UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Professor Doutor Ney Lemke

Construção, caracterização global e local de redes biológicas

Botucatu – SP

2011

Prof. Dr. Ney Lemke

Construção, caracterização global e local de redes biológicas

Tese apresentada ao Instituto de Biociências de Botucatu da Universidade Estadual Paulista "Júlio de Mesquita Filho" para obtenção de título de Livre-Doutor em Física Computacional

Botucatu – SP

2011

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. E TRAT. DA INFORMAÇÃO
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: *Rosemeire Aparecida Vicente*

Lemke, Ney.
Construção, caracterização global e local de redes biológicas / Ney Lemke. – Botucatu : [s.n], 2011

Tese (livre-docência) - Universidade Estadual Paulista, Instituto de Biociências
Capes: 10507000

1. Física. 2. Matéria condensada. 3. Sistemas biológicos.

Palavras-chave: Biologia sistêmica; Redes complexas; Sistemas complexos.

A todas mulheres da minha vida ...

Agradecimentos

À minha família pelo suporte em todos os momentos.

Aos professores Rita Maria Cunha de Almeida e José Roberto Iglesias por terem me convencido a trabalhar com Mecânica Estatística.

A José Carlos Merino Mombach pelos anos de parceria e por ter feito a diferença nos momentos críticos.

A Jeferson Arenzon pela cumplicidade com o modelo das minhocas canibais, que felizmente nunca foi acabado.

To Ian Campbell for showing me the power of kindness.

A Sergio Novaes pelo incentivo e pela confiança.

A todos os meus colegas e professores do Instituto de Física.

A todos os meus amigos da Unisinos.

A todos os meus alunos pela coragem.

A meus orientados e orientandos pela perseverança e empenho. Não percam as esperanças vou seguir me esforçando para eventualmente sugerir algo que dê certo.

A Joel Mesa e a Roberto Morato pela triangulação.

A Marcos Fontes, José Ricardo Miranda e José Roberto Saglietti pelos ensinamentos sobre a Unesp e o IB.

A Sílvia Helena Ramos pela competência.

A Murilo Stelzer pelos IPs.

Aos professores e funcionários do Instituto de Biociências pela colaboração.

A Fapesp, Capes, CNPq, Fapergs e HP por terem financiado a realização dos trabalhos descritos nesta tese.

Finalmente a Douglas Adams, por me ensinar que o importante são as perguntas e que a regra número 1 em qualquer situação é “Don’t Panic”.

"Au départ l'art du puzzle semble un art bref, un art mince, tout entier contenu dans un maigre enseignement de la Gestalttheorie : l'objet visé - qu'il s'agisse d'un acte perceptif, d'un apprentissage, d'un systeme physiologique ou, dans le cas qui nos occupe, d'un puzzle de bois - n'est pas une somme d'éléments qu'il faudrait d'abord isoler et analyser, mais un ensemble, c'est-à-dire une forme, une structure: l'élément ne préexiste pas à l'ensemble, il n'est ni plus immédiat ni plus ancien, ce ne sont pas les éléments qui déterminent l'ensemble, mais l'ensemble qui détermine les éléments : la connaissance du tout et de ses lois, de l'ensemble et de sa structure, ne saurait être déduite de la connaissance séparé des parties qui le composent (...)" [1]

Georges Perec ("la vie mode démploi" ,1978)

[1]Inicialmente a arte dos quebra-cabeças parece uma arte breve, uma arte menor, inteiramente contida em um ensinamento básico da *Gestlatterapie*: o objeto em questão - seja ele um ato de percepção, uma aprendizagem, um sistema psicológico, ou no caso de que nos ocupamos, de um quebra-cabeça em madeira - não é a soma dos elementos que devemos no início isolar e analisar, mas um conjunto, por assim dizer uma forma, uma estrutura: o elemento não precede o conjunto, ele não é nem mais imediato nem mais antigo, não são os elementos que determinam o conjunto, mas o conjunto que determina os elementos: o conhecimento do todo e de suas leis, do conjunto e de sua estrtura, não será deduzida do conhecimento separado das partes que o compõe (...) Tradução livre do autor

Resumo

A Biologia Sistêmica visa a compreensão da vida através de modelos integrativos que enfatizem as interações entre os diferentes agentes biológicos. O objetivo é buscar por leis universais, não nas partes componentes dos sistemas mas sim nos padrões de interação dos elementos constituintes. As redes complexas biológicas são uma poderosa abstração matemática que permite a representação de grandes volumes de dados e a posterior formulação de hipóteses biológicas. Nesta tese apresentamos as redes biológicas integradas que incluem interações oriundas do metabolismo, interação física de proteínas e regulação. Discutimos sua construção e ferramentas para sua análise global e local. Apresentamos também resultados do uso de ferramentas de aprendizado de máquina que nos permitem compreender a relação entre propriedades topológicas e a essencialidade gênica e a previsão de genes mórbidos e alvos para drogas em humanos.

Abstract

Systems Biology aims to understand life process through integrative models that emphasize the interactions among biological agents. The goal is to search for universal laws, not in the description of systems parts, but on interactions patterns among systems elements. Biological networks are a powerful mathematical abstraction that allows the representation of large experimental datasets and the formulation of biological hypothesis. In this thesis we investigate biological networks that assemble information of: metabolism, protein interaction and regulation. We discuss the network construction methodology and mathematical tools to describe its local and global properties. We also present results, obtained through machine learning tools, expressing the implication of topological properties on gene essentiality, morbidity, and drugability on human genes.

Sumário

Lista de Figuras

Lista de Tabelas

Introdução p. 15

1 Construção de Redes Biológicas p. 21

1.1 Grafos . . . . . p. 21

1.2 Redes Biológicas . . . . . p. 23

1.2.1 Redes Biológicas Integradas . . . . . p. 24

1.3 Construção de redes biológicas . . . . . p. 26

1.4 Ontologias . . . . . p. 27

1.5 Rede Integrada do *H. sapiens* . . . . . p. 30

1.6 Sub-redes . . . . . p. 31

1.6.1 Construção da G_{ccam} . . . . . p. 31

2 Caracterização Global de Redes Biológicas p. 36

2.1 Medidas . . . . . p. 39

2.2 Redes Integradas da *E. coli* e *S. cerevisiae* . . . . . p. 44

2.2.1 Rede integrada da *E. coli* . . . . . p. 45

2.2.2 Rede integrada da *S. cerevisiae* . . . . . p. 48

2.2.3 Análise comparativa do grau de proximidade e do grau de intermediação para as redes integradas da *S. cerevisiae* e da *E. coli* . p. 50

2.3 *Motifs* . . . . . p. 53

2.4 Comunidades . . . . . p. 58
2.5 Caracterização da Rede Integrada de Humanos . . . . . p. 66
2.5.1 Características gerais da G_{ccam} . . . . . p. 67

3 Caracterização Local de Redes Biológicas p. 71
3.1 Dano e Essencialidade . . . . . p. 71
3.1.1 Dano em *E. coli* . . . . . p. 72
3.1.2 Dano em S. cerevisiae . . . . . p. 73
3.2 Aprendizado de Máquina . . . . . p. 78
3.2.1 Aprendizado de Máquina em *E. coli* . . . . . p. 78
3.3 Essencialidade na *S. cerevisiae* . . . . . p. 81
3.4 Morbidade e Drogabilidade . . . . . p. 85
3.5 Predição de alvos para drogas na G_{ccam} . . . . . p. 88

4 Perspectivas p. 91

Referências p. 93

Apêndice A – Fonte dos dados p. 102

Apêndice B – Mineração de Dados p. 103
B.0.1 Estatísticas Utilizadas na Mineração de Dados . . . . . p. 104
B.0.1.1 Valores de Desempenho do Classificador . . . . . p. 104
B.0.1.2 Processo de Validação Cruzada . . . . . p. 104
B.0.1.3 Índice Kappa – κ . . . . . p. 105
B.0.1.4 Teste de Wilcoxon – W . . . . . p. 105

Trabalhos do Autor p. 107

Lista de Figuras

1 Representação esquemática da integração das diferentes ômicas . Extraído de (JOYCE; PALSSON, 2006). . . . . p. 18

2 As pontes de Königsberg em três representações. Fonte: Wikipedia . . p. 22

3 A árvore da vida de Darwin. . . . . p. 23

4 Esquema representando os três tipos de interação que compõe a rede integrada. . . . . p. 25

5 Rede biológica integrada para a *E. coli*. Em vermelho representamos os genes essenciais. . . . . p. 26

6 Os principais conceitos de Ontologia MONET que integram as redes metabólicas, regulatórias e de interação fisica proteína-proteína. . . . . p. 29

7 Parte da rede das relações hierárquicas entre termos da categoria *biological process* (processo biológico) do *Gene Ontology* . . . . . p. 32

8 Esquema de funcionamento do algoritmo *busca_cg* . . . . . p. 34

9 Dígrafo que representa a rede metabólica, os círculos grandes representam metabólitos e os pequenos as reações. Em negro representamos os vértices que são excluídos quando uma reação deixa de ocorrer. . . . . p. 41

10 Modelos de redes.(Aa) Rede aleatória. (Ab) Distribuição de Poisson das conectividades de uma rede aleatória. (Ac) Coeficiente de agrupamento $C(k)$ independe de k em uma rede aleatória. (Ba) Rede com distribuição de conectividades obedecendo lei de potência. (Bb) A distribuição de conectividades representada em um gráfico Log-Log é linear. (Ac) Coeficiente de agrupamento $C(k)$ independe da conectividade k em uma rede com distribuição de conectividades obedecendo uma lei de potência. (Ca) Rede hierárquica: possui, simultaneamente, distribuição de conectividades obedecendo uma lei de potência e $C(k)$ dependente de k. (Cb) A distribuição de conectividades representada em um gráfico Log-Log é linear. (Cc) $C(k)$ tem dependência em k. Extraído de (BARABASI; OLTVAI, 2004). . . . . p. 43

11 Distribuição de conectividades da rede integrada de interações moleculares entre genes da *E. coli* construída neste projeto. Essa distribuição segue uma lei de potência, o que a caracteriza como uma rede livre de escala. . . . . p. 45

12 Dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k. Pode-se observar que o $C(k)$ é ajustado por um modelo linear quadrático, o que indica que a rede integrada de interações moleculares entre genes da *E. coli* construída neste projeto é linear quadrática. . . . p. 46

13 Distribuição de conectividades da rede integrada de interações moleculares entre genes da *S. cerevisiae* construída neste projeto. Essa distribuição é ajustada por um modelo livre de escala diferenciado. . . . . p. 48

14 Dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k. Pode-se observar que o $C(k)$ é ajustado por um modelo linear quadrático, o que indica que a rede integrada de interações moleculares entre genes da *S. cerevisiae* construída neste projeto é hierárquica, mas com uma estrutura diferente das redes contendo apenas um tipo de interação. . . . . p. 49

15 Grau de proximidade dos vértices em função de k para as redes integradas da *E. coli* e da *S. cerevisiae*. Ambos organismos tiveram suas medidas ajustadas por uma lei de potência. . . . . p. 50

16 Grau de intermediação dos vértices em função de k para as redes integradas da *E. coli* e da *S. cerevisiae*. O ajuste utilizado para ambos organismos foi feito através de uma lei de potência. . . . . p. 51

17 Grafico de $P(k)$ em relação k para *E. coli* e *S. cerevisiae* e de 100 redes aleatórias (usando o método I) da mesma. Nestes gráficos é possível observar que o $P(k)$ das redes experimentais e das Redes aleatórias possuem o mesmo comportamento. . . . . p. 55

18 Grafico de $C(k)$ em relação a k da *E. coli* e de 100 Redes Aleatórias (Método I) da mesma. Observa-se que a dependência de $C(k)$ em relação ao grau de conectividade k da *E. coli* fica reduzido após sua rede ser aleatorizada. . . . . p. 56

19 Representação da entropia para as comunidades encontradas pelo método proposto por *Clauset* para a rede da bactéria *Escherichia coli*. . . . . p. 62

20 Representação da entropia para as comunidades encontradas pelo método *Clique Percolation* para a rede da bactéria *Escherichia coli*. . . . . p. 63

21 Representação da entropia para as comunidades encontradas pelo método *MinCut* para a rede da levedura *Escherichia coli*. . . . . p. 63

22 Representação da entropia para as comunidades encontradas pelo método proposto por *Clauset* para a rede da levedura *Saccharomyces cerevisiae*. p. 63

23 Representação da entropia para as comunidades encontradas pelo método *Clique Percolation* para a rede da levedura *Saccharomyces cerevisiae*. . p. 64

24 Representação da entropia para as comunidades encontradas pelo método *MinCut* para a rede da levedura *Saccharomyces cerevisiae*. . . . . p. 64

25 Representação da entropia para os agregados identificados, por coexpressão, na rede da levedura *Saccharomyces cerevisiae*. . . . . p. 65

26 Histograma dos tamanhos das comunidades, obtidas por diversos métodos, para a rede da levedura *Saccharomyces cerevisiae*. . . . . p. 65

27 Distribuições dos graus de conectividade da $RIGH$ e da G_{ccam} . . . . . p. 68

28 Distribuições dos coeficientes de agrupamento médios, $C(k)$, em relação à conectividade k . . . . . p. 69

29 Relação entre dano e fração de genes essenciais. A reta representa o ajuste linear. . . . . p. 72

30 A função L_d e L_k contra a classificação normalizada para os dados Uetz sem enzimas. A curvatura indicam uma muito fraca correlação entre os dois conectividade e os danos letalidade. . . . . p. 76

31 Árvore de decisão gerada pela aplicação do algoritmo J48 para classificação de genes essenciasi em ıE. coli com *F-measure* de 83.4% para genes essenciais. . . . . p. 80

32 Árvore de decisão foi gerado pelo treinamento do algoritmo J48 na conjunto de dados balanceado com todos os dados disponíveis. A elipse superior é o nó raiz da árvore que representa a condição mais importante para discriminar genes essenciais de genes não-essenciais. Neste caso, tal condição é o número de interações proteína física (*ppi*). As elipses remanescentes são nós internos que representam condições suplementares para um gene ser considerado como essencial ou não essencial. No ramo esquerdo de árvore, tais condições são o envolvimento em um processo metabólico (*met-proc*) e localização nuclear (*nucleus*). No ramo direito, tais condições são a localização nuclear (*nucleus*) e o número de fatores de transcrição que regulam o gene (*regin*). Os retângulos são os nós folha que representam a classificação final. Retângulos vermelho e verde retratam genes que, sob certas condições (representado pelo nó raiz e nós internos nós), são classificados como essenciais (**E**) e não-essenciais (**N**). No parênteses dentro dos rectângulos, o número antes da barra indica a quantidade de genes que são realmente essenciais ou não essenciais e os número depois da barra indica quantos genes foram corretamente previstos. . . . . p. 84

33 Árvore de decisão para drogabilidade . . . . . p. 86

34 Árvore de decisão para morbidade . . . . . p. 87

35 Genes conhecidamente e potencialmente drogáveis na sub-rede $EGFR-CDC6$ . . . . . p. 89

Lista de Tabelas

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. 35

2 Resultados para a rede integrada da bactéria *E. coli* . . . . . p. 52

3 Resultados para a rede integrada da levedura *S. cerevisiae* . . . . . p. 53

4 Legenda das interações . . . . . p. 54

5 Porcentagem dos tipos de interações da rede experimental da *E. coli* e comparação com os resultados obtidos para redes aleatorizadas usando os Métodos I e II. . . . . p. 57

6 Porcentagem dos tipos de interações da *S. cerevisiae* e comparação com os resultados obtidos para redes aleatorizadas usando os Métodos I e II. p. 57

7 Tabela comparativa para os métodos de divisão de comunidades utilizados. p. 61

8 Quantia de comunidades obtidas por cada método de divisão de comunidades utilizado. . . . . p. 61

9 Comparação dos dados de diferentes conjuntos de proteína-proteína interações, S mede a correlação entre um topológica parâmetro com a letalidade, para um parâmetro não correlacionadas $S = 0{,}5$, neste cenário valores maiores de S indicam forte correlação, ver texto para detalhes. p. 77

10 Tabela com os resultados obtidos durante o treinamento do modelo. . . p. 87

11 Lista dos bancos de dados biológicos utilizados para o processo de aquisição dos dados. . . . . p. 102

12 Valores de W_c para o teste de Wilcoxon . . . . . p. 106

Introdução

Não se trata de nenhuma novidade a percepção de que abordagem reducionista traz em sua própria essência limitações sérias que são um entrave a compreensão de muitos dos mais maravilhosos objetos que povoam nosso universo. O caráter inovador da abordagem sistêmica dos nossos dias está em conseguir propor abstrações que consigam ser genéricas e que possuam capacidade preditiva, ou seja gerem modelos e conceitos que iluminem resultados experimentais e nos permitam perceber problemas antigos com uma nova perspectiva.

Mas se essa preocupação é antiga por que essas teorias não foram desenvolvidas antes, por que somente hoje estamos conseguindo fazer avançar essas metodologias? A razão para isso possui quatro vertentes tecnológicas: a existência de sistemas computacionais baratos, o desenvolvimento de linguagens de programação simples e poderosas, o desenvolvimento de sensores baratos e ubíquos e finalmente uma forma de disponibilizar informação praticamente universal.

Esta consonância de fatores muitas vezes é pejorativamente denotada por inundação de dados, como sendo algo negativo. De fato desde a nossa perspectiva de membros da geração 1.0, com nossos sistemas cognitivos treinados em sistemas não automatizados de busca e organização da informação, o estado atual de coisas pode ser assustador. Mas para a geração 3.0 que vai se alfabetizar com o teclado de celulares, que vai interagir de forma totalmente natural usando as redes sociais que estamos moldando neste início de século, a infovia será algo tão natural como os produtos tecnológicos do século XX, são para a geração 1.0.

A Biologia Sistêmica foi proposta inicialmente por Ludwig von Bertalanffy (BERTALANFFY, 1968) como parte de um movimento filosófico e científico que visava estudar a natureza desde um ponto de vista sistêmico ou holístico. Este movimento foi chamada de Teoria Geral de Sistemas e incluía todas as áreas do conhecimento humano.

Existem duas possíveis traduções para o termo em inglês *System Biology*, "Biologia de Sistemas" e "Biologia Sistêmica". Eu opto pela segunda para enfatizar a diferença metodológica implicada por essa abordagem. Não se trata de investigar "sistemas", mas

sim estudar de forma "sistêmica" a Biologia. O objetivo é buscar por leis universais, não nas partes componentes dos sistemas mas sim nos padrões de interação dos elementos constituintes.

Edgar Morin na série de livros "O Método" explora exaustivamente esse viés epistemológico, enfatizando que só é possível compreender a natureza através de modelos integrativos que considerem inclusive os aspectos sociológicos da prática científica (MORIN, 2003).

Uma de suas aplicações iniciais foi o estudo do metabolismo e das curvas de crescimento de diferentes seres vivos. Mas a pesquisa de sistemas nestes primeiros passos foi severamente prejudicada pela dificuldade de coletar dados, modelar e extrair predições de modelos que estivessem focados na forma como as partes de um sistema interagem, do que na descrição minuciosa das partes. A solução encontrada foi em muitos casos a construção de modelos excessivamente simplificados com base em pressupostos que não podiam ser testados. Talvez isto explique por que este programa de pesquisa acabou por se tornar minoritário na comunidade científica.

A Biologia Sistêmica engloba o estudo de redes biológicas complexas, como metabólicas, gênicas, proteicas, dentre outras, e visa, a partir de fenômenos complexos, prover uma estrutura abstrata que aumente a compreensão da dinâmica de redes biológicas (FERRELL, 2009).

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).

Seguindo a mesma linha de von Bertalanffy, Stuart Kauffman publicou, em 1969, um trabalho onde ele propôs que as principais caraterísticas 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).

Nos últimos anos temos assistido uma retomada da abordagem sistêmica, revigorada pelo surgimento de técnicas experimentais e computacionais de larga escala que nos permitem construir, simular e testar sistemas de forma muito mais adequada. Esta retomada está relacionada a uma "transição de fase" da biologia (JOYCE; PALSSON, 2006) de uma ciência pobre para uma ciência rica em dados. Mas mais do que isso, o surgimento da internet e a pressão das agências de fomento possibilitaram a distribuição de dados para toda a comunidade científica. Nos últimos anos a comunidade científica vem discutindo como padronizar estes dados para permitir sua análise em larga escala. Um último desafio ainda deve ser vencido, convencer a comunidade da relevância dos resultados negativos, ingrediente que melhoria em muito a capacidade preditiva dos modelos gerados utilizando aprendizado de máquina.

Um dos marcos experimentais que permitiu a retomada da abordagem sistêmica foi o projeto Genoma Humano que determinou a seqüência de bases que compõe o DNA dos seres humanos. Mas o conhecimento desta sequência, ainda que um passo fundamental, é por si só insuficiente para responder as perguntas que nos inquietam e que motivaram este avanço. Para conseguirmos responder estas perguntas temos que compreender como a informação biológica armazenada no DNA é lida e transformada nas características humanas que nos definem. Este raciocínio pode ser aplicado também para outros organismos cujos genomas estão sendo determinados ou serão determinados.

A genômica mais do que um fim em si mesma, é o primeiro exemplo de uma nova classe de tecnologias, que estão sendo chamadas de Ômicas, e incluem a proteômica, metabolômica, transcriptômica, entre outras. Os dados produzidos através destes experimentos de larga escala estão permitindo a construção de modelos matemáticos cada

vez mais complexos, e que geram previsões quantitativamente corretas. De acordo com Joyce e Palsson (JOYCE; PALSSON, 2006) para podermos avançar nesta direção não basta simplesmente coletar dados com alto desempenho temos que integrar bases de dados construídas utilizando diferentes metodologias experimentais. Na Figura 1 apresentamos de forma esquemática como as diferentes ômicas estão sendo integradas para a construção de modelos para sistemas biológicos.

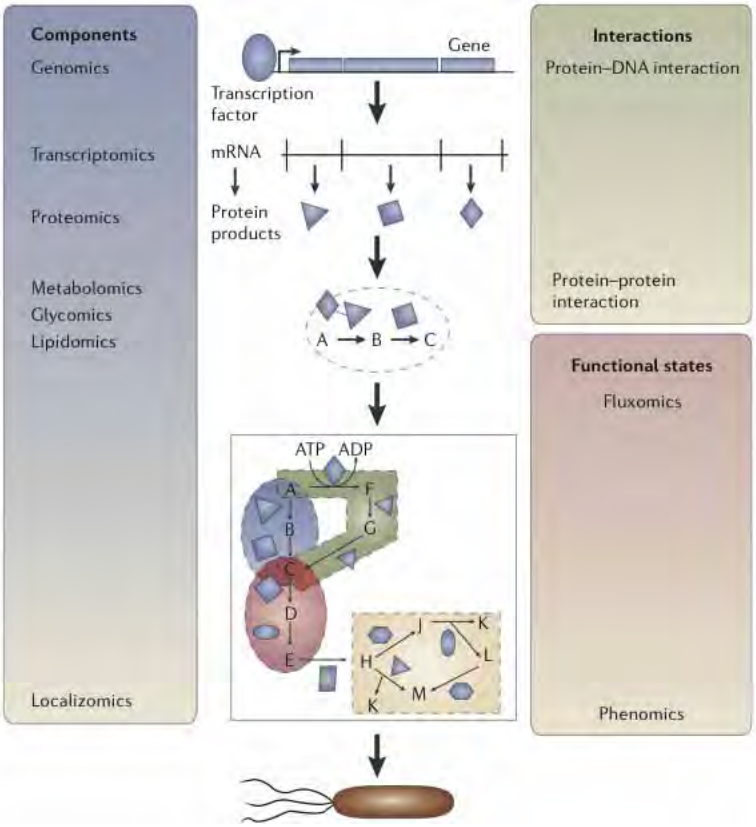


Figura 1: Representação esquemática da integração das diferentes ômicas . Extraído de (JOYCE; PALSSON, 2006).

O impacto aplicado destas novas tecnologias ainda está apenas começando e depende da criação de ferramentas de análise de dados, de geração e teste de modelos e finalmente da proposição de novos experimentos. Tarefas estas que dependem de avanços em ciências como a Física e a Ciência da Computação, que deverão aprender a lidar com a rica complexidade dos dados biológicos.

O processamento destes dados só será possível através da integração de dados gerados por laboratórios dispersos geograficamente armazenados em bases de dados públicos. A simulação e o ajuste dos complexos modelos necessários também demandará por processamento distribuído. Ou seja a elucidação das complexas redes de interações moleculares

que existem no interior de cada célula está induzindo uma rede igualmente complexa de serviços computacionais que vêm sendo chamada de grade computacional.

Um marco importante nessa área é o trabalho de Jeong et al (JEONG et al., 2000) que conseguiu convencer a comunidade acadêmica que, apesar da inegável eficácia e utilidade das abordagens reducionistas abordagens integradoras que incorporassem dados de experimentos de larga escala em modelos puramente topológicos lançam uma nova luz sobre os sistemas biológicos. Estas abordagens apesar de estarem em sua infância, hoje são o principal arcabouço teórico utilizado para extrair conhecimento dos novos dados.

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 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).

Nesta tese iremos seguir nesta direção definindo, caracterizando e simulando redes biológicas integradas. Nossa abordagem é inerentemente pragmática, uma vez construídas as redes integradas nos valemos de ferramentas computacionais oriundas da Física, da Inteligência Artificial e da Estatística para obter informações de interesse biológico com base em dados disponíveis em bancos de dados públicos. Visamos unificar dados provenientes de diferentes fontes experimentais tais como:

1. interação física de proteínas,
2. interações regulatórias,
3. interações metabólicas,

Com base nestas informações construímos modelos topológicos, em geral hipergrafos com arestas de diferentes cores que serão caracterizados visando relacionar propriedades topológicas com características de interesse biológico, tais como essencialidade dos genes (no caso de bactérias) ou envolvimento em alguma doença humana ou ainda ser alvo para drogas.

Esta tese está organizada na seguinte forma no capítulo 1 vamos discutir a construção de redes biológicas, no capítulo 2 vamos discutir a caracterização global da topologia de redes, no capítulo 3 discutimos as ferramentas de análise que podemos utilizar para extrair informações biológicas usando informações topológicas locais. Nosso objetivo não é o de apresentar uma revisão sobre os avanços nessa área, mas apresentar de forma ordenada as minhas contribuições.

1 Construção de Redes Biológicas

1.1 Grafos

Grafos são estruturas matemáticas definidos por dois conjuntos: $G = \{V, A\}$ onde V é um conjunto de vértices e A um conjunto de pares formados por elementos de V, chamados arestas. Se os pares forem ordenados o grafo é dito direcionado e não direcionado caso contrário.

No caso em que as arestas podem ser de diferentes tipos o objeto matemático é chamado de hipergrafo. Quando os vértices de um grafo podem ser divididos em dois conjuntos V_1 e V_2 de forma que só existem arestas entre V_1 e V_2 o grafo é chamado de dígrafo.

Grafos são elementos úteis pois são abstrações de sistemas interagentes, por exemplo considerere um sistema dinâmico descrito pelas equações:

$$\dot{x_i} = f_i(x_1, \ldots, x_n)$$

podemos induzir um grafo considerando como vértices as variáveis x_i e dizendo que x_i interage com x_j se f_i depende explicitamente de x_j. Ou ainda podemos proceder experimentalmente e analisar a correlação entre variáveis e induzir um grafo considerando como interagentes as variáveis com correlação alta o suficiente.

Como grafos são estruturas matemáticas muito estudadas temos a nossa disposição muitas ferramentas para sua análise e processamento. Por outro lado resultados válidos para grafos possuem repercussão em muitas outras áreas do conhecimento (BOLLOBÁS, 1979).

Os grafos são estruturas muito antigas, remontando pelo menos a Euler e o problema da ponte de Königsberg. Na Figura 2 mostramos três representações para o problema. Na primeira representação temos uma visualização realista, na segunda abstraímos quase todos os detalhes, mas mantemos os detalhes geográficos já no terceiro caso mantemos

apenas os aspectos topológicos, cada região é representada como um ponto e cada ponte como uma aresta. Essa última representação se restringe aos aspectos topológicos do problema, por exemplo ignoramos as áreas e a forma das ilhas, a capacidada das pontes etc.

Na formulação original do problema, o objetivo era traçar um percurso que visitasse todas as pontes uma única vez. Esse caminho é chamado ciclo euleriano. Euler demonstrou que para esse grafo não existem ciclos eulerianos. Ele resolveu uma longa disputa entre motoristas de charretes e turistas.

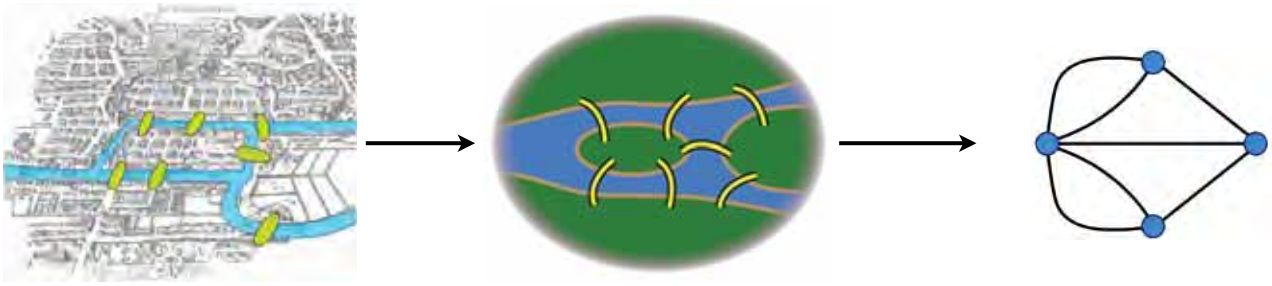


Figura 2: As pontes de Königsberg em três representações. Fonte: Wikipedia

O que fazemos nos dias de hoje não difere essencialmente da abordagem de Euler. Mas é conveniente perceber que o sucesso desta abordagem está condicionada a pergunta que gostaríamos de responder. A abordagem topológica é poderosa e permite entender o sistema estudado sob uma perspectiva bastante profunda mas não é uma teoria de primeiros princípios.

Em Biologia os grafos aparecem na área desde Darwin com o seu desenho de 1837 (ver Figura 3) representando a árvore da vida. Grafos são usados rotineiramente em problemas chaves de Bioinformática: comparação entre sequências de nucleotídeos (SMITH; WATERMAN, 1981) e em árvores filogenéticas.

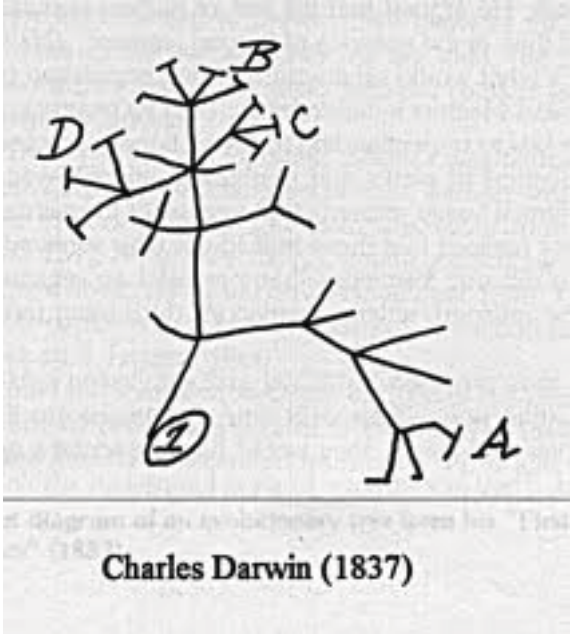


Figura 3: A árvore da vida de Darwin.

1.2 Redes Biológicas

Os sistemas biológicos possuem um número muito grande de agentes: proteínas, genes, compostos químicos que estão organizados em redes complexas de interação que são o produto da evolução.

O metabolismo de todos os organismos é caracterizado por uma rede complexa de reagentes conectados por reações químicas catalisadas por enzimas. Existem muitos organismos com genomas seqüenciados onde as proteínas codificadas estão sendo determinadas. Utilizando estas informações podemos construir as redes metabólicas de um dado microrganismo. Barabási e colaboradores (JEONG et al., 2000) propuseram uma representação gráfica da rede metabólica onde os nós representam os substratos, que estão ligados uns aos outros através de conexões que representam as reações metabólicas propriamente ditas. Uma estrutura topológica universal de larga escala foi descoberta nestas redes; a probabilidade de que um dado composto em uma rede metabólica participe em um certo número de conexões (reações) obedece uma lei de escala do tipo: $P(k) = k^{-\gamma}$, onde k é o número de conexões e γ é um expoente com valor 2.2.

Outra alternativa interessante é considerar as enzimas como vértices e considerar duas enzimas como interagentes se o produto da reação da enzima A for usado como reagente por uma enzima catalisada pela enzima B (LEMKE et al., 2004). A maior vantagem desta

abordagem é que enfatizamos os aspectos biológicos das redes metabólicas.

Em outro trabalho Barabási e colaboradores investigaram a rede de interação de proteínas. Utilizando o método experimental dos dois híbridos aplicados à levedura *S. cerevisiae* é possível determinar quando duas proteínas interagem fisicamente. Também neste caso foi possível determinar que a distribuição de conectividades das proteínas obedecia a uma lei de potência com expoente próximo a dois. Essa rede é usualmente denominada *protein-protein interaction* ou PPI. Além disso, foi possível comprovar que as proteínas com um número maior de interações tendem a ser mais importantes para o organismo, pois organismos nos quais estas proteínas não são produzidas possuem uma probabilidade maior de serem inviáveis (JEONG; MASON; BARABÁSI, 2001).

Outra rede biológica interessante é a rede regulatória onde genes interagem se um gene regula a expressão do outro gene, estes dados podem ser obtidos através de experimentos de micro-arranjos (FEATHERSTONE; BROADIE, 2002; AGRAWAL, 2002), proteômica (GONZÁLEZ-DíAZ et al., 2008) ou ainda experimentos do tipo chip-on-chip (LUSCOMBE et al., 2004).

1.2.1 Redes Biológicas Integradas

A expressão dos fenótipos depende geralmente da interação entre as redes descritas na seção anterior. Assim nos pareceu natural definir uma construção matemática que nos permitisse analisar de forma integrada os organismos de interesse (SILVA et al., 2008). Em uma rede integrada os genes representam os vértices, e considerando que os genes g_1 e g_2 codificam as proteínas p_1 e p_2, estes são conectados na rede, se

interação física p_1 e p_2 interagem fisicamente

interação regulatória g_1 regula a transcrição do gene g_2

interação metabólica um produto gerado pela reação catalisada pela proteína p_1 é consumido na reação catalisada pela proteína p_2 (nesta análise são excluídos os compostos mais usados, tais como, ATP, ADP, NAD, etc).

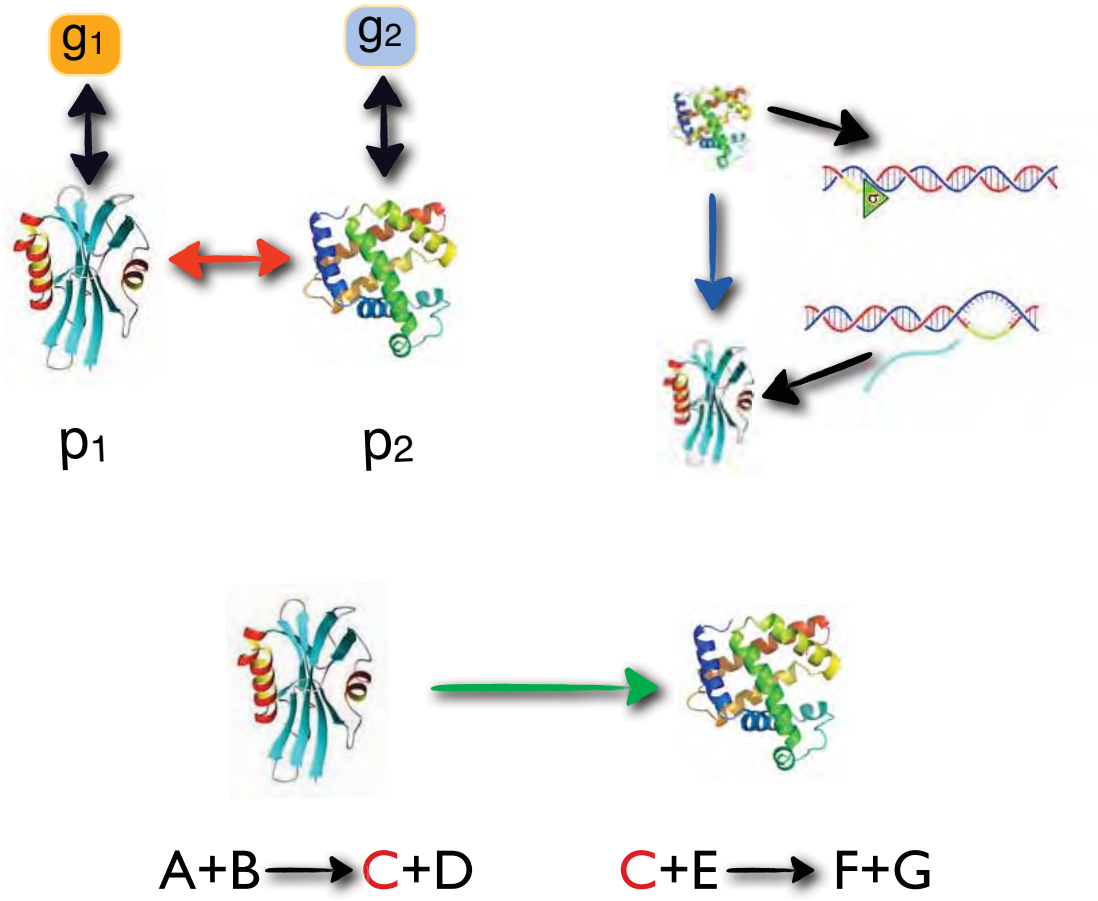


Figura 4: Esquema representando os três tipos de interação que compõe a rede integrada.

Na Figura 5 apresentamos uma representação da rede integrada da bacteria *E. coli*.

Figura 5: Rede biológica integrada para a *E. coli*. Em vermelho representamos os genes essenciais.

1.3 Construção de redes biológicas

No primeiro trabalho (LEMKE et al., 2004) que publicamos em redes biológicas tivemos de reconstruir a rede metabólica da *E. coli*. Neste trabalho entramos em contato com alguns dos desafios da Bioinformática. Este trabalha implicava em construir bases de dados que conseguissem reunir informações sobre milhares de genes e milhares de reagentes.

O problema inicial que encontramos foi a falta de consistencia na nomenclatura dos dados biológicos. Genes e compostos possuiam vários nomes e a tarefa de integração era complexa. O exemplo clássico é o composto água, que poderia surgir com vários nomes diferentes tais como “Water”, “water”, “H20” entre outros.

Neste trabalho inicial construímos os dados utilizando dados oriundos de várias fontes:

Palsson (EDWARDS; IBARRA; PALSSON, 2001)

Kegg (KANEHISA; GOTO, 2000)

Ergo (OVERBEEK et al., 2003)

Ecocyc (OUZOUNIS; KARP, 2000)

O trabalho de construção de um rede pode ser dividido em duas grandes etapas: a determinação dos vértices e a determinação das interações. Como as interações podem

ser oriundas de várias fontes é imprecindível elaborar uma lista de sinônimos para cada um dos genes e no caso de redes matabólicas para cada um dos compostos.

A lista de possíveis reações bioquímicas que podem ocorrer em uma célula são definidos pela física e química sendo considerada universal. Ou seja se pretendemos trabalhar com muitos organismos é bastante conveniente definir de forma criteriosa essas reações.

Apesar da diversidade dos organismos e dos bancos de dados percebemos que poderíamos facilitar a construção de redes biológicas se sistematizassemos o processo através da definição de uma ontologia. Este processo foi desenvolvido e implicou na produção de alguns trabalhos (BATTISTELLA et al., 2004, 2005) e uma dissertação de mestrado (SILVA, 2006).

1.4 Ontologias

O ponto-chave para compreender a estrutura e o comportamento da célula é integrar os dados disponíveis de forma a aumentar a nossa compreensão dos processos biológicos subjacentes que operam no interior da célula (BARABASI; OLTVAI, 2004; UETZ; IDEKER; SCHWIKOWSKI, 2002; YEGER-LOTEM et al., 2004; IDEKER et al., 2001a). Construir modelos biológicos que assimilem esse conhecimento são essenciais para formulação de novas hipóteses e prever comportamentos celulares que podem ser testado experimentalmente (IDEKER et al., 2001a). Mas a tarefa de integração não é simples.

Dados biológicos são disseminados em muitos bancos de dados diferentes. Estes bases de dados usam diferentes sistemas de gestão, formatos e visões de como representar os dados armazenados. A maioria deles são acessíveis por arquivos *flat* ou por interfaces web que permite algum tipo de consulta sobre ele. Os dois principais problemas envolvidos aqui são a dificuldade em analisar o dados quando se lida com formatos de arquivos heterogêneos e a inconsistência devido à ausência de um vocabulário unificado, que faz com que a mesma informação a ser representada em mais de uma maneira. Felizmente, formas de melhorar este cenário já existem.

Ontologias são uma abordagem importante para trazer ordem a este cenário e para permitir uma visão integrada destes dados. Uma ontologia é um especificação explícita de uma conceituação (GRUBER, 1993). Enquanto vocabulários controlados (por exemplo, RDF, XML Schema) apenas restringem as palavras usadas para descrever um domínio, ontologias extendem este controle de recursos de vocabulário e permitem a especificação formal dos termos e as relações entre eles. Em bioinformática, ontologias são cruciais para

manter a coerência de uma grande coleção de conceitos complexos e suas relações (AL, 1999).

Neste contexto, desenvolvemos a ontologia MONET: *MOlecular NETwork*. A Ontologia MONET é uma proposta para integrar dados de diferentes redes biológicas que existem no interior da célula.

Há uma necessidade de propostas de ontologias que permitem a compreensão da como as redes molecular dentro de uma célula determinam o comportamento das células (BARABASI; OLTVAI, 2004; UETZ; IDEKER; SCHWIKOWSKI, 2002; YEGER-LOTEM et al., 2004; IDEKER et al., 2001a). Entre outros requisitos, a proposta deve ser capaz de minimizar redundâncias de dados e inconsistências. O problema da distribuição de dados implica na adoção de padrões livres e abertos. Ele também precisa ser extensível, portanto, novos conhecimentos podem ser facilmente implementado pela agregação de novos conceitos.

A ontolgia MONET integra informações de transcrição-regulação, metabolismo e interações físicas proteína-proteína através de uma visão de que estabelece um modelo capaz de minimizar redundâncias de dados e inconsistências. É expansível pois a integração de novos conhecimentos pode ser facilmente implementada. Desta forma, MONET facilita a construção de modelos topológicos de redes integradas e a extração de conhecimento.

A definição de uma ontologia é demorada. Uma ferramenta adequada pode representar um ganho de produtividade significativa. Entre as ferramentas para construção de ontologia disponíveis optamos pelo Protégé-2000 [1]. As duas principais razões para a escolha de Protégé foram: (a) a necessidade, não só de um editor de ontologias, mas de uma ferramenta de Gestão da Base de Conhecimento já que queremos preencher o banco de dados com instâncias a partir de microrganismos diversos, e (b) a sua fonte aberta e extensível de arquitetura Java que permite melhorias em suas funcionalidades através do agregação de novos *plugins*. Esta última característica permite que o ontologia seja exportada para diferentes formatos exigidos por diferentes grupos de pesquisa.

Na Figura 6 podemos ver uma representação esquemática do os principais conceitos implementados para concretizar esta abordagem integrada. Vários tipos de conceitos relacionados à mollecules químicos, tais como DNA, RNA, mRNA, rRNA, tRNA e snRNA.

A rede de transcrição-regulação implementa conceitos como *Operon* (um conjunto de genes transcritos sob o controle de um gene operador), *Unidade de transcrição* (parte do DNA que será transcrito em um RNA), *Terminator* região do DNA (onde o transcrição

[1] http://protege.stanford.edu

supostamente paragens), *ORF* (uma parte de um seqüência que potencialmente codifica uma proteína), *site DNA* (seqüência cuja localização e seqüência de bases são conhecidos), *Promotor* (Um segmento de DNA que fornece um local onde as enzimas envolvidas na o processo de transcrição podem se ligar a uma molécula de DNA, e iniciar transcrição), e informações *Interação Regulatória* (relativa aos dados de transcrição-regulação que está sendo mapeada).

A rede de transcrição-regulação está envolvido com interações entre DNA e proteínas, e conseqüente produção de proteínas. A rede metabólica também envolve proteínas caracterizadas por sua função enzimática. As proteínas são o elo comum entre essas redes.

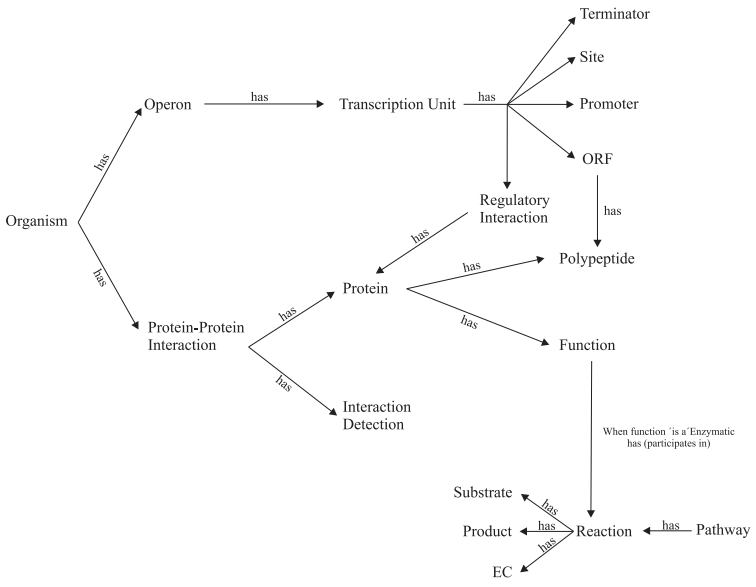


Figura 6: Os principais conceitos de Ontologia MONET que integram as redes metabólicas, regulatórias e de interação fisica proteína-proteína.

A rede de interação proteína-proteína tem pares de proteínas cuja interação foi de-

tectada experimentalmente ou por métodos *in silico*. Este conhecimento também foi mapeada na ontologia MONET. Para cada interação proteína-proteína adotamos a partir da ontologia PSI MI o conceito de *detecção de interação (id MI: 0001)* e sua sub-árvore de conceitos. O método para determinar a interação foi dividido em sub-métodos de *experimental* e *in silico* e cada um com suas respectivas notações.

A rede via metabólica da MONET é responsável pelos dados envolvidos em reações do metabolismo. O metabolismo de moléculas pequenas da MONET é um subconjunto do metabolismo completo que exclui a replicação do DNA e síntese proteíca. Além dos conceitos de *reação*, *substrato* e *CE* (o número da enizma atribuído pela comissão enzima) outros conceitos como *inibidor*, Embora as estruturas da rede metabólica e redes de interação proteína sejam semelhantes, há um certo número de diferenças significativas. Enquanto vias metabólicas concentram-se na conversão de moléculas pequenas e a enzima responsável para estas conversões, mapas de interação de proteínas concentram-se principalmente na contatos físicos, sem conversões químicas (UETZ; IDEKER; SCHWIKOWSKI, 2002).

O aspecto espacial também foi levado em consideração. MONET implementa um conceito intitulado *Compartimento* para indicar a localização subcelular. Considerar a localização de uma proteína e outros produtos químicos é uma característica importante para permitir conclusões mais precisas.

Usando abordagens parcilamente inspiradas na ontologia Monet conseguimos construir redes integradas para a *E. coli*, para a *S. cerevisae*, e metabólicas para *M. pneumoniae* (BARCELLOS; HERÉDIA; SCHMITH, 2008) e para o Eucalipto (MOMBACH et al., 2005).

1.5 Rede Integrada do *H. sapiens*

Para a tese de Doutorado de Marcio Acencio necessitávamos construir a rede integrada para o *H. sapiens*: Rede Integrada do Genoma Humano *RIGH* (ACENCIO, 2010).

A *RIGH* 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 outras informações, as relações entre nomes oficiais, apelidos e *GeneIDs* dos genes humanos.

1.6 Sub-redes

Em determinadas situações lidar com os dados referentes ao organismo completo pode ser bastante dificil nestes casos é interessante construir sub-redes. Após a construção da $RIGH$ estávamos interessados em entender a relação entre duas características de células cancerosas (as fontes dos dados estão descritas no apêndice A).

1.6.1 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 matrix extra-celular.

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 organis-

mos. 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 7 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).

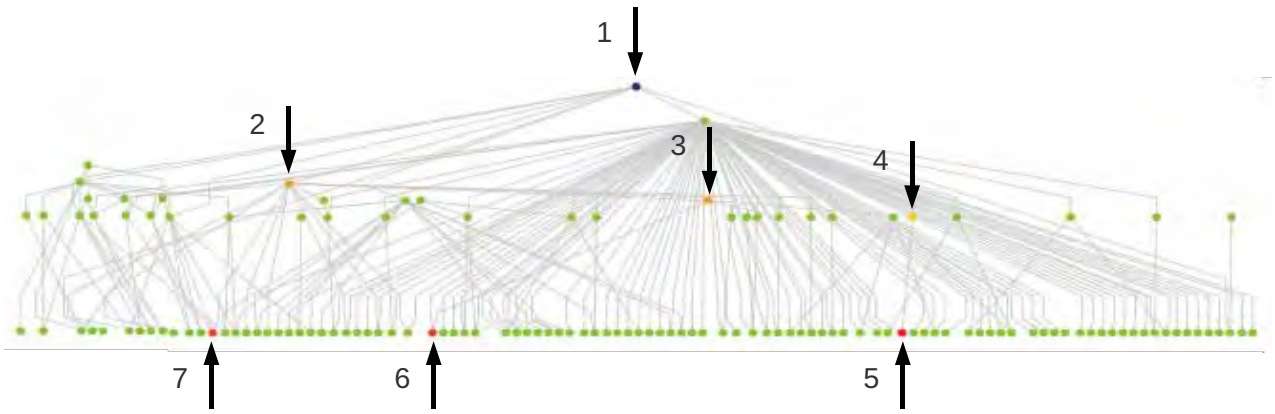


Figura 7: Parte da rede das relações hierárquicas entre termos da categoria *biological process* (processo biológico) do *Gene Ontology* com o termo *cell cycle* (ciclo celular) no topo da hierarquia. Vértice 1: *cell cycle* (ciclo celular); vértice 2: *mitotic cell cycle* (ciclo celular mitótico); vértice 3: *interphase of mitotic cell cycle* (intérfase do ciclo celular mitótico); vértice 4: *G1/S transition of mitotic cell cycle* (transição G1/S do ciclo celular mitótico); vértice 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); vértice 6: *M/G1 transition of mitotic cell cycle* (transição M/G1 do ciclo celular mitótico); vértice 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 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 1 foram considerados g_{cc} (54 genes) ou g_{am} (66 genes).

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 vértice de partida A e um vértice alvo B, realiza a busca de vértices adjacentes ao B localizados nos caminhos geodésicos entre A e B (Figura 8).

A primeira parte do *busca_cg* consiste na execução do *predecessor* tendo como vértices de partida g_{cc} e g_{am} e como vértices alvos todos os outros genes da $RIGH$. O resultado é a geração, para cada vértice 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 8). A segunda parte do *busca_cg* consiste na busca sequencial, em cada uma das listas geradas 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 8). 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 $RIGH$.

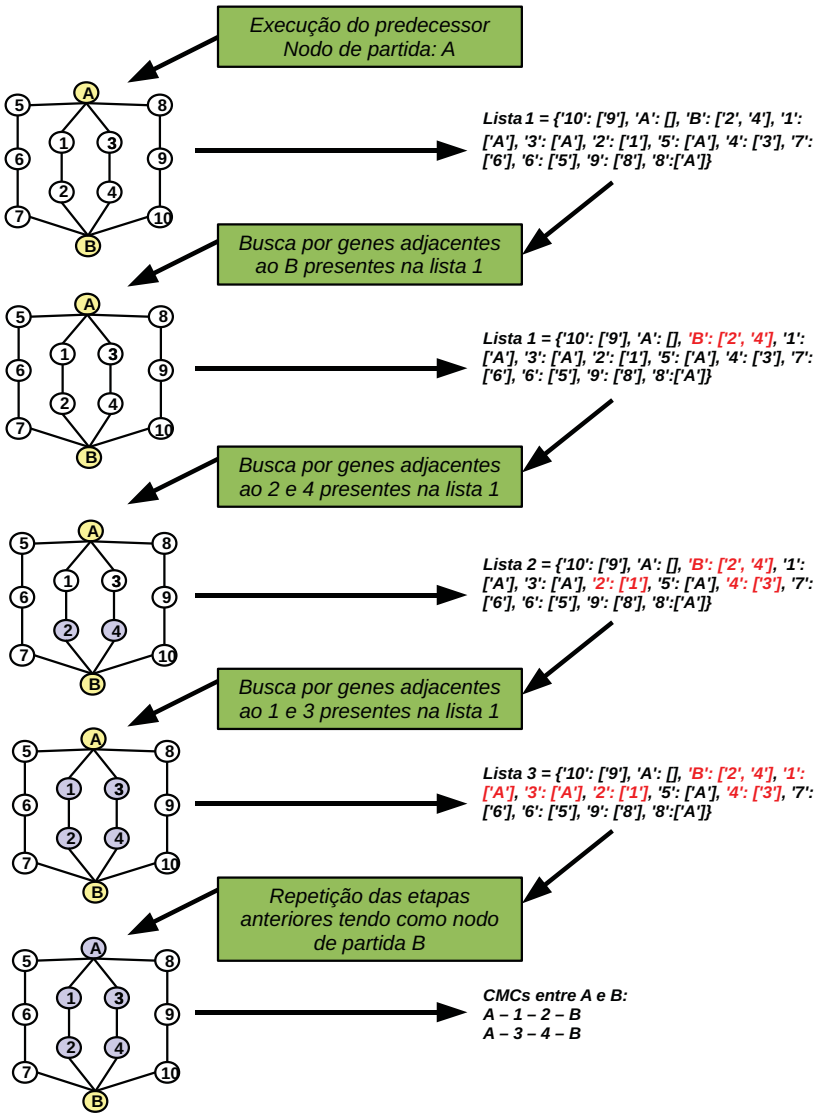


Figura 8: Esquema de funcionamento do algoritmo *busca_cg*. O vértice A representa o g_{cc} ou g_{am} de partida, o vértice B representa o g_{cc} ou g_{am} alvo e os outros vértices representam outros genes da $RIGH$. A primeira etapa do *busca_cg* consiste na execução do algoritmo *predecessor* do *Networkx* tendo A como vértice de partida e todos os outros genes como vértices 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 vértice de partida.

Tabela 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

2 Caracterização Global de Redes Biológicas

Kauffman propôs no final dos anos 70 que as principais características dos seres vivos poderiam ser compreendidas não através de uma análise minuciosa de cada um dos atores bioquímicos que operam em um dado organismo. Pois dado o proibitivo número de agentes e de suas complexas interações seria impossível experimentalmente caracterizá-las completamente. Por outro lado, a proposta reducionista de isolar algumas funcionalidades e estudá-las isoladamente seria uma aproximação perigosa.

A proposta metodológica de Kauffman consistia em propor modelos genéricos que incluíssem apenas algumas propriedades essenciais dos agentes bioquímicos e de suas inter-relações. As principais características dos seres vivos seriam previstas analisando o comportamento de elementos típicos de modelos aleatórios construídos com base nas propriedades determinadas experimentalmente de organismos vivos. As propriedades mais relevantes de acordo com Kauffman seriam K e N, o número de agentes e a quantidade média de agentes que interagem diretamente com um determinado agente (KAUFFMAN, 1993).

Na época de sua proposição mesmo a determinação destes parâmetros era impossível. Com o avanço de áreas como a proteômica e a genômica, o número de dados experimentais começou a crescer de forma exponencial e as idéias de Kauffman puderam ser testadas de forma explícita em diversos contextos biológicos, entre eles o metabolismo, a rede de interação de proteínas e a rede regulatória.

Uma das conclusões de Kauffmann era de que os sistemas biológicos viviam na fronteira entre os sistemas caóticos e os sistemas periódicos. Isso ocorreria quando o parâmetro K fosse igual a 3. Nestas condições a evolução ocorreria de forma otimizada, os sistemas possuiriam comportamento mais complexo entre outras propriedades interessantes.

Nós investigamos de forma explicita se as redes de Kauffman de fato se adaptariam de

forma mais eficaz se $K = 3$ (LEMKE; MOMBACH, 2001) e concluímos que para o modelo que consideramos isso não ocorria.

A proposta de Kauffman apesar de ser muito interessante está em direta contradição com os resultados experimentais. Nesta seção iremos mostrar resultados que mostram que as premissas do modelo de Kauffman não são adequadas. Recentemente suas idéias estão sendo repensadas e as Redes Booleanas estão sendo utilizadas com sucesso para modelar sistemas biológicos (ALBERT et al., 2008; CHAVES; ALBERT, 2008).

A abordagem de Kauffman estava baseada em um modelo clássico de grafos aleatórios proposto por Erdös (ERDÖS; RÉNYI, 1960). Um grafo aleatório é definido como um conjunto de N vértices conectados por n arestas que são escolhidas aleatoriamente entre as $N(N-1)/2$ possíveis arestas. Existe um total de:

$$C^{n}_{N(N-1)/2} = \frac{(N(N-1)/2)!}{n!(N(N-1)/2-n)!} \tag{2.1}$$

grafos diferentes que podem ser criados com N vértices e n arestas, todos equiprováveis estatisticamente.

Um grafo aleatório também pode ser definido pelo modelo binomial, no qual um par de vértices qualquer é conectado com uma probabilidade p. Desta maneira, o número esperado de arestas no grafo será:

$$E(n) = p[N(N-1)/2]. \tag{2.2}$$

A distribuição de conectividades de um grafo aleatório obedece uma distribuição binomial, mas para N grande, ela é bem aproximada por uma distribuição de Poisson:

$$P(k) \simeq e^{-\langle k \rangle} \frac{\langle k \rangle^{k}}{k!}. \tag{2.3}$$

Onde $\langle k \rangle$ é a conectividade média dos vértices no grafo.

Antes de começar a trabalhar com Bioinformática eu havia trabalhado com um modelo correlato de grafos aleatórios (ERDÖS; SPENCER, 1979). Neste modelo consideramos inicialmente um hipercubo, onde os vértices são os vértices do hipercubo e as ligações são as arestas do hipercubo. No nosso caso estudamos o que ocorre quando retiramos com probabilidade p os vértices desse sistema.

Em $p = p_c$ ocorre um fenômeno interessante, o sistema deixa de formar um conjunto conexo de vértices e passa a ser composto por vários conjuntos disconexos de vértices

chamados de agregados. p_c é chamado de limiar percolativo. Nós investigamos processos difusivos nesses sistemas em vários artigos (LEMKE; CAMPBELL, 1996; ALMEIDA; LEMKE; CAMPBELL, 2001b, 2001a; ALMEIDA; LEMKE, 2000; LEMKE, 2003) e a experiência obtida nessa área pôde ser aproveitada para desenvolver os algoritmos em redes biológicas.

O modelo de grafos aleatórios de Erdös era tão influente que muitas aplicações assumiam sua validade a priori. A situação mudou radicalmente quando Barabási mediu efetivamente a distribuição para grafos reais (BARABASI; ALBERT, 1999; DOROGOVTSEV, 2002; BARABASI, 2002) e descobriu que em muitos casos $P(k)$ era uma distribuição de cauda longa[1]. Estes grafos passaram a ser chamados de grafos "livres de escala". Essa nomenclatura faz referência ao fato que distribuições que possuem segundo momento definido, (como a distribuição gaussiana e a distribuição de Poisson), o desvio padrão define uma escala natural para o sistema, mas quando esse requisito não é satisfeito o sistema não possui uma escala característica.

As medidas realizadas por Barabási e por muitos outros pesquisadores mostravam que $P(k) \sim k^{-\gamma}$ (DOROGOVTSEV; MENDES, 2003). O que implica que existiam alguns vértices com muitas conexões, eles foram chamados de *hubs* e a grande maioria dos demais vértices possuem um número muito pequeno de conexões.

As leis de potência possuem uma longa história na Física e na Biologia Sistêmica. Von Bertalanffy iniciou suas investigações em Biologia de Sistemas observando a dependência em forma de lei de potência da relação entre metabolismo e massa de um organismo (BERTALANFFY, 1968). A geometria Fractal de Mandelbrot está relacionada a uma lei de potência que descreve a dependência da massa de um objeto com sua escala (MANDELBROT, 1982). Ou ainda a Lei da Gravitação de Newton que relaciona força com distância. Podemos dizer informamente que as redes livres de escala compartilham muitas das propriedades dos fractais, entre elas a invariância por escala e a presença de desordem estrutural.

Dentro de minha perspectiva pessoal eu possuía alguma experiência com fractais (LEMKE et al., 1993; MOMBACH; LEMKE; BODMANN, 2002) e havia publicado um trabalho sobe evolução (LEMKE, 1998). Eu possuía interesse em Bioinformática e foi bastante natural focar meus interesses nesta nova área.

[1]Distribuições de cauda longa possuem o primeiro momento ou o segundo momento não definido.

2.1 Medidas

Uma questão importante é a de organizar as diferentes redes aleatórias em classes que compartilhem propriedades relevantes e classificar os sistemas de interesse nestas classes.

Os físicos da área de materia condensada desenvolveram um ferramental matemático sofisticado para lidar com transições de fase, que são um exemplo clássico de fenômeno emergente. Em linhas gerais nestes sistemas as grandezas termodinâmicas divergem nas proximidades de uma transição de fase. Essas divergências são caracterizadas por expoentes críticos. Os expoentes críticos não dependem dos detalhes das interações, mas dependem da dimensão dos sistemas e das simetrias presentes. Essas propriedades em nosso contexto são dadas por parâmetros topológicos. O que se espera é que em analogia com os sistemas termodinâmicos é que as redes também se organizam em classes de universalidade caracterizadas por alguns expoentes críticos.

As redes complexas são fortemente não homogêneas, o que não é tão usual em sistemas desordenados na área de Física como os vidros de spin e as redes neurais (MEZARD; PARISI; VIRASORO, 1987). Estas heterogeneidades são importantes para caracterizar os vértices da rede.

Considere a rede formadas pelas páginas da internet e as arestas são os *links* hipertextuais entre as páginas. Quais são as páginas mais relevantes? Uma alternativa simples é contar o número de arestas, pois páginas que são mais citadas devem ser mais interessantes. Considere agora uma rede biológica, quais são os genes mais "importantes"? Quais genes que uma vez perturbados vão influenciar mais fortemente os sistemas biológicos?

Para responder essas e outras perguntas vamos apresentar uma série de medidas que nos permitam caracterizar vértices e redes.

Grau O conceito chave é o grau de um vértice, $deg(v)$ é o número de vizinhos do vértice v no caso de grafos direcionados deg_{in} e deg_{out} medem o número de ligações que chegam e saem do vértice respectivamente.

Distribuição de conectividades A distruibuição de conectividades, $P(k)$, é uma função que calcula o número total de vértices com determinado grau de conexão, em um dado grafo. Formalmente, a distribuição de conectividades é:

$$P(k) = \sum_{v \in V | deg(v)=k} 1 \tag{2.4}$$

onde v é um vértice do grupo de vértices V pertencentes ao grafo, e $deg(v)$ é o grau do vértice v. Essa mesma informação é freqüentemente representada como uma distribuição de conectividades cumulativa:

$$p(k) = \sum_{k'=k} P(k'). \tag{2.5}$$

Coeficiente de agrupamento de um vértice O coeficiente de agrupamento médio, $C(v)$, caracteriza a densidade de conexões próxima a um determinado vértice, v e seja $z = deg(v)$ (WATTS; STROGATZ, 1998; NEWMAN; WATTS, 1999; NEWMAN; STROGATZ; WATTS, 2001). Se todos os seus vizinhos também forem vizinhos entre si, existiriam $z(z+1)/2$ ligações entre eles. Considere o número efetivo de ligações entre primeiros vizinhos de v, c, definimos então

$$C(v) = \frac{2c}{z(z+1)}$$

Dependência de C com k Podemos calcular também a dependência do Coeficiente de agrupamento com k, definido por:

$$C(k) = \frac{\sum_{v \in V | deg(v)=k} C(v_i)}{\sum_{v \in V | deg(v)=k} 1}$$

Através da análise da dependência do coeficiente de agrupamento com a conectividade, pode-se definir se uma rede é hierárquica. Essas redes obedecem a relação $C(k) \sim k^{-\alpha}$

Grau de proximidade Grau de proximidade, $C_C(v)$, mede quanto um vértice particular está próximo de todos os outros vértices da rede. Para calcular o grau de proximidade soma-se a distância geodésica do vértice em relação a todos os demais vértices do grafo e depois inverte-se, uma vez que quanto maior a distância menor a proximidade, e esse valor ainda é normalizado em relação ao vértice de menor valor. Matematicamente, ela é representada por:

$$C_C(v) = \frac{1}{\sum_{t \in V/v} d_G(v,t)} \tag{2.6}$$

onde $d_G(v,t)$ é a distância entre v e t.

Grau de intermediação Grau de intermediação, $C_B(k)$, mede quanto um vértice está nos menores caminhos entre outros vértices na rede, ou seja, a fração dos trajetos mais curtos que incluem o vértice v. Também é considerado uma medida de

relevância do vértice, e é representado matematicamente por:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.7}$$

onde $\sigma_{st}(v)$ é o número de trajetos mais curtos de s até t que inclui v e σ_{st} é o número de trajetos de s até t.

$\boldsymbol{C_C(k)}$ Esta quantidade mede a dependência do grau de proximidade com o grau de um vértice definida por:

$$C(k) = \frac{\sum_{v \in V | deg(v)=k} C_C(v_i)}{\sum_{v \in V | deg(v)=k} 1}$$

$\boldsymbol{C_C(k)}$ Esta quantidade mede a dependência do grau de intermediação com o grau de um vértice definida por:

$$C(k) = \frac{\sum_{v \in V | deg(v)=k} C_B(v_i)}{\sum_{v \in V | deg(v)=k} 1}$$

Dano em Redes Metabólicas Para calcular d, usamos uma representação gráfica do metabolismo da *E. coli*. O gráfico é dirigido e tem dois tipos de vértices: um representa um produto químico e o outro uma reação envolvendo metabólitos de baixo peso molecular (esta rede está representada na Figura 9). A ligação entre um reação e um metabólito é dirigido para o metabólito, se o metabólito é um produto, e no sentido oposto, se o metabólito é um reagente.

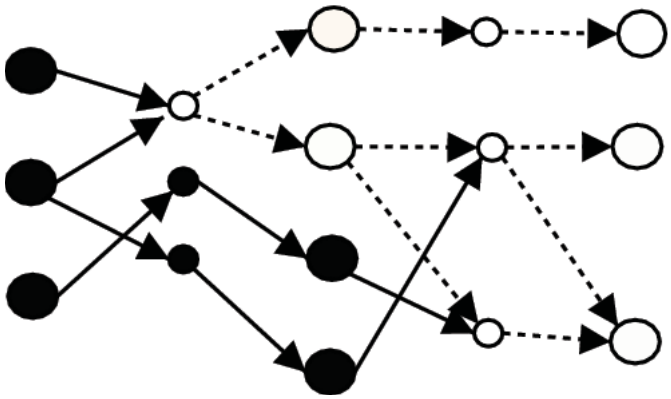


Figura 9: Dígrafo que representa a rede metabólica, os círculos grandes representam metabólitos e os pequenos as reações. Em negro representamos os vértices que são excluídos quando uma reação deixa de ocorrer.

Nós tratamos reações reversíveis como duas reações separadas. Propomos o seguinte

algoritmo:

- escolhemos uma enzima e determinamos todas as reações que ela catalisa
- se a reação é irreversível apagamos todos os metabólitos que ela produz exclusivamente
- se a reação é reversível deletar todos os metabólitos produzidos exclusivamente pelas reações direta e inversa catalisada pela enzima
- determinamos o conjunto de reações que ocorrem com o restante os metabólitos disponíveis
- iteramos o algoritmo até atingirmos um ponto fixo.
- o número total de metabólitos excluído é o dano d.

Dano em Redes de Interação Física Considere um grafo G e seu maior agregado G com n vértices. Agora considere $G'(v)$ o maior agregado obtido de G depois da deleção do vértice v que possui n' vértices. Nós definimos o dano d como sendo $d = n - n'$.

Na Figura 10 apresentamos uma representação esquemática das diferentes classes de redes, observe que a classificação proposta depende basicamente da medida de $P(k)$ e $C(k)$.

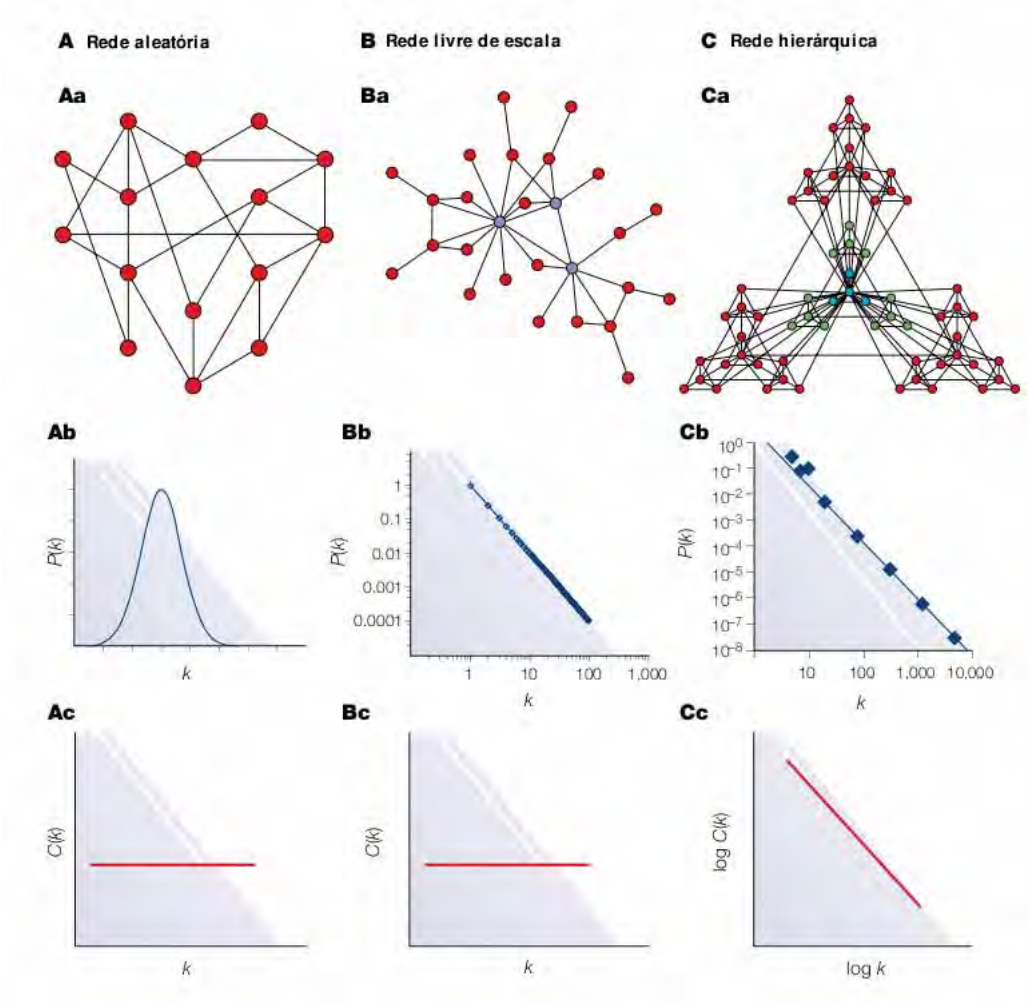


Figura 10: Modelos de redes.(Aa) Rede aleatória. (Ab) Distribuição de Poisson das conectividades de uma rede aleatória. (Ac) Coeficiente de agrupamento $C(k)$ independe de k em uma rede aleatória. (Ba) Rede com distribuição de conectividades obedecendo lei de potência. (Bb) A distribuição de conectividades representada em um gráfico Log-Log é linear. (Ac) Coeficiente de agrupamento $C(k)$ independe da conectividade k em uma rede com distribuição de conectividades obedecendo uma lei de potência. (Ca) Rede hierárquica: possui, simultaneamente, distribuição de conectividades obedecendo uma lei de potência e $C(k)$ dependente de k. (Cb) A distribuição de conectividades representada em um gráfico Log-Log é linear. (Cc) $C(k)$ tem dependência em k. Extraído de (BARABASI; OLTVAI, 2004).

Apesar do esforço dos pesquisadores em diversas áreas, existe ainda um vigoroso debate em torno da caracterização de redes biológicas e de outras redes complexas (DOROGOVTSEV; MENDES, 2003; BARABASI, 2009). Entre outras questões ainda persistem

dúvidas se $P(k)$ é de fato descrito por uma lei de potência para as redes de interação física entre proteínas. Como veremos a situação para a função $C(k)$ é ainda mais incerta.

2.2 Redes Integradas da *E. coli* e *S. cerevisiae*

Uma vez construídas uma rede biológica o primeiro passo consiste em buscar sua caracterização global (o apêndice A lista as fontes dos dados). Inicialmente nós realizamos essa caracterização no trabalho de conclusão de Tiago Andrade (ANDRADE, 2008).

As redes integradas contendo interações físicas entre proteínas, interações metabólicas e interações de regulação transcricional dos organismos *E. coli* e *S. cerevisiae* construídas possuem 2.349 vértices e 19.908 interações e 6.115 vértices e 82.994 interações, respectivamente.

Com o intuito de caracterizar a estrutura topológica dessas redes, ou seja, classificá-las em aleatórias ou livres de escala e determinar se são hierárquicas, nós analisamos quatro parâmetros: a distribuição de conectividades, o coeficiente de agrupamento médio, o grau de proximidade e o grau de intermediação.

2.2.1 Rede integrada da *E. coli*

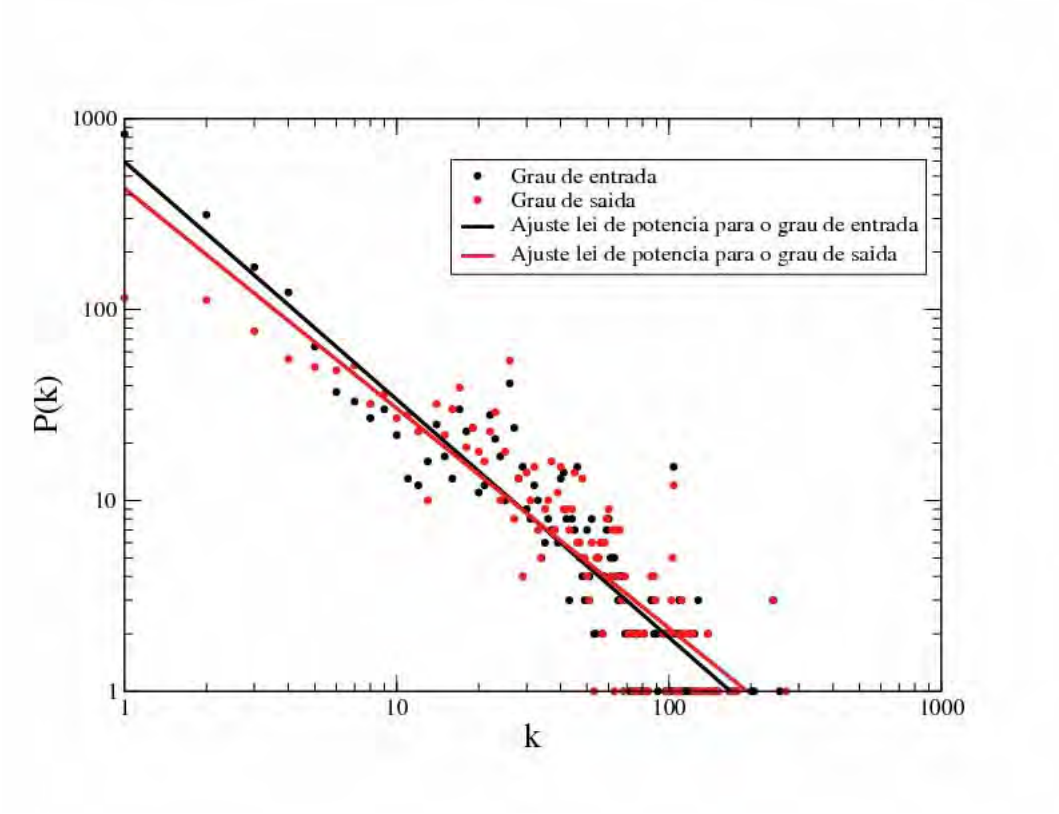


Figura 11: Distribuição de conectividades da rede integrada de interações moleculares entre genes da *E. coli* construída neste projeto. Essa distribuição segue uma lei de potência, o que a caracteriza como uma rede livre de escala.

Os ajustes realizados nas distribuições de conectividades da rede integrada de interações moleculares entre genes da *E. coli* seguem uma lei de potência representada por:

$$y = \alpha x^{-\beta} \tag{2.8}$$

onde α e β são constantes.

Os ajustes utilizados nas distribuições de conectividades da rede integrada de interações moleculares mostrados na Figura 11 foram:

Grau de entrada:

$$P(k)_{in} = 595,09k^{-1.24} \tag{2.9}$$

Grau de saída:

$$P(k)_{out} = 430,8k^{-1.15} \tag{2.10}$$

Na análise desse resultado, verificamos que a rede integrada da *E. coli* é livre de escala, o que indica que ela segue a lei de potência, assim como foi observado nas redes contendo somente um tipo de interação, ou seja, redes somente com interações proteína-proteína ou interações metabólicas ou interações de regulação transcricional (BARABÁSI; OLTVAI, 2004).

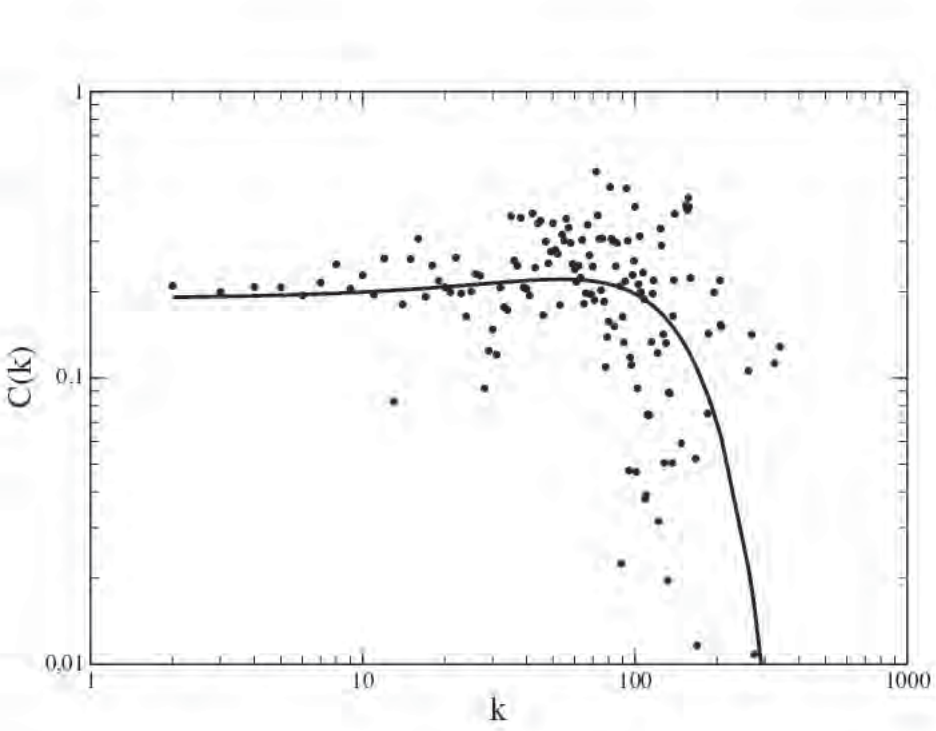


Figura 12: Dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k. Pode-se observar que o $C(k)$ é ajustado por um modelo linear quadrático, o que indica que a rede integrada de interações moleculares entre genes da *E. coli* construída neste projeto é linear quadrática.

O ajuste realizado na dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k dos dados da *E. coli* segue um modelo linear quadrático, descrito por:

$$y = \alpha e^{\beta x - \gamma x^2} \tag{2.11}$$

onde α, β, γ são constantes.

O ajuste utilizado na análise da dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k mostrado na Figura 12 foi:

$$C(k) = 0.19e^{0,0058k-0,036k^2} \tag{2.12}$$

Esse modelo linear quadrático para $C(k)$ indica que a rede integrada da *E. coli* não segue o modelo clássico de classificação proposto por Ravasz et al. (RAVASZ et al., 2002), onde vértices escassamente conectados fazem parte de áreas altamente conectadas com comunicação entre as diferentes áreas vizinhas altamente conectadas mantidas por alguns vértices altamente conectados denominados *hubs* (BARABÁSI; OLTVAI, 2004).

Nessa rede integrada, podemos observar que para baixos valores de k, existe uma fraca depêndencia em relação ao $C(k)$ e também que o $C(k)$ aumenta gradativamente a medida que o grau de conectividade aumenta até um valor de aproximadamente 100 conexões. Acima desse valor, os valores de $C(k)$ decaem abruptamente a medida que o grau de conectividade aumenta. Classificamos esse modelo como linear quadrático devido ao ajuste utilizado na dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k dos dados da *E. coli*.

2.2.2 Rede integrada da *S. cerevisiae*

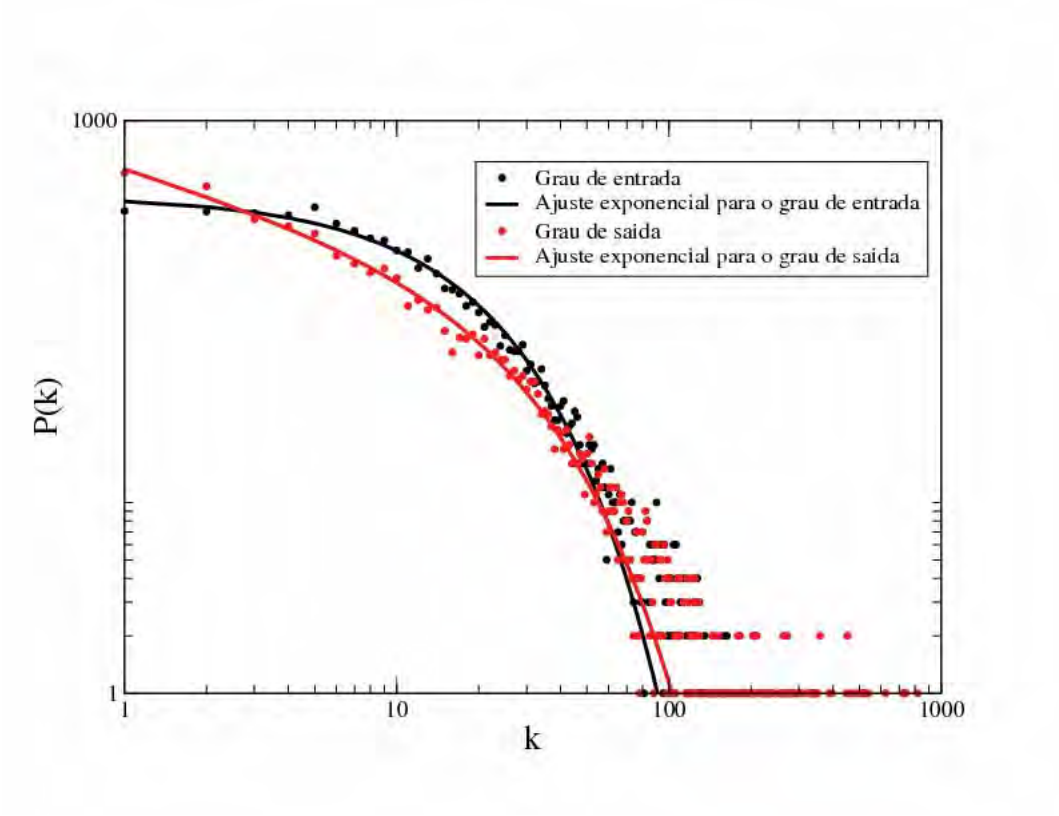


Figura 13: Distribuição de conectividades da rede integrada de interações moleculares entre genes da *S. cerevisiae* construída neste projeto. Essa distribuição é ajustada por um modelo livre de escala diferenciado.

Os ajustes realizados na distribuição de conectividades da rede de interações moleculares entre genes da *S. cerevisiae* seguem um modelo diferenciado representado por:

$$y = \alpha x^{\beta} e^{-\gamma x} \tag{2.13}$$

onde α e β são constantes.

Os ajustes utilizados na distribuição de conectividades da rede integrada de interações moleculares mostrado na Figura 13 foram:

Grau de entrada:

$$P(k)_{in} = 1141,67 k^{0,037} e^{-0,57k} \tag{2.14}$$

Grau de saída:

$$P(k)_{out} = 582,1k^{-0,43}e^{-0,043k} \tag{2.15}$$

Na análise desse resultado, verificamos que a rede da *S. cerevisiae* é representada por uma lei de potência com um pré fator exponencial. Se esse ajuste fosse puramente exponencial, diríamos que o grafo seria compatível com o modelo de Erdös-Rényi (ERDÖS; RÉNYI, 1959), mas neste caso dizemos que o grafo é compatível com o modelo livre de escala, no entanto, com um *cutoff* exponencial.

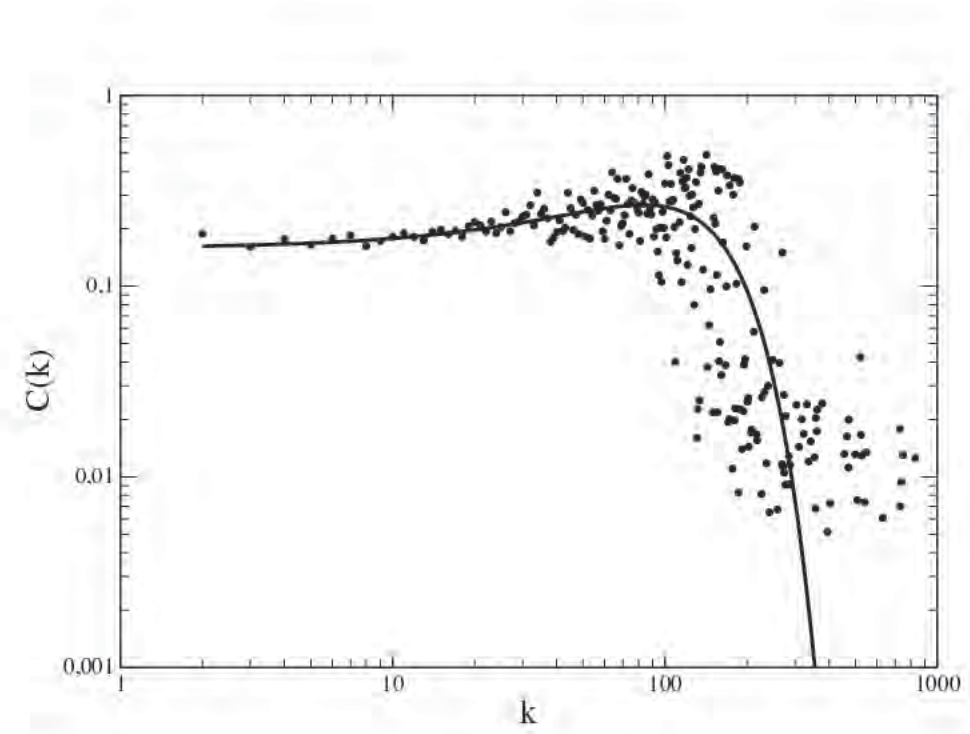


Figura 14: Dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k. Pode-se observar que o $C(k)$ é ajustado por um modelo linear quadrático, o que indica que a rede integrada de interações moleculares entre genes da *S. cerevisiae* construída neste projeto é hierárquica, mas com uma estrutura diferente das redes contendo apenas um tipo de interação.

O ajuste realizado na dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k dos dados da *S. cerevisiae* segue um modelo linear quadrático, descrito por:

$$y = \alpha e^{\beta x - \gamma x^2} \tag{2.16}$$

onde α, β, γ são constantes.

O ajuste utilizado na análise da dependência do coeficiente de agrupamento $C(k)$ em relação à conectividade k mostrado na Figura 14 foi:

$$C(k) = 0,16e^{0,012k-0,051k^2} \tag{2.17}$$

Esse modelo linear quadrático para $C(k)$ indica que a rede integrada da *S. cerevisiae* também não segue o modelo clássico de classificação proposto por Ravasz et al. (RAVASZ et al., 2002).

2.2.3 Análise comparativa do grau de proximidade e do grau de intermediação para as redes integradas da *S. cerevisiae* e da *E. coli*

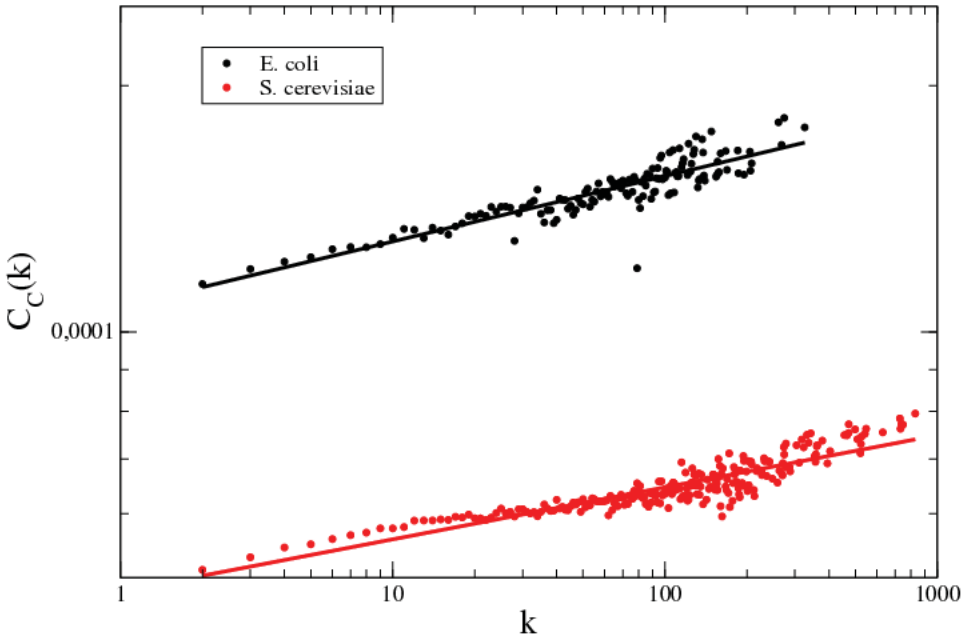


Figura 15: Grau de proximidade dos vértices em função de k para as redes integradas da *E. coli* e da *S. cerevisiae*. Ambos organismos tiveram suas medidas ajustadas por uma lei de potência.

Os ajustes realizados no grau de proximidade da rede integrada de interações moleculares entre genes da *E. coli* e da *S. cerevisiae* seguem uma lei de potência representada por:

$$y = \alpha x^{\beta} \tag{2.18}$$

onde α e β são constantes.

Os ajustes utilizados no grau de proximidade da rede integrada de interações moleculares mostrados na Figura 15 foram:

E. coli:

$$C_C(k) = 0,00011k^{0,079} \tag{2.19}$$

S. cerevisiae:

$$C_C(k) = 0,032k^{0,063} \tag{2.20}$$

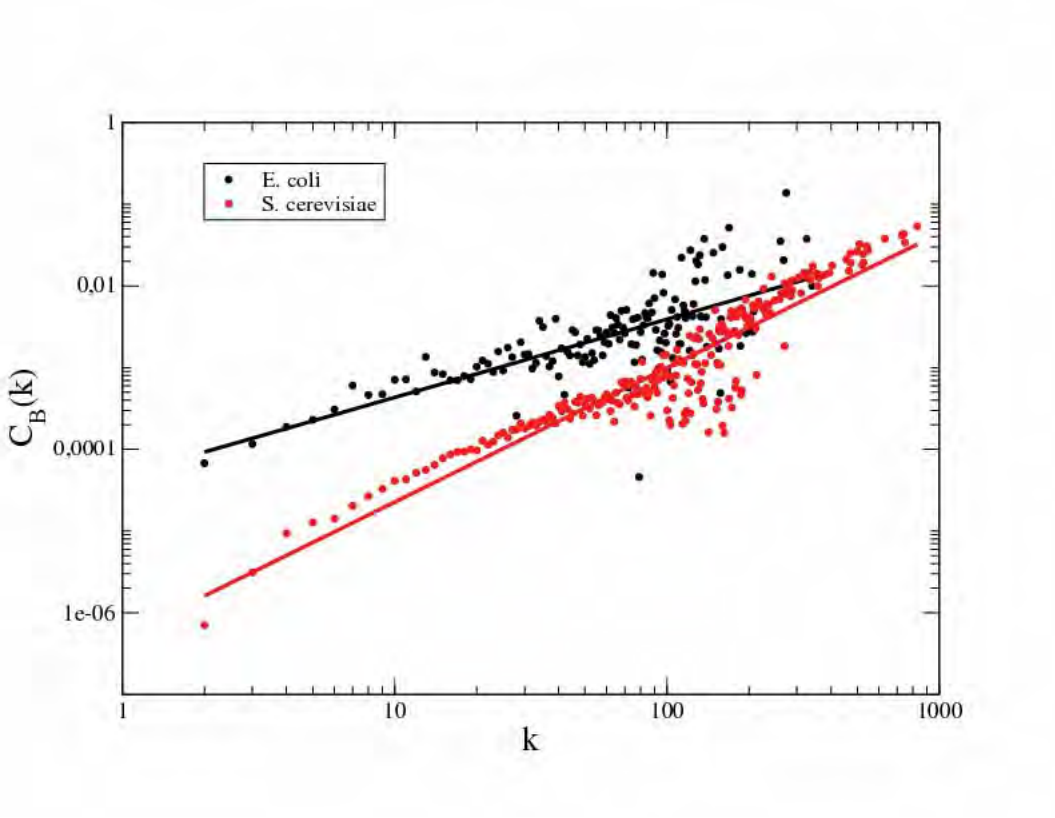


Figura 16: Grau de intermediação dos vértices em função de k para as redes integradas da *E. coli* e da *S. cerevisiae*. O ajuste utilizado para ambos organismos foi feito através de uma lei de potência.

Os ajustes realizados no grau de intermediação da rede integrada de interações moleculares entre genes da *E. coli* e da *S. cerevisiae* seguem uma lei de potência representada por:

$$y = \alpha x^{\beta} \tag{2.21}$$

onde α e β são constantes.

Os ajustes utilizados no grau de intermediação da rede de interações moleculares mostrados na Figura 16 foram:

E. coli:

$$C_B(k) = 0,032k^{0,95} \tag{2.22}$$

S. cerevisiae:

$$C_B(k) = 0,0047k^{1,64} \tag{2.23}$$

Em relação aos graus de proximidade e de intermediação dos vértices das redes integradas da *E. coli* e da *S. cerevisiae*, pode-se observar nas Figuras 15 e 16 que existe uma dependência como lei de potência entre essas quantidades. Estes parâmetros não foram medidos sistematicamente em outras redes assim não possuímos elementos para estabelecer uma comparação. Mas em linhas gerais podemos dizer que k está fortemente correlacionado com as demais centralidades.

Em suma, podemos observar os resultados obtidos nas Tabelas 2 e 3 para cada organismo estudado.

Tabela 2: Resultados para a rede integrada da bactéria *E. coli*

Parâmetro	Equação	Coeficientes
$P(k)$	$y = \alpha x^{-\beta}$	$\beta_{in} = -0,9 \pm 0,1$; $\beta_{out} = -1,0 \pm 0,1$
$C(k)$	$y = \alpha e^{\beta x - \gamma x^2}$	$\beta = -0,007 \pm 0,005$; $\gamma = -9,2 \times 10^{-6} \pm 3 \times 10^{-5}$
$C_C(k)$	$y = \alpha x^{\beta}$	$\beta = -0,081 \pm 0,007$
$C_B(k)$	$y = \alpha x^{\beta}$	$\beta = 1,6 \pm 0,4$

Os resultados obtidos indicam que a rede biológica integrada da *E. coli* é livre de escala e a dependência do $C(k)$ nos valores de k indica um novo modelo de ajuste, o que nos conduziu a um método novo de classificação, definido como linear quadrático. Já para a rede integrada da *S. cerevisiae* os resultados indicam que, ela é classificada como livre de escala com *cutoff* exponencial. Além disso ela também é classificada segundo o novo método de ajuste que propusemos, como linear quadrático.

Tabela 3: Resultados para a rede integrada da levedura *S. cerevisiae*

Parâmetro	Equação	Coeficientes
$P(k)$	$y = \alpha x^{-\beta} e^{\gamma x}$	$\beta_{in} = 0,037 \pm 0,01$; $\beta_{out} = 0,023\ pm 0,005$
$C(k)$	$y = \alpha e^{\beta x - \gamma x^2}$	$\beta = 0,017 \pm 0.005$; $\gamma = 1,5 \times 10^{-4} \pm 4 \times 10^{-5}$
$C_C(k)$	$y = \alpha x^{\beta}$	$\beta = 0,06 \pm 0,003$
$C_B(k)$	$y = \alpha x^{\beta}$	$\beta = 1,7 \pm 0,6$

2.3 *Motifs*

Um subgrafo é definido por um subconjunto de vértices e um subconjunto de arestas de um grafo. Os exemplos mais simples de subgrafos são triângulos (subgrafo de ordem 3), quadrados (subgrafo de ordem 4) e assim por diante.

Motivos são subgrafos que ocorrem significativamente mais vezes na rede real do que em redes aleatorizadas (MILO et al., 2002), que preservam as características dos vértices da rede real, ou seja, o número de ligações dos vértices deve permanecer o mesmo. A detecção de motivos em redes biológicas é importante. Por exemplo, foram encontrados três principais motivos nas redes de regulação transcricional da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae* ((MILO et al., 2002), (SHEN-ORR et al., 2002)). Dentre eles temos o:

SIM (*single-input module*), que foi demonstrado teoricamente (SHEN-ORR et al., 2002) e experimentalmente ((KALIR et al., 2001), (RONEN et al., 2002), (ZASLAVER et al., 2004)) por gerar programas temporais de expressão, tal ordem temporal pode ser útil em processos que requerem vários estágios para se completarem.

FFL (*feedforward loop*), ocorre quando um sinal externo causa uma resposta rápida em várias sistemas, tal como a repressão do sistema de utilização de açucar em reação a glicose (SHEN-ORR et al., 2002).

Esse tipo de detecção não foi feita em redes integradas, ou seja, as detecções dos motivos, até o presente momento foram feitas em redes que somente possuem apenas um tipo de interação. Portanto, nós propusemos a busca de motivos nas redes biológica integrada da *E. coli* e *S. cerevisiae*, sem levar em consideração as direções das ligações entre os vértices (GERMEK, 2009)

Nesta seção serão descritas algumas propriedades dos grafos, as quais foram utilizadas

no estudo, os bancos de dados utilizados para a formação das redes biológicas integradas e os algoritmos criados e utilizados para o estudo dos motivos.

As ligações foram classificadas de acordo com a Tabela 4:

Tabela 4: Legenda das interações

Letra	Tipo de interação
p	proteína-proteína
m	metabólica
r	regulatória
pm	proteína-proteína e metabólica
pr	proteína-proteína e regulatória

Para determinar os *motifs* precisamos de algoritmos que gerem versões aleatórias das redes de interesse. Existem diferentes formas de gerar essas versões:

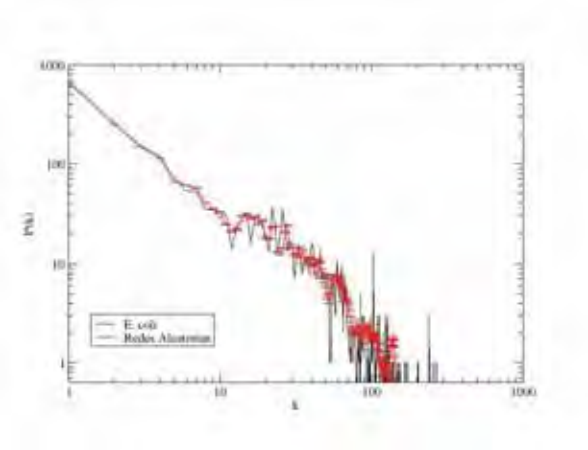
Método I: utiliza a lista de entrada com as interações e seus respectivos tipos, armazena os primeiros vértices da lista, depois os tipos de interações e, em seguida, os segundos vértices da lista. Com estes conjuntos de listas, é feita uma escolha aleatória de um termo em cada lista; estas ecolhas tornam-se os novos vértices com seus novos tipos de ligações. Assim podemos garantir que os vértices mantenham os seus graus de conectividades. Para observar se os graus de conectividade não mudaram, foi feito o gráfico de $P(k)$ em função de k.

Método II: utiliza a lista de entrada com as interações e seus respectivos tipos, armazena os tipos de interações em uma lista. Em seguida, é feita uma escolha aleatória de um termo da lista do tipo de interações, mantendo-se os vértices das antigas interações. Com isso apenas as interações são aleatorizadas.

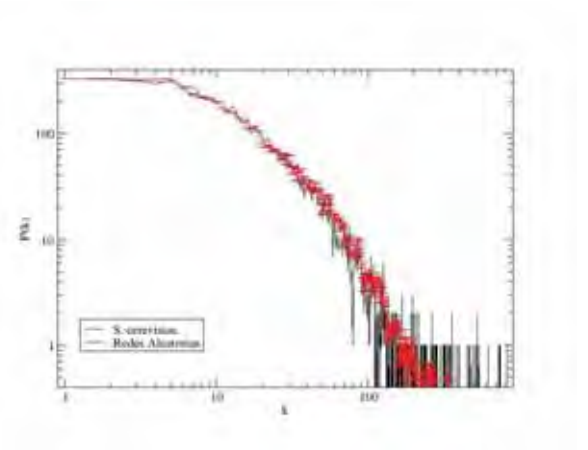
Antes da detecção dos motivos triangulares nas redes integradas da *E. coli* e da *S. cerevisiae*, nós geramos, primeiramente, dois grupos de 100 redes aleatórias: um grupo no qual somente os tipos de ligação entre os vértices foram aleatorizados e outro grupo no qual tanto o tipo de ligação quanto os pares de vértices ligados foram aleatorizados.

Nos dois casos, os graus de conectividade dos vértices nas redes reais foram preservados nas redes aleatórias. De acordo com a Figura 17, o algoritmo utilizado para a geração dos

dois grupos de 100 redes aleatórias foi bem sucedido, já que tanto para a *E. coli* quanto para a *S. cerevisiae*, os ajustes dos pontos referentes aos valores de $P(k)$ em função de k das redes aleatórias mostraram-se semelhantes aos ajustes das respectivas redes reais, indicando que as distribuições de conectividades foram preservadas.



(a) *E. coli*

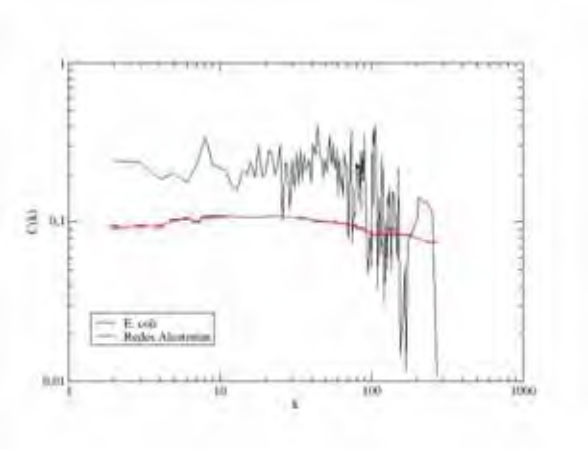


(b) *S. cerevisiae*

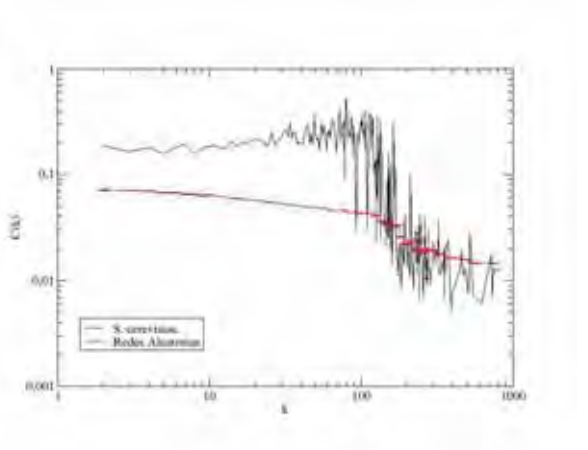
Figura 17: Grafico de $P(k)$ em relação k para *E. coli* e *S. cerevisiae* e de 100 redes aleatórias (usando o método I) da mesma. Nestes gráficos é possível observar que o $P(k)$ das redes experimentais e das Redes aleatórias possuem o mesmo comportamento.

As aleatorizações dos tipos de ligações e dos pares de vértices ligados das redes integradas da *E. coli* e da *S. cerevisiae* afetaram os valores de $C(k)$ dos vértices e a dependência desses valores em relação aos seus respectivos valores de k. Como pode ser observado na Figura 18, os valores de $C(k)$ desse grupo de redes aleatórias variaram pouco a medida que os valores de k aumentaram, comportamento distinto ao das redes reais, onde os valo-

res de $C(k)$ mantêm-se com pouca variação até próximo de k=100 e caem abruptamente após esse valor. Como o coeficiente de agrupamento representa, de fato, o número de triângulos aos quais um determinado vértice pertence, então os valores de $C(k)$ das redes aleatórias indicam que, quando o tipo de ligação e os pares de vértices conectados foram aleatorizados, o número de triângulos diminui, mesmo com a manutenção da distribuição de conectividades. De fato, enquanto o número de triângulos encontrados na rede real da *E. coli* foi de 72572, foram contados uma média de 42541,7 triângulos nas suas respectivas redes aleatórias onde o tipo de ligação e os pares de vértices conectados foram aleatorizados (Tabela 5). O mesmo pode ser observado para *S. Cerevisiae* (Tabela 6).



(a) *E. coli*



(b) *S. cerevisiae*

Figura 18: Grafico de $C(k)$ em relação a k da *E. coli* e de 100 Redes Aleatórias (Método I) da mesma. Observa-se que a dependência de $C(k)$ em relação ao grau de conectividade k da *E. coli* fica reduzido após sua rede ser aleatorizada.

Tabela 5: Porcentagem dos tipos de interações da rede experimental da *E. coli* e comparação com os resultados obtidos para redes aleatorizadas usando os Métodos I e II.

Interação	Exp. (%)	Mét. I (%)	Mét. II (%)
um tipo de ligação	94,81	30,73	30,76
uma ligação de um tipo e duas do outro	3,35	59,46	59,46
três tipos de ligações	0,0082	8,82	8,81
ligações do tipo pm	1,8	0,92	0,91

Tabela 6: Porcentagem dos tipos de interações da *S. cerevisiae* e comparação com os resultados obtidos para redes aleatorizadas usando os Métodos I e II.

Interação	Exp. (%)	Met. I (%)	Met. II (%)
um tipo de ligação	81,65	18,57	18,54
uma ligação de um tipo e duas do outro	17,7	66,59	66,64
três tipos de ligações	0,022	14,48	14,46
ligações do tipo pm ou pr	0,59	0,30	0,30

Como pode ser observado nas Tabelas 5 e 6, uma super-representação de triângulos contendo somente um tipo de interação perfazem cerca de 93% dos triângulos na *E. coli* e cerca de 90% dos triângulos na *S. cerevisiae* indicando que, embora essas redes sejam integradas, elas parecem ser altamente modularizadas, ou seja, pares de vértices ligados através de um mesmo tipo de interação tendem a formar módulos dentro das redes integradas. Os triângulos que possuem mais de um tipo diferente de interação correspondem a aproximadamente 10% dos triângulos e poderiam, por sua vez, representar "pontes"de ligação entre os módulos mencionados acima.

A composição e os tipos de motivos detectados nas redes integradas da levedura *Saccharomyces cerevisiae* e da bactéria *Escherichia coli* indicam que tais redes são organizadas em três módulos principais que são interligados entre si. Os módulos principais são constituídos pelos motivos contendo somente um tipo de interação e as interligações entre esses módulos contém os motivos que possuem mais de um tipo diferente de interação.

O processo de aleatorização diminiu sensivelmente o número de triângulos, ou seja

deve haver pressão seletiva para que surjam triângulos envolvendo elementos de uma mesma rede.

O alto grau de divisão das redes possui um efeito no comportamento de $C(k)$ podendo explicar a dependência observada. Temos basicamente três redes com diferentes comportamentos $C(k)$ amalgamados em uma única estrutura ligados fracamente por pontes não homogêneas, o que poderia explicar a forma complexa encontrada.

2.4 Comunidades

A detecção de comunidades pode ser realizada através de diversos algoritmos, mas a maioria se baseia na alta densidade de ligações entre os vértices de uma mesma comunidade. Até o momento, a detecção e análise das comunidades em redes biológicas foram realizadas apenas para redes compostas com apenas um tipo de interação (ex.: interações proteína-proteína ou interações metabólicas).

Para este propósito, nós detectamos as comunidades em redes biológicas integradas da bactéria *E. coli* e da levedura *S. cerevisiae* através do método de percolação de cliques e verificamos, através do cálculo da frequência de cada tipo de interação e sua respectiva entropia, se os componentes dessas tendem a ser interligados com um tipo preferencial de interação. Posteriormente, comparamos estes resultados com obtidos de redes aleatórias para determinar se essas comunidades têm relevância biológica. Por fim, foi obtida a entropia para os agregados devido a similaridades no perfil de expressão na estrutura da rede da *Saccharomyces cerevisiae* gerada através de dados de microarranjos (GE et al., 2001), e os resultados comparados com os valores obtidos para as comunidades encontradas anteriormente (GIGLIOLI, 2009).

Muitas pesquisas são, atualmente, realizadas a partir do uso de dados de microarray para a obtenção de estrutura de redes biológicas. No caso de redes de interações gênicas, obtém-se o perfil de expressão, a análise de agregação é realizada e, assim, os genes são agrupados em agregados de acordo com suas similidades de expressão. Este tipo de aglomeração é utilizada devido a hipótese de que genes co-expressos estão, provávelmente, relacionados a funções biológicas (GE et al., 2001).

A simples identificação de grupos bem conectados em uma rede pode transmitir informações úteis: se uma rede metabólica esta dividida em grupos, por exemplo, isto pode gerar evidencias de uma visão modular da dinâmica da rede, com cada grupo de vértices executando uma função distinta com algum grau de independência (NEWMAN, 2006).

Se as comunidades estão claramente separadas no sistema, a maioria dos métodos de aglomeração pode identificá-las. De fato, muitos métodos têm sido introduzidos recentemente para identificar módulos de redes variadas (NEWMAN, 2006) (CLAUSET; NEWMAN; MOORE, 2004) (ADAMCSEK et al., 2006) (PALLA et al., 2005) (KUMPULA et al., 2008), usando ainda a descrição topológica das redes ou combinando a topologia com os dados genomicos funcionais integrados. Entretanto, métodos diferentes predizem divisões diferentes entre módulos que não estão claramente separados.

Atualmente, não há critério matemático objetivo para decidir que uma divisão é melhor que outra, assim, alguns parâmetros internos controlam o tamanho típico de módulos não conhecidos e a mudança destes gera as diferenças entre métodos distintos de divisão em comunidades. Uma solução, portanto, é evitar procurar uma separação em um grupo absoluto de módulos, mas sim visualizar uma relação hierárquica entre comunidades de tamanhos diferentes.

Devido a esta diversidade de métodos de divisão de redes em comunidades, seguem duas linhas de pesquisa: partição de grafos e detecção de estrutura de comunidade. A primeira é adotada, particularmente, nas áreas de ciência da computação e campos relacionados, com aplicação em computação paralela e elaboração de circuitos integrados, dentre outras áreas. A segunda é utilizada por sociólogos e, recentemente, por físicos, biólogos e matemáticos, com aplicação, especialmente, em redes sociais e biológicas (NEWMAN, 2006). Ambas as linhas de pesquisa descritas são direcionadas a mesma questão, embora por meios um pouco diferentes. Há, entretanto, diferenças entre os objetivos dos dois campos que originam a diferença técnica das aproximações desejadas.

Uma aplicação típica da partição de grafos é a divisão de tarefas entre os processadores na computação paralela para minimizar a quantidade necessária de comunicação entre os processadores. Em cada cálculo, o número de processadores e a quantia de tarefas que cada um destes pode suportar são, normalmente, conhecidos e, deste modo, pode-se identificar em quantos grupos a rede será dividida e o tamanho de cada grupo. Além disso, o objetivo é, normalmente, encontrar a melhor divisão da rede independente se uma boa divisão existe (NEWMAN, 2006).

A detecção de estrutura de comunidades, pelo contrário, é, talvez, o melhor método para identificar a estrutura de redes de larga escala, como redes sociais, da internet e dados da web, ou redes bioquímicas. Este método, normalmente, assume que a rede de interesse está dividida naturalmente em subgrupos e o trabalho dos pesquisadores é identificar estes grupos. O número e o tamanho dos grupos são determinados pela própria rede e não pelo

pesquisador. Além disso, este método pode, explicitamente, admitir a possibilidade de que não haja uma boa divisão para a rede (NEWMAN, 2006).

Até agora, a detecção e análise de comunidades são realizadas em redes biológicas envolvendo apenas um tipo de interação, por exemplo, interação proteína-proteína ou interação metabólica. Como uma rede biológica real é composta, simultaneamente, por interações proteína-proteína, metabólicas e de regulação transcricional, esta pesquisa aborda como as comunidades são organizadas em redes que possuam os três tipos de interações, identificando-as pelo método de detecção de estrutura e analisando-as a partir do cálculo da frequência de cada tipo de interação e sua respectiva entropia.

Ambas as redes foram divididas em comunidades utilizando três métodos distintos, *CommunityStructureAssignment* e *MinCut* e o utilizado pelo *software CFinder*, denominado *Clique Percolation Method*.

Um alto valor de entropia medida na comunidade indica uma distribuição aleatória dos tipos de interações, ou seja, presença dos três tipos de interações existentes, sem predominância de nenhum dos três tipos (BUTTE; KOHANE, 2000).

Após o cálculo da entropia, as comunidades foram aleatorizadas. Isto é realizado alterando os tipos de interações entre os genes de maneira aleatória. Em seguida, as entropias das comunidades das redes aleatórias foram medidas e foi realizada uma comparação entre os resultados obtidos para as redes reais e as aleatorizadas.

Três métodos distintos foram utilizados na determinação das comunidades para as redes biológicas integradas da *E. coli* e da *S. cerevisiae* e como cada algoritmo utiliza um método de identificação, diferenças podem ser notadas e estão apresentadas nas Tabelas 7 e 8 e figura 26.

As diferenças mais notáveis entre os tipos de divisões em comunidades são o número de comunidades obtidas por cada método e o tamanho destas. As diferenças mostram claramente que dada uma rede complexa não existem ainda critérios objetivos para dividí-la em grupos de genes fortemente interagentes.

Tabela 7: Tabela comparativa para os métodos de divisão de comunidades utilizados.

Método	Tipo de divisão (número de comunidades)	Intersecção entre comunidades	Uso de CPU
Clauset	Detecção de estrutura	Vazia	20 minutos
Clique Percolation	Detecção de estrutura	Não vazia	12 horas
MinCut	Partição de grafos	Vazia	2 minutos

Tabela 8: Quantia de comunidades obtidas por cada método de divisão de comunidades utilizado.

	Número de comunidades	
Método	*E.coli*	S. cerevisiae
Clauset	127	28
Clique Percolation	220	1691
MinCut	30	30

O algoritmo proposto por *Clauset* (CLAUSET; NEWMAN; MOORE, 2004) divide as redes pelo método de estrutura de comunidades, isto é, assume que a rede possui uma estrutura naturalmente dividida em módulos e os identifica. Este método dividiu a rede da *E. coli* em 127 comunidades e a da *S. cerevisiae* em 28 comunidades (ver tabela 8).

A divisão em comunidades realizada pelo *CFinder* com o método *Clique Percolation* é, também, feita por estrutura de comunidades. O método também foi aplicado para a *E. coli* e para a *S. cerevisiae* e identificou, respectivamente, 220 e 1691 comunidades. Pode-se notar uma grande diferença no número de comunidades encontradas plo algoritmo de *Clauset* e o *Clique Percolate* e isso ocorre devido a aplicação de critérios distintos de detecção de comunidades.

Já a divisão pelo método *MinCut* é realizada por partição de grafos, ou seja, o número de comunidades no qual a rede será dividida,é previamente definido e o tamanho é aproximadamente o mesmo para todas. Neste trabalho, foi realizada a divisão das redes da *E. coli* e da *S. cerevisiae* em 30 comunidades.

Para as comunidades obtidas pelos métodos utilizados foram medidas as frequencias

de cada tipo de interação e sua repectiva entropia, para verificar se as há predomínio de algum tipo de interação nos módulos determinados. Apesar das diferenças entre os métodos, os resultados apresentaram um certo padrão, pois os valores de entropia medidos foram, em geral, bem próximos de zero, o que indica a existência deum tipo predominante de interação nas comunidades identificadas, independente do método utilizado. A fim de verificar se esses valores de entropia não são independentes da estrutura de comunidades, as redes foram aleatorizadas e a entropia medida novamente; os valores obtidos para as redes aleatórias foram próximos de *um*, portanto, distintos dos obtidos para as redes experimentais, o que indica que o predomínio de um tipo de interação está relacionado a estrutura modular das redes. Estes resultados estão apresentados nas Figuras 19, 20, 21, 22, 23, e 24.

Estes resultados estão de acordo com os apresentados no presente trabalho na seção anterior, pois ambos mostram que há predomínio de um tipo de interação em grupo com alta densidade de interações entre seus componentes. Possivelmente, essa concordância de resultados está relacionada ao fato de alguns algoritmos de detecção de comunidades utilizarem a estrutura de cliques na determinação da estrutura da rede e, os motivos triangulares são clique de ordem 3 podem se sobrepor para formar grupos de motivos, os módulos ou comunidades.

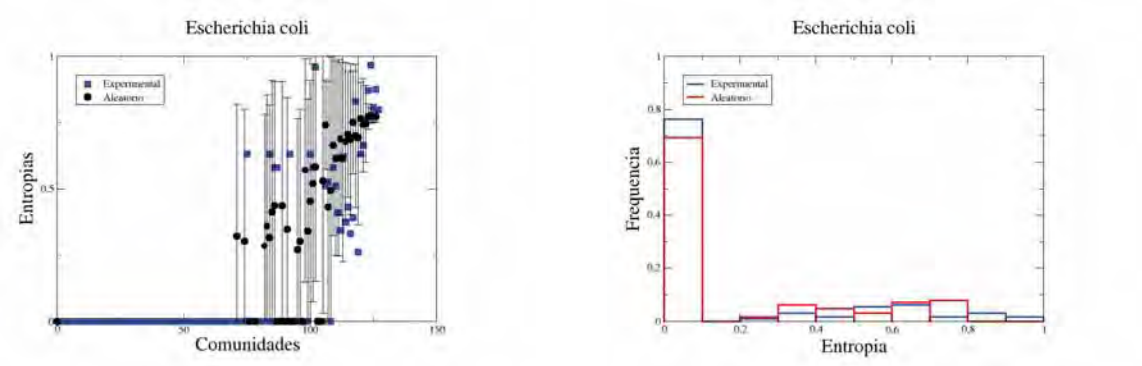


Figura 19: Representação da entropia para as comunidades encontradas pelo método proposto por *Clauset* para a rede da bactéria *Escherichia coli*.

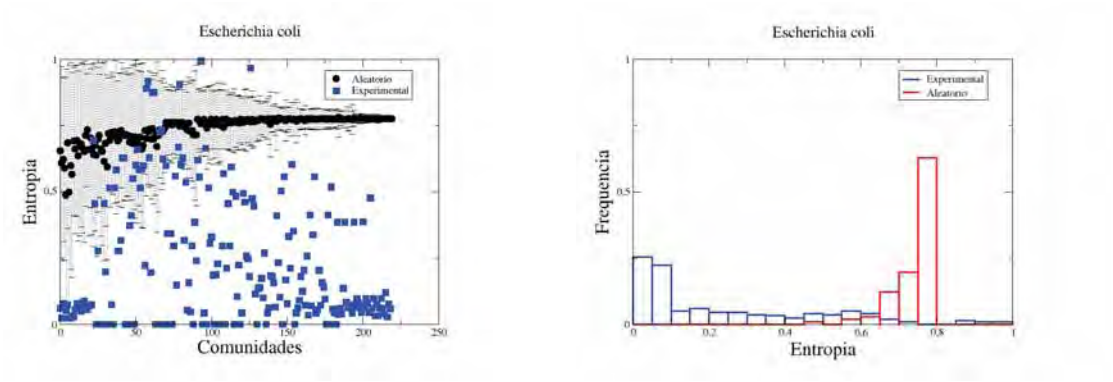


Figura 20: Representação da entropia para as comunidades encontradas pelo método *Clique Percolation* para a rede da bactéria *Escherichia coli*.

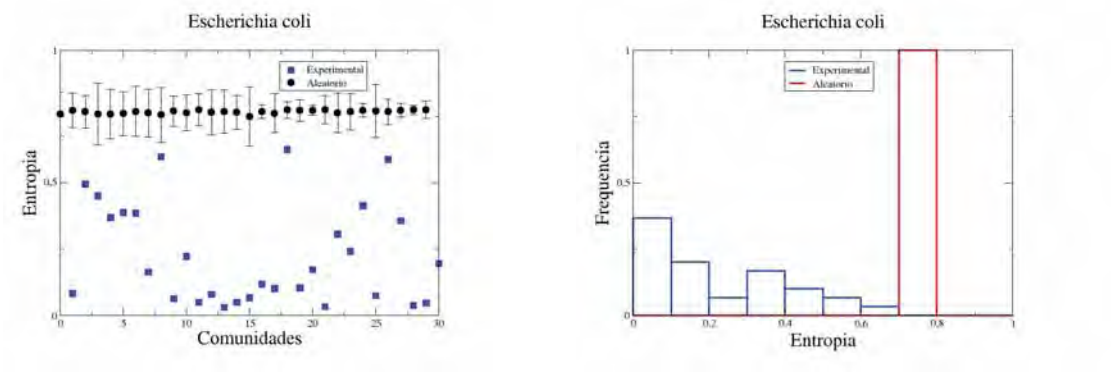


Figura 21: Representação da entropia para as comunidades encontradas pelo método *MinCut* para a rede da levedura *Escherichia coli*.

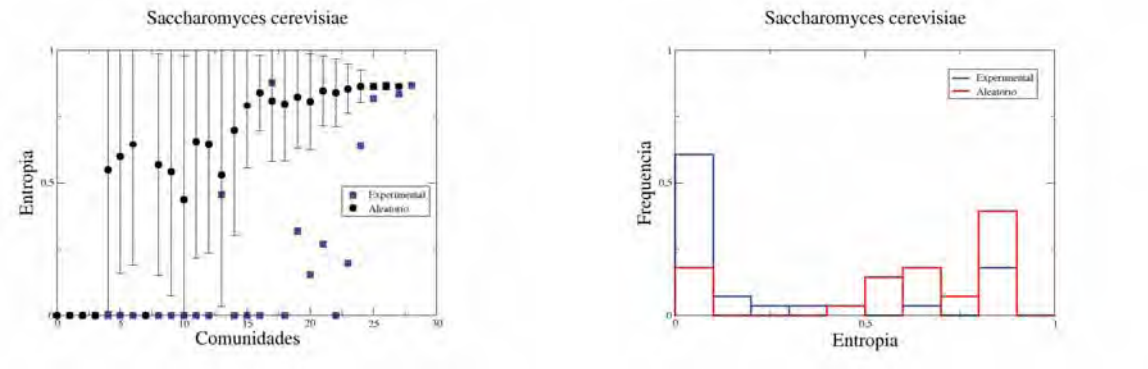


Figura 22: Representação da entropia para as comunidades encontradas pelo método proposto por *Clauset* para a rede da levedura *Saccharomyces cerevisiae*.

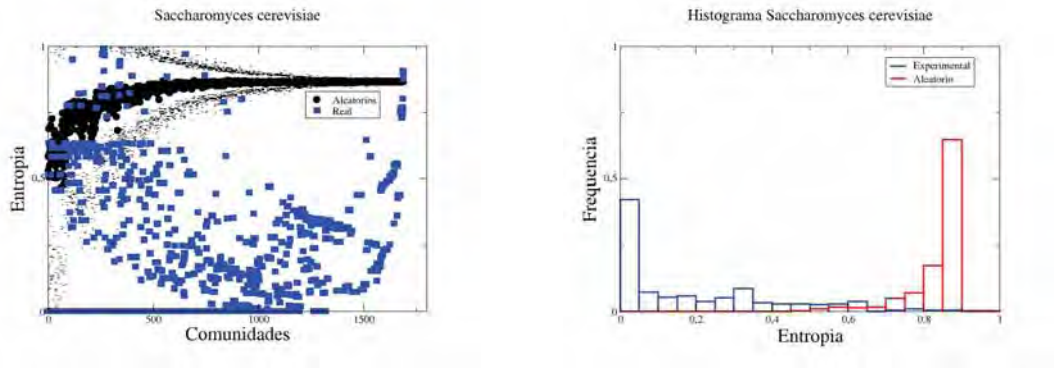


Figura 23: Representação da entropia para as comunidades encontradas pelo método *Clique Percolation* para a rede da levedura *Saccharomyces cerevisiae*.

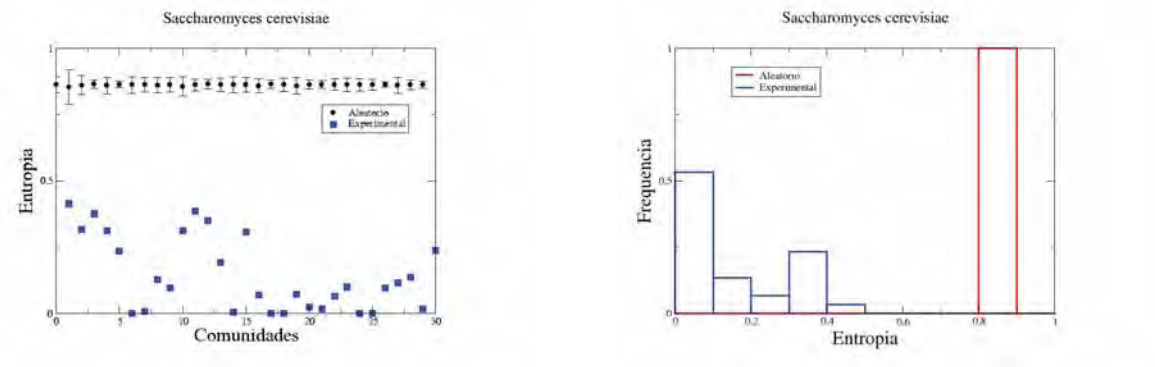


Figura 24: Representação da entropia para as comunidades encontradas pelo método *MinCut* para a rede da levedura *Saccharomyces cerevisiae*.

A fim de investigar uma possível relação entre grupos funcionais e a estrutura de comunidades em redes com predomínio de um tipo de interação, foram medidas as entropias dos agregados identificados na rede da *S. cerevisiae* (GE et al., 2001). Essas medidas foram realizadas porque, em geral, genes coexpressos estão envolvidos em um mesmo processo biológico, seja direta ou indiretamente, e os resultados estão apresentados na Figura 25. Como pode ser observado, os valores obtidos não possuem o mesmo padrão dos obtidos anteriormente para as divisões em comunidades, pois o histograma mostra que os agregados possuem distribuições distintas das frequências dos tipos de interações, o que indica ausência de um tipo predominante de interação.

Enfim, nota-se o predomínio de um tipo de interação para as comunidades indenpendente do método utilizado neste trabalho, mas não para os agregados. Portanto, o fato de uma comunidade possuir um tipo predominante de interação não indica que os genes

presentes nesta se agruparam de acordo com a função na qual estão envolvidos e nem a formação de comunidades como módulos funcionais.

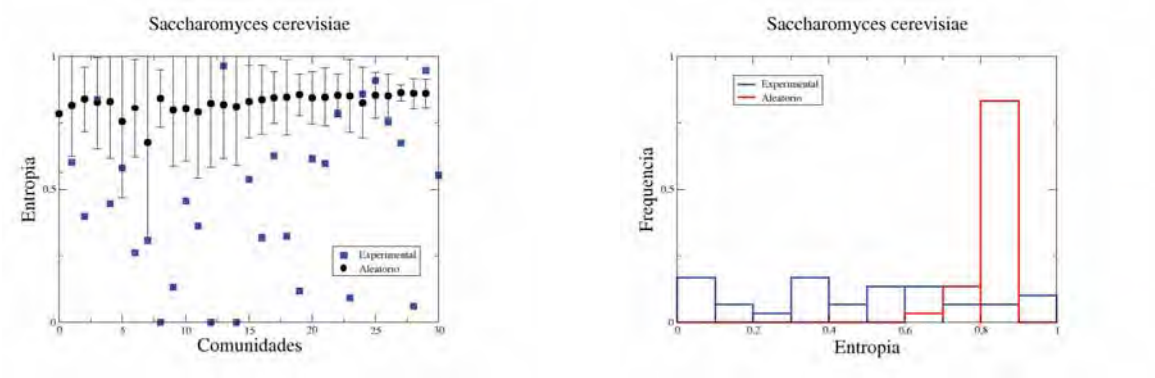


Figura 25: Representação da entropia para os agregados identificados, por coexpressão, na rede da levedura *Saccharomyces cerevisiae*.

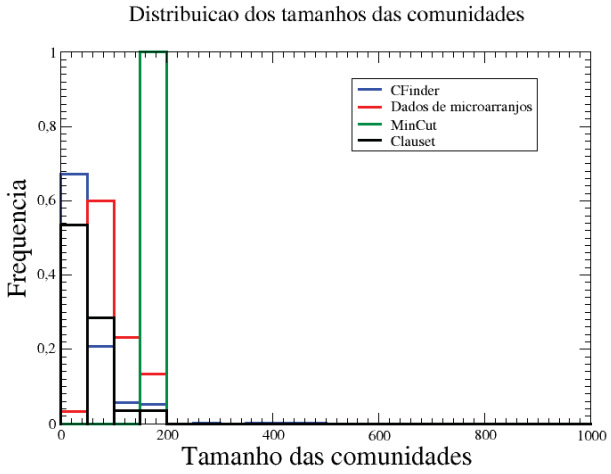


Figura 26: Histograma dos tamanhos das comunidades, obtidas por diversos métodos, para a rede da levedura *Saccharomyces cerevisiae*.

Neste trabalho foram determinadas as estruturas de comunidades para as redes da bactéria *Escherichia coli* e da levedura *Saccharomyces cerevisiae*. Podemos concluir que métodos distintos de divisão das redes em comunidades geram divisões diferentes das redes, com números distintos de comunidades e do tamanho destas. Porém, apesar destas diferenças pode-se afirmar que para as comunidades identificadas, independentemente do método utilizado, há o predomínio de um tipo de interação entre os genes, em todos os

casos essa organização é muito maior do que a obtida em versões aleatorizadas das redes. Finalmente as comunidades obtidas através da co-expressão em *S. cerevisiae* diferem bastante das comunidades obtidas usando métodos baseados exclusivamente em métodos topológicos.

2.5 Caracterização da Rede Integrada de Humanos

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.

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.

2.5.1 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 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 vértices com k conexões, $C(k)$.

Como pode ser observado na Figura 27, 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 $\gamma \approx 1,6$. 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).

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 27: 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

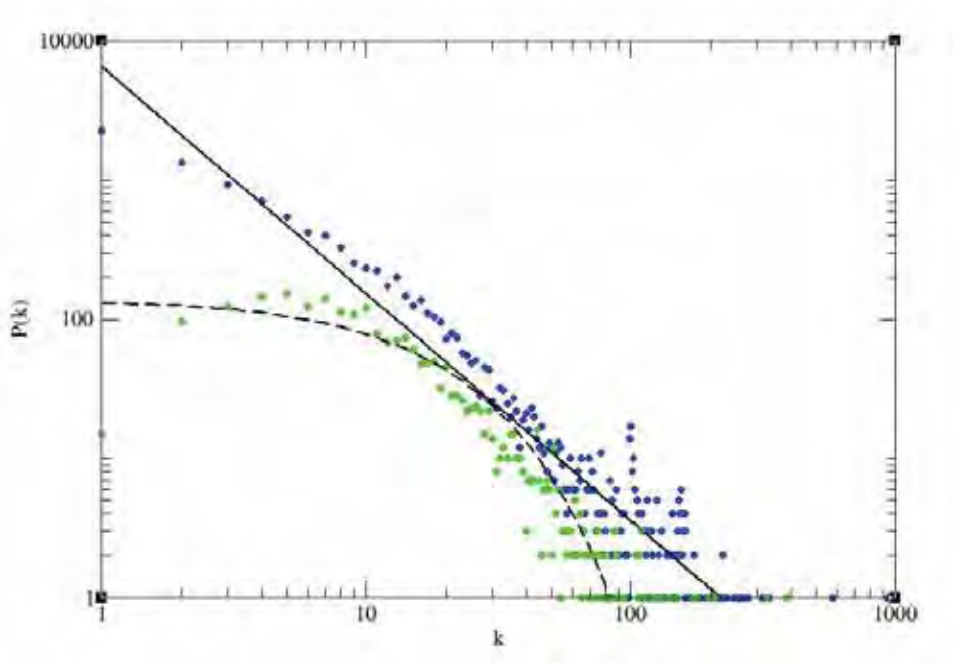


Figura 27: Distribuições dos graus de conectividade, $P(k)$, da $RIGH$ e da G_{ccam}. A distribuição da $RIGH$ (círculos azuis) parece seguir uma lei de potência $P(k) = Ak^{-\gamma}$ com $\gamma \approx 1,6$, o que a caracteriza como uma rede livre de escala. Já a distribuição da G_{ccam} (círculos verdes) não segue uma lei de potência.

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 vértices com k conexões, $C(k)$, podemos observar na Figura 28 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).

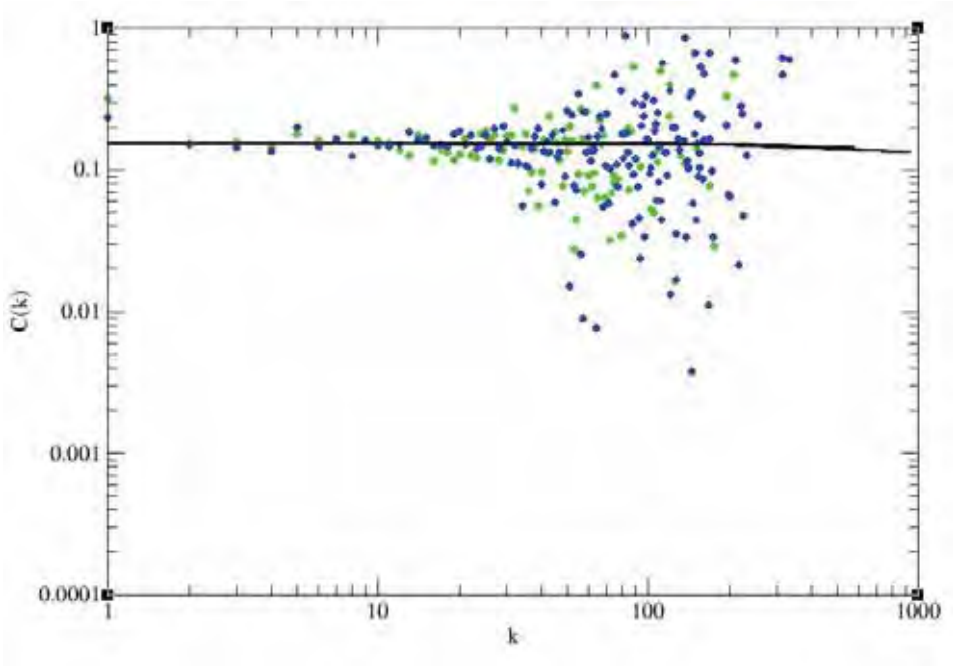


Figura 28: Distribuições dos coeficientes de agrupamento médios $C(k)$ em relação à conectividade k. Pode-se observar que o $C(k)$ tende a ser constante em ambas as redes (círculos azuis: $RIGH$; círculos verdes: G_{ccam}), o que indica que as duas não são hierárquicas.

A organização em larga escala das redes biológicas é ainda um tópico sob intensa investigação. A visão inicial de redes livres de escala organizadas de forma hierarquica parece ser simples em demasia para caracterizar de fato as redes observadas experimentalmente.

Para as redes integradas estudadas com excessão da rede integrada da *E. coli* a função $P(k)$ é descritas por uma lei de potência com *cutoff*. Isto parece indicar que existe um limite físico para o número de interações que um gene pode comportar.

As redes integradas investigadas possuem elevado número de triângulos em comparação com redes aleatorizadas, além disso parecem estar organizadas em módulos compostos por um único tipo de interação, fracamente conectados entre si.

Uma questão interessante é sobre a estrtura hierarquica das redes integradas, as curvas $C(k)$ possuem um comportamento complexo e além disso diferentes métodos computacionais encontram conjuntos de comunidades bastante distintos. Lançando dúvidas se de fato essas redes possuem uma estrutura modular mais fina do que a formada por grandes comunidades formadas apenas por genes interligados por um único tipo de interação.

No próximo capítulo vamos iniciar uma investigação sobre os vértices e que tipo de

informações biológicas a estrutura topológica local pode nos indicar.

3 *Caracterização Local de Redes Biológicas*

Nesse capítulo vamos apresentar a parte mais significativa de nossa produção na área. O nosso objetivo principal nesses trabalhos foi o de relacionar características topológicas e outras propriedades dos genes e de seus produtos com alguma propriedade fenotípica dos organismos. Visando entender como as redes gênicas determinam o comportamento dos organismos.

3.1 Dano e Essencialidade

Genes essenciais são genes que são indispensáveis para que os organismos permaneçam vivos e sejam capazes de se reproduzir em um meio de cultura rico com diferentes nutrientes e estável. Portanto, as funções codificadas por este genes podem ser consideradas essenciais para a fundação da própria vida (KOBAYASHI; EHRLICH, 2003; ITAYA, 1995).

A identificação de genes essenciais é importante não só para a compreensão da requisitos mínimos para a vida celular, mas também para finalidades práticas. Por exemplo, uma vez que a maioria dos antibióticos visam interromper processos celulares essenciais, gene essenciais das células microbianas são promissores alvos para drogas (AKERLEY et al., 2002).

A descoberta de genes essenciais é realizada por procedimentos experimentais, como gene *knockouts* (GIAEVER et al., 2002), interferência de RNA (CULLEN; ARNDT, 2005) e *knockouts* condicionais (ROEMER et al., 2003), mas cada uma destas técnicas requer um grande investimento de tempo e recursos. Considerando estas restrições experimentais, uma abordagem computacional capaz de predizer com precisão a essencialidade de genes seria de grande valor.

3.1.1 Dano em *E. coli*

Nossa primeira abordagem (LEMKE et al., 2004) para esse problema consistiu em recolocar a pergunta (JEONG; MASON; BARABÁSI, 2001) no contexto de redes metabólicas envolvendo apenas enzimas. A nossa percepção sobre o trabalho de (JEONG et al., 2000) era de que as conclusões não eram relevantes do ponto de vista biológico, pois os metabólitos de baixo peso molecular não eram alvos da seleção natural.

Nós investigamos a relação entre o dano d (ver sessão 2.1) na rede metabólica construída seguindo os passos descritos na seção 1.2. e a fração f de genes essenciais. Essa fração é definida por:

$$f = \frac{\text{número de genes cuja deleção causa dano } d}{\text{número de genes}} \tag{3.1}$$

Através de um teste-f obtivemos uma correlação estatisticamente significativa entre essas quantidade com um P-value de 0,0228. A Figura 29 mostra os resultados.

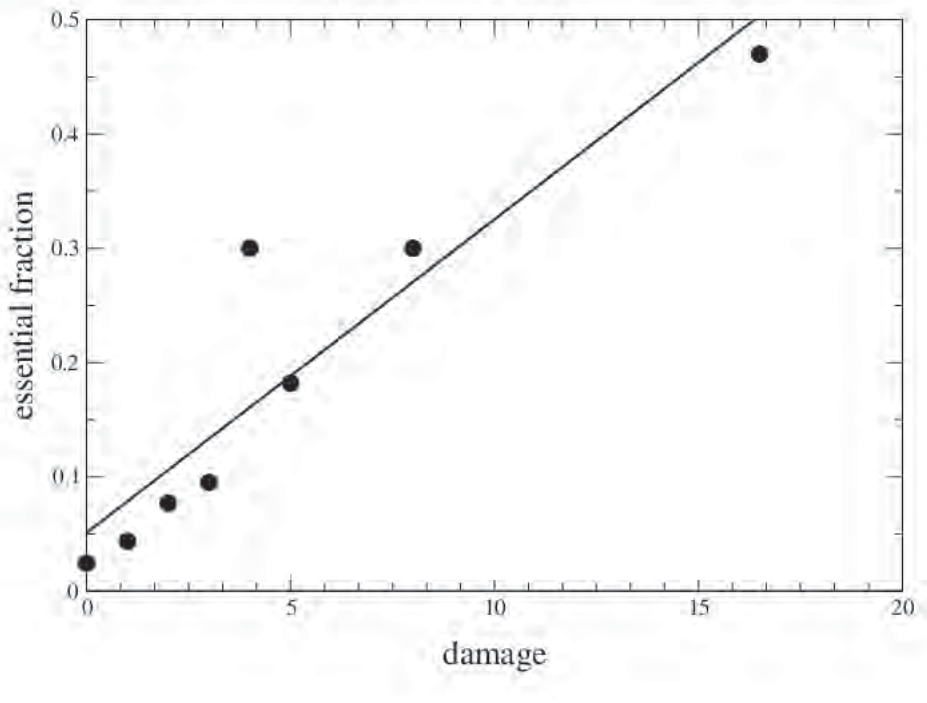


Figura 29: Relação entre dano e fração de genes essenciais. A reta representa o ajuste linear.

Enzimas associadas com danos elevados estão envolvidos com a produção de com-

postos com baixa conectividade mas que ligam importantes partes do metabolismo. Por outro lado, compostos altamente conectados tendem a ser redundantes, uma vez que são produzidos por muitas reações. Surpreendentemente, algumas enzimas essenciais possuem um valor de d baixo e ao contrário, algumas enzimas não essenciais podem causar sérios danos.

No caso de enzimas essenciais com dano baixo, analisamos o bibliografia fornecida pelo banco de dados do PEC (PROFILING...,) , e verificou-se que a maioria pode estar envolvido em outras importantes funções biológicas além do metabolismo de moléculas pequenas. Enzimas não essenciais com alto dano têm influência restrita a um módulo de metabolismo que é não é essencial no ambiente ideal em que crescem as bacterias onde os genes essenciais são determinados.

Nosso próximo paso foi investigar se a conceito de dano também poderia se aplicar em outros contextos biológicos como redes de interações físicas entre as proteínas. Na próxima subsessão investigamos esse cenário.

3.1.2 Dano em S. cerevisiae

As redes de interação física entre proteínas são consideradas robustas frente a perturbações externas, desde que as perturbações não atinjam *hubs*, neste caso podemos esperar pertubações severas na rede. A medida de dano introduzida na sessão 2.1 fornece uma medida quantitativa para essa perturbação.

Se o funcionamento de um célula for de fato um todo integrado, genes isolados deixariam de influenciar o funcionamento da rede. Os dados sobre essencialidade na levedura forneciam uma boa oportunidade para testar esse cenário (SCHMITH et al., 2005).

Nossa idéia foi testar se a fração de genes essenciais estaria correlacionado com o parâmetro dano para a rede de interações físicas e uma rede integrada proposta por (IDEKER et al., 2001b). Os dados que utilizamos foram:

Ito Os dados produzidos a partir de experimentos de dois híbridos. Nesta experimentos pares de proteínas a ser testada para a interação são expressos como proteínas de fusão (“ híbridos”) em levedura: uma proteína é fundido a uma domínios DNA-binding: o outro a uma transcrição domínios ativador. Qualquer interação entre eles é detectado pelo formação de um fator de transcrição. As experiências obtidas 4549 interações entre proteínas 3.278 (ITO et al., 2001).

Uetz Uetz conjunto de dados também foi obtida utilizando os dois híbridos experimentos, obtiveram 957 interações entre proteínas 1003 (UETZ; IDEKER; SCHWIKOWSKI, 2002). Mesmo que os dois conjuntos de dados foram obtidos utilizando a mesma técnica experimental que eles compartilham uma pequena inesperado número de par de proteínas que interagem: 141.

Mering com confiança alta Mering conjunto de dados foram obtidos utilizando contribuições de três métodos diferentes: o gene conservado bairro, co-ocorrência de genes e eventos de fusão gênica. O conjunto de dados contidos 988 proteínas e 2455 interações.

Mering com confiança de Alta e Média Este conjunto de dados foi obtidos usando os mesmos métodos da outra, mas usando mais condições restritivas para mais detalhes veja (MERING et al., 2002) e material suplementar. O conjunto de dados contidos 2.617 proteínas e 11.855 interações.

Ideker Este conjunto de dados foi proposto por Ideker *et al.* Que utilizada uma abordagem integrada para investigar o metabolismo galactose, utilizando DNA microarrays, proteômica quantitativa e bases de dados de interações físicas das proteínas. O conjunto de dados contém 722 proteínas e 612 interações.

Nós avaliamos e comparamos a relevância biológica do dano e da conectividade medindo a correlação entre o dano d e a viabilidade de o organismo quando a proteína é removido de seu proteoma (JEONG; MASON; BARABÁSI, 2001).

Para determinar se duas quantidades são correlacionados calculamos o função L_d proposta originalmente por Jeong e colaboradores (JEONG; OLTVAI, 2003). Esta função pode ser calculada segundo os passos:

- Ordene os genes em ordem decrescente de d.
- Atribua δ_i=0,1 para cada gene cuja letalidade é conhecida (1 para letal e 0 não letal).
- Determine a curva $L_d(R)$ somando δ_i para cada gene
- Normalize $L_d(R)$ dividindo pelo número de genes essenciais.
- Plote L_d contra R/N.

Se os dados não são correlacionados $L_{d,r}$ é uma linha reta, se os dados foram correlacionados positivamente ou negativamente, vamos obter uma curva côncava ou convexa, respectivamente. Para comparar a correlação entre os diferentes parâmetros e conjuntos de dados, integramos as curvas $L_{d,r}$ e determinamos $S_{d,r}$.

Calculamos o dano como uma medida diferente topológica do influência de uma determinada proteína na rede. Ele calcula o número de proteínas que estão desconectados da rede quando a proteína em estudo é excluída. Supondo que a informação é distribuída através da rede, proteínas desconectado param de compartilhar informações com o resto da rede. Sob essa perspectiva, mesmo proteínas com poucas conexões podem causar grandes danos ao serem removidas.

Assumindo que a topologia livre de escala da rede PPI é uma estratégia evolutiva para garantir a robustez do sistema. Nós esperamos que o parâmetro dano seja capaz de identificar as proteínas mais importantes para o funcionamento do sistema.

Para comparar a relevância biológica da conectividade e do dano determinamos qual deles possuem maior correlação com essencialidade. Os resultados são apresentados na Figura 30. Também estudamos os conjuntos Uetz e Ito com e sem as enzimas. Isto foi feito porque a relevância de enzimas pode estar relacionado não só com propriedades topológicas, mas também com o papel que desempenham na rede metabólica.

Nós medimos S para cada curva, os resultados são apresentados na Tabela 9, quanto maior o desvio de 0,5 maior a correlação entre duas grandezas. Com base nesta análise podemos concluir que:

- Connectividade está correlacionada com a essencialidade para todos os casos, exceto para o conjunto Ideker.
- A exclusão de enzimas a partir da construção de L implica uma aumento da correlação, mas o ganho é pequeno.
- No conjunto de dados de Mering S_d aumenta se considerarmos a conjuntos de dados com alta confiança sets, enquanto S_k diminui.

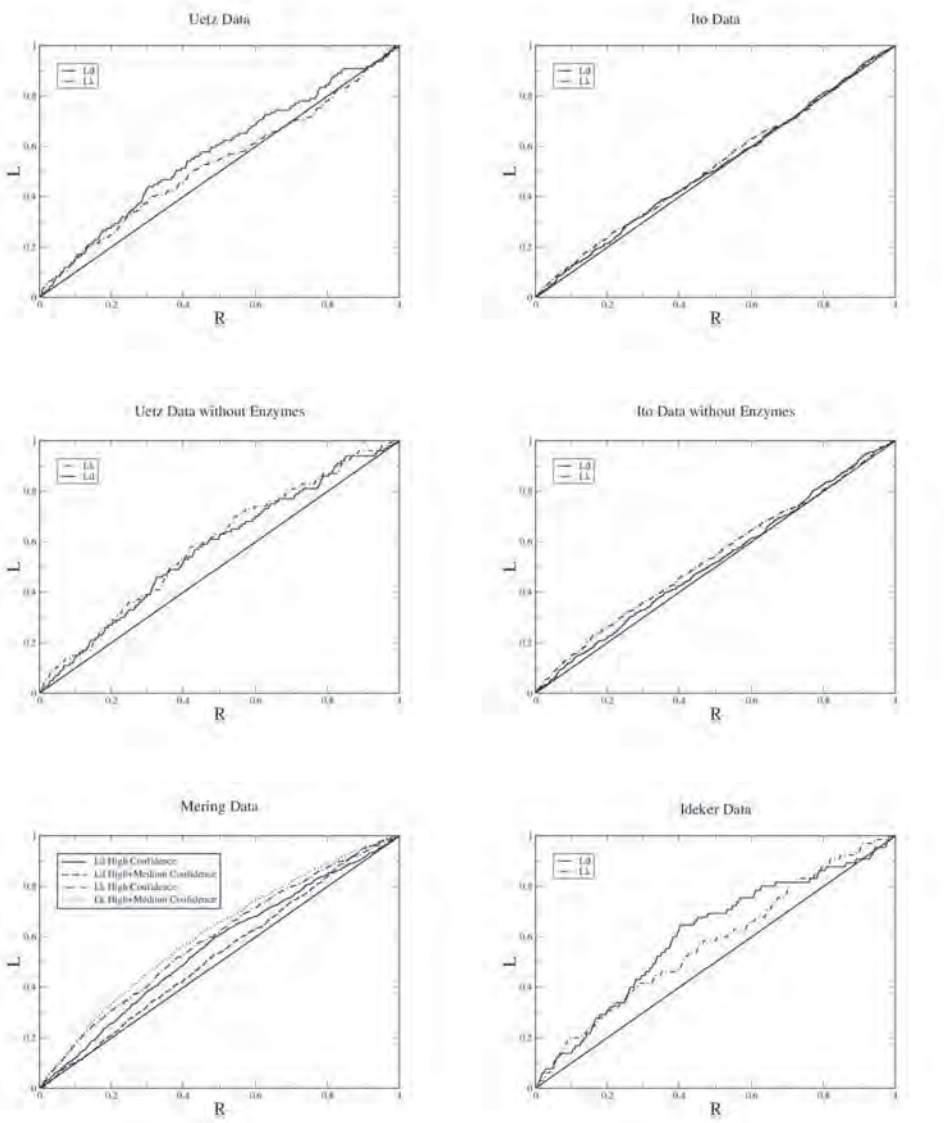


Figura 30: A função L_d e L_k contra a classificação normalizada para os dados Uetz sem enzimas. A curvatura indicam uma muito fraca correlação entre os dois conectividade e os danos letalidade.

Banco de dados	S_d	S_k
Mehring Alta confiança	0,56	0,59
Mehring Confiança alta e média	0,53	0,61
Ito	0,51	0,52
Ito com Enzimas	0,52	0,54
Uetz	0,53	0,57
Uetz com Enzimas	0,57	0,59
Ideker	0,60	0,59

Tabela 9: Comparação dos dados de diferentes conjuntos de proteína-proteína interações, S mede a correlação entre um topológica parâmetro com a letalidade, para um parâmetro não correlacionadas $S = 0{,}5$, neste cenário valores maiores de S indicam forte correlação, ver texto para detalhes.

Os resultados mostram conclusivamente que tanto dano e conectividade são correlacionada com a essencialidade. Se considerarmos apenas as redes PPI, conectividade apresenta maior correlação com a essencialidade do que o dano, no entanto, para redes integradas, como o conjunto de Ideker, o dano apresenta maior correlação. A interpretação desses resultados não é totalmente claro e propomos duas hipóteses:

- A influência de uma proteína é localizada e não se espalha através da rede que está sendo concentrado para as proteínas a que interage diretamente.
- O dano é muito sensível a erros nos conjuntos de dados.

Apesar de existir correlação estatisticamente significativa entre os dados, ela possui um poder explicativo muito limitado. Outro aspecto importante é que o dano, não estava mais correlacionado do que a conectividade. Ou seja as influências dos genes sobre outros genes não parecia influir em genes topologicamente mais distantes. Os resultados obtidos usando os dados de Ideker indicavam que a inclusão de outras formas de interação poderiam ser importantes.

Mas a inclusão de mais dados implicaria que abordagens mais sofisticadas deveriam ser utilizadas. O problema que estavamos considerando é um problema de classificação e existem muitos algoritmos na área de Aprendizado de Máquina eficazes para abordar esse problema.

Ou seja se pretendíamos avançar neste problema deveríamos usar dados que incluíssem

mais informações sobre as interações e algoritmos mais sofisticados que nos permitissem determinar o que torna um gene essencial.

3.2 Aprendizado de Máquina

Nosso próximo passo foi então montar a rede integrada para a bactéria *E. coli* e com base nas informações sobre a topologia determinar quais eram as condições para que um gene fosse essencial (SILVA et al., 2008).

3.2.1 Aprendizado de Máquina em *E. coli*

Nossa metodologia foi implementada utilizando o aplicativo WEKA (*Waikato for Knowledge Analysis*) (WITTEN, 2005; HALL et al., 2009). O aplicativo WEKA é uma coleção de algoritmos de aprendizagem de máquina e mineração de dados que contém ferramentas para pré-processamento de dados, classificação, regressão, clusterização, associação de regras, e visualização.

Dentre os possíveis algoritmos, foi utilizado o algoritmo J48, implementação do WEKA do algoritmo C4.5 (QUINLAN, 1993) que gera árvores de decisão para o problema de classificação. A árvore de decisão do modelo é construído através da análise de dados de treino e o modelo é avaliado sendo aplicado para um conjunto de dados de teste.

Nós treinamos o algoritmo J48 com conjuntos de de treinamento que eram um conjunto de genes essenciais e não essenciais da *Escherichia coli* retirado do banco de dados do PEC(PROFILING...,). Em todas as configurações de treinamento, para um determinado gene, o aprendizado os atributos utilizados foram os seguintes:

- número de interações físicas para a proteína codificada pelo gene;
- número de genes alvo transcricionalmente regulados pela fator de transcrição correspondente codificado pelo gene (`regulacao-out`);
- número de fatores de transcrição que regulam o gene; (`regulação-in`);
- número de enzimas que usam metabólitos produzidos pelo enzima correspondente como reagentes (`metabolismo-out`);
- número de enzimas que produzem metabólitos usados como reagentes pela enzima correspondente (`metabolismo-in`);

O algoritmo J48 foi avaliado usando validação cruzada (veja o apêndice B). O desequilíbrio de dados é uma das causas que degradam o desempenho de algoritmos de aprendizado de máquina (CHO, 2006). Para minimizar esses problemas repetimos os dados relacionadas com os genes essenciais para corrigir o desbalanceamento dos dados, pois o número de genes não-essenciais é muito maior que o número de genes essenciais.

O desempenho do método foi avaliado testando os classificadores criados pelo algoritmo J48, como descrito acima, sobre os dados de treinamento em si. A seleção da melhor configuração de treinamento foi realizada com base na *F-measure* do classificador gerado, este parâmetro fornece uma média harmônica da *precision* e do *recall* e é definido como:

$$F = \frac{2 \times precision \times recall}{(precision + recall)} , \tag{3.2}$$

onde *precision* é o percentual de casos corretamente classificados e *recall* a percentagem de casos positivos que foram rotulados classificado como tal. Estes indicadores foram calculados a partir de matrizes confusão do classificadores obtidos a partir do treinamento de diferentes conjuntos de dados.

O melhor resultado obtido através de nossa metodologia foi uma *F-measure* de 83,4% para genes essenciais e 79,7% para genes não-essenciais genes. Figura 31 mostra o conjunto de regras do árvore de decisão gerada. O nó superior da árvore corresponde às interações físicas entre proteínas. isso significa que que o algoritmo de árvore de classificação concluiu que a melhor maneira de explicar essencialidade em *E. coli* foi considerar o número de interações físicas entre proteínas. O grau de uma proteína tem sido documentada na literatura como indicativo de essencialidade em vários organismos (JEONG et al., 2001).

Em nossa abordagem, uma combinação de um número intermediário de interações físicas com pelo menos uma interação do tipo `metabolismo-in` é um indicativo de essencialidade. Genes com uma interação do tipo `regulação-out` foram classificados como não-essenciais.

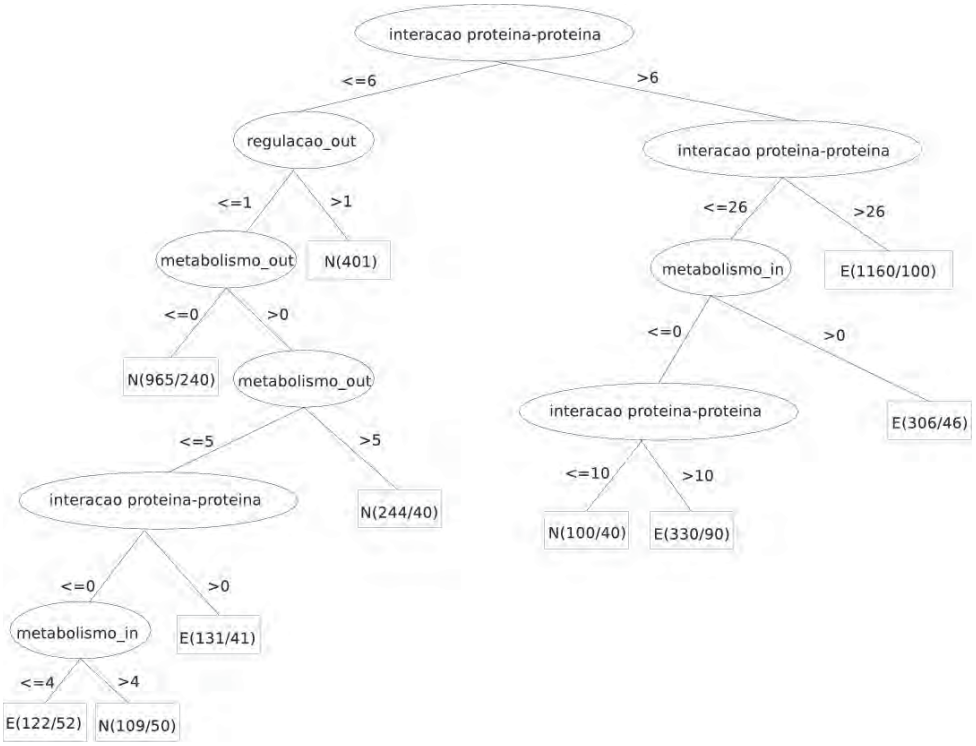


Figura 31: Árvore de decisão gerada pela aplicação do algoritmo J48 para classificação de genes essenciasi em ıE. coli com *F-measure* de 83.4% para genes essenciais.

Os resultados obtidos nesse trabalho indicava que a abordagem desse problema usando ferramentas de Inteligência Artificial permitia uma compreensão mais aprofundada do problema. Ressalto que o fato de que genes com um maior número de interações regulatórias fosse não essencial indicava que *hubs* não seriam necessariamente os genes mais importantes. Além disso os resultados mostravam que o peso de cada uma das interações na determinação da essencialidade era diferente.

A abordagem proposta era bastante simples, nós utilizamos um algoritmo que gerava classificadores com baixo desempenho e o método de balancemanto das classes acabava incrementando de forma espúria o desempenho do classificador. Ainda existiam alguns problemas que estavam diretamente relacionados com a bactéria *E. coli* os dados de interação física eram incompletos. Ou seja havia espaço para realizar melhoramentos.

Nosso próximo passo foi sanar várias várias dessas deficiências em um novo artigo buscando determinar o que torna os genes essenciais (ACENCIO; LEMKE, 2009).

3.3 Essencialidade na *S. cerevisiae*

Para predição de genes essenciais para a levedura *Saccharomyces cerevisiae*, persquisadores implementaram abordagens computacionais que são baseados em características da sequência de genes e proteínas, com ou sem comparação de homologia (SERINGHAUS et al., 2006; GUSTAFSON et al., 2006). Com o acúmulo de dados experimentais derivados de estudos de pequena escala e de técnicas de *high-troughput*, agora é possível construir redes de interação gênica e investigar se a propriedades topológicas dessas redes seriam úteis para prever a essencialidade de genes. Embora muitos estudos tenham sido realizados para responder essa questão (JEONG et al., 2001; SCHWIKOWSKI; UETZ; FIELDS, 2000 Dec; DUARTE; HERRGARD; PALSSON, 2004; GUELZIM et al., 2002 May; JEONG et al., 2001; PALUMBO et al., 2005 Aug 29), a maioria dos estudos tem se concentrado em responder quais propriedades topológicas isoladas são preditivas de essencialidade.

Neste trabalho investigamos a inter-relação entre a topologia, a função dos genes e a localização celular desempenham para determinar a essencialidade dos genes. Nós consideramos os compartimentos celulares: citoplasma, retículo endoplasmático, mitocôndria, núcleo, ou outras localizações e os processos celulares: ciclo celular, metabolismo, transdução de sinal, transcrição, transporte e outros processos (ACENCIO; LEMKE, 2009).

Outro aspecto que investigamos nesse trabalho foi a relevância de parâmetros topológicos que levassem em conta não apenas o número de interações dos genes, mas também sue papel global na rede. Nós consideramos os parâmetros: grau para cada tipo de interação, coeficiente de clusterização, e as centralidades de proximidade e intermediação para cada tipo de interação e um parâmetro que verificava que indicava quantos genes se relacionavam com os mesmos genes que o gene de interesse (ver sessão 2.1).

Para gerar os preditores de essencialidade foram construídos dois diferentes conjuntos de dados balanceados para treinamento: um conjunto de dez grupos de treinamento contendo as instâncias positivas e negativas corretamente associadas às suas classes que, nesse caso, são a classe "essencial" (*essenc*) e a classe "não essencial" (*nessenc*). Esses grupos de treinamento são ditos "balanceados" por que eles contêm o mesmo número de genes essenciais e não essenciais. Essa estrategia depende da escolha dos conjuntos de treinamento, para validação estatística devemos repetir esse processo várias vezes e tomar médias das quantidades de interesse.

Outra mudança introduzida nesse trabalho foi comparar o desempenho dos classificadores utilizando a métrica AUC (*Area Under Curve*) que basicamente indica qual é a

probabilidade de dado uma instância qualquer ela seja classificada corretamente (ver o apêndice B).

Nesse trabalho nós aplicamos uma estratégia mais avançada para melhorar o desempenho dos classificadores. Nessa estrategia inicialmente determinamos vários classificadores com bom desempenho, utilizamos então para todos eles um meta-classificador chamado de *bagging* que aprende com os erros do classificador melhorando ainda mais o desempenho individual desses classificadores. Posteriormente integramos os resultados de todos os classificadores usando outro meta classificador chamado de *Vote*. Os classificadores usados foram: (1) REPtree (WITTEN; FRANK, 2000), (2) naive bayes tree (KOHAVI, 1996), (3) random tree (WITTEN; FRANK, 2000), (4) random forest (BREIMAN, 2001), (5) J48, (6) best-first decision tree (SHI, 2007, The University of Waikato), (7) logistic model tree (LANDWEHR; HALL; FRANK, 2005) and (8) alternating decision tree (FREUND; MASON, 1999).

Neste trabalho realizamos uma análise exaustiva da influência dos diferentes parâmetros na capacidade preditiva do classificador. Nessa análise comprovamos que a localização celular e os processos biológicos isolados são preditores razoáveis para a essencialidade. A conclusão mais relevante do trabalho é que a integração de todos esses elementos gera classificadores com melhor desempenho AUC=0,808, do que se tomarmos os parâmetros de forma isolada.

Além da capacidade de previsão, as técnicas de aprendizagem de máquina podem ser usadas para aquisição de conhecimento. Portanto, a fim de descobrir as regras para a essencialidade dos gene em *S. cerevisiae*, analisamos árvores de decisão geradas pelo algoritmo J48. Como as árvores de decisão geradas a partir de diferentes conjuntos de dados equilibrados produzem árvores levemente diferentes umas das outras. Tivemos de analisar alguns exemplos para podermos ser capazes de elaborar as regras gerais para a essencialidade dos genes.

A Figura 32 mostra a árvore de decisão que melhor ilustra as regras gerais para a essencialidade. Como pode-se observar na Figura 32, o nó raiz da árvore de decisão é o número de interações físicas por isso, este atributo pode ser considerado a característica mais importante entre as analisadas. O preditor contendo apenas o número de interações física como recurso de treinamento é o que melhor prediz (AUC = 0,747) essencialdade. Isto está em sintonia com estudos anteriores que têm demonstrado que o número de interações física é indicativo da essencialidade (JEONG et al., 2001; ESTRADA, 2006; WUCHTY, 2004).

Vários hipóteses sobre a ligação entre essencialidade e número de interações físicas têm

sido propostas. Coulomb *et al.* (COULOMB et al., 2005) têm sugerido que a relação entre esse atributo e essencialidade é em parte devido a vieses nos dados de interação que são enriquecido em experimentos de pequena escala. Por outro lado, Zotenko *et al.* (ZOTENKO et al., 2008 Aug) propôs recentemente a hipótese de que a conexão essencialidade entre gene e número de interações físicas é provavelmente devido ao envolvimento de proteínas codificadas por genes essenciais em sub-redes de proteínas densamente conectados com funções biológicas que são ricas em proteínas codificadas por genes essenciais.

Seguindo o caminho do nó raiz ao nó na primeira folha o através do ramo direito, encontramos a seguinte regra para essencialidade: se as proteínas interagem com mais de 7 outras proteínas e estão localizados no núcleo, os genes que codificam são classificados como essenciais. Isso nos fornece outros elementos para se compreender por que genes com muitas interações físicas tendem a ser essenciais. Esses genes em muitos casos estão envolvidos com operações chaves como a replicação da fita do DNA, que são realizados por estruturas moleculares grandes que acabam sendo compostas por muitas proteínas que devem interagir. isso explica por que o atributo mais importante é de fato o grau do vértice e não sua centralidade.

O caminho do nó raiz para os nós folheares no ramo esquerdo (Figura 32) nos levam a descobrir uma outra regra para o gene essencialidade: se as proteínas interagem fisicamente com 6 ou menos proteínas e participam de um processo metabólico no interior do núcleo, os genes que codificam essas proteínas são tendem a ser essenciais.

De acordo com estas regras, a condição final para essencialidade é a localização de proteínas no núcleo, sugerindo que este componente celular é de alguma forma importante para a essencialidade. A importância do núcleo para a essencialidade também sido sugerido por Seringhaus *et al.* (SERINGHAUS et al., 2006) que mostraram que a localização nuclear tem maior correlação positiva com a essencialidade se comparado com outros componentes celulares. A relação entre núcleo e essencialidade pode ser explicada pelo o fato de que cerca de um terço das proteínas nucleares são codificados por genes essenciais e mais processos biológicos essenciais para a viabilidade celular ocorrem dentro do núcleo (KUMAR et al., 2002).

Descobrimos uma regra adicional interessante para essencialidade gene em levedura: genes regulados por mais de 3 fatores de transcrição tendem a ser não-essenciais. Nossa conclusão é corroborada por Yu *et al.* (YU et al., 2004 Jun) que descobriram que entre os genes regulados por fatores de transcrição há proporcionalmente um menor número de genes essenciais. Isto pode ser explicado pelo fato de que os genes codificam proteínas

envolvidas com *house-keeping* tais como processamento metabólico de mRNA e iniciação de transcrição (ZOTENKO et al., 2008 Aug) não são regulados.

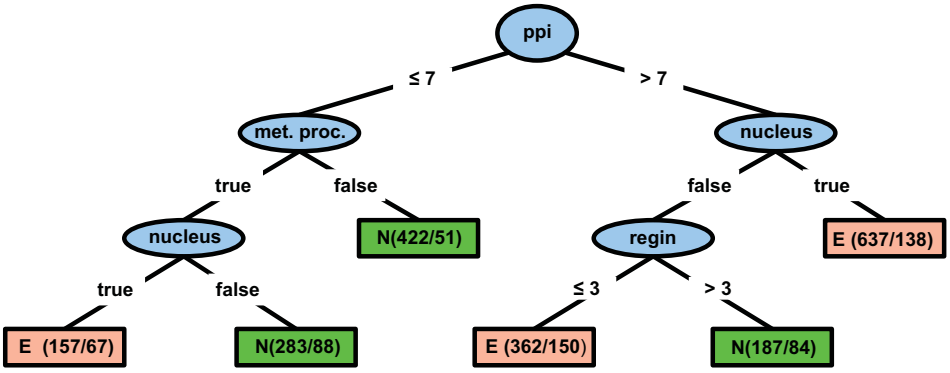


Figura 32: Árvore de decisão foi gerado pelo treinamento do algoritmo J48 na conjunto de dados balanceado com todos os dados disponíveis. A elipse superior é o nó raiz da árvore que representa a condição mais importante para discriminar genes essenciais de genes não-essenciais. Neste caso, tal condição é o número de interações proteína física (*ppi*). As elipses remanescentes são nós internos que representam condições suplementares para um gene ser considerado como essencial ou não essencial. No ramo esquerdo de árvore, tais condições são o envolvimento em um processo metabólico (*met-proc*) e localização nuclear (*nucleus*). No ramo direito, tais condições são a localização nuclear (*nucleus*) e o número de fatores de transcrição que regulam o gene (*regin*). Os retângulos são os nós folha que representam a classificação final. Retângulos vermelho e verde retratam genes que, sob certas condições (representado pelo nó raiz e nós internos nós), são classificados como essenciais (**E**) e não-essenciais (**N**). No parênteses dentro dos rectângulos, o número antes da barra indica a quantidade de genes que são realmente essenciais ou não essenciais e os número depois da barra indica quantos genes foram corretamente previstos.

Em linhas gerais esse trabalho ilustra as potencialidades do uso de ferramentas de aprendizado de máquina em Bioinformática e sua capacidade de gerar hipóteses biologicamente plausíveis.

Nosso próximo passo foi o de testar essa metodologia em outros problemas para verificar sua generalidade com esse objetivo nos voltamos para o problema de classificar genes mórbidos e alvos para droga, para denominar essa característica usamos o neologismo drogabilidade (COSTA, 2009; COSTA; ACENCIO; LEMKE, 2010).

3.4 Morbidade e Drogabilidade

Atualmente, a identificação em larga escala experimental dos genes mórbidos, ou seja, aqueles genes cujas mutações causem doenças hereditárias e genes drogáveis, ou seja, genes que codificam proteínas cuja modulação por pequenas moléculas provocam efeitos fenotípicos, é complexa e muito custosa. A descoberta de genes mórbidos, por exemplo, requer um grande esforço para reunir padrões de herança de famílias com a doença e análises sofisticadas para determinar as mutações envolvidas (KESHAVA PRASAD et al., 2009). De maneira semelhante, a descoberta de novos medicamentos também requer um grande esforço que envolve uma variedade de técnicas oriundas da genômica, proteômica associação genética entre outros (LINDSAY, 2003).

À luz dos fatos mencionados acima, uma abordagem computacional que preveja com alguma precisão genes mórbidos e drogáveis, especialmente em uma escala genômica, seria, portanto, de valor inestimável, pois o número de técnicas experimentais a serem realizados para descobrir esses genes poderia ser minimizada. Com a grande quantidade de dados disponíveis tais como dados de interação molecular e expressão gênica, temos agora a oportunidade para o desenvolvimento de uma abordagem computacional baseada em ferramentas de mineração, tais como aprendizagem de máquina, para extrair padrões que possam ser usados como preditores em escala genômica de genes mórbidos e drogáveis.

Neste sessão utilizamos todos os atributos utilizados na sessão anterior Além disso, foram incluídos dados sobre o número de tecidos onde o gene possui um nível de expressão maior que 5 transcrições por milhão em média (tpm, 32 tecidos estudados), nível de expressão médio nesses tecidos, em tpm (REVERTER; INGHAM; DALRYMPLE, 2008).

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

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

O algoritmo J48 retorna *árvores de decisão* para mostrar quais as regras utilizadas para sua classificação. Para cada conjunto teste, a árvore de decisão possuía detalhamento diferente, mas as Figuras 33 e 34 mostram as árvores mais significativas para cada caso. *Verd.* significa que o algoritmo identificou os genes como portadores da característica (para Figura 33, drogável, e para a Figura 34, mórbido), e *Falso* que indica que o algoritmo identificou os genes como não portadores da característica. Entre parênteses temos o número de acertos seguido do número de erros que o algoritmo cometeu seguindo o critério anterior no fluxograma. Para a Figura 33, temos que o algoritmo considerou necessários os dados de apenas 3 características para fazer a previsão a respeito da drogabilidade: presença da proteína codificada pelo gene na membrana plasmática (*PlasmaMembrane*), o grau de intermediação relacionado a interações regulatórias transcricionais (*InBetReg*) e o número de metabólitos que são utilizados como reagentes em uma reação metabólica catalisada por uma enzima codificada por esse gene (entradas metabólicas, *metin*). Para a Figura 34, foram necessários dados de 5 características: número de fatores de transcrição que controlam o gene (*regin*), grau de intermediação relacionado a interações metabólicas (*InBetMet*), se as proteínas codificadas pelo gene estão presentes no meio extracelular (*Extracellular*) e na membrana plasmática (*PlasmaMembrane*) e seu valor de coeficiente de agregação.

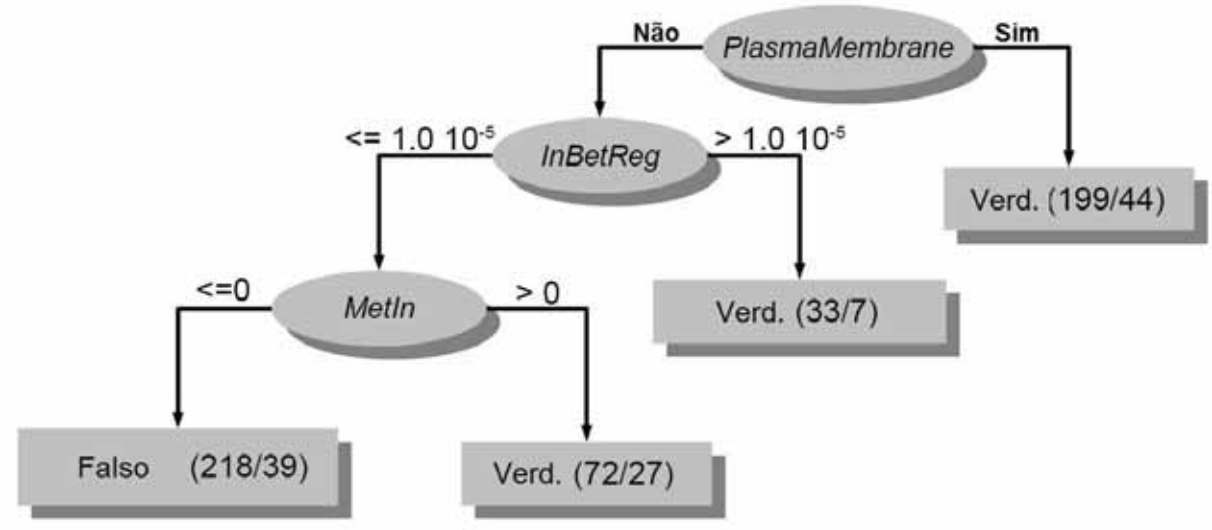


Figura 33: Árvore de decisão para drogabilidade

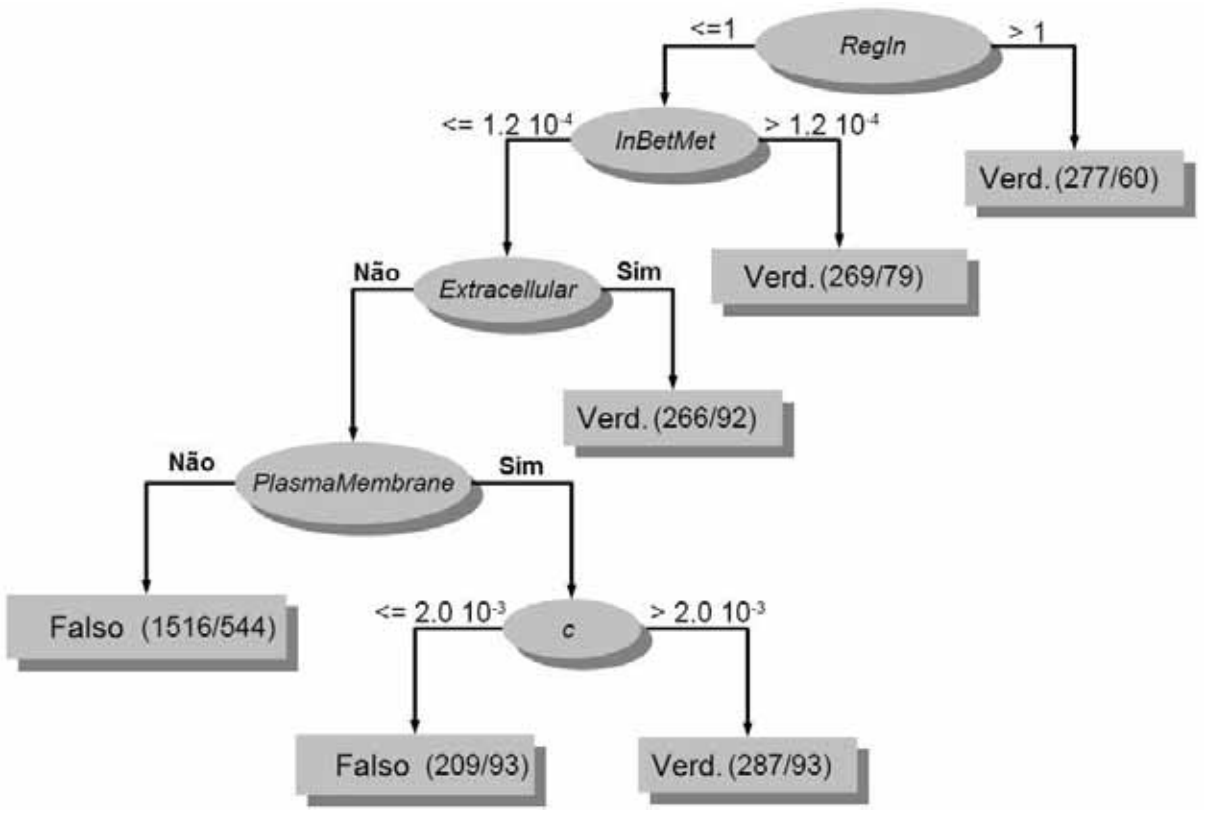


Figura 34: Árvore de decisão para morbidade

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

Tabela 10: Tabela com os resultados obtidos durante o treinamento do modelo.

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

Os resultados obtidos para *recall* e precisão médios dos modelos gerados para determinação da *morbidade* (65% e 66%, respectivamente) indicam grande quantidade de ruído nos conjuntos de teste, devido a possíveis características compartilhadas entre os

genes classificados como "mórbidos" e "não mórbidos", que induziram o classificador ao erro. Isso se deve ao fato de não existir um conjunto negativo para essa classificação, ou seja, não existem evidências de que os genes apresentados como "não mórbidos" possam ser classificados como tal. Logo, possíveis genes mórbidos participaram do conjunto de treinamento.

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

3.5 Predição de alvos para drogas na G_{ccam}

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. 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.

Na tese de Marcio Luiz Acencio propomos um método para detectar vias de sinalização e aplicamos este método para a criação de uma possível rede de sinalização envolvida com a adesão a matriz celular e o controle do ciclo celular. Dada a complexidade dessa metodologia não vamos descrevê-la aqui (interessados podem consultar a referência (ACENCIO, 2010)). Esta sub-rede é chamada de $EGFR - CDC6$.

Dentre os 21 genes que formam a sub-rede $EGFR - CDC6$, 17 podem ser conside-

rados alvos para drogas. 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 35).

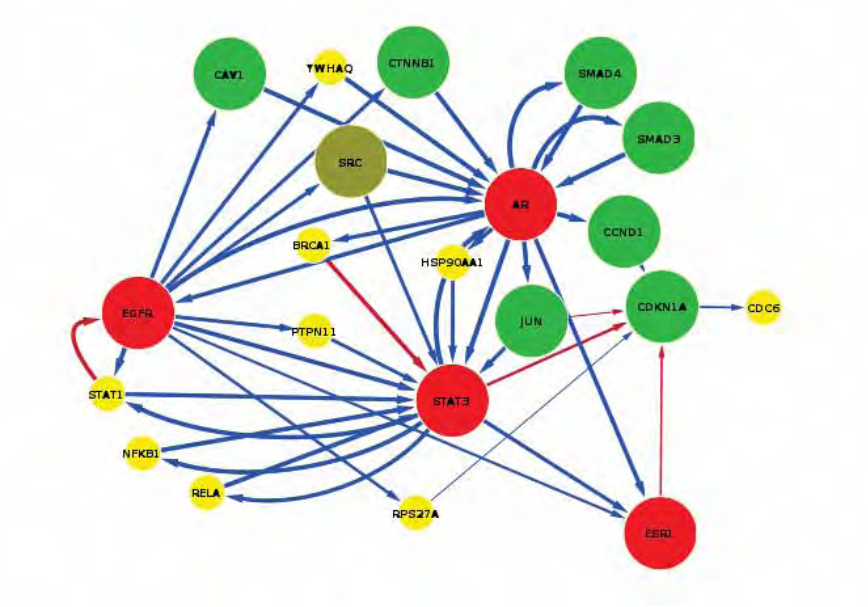


Figura 35: Genes conhecidamente e potencialmente drogáveis na sub-rede $EGFR - CDC6$. Os genes coloridos em vermelho, marrom e verde são, respectivamente, genes conhecidamente drogáveis, genes reconhecidos como oficialmente drogáveis somente após a coleta dos dados para construção dos grupos de treinamento e genes potencialmente drogáveis.

Nesta seção demonstramos que informação topológica associada a informações sobre localização celular e função permitem determinar regras que tornam os genes essenciais, mórbidos ou alvos para drogas. A extração desse conhecimento pode ser realizada utilizando algoritmos oriundos da área de aprendizado de máquina e permitem uma melhor

compreensão sobre os mecanismos moleculares que determinam estas características.

Olhando em retrospecto podemos perceber que as idéias iniciais sobre o papel dos *hubs* no comportamento das redes foram revisados. O número de relações que um gene possui nem sempre é um indicativo de sua relevância, as interações regulatórias por exemplo indicam em geral que os genes são não essenciais. Por outro lado genes regulados em humanos podem estar relacionados com morbidade, isso ocorre por que existe uma complementaridade entre essencialidade e morbidade. A deleção de genes essenciais em humanos, implicaria na morte das células, o que levaria a indivíduos totalmente disfuncionais.

4 Perspectivas

Nesta tese apresentamos as redes biológicas integradas e discutimos sua caracterização local e global e o impacto dessas propriedades nas características biológicas de alguns organismos. Também demonstramos que a aplicação de ferramentas de aprendizado de máquina é útil para analisar esses sistemas e produzir hipóteses biológicas plausíveis.

A investigação sobre a essencialidade em diversos organismos nos mostra que as analogias de sistemas biológicos com sistemas físicos ou objetos tecnológicos pode ser perigosa. Ainda que "hubs" possam ser importantes tanto em redes biológicas como em redes sociais, os motivos que determinam essa relevância podem ser inteiramente diferentes. Acredito que as ferramentas de aprendizado de máquinas são úteis para descobrir estas relações, que não são óbvias.

As abordagens topológicas como as apresentadas nessa tese são instrumentos poderosos, mas certamente não são a única ferramenta disponível. Por outro lado talvez sejam a única abordagem disponível hoje para investigar modelos holísticos para as células, apesar de que as redes biológicas permaneçam ainda bastante incompletos. Esperamos especialmente no caso de dados sobre humanos, um incremento substancial nos próximos anos.

Além disso estão surgindo também novas classes de dados, entre elas dados que comparem fenótipos em diferentes meios. Acredito que a análise destes dados é uma fronteira instigante de pesquisa, pois permitirá elucidar mecanismos de interação entre o meio e as redes biológicas subjacentes.

O manejo dos novos dados vai demandar por estratégias cada vez mais sofisticadas para permitir não apenas sua análise, mas até mesmo sua armazenagem em bancos de dados está se tornando um importante desafio tecnológico. Ferramentas de mineração de dados e aprendizado de máquinas deverão ser cada vez mais presentes nesta área.

Finalmente as abordagens topológicas como as descritas nesse trabalho acabaram sendo rapidamente incorporadas ao nosso cotidiano, não apenas na literatura científica,

mas também na forma como nos relacionamos e como medimos nossa produtividade.

Mas os modelos topológicos certamente são apenas um passo na descrição dos sistemas biológicas. A fronteira nessa área está em propor abordagens integrativas que explorem a intensidade das interações e o estado dos vértices e que consigam gerar modelos que simulem a dinâmica desses sistemas.

Referências

ABDULREHMAN, D. et al. Yeastract: providing a programmatic access to curated transcriptional regulatory associations in saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res*, v. 34, p. D446–451, 2010.

ACENCIO, M. L. *Metodologia para construção de redes moleculares integradas para organismos eucariotos e sua aplicação na construção e análise da rede envolvida com a regulação do ciclo celular pela adesão à matriz extracelular em H. sapiens.* Tese (Doutorado) — Biologia Geral e Aplicada - Instituto de Biociências, 2010.

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.

ADAMCSEK, B. et al. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, v. 22, p. 1021–1023, Apr 2006.

AGRAWAL, H. Extreme self-organization in networks constructed from gene expression data. *Phys. Rev. Lett*, v. 89, n. 26, p. 268702, dez. 2002.

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.

AKERLEY, B. J. et al. A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, v. 99, n. 2, p. 966–971, jan. 2002.

AL, P. G. B. et. An ontology for bioinformatics applications. *Bioinformatics*, v. 15, n. 6, p. 510–520, 1999.

ALBERT, I. et al. Boolean network simulations for life scientists. *Source Code for Biology and Medicine*, v. 3, p. 16, 2008.

ALMEIDA, R. de; LEMKE, N. Stretched exponential relaxation on the hypercube and the glass transition. *The European Physical . . .*, 2000.

ALMEIDA, R. de; LEMKE, N.; CAMPBELL, I. Stretched exponential relaxation and fractal phase space. In: UNIV MONTPELLIER 2, LAB VERRES, F-34095 MONTPELLIER, FRANCE. *Journal of Magnetism and Magnetic Materials*. Univ Montpellier 2, Lab Verres, F-34095 Montpellier, France, 2001. p. 1296–1297.

ALMEIDA, R. M. C. de; LEMKE, N.; CAMPBELL, I. A. Stretched exponential relaxation on the hypercube and the glass transition. *Eur. Phys. J. B*, v. 18, p. 513, 2001.

ANDRADE, T. F. *Análise de propriedades topológicas das redes biológicas integradas da Escherichia coli e da Saccharomyces cerevisiae*. 2008. Trabalho de Conclusão de Curso de Física Médica.

BARABASI, A. *Linked: The New Science of Networks. 2002*. [S.l.]: Cambridge, 2002.

BARABASI, A. Scale-Free Networks: A Decade and Beyond. *Science*, jan. 2009.

BARABASI, A.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, n. 5439, p. 509–512, out. 1999.

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.

BARABÁSI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, v. 5, n. 2, p. 101–13, fev. 2004. ISSN 1471-0056.

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.

BARCELLOS, C.; HERÉDIA, F.; SCHMITH, J. Analysis of Functional Interactions of Enzymes in Mycoplasma pneumoniae. . . . , n. 7674832, p. 199, 2008.

BATTISTELLA, E. et al. An integrated model for cellular analysis. *Genetics And Molecular Research*, v. 4, n. 3, p. 506–513, 2005.

BATTISTELLA, E. et al. Bioinformatics: A Growing Field for Ontologies. In: *Workshop on Ontologies and their Applications*. São Luis, Brasil: [s.n.], 2004. To appear.

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.

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

BOLLOBÁS, B. *Graph Theory: an introductory course*. 1st. ed. New York: Springer Verlag, 1979. ISBN 0387903992.

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

BREITKREUTZ, B.-J. et al. The biogrid interaction database: 2008 update. *Nucleic Acids Res*, v. 36, n. Database issue, p. D637–40, jan. 2008. ISSN 1362-4962.

BUTTE, A. J.; KOHANE, I. S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, p. 418–429, 2000.

CHAVES, M.; ALBERT, R. Studying the effect of cell division on expression patterns of the segment polarity genes. *Journal of The Royal Society Interface*, v. 5, n. Suppl 1, p. S71, 2008.

CHO, S. EUS SVMs: ensemble of under-sampled SVMs for data imbalance problems. *Neural Information Processing*, 2006.

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.

CLAUSET, A.; NEWMAN, M. E.; MOORE, C. Finding community structure in very large networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 70, p. 066111, Dec 2004.

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

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

COSTA, P. R. *Determinação de genes mórbidos e drogáveis a partir da construção e análise da rede integrada de interações moleculares entre genes humanos*. 2009. Trabalho de Conclusão de Curso de Física Médica.

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 Suppl 5, p. S9, 2010.

COULOMB, S. et al. Gene essentiality and the topology of protein interaction networks. *Proc Biol Sci*, v. 272, n. 1573, p. 1721–5, 2005.

CULLEN, L. M.; ARNDT, G. M. Genome-wide screening for gene function using RNAi in mammalian cells. *Immunology and cell biology*, v. 83, n. 3, p. 217–223, jun. 2005.

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

DOROGOVTSEV, S. *Evolution of networks*. [S.l.]: Advances in Physics, 2002.

DOROGOVTSEV, S. N.; MENDES, J. F. F. Evolution of networks: from biological nets to the internet and www? p. 264, jan. 2003.

DUARTE, N. C.; HERRGARD, M. J.; PALSSON, B. O. Reconstruction and validation of saccharomyces cerevisiae ind750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, v. 14, n. 7, p. 1298–1309, 2004.

EDWARDS, J. S.; IBARRA, R. U.; PALSSON, B. O. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature Biotechnology*, v. 19, n. 2, p. 125–130, fev. 2001.

ERDÖS, P.; RÉNYI, A. *On Random Graphs*. 1st. ed. [S.l.], out. 1959. v. 1, 290–297 p.

ERDÖS, P.; RÉNYI, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, v. 5, p. 17, 1960.

ERDÖS, P.; SPENCER, J. Evolution of the n-cube. *Comp. and Maths with Appls.*, v. 5, p. 33, 1979.

ESTRADA, E. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, v. 6, n. 1, p. 35–40, 2006.

FEATHERSTONE, D. E.; BROADIE, K. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *BioEssays*, v. 24, n. 3, p. 267–274, mar. 2002.

FERRELL, J. E. Q&A: Systems biology. *Journal of Biology*, v. 8, n. 1, p. 2, jan. 2009.

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.

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.

GE, H. et al. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.*, v. 29, p. 482–486, Dec 2001.

GERMEK, G. R. *Caracterização topológica dos motivos das Redes Biológicas Integradas da Escherichia coli e da Saccharomyces cerevisiae*. 2009. Trabalho de Conclusão de Curso de Física Médica.

GIAEVER, G. et al. Functional profiling of the saccharomyces cerevisiae genome. *Nature*, v. 418, n. 6896, p. 387–91, 2002.

GIGLIOLI, M. *Determinação de comunidades em redes biológicas integradas*. 2009. Trabalho de Conclusão de Curso de Física Médica.

GONZÁLEZ-DíAZ, H. et al. Proteomics, networks and connectivity indices. *PROTEOMICS*, WILEY-VCH Verlag, v. 8, n. 4, p. 750–778, 2008. ISSN 1615-9861. Disponível em: <http://dx.doi.org/10.1002/pmic.200700638>.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. In: _____. *Formal Ontology in Conceptual Analysis and Knowledge Representation*. [S.l.]: Kluwer Academic, 1993.

GUELZIM, N. et al. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, ATelier de Genomique Cognitive, Centre National de la Recherche Scientifique ESA 8071, genopole(R), 523 Terrasses de l'Agora, 91000 Evry, France., v. 31, n. 1, p. 60–63, 2002 May.

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

GUSTAFSON, A. M. et al. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics*, v. 7, p. 265, 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.

HALL, M. et al. The weka data mining software: An update. *SIGKDD Explorations*, v. 11, 2009.

HUERTA, A. M. et al. RegulonDB: a database on transcriptional regulation in *Escherichia coli* K-12. *Nucleic Acids Research*, v. 26, n. 1, p. 55–59, 1998.

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

IDEKER, T. et al. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science*, v. 292, n. 5518, p. 929–934, 2001. Disponível em: <http://www.sciencemag.org/cgi/content/abstract/292/5518/929>.

IDEKER, T. et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, v. 292, n. 5518, p. 929–934, maio 2001.

ITAYA, M. An estimation of minimal genome size required for life. *FEBS letters*, 1995.

ITO, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, v. 98, n. 8, p. 4569–4574, abr. 2001.

JEONG, H.; MASON, S.; BARABÁSI, A. Lethality and centrality in protein networks. *Nature*, jan. 2001.

JEONG, H. et al. Lethality and centrality in protein networks. *Nature*, v. 411, n. 6833, p. 41–2, maio 2001. ISSN 0028-0836.

JEONG, H.; OLTVAI, Z. Prediction of protein essentiality based on genomic data. *ComPlexUs*, 2003.

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

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

JOYCE, A. R.; PALSSON, B. Ø. The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, v. 7, n. 3, p. 198–210, mar. 2006.

KALIR, S. et al. Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science*, v. 292, n. 5524, p. 2080–3, jun. 2001. ISSN 0036-8075.

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.

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.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, v. 28, n. 1, p. 27–30, jan. 2000.

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

KAUFFMAN, S. A. *The Origins of order: self organization and selection in evolution*. New York: Oxford University Press, 1993.

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

KOBAYASHI, K.; EHRLICH, S. Essential Bacillus subtilis genes. In: *Proceedings of the ...* [S.l.: s.n.], 2003.

KOHAVI, R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 1996.

KUMAR, A. et al. Subcellular localization of the yeast proteome. *Genes Dev*, v. 16, n. 6, p. 707–19, 2002.

KUMPULA, J. M. et al. Sequential algorithm for fast clique percolation. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 78, p. 026109, Aug 2008.

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

LEMKE, N. Phenotipic space approach to prey-predator coevolution. *Theory in Biosciences*, v. 117, p. 321–333, 1998.

LEMKE, N. Tsallis entropy production for diffusion on the diluted hypercube. *Physica A*, 2003.

LEMKE, N.; CAMPBELL, I. A. Random walks on closed spaces. *Physica A*, v. 230, p. 554, 1996.

LEMKE, N. et al. Essentiality and damage in metabolic networks. *Bioinformatics*, v. 20, n. 1, p. 115–119, 2004.

LEMKE, N. et al. Growth and form of two-dimensional rotating aggregates. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, v. 47, n. 5, p. 3218–3224, maio 1993.

LEMKE, N.; MOMBACH, J. A numerical investigation of adaptation in populations of random boolean networks. *Physica A*, 2001.

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

LUSCOMBE, N. M. et al. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, v. 431, n. 7006, p. 308–312, set. 2004.

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

MANDELBROT, B. *The Fractal Geometry of Nature*. [S.l.]: W. H. Freeman and Co., 1982.

MCKUSICK, V. A. Mendelian inheritance in man and its online version, omim. *Am J Hum Genet*, v. 80, n. 4, p. 588–604, 2007.

MERING, C. von et al. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, jan. 2002.

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.

MEZARD, M.; PARISI, G.; VIRASORO, M. A. *Spin Glass Theory and Beyond*. Singapore: World Scientific, 1987.

MILO, R. et al. Network motifs: simple building blocks of complex networks. *Science*, v. 298, n. 5594, p. 824–7, out. 2002. ISSN 1095-9203.

MOMBACH, J.; LEMKE, N.; BODMANN, B. A mean-field theory of cellular growth. *Europhysics Letters*, jan. 2002.

MOMBACH, J. C. M. et al. Using the FORESTS and KEGG databases to investigate the metabolic network of Eucalyptus. *Genetics And Molecular Biology*, v. 28, n. 3, p. 630–633, 2005.

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

MORIN, E. *Método I - A Natureza da Natureza*. Porto Alegre: Sulina, 2003.

NEWMAN, M. E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.*, v. 103, p. 8577–8582, Jun 2006.

NEWMAN, M. E.; STROGATZ, S. H.; WATTS, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 64, n. 2 Pt 2, p. 026118, ago. 2001. ISSN 1539-3755.

NEWMAN, M. E.; WATTS, D. J. Scaling and percolation in the small-world network model. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, v. 60, n. 6 Pt B, p. 7332–42, dez. 1999. ISSN 1063-651X.

OUZOUNIS, C. A.; KARP, P. D. Global properties of the metabolic map of Escherichia coli. *Genome Research*, v. 10, n. 4, p. 568–576, abr. 2000.

OVERBEEK, R. et al. The ERGO genome analysis and discovery system. *Nucleic Acids Research*, v. 31, n. 1, p. 164–171, jan. 2003.

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

PALLA, G. et al. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, v. 435, p. 814–818, Jun 2005.

PALUMBO, M. C. et al. Functional essentiality from topology features in metabolic networks: a case study in yeast. *FEBS Lett*, Physiology and Pharmacology Department, University of Rome, La Sapienza, Rome, Italy., v. 579, n. 21, p. 4642–4646, 2005 Aug 29.

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

PROFILING of E. coli Chromosome (PEC) database.

QUINLAN, J. *C4. 5: programs for machine learning*. 1993.

RAVASZ, E. et al. Hierarchical organization of modularity in metabolic networks. *Science*, v. 297, n. 5586, p. 1551–5, ago. 2002. ISSN 1095-9203.

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.

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.

ROEMER, T. et al. Large-scale essential gene identification in candida albicans and applications to antifungal drug discovery. *Mol Microbiol*, v. 50, n. 1, p. 167–81, 2003.

RONEN, M. et al. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A*, v. 99, n. 16, p. 10555–60, ago. 2002. ISSN 0027-8424.

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

SCHMITH, J. et al. Damage, connectivity and essentiality in protein-protein interaction networks. *Physica A*, v. 349, n. 3-4, p. 675 – 684, 2005. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437104013627>.

SCHWIKOWSKI, B.; UETZ, P.; FIELDS, S. A network of protein-protein interactions in yeast. *Nat Biotechnol*, The Institute for Systems Biology, 4225 Roosevelt Way NE, Suite 200, Seattle, WA 98105, USA., v. 18, n. 12, p. 1257–1261, 2000 Dec.

SERINGHAUS, M. et al. Predicting essential genes in fungal genomes. *Genome Res*, v. 16, n. 9, p. 1126–35, 2006.

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

SHEN-ORR, S. S. et al. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, v. 31, n. 1, p. 64–8, maio 2002. ISSN 1061-4036.

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

SILVA, J. P. M. da. *Construção e análise de modelos topológicos de redes biológicas usando a ontologia MONET*. Dissertação (Mestrado) — (Computação Aplicada) - Universidade do Vale do Rio dos Sinos, 2006.

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

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology*, v. 147, n. 1, p. 195 – 197, 1981. ISSN 0022-2836. Disponível em: <http://www.sciencedirect.com/science/article/pii/0022283681900875>.

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.

SU, C. et al. Bacteriome.org–an integrated protein interaction database for E. coli. *Nucleic Acids Research*, v. 36, n. Database issue, p. D632–6, jan. 2008.

UETZ, P.; IDEKER, T.; SCHWIKOWSKI, B. Visualization and integration of protein-protein interactions. In: _____. *Protein-Protein Interactions - A Molecular Cloning Manual*. [S.l.]: Cold Spring Harbor Laboratory Press, 2002. p. 623–646.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. *Nature*, v. 393, n. 6684, p. 440–2, jun. 1998. ISSN 0028-0836.

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

WITTEN, I. *Data Mining: Practical machine learning tools and techniques*. 2005.

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

WUCHTY, S. Evolution and topology in the yeast protein interaction network. *Genome Res*, v. 14, n. 7, p. 1310–4, 2004.

YEGER-LOTEM, E. et al. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *PNAS*, v. 101, n. 16, p. 5934–5939, 2004. Disponível em: <http://www.pnas.org/cgi/content/abstract/101/16/5934>.

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

YU, H. et al. Genomic analysis of essentiality within protein networks. *Trends Genet*, Department of Molecular Biophysics and Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA., v. 20, n. 6, p. 227–231, 2004 Jun.

ZASLAVER, A. et al. Just-in-time transcription program in metabolic pathways. *Nat Genet*, v. 36, n. 5, p. 486–91, maio 2004. ISSN 1061-4036.

ZOTENKO, E. et al. 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*, Max-Planck Institute for Informatics, Saarbruecken, Germany., v. 4, n. 8, p. e1000140, 2008 Aug.

APÊNDICE A – Fonte dos dados

Os dados utilizados nessa tesa são oriundos de várias fontes. Na Tabela 11 apresentamos a origem das várias bases de dados usados nessa tese.

Tabela 11: Lista dos bancos de dados biológicos utilizados para o processo de aquisição dos dados.

Banco de dados	Organismo	Dados	Referência
BIGG	*S. cerevisiae*	Metabolismo	(SCHELLENBERGER;2010)
YEASTRACT	*S. cerevisiae*	Regulação transcricional	(ABDULREHMAN et al., 2010)
KEGG	Vários	Metabolismo	(KANEHISA et al., 2008)
RegulonDB	*E. coli*	Regulação transcricional	(HUERTA et al., 1998)
Bacteriome.org	*E. coli*	Proteína-proteína	(SU et al., 2008)
PEC	*E. coli*	Essencialidade	(PROFILING...,)
BRITE	S. cerevisiae	Essencialidade	(KANEHISA et al., 2008)
BioGRID	Vários	Interações moleculares	(BREITKREUTZ et al., 2008)
DIP	Vários	Interação de Proteínas	(SALWINSKI et al., 2004)
HPRD	Humanos	Proteínas	(KESHAVA PRASAD et al., 2009)
MIPS	Mamíferos	Interação entre proteínas	(PAGEL et al., 2005)
TRED	Humanos	Regulação Transcripcional	(JIANG et al., 2007)
Reverter *et al*	Humanos	Expressão em Tecidos	(REVERTER; 2008)
OMIM	Humanos	Genes Mórbidos	(MCKUSICK, 2007)
Yildrim *et al*	Humanos	Genes Drogáveis	(YILDIRIM et al., 2007)

APÊNDICE B – Mineração de Dados

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

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

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

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

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

B.1 Estatísticas Utilizadas na Mineração de Dados

B.1.1 Valores de Desempenho do Classificador

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

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

B.1.2 Processo de Validação Cruzada

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

B.1.3 Índice Kappa – κ

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

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

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

B.1.4 Teste de Wilcoxon – W

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

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

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

Determinamos, então, o valor de W, dado por $W = \min(R^+, R^-)$. Se possuirmos mais de

15 diferenças, excluindo um termo caso o número de $d_i = 0$ seja ímpar, segue-se o teste de hipótese nula calculando o valore de z, dado por:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)N(2N+1)}} \tag{B.3}$$

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

Caso $N < 25$, costuma-se comparar o valor de W com os valores críticos para determinado α, W_c, propostos por Wilcoxon em seu artigo (WILCOXON, 1947). Se $W \leq W_c$, a hipótese nula pode ser rejeitada. Alguns desses valores estão apresentados na Tabela 12.

Tabela 12: Valores de W_c dependendo do número de diferenças utilizadas N, com com probabilidade $\alpha < 0,05$ e $\alpha < 0,01$ para hipótese nula.

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

Trabalhos do Autor

ELSEVIER

Physica A 301 (2001) 589–600

PHYSICA A

www.elsevier.com/locate/physa

A numerical investigation of adaptation in populations of random boolean networks

Ney Lemke, José C.M. Mombach*, Bardo E.J. Bodmann

Centro de Ciências Exatas e Tecnológicas, Universidade do Vale do Rio dos Sinos, 93022-000 São Leopoldo, RS, Brazil

Received 6 March 2001

Abstract

We investigate the adaptation of random boolean networks that are a model for regulatory gene networks. The model considers a general genetic algorithm and a fitness function that takes into account the full network dynamical behavior. We propose a mathematical function to quantify the complexity catastrophe. We also find that the latter occurs when the task complexity increases, i.e., using networks with longer periods. Finally, we discuss a scenario that describes the adaptation on the proposed fitness landscape. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Random boolean networks; Genetic regulatory networks; Cellular automata; Disordered systems

1. Introduction

The cell genome stores all information required for the construction and function of an organism. Its basic units, the genes, interact with each other to perform these tasks in an orchestrated way. Kauffman proposed a cellular automata model for the functioning genome where the dynamics are due to mutuous activations and inactivations of regulatory genes represented by a network of boolean variables. A Kauffman *NK* network is a set of N boolean variables each connected randomly to K other variables in the set. The state of each variable is determined from a random logical function of the K inputs.

* Corresponding author. Fax: +55-51-590-8162.
E-mail address: mombach@exatas.unisinos.br (J.C.M. Mombach).

PII: S0378-4371(01)00372-7

The underlying dynamics are set up as follows: The state of a variable S_i at instant $(t+1)$ is determined from a logical function (B_i) evaluating the states of the K input variables connected to it at instant t,

$$S_i(t+1) = B_i(S_{j_1}(t), S_{j_2}(t), S_{j_3}(t), \ldots, S_{j_K}(t)) \,. \tag{1}$$

Since the phase space of the networks is discrete and finite, the attractors are cycles with period length between 1 and 2^N (the total number of states of a network of size N). The networks are known to possess distinct dynamical behaviors dependent on K including a dynamical transition that separates an ordered phase at $K=2$ from a disordered phase for $K>2$. For $K=2$ the average period length and the number of cycles scale with $\sim\sqrt{N}$ while for $K>2$ they scale with $\sim e^N$ and $\sim N$, respectively [1–6].

An important application of random boolean networks (RBN) is the study of evolution where we can investigate the relation between genotype represented by an RBN and its phenotype defined by a fitness function. Despite of its relevance, the adaptation of RBN has received little attention from the scientific community. The only study on the subject was conducted by Kauffman [1]. The issue has both biological and technological appeal. In biology, boolean networks are a powerful model to understand the behavior and evolution of gene networks. In technology research, they provide a method to understand and develop, through trial and error, real intensive parallel computer programming [7–9].

In Kauffman's simulations [1] an initial population of N_{pop} networks with same connectivity K, size N, random boolean functions and connections are evolved. Fitness f of a network is defined as the highest fraction of the N elements ($f \in [0,1]$) of any state cycle of the network that matches a steady arbitrary pattern. At each generation (time step) N_{pop} individuals are created by changing M bits defining the boolean functions. The only adaptation mechanism considered is mutation. If a fitter net is found, it replaces all other individuals and the process is repeated. If no fitter net is found, the search is repeated from the current best-fit network. The main conclusions of the simulations are:

1. Adaptation for large values of M is faster.
2. Population fitness after a number of generations decreases with N for a constant K and approximates 0.5 assimptotically. This is called *complexity catastrophe* since these parameters are related to network complexity.
3. The complexity catastrophe intensifies for large K since they imply a more rugged fitness landscape where trapping the evolution is more likely.
4. Mutations, as the only adaptive mechanism, seem not sufficient for a population to explore the fitness space and reach maximum fitness.
5. $K=2$ RBN also exhibit the complexity catastrophe, however it is less intense than for large values of K.

Kauffman considers these results as an indication that the random boolean network evolution is similar to another model proposed by him, the NK model [1]. In the

NK model, we consider N genes as a bit-string. Each gene interacts epistatically with K other genes. The fitness of an individual of the population is determined by the interaction among the genes. In order to maintain the model tractable, the fitness of a particular bit-string is given by the average of $K+1$ different random numbers, uniformly distributed in the interval $[0,1]$, where parameter K is a direct measure of the ruggedness of the fitness space. We list below Kauffman's most important conclusions about the *NK* model for further reference:

- The number of local fitness optima is huge.
- The lengths of adaptive walks to optima are short and increase as logarithm of N (this is valid only in the limit $K=N-1$).
- In the limit $K=N-1$, fitness grows logarithmically with time.
- A genotype can climb only to a small fraction of the local optima.
- Only a small selection of genotypes can climb to any given optimum.

Based on these results Kauffman suggests that $K=2$ RBN adapts on "well correlated, good landscapes" [1]. In other words the fitness landscape for $K=2$ is equivalent to an *NK* surface of low K. This property and the dynamical behavior of these networks inspired Kauffman to propose the hypothesis that "living systems exist in the solid regime near the edge of chaos …", since these networks can have reasonable complex behavior and are also highly evolvable.

Both models are relevant as they provide a general framework to understand the evolution process by constructing a bridge between the micro- and the macroscopic evolution scales. Given the importance of these conclusions, and the fact that they are based on small scale simulations we understand that further work on this topic is essential to answer some open questions:

1. Are these results general, or do they depend on the way the fitness function is defined or in the specific details of the adaptation process?
2. Does the phase transition at $K=2$ change qualitatively the adaptation process?
3. Why does the complexity catastrophe set in?

2. The model

We investigate the capacity of the networks to be "programmed" to reach an arbitrary target cycle. This arbitrary dynamical state represents the goal of the selective process. For example, it may represent the cellular cascade of differentiations during ontogeny, the mutations of tumor cells, the selection of clone cells of the immune system [1], etc. Differently from Kauffman's work, in our model the goal of adaptation is not a steady pattern of activity of the boolean networks but a state cycle, which seems more realistic.

We start by defining a target net and its initial state, S_0. The target net evolves from S_0 and eventually reaches a cycle in which it remains. The goal is to find boolean

networks with cycles similar to the target's. To this end we employ a genetic algorithm (GA) [10] that mutates the connections and boolean functions of the networks and also performs crossovers. This is achieved by representing the boolean networks as a bit string and defining a fitness function (see Eq. (2)) to control the differential reproduction rate in the population, favoring networks with cycles closest to the target.

The following criteria yield a guide to construct the fitness function: The highest possible fit ($f=1$) will be obtained only if the asymptotic dynamical behavior of a given net is exactly the same of the target net. If both networks have cycles with the same length (period), we calculate the Hamming distance (the number of different bits) between the states belonging to each cycle. If the two cycles have different periods, the fitness function depends on the difference between the two periods and the similarity of both cycles. The fitness decreases exponentially with the difference between the two periods simulating the effect of deleterious mutations.

To calculate the fitness of a given net η, we start both η and the target net on S_0 and let them evolve until they reach cycles. Let P_η and P_τ be the cycle length of η and of the target net, respectively. We choose the longest one and call it P_m. In the next step both networks are evolved from their first state after the transient (labeled t_η and t_τ). The fitness f is calculated according to

$$f(\eta)=\frac{\mathrm{e}^{-A(P_\eta-P_\tau)^2}}{P_m N}\sum_{i=1}^{P_m} d_H(\vec{S}_\eta(t_\eta+i),\vec{S}_\tau(t_\tau+i))\,, \tag{2}$$

where A is a constant, which without qualitative restrictions on the results is defined as 1 and d_H is the Hamming distance between any two states. The exponential factor was included to favor networks with cycle lengths closer to the target.

Once the fitness function is defined we use a GA to simulate the adaptation process. We choose a standard GA with the following characteristics (operators):

Crossover: Two individuals are chosen from the population and a position is tagged on the genome with probability P_{cross}. The new individual genome will be composed by the first individual genome until the tagged position and by the second one afterwards.

Mutation: After the new individual is formed, its genome bits are flipped with probability P_{mut}.

Overlapping: A given population fraction P_{rep} is maintained from one generation to the next.

Fitness rescale: The fitness function rescale is a standard procedure to prevent the fast decrease in population variability at the initial stages of evolution [10]. It is implemented by a linear transformation

$$f'=af+b\,,$$

where a and b are chosen such that $f'_{av}=f_{av}$ and $\max(f')=Cf'_{av}=Cf_{av}$. In our simulations we use the standard value $C=2$, a more detailed description is given in Ref. [10].

The simulation code was implemented in C++ and uses the Galib package developed at Massachussets Institute of Technology (see `http://lancet.mit.edu/ga/`).

At this point we stress the differences between this study and the one originally introduced by Kauffman [1]:

- In Kauffman's simulation the target net is chosen with no restriction imposed on its cycle length, this decreases simulation performance while considering very large cycle lengths.
- The fitness function used by Kauffman is fairly simple and calculated by the Hamming distance between the best matching states of the target and the given net.
- Kauffman considered only $1, 2, \ldots, n$ point mutations whereas our implementation mutates each site on the genome with a given flipping probability. Our GA also uses crossover, overlapping populations and linear rescaling operators.

These modifications were included to investigate whether different GAs could obtain better performance and the quantitative importance of K, N and the target cycle length on the GA performance. Our fitness function takes into account the entire cycle and thus genuinely discriminates different dynamical behaviors.

3. Results

We have performed simulations using the following parameters:

- population size: 30;
- mutation probability: 0.1;
- crossover (P_{cross}): 0 and 0.9;
- target cycle length: 1,2,3,4,8,16 and 32;
- replacement probability (P_{rep}): 1 and 0.9;
- $N = 8, 16, 32, 64, 128$;
- $K = 1, 2, 3, 4, 5, 6$.

The results are averages calculated over five different runs for a given set of parameters.

Our concern is to investigate the GA performance to find solutions for the proposed problem. Our analysis was based on parameter F, the best fitness found in a population. In order to characterize the performance we investigate F time dependence and its value attained after 50,000 generations (this value was chosen to maintain our computational time manageable—the longer simulations took 2 h on a 400 MHz Pentium III computer).

In Fig. 1(a) we present the F dependence on N for target period 4 for different K values. The graph shows clearly that as K and N grow the GA performance dwindles. Since fitness is a central quantity reflecting dynamical properties of the complex system, we can synthesize the set of graphs in Fig. 1(a) with different K described by a single parameter a:

$$F(N, K) = \frac{1 + \mathrm{e}^{-aNK}}{2} \,. \tag{3}$$

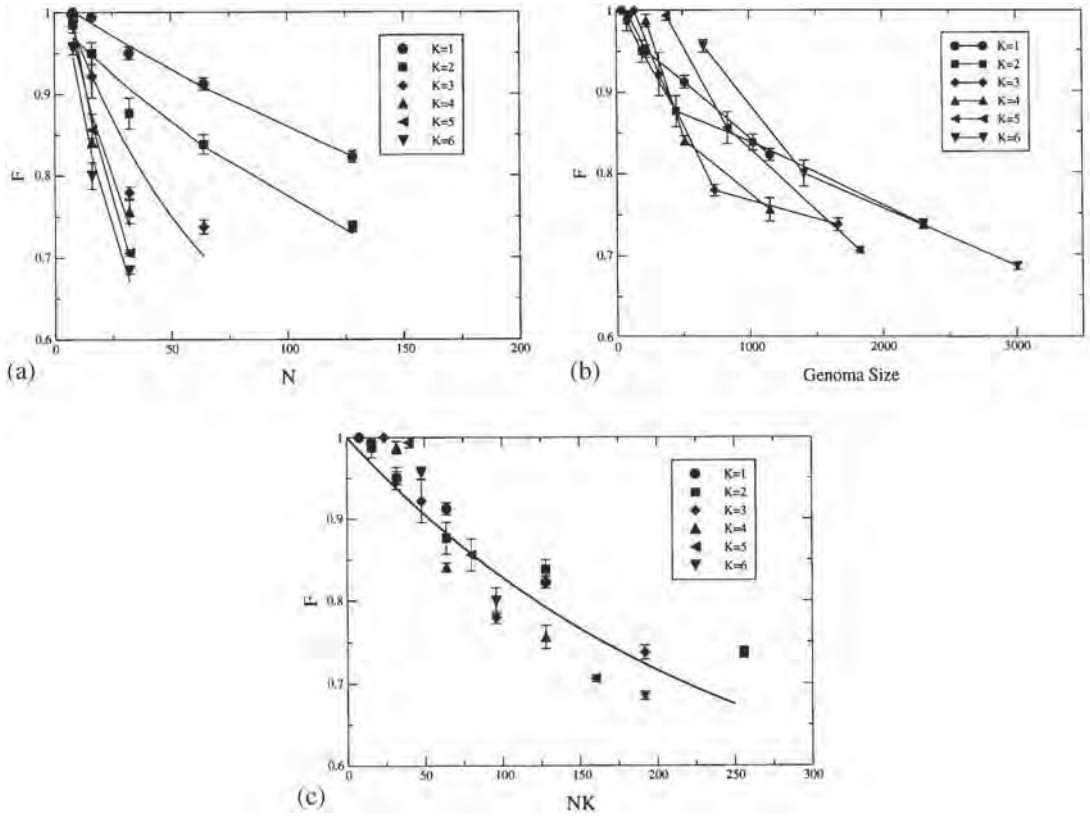


Fig. 1. The F dependence on N, genome size and NK for $K = 1, 2, 3, 4, 5$ and 6. Full lines in (a) and (c) correspond to the function $(1 + e^{-0.0042NK})/2$. The results illustrate the complexity catastrophe, F shows a fast decrease with N and K.

From parametric inference using the maximum likelihood [12] $a = 0.0042 \pm 0.0004$, where the uncertainty is based on a 95% confidence level (CL). The fairly good agreement of the multiple fit with the simulation data at the given CL may be interpreted as a confirmation of the proposed parametrization of the fitness function, as well as its significance.

Another issue of interest in applications using RBN is the interplay between F and the binary genome size. Fig. 1(b) shows the relation of F with the number of bits necessary to characterize the RBN genome:

$$G = N(K \log N + 2^K) . \tag{4}$$

The graph does not support Kauffman's conclusion that $K = 2$ is favored in comparison to the others since on comparing $K = 2$ RBN with other networks with different K values but roughly the same G, no conclusive evidence of improved performance can be found. This implies that the decaying performance of the GA is related both to the ruggedness and the volume of the phase space.

In Fig. 1(c) we present the same data from a different point of view, we plot F against NK. The proposed function captures the decay of GA efficiency and RBN complexity.

In Fig. 2 we present the dependence of F on the target net cycle period length for $K = 2$ and 3, and $N = 16$. We also compare the performance by changing P_{cross} and P_{rep}.

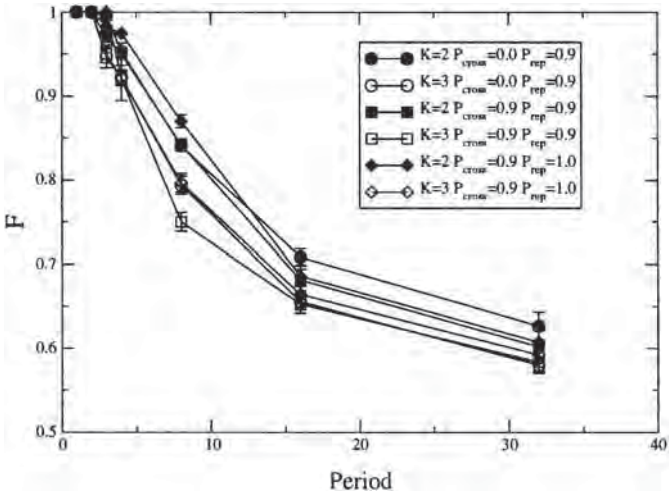


Fig. 2. The fitness dependence on the target period for crossover probability $P_{cross} = 0$ and 0.9, replacement probability $P_{rep} = 1$ and 0.9 and $K = 2$ and 3. The GA performance is slightly better for $K = 2$. The system size is 16.

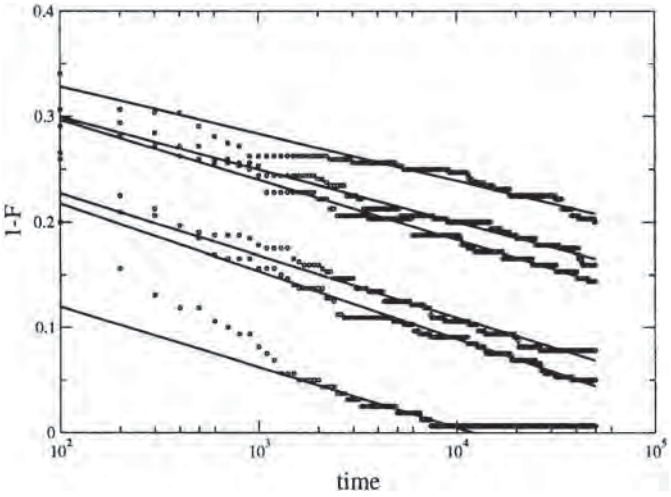


Fig. 3. Time evolution of the best individual for a GA with $N = 16$, $P = 4$, $P_{rep} = 0.9$ and from bottom to top $K = 1, 2, 3, 4, 5$ and 6. The straight lines are the best fit obtained using the function $A \ln t + B$.

The GA performance is slightly better for $K = 2$, but remains effectively unchanged as we vary other parameters. This is an indication that the fitness landscape presents no long range correlations [1].

From Figs. 1 and 2 we observe that performance decreases as we get more complex landscapes or longer periods, respectively. This is the manifestation of a complexity catastrophe caused by the complexity of the imposed task differently from the catastrophe caused by a structural property of the network. This result appears to be universal, being robust against changes P_{cross} and P_{rep}, as well as modifications in the crossover schema used.

As shown in Fig. 3, the $1 - F$ time dependence is well approximated by a function of the type $A \ln t + B$, suggesting a very slow relaxation towards equilibrium, where A

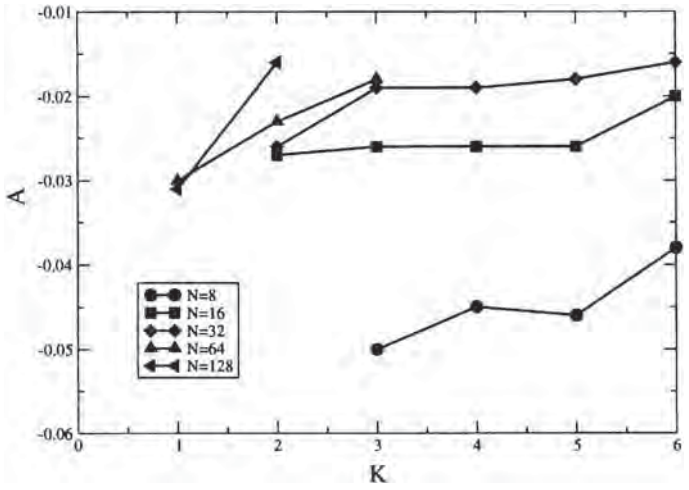


Fig. 4. Values of parameter A from the fitting function: $1 - F(t) = A \ln(t) + B$ for different values of N and K for $P = 4$. The data illustrate the complexity catastrophe, as N and K grow $|A|$ gets smaller, implying a divergent relaxation time.

measures the rapidity the GA takes to find an optimum solution. Thus the parameter dependence of A on N and K represents the network structure influence on the GA performance. As N and K grow A tends to 0 which implies a divergent relaxation time. For all parameters tested the proposed fit is accurate over two or three orders of magnitude in time, but in some cases, where the evolution is faster, the parametrization is no longer adequate. We believe that this behavior is due to a change in the landscape topology close to global maxima.

The logarithmic decay of F, is equivalent to say that the time needed to achieve a given F scales like $C \exp(-F/A)$. Since A increases towards zero when the phase space volume diverges, this suggests that the algorithm complexity for this problem is exponential.

The results presented here (Fig. 4) do not support the idea of a qualitative change of behavior for the critical $K = 2$ [3]. Actually, evolution is faster for $K = 2$ as compared to networks with the same N and larger K values but this has no influence on the functional dependence of $F(t)$. Comparing $K = 2$ networks to others with the same phase space volume (see Fig. 1(b)), we do not find any conclusive indication for a better GA performance. Hence, our results give no support to the idea that the adaptation on the edge of chaos is facilitated opposing Kaufmann's prediction [1].

In order to get a better insight in the complexity catastrophe phenomenon, a further analysis was performed. In the traditional view the population sticks to local maxima, there the variability decreases and the system has to cross a fitness barrier to reach another maximum. The complexity catastrophe in such scenario occurs because the number of local maxima grows exponentially with system size and so does the time spent at a given maximum.

To verify whether this picture is adequate to explain our system behavior, we quantified some features of the late stage evolving population. After performing a series of mutations on the best individual on a typical population we verified that it was not

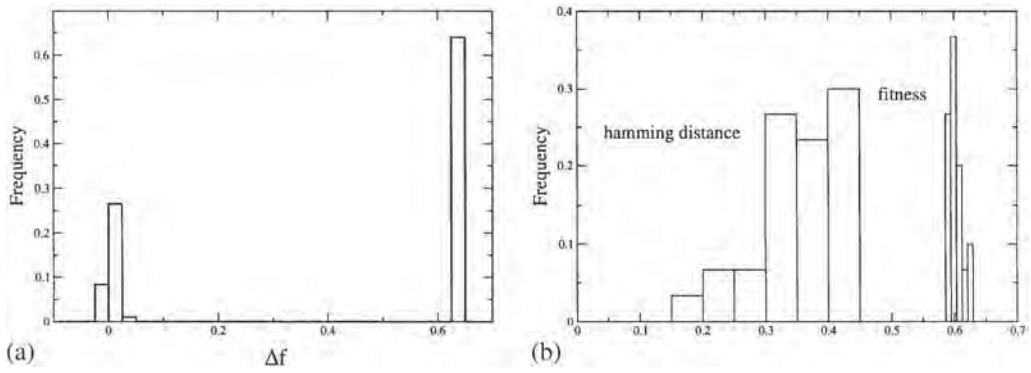


Fig. 5. Characterization of the late stage population. (a) The best individual is selected and all possible mutations are executed. The figure shows histogram for the fitness variations Δf between the best individual and its mutants ($N = 16$, $K = 2$, and $P = 4$). (b) We consider the distribution of fitness values for a given population and the hamming distance between the best individual and other individuals of the population. The figure shows that while the fitness distribution is concentrated around the maximum value, the hamming distance presents a broad distribution.

located on a local maxima. As shown in Fig. 5(a), approximately 40% of point mutations are adaptive or neutral. Fig. 5(b) shows for a given population the fitness and the hamming distance histogram. The graph suggests that while fitness values are strongly concentrated around the mean, the hamming distances have a spread histogram, indicating that the population is not concentrated around a given maximum. These results contradict the hypothesis of "population climbing" to fitness maxima, since we should obtain a large frequency of small (~ 0) hamming distances, corresponding to mutants of the best fit network.

These data support a new picture for the fitness landscape, although the landscape has a large number of local maxima there always exist some directions where the change in fitness is small or even zero. This is a direct consequence of the fitness function definition. Following the classical argument of Eigen and Schuster [13] we conclude that the *error catastrophe* sets in. Instead of being strongly located at the maxima, the population diffuses through the phase space. The image we have in mind is that the population diffuses on plateaus with small variation in fitness.

This phenomenon may be understood from the comparison of the two random surfaces in Fig. 6 (generated by MATHEMATICA 4.1 software as described by Maeder [14]) that are meant to be an instructive sketch of the high-dimensional phase space where the evolution occurs. As the population reaches higher plateaus with small variations in fitness values, diffusion slows down. This happens since the plateaus are irregular structures, possibly fractals. For higher fitness values they get more labirinthyne increasing substantially the relaxation times. The situation is equivalent to diffusion occurring in percolating clusters over the diluted hypercube as described in Refs. [15,16].

This model considers an hypercube in high dimension N with a fraction p of its sites occupied at random. As the critical dilution for percolation $p_c \sim 1/(N-1)$ is approached, the characteristic time for the relaxation diverges in the thermodynamic limit.

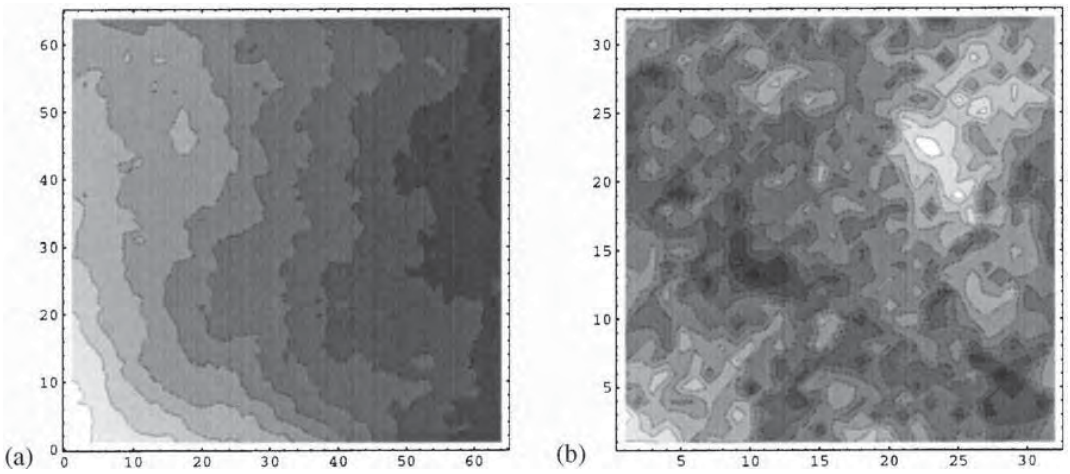


Fig. 6. Contour plot for two random surfaces with different roughnesses. Higher fitness values are represented by darker colors. These pictures are understood as a rough representation of the phase space.

The higher plateaus have shrinking volumes, which is equivalent to taking smaller p values on the hypercube. As the population reaches higher plateaus, one expects that phase space exploration slows down, while the typical hamming distance between two individuals remains considerably large. This explains the small variation in fitness values and the high hamming distances values depicted in Fig. 5. To the best of our knowledge there is no alternative explanation for this feature within the traditional picture of the population trapped in local maxima.

4. Discussion

In the present work we analyzed the adaptation of random boolean networks with genetic algorithms (GAs). The principal findings of our investigation are:

1. Convergence is slow (logarithmic) with time, suggesting that the system could reach arbitrarily high best-fitness values ($F \sim 1$) at the cost of an exponential increase of time. The time evolution of the best individual shows a decreasing improvement rate A, with increasing K, which tends towards zero when the phase space volume diverges, suggesting that the algorithm complexity for this problem is exponential.
2. The maximum fitness found in a population after a fixed number of generations is an exponentially decreasing function of N, K and P. Since the functional F dependence may be parametrized for all values of K, there is no evidence from our numerical data which might classify the behavior for $K = 2$ networks as qualitatively different from the ones with other K values. Especially, the complexity catastrophe appears not to be exclusively related to the structure of the network, but also to the task complexity imposed. Further, there is no significant GA parameter influence on the population evolution.

3. From the series of mutations performed on the best individual on a typical population our data show that, while fitness values are strongly concentrated around a mean, the Hamming distance histogram is characterized by a broad distribution. Hence there is no population concentration around one preferred maximum. Since approximately 40% of point mutations are adaptive or neutral this indicates that even the best individual is not located on a local maxima. Further, there is no specific correlation among the individuals having a similar fitness at a given evolution state.

The last cited result apparently contradicts the hypothesis of "population climbing" to fitness maxima, which is manifest in the missing of predominance in the zero frequency range of the Hamming distances. In order to understand more details of the model we make contact to an analog situation in a model studied by Mitchell and coworkers [11], who consider a mutation only GA evolving on a simple fitness landscape called the Royal Road Genetic Algorithm. In their study they find that the major contribution to the collapse of adaptation is the fast increase of phase space with GA parameters for a finite population. Similar to our findings in their model mutations are mostly neutral, and the population diffuses on plateaus until a higher one is found. The time involved in this process increases rapidly with the increasing phase space volume, decreasing the efficiency of the GA. In this case the complexity catastrophe is the result of an increasing phase space volume and not due to the ruggedness of the fitness landscape. They also find that crossover operators are inefficient in such a scenario as we verify in our model. The adaptation scenario proposed here has common features with other models [9,17], which indicates that the complexity catastrophe is a general phenomenon and may not be exclusively related to the ruggedness of the fitness profile but can also happen even in well behaved landscapes.

The combination of a small best-fitness distribution but a broad distribution of Hamming distances indicates the presence of fitness plateaus with the crucial difference that while in the model of Ref. [9] the plateaus correspond to a constant phase space volume, whereas in our model the higher the plateau, the smaller is the occupied phase space volume (this is a direct consequence of the fitness function definition or equivalently the selection criterion). Paradoxically the relaxation is slower as evolution proceeds. All indications point towards the existence of plateaus having a complex topology rather than the existence of a multitude of pronounced local maxima. Higher fitness values imply more labirinthyne plateaus and thus relaxation times increases. The scenario is then analogous to the diffusion occurring on percolating clusters over the hypercube as described in Refs. [15,16]. There is clearly the need for a more detailed comparison of the present model to the percolating cluster scenario on the hypercube. The present work may be considered as the first step in this direction. However, to establish a more rigorous connection between the present model and percolation clusters on a hypercube a more exhaustive characterization of those plateaus is in order, which is the topic of future work.

Acknowledgements

Work partially supported by FAPERGS and CNPq (Project 469445/2000-9).

References

[1] S.A. Kauffman, The Origins of Order: Self-Organization and Selection in Evolution, Oxford University Press, New York, 1993.
[2] A. Bhattacharjya, S. Liang, Physica D 95 (1996) 29.
[3] A. Bhattacharjya, S. Liang, Phys. Rev. Lett. 77 (1996) 1644.
[4] B. Derrida, Y. Pomeau, Europhys. Lett. 1 (1986) 45.
[5] B. Derrida, G. Weisbuch, J. Phys. 47 (1987) 1297.
[6] R. Albert, A.L. Barabási, Phys. Rev. Lett. 84 (2000) 5660.
[7] J.P. Crutchfield, M. Mitchell, Proc. Natl. Acad. Sci. USA 92 (1995) 10,742.
[8] S. Wolfram, Cellular Automata and Complexity: Collected Papers, 1st Edition, Addison Wesley, Reading, MA, 1994.
[9] M. Mitchell, J.P. Crutchfield, P.T. Hraber, Physica D 75 (1994) 361.
[10] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Learning, Addison-Wesley, Reading, MA, USA, 1989.
[11] E. Van Nimwegen, J.P. Crutchfield, M. Mitchell, Theor. Comput. Sci. 229 (1999) 41.
[12] M. Tanner, Tools for Statistical Inference, Springer, New York, 1996.
[13] M. Eigen, J. McCaskill, J. Chem. Phys. 92 (1988) 6881.
[14] R.E. Maeder, The Mathematica Programmer, Academic Press Inc., San Diego, 1996.
[15] I.A. Campbell, J.M. Flesseles, J.R. Jullien, R. Botet, J. Phys. C 20 (1987) L47.
[16] N. Lemke, I.A. Campbell, Physica A 230 (1996) 554.
[17] P. Schuster, W. Fontana, Physica D 133 (1999) 427.

BIOINFORMATICS

Vol. 20 no. 1 2004, pages 115–119
DOI: 10.1093/bioinformatics/btg386

Essentiality and damage in metabolic networks

Ney Lemke, Fabiana Herédia, Cláudia K. Barcellos, Adriana N. dos Reis and José C. M. Mombach

Laboratório de Bioinformática e Biologia Computacional, Centro de Ciências Exatas e Tecnológicas, Universidade do Vale do Rio dos Sinos, 93022-000 São Leopoldo, RS, Brazil

Received on January 9, 2003; revised on March 14, 2003; accepted on June 3, 2003

ABSTRACT

Understanding the architecture of physiological functions from annotated genome sequences is a major task for postgenomic biology. From the annotated genome sequence of the microbe *Escherichia coli*, we propose a general quantitative definition of enzyme importance in a metabolic network. Using a graph analysis of its metabolism, we relate the extent of the topological damage generated in the metabolic network by the deletion of an enzyme to the experimentally determined viability of the organism in the absence of that enzyme. We show that the network is robust and that the extent of the damage relates to enzyme importance. We predict that a large fraction (91%) of enzymes causes little damage when removed, while a small group (9%) can cause serious damage. Experimental results confirm that this group contains the majority of essential enzymes. The results may reveal a universal property of metabolic networks.

Contact: lemke@exatas.unisonos.br

INTRODUCTION

Understanding the architecture of the cell's biochemical network is a fundamental problem in modern science (Schuster *et al*., 1999; Jeong *et al*., 2000; Ravasz *et al*., 2002; Wagner and Fell, 2001). The genome sequences of several organisms are available (Kanehisa and Goto, 2000) and the prediction of function from the metabolic networks reconstructed from the gene products is an essential step in the postgenomic era (Karp *et al*., 1999). Here, we present a novel graph analysis of metabolism that determines the most important enzymes for the survival of an organism (Jeong *et al*., 2001) and apply our bioinformatics approach to the metabolic network of *Escherichia coli*. The method predicts quantitatively essentiality of enzymes from the topological damage their removal causes to a simulated metabolic network of the organism.

Cellular metabolism has a characteristic complex network of reactants connected by chemical reactions catalyzed by specialized proteins called enzymes. The reactions cluster in modules called metabolic maps with specific catabolic or anabolic functions. The complete set of metabolic maps forms the metabolic network. Developing realistic models for metabolic networks is still beyond our capabilities, since the experimental determination of the set of kinetic parameters that characterize each of the hundreds of reactions in a cell is a challenging problem. On the other hand, an exponentially growing number of organisms have sequenced genomes with many determined encoded proteins (Devos and Valencia, 2001). Assuming that the annotated enzymes are expressed, we can construct the metabolic network of the organism (Karp *et al*., 1999). A possible approach is to analyze the static structure of the components of the network to infer causal and physiological relationships.

Barabási and co-workers (Jeong *et al*., 2000) have introduced a graph representation of the metabolic network, where the nodes and the links connecting the nodes denote metabolites and chemical reactions, respectively. Their analysis of the networks of 43 organisms have shown that they have an inhomogeneous connectivity. A few nodes are highly connected while the majority have few connections, obeying a power law distribution of the type: $P(k) \propto k^{-\gamma}$, where k is the node connectivity and γ is an exponent with value close to 2. γ is approximately conserved among the organisms studied, suggesting than it is a universal feature. In these networks, highly connected nodes, i.e. those that participate in many reactions (referred to as hubs) are central to the integrity of the network, connecting the majority of the nodes. In order of decreasing connectivity, some of the main hubs are H_2O, ATP, ADP, phosphate, pyrophosphate, NAD, etc. (Jeong *et al*., 2000). More recently Ravasz *et al*. (2002), found that metabolic networks are a special type of scale-free network. The essential difference is that they are hierarchically organized in strongly interacting modules corresponding closely to the well-known metabolic maps.

Previous works focused on the relevance of specific metabolites. However, from a practical and biological point of view, is more important to investigate the influence of enzymes on the network. In contrast to metabolites that are not encoded in the genome, enzymes are subject to evolution and can be genetically engineered to change metabolic output. They can also be targets for drugs (Karp *et al*., 1999) so identifying important enzymes is a critical issue. Our main aim is to

*To whom correspondence should be addressed.

show that we can determine quantitatively the importance of enzymes by analyzing a graphical representation of the network.

METHODOLOGY

Our quantitative criterion for enzyme's importance is the deleterious effect of its removal from the network. Since the complete set of kinetic parameters is not known and we have rather incomplete information on the regulatory network, we cannot predict all the consequences of the deletion of a specific enzyme. However, we can determine the number of metabolites whose production the absence of the enzyme prevents, which we define as the damage d to the network. We can interpret d as a quantitative measure of the lack of alternative metabolic pathways that employ the enzyme. We validate our results by correlating d with the experimentally determined viability of the organism when the enzyme is removed from its proteome (Jeong *et al*., 2001). Experiments using systematic mutagenesis determine whether or not an organism is viable in the absence of the protein produced by a given gene. If the organism is not viable, the gene is classified as essential, if not, non-essential. For *E.coli* this information is publicly available, see *Profiling of the Escherichia coli chromosome* (PEC database, http://www.shigen.nig.ac.jp/ecoli/pec/) maintained by the Genetic Resource Committee of Japan. The PEC database compiles experimental information from the literature prior to 1998 that characterizes *E.coli* genome strains, including gene classification based on essentiality for cell growth. The genes are classified into three groups: essential, non-essential and unknown.

To calculate d, we use a special graphical representation of the metabolism of *E.coli*. The graph is directed and has two types of nodes (bipartite digraph, Chartrand, 1977). One type represents chemical reactions and the other metabolites (Fig. 1). A link between a reaction and a metabolite is directed towards the metabolite, if the metabolite is a product, and in the opposite direction, if the metabolite is a reactant. We treat reversible reactions as two separate reactions.

We constructed a graph from the list of reactions catalyzed by the enzymes involved in the small molecule metabolism of *E.coli* proposed by Palsson (Edwards *et al*., 2001; http://gcrg.ucsd.edu/downloads/index.html). We checked the reaction set in the KEGG (Kanehisa and Goto, 2000; http://www.genome.ad.jp/kegg/kegg2.html), ERGO (Overbeek *et al*., 2000; http://ergo.integratedgenomics.com/ERGO/) and EcoCyc (Ouzounis and Karp, 2000; http://BioCyc.org/ecocyc/) databases and formatted it for our use.

We propose the following algorithm (Fig. 1):

- First, we choose an enzyme and determine all reactions it catalyzes.
- If the reaction is irreversible we delete all the metabolites the enzyme produces exclusively.

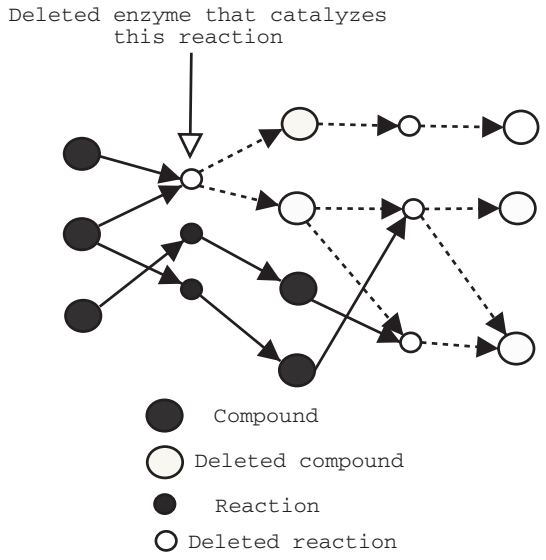


Fig. 1. Schematic representation of damage in a metabolic network. Small and large nodes represent chemical reactions and metabolites, respectively. Open symbols represent the absence of a chemical reaction or metabolite. The figure illustrates the effect of a deleted chemical reaction at left and the subsequent metabolites and reactions affected.

- If the reaction is reversible we delete all metabolites produced exclusively by the forward and reverse reactions catalyzed by the enzyme.
- We determine the set of remaining reactions that occur with the available metabolites.
- We iterate until we reach a fixed point.

The total number of deleted metabolites is the damage d.

RESULTS

Figure 2 presents a histogram of the number of enzymes with a given d. The distribution fits a power law of the form $P(d) = \alpha d^{-\lambda}$, with $\alpha = 0.54$ and $\lambda = 2.04$. This result supports the idea of metabolic network robustness (Jeong *et al*., 2001) since it shows that the vast majority of the enzymes, when deleted, cause little damage to the network. Table 1 lists the 30 enzymes for which the simulated network predicts more serious associated damage. From this set of 30, 13 are essential according to data in the PEC database, suggesting that our definition of damage is useful.

To compare our results with the database of essential enzymes of *E.coli*, we sorted the enzymes according to their d values and determined the fraction which were essential in each group. Figure 3 shows the results. Using an F-test, we determined that the correlation between the fraction of essentials and d is statistically significant with a P-value of 0.0228.

Table 1. List of 30 enzymes predicted to be associated with the highest damage

No.	Enzyme name	d	C	Product
1	Ribose-phosphate pyrophosphokinase	22	E	Phosphoribosyl pyrophosphate
2	3-Dehydroquinate dehydratase	21	N	Dehydroshikimate
3	Phosphoglucosamine mutase	20	E	Glucosamine 1-phosphate
4	Shikimate 5-dehydrogenase	20	N	Shikimate
5	UDP-*N*-acetylglucosamine pyrophosphorylase	19	E	UDP *N*-acetyl glucosamine
6	3-Phosphoshikimate 1-carboxyvinyltransferase	18	N	3-Phosphate-shikimate
7	Acetyl-CoA carboxylase carboxyl transferase	18	E	Malonyl-CoA
8	Malonyl CoA-ACP transacylase	17	E	Malonyl-ACP
9	3-Oxoacyl-ACP synthase	17	E	Acetyl-ACP
10	Chorismate synthase	17	N	Chorismate
11	*S*-adenosylmethionine synthase	16	N	*S*-Adenosyl-L-methionine
12	Fatty acid biosynthesis enzymes	16	E	Fatty acid-ACP
13	Geranyltranstransferase	14	E	Geranyl diphosphate
14	Glutamyl-tRNA synthetase	14	E	L-Glutamyl-tRNA
15	Glutamyl-tRNA reductase	13	N	L-Glutamate-semialdehyde
16	Glutamate-1-semialdehyde 2,1-aminomutase	12	N	5-Aminolevulinate
17	Delta-aminolevulinic acid dehydratase	11	N	Porphobilinogen
18	UDP-*N*-acetylglucosamine acyltransferase	10	E	UDP-3-*O*-(3-hydroxytetradecanoyl) *N*-acetylglucosamine
19	Octaprenyl pyrophosphate synthetase	10	N	All-trans-Octaprenyl diphosphate
20	Porphobilinogen deaminase	10	U	Hydroxymethylbilane
21	Amidophosphoribosyltransferase	9	N	5-Phosphoribosylamine
22	UDP-3-*O*-[3-hydroxymyristoyl] *N*-acetylglucosamine deacetylase	9	E	UDP-3-*O*-(3-hydroxytetradecanoyl)-D-glucosamine
23	Uroporphyrinogen-III synthase	9	U	Uroporphyrinogen III
24	Aspartate-semialdehyde dehydrogenase	9	N	Aspartate beta-semialdehyde
25	Chorismate–pyruvate lyase	8	N	Hydroxybenzoic acid
26	Phosphoribosylamine–glycine ligase	8	N	5′-Phosphoribosylglycinamide
27	UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase	8	E	UDP-*N*-acetyl-3-(1-carboxyvinyl)-D-glucosamine
28	UDP-3-*O*-[3-hydroxymyristoyl] gluc. *N*-acyltransferase	8	E	UDP-2,3-bis(3-hydroxytetradecanoyl)glucosamine
29	Ketol-acid reductoisomerase	8	N	2,3-Dihydroxy-isovalerate
30	ATP phosphoribosyltransferase	8	N	Phosphoribosyl-ATP

First column: Enzyme name. Second column: Damage d. Third column: Experimentally determined essentiality: E (essential), N (non-essential), U (unknown). Fourth column: Catalyzed metabolites.

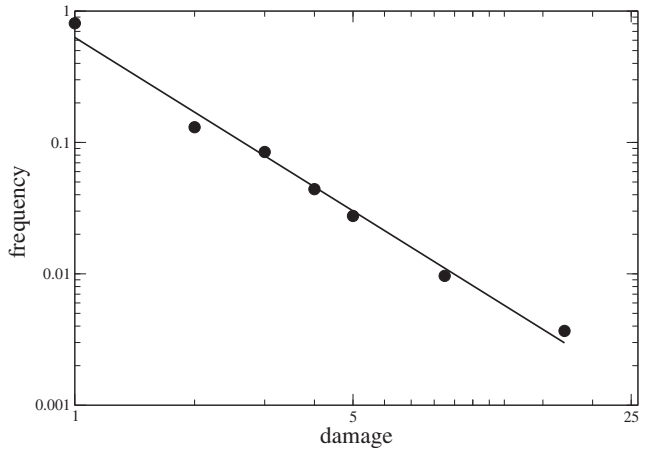


Fig. 2. Frequency of enzymes with damage d. The fitted line corresponds to a power law of the form $P(d) = \alpha d^{-\lambda}$, with $\alpha = 0.54$ and $\lambda = 2.04$. The correlation coefficient of the fit is 0.994.

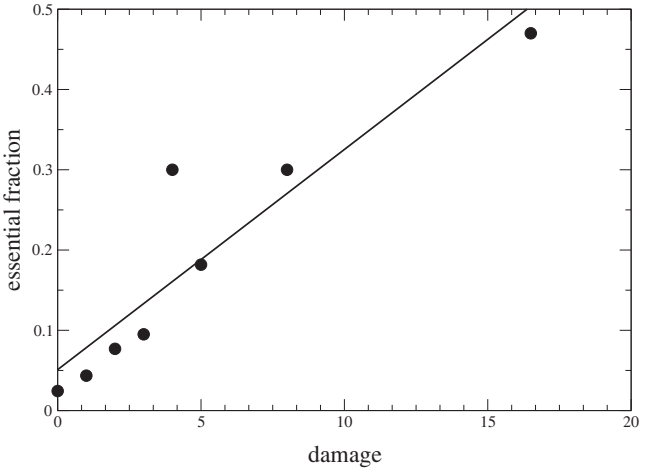


Fig. 3. Plot of the fraction of essentials enzymes in groups sorted according to the damage d. The P-value of the fit is 0.0228.

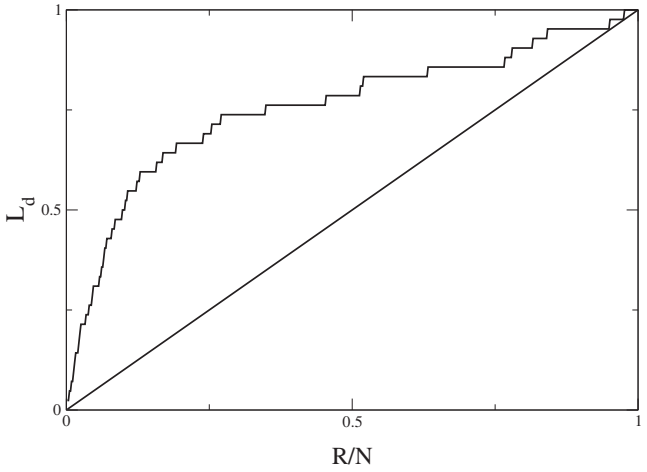


Fig. 4. The function L_d defined in the text versus the normalized rank. The concavity of this function implies that d correlates positively with the damage.

In another statistical analysis, we separated the enzymes into two groups. One set for $d < 5$ and another for $d \geq 5$. The first group has 91% of the total number of enzymes and the second 9%. We found that the second group contains 50% of all experimental essential enzymes. According to chance, the probability of finding this frequency of essential enzymes or higher in such a small subset of enzymes is 10^{-7}.

Another way to verify if two quantities correlate is to calculate the function L_d (Jeong *et al*., 2002):

- Rank the ORFs in decreasing order of d.
- Assign a $\delta_i = 0, 1$ variable to each ORF, i, whose lethality is known.
- Determine the $L_d(R)$ curve by summing δ_i starting from the bottom of the list and moving towards $R = N$ (the number of ORFs).
- Normalize $L_d(R)$ by dividing by the number of essential ORFs.
- Plot L_d against R/N.

If the data were uncorrelated, L_d, would be a straight line, if the data were positively or negatively correlated we would obtain a convex or concave curve. Figure 4 shows a positive correlation between d and lethality.

These analyses indicate that the damage is a quantitative measure of enzyme importance.

Our analysis of the thirty highest damage enzymes in Table 1 confirms that our model can identify potential targets for drugs since many of the listed enzymes are subjects of investigation for this purpose (Schroeder *et al*., 2002). A careful analysis of the 10 first enzymes shows that four enzymes (enzymes number 1, 5, 7 and 10, Table 1) are key compounds that link different metabolic pathways. Four enzymes (enzymes number 2, 4, 6 and 10) are involved in the production of chorismate, which is an important link to the biosynthesis of aromatic aminoacids, folate and ubiquinone. The enzyme with the highest damage, ribose-phosphate-pyrophosphokinase, generates phosphoribosyl pyrophosphate, which is the initial compound of four different pathways and is involved in intermediate metabolism. Finally, two enzymes (enzymes 3 and 5) are involved in the biosynthesis of the cell wall.

Several antibacterial agents target the cell wall since mammals do not synthesize walls and therefore are immune to the toxic effects of these agents. Many antibiotics inhibit normal synthesis of peptidoglycan by bacteria, causing them to burst as a result of osmotic lysis. Examples include the penicillins, the cephalosporins, the carbapenems and the glycopeptides (Chopra, 2002). About 40% of all essential *E.coli* ORFs in the PEC database, are involved in the metabolism of cell wall production. Six out of the 30 ORFs analyzed (Table 1) are involved with peptidoglycan biosynthesis or relate to aminosugar metabolism and lipopolysaccharide biosynthesis. These ORFs could be attractive targets for the development of antibacterial drugs (Schroeder *et al*., 2002; Campbell *et al*., 2001).

DISCUSSION

Enzymes associated with high damage are involved in the production of compounds of small connectivity that connect important parts of the metabolism. On the other hand, highly connected compounds tend to be redundant since they are produced by many reactions.

Surprisingly, some essential enzymes cause little damage and conversely, some non-essential enzymes cause serious damage. In the case of essential enzymes with low damage, we analyzed the bibliography provided by the PEC database, and verified that the majority may be involved in other important biological functions besides the metabolism of small molecules. Non-essential enzymes with high damage have influence restricted to a metabolism module that is not necessary in the bacterial environment; such micro-organism may not be viable in the wild.

Our method also applies to genetic engineering. We can investigate the reactions and enzymes responsible for the production of a given compound to help determine possible routes to its overproduction or elimination.

We believe that the damage concept may also apply in other biological contexts like networks of physical interactions among proteins. We are currently investigating this hypothesis.

ACKNOWLEDGEMENTS

We acknowledge the support of CNPq and FAPERGS. We thank the KEGG and PEC databases for providing public access to their data.

REFERENCES

Campbell,J.W. and Cronan,J.E.,Jr. (2001) Bacterial fatty acid biosynthesis: targets for antibacterial drug discovery. *Annu. Rev. Microbiol.*, **55**, 305–332.

Chartrand,G. (1977) *Introductory Graph Theory*. Dover Publications, New York.

Chopra,I., Hesse,L. and ONeill,A.J. (2002) Exploiting current understanding of antibiotic action for discovery of new drugs. *Symp. Ser. Soc. Appl. Microbiol.*, **31**, 4S–15S.

Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.

Edwards,J.S., Ibarra,R.U. and Palsson,B.O. (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.*, **19**, 125–130.

Jeong,H., Mason,S.P., Barabási,A.-L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Jeong,H., Oltvai,Z.N. and Barabási,A.-L. (2002) Prediction of protein essentiality based on genomic data. *ComPlexUs*, **1**, 19–28.

Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabási,A.-L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Karp,P.D., Krummenacker,M., Paley,S. and Wagg,J. (1999) Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol.*, **17**, 275–281.

Ouzounis,C.A. and Karp,P.D. (2000) Global properties of the metabolic map of *Escherichia coli*. *Genome Res.*, **10**, 568–576.

Overbeek,R. *et al.* (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.

Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. and Barabási,A.-L. (2002) Hierarquical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Schroeder,E.K., de Souza,O.N., Santos,D.S., Blanchard,J.S. and Basso,L.A. (2002) Drugs that inhibit mycolic acid biosynthesis in *Mycobacterium tuberculosis*. *Curr. Pharmaceut. Biotechnol.*, **3**, 197–225.

Schuster,S., Dandekar,T. and Fell,D.A. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.

Wagner,A. and Fell,D.A. (2001) The small world inside large metabolic networks. *Proc. R. Soc. Lond. B*, **268**, 1803–1810.

Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

An integrated model for cellular analysis

Eduardo Battistella, José G.C. de Souza, Cláudia K. Barcellos, Ney Lemke and José C.M. Mombach

Laboratório de Bioinformática e Biologia Computacional,
Universidade do Vale do Rio dos Sinos, Unisinos, Av. Unisinos, 950,
Caixa Postal 270, 93022-000 São Leopoldo, RS, Brasil
Corresponding author: J.G.C. de Souza
E-mail: jose@exatas.unisinos.br

Genet. Mol. Res. 4 (3): 506-513 (2005)
Received May 20, 2005
Accepted July 8, 2005
Published September 30, 2005

ABSTRACT. We present the MOlecular NETwork (MONET) ontology as a model to integrate data from different networks that govern cell function. To achieve this, different existing ontologies were analyzed and an integrated ontology was built in a way to make it possible to share and reuse knowledge, support interoperability between systems, and also allow the formulation of hypotheses through inferences. By studying the cell as an entity of a myriad of elements and networks of interactions, we aim to offer a means to understand the large-scale characteristics responsible for the behavior of the cell and to enable new biological insights.

Key words: Ontology, Cellular function, Knowledge data discovery, Data integration

INTRODUCTION

One of the most important challenges for biology in the postgenomic era is to understand the structure and behavior of the complex intercellular Web of molecular interactions that control cell behavior (Barabási and Oltvai, 2004). The basis to achieve this goal has already been built.

There are more than 548 biological data sources available on the Internet (Bateman, 2004). They hold data, such as genomes, mRNA, protein structures, protein-protein interactions, cellular signaling, metabolic pathways, and transcription-regulatory networks. This huge and complex set of data collected during recent years harbors information that requires an integrative approach (Uetz et al., 2002). Computer scientists and biologists will need to use innovative methodologies to deal with them.

The key to understand the structure and behavior of the cell is to integrate the available data in a way that it increases our understanding of the underlying biological processes that operate inside the cell (Ideker et al., 2001; Uetz et al., 2002; Barabási and Oltvai, 2004; Yeger-Lotem et al., 2004). Integrated biological models that assimilate this knowledge are essential to formulate new hypotheses, to predict cellular behaviors that can be tested experimentally (Ideker et al., 2001), and for a complete understanding of the cell. But the integration task is not simple.

Biological data are disseminated in many different databases. These databases have different management systems, formats and views of how to represent the data stored. Most of them are accessible by flat files or by web interfaces that allow some kind of query. The two main problems are the difficulty in parsing the data when dealing with heterogeneous flat file formats and the inconsistency due to the absence of a unified vocabulary, which means that the same information is represented in more than one way. Fortunately, ways to improve this scenario already exist.

Ontologies are an important approach to bring order to this scenario and to enable an integrated view of these data. An ontology is an explicit specification of a conceptualization (Gruber, 1993). While controlled vocabularies (e.g., Resource Description Framework (RDF), Extensible Markup Language (XML) Schema) only restrict the words used to describe a domain, ontologies extend this simple control vocabulary feature and allow the formal specification of the terms and the relations among them. They make sharing and reusing the knowledge possible, support the interoperability between systems, and also allow inferences from them. In bioinformatics, ontologies are crucial for maintaining the coherence of a large collection of complex concepts and their relationships (Backer et al., 1999).

In this context, we present the MOlecular NETwork (MONET) ontology. MONET ontology is a proposal to integrate data from the "network of networks" (Barabási and Oltvai, 2004) that exist inside the cell, helping us to understand the large-scale characteristics responsible for the behavior of the cell and enabling new biological insights. In short, it provides a way to cross the bridge between data and knowledge.

DOMAIN ANALYSIS

Bioinformatics is a growing field for ontologies (Battistella et al., 2004). As in other hotspot areas, new ontologies are frequently proposed, but the "infant mortality" is high. We pres-

ent some of the ontologies available for the molecular biology domain. We opted for the ones that have shown a continuous investment in research, resulting in new features/tools, and those whose proposals seem to have a promising future and could be adopted broadly.

One of the most ambitious projects of ontology applied to biology is the Gene Ontology Consortium (GO) (http://www.geneontology.org). GO aims to provide an ontology that covers several domains of molecular and cellular biology (Gene Ontology Consortium, 2004). It is structured into three sub-ontologies: biological processes (formed by one or more assemblies of molecular functions), molecular function (describes activities at a molecular level), and cellular component (enumerates the locations in a cell, considering subcellular structures). These sub-ontologies have been built to be used in the annotation of genes, gene products and sequences.

The Sequence Ontology Project (SO) (http://song.sourceforge.net) is a joint effort by genome annotation centers (WormBase, the Berkeley *Drosophila* Genome Project, FlyBase, the Mouse Genome Informatics group, and the Sanger Institute) that aim to offer an ontology suitable for sequence annotation and for data exchange of this annotation. It is under development, and its interim releases are made available as soon as they are considered to be usable. Examples of concepts available at SO are: intron, exon, gene, polypeptide, protein, DNA, RNA, mRNA, tRNA, and rRNA.

The Proteomics Standards Initiative (PSI) Molecular Interaction (MI) (http://psidev.sourceforge.net) ontology aims to represent interactions among proteins. PSI MI, an effort of the Human Proteome Organization (HUPO), was implemented through a specification of an ontology and an XML Schema. Both are being developed with a multi-level approach (Orchard et al., 2003; Hermjakob et al., 2004). The current level implements declarative representations of molecular interaction concepts divided into: interaction type, sequence feature type, feature detection, participant detection, and interaction detection. The interaction type vocabulary describes the type of connection found between molecules. The sequence feature type describes the relevant properties for the binding of proteins. The other three vocabularies describe the method by which the feature was detected.

The primary purpose of the Microarray Gene Expression Data (MAGE) (http://www.mged.org) ontology is to provide standard terms for the annotation of microarray experiments. Microarray data require complex structures, making some processes difficult, such as data-interchange and data documentation (Spellman et al., 2002). There have been various types of representations for microarray data, which make the reproduction of experiments a problematic task (Brazma et al., 2001). This ontology, which is currently under development, enables unambiguous descriptions of how the experiment was performed.

Other ontologies can be found at Open Biological Ontologies (OBO) (http://obo.sourceforge.net). OBO is an effort focused on the production of research that intends to facilitate the sharing of ontologies from different biological domains. All ontologies are open for use by the scientific community, and they are a useful starting point for new ones.

The ontologies presented here are not the only ones. There are other proposals, such as Transparent Access to Multiple Bioinformatics Information Sources (Goble et al., 2001), and proprietary ones, such as EcoCyc (Karp, 2000) ontology. New efforts are being launched, such as BioBabel (a new European project being coordinated by the European Bioinformatics Institute - EBI - http://www.ebi.ac.uk/biobabel). BioBabel aims to enhance the data interchange of biological databases by standardization of biochemical terminology.

All these ontologies show how the efforts to cover the vast area of molecular biology were developed until now. Based on these ontologies, and because of the need for an integrated approach, we introduce MONET.

MONET ONTOLOGY

There is a need for ontology proposals that allow an understanding of how the molecular networks inside a cell determine cell behavior (Ideker et al., 2001; Uetz et al., 2002; Barabási and Oltvai, 2004; Yeger-Lotem et al., 2004). Among other requirements, the proposal must be able to minimize data redundancies and inconsistencies. The data-interchange problem must be taken into consideration through the adoption of free and open standards. It also needs to be extensible, so new knowledge can be easily implemented by the aggregation of new concepts.

MONET integrates information from transcription-regulatory, metabolic pathway, and protein-protein interaction networks through a strategy that aims to establish a model able to minimize data redundancies and data inconsistencies. It is expandable, so new knowledge can be easily implemented. Even whole ontologies can be incorporated into MONET, which allows unlimited possibilities concerning the coverage of domains. Consequently, MONET allows the construction of topological models of cells of microorganisms, and the extension of these models becomes available as new knowledge.

The definition of an ontology is time consuming. An editor can result in a significant productivity profit. Among the available ontology editors we chose Protégé-2000 (http://protege.stanford.edu). The two main reasons for choosing Protégé were: a) the need, not only for a ontology editor, but for a Knowledge Base Management System, since we want to populate the database with examples from various microorganisms, and b) its open-source Java extensible architecture allows improvements in its functionality through the aggregation of new plugins. This latter characteristic allows the ontology to be exported in the different formats required by different research groups. A variety of import/export plugins can be used to automatically read/write the ontology in different representation data standards, such as Web Ontology Language (OWL), RDF, XML, and XML Schema.

The technical vocabulary used to describe MONET, concerning the ontology (not the biological knowledge), is based on Protégé. Its frame-based representation defines an ontology as a formal explicit description of concepts in a domain of discourse (concepts or classes), the properties of each concept describing various features, the attributes of the concept (slots or properties), and restrictions on slots (facets).

Figure 1 is a schematic representation of the main concepts implemented to achieve this integrated approach. Various types of concepts related to chemical molecules, such as DNA, RNA, mRNA, rRNA, tRNA, snRNA, and small metabolites, were omitted for simplification. We also omitted the slots of all concepts.

The transcription-regulatory network implements concepts, including operon (a set of genes transcribed under the control of an operator gene), transcription unit (part of DNA that will be transcribed into an RNA), terminator (DNA region where the transcription supposedly stops), ORF (a portion of a gene sequence that potentially encodes a protein), site (DNA sequence whose location and base sequence are known), promoter (a segment of DNA which provides a site where the enzymes involved in the transcription process can bind to a DNA

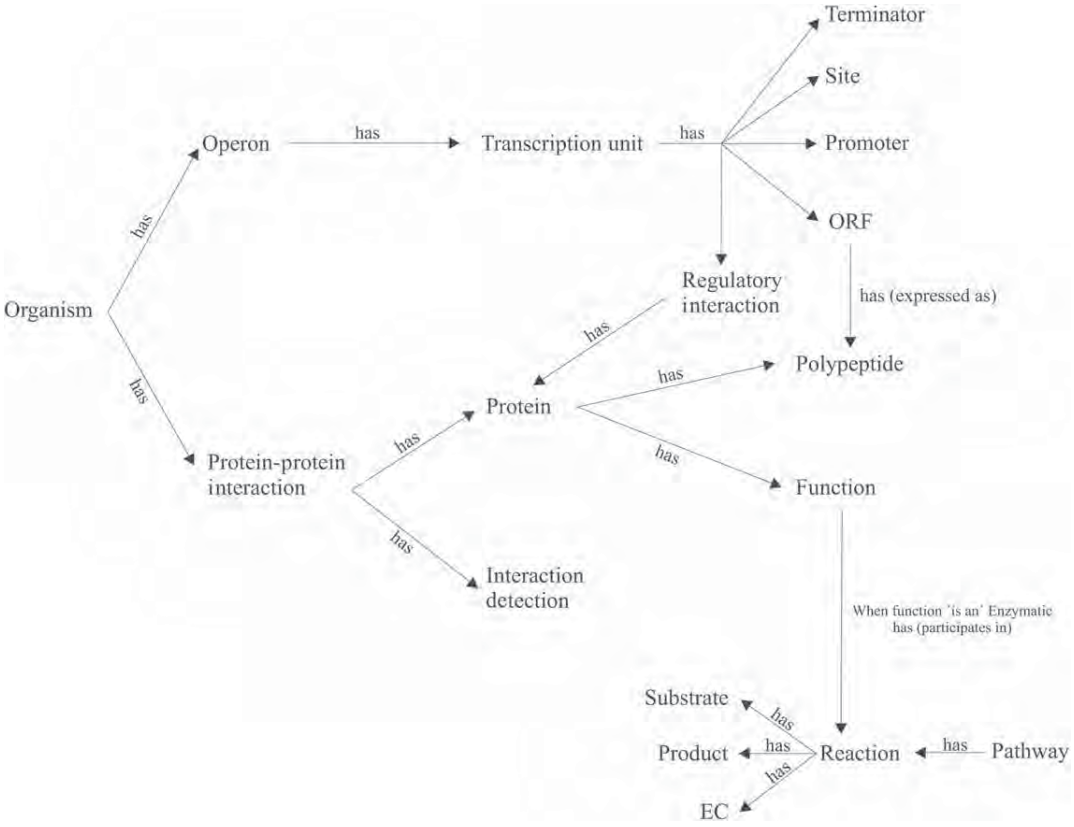


Figure 1. The main concepts of MONET ontology that integrate metabolic pathway networks, transcription-regulatory networks, and protein-protein interaction networks. ORF = open reading frame; EC = enzyme commission number.

molecule, and initiate transcription), and regulatory interaction (general information concerning the transcription-regulatory data being mapped).

The transcription-regulatory network is involved with interactions between DNA and proteins, and with the consequent production of proteins. The metabolic pathway network also involves proteins characterized by their enzymatic function. Proteins are the link between these networks.

The protein-protein interaction network has pairs of proteins whose interaction was detected experimentally or by an *in silico* process. This knowledge was also mapped into MONET. For each protein-protein interaction, we adopted from PSI MI ontology the concept of interaction detection (id MI:0001) and its subtree of concepts. The method to determine the interaction was divided into the sub-methods experimental and *in silico*, each with their corresponding possible notations.

The small molecule metabolism (metabolic pathway network) of MONET is a subset of the complete metabolism that excludes DNA replication and protein synthesis reaction. Beyond the concepts of reaction, substrate and EC (the enzyme commission number), other con-

cepts, such as inhibitor, activator, kinetic, and chemicals, are involved. Although the structures of metabolic pathway networks and protein interaction networks are similar, there are a number of significant differences. While metabolic pathways focus on the conversion of small molecules and on the enzymes responsible for these conversions, protein interaction maps concentrate mainly on physical contacts, without obvious chemical conversions (Uetz et al., 2002).

The spatial aspect was also taken into consideration. MONET implements a concept entitled compartment to indicate the protein's subcellular location. Consideration of the location of a protein and other chemicals is an important feature that allows more precise conclusions.

DISCUSSION

In our view, this model is a way to understand the internal organization and evolution of cells. It is not static, nor is it complete. But it is an important step in a direction that can lead us to a comprehensive modeling of the various networks that control the behavior of the cell.

The current version of this model implements metabolic pathway, transcription-regulatory, and protein-protein interaction networks. This model is being improved through the incorporation of a cell-signaling network.

MONET is neither better nor worse than GO, PSI MI, MAGE, SO, or other ontologies. It has a different point of view of how to model the knowledge. GO attacks the annotation problem; MONET is not in this stage yet. PSI MI deals with molecular interactions; MONET also deals with this problem, and it incorporates most of the concepts available in PSI MI. MAGE covers microarray experiments; MONET does not. SO offers sequence annotation and provides for data interchange of this annotation; MONET also does so by incorporating most SO concepts.

While these other ontologies are specific to a particular aspect of the molecular biology domain, MONET extends them and integrates them as a whole, giving a holistic view of the cell, allowing for "functional bioinformatics" (Karp, 2000). This bioinformatics makes the development of new algorithms, graphical visualization interface, and many other tools that aid in the investigation of the principles that govern cellular function, possible.

We intend to populate our knowledge base with information from some microorganisms. We already started this process with the incorporation of the KEGGs Ligand database (http://www.genome.ad.jp/kegg) as part of the metabolic pathway networks. This was not a simple task. To achieve this, we developed Python scripts to normalize the data available in the flat files, executed a series of consistency checks to correct the inconsistencies, and automated the generation of the instance flat file of Protégé (*pins* file). This process resulted in 21,430 small metabolites, 6,135 reactions, 4,327 enzymes, and 120 metabolic pathways.

We have also populated the knowledge base with protein and metabolic data from the microorganism *Ureaplasma urealyticum*. By doing so, we were able to export the metabolic data in XML format and load it in Mathematica 5.0 (http://www.wolfram.com) software, which allowed us to build up the metabolic network (Figure 2).

CONCLUSION

We present the MONET ontology as an integrated approach to build, test, and refine a model of the cellular pathway of organisms. It remains a challenge to integrate data from the

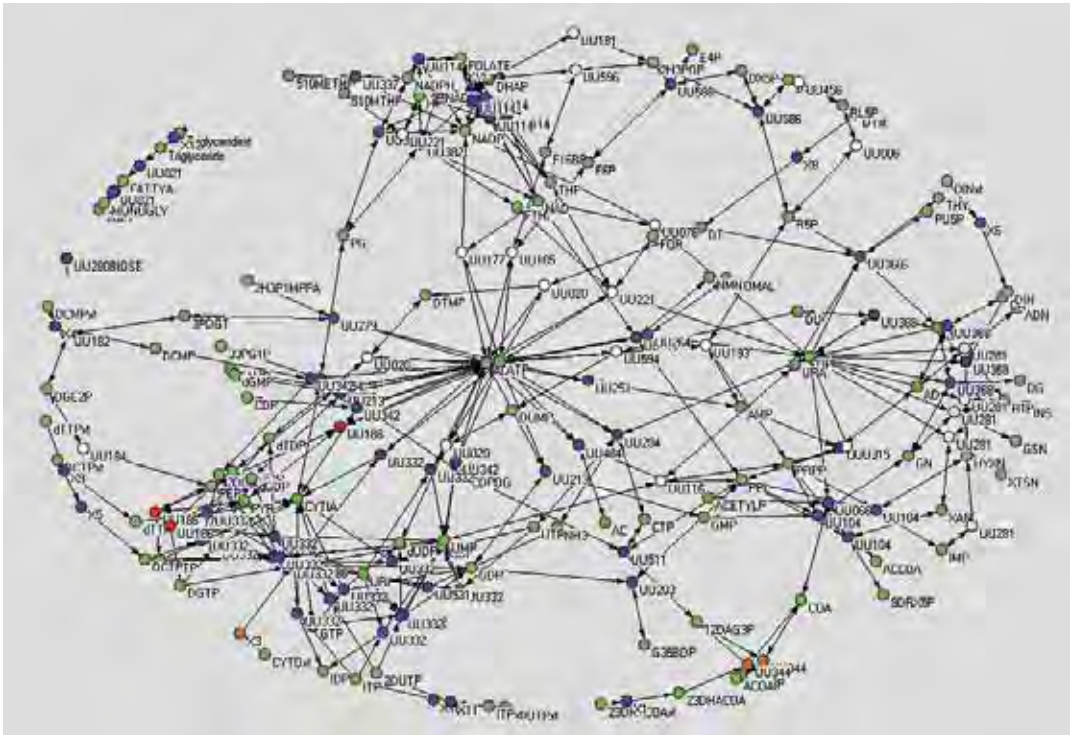


Figure 2. Bipartite graph of the metabolic network of *Ureaplasma urealyticum*. Dark gray and white nodes represent enzymes and light gray nodes represent metabolites (Lemke et al., 2004).

myriad interactions of the cellular constituents. One may contest our view of how to model these networks and to integrate them. This is one of the possible variations that concern this complex, constantly changing, and not yet completely understood, area of molecular biology.

The future will bring new graphical interfaces to visualize and to analyze these networks, and will also bring new integrative models on which simulations may be performed, fundamentally improving our view of cell biology.

The next steps in our work are to refine MONET, including concepts such as cellular signaling, and to use this ontology to build a knowledge base for the microorganisms *Escherichia coli*, *Helicobacter pylori* and *Mycoplasma pneumonia*. We expect to simplify and speed up the extraction of relevant biological knowledge with this topological integrated model of an organism.

Copies of the MONET ontology (in Protégé, OWL, RDF, and XML Schema formats) are available upon request from the authors.

ACKNOWLEDGMENTS

Research supported by FAPERGS and CNPq, process number 401999/2003-3. This work was developed in collaboration with HP Brazil R&D.

REFERENCES

Baker, P.G., Goble, C.A., Bechhofer, S., Paton, N.W., Stevens, R. and **Brass, A.** (1999). An ontology for bioinformatics applications. *Bioinformatics 15*: 510-520.

Barabási, A.-L. and **Oltvai, Z.N.** (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet. 5*: 101-113.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and **Eddy, S.R.** (2004). The Pfam protein families database. *Nucleic Acids Res. 32* (Database issue): D138-D141.

Battistella, E., Souza, J.G.C., Ferreira, R.A., Vieira, R., Lemke, N. and **Mombach, J.C.M.** (2004). Bioinformatics: A Growing Field for Ontologies. In: *Proceedings of the Workshop on Ontologies and their Applications* (WONTO'2004), São Luis, MA, Brazil, pp. 93-103.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and **Vingron, M.** (2001). Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nat. Genet. 29*: 365-371.

Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res. 32*: 258-261.

Goble, C.A., Stevens, R., Bechhofer, G., Ng, S., Paton, N.W., Baker, P.G., Peim, M. and **Brass, A.** (2001). Transparent access to multiple bioinformatics information sources. *IBM Systems Journal Special Issue on Deep Computing for the Life Sciences 40*: 532-552.

Gruber, T.R. (1993). Toward principles for the design of ontologies used for knowledge sharing. In: *Formal Ontology in Conceptual Analysis and Knowledge Representation* (Guarino, N. and Poli, R., eds.). Kluwer Academic, Deventer, The Netherlands.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, R., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y.X., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G.N., Sander, C., Bork, P., Zhu, W.M., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, L., Eisenberg, D., Steipe, B., Hogue, C. and **Apweiler, R.** (2004). The HUPO PSI's Molecular Interaction format - a community standard for the representation of protein interaction data. *Nat. Biotechnol. 22*: 177-183.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. and **Hood, L.** (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science 292*: 929-934.

Karp, P.D. (2000). An ontology for biological function on molecular interactions. *Bioinformatics 16*: 269-285.

Lemke, N., Heredia, F., Barcellos, C.K., Reis, A.N. and **Mombach, J.C.M.** (2004). Essentiality and damage in metabolic networks. *Bioinformatics 20*: 115-119.

Orchard, S., Kensey, P., Hermjakob, H. and **Apweiler, R.** (2003). Meeting review: The HUPO Proteomic Standard Initiative meeting: towards common standards for exchanging proteomic data. *Comp. Funct. Genomics 4*: 16-19.

Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W.L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B.J., Robinson, A., Bassett, D., Stoeckert Jr., C.J. and **Brazma, A.** (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol. 3*: 461-469.

Uetz, P., Ideker, T. and **Schwikowski, B.** (2002). *Visualization and Integration of Protein-Protein Interactions*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, pp. 623-646.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U. and **Margalit, H.** (2004). Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc. Natl. Acad. Sci. USA 101*: 5934-5939.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 349 (2005) 675–684
PHYSICA A
www.elsevier.com/locate/physa

Damage, connectivity and essentiality in protein–protein interaction networks

Jean Schmith, Ney Lemke*, José C.M. Mombach, Patrícia Benelli, Cláudia K. Barcellos, Guilherme B. Bedin

Laboratório de Bioinformática e Biologia Computacional, Programa Interdisciplinar de Computacõ Aplicada, Universidade do Vale do Rio dos Sinos, 93022-000 São Leopoldo, RS, Brazil

Received 8 June 2004
Available online 24 November 2004

Abstract

The proteome, a set of proteins expressed in an organism, is organized in an intricate web called the protein–protein interaction (PPI) network. In this article we propose a topological parameter called damage, that measures the consequences of the deletion of a protein from the network. We investigate different PPI data sets using this parameter and also traditional ones: the connectivity and the clusterization coefficient. We show that damage histogram obeys a power law in all data sets, and that proteins that cause a large damage are, with high probability, essential. For data sets that consider physical interactions the PPI network is a hierarchical scale-free network, while for data sets that consider functional interactions the PPI network is a scale-free graph.

PACS: 87.10.+e; 87.17.Aa

1. Introduction

The development of high throughput methods to determine complete genomes of organisms is a major step towards the understanding of biological systems. However,

*Corresponding author. Centro de Ciencias Exatas e da Terra-Unisinos, Av. Unisinos 950, Sao Leopoldo, RS 93022-000, Brazil.
E-mail address: lemke@exatas.unisinos.br (N. Lemke).

0378-4371/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2004.10.038

determining the proteins of an organism is only the start of a full description that requires the dynamics of how they interact and perform networked functions [1]. Aiming at these objectives other initiatives employing large-scale methods are investigating other major systems of the cell, challenging theoretical biologists to propose frameworks to extract useful biological information from them.

Among these systems, one of the most important is the proteome, which encompasses the expressed proteins of an organism and their interactions. Understanding this network that performs computations within the cell is a central problem in molecular biology. From the point of view of physics, it offers the challenge of understanding the collective behavior of interacting molecular machines designed to operate with remarkable precision under strong biological constraints. The recent abundance of genome sequence data brought an urgent need for systematic proteomics to decipher the encoded protein networks that dictate cellular function.

The image we obtain from these large-scale experiments and from more traditional ones, is that the proteome is a complex network formed by thousands of interactions, this network of protein interactions is here called PPI for protein–protein interaction. Given the intricacies of this web, a detailed and complete description that takes into account the dynamical aspects of these interactions is beyond our reach. A natural approach is to consider simplified models that consider only the topological aspects of the interaction network, but even this task is challenging since determining protein interactions remains a difficult and imprecise task. To illustrate this point we have $\sim 80,000$ reported interactions between yeast proteins, only $\sim$ 2400 are supported by more than one data set [1]. So, a natural question is whether these interactions have a truly biological and physical meaning.

Mering et al. [1] proposed a computational approach that captures functional relationships between proteins based on their conservation in different organisms, on the position their corresponding ORFs occupy in the genome, and also from proteins fused in a single polypeptide chain on other organisms [2]. The data sets obtained using this in silico technique cover a substantial portion of the proteome and the predicted interactions are compatible with experimental results [1].

Barabási and coworkers were the first to characterize topologically PPI networks. In this model, each protein is represented as a node and the interactions as arcs. Jeong et al. [3] have shown that the probability that a yeast *Sacharomyces cerevisiae* protein interacts with k other proteins follows a power law (this property is called scale freeness) with an exponential cut-off at $k_c = 20$. This topological feature is also shared by the PPI network of the bacterium *Helicobacter pylori* [5]. Yook et al. [6] have shown that in these networks the clusterization coefficient decays as a power law with the connectivity k a property called hierarchy. The work of Maslov and Sneppen [7] has shown the tendency of highly connected nodes (hubs) to associate with nodes of low connectivity, while hubs have reduced probability of being directly linked to each other. This property is called assortativity and its biological significance remains unknown [8].

Networks of this class are considered to be robust against random perturbations, since they tend to affect the more numerous weakly connected nodes, producing only

light perturbations on the network. In contrast, perturbations on hubs are prone to kill the organism since they affect a considerable part of the network. Thus connectivity is supposed to be a measure of gene importance or essentiality.

In this article we introduce damage, which is an extension of the topological measure of the importance of a node on a network. Damage takes into account the influence of a node not only on its first neighbors but also on the integrity of the whole network. A similar measure on metabolic networks was proposed by Lemke et al. [9] and showed correlation between damage and essentiality. In this paper, we compare connectivity and damage as topological measures of importance on networks of protein interactions (generated from different data sets) in order to determine which is the more efficient measure of gene importance.

In a cell the whole network of interactions among its molecular constituents include interactions among proteins, proteins and metabolites (enzymes and substrates or products) in metabolism, and among proteins and DNA in transcription, replication, and regulation. This integrated network is the circuitry that controls cell behavior. In this work, we investigate predictions of gene essentiality in *Sacharomyces cerevisiae* using the PPI network and a version of its integrated network generated by Ideker et al. [10].

This article is divided as follows, in Section 2 we describe our methodology and different data sets used in this work, in Section 3 we present our results, finally in Section 4 we draw our conclusions.

2. Methodology

We measure three standard topological parameters: connectivity, clusterization coefficient and damage [9]. The connectivity k_i is the number of nodes connected to node i. 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. It counts the number of triangles we can construct having the node i as a vertex. A hierarchical network has $C(k) \sim k^{\alpha}$ [8].

Damage is a quantitative criterion for the deleterious effect of the removal of a node from the network. In analogy to our definition of damage in metabolic networks [9], consider a PPI graph and its largest cluster G with n vertices. Now consider $G'(i)$ the largest cluster of the subgraph obtained from G after the deletion of the node i. We define damage d as $d = n - n'$, where n' is the number of vertexes of G'.

We assess the biological relevance of d and k determining the correlation between d and the experimentally verified viability of the organism when the protein is removed from its proteome [3]. Experiments using systematic mutagenesis determine whether or not an organism is viable in the absence of the protein produced by a given gene. If the organism is not viable the gene is classified as essential, if not, as non-essential. For *S. cerevisiae* this information is publicly available, see Ref. [4].

To determine if two quantities are correlated we calculate the function L_d proposed originally by Barabási and coworkers [11]. This function can be calculated following the steps below:

- Rank in decreasing order the ORFs based on d or k.
- Assign a $\delta_i = 0,1$ variable to each ORF i whose lethality is known.
- Determine $L_{d,k}(R)$ curve by summing δ_i starting from the smallest rank moving toward $R = N$ (the number of ORFs).
- Normalize $L_{d,r}(R)$ dividing by the number of essential ORFs.
- Finally plot L_d against R/N.

If the data were uncorrelated $L_{d,r}$ would be a straight line, if the data were positively or negatively correlated we would obtain a concave or convex curve, respectively. To compare the correlation we integrate the $L_{d,r}$ curves, obtaining a quantity that we call $S_{d,r}$.

2.1. Data sets

We use the five data sets below in our analysis:

Ito: Data produced from two-hybrid experiments. In these experiments pairs of proteins to be tested for interaction are expressed as fusion proteins (''hybrids'') in yeast: one protein is fused to a DNA-binding domain: the other to a transcriptional activator domain. Any interaction between them is detected by the formation of a transcription factor. The experiments obtained 4549 interactions among 3278 proteins [12].

Uetz: Data set was also obtained using the two-hybrid experiments, they obtained 957 interactions among 1003 proteins [13]. Even though the two data sets were obtained using the same experimental technique they share an unexpected small number of pair of interacting proteins: 141.

Mering with high confidence: Mering data set was obtained using contributions from three different methods: conserved gene neighborhood, co-occurrence of genes and gene fusion events. The data set contained 988 proteins and 2455 interactions.

Mering with high and medium confidence: This data set was obtained using the same methods from the other one but using more restrictive conditions, for details see Ref. [1] and supplementary material. The data set contained 2617 proteins and 11,855 interactions.

Ideker: This data set was proposed by Ideker et al. that used an integrated approach to investigate galactose metabolism, using DNA micro-arrays, quantitative proteomics and databases of physical interactions of proteins. The data set contains 722 proteins and 612 interactions.

3. Results

In Fig. 1 we show the histogram of connectivity for each data set. The curves are well approximated by power laws $P(k) = Ak^{-\gamma}$ for all data sets. In Table 1 the γ

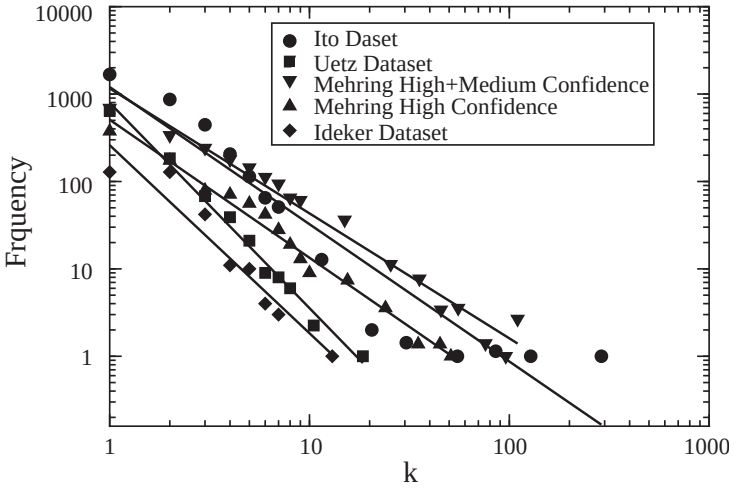


Fig. 1. Histogram of connectivity for the different data sets used in this work. The data are well described by power laws: $\sim k^{-\gamma}$. The slopes (γ) of the curves were obtained using a least-squares algorithm.

Table 1
Slopes for histogram of connectivity, damage, and clusterization coefficient $C(K)$

Data sets	Connectivity	Damage	$C(k)$
Ito	1.6	2.6	1.2
Uetz	2.3	1.9	0.7
Mering high	1.6	2.3	0.1
Mering high + medium	1.4	3.00	0.03
Ideker	2.1	1.4	0.6

values are presented. These values are coherent with results in the literature [8] indicating that the graphs belong to the class of scale-free networks. The results are robust and consistently observed on an integrated model such as the one represented by Ideker et al. model, suggesting that integrated networks are also scale-free.

We also investigated the dependence of C, the clusterization coefficient on the connectivity k. For a traditional scale-free graph we expect $C(k)$ not to depend on k, while for hierarchical networks, $C(k) \sim k^{-\beta}$. Fig. 2 shows that the results from the two-hybrid methods are consistent with hierarchical networks, however, the Mering data set does not seem so. These incompatibility is credited to the low accuracy of the two-hybrid experiments or to the fact that the Mering data set reflects functional relationships between two proteins. In Table 1 we present best fits assuming a functional dependency of the type $C(k) \sim k^{\gamma}$.

We calculate the damage as a different topological measure of the influence of a given protein in the network. It calculates the number of proteins that are disconnected from the network when the protein under study is deleted. Assuming that information is distributed through the network, disconnected proteins stop sharing information with the remaining network. From this perspective, even

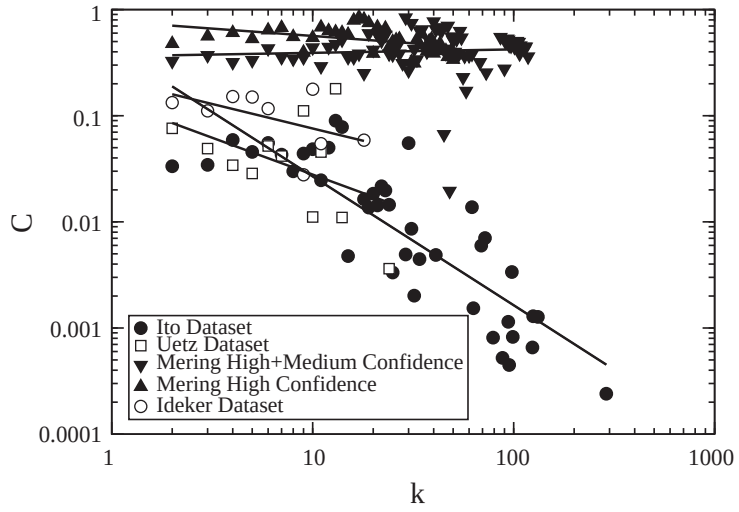


Fig. 2. The dependence of the clusterization coefficient C on the connectivity k. The results show that the Uetz, Ito, and Ideker sets are compatible with an hierarchical scale-free model. The graph generated from the Mering set does not seem hierarchical.

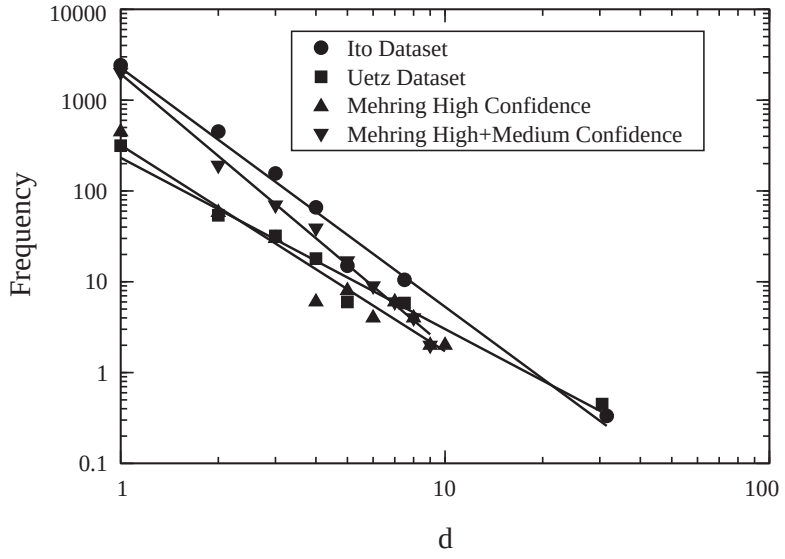


Fig. 3. Histogram for damage for the different data sets. The data are well described by power laws: $Ak^{-\gamma}$. The γ values were obtained by fitting the curves using a least-squares algorithm.

proteins with few connections can cause large damage when removed if they generate large disconnected clusters of proteins, assuming that the scale-free topology of the PPI network is an evolutionary strategy towards the robust functioning of the system. Robustness in proteins networks is represented by a large degree of redundancy in the system. We expect the damage to be able to identify redundant and non-redundant proteins.

In Fig. 3 we present the histogram for the damage, showing that the damage caused by the majority of the proteins is small and obeys a power law. A similar result was found in our analysis of metabolism [9].

We have investigated the dependence of damage on connectivity, the results are presented in Fig. 4. Damage and connectivity are correlated quantities, we observe that only highly connected proteins cause a considerable damage, this is a direct consequence of the assortative property, since the excluded proteins are usually weakly connected. So the scale-freeness and the assortativeness collaborate to reduce damage [8]. The graph also presents a very small number of bridges (proteins with two connections) whose deletion causes appreciable damage.

To compare either connectivity and damage as measures of protein importance, we determine which one has higher correlation with essentiality. The results are presented in Fig. 5. We also studied the Uetz and Ito sets without the enzymes. This was done because the relevance of enzymes could be related not only to its topological properties, but also to the role they play on the metabolic network, so we might expect that even weakly connected enzymes on PPI networks may cause severe disruption on metabolic networks.

We have measured the S values for each curve, the results are presented in Table 2, S deviation from 0.5 implies larger correlation between two quantities. Based on this analysis we conclude that

- connectivity is strongly correlated to essentiality for all cases except for the Ideker set,
- the exclusion of enzymes from the construction of L implies an increase on the correlation, but the gain is marginally small,
- on the Mering data set, S_d increases if we consider the high confidence data sets, while S_k decreases.

If damage is correlated to essentiality we expect that we consider the set of proteins deleted by the exclusion of an enzyme i, if this set contains an essential

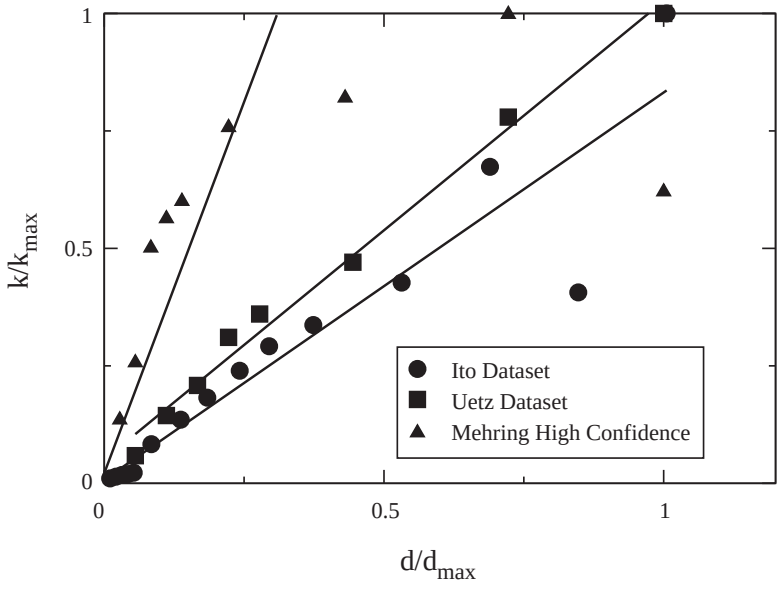


Fig. 4. The dependence of damage on connectivity. The graph shows that damage is correlated with connectivity.

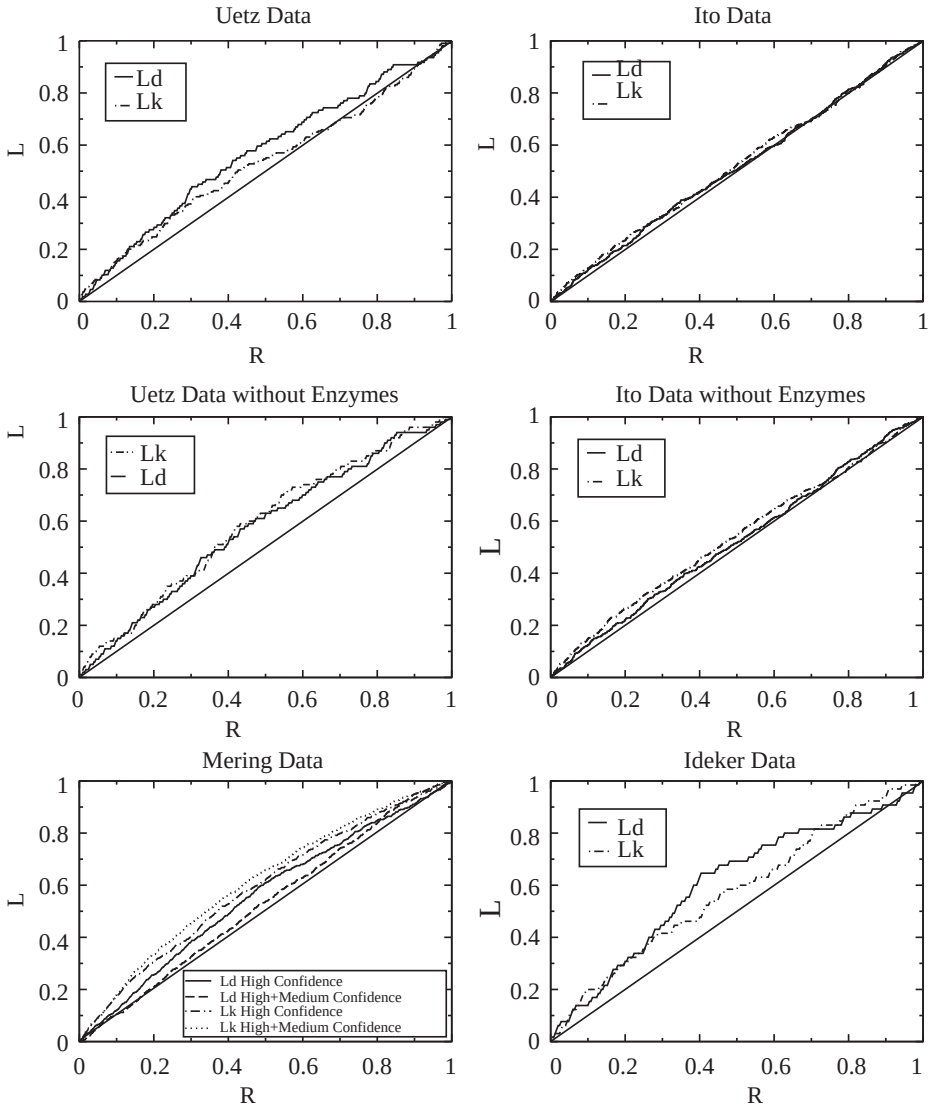


Fig. 5. The L_d and L_k function against the normalized rank for Uetz data without enzymes. The curvatures indicate a very weak correlation between both connectivity and damage to lethality.

Table 2
Comparison of different data sets of protein–protein interactions, S measures the correlation between a topological parameter with lethality, for an uncorrelated parameter $S = 0.5$

Data sets	S_d	S_k
Mehring high confidence	0.56	0.59
Mehring high + medium confidence	0.53	0.61
Ito	0.51	0.52
Ito without enzymes	0.52	0.54
Uetz	0.53	0.57
Uetz without enzymes	0.57	0.59
Ideker	0.60	0.59

In this scenario larger values of S indicate stronger correlation, see text for details.

Table 3
The number of essential/non-essential (n/e) ORFs whose deletion implies the deletion of at least one essential ORF (denoted as e) or only non-essential ORFs (denoted as n)

	# ORFs	Random model
High confidence Mering		
e → e	46	25 ± 2
n → e	2	17 ± 2
e → n	16	36 ± 2
n → n	36	22 ± 2
Ideker		
e → e	14	5 ± 2
n → e	0	4 ± 1
e → n	26	31 ± 1
n → n	41	37 ± 1
Uetz		
e → e	14	5 ± 2
n → e	4	12 ± 2
e → n	36	44 ± 2
n → n	101	93 ± 2

We compare these numbers with the expected values for a random model, obtained by Monte Carlo simulation. The data were obtained using high confidence Mering, Ideker and Uetz data sets. These data show that if the set formed by the proteins excluded by the deletion of a protein contains an essential protein, the excluded one has a high probability of also being essential.

protein, i will have a high probability of also being essential. We tested this idea on Mering high confidence, Uetz and Ideker data sets, by measuring the fraction of essential proteins that deleted some essential proteins ($e \to e$ in Table 3), the fraction of essential proteins that deleted only non-essential ones ($e \to n$), and the same quantities for non-essential enzymes. The results are presented in Table 3. These results were compared to random models where we have randomly exchanged the character of the proteins. In all cases, we show that the results have a strong bias toward our expectation.

4. Discussion

The results show conclusively that both damage and connectivity are correlated to essentiality. If we consider only PPI networks, connectivity presents higher correlation with essentiality than damage, however, for integrated networks, such as the Ideker set, damage presents higher correlation. The interpretation of these results is not totally clear and we propose two hypotheses:

- The influence of a protein is localized and does not spread through the network being concentrated on proteins with which it interacts directly.

- Damage is very sensitive to errors in the data sets. For example, consider a highly connected protein with high damage. A single missing interaction does not appreciably change its connectivity, however, if this missing interaction is a bridge involving the protein in the network its damage is actually null.

We believe that the second hypothesis is true to a given extent. We have no conditions to test the second one yet. If we consider integrated networks, damage seems to be a better measure of protein importance.

We have also addressed the question of classifying the topology of these networks. For all data sets, they seem to belong to the class of hierarchical scale-free networks with the exception of the Mering set that is determined from functional interactions and not from physical interactions. Even if our results are not totally conclusive, we have raised important questions to be answered in additional investigations in the field to contribute to the understanding of the machinery underlying life.

Acknowledgements

We acknowledge the support of CNPq and FAPERGS. This work was developed in collaboration with HP Brazil R&D.

References

[1] C. von Mering, et al., Nature 417 (2002) 399.
[2] E.M. Marcotte, et al., Nature 402 (1999) 83.
[3] H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, Nature 411 (2001) 41.
[4] http://www.genome.ad.jp/brite.
[5] J.C. Rain, et al., Nature 409 (2001) 211.
[6] S.-H. Yook, Z.N. Oltvai, A.-L. Barabasi, Proteomics 4 (2004) 928.
[7] S. Maslov, K. Sneppen, Science 296 (2002) 910.
[8] A.-L. Barabasi, Z.N. Oltvai, Nat. Rev. Genetics 5 (2004) 101.
[9] N. Lemke, et al., Bioinformatics 20 (2004) 115 (Evaluation Studies).
[10] T. Ideker, et al., Science 292 (2001) 929.
[11] H. Jeong, Z.N. Oltvai, A.L. Barabasi, ComplexUS 1 (2002) 19.
[12] T. Ito, et al., Proc. Natl. Acad. Sci. USA 98 (2001) 4569.
[13] P. Uetz, et al., Nature 403 (2000) 623.

ELSEVIER

Computers in Biology and Medicine 36 (2006) 542–552

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Bioinformatics analysis of mycoplasma metabolism: Important enzymes, metabolic similarities, and redundancy

José C.M. Mombach*, Ney Lemke, Norma M. da Silva, Rejane A. Ferreira, Eduardo Isaia, Cláudia K. Barcellos

Laboratório de Bioinformática e Biologia Computacional, Universidade do Vale do Rio dos Sinos, 93022-000 São Leopoldo, RS, Brazil

Received 1 November 2004; accepted 9 March 2005

Abstract

In this work we apply a bioinformatics approach to determine the most important enzymes of the metabolic network of mycoplasmas. The genomes of several mycoplasmas shared predicted important enzymes. Our method allows us to determine both enzymes that are isolated from the metabolic network of the organism and those that are redundant. We also compare the similarities of the mycoplasmas metabolic networks with the phylogenetic relationships predicted from their 16s rRNA sequences.

Keywords: Mycoplasma; Metabolism; Drug targets

1. Introduction

The complete genome sequences of several organisms are now available and a major challenge is the extraction of relevant physiological information from these data. The metabolic networks reconstructed from the annotated genomes provide a plethora of information to explore by bioinformatics approaches [1].

Here we apply to mycoplasmas our recently developed graph based technique for the analysis of metabolism that predicts important enzymes [2]. Genomic and metabolic data of mycoplasmas were obtained from the KEGG database [3].

* Corresponding author. Tel.: +55 51 5908 161; fax: +55 51 590 8162.
E-mail address: mombach@exatas.unisinos.br (J.C.M. Mombach).

0010-4825/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2005.03.004

Table 1
Phenotypic features of the mycoplasmas investigated in this paper

Species	Host	% G+C in DNA	Genome size (Kbp)
M. penetrans	Human	26	1358
M. gallisepticum	Bird	31	996
U. urealyticum	Animals	25.5	751
M. pulmonis	Rodent	26.6	963
M. genitalium	Human	32	580
M. pneumoniae	Human	40	816

Cellular metabolism characterizes a complex network of reactants connected through chemical reactions catalyzed by specialized proteins, the enzymes. The reactions organize into modules with specific catabolic or anabolic functions called metabolic maps and the complete set of metabolic maps forms the metabolic network. An exponentially growing number of organisms have sequenced genomes [4]. Assuming that the annotated proteins are expressed, we can reconstruct the metabolic network of the organism [1].

Determining of the influence of enzymes on the network is a critical issue for bioengineering and pharmaceutical industry since important enzymes can be targets for drugs [1] or genetically engineered to change the production of specific metabolites [5].

Our approach analyzes the static structure of the components of the network to infer causal and physiological relationships. In a previous paper we applied our method to *Escherichia coli* which has a well studied metabolism and found a correlation between our quantitative definition of the importance of an enzyme in the metabolic network and the probability that the enzyme is essential [2] demonstrating that the method has predictive power to determine important enzymes.

We now apply our method to several sequenced genomes of mycoplasmas, prokaryotes belonging to the Mollicute class with a reduced genome (varying from 580 to 1350 Kbp) with low G-C content. For comparison, *E. coli*, *Bacillus subtilis*, and *C. perfringens* have a genome size of 4.6 Mbp, 4.2 Mbp, 3.08 Mbp, respectively. The mycoplasmas differ phenotypically from other bacteria due to their reduced size and absence of a cellular wall. Some species infect humans and animals, making them an important issue for health organizations. These aspects make mycoplasmas a very interesting bacterial group to apply our methodology, since we can use our method to identify potential candidate drug targets. In [2] the methodology identified essential enzymes, in this work we extend the method to determine the most important biochemical reactions in the organism.

Mycoplasmas seem to have evolved more rapidly than other bacteria since some highly variable positions in their rRNA sequences are strongly conserved in other bacteria species [6]. Their interesting evolutionary history showing massive reduction in genome size motivated us to investigate their metabolism. The general belief is that they have minimal genomes, so we would expect a low level functional redundancy.

Table 1 presents some phenotypic characteristics of the six mollicutes investigated in this study listing their host, genome G-C content, and genome size in kilobase pairs.

2. Methodology

We have introduced a new quantitative criterion for enzyme importance: the damage its removal causes to the metabolic network [2]. In the absence of complete information about kinetic parameters and the

influence of the regulatory network, we cannot predict all consequences of the deletion of a specific enzyme, however, we assume that the essentiality of a protein need not depend on its level of expression. Rocha and Danchin [7] have shown for *E. coli* and *B. subtilis* that the level of expression of a gene does not relate to its essentiality, supporting this hypothesis.

Using only information about the reactions occurring in an organism we can determine the number of metabolites whose production the absence of the enzyme prevents, which we define as the damage d to the network. d is a measure of the number of disrupted pathways that the removal of the open reading frame (ORF) coding an enzyme in the metabolism generates.

To build the network of metabolic reactions for an organism, we collect all annotated enzyme codes (EC numbers) from the databases cited above and collect all reactions associated with each EC in the corresponding metabolic map using the KEGG database. This initial metabolic set of reactions can present inconsistencies, since the annotation of biological sequences is error prone [4].

In order to improve the consistency of biological information available, we are developing an ontology called MOlecular NETwork (MONET) as a model to integrate information from different databases [8].

Among the most common problems we find in annotated genomes are wrong or missing attribution of function to genes. For metabolic analysis these errors generate inconsistent sets of chemical reactions with missing or false reactions. In our analysis we verify the consistency of this set, remove impossible reactions and introduce highly probable missing reactions. The reactions produce or use metabolites from other reactions helping our consistency analysis. For example, if the reaction corresponding to an annotated EC may require a metabolite which no other reaction produces and that is not used in any other reaction, we do not include it in the input reaction set for the simulation. In some cases a metabolite that is used by the cell can come from an external source and is difficult to determine whether this is the case or not. In this situation we analyze the reactions that use it to decide.

In addition, some reactions require metabolites that are not produced inside the cell, however if what is produced is used in other reactions, we determine all external metabolites required and use them as inputs for the simulation. Consequently, our generated set of chemical reactions has higher confidence than one obtained solely from annotations, though it still depends on the accuracy of the annotations.

Our metabolic analysis uses a graphical representation of metabolism. The graph is directed and has two types of nodes (formally we classify this structure as a bipartite graph) [9]. One node type represents chemical reactions and the other metabolites (see Fig. 1). A link between a reaction and a metabolite points towards the metabolite, if the metabolite is a product, and in the opposite direction, if the metabolite is a reactant. We treat reversible reactions as two separate reactions. We constructed graphs from the list of reactions catalyzed by all enzymes involved in the small molecule metabolism of mycoplasmas using the reaction set in the KEGG [3] database. Small molecule metabolism is a subset of the complete metabolism that excludes DNA replication and protein synthesis reactions [10].

To calculate the damage d, we use two different methods (methods A and B below). In method A we choose an ORF and delete all reactions catalyzed by the enzyme it produces. In method B we delete reactions in the network sequentially. We define d to be the number of deleted metabolites is defined as d and calculate it using the algorithms following:
Method A:

- First we choose an ORF coding an enzyme and determine the reaction(s) it catalyzes.
- If the reaction is irreversible we delete the nodes representing the metabolites the enzyme produces.

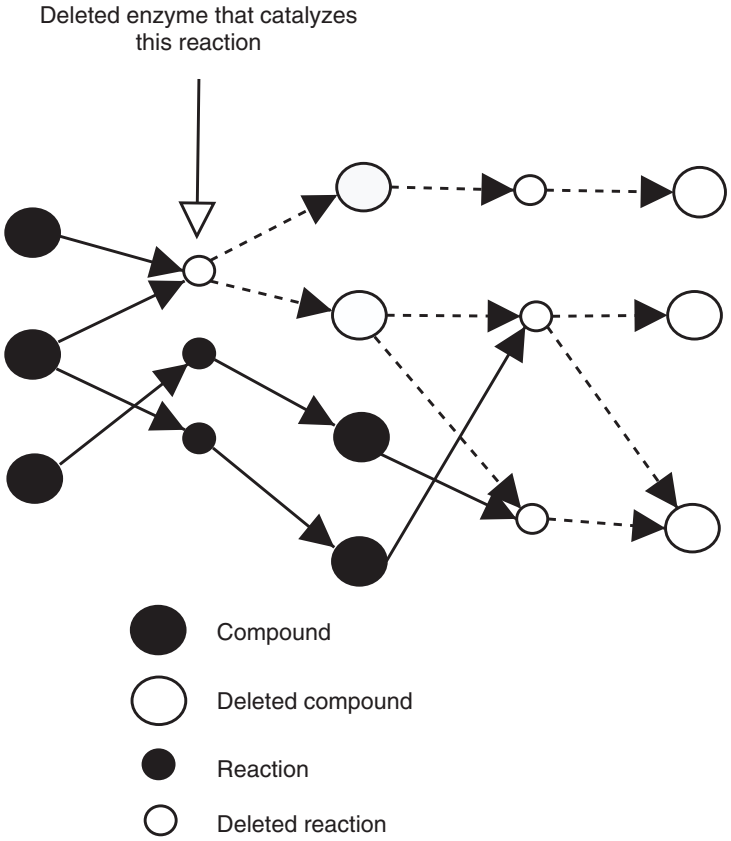


Fig. 1. Schematic representation of damage in a metabolic network. Small and large nodes represent chemical reactions and metabolites, respectively. Open symbols represent the absence of a chemical reaction or metabolite. The figure illustrates the effect on reactions and metabolites when a delete a chemical reaction at the left.

- If the reaction is reversible we delete all metabolites produced which the forward and reverse reactions the enzyme catalyzes.
- We determine the remaining set of reactions that occur with the available metabolites.
- We iterate until we reach a fixed point.

Method B:

- We choose a reaction and delete its node from the network.
- If the reaction is irreversible we delete all metabolites it produces.
- If the reaction is reversible we delete all metabolites produced which the forward and reverse reactions the enzyme catalyzes.
- We also delete repeated reactions catalyzed by the same enzyme which are coded for by different ORFs.
- We determine the set of remaining reactions that occur with the available metabolites.
- We iterate until we reach a fixed point.

Systematic mutagenesis experiments which determine the importance of each gene for the survival of an organism motivated Method A. We can think of this algorithm as an in silico knockout of the ORF coding an enzyme.

Method B simulates the action of drugs on enzymes. Molecules designed to act on proteins, dock to the active site of the protein, blocking its function. Enzymes catalyze reactions so when their activity is blocked, the reaction becomes completely inefficient for biological purposes. Thus we remove reactions which one or more ORFs can catalyze and determine the damage each one generates.

Our method identifies enzymes whose reactions are isolated from the metabolic network of the organism. Such enzymes are interesting examples to investigate evolutive questions that can explain their presence in an organism.

We also used the phylogenetic relationships among the organisms to compare their metabolic networks. The phylogenetic analysis used the software MEGA 2.1 [11]. We built the trees using neighbor-joining with the *p*-distance method and maximum parsimony. Bootstrap tests we did with 1000 samples. We used 16s rRNA sequences from GenBank with the following access numbers: AP004174 (*Mycoplasma penetrans*), L35043 (*Mycoplasma gallisepticum*), AF073455 (*Ureaplasma urealyticum*), AF132741 (*Mycoplasma pneumoniae*), U39694 (*Mycoplasma genitalium*), NC002771 (*Mycoplasma pulmonis*) and AB055007 (*B. subtilis*).

The next section presents our results for the mycoplasmas in Table 1. The data correspond to December 2002 update of KEGG. We corrected errors in the data manually.

3. Results and analysis

Little data is available on the essentiality of ORFs in mycoplasmas, so our simulation predictions require experimental validation [12].

Tables 2 and 3 lists the enzymes and reactions for which Method A and B, respectively, predict high damage for all mycoplasmas investigated. To summarize the results in Tables 2 and 3, we selected only enzymes with damage higher than one for at least one mycoplasma. Enzyme 1 in Table 2 is known to be essential for *M. genitalium* [12].

Next, we describe briefly the function of the enzymes in Tables 2 and 3.

Table 2
Damage caused by ORF deletion calculated using Method A (see text)

Enzyme	Mpe	Mga	Uur	Mpu	Mge	Mpn
1. 2.4.2.1	5	**7**	7	5	7^{e}	**9**
2. 2.7.1.69	**7**, 4	**7**, 4	♯	0, 2	**9**, 4	4, 0
3. 2.3.1.12	3	5	♯	5	5	5
4. 2.7.8.7	5	4	♯	5	?	4
5. 1.8.1.4	3	5	♯	0	5	5
6. 3.1.4.14	5	4	♯	5	♯	4
7. 2.7.1.40	0	3	**8**	3	3	3
8. 3.1.3.5	♯	♯	♯	**8**	♯	♯

Bold face numbers indicate the enzymes with the highest damage for each species. Damage 0 means that the catalyzed reaction is redundant, i.e. other reactions replace it. Two different values for damage imply that two different ORFs code for the same enzyme in the genome and that they function in different metabolic reactions. ♯ means that the ORF is not found in the genome. *e* means that the ORF is essential for the organism. ? means that the reaction is isolated in the metabolic network of the corresponding organism and that we removed it from our calculation. Mpe: *M. penetrans*; Mga: *M. gallisepticum*; Uur: *U. urealiticum*; Mpu: *M. pulmonis*; Mge: *M. genitalium*; Mge: *M. pneumoniae*.

Table 3
Damage of enzyme deletion calculated using Method B (see text)

Enzyme EC	Mpe	Mga	Uur	Mpu	Mge	Mpn
1. 2.7.1.69	0, 1, 2, **7**	0, 1, 2, **7**	∄	0, 1, 2, 3, **8**	0, 1, 2, 3, **9**	0, 1, 2, 3, **8**
2. 2.3.1.12	3, 0	5, 0	∄	5, 0	5, 0	5, 0
3. 1.8.1.4	3, 0	5, 0	∄	5, 0	5, 0	5, 0
4. 2.7.8.7	5	4	∄	5	?	4
5. 3.1.4.14	5	4	∄	5	∄	4
6. 2.3.1.51	0, 2	0, 2	0, **6**	0, 2	0, 2	0, 2
7. 3.1.1.3	∄	∄	4, 1	∄	∄	∄
8. 2.7.7.41	∄	2	4	2	2	2
9. 3.5.4.9	0, 2	1, 2	4, 0	1, 2	1, 2	1, 2

Bold face numbers indicate the enzymes with highest damage for each species. Damage 0 means that the catalyzed reaction is redundant, i.e. other reactions replace it. More than one value for damage implies that different reactions catalyzed by an enzyme coded for by different ORFs were deleted separately, generating a different value for damage. ∄ means that the ORF is not found in the genome. ? means that the reaction is isolated in the metabolic network of the corresponding organisms and that we removed it from our calculation. Mpe: *M. penetrans*; Mga: *M. gallisepticum*; Uur: *U. urealiticum*; Mpu: *M. pulmonis*; Mge: *M. genitalium*; Mge: *M. pneumoniae*.

The enzyme purine nucleoside phosphorylase (EC 2.4.2.1) is central to all purine and pyrimidine metabolism. It can also catalyze ribosyltransferase reactions of the type nucleoside ribosyltransferase (EC 2.4.2.5) catalyzes. The protein-Np-phosphohistidine-sugar phosphotransferase (EC 2.7.1.69) consists of a group of related enzymes. This enzyme is a transmembrane protein that transports mainly fructose, mannose and glucose into the cell. Each enzyme translocates the sugar it phosphorylates into bacteria. Aldohexoses and their glycosides and alditols are phosphorylated on O-6, whereas fructose and sorbose are phosphorylated on O-1. Glycerone and disaccharides are also substrates. The protein substrate is a phosphocarrier protein of low molecular mass (9.5 kDa).

Dihydrolipoamide S-acetyltransferase (EC 2.3.1.12) forms a multimer (24-mer or 60-mer, depending on the source) core of the pyruvate dehydrogenase multienzyme complex, and binds tightly to both pyruvate dehydrogenase (acetyl-transferring) (EC 1.2.4.1) and dihydrolipoyl dehydrogenase (EC 1.8.1.4). EC 1.2.4.1 reductively acetylates the lipoyl group of this enzyme, and the only observed direction catalyzed by EC 2.3.1.12 is that where the acetyl group passes to coenzyme A. The enzyme 2.3.1.12 produces CoA that connects different metabolic pathways. We call these enzymes bridges and in *E. coli* we found that many enzymes with this feature were essential [2].

Holo-[acyl-carrier protein] synthase (EC 2.7.8.7) is a transferase for other substituted phosphate groups. This enzyme produces acyl-carrier-protein (acp). It is a bridge between metabolic pathways for fatty acids metabolism.

Dihydrolipoamide dehydrogenase (EC 1.8.1.4) is an oxidoreductase acting on a sulfur group of donors with NAD+ or NADP+ as acceptors. This enzyme is central in pyruvate metabolism and produces acetil-CoA, which is necessary for many other pathways, including fatty acids metabolism.

Acyl-carrier-protein phosphodiesterase (EC 3.1.4.14) is a hydrolase acting on ester bonds, producing and using acyl-carrier-protein.

Pyruvate kinase (EC 2.7.1.40) is a phosphotransferase with an alcohol group as acceptor. UTP, GTP, CTP, ITP, and dATP can also act as donors. It also phosphorylates hydroxylamine and fluoride in the presence of CO_2. This enzyme in central to glycolysis and pyruvate metabolism. It produces ATP as well.

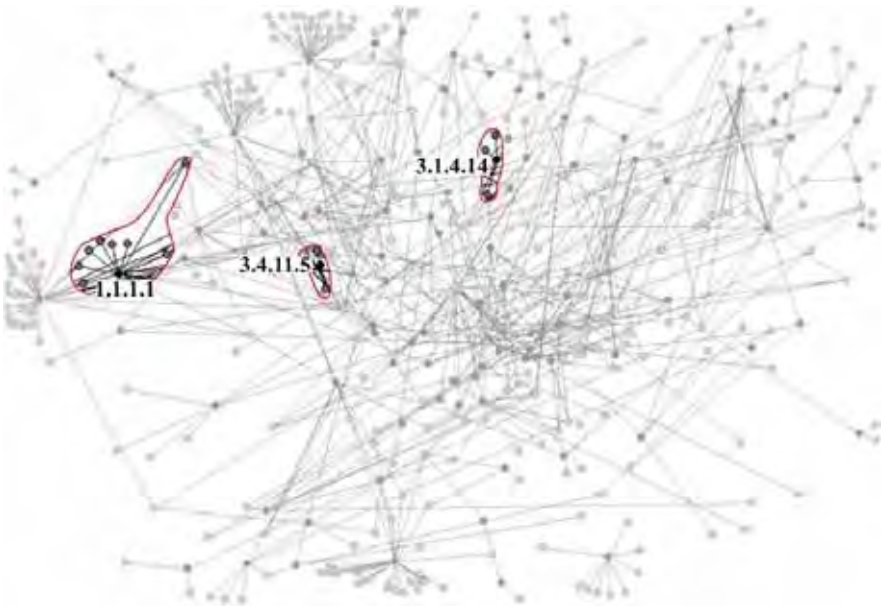


Fig. 2. Graph of the metabolism of *M. penetrans*. Dark gray circles represent enzymes and light gray circles metabolites the enzymes in the reactions (represented by edges) use or produce. The text discuss marked enzymes and their corresponding metabolites.

5′-Nucleotidase (EC 3.1.3.5) is a hydrolase acting on ester bonds. This enzyme is central to all purine and pyrimidine metabolism.

1-acylglycerolphosphate acyltransferase (EC 2.3.1.51) transfers groups other than amino-acyl groups. Acyl-[acyl-carrier protein] can also act as an acyl donor. It acts in glycerolipid metabolism.

Triacylglycerol lipase (EC 3.1.1.3) is involved in the hydrolysis of carboxylic esters.

Phosphatidate cytidylyltransferase (EC 2.7.7.41) transfers phosphorus-containing groups and acts on glycerolipid metabolism.

Methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9) is a hydrolase that acts on carbon–nitrogen bonds. In some prokaryotes, it occurs as a bifunctional enzyme that also has dehydrogenase (EC 1.5.1.5) activity or formiminotetrahydrofolate cyclodeaminase (EC 2.3.1.7) activity.

Fig. 2 shows the metabolic network of *M. penetrans*, as an example of the graph our method generates. We also include the enzymes whose catalyzed reactions are isolated in the metabolic network of the organism. Comparison of the mycoplasmas networks reveals interesting differences. For instance, only *M. genitalium* and *U. urealyticum* lack enzyme acyl-carrier-protein phosphodiesterase (EC 3.1.4.14). The reaction catalyzed by enzyme prolyl aminopeptidase (EC 3.4.11.5) is isolated in all organisms where it occurs: *M. pneumoniae*, *M. penetrans*, and *M. genitalium*. The enzyme alcohol dehydrogenase (1.1.1.1) is present only in *M. penetrans*.

Using the information in Figs. 2 and 3 we compared the mycoplasmas metabolism and phylogeny. The phylogenetic relationship in Fig. 3 is the same obtained using maximum parsimony (not shown). The results show that species that are distant phylogenetically share enzymes that close do not share. Independent loss events in the evolution of these organisms may explain such surprising differences. According to Weisburg [6], genome size reductions seem to have occurred more than once in the Mollicute class. Mollicutes have evolved by reductive or degenerative evolution, followed by significant losses

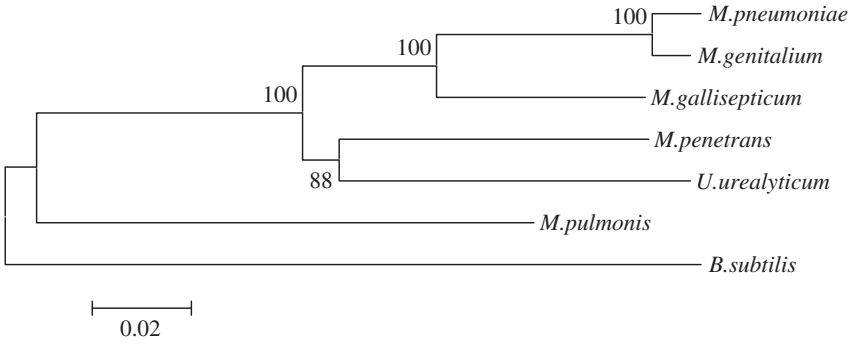


Fig. 3. Phylogenetic tree obtained using the neighbor-joining method on the 16S rRNA sequences of Mollicutes. We use the sequence from *B. subtilis* as an outgroup.

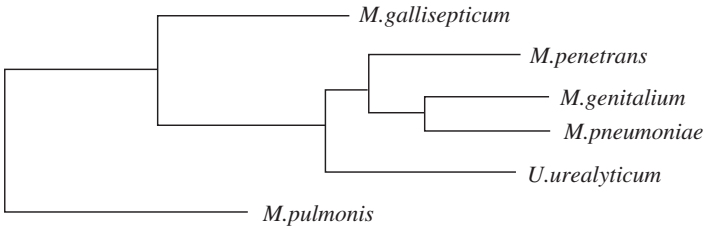


Fig. 4. Dendogram generated by a distance matrix calculated from the number of enzymes each pair of species shares.

of genomic sequences. Smaller chromosomes allow faster bacterial replication providing significantly evolutionary pressure to shrink bacterial genomes [13]. For instance, *M. pneumoniae* and *M. genitalium* apparently lost all genes involved in amino acid biosynthesis and need to obtain amino acids from the host or artificial culture medium. Mycoplasmas also lost a large fraction of the genes involved in the biosynthesis of cofactors and many do not synthesize fatty acids [14].

Another phenomenon that can explain partially these metabolic and phylogenetic differences is horizontal DNA transfer. Studies of genes and genomes indicate that a considerable amount of horizontal transfer has occurred among prokaryotes, more of genes involved in housekeeping (operational genes) than of genes involved in transcription, translation, and related processes (informational genes) [15]. A possible example of horizontal transfer is the ORF of the enzyme carnitine O-acetyltransferase (EC number 2.3.1.7) found only in the genomes of *M. pneumoniae* and *M. pulmonis* which are phylogenetically distant according to Fig. 3. These mycoplasmas parasitize humans and rodents, respectively, which also possess this gene. Thus the mycoplasmas may have acquired the gene from their vertebrate hosts. Interestingly, bacteria from the genera Bacillus and Clostridium, that are closer to Mollicute class, do not possess this gene which would be surprising if it has been inherited from a common ancestor. Only 10 organisms in KEGG, 7 eukaryotes and 3 prokaryotes have this gene, the third prokaryote being the bacterium *Streptomyces coelicolor*. An alternative explanation for the presence of this ORF in the annotated genome organism is incorrect attribution of gene function.

In order to infer phylogeny from the metabolic networks, we generated a dendogram (Fig. 4) using a distance matrix whose elements are the ratios of the number of enzymes shared by each pair of species to the total number of enzymes in both species. We generated the dendogram using the software PHYLIP 3.6a2 [16]. Our dendogram is close to the phylogenetic tree in Fig. 3. For example, *M. genitalium* and

Table 4
Estimate of metabolic redudancy in mycoplasmas

Organism	Genome size (Kbp)	ORF deletion (%)	Reaction deletion (%)
M. penetrans	1358	42.35	62.57
M. gallisepticum	996	35.08	64.18
U. urealyticum	751	35.71	65.68
M. pulmonis	963	55.69	54.94
M. genitalium	580	39.98	61.64
M. pneumoniae	816	48.61	54.82

The percentages represent the fraction of the total number of ORFs (method A) or reactions (method B) that produce null damage in our calculation.

M. pneumoniae correctly group in the same clade showing that this pair of species share both 16s rRNA sequence and metabolic similarities.

Table 4 estimates the metabolic redundancy for the mycoplasmas. We consider a reaction redundant if other reactions can replace it, which in our simulation implies no damage. We determined the percentage of the total number of ORFs and reactions that generate null damage in the simulated network of each organism using both method A and B. The percentages in Table 4 show more redundancy than we would expect for organisms with minimal genomes. We find no correlation between genome size and the amount of redundancy.

Our integrated analysis suits investigations of mycoplasma metabolism, indicating enzymatic drug targets and by identifying isolated enzymes revealing both possible horizontal gene transfer and biochemical artifacts of evolutionary genomic reduction.

4. Summary

The complete genome sequences of several organisms are now available and a major challenge is the extraction of relevant physiological information from these data. Here we apply to mycoplasmas, that are small bacteria, our recently developed graph based technique for the analysis of metabolic networks that predicts important enzymes. Determining the influence of enzymes on the network is a critical issue since important enzymes can be targets for drugs or genetically engineered to change the production of specific metabolites. In this work we apply a bioinformatics approach to determine the most important enzymes of the metabolic network of mycoplasmas. Our results show that the genomes of several mycoplasmas shared predicted important enzymes. Our method allows us to determine both enzymes that are isolated from the metabolic network of the organism and those that are redundant. We also compare the similarities of the mycoplasmas metabolic networks with the phylogenetic relationships predicted from their 16s rRNA sequences. Our integrated analysis suits investigations of mycoplasma metabolism, indicating enzymatic drug targets and by identifying isolated enzymes revealing both possible horizontal gene transfer and biochemical artifacts of evolutionary genomic reduction.

Acknowledgements

We acknowledge the support of CNPq Grants 521089/2001-8 and 550042/2003-2, FAPERGS and the KEGG database for providing public access to its data.

References

[1] P.D. Karp, M. Krummenacker, S. Paley, J. Wagg, Integrated pathway-genome databases and their role in drug discovery, Trends Biotechnol. 17 (1999) 275–281.
[2] N. Lemke, F. Herédia, C.K. Barcellos, A.N. Reis, J.C.M. Mombach, Essentiality and damage in metabolic networks, Bioinformatics 20 (2004) 115–119.
[3] M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30.
[4] D. Devos, A. Valencia, Intrinsic errors in genome annotation, Trends Genet. 17 (2001) 429–431.
[5] J.S. Edwards, B.O. Palsson, How will bioinformatics influence metabolic engineering, Biotechnol. Bioeng. 58 (1997) 162–169.
[6] W.G. Weisburg, J.G. Tully, D.L. Rose, J.P. Petzel, H. Oyaizu, D. Yang, L. Mandelco, J. Sechrest, T.G. Lawrence, J. Van Etten, J. Maniloff, C.R. Woese, A phylogenetic analysis of the mycoplasmas: basis for their classification, J. Bacteriol. 171 (1989) 6455–6467.
[7] P.C.R. Rocha, A. Danchin, Essentiality, not expressiveness drives gene-strand bias in bacteria, Nat. Genet. 34 (2003) 377–378.
[8] E. Battistella, J.G.C. de Souza, C.K. Barcellos, N. Lemke, J.C.M. Mombach, MONET: the MOlecular NETwork Ontology, Genet. Mol. Res., to be published.
[9] G. Chartrand, Introductory Graph Theory, Dover Publications, New York, 1977.
[10] A. Masanori, The metabolic world of *Escherichia coli* is not small, Proc. Natl. Acad. Sci. USA 101 (2004) 1543–1547.
[11] S. Kumar, K. Tamura, I.B. Jakobsen, M. Nei, MEGA 2: molecular evolutionary genetics analysis software, Bioinformatics 17 (2001) 1244–1245.
[12] C.A. Hutchison III, S.N. Peterson, S.R. Gill, R.T. Cline, O. White, C.M. Fraser, H.O. Smith, J.C. Venter, Global transposon mutagenesis and a minimal mycoplasma genome, Science 286 (1999) 2165–2169.
[13] J. Maniloff, The minimal cell genome: on being the right size, Proc. Natl. Acad. Sci. USA 93 (1996) 10004–10006.
[14] S. Razin, D. Yogev, Y. Naot, Molecular biology and pathogenicity of mycoplasmas, Microbiol. Mol. Biol. Rev. 62 (1998) 1094–1156.
[15] R. Jain, M.C. Rivera, J.A. Lake, Horizontal gene transfer among genomes: the complexity hypothesis, Proc. Natl. Acad. Sci. USA 96 (1999) 3801–3806.
[16] J. Felsenstein, PHYLIP (Phylogeny Inference Package), version 3.6a2, Distributed by the author, Department of Genetics, University of Washington, Seattle, 1993.

José Carlos Merino Mombach was born in Porto Alegre, Brazil, in 1964. He received the degree of Doctor in Sciences from the Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, in 1997. Dr. Mombach has been an Adjunct professor at the Graduate School in Applied Computing, Universidade do Vale do Rio dos Sinos, São Leopoldo, Brazil, since 1999. Since 2002 he has been a researcher of the Laboratório de Bioinformática e Biologia Computacional. His main research interests are bioinformatics, computational biology, and complex systems.

Ney Lemke was born in Porto Alegre, Brazil, in 1969. He received the degree of Doctor in Sciences from the Universidade Federal do Rio Grande do Sul in 1997. Dr. Lemke has been an Adjunct professor at the Graduate School in Applied Computing, Universidade do Vale do Rio dos Sinos since 1999. Since 2002 he has been a researcher of the Laboratório de Bioinformática e Biologia Computacional at the Universidade do Vale do Rio dos Sinos. His main research interests are bioinformatics, computational biology, disordered systems, artificial intelligence, and complex systems.

Norma Machado da Silva was born in Porto Alegre, Brazil, in 1973. She received the degree of Master in Genetics and Molecular Biology from the Universidade Federal do Rio Grande do Sul in 2003. Since 2003 she is a researcher of the Laboratório de Bioinformática e Biologia Computacional at the Universidade do Vale do Rio dos Sinos. Her main research interests are genetics and bioinformatics.

Eduardo Isaia Filho was born in Santa Maria, Brazil, in 1978. He received the degree of Master in Computer Science from the Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil, in 2004. Ms. Isaia has been an Adjunct professor at the Santo Inácio Technical School, Porto Alegre, Brazil, since 2004. From 2003 to 2004 he was a researcher of the Laboratório

de Bioinformática e Biologia Computacional at the Universidade do Vale do Rio dos Sinos. His main research interests are bioinformatics, DNA computing, and DNA programming methodologies.

Rejane Apolo Ferreira was born in São Leopoldo, Brazil, in 1974. She is an undergraduate student in computer science at the Universidade do Vale do Rio dos Sinos. From 2002 to 2004 she was a research assistant at the Laboratório de Bioinformática e Biologia Computacional at the Universidade do Vale do Rio dos Sinos.

Cláudia Kuplich Barcellos was born in Porto Alegre, Brazil, in 1967. She received the Ph.D. degree in Neurobiology from the University of Newcastle Upon Tyne, Great Britain, in 1998. Dr. Barcellos has been assistant professor at Universidade do Vale do Rio dos Sinos, Sao Leopoldo, Brazil, since 2001. Since 2002 she has collaborated with the Laboratório de Bioinformática e Biologia Computacional. Her main research interests are bioinformatics, computational biology, and biochemistry.

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 © 2007 Elsevier B.V. All rights reserved.
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

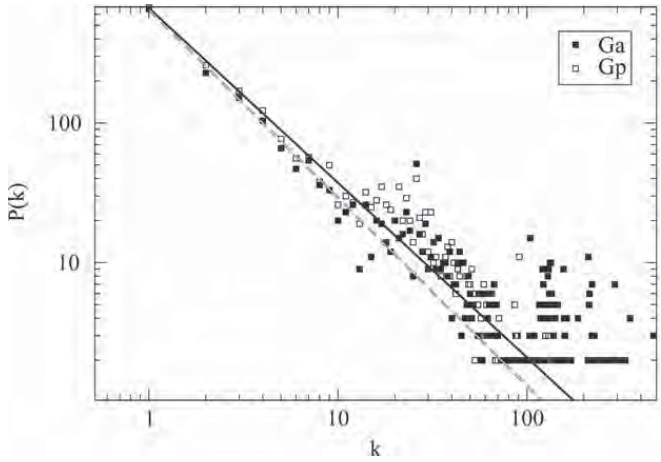


Fig. 1. Histogram of the degree distribution for G_a and G_p used in this work. Both G_a (solid line) and G_p (dashed line) are well-described by a power law function $P(k) = Ak^{-\gamma}$ that characterizes them as scale-free networks.

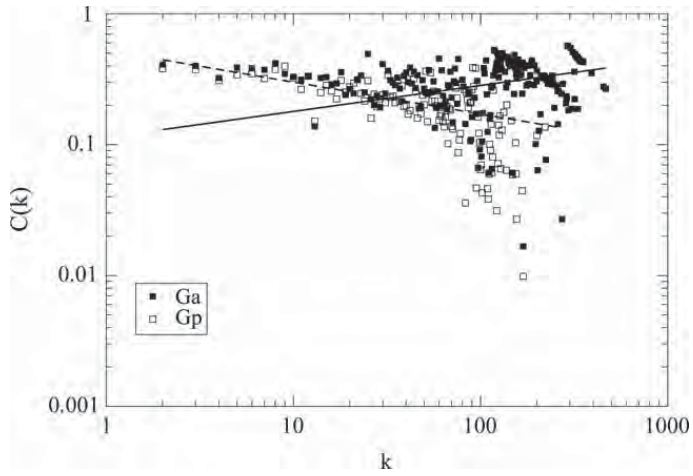


Fig. 2. The dependence of the average clusterization coefficient C on the connectivity k. The best-fit regression line for G_a (solid line) has a regression slope of -0.03 with a confidence interval of $[-0.08, 0.01]$, while the best-fit regression line for G_p (dashed line) has a regression slope of 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*.

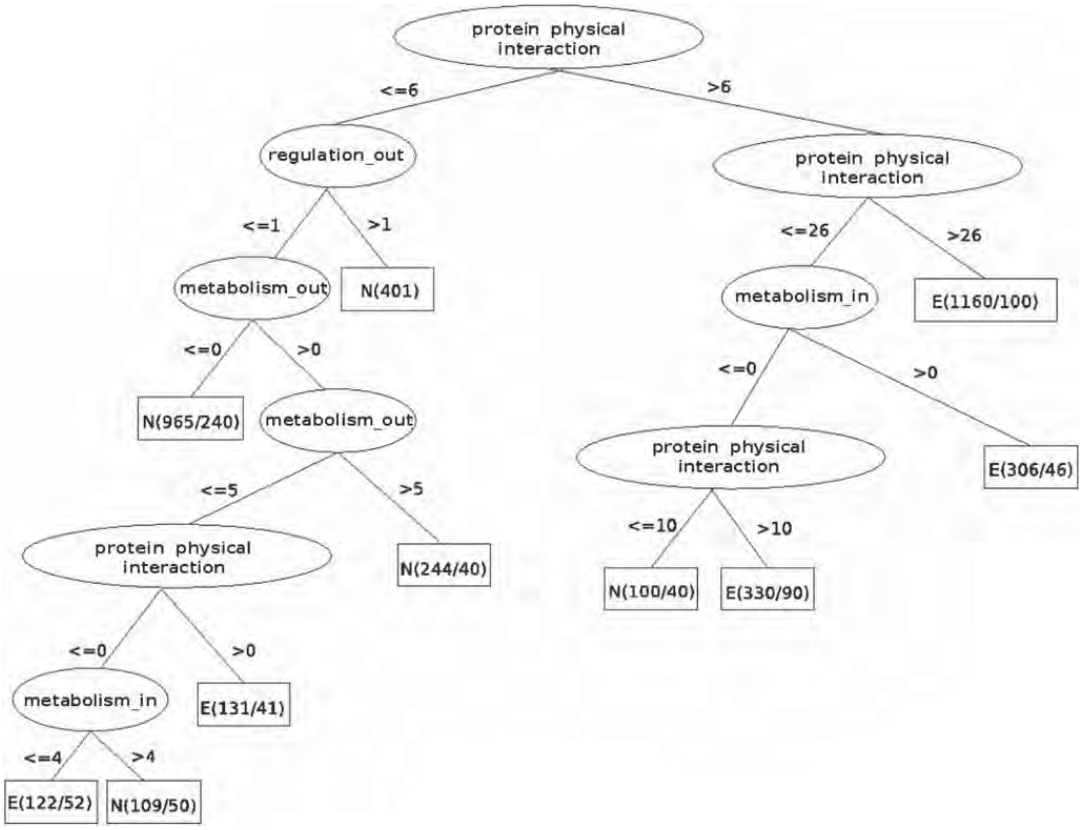


Fig. 3. Decision tree generated from t_1 with an *F-measure* of 83.4% for essential genes (E) and 79.7% for non-essential genes (N). The (x/y) inside rectangles denotes the number of correctly classified genes (x) and the number of incorrectly classified genes (y).

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.

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

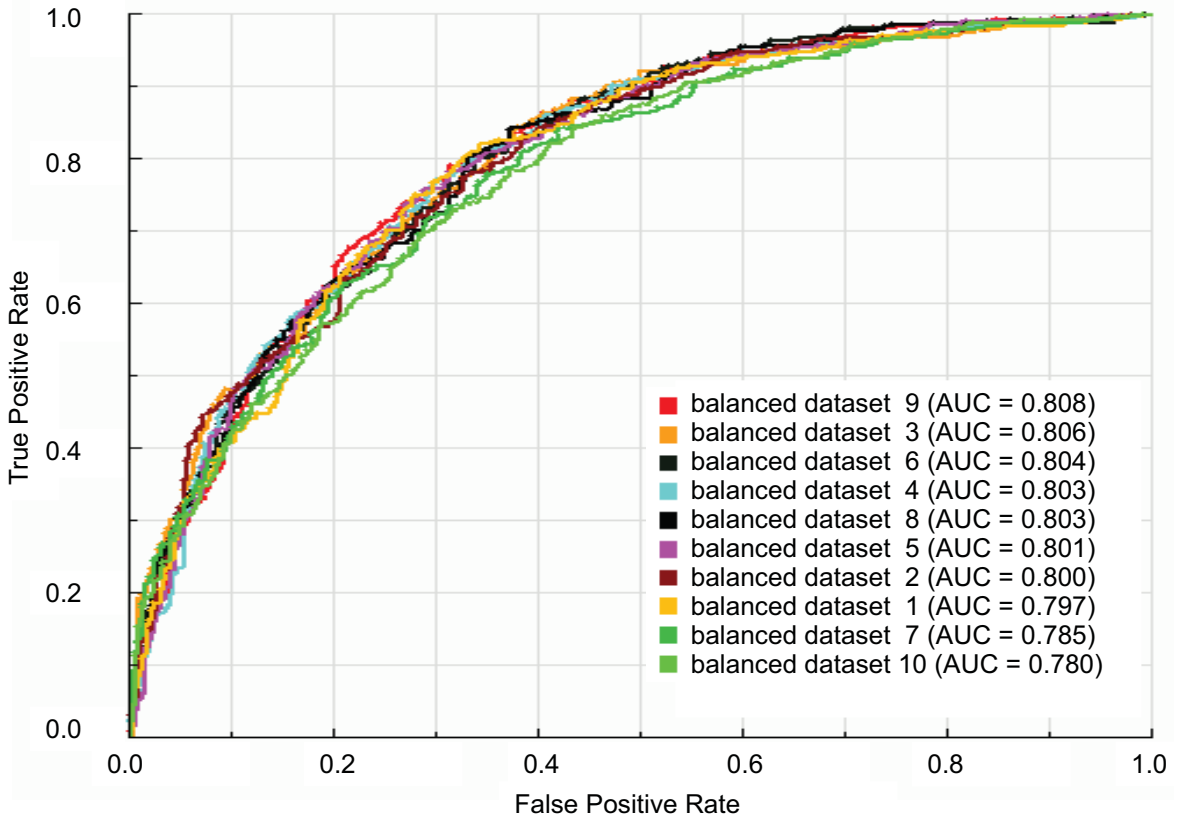


Figure 1
ROC curves and AUC values for classifiers trained on the ten balanced datasets with all available learning attributes. Balanced datasets 1-10: datasets with all available learning attributes prepared by an undersampling scheme as described in "Methods".

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-

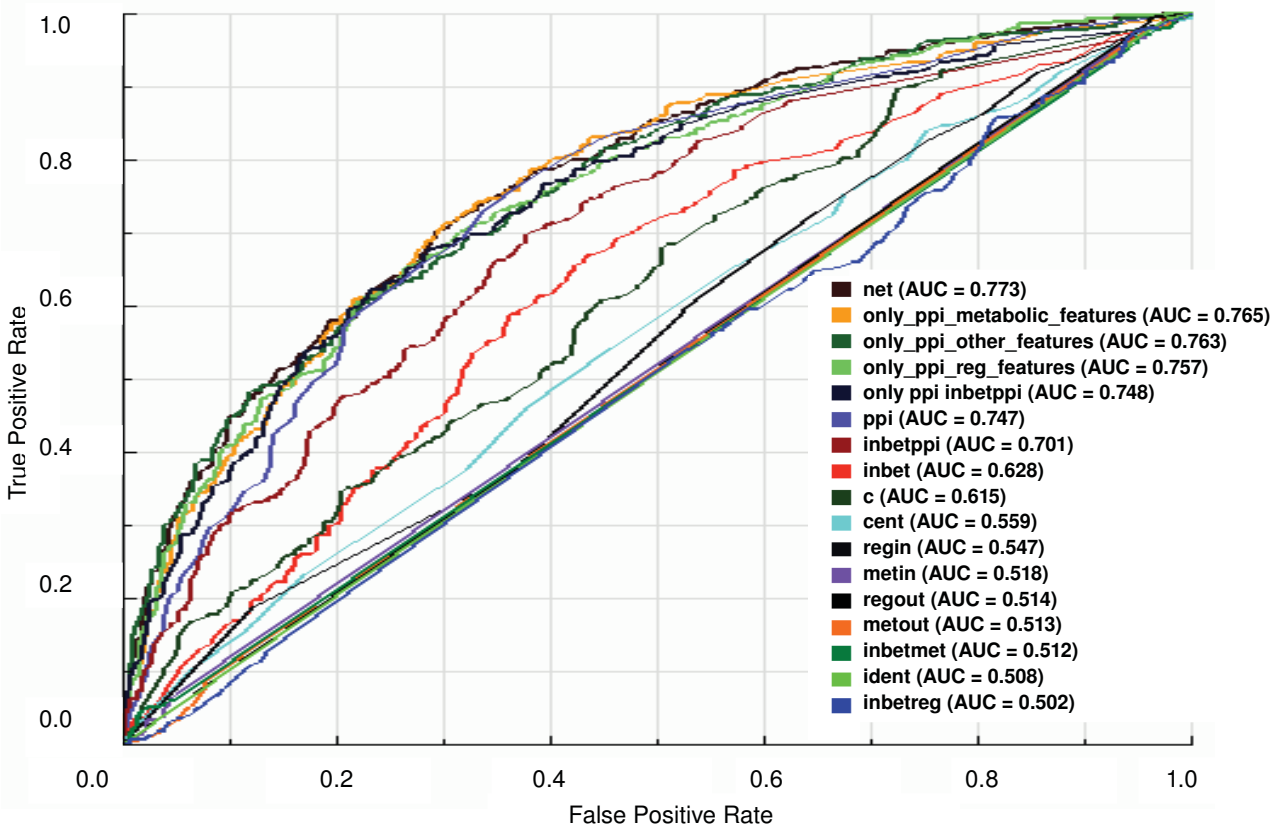


Figure 2
ROC curves and AUC values for the classifiers trained on balanced datasets with individual or grouped network topological features. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with one or groups of network topological features as learning attributes 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-

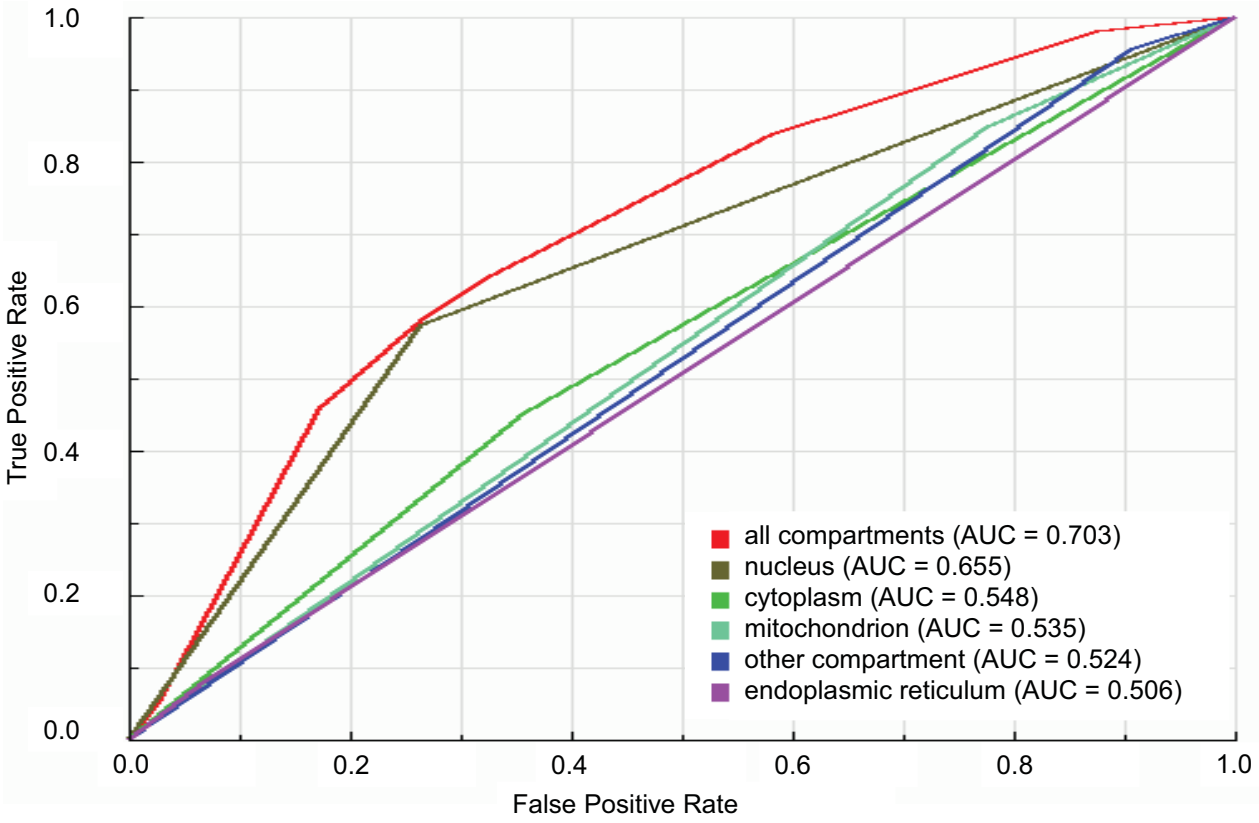


Figure 3
ROC curves and AUC values for the classifiers trained on balanced datasets with individual or grouped cellular compartments. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with one or all cellular compartments as learning attributes. "all compartments" 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

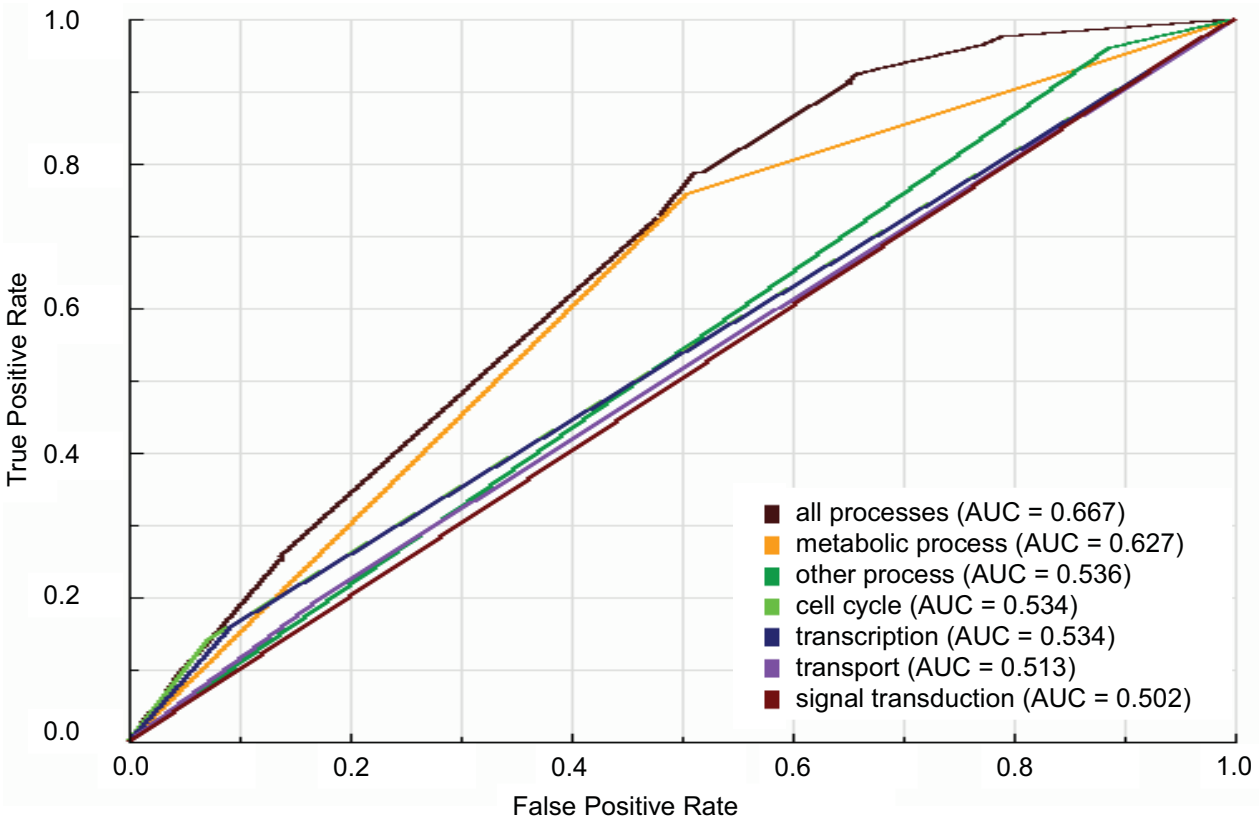


Figure 4
ROC curves and AUC values for the classifiers trained on balanced datasets with individual or grouped biological processes. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with one or all biological processes as learning attributes. "all processes" is the 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.

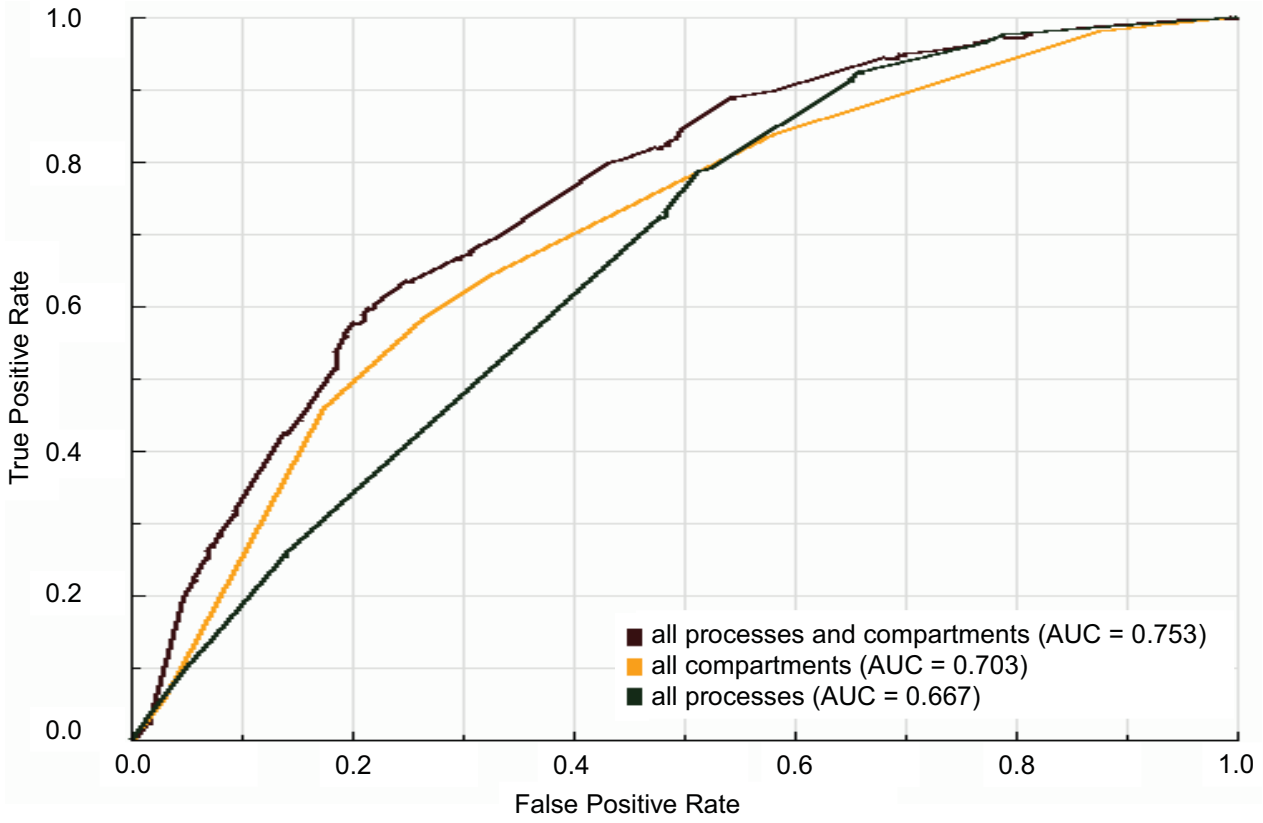


Figure 5
ROC curves and AUC values for the integrated predictors with cellular localization and biological process information. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with all biological processes ("all processes"), all cellular compartments ("all compartments")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

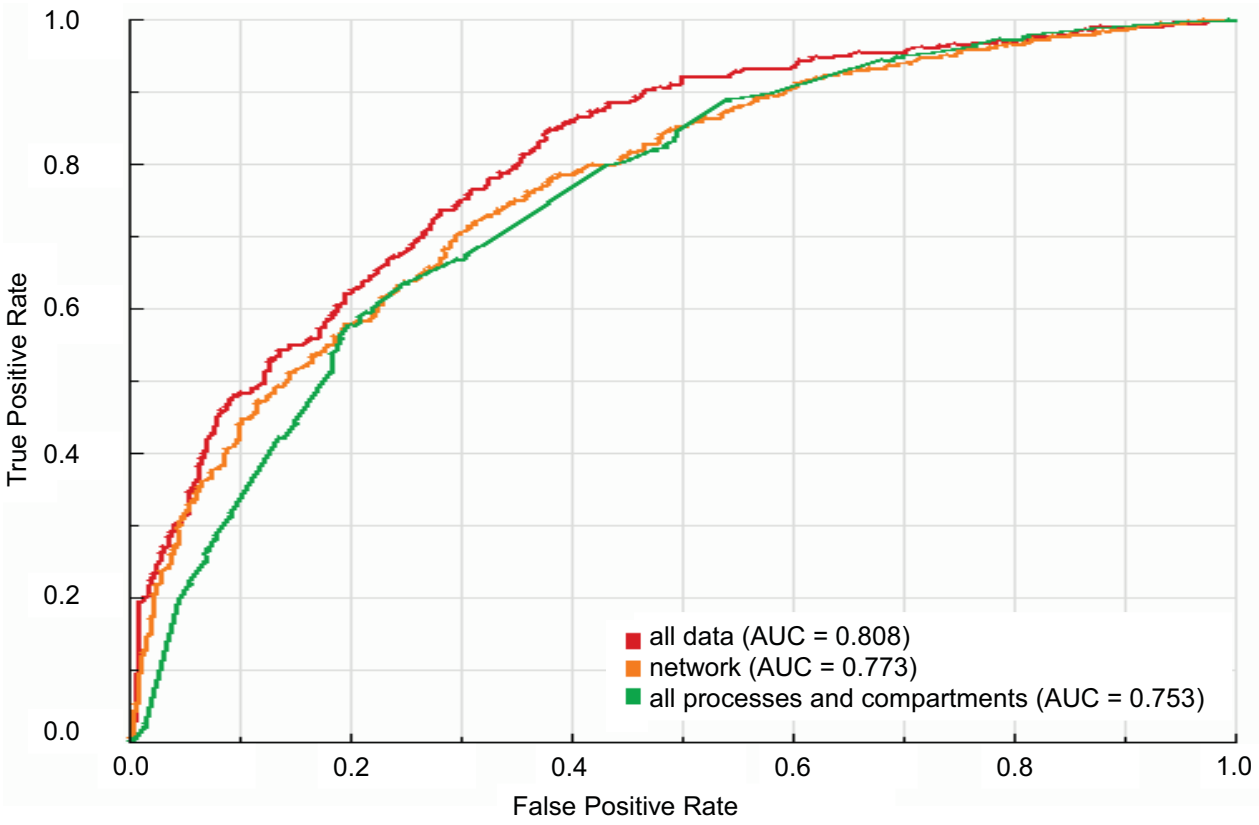


Figure 6
ROC curves and AUC values for the integrated predictors with available data. ROC curves and AUC values of classifiers trained on balanced dataset 9 (see Figure 1) with all network topological features, cellular compartments and biological processes ("all data"), all biological processes and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

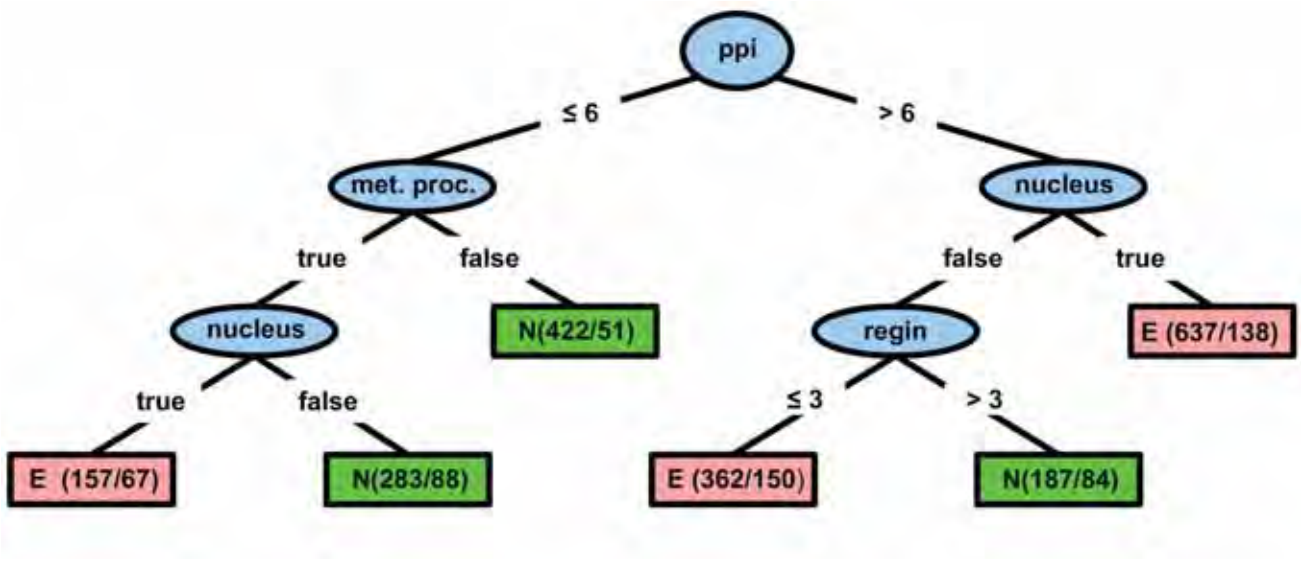


Figure 7
Decision tree generated by training the J48 algorithm on the balanced dataset 8 with all available data. This decision tree was generated by training the J48 algorithm on the balanced dataset 8 with all available data (see "Methods"). The uppermost ellipse is the node root of tree that represents 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.

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

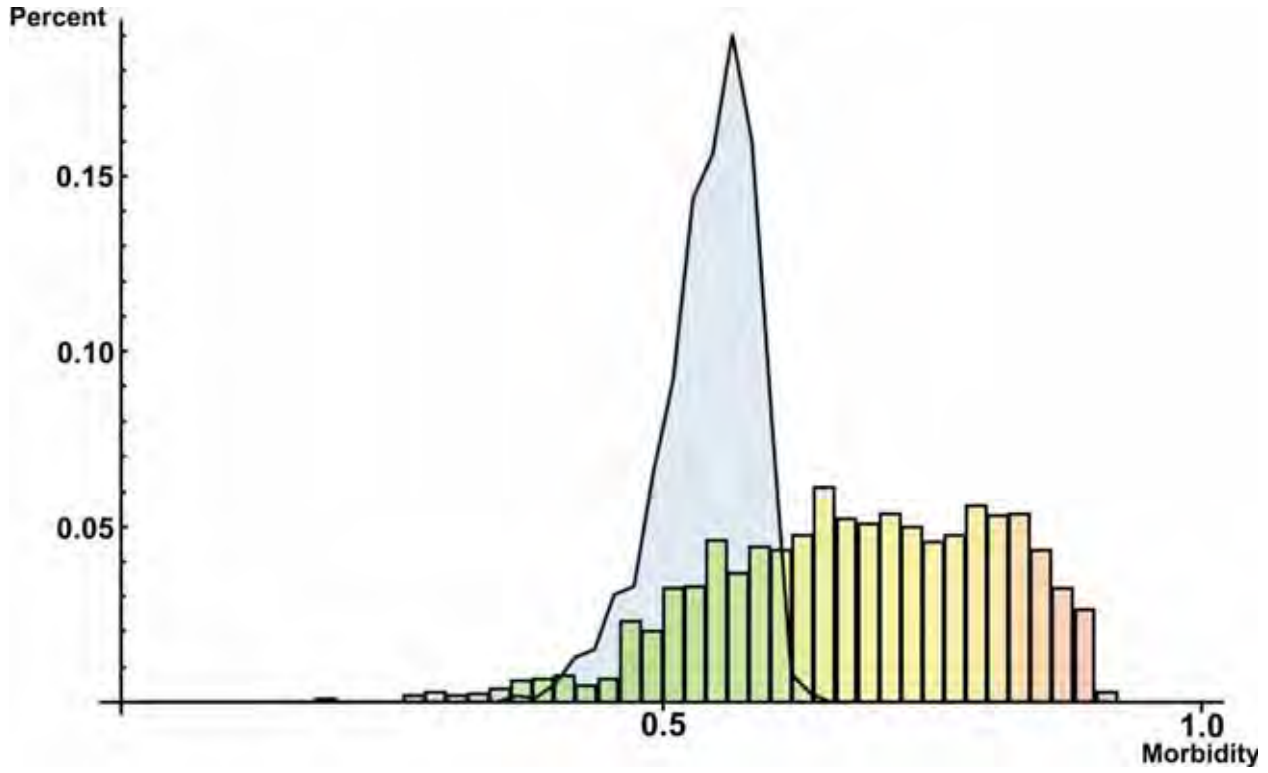


Figure 1 Frequency distribution of known morbid genes per intervals of morbidity scores Bars show the frequency distribution of known morbid genes (in percent) per 0.2 intervals of normal morbidity scores.The blue-shaded area represents the frequency distribution of known morbid genes (in percent) per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

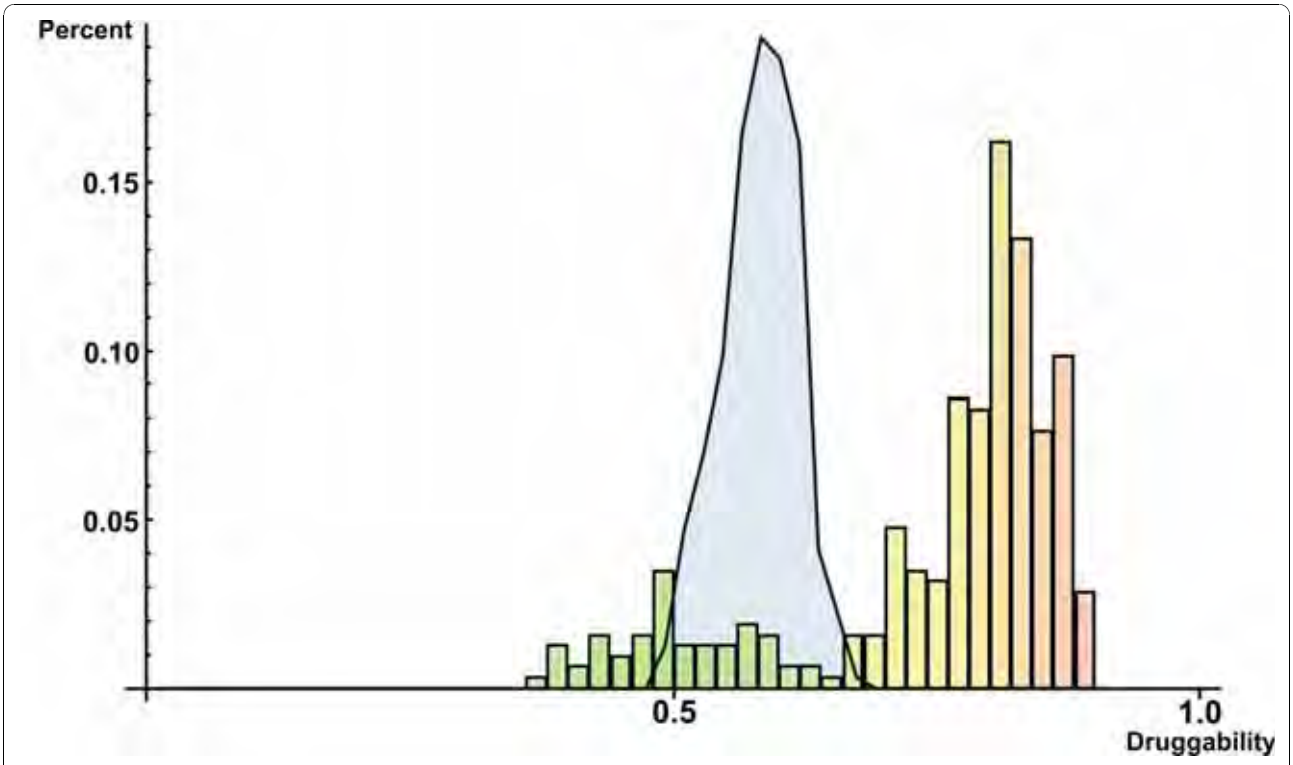


Figure 2 Frequency distribution of known druggable genes per intervals of druggability scores Bars show the frequency distribution of known druggable genes (in percent) per 0.2 intervals of normal druggability scores.The blue-shaded area represents the frequency distribution of known druggable genes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

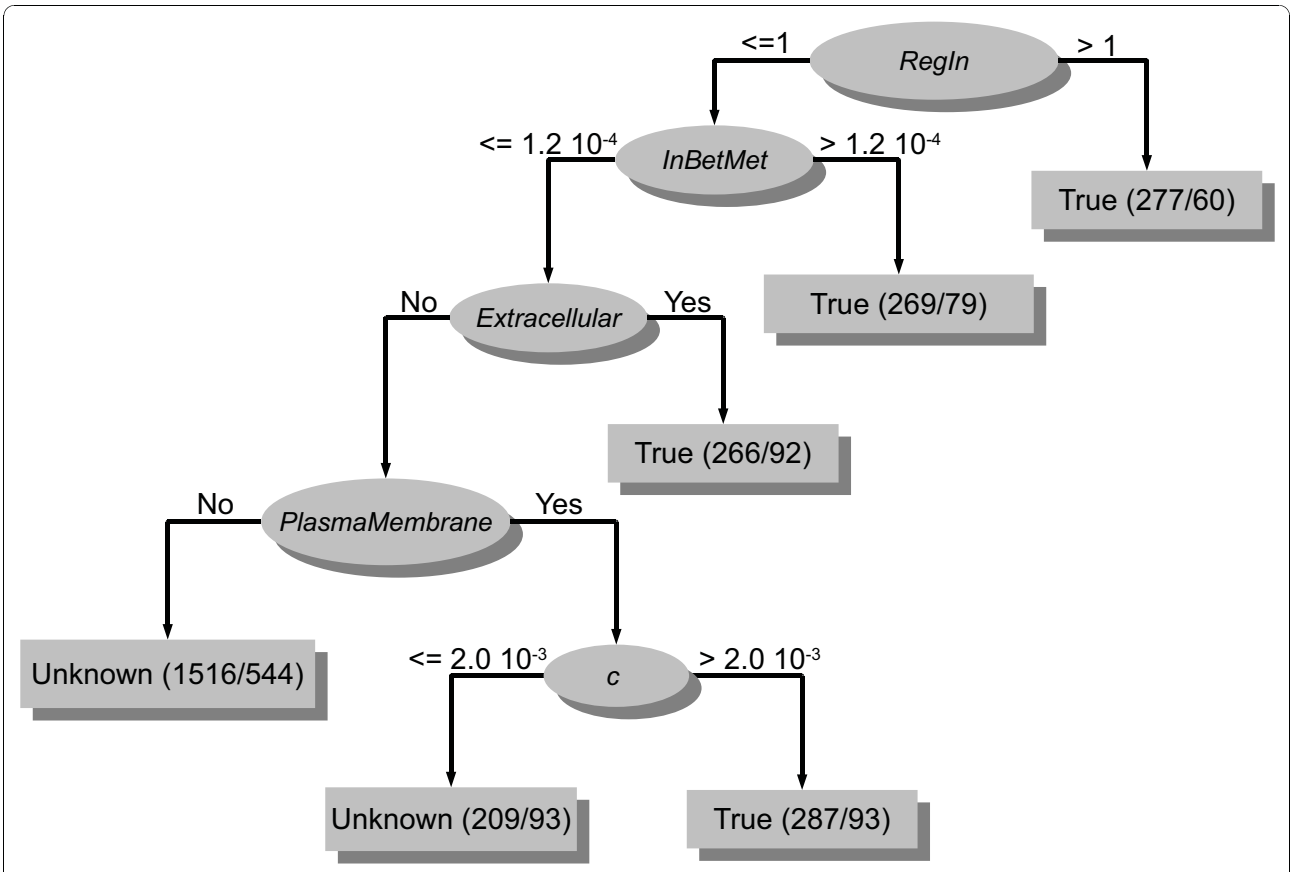


Figure 3 Decision tree generated by training the J48 algorithm on the normal morbidity datasets This decision tree was generated by training the J48 algorithm on the normal morbidity datasets (see "Methods"). The uppermost ellipse is the node root of tree that represents the most important condition 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

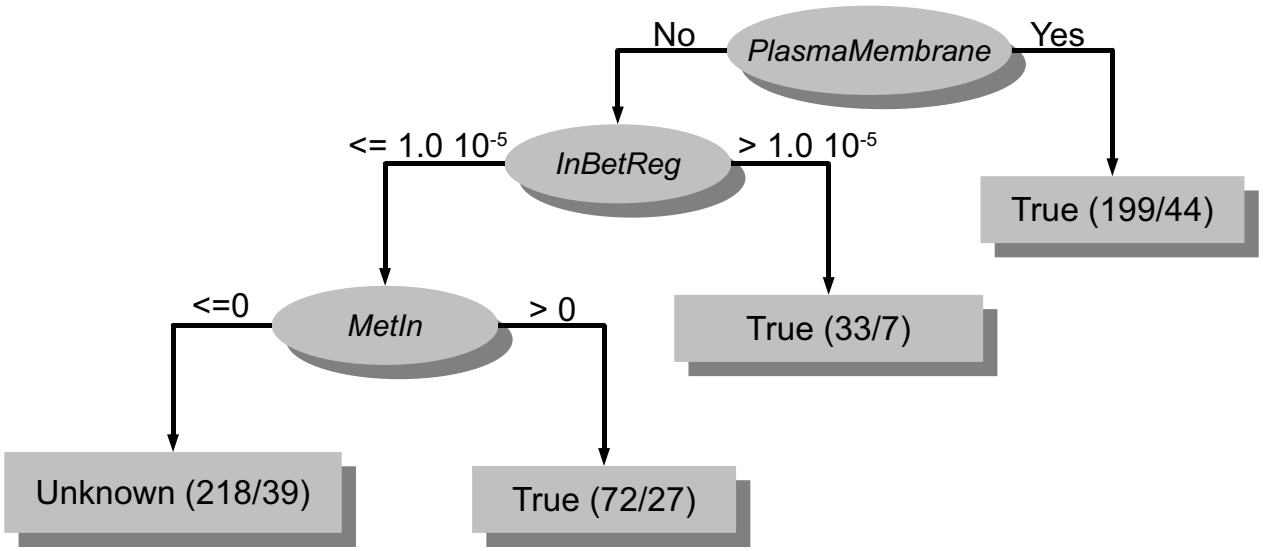


Figure 4 Decision tree generated by training the J48 algorithm on the normal druggability datasets This decision tree was generated by training the J48 algorithm on the normal druggability datasets (see "Methods"). The uppermost ellipse is the node root of tree that represents the most important condition 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.