modular hierárquica, isto é, os módulos são formados por módulos menores e mais coesos e assim por diante. Portanto, nessa condição de $P(k)$ livre de escala, $C(k)$ constante à medida que k aumenta indica que a rede não é hierárquica e $C(k)$ decrescente à medida que k aumenta indica que a rede é hierárquica (BARABASI; OLTVAI, 2004) (Figura 2.5).

2.3 Aprendizado de máquina

O aprendizado de máquina é uma subárea da inteligência artificial dedicado ao desenvolvimento de algoritmos, chamados de algoritmos de aprendizado (AA), que permitem ao computador extrair padrões relevantes de um conjunto de dados e utilizar esses padrões para prever ou descrever certas características de interesse. Os processos de aprendizado utilizados pelos AAs podem ser divididos em dois grupos principais: *aprendizado supervisionado*, onde o processo de aprendizado consiste na geração de um modelo de predição (ou simplesmente preditor) através do treinamento do AA com um conjunto de dados contendo exemplos conhecidos do que se pretende prever ou descrever e *aprendizado não supervisionado*, onde o AA tenta descobrir padrões a partir de algum critério de similaridade entre os dados de forma que eles possam ser agrupados (WITTEN; FRANK, 2000). Nesta seção será descrito com mais detalhes apenas o processo de aprendizado supervisionado.

No aprendizado supervisionado, de forma geral, a criação de um preditor envolve *(i)* a seleção de atributos de treinamento, isto é, características associadas às instâncias analisadas pelos AAs para o aprendizado de padrões; *(ii)* a construção de um grupo de treinamento, isto é, grupo de instâncias com classificação conhecida e seus atributos e *(iii)* seleção de um AA ou de uma combinação de AAs (Figura 2.6). Depois de gerado, o preditor é avaliado para verificar seu desempenho em classificar as instâncias pertencentes ao grupo de treinamento. Se esse desempenho for satisfatório, o preditor é então utilizado para classificar instâncias desconhecidas. De forma geral, os preditores atribuem às instâncias probabilidades estimadas de classificação em uma determinada classe i ($D(i)$); a decisão final de classificar uma dada instância em uma certa classe i depende de um valor de corte para $D(i)$.

A etapa de seleção dos atributos de treinamento depende do problema sob investigação. Se o objetivo do estudo for, por exemplo, investigar se a posição de uma proteína em uma rede de interações físicas entre proteínas é capaz de prever se essa proteína é essencial, selecionam-se medidas de centralidade como atributos de treinamento. Na etapa de construção de um grupo de treinamento, deve-se primeiramente selecionar as instâncias positivas, isto é, instâncias classificadas em uma classe de interesse, e as instâncias negativas, isto é, instâncias que não pertencem à classe de interesse. As instâncias positivas e negativas, juntamente com seus atributos de treinamento, devem ser então combinadas em um mesmo conjunto de dados – o grupo de treinamento – que é fornecido ao AA que, então, tentará encontrar padrões nos atributos de treinamento capazes de distinguir da melhor forma possível as instâncias positivas das negativas. O treinamento de um AA deve ser feito com grupos de treinamento balanceados, isto é, contendo a mesma quantidade ou quantidades semelhantes de instâncias positivas e negativas

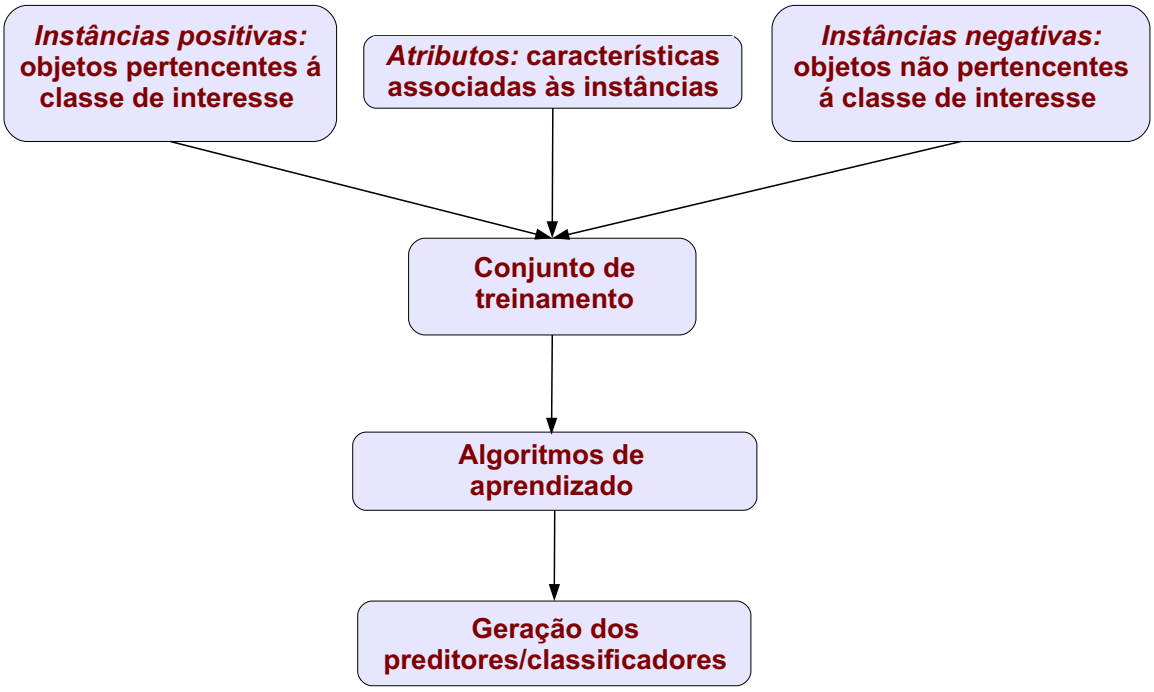

Figura 2.6: Etapas da construção de preditores.

já que, quando essa diferença é muito grande, os AAs podem encontrar dificuldades em gerar regras para a classe minoritária (VISA; RALESCU, 2005) e isso degrada o desempenho desses algoritmos.

A terceira etapa do processo de construção de um preditor consiste na seleção dos AAs a serem treinados. A abordagem tradicional para se determinar qual é o melhor AA para uma dada aplicação é selecionar aquele que fornece um modelo de predição com as melhores medidas de desempenho para essa aplicação. Outra estratégia é treinar não somente um único AA, mas, sim, uma combinação de AAs. Estudos têm mostrado que é possível obter, através dessa combinação, modelos de predição com melhores desempenhos do que modelos de predição obtidos com um único AA (LEBLANC; TIBSHIRANI, 1996; BREIMAN, 2000; OPITZ; MACLIN, 1999; POLIKAR, 2006). Isso ocorre por que padrões valiosos descobertos pelos demais AAs podem estar sendo ignorados pelo AA selecionado. A combinação de AAs, por outro lado, favorece a captura de uma maior gama de padrões úteis para a construção de modelos de predição mais precisos e sensíveis. Dentre os AAs disponíveis atualmente, os mais utilizados são as árvores de decisão, as máquinas de suporte vetorial, as redes neurais e o *naïve Bayes* (WITTEN; FRANK, 2000). Como os AAs utilizados no presente trabalho (Capítulos 3 e 4) são árvores de decisão, somente os detalhes sobre esses algoritmos serão descritos.

Os algoritmos de árvore de decisão são AAs que geram preditores cujas estruturas podem

ser visualizadas na forma de uma árvore. Cada ramo da árvore é uma condição para classificação e cada folha é uma partição do conjunto de dados com sua classificação. Dado todo o conjunto de treinamento, o algoritmo de árvore de decisão primeiro encontra a condição principal, isto é, um certo valor de algum atributo que melhor separa as instâncias de todo o conjunto de treinamento nas classes de interesse e coloca o atributo correspondente como o primeiro ramo (chamado de raiz) (Figura 2.7). O algoritmo, então, encontra a condição que melhor separa as instâncias sob a condição principal nas classes de interessante, e assim por diante, até que se chegue ao último nível, chamado de folha (Figura 2.7). Diferentes algoritmos de árvore de decisão utilizam diferentes estratégias para encontrar a condição que melhor separa as instâncias nas classes de interesse. Por exemplo: enquanto o *J48* (QUINLAN, 1993) encontra a condição que melhor separa as instâncias nas classes de interesse utilizando entropia da informação, o *logistic model tree* (LANDWEHR; HALL; FRANK, 2005) utiliza regressão logística para esse fim.

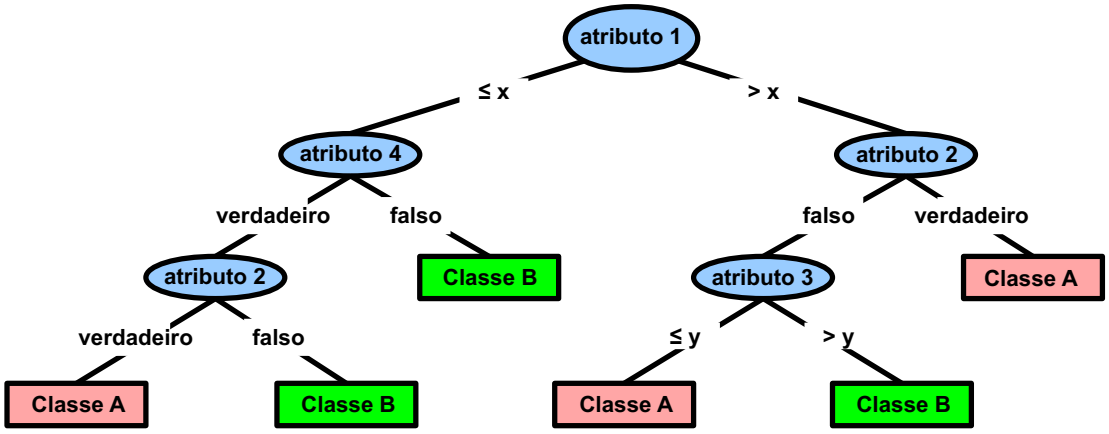

Figura 2.7: Exemplo de árvore de decisão. Neste exemplo, o algoritmo de árvore de decisão determina que a condição que melhor separa as instâncias de todo conjunto de treinamento nas classes A e B é um certo valor de x do atributo 1. Seguindo o ramo da direita, o algoritmo determina que se o atributo 2 for verdadeiro sob a condição de atributo 1 $>$ x, as instâncias podem ser classificadas em A e se o atributo 2 for falso, a classificação só pode ser feita se for considerado o valor de y do atributo 3.

Um aspecto fundamental em aprendizado de máquina é avaliação do desempenho dos preditores construídos. De forma geral, são estimadas, para esse fim, duas medidas de desempenho: precisão e sensibilidade.

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 todos as instâncias classificadas como i ($VP+FP$):

$$Precisão = \frac{VP}{VP+FP} \tag{2.4}$$

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 instâncias 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 ($VP+FN$):

$$Sensibilidade = \frac{VP}{VP+FN} \tag{2.5}$$

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

As precisões e as sensibilidades dos preditores podem ser estimadas através de várias técnicas, sendo a mais utilizada a técnica de validação cruzada. 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. O valor selecionado de v varia de acordo com o problema e com o tipo de AA utilizado. Para árvores de decisão, por exemplo, $v = 10$ por que a variação das medidas de desempenho estimadas entre os v subgrupos é menor para $v = 10$ do que para outros valores de v (KOHAVI, 1995).

3 *Análise da estrutura da rede integrada de interações gênicas envolvidas com o controle da transição G1/S pela adesão à matriz extracelular (G_{ccam})*

3.1 Introdução

No Capítulo 1, foram mencionados alguns exemplos que mostram que a modelagem de processos ou sistemas biológicos em redes pode revelar como propriedades emergentes surgem a partir das interações entre os componentes do sistema. Como o objetivo principal do trabalho apresentado nesta tese é utilizar essa estratégia para tentar elucidar alguns aspectos do controle da transição G1/S do controle celular pela adesão à matriz extracelular, a primeira etapa do trabalho é justamente a construção de uma rede potencialmente envolvida nessa regulação.

Neste Capítulo, apresentamos a construção dessa rede, batizada de G_{ccam}, e a análise de suas características estruturais e funcionais. Além disso, apresentamos também a construção e a análise das características estruturais da *RIGH*, rede integrada de interações entre genes humanos a partir da qual se extraiu a G_{ccam}. Embora a análise das características estruturais da *RIGH* não seja imprescindível para a construção e análise da G_{ccam}, ela é importante para determinar como é a estrutura de uma rede integrada contendo simultaneamente interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional já que, até o momento, esse tipo de análise só foi realizado somente com redes contendo somente um tipo de interação (BARABASI; OLTVAI, 2004).

3.2 Métodos

3.2.1 Construção da rede integrada de interações entre genes humanos (*RIGH*)

Para a construção da *RIGH*, dois genes, g_1 e g_2, que codificam, respectivamente, as proteínas p_1 e p_2, foram 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 interações físicas entre proteínas foram obtidas a partir dos seguintes bancos de dados:

- *The Biological General Repository for Interaction Datasets* (*BioGRID*) (STARK et al., 2010): banco que contém dados sobre *ppis* de vários organismos, incluindo os tipos de experimentos utilizados para a detecção das interações e os artigos que descrevem a detecção das interações. Neste banco, todas as interações são obtidas manualmente a partir de artigos publicados na literatura biomédica;

- *Database of Interacting Proteins* (*DIP*) (SALWINSKI et al., 2004): idem ao *BioGRID*, exceto pela ausência de dados sobre interações genéticas;

- *Human Protein Reference Database* (*HPRD*) (KESHAVA PRASAD et al., 2009): banco que contém informações variadas sobre proteínas humanas, todas com evidência experimental descrita na literatura, incluindo modificações pós-traducionais, localização subcelular, domínios presentes, perfil tecidual de expressão, associação com doenças e *ppis*. Assim como o *BioGRID* e o *DIP*, as *ppis* são acompanhadas dos tipos de experimentos utilizados para a detecção das interações e dos artigos que descrevem a detecção das interações;

- *IntAct* ((HERMJAKOB et al., 2004)): Idem ao *BioGRID*, mas não possui dados sobre interações genéticas;

- *Molecular Interactions Database* (*MINT*) (CHATR-ARYAMONTRI et al., 2007): Idem ao *BioGRID*, exceto pela ausência de dados sobre interações genéticas;

- *Mammalian Protein Interaction Database* (*MPPI*) do banco de dados *Munich Information Center for Protein Sequences* (*MIPS*) (PAGEL et al., 2005): Idem ao *BioGRID*, exceto por possuir somente *ppis* de mamíferos e pela ausência de dados sobre interações genéticas.

As interações de regulação transcricional foram obtidas a partir do *Transcriptional Regulatory Element Database* (*TRED*) (JIANG et al., 2007)), repositório de dados sobre interações entre fatores de transcrição e seus genes alvos de humanos, ratos e camundongos. Como no *TRED* há tanto *regs* experimentalmente verificadas quanto previstas por métodos computacionais, foram consideradas para a construção da *RIGH* somente as interações explicitamente declaradas como experimentalmente verificadas.

As interações metabólicas foram extraídas da rede metabólica humana *Recon 1* (DUARTE et al., 2007). A *Recon 1* foi construída a partir da avaliação manual de artigos sobre metabolismo humano publicados nos últimos 50 anos. Obtivemos a *Recon 1* a partir do banco de dados *BiGG* (SCHELLENBERGER et al., 2010), banco que contém, além da rede metabólica humana, outras seis redes metabólicas de outros organismos. Como a *Recon 1* é uma rede onde as interações representam as reações metabólicas e os nodos representam os metabólitos e as enzimas e, neste trabalho, definimos as interações metabólicas como os metabólitos compartilhados entre as enzimas, foi desenvolvido um código no programa *Mathematica*® 7.0 (*Wolfram Research, Inc.*) para converter os metabólitos da *Recon 1*em interações. Foram desconsideradas as interações metabólicas geradas a partir de "metabólitos de troca", isto é, metabólitos abundantes presentes nas células a maior parte do tempo e que, devido a essa presença ubíqua, não impõem limites à dinâmica das reações metabólicas (HUSS; HOLME, 2007). Foram considerados metabólitos de troca os oito metabólitos mais conectados, isto é, os metabólitos com os oito maiores graus de conectividade (ADP, ATP, H^+, H_2O, $NADP^+$, NADPH, ortofosfato e pirofosfato).

A *RIGH* final é o resultado da integração das interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional através dos genes comuns a esses conjuntos de interações. Antes da integração propriamente dita, todos os nomes dos genes humanos foram convertidos para seus *GeneIDs*, códigos identificadores únicos dos genes fornecidos pelo banco de dados *Entrez Gene* (MAGLOTT et al., 2007), para evitar a criação de falsas interações devido a nomes ambíguos. Essa conversão foi feita através de um código desenvolvido na linguagem de programação *Python* (*http://www.python.org*) com a utilização de dicionários criados a partir do arquivo "Homo_sapiens.gene_info" obtido no *Entrez Gene* (*ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/*). Esse arquivo contém, dentre ou-

tras informações, as relações entre nomes oficiais, apelidos e *GeneIDs* dos genes humanos.

3.2.2 Construção da G_{ccam}

A construção da sub-rede de interações gênicas envolvidas com o controle da transição G1/S pela adesão à matriz extracelular, G_{ccam}, baseou-se na procura dos caminhos geodésicos entre os genes anotados como participantes de processos biológicos relacionados à transição da fase G1 para a fase S do ciclo celular, g_{cc}, e adesão das células à matriz extracelular, g_{am}, na *RIGH*. O conjunto de todos os genes localizados nesses caminhos geodésicos e os g_{cc} e g_{am} formam, então, a G_{ccam}.

Essa estratégia de construção da G_{ccam} fundamenta-se na premissa de que o comprimento de um caminho geodésico entre dois genes em uma rede está inversamente correlacionado com a similaridade funcional entre esses genes. De fato, foi demonstrado que a similaridade semântica – no caso de genes, trata-se do grau de similaridade entre termos utilizados para caracterizar funcionalmente os genes – entre dois genes em uma rede diminui à medida que a distância entre esses genes aumenta (SHARAN; ULITSKY; SHAMIR, 2007; GUO et al., 2006). Portanto, dentre todos os caminhos que interligam g_{cc} e g_{am} na *RIGH*, os caminhos geodésicos são aqueles que provavelmente têm a maior proporção de genes funcionalmente semelhantes aos g_{cc} e aos g_{am}. Isso significa que os genes localizados nos caminhos geodésicos utilizados para construir a G_{ccam} tendem, portanto, a participar simultaneamente dos processos de transição G1/S e de adesão à EM.

3.2.2.1 Seleção dos g_{cc} e g_{am}

A seleção dos g_{cc} e g_{am} foi feita com a utilização do *Gene Ontology Consortium* (*GO*, *http://www.geneontology.org*) (BERARDINI et al., 2010), projeto que define termos que representam as propriedades dos genes e de seus produtos gênicos em vários organismos. Esses termos descrevem propriedades dos genes e de seus produtos gênicos, como localização subcelular, funções moleculares e processos biológicos, e estão agrupados em três diferentes categorias de acordo com cada tipo de propriedade: *cellular component* (localização subcelular), *molecular function* (função molecular) e *biological process* (processo biológico). Dentro de cada categoria, esses termos relacionam-se entre si de forma hierárquica: termos mais gerais estão no topo da hierarquia e termos mais específicos encontram-se na base da hierarquia.

A Figura 3.1 exemplifica essa relação hierárquica entre os termos do *GO* mostrando parte da estrutura hierárquica dos termos da categoria *biological process* com o termo *cell cycle*

no topo. Como se pode observar, dentro da hierarquia de termos envolvidos com processos biológicos, os termos *mitotic cell cycle* (ciclo celular mitótico), *interphase of mitotic cell cycle*(intérfase do ciclo celular mitótico) e *G1/S transition of mitotic cell cycle* (transição G1/S do ciclo celular mitótico) estão hierarquicamente abaixo do termo *cell cycle* (ciclo celular) e hierarquicamente acima dos termos *regulation of transcription involved in G1/S-phase of mitotic cell cycle* (regulação da transcrição envolvida na fase G1/S do ciclo celular mitótico), *M/G1 transition of mitotic cell cycle* (transição M/G1 do ciclo celular mitótico) e *negative regulation of mitotic cell cycle* (regulação negativa do ciclo celular mitótico).

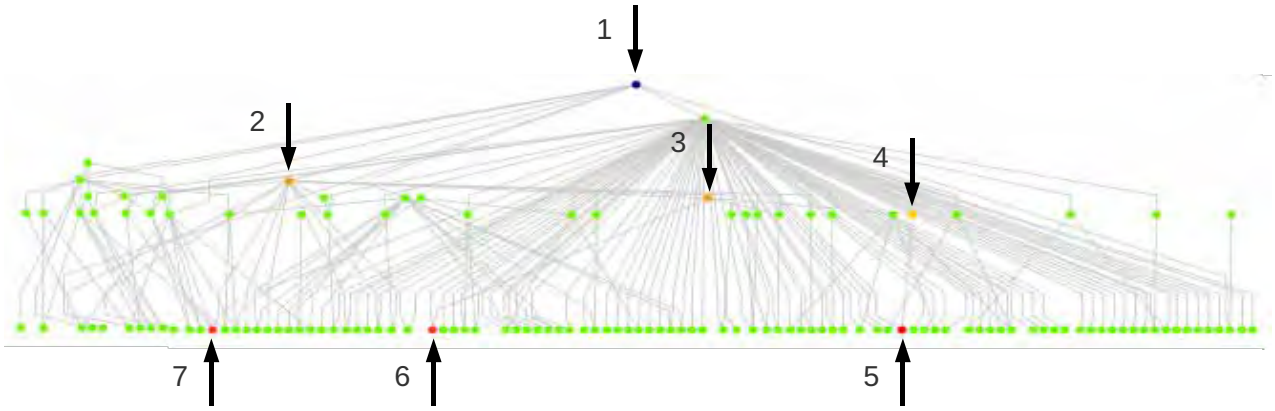

Figura 3.1: Parte da rede das relações hierárquicas entre termos da categoria *biological process* (processo biológico) do *Gene Ontology* com o termo *cell cycle* (ciclo celular) no topo da hierarquia. Nodo 1: *cell cycle* (ciclo celular); nodo 2: *mitotic cell cycle* (ciclo celular mitótico); nodo 3: *interphase of mitotic cell cycle* (intérfase do ciclo celular mitótico); nodo 4: *G1/S transition of mitotic cell cycle* (transição G1/S do ciclo celular mitótico); nodo 5: *regulation of transcription involved in G1/S-phase of mitotic cell cycle* (regulação da transcrição envolvida na fase G1/S do ciclo celular mitótico); nodo 6: *M/G1 transition of mitotic cell cycle* (transição M/G1 do ciclo celular mitótico); nodo 7: *negative regulation of mitotic cell cycle* (regulação negativa do ciclo celular mitótico)

No *GO*, cada gene está associado a pelo menos um termo de cada categoria. O gene *CDC6* humano, por exemplo, está associado a nove termos da categoria dos processos biológicos, cinco termos da categoria das localizações subcelulares e cinco termos da categoria das funções moleculares. Os genes podem estar associados a termos mais gerais ou a termos mais específicos dependendo da quantidade de dados disponíveis sobre os genes; na ausência de qualquer tipo de informação, o gene é associado ao termo mais geral de cada categoria. Por exemplo, enquanto o gene *CDC6*, cuja quantidade de dados disponíveis é grande, está associado a mais termos específicos do que gerais, o gene *TMEM62*, cuja quantidade de dados disponíveis ainda é pequena, está associado somente a termos mais gerais em cada categoria: *integral to membrane* (integral à membrana) e *membrane* (membrana) na categoria das localizações subcelulares e *biological process* e *molecular function* nas categorias dos processos biológicos e das funções moleculares, respectivamente.

A Tabela 3.1 mostra os termos da categoria *biological process* do *GO* relacionados com a

transição da fase G1 para a fase S do ciclo celular e adesão à matriz extracelular utilizados para selecionar os g_{cc} e g_{am}. Os genes da *RIGH* associados a pelo menos um dos termos apresentados na Tabela 3.1 foram considerados g_{cc} (54 genes) ou g_{am} (66 genes).

Tabela 3.1: Termos da categoria *biological process* do *GO* relacionados com a transição da fase G1 para a fase S do ciclo celular e adesão à matriz extracelular utilizados para selecionar os g_{cc} e g_{am}

Termo	Tradução livre	Código identificador do *GO*
G1/S transition of mitotic cell cycle	Transição G1/S do ciclo celular mitótico	GO:0000082
Regulation of transcription involved in G1/S-phase of mitotic cell cycle	Regulação transcricional da transição G1/S do ciclo celular mitótico	GO:0000083
Traversing start control point of mitotic cell cycle	Ponto de controle do início do ciclo celular mitótico	GO:0007089
G1/S transition checkpoint	Ponto de verificação da transição G1/S	GO:0031575
Regulation of cell adhesion mediated by integrin	Regulação da adesão celular mediada por integrinas	GO:0033628
Negative regulation of cell-substrate adhesion	Regulação negativa da adesão célula-substrato	GO:0010812
Positive regulation of cell-substrate adhesion	Regulação positiva da adesão célula-substrato	GO:0010811
Cell-matrix adhesion	Adesão célula-matriz	GO:0007160
Negative regulation of cell-matrix adhesion	Regulação negativa da adesão célula-matriz)	GO:0001953
Positive regulation of cell-matrix adhesion	Regulação positiva da adesão célula-matriz	GO:0001954
Regulation of cell-matrix adhesion	Regulação da adesão célula-matriz)	GO:0001952

3.2.2.2 Determinação dos caminhos geodésicos

Os caminhos geodésicos entre g_{cc} e g_{am} na *RIGH* foram determinados com a utilização de um algoritmo batizado de *busca_cg*. Esse algoritmo foi implementado em *Python* com base no algoritmo *predecessor* do pacote *Networkx* (HAGBERG; SCHULT; SWART, 2008). O *Networkx* é um pacote que contém centenas de algoritmos para criação, manipulação e análise da estrutura, dinâmica e funções de redes complexas, incluindo o *predecessor*, algoritmo que, dado um nodo de partida *A* e um nodo alvo *B*, realiza a busca de nodos adjacentes ao *B* localizados nos caminhos geodésicos entre *A* e *B* (Figura 3.2).

A primeira parte do *busca_cg* consiste na execução do *predecessor* tendo como nodos de partida g_{cc} e g_{am} e como nodos alvos todos os outros genes da *RIGH*. O resultado é a geração, para cada nodo de partida, de uma lista contendo os genes adjacentes aos genes alvos localizados nos caminhos geodésicos entre o g_{cc} ou g_{am} de partida e todos os outros genes da *RIGH* (Figura 3.2). A segunda parte do *busca_cg* consiste na busca sequencial, em cada uma das listas geradas

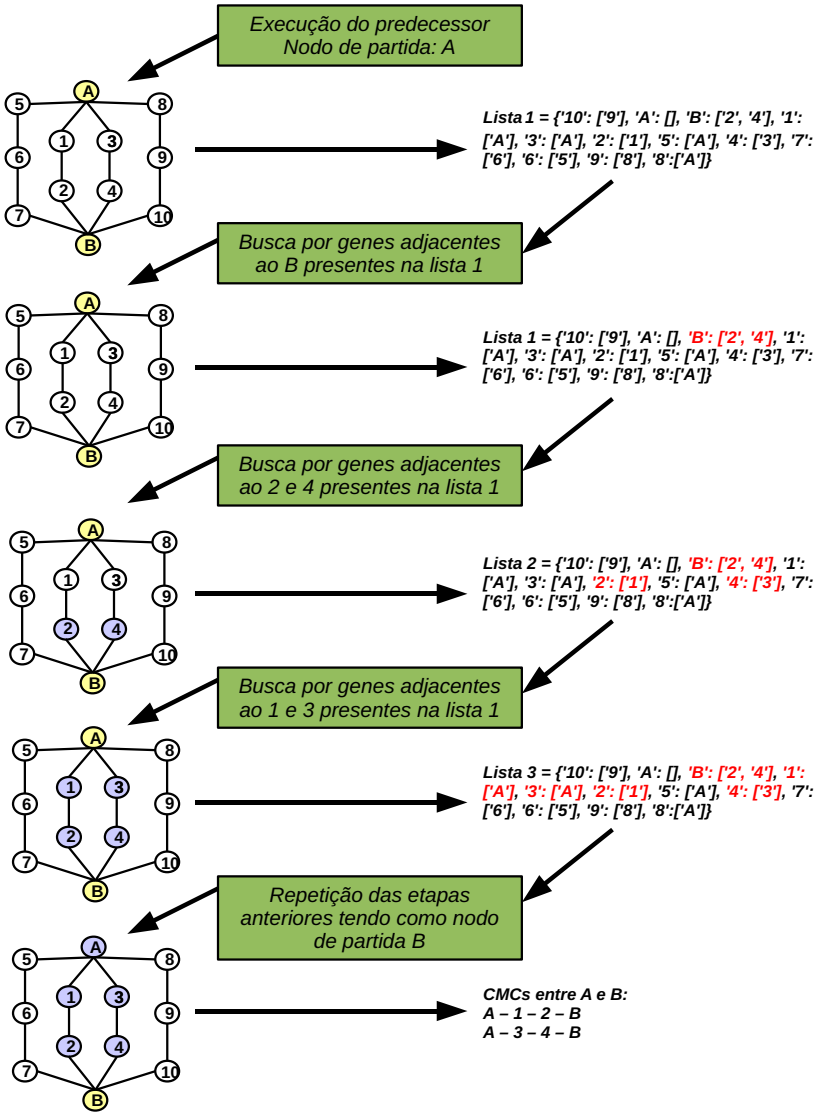

Figura 3.2: Esquema de funcionamento do algoritmo *busca_cg*. O nodo *A* representa o g_{cc} ou g_{am} de partida, o nodo *B* representa o g_{cc} ou g_{am} alvo e os outros nodos representam outros genes da *RIGH*. A primeira etapa do *busca_cg* consiste na execução do algoritmo *predecessor* do *Networkx* tendo *A* como nodo de partida e todos os outros genes como nodos alvos. Gera-se uma lista (lista 1) com os genes adjacentes localizados nos caminhos geodésicos entre *A* e todos os outros genes; a segunda etapa consiste na busca sequencial dos genes adjacentes aos adjacentes até o fechamento do caminho entre *A* e *B*: na lista 1, buscam-se os genes adjacentes a *B* (2 e 4) e os genes adjacentes a 2 e 4 (1 e 3); na lista 2, buscam-se os genes adjacentes a 1 e 3 que, nesse caso, trata-se de *A*. A inclusão de *B* nos caminhos geodésicos entre *A* e *B* ocorre quando o *busca_cg* é executado tendo como *B* como nodo de partida.

na primeira parte do algoritmo, dos genes que são adjacentes aos genes adjacentes aos genes alvos, e assim por diante, até que um outro g_{cc} ou g_{am} seja encontrado (Figura 3.2). Para cada lista, o encontro de um outro g_{cc} ou g_{am} marca o fechamento do caminho geodésico entre o g_{cc} ou g_{am} que originou a lista e os outros g_{cc} ou g_{am} da *RIGH*.

3.2.3 Cálculo das medidas de centralidade

As medidas de centralidade utilizadas para analisar as características estruturais da *RIGH* e da G_{ccam} foram os graus de conectividade, de agrupamento e de intermediação, descritos detalhadamente no Capítulo 2. Essas medidas de centralidade foram calculadas com a utilização de um programa desenvolvido no Mathematica® 7.0. Para o cálculo dos graus de intermediação, foi utilizado o pacote *NetworkX* (HAGBERG; SCHULT; SWART, 2008) para Python.

3.3 Resultados e discussão

3.3.1 Características gerais da *RIGH*

Até onde sabemos, a *RIGH* é a primeira rede de interações entre genes humanos já construída que possui simultaneamente interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional. Geralmente, os investigadores interessados em modelar qualitativamente um dado processo biológico em forma de rede o faz através da construção de uma rede de interações físicas entre proteínas. Embora têm-se obtido dados interessantes sobre o funcionamento de alguns processos biológicos usando essa abordagem, o funcionamento real dos processos biológicos implica na presença concomitante de interações entre interações físicas entre proteínas, metabólicas e de regulação transcricional. A nossa opção em construir uma rede integrada para a posterior extração de uma sub-rede de interesse fundamenta-se, portanto, nesse cenário real.

A *RIGH* em uma possui 10.161 genes e 70.932 interações. Desse total de interações, 43.169 correspondem às interações físicas entre proteínas, 24.547 correspondem às interações metabólicas e 3.012 correspondem às interações de regulação transcricional (113 fatores de transcrição regulando 1502 genes). A *RIGH* cobre cerca de 30% da quantidade estimada de genes humanos (cerca de 30.000 genes de acordo com o NCBI). Essa baixa cobertura deve-se ao fato de que foram consideradas para a construção da *RIGH* somente interações experimentalmente verificadas. Em relação aos genes que codificam fatores de transcrição, estão presentes na *RIGH* cerca de 8% de todos os fatores de transcrição humanos conhecidos que, de acordo

com Messina e colaboradores (MESSINA et al., 2004), totalizam cerca de 1500 fatores. Ainda, somente 1502 genes da rede (cerca de 15%) possuem interações de regulação transcricional. Considerando que todos os genes são controlado por pelo menos um fator de transcrição, então podemos estimar que pelo menos 8.500 interações de regulação transcricional ainda faltam ser adicionadas à rede.

3.3.2 Características gerais da G_{ccam}

A G_{ccam} construída a partir dos genes localizados nos caminhos geodésicos entre os 54 genes diretamente envolvidos com a transição da fase G1 para a fase S do ciclo celular (g_{cc}) e os 66 genes diretamente envolvidos com a adesão da célula à matriz extracelular na rede total (g_{am}), tem 2.212 genes (incluindo os genes g_{cc} e g_{am}) e 20.569 interações. Desse total de interações, 16.715 são interações físicas entre proteínas, 1.941 são interações metabólicas e 1.913 são interações de regulação transcricional (82 fatores de transcrição regulando 705 genes). A G_{ccam} possui cerca de 20% dos genes da rede original e cerca de 35% da quantidade de interações da rede original. Essa proporção relativamente alta de genes e interações presentes em relação à rede original nos sugere que os genes capturados e que formam a G_{ccam} devem participar de processos biológicos intensamente investigados pela comunidade científica, já que a rede original possui somente interações experimentalmente verificadas. A quantidade de genes na G_{ccam} com pelo menos uma interação de regulação transcricional reforça essa hipótese: 705 genes (32% do total dos 2.212 genes) possuem pelo menos uma interação de regulação transcricional contra cerca de 15% dos 10.161 genes da rede original. Além disso, dos 113 fatores de transcrição presentes na rede original, 82 (73%) deles estão presentes na G_{ccam}.

Para verificarmos se a G_{ccam} está potencialmente envolvida na regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular, nós compilamos uma lista de 41 processos biológicos conhecidamente envolvidos na regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular e os mapeamos nos termos relacionados a processos biológicos do *GO* associados aos 2.212 genes da G_{ccam}. Dos 41 processos, 32 (78%) estão representados na G_{ccam} (Tabela 3.2), o que sugere que a utilização de busca dos caminhos geodésicos entre parece ter capturado genes relevantes para a regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular.

Tabela 3.2: Termos GO relacionados a processos biológicos conhecidamente envolvidos com a regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular. Os identificadores GO e descrições correspondentes em vermelho indicam os processos presentes na G_{ccam}

Identificador GO	Descrição do processo biológico (traduzido do original em inglês)
GO:0000114	**Regulação da transcrição durante a fase G1 do ciclo celular mitótico**
GO:0000080	**Fase G1 do ciclo celular mitótico**
GO:0045737	**Regulação positiva da atividade das ciclinas dependentes de quinases**
GO:0051439	Regulação da atividade da proteína ligase de ubiquitina durante o ciclo celular mitótico
GO:0007050	**Parada do ciclo celular**
GO:0006927	**Apoptose em célula transformada**
GO:0045767	Regulação da anti-apoptose
GO:0008633	**Ativação de produtos gênicos pró-apoptóticos**
GO:0045736	**Regulação negativa da atividade das quinases dependentes de ciclinas**
GO:0000079	**Regulação da atividade das quinases dependentes de ciclinas**
GO:0051318	Fase G1
GO:0030308	**Regulação negativa do crescimento celular**
GO:0016049	**Crescimento celular**
GO:0022407	Regulação da adesão célula-célula
GO:0033631	**Adesão célula-célula mediada por integrina**
GO:0016337	**Adesão célula-célula**
GO:0007229	**Via de sinalização mediada por integrina**
GO:0007173	**Via de sinalização do receptor de fator de crescimento epidermal**
GO:004205	Regulação negativa da via de sinalização do receptor de fator de crescimento epidermal
GO:0045742	**Regulação positiva da via de sinalização do receptor de fator de crescimento epidermal**
GO:0042058	Regulação da via de sinalização do receptor de fator de crescimento epidermal
GO:0008543	**Via de sinalização do receptor do fator de crescimento de fibroblasto**
GO:0040036	**Regulação da via de sinalização do receptor do fator de crescimento de fibroblasto**
GO:0008286	**Via de sinalização do receptor de insulina**
GO:0046627	**Regulação negativa da via de sinalização do receptor de insulina**

Continua na próxima página

Tabela 3.2 – continuação

Identificador G0	Descrição do processo biológico (traduzido do original em inglês)
GO:0046628	**Regulação positiva da via de sinalização do receptor de insulina**
GO:0046626	**Regulação da via de sinalização do receptor de insulina**
GO:0048008	**Via de sinalização do receptor do fator de crescimento derivado de plaqueta**
GO:0048010	**Via de sinalização do receptor do fator de crescimento endotelial vascular**
GO:0030948	Regulação negativa da via de sinalização do receptor do fator de crescimento endotelial vascular
GO:0030949	**Regulação positiva da via de sinalização do receptor do fator de crescimento endotelial vascular**
GO:0030947	Regulação da via de sinalização do receptor do fator de crescimento endotelial vascular
GO:0007254	**Cascata do JNK**
GO:0007242	**Cascata de sinalização intracelular**
GO:0007265	**Transdução de sinal da Ras**
GO:0007179	**Via de sinalização do receptor do fator de crescimento e transformação beta**
GO:0030511	**Regulação positiva da via de sinalização do receptor do fator de crescimento e transformação beta**
GO:0017015	**Regulação da via de sinalização do receptor do fator de crescimento e transformação beta**
GO:0007181	Montagem do complexo do receptor do fator de crescimento e transformação beta
GO:0030509	**Via de sinalização da BMP**
GO:0030512	**Regulação negativa da via de sinalização do receptor do fator de crescimento e transformação beta**

3.3.3 Análise das estruturas globais da *RIGH* e da G_{ccam}

Para determinar as estruturas globais da *RIGH* e da G_{ccam}, foram analisadas suas distribuições dos graus de conectividade, $P(k)$, e suas distribuições dos coeficientes de agrupamento médios de todos os nodos com k conexões, $C(k)$.

Como pode ser observado na Figura 3.3, a *RIGH* parece pertencer às redes do tipo "livre de escala", já que seu $P(k)$ segue uma função de lei de potência, onde $P(k) = Ak^{-\gamma}$ e $\gamma \approx$

1,6 (ver Capítulo 2). Isso significa que a *RIGH* não tem um grau de conectividade médio típico que possa caracterizá-la; em vez disso, a *RIGH* possui poucos genes com altos graus de conectividade e muitos genes com baixos graus de conectividade. Essa é a primeira vez que se determina a $P(k)$ de uma rede integrada composta por interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional. Portanto, dado que seu $P(k)$ segue uma lei de potência, parece que a *RIGH* tem uma estrutura semelhante às redes contendo somente interações físicas entre proteínas ou interações metabólicas (BARABASI; OLTVAI, 2004).

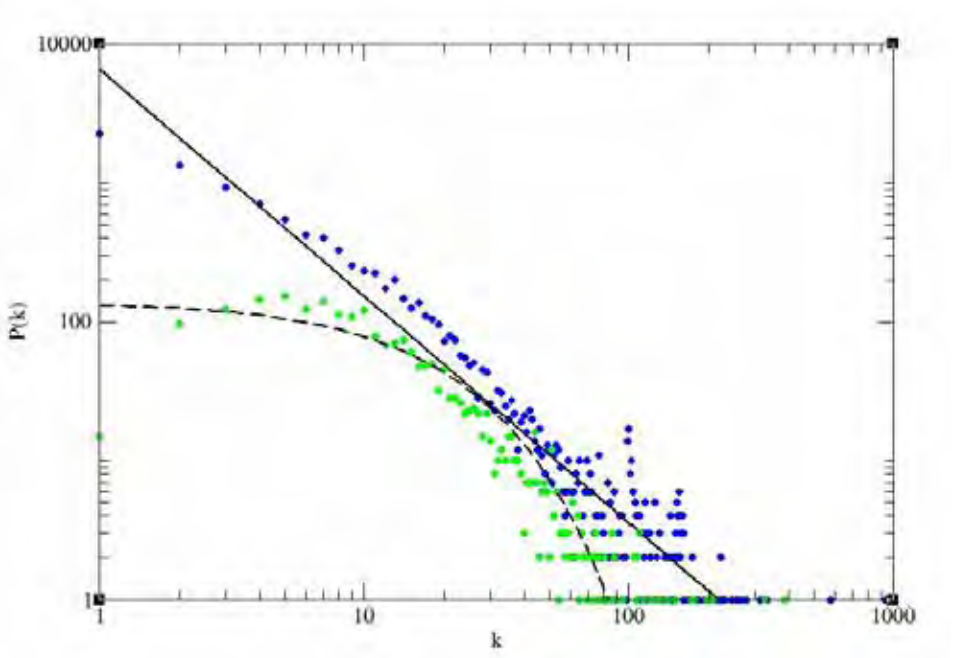

Figura 3.3: Distribuições dos graus de conectividade, $P(k)$, da *RIGH* e da G_{ccam}. A distribuição da *RIGH* (círculos azuis) parece seguir uma lei de potência $P(k) = Ak^{-\gamma}$ com $\gamma \approx 1,6$, o que a caracteriza como uma rede livre de escala. Já a distribuição da G_{ccam} (círculos verdes) não segue uma lei de potência.

A distribuição dos graus de conectividade da G_{ccam}, por sua vez, parece ter um comportamento distinto à da distribuição da *RIGH* como pode ser observado na Figura 3.3: se forem considerados valores de k entre 1 e aproximadamente 20, a $P(k)$ da G_{ccam} parece seguir uma função exponencial; se forem considerados valores de k maiores do que aproximadamente 20, a $P(k)$ parece seguir uma lei de potência. Para verificarmos se a distribuição da G_{ccam} segue ou não uma lei de potência, nós utilizamos um método estatístico desenvolvido recentemente por Clauset e colaboradores (CLUASET; SHALIZI; NEWMAN, 2009) que combina métodos de ajuste por máxima verossimilhança com testes de adequação dos ajustes baseados na estatística de Kolmogorov-Smirnov. Considerando como estatisticamente significativo o método de ajuste com $p > 0,1$, então a distribuição dos graus de conectividade da G_{ccam} tende a seguir uma lei

de potência com corte exponencial ($p = 0,9$ com ajuste para genes com $k \geq 23$) em vez de uma lei de potência ($p = 0,4$ com ajuste para genes com $k \geq 48$) ou uma exponencial ($p = 0$ com ajuste para genes com $k \geq 28$).

Em relação à distribuição dos coeficientes de agrupamento médios de todos os nodos com k conexões, $C(k)$, podemos observar na Figura 3.4 que tanto a *RIGH* quanto a G_{ccam} parecem não ser hierárquicas, já que o $C(k)$ tende a ser constante à medida que k aumenta. Esse resultado indica que o agrupamento dos genes em módulos dentro da rede não depende da quantidade de interações que eles possuem e que, embora a rede possa ser modular, tais módulos não devem se sobrepor e devem estar conectados uns aos outros por genes com alto grau de conectividade. Essa independência da $C(k)$ dos valores de k pode ser uma característica particular de redes integradas, já que, em redes biológicas contendo somente um tipo de interação, $C(k)$ depende de k e segue uma lei de potência (BARABASI; OLTVAI, 2004).

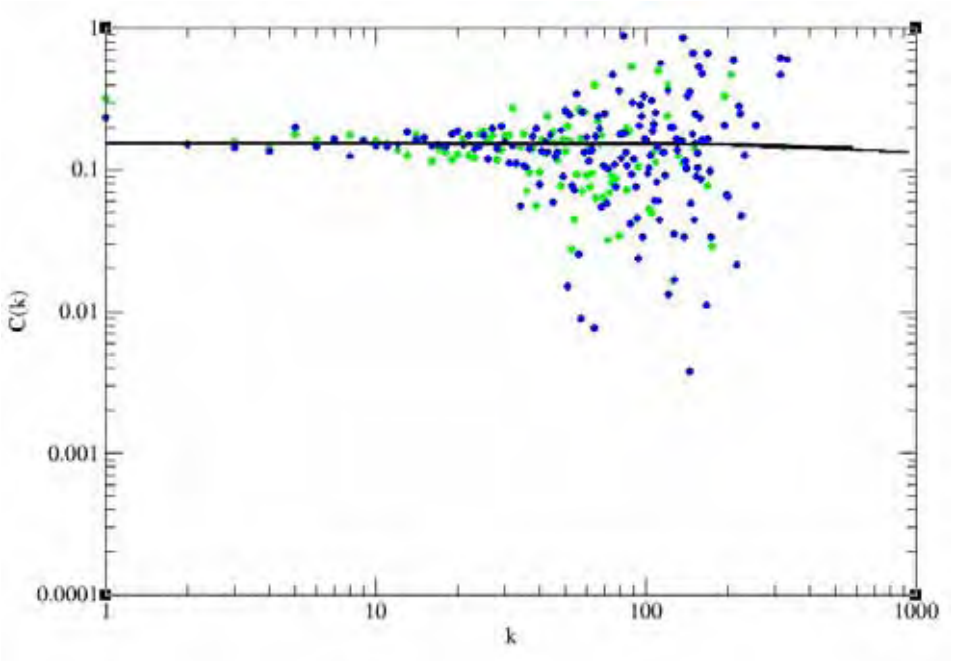

Figura 3.4: Distribuições dos coeficientes de agrupamento médios $C(k)$ em relação à conectividade k. Pode-se observar que o $C(k)$ tende a ser constante em ambas as redes (círculos azuis: *RIGH*; círculos verdes: G_{ccam}), o que indica que as duas não são hierárquicas.

3.3.4 Conclusões

Os resultados obtidos no presente capítulo permitem-nos concluir que:

- A construção de uma sub-rede que represente a regulação entre dois processos biológicos de interesse através da determinação dos caminhos geodésicos entre grupos de genes que

participam desses processos biológicos de interesse – de acordo com termos do *Gene Ontology* – é viável: $\approx 78\%$ dos termos do *Gene Ontology* relacionadas aos processos biológicos conhecidamente envolvidos na regulação da transição G1/S pela adesão á matriz extracelular estão presentes na G_{ccam};

- Embora a *RIGH* esteja incompleta, principalmente em relação às interações de regulação transcricional, a análise de sua distribuição de conectividades sugere que uma rede integrada contendo, simultaneamente, interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional, segue uma lei de potência e, portanto, pode ser classificada como livre de escala;

- A distribuição dos coeficientes de agrupamento médios de uma rede integrada parece não depender dos graus de conectividade dos genes, sugerindo, portanto, que esse tipo de rede não é hierárquica;

- A G_{ccam} não pode ser classificada como uma rede livre de escala, já que sua distribuição dos graus de conectividade tende a ser exponencial ou seguir uma lei de potência com corte exponencial;

- Assim como a *RIGH*, a G_{ccam} não é uma rede hierárquica: sua distribuição dos coeficientes de agrupamento médios parece não depender dos graus de conectividade dos genes.

4 *Predição de potenciais vias de sinalização na G_{ccam} usando o graph2sig*

4.1 Introdução

A G_{ccam} é uma rede formada pela sobreposição de interações que foram detectadas em diferentes situações aos quais as células foram expostas. Portanto, a G_{ccam} pode ser considerada uma representação global do controle da transição G1/S do ciclo celular pela adesão à matriz extracelular. Mas quais são as vias de sinalização na G_{ccam} que estão potencialmente ativas em células cancerosas? Considerando que uma via de sinalização, em uma rede biológica, é uma sequência de interações adjacentes que transmitem sinais – modificações estruturais em proteínas ou em regiões promotoras de genes que induzem modificações estruturais em outras proteínas ou o processo de transcrição – de um ponto a outro da rede, a detecção de interações adjacentes ativas na G_{ccam} em células cancerosas poderia levar à descoberta de potenciais vias de sinalização envolvidas com o controle da transição G1/S do ciclo celular pela adesão à matriz extracelular nesses tipos de células.

Como mencionado no Capítulo 2, a proliferação independente da adesão à matriz extracelular é pré-requisito para que as células cancerosas adquiram capacidade metastática (CIFONE, 1982; FREEDMAN; SHIN, 1974; STEIN, 1979; MORI et al., 2009). Logo, a detecção das vias de sinalização na G_{ccam} que estão ativas em células cancerosas é importante, por exemplo, para o desenvolvimento de novas drogas ou a utilização concomitante de múltiplas drogas já conhecidas com o intuito de controlar os membros dessas vias ativas. Ainda, a detecção de vias de sinalização na G_{ccam} ativas em células cancerosas pode indicar as prováveis vias de sinalização independentes das proteínas da família do retinoblastoma (RB1, RBL1 e RBL2) que conectam a proteína EGFR à CDC6 (ver Capítulo 2 para mais detalhes).

A abordagem comumente utilizada para a detecção de vias de sinalização ativas entre

duas proteínas de interesse em uma determinada condição a partir de uma rede biológica é a atribuição de pesos às interações que, de forma geral, são correlações entre os perfis de expressão dos genes que codificam as proteínas que interagem entre si na condição sob estudo, e posterior utilização de algoritmos de busca de caminhos em redes que levam em consideração esses pesos para extração das vias (STEFFEN et al., 2002; SCOTT et al., 2006; ZHAO et al., 2008; SUPPER et al., 2009; REN et al., 2010).

Neste Capítulo, nós descrevemos o desenvolvimento de um método baseado em aprendizado de máquina e medidas de centralidade da rede com o mesmo objetivo dos métodos descritos acima. Porém, nosso método, batizado de *graph2sig*, utiliza como peso das interações os valores de probabilidade de existência das interações no processo biológico de interesse em vez das correlações entre perfis de expressão. Ainda, o *graph2sig*, além de ser uma alternativa à detecção de vias de sinalização relacionadas aos processos biológicos cujos dados de expressão não são suficientes para a determinação de correlações, extrai não somente uma via, mas, sim, uma sub-rede de vias de sinalização entre os genes de interesse.

4.2 Métodos

O *graph2sig* consiste nas seguintes etapas: (1) construção da rede biológica de interesse; (2) cálculo de medidas de centralidade dos genes da rede; (3) utilização de aprendizado de máquina com base nas medidas de centralidade calculadas na etapa (2) para a geração de pesos para as interações da rede, isto é, geração de estimativas de probabilidades de envolvimento das interações na condição de interesse; (3) utilização de um algoritmo de busca de caminhos na rede que estejam potencialmente envolvidos com a condição de interesse que leva em consideração os pesos das interações determinados na etapa anterior e (4) união dos caminhos induzidos na etapa (3) para a formação do sub-rede.

Para testar o *graph2sig*, a rede utilizada foi a *RIGH* e a condição selecionada foi o câncer. O algoritmo de busca de caminhos em redes selecionado foi o *recursive enumeration algorithm* (REA) (JIMENEZ; MARZAL, 1999) descrito em detalhes adiante

4.2.1 Construção da rede biológica de interesse e cálculo das medidas de centralidade

A primeira etapa do *graph2sig* consiste na construção da rede biológica de interesse e, como citado anteriormente, a rede utilizada foi a *RIGH*.

A segunda etapa do *graph2sig*, isto é, o cálculo de medidas de centralidade dos genes da rede, foi realizado com base na *RIGH*. Para cada gene da *RIGH*, foram calculadas 12 medidas de centralidade, a saber: (1) grau de conectividade das interações físicas entre proteínas; (2) grau de conectividade de entrada das interações metabólicas; (3) grau de conectividade de saída das interações metabólicas; (4) grau de conectividade de entrada das interações de regulação transcricional; (5) grau de conectividade de saída das interações de regulação transcricional; (6) grau de intermediação sem distinção entre os tipos de interações; (7) grau de intermediação das interações físicas entre proteínas; (8) grau de intermediação das interações metabólicas; (9) grau de intermediação das interações de regulação transcricional; (10) grau de proximidade; (11) coeficiente de agrupamento e (12) sósias.

Cinco dessas medidas de centralidade são derivações do grau de conectividade (1 a 5) e três são derivações do grau de intermediação (7 a 9), ambas descritas em detalhes no Capítulo 2. No Capítulo 2 também estão descritos em detalhes o grau de proximidade e o coeficiente de agrupamento. A medida definida como sósias é o número de genes que possuem exatamente os mesmos valores de todas as outras medidas de centralidade. Essas medidas foram calculadas através de um programa desenvolvido no Mathematica® 7.0. Para o cálculo dos graus de intermediação, foi utilizado o pacote *NetworkX* (HAGBERG; SCHULT; SWART, 2008) para Python.

4.2.2 Geração dos pesos das interações

A terceira etapa do *graph2sig* trata-se da geração dos pesos das interações através da utilização de aprendizado de máquina. Os pesos das interações são estimativas de probabilidades de envolvimento das interações na condição de interesse (nesse caso, câncer) geradas por um modelo de predição (ou simplesmente preditor). Como mostrado no Capítulo 2, a criação de um preditor envolve *(i)* a seleção de atributos de treinamento, isto é, características associadas às instâncias (nesse caso, as interações) analisadas pelos algoritmos de aprendizado (AA) para extração de padrões; *(ii)* a construção de um grupo de treinamento, isto é, grupo de instâncias com classificação conhecida e seus atributos de aprendizado e *(iii)* seleção de um AA ou de uma combinação de AAs. Assim que o preditor é gerado, avalia-se seu desempenho e, caso esse desempenho atenda às expectativas, ele é finalmente utilizado para a geração das estimativas de probabilidade de interesse. Essas etapas estão descritas adiante.

4.2.2.1 Seleção dos atributos de treinamento

Foram utilizados como atributos de treinamento para a construção dos preditores as 12 medidas de centralidade de cada um dos genes presentes em cada uma das interações da *RIGH*.

4.2.2.2 Construção dos grupos de treinamento

Para gerar os preditores de interações oncogênicas, isto é, interações que promovem a transformação de células normais em células cancerosas ou promovem a manutenção do fenótipo das células cancerosas, foram construídos dois diferentes conjuntos de grupos balanceados de treinamento (Figura 4.1): (1) um conjunto de dez grupos de treinamento contendo as instâncias positivas e negativas (interações) corretamente associadas às suas classes que, nesse caso, são a classe "oncogênica" (*encar*) e a classe "sem oncogenicidade conhecida" (*sd_encar*) (Uma classe "não oncogênica" não pode ser criada por que, atualmente, não é possível afirmar inequivocamente que uma interação não seja oncogênica) e (2) outro conjunto de dez grupos de treinamento nos quais as classes *encar* e *sd_encar* foram aleatoriamente atribuídas às interações. O conjunto 1 foi chamado de "treinamento normal" e o conjunto 2 foi chamado de "treinamento permutado" (Figura 4.1). A construção de um grupo de treinamento no qual as classes são aleatoriamente atribuídas às instâncias serve para verificar se os AAs treinados com o grupo de treinamento normal aprendem características realmente associadas às interações oncogênicas em vez de características associadas a quaisquer subgrupos aleatórios de interações. Ainda, esses grupos de treinamento são ditos "balanceados" por que eles contêm o mesmo número de interações oncogênicas e e interações sem oncogenicidade conhecida.

A primeira etapa de construção dos grupos de treinamento foi a compilação de uma lista de interações oncogênicas, isto é, as interações classificadas com a classe *encar*. As interações oncogênicas foram extraídas a partir de duas fontes: *Netpath* (KANDASAMY et al., 2010) e *KEGG PATHWAY* (KANEHISA et al., 2008). O *Netpath* é um banco de dados que contém 20 vias de sinalização humanas manualmente extraídas da literatura, sendo 10 vias envolvidas em câncer e 10 vias envolvidas no sistema imune (KANDASAMY et al., 2010) (Tabela 4.1). O *KEGG PATHWAY* é um banco dados que contém centenas de vias metabólicas e de sinalização de vários organismos que foram manualmente extraídas da literatura. Dentro da categoria de vias de sinalização, há 14 vias de sinalização envolvidas em câncer (Tabela 4.1). Depois de extraídas, as interações foram mapeadas na *RIGH* e aquelas presentes na rede – 1.479 interações – foram utilizadas como instâncias positivas nos grupos de treinamento.

A segunda etapa da construção dos grupos de treinamento foi a compilação de uma lista de

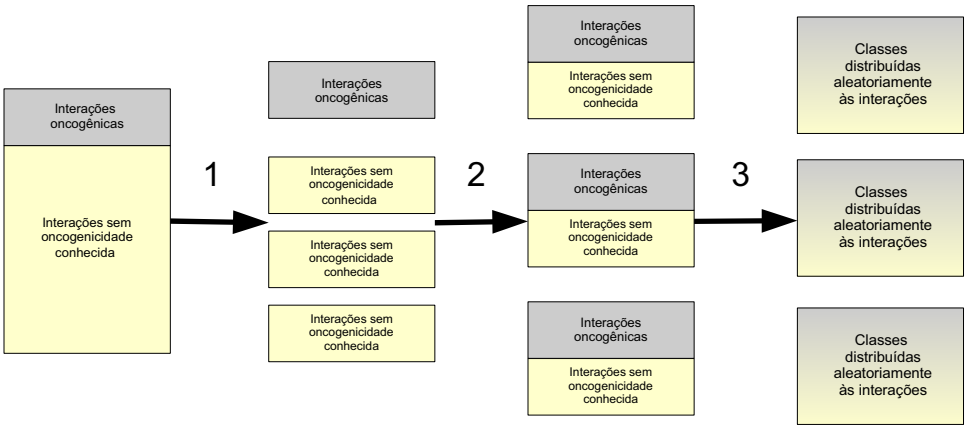

Figura 4.1: Esquema representativo da construção dos grupos de treinamento no *graph2sig*. **1**: seleção aleatória de *n* diferentes grupos de interações sem oncogenicidade conhecida (*sd_encar*); **2**: formação dos grupos de treinamento contendo a mesma quantidade de interações oncogênicas (*encar*) e *sd_encar*; **3**: atribuição aleatória das classes às interações

interações pertencentes à classe *sd_encar*. Para isso, foram consideradas interações classificadas como *sd_encar* todas as 102.047 interações que sobraram após a remoção das 1.479 interações classificadas como *encar*. Isso foi feito pois, de fato, levando em consideração os bancos de dados *Netpath e* KEGG PATHWAY, não há dados sobre envolvimento dessas 102.047 interações em câncer. A terceira etapa da construção dos grupos balanceados de treinamento foi a seleção aleatória de 10 diferentes grupos de 1.479 interações dentre as 102.047 interações classificadas como *sd_encar* e a combinação de cada um desses grupos com as 1.479 interações classificadas como *encar* formando um conjunto com 10 diferentes grupos de treinamento contendo as 1.479 interações classificadas como *encar* e 1.479 interações classificadas como *sd_encar* (Figura 4.1). O conjunto de 10 grupos de treinamento nos quais as classes *encar* e *sd_encar* foram aleatoriamente atribuídas às interações foi gerado a partir do conjunto acima.

4.2.2.3 Seleção dos algoritmos de aprendizado

Levando em consideração os estudos que têm mostrado que é possível obter, através da combinação de AAs, modelos de predição com melhores desempenhos do que modelos de predição obtidos com um único AA (LEBLANC; TIBSHIRANI, 1996; BREIMAN, 2000; OPITZ; MACLIN, 1999; POLIKAR, 2006), foi utilizada uma abordagem baseada na combinação de vários AAs para a construção do preditor de interações oncogênicas.

Usando o *WEKA* (*Waikato Environment for Knowledge Analysis*), programa escrito em JAVA desenvolvido na Universidade de Waikato, Nova Zelândia, que agrega ferramentas de

Tabela 4.1: Vias de sinalização envolvidas em câncer no *Netpath* e no *KEGG PATHWAY*

Netpath[1]		*KEGG PATHWAY*[2]	
Nome da via	**Código**	**Nome da via**	**Código**
Receptor do fator de crescimento epidermal 1 (*EGFR1)*	NetPath_4	Câncer colorretal	hsa05210
Receptor do fator de crescimento tumoral β (*TGF*β*)*	NetPath_7	Câncer pancreático	hsa05212
Fator de necrose tumoral α / Fator nuclear κ B (*TNF*α */ NF-*κ *B)*	NetPath_9	Glioma	hsa05214
Integrina $\alpha 6\beta 4$	NetPath_1	Câncer de tireoide	hsa05216
Inibidores de proteínas de ligação ao DNA (*ID)*	NetPath_5	Leucemia mieloide aguda	hsa05221
Proteínas *Hedgehog*	NetPath_10	Leucemia mieloide crônica	hsa05220
Proteínas *Notch*	NetPath_3	Carcinoma de células basais	hsa05217
Proteínas da família *Wnt*	NetPath_8	Melanoma	hsa05218
Receptor de andrógeno (*AR)*	NetPath_2	Carcinoma de células renais	hsa05211
Receptor *Kit* (antígeno *CD117)*	NetPath_6	Câncer de bexiga	hsa05219
		Câncer de próstata	hsa05215
		Câncer endometrial	hsa05213
		Câncer pulmonar de células pequenas	hsa05222
		Câncer pulmonar de células não-pequenas	hsa05223

[1] Acesso pelo *link http://www.netpath.org/browse*

[2] Acesso pelo *link http://www.genome.jp/kegg/pathway.html*

visualização e algoritmos de análise de dados, incluindo centenas de algoritmos de aprendizado de máquina (WITTEN; FRANK, 2000), foi selecionada uma combinação de sete algoritmos de aprendizado para a construção de preditores de interações oncogênicas: (1) *REPtree* (WITTEN; FRANK, 2000), (2) *random tree* (WITTEN; FRANK, 2000), (3) *random forest* (BREIMAN, 2001), (4) *J48* (QUINLAN, 1993), (5) *best-first decision tree*, (SHI, 2007, The University of Waikato), (6) *logistic model tree* (LANDWEHR; HALL; FRANK, 2005) e (7) *alternating decision tree* (FREUND; MASON, 1999). A seleção desses algoritmos foi realizado empiricamente através da análise do desempenho dos preditores gerados pela combinação de diferentes AAs disponíveis no WEKA. Foi utilizado o algoritmo *meta.Vote* (KITTLER et al., 1998), também presente no *WEKA*, para combinar as probabilidades estimadas de classificação de uma interação ser oncogênica geradas pelos sete AAs. A probabilidade estimada final de classificação de uma interação ser oncogênica, $\mu(i)$, é a média aritmética das $D(i)s$ geradas pelos AAs:

$$\mu(i) = \frac{1}{N}\sum_{n=1}^{N} D_n(i) \tag{4.1}$$

onde N é o número de algoritmos.

Os AAs selecionados e a estratégia de combinação de suas $D(i)s$ para cada instância no presente trabalho foram os mesmos utilizados pelo nosso grupo para a construção de modelos de predição de genes essenciais em *Saccharomyces cerevisiae* (ACENCIO; LEMKE, 2009) (ver Apêndice E).

Ainda, como em (ACENCIO; LEMKE, 2009), antes de utilizar o *meta.Vote*, foi aplicado a cada AA o algoritmo chamado *bootstrap aggregating* (*bagging*) (BREIMAN, 1996a). Esse método aumenta o desempenho de AAs instáveis (BREIMAN, 1996a), isto é, AAs que geram, para uma dada instância, $D(i)s$ significativamente diferentes quando apenas pequenas alterações são feitas no grupo de treinamento (BREIMAN, 1996b). Dentre os AAs considerados instáveis em (BREIMAN, 1996b), estão os de indução de árvore de decisão e os sete AAs selecionados neste estudo são todos indutores de árvores de decisão.

4.2.2.4 Avaliação dos preditores

O desempenho dos preditores de interações oncogênicas foi avaliado estimando-se a precisão e a sensibilidade desses preditores através da técnica de validação cruzada com $v = 10$. As precisões e as sensibilidades estimadas pela validação cruzada foram expressas como medianas das 10 medidas de desempenho dos 10 preditores gerados pelo treinamento da combinação de AAs pelos 10 grupos de treinamento do conjunto normal e das 10 medidas de desempenho dos 10 preditores gerados pelo treinamento da combinação de AAs pelos 10 grupos de treinamento do conjunto permutado.

4.2.2.5 Geração dos valores de potencial oncogênico

Os preditores construídos a partir dos grupos de treinamento normal e permutado foram utilizados para atribuir valores normais e permutados de potencial oncogênico (p_{carc}) a todos as interações da $RIGH$. Atribuíram-se valores permutados de p_{carc} para cada interação para verificar se os valores normais de p_{carc} não foram gerados simplesmente por que a combinação utilizada de AAs aprendeu características associadas a quaisquer subgrupos aleatórios de interações em vez de características associadas às duas classes de interações.

Para cada interação, o valor normal final de p_{carc} é a mediana dos 10 valores normais de p_{carc} atribuídas pelos 10 preditores gerados pelo conjunto de treinamento normal e o valor permutado final de p_{carc} é a mediana dos 10 valores permutados de p_{carc} atribuídas pelos 10 preditores gerados pelo conjunto de treinamento permutado. Só foram consideradas para fu-

turas análises as interações cujos valores normais e permutados de p_{carc} foram considerados estatisticamente diferentes pelo teste de Wilcoxon (ver próxima seção).

4.2.2.6 Comparações estatísticas

As comparações estatísticas entre *(i)* as medidas de desempenho estimadas dos preditores gerados pelo treinamento da combinação de AAs pelo conjunto de treinamento normal e as medidas de desempenho estimadas dos preditores gerados pelo treinamento da combinação de AAs pelo conjunto de treinamento permutado e *(ii)* os valores normal e permutado de p_{carc} para cada interação da *RIGH* foram realizadas pelo teste de Wilcoxon (WILCOXON, 1947) descrito detalhadamente no Apêndice A.

O teste de Wilcoxon foi utilizado seguindo recomendações da comunidade envolvida com aprendizado de máquina (DEMSAR, 2006). Recomenda-se o teste de Wilcoxon para fazer comparações entre medidas de desempenho de dois ou mais preditores ou entre as $D(i)$s geradas por dois ou mais preditores por que esse teste, sendo não-paramétrico, não requer o conhecimento prévio do tipo de distribuição dos dados (DEMSAR, 2006).

Como, no nosso caso, $N < 15$, já que temos 10 grupos de treinamento para cada um dos conjuntos de treinamento (normal e e permutado), foi feita a comparação entre os valores de W calculados com W_c na Tabela 6.1 no Apêndice A considerando $\alpha = 0,05$. Portanto, se $W \leq W_c$ para um dado N em $\alpha = 0,05$, as diferenças foram consideradas estatisticamente significativas.

4.2.3 Busca de caminhos potencialmente oncogênicos

Para a busca de caminhos entre dois genes de interesse na *RIGH* levando em consideração como pesos das interações os valores normais de p_{carc}, foi utilizado o *recursive enumeration algorithm* (REA) (JIMENEZ; MARZAL, 1999). Em uma rede G cujas interações estão associadas a pesos, o *REA* ordena os n caminhos entre os nodos s e v em ordem crescente de custo do caminho, sendo o custo do caminho a soma dos pesos das interações desse caminho. Por exemplo: vamos supor que se deseja ordenar os quatro caminhos menos custosos entre os nodos 1 e 5 na rede hipotética apresentada na Figura 4.2. O *REA* retorna os seguintes caminhos em ordem crescente de custo: (1) caminho $1 \rightarrow 2 \rightarrow 4 \rightarrow 5$ (custo: 9), (2) caminho $1 \rightarrow 4 \rightarrow 5$ (custo: 12), (3) caminhos $1 \rightarrow 2 \rightarrow 4 \rightarrow 3 \rightarrow 4 \rightarrow 5$ e $1 \rightarrow 2 \rightarrow 4 \rightarrow 3 \rightarrow 5$, ambos com custo 13, e (4) caminho $1 \rightarrow 2 \rightarrow 5$ (custo: 15).

No caso da *RIGH*, o custo de cada caminho c entre dois genes de interesse foi calculado

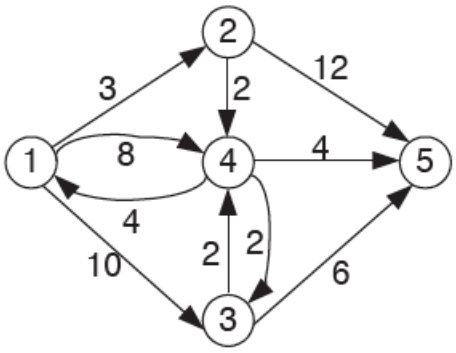

Figura 4.2: Rede hipotética para ilustrar o *REA*: rede direcionada composta por cinco nodos e 10 interações com pesos associados.

através da seguinte fórmula:

$$C(c) = \sum_{i=1}^{I} 1 - p_{carc}(i) \tag{4.2}$$

onde $C(c)$ é o custo do caminho c, $p_{carc}(i)$ é a probabilidade estimada da interação i ser oncogênica e I é o número total de interações no caminho. Utilizou-se $1 - p_{carc}(i)$ em vez do próprio p_{carc} como peso da interação pois o *REA* considera esse peso como custo e não como "força" da interação.

Foi utilizado uma implementação do *REA* na linguagem de programação C gentilmente fornecido pelo Dr. Victor M. Jimenez, criador do *REA*. Nessa implementação, o único parâmetro a ser configurado é o número n de caminhos a serem ordenados de acordo com seus valores de $C(c)$ entre os nodos de interesse. Nesse caso, o *REA* foi configurado para ordenar os 20.000 caminhos menos custosos entre os genes de interesse na *RIGH*.

4.2.4 Construção das sub-redes de vias de sinalização

As sub-redes de vias de sinalização entre dois genes de interesse, g_i e g_j, foram construídas a partir da união de m caminhos menos custosos selecionados dentre os 20.000 retornados pelo *REA* entre g_i e g_j na *RIGH* (conjunto M). Para a seleção desses m caminhos, foi determinado um valor de corte v. Porém, antes da determinação desse v, foi necessário realizar algumas transformações nos valores de $C(c)$ para que (i) os custos fossem transformados em forças e (ii) as distribuições dos valores de força dos conjuntos M pudessem ser comparadas, já que os valores mínimo e máximo de força são diferentes entre os conjuntos M.

Para a transformação dos custos em forças, foram calculados, primeiramente, os inversos dos valores de $C(c)$ ($F(c)$); esses valores de $F(c)$, então, foram normalizados de forma que, para cada M, $\max(F(c)) = 1$ e $\min(F(c)) = 0$. A normalização foi feita através da seguinte

fórmula:

$$nF(c) = \frac{F(c) - \min(F(c))}{\max(F(c)) - \min(F(c))} \tag{4.3}$$

onde, para cada c em M, $nF(c)$ é o valor de $F(c)$ normalizado, $F(c)$ é o inverso do valor de $C(c)$ calculado pela Equação 4.2 e $\max(F(c))$ e $\min(F(c))$ são, respectivamente, os valores máximo e mínimo de $F(c)$ para M.

Obtidos os valores de $nF(c)$, foi possível determinar v, ou seja, um valor de $nF(c)$ a partir do qual m caminhos foram selecionados para a construção da sub-rede. Para isso, construíram-se 20 sub-redes de vias de sinalização através da união de m caminhos com $nF(c)$ acima de 20 diferentes valores (Figura 4.3). As sub-redes formadas por menos de 10 interações foram descartadas e foram calculados os coeficientes de agrupamento médio das sub-redes restantes. Foi considerado como v o valor de $nF(c)$ a partir da qual a sub-rede formada apresenta o maior coeficiente de agrupamento médio dentre todas as outras sub-redes.

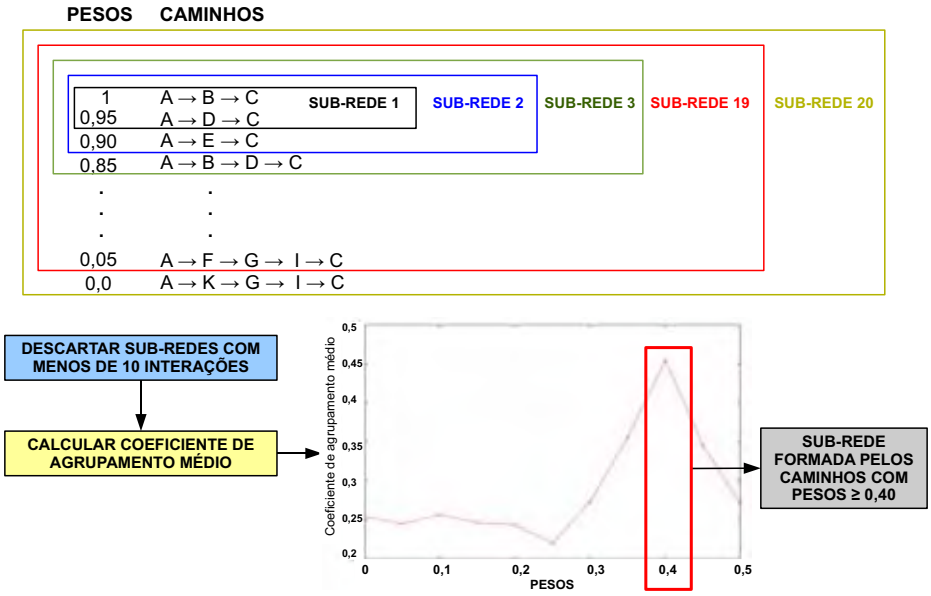

Figura 4.3: Esquema da determinação do valor de corte v no *graph2sig*. v é o valor de $nF(c)$ a partir do qual m caminhos são selecionados para a construção da sub-rede de vias de sinalização entre dois genes de interesse.

4.3 Resultados e discussão

4.3.1 Análise do desempenho dos preditores utilizados no *graph2sig*

Antes de avaliar o desempenho dos preditores de interações oncogênicas gerados a partir do conjunto de treinamento normal, comparamos estatisticamente as medianas dos valores

de sensibilidade e precisão dos 10 preditores gerados a partir do conjunto de treinamento permutado com as medianas dos valores de sensibilidades e precisão dos 10 preditores gerados a partir do conjunto de treinamento normal. Como já mencionado na seção "Métodos", esse procedimento serve para verificar se os AAs treinados com o grupo de treinamento normal detectaram padrões nos atributos de treinamento realmente associados aos genes drogáveis em vez de padrões associados a quaisquer subgrupos aleatórios de genes. Como pode-se observar na Tabela 4.2, as medianas das medidas de desempenho dos 10 preditores gerados a partir do conjunto de treinamento permutado são significativamente menores do que as medidas de desempenho dos 10 preditores gerados a partir do conjunto de treinamento normal ($W \leq W_c$ para N = 10 em $\alpha = 0,05$; ver Tabela 6.1 no Apêndice A). Esse resultado indica, portanto, que padrões extraídos dos atributos de treinamento realmente associados às interações oncogênicas foram detectados pelos preditores construídos a partir do conjunto de treinamento normal.

Tabela 4.2: Medidas de desempenho dos preditores de interações oncogênicas

Medida de desempenho	Mediana [min,max] [1]	Mediana [min,max] [1]	N	W	W_c ($\alpha = 0,05$)[2]
	Normal	Permutado			
Precisão	0,806 [0,795,0,826]	0,500 [0,483,0,509]	10	0	8 *
Sensibilidade	0,842 [0,836,0,856]	0,491 [0,474,0,541]	10	0	8 *

[1] Conjunto de 10 preditores

[2] De acordo com a Tabela 6.1 do Apêndice A

* Diferença estatisticamente significativa

Baseado no resultado acima, podemos concluir, portanto, que os valores de sensibilidade e precisão do conjunto de preditores gerados a partir do conjunto de treinamento normal observados na Tabela 4.2 refletem, de fato, o desempenho desse conjunto de preditores em discernir interações oncogênicas de interações sem oncogenicidade conhecida a partir de padrões dos dados sobre parâmetros topológicos. O desempenho desse conjunto de preditores, aliás, foi satisfatório: ele foi capaz de recuperar 84,2% das interações oncogênicas (sensibilidade) com uma precisão de 80,6% (Tabela 4.2).

A observação de que nosso conjunto de preditores não recuperou cerca de 16% das interações conhecidamente oncogênicas (interações *encar*) e classificou cerca de 20% das interações sem oncogenicidade conhecida (interações *sd_encar*) como interações *encar* indica que há existência de características comuns compartilhadas entre as interações *encar* e as interações *sd_encar*. Provavelmente, essas características comuns devem-se parcialmente à abordagem que adotamos para a seleção das interações oncogênicas: como, atualmente, não é possível construir uma lista com interações que não sejam inequivocamente oncogênicas, foram consideradas como interações não oncogênicas todos as interações da *RIGH*, exceto as 1.479 interações *encar* (ver

"Métodos"). Portanto, algumas dessas interações podem ser interações *encar* ainda desconhecidos que compartilham características em comum com as interações *encar* conhecidos. Outro fator que pode contribuir com a existência dessas características comuns compartilhadas entre as interações *encar* e *sd_encar* é o fato da *RIGH* ainda ser incompleta: Stumpf e colaboradores (STUMPF et al., 2008) estimaram, por exemplo, que a rede de interações físicas entre proteínas humanas têm cerca de 650.000 interações. A *RIGH* contém cerca de 43.000 interações físicas entre proteínas e, portanto, podemos prever que os valores de todos os parâmetros topológicos provavelmente mudarão com o aumento do tamanho da rede e isso poderá fazer com que características comuns compartilhadas entre interações *encar* e sd_encar, pelo menos aquelas relacionadas com os parâmetros topológicos, desapareçam.

Apesar das limitações discutidas acima, nosso conjunto de preditores de interações oncogênicas parece, de fato, ser confiável, como mostrado a seguir. Atribuímos a cada um das interações da *RIGH* um valor normal e outro permutado de potencial oncogênico (p_{carc}), como descrito em "Métodos", e analisamos se as interações *encar* seriam mais frequentes em intervalos de valores normais de p_{carc} mais elevados. Para isso, determinamos as distribuições de frequências das interações *encar* com valores normais e permutados de p_{carc} em 20 intervalos de valores p_{carc} (Figura 4.4). Essas frequências representam as razões entre a quantidade de interações *encar* presentes nos intervalos de valores de p_{carc} e a quantidade total dessas interações presentes na *RIGH*. Podemos observar na Figura 4.4 que as interações *encar* com valores normais de p_{carc} tendem a ser mais frequentes em intervalos com valores crescentes de p_{carc} (Figura 4.4), o que indica que o nosso conjunto de preditores conseguiu atribuir às interações *encar* valores normais elevados de p_{carc} ($\approx 79\%$ das interações receberam um $p_{carc} \geq 0.7$). Esse resultado indica, juntamente com as medidas de desempenho dos preditores, que os valores de p_{carc} gerados podem ser utilizados na extração de sub-redes de vias de sinalização oncogênica como mostrado na próxima seção.

4.3.2 Validação do *graph2sig*: extração da sub-rede global de vias de sinalização oncogênica

Para verificar se os valores de p_{carc} gerados pelos preditores são adequados para serem utilizados como pesos das interações para a extração de sub-redes de vias de sinalização oncogênica entre dois genes de interesse, nós utilizamos o *graph2sig* para extrair a sub-rede global de vias de sinalização oncogênica (*ONCO*) a partir da *RIGH*. A *ONCO*, com 1.400 interações, foi construída através da união de todas as interações conhecidamente oncogênicas, isto é, aquelas presentes nos bancos de dados *Netpath* e *KEGG PATHWAY*, mapeadas na *RIGH*. A opção em

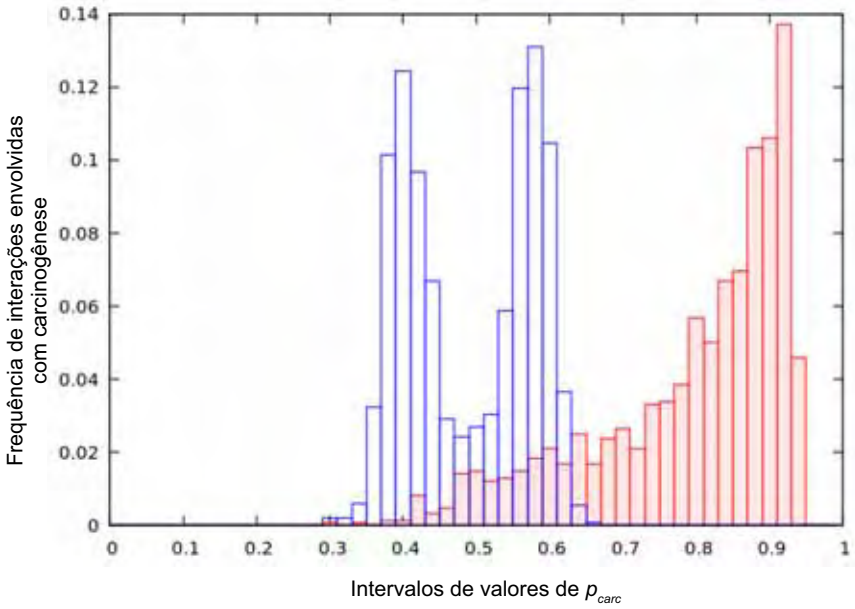

Figura 4.4: Distribuição de frequências de interações conhecidamente oncogênicas em intervalos de valores de p_{carc}. As barras coloridas em vermelho claro representam a distribuição de frequências de interações conhecidamente oncogênicas com valores normais de p_{carc} e as barras transparentes representam a distribuição de frequências de interações conhecidamente oncogênicas com valores permutados de p_{carc}. Os intervalos são de 0,02.

extrair a *ONCO* em vez de outra sub-rede menor entre dois genes específicos deve-se à plasticidade das vias de sinalização: não é possível considerar uma certa sub-rede de vias de sinalização entre dois genes específicos como a sub-rede unicamente verdadeira ou de referência.

Selecionamos todos os genes na extremidade da *ONCO*, isto é, genes que só recebem ou só transmitem sinais oncogênicos ("genes extremos") e utilizamos o *graph2sig* para extrair as sub-redes de vias de sinalização oncogênica entre todos os 1.439 pares desses genes extremos. As 1.439 sub-redes construídas foram concatenadas formando uma sub-rede final contendo 4.593 interações (*pred_ONCO*). Das 1.400 interações da *ONCO*, 933 ($\approx$ 67%) estão presentes na *pred_ONCO*. Ainda, foi realizada uma análise de enriquecimento de vias do *KEGG PATHWAY* com a utilização da versão *online* do programa *GeneTrail* (BACKES et al., 2007) (os detalhes sobre o método estão no Apêndice B) com o intuito de determinar os processos biológicos significativamente mais frequentes do que o esperado na *pred_ONCO* em comparação com a *RIGH*. Dentre as 69 vias do *KEGG PATHWAY* que são significativamente mais frequentes na *pred_ONCO* do que na *RIGH*, estão todas as 14 vias envolvidas com câncer (ver Apêndice C). Esses resultados sugerem, portanto, que o *graph2sig* parece ser um método relativamente confiável para a extração de sub-redes de vias de sinalização oncogênica entre genes de interesse.

4.3.3 Aplicação do *graph2sig* na G_{ccam}

Como um dos objetivo da presente tese é a determinação das potenciais vias de sinalização oncogênica que regulam a expressão da proteína CDC6 pela EGFR independentemente das proteínas da família do retinoblastoma e dos complexos EF2-DP na transição G1/S quando as células não estão aderidas à matriz extracelular, nós extraímos da G_{ccam}, através da utilização do *graph2sig*, a sub-rede de vias de sinalização potencialmente oncogênicas que ligam essas duas proteínas (sub-rede $EGFR-CDC6$). Foram consideradas vias potencialmente oncogênicas para a extração da $EGFR-CDC6$ por que, como já mencionado no Capitulo 2, a CDC6 é protegida contra degradação em células em suspensão através da ativação contínua da EGFR (JINNO et al., 2002), característica comum há diferentes linhagens de células cancerosas (MARMOR; SKARIA; YARDEN, 2004). Portanto, as vias de sinalização com maior potencial oncogênico que ligam a EGFR à CDC6 são aquelas potencialmente envolvidas na proteção da CDC6 contra degradação e, por consequência, no surgimento do fenótipo de transição da fase G1 para a fase S do ciclo celular sem adesão à matriz extracelular.

A sub-rede $EGFR-CDC6$ foi construída a partir da aplicação do *REA* na G_{ccam} considerando como nodo inicial o gene *EGFR* e nodo final o gene *CDC6*. Foi considerada como a sub-rede final aquela formada pelos caminhos com $nF(c) \geq 0,4$ já que, como pode ser observado na Figura 4.5, essa sub-rede é aquela com o maior coeficiente de agrupamento médio dentre as demais.

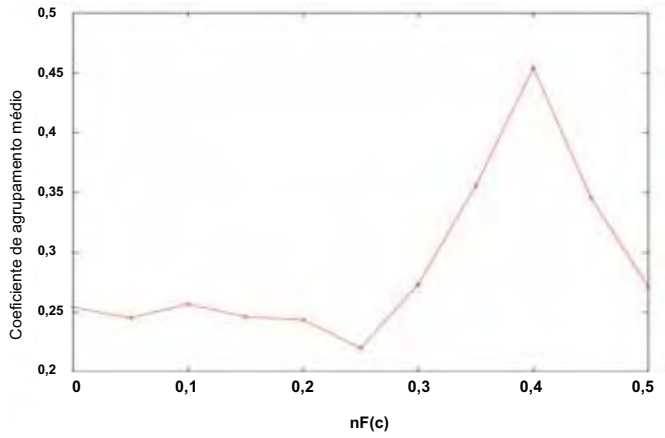

Figura 4.5: Coeficientes de agrupamento médios das sub-redes construídas a partir de *m* caminhos entre *EGFR* e *CDC6* com $nF(c)$ acima de diferentes valores.

A sub-rede $EGFR-CDC6$ (Figura 4.6) contém 21 genes e 47 interações, sendo 42 interações físicas entre proteínas e cinco interações de regulação transcricional. Das 47 interações, 17 são conhecidamente oncogênicas (*encar*) de acordo com os bancos de dados *KEGG PATHWAY* e *Netpath* e, das 30 interações ausentes no *KEGG PATHWAY* e no *Netpath* (*sd_encar*), seis

(*EGFR-AR*,*SMAD4-AR*,*YWHAQ-AR*,*SRC-AR*,*AR-SMAD4* e *NFKB1-STAT3*) pertencem ao grupo de 21 interações *sd_encar* com os 10 maiores valores de p_{carc}. A observação de que aproximadamente 36% das interações da sub-rede $EGFR-CDC6$ são *encar* e aproximadamente 30% das interações *sd_encar* com os 10 maiores valores de p_{carc} estão presentes na sub-rede $EGFR-CDC6$ sugere que essa sub-rede tem um papel importante no surgimento do fenótipo de transição da fase G1 para a fase S do ciclo celular sem adesão à matriz extracelular.

Como pode ser observado na Figura 4.6, o aspecto mais marcante da $EGFR-CDC6$ é que, enquanto o gene *EGFR* sinaliza em direção ao gene *CDC6* através de interações físicas com proteínas codificadas por nove genes (*CAV1*, *YHHAQ*, *CTNNB1*, *SRC*, *PTPN11*, *AR*, *STAT3*, *ESR1* e *RPS27A*), o *CDC6* recebe sinal oncogênico do *EGFR* através de sua única interação com $p_{carc} \geq 0,700$, que é a interação física com a proteína codificada pelo *CDKN1A*. De acordo com essa característica da $EGFR-CDC6$, podemos, então, levantar a hipótese de que a CDKN1A tem um papel importante sobre a CDC6 em células cancerosas em reposta à ativação contínua da EGFR.

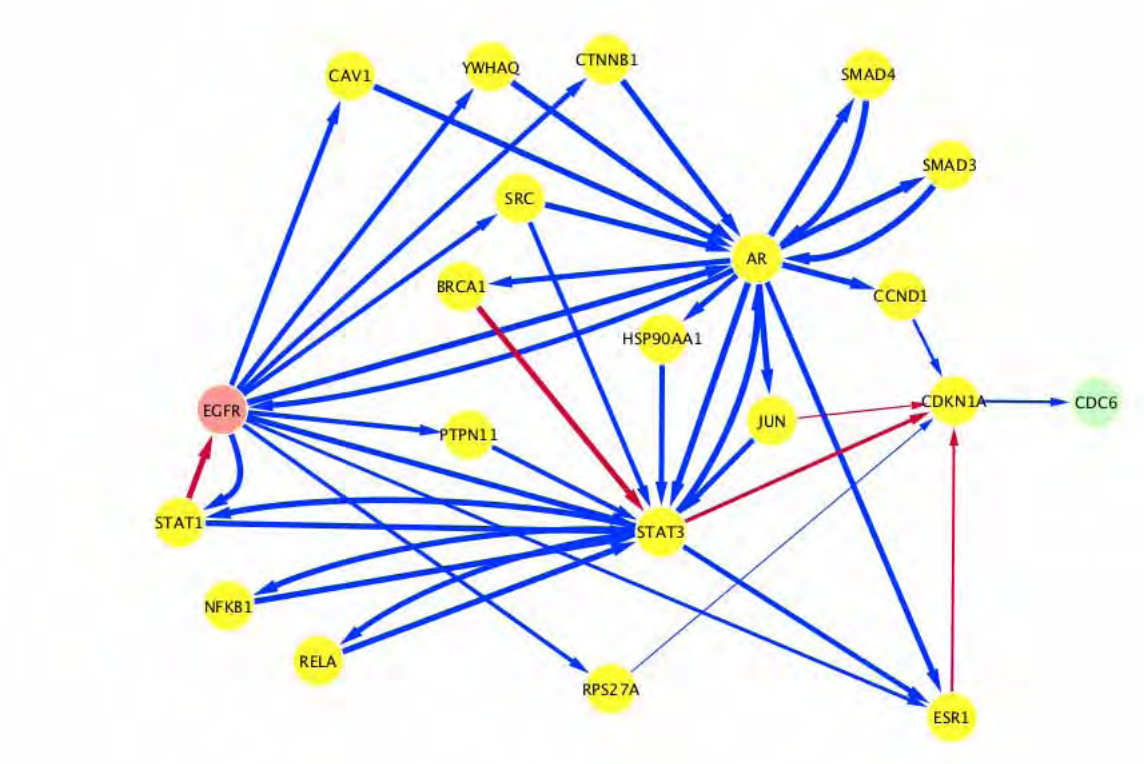

Figura 4.6: Sub-rede de vias de sinalização oncogênica entre *EGFR* e *CDC6* extraída da G_{ccam} pelo *graph2sig*. O nodos rosa e verde claro indicam, respectivamente, os nodos inicial e final no *REA*; o restante dos nodos (em amarelo) são aqueles localizados no conjunto de caminhos gerados pelo *REA* com $nF(c) \geq 0,4$. As arestas direcionadas indicam as interações entre os genes, sendo que a direção da aresta indica a direção do sinal entre os genes. As diferentes larguras das arestas indicam diferentes valores de p_{carc}, sendo as larguras crescentes correspondentes aos valores crescentes de p_{carc} (mínimo = 0,792; máximo: 0,959). As arestas em azul indicam interações físicas entre proteínas e as interações em vermelho indicam interações de regulação transcricional.

A CDKN1A, também conhecida como p21CIP1, é uma proteína multifuncional cuja função mais descrita é o controle tanto positivo quanto negativo da passagem da fase *G1* para a fase *S* do ciclo celular. Até o momento, estudos demonstraram que a CDKN1A controla negativamente a transição G1/S através da inibição da atividade dos complexos formados entre ciclinas e CDKs em geral e controla positivamente essa transição através da facilitação na formação dos complexos entre as ciclinas do tipo D e a CDK4 e CDK6 e também através da inibição da apoptose (DOTTO, 2000; BESSON; DOWDY; ROBERTS, 2008). Essas funções aparentemente antagônicas da CDKN1A se traduzem em seu envolvimento igualmente antagônico no câncer: enquanto alguns investigadores ressaltam a CDKN1A como supressora de câncer, outros detectaram superexpressão dessa proteína em diferentes tipos de cânceres (BIANKIN et al., 2001; RONINSON, 2002).

Baseado na $EGFR - CDC6$, propomos que o papel importante que a CDKN1A exerce sobre a CDC6 em células cancerosas em reposta à ativação contínua da EGFR poderia contribuir para a função oncogênica da CDKN1A através da proteção da CDC6 contra degradação na ausência de adesão à matriz extracelular. A proposta desse mecanismo fundamenta-se nas seguintes evidências descritas adiante. A primeira evidência é que, na ausência de adesão à matriz, sinais deflagrados pela interação entre as integrinas e componentes da matriz extracelular que estimulam a degradação da CDKN1A deixam de existir e, portanto, ocorre aumento de CDKN1A nas células (BOTTAZZI et al., 1999; BAO et al., 2002). A segunda evidência é que, na ausência de adesão e presença de sinais mitogênicos, o aumento de CDKN1A parece não inibir a atividade dos complexos formados pelas ciclinas do tipo E e a CDK2 (JINNO et al., 1999); essa inibição, por outro lado, ocorre na ausência de adesão e de sinais mitogênicos, o que faz com que as células não progridam para a fase S (BOTTAZZI et al., 1999). A terceira evidência é que, embora os complexos formados pelas ciclinas do tipo E e a CDK2 não sejam inibidos pela CDKN1A na ausência de adesão e presença de sinais mitogênicos, essa condição não é suficiente para que as células transitem da fase G1 para a fase S já que, mesmo com a manutenção de uma atividade constitutiva das CDKs na ausência de sinais mitogênicos, como mencionado no Capítulo 2, as células não conseguem progredir para a fase S (JINNO et al., 1999). A quarta evidência é que a CDC6 requer tanto adesão à matriz extracelular quanto presença de sinais mitogênicos para sua expressão independentemente da atividade das CDKs (JINNO et al., 2002). Portanto, na ausência de adesão à matriz e presença de sinais mitogênicos, particularmente ativação contínua da EGFR, a CDKN1A não exerceria sua função oncogênica somente sobre os complexos formados por ciclinas e CDKs mas, sim, sobre a CDC6.

Como o *CDKN1A*, por sua vez, seria regulado pelo *EGFR* de forma a garantir a estabilidade da CDC6 na ausência de adesão à matriz extracelular? Como podemos observar na Figura

4.6, os sinais oncogênicos enviados do *EGFR* para o *CDKN1A* podem trafegar por diferentes interações. Como a sub-rede $EGFR-CDC6$ pode ser considerada uma representação global do controle da transição G1/S do ciclo celular pela adesão à matriz extracelular em diferentes células cancerosas, o tráfego através dessas interações pode, então, ser diferente em tipos distintos de células cancerosas. Portanto, o *CDKN1A* poderia ser regulado de diferentes formas por diferentes tipos de células cancerosas. Por exemplo, em células cancerosas de próstata, o *CDKN1A* poderia ser regulado por sinais oncogênicos que passariam pelo *AR* (Figura 4.6), gene que codifica o receptor de andrógeno. Há evidências indiretas para essa regulação: Shigemura e colaboradores (SHIGEMURA et al., 2009) mostraram que a ativação direta ou indireta do receptor de andrógeno pela EGFR aumenta a capacidade de proliferação independente de adesão à matriz extracelular das células cancerosas prostáticas. Em outros tipos de células cancerosas, por exemplo, o *CDKN1A* poderia ser regulado por sinais que passariam pelo *STAT3* (Figura 4.6), gene que codifica um fator de transcrição que regula a expressão de vários genes em resposta a variados estímulos. O estudo realizado por Barbieri e colaboradores (BARBIERI et al., 2010) mostra uma evidência indireta dessa regulação: a proteína codificada pelo *STAT3* parece necessária para a aquisição da capacidade de proliferação sem adesão à matriz extracelular por células cancerosas de mama quando essas células superexpressam a proteína ERBB2, proteína que estabiliza e mantém a EGFR em contínua atividade (SCHAFER et al., 2009; GRASSIAN; SCHAFER; BRUGGE, 2011).

A análise exaustiva das várias possibilidades de caminhos utilizados por diferentes tipos de células cancerosas para a transmissão dos sinais oncogênicos entre o *EGFR* e o *CDKN1A* poderia nos levar a encontrar evidências sobre o envolvimento de cada uma delas com a proliferação independente de adesão à matriz extracelular. Porém, como cada tipo de célula cancerosa pode usar um certo grupo particular de interações para a sinalização oncogênica entre o *EGFR* e o *CDKN1A*, essa análise não seria útil para encontrar genes em comum que estariam envolvidos na transição G1/S sem adesão à matriz extracelular em todos os tipos de cânceres. Nesse sentido, é notável observar que todos esses sinais oncogênicos, independentemente do tipo de célula cancerosa, convergem no *CDKN1A* (Figura 4.6), indicando que a estabilização da CDC6 pela CDKN1A na ausência de adesão à matriz extracelular, hipótese que levantamos através da análise da sub-rede $EGFR-CDC6$, pode ser um evento comum a todos os tipos de células cancerosas com EGFR permanentemente ativo e até mesmo o fator preponderante que tornam essas células cancerosas aptas à proliferação sem adesão à matriz extracelular.

É interessante observar que, dentre as interações pelas quais os sinais oncogênicos podem trafegar entre o *EGFR* e os genes *CDKN1A* e *CDC6* na sub-rede $EGFR-CDC6$, não há nenhuma envolvida na via clássica de controle da transição G1/S, ou seja, a liberação de comple-

xos E2F-DP das proteínas da família do retinoblastoma (RB1, RBL1, RBL2) inativadas pelas CDK2, CDK4 e CDK6 ativadas por ciclinas dos tipos D e E. Essa observação reforça a ideia de que o controle da transição G1/S é uma propriedade emergente que surge a partir de uma rede intrincada de interações gênicas e que, em células cancerosas, esse controle pode ser prioritariamente realizado por conjuntos de interações distintos aos da via clássica, como sugerido por alguns estudos experimentais (JINNO et al., 1999, 2002; GAD et al., 2004).

4.4 Conclusões

Neste Capítulo, descrevemos o desenvolvimento e os resultados de um método baseado em medidas de centralidade de redes e aprendizado de máquina para a extração de sub-redes de vias de sinalização envolvidas em um certo processo biológico de interesse entre dois genes localizados em uma rede. Batizado de *graph2sig*, esse método foi testado na extração de sub-rede conhecida de vias de sinalização oncogênica na *RIGH* e também na extração, a partir da G_{ccam}, da sub-rede de vias de sinalização oncogênica entre os genes *EGFR* e *CDC6* para a descoberta de potenciais vias de sinalização envolvidas na regulação da proteína CDC6 pela adesão à matriz extracelular em células cancerosas.

Levando em consideração os limites impostos pela incompletude da *RIGH*, como discutido no Capítulo 3 e na seção "Análise do desempenho dos preditores utilizados no *graph2sig*" deste Capítulo, os resultados obtidos nos permitem tecer as seguintes conclusões:

- O conjunto das medidas de centralidade dos dois genes que participam de uma interação são atributos de treinamento capazes de gerar preditores de interações oncogênicas já que *(i)* a sensibilidade mediana desses preditores foi de aproximadamente 84%, isto é, esses preditores conseguiram recuperar 84% de todas as interações envolvidas em carcinogênese, e *(ii)* os valores normais de p_{carc} atribuídos às interações conhecidamente oncogênicas se distribuem com maior frequência em intervalos de valores de $p_{carc} \geq 0,700$;

- O *graph2sig* é capaz de extrair sub-redes de vias de sinalização potencialmente oncogênicas: a sub-rede entre os genes *EGFR* e *CDC6*, embora ainda não descrita em nenhum banco de dados de vias de sinalização, como o *KEGG PATHWAY* e o *Netpath*, possui interações oncogênicas que têm evidências na literatura biomédica;

- A sub-rede $EGFR-CDC6$ parece potencialmente envolvida no controle da expressão da CDC6 pelos sinais oncogênicos deflagrados pela EGFR em células cancerosas;

- A interação entre as proteínas codificadas pelos genes *CDKN1A* e *CDC6* parece ser importante para que células cancerosas com EGFR constitutivamente ativo adquiram a capacidade de proliferação na ausência de adesão à matriz extracelular: os sinais oncogênicos deflagrados pela EGFR chegam à CDC6 através de sua interação com a CDKN1A.

5 *Predição de alvos para drogas na G_{ccam}*

5.1 Introdução

A utilização de drogas para modular o controle da transição G1/S do ciclo celular pela adesão à matriz extracelular poderia ser uma estratégia interessante para impedir a formação de tumores metastáticos. Como citado no Capítulo 2, a proliferação independente da adesão à matriz extracelular é pré-requisito para que as células neoplásicas adquiram capacidade metastática (CIFONE, 1982; FREEDMAN; SHIN, 1974; STEIN, 1979; MORI et al., 2009). Ao cruzarmos a G_{ccam} com a rede de interações entre drogas aprovadas pela FDA (agência reguladora de medicamentos e alimentos nos Estados Unidos) e seus alvos construída por Yildirim e colaboradores (YILDIRIM et al., 2007), verificamos que 103 dos 2.212 genes presentes na G_{ccam} codificam proteínas que já são alvos terapêuticos para diversas doenças, sendo que 38 são alvos terapêuticos específicos para o tratamento de câncer. É possível que uma parte dos restantes 2.174 genes da G_{ccam} ainda não associados a nenhuma droga codifiquem alvos de drogas anti-câncer e a descoberta desses genes poderia expandir as possibilidades de tratamento de neoplasias.

A identificação experimental em larga escala de genes que codificam proteínas alvos para drogas, a partir daqui denominados como "genes drogáveis", envolve a execução de técnicas variadas de genômica, proteômica e genética direta e reversa (LINDSAY, 2003) que demandam tempo e são muito custosas. Com o intuito de tornar essa tarefa mais rápida e mais barata, o desenvolvimento de uma técnica computacional capaz de prever acuradamente genes drogáveis em larga escala seria valiosa. A disponibilização pública de uma vasta quantidade de dados abrangendo, por exemplo, interações físicas entre proteínas, interações metabólicas, interações regulatórios, expressão gênica em larga escala e sublocalização celular dos produtos gênicos, dentre outros, torna possível o desenvolvimento um método computacional baseado em aprendizagem de máquina para extrair padrões desses dados que poderiam, então, ser usa-

dos como preditores de genes drogáveis em escala genômica. Utilizando essa abordagem, nosso grupo desenvolveu, recentemente, um método baseado em aprendizagem de máquina para extrair padrões de dados variados e então aplicar esses padrões para diferenciar genes essenciais de genes não-essenciais em escala genômica em *Escherichia coli* (DA SILVA et al., 2008) e *Saccharomyces cerevisiae* (ACENCIO; LEMKE, 2009) (Ver Apêndices D e E).

Como o desempenho desse método descrito acima foi satisfatório na predição de genes essenciais em *Escherichia coli* e *Saccharomyces cerevisiae*, decidimos verificar se um método semelhante poderia ser útil na predição de genes drogáveis em seres humanos. Neste capítulo, descrevemos a criação de um preditor de genes drogáveis em escala genômica baseado em aprendizagem de máquina que utiliza, como atributos de treinamento, dados de interações, expressão gênica em larga escala e sublocalização celular dos genes. Ainda, descrevemos os resultados dos testes de avaliação desse preditor e sua aplicação para a previsão de potenciais genes drogáveis na *RIGH* e na G_{ccam}. Todos os resultados apresentados aqui, exceto a previsão de genes potencialmente drogáveis na G_{ccam}, fazem parte de um artigo recentemente publicado no periódico *BMC Genomics* (COSTA; ACENCIO; LEMKE, 2010) (ver Apêndice F).

5.2 Métodos

As etapas para a criação de preditores de genes drogáveis são as mesmas utilizadas para a criação de preditores de interações oncogênicas mostradas no Capítulo 4, exceto pela geração de alguns atributos de treinamento e seleção das instâncias positivas (genes drogáveis).

5.2.1 Geração dos atributos de treinamento

Foram utilizados como atributos de treinamento para a construção do preditor de genes drogáveis *(i)* parâmetros topológicos, *(ii)* localização subcelular e *(iii)* perfil de expressão tecidual dos genes presentes na *RIGH*.

5.2.1.1 Obtenção dos parâmetros topológicos da rede integrada de interações moleculares de genes humanos (*RIGH*)

Os 12 parâmetros topológicos utilizados como atributos de treinamento para a criação do preditor de genes drogáveis foram os mesmos parâmetros utilizados para a predição de interações oncogênicas (Capítulo 4).

5.2.1.2 Localização subcelular dos genes humanos

A localização subcelular das proteínas codificadas pelos genes na *RIGH* foi realizada através da utilização do *QuickGO*, ferramenta disponibilizada no banco de dados *InterPro* do *European Bioinformatics Institute* (BINNS et al., 2009) que busca termos do *Gene Ontology (GO)* associados aos genes. Os termos do *GO*, como visto anteriormente no Capítulo 2, estão relacionados hierarquicamente onde termos mais gerais estão no topo da hierarquia e termos mais específicos encontram-se na base da hierarquia. Para a determinação da localização subcelular das proteínas codificadas pelos genes da *RIGH*, foram selecionados termos do *GO* relacionados à localização subcelular dispostos em uma posição intermediária da hierarquia para evitar termos extremamente gerais ou específicos. Os termos selecionados foram: *cytoplasm* (citoplasma), *endoplasmic reticulum* (retículo endoplasmático), *mitochondrion* (mitocôndria), *nucleus* (núcleo), *extracellular space* (espaço extracelular), *Golgi apparatus* (complexo de Golgi), *plasma membrane* (membrana plasmática) e *cellular component* (qualquer local na célula; ainda sem localização específica). Os genes que foram anotados com outros termos na posição intermediária da hierarquia do *GO* foram reanotados com o termo *other localization* (outra localização) e genes cuja localização subcelular de suas proteínas ainda não foi determinada foram anotados como *unknown* (desconhecida).

5.2.1.3 Perfil de expressão tecidual dos genes humanos

Os perfis de expressão tecidual dos genes da *RIGH* foram obtidos a partir do estudo realizado por Reverter e colaboradores (REVERTER; INGHAM; DALRYMPLE, 2008). Nesse estudo, Reverter e colaboradores analisaram dados de expressão obtidos por sequenciamento paralelo de transcritos (mRNA) de 32 tecidos com o objetivo de identificar genes com expressão ubíqua ou tecido-específicos e, então, relacionar esses dados com interações e doenças. Foram considerados como atributos de treinamento relacionados ao perfil de expressão tecidual dos genes humanos (*i*) o número de tecidos nos quais o gene é expresso pelo menos em 5 transcritos por milhão (tpm) e *(ii)* a expressão média em tpm entre todos os tecidos nos quais o gene é expresso (REVERTER; INGHAM; DALRYMPLE, 2008).

5.2.2 Construção e avaliação dos preditores

5.2.2.1 Construção dos grupos de treinamento

O esquema de construção dos grupos de treinamento para a geração dos preditores de genes drogáveis foi o mesmo utilizado para a construção de preditores de interações oncogênicas

no *graph2sig*, isto é, foram construídos dois diferentes conjuntos de grupos balanceados de treinamento: (1) um conjunto de dez grupos de treinamento contendo as instâncias positivas e negativas (genes) corretamente associadas às suas classes (drogável e não-drogável), conjunto chamado de "treinamento normal" e (2) outro conjunto de dez grupos de treinamento nos quais as classes "drogável" e "não-drogável" foram aleatoriamente atribuídas aos genes, conjunto chamado de "treinamento permutado".

A primeira etapa de construção dos grupos de treinamento foi a compilação de uma lista de genes drogáveis. Esses genes foram extraídos da rede de interações entre drogas aprovadas pela FDA e seus alvos construída por Yildirim e colaboradores (YILDIRIM et al., 2007) e então mapeados na *RIGH*. A lista final de genes drogáveis presentes na *RIGH* e utilizados como instâncias positivas nos grupos de treinamento contém 257 genes drogáveis. A segunda etapa da construção dos grupos de treinamento foi a compilação de uma lista de genes não-drogáveis. Como não é possível, no presente momento, construir uma lista de genes que codificam proteínas que não sejam, inequivocamente, alvos de drogas, foram considerados genes não-drogáveis todos os 9.894 genes que sobraram após a remoção dos 257 genes drogáveis. A terceira etapa da construção dos grupos balanceados de treinamento foi a seleção aleatória de 10 diferentes grupos de 257 genes dentre os 9.894 genes não-drogáveis e a combinação de cada um desses grupos com os 257 genes drogáveis formando um conjunto com 10 diferentes grupos de treinamento contendo os 257 genes drogáveis e 257 genes não-drogáveis. O conjunto de 10 grupos de treinamento nos quais as classes "drogável" e "não-drogável" foram aleatoriamente atribuídas aos genes foi gerado a partir do conjunto acima.

5.2.2.2 Seleção dos algoritmos de aprendizagem

Foi utilizado a mesma combinação de sete algoritmos de aprendizagem (AA) utilizada para a construção de preditores de interações oncogênicas (Capítulo 4). As probabilidades estimadas pelos AAs foram combinadas pelo algoritmo *meta.Vote* (KITTLER et al., 1998) e, antes de utilizar o *meta.Vote*, foi aplicado a cada AA o *bootstrap aggregating* (*bagging*) (BREIMAN, 1996a).

5.2.2.3 Avaliação dos preditores

O desempenho dos preditores de genes drogáveis foi avaliado estimando-se a precisão e a sensibilidade desses preditores através da técnica de validação cruzada com $v = 10$ (ver Capítulo 2 para descrição detalhada da técnica de validação cruzada). As precisões e as sensibilidades estimadas pela validação cruzada foram expressas como medianas das 10 medidas de desem-

penho dos 10 preditores gerados pelo treinamento da combinação de AAs pelos 10 grupos de treinamento do conjunto normal e das 10 medidas de desempenho dos 10 preditores gerados pelo treinamento da combinação de AAs pelos 10 grupos de treinamento do conjunto permutado.

5.2.3 Predição de novos alvos de drogas

Como não é possível garantir que não há genes absolutamente drogáveis nem não-drogáveis, atribuiu-se a cada gene da *RIGH* um "grau de drogabilidade normal". Para verificar se o grau de drogabilidade normal atribuído ao gene não é resultado do aprendizado de características associadas a quaisquer subgrupos aleatórios de genes pelos AAs, foi gerado também um "grau de drogabilidade permutado" para cada gene. Para isso, os 10 preditores gerados pelo conjunto de treinamento normal e os 10 preditores gerados pelo conjunto de treinamento permutado foram aplicados a todos os genes da *RIGH*. Nesse processo, a classe de cada gene é omitida e os preditores atribuem para cada gene uma probabilidades estimada do gene ser drogável ($D(i)$). Para cada gene, o grau de drogabilidade normal final é a mediana das 10 $D(i)$s atribuídas pelos 10 preditores gerados pelo conjunto de treinamento normal e o grau de drogabilidade permutado final é a mediana das 10 $D(i)$s atribuídas pelos 10 preditores gerados pelo conjunto de treinamento permutado.

5.2.4 Comparações estatísticas

As comparações estatísticas entre *(i)* as medidas de desempenho estimadas dos preditores gerados pelo treinamento da combinação de AAs pelo conjunto de treinamento normal e as medidas de desempenho estimadas dos preditores gerados pelo treinamento da combinação de AAs pelo conjunto de treinamento permutado e *(ii)* os graus de drogabilidade normal e permutado para cada gene da *RIGH* foram realizadas pelo teste de Wilcoxon (WILCOXON, 1947), descrito detalhadamente no Apêndice A. Como, no nosso caso, $N < 15$, já que temos 10 grupos de treinamento para cada um dos conjuntos de treinamento (normal e permutado), foi feita a comparação entre os valores de W calculados com W_c na Tabela 6.1 no Apêndice A considerando $\alpha = 0,05$. Portanto, se $W \leq W_c$ para um dado N em $\alpha = 0,05$, as diferenças foram consideradas estatisticamente significativas.

5.3 Resultados e discussão

5.3.1 Avaliação do desempenho dos preditores

Antes de avaliar o desempenho dos preditores de genes drogáveis gerados a partir do conjunto de treinamento normal, comparamos estatisticamente as medianas dos valores de sensibilidade e precisão dos 10 preditores gerados a partir do conjunto de treinamento permutado com as medianas dos valores de sensibilidades e precisão dos 10 preditores gerados a partir do conjunto de treinamento normal. Como já mencionado na seção "Métodos", esse procedimento serve para verificar se os AAs treinados com o grupo de treinamento normal detectaram padrões nos atributos de treinamento realmente associados aos genes drogáveis em vez de padrões associados a quaisquer subgrupos aleatórios de genes. Como pode-se observar na Tabela 5.1, as medianas das medidas de desempenho dos 10 preditores gerados a partir do conjunto de treinamento permutado são significativamente menores do que as medidas de desempenho dos 10 preditores gerados a partir do conjunto de treinamento normal ($W \leq W_c$ para N = 10 em $\alpha = 0,05$; ver Tabela 6.1 no Apêndice A). Esse resultado indica, portanto, que padrões extraídos dos atributos de treinamento realmente associados aos genes drogáveis foram detectados pelos preditores construídos a partir do conjunto de treinamento normal.

Tabela 5.1: Medidas de desempenho dos preditores de genes drogáveis

Medida de desempenho	Mediana [min,max] [1]	Mediana [min,max] [1]	N	W	W_c ($\alpha = 0,05$)[2]
	Normal	Permutado			
Precisão	0,748 [0,72,0,763]	0,5 [0,451,0,556]	10	0	8 *
Sensibilidade	0,782 [0,732,0,809]	0,492 [0,447,0,564]	10	0	8 *

[1] Conjunto de 10 preditores

[2] De acordo com a Tabela 6.1 do Apêndice A

* Diferença estatisticamente significativa

Baseado no resultado acima, podemos concluir, portanto, que os valores de sensibilidade e precisão do conjunto de preditores gerados a partir do conjunto de treinamento normal observados na Tabela 5.1 refletem, de fato, o desempenho desse conjunto de preditores em discernir genes drogáveis de genes não-drogáveis a partir de padrões extraídos da integração dos dados sobre parâmetros topológicos, perfis de expressão tecidual e localização subcelular dos genes. O desempenho desse conjunto de preditores, aliás, foi satisfatório: ele foi capaz de recuperar 78,2% dos genes drogáveis (sensibilidade) com uma precisão de 74,8%.

A observação de que nosso conjunto de preditores não recuperou cerca de 22% dos genes drogáveis e classificou cerca de 25% dos genes não-drogáveis como drogáveis indica que o nível

de ruído nos dados de treinamento é grande e está provavelmente associado com a existência de características comuns compartilhadas entre os gene drogáveis e não-drogáveis. Provavelmente, essas características comuns devem-se parcialmente à abordagem que adotamos para a seleção dos genes não-drogáveis: como, atualmente, não é possível construir uma lista com genes que codifiquem proteínas que não sejam inequivocamente alvos de drogas, foram considerados genes não-drogáveis todos os genes da *RIGH*, exceto os 257 genes conhecidamente drogáveis. Portanto, alguns desses genes não-drogáveis podem ser genes drogáveis ainda desconhecidos que compartilham características em comum com os genes drogáveis conhecidos. Outro fator que pode contribuir com a existência dessas características comuns compartilhadas entre genes drogáveis e não-drogáveis é o fato da *RIGH* ainda ser incompleta: Stumpf e colegas (STUMPF et al., 2008) estimaram, por exemplo, que a rede de interações físicas entre proteínas humanas têm cerca de 650.000 interações. A *RIGH* contém cerca de 43.000 interações físicas entre proteínas e, portanto, podemos prever que os valores de todos os parâmetros topológicos provavelmente mudarão com o aumento do tamanho da rede e isso poderá fazer com que características comuns compartilhadas entre genes drogáveis e não-drogável, pelo menos aquelas relacionadas com os parâmetros topológicos, desapareçam.

Apesar das limitações discutidas acima, nosso conjunto de preditores de genes drogáveis parece, de fato, ser confiável, como mostrado a seguir. Atribuímos a cada um dos genes da *RIGH* um grau de drogabilidade normal e um grau de drogabilidade permutado, como descrito em “Métodos”, e analisamos se os genes drogáveis conhecidos seriam mais frequentes em intervalos de valores normais de grau de drogabilidade mais elevados. Para isso, determinamos as distribuições de frequências dos genes drogáveis conhecidos com graus de drogabilidade normal e permutado em 20 intervalos de valores de grau de drogabilidade (Figura 5.1). Essas frequências representam as razões entre a quantidade de genes drogáveis conhecidos presentes nos intervalos de valores de grau de drogabilidade e a quantidade total desses genes presentes na *RIGH*. Podemos observar na Figura 5.1 que os genes drogáveis conhecidos com grau de drogabilidade normal tendem a ser mais frequentes em torno de um grau de drogabilidade de aproximadamente 0,82 (Figura 5.1), o que indica que o nosso conjunto de preditores conseguiu atribuir aos genes drogáveis conhecidos valores elevados de grau de drogabilidade. Esse resultado reforça, juntamente com as medidas de desempenho satisfatórias, que nosso conjunto de preditores tem potencial para ser utilizado para prever potenciais genes drogáveis, como mostrado na seção seguinte.

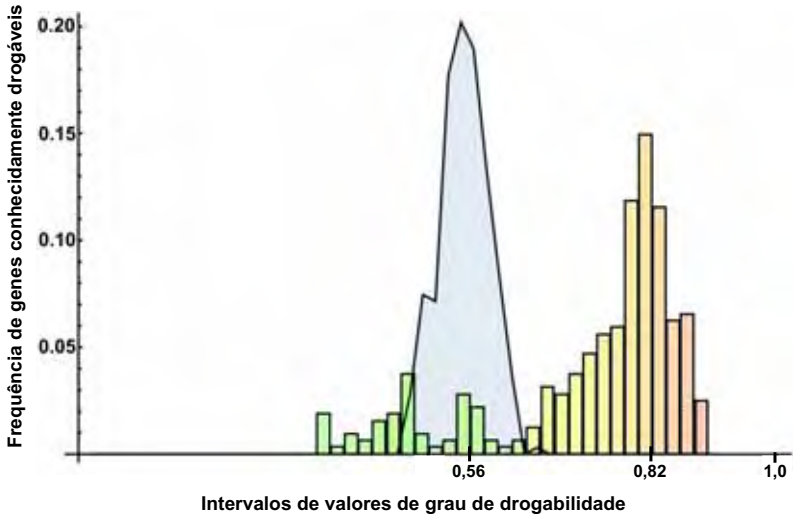

Figura 5.1: Distribuição de frequências de genes conhecidamente drogáveis em intervalos de valores de grau de drogabilidade. As barras coloridas representam a distribuição de frequências de genes conhecidamente drogáveis com valores normais de grau de drogabilidade e a área em azul-claro semitransparente representa a distribuição de frequências de genes conhecidamente drogáveis com valores permutados de grau de drogabilidade. Os intervalos são de 0,02.

5.3.2 Predição de potenciais alvos de drogas na *RIGH*

O poder preditivo que o nosso conjunto de preditores exibiu na seção anterior (Tabela 5.1 e Figura 5.1) nos impulsionou a aplicá-lo na *RIGH* para a predição de genes potencialmente drogáveis. Atribuímos a cada um dos genes da *RIGH* um grau de drogabilidade normal e um grau de drogabilidade permutado e comparamos estatisticamente esses dois valores para cada gene da rede. Somente os genes que apresentaram um grau de drogabilidade normal significativamente diferente ($W \leq W_c$ para $N = 10$ em $\alpha = 0,05$; ver Tabela 6.1 no Apêndice A) do grau de drogabilidade permutado foram considerados. Dos 10.151 genes da *RIGH*, 8.967 ($\approx 88\%$) apresentaram graus de drogabilidade normais significativamente diferentes dos graus de drogabilidade permutados. Desses 8.967 genes, selecionamos os genes com os 10 maiores graus de drogabilidade, excluindo os drogáveis conhecidos, para uma avaliação mais detalhada (Tabelas 5.2 e 5.3).

Todos os 11 genes com os 10 maiores graus de drogabilidade na *RIGH* codificam proteínas extracelulares ou associadas à membrana plasmática (Tabela 5.3). Esse resultado reflete a natureza da lista de genes drogáveis utilizadas para o treinamento dos algoritmos de aprendizagem: cerca de 60% das proteínas codificadas por esses genes estão presentes na matriz extracelular ou associadas à membrana plasmática. Esse enriquecimento em proteínas extracelulares ou as-

sociadas à matriz extracelular entre proteínas drogáveis, por sua vez, reflete a tendência em se desenvolver drogas que atuem em alvos mais acessíveis (YILDIRIM et al., 2007).

Buscamos na literatura biomédica artigos que mostrassem explicitamente que as proteínas codificadas pelos genes com os 10 maiores graus de drogabilidade na *RIGH* (Tabelas 5.2 e 5.3), excluindo os drogáveis conhecidos, fossem consideradas potenciais alvos de drogas, o que chamamos de "evidência de drogabilidade". Como se pode observar na Tabela 5.2, oito ($\approx 73\%$) dos 11 genes com os 10 maiores graus de drogabilidade têm evidência de drogabilidade. Dentre os três genes cujas proteínas ainda não foram identificadas explicitamente como possíveis alvos de drogas (*HLA-F*, *CD8A* e *ITGAX*; Tabela 5.2), somente o *ITGAX* apresenta evidência indireta de drogabilidade. Foi demonstrado que esse gene, cuja expressão se restringe às células hematopoieticas da linhagem mieloide, também é expresso em linfócitos neoplásicos na tricoleucemia, um tipo raro de leucemia de linfócitos B (NICOLAOU et al., 2003). Portanto, essa especificidade de expressão torna a proteína ITGAX, por exemplo, candidata a alvo terapêutico contra esse tipo de leucemia.

Tabela 5.2: Genes da *RIGH* com os 10 maiores graus de drogabilidade

Gene[1]	Grau de drogabilidade (Mediana [min,max])[2]		N	W	W_c[3] ($\alpha = 0,05$)	Evidência drogabilidade[4]
	Normal	Permutado				
HLA-F	0,887[0,803,0,915]	0,530[0,427,0,584]	10	0	8*	Sem evidência
PLAU	0,886[0,808,0,907]	0,561[0,387,0,675]	10	0	8*	19301652
CD8A	0,885[0,871,0,902]	0,56[0,37,0,664]	10	0	8*	Sem evidência
CD19	0,880[0,751,0,907]	0,562[0,38,0,628]	10	0	8*	19509168
ITGAM	0,878[0,614,0,887]	0,534[0,36,0,656]	10	1	8*	11931348
THBS1	0,875[0,53,0,9]	0,532[0,293,0,592]	10	0	8*	17878288
ITGAX	0,873[0,784,0,897]	0,539[0,422,0,691]	10	0	8*	Sem evidência
CXCR5	0,871[0,755,0,895]	0,537[0,49,0,59]	10	0	8*	17652619
EBI3	0,871[0,801,0,888]	0,529[0,391,0,626]	10	0	8*	19556516
IL6	0,87[0,766,0,893]	0,591[0,361,0,643]	10	0	8*	17465721
TIMP2	0,869[0,645,0,916]	0,584[0,34,0,701]	10	0	8*	10985804

[1] Símbolos oficiais de acordo com o *Human Gene Nomenclature Committee* (*HGCN*) (BRUFORD et al., 2008)

[2] Conjunto de 10 preditores

[3] De acordo com a Tabela 6.1 do Apêndice A

[4] *Pudmed IDs* dos artigos mais recentes que mostram claramente que as proteínas codificadas pelos referidos genes podem ser potenciais alvos de drogas.

[*] Diferença estatisticamente significativa

Tabela 5.3: Descrição e função dos genes da *RIGH* com os 10 maiores graus de drogabilidade

Símbolo oficial[1]	Descrição	Processos biológicos [2]
HLA-F	Antígeno leucocitário humano F	Processamento e apresentação de antígeno.
PLAU	Uroquinase	Quimiotaxia; transdução de sinal; regulação de adesão celular mediada por integrinas; regulação de migração de células musculares lisas; cicatrização.
CD8A	Cadeia α do antígeno CD8	Ativação de linfócitos T; processamento e apresentação de antígeno; transdução de sinal
CD19	Antígeno CD19	Transdução de sinal; defesa celular
ITGAM	Cadeia α M da integrina	Adesão celular; transdução de sinal; proliferação de células T ativadas.
THBS1	Trombospondina 1	Regulação positiva de migração celular, angiogênese e coagulação; regulação negativa de proliferação celular e adesão à matriz extracelular.
ITGAX	Cadeia α X da integrina	Adesão celular; transdução de sinal; morfogênese.
CXCR5	Receptor de quimiocina 5	Transdução de sinal acoplada à proteína G; ativação e migração de linfócitos B.
EBI3	Induzido pelo vírus Epstein-Barr	Regulação da proliferação de linfócitos T; regulação da biossíntese de interferon gama.
IL6	Interleucina 6	Regulação de cascatas de sinalização envolvidas com repostas imunológicas e inflamatórias;
TIMP2	Inibidor de metaloproteinase 2	Regulação do metabolismo de cAMP; regulação de proliferação celular; regulação da diferenciação neuronal.

[1] Símbolos oficiais de acordo com o *Human Gene Nomenclature Committee* (*HGCN*) (BRUFORD et al., 2008)

[2] De acordo com os termos do *Gene Ontology* (BERARDINI et al., 2010) relacionados aos processos biológicos

5.3.3 Predição de potenciais alvos de drogas na G_{ccam}

Dos 2.212 genes da G_{ccam}, 1.783 ($\approx$ 81%) apresentaram graus de drogabilidade normais significativamente diferentes dos graus de drogabilidade permutados. Desses 1.783 genes, selecionamos aqueles com os 10 maiores graus de drogabilidade, excluindo os drogáveis conhecidos, para uma avaliação mais detalhada (Tabela 5.4).

Dentre os 11 genes com os 10 maiores graus de drogabilidade na G_{ccam}, oito (*PLAU*, *CD8A*, *ITGAM*, *THBS1*, *ITGAX*, *EBI3*, *IL6* e *TIMP2*) estão entre os genes com os 10 maiores graus de drogabilidade na *RIGH* (Tabela 5.4). Desses 11 genes com os maiores graus de drogabilidade

na G_{ccam}, nove (81%) apresentaram evidência de drogabilidade (Tabela 5.4). É interessante observar que, dentre esses noves genes, o *FLT1* é alvo da Ranibizumab, droga aprovada pela FDA em 2009 (ROUVAS et al., 2009). Por que o *FLT1*, então, não entrou na nossa lista de genes drogáveis conhecidos? A lista de genes drogáveis construída para treinar os algoritmos de aprendizagem contém genes drogáveis conhecidos até 2007 e, naquele momento, o *FLT1* ainda era um alvo potencial. Essa fato, em conjunto com a descoberta de que grande parte dos genes com os maiores graus de drogabilidade têm evidência de drogabilidade, reforça o poder de predição do nosso conjunto de preditores.

Tabela 5.4: Genes da G_{ccam} com os 10 maiores graus de drogabilidade

Gene[1]	Grau de drogabilidade (Mediana [min,max])[2]		N	W	W_c[3] ($\alpha = 0,05$)	Evidência drogabilidade[4]
PLAU	0,886[0,808,0,907]	0,561[0,387,0,675]	10	0	8*	19301652
CD8A	0,885[0,871,0,902]	0,56[0,37,0,664]	10	0	8*	Sem evidência
ITGAM	0,878[0,614,0,887]	0,534[0,36,0,656]	10	1	8*	11931348
THBS1	0,875[0,53,0,9]	0,532[0,293,0,592]	10	0	8*	17878288
ITGAX	0,873[0,784,0,897]	0,539[0,422,0,691]	10	0	8*	Sem evidência
EBI3	0,871[0,801,0,888]	0,529[0,391,0,626]	10	0	8*	19556516
IL6	0,87[0,766,0,893]	0,591[0,361,0,643]	10	0	8*	17465721
TIMP2	0,869[0,645,0,916]	0,584[0,34,0,701]	10	0	8*	10985804
FLT1	0.863[0.748,0.903]	0.563[0.365,0.691]	10	0	8*	19602910[5]
ITGAV	0.863[0.819,0.9]	0.556[0.401,0.63]	10	0	8*	11911248
AREG	0.861[0.797,0.888]	0.517[0.317,0.669]	10	0	8*	20383197

[1] Símbolos oficiais de acordo com o *Human Gene Nomenclature Committee* (*HGCN*) (BRUFORD et al., 2008)

[2] Conjunto de 10 preditores

[3] De acordo com a Tabela 6.1 do Apêndice A

[4] *Pudmed IDs* dos artigos mais recentes que mostram claramente que as proteínas codificadas pelos referidos genes podem ser potenciais alvos de drogas.

[5] Drogas aprovadas pela FDA após a obtenção da lista de genes drogáveis.

* Diferença estatisticamente significativa

O gene *FLT1*, além de pertencer ao grupo de genes com evidência de drogabilidade dentre os 11 genes com os 10 maiores graus de drogabilidade na G_{ccam}, também se encontra no grupo de genes participantes de interações na G_{ccam} com alto potencial de oncogenicidade ($p_{carc} \geq 0,7$) determinados pelo *graph2sig* (ver Capítulo 4). Os outros genes presentes na G_{ccam} que participam desse grupo são o *AREG*, o *IL6* e o *THBS1*. A participação simultânea desses genes no grupo dos genes com os 10 maiores graus de drogabilidade e no grupo de genes com interações com alto potencial de oncogenicidade os tornam potenciais alvos de drogas no controle da transição G1/S pela adesão à matriz extracelular em células cancerosas.

Nenhum desses genes, porém, está presente na sub-rede $EGFR - CDC6$ extraída pelo

graph2sig a partir da G_{ccam}. Como mencionado no Capítulo 4, essa sub-rede parece estar potencialmente envolvida no controle da expressão da CDC6 pelos sinais oncogênicos deflagrados pela EGFR em células cancerosas e, por esse motivo e de acordo com estudos anteriores (JINNO et al., 2002), a sub-rede $EGFR-CDC6$ também pode estar envolvida na transição G1/S do ciclo celular na ausência de adesão à matriz extracelular em células cancerosas com atividade contínua da EGFR. Portanto, esses genes até podem regular o controle da transição G1/S pela adesão à matriz extracelular em células cancerosas, mas parece que eles não tem potencial para regular a transição G1/S do ciclo celular na ausência de adesão à matriz extracelular particularmente quando a EGFR está constitutivamente ativada.

Como um dos objetivos dessa tese é encontrar potenciais novos alvos de drogas que possam inibir a capacidade de células cancerosas em progredir para a fase S sem adesão à matriz extracelular, fenótipo aparentemente necessário para a ocorrência de metástases (CIFONE, 1982; FREEDMAN; SHIN, 1974; STEIN, 1979; MORI et al., 2009), investigamos os graus de drogabilidade dos genes da sub-rede $EGFR-CDC6$. Foram considerados genes potencialmente alvos de drogas que possam inibir a capacidade de células cancerosas em progredir para a fase S sem adesão à matriz extracelular, chamados a partir daqui de genes $alvo\_pot$, todos os genes, exceto os conhecidamente drogáveis, com valores normais de grau de drogabilidade significativamente diferentes dos valores permutados ($W \leq W_c$ para $N = 10$ em $\alpha = 0,05$) e maiores do que 0,5.

Dentre os 21 genes que formam a sub-rede $EGFR-CDC6$ (Figura 4.6), 17 ($\approx 81\%$) têm valores normais de drogabilidade significativamente diferentes dos valores permutados. Desses 17 genes, 12 (*AR*, *CAV1*, *CCND1*, *CDKN1A CTNNB1*, *EGFR*, *ESR1*, *JUN*, *SMAD3*, *SMAD4*, *SRC* e *STAT3*) têm grau de drogabilidade maior do que 0,5. Como os genes *AR*, *ESR1*, *EGFR*, *SRC* e *STAT3* são genes conhecidamente drogáveis, foram considerados como $alvo\_pot$ os sete genes restantes (*CAV1*, *CCND1*, *CDKN1A*, *CTNNB1*, *JUN*, *SMAD3* e *SMAD4*). A lista final de genes $alvo\_pot$ é composta, portanto, por esses sete genes. Note que, embora o gene *SRC* não esteja presente no grupo de treinamento original, ele foi considerado como conhecidamente drogável por que a droga anti-câncer (Dasatinib) que atua na proteína SRC foi considerada oficial pela FDA somente após a data da coleta dos dados (KAMATH et al., 2008).

Nós podemos levantar a hipótese, portanto, que a supressão total ou parcial da capacidade de proliferação na ausência de adesão à matriz extracelular das células cancerosas que carregam a proteína EGFR continuamente ativada poderia ser feita com a utilização de drogas que atuem isoladamente ou conjuntamente sobre os genes *CDKN1A*, *JUN*, *SMAD3*, *SMAD4*, *CAV1*, *CCND1* e *CTNNB1* (Figura 5.2).

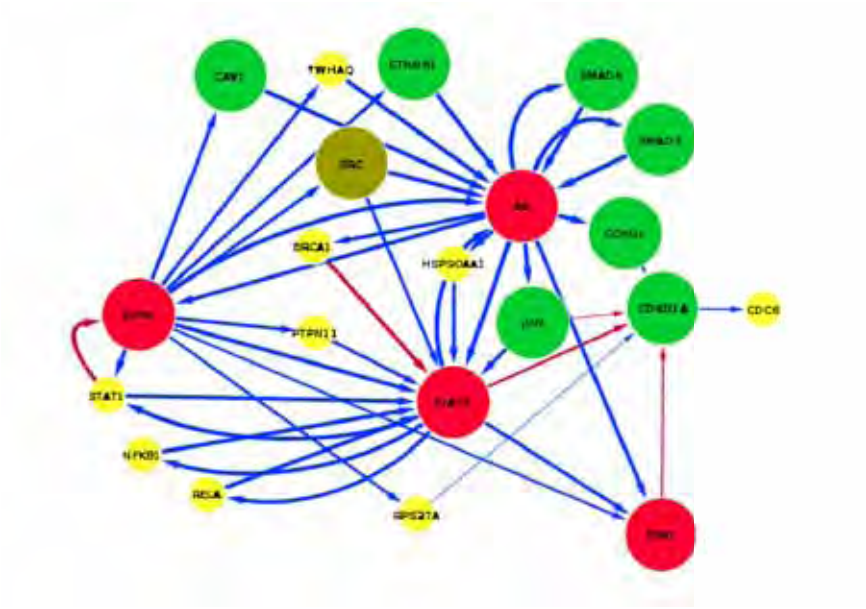

Figura 5.2: Genes conhecidamente e potencialmente drogáveis na sub-rede $EGFR-CDC6$. Os genes coloridos em vermelho, marrom e verde são, respectivamente, genes conhecidamente drogáveis, genes reconhecidos como oficialmente drogáveis somente após a coleta dos dados para construção dos grupos de treinamento e genes potencialmente drogáveis.

5.4 Conclusões

Neste Capítulo, descrevemos o desenvolvimento de um método computacional baseado em aprendizado de máquina e dados de medidas de centralidade, perfil de expressão tecidual e localização subcelular dos genes para a predição de genes drogáveis na *RIGH* e na G_{ccam}. Levando em consideração os limites impostos pela incompletude da *RIGH*, como discutido nos Capítulos 3 e 4 e na seção "Análise do desempenho dos preditores" deste Capítulo, os resultados obtidos nos conduzem às seguintes conclusões:

- Conjuntamente, os dados de medidas de centralidade, perfil de expressão tecidual e localização subcelular dos genes são capazes de gerar preditores de gene drogáveis já que *(i)* a sensibilidade mediana desses preditores foi de aproximadamente 78%, isto é, esses preditores conseguiram recuperar 78% de todos os genes conhecidamente drogáveis, e *(ii)* os valores normais de grau de drogabilidade atribuídos aos genes conhecidamente drogáveis se distribuem com maior frequência em torno de um grau de drogabilidade aproximadamente igual a 0,82;

- Esse método é capaz de prever genes potencialmente drogáveis: tanto na *RIGH* quanto na G_{ccam}, cerca de 2/3 dos genes com os 10 maiores graus de drogabilidade têm evidência de drogabilidade;

- Os genes *AREG*, *IL6* e *THBS1* são potenciais alvos de drogas que podem atuar na modulação do controle da transição G1/S do ciclo celular pela adesão à matriz extracelular em células cancerosas;
- Os genes *CDKN1A*, *JUN*, *SMAD3*, *SMAD4*, *CAV1*, *CCND1* e *CTNNB1* são potenciais alvos de drogas para inibir a capacidade proliferativa na ausência de adesão à matriz extracelular de células cancerosas que possuem EGFR constitutivamente ativada.

6 *Considerações finais*

Os resultados mostrados nesta tese indicam que a combinação entre *(i)* modelagem de processos biológicos em redes e *(ii)* a utilização de aprendizado de máquina para predizer ou descrever como propriedades emergentes podem surgir a partir das interações entre os componentes dessas redes é promissora e passível de ser utilizada para a geração de hipóteses biologicamente plausíveis sobre processos biológicos de interesse. Essa combinação é promissora por que conseguimos prever com sucesso tanto genes conhecidamente drogáveis quanto interações conhecidamente oncogênicas na *RIGH* e na G_{ccam}. E é passível de ser utilizada para a geração de hipóteses biologicamente plausíveis por que as hipóteses que levantamos sobre o funcionamento de parte do mecanismo molecular subjacente ao controle da transição G1/S do ciclo celular pela adesão à matriz extracelular em células cancerosas e sobre os potenciais alvos de drogas para inibir a capacidade da proliferação das células cancerosas na ausência de matriz extracelular respaldam-se amplamente em dados da literatura biomédica.

Além do que foi exposto acima, outras contribuições desta tese foram:

- Construção pioneira de uma rede de interações entre genes humanos contendo simultaneamente interações físicas entre proteínas, metabólicas e de regulação transcricional;
- Criação de uma combinação de algoritmos de aprendizado de máquina que aumenta o desempenho de predição em relação aos algoritmos individuais;
- Desenvolvimento de um método, o *graph2sig*, para extrair vias de sinalização potencialmente envolvidas em um processo biológico entre dois genes:
  - Virtualmente, o *graph2sig* pode ser utilizado com qualquer processo biológico de interesse.
- Desenvolvimento de um método para prever genes potencialmente drogáveis;
- Criação de uma plataforma “geradora de hipóteses” relacionadas ao câncer:

- O cruzamento das informações sobre as interações potencialmente oncogênicas com as informações sobre os genes potencialmente drogáveis pode revelar novos alvos de drogas anti-câncer;

Em relação à regulação da transição G1/S do ciclo celular pela adesão à matriz extracelular, as hipóteses geradas nesta tese foram as seguintes:

1. Parte da capacidade das células cancerosas que carregam a proteína EGFR continuamente ativada de transitarem da fase G1 para a fase S do ciclo celular sem adesão à matriz extracelular deve-se à estabilização da proteína CDC6 pela CDKN1A. Essa estabilização é comum a todas as células cancerosas, mas a regulação da expressão da CDKN1A pela EGFR depende do tipo de célula cancerosa ou da situação a qual a célula cancerosa está submetida;

2. A supressão total ou parcial da capacidade de proliferação na ausência de adesão à matriz extracelular das células cancerosas que carregam a proteína EGFR continuamente ativada poderia ser feita com a utilização de drogas que possam atuar, isoladamente ou conjuntamente, sobre os genes *CDKN1A*, *JUN*, *SMAD3*, *SMAD4*, *CAV1*, *CCND1* e *CTNNB1*.

Referências Bibliográficas

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.

ALBERTS, B. et al. *Molecular Biology of the Cell*. [S.l.]: Garland Science, 2002.

ANTHONISSE, J. *The rush in a directed graph*. [S.l.], 1971. Technical Report BN 9/71.

ASSOIAN, R. K. Control of the G1 phase cyclin-dependent kinases by mitogenic growth factors and the extracellular matrix. *Cytokine Growth Factor Rev*, v. 8, n. 3, p. 165–170, 1997.

BACKES, C. et al. GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res*, v. 35, n. Web Server issue, p. W186–92, 2007.

BALDIN, V. et al. Cyclin D1 is a nuclear protein required for cell cycle progression in G1. *Genes Dev*, v. 7, n. 5, p. 812–821, 1993.

BAO, W. et al. Cell attachment to the extracellular matrix induces proteasomal degradation of p21(CIP1) via Cdc42/Rac1 signaling. *Mol Cell Biol*, v. 22, n. 13, p. 4587–4597, 2002.

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.

BARBERIS, M. et al. Cell size at S phase initiation: an emergent property of the G1/S network. *PLoS Comput Biol*, v. 3, n. 4, p. e64, 2007.

BARBIERI, I. et al. Stat3 is required for anchorage-independent growth and metastasis but not for mammary tumor development downstream of the ErbB-2 oncogene. *Mol Carcinog*, v. 49, n. 2, p. 114–120, 2010.

BERARDINI, T. Z. et al. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*, v. 38, n. Database issue, p. D331–5, 2010.

BERTALANFFY, L. V. *General System theory: Foundations, Development, Applications*. [S.l.: s.n.], 1968.

BESSON, A.; DOWDY, S. F.; ROBERTS, J. M. CDK inhibitors: cell cycle regulators and beyond. *Dev Cell*, v. 14, n. 2, p. 159–169, 2008.

BHALLA, U. S.; IYENGAR, R. Emergent properties of networks of biological signaling pathways. *Science*, v. 283, n. 5400, p. 381–387, 1999.

BIANKIN, A. V. et al. Overexpression of p21(waf1/cip1) is an early event in the development of pancreatic intraepithelial neoplasia. *Cancer Res*, v. 61, n. 24, p. 8830–8837, 2001.

BINNS, D. et al. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, v. 25, n. 22, p. 3045–6, 2009.

BOTTAZZI, M. E. et al. Regulation of p21(cip1) expression by growth factors and the extracellular matrix reveals a role for transient erk activity in g1 phase. *J Cell Biol*, v. 146, n. 6, p. 1255–1264, 1999.

BREIMAN, L. Bagging predictors. *Mach Learn*, v. 24, n. 2, p. 123, 1996.

BREIMAN, L. Heuristics of instability and stabilization in model selection. *Ann. Stat.*, v. 24, n. 6, p. 2350–2383, 1996.

BREIMAN, L. *Some infinity theory for predictor ensembles*. [S.l.], 2000.

BREIMAN, L. Random forests. *Mach Learn*, v. 45, n. 1, p. 5–32, 2001.

BRUFORD, E. A. et al. The HGNC database in 2008: a resource for the human genome. *Nucleic Acids Res*, v. 36, n. Database issue, p. D445–8, 2008.

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

CIFONE, M. A. In vitro growth characteristics associated with benign and metastatic variants of tumor cells. *Cancer Metastasis Rev*, v. 1, n. 4, p. 335–47, 1982.

CLUASET, A.; SHALIZI, C. R.; NEWMAN, M. E. J. Power-law distribution in empirical data. *SIAM Rev*, v. 51, p. 661–703, 2009.

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.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, v. 7, p. 1–30, 2006.

DEPAMPHILIS, M. L. et al. Regulating the licensing of dna replication origins in metazoa. *Curr Opin Cell Biol*, v. 18, n. 3, p. 231–239, 2006.

DOTTO, G. P. p21(waf1/cip1): more than a break to the cell cycle? *Biochim Biophys Acta*, v. 1471, n. 1, p. M43–M56, 2000.

DUARTE, N. C. et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *PNAS*, v. 104, p. 1777–1782, 2007.

FREEDMAN, V. H.; SHIN, S. I. Cellular tumorigenicity in nude mice: correlation with cell growth in semi-solid medium. *Cell*, v. 3, n. 4, p. 355–9, 1974.

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

FREUND, Y.; MASON, L. The alternating decision tree learning algorithm. In: *Proceedings of the Sixteenth International Conference on Machine Learningg*. San Francisco: Morgan Kaufmann, 1999. p. 124–133.

GAD, A. et al. Retinoblastoma susceptibility gene product (pRb) and p107 functionally separate the requirements for serum and anchorage in the cell cycle G1-phase. *J Biol Chem*, v. 279, n. 14, p. 13640–13644, 2004.

GRASSIAN, A. R.; SCHAFER, Z. T.; BRUGGE, J. S. ErbB2 stabilizes epidermal growth factor receptor (EGFR) expression via Erk and Sprouty2 in extracellular matrix-detached cells. *J Biol Chem*, v. 286, n. 1, p. 79–90, 2011.

GUO, X. et al. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, v. 22, n. 8, p. 967–973, 2006.

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.

HARBOUR, J. W.; DEAN, D. C. The Rb/E2F pathway: expanding roles and emerging paradigms. *Genes Dev*, v. 14, n. 19, p. 2393–2409, 2000.

HENGSTSCHLAGER, M. et al. Cyclin-dependent kinases at the G1-S transition of the mammalian cell cycle. *Mutat Res*, v. 436, n. 1, p. 1–9, 1999.

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

HUSS, M.; HOLME, P. Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. *IET Syst Biol*, v. 1, n. 5, p. 280–285, 2007.

HWANG, S. et al. A protein interaction network associated with asthma. *J Theor Biol*, v. 252, n. 4, p. 722–731, 2008.

JEONG, H. et al. Lethality and centrality in protein networks. *Nature*, v. 411, n. 6833, p. 41–2, 2001.

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

JIMENEZ, V. M.; MARZAL, A. Computing the k shortest paths: a new algorithm and an experimental comparison. *Lect Notes Comput Sc*, v. 1668, p. 15–29, 1999.

JINNO, S. et al. Oncogenic stimulation recruits cyclin-dependent kinase in the cell cycle start in rat fibroblast. *Proc Natl Acad Sci U S A*, v. 96, n. 23, p. 13197–13202, 1999.

JINNO, S. et al. Cdc6 requires anchorage for its expression. *Oncogene*, v. 21, n. 11, p. 1777–1784, 2002.

KAMATH, A. V. et al. Preclinical pharmacokinetics and in vitro metabolism of dasatinib (BMS-354825): a potent oral multi-targeted kinase inhibitor against SRC and BCR-ABL. *Cancer Chemother Pharmacol*, v. 61, n. 3, p. 365–376, Mar 2008.

KANDASAMY, K. et al. Netpath: a public resource of curated signal transduction pathways. *Genome Biol*, v. 11, n. 1, p. R3, 2010.

KANEHISA, M. et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, v. 36, n. Database issue, p. D480–4, 2008.

KAUFFMAN, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*, v. 22, n. 3, p. 437–467, 1969.

KAWADA, M. et al. Induction of p27Kip1 degradation and anchorage independence by Ras through the MAP kinase signaling pathway. *Oncogene*, v. 15, n. 6, p. 629–637, 1997.

KESHAVA PRASAD, T. S. et al. Human Protein Reference Database–2009 update. *Nucleic Acids Res*, v. 37, n. Database issue, p. D767–72, 2009.

KITTLER, J. et al. On combining classifiers. *IEEE T Pattern Anal*, 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.

KRULL, M. et al. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res*, v. 34, n. Database issue, p. D546–D551, 2006.

LANDWEHR, N.; HALL, M.; FRANK, E. Logistic model trees. *Mach Learn*, v. 95, n. 1-2, p. 161–205, 2005.

LEBLANC, M.; TIBSHIRANI, R. Combining estimates in regression and classification. *J Am Stat Assoc*, v. 91, n. 436, p. 1641–1650, 1996.

LINDSAY, M. A. Target discovery. *Nat Rev Drug Discov*, v. 2, n. 10, p. 831–8, 2003.

MAGLOTT, D. et al. Entrez Gene: gene-centered information at ncbi. *Nucleic Acids Res*, v. 35, p. D26–D31, 2007.

MARMOR, M. D.; SKARIA, K. B.; YARDEN, Y. Signal transduction and oncogenesis by ErbB/HER receptors. *Int J Radiat Oncol Biol Phys*, v. 58, n. 3, p. 903–913, 2004.

MESSINA, D. N. et al. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res*, v. 14, n. 10B, p. 2041–2047, 2004.

MORI, S. et al. Anchorage-independent cell growth signature identifies tumors with metastatic potential. *Oncogene*, v. 28, n. 31, p. 2796–805, 2009.

NICOLAOU, F. et al. CD11c gene expression in hairy cell leukemia is dependent upon activation of the proto-oncogenes ras and jund. *Blood*, v. 101, n. 10, p. 4033–41, 2003.

OBAYA, A. J.; SEDIVY, J. M. Regulation of cyclin-Cdk activity in mammalian cells. *Cell Mol Life Sci*, v. 59, n. 1, p. 126–142, 2002.

OPITZ, D.; MACLIN, R. Popular ensemble methods: An empirical study. *J Artif Intell Res*, v. 11, p. 169–198, 1999.

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

POLIKAR, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, v. 6, n. 3, p. 21–45, 2006.

QUINLAN, J. R. *C4.5: programs for machine learning*. San Francisco: Morgan Kaufmann, 1993.

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.

REN, X. et al. An information-flow-based model with dissipation, saturation and direction for active pathway inference. *BMC Syst Biol*, v. 4, p. 72, 2010.

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

RONINSON, I. B. Oncogenic functions of tumour suppressor p21(Waf1/Cip1/Sdi1): association with cell senescence and tumour-promoting activities of stromal fibroblasts. *Cancer Lett*, v. 179, n. 1, p. 1–14, 2002.

ROUVAS, A. et al. The effect of intravitreal ranibizumab on the fellow untreated eye with subfoveal scarring due to exudative age-related macular degeneration. *Ophthalmologica*, v. 223, n. 6, p. 383–389, 2009.

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

SALWINSKI, L. et al. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, v. 32, n. Database issue, p. D449–51, 2004.

SCHAFER, Z. T. et al. Antioxidant and oncogene rescue of metabolic defects caused by loss of matrix attachment. *Nature*, v. 461, n. 7260, p. 109–113, 2009.

SCHELLENBERGER, J. et al. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, v. 11, p. 213, 2010.

SCOTT, J. et al. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol*, v. 13, n. 2, p. 133–144, 2006.

SHARAN, R.; ULITSKY, I.; SHAMIR, R. Network-based prediction of protein function. *Mol Syst Biol*, v. 3, p. 88, 2007.

SHI, H. Best-first decision tree learning. *Master Thesis*, 2007, The University of Waikato.

SHIGEMURA, K. et al. Soluble factors derived from stroma activated androgen receptor phosphorylation in human prostate LNCaP cells: roles of ERK/MAP kinase. *Prostate*, v. 69, n. 9, p. 949–955, 2009.

STARK, C. et al. The BioGRID interaction database: 2011 update. *Nucleic Acids Res*, 2010.

STEFFEN, M. et al. Automated modelling of signal transduction networks. *BMC Bioinformatics*, v. 3, p. 34, 2002.

STEIN, G. H. T98G: an anchorage-independent human tumor cell line that exhibits stationary phase G1 arrest in vitro. *J Cell Physiol*, v. 99, n. 1, p. 43–54, 1979.

STUMPF, M. P. H. et al. Estimating the size of the human interactome. *Proc Natl Acad Sci U S A*, v. 105, n. 19, p. 6959–64, 2008.

SUPPER, J. et al. BowTieBuilder: modeling signal transduction pathways. *BMC Syst Biol*, v. 3, p. 67, 2009.

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.

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

WILCOXON, F. Probability tables for individual comparisons by ranking methods. *Biometrics*, v. 3, n. 3, p. 119–22, 1947.

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

YILDIRIM, M. A. et al. Drug-target network. *Nat Biotechnol*, v. 25, n. 10, p. 1119–26, 2007.

ZHAO, X.-M. et al. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Res*, v. 36, n. 9, p. e48, 2008.

ZHU, X. et al. Adhesion-dependent cell cycle progression linked to the expression of cyclin D1, activation of cyclin E-cdk2, and phosphorylation of the retinoblastoma protein. *J Cell Biol*, v. 133, n. 2, p. 391–403, 1996.

Apêndices

Apêndice A - Teste estatístico de Wilcoxon

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

$$R^+ = \sum_{d_j>0} r(d_j) + \frac{1}{2}\sum_{d_j=0} r(d_j) \qquad R^- = \sum_{d_j<0} r(d_j) + \frac{1}{2}\sum_{d_j=0} r(d_j) \tag{6.1}$$

Determina-se, então, o valor de W, dado por $W = \min(R^+, R^-)$. Se houver mais de 15 diferenças, excluindo um termo caso o número de $d_j = 0$ seja ímpar, segue-se o teste de hipótese nula – os resultados entre os dois conjuntos de dados sob comparação são iguais – calculando-se o valor de z, dado por:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)N(2N+1)}} \tag{6.2}$$

onde N é o número final de diferenças utilizadas. Considerando $\alpha = 0.05$ como a probabilidade da hipótese nula ser verdadeira, podemos descartá-la se $z < -1,96$ (WILCOXON, 1947).

Caso $N < 15$, 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 6.1.

Tabela 6.1: Valores críticos (W_c) para o teste estatístico de Wilcoxon

N	W_c $\alpha = 0,05$	W_c $\alpha = 0,01$
6	1	–
7	2	–
8	4	0
9	6	2
10	8	3
11	11	5
12	14	7
13	18	9
14	22	12
15	26	15

Apêndice B - *GeneTrail*

O *GeneTrail* é uma ferramenta que compara estatisticamente as frequências de categorias funcionais de três diferentes projetos – *Kyoto Encyclopedia of Genes and Genomes* (*KEGG* (KANEHISA et al., 2008), *TRANSPATH* (KRULL et al., 2006) e *GO* (BERARDINI et al., 2010) – entre dois grupos de genes.

O método estatístico disponibilizado pelo *GeneTrail* para a comparação das frequências de categorias funcionais entre dois grupos de genes, especificamente quando um dos grupos é um subgrupo do outro, chamado de grupo de referência, é o teste hipergeométrico. Esse teste se baseia na distribuição hipergeométrica, distribuição de probabilidades discretas que descreve a probabilidade de se retirar x elementos do tipo C numa sequência de n extrações de uma população finita de tamanho N, com K elementos do tipo C e $N-K$ elementos de outro tipo, sem reposição.

O problema do enriquecimento de uma categoria funcional em um subgrupo em relação ao grupo de referência pode ser modelada por uma distribuição hipergeométrica e, portanto, o cálculo da significância estatística da comparação das frequências de um determinado processo biológico entre dois grupos de genes pode ser feita por um teste hipergeométrico. Essa significância equivale à probabilidade de se obter x genes pertencentes a uma certa categoria funcional C em um subgrupo com n genes retirado do grupo de referência de tamanho N com K genes pertencentes à C:

$$P_c = \sum_{i=x}^{n} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \tag{6.3}$$

P_c é a probabilidade de se obter um número observado de genes pertencentes à C no subgrupo de n genes igual ou maior ao número esperado de genes pertencentes à C em um subgrupo de n genes selecionado aleatoriamente a partir do grupo de referência com N genes.

Apêndice C - Análise de enriquecimento de vias do *KEGG PATHWAY* no *pred_ONCO* pelo *GeneTrail*

Segue abaixo tabela com o resultado da análise de enriquecimento de vias do *KEGG PATHWAY* no *pred_ONCO* tendo como referência a *RIGH*. Essa análise foi realizada pelo *GeneTrail* através do teste hipergeométrico como descrito acima. Foram consideradas vias significativamente enriquecidas na *pred_ONCO* em relação à *RIGH* aquelas com $p < 0,05$.

As vias envolvidas com câncer do *KEGG PATHWAY* estão destacadas em amarelo. O significado de cada coluna está descrito abaixo:

- Coluna 1: nome do banco de dados de origem das categorias sob análise (nesse caso, o KEGG);
- Coluna 2: nome da categoria (nesse caso, as vias de sinalização do *KEGG PATHWAY*);
- Coluna 3: código da via de sinalização segundo o *KEGG PATHWAY*;
- Coluna 4: número esperado de genes na categoria considerando como referência a *RIGH*;
- Coluna 5: número observado de genes na categoria presentes na *pred_ONCO*;
- Coluna 6: valor de p produzido pelo teste hipergeométrico e corrigido pelo método da taxa de falsas descobertas (fdr).

Categoria	Subcategoria	Subcategoria, nome alternativo	esperado	observado	p-value (fdr)
KEGG	Pathways in cancer	05200	16.3318	102	9.95472e-60
KEGG	Chronic myeloid leukemia	05220	3.94593	53	5.68984e-52
KEGG	Prostate cancer	05215	4.65839	45	2.80262e-34
KEGG	Pancreatic cancer	05212	3.83632	40	3.32228e-32
KEGG	ErbB signaling pathway	04012	4.60359	43	4.51325e-32
KEGG	T cell receptor signaling pathway	04660	5.64488	45	7.6442e-30
KEGG	Neurotrophin signaling pathway	04722	6.74097	45	6.79314e-26
KEGG	MAPK signaling pathway	04010	12.9887	59	2.82826e-24
KEGG	Glioma	05214	3.45269	32	1.01889e-23
KEGG	Cell cycle	04110	6.57655	42	2.37072e-23
KEGG	Colorectal cancer	05210	3.34308	30	9.20986e-22
KEGG	Endometrial cancer	05213	2.74023	27	5.042e-21
KEGG	Focal adhesion	04510	9.91964	47	4.70127e-20
KEGG	B cell receptor signaling pathway	04662	3.78152	30	6.71696e-20
KEGG	Renal cell carcinoma	05211	3.83632	30	1.046e-19
KEGG	Small cell lung cancer	05222	4.49398	32	1.51136e-19
KEGG	Acute myeloid leukemia	05221	3.06906	27	1.82278e-19
KEGG	Metabolic pathways	01100	51.626	5	3.87562e-19 down
KEGG	Non-small cell lung cancer	05223	2.95945	26	8.88684e-19
KEGG	Chemokine signaling pathway	04062	9.59081	43	1.81544e-17
KEGG	Toll-like receptor signaling pathway	04620	4.93242	30	3.49368e-16
KEGG	Melanoma	05218	3.34308	25	4.96162e-16
KEGG	Fc epsilon RI signaling pathway	04664	4.00074	26	6.68637e-15
KEGG	Adherens junction	04520	3.78152	24	1.5453e-13
KEGG	Bladder cancer	05219	2.24699	19	1.5453e-13
KEGG	Chagas disease	05142	5.48046	28	4.96049e-13
KEGG	Bacterial invasion of epithelial cells	05100	3.67191	23	6.7152e-13
KEGG	Insulin signaling pathway	04910	6.90538	31	9.8379e-13
KEGG	RIG-I-like receptor signaling pathway	04622	2.84984	20	2.12005e-12
KEGG	Jak-STAT signaling pathway	04630	7.39862	31	6.55949e-12
KEGG	TGF-beta signaling pathway	04350	4.05554	23	6.55949e-12
KEGG	NOD-like receptor signaling pathway	04621	3.23347	20	3.0215e-11
KEGG	Leishmaniasis	05140	3.78152	21	9.15856e-11
KEGG	Epithelial cell signaling in Helicobacter pylori infection	05120	3.17867	19	1.90667e-10
KEGG	Adipocytokine signaling pathway	04920	3.5623	20	2.08593e-10
KEGG	Apoptosis	04210	4.49398	22	4.4077e-10
KEGG	Endocytosis	04144	9.64561	32	1.65981e-09
KEGG	Wnt signaling pathway	04310	6.63136	26	1.7785e-09
KEGG	Fc gamma R-mediated phagocytosis	04666	4.87761	22	2.28641e-09
KEGG	Natural killer cell mediated cytotoxicity	04650	6.30253	25	2.83842e-09
KEGG	GnRH signaling pathway	04912	5.09683	22	5.36995e-09
KEGG	Shigellosis	05131	3.23347	17	1.50764e-08
KEGG	Ubiquitin mediated proteolysis	04120	6.41214	23	9.68488e-08
KEGG	VEGF signaling pathway	04370	3.67191	17	1.17821e-07
KEGG	Thyroid cancer	05216	1.53453	11	1.97278e-07
KEGG	Regulation of actin cytoskeleton	04810	10.1937	29	3.2834e-07
KEGG	Progesterone-mediated oocyte maturation	04914	4.54878	18	5.94559e-07
KEGG	Oocyte meiosis	04114	5.59007	19	3.26394e-06
KEGG	Cytosolic DNA-sensing pathway	04623	2.02777	11	4.64475e-06
KEGG	Leukocyte transendothelial migration	04670	5.31605	18	6.29447e-06
KEGG	Type II diabetes mellitus	04930	2.52101	12	7.13691e-06
KEGG	Dorso-ventral axis formation	04320	1.09609	8	8.83601e-06
KEGG	p53 signaling pathway	04115	3.28828	13	2.4946e-05
KEGG	Notch signaling pathway	04330	2.19218	10	6.59803e-05
KEGG	Amyotrophic lateral sclerosis (ALS)	05014	2.68543	11	7.85088e-05

KEGG	Melanogenesis	04916	4.7132	15	8.10008e-05				
KEGG	Axon guidance	04360	5.69968	15	0.000718994				
KEGG	Pathogenic Escherichia coli infection			05130	2.90464	10	0.000753058		
KEGG	Gap junction	04540	4.60359	13	0.00083727				
KEGG	mTOR signaling pathway			04150	2.68543	8	0.00689705		
KEGG	Basal cell carcinoma	05217	2.24699	7	0.00906738				
KEGG	Phosphatidylinositol signaling system			04070	3.891139	0.0210308			
KEGG	Tight junction	04530	5.9737	12	0.0216966				
KEGG	Amoebiasis	05146	5.37085	11	0.0242674				
KEGG	Aldosterone-regulated sodium reabsorption				04960	2.19218	6	0.0279161	
KEGG	Long-term depression	04730	3.45269	8	0.0279161				
KEGG	Phagosome	04145	6.85058	2	0.0378275				
KEGG	Long-term potentiation	04720	3.67191	8	0.0378724				
KEGG	Prion diseases	05020	1.86336	5	0.0467517				

Apêndice D - Trabalho publicado no periódico *Physica A*

Trabalho publicado em 2008 onde os autores descrevem a utilização de aprendizado de máquina e medidas de centralidade da rede integrada de interações gênicas da bactéria *Escherichia coli* para a predição de genes essenciais e descoberta de regras para essencialidade nesse organismo. As contribuições do autor desta tese para esse trabalho foram a interpretação dos resultados e a preparação do manuscrito.

Available online at www.sciencedirect.com
ScienceDirect
Physica A 387 (2008) 1049–1055
PHYSICA A
www.elsevier.com/locate/physa

In silico network topology-based prediction of gene essentiality

João Paulo Müller da Silva[a], Marcio Luis Acencio[a], José Carlos Merino Mombach[b], Renata Vieira[c], José Camargo da Silva[c], Ney Lemke[a,*], Marialva Sinigaglia[c]

[a] *Department of Physics and Biophysics, Institute of Biosciences, São Paulo State University, UNESP, 18618-000, Botucatu, SP, Brazil*
[b] *Centro de Ciências Rurais, Unipampa/São Gabriel - Pós-Graduação em Física, Prédio 13, Universidade Federal de Santa Maria, 97105-900, Santa Maria, Brazil*
[c] *Programa Interdisciplinar de Computação Aplicada, Universidade do Vale do Rio dos Sinos, 93022-000 São Leopoldo, RS, Brazil*

Received 20 September 2007
Available online 26 October 2007

Abstract

The identification of genes essential for survival is important for the understanding of the minimal requirements for cellular life and for drug design. As experimental studies with the purpose of building a catalog of essential genes for a given organism are time-consuming and laborious, a computational approach which could predict gene essentiality with high accuracy would be of great value. We present here a novel computational approach, called *NTPGE* (Network Topology-based Prediction of Gene Essentiality), that relies on the network topology features of a gene to estimate its essentiality. The first step of *NTPGE* is to construct the integrated molecular network for a given organism comprising protein physical, metabolic and transcriptional regulation interactions. The second step consists in training a decision-tree-based machine-learning algorithm on known essential and non-essential genes of the organism of interest, considering as learning attributes the network topology information for each of these genes. Finally, the decision-tree classifier generated is applied to the set of genes of this organism to estimate essentiality for each gene. We applied the *NTPGE* approach for discovering the essential genes in *Escherichia coli* and then assessed its performance.

PACS: 87.16.dr; 87.16.Yc; 87.18.Cf

Keywords: Biological networks; Complex systems; Gene essentiality; Machine learning

1. Introduction

Essential genes are genes that are indispensable to support cellular life. These genes constitute a minimal set of genes required for a living cell. Therefore, the functions encoded by this gene set are essential and could be considered as a foundation of life itself [1,2]. The identification of genes which are essential for survival is important not only for the understanding of the minimal requirements for cellular life, but also for practical purposes. For example, since most antibiotics target essential cellular processes, essential gene products of microbial cells are promising new targets

* Corresponding author. Tel.: +55 5138153263.
E-mail address: lemke@ibb.unesp.br (N. Lemke).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2007.10.044

for such drugs [3]. The prediction and discovery of essential genes have been performed by experimental procedures such as single gene knockouts [4], RNA interference [5] and conditional knockouts [6], but each of these techniques require a large investment of time and resources and they are not always feasible.

Considering these experimental constraints, a computational or *in silico* approach capable of accurately predicting gene essentiality would be of great value. Some such predictors have already been developed in which sequence features of genes and proteins with or without homology comparison have been utilized as parameters for training machine-learning classifiers for gene essentiality prediction [7,8]. In addition, predictors of gene essentiality based on network topology features, such as the physical interactions of a protein [9] or the number of biochemical species that are knocked out from the metabolic network following a gene deletion [10,11] have also been developed.

The currently available network topology-based methodologies of gene essentiality prediction use only one type of network topology feature, i.e. protein physical interactions or metabolic interactions, for performing such predictions. Actual molecular interaction networks, however, are composed by entities that are intricately connected with diverse types of interactions, such as protein physical, metabolic and transcriptional regulation interactions.

We therefore propose here a novel machine-learning-based *in silico* approach, called *NTPGE* (Network Topology-based Prediction of Gene Essentiality), that relies on multiple topological network features of a given gene to estimate its essentiality. For the generation of the decision-tree classifier, NTPGE employs the following network topological features as learning attributes: number of physical interactions for the corresponding encoded protein, number of target genes transcriptionally regulated by the corresponding encoded transcription factor, number of transcription factors that regulate it, number of enzymes that use metabolites produced by the corresponding encoded enzyme as reactants and number of enzymes that produce metabolites used as reactants by the corresponding encoded enzyme. To assess the performance of the *NTPGE* approach, we used it for the discovery of essential genes in the bacterium *E. coli*, a model organism whose majority of genes have already been characterized experimentally as essential or non-essential.

2. Construction of the IMN of *E. coli*

As *NTPGE* relies on topological features of molecular network, the first step was to construct the *E. coli* integrated molecular network (IMN) comprising protein physical, metabolic and transcriptional regulation interactions. For this purpose, we used MONET (MOlecular NETwork) ontology, a tool developed by our group that facilitates the construction of IMNs of organisms via integration of information from metabolic pathways, protein–protein interaction networks and transcriptional regulation interactions through a model able to minimize data redundancy and inconsistency [12]. As previously described, two genes of a given organism, g_1 and g_2, coding for proteins p_1 and p_2 are linked if:

- p_1 and p_2 interact physically,
- g_1 regulates the transcription of gene g_2,
- or one metabolite generated by a reaction catalyzed by p_1 is consumed in a reaction catalyzed by p_2 (we may exclude from this analysis the most used compounds such as ATP, NAD, H2O, etc.).

The data sources present in MONET ontology used for the construction of the *E. coli* IMN were KEGG (Kyoto Encyclopedia of Genes and Genomes) [13] for metabolic interactions, RegulonDB [14] for transcriptional regulation interactions, and Butland et al. [15] for protein physical interactions.

Using MONET, we constructed two directed IMNs of *E. coli*, G_a and G_p. G_a contained all possible interactions among genes with 1998 genes and 51,642 interactions. G_p was similar to G_a, except that the connections through the ten most frequently used compounds on the metabolism were deleted producing a network with 1987 genes and 21,338 interactions, since connections via these common compounds are not likely to be important for the determination of gene essentiality due to their promiscuity.

3. Brief analysis of the *E. coli* IMNs

Prior to the use of *E. coli* IMNs for the validation of the *NTPGE* approach, we present here a brief analysis of the most common network measures, i.e. degree distribution and clustering coefficient, of these IMNs. The degree distribution, $P(k)$, gives the probability that a selected node has exactly k links. $P(k)$ is obtained by counting the

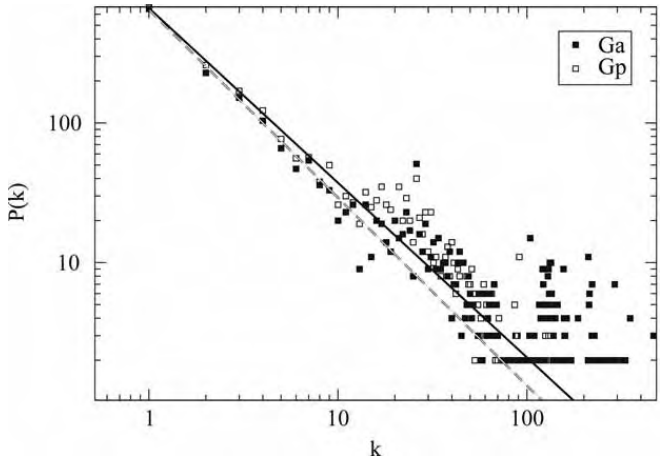

Fig. 1. Histogram of the degree distribution for G_a and G_p used in this work. Both G_a (solid line) and G_p (dashed line) are well-described by a power law function $P(k) = Ak^{-\gamma}$ that characterizes them as scale-free networks.

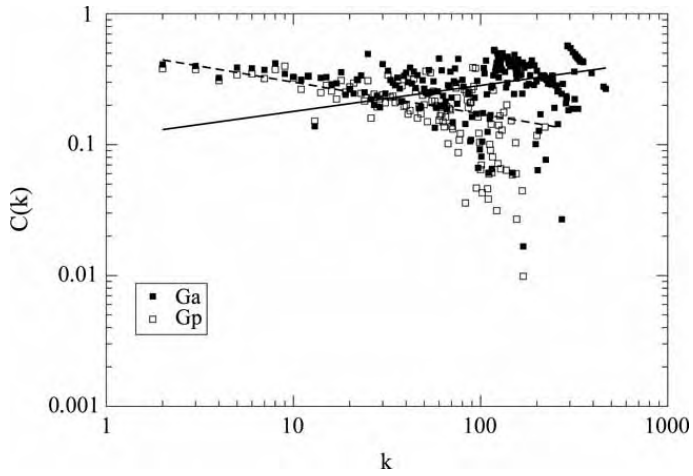

Fig. 2. The dependence of the average clusterization coefficient C on the connectivity k. The best-fit regression line for G_a (solid line) has a regression slope of -0.03 with a confidence interval of $[-0.08, 0.01]$, while the best-fit regression line for G_p (dashed line) has a regression slope of 0.28 with a confidence interval of [0.22, 0.33]. The results show that G_a is a non-hierarchical scale-free network, whereas G_p is a hierarchical scale-free network.

number of nodes $N(k)$ with $k = 1, 2, \ldots$ links and dividing by the total number of nodes N. The clustering coefficient, C_i, gives the density of triangles we can construct in the network having the node i as a vertex. The clusterization coefficient is defined as:

$$C_i = \frac{2n_i}{k_i(k_i - 1)}, \tag{1}$$

where n_i is the number of links connecting the k_i neighbors of the node i. The average clustering coefficient C is the clustering coefficient for the whole network and characterizes the overall tendency of nodes to form clusters or groups.

In Fig. 1 we show the histogram of degree distribution for G_a and G_p. The curves are well-approximated by a power law function, $P(k) = Ak^{-\gamma}$ for both the IMNs, suggesting that G_a and G_p are scale-free networks.

We also analyzed the dependence of the average clusterization coefficient, C, on the connectivity k, defined as $C(k)$. For a traditional scale-free network, we expect $C(k)$ not to depend on k, while for hierarchical networks we expect $C(k) \sim k^{-\alpha}$. Fig. 2 shows the $C(k)$ for G_a and G_p. These results point to a $C(k)$ not dependent on k for G_a and a $C(k)$ dependent on k for G_p, thus indicating that G_a is a non-hierarchical IMN and G_p is a hierarchical

Table 1
Parameters used to run the J48 algorithm on training data

Parameter	Value
binarySplit	False
confidenceFactor	0.25
debug	False
minNumObj	100
numFolds	3
reduceErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False

IMN. This shift from a non-hierarchical topology for G_a to a hierarchical topology for G_p seems to be caused by the deletion of the connections through the ten most frequently used compounds in the metabolism on the construction of G_p. Such compounds induce a strongly connected IMN due to their promiscuity.

4. Description of the *NTPGE* approach

The *NTPGE* approach was performed using WEKA (*Waikato Environment for Knowledge Analysis*) system [16]. WEKA is a collection of machine-learning algorithms for data mining tasks. It also provides means for data pre-processing, classification, regression, clustering, association rules, and visualization [16]. Among these algorithms, we used the J48 [16], which is the Weka's implementation of the well-known C4.5 [17] that uses the greedy technique to induce decision trees for classification. A decision-tree model is built by analyzing training data, which is then used to classify unseen data.

We trained the J48 algorithm on four different training configurations (t_1, t_2, t_3 and t_4). In all the configurations, the training data was a set of known essential and non-essential genes of *E. coli* taken from the PEC database (*Profiling of E. coli chromosome*, http://www.shigen.nig.ac.jp/ecoli/pec/). The PEC database has been compiled on experimental information on *E. coli* strains from research reports and deletion mutation studies prior to 1998, including gene essentiality for cell growth. Based on these reports about gene essentiality for cell growth, the *E. coli* genes are classified in essential, non-essential and unknown. In all the training configurations, for a given gene, the learning attributes used were as follows:

- number of physical interactions for the corresponding encoded protein;
- number of target genes transcriptionally regulated by the corresponding encoded transcription factor (`regulation_out`);
- number of transcription factors that regulate it; (`regulation_in`);
- number of enzymes that use metabolites produced by the corresponding encoded enzyme as reactants (`metabolism_out`);
- number of enzymes that produce metabolites used as reactants by the corresponding encoded enzyme (`metabolism_in`);

In t_1 and t_2, the above mentioned attributes were extracted from G_a, whereas these same attributes were extracted from G_p in t_3 and t_4. Moreover, the attribute *damage*, which was not originally present in G_a and G_p, was included in t_2 and t_4, The damage d is defined as the number of metabolites whose production was prevented by the deletion of the enzyme. For a given enzyme, its damage d has been shown to be strongly correlated to its essentiality [18].

The J48 algorithm was trained with the parameters presented in Table 1. As it is known that data imbalance is one of the causes that degrade the performance of machine-learning algorithms [19], we replicated the data related to the essential genes in order to correct data imbalance as the number of non-essential genes is much larger than the number of essential genes.

Table 2
Confusion matrices of the classifiers generated from t_1, t_2, t_3 and t_4

	Predicted		Actual
	Non-essential	Essential	
t_1	1392	397	Non-essential
	310	1780	Essential[a]
t_2	1348	405	Non-essential
	313	1777	Essential[a]
t_3	1346	432	Non-essential
	298	1792	Essential[a]
t_4	1348	430	Non-essential
	300	1790	Essential[a]

[a] The number of essential genes were replicated to avoid data imbalance. Actually, the number of essential genes is 209.

Table 3
Features of the training configurations and performance measures of their corresponding generated classifiers

Features and performance measures	Training configurations			
	t_1	t_2	t_3	t_4
Number of Genes[a]	3879	3879	3868	3868
Damage d	No	Yes	No	Yes
Correctly Predicted Genes (%)	81.8	81.5	81.1	81.1
Incorrectly Predicted Genes (%)	18.2	18.5	18.9	18.9
F-measure (N) (%)	79.7	79.4	78.7	78.7
F-measure (E) (%)	83.4	83.2	83.1	83.1
Recall (N) (%)	77.8	77.4	75.7	75.8
Recall (E) (%)	85.2	85.0	85.7	85.6
Precision (N) (%)	81.8	81.6	81.9	81.8
Precision (E) (%)	81.8	81.4	80.6	80.6

[a] The number of essential genes were replicated to avoid data imbalance; number of non-essential genes remained unchanged. Actually, the number of essential genes is 209 and non-essential genes is 1789 for G_a and the number of essential genes is 209 and non-essential genes is 1778 for G_p.

5. Performance of the *NTPGE* approach and related discussion

The performance of the *NTPGE* approach was evaluated by testing the classifiers created by the J48 algorithm, as described above, on the training data itself. The selection of the best training configuration to be considered as default by the *NTPGE* approach was performed based on the *F-measure* of the corresponding generated classifier. The *F-measure* provides a harmonic mean of precision and recall and is defined as:

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{2}$$

Precision (the percentage of correctly classified instances) and recall (the percentage of positive labeled instances that were classified as such) were calculated from the confusion matrices of the classifiers obtained from the training configurations t_1, t_2, t_3 and t_4 (Table 2) and are shown in Table 3. Table 3 also shows the *F-measure* as well as the features of the training configurations, as the number of instances (genes plus metabolites) and presence or absence of the learning attribute damage d on training.

According to Table 3, the best training configuration was t_1 (all genes and metabolites with the attribute damage). Its corresponding generated classifier had an *F-measure* of 83.4% for essential genes and 79.7% for non-essential genes. In fact, all the generated classifiers yielded similar results, suggesting that the presence or absence of the ten most used compounds in metabolism or the presence or absence of the attribute damage d did not affect the classification of genes as essential or non-essential by the *NTPGE* approach. Therefore, any training configuration could be selected as default by *NTPGE*.

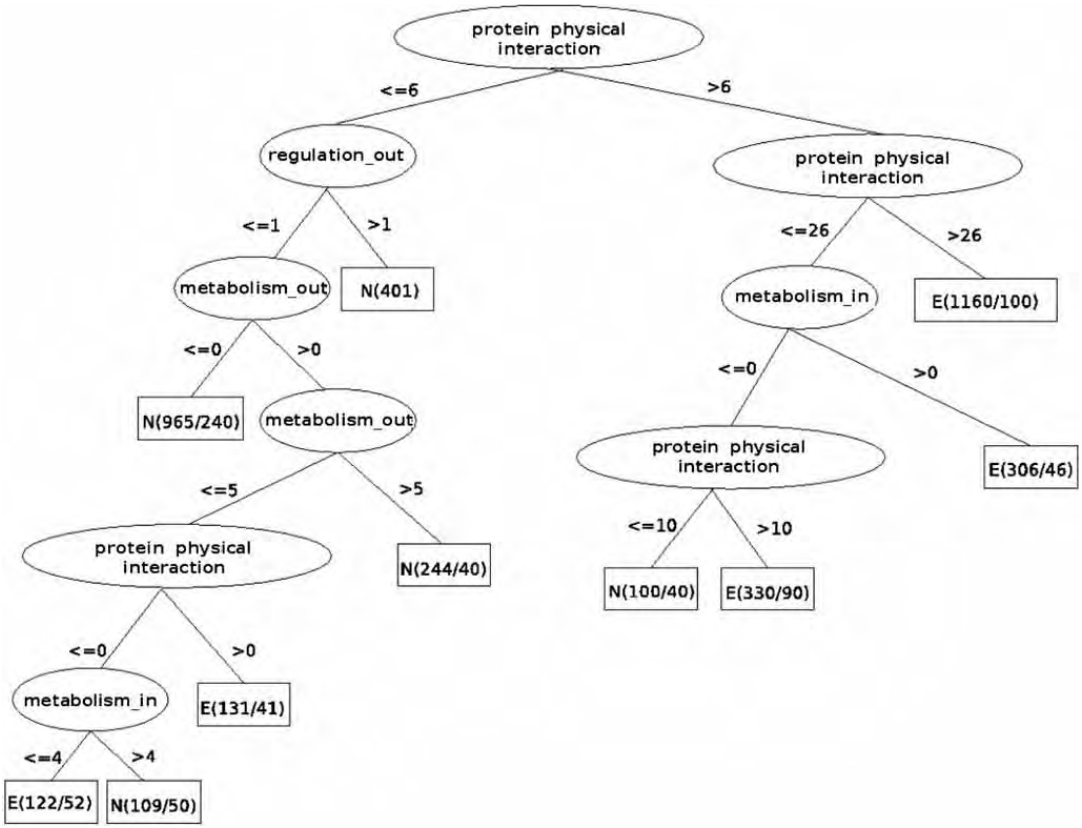

Fig. 3. Decision tree generated from t_1 with an *F-measure* of 83.4% for essential genes (E) and 79.7% for non-essential genes (N). The (x/y) inside rectangles denotes the number of correctly classified genes (x) and the number of incorrectly classified genes (y).

Fig. 3 shows the set of rules of the decision tree generated from t_1. The top node of the tree corresponds to the attribute protein physical interaction. This means that the classification-tree algorithm concluded that the main factor to define essentiality in *E. coli* was the protein physical interaction. In fact, the degree of a protein has been documented in the literature as being indicative of essentiality in various organisms [9,20,21]. In our approach, a combination of intermediate number of protein physical interactions with at least one interaction of the type `metabolism_in`, i.e. number of enzymes that produce metabolites used as reactants by the corresponding encoded enzyme, was also indicative of essentiality. Transcriptional regulation interactions seems not to be a good predictor for gene essentiality, since genes with at least one interaction of the type `regulation_out`, i.e. number of target genes transcriptionally regulated by the corresponding encoded transcription factor, were classified as non-essential. Moreover, the attribute (`regulation_in`, i.e. the number of transcription factors that regulate a given gene, was not even included in the decision tree. These results regarding gene essentiality and transcriptional regulation are not surprising, since transcription factors are usually not essential under the conditions in which the knockout experiments for determining gene essentiality are performed (PEC database, http://www.shigen.nig.ac.jp/ecoli/pec/).

6. Concluding remarks

We proposed here a novel machine-learning-based computational approach, called *NTPGE* (Network Topology-based Prediction of Gene Essentiality), that relies on network topology features of a gene to estimate its essentiality. Distinct from previous network topology-based gene essentiality predictors, *NTPGE* employs multiple topological network features of a given gene to estimate its essentiality, namely physical interactions for the corresponding encoded protein, number of target genes transcriptionally regulated by the corresponding encoded transcription factor, number of transcription factors that regulate it, number of enzymes that use metabolites produced by the corresponding encoded enzyme as reactants and number of enzymes that produce metabolites used as reactants by the corresponding encoded enzyme.

We verified the performance of *NTPGE* by applying it to the discovery of essential genes in the bacterium *E. coli*, a model organism whose majority of genes have already been characterized experimentally as essential or non-essential. Among the interactions considered as learning attributes, *NTPGE* relied mostly on protein physical and metabolic interactions for gene essentiality prediction. In addition, the presence or absence of the ten most used compounds in metabolism or the presence or absence of the attribute damage d did not likely influence the classification of genes as essential or non-essential by *NTPGE*. This can be concluded because the *F-measure* values of all generated decision trees were similar. Anyway, the best classifier was generated from t_1 (all genes and metabolites with the attribute damage) with an *F-measure* of 83.4% for essential genes and 79.7% for non-essential genes.

In conclusion, the *NTPGE* seems to be a reliable method of gene essentiality discovery that may be applied to the gene set of other organisms. However, *NTPGE* is limited to organisms whose corresponding IMN has already been constructed. The construction of the IMN of a given organism involves the gathering of experimentally determined data that are not always available, particularly for a newly sequenced organism. To overcome this limitation, future developments would be the integration of *NTPGE* with sequence-based methods of IMN construction, thus creating a purely *in silico* network topology information-based methodology of gene essentiality discovery.

Acknowledgements

We would like to thank CNPq (research grants 474278/2006-9 and 506414/2004-3), FAPESP (research grant 2007/02827-9) and FAPERGS (05600005-BRD) for supporting this work. We would also like to thank HP Brazil R&D for the collaboration.

References

[1] K. K, S. Ehrlich, A. Albertini, G. Amati, K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, S. Aymerich, P. Bessieres, et al., Proc. Natl. Acad. Sci. USA 100 (2003) 4678.
[2] I. M, FEBS Lett. 362 (1995) 257.
[3] J. N, J. Mekalanos, Nature Biotechnol. 18 (2000) 740.
[4] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A.P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D.J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J.H. Hegemann, S. Hempel, Z. Herman, D.F. Jaramillo, D.E. Kelly, S.L. Kelly, P. Kötter, D. LaBonte, D.C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S.L. Ooi, J.L. Revuelta, C.J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D.D. Shoemaker, S. Sookhai-Mahadeo, R.K. Storms, J.N. Strathern, G. Valle, M. Voet, G. Volckaert, C. yun Wang, T.R. Ward, J. Wilhelmy, E.A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J.D. Boeke, M. Snyder, P. Philippsen, R.W. Davis, M. Johnston, Nature 418 (2002) 387.
[5] L.M. Cullen, G.M. Arndt, Immunol. Cell. Biol. 83 (2005) 217.
[6] T. Roemer, B. Jiang, J. Davison, T. Ketela, K. Veillette, A. Breton, F. Tandia, A. Linteau, S. Sillaots, C. Marta, N. Martel, S. Veronneau, S. Lemieux, S. Kauffman, J. Becker, R. Storms, C. Boone, H. Bussey, Mol. Microbiol. 50 (2003) 167.
[7] M. Seringhaus, A. Paccanaro, A. Borneman, M. Snyder, M. Gerstein, Genome Res. 16 (2006) 1126.
[8] A.M. Gustafson, E.S. Snitkin, S.C.J. Parker, C. DeLisi, S. Kasif, BMC Genomics 7 (2006) 265.
[9] H. Jeong, S.P. Mason, A.L. Barabási, Z.N. Oltvai, Nature 411 (2001) 41.
[10] M. Imieliński, C. Belta, A. Halász, H. Rubin, Bioinformatics 21 (2005) 2008.
[11] M.C. Palumbo, A. Colosimo, A. Giuliani, L. Farina, FEBS Lett. 579 (2005) 4642.
[12] J.P.M. daSilva, N. Lemke, J.C. Mombach, J.G.C. deSouza, M. Sinigaglia, R. Vieira, Genet. Mol. Res. 5 (2006) 182.
[13] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, Nucleic Acids Res. 34 (2006) D354.
[14] H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Díaz-Peredo, F. Sánchez-Solano, A. Santos-Zavaleta, I. Martínez-Flores, V. Jiménez-Jacinto, C. Bonavides-Martínez, J. Segura-Salazar, A. Martínez-Antonio, J. Collado-Vides, Nucleic Acids Res. 34 (2006) D394.
[15] G. Butland, J.M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, A. Emili, Nature 433 (2005) 531.
[16] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, 2000.
[17] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, 1993.
[18] N. Lemke, F. Herédia, C.K. Barcellos, A.N.D. Reis, J.C.M. Mombach, Bioinformatics 20 (2004) 115.
[19] P. Kang, S. Cho, Lecture Notes in Comput. Sci. 4232 (2006) 837.
[20] E. Estrada, Proteomics 6 (2006) 35.
[21] S. Wuchty, Genome Res. 14 (2004) 1310.

Apêndice E - Trabalho publicado no periódico *BMC Bioinformatics*

Trabalho publicado em 2009 onde os autores descrevem a utilização de aprendizado de máquina e dados de sublocalização celular, participação em processos biológicos e medidas de centralidade da rede integrada de interações gênicas do fungo *Saccharomyces cerevisiae* para a predição de genes essenciais e descoberta de condições associadas à essencialidade dos genes nesse organismo. O autor desta tese é o autor principal do trabalho.

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information

Marcio L Acencio* and Ney Lemke

Address: Department of Physics and Biophysics, São Paulo State University, Distrito de Rubiao Jr. s/n, Botucatu, São Paulo, Brazil

Email: Marcio L Acencio* - mlacencio@ibb.unesp.br; Ney Lemke - lemke@ibb.unesp.br

* Corresponding author

Published: 16 September 2009

BMC Bioinformatics 2009, **10**:290 doi:10.1186/1471-2105-10-290

Received: 31 October 2008

Accepted: 16 September 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/290

Abstract

Background: The identification of essential genes is important for the understanding of the minimal requirements for cellular life and for practical purposes, such as drug design. However, the experimental techniques for essential genes discovery are labor-intensive and time-consuming. Considering these experimental constraints, a computational approach capable of accurately predicting essential genes would be of great value. We therefore present here a machine learning-based computational approach relying on network topological features, cellular localization and biological process information for prediction of essential genes.

Results: We constructed a decision tree-based meta-classifier and trained it on datasets with individual and grouped attributes-network topological features, cellular compartments and biological processes-to generate various predictors of essential genes. We showed that the predictors with better performances are those generated by datasets with integrated attributes. Using the predictor with all attributes, i.e., network topological features, cellular compartments and biological processes, we obtained the best predictor of essential genes that was then used to classify yeast genes with unknown essentiality status. Finally, we generated decision trees by training the J48 algorithm on datasets with all network topological features, cellular localization and biological process information to discover cellular rules for essentiality. We found that the number of protein physical interactions, the nuclear localization of proteins and the number of regulating transcription factors are the most important factors determining gene essentiality.

Conclusion: We were able to demonstrate that network topological features, cellular localization and biological process information are reliable predictors of essential genes. Moreover, by constructing decision trees based on these data, we could discover cellular rules governing essentiality.

Background

Essential genes are those genes required for growth in a rich medium, i.e., medium containing all nutrients required for growth. The deletion of only one of these genes is sufficient to confer a lethal phenotype on an organism regardless the presence of remaining genes. Therefore, the functions encoded by essential genes are crucial for survival and could be considered as a foundation of life itself [1,2]. The identification of essential genes is important not only for the understanding of the mini-

mal requirements for cellular life, but also for practical purposes. For example, since most antibiotics target essential cellular processes, essential gene products of microbial cells are promising new targets for such drugs [3]. The prediction and discovery of essential genes have been performed by experimental procedures such as single gene knockouts [4], RNA interference [5] and conditional knockouts [6], but these techniques require a large investment of time and resources and they are not always feasible. Considering these experimental constraints, a computational approach capable of accurately predict essential genes would be of great value.

For prediction of essential genes, some investigators have implemented computational approaches in which most are based on sequence features of genes and proteins with or without homology comparison [7,8]. With the accumulation of data derived from experimental small-scale studies and high-throughput techniques, however, it is now possible to construct networks of gene and proteins interaction and then investigate whether the topological properties of these networks would be useful for predicting essential genes. Although many interaction networks have been built to date [9-12], most of studies relating essentiality with topological properties of these networks have been limited to indicate what topological properties are predictive of essentiality instead of using them as predictors of essential genes [9,13]. We have previously shown the feasibility of using network topological features for predicting essential genes in the bacterium *Escherichia coli* [14]. We have chosen *E. coli* as starting point for evaluating the prediction performance of essential genes by network topological features due to two reasons: the completeness of the catalog of *E. coli* essential genes [15] and the vast amount of interaction data available for this organism. In this present work, we sought to evaluate if network topological features can also be used as predictors of essential genes in the yeast *S. cerevisiae* since most of its genes have already been classified as essential or non-essential [4] and there are copious amounts of available interaction data for this organism.

For this purpose, we constructed a *S. cerevisiae* integrated network of gene interactions containing simultaneously protein physical, metabolic and transcriptional regulation interactions and used the topological features of this network as learning attributes in a machine learning-based prediction system. We tested individual and grouped network topological features as predictors of essential genes and showed that essential genes are best predicted by integrating the topological features in a single predictor. Although the prediction performance of topological features was shown to be acceptable, we added to this set of learning attributes data on cellular localization and biological process of genes in order to increase the predictability of essential genes. We found that the integration of network topology, cellular localization and biological process information in a single predictor increased the predictability of essential genes in comparison with the predictor containing only network topological features. Moreover, we observed that the predictability of essential genes by integration of cellular localization and biological process data in a single predictor was comparable to that of predictor containing network topological features.

Finally, in addition to study the predictability of essential genes, we tried to define some general rules governing essentiality in *S. cerevisiae* by analyzing decision trees generated by a machine learning-based technique. Using network topology, cellular localization and biological process information as training attributes, we discovered that essentiality depends on the number of protein physical interactions, the nuclear localization of proteins and the number of regulating transcription factors. Taken together, all these findings show that the integration of network analysis along with cellular localization and biological process information is a powerful tool for both predicting biological characteristics of genes, such as essentiality, and discovering the biological determinants of phenotypes.

Results and Discussion

Integrated network of gene interactions in S. cerevisiae and calculation of topological features

For obtaining the network topological features used as training data for predicting essential genes, we first constructed an integrated network of gene interactions (INGI) of *Saccharomyces cerevisiae* simultaneously containing experimentally verified protein physical interactions, metabolic interactions and transcriptional regulation interactions (definitions for each type of interaction are detailed in "Methods"). This network is comprised by 5,667 genes interacting with one another via 42,893 protein physical interactions, 11,192 metabolic interactions and 18,721 transcriptional regulation interactions. Of 5,667 genes in the network, 5,637 (99,5%) are protein-coding genes (including transposable elements), 15 (0.26%) are transfer RNA-coding genes, 13 (0.23%) are small nucleolar RNA-coding genes and 2 (0.01%) are RNA-coding genes of unknown function. Regarding protein-coding genes, including transposable elements, our network contains 96% of the total 5,884 protein-coding genes of *S. cerevisiae* according to the current status of the yeast genome provided by the *Saccharomyces* Genome Database (SGD) [16].

We calculated 12 different topological features for each gene in the INGI, including degree centralities for each type of interaction, clustering coefficient, betweenness centralities for each type of interaction, closeness centrality and identicalness. The detailed description of these

topological features and how they were calculated are found in the Additional file 1 and "Methods".

Comparison of the classification performance among balanced datasets

The performance of machine learning-based approaches is known to be affected by imbalanced data [17]. As the dataset containing yeast genes classified into essential and non-essential genes intended to be used as training data for our classifier is an imbalanced dataset, we used an undersampling scheme to generate ten balanced datasets from the original data (see "Methods"). Each balanced dataset contains different subsets of non-essential genes as a result of the sampling approach. Due to these different subsets of non-essential genes, therefore, we statistically compared the prediction performance of balanced datasets before assessing the predictability of essential genes by the different features. We trained our classifier on each of the balanced dataset with all available training data (network topological features and cellular localization and biological process information) and evaluated the prediction performance of each balanced dataset. Comparing the Area Under the receiver operating characteristic (ROC) Curve (AUC) values among all the balanced datasets (Figure 1 and Additional file 2), we verified that their prediction performances are not statistically different as evaluated by a nonparametric statistical method based on the Mann-Whitney U-statistic [18] (see more details in "Methods"). Based on these results, we selected one of the balanced datasets to perform the following analyses.

Prediction of essential genes by network topological features

We started the analyzes by assessing the predictability of essential genes by each of the 12 network topological features (computed as described in " Methods") and by all 12 network topological features integrated in a single predictor. For this purpose, we trained our classifier on a balanced dataset with all network topological features as training data and on a dataset containing only one of the network topological features as training data (see "Methods" for detailed information on construction of the balanced datasets). The ROC plot shown in Figure 2 indicates that integration of all networks topological features in a single predictor outperforms the predictability of essential genes by the individual network topological features. By

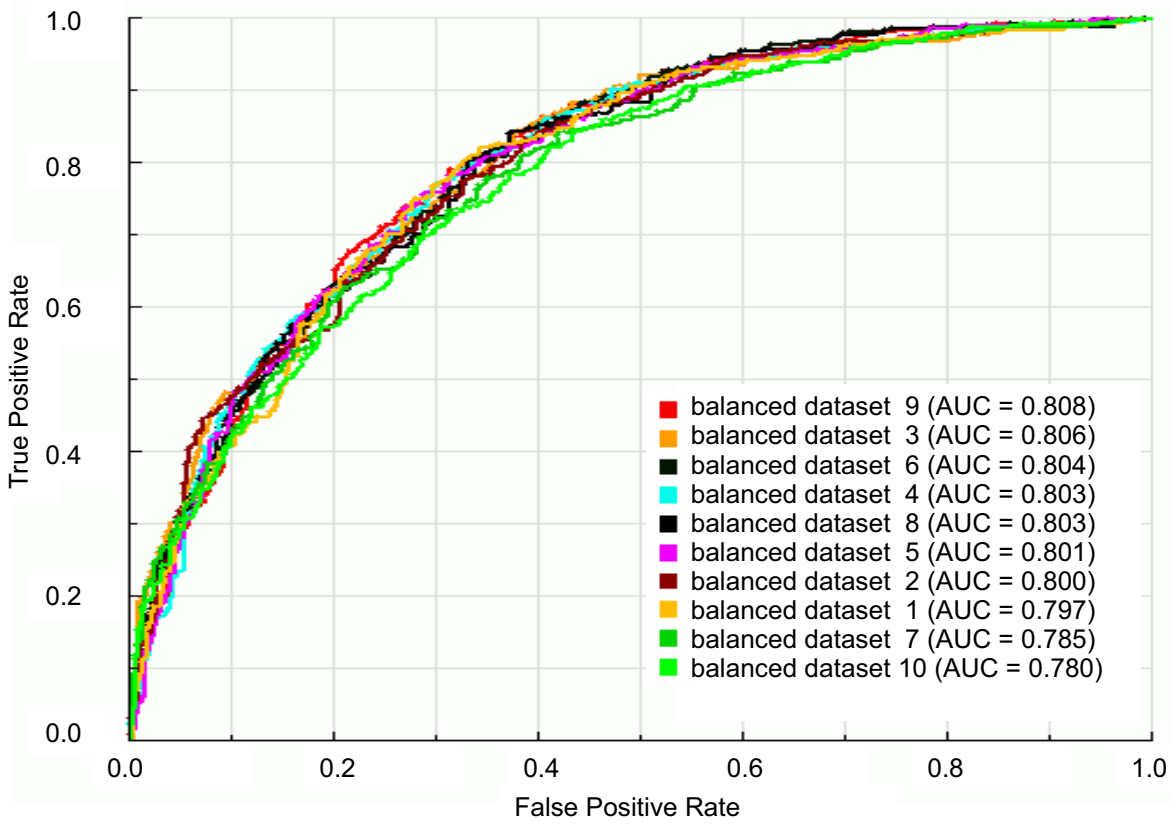

Figure 1
ROC curves and AUC values for classifiers trained on the ten balanced datasets with all available learning attributes. Balanced datasets 1-10: datasets with all available learning attributes prepared by an undersampling scheme as described in "Methods".

comparing the AUC values of grouped and individual network topological features, we verified that the AUC value of grouped network topological features (AUC = 0.773) is statistically significantly higher ($P < 0.002$) than AUC value of any individual network topological feature (Figure 2 and Additional file 2).

We then verified if different combinations of grouped network topological features could show prediction performances comparable to that of all grouped network topological features. We found that the combination of protein physical interactions-related features with metabolic interactions-related features has the same performance (AUC = 0.765, $P = 0.302$; see Additional file 2 and Figure 2) seen for the predictor containing all grouped network topological features (AUC = 0.773). Also, the combination of protein physical interactions-related features with clustering coefficient, identicalness and betweenness and closeness centralities has the same prediction performance (AUC = 0.763, $P = 0.071$; see Addi-

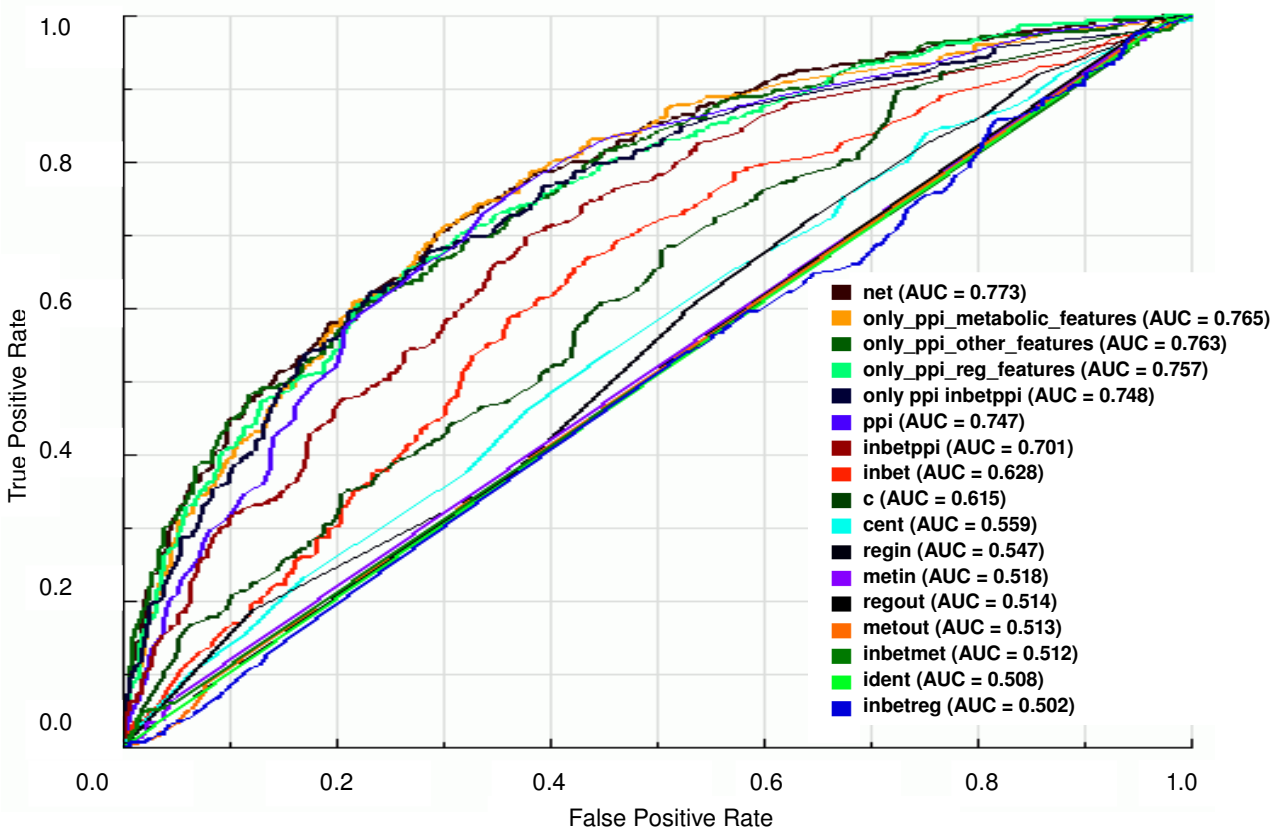

Figure 2
ROC curves and AUC values for the classifiers trained on balanced datasets with individual or grouped network topological features. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with one or groups of network topological features as learning attributes as follows: "net": all network topological features as learning attributes; "ppi", "inbetppi", "inbet", "c", "cent", "regin", "metin", "regout". "metout", "inbetmet", "ident" and "inbetreg": datasets with only one of the following network topological features as learning attribute: number of protein physical interactions (*ppi*), betweenness centrality for the protein physical interactions (*inbetppi*), betweenness centrality for all types of interactions (*inbet*), clustering coefficient (*c*), closeness centrality (*cent*, number of regulating transcription factor (*regin*), number of reactants participating in a metabolic reaction catalyzed by the enzyme encoded by the gene (*metin*), number of genes regulated by the transcription factor encoded by the gene (*regout*), number of products generated in a metabolic reaction catalyzed by the enzyme encoded by the gene (*metout*), betweenness centrality for the metabolic interactions (*inbetmet*), number of genes with identical topological features (*ident*) and betweenness centrality for the transcriptional regulation interactions (*inbetreg*). "only _ppi_metabolic_features" and "only_ppi_reg_features": datasets containing protein physical interactions-related features (*ppi* and *inbetppi*) and, respectively, metabolic (*met*, *metin*, *metout* and *inbetmet*) and transcriptional regulatory interactions-related features (*reg*, *regin*, *regout* and *inbetreg*). "only_ppi_other_features": dataset containing protein physical interactions-related features (*ppi* and *inbetppi*) and *c*, *ident*, *cent* and *inbet*. "only_ppi_inbetppi": dataset containing only the indicated network topological features as learning attributes. For more details on network topological features, see Additional file 1.

tional file 2 and Figure 2) observed for all grouped network topological features (AUC = 0.773). Therefore, smaller sets of network topological features can be used to predict essential genes, thus making the calculation of all topological features dispensable.

To verify if the predictive power of all grouped network topological features could be improved by exclusion of topological features with marginal AUC values, i.e., AUC values ranging from 0.500 to 0.600, we compared the prediction performance of all grouped network topological features (AUC = 0.773) with those of the combinations of features in which one feature or a small set of features was excluded (see the correspondent ROC curves in the Additional file 3 and the pairwise comparison of predictors with the p-values of AUC differences between each pair of predictors in Additional file 2). We discovered that the prediction performance of all grouped network topological features is not improved by the removal of any individual or small sets of topological features (see Additional files 2 and 3). As expected, the exclusion of grouped features related to metabolic interactions or grouped features related to protein physical interactions diminishes (AUC = 0.764; $P = 0.002$ and for metabolic interaction-related features and AUC = 0.749; $P = 0.001$ for protein physical interaction-related features) the prediction performance of all grouped network topological features (AUC = 0.773).

Among all individual network topological features, the number of protein physical interactions is that one that best predicts essential genes (AUC = 0.747). As further discussed in "Cellular rules for essentiality", other investigators have shown that the number of physical interactions is indicative of essentiality [9,19,20]. To our knowledge, we are the first to compare the number of protein physical interactions with other network topological features. Despite the good performance of number of protein physical interactions on predicting essential genes among other individual network topological features, the best predictors are those integrating other groups of topological features with the number of protein physical interactions. This indicates that essentiality depends more or less on each network topological feature and, therefore, the gene location in the network seems to be important for determining its essentiality.

Prediction of essential genes by cellular localization and biological process data

Although the prediction performance of the integrated network topological features in a single predictor can be considered acceptable for predicting essential genes, we decided to check if the addition of information on cellular localization and biological process as training data would increase the predictability of essential genes. Before integrating cellular localization and biological process data with network topological data, we assessed the individual performance of each cellular component and each biological process, as well as the collective performance of all cellular components and all biological processes on predicting essential genes, in order to verify if any individual feature or grouped features related to cellular localization or biological process are good predictors of essential genes.

Regarding cellular localization, we trained our classifier on balanced datasets with all cellular compartments as training data (cytoplasm, endoplasmic reticulum, mitochondrion, nucleus or other localization) and on datasets containing only one of the cellular compartments as training data. We can observe in the ROC plot shown in Figure 3 that the best predictor of essential genes seems to be the integrated set of cellular compartments. This is confirmed by the statistical comparison of the AUC value of the integrated set of cellular compartments with those of individual cellular compartments: the AUC value of grouped cellular compartments (AUC = 0.703) is significantly ($P < 10^{-5}$) higher than AUC values of any individual cellular compartment (Figure 3 and Additional file 2), although such AUC value characterizes the set of all cellular components as fair predictors of essential gene prediction. With regard to biological processes, we trained our classifier on balanced datasets with all biological processes as training data (cell cycle, metabolic process, signal transduction, transcription, transport or other process) and on datasets containing only one of the biological processes as training data. The ROC curves for biological processes (Figure 4) show the same behavior observed for the prediction of essential genes by both network topological features and cellular compartment: the integration of attributes in a single predictor increases the predictability of essential genes in comparison with predictability by individual attributes. The AUC value of the integrated set of biological processes (AUC = 0.667) is statistically significantly ($P < 0.001$) higher than AUC values of any individual biological process (Figure 4 and Additional file 2). With the AUC value of 0.667, however, the set of biological processes can be considered a poor predictor of essential genes.

The moderate and poor performances of cellular localization and biological processes as predictors of essential genes, respectively, suggest that essentiality, as further discussed in "Cellular rules for gene essentiality", is probably a result of multiple factors, reinforcing what we found by analyzing the prediction performance of network topological features. Therefore, we decided to evaluate the prediction performance of the integration of cellular localization and biological process information in a single predictor. We then trained our classifier on balanced datasets with all cellular compartments and biological proc-

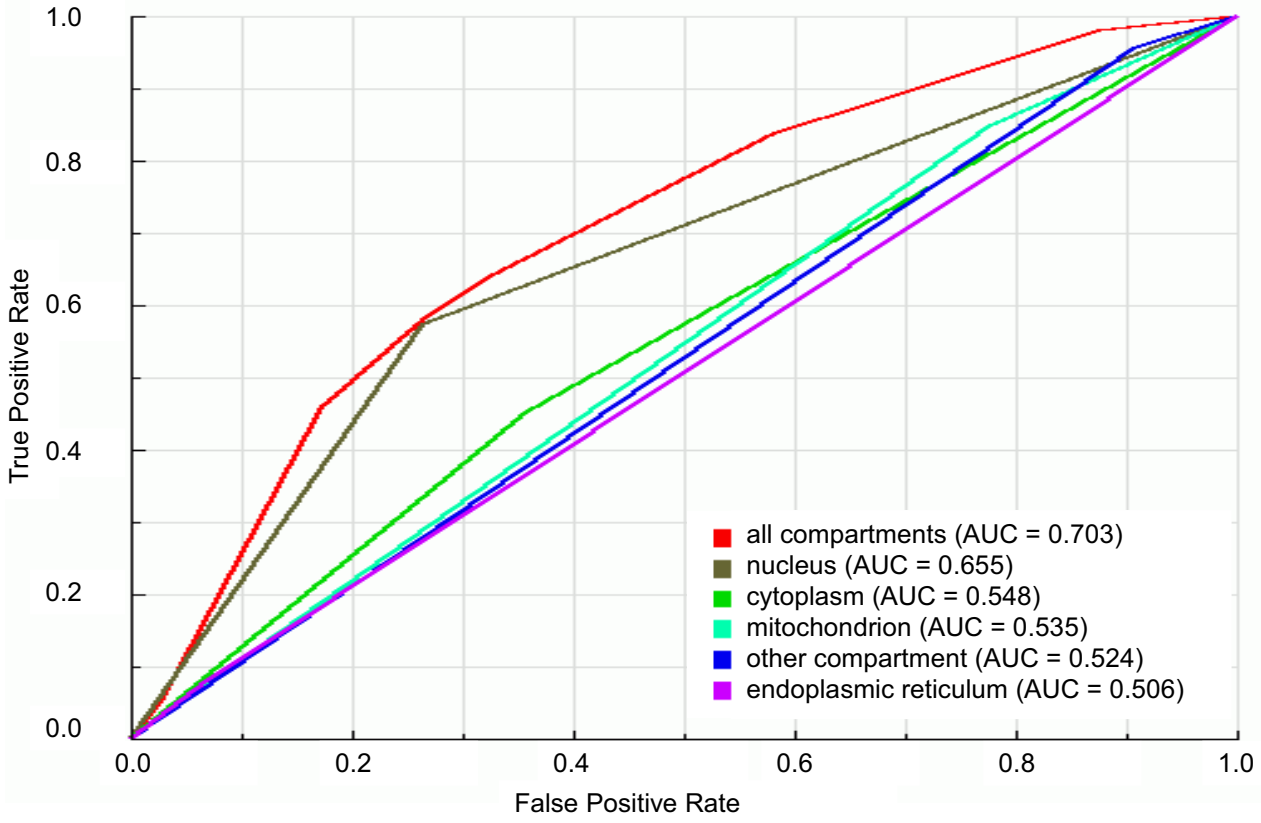

Figure 3
ROC curves and AUC values for the classifiers trained on balanced datasets with individual or grouped cellular compartments. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with one or all cellular compartments as learning attributes. "all compartments" is the dataset with all cellular compartments as learning attributes; "nucleus", "cytoplasm", "mitochondrion", "other compartment" and "endoplasmic reticulum" are datasets with only the respective cellular compartment as learning attribute.

esses as training data. Figure 5 indicates that the performance of integration of cellular localization and biological process data on predicting essential genes is better than other predictors. In fact, the AUC value of predictor containing all cellular localization and biological processes data (AUC = 0.753) is statistically higher ($P < 10^{-5}$) than AUC values of other predictors (see Additional file 2).

Prediction of essential genes by integrating network topological features, cellular localization and biological process information

After determining the predictive power of individual and grouped cellular localization and biological process data, we sought to verify if integration of network topological features with cellular localization and biological process data in a single predictor would improve predictability of essential genes. Moreover, we also sought to compare the predictability of essential genes by all network topological features integrated in a single predictor with that by all cellular compartments and all biological processes integrated in a single predictor. It is worth to mention that although we choose the predictor containing all network topological features to perform the following comparisons, the sets containing protein physical interactions-related features with metabolic interactions-related features or other features (see "Prediction of essential genes by network topological features" for details) also could be used since their prediction performances are comparable to that of all grouped network topological features.

For evaluating the integration of all data in a single predictor and comparing it with the predictor containing only cellular localization and biological process information and with the predictor containing only network topological features, we trained our classifier on balanced datasets with all available data as training data, all cellular compartments and biological processes as training data and all

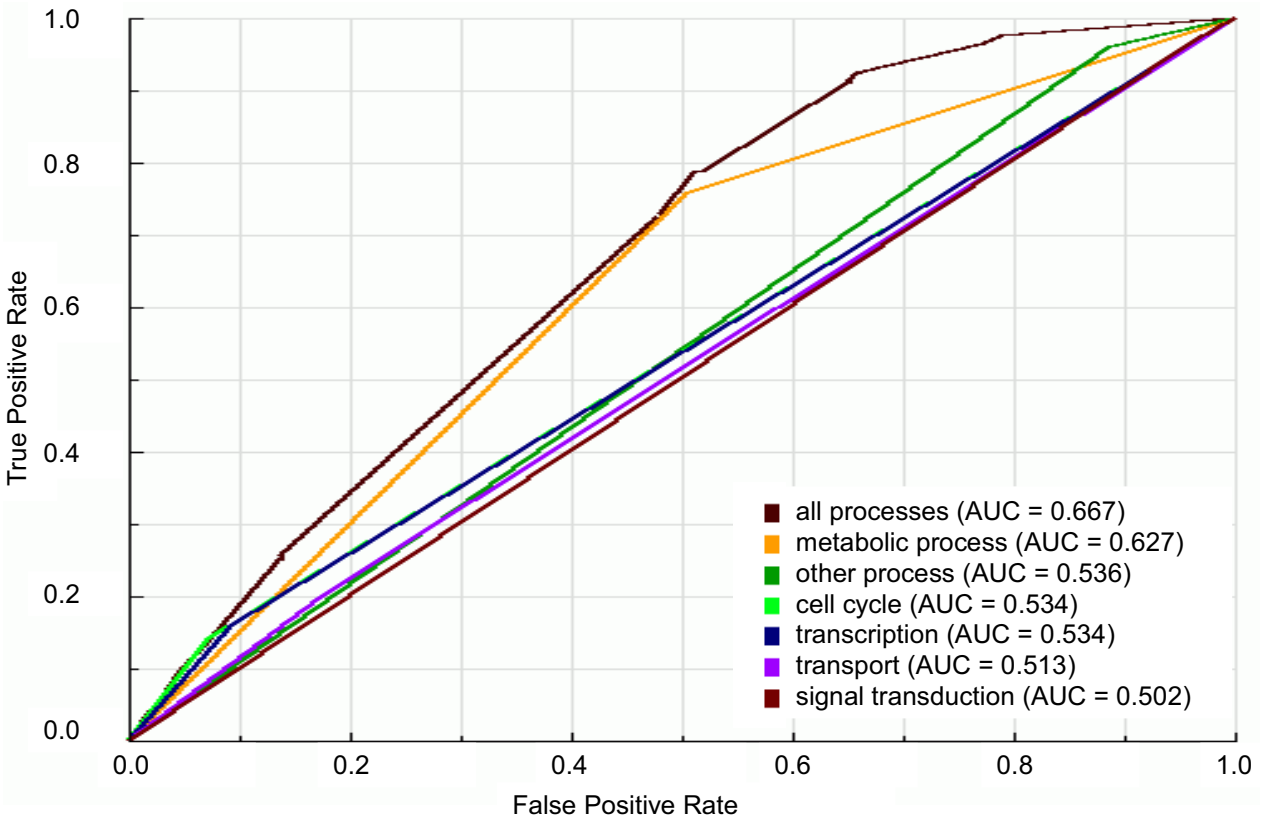

Figure 4
ROC curves and AUC values for the classifiers trained on balanced datasets with individual or grouped biological processes. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with one or all biological processes as learning attributes. "all processes" is the dataset with all biological processes as learning attributes; "metabolic process", "other process", "cell cycle", "transcription" and "transport" are datasets with only the respective biological process as learning attribute.

network topological features, cellular components and biological processes as training data. As expected, the ROC curves in Figure 6 indicate that integration of all network topological features with cellular compartments and biological processes information in a single predictor increases the predictability of essential genes in comparison with predictors containing only network topological features or cellular compartments and biological processes information. Indeed, comparing the AUC value of predictor containing all network topological features and all cellular compartments and biological processes information with that of predictor containing only network topological features or cellular compartments and biological processes information, we confirmed that predictability of essential genes by the integrated predictor (AUC = 0.808) is statistically significantly ($P < 10^{-4}$) higher than that by others predictors (Figure 6 and Additional file 2).

Regarding the comparison of the predictive power of integrated topological network features with that of integrated cellular localization and biological process data, we observed that the difference between the AUC value of predictor containing all cellular compartments and biological processes information (AUC = 0.753) and the AUC value of predictor containing all network topological features (AUC = 0.773) is not statistically significant (P = 0.269) (see Additional file 2). Considering that the function of a protein is intimately linked to its cellular localization [21] and that both the biological process in which a protein is involved and the cellular localization in which a protein acts are predictable by network topological features [10,22], it is not surprising that the predictabilities of essential genes by both the predictor containing all network topological features and the predictor containing all cellular localization and biological process data are similar.

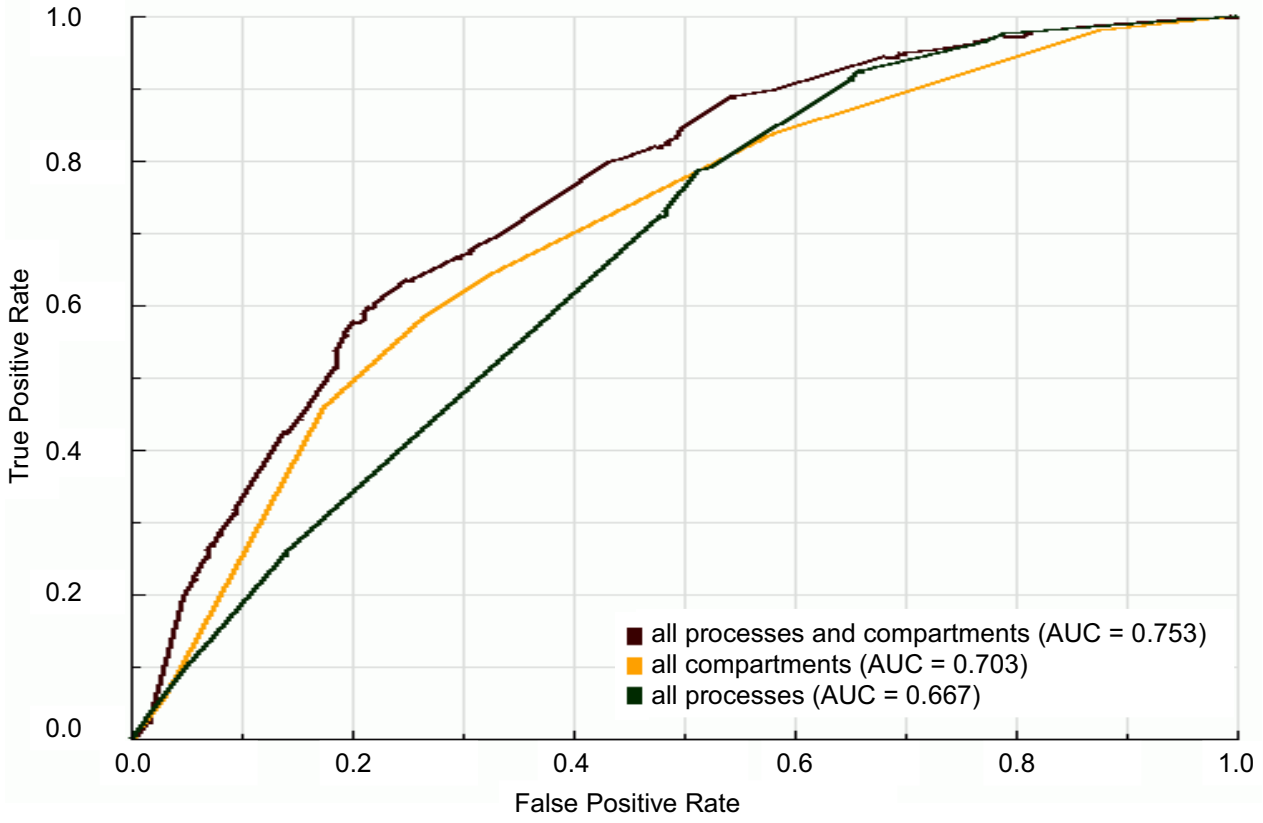

Figure 5
ROC curves and AUC values for the integrated predictors with cellular localization and biological process information. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with all biological processes ("all processes"), all cellular compartments ("all compartments") or all biological processes and cellular compartments ("all processes and compartments") as learning attributes.

Classification of yeast genes not known to be essential

We obtained the list of genes classified as essential and non-essential used for training our classifier from Giaever *et al.* [4] (see "Methods"). Giaever *et al.* have systematically constructed a nearly complete collection of yeast gene-deletion mutants covering about 96% of all genes. However, about 430 genes of this collection were removed from the yeast genome after a comprehensive reannotation process of the *S. cerevisiae* genome performed in 2006 [23]. In addition, new genes were annotated to yeast genome as a result of this reannotation process. In order to classify these genes not analyzed by Giaever *et al.*, we used our best classifier, that is, the one that containing all network topological features, cellular components and biological processes information as training attributes. For each gene, the predictor output the probability of classifying it as essential and non-essential, which we called, respectively, "essentiality score" and "non-essentiality score".

To predict a gene as essential, we defined an essentiality score of 0.654 as the cutoff value, i.e., genes with essentiality score above 0.654 were considered to be essential. This cutoff value was based on the optimal threshold, which is the score value that leads to the maximal accuracy of classification, calculated by the software StAR [24] for the predictor containing all features (network topological, cellular component and biological process; see Figure 6 and Additional file 2). Among the 514 genes with the essentiality status not defined by Giaever *et al.*, 44 genes were predicted as essential (Table 1). Analyzing these genes, we found that 9 genes have been previously demonstrated to be essential (*YHR165C*, *YHR089C*, *YHR052W*, *YCR042C*, *YDR320C-A*, *YHR169W*, *YKL138C-A*, *YGL106W* and *YHR099W*) and other 14 genes (*YGR252W*, *YHR027C*, *YOL012C*, *YNL147W*, *YGL100W*, *YNL096C*, *YOL148C*, *YFL007W*, *YOL145C*, *YBR111W-A*, *YNL055C*, *YHR216W*, *YBL071W-A* and *YHR039C-A*) have been previously demonstrated to be non-essential by

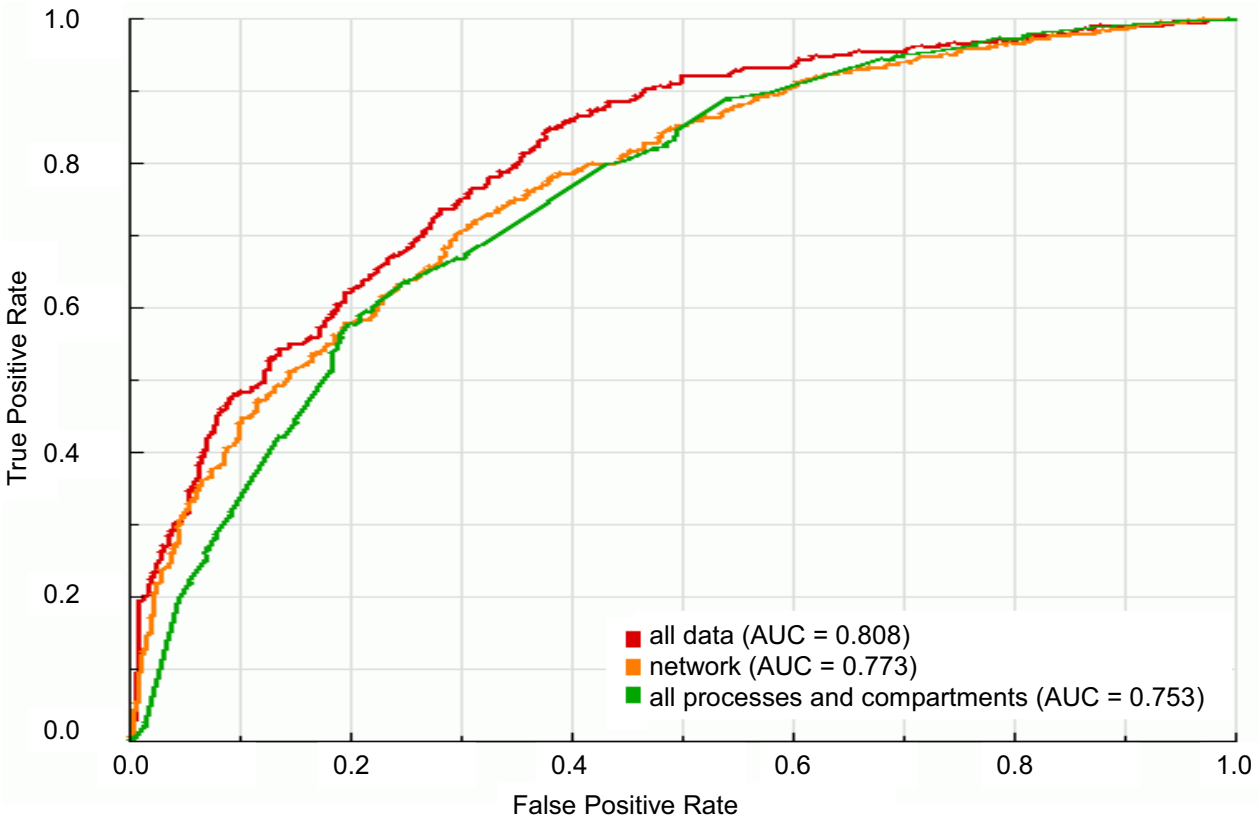

Figure 6
ROC curves and AUC values for the integrated predictors with available data. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with all network topological features, cellular compartments and biological processes ("all data"), all biological processes and cellular compartments ("all processes and compartments") or all network topological features ("network") as learning attributes.

other investigators through small-scale gene deletion experiments in functional characterization studies [25-36] (Table 1). Among non-essential genes, 10 genes (*YGR252W*, *YHR027C*, *YOL012C*, *YNL147W*, *YNL096C*, *YOL148C*, *YOL145C*, *YBR111W-A*, *YNL055C* and *YHR039C-A*) have been shown to impair substantially the growth of *S. cerevisiae* when they are completely deleted [33,36-40], whereas the 4 remaining non-essential genes (*YGL100W*, *YFL007W*, *YHR216W* and *YBL071W-A*) have been shown not to affect the growth phenotype of yeast when they are deleted [34,35,41,42]. Although roughly 1/3 of the these genes predicted to be essential have been previously classified as non-essential, the complete deletions of most of them have been shown to severely reduce the fitness of organisms [33,36-40], suggesting that our predictor, even when directly contradicted by these experimental findings, can nonetheless identify genes important to cellular function. Regarding the 4 non-essential genes whose deletion has been shown not to affect the growth phenotype of yeast (*YFL007W* and *YGL100W*), we hypothesize that our classifier assigned a high essentiality score to these genes due to the following features: (*i*) their encoded proteins interact with more than 12 other proteins, (*ii*) they are regulated by less than 4 transcription factors and (*iii*) their encoded proteins are located in the nucleus. These characteristics are in accordance with two cellular rules for essentiality discovered by our approach as demonstrated in the section "Cellular rules for gene essentiality": if proteins interact with more than 7 other proteins and are located in the nucleus, genes encoding them are likely to be essential and genes regulated by more than 3 transcription factors tend to be non-essential.

Among the 44 genes predicted to be essential, 21 genes have not yet been investigated for essentiality to date (Table 1). One of these genes is the *YER029C* whose encoded protein (Yer029cp) binds to other 6 proteins to form the heteroheptameric complex that is required for the biogenesis of the spliceosomal U1, U2, U4, and U5 snRNPs [43]. These spliceosomal snRNPs are involved in

Table 1: List of the 44 yeast genes predicted to be essential in S. cerevisiae

Rank	Gene	Essentiality Score	Essentiality Status	Deletion phenotype	Reference
1	YHR165C	0.940	essential	lethality	[32]
2	YGR252W	0.939	non-essential	defective growth	[33]
3	YHR089C	0.937	essential	lethality	[25]
4	YHR052W	0.065	essential	lethality	[26]
5	YER029C	0.930	not defined	not defined	-
6	YHR027C	0.930	non-essential	defective growth	[37]
7	YHR099W	0.929	essential	lethality	[27]
8	YOL012C	0.925	non-essential	defective growth	[33]
9	YHR169W	0.921	essential	lethality	[28]
10	YCR042C	0.920	essential	lethality	[29]
11	YDR320C-A	0.897	essential	lethality	[30]
12	YNL147W	0.885	non-essential	defective growth	[33]
13	YGL100W	0.866	non-essential	not related to growth	[41]
14	YNL096C	0.865	non-essential	defective growth	[33]
15	YOL148C	0.859	non-essential	defective growth	[38]
16	YOR145C	0.856	essential	lethality	[31]
17	YFL007W	0.839	non-essential	not related to growth	[42]
18	YKL138C-A	0.837	essential	lethality	[30]
19	YOL145C	0.824	non-essential	defective growth	[39]
20	YBR111W-A	0.822	non-essential	defective growth	[40]
21	YLL022C	0.816	not defined	not defined	-
22	YNL209W	0.816	not defined	not defined	-
23	YGL106W	0.813	not defined	not defined	-
24	YPR080W	0.813	not defined	not defined	-
25	YER105C	0.794	not defined	not defined	
26	YNL055C	0.783	non-essential	defective growth	[33]
27	YOL142W	0.781	not defined	not defined	-
28	YAL024C	0.770	not defined	not defined	
29	YHR216W	0.768	non-essential	defective growth	[34]
30	YHL004W	0.743	not defined	not defined	-
31	YHR072W-A	0.741	not defined	not defined	-
32	YGL190C	0.738	not defined	not defined	-
33	YDR079C-A	0.731	not defined	not defined	-
34	YNL186W	0.731	not defined	not defined	-
35	YJR132W	0.716	not defined	not defined	-
36	YDR261W-A	0.713	non-essential	defective growth	[33]
37	YHR119W	0.696	not defined	not defined	-
38	YBL071W-A	0.693	non-essential	defective growth	[35]
39	YDR261W-B	0.682	non-essential	defective growth	[34]
40	YHR039C-A	0.680	non-essential	defective growth	[36]
41	YHR090C	0.680	not defined	not defined	-
42	YER026C	0.675	not defined	not defined	-
43	YHR056C	0.665	not defined	not defined	-
44	YCL019W	0.659	not defined	not defined	-

splicing of nuclear pre-mRNAs [44], an essential biological process for cell viability, and, interestingly, all proteins forming the heteroheptameric complex along with Yer029cp have been demonstrated to be essential [4]. Therefore, the presence of this gene among ones predicted to be essential reinforces the fact that our predictor is able to identify genes that are important to cellular function.

Finally, regarding the remaining 470 genes predicted as non-essential, we verified that 129 of these genes have been previously tested for essentiality by other studies (see Additional file 4). Among them, 124 have been demonstrated to be non-essential genes and only 5 have been demonstrated to be essential genes. Thus, about 4% of genes with known essentiality status and predicted as non-essential are actually essential genes (Additional file 4). Providing that 38% (9 of 14; see Table 2) of the genes with known essentiality status and predicted as essential are actually essential genes, the predictor integrating all available features (network topological, cellular component and biological process; see Figure 6 and Additional file 2) leads to an enrichment of actual essential genes in the set of genes predicted as essential. This suggests that

this predictor is committed to minimize the false negative rate thus avoiding the loss of essential genes.

Cellular rules for gene essentiality

Beyond the prediction capability, machine learning techniques can be used for knowledge acquisition in order to describe patterns in datasets. The machine learning algorithms most used for knowledge acquisition are those that generate decision trees. Decision trees are decision support tools inferred from the training data that use a graph of conditions and their possible consequences. The structure of a decision tree consists of a root node representing the most important condition for discriminating classes, internal nodes representing additional conditions for class discrimination under the main condition, and leaf nodes representing the final classification. So, one can learn the conditions for classifying instances in a given class by following the path from the root node to the leaf node [45].

Therefore, in order to discover the rules for gene essentiality in *S. cerevisiae*, we analyzed decision trees generated by training the J48 algorithm, a WEKA's implementation of the C4.5 algorithm [46] (for more details, see "Methods"), on the ten balanced datasets containing all network topological features, cellular components and biological processes as training data (the construction of balanced datasets are detailed in "Methods"). As decision trees generated from the balanced datasets could be slightly different from one another due to the undersampling scheme used to balance the original set of classified genes--each balanced dataset contains a different set of 1,024 non-essential genes, 1/8 of the total amount in the original imbalanced dataset--we generated one detailed (64 instances per leaf) and one simplified (128 instances per leaf) decision tree for each balanced dataset (see "Methods" for details) and then we manually inspected them in order to discover the general rules for gene essentiality.

From the 20 slightly different generated decision trees, we were able to devise the general rules for gene essentiality in *S. cerevisiae*. Figure 7 shows the decision tree that best illustrates the general rules for gene essentiality (all decision trees are available in text format in the Additional file 5). As we can observe in Figure 7, the root node of decision tree is the number of protein physical interactions (all generated decision trees exhibit this feature; see Additional file 5); so, this attribute can be considered the most important feature among all network topological features and cellular localization and biological process information for gene essentiality. Accordingly, the predictor containing only the number of protein physical interaction as training feature is the one that best predicts (AUC = 0.747) essential genes among all other individual features as we can observe in Figure 2. This is in concert with previous studies that have demonstrated that the number of protein physical interactions is indicative of essentiality [9,19,20]. Several hypotheses about the connection between gene essentiality and number of protein physical interactions have been proposed. Coulomb *et al*. [47] have suggested that the relationship between this network feature and gene essentiality is partly due to biases in the interaction data that are enriched in small-scale experiments which are partial towards essential genes. On the other hand, Zotenko *et al*. [48] have recently hypothesized that the connection between gene essentiality and number of protein physical interactions is likely due to the involvement of proteins encoded by essential genes in subnetworks of densely connected proteins with shared biological functions that are enriched in proteins encoded by essential genes.

Following the path from root node to first leaf node through the right branch (Figure 7), we found the following rule for gene essentiality: if proteins interact with more than 7 other proteins (average of number of interactions ranging from 6 to 12 in all decision trees) and are located in the nucleus, genes encoding them are likely to be essential. This rule can be observed in 9 of 10 decision trees with 128 instances per leaf and 8 of 10 decision trees with 64 instances per leaf (see Additional file 5). If these proteins are located in cellular compartments other than the nucleus, essentiality of their corresponding genes depends on conditions particular to each decision tree (Figure 7 and Additional file 5). The path from root node to the leaf nodes through the left branch (Figure 7) drove us to discover another rule for gene essentiality: if proteins interact with 6 or fewer proteins and participate in a metabolic process inside the nucleus, genes encoding these proteins are likely to be essential. This rule can be observed in 7 of 10 decision trees with both 128 and 64 instances per leaf (Additional file 5).

According to these rules, the ultimate condition for gene essentiality is the localization of proteins in the nucleus, suggesting that this cellular component is somehow important for essentiality. The importance of nucleus for essentiality has also been suggested by Seringhaus *et al*. [7] that have shown that nuclear localization has the strongest positive correlation with essentiality among other cellular components. The relationship between nucleus and essentiality can be explained by the fact that roughly one third of nuclear proteins are encoded by essential genes and most of essential biological processes for cell viability take place within the nucleus [49]. Therefore, the participation of proteins in these nuclear-localized essential processes, such as DNA replication, transcription and DNA repair, should be a pivotal condition for essentiality in the rules defined by both the paths via the left and right branches of decision tree. It is worth to mention that, as a

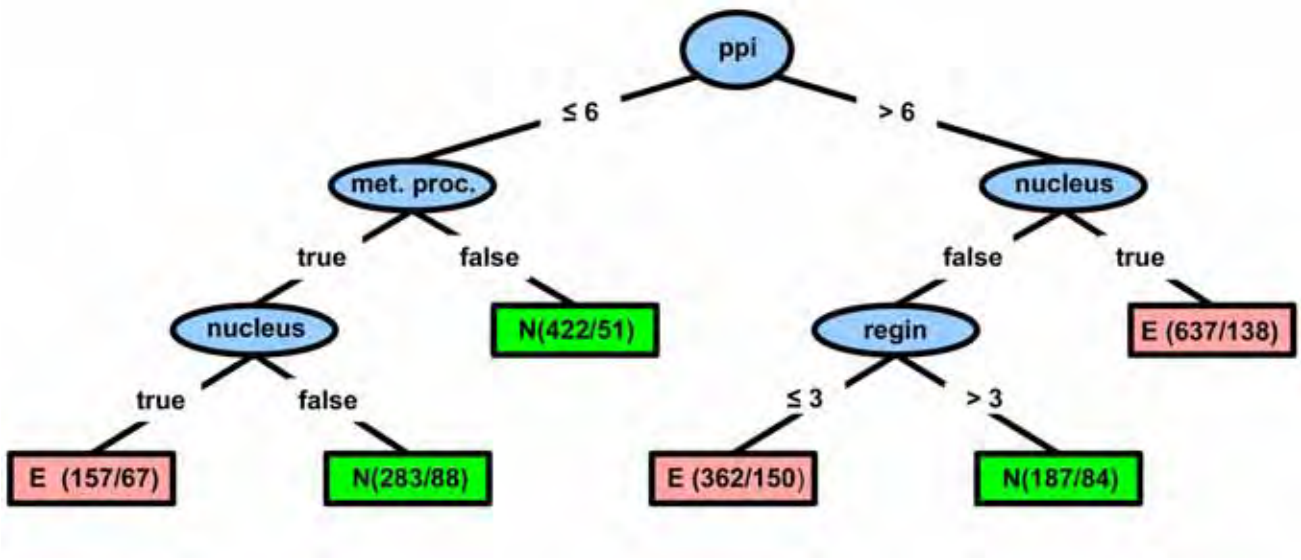

Figure 7

Decision tree generated by training the J48 algorithm on the balanced dataset 8 with all available data. This decision tree was generated by training the J48 algorithm on the balanced dataset 8 with all available data (see "Methods"). The uppermost ellipse is the node root of tree that represents the most important condition for discriminating essential genes from non-essential genes. In this case, such condition is the number of protein physical interactions (*ppi*). The remaining ellipses are internal nodes that represent additional conditions for considering a gene as essential or non-essential. In the left branch of tree, such conditions are involvement in a metabolic process (*met. proc.*) and nuclear localization (*nucleus*). In the right branch, such conditions are nuclear localization (*nucleus*) and number of regulating transcription factors (*regin*). The rectangles are the leaf nodes that represent the final classification. Red and green rectangles depict genes that, under certain conditions (represented by the root node and internal nodes), are respectively and predominantly classified as essential (**E**) and non-essential (**N**). In the round brackets inside rectangles, the number before the slash indicates the total number of genes that are actually essential or non-essential and the number after the slash indicates how many genes were incorrectly predicted.

result of the annotation method we used (see more details in "Methods"), these nuclear-localized essential processes are embedded in the biological process "metabolic process", one of the conditions for essentiality along with nuclear localization and number of protein physical interactions equal or less than 6 in the rule defined by the path via the left branch of decision tree (Figure 7). In the rule defined by the path via the right branch, although essentiality is apparently not dependent on the involvement of proteins in metabolic processes inside the nucleus, the nuclear proteins encoded by genes classified as essential according to this rule may be actually involved in a nuclear metabolic process. In this case, however, the involvement in nuclear metabolic processes is overwhelmed by the number of protein physical interactions.

We discovered an additional interesting rule for gene essentiality in yeast: genes regulated by more than 3 transcription factors tend to be non-essential (Figure 7). This rule can be observed in 6 of 10 decision trees with 128 instances per leaf and in all decision trees when the number of instances per leaf is set to 64 (see "Methods" for details and Additional file 5). Our finding is corroborated by Yu *et al*. [50] that have found that genes regulated by > 10 transcription factors are less likely to be essential than those regulated by 2-9 transcription factors, whereas these genes are less likely to be essential than those with only one transcription factor. At first glimpse, the fact that essential genes tend to be regulated by a few transcription factors seems contradictory since one would expect that gene essentiality is correlated with a high level of transcriptional regulation. However, most essential genes encode housekeeping proteins, i.e., proteins involved in housekeeping functions, such as rRNA metabolic process and transcription initiation [48]. As housekeeping functions are the most basic and important functions within cell, genes encoding housekeeping proteins are ubiquitously expressed and, consequently, they tend to be regulated by fewer transcription factors than genes encoding non-housekeeping proteins. Therefore, this phenomenon is likely due to the enrichment of genes encoding housekeeping proteins in the set of essential genes.

Conclusion

The identification of essential genes has largely been an experimental effort mostly performed by time-consuming whole-genome knockout experiments. In an effort to accelerate the pace of discovery of essential genes, we designed a machine learning-based computational approach that relies on network topological features, cellular localization and biological process information for predicting essential genes and evaluated it in the yeast *Saccharomyces cerevisiae*.

We therefore constructed an integrated network of gene interactions for *S. cerevisiae* containing protein physical, metabolic and transcriptional regulation interactions and computed 12 different network topological features (as described in Additional file 1 and "Methods") that were individually and collectively evaluated for their ability to predict essential genes. We showed that the predictors containing all 12 network topological features or different combinations of protein physical interactions-related features with other groups of topological features as training data are reliable predictors (AUC = 0.763-0.773) of essential genes in *S. cerevisiae*, thus reinforcing the fact that an integrated network of gene interactions can be an useful tool for the prediction of essential genes.

Although the performance of predictors containing only network topological features can be considered acceptable for predicting essential genes, we decided to check if the addition of cellular localization and biological process information to these predictors would increase the predictability of essential genes. In fact, we verified that the performance of the predictor containing all network topological features, cellular localization and biological process information as training data is better than those of the predictors containing only network topological features or only cellular localization and biological process information. Interestingly, we also showed that the prediction performances of the predictor containing only network topological predictions and the predictor containing only cellular localization and biological process information are similar. To our knowledge, this is the first time that Gene Ontology terms related to cellular localization and biological process are shown to be useful predictors of essential genes.

In addition to prediction of essential genes, we could also devise some cellular rules for gene essentiality using all network topological features, cellular localization and biological process information as training data for generation of decision trees (see details in section "Cellular rules for gene essentiality"). We discovered that the number of protein physical interactions, the nuclear localization and the number of regulating transcription factors are important factors determining gene essentiality. Although these findings have previously been demonstrated by other investigators [7,9,19,20,50], it is interesting to notice that we were able to obtain these same results by simply inspecting the decision tree generated as shown in section "Cellular rules for gene essentiality". So, decision trees are useful tools for extracting knowledge from complex biological data.

Besides confirming previous findings, the exploration of decision trees can also lead to new discoveries. This can be exemplified by an additional analysis that we performed due to a referee' s suggestion regarding the nuclear localization of essential proteins. The referee has suggested us to analyze the influence of some children terms of GO term "nucleus" on the nuclear localization-related gene essentiality. For this purpose, we generated a decision tree by training the J48 algorithm on one of the ten balanced datasets (see "Methods" for details) with all features plus the GO terms "nucleolus", "nucleoplasm", "nuclear chromosome" and "nuclear envelope" and, as can be observed in the Additional file 5, an entirely new rule can be devised from the generated decision tree: the nucleolar localization of proteins is the most important factor for gene essentiality. We did not mention this potential and interesting rule for gene essentiality in the section "Cellular rules for gene essentiality" since this rule *per se* is interesting enough to deserve a more exhaustive analysis that can be reported in a future paper.

Albeit the good prediction performance and the ability to discover cellular rules for essentiality, our approach suffers from two limitations. First, it depends on existing Gene Ontology annotation and protein physical interaction data which are likely to be enriched in small-scale experiments involving essential genes. Second, the construction of an integrated network of gene interactions requires a large amount of experimental interaction data that are currently available only to a limited number of organisms.

Therefore, the prediction of essential genes in newly sequenced organisms, for example, is impractical by our approach. However, the integration of our approach with (*i*) computational-based methods for gene annotation and (*ii*) computational-based methods for the construction of integrated networks of predicted gene interactions in which each type of interaction (protein physical, metabolic and transcriptional regulation interactions) can be distinguished from one another could give rise to a purely *in silico* network topology, cellular localization and biological process information-based methodology for prediction of essential genes. Such a methodology would be totally independent on experimental interaction data and, accordingly, unbiased in essential genes-driven experiments.

In summary, despite the limitations discussed above, we could demonstrate that the integration of network topological features, cellular localization and biological process information is capable to predict essential genes. In this work, we tested the predictive performance of this integration in *S. cerevisiae*, but we envisage that it might be useful to predict essential genes in any other organism if a purely computational-based prediction approach, as suggested above, is used.

Methods

Generation of the set of training features

Network topological features

In order to compute the network topological features used as training features for predicting essential genes, we first constructed an integrated network of gene interactions of *S. cerevisiae* based on assumption that two genes, g_1 and g_2, coding respectively for proteins p_1 and p_2, are interacting genes if *(i)* p_1 and p_2 interact physically (protein physical interaction), (*ii*) the transcription factor p_1 directly regulates the transcription of gene g_2, i.e., p_1 binds to the promoter region of g_2 (transcriptional regulation interaction), or (*iii*) the enzymes p_1 and p_2 share metabolites, i.e., a product generated by a reaction catalyzed by enzyme p_1 is used as reactant by a reaction catalyzed by enzyme p_2 (metabolic interaction).

Yeast protein physical interactions data were obtained from The Biological General Repository for Interaction Datasets (BioGRID) database, a repository of literature-curated protein physical and genetic interactions [51]. We downloaded the database release 2.0.42 of July 2008 and removed the entries related to genetic interactions. Yeast transcriptional regulation interactions were obtained from the Yeast Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT) database, a curated repository of regulatory associations between transcription factors and target genes in *Saccharomyces cerevisiae* [52]. By using the utility "Generate Matrix Regulation" in the YEASTRACT website, we generated and downloaded a regulation matrix containing only documented transcriptional regulation interactions determined by direct experimental evidence.

Yeast metabolic interactions were extracted from the metabolic model iND750 of *Saccharomyces cerevisiae* [11] by a code implemented in Mathematica® 6.0 (Wolfram Research, Inc.). We excluded those metabolic interactions generated by the so-called "currency metabolites", abundant molecular species present throughout the cell most of the time and, therefore, unlikely to impose any constraints on the dynamics of metabolic reactions. Due to this feature of currency metabolites, the functionality of the network would be better represented without them [53]. We considered currency metabolites the eight most connected metabolites (ADP, ATP, H^+, H_2O, $NADP^+$, NADPH, orthophosphate and pyrophosphate) in the original metabolic model iND750.

The final integrated network of gene interactions (INGI) of yeast is the result of integration of the protein physical, metabolic and transcriptional regulation interactions datasets through genes common to these datasets. Before performing the integration, we converted all yeast gene names to their systematic names--as provided by the Saccharomyces Genome Database (SGD) Nomenclature Conventions [23]--to avoid the creation of false interactions due to gene name ambiguity. Genes classified as dubious, i.e., genes unlikely to encode an expressed protein and not considered biologically significant by SGD, were removed from the final INGI.

For each gene g in the yeast INGI, we computed twelve network topological features as listed in Additional file 1. Briefly, degree centrality is defined as the number of links to node (in our case, gene). We considered each type of interaction as a distinct measure of degree as described in Additional file 1. Clustering coefficient (c) of a node (in our case, a gene) quantifies how close the node and its neighbors are to being a clique, i.e., all nodes connected to all nodes. For yeast INGI, c is defined as the proportion of links between the genes within the neighborhood of g divided by the number of links that could possibly exist between them. Betweenness centrality reflects the role played by a node (in our case, a gene) in the global network architecture and, for the yeast INGI, is defined as the fraction of shortest paths between g_i and g_j passing through g. We computed the betweenness centrality based on shortest paths via all types of interaction (*inbet*) as well as based on shortest paths via each type of interaction (*inbetppi*, *inbetmet* and *inbetreg*). Closeness centrality (*cent*) measures how close a node (in our case, a gene) is to all others in the network and, for the yeast INGI, is defined as the mean shortest path between g and all other genes reachable from it. Identicalness is the number of genes with identical network topological characteristics. All these network topological features, except for the betweenness centrality-related features, were calculated by a program written in a Mathematica® 6.0 notebook. The betweenness centrality-related features were calculated by the Python package *NetworkX* [54].

Cellular localization and biological process annotation of yeast genes

We determined the cellular component in which a yeast gene product acts and the biological process in which a yeast gene is involved by using the Saccharomyces Genome Database (SGD) Gene Ontology (GO) Slim Mapper [55]. The SGD GO-Slim Mapper maps annotations of a group of genes to more general GO terms. Among GO Slim sets available at SGD, we selected cellular

component and biological process terms from the Super GO-Slim set, a collection of high-level GO terms. For cellular localization annotation, genes annotated to terms rather than "cytoplasm", "endoplasmic reticulum", "mitochondrion" and "nucleus" were reannotated to one of these terms or to a new term named "other localization". For biological process annotation, genes annotated to terms rather than "cell cycle", "metabolic process", "signal transduction", "transcription" and "transport" were reannotated to one of these terms or to a new term named "other process".

Classifier design, training and evaluation

Construction of datasets for classifier training and evaluation

We defined "essential genes" as those genes whose deletion leads to an inviable yeast organism cultured on rich glucose medium. We obtained the dataset containing the classification of yeast genes in essential or non-essential from Giaever *et al*. [4]. After downloading the dataset, we removed from it genes classified as dubious in SGD and converted the name of remaining genes to their systematic names as provided by the SGD Nomenclature Conventions [23].

As this dataset of classified genes is an imbalanced dataset, i.e., the number of non-essential genes is much larger than the number of essential genes, and it has been known that data imbalance degrades the performance of machine learning algorithms [17], we built balanced datasets from the original imbalanced dataset by using an undersampling scheme as follows: (1) first, we split the dataset of classified genes into two subsets: "essential genes set", containing 1,024 essential gene entries, and "non-essential genes set", containing 4,097 non-essential gene entries; (2) second, we selected all entries from the essential genes set (1,024 entries) and randomly selected 1,024 entries from the non-essential genes set; (3) we then created the balanced dataset containing the 2,048 selected entries with random distribution of the essential gene and non-essential gene entries. This procedure was repeated 10 times in order to generate 10 different balanced datasets containing different sets of non-essential gene entries.

To compare the predictability of essential genes by individual training features with that of different groups of training features, we generated, from the balanced datasets, different subsets containing different combinations of training features as detailed in Additional file 2.

Classifier design

We used WEKA (Waikato Environment for Knowledge Analysis) software package, a collection of machine learning algorithms for data mining tasks [56], for designing, training and evaluating the classifiers applied to prediction of essential genes. Among classifiers that we evaluated, the one that provided the best performance was an ensemble of eight decision tree algorithms using the meta-classifier "Vote", a WEKA's implementation of the voting algorithm that combines the output predictions of each classifier by different rules [57]. We combined the classifiers by the average rule, where the output predictions computed by the individual classifiers for each class are averaged and this average is used in its decision [57]. The classifiers composing our model were: (1) REPtree [56], (2) naive bayes tree [58], (3) random tree [56], (4) random forest [59], (5) J48, a WEKA's implementation of the C4.5 decision tree [46], with minimum number of 32 instances per leaf, (6) best-first decision tree with minimum number of 32 instances at the terminal nodes [60], (7) logistic model tree [61] and (8) alternating decision tree with 25 boost iterations [62]. In addition, we applied the bootstrap aggregating (bagging) approach [63] to each classifier. Parameters values for each classifier are provided in the Additional file 6.

Classifier training and evaluation

For each of the 10 balanced datasets, we trained our classifier on half of entries and the other half was used to evaluate the classifier performance, totaling 10 runs of training and evaluation. For these runs, we generated a receiver operating characteristic (ROC) curve and calculated the area under the ROC curve (AUC). The ROC curve is a plot of the true positive rate versus false positive rate and indicates the probability of a true positive prediction as a function of the probability of a false positive prediction for all possible threshold values [64]. AUC is a widely used summary measure of the ROC curve and is equivalent to the probability that a randomly chosen negative example (in our case, a non-essential gene) will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example (in our case, an essential gene) [65].

We used the web server version of the StAR (**St**atistical **A**nalysis of **R**OC curves) software [24] for calculating the true and false positive rates and the AUC values and for generating the ROC curves. The statistical comparison of AUC values derived from the different datasets was also performed by StAR by means of a nonparametric statistical method based on the Mann-Whitney U-statistic for comparing distributions of values from two samples [18] with a significance level (P) of 0.01.

Determination of rules for gene essentiality

The determination of rules for gene essentiality was performed by analyzing decision trees generated through the training of J48 algorithm on balanced datasets containing all training data. We used two different values of the

parameter "number of objects per leaf" of J48 algorithm for generating two different types of decision trees: 64 for more detailed trees and 128 for more simplified trees [56]. For each balanced dataset, then, we obtained two decision trees (detailed and simplified) and manually inspected all the 20 generated decision trees for determining the general rules for gene essentiality. The remaining parameters values for producing decision trees by J48 algorithm training are provided in the Additional file 6 and all decision trees are provided in text format in the Additional file 5.

Authors' contributions

MLA obtained all interaction data, constructed the network, designed and analyzed the classifier performance, pursued the biological interpretation of results and drafted the manuscript. NL conceived, designed and directed the project and implemented the program for calculation of network topological features. All authors read and approved the final manuscript.

Additional material

Additional file 1

***Network topological features**. This file includes a table showing the functions and descriptions of the twelve network topological features used as learning attributes for training the classifier algorithm*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-290-S1.PDF]

Additional file 2

***Statistical pairwise comparison of predictors**. This file includes tables showing the pairwise comparison of predictors with the p-values of AUC differences between each pair of predictors.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-290-S2.XLS]

Additional file 3

***ROC curves and AUC values demonstrating the effect of removal of individual or small sets of network topological features**. File containing ROC curves for classifiers trained on datasets whose learning attributes were different sets of network topological features in which each set lacks one of the topological features or a small group of 2-4 topological features.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-290-S3.PDF]

Additional file 4

***List of the 470 yeast genes predicted to be non-essential**. Tab-limited text file containing the 470 genes classified as non-essential with their essentiality scores, actual essentiality statuses and, if applicable, the Pubmed references showing their essentiality statuses.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-290-S4.TXT]

Additional file 5

***J48 decision trees**. This file contains all 10 decision trees generated by training the J48 algorithm on the 10 balanced datasets with all available data as learning attributes. Decision trees are represented in text format (raw output generated by WEKA).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-290-S5.PDF]

Additional file 6

***Parameters used to train the meta-classifier and J48**. File containing all parameters values used to train the meta-classifier for essential gene prediction and all parameters values used to train the J48 algorithm to generate decision trees for discovery of cellular rules for essentiality.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-290-S6.PDF]

Acknowledgements

The authors would like to thank the anonymous referee for the helpful suggestions that greatly improved this manuscript. The authors would also like to thank FAPESP (The State of Sao Paulo Research Foundation) and CNPq (National Council of Technological and Scientific Development) for the financial support through the FAPESP research grants 2007/02827-9 and 2007/01213-7 and CNPq research grant 474278/2006-9.

References

1. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Débarbouille M, Dervyn E, Deuerling E, Devine K, Devine SK, Dreesen O, Errington J, Fillinger S, Foster SJ, Fujita Y, Galizzi A, Gardan R, Eschevins C, Fukushima T, Haga K, Harwood CR, Hecker M, Hosoya D, Hullo MF, Kakeshita H, Karamata D, Kasahara Y, Kawamura F, Koga K, Koski P, Kuwana R, Imamura D, Ishimaru M, Ishikawa S, Ishio I, Coq DL, Masson A, Mauël C, Meima R, Mellado RP, Moir A, Moriya S, Nagakawa E, Nanamiya H, Nakai S, Nygaard P, Ogura M, Ohanan T, O'Reilly M, O'Rourke M, Pragai Z, Pooley HM, Rapoport G, Rawlins JP, Rivas LA, Rivolta C, Sadaie A, Sadaie Y, Sarvas M, Sato T, Saxild HH, Scanlan E, Schumann W, Seegers JFML, Sekiguchi J, Sekowska A, Séror SJ, Simon M, Stragier P, Studer R, Takamatsu H, Tanaka T, Takeuchi M, Thomaides HB, Vagner V, van Dijl JM, Watabe K, Wipat A, Yamamoto H, Yamamoto M, Yamamoto Y, Yamane K, Yata K, Yoshida K, Yoshikawa H, Zuber U, Ogasawara N: **Essential Bacillus subtilis genes.** *Proc Natl Acad Sci USA* 2003, **100(8):**4678-83.
2. Itaya M: **An estimation of minimal genome size required for life.** *FEBS Lett* 1995, **362(3):**257-60.
3. Judson N, Mekalanos JJ: **TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes.** *Nat Biotechnol* 2000, **18(7):**740-5.
4. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Güldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kötter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, yun Wang C, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M: **Functional profiling of the Saccharomyces cerevisiae genome.** *Nature* 2002, **418(6896):**387-91.

5. Cullen LM, Arndt GM: **Genome-wide screening for gene function using RNAi in mammalian cells.** *Immunol Cell Biol* 2005, **83(3):**217-23.
6. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C, Martel N, Veronneau S, Lemieux S, Kauffman S, Becker J, Storms R, Boone C, Bussey H: **Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery.** *Mol Microbiol* 2003, **50:**167-81.
7. Seringhaus M, Paccanaro A, Borneman A, Snyder M, Gerstein M: **Predicting essential genes in fungal genomes.** *Genome Res* 2006, **16(9):**1126-35.
8. Gustafson AM, Snitkin ES, Parker SCJ, DeLisi C, Kasif S: **Towards the identification of essential genes using targeted genome sequencing and comparative analysis.** *BMC Genomics* 2006, **7:**265.
9. Jeong H, Mason SP, Barabási AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411(6833):**41-2.
10. Schwikowski B, Uetz P, Fields S: **A network of protein-protein interactions in yeast.** *Nat Biotechnol* 2000, **18(12):**1257-1261.
11. Duarte NC, Herrgard MJ, Palsson BO: **Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model.** *Genome Res* 2004, **14(7):**1298-1309.
12. Guelzim N, Bottani S, Bourgine P, Kepes F: **Topological and causal structure of the yeast transcriptional regulatory network.** *Nat Genet* 2002, **31:**60-63.
13. Palumbo MC, Colosimo A, Giuliani A, Farina L: **Functional essentiality from topology features in metabolic networks: a case study in yeast.** *FEBS Lett* 2005, **579(21):**4642-4646.
14. Muller da Silva JP, Acencio ML, Merino Mornbach JC, Vieira R, da Silva JC, Lemke N, Sinigagliac M: **In silico network topology-based prediction of gene essentiality.** *PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS* 2008, **387(4):**1049-1055.
15. **Profiling of E. coli Chromosome (PEC) database** [http://shigen.lab.nig.ac.jp/ecoli/pec/]
16. **SGD: Saccharomyces cerevisiae Genome Snapshot/Overview** [http://www.yeastgenome.org/cache/genomeSnapshot.html]
17. Visa S, Ralescu A: **Issues in Mining Imbalanced Data Sets - A Review Paper.** *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference* 2005:67-73.
18. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988, **44(3):**837-845.
19. Estrada E: **Virtual identification of essential proteins within the protein interaction network of yeast.** *Proteomics* 2006, **6:**35-40.
20. Wuchty S: **Evolution and topology in the yeast protein interaction network.** *Genome Res* 2004, **14(7):**1310-4.
21. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425(6959):**686-91.
22. Lee K, Chuang HY, Beyer A, Sung MK, Huh WK, Lee B, Ideker T: **Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species.** *Nucleic Acids Res* 2008, **36(20):**e136.
23. **SGD: SGD Gene Nomenclature Conventions** [http://www.yeastgenome.org/help/yeastGeneNomenclature.shtml]
24. Vergara IA, Norambuena T, Ferrada E, Slater AW, Melo F: **StAR: a simple tool for the statistical comparison of ROC curves.** *BMC Bioinformatics* 2008, **9:**265.
25. Girard JP, Lehtonen H, Caizergues-Ferrer M, Amalric F, Tollervey D, Lapeyre B: **GAR1 is an essential small nucleolar RNP protein required for pre-rRNA processing in yeast.** *EMBO J* 1992, **11(2):**673-682.
26. Jager S, Strayle J, Heinemeyer W, Wolf DH: **Cic1, an adaptor protein specifically linking the 26S proteasome to its substrate, the SCF component Cdc4.** *EMBO J* 2001, **20(16):**4423-4431.
27. Saleh A, Schieltz D, Ting N, McMahon SB, Litchfield DW 3rd, Yates JR, Lees-Miller SP, Cole MD, Brandl CJ: **Tra1p is a component of the yeast Ada.Spt transcriptional regulatory complexes.** *J Biol Chem* 1998, **273(41):**26559-26565.
28. Daugeron MC, Linder P: **Characterization and mutational analysis of yeast Dbp8p, a putative RNA helicase involved in ribosome biogenesis.** *Nucleic Acids Res* 2001, **29(5):**1144-1155.
29. Ray BL, White CI, Haber JE: **The TSM1 gene of Saccharomyces cerevisiae overlaps the MAT locus.** *Curr Genet* 1991, **20(1-2):**25-31.
30. mei Li J, Li Y, Elledge SJ: **Genetic analysis of the kinetochore DASH complex reveals an antagonistic relationship with the ras/protein kinase A pathway and a novel subunit required for Ask1 association.** *Mol Cell Biol* 2005, **25(2):**767-778.
31. Grava S, Dumoulin P, Madania A, Tarassov I, Winsor B: **Functional analysis of six genes from chromosomes XIV and XV of Saccharomyces cerevisiae reveals YOR145c as an essential gene and YNL059c/ARP5 as a strain-dependent essential gene encoding nuclear proteins.** *Yeast* 2000, **16(11):**1025-1033.
32. Jackson SP, Lossky M, Beggs JD: **Cloning of the RNA8 gene of Saccharomyces cerevisiae, detection of the RNA8 protein, and demonstration that it is essential for nuclear pre-mRNA splicing.** *Mol Cell Biol* 1988, **8(3):**1067-1075.
33. Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, Giaever G: **Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast.** *Genetics* 2005, **169(4):**1915-1925.
34. Hyle JW, Shaw RJ, Reines D: **Functional distinctions between IMP dehydrogenase genes in providing mycophenolate resistance and guanine prototrophy to yeast.** *J Biol Chem* 2003, **278(31):**28470-28478.
35. Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, Carter CD, Wheeler D, Davis RW, Boeke JD, Snyder MA, Basrai MA: **Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae.** *Genome Res* 2006, **16(3):**365-373.
36. Dudley AM, Janse DM, Tanay A, Shamir R, Church GM: **A global view of pleiotropy and phenotypically derived gene function in yeast.** *Mol Syst Biol* 2005, **1:**. 2005.0001
37. Tsurumi C, Shimizu Y, Saeki M, Kato S, Demartino GN, Slaughter CA, Fujimuro M, Yokosawa H, Yamasaki M, Hendil KB, Toh-e A, Tanahashi N, Tanaka K: **cDNA cloning and functional analysis of the p97 subunit of the 26S proteasome, a polypeptide identical to the type-1 tumor-necrosis-factor-receptor-associated protein-2/55.11.** *Eur J Biochem* 1996, **239(3):**912-921.
38. Roberts SM, Winston F: **SPT20/ADA5 encodes a novel protein functionally related to the TATA-binding protein and important for transcription in Saccharomyces cerevisiae.** *Mol Cell Biol* 1996, **16(6):**3206-3213.
39. Imbeault D, Gamar L, Rufiange A, Paquet E, Nourani A: **The rtt106 histone chaperone is functionally linked to transcription elongation and is involved in the regulation of spurious transcription from cryptic promoters in yeast.** *J Biol Chem* 2008, **283(41):**27350-27354.
40. Gonzalez-Aguilera C, Tous C, Gomez-Gonzalez B, Huertas P, Luna R, Aguilera A: **The THP1-SAC3-SUS1-CDC31 complex works in transcription elongation-mRNA export preventing RNA-mediated genome instability.** *Mol Biol Cell* 2008, **19(10):**4310-4318.
41. Lillo JA, Andaluz E, Cotano C, Basco R, Cueva R, Correa J, Larriba G: **Disruption and phenotypic analysis of six open reading frames from the left arm of Saccharomyces cerevisiae chromosome VII.** *Yeast* 2000, **16(4):**365-375.
42. Febres DE, Pramanik A, Caton M, Doherty K, McKoy J, Garcia E, Alejo W, Moore CW: **The novel BLM3 gene encodes a protein that protects against lethal effects of oxidative damage.** *Cell Mol Biol (Noisy-le-grand)* 2001, **47(7):**1149-1162.
43. Walke S, Bragado-Nilsson E, Séraphin B, Nagai K: **Stoichiometry of the Sm proteins in yeast spliceosomal snRNPs supports the heptamer ring model of the core domain.** *J Mol Biol* 2001, **308:**49-58.
44. Salgado-Garrido J, Bragado-Nilsson E, Kandels-Lewis S, Séraphin B: **Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin.** *EMBO J* 1999, **18(12):**3451-62.
45. Kingsford C, Salzberg SL: **What are decision trees?** *Nat Biotechnol* 2008, **26(9):**1011-1013.
46. Quinlan JR: *C4.5: programs for machine learning* San Francisco: Morgan Kaufmann; 1993.
47. Coulomb S, Bauer M, Bernard D, Marsolier-Kergoat MC: **Gene essentiality and the topology of protein interaction networks.** *Proc Biol Sci* 2005, **272(1573):**1721-5.
48. Zotenko E, Mestre J, O'Leary DP, Przytycka TM: **Why do hubs in the yeast protein interaction network tend to be essential:**

reexamining the connection between the network topology and essentiality. *PLoS Comput Biol* 2008, **4(8):**e1000140.
49. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, Cheung KH, Miller P, Gerstein M, Roeder GS, Snyder M: **Subcellular localization of the yeast proteome.** *Genes Dev* 2002, **16(6):**707-19.
50. Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M: **Genomic analysis of essentiality within protein networks.** *Trends Genet* 2004, **20(6):**227-231.
51. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, Dolinski K, Tyers M: **The BioGRID Interaction Database: 2008 update.** *Nucleic Acids Res* 2008:D637-40.
52. Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, Mira NP, Alenquer M, Freitas AT, Oliveira AL, Sa-Correia I: **The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae.** *Nucleic Acids Res* 2006:D446-51.
53. Huss M, Holme P: **Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks.** *IET Syst Biol* 2007, **1(5):**280-285.
54. **NetworkX package** [https://networkx.lanl.gov]
55. **SGD: SGD Gene Ontology Slim Mapper** [http://db.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl]
56. Witten IH, Frank E: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* San Francisco: Morgan Kaufmann; 2000.
57. Kittler J, Hatef M, Duin RP, Matas J: **On Combining Classifiers.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1998, **20(3):**226-239.
58. Kohavi R: **Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid.** *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 1996 [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.4952].
59. Breiman L: **Random forests.** *Machine Learning* 2001, **45:**5-32.
60. Shi H: **Best-first Decision Tree Learning.** In *Master Thesis* The University of Waikato; 2007.
61. Landwehr N, Hall M, Frank E: **Logistic Model Trees.** *Machine Learning* 2005, **95(1-2):**161-205.
62. Freund Y, Mason L: **The alternating decision tree learning algorithm.** In *Proceeding of the Sixteenth International Conference on Machine Learning* San Francisco: Morgan Kaufmann; 1999:124-133.
63. Breiman L: **Bagging predictors.** *Machine Learning* 1996, **24(2):**123.
64. Huang J, Ling CX: **Using AUC and Accuracy in Evaluating Learning Algorithms.** *IEEE Trans on Knowl and Data Eng* 2005, **17(3):**299-310.
65. Hand DJ, Till RJ: **A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems.** *Mach Learn* 2001, **45(2):**171-186.

Apêndice F - Trabalho publicado no periódico *BMC Genomics*

Trabalho publicado em 2010 onde os autores descrevem a utilização de aprendizado de máquina e dados de expressão gênica em larga escala, sublocalização celular e medidas de centralidade da rede integrada de interações entre genes humanos para a predição de genes mórbidos e drogáveis em humanos, assim como a descoberta das principais condições associadas à morbidade e drogabilidade dos genes. As contribuições do autor desta tese para esse trabalho foram *(i)* a construção da rede de interações gênicas de humanos, *(ii)* a determinação da melhor combinação de algoritmos de aprendizado de máquina para a predição dos genes mórbidos, *(iii)* a interpretação biológica dos resultados e *(iv)* a preparação do manuscrito. Parte desse trabalho faz parte do Capítulo 5 desta tese.

Costa et al. BMC Genomics 2010, 11(Suppl 5):S9
http://www.biomedcentral.com/1471-2164/11/S5/S9

BMC Genomics

PROCEEDINGS Open Access

A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data

Pedro R Costa, Marcio L Acencio*, Ney Lemke

From 5th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2009)
Angra Dos Reis, RJ, Brazil. 18-22 October 2009

Abstract

Background: The genome-wide identification of both morbid genes, i.e., those genes whose mutations cause hereditary human diseases, and druggable genes, i.e., genes coding for proteins whose modulation by small molecules elicits phenotypic effects, requires experimental approaches that are time-consuming and laborious. Thus, a computational approach which could accurately predict such genes on a genome-wide scale would be invaluable for accelerating the pace of discovery of causal relationships between genes and diseases as well as the determination of druggability of gene products.

Results: In this paper we propose a machine learning-based computational approach to predict morbid and druggable genes on a genome-wide scale. For this purpose, we constructed a decision tree-based meta-classifier and trained it on datasets containing, for each morbid and druggable gene, network topological features, tissue expression profile and subcellular localization data as learning attributes. This meta-classifier correctly recovered 65% of known morbid genes with a precision of 66% and correctly recovered 78% of known druggable genes with a precision of 75%. It was than used to assign morbidity and druggability scores to genes not known to be morbid and druggable and we showed a good match between these scores and literature data. Finally, we generated decision trees by training the J48 algorithm on the morbidity and druggability datasets to discover cellular rules for morbidity and druggability and, among the rules, we found that the number of regulating transcription factors and plasma membrane localization are the most important factors to morbidity and druggability, respectively.

Conclusions: We were able to demonstrate that network topological features along with tissue expression profile and subcellular localization can reliably predict human morbid and druggable genes on a genome-wide scale. Moreover, by constructing decision trees based on these data, we could discover cellular rules governing morbidity and druggability.

Background

Currently, the large-scale experimental identification of both morbid genes, i.e. those genes whose mutations cause hereditary human diseases, and druggable genes, i.e. genes coding for proteins whose modulation by small molecules elicits phenotypic effects, demands time-consuming and laborious approaches that are impractical for rapidly revealing the causal relationships between genes and diseases and determining the druggability of gene products. The discovery of morbid genes, for instance, requires a large effort to gather inheritance patterns from families with the disease and to perform linkage and mutation analyses in order to identify

* Correspondence: mlacencio@ibb.unesp.br
Departamento de Física e Biofísica, Instituto de Biociências de Botucatu, UNESP - Univ Estadual Paulista, Distrito de Rubião Jr. s/n, Botucatu, São Paulo, 18618-970, Brazil
Full list of author information is available at the end of the article

candidate gene(s) involved in a particular hereditary disorder [1]. In similar fashion, the discovery of new drug targets also requires a large effort involving a variety of genomics, proteomics, genetic association and forward and reverse genetics-related techniques [2] in order to find drugs capable to modulate disease processes.

In the light of above mentioned facts, a computational approach which could accurately predict morbid and druggable genes, especially on a genome-wide scale, would be thus invaluable since the number of experimental techniques to be performed to discover these genes could be minimized. With the vast amount of current available systems-level data, such as molecular interaction data and genome-wide gene expression and subcellular localization data, we have now the opportunity for developing a computational approach based on data mining tools, such as machine learning, to extract patterns that could be used as genome-wide predictors of morbid and druggable genes. Based on this assumption, we have previously used a machine learning-based methodology as a data mining tool to extract knowledge from systems-level data and then apply this knowledge to predict essential genes on a genome-wide scale and determine cellular rules for essentiality on *Escherichia coli*[3] and *Saccharomyces cerevisiae*[4]. In addition to attain successful prediction rates, we have also obtained biologically plausible cellular rules for gene essentiality using this machine learning approach.

Due this successful prediction of essential genes and determination of cellular rules for gene essentiality in *Escherichia coli* and *Saccharomyces cerevisiae*, we sought to verify in this present work whether a similar machine learning-based approach is able to predict human morbid and druggable genes on a genome-wide scale and to reveal cellular rules governing morbidity and druggability of genes. Using knowledge acquired from network topological features, tissue expression profile and subcellular localization data, we show here that the classifiers trained on these systems-level data can reliably predict morbid and druggable genes on a genome-wide scale and also can define some general rules governing morbidity and druggability in human.

Results and Discussion

The integrated network of human gene interactions and calculation of topological features

For obtaining the network topological features used as training data for predicting morbid and druggable genes, we first constructed an integrated network of human gene interactions (INHGI) simultaneously containing experimentally verified protein physical interactions, metabolic interactions and transcriptional regulation interactions (definitions for each type of interaction are detailed in "Methods"). This network is comprised by 10,241 genes interacting with one another via 43,342 protein physical interactions, 24,540 metabolic interactions and 3,015 transcriptional regulation interactions. INHGI contains approximately 25% of the already identified ≈ 45,000 human genes according to the Entrez-Gene database [5].

From the INHGI, we calculated 12 different topological features for each gene, including degree centralities for each type of interaction, clustering coefficient, betweenness centralities for each type of interaction, closeness centrality and identicalness. The detailed description of these topological features and how they were calculated are found in the Additional file 1 and "Methods".

Evaluation of classifier performance

To examine how well a machine learning-based approach is able to predict human morbid and druggable genes on a genome-wide scale using knowledge acquired from systems-level data, we designed a meta-classifier similar to that used to predict essential genes in *Escherichia coli*[3] and *Saccharomyces cerevisiae*[4] and trained it on network topological features, tissue expression profile and subcellular localization data of known morbid and druggable genes (see "Methods" for details). We then assessed its performance by measuring its median recall, precision and area under the curve (AUC) of the receiver operating characteristic (ROC) curve across 10 different normal morbidity datasets and 10 different normal druggability datasets (see "Methods" for more details).

Before analyzing the performance measures of our meta-classifier trained on the datasets described above, we decided to estimate the performance measures of our meta-classifier on equivalent normal morbidity and druggability datasets where the class labels—morbid and druggable—were randomly shuffled among genes (shuffled morbidity and shuffled druggability datasets) and then compared them with our meta-classifier trained on the normal morbidity and druggability datasets. This was done to check whether the meta-classifier trained on non-shuffled datasets learned the traits actually associated with morbidity and druggability instead of traits associated with any random subset of genes. For this comparison, we used the Wilcoxon signed-rank statistical test as described in "Methods". As can be observed in Table 1, all performance measures of our meta-classifier trained on the correspondent shuffled datasets were statistically different from measures of meta-classifier trained on normal datasets (for all performance measures, $W \leq W_c$ with $N = 10$ at the $p = 0.05$ level; see "Methods" and [6]), thereby indicating that the

Table 1 Classifier performance measures for prediction of morbid and druggable genes

Prediction of morbid genes					
Performance measure	Median [min,max] [1]	Median [min,max] [1]	*N*	*W*	W_c (two-tailed $p = 0.05$)[2]
	Normal	Shuffled			
Precision	0.658 [0.648,0.679]	0.495 [0.473,0.522]	10	0	8 *
Recall	0.648 [0.632,0.657]	0.502 [0.471,0.521]	10	0	8 *
AUC	0.716 [0.706,0.729]	0.498 [0.462,0.526]	10	0	8 *
Prediction of druggable genes					
Performance measure	Median [min,max] [1]	Median [min,max] [1]	*N*	*W*	W_c (two-tailed $p = 0.05$)[2]
	Normal	Shuffled			
Precision	0.748 [0.72,0.763]	0.5 [0.451,0.556]	10	0	8 *
Recall	0.782 [0.732,0.809]	0.492 [0.447,0.564]	10	0	8 *
AUC	0.820 [0.801,0.835]	0.500 [0.43,0.546]	10	0	8 *

[1] Of 10 datasets

[2] According to table of critical values for *W* in [6]

* Difference statistically significant

traits actually associated with morbidity and druggability were learned by our meta-classifier.

After confirmation that our meta-classifier trained on normal datasets was likely to learn the traits actually associated with morbidity and druggability, we aimed to analyze its performance measures. As shown in Table 1, for the genome-wide prediction of morbid genes, our meta-classifier achieved a median recall of 0.648 and a median precision of 0.658, i.e., it correctly recovered 64.8% of known morbid genes with a precision of 65.8%. Furthermore, the probability of a gene predicted as morbid belongs to the set of known morbid genes is 71.2% as indicated by the median AUC. For the genome-wide prediction of druggable genes, our meta-classifier achieved a median recall of 0.782 and a median precision of 0.748, i.e, it correctly recovered 78.2% of known druggable genes with a precision of 74.8% (Table 1). Furthermore, the probability of a gene predicted as druggable belongs to the set of known druggable genes is 82.0% as indicated by the median AUC.

The moderate values for both median recall (0.648) and median precision (0.662) for genome-wide prediction of morbid genes indicate that the level of noise in the training data is high and likely associated with existence of shared common features between morbid and non-morbid genes that induced our meta-classifier to yield a moderate performance in discriminating morbid from non-morbid genes. This could be partially due to the approach used to select non-morbid genes: since it is impossible at present to compile a list of genes not known to cause any hereditary disease, we selected genes not known to be morbid, i.e., all genes in INHGI except the known morbid genes, as non-morbid genes. Thus, some of these non-morbid genes may actually be existing unknown morbid genes sharing common characteristics with the existing known morbid genes. Other contributing factor for the existence of shared common features between morbid and non-morbid genes could be the incompleteness of INHGI: Stumpf *et al.*[7], for example, have estimated that the size of human interactome (only protein-protein interactions) is about 650,000 interactions. Since our network contains about 43,000 protein-protein interactions, we could envisage that the values of all network topological parameters might change with the enlargement of network size and, therefore, some of the network topological parameters-related shared common features between morbid and non-morbid might disappear as a consequence. The existence of shared common features between druggable and non-druggable genes also seems to affect the performance of our meta-classifier, but to a lesser extent: our meta-classifier achieved reliables values for the median recall (0.782) and precision (0.748) for genome-wide prediction of druggable genes (Table 1).

Despite these limitations discussed above, our meta-classifier trained on network topological features, tissue expression profile and subcellular localization data seems indeed to be a reliable predictor of morbid and druggable genes on a genome-wide scale as shown by Figures 1 and 2: the frequency distribution of known morbid and known druggable genes per intervals of morbidity and druggability scores—probabilities of classifying genes as morbid and druggable, respectively, as output by the meta-classifier (see "Prediction of novel morbid and druggable genes" and "Methods" for more details)—tend to increase as morbidity (Figure 1) and druggability (Figure 2) scores increase.

Evaluation of individual features on classifier performance

We sought to verify the influence of individual features on the meta-classifier performance. To achieve this goal, we first trained our meta-classifier on normal morbidity

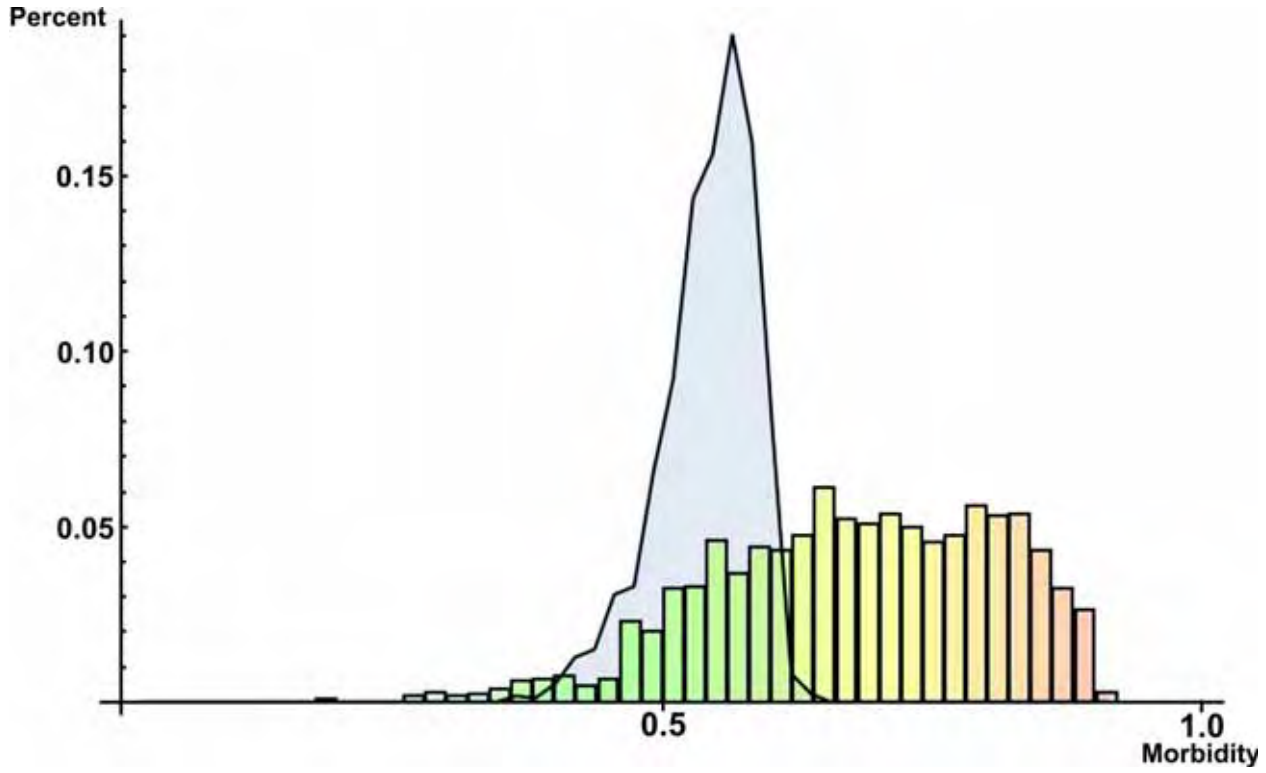

Figure 1 Frequency distribution of known morbid genes per intervals of morbidity scores Bars show the frequency distribution of known morbid genes (in percent) per 0.2 intervals of normal morbidity scores.The blue-shaded area represents the frequency distribution of known morbid genes (in percent) per intervals of shuffled morbidity scores.

and druggability datasets without one of the features, which we call "without-one-feature" datasets as described in "Methods" . We then compared the output AUC values with those of meta-classifier trained on datasets with all features by using the Wilcoxon signed-rank statistical test [6]. A difference is considered statistically significant If the obtained W is lower than or equal to W_c with a given N at the p = 0.05 level (see "Methods"). Note that we use AUC instead of recall or precision to compare the overall performances of meta-classifiers because it represents the meta-classifier performance across all combinations of recall and precision (see "Methods"). Table 2 shows that the median AUC of our meta-classifier trained on morbidity datasets without the number of tissues in which the gene is expressed at least 5 transcripts per million (tpm) (see "Methods" for details) was statistically lower than the median AUC for normal morbidity datasets (W = 7 *versus* W_c = 8 for N = 10 at p = 0.05). So, the tissue expression profile seems to be an important feature to distinguish morbids from non-morbids genes.

As shown in Table 3, for prediction of druggable genes, the overall performance (AUC) of our meta-classifier was statistically lower following the removal of the plasma membrane feature (W = 1 *versus* W_c = 8 for N = 10 at p = 0.05). This result is in concert with the most important cellular rule for druggability derived from the analysis of decision trees (see more details in "Methods") that we will show in the section"Cellular rules for gene morbidity and druggability"): if proteins are located in plasma membrane, their encoding genes are likely to be druggable. This rule is supported by Bakheet and Doig [8] that demonstrated that proteins encoded by druggable genes had more transmembrane helices than proteins encoded by non-druggable ones which suggests that proteins encoded by druggable genes are more likely to be found in plasma membrane.

Comparison with other methods

Regarding prediction of morbid genes, there have been several methods available for predicting morbid genes [9-16]. However, our method can not be directly compared to most of them since they have been constructed to predict only small sets of disease-specific candidate genes, such as ENDEAVOUR [13] and ToppGene [15], while our method has been constructed for the genome-wide prediction of morbid genes. We can, however, compare our method to PROSPECTR [9], CIPHER [14]

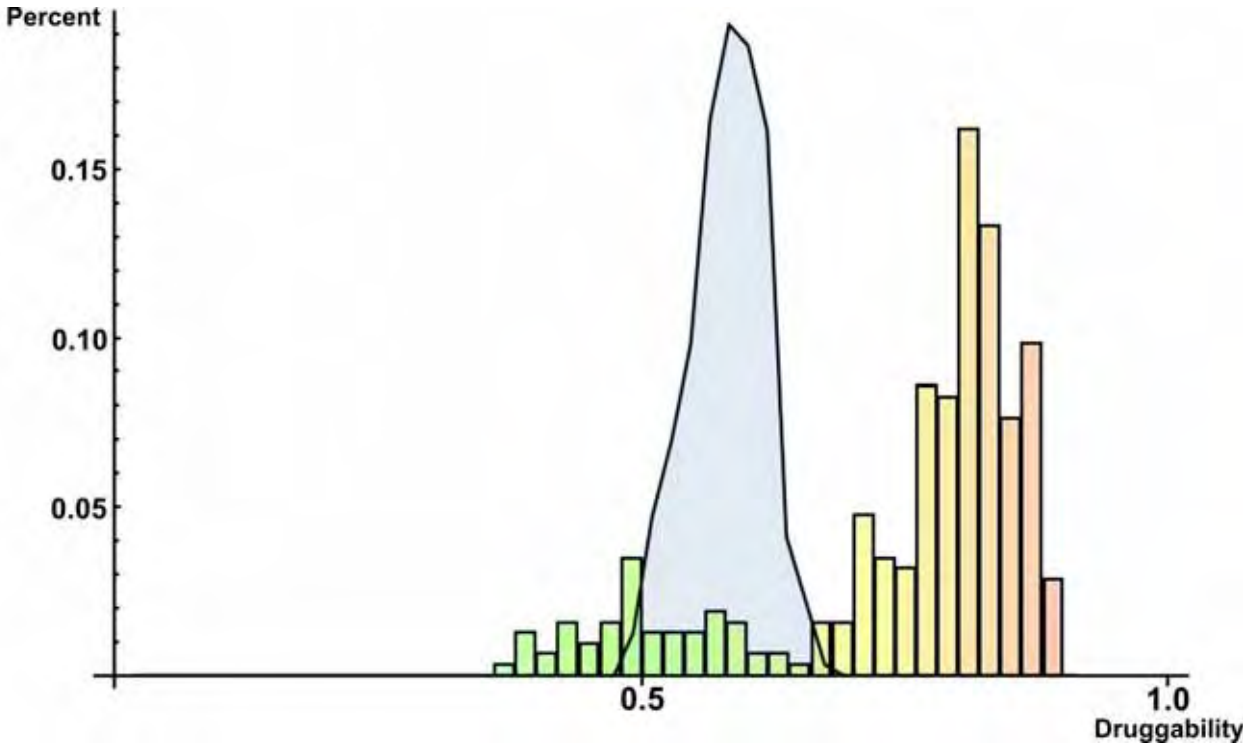

Figure 2 Frequency distribution of known druggable genes per intervals of druggability scores Bars show the frequency distribution of known druggable genes (in percent) per 0.2 intervals of normal druggability scores.The blue-shaded area represents the frequency distribution of known druggable genes (in percent) per intervals of shuffled druggability scores

and that developed by Xu and Li [16]. Our method outperforms CIPHER (this method, for genome-wide prediction, yields a precision of about 0.1; there is no value of recall reported) and is comparable to PROSPECTR that achieves a recall of 0.70, a precision of 0.62 and an AUC of 0.70. Although PROSPECTR has a higher recall, we considered our method comparable to it as the precision and AUC values of our method are higher than those of PROSPECTR. Moreover, our performance measures are medians of 10 runs of 10-cross-fold validation (see "Methods" for more details), while the performance measures of PROSPECTR were obtained by only one run of 10-cross-fold validation.

The method developed by Xu and Li is the only genome-wide prediction method that apparently outperforms our method (this method achieves, for genome-wide prediction, an average recall about 0.78 and an average precision about 0.77). Their method is also based on network topological parameters, but while we trained our meta-classifier on various features, including 12 network topological parameters (see "Methods" and Additional file 1), they trained their classifiers on only five network topological parameters: degree, defined as the number of links to node *i*; 1N index, defined as the proportion of the number of links to morbid genes among all links to node *i*; 2N index, defined as the proportion of the number of links to morbid genes among all links to neighbors of node *i*; the average distance to morbid genes; and positive topological coefficient, a variant of the classical topological coefficient [17]. The apparent success of Xu and Li approach in predicting morbid genes mostly relies on the 2N index: when node *i* is a morbid gene, 2N index is always higher than zero since at least one neighbor of node *i*'s neighbor—the node *i* itself—is a morbid gene; if node *i* is a non-morbid gene, 2N index is higher than or equal to zero. Thus, this parameter induces a spurious correlation on dataset that is captured by classifiers that, in turn, achieve high performance measures. Therefore, the Xu and Li method can be disregarded for comparison purposes and, accordingly, our approach, although showing moderate recall and precision values, is currently, along with PROSPECTR, the most accurate predictor of morbid genes on a genome-wide scale.

Concerning the prediction of druggable genes, as for prediction of morbid genes, we can compare our method only with those developed to predict druggable genes on a genome-wide scale. Therefore, to our knowledge, we can compare our methodology with that developed by Sugaya and Ikeda [18]. Using support vector

Table 2 Statistical comparison of performances of classifiers trained on normal and without-one-feature morbidity datasets

Missing feature [1]	Median AUC [min,max][2]	N	W	W_c (two-tailed $p = 0.05$)[3]
ppi	0.715 [0.705,0.726]	10	26	8
metin	0.714 [0.707,0.727]	10	26	8
metout	0.713 [0.707,0.729]	10	25	8
regin	0.714 [0.703,0.726]	9	18	6
regout	0.716 [0.705,0.729]	10	26	10
c	0.713 [0.701,0.724]	10	13	8
identicalness	0.711 [0.704,0.727]	10	24	8
cent	0.714 [0.707,0.727]	10	25	8
inbet	0.716 [0.708,0.731]	10	25	8
inbetppi	0.714 [0.707,0.727]	9	21	6
inbetmet	0.714 [0.707,0.728]	9	21	6
inbetreg	0.715 [0.706,0.727]	10	25	8
numtissuesexp[4]	0.709 [0.701,0.719]	10	7	8*
avegexptec[5]	0.715 [0.704,0.727]	10	27	8
Unknown	0.713 [0.701,0.725]	10	18	8
Cytoplasm	0.715 [0.706,0.728]	10	26	8
Endoplasmic reticulum	0.716 [0.705,0.727]	10	26	8
Mitochondrion	0.714 [0.706,0.728]	10	24	8
Nucleus	0.715 [0.704,0.728]	10	24	8
Other localization	0.714 [0.704,0.726]	10	21	8
Cellular component	0.714 [0.705,0.727]	9	21	6
Extracellular space	0.710 [0.7,0.723]	10	14	8
Golgi apparatus	0.715 [0.706,0.728]	10	26	8
Median AUC [min,max] for normal datasets: 0.716 [0.706,0.729]				

[1] See "Methods" and Additional file 1 for a description of features

[2] Of 10 datasets

[3] According to table of critical values for W in [6]

[4] The number of tissues (out of 32) in which the gene is expressed at least 5 transcripts per million (tpm) according to Reverter *et al.* [33]

[5] The average expression in tpm among all the tissues in which the gene is expressed according to Reverter *et al.* [33]

* Difference statistically significant

machines trained on 69 different features covering structural, drug and chemical, and functional information on protein-protein interactions, Sugaya and Ikeda classifiers achieved an average recall of 75%, an average precision of 70% and an average AUC of 72%, performance measures comparable to those obtained by our meta-classifier.

Prediction of novel morbid and druggable genes

Since the morbidity and druggability of most of genes in INHGI are unknown—only ≈ 14% and ≈ 3% are known to be morbid and druggable, respectively—we applied our trained meta-classifier to determine the morbidity and druggability statuses of these genes. Instead of simply predicting genes as morbid or druggable, we decided to assign a "morbidity score" and a "druggability score" (see "Methods") to each gene since we understand that there is no gene that is absolutely non-morbid or non-druggable. We also assigned to each gene a "shuffled morbidity score" and a "shuffled druggability score" to test the significance of normal scores. For this purpose, we used the Wilcoxon signed-rank statistical test as described in "Methods".

Table 4 shows genes not known to be morbid with the 10 highest morbidity scores (see Additional file 2 for the normal and shuffled morbidity scores of all genes in INHGI). All these scores are significantly higher than the shuffled scores ($W \leq W_c$ with $N = 10$ at the p = 0.05 level; see "Methods" and [6]). With the purpose of investigating whether the assigned scores resemble the potential morbidities of these genes, we mined the Human Genome Epidemiology Network (HuGENet) database [19] for articles clearly stating that such genes may be associated with some disease, which we call as "morbidity evidences". According to this approach, we found that 10 of 11 (≈ 90%) genes with the 10 highest morbidity scores are considered to be associated with some disease (Table 4). This shows that our meta-classifier is quite capable of assigning high morbidity scores to genes potentially morbid.

Table 3 Statistical comparison of performances of classifiers trained on normal and without-one-feature druggability datasets

Missing feature [1]	Median AUC [min,max][2]	*N*	*W*	W_c (two-tailed $p = 0.05$)[3]
ppi	0.819 [0.798,0.835]	10	27	8
metin	0.817 [0.803,0.834]	10	26	8
metout	0.817 [0.801,0.832]	9	20	6
regin	0.818 [0.799,0.83]	9	18	6
regout	0.818 [0.801,0.833]	10	26	8
c	0.821 [0.799,0.836]	10	21	8
identicalness	0.819 [0.8,0.836]	10	27	8
cent	0.814 [0.797,0.832]	10	18	8
inbet	0.821 [0.804,0.837]	10	25	8
inbetppi	0.819 [0.803,0.833]	10	25	8
inbetmet	0.82 [0.791,0.833]	10	26	8
inbetreg	0.818 [0.802,0.83]	9	19	6
numtissuesexp[4]	0.806 [0.795,0.832]	9	11	6
avegexptec[5]	0.814 [0.799,0.835]	10	23	8
Unknown	0.816 [0.796,0.832]	9	12	6
Cytoplasm	0.814 [0.794,0.834]	10	20	8
Endoplasmic reticulum	0.820 [0.799,0.834]	10	27	8
Mitochondrion	0.820 [0.796,0.831]	9	22	6
Nucleus	0.816 [0.793,0.831]	10	20	8
Other localization	0.821 [0.802,0.837]	9	20	6
Cellular component	0.82 [0.801,0.835]	10	25	8
Extracellular space	0.817 [0.8,0.837]	10	26	8
Golgi apparatus	0.812 [0.8,0.834]	10	24	8
Plasma membrane	0.781 [0.762,0.816]	10	1	8*

Median AUC [min,max] for normal datasets : 0.820 [0.801,0.835]

[1] See "Methods" and Additional file 1 for a description of features

[2] Of 10 datasets

[3] According to table of critical values for *W* in [6]

[4] The number of tissues (out of 32) in which the gene is expressed at least 5 transcripts per million (tpm) according to Reverter *et al.* [33]

[5] The average expression in tpm among all the tissues in which the gene is expressed according to Reverter *et al.* [33]

* Difference statistically significant

Table 5 shows genes not known to be druggable with the 10 highest druggability scores (see Additional file 2 for the normal and shuffled druggability scores of all genes in INHGI). All these scores are significantly higher than the shuffled scores ($W \leq W_c$ with $N = 10$ at the p = 0.05 level; see "Methods" and [6]). With the purpose of investigating whether the assigned scores resemble the potential druggabilities of these genes, we mined the literature for articles clearly stating that such genes may be drug target candidates, which we call as "druggability evidences". According to this approach, we found that 8 of 11 (≈ 73%) genes with the 10 highest druggability scores are considered to be drug target candidates (Table 5). This shows that our meta-classifier is quite capable of assigning high druggability scores to genes potentially druggable. Among these candidates, five (*PLAU*, *CD8A*, *CD19*, *ITGAM* and *IL6*) are known morbid genes and two (*THBS1* and *TIMP2*) are within the list of genes with the 10 highest morbidity scores. About the known morbid genes with druggability evidence—*PLAU*, *CD19*, *ITGAM* and *IL6*—, it is interesting to note that the druggabilities assigned to these genes by our classifier are not related to the diseases caused by their corresponding mutated versions. The gene *PLAU* is a susceptibility gene for late-onset Alzheimer disease according to the Online Mendelian Inheritance in Man (OMIM) database [20] (MIM # 191840), but the protein encoded by this gene seems to be a good candidate target for treatment of cancer in combination with conventional therapeutics such as chemotherapy or radiation [21]. Similarly, mutations in the gene *CD19* cause antibody deficiency that increases susceptibility to infection ([22] (MIM #107265), but its encoded protein has proven to be a promise as a novel and well-tolerated therapy in B-cell non-Hodgkin's lymphoma [23]. Regarding *ITGAM*, while Yang *et al.*[24] have confirmed the association of the this gene with disease susceptibility and renal nephritis of systemic lupus erythematosus (MIM # 609939), Romano *et al.*[25], on the other hand, have suggested that the protein encoded by *ITGAM* is a potential target of the femtomolar-acting eight-amino-acid peptide for protection against the deleterious effects of closed head injury in mice. Finally, according to OMIM database (MIM # 147620), the gene *IL6* mediates growth failure in Crohn disease [26], but we found that its encoded protein is a promising target for therapy of several chronic inflammatory and autoimmune diseases as well as in cancer [27]. These findings show that our classifier, besides discovering new druggable genes, can also reveal unexpectedly roles for known morbid genes in the modulation of diseases caused by other seemingly unrelated genes.

Two potential morbid genes, *THBS1* and *TIMP2*, reinforce the fact that our meta-classifier is able to reveal unexpectedly roles for morbid genes in the modulation of diseases caused by other seemingly unrelated genes. Mutations in the gene *THBS1* have been suggested to play a role in atherosclerosis and thrombosis [28], but its encoded protein may be considered a promising therapeutic target for diabetic nephropathy [29]; alterations in *TIMP2* has been demonstrated to be one of the causes of chronic obstructive pulmonary disease [30], but targeting its encoded protein may be a therapeutic intervention against connective amino acid tissue degradation [30].

Cellular rules for gene morbidity and druggability

Beyond the prediction capability, machine learning techniques can be used for knowledge acquisition in order to

Table 4 List of the human genes in the INHGI with the 10 highest morbidity scores

Gene	Morbidity score	(Median [min,max])[1]	N	W	W_c[2] (two-tailed $p = 0.05$)	Morbidity evidence[3]
	Normal	Shuffled				
TFRC	0.880 [0.576,0.939]	0.568 [0.447,0.678]	10	1	8*	5941956
ITGA5	0.875 [0.635,0.916]	0.491 [0.377,0.631]	10	0	8*	No evidence
LTF	0.868 [0.803,0.913]	0.509 [0.356,0.642]	10	0	8*	19258923
SFTPD	0.866 [0.618,0.923]	0.565 [0.458,0.682]	10	2	8*	19590686
THBS1	0.865 [0.831,0.918]	0.511 [0.354,0.566]	10	0	8*	18178577
TIMP2	0.860 [0.603,0.92]	0.574 [0.388,0.609]	10	0	8*	19933216
TGFB2	0.857 [0.565,0.918]	0.526 [0.407,0.707]	10	3	8*	19258923
CGA	0.856 [0.62,0.916]	0.535 [0.283,0.656]	10	0	8*	19730683
SPP1	0.856 [0.577,0.887]	0.564 [0.34,0.696]	10	0	8*	15868370
FLT1	0.854 [0.61,0.931]	0.527 [0.424,0.715]	10	3	8*	19741061
NOL3	0.850 [0.647,0.875]	0.576 [0.31,0.651]	10	1	8*	19773279

[1] Of 10 scores

[2] According to table of critical values for *W* in [6]

[3] Pudmed IDs of most recent article(s) clearly stating a gene-disease association

* Difference statistically significant

describe patterns in datasets. The machine learning algorithms most used for knowledge acquisition are those that generate decision trees. Decision trees are decision support tools inferred from the training data that use a graph of conditions and their possible consequences. The structure of a decision tree consists of a root node representing the most important condition for discriminating classes, internal nodes representing additional conditions for class discrimination under the main condition, and leaf nodes representing the final classification. So, one can learn the conditions for classifying instances in a given class by following the path from the root node to the leaf node [31].

Therefore, in order to discover the rules for gene morbidity and druggability, we analyzed decision trees generated by training the J48 algorithm, a WEKA's implementation of the C4.5 algorithm [32] (for more details, see "Methods"), on the normal morbidity and druggability datasets containing all network topological features, tissue expression profiles and subcellular localization as training data. The decision trees in Figures 3 and 4 are the best representative tree among the 10 generated decision trees for morbidity (Figure 3) and the 10 generated decision trees for druggability (Figure 4).

From the best representative decision tree for morbidity, we were able to devise some general rules for

Table 5 List of the human genes in the INHGI with the 10 highest druggability scores

Gene	Druggability score	(Median [min,max])[1]	N	W	W_c[2] (two-tailed $p = 0.05$)	Druggability evidence[3]
	Normal	Shuffled				
HLA-F	0.887[0.803,0.915]	0.530[0.427,0.584]	10	0	8*	No evidence
PLAU[4]	0.886[0.808,0.907]	0.561[0.387,0.675]	10	0	8*	19301652
CD8A[4]	0.885[0.871,0.902]	0.56[0.37,0.664]	10	0	8*	No evidence
CD19[4]	0.880[0.751,0.907]	0.562[0.38,0.628]	10	0	8*	19509168
ITGAM[4]	0.878[0.614,0.887]	0.534[0.36,0.656]	10	1	8*	11931348
THBS1[5]	0.875[0.53,0.9]	0.532[0.293,0.592]	10	0	8*	17878288
ITGAX	0.873[0.784,0.897]	0.539[0.422,0.691]	10	0	8*	No evidence
CXCR5	0.871[0.755,0.895]	0.537[0.49,0.59]	10	0	8*	17652619
EBI3	0.871[0.801,0.888]	0.529[0.391,0.626]	10	0	8*	19556516
IL6[4]	0.87[0.766,0.893]	0.591[0.361,0.643]	10	0	8*	17465721
TIMP2[5]	0.869[0.645,0.916]	0.584[0.34,0.701]	10	0	8*	10985804

[1] Of 10 scores

[2] According to table of critical values for *W* in [6]

[3] Pudmed IDs of most recent articles clearly stating that such genes may be drug target candidates

[4] Morbid genes according to Morbid Map [46]

[5] Genes among those with 10 highest morbidity scores (Table 4)

* Difference statistically significant

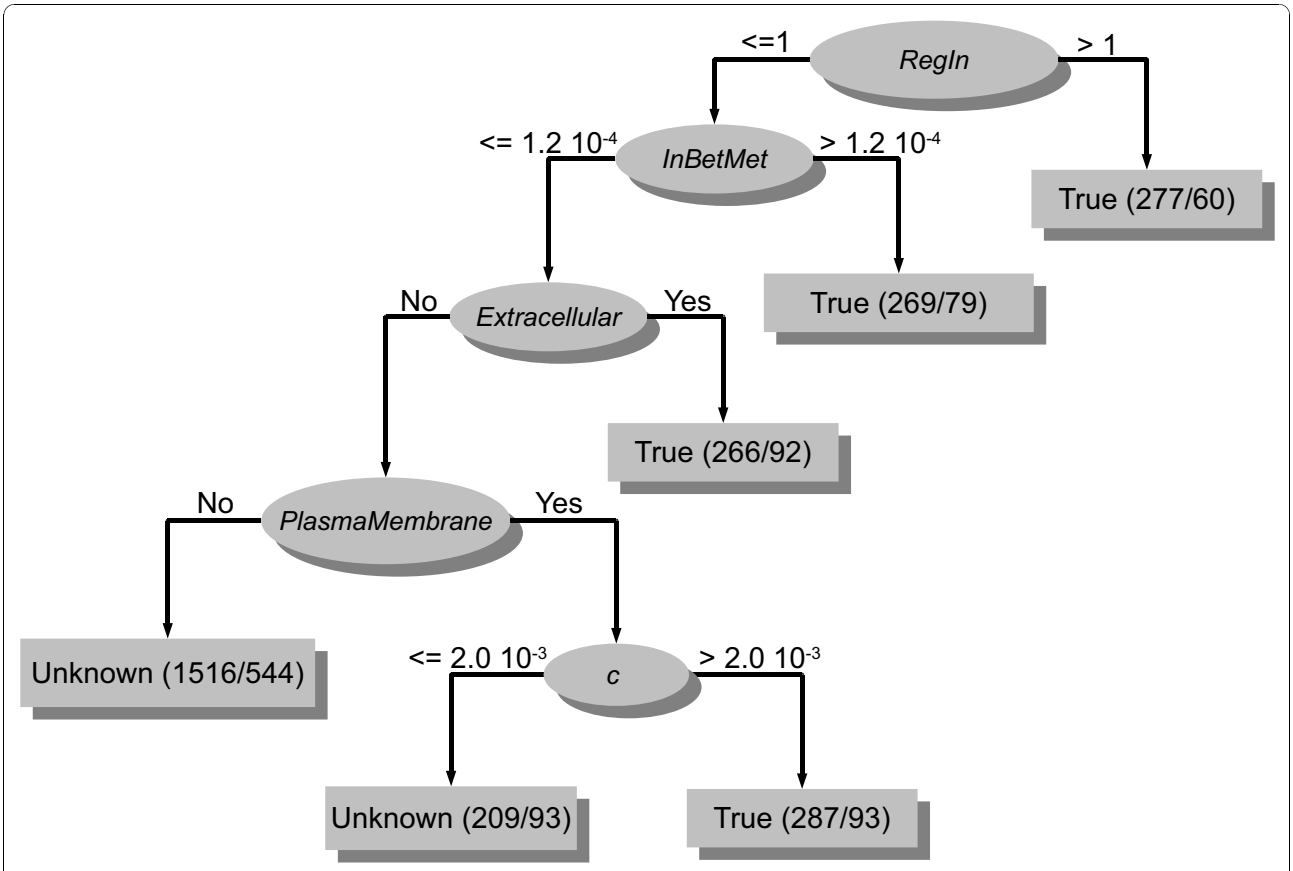

Figure 3 Decision tree generated by training the J48 algorithm on the normal morbidity datasets This decision tree was generated by training the J48 algorithm on the normal morbidity datasets (see "Methods"). The uppermost ellipse is the node root of tree that represents the most important condition for discriminating morbid genes from non-morbid genes. In this case, such condition is the number of transcription factors regulating the gene (*regin*). The remaining ellipses are internal nodes that represent additional conditions for considering a gene as morbid or non-morbid. In the left branch of tree, such conditions are a central position in a metabolic pathway (*inbetmet*), the extracellular or plasma membrane localization of respective encoded proteins and tendency of encoded proteins to form clusters with others (*c*). The rectangles depict genes that, under certain conditions (represented by the root node and internal nodes), are respectively and predominantly classified as morbid (**True**) and non-morbid (**Unknown**). In the round brackets inside rectangles, the number before the slash indicates the total number of genes that are actually morbid or non-morbid and the number after the slash indicates how many genes were incorrectly predicted.

morbidity in human. As we can observe in Figure 3, the root node of decision tree is the number of transcription factors that regulate a given gene (*regin*). So, this attribute can be considered the most important feature, among those used to train the J48 algorithm, for discriminating a morbid from a non-morbid gene. To reinforce this, we found, by walking the path from root node to first leaf node through the right branch, the following rule for morbidity: if genes are regulated by more than one transcription factor, they are likely to be morbid (Figure 3). The study by Reverter *et al.*[33] supports this rule as they showed that morbid genes are more likely to show tissue specific expression than non-morbid ones. Genes whose expression is tissue specific tend to be regulated by more transcription factors than those that are ubiquitously expressed, e.g. housekeeping genes, since a high level of transcriptional regulation is needed in this case.

Walking the path from root node to first and second leaf nodes through the left branch (Figure 4), we found the following rule for morbidity: if genes are regulated by one transcription factor and their encoded proteins are either centrally located in metabolic pathways (*inbetmet* is the betweenness centrality via metabolic interactions; see "Methods" and Additional file 1) or play a role in the extracellular region, genes are likely to be morbid. This rule is supported by Jimenez-Sanchez and colleagues [34] that showed that morbid genes are more likely to be enzymes than non-morbid ones and by Winter *et al.* [35] that demonstrated that ≈ 40% of proteins encoded by morbid genes are predicted to be secreted. Furthermore, if proteins are neither centrally located in

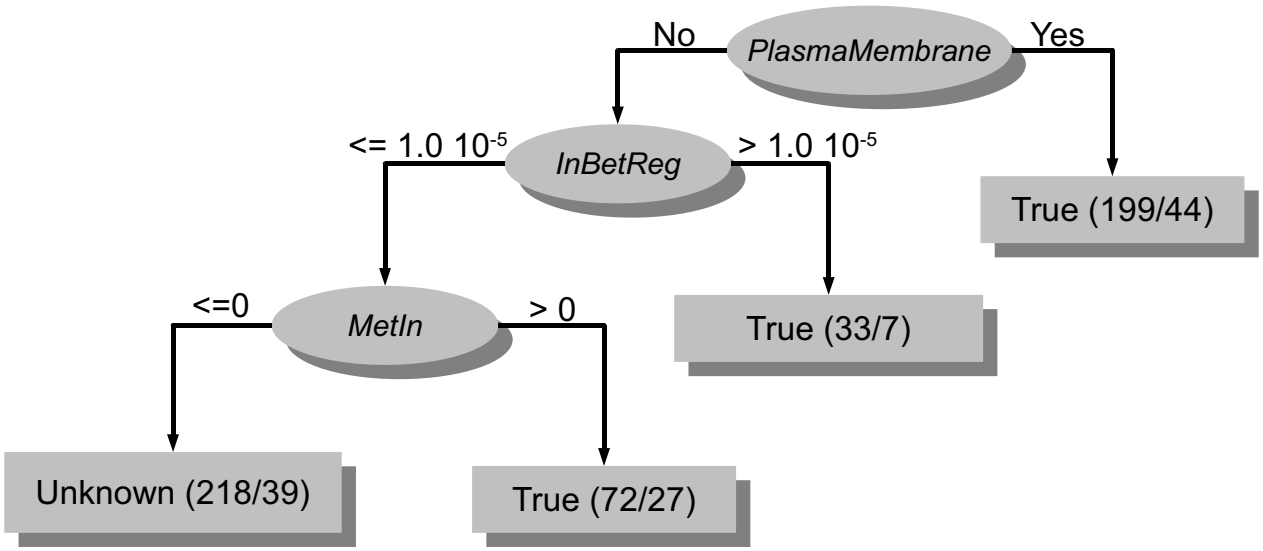

Figure 4 Decision tree generated by training the J48 algorithm on the normal druggability datasets This decision tree was generated by training the J48 algorithm on the normal druggability datasets (see "Methods"). The uppermost ellipse is the node root of tree that represents the most important condition for discriminating druggable genes from non-druggable genes. In this case, such condition is the plasma membrane localization of encoded proteins. The remaining ellipses are internal nodes that represent additional conditions for considering a gene as druggable or non-druggable. In the left branch of tree, such conditions are a central position in a transcriptional regulatory circuitry (*inbetreg*) and being an enzyme (*metin*). The rectangles depict genes that, under certain conditions (represented by the root node and internal nodes), are respectively and predominantly classified as druggable (**True**) and non-druggable (**Unknown**). In the round brackets inside rectangles, the number before the slash indicates the total number of genes that are actually druggable or non-druggable and the number after the slash indicates how many genes were incorrectly predicted.

metabolic pathways nor play a role in the extracellular region but are located in plasma membrane and tend to form clusters with other proteins (recall that *c* is the clustering coefficient, a network feature the measures the local group cohesiveness; see "Methods" and Additional file 1), their encoding genes are likely to be morbid. For this rule, we could not find any article supporting it. Therefore, the plasma membrane localization of proteins encoded by morbid genes as well as the tendency of these proteins to form clusters with other proteins are issues to be examined.

From the best representative decision tree for druggability, we were able to devise some general rules for druggability in human. As we can observe in Figure 4, the root node of decision tree is the plasma membrane localization of proteins. So, this attribute can be considered the most important feature, among those used to train the J48 algorithm, for discriminating a druggable from a non-druggable gene. To reinforce this, we found, by walking the path from root node to first leaf node through the right branch, the following rule for druggability: if proteins are located in plasma membrane, their encoding genes are likely to be druggable (Figure 4). This rule is supported by Bakheet and Doig [8] that demonstrated that proteins encoded by druggable genes had more transmembrane helices than proteins encoded by non-druggable ones which suggests that proteins encoded by druggable genes are more likely to be found in plasma membrane. Walking the path from root node to first and second leaf nodes through the left branch (Figure 4), we found the following rule for druggability: if proteins are not located in plasma membrane but are either centrally located in a transcriptional regulatory circuitry (*inbetreg* is the betweenness centrality via transcriptional regulation interactions; see "Methods" and Additional file 1) or are enzymes (*metin* is the number of metabolites catalyzed by a given enzyme; see Additional file 1), their encoding genes are likely to be druggable. This rule is partially supported by Bakheet and Doig [8] as they showed that druggable proteins are more likely to be enzymes than non-morbid ones. In respect to central position in a transcriptional regulatory circuitry, this is an issue that remains to be elucidated.

Conclusions

The identification of morbid and druggable genes has largely been an experimental effort mostly performed by time-consuming experiments. In an effort to accelerate the pace of discovery of such genes, we designed a machine learning-based computational approach that relies on network topological features, tissue expression profile and subcellular localization information for predicting morbid and druggable genes in human on a genome-wide scale.

We could demonstrate that our method is able to reliable predict morbid and druggable genes on a genome-wide scale as demonstrated by (i) the moderate to high performance measures achieved by the meta-classifiers (Table 1), (ii) the observation that the designed meta-classifiers learned traits actually related to morbidity and druggability instead of traits associated with any random sets of genes (Table 1) and (iii) the fact that known morbid and druggable genes tend to have high morbidity and druggability scores, respectively (Figures 1 and 2). Furthermore, in comparison with other available genome-wide prediction methods, the performance of our method proved to be equal or superior . We could also devise some cellular rules for gene morbidity and druggability using all network topological features, tissue expression profile and subcellular localization information as learning attributes for generation of decision trees (see details in section "Cellular rules for gene morbidity and druggability"). We discovered that number of regulating transcription factors, the central position in metabolic pathways, the localization of their encoded proteins in extracellular region and plasma membrane and tendency to form clusters with other proteins are important factors determining gene morbidity. In respect to druggability, the important factors determining druggability are plasma membrane localization, a central position in a transcriptional regulatory circuitry and being an enzyme. The fact that almost all discovered rules are supported by some additional evidences solidifies decision trees as useful tools for extracting knowledge from complex biological data. Albeit the good prediction performance and the ability to discover cellular rules for morbidity and druggability, our approach suffers from three limitations. First, it depends on existing Gene Ontology annotation and interaction data which are likely to be enriched in small-scale experiments involving morbid and druggable genes. Second, the construction of an integrated network of gene interactions requires a large amount of experimental interaction data that are currently available only to a limited number of human genes—our INHGI, for example, covers only ≈ 25% of already identified human genes. Third, the lack of negative examples to train the classifier forces us to consider all genes not known to be morbid or druggable as *de facto* non-morbid and non-druggable genes. We expect, however, that such limitations will be soon addressed as more systems-level data are generated.

Methods

Generation of the set of training features

Network topological features

In order to compute the network topological features used as training features for predicting morbid and druggable genes, we first constructed an integrated network of human gene interactions (INHGI) based on assumption that two genes, g_1 and g_2, coding respectively for proteins p_1 and p_2, are interacting genes if *(i)* p_1 and p_2 interact physically (protein physical interaction), *(ii)* the transcription factor p_1 directly regulates the transcription of gene g_2, i.e., p_1 binds to the promoter region of g_2 (transcriptional regulation interaction), or *(iii)* the enzymes p_1 and p_2 share metabolites, i.e., a product generated by a reaction catalyzed by enzyme p_1 is used as reactant by a reaction catalyzed by enzyme p_2 (metabolic interaction). Experimentally verified human protein physical interactions data were obtained from the following databases: the Biological General Repository for Interaction Datasets (BioGRID) database (release 2.0.47; [36]), the Database of Interacting Proteins (DIP; release Hsapi20081014; [37]), the Human Protein Reference Database (HPRD; release 7; [1]), IntAct (release 91; [38]), the Molecular Interactions Database (MINT; October 2008 release; [39]) and The Munich Information Center for Protein Sequences (MIPS) Mammalian Protein Interaction Database (MPPI; downloaded in December 2008; [40]. Experimentally verified human transcriptional regulation interactions were obtained from the Transcriptional Regulatory Element Database (TRED; [41]).

Experimentally verified human metabolic interactions were extracted from the human metabolic model Recon 1 [42] by a code implemented in Mathematica® 7.0 (Wolfram Research, Inc.). We excluded those metabolic interactions generated by the so-called "currency metabolites", abundant molecular species present throughout the cell most of the time and, therefore, unlikely to impose any constraints on the dynamics of metabolic reactions. Due to this feature of currency metabolites, the functionality of the network would be better represented without them [43]. We considered currency metabolites the eight most connected metabolites (ADP, ATP, H+, H_2O, NADP+, NADPH, orthophosphate and pyrophosphate) in the original metabolic model Recon 1.

The final INHGI is the result of integration of the protein physical, metabolic and transcriptional regulation interactions datasets through genes common to these datasets. Before performing the integration, we converted all human gene names to their GeneIDs—as provided by the Entrez Gene database [5]—to avoid the creation of false interactions due to gene name ambiguity.

For each gene g in INHGI, we computed 12 network topological features as listed in Additional file 1. Briefly, degree centrality is defined as the number of links to node (in our case, gene). We considered each type of interaction as a distinct measure of degree as described

in Additional file 1. Clustering coefficient (c) of a node (in our case, a gene) quantifies how close the node and its neighbors are to being a clique, i.e., all nodes connected to all nodes. For the INHGI, c is defined as the proportion of links between the genes within the neighborhood of g divided by the number of links that could possibly exist between them. Betweenness centrality reflects the role played by a node (in our case, a gene) in the global network architecture and, for the INHGI, is defined as the fraction of shortest paths between g_i and g_j passing through g. We computed the betweenness centrality based on shortest paths via all types of interaction (*inbet*) as well as based on shortest paths via each type of interaction (*inbetppi, inbetmet* and *inbetreg*). Closeness centrality (*cent*) measures how close a node (in our case, a gene) is to all others in the network and, for the INHGI, is defined as the mean shortest path between g and all other genes reachable from it. Identicalness is the number of genes with identical network topological characteristics.

All these network topological features, except for the betweenness centrality-related features, were calculated by a program written in a Mathematica® 7.0 notebook. The betweenness centrality-related features were calculated by the Python package *NetworkX* 0.99 [44].

Subcellular localization of human genes

We determined the subcellular localization of proteins encoded by the genes in the INHGI by using the QuickGO tool, a Gene Ontology (GO) browser associated with the integrated database resource for protein families (InterProt) at the European Bioinformatics Institute [45]. We selected GO slim terms—subsets of GO terms consisting of a limited number of high-level GO terms that cover some or all of the content of GO—related to cellular components provided by QuickGO to annotate genes in the INHGI. Genes were annotated to the following slim terms:"cytoplasm", "endoplasmic reticulum", "mitochondrion", "nucleus", "extracellular space", "Golgi apparatus", "plasma membrane" and "cellular component". Genes annotated to other slim terms were reannotated to one of these terms or to a new term named "other localization" and genes with no GO cellular component slim term annotation was annotated to the term "unknown".

Tissue expression profile of human genes

We retrieved the tissue expression profiles of genes in the INHGI from the study performed by Reverter and colleagues [33]. In their study, Reverter and colleagues mined three large datasets comprising expression data obtained from massively parallel signature sequencing across 32 tissues in order to classify genes as housekeeping or tissue-specific genes and then relate this tissue specificity with gene interactions and disease states. According to Reverter and colleagues, tissue expression profile of a given gene is (i) the number of tissues (out of 32) in which the gene is expressed at least 5 transcripts per million (tpm) and (ii) the average expression in tpm among all the tissues in which the gene is expressed [33].

Classifier design and evaluation

Construction of training datasets

For evaluating the performance of the chosen training features–network topological features, subcellular localization and tissue expression profile–in predicting morbid and druggable genes, we constructed four different groups of balanced training datasets, i.e., datasets containing the same number of positive (in our case, morbid or druggable genes) and negative (in our case, non-morbid or non-druggable genes) examples: (1) "normal morbidity datasets", (2) "shuffled morbidity datasets", (3) "normal druggability datasets" and (4) "shuffled druggability datasets".

For the construction of the morbidity datasets, we first gathered a list of "morbid genes"—genes whose mutations cause hereditary diseases—from the morbid map table in the Online Mendelian Inheritance in Man (OMIM) [46] and then mapped them to the INHGI. The final list of morbid genes used as positive examples to train our classifier is comprised by 1,412 morbid genes present in the INHGI. Regarding the negative examples, we considered as "non-morbid genes" the remaining genes present in the INHGI; this was done since building a list of genes not known to be involved in hereditary diseases is impossible currently. We randomly selected 10 different sets of 1,412 of these non-morbid genes and combine them with the list of morbid genes to build 10 different training datasets which we call "normal morbidity datasets". From these normal morbidity datasets, we generate 10 different "shuffled morbidity datasets" by randomly shuffling the class labels (morbid and non-morbid) among genes.

For the construction of the druggability dataset, we first built a list of "druggable genes"—genes coding for proteins whose modulation by small molecules elicits phenotypic effects—from the drug-target network constructed by Yildirim and colleagues [47] and then mapped them to the INHGI. The final list of druggable genes used as positive examples to train our classifier is comprised by 257 druggable genes present in the INHGI. Regarding the negative examples, we considered as "non-druggable genes" the remaining genes present in the INHGI; this was done since, similar to non-morbid genes, it is also impossible to construct a list of genes coding for proteins whose modulation by small molecules do not elicits phenotypic effects. We randomly selected 10 different sets of 257 of these non-druggable genes and combine them with the list of

druggable genes to build 10 different training datasets which we call "normal druggability datasets". From these normal druggability datasets, we generate 10 different "shuffled druggability datasets" by randomly shuffling the class labels (druggable and non-druggable) among genes. We also constructed 25 additional morbidity and 25 additional druggability datasets lacking one of the 25 features used as training attributes. We call these datasets as "without-one-feature" datasets, where *one* can be replaced by the name of feature.

Classifier design

Using WEKA (Waikato Environment for Knowledge Analysis) software package, a collection of machine learning algorithms for data mining tasks [48], we designed the classifier used for predicting morbid and druggable genes in the INHGI. This classifier is an ensemble of seven decision tree algorithms using the meta-classifier "Vote", a WEKA's implementation of the voting algorithm that combines the output predictions of each classifier by different rules [49]. We combined the classifiers by the average rule, where the output predictions computed by the individual classifiers for each class are averaged and this average is used in its decision [49]. The classifiers composing our model were: (1) REPtree [48], (2) random tree [48], (3) random forest [50], (4) J48, a WEKA's implementation of the C4.5 decision tree [32], with minimum number of 32 instances per leaf, (5) best-first decision tree with minimum number of 32 instances at the terminal nodes [51], (6) logistic model tree [52] and (7) alternating decision tree with 25 boost iterations [53]. In addition, we applied the bootstrap aggregating (bagging) approach [54] to each classifier. Parameters values for each classifier are provided in the Additional file 3.

Classifier evaluation

We assessed the performance of our classifier by estimating the following measures: recall, precision and area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Recall is the proportion of actual morbid or druggable genes which are correctly predicted as such against all actual morbid or druggable genes:

$$Recall = \frac{TP}{TP + FN}$$

TP (true positive) denotes the amount of actual morbid or druggable genes correctly predicted as such and FN (false negative) denotes the amount of actual morbid or druggable genes incorrectly predicted as non-morbid or non-druggable, respectively.

Precision is the proportion of actual morbid or druggable genes which are correctly predicted as such against all genes predicted as morbid or druggable:

$$Precision = \frac{TP}{TP + FP}$$

FP denotes the amount of actual non-morbid or non-druggable genes incorrectly predicted as morbid or druggable, respectively.

The AUC is a widely used summary measure of the ROC curve–a plot of the true positive rate versus false positive rate that indicates the probability of a true positive prediction as a function of the probability of a false positive prediction for all possible threshold values [55]–and is equivalent to the probability that a randomly chosen negative example (in our case, a non-morbid or non-druggable gene) will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example (in our case, a morbid or druggable gene) [56].

We estimated the above-mentioned performance measures by performing a 10-fold cross-validation test–using WEKA–on the 10 normal and 10 shuffled morbidity datasets and on the 10 normal and 10 shuffled druggability datasets constructed as described in the section "Construction of training datasets". During the 10-fold cross-validation test process, each dataset is randomly partitioned into 10 subsets. Of the 10 subsets, a single subset is retained as the validation data for testing the model, and the remaining 9 subsets are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsets used exactly once as the validation data. The 10 results from the folds are then averaged to produce a single estimation for each performance measure for each dataset. We reported the performance measures estimated by the 10-fold cross-validation as medians of the 10 datasets for each category (normal morbidity, shuffled morbidity, normal morbidity and shuffled morbidity).

The statistical comparisons of (i) the performance measures estimated by our classifier trained on normal and shuffled datasets, (ii) the AUC values estimated by our classifier trained on normal datasets and without-one-feature datasets, and (iii) the normal and shuffled morbidity and druggability scores for each gene in INHGI were performed by the Wilcoxon signed-rank test [6]. Following established conventions in the machine learning community, we used this test since it makes minimal assumptions about the underlying distribution of performance measures used to evaluate classifiers [57]. The differences were statistically significant if the obtained Wilcoxon's test statistic value (W) was equal to or smaller than the critical Wilcoxon's test statistic value (W_c) for a given sample size (N) at the two-tailed significance level of 0.05 (p = 0.05) according to the table of critical values for the Wilcoxon test [6].

Prediction of novel morbid and druggable genes

The "normal morbidity scores" and the "normal druggability scores" were generated by applying the models constructed by training our meta-classifier on the normal datasets to the entire set of genes in INHGI where the class labels were removed. These scores are the probability values of classifying each gene as morbid or druggable as returned by the models. The final normal morbidity and druggability scores are median scores of 10 scores. We also obtained "shuffled morbidity scores" and "shuffled druggability scores" that were generated by models trained on the shuffled datasets.

Determination of rules for gene morbidity and druggability

The determination of rules for gene morbidity and druggability was performed by analyzing the best representative decision tree for each category among the 10 decision trees generated through the training of J48 algorithm [32] on the 10 normal morbidity and 10 normal druggability datasets. The parameters values for producing decision trees by J48 algorithm training are provided in the Additional file 3.

Additional material

Additional file 1: Network topological features Description: This file includes a table showing the functions and descriptions of the 12 network topological features used as learning attributes for training the classifier algorithm

Additional file 2: Morbidity and druggability scores of genes in INHGI Description: Tab-limited text file containing all genes (Entrez GeneIDs) in the INHGI with their morbidity and druggability scores.

Additional file 3: Parameters used to train the meta-classifier and J48 Description: File containing all parameters values used to train the meta-classifier for prediction of morbid and druggable genes and all parameters values used to train the J48 algorithm to generate decision trees for discovery of cellular rules for morbidity and druggability.

Competing interests statement

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank FAPESP (The State of Sao Paulo Research Foundation) for the financial support through the FAPESP research grants 2007/02827-9, 2007/01213-7 and 2007/08466-8. This research was supported by resources supplied by the Center for Scientific Computing (NCC/GridUNESP) of the Univ Estadual Paulista (UNESP).
This article has been published as part of *BMC Genomics* Volume 11 Supplement 5, 2010: Proceedings of the 5th International Conference of the Brazilian Association for Bioinformatics and Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/11?issue=S5.

Authors contributions

PRC obtained the tissue expression profile and gene ontology data, analyzed the meta-classifiers' performances, implemented the program for calculation of network topological features and drafted the manuscript. MLA obtained all interaction data, constructed the network, designed the meta-classifier, pursued the biological interpretation of results and drafted the manuscript. NL conceived, designed and directed the project. All authors read and approved the final manuscript.

Published: 22 December 2010

References

1. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A: **Human Protein Reference Database–2009 update.** *Nucleic Acids Res* 2009, **37(Database issue)**:D767-72.
2. Lindsay MA: **Target discovery.** *Nat Rev Drug Discov* 2003, **2(10)**:831-8.
3. da Silva JPM, Acencio ML, Mombachb JCM, Vieirac R, da Silva J, Lemke N, Sinigagliac M: **In silico network topology-based prediction of gene essentiality.** *Physica A* 2008, **387**:1049-1055.
4. Acencio ML, Lemke N: **Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information.** *BMC Bioinformatics* 2009, **10**:290.
5. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2007, **35**:D26-D31.
6. Wilcoxon F: **Probability tables for individual comparisons by ranking methods.** *Biometrics* 1947, **3(3)**:119-22.
7. Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C: **Estimating the size of the human interactome.** *Proc Natl Acad Sci U S A* 2008, **105(19)**:6959-64.
8. Bakheet TM, Doig AJ: **Properties and identification of human protein drug targets.** *Bioinformatics* 2009, **25(4)**:451-7.
9. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene discovery by sequence based candidate prioritization.** *BMC Bioinformatics* 2005, **6**:55.
10. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited diseases using data mining.** *Nat Genet* 2002, **31(3)**:316-9.
11. Turner FS, Clutterbuck DR, Semple CAM: **POCUS: mining genomic sequence annotation to predict disease genes.** *Genome Biol* 2003, **4(11)**: R75.
12. Van Driel MA, Cuelenaere K, Kemmeren PPCW, Leunissen JAM, Brunner HG: **A new web-based data mining tool for the identification of candidate genes for human genetic disorders.** *Eur J Hum Genet* 2003, **11**:57-63.
13. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De-Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y: **Gene prioritization through genomic data fusion.** *Nat Biotechnol* 2006, **24(5)**:537-44.
14. Wu X, Jiang R, Zhang MQ, Li S: **Network-based global inference of human disease genes.** *Mol Syst Biol* 2008, **4**:189.
15. Chen J, Bardes EE, Aronow BJ, Jegga AG: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.** *Nucleic Acids Res* 2009, **37(Web Server issue)**:W305-11.
16. Xu J, Li Y: **Discovering disease-genes by topological features in human protein-protein interaction network.** *Bioinformatics* 2006, **22(22)**:2800-5.
17. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world.** *Proc Natl Acad Sci U S A* 2003, **100(8)**:4372-6.
18. Sugaya N, Ikeda K: **Assessing the druggability of protein-protein interactions by a supervised machine-learning method.** *BMC Bioinformatics* 2009, **10**:263.
19. Lin BK, Clyne M, Walsh M, Gomez O, Yu W, Gwinn M, Khoury MJ: **Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database.** *Am J Epidemiol* 2006, **164**:1-4.
20. Finckh U, van Hadeln K, Müller-Thomsen T, Alberici A, Binetti G, Hock C, Nitsch RM, Stoppe G, Reiss J, Gal A: **Association of late-onset Alzheimer disease with a genotype of PLAU, the gene encoding urokinase-type plasminogen activator on chromosome 10q22.2.** *Neurogenetics* 2003, **4(4)**:213-7.
21. Gondi CS, Rao JS: **Therapeutic potential of siRNA-mediated targeting of urokinase plasminogen activator, its receptor, and matrix metalloproteinases.** *Methods Mol Biol* 2009, **487**:267-81.
22. van Zelm MC, Reisli I, van der Burg M, Castaño D, van Noesel CJM, van Tol MJD, Woellner C, Grimbacher B, Patiño PJ, van Dongen JJM, Franco JL:

An antibody-deficiency syndrome due to mutations in the CD19 gene. *N Engl J Med* 2006, **354(18)**:1901-12.
23. Al-Katib AM, Aboukameel A, Mohammad R, Bissery MC, Zuany-Amorim C: **Superior antitumor activity of SAR3419 to rituximab in xenograft models for non-Hodgkin's lymphoma.** *Clin Cancer Res* 2009, **15(12)**:4038-45.
24. Yang W, Zhao M, Hirankarn N, Lau CS, Mok CC, Chan TM, Wong RWS, Lee KW, Mok MY, Wong SN, Avihingsanon Y, Lin IO, Lee TL, Ho MHK, Lee PPW, Wong WHS, Sham PC, Lau YL: **ITGAM is associated with disease susceptibility and renal nephritis of systemic lupus erythematosus in Hong Kong Chinese and Thai.** *Hum Mol Genet* 2009, **18(11)**:2063-70.
25. Romano J, Beni-Adani L, Nissenbaum OL, Brenneman DE, Shohami E, Gozes I: **A single administration of the peptide NAP induces long-term protective changes against the consequences of head injury: gene Atlas array analysis.** *J Mol Neurosci* 2002, **18(1-2)**:37-45.
26. Sawczenko A, Azooz O, Paraszczuk J, Idestrom M, Croft NM, Savage MO, Ballinger AB, Sanderson IR: **Intestinal inflammation-induced growth retardation acts through IL-6 in rats and depends on the -174 IL-6 G/C polymorphism in children.** *Proc Natl Acad Sci U S A* 2005, **102(37)**:13260-5.
27. Rose-John S, Waetzig GH, Scheller J, GrÖtzinger J, Seegert D: **The IL-6/sIL-6R complex as a novel target for therapeutic approaches.** *Expert Opin Ther Targets* 2007, **11(5)**:613-24.
28. Koch W, Hoppmann P, de Waha A, SchÖmig A, Kastrati A: **Polymorphisms in thrombospondin genes and myocardial infarction: a case-control study and a meta-analysis of available evidence.** *Hum Mol Genet* 2008, **17(8)**:1120-6.
29. Daniel C, Schaub K, Amann K, Lawler J, Hugo C: **Thrombospondin-1 is an endogenous activator of TGF-beta in experimental diabetic nephropathy in vivo.** *Diabetes* 2007, **56(12)**:2982-9.
30. Castaldi PJ, Cho MH, Cohn M, Langerman F, Moran S, Tarragona N, Moukhachen H, Venugopal R, Hasimja D, Kao E, Wallace B, Hersh CP, Bagade S, Bertram L, Silverman EK, Trikalinos TA: **The COPD genetic association compendium: a comprehensive online database of COPD genetic associations.** *Hum Mol Genet* 2010, **19(3)**:526-34.
31. Kingsford C, Salzberg SL: **What are decision trees?** *Nat Biotechnol* 2008, **26(9)**:1011-1013.
32. Quinlan JR: **C4.5: programs for machine learning.** San Francisco: Morgan Kaufmann; 1993.
33. 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 Min.* 2008, **1**:8.
34. Jimenez-Sanchez G, Childs B, Valle D: **Human disease genes.** *Nature* 2001, **409(6822)**:853-5.
35. Winter EE, Goodstadt L, Ponting CP: **Elevated rates of protein secretion, evolution, and disease among tissue-specific genes.** *Genome Res* 2004, **14**:54-61.
36. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, Dolinski K, Tyers M: **The BioGRID Interaction Database: 2008 update.** *Nucleic Acids Res* 2008, **36(Database issue)**:D637-40.
37. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32(Database issue)**:D449-51.
38. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R: **IntAct: an open source molecular interaction database.** *Nucleic Acids Research* 2004, **32**:D452-D455.
39. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database.** *Nucleic Acids Res.* 2007, **35**:D572-D574.
40. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes HW, Ruepp A, Frishman D: **The MIPS mammalian protein–protein interaction database.** *Bioinformatics* 2005, **21**:832-834.
41. Jiang C, Xuan Z, Zhao F, Zhang MQ: **TRED: a transcriptional regulatory element database, new entries and other development.** *Nucleic Acids Res.* 2007, **35**:D137-D140.
42. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson B: **Global reconstruction of the human metabolic network based on genomic and bibliomic data.** *PNAS* 2007, **104**:1777-1782.
43. Huss M, Holme P: **Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks.** *IET Syst Biol* 2007, **1(5)**:280-285.
44. **NetworkX package.** [https://networkx.lanl.gov].
45. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R: **QuickGO: a web-based tool for Gene Ontology searching.** *Bioinformatics* 2009, **25(22)**:3045-6.
46. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM.** *Am J Hum Genet* 2007, **80(4)**:588-604.
47. Yildirim MA, Goh KI, Cusick ME, Barabási AL, Vidal M: **Drug-target network.** *Nat Biotechnol* 2007, **25(10)**:1119-26.
48. Witten IH, Frank E: **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.** San Francisco: Morgan Kaufmann; 2000.
49. Kittler J, Hatef M, Duin RP, Matas J: **On Combining Classifiers.** *IEEE Trans Pattern Anal Mach Intell.* 1998, **20(3)**:226-239.
50. Breiman L: **Random forests.** *Mach Learn* 2001, **45**:5-32.
51. Shi H: **Best-first Decision Tree Learning.** *Master Thesis* The University of Waikato; 2007.
52. Landwehr N, Hall M, Frank E: **Logistic Model Trees.** *Mach Learn* 2005, **95(1-2)**:161-205.
53. Freund Y, Mason L: **The alternating decision tree learning algorithm.** *Proceedings of the Sixteenth International Conference on Machine Learning* San Francisco: Morgan Kaufmann; 1999, 124-133.
54. Breiman L: **Bagging predictors.** *Mach Learn* 1996, **24(2)**:123.
55. Huang J, Ling CX: **Using AUC and Accuracy in Evaluating Learning Algorithms.** *IEEE Trans. on Knowl. and Data Eng* 2005, **17(3)**:299-310.
56. Hand DJ, Till RJ: **A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems.** *Mach. Learn* 2001, **45(2)**:171-186.
57. Demšar J: **Statistical Comparisons of Classifiers over Multiple Data Sets.** *J. Mach. Learn. Res.* 2006, **7**:1-30.

doi:10.1186/1471-2164-11-S5-S9
Cite this article as: Costa *et al.*: A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC Genomics* 2010 **11**(Suppl 5):S9.