FACULDADE DE FILOSOFIA CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
UNIVERSIDADE DE SÃO PAULO

Daniane Silva de Paula

Método para Auxiliar a Interpretação de Clusters de Expressão Gênica considerando Sumarização Automática

MESTRADO EM BIOINFORMÁTICA

RIBEIRÃO PRETO
2012

Daniane Silva de Paula

Método para Auxiliar a Interpretação de Clusters de Expressão Gênica considerando Sumarização Automática

Dissertação apresentada à Banca Examinadora para obtenção do título de mestre em Bioinformática pela Faculdade de Filosofia Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, sob a orientação da Prof. Dra. Alessandra Alaniz Macedo.

Ribeirão Preto, Maio de 2012

Resumo

A evolução da tecnologia permitiu o desenvolvimento de novas técnicas e métodos para analisar DNA, RNA e proteínas. Assim, houve um crescimento da quantidade de dados biomoleculares disponíveis, por exemplo, em um único experimento de microarray milhares de genes podem ser monitorados de uma só vez. Os genes são responsáveis pela produção de proteínas e são elas que movem a maquinaria celular. Portanto, estudar o comportamento dos genes é imprescindível para entender os processos celulares. Microarrays de DNA são uma técnica poderosa para obter dados de expressão, pois permitem que todos os genes sejam monitorados em um momento celular como, por exemplo, divisão, exposição a hormônios ou fármacos, etc. Para gerar informação útil a partir de dados biomoleculares, precisa-se de técnicas eficientes de análises de dados. Clusterização é muito difundida na análise de dados obtidos em experimentos de microarrays, pois permite agrupar genes com padrões de expressão similares. Porém, os cluster de genes obtidos precisam ser analisados dentro de um contexto, implicando em consultas a literatura. O volume de publicações científicas na literatura biomédica tem crescido também em consequência do crescimento do volume de dados. Assim, uma referência cruzada pode ser estabelecida entre os clusters gênicos (dados biomoleculares) e o conhecimento previamente publicado em artigos. No entanto, não é simples encontrar e relacionar informações de interesse na literatura, sem gastar quantidades inviáveis de tempo. Tecnologias de software podem colaborar nesse contexto, a partir da disponibilidade de muitas publicações surgiu o interesse em automatizar o processo de sumarização de textos. Nesta monografia, propõe-se o método SARI (**S**umarização Automática de **A**rtigos Científicos para **R**epresentar o significado de **I**nterações Gênicas), cujo objetivo é auxiliar a análise e a interpretação de clusters de expressão gênica, por meio de consultas rápidas à literatura com a sumarização automática de artigos científicos relacionados. Para realizar a sumarização, utilizou-se a nomenclatura gênica para identificar as sentenças mais relevantes nos artigos científicos. A sumarização implementada foi extrativa, em abordagens mono-documentos e multi-documentos. Os resultados a capacidade de aplicações que utilizaram o SARI em relacionar conhecimento da literatura com dados biomoleculares. Os resultados indicaram também que a qualidade e o poder de informação dos sumários são mais relevantes do que o tamanho do sumário.

Palavras-chave: sumarização automática, interação gênica, clusterização.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

As ciências biológicas sofreram uma revolução nas últimas décadas, pricipalmente, devido ao sucesso no sequenciamento do DNA completo de diversos organismos (MORGAN et al., 2004). O desenvolvimento de técnicas de análise de proteínas, DNA e RNA tem provocado o crescimento exponencial de dados biomoleculares. Assim, para a promoção avanços científicos, torna-se fundamental a transformação dos dados gerados em informação e conhecimento. Por exemplo, diversos métodos de análise de dados permitem que informações úteis sejam extraídas a partir de grande quantidades dados.

Experimentos clássicos da área de genética revelaram que todas as células de um organismo possuem o mesmo conteúdo de DNA, ou seja, a mesma informação genética (STRACHAN; READ, 1999). Apesar de possuírem exatamente o mesmo DNA, as células de um organismo complexo se diferenciam e executam funções diferentes. As células executam as diversas funções necessárias para a manutenção da vida do organismo ao expressar genes diferentes, os quais são apropriados para cada situação, tecido, etc. Os genes são segmentos de DNA, que contêm as informações para codificar as proteínas e RNAs necessários para o funcionamento da célula. O processo em que um gene sintetiza um produto, RNA ou proteína, é chamado de expressão gênica. As células possuem diversos mecanismos para regular a expressão dos genes. Os padrões de expressão gênica se alteram de acordo com o estado fisiológico da célula, assim genes são ativados ou inativados nos processos de crescimento, divisão, respostas ao ambiente (hormônios, toxinas, etc). Para monitorar a expressão gênica pode-se utilizar técnicas de microarray de DNA. Os dados provenientes de um experimento de microarray representam o nível de atividade de milhares de genes simultaneamente em um ambiente bioquímico. A possibilidade de medir como os genes se comportam em um dado momento contribuiu para o entendimento de processos celulares, tratamento e diagnóstico de doenças e desenvolvimento de drogas (KANKAR et al., 2002). Cada experimento de microarray gera uma quantidade enorme de dados. Um conjunto de dados de expressão gênica humana pode conter valores de expressão de até trinta e nove mil genes (KOSCHMIEDER et al., 2011)(MOHAN, 2004). Um dos principais objetivos da análise de dados de microarray é agrupar genes com perfil de expressão gênica similares. A clusterização é um tipo de aprendizado de máquina não-supervisionado que é bastante utilizado na análise de microarrays de DNA. Na clusterização, dados são agrupados de acordo com similaridades, contudo, métodos não-supervisionados exigem análises posteriores dos grupos gerados (MONARD; BARANAUSKAS, 2003). Assim, clusterização é método de análise de dados que pode ser utilizada para classificar genes em grupos de padrões parecidos.

Técnicas como clusterização envolvem grande quantidade de dados, os quais necessitam ser analisados dentro de um contexto, implicando em eventuais consultas a literatura, por exemplo, na internet. O advento da internet permitiu o armazenamento em massa e a distribuição rápida de todo tipo de informação como, informações e conhecimentos em artigos científicos das ciências biológicas. Consequentemente, houve aumento na quantidade de estudos científicos acessíveis, desenvolvimento de grandes repositórios e bases de dados especializadas como, bases de dados com informações de genes, proteínas, interações gene/proteína, organismos, etc (AFANTENOS et al., 2005) (MORGAN et al., 2004). Nesse enorme volume de dados e de literatura disponível, há dificuldades de pesquisar informações desejadas. Por exemplo, o desuso da nomenclatura gênica oficial é problema comum para busca de informação de genes e seus produtos na literatura científica. Há casos de (i) artigos com nomenclatura obsoleta e (ii) autores que não especificam se fazem referência ao gene ou a proteína resultante, etc (SPLENDORE, 2005). Nesse cenário, a utilização de buscas avançadas, ferramentas de relacionamento automático de informações e sumários automáticos podem se tornar interessantes, pois ao manipular muitos documentos, há dificuldade de se encontrar a informação desejada. A sumarização automática busca extrair conteúdo de uma fonte de informação e apresentar somente o assunto mais importante. Considerando o grande volume de publicações científicas, a tarefa de identificar, selecionar e analisar textos de interesse tornou-se uma tarefa difícil. Assim, utilizar sumários é um recurso interessante para consulta de informação na Internet, pois torna possível obter o conteúdo mais relevante de um texto, de forma condensada e rápida(MANI, 2001) (PARDO et al., 2002).

Este trabalho apresenta o método SARI (**S**umarização Automática de **A**rtigos Científicos para **R**epresentar o significado de **I**nterações Gênicas), o qual foi desenvolvido com objetivo de auxiliar na definição de significado a genes agrupados segundo algum critério. Esse auxílio ocorre com a definição de relações semânticas de artigos científicos da literatura online com a representação de interações gênicas em clusters. Para alcançar esse objetivo, o SARI foi proposto pela composição dos seguintes processos: (i) obtenção de dados de expressão gênica; (ii) análise dos dados; (iii) consulta a literatura científica, para estabelecer referências cruzadas com os resultados do processo (ii); e (iv) apresentação sumarizada dos resultados. A modularidade das etapas do método SARI permite que pesquisadores obtenham e analisem dados com os métodos que lhe são mais convenientes. Nesse caso, os processos (iii) e (iv) são suficientes para definir relações entre dados e literatura científica. Para aplicar o SARI com todos os seus processos, o método pode ser instanciado pelas seguintes etapas no cenário de auxílio a análise de clusters de expressão gênica: (a) entrada de dados de expressão gênica; (b) conversão de formato dos dados; (c) clusterização; (d) consultas de interações gênicas na literatura científica, segundo o processo de curação do BioGRID; (e) busca e (f) recuperação do conteúdo dos artigos no PubMed; e (g) sumarização de artigos para facilitar a visualização dos resultados.

Em relação a materiais e métodos, diferentes abordagens de sumarização automática foram investigadas para verificar suas adaptações a artigos científicos que contêm nomenclatura de genes. Uma aplicação foi no auxílio ao processo de atribuição de significado aos clusters gerados a partir

de dados de expressão gênica. Quando a literatura científica indicava relacionamento entre genes de um cluster, pôde-se inferir que o cluster não foi formado por aleatoriedade e que o algoritmo estava classificando de acordo com estudos científicos previamente publicados. Diversos trabalhos apresentam métodos de sumarização baseados na frequência de termos na sentença. Nesta dissertação, a sumarização foi guiada pela presença dos nomes dos genes nas sentenças. Para identificar artigos científicos que tratassem sobre interações gênicas, utilizou-se o BioGRID, base de dados curada com artigos contendo identificadores de pares de genes ou proteínas interagindo. Após obter os identificadores de artigos no BioGRID, título e *abstract* dos artigos foram buscados no PubMed e usados como os textos fontes da sumarização.

Esta dissertação está organizada da seguinte maneira: o Capítulo 2 apresenta os fundamentos teóricos que embasaram o desenvolvimento desta pesquisa e contém uma breve descrição de pesquisas que possuem alguma relação com o trabalho aqui apresentado; o Capítulo 3 descreve o método SARI; o Capítulo 4 apresenta os experimentos realizados e alguns dos resultados obtidos; e finalmente, o Capítulo 5 apresenta as conclusões e os trabalhos futuros.

2 *Fundamentos Teóricos*

O advento da internet, o desenvolvimento da capacidade de armazenamento e a distribuição da informação são fatores que contribuíram para a grande quantidade de informação on-line disponível. Paralelamente, o desenvolvimento da tecnologia na área de biologia, por exemplo, microarrays, permitiu medir os níveis de expressão de milhares de genes simultaneamente. Cada experimento de microarray gera uma quantidade enorme de dados, levando pesquisadores a analisarem esses dados por meio de métodos computacionais, além de armazená-los em grandes bases de dados científicas. A revolução das ciências biológicas com os genomas levaram ao crescimento da pesquisa e publicações científicas, gerando interesse em sumarização automática de textos das ciências biológicas, relacionamento de informação e métodos de análise de dados. Este capítulo apresenta fundamentos teóricos da área biológica, como expressão gênica e bases de dados biológicos e da área de computação.

2.1 Expressão Gênica

2.1.1 DNA

Em todas a células, o material hereditário está retido nas moléculas de DNA (ácido desoxirribonucléico), exceto em alguns vírus. O DNA é um longo polímero formado por uma sequência linear de nucleotídeos. Os nucleotídeos são a unidade básica de repetição da fita de DNA e possuem um grupo fosfato, um açúcar desoxirribose e uma base nitrogenada, como pode ser observado na Figura 1. As bases nitrogenadas dos nucleotídeos de DNA podem ser de quatro tipos: adenina (A), citosina (C), guanina (G) e timina (T) (ALBERTS et al., 2004) (STRACHAN; READ, 1999). Os açúcares ligam-se uns aos outros por meio do grupo fosfato, formando uma longa cadeia.

A molécula de DNA é formada por duas fitas complementares de nucleotídeos, as fitas sofrem uma torção assumindo a forma de dupla hélice. As moléculas (fitas) de DNA unem-se por fortes ligações químicas de acordo com as regras de Watson-Crick: adenina (A) pareia com timina (T) e citosina (C) com guanina (G) (STRACHAN; READ, 1999) (WATSON; CRICK, 1953). A Figura 2 mostra a estrutura do DNA em dupla hélice e como ele é formado pelas subunidades, os nucleotídeos.

A sequência de bases nitrogenadas é o código que carrega a informação genética. Define-se como genoma a série completa de informações genéticas contidas no DNA (ALBERTS et al.,

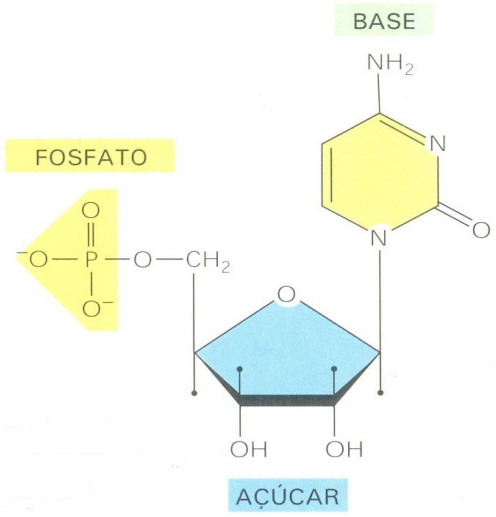


Figura 1: *Nucleotídeo (ALBERTS et al., 2004)*

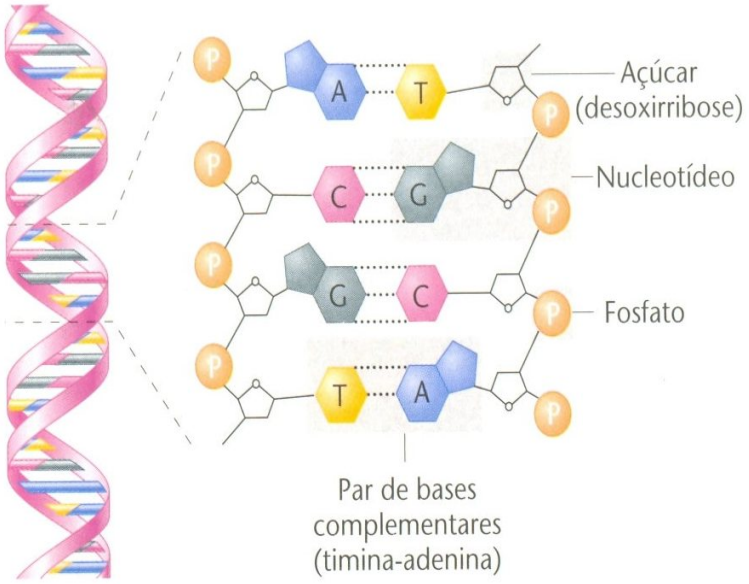


Figura 2: *Estrutura do DNA, dupla hélice (à esquerda) e nucleotídeos que compõem cada fita (à direita) (KLUG et al., 2010).*

2004). Geralmente, o fluxo da informação genética segue os seguintes passos: o DNA especifica a síntese do RNA (ácido ribonucléico) e o RNA especifica a síntese de proteínas (STRACHAN; READ, 1999). Esta via, DNA $\rightarrow$ RNA $\rightarrow$ proteína, costuma ser descrita como dogma dentral da biologia molecular, representado na Figura 3. O processo pelo qual um DNA produz um RNA é chamado transcrição e a produção de uma proteína a partir do RNA é chamada de tradução.

2.1.2 Controle da Expressão Gênica

As células acumulam diferentes conjuntos de RNA e de proteína, diferenciado-se umas das outras, porém todas células de um organismo possuem o mesmo DNA (ALBERTS et al., 2004). Em organismos complexos, como o ser humano, há uma grande quantidade de DNA não-codificador (STRACHAN; READ, 1999). Os genes são as porções codificadoras do DNA. Genes são sequências

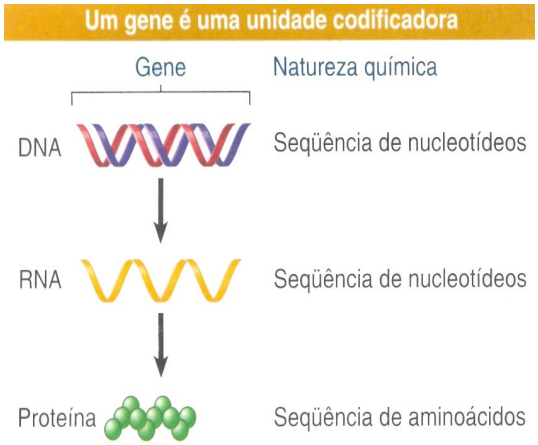


Figura 3: *Dogma central da biologia molecular: um gene codifica um RNA, que pode codificar uma proteína (LEWIN, 2009).*

de nucleotídeos da molécula de DNA, que atuam como unidade funcional para a produção de uma proteína ou de um RNA. O conjunto de ações em que a partir de um gene sintetiza-se uma proteína se chama expressão gênica. A expressão dos genes sempre é regulada, a produção de proteínas ocorre de acordo com a necessidade da célula. Não há necessidade de sintetizar todo o repertório de proteínas constantemente (ALBERTS et al., 2004).

Sinais externos podem alterar a expressão dos genes de uma célula. Por exemplo, células do fígado aumentam a produção de diversas proteínas quando são expostas ao hormônio glicocorticóide. Já as células adiposas diminuem a produção de tirosina aminotransferase, enquanto tipos celulares não apresentam alteração quando expostos ao glicocorticóide. Células se diferenciam pelos genes expressos, na Figura 4, observa-se os pontos de controle da expressão dentro da célula. Existem diversos passos que levam do DNA à proteína e em tese qualquer um desses passos pode ser regulado.

A célula pode regular a produção de proteínas ao (1) controlar quando e como um gene será transcrito, (2) controlar o processamento que um mRNA sofre, (3) selecionar quais mRNAs são levados do núcleo para determinada localidade do citoplasma, (4) selecionar mRNAs para serem traduzidos nos ribossomos, (5) desestabilizar seletivamente moléculas de mRNA, tornando-as inativas ou (6) ativar, desativar ou degradar moléculas de proteínas previamente produzidas (ALBERTS et al., 2004). Como visto, há diversos mecanismos para controlar a expressão dos genes, contudo o controle transcricional (DNA $\rightarrow$ RNA) é o principal ponto de controle da expressão gênica (LEWIN, 2009).

2.1.3 Ferramentas e Métodos da Biologia Molecular

A maioria das funções biológicas advém de interações entre muitos componentes da célula (proteínas, DNA, RNA e pequenas moléculas). Portanto, estudar isoladamente as propriedades dos componentes celulares não facilita o entendimento das funções que desempenham (HARTWELL et al., 1999). O desenvolvimento de novas técnicas e metologias científicas tem auxiliado a visu-

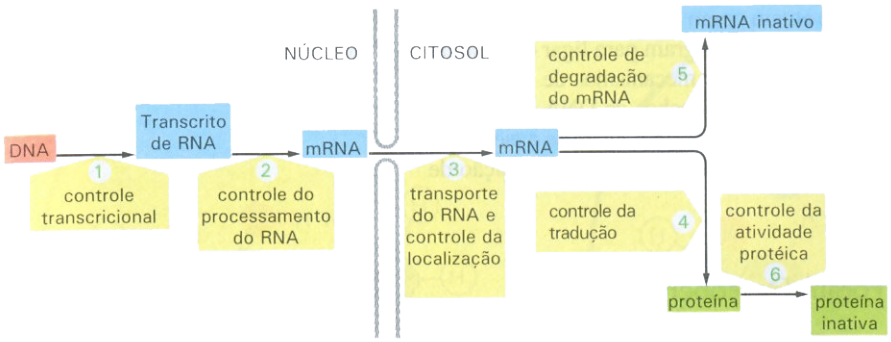


Figura 4: *Pontos de controle da expressão gênica (ALBERTS et al., 2004)*

alização de como as moléculas atuam conjuntamente na célula. É crucial entender quais genes são ativados ou silenciados, quais mRNAs estão sendo transcritos e quais proteínas estão atuando (ALBERTS et al., 2004).

A obtenção de dados biológicos pode ocorrer de diversas formas e a partir de diferentes fontes. Uma técnica frequentemente utilizada para analisar proteínas é o *Western blotting*. As proteínas são separadas umas das outras pelo tamanho (massa molecular) e colocadas em uma membrana. Esta membrana é incubada com anticorpos que identificam particularmente a proteína a ser analisada. Os anticorpos são marcados por fluorescência, radioatividade ou alguma outra substância que os torne detectáveis. O *Western blotting* pode ser utilizado para detectar a presença e o tamanho de uma ou mais proteínas (NUSSBAUM et al., 2008).

Dados úteis podem ser obtidos pela observação de quais genes estão expressando em uma situação celular. A expressão gênica comumente é monitorada por meio de ensaios hibridização de ácidos nucléicos (DNA e RNA). A hibridização de ácidos nucléicos é uma ferramenta essencial em biologia molecular. Quando estão no estado de fita simples, os ácidos nucléicos possuem a capacidade de formarem uma fita dupla, hibrizando entre si. A hibridização ocorre seguindo as regras do pareamento de bases: guanina (G) pareia com citosina (C) e adenina (A) pareia com timina (T), no DNA e (A) pareia com uracila (U), no RNA (NUSSBAUM et al., 2008) (STRACHAN; READ, 1999).

O *Southern blotting* é uma técnica de hibridização desenvolvida na década de 70 e é o método padrão para análise de fragmentos específicos de DNA. Nesta técnica, o DNA é fragmentado por enzimas de restrição e os fragmentos são separados por tamanho. Os fragmentos de DNA são desnaturados, assim o DNA torna-se uma fita simples. Estas fitas simples de DNA são incubadas com outras moléculas de DNA simples marcadas. As moléculas marcadas irão formar uma fita dupla com os fragmentos de interesse (NUSSBAUM et al., 2008). Uma variação da técnica de *Southern blotting* é o *Northern blotting*, onde o ácido nucléico utilizado é o RNA. O *Northern blotting* permite avaliar o tamanho e a quantidade de um ou um pequeno grupo mRNAs, possibilitando a obtenção de dados da expressão de genes específicos (NUSSBAUM et al., 2008)(STRACHAN; READ, 1999).

Hibridizações *Southern* e *Northern* são úteis para o estudo de um número pequeno de genes ou

mRNAs (NUSSBAUM et al., 2008). Porém, uma tecnologia mais poderosa de hibridização é a técnica de microarrays de DNA, que revolucionaram o modo de analisar a expressão gênica. Em 1995, o resultado dos estudos de Schena (95) trouxe o advento do microarray, permitindo pela primeira vez medir simultaneamente a expressão de todo ou a maior parte do genoma de um organismo (KANKAR et al., 2002). Em uma lâmina de vidro, pode-se arranjar amostras de sequências de DNA que um organismo codifica. Uma molécula de DNA irá hibridizar suas bases com - DNA ou RNA - que tenham uma sequência complementar. Desde que as moléculas da mistura estejam marcadas com fluorescência, tem-se como resultado várias manchas fluorescentes em pontos específicos do arranjo. Logo, pode-se monitorar a quantidade de mRNA (RNA mensageiro) produzida por cada gene do genoma sob condições escolhidas e observar como este padrão se altera ao modificar as condições iniciais. Técnicas de microarrays tornaram possível a produção de dados que mostram, de uma forma geral, o sistema que regula a expressão gênica (ALBERTS et al., 2004). Dados que representem a expressão total dos genes, auxiliam a identificar genes pertencentes a um mesmo processo biológico (BOLSHAKOVA; AZUAJE, 2003).

Mudanças na expressão gênica estão associadas com fenômenos biológicos importantes, como, morfogênese, envelhecimento, câncer, doenças e respostas adaptadas ao ambiente. Os perfis de expressão gênica produzidos com microarray podem ajudar na compreenssão de processos celulares, desenvolvimento de alvos terapêuticos, diagnóstico e tratamento de doenças. A ampla quantidade de dados gerados a partir do uso de microarrays ofereceu aos pesquisadores a oportunidade de utilizar, em ciências biológicas e médicas, métodos computacionais para análise de dados com possível geração de conhecimento (KANKAR et al., 2002). Para identificar os padrões similares no conjunto de dados biológicos existem diversos tipos de análise. A seguir, serão detalhados alguns métodos que permitem analisar dados biológicos.

2.2 Clusterização

Um cluster é um conjunto de objetos que são semelhantes entre si e que são diferentes dos objetos de outros clusters (HAN; KAMBER, 2006). Portanto, o processo de agrupar um conjunto de objetos em classes de objetos similares é conhecido como clusterização. Os objetos sãos descritos por um conjunto de atributos numéricos ou categóricos (JAIN; DUBES, 1988).

Agrupar de acordo com similaridades é atividade fundamental na tentativa de representar e analisar informações. Organizar dados em agrupamentos sensíveis é uma das maneiras fundamentais de discernimento (JAIN; DUBES, 1988). A clusterização tem raízes em diversas áreas, incluindo estatística, biologia, mineração de dados e aprendizado de máquina (HAN; KAMBER, 2006). O objetivo da clusterização é encontrar uma organização conveniente e válida de dados para os propósitos de pesquisa e análise em questão (JAIN; DUBES, 1988) (D'HAESELEER, 2005).

Em aprendizado de máquina[1], clusterização é um exemplo de aprendizado não-supervisionado.

[1] Aprendizado de Máquina é a área da Inteligência Artificial que trata de algoritmos capazes de aprender de forma

No aprendizado supervisionado, o algoritmo é treinado com um conjunto de exemplos cujo rótulo da classe é conhecido. Cada exemplo possui um vetor de atributos e uma classificação associada aos atributos. Os dados futuramente analisados pelo algoritmo são classificados de acordo com a experiência obtida com o conjunto de treinamento. A Tabela 1 apresenta um conjunto de exemplos que contêm atributos e uma classe associada.

Tabela 1: *Conjuntos de exemplos rotulados*

	$Atributo_1$	$Atributo_2$	...	$Atributo_m$	Classe
$Exemplo_1$	e_{11}	e_{12}	...	e_{1m}	c_A
$Exemplo_2$	e_{21}	e_{22}	...	e_{2m}	c_B
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$Exemplo_n$	e_{n1}	e_{n2}	...	e_{nm}	c_K

Já no aprendizado não-supervisionado, o algoritmo analisa os dados fornecidos e tenta agrupá-los de acordo com algum critério. Por esta razão, clusterização é uma forma de aprendizado por observação e não aprendizado por exemplo (HAN; KAMBER, 2006). Não há uma definição de classes para cada exemplo, como observa-se na Tabela 2. Portanto depois da definição dos agrupamentos, é necessário analisar qual o significado dos grupos dentro do contexto do problema (MONARD; BARANAUSKAS, 2003). O agrupamento de objetos em um mesmo cluster normalmente representa algum mecanismo do mundo real durante o processo de aquisição dos dados (MONARD; BARANAUSKAS, 2003).

Tabela 2: *Conjunto de exemplos não rotulados*

	$Atributo_1$	$Atributo_2$	...	$Atributo_m$
$Exemplo_1$	e_{11}	e_{12}	...	e_{1m}
$Exemplo_2$	e_{21}	e_{22}	...	e_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$Exemplo_n$	e_{n1}	e_{n2}	...	e_{nm}

A propriedade mais importante da clusterização é que um objeto que está em um cluster seja mais parecido com os demais objetos do mesmo cluster do que com os objetos fora do seu cluster (DUNHAM, 2002). Para medir a diferença ou dissimilaridade entre os objetos, calcula-se a distância entre o par de objetos. Em geral, a distância entre um objeto *i* e *j* é um número não negativo que é próximo de zero, quando *i* e *j* são altamente similares. Este número torna-se maior à medida que *i* e *j* diferem entre si (HAN; KAMBER, 2006) (DUNHAM, 2002). A medida de distância mais popular é a distância Euclidiana, que é definida como:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{in} - x_{jn}^2}. \tag{2.1}$$

Existem muitos algoritmos de clusterização na literatura, mas em geral eles são categorizados como hierárquicos ou particionais (DUNHAM, 2002) (HAN; KAMBER, 2006). Um método hierár-

automática (DUNHAM, 2002) (MONARD; BARANAUSKAS, 2003)

quico cria uma decomposição hierárquica do conjunto de objetos e não é necessário estabelecer o número de clusters desejado (HAN; KAMBER, 2006). Na hierarquia, cada nível tem um conjunto de clusters separado. Uma estrutura chamada dendograma pode ser usada para ilustrar a técnica hierárquica de clusterização, um exemplo de dendograma é mostrado na Figura 5 (DUNHAM, 2002). O impacto visual do dendograma é importante, pois permite analisar como os objetos se fundem ou se separam em cada nível (JAIN; DUBES, 1988).

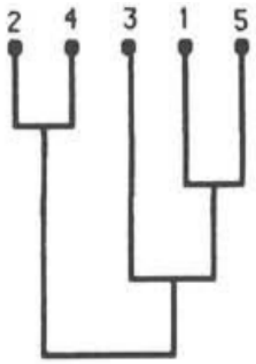


Figura 5: *Exemplo de dendograma (JAIN; DUBES, 1988)*

A clusterização hierárquica apresenta dois tipos de comportamento: aglomerativo ou divisivo. Na clusterização hierárquica aglomerativa, os clusters são criados de baixo para cima. O algoritmo inicia com cada objeto em seu próprio cluster e os funde iterativamente até que todos objetos estejam dentro de um cluster. Já a clusterização hierárquica divisiva cria clusters de cima para baixo, inicialmente todos os objetos estão no mesmo cluster e os clustes são divididos em dois até que todos os objetos estejam em seu próprio cluster (DUNHAM, 2002). Clusters hierárquicos tem a seguinte desvantagem: um passo (aglomeração ou divisão) feito não pode ser desfeito para ajustes (HAN; KAMBER, 2006).

Na clusterização aglomerativa os objetos são fundidos até que todos os objetos pertençam a um cluster. A Figura 6 apresenta um exemplo de clusterização hierárquica aglomerativa utilizando *single linkage*. Neste exemplo, existem cinco genes, aqueles que são próximos um do outro são agrupados até que todos os genes pertençam a um cluster. A distância entre os genes é calculada pelo algoritmo de *single linkage*, no qual a distância entre dois clusters é a distância entre seus pontos mais próximos (BABU, 2004) (DUNHAM, 2002) (HAN; KAMBER, 2006).

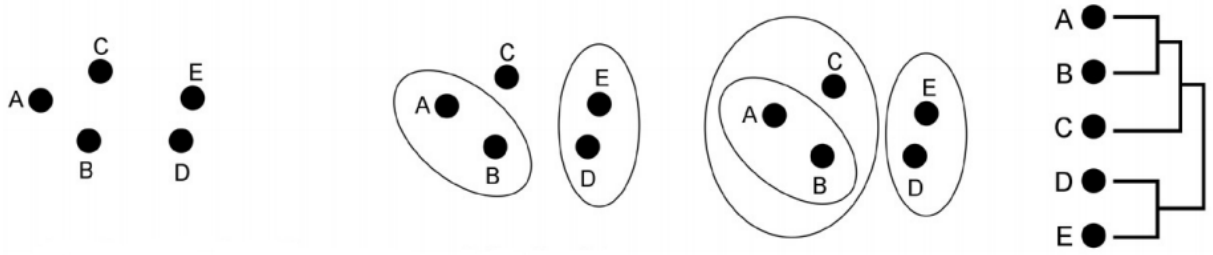


Figura 6: *Clusterização hierárquica aglomerativa utilizando single linkage (imagem adaptada de (BABU, 2004))*

Como alternativa, existem métodos de clusterização não-hierárquicos, também chamados de

particionais. A clusterização particional gera partições simples (que não se ligam como na hierárquica) em uma tentativa de recuperar grupos naturais presentes nos dados (JAIN; DUBES, 1988). Dado um conjunto de dados com n objetos, um algoritmo de clusterização particional irá formar k partições dos dados, cada partição é um cluster e $k \leq n$. Cada cluster contém pelo menos um objeto e cada objeto pertence a somente um cluster. O critério que define uma boa partição é que os objetos do mesmo cluster são próximos, ao passo que os objetos de clusters diferentes são distantes e distintos. Métodos populares de clusterização particional são o k-means e o mapas auto-organizáveis (*Self Organizing Maps* - SOM).

A Figura 7 representa o passo-a passo do k-means e de mapas auto-organizáveis. No k-means, os genes são agrupados em um número pré-definido de clusters. Calcula-se um centróide para cada cluster e os genes são rearranjados de acordo com a proximidade com os centróides. Este passo é calculado iterativamente até que ocorra a convergência ou até um número de iterações determinado. Na inicialização do mapa auto-organizável, uma grade de nós é projetada no espaço de expressão e cada gene é associado ao nó mais próximo. Um gene é escolhido aleatoriamente, o nó da grade ao qual o gene pertence é movido em direção a este gene. Os demais nós são movidos uma pequena extensão, que varia de acordo com proximidade com o gene escolhido. Realiza-se sucessivas iterações, escolhendo-se genes aleatórios, até que a convergência ou até um número fixo de iterações (BABU, 2004). Na próxima subseção, apresenta-se o k-means, pois ele foi utilizado pelo método proposto nesta dissertação.

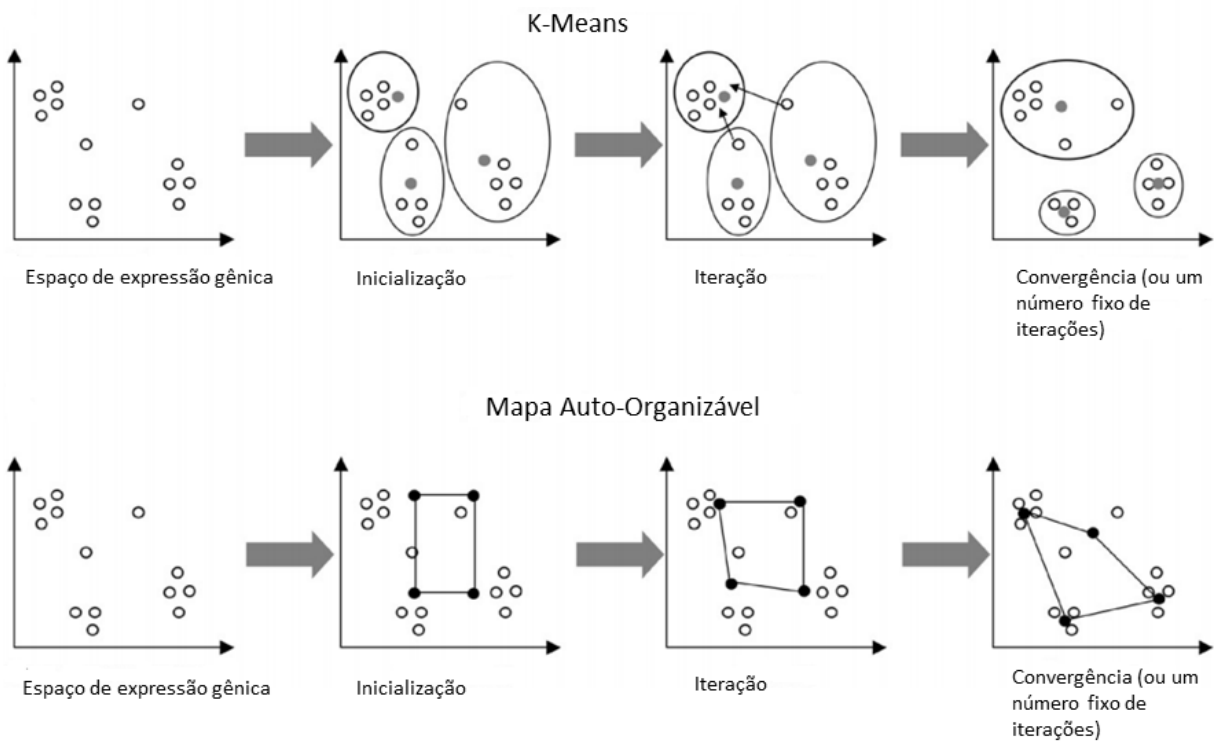


Figura 7: *Clusterização particional: k-means e mapas auto-organizáveis (imagem adaptada de (BABU, 2004))*

2.2.1 K-Means

O k-means é um algoritmo iterativo cujo os objetos são arranjados no conjunto de clusters até que o conjunto desejado seja alcançado (DUNHAM, 2002). Primeiro, o algoritmo seleciona randomicamente *k* objetos, cada objeto representa um centróide do cluster. O próximo passo é atribuir cada objeto restante ao cluster que é mais similar, baseado na distância entre o objeto e o centróide. Calcula-se um novo centróide para cada cluster (HAN; KAMBER, 2006). Este processo repete-se até que a função critério convergir. O algoritmo k-means está representado no pseudocódigo da Figura 8.

```
Algoritmo: K-means
Entrada:
     • k: o número de clusters,
     • D: conjunto de dados contendo n objetos
Saída: Um conjunto de k clusters
Método:
     (1) de forma arbitrária escolha k objetos de D como os centróides
     iniciais dos clusters;
     (2) repita
     (3)   (re)atribua cada objeto ao centróide mais próximo
     (4)    atualize o centróide de cada cluster
     (5) até estabilidade;
```

Figura 8: *O algoritmo particional k-means*

O k-means é escalável e eficiente em processar conjunto de dados grandes, pois sua complexidade é $O(tkn)$, onde t é o número de iterações, k é o número de clusters e n é o número total de objetos. No entanto, a necessidade de especificar o valor de k pode ser visto como uma desvantagem do método.

Devido ao grande número de genes e à complexidade das redes biológicas, clusterização é uma técnica útil para a análise de dados de expressão gênica (YEUNG et al., 2001). Algoritmos de clusterização são uma das ferramentas essenciais para análise de dados de expressão gênica. Eles são úteis para elucidar vários aspectos da maquinaria genética como, por exemplo, identificar a funcionalidade de genes, encontrar genes co-regulados, distinguir níveis de expressão gênica em tecidos normais e anormais (HU, 2006). A clusterização pode auxiliar na identificação de relacionamentos ocultos existentes no conjunto de genes, ou seja, clusterização não é somente para classificar em categorias distintas, mas também para descoberta de novas classes relevantes (BOLSHAKOVA; AZUAJE, 2003). Assim, pesquisadores podem identificar potenciais relacionamentos significativos entre genes. Apesar de revelar relacionamentos potenciais entre genes, por meio de técnicas de clusterização, não se pode explicar os mecanismos biológicos subjacentes (HU, 2006).

2.3 Base de Dados Biológicos

É um grande desafio seguir o atual desenvolvimento da bioinformática, genômica, proteômica e outras ”ômicas” (KLUG et al., 2010). Em decorrência, cientistas envolvidos com as ciências biológicas desenvolveram centenas de bases de dados para administrar, organizar e disponiblizar a grande quantidade de informação produzida (MORGAN et al., 2004). Alguns exemplos de tipos de bases de dados mantidas e acessíveis pela e para a comunidade científica são:

- FlyBase, um exemplo de base de organismo específico, a *Drosophila* (TWEEDIE et al., 2009);
- *Saccharomyces* Genome Database, outro exemplo de base de organismo, neste caso o *Saccharomyces cerevisiae* (CHERRY et al., 1997);
- UniProtKB/Swiss-Prot, exemplo base de dados curada de sequências de proteínas (JAIN et al., 2009);
- PubMed, exemplo de bases de dados de literatura científica (NCBI, 2011);
- BioGRID, exemplo de base de dados de interações gênicas e proteicas (STARK et al., 2005).

Nas subseções seguintes serão apresentadas duas bases de dados biológicos utilizadas na presente pesquisa.

2.3.1 BioGRID

As interações proteicas são a base da dinâmica celular, pois elas realizam o trabalho da maquinaria molecular da célula. As interações genéticas revelam relacionamentos funcionais dentro dos módulos regulatórios da expressão gênica. A soma de todas as interações proteicas e genéticas define a rede global regulatória da célula. Portanto, acessar bases de dados que contenham informações sobre interações gênicas (e seus produtos, as proteínas) é crucial no processo de análise das características e funcionalidades dos genes. Para entender as funções de cada gene, é necessário saber como e com quais outros genes ele interage. O BioGRID (*Biological General Repository for Interaction Datasets*) é um repositório on-line de acesso livre de interações genéticas e físicas. A versão 2.0 do BioGRID inclui mais de 116 mil interações dos organismos *Saccharomyces cerevisae*, *Caenorhabditis elegans*, *Drosophila melanogaster* e *Homo sapiens*. As interações são identificadas em um exaustivo processo de curação na literatura (STARK et al., 2005).

As informações do BioGRID são registradas como interações entre duas proteínas ou genes (relacionamentos binários) com uma publicação científica de referência. O termo interação refere-se ao ligamento físico entre duas proteínas, co-existência em um complexo proteico estável e interação genética. Todos os dados de interações presentes no BioGRID são disponibilizados para

download, que podem ser customizados por organismos, publicação, etc. Os dados estão disponíveis em diversos formatos, como em arquivo texto (delimitado por tabulações) e PSI MI XML[2] (HERMJAKOB et al., 2004). O conteúdo dos arquivos são:

1. identificador único do interagente A;
2. identificador único do interagente B;
3. identificador alternativo para o interagente A, como o símbolo oficial do gene;
4. identificador alternativo para o interagente B, mesma estrutura de 3;
5. aliases[3] para A;
6. aliases para B;
7. método de detecção da interação;
8. sobrenome do primeiro autor da publicação, por exemplo "'Stephenson A (2005)"';
9. identificador da publicação científica em que a interação aparece;
10. identificador de taxonomia de A;
11. identificador de taxonomia de B;
12. tipo de interação;
13. base de dados fonte;
14. identificador da interação;
15. grau de confiança.

A Figura 9 apresenta um exemplo editado de um arquivo extraído do BioGRID com informações de interações de genes humanos. Neste exemplo, pode-se ver nas colunas alguns dos itens acima enumerados.

#ID Interactor A	#ID Interactor B	Symbol Interactor A	Symbol Interactor B	Aliases Interactor A	Aliases Interactor B	Publication Identifiers	...
6416	2318	MAP2K4	FLNC	JNKK1, ...	ABPL, ...	9006895	...
84665	88	MYPN	ACTN2	MYOP	CMD1AA	11309420	...
...	...	...	...	...	...	...	...

Figura 9: *Exemplo de dados disponibilizados pelo BioGRID para genes humanos (STARK et al., 2005)*

[2]PSI MI formato de arquivo proposto para representar interações proteína-proteína.
[3]alias: outros nomes e símbolos não oficiais usados para designar um gene

2.3.2 GEO

No início da década de 2000, o *Gene Expression Omnibus* (GEO) foi criado pelo *National Center for Biotechnology Information* (NCBI) na *National Library of Medicine* (NLM). O GEO é um repositório público que armazena e distribui livremente dados de microarray entre outros dados genômicos de larga escala. Atualmente, o GEO armazena mais de 20.000 microarrays e estudos genômicos. Há vários mecanismos disponíveis que permitem busca, navegação, download e visualização de dados desde o nível de genes individuais até estudos completos. Em 2002, a revista Nature anunciou que doravante os autores cujos trabalhos utilizassem microarrays deveriam depositar os dados no GEO ou no *ArrayExpress* (PARKINSON et al., 2011), pois são repositórios públicos que qualquer pessoa pode acessar livremente e avaliar criticamente os dados apresentados nos artigos científicos. Diversas revistas acompanharam a Nature e passaram a fazer a mesma exigência. Por esta razão o GEO e o *ArrayExpress* tiveram um crescimento notável em submissão de dados e acessos (BARRETT et al., 2011)(EDGAR et al., 2002).

Os conjuntos de dados do GEO estão disponíveis no formato SOFT (*Simple Omnibus Format in Text*). Este formato foi desenvolvido para ser fácil de manipular por algoritmos e de importar para planilhas e bases de dados (EDGAR et al., 2002). A Figura 10 ilustra o formato SOFT disponibilizado pelo GEO. Em IDENTIFIER, tem-se o símbolo dos genes, as colunas - GSM455115, GSM455116, etc. - possuem os valores da expressão dos genes quando as células são submetidas aos ambientes bioquímicos de interesse, no caso da Figura 10, células tratadas com DMSO ou SAHM1.

```
^DATASET = GDS3717
#ID_REF = Platform reference identifier
#IDENTIFIER = identifier
#GSM455115 = Value for GSM455115: KOPT-K1_DMSO_01; src: KOPT-K1 cells treated with DMSO
#GSM455116 = Value for GSM455116: KOPT-K1_DMSO_02; src: KOPT-K1 cells treated with DMSO
#GSM455117 = Value for GSM455117: KOPT-K1_DMSO_03; src: KOPT-K1 cells treated with DMSO
#GSM455121 = Value for GSM455121: KOPT-K1_SAHM1_01; src: KOPT-K1 cells treated with SAHM1
#GSM455122 = Value for GSM455122: KOPT-K1_SAHM1_02; src: KOPT-K1 cells treated with SAHM1
#GSM455123 = Value for GSM455123: KOPT-K1_SAHM1_03; src: KOPT-K1 cells treated with SAHM1
#GSM455118 = Value for GSM455118: HPB-ALL_DMSO_01; src: HPB-ALL cells treated with DMSO
#GSM455119 = Value for GSM455119: HPB-ALL_DMSO_02; src: HPB-ALL cells treated with DMSO
#GSM455120 = Value for GSM455120: HPB-ALL_DMSO_03; src: HPB-ALL cells treated with DMSO
#GSM455124 = Value for GSM455124: HPB-ALL_SAHM1_01; src: HPB-ALL cells treated with SAHM1
#GSM455125 = Value for GSM455125: HPB-ALL_SAHM1_02; src: HPB-ALL cells treated with SAHM1
#GSM455126 = Value for GSM455126: HPB-ALL_SAHM1_03; src: HPB-ALL cells treated with SAHM1
!dataset_table_begin
ID_REF      IDENTIFIER  GSM455115  GSM455116  GSM455117  GSM455121  GSM455122  GSM455123  GSM455118  GSM455119  GSM455120  GSM455124  GSM455125  GSM455126
1007_s_at   DDR1        0.953      1.006      1.041      1.245      0.844      1.052      0.968      1.053      0.980      1.124      1.064      1.112
1053_at     RFC2        1.064      0.940      0.996      0.932      0.835      0.908      1.076      0.949      0.975      1.002      1.056      1.061
117_at      HSPA6       1.015      0.902      1.083      0.952      1.015      1.036      0.925      1.029      1.046      0.839      0.936      0.993
```

Figura 10: *Parte de um arquivo no formato SOFT do GEO*

O GEO disponbiliza uma ferramenta de visualização para exibir os clusters em forma gráfica

de *heat maps*[4]. As colunas representam as amostras e as linhas representam os genes. Os níveis de expressão são representadas por duas cores. Os clusters produzidos pelo GEO utilizam algoritmos de clusterização hierárquica pré-computados e K-Means com parâmetros a serem definidos pelo usuário. A Figura 11 apresenta um exemplo de visualização dada pelo GEO com o algoritmo k-means na 11(a) e de um cluster hierárquico na Figura 11(b).

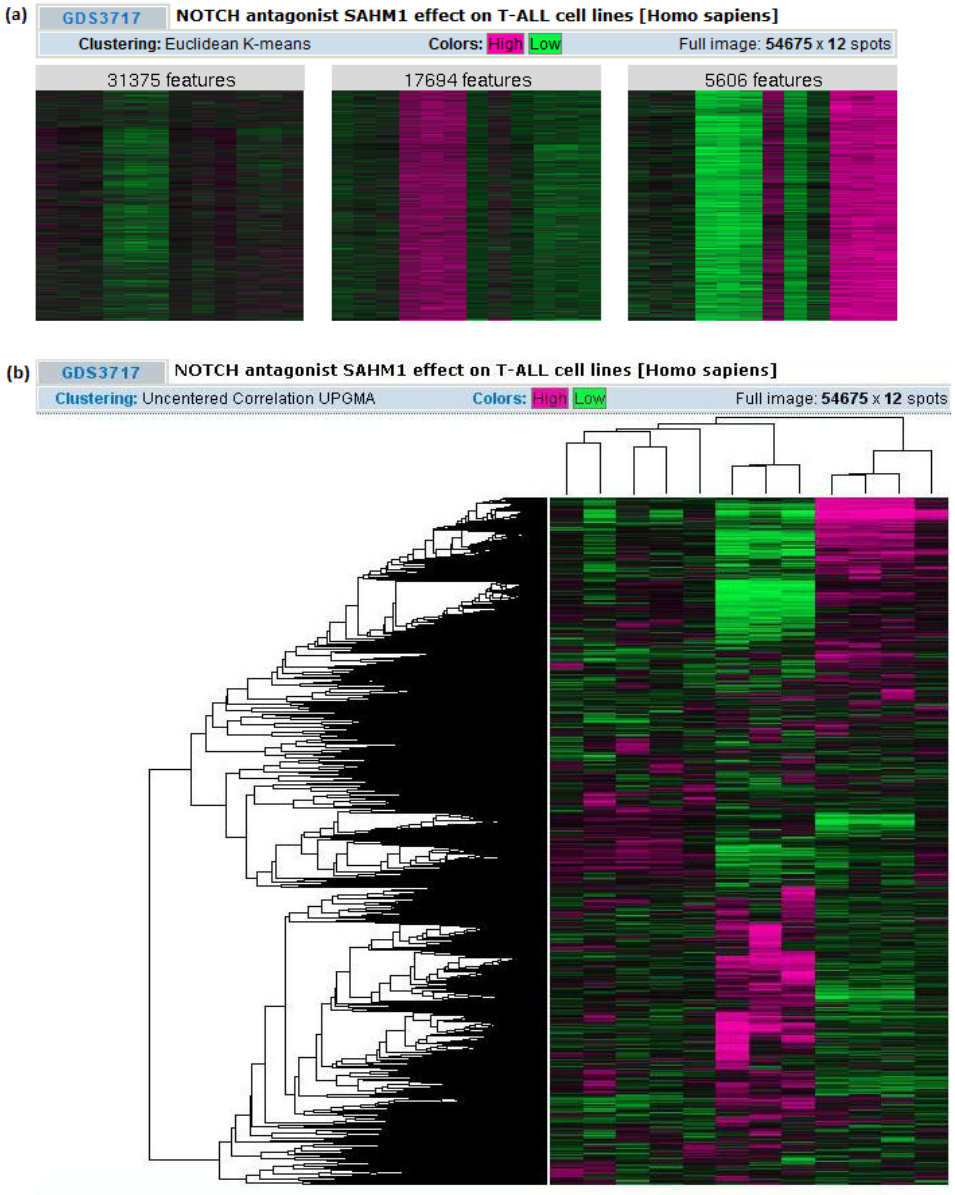


Figura 11: *Visualização de cluster disponibilizada pelo GEO*

2.3.3 PubMed

O MEDLINE é a base de dados bibliográficos do NLM, contém mais 19 milhões de referências a artigos de revistas de ciências da vida, principalmente biomedicina. No MEDLINE, todos

[4]Cluster *heat map* é uma maneira de apresentar um cluster na estrutura de uma matriz. Consiste em um conjunto de blocos, onde cada bloco recebe uma escala de cor que representa o valor correspondente ao elemento da matriz (FRIENDLY, 2009)

os registros estão indexados pelo *Medical Subjects Headings* (MeSH[5]) (NLM, 2011b). A Figura 12 apresenta a quantidade de artigos adicionados ao MEDLINE ao longo dos anos, ilustrando o grande crescimento da literatura biomédica (NLM, 2011c). O PubMed é uma interface para buscas no MEDLINE, disponibilizada pelo NLM para livre acesso do conteúdo do MEDLINE via Internet. No PubMed, também é possível acessar conteúdo não presente no MEDLINE, como, livros on-line e artigos ainda não indexados com os termos MeSH (MOTSCHALL; FALCK-YTTER, 2005). Atualmente, o PubMed contém mais 21 milhões de citações (NCBI, 2011).

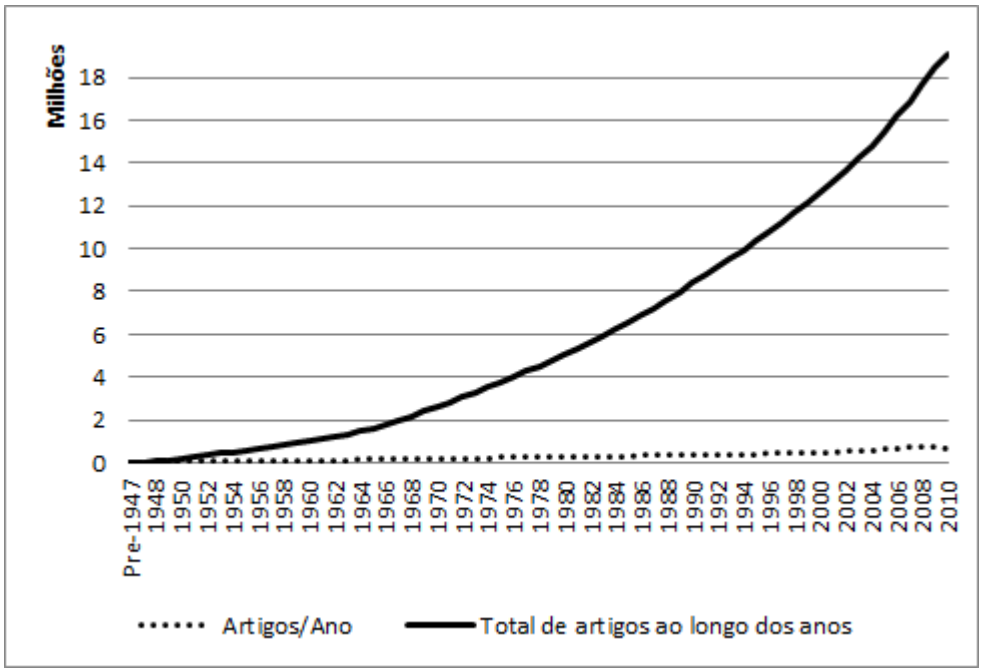


Figura 12: *Crescimento da quantidade artigos no MEDLINE*

2.3.3.1 *Entrez Programming Utilities*

O Entrez é o sistema de recuperação e busca de texto do NCBI, que integra o PubMed com trinte e nove outras bases de literatura e moleculares, incluindo sequências de DNA e proteínas, estruturas, genes, genomas, variações genéticas e expressão gênica (NCBI, 2005). Segundo Klug (2010), se uma informação genômica não pode ser encontrada no Entrez, isso significa que ela provavelmete não existe na Internet (KLUG et al., 2010). O Entrez possui o *Entrez Programming Utilities* (E-Utilities), oito programas *server-side*[6] que provêem acesso ao Entrez fora da interface regular de busca. Esses programas também chamados de *NCBI Entrez Utilities Web Services* permitem que desenvolvedores acessem o E-Utilities via SOAP (*Simple Object Access Protocol*). Os oito programas do E-Utilities são (NCBI, 2010):

- **EInfo:** recupera a quantidade de registros de cada campo da base de dados em questão, a

[5] *Medical Subject Headings* (MeSH) - é um tesauro de vocabulário controlado do NLM. Consiste em um conjunto de termos descritores organizados em estrutura hierárquica (NLM, 2011a).

[6] Programas *server-side* são aqueles que executam suas operações na entidade servidora, em um sistema distribuído que siga o modelo cliente-servidor.

data da última atualização da base e os links disponíveis da base de dados em outras bases do Entrez;

- **EGQuery:** responde a uma consulta textual com o número de registros que combinam com a consulta em cada base de dados do Entrez;
- **ESearch:** responde a uma consulta textual com a lista de identificadores de registros da base que combinam com a consulta feita, junto com a tradução dos termos da consulta;
- **EPost:** recebe uma lista de identificadores de uma base dados e responde com uma *query key* e um *web environment* para envio de conjunto de dados;
- **EFetch:** recebe uma lista de identificadores e responde com os dados dos registros correspondentes;
- **ELink:** recebe uma lista de identificadores e responde com outra lista de identificadores relacionados na base de dados consultada ou em outras bases do Entrez;
- **ESpell:** recupera sugestões ortográficas para uma consulta textual;
- **ESummary:** recebe um lista de identificadores e retorna informações do registro (nome, data de inserção na base, taxonomia, etc).

2.4 Nomenclatura Gênica

Problemas na nomenclatura de genes humanos é um obstáculo antigo para pesquisadores. É essencial utilizar termos precisos, que transmitam com clareza determinados conceitos (SPLENDORE, 2005). Reconhecendo essa necessidade, em 1979, foi apresentada uma diretriz para a nomenclatura de genes humanos em Edimburgo no HGM (*Human Genome Meeting*) (SEAL et al., 2011). No entanto, há muitos casos em que autores citam um mesmo gene utilizando nomes distintos ou não demonstram se estão se referindo ao gene ou a proteína resultante (SPLENDORE, 2005). A entidade *Human Genome Organization* (HUGO), através do subcomitê HGNC (*HUGO Gene Nomenclature Information Committee*), regulamenta a aprovação de nomes e do símbolo oficial dos genes, disponibilizando on-line e para download essas informações. Além dos nomes e símbolos oficiais, estão disponíveis vários nomes alternativos e símbolos alias, e nomes e símbolos previamente aprovados. Para cada gene humano conhecido, o HGNC aprova um nome e um símbolo. O HGNC já aprovou mais de 31.000 nomes e símbolos, sendo a maioria para genes codificantes de proteínas (SEAL et al., 2011). A Figura 13 apresenta um exemplo reformulado dos dados presentes no HGNC.

HGNC ID	Approved Symbol	Approved Name	Previous Symbol	Previous Name	Aliases	Name Aliases	Entrez Gene ID
25100	A2LD1	AIG2-like domain 1	-	-	ACF,...	-	29974
23336	A2ML1	alpha-2-macroglobulin	CPAMD9	C3 and PZP-like ...	FLJ25179	-	87769
...	...	...	...	...	...	...	...

Figura 13: *Exemplo reformulado dos dados de nomenclatura do HGNC*

2.5 Sumarização Automática

Resumir é um hábito bastante comum e facilita o processo de obter informações em grande volume. No cotidiano, sumários são elaborados e utilizados constantemente tanto na língua falada e quanto na escrita. Inconscientemente, pessoas estão sempre resumindo a história quando narram qualquer evento. Há muito interesse em automatizar o processo de sumarização, já que sumários possuem várias aplicações e são úteis em diversos contextos (MARTINS et al., 2001). A sumarização automática permite elaborar automaticamente sumários a partir de um ou mais textos-fonte. Segundo Spärk-Jones, um sumário é a redução de um texto pela seleção e generalização de suas principais informações (SPARCK-JONES, 1999). A característica crucial da sumarização é a noção de condensação da informação de um documento para o benefício do leitor (MANI, 2001). Com sumarização, pode-se reduzir a quantidade de dados textuais, desvendando dentro do texto o que é imprescindível e o que pode ser descartado (MARTINS et al., 2001) (REEVE et al., 2007).

A sumarização de informação textual pode ser utilizada, por exemplo, para (i) adaptar as informações para um formato adequado para aparelhos móveis e pequenos, como, celulares e PDAs; e (ii) ferramentas de busca, apresentando uma descrição resumida dos resultados da busca. Uma aplicação da sumarização de dados é a redução da dimensão de matrizes semânticas, reduzindo o custo computacional para processar essas matrizes.

2.5.1 Conceitos Básicos de Sumarização

Existem diversos parâmetros para nortear um sistema de sumarização e a importância deles varia de acordo com o objetivo da aplicação. Um parâmetro importante é a taxa de compressão dos sumários, que consiste no tamanho do sumário (S) sobre o tamanho do texto fonte (T):

$$TC = (tamanhoS)/(tamanhoT) \tag{2.2}$$

Em sistemas de sumarização automática, a taxa de compressão costuma assumir valores entre 5% e 30%.

Um sumário pode ser classificado em extrato ou *abstract* (III, 2004). Um extrato é um sumário composto exclusivamente de material copiado do texto fonte. Um extrato não necessariamente é

composto por sentenças do texto original, por exemplo, pode ser uma lista de termos (MANI, 2001). Já um *abstract* é um sumário que funde e reescreve as partes importantes do texto fonte, pode conter paráfrases[7], rearranjos, generalizações e especializações do texto original (ANTIQUEIRA, 2007).

Sumários também podem ser classificados de acordo com sua função: indicativos ou informativos. Um sumário indicativo provê somente um indicativo dos principais tópicos de um texto (MARCU, 2000). O objetivo de sumários indicativos é auxiliar leitores a decidirem se devem ou não ler o texto original (MANI, 2001). Em contraste, um sumário informativo contém todas as informações relevantes do texto original, permitindo que o leitor não tenha que recorrer ao texto original para obter a informação de interesse (MANI, 2001).

Pode-se distinguir sumários de acordo com a audiência, tipo de usuário. Os sumários podem ser genéricos ou focados em uma pergunta, tópico ou usuário. Sumários genéricos costumam refletir o ponto de vista do autor sobre os tópicos principais do texto. Já sumários orientados a pergunta contém os tópicos principais, que podem responder determinada questão de interesse (MANI, 2001) (MARCU, 2000).

O idioma também é um fator importante para a sumarizaçção, que pode ser monolíngue, quando utiliza-se somente um idioma no texto fonte e no sumário resultante. Sumários multilíngues são produzidos em diversos idiomas, porém o texto fonte e o sumário pertencem a mesma língua. Na sumarização *cross-lingual*, o sumário tem o idioma diferente do texto fonte. Os sumários ainda podem estar restritos a uma *sublanguage*, por exemplo, linguagem técnica, vocabulário especializado (MANI, 2001). Um sumarizador também deve considerar o gênero dos seus textos fontes, que podem ser textos científicos, notícias, editoriais, livros, etc.

O modelo básico de um sistema de sumarização automática pode ser representado em três estágios (SPARCK-JONES, 1999)(MANI, 2001):

- **Análise:** analisa o texto fonte (entrada), construindo uma representação interna deste;
- **Transformação:** transforma a representação interna do texto fonte em uma representação do sumário. Esta fase se aplica melhor a sistemas que produzem *abstracts* ou sumarização multi-documentos;
- **Síntese:** gera um sumário a partir da representação interna do sumário.

Outra distinção importante em sumarização é se o sumário foi elaborado a partir de somente um texto fonte ou de múltiplos documentos. Quando a sumarização é multi-documentos é importante encontrar os pontos em comum e as particularidades dos textos, para não produzir sumários com informações redundantes (MANI, 2001).

[7]Paráfrase: elementos textuais distintos semanticamente equivalentes. Geralmente, utiliza-se paráfrase para esclarer, tornar mais claro um texto (MANI, 2001).

2.5.1.1 Sumarização Multi-Documentos

O objetivo da sumarização é selecionar uma fonte de informação, extrair conteúdo dessa fonte e apresentar o conteúdo relevante, de forma condensada e adaptada as necessidades do usuário ou da aplicação. Na Sumarização Multi-Documentos (MDS - *Multi-Document Summarization*, a fonte de informação é um conjunto de documentos relacionados e ao extrair o conteúdo, as redundâncias devem ser removidas e as similaridades e as diferenças presentes no conteúdo devem ser levadas em consideração (MANI, 2001).

A explosão da World Wide Web proporcionou um tesouro imenso de informação, em sua maioria apresentada de forma não estruturada, em linguagem natural (MANI, 2001). Esta quantidade de informação criou uma demanda para o desenvolvimento de novos mecanismos para buscar e apresentar informação textual adequadamente (GOLDSTEIN et al., 2000). Ademais, existe muita informação repetida ou reciclada em fontes de informação distintas, por exemplo, uma descoberta científica pode ser tratada em diversos artigos da literatura. Se buscar no google ... Logo, vislumbra-se como podem ser úteis sumarizadores que identifiquem os aspectos em comum em documentos relacionados, e que também estabeleçam como os documentos se diferenciam. Apresentar as similaridades e diferenças pode ou não ser necessário, depende do tipo de aplicação. O que é realmente necessário é que estas similaridades e diferenças sejam consideradas, para que redundâncias sejam evitadas em sumários elaborados a partir de múltiplos documentos (MANI, 2001).

2.5.2 Níveis e Métodos de Sumarização Automática

A análise linguística de um texto pode ser feita em vários níveis: morfológico, sintático, semântico e discursivo. A relação entre os elementos do texto, o nível de análise linguística e a posição dos elementos pode ser vista como um gráfico multidimensional. A Figura 14 representa a idéia proposta por (BARNETT et al., 1990). No eixo vertical, estão os elementos do texto e o eixo horizontal representa a ordem de aparecimento do elemento textual no texto fonte. O terceiro eixo contém os níveis de análise linguística que vão do mais superficial (morfológico) ao mais profundo (discursivo/contextual) (MANI, 2001).

Os métodos básicos de sumarização automática podem ser tomados em termos do nível no Espaço Linguístico. Duas abordagens podem ser identificadas: superficial e profunda. A abordagem superficial não vai além do nível sintático. Sumários do tipo extrato são produzidos com abordagens superficiais, geralmente por meio de extração de sentenças (que podem ficar descontextualizadas). A principal vantagem de abordagens superficiais é a robustez. Abordagens profundas produzem *abstracts*, tipicamente gera-se texto em liguagem natural. Este tipo de abordagem está em níveis semânticos e discursivos. Métodos profundos prometem produzir sumários mais informativos, mas necessitam que seus textos pertençam a um domínio específico (MANI, 2001).

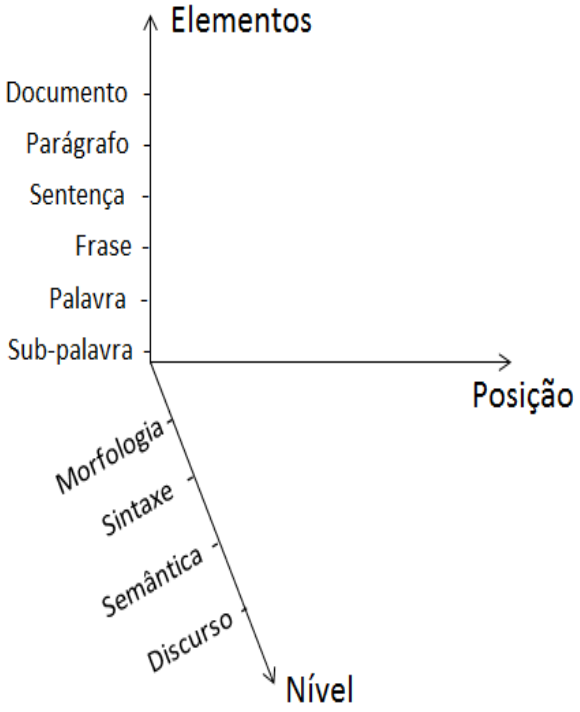


Figura 14: *Espaço linguístico [adaptado de (MANI, 2001)]*

2.5.3 Extração

Do ponto de vista computacional, produzir sumários do tipo extrato é muito mais econômico do que elaborar *abstracts*, já que não exige criar um novo texto. Por ser uma solução de baixo custo, muito mais atenção é dada a produção automática de extratos. Em abordagens extrativas domina a fase de Análise, que é superficial e raramente ultrapassam o nível do semântico. A unidade básica da extração é a sentença, existem diversos motivos para a preferência da sentença como unidade. Historicamente, sentenças têm servido como unidade em análises sintáticas e semânticas (MANI, 2001). O principal desafio de um sistema de sumarização extrativa é decidir qual conteúdo selecionar, ou seja, decidir quais sentenças do texto original são importantes para serem incluídas no sumário (NENKOVA; VANDERWENDE, 2005).

2.5.3.1 Método das palavras-chave

Enquanto desenvolve-se idéias ao longo de um texto, alguns termos importantes se repetem. Assim, pode-se classificar as sentenças pela premissa de que termos mais frequentes são mais relevantes. Se um texto trata de um tema, espera-se muitas referências ao tema tratado (MANI, 2001). Portanto, as palavras-chaves são aquelas mais frequentes, o método extrai as senteças que contenham essas palavras e agrupa as sentenças selecionadas para compor o sumário. Ou seja, pode-se ver que o método de palavras-chave consiste nos seguintes passos:

1. identificar sentenças relevantes;

2. extrair do texto original as sentenças de interesse, identificadas no passo anterior;

3. justapor as sentenças para formar o sumário.

O trabalho pioneiro de Luhn (1958) sugere encontrar os termos frequentes, que sejam substantivos, verbos, advérbios e adjetivos, ignorando a frequência de preposições, conjunções, etc (MANI, 2001)(LUHN, 1958)(MARTINS et al., 2001). Há variações deste método, como ordenar as frases de acordo com a quantidade de palavras-chave presentes e construir o sumário com as sentenças de maior pontuação. Outra variação é considerar palavras-chave somente as palavras que aparecem no título do texto (PARDO et al., 2002).

2.5.4 Avaliação

Não há na literatura uma resposta exata de como se avaliar a qualidade de um sumário (HOVY; LIN, 1996). Há vários desafios para a avaliação de sumários. No processo de sumarização, uma máquina produz uma saída em linguagem natural, o que torna difícil a noção de quão correta é esta saída. A sumarização envolve taxas de compressão, portanto a avaliação deve considerar a habilidade do sumário transmitir a informação de acordo com as taxas aplicadas. Outros aspectos que dificultam a avaliação são legibilidade de um sumário e se a informação está apresentada de maneira adequada ao usuário (MANI, 2001).

Apesar da inexatidão, é possível desenvolver algumas *guidelines* e abordagens, e a partir delas realizar avaliações de resultados de processos de sumarização. Algumas regras são óbvias, por exemplo, um sumário deve ser menor que o texto original e tratar da mesma informação que o texto original. Existem algumas medidas que auxiliam avaliar a sumarização. Uma medida bastante simples é comparar o tamanho do texto original com o sumário produzido. A métrica utilizada pode ser a quantidade letras, palavras ou sentenças (HOVY; LIN, 1996). Este tipo de medida foi utilizada no Capítulo 4, para comparar o tamanho do texto original com o tamanho do sumário produzido.

Métodos de avaliação podem ser classificados como intrínsecos ou extrínsecos. Os métodos intrínsecos avaliam o sistema de sumarização em si. Exemplos de métodos intrínsecos são a avaliação de Qualidade e de *Informativeness*. A avaliação de Qualidade tenta estimar quão legível e fluente são os sumários. Essa avaliação leva em consideração a gramática, presença de redundâncias, estrutura e coerência. A avalição de *Informativeness* analisa a quantidade de informação do texto fonte que é preservada no sumário. Os métodos extrínsecos avaliam o sistema em relação a alguma tarefa. A avaliação extrínseca determina o efeito da sumarização em alguma tarefa como, por exemplo:

- avaliar eficiência do sumário em auxiliar o usuário a executar um conjunto de instruções. Por exemplo, suponha um texto que seja um manual extremamente detalhado e que um sumário

foi produzido a partir deste texto. Deseja-se avaliar se o sumário é tão bom ou melhor que o manual original ao auxiliar o usuário a executar as instruções;

- examinar a utilidade de um sumário em relação a alguma informação necessária ou objetivo, como encontrar um documento relevante dentro de uma grande coleção;
- analisar o impacto de um sumarizador que está incorporado a outro sistema, como, o quanto a sumarização é útil em um sistema que responda perguntas.

2.6 Trabalhos Relacionados

Na literatura, foram identificadas pesquisas utilizando sumarização automática em textos de diversas áreas. O sistema LAKE (*Learning Algorithm for Keyphrase Extraction*) considera as características TF/IDF[8] e a posição do termo no documento para realizar a sumarização. Primeiramente, identificam-se as frases candidatas. Uma frase é candidata quando se enquadra dentro um padrão sintático. Alguns exemplos de padrões são: substantivo; adjetivo+substantivo; substantivo+verbo+adjetivo+substantivo. Um algoritmo de aprendizado de máquina (classificador *Naive Bayes* da Weka[9] (HALL et al., 2009)) seleciona entre as candidatas quais serão as frases-chave do documento (D'AVANZO et al., 2004). Em vez de usar a própria frase, utiliza-se o *head* da frase candidata. De acordo com o princípio de *headedness*, qualquer frase possui uma única palavra (*head*) que determina as propriedades da frase, o *head* pode ser um verbo ou um substantivo. Para escolher as frases-chave, o classificador da Weka também utilizou TF/IDF (o produto entre a frequência do *head* em um documento e a frequência inversa do *head* em todos documentos) e, a Primeira Ocorrência (a distância da frase candidata do começo do documento). Os textos utilizados na experimentação foram da coleção disponbilizada pelo DUC-2003 (*Document Understand Conference*), o idioma foi o inglês e a maioria dos documentos eram de origem jornalística. Como resultado, o classificador deu prioridade as frases candidatas cujo *head* maximizava seu TF/IDF e tendia a ocorrer no começo do documento. Os autores concluiram que utilizar padrões sintáticos contendo formas verbais introduziu ruído e que a relevância de uma frase-chave não pôde ser definida exclusivamente pelo ponto de vista sintático. Apesar das dificuldades para escolher as frases candidatas, os autores afirmaram que as características utilizadas pelo classificador para encontrar as frases-chaves foram efetivas (D'AVANZO et al., 2004).

Outro trabalho relacionado é o SUMMARIST. Este sistema apresenta um plano de módulos para identificar e interpretar os tópicos centrais do texto fonte e para gerar sumários. No processo de identificação, filtra-se do texto fonte somente os tópicos centrais. Na fase de interpretação, compacta-se os tópicos extraídos em tópicos mais sucintos, por exemplo, pera e maçã, são

[8] A medida TF/IDF assume que a importância de um termo é proporcional a frequência do termo em um documento e inversamente proporcional ao número total de documentos de uma coleção em que termo ocorre(JONES, 1972)

[9] Weka é uma coleção de algoritmos de aprendizado de máquina para mineração de dados. A Weka contém ferramentas para pré-processamento de dados e algoritmos de classificação, regressão, clusterização, regras de associação, além de uma interface para visualização

generalizadas para o conceito fruta. Essa generalização de conceitos é realizada pelo WordNet (FELLBAUM, 1998). Primeiro, identifica-se o sinônimo no WordNet para cada palavra considerada central do texto, e em seguida localiza-se uma generalização apropriada para o conceito (HOVY; LIN, 1996) (FELLBAUM, 1998). O WordNet é uma grande base de dados lexical da língua inglesa. No WordNet, pronomes, verbos, adjetivos e advérbios são agrupados em conjuntos de sinônimos cognitivos, cada um expressando um conceito distinto. Esses conjuntos estão interligados por significados semânticos e relações lexicais. Os textos utilizados para testar o sumarizador foram dois artigos jornalísticos em língua inglesa, um sobre roubo de arte e outro sobre urbanização. Uma das medidas usadas na avaliação mostrou que (i) texto completo, (ii) *abstract* feito por humano e (iii) extrato feito pelo SUMMARIST foram igualmente bons em fornecer o conteúdo da informação. Uma outra medida utilizada mostrou que o extrato do SUMMARIST foi 30 porcento melhor que os *abstracts* feitos por humano e que seleção randômica de sentenças é tão boa quanto os *abstracts*. Os autores consideraram este últimos resultados obscuros e concluíram que será necessário de um estudo aprofundado para determinar a validade dos resultados (FELLBAUM, 1998).

Nenkova e Vanderwende discutem o impacto da frequência de termos no processo de sumarização e o papel da frequência na criação de um sistema de sumarização. Para criar sumários, o sistema SumBasic explora exclusivamente a frequência de termos e de unidades de conteúdo no texto fonte e a probabilidade destes aparecerem em um sumário feito por um humano. A coleção de textos é em inglês e são do DUC-2003 e 2004. O trabalho conclui que a frequência no texto fonte é um forte indicativo se uma palavra seria usada em um sumário feito por um humano. No entanto, a frequência não explica completamente as escolhas de um humano e muitas palavras com baixa frequência no texto fonte aparecem nos sumários (NENKOVA; VANDERWENDE, 2005).

O MedMeSH Summarizer foi desenvolvido com o objetivo de auxiliar pesquisadores a estabelecer referências cruzadas experimentais e analíticas dos resultados obtidos de experimentos com microarrays (KANKAR et al., 2002). O MedMeSH Summarizer sumariza informação de um grupo de genes pela filtragem da literatura biomédica e atribui palavras-chave que descrevem a funcionalidade desse grupo de genes. O sistema requer do usuário uma lista de genes como entrada e o PubMed é utilizado como base de dados. Outras bases de dados de organismos específicos também são utilizadas para obter os sinônimos (alias) dos nomes dos genes. O MedMeSH Summarizer não utiliza título, *abstract* ou texto completo. O texto fonte do MedMeSH Summarizer são os termos MeSH usado para indexar artigos do PubMed associados a um artigo. Logo, o resultado é uma lista ordenada por frequência de termos Mesh para a lista de genes que o usuário inseriu na entrada (KANKAR et al., 2002).

É interessante para pesquisadores fazer comparação de novos resultados com fatos biológicos previamente conhecidos, teorias bem estabelecidas e resultados anteriores. Bases de dados com literatura biológica e médica fornecem um grande depósito de conhecimento para estas comparações. No entanto, o tamanho dessas bases de dados torna a tarefa de fazer referências cruzadas muito lenta, tediosa e desencorajadora. Pôde-se observar que maior parte dos trabalhos relacio-

nados explora a frequência dos termos para identificar o assunto principal do texto fonte e assim construir o sumário. O presente trabalho difere em alguns pontos, pois foram eleitas como termos-chaves a nomenclatura dos genes (nomes e símbolos oficiais ou não). No contexto deste trabalho, o fundamental não é identificar o conteúdo principal do texto fonte, mas buscar dentro do texto informações sobre interações gênicas.

3 *Método SARI e Aplicações*

Neste capítulo, são apresentados os materiais e os métodos utilizados para desenvolver a proposta de auxiliar a interpretação de interações entre genes e entre produtos gênicos pela atribuição de significado a grupos de genes. Esta interpretação é apoiada por meio de consultas a literatura, que é sumarizada devido a sua grande extensão, visando o benefício do leitor. Enfim, neste capítulo, descreve-se genericamente o método proposto, o qual foi nomeado de SARI (**S**umarização Automática de **A**rtigos Científicos para **R**epresentar o significado de **I**nterações Gênicas). Apresenta-se também um cenário específico de aplicação do SARI, descrevendo desde a escolha de um conjunto de dados de expressão gênica até a sumarização dos textos que tentam caracterizar os clusters formados. Ilustra-se também uma forma genérica de utilizar a sumarização no contexto biológico, desconsiderando o método utilizado para a obtenção do grupo de genes a ser manipulado pelo SARI.

3.1 Método SARI

O método SARI foi desenvolvido para auxiliar a interpretação da interação entre genes e da interação entre produtos de genes. O SARI, ilustrado na Figura 15, possui quatro passos principais: (1a) obtenção de dados gênicos, que são submetidos a processos de análise de dados ou (1b) inserção de um conjunto de genes obtidos externamente; (2) consultas a literatura científica, que reforcem ou contradigam os resultados da análise de dados; (3) sumarização automática dos textos consultados e (4) apresentação ao usuário a literatura de interesse de forma condensada.

No método SARI, o primeiro passo para analisar e interpretar possíveis interações gênicas e proteicas é obter os dados biológicos. Pesquisadores podem obter dados biológicos em experimentos feitos em seus próprios laboratórios ou em laboratórios parceiros. Nesses experimentos, um grande volume de dados biológicos é gerado, os quais podem ser armazenados em grandes bases de dados online, mantidas pela própria comunidade científica. Por exemplo, existem dados de expressão gênica disponíveis em bases de dados online como o GEO ou o *ArrayExpress*.

Após obter os dados biológicos, o método SARI encontra-se "rico em dados, mas pobre em informação". Assim, a análise de dados deve ser executada, a partir de métodos que permitam descrever fatos, detectar padrões e desenvolver explicações (LEVINE, 1996). Com a análise de

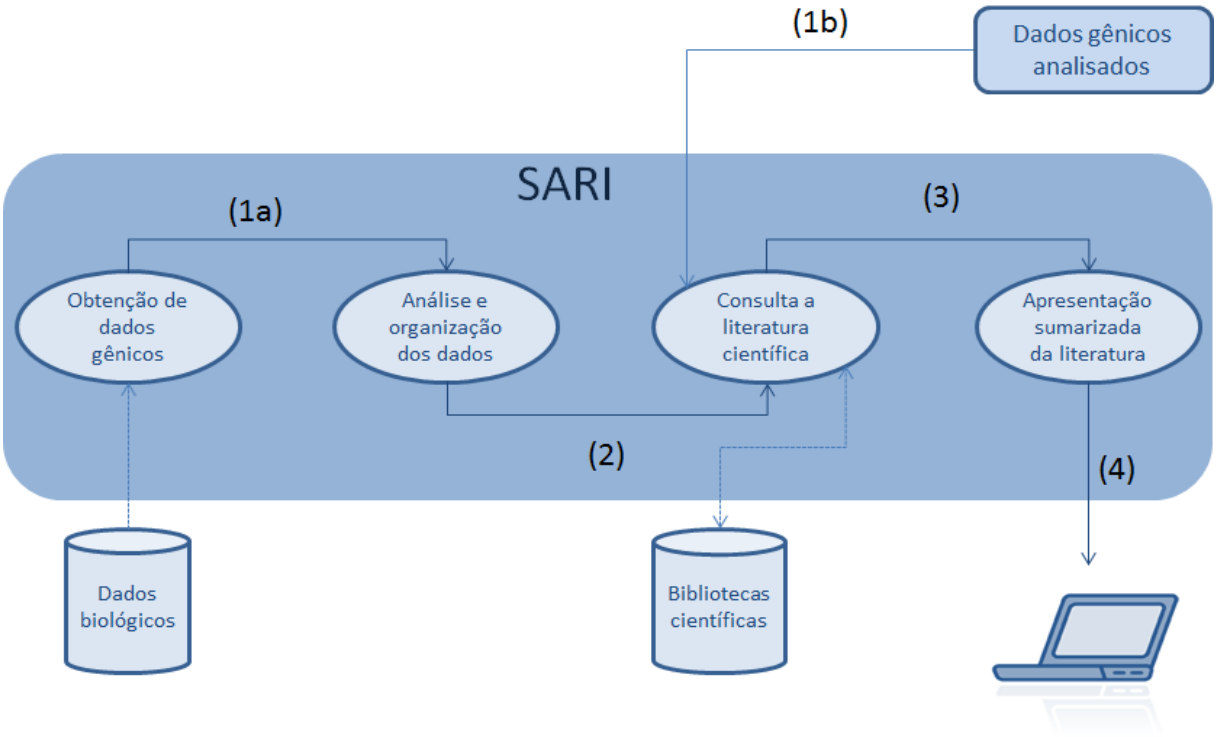


Figura 15: *Método SARI: (1a) dados de genes a serem analisados, (1b) busca de referências na literatura para resultados obtidos na análise gênica , (2) busca de referências na literatura para resultados obtidos na análise gênica, (3) submeter resultados da literatura ao processo de sumarização e (4) apresentar sumários*

dados, espera-se obter informação útil.

Um dos principais objetivos da análise de microarrays de DNA é agrupar genes com perfis de expressão similares (BABU, 2004). Para analisar dados de expressão gênica, ferramentas de mineração de dados são muito utilizadas. Alguns dos principais métodos de mineração de dados: classificação, clusterização, regressão, regras de associação, análise de séries temporais, entre outros. Na Seção 2.2, foram apresentados resumidamente fundamentos teóricos sobre clusters e suas classificações.

A consulta a literatura é o terceiro passo do SARI, cujo objetivo é verificar os resultados obtidos na etapa de análise. É importante relacionar os dados de expressão gênica analisados com outras informações biológicas presentes na literatura científica para verificar a fundamentação científica dos genes agrupados. Ao estabelecer uma relação entre os dados de expressão com informações externas, consegue-se ganhar conhecimento ou incentivar novas descobertas sobre os processos biológicos (BABU, 2004). O conhecimento prévio publicado na literatura pode reforçar ou contradizer os resultados obtidos na análise de dados. Desta maneira, pode-se também gerar novos focos de estudo para resultados obtidos, mas não presentes no estado da arte. Algumas questões que podem ser abordadas após a análise de dados são: predizer sítios de ligação; predizer interações e funções gênicas e proteicas; predizer módulos conservados e inferir redes regulatórias.

Ao consultar um assunto de interesse na literatura, a quantidade de informação retornada pode ser bastante extensa. Consequentemente, é difícil para o usuário assimilar a informação sem se sobrecarregar ou até mesmo ficar perdido. Portanto, propôs-se uma aplicação de sumarização auto-

mática de textos como a quarta etapa do método SARI. A sumarização automática permite reduzir a quantidade de conteúdo textual, sem que a informação principal do texto seja perdida. Na Seção 2.5, foram exibidos o conceito de sumarização e as várias classificações em que um sumário pode assumir.

Finalmente, a última etapa do SARI é apresentar os sumários ao usuário. Além dos sumários, elaborou-se uma rede das interações identificadas. Assim, além do texto, exibe-se uma visualização gráfica dos resultados. Os nós da rede são os genes e as arestas que ligam dois genes são os artigos cujas interações estão descritas. Uma integração visual e textual que auxilia na atribuição de significado aos clusters pode ser visualizada na Figura 16. Nas subseções seguintes, apresenta-se cenários de instanciação do método SARI.

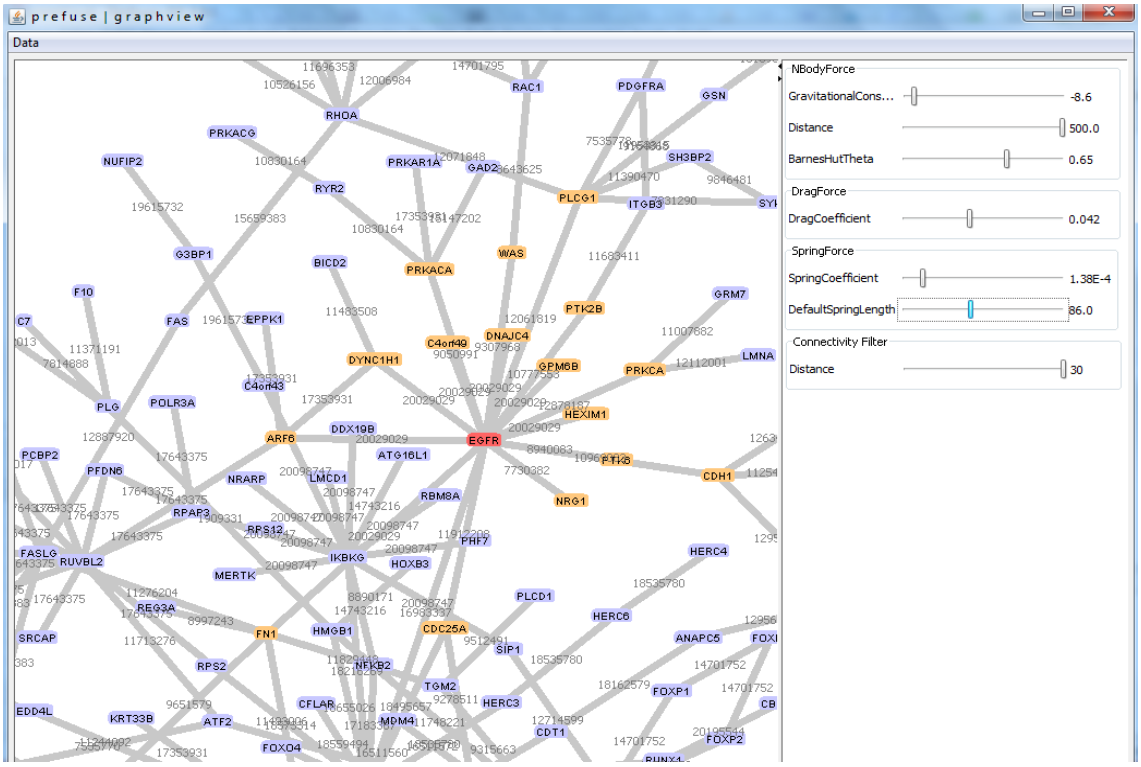


Figura 16: *Exemplo de rede de interações*

3.2 Aplicação do SARI: expressão gênica do GEO, clusterização com Weka e sumarização mono-documento

Esta seção apresenta uma instanciação do SARI utilizando dados gênicos do GEO, algoritmos de clusterização da Weka e sumarização mono-documentos, ou seja, apenas um artigo científico para cada interação . A Figura 17 apresenta os passos dessa instanciação do SARI.

Seguindo o método SARI, o primeiro passo foi obter um conjunto de dados de expressão gênica, o qual obteve-se na base de dados GEO, conforme apresentado na Figura 17(**a**). Houve um pré-processamento dos dados para analisá-los com as ferramentas da Weka. A Weka manipula

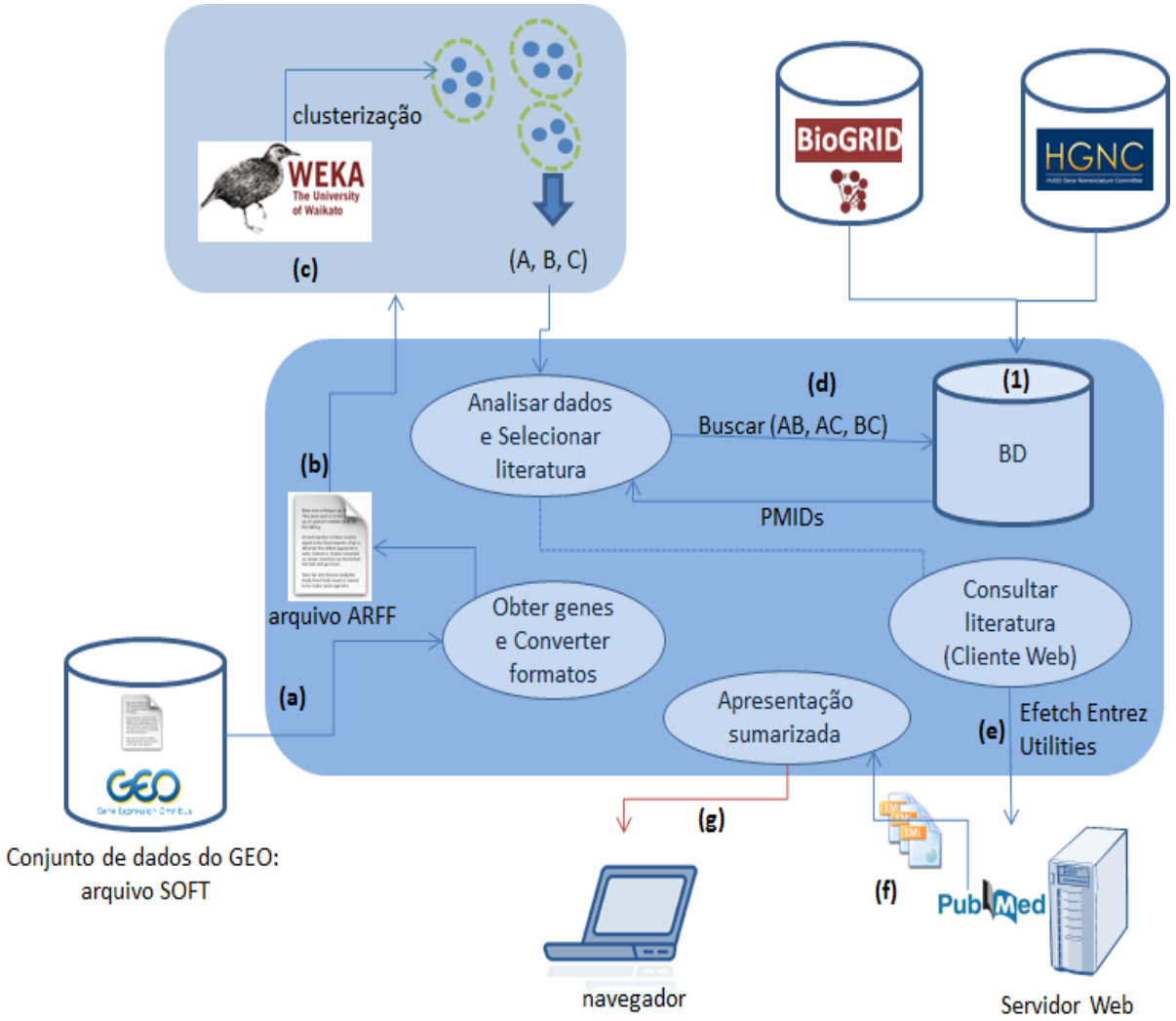


Figura 17: *Instanciação do método SARI com dados de expressão gênica do GEO, clusterização com algoritmos da Weka e sumarização mono-documento*

arquivos no formato ARFF (*Attribute-Relation File Format*) e o GEO disponibiliza seus dados no formato SOFT (ver Figura 10). A primeira etapa do SARI nessa aplicação, consistiu em adquirir um conjunto de dados e aplicar um algoritmo Java de conversão de formato SOFT para formato ARFF (Figura 17**(b)**). Um arquivo resultante da conversão do formato SOFT para o formato ARFF está ilustrado na Figura 18. O arquivo ARFF possui duas seções: cabeçalho e dados. O cabeçalho contém o nome da relação (*@relation <nome_relação>*) e uma lista de atributos (*@attribute <nome_atributo> <tipo_atributo>*). A ordem, em que os atributos são listados, indica o posicionamento da coluna na seção de dados. Quando um atributo é o terceiro a ser declarado na lista de atributos, a Weka assume que os valores desse atributo serão encontrados na terceira coluna delimitada por vírgulas. A linha *@data* delimita o início da seção de dados. Cada linha representa um exemplo (ver Figura 2).

O próximo passo do SARI, representado na Figura 17**(c)**, foi a clusterização com o algoritmo k-means da Weka. A Weka foi utilizada, pois possui vários algoritmos e softwares de licença livre para aprendizado de máquina, incluindo diversos algoritmos de clusterização (HALL et al., 2009). Nessa etapa, a entrada foi o conjunto de dados de expressão gênica (pré-processado na etapa

```
@relation expressao_GDS3717

@attribute IDENTIFIER string
@attribute GSM455115 numeric
@attribute GSM455116 numeric.
.
.
.

@data
DDR1,0.953,1.006,1.041,1.245...
RFC2,1.064,0.940,0.996,0.932...
```

Figura 18: *Exemplo de arquivo ARFF*

anterior) e a saída são os genes agrupados em clusters.

Para a instanciação do método SARI, criou-se um banco de dados seguindo o modelo de entidade-relacionamento, conforme apresentado na Figura 19. O banco de dados foi construído para armazenar as seguintes informações do BioGRID: (i) nome, símbolo oficial e aliases; (ii) informações dos genes que interagem entre si e em qual publicação científica essas interações estão demonstradas; e (iii) as informações de nomenclatura gênica do HGNC. A entidade `Gene` possui um símbolo oficial e um nome aprovado pelo HGNC. Os genes possuem vários nomes e símbolos alternativos ainda utilizados, representados pelas subentidades `AliasName` e `Alias Symbol`. Há também nomes e símbolos adotados antes da oficialização dos nomes pelo HGNC, representados pelas subentidades `PreviousName` e `PreviousSymbol`. Genes que interagem entre si estão representados na entidade `Interaction`. Uma interação é descrita na entidade `Article`, um artigo pode ter várias interações e uma interação pode ser abordada em artigos diferentes. As informações inseridas no banco de dados advém dos dados dos arquivos provenientes do BioGRID e do HGNC apresentadas respectivamente nas Figuras 9 e 13. A base de dados está representada na Figura 17(**1**). O sistema gerenciador de banco de dados utilizado foi o PostgreSQL.

Após agrupar os genes em clusters com o algoritmo de clusterização da Weka, consultou-se o banco de dados para identificar artigos que contivessem interações entre os genes. Esta etapa está demonstrada na Figura 17(**d**). Dado um cluster que contenha os genes [*A, B, C*], procurou-se no banco de dados as possíveis combinações de interações sem repetições dos genes do cluster, ou seja, artigos que apresentassem textualmente as interações para as combinações: *AB*, *AC*, *BC*. O resultado da consulta foi uma lista de PMIDs (*PubMed Identifier*) de artigos.

A etapa **(e)**, na Figura 17, representa a consulta ao PubMed utilizando a ferramenta EFetch. Criou-se um cliente Web para consumir o serviço EFetch. Na Figura 20, apresenta-se a arquitetura cliente-servidor desenvolvida. Um serviço Web é um software que espera requisições de

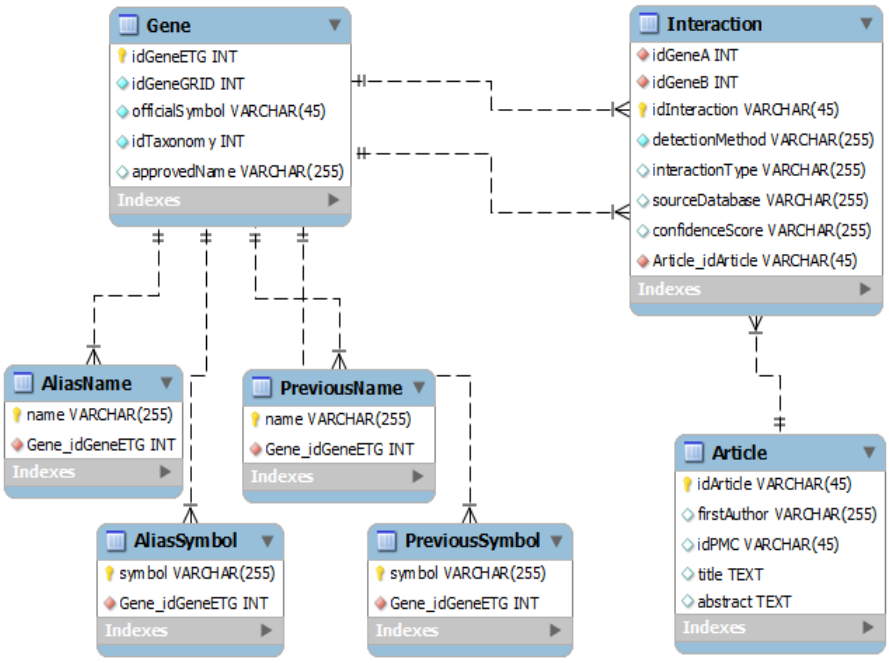


Figura 19: *Diagrama de entidade-relacionamento*

um outro software, que é o cliente Web. No cliente, codificado em Java, especificou-se a URL para o serviço Web: http://eutils.ncbi.nlm.nih.gov/soap/v2.0/efetch_pubmed.wsdl. O serviço Web `eFetchPubmedService` disponibilizou a operação `run_eFetch`. Utilizando o assistente de Serviços Web JAX-WS[1], os stubs cliente foram criados, os quais são objetos java remotos produzidos a partir das descrições contidas no arquivo WSDL (*Web Services Descripiton Language*). Os *stubs*[2] são os *proxys* no cliente e os *skeletons*[3] são os *proxys* no servidor. O ambiente utilizado para desenvolver o serviço Web foi o contâiner Web Tomcat (APACHE, 2012b) e o conjunto de ferramentas do Axis2 (APACHE, 2012a). O Axis2 implementa o envio de mensagens SOAP(GUDGIN et al., 2007) sobre o HTTP. As saídas resultantes, após consumir o serviço Web EFetch, estão ilustradas na Figura 17**(f)**.

Consumindo o serviço EFetch, adquiriu-se os arquivos XML dos artigos, que passaram pelo processo de sumarização. A Figura 21 mostra um exemplo de artigo recuperado do PubMed em formato XML. Pode-se observar que é necessário um processamento desses arquivos XML, para que o texto de interesse seja extraído. Na Figura 21, estão circulados os marcadores cujo conteúdo textual será extraído para passar pelo processo de sumarização: o título (*ArticleTitle*) e o abstract (*AbstractText*) do artigo.

Aplicou-se uma abordagem superficial de sumarização nos artigos recuperados. A sumarização está citada na Figura 17**(g)**. Uma interação entre dois genes pode estar relatada em diversos artigos, mas nessa aplicação do SARI utilizou-se apenas um artigo para cada interação, escolhido arbitrariamente, como apresentado na Figura 22.

[1] *Java API for XML Web Services*

[2] Stub (proxy cliente) - encapsula as invocações dos métodos que serão enviadas para o servidor e desencapsula as respostas do servidor

[3] Skeleton (proxy servidor) - desencapsula invocações remotas do cliente e encapsula os resultados da invocação

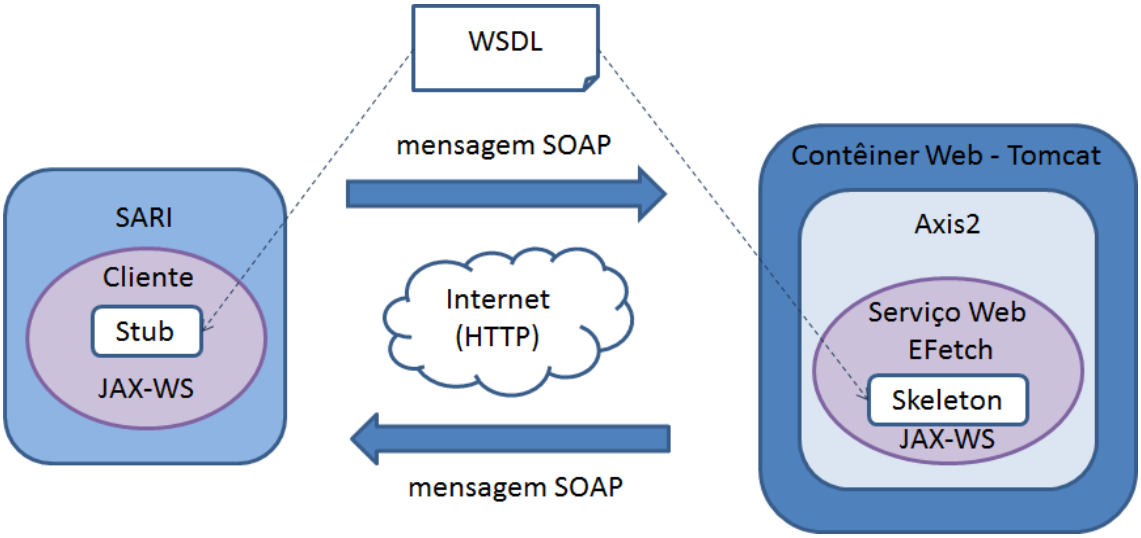


Figura 20: *Arquitetura serviço web*

```
<PubmedArticle>
    <MedlineCitation Owner="NLM" Status="In-Data-Review">
        <PMID Version="1">22450142</PMID>
        <DateCreated>
            <Year>2012</Year>
            <Month>03</Month>
            <Day>27</Day>
        </DateCreated>
        <Article PubModel="Print">
            <Journal>
                <ISSN IssnType="Print">0030-6002</ISSN>
                <JournalIssue CitedMedium="Print">
                    <Volume>153</Volume>
                    <Issue>14</Issue>
                    <PubDate>
                        <Year>2012</Year>
                        <Month>Apr</Month>
                        <Day>8</Day>
                    </PubDate>
                </JournalIssue>
                <Title>Orvosi hetilap</Title>
                <ISOAbbreviation>Orv Hetil</ISOAbbreviation>
            </Journal>
            <ArticleTitle>Leukemia- and lymphoma-associated flow cytometric, cytogenetic, and molecular genetic aberrations
            <Pagination>
                <MedlinePgn>531-40</MedlinePgn>
            </Pagination>
            <Abstract>
                <AbstractText>Most leukemia and lymphoma cases are characterized by specific flow cytometric, cytogenetic and
            </Abstract>
            <Affiliation>Jósa András Oktatókórház Egészségügyi Szolgáltató Nonprofit Kft. Hematológiai Osztály Nyíregyháza Lu
            <AuthorList CompleteYN="Y">
                <Author ValidYN="Y">
                    <LastName>Jakó</LastName>
                    <ForeName>János</ForeName>
                    <Initials>J</Initials>
                </Author>
                <Author ValidYN="Y">
                    <LastName>Szerafin</LastName>
```

Figura 21: *Fragmento de artigo em formato XML*

A sumarização foi executada a partir de uma adaptação do método de palavras-chave. O método de palavras-chave é baseado na premissa de que autor do texto usa algumas palavras-chave para expressar suas idéias e essas palavras se repetem ao longo do texto. No entanto, o interesse neste trabalho é identificar as interações nos artigos, e não o tópico central da publicação. Portanto, considerou-se como palavra-chave os nomes dos genes, ou quaisquer palavras que identifiquem um gene, como símbolo oficial, símbolos e nomes alias, símbolos e nomes prévios ou não oficiais. Logo, a sumarização consiste nos seguintes passos:

Figura 22: *Sumarização com documento único*

1. identificar sentenças relevantes, ou seja, as sentenças que possuem os nomes ou símbolos ou quaisquer palavras que identifiquem um gene;
2. extrair do texto original as sentenças de interesse, identificadas no passo anterior;
3. justapor as sentenças para formar o sumário.

Todas as sentenças classificadas como de interesse fizeram parte do sumário, ou seja, não foi aplicada nenhum tipo de taxa de compressão. Esse método será referido doravante no presente documento como sumarização mono-documento, isto é, em que um artigo resulta em um sumário.

3.3 Aplicação do SARI: grupos de genes e sumarização multi-documentos

Pesquisadores podem aplicar o SARI utilizando um conjunto particular de dados expressão gênica, o qual não está em base de dados disponibilizadas na Internet. Também há diversos algoritmos de clusterização que podem ser aplicados além do k-means. Portanto, diferentemente do cenário apresentado na Seção 3.2, deseja-se que a análise de interações gênicas por meio de consulta a literatura e sumarização seja aplicada não importando os métodos pelos quais foram obtidos os clusters de genes. Para a instanciação do método SARI neste caso, deseja-se obter sumários de artigos que contenham interações entre genes, sendo invisível o método utilizado para chegar nesse conjunto de genes. Na Figura 23, pode-se observar a abstração do método SARI para o novo cenário.

Nessa segunda instanciação, um usuário insere um grupo de genes em uma interface web, a qual foi codificada com tecnologia JSP (*JavaServer Pages*), como pode-se ver na etapa **(a)** da Figura 23. Na etapa (**b**), observa-se a transição dos genes inseridos da interface para a camada de negócios do software. Busca-se quais dentre os genes inseridos possuem interações. Esta interações encontram-se nas informações do BioGRID armazenadas na base de dados, a consulta está representada na etapa (**c**). A resposta retornada do banco de dados são os PMIDs das interações identificadas, etapa mostrada pela etapa (**d**). O cliente web é um outro módulo codificado, o qual é responsável por buscar as informações dos artigos com a lista de PMIDs obtida na etapa (**d**). A

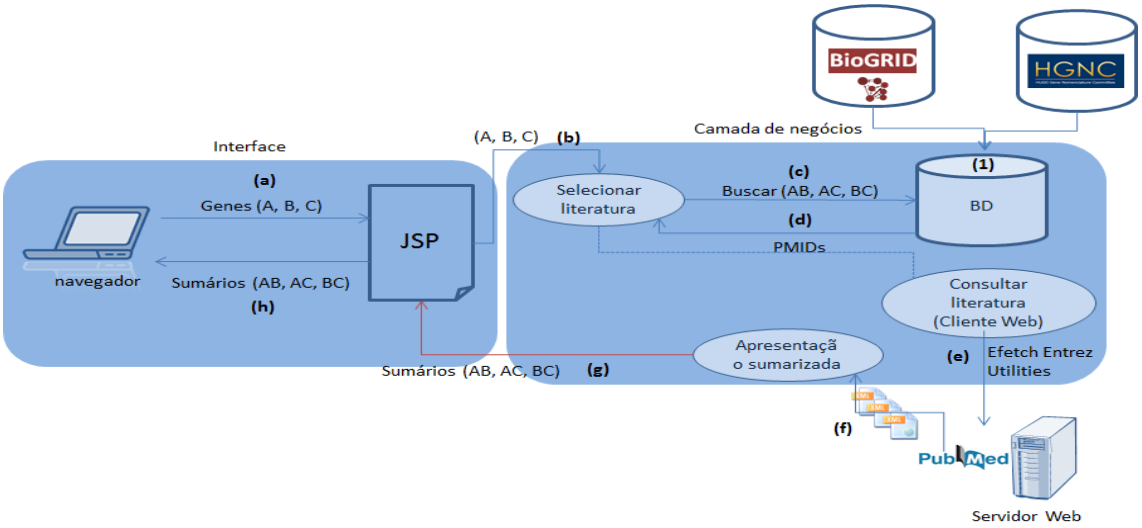


Figura 23: *Instanciação do método SARI com sumarização de múltiplos documentos*

etapa (**e**) consiste na requisição dos artigos pelo cliente web. O funcionamento do cliente web foi abordado na Figura 20. Na etapa (**f**), os artigos no formato XML retornados pelo servidor web são enviados para o sumarizador. Após o processo de sumarização, os sumários são enviados para a interface web, representado na etapa (**g**). Finalmente, os sumários são apresentados para o usuário, etapa (**h**). Resumidamente, o usuário fornece como entrada um conjunto de genes e como saída obtém as interações e os sumários dos diversos artigos em que as interações aparecem.

Uma interação entre dois genes pode estar descrita em vários artigos. Assim, a sumarização nessa instanciação foi projetada para suportar múltiplos documentos, como pode-se observar na Figura 24. Utilizando-se múltiplos documentos, há uma cobertura maior da literatura sobre determinada interação.

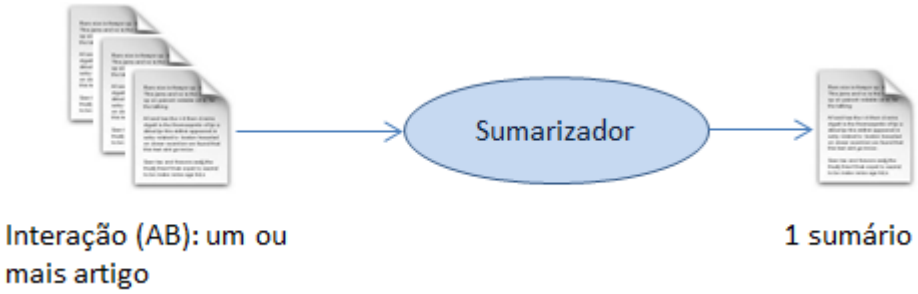


Figura 24: *Sumarização com múltiplos documentos*

Na sumarização codificada para essa instanciação, adotou-se o conceito de taxa de compressão. Como foram utilizados múltiplos documentos, a falta de compressão poderia ocasionar a produção de sumários extensos. Para selecionar as sentenças de interesse (que contêm nomenclatura gênica) para formação do sumário, utilizou-se três critérios de seleção das sentenças: quantidade de palavras-chave, data da publicação do artigo e tamanho da sentença. As sentenças em que aparecem mais vezes nomes, símbolos ou alias são consideradas mais importantes. As sentenças que pertencem a artigos mais recentes contém informação mais atualizada. As sentenças menores foram privelegiadas, pois elas transmitem informação de forma mais ágil e concisa para o usuário.

A segunda aplicação do SARI, que considera a sumarização de múltiplos documentos, é mais abrangente para a etapa de sumarização deo método SARI, já que recupera, sumariza e apresenta conteúdo da literatura científica. Além disso, o segundo cenário permite mais liberdade ao usuário, que pode agrupar os genes seguindo métodos que lhe pareça mais convenientes. Desse modo, essa aplicação não instanciou os passos de coleta e análise de dados biológicos do método SARI. No próximo capítulo, são demonstrados os experimentos e os resultados obtidos com as instanciações do SARI.

4 Experimentação e Resultados

Resumidamente, as principais etapas do método SARI são análise de dados e processamento da literatura para apresentação. Neste capítulo, são apresentados os experimentos realizados e os resultados obtidos para as duas aplicações do método SARI apresentados no capítulo anterior. A primeira aplicação (Seção 3.2) teve seus experimentos divididos em: clusterização e sumarização. Na etapa de clusterização, apresenta-se o conjunto de dados de expressão gênica utilizado e a clusterização aplicada. Já na etapa de sumarização, mostra-se as interações gênicas encontradas na literatura para genes que pertencem ao mesmo cluster, além da sumarização de artigos em que interações estejam presentes. Na segunda aplicação do SARI (Seção 3.3), não há experimentos de clusterização nesta instanciação, pois o usuário tem a liberdade de obter os clusters gênicos da maneira que lhe for mais conveniente. A etapa de sumarização utilizou todos os documentos em que uma interação é relatada, como as interações podem estar especificadas em mais de um artigo, aplicou-se a sumarização multi-documentos. Apresenta-se também a avaliação de usabilidade dos sumários realizada. Finalmente, apresenta-se uma comparação dos resultados da aplicação SARI com algumas ferramentas de busca.

4.1 Experimentação da primeira aplicação do SARI: expressão gênica do GEO, clusterização com Weka e sumarização mono-documento

Nessa experimentação do SARI, considerou-se dados de expressão gênica obtidos na base de dados online para posterior análise e sumarização de único artigo científico, o qual foi selecionado aleatoriamente, para apresentação.

4.1.1 Análise de dados utilizando clusterização

Na experimentação desta instanciação, utilizou-se o seguinte conjunto de dados de expressão gênica proveniente do GEO: *NOTCH antagonist SAHM1 effect on T-ALL cell lines*. Esse conjunto de dados é originário de uma pesquisa que relata o processo de desenho de peptídeos, os quais são alvos críticos na interface proteína-proteína do complexo NOTCH. Proteínas NOTCH participam de vias conservadas que regulam a diferenciação, proliferação e morte celular. Normalmente, a du-

ração e a força da sinalização do NOTCH é rigidamente controlada. Quando ocorrem mutações de perda de função, são observadas diversas doenças. Já mutações de ganho de função na via NOTCH são relacionadas ao desenvolvimento de câncer. A ativação inapropriada do receptor NOTCH está diretamente ligada a várias patogêneses, inclusive a leucemia linfoblástica aguda. O tratamento de células leucêmicas com o peptídeo SAHM1 resultou na supressão dos genes NOTCH ativados. É demonstrado que o peptídeo SAHM1 previne a montagem do complexo de transcrição ativo (MOELLERING et al., 2009).

O conjunto de dados de expressão gênica foi convertido do formato SOFT para o formato ARFF. Aplicou-se o algoritmo de clusterização k-Means da Weka, conforme apresentado na Seção 3.2. Os parâmetros utilizados para a execução do algoritmo de clusterização foram similares aos da ferramenta de clusterização e visualização do GEO. O valor escolhido para k foi 15, logo os genes foram agrupados em 15 clusters distintos e utilizou-se a distância euclidiana para calcular a distância entre os genes. A Tabela 3 apresenta a quantidade de genes agrupados em cada cluster. O cluster 1 possui mais grupos e o cluster 9 agrupou o menor número de genes. O cluster 9 possui poucos genes, isto pode significar que estes genes diferenciam-se bastante dos demais. Logo, podem ser genes com importância para o estudo do NOTCH. Já, clusters muito grandes representam genes com expressão mediana, ou seja, que não sofreram grandes alterações quando expostos ao peptídeo SAHM1.

Tabela 3: *Número de genes em cada cluster*

Cluster	Número de genes
1	2952
2	989
3	1043
4	1209
5	757
6	1106
7	258
8	1054
9	73
10	1025
11	1909
12	938
13	1119
14	1068
15	1340

4.1.2 Experimentação da sumarização automática

Para a experimentação da sumarização, precisou-se identificar os genes de um mesmo cluster que continham interação. As interações foram buscadas na base de dados modelada para armazenar as informações provenientes do BioGRID e do HGNC. Como já apresentado na Figura 9, o

BioGRID disponibiliza pares de genes junto de um número identificador PubMed de um artigo, o qual contém a descrição da interação entre esses genes. Assim, modelou-se uma base de dados para armazenar a informação do BioGRID (ver Figura 19). Já que recuperar a informação da descrição da interação entre genes em um grande arquivo texto não era viável. A nomenclatura dos genes humanos fornecida pelo HGNC também foi inserida na base de dados.

Codificou-se um algoritmo, que tem como entrada os clusters formados com o k-means. Como resposta da consulta na base de dados, tem-se resultados semelhantes aos da Figura 25, apresentando quais genes do cluster tem sua interação confirmada pela literatura, segundo informações do BioGRID. Na primeira linha da Figura 25, observa-se que o algoritmo de clusterização k-Means, agrupou os genes *HCN4* e *RYR2* no mesmo cluster. No entanto, segundo as informações do BioGRID não há nenhum artigo do PubMed que demonstre interação entre esses genes ou entre seus produtos. Já a interação entre *HCN4* e *HCN2* (segunda linha da Figura 25) aparece em um artigo científico, cujo identificador no PubMed é: PMID 10197448.

```
Não há artigo relacionando 10021 HCN4 com 6262 RYR2
Interação 10021 HCN4 com 610 HCN2
PMID: 12034718
Interação 3925 STMN1 com 3312 HSPA8
PMID: 10197448
Interação 1956 EGFR com 5578 PRKCA
PMID: 12878187
Interação 1956 EGFR com 2335 FN1
PMID: 20029029
```

Figura 25: *Exemplo interações de um cluster*

A Tabela 4 apresenta o número de interações confirmadas por artigos científicos do PubMed em cada cluster. Cada interação está contida em um artigo, em alguns casos, um mesmo artigo pode conter mais de um par de genes interagindo. Nesse caso, o artigo é contabilizado duas vezes, uma vez que as sentenças de interesse irão mudar de acordo com a mudança dos genes. Com esses números de interações por cluster, pode-se observar a grande quantidade de informação na literatura envolvida na análise de um conjunto de dados de expressão gênica clusterizado.

Fazendo a consulta na base de dados, obtém-se uma lista de identificadores PubMed (PMIDs) das interações de cada cluster. Com os PMIDs, o cliente Web do SARI requisita informações do serviço Web EFetch (ver Figura 20). O servidor recupera registros de uma base de dados do NCBI, a partir de uma lista de um ou mais identificadores. O cliente Web envia uma lista de PMIDs e o serviço Web EFetch retorna as informações dos respectivos artigos.

Neste trabalho, as informações de interesse foram o título e o resumo dos artigos da lista. Nos experimentos das Seções 4.1.2.1 e 4.1.2.2, o texto original é a concatenção do título e do resumo do artigo. Um exemplo de um texto recuperado do PubMed está ilustrado na Figura 26. Essa figura apresenta o artigo de PMID 10197448, indicando interação entre os genes *STMN1* e *HSPA8*.

Tabela 4: *Número de interações em cada cluster*

Cluster	Número de interações
1	2194
2	220
3	203
4	317
5	98
6	255
7	25
8	290
9	3
10	218
11	996
12	173
13	377
14	202
15	482

TEXTO:

"Stathmin interaction with HSC70 family proteins. Stathmin is a ubiquitous cytosolic phosphoprotein participating in the relay and integration of diverse intracellular signaling pathways involved in the control of cell proliferation, differentiation, and activities. It is phosphorylated in response to diverse extracellular signals including hormones and growth factors, and it is highly expressed during development and in diverse tumoral cells and tissues. Stathmin interacts with tubulin and other potential protein partners such as BiP, KIS, CC1 and CC2/tsg101. In our present search for further functional partners of stathmin, we identified proteins in the Hsp70 family, and in particular Hsc70, as interacting with stathmin in vitro. Hsc70 is among the proteins coimmunoprecipitated with stathmin, and it is the main protein retained specifically on stathmin-Sepharose beads identified by one- and two-dimensional electrophoresis and immunoblots. Bovine serum albumin (BSA)-Sepharose did not bind Hsc70, and anti-stathmin antisera specifically inhibited the interaction of Hsc70 with stathmin-Sepharose. The binding of Hsc70 to stathmin is dependent on the phosphorylation status of stathmin, as it did not occur with a "pseudophosphorylated" mutant form of stathmin. This interaction is further dependent on the ATP status of Hsc70. It was inhibited in the presence of ATP-Mg++ but not in the presence of ATP-Mg++ and ethylenediaminetetraacetic acid (EDTA) or of ADP. Our results suggest that the interaction of stathmin with Hsc70 is specific in both proteins and most likely biologically relevant in the context of their functional implication in the control of numerous intracellular signaling and regulatory pathways, and hence of normal cell growth and differentiation."

Figura 26: *Texto original: título concatenado com resumo do artigo*

Considerando o título e o resumo dos artigos de interesse, o sumarizador compõe o vetor de palavras-chave para cada artigo. Para construir este vetor, consulta-se na base de dados, a partir do símbolo oficial, todos os possíveis sinônimos do gene: os aliases. Assim, o vetor de palavras-chave é formado pelo símbolo oficial, nome oficial, símbolos aliases, nomes aliases, símbolos prévios e nomes prévios. Por exemplo, o vetor de palavras-chave do artigo apresentado na Figura 26 é: *[stathmin 1, STMN1, oncoprotein 18, MGC138870, MGC138869, C1orf215, Lag, LAP18, FLJ32206, OP18, PR22, PP19, PP17, SMN, chromosome 1 open reading frame 215, stathmin 1 oncoprotein 18, heat shock 70kDa protein 8, HSPA8, HSC70, HSP73, HSPA10, HSC54, HSC71, HSP71, MGC29929, MGC131511, NIP71, LAP1, heat shock 70kD protein 8]*. Esse vetor contém

o nome e o símbolo oficial dos genes STMN1 e HSPA8 e os seus nomes e símbolos aliases e prévios. O processo de sumarização consiste na identificação de sentenças que possuem uma ou mais palavras-chave do vetor. As sentenças identificadas são concatenadas e as sentenças que não contém palavras-chave são descartadas. A Figura 27 apresenta o sumário do texto exibido na Figura 26.

SUMÁRIO DO TEXTO:

"Stathmin interaction with HSC70 family proteins. In our present search for further functional partners of stathmin, we identified proteins in the Hsp70 family, and in particular Hsc70, as interacting with stathmin in vitro. Hsc70 is among the proteins coimmunoprecipitated with stathmin, and it is the main protein retained specifically on stathmin-Sepharose beads identified by one- and two-dimensional electrophoresis and immunoblots. Bovine serum albumin (BSA)-Sepharose did not bind Hsc70, and anti-stathmin antisera specifically inhibited the interaction of Hsc70 with stathmin-Sepharose. The binding of Hsc70 to stathmin is dependent on the phosphorylation status of stathmin, as it did not occur with a "`pseudophosphorylated"' mutant form of stathmin. This interaction is further dependent on the ATP status of Hsc70. Our results suggest that the interaction of stathmin with Hsc70 is specific in both proteins and most likely biologically relevant in the context of their functional implication in the control of numerous intracellular signaling and regulatory pathways, and hence of normal cell growth and differentiation."

Figura 27: *Sumário do Texto*

Foram realizados dois experimentos para medir a redução dos textos utilizando os clusters e as interações mostradas nas Tabelas 3 e 4. No primeiro experimento, o vetor de palavras-chave possuía somente os nomes e símbolos oficiais dos genes. Este experimento está detalhado na Seção 4.1.2.1. No segundo experimento, o vetor de palavras possuía nomes e símbolos oficiais e não oficiais, apresentado na Seção 4.1.2.2. Os dois experimentos foram feitos com o mesmo conjunto de dados *NOTCH antagonist SAHM1 effect on T-ALL cell lines*, proveniente do GEO. Nas subseções a seguir, são apresentados os resultados dos experimentos de sumarização automática nos artigos em que interações foram detectadas.

4.1.2.1 Sumarização sem alias

No experimento sem considerar os aliases, o vetor de palavras-chave usado para identificar as sentenças de interesse continha apenas nomes e símbolos oficiais dos genes que estão interagindo (um vetor de tamanho constante de quatro palavras-chave). Assim, o sumário foi formado somente com sentenças que continham alguma nomenclatura oficial do par de genes de interesse.

A Figura 28 apresenta a frequência média de sentenças nos textos originais e nos sumários. O tamanho médio dos textos fontes é 9,4 sentenças. Com a sumarização, obtém-se textos com tamanho médio de 4,7 sentenças. O cluster 9 é pequeno, pois possui apenas 73 genes (ver Tabela 3) e com apenas três pares de genes interagindo (ver Tabela 4). Portanto, ao utilizar somente os nomes e os símbolos oficiais, nenhum dos três artigos do cluster 9 formou sumário, ou seja, título e resumo não citam o nome ou o símbolo oficial dos genes que possuem a interação caracterizada no artigo. Logo, infere-se que os autores utilizaram nomes e símbolos não oficiais para identificar os genes nesses artigos.

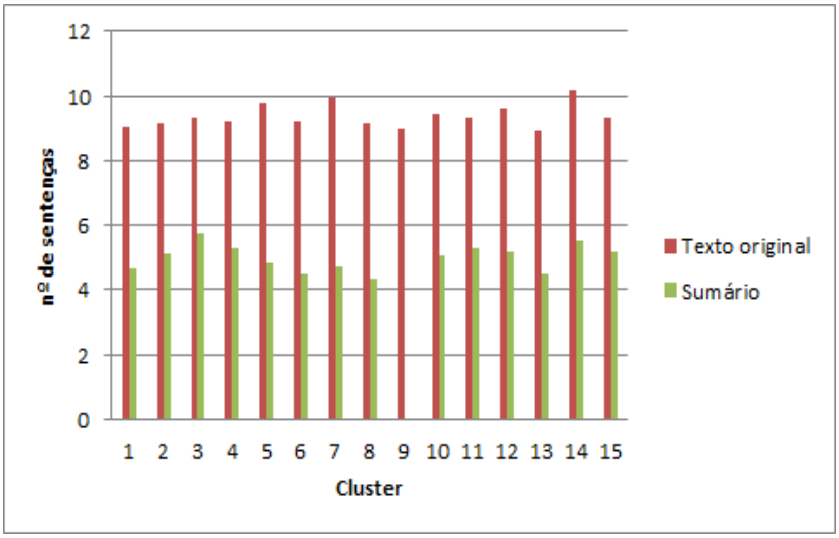


Figura 28: *Quantidade média de sentenças nos textos originais e nos sumários - sem alias*

4.1.2.2 Sumarização com alias

No experimento de sumarização do SARI considerando os alias dos genes, o vetor de palavras-chave continha nomes e símbolos oficiais dos genes, assim como outros nomes e símbolos não oficiais que são utilizados para identificar os genes, os aliases. Na Figura 29, observa-se a quantidade média de sentenças nos textos originais e nos sumários, a quantidade média de sentenças no texto original é 9,4, no sumário foi de 6,2.

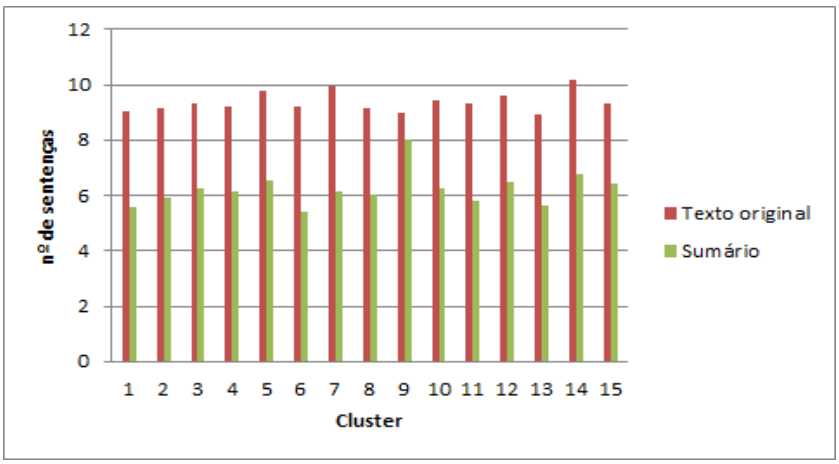


Figura 29: *Quantidade média de sentenças nos textos originais e nos sumários - com alias*

O gráfico da Figura 30 ilustra a quantidade de artigos, que segundo o BioGRID, possuem a interação entre dois genes, só que nenhum sumário foi formado. Portanto, não há sentenças que possuam termos identificadores de genes no título ou no *abstract*. A coluna *oficiais* indica quando o vetor é formado pelo nome e símbolo oficial do par de genes que interagem; a coluna *oficiais e alias* representa quando o vetor de palavras-chave contém nomenclatura oficial e alias. Pode-se observar que, quando se utiliza exclusivamente a nomenclatura oficial, há uma grande quantidade de sumários não formados. Porém, quando utiliza-se os *aliases*, o número de sumários não formados diminui. Essa constatação sugere que a nomenclatura oficial é pouco utilizada e os pseudônimos dos genes estão muito presentes na literatura. Por exemplo, observa-se na Figura 28,

que não foram formados sumários no cluster 9. Essa situação não ocorreu no experimento somente com nomenclatura oficial. Esse resultado pode demonstrar que muitos artigos citam os genes sem utilizar a nomenclatura oficial.

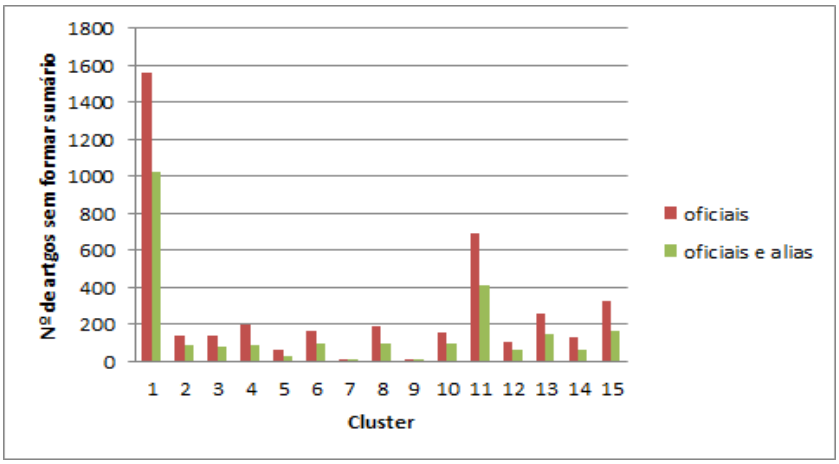


Figura 30: *Quantidade de textos que não formaram sumários*

Na Tabela 5, expõe-se um resumo dos resultados dos experimentos de sumarização com e sem alias. Quando foi utilizado vetor de palavras-chave somente com nomenclatura oficial, a quantidade média de sentenças, 4,7, foi menor do que utilizando alias, 6,2. Se o intuito é reduzir a quantidade de texto apresentado ao usuário, pode-se afirmar que os sumários com palavras-chave somente oficiais foi mais eficiente na redução dos textos. No entanto, ao comparar a quantidade de sumários não formados ao utilizar somente a nomenclatura oficial, 275,1, e ao utilizar os alias, 163,3, observa-se que ignorar a nomenclatura não oficial resulta em grande perda de informação. Assim, pode-se concluir que quanto mais específicas são as palavras-chave, melhores serão os resultados na redução da quantidade de texto. Observa-se também que não se pode ignorar termos de nomenclatura não oficiais, pois eles são amplamente utilizados pelos autores.

Tabela 5: *Resumo dos Resultados dos Experimentos*

	Texto original	**Oficiais**	**Oficiais e alias**
nº de sentenças	9,3	4,7	6,2
nº de zeros	-	275,1	163,3

4.2 Experimentação da segunda aplicação do SARI: grupos de genes e sumarização multi-documentos

Na instanciação do método SARI, apresentado na Seção 3.3, o método pelo qual os genes foram agrupados não é uma preocupação. Portanto, foram feitos somente experimentos de sumarização automática. Buscou-se interações entre os seguintes genes: *MLL, MEN1, CREBBP, PPIE, XAB2, XPA, BRCA1*. Nesse grupo de sete genes, foram encontradas sete interações gênicas.

Aplicou-se o método de sumarização com taxa de compressão de 20%. Na Figura 31, pode-se observar a quantidade de sentenças obtidas em todos os artigos fonte, os quais apresentam a inte-

ração entre dois genes de interesse. Na mesma figura, há também a quantidade de sentenças que possuem palavras-chave e a quantidade de sentenças resultante após aplicar uma taxa de compressão de 20%. A média da quantidade de sentenças nos sumário foi 4,25 para essa experimentação do SARI.

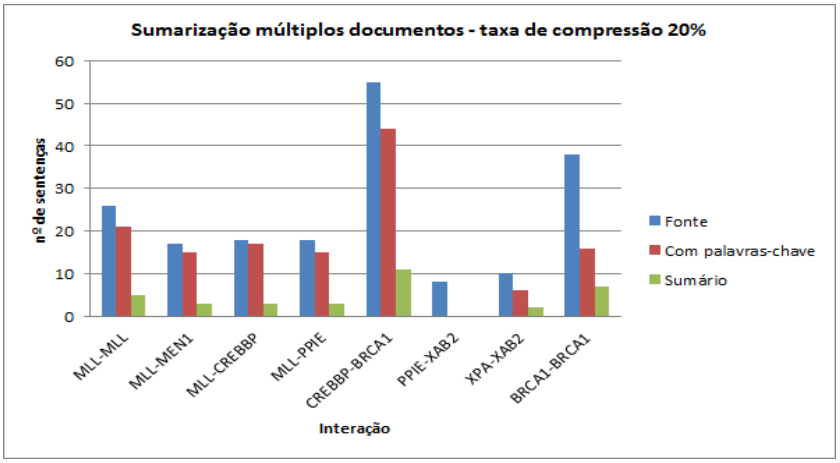


Figura 31: *Quantidade de sentenças nos textos originais e nos sumários - taxa de compressão 20%*

Na Figura 32, apresenta-se a quantidade de sentenças resultante quando a taxa de compressão aplicada é de 10%. Nesse caso, os sumário ficaram pequenos e a média de sentenças foi 1,75. Nas Figuras 31 e 32, observa-se que a interação entre PPIE e XAB2 não formou nenhum sumário, pois os artigos em que a interação está relatada não cita os genes no título e no *abstract*. Assim, a interação entre PPIE e XAB2, provavelmente, só pode ser encontrada ao consultar os artigos completos. Nas Figuras 31 e 32, observa-se a quantidade de texto que pode ser reduzida e ser apresentada de forma condensada para o usuário. Suponha que um usuário deseja buscar na literatura artigos que contenham informações sobre a interação entre os genes CREBBP e BRCA1. Encontrar todos os artigos que possuem informação da interação já não seria uma tarefa trivial. Adicionalmente, ele teria mais de cinquenta sentenças para ler, ao consultar somente título e *abstract* do artigo. Com sumarização automática para esse cenário, pode-se reduzir para cerca de dez sentenças ou até menos. No entanto, é necessário avaliar se estas poucas sentenças são realmente úteis para o usuário. Na próxima seção, apresenta-se a avaliação de utilidade, feita por usuários, para os sumários multi-documentos elaborados com taxas de compressão de 20% e 10% e também para os sumários mono-documentos.

4.3 Experimentação da avaliação da sumarização do SARI com usuário

A experimentação da sumarização considerando a opinião de julgadores tinha como objetivo avaliar os métodos de sumarização propostos em termos de utilidade (*usefulness*), em relação à tarefa de informar sobre a interação de pares de genes ou proteínas.

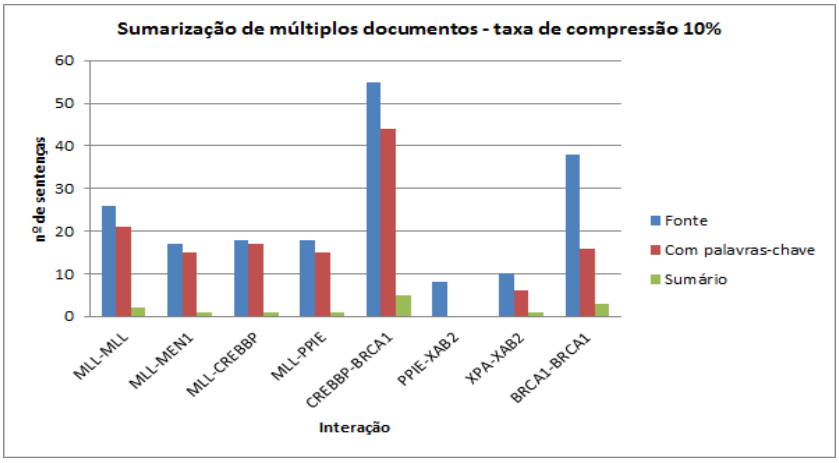


Figura 32: *Quantidade de sentenças nos textos originais e nos sumários - taxa de compressão de 10%*

4.3.1 Contextualização do experimento

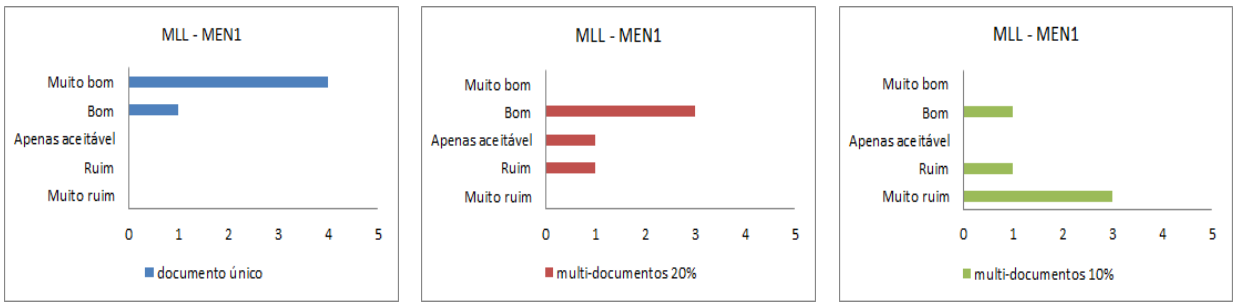
A avaliação aplicada aos sumários foi extrínseca. Esse tipo de avaliação é utilizado para determinar o efeito da sumarização em alguma tarefa. Cinco julgadores avaliaram um conjunto de sumários elaborados por métodos diferentes de sumarização no SARI. Como o contexto da aplicação era bastante específico, foram escolhidos julgadores com conhecimento em genética, biologia molecular, bioinformática, etc. Apresenta-se o perfil acadêmico dos julgadores no Apêndice B.

Simulou-se uma busca para encontrar e analisar as interações presentes no grupo de genes: *MLL, MEN1, CREBBP, PPIE, XAB2, XPA, BRCA1*. No banco de dados modelado, foram encontradas as seguintes interações entre estes genes: *MLL - MLL, MLL - MEN1, MLL - CREBBP, MLL - PPIE, CREBBP - BRCA1, PPIE - XAB2, XPA - XAB2, BRCA1 - BRCA1*. Para cada interação, foram criados três sumários. Um sumário foi criado a partir de apenas um texto fonte (Figura 22). Os outros dois sumários foram elaborados a partir de múltiplos documentos, com taxa de compressão de 20% e 10% (Figura 24). Os julgadores não sabiam quais sumários eram mono-documentos ou multi-documentos, apenas sabiam que haviam três tipos de sumário para cada interação gênica. Aos julgadores, solicitou-se a classificação dos sumários em: “Muito bom”, “Bom”, “Apenas aceitável”, “Ruim” ou “Muito ruim”. Para a avaliação, os sumários acompanhados de algumas instruções foram encaminhados por email aos julgadores, sem exigência de tempo máximo de avaliação. As instruções completas e os sumários avaliados podem ser vistos no Apêndice A. A seguir, apresenta-se a classificação de utilidade que os julgadores associaram a cada sumário.

4.3.2 Resultados

Nas Figuras 33(a), 33(b) e 33(c), apresentam-se as respostas dos julgadores para os três métodos de sumarização aplicados para a interação entre os genes MLL e MEN1. Os julgadores avaliaram como mais útil o sumário que foi feito a partir de um único documento. Quatro julgadores consideraram este sumário “Muito bom” e um julgador o considerou “Bom”. O segundo melhor sumário foi o elaborado a partir de múltiplos documentos com taxa de compressão de 20%.

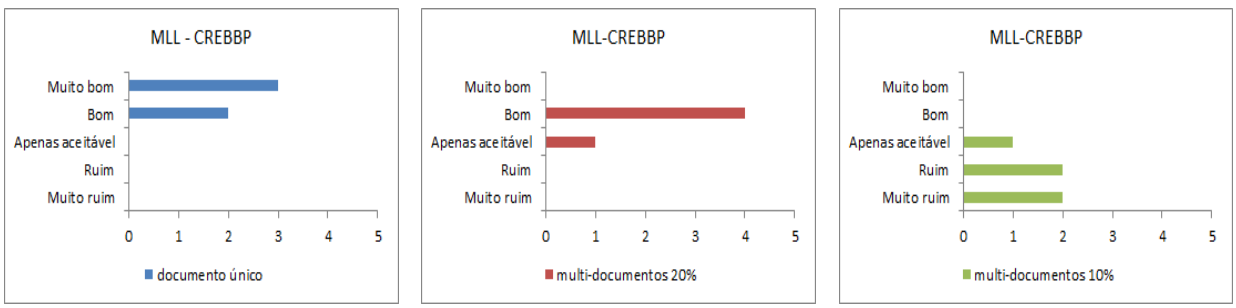
Este foi julgado como “Bom” três vezes, os outros dois julgadores classificaram o sumário como “Apenas aceitável” e “Ruim”. A pior utilidade foi do sumário feito a partir múltiplos documentos e taxa de compressão de 10%. Três julgadores o consideraram “Muito ruim”, os outros dois julgadores classificaram como “Ruim” e “Bom”. O sumário feito a partir de um único documento e o sumário com taxa de compressão 20% tiveram boa avaliação, pois em suas sentenças a interação entre os genes MLL e MEN1 foi realmente abordada. Já, o sumário com taxa de compressão de 10% ficou pequeno, apenas com uma sentença. Nesse sumário, o MEN1 não foi citado. Assim, o sumário não foi capaz de revelar a interação entre os dois genes.



(a) documento único (b) multi-documento 20% (c) multi-documento 10%

Figura 33: *Avaliação dos sumários da interação MLL - MEN1*

Nas Figuras 34(a), 34(b) e 34(c), apresentam-se as respostas dos julgadores para os três métodos de sumarização aplicados, para a interação entre os genes MLL e CREBBP. O sumário elaborado a partir de apenas um documento foi eleito o melhor pelos julgadores. Foi classificado como “Muito bom” três vezes e “Bom” duas. O sumário feito a partir de múltiplos documentos e taxa de compressão de 20% foi considerado “Bom” por quatro julgadores e “Apenas aceitável” por um. O pior desempenho foi do sumário com taxa de compressão de 10%. Foi considerado “Apenas aceitável” uma vez, “Ruim” duas vezes e “Muito ruim” duas vezes também. Os três sumários apresentavam a interação entre os genes MLL e CREBBP. No sumário feito de um único documento e no sumário com taxa de compressão de 20%, a referência ao gene CREBBP foi feita utilizando seu nome oficial: *CREB binding protein*. Enquanto, que no sumário com taxa de compressão de 10%, o gene CREBBP foi citado por um de seus alias: CBP.

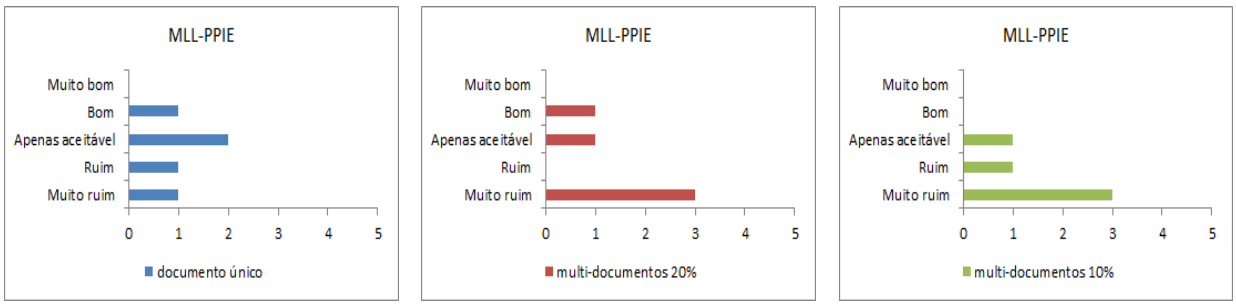


(a) documento único (b) multi-documento 20% (c) multi-documento 10%

Figura 34: *Avaliação dos sumários da interação MLL - CREBBP*

Nas Figuras 35(a), 35(b) e 35(c), apresenta-se as respostas dos julgadores para os três métodos

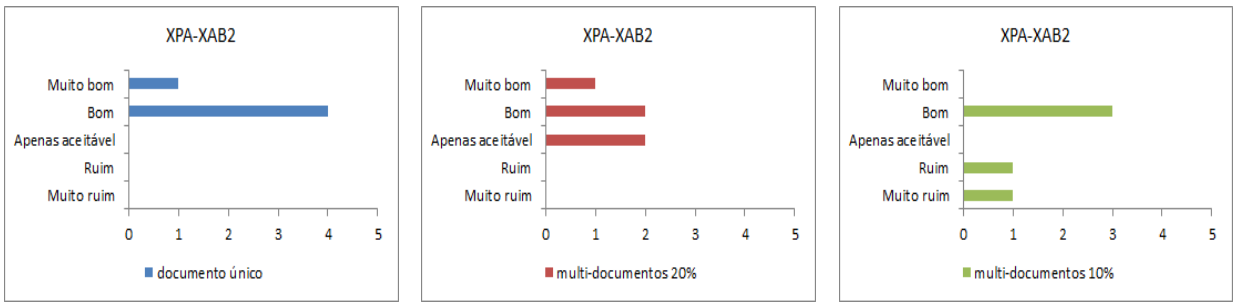
de sumarização aplicados, para a interação entre os genes MLL e PPIE. O sumário que foi feito a partir de um único documento e não sofreu taxa de compressão foi classificado como “Apenas aceitável” duas vezes e “Bom”, “Ruim” e “Muito ruim” uma vez cada. O sumário com taxa de compressão de 20% foi definido como “Muito ruim” por três julgadores, “Bom” por um julgador e “Apenas aceitável” por outro julgador. O sumário com taxa de compressão de 10% teve uma classificação similar ao sumário de 20%. Também foi classificado como “Muito ruim” por três julgadores e foi classificado como “Apenas aceitável” e “Ruim” por outros dois julgadores. Somente no sumário a partir de um documento a interação entre os genes MLL e PPIE aparece e o PPIE só é citado por meio de seu alias: Cyp33. Assim, o sumário mono-documento recebeu a melhor avaliação dentre os três, contudo o melhor sumário não foi unanimidade, provavelmente, devido ao fato do PPIE não ser tratado por meio da nomenclatura oficial.



(a) documento único (b) multi-documento 20% (c) multi-documento 10%

Figura 35: *Avaliação dos sumários da interação MLL - PPIE*

Nas Figuras 36(a), 36(b) e 36(c), apresenta-se as respostas dos julgadores para os três métodos de sumarização aplicados, para a interação entre os genes XPA e XAB2. Quatro julgadores classificaram como “Muito bom” e um como “Bom” o sumário cuja fonte foi apenas um artigo. O sumário com taxa de compressão de 20% foi declarado “Bom” por dois julgadores, “Apenas aceitável” por dois outros julgadores e “Muito bom” por um julgador. O sumário com taxa de compressão de 10% foi considerado “Bom” três vezes, “Ruim” uma vez e “Muito ruim” uma vez também. Em todos os sumários, a interação entre os genes XPA e XAB2 foi descrita e sempre utilizando símbolos oficiais. Esses aspectos determinaram o bom desempenho de todos os sumários.

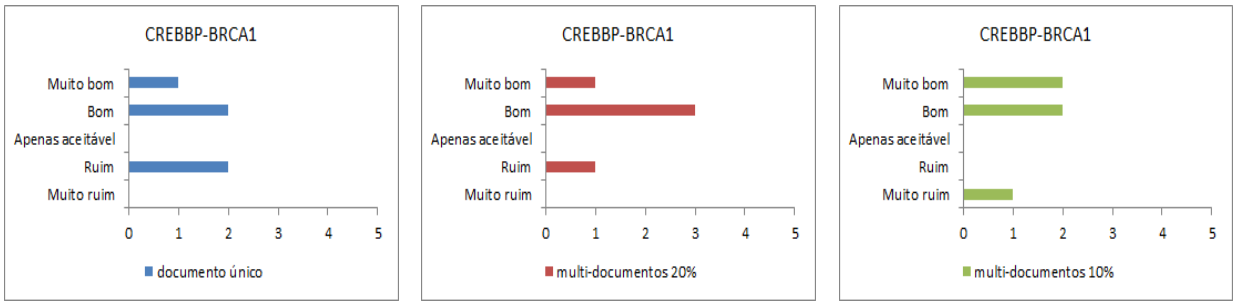


(a) documento único (b) multi-documento 20% (c) multi-documento 10%

Figura 36: *Avaliação dos sumários da interação XPA - XAB2*

As Figuras 37(a), 37(b) e 37(c) apresentam as respostas dos julgadores para os três métodos

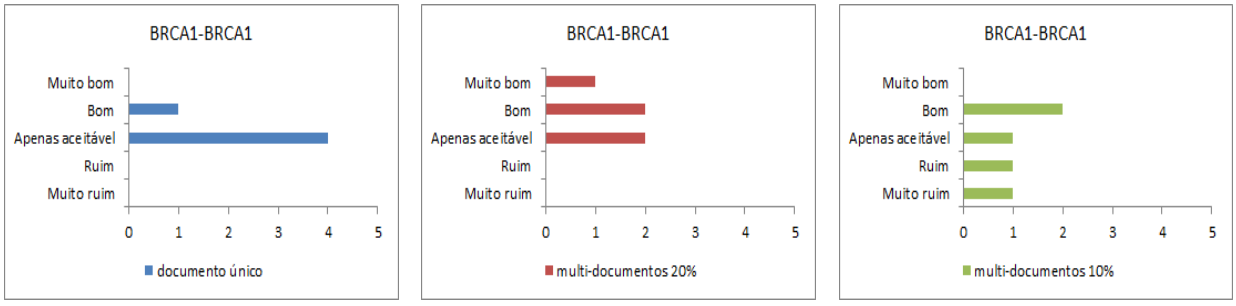
de sumarização aplicados, para a interação entre os genes CREBBP e BRCA1. Dois julgadores consideraram o sumário feito a partir de um único documento “Bom”, outros dois julgadores o consideraram “Ruim” e um julgador “Muito bom”. O sumário com múltiplos artigos fonte e taxa de compressão de 20% foi julgado com “Bom” três vezes, “Muito bom” uma vez e “Ruim” uma vez. O sumário com taxa de compressão de 10% foi considerado “Muito bom” duas vezes, “Bom” duas vezes e “Muito ruim” uma vez. Em todos os sumários, a interação entre os genes CREBBP e BRCA1 foi retratada. Este fato explica a boa avaliação de todos os sumários pelos julgadores.



(a) documento único (b) multi-documento 20% (c) multi-documento 10%

Figura 37: *Avaliação dos sumários da interação CREBBP - BRCA1*

Nas Figuras 38(a), 38(b) e 38(c), apresenta-se as respostas dos julgadores para os três métodos de sumarização aplicados, para a interação do gene BRCA1 com ele mesmo. Quatro julgadores consideraram o sumário feito a partir de apenas um documento “Apenas aceitável” e um julgador o considerou “Bom”. O sumário com taxa de compressão de 20% foi considerado “Bom” duas vezes, “Apenas aceitável” duas vezes e “Muito bom” uma vez. O sumário com taxa de compressão de 10% foi avaliado como “Bom” duas vezes, “Apenas aceitável” uma vez, “Ruim” uma vez e “Muito ruim” uma vez. O melhor desempenho foi do sumário multi-documentos com taxa de compressão de 20%.



(a) documento único (b) multi-documento 20% (c) multi-documento 10%

Figura 38: *Avaliação dos sumários da interação BRCA1 - BRCA1*

Além de classificar os sumários em questão de utilidade, disponibilizou-se um espaço para que os julgadores fizessem eventuais comentários. Nos comentários, um dos julgadores considerou resumos muito pequenos bastante práticos e úteis quando se quer estudar apenas a interação gênica, apesar de serem menos informativos. Em outro comentário, o julgador encontrou dificuldades em detectar os alias dos genes. Confirmando o fato que a não ou má utilização da nomenclatura pode

dificultar o entendimento de artigos.

4.3.3 Lições Aprendidas

Os resultados da avaliação de utilidade mostraram que os sumários bastante pequenos só são úteis se forem capazes de trazer a informação de interesse, a interação gênica, como no exemplo da interação entre XPA e XAB2. Em geral, houve uma preferência pelos sumários com mais sentenças e consequentemente com informação mais detalhada. Para os julgadores foi informado somente os símbolos oficiais dos genes da interação. Pode-se observar que quando os genes foram citados por meio de aliases ou o nome oficial, a utilidade atribuída foi menor. Este fato também pôde ser observado pelos comentários dos julgadores, demonstrando que a comunidade científica da área, além de não utilizar a nomenclatura oficial, muitas vezes também não conhece todos os pseudônimos dos genes.

Após a avaliação de utilidade pelos julgadores, conclui-se que um método de sumarização que ordene as sentenças pela presença dos dois genes da interação pode produzir sumários melhores. Verificou-se também que o uso de palavras-chave como "*bind*" e "*interaction*" pode ser mais interessante do que somente contabilizar o número de vezes os nomes, os símbolos ou os aliases dos genes aparecem. Existem sentenças que repetem mesmo gene diversas vezes, fazendo com que a sentença seja melhor ordenada.

4.4 Avaliação do uso de ferramentas de busca

A maior parte das interações biomoleculares e das vias biológicas estão representadas na literatura científica, ou seja, estão presentes em textos não estruturados (DONALDSON et al., 2003). Atualmente, acessar a literatura científica é bastante simples na internet. Porém, encontrar uma informação de interesse geralmente não é uma tarefa trivial, dada a grande quantidade de material publicado. Existem diversas ferramentas para acessar a literatura biomédica online. Nesta seção, buscou-se por um conjunto de três genes em ferramentas de busca conhecidas, como, Google, PubMed e BioGRID. O objetivo era encontrar textos que retratassem a interação entre genes ou entre seus produtos. Os resultados das ferramentas de busca foram comparados aos resultados do SARI.

4.4.1 Google

O Google é a ferramenta de busca mais difundida e é amplamente utilizado, apesar de não ser uma ferramenta específica para a literatura biomédica (STEINBROOK, 2006). Foram escolhidos um grupo de três genes que interagem entre si e utilizando o Google, buscou-se pelo símbolo oficial desses genes: *CALR*, *CANX* e *TF*. Os resultados do Google são ordenados por relevância e mais de

2 milhões de resultados foram encontrados. Na Figura 39, pode-se observar os quatro resultados mais relevantes. Apenas o terceiro parece estar relacionado a biologia molecular e genética.

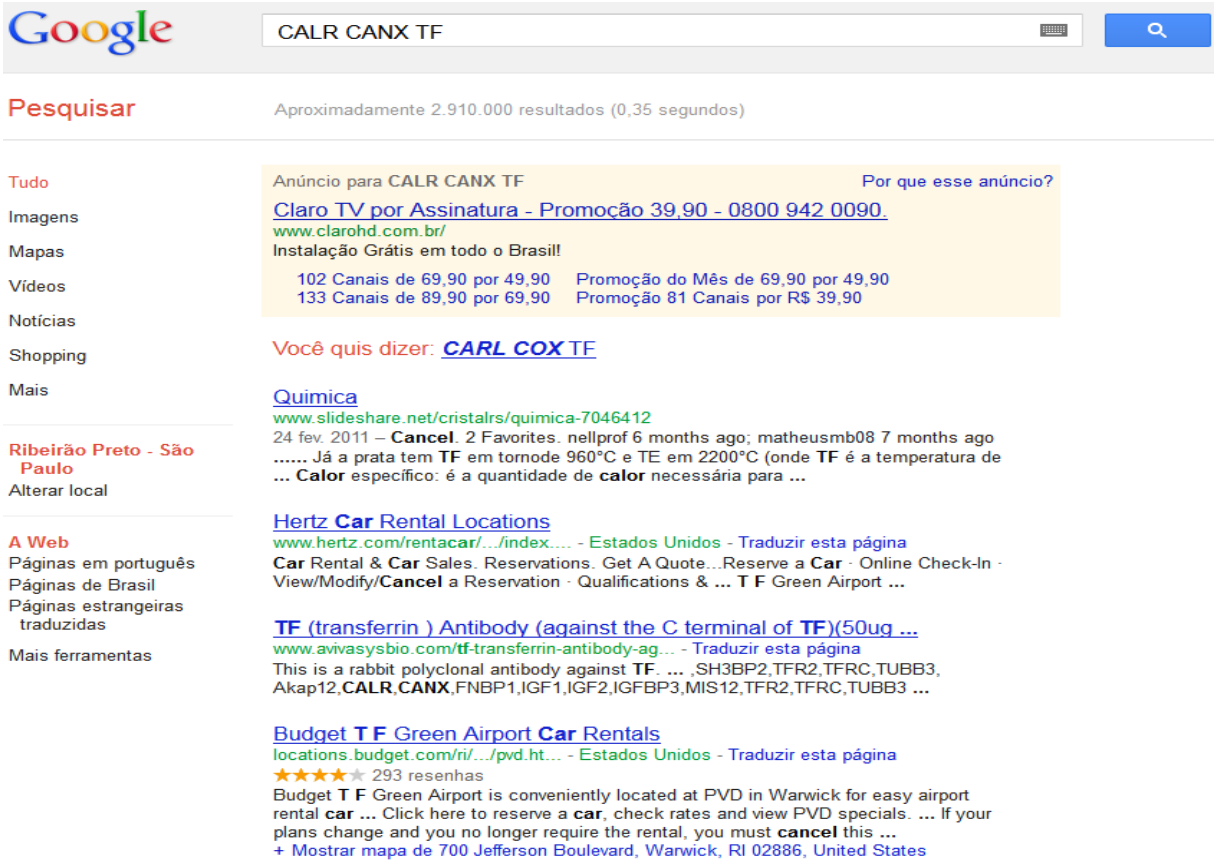


Figura 39: *Pesquisa no Google com símbolo oficial dos genes: CALR, CANX e TF*

Ao consultar o terceiro documento, encontra-se uma propaganda de uma empresa de bioinformática, que anuncia o anticorpo policlonal de coelho contra o *TF*. Somente no final da página, observa-se a seguinte informação: *CALR* e *CANX* são proteínas associadas ao *TF*, destacado no quadro da Figura 40(a). Apesar de não explicar as circunstâncias em que essas proteínas estão associadas, o resultado pode ser considerado relevante e informativo, mesmo não pertencendo a literatura científica.

Como o símbolo oficial dos genes pode ser uma sigla com outro significado em outros contextos, ferramentas não específicas como o Google[1] pode trazer sinomímia para esse tipo de pesquisa. Portanto, realizou-se a pesquisa utilizando o nome oficial dos genes: *calreticulin*, *calnexin* e *transferrin*. Esta busca recuperou 230 mil resultados, parte dos primeiros resultados da busca com os nomes oficiais é apresentado na Figura 41. Todos os resultados da primeira página apontaram para artigos da literatura biomédica, sendo que os três primeiros resultados são para a mesma publicação em endereços web diferentes. Ao retornar o mesmo artigo nas três primeiras posições, presume-se que este artigo realmente é relevante ao tratar dos genes *calreticulin*, *calnexin* e *transferrin*. Porém, seria mais interessante apresentar para os usuários artigos diferentes, que lhe proporcionassem uma pesquisa mais vasta.

[1] As buscas no Google foram realizadas com a personalização de histórico desativada.

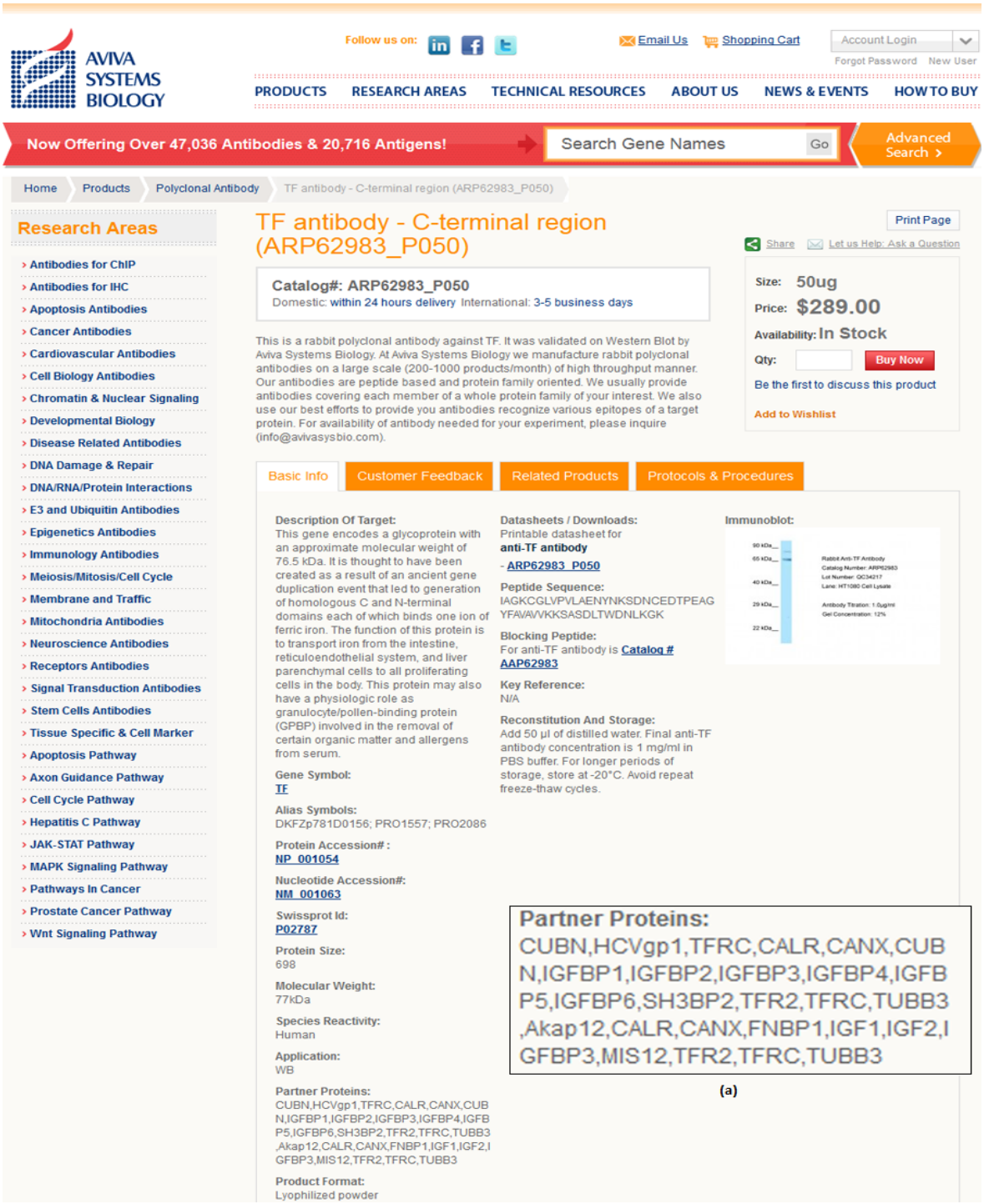


Figura 40: *Terceiro pesquisa no Google: propaganda de empresa de bioinformática*

O título do artigo três vezes retornado é *"Promotion of transferrin folding by cyclic interactions with calnexin and calreticulin"* e o PMID é *9312001*. Na tabela `interaction` do banco de dados modelado (ver Figura 19), buscou-se interações curadas pelo BioGRID cujo PMID fossem igual a *9312001*. A busca retornou dois resultados: interação entre *CALR* (identificador *811*) e *TF* (identificador *7018*) e interação entre *CANX* (identificador *821*) e *TF*, este resultado é apresentado na Figura 42.

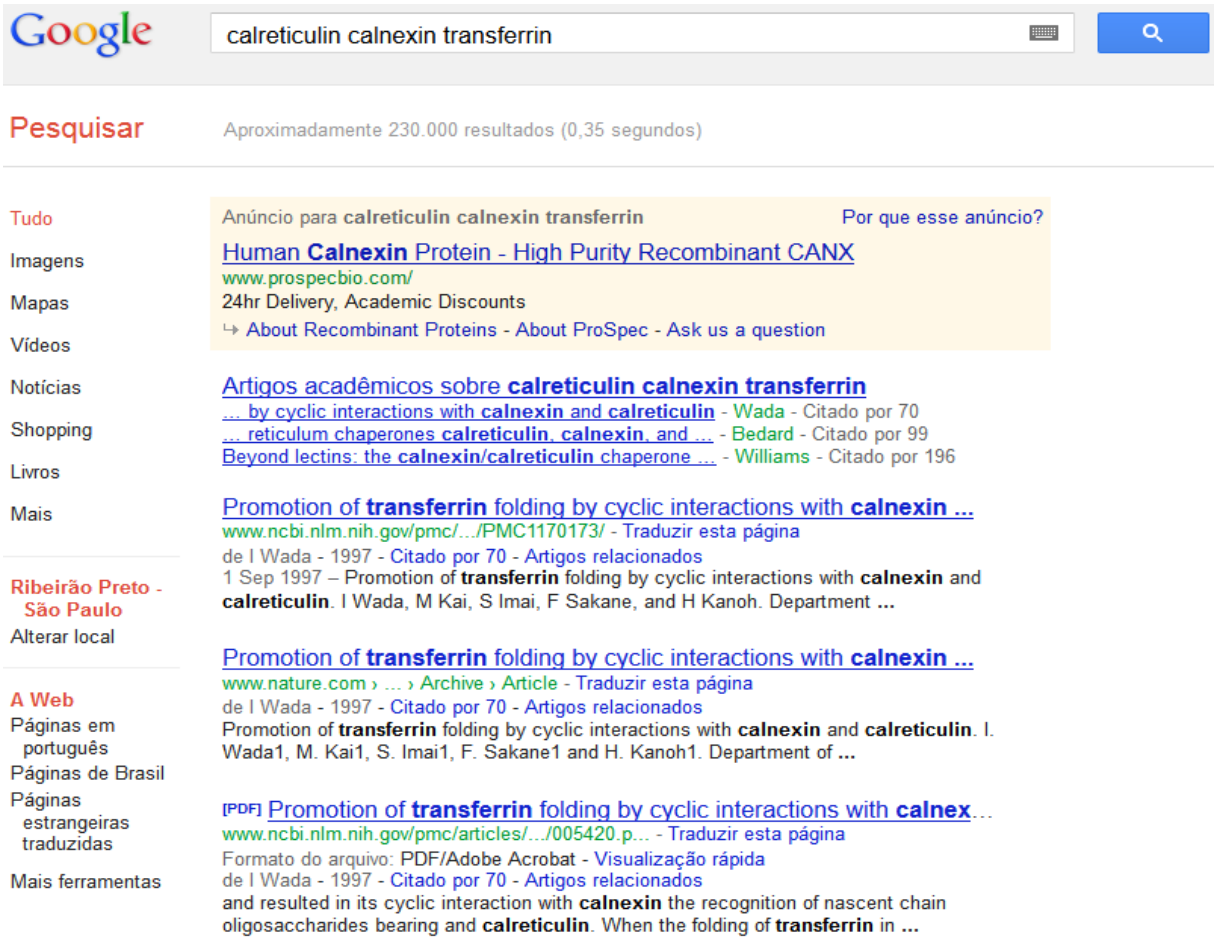
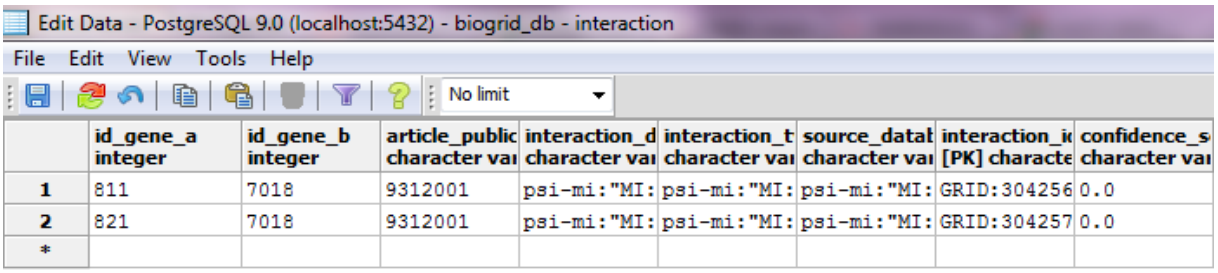


Figura 41: *Pesquisa no Google com nome oficial dos genes: calreticulin, calnexin e transferrin*



	id_gene_a integer	id_gene_b integer	article_public character var	interaction_d character var	interaction_t character var	source_datat character var	interaction_i [PK] characte	confidence_s character var
1	811	7018	9312001	psi-mi:"MI:	psi-mi:"MI:	psi-mi:"MI:	GRID:304256	0.0
2	821	7018	9312001	psi-mi:"MI:	psi-mi:"MI:	psi-mi:"MI:	GRID:304257	0.0
*								

Figura 42: *Consulta na tabela* `interaction` *com PMID igual a 9312001*

4.4.2 Google Acadêmico

O Google Acadêmico permite realizar buscas especificamente na literatura acadêmica e não em todo universo online, como o Google. É possível realizar pesquisas de diversas disciplinas e em fontes variadas, como, artigos, teses, livros, resumos, etc (GOOGLE, 2011). Na Figura 43, apresenta-se o resultado obtido ao pesquisar os genes *CALR*, *CANX* e*TF* por meio de seus símbolos oficiais no Google Acadêmico. Nos cinco primeiros resultados obtidos, ou seja, nos cinco documentos com maior similaridade à consulta apenas o terceiro parece apresentar informação relacionada à busca. Os demais resultados não pertencem a literatura biomédica.

Ao realizar a busca no Google Acadêmico com o nome oficial dos genes, obtém-se o resultado da Figura 44. O primeiro resultado é o mesmo artigo que retornou três vezes ao realizar a

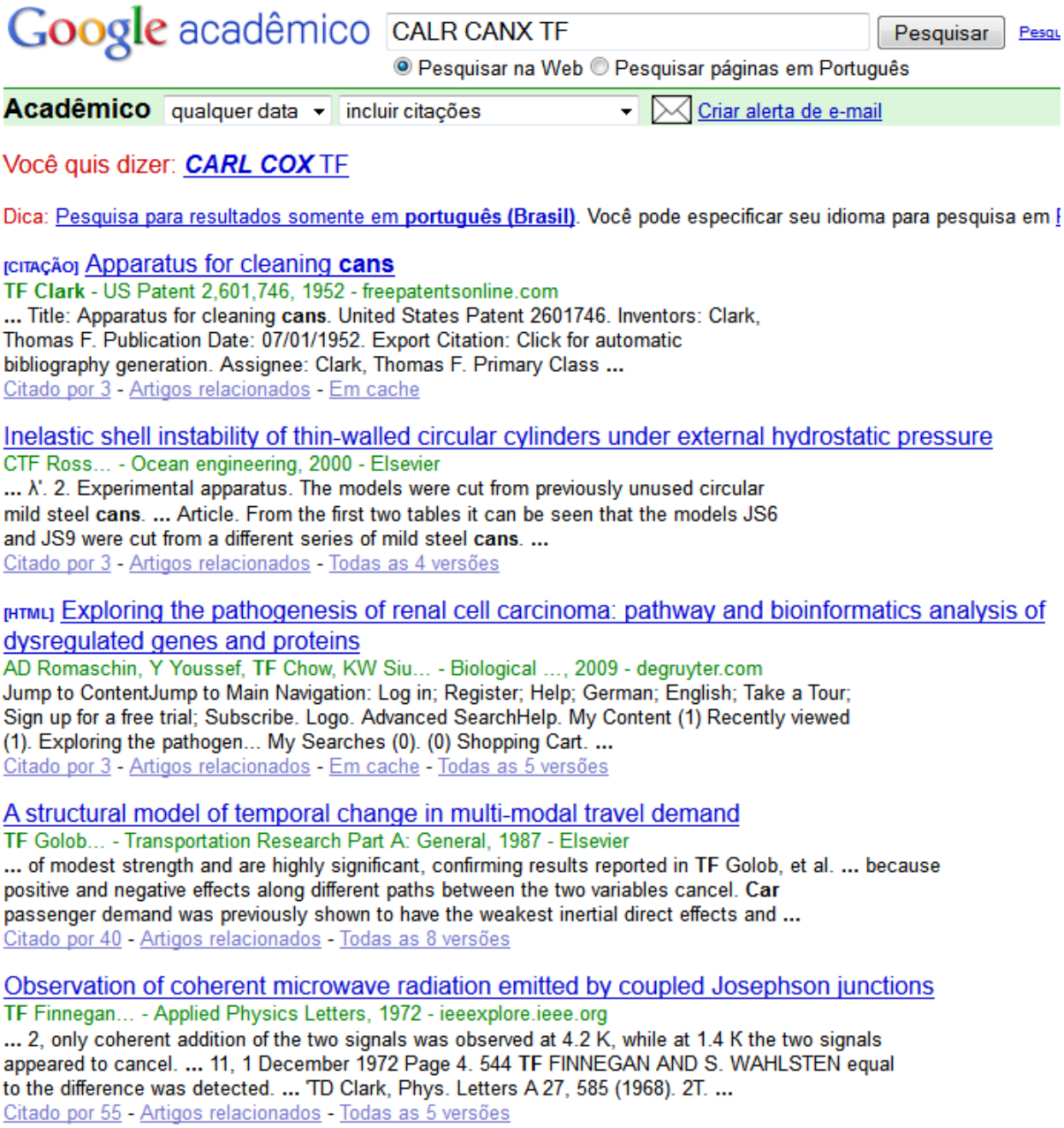


Figura 43: *Pesquisa no Google Acadêmico com símbolo oficial dos genes: CALR, CANX e TF*

busca no Google. Os demais resultados estão relacionados aos genes buscados e possuem potencial para informar sobre a interação entre estes genes. Em alguns dos resultados, antes selecionar o documento resultante pode-se ver que o texto relata a interação entre os genes.

4.4.3 PubMed

O PubMed oferece acesso a base de dados da *National Library of Medicine*, livros online e revistas de ciências da vida. Os usuários alvo do PubMed são pesquisadores, profissionais da saúde e público em geral, que precisem de informações da literatura biomédica (ver Seção 2.3.3) (LU, 2011). Apesar de conter mais de 21 milhões de citações para publicações científicas das ciências biológicas e médicas, ao pesquisar pelo símbolo de três genes com intuito de encontrar relações

Figura 44: *Pesquisa no Google Acadêmico com nome oficial dos genes: calreticulin, calnexin e transferrin*

gênicas na literatura não se obtém nenhum resultado (ver Figura 45). O PubMed sugere que se pesquise na base de dados *Gene* [2]. Já ao pesquisar com os nomes oficiais dos genes, a pesquisa retorna três resultados. A Figura 46 apresenta a pesquisa com os nomes oficiais de três genes no PubMed, em dois resultados os artigos completos podem ser acessados. O segundo resultado é o artigo *"Promotion of transferrin folding by cyclic interactions with calnexin and calreticulin"*, o qual retornou nos três resultados mais relevantes do Google e foi o resultado mais relevante no Google Acadêmico. Este artigo também foi curado pelo BioGRID (Figura 42), que afirma que o artigo relata as interações: *CALR - TF* e *CANX - TF*.

[2] A base de dados Gene integra informações de várias espécies. Os registros desta base podem conter as seguintes informações sobre um gene: nomenclatura, RefSeqs, mapas, vias, variações, fenótipos, variações, etc.

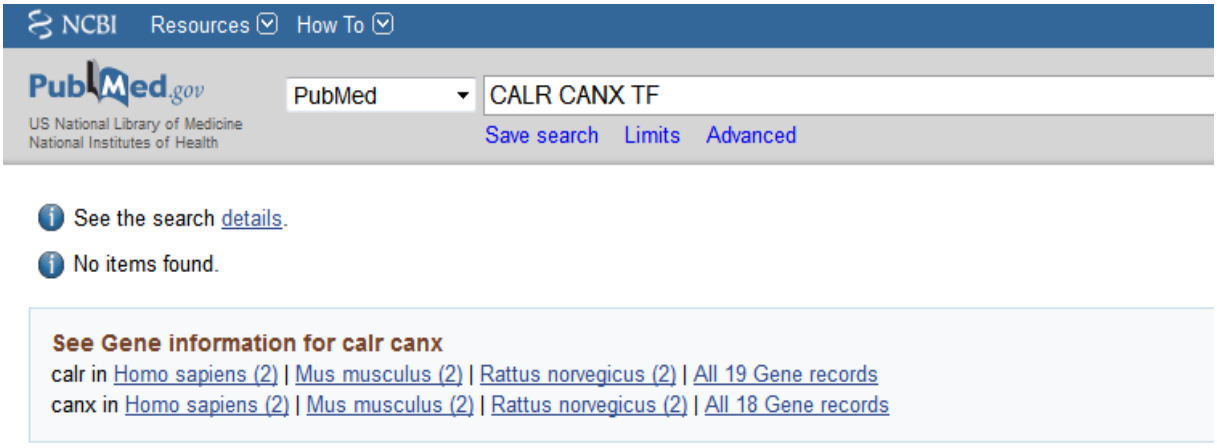


Figura 45: *Pesquisa no PubMed com símbolo oficial dos genes: CALR, CANX e TF*

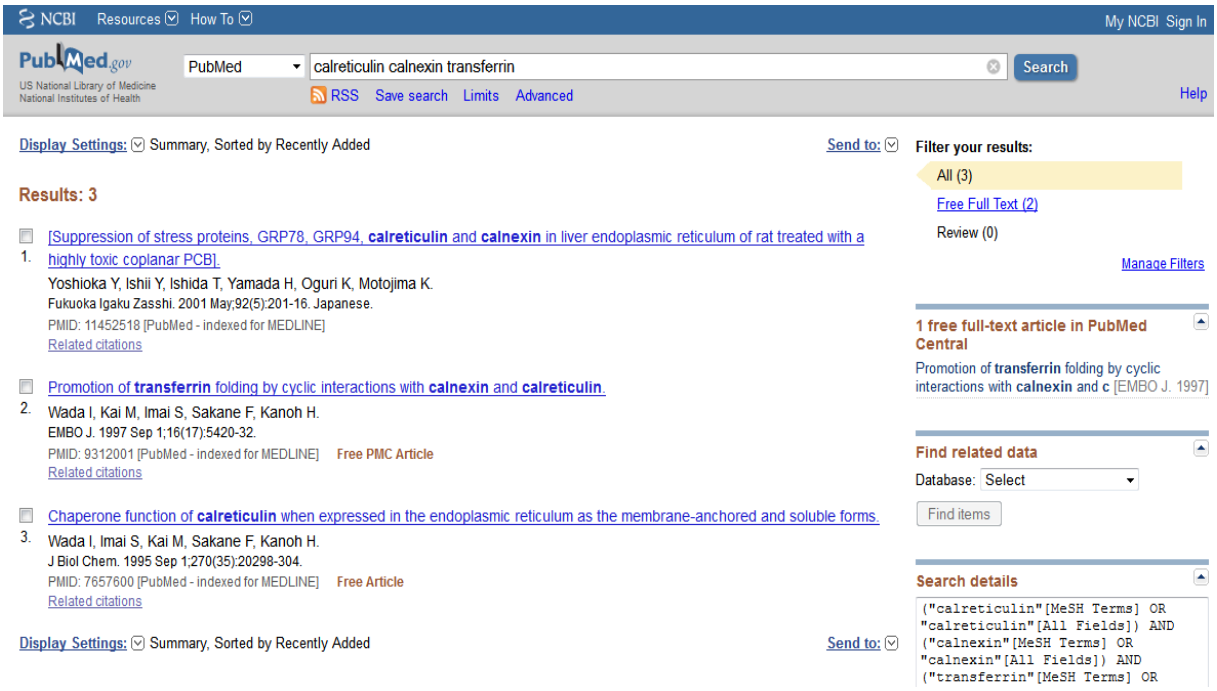


Figura 46: *Pesquisa no PubMed com nome oficial dos genes: calreticulin, calnexin e transferrin*

4.4.4 BioGRID

O BioGRID é um grande repositório online de interações entre genes ou entre seus produtos, todas interações foram detectadas por meio de curadoria. Esse repositório possui mais de 116 mil interações gênicas e proteicas (ver Seção 2.3.1). A interface do BioGRID, apresentada na Figura 47, possui uma aba para busca por genes e uma aba para busca por publicação.

Ao pesquisar pelo símbolo dos genes *CALR*, *CANX* e *TF* para Homo sapiens na aba *"by Gene"*, não se consegue nenhum resultado. Esta pesquisa é apresentada na Figura 48. O mesmo acontece ao pesquisar pelos nomes oficiais dos genes na aba *"by Gene"*, nenhum resultado é apresentado, como pode ser visto na Figura 49. Nesta aba, só é possível obter resultados ao buscar por somente um gene de cada vez, ou seja, ao combinar um grupo de genes na pesquisa sempre é inválida.

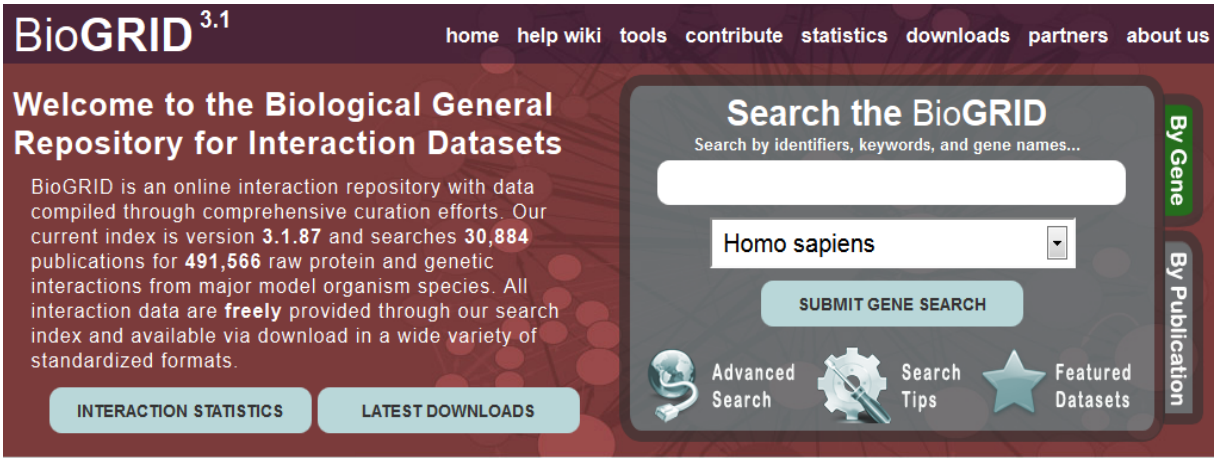


Figura 47: *Página inicial do BioGRID*

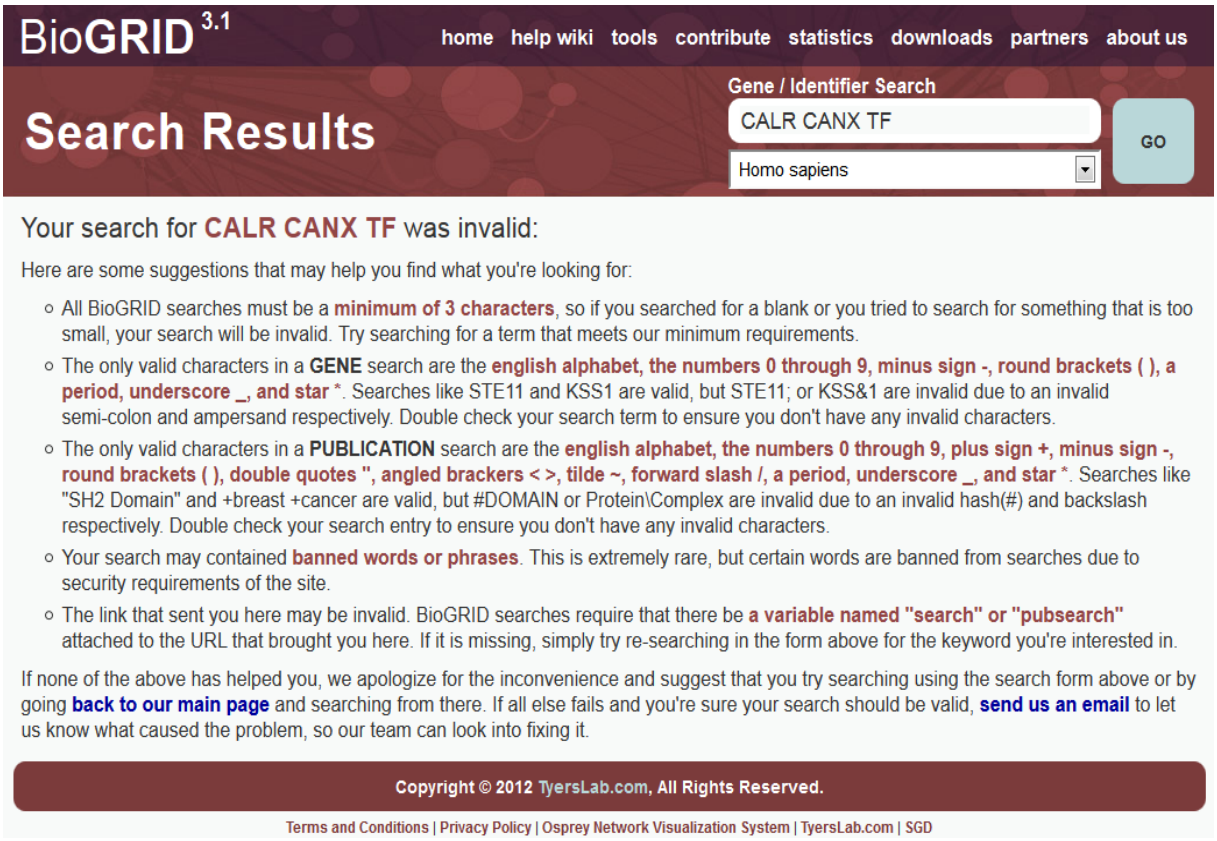


Figura 48: *Pesquisa no BioGRID na aba "by Gene" com símbolo oficial dos genes: CALR, CANX e TF*

Figura 49: *Pesquisa no BioGRID na aba "by Gene" com nome oficial dos genes: calreticulin, calnexin e transferrin*

Utilizando a aba *"by Publication"* para procurar os três símbolos dos genes *CALR*, *CANX* e *TF*, o resultado pesquisa é inválido assim como nas pesquisas anteriores. Já, ao procurar pelos

nomes oficiais nesta aba obtém-se 132 resultados, como mostrado na Figura 50. No entanto, estes resultados não estão separados por organismos.

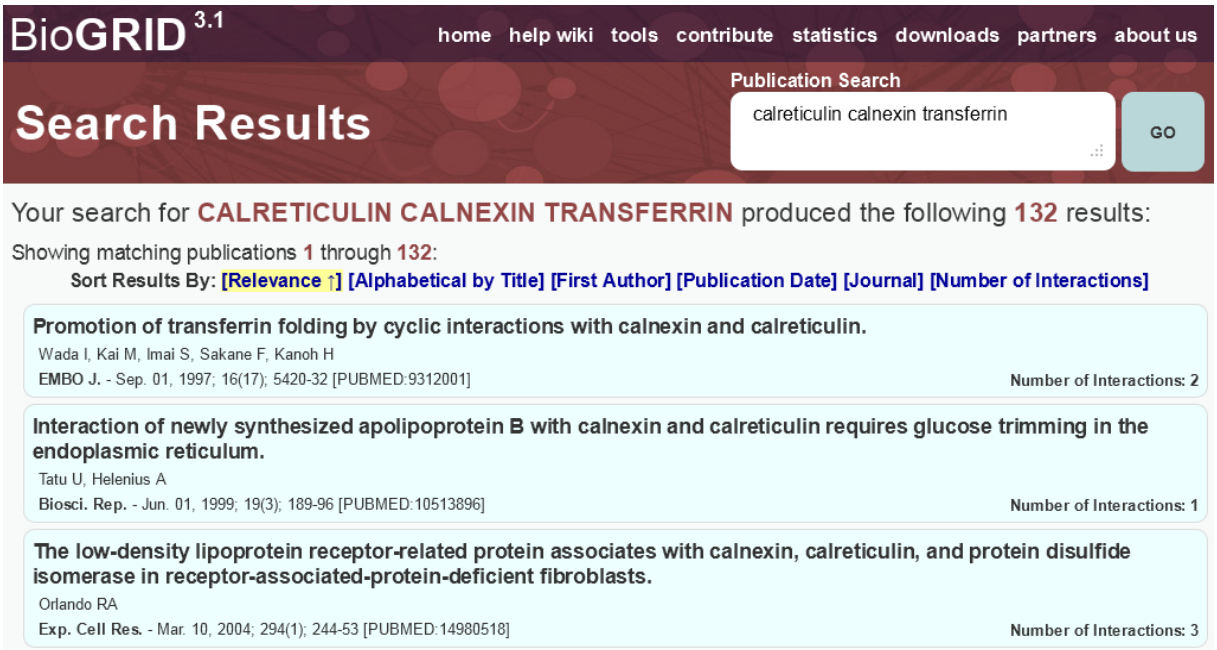


Figura 50: *Pesquisa no BioGRID na aba "by Publication" com nome oficial dos genes: calreticulin, calnexin e transferrin*

O BioGRID apresenta título, resumo e as interações detectadas ao entrar nos resultados obtidos. A Figura 51 mostra o resultado mais relevante da pesquisa na aba *"by Publication"* com nome oficial dos genes *calreticulin*, *calnexin* e *transferrin*.

4.4.5 SARI

Utilizando a implementação do SARI com múltiplos documentos (ver Seção 3.3) pesquisou-se pelos genes *CALR*, *CANX* e *TF*. Na implementação do SARI, não importa se a pesquisa foi feita por meio dos símbolos ou nomes oficiais dos genes, o mesmo resultado é obtido. O resultado da pesquisa é apresentado na Figura 52.

4.4.6 Comparação entre as ferramentas de busca

A Tabela 6 apresenta um quadro comparativo dos resultados das buscas para símbolos oficiais e nomes oficiais dos genes *CALR*, *CANX* e *TF* nas ferramentas: Google, Google Acadêmico, PubMed, BioGRID e SARI. Dos três primeiros resultados analisados como retorno para a ferramenta Google ao buscar com o símbolo oficial dos genes tem-se um resultado relevante. Quando buscou-se com o símbolo oficial dos genes no Google, os três primeiros resultados foram relevantes, mas esses três resultados são o mesmo artigo científico armazenado em endereços web distintos. Ao utilizar o símbolo oficial dos genes no Google Acadêmico, entre os três primeiros resultados um mostrou-se relevante. Já ao utilizar o nome oficial, os três primeiros resultados obtidos no Google

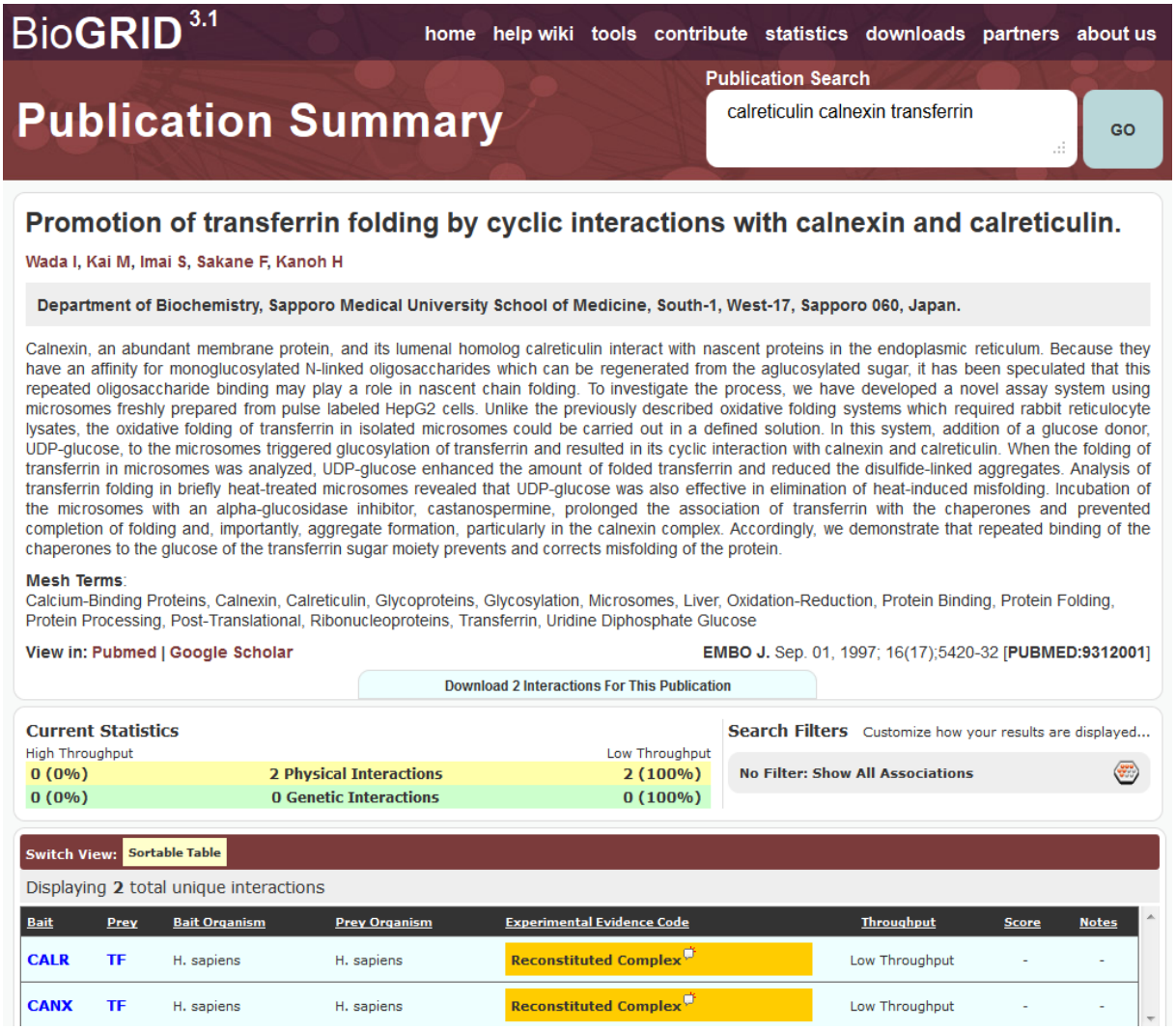


Figura 51: *Resultado mais relevante na busca dos genes calreticulin, calnexin e transferrin*

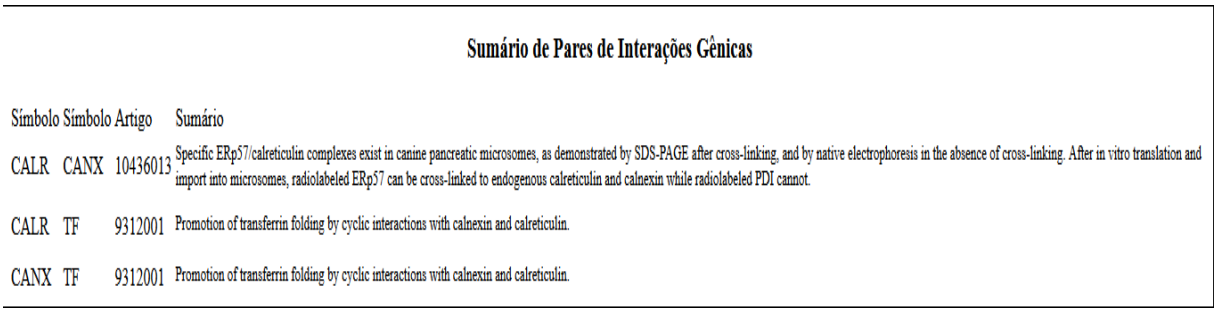


Figura 52: *Resultado da pesquisa no SARI dos genes CALR, CANX e TF*

Acadêmico mostraram-se relevantes. No PubMed, ao buscar com símbolo oficial dos genes não foi obtido nenhum resultado e ao buscar pelo nome oficial dos genes, a pesquisa retornou três resultados relevantes. Ao buscar tanto por símbolo quanto por nome não obteve-se nenhum resultado no BioGRID na aba *by Gene*. Quando buscou-se o símbolo oficial dos três genes no BioGRID na aba *by Publication* também não houve nenhum resultado. Mas ao buscar pelos nomes oficiais os três primeiros resultados foram relevantes. No SARI, os três primeiros resultados foram relevantes utilizando símbolo ou nome oficial.

A escolha da ferramenta de busca mais adequada depende da informação de interesse do usá-

Tabela 6: *Comparação entre as ferramentas de busca*

Consulta / Ferramenta	símbolo oficial	nome oficial
Google	1	3
Google Acadêmico	1	3
PubMed	0	3
BioGRID *by Gene*	0	0
BioGRID *by Publication*	0	3
SARI	3	3

rio e de suas preferências pessoais. Todas as ferramentas possuem vantagens e desvantagens. O Google é uma ferramenta popular e fácil de usar, mas os resultados obtidos dependem do cuidado com a escolha dos termos da busca. É comum obter resultados totalmente inesperados ao fazer buscas no Google. Já o Google Acadêmico agrega a vantagem das buscas serem feitas somente na literatura científica. Assim como o Google Acadêmico, o PubMed somente faz buscas na literatura científica. Porém, o PubMed é específico para literatura biológica e médica. O PubMed pode oferecer resultados mais profundos, contudo requer mais esforço do usuário, por exemplo, treinamento nas buscas avançadas pode ajudar aos usuários obterem melhores resultados (STEINBROOK, 2006).

A busca no BioGRID tem o foco desejado desta monografia, isto é, buscar interações entre os genes ou entre seus produtos. Porém, o BioGRID não trata os termos da busca, ou seja, não faz um mapeamento dos símbolos ou dos alias dos genes, somente reconhece quando o nome oficial é utilizado na busca. Uma característica em comum as ferramentas de busca exploradas é retornar uma quantidade inviável de resultados a serem analisados por completo. Como esperado, o método SARI, por tratar especificamente de interações gênica e manipular os termos da consulta cuidadosamente, obteve melhores resultados. Além disso, no método SARI buscou-se sumarizar os textos diminuindo a sobrecarga cognitiva para leitura da informação que retorna das buscas.

5 Conclusão

Nas últimas décadas, as ciências biológicas passaram por grandes mudanças. O desenvolvimento de técnicas para análise de proteínas, DNA e RNA têm permitido o estudo destas macromoléculas de uma forma antes não imaginada pelos pesquisadores. Por exemplo, técnicas de microarrays permitem monitorar a expressão gênica de mais de trinta mil genes simultaneamente em somente um experimento. Logo, a quantidade de dados gerados por estas novas técnicas são enormes e tornou-se imprescíndivel a utilização de métodos computacionais para analisar os dados biológicos. Algoritmos de clusterização são um clássico na análise de dados de expressão gênica, pois possibilitam a detecção de interações e co-regulação gênica. Contudo, apenas obter e analisar dados não é suficiente. Para produzir conhecimento e inferir novos alvos de pesquisa, é necessário fazer referências cruzadas dos resultados obtidos com resultados já estabelecidos na literatura científica. No entanto, a literatura disponível é muito vasta e encontrar uma informação de interesse pode se tornar uma tarefa demorada e cansativa. Assim, no contexto de interações gênicas e proteicas, buscar por artigos curados e utilizar sumarização para consultá-los pode ser uma boa estratégia.

Nessa monografia, propôs-se o método SARI para dar significado a genes agrupados, por algoritmos de clusterização ou por algum outro método de agrupamento. Utilizando a literatura científica curada pelo BioGRID e disponível no PubMed, foram feitos sumários com objetivo de produzir referências cruzadas entre resultados de análise de expressão gênica com a literatura. O método SARI foi desenvolvido para ilustrar uma sequência de etapas desde a obtenção dos dados até os sumários e rede de interação gênica e proteica resultante.

Os resultados da avaliação de utilidade dos sumários mostraram que os julgadores, que são possíveis usuários, consideram mais útil os sumários que realmente tratam da interação gênica. Os julgadores não deram muita importância ao tamanho do sumário ou se ele foi produzido a partir de somente um ou vários documentos fonte.

Uma possível contribuição do trabalho aqui apresentado é o auxílio a validação e a atribuição de significado aos clusters gerados a partir de dados de expressão gênica. Para isso considera-se a premissa que quando a literatura científica indica relacionamento entre genes de um cluster, pode-se inferir que o cluster não foi formado por aleatoriedade e que o algoritmo está classificando de acordo com o esperado. Quando um cluster aponta relacionamento entre genes que nunca foram citados na literatura, isso pode indicar um novo foco de estudo ou um indicativo de problemas no

algoritmo de clusterização.

O desenvolvimento de um projeto de interfaces e análise de usabilidade de software pode ser um trabalho futuro aplicado a instanciação do método SARI, principalmente para a instanciação web. Em relação à eficiência computacional, a utilização de computação paralela também seria interessante quando um grupo de genes muito grande precisa ser analisado. Um conjunto grande de genes pode tornar a resposta da aplicação SARI demorada. Outra solução possível para conjuntos grandes de genes é a codificação assíncrona do serviço web.

Trabalhos futuros também podem envolver adaptações e novos métodos de sumarização. Por exemplo, a utilização de outros termos além da nomenclatura gênica como palavras-chave. Palavras indicadoras interação como “bind”, “interaction”, “co-regulation” podem ser incluídas na lista de palavras-chave. Um trabalho de iniciação científica para identificar esses termos foi realizado no grupo de pesquisa, o qual pertencem autor e orientador do presente trabalho. Portanto, a colaboração entre as duas pesquisas pode obter resultados interessantes na adaptação dos métodos de sumarização.

Um outro trabalho de mestrado desenvolvido pelo grupo de pesquisa, envolve a apresentação de artigos científicos para médicos que necessitam diagnosticar fatores de risco com causas epigenéticas de pacientes. Neste contexto, a sumarização para apresentação dos artigos para o usuário, no caso os médicos, também pode ser útil. Assim, pode-se auxiliar na atualização do conhecimento do profissional e auxiliar o diagnóstico de uma forma mais rápida.

6 *Referências Bibliográficas*

AFANTENOS, S.; KARKALETSIS, V.; STAMATOPOULOS, P. Summarization from medical documents: a survey. *Artif. Intell. Med.*, Elsevier Science Publishers Ltd., Essex, UK, v. 33, p. 157–177, February 2005. ISSN 0933-3657. Disponível em: <http://portal.acm.org/citation-.cfm?id=1644738.1644824>.

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. *Biologia molecular da célula*. 4ª edição. ed. Porto Alegre: Artmed, 2004.

ANTIQUEIRA, L. *Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos*. 124 p. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação (ICMC), São Carlos, February 2007.

APACHE, S. F. *Apache Axis2/Java*. April 2012. Disponível em: <http://axis.apache.org/axis2-/java/core/index.html>.

APACHE, S. F. *Apache Tomcat*. April 2012. Disponível em: <http://tomcat.apache.org/index-.html>.

BABU, M. M. Computational genomics. In: ____. [S.l.]: Horizon press, 2004. cap. Chapter 11 An Introduction to Microarray Data Analysis.

BARNETT, J.; KNIGHT, K.; MANI, I.; RICH, E. Knowledge and natural language processing. *Commun. ACM*, ACM, New York, NY, USA, v. 33, p. 50–71, August 1990. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/79173.79177>.

BARRETT, T.; TROUP, D.; WILHITE, S.; LEDOUX, P.; EVANGELISTA, C.; KIM, I.; TOMASHEVSKY, M.; MARSHALL, K.; PHILLIPPY, K.; SHERMAN, P.; MUERTTER, R.; HOLKO, M.; AYANBULE, O.; YEFANOV, A.; SOBOLEVA, A. Ncbi geo: archive for functional genomics data sets-10 years on. *Nucleic Acids Research*, v. 39, n. suppl 1, p. D1005–D1010, 2011. Disponível em: <http://nar.oxfordjournals.org/content/39/suppl_1/D1005.abstract>.

BOLSHAKOVA, N.; AZUAJE, F. Cluster validation techniques for genome expression data. *Signal Process.*, Elsevier North-Holland, Inc., Amsterdam, The Netherlands, The Netherlands, v. 83, p. 825–833, April 2003. ISSN 0165-1684. Disponível em: <http://dx.doi.org/10.1016-/S0165-1684(02)00475-9>.

CHERRY, J. M.; BALL, C.; WENG, S.; JUVIK, G.; SCHMIDT, R.; ADLER, C.; DUNN, B.; DWIGHT, S.; RILES, L.; MORTIMER, R. K.; BOTSTEIN, D. Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, Department of Genetics, Stanford University School of Medicine, California 94305-5120, USA, v. 387, n. 6632 Suppl, p. 67–73, maio 1997. ISSN 0028-0836. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3057085/>.

D'AVANZO, E.; MAGNINI, B.; VALLIN, A. Keyphrase extraction for summarization purposes: The LAKE system at DUC-2004. In: *DUC2004*. [S.l.: s.n.], 2004.

D'HAESELEER, P. How does gene expression clustering work? *Nature Biotechnology*, Nature Publishing Group, v. 23, n. 12, p. 1499–1501, dez. 2005. ISSN 1087-0156. Disponível em: <http://dx.doi.org/10.1038/nbt1205-1499>.

DONALDSON, I.; MARTIN, J.; BRUIJN, B. de; WOLTING, C.; LAY, V.; TUEKAM, B.; ZHANG, S.; BASKIN, B.; BADER, G.; MICHALICKOVA, K.; PAWSON, T.; HOGUE, C. Prebind and textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, v. 4, n. 1, p. 11, 2003. ISSN 1471-2105. Disponível em: <http://www.biomedcentral.com/1471-2105/4/11>.

DUNHAM, M. H. *Data Mining: Introductory and Advanced Topics*. 1. ed. [S.l.]: Prentice Hall, 2002. Paperback. ISBN 0130888923.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, v. 30, n. 1, p. 207–210, 2002. Disponível em: <http://nar.oxfordjournals.org/content/30/1/207.abstract>.

FELLBAUM, C. (Ed.). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998. ISBN 978-0-262-06197-1.

FRIENDLY, M. The history of the cluster heat map. *The American Statistician*, 2009.

GOLDSTEIN, J.; MITTAL, V.; CARBONELL, J.; KANTROWITZ, M. Multi-document summarization by sentence extraction. In: *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization - Volume 4*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. (NAACL-ANLP-AutoSum '00), p. 40–48. Disponível em: <http://dx.doi.org-/10.3115/1117575.1117580>.

GOOGLE. *Sobre o Google Acadêmico*. 2011. Disponível em: <http://scholar.google.com.br/intl-/pt-BR/scholar/about.html>.

GUDGIN, M.; HADLEY, M.; MENDELSOHN, N.; LAFON, Y.; MOREAU, J.-J.; KARMARKAR, A.; NIELSEN, H. F. *SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)*. [S.l.], abr. 2007. Http://www.w3.org/TR/2007/REC-soap12-part1-20070427/.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 11, p. 10–18, November 2009. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/1656274.1656278>.

HAN, J.; KAMBER, M. *Data mining: Concepts and Techniques*. [S.l.]: Morgan Kaufmann Publishers, 2006. (The Morgan Kaufmann series in data management systems).

HARTWELL, L. H.; HOPFIELD, J. J.; LEIBLER, S.; MURRAY, A. W. From molecular to modular cell biology. *Nature*, Nature Publishing Group, Fred Hutchinson Cancer Center, Seattle, Washington 98109, USA., v. 402, n. 6761 Suppl, p. C47–C52, dez. 1999. ISSN 0028-0836. Disponível em: <http://dx.doi.org/10.1038/35011540>.

HERMJAKOB, H.; MONTECCHI-PALAZZI, L.; BADER, G.; WOJCIK, J.; SALWINSKI, L.; CEOL, A.; MOORE, S.; ORCHARD, S.; SARKANS, U.; MERING, C. V.; ROECHERT, B.; POUX, S.; JUNG, E.; MERSCH, H.; KERSEY, P.; LAPPE, M.; LI, Y.; ZENG, R.; RANA, D.; NIKOLSKI, M.; HUSI, H.; BRUN, C.; SHANKER, K.; GRANT, S. G. N.; SANDER, C.; BORK, P.; ZHU, W.; PANDEY, A.; BRAZMA, A.; JACQ, B.; VIDAL, M.; SHERMAN, D.; LEGRAIN, P.; CESARENI, G.; XENARIOS, I.; EISENBERG, D.; STEIPE, B.; HOGUE, C.; APWEILER, R. The hupo psi's molecular interaction format - a community standard for the representation of protein interaction data. *Nature biotechnology*, v. 22, n. 2, p. 177–183, 2004. Cited By (since 1996): 248. Disponível em: <www.scopus.com>.

HOVY, E.; LIN, C.-Y. Automated text summarization and the summarist system. In: *Proceedings of a workshop on held at Baltimore, Maryland*. Morristown, NJ, USA: Association for Computational Linguistics, 1996. p. 197–214.

HU, X. Ge-miner: integration of cluster ensemble and text mining for comprehensive gene expression analysis. *Int. J. Bioinformatics Res. Appl.*, Inderscience Publishers, Inderscience Publishers, Geneva, SWITZERLAND, v. 2, n. 3, p. 325–338, 2006. ISSN 1744-5485.

III, H. D. Book review, inderjeet mani: Automatic summarization, john benjamins publishing, amsterdam, the netherlands, 2001, xi + 286 pp. *Machine Translation*, v. 18, n. 4, p. 343–347, 2004.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. ISBN 0-13-022278-X.

JAIN, E.; BAIROCH, A.; DUVAUD, S.; PHAN, I.; REDASCHI, N.; SUZEK, B.; MARTIN, M.; MCGARVEY, P.; GASTEIGER, E. Infrastructure for the life sciences: design and implementation of the uniprot website. *BMC Bioinformatics*, v. 10, n. 1, p. 136, 2009. ISSN 1471-2105. Disponível em: <http://www.biomedcentral.com/1471-2105/10/136>.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, v. 28, p. 11–21, 1972.

KANKAR, P.; ADAK, S.; SARKAR, A.; MURARI, K.; SHARMA, G.; EXPRESSION, G. Medmesh summarizer: text mining for gene clusters. In: *in the Proceedings of the Second SIAM International Conference on Data Mining*. [S.l.: s.n.], 2002.

KLUG, W.; CUMMINGS, M.; PALLADINO, M.; SPENCER, C. *Conceitos de Genética*. 9. ed. [S.l.]: Artmed, 2010. ISBN 978-85-363-2115-8.

KOSCHMIEDER, A.; ZIMMERMANN, K.; TRIßL, S.; STOLTMANN, T.; LESER, U. Tools for managing and analyzing microarray data. *Briefings in Bioinformatics*, 2011. Disponível em: <http://bib.oxfordjournals.org/content/early/2011/03/20/bib.bbr010.abstract>.

LEVINE, J. H. *Introduction to data analysis: rules of evidence volume I: well-behaved variables*. Dartmouth: Dartmouth College, 1996.

LEWIN, B. *Genes IX*. [S.l.]: Artmed, 2009. ISBN 978-85-363-1754-0.

LU, Z. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, v. 2011, 2011. Disponível em: <http://database.oxfordjournals.org/content/2011/baq036-.abstract>.

LUHN, H. The automatic creation of literature abstracts. *IBM Journal*, v. 2, p. 159–165, 1958. Disponível em: <http://www.db.dk/bh/core%20concepts%20in%20lis/articles%20a-z/luhn-.htm>.

MANI, I. *Automatic Summarization*. [S.l.]: John Benjamins Publishing Company, 2001. (Natural Language Processing).

MARCU, D. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA, USA: MIT Press, 2000. ISBN 0262133725.

MARTINS, C. B.; PARDO, T. A. S.; ESPINA, A. P.; RINO, L. H. M. *Introdução À Sumarização Automática*. [S.l.], 2001.

MOELLERING, R. E.; CORNEJO, M.; DAVIS, T. N.; BIANCO, C. D.; ASTER, J. C.; BLACKLOW, S. C.; KUNG, A. L.; GILLILAND, D. G.; VERDINE, G. L.; BRADNER, J. E. Direct inhibition of the notch transcription factor complex. *Nature*, v. 462, n. 7270, p. 182–8, 2009. ISSN 1476-4687. Disponível em: <http://www.biomedsearch.com/nih/Direct-inhibition-NOTCH-transcription-factor/19907488.html>.

MOHAN, M. B. Computational genomics: Theory and application. In: ____. [S.l.]: Horizon Scientific Press, Norwich, UK, 2004. cap. An introduction to microarray data analysis, p. 225–249.

MONARD, M. C.; BARANAUSKAS, J. A. Sistemas inteligentes: Fundamentos e aplicações. In: ____. [S.l.]: Editora Manole Ltda, 2003. cap. Conceitos sobre aprendizado de máquina.

MORGAN, A. A.; HIRSCHMAN, L.; COLOSIMO, M.; YEH, A. S.; COLOMBE, J. B. Gene name identification and normalization using a model organism database. *J. of Biomedical Informatics*, Elsevier Science, San Diego, USA, v. 37, n. 6, p. 396–410, 2004. ISSN 1532-0464.

MOTSCHALL, E.; FALCK-YTTER, Y. Searching the medline literature database through pubmed: a short guide. *Onkologie*, v. 28, n. 10, p. 517–522, 2005. Disponível em: <http://www-.ncbi.nlm.nih.gov/pubmed/16186693>.

NCBI. *Entrez Help*. [S.l.], 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK3837-/>.

NCBI. *Entrez Programming Utilities Help*. [S.l.], 2010. Disponível em: <http://www.ncbi.nlm-.nih.gov/books/NBK25497/>.

NCBI. *PubMed*. junho 2011. Disponível em: <"http://www.ncbi.nlm.nih.gov/pubmed/">.

NENKOVA, A.; VANDERWENDE, L. *The Impact of Frequency on Summarization*. [S.l.], January 2005.

NLM. *Medical Subjects Headings (MeSH)*. janeiro 2011. Disponível em: <http://www.nlm.nih-.gov/pubs/factsheets/mesh.html>.

NLM. *MEDLINE*. dezembro 2011. Disponível em: <http://www.nlm.nih.gov/pubs/factsheets-/medline.html>.

NLM. *MEDLINE Citation Counts by Year of Publication*. maio 2011. Disponível em: <http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html>.

NUSSBAUM, R. L.; MCINNES, R. R.; WILLARD, H. F. *THOMPSON & THOMPSON GENÉTICA MÉDICA*. 7. ed. Elsevier Health Sciences, 2008. ISBN 9788535245752. Disponível em: <http://books.google.com.br/books?id=vjOET7ul_R0C>.

PARDO, T.; RINO, L.; NUNES, M. Extractive summarization: how to identify the gist of a text. In: CITESEER. *the Proceedings of the 1st International Information Technology Symposium–I2TS*. [S.l.], 2002. p. 1–6.

PARKINSON, H.; SARKANS, U.; KOLESNIKOV, N.; ABEYGUNAWARDENA, N.; BURDETT, T.; DYLAG, M.; EMAM, I.; FARNE, A.; HASTINGS, E.; HOLLOWAY, E.; KURBATOVA, N.; LUKK, M.; MALONE, J.; MANI, R.; PILICHEVA, E.; RUSTICI, G.; SHARMA, A.; WILLIAMS, E.; ADAMUSIAK, T.; BRANDIZI, M.; SKLYAR, N.; BRAZMA, A. Arrayexpress update-an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Research*, v. 39, n. suppl 1, p. D1002–D1004, 2011. Disponível em: <http://nar.oxfordjournals.org/content/39/suppl_1/D1002.abstract>.

REEVE, L. H.; HAN, H.; BROOKS, A. D. The use of domain-specific concepts in biomedical text summarization. *Information Processing and Management: an International Journal*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 43, n. 6, p. 1765–1776, 2007. ISSN 0306-4573.

SEAL, R. L.; GORDON, S. M.; LUSH, M. J.; WRIGHT, M. W.; BRUFORD, E. A. genenames.org: the hgnc resources in 2011. *Nucleic Acids Research*, v. 39, n. suppl 1, p. D514–D519, 2011. Disponível em: <http://nar.oxfordjournals.org/content/39/suppl_1/D514.abstract>.

SPARCK-JONES, K. Automatic summarizing: factors and directions. In: MANI, I.; MAYBURY, M. T. (Ed.). *Advances in automatic text summarization*. [S.l.]: The MIT Press, 1999. cap. 1, p. 1 – 12.

SPLENDORE, A. Para que existem as regras de nomenclatura genética? *Revista Brasileira de Hematologia e Hemoterapia*, scielo, v. 27, p. 148–152, 06 2005. ISSN 1516-8484.

STARK, C.; BREITKREUTZ, B.-J.; REGULY, T.; BOUCHER, L.; BREITKREUTZ, A.; TYERS, M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, v. 34, n. suppl 1, p. D535–D539, 2005.

STEINBROOK, R. Searching for the right search - reaching the medical literature. *New England Journal of Medicine*, v. 354, n. 1, p. 4–7, 2006. Disponível em: <http://www.nejm.org/doi/full-/10.1056/NEJMp058128>.

STRACHAN, T.; READ, A. P. *Human Molecular Genetics*. 2. ed. [S.l.]: Garland Science, 1999.

TWEEDIE, S.; ASHBURNER, M.; FALLS, K.; LEYLAND, P.; MCQUILTON, P.; MARYGOLD, S.; MILLBURN, G.; OSUMI-SUTHERLAND, D.; SCHROEDER, A.; SEAL, R.; ZHANG, H.; CONSORTIUM, T. F. Flybase: enhancing drosophila gene ontology annotations. *Nucleic Acids Research*, v. 37, n. suppl 1, p. D555–D559, 2009. Disponível em: <http://nar.oxfordjournals.org-/content/37/suppl_1/D555.abstract>.

WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, Am Med Assoc, v. 171, n. 4356, p. 737–738, 1953. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17804965>.

YEUNG, K. Y.; HAYNOR, D. R.; RUZZO, W. L. Validating clustering for gene expression data. *Bioinformatics*, v. 17, n. 4, p. 309–318, 2001. Disponível em: <http://bioinformatics-.oxfordjournals.org/content/17/4/309.abstract>.

APÊNDICE A – Questionário de Usefulness

Usefulness

Imagine que deseja pesquisar na literatura se há interações, físicas ou genéticas, entre alguns genes. Para economizar tempo, ao invés de ler um conjunto de documentos completos em busca de alguns documentos que lhe interesse. Você pode primeiro ler uma lista de resumos curtos desses documentos e nesses resumos encontrar a informação que buscava ou escolher quais documentos serão lidos na íntegra.

É possível criar vários resumos de um determinado documento e alguns resumos podem ser mais útil do que outros (por exemplo, lhe dizer mais sobre o conteúdo relevante para o assunto, mais fácil de ler, etc). Aqui apresentamos 3 resumos diferentes para cada interação gênica e deseja-se saber qual foi mais útil ao informar sobre esta interação. Sua tarefa é nos ajudar a entender quão útil podem ser esses resumos para suas pesquisas na literatura científica.

Por favor, leia todos os seguintes resumos que lhe foram dados. Suponha que você quer buscar informações sobre as interações entre genes x e y . Classifique cada resumo de acordo a sua utilidade: Muito Ruim (não tem utilidade alguma), Ruim, Apenas aceitável, Bom, Muito Bom (tão bom quanto ler o documento na íntegra, esclareceu a relação entre os dois genes). Não se aborreça pois existem interações repetidas com resumos diferentes.
*Obrigatório

Nome *

Link Currículo Lattes
(para caracterizarmos o perfil dos avaliadores)

Interação entre genes: MLL - MEN1 *
Leukemia proto-oncoprotein MLL forms a SET1-like histone methyltransferase complex with menin to regulate Hox gene expression.Two other members of the novel MLL complex identified here are host cell factor 1 HCF-1, a transcriptional coregulator, and the related HCF-2, both of which specifically interact with a conserved binding motif in the MLLN p300 subunit of MLL and provide a potential mechanism for regulating its antagonistic transcriptional properties. Menin, a product of the MEN1 tumor suppressor gene, is also a component of the 1-MDa MLL complex.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Figura 53: *Questionário de utilidade - página 1*

Interação entre genes: MLL - MEN1 *

Two other members of the novel MLL complex identified here are host cell factor 1 HCF-1, a transcriptional coregulator, and the related HCF-2, both of which specifically interact with a conserved binding motif in the MLLN p300 subunit of MLL and provide a potential mechanism for regulating its antagonistic transcriptional properties.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: MLL - MEN1 *

Leukemia proto-oncoprotein MLL forms a SET1-like histone methyltransferase complex with menin to regulate Hox gene expression. MLL for mixed-lineage leukemia is a proto-oncogene that is mutated in a variety of human leukemias. Here we report the biochemical purification of MLL and demonstrate that it associates with a cohort of proteins shared with the yeast and human SET1 histone methyltransferase complexes, including a homolog of Ash2, another Trx-G group protein. Two other members of the novel MLL complex identified here are host cell factor 1 HCF-1, a transcriptional coregulator, and the related HCF-2, both of which specifically interact with a conserved binding motif in the MLLN p300 subunit of MLL and provide a potential mechanism for regulating its antagonistic transcriptional properties. Menin, a product of the MEN1 tumor suppressor gene, is also a component of the 1-MDa MLL complex. Abrogation of menin expression phenocopies loss of MLL and reveals a critical role for menin in the maintenance of Hox gene expression. Oncogenic mutant forms of MLL retain an ability to interact with menin but not other identified complex components. These studies link the menin tumor suppressor protein with the MLL histone methyltransferase machinery, with implications for Hox gene expression in development and leukemia pathogenesis.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: MLL - CREBBP *

A fragment of the mixed-lineage leukemia MLL gene MII, HRX, ALL-1 was identified in a yeast genetic screen designed to isolate proteins that interact with the CREB-CREB-binding protein CBP complex. When tested for binding to CREB or CBP individually, this MLL fragment interacted directly with CBP, but not with CREB. The transactivation activity of MLL was dependent on CBP, as either adenovirus E1A expression, which inhibits CBP activity, or alteration of MLL residues important for CBP interaction proved effective at inhibiting MLL-mediated transactivation.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Figura 54: *Questionário de utilidade - página 2*

Interação entre genes: MLL - CREBBP *

MLL and CREB bind cooperatively to the nuclear coactivator CREB-binding protein. A fragment of the mixed-lineage leukemia MLL gene MII, HRX, ALL-1 was identified in a yeast genetic screen designed to isolate proteins that interact with the CREB-CREB-binding protein CBP complex. When tested for binding to CREB or CBP individually, this MLL fragment interacted directly with CBP, but not with CREB. In vitro binding experiments refined the minimal region of interaction to amino acids 2829 to 2883 of MLL, a potent transcriptional activation domain, and amino acids 581 to 687 of CBP the CREB-binding or KIX domain. The transactivation activity of MLL was dependent on CBP, as either adenovirus E1A expression, which inhibits CBP activity, or alteration of MLL residues important for CBP interaction proved effective at inhibiting MLL-mediated transactivation. Single amino acid substitutions within the MLL activation domain revealed that five hydrophobic residues, potentially forming a hydrophobic face of an amphipathic helix, were critical for the interaction of MLL with CBP. Using purified components, we found that the MLL activation domain facilitated the binding of CBP to phosphorylated CREB. In contrast with paradigms in which factors compete for limiting quantities of CBP, these results reveal that two distinct transcription factor activation domains can cooperatively target the same motif on CBP.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: MLL - CREBBP *

The transactivation activity of MLL was dependent on CBP, as either adenovirus E1A expression, which inhibits CBP activity, or alteration of MLL residues important for CBP interaction proved effective at inhibiting MLL-mediated transactivation.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: MLL - PPIE *

Loss of MLL PHD finger 3 is necessary for MLL-ENL-induced hematopoietic stem cell immortalization.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: MLL - PPIE *

Loss of MLL PHD finger 3 is necessary for MLL-ENL-induced hematopoietic stem cell immortalization. Reciprocal chromosomal translocations at the MLL gene locus result in expression of novel fusion proteins, such as MLL-ENL, associated with leukemia. Protein interactions of the MLL PHD fingers modulate MLL target gene regulation in human cells.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Figura 55: *Questionário de utilidade - página 3*

Interação entre genes: MLL - PPIE *

Protein interactions of the MLL PHD fingers modulate MLL target gene regulation in human cells. The PHD fingers of the human MLL and Drosophila trx proteins have strong amino acid sequence conservation but their function is unknown. We have determined that these fingers mediate homodimerization and binding of MLL to Cyp33, a nuclear cyclophilin. Overexpression of the Cyp33 protein in leukemia cells results in altered expression of HOX genes that are targets for regulation by MLL. These alterations are suppressed by cyclosporine and are not observed in cell lines that express a mutant MLL protein without PHD fingers. These results suggest that binding of Cyp33 to MLL modulates its effects on the expression of target genes.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: XPA - XAB2 *

Here we report the identification of a novel protein designated XAB2 (XPA-binding protein 2) that was identified by virtue of its ability to interact with XPA, a factor central to both nucleotide excision repair subpathways. In addition to interacting with XPA, immunoprecipitation experiments demonstrated that a fraction of XAB2 is able to interact with the transcription-coupled repair-specific proteins CSA and CSB as well as RNA polymerase II.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: XPA - XAB2 *

Here we report the identification of a novel protein designated XAB2 (XPA-binding protein 2) that was identified by virtue of its ability to interact with XPA, a factor central to both nucleotide excision repair subpathways

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: XPA - XAB2 *

XAB2, a novel tetratricopeptide repeat protein involved in transcription-coupled DNA repair and transcription. Here we report the identification of a novel protein designated XAB2 (XPA-binding protein 2) that was identified by virtue of its ability to interact with XPA, a factor central to both nucleotide excision repair subpathways. The XAB2 protein of 855 amino acids consists mainly of 15 tetratricopeptide repeats. In addition to interacting with XPA, immunoprecipitation experiments demonstrated that a fraction of XAB2 is able to interact with the transcription-coupled repair-specific proteins CSA and CSB as well as RNA polymerase II. Furthermore, antibodies against XAB2 inhibited both transcription-coupled repair and transcription in vivo but not global genome repair when microinjected into living fibroblasts. These results indicate that XAB2 is a novel component involved in transcription-coupled repair and transcription.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Figura 56: *Questionário de utilidade - página 4*

Interação entre genes: CREBB - BRCA1 *

The second BRCT domain of BRCA1 proteins interacts with p53 and stimulates transcription from the p21WAF1/CIP1 promoter. Inherited mutations in the breast and ovarian cancer susceptibility gene BRCA1 are associated with high risk for developing breast and ovarian cancers. Several studies link BRCA1 to transcriptional regulation, DNA repair, apoptosis and growth/tumor suppression. BRCA1 associates with p53 and stimulates transcription in both p53 dependent and p53-independent manners. BRCA1 splice variants BRCA1a (p110) and BRCA1b (p100) associates with CBP/p300 co-activators. Here we show that BRCA1a and BRCA1b proteins stimulate p53-dependent transcription from the p21WAF1/CIP1 promoter. In addition, the C-terminal second BRCA1 (BRCT) domain is sufficient for p53 mediated transactivation of the p21 promoter. We also found that BRCA1a and BRCA1b proteins interact with p53 in vitro and in vivo. The p53 interaction domain of BRCA1a/1b maps, in vitro, to the second BRCT domain (aa 1760-1863). These results demonstrate for the first time the presence of a second p53 interaction domain in BRCA1 proteins and suggests that BRCA1a and BRCA1b proteins, like BRCA1, function as p53 co-activators. This BRCT domain also binds in vitro to CBP. These results suggest that one of the mechanisms by which BRCA1 proteins function is through recruitment of CBP/p300 associated HAT/FAT activity for acetylation of p53 to specific promoters resulting in transcriptional activation.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: CREBBP - BRCA1 *

On the other hand, BRCA1 failed to alter the expression of the CREB binding protein CBP, the structural and functional homologue of p300, in any of these cell types. Ectopic expression of either p300 or CBP "rescued" i.e., reversed the BRCA1 inhibition of ER-alpha activity, whereas two other nuclear receptor coactivators, the p300/CBP-associated factor PCAF and the glucocorticoid receptor-interacting protein-1 GRIP1, failed to rescue the ER-alpha activity. These findings suggest that the cofactors p300 and CBP modulate the ability of the BRCA1 protein to inhibit ER-alpha signaling. CBP/p300 interact with and function as transcriptional coactivators of BRCA1.We now describe that BRCA1-mediated transactivation is enhanced by p300/CBP CREB binding protein and that this effect was suppressed by the adenovirus E1A oncoprotein. We show a physical association of BRCA1 with the transcriptional coactivators/acetyltransferases p300 and CBP. BRCA1 splice variants BRCA1a (p110) and BRCA1b (p100) associates with CBP/p300 co-activators. These results demonstrate for the first time the presence of a second p53 interaction domain in BRCA1 proteins and suggests that BRCA1a and BRCA1b proteins, like BRCA1, function as p53 co-activators. This BRCT domain also binds in vitro to CBP. These results suggest that one of the mechanisms by which BRCA1 proteins function is through recruitment of CBP/p300 associated HAT/FAT activity for acetylation of p53 to specific promoters resulting in transcriptional activation.Earlier experiments have demonstrated that the breast cancer-associated tumor suppressor BRCA1 and the CREB binding protein CBP were associated with the holoenzyme complex.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: CREBB - BRCA1 *

On the other hand, BRCA1 failed to alter the expression of the CREB binding protein CBP, the structural and functional homologue of p300, in any of these cell types. Ectopic expression of either p300 or CBP "rescued" i.e., reversed the BRCA1 inhibition of ER-alpha activity, whereas two other nuclear receptor coactivators, the p300/CBP-associated factor PCAF and the glucocorticoid receptor-interacting protein-1 GRIP1, failed to rescue the ER-alpha activity. BRCA1 splice variants BRCA1a (p110) and BRCA1b (p100) associates with CBP/p300 co-activators. These results demonstrate for the first time the presence of a second p53 interaction domain in BRCA1 proteins and suggests that BRCA1a and BRCA1b proteins, like BRCA1, function as p53 co-activators. Earlier experiments have demonstrated that the breast cancer-associated tumor suppressor BRCA1 and the CREB binding protein CBP were associated with the holoenzyme complex.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Figura 57: *Questionário de utilidade - página 5*

Interação entre genes: BCRA1 - BCRA1 *

Phosphopeptide binding specificities of BRCA1 COOH-terminal BRCT domains.Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1.The structure provides a basis to predict the structural consequences of uncharacterized BRCA1 mutations.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: BRCA1 - BRCA1 *

Phosphopeptide binding specificities of BRCA1 COOH-terminal BRCT domains.Structural analysis of BRCA1 BRCT repeats also predicted conserved residues that may form the phosphopeptide-binding pocket. Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1.Here we determine the crystal structure of the BRCT domain of human BRCA1 at 2.5 A resolution. The structure provides a basis to predict the structural consequences of uncharacterized BRCA1 mutations.The cancer-predisposing mutation C61G disrupts homodimer formation in the NH2-terminal BRCA1 RING finger domain.Analytical gel-filtration chromatography and chemical cross-linking experiments demonstrate that the BRCA1 NH2-terminal domain readily homodimerizes in solution.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Interação entre genes: BRCA1 - BRCA1 *

Phosphopeptide binding specificities of BRCA1 COOH-terminal BRCT domains. The BRCA1 COOH-terminal BRCT domains are protein modules found in many proteins that regulate DNA damage responses Koonin, E. Oriented peptide library analysis indicated that the BRCT domains from BRCA1, MDC1, BARD1, and DNA Ligase IV preferred distinct phosphoserine-containing peptides. In addition, the interaction between BRCA1 and the BRCT binding motif of BACH1 was required for BACH1 checkpoint activity. Structural analysis of BRCA1 BRCT repeats also predicted conserved residues that may form the phosphopeptide-binding pocket.

- Muito ruim
- Ruim
- Apenas aceitável
- Bom
- Muito bom

Comentários

APÊNDICE B – Perfil dos julgadores

Abaixo apresenta-se um breve perfil acadêmico dos julgadores.

- Julgador 1
  - Bacharel em Administração de Empresas
  - Licenciatura (em andamento) em Ciência Biológicas
  - Mestrado (em andamento) em Engenharia de Produção e Sistemas
    - Foco do mestrado: bioinformática, biologia, biotecnologia, mineração de dados, dna
- Julgador 2
  - Especialização Técnica em Biologia Molecular
  - Especialização Lato Sensu em Análises Clínicas
  - Mestre e Doutor em Ciências Médicas
  - Cursando pós-doutorado em Genética
  - Áreas de atuação
    - Genética Humana e Médica
      - expressão gênica
      - células precursoras hematopoéticas
      - células estromais mesenquimais
      - microRNAs
- Julgador 3
  - Bacharel em Informática Biomédica
  - Mestre em Bioinformática
  - Cursando doutorado em Bioinformática
  - Áreas de atuação
    - Informática Biomédica e Bioinformática
      - estrutura de proteínas
      - algoritmos evolutivos
- Julgador 4

–Graduação em Ciências Biológicas

–Mestre em Biologia Celular e Molecular

–Especialista de Laboratório Fisiologia e Biologia Molecular

–Áreas de atuação

∗Bioquímica e Biologia Celular e Molecular